DOE Joint Genome Institute FY15 - End of Year Progress Report

Overview

The present FY15 End-of-Year Progress Report concludes the FY15 series of "Quarterly Metrics" progress reports from the JGI in which we described several new experimental or computational capabilities for the assembly, analysis, and functional annotation of genomes.

The topic of the Q1 report was "New computational developments for improving microbial, metagenomic, or plant genome assemblies". We described the development of Meraculous2, a state-of-the-art de novo assembler for short reads; its parallelization for use in high performance computing settings; the development of efficient methods for shotgun genetic mapping; whole-genome shotgun assembly of hexaploid wheat; and the development of merAligner, a highly parallel sequence aligner.

In Q2, we reported on "Progress developing new experimental capabilities to analyze complex genomic or metagenomic datasets." We described the development of a general strategy for the genome-wide functional annotation of microbial genes using saturation transposon mutagenesis, followed by sequencing (TnSeq). In particular, we reported on the implementation of a new random barcode-enabled transposon site-sequencing approach (RB-TnSeq) and its application in the context of the Functional Encyclopedia of Bacteria and Archaea (FEBA) project.

A report on "New computational approaches for the annotation of genomic data," published in Q3, described the development of ProDeGe, a fully automated protocol for decontamination of single amplified genomes derived from single uncultivated microbial cells, as well as genomes assembled from metagenomic data.

In Q4, we provided a report on a "New computational method for improving the interpretation of microbial, metagenomic, or plant genomes." We described the development of a new data mart for the Integrated Microbial Genomes (IMG) resource, the Atlas of Biosynthetic gene Clusters (IMG-ABC). The IMG-ABC system is seamlessly integrated with the user interface of IMG and provides users with

immediate access to computationally predicted and experimentally validated biosynthetic gene clusters, as well as their associated secondary metabolites.

In this final report of the series, we describe our progress towards our end-of-year goal of developing a "New computationally enabled approach to analyze complex genomic datasets." Specifically, we describe the development of Elviz, an interactive tool for the visualization and exploration of metagenome assemblies. This effort represents another manifestation of the general direction defined by the prior four quarterly reports and further enhances the portfolio of computational and experimental capabilities provided by the DOE JGI.

Development of Elviz – an interactive visualization tool for exploration of metagenome assemblies[1]

Background

Metagenomics data can provide previously unattainable insights into the types of microorganisms that make up a system and the processes that they mediate. Metagenome sequencing has changed our understanding of energy metabolism in the oceans [2], biomass degradation in the gut of termites and cows [3,4], microbial bioremediation of metals and hydrocarbons [5,6], and the human microbiome [7], but its power has been limited by the difficulty to explore microbial communities at multiple levels of granularity to gain a systems-level understanding of their role in the habitat.

Visual analysis of metagenome data at the level of gene, genome, and ecosystem is of critical importance due to the huge volumes and complexity of the data produced, yet relatively few interactive visualization tools are specifically geared toward microbial community data. Several visualization tools or packages exist for microbial community data, but most (e.g., Mothur [8] and QIIME ([8,10]) focus on phylogenetic profiling using 16S rRNA or other marker genes. Others visualize coverage data integrated with alignment information (e.g., MGAviewer [1]) or comparative analysis of complex metagenome data (e.g., Megan [12], MG-RAST [13] IMG/M-ER [14]), but produce mostly static images.

[1] A corresponding manuscript describing Elviz has been accepted and published by *BMC Bioinformatics* [1]

These tools, as well as current non-visual metagenome analysis platforms, treat metagenomes essentially as low quality genome and annotation data. Adding consideration of the rich information on organism abundance and adaptation conveyed by contig (and gene) sequence depth and heterogeneity [14-18] offers multiple advantages. To explore the relationships among these metadata (e.g., to investigate coverage vs. phylogenetic prediction in a sample) investigators must still create a static plot, which, using conventional methods, requires time-consuming manual steps. Changing display parameters or exploring different relationships within the data requires repeating these steps. This process is slow, and it requires that the investigator knows exact questions to ask beforehand.

