The Role of High Performance, Grid and Cloud Computing in High-Throughput Sequencing

Gaye Lightbody, Fiona Browne, Huiru Zheng
School of Computing and Mathematics
University of Ulster
Newtownabbey, United Kingdom
g.lightbody@ulster.ac.uk,
f.browne@ulster.ac.uk,
h.zheng@ulster.ac.uk

Valeriia Haberland
Tungsten Centre for Intelligent Data Analytics
Goldsmiths, University of London
London, United Kingdom
v.haberland@gold.ac.uk

Jaine Blayney
School of Medicine, Dentistry and Biomedical Sciences,
Queen's University Belfast
j.blayney@qub.ac.uk

***Abstract*— We have reached the era of full genome sequencing using high throughput sequencing technologies pouring out gigabases of reads in a day. To fully benefit from such a profusion of data high performance tools and systems are needed to extract the information lying within the sequences. This paper provides an overview of the evolution of high-throughput sequencing and the tools, infrastructure and data management developing in this space to support a key area in personalized medicine. The paper concludes by providing an outlook in the future of such technologies and their applications and how they might shape clinical governance.**

Keywords—high-throughput sequencing; grid; cloud; personalised medicine

I. INTRODUCTION

Technological advances of High-throughput sequencing (HTS) technologies over the past decade have been pivotal in DNA sequencing. Compared to the conventional Sanger sequencing approach pioneered by Edward Sanger in 1975 using capillary electrophoresis, HTS technology provides massive parallel sequencing producing larger throughput at lower costs [1]. We are now at the stage where it is possible to sequence a whole human genome using a single instrument in 26 hours [2]. There are a number of small and relatively cost effective, HTS platforms available including the Illumina® MiSeq[1], Ion PGM™ (Personal Genome Machine)[2] and the PacBio RS II[3]. These platforms differ in terms of protocol, technology, throughput and read length with the selection of technology dependent on application. A comprehensive review along with additional technological solutions can be found in [3] and [4].

HTS technologies have been used in a broad range of applications including diagnostic testing for hereditary disorders, high-throughput polymorphism detections, comparative genomics, transcriptome analysis and therapeutic decision-making for somatic cancers [5]. The reduction in costs has made HTS technologies accessible to labs to perform high-throughput sequencing experiments, which can generate enormous datasets. For instance, recent advances in HTS technologies have resulted in instruments that are capable of producing >100 gigabases (Gb) of reads in a day [6]. Furthermore, the data generated is not noise free. For these reasons, coupled with the challenges of integrating heterogeneous datasets HTS sequencing data it is considered to be characterized as Big Data, and as with such examples there lies a significant computational challenge. High performance, cloud and grid computing are aspects of computing that have become ubiquitous with processing and analysis of HTS data generated at ever increasing momentum.

With a myriad of options available it is not a straightforward task in selecting a computing technology to suit a specific HTS pipeline. To address this issue, we focus on the computational challenges associated with HTS and review how high performance, cloud and grid based systems can offer solutions. As part of the discussion a brief view on data governance will be presented with respect to HTS in terms of the creation, handling, security and sharing of sequencing data. The paper then concludes with a closing review of the future of HTS technology and its role in personalized medicine.

II. HTS PIPELINE, APPLICATIONS AND PLATFORMS

This technology has been applied to a diverse range of biological science applications including Human Genetics, agriculture, microbes, viruses and infectious diseases and Environmental Genomics. HTS platforms perform parallel sequencing at a large scale. This process allows the sequencing of millions of fragments of DNA from a single sample in unison. The parallel sequencing technology facilitates high-throughput sequencing, which allows an entire genome to be sequenced in about one day [2]. The scale and speed of these technologies are aiding the analysis of genomes and improving our understanding of proteins and their interaction with nucleic acids. HTS has been applied to varied

[1] http://www.illumina.com/systems/miseq.html

[2] https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing.html

[3] http://www.pacb.com/products-and-services/pacbio-systems/rsii/

applications (Table I) from understanding the role of non-coding sequence variants in cancer [7], identifying genomic targets of small molecules [8] to the reconstruction of ancient genomes and epigenomes [9].

