# Genomics

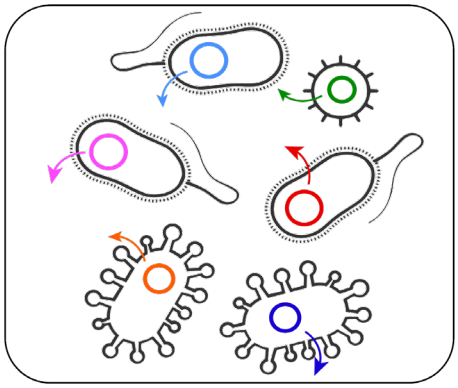
Do you want **state-of-the-art** Genomics?
Do you want a workflow **tailored to your specific bioinformatic requirements**?

- DNASense uses cutting-edge workflows for generating highly contiguous reference-grade assemblies.
- DNASense offers a range of bioinformatic add-on services, including variant analysis, core genome SNP analysis, functional annotation, and functional enrichment analysis.
- Fast-track (≤ 7 days turn-around time) and ultra fast-track (≤ 3 days turn-around time) options are available.

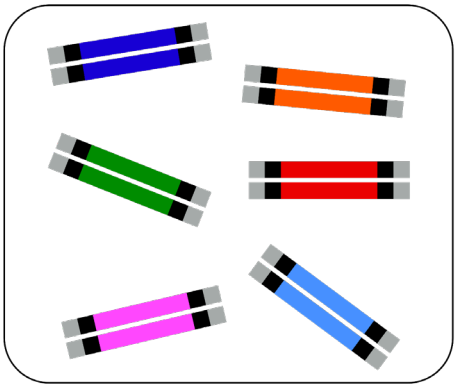
**DNASense** provides a complete **sample-to-answer** service for genomics and an endless list of customizable bioinformatic add-on options

## State-of-the-art workflow

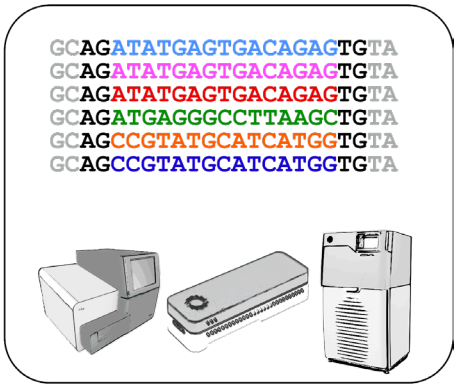
DNA extraction



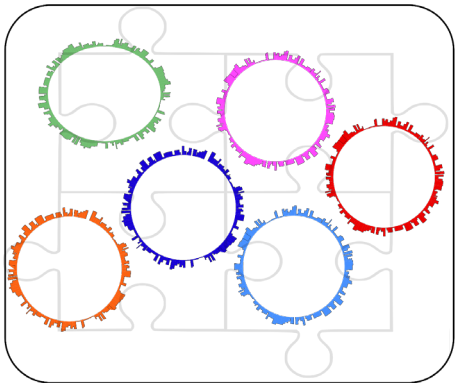
Library preparation



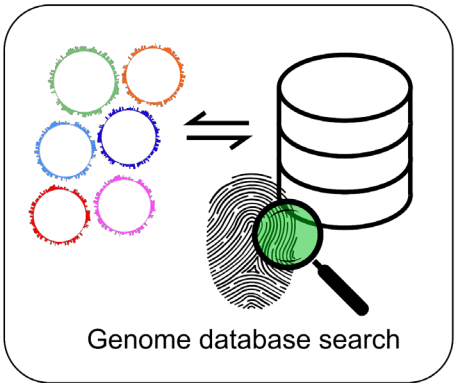
DNA sequencing



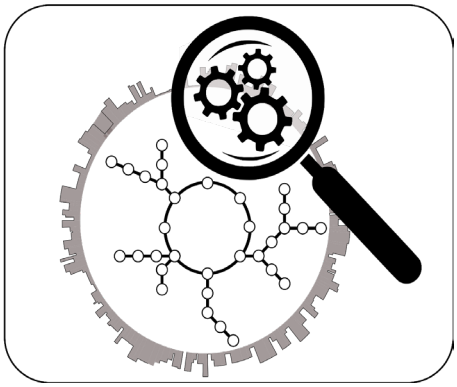

*De novo* assembly



Taxonomy assignment




Annotation / Downstream analysis



## Customized solutions

**Our standard package includes:** Optional pre- and post-project meeting with a DNASense specialist, DNA extraction, library preparation, sequencing, pre- and post-sequencing quality control, de novo assembly, taxonomic profiling, gene annotation, rDNA extraction, access to raw data, result files and a detailed project report.

