Supercomputing for the parallelization of whole genome analysis

Megan J. Puckelwartz1, Lorenzo L. Pesce2, Viswateja NelaKuditi3, Lisa Dellefave-Castillo1, Jessica R. Golbus1, Sharlene M. Day4, Thomas P. Cappola5, Gerald W. Dom, II6, Ian T. Foster2 and Elizabeth M. McNally1,3,*

1Department of Medicine, 2Computation Institute and Argonne National Laboratory, 9700 S. Cass Ave. Argonne, IL 60439, USA, 3Department of Human Genetics, The University of Chicago, 5841 S. Maryland Ave Chicago, IL 60637, USA, 4Department of Internal Medicine, The University of Michigan, 1150 W Medical Center Dr. Ann Arbor, MI 48109, USA, 5Perelman School of Medicine, Penn Cardiovascular Institute and Department of Medicine, University of Pennsylvania, 3400 Civic Center Blvd. Philadelphia, PA 19104, USA and 6Washington University School of Medicine, 660 S. Euclid Ave. St. Louis, MO 63110, USA

ABSTRACT

Motivation: The declining cost of generating DNA sequence is promoting an increase in whole genome sequencing, especially as applied to the human genome. Whole genome analysis requires the alignment and comparison of raw sequence data, and results in a computational bottleneck because of limited ability to analyze multiple genomes simultaneously.

Results: We now adapted a Cray XE6 supercomputer to achieve the parallelization required for concurrent multiple genome analysis. This approach not only markedly speeds computational time but also results in increased usable sequence per genome. Relying on publicly available software, the Cray XE6 has the capacity to align and call variants on 240 whole genomes in ~50h. Multisample variant calling is also accelerated.

Availability and implementation: The MegaSeq workflow is designed to harness the size and memory of the Cray XE6, housed at Argonne National Laboratory, for whole genome analysis in a platform designed to better match current and emerging sequencing volume.

Contact: emcnally@uchicago.edu

Received on December 2, 2013; revised on January 26, 2014; accepted on January 27, 2014

1 INTRODUCTION

With the advent of massively parallel DNA sequencing, the rate at which human genome variation can be determined is limited less by sequence generation but instead by the computational tools required to analyze these data. With current sequencing technology using short sequence reads of ~100 bp, whole genome analysis (WGA) requires the cleaning, aligning and interpreting of a billion sequence reads per single genome. With focus on scalability, we sought to improve the timeline required to process whole genome sequencing (WGS) by optimizing extraction, alignment, processing and variant calling. We reasoned that supercomputing capacity was better suited to parallelize WGA and allow for the rapid simultaneous analysis of multiple genomes.

Beagle is a Cray XE6 supercomputer housed at Argonne National Laboratory and administered by the Computation Institute at the University of Chicago. Beagle has ~726 compute nodes each with 32 GB of memory. Each node has 24 cores, 2.1 GHz cores on two AMD “Magny-Cours” processors. The XE6 can work in both Extreme Scalability Mode for scalable applications and Cluster Compatibility Mode for use with programs that are designed for smaller machines or clusters, such as the freely available genomics tools that are now routinely implemented for WGA (Li andDurbin, 2009; McKenna et al., 2010). While parallelization is possible on smaller systems, both memory and computational core number limit the capacity for simultaneous computation. Beagle uses a parallel computation environment and a parallel file system (Lustre) based on shared storage. Having both a parallel computation environment and external disk storage based on a parallel file system ensures that each node (and core) is able to access all data at any time without waiting for transfer of data across nodes. In this system, nodes have no local storage, and therefore no disk-to-disk transfer is required. On clusters without a shared file system, data transfer across nodes during analysis can be a time-intensive process. Here, we describe a workflow referred to as MegaSeq that uses the MapReduce (Dean and Ghemawat, 2008) approach to take advantage of supercomputing size and memory.

