## Contents

1 EDGE ABCs
   1.1 About EDGE Bioinformatics .................................................. 1
   1.2 Bioinformatics overview .......................................................... 1
   1.3 Computational Environment ...................................................... 4

2 Introduction
   2.1 What is EDGE? ................................................................. 5
   2.2 Why create EDGE? .............................................................. 5

3 System requirements
   3.1 Hardware Requirements .......................................................... 7
   3.2 Ubuntu 18.04 ........................................................................... 7
   3.3 CentOS 7 ............................................................................... 8

4 Installation
   4.1 EDGE Installation .................................................................... 10
   4.2 Configure SELinux on CentOS .................................................... 27
   4.3 EDGE Docker image .................................................................. 27

5 Graphic User Interface (GUI)
   5.1 User Login .............................................................................. 28
   5.2 Upload Files ............................................................................. 29
   5.3 Initiating an analysis job ............................................................ 30
   5.4 Choosing processes/analyses ....................................................... 35
   5.5 Submission of a job .................................................................. 43
   5.6 Checking the status of an analysis job ......................................... 43
   5.7 Monitoring the Resource Usage .................................................. 45
   5.8 Management of Jobs .................................................................. 45
   5.9 Project List Table ...................................................................... 46
   5.10 Other Methods of Accessing EDGE ............................................. 47

6 Command Line Interface (CLI)
   6.1 Configuration File ..................................................................... 50
   6.2 Test Run .................................................................................. 50
   6.3 Descriptions of each module ....................................................... 55
   6.4 Other command-line utility scripts .............................................. 62
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 publically 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 stitch paired-end(PE) reads and use joined PE reads for downstream analysis. The host removal step requires the input of one or more reference genomes as FASTA. 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 for PacBio/Nanopore reads, and Unicycler for bacteria genomes hybrid assembly. These assembly tools are to accommodate a range of sample types and data sizes. When the user selects to perform an assembly, all subsequent workflows can execute analysis with either the reads, the contigs, or both (default). For annotation, Prokka and RATT are provided for ab initio or transfer annotation from close-related reference genome. Start from version 2.4, EDGE use 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 assessed the quality of binning result 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 if there are 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 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 Huttenhower lab’s program ShortBRED. The antibiotic resistance gene database was generated by the developers of ShortBRED using genes from ARDB and Refsams. 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 (TargetedNGS) 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 allow 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.

### 1.2. Bioinformatics overview
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 institute that provides sequencing services or that produces or hosts NGS data. When installed locally, EDGE can access the raw FASTQ files from within the institute, thereby providing immediate access by the biologist for analysis. EDGE is available in a variety of packages to fit various institute needs. EDGE source code can be obtained via our GitHub page. To simplify installation, a Docker image can also be obtained. A demonstration version of EDGE is currently available at https://edgebioinformatics.org/.
2.1 What is EDGE?

EDGE is a highly adaptable bioinformatics platform that allows laboratories to quickly analyze and interpret genomic sequence data. The bioinformatics platform allows users to address a wide range of use cases including assay validation and the characterization of novel biological threats, clinical samples, and complex environmental samples. EDGE is designed to:

- Align to real world use cases
- Make use of open source (free) software tools
- Run analyses on small, relatively inexpensive hardware
- Provide remote assistance from bioinformatics specialists

2.2 Why create EDGE?

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: quality trimming and host removal, assembly and annotation, comparisons against known references, taxonomy classification of reads and contigs, whole genome SNP-based phylogenetic analysis, and PCR 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 the design of EDGE was intentionally done 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 insight into how each tool or workflow functions, and how the results should best be interpreted.
Fig. 1: Four common Use Cases guided initial EDGE Bioinformatic Software development.
CHAPTER 3

System requirements

NOTE: There is a demo version of EDGE, found on https://edgebioinformatics.org/ is run on our own internal servers and is recommended only for testing and demo purposes only.

The current version of the EDGE pipeline has been extensively tested on a Linux Server with Ubuntu 18.04 and CentOS 6.5/7 operating system and will work on 64bit Linux environments.

### 3.1 Hardware Requirements

Due to the involvement of several high memory and high cpu consuming steps Minimum requirement: 24GB memory, at least 8 computing CPUs and 1 TB disk space. A higher computer spec is strongly recommended: 256GB memory, 64 computing CPUs and > 4 TB disk space. Please ensure that your system has the essential software packages installed properly before running the installing script. The following should be installed by a system administrator (requires sudo).

**Note:** If your system OS is neither Ubuntu 18.04 or CentOS 7.0, it may have different packages/libraries name and the newer compiler on newer OS may fail on compiling some of third party bioinformatics tools. We would suggest to use EDGE Docker container.

#### 3.2 Ubuntu 18.04

1. Install build essential libraries and dependancies:
sudo apt-get update

sudo apt-get install -y build-essential libreadline-gplv2-dev libx11-dev \ libgsl-dev libfreetype6-dev libncurses5-dev gfortran \ inkscape libwww-perl libxml1-perl libperlzip-perl \ zlib-perl zip unzip libjson-perl libpng-dev cpanminus default-jre \ firefox wget csh liblapack-dev libblas-dev libatlas-base-dev \ libcairo2-dev libssh2-1-dev libssl-dev libcurl4-openssl-dev bzip2 \ bioperl rsync libbz2-dev liblzma-dev time libterm-readkey-perl \ liblwp-protocol-https perl gnuplot libjson-xs-perl libio-socket-ip-perl \ vim php sendmail mysql-client mysql-server libgfortran3 texinfo \ openssh-server openssh-client zlib-dev openjdk-11-jdk texlive \ texlive-fonts-extra libboost-all-dev cron less libxml2-dev

2. Install Apache2 for EDGE UI:

sudo apt-get install apache2
sudo a2enmod cgid proxy proxy_http headers rewrite

3. Install packages for user management system:

sudo apt-get install sendmail mysql-client mysql-server

cd /usr/share
wget https://archive.apache.org/dist/tomcat/tomcat-7/v7.0.92/bin/apache-tomcat-7.0.92.tar.gz
tar xzf apache-tomcat-7.0.92.tar.gz
rm apache-tomcat-7.0.92.tar.gz
mv apache-tomcat-7.0.92 tomcat7
echo "export CATALINA_HOME="/usr/share/tomcat7" >> /etc/profile

4. Change the image conversion policy:

sed -i.bak 's/rights="none" pattern="PDF"/rights="read|write" pattern="PDF"/g' /etc/ImageMagick-6/policy.xml

3.3 CentOS 7

1. Install libraries and dependencies by yum:

# add epel repository
sudo yum -y install epel-release

sudo yum install -y libX11-devel readline-devel libXt-devel ncurses-devel \ inkscape \ expat expat-devel freetype freetype-devel zlib zlib-devel perl-App-cpanminus \ perl-Test-Most blas-devel atlas-devel lapack-devel libpng12 libpng12-devel \ perl-XML-Simple perl-JSON csh gcc gcc-c++ make binutils gd gsl-devel git \ graphviz \ java-1.7.0-openjdk perl-Archive-Zip perl-CGI curl perl-CGI-Session

(continues on next page)
2. Update perl tools:

   ```bash
   sudo cpanm App::cpanoutdated
   sudo su -
   cpan-outdated -p | cpanm
   exit
   ```

3. Install perl modules by cpanm:

   ```bash
   sudo cpanm -f Bio::Perl Net::Ping
   sudo cpanm Graph Time::Piece Hash::Merge PerI1O::gzip Heap::Simple::XS File::Next
   sudo cpanm Algorithm::Munkres Archive::Tar Array::Compare Clone Convert::Binary::C
   sudo cpanm HTML::Template HTML::TableExtract List::MoreUtils PostScript::TextBlock
   sudo cpanm SOAP::Lite SVG SVG::Graph Set::Scalar Sort::Naturally
   Spreadsheet::ParseExcel
   sudo cpanm CGI::Simple GraphViz XML::Parser::PerlSAX XML::Simple Term::ReadKey
   ```

4. Install package for httpd for EDGE UI:

   ```bash
   sudo yum -y install httpd
   sudo systemctl enable httpd && sudo systemctl start httpd
   ```

5. Install packages for user management system:

   ```bash
   sudo yum -y install sendmail mariadb-server mariadb php phpMyAdmin tomcat
   sudo systemctl enable tomcat && sudo systemctl start tomcat
   ```

6. Configure firewall for ssh, http, https, and smtp:

   ```bash
   sudo firewall-cmd --permanent --add-service=ssh
   sudo firewall-cmd --permanent --add-service=http
   sudo firewall-cmd --permanent --add-service=https
   sudo firewall-cmd --permanent --add-service=smtp
   sudo firewall-cmd --reload
   ```

7. Disable SELinux:

   ```bash
   As root edit /etc/selinux/config and set SELINUX=disabled
   Restart the server to make the change
   ```

**Warning:** This is for development version of EDGE. Stable version (v2.3) is here.
Note: These instructions assumes Ubuntu 18 and CentOS 7

4.1 EDGE Installation

Note: A base install is ~12GB for the code base and ~500GB for the databases.

1. Please ensure that your system has the essential software building packages (page 7) installed properly before proceeding following installation.

2. Download the codebase, databases and third party tools:

```bash
## Codebase is ~207Mb and contains all the scripts and HTML needed to make EDGE run
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_main.tgz

## Third party tools is ~2.8Gb and contains the underlying programs needed to do the analysis

## Pipeline database is ~17Gb and contains the other databases needed for EDGE
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_pipeline_databases.tgz

## BWA index is ~41Gb and contains the databases for bwa taxonomic identification
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_bwa_index.tgz

## HOST genomes BWA index is ~41Gb for Host removal, including human, bacteria, phiX, viruses, invertebrate vectors of human pathogens
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_HostIndex.tgz
```

(continues on next page)
## NCBI Genomes is ~21Gb and contain the full genomes for prokaryotes and some viruses
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_NCBI_genomes.tgz

## GOTTCHA database is ~16Gb and contains the custom databases for the GOTTCHA taxonomic identification pipeline
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_GOTTCHA_db.tgz

## Amplicon database is ~78Mb and contains the databases for Qiime 16s and 18s ITS pipeline
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_amplicons_db.tgz

## NT database is ~25Gb and contains the NCBI nt database for contig identification
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_nt_20160426.tgz

## ShortBRED database is ~27Mb and contains the databases used by ShortBRED for virulence factors and read based antibiotic resistance analysis
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_ShortBRED_Database.tgz

## Diamond database is ~16Gb and contains the databases from RefSeq for protein based taxonomic identification
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_diamond_db.tgz

## MetaPhlAn2 database is 1.1Gb contains the databases used for the MetaPhlAn2 taxonomic identification pipeline
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_metaphlan2DB.tgz

## GOTTCHA2 databases are 38Gb and contains the custom databases for the GOTTCHA2 taxonomic identification pipeline
wget -c https://edge-dl.lanl.gov/EDGE/DB/edge_GOTTCHA2_db_20190729.tgz

## Kraken2 database is 26Gb contains the databases used for the Kraken2 taxonomic identification pipeline
wget -c https://edge-dl.lanl.gov/EDGE/DB/edge_Kraken2_db_20190104.tgz

## MICCR database is 48GB contains the databases used for the contig taxonomic identification pipeline
wget -c https://edge-dl.lanl.gov/EDGE/DB/edge_ContigTax_db_20190114.tgz

## CheckM database is 275MB contains the databases used for the Metagenome Binned contig quality assessment.
wget -c https://edge-dl.lanl.gov/EDGE/DB/edge_checkM_db_20190213.tgz

## Qiime2 database is 910MB contains 16s,18s and ITS db.
wget -c https://edge-dl.lanl.gov/EDGE/DB/edge_qiime2_db_20190227.tgz

(Optional)
## Other Host bwa index ~18Gb for host removal, including pig, sheep, cow, monkey, hamster, and goat.
wget -c https://edge-dl.lanl.gov/EDGE/DB/edge_dev_otherHostIndex.tgz

## For machine with < 32Gb memory, we suggest to use the smaller BWA index (~14Gb) and contains the databases for bwa taxonomic identification pipeline
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_bwa_mini_index.tgz

4.1. EDGE Installation
Warning: Be patient; the database files are huge.

3. Unpack main archive:

```
tar -xvzf edge_dev_main.tgz
```

Note: The main directory, edge_dev, will be created.

Create a link from edge to that directory:

```
ln -sf edge_dev edge
```

4. Unpack the third party software into main directory (edge):

```
tar -xvzf edge_dev_thirdParty_softwares.tgz -C edge/
```

Note: You should see a thirdParty directory inside the edge directory.

Note: If you are updating from old version, you should still expand the new thirdParty tgz file into the existing thirdParty directory.

5. Unpack the databases:

```
# unpack databases
   tar -xvzf edge_dev_pipeline_databases.tgz
   tar -xvzf edge_dev_GOTTCHA_db.tgz
   tar -xzvf edge_dev_bwa_index.tgz
   tar -xzvf edge_dev_NCBI_genomes.tar.gz
   tar -xzvf edge_dev_amplicons_db.tgz
   tar -xzvf edge_dev_nt_20160426.tgz
   tar -xzvf edge_dev_ShortBRED_Database.tgz
   tar -xzvf edge_dev_HostIndex.tgz
   tar -xzvf edge_dev_diamond_db.tgz
   tar -xzvf edge_dev_metaphlan2DB.tgz
   tar -xzvf edge_dev_GOTTCHA2_db_20181115.tgz
   tar -xzvf edge_dev_Kraken2_db_20190104.tgz
   tar -xzvf edge_contigTax_db_20190114.tgz
   tar -xzvf edge_checkM_db_20190213.tgz
   tar -xzvf edge_qiime2_db_20190227.tgz
```

Note: At this point, you should see a database directory and the edge directory.

6. Create the symlink from edge to the database directory:

```
ln -s `pwd`/database edge/database
```

Note: This will keep the database directory outside of the edge install location. Should you need to reinstall the code base you will not need to redownload/install the databases.

4.1. EDGE Installation
7. Installing pipeline:

```bash
cd edge
./INSTALL.sh
```

**Note:** When installing JBrowse, it may require internet connection.

**Note:** If the machine is shared and used with others, the system installed tools version may not be compatible with EDGE. In this case, we would suggest to use force option `./INSTALL.sh force` to install all list tools locally.

It will install the following depended tools (page 76).

