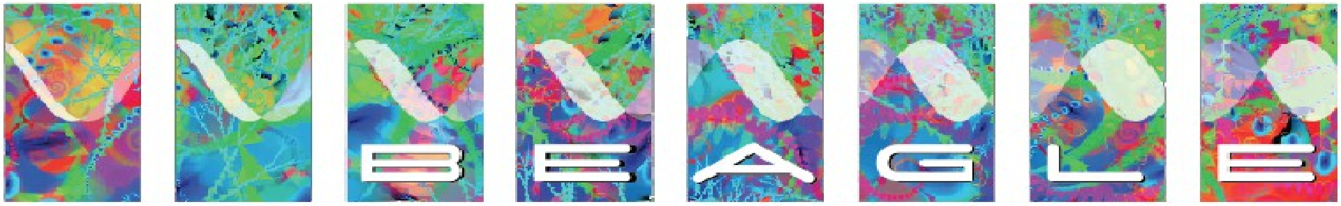
Volume 7, May 2014



Science Highlight

SwiftSeq: A High-Performance Workflow for Processing DNA Sequencing Data

Jason Pitt, Kevin White Lab
Committee on Genetics, Genomics, and Systems Biology, University of Chicago

Summary:

With increasing throughput and decreasing cost of next generation sequencing (NGS) technologies, investigators and consortia are now producing extraordinary amounts of genomics data. This transformation of genomics to a data-intensive science has left a significant portion of researchers unable to reap the benefits of their own data, let alone the troves available from public repositories. In the Kevin White lab, much of our data comes from The Cancer Genome Atlas, a project which strives to characterize the genomic properties of over 20 different cancer types. To date, this project has generated over a petabyte of compressed sequencing data. While significant hardware resources such as Beagle are necessary to perform large-scale analyses, insufficient software infrastructure is a substantial bottleneck to data-intensive genomics. Using Beagle as a test-bed, and in collaboration with the Computation Institute, we've developed SwiftSeq, a highly-parallel and scalable next generation DNA sequencing workflow underpinned by the parallel scripting language Swift (Figure 1).

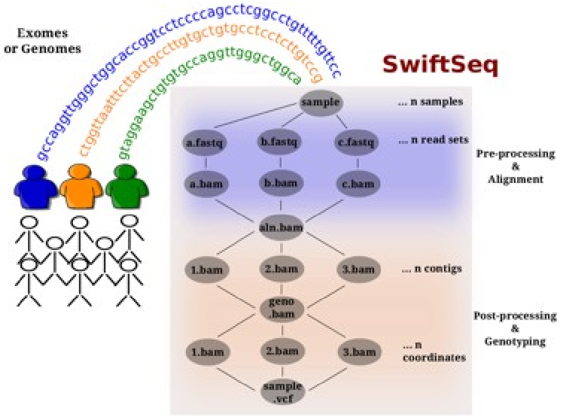


Figure 1. Schematic of the SwiftSeq workflow. By employing multiple map-reduce steps, an optimized job packing scheme, and the parallel nature of Swift, SwiftSeq provides both a fast and efficient way to process next-generation DNA sequencing data.

SwiftSeq is able to perform end-to-end analysis beginning with sequencing reads and ending with annotated genotypes. Independent samples, tumor-normal pairs, and either bam or fastq files are accepted as input. Samples are pre-processed, aligned, post-processed, and genotyped in a manner consistent with the 1,000 Genomes Project and Genome Analysis Toolkit's best practices. To maintain workflow flexibility, algorithms for key processes such as alignment and genotyping are interchangeable, which is critical given dynamic nature of genome analysis. Computational adversities such as execution site selection, parallelization, and synchronization are all managed under-the-hood by capitalizing on Swift's implicitly parallel framework. Executed tasks are robust to transient software and localized hardware failures, keeping user intervention to a minimum. Due to an optimized packing scheme, SwiftSeq requires less than a third of the computational resources necessary for standard pipeline-based approaches. Importantly, the workflow is easily portable and has been successfully deployed to clusters and Bionimbus clouds in addition to Beagle. Both members of the Beagle team (Lorenzo Pesce, Ana Marija Sokovic & Joe Urbanski) and the Swift developers have provided significant technical and development support to make this software possible. With their continued assistance, we hope to provide SwiftSeq as an analysis resource to the University of Chicago as well as the genomics community at large.

