LA-UR-17-23029
Approved for public release; distribution is unlimited.

Title: EDGE 2017 R&D 100 Entry with Appendix

Author(s): Chain, Patrick Sam Guy; Davenport, Karen Walston; Li, Po-E; Lo, Chien-Chi; Xu, Yan; Senin, Pavel; Shakya, Migun; Ahmed, Sanaa Afroz; Hamilton, Theron C.; Bishop-Lilly, Kimberly A.; Anderson, Joseph J.; Voegtly, Logan J.; Philipson, Casandra W.

Intended for: R&D 100 Award Application

Issued: 2017-04-13

Joint 2017
R&D 100 entry

Los Alamos National Laboratory and the Naval Medical Research Center

BIOINFORMATICS

EDGE

Empowering the Development of Genomics Expertise

Making Genomics Accessible to Everyone

- Provides an intuitive, web-based interface so that even scientists with *no* bioinformatics experience can easily process genomics data
- Creates for the first time a detailed panoramic view of samples from various analytical standpoints
- Runs a variety of bioinformatics analyses for viral, bacterial, archaeal, and complex metagenomic samples
- Enables users to visualize and interact with selected results

Los Alamos NATIONAL LABORATORY EST. 1943

NAVAL MEDICAL RESEARCH CENTER U.S. NAVY

2017 R&D 100 Award Entry Form

Title

EDGE Bioinformatics: Empowering the Development of Genomics Expertise
Making genomics accessible to everyone

LA-UR-17-

Categories

Analytical Instrumentation
Life Sciences
Software
Special Recognition: Market Disruptor—Products

Product/Service Brand Name

EDGE Bioinformatics

Name of Primary Submitting Organization

Los Alamos National Laboratory

Name of Co-developing Organization

Naval Medical Research Center

Was the product/service introduced to the market between January 1, 2016, and March 31, 2017?

Yes ☑ No □

If your submission is subject to regulatory approval: Has the product been approved?

Yes □ No □ Not applicable to this product ☑

Product Price (U.S. Dollars)

EDGE bioinformatics is free, open-source software.

Product Description

EDGE Bioinformatics "democratizes" the genomics revolution by enabling *any* biologist (researcher or physician) to quickly and easily analyze complex genomics data.

Indicate the type of institution you represent

Government Lab

Submitter's relation to entered product/service

Product Developer

Product Photos

EDGE 2016 R&D 100 Cover
EDGE Flowchart
EDGE Project Page
Genome Browser View in EDGE

Video Links or Files—List File Names of up to Three Videos

EDGE Bioinformatics: An Interview with Patrick Chain
EDGE Tutorial Series:
https://www.youtube.com/playlist?list=PL7DNo6h5wJsTh2l2GK3N86Imb-9fYQFfH

What does the product or technology do? Describe the principal applications of this product.

Diabetes, infertility, cancer, and Alzheimer's disease—the key to one day preventing or even curing such afflictions and diseases (both infectious and genetically driven) may be locked in our own genetic code and the code of microorganisms that inhabit our bodies. The study of this code, known as genomics, has recently become much more promising as a result of two things: (1) vast improvements in high-throughput, next-generation sequencing (NSG), and (2) an exponential decrease in the cost of such sequencing. For example, it originally cost approximately $3 billion to sequence the human genome; today, this genome could be resequenced for less than $1,000.

Given the rise in throughput and the decline in sequencing costs, there remain two key problems that serve as a significant bottleneck when it comes to data analysis: (1) the sheer volume of data available (known in the computing industry as Big Data) and (2) the

paucity of bioinformatics expertise necessary to understand and act on these new types of Big Data.

To help address both of these problems, scientists at Los Alamos National Laboratory and the Naval Medical Research Center have come together to create EDGE (short for Empowering the Development of Genomics Expertise) Bioinformatics. An intuitive, web-based platform, EDGE Bioinformatics consists of a broad variety of fully integrated and innovative bioinformatic software and algorithms, incorporated into a user-friendly, web-based system with preconfigured workflows. All these workflows can be applied quickly and easily—just point and click—to wide variety of genome-sequencing projects ranging from individual isolates (from a culture of a single organism) to much more complex metagenomics (microbiome) projects.

EDGE Bioinformatics addresses both the problem of handling Big Data, and it does so without users having to possess bioinformatics expertise.

Handling Big Data. The human genome alone is approximately 6×10^9 base pairs—that's 6 billion base pairs of human DNA in a single human cell. Although bacterial cells and viruses have orders of magnitude less DNA than human cells, microbes live in communities. For example, it has been estimated that the human body harbors at least as many bacterial cells as human cells. Targeting a microbial community or a microbiome—such as in a clinical sample—may reveal thousands or millions of different bacterial and viral species, represented by many millions of cells, all of them residing in a mixture with human cells. Imagine attempting to characterize the volume of data from such a sample. In fact, taking into account the diversity of microbes inhabiting the human body, researchers have estimated that the gene content of the typical microbiome is over 300 times larger than the human genome.

Typically, a user with raw sequence data from this type of mixed sample would hire experts to develop cryptic command-line tools or write computer code to process such Big Data. With EDGE Bioinformatics, users are provided several preconfigured workflows with default parameters for many of the standard analyses typically needed for genomic data. These workflows use raw data from sequencers as input and create reports and graphics based on the data, providing integrated and interactive views for the user who in turn can delve further into the data and results.

EDGE makes performing Big Data analyses enormously easier; it is capable of aligning millions of sequence reads to databases of thousands of genomes and identifying which organisms are present, and/or listing differences found between the genome(s) in the sample and reference genomes.

"Democratizing" the genomics revolution. Perhaps more problematic than handling Big Data to achieve results is the dearth of expertise in bioinformatics. To manage and interpret data, most bioinformaticians rely on developing a custom code of command-line tools. Realistically, very few biologists have the computational expertise or resources to accomplish such sophisticated analyses. Instead, they must rely on experts for analysis, often sending the data elsewhere, which contributes to the cost and loss of valuable time. This "black-box" approach can also defeat reproducibility and lends itself to a lack of analytic standardization. Even with such a solution, because of the enhanced pace of new and improved technology development coupled with breadth of genomic applications, it is even difficult for seasoned genomic veterans to keep pace with the variety of tools and algorithms optimally combined to address specific questions. As genomics becomes more commonplace and applied to everyday scenarios, such as analyzing a sample from a patient with an unknown infectious disease, any delay or error in analysis could significantly affect the outcome.

EDGE Bioinformatics addresses this key problem by making its software readily available and easy to use.

Availability. It is expected that most hospitals and labs around the world will be have sequencing technology in just a few years. Some hospitals already perform routine sequencing of clinical samples. EDGE Bioinformatics makes it possible for nearly *any* biologist/physician with access to genomic data to use this technology. As the cost of sequencing continues to drop, the availability of sequencers to a variety of researchers continues to rise. However, conventional bioinformatic tools required to process and analyze genomic data are not as readily accessible or easy to use. EDGE is open source and a demonstration server is available for broad public use.

Easy to use. Running EDGE Bioinformatics is far from complex, typically requiring only a few mouse clicks to launch jobs and achieve results. The tools in EDGE were selected to achieve robust and accurate analysis together with rapid computational

processing. Using EDGE, most analyses take minutes or hours rather than days or weeks. EDGE Bioinformatics does the bulk of the work, so that users with little or no bioinformatics expertise can readily view, analyze, and understand the results of genomic data.

EDGE Bioinformatics has already helped streamline data analysis for groups in the United States, as well as a number of countries outside the United States, such as Australia, Cambodia, Canada, Egypt, Gabon, Kenya, Peru, Republic of Georgia, Republic of Korea, Thailand, Uganda, and the United Kingdom. Because the software environment is open source, there is no license or payment required to download and use it locally or remotely. Computational requirements are dependent on the complexity and size of dataset and the particular workflows selected. Servers with at least 256 gigabytes of memory and 24 CPUs are recommended for full operability on a wide range of samples and applications, although it is possible to run the software suite with a minimum of 16 gigabytes with 1 CPU.

EDGE Bioinformatics stands to revolutionize the way individuals can analyze genomic data by making sophisticated tools available via an intuitive and easy-to-use web-based interface. No longer will such analysis require extensive training in computer science—all it takes is a few simple clicks on a computer and the tools perform the analysis for you. Thus, EDGE Bioinformatics make the following possible:

Allows almost anyone to conduct sophisticated genomic analysis. With EDGE implemented on a local computer server or in the cloud, it is now possible to bring the power of complex, big-data NGS analysis to smaller research laboratories, including clinics, hospitals, and university laboratories. As the applications of sequencing grow from looking only at the human genome to looking at the organisms that reside in and on humans, so too does the user-base for EDGE Bioinformatics, which already addresses both types of analyses. For example, scientists are using genomics to assess human genome mutations contributing to cancer or to analyze human microbiome shifts associated with Crohn's disease, irritable bowel syndrome, allergies, and even Alzheimer's disease.

Brings genomics to everyday use. Armed with NGS and EDGE bioinformatics, it will be possible in the future for a nurse to swab your saliva and later tell you if your

symptoms are more likely caused by a viral or bacterial infection. Such an analysis would be a tremendous advancement because if it is a viral infection, you avoid unnecessary "just-in-case" antibiotics; taking such unnecessary medicine contributes to the ongoing antibiotic resistance crisis, which makes bacterial infections much more difficult for modern medicine to treat. Similarly, it is envisioned that hospital staff could routinely use EDGE to ascertain which pathogens inhabit the hospital environment and thus implement countermeasures or inform infection-control procedures.

Applies to more than health and medicine. Outside the realm of medicine, EDGE has been used to understand differences among algal strains, an understanding that can help in the development and engineering of algae to produce more oil for biofuel production. EDGE has also helped identify differences in microbial communities residing in soil and water, which can help better understand the role of microorganisms in fixing carbon from the atmosphere and how such populations adapt to changes in climate (the fluxes of both heat and moisture).

Los Alamos National Laboratory has created a website with a demonstration version of EDGE (v1.1), which is located at https://bioedge.lanl.gov. This demo version can process data from public repositories, such as the National Center for Biotechnology Information Sequence Read Archive. The Naval Medical Research Center has also made a full version of EDGE available (v1.5 at http://hobo-nickel.getedge.org/) that allows users to upload their own data and run EDGE.

The following letters of support, included in this entry, address the benefits and applications of EDGE Bioinformatics: U.S. Department of Defense: Defense Threat Reduction Agency, U.S. Army Medical Research Institute of Infectious Diseases, Department of Health & Human Services, Armed Forces Research Institute of Medical Sciences, U.S. Naval Medical Research Unit #6, National Center for Disease Control and Public Health (Republic of Georgia), Agency for Defense Development (Republic of Korea), and Viome, Inc.

Testimonials from EDGE users titled "Comments from Current EDGE Bioinformatics Users" and a "List of EDGE Bioinformatics Users" are found in the Appendix.

The following articles in the Appendix provide information about EDGE Bioinformatics and its applications: "Science on the Hill: Bringing the power of genetic research to an office near you," "How bioinformatics tools are bringing genetic analysis to the masses," "LANL's EDGE Offers Easy-to-Use Bioinformatics Pipelines for Microbial Sequence Analysis," and "Edge bioinformatics brings genomics to everyone."

How does the product operate? Describe the materials, composition, construction, theories, or mechanism of action.

EDGE Bioinformatics was designed to function as a highly-integrated web-based platform that runs many of the standard analyses biologists use on viral, bacterial, archaeal, and metagenomic samples. Many of these integrated tools are packaged as pre-constructed modules (workflows) that are easily activated by point-and-click. Default parameters can be changed, and all modules selected can start with a simple click on "Submit." At this point, the user can track progress and view any completed tasks on the platform's project base. EDGE Bioinformatics consists of the following key modules:

- **Preprocessing**: Low-quality DNA and inherent limitations of sequencing technologies can produce less-than-optimal data. To optimize such data, the FaQCs tool was developed to assess the genomic data and trims and/or filters data of low quality. EDGE Bioinformatics also provides tools that remove non-informative genomic data, such as host data or generic adaptors. Removing such data streamlines the analysis of the microorganism or the microbial community in the sample.

- **Assembly and Annotation**: Discontinuous sequencing data are typical because DNA must be fragmented into smaller pieces before sequencing. EDGE Bioinformatics provides assembly tools that help reassemble genomes and accommodate a variety of sample types and read lengths. Assembled data provide more complete information about the genes in a sample. When a user performs an assembly, all subsequent analyses are done with the sequencing reads, the assembled data, or both (default). Annotation identifies genes from the assembly and provides their functional information and location.

- **Reference-Based Analysis**: Users can compare a sample with one or more known reference genomes from public databases or from their own research. This may be performed to understand the differences between organisms or to characterize known organisms from within complex samples. Results include lists of missing regions (gaps), new inserted sequences or sequence changes (single nucleotide polymorphisms, or SNPs), and coverage graphs showing this

information. A genome viewer provides interactive access to all the output, as well as the ability to zoom in or out of the data compared to the reference(s).

- **Taxonomy Classification**: All organisms are classified into categories and subcategories known as taxonomic levels (kingdom, phylum, order, class, family, genus, species); for example, the genus and species name for humans is *Homo sapiens*. To determine the identity of the organism(s) that have been sequenced, EDGE provides a number of classification tools (some developed at Los Alamos National Laboratory) that conduct analyses on individual reads or on the assembled data. The results are provided as heatmaps (to show relative abundance), interactive radar plots (or Krona plots), and as tables and graphs showing the organisms found at each taxonomy level.

- **Phylogenetic Analysis**: All taxonomies are based on evolutionary relatedness. To reconstruct the evolutionary history, or "phylogenetic," tree, EDGE Bioinformatics includes software to help build the phylogeny of a target organism using either sequencing reads, the assembled data, or both, regardless of whether the sequenced sample represents a single organism or one that is mixed with other organisms (i.e., a metagenome).

- **Polymerase Chain Reaction (PCR) Analysis**: PCR is often used as a proxy for more complete (and more expensive) sequencing data. EDGE Bioinformatics provides two utilities with respect to PCR. The first enables users to interpret the expected results of a known PCR assay (such as those used to identify pathogens in clinical samples), whereas the second enables users to design a novel PCR assay that would uniquely identify the target organism based on unique sequences found in the assembled genome.

Other modules that run on EDGE Bioinformatics version 1.5 offer the following capabilities:

- **Identifies Special Genes of Interest**: EDGE Bioinformatics can help identify specific marker genes that correspond to virulence or to resistance to antibiotics. Given the prevalence of antibiotic resistance and the need to predict potential therapeutics in the case of infections, EDGE now provides users with the ability to determine if specific virulence genes or antibiotic resistance genes are present in a sample. Output consists of a list of the specific genes and their determined abundance.

- **Performs Community Profiling of Archaea, Bacteria, and Fungi**: Because microbial communities can be highly diverse (many thousands of species in a single gram of soil or liter of lake or ocean water), and because all free-living organisms share ribosomal ribonucleic acid (rRNA), genes that encode the machinery that makes proteins from genes, researchers developed a shortcut to "profile" an entire community simply by looking at fragments of this one gene

region (16S, 18S, or ITS). EDGE Bioinformatics provides tools that help cluster, sort, and identify the taxonomies of all organisms within the sample from sequencing targeting these regions. The output consists of a list of identified taxa and their representative abundance.

- **Provides Metadata Collection/Storage**: Analysis of sequencing data by itself can provide information on the organisms and genes (functions) that the genomes encode. For some studies, the context of when, where, and how the sample was taken (i.e., metadata) is just as critical. For example, epidemiologists associate an organism's genotype with the time and location of the sampling to help with their investigations. For this reason, EDGE provides the ability to collect and report metadata on a per-sample basis to enable users to track additional information related to each sample.

- **Compares Metagenome Information from Two or More Samples**: Although EDGE already provides the capability to compare samples with one or more reference genomes, there is often a need to compare among samples. EDGE Bioinformatics begins to provide such functionality by focusing on the ability to compare the taxonomic identities and abundances of the organisms found within one sample, with those from other samples. This information is provided to users in the form of heatmaps.

This collective of modules is just the tip of the iceberg when it comes to EDGE's functionality. Already the EDGE Bioinformatics Team is working on developing new modules. Examples of these modules are discussed in "EDGE 2.0 Modules Under Development" in the Appendix.

EDGE Bioinformatics' web-based interface produces easy-to-understand visuals. EDGE helps users to interact with the results by using its project page. EDGE also provides a final, detailed report in PDF format that encompasses all the graphical results. EDGE users can download any result and intermediate file(s) in a variety of formats, such as tables, text files, various graphic files, and PDFs. EDGE's user-management system enables users to track all their own EDGE analyses; this system also enables users to share results, post publicly, and even delete or archive any of their projects.

Additional information on EDGE, the bioinformatics workflows, the open-source tools (developed by Los Alamos National Laboratory or third parties), and the computational environment requirements can be found at the following URL: https://edge.readthedocs.io/en/v1.1/. This website provides a comprehensive explanation

of EDGE, including system requirements, installation, using the graphic user interface, the command line interface, databases, third-party tools, and FAQs and troubleshooting.

Further details and examples of uses of EDGE Bioinformatics can be found in the "EDGE Fact Sheet" and the technical manuscript "Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform" in the Appendix.

EDGE BIOINFORMATICS COMPARISON MATRIX

Parameter	EDGE	Galaxy	CLC Genomics	Web Services (such as IMG or MG-RAST)
Removes host sequences	Yes	Difficult	Difficult	Some have this capability
Comments: EDGE is the only technology that can select and remove host sequences natively. It is possible to do so with the Galaxy and CLC Genomics platforms, but users must build a workflow and indexes to perform this function successfully. MG-RAST (Web Services) has limited functionality, selecting only some reference genomes.				
Compares reads and contigs to references	Yes	Difficult	Difficult	No
Comments: Galaxy and CLC Genomics platforms may have tools available that natively allow reads compared with references, as well as tools capable of aligning contigs to references. However, contig alignment tools may be suboptimal and computationally expensive. Moreover, the data from read- and contig-based alignments are not integrated. Web Services do not allow comparisons to individual genomes. EDGE is natively capable of performing both read and contig comparisons to references. A contig is a set of overlapping DNA segments that together make up a consensus region of DNA.				
Provides functional annotation	Yes	Difficult	No	Yes
Comments: The Galaxy platform can provide this feature. However, users must find and install the appropriate annotation software. Such software often consists of many tools and has many additional dependencies, and the user must build a workflow from there.				
Provides taxonomy profiling of contigs	Yes	No	No	Yes
Comments: It is possible for the Galaxy and CLC Genomics platforms to do such profiling, but users must develop an algorithm for contig classification and then integrate it within the platform.				
Provides taxonomy profiling with reads	Yes	Yes	No	Yes
Comments: EDGE uses multiple tools, whereas Galaxy natively uses only one tool and thus cannot compare among tools. Web Services use gene annotations, which will not necessarily reflect all the data.				
Performs phylogenetic reconstruction from contigs or genomes	Yes	No	No	No
Comments: It is possible for the Galaxy and CLC Genomics platforms to do such reconstruction, but users would be required to link alignment of long sequences with a program for phylogenetic inference.				

EDGE BIOINFORMATICS COMPARISON MATRIX (continued)

Parameter	EDGE	Galaxy	CLC Genomics	Web Services (such as IMG or MG-RAST)

Performs phylogenetic reconstruction from reads	Yes	No	Yes	No
Comments: EDGE is based on SNPs (single nucleotide polymorphisms), whereas CLC Genomics is based on Kmer. SNPs have the advantage of better reflecting biological evolution, whereas Kmers are more of an approximation. It is possible for Galaxy to perform this function, but users must develop algorithm to identify SNPs or Kmers and link these with a program for alignment or phylogenetic inference.				
Performs PCR amplicon analysis	Yes	Primers only	Primers only	No
Comments: With Galaxy and CLC Genomics, users must develop an algorithm of find other tools to identify where primers align to a genome, report presences vs. absence, and other issues. An amplicon is an amplified segment of specific DNA or RNA sequences within which multiple copes of nucleic acid sequences are found. Amplicons are made during PCR or they can occur spontaneously, such as in the nucleic acid content of certain organisms or in tumors.				
Integrates results from all tools/workflows into a uniform sample or project page	Yes	No	No	Yes
Comments: Because Web Services only focus on one type of application (such as functional and taxonomic diversity within samples and differences among samples), the integration on a project page is centered on that application and the project page does not provide multiple views of the sample (e.g., only focused on assembly taxonomy and function). EDGE provides a more detailed panoramic view of the data allowing users to better understand and interpret their results.				
Provides PDF report	Yes	No	No	No
Comments: PDF reports may be useful as a means to share data, if it is not possible to allow access to local analyses (such as may be the case with CLC Genomics or local Galaxy instances). EDGE provides a succinct report of the analyses and results if internet services are not available.				
Databases available for complex analyses	Yes	No	No	Limited
Comments: The databases for Web Services are privately owned and cannot be readily updated by users. EDGE's databases are publically available and can be easily replaced by users.				
Focuses on tools or workflows	Workflows	Tools	Tools	Workflows
Comments: Galaxy and CLC Genomics users can link tools together to create workflows. However, the outputs of each tool are provided separately, making a single-source view of the results impossible. Because Web Services focus on specific applications, a single workflow is generally available, while EDGE provides a number of adaptable workflows that can be used for a number of different applications. Workflows are most convenient for non-bioinformatics users.				
Provides third-party tool additions?	Yes	Yes	Yes	No
Comments: EDGE Bioinformatics allows such additions, so long as they are added to the source code. Galaxy allows such additions, but users must develop ways of how to incorporate such new tools into				

the platform. CLC Genomics offers a plug-in method for incorporating command line tools.

EDGE BIOINFORMATICS COMPARISON MATRIX (continued)

Parameter	EDGE	Galaxy	CLC Genomics	Web Services (such as IMG or MG-RAST)
Installation ease and accessibility to online users	Experienced admin. Online demo servers	Experienced admin. Online servers	Experienced admin. No webservers	Available online only.
Comments: EDGE must be installed by an experienced system administrator. However, EDGE can be used immediately via our demo webservers. Galaxy must be installed by experienced system administrator. However, Galaxy can be used immediately via one of their servers. CLC Genomics must be purchased and can be readily installed by almost any user with authority to install software. Setting up a larger cluster-based CLC Genomics server requires an experienced system administrator. No webservers. Web Services are only available online and are essentially closed, black box options. Submissions are free but runtimes can be very long.				
Easy to use	Point-and-click	Multistep	Multistep	Point-and-click
Comments: Point and click means that users select data/project name and submit for analysis—results are presented as a "project" view. Multistep means that users must find and/or install tools, learn to create workflows, and navigate to the proper directory to find individual analyses—results are presented as a "tool-based" view.				
Level of expertise and turnaround time	Anyone Minutes to hours	Experienced Hours	Experienced Hours	Experienced Days to Weeks
Comments: Turnaround times for EDGE vary from minutes to hours based on workflows selected, complexity of the data, and the computational server used. Viewing results for the entire project is intuitively organized by workflow and is presented as a project webpage. Galaxy does not have the capability to reconstruct workflows comparable to EDGE. Users could spend hours on reconstructing EDGE workflows. Viewing results is less intuitive because each tool provides its own output in a different location, and there is only limited capability to view results within Galaxy, as most results would be native output text files. CLC Genomics does not have the capability to reconstruct workflows comparable to EDGE. Building some of the workflows similar to EDGE would take an experienced CLC user one or more hours. Runtime for users could be similar to users of EDGE, despite less functionality. Similar to Galaxy, viewing results is less intuitive, with each tool providing results separately. Because Web Services focuses on genes/proteins and annotations, most of the functionality demonstrated				

in EDGE cannot be performed for Web Services. Users select dataset(s) and submit them to the web service, but turnaround time often takes weeks and is never less than a day. An intuitive web-based integrated view of the results is presented to the user.

Describe how your product improves upon competitive products or technologies. Be SPECIFIC. Include such items as how much faster, how much less cost, etc.

***Anyone* can use EDGE Bioinformatics, and run it for a variety of analyses**. Any researcher or technician who either generates genomic data or uses sequencing data, can run EDGE Bioinformatics software via the publicly accessible web servers or by installing the open-source software on a local server or in the cloud. Once logged in, EDGE's breadth of application coupled with its ease of use is unparalleled.

Although Galaxy is also an open-source software platform, it is not natively point-and-click. In addition, to match EDGE's versatility, many tools would first need to be incorporated within Galaxy, a process that requires advanced bioinformatics knowledge. Even if workflows are created or selected, the outputs of each tool reside in their own directory. Thus, users must navigate to many different directories to find the results of the many analyses conducted on a single sample in order to provide the multidimensional view that EDGE provides within a single webpage. Lastly, several additional tools would need to be used independently outside of Galaxy, to view some of the results.

CLC Genomics, like other commercial options, is costly software and because it is proprietary, does not have the capability to incorporate many of the advanced tools that have recently been invented to address critical areas such as microbiome taxonomy classification. To reconstruct workflows that would be somewhat comparable to what is natively in EDGE, an experienced CLC user would need one or more hours. The results of a single workflow can be presented in graphical form, but the integration of many different analyses requiring multiple workflows, would be presented separately.

