1000 Genomes, 1 Approach: Analyzing Large Public Genomic Datasets using Cloud Computing

Maya Anand
Department of Computer Science, Columbia University
mva2112@columbia.edu

Abstract

As DNA sequencing becomes more affordable, there is an increased need for time- and cost-effective means of computing to process and analyze genomic data for research purposes. A large, publicly available dataset of anonymized genomic data was analyzed using new and existing programs, scaled horizontally with AWS. This approach demonstrates an effective means of analysis of big-data without need for a large amount of existing infrastructure.

Introduction

With the cost of DNA sequencing becoming more and more affordable (reaching the milestone of the $1000 genome), new time- and cost-effective means of computing are needed to analyze the wealth of information that has become available. Within the bioinformatics community, there are a variety of file formats and open-source tools being used to store and draw insights from genomic data. However, developing pipelines to work with the varied sources of data that are available can be a complex endeavor due to subtle differences between files from various sources and the lack of established methods for many tasks. The goal of this project was to extract samples from the a large public genomic dataset and process them in such a way as to augment an existing data set of the Erlich Lab. Open-source tools, in-house pipelines and the cloud computing infrastructure of Amazon Web Services were harnessed to create a reproducible series of scripts for this task.

The 1000 Genomes Project

The 1000 Genomes Project ran from 2008 until 2015 with the goal of sequencing at least 1000 individuals to discover and characterize over 95% of genetic variants with an allele frequency of 1% or higher in multiple major human populations[1, 2]. The project was completed in a three phases plus a pilot phase, with the Phase 3 release containing variants from 2504 individuals from 26 populations (these samples will be referred to as the 1KGP-Phase3 samples). A supplementary dataset includes 31 additional individuals related to some of the 1KGP-Phase3 samples (these samples will be referred to as the 1KGP-relatives). Finally, an additional dataset from the Phase 3 supporting data included an additional 595 individuals not included in the other

2 datasets (these samples will be referred to as the 1KGP-obs samples)[3].

Cloud Computing and Amazon Web Services (AWS)

Cloud computing involves on-demand IT resources accessed through the internet. Cloud computing eliminates the need for up-front investments in infrastructure and hardware as well as the need for an individual lab to maintain those resources. Cloud computing allows labs without a large amount of existing infrastructure to access the amount of resources needed for a specific analysis when it is needed and only pay for the amount of resources used. Amazon Web Services has been offering cloud computing services since 2006. In this project, I utilized Amazon S3 and Amazon EC2 for their horizontal scalability in order to process a large amount of data much more efficiently[4].

Amazon Simple Storage Service (Amazon S3) provides scalable data storage that allows read/ write access to the data from multiple clients or application threads. It allows for large amounts of data to be stored and accessed from multiple computing nodes, while only paying for the amount of storage used up to a virtually unlimited number of files[5]. Amazon Elastic Compute Cloud (Amazon EC2) provides resizable compute capacity in the cloud, allowing new server instances to be booted almost immediately so capacity can be scaled[6]. A specialized environment for the analysis of big data can be created using custom applications run on Amazon EC2 and scaled according to needs[7].

Methods and Results

1KGP-Phase3, 1KGP-relatives and 1KGP-obs data were downloaded from the 1000 Genome Project in the form of VCF (Variant Call Format) files which store information about a position in the genome on each data line[8]. To speed up future processing steps of the files, all VCF files were converted to a binary format, BCF. Using bcftools, individual samples were extracted as VCFs and BCFs for 1KGP-Phase3 and 1KGP-relatives[9]. The 1KGP-obs data, as downloaded from the 1000 Genomes Project, was not phased. The data was converted to Oxford GEN format for phasing with ShapeIt (currently in progress)[10].

Much of the genomic data available today, is in the form of 23andMe-like files generated by direct-to-consumer testing services. For the samples in this project, simulated 23andMe-like files were generated. These files generally include roughly 500,000 SNPs (Single Nucleotide Polymorphisms), however the data from the 1000 Genomes Project contained roughly 81 million SNPs per sample. To pick a smaller subset of these 81 million SNPs to include in the 23andMe-like file, a list of the top 500,000 SNPs most commonly observed in a subset of OpenSNP files

was regenerated and used in creating the 23andMe-like files.

