HYPERION RESEARCH

HPC Profiles in Leadership

Custom Genotyping Chip for African Populations

Organizations: H3Africa and H3ABioNet

Contributors: Gerrit Botha, Ayton Meintjes, Nicola Mulder, Sumir Panji, Liudmila Mainzer, Gloria Rendon, Victor Jongeneel, Adebowale Adeyemo and Zané Lombard

August 2017

HYPERION RESEARCH OPINION

This computational project lead to the design of a cost-effective genotyping chip "*that can capture the genetic diversity in populations of African origin.*" Working under the leadership of Professor Nicola Mulder of the Human Heredity and Health in Africa (H3Africa) initiative and the H3ABioNet group working on the H3Africa Genome Analysis project, Gerrit Botha was asked to lead the alignment and variant calling part of the process. The work will enable the identification of genetic variations specific to African populations, which will help to clarify the links between genotype and disease in people of African origin by extending the principles of personalized medicine to these sometimes underserved populations.

Why a customized African genotype chip? Because Botha and the consortium are driven to sequence the African genome to represent African populations not well represented in other datasets used in the chip design process.

- African genomes are more diverse and have smaller haplotype blocks.
- A common variant in a specific European population might have a lesser frequency in an African population.
- African populations have higher average numbers of variant SNP sites (3.3 million SNPs per individual), compared with European (2.9 million) and Japanese/Chinese (2.8 million) individuals.

Botha and the consortium also believe that data gathered using the African genotype chip can be used to support planned future studies, including:

- Genetics of Rheumatic Heart Disease
- Susceptibility to Trypanosomiasis
- Burden, Spectrum and Aetiology of Type 2 Diabetes
- Kidney Disease
- Stroke Investigative Research
- Risk Factors for Cardiometabolic Disease
- Molecular Analysis of Tuberculosis
- Genetic Determinants of Febrile Illness

- Hereditary Neurological Disorders
- Nasopharyngeal Microbiome and Respiratory Disease
- Genomics of Schizophrenia

Note: this page is intentionally blank.

In This HPC Profiles in Leadership Report

Searching for the Origins of Disease: Creating a Customized Genotype Chip for Africans
Gerrit Botha, a bioinformatics engineer with the University of Capetown's Computational Biology Group, was part of a computational project that created a genotype chip designed to reflect the diversity of the African population and enable researchers to more precisely understand the genetic origins of disease there. Working under the leadership of Professor Nicola Mulder of the H3ABioNet group working on the H3Africa Genome Analysis project, Botha led the alignment and variant-calling part of the process.

Other teams contributing to the creation of the chip included the Wellcome Trust, Sanger Institute, and the University of Witwatersrand, as well as several data providers within the H3Africa consortium who provided samples for the design. Others who contributed to the chip design included Adebowale Adeyemo, Zane Lombard, Victor Jongeneel, Liudmila Mainzer, Gloria Rendon, Ayton Meintjes, and Sumir Panji.

For their work on this project using the Blue Waters supercomputer at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign, Botha won a Hyperion Research HPC Innovation Award in 2016.

SITUATION OVERVIEW

Blue Waters Runs Deep

One of the fastest and most powerful supercomputers in the world and located at the NCSA, Blue Waters is an open source petascale supercomputer created to be accessible to researchers in need of a machine with massive computational abilities. Designed by Cray and installed in 2012, it possesses more than 1.5 petabytes of memory; more than 25 petabytes of disk storage, and about 500 petabytes of tape storage.

Funded by the National Science Foundation and the University of Illinois, Blue Waters is capable of addressing myriad scientific problems – from simulating the origins of the cosmos to predicting the behavior of biological systems, to dealing with complex physics. Since it went into production in March 2013, it has provided 17.6 billion core hours to scientists and engineers – and that number is growing hourly, as the machine takes on additional work.

The NCSA describes itself as a 'hub of trans-disciplinary research and digital scholarship. "It is a place where University of Illinois faculty, staff, students, and collaborators from around the world come together to investigate some of the biggest issues and greatest challenges we face today, with the help of NCSA's advanced HPC technology."

- The NCSA's current research focus areas include bioinformatics and health sciences, computing and data sciences, culture and society, earth and environment, materials and manufacturing and physics and astronomy.
- The NCSA also provides integrated cyber-infrastructure such as computing, software, data, networking and visualization resources and expertise that are essential to the work of the scientists, engineers and scholars from the University of Illinois at Urbana-Champaign and other universities across the country and the world.

WHY HPC IS IMPORTANT TO GERRIT BOTHA AND THE H3 AFRICA PROJECT

Because of the sheer volume of the data and the complexity of storage and analysis of the information that was accruing in the H3Africa Genome Project, Botha realized it was time to search for alternatives. “It became necessary to use HPC when I got to the point where it was not realistic to store and analyze some of our genomic datasets on local desktops or laptops,” he said.

Botha reached out to HPCBio for advice and they suggested his team put in an application to work at Blue Waters for processing the Variant Calling of the H3Africa Set that was being analyzed at Baylor College of Medicine, one of the participants in the study. The sequencing, processing, and transfers were done in batches at Baylor.

Botha chose Blue Waters because he said the transfer issues would likely be less and the team already had a variant pipeline calling there that was proved and tested. “H3ABioNet consists of 32 nodes and one of the nodes is HPCBio in Illinois,” Botha explained. “The HPCBio team already had a robust alignment and variant calling pipeline built on Blue Waters, so the suggestion sounded promising,” he said.

