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)

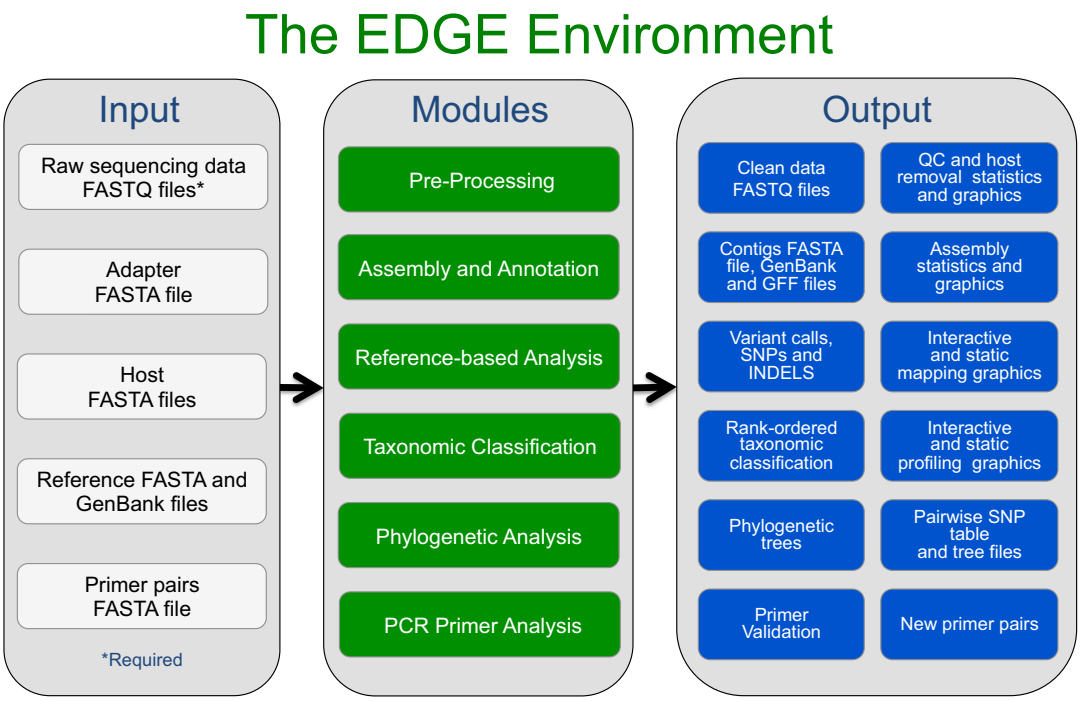

Figure 1. An overview of the EDGE Bioinformatics Environment. The only inputs required from the user are raw sequencing data and a project name. The user can create specific workflows with any combination of the modules. In addition, tailored parameters dictating how each module functions can be modified 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-

