CHAPTER 1

EDGE ABCs

A quick About EDGE, overview of the Bioinformatic workflows, and the Computational environment

1.1 About EDGE Bioinformatics

EDGE bioinformatics was **developed to help biologists process Next Generation Sequencing data** (in the form of **raw FASTQ** files), even if they have little to no bioinformatics expertise. EDGE is a **highly integrated and interactive web-based platform** that is capable of running many of the standard analyses that biologists require for viral, bacterial/archaeal, and metagenomic samples. EDGE provides the following analytical workflows: **pre-processing, assembly and annotation, reference-based analysis, taxonomy classification, phylogenetic analysis, Gene Family Analysis, PCR analysis, Qiime2 amplicon data analysis, targeted sequencing adjudication and RNA-Seq analysis**. EDGE provides an intuitive web-based interface for user input, allows users to visualize and interact with selected results (e.g. JBrowse genome browser), and generates a final detailed PDF report. Results in the form of tables, text files, graphic files, and PDFs can be downloaded. A user management system allows tracking of an individual's EDGE runs, along with the ability to share, post publicly, delete, or archive their results.

While EDGE was intentionally designed to be as simple as possible for the user, there is still no single 'tool' or algorithm that fits all use-cases in the bioinformatics field. Our intent is to provide a detailed panoramic view of your sample from various analytical standpoints, but users are encouraged to have some knowledge of how each tool/algorithm workflow functions, and some insight into how the results should best be interpreted.

1.2 Bioinformatics overview

1.2.1 Inputs:

The input to the EDGE workflows begins with one or more **Illumina FASTQ files** for a single sample. (There is currently limited capability of incorporating PacBio and Oxford Nanopore data into the Assembly module) The user can also enter SRA/ENA accessions to allow processing of publicly available datasets. Comparison among samples is not yet supported but development is underway to accommodate such a function for assembly and taxonomy profile comparisons.

1.2.2 Workflows:

Pre-Processing

Assessment of quality control is performed by FAQCS. Users can optionally find and remove adapters from Oxford Nanopore reads using Porechop. In addition, users can optionally stitch paired-end(PE) reads using fastq-join and use joined PE reads for downstream analysis. The host removal step requires the input of one or more reference genomes as FASTA file(s). Several common references are available for selection. Trimmed and host-screened FASTQ files are used for input to the other workflows.

Assembly and Annotation

We provide the IDBA, Spades, MegaHit for Illumina reads, LRASM includes miniasm and wtdbg2 algorithm and (meta)flye for PacBio/Nanopore reads, and Unicycler for bacteria genomes hybrid assembly. These assembly tools can accommodate a range of sample types and data sizes. When the user chooses to perform an assembly, all subsequent workflows can execute the analysis with either the reads, the contigs, or both (default). For annotation, Prokka and RATT are provided for *ab initio* or transfer annotation from a closely-related reference genome. Starting from version 2.4, EDGE uses antiSMASH v4.1.0 for the rapid genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes. In addition, the assembled contigs can be binned by Maxbin2 and the quality of the binning result can be assessed by CheckM .

Reference-Based Analysis

For comparative reference-based analysis with reads and/or contigs, users must input one or more references (as FASTA or multi-FASTA files if there is more than one replicon) and/or select from a drop-down list of RefSeq complete genomes. Results include lists of missing regions (gaps), inserted regions (with input contigs if assembly was performed), SNPs (and coding sequence changes with GenBank information), as well as genome coverage plots and interactive access via JBrowse. There is an option to output the consensus FASTA file from the mapping result.

Taxonomy Classification

For taxonomy classification with reads, multiple tools are used and the results are summarized in heat map and radar plots. Individual tool results are also presented with taxonomy dendograms and Krona plots. Contig classification occurs by assigning taxonomies to all possible portions of contigs. For each contig, the longest and best match (using minimap2) is kept for any region within the contig and the region covered is assigned to the taxonomy of the hit. The next best match to a region of the contig not covered by prior hits is then assigned to that taxonomy. The contig results can be viewed by length of assembly coverage per taxa or by number of contigs per taxa.