2 MATERIALS AND METHODS/RESULTS

2.1 Workflow-alignment

A summary of the MegaSeq workflow is shown in Figure 1. We adapted Beagle for WGA using a test dataset of 61 genomes that had been determined by Illumina Inc. Illumina provided WGS from 100 bp paired end reads as bam files for data transfer. FASTQ files were extracted from bam files using the Picard tool, SamToFastq (http://picard.sourceforge.net), and extraction was performed by readgroup (Table 1). The readgroup tag provides information on sample identity, library of origin and sequencing machine lane (see SAM Format Specification, samtools.sourceforge.net/SAM1.pdf). The extraction step is unnecessary if FASTQ sequence data are directly available. In the present cohort, each genome was represented by ~3–4 readgroups, creating a natural division of the reads. Alignment was performed with the Burrows-Wheeler Aligner (BWA) on 2^n nodes, with n equal to the number of readgroups (Li and Durbin, 2009). Each readgroup alignment was concurrent on 24 cores (Table 1). Alignment displayed a linear speedup and therefore
tools (Li et al., 2011). At this stage, the data are still split by chromosome. The efficiency of BaseRecalibrator is proportional to the number of base pairs provided. Therefore, indels within an individual genome may misalign. Because smaller datasets may have poor recalibration, splitting should be avoided with smaller datasets. On smaller datasets, the CRAY XE6 has the capability to perform recalibration across the genome in a reasonable time frame. Regardless of dataset size, recalibration results should always be reviewed by the end-user.

Table 1 details the programs and flags used for each operation. Each step was performed on 25 cores concurrently, with approximately six nodes dedicated to each genome (Table 1). On Beagle, running multiple genomes simultaneously is more efficient than running a single genome because of the structure of both Beagle and the genome. For the cleaning steps, each genome was run on 25 cores (24 chromosomes plus the mitochondrial genome). Four chromosomes were analyzed on each node, and therefore each genome needed six nodes, plus one core. During testing of the pipeline, we noted that node failure occurred as a result of java errors, often associated with memory allocation. To improve memory usage and avoid node failure, we hard coded the chromosomes into groups that are always sent out together. This scheme allows us to send the largest and most memory intensive chromosomes with smaller, less taxing ones resulting in fewer memory issues. For java run programs we used 28 GB memory/number of jobs per node. We also found that for java programs, using two threads for garbage collection better managed memory issues allowing us to pack more jobs per node. For each clean-up step, threading was used, where available (Table 1).

2.3 Workflow—variant calling

Variants were called using the HaploTypeCaller from GATK. HaploTypeCaller calls both single nucleotide polymorphisms and indels using de novo assembly of haplotypes in the active region. Haplotypes are evaluated using an affine gap penalty pair hidden Markov model (DePristo et al., 2011). Table 1 provides flags used for each step. MegaSeq identifies both single nucleotide variants (SNVs) and Indels simultaneously on ~6 nodes with 25× concurrency per genome. After variants were called and exported in variant call format (VCF), we used the GATK tool VariantFiltering to filter variants (DePristo et al., 2011). Variants were removed from the analysis using these criteria: biallelic balance >0.75; quality score <30; depth of coverage >360; strand bias more than −0.01 and mapping quality zero reads ≥10. Variants were then annotated using the default parameters of snpEff. snpEff is a fast variant annotation and effect prediction tool that is integrated with GATK (Cingolani et al., 2012).

