CUSTOM GENOTYPING CHIP FOR AFRICAN POPULATIONS

Allocation: Illinois/250 Knh
PI: Liudmila Sergeevna Mainzer[1]
Collaborators: Gerrit Botha[2], Victor Jongeneel[1], Ayton Meintjes[2], Nicola Mulder[2], Gloria Rendon[1]

[1]University of Illinois at Urbana-Champaign
[2]University of Cape Town

EXECUTIVE SUMMARY

Our goal was to aid in the design of a cost-effective genotyping chip that would capture the genetic diversity in populations of African origin, including African-Americans. This collaborative project was part of the H3Africa Custom Chip Design task force led by Zane Lombard at the University of Witwatersrand and Debo Adeyemo of the National Human Genome Research Institute. This work will enable the identification of genetic variation specific to African populations, improve understanding of the links between genotype and disease in people of African origin, and extend the principles of personalized medicine to these underserved populations. It will also permit deeper study of African genetic diversity, bringing important insights into the history and evolution of humans in general.

INTRODUCTION

Much of what is known about the genetics of diseases is based on people with European ancestry. The Consortium for Human Heredity and Health in Africa—H3Africa—aims to change that by promoting health research that takes into account the genetic diversity of African populations. Genomic variation plays a large role in disease predisposition and drug response. Thus, it is important to develop tools for genomic variant discovery specifically in people of African descent, who have been underrepresented in many worldwide genetic diversity measurement projects. The Genome Analysis Working Group from the H3Africa consortium partnered with H3ABioNet and the Wellcome Trust Sanger Institute to construct a genotyping chip to test for genomic variants found specifically in African populations. The chip will be a tool for rapid and inexpensive genotyping of individuals, to aid in the studies of human evolution and to identify genomic bases of disease. The data include publicly available sequence data from the 1000 Genomes Project and over 2,000 low-depth whole genome sequences from the Sanger Institute's African Genome Variation Project. An additional set of 348 samples was sent for deep sequencing (30X) at the Baylor College of Medicine thanks to a funding supplement from the National Institutes of Health. Blue Waters was used to perform genomic variant calling analysis on this set of 348 deeply-sequenced whole human genomes.

METHODS & RESULTS

Scientists at the University of Cape Town led by Nicola Mulder, together with Dr. Manj Sandhu's team at the Wellcome Trust Sanger Institute, developed the computational workflow to extract genomic variants from the 348 samples sequenced at Baylor. The workflow followed best practices [1], with additional steps ensuring quality control, robustness and mechanisms of recovery from failure added by collaborators at the University of Illinois. This workflow was instantiated on Blue Waters and used to analyze the data in six batches, ranging from eight to 87 samples. Production was completed in 51 days, producing genomic variant calls on the entire dataset of 348 individuals, and those calls are now being used to make the final design for the genotyping chip. The final product, the chip itself, will be used in biomedical research throughout the world.

A number of data management issues had to be resolved to make this possible. The Blue Waters team engaged in debugging data transfers from Baylor to Illinois and from Illinois to South Africa. The data-transfer challenges we faced led directly to learning and evaluating community data transfer tools, such as bbftp. Ultimately all data transfers were successfully completed using Globus, but our work to test the functionality and performance of additional tools will provide useful alternatives, should the need arise.

We also demonstrated, in a production-grade project, the capability of Blue Waters to **conduct high-throughput analysis** of human genomes. Extensive benchmarking data were collected, and the computational workflow was hardened with many quality control steps to ensure delivery of correct results. The code is posted on GitHub, to be shared with the community: https://github.com/HPCBio/BW_VariantCalling.

WHY BLUE WATERS

Several hundred deeply sequenced human samples is a lot of genetic data, which have a large disk footprint and take a long time to analyze. The available capacity of the University of Cape Town's computing cluster was not sufficient to process all of the data in the required time. Blue Waters enabled them to process this wealth of data in a timely manner, using nearly 250,000 node-hours, even with extensive use of backfill and accuracy discounts. The total disk footprint was nearly 600 TB, while only 200 TB total was available on the Cape Town cluster.

The Blue Waters team has been **instrumental** to the success of this project. **Extremely high uptime**, rapid resolution of failures (in one case under 20 minutes) and close collaboration on data transfers between the United States and South Africa have made the Blue Waters support team an **integral** part of the project. Without their involvement and professionalism, this work would have been very difficult to complete. In our case, the team was as important as the compute, storage, and networking resources.

