D-JRP15-FED-AMR-WP2.1

Bioinformatic analysis of shotgun metagenomic sequences

Version 1
September /2020

Table of Contents

Bioinformatic analysis of shotgun metagenomic sequences

A. Description

This protocol is for the analysis of DNA sequence data from the shotgun sequencing of soil, water, manure and faeces samples from the FED-AMR consortium. This protocol does not cover microbiome analysis of 16S rRNA gene sequencing. The protocols assume that the sequencing will be carried out using Illumina sequencing technologies (HiSeq, NovaSeq) and it is applicable to samples sequenced with or without gene enrichment. When performed gene enrichment will be carried out by an external Company using ARESdb (1).

B. Computing resources requirements

The large number of samples will require very large storage capacity, large memories and processors speeds: storage: > 10 TB; memory: 128 GB RAM; processors: at least 24 threads.

C. Linking metadata with sequence data

Sequence data files should be linked to all metadata collected (location, GPS coordinates, date, qPCR data, heavy metal and other contaminants data). This should be done through a spreadsheet in which DNA sequence file names are linked to identifiers and the relevant sample metadata.

D. Processing FastQ files (raw Illumina sequence data)

Prior to processing, Illumina adaptors will be removed using appropriate tools e.g. Cutadapt (2). Individual reads with phred score ≤ 20 and length ≤ 70 will be removed using tools such as NGS QC Toolkit and/or Sickle (3). Overlapping paired end reads will be merged using SeqPrep or similar tools. Sequences with more than 10% undetermined nucleotides will be removed using Trimmomatic (4).

E. Determining abundance of AMR genes in metagenomes

1. Standard shotgun metagenomic pipeline

i. Metagenome co-assembly

The merged and trimmed sequences will be assembled using sequences from all samples with MEGAHITassembler (5) with a minimum contig length of 1000 bp. Metagenome assemblies will be assessed with MetaQUAST(6). Metagenome co-assembly will generate longer contigs and possibly Metagenome-Assembled Genomes (MAGs). The longer contigs/MAGs will allow characterisation of AMR gene flanking regions, which may shed light on the molecular basis of AMR spread, i.e. possible role of different types of mobile elements in the horizontal gene transfer via exDNA.

ii. Coding sequence (CDS) and Open Reading Frame (ORF) detection and annotation

Gene prediction will be carried out using Prodigal (7) using default parameters for metagenomic analyses. Predicted genes will be compared against AMR gene databases, and sequences with strong matches to AMR genes (*e* value of 10^{-100}, 80% coverage and 95% identity match to reference gene) will be considered as matches to AMR genes. The remaining (non-AMR gene) predicted coding sequences will be translated to proteins, and these will be characterised by searches against InterPro and Gene Ontology (GO) databases using blastp.

iii. Quantification of AMR genes in shotgun metagenomes

Abundance of AMR or other genes of interest will be determined by mapping individual reads to contigs as determined above. Mapping of reads against the contigs will be carried out using the bwt tool from the Resistant Gene Identified (RGI) package (https://github.com/arpcard/rgi). The abundance of AMR sequence reads in each sample will be determined using the "featureCount" algorithm of the Subread package (8). AMR reads with ambiguous mapping (i.e., reads that map to more than one contig) will be discarded from the count.

2. AMR gene abundance and diversity determination by gene allele network analysis

Metagenomic datasets can contain a large number of ambiguous reads (i.e. sequence reads that match more than one reference gene). In order to circumvent this issue, AMR genes can also be analysed using a gene allele network as outlined by Lanza et al (2018).

i. Mapping genes to allele network

Individual high-quality merged reads will be mapped against AMR gene databases using Bowtie2 (9) or similar tools. The outputs from Bowtie2 will be used to generate the following indices: 1) the number of reads per gene mapped to an AMR gene (RPG, or gene depth coverage); 2) the number of reads per kb of gene ("RPK"); 3) the number of reads mapped unequivocally to a given AMR gene (unique matches); 4) the percentage of coverage of the gene sequence. The RPG, RPK and "unique" indices will be normalised against the total number of reads in each sample to generate RPGM (reads per gene per million reads), RPKM (reads per kb of gene per million reads) and "unique" gene reads per million reads. This normalisation step is introduced in order to evaluate the proportion of AMR genes among samples per unit of sample DNA (i.e. total number of reads).

