DRY LAB PROTOCOL – METAGENOMICS SEQUENCING AND DATA ANALYSIS

Metagenomics sequencing

Metagenomic DNA sequencing of the sewage samples is outsourced to an external partner/company.

After DNA extraction and purification at DTU, aliquots of 100+ ng Qubit-quantified metagenomic DNA is sent on dry ice for sequencing.

The method and kit used for library preparation is dependent on the intended sequencing platform, but a PCR-free kit should always be used when sequencing sewage. The currently used kit is the PCR-free KAPA HyperPrep Kit. DNA is sheared, targeting a fragment size of 350 bp. The choice of flow cell and sequencing platform depends on the specific number of samples sequenced, as well as the amount of data coverage generated per sample.

At the current time, sequencing is performed on the Illumina NovaSeq instrument with S2 or S4 flow cells using a paired-end 2x150 cycle, yielding at least 35,000,000 read pairs per sample. Protocols of sequencing depends on the chosen platform, but can be found on the Illumina website.

After sequencing is complete, de-multiplexed FASTQ files stored on the sequence providers webpage are downloaded. If the name of the datasets delivered by the sequence provider is different than the original sample names, a *samplekey.txt* document supplied by sequence provider is used to remap the datasets back to the original sample names.

Assessment of sequence data quality

The first step after downloading sequence data is to assess data quality. This is essential to guarantee quality of downstream analysis, as poorly sequenced samples can lead to errors in AMR prediction where pathogens, genes or mutations may fail to be detected.

A standard, minimum of 35,000,000 paired-end reads with a length of 150 base pairs must be delivered per sample, when sequencing on Illumina NovaSeq/HiSeq machines. Any samples below this threshold must be re-sequenced by the provider, and subsequently undergo quality assessment again. Exceptions are negative controls, as well as samples for which less than the required *100 ng* of metagenomic DNA was shipped to the sequence provider.

Quality control procedures include filtration of sequencing artefacts such as low-quality reads, sequencing adapters, short reads and potentially contaminating reads. The QC-tool used takes the raw paired-end sequence data as input and trims data utilizing BBduk2. This includes removing common adaptor sequences

and trimming low quality data off the 3'-ends of the reads to a sliding window using a Phred-score of Q20, corresponding to a 1% error rate.

In addition to trimming, the QC-tool produces QC reports using FastQC both pre- and post-trimming. Illumina adaptor sequences are identified using a curated adaptor sequence database contained as part of the tool. As a standard, the pipeline is run with default parameters to ensure reproducibility.

The output of the QC-tool is trimmed data (forward-, reverse- and singleton reads) as well as FastQC reports for all datasets before and after the trimming step.

For simultaneous QC inspection of multiple samples, a multiQC-tool produces aggregated reports for all samples, which are inspected to verify data quality both pre- and post-trim, and the reports along with the trimmed data (forward-, reverse- and singleton reads) are stored.

Only datasets that pass the QC demands are used in subsequent analysis. Specific demands depend on the data, examples of low-quality data could include:

- Low average nucleotide PHRED score
- High sequencing error rate
- High amounts of adapter contamination (i.e. significant loss of data post-trim)

Bioinformatics analysis

Most bioinformatics analysis of samples in the Global Sewage Surveillance Project is performed on the Danish National Life Science Supercomputer, Computerome2.

Due to the nature of research, no protocol can span all the scientific methods, visualizations or bioinformatics approaches applied in the exploration and interpretation of data. As such the data analysis and interpretation section of this protocol is intentionally left open, and the sections included are merely some examples of how initial analysis may be performed.

Data sharing on BitBucket

Both data products as well as sample metadata can be found and shared with partners on the dedicated Global Sewage BitBucket repository. Each user of the repository may create her/his own branch of the repository in which they are free to work on the data available.

Mapping sequence data against reference databases

Datasets passing QC demands move on to the bioinformatics analysis, where the trimmed paired-end and singleton reads are used as input to the reference-based mapping and taxonomy-assignment tool KMA (**link 30**). KMA is the preferred

mapping software due to its performance when mapping short reads against highly redundant databases, such as the resistance gene database ResFinder.

Running on Computerome2, KMA is used to map forward-, reverse- and singleton reads against a selected set of reference databases that include but are not limited to:

- ResFinder/PointFinder (acquired and intrinsic antibiotic resistance)
- Silva (small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life

When mapping multiple sewage sample datasets against one or several databases, KMA can be called using the KMA_wrapper tool, which simplifies the amount of commands to be run. KMA_wrapper will then call KMA for all samples, run KMA with default parameters. Documentation of KMA_wrapper and calls to KMA can be found online.