2.4 Testing the workflow

The above workflow was tested using data from 61 human genomes. The starting FASTQ file size of each genome was ~300 GB, requiring ~18 TB of space to process all individuals simultaneously. Reads were aligned to NCBI reference genome 37.1 (hg19). We compared alignment output of MegaSeq with that produced by Illumina using the proprietary alignment/variant calling software Eland/Casava because this software is designed to efficiently perform WGA (Cox, 2007). MegaSeq alignment using BWA resulted in greater coverage with a mean coverage of 40.0× compared with 37.2× for ELAND/Casava’s alignment across all 61 genomes (paired t-test, P < 0.0004, Fig. 2a). The mean percent of the non-N reference genome covered was also greater with MegaSeq compared with Eland/Casava (98.7 versus 98.0) using MegaSeq versus Illumina (paired t-test, P < 0.0001, Fig. 2b). In total, 285,896,445 variants were called by MegaSeq across the 61 genomes identifying ~4.5 million variants per individual. To compare variants between MegaSeq and Illumina’s software, only variants with a quality score ≥30 were included. The mean number of SNVs called per genome differed between MegaSeq and ELAND/Casava data (3.670 × 10^6 versus 3.736 × 10^6, MegaSeq and ELAND/Casava respectively, paired t-test, P < 0.0001, Fig. 2c). ELAND/Casava also called more indels (536,853 versus 618,779, MegaSeq and ELAND/Casava respectively, paired t-test, P < 0.0001) (Fig. 2d). There was 88% concordance between MegaSeq and ELAND/Casava SNV calls and only 64.7% concordance for indels (Fig. 3a).
To estimate validity of the calls, we examined the normalized density of quality scores. Concordant SNVs, called by both MegaSeq and ELAND/Casava, had higher quality scores compared with non-concordant SNVs (Fig. 3b, MegaSeq, light purple; ELAND/CASAVA, dark purple). Non-concordant SNVs had lower quality scores, especially in the ELAND/Casava call set (Fig. 3b, MegaSeq, red; ELAND/Casava, blue). This same pattern was evident for indels (Fig. 3c). Non-concordant SNVs identified by MegaSeq have quality scores more closely resembling the concordant calls. These data indicate that the ELAND/Casava call set contained more low quality variants than the MegaSeq call set. We next compared depth of sequence reads called by MegaSeq or ELAND/Casava. Concordant SNVs called by both MegaSeq and ELAND/Casava have similar normalized depths, with the highest density of calls occurring between 35–40×. The non-concordant SNVs have markedly different depth distributions (Fig. 3d). SNVs identified solely by ELAND/Casava had the highest density of calls at ~20× depth. In contrast, those SNVs identified solely by MegaSeq had a depth density that more closely matched the concordant SNV distribution, with the majority of calls occurring between 30–35× depth. These data

<table>
<thead>
<tr>
<th>Step</th>
<th>Program</th>
<th>Module/Call</th>
<th>Input</th>
<th>Parameters</th>
<th>Output</th>
<th>Number per genome</th>
<th>Number of active cores/node</th>
<th>Number nodes</th>
<th>Concurrency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extract fastq</td>
<td>Picard 1.98</td>
<td>SamToFastq</td>
<td>bam</td>
<td>default</td>
<td>fastq</td>
<td>2(#RGs)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Alignment</td>
<td>BWA 0.5.9</td>
<td>bwa aln</td>
<td>fastq</td>
<td>-qtrim 15</td>
<td>sai file</td>
<td>2(#RGs)</td>
<td>24</td>
<td>2(#RGs)</td>
<td>24(#RGs)</td>
</tr>
<tr>
<td>Convert sai to sam</td>
<td>BWA 0.5.9</td>
<td>bwa sampe</td>
<td>sai</td>
<td>-T -X -P (BWA)</td>
<td>sam</td>
<td>2(#RGs)</td>
<td>~24</td>
<td>#RGs</td>
<td>24(#RGs)</td>
</tr>
<tr>
<td>Compression</td>
<td>samtools 0.1.18</td>
<td>view</td>
<td>sam</td>
<td>-bo (samtools)</td>
<td>bam</td>
<td>2(#RGs)</td>
<td>3</td>
<td>#RGs</td>
<td>1(#RGs)</td>
</tr>
<tr>
<td>Split</td>
<td>samtools 0.1.18</td>
<td>view/merge</td>
<td>bam</td>
<td>-h (preserve readgroup)</td>
<td>bam</td>
<td>25</td>
<td>24</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>Sort</td>
<td>Picard 1.98</td>
<td>SortSam</td>
<td>bam</td>
<td>-coordinate</td>
<td>bam</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25</td>
</tr>
<tr>
<td>Mark duplicates</td>
<td>Picard 1.98</td>
<td>MarkDuplicates</td>
<td>bam</td>
<td>-REMOVE_DUPLICATES false</td>
<td>bam</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (each with 3 threads)</td>
</tr>
<tr>
<td>Reorder</td>
<td>Picard 1.98</td>
<td>ReorderSam</td>
<td>bam</td>
<td>default</td>
<td>bam</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (GC*)</td>
</tr>
<tr>
<td>Identify indel realignment targets</td>
<td>GATK 2.7-1</td>
<td>GATK</td>
<td>bam</td>
<td>-T RealignerTargetCreator</td>
<td>intervals</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (GC)</td>
</tr>
<tr>
<td>Realign targeted intervals</td>
<td>GATK 2.7-1</td>
<td>GATK</td>
<td>bam</td>
<td>-T IndelRealigner</td>
<td>intervals</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (GC)</td>
</tr>
<tr>
<td>Base recalibrator</td>
<td>GATK 2.7-1</td>
<td>GATK</td>
<td>bam</td>
<td>-T BaseRecalibrator</td>
<td>csv</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (GC)</td>
</tr>
<tr>
<td>Print reads</td>
<td>GATK 2.7-1</td>
<td>GATK</td>
<td>bam</td>
<td>-T PrintReads</td>
<td>bam</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (GC)</td>
</tr>
<tr>
<td>Call variants</td>
<td>GATK 2.7-1</td>
<td>GATK</td>
<td>bam</td>
<td>-T Haplotype Caller</td>
<td>vcf</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (GC)</td>
</tr>
<tr>
<td>Filter variants</td>
<td>GATK 2.7-1</td>
<td>GATK</td>
<td>bam</td>
<td>-T VariantFiltration</td>
<td>vcf</td>
<td>25</td>
<td>4-GC</td>
<td>~6</td>
<td>25 (GC)</td>
</tr>
<tr>
<td>Annotate variants</td>
<td>snpEff 2.0.5</td>
<td>snpEff</td>
<td>vcf</td>
<td>default</td>
<td>vcf</td>
<td>25</td>
<td>4</td>
<td>~6</td>
<td>25 (GC)</td>
</tr>
</tbody>
</table>

