Application Note

GenomeVISTA – an integrated software package for whole-genome alignment and visualization

Alexandre Poliakov1,*  Justin Foong2  Michael Brudno2,3  Inna Dubchak1,4,*

1 US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598.
2 Centre for Computational Medicine, Hospital for Sick Children, Toronto, ON M5G 1X8 Canada
3 Department of Computer Science, University of Toronto, Toronto, ON M5S 3G4 Canada
4 Genomics Division, LBNL, Berkeley, CA, 94720.

ABSTRACT
Summary: With the ubiquitous generation of complete genome assemblies for a variety of species, efficient tools for whole-genome alignment along with user-friendly visualization are critically important. Our VISTA family of tools for comparative genomics, based on algorithms for pairwise and multiple alignments of genomic sequences and whole-genome assemblies have become one of the standard techniques for comparative analysis. Most of the VISTA programs have been implemented as Web-accessible servers and are extensively used by the biomedical community. In this manuscript we introduce GenomeVISTA: a novel implementation that incorporates most features of the VISTA family – fast and accurate alignment, visualization capabilities, GUI, and analytical tools within a standalone software package. GenomeVISTA thus provides flexibility and security for users who need to conduct whole genome comparisons on their own computers.

Availability and implementation: Implemented in Perl, C/C++ and Java, the source code is freely available for download at the VISTA website: http://genome.lbl.gov/vista/

Contact: avpoliakov@lbl.gov, idubchak@lbl.gov

Supplementary information: Available at the VISTA website.

1 INTRODUCTION
Comparing genomic sequences across related species has become a source of invaluable data on the functional elements in various genomes (Ponting and Hardison, 2011). There are a number of individual programs developed separately for genome alignment (Chen and Tompa, 2010) and visualization of comparative information (Chan, et al., 2012), but few programs and Web servers integrate the two, giving researchers an opportunity to analyze results interactively. VISTA (Frazer, et al., 2004), Dcode.org (Loots and Ovcharenko, 2005), the PipMaker suite of tools (Schwartz, et al., 2000) and Mauve (Darling, et al., 2010) are some examples of such integration.

VISTA on-line servers provide a wide range of services, which allow a user to: align and compare sequences from multiple species up to 10 Mb long using different algorithms (Bray, et al., 2003; Brudno, et al., 2003; Dubchak, et al., 2009), locate regulatory sequences using comparative sequence analysis and transcription factor binding site search (Loots, et al., 2002), compare user’s sequences against whole-genome assemblies, and browse pre-computed alignments of hundreds of microbial, fungal, plants, eukaryotic and other genomes. Comparative results can be examined through a highly interactive Graphic User Interface (GUI) featuring the visualization of the level of conservation in the format of a continuous VISTA curve based on the conservation in a sliding window. This concept proved to be extremely successful due to the easy interpretation of the resulting plots.

A novel stand-alone software GenomeVISTA integrates all well-established popular features of the VISTA family of tools (Dubchak, et al., 2009; Frazer, et al., 2004) in one package, and provides users the opportunity to carry out comparative analysis of whole genomes on their own computers allowing for more flexibility and security of computations. It runs an extensively tested and recently improved alignment algorithm (Dubchak, et al., 2009; Earl, et al., 2014). Simultaneously, the built-in interactive GUI allows for real-time examination of results of the comparative analysis. VISTA Point, a novel visualization program, is provided as a part of the GenomeVISTA package.

2 IMPLEMENTATION
Architecture. The whole-genome alignment pipeline is a combination of Perl and C/C++ programs, and MySQL relational database to store both input genomic sequences and generated alignments. The pipeline uses the open source BLAT program (Kent, 2002) to obtain local hits. The interactive GUI for data input and the examination of results was written in Java. GenomeVISTA can be run on any major platform - Windows, Mac OS X, and Linux. In its minimal setup it

*To whom correspondence should be addressed.
requires only a single machine, but it can also be configured to utilize a computer cluster using SGE/UGE, Condor, or Torque batch systems.

**Design.** Figure 1 shows the workflow of pairwise and multiple whole genome alignment computations performed by GenomeVISTA. In the pairwise alignment, the local anchors between all sequences are computed using BLAT, which is run in a translated DNA mode, indexing all 5-amino acid words. Then, Supermap (Dubchak, et al., 2009), the fully symmetric whole-genome extension to the original Shuffle-LAGAN chaining algorithm (Brudno, et al., 2003), is used to obtain a map of large blocks of conserved synteny between the two species. Finally, regions of conserved synteny are aligned using Shuffle-LAGAN. The major differences for the multiple alignment pipeline are: 1) the use of PROLAGAN, which is a variation of the original Multi-LAGAN program (Brudno, et al., 2003) that allows for the alignment of two alignments (profiles), and 2) an additional step of predicting ancestral contigs using a maximum matching algorithm. The four stages (local hits, chaining, global alignment, and ancestral reconstruction) are repeated for every node in the phylogenetic tree.

**Runtime.** Estimated runtimes for GenomeVISTA depend on the length and the number of genomic regions submitted to the program. It varies from several minutes to several hours for genomes from 1 to 50 MB long (see Supplementary Table 1). We recommend utilizing a computer cluster for improved run times.

**Output.** GenomeVISTA provides users with an interactive Graphical User Interface (GUI) similar to VISTA Point used for analysis and visualization of alignments in all on-line VISTA applications. It displays a level of conservation in the format of a conventional VISTA plot and allows for an interactive change of parameters, such as level of conservation and resolution of a plot. It also gives convenient access to all data used and produced in the alignment (Figure 1).

### 3. DISCUSSION

GenomeVISTA unifies in one package multiple capabilities necessary to carry out various types of comparative analysis of genomic sequences and whole-genome assemblies. It aligns sequences both in finished and draft format thus allowing to use it for multiple application such as genome assembly, mapping newly sequence reads on the reference genome, calculating syntenic regions on complete genome assemblies, and many others. Importantly, it also gives access to the results of the alignment through a highly interactive interface that makes comparative analysis of genomic data fast and efficient.

### ACKNOWLEDGEMENTS

The authors are grateful to all VISTA developers, collaborators and users for support and suggestions for its ongoing development.

**Funding:** National Heart, Lung and Blood Institute, National Institute of Health, Grant R01GM081080A. 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).

---

**REFERENCES**


Kent, W.J. (2002) BLAT--the BLAST-like alignment tool, **Genome Res.** 12, 656-664.