To facilitate comparative metagenomics analyses, we designed a web-based interactive tool, Elviz, that eliminates time-consuming manual step in the analysis of metagenome assemblies [1]. Elviz enables the interpretation and visual exploration of assembled metagenome data, including sequence composition, assembly metrics, preliminary functional predictions, and phylogenetic affiliations. Integration of this information can aid in quickly defining microbial community structure and retrieving sequences and annotations of specific subsets of the data. These capabilities create a true discovery tool that allows for the recognition of phenomena before they can be quantified. Similar recognition tools have been revolutionary for other data-intensive fields (see http://www.nsf.gov/news/special_-reports/scivis/winners_2012.jsp) and while Elviz has been developed to address questions predominately relevant to microbiologists and specifically to provide the infrastructure necessary to explore metagenome datasets, most of the framework, libraries, and user interface of Elviz can also be utilized for visualizing data from areas other than microbiology.

Progress

1. Elviz architecture

Elviz is a web application, written primarily in AngularJS, JavaScript, and WebGL, and nearly all of the logic and computation occurs on the “client” side, in the browser. Users can load their own data into Elviz or explore metagenome assemblies created at the Joint Genome Institute (JGI) and provided

through the “server” side of Elviz, a thin REST server, written in Java, that sends data to the client in JSON or tabular text format.

Web browser vendors have put great effort into making the web platform a viable environment for fully featured applications that previously were squarely in the domain of the desktop operating system. This creates the opportunity to develop tools that harness the benefits of the internet (e.g., platform-independence, no need to install or setup software, connectivity to other resources, and the ability to share views with other users) while preserving the computational power and graphical interactive interfaces that were previously limited to the desktop environment.

Elviz takes advantage of two recent technological developments that have greatly accelerated the ability to create rich, efficient, and interactive visual tools on the web, namely WebGL (http://www.khronos.org/webgl/) and HTML5, in particular the LocalStorage API (http://www.w3.org/html/wg/; http://www.w3.org/TR/webstorage/). http://www.w3.org/TR/webstorage/

WebGL is a web-based implementation of the GL framework, which allows a web application to execute graphics commands using the client computer’s Graphical Processing Unit (GPU). The GPU is specialized hardware for graphics processing, with most graphics cards now supporting hardware-accelerated 3D rendering. Elviz uses this capability to increase the number of objects (i.e., metagenomic contigs) that can be displayed and manipulated in a responsive fashion. Elviz also leverages the wide variety of advanced visual effects available in the WebGL API to differentiate selections and show varying data parameters. The LocalStorage API provides access by the browser to the native file system. With the ability to store and access files on the user’s computer, an application can minimize expensive transfers of data between the server and the client. Additionally, this enables the local exploration of private datasets. LocalStorage also offers the possibility of caching remote datasets so that the user can revisit work in progress without retrieving information from the server.

Elviz has been successfully tested with Google Chrome, Mozilla Firefox, and Opera on OS/X and Windows, with Internet Explorer 9 on Windows, and with Safari on OS/X.

2. The Elviz Graphical interface

The Elviz interface is comprised of two primary components: (1) an interactive bubble-plot displaying metagenomic contigs (Figure 1A), and (2) a floating panel of controls (“Application Tools”) for configuring and manipulating the plot (Figure 1B).