Note: RG = readgroup. *GC = 2 threads used for java Garbage Collection.
omes on a single core to be reference, we predicted total CPU time required for WGA of 240 gen-
run times on a single core, a 3 (24-core) node cluster and compared them
the Beagle supercomputer.
suggest that these single genome times may reflect the maximum speed for
bam format and, as such, will not be a consideration for many users.
necessary when genomes are delivered as the compressed pre-aligned
Computational time is a major bottleneck in WGA. We calculated the
speed, memory speed and disk speed. These
clock speed. The SamToFastq step requires
2.1 GHz processor as 1701 h (0.20 years) for a single genome. This time
central processing unit (CPU) time that would be required on a single
2.5 Speed of analysis
Computational time is a major bottleneck in WGA. We calculated the
central processing unit (CPU) time that would be required on a single
2.1 GHz processor as 1701 h (0.20 years) for a single genome. This time
can vary based on clock speed, memory speed and disk speed. These
calculations are bound by disk and memory much more than by CPU
clock speed. The SamToFastq step requires ~48 CPU h, but is only
necessary when genomes are delivered as the compressed pre-aligned
bam format and, as such, will not be a consideration for many users.
Alignment scales perfectly and therefore cannot be easily accelerated. We
suggest that these single genome times may reflect the maximum speed for
the Beagle supercomputer.
To highlight the power of a parallel system, we calculated hypothetical
run times on a single core, a 3 (24-core) node cluster and compared them
with the CRAY XE6 run times. Using CRAY XE6 execution time as a reference, we predicted total CPU time required for WGA of 240 gen-
omeomes on a single core to be ~47 years. A total of ~11.8 years of CPU
time is required to complete the workflow for 61 genomes. A biological
computing cluster with 3 nodes can accelerate this time to ~7.2 months.
Using the MegaSeq workflow on Beagle, 61 genomes were analyzed in
50.3 real time hours (Fig. 4a). Based on the size and number of nodes
available, Beagle has the capacity to perform WGA on 240 genomes in
Fig. 2. MegaSeq identifies more usable sequence and fewer SNVs and
Indels per genome. WGA from 61 individual genomes was compared
between MegaSeq using BWA/GATK on the Cray XE6 and ELAND/
CASAVA from Illumina because ELAND/CASAVA is aimed at effi-
ciency (Cox, 2007). (a) The mean coverage for each genome was higher
with MegaSeq (light gray triangles) compared with ELAND/Casava
(black squares) (40.0 × for MegaSeq and 37.2× for ELAND/Casava
[paired t-test, P < 0.0004]). (b) The percentage of non-N genome covered
by MegaSeq (light gray triangles) was greater than ELAND/Casava
(black squares) for each genome (98.7 and 98.0, MegaSeq and
ELAND/Casava, respectively [paired t-test, P < 0.0001]). (c) The total
number of SNVs identified per genome with MegaSeq (light gray tri-
angles) was less than with ELAND/Casava (black squares) [3.670 × 10^6
and 3.736 × 10^6 for MegaSeq and ELAND/Casava, respectively, paired
t-test, P < 0.0001]. (d) MegaSeq (light gray triangles) identified fewer
indels compared with ELAND/Casava (black squares) in each genome
[mean number of indels 536 853 and 618 779 for MegaSeq and ELAND/
Casava, respectively, paired t-test, P < 0.0001]]

