GAViT: Genome Assembly Visualization Tool for Short Read Data

Aijazuddin Syed[*1], Harris J. Shapiro[1], Hank Tu[1], Jasmyn Pangilinan[1], Eugene Goltsman[1], Kurt LaButti[1], Alla Lapidus[1], Anthony Kosky[1] and Stephan Trong[2]

[1] Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA
[2] Lawrence Livermore National Laboratory – Livermore, CA

May 2008

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

GAViT: Genome Assembly Visualization Tool For Short Read Data

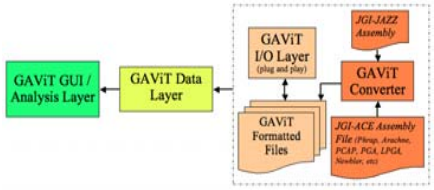
Aijazuddin Syed, Stephan Trong, Harris J Shapiro, Eugene Goltsman, Kurt LaButti, Alla Lapidus, and Anthony Kosky.

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

INTRODUCTION

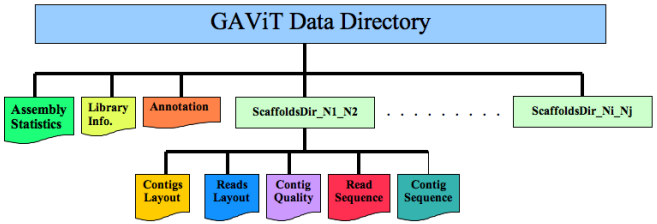
It is a challenging job for genome analysts to accurately debug, troubleshoot, and validate genome assembly results. Genome analysts rely on visualization tools to help validate and troubleshoot assembly results, including such problems as mis-assemblies, low-quality regions, and repeats. Short read data adds further complexity and makes it extremely challenging for the visualization tools to scale and to view all needed assembly information. As a result, there is a need for a visualization tool that can scale to display assembly data from the new sequencing technologies. We present Genome Assembly Visualization Tool (GAViT), a highly scalable and interactive assembly visualization tool developed at the DOE Joint Genome Institute (JGI).

GAViT OVERVIEW



High-level diagram of GAViT: GAViT is implemented with a flexible architecture to accommodate data from a variety of data sources. Further GUI layer is separated from the Data layer. GAViT is OS independent.

GAViT DATA ORGANIZATION



GAViT Data Organization: Assembly data is organized and formatted in a special GAViT format. Non overlapping scaffold ranges are binned into directories. These directories contain read layout, contig layout, sequence and quality information files. For example, ScaffoldsDir_Ni_Nj contains assembly information for scaffolds Ni to Nj. With such divide and conquer approach, GAViT can easily scale to view large genome assemblies. Also, this facilitates easy access of assembly data at various levels of resolution, including scaffold, contig, read, and consensus.

GAViT IMPLEMENTATION

GAViT is implemented in Java. A separate 'GAViT Converter' converts JGI's JAZZ/ACE assemblies into GAViT formatted data. Initially GAViT was developed to visualize DOE JGI's eukaryotic assemblies, however, now GAViT contains variety of finishing capabilities. Now a stable version of GAViT is available, and testing and new feature additions is ongoing. 'Ant' is used for project configuration management.

Usage: GAViT could be used in different modes, depending on the type of input.

(1) Directly start up GAViT and use file browser to load up the Assembly
(2) Directly start up GAViT with GAViT formatted data
(3) Directly start up GAViT with ACE file

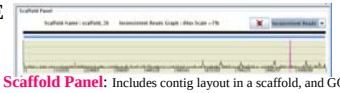
Display Optimization: In addition to data handling capabilities GAViT incorporates display optimizations techniques which make it interactive and scalable to view large genome assemblies and short read data assemblies. A few of such optimization techniques are:

- Selective rendering of the consensus sequence.
- Selective display of Reads in 'contig panel' and graphs in 'reports panel'. Layout is divided as layers of read channels. Rendering of the data is performed depending on vertical/horizontal scroll position.

GAViT FEATURES

- Birds eye view of entire Assembly. Browse through Scaffolds and Contigs through easily navigable tree. Scaffold-Contig-Read view of assembly.
- Edit Library information on the fly without restarting the tool
- Scalable to accommodate large data sets (tested for 1.8 GB) and short read data from new sequencing technologies. Divide-and-conquer approach for data.
- GAViT is a Java tool, hence, platform independent.
- Display reads in either Mesh or Tiling style. Supports flipping strands.
- Highlight reads depending on read categories (linking reads, mis-oriented reads, misplaced reads, stretched/compressed reads).
- Highlight reads by one or more libraries.
- Displaying various assembly statistics graphs (GC, clone coverage, read depth), and summary table for all reads in a scaffold.
- Link Analyzer – high level view of scaffold with contigs connecting through linking reads. Contigs colored by GC.
- Annotation, contig summary, and low/high GC, read depth and clone coverage reports
- User may choose to customize his/her own color scheme.
- Interactive Scaffold GC plotter – to identify potential contamination.

GAViT MAIN CONSOLE



Assembly Explorer Panel: A tree like organization of scaffolds and contigs. Click to load a contig/ scaffold.

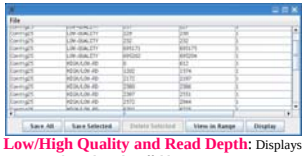
Scaffold Panel: Includes contig layout in a scaffold, and GC content, Read depth and Clone coverage graphs for the scaffold.

Contig Read Panel: Read layout can be portrayed in a mesh/tiling style (current display is mesh style). Reads can be highlighted depending on the type, library, etc. It is possible to Zoom in/out. The display region corresponds to the display region in graphs panel in the 'Reports Panel'. Annotations could be displayed/hidden, mark reads, etc.

Consensus Panel: Scaffold consensus sequence, and quality score bar.

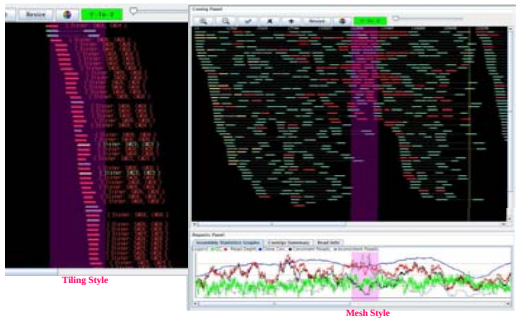
Reports Panel: Includes various assembly statistics graphs (**GC content**, **read depth**, **clone coverage**, **consistent reads**, inconsistent reads), read information, contigs summary table, and zoom in/out feature. Contig region displayed in 'assembly statistics graphs' corresponds to the display in the 'Read Panel'.

Library Information Edit Panel: Edit library information and apply changes, does not require a restart of the tool.

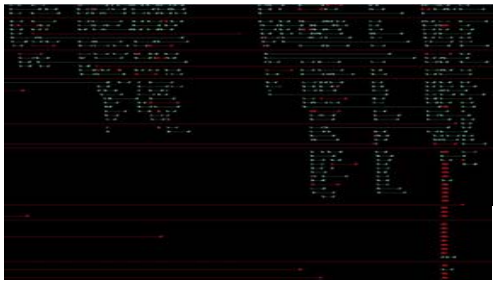


Low/High Quality and Read Depth: Displays regions within selected scaffold

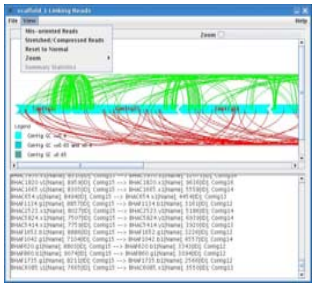
Assembly reads statistics: Displays total number of reads by category (library, sequencing technology). Further, we may zoom into each category.



Potential Mis-assembly: Mis-assembly in both Mesh and Tiling display styles. This facilitates easy detection of a false-positive assembly. As we can see in the highlighted region all the reads are either misplaced or stretched/compressed. This region is potential mis-assembly.

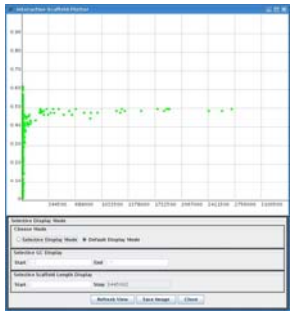


Zoomed out view of the reads



Link Analyzer: Displays contigs in a Scaffold according to their position and colors them by their GC content. Consistent and inconsistent linking reads are displayed between the linking contigs, they are colored and oriented to clearly depict the linking information. With this display users can easily point out poorly linked contigs. By selecting a contig, GAViT will load corresponding contig in GAViT main console.

Interactive Scaffold GC plotter: Plots GC for each Scaffold in the assembly. Assembly results are combined with the MegaBLAST to identify any contamination. The plots are colored by the contamination. There is a provision to Zoom in/out by selecting GC range or Scaffold length range user wants to view. By selecting a plot GAViT will load corresponding Scaffold in GAViT main console.



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