Because Web Services focuses on genes and proteins, and their annotations, most of the functionality demonstrated in EDGE Bioinformatics cannot be performed for Web Services. In addition, users select dataset(s), typically already assembled (requiring prior data manipulation) and submit them to the web service, but turnaround time often takes weeks and is never less than a day. Limited options are provided for analysis, since the Web Services are generally uniquely focused on a single type of analysis, thus only one workflow is provided.

EDGE Bioinformatics is easy to use. EDGE provides a wide array of sophisticated bioinformatics tools in easy-to-use workflows that are presented within a user-friendly

GUI (Graphical User Interface). This web-based GUI provides easy access from any internet-enabled computer with a browser and provides a standardized way to analyze and view results for genomic data, from anywhere in the world.

Although Galaxy provides a large array of bioinformatics tools along with some pre-configured workflows that one can manipulate in a web-based GUI, it does not provide a web-based interface for a per-sample analysis, nor does it provide visualization or interaction with the results of the selected workflows. In addition, most state-of-the-art workflows provided within EDGE are not natively available with Galaxy.

CLC Genomics aims to offer extremely flexible, tool-specific analysis. However, these solutions, like Galaxy, do not provide an integrated view of results from multiple workflows for a single sample, resulting in a variety of different folders holding different results. Additionally, many recent advances in read-based metagenomic taxonomy profiling or read-based phylogenetics are not options for this platform.

Web Services provide easy web-access to results of data and are generally easy to use. However, because they are singularly focused on highly specific types of analyses, they act more like individual workflows that are provided by EDGE, Galaxy, or CLC Genomics.

EDGE Bioinformatics is a fully integrated system. In EDGE, the results of the various workflows are integrated into a single, sample-based project page with visual and interactive displays, making it easy to investigate and interpret sequencing data from a sample. The various workflows effectively provide different "views" of data, thus presenting a more complete picture of the sample. Reports are automatically generated for each module, both graphic and tabular, and a full PDF report for each run (and each workflow) is provided, making it easy to rapidly share results, even via email.

Galaxy is an excellent workflow management system, but does not provide easy access to preconfigured workflows for state-of-the-art analysis of sequencing data for novice bioinformaticians or biologists. Galaxy is focused on workflows and tool management, but is not focused on the analysis of a sample.

Similar to Galaxy and EDGE, CLC Genomics offers several commercial software options and also provides a large array of bioinformatics tools. Like EDGE, CLC Genomics can provide some visual and interactive results. However, these solutions are

often very expensive, and many analyses are not sufficiently or adequately described to be validated or reproduced using other commercial, or open-source software.

Web Services provides analytics free for users. However, the software itself is not available, thus the inner-workings of the system are not fully explained and are not reproducible elsewhere. In addition, the features of Web Services are more specialized than EDGE. For example, MG-RAST is metagenome-specific, and provides users with a sample-based analysis of metagenome contents, but it will not allow the reconstruction of phylogenies, perform comparative analyses, etc. The IMG family of websites is annotation-centric, performing analysis of only genes that have been annotated and only allowing comparisons of these annotations. PATRIC is a pathogen-focused Web Service, and provides limited analysis of sequencing reads against a database of only some pathogens. Lastly, even those web servers with a metagenome focus do not allow more advanced analytics of read-based taxonomy classification or read-based phylogenetics.

EDGE offers unprecedented integration and functionality. EDGE provides a number of analyses that use both reads and/or assembled data, and can do so with metagenomic data (i.e., a community of organisms). No other software or platform provides such functionality. Given the recently reported importance of microbial communities in human and environmental health, and the relevance of such investigations (organisms natively exist in complex communities, thus the investigations more realistically represent nature when examining communities rather than cultures of individual species/strains), such types of analyses are becoming increasingly important. EDGE provides the widest array of tools among any other platform, in order to analyses these types of microbiome data.

Describe the limitations of your product/service. What criticisms would your competitors offer?

Requires some technical understanding. Although EDGE Bioinformatics makes analysis easy for even novices, users still must have an understanding of what the platform's tools do and understand the sequencing process to grasp the basics of interpreting results. Users should read all online documentation and understand how parameters can affect the outcome of analyses. To help ease biologists and analysts into genomics, Los Alamos National Laboratory and the Naval Medical Research Center

provide reach-back support and training to facilitate the use of EDGE Bioinformatics and the interpretation of results. Note: All other competing products or platforms require a similar level of expertise for interpretation of results and manipulation of parameters, if it is an option.

Requires expertise to integrate new tools. Integrating novel tools or workflows into the EDGE platform requires that users modify or contribute to EDGE's underlying code. Such work requires expertise in bioinformatics and programming. To address this limitation, the future version of EDGE Version 3.0 will have the ability to more readily accept third-party tools and workflows as either plug-and-play or with simple modification of configuration files and output design.

Galaxy does have the capability to integrate new tools, however, this process still requires expertise in bioinformatics. CLC Genomics provides programmers with the ability to write plug-ins, or users to purchase plug-ins available from CLC Genomics. Again, this requires experts in bioinformatics to write code to incorporate command-line scripts and programs. Because the software behind Web Services is not open source, it is not possible to integrate new tools or to modify workflows.

Installation can be involved. Users installing EDGE locally on a compute server or cluster may find that the installation can be involved and requires expertise with Unix. For easier installations, users can more readily install EDGE with an EDGE Docker or EDGE VMware version, however these installations come with an additional computing cost when using EDGE. Galaxy, like EDGE, offers the ability to install locally with similar complications. Galaxy also offers Docker but not VMware, for easier installation with similar added computing overhead. Fee-based commercial platforms like CLC Genomics do offer single executables that allow for easier installation on common computers (e.g., Windows OS). Web Services do not allow any local installations.

Summary

Important societal issues such as climate change and eradicating the spread of disease require an in-depth understanding of microorganisms and how they work. The key to understanding how best to use biological organisms, or how to treat and one day cure disease is locked in the genetic code. Sequencing as a technology is now readily available

to all, yet the tools for understanding big genomic data are not.

But now we have EDGE.

Developed by Los Alamos National Laboratory and the Naval Medical Research Center, EDGE Bioinformatics essentially "democratizes" the genomics revolution by enabling *any* researcher or physician to quickly and easily analyze complex genomic data.

With EDGE, it is now possible to bring the power of complex, big-data sequencing analysis to smaller research laboratories, including clinics, hospitals, and university laboratories. Applications span the entire field of biology, including medicine and infectious disease research, algal and other enhanced oil and biofuel production, and determining the role microbial communities play in fixing carbon from the atmosphere under the continuous changes in climate and their environment.

Support Letters

Defense Threat Reduction Agency

U.S. Army Medical Research Institute of Infectious Diseases (Department of the Army)

Department of Health & Human Services (Centers for Disease Control and Prevention)

Armed Forces Research Institute of Medical Sciences (United States Army Medical Directorate)

U.S. Naval Medical Research Unit #6 (Research Science Directorate)

National Center for Disease Control and Public Health (Republic of Georgia)

Agency for Defense Development (Republic of Korea)

Viome, Inc.

Appendix: Supporting Information

About the Cover and Captions for Product Photos

Comments from Current EDGE Bioinformatics Users

"List of EDGE Bioinformatics Users"

"Science on the Hill: Bringing the power of genetic research to an office near you," *Santa Fe New Mexican*, December 2016

"How bioinformatics tools are bringing genetic analysis to the masses," *Nature*, February 2017

"LANL's EDGE Offers Easy-to-Use Bioinformatics Pipelines for Microbial Sequence Analysis," *genomeweb*, December 2016

"Edge bioinformatics brings genomics to everyone," Los Alamos National Laboratory, November 2016

"Edge bioinformatics brings genomics to everyone," *EurekAlert (American Association for the Advancement of Science)*, November 2016

"EDGE 2.0 Modules Under Development"

"EDGE Fact Sheet"

"Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform," *Nucleic Acids Research* **45** (1), November 2016

Los Alamos Profile

Principal Investigators from Each of the Submitting Organizations

PI Name:	Patrick Chain
Title:	Scientist
Organization	Los Alamos National Laboratory
Phone:	505-665-4019
Email:	pchain@lanl.gov

PI Name:	Theron Hamilton
Title:	Head, Genomics & Bioinformatics
Organization	Naval Medical Research Center
Phone:	301-619-1265
Email:	theron.c.hamilton.mil@mail.mil

Full Development Team

Team Member Name:	Po-E Li
Title:	Research Technologist
Organization	Los Alamos National Laboratory
Phone:	505-664-0943
Email:	po-e@lanl.gov

Team Member Name:	Chien-Chi Lo
Title:	Research Technologist
Organization	Los Alamos National Laboratory
Phone:	505-665-7798
Email:	chienchi@lanl.gov

Team Member Name:	Karen Davenport
Title:	Research Technologist
Organization	Los Alamos National Laboratory
Phone:	505-667-8676
Email:	kwdavenport@lanl.gov

Team Member Name:	Yan Xu
Title:	Software Developer
Organization:	Los Alamos National Laboratory
Phone:	505-667-6274
Email:	yxu@lanl.gov

Team Member Name:	Sanaa Ahmed
Title:	Graduate Student
Organization:	Los Alamos National Laboratory
Phone:	505-665-4019
Email:	sahmed@lanl.gov

Team Member Name:	Pavel Senin
Title:	Post–Doctoral Researcher
Organization	Los Alamos National Laboratory
Phone:	505-667-4847
Email:	psenin@lanl.gov

Team Member Name:	Migun Shakya
Title:	Post–Doctoral Researcher
Organization	Los Alamos National Laboratory
Phone:	505-667-5571
Email:	migun@lanl.gov

Team Member Name:	Kimberly A. Bishop-Lilly
Title:	Deputy Head, Genomics & Bioinformatics Dept.
Organization	Naval Medical Research Center
Phone:	301-619-1490
Email:	kimberly.a.bishop-lilly.ctr@mail.mil

Team Member Name:	Joseph J. Anderson
Title:	Bioinformatics Analyst
Organization	Naval Medical Research Center BDRD
Phone:	804-504-1486
Email:	joe@getedge.org

Team Member Name:	Logan J. Voegtly
Title:	Software Developer
Organization	Naval Medical Research Center
Phone:	301-619-1486
Email:	logan.j.voegtly.ctr@mail.mil

Team Member Name:	Casandra W. Philipson
Title:	Computational Biologist
Organization	Naval Medical Research Center
Phone:	301-619-1696
Email:	casandra.w.philipson.civ@mail.mil

Marketing and Media Information

Contact person to handle all arrangements on exhibits, banquet, and publicity.

First Name:	Janet
Last Name:	Mercer-Smith
Title:	R&D 100 Coordinator
Organization:	Los Alamos National Laboratory
Email:	mercer-smith_janet@lanl.gov
Phone:	505-665-9574

Contact person for media and editorial inquiries.

First Name:	Patrick
Last Name:	Chain
Title:	Scientist
Organization:	Los Alamos National Laboratory
Email:	665-4019
Phone:	pchain@lanl.gov

Company Logo (CAS will take care of this)

LANL Facebook Page URL: https://www.facebook.com/LosAlamosNationalLab

About the Cover

The sequence of DNA is the genetic blueprint of life and is present in the cells of all free-living organisms. Understanding an organism's DNA code, or genome, provides insight into what an organism is capable of doing or of mutations that result in disease. Sequencing and associated technologies now make it possible to take any type of sample with biological cells, and determine the DNA within those cells, and hence identify the organism(s). While there exist innumerable applications for these technologies, the remaining high-tech barrier has been the computational interpretation of sequencing data. EDGE Bioinformatics provides a novel, user-friendly, automated solution to this complex Big Data challenge, and allows regular scientists, physicians, and technicians to interpret and gain insights from the sequencing data.

The top portion of the cover shows a clinician evaluating the results of today's Big Data-generating sequencing instruments.

The bottom portion of the cover shows physicians taking body-fluid samples to identify either genetic mutations or to uncover infectious diseases. The method relies on extracting and determining the DNA sequence of the cells within the sample. Each different type of cell, or organism, has a unique genome, which allows it to be identified or allows mutations to be identified.

Caption for EDGE Flowchart

This flowchart presents an overview of the components that make up EDGE Bioinformatics. Inputs from the user (light blue parallelograms) are provided by point-and-click (I is required, II–V are optional). EDGE modules (green rectangles) perform distinct computational operations and can be selected by point-and-click. Some of the outputs (in text format or as graphics) are shown in darker blue. Some of the output text files, serve as optional inputs to other modules.

Caption for EDGE Project Page

The EDGE platform is principally divided into the home page, the page that sets up sequencing analytic runs, and the project pages which allow users to view the results of analyses. These main displays are selectable to the left of this website. Here, the user has selected a project in the project list, which is highlighted in yellow. The projects displayed in the drop-down list are also colored according to whether the project is running (orange), is completed (green) or is in the queue to be run (grey). Because of the user's selection, the current results of a project that is undergoing analysis is displayed in the middle of the page. The project name is displayed at the top along with a summary of the project. Underneath the summary lie a number of sections with details of the results of the project. The first section displayed is the "General" section, which outlines the various pipelines selected to be run by the user and their status. Several links and reports are also available in this section. Additional sections below this provide the actual results of any of the pipelines and workflows selected, including graphics, tables, and links to results, reports, and log files (not displayed). A panel to the right allows users to see real-time the progress of their sample, together with the status of the computational server that is being used to process the data. Additional actions allow users to view a live log of the command lines used to perform the analyses, to rerun, interrupt (pause), or delete the project, to empty the results of the project or archive the project to a different storage server, or to share the project with individual EDGE users, or with the general public.

Caption for Genome Browser View in EDGE

This figure shows an example of one type of output derived from selecting an Ebola genome under the "Reference-Based Analysis" module (see the figure titled "EDGE input") in EDGE. Scientists can view the genomes as lines and/or strings of letters representing the DNA code (A, T, G, C) using genome "browsers" such as JBrowse. The translated code of triplets of DNA is represented by 26 possible amino acids (also represented by letters and lines).

Top: A zoomed out view of a reference Ebola genome (~18,000 DNA characters, or 'nucleotides'), with coding sequences (CDS) outlined in green (and in blue, the named description of the translated protein code). The main input into EDGE was a new sample

with suspected Ebola in it, and both the assembly (long blue lines) and the reads (short red and blue lines; the colors indicate different directions of alignment) from the sample are aligned to the Ebola reference. Here, two assembled segments align to nearly the complete Ebola genome, strongly suggesting that Ebola was indeed in the sample. The reads support this assessment with lots of data.

Bottom: A zoomed in view of a small section of the Ebola reference genome. Two additional tracks (horizontal sections) are displayed that indicate suspected mutations (called single nucleotide polymorphisms, or SNPs, identified by comparing the sequence of assembled segments aligned to the reference; or single nucleotide variants, or SNVs, that are identified by comparing the reads aligned to the reference). In the reads track (called "Mapping reads to reference"), regions of each read that do not match the reference are highlighted by distinguishing colors (A: green, T: red, G: yellow, C: blue).

A true SNP is identified at position 14,212 where the assembly and all supporting reads agree that the Ebola genome found in the sample contains a C instead of a T at this position. A true variant is outlined at position 14,190 where the assembled segment suggests the Ebola in the sample is identical to the reference, but a fraction of the reads clearly show that a mutation has occurred at this position in the Ebola genome (a T instead of a C).

Such mutations in infecting viruses are commonly found during the course of infection and are known as viral quasispecies. Other differences between individual reads and the reference are likely sequencing errors and not true variants. All these analyses occur behind the scenes when using EDGE, making it very easy and fast for scientists or physicians to understand what may reside within specific samples.

Comments from Current EDGE Users

"EDGE provides web-based interfaces that are ready to use and user-friendly EDGE provides clear and convenient visualization of various results such as QC results. For metagenomics, it simplifies running multiple analyses that would normally be required. These tools have different strengths and samples often require several analyses for pathogen identification. It simplifies the workflow of data analysis. It provides a "one stop shop" for results."

–Sue Tong, Centers for Disease Control and Prevention, Atlanta, GA

—

"We are very grateful for EDGE, particularly for those folks that do not do command line [tools] and can now process their own data."

– Mariana Leguia, United States Naval Medical Research Unit No. 6, Lima, Peru

—

"Edge helps in detecting rapidly pathogens possibly present in sample from patient with characterized or unknown etiology syndromes. The taxonomic classification, reference-based analysis, amplicon analysis, assembly modules and host removal tool are the most applicable to our work."

– Andy Nkili, Centre International de Recherches Médicales de Franceville, Gabon

—

"[EDGE] simplifies the workflow of data analysis. It provides a "one stop shop" for results."

–Krista Queen, Centers for Disease Control and Prevention

—

"We use EDGE to process genomic data from viruses and metagenomic data to look for new viruses. [The] EDGE host removal module is one of the most easy-going tools because this module has preloaded a short list of genomes that facilitates its use."

– Armando Torre, United States Naval Medical Research Unit No. 6, Lima, Peru

—

"Thank you for your kind [help] about Hanta virus tree building (and for help with the Hanta virus SNPdb)."

–Daesang Lee, Agency for Defense Development, Republic of Korea

—

"You guys hit the nail on the head with depth of coverage [to identify plasmids]. What you explained on the disparity between chromosome and plasmid hits and taxon reporting is exactly what I was referring to in our results."

– Captain Turner Conrad, US Army Medical Research Institute of Infectious Disease

—

"I don't have to know PERL or Python scripting! You guys have done a wonderful job of grabbing Illumina sequencing run output, taking it through . . . several analysis tools [which] have been bundled [and] which are applied to the data in a sequential manner to get a nice final output."

– Raju Lathigra, United States Army Medical Research Institute of Infectious Disease

—

"I did test the pipeline. It installed without any hitches and [it was] very easy to run. Got very interesting results from some Illumina data that had poor indexing done. The samples were from Mosquitoes collected around Lake Baringo when it had flooded during a dry season. Thank you."

– George Ngondi, International Livestock Research Institute and The Africa Genomics Centre and Consultancy, Nairobi, Kenya

—

"I am a pathologist with a background in molecular biology and molecular-based detection methods for diseases. I have done some Illumina-based transcriptomics and I am now working with Dr. Afonso's group on some metagenomic-based analyses for NDV, but with an interest in applying [next-generation sequencing] more broadly to diseases.

"Yes, I tried out the software with one of our other Illumina samples to see how it works. I liked it. It is easy to use, but parameter options are still there. I also liked that I could easily extract the Fastq/Fasta from the Taxonomy Classification for each of the methods. That is very handy to be able to do that with just a click. Very nice work!"

– James Stanton, The University of Georgia, College of Veterinary Medicine, Department of Pathology, Athens, Georgia

—

"Your recent bioinformatics programs (EDGE) to perform common tasks is quite useful . . . You guys have a wonderful team and really are doing excellent jobs! However, I am the lone guy here doing bioinformatics (just joking, no complaint here) . . . I would love to have a collaboration with you and Dr. Chain's group. Having an EDGE module for antibiotics gene detection and [single nucleotide polymorphism] analysis will be really helpful to scientific community. I have not done any bioinformatics pipeline development for the last 10 years, more like a data analyst right now . . . This software pipeline is great after I did several rounds of testing! I like it very much!"

– Xianghe Yan, PhD, Environmental Microbial and Food Safety Laboratory, USDA Agricultural Research Service (ARS), Beltsville, Maryland

List of EDGE Bioinformatics Users
List One: Users within the United States

Private Industry	Healthcare Providers	National/Government Labs (within US and overseas)	US Military (within US and overseas)	Academia
▪ Battelle Memorial Institute ▪ Becton, Dickinson and Company ▪ Biocept, Inc. ▪ Conagen, Inc. ▪ Digital Infuzion ▪ FourPartsWater, Ltd. ▪ Harris Corporation ▪ HudsonAlpha Institute for Biotechnology ▪ Mitre Corporation ▪ MRIGlobal Research Institute ▪ One Codex ▪ Signature Science ▪ Viome, Inc.	▪ Children's National Health System ▪ Nationwide Children's Hospital ▪ Providence Health and Services	▪ Centers for Disease Control and Prevention, Atlanta, GA ▪ Centers for Disease Control and Prevention, Nairobi, Kenya ▪ Food and Drug Administration ▪ Lawrence Berkeley Laboratory ▪ Lawrence Livermore National Laboratory ▪ Los Alamos National Laboratory ▪ National Center for Biotechnology Information ▪ New Mexico Department of Health ▪ USDA/ARS-Animal and Plant Health Inspection Service	▪ Naval Medical Research Center ▪ US Armed Forces Research Institute of Medical Sciences-Thailand ▪ US Army Medical Research Institute for Infectious Diseases ▪ US Army Medical Research – Kenya ▪ US Naval Medical Research Unit #2-Cambodia ▪ US Naval Medical Research Unit #3- Egypt ▪ US Naval Medical Research Unit #6- Peru	▪ Broad Institute of MIT and Harvard ▪ California Academy of Sciences ▪ California State University ▪ Cornell University ▪ Georgia Institute of Technology ▪ James Madison University ▪ Johns Hopkins University ▪ Missouri State University ▪ Montana State University ▪ New Mexico State University ▪ Ohio State University Wexner Medical Center ▪ Rutgers University School of Dental Medicine ▪ University of Arkansas ▪ University of California- San Diego ▪ University of Colorado- Boulder ▪ University of Florida ▪ University of New Mexico ▪ University of North Carolina-Charlotte ▪ University of Pennsylvania ▪ University of Texas Medical Branch ▪ University of Washington ▪ Virginia Commonwealth University

List of EDGE Bioinformatics Users
List Two: Users outside the United States

Private Industry	Healthcare Providers	National/Government Labs	Academia
▪ Bionivid, India ▪ Biotech Diagnostics, Germany ▪ Genome Life Sciences, India ▪ Innov4Sight (I4S), Singapore ▪ RASA Life Science Informatics, India	▪ Hôpital Universitaire Pitié-Salpêtrière, France ▪ National University Health System, Singapore ▪ Newcastle upon Tyne Hospitals, England ▪ Tan Tock Seng Hospital, Singapore	▪ Agency for Defense Development, Republic of Korea ▪ Centre for Ecology & Hydrology, United Kingdom ▪ Defence Canada, Canada ▪ Defence Science and Technology Laboratory, United Kingdom ▪ Defence Research Establishment (FFI), Norway ▪ DSO National Laboratories, Singapore ▪ International Centre for Medical Research in Franceville, Gabon ▪ Kazakh Scientific Center of Quarantine and Zoonotic Diseases, Kazakhstan ▪ Kenya Medical Research Institute, Kenya ▪ Ministry of Defense, Australia ▪ Ministry of Health, Singapore ▪ National Center for Disease Control, Republic of Georgia ▪ National Institute of Advanced Industrial Science and Technology, Japan ▪ Research Institute for Biological Safety Problems, Kazakhstan ▪ Uganda Virus Research Institute, Uganda	▪ Bioinformatics Institute A*STAR, Singapore ▪ Carleton University, Canada ▪ Curtain University, Australia ▪ Dublin City University, Ireland ▪ Eberhard Karls Universität Tübingen, Germany ▪ Jeju National University, Republic of Korea ▪ Jordan University of Science and Technology, Jordan ▪ Korea Advanced Institute of Science and Technology, Republic of Korea ▪ Korea University, Republic of Korea ▪ McGill University, Canada ▪ Memorial University of Newfoundland, Canada ▪ National University of Ireland- Galway, Ireland ▪ National University of Singapore, Singapore ▪ Ruhr University Bochum, Germany ▪ Universidad Austral de Chile, Chile ▪ Universidad Complutense de Madrid, Spain ▪ Universidad de Sevilla, Spain ▪ Universidad de Vigo, Spain ▪ Universidade de Aveiro, Portugal ▪ Universidade de Sao Paulo, Brazil ▪ Universitair Medisch Centrum Groningen, Netherlands ▪ Universität Basel, Switzerland ▪ Universitat Zurich Institut fur Medizinische Virologie, Switzerland ▪ Université de Lausanne, Switzerland ▪ Université de Strasbourg, France ▪ Universiti Malaysia Kelantan, Malaysia ▪ University College London, England ▪ University of Alberta, Canada ▪ University of Liverpool, England ▪ University of Melbourne, Australia ▪ University of Turku, Finland ▪ University of Waikato, New Zealand ▪ Zhejiang University, China

Science on the Hill: Bringing the power of genetic research to an office near you

By Patrick Chain
For the New Mexican | Posted: Sunday, December 11, 2016 11:30 pm

Most of us have gone to the doctor ourselves or taken our children with a sore throat and sinus congestion, only to find our physician couldn't readily tell whether we had a cold virus or a bacterial infection. Just in case, we might have walked away from the appointment with a possibly unnecessary prescription for antibiotics.

Now imagine that a nurse could swipe your saliva and run a quick genetic test for bacteria. If it comes back negative, this time you walk out with just a script for decongestant and orders to get some rest instead of buying an unnecessary prescription and contributing to the antibiotic resistance crisis.