- **Assembly**
  - idba
  - spades
  - megahit
  - long_read_assembly
  - racon
  - unicycler

- **Annotation**
  - prokka
  - RATT
  - tRNAscan
  - barrnap
  - BLAST+
  - blastall
  - phageFinder
  - glimmer
  - aragorn
  - prodigal
  - tbl2asn
  - ShortBRED
  - antismash

- **Alignment**
  - hmmer
  - infernal
  - bowtie2
  - bwa
  - mummer
- RAPSearch2
- diamond
- minimap2

- **Taxonomy**
  - kraken2
  - metaphlan2
  - kronatools
  - gottcha
  - gottcha2
  - centrifuge
  - miccr

- **Phylogeny**
  - FastTree
  - RAxML

- **Metagenome**
  - MaxBin
  - checkM

- **Utility**
  - FaQCs
  - bedtools
  - R
  - GNU_parallel
  - tabix
  - JBrowse
  - bokeh
  - primer3
  - samtools
  - bcftools
  - sratoolkit
  - ea-utils
  - omics-pathway-viewer
  - NanoPlot
  - Porechop
  - seqtk
  - Rpackages
  - Chromium

4.1. EDGE Installation
• Perl_Modules
  – perl_parallel_forkmanager
  – perl_excel_writer
  – perl_archive_zip
  – perl_string_approx
  – perl_pdf_api2
  – perl_html_template
  – perl_html_parser
  – perl_JSON
  – perl_bio_phylo
  – perl_xml_twig
  – perl_cgi_session
  – perl_email_valid
  – perl_mailtools

• Python_Packages
  – Anaconda2
  – Anaconda3

• Pipeline_Tools
  – DETEQT
  – reference-based_assembly
  – PyPiReT
  – qiime2

8. Restart the Terminal Session to allow $EDGE_HOME to be exported.

Note: After running INSTALL.sh successfully, the binaries and related scripts will be stored in the ./bin and ./scripts directory. It also writes EDGE_HOME environment variable into .bashrc or .bash_profile.

4.1.1 Testing the EDGE Installation

After installing the packages above, it is highly recommended to test the installation:

```bash
> cd $EDGE_HOME/testData
> ./runAllTest.sh
```
There are 20 module/unit tests which took around 2 hours 07 mins in our testing environments. (64 cores 2.30GHz, 512GB ram with CentOS-7.1.1503). You will see test output on the terminal indicating test successes and failures. The **Specialty Genes Profiling test** will fail in this stage since it requires virulence database imported and configured. You can test it again after database created and configured. Some tests may fail due to missing external applications/modules/packages or failed installation. These will be noted separately in the $EDGE_HOME/testData/runXXXXTest/TestOutput/error.log or log files in each modules. If these are related to features of EDGE that you are not using, this is acceptable. Otherwise, you’ll want to ensure that you have the EDGE installed correctly. If the output doesn’t indicate any failures, you are now ready to use EDGE through command line.

To take advantage of the user friendly GUI, please follow the section below to configure the EDGE Web server.

### 4.1.2 Apache Web Server Configuration

1. Modify/Check sample apache configuration file:

```bash
For Ubuntu

Double check $EDGE_HOME/edge_ui/apache_conf/edge_apache.conf alias directories:

- → the match EDGE installation path at line 2,5,6,16,17,29,38,69.

The default is configured as http://localhost/edge_ui/ or http://www.yourdomain.com/edge_ui/

For CentOS

Double check $EDGE_HOME/edge_ui/apache_conf/edge_httpd.conf alias directories the:

- → match EDGE
```

(continues on next page)
2. Confirm apache/httpd user and groups are edge:

For Ubuntu

The user and group can be edited at /etc/apache2/envvars and the variables are APACHE_RUN_USER and APACHE_RUN_GROUP.

For CentOS

The User and Group on lines 66 and 67 in $EDGE_HOME/edge_ui/apache_conf/centos_httpd.conf should be edge

# Make APACHE_RUN_USER have Permission to write
> sudo chown -R xxxxx $EDGE_HOME/edge_ui $EDGE_HOME/edge_ui/JBrowse/data
#(xxxxx is the APACHE_RUN_USER value)
> sudo chgrp -R xxxxx $EDGE_HOME/edge_ui $EDGE_HOME/edge_ui/JBrowse/data
#(xxxxx is the APACHE_RUN_GROUP value)

3. (Optional) If users are behind a corporate proxy for internet:

Please add proxy info into $EDGE_HOME/edge_ui/apache_conf/edge_apache.conf or $EDGE_HOME/edge_ui/apache_conf/edge_httpd.conf

# Add following proxy env
SetEnv http_proxy http://yourproxy:port
SetEnv https_proxy http://yourproxy:port
SetEnv ftp_proxy http://yourproxy:port

4. Copy configuration files to the appropriate directories:

For Ubuntu

  > sudo cp $EDGE_HOME/edge_ui/apache_conf/edge_apache.conf /etc/apache2/conf-available/
  > sudo ln -s /etc/apache2/conf-available/edge_apache.conf /etc/apache2/conf-enabled/
  > sudo cp $EDGE_HOME/edge_ui/apache_conf/pangia-vis.conf /etc/apache2/conf-available/
  > sudo ln -s /etc/apache2/conf-available/pangia-vis.conf /etc/apache2/conf-enabled/

For CentOS

  > sudo cp $EDGE_HOME/edge_ui/apache_conf/edge_httpd.conf /etc/httpd/conf.d/
  > sudo cp $EDGE_HOME/edge_ui/apache_conf/centos_httpd.conf /etc/httpd/conf/httpd.conf
  > sudo cp $EDGE_HOME/edge_ui/apache_conf/pangia-vis.conf /etc/httpd/conf.d/

5. (Optional) HTTPS / SSL configuration:
1. Please add redirect conditions into `$EDGE_HOME/edge_ui/apache_conf/edge_apache.conf` or `$EDGE_HOME/edge_ui/apache_conf/edge_httpd.conf`

```bash
# Add redirect to https
RewriteEngine on
RewriteCond %{HTTPS} !=on
RewriteRule ^(.*) https://%{SERVER_NAME}$1 [R,L]
```

1i. Use `pangia-vis-https.conf` instead of `pangia-vis.conf`

For Ubuntu

```bash
> sudo cp $EDGE_HOME/edge_ui/apache_conf/pangia-vis-https.conf /etc/apache2/conf-available/pangia-vis.conf
```

For CentOS

```bash
> sudo cp $EDGE_HOME/edge_ui/apache_conf/pangia-vis-https.conf /etc/httpd/conf.d/
```

1ii. Use `pangia-vis-https.conf` instead of `pangia-vis.conf`

For Ubuntu

```bash
> sudo cp $EDGE_HOME/edge_ui/apache_conf/pangia-vis-https.conf /etc/apache2/conf-available/pangia-vis.conf
```

For CentOS

```bash
> sudo cp $EDGE_HOME/edge_ui/apache_conf/pangia-vis-https.conf /etc/httpd/conf.d/
```

1iii. Add SSL configuration:

see `edge_ssl.conf` using letsencrypt (https://letsencrypt.org/) as an example. Please modify it as your environments and copy modified `$EDGE_HOME/edge_ui/apache_conf/edge_ssl.conf` to `/etc/httpd/conf.d/` for CentOS or `/etc/apache2/conf-enabled/` for Ubuntu.

6. Restart the `apache2/httpd` to activate the new configuration:

For Ubuntu

```bash
> sudo service apache2 restart
```

For CentOS

```bash
> sudo systemctl restart httpd
```

### 4.1.3 User Management System Installation: MySQL

**Note:** Setup two temporary environmental variables:

```bash
UN=username
PW=password
```

These will be used when setting up the user management system

**Note:** If you were using the user management system and are updating from old EDGE version to this version. You only need to run the commands below and continue to install tomcat:

```bash
cd $EDGE_HOME/userManagement
mysql -u $UN -p userManagement
mysql> source update_userManagement_db.sql
```
Note: For MySQL 5.7 (Ubuntu 18.04), please append following content to /etc/mysql/my.cnf

```ini
[mysqld]
show_compatibility_56 = on
sql-mode=""
```

1. Start mysql (if it is not already running):

   For Ubuntu
   ```
   > sudo service mysql start
   ```

   For CentOS
   ```
   > sudo systemctl start mariadb.service && sudo systemctl enable mariadb.service
   ```

2. Secure mysql:

   Note: The root password here is for the mysql root and not the system root.

   ```
   > sudo mysql_secure_installation
   ```

   1. Enter root password (likely none)
   2. Set root password? Yes
   3. Enter new root password.
   4. Re-enter new root password.
   5. Remove anonymous users? Yes
   6. Disallow root login remotely? Yes
   7. Remove test database and access to it? Yes
   8. Reload privilege table now? Yes

3. Create database: userManagement:

   ```
   > cd $EDGE_HOME/userManagement
   > mysql -p -u root
   ```

   ```
   mysql> create database userManagement;
   mysql> use userManagement;
   ```

4. Load userManagement_schema.sql:

   ```
   mysql> source userManagement_schema.sql;
   ```

5. Load userManagement_constrains.sql:

   ```
   mysql> source userManagement_constrains.sql;
   ```

6. Create an user account and grant all privileges to user:

   Note: This is the database user (not an individual EDGE user account).

4.1. EDGE Installation
Replace with the appropriate values:

<table>
<thead>
<tr>
<th>username: yourDBUsername</th>
</tr>
</thead>
<tbody>
<tr>
<td>password: yourDBPassword</td>
</tr>
</tbody>
</table>

```
mysql> CREATE USER 'yourDBUsername'@'localhost' IDENTIFIED BY 'yourDBPassword';
mysql> GRANT ALL PRIVILEGES ON userManagement.* to 'yourDBUsername'@'localhost';
mysql> exit;
```

### 4.1.4 User Management System Installation: Tomcat

**Note:** If you were using the user management system and are updating from old EDGE version to this version. You only need continue from step 6.

1. Configure tomcat basic auth to secure /user/admin/register web service:

   **Warning:** Run this code only once!

   **Note:** The username and password here should be the same as the database user.

   Update the values for the username and password accordingly before running the code.

   This adds the following to /usr/share/tomcat/conf/tomcat-users.xml or /usr/share/tomcat7/conf/tomcat-users.xml:

   ```
   <role rolename="admin"/>
   <user username="yourAdminName" password="yourAdminPassword" roles="admin"/>
   ```

   **For Ubuntu**

   ```
   sudo sed -i 's@</tomcat-users>@<role rolename="admin"/>
   <user username="'$UN'" password="'$PW'" roles="admin"/>
   </tomcat-users>@g' /usr/share/tomcat7/conf/tomcat-users.xml
   ```

   **For CentOS**

   ```
   sudo sed -i 's@<role rolename="admin"/> --><role rolename="admin"/>
   <user username="'$UN'" password="'$PW'" roles="admin"/>
   </tomcat-users>' /usr/share/tomcat/conf/tomcat-users.xml
   ```

2. Update inactive timeout to a more reasonable number 4320 min (3 days) from default (30mins) in /usr/share/tomcat7/conf/web.xml or /etc/tomcat/web.xml

   **Note:** This is modifying the following code:
3. Add memory constrains to Java:

**Warning:** Run this code only once!

**Note:** This will add the following line to the appropriate file:

```
JAVA_OPTS=" -Xms256M -Xmx1024M -XX:PermSize=256m -XX:MaxPermSize=512m"
```

4. Restart tomcat server:

For Ubuntu
```
sudo /usr/share/tomcat7/bin/startup.sh
```

For CentOS7
```
sudo systemctl restart tomcat
```

5. Copy database connector clients to appropriate lib directory:

For Ubuntu
```
sudo cp mysql-connector-java-5.1.34-bin.jar /usr/share/tomcat7/lib/
sudo chmod 744 /usr/share/tomcat7/lib/mysql-connector-java-5.1.34-bin.jar
```

For CentOS
```
sudo cp mariadb-java-client-1.2.0.jar /usr/share/tomcat/lib/
sudo chmod 744 /usr/share/tomcat/lib/mariadb-java-client-1.2.0.jar
```
6. Centos Only: Update the MySQL database driver to be used:

```
sed -i 's@driverClassName=.*$@driverClassName="org.mariadb.jdbc.Driver"@' $EDGE_HOME/userManagement/userManagementWS.xml
```

7. Deploy userManagement to tomcat server:

**Note:** For CentOS the userManagementWS.xml should have:

```
driverClassName="org.mariadb.jdbc.Driver"
```

Please check and confirm this before deploying userManagement.

For Ubuntu

```bash
sudo rm -rf /usr/share/tomcat7/webapps/userManagementWS
sudo cp userManagementWS.war /usr/share/tomcat7/webapps/
sudo rm -rf /usr/share/tomcat7/webapps/userManagement
sudo cp userManagement.war /usr/share/tomcat7/webapps/
sudo chmod 755 /usr/share/tomcat7/webapps/*war
sudo cp userManagementWS.xml /usr/share/tomcat7/conf/Catalina/localhost/
sudo chmod 744 /usr/share/tomcat7/conf/Catalina/localhost/userManagementWS.xml
```

For CentOS

```bash
sudo rm -rf /var/lib/tomcat/webapps/userManagementWS
sudo cp userManagementWS.war /var/lib/tomcat/webapps/
sudo rm -rf /var/lib/tomcat/webapps/userManagement
sudo cp userManagement.war /var/lib/tomcat/webapps/
sudo chmod 755 /var/lib/tomcat/webapps/*war
sudo cp userManagementWS.xml /etc/tomcat/Catalina/localhost/
sudo chmod 744 /etc/tomcat/Catalina/localhost/userManagementWS.xml
```

8. Modify the username/password in userManagementWS.xml:

For Ubuntu

```bash
sudo sed -i 's@username=.*$@username="'${UN}'"@' /usr/share/tomcat7/conf/Catalina/localhost/userManagementWS.xml
sudo sed -i 's@password=.*$@password="'${PW}'"@' /usr/share/tomcat7/conf/Catalina/localhost/userManagementWS.xml
```

For CentOS

```bash
sudo sed -i 's@username=.*$@username="'${UN}'"@' /etc/tomcat/Catalina/localhost/userManagementWS.xml
sudo sed -i 's@password=.*$@password="'${PW}'"@' /etc/tomcat/Catalina/localhost/userManagementWS.xml
```

9. Update sys.properties in the userManagement deployment:

**Note:** Tomcat should automatically unarchive the .war files.

The default configuration is to have the user management system on localhost with email notifications turned off.

4.1. EDGE Installation
For “Forgot your password” reset function, the ‘email_notification’ should be on and a valid email address for ‘email_sender’

Modify the user management sys.properties if you want to change the default behavior. (make sure port match with tomcat server)

You will need to copy the sys.properties files to the directory of the userManagement deployment.

