This is a repository copy of *Scalability and cost-effectiveness analysis of whole genome-wide association studies on Google Cloud Platform and Amazon Web Services*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/164194/

Version: Published Version

AMIA

INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Brief Communications

Scalability and cost-effectiveness analysis of whole genome-wide association studies on Google Cloud Platform and Amazon Web Services

Inès Krissaane, Carlos De Niz, Alba Gutiérrez-Sacristán, Gabor Korodi, Nneka Ede, Ranjay Kumar, Jessica Lyons, Arjun Manrai, Chirag Patel, Isaac Kohane, and Paul Avillach (iD)

Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

Corresponding Author: Paul Avillach, Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston 02115, MA, USA; paul_avillach@hms.harvard.edu

ABSTRACT

Objective: Advancements in human genomics have generated a surge of available data, fueling the growth and accessibility of databases for more comprehensive, in-depth genetic studies.

Methods: We provide a straightforward and innovative methodology to optimize cloud configuration in order to conduct genome-wide association studies. We utilized Spark clusters on both Google Cloud Platform and Amazon Web Services, as well as Hail (http://doi.org/10.5281/zenodo.2646680) for analysis and exploration of genomic variants dataset.

Results: Comparative evaluation of numerous cloud-based cluster configurations demonstrate a successful and unprecedented compromise between speed and cost for performing genome-wide association studies on 4 distinct whole-genome sequencing datasets. Results are consistent across the 2 cloud providers and could be highly useful for accelerating research in genetics.

Conclusions: We present a timely piece for one of the most frequently asked questions when moving to the cloud: what is the trade-off between speed and cost?

Key words: whole genome, genome-wide association study, cloud computing, distributed systems

INTRODUCTION

As datasets become increasingly larger and more abundant, science faces a new challenge: how to overcome the economic and technological barriers that arise when trying to store and analyze the data generated by large sample sizes. Every year, the scale of available genomic variant datasets nearly doubles.[1–3] This has led to a recent broad interest in genomics analyses using cloud computing.[4–6] For example, investigators have launched a new large-scale initiative, called the Trans-Omics for Precision Medicine (TOPMed) program, as part of the Precision Medicine Initiative. TOPMed focuses on the integration of thousands of whole genomes[7,8] gathered across several studies. The processing of such large amounts of data[9,10] is un-

precedented and requires significant funding for both storage and computation.

A solution, perhaps the only sustainable one currently available, is cloud-distributed computing systems.[11–14] Because the costs of such a solution remain obscure for common genomic operations, many investigators remain tentative or unsure of the suitability of cloud computing for their purpose; therefore, we undertook this study to clarify those costs.

We present an adaptable and reproducible method to deploy Spark clusters using Hail, an open-source, scalable framework for exploring and analyzing genetic data, as well as variant storage. We also utilized the cloud service Google Dataproc in the Google Cloud

Table 1. Whole-genome sequencing datasets description used to conduct the genome-wide association study

Project releases	VCF file size in GB	MT file size in GB	SNPs	Samples
1KG Phase 1	1231	250	38 248 779	1092
1KG Phase 3	853	12	77 253 690	2535
COPD Freeze 4	52[a]	102	69 023 355	1886
Jackson Freeze 5	29[a]	34	74 623 050	3406

1KG: 1000 Genomes Project; GB: gigabytes; MT: Matrix Table; SNP: single nucleotide polymorphism; VCF: variant call format.

[a]Compressed VCF file size.

Platform (GCP) and the cloud platform Amazon Elastic MapReduce (EMR) in Amazon Web Services (AWS) for performing genomic variant analysis with whole-genome sequencing (WGS) data. Therefore, we offer a promising strategy to accelerate functional interpretation of genetic variants[15,16] and discover their association with human disease in particular for genome-wide association study (GWAS) analysis.[17–19]

In order to estimate the required computational infrastructure needed, we performed cost analyses of GWAS[20] using 4 different datasets from the 1000 Genomes Project[21,22] and the TOPMed WGS program. Our goal was to optimize and customize cloud resources to fit computation and storage needs. We further offered appropriate strategies for using cloud resources by assessing the best cluster configuration for a GWAS analysis based on total cost[23] and runtime.