That's just one example of the benefits of rapid genetic screening on a personal level. On a grander scale, the ability to quickly analyze genetic data stands to revolutionize research into everything from the mutations causing various cancers to the "Second You," your microbiome, or the bacteria living inside you.

Genomics can also revolutionize our understanding of a range of diseases — Alzheimer's, irritable bowel syndrome, Crohn's disease, for instance — as well as how to grow algae to best produce oil to make gasoline. In medicine, genetic screening can tell hospital staff what pathogens inhabit the hospital environment. In environmental research, it can clarify how communities of microorganisms fix carbon from the atmosphere and how their populations adapt to less rain and hotter summers.

Bringing the power of genetic research to an office near you

The DNA code in a genome is built from molecular units called bases (identified by the letters A, T, G, and C) that pair with each other to form the iconic double helix. When strung together, these base pairs form the instructional code for all life to reproduce and grow. By sequencing genomes of various organisms and comparing them to each other, scientists are making breakthroughs in understanding infectious disease, cancer, and even climate change.

Genomics — the genetic mapping and DNA sequencing of sets of genes or the complete genomes of organisms, along with related genome analysis and database work — is emerging as one of the

transformative sciences of the 21st century. Partly that's resulting from the rapid spread of so-called next-generation sequencing instruments, which have become accessible to the average biologist and, eventually, to the physician. Gene sequencing has become much more democratized over the last few years.

Decreasing costs for sequencing instruments is driving their spread to new users, making them available to the common scientist. Today you'll find sequencers not only in most universities and other large research institutions, but also in hospitals, individual clinics and the small labs of individual researchers. Genomics has become the cornerstone of all biological research, which almost always involves sequencing all the genes (the genome) of the organism under study or the many species forming a community to see what's going on. So everyone wants their own capability in-house.

All this easily and rapidly generated data has caused a new bottleneck, as the ability to analyze the data — and it's very big Big Data —i s swamping genomics. Bioinformatics tools use computers to pull together, classify, store, process and analyze molecular genetic and genomic data. Unfortunately, the current tools are not entirely user-friendly or accessible to most biological researchers, who have more expertise in biology than in crunching data.

Seeing a need that the unique expertise at Los Alamos National Laboratory could fill, a team in the Biosecurity and Public Health group, collaborating with the Naval Medical Research Center, has developed a new computational and web-based tool called EDGE Bioinformatics to fulfill the promise of democratizing genomics.

Funded by the Department of Defense's Defense Threat Reduction Agency, the work comes out of the lab's decades of research in genetics and life sciences. Long interested in the link between radiation and genetic mutations, the U.S. Department of Energy and the National Institutes of Health received federal funding in 1998 to begin the Human Genome Project to sequence, or map, the genome of the species Homo sapiens — us.

Los Alamos was a key player, contributing its expertise in life sciences, particularly genetics and its world-class computing resources to the task of unraveling the human genetic code. By June 2003, the map was mostly complete. Since then, the lab has applied its expertise to a range of related genetic research, from illuminating the causes of cancer to perfecting algae for biofuel production.

For the Los Alamos EDGE team, it was a natural step from this background to creating a handy, easy-to use, web-based computer program with a wide assortment of integrated and pioneering bioinformatics tools. EDGE includes several pre-configured workflows to analyze sequencing data, identify genomes and create reports and graphics based on the data. Using EDGE, with a few mouse clicks a novice in bioinformatics can create sophisticated analyses of a sample in minutes instead of days or weeks.

This bioinformatics platform was designed as an initial attempt at empowering the development of genomics expertise — that's what EDGE stands for. EDGE has already helped streamline data analysis for groups in multiple countries worldwide as well as within several government laboratories in the United States. Because the program is "open source," anyone can use it or even modify it to suit their needs and bring the power of Big Data Analysis to even the smallest research lab — or doctor's office.

Genomics researcher Patrick Chain is the EDGE team leader in the Biosecurity and Public Health group at Los Alamos National Laboratory. With a background in microbial ecology, evolution, genomics and bioinformatics, Chain has spent the past 20 years using genomics to study various microbial systems.

He currently leads a team of researchers whose charge is to devise novel methods, algorithms and strategies for the biological interpretation of massively parallel sequencing data.

TOOLBOX

DEMOCRATIZING BIOINFORMATICS

Computational biologists are starting to develop platforms that open up the ability to analyse and interpret genetic-sequence data.

ILLUSTRATION BY THE PROJECT TWINS

BY JEFFREY M. PERKEL

For doctors trying to treat people who have symptoms that have no clear cause, gene-sequencing technologies might help in pointing them to a diagnosis. But the vast amount of data generated can make it hard to get to the answer quickly.

Until a couple of years ago, doctors at US Naval Medical Research Unit-6 (NAMRU-6) in Lima had to send their sequence data to the United States for analysis, a process that could take weeks — much too long to make pressing decisions about treatment. "If all you could do was get the data that you then have to ship to the US, it's almost useless," says Mariana Leguia, who heads the centre's genomics and pathogen-discovery unit.

But Leguia no longer has to wait for the analyses; she can get results in days or even hours — and she can do them in her own lab. Her unit makes use of EDGE (Empowering the Development of Genomics Expertise), a bioinformatics tool that hides common microbial-genomics tasks, such as sequence assembly and species identification, behind a slick interface that allows users to generate polished analyses. "We can have actionable information on site that allows us to make decisions very quickly about how to go forward," Leguia says.

EDGE isn't the first tool to simplify informatics with a point-and-click interface. Indeed, it lacks much of the flexibility and ▶

scope of more established alternatives such as Galaxy and Illumina's BaseSpace platform. But its simplicity is drawing in users who might otherwise shun bioinformatics. "People have used [EDGE] who would never have bothered learning command-line tools," says Clinton Paden, who uses EDGE in his work on virus pathogenesis at the US Centers for Disease Control and Prevention in Atlanta, Georgia. As such, it represents a case study in democratizing genome informatics — one that could help to accelerate uptake of the field by pure biologists.

INFORMATICS IN THE FIELD

Patrick Chain, who led the development of the software[1], at Los Alamos National Laboratory (LANL) in New Mexico, says that EDGE was created to try to square the rapidly growing availability of low-cost DNA sequencers with the relative paucity of know-how required to make sense of the data. It is designed for use in facilities that lack expertise in bioinformatics, says Joe Anderson, a computational biologist who honed the software for military applications at the Biological Defense Research Directorate (BDRD) at the Naval Medical Research Center in Frederick, Maryland.

It is also open-source, self-contained and provides end-to-end analyses for microbial genomics, from raw sequence reads to species identification and phylogeny in a single click. The system is also relatively cheap to run because the recommended hardware configuration (256 gigabytes of memory and 64 processors) can be bought for less than US$10,000, says Anderson. This means that most labs that can afford to run sequencing projects can afford the hardware. "That's not throw away money, but it's cheap enough," he says. It also helps that the set-up doesn't rely on an Internet connection and can be powered by a generator.

Users with reliable network connections can install the system to a cloud network. Nicholas Loman, a bioinformatician at the University of Birmingham, UK, points to CLIMB, the Cloud Infrastructure for Microbial Bioinformatics, which he helped to develop. CLIMB is a free service specifically dedicated to academics in the United Kingdom who are working on microbial genomics.

CLIMB was supported by £8.4 million (US$10.5 million) from the UK Medical Research Council and incorporates several informatics tools, including sequence databases and an analysis workbench known as the Genomics Virtual Laboratory. "I'm definitely thinking about having EDGE as a possible option on there as well," Loman says.

Overall, EDGE has been officially installed at 18 US Department of Defense and partner-nation labs, and on every continent except Antarctica, says Theron Hamilton, who is head of genomics and bioinformatics at the BDRD.

One of those is in Phnom Penh at the NAMRU-2 facility, which uses the system to track vector-borne diseases. "It's not traditionally the kind of place you would go to do bioinformatics," says Anderson. But EDGE is changing that. "One of the things I've realized is that, if you give [researchers] tools and get out of the way, they will amaze you," Anderson says.

The latest version of EDGE — version 1.5, released last October — includes 54 third-party tools. All components, including algorithms, databases, visualization tools and reference genomes, are housed on a server that drives six interlocking analysis modules: sequence clean-up; assembly and annotation; comparison to reference genomes; taxonomic identification; evolutionary analysis; and PCR primer design. Additional modules, including RNA analysis and pathogen detection, are slated for the upcoming EDGE 2.0, Chain says.

Last November, Chain and his colleagues demonstrated EDGE's capabilities in a study in which they used the platform to assemble, classify and map the evolutionary relationships in isolates of the bacteria *Bacillus anthracis* and *Yersinia pestis*; to untangle a mock human microbiome; and to analyse a series of human clinical samples, including cases of Ebola virus and *Escherichia coli* infection[1]. But the first published use of the system actually pre-dates that study by several months. Leguia's lab used EDGE to optimize methods for whole-genome sequencing of dengue virus — in a study published last June[2].

"People have used EDGE who would never have bothered learning command-line tools."

Users can explore those and other data sets using a free demo hosted on the LANL server. Researchers who wish to analyse their own sequences must install the software on their own systems. The code is freely downloadable from GitHub, and a Docker container and virtual machine image are available, but an information-technology expert will probably be required to handle the installation, says Chain. It is possible to tweak the source code to add other tools and workflows, but that's beyond the capabilities of many users, Chain acknowledges. A mechanism to simplify the process is in development, he says.

Paden, who has a background in computer science, says that the tool's simplicity makes computational biology accessible to researchers who might otherwise be intimidated by the usual tool for bioinformatics work — the computer's text-based command line.

But Titus Brown, a computational scientist at the University of California, Davis, warns that some of the benefits of EDGE are tempered by shortcomings that could limit the software's long-term use. He describes EDGE as an example of "opinionated software". "It gives you a small set of software to run that's been tuned to a specific set of examples," he says, "and it gives nice graphical summaries and outputs." But, he notes, it isn't clear how other researchers might help to improve the tool, nor what will happen should its funding dry up.

Chain says that the team made EDGE open-source partly because of concerns over future funding, which are also informing future development plans. "Sustainability is a question we have to think about," Chain says, "which is why we're going to try to allow third-party implementers to much more easily plug-and-play their projects, most likely using Docker."

A GALAXY OF TOOLS

EDGE is not the first bioinformatics system to offer a user-friendly interface. Galaxy, first published[3] in 2005, allows researchers to assemble informatics pipelines from a vast and flexible toolbox of free software offered through a web-based interface. Users can solve nearly any problem they can dream up by combining these tools in different ways.

But Galaxy can be intimidating to use. And, unlike the graphical representations generated by EDGE, such as phylogenetic trees or interactive 'Krona' plots of taxonomic data in hierarchical pie charts, Galaxy's output tends to take the form of processed data files, which the user then needs to take elsewhere to visualize.

"Galaxy is more like a kitchen, but there's no dining room," says Jeremy Leipzig, a software developer in the Department of Biomedical and Health Informatics at the Children's Hospital of Philadelphia, Pennsylvania. "The system is not really there for coming up with a way of delivering that output in an appealing way," he says. "With EDGE, they've actually thought about what the reports should look like."

Nathan Watson-Haigh, a bioinformatician at the University of Adelaide in Australia, says that EDGE could help to ease pressure on overworked bioinformaticians. But he cautions that it remains a complicated bioinformatics tool, and biologists who are inexperienced in computation would be wise to consult an expert before placing too much certainty in their results.

As with any tool, they need to understand what the algorithms are doing, and how different parameters affect their output, adds Kathleen Fisch, interim director of the Center for Computational Biology and Bioinformatics at the University of California, San Diego. "Just because you can run the tools doesn't mean that you should run the tools."

Still, as bioinformatics tools get ever easier, informatics could lose some of its aura of complexity. And for biologists, that could lead to wider adoption — and democratization. ■

1. Li, P.-E. *et al. Nucleic Acids Res.* **45,** 67–80 (2017).
2. Cruz, C. D. *et al. J. Virol. Methods* **235,** 158–167 (2016).
3. Giardine, B. *Genome Res.* **15,** 1451–1455 (2005).

https://www.genomeweb.com/informatics/lanls-edge-offers-easy-use-bioinformatics-pipelines-microbial-sequence-analysis

LANL's EDGE Offers Easy-to-Use Bioinformatics Pipelines for Microbial Sequence Analysis

Dec 22, 2016 | Uduak Grace Thomas

NEW YORK (GenomeWeb) – Researchers at the Los Alamos National Laboratory and the Naval Medical Research Center have developed a web-based bioinformatics platform called Empowering the Development of Genomics Expertise (EDGE) that is designed to help users with limited or no bioinformatics expertise use existing tools to analyze and interpret microbial genomic sequence data.

EDGE Bioinformatics integrates hundreds of public, open-source software and internally developed tools that are designed to process primarily Illumina raw reads. Available pipelines allow users to assemble, annotate, and compare genomes as well as characterize complex clinical or environmental samples including data from bacterial, archaeal, and viral isolates or shotgun metagenome samples. There are also methods for visualizing the output of taxonomy classification tools for easy comparison as well as links to output directories where data from each pipeline is stored.

According to a paper published recently in *Nucleic Acids Research*, the tools available in EDGE were selected for the quality of results that they provide across sample types, their speed, and the computational resources that are required to run them. It packages publicly available open-source software into six modules that can be run individually or in combination. "We've done a robust comparison between a number of different tools and we've assembled together some basic workflows where we are aiming to get 80 to 90 percent of the questions answered for 80 to 90 percent of the problems that the user might have," Patrick Chain, leader of the bioinformatics and analytics team and the metagenomics program within LANL's Biosecurity and Public Health group, and lead for the EDGE development team, said in an interview.

The list includes well-known tools such as Blast, BowTie, Burrows-Wheeler Aligner, Kraken, MetaPhlan, IDBA, and SAMtools. Full details of the tools included in the platform are provided in accompanying documents on the EDGE website. These tools have been assembled into ready-to-run pipelines for sample pre-processing, *de novo* assembly and annotation, comparing samples to reference genomes, taxonomic classification, phylogenetics, and PCR primer analysis. It also includes pre-processing and reference-based analysis functions for eukaryotic genomes. Users can tweak the default settings of the pipelines as well as activate or deactivate some steps depending on their needs. They can also view the results of their analysis at the genus, species, or strain level.

Compared to available alternative environments for NGS data analysis such as Galaxy, "EDGE is the only open-source platform that can be used locally and that integrates both the processing of individual

samples and the presentation of results in a seamless web-based interface." It's also unique because it provides pre-selected algorithms and parameters for users rather than letting them choose and combine tools into workflows themselves which can be daunting for novice bioinformaticians. "You have to know what tools you want to pick for your particular analysis [and] that's not always intuitive," Chain said.

Furthermore, compared to EDGE, Galaxy doesn't provide much visualization. "You can create workflows and you can use that workflow to run your data through [but] then you have to run around for another program to feed your outputs [into for visualization]," he noted. In contrast EDGE provides users with quality control graphics, assembly summary charts, heat maps, and phylogenetic trees. It also links to third-party visualization tools such as the JBrowse genome browser.

EDGE is also a cheaper alternative to commercial packages that can be "inflexible" and can affect interpretation results if users don't know the details of the proprietary algorithms that the packages use, the developers wrote.

The paper also describes the results of a few analysis experiments performed to demonstrate the efficacy of the EDGE platform. One of these focused on two sequence datasets from separate isolate genome sequencing projects involving *Bacillus anthracis* and *Yersinia pestis* strains. According to the paper, results from EDGE's assembly and annotation module were consistent with known genomic elements from the microbes including known insertion sequences and rRNA operons. The assembled sequences were also consistent with the known genome size and number of genes found in the microbes. They were also able to confirm the expected identities of the sequenced organisms using taxonomy classification tools available in EDGE.

The researchers also used EDGE to successfully characterize pathogenic sequences in a number of clinical samples including one from the recent Ebola outbreak and one from a fecal sample collected from a patient infected with *Escherichia coli*.

Currently, EDGE is used by research groups around the world as well as in several government laboratories in the United States. "There are some collaborators that are using this to teach individuals how some tools work," Chain said. For example, "there are a number of funded programs to teach graduate students to collect various organisms and analyze them."

Turner Conrad, a research microbiologist in the diagnostics systems division of the United States Army Medical Research Institute for Infectious Diseases (USAMRIID) and one of the platform's beta testers, highlighted EDGE's ease of use compared to some existing workflow environments. "When you look at something like Galaxy or any of these other workflow managers or workflow software ... they are so broad and open that you have to figure out how to make your own pipelines," he said. "The advantage that EDGE puts forth is ... they've still made it general use enough while offering a more universal type of workflow where you just pick and select what you want out of that whole thing to do [but] it's still all one big workflow."

EDGE's source code is available from GitHub. Researchers can also access the code in Docker containers and virtual machine images for local installation. The developers have provided a publicly accessible webserver that can be run with publicly available data from repositories such as the National Center for Biotechnology Information's Sequence Read Archive and the European Molecular Biology Laboratory's European Nucleotide Archive — the webserver does not support upload of personal datasets for security reasons. EDGE's modular design and open source license allow other researchers to expand its capabilities beyond the initial implementation, according to the developers. They can also

integrate the platform into their existing workflows.

The developers recommend that researchers running EDGE use computers that have at least 16GB of memory and eight central processing units available to run pipelines — using more CPUs will reduce run times. For their next steps, the developers hope to add more tools to the EDGE platform including RNA- and 16S- sequence data analysis pipelines, Chain said. They are also currently testing an amplicon sequence analysis pipeline that they hope to integrate into the platform. Also, some current users have requested new visualization tools, he said. They will also work on creating definitions and methods that will allow third-party developers to contribute best-practice tools and workflows to the platform.

Filed Under Informatics software developers LANL USAMRIID bioinformatics microbial sequencing software

EDGE bioinformatics brings genomics to everyone

A new bioinformatics platform will help democratize the genomics revolution by allowing users with limited bioinformatics expertise to quickly analyze and interpret genomic sequence data.

November 29, 2016

Contact

Nick Njegomir
Communications Office
(505) 665-9394
Email

"We realized that while next-generation sequencing instruments are becoming more widespread and more accessible to the average biologist or physician, the bioinformatics tools required to process and analyze the data were not as user-friendly or accessible," said Patrick Chain.

New platform enables rapid genomic-sequence data analysis

LOS ALAMOS, N.M., Nov. 29, 2016 – A new bioinformatics platform called Empowering the Development of Genomics Expertise (EDGE) will help democratize the genomics revolution by allowing users with limited bioinformatics expertise to quickly analyze and interpret genomic sequence data. Researchers at Los Alamos National Laboratory and their collaborators at the Naval Medical Research Center developed EDGE, which is described in a paper recently published in *Nucleic Acids Research*.

"We realized that while next-generation sequencing instruments are becoming more widespread and more accessible to the average biologist or physician, the bioinformatics tools required to process and analyze the data were not as user-friendly or accessible," said Patrick Chain, of Los Alamos' Biosecurity and Public Health group and EDGE team lead. "Given the large number of applications where sequencing is now used, a robust bioinformatics platform that encapsulates a broad array of algorithms is required to help address questions a researcher may have. We sought to develop a web-based environment where non-bioinformatics experts could easily select what pipelines they need and rapidly obtain results and interact with their data."

Stopping the spread of disease—from naturally occurring or manmade threats—requires an in-depth understanding of pathogens and how they work. To this end, the ability to characterize organisms through accurately and rapidly comparing genomic data is an important part of Los Alamos' national security mission.

Technology advancements have fueled the development of new sequencing applications and will flood current databases with raw data. A number of factors limit the use of these data, including the large number of associated software and hardware dependencies and the detailed expertise required to perform this analysis. To address these issues, Chain and his team have developed an intuitive web-based environment with a wide assortment of integrated and pioneering bioinformatics tools in pre-configured workflows, all of which can be readily applied to isolate genome sequencing projects or metagenomics projects.

EDGE is a user-friendly and open-source platform that integrates hundreds of cutting-edge tools and helps reduce data analysis times from days or weeks to minutes or hours. The workflows in EDGE, along with its ease of use, provide novice next-generation sequencing users with the ability to perform many complex analyses with only a few mouse clicks. This bioinformatics platform is described as an initial attempt at empowering the development of genomics expertise, as its name suggests, for a wide range of applications in microbial research.

EDGE has already helped streamline data analysis for groups in Thailand, Georgia, Peru, South Korea, Gabon, Uganda, Egypt and Cambodia, as well as within several government laboratories in the United States.

The paper "Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform" was published in Nucleic Acids Research in partnership with the Defense Threat Reduction Agency, the Naval Medical Research Center-Frederick and the Henry M. Jackson Foundation.

About Chain

Patrick Chain earned his master's of science in microbial genomics from McMaster University and his doctoral degree in molecular microbiology and molecular genetics at Michigan State University. He is currently leading the Bioinformatics and Analytics Team and the Metagenomics Program within the Biosecurity and Public Health group at Los Alamos National Laboratory.

His background is in microbial ecology, evolution, genomics and

bioinformatics, having spent the past 20 years using genomics to study various microbial systems, including the human microbiome, other environmental metagenomic communities, various isolate microbes or single cells, including bacterial and viral pathogens as well as fungal, algal, plant and animal systems. He currently leads a team of researchers whose charge is to devise novel methods, algorithms and strategies for the biological interpretation of massively parallel sequencing data.

About Los Alamos National Laboratory

Los Alamos National Laboratory, a multidisciplinary research institution engaged in strategic science on behalf of national security, is operated by Los Alamos National Security, LLC, a team composed of Bechtel National, the University of California, BWXT Government Group, and URS, an AECOM company, for the Department of Energy's National Nuclear Security Administration.

Los Alamos enhances national security by ensuring the safety and reliability of the U.S. nuclear stockpile, developing technologies to reduce threats from weapons of mass destruction, and solving problems related to energy, environment, infrastructure, health, and global security concerns.

PUBLIC RELEASE: 29-NOV-2016

EDGE bioinformatics brings genomics to everyone

New platform enables rapid genomic-sequence data analysis

DOE/LOS ALAMOS NATIONAL LABORATORY

Media Contact

Nick Njegomir
njegomir@lanl.gov
505-665-9394

@LosAlamosNatLab

http://www.lanl.gov

LOS ALAMOS, N.M., Nov. 29, 2016 -- A new bioinformatics platform called Empowering the Development of Genomics Expertise (EDGE) will help democratize the genomics revolution by allowing users with limited bioinformatics expertise to quickly analyze and interpret genomic sequence data. Researchers at Los Alamos National Laboratory and their collaborators at the Naval Medical Research Center developed EDGE, which is described in a paper recently published in *Nucleic Acids Research*.

IMAGE: LOS ALAMOS'S CHERYL GLEANER (LEFT) DEMONSTRATES TO STUDENTS HOW TO USE EDGE BIOINFORMATICS TO ANALYZE THEIR SEQUENCE DATA. AT ITS ANNUAL SEQUENCING AND BIOINFORMATICS TRAINING IN JUNE, LOS ALAMOS HOSTED... **view more ›**

CREDIT: LOS ALAMOS NATIONAL LABORATORY

"We realized that while next-generation sequencing instruments are becoming more widespread and more accessible to the average biologist or physician, the bioinformatics tools required to process and analyze the data were not as user-friendly or accessible," said Patrick Chain, of Los Alamos' Biosecurity and Public Health group and EDGE team lead. "Given the large number of applications where sequencing is now used, a robust bioinformatics platform that encapsulates a broad array of algorithms is required to help address questions a researcher may have. We sought to develop a web-based environment where non-bioinformatics experts could easily select what pipelines they need and rapidly obtain results and interact with their data."

Stopping the spread of disease--from naturally occurring or manmade threats -- requires an in-depth understanding of pathogens and how they work. To this end, the ability to characterize organisms through accurately and rapidly comparing genomic data is an important part of Los Alamos' national security mission.

Technology advancements have fueled the development of new sequencing applications and will flood current databases with raw data. A number of factors limit the use of these data, including the large number of associated software and hardware dependencies and the detailed expertise required to perform this analysis. To address these issues, Chain and his team have developed an intuitive web-based environment with a wide assortment of integrated and pioneering bioinformatics tools in pre-configured workflows, all of which can be readily applied to isolate genome sequencing projects or metagenomics projects.

EDGE is a user-friendly and open-source platform that integrates hundreds of cutting-edge tools and helps reduce data analysis times from days or weeks to minutes or hours. The workflows in EDGE, along with its ease of use, provide novice next-generation sequencing

users with the ability to perform many complex analyses with only a few mouse clicks. This bioinformatics platform is described as an initial attempt at empowering the development of genomics dxpertise, as its name suggests, for a wide range of applications in microbial research.

###

EDGE has already helped streamline data analysis for groups in Thailand, Georgia, Peru, South Korea, Gabon, Uganda, Egypt and Cambodia, as well as within several government laboratories in the United States.

The paper "Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform" was published in Nucleic Acids Research in partnership with the Defense Threat Reduction Agency, the Naval Medical Research Center-Frederick and the Henry M. Jackson Foundation.