genes would need to be run consecutively as the number of nodes in
use is exhausted.
Multisample variant calling can also be performed using the MegaSeq
workflow. Multisample calling reduces the false discovery rate, but is
computationally intensive. We performed multisample calling on 61 gen-
omeomes using HaplotypeCaller from GATK with the flags noted in Table 1.
To take advantage of the size of Beagle, we split the genomes by chromo-
some, then further split chromosomes into overlapping subunits that
varied based on total chromosome size resulting in 2400 total subunits.
For 61 genomes, ~600 nodes were used with four jobs per node, requiring
~40 000 CPU hours. Calls were completed in ~16h, real time. A bio-
ological computing cluster with three nodes would require ~4.4 months real
time to perform multisample calling on 61 genomes using HaplotypeCaller.
We estimate that ~154 224 CPU h (~17.8 years) would be required to complete 240 genomes. We estimate that a 3
node cluster would take ~1.3 years to complete multisample calling on
240 genomes. In real time using 600 nodes with four jobs per node, we
estimate Beagle can complete multisample calling using HaplotypeCaller
on 240 genomes in ~2.5 days, making Beagle an excellent resource for
multisample calling. After calling the resulting 2400 vcf files were merged
and overlapping calls were removed.

2.6 Space management
Space constraints are another major hurdle in large volume WGA. We
estimate that approximately 1 TB of space per genome was needed to com-
plete the MegaSeq analysis (Fig. 4b). Although it is possible to discard
output at each successive step, there is still a large volume of data that were active at any single step. For instance, at the SamToFastq step, 18 TB was required for the FASTQ files. During the cleanup steps, each genome required ~85 GB of space; ~5 TB of space was needed to process each step for 61 genomes. Because each step requires computational time, discarding data from earlier steps can be costly if there is an analysis failure. These space requirements are generally not met by small computing clusters or by larger clusters with many users. The final merged VCF file for each individual genome required ~1 GB of space.

3 CONCLUSION

The development of next-generation sequencing technology has transformed the genetic analysis paradigm, from examination of the coding region of a handful of genes, to the current methods of interrogation of the entire coding region of the genome and, finally, to analysis of the entire 3 billion base pair genome. While cost and time are no longer major barriers to whole genome sequencing, data analysis and storage are major bottlenecks in using whole genome data (Majewski et al., 2011). Currently, whole genome sequencing using the Illumina HiSeq 2000 platform at moderate coverage (30–50x) yields >100 GB of data. This can be an overwhelming amount of data to process and store, and therefore many have turned to exome sequencing. Exome sequencing focuses on evaluating variation in the coding portion of the genome (~1–2%), and therefore provides smaller and less complex data to manage than whole genome sequencing. Exome sequencing is useful, as ~85% of recognized disease-causing mutations are located in protein coding regions of the genome (Majewski et al., 2011). However, this finding reflects a bias in our ability to evaluate non-coding variation.