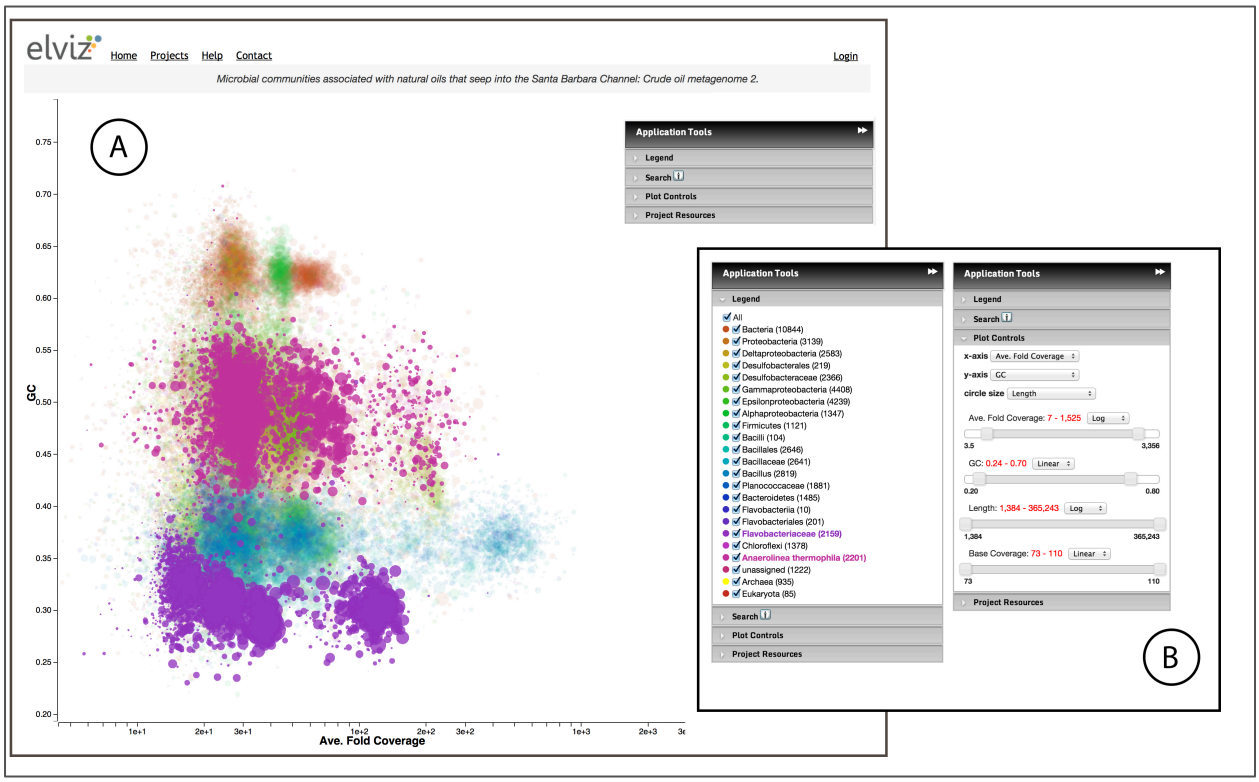

Figure 1: The Elviz Interface. A. The bubble-plot displays assembled contigs. The floating "Application Tools" panel in the upper right controls plotting parameters. **B.** "Application Tools" are shown with two different panels expanded. The Legend Panel (Fig. 1B left) controls the coloring and highlighting of different data groups in the plot. Here, contigs predicted to belong to *Flavobacteriaceae* and *Anaerolinea thermophilia* are highlighted. The Plot Controls panel (Fig. 1B right) controls axis selection, plot navigation, and data filtering.

The Plot and the Legend panel. Each point in the Elviz plot represents a single contig of the assembled metagenome. Points are displayed along four user-controllable dimensions: x-axis, y-axis, point size, and point color. For instance, in Figure 1A, GC content is plotted on the y-axis while “Ave.

Fold Coverage" – a measure of contig abundance – is plotted on the x-axis. The color of each point indicates the predicted taxonomic assignment of the contig (shown in the Legend Panel of the Application Tools; Figure 1B) while the point size is proportional to the length of each contig. The particular choice of parameters in this example is designed to support a visual assessment of the quality of taxonomic assignments in the sample. Contigs that derive from genomes of the same organism should have similar GC content and read coverage as compared to contigs that derive from separate organisms. Thus, if contigs of a particular color cluster together over these axes (GC content and coverage), the corresponding taxonomic assignment is corroborated.

The Legend panel allows the user to hide, highlight (brighten), and specify colors for contigs assigned to specific phylogenetic groups. These groups can be specified in a variety of ways, as will be described later. In addition, hovering the mouse over legend entries temporarily highlights the corresponding points in the plot, allowing for quick identification.

With JGI metagenomes, such as the example shown in Figure 1, colors are assigned for a finite set of phylogenetic classifications. This set is determined by an algorithm run during preparation of the datasets that determines the taxonomically "deepest" set of 25-30 taxa that can account for all of the contigs. The complete phylogenetic classification for each contig is, however, preserved in a metadata field called "Complete lineage" which can be seen when hovering the mouse over a contig (Figure 2A). This field, like all other contig metadata, is scanned when using Search (see below), making it possible to locate the set of contigs belonging to a taxon at any phylogenetic level.