INSTRUMENTING HUMAN VARIANT CALLING WORKFLOW AT SCALE

Allocation: Illinois/619 Knh
PI: Liudmila Sergeevna Mainzer[1]
Collaborators: Arjun Athreya[1], Subho Banerjee[1], Ravishankar K. Iyer[1], Victor C. Jongeneel[1], Volodymyr Kindratenko[1], and Zachary Stephens[1]

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

Whole genome sequencing and analysis are becoming part of the clinical standard of care. President Obama's 2015 Precision Medicine Initiative included genomics as an inextricable component in development of medical treatment and prevention strategies. Understanding the associated computational challenges is necessary in order to plan and construct the computing infrastructure that will support very high-throughput analyses in regional genomic sequencing centers across the country. The information we obtained in this project will help make such design recommendations, from relatively small clusters to large supercomputers. Specifically, we addressed problems associated with workflow scheduling, job management and recovery, file distribution, auto-archiving, and workflow scalability. We identified, documented, and resolved the bottlenecks associated with the large number of small files created by the workflow, saturated I/O bandwidth for part of the workflow, and potential for unbalanced data load on the file system. The resultant codes are available on GitHub at https://github.com/HPCBio/BW_VariantCalling.

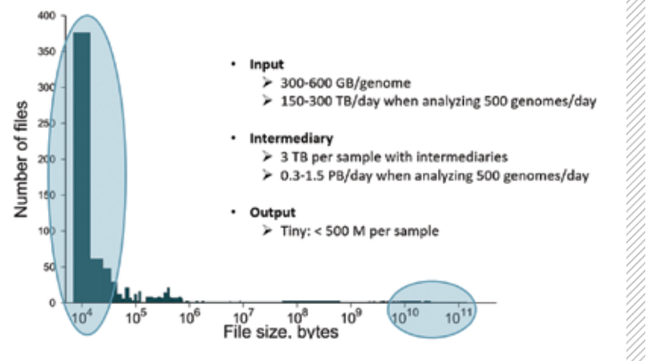

FIGURE 1: File size distribution for all the intermediate and output files produced in a variant calling workflow on one sample. Inset shows total disk footprint per sample, and per batch of 500 samples.

INTRODUCTION

Human variant calling (genotyping) is a process of searching for the differences between an individual's genome and that of the average in a human population. It is widely used to diagnose diseases and study human diversity and evolution. Genotyping every patient arriving at a given hospital is likely to become routine within our lifetime. This will require powerful computational facilities. Genotyping every newborn in Illinois would require analysis of ~500 genomes daily. At this scale, the standard workflow accepted by the community would use thousands of nodes in parallel and have computational bottlenecks affecting performance and reliability of the system. Our goal in this project was to identify such bottlenecks, construct workarounds, and come up with recommendations for the kind of computational/analysis systems that could handle this high throughput.

METHODS & RESULTS

We set up the "standard practices" workflow [1] and tested a number of tools [2,3] to shorten the wall time required to complete computation of each genome. We were able to accomplish a total run time of ~24 hours to process a single whole human genome, sequenced at the depth of 50X, using 25 nodes in parallel. In addition, we benchmarked the CPU, RAM, and I/O utilization across the workflow using Perfsuite [4], Cray Profiler [5], OVIS [6], Valgrind [7], and our own software. The data suggest two classes of problems are associated with genomic variant calling at-scale: workflow management and data management.

Workflow management: The workflow consists of multiple steps, each taking substantial time. It is therefore common practice to separate the steps into a series of independent jobs. When analyzing hundreds of genomes simultaneously, the generated deluge of jobs causes significant challenges for resource management and job scheduling systems. Indeed, on a system as large as Blue Waters, submitting more than 2,000 to 3,000 jobs at a time causes longer than desired scheduler planning cycles. For a big portion of the workflow, that would only accommodate analysis of 100 genomes, while we were targeting 500 genomes for simultaneous processing. To solve this problem, we worked with the Blue Waters team to incorporate a job launcher, which serves as an MPI wrapper around our OpenMP software. This adaption has been fully tested and works well.

Data management: Variant calling is a big-data workflow. When used on 500 deeply-sequenced, whole genome samples, it can create hundreds of thousands of files that use over a petabyte of disk space (Figure 1). The vast majority of these data are generated in the form of small files. Thankfully, we found that the Lustre file system on Blue Waters placed files evenly across disks, preventing a problematic load imbalance. Handling large numbers of files can strain the metadata servers and result in uneven performance. We worked with the Blue Waters team to create a parallel packaging utility, similar to TAR that could efficiently bundle files, thus preventing such issues.