MATERIALS AND METHODS

Study sample and variant calling format

For this study, we used 4 different WGS datasets from the 1000 Genomes Project and TOPMed project, as well as the COPDGene Study and Jackson Heart Study. First, phases 1 and 3 of the 1000 Genomes Project were publicly and readily available in Google cloud buckets (gs://1000-genomes on Google Cloud Storage and s3://1000genomes/in Amazon S3). Freeze 4 (COPDGene Study) and freeze 5 (Jackson Heart Study) obtained variant data in variant call format (VCF) files for every sample in a specific freeze. These corresponded to aggregate single nucleotide polymorphisms (SNPs) for each study. We combined VCF files using the function merge in bcftools[24] of the dbGap database (see Table 1 and Supplementary Appendix File 1). We imported VCF files and transformed them into a Hail Matrix Table object (.mt). We note that we found it advantageous to use .mt files in Hail, as they are written and read faster than VCF files (see Supplementary Appendix 1).

GWAS analysis

We chose the variable gender present in all 4 GWAS datasets: female vs male as case and control group. Though both Python and Jupyter notebook scripts applied to the 4 datasets, we executed the necessary steps one would use to perform a GWAS (see Figure 1). Specifically, we utilized genotype information (GT) for many genetic markers, divided upon chromosomes. We deployed the standard quality procedures for genomics data.[25] Then, we filtered the results based upon minor allele frequency of the most common SNPs representing more than 1%. We further checked for missing values. We then corrected for population structure by performing a principal component analysis from the Hardy-Weinberg normalized genotype call matrix[26] method. We conducted a logistic regression using as a covariate the 2 first principal components obtained previously to predict the gender (see Supplementary Appendix 1, part 1) with the genotype call (GT):

$$\text{Prob(Gender)} = \text{sigmoid}(\beta0 + \beta1\text{gt} + \beta2\text{pca}[0] + \beta3\text{pca}[1] + \varepsilon),$$
$$\varepsilon \sim N(0, \sigma2)$$

Results are plotted in a Manhattan plot (Figure 1), showing the significant SNPs (Wald test per variant). The horizontal line represents the significance threshold after Bonferroni correction (P value $\leq 5.0 \times 10^{-8}$).

Cloud deployment

In GCP, we executed a shell script to automate the process of deployment and deletion of our clusters (cluster creation code supported by the Hail team [https://github.com/Nealelab/cloudtools]). For AWS, we used an in-house script to manage EMR cluster generation. We used 2 cloud formation tools to create Spark cloud clusters with master node and several worker nodes: (1) Google Dataproc with the image 1.2-deb9 and Spark version 2.2.1 and (2) Amazon EMR 5.13 with Spark version 2.3.0. The use of 2 different cloud providers created only minor variation, in terms of hardware (instance characteristics), cluster configurations, and network connection.

However, as shown in the Results, the outcome and results are very consistent for both platforms. Both providers have similarities in terms of the computing environment, including number of central processing units (CPUs) per instance, storage, memory (random access memory [RAM]), networking, and operating systems. For the worker nodes, we performed GWAS with preemptible instances provided in the GCP and spot instances in AWS. These represented significant cost reduction, while at the same time meeting performance requirements (see Supplementary Appendix 1, part 4).