For Ubuntu

```
sudo cp $EDGE_HOME/userManagement/sys.properties /usr/share/tomcat7/
   →webapps/userManagement/WEB-INF/classes/sys.properties
sudo chmod 744 /usr/share/tomcat7/webapps/userManagement/WEB-INF/classes/
   →sys.properties
```

For CentOS

```
sudo cp $EDGE_HOME/userManagement/sys.properties /usr/share/tomcat/
   →webapps/userManagement/WEB-INF/classes/sys.properties
sudo chmod 744 /usr/share/tomcat/webapps/userManagement/WEB-INF/classes/
   →sys.properties
```

10. Restart tomcat server:

For Ubuntu

```
sudo /usr/share/tomcat7/bin/shutdown.sh
sudo /usr/share/tomcat7/bin/startup.sh
```

For CentOS 7

```
sudo systemctl restart tomcat
```

11. Setup admin user:

**Note:** The script createAdminAccount.pl creates an admin user account for EDGE userManagement.

Update email (-e), First Name (-fn), and Last Name (-ln) appropriately.

It will ask tomcat service username and password (tomcat-users.xml:) before creating EDGE user account (email).

If “HTTP Status 401” error shows, please make sure the tomcat username and password in the first step match with what entered here.

If “HTTP Status 403” error shows, please make sure the tomcat rolename in the first step match with /var/lib/tomcat/webapps/userManagementWS/WEB-INF/web.xml and where the web.xml file existed or not.

If “HTTP Status 500” error shows, please make sure the port (default: 8080) for tomcat service are matched in tomcat server.xml, $EDGE_HOME/edge_ui/sys.properties and $EDGE_HOME/userManagement/sys.properties (need to redo step 9).

Should this script fail, the userManagement is not set up correctly.

```
perl createAdminAccount.pl -e <email> -fn <first name> -ln <last name>
```

12. Enable userManagement in EDGE sys.properties:
Note: See EDGE Configuration (page 25) below

```
> sed -i 's@user_management=.*$@user_management=1@g' $EDGE_HOME/edge_ui/sys.properties
> sed -i 's@edge_user_management_url=.*$@edge_user_management_url=http://localhost/userManagement@g' $EDGE_HOME/edge_ui/sys.properties
```

13. Optional: configure social (facebook, google, windows live, LinkedIn) login function:
   • modify $EDGE_HOME/edge_ui/javascript/social.js, change apps id you created on each social media.

Note: This allow users to use their social media account to login EDGE. You need to register your EDGE’s domain on each social media to get apps id. e.g.: A FACEBOOK app needs to be created and configured for the domain and website set up by EDGE. see https://developers.facebook.com/ and StackOverflow Q&A

Google+
Windows
LinkedIn

14. Optional: configure sendmail to use SMTP to email out of local domain:
   • edit /etc/mail/sendmail.cf and edit this line:
     ```
     # “Smart” relay host (may be null) DS
     # “Smart” relay host (may be null) DS mail.yourdomain.com
     ```
   • and append the correct server right next to DS (no spaces);
   • Then, restart the sendmail service
   ```
   > sudo service sendmail restart
   ```

4.1.5 MYSQL Databases CREATION

Note: This requires that MySQL is installed and running.

Note: EDGE provides Virulence Factors, Metadata, and Pathogen sql dump files which will be used for Speciality Gene Profiling module, Sample MetaData module and Pathogen Detection module, respectively. You will need configure the database info in the $EDGE_HOME/edge_ui/sys.properties. See EDGE Configuration (page 25) below

1. Change directory into database:
   ```
   cd $EDGE_HOME/SQLdbfile
   ```

2. Run install script for databases and Grant privilege database user to have access to the databases:

4.1. EDGE Installation
mysql -u root -p
mysql> source virulence_db.sql ;
mysql> GRANT ALL PRIVILEGES ON virulenceFactors.* to 'yourDBUsername'@'localhost';

mysql> create database edgeDB;
mysql> use edgeDB;
mysql> source edge_db.sql ;
mysql> GRANT ALL PRIVILEGES ON edgeDB.* to 'yourDBUsername'@'localhost';

mysql> create database pathogens ;
mysql> use pathogens;
mysql> source pathogen_db.sql ;
mysql> GRANT ALL PRIVILEGES ON pathogens.* to 'yourDBUsername'@'localhost';
mysql> exit;

3. Configure Virulence, Metadata and Pathogen Database information:

```properties
# Virulence Factor database
VFDB_dbhost = localhost
VFDB_dbport = 3306
VFDB_dbname = virulenceFactors
VFDB_dbuser = edge_user
VFDB_dbpasswd = edge_user_password

##configure edge pathogen detection 1: with 0: without
edge_pathogen_detection=0
pathogen_dbhost=localhost
pathogen_dbname=pathogens
pathogen_dbuser=edge_user
pathogen_dbpasswd=edge_user_password

##configure edge sample metadata option 1: with 0: without
edge_sample_metadata=0
edge_dbhost=localhost
edge_dbname=edgeDB
edge_dbuser=edge_user
edge_dbpasswd=edge_user_password
```

4.1.6 EDGE configuration

**Note:** EDGE system configuration file is $EDGE_HOME/edge_ui/sys.properties. You can edit this file to turn on/off EDGE functions/modules here. (on=1, off=0);

1. Add EDGE GUI admin info:

   According to User Management system installation step 11:

   ```properties
   edgeui_admin=admin@my.com
   edgeui_admin_password=admin
   ```

2. Turn on user management system:
Note: This assumes localhost is the domain. Update the domain as necessary. If user management system is not in the same domain with EDGE:

```plaintext
edge_user_management_url=http://www.someother.com/userManagement
```

# If you have User Management system enabled.
user_management=1
edge_user_management_url=http://localhost/userManagement

3. Turn on upload function:

```plaintext
user_upload=1
user_upload_maxFileSize='5gb'
```

4. Turn on project intermediate files clean up:

```plaintext
#Clean up old bam/sam/fastq/gz files (based on file age) in project directories
degenui_proj_store_days=10
```

5. Set up the archive directory:

```plaintext
#The archive space is for offload the main computational disk space
degenui_archive=/path/to/archive_SPACE
```

6. Turn on/off Social Login function:

```plaintext
#If you have User Management system installation step 18 done.
user_social_login=0
```

7. Turn on job submission for SGE/UGE cluster environment:

Note: make sure the user/apache user running EDGE is a cluster user.

```plaintext
qconf -suserl to check cluster user list
```

Edit the sge_bin, sge_root and sge_cell corresonding to your cluster environment

```plaintext
#Configure cluster system 1: with 0: without
cluster=1

## sge environment configuration
sge_bin=/cm/shared/apps/sge/2011.11p1/bin/linux-x64
sge_root=/cm/shared/apps/sge/2011.11p1
sge_cell=default

## edge job submission configuration
cluster_job_notify=edge@yourdomain.com
cluster_job_prefix=EDGE_pipeline_
cluster_qsub_options=
cluster_job_resource=h_vmem=6G -pe smp <CPU> -binding linear:<CPU/2>
cluster_job_max_cpu=64
```
4.2 Configure SELinux on CentOS

Warning: This is not complete.

1. Install semanage (if not already installed):

```bash
> sudo yum install -y policycoreutils-python setroubleshoot
```

2. Allow httpd to access $EDGE_HOME, the databases, and read/write to the EDGE_input/EDGE_output:

```bash
> sudo semanage fcontext -a -t httpd_sys_content_t "$EDGE_HOME(/.*)?"
> sudo semanage fcontext -a -t httpd_sys_content_t "$EDGE_HOME/database(/.*)?"
> sudo semanage fcontext -a -t httpd_sys_content_t "$EDGE_HOME/edge_ui/EDGE_input(/.*)?"
> sudo semanage fcontext -a -t httpd_sys_content_t "$EDGE_HOME/edge_ui/EDGE_output(/.*)?"
```

3. Allow httpd to execute cgi-scripts in $EDGE_HOME/edge_ui/cgi-bin/:

```bash
> sudo semanage boolean -m --on httpd_enable_cgi
> sudo semanage fcontext -a -t httpd_sys_script_exec_t "$EDGE_HOME/edge_ui/cgi-bin(/.*)?"
```

4. Allow httpd to connect to mysql database:

```bash
> sudo semanage boolean -m --on httpd_can_network_connect_db
```

5. Optional: Allow httpd to work with nfs and send mail:

```bash
> sudo semanage boolean -m --on httpd_use_nfs
> sudo semanage boolean -m --on httpd_can_sendmail
```

6. REQUIRED: Apply the rules:

```bash
> sudo restorecon -R $EDGE_HOME
> sudo restorecon -R $EDGE_HOME/database/
> sudo restorecon -R $EDGE_HOME/edge_ui/EDGE_input/
> sudo restorecon -R $EDGE_HOME/edge_ui/EDGE_output/
```

4.3 EDGE Docker image

EDGE has a lot of dependencies and can (but doesn’t have to) be very challenging to install. The EDGE docker gets around the difficulty of installation by providing a functioning EDGE full install on top of official Ubuntu Base Image (18.04.2). You can find the image and usage at docker hub. We would recommend to use Docker container for easy update in the future.
The User Interface was mainly implemented in JQuery Mobile, CSS, javascript and perl CGI. It is a HTML5-based user interface system designed to make responsive web sites and apps that are accessible on all smartphone, tablet and desktop devices. (see *How to make an app icon on the mobile device* (page 91))

See GUI page

### 5.1 User Login

A user management system has been implemented to provide a level of privacy/security for a user’s submitted projects. When this system is activated, any user can view projects that have been made public, but other projects can only be accessed by logging into the system using a registered local EDGE account or via an existing social media account (Facebook, Google+, Windows, or LinkedIn). The users can then run new jobs and view their own previously run projects or those that have been shared with them. Click on the upper-right user icon will pop up an user login window.
5.2 Upload Files

EDGE supports input from NCBI Sequence Reads Archive (SRA) and select files from the EDGE server. To analyze users’ own data, EDGE allows user to upload fastq, fasta and genbank (which can be in gzip format) and text (txt). Max file size is ‘5gb’ and files will be kept for 7 days. Choose “Upload files” from the navigation bar on the left side of the screen. Add users files by clicking “Add Files” button or drag files to the upload feature window. Then, click “Start Upload” button to upload files to EDGE server.
5.3 Initiating an analysis job

Choose “Run EDGE” or “Run Qiime” from the navigation bar on the left side of the screen.

5.3.1 Run EDGE

Click “Run EDGE” will cause a section to appear called “Input Raw Reads.” Here, you may browse the EDGE Input Directory and select FASTQ files containing the reads to be analyzed. EDGE supports gzip compressed fastq files. At minimum, EDGE will accept two FASTQ files containing paired reads and/or one FASTQ file containing single reads as initial input. Alternatively, rather than providing files through the EDGE Input Directory, you may decide to use as input reads from the Sequence Read Archive (SRA). In this case, select the “yes” option next to “Input from NCBI Sequence Reads Archive” and a field will appear where you can type in an SRA accession number.

In addition to the input read files, you have to specify a project name. The project name is restricted to only alphanumerical characters and underscores and requires a minimum of three characters. For example, a project name of “E.
coli. Project” is not acceptable, but a project name of “E_coli_project” could be used instead. In the “Description” fields you may enter free text that describes your project. If you would like, you may use as input more reads files than the minimum of 2 paired read files or one file of single reads. To do so, click “additional options” to expose more fields, including two buttons for “Add Paired-end Input” and “Add Single-end Input”.

In the “additional options”, there are several more options, for output path, number of CPUs, and config file. In most cases, you can ignore these options, but they are described briefly below.

### 5.3.2 Run Qiime

Click “Run Qiime2” will cause a section to appear for Qiime input and parameters. Currently, EDGE supports three amplicon types, 16s using GreenGenes database, 16s/18s using SILVA database, and Fungal ITS. Similar to “Run EDGE”, input can be either from the Sequence Read Archive (SRA, internet required) or browse the EDGE Input Directory based on the reads type. The Qiime pipeline supports one Reads Type in a run, paired-reads, single end reads, or de-multiplexed reads directory. There is also a mapping file input requirement which is adapted from QIIME Metadata mapping file. This mapping file contains all of the information about the samples necessary to perform the
data analysis. It is in tab-delimited format. In general, the header for this mapping file starts with a pound (#) character, and generally requires a “SampleID”, “BarcodeSequence”, and a “Description”.

Mapping File Example:

<table>
<thead>
<tr>
<th>#SampleID</th>
<th>BarcodeSequence</th>
<th>SampleType</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample1</td>
<td>ACATACCGTCTA</td>
<td>Stool</td>
<td>MiSeq,metagenome</td>
</tr>
<tr>
<td>Sample2</td>
<td>ACCATGCCTCTA</td>
<td>Blood</td>
<td>MiSeq,clinical</td>
</tr>
<tr>
<td>Control1</td>
<td>AGCCATCGTCTA</td>
<td>Control</td>
<td>Negative</td>
</tr>
<tr>
<td>Control2</td>
<td>CGTCTAAACCATG</td>
<td>Control</td>
<td>Spike-in Control</td>
</tr>
</tbody>
</table>

When the reads type is “De-multiplexed Reads Directory “, the mapping file needs a ‘Files’ column with FASTQ filenames for each sampleID. It can be paired-end or single-end FASTQ file and paired-end FASTQ files are comma-separated.