Plot interaction and navigation. The ability to distinguish and visually separate points or groups of points within a dense and overlapping plot represents a central challenge in the visualization of large datasets. Elviz provides a number of features, including plot navigation, filtering, and search, that help to identify contigs of interest.

The Elviz plot can be navigated with mouse operations as well as through the "Plot Controls" panel (Figure 1B). Zooming and panning are accomplished by using the mouse scroll wheel and by click-dragging within the plot, respectively. The same result can be accomplished by setting the boundaries

of the axes using their respective sliders. Sliders for variables other than the x- and y-axis operate as filters, which allow the user to reduce visual noise (e.g., filtering out smaller contigs) in the plot or to focus only on contigs within specific parameter ranges.

Within the plot, hovering over any contig brightens this point and displays a panel showing the details of the given contig. When using Elviz with JGI projects, clicking on a point opens the "Contig Detail Viewer" in which the user can navigate along the contig to explore predicted genes and other functional annotations (Figure 2).

Figure 2: Exploring individual contigs. A. When the mouse moves over contig_11 in the plot, a panel appears showing all of the metadata for this contig. **B.** Clicking on the contig brings up the Contig Detail Viewer. Here the user can explore gene annotations on the contig, navigating with the slider at the top of the panel. Red and black glyphs represent predicted gene models in forward and reverse orientation respectively. The yellow-filled gene model is currently selected. Details for this annotation appear in tabular form, including the feature name and position, and predicted COG and Pfam classification

Contigs on the Elviz plot are searchable. Figure 3 shows the results of using the “Search Controls” to locate contigs containing a particular Pfam annotation. Matching contigs appear as black-outlined circles in the plot and are presented in two tables. The first table shows hit counts for each “group”, which can be shown or hidden using the associated check boxes. The second table lists the individual contigs. When viewing JGI projects, clicking on these contigs will bring up the Contig Detail Viewer.

Via the "Download Panel", Elviz supports the export of subsets of contigs from search or visual selection (enabling or disabling groups or group search results) in a variety of formats, including CSV, and, when available for JGI datasets, GFF (annotations) and FASTA (contig sequences).

3. Elviz data

A number of metagenomes produced by DOE’s JGI and annotated in the Integrated Microbial Genomes with Microbiome Samples (IMG/M) database [19] are currently accessible for exploration in Elviz. These projects can be browsed via the "Projects" link at the top of the application.

In addition, users can import their own metagenomic assemblies and annotation into Elviz in an easy and highly customizable fashion using the Elviz upload wizard. The user simply provides a tab or comma delimited metadata file in which each row represents a contig and each column defines a feature (e.g., length or GC content) of the contigs. Column headings must be located in the first row of the table. After uploading the file, the Elviz upload wizard guides the user through a process of assigning columns in the data to the contig id, and default x-axis, y-axis, and point size properties of the plot (these can be changed dynamically once the data is loaded). In this step the user also specifies which columns should be included in the upload. Elviz will automatically assign numerical columns not assigned to plot properties, as "filter" parameters, for which a filter slider will be created. All parameters marked as included (numerical or descriptive) will be displayed in contig popups.

Next, the wizard asks the user to name the column to be used for point color and the method by which the color should be assigned. In the simplest case, the selected column contains ordinal names (e.g., phylogenetic assignments) to which colors can be assigned. In the case of columns containing quantitative values, Elviz supports (1) statistical binning of these values, with a single color then assigned to each bin or (2) creation of a "heat map" such that each point in the plot will be colored along a gradient representing the range of values in the chosen column.

Finally, the wizard provides the user with the option to load annotation data corresponding to the contigs in their dataset. Annotation files are accepted in GFF format.

The user is also given the option to store uploaded data securely and privately on JGI's Elviz server, provided that a login with the JGI is created. This will allow the user to revisit the imported project later from any computer without having to repeat the upload.

