Exploration of Metagenome Assemblies with an Interactive Visualization Tool

Michael Cantor1, Henrik Nordberg1, Tatyana Smirnova1, Evan Andersen1, Susannah Tringe1, Matthias Hess2, Inna Dubchak 1,3

[1]DOE Joint Genome Institute, Walnut Creek, CA
[2]Washington State University, Richland, WA
[3]Lawrence Berkeley National Laboratory, Berkeley, CA

July 2014

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

LBNL-178516

Exploration of metagenome assemblies with an interactive visualization tool

Michael Cantor, Henrik Nordberg, Tatyana Smirnova, Evan Andersen, Susannah Tringe, Matthias Hess, Inna Dubchak

JOINT GENOME INSTITUTE
DEPARTMENT OF ENERGY

ABSTRACT

Metagenomics, one of the fastest growing areas of modern genomic science, is the genetic profiling of the entire community of microbial organisms present in an environmental sample.

Elviz is a web-based tool for the interactive exploration of metagenome assemblies. Elviz can be used with publicly available data sets from the Joint Genome Institute or with custom user-loaded assemblies.

Elviz is available at
genome.jgi.doe.gov/viz

MOTIVATION

Metagenomics datasets consist of millions of reads, partially assembled into 10s or 100s of thousands of contig consensus sequences (“contigs”). The assembly is partial due to the challenge of performing an enormous alignment over a variety of species (some unknown) at varying levels of abundance.

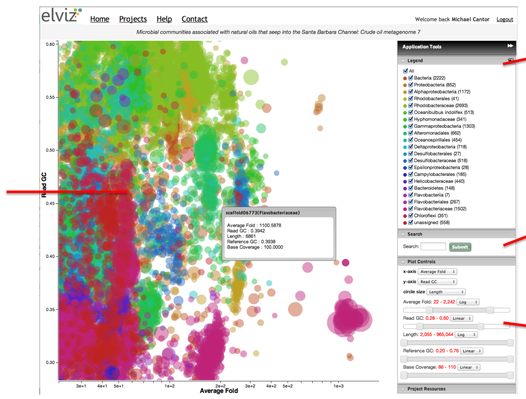
From a metagenomic datatset, investigators seek to determine:

- The composition of species in the sample.
- The relative abundance of these species.
- The metabolic functions performed within the community, as inferred from the presence of gene or protein families among the contigs.
- The distribution of these functions among species; the roles different species play in the community.
- The existence of novel members of gene families of interest (e.g. cellulases).

Answering these questions currently requires an iterative and labor-intensive process. Investigators sift through contigs, producing figure after figure in an attempt to identify species clusters by a variety of markers such as GC content, sample coverage, k-mer frequency, ribosomal sequences, or gene homology, and to generate distributions for these clusters for particular gene families.

Elviz seeks to accelerate this process by giving scientists the means to nimbly create and move among different plots of a metagenomic assembly and to explore them interactively. On the fly, users can use axes, color, and point size to cluster contigs by different variables in the sample, zoom in and and out of the plot, show or hide different groups of contigs, search for functional markers, and drill into the gene annotation of individual contigs. As compared with static plotting, the rapid visual feedback experienced with Elviz yields an accelerated process of hypothesis generation and discovery.

EXPLORING METAGENOME ASSEMBLIES



The Plot
The plot can be zoomed and panned by dragging and scrolling with the mouse. Hovering over each contig will display a small detail panel. Clicking on a contig brings up an annotation browser (see below).

Interactive Legend
Each contig group can be assigned to a color, hidden, shown, or high-lighted.

Search
Contigs can be searched for genes of interest, providing a powerful visual indication of the distribution of function across category in the sample.

Plot Configuration
Each variable in the data can be selected as a plot axis. Range sliders for each variable operate in one of two modes. For axes variables they operate to zoom and pan. For all other variables they act as filters.

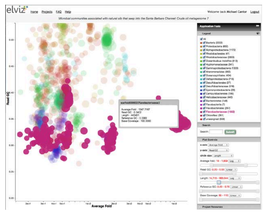
VISUALIZATION TASKS

RESOLVE

Dense, over-lapping plots are typical for metagenomic studies. Elviz offers multiple strategies for visually discriminating features of interest in the data.

Clusters of Flavobacteria are resolved by:

1. High-lighting this group in the Legend.
2. Zooming and panning in the plot
3. Using the Length filter to hide smaller contigs.

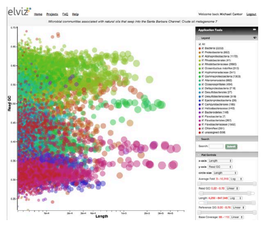


REORIENT

Elviz allows the user to explore relationships among different variables by dynamically assigning them to x-axis, y-axis, and point size, and by binning.