For each sample-database combination, KMA produces several output. For most applications, the mapping statistics summary file format dubbed "*.mapstat*" is required. The *.mapstat* files contain mapped *fragmentCount* (reads that mapped to a reference sequence) and aligned *fragmentCountAln* (reads that aligned to a reference sequence). As a standard, *fragmentCountAln* is used when determining abundance of sequences from gene catalogues such as the Silva or CGE pipeline ResFinder, while *fragmentCount* is used when determining taxonomic assignment abundance to larger databases containing full genomes.

For taxonomical annotation of *.mapstat* files, the tool *mapstat2refdata* can be used to produce taxonomy-annotated *.refdata* files for the batch of samples that were mapped with KMA. To obtain database abundance summaries, *mapstat2overview* can be used.

Data transformation
Utilizing *.refdata* and *.mapstat* files, a selection of different data transformations can be performed, such as centered-log ratio transformation (CLR), fragments per kilobase million (FPKM) and transcripts per kilobase million (TPM), allowing for adjustments being made based on various parameters such as differences in gene lengths, bacterial sequence abundance or sequencing depth. These transformations can be done with the Mapstat2Abundancetable.

Abundance table products and thereof derived plots may serve as a starting point for the analysis, but all results arising thereof should be further examined in the subsequent main part of the bioinformatics analysis.

The following data products may be used for the diagnostic and descriptive plots mentioned in the sections about **Verifying data** and **Reporting Data**:

- Resistome composition:
  CLR-transformed fragmentCountAln abundance of homology reduced* resistance gene clusters, as well as genes aggregated into classes, based on ResFinder KMA mapping results

- AMR/bacterial sample load:
  CLR-transformed total fragmentCountAln abundance of resistance genes (ResFinder) and 16S bacterial sequences (Silva) over total fragments per sample, i.e. abundance of AMR/bacterial sequences relative to sequencing depth

- Relative AMR sample load:
  ALR-transformed fragmentCountAln abundance of AMR genes/classes over bacterial 16S fragmentCount abundance, i.e. abundance of resistance genes relative to bacterial load

* Reduced referring to fragment counts that have been gene-length corrected, homology-reduced or have undergone other types of standard data adjustments. For homology reduction, the tool CD-HIT-EST can be used to aggregate resistance gene abundances to 90% gene identity clusters.

The sewage samples contain variable proportions of non-bacterial DNA, which will be included in KMA mapping results to Silva. The number of sequence fragments assigned to individual ARGs should therefore only be related to the number of ARG-assigned fragments quantified through the bacteria-assigned hits of the Silva database.

Verifying data - diagnostic plots
Exploratory verification of mapping results using KMA may be conducted by producing a series of diagnostic plots, with the main objective of identifying sample outliers as well as general signs of data quality issues.

These plots below may be created from the data products mentioned in the section "Data transformation" above. Both raw fragment count abundance as well as adjusted abundances of taxonomies and AMR derived from *.mapstat* files are used to produce the following visualizations:

- Sequencing depth:
  Total number of raw fragments per sample, sorted according to sample size

- Sample AMR load:
  CLR-transformed abundances of resistance genes (ResFinder) vs. 16S bacterial sequences (Silva)

- Sample bacterial load:

CLR-transformed abundances of 16S bacterial sequences (Silva) vs. 18S eucaryotic sequences (Silva)

Reporting data - descriptive plots

With the aim of producing descriptive plots of KMA results to gain an initial exploration of the data, the following plots can be produced from the data mentioned in section "Data transformation":

- Resistome compositions:
  Stacked bar plots visualizing sample resistome composition on gene and class level, based on CLR-transformed fragmentCountAln values.

- Resistome PCA:
  A principal components plot (PCA) of sample resistomes (gene and class level), showing sample clustering based on CLR-transformed fragmentCountAln values.

- Diversity indicators:
  Alpha diversity and richness plots based on CLR-transformed values of 16S and ResFinder gene abundance.

Additional bioinformatics analysis

Additional bioinformatics analysis may be conducted with a range of other tools and pipelines, many of which are developed by CGE. Such analyses could include draft genome construction via metagenomics assembly software such as metaSPAdes (as performed in the latest Global Sewage report manuscript 2021), extraction of flanking regions of ARGs using BEDtools, sub-typing bacteria, cluster analysis, MLST/cgMLST determination, serotyping, MGE identification, identification of virulence factors and AMR-carrying plasmids, and phylogenetic SNP analysis. The recommended CGE tools for these types of analysis can be found on the CGE website.