<table>
<thead>
<tr>
<th>#SampleID</th>
<th>Files</th>
<th>SampleType</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample1</td>
<td>S1.R1.fastq,S1.R2.fastq</td>
<td>Stool</td>
<td>MiSeq,metagenome</td>
</tr>
<tr>
<td>Sample2</td>
<td>S2.R1.fastq,S2.R2.fastq</td>
<td>Blood</td>
<td>MiSeq,clinical</td>
</tr>
<tr>
<td>Control1</td>
<td>C1.R1.fastq,C1.R2.fastq</td>
<td>Control</td>
<td>Negative</td>
</tr>
<tr>
<td>Control2</td>
<td>C2.R1.fastq,C2.R2.fastq</td>
<td>Control</td>
<td>Spike-in Control</td>
</tr>
</tbody>
</table>

5.3.3 Run DETEQT

Click “Run DETEQT” will cause a section to appear for DETEQT input and parameters. The DETEQT is a pipeline for diagnostic targeted sequencing adjudication. You may find more information from here. The DETEQT pipeline required user to select a directory, a metadata mapping file and a targeted amplicon references. The metadata mapping file is a tab-delimited file or excel file which header or first row includes #SampleID and Files. In the Files column, the paired-end fastq files are separated by a comma(,) and all the fastq files should be located in the input directory. The reference is comprised of only target regions in FASTA format in the assay.
Metadata Mapping File example:

<table>
<thead>
<tr>
<th>SampleID</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dengue</td>
<td>sample.1.fq,sample.2.fq</td>
</tr>
<tr>
<td>Flu</td>
<td>flu.1.fq,flu.2.fq</td>
</tr>
<tr>
<td>Ebola</td>
<td>ebola.1.fq,ebola.2.fq</td>
</tr>
<tr>
<td>MERS</td>
<td>mers.1.fq,mers.2.fq</td>
</tr>
<tr>
<td>SARS</td>
<td>sars.1.fq,sars.2.fq</td>
</tr>
<tr>
<td>Zika</td>
<td>zika.1.fq,zika.2.fq</td>
</tr>
<tr>
<td>Rota</td>
<td>rota.1.fq,rota.2.fq</td>
</tr>
<tr>
<td>HIV</td>
<td>hiv.1.fq,hiv.2.fq</td>
</tr>
<tr>
<td>Hanta</td>
<td>hanta.1.fq,hanta.2.fq</td>
</tr>
<tr>
<td>HCV</td>
<td>hcv.1.fq,hcv.2.fq</td>
</tr>
</tbody>
</table>

5.3.4 Run PiReT

Click “Run PiReT” will cause a section to appear for PiReT input and parameters. The PiReT is a pipeline for Reference based Transcriptomics analysis. You may find more information from PiReT github. The PiReT pipeline required user to select a directory, a experimental design file and references FASTA and GFF files in the parameters section. The experimental file is a tab-delimited file or excel file which header or first row includes #SampleID, Files, and Group. In the Files column, the paired-end fastq files are separated by a colon(;) and all the fastq files should be located in the input directory. The feature ID in the reference GFF files should be unique within the scope of the GFF file.

5.3. Initiating an analysis job
Experimental Design File example:

<table>
<thead>
<tr>
<th>#SampleID</th>
<th>Files</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>samp1</td>
<td>samp1_R1.fastq; samp1_R2.fastq</td>
<td>liver</td>
</tr>
<tr>
<td>samp2</td>
<td>samp2_R1.fastq; samp2_R2.fastq</td>
<td>spleen</td>
</tr>
<tr>
<td>samp3</td>
<td>samp3_R1.fastq; samp3_R2.fastq</td>
<td>spleen</td>
</tr>
<tr>
<td>samp4</td>
<td>samp4_R1.fastq; samp4_R2.fastq</td>
<td>liver</td>
</tr>
<tr>
<td>samp5</td>
<td>samp5_R1.fastq; samp5_R2.fastq</td>
<td>liver</td>
</tr>
<tr>
<td>samp6</td>
<td>samp6_R1.fastq; samp6_R2.fastq</td>
<td>spleen</td>
</tr>
</tbody>
</table>

5.3.5 Number of CPUs

Additionally, you may specify the number of CPUs to be used. The default and minimum value is one-fourth of total number of server CPUs. You may adjust this value if you wish. Assuming your hardware has 64 CPUs, the default is 16 and the maximum you should choose is 62 CPUs. Otherwise, if the jobs currently in progress use the maximum number of CPUs, the new submitted job will be queued (and colored in grey. Color-coding see Checking the status of an analysis job (page 43)). For instance, if you have only one job running, you may choose 62 CPUs. However, if you are planning to run 6 different jobs simultaneously, you should divide the computing resources (in this case, 10 CPUs per each job, totaling 60 CPUs for 6 jobs).
5.3.6 Config file

Below the “Use # of CPUs” field is a field where you may select a configuration file. A configuration file is automatically generated for each job when you click “Submit.” This field could be used if you wanted to restart a job that hadn’t finished for some reason (e.g. due to power interruption, etc.). This option ensures that your submission will be run exactly the same way as previously, with all the same options.

See also:

Example of config file (page 51)

5.3.7 Batch project submission

The “Batch project submission” section is toggled off by default. Clicking on it will open it up and toggle off the “Input Sequence” section at the same time. When you have many samples in “EDGE Input Directory” and would like to run them with the same configuration, instead of submitting several times, you can compile a Excel file with project name, fastq inputs and optional project descriptions (you can download the example excel file and fill it with your own data) and submit through the “Batch project submission” section.

![Batch project submission image]

5.4 Choosing processes/analyses

Once you have selected the input files and assigned a project name and description, you may either click “Submit” to submit an analysis job using the default parameters, or you may change various parameters prior to submitting the job. The default settings include quality filter and trimming, assembly, annotation, and community profiling. Therefore, if you choose to use default parameters, the analysis will provide an assessment of what organism(s) your sample is composed of, but will not include host removal, primer design, etc. Below the “Input Your Sample” section is a section called “Choose Processes / Analyses”. It is in this section that you may modify parameters if you would like to use settings other than the default settings for your analysis (discussed in detail below).
5.4.1 Pre-processing

Pre-processing is by default on, but can be turned off via the toggle switch on the right hand side. The default parameters should be sufficient for most cases. However, if your experiment involves specialized adapter sequences that need to be trimmed, you may do so in the Quality Trim and Filter subsection. There are two options for adapter trimming. You may either supply a FASTA file containing the adapter sequences to be trimmed, or you may specify N number of bases to be trimmed from either end of each read.
Note: Trim Quality Level can be used to trim reads from both ends with defined quality. “N” base cutoff can be used to filter reads which have more than this number of continuous base “N”. Low complexity is defined by the fraction of mono-/di-nucleotide sequence. Ref: FaQCs.

The host removal subsection allows you to subtract host-derived reads from your dataset, which can be useful for metagenomic (complex) samples such as clinical samples (blood, tissue), or environmental samples like insects. In order to enable host removal, within the “Host Removal” subsection of the “Choose Processes / Analyses” section, switch the toggle box to “On” and select either from the pre-build host list (Human, Invertebrate Vectors of Human Pathogens, PhiX, RefSeq Bacteria and RefSeq Viruses) or the appropriate host FASTA file for your experiment from the navigation field. The Similarity (%) can be varied if desired, but the default is 90 and we would not recommend using a value less than 90.

5.4. Choosing processes/analyses
5.4.2 Assembly And Annotation

The Assembly option by default is turned on. It can be turned off via the toggle button. EDGE performs iterative kmers de novo assembly by IDBA-UD. It performs well on isolates as well as metagenomes but it may not work well on very large genomes. By default, it starts from kmer=31 and iterative step by adding 20 to maximum kmer=121. When the maximum k value is larger than the input average reads length, it will automatically adjust the maximum value to average reads length minus 1. User can set the minimum cutoff value on the final contigs. By default, it will filter out all contigs with size smaller than 200 bp.

The Annotation module will be performed only if the assembly option is turned on and reads were successfully assembled. EDGE has the option of using Prokka or RATT to do genome annotation. For most cases, Prokka is the appropriate tool to use, however, if your input is a viral genome with attached reference annotation (GenBank file), RATT is the preferred method. If for some reason the assembly fails (ex: run out of Memory), EDGE will bypass any modules requiring a contigs file including the annotation analysis.

5.4.3 Reference-based Analysis

The reference-based analysis section allows you to map reads/contigs to the provided references, which can be useful for known isolated species such as cultured samples, to get the coverage information and validate the assembled contigs. In order to enable reference-based analysis, switch the toggle box to “On” and select either from the pre-
Given a reference genome fasta file, EDGE will turn on the analysis of the reads/contigs mapping to reference and JBrowse reference track generation. If a GenBank file is provided, EDGE will also turn on variant analysis.

**Note:** If there are more than one sequence in the reference genome fasta (mulit > ), the fasta header must have unique id for each sequence which is defined in the beginning non space words. ex: >unique_id any other annotation

### 5.4.4 Taxonomy Classification

Taxonomic profiling is performed via the “Taxonomy Classification” feature. This is a useful feature not only for complex samples, but also for purified microbial samples (to detect contamination). In the “Community profiling” subsection in the “Choose Processes / Analyses section,” community profiling can be turned on or off via the toggle button.
There is an option to “Always use all reads” or not. If “Always use all reads” is not selected, then only those reads that do not map to the user-supplied reference will be shown in downstream analyses (i.e. the results will only include what is different from the reference). Additionally, the user can use different profiling tools with checkbox selection menu. EDGE uses multiple tools for taxonomy classification including GOTTCHA (bacterial & viral databases), MetaPhlAn, MetaPhyler (short read version), Kraken, MetaScope and reads mapping to NCBI RefSeq using BWA.

Turning on the “Contig-Based Taxonomy Classification” section will initiate mapping contigs against NCBI databases for taxonomy and functional annotations.

### 5.4.5 Phylogenomic Analysis

EDGE supports 5 pre-computed pathogen databases (E.coli, Yersinia, Francisella, Brucella, Bacillus (page 68)) for SNP phylogeny analysis. You can also choose to build your own database by first selecting a build method (either FastTree or RAxML), then selecting a pathogen from the “Search Genomes” search function. You can also add FASTA files or SRA Accessions.
5.4.6 Specialty Genes Profiling

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 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.
5.4.7 PCR Primer Tools

EDGE includes PCR-related tools for use by those who want to use PCR data for their projects.

- **Primer Validation**
  
The “Primer Validation” tool can be used to verify whether and where given primer sequences would align to the genome of the sequenced organism. Prior to initiating the analysis, primer sequences in FASTA format must be deposited in the folder on the desktop in the directory entitled “EDGE Input Directory.”

  In order to initiate primer validation, within the “Primer Validation” subsection switch the “Run Primer Validation” toggle button to “On”. Then, within the “Primer FASTA Sequences” navigation field, select your file containing the primer sequences of interest. Next, in the “Maximum Mismatch” field, choose the maximum number of mismatches you wish to allow per primer sequence. The available options are 0, 1, 2, 3, or 4.

- **Primer Design**
  
If you would like to design new primers that will differentiate a sequenced microorganism from all other bacteria and viruses in NCBI, you can do so using the “Primer Design” tool. To initiate primer design switch the “Run Primer Design” toggle button to “On”. There are default settings supplied for Melting Temperature, Primer Length, Tm Differential, and Number of Primer Pairs, but you can change these settings if desired.
5.5 Submission of a job

When you have selected the appropriate input files and desired analysis options, and you are ready to submit the analysis job, click on the “Submit” button at the bottom of the page. Immediately you will see indicators of successful job submission and job status below the submit button, in green. If there is something wrong with the input, it will stop the submission and show the message in red, highlighting the sections with issues.

![Image of submit button and job status indicators]

5.6 Checking the status of an analysis job

Once an analysis job has been submitted, it will become visible in the left navigation bar. There is a grey, red, orange, green color-coding system that indicates job status as follow:

<table>
<thead>
<tr>
<th>Status</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not yet begun</td>
<td>Grey</td>
</tr>
<tr>
<td>Error</td>
<td>Red</td>
</tr>
<tr>
<td>In progress (running)</td>
<td>Orange</td>
</tr>
<tr>
<td>Completed</td>
<td>Green</td>
</tr>
</tbody>
</table>

While the job is in progress, clicking on the project in the left navigation bar will allow you to see which individual steps have been completed or are in progress, and results that have already been produced. Clicking the job progress widget at top right opens up a more concise view of progress.
### 5.6. Checking the status of an analysis job

#### General

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Run</th>
<th>Status</th>
<th>Running Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Download SRA</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>01:12:21</td>
</tr>
<tr>
<td>Quality Trim and Filter</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>01:10:35</td>
</tr>
<tr>
<td>Host Removal</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>01:10:35</td>
</tr>
<tr>
<td>IDBA Assembly</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>01:06:23</td>
</tr>
<tr>
<td>Reads Mapping To Contigs</td>
<td>Auto</td>
<td>Skipped (result exists)</td>
<td>00:47:30</td>
</tr>
<tr>
<td>Reads Mapping To Reference</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>01:34:40</td>
</tr>
<tr>
<td>Reads Taxonomy Classification</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>01:20:40</td>
</tr>
<tr>
<td>Contigs Mapping To Reference</td>
<td>Auto</td>
<td>Skipped (result exists)</td>
<td>00:00:07</td>
</tr>
<tr>
<td>Variant Analysis</td>
<td>Auto</td>
<td>Skipped (result exists)</td>
<td>00:00:00</td>
</tr>
<tr>
<td>Contigs Taxonomy Classification</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>00:00:31</td>
</tr>
<tr>
<td>Contigs Annotation</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>00:02:12</td>
</tr>
<tr>
<td>ProPhage Detection</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>00:00:45</td>
</tr>
<tr>
<td>Generate, Browse Tracks</td>
<td>On</td>
<td>Skipped (result exists)</td>
<td>00:43:17</td>
</tr>
<tr>
<td>HTML Report</td>
<td>On</td>
<td>Complete</td>
<td>00:01:00</td>
</tr>
</tbody>
</table>

**Report Info**

**Location**

---

**Project Summary**

Description: Transcriptomic analysis of the Novel Middle East Respiratory Syndrome Coronavirus (MERS-CoV)

Human VMERS_MERS-MRC5-HighMCI-24hr-2

Submission Time: 2015 Jun 10 16:05:23

Number of CPUs: 8

Project Status: Complete

Total Analysis Run Time: 00:10:50

Last Run Time: 00:00:10
5.7 Monitoring the Resource Usage

In the job project sidebar, you can see there is an “EDGE Server Usage” widget that dynamically monitors the server resource usage for %CPU, %MEMORY and %DISK space. If there is not enough available disk space, you may consider deleting or archiving the submitted job with the Action tool described below.

5.8 Management of Jobs

Below the resource monitor is the “Action” tool, used for managing jobs in progress or existing projects.
The available actions are:

- **View live log** A terminal-like screen showing all the command lines and progress log information. This is useful for troubleshooting or if you want to repeat certain functions through command line at edge server.
- **Force to rerun this project** Rerun a project with the same inputs and configuration. No additional input needs.
- **Interrupt running project** Immediately stop a running project.
- **Delete entire project** Delete the entire output directory of the project.
- **Remove from project list** Keep the output but remove project name from the project list.
- **Empty project outputs** Clean all the results but keep the config file. User can use this function to do a clean rerun.
- **Move to an archive directory** For performance reasons, the output directory will be put in local storage. User can use this function to move projects from local storage to a slower but larger network storage, which are configured when the edge server is installed.
- **Share Project** Allow guests and other users to view the project.
- **Make project Private/Public** Restrict access to viewing the project to only yourself. Or open it everyone.