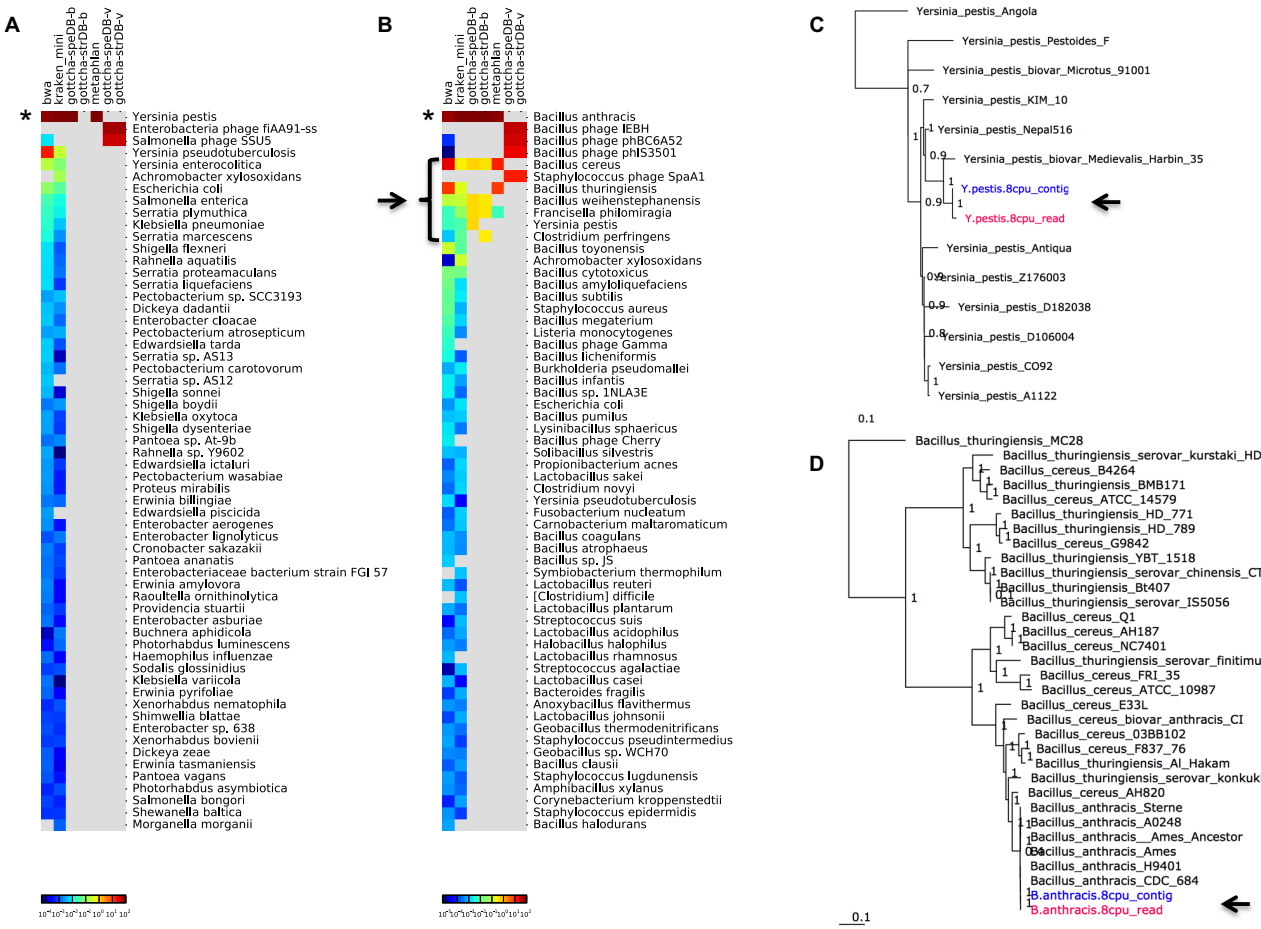

Figure 2. Taxonomy and phylogenetic evaluations of bacterial isolates. Panels **A** and **B** show taxonomic classification of reads for (A) the *Y. pestis* Harbin35 sample and (B) the *B. anthracis* SK-102 sample. The stars indicate the consistent dominant taxonomic calls for all tools, while the black 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