Conclusion

Data-intensive fields are often limited by the ability to extract meaning from large datasets. Easily maneuverable software to process and visualize biological (e.g., metagenome) data is critical to leveraging biological meaning from the "omics" datasets now generated in large amounts by thousands of individual investigators around the world. Elviz is a general-purpose tool for visualization of multidimensional data with a set of features that make it of particular value for a visual and efficient exploration of metagenomic data. In addition, Elviz allows one to explore and mine private as well as already published databases. In the example illustrated above, a simple assessment of a metagenomic data set comprised of 803,203 contigs totaling ~495 Mbp generated an instant picture of the distribution of function and phylogeny across the sample, and the immediate visual identification of outliers. This exercise required no bioinformatics expertise or software configuration beyond using a computer with a web browser. With the ability to search and export data from the tool, it was possible to further investigate

hypotheses generated from visual exploration with statistical means.

The versatility of Elviz will facilitate metagenomic analyses that would otherwise require extensive bioinformatic skills and a substantial infrastructure, both not readily available to many individual Principal Investigators. Elviz thus represents a valuable contribution to the scientific community, in particular to the field of microbiology and microbial ecology.

The interface and backend of Elviz are a platform from which a wide range of exploration capabilities will be added in the future. Many metagenomics research projects involve the collection of multiple samples, either from different environments or as part of time series surveys in order to understand dynamics of microbial communities in natural and laboratory conditions. Hence we have identified comparative metagenomics as the most critical next direction for Elviz, and we are currently looking into ways to efficiently visualize the similarities and differences among two or more metagenomic assemblies. Additionally, we recognize a need to integrate metagenomics data with complementary "omics" datasets (e.g., metatranscriptomics and metaproteomics). We are thus exploring methods for visually overlaying these modalities onto assembly data.

Elviz is freely available at http://genome.jgi.doe.gov/viz. Elviz requires a computer containing a Graphical Processing Unit (GPU) compatible with WebGL rendering in the browser (see http://get.webgl.org/) and runs in web browsers that support WebGL, including Chrome v31+, Firefox v35+, and IE v.11+ (see "http://caniuse.com/#feat=webgl"). Safari, which currently has only partial support for WebGL, is not recommended for use with Elviz. We have successfully tested Elviz using Chrome, Firefox, and Opera in Mac OS/X, Internet Explorer in Windows, and Firefox and Chrome in Linux operating systems. With fairly large datasets (50-100K contigs), we find that the initial load of the Elviz plot takes from 3-10 seconds over a DSL or Broadband connection (2-15 Mbs download speed). Subsequent loads of the same project take 2 seconds or less when the project has successfully been cached on the client using localStorage.

DOE JGI team

Michael Cantor, Henrik Nordberg, Tatyana Smirnova, Susannah Tringe, Inna Dubchak

References

1. Cantor M, Nordberg H, Smirnova T, Hess M, Tringe S, Dubchak I. (2015) Elviz - exploration of metagenome assemblies with an interactive visualization tool. BMC Bioinformatics. 2015 Apr 28;16:130
2. Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. Nature 411: 786-789.
3. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, et al. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331: 463-467.
4. Warnecke F, Luginbuhl P, Ivanova N, Ghassemian M, Richardson TH, et al. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450: 560-565.
5. Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, et al. (2010) Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. ISME J 4: 660-672.
6. Mason OU, Hazen TC, Borglin S, Chain PS, Dubinsky EA, et al. (2012) Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. ISME J 6: 1715-1727.
7. Human Microbiome Project C (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214.
8. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537-7541.
9. Caporaso JG, Knight R, Kelley ST (2011) Host-associated and free-living phage communities differ profoundly in phylogenetic composition. PLoS One 6: e16900.
10. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7: 335-336.
11. Zhu Z, Niu B, Chen J, Wu S, Sun S, et al. (2013) MGAviewer: a desktop visualization tool for analysis of metagenomics alignment data. Bioinformatics 29: 122-123.
12. Huson DH, Weber N (2013) Microbial community analysis using MEGAN. Methods Enzymol 531: 465-485.
13. Wilke A, Glass EM, Bartels D, Bischof J, Braithwaite D, et al. (2013) A metagenomics portal for a democratized sequencing world. Methods Enzymol 531: 487-523.
14. Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, et al. (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res 42: D568-573.
15. Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc 2010: pdb prot5368.

16. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, et al. (2014) Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci U S A 111: 4904-4909.
17. Mitra S, Rupek P, Richter DC, Urich T, Gilbert JA, et al. (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics 12 Suppl 1: S21.
18. Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2: 3.
19. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, et al. (2008) IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res 36: D534-538.