Phylogenetic Analysis

For phylogenetic analysis, the user must select datasets from near neighbor isolates for which the user desires a phylogeny. A minimum of two additional datasets are required to draw a tree. At least one dataset must be an assembly or complete genome. RefSeq genomes (Bacteria, Archaea, Viruses) are available from a dropdown menu, SRA and FASTA entries are allowed, and previously built databases for some select groups of bacteria are provided. This workflow (see PhaME) is a whole genome SNP-based analysis that uses one reference assembly to which both reads and contigs are mapped. Because this analysis is based on read alignments and/or contig alignments to the reference genome(s), we **strongly recommend only selecting genomes that can be adequately aligned at the nucleotide level (i.e. ~90% identity or better)**. The number of 'core' nucleotides able to be aligned among all genomes, and the number of SNPs within the core, are what determine the resolution of the phylogenetic tree. Output phylogenies are presented along with text files outlining the SNPs discovered.

Gene Family Analysis

For specialty gene analysis, the user selects read-based analysis and/or ORF(contig)-based analysis.

For read-based analysis, antibiotic resistance genes and virulence genes are detected using the Huttenhower lab's program ShortBRED. The antibiotic resistance gene database was generated by the developers of ShortBRED using genes from ARDB and Resfams. The virulence genes database was generated by the developers of EDGE using VFDB.

For ORF-based analysis, antibiotic resistance genes are detected using CARD's (Comprehensive Antibiotic Resistance Database) program RGI (Resistance Gene Identifier). RGI uses CARD's custom database of antibiotic resistance genes. The virulence genes are detected using ShortBRED with a database generated by the developers of EDGE using VFDB.

Primer Analysis

For primer analysis, if the user would like to validate known PCR primers *in silico*, a FASTA file of primer sequences must be input. New primers can be generated from an assembly as well.

Qiime2 analysis

QIIME2 is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data. EDGE implementation is based on Qiime 2 core 2019.1 and includes demultiplexing and quality control/filtering, feature table construction, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations. Currently, EDGE supports three amplicon types, 16s using GreenGenes database, 16s/18s using SILVA database, and Fungal ITS.

DETEQT (Targeted NGS) analysis

DETEQT is a pipeline for diagnostic targeted sequencing adjudication.

This tool been designed to be robust enough to handle a range of assay designs. Therefore, no major assumptions of input reads are made except that they represent amplicons from a multiplexed targeted amplification reaction and that the **reference is comprised of only target regions** in the assay, instead of whole genomes. The idea is to survey the reads and delineate whether each reference sequence, or target, is present or absent.

PiReT analysis

EDGE integrated PiReT (Pipeline for Reference based Transcriptomics) which is an open-source bioinformatics pipeline for performing RNA-Seq analysis. The workflow written mostly in Python on a popular workflow manager package luigi (developed by spotify). It allows users to find differentially expressed transcripts (genes, sRNAs), discover novel non-coding RNAs, co-expressed genes and pathways from raw FASTQ, reference sequence, and experimental design files.

All commands and tool parameters are recorded in log files to make sure the results are repeatable and traceable. The main output is an integrated interactive web page that includes summaries of all the workflows run and features tables, graphical plots, and links to genome (if assembled, or of a selected reference) browsers and to access unprocessed results and log files. Most of these summaries, including plots and tables are included within a final PDF report.

1.2.3 Limitations

Pre-processing

For host removal/screening, not all genomes are available from a drop-down list, however users can provide their own genome FASTA file as host input.

Assembly and Taxonomy Classification

EDGE has been primarily designed to **analyze microbial (bacterial, archaeal, viral) isolates or (shotgun) metagenome samples**. Due to the complexity and computational resources required for eukaryotic genome assembly, and the fact that the most taxonomy classification tools do not support eukaryotic classification (except Metaphlan2), EDGE does not fully support eukaryotic samples. The combination of large NGS data files and complex metagenomes may also run into computational memory constraints.

Reference-based analysis

We recommend only aligning against (a limited number of) most closely related genome(s) (default on GUI limit up to 200 fragments). If this is unknown, the Taxonomy Classification module is recommended as an alternative. If the user

selects too many references, this may affect runtimes or require more computational resources than may be available on the user's system.

Phylogenetic Analysis

Because this pipeline provides SNP-based trees derived from whole genome (and contig) alignments or read mapping, **we recommend selecting genomes within the same species or at least within the same genus**.

1.3 Computational Environment

1.3.1 EDGE source code, images, and webservers

EDGE was designed to be installed and implemented from within any institution that provides sequencing services or that produces or hosts NGS data. When installed locally, EDGE can access the raw FASTQ files from within the institution, thereby providing immediate access by the biologist for analysis. EDGE is available in a variety of packages to fit various institutional needs. **EDGE source code** can be obtained via our GitHub page. To simplify installation, a Docker image can also be obtained. An **online version of EDGE** is currently available at https://edgebioinformatics.org/.