The Goby framework: towards efficient next-generation sequencing data analysis

Nyasha Chambwe [1,2](nyc2003@med.cornell.edu), Kevin C. Dorff[1], Marko Srdanovic[1], Xutao Deng[1], Stuart J.D. Andrews[1,2], Fabien Campagne[1,2*] (fac2003@med.cornell.edu)
[1]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine; Weill Medical College of Cornell University, New York, NY 10021; [2]Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, NY 10021

Project URL: http://goby.campagnelab.org/
Source Code URL: http://campagnelab.org/software/goby/download-goby/
Open Source License: GNU General Public License

Next-Generation Sequencing (NGS) technologies continue to evolve rapidly, generating massive amounts of short-read sequence data. The sheer volume of NGS data introduces computational challenges in its processing and storage.

NGS projects often require the ability to store sequence reads, alignments, base-resolution histograms, and to record specific subsets of reads. Current file formats support each of these types of information. NGS reads are often stored as Fasta or Fastq format. Alignment formats encode information about where reads map to a reference sequence as well as sequence variations (e.g., SAM/BAM format[1]). Base resolution histograms can be used to represent read abundance at particular positions of the reference sequence (e.g., Wig/Bed format are often shown as tracks in genome viewers). Files that represent a subset of reads help mark reads with specific properties, such as reads that do not match a reference and are often stored as text files. Files encoded in these formats can become very large when analyzing datasets that contain tens of millions to billions of reads. This creates scalability issues in most NGS analysis pipelines.

We have developed the Goby software framework to address these problems and provide a test-bed for alternative file formats and algorithms. Goby combines Gzip compression with the open-source Google Protocol Buffers library (http://code.google.com/p/protobuf/). Protocol buffers are advantageous because they support multiple languages (i.e., Python, Java, C++ and others), are cross-platform, flexible, and extensible. For instance, they allow older versions of software to read newer versions of a file format (forward compatibility) or new versions of the software to read older versions of a file format (backward compatibility). Goby files are organized as chunks and support semi-random access in a very large file, which is useful to retrieve only slices of a read or alignment file for processing on a cluster. Goby file formats are precisely specified and are often significantly smaller than current standards. For instance, we find that Goby alignment files can be 2 to 5 times smaller than an equivalent alignment encoded in BAM format (compressed binary version of a SAM file). The size ratio depends on the lengths of the reads, the amount of sequence variation and the platform and can only be estimated empirically. We have tested the Goby file formats by constructing an RNA-Seq analysis pipeline. This pipeline supports multiple aligners, including the Burrows Wheeler Aligner[2] (BWA) and the Last[3] aligner and has been tested using data from the Sequencing Quality Control (SEQC) (http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.htm), from four major sequencing platforms. The talk will present compression ratios obtained with the Goby file formats and provide an overview of this Java software framework.

References:

1. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. Epub 2009 Jun 8
2. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.
3. Incorporating sequence quality data into alignment improves DNA read mapping Martin C. Frith, Raymond Wan, Paul Horton Nucleic Acids Research, 2010 (in press)