About Chain

Patrick Chain earned his master's of science in microbial genomics from McMaster University and his doctoral degree in molecular microbiology and molecular genetics at Michigan State University. He is currently leading the Bioinformatics and Analytics Team and the Metagenomics Program within the Biosecurity and Public Health group at Los Alamos National Laboratory.

His background is in microbial ecology, evolution, genomics and bioinformatics, having spent the past 20 years using genomics to study various microbial systems, including the human microbiome, other environmental metagenomic communities, various isolate microbes or single cells, including bacterial and viral pathogens as well as fungal, algal, plant and animal systems.

He currently leads a team of researchers whose charge is to devise novel methods, algorithms and strategies for the biological interpretation of massively parallel sequencing data.

About Los Alamos National Laboratory

Los Alamos National Laboratory, a multidisciplinary research institution engaged in strategic science on behalf of national security, is operated by Los Alamos National Security, LLC, a team composed of Bechtel National, the University of California, BWXT Government Group, and URS, an AECOM company, for the Department of Energy's National Nuclear Security Administration. Los Alamos enhances national security by ensuring the safety and reliability of the U.S. nuclear stockpile, developing technologies to reduce threats from weapons of mass destruction, and solving problems related to energy, environment, infrastructure, health, and global security concerns.

EDGE 2.0 Modules Under Development

EDGE 2.0 will likely be released in late 2017 or early 2018. The team plans to roll out three new modules at this time.

- **Pathogen Analysis and Characterization**: Although EDGE already provides a number of analysis methods that can help identify pathogens, the modules have not been optimized to identify or characterize pathogens. This will help streamline efforts by scientists who wish to understand the nature of any pathogen found within the sample and provide them with a detailed view of the pathogen(s) in question.

- **Differential Gene Expression (RNA-Seq) Analysis**: Although DNA can help identify all the genes in an organism, RNA is produced as a result of gene activation (i.e., "expression"), which represents what the cell(s) are doing at a given time within a given set of environmental conditions. This information provides scientists with crucial information about what genes are involved in specific processes. The main methods used in these types of studies are to examine the differences in expression of genes over time or given different environmental conditions. EDGE will provide users with the ability to compare different samples to provide a list of genes that are differentially expressed under different conditions or at different times.

- **Presence/Absence of Targeted Amplicons**: Although EDGE currently provides the ability to determine the outcome of a PCR assay given the sequencing data of a sample, some scientists have specifically used multiple PCR assays on a sample and wish to determine which, if any, of the assays have positive results. For example, some users wish to streamline sequencing efforts on many hundreds of samples to screen for the presence of many (up to hundreds or thousands) of different pathogens. Therefore, EDGE will provide the ability for users to specify the type of assay they are performing, and determine/report if any of the assays are positive and in which samples they are found.

EDGE Bioinformatics

Los Alamos NATIONAL LABORATORY EST. 1943

What is EDGE?

EDGE bioinformatics was developed to help biologists process NGS data 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 an intuitive web-based interface for user input, allows users to visualize and interact with selected results, 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.

The initial release of EDGE provides the following analytical workflows:

- **Pre-processing (data QC and host removal)**
- **Assembly and annotation**
- **Reference-based analysis**
- **Taxonomy classification**
- **Phylogenetic analysis**
- **PCR analysis**

The latest release (version 1.5) includes several new features:

- **AMR and virulence genes identification**
- **16S/18S/fungal ITS analysis using QIIME**
- **Metadata collection/storage**
- **Comparative analysis of taxonomic classification of multiple metagenomic samples**

EDGE 2.0 will include these modules:

- **Pathogen analysis and characterization**
- **Differential gene expression (RNA-Seq)**
- **Presence/absence of targeted amplicons**

Implementing a strategy for allowing rapid incorporation of 3rd party tools is an ongoing effort.

EDGE availability:

A complete version of EDGE is available as a variety of packages that can fit individual needs, including:

- **source code** https://github.com/LANL-Bioinformatics/EDGE/releases

- image in **VMware** http://edge.readthedocs.io/en/v1.5/installation.html#edge-vmware-ovf-image

- **Docker** container http://edge.readthedocs.io/en/v1.5/installation.html#edge-docker-image

A **demonstration** webserver (with version 1.1) for use with publicly available data (download from SRA/ENA) can be found at: https://bioedge.lanl.gov/edge_ui/

EDGE requirements

The current version of EDGE pipeline has been extensively tested on Linux platforms with Ubuntu 14.04 and CentOS 6/7 operation system and will only work on 64bit Linux environments.

Due to the involvement of several memory/time consuming steps, we normally recommend computers with at least 16GB memory and 8 CPUs, though we typically use servers with a minimum of 256GB memory with 16 CPUs.

Detailed documentation for installation on a local server is provided online:
http://edge.readthedocs.io/en/v1.5/installation.html

For more information about EDGE:

EDGE is described in our paper recently published in *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1027

Published online 24 November 2016

Nucleic Acids Research, 2017, Vol. 45, No. 1 **67–80**
doi: 10.1093/nar/gkw1027

Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform

Detailed documentation for the software can be found online: http://edge.readthedocs.io/en/v1.5/

LA-UR-16-26934

Published online 24 November 2016

Nucleic Acids Research, 2017, Vol. 45, No. 1 ***67–80***
doi: 10.1093/nar/gkw1027

Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform

Po-E Li[1,†], Chien-Chi Lo[1,†], Joseph J. Anderson[2,3], Karen W. Davenport[1], Kimberly A. Bishop-Lilly[3,4], Yan Xu[1], Sanaa Ahmed[1], Shihai Feng[1], Vishwesh P. Mokashi[3] and Patrick S.G. Chain[1,*]

[1]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, [2]Defense Threat Reduction Agency, Fort Belvoir, VA 22060, USA, [3]Naval Medical Research Center-Frederick, Fort Detrick, MD 21702, USA and [4]Henry M. Jackson Foundation, Bethesda, MD 20817, USA

Received June 16, 2016; Revised October 12, 2016; Editorial Decision October 17, 2016; Accepted October 18, 2016

ABSTRACT

Continued advancements in sequencing technologies have fueled the development of new sequencing applications and promise to flood current databases with raw data. A number of factors prevent the seamless and easy use of these data, including the breadth of project goals, the wide array of tools that individually perform fractions of any given analysis, the large number of associated software/hardware dependencies, and the detailed expertise required to perform these analyses. To address these issues, we have developed an intuitive web-based environment with a wide assortment of integrated and cutting-edge bioinformatics tools in pre-configured workflows. These workflows, coupled with the ease of use of the environment, provide even novice next-generation sequencing users with the ability to perform many complex analyses with only a few mouse clicks and, within the context of the same environment, to visualize and further interrogate their results. This bioinformatics platform is an initial attempt at Empowering the Development of Genomics Expertise (EDGE) in a wide range of applications for microbial research.

INTRODUCTION

The field of genomics has made tremendous technological leaps in recent years, and the combined decrease in sequencing costs and expansion in applications (transcriptomics, metagenomics, single cell genomics) have truly revolutionized the way scientists approach biological questions (for a recent review, see (1)). Now that a trained technician can single-handedly produce gigabases of sequence data in essentially a day's work, 'next generation sequencing' (NGS) is being applied by many smaller laboratories, as well as the large traditional sequencing centers, across a wide range of disciplines in order to answer a variety of complex problems. For instance, NGS is being applied to the characterization and attribution of outbreaks in clinical environments (2), food safety (3), the development of alternative energy sources (4,5) and many other fields.

Although many advances have been made in bioinformatics methods development, the so-called 'democratization of genomics' (6) has not yet fully expanded to the bioinformatic realm, making it difficult for investigators to adequately analyze genomic big data (7,8). While NGS no longer seems new, it has really only been since 2005 that a revolutionary new technology (pyrosequencing) (9) was introduced after more than twenty years of chemical degradation (10) and chain termination (Sanger (11)) sequencing. Some of these NGS technologies have already been abandoned even after strong market performance; other new technologies are only now emerging, and the ones that have thus far survived continue to undergo improvement. Despite reads of limited length, Illumina® (12) currently dominates the market, in part due to its very high throughput and low cost.

Analysis of the massive datasets produced in NGS studies and interpretation of the results requires expertise in both computer science and biology and often experience in statistics, applied math, or other fields such as biochemistry and ecology depending on the experiment at hand and goals of the project. Bioinformatics is always the first step to transform a sample's raw NGS data into interpretable data that can be further analyzed or compared with data collected from other samples. Although the decreasing cost and decreasing laboratory footprint of NGS technologies make

*To whom correspondence should be addressed. Tel: +1 505 665 4019; Fax: +1 505 665 3024; Email: pchain@lanl.gov
†These authors contributed equally to the work as first authors.

the production of these datasets a more realistic goal for many laboratories, there still remain a number of core issues in bioinformatics that hamper the broader use of NGS data, including the broad range of questions that can now be asked with NGS (i.e. different goals), the plethora of highly specific tools to choose from, and the expertise required to install and use these tools. The numerous and diverse specific questions being asked of NGS data often require highly specialized algorithms and pipelines. While any given question can sometimes make use of the same basic tool(s) with different parameters and post-processing, other questions may require similar bioinformatic manipulation but are optimally answered using different tools, and further questions may require developing entirely new methods or adapting existing algorithms that were originally designed for other purposes. The related issue of having numerous available (and somewhat redundant) options for extremely complex data analysis requires users to become familiar with these options as well as their computational and algorithmic limitations. Because NGS data and their formats can change frequently, the analytical tools must also adapt; new tools arise frequently through efforts to improve upon initially developed algorithms, or to complement other methods. One can often identify dozens of individual tools that can perform similar types of analyses, and it has been an increasing challenge to decide which tools are best for which specific applications. In addition, some tools are tailored to specialized hardware architectures. Lastly, many laboratories do not have the degree of expertise required to implement robust methods, install the appropriate tools, or construct standardized pipelines for processing data. The need for such expertise can delay studies and make comparisons of disparate studies very difficult.

Because we view bioinformatics as the key bottleneck in the use and interpretation of NGS data, we present an integrated platform toward Empowering the Development of Genomics Expertise (EDGE). This bioinformatics effort is intended to truly democratize the use of NGS for exploring microbial genomes and metagenomes. EDGE also provides limited capability of analyzing eukaryotic data as well (e.g. reference-based alignments can be performed, but assembly/annotation is not currently supported). We developed EDGE Bioinformatics as an initial suite of preconfigured bioinformatics workflows that allow rapid analysis of raw (FASTQ) NGS data, coupled with result visualization and interactive features (Figure 1, Supplementary Figure S1). This software lowers the barrier to NGS bioinformatic analysis by providing a down-selected array of tools using well-tested parameter settings across an array of different sample types. Best of breed software tools were selected for the quality of their results among various sample types, for their speed, and for the computational resources required to run them. The interactive results are presented on a sample-by-sample basis and allow users to explore ongoing data processing within an intuitive and user-friendly web-based environment. 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 the user's sample from various analytical standpoints, but biologists are always encouraged to understand how each tool and algorithm functions, and to have some insight into how the results should best be interpreted.

Alternative platforms for NGS data analysis do exist, however EDGE is the only open source platform that can be used locally and that integrates both the processing of individual samples and the presentation of results in a seamless web-based interface. The most similar platform is the Galaxy environment (13), which is also open source and can perform a multitude of different analyses of both isolate genomes or metagenomes, allowing users to select from a large number of pre-integrated tools to construct workflows (some preconstructed workflows are also available). However, the selection amongst so many seemingly similar tools can be daunting for novice bioinformaticians and the installation of additional capabilities, such as read-based taxonomic classification algorithms, can be challenging. While the raw result files can be accessed for each individual analysis, Galaxy also does not currently support a full integration of post-processed graphics, tables or other results from orthogonal analyses of individual samples. EDGE provides a single, integrated results page for each processed sample, and for novel analyses such as read-based taxonomic classification, the results of multiple tools can be displayed. A more costly option includes commercial packages that can perform many similar operations to Galaxy and EDGE, and also allow visualization of results, however these packages often use proprietary software that can be inflexible (e.g. word size used for assembly), and can impact interpretation of results if one does not know the details of the algorithm used. While several useful web services do exist, these are generally focused on specific organisms such as pathogens (e.g. PATRIC (14)), or specific types of NGS analyses such as differential gene expression (e.g. GenePattern (15)), isolate genome annotation and annotation comparisons (e.g. IMG (16), RAST (17)), or metagenomic annotation and annotation comparisons (e.g. IMG/M (18), MG-RAST (19)). The webservices that provide comparative genomic capabilities generally rely on private databases and the software is not open source. EDGE provides a complementary suite of NGS analysis capabilities, is freely available, and is designed to be locally installed to provide an array of analytical tools for microbial isolates or metagenomes.

To fit diverse institution-specific needs, EDGE Bioinformatics is available in a variety of options. For full installation, EDGE source code can be obtained via GitHub. Both a Docker container and a VMware (OVF) virtual machine image are provided to simplify local installation. For demonstration purposes, a publicly accessible EDGE webserver (https://bioedge.lanl.gov/) is also provided for use with publicly available data.

METHODS

EDGE Bioinformatics computational design

EDGE Bioinformatics is built around a collection of publicly available, open-source software packaged in six modules. The main wrapper script is written in Perl, while the various tools currently include BLAST (version 2.2.26) (20), BowTie2 (version 2.1.0) (21), BWA (version 0.7.9)

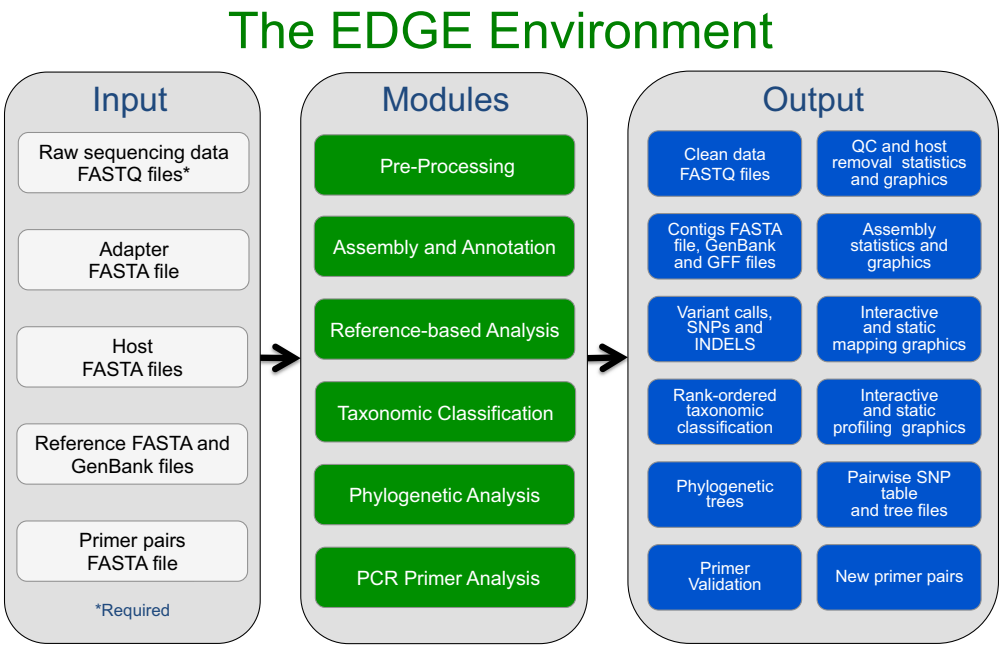


Figure 1. An overview of the EDGE Bioinformatics Environment. The only inputs required from the user are raw sequencing data and a project name. The user can create specific workflows with any combination of the modules. In addition, tailored parameters dictating how each module functions can be modified by the user. EDGE outputs a variety of files, tables and graphics which can be viewed on screen or downloaded. A more detailed overview is shown in Supplementary Figure S1. All Modules are described in the Methods section.

(22), FaQCs (version 1.33) (23), FastTree (version 2.1) (24), GOTTCHA (version 1.0b) (25), IDBA_UD (version 1.1.1) (26), SPAdes (version 3.5.0) (27), JBrowse (version 1.11.6) (28), jsPhyloSVG (version 1.55) (29), Kraken (version 0.10.4-beta) (30), KronaTools (version 2.4) (31), MetaPhlAn(version 1.7.7) (32), MUMmer3 (version 3.23) (33), Phage_Finder (version 2.1) (34), PhaME (*bioRxiv 032250; doi:* http://dx.doi.org/10.1101/032250), Primer3 (version 2.3.5) (35), Prokka (version 1.11) (36), RATT (version 08-Oct-2010) (37), RAxML (version 8.0.26) (38) and SAMtools (version 0.1.19) (39).

All tools and modules can be run on the Unix command line, however we provide a user-friendly web-based graphic user interface (GUI). The GUI is primarily implemented using the JQuery Mobile javascript framework and HTML5 on the client-side, and implements Perl CGI using Apache or Python on the server-side. This implementation makes EDGE accessible on any platform, including all smartphones, tablets, and desktop devices. The EDGE software tools were selected or developed based on the desire (and need) for both accuracy and speed, with the assumption of moderate computational hardware resources. Additional detail regarding the installation, implementation, and the tools encompassed within EDGE can be found at http://edge.readthedocs.org/.

The modular design and open source license also allow other researchers to expand the available capabilities beyond our initial implementation. For expert bioinformaticians, another benefit is that EDGE can also be integrated into other workflows and be used via command line to submit jobs on a cluster. More information can be found at the EDGE homepage (https://lanl-bioinformatics.github.io/EDGE/), and the software is available at https://github.com/LANL-Bioinformatics/edge. To simplify installation, a VM in OVF (https://edge.readthedocs.io/en/latest/installation.html#edge-vmware-ovf-image) or a Docker image (https://edge.readthedocs.io/en/latest/installation.html#edge-docker-image) can also be obtained. The EDGE demonstration webserver is available at https://bioedge.lanl.gov/ with the example data sets from this manuscript available to the public to view and/or re-run and also allows users to run publically available data (Supplementary Figure S2) deposited in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) or the European Nucleotide Archive of the European Molecular Biology Laboratory (EMBL ENA). This webserver does not currently support upload of any other data (due in part to LANL security regulations), however local installations and the available images are fully functional. The EDGE software is intended to be run while connected to the internet, but can be run entirely offline, with only a few links to third party websites that would be non-functional. EDGE was designed to be implemented within an institution and linked to local raw data (FASTQ) repositories, meaning that the user's data can remain private.

EDGE Bioinformatics has been primarily designed to analyze microbial (bacterial, archaeal, viral) isolates or shotgun metagenome samples. The optional analytical pipelines include pre-processing quality control, assembly and annotation, comparison to reference genomes, taxonomic classification, phylogenetics and primer analysis. Due to the

complexity and computational resources required for eukaryotic genome assembly and annotation, and the fact that several of the current taxonomy classification tools do not support eukaryotic classification, EDGE does not fully support eukaryotic samples. However, pre-processing and reference-based analysis functions are able to support eukaryotic genomes.

One of the key features of the EDGE Bioinformatics platform is that the visualization of the results is fully integrated with, and accessible directly on, the webpage in real time. Many graphics are displayed on each project page as thumbnails that link to either a full-page view or a lightbox (quick zoom) view, including quality control graphics, assembly summary charts, heat maps, phylogenetic trees, etc. In addition, there are links to the interactive genome browser JBrowse and to interactive classification results via Krona, as well as links to output directories where all resulting data for each pipeline are stored.

Because some of the most challenging aspects of genomics involve the exponentially increasing size of datasets and the resources required to move large datasets, a key benefit of the EDGE Bioinformatics software is that it can be implemented on a stand-alone server that can access datasets in local storage or in network-mounted space. We have tested EDGE Bioinformatics with datasets of up to hundreds of millions of reads, on a variety of servers (e.g. 12–64 core servers with 64–512GB of RAM), with run times ranging from minutes to hours. Using more CPUs will decrease runtime (see Table 1). All analyses described in this study were performed on our demonstration server which is a Dell PowerEdge R720 with 24 cores, 512GB RAM, and 7TB disk space. On this particular webserver, we allow any user to sign up for an account and run publicly accessible FASTQ files (from SRA/ENA).

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.

The project page layout

A left navigation menu on the EDGE website provides access to the Home page, the Run EDGE page (to initiate a new project) and the Projects list, allowing users to navigate to any desired project page (Supplementary Figure S3). A page for each project is produced as soon as it is launched within EDGE and allows the user to monitor the progress of the run and access the output summaries of each pipeline as they complete in real time. Each project page provides a summary of the project, and under a 'General' tab, a description of the input(s) provided, the modules selected for the run along with their run time statistics, and access to log files, the output directory, and a final PDF report.

A link in the upper right corner provides access to a sliding panel that contains a job progress widget, a resource monitoring widget, and an action widget. Once the job is submitted, the job progress widget reports the status for each analysis step in real time. The resource monitoring widget provides a real time view of the computational system running EDGE, and allows the user to anticipate whether there are sufficient resources to simultaneously run additional jobs, or if some projects should be moved to a different storage location. For example, projects will fail to complete one or more of the modules if there is insufficient storage for the outputs. The action widget provides the user some flexibility over the project, including allowing a user to interrupt, rerun, delete, and move his or her submitted jobs. The user can also share the project with other users, publish the project such that any user can access the results, or make the project private again ('unpublish'). In addition, there is a command line 'live log' view, which displays the real time actions and the Unix commands launched by EDGE.

The EDGE modules and their outputs

All of the six main modules within the EDGE Bioinformatics environment are optional and can be selectively run as individual modules or in any combination, thus affording the user maximum flexibility in customizing each analysis to particular specifications. These consist of: (i) a pre-processing module that performs quality control, trimming, and removal of sequences matching an unwanted target (e.g. host removal); (ii) a *de novo* assembly module which assembles the data, validates the assembly, and annotates the resulting contigs; (iii) a reference-based analysis module, which allows users to select one or more references to which reads (and contigs) are compared; (iv) a taxonomy classification module, which classifies reads (and contigs); (v) a phylogenetics module, which calculates a core genome, determines all SNPs, and infers a phylogenetic tree from a number of input genomes and (vi) a primer and assay module which allows users to validate *in silico* known primers against the *de novo* assembly, or to design new primers that uniquely amplify short sequences within the *de novo* assembly. The latter module does require an assembly for primer analysis.

Each module comprises a Perl wrapper with one or more bioinformatics tools tailored to handle NGS reads and/or contigs, as well as several scripts to parse and post-process the results. The users can also adjust a limited set of parameters or toggle options within each module. EDGE produces a web page for each project with many different summaries of the results for each module, including the statistics of the run (each module and time to completion), summary log files and a PDF summary of all results, along with more detailed results of each individual module. Each module outputs a number of files, which are accessible via a directory link and are summarized with both text and figures along with some interactive graphics all within the context of the website.

Pre-processing (Supplementary Figure S1, module 1). This module consists of two independent, selectable pipelines. For data quality control, the FaQCs software is used to analyze all reads for quality and to trim or filter out reads using default parameters, unless these are changed by the user (optional). Using an input reference FASTA, EDGE

Table 1. Descriptions of samples and EDGE modules tested

Sample description	Sample type (material)	# of reads (millions)	Sequence type	EDGE Modules[a]						CPUs	Run time (h)
				1	2	3	4	5	6		
Bacillus anthracis strain SK-102 SRR1993644	Isolate (gDNA)	28.6	HiSeq 2×101 nt	X	X	X	X	X	X	8	04:12:03
Bacillus anthracis strain SK-102 SRR1993644	Isolate (gDNA)	28.6	HiSeq 2×101 nt	X	X	X	X	X	X	20	03:33:52
Yersinia pestis strain Harbin 35 SRR1993645	Isolate (gDNA)	15.0	GAII 2×110 nt	X	X	X	X	X	X	8	03:35:39
Human Microbiome Project (staggered mock community) SRR172903	Metagenome (DNA)	7.93	GAII 75 nt	X	X		X			8	00:53:59
Patient plasma sample 2014 *Ebola* outbreak (IDBA assembly) SRR1553609[b]	Metagenome (RNA)	0.930	HiSeq 2×100 nt	X	X	X	X			12	00:38:07
Patient plasma sample 2014 *Ebola* outbreak (SPAdes assembly) SRR1553609[b]	Metagenome (RNA)	0.930	HiSeq 2×100 nt	X	X	X	X			12	00:47:24
Patient fecal sample 2011 *E. coli* outbreak SRR2164314	Metagenome (DNA)	273	HiSeq 2×100 nt	X	X		X			8	34:43:30
Patient nasal swab acute respiratory illness SRP062772[b]	Metagenome (DNA)	2.52	MiSeq 2×300 nt	X	X		X			8	00:20:59

[a]EDGE Modules are described in Materials and Methods: 1. Pre-Processing; 2. Assembly and Annotation; 3. Reference-Based Analysis; 4. Taxonomic Classification; 5. Phylogenetic Analysis; 6. PCR Primer Analysis.
[b]These samples were retrieved directly from the NCBI SRA.

can also filter unwanted reads that align to a selected reference. While this 'Host Removal' function was originally envisioned to exclude host reads when inputting clinical samples or those derived from known animals, this component can remove any data that aligns to the input reference, allowing users to selectively remove any other target genome(s). Some built-in references include the most recently updated GRCh38 Human reference and the Enterobacteriophage phiX 174 ('PhiX'), which is often used as a control within Illumina sequencing runs. This module aims to provide high quality, clean reads for any subsequent analysis by EDGE. If this module is not selected, the raw data will be used for all downstream process modules.

