Title: BioHDF: Open binary file formats for NGS data – current status and future directions

Authors: Dana Robinson(derobins@hdfgroup.org) [1], Mark Welsh [2], Todd M. Smith [2], Mike Folk [1]

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

Project URL: http://www.biohdf.org/

Download URL: http://www.biohdf.org/biohdf_downloads.html

Licensing: BSD-style licensing

Abstract:

The huge volume of data produced by the latest generation of sequencing technologies presents significant challenges in 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 (http://www.biohdf.org) aims to solve some of these data storage and manipulation challenges by using the established, open-source HDF5 (http://www.hdfgroup.org) binary file format to store NGS data.

BioHDF extends HDF5 data structures and library routines with new features to support the high-performance data storage and computation requirements of next-generation sequencing. The open-source, BSD-licensed tools support the storage of sequences, their alignments against reference data sources and annotations such as SNP or splice variation analysis. Multiple NGS platforms are supported including Applied Biosystems SOLiD, Illumina Genome Analyzer, Roche 454, and Helicos.

Recent developments in the BioHDF toolset include NCList-like indexing(Alekseyenko and Lee 2007) for correct, efficient retrieval of gene annotation and alignment data and the ability to read and write SAM data(Li, Handsaker et al. 2009) for easier integration into existing toolchains. Our roadmap for the next year is also presented, including the new BioHDF API and scalability/performance improvements.

Alekseyenko, A. V. and C. J. Lee (2007). "Nested Containment List (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases." Bioinformatics **23**(11): 1386-1393.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.