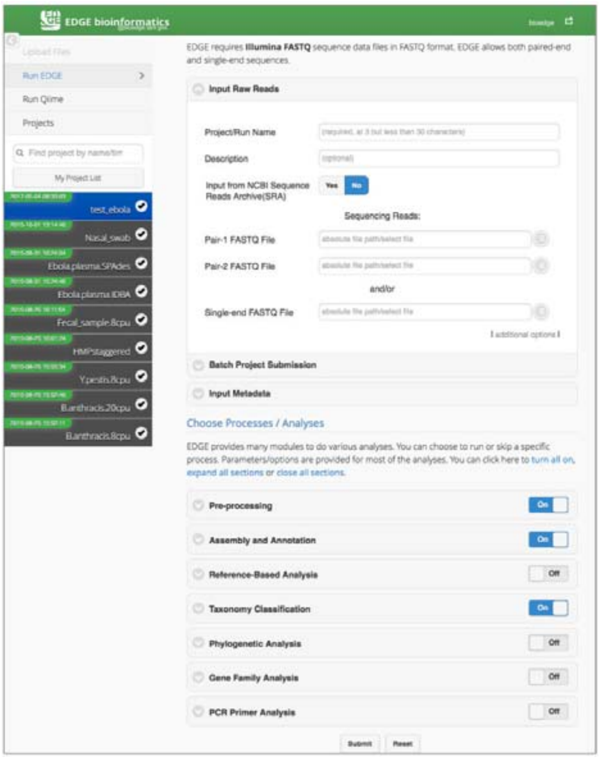
Atlas of Science
another view on science
http://atlasofscience.org

Bioinformatics at the cutting EDGE: Empowering the Development of Genomics Expertise

An in-depth understanding of how microorganisms function is required to address many challenging issues facing society today such as understanding how to eradicate infectious diseases and mitigate climate change. Answers are held in their genetic code. Advancements in next-generation sequencing (NGS) technologies over the past decade promise a better understanding of the living world by accessing their DNA. There are however, significant bottlenecks when it comes to analyzing and interpreting the sequencing data, e.g. computational resources and expertise required to handle the enormous volumes of sequencing data. To help address these problems, scientists at Los Alamos National Laboratory and the Naval Medical Research Center have developed EDGE Bioinformatics (Empowering the Development of Genomics Expertise), which aims to lower the barrier to better understand NGS data by making complex algorithms available via an intuitive web-based interface.



EDGE requires **Illumina FASTQ** sequence data files in FASTQ format. EDGE allows both paired-end and single-end sequences.

Input Raw Reads

Project/Run Name: (required, at 3 but less than 30 characters)

Description: (optional)

Input from NCBI Sequence Reads Archive(SRA): Yes [✓ No]

Sequencing Reads:

Pair-1 FASTQ File: absolute file path/select file

Pair-2 FASTQ File: absolute file path/select file

and/or

Single-end FASTQ File: absolute file path/select file

[additional options]

Batch Project Submission

Input Metadata

Choose Processes / Analyses

EDGE provides many modules to do various analyses. You can choose to run or skip a specific process. Parameters/options are provided for most of the analyses. You can click here to turn all on, expand all sections or close all sections.

Pre-processing: On

Assembly and Annotation: On

Reference-Based Analysis: Off

Taxonomy Classification: On

Phylogenetic Analysis: Off

Gene Family Analysis: Off

PCR Primer Analysis: Off

Fig. 1. A screenshot of EDGE Bioinformatics and all available modules. Different workflows and available options of EDGE bioinformatics are shown here as collapsible tabs that users can expand and turn ON or OFF as per the need of analysis. The menu on the top left shows options to either run EDGE pipelines or QIIME through the webpage. My project List shows the list of projects within individual boxes with the project name, date, and time it was started. A green date/time stamp and a check mark indicate that the project has finished running. The blue colored box indicates the selected project’s results page is open. The workflows are briefly described in the main part of the paper.