Recently, great strides have been made to decipher the other 98% of genome sequence. The ENCODE project (Encyclopedia of DNA Elements) has assigned biochemical functions to 80% of the genome (Bernstein et al., 2012). These annotations will prove valuable tools in evaluating non-coding variation. Only whole genome sequencing can be used to interrogate non-coding variation, although the complexity of WGA has limited this possibility. Whole genome sequencing may also be better suited than exome sequencing to assess structural variation in genomes. Structural variants are emerging as important factors in human disease, making them an important factor in weighing the benefits of whole genome sequencing in relation to the challenges of computation (Snyder et al., 2010; Spielmann and Mundlos, 2013). Thus, whole genome sequencing may be the method of choice for many researchers if not for the tremendous computational bottleneck.

The deluge of genetic data is appropriate for high-performance computing and large-scale storage options (Koboldt et al., 2010). Sequence analysis includes read alignment to a reference genome, alignment clean-up and variant calling. A number of resources are freely available for analysis of genome sequencing, including BWA, GATK and snpEff. These tools can be used to align and call variants from a single genome by most laboratories, even those with limited computational experience and resources. However, high-throughput analysis of many genomes is significantly accelerated by parallelization and better meets the needs of the genetics community.

A common approach for analysis has relied on computing clusters, and more recently, cloud-based computing. By transitioning WGS to a supercomputing environment, we achieved high reliability with accelerated speed. One of the more cumbersome problems with clusters and cloud-based computing involves long wait times for data transfer between nodes (Zhao et al., 2013). The Cray XE6 supercomputing environment described here eliminates these wait times by using a parallel file system (Lustre) without creating a resource conflict bottleneck. A parallel file system, like Lustre, removes the need for tracking of data location, leaving only the issues of cache, RAM and disk hierarchy (Eijkhout, 2013). The demonstration that whole genome sequences can be aligned, cleaned and interpreted in parallel was achieved by using BWA/GATK, robust, publicly available software packages, in the Cray XE6 environment. Notably, this method uses the same software packages commonly used in computing clusters, but takes advantage of the Cray platform to parallelize the analysis. The ability to apply multisample variant calling significantly improves reliability and begins to extend analysis to beyond what is possible in a cluster environment. The application of the Cray XE6 has the capacity to analyze, in parallel, as many as 240 genomes in ~50h. This is a platform-dependent workflow that serves as proof of principle that genome analysis is greatly accelerated when performed on a supercomputer. More importantly, this work demonstrates that the publically available software currently in use for genome analysis is amenable to the supercomputing environment and can be installed as is on a CRAY XE6 and likely other systems, although we have not tested those systems. Currently, MegaSeq is available on the Beagle supercomputer at the University of Chicago.
The MegaSeq workflow backbone is based on bash shell instruction, and the submission subscripts are based on Portable Batch System (PBS) commands and are adaptable to other batch systems including Sun grid engine or SLURM, because the parallel logical structure of the workflow is compatible. Disk, memory and CPU usage will likely require optimization because of differences in machine design, which may affect bottlenecks and stability. The workflow should port directly with only minimal modifications to any Cray XE6, CRAY XC30 and related systems. The Beagle supercomputer is an NIH supported resource and provides an opportunity for large-scale genome projects. This computing application provides a format where human WGS can be rapidly analyzed relieving major constraint for better defining the range and utility of human genome variation.

Funding: This work was supported in part by National Institutes of Health through resources provided by the Computation Institute and the Biological Sciences Division of the University of Chicago and Argonne National Laboratory [S10 RR029030-01], and NIH AR052646, NIH HL61322, NIH NS072027, and the Doris Duke Charitable Foundation.

Conflict of Interest: Spouse receives patent royalties related to DNA sequencing (EMM).

REFERENCES


Cingolani,P. et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin), 6, 80–92.