### 5.9 Project List Table

When you click “My Project List”, all your projects or projects shared to you will show in a table. It lists the projects status, submission time, running time, type and owner. User can select one or more jobs from the checkbox in the project table and perform actions similar to “Action” Widget described in the previous section. The action will apply to all checked projects.
When mouse over the action buttons on the project list page, it will show a pop up info for the action buttons. There is a special action button for multiple projects, “Compare Selected Projects Taxonomy Classification (HeatMap)” which will draw heatmaps of taxonomy profiling results for multiple projects using MetaComp.

5.10 Other Methods of Accessing EDGE

5.10.1 Internal Python Web Server

EDGE includes a simple web server for single-user applications or other testing. It is not robust enough for production usage, but it is simple enough that it can be run on practically any system.

To run gui, type:

```
$EDGE_HOME/start_edge_ui.sh
```

This will start a localhost and the GUI html page will be opened by your default browser.

5.10.2 Apache Web Server

The preferred installation of EDGE uses Apache 2 (See Testing the EDGE Installation (page 15)), and serves the application as a proper system service. A sample httpd.conf (or apache2.conf, depending on your operating system) is provided in the root directory of your installation. If this configuration is used, EDGE will be available on any IP or hostname registered to the machine, on ports 80 and 8080.

You can access EDGE by opening either the desktop link (below), or your browser, and entering http://localhost:80 in the address bar.
Note: If the desktop environment is available, after installation, a “Start EDGE UI” icon should be on the desktop. Click on the green icon and choose “Run in Terminal.” Results should be the same as those obtained by the above method to start the GUI.

The URL address is 127.0.0.1:8080/index.html. It may not be that powerful, as it is hosted by Apache HTTP Server, but it works. With system administrator help, the Apache HTTP Server is the suggested method to host the gui interface.

Note: You may need to configure the edge_wwwroot and input and output in the edge_ui/edge_config.tmpl file while configuring the Apache HTTP Server and link to external drive or network drive if needed.

A Terminal window will display messages and errors as you run EDGE. Under normal operating conditions you can minimize this window. Should an error/problem arise, you may maximize this window to view the error.
Warning: IMPORTANT: Do not close this window!

The Browser window is the window in which you will interact with EDGE.
CHAPTER 6

Command Line Interface (CLI)

The command line usage is as followings:

Usage: perl runPipeline.pl [options] -c config.txt -p 'reads1.fastq reads2.fastq' -o <out_directory>
Version 1.1

Input File:
- u Unpaired reads, Single end reads in fastq
- p Paired reads in two fastq files and separate by space in quote
- c Config File

Output:
- o Output directory.

Options:
- ref Reference genome file in fasta
- primer A pair of Primers sequences in strict fasta format
- cpu number of CPUs (default: 8)
- version print version

A config file (example in the below section, the Graphic User Interface (GUI) (page 28) will generate config automatically), reads Files in fastq format, and a output directory are required when run by command line. Based on the configuration file, if all modules are turned on, EDGE will run the following steps. Each step contains at least one command line scripts/programs.

1. Data QC
2. Host Removal QC
3. De novo Assembling
4. Reads Mapping To Contig
5. Reads Mapping To Reference Genomes
6. Taxonomy Classification on All Reads or unMapped to Reference Reads
7. Map Contigs To Reference Genomes
8. Variant Analysis
9. Contigs Taxonomy Classification
10. Contigs Annotation
11. ProPhage detection
12. PCR Assay Validation
13. PCR Assay Adjudication
14. Phylogenetic Analysis
15. Generate JBrowse Tracks
16. HTML report

6.1 Configuration File

The config file is a text file with the following information. If you are going to do host removal, you need to build host index (page 68) for it and change the fasta file path in the config file.

```plaintext
[Count Fastq]
DoCountFastq=auto

[Quality Trim and Filter]
## boolean, 1=yes, 0=no
DoQC=1
## Targets quality level for trimming
q=5
## Trimmed sequence length will have at least minimum length
min_L=50
## Average quality cutoff
avg_q=0
## "N" base cutoff. Trimmed read has more than this number of continuous base "N" will be discarded.
n=1
## Low complexity filter ratio, Maximum fraction of mono-/di-nucleotide sequence
lc=0.85
## Trim reads with adapters or contamination sequences
adapter=/PATH/adapter.fasta
## phiX filter, boolean, 1=yes, 0=no
phiX=0
## Cut # bp from 5 end before quality trimming/filtering
5end=0
## Cut # bp from 3 end before quality trimming/filtering
3end=0

[Host Removal]
## boolean, 1=yes, 0=no
DoHostRemoval=1
## Use more Host= to remove multiple host reads
Host=/PATH/all_chromosome.fasta
similarility=90
```

(continues on next page)
[Assembly]
## boolean, 1=yes, 0=no
DoAssembly=1
## Bypass assembly and use pre-assembled contigs
assembledContigs=
minContigSize=200
## spades or idba_ud
assembler=idba_ud
idbaOptions="--pre_correction --mink 31"
## for spades
singleCellMode=
pacbioFile=
nanoporeFile=

[Reads Mapping To Contigs]
# Reads mapping to contigs
DoReadsMappingContigs=auto

[Reads Mapping To Reference]
# Reads mapping to reference
DoReadsMappingReference=0
bowtieOptions=
# reference genbank or fasta file
reference=
MapUnmappedReads=0

[Reads Taxonomy Classification]
## boolean, 1=yes, 0=no
DoReadsTaxonomy=1
## If reference genome exists, only use unmapped reads to do Taxonomy Classification.
## Turn on AllReads=1 will use all reads instead.
AllReads=0
enabledTools=gottcha-genDB-b,gottcha-speDB-b,gottcha-strDB-b,gottcha-genDB-v,gottcha-speDB-v,gottcha-strDB-v,metaphlan,bwa,kraken_mini

[Contigs Mapping To Reference]
# Contig mapping to reference
DoContigMapping=auto
## identity cutoff
identity=85
MapUnmappedContigs=0

[Variant Analysis]
DoVariantAnalysis=auto

[Contigs Taxonomy Classification]
DoContigsTaxonomy=1

[Contigs Annotation]
## boolean, 1=yes, 0=no
DoAnnotation=1
# kingdom: Archaea Bacteria Mitochondria Viruses
kingdom=Bacteria
contig_size_cut_for_annotation=700
## support tools: Prokka or RATT
annotateProgram=Prokka

(continues on next page)
6.2 Test Run

EDGE provides an example data set which is an E. coli MiSeq dataset and has been subsampled to ~10x fold coverage reads.

In the EDGE home directory,

```bash
cd testData
sh runTest.sh
```

See Output (page 63)
6.2. Test Run

Fig. 1: Snapshot from the terminal.
6.3 Descriptions of each module

Each module comes with default parameters and user can see the optional parameters by entering the program name with –h or -help flag without any other arguments.

1. **Data QC**
   - Required step? **No**
   - Command example
   ```
   perl $EDGE_HOME/scripts/illumina_fastq_QC.pl -p 'Ecoli_10x.1.fastq Ecoli_10x.2.fastq' -q 5 -min_L 50 -avg_q 5 -n 0 -lc 0.85 -d QcReads -t 10
   ```
   - What it does
     - Quality control
     - Read filtering
     - Read trimming
   - Expected input
     - Paired-end/Single-end reads in FASTQ format
   - Expected output
     - QC.1.trimmed.fastq
     - QC.2.trimmed.fastq
     - QC.unpaired.trimmed.fastq
     - QC.stats.txt
     - QCqc_report.pdf

2. **Host Removal QC**
   - Required step? **No**
   - Command example
   ```
   perl $EDGE_HOME/scripts/host_reads_removal_by_mapping.pl -p 'QC.1.trimmed.fastq QC.2.trimmed.fastq' -u QC.unpaired.trimmed.fastq -ref human_chromosomes.fasta -o QcReads -cpu 10
   ```
   - What it does
     - Read filtering
   - Expected input
     - Paired-end/Single-end reads in FASTQ format
   - Expected output
     - host_clean.1.fastq
     - host_clean.2.fastq
     - host_clean.mapping.log
     - host_clean.unpaired.fastq
     - host_clean.stats.txt
3. **IDBA Assembling**
   - Required step? No
   - Command example
     ```
     fq2fa --merge host_clean.1.fastq host_clean.2.fastq pairedForAssembly.fasta
     idba_ud --num_threads 10 -o AssemblyBasedAnalysis/idba --pre_correction
     pairedForAssembly.fasta
     ```
   - What it does
     - Iterative kmers de novo Assembly, it performs well on isolates as well as metagenomes. It may not work well on very large genomes.
   - Expected input
     - Paired-end/Single-end reads in FASTA format
   - Expected output
     - contig.fa
     - scaffold.fa (input paired end)

4. **Reads Mapping To Contig**
   - Required step? No
   - Command example
     ```
     perl $EDGE_HOME/scripts/runReadsToContig.pl -p 'host_clean.1.fastq host_clean.2.
     fastq' -d AssemblyBasedAnalysis/readsMappingToContig -pre readsToContigs -ref
     AssemblyBasedAnalysis/contigs.fa
     ```
   - What it does
     - Mapping reads to assembled contigs
   - Expected input
     - Paired-end/Single-end reads in FASTQ format
     - Assembled Contigs in Fasta format
     - Output Directory
     - Output prefix
   - Expected output
     - readsToContigs.alnstats.txt
     - readsToContigs_coverage.table
     - readsToContigs_plots.pdf
     - readsToContigs.sort.bam
     - readsToContigs.sort.bam.bai

5. **Reads Mapping To Reference Genomes**
   - Required step? No
   - Command example:
• What it does
  – Mapping reads to reference genomes
  – SNPs/Indels calling

• Expected input
  – Paired-end/Single-end reads in FASTQ format
  – Reference genomes in Fasta format
  – Output Directory
  – Output prefix

• Expected output
  – readsToRef.alnstats.txt
  – readsToRef_plots.pdf
  – readsToRef_refID.coverage
  – readsToRef_refID.gap.coords
  – readsToRef_refID.window_size_coverage
  – readsToRef.ref_windows_gc.txt
  – readsToRef.raw.bcf
  – readsToRef.sort.bam
  – readsToRef.sort.bam.bai
  – readsToRef.vcf

6. Taxonomy Classification on All Reads or unMapped to Reference Reads

• Required step? No

• Command example:

```
perl $EDGE_HOME/scripts/microbial_profiling/microbial_profiling_configure.pl
  →$EDGE_HOME/scripts/microbial_profiling/microbial_profiling.settings.tmpl
  →gottcha-speDB-b > microbial_profiling.settings.ini
perl $EDGE_HOME/scripts/microbial_profiling/microbial_profiling.pl -o Taxonomy -
  →s microbial_profiling.settings.ini -c 10 UnmappedReads.fastq
```

• What it does
  – Taxonomy Classification using multiple tools, including BWA mapping to NCBI Refseq, metaphlan, kraken, GOTTCHA.
  – Unify varies output format and generate reports

• Expected input
  – Reads in FASTQ format
  – Configuration text file (generated by microbial_profiling_configure.pl)

• Expected output
– Summary EXCEL and text files.
– Heatmaps tools comparison
– Radarchart tools comparison
– Krona and tree-style plots for each tool.

7. Map Contigs To Reference Genomes

• Required step? No

• Command example:

```perl
perl $EDGE_HOME/scripts/nucmer_genome_coverage.pl -e 1 -i 85 -p contigsToRef_
Reference.fna contigs.fa
```

• What it does
  – Mapping assembled contigs to reference genomes
  – SNPs/Indels calling

• Expected input
  – Reference genome in Fasta Format
  – Assembled contigs in Fasta Format
  – Output prefix

• Expected output
  – contigsToRef_avg_coverage.table
  – contigsToRef.delta
  – contigsToRef_query_unUsed.fasta
  – contigsToRef.snps
  – contigsToRef.coords
  – contigsToRef.log
  – contigsToRef_query_novel_region_coord.txt
  – contigsToRef_ref_zero_cov_coord.txt

8. Variant Analysis

• Required step? No

• Command example:

```perl
perl $EDGE_HOME/scripts/SNP_analysis.pl -genbank Reference.gbk -SNP contigsToRef_
snps -format nucmer
perl $EDGE_HOME/scripts/gap_analysis.pl -genbank Reference.gbk -gap contigsToRef_
ref_zero_cov_coord.txt
```

• What it does
  – Analyze variants and gaps regions using annotation file.

• Expected input
  – Reference in GenBank format
  – SNPs/INDELs/Gaps files from “Map Contigs To Reference Genomes”
• Expected output
  – contigsToRef.SNPs_report.txt
  – contigsToRef.Indels_report.txt
  – GapVSReference.report.txt

9. **Contigs Taxonomy Classification**

• Required step? No

• Command example:

```
perl $EDGE_HOME/scripts/contig_classifier_by_bwa/contig_classifier_by_bwa.pl --db
$EDGE_HOME/database/bwa_index/NCBI-Bacteria-Virus.fna --threads 10 --prefix
OutputCT --input contigs.fa
```

• What it does
  – Taxonomy Classification on contigs using BWA mapping to NCBI Refseq

• Expected input
  – Contigs in Fasta format
  – NCBI Refseq genomes bwa index
  – Output prefix

• Expected output
  – prefix.assembly_class.csv
  – prefix.assembly_class.top.csv
  – prefix.ctg_class.csv
  – prefix.ctg_class.LCA.csv
  – prefix.ctg_class.top.csv
  – prefix.unclassified.fasta

10. **Contig Annotation**

• Required step? No

• Command example:

```
prokka --force --prefix PROKKA --outdir Annotation contigs.fa
```

• What it does
  – The rapid annotation of prokaryotic genomes.

• Expected input
  – Assembled Contigs in Fasta format
  – Output Directory
  – Output prefix

• Expected output
  – It produces GFF3, GBK and SQN files that are ready for editing in Sequin and ultimately submitted to Genbank/DDJB/ENA.

