CeGaT

Research &
Pharma Solutions

HiFi WGS De Novo

Bioinformatic Note

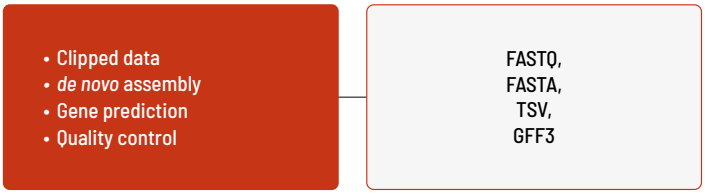
HiFi WGS De Novo

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. With a *de novo* genome assembly, millions of sequenced DNA fragments are computationally pieced together to reconstruct the original sequence from scratch – without a template.

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

In the following, we will focus on the bioinformatic analysis of the HiFi WGS De Novo product:



Only one level of bioinformatic analysis is available for the HiFi WGS De Novo product.

Clipped data

The sequencing reads are demultiplexed. HiFi reads (CCS reads with a predicted accuracy ≥ Q20) are clipped and provided in FASTQ format. Additionally, the quality of the FASTQ files is analyzed.

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 platform. For the demultiplexed data, the number of HiFi reads and bases is reported. The read length and the GC content per read are additionally illustrated in bar plots.

De novo assembly

During the *de novo* assembly, the quality-checked long reads are computationally pieced together to reconstruct the original sequence without an existing template. In this process, overlapping reads are assembled into continuous contigs. Assembled contigs below a defined minimum length are excluded from downstream analyses.

The *de novo* assembled genome is provided as a FASTA file. A FASTA file is a plain-text format used to store DNA, RNA, or protein sequences. It can be opened with a standard text editor. Together with the FASTA file, we provide an assembly summary in TSV format. This assembly summary contains various assembly metrics, such as the number of contigs, the mean read length, the mean quality, the length of the total assembly, and the length of the longest contig. The TSV file consists of three columns, indicating the sample, the metric, and its corresponding value.

In the following, we highlight a few of the assembly summary statistics that are also stated in the project report:

- The **number of contigs** indicates how fragmented the assembly is: fewer contigs generally suggest a more contiguous assembly, whereas many contigs may indicate fragmentation, unresolved repeats, contamination, or insufficient read support.
- The **length of the total assembly** is the total number of bases in all assembled contigs and gives an estimate of the reconstructed genome size. This value can be compared with the expected genome size to assess whether the assembly is broadly plausible.
- The **N50** is a contiguity statistic: it is the contig length at which 50% of the total assembly length is contained in contigs of that size or longer. Higher N50 values usually indicate a more contiguous assembly, although they should always be interpreted together with the total assembly length and the number of contigs.

In a second step, we assess assembly support by mapping the original reads to the newly generated assembly to calculate the assembly's mean depth and breadth.

- The **assembly mean depth** describes the average sequencing depth when the reads are mapped back to the assembly. It reflects how strongly the assembly is supported by the underlying data; very low depth may indicate weak support, while highly uneven depth can suggest collapsed repeats, duplicated regions, or contamination.
- The **assembly breadth** describes the proportion of assembly bases covered by at least one mapped read. High breadth indicates that most of the assembly is supported by the sequencing reads, whereas lower breadth suggests that some assembled regions may be poorly supported or spurious.

Gene prediction

Genes are predicted for the assembled genomes. The gene prediction results are provided in two FASTA files: one for the nucleotide sequence and one for the amino acid sequence, following the already explained principles for FASTA files. Thus, for each predicted gene, the nucleic acid sequence and the protein sequence are reported.

Additionally, the gene prediction results are stored in a GFF3 file. The GFF3 format is the Generic Feature Format Version 3. It is a standard tab-delimited plain-text file for describing genomic features.

It consists of exactly nine specific columns: the name of the chromosome or scaffold ("seqid"); the program or database that generated the feature ("source"); the type of the feature ("type"); the start position and end position of the feature ("start" and "end"); a score as a floating point value ("score"); the strand defined as +/forward or -/reverse ("strand"); the phase defined as 0, 1 or 2, where 0 indicates that the first base of the feature is the first base of a codon, 1 that the second base is the first base of a codon etc. ("phase"); and finally the attributes, which is a semicolon-separated list of tag-value pairs, providing additional information about each feature ("attributes"). An excerpt of a gene prediction GFF3 file can be found in table 1.