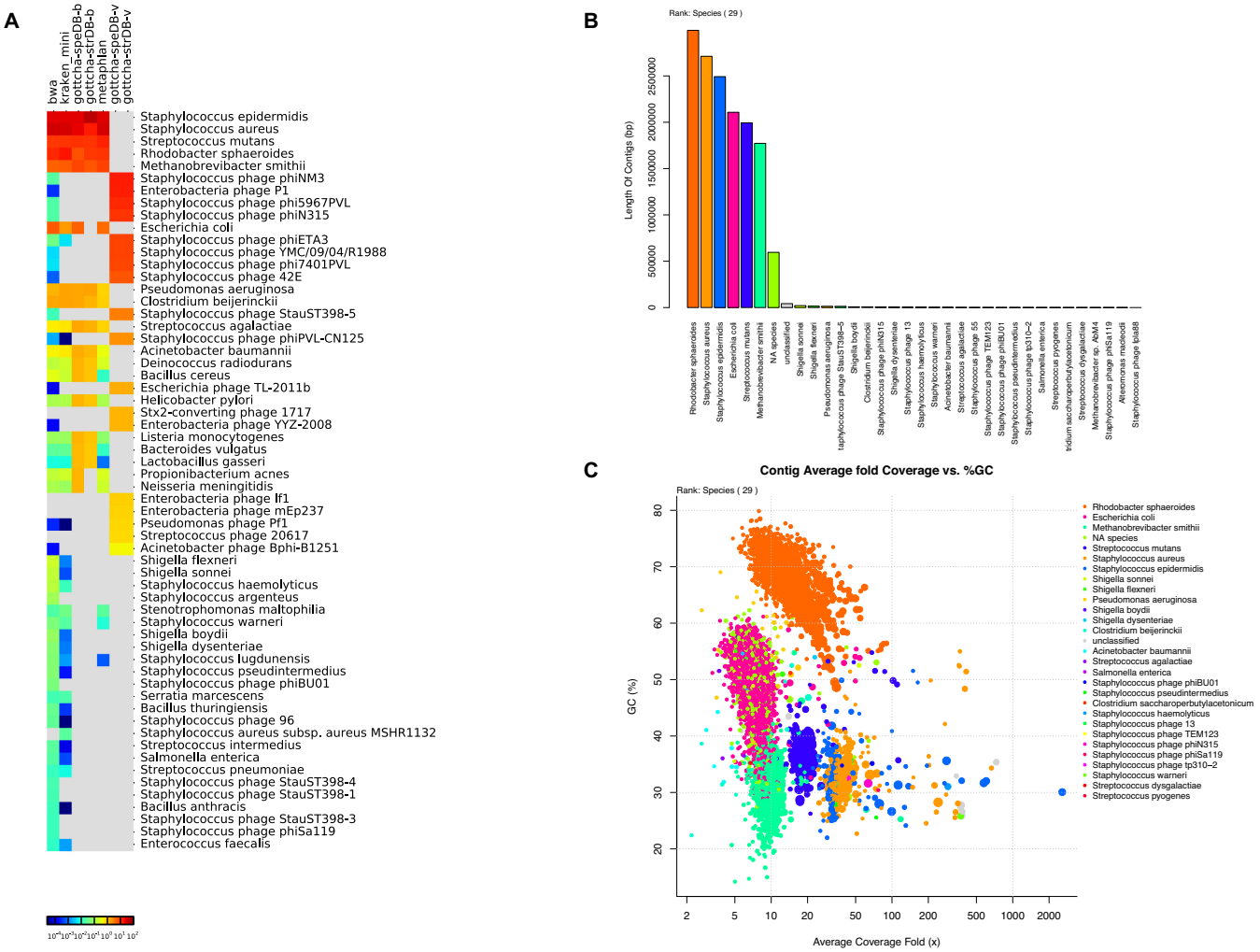

Figure 3. Taxonomic Classification of the HMP staggered mock sample. (**A**) Read-based classification using various taxonomy profiling tools; (**B**) contig-based classification displaying length of all classified contigs per taxon and (**C**) a scatterplot of contig % GC versus fold coverage of the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