Re-sorting intermediary files is common in a number of places along the workflow. The fastest software for it is Novosort [8], which involves two phases: sorting data in individual fragments, then merging the sorted fragments to produce the final output file. Ideally, the algorithm keeps all the intermediary fragments in RAM. When the available RAM is insufficient, it will write the fragments to disk, which can create enough I/O to saturate the bandwidth of the network routers and object storage servers. How many genomes can be feasibly analyzed in a single batch, before these bottlenecks begin to affect performance? We conducted Novosort scalability studies on three filesystems of different sizes (Figure 2) and concluded that the Blue Waters scratch filesystem can handle up to 1,000 genomes analyzed in parallel, without detrimental consequences. This is twice the required throughput for daily genotyping of all newborns in Illinois. The team reported the resultant I/O activity at the incredible rate of 4 TB/sec, likely due to re-reading from the Lustre cache, which functioned to compensate for the lack of RAM on the compute nodes (64 GB instead of the desired 256 GB).

Another I/O bottleneck resides in the alignment step of the workflow, when the algorithm mapping the sequencing reads against the reference attempts to access the input data on disk. Doing so on hundreds of samples at once engages the entire filesystem, forcing the processes to compete for disk access. We were able to overcome this bottleneck by striping the input data (width 3) and staggering the analysis in groups of 100 genomes. Configured this way, a massively parallel analysis of 500 genomes (30X coverage) completes in about 30-35 hours (Figure 3).

WHY BLUE WATERS

Our target problem size, 500 genomes, would use half the nodes and 1/20th of the file system daily on Blue Waters. The sheer size of the system made it possible for us to even consider such a problem. Our work demonstrated that Blue Waters can sustain **very high throughput** of genomic analyses without much impact on other users, and will remain very attractive for large genomics projects. The Blue Waters support staff have been instrumental in helping us figure out and eliminate issues with computational performance.

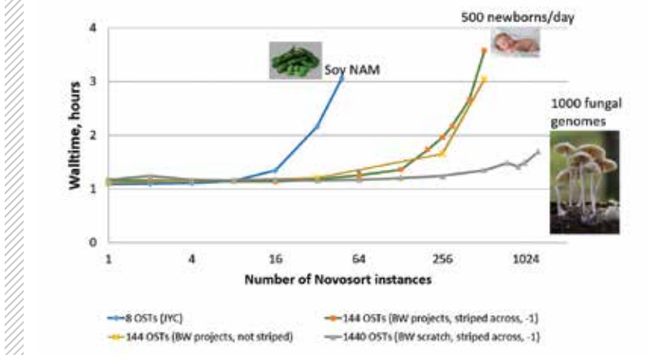

FIGURE 2: Scalability of Novosort on filesystems of different sizes. Images illustrate example problems feasible on each filesystem: 42 Soybean Nested Association Mapping (NAM) parents on the Blue Waters test cluster JYC, 500 newborns per day on /projects and up to 1,000 samples on /scratch (e.g. the 1,000 fungal genomes project). Legend lists the filesystem size as the number of object storage targets (OSTs).

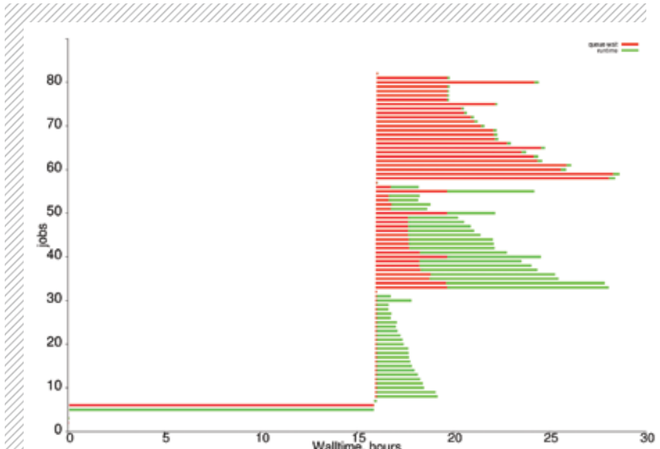

FIGURE 3: A Gantt chart displaying job execution in a single 100-genome batch. Each bar is a job. Red denotes waiting in the queue due to a job dependency, and green denotes the time of active execution. Five such batches staggered within a few minutes of each other permit 500 genomes (30X) to run in ~30-35 hours.