Table 1 | Excerpt of the gene prediction file in GFF3 format. (Header line is added for the sake of clarity).

seqid	source	type	start	end	score	strand	phase	attributes
ptg000001c	Prodigal_v2.6.3	CDS	2	808	82.8	-	0	ID=1_1;partial=10;start_type=ATG;rbs_motif=AGGAG; rbs_spacer=5-10bp;gc_cont=0.478;conf=100.00; score=82.81;cscore=62.22;sscore=20.59;rscore=14.48; uscore=0.91;tscore=3.91;
ptg000001c	Prodigal_v2.6.3	CDS	808	4167	181.7	-	0	ID=1_2;partial=00;start_type=TTG;rbs_motif=AGGA; rbs_spacer=5-10bp;gc_cont=0.396;conf=99.99; score=181.70;cscore=184.00;sscore=-2.29;rscore=10.60; uscore=0.63;tscore=-12.88;
ptg000001c	Prodigal_v2.6.3	CDS	4379	6031	156.3	+	0	ID=1_3;partial=00;start_type=ATG;rbs_motif=GGA/GAG/AGG; rbs_spacer=5-10bp;gc_cont=0.381;conf=100.00; score=156.28;cscore=146.45;sscore=9.83;rscore=2.52; uscore=3.39;tscore=3.91;

The number of predicted genes represents the total set of gene models identified in the assembly and provides a general estimate of coding content, but it should be interpreted alongside the assembly's gene content completeness (see below) and the expected gene content of related organisms, rather than as a standalone quality measure.

The predicted genes are compared to a set of expected conserved marker genes to evaluate the completeness of the assembly gene content. The gene matches are categorized as Complete (C) and Single-copy (S), Complete (C) and Duplicate (D), Fragmented (F), or Missing (M):

- (C) and (S): A high proportion of complete gene matches, especially the single-copy counts, indicates that most expected genes are present and covered in the expected copy number.
- (C) and (D): Complete and duplicated matches indicate that a conserved gene was detected more than once, which can reflect true biological duplication but may also suggest assembly redundancy or contamination.
- (F): Fragmented matches indicate that only part of an expected gene could be identified. This scenario is often consistent with assembly fragmentation or incomplete reconstruction of that region.
- (M): Missing marker genes indicate that no convincing match to the expected conserved gene was found and may therefore point to incompleteness of the assembly.

The full table in TSV lists all marker genes and their comparison result. The table contains eight columns: the marker gene id, the status of the marker gene, the sequence where the predicted gene was found, the gene start and end coordinates, the strand, the bit score of the comparison, and the length of the predicted sequence. An excerpt of such a full table can be found in table 2.

A summary of the results of this comparison can be found in the short summaries in TXT and JSON format. The values of the four categories (C and S, C and D, F, and M) are additionally illustrated in a bar plot, as shown in figure 1.

Additionally, a list of the missing marker genes is provided in TSV format. This file contains the identifier of all missing marker genes.

Table 2 | Full table of the comparison between the predicted genes and the set of expected conserved marker genes.

# Busco id	Status	Sequence	Gene Start	Gene End	Strand	Score	Length
5293449at2	Complete	ptg000001c	2825011	2825622	+	176.2	165
9759736at2	Complete	ptg000001c	2104813	2106828	+	540.3	530
9767721at2	Complete	ptg000001c	4965778	4964480	-	481.5	425

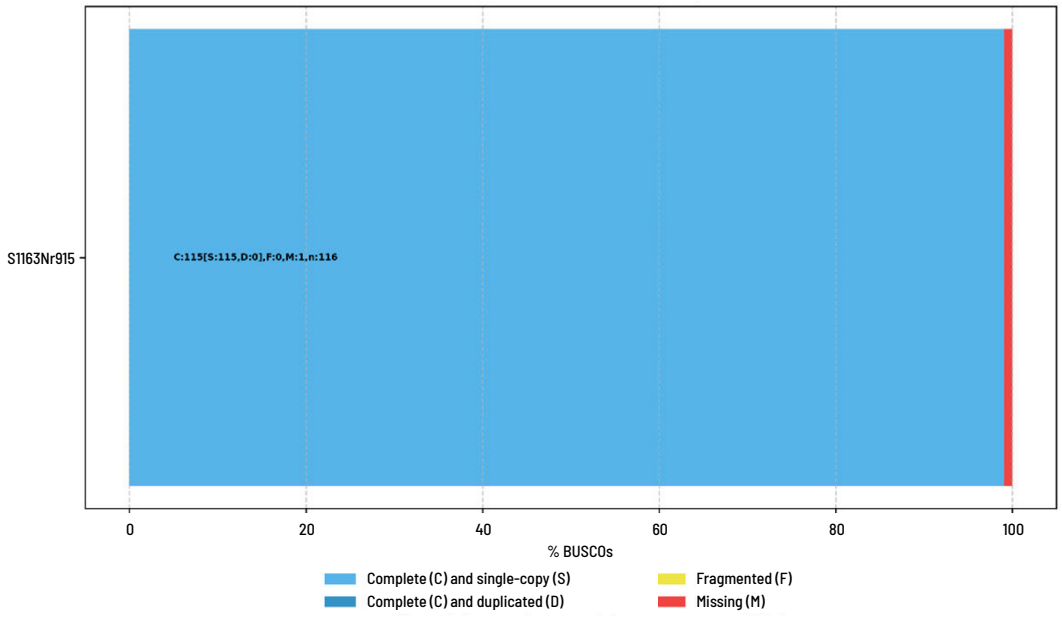


Figure 1 | Comparison results of the predicted genes to a set of expected conserved marker genes.