Hail cloud testing and workflow

For the sake of reproducibility, we decided to use standard instance types—preferably the most common and accessible—rather than customizing our own. In GCP, we varied only 2 parameters: the typical instance for worker nodes and the number of nodes. These directly impacted the total number of CPU and memory (in gigabytes [GB]) of the cluster. We tested 2 instance types: n1-standard and n1-highmem among those possible for a total of 6 different Google Cloud Engine (GCE) virtual instance machines. We used clusters, including 16 to 64 CPUs and 60 to 416 GB of RAM per worker nodes in GCP. For AWS, we used a cluster with worker nodes with 16 CPUs and 64 GB of RAM. We calculated the total cost of each cluster during end-to-end processing (from the instantiation to deletion). The total cost was calculated based on the prices applied by Google Cloud. These include the price per instance and product Google Cloud Dataproc. These rates were applicable to the North Virginia zone (January 31, 2019). The process described in Figure 2 was performed using a bash script that parallels the creation of all clusters
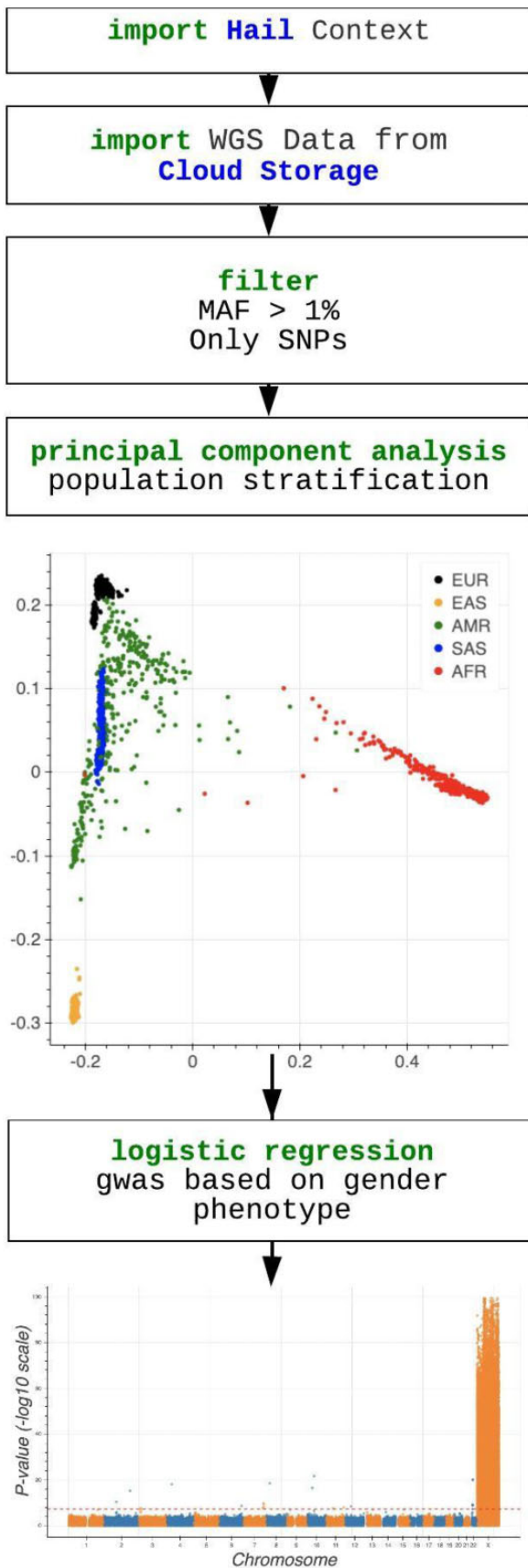
Figure 1. Overview of a genome-wide association study using Hail variant store using 1000 Genome Project dataset Phase 3. EUR, EAS, AMR, SAS and AFR designed the European, East Asian, American, South Asian and African populations.

and automates their deletion after the Hail operation finished. The code is available online (https://github.com/hms-dbmi/Hail-on-Google-Cloud/tree/master/Bash_script).

Availability and implementation
The workflows to deploy Hail cloud clusters are available online (https://github.com/hms-dbmi/Hail-on-Google-Cloud and https://github.com/hms-dbmi/hail-on-AWS-spot-instances) and the Jupyter notebook to launch analyses with the 1000 Genomes Project can be accessed online (https://github.com/hms-dbmi/Hail-on-Google-Cloud/blob/master/Analysis/GWAS_Gender_Phase1.ipynb).