**Add-on services (non-exhaustive list):** Customized DNA extraction and purification, SNV/SV analysis, core genome SNP analysis, core genome MLST analysis, Functional annotation (KO, GO and KEGG), functional enrichment analysis, manual curation of metabolic pathways, gene mining, custom annotation and data submission.


info@dnasense.com | +45 71 99 20 20

# Working with the DNASense team

- Extensive experience from hundreds of projects and challenging samples
- Detailed documentation and full method transparency
- State-of-the-art sample preparation, DNA sequencing and bioinformatics
- Extensive expert consultant services

## Encompassing report with actionable results



### 3 Results

#### 3.1 DNA sequencing

Table 2 shows the outcome of the DNA sequencing in terms of data yield, median read quality scores, and read N50. The quality score was *on par* with the raw read accuracy supported by the current kit chemistry (V14 kit), flow cell (R10.4.1), and basecalling algorithms (further details in Materials and Methods section). All samples generated more than 100x genome data depth or 300 megabases (Mb) (assuming a *Flavobacterium psychrophilum* genome size of 3.0 Mb) which is the R10.4.1 data coverage recommended by DNASense for downstream *de novo* assembly and genome polishing workflows.

**Table 2:** ***Sequencing statistics***. ***Samplename*** *denotes customer-assigned sample identification nomenclature.* ***Sequencing ID*** *denotes the DNASense-assigned barcode identification number.* ***Data (Mbp)*** *and* ***Reads*** *denote the sequencing data yield in total basepairs and number of reads, respectively. The* ***Read N50 (bp)*** *value denotes that half of the data is contained within reads of length N50 or greater. The* ***q-score*** *denotes the median Phred scaled read quality score.*

| | | Raw | | | Trimmed | | | |
|---|---|---|---|---|---|---|---|---|
| Samplename | Sequencing ID | Data (Mbp) | Read N50 (bp) | q-score | Reads | Data (Mbp) | Read N50 (bp) | q-score |
| SF00A1 | barcode01 | 856 | 7734 | 18.0 | 141388 | 780 | 7737 | 19.3 |
| SF00A2 | barcode02 | 562 | 7692 | 17.9 | 91947 | 511 | 7670 | 19.0 |
| SF00A3 | barcode03 | 696 | 6516 | 17.7 | 128981 | 628 | 6479 | 19.0 |
| SF00A4 | barcode04 | 550 | 8444 | 17.8 | 84515 | 499 | 8442 | 19.0 |
| SF00A5 | barcode05 | 1231 | 7381 | 18.0 | 211121 | 1117 | 7370 | 19.2 |
| SF00A6 | barcode06 | 855 | 6434 | 18.0 | 158724 | 777 | 6418 | 19.3 |
| SF00A7 | barcode07 | 432 | 9321 | 18.0 | 63187 | 394 | 9362 | 19.1 |
| SF00A8 | barcode08 | 584 | 7968 | 17.8 | 94952 | 531 | 7956 | 19.0 |

#### 3.2 *de novo* assembly

The starting point for downstream analyses is a state-of-the-art *de novo* assembly workflow, in which DNA long reads are used to assemble the genome(s) present within each sample. Ideally, the workflow generates a single circular chromosome and (if present) extra-chromosomal elements such as circular plasmids (see figure 1) or viral DNA phages. These assembled elements or contigs may sometimes reflect multiple (co-assembled) genomes and numerous contigs, e.g. if the sample has been contaminated. The contaminating genomes or contigs can be removed prior to analysis if the overall assembly is of sufficient quality and contiguity.





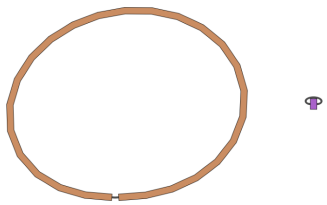


**Figure 1:** ***SF00A1 - assembly graph***. *Draft assembly graph resulting from the draft de novo assembly of SF00A1, highlighting assembly content and contiguity. The large enclosed circle represents a fully-closed genome. Colored lines denote size-scaled contiguous DNA elements. Line thickness is proportional to contig coverage. Thin black lines denote connectivity between segments. All draft assembly graphs are available from the misc folder and in the Supplementary information section.*

Tables 3-4 show the outcome of the *de novo* assembly workflow, which generated fully-closed genomes (i.e circular). A total of 0 out of the 8 (0 %) samples were contaminated above a 5 % contamination threshold (adopted MIMAG criteria (Bowers et al., 2017)). Specifically, a given prokaryotic lineage can be associated with a specific set of single-copy and lineage-specific marker genes. If a given sample contains the entire set of single-copy marker genes, the sample-associated genome is considered complete (100 %). If no additional marker genes are located, the genome is said to have 0 % contamination. If more single-copy marker genes than expected from classification to a single organism are found, the genome is said to be contaminated in a ratio proportional to the number of additionally identified marker genes. For substantially complete genomes (≥70 – 90 %) with medium contamination (5 % - 10 %), completeness and contamination estimates generally have an absolute error of ≤6 %, and the error in the quality estimates tends to decrease as the quality of a genome improves (Parks et al., 2015). The concept of strain heterogeneity is used to indicate the relatedness of an isolate contaminant based on the identity of the duplicate marker genes. High strain heterogeneity suggests that the majority of the reported contamination is from one or more closely related organisms, potentially the same species, while low strain heterogeneity suggests the majority of contamination is from more phylogenetic diverse sources.