Statistics and graphical outputs of the data, prior to and after processing, are provided for user interpretation, along with access to the cleaned data files. The major outputs of this module are shown in Supplementary Figure S1A–C and example screen shots of output from the EDGE webpage can be found in Supplementary Figure S4.

Assembly and annotation (Supplementary Figure S1, module 2). EDGE performs *de novo* assembly with the input reads using either IDBA-UD or SPAdes. Because each of these assemblers performs and combines multiple assemblies, both tools are capable of providing reasonable assemblies from a wide variety of sample types, including isolate genomes, single cell projects, and metagenomes. IDBA-UD is used by default (due to time and memory considerations—SPAdes is more RAM-intensive), and the assembly parameter option for kmer sizes begins with $k = 31$ with a step size of 20, until a maximum kmer size is reached (dependent on the read lengths). When this module is selected, assembly validation is performed by mapping the short read input data to the assembled contigs using Bowtie2. Additionally, the user can select to have the assembly annotated (default behavior) using a modified Prokka tool (for the rapid annotation of prokaryotic genomes), and prophages within microbial genomes are detected using Phage_Finder. If there is an available reference that is sufficiently similar to the target genome assembly, EDGE can also use a modified version of the Rapid Annotation Transfer Tool (RATT) to transfer the annotation from the reference GenBank file (a required input for this step) to the assembly. When SPAdes is selected as the assembler, there exists an additional option to input long read data (PacBio or Nanopore) which can help in gap closure and repeat resolution.

The results of this module include the assembled contigs FASTA file, assembly and assembly validation statistics and graphics, the annotation files (gbk and gff), and an interactive JBrowse implementation, which provides visualization of the contigs and their annotation. The major outputs of this module are displayed in Supplementary Figure S1D–G and example screenshots can be found in Supplementary Figures S5 and S6.

Reference-based analysis (Supplementary Figure S1, module 3). When this module is selected, the user must choose one or more reference genomes (FASTA or Genbank formats) to which the reads (and contigs, if assembly was performed) are compared. RefSeq genomes (Bacteria, Archaea, Viruses) are available from a dropdown menu or the user can provide a path to one or more input references. Reads are aligned to the input reference using BowTie2 and variants are identified using SAMtools. Any regions left uncovered by reads are also identified and reported in text files. Similarly, contigs are aligned to the same reference(s) using MUMmer and the results parsed using Perl scripts to catalogue SNPs and small insertions or deletions (indels), as well as regions within the contigs that may be novel and do not align to the reference. If Genbank reference files are provided, the variants, SNPs, and uncovered regions of the reference are further analyzed to output any affected genes and reports are generated to display whether the changes also contribute to synonymous or non-synonymous substitutions within coding regions. Reads and contigs that do not map to the reference are parsed into separate FASTA/Q files and an option is available to align these reads and contigs to RefSeq for taxonomic identification.

In addition to the output text files, several graphics along with statistics are provided that outline linear coverage of

the reference, depth of coverage along the reference, number of variants, as well as percentages of input reads and contigs mapped to the reference. Interactive JBrowse views allow for the display of the reference and associated annotation (genes, rRNAs, etc.), along with detailed views of the aligned reads and contigs, as well as any SNPs or small indels that have been discovered. The major outputs of this module are displayed in Supplementary Figure S1G–I, while an example output can be found in Supplementary Figure S7.

Taxonomy classification (Supplementary Figure S1, module 4). Envisioned primarily for use with metagenomic datasets or with novel genomes, this module allows both read-based and contig-based classification (the latter performed if assembly was also selected). For taxonomic classification of the reads, the user can select one or more of several available metagenome tools (currently GOTTCHA, Kraken and MetaPhlAn) along with BWA, a read mapper used against RefSeq. The default is to run all tools to take advantage of their different strengths, and to provide users with additional information to help interpret their data. Each of these classifiers has its own algorithm and database, parameters for the search, and required input format, all of which are automatically managed within the EDGE platform. The specific output formats of each tool are unified into a common framework to generate the reports/graphs displayed by EDGE. There is also an option to classify only unassembled reads, if assembly is selected and the user desires to only classify unassembled data.

The results of each read-based taxonomy profiling method are summarized in comparative views (heatmap plots and radar charts summarize the top hits of each tool) at the user-selected level of taxonomy (genus, species, strain). Results are also presented in more detail in individual tool-based views with taxonomy tree dendrograms and Krona charts while more detailed outputs can be found within the directory links.

For contig classification, EDGE aligns contigs to NCBI's RefSeq database using BWA-mem. While contigs can match multiple taxa, each segment within a contig is assigned to a unique taxon based on best hit score. While the total length within all contigs is calculated per taxon, each contig is also assigned to a unique taxon based on linear coverage. Both the total length per taxon (Length barplot) and the number of contigs (Count barplot) assigned to a taxon are reported, along with a scatterplot showing the identity of the contig, its fold coverage by reads, and its G+C content. These results are reported at all levels of taxonomy using the last common ancestor algorithm.

The major outputs of this module are displayed in Supplementary Figure S1J and K, while example outputs can be found in Figures 2 and 3, and Supplementary Figures S8 and S9.

Phylogenetic analysis (Supplementary Figure S1, module 5). Because phylogenetic analysis is a highly desired feature for many genomic investigations, we utilize a portion of a newly developed tool, PhaME, which provides the ability to infer a whole genome SNP-based tree from completed genomes, genome assemblies, and even from reads. This tool works with viruses, bacteria, archaea and single cell eukaryotes, but should not be used for multi-ploidy organisms. Because this tool is based on nucleotide alignments and SNP identification, the recommended use of this module is to select the genomes or assemblies of closely related strains or species for the alignments in order to appropriately place the user's target genome within the context of a species or genus tree. Briefly, contigs and completed genomes are compared with one another to identify conserved segments while ignoring repeated regions, and reads are mapped to one of these references to continue the identification of a conserved core genome. The core genome alignment is used to identify all SNPs from all datasets (reads, contigs, genomes) and FastTree (default, for speed considerations) or RAxML can be used to generate a phylogenetic tree. This module was envisioned for use primarily with isolate genome projects (however metagenomes have also been successfully used), where a target genome comprises the majority of the sequencing data (thus allowing for genome assembly and sufficient read-mapping to allow accurate SNP calling) and the user desires to accurately place this target genome within the context of near neighbor genomes. The user must select datasets from near neighbor isolates as references to which the sample's reads and contigs (if assembly was selected) will be added to infer a phylogeny. Three additional datasets (at minimum) 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.

The Newick format tree files, core genome FASTA, and SNP statistics are available in the directory link and the phylogenetic trees, generated using jsPhyloSVG, are provided for easy viewing in either rectangular or circular tree formats (Outputs L and M in Supplementary Figure S1). The input sample (reads and/or contigs) is highlighted within the trees. An output screenshot can be found in Supplementary Figure S10.

PCR primer analysis (Supplementary Figure S1, module 6). EDGE also supports both the design and validation of PCR primers based on the assembly. In the validation pipeline, known primers within a user-specified input file are mapped to the assembly using BWA, given a user-defined number of mismatches (default of 1) to determine if an amplicon would be generated. The user can also select a pipeline to design new primers based on the assembly, that will differentiate the input sequenced sample from all other bacteria, archaea, and viruses in NCBI's RefSeq database. In this design component, unique regions are identified using BWA, and Primer3 is used to select primer pairs. All primers are further filtered by melting temperature (T_m) difference to the nearest neighbor background, within a user-specified value (5°C by default).

For primer validation, the primer binding location(s) and product sizes are reported for any submitted primers (output *N* in Supplementary Figure S1). For primer design, a full list of primers that uniquely amplify a product within the assembled contigs is reported (only five are displayed by default on the project page), along with information on

the nearest neighbor amplicon (output O in Supplementary Figure S1). Examples of output for both primer validation and primer design can be found in Supplementary Figure S11.

RESULTS

The EDGE bioinformatics overview

An overview of the EDGE Bioinformatics workflow is shown in Figure 1, with a more detailed workflow shown in Supplementary Figure S1. Because most sequencers can now output data as one or more FASTQ files (or are readily converted to FASTQ files) we opted for this format (full or compressed) as the required input for raw sequencing data. EDGE can use files derived from multiple libraries, runs or lanes by specifying the location of one or more FASTQ files or by retrieving them from the SRA (Supplementary Figure S2). EDGE was originally designed for use with raw Illumina® FASTQ data and performs best with these short sequence data types, but the development of alternative workflows are envisioned for future versions to better handle other types of data (e.g. longer reads, different error models, etc.). There are a number of additional options such as specifying number of CPUs to use, inputting multiple runs of the same sample, or allowing batch submission of many samples using the same modules and parameters.

Optional inputs depend on the selected modules (see Materials and Methods) and can include an adapter FASTA file for adapter filtering, a host FASTA file for removal of host reads, PacBio/Nanopore long read FASTA/FASTQ files for use with the SPAdes assembler, one or more reference genomes for comparative genomic analysis, and a primer pair(s) file in FASTA format for *in silico* primer validation. While there are several optional environmental parameters that can control the way EDGE runs, the users need only specify a project name, select the input file(s), toggle which modules they would like to use, and click Submit. The results of each project are displayed within its own project page (see Materials and Methods and Supplementary Figure S3). Descriptions of all modules are in the Methods section and in the online documentation.

Analysis in EDGE

To demonstrate the utility and versatility of EDGE, we tested this platform using a number of different samples that represent varied scenarios, including examples of isolate sequencing and analysis of several clinical metagenome samples with known, suspected, and unknown etiologic agents (Table 1). Not all results are described in depth, but the different datasets are used to highlight some of the various modules and analytic capabilities encompassed within the EDGE Bioinformatics platform. All datasets and project pages with full results are publicly available on our demonstration webserver. There, users can view or select and run their own analyses of these data or other publicly accessible SRA data.

Analysis of isolate genome sequencing projects

To highlight and validate some of the features and integration of utilities within EDGE, we tested the various modules using two datasets (sequenced at two different institutions) from recently completed isolate genome sequencing projects: *Bacillus anthracis* strain SK-102 (40) and *Yersinia pestis* strain Harbin 35 (41). After quality control, 96–98% of the reads were retained for *B. anthracis* and *Y. pestis* (Supplementary Figure S4). Results from the Assembly and Annotation module were consistent with known genome complexity (repeated elements such as insertion sequences and rRNA operons), genome size, and associated number of genes. The *B. anthracis* assembly was 5.5 Mb in size, consisting of 89 contigs with a maximum contig size of 450 kb and an average contig fold coverage of 328×, consistent with the amount of data sequenced (Supplementary Figure S5). The *Y. pestis* assembly (4.6 Mb with 306× fold coverage) was more fragmented (329 contigs) with smaller contig sizes (maximum contig size of 115 kb) owing to the large number of repeat sequences within the genome. However, using the reference-based analysis module, all of the *Y. pestis* contigs, and all but a single contig of the *B. anthracis* assembly, could be mapped to the selected reference genome (*Y. pestis* CO92 and *B. anthracis* Ames Ancestor, respectively). More than 98% of the reads of either sample could also be mapped, covering 97–100% of the reference chromosomes and plasmids (Supplementary Figure S7).

While the identities of the organisms sequenced in this case are not in question, the taxonomy classification module can be used to identify a contaminant, or otherwise suggest similarity to another taxon. The consensus for all the taxonomy classification tools encompassed in EDGE confirmed the presumed identities of the organisms sequenced. With *Y. pestis*, both GOTTCHA (25) and Metaphlan (32) provided the cleanest results, suggesting only *Y. pestis* reads comprise the dataset (Figure 2A), however with *B. anthracis*, a number of different organisms were found by these tools (Figure 2B), even at the genus level. At the species level, both GOTTCHA and Metaphlan identified *B. cereus* and *Francisella philomiragia* in addition to the dominant *B. anthracis*. In addition, GOTTCHA found signatures of *Y. pestis* and *B. weihenstephanensis*, while Metaphlan suggested *B. thuringiensis* was present. Upon further investigation, we discovered that the *B. anthracis* SK-102 sample was sequenced within the same Illumina lane as many other samples, including *F. philomiragia* ATCC25018, two *Y. pestis* strains (771 and 790), *B. cereus* BACI291, *B. mycoides* BACI084 (a near neighbor to *B. weihenstephanensis* (42)), and several fecal samples from Condors (found to contain dominant amounts of *Clostridia* sequences, consistent with dominance of *Clostridia* in the Vulture hindgut (43)). Therefore, these additional identifications are likely the result of index cross contamination (or other mis-assignment) of barcodes to sample, often found among samples run within the same lane (44). In addition, and consistent with the bacteria in this sample, GOTTCHA viral analysis suggested three *Bacillus* phages as well as *Staphylococcus* phage SpaA1, which is similar to *Bacillus* prophages and can infect *Bacillus* spp. (45).

Phylogenetic analysis was performed for each dataset, selecting all available NCBI RefSeq genomes for either *Y. pestis*, or for *B. anthracis*, *B. cereus*, and *B. thuringiensis*. This phylogenetic module, based on PhaME, independently treats the input reads and resulting contigs (when assem-

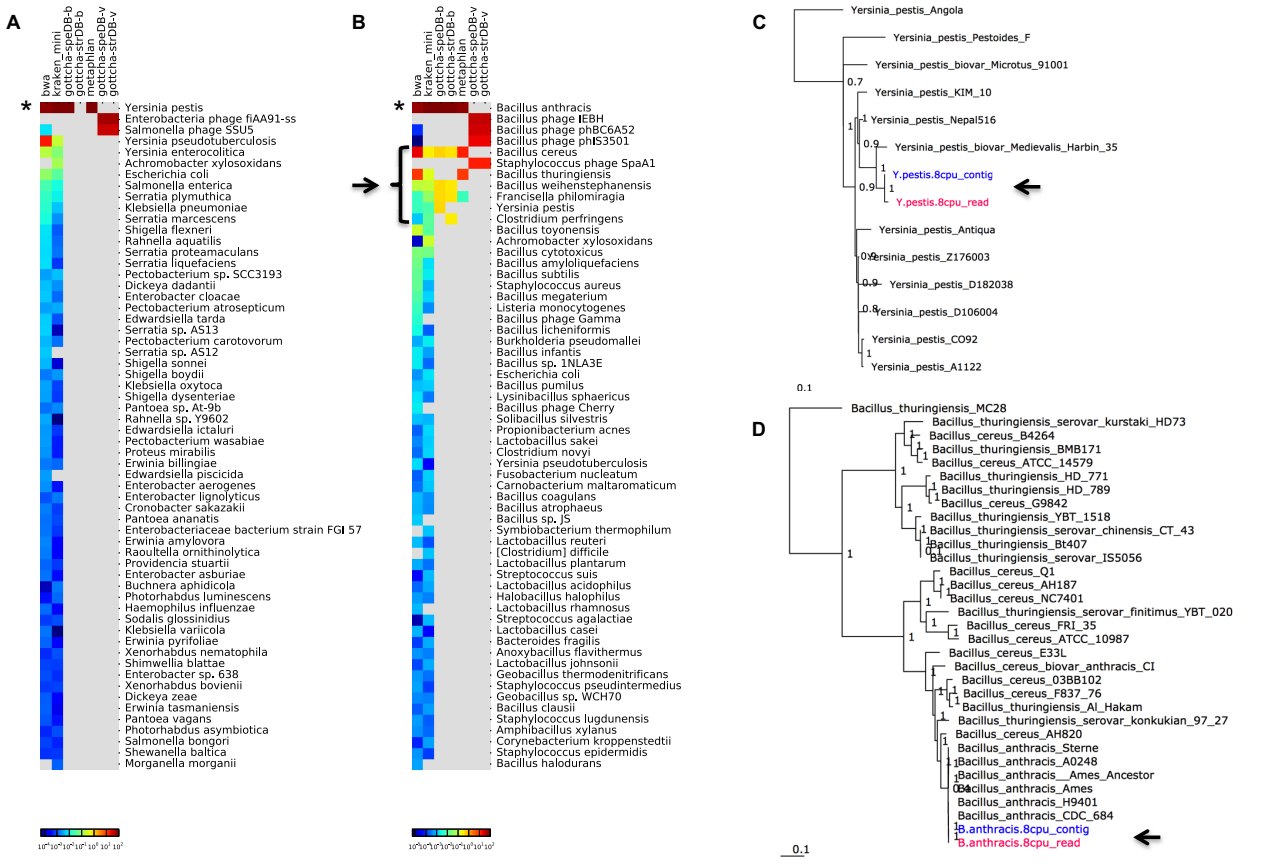


Figure 2. Taxonomy and phylogenetic evaluations of bacterial isolates. Panels **A** and **B** show taxonomic classification of reads for (A) the *Y. pestis* Harbin35 sample and (B) the *B. anthracis* SK-102 sample. The stars indicate the consistent dominant taxonomic calls for all tools, while the black arrow and bracket indicate identified contamination in the *B. anthracis* sample. Panels **C** and **D** indicate the inferred phylogenetic trees for the (C) *Y. pestis* and (D) *B. anthracis*; black arrows point to the read dataset (pink) and contigs (blue) that were placed in these trees.

bly is selected) for whole genome SNP analysis, and consistently placed the datasets within their respective phylogenetic trees (Figure 2C and D). The *Y. pestis* tree was inferred from a 4.0 Mb core genome with 2077 SNPs and the *Y. pestis* sample was placed nearest a previously sequenced *Y. pestis* Harbin35. The *Bacillus* tree was based on a core genome of 3.1 Mb with 384 568 SNPs, is fully consistent with known *Bacillus* relationships (42), and placed the reads and the resulting contigs of the *B. anthracis* SK-102 closest to *B. anthracis* CDC684.

Using the PCR Primer Tools module, published primers that have been used to detect either *Y. pestis* (46,47) or *B. anthracis* (48,49) were input for validation against these isolates and confirmed the appropriate amplicon sizes using electronic PCR against the respective assemblies. For *B. anthracis*, the primer design software suggested two PCR primer pairs that would specifically amplify only this strain compared with all other NCBI genomes (Supplementary Figure S11).

Analysis of a mock human microbiome sample of known complexity

The Human Microbiome Project's (HMP) staggered mock community (50) was used to evaluate the metagenome analysis potential of EDGE. This dataset, consisting of sequencing reads derived from a mixture of 21 known bacterial strains and one eukaryotic strain, was analyzed using the Pre-processing, Assembly, and Taxonomy classification modules with default parameters. The FaQCs (23) quality control pipeline retained 81.2% of the reads and 76.7% of the data from the 7.9M read dataset, while the subsequent assembly produced 13 097 contigs totaling 14.8 Mb. Read mapping validation suggested that the assembly represents 77.6% of the reads with a contig average fold coverage of 24× (Supplementary Figure S6). Both the read- (Figure 3A), and contig-based (Figure 3B) taxonomy classification tools accurately identified most of the known community members of this sample with the exception of the eukaryote since these tools are currently implemented with the objective of identifying bacteria, archaea, and viruses only. The contig plot of average G+C (%) versus average fold coverage can also help distinguish groups of contigs that belong

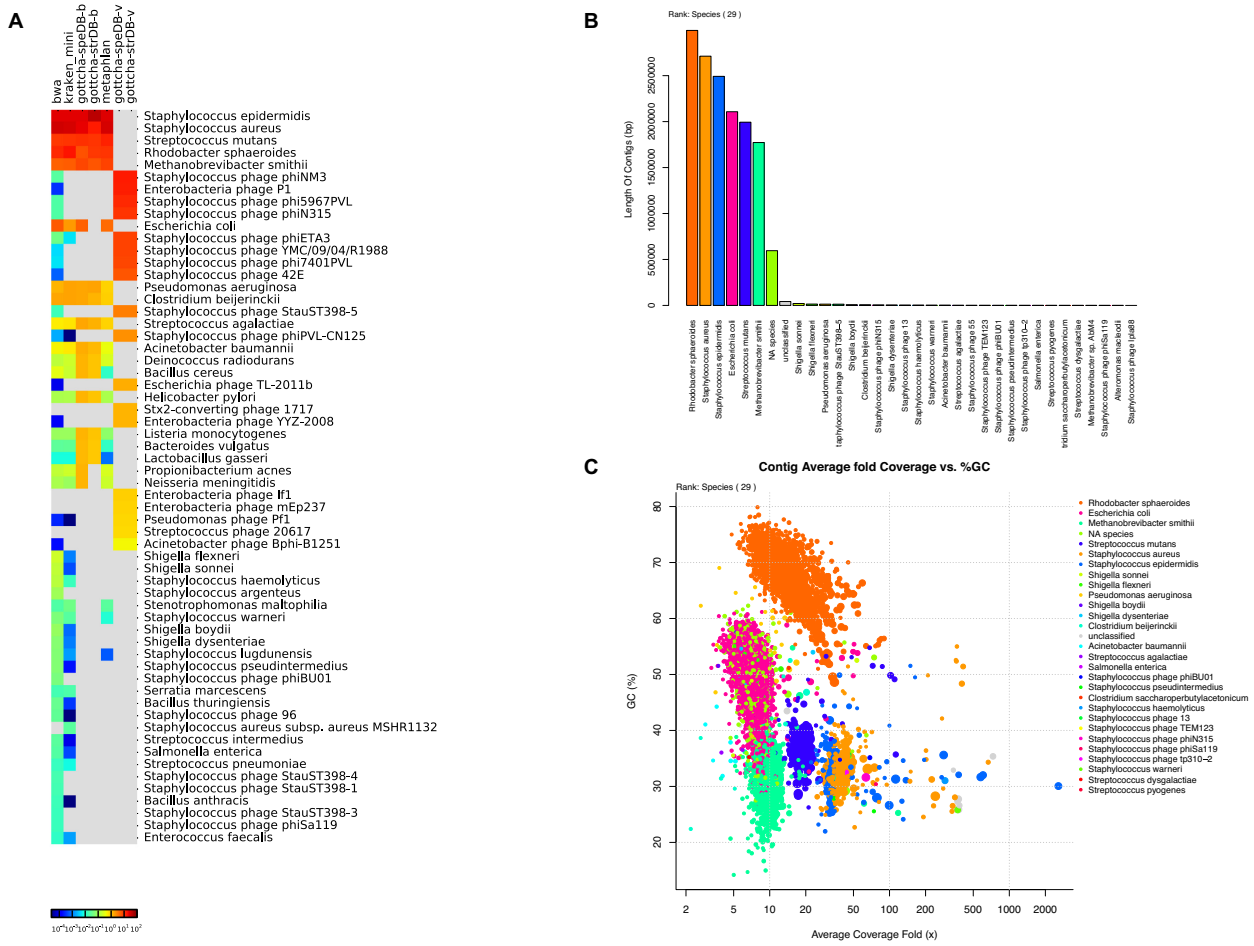


Figure 3. Taxonomic Classification of the HMP staggered mock sample. (**A**) Read-based classification using various taxonomy profiling tools; (**B**) contig-based classification displaying length of all classified contigs per taxon and (**C**) a scatterplot of contig % GC versus fold coverage of the contigs, colored by taxon.

to different organisms (Figure 3C). Similar graphics and results can be found at various taxonomic levels.