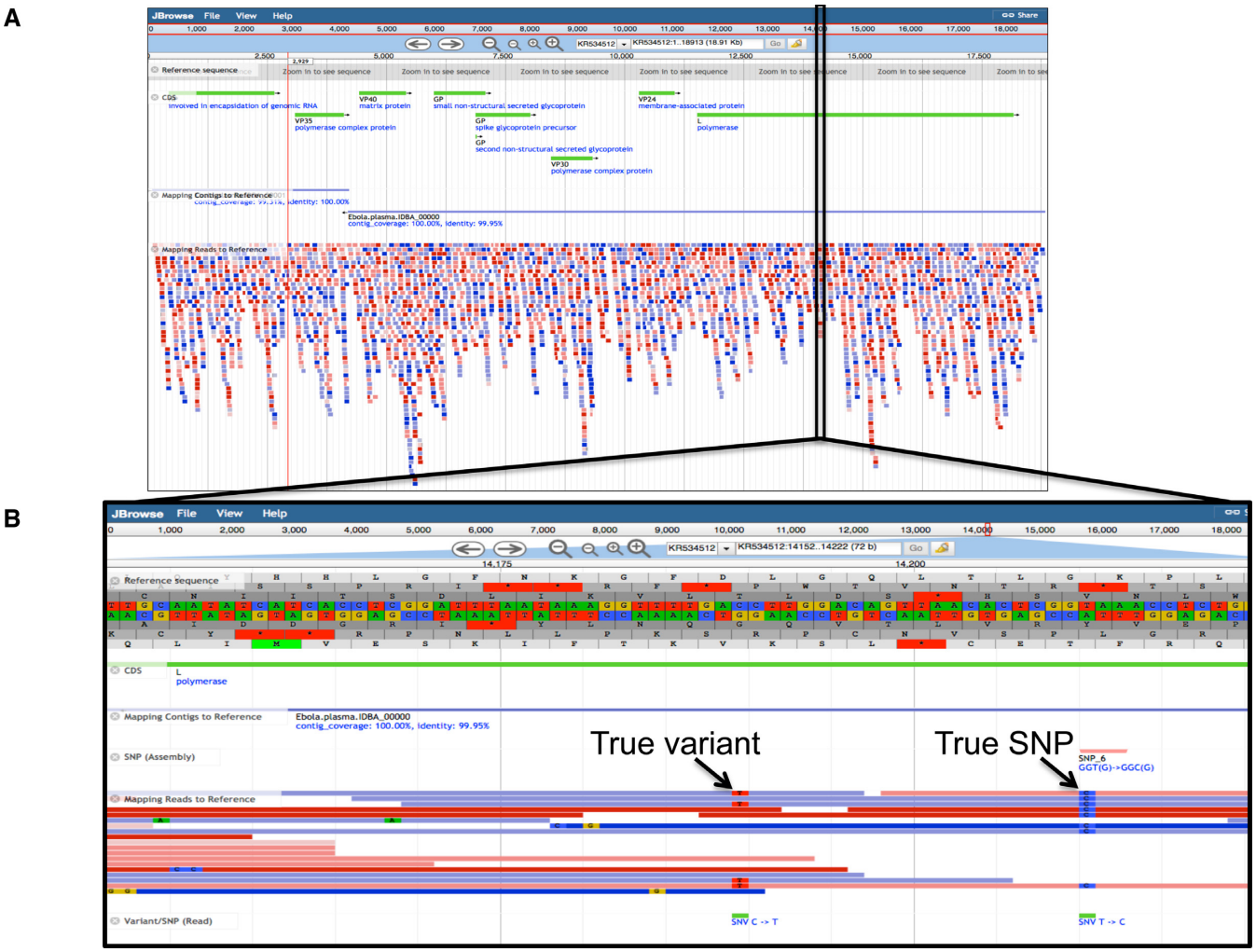

Figure 4. Interactive genome browsing view of a reference-based analysis in EDGE with a human clinical sample containing Ebolavirus. (**A**) An Ebola reference genome and its genes (green lines) are displayed together with contig-based (using IDBA) and read-based comparisons. The two contigs (blue lines) 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.