RESULTS

Large-scale genomic data analyses on GCP
Focusing first on clusters generated in GCP, we analyzed the total cost and the runtime necessary to perform GWAS analyses for each cluster, from creation to deletion. We instantiated more than 100 clusters (see Supplementary Appendix 3 and Figure 3), resulting in high variability for both total time and cost necessary to conduct a GWAS analysis. Overall, total time for GWAS analysis of each dataset was <2 hours, evidencing the high-performance capacity of cloud parallel processing at scale (see Supplementary Appendices 2 and 3). When trying to optimize our method, we tested the current mindset around cloud-based resources: that to be the most efficient, one should group data into the largest-size clusters. However, our results (Figure 3) showed that each instance type had a breaking point, after which increasing the size of a cluster yielded no further benefit. When facing limitation in terms of cost, we demonstrated that it was advantageous to sample larger clusters (ie, those with a high number of nodes). This reduced the time and consequently, the total cost. However, we also noted a trade-off. Once a particular number of nodes (a large enough cluster) was reached, performance plateaued or even decreased. This manifested as a significant inflexion point across all 4 datasets (Figure 3), where one gives up speed for cost. We determined the best configuration for the 4 distinct datasets (Table 2) based on our primary goal: achieving the lowest total cost (see Supplementary Appendix 2).

Validation with AWS
As described previously, cluster setup required more time in Amazon Web Services. Therefore, we compared the performance obtained with GCP by running the same GWAS script in Jupyter notebook, without considering cluster preparation time and cluster deletion. When choosing the same configuration for both cloud services, we obtained identical execution runtimes for all GWAS (see Supplementary Appendix 3). Our approach worked on both cloud services and with identical computational runtimes and cost (not factoring the cluster setup).

DISCUSSION

Although using distributed computing in research is becoming increasingly common, information concerning cost and the computational power required to perform any specific study (from storage and loading data through computation) is lacking. Moreover, distributed system tools like Spark and Hadoop require specific knowledge that is not yet not widely utilized by bioinformaticians. In this study, Hail, a cloud-compatible analytic tool can be harnessed to address scalability challenges arising from large genomic data analyt-
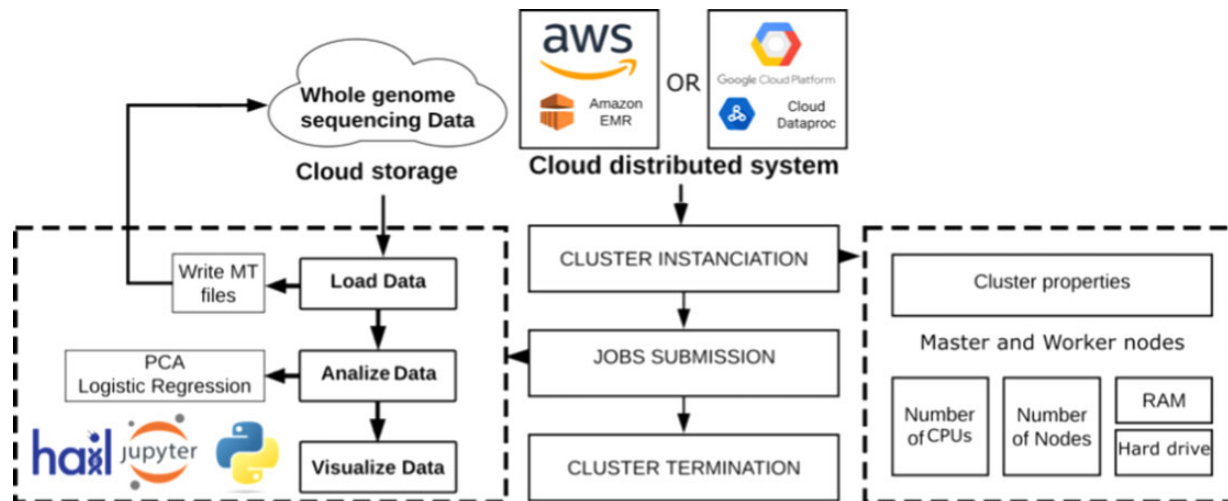
Figure 2. A distributed computational framework for large genomics analysis. Cloud computing setup for executing Hail jobs on Google Cloud Platform and Amazon Web Services with Spark and Hadoop-distributed systems. CPU: central processing unit; MT: Matrix Table; RAM: random access memory.

ics. We described a simple and relatively effortless way (with line of command) to set up a Spark cluster via Hail on both GCP and AWS.