ii. Quantitative analysis

Following the bioinformatic pipeline outlined previously (10), a gene allele network will be generated in order to allow quantitative analysis of AMR gene abundances in each sample. The gene allele network is represented by the nodes (AMR genes) and the network edges (the number of reads mapping to two or more genes). Clusters of nodes in the allele network will form the "mapping gene clusters" (MGC), which will be used to determine the changes in abundance of each AMR gene among different samples. The abundance of each AMR gene/allele in each sample will be set as the highest value observed by a node in each MGC. The allele abundance value will then be used to perform differential abundance analysis using DESeq2 (11).

F. Taxonomic analysis of metagenomic reads

The taxonomic affiliation of AMR and non-AMR genes will be determined by aligning reads to the NCBI RefSeq database (12), using tools such as Diamond (13), LAST (14) and FASTLSA (15). Alignment results will be analysed using the Lowest Common Ancestor (LCA) algorithm of the MEGAN6 software package (Huson et al., 2007), using a minimum score of 50 and maximum e-value of 0.001.

G. Reference list

1. Ferreira I, Beisken S, Lueftinger L, Weinmaier T, Klein M, Bacher J, Patel R, von Haeseler A, Posch AE. Species identification and antibiotic resistance prediction by analysis of whole-genome sequence data by use of ARESdb: an analysis of isolates from the Unyvero lower respiratory tract infection trial. J Clin Microbiol, 2020. 58:e00273- 20. https://doi.org/10.1128/JCM.00273-20

2. Martin M. Cutadapt removes adapter sequences from high-throughput

sequencing reads. EMBnet Journal. 2011;17(1):10–2.

3. Joshi N, Fass J. Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files. Available at https://github.com/najoshi/sickle. 2011.

3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014; 30:2114-20. doi: 10.1093/bioinformatics/btu170

5. Li D, Liu C, Luo R, Sadakane K, Lam T. Sequence analysis MEGAHIT : an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 2015; 31:1674–76. doi: 10.1093/bioinformatics/btv033

6. Mikheenko A, Saveliev V, Gurevich A. Genome analysis MetaQUAST : evaluation of metagenome assemblies. Bioinformatics, 2016; 32:1088–90. doi: 10.1093/bioinformatics/btv697

7. Hyatt D, Chen G, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal : prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 2010; 11:119. doi: 10.1186/1471-2105-11-119

8. Liao Y, Smyth GK, Shi W. The Subread aligner : fast , accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 2013; 41:e108. doi: 10.1093/nar/gkt214

9. Pongor LS, Vera R, Ligeti B. Fast and sensitive alignment of microbial whole genome sequencing reads to large sequence datasets on a desktop PC: application to metagenomic datasets and pathogen identification. PLoS One. 2014; 9:e103441. doi: 10.1371/journal.pone.0103441

10. Lanza VF, Baquero F, Martínez JL, Ramos-Ruíz R, González-Zorn B, Andremont A, et al. In-depth resistome analysis by targeted metagenomics. Microbiome. 2018;6:1–14. doi: 10.1186/s40168-017-0387-y

11. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15:550. doi: 10.1186/s13059-014-0550-8

12. Leary NAO, Wright MW, Brister JR, Ciufo S, Haddad D, Mcveigh R, et al. Reference sequence (RefSeq) database at NCBI: current status , taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016; 44:D733–D745. doi: 10.1093/nar/gkv1189

13. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods. 2015;12:59–60.doi: 10.1038/nmeth.3176

14. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Research. 2011; 21:487–93. doi: doi/10.1101/gr.113985.110

15. Durno WE, Hanson NW, Konwar KM, Hallam SJ. Expanding the boundaries of local similarity analysis. BMC Genomics. 2013;14(Suppl 1):1–14. doi: 10.1186/1471-2164-14-S1-S3