6.3. **Descriptions of each module**
11. **ProPhage detection**

- Required step? **No**
- Command example:

```bash
perl $EDGE_HOME/scripts/phageFinder_prepare.pl -o Prophage -p Assembly Annotation/
  →PROKKA.gff Annotation/PROKKA.fna
$EDGE_HOME/thirdParty/phage_finder_v2.1/bin/phage_finder_v2.1.sh Assembly
```

- What it does
  - Identify and classify prophages within prokaryotic genomes.
- Expected input
  - Annotated Contigs GenBank file
  - Output Directory
  - Output prefix
- Expected output
  - phageFinder_summary.txt

12. **PCR Assay Validation**

- Required step? **No**
- Command example:

```bash
perl $EDGE_HOME/scripts/pcrValidation/validate_primers.pl -ref contigs.fa -primer, ˓
  →primers.fa -mismatch 1 -output AssayCheck
```

- What it does
  - In silico PCR primer validation by sequence alignment.
- Expected input
  - Assembled Contigs/Reference in Fasta format
  - Output Directory
  - Output prefix
- Expected output
  - pcrContigValidation.log
  - pcrContigValidation.bam

13. **PCR Assay Adjudication**

- Required step? **No**
- Command example:

```bash
perl $EDGE_HOME/scripts/pcrAdjudication/pcrUniquePrimer.pl --input contigs.fa -- ˓
  →gff3 PCR.Adjudication.primers.gff3
```

- What it does
  - Design unique primer pairs for input contigs.
- Expected input
Assembled Contigs in Fasta format
Output gff3 file name

• Expected output
  • PCR.Adjudication.primers.gff3
  • PCR.Adjudication.primers.txt

14. **Phylogenetic Analysis**

• Required step? **No**

• Command example:

```bash
perl $EDGE_HOME/scripts/prepare_SNP_phylogeny.pl -o output/SNP_Phylogeny/Ecoli -tree FastTree -db Ecoli -n output -cpu 10 -p QC.1.trimmed.fastq QC.2.trimmed.fastq -c contigs.fa -s QC.unpaired.trimmed.fastq
perl $EDGE_HOME/scripts/SNPphy/runSNPphylogeny.pl output/SNP_Phylogeny/Ecoli/SNPphy.ctrl
```

• What it does
  • Perform SNP identification against selected pre-built SNPdb or selected genomes
  • Build SNP based multiple sequence alignment for all and CDS regions
  • Generate Tree file in newick/PhyloXML format

• Expected input
  • SNPdb path or genomesList
  • Fastq reads files
  • Contig files

• Expected output
  • SNP based phylogenetic multiple sequence alignment
  • SNP based phylogenetic tree in newick/PhyloXML format.
  • SNP information table

15. **Generate JBrowse Tracks**

• Required step? **No**

• Command example:

```bash
perl $EDGE_HOME/scripts/edge2jbrowse_converter.pl --in-ref-fa Reference.fna --in-ref-gff3 Reference.gff --proj_outdir EDGE_project_dir
```

• What it does
  • Convert several EDGE outputs into JBrowse tracks for visualization for contigs and reference, respectively.

• Expected input
  • EDGE project output Directory

• Expected output
  • EDGE post-processed files for JBrowse tracks in the JBrowse directory.
  • Tracks configuration files in the JBrowse directory.
16. **HTML Report**

- Required step? No
- Command example:
  ```bash
  perl $EDGE_HOME/scripts/munger/outputMunger_w_temp.pl EDGE_project_dir
  ```
- What it does
  - Generate statistical numbers and plots in an interactive html report page.
- Expected input
  - EDGE project output Directory
- Expected output
  - report.html

### 6.4 Other command-line utility scripts

1. To extract certain taxa fasta from contig classification result:
   ```bash
   cd /home/edge_install/edge_ui/EDGE_output/41/AssemblyBasedAnalysis/Taxonomy
   perl /home/edge_install/scripts/contig_classifier_by_bwa/extract_fasta_by_taxa.pl
   -fasta ../contigs.fa -csv ProjectName.ctg_class.top.csv -taxa "Enterobacter cloacae"
   > Ecloacae.contigs.fa
   ```

2. To extract unmapped/mapped reads fastq from the bam file:
   ```bash
   cd /home/edge_install/edge_ui/EDGE_output/41/AssemblyBasedAnalysis/
   readsMappingToContig
   # extract unmapped reads
   perl /home/edge_install/scripts/bam_to_fastq.pl -unmapped readsToContigs.sort.bam
   # extract mapped reads
   perl /home/edge_install/scripts/bam_to_fastq.pl -mapped readsToContigs.sort.bam
   ```

3. To extract mapped reads fastq of a specific contig/reference from the bam file:
   ```bash
   cd /home/edge_install/edge_ui/EDGE_output/41/AssemblyBasedAnalysis/
   readsMappingToContig
   perl /home/edge_install/scripts/bam_to_fastq.pl -id ProjectName_00001 -mapped
   ```
The output directory structure contains ten major sub-directories when all modules are turned on. In addition to the main directories, EDGE will generate a final report in portable document file format (pdf), process log and error log file in the project main directory.

- AssayCheck
- AssemblyBasedAnalysis
- HostRemoval
- HTML_Report
- JBrowse
- QcReads
- ReadsBasedAnalysis
- ReferenceBasedAnalysis
- Reference
- SNP_Phylogeny

In the graphic user interface, EDGE generates an interactive output webpage which includes summary statistics and taxonomic information, etc. The easiest way to interact with the results is through the web interface. If a project run finished through the command line, user can open the report html file in the HTML_report subdirectory off-line. When a project run is finished, user can click on the project id from the menu and it will generate the interactive html report on the fly. User can browse the data structure by clicking the project link and visualize the result by JBrowse links, download the pdf files, etc.
7.1 Example Output

See http://lanl-bioinformatics.github.io/EDGE/example_output/report.html

Note: The example link is just an example of graphic output. The JBrowse and links are not accessible in the example links.
8.1 EDGE provided databases

8.1.1 Taxonomy Database Info Table

https://lanl-bioinformatics.github.io/EDGE/docs/taxonomyDBtable.html

8.1.2 NCBI Refseq

EDGE prebuilt blast db and bwa_index of NCBI RefSeq genomes.

**Warning**: NCBI restructure the ftp site. The link for Bacteria below is an archive.

- **Bacteria**: NCBI all complete bacteria download method
  - Version: NCBI 2017 Oct 3
  - 245 Archaea + 7917 Bacteria genomes
- **Virus**: NCBI Virus
  - Version: NCBI 2017 Oct 3
  - 7458 complete genomes + Neighbor Nucleotoides (118039 sequences)

see $EDGE_HOME/database/bwa_index/id_mapping.txt for all gi/accession to genome name lookup table.

8.1.3 Krona taxonomy

- **website**: http://sourceforge.net/p/krona/home/krona/
**Update Krona taxonomy db**

Download these files from ftp://ftp.ncbi.nih.gov/pub/taxonomy:

```bash
```

Transfer the files to the taxonomy folder in the standalone KronaTools installation and run:

```bash
```

### 8.1.4 Metaphlan2 database

MetaPhlAn2 relies on unique clade-specific marker genes identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic)

- **paper:** http://www.ncbi.nlm.nih.gov/pubmed/?term=22688413
- **website:** http://huttenhower.sph.harvard.edu/metaphlan2

### 8.1.5 Human Genome

The bwa index is prebuilt in the EDGE. The human hs_ref_GRCh38 sequences from NCBI ftp site.


### 8.1.6 Kraken2 DB

Kraken2 is a system for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomic studies. Kraken2 database in EDGE is a pre-built database constructed from Refseq bacteria, archaea, and viral libraries and the GRCh38 human genome and UniVec_Core in RefSeq (as of Dec 20, 2018).

- **Kraken1 paper:** http://www.ncbi.nlm.nih.gov/pubmed/?term=24580807
- **website:** http://ccb.jhu.edu/software/kraken2/

### 8.1.7 GOTTCHA DB

A novel, annotation-independent and signature-based metagenomic taxonomic profiling tool.

- **website:** http://lanl-bioinformatics.github.io/GOTTCHA/
- **ftp:** https://edge-dl.lanl.gov/gottcha/
- **version:** v20150825

### 8.1.8 SNPdb

SNP database based on whole genome comparison. Current available db are *Ecoli, Yersinia, Francisella, Brucella, Bacillus* (page 68).  

8.1. EDGE provided databases
8.1.9 Invertebrate Vectors of Human Pathogens

The bwa index is prebuilt in the EDGE.

- website: https://www.vectorbase.org

version: 2014 July 24

8.1.10 NCBI Nucleotide database (NT) database

- version: 2016 April 26

8.1.11 VFDB

A Microbial database of virulence factors

- website: http://www.mgc.ac.cn/VFs/main.htm
- version: 20160818

8.1.12 ARDB

Antibiotic Resistance Genes Database

- website: http://ardb.cbcb.umd.edu/index.html
- version: 1.1

8.1.13 CARD

The Comprehensive Antibiotic Resistance Database

- website: https://card.mcmaster.ca/
- Version: 3.0.0

8.1.14 Amplicon: 16s/18s/ITS

For QIIME (Quantitative insights into Microbial Ecology) analysis

- Greengenes OTUs (16s)
  - website: http://greengenes.secondgenome.com/
  - version: 2013 May
- SILVA OTUs (16S/18S)
  - website: http://www.arb-silva.de/download/archive/qiime/
  - version: 119
- UNITE OTUs (ITS)
8.2 Building bwa index

Here take human genome as example.

1. Download the human hs_ref_GRCh38 sequences from NCBI ftp site.


   perl $EDGE_HOME/scripts/download_human_refseq_genome.pl output_dir

2. Gunzip the downloaded fasta file and concatenate them into one human genome multifasta file:

   gunzip hs_ref_GRCh38.*.fa.gz
   cat hs_ref_GRCh38.*.fa > human_ref_GRCh38.all.fasta

3. Use the installed bwa to build the index:

   $EDGE_HOME/bin/bwa index human_ref_GRCh38.all.fasta

   Now, you can configure the config file with “host=/path/human_ref_GRCh38.all.fasta” for host removal step.

8.3 SNP database genomes

SNP database was pre-built from the below genomes.

8.3.1 Ecoli Genomes

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecoli_042</td>
<td>Escherichia coli 042, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a></td>
</tr>
<tr>
<td>Ecoli_APEC_O1</td>
<td>Escherichia coli APEC O1 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a></td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>URL</td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>Ecoli_CFT073</td>
<td>Escherichia coli CFT073 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/74310614">http://www.ncbi.nlm.nih.gov/nuccore/74310614</a></td>
</tr>
<tr>
<td>Ecoli_DH1</td>
<td>Escherichia coli DH1, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/26245917">http://www.ncbi.nlm.nih.gov/nuccore/26245917</a></td>
</tr>
<tr>
<td>Ecoli_E24377A</td>
<td>Escherichia coli E24377A chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/30061571">http://www.ncbi.nlm.nih.gov/nuccore/30061571</a></td>
</tr>
<tr>
<td>Ecoli_ED1a</td>
<td>Escherichia coli ED1a chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/377520096">http://www.ncbi.nlm.nih.gov/nuccore/377520096</a></td>
</tr>
<tr>
<td>Ecoli_ETEC_H10407</td>
<td>Escherichia coli ETEC H10407, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386703215">http://www.ncbi.nlm.nih.gov/nuccore/386703215</a></td>
</tr>
<tr>
<td>Ecoli_HS</td>
<td>Escherichia coli HS, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/387610477">http://www.ncbi.nlm.nih.gov/nuccore/387610477</a></td>
</tr>
<tr>
<td>Ecoli_IAI1</td>
<td>Escherichia coli IAI1 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/218687878">http://www.ncbi.nlm.nih.gov/nuccore/218687878</a></td>
</tr>
<tr>
<td>Ecoli_IHE3034</td>
<td>Escherichia coli IHE3034 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/209917191">http://www.ncbi.nlm.nih.gov/nuccore/209917191</a></td>
</tr>
<tr>
<td>Ecoli_KO11FL</td>
<td>Escherichia coli KO11FL chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/82775382">http://www.ncbi.nlm.nih.gov/nuccore/82775382</a></td>
</tr>
<tr>
<td>Ecoli_LF82</td>
<td>Escherichia coli LF82, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/157159467">http://www.ncbi.nlm.nih.gov/nuccore/157159467</a></td>
</tr>
<tr>
<td>Ecoli_NA114</td>
<td>Escherichia coli NA114 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/218698419">http://www.ncbi.nlm.nih.gov/nuccore/218698419</a></td>
</tr>
<tr>
<td>Ecoli_P12b</td>
<td>Escherichia coli P12b chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386611756">http://www.ncbi.nlm.nih.gov/nuccore/386611756</a></td>
</tr>
<tr>
<td>Ecoli_S88</td>
<td>Escherichia coli S88 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/387504934">http://www.ncbi.nlm.nih.gov/nuccore/387504934</a></td>
</tr>
<tr>
<td>Ecoli_Sakai</td>
<td>Escherichia coli SMS-3-5 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/388476123">http://www.ncbi.nlm.nih.gov/nuccore/388476123</a></td>
</tr>
<tr>
<td>Ecoli_UM146</td>
<td>Escherichia coli UM146 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/384541581">http://www.ncbi.nlm.nih.gov/nuccore/384541581</a></td>
</tr>
<tr>
<td>Ecoli_UMN026</td>
<td>Escherichia coli UMN026 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386632422">http://www.ncbi.nlm.nih.gov/nuccore/386632422</a></td>
</tr>
<tr>
<td>Ecoli_UMNK88</td>
<td>Escherichia coli UMNK88 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/82775382">http://www.ncbi.nlm.nih.gov/nuccore/82775382</a></td>
</tr>
<tr>
<td>Ecoli_UTI89</td>
<td>Escherichia coli UTI89 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/387610477">http://www.ncbi.nlm.nih.gov/nuccore/387610477</a></td>
</tr>
<tr>
<td>Ecoli_W</td>
<td>Escherichia coli W chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/387610477">http://www.ncbi.nlm.nih.gov/nuccore/387610477</a></td>
</tr>
<tr>
<td>Ecoli_Xuzhou21</td>
<td>Escherichia coli Xuzhou21 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/218552585">http://www.ncbi.nlm.nih.gov/nuccore/218552585</a></td>
</tr>
<tr>
<td>Sboydii_Sb227</td>
<td>Shigella boydii Sb227 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386612163">http://www.ncbi.nlm.nih.gov/nuccore/386612163</a></td>
</tr>
<tr>
<td>Sdysenteriae_Sd197</td>
<td>Shigella dysenteriae Sd197, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386612163">http://www.ncbi.nlm.nih.gov/nuccore/386612163</a></td>
</tr>
<tr>
<td>Sflexneri_2002017</td>
<td>Shigella flexneri 2002017 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386612163">http://www.ncbi.nlm.nih.gov/nuccore/386612163</a></td>
</tr>
<tr>
<td>Sflexneri_2a_2457T</td>
<td>Shigella flexneri 2a str. 2457T, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386612163">http://www.ncbi.nlm.nih.gov/nuccore/386612163</a></td>
</tr>
<tr>
<td>Sflexneri_2a_301</td>
<td>Shigella flexneri 2a str. 301 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386612163">http://www.ncbi.nlm.nih.gov/nuccore/386612163</a></td>
</tr>
<tr>
<td>Sflexneri_5_8401</td>
<td>Shigella flexneri 5 str. 8401 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386612163">http://www.ncbi.nlm.nih.gov/nuccore/386612163</a></td>
</tr>
<tr>
<td>Ssonnei_Ss046</td>
<td>Shigella sonnei Ss046 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/386612163">http://www.ncbi.nlm.nih.gov/nuccore/386612163</a></td>
</tr>
</tbody>
</table>