Using this framework, we facilitated the downloading and preprocessing of data via an optimized pipeline for large scale genomic variant analytics. The method is highly scalable and shows that cloud-based distributed systems are, indeed, an effective and novel way to perform cost-effective computational analysis with data sizes higher than several terabytes. The cost of cloud commercial services alone can deter many researchers from transitioning to a cloud infrastructure. We showed that this cost can be reduced by deploying an optimized strategy of cluster size choice, aligned with submission of Hail jobs to the cloud.

We acknowledge that cloud computing still needs to overcome many challenges (ie, cost that is subject to abrupt change and problems with network speed between components). Given these realities, future work might focus on finding ways to estimate the best upstream cluster configuration before launch, specifically optimizing both cost and time. Future studies might delve more deeply into the complex mechanisms of cloud computation, thus further enhancing optimization and driving down cost.

Looking toward the future of precision medicine, a daunting challenge lies in the handling of the massive genomics datasets being generated, as well as the ability to perform extensive interrogation of whole-genome sequences.[27] With an eye toward enabling new biological discovery, we champion the performance benefits provided by the cloud, while emphasizing the boundaries of cluster size and utilization of computational resources. We anticipate that researchers will increasingly utilize cloud computing, especially as the challenges mount around prominent initiatives, such as the 100 000 Genomes Project, the Cancer Genomics Cloud,[28] and the Precision Medicine Initiative.[29] We propose that our method and framework will be an applicable and a powerful addition to these and other future large-scale genomic datasets.

AUTHOR CONTRIBUTIONS

IKr led the design and implementation of the framework in Google Cloud Platform and wrote the majority of the manuscript. CDN worked on the implementation of the method in Amazon Web Services and carried out most of the benchmark for Amazon Web Services. AG-S and NE participated in the analysis and understanding of genetic data. GK and RK helped in the implementation of cloud tools needed for the manuscript as well as for the management of the data used. IKo, JL, AM, and CP have made substantial contributions in the conception and initiation of the project. PA was responsible for the conception of the study, overseeing the analysis, revising the manuscript, and approving the final manuscript. All authors reviewed and approved the final version of the manuscript.