Analysis of complex clinical samples

We also used EDGE to evaluate datasets from several clinical samples with suspected pathogens. In the first example, we used EDGE to characterize one of the recent 2014 Ebola outbreak samples. Using the Sierra Leone human plasma RNA sequencing sample SRR1553609 retrieved directly from the SRA, we ran all EDGE modules with the exception of phylogenetic and primer analyses. Preprocessing removed ∼25% of the data, and human host removal only identified 605 reads that matched the human reference. IDBA (26) assembly of the remaining reads resulted in 1588 contigs, a total assembly size of 665 kb and a largest contig of 14.6 kb. Due to the complexity of the sample, only 15% of the data assembled. We examined the use of the alternate assembler, SPAdes (27), with this sample and found an increased run time (Table 1) balanced by an improved 36% read incorporation (versus 15%) into the assembly, resulting in 12 105 contigs, a total assembly size of >3.8 Mb and a largest contig of 18.6 kb. Using as reference the *Homo sapiens*-wt/GIN/2014/Makona-Gueckedou-633 Zaire ebolavirus (a sequence from Guinea, 2014), we found that only 3228 reads (0.43% of the input reads) could be mapped to the genome, covering 98.9% of the length with 10 potential single nucleotide variants. Two of the IDBA contigs overlapped and together covered 99.2% of the genome, while a single SPAdes contig covered 97.8% of the reference. Both assemblies identified the same 8 SNPs with respect to the reference genome. The genome browser in EDGE helped resolve the disparate variant analysis found between the reads and the contigs (Figure 4). While almost all of the reads confirmed all eight SNPs found within the contigs, the two additional variants identified with read-based analysis likely reflected the quasispecies nature of the virus, with strong support but fewer than 50% of the reads at those positions carrying the additional point mutations. This shows the utility of a multi-pronged approach when performing such comparisons. The

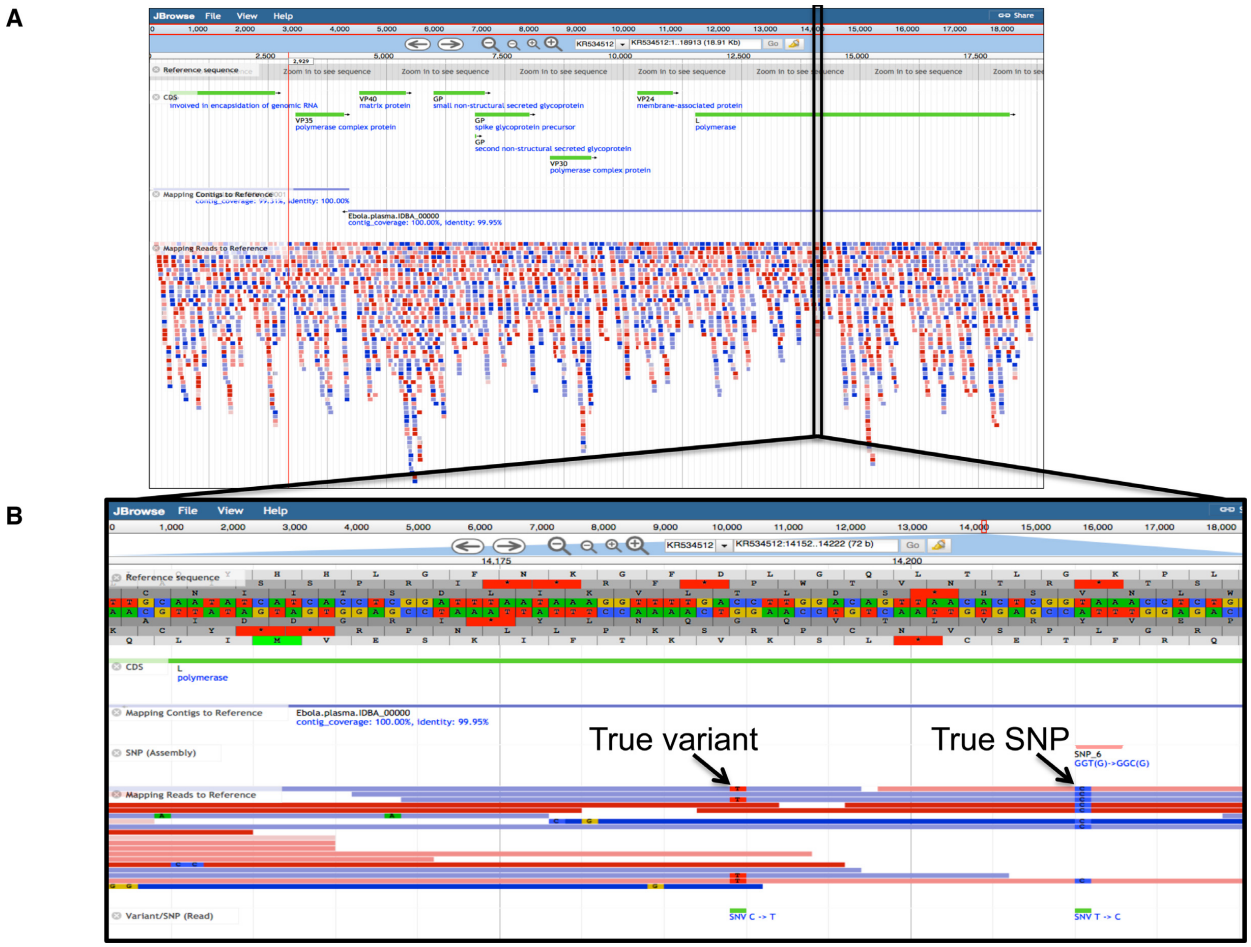


Figure 4. Interactive genome browsing view of a reference-based analysis in EDGE with a human clinical sample containing Ebolavirus. (**A**) An Ebola reference genome and its genes (green lines) are displayed together with contig-based (using IDBA) and read-based comparisons. The two contigs (blue lines) from IDBA are shown aligned along the length of the reference as well as the reads (red and blue). (**B**) A zoomed-in view of one section of the genome where SNPs were identified. The SNP and coding difference is outlined under the contig alignment, while the variants are indicated under the read alignments.

taxonomy classification module showed that Ebola could indeed be found within the reads, though only with the GOTTCHA and BWA pipelines. Unexpectedly, a number of bacteria were also identified as present within the sequenced sample including *Ralstonia*, *Bradyrhizobium*, *Propionibacterium* and *Pseudomonas* (Supplementary Figure S8). It is unknown whether these bacterial organisms were actually present within the patient or alternatively their nucleic acids were introduced via laboratory reagents (51) or were sample carryover from a prior sequencing run. However, some of the detected bacteria such as *Propionibacterium*, a common skin inhabitant, or *Ralstonia* have been shown before to be present in human blood (52,53). The contig-based taxonomy analyses also clearly showed Ebola virus to be present, and confirmed that many contigs belonged to the same bacterial groups identified by read-based analyses.

In the second clinical example, we analyzed data derived from a fecal sample of a patient returning from Germany during the 2011 enterohemorrhagic *Escherichia coli* outbreak, and who was suspected of harboring *E. coli* O104:H4. Trimming and filtering removed 13.3% of the bases while host removal identified only 0.15% of the reads as human and 0.02% as PhiX (a spike-in control commonly used in Illumina sequencing). Assembling the remaining 253M reads resulted in 2957 contigs totaling 10.5 Mb, comprising 23.9% of the reads. The single chromosome and three plasmids of *E. coli* O104:H4 2011C-3493 were used as reference for both read- and contig-based comparisons. Using reads, 99.99% of the reference chromosome was covered at 115×, while the three plasmids were covered 100% at fold-coverages ranging from 250× for the largest plasmid to 7.6 million fold coverage for the smallest plasmid. Using contigs, all replicons were covered >99.7% with the exception of the small plasmid which was absent from the assembly (this absence is likely due to the excessive fold coverage known to create assembly issues). All taxonomy profiling tools clearly showed that *E. coli* (or *Shigella*) was the dominant organism

and that the Shiga-toxin phage was also present (Supplementary Figure S9). Whole genome SNPs were identified and phylogenetic analysis was performed with both reads and contigs, easily done within EDGE using the drop down menu to select 68 *E. coli* and *Shigella* genomes. Both the predominantly *E. coli* metagenome reads and the assembled contigs were placed within the same clade as the other *E. coli* O104 strains, reaffirming the initial suspicion of *E. coli* O104:H4 as the etiologic agent (Figure 5A).

A nasal swab sample from a patient with acute respiratory illness of unknown etiology was used as a final test of EDGE's utility for analysis of clinically derived metagenomic datasets. In this case, while >99% of the data passed FaQCs quality control, the majority of sequence reads (78.9%) were human-derived and removed (data not shown). The remaining reads were submitted to SRA and used for assembly and taxonomy classification. A number of expected organisms (54,55) ranked among the most abundant genera identified, including *Prevotella, Veillonella* and *Streptococcus*. Unexpectedly, *E. coli* was identified by GOTTCHA, and also detected (at a substantially lower level) by BWA and Kraken mini (Figure 5B). Upon closer inspection, the mapping results demonstrated that all of the *E. coli* hits were to the plasmid (with no matches to the chromosome) in *E. coli* strain ABU83972, covering ∼80% of this replicon. Interestingly, this plasmid is very similar (>90% identity) to a number of enteric plasmids, as well as to the *Corynebacterium renale* plasmid pCR1, suggesting that the presence of this plasmid might be the result of colonization or infection by a *Cornyebacterium* species, which are common in nasal cavities (55). This hypothesis is partially supported by BWA and Kraken, which identified a different *Cornyebacterium* at low levels, as well as by 16S sequence data in which *E. coli* is not detected but the genus *Cornyebacterium* is found (Supplementary Table S1). As a result of these findings a new feature now present in EDGE separates plasmid from chromosomal hits for GOTTCHA, thereby allowing for greater specificity in evaluating taxonomic profiling results (Figure 5C). The differences in bacterial species found by Metaphlan compared with all other tools can be explained by the additional draft genome references included within the Metaphlan database (32), and which are not yet available in RefSeq.

DISCUSSION

As the number of investigations that apply sequencing continues to climb, the wider genomics community will greatly benefit from a user-friendly bioinformatics environment of integrated tools and pipelines designed to address a large number of scenarios and scientific end-goals. The initial system and the tools we developed and used in EDGE are available as open source software, and we encourage other developers to contribute best-practice tools and pipelines, as there are yet a number of use cases not addressed within this initial platform. For the tools in current use, the focus was on accuracy, speed, flexibility and ability to run within a modest computational environment for analysis of individual microbial samples (isolates or metagenomes). In some cases, like with read-based taxonomy profiling, given that this is a still emerging field of exploration, we provide a suite of tools based on different algorithms, and present a comparative view of the results for further scrutiny by researchers. In other cases, tools were selected that perform well under a diverse set of circumstances, and are computationally friendly with respect to speed and memory considerations. While novel tools continue to be developed and databases continue to grow, future focus will be on the systematic incorporation of better tools and updating of databases alongside the development of new modules and new visualizations.

Collectively, our results and experiences suggest that EDGE provides significant advantages over the current status quo. EDGE assists non-expert users by providing predefined pipelines to run cutting-edge tools and a web interface that makes inspection of results quick and easy through a series of interactive visualizations provided within a single user-friendly interface. Comparative views of results output by complex metagenome taxonomy profiling tools distinguish this system from all others along with the ability to easily perform whole genome SNP phylogenies with user-selected genomes. The ability to integrate read-based with assembly-based analyses is natively provided in EDGE and affords complimentary views of genomic data. While analysis times differ depending on the amount of data input, the computational hardware available, the modules selected, and the complexity of the sample, EDGE was designed to provide rapid analysis of NGS data. As shown with the examples in this manuscript, run on our publicly available server, individual isolate or metagenome projects generally complete within hours, even when selecting all analysis modules. Very large and complex datasets will invariably take longer, however real-time tracking of projects and system resources allows for monitoring progress and job queuing. With embedded log files detailing the specifics of each run, a wide adoption of systems like EDGE can also provide a form of standardized data analysis which would allow for more robust comparisons to be made across different independent projects and laboratories.

EDGE is a unique bioinformatic software package both for the variety of open-source tools that are encompassed, for its ease of use, and for the integration of all analysis results for the sample within a single web page. We selected specific isolate and metagenome examples to present within this manuscript to highlight the versatility of the EDGE platform, including quality assessment and trimming, assembly and annotation, reference-based comparisons, taxonomy classification, phylogenetic analysis, and PCR primer analysis. To our knowledge, there is no other freely available bioinformatic software package that incorporates these types of analyses and tools within a sample-centric framework of intuitive pipelines and interactive graphical and tabular results. Because EDGE can be installed locally, all analyses and raw sequencing data can be kept entirely private. This software package is designed to enable scientists with limited experience in bioinformatics to perform a variety of genomic analyses on microbial isolates or metagenomes, with resources that can be housed in smaller laboratories rather than requiring extensive computational and personnel infrastructure. Therefore, we believe the EDGE Bioinformatics software represents a critical step forward in democratizing genomics analyses.

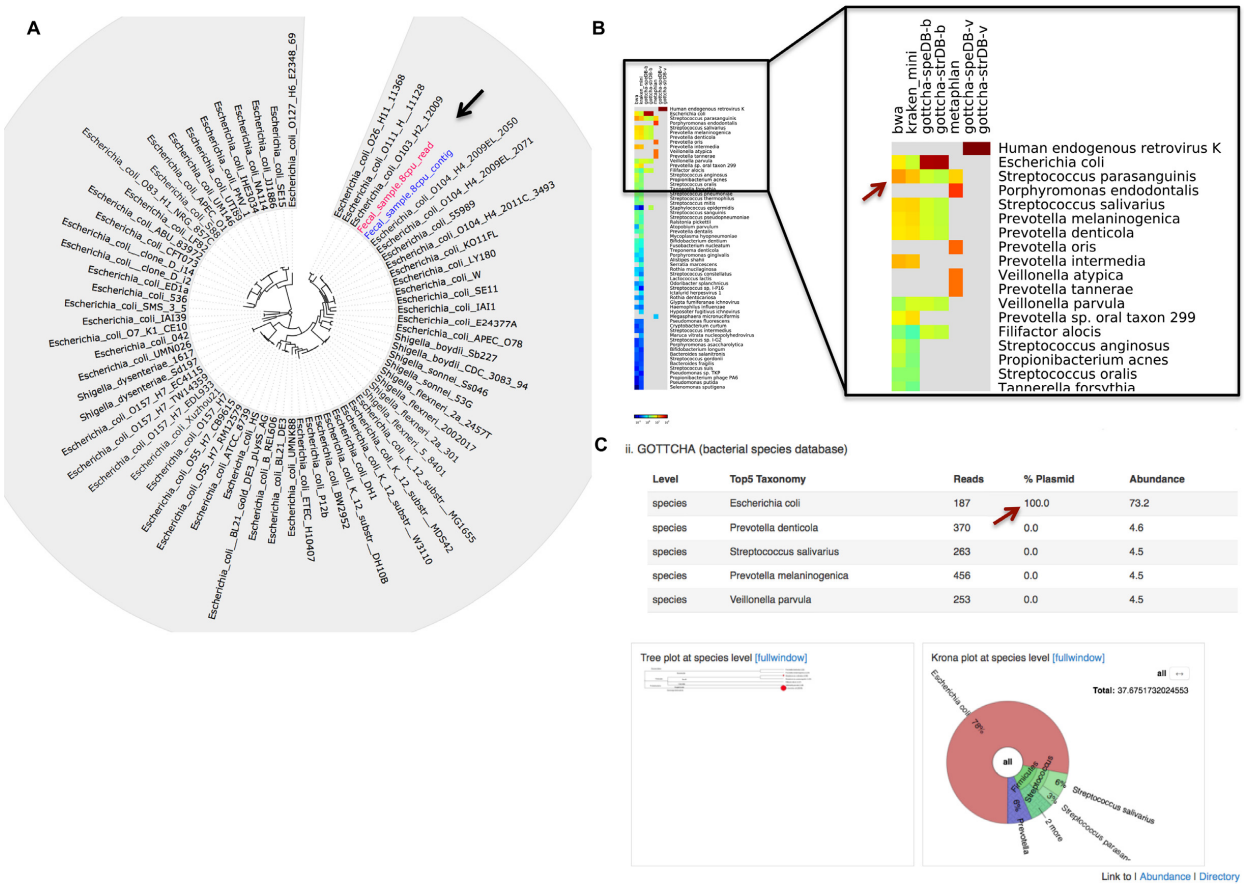


Figure 5. Phylogenetic and taxonomic analysis of human clinical samples with suspected and unknown causative agents. (**A**) Circular phylogenetic tree clearly places within the *E. coli* O104 group both the raw reads and the contigs obtained from a clinical fecal sample. (**B**) A comparative heatmap view of identified taxa from a nasal swab sample demonstrates the abundance of typical nasal cavity organisms. (**C**) The *E. coli* identified with GOTTCHA in the nasal swab sample (in B) is described in greater detail under the tool-specific EDGE view (red arrow), showing the percent of hits to plasmids for each identified taxon; below are a taxonomic dendrogram featuring the taxa detected with circles representing relative abundance, and a Krona plot view of the same data.

AVAILABILITY

The software is freely available (https://lanl-bioinformatics.github.io/EDGE/) and a demonstration webserver is provided (https://bioedge.lanl.gov/) for use with the data from this manuscript and any publicly available data via the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) or European Molecular Biology Laboratory European Nucleotide Archive (EMBL ENA).

ACCESSION NUMBERS

Accession numbers for all data can be found in Table 1.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank our beta test users for their valuable feedback. Many thanks to the LANL Genome Programs group, and in particular to the informatics support team at LANL for their great help and feedback with both sequencing and bioinformatics. We thank Jason Gans for his careful reading of the manuscript and his helpful suggestions. We also thank Gerald Quinnan, Pengfei Zhang, Regina Cer, Cassie Redden, Kenneth Frey, Eugene Millar and the IDCRP who were involved in production of nasal swab sequence data (metagenome and 16S). The views expressed in this manuscript are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, the Department of Defense, the National Institutes of Health, the Department of Health and Human Services, nor the U.S. Government. VPM is a military service member of the U.S. Government. This work was prepared as part of his official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties.

FUNDING

Defense Threat Reduction Agency [CB4026 to Naval Medical Research Center]; Defense Threat Reduction Agency [CB10152 to Los Alamos National Laboratory]. Funding for open access charge: Defense Threat Reduction Agency [CB10152].

Conflict of interest statement. None declared.

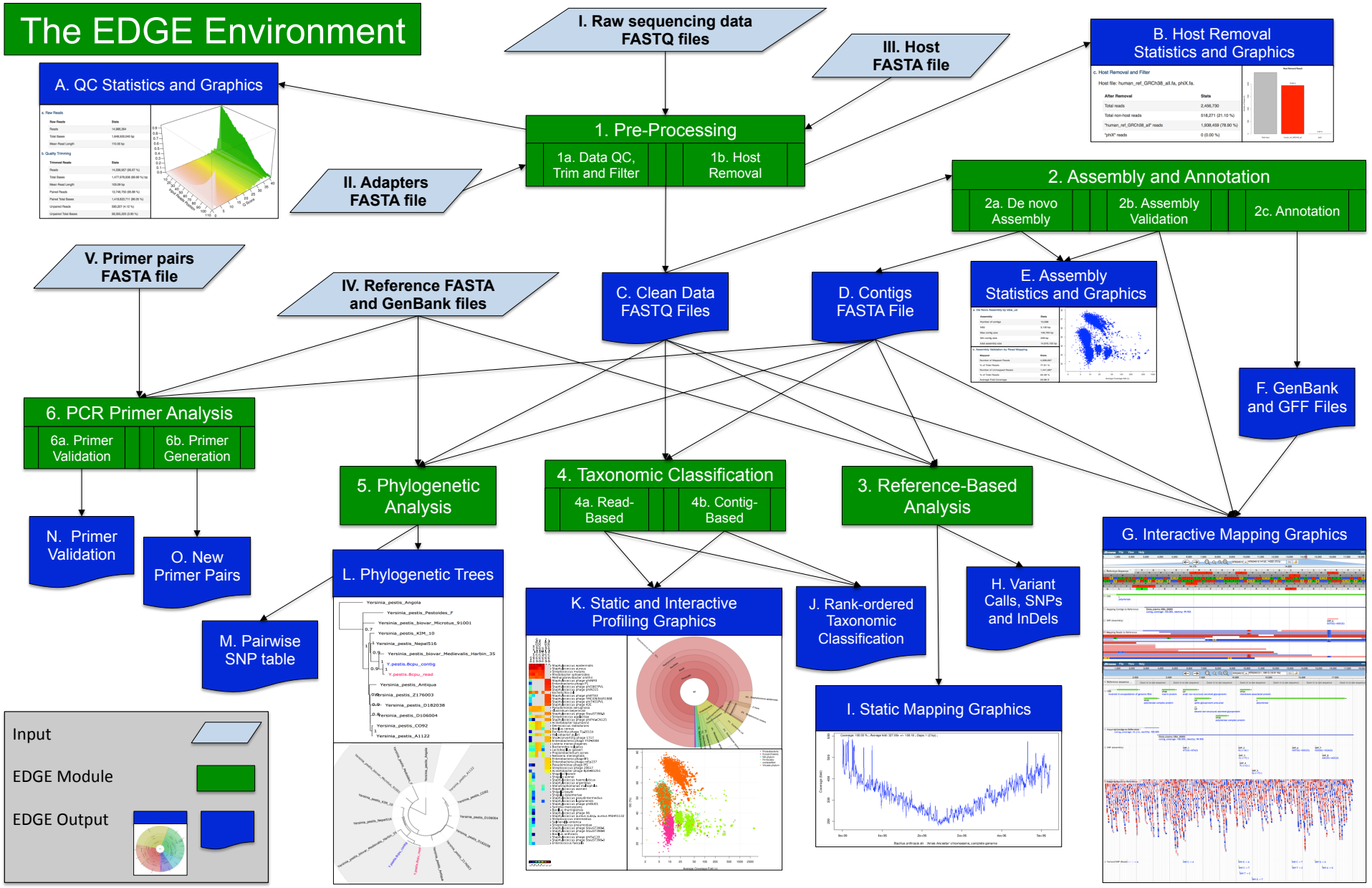
REFERENCES

1. Buermans,H.P. and den Dunnen,J.T. (2014) Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta*, **1842**, 1932–1941.
2. Conlan,S., Thomas,P.J., Deming,C., Park,M., Lau,A.F., Dekker,J.P., Snitkin,E.S., Clark,T.A., Luong,K., Song,Y. *et al.* (2014) Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing Enterobacteriaceae. *Sci. Transl. Med.*, **6**, 254ra126.
3. den Bakker,H.C., Allard,M.W., Bopp,D., Brown,E.W., Fontana,J., Iqbal,Z., Kinney,A., Limberger,R., Musser,K.A., Shudt,M. *et al.* (2014) Rapid whole-genome sequencing for surveillance of Salmonella enterica serovar enteritidis. *Emerg. Infect. Dis.*, **20**, 1306–1314.
4. Wohlbach,D.J., Rovinskiy,N., Lewis,J.A., Sardi,M., Schackwitz,W.S., Martin,J.A., Deshpande,S., Daum,C.G., Lipzen,A., Sato,T.K. *et al.* (2014) Comparative genomics of Saccharomyces cerevisiae natural isolates for bioenergy production. *Genome Biol. Evol.*, **6**, 2557–2566.
5. Wang,J., Chen,L., Huang,S., Liu,J., Ren,X., Tian,X., Qiao,J. and Zhang,W. (2012) RNA-seq based identification and mutant validation of gene targets related to ethanol resistance in cyanobacterial Synechocystis sp. PCC 6803. *Biotechnol. Biofuels*, **5**, 89.
6. Koren,S., Treangen,T.J., Hill,C.M., Pop,M. and Phillippy,A.M. (2014) Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics*, **15**, 126.
7. Watson-Haigh,N.S., Shang,C.A., Haimel,M., Kostadima,M., Loos,R., Deshpande,N., Duesing,K., Li,X., McGrath,A., McWilliam,S. *et al.* (2013) Next-generation sequencing: a challenge to meet the increasing demand for training workshops in Australia. *Brief. Bioinformatics*, **14**, 563–574.
8. Daber,R., Sukhadia,S. and Morrissette,J.J. (2013) Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. *Cancer Genet.*, **206**, 441–448.
9. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.J., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
10. Maxam,A.M. and Gilbert,W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 560–564.
11. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5463–5467.
12. Bennett,S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433–438.
13. Blankenberg,D., Von Kuster,G., Coraor,N., Ananda,G., Lazarus,R., Mangan,M., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a web-based genome analysis tool for experimentalists. In: Frederick,MA (ed). *Current Protocols in Molecular Biology*. doi:10.1002/0471142727.mb1910s89.
14. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D., Kenyon,R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
15. Reich,M., Liefeld,T., Gould,J., Lerner,J., Tamayo,P. and Mesirov,J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.
16. Chen,I.M., Markowitz,V.M., Palaniappan,K., Szeto,E., Chu,K., Huang,J., Ratner,A., Pillay,M., Hadjithomas,M., Huntemann,M. *et al.* (2016) Supporting community annotation and user collaboration in the integrated microbial genomes (IMG) system. *BMC Genomics*, **17**, 307.
17. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M. *et al.* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
18. Markowitz,V.M., Chen,I.M., Chu,K., Szeto,E., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Pagani,I., Tringe,S. *et al.* (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.*, **42**, D568–D573.
19. Keegan,K.P., Glass,E.M. and Meyer,F. (2016) MG-RAST, a Metagenomics Service for analysis of microbial Community Structure and Function. *Methods Mol. Biol.*, **1399**, 207–233.
20. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
21. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
22. Chen,P.E., Cook,C., Stewart,A.C., Nagarajan,N., Sommer,D.D., Pop,M., Thomason,B., Thomason,M.P., Lentz,S., Nolan,N. *et al.* (2010) Genomic characterization of the Yersinia genus. *Genome Biol.*, **11**, R1.
23. Lo,C.C. and Chain,P.S. (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics*, **15**, 366.
24. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
25. Freitas,T.A., Li,P.E., Scholz,M.B. and Chain,P.S. (2015) Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.*, **43**, e69.
26. Peng,Y., Leung,H.C., Yiu,S.M. and Chin,F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.
27. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
28. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
29. Smits,S.A. and Ouverney,C.C. (2010) jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One*, **5**, e12267.
30. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
31. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
32. Segata,N., Waldron,L., Ballarini,A., Narasimhan,V., Jousson,O. and Huttenhower,C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–814.
33. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
34. Fouts,D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.*, **34**, 5839–5851.
35. Untergasser,A., Cutcutache,I., Koressaar,T., Ye,J., Faircloth,B.C., Remm,M. and Rozen,S.G. (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.
36. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
37. Otto,T.D., Dillon,G.P., Degrave,W.S. and Berriman,M. (2011) RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.*, **39**, e57.
38. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
39. Scott,L.J., Muglia,P., Kong,X.Q., Guan,W., Flickinger,M., Upmanyu,R., Tozzi,F., Li,J.Z., Burmeister,M., Absher,D. *et al.* (2009) Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 7501–7506.
40. Johnson,S.L., Daligault,H.E., Davenport,K.W., Jaissle,J., Frey,K.G., Ladner,J.T., Broomall,S.M., Bishop-Lilly,K.A., Bruce,D.C., Gibbons,H.S. *et al.* (2015) Complete genome sequences for 35