The ancestry of each sample was estimated using Dr. Joe Pickrell's Ancestry program[11]. The 41 ancestry categories from this program were mapped back to the five super populations supplied by the 1000 Genomes project: AFR (African), AMR (Ad Mixed American), EAS (East Asian), EUR (European) and SAS (South Asian). The mapped super populations were compared to the (actual) super populations for each sample provided by 1000 Genomes to check the accuracy of the Ancestry algorithm and mapping. Over 90% of samples matched the categorization provided by the 1000 Genomes Project with the 10% not matching falling into the 1000 Genomes Project category of AMR, an interesting result which could warrant a closer look in future work.

PED files were created including all the samples and IBD (Identical by Descent) analysis was conducted using GERMLINE and ERSA[12, 13]. Our analysis was able to confirm most of the familial relationships between samples stated by the 1000 Genomes Project.

Finally, the 23andMe-like files for several samples were run through an in-house imputation pipeline using IMPUTE2[14]. The approximately 39 million SNPs returned by the imputation pipeline were compared to the original 81 million SNPs from the 1000 Genomes Project data to confirm the accuracy of the imputation. The lab plans on updating the pipeline in the near future to use a different imputation algorithm and the scripts for this analysis will be useful for benchmarking the difference in accuracy between the old and new algorithms.

Conclusion

The genomes of 2535 individuals from the 1000 Genomes Project were processed in using horizontal scaling with Amazon Web Services and the genomes of 595 additional individuals are still in progress of being processed. Genomes for each individual were extracted into individual VCF and BCF files. 23andMe-like files were generated for each individual by extracting 500,000 SNPs at loci commonly included in 23andMe files. Existing algorithms for ancestry detection and IBD (Identity by Descent) were run for each individual, confirming the relationships between trios (mother, father, child) identified by the 1000 Genomes Project. Through this analysis 9.3 TB of data were generated and stored using Amazon S3. The analysis used roughly 5,000 hours of Amazon EC2 instances, generally with machines with 16 or 40 CPUs each. This analysis demonstrates the use of horizontal scaling and cloud computing for the analysis of large quantities of genomic information without the need for maintenance of and investment in large amounts of computing infrastructure.

Acknowledgements

I would like to thank Dr. Yaniv Erlich, Assistant Professor of Computer Science at Columbia University and Core Member at the New York Genome Center, for his guidance throughout my research. This work would not have been possible without the tremendous help of Assaf Gordon, a staff programmer at the New York Genome Center. Finally, I would like to express my gratitude to DREU (Distributed Research Experience for Undergraduates) for facilitating and funding this research experience.

References

1. The 1000 Genomes Project Consortium (2010) A Map of Human Genome Variation from Population-Scale Sequencing. Nature. 467(7319): 1061–73. doi:10.1038/nature09534.

2. The 1000 Genomes Project Consortium (2007) "Meeting Report: A Workshop to Plan a Deep Catalog of Human Genetic Variation." 1000genomes.org. http://www.1000genomes.org/sites/1000genomes.org/files/docs/1000Genomes-MeetingReport.pdf (accessed August 10, 2016).

3. The 1000 Genomes Project Consortium (2015) "A Global Reference for Human Genetic Variation." Nature 526 (7571): 68–74.

4. Amazon Web Services (Dec 2015) "Overview of Amazon Web Services" aws.amazon.com. https://d0.awsstatic.com/whitepapers/aws-overview.pdf (accessed August 11, 2016).

5. Amazon Web Services (Nov 2015) "AWS Storage Services Overview" aws.amazon.com. https://d0.awsstatic.com/whitepapers/Storage/AWS%20Storage%20Services%20Whitepaper-v9.pdf (accessed August 11, 2016).

6. Amazon Web Services "Amazon EC2 - Virtual Server Hosting" aws.amazon. https://aws.amazon.com/ec2/ (accessed August 11, 2016)

7. Amazon Web Services (Jan 2016) "Big Data Analytics Options on AWS" aws.amazon.com. https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS.pdf (accessed August 11, 2016).

8. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) “The variant call format and VCFtools” Bioinformatics. 27:2156–8.

9. Heng Li (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 27 (21): 2987-2993. doi: 10.1093/bioinformatics/btr509

10. O. Delaneau, J. Marchini, JF. Zagury (2012) A linear complexity phasing method for thousands of genomes. Nat Methods. 9(2):179-81. doi: 10.1038/nmeth.1785

11. Pickerell, J (2016) Ancestry. unpublished. https://bitbucket.org/joepickrell/ancestry.git

12. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I (2008) Whole population, genomewide mapping of hidden relatedness. Genome Research.

13. Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, Tuohy TM, Neklason DW, Burt RW, Guthery SL, Woodward SR, Jorde LB (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Research.

14. IB. N. Howie, P. Donnelly, and J. Marchini (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5(6): e1000529