SUPPLEMENTARY APPENDIX

Supplementary material is available at *Journal of the American Medical Informatics Association* online.
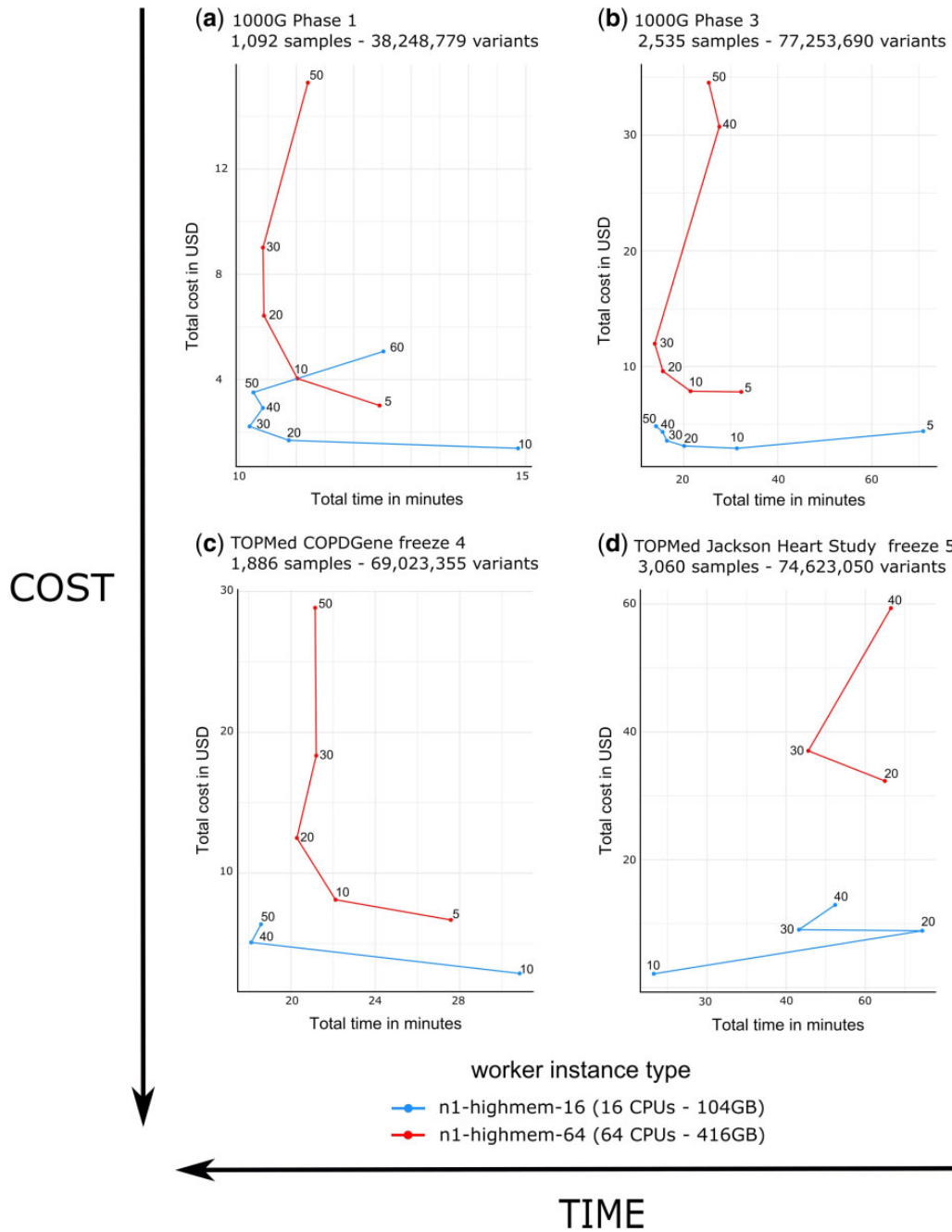
Figure 3. Total cost and performance (from cluster instantiation to deletion) of a Google Cloud, DataProc computing cluster. Master node was instance type: n1-standard-4. Worker nodes had 2 different instance types: n1-highmem-16 and n1-highmem-64. Analysis of genome-wide association studies were performed on 4 different datasets. The numbers near each point indicate the number of worker nodes per cluster. Lines link clusters with the same instance, with increasing numbers of instances per cluster. 1000G: 1000 Genomes Project; CPU: central processing unit; TOPMed: Trans-Omics for Precision Medicine.

Table 2. Best cluster configuration based on the total cost to conduct the genome-wide association study across 4 different datasets

Project releases	Instance type	Nodes	Total runtime (min)	Total cost ($)
1KG Phase 1	n1-standard-16	10	14	1.1
1KG Phase 3	n1-standard-16	10	32	2.6
COPD Freeze 4	n1-highmem-16	10	30	2.9
Jackson Freeze 5	n1-highmem-16	10	23	2.2

1KG: 1000 Genomes Project.

