Title: BioHDF: Open binary file formats for large-scale data management – Project Update

Authors: Todd M. Smith [1], N. Eric Olson [1], Mark Welsh [1], Mike Folk [2]

Affiliations: [1] Geospiza, Inc. 100 West Harrison St. North Tower #330, Seattle WA 98119
[2] The HDF Group, 1901 S. First St., Suite C-2 Champaign IL 61820

Presenting Author: Mark Welsh – markw@geospiza.com

Project URL: http://www.hdfgroup.org/projects/bioinformatics/

Download URL: http://www.hdfgroup.org/projects/bioinformatics/bio_software.html

Licensing: BSD-style licensing for HDF5 libraries and BioHDF data access software;
GPL licensing for BioHDF data analysis software.

Abstract: The huge volume of data produced by novel sequencing technologies presents significant challenges around data transmission, storage, bioinformatics analysis, visualization, and archiving. Widespread adoption of Next Generation DNA Sequencing (NGS) will be hindered if bioinformatics software cannot scale to meet these challenges.

The BioHDF project is investigating the use of a mature technology for the storage and retrieval of very large scientific datasets – Hierarchical Data Format. Supported by The HDF Group, HDF5 provides a binary file format, a collection of APIs supporting data access (C-based, with bindings to Fortran and other languages), as well as comprehensive test cases to ensure the stability of this software. HDF has been in use for over 20 years in fields that have faced similar data challenges to those now being experienced in biology, such as aircraft flight test data (Boeing), Earth observation data (NASA, NOAA), and CAD/CAM data (EU).

BioHDF will extend HDF5 data structures and library routines with new features to support the high-performance data storage and computation requirements of Next Gen Sequencing. To enable its use within the bioinformatics and research communities, BioHDF will be delivered as an Open Source technology that will include: a data model supporting storage of sequences, their alignments against reference data sources, and annotations such as SNP or splice variation analysis; indexing for fast random-access into this data; analysis tools that target specific applications such as RNA-Seq, Tag Profiling, novel micro-RNA identification, and many others on multiple NGS platforms including Applied Biosystems SOLiD, Illumina Genome Analyzer, Roche 454, and Helicos; visualization tools that provide reporting and interactive display of these very large data sets.

Initial prototyping of BioHDF has demonstrated clear benefits. Sequences and their alignments against reference data sources, which occupy 10s of Gb stored as highly redundant text-format files, are just a few percent of that size when stored in BioHDF's compressed, structured binary format. Indexing in BioHDF enables very rapid (typically, few millisecond) random access into these sequence and alignment datasets, essentially independent of the overall HDF5 file size. Additionally, through our prototyping activities we have identified key architectural elements and tools that will form BioHDF.