**Table 3:** ***Genome QC***. ***Samplename*** *denotes the customer-assigned sample nomenclature.* ***Completion (%)*** *is the estimated genome completeness based on the presence or absence of essential lineage-specific marker genes.* ***Contamination (%)*** *is the estimated contamination based on the presence of multiple single-copy marker genes.* ***Strain heterogeneity (%)*** *is the estimated strain heterogeneity as determined from the number of multi-copy marker pairs which exceed a specified amino acid identity threshold.* ***GTDB taxonomy*** *refers to the GTDB (Genome Taxonomy Database) classification at the highest taxonomic resolution assigned to a specific genome.*

| Sample name | Completion (%) | Contamination (%) | Strain heterogeneity (%) | GTDB taxonomy |
|---|---|---|---|---|
| SF00A1 | 99.12 | 0.44 | 0 | s__Flavobacterium psychrophilum |
| SF00A2 | 99.29 | 0.44 | 0 | s__Flavobacterium psychrophilum |
| SF00A3 | 99.29 | 0.44 | 0 | s__Flavobacterium psychrophilum |
| SF00A4 | 98.82 | 0.71 | 0 | s__Flavobacterium psychrophilum |
| SF00A5 | 99.29 | 0.44 | 0 | s__Flavobacterium psychrophilum |
| SF00A6 | 99.29 | 0.44 | 0 | s__Flavobacterium psychrophilum |
| SF00A7 | 98.92 | 0.71 | 0 | s__Flavobacterium psychrophilum |
| SF00A8 | 98.20 | 0.71 | 0 | s__Flavobacterium psychrophilum |

**Table 4:** ***Basic genome assembly statistics***. ***Samplename*** *denotes the customer-assigned sample nomenclature.* ***Genome size (Mbp)*** *and* ***Contigs*** *denote the size (in megabasepairs) and the number of contiguous DNA elements associated with each genome assembly, respectively.* ***GC content (%)*** *and* ***Prokka CDS*** *denote the mean content of the G- and C-nucleotides and the number of Prokka-identified coding sequences in the assembled genome, respectively.* ***16S rRNA***, ***23S rRNA*** *and* ***5S rRNA*** *denote the number of different ribosomal RNA genes identified by Barrnap.*

| Sample name | Genome size (Mbp) | Contigs | GC content (%) | Prokka CDS | 16S rRNA | 23S rRNA | 5S rRNA |
|---|---|---|---|---|---|---|---|
| SF00A1 | 2.80 | 2 | 32.5 | 2466 | 6 | 6 | 6 |
| SF00A2 | 2.93 | 2 | 32.5 | 2576 | 6 | 6 | 6 |
| SF00A3 | 2.96 | 5 | 32.5 | 2607 | 6 | 6 | 6 |
| SF00A4 | 2.81 | 1 | 32.8 | 2534 | 6 | 6 | 6 |
| SF00A5 | 2.82 | 2 | 32.6 | 2456 | 6 | 6 | 6 |
| SF00A6 | 2.89 | 3 | 32.5 | 2531 | 6 | 6 | 6 |
| SF00A7 | 2.81 | 1 | 32.7 | 2510 | 6 | 6 | 6 |
| SF00A8 | 2.86 | 1 | 32.7 | 2586 | 6 | 6 | 6 |

#### 3.3 Core genome SNP analysis

A cgSNP analysis can be conducted to infer the relatedness of strains (i.e. same species). The core genome is defined as DNA regions that are present across all isolates. A cgSNP analysis reflects an analysis of the number of identified SNPs with respect to the chosen reference, i.e. across a reference-based alignment of the core genomic region (see figure 2). The size of the analysed core genome depends on the strain diversity, i.e. the inclusion of highly diverse strains will result in a smaller core genome and less SNP positions analysed. Therefore, for a more detailed analysis specific clades can be subjected to individual cgSNP analyses.

A phylogenomic tree reflecting a cgSNP analysis is presented in figure 3 (hierarchial clustering). *SF00A1* has (arbitrarily) been set as the reference (i.e. SNPs are reported with respect to SF00A1).



## Price example*

| Service | Analysis | Sample fee (pr. isolate) | Fast-track fee | Turn-around-time** | 24 isolate price example |
|---|---|---|---|---|---|
| Normal | 1000 EUR | 250 EUR | 0 EUR | ≤ 15 days | 7000 EUR |
| Fast-track | 1000 EUR | 250 EUR | 1150 EUR | ≤ 5 days | 8150 EUR |
| Ultra fast-track | 1000 EUR | 250 EUR | 2000 EUR | ≤ 3 days | 9000 EUR |

*Prices assume that isolates are pure culture isolates (~ 500 Mbp/sample). ** Working days

**Contact us today at**
**info@DNASense.com +45 7199 2020**

Founded in 2014, DNASense's mission is to make cutting-edge DNA sequencing and bioinformatics readily available to life science research and industry