Length is plotted on the x-axis in order to quickly identify the largest contigs in the sample.

The overlapping clump of large Flavobacteria contigs is now stretched out so that they can be explored individually.

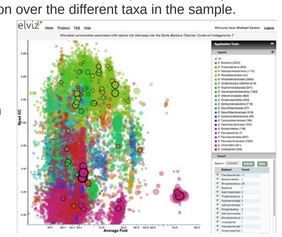


SEARCH

With Search, the user can identify contigs containing specific genes or functional annotations. The ability to overlay search results on the plot immediately illuminates the distribution of function over the different taxa in the sample.

Contigs are searched for the lipase COG model.

The majority of lipases appear in the Gammaproteobacteria and Flavobacteriaceae, both previously observed to arise late in the ecological succession following an oil spill.

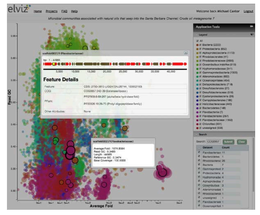


DRILL DOWN

Clicking on an individual contig opens an annotation browser allowing the user to examine the contig in detail.

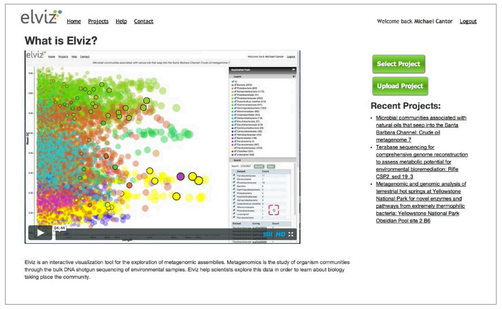
A Flavobacteriaceae contig containing the lipase is examined in detail allowing the user to reach the individual annotation and examine its gene neighborhood.

The user can also click on an annotation and search for it across all of the other contigs.



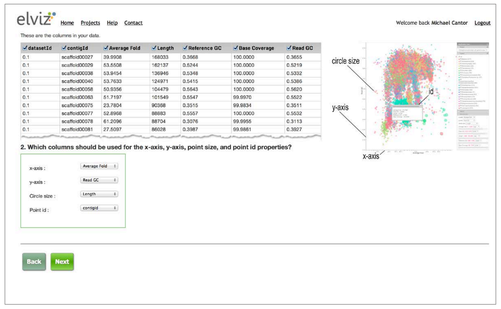
USER FEATURES

Landing Page



When logged in to the JGI, The home page provides convenient links to recently visited projects. From here the user selects whether to upload their own data or browse JGI projects.

Upload wizard



To load a project into Elviz, provide a tab or comma delimited file with variable names as the first row. The upload wizard provides a preview of the data file and allows the user to include and exclude columns, as well as to specify initial assignments for x-axis, y-axis, point size and point color.

IMPLEMENTATION

Current advances in web standards have ushered in a new breed of Internet applications that approach the performance and interactivity of desktop tools while maintaining the platform independence and sharing capacity of the web.

Elviz employs **WebGL**, harnessing the client’s graphical hardware (GPU), to render performant interactive displays of tens of thousands of data points. These data are stored on the client (using the HTML5 **localStorage** API). The client side of the application is written using **Angular.js**. This combination enables the user to operate Elviz without repeated round-trips to the server for data. The server side of Elviz consists of a RESTful API written in Java to provide data to the application.

When users choose to examine their own assemblies with Elviz (see below) they are given the choice to privately and securely store their uploaded data and project settings on the JGI Elviz server, allowing them to return to it in the future from any computer.

ELVIZ DATA

Using Elviz with JGI data
Investigators can use Elviz to explore either their own metagenomics datasets or datasets created from metagenomics sequencing projects at the JGI. Assembly, annotation, and phylogenetic prediction for JGI projects in Elviz originate from the JGI Integrated Microbial Genome metagenomics pipeline, IMG/M[1]. All publicly available IMG/M metagenome projects released after Jan 1, 2014 are available for analysis with Elviz.

Using Elviz with user data
Elviz supports exploration of user data in a simple and highly customizable fashion. Users provide a tabular file (tab or comma delimited) in which each row represents a contig and each column defines a variable (e.g. length, GC content). The first row of this file must contain the column headings. The Elviz upload wizard then allows the user to assign columns in their data to the various dimensions of the plot including x-axis, y-axis, point size, and point color. Elviz also provides the capability to define color groups by binning of a user-selected column.

1. Markowitz VM et al, Nucleic Acids Res. Jan 2012; 40(D1)

ACKNOWLEDGEMENTS

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. (DE-AC02-05CH11231).

CONTACT

jgi-viz@lists.lbl.gov

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.