Speaking at the Blue Waters Symposium in 2016, Botha said it was a “privilege to be exposed to such resources on such a scale,” referring both to the supercomputing capabilities of Blue Water and the Blue Waters Team. “On Blue Waters, we did the processing on the XE nodes,” Botha said. “We basically followed the GATK best practices for human variant calling and scripted the steps together with Bash. The steps involved running the software FastQC, bwa, samtools, GATK and verifyBamID. We also used the Anisimov launcher to pack single core jobs onto the multi-core nodes.”

JOINING A REVOLUTION: HOW HPC IS HELPING AFRICA GET HEALTHIER *SUCCESS STORY*

In the United States and Europe, precision medicine, which includes scanning patients’ genes to create treatment programs specific to their needs, is emerging as the preferred approach to treat many illnesses. But it is only recently that Africans and people of African descent have been able to join in this radical new way to approach disease. Most genetic screening studies have targeted people of European ancestry. There have been some studies of Asians, but only 3% of existing studies have considered people of African origin. There were already chips manufactured to support

genetic/disease studies, but for Africans, who are considered the most genetically diverse people in the world, there was little information.

Enter the H3Africa initiative, a program focused on fueling genetic research in Africa and which has now produced a chip that will eventually contain 2.5 million different genetic markers specific to the African population. To amass this data, researchers in the program collected thousands of samples from Africans to study genetic links for such blights as diabetes, kidney disease, sleeping sickness, tuberculosis and rheumatic heart disease.

A CHIP IS BORN

Using the data gathered to develop a useable chip required the innovative talents of researchers and bio-informatics engineers working under the umbrella of H3ABioNet -- a Pan-African bioinformatics network responsible for supporting H3Africa. Groups from H3Africa submitted blood samples from participants across Africa to Baylor College of Medicine (BCM), Botha said. The sequenced data was then transferred from BCM to Blue Waters using Globus Online. Botha said he also used local HPC facilities in South Africa to do some of the work.

Because of the complex efforts of diverse teams of researchers, Illumina announced in October, 2016 that it would be building the chip. The company, which says its goal is to “apply innovative technologies to the analysis of genetic function and variation,” indicated that it was creating an “array for genome-wide association studies” by the H3Africa initiative. At the time of the announcement, the company expected to have the chip available commercially in early 2017. In its announcement, Illumina’s Senior Manager of Market- Development Julie Collens praised the work of the H3Africa initiative and spoke about its importance.

“In total from the exomes and genomes, I believe they have discovered something like 100 million or more variants that are specific to the African populations,” she said in a company press release. Because the Illumina chip will represent a subset of those 100 million variants, she explained, “that basically does the best job of representing those different populations.” Collens also said that because the array will have multiple different populations or disease areas on it, “you can find regions of overlap or you can find commonalities that are really beneficial. It is a cost-effective way to begin addressing health disparities.”

In no small part because of this work, Africa has now been able to join the Genome Revolution.

PARTNERSHIPS AMID CHALLENGES

One of the biggest difficulties of the project was dealing with the time difference between the United States and South Africa, Botha recalled.

“It was challenging in terms of staying in sync and keeping things productive. There were some delays in terms of sequencing because of reagent issues and this in turn delayed runs on the cluster,” Botha said. “It took some coordinating to be sure that Blue Waters' implementation of the pipeline was consistent with the processing that was done on some of the other samples. Transferring the total data of about 140TB to South Africa took months, and it was very challenging keeping track of transfers and ensuring all data was intact. Looking back at things now we would probably ship a big enough storage device to transfer the data next time.”

But the difficulties helped create new partnerships and ways of working, Botha said. “The South African H3ABioNet node at UCT (CBIO) and the HPCBio node built a good working relationship during the life of the project,” Botha said. “The data transfer challenges helped us build a good relationship with the local network team (TENET). What we have learned from the setup at Blue Waters and working with their

system administrators was shared with the facilities locally and enabled us to improve them. For example, we know of a GO endpoint at one of the facilities that directly connects to our Lustre storage."

Why the Research Is Important to the World

"Current genotyping arrays are sub-optimal for African studies," Botha said. "A genotyping chip that better captures linkage disequilibrium (LD) structure of the African genome would allow for better research. Since all populations are derived from Africa, the chip should be applicable to other populations."

Future Outlook

Future Research Plans

"The genotype chip is currently in manufacturing," Botha said. "What we need to work on now is the reference panel that will be used in combination with the chip data to impute full genome data. We are thinking of providing some kind of imputation service. On top of that I also have some other human variant calling and microbiome projects that keep me busy."

Future HPC Plans

Botha says he definitely plans to use HPC facilities in the future for many of the projects he will be working on.

What Botha Would Like to See Improved the Most in The Future

As projects get larger and more ambitious, Botha sees the need to improve the speed and storage capacities of HPC systems. "The sequencing projects will only get larger," he said. "We need to think of better ways to transfer and process the data. This includes possibly processing the data where it is sequenced, setting up a better software stack, for example by using something such as Docker or doing better management of data and transfers.

"The work that I'm doing needs HPC resources but it falls more into the category of high throughput processing (not MPI based). So you run many jobs and the duration of the job sometimes can exceed the default cluster of wall clock times. There are sometimes ways around it, as facilities have different infrastructure, schedulers, and resource managers."

Essential Guidance To Other HPC Sites

Best Practices and Advice

Botha says some important things learned from the setup at Blue Waters are:

- Have a GO endpoint that is directly attached to processing and archival space.
- Have an archival space such as nearline available for researchers.
- Have a portal where you as a user can view various resources and run statistics of your project
- Have a help desk system with a team that is responsive
- Have good documentation and keep users updated on any changes on the system.

About Hyperion Research, LLC

Hyperion Research, consisting of the former IDC high performance computing (HPC) analyst team, provides HPC information, analysis, and recommendations based on technology and market trends. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.hpcuserforum.com and www.HPCatHyperion.com