EDGE Bioinformatics combines multiple tools and algorithms into standardized workflows. It empowers users without computational expertise to analyze complicated NGS datasets through simple point and click. A user only needs to input raw reads (directly from a sequencer) and select one or more of the available custom-designed workflows in order to achieve robust and reproducible analysis. The current version of EDGE includes seven workflows that accommodate a variety of use cases (Fig. 1). The workflows include (i) *Pre-processing*, used to remove low quality or undesired data (e.g. host sequences or known contamination), (ii) *de novo assembly and annotation*, used to reconstruct genome(s) from raw reads, (iii) *Reference-Based Analysis,* used to compare reads/contigs to a known genome, (iv) *Taxonomy Classification* provides a taxonomic profile of a sample based on assigning reads/contigs to known organism, (v) *Phylogenetic Analysis* places the sample within phylogenetic context of sequenced members of the same species. (vi*) Gene Family Analysis* searches for virulent and antibiotic resistant genes, and (vii*) PCR Primer Analysis* checks the validity of primer pairs or designs them *de novo*. Additional capabilities recently added to EDGE include incorporation of sample metadata (e.g., date, geo-location, patient symptoms, etc.) and the ability to compare the taxonomic profiles among multiple samples.

EDGE provides a variety of static and interactive outputs presented as PDF reports, data tables and publication quality figures, so that the user can delve further into the results (Fig. 2). These outputs are provided in real-time, so users can explore their results within the user-friendly, intuitive web-based environment as the sample is being processed. All results for samples (projects) can also be shared with collaborators or made public.

While EDGE was 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, therefore some expertise is still required to derive adequate conclusions from the results. In addition, while the tools used in EDGE are optimized to run in parallel environment and can reduce the analysis time from days to minutes, many computational steps yet require significant computational resources. EDGE Bioinformatics is an ongoing effort to provide best of breed bioinformatics tools for NGS data analysis, thus updates to modules, tools and overall functionality, are continuously under development.

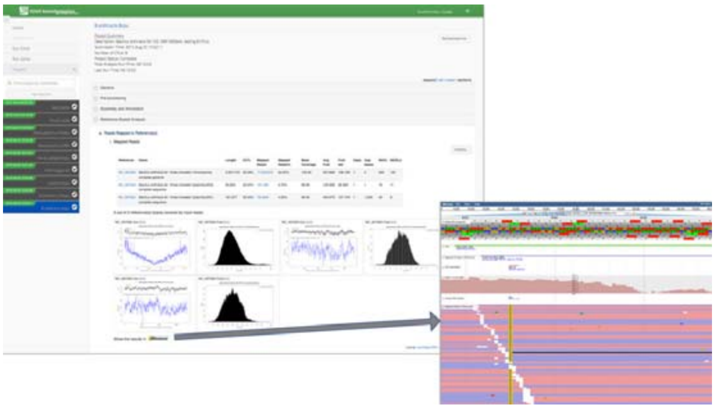


Fig. 2. An example results page from EDGE analysis. The top part of the results page shows summaries of the run that includes sample description, submission time, and number of CPUs used, status of the project, and run

time. Results from the different workflows are displayed within collapsible tabs. Links are provided for more in-depth results, as shown here from Reference Based Analysis. For example, the inset figure shows an instance of JBrowse results.

Additional information on EDGE bioinformatics workflows and the computational environment requirements can be found at https://edge.readthedocs.io/ and a video tutorial series on running each EDGE module can be found at http://tutorial.getedge.org. The software is freely available to install from https://lanl-bioinformatics.github.io/EDGE/ and a demonstration webserver at https://bioedge.lanl.gov/ is available to analyze publicly available data. Users with private data can contact the authors for access to private servers or can register at http://hobo-nickel.getedge.org to upload and process their data.

Migun Shakya [3], ***Chien-Chi Lo*** [3], ***Karen Davenport*** [3], ***Logan Voegtly*** [2, 4], ***Po-E Li*** [3], ***Yan Xu*** [3],
Casandra Philipson [1, 2], ***Regina Z. Cer*** [2, 4], ***Kimberly A. Bishop-Lilly*** [1],
Theron Hamilton [1], ***Patrick S. G. Chain*** [3]
[1]*Genomics and Bioinformatics Department, Biological Defense Research Directorate, Naval Medical Research Center-Frederick, 8400 Research Plaza, Fort Detrick, MD, USA*
[2]*Defense Threat Reduction Agency, Fort Belvoir, VA, USA*
[3]*Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA*
[4]*Leidos, 11955 Freedom Drive, Reston VA, USA*

Publication

Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform.
Li PE, Lo CC, Anderson JJ, Davenport KW, Bishop-Lilly KA, Xu Y, Ahmed S, Feng S, Mokashi VP, Chain PS
Nucleic Acids Res. 2017 Jan 9