TABLE I. COMMON APPLICATIONS OF HTS TECHNOLOGY

Application	Description
Whole Genome Sequencing (WGS)	WGS can be of particular importance when the exact genetic cause of a disease is not fully understood. It allows for a more open discovery of mutations or polymorphisms. Furthermore, there is the potential to derive insights into future potential health issues. There has been recent investigation into how WGS could be integrated with clinical medicine [10].
Whole Exome Sequencing	The exome comprises just over 1% of the genome but offers valuable information in gene discovery research. Studies have resulted in the identification of genes that are relevant to certain diseases such as inherited skin disease [11] or indicate conditions such as inherited autism [12]. Exome sequencing [13] can also be applied to identify disease-causing mutations in pathogenic presentations where the exact genetic cause is unknown.
Targeted Sequencing	In cases where a suspected disease or condition has been identified, sequencing full genome or exome is not necessary. A more efficient application of HTS would be to use targeted sequencing of specific genes or genomic regions. This yields a more affordable and efficient solution while maintaining a high level of resolution [14].
Chromatin immune-precipitation sequencing	ChiP-Sequencing is used to investigate the interactions between proteins and DNA [15]. It has been used to measure transcription binding and protein interactions. miRNA-Seq NGS (Next Generation Sequencing, a form of HTS) technology has been an enabler for analysis of transcriptomes [16].
Large-scale analysis of DNA methylation	DNA methylation (by deep sequencing of bisulfite-treated DNA) acts to investigate genes epigenetic behavior. Whole-genome bisulfite sequencing (WGBS) combined with NGS and genome-wide analysis used to provide a comprehensive view of methylation patterns at single-base resolution across the genome.
Variant detection	Determining single nucleotide variants (SNVs) from NGS results.
de novo assembly sequencing	NGS alignment without a reference genome.

Within the application area of human health, the application of HTS has provided evidence on the context and complexity of cancer genomic alterations, including point mutations, small insertions or deletions, copy number alternations and structural variations. The Cancer Genome Atlas (TCGA) (a coordinated project with the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI)) has applied HTP technologies to profile and analyze large numbers of human tumors to uncover the molecular basis of cancer. A recent study investigated the genomic diversity of primary prostate cancers [17].

A number of platforms are currently used in HTS workflows. These platforms differ in terms of sequencing, imaging and output of data. Popular platforms include: Ion Torrent PGM™ (LifeTechnologies, Carlsbad, CA), MiSeq™ (Illumina, San Diego, CA) and Roche 454. The selection of platform is dependent upon the sequencing problem being investigated. A comprehensive review can be found in [14]. They provide a thorough summary of NGS platforms and applications. In particular, they also provide a synopsis of discoveries of cancer driver mutations achieved through NGS technology.

III. HTS PLATFORM CHALLENGES

With the complexity of genetic sequencing data, coupled with the inevitability of combining it with other relevant datasets so to obtain a complete disease profile, the outcome is highly technically challenging. Cloud, high performance and grid computing offer some answers to meeting these computational needs. However, there are major efforts in developing the computing platforms, infrastructure and data governance policy required to meet growth in use and need of high performance computing solutions.

All platforms have their strengths and weaknesses, which must be considered to enable appropriate use of sequencing data [3]. Different HTS technologies apply different protocols, which in turn determine the type of data produced from each platform [18]. Comparing data output can be challenging in terms of quality scores and accuracy estimates as there is no quality base consensus between different manufacturers. The volume of HTS data is of greater magnitude than that generated by earlier techniques therefore analytical algorithms need to be optimized for speed and memory usage. In addition, HTS techniques produce short read data with error profiles that differ from previous-generation technology. This has resulted in the development of new algorithms to process short reads for sequence alignment, assembly, and read annotation [3]. Consensus of standardized methodologies for HTS analysis is still lacking [5].

Furthermore, using HTS platforms can necessitate planning for at least several hundred gigabytes of data storage per sample. A recent study by Baker [1], investigated the cost of generating sequencing data compared to the cost of storing this data. Interestingly, the cost of generating sequencing data per base is reducing much faster than the cost of data storage per byte. This highlights the disconnect between data production and data storage and the resource to process these data. However, compression techniques [19]–[21] could have the potential to help with the storage and retrieval of these huge data files.

As highlighted, any HTS research in genomics will require significant computational resources, however, with this comes the need for bioinformaticians with skills to install, update, and run the latest tools. The skillset of the bioinformatician needs to be diverse as many of the stages in a typical HTS pipeline employ varied tools, platforms and developmental languages for each specific task, and much of these are open source supported through community forums. Furthermore, as also discussed, datasets themselves are complex and a challenge to incorporate so to best mitigate for inconsistencies in the creation of the datasets. Below gives a small example of the technologies and tools employed, a more comprehensive review can be found in [22]:

- Languages: R/Bioconductor, Python, Bash, Perl.
- Tools: BWA, samtools, vcftools, Picard, Genome Analysis Toolkit (GATK) [23], Bowtie, Tophat, Cufflinks.

- Platforms: BIOVIA ScienceCloud [24], DNAnexus [25], CloVR [26].

IV. High Performance Computing Solutions

Each application will have a varied computational demand. The sections below discuss high performance computing solutions ranging in computational performance.

A. Commodity Clusters

Commodity clusters became popular in bioinformatics, because they offer low-budget elements and scalability with regard to the user's requirements [27], [28]. Commodity clusters consist of regular computers (servers) which are connected through the network as compared to a supercomputer with many processors [29]. Here, a regular computer may also have multiple processing cores. One of the most well-known open-source frