# SeqWare Query Engine: Storing and Searching Sequence Data in the Cloud

Brian D. O'Connor[1], Jordan Mendler[2], and Stanley F. Nelson[2]

1. Lineberger Comprehensive Cancer Center, UNC, Chapel Hill, NC, brianoc@email.unc.edu
2. Department of Human Genetics, UCLA, Los Angeles, CA

website: http://seqware.sourceforge.net
code: http://sourceforge.net/projects/seqware/develop
license: GPLv3

The SeqWare Query Engine project was started to provide a scalable means to store and query large amounts of data derived from massively parallel sequencing technologies. This project provides two components, first, a backend to store variant, coverage, and other key information in a scalable database cluster and, second, a web service to provide both a programmatic and web-based interface to query and filter this data. The SeqWare Query Engine backend is built using the open source HBase project which offers a truly distributed, NOSQL database that is capable of supporting thousands of genomes. HBase, as a Hadoop subproject, has tight integration with the Map/Reduce framework which provides a convenient and robust mechanism to traverse and process entries in the query engine database. The query engine webservice was written in Java using the RESTlet framework which provides both a user interface and a RESTful API for programmatic access. The SeqWare Query Engine has been successfully deployed on a small (4-8 node) HBase cluster within our group and is a key part of sequencing projects currently underway at UCLA as well as UNC. The SeqWare project (which includes SeqWare Query Engine) is freely available under the GNU General Public License (GPLv3) and has been written to allow for easy customization and community driven development through the project website (http://seqware.sourceforge.net).