8.3. SNP database genomes
# 8.3.2 Yersinia Genomes

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ypestis_A1122</td>
<td>Yersinia pestis A1122 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/384137007">http://www.ncbi.nlm.nih.gov/nuccore/384137007</a></td>
</tr>
<tr>
<td>Ypestis_Angola</td>
<td>Yersinia pestis Angola chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/162418099">http://www.ncbi.nlm.nih.gov/nuccore/162418099</a></td>
</tr>
<tr>
<td>Ypestis_Antiqua</td>
<td>Yersinia pestis Antiqua chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/108805998">http://www.ncbi.nlm.nih.gov/nuccore/108805998</a></td>
</tr>
<tr>
<td>Ypestis_CO92</td>
<td>Yersinia pestis CO92 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/16120353">http://www.ncbi.nlm.nih.gov/nuccore/16120353</a></td>
</tr>
<tr>
<td>Ypestis_D106004</td>
<td>Yersinia pestis D106004 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/384120592">http://www.ncbi.nlm.nih.gov/nuccore/384120592</a></td>
</tr>
<tr>
<td>Ypestis_D182038</td>
<td>Yersinia pestis D182038 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/384124469">http://www.ncbi.nlm.nih.gov/nuccore/384124469</a></td>
</tr>
<tr>
<td>Ypestis_KIM_10</td>
<td>Yersinia pestis KIM 10 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/22123922">http://www.ncbi.nlm.nih.gov/nuccore/22123922</a></td>
</tr>
<tr>
<td>Ypestis_Microtus_91001</td>
<td>Yersinia pestis biovar Microtus str. 91001 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/45439865">http://www.ncbi.nlm.nih.gov/nuccore/45439865</a></td>
</tr>
<tr>
<td>Ypestis_Pestoides_F</td>
<td>Yersinia pestis Pestoides F chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/145597324">http://www.ncbi.nlm.nih.gov/nuccore/145597324</a></td>
</tr>
<tr>
<td>Ypestis_Z176003</td>
<td>Yersinia pestis Z176003 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/294502110">http://www.ncbi.nlm.nih.gov/nuccore/294502110</a></td>
</tr>
<tr>
<td>Ypseudotuberculosis_PB1</td>
<td>Yersinia pseudotuberculosis PB1/+ chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/186893344">http://www.ncbi.nlm.nih.gov/nuccore/186893344</a></td>
</tr>
<tr>
<td>Ypseudotuberculosis_YPIII</td>
<td>Yersinia pseudotuberculosis YPIII chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/170022262">http://www.ncbi.nlm.nih.gov/nuccore/170022262</a></td>
</tr>
</tbody>
</table>
## 8.3.3 Francisella Genomes

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ftularensis_holarctica_F92</td>
<td>Francisella tularensis subsp. holarctica F92 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/423049750">http://www.ncbi.nlm.nih.gov/nuccore/423049750</a></td>
</tr>
<tr>
<td>Ftularensis_holarctica_FSC200</td>
<td>Francisella tularensis subsp. holarctica FSC200 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/422937995">http://www.ncbi.nlm.nih.gov/nuccore/422937995</a></td>
</tr>
<tr>
<td>Ftularensis_holarctica_FTNF00200</td>
<td>Francisella tularensis subsp. holarctica FTNF002-00 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/156501369">http://www.ncbi.nlm.nih.gov/nuccore/156501369</a></td>
</tr>
<tr>
<td>Ftularensis_holarctica_LVS</td>
<td>Francisella tularensis subsp. holarctica LVS chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/89255449">http://www.ncbi.nlm.nih.gov/nuccore/89255449</a></td>
</tr>
<tr>
<td>Ftularensis_holarctica_OSU18</td>
<td>Francisella tularensis subsp. holarctica OSU18 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/115313981">http://www.ncbi.nlm.nih.gov/nuccore/115313981</a></td>
</tr>
<tr>
<td>Ftularensis_TIGB03</td>
<td>Francisella tularensis TIGB03 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/379716390">http://www.ncbi.nlm.nih.gov/nuccore/379716390</a></td>
</tr>
<tr>
<td>Ftularensis_tularensis_FSC198</td>
<td>Francisella tularensis subsp. tularensis FSC198 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/110669657">http://www.ncbi.nlm.nih.gov/nuccore/110669657</a></td>
</tr>
<tr>
<td>Ftularensis_tularensis_NE061598</td>
<td>Francisella tularensis subsp. tularensis NE061598 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/385793751">http://www.ncbi.nlm.nih.gov/nuccore/385793751</a></td>
</tr>
<tr>
<td>Ftularensis_tularensis_SCHU_S4</td>
<td>Francisella tularensis subsp. tularensis SCHU S4 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/255961454">http://www.ncbi.nlm.nih.gov/nuccore/255961454</a></td>
</tr>
<tr>
<td>Ftularensis_tularensis_TI0902</td>
<td>Francisella tularensis subsp. tularensis TI0902 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/379725073">http://www.ncbi.nlm.nih.gov/nuccore/379725073</a></td>
</tr>
<tr>
<td>Ftularensis_tularensis_WY9634</td>
<td>Francisella tularensis subsp. tularensis WY96-3418 chromosome, complete genome</td>
<td><a href="http://www.ncbi.nlm.nih.gov/nuccore/134301169">http://www.ncbi.nlm.nih.gov/nuccore/134301169</a></td>
</tr>
</tbody>
</table>
### 8.3.4 Brucella Genomes

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bmelitensis_i_1_16M</td>
<td>Brucella melitensis bv. 1 str. 16M</td>
<td><a href="http://www.ncbi.nlm.nih.gov/bioproject/">http://www.ncbi.nlm.nih.gov/bioproject/</a></td>
</tr>
</tbody>
</table>
## 8.3.5 Bacillus Genomes

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
</table>

### 8.3. SNP database genomes
8.4 Ebola Reference Genomes

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
</table>
9.1 Assembly

- **IDBA-UD**
  - Citation: Peng, Y., et al. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, Bioinformatics, 28, 1420-1428.
  - Site: http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/
  - Version: 1.1.1
  - License: GPLv2

- **SPAdes**
  - Site: http://bioinf.spbau.ru/spades
  - Version: 3.13.1
  - License: GPLv2

- **MEGAHIT**
  - Citation: Li D. et al. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015 May 15;31(10):1674-6
  - Site: https://github.com/voutcn/megahit
  - Version: 1.1.3
  - License: GPLv3

- **LRASM: Long Read Assembler**
  - Citation:
  - Site: https://gitlab.com/chienchi/long_read_assembly
• RACON
  – Site: https://github.com/isovic/racon
  – Version: 1.3.1
  – License: MIT
• Unicycler
  – Site: https://github.com/rrwick/Unicycler
  – Version: 0.4.7
  – License: GPLv3

9.2 Annotation

• RATT
  – Site: http://ratt.sourceforge.net/
  – Version:
  – License: GPLv3
  – Note: The original RATT program does not deal with reverse complement strain annotations transfer. We edited the source code to fix it.
• Prokka
  – Citation: Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation, Bioinformatics, 30,2068-2069.
  – Site: http://www.vicbioinformatics.com/software.prokka.shtml
  – Version: 1.14
  – License: GPLv2
  – Note: The NCBI tool tbl2asn included within PROKKA can have very slow runtimes (up to several hours) while it is dealing with numerous contigs, such as when we input metagenomic data. We modified the code to allow parallel processing using tbl2asn.
• tRNAscan
  – Site: http://lowelab.ucsc.edu/tRNAscan-SE/
  – Version: 1.3.1
  – License: GPLv2
• Barnmap
  – Citation:
  – Version: 0.9
  – License: GPLv3
• BLAST+
  – Citation: Camacho, C., et al. (2009) BLAST+: architecture and applications, BMC bioinformatics, 10, 421.
  – Version: 2.9.0
  – License: Public domain
• blastall
  – Version: 2.2.26
  – License: Public domain
• Phage_Finder
  – Site: http://phage-finder.sourceforge.net/
  – Version: 2.1
  – License: GPLv3
• Glimmer
  – Site: http://ccb.jhu.edu/software/glimmer/index.shtml
  – Version: 302b
  – License: Artistic License
• ARAGORN
  – Citation: Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences, Nucleic acids research, 32, 11-16.
  – Site: http://mbio-serv2.mbioekol.lu.se/ARAGORN/
  – Version: 1.2.36
  – License: GPLv2
• Prodigal
• tbl2asn
  – Citation:
  – Version: 25.7 (2019 Feb 26)
  – License: Public Domain

Warning: tbl2asn must be compiled within the past year to function. We attempt to recompile every 6 months or so. Most recent compilation is 27 Feb 2018

• AntiSmash
  – Citation: Kai Blin et al. (2017) antiSMASH 4.0 — improvements in chemistry prediction and gene cluster boundary identification, Nucleic Acids Research Volume 45, Issue W1, 3 July 2017, Pages W36–W41
  – Site: https://antismash.secondarymetabolites.org/#/start
  – Version: 2.4.1
  – License: AGPL-3.0

9.3 Alignment

• HMMER3
  – Citation: Eddy, S.R. (2011) Accelerated Profile HMM Searches, PLoS computational biology, 7, e1002195
  – Site: http://hmmer.janelia.org/
  – Version: 3.1b1
  – License: GPLv3

• Infernal
  – Citation: Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches, Bioinformatics, 29, 2933-2935.
  – Site: http://infernal.janelia.org/
  – Version: 1.1rc4
  – License: GPLv3

• Bowtie 2
  – Version: 2.2.6
  – License: GPLv3
• BWA
  – Citation: Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics, 25, 1754-1760.
  – Site: http://bio-bwa.sourceforge.net/
  – Version: 0.7.12
  – License: GPLv3

• MUMmer3
  – Citation: Kurtz, S., et al. (2004) Versatile and open software for comparing large genomes, Genome biology, 5, R12.
  – Site: http://mummer.sourceforge.net/
  – Version: 3.23
  – License: GPLv3

• RAPSearch2
  – Site: http://omics.informatics.indiana.edu/mg/RAPSearch2/
  – Version: 2.23
  – License: GPL

• minimap2
  – Site: https://github.com/lh3/minimap2
  – Version: 2.16
  – License: MIT

• diamond
  – Citation: Buchfink, Xie C., D. Huson (2015) Fast and sensitive protein alignment using DIAMOND, Nature Methods 12, 59-60
  – Site: https://github.com/bbuchfink/diamond
  – Version: v0.9.22.123
  – License: GPLv3

### 9.4 Taxonomy Classification

• Kraken2
  – Site: http://ccb.jhu.edu/software/kraken2/
  – Version: 2.0.7-beta
9.5 Phylogeny

- **FastTree**
  - Site: http://www.microbesonline.org/fasttree/
  - Version: 2.1.9
  - License: GPLv2

- **RAxML**
  - Citation: Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30:1312-1313
  - Site: http://sco.h-its.org/exelixis/web/software/raxml/index.html
  - Version: 8.0.26
  - License: GPLv2

- **Bio::Phylo**
  - Site: http://search.cpan.org/~rvosa/Bio-Phylo/
9.6 Specialty Genes

**PhaME**
- Citation: Sanaa Afroz Ahmed, Chien-Chi Lo, Po-E Li, Karen W Davenport, Patrick S.G. Chain. From raw reads to trees: Whole genome SNP phylogenetics across the tree of life. bioRxiv doi: http://dx.doi.org/10.1101/032250
- Site: https://github.com/LANL-Bioinformatics/PhaME/
- Version: 1.0
- License: GPLv3

9.7 Metagenome

**MaxBin2**
- Site: https://downloads.jbei.org/data/microbial_communities/MaxBin/MaxBin.html
- Version: 2.2.6
- License: BSD

**CheckM**
- Site: https://ecogenomics.github.io/CheckM/
- Version: 1.0.13
- License: GPLv3
9.8 Visualization and Graphic User Interface

- jsPhyloSVG
  - Site: http://www.jsphylosvg.com
  - Version: 1.55
  - License: GPL

- JBrowse
  - Citation: Skinner, M.E., et al. (2009) JBrowse: a next-generation genome browser, Genome research, 19, 1630-1638.
  - Site: http://jbrowse.org
  - Version: 1.11.6
  - License: Artistic License 2.0/LGPLv.1

- KronaTools
  - Citation: Ondov, B.D., Bergman, N.H. and Phillippy, A.M. (2011) Interactive metagenomic visualization in a Web browser, BMC bioinformatics, 12, 385.
  - Site: http://sourceforge.net/projects/krona/
  - Version: 2.7
  - License: BSD

- JQuery
  - Site: http://jquery.com/
  - Version: 1.10.2
  - License: MIT

- JQuery Mobile
  - Site: http://jquerymobile.com
  - Version: 1.4.3
  - License: CC0

- DataTables
  - Site: https://datatables.net
  - Version: 1.10.11
  - License: MIT

- jQuery File Tree
  - Site: http://www.abeautifulsite.net/jquery-file-tree/
  - Version: 1.01
  - License: GPL and MIT

- Raphael - JavaScript Vector Library
9.9 Utility

- Chromium * Citation: * Site: https://www.chromium.org * Version: 75.0.3767.0 * License: Google-authored portion is released under the BSD license.