Resouces

Beagle Wiki
Get detailed usage information from the Beagle team

Beagle Support
Contact the Beagle experts for help

Globus Online for file transfer
Get started moving files to/from Beagle using this fast service

Other CI resources
Learn about other computing resources available at the Computation Institute

Scientifically, SwiftSeq has become a workhorse for many collaborative projects such as whole genome sequencing of Nigerian individuals with breast cancer (Funmi Olopade, University of Chicago), exome sequencing of autism families (Andrey Rzhetsky, University of Chicago), and the ENCODE-Cancer working group. To date, over 5,500 germline exomes from The Cancer Genome Atlas have been uniformly processed with SwiftSeq using approximately 4.8 million core hours. By integrating these exonic genotype calls with somatic copy number alterations, we've shown that inherited and acquired DNA changes can interact to reveal novel tumor suppressor genes. Using cells engineered to carry our mutations of interest, we're performing experiments to validate if these aberrations can indeed induce cancer phenotypes. With SwiftSeq and Beagle we intend to continue discovering and validating novel genomic features, as well as exploring biological questions that would otherwise be computationally infeasible.

Please send us your most recent publications made using Beagle!

Feel free to send papers that you are not sure you might have sent already, we can deal with duplications. We use these papers when we work on allocations and related matters..

Beagle Related Publications

R. Bamba, J.M. Lorenz, A.J. Lale, B.S. Funaki, S.M. Zangan. Clinical Predictors of Port Infections within the First 30 Days of Placement. J Vasc. Interv. Radiol. 2014;25:419–423

L.L. Pesce, H.C. Lee, M. Hereld, S. Visser, R.L. Stevens, A. Wildeman, W. van Drongelen, "Large-Scale Modeling of Epileptic Seizures: Scaling Properties of Two Parallel Neuronal Network Simulation Algorithms," Computational and Mathematical Methods in Medicine, vol. 2013, Article ID 182145, 10 pages, 2013.

Y. Meng, B. Roux. Locking the active conformation of c-Src kinase through the phosphorylation of the activation loop. J Mol Biol., 426(2), 2014

Y. Lin, B. Roux. Computational analysis of the binding specificity of Gleevec to Abl, c-Kit, Lck, and c-Src tyrosine kinases. J Am Chem Soc. 135(39), 2013.

J. Elliott, D. Deryng, C. Müller, K. Frieler, M. Konzmann, D. Gerten, et al. Constraints and potentials of future irrigation water availability on agricultural production under climate change. Proceedings of the National Academy of Sciences of the United States of America, PNAS.111(9):3241 (2014).

J.C. Gumbart, Morgan Beeby, G.J. Jensen, B. Roux. Escherichia coli Peptidoglycan Structure and Mechanics as Predicted by Atomic-Scale Simulations. PLOS Computational Biology.10(2), (2014)

A.M. Fluitt, J.J. de Pablo, "Atomistic simulation of polyglutamine in solution." American Institute of Chemical Engineers Annual Meeting. San Francisco, CA. (11/6/2013).

T.G. Armstrong, J.M. Wozniak, M. Wilde, I.T. Foster, Compiler Optimization for Data-Driven Task Parallelism on Distributed Memory Systems, Argonne National Laboratory Tech Report ANL/MCS-P5080-0214, 2014.

T.G. Armstrong, J.M. Wozniak, M. Wilde, I.T. Foster, Compiler Optimization for Extreme-Scale Scripting, To appear at CCGrid 2014 Doctoral Symposium, 2014

Searle Chemistry Laboratory
5735 South Ellis Avenue
Chicago, IL 60637

Help E-Mail:
beagle-support@ci.uchicago.edu
Beage CI Website:
beagle.ci.uchicago.edu