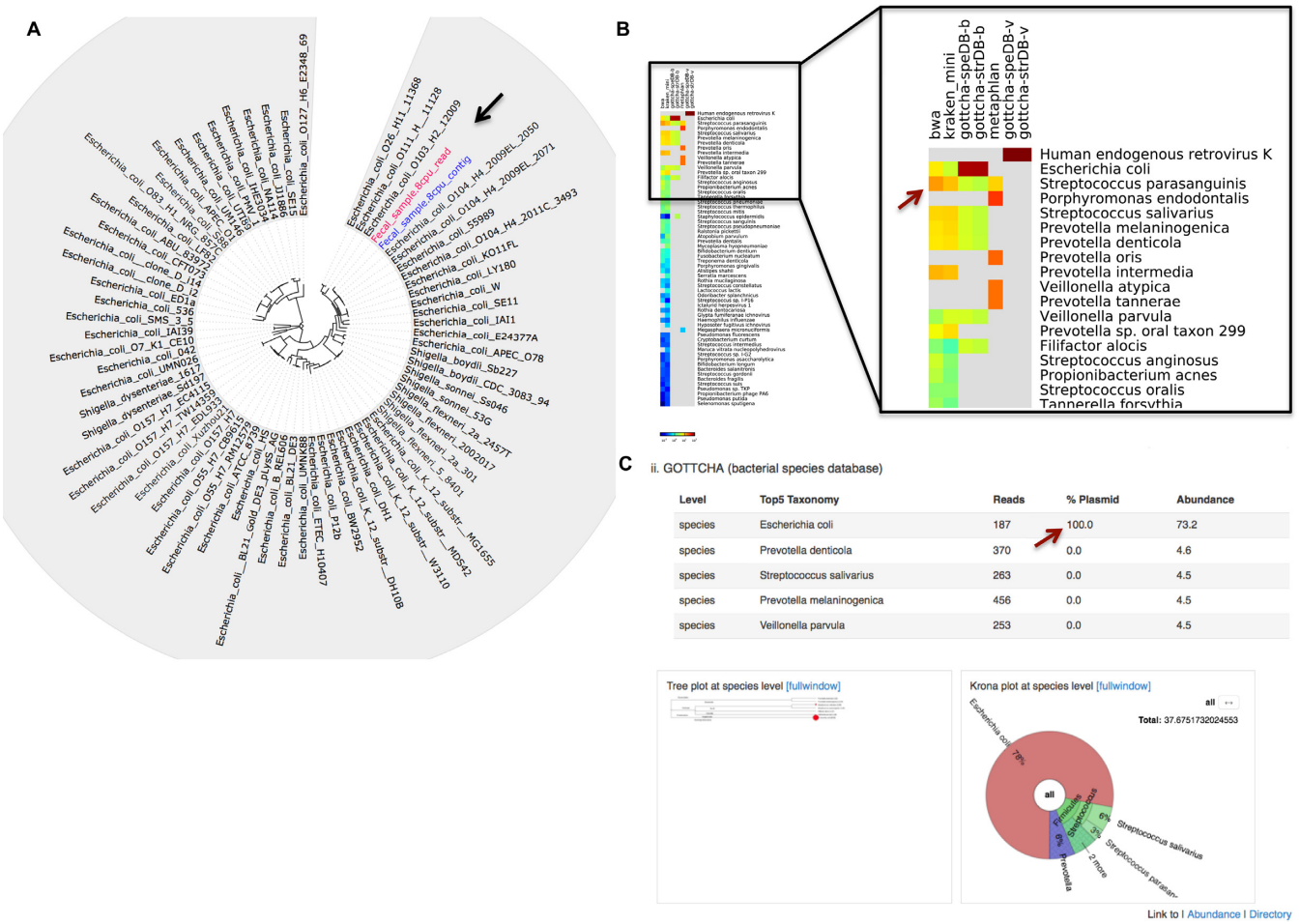

Figure 5. Phylogenetic and taxonomic analysis of human clinical samples with suspected and unknown causative agents. (**A**) Circular phylogenetic tree clearly places within the *E. coli* O104 group both the raw reads and the contigs obtained from a clinical fecal sample. (**B**) A comparative heatmap 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.