Quality control

The assembly contiguity and additional quality metrics are evaluated. If a reference genome of a close relative of the assembled organism is available, reference-based statistics can be examined. These statistics help to assess the assembly quality beyond basic contiguity statistics.

Reports in TSV format

Several reports in TSV format are created that contain various metrics for evaluating the genome assembly. The **summary report** includes, amongst others, the number of contigs with a certain length, the total length, indicating the total number of bases in the contigs of a certain length, the length of the largest contig, the total length of the assembly together with its GC content, if provided, the length of the reference genome together with its GC content, and the duplication ratio. The summary report also includes information about misassemblies and unaligned reads.

Selected metrics are also provided in the project report, including

- the **duplication ratio**, which indicates whether regions of the reference are represented more than once in the assembly and can therefore highlight redundancy or over-assembly.
- the **largest alignment** and the **total aligned length**, which describe how much of the assembly aligns to the reference and how large the longest continuous aligned block is. These metrics help to distinguish true reconstruction from extra or poorly matching sequences.
- the **number of misassemblies**, which reflects structural inconsistencies relative to the reference,
- the **number of unaligned contigs** and the unaligned length, which indicate how much of the assembled sequence does not match the reference at all.
- the **NGA50** value, which is a reference-aware contiguity measure that reflects the size of aligned assembly blocks after accounting for structural errors. It is often more informative than N50 when a suitable reference genome is available. More information on related N-values is provided in the section on interactive reports.

For the already mentioned misassemblies, an additional **misassemblies report** in TSV format is provided, not only containing the number of misassemblies as noted in the summary report, but also a classification of all misassembly events into the following three groups: relocations, translocations, and inversions. Additionally, breakpoints containing scaffold-like gaps and breakpoints without those gaps are separately counted.

The same applies to the unaligned reads: in addition to the information in the summary report, a dedicated **unaligned report** in TSV format is provided. In addition to the total number of uncalled bases in the assembly, it also contains information about the number of fully and partially unaligned contigs and their length.

All three report files can be opened with a standard text editor and contain the metric and its corresponding value.

Interactive reports in HTML format

In addition to the TSV files, we provide three interactive reports: the first **report** summarizes the alignment-based statistics, the misassemblies and unaligned information, the per-base quality, the statistics without a reference, and the similarity statistics. Additionally, different plots can be interactively explored, visualizing the cumulative length, the misassemblies, the GC content, and several contiguity measure values, such as the Nx, NAx, NGx, and NGAx plots.

Nx is similarly defined as the N50, but x ranges from 0% to 100% - it is not fixed to 50% as for the N50 value. NGx is only provided if a reference genome is available. It is similarly defined as Nx but refers to the reference genome: It represents the length of the shortest contig at which longer or equal-length sequences cover at least 50% of the reference genome. For the NAx and NGAx values, aligned blocks, rather than contigs, are considered. Aligned blocks are obtained by breaking contigs and misassembly events and removing all unaligned bases. The corresponding plots are only generated if a reference genome is available.

In addition to the plots in the interactive report, we provide the cumulative length plot of the aligned contigs in PDF format. This plot shows the growth of the lengths of the aligned blocks. On the x-axis, blocks are ordered from largest to smallest, the y-axis gives the size of the x largest aligned blocks. This plot is only created if a reference genome is provided. Furthermore, the GC content plot, the NAx plot, and the NGAx plot are provided as additional files in PDF format.

The second interactive report is the **contig alignment viewer.** The viewer aligns contigs according to their mapping to the reference genome.

The third interactive report is the **contig size viewer**. In this viewer, the contigs are ordered from longest to shortest, facilitating the comparison of only large contigs or a number of contigs longer than a specific threshold. The N50 value is also indicated in the visualization.

The latter two interactive viewer reports are only provided if a reference genome is available. All three reports in HTML format can be opened with a web browser to interactively explore and evaluate the assembly result.

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 bioinformaticians 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 according to DIN EN ISO/IEC 17025

For more details please visit
www.cegat.com/transcriptome-sequencing

2026/08