biothreat assay-relevant bacillus species. *Genome Announcements*, **3**, doi:10.1128/genomeA.00151-15.
41. Johnson,S.L., Daligault,H.E., Davenport,K.W., Jaissle,J., Frey,K.G., Ladner,J.T., Broomall,S.M., Bishop-Lilly,K.A., Bruce,D.C., Coyne,S.R. *et al.* (2015) Thirty-two complete genome assemblies of nine yersinia species, including Y. pestis, Y. pseudotuberculosis, and Y. enterocolitica. *Genome Announcements*, **3**, doi:10.1128/genomeA.00148-15.
42. Soufiane,B. and Cote,J.C. (2013) Bacillus weihenstephanensis characteristics are present in Bacillus cereus and Bacillus mycoides strains. *FEMS Microbiol. Lett.*, **341**, 127–137.
43. Roggenbuck,M., Baerholm Schnell,I., Blom,N., Baelum,J., Bertelsen,M.F., Ponten,T.S., Sorensen,S.J., Gilbert,M.T., Graves,G.R. and Hansen,L.H. (2014) The microbiome of New World vultures. *Nat. Commun.*, **5**, 5498.
44. Kircher,M., Sawyer,S. and Meyer,M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.*, **40**, e3.
45. Swanson,M.M., Reavy,B., Makarova,K.S., Cock,P.J., Hopkins,D.W., Torrance,L., Koonin,E.V. and Taliansky,M. (2012) Novel bacteriophages containing a genome of another bacteriophage within their genomes. *PLoS One*, **7**, e40683.
46. Hinnebusch,J. and Schwan,T.G. (1993) New method for plague surveillance using polymerase chain reaction to detect Yersinia pestis in fleas. *J. Clin. Microbiol.*, **31**, 1511–1514.
47. Begier,E.M., Asiki,G., Anywaine,Z., Yockey,B., Schriefer,M.E., Aleti,P., Ogden-Odoi,A., Staples,J.E., Sexton,C., Bearden,S.W. *et al.* (2006) Pneumonic plague cluster, Uganda, 2004. *Emerg. Infect. Dis.*, **12**, 460–467.
48. Francy,D.S., Bushon,R.N., Grady,A.M.G., Bertke,E.E., Kephart,C.M., Likirdopulos,C.A., Mailot,B.E., Schaefer,F.W. III and Lindquist,H.D.A. (2009). U.S. Department of the Interior, U.S. Geological Survey.
49. Fasanella,A., Losito,S., Adone,R., Ciuchini,F., Trotta,T., Altamura,S.A., Chiocco,D. and Ippolito,G. (2003) PCR assay to detect Bacillus anthracis spores in heat-treated specimens. *J. Clin. Microbiol.*, **41**, 896–899.
50. Consortium, H.M.P. (2012) A framework for human microbiome research. *Nature*, **486**, 215–221.
51. Salter,S.J., Cox,M.J., Turek,E.M., Calus,S.T., Cookson,W.O., Moffatt,M.F., Turner,P., Parkhill,J., Loman,N.J. and Walker,A.W. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.*, **12**, 87.
52. Grumaz,S., Stevens,P., Grumaz,C., Decker,S.O., Weigand,M.A., Hofer,S., Brenner,T., von Haeseler,A. and Sohn,K. (2016) Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med.*, **8**, 73.
53. Stelzmueller,I., Biebl,M., Wiesmayr,S., Eller,M., Hoeller,E., Fille,M., Weiss,G., Lass-Floerl,C. and Bonatti,H. (2006) Ralstonia pickettii-innocent bystander or a potential threat? *Clin. Microbiol. Infect.*, **12**, 99–101.
54. Rawlings,B.A., Higgins,T.S. and Han,J.K. (2013) Bacterial pathogens in the nasopharynx, nasal cavity, and osteomeatal complex during wellness and viral infection. *Am. J. Rhinol. Allergy*, **27**, 39–42.
55. Bassis,C.M., Erb-Downward,J.R., Dickson,R.P., Freeman,C.M., Schmidt,T.M., Young,V.B., Beck,J.M., Curtis,J.L. and Huffnagle,G.B. (2015) Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. *mBio*, **6**, e00037.

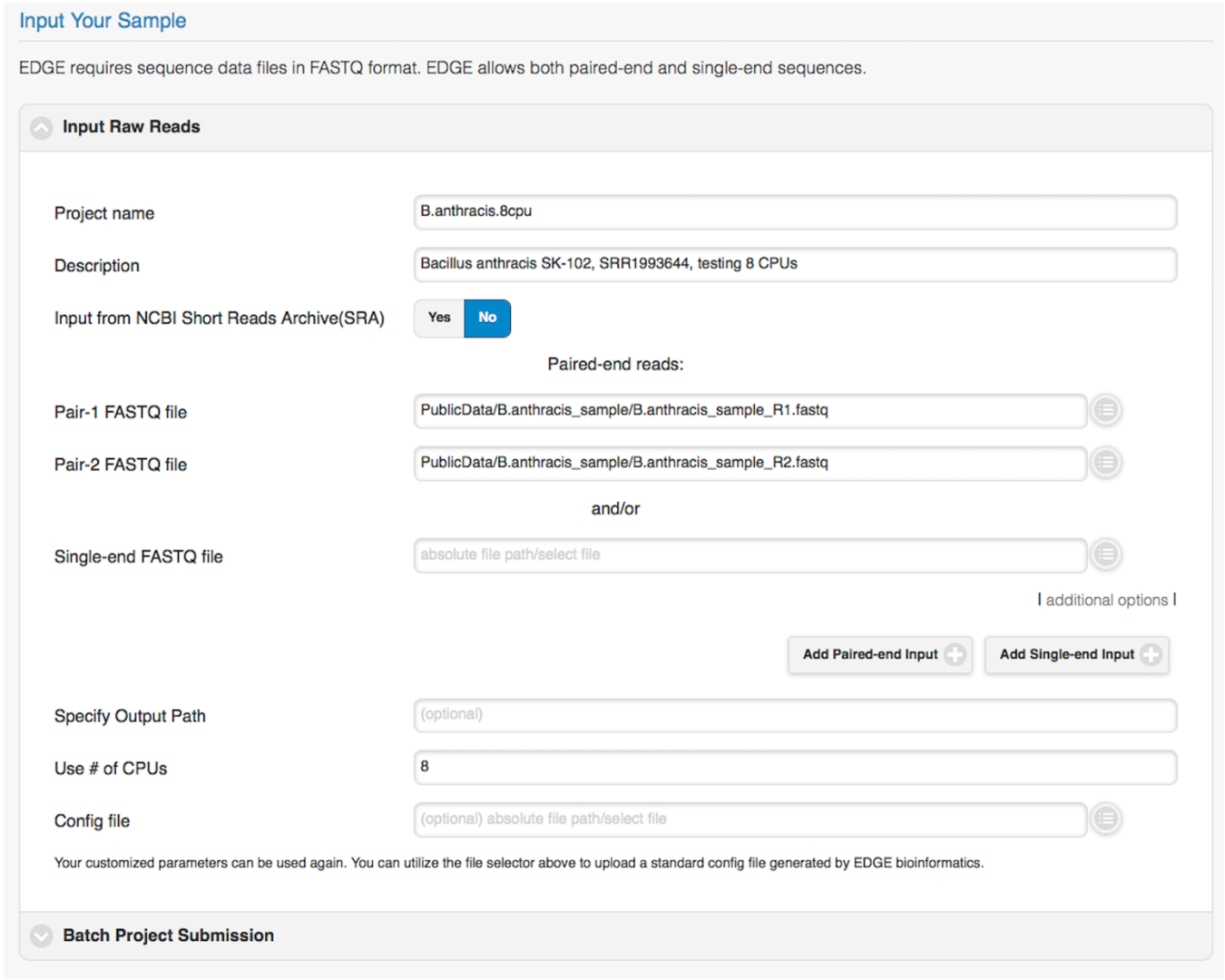
Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform

Po-E Li*[1], Chien-Chi Lo*[1], Joseph J. Anderson[2,3], Karen W. Davenport[1], Kimberly A. Bishop-Lilly[3,4], Yan Xu[1], Sanaa Ahmed[1], Shihai Feng[1], Vishwesh P. Mokashi[3], and Patrick S. G. Chain[1]

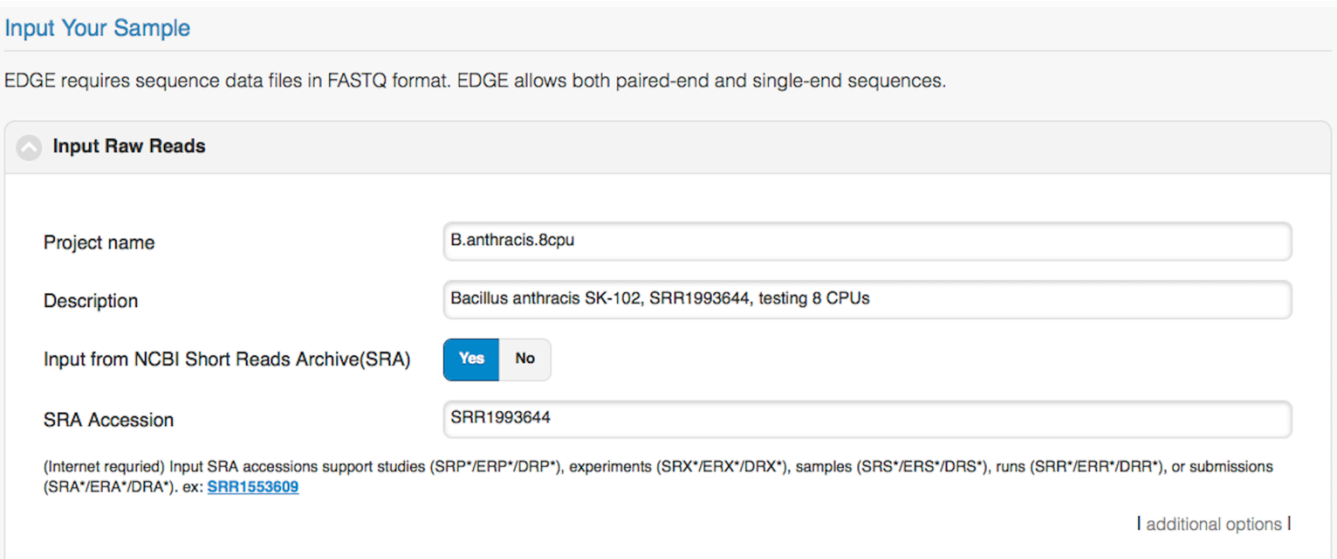


Supplementary Figure S1. A detailed overview of the EDGE Bioinformatics environment. Inputs from the user are given successive Roman numerals (I-V) and are shown in light blue. All EDGE process modules are shown in green and are numbered sequentially (1-6). Examples of resulting tabular and graphic outputs, shown in blue and lettered (A-O), are representative of EDGE outputs, but are not comprehensive. The user can create specific workflows with any combination of the modules. In addition, tailored parameters dictating how each module functions can be modified by the user.

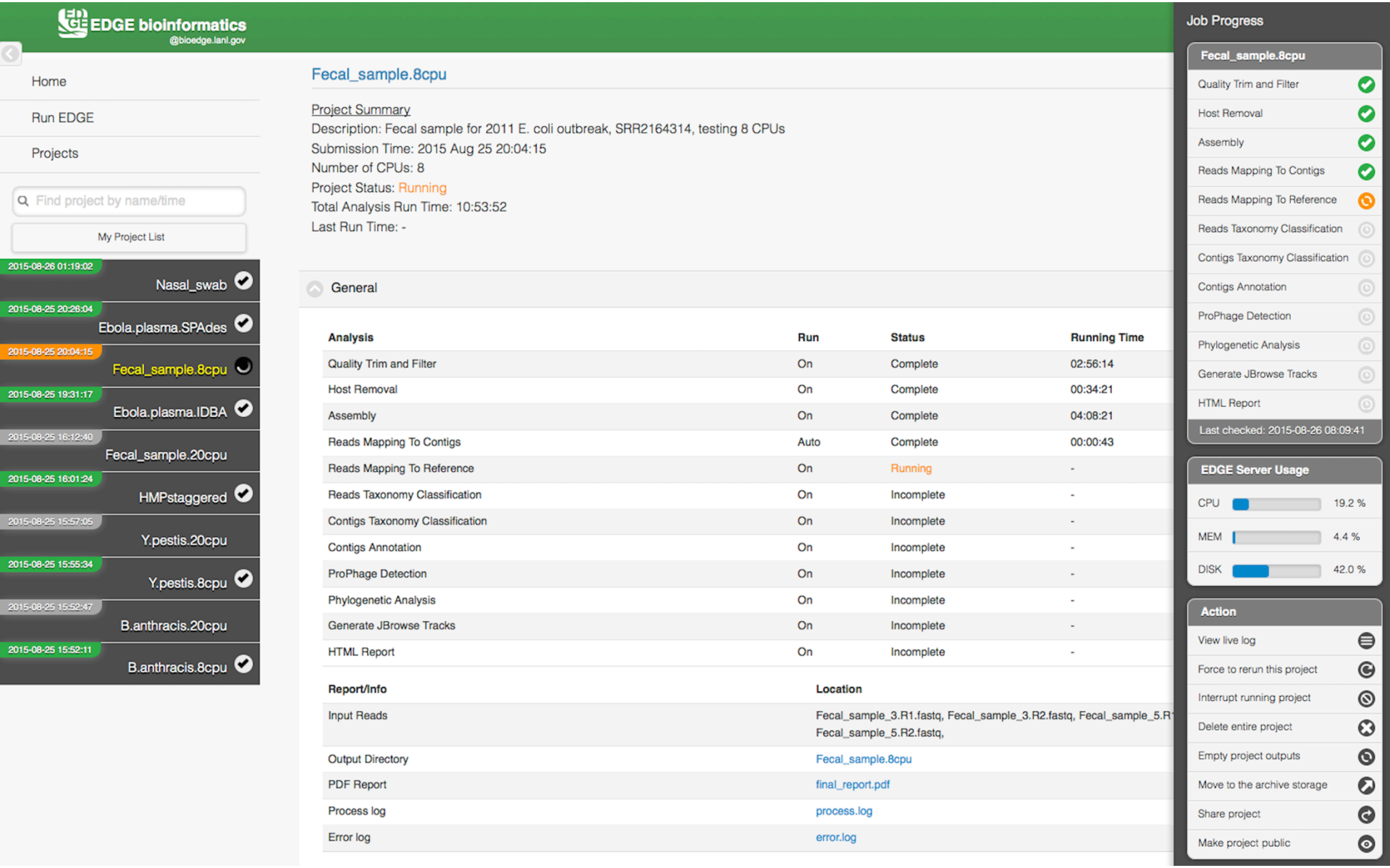
a



b

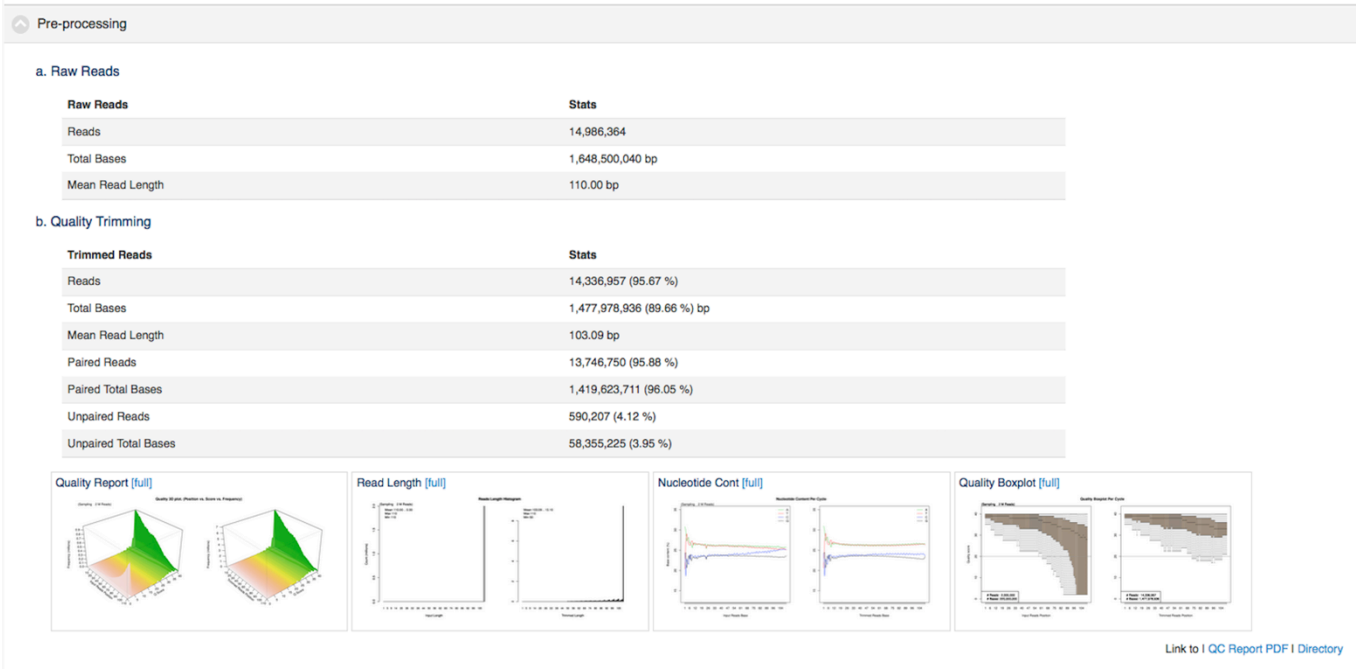


Supplementary Figure S2. Inputting data into EDGE. a) Data input can be achieved by using the absolute path to a local file. **b**) Alternatively, data input can be achieved by downloading files from the Sequence Read Archive at NCBI. These are actual screen shots of the data entry section from https://bioedge.lanl.gov/.

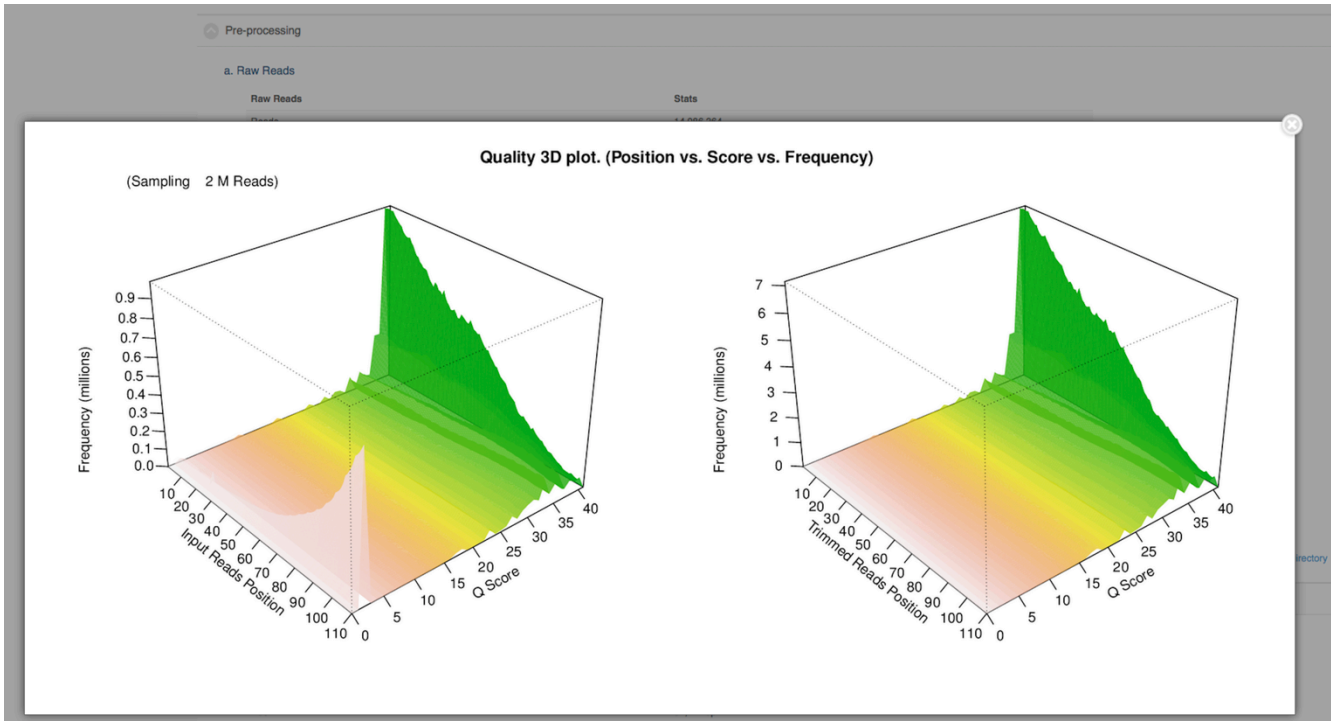


Supplementary Figure S3. The EDGE Project page displays an analysis in progress. The Project page has links to other pages and the project list on the left (which can be hidden with a link in the upper left corner). Completed jobs, running jobs and queued jobs are tagged in green, orange, and gray, respectively. Project information and results are shown in the center section. The information on this page is static and allows users to access portions of the run that are already complete, however the page needs to be refreshed for any updates to the project. Active monitoring and action widgets are within a sliding panel on the right that is refreshed every 5 seconds. This is an actual screen shot of a project page from https://bioedge.lanl.gov/.

a



b



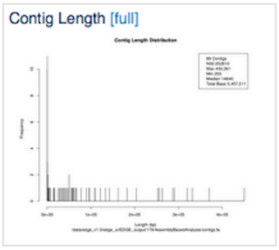
Supplementary Figure S4. The EDGE Pre-processing results. a) An example of results from the Pre-Processing module shown on the project page of the *Yersinia* isolate genome. **b**) Clicking on any of the thumbnails of the graphs shown below the statistical results will open a light box version of the graph. These results are for the *Yersinia* isolate genome showing that low quality data has been removed. These are actual screen shots from https://bioedge.lanl.gov/.

