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.