- BEDTools
  - Citation: Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, 26, 841-842.
  - Site: https://github.com/arq5x/bedtools2
  - Version: 2.19.1
  - License: GPLv2

- Pilon

Site: https://github.com/broadinstitute/pilon

Version: 1.23

License: GPLv2 & MIT

**R**


Site: http://www.R-project.org/

Version: 3.5.1

License: GPLv2

**R Packages**

- Custom built directory containing all the packages required to install R packages offline
- The majority of the packages were downloaded automatically using the following R commands.

```r
# Function to get dependencies and imports for a given list of packages.

getPackages <- function(packs){
  packages <- unlist(
    tools::package_dependencies(packs, available.packages(), which=c(
      "Depends", "Imports"), recursive=TRUE)
  )
  packages <- union(packs, packages)
  packages
}

# Use the function by providing the names of the desired packages.

packages <- getPackages(c("packageName", "packageName2"))

# For example

packages <- getPackages(c("MetaComp", "gtable", "gridExtra", "devtools", 
  "phyloseq", "webshot", "plotly", "shiny", "DT", "ape", "igraph", "vegan", "BH 
  plogr", "dplyr", "ade4", "iterators"))

# Download packages to current/desired directory.

download.packages(packages, destdir="./", type="source")
```

- stringi defaults to downloading icudt55I.zip from online, the following method, from their documentation, was used to build a custom stringi package to avoid connecting to the internet:

1. File the `git clone https://github.com/gagolews/stringi.git` command.
2. Edit the `.Rbuildignore` file and get rid of the `^src/icu55/data` line.
3. Run `R CMD build stringi_dir_name`.

# index the downloaded packages into PACKAGES files.
```r
require(tools)
write_PACKAGES('.
')
```

- **MetaComp:** EDGE Taxonomy Assignments Visualization
  - Citation:
  - Site: https://cran.r-project.org/
  - Version: 1.0.2
  - License: BSD 3-Clause

- **GNU_parallel**
  - Site: http://www.gnu.org/software/parallel/
  - Version: 20190422
  - License: GPLv3

- **tabix**
  - Citation:
  - Site: http://sourceforge.net/projects/samtools/files/tabix/
  - Version: 0.2.6
  - License: MIT/Expat License

- **Primer3**
  - Site: http://primer3.sourceforge.net/
  - Version: 2.3.5
  - License: GPLv2

- **SAMtools**
  - Citation: Li, H., et al. (2009) The Sequence Alignment/Map format and SAMtools, Bioinformatics, 25, 2078-2079.
  - Site: http://www.htslib.org/
  - Version: 1.9
  - License: MIT

- **FaQCs**
  - Citation: Chienchi Lo, PatrickS.G. Chain (2014) Rapid evaluation and Quality Control of Next Generation Sequencing Data with FaQCs. BMC Bioinformatics. 2014 Nov 19:15
  - Site: https://github.com/LANL-Bioinformatics/FaQCs
  - Version: 2.08
  - License: GPLv3

- **Seqtk**
- Citation: Heng Li https://github.com/lh3/seqtk
- Site: https://github.com/lh3/seqtk
- Version: 1.3
- License: MIT

**NanoPlot**
- Site: https://github.com/wdecoster/NanoPlot
- Version: 1.13.0
- License: GPLv3

**Porechop**
- Citation:
- Site: https://github.com/rrwick/Porechop
- Version: 0.2.3
- License: GPLv3

**wigToBigWig**
- Citation: Kent, W.J., et al. (2010) BigWig and BigBed: enabling browsing of large distributed datasets, Bioinformatics, 26, 2204-2207.
- Site: https://genome.ucsc.edu/goldenPath/help/bigWig.html#Ex3
- Version: 4
- License: Free for academic, nonprofit, and personal use. A license is required for commercial usage.

**sratoolkit**
- Citation:
- Site: https://github.com/ncbi/sra-tools
- Version: 2.9.6
- License: Public Domain

**ea-utils**
- Citation: Erik Aronesty (2011) ea-utils : “Command-line tools for processing biological sequencing data”
- Site: https://code.google.com/archive/p/ea-utils/
- Version: 1.1.2-537
- License: MIT License

**Anaconda2 (Python 2)**
- Citation:
- Site: https://anaconda.org
- Version: 4.1.1
- License: 3-clause BSD

**ANAconda2 Packages**
Edge Documentation, Release Notes development

- Custom built directory containing all the required python2 packages for offline installation.
- This was generated primarily using the command:

```
  pip download packageName
```
- Some packages were manually downloaded into the directory to install via conda
- Dependencies were manually downloaded as they were discovered

- Anaconda3 (Python 3)
  - Citation:
  - Site: https://anaconda.org
  - Version: 5.1.0
  - License: 3-clause BSD

- Anaconda3 Packages
  - Custom built directory containing all the required python3 packages for offline installation.
  - This was generated primarily using the command:

```
  pip download packageName
```

9.10 Amplicon Analysis

- QIIME2
  - Citation: Caporaso et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335-6
  - Site: http://qiime2.org/
  - Version: 2019.1
  - License: BSD 3-Clause

- DETEQT: Diagnostic targeted sequencing adjudication
  - Citation: Conrad TA et al. (2019) Diagnostic targetEd seQuencing adjudicaTion (DETEQT): Algorithms for Adjudicating Targeted Infectious Disease Next-Generation Sequencing Panels.
  - Site: https://github.com/LANL-Bioinformatics/DETEQT
  - Version: 0.3.0
  - License: GPLv3

9.11 RNA-Seq Analysis

- PyPiReT: Pipeline for Reference based Transcriptomics.
  - Citation:
  - Site: https://github.com/mshakya/PyPiReT
  - Version: 0.3.2
10.1 FAQs

• Can I speed up the process?
  You may increase the number of CPUs to be used from the “additional options” of the input section. The default and minimum value is one-eighth of total number of server CPUs.

• There is no enough disk space for storing projects data. How do I do?
  There is an archive project action which will move the whole project directory to the directory path configured in the $EDGE_HOME/sys.properties. We also recommend a symbolic link for the $EDGE_HOME/edge_ui/EDGE_input/public/ directory which points to the location where the users’ (or sequencing centers’) raw data are stored, obviating unnecessary data transfer via web protocol and saving local storage.

• How to decide various QC parameters?
  The default parameters should be sufficient for most cases. However, if you have very depth coverage of the sequencing data, you may increase the trim quality level and average quality cutoff to only use high quality data.

• How to set K-mer size for IDBA_UD assembly?
  By default, it starts from kmer=31 and iterative step by adding 20 to maximum kmer=121. Larger K-mers would have higher rate of uniqueness in the genome and would make the graph simpler, but it requires deep sequencing depth and longer read length to guarantee the overlap at any genomic location and it is much more sensitive to sequencing errors and heterozygosity. Professor Titus Brown has a good blog on general k-mer size discussion.

• How many reference genomes for Reference-Based Analysis and Phylogenetic Analysis can be used from the EDGE GUI?
  The default maximum is 20 and there is a minimum 3 genomes criteria for the Phylogenetic Analysis. But it can be configured when installing EDGE.

• Which aligner should I choose?
We use default setting of the aligner. Bowtie2 default is for global alignment and BWA mem algorithm will do local alignment. If users would like to overwrite the setting, users can use “Aligner Options” to do so. For example, use “--local” to run bowtie2 with local alignment mode. Or, use “-x ont2d” to run BWA mem with Nanopore reads.

- How to make an app icon on the mobile device?

Launch the Safari browser on Apple’s iOS and navigate to the https://bioedge.lanl.gov/edge_ui/ or your EDGE instance website. (Please refresh the page few times to update the cache) Tap the Share button on the browser’s toolbar — that’s the rectangle with an arrow pointing upward. It’s on the bar at the top of the screen on an iPad, and on the bar at the bottom of the screen on an iPhone or iPod Touch. Tap the Add to Home Screen icon in the Share menu.

Launch Chrome for Android and open the https://bioedge.lanl.gov/edge_ui/ or your EDGE instance website. (Please refresh the page few times to update the cache) Tap the menu button and tap Add to homescreen. You’ll be able to enter a name for the shortcut and then Chrome will add it to your home screen. Alternatively, we have bioedge Web App as APK file to download and install in your android device too. You can download by scan the QR code below.

- Why a job is queued and never autostarted?

The queued job had too much CPUs request. The autorun feature will start running queued job when there is available CPU resource. The queued job CPUs usage plus running jobs CPUs usage should be less than (<) edgeui_tol_cpu configured in the $EDGE_HOME/edge_ui/sys.properties.

- Why some of the taxonomy profiling result are N.A.?

Please check the log file to give us more information. For above example on BWA result, at web
UI, you can open log file by Clicking the link next to “Output Directory” at “General” section ->
ReadsBasedAnalysis -> Taxonomy -> log -> allReads-bwa.log.

In this case, it is out of memory. EDGE requires at least 16G memory. see System requirements
For machine with < 32Gb memory, we suggest to use the smaller BWA index (~13Gb) and contains
the databases for bwa taxonomic identification pipeline

```
wget -c https://edge-dl.lanl.gov/EDGE/dev/edge_dev_bwa_mini_index.tgz
```

10.2 Troubleshooting

- Process.log and error.log files may help on the troubleshooting.

10.2.1 EDGE WEB GUI

- In the GUI, if you are trying to enter information into a specific field and it is grayed out or won’t let you, try
refreshing the page by clicking the icon in the right top of the browser window.
- After installation, I can login but cannot select any files for input. the selection pop-up is empty.
This could be the permission issue on the EDGE_input/EDGE_output directory for Apache user. Please see Apache Web Server Configuration

- I can not login to EDGE, it keeps saying Session expired.

The login session will expire in 12 hours. If you keep get session expired message. It may indicate the '/' (root) space is full. Please try to clean up log files or others you/admins can delete. For example, /var/log/message-2016xxxx is the archived log rotations which can be deleted for the space.

### 10.2.2 Coverage Issues

- Average Fold Coverage reported in the HTML output and by the output tables generated in {output directory}/AssemblyBasedAnalysis/ReadsMappingToContigs/ are calculated with mpileup using the default options for metagenomes. These settings discount reads that are unpaired within a contig or with an insert size out of the expected bounds. This will result in an underreporting of the average fold coverage based on the generated BAM file, but one that the team feels is more accurate given the intended use of this environment.

### 10.2.3 Data Migration

- The preferred method of transferring data to the EDGE appliance is via SFTP. Using an SFTP client such as FileZilla, connect to port 22 using your system’s username and password.
- In the case of very large transfers, you may wish to use a USB hard drive or thumb drive.
- If the data is being transferred from another LINUX machine, the server will recognize partitions that use the FAT, ext2, ext3, or ext4 filesystems.
- **If the data is being transferred from a Windows machine, the partition may use the NTFS filesystem. If this is the case, then**
  
  - Open the command line interface by clicking the Applications menu in the top left corner (or use SSH to connect to the system).
  - Enter the command: ‘`sudo yum install ntfs-3g ntfs-3g-devel -y`’
Enter your password if required.

- After a reboot, you should be able to connect your Windows hard drive to the system, and it will mount like a normal disk.

10.2.4 Known Issues

- Installations on CentOS 6.4 with Apache 2.2 are known to have difficulty executing jobs that have “.real” anywhere in the name. This is due to a CGI execution issue. The recommended resolution is to use an underscore in place of the period, or simply name your job something else.

10.3 Discussions / Bugs Reporting

- We welcome questions, feedback and bug reports that may help us improve the EDGE platform. Ideally, any bug report should include the process.log, the error.log and the failed module log files. If it is a system error, the tomcat, apache and mysql logs are helpful. If possible, please share a subset of the input dataset/files for us to re-create the bug. Here is a bug report template for your reference:

**Describe the bug**
A clear and concise description of what the bug is.

**To Reproduce**
Steps to reproduce the behavior:

Go to '...
Click on '....'
Scroll down to '....'
See error

Expected behavior
A clear and concise description of what you expected to happen.

**Screenshots**
If applicable, add screenshots to help explain your problem.

**Desktop (please complete the following information if applicable):**
OS: [e.g. iOS]
Browser [e.g. chrome, safari]
Browser Version [e.g. 22]
EDGE Version [e.g. 2.3.1]

**Additional context**
If you ran a project, can you provide process.log, error.log and the failed module log files. If it is system error, the tomcat, apache and mysql logs are helpful.

- We have created a mailing list for EDGE users. If you would like to receive notifications about the updates and join the discussion, please join the mailing list by becoming the member of edge-users groups.

  EDGE user's google group

- We appreciate any feedback or concerns you may have about EDGE. If you encounter any bugs, you can report them to our GitHub issue tracker.

  Github issue tracker
• Any other questions? You are welcome to Contact Us and Citation (page 97)
Copyright (2018). Triad National Security, LLC. All rights reserved.

This program was produced under U.S. Government contract 89233218CNA000001 for Los Alamos National Laboratory (LANL), which is operated by Triad National Security, LLC for the U.S. Department of Energy/National Nuclear Security Administration.

All rights in the program are reserved by Triad National Security, LLC, and the U.S. Department of Energy/National Nuclear Security Administration. The Government is granted for itself and others acting on its behalf a nonexclusive, paid-up, irrevocable worldwide license in this material to reproduce, prepare derivative works, distribute copies to the public, perform publicly and display publicly, and to permit others to do so.

This is open source software; you can redistribute it and/or modify it under the terms of the GPLv3 License. If software is modified to produce derivative works, such modified software should be clearly marked, so as not to confuse it with the version available from LANL. Full text of the GPLv3 License can be found in the License file in the main development branch of the repository.
Contact Us and Citation

Questions? Concerns? Please feel free to email our google group at edge-users@googlegroups.com or contact a dev team member listed below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patrick Chain</td>
<td><a href="mailto:pchain@lanl.gov">pchain@lanl.gov</a></td>
</tr>
<tr>
<td>Chien-Chi Lo</td>
<td><a href="mailto:chienchi@lanl.gov">chienchi@lanl.gov</a></td>
</tr>
<tr>
<td>Paul Li</td>
<td><a href="mailto:po-e@lanl.gov">po-e@lanl.gov</a></td>
</tr>
<tr>
<td>Karen Davenport</td>
<td><a href="mailto:kwdavenport@lanl.gov">kwdavenport@lanl.gov</a></td>
</tr>
<tr>
<td>Logan Voegtly</td>
<td><a href="mailto:logan.j.voegtly.ctr@mail.mil">logan.j.voegtly.ctr@mail.mil</a></td>
</tr>
<tr>
<td>Kim Bishop-Lilly</td>
<td><a href="mailto:kimberly.a.bishop-lilly.ctr@mail.mil">kimberly.a.bishop-lilly.ctr@mail.mil</a></td>
</tr>
</tbody>
</table>

12.1 Citation

Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform
Po-E Li; Chien-Chi Lo; Joseph J. Anderson; Karen W. Davenport; Kimberly A. Bishop-Lilly; Yan Xu; Sanaa Ahmed; Shihai Feng; Vishwesh P. Mokashi; Patrick S.G. Chain
Nucleic Acids Research 2016;
doi: 10.1093/nar/gkw1027