Assembly and Annotation

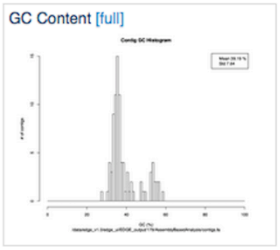
a. De Novo Assembly by idba_ud

Assembly	Stats
Number of contigs	89
N50	252,610 bp
Max contig size	450,361 bp
Min contig size	203 bp
total assembly size	5,457,511 bp

Contig Length [full]



GC Content [full]

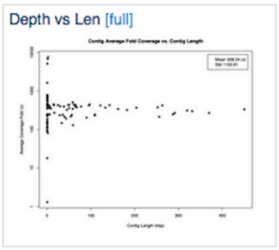


Link to | Report PDF | Contigs Fasta | JBrowse | Directory

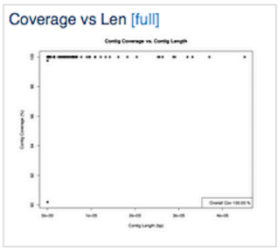
b. Assembly Validation by Read Mapping

Mapped	Stats
Number of Mapped Reads	18,413,766
% of Total Reads	98.68 %
Number of Unmapped Reads	246,011
% of Total Reads	1.32 %
Average Fold Coverage	328.23 X

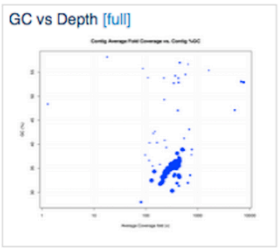
Depth vs Len [full]



Coverage vs Len [full]



GC vs Depth [full]

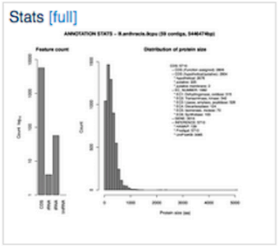


Link to | Report PDF | JBrowse | Directory

c. Annotation

Annotation	Stats
CDS	5,710
rRNA	4
tRNA	57

Stats [full]



Show the results in JBrowse

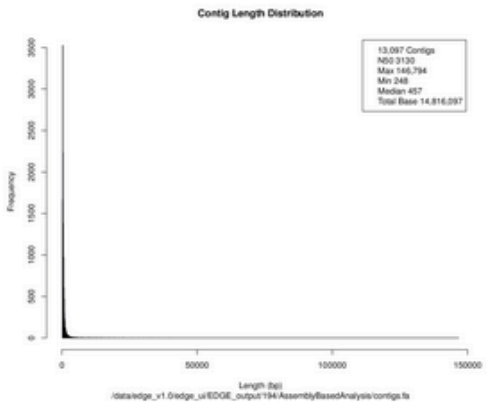
Supplementary Figure S5: The EDGE Assembly and Annotation results. Example results for the Assembly and Annotation module shown on the project page of the *Bacillus* isolate genome. This is an actual screen shot from https://bioedge.lanl.gov/

Assembly and Annotation

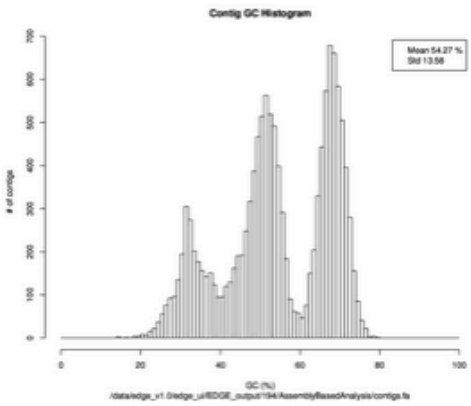
a. De Novo Assembly by idba_ud

Assembly	Stats
Number of contigs	13,097
N50	3,130 bp
Max contig size	146,794 bp
Min contig size	248 bp
total assembly size	14,816,097 bp

Contig Length [full]



GC Content [full]



b. Assembly Validation by Read Mapping

Mapped	Stats
Number of Mapped Reads	4,998,745
% of Total Reads	77.61 %
Number of Unmapped Reads	1,442,065
% of Total Reads	22.39 %
Average Fold Coverage	23.96 X

Supplementary Figure S6. Assembly results for a Human Microbiome Project (HMP) mock community. Results for Assembly (and assembly validation) of the HMP staggered mock community sample. Assembly provided a largest contig of 146.8 kb and a total assembly size of 14.8 Mb, while validation by read mapping indicates that only 77.6% of the data assembled and the contigs have an average coverage of 24X. This is an actual screen shot from https://bioedge.lanl.gov/.

a

a. Reads Mapped to Reference(s)

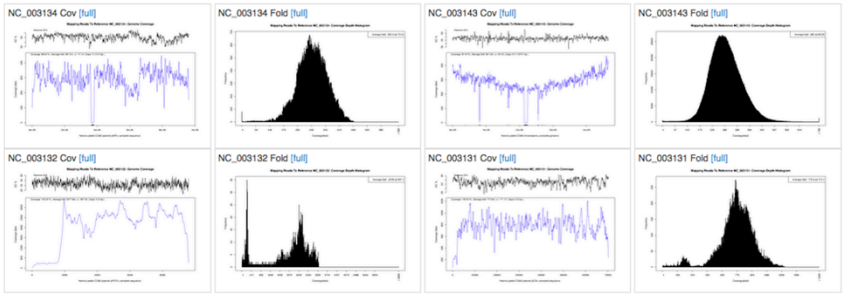
i. Mapped Reads

Analysis	Stats
Number of Mapped Reads	14,104,624
% of Total Post-QC Reads	98.38 %
Average Fold	291.98X
Linear Coverage	97.51%
SNPs	655
InDels	228

Columns...

Reference	Name	Length	GC%	Mapped Reads	Base Coverage	Avg Fold	Fold std.	Gaps	Gap bases	SNPs	INDELs
NC_003134	Yersinia pestis CO92 plasmid pMT1, complete sequence	96,210	50.23%	283,164	98.63	291.91X	73.14X	4	1,312	17	4
NC_003143	Yersinia pestis CO92 chromosome, complete genome	4,653,728	47.64%	13,084,113	97.44	281.02X	85.35X	97	118,751	624	222
NC_003132	Yersinia pestis CO92 plasmid pPCP1, complete sequence	9,612	45.27%	198,084	100.00	2077.96X	957.01X	0	0	0	0
NC_003131	Yersinia pestis CO92 plasmid pCD1, complete sequence	70,305	44.84%	539,263	100.00	773.64X	171.13X	0	0	14	2

4 out of 4 reference(s) is(are) covered by input reads.



Show the results in JBrowse

Link to I All Plots PDF I SNP Report I JBrowse I Directory

ii. Unmapped Reads

Unmapped	Stats
Number of Unmapped Reads	232,333
% of total reads	1.62%

Taxonomy ID of unmapped reads with BWA

Columns...

Organism	Length	GC	Avg Fold	Fold std.	Base Coverage	Mapped Reads	Linear Length
Yersinia pestis	4,532,063	47.58%	0.36X	4.93	1.51%	16,410	68,443
Yersinia pestis	4,534,590	47.58%	0.24X	3.04	1.25%	11,413	56,927
Yersinia pestis	4,595,065	47.65%	0.19X	2.49	1.07%	8,999	49,282
Yersinia pestis	4,553,586	47.67%	0.18X	2.58	0.96%	8,721	43,996
Yersinia pestis	4,640,720	47.62%	0.18X	2.50	0.92%	8,515	42,857

Only top 5 results in terms of "Mapped Reads" are listed in the table.

Link to I Mapping Result I Directory

b

b. Contigs Mapped to Reference(s)

Analysis	Stats
Number of Mapped Contigs	329
Proportion	100.00%
Average Fold	1.41X
Linear Coverage	97.29%
Average Identity	99.19%
SNPs	524
InDels	297

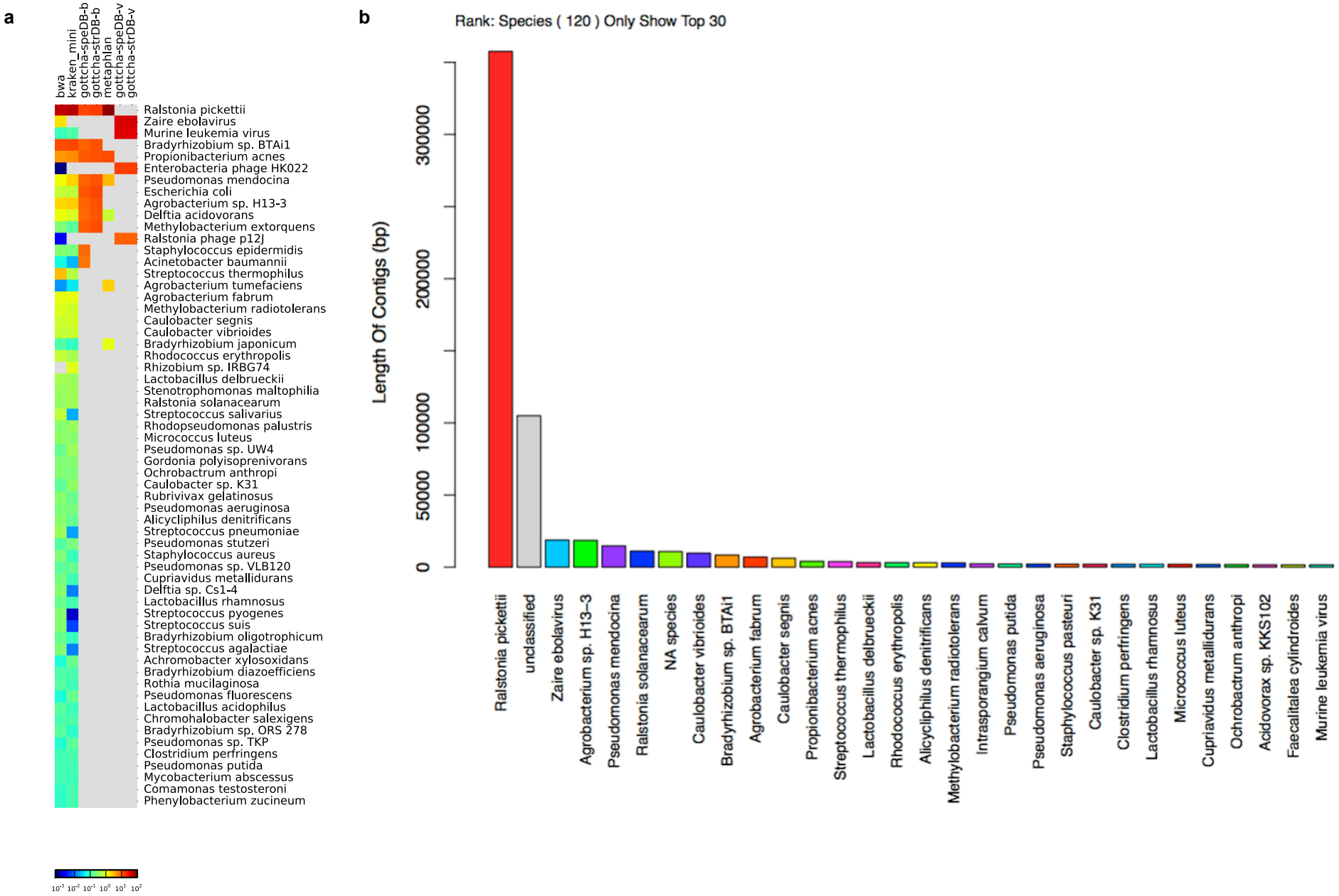
Columns...

Reference	Name	Length	GC%	Mapped Contigs	Base Coverage	Avg Fold	Gaps	Gap bases	SNPs	INDELs
NC_003143	Yersinia pestis CO92 chromosome, complete genome	4,653,728	47.64%	318	97.22%	1.42X	90	129,066	506	1,546
NC_003134	Yersinia pestis CO92 plasmid pMT1, complete sequence	96,210	50.23%	164	98.15%	1.28X	2	1,779	10	72
NC_003131	Yersinia pestis CO92 plasmid pCD1, complete sequence	70,305	44.84%	75	100%	1.19X			8	6
NC_003132	Yersinia pestis CO92 plasmid pPCP1, complete sequence	9,612	45.27%	54	100%	1.55X			0	2

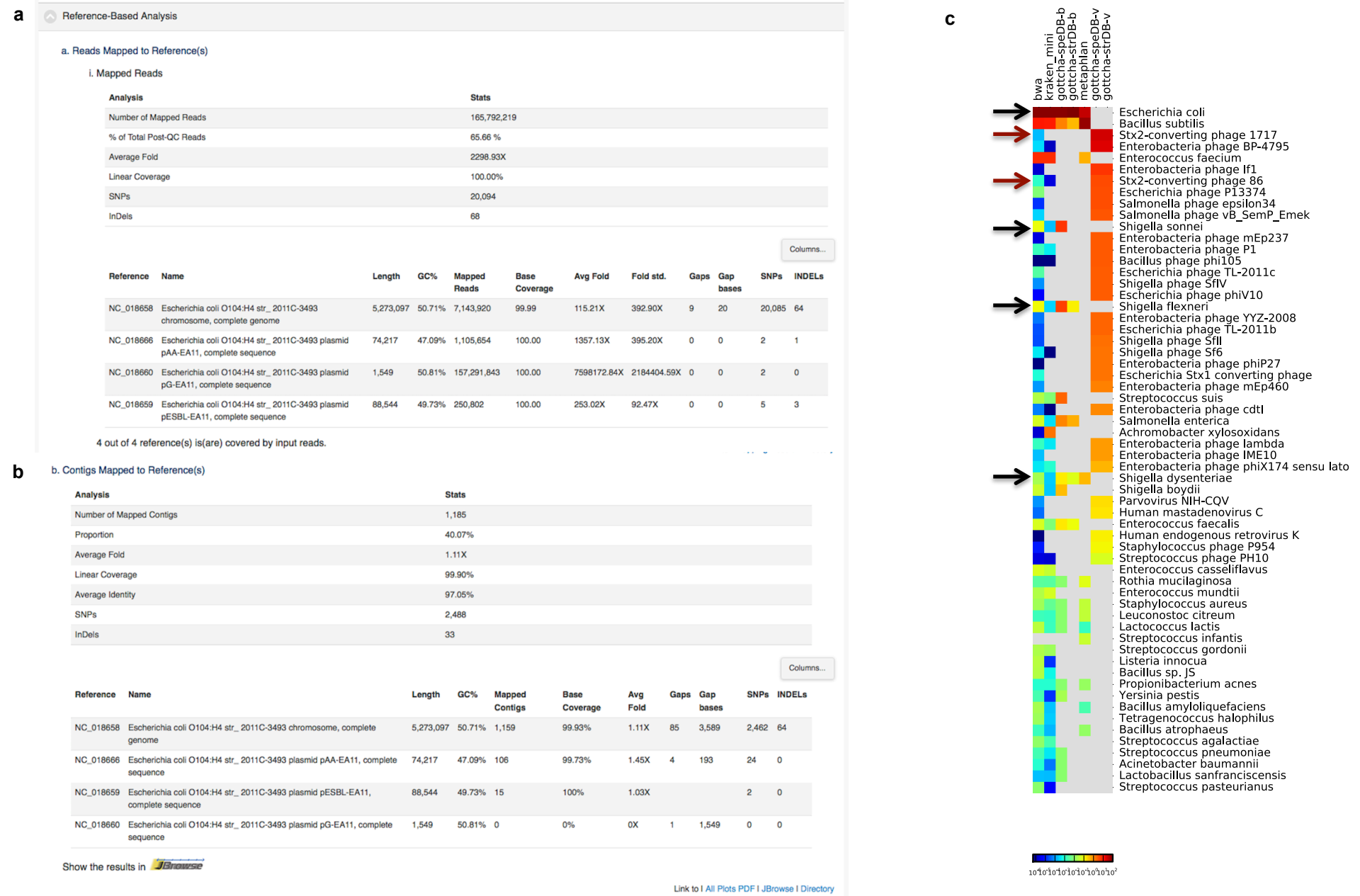
Show the results in JBrowse

Link to I All Plots PDF I JBrowse I Directory

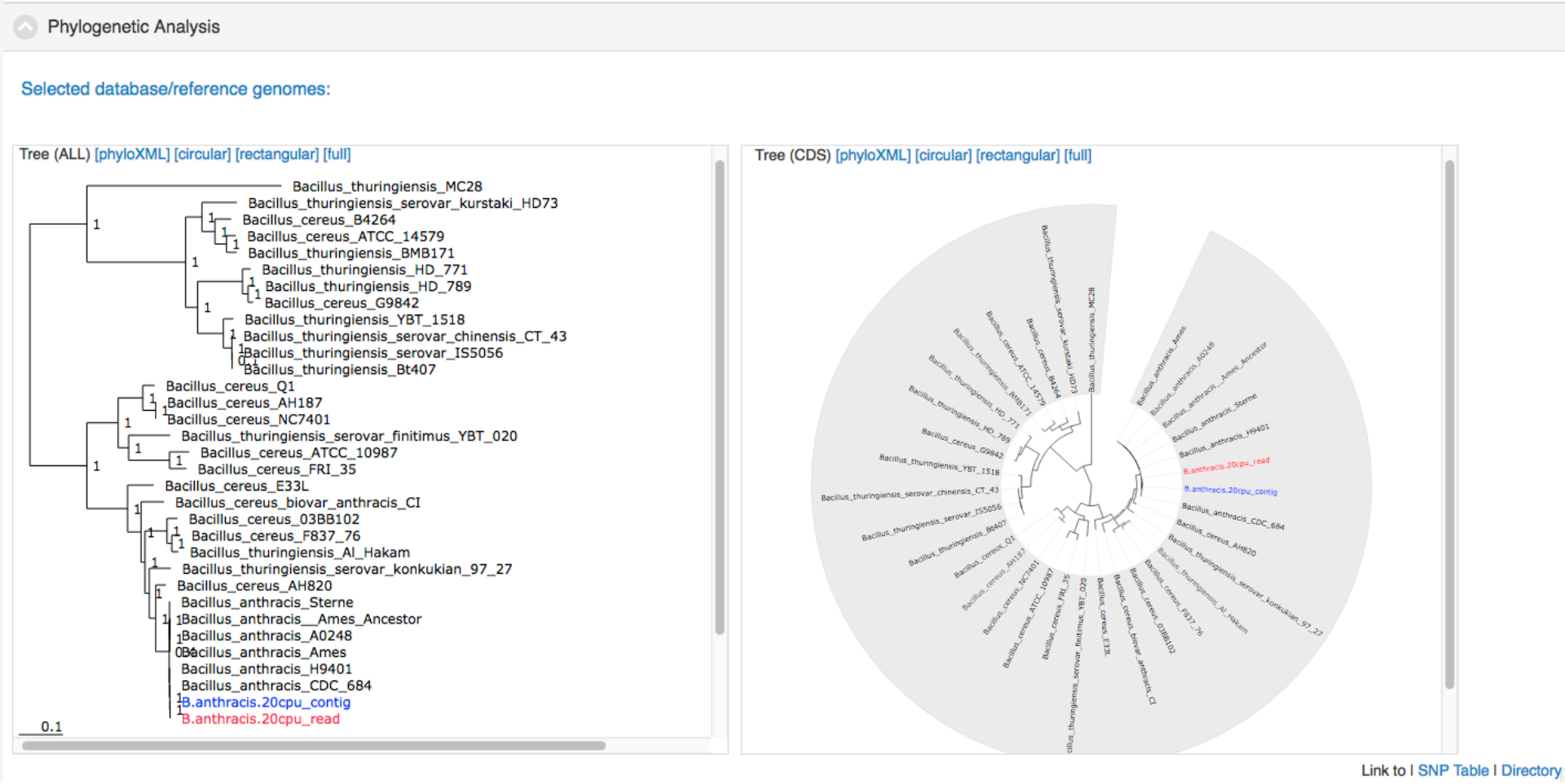
Supplementary Figure S7. The EDGE Reference-Based Analysis results. Example results for the read mapping of the Reference-Based Analysis module shown on the project page for the *Yersinia* isolate genome. B) Example results for the contig mapping of the Reference-Based Analysis module shown on the project page for the *Yersinia* isolate genome. These are actual screen shots from https://bioedge.lanl.gov/. Similar results for the *Bacillus* isolate can be found there.



Supplementary Figure S8. Taxonomy Classification results for a human clinical sample. a) Read-based taxonomic classification of a clinical sample from an individual carrying Ebola. Ebola is clearly identified as the top viral hit with GOTTCHA and is also found with BWA. **b**) Contig-based taxonomic classification of the same clinical sample. Ebola was clearly identified as one of the top hits based on cumulative contig length. These are actual screen shots from https://bioedge.lanl.gov/.



Supplementary Figure S9. Reference-Based Analysis and Read-based Taxonomy classification of the fecal sample. a) Reads mapped to the *E. coli* O104:H4 reference cover >99.99% of the chromosome and 100% of all three plasmids at very high fold coverage. **b**) Forty percent of the metagenome's assembled contigs aligned to the same reference cover >99.93% of the chromosome and 99.73-100% of the two larger plasmids. **c**) Taxonomic classification of the fecal sample shows dominant *E. coli* and *Shigella* (black arrows) and the presence of two Shiga-toxin converting phages (red arrows) along with commensal organisms. These are actual screen shots from https://bioedge.lanl.gov/.



Supplementary Figure S10. Phylogenetic trees of an isolate sample. Rectangular (traditional) and circular phylogenetic trees are constructed showing placement of both the reads and the assembled contigs of the *Bacillus anthracis* sample within the reference *Bacillus* genomes selected.

PCR Primer Analysis

a. Primer Validation

Primers	Target	Location	Product Size
pag_1_fwd_primer, pag_1_rev_primer	B.anthracis.8cpu_0033	17505..18251	747
pag_2_fwd_primer, pag_2_rev_primer	B.anthracis.8cpu_0033	18020..18170	151
cya_1_fwd_primer, cya_1_rev_primer	B.anthracis.8cpu_0033	37898..38826	929
cya_2_fwd_primer, cya_2_rev_primer	B.anthracis.8cpu_0033	38077..38622	546
lef_1_fwd_primer, lef_1_rev_primer	B.anthracis.8cpu_0033	11113..11497	385
lef_2_fwd_primer, lef_2_rev_primer	B.anthracis.8cpu_0033	10824..11816	993
capC_fwd_primer, capC_rev_primer	B.anthracis.8cpu_0025	48039..48302	264
capBCA_fwd_primer, capBCA_rev_primer	B.anthracis.8cpu_0025	47666..48538	873
Ba813_fwd_primer, Ba813_rev_primer	B.anthracis.8cpu_0021	21869..22020	152
Ba_EPA_2F_capB_fwd_primer, Ba_EPA_2R_capB_rev_primer	B.anthracis.8cpu_0025	46921..46997	77
Ba_EPA_1F_pagA_fwd_primer, Ba_EPA_1R_pagA_rev_primer	B.anthracis.8cpu_0033	16677..16777	101
Ba_BC3_F_fwd_primer, Ba_BC3_R_rev_primer	B.anthracis.8cpu_0006	195214..195318	105
pag_1_fwd_primer, pag_1_rev_primer	NC_007322	143900..144646	747
pag_2_fwd_primer, pag_2_rev_primer	NC_007322	143981..144131	151
cya_1_fwd_primer, cya_1_rev_primer	NC_007322	123319..124247	929
cya_2_fwd_primer, cya_2_rev_primer	NC_007322	123523..124068	546
lef_1_fwd_primer, lef_1_rev_primer	NC_007322	150645..151029	385
lef_2_fwd_primer, lef_2_rev_primer	NC_007322	150326..151318	993
capC_fwd_primer, capC_rev_primer	NC_007323	55208..55471	264
capBCA_fwd_primer, capBCA_rev_primer	NC_007323	54972..55844	873
Ba813_fwd_primer, Ba813_rev_primer	NC_007530	4564808..4564959	152
Ba_EPA_2F_capB_fwd_primer, Ba_EPA_2R_capB_rev_primer	NC_007323	56513..56589	77
Ba_EPA_1F_pagA_fwd_primer, Ba_EPA_1R_pagA_rev_primer	NC_007322	145374..145474	101
Ba_BC3_F_fwd_primer, Ba_BC3_R_rev_primer	NC_007530	4851870..4851974	105

b. Primer Design

Primer Name	Location	Forward Primer	Forward Tm	Reverse Primer	Reverse Tm	Size	Background
B.anthracis.8cpu_0002-1	301729..302064	CGCTTCTTGCACTGGATCTC	59.0 C	ATACTGGCCGGAGCGTTAAT	59.2 C	336 bp	[48.79 C] Bacillus cereus AH187 chromosome
B.anthracis.8cpu_0068-1	36..257	CACCCAATGGAATGGTCACC	58.8 C	GCTGACCTCTCCTAACTGGA	58.1 C	222 bp	[39.08 C] Pseudomonas stutzeri RCH2 chromosome

Supplementary Figure S11. The EDGE PCR Primer Analysis results. Results for PCR Primer Analysis for the *Bacillus* isolate. Twelve previously published primers for *B. anthracis* were validated and two new primer pairs for this particular isolate were generated. This is an actual screen shot from https://bioedge.lanl.gov/.

Supplementary Table S1. 16S sequence data from the nasal swab sample. *E. coli* is not detected but an organism in the genus *Cornyebacterium* is found.

Organism Name	# of 16S Hits		Organism Name	# of 16S Hits
Veillonella parvula	2521		Actinomyces graevenitzii	55
Atopobium parvulum	1091		Erysipelothrix rhusiopathiae	54
Actinomyces odontolyticus	846		**Corynebacterium ciconiae**	**50**
Filifactor alocis	743		Alloprevotella tannerae	50
Parvimonas micra	723		Geobacillus thermoglucosidasius	49
Prevotella nigrescens	573		Cryptobacterium curtum	48
Rhodoferax antarcticus	559		Lactobacillus reuteri	48
Streptococcus sanguinis	497		Shuttleworthia satelles	48
Elizabethkingia meningoseptica	474		Selenomonas sputigena	45
Brevundimonas vesicularis	423		Olsenella uli	44
Megasphaera micronuciformis	399		[Eubacterium] sulci	39
Dialister pneumosintes	374		Schlegelella thermodepolymerans	32
Porphyromonas endodontalis	322		Rothia mucilaginosa	30
Prevotella histicola	289		Prevotella denticola	28
Veillonella rodentium	230		Parapedobacter pyrenivorans	28
Solobacterium moorei	225		Prevotella oulorum	26
Dialister propionicifaciens	205		Gemella cuniculi	26
Bifidobacterium dentium	176		Eubacterium saphenum	26
Flavobacterium ceti	157		Dialister micraerophilus	24
Pedobacter steynii	132		Vasilyevaea enhydra	23
Stenotrophomonas maltophilia	132		Thermomonas hydrothermalis	22
Selenomonas flueggei	125		Slackia exigua	21
Bifidobacterium pseudolongum	119		Agrobacterium fabrum	20
Prevotella oris	116		Bosea lathyri	19
Bulleidia extructa	112		Sphingomonas paucimobilis	19
Fretibacterium fastidiosum	99		Dialister succinatiphilus	16
Propionibacterium acnes	91		Prevotella oralis	15
Rothia dentocariosa	77		Geobacter luticola	14
Tannerella forsythia	74		Campylobacter fetus	14
Anaeroglobus geminatus	61		Marinobacterium georgiense	13