Whole Genome Resequencing Analysis in the Clouds

Michael C. Schatz[1], Ben Langmead[1, 2], Jimmy Lin[3], Mihai Pop[1], Steven L. Salzberg[1]

[1] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[3] The iSchool, College of Information Studies, University of Maryland, College Park MD, USA

http://bowtie-bio.sf.net/crossbow

Abstract

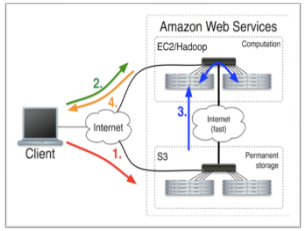
As growth in short read sequencing throughput vastly outpaces improvements in microprocessor speed, there is a critical need to accelerate common tasks, such as short read alignment and SNP calling, via large-scale parallelization.

Crossbow is a software tool that combines the speed of the short read aligner Bowtie[1] and the accuracy of the SOAPsnp[2] consensus and SNP caller within a cloud computing environment. Crossbow aligns reads and makes highly accurate SNP calls from a dataset comprising 38-fold coverage of the human genome in under 1 day on a local 40 core cluster, and under 3 hours using a 320-core cluster rented from Amazon's Elastic Compute Cloud[3] (EC2) service. Crossbow's ability to run on EC2 means that *users need not own or operate an expensive computer cluster in order to run Crossbow*. Crossbow is available at http://bowtie-bio.sf.net/crossbow under the Artistic license.

Crossbow Design

Crossbow builds upon a parallel software framework called Hadoop[4]. Hadoop is an open source implementation of the MapReduce programming model that was first described by scientists at Google[5]. Hadoop has become a popular tool for computation over very large datasets, used at companies including Google, Yahoo, IBM, and Amazon. Hadoop requires that programs be expressed as a series of Map and Reduce steps operating on tuples of data. Though not all programs are easily expressed this way, Hadoop programs gain many benefits. In general, *Hadoop programs need not deal with particulars of how work and data are distributed across a cluster or how to recover from failures*. Hadoop handles this.

The insight behind Crossbow is that alignment and SNP calling can be framed as a series of Map, Sort and Reduce steps. The Map step is short read alignment, the Sort step bins and sorts alignments according to the genomic position aligned to, and the Reduce step calls SNPs for a given partition.

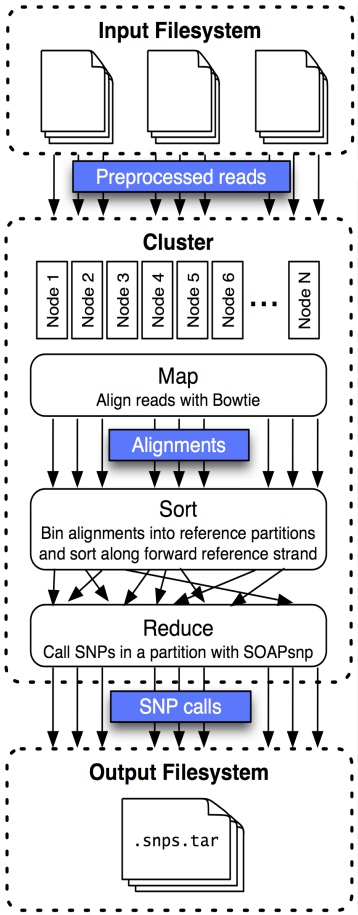


Steps involved in running Crossbow using Amazon's EC2 and S3 services

References

1. Langmead B, Trapnell C, Pop M, Salzberg SL. **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 10 (3): R25.
2. Li, R., Y. Li, et al. (2009). **SNP detection for massively parallel whole-genome resequencing**. *Genome Res* 19 (6): 1124-32.
3. http://aws.amazon.com/
4. http://hadoop.apache.org
5. Dean, J. and Ghemawat, S. 2008. **MapReduce: simplified data processing on large clusters**. Communications of the ACM 51, 1 (Jan. 2008), 107-113.
6. Wang, J., W. Wang, et al. (2008). **The diploid genome sequence of an Asian individual**. *Nature* 456 (7218): 60-5.

Crossbow Flow



The user first uploads reads to a filesystem visible to the Hadoop cluster. If the Hadoop cluster is in EC2, the filesystem might be an S3 bucket. If the Hadoop cluster is local, the filesystem might be an NFS share.

A cluster may consist of any number of nodes. *Hadoop handles the details of routing data, distributing and invoking programs, providing fault tolerance, etc.*

Map step is short read alignment. Many instances of **Bowtie** run in parallel across the cluster. Input tuples are reads and output tuples are alignments.

Sort step bins alignments according to primary key (genome partition) and sorts according to a secondary key (offset into partition). *This is handled efficiently by Hadoop.*

Reduce step calls SNPs for each reference partition. Many instances of **SOAPsnp** run in parallel across the cluster. Input tuples are sorted alignments for a partition and output tuples are SNP calls.

Results are stored in cluster's filesystem, then automatically archived and downloaded to the client machine. SNP calls are provided in SOAPsnp's format.

Whole-Human Resequencing with Crossbow

Crossbow was used to align and call SNPs from the set of 2.7 billion reads sequenced from a Han Chinese male by Wang *et al*[6]. Previous work demonstrated SNPs called from this dataset by SOAPsnp are highly concordant with genotypes determined via an Illumina 1M BeadChip assay of the same individual[2]. Reads were downloaded from a mirror of the YanHuang site (http://yh.genomics.org.cn). The reads cover the assembled human genome sequence to 38-fold coverage. They consist of 2.02 billion unpaired reads with sizes ranging from 25 to 44 bps, and 658 million paired-end reads. The most common unpaired read lengths are 35 and 40 bps, comprising 73.0% and 17.4% of unpaired reads respectively. Cost timing, and accuracy results are summarized below. SNPs produced by Crossbow exhibit similar agreement with the BeadChip calls as did the SOAPsnp study.

	1 Master 10 Workers	1 Master 20 Workers	1 Master 40 Workers
Worker CPU cores	80	160	320
Wall clock time	6h:30m	4h:33m	2h:53m
Cost	$61.60	$84.00	$98.40

Simulated	True # sites	Crossbow sensitivity	Crossbow precision
Human Chr. 22, simulated SNPs	46,586	99.01%	99.14%
Human Chr. X, simulated SNPs	102,219	98.97%	99.64%
Real	**# sites**	**Autosomal agreement**	**Chr. X agreement**
Whole human, versus Illumina 1M BeadChip	1.04M	99.5%	99.6%

Simulated reads are paired-end reads with simulated SNPs, including known HapMap SNPs, which SOAPsnp handles specially, and novel SNPs. "Real" reads are the reads from the Wang *et al* study. Agreement is calculated as correct calls at genotyped sites divided by number of genotyped sites.

Crossbow is a new software tool for efficient and accurate whole genome genotyping. Crossbow aligns and calls SNPs from 38-fold coverage of short reads from a human in less than 3 hours on a 320-core cluster rented from Amazon's EC2 service. **Crossbow condenses over 1,000 hours of resequencing computation into a few hours without requiring the user to own or operate a computer cluster**. Running on standard software (Hadoop) and hardware (EC2 instances) makes it easier for other researchers to reproduce our results or other results obtained with Crossbow. Crossbow is freely available from http://bowtie-bio.sf.net/crossbow