tutes of Health, National Heart, Lung, and Blood Institute, Amazon Web Services, Google, or affiliated organizations and institutions.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Pan C, McInnes G, Deflaux N, *et al*. Cloud-based interactive analytics for terabytes of genomic variants data. *Bioinformatics* 2017; 33 (23): 3709–15.
2. Lacaze P, Pinese M, Kaplan W, *et al*. The Medical Genome Reference Bank: a whole-genome data resource of 4,000 healthy elderly individuals. Rationale and cohort design. *Eur J Hum Genet* 2019; 27: 308–16.
3. Bycroft C, Freeman C, Petkova D, *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; 562: 203–9.
4. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. *Nat Rev Genet* 2018; 19 (4): 208–19.
5. Mashl RJ, Scott AD, Huang K-L, *et al*. GenomeVIP: a cloud platform for genomic variant discovery and interpretation. *Genome Res* 2017; 27 (8): 1450–9.
6. Wang T, Wu YI, Moore D, Russell SJ. Meta-learning MCMC proposals. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Advances in Neural Information Processing Systems 31*. New York, NY: Curran Associates; 2018: 4146–56.
7. Qiao D, Ameli A, Prokopenko D, *et al*. Whole exome sequencing analysis in severe chronic obstructive pulmonary disease. *Hum Mol Genet* 2018; 27 (21): 3801–12.
8. Heath AP, Greenway M, Powell R, *et al*. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *J Am Med Inform Assoc* 2014; 21 (6): 969–75.
9. Stephens ZD, Lee SY, Faghri F, *et al*. Big data: astronomical or genomical? *PLoS Biol* 2015; 13 (7): e1002195.
10. Cirulli ET, White S, Read RW, *et al*. Genome-wide rare variant analysis for thousands of phenotypes in 54,000 exomes. *Nat Commun* 2020; 11: 542.
11. Wiewiórka MS, Messina A, Pacholewska A, Maffioletti S, Gawrysiak P, Okoniewski MJ. SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision. *Bioinformatics* 2014; 30 (18): 2652–3.
12. Maarala AI, Bzhalava Z, Dillner J, Heljanko K, Bzhalava D. ViraPipe: scalable parallel pipeline for viral metagenome analysis from next generation sequencing reads. *Bioinformatics* 2018; 34 (6): 928–35.
13. Chung W-C, Chen C-C, Ho J-M, *et al*. CloudDOE: a user-friendly tool for deploying Hadoop clouds and analyzing high-throughput sequencing data with MapReduce. *PLoS One* 2014; 9 (6): e98146.
14. Fjukstad B, Bongo LA. A review of scalable bioinformatics pipelines. *Data Sci Eng* 2017; 2 (3): 245–51.
15. Zhao S, Prenger K, Smith L, *et al*. Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing. *BMC Genomics* 2013; 14 (1): 425.
16. Wall DP, Kudtarkar P, Fusaro VA, Pivovarov R, Patil P, Tonellato PJ. Cloud computing for comparative genomics. *BMC Bioinformatics* 2010; 11 (1): 259.
17. Hindorff LA, Bonham VL, Brody LC, *et al*. Prioritizing diversity in human genomics research. *Nat Rev Genet* 2018; 19 (3): 175–85.
18. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. *Genome Biol* 2009; 10 (11): R134.
19. Howard DM, Adams MJ, Shirali M, *et al*. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat Commun* 2018; 9 (1): 1470.
20. Turner S, Armstrong LL, Bradford Y, *et al*. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* 2011; Chapter 1: Unit1.19. Chapter 1: Unit1.19.
21. Sudmant PH, Rausch T, Gardner EJ, *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015; 526 (7571): 75–81.
22. 1000 Genomes Project Consortium, Abecasis GR, Auton A, *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; 491 (7422): 56–65.
23. Mardis ER. The 1,000 genome, the 100,000 analysis? *Genome Med* 2010; 2 (11): 84.
24. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009; 25 (16): 2078–9.
25. Ellingson SR, Fardo DW. Automated quality control for genome wide association studies. *F1000Res* 2016; 5: 1889.
26. Patterson N, Price AL, Reich D. Population structure and eigen analysis. *PLoS Genet* 2006; 2 (12): e190.
27. Carter TC, He MM. Challenges of identifying clinically actionable genetic variants for precision medicine. *J Healthc Eng* 2016; 2016: 3617572.
28. Lau JW, Lehnert E, Sethi A, *et al*. The cancer genomics cloud: collaborative, reproducible, and democratized-a new paradigm in large-scale computational research. *Cancer Res* 2017; 77 (21): e3–6.
29. Turnbull C, Scott RH, Thomas E, *et al*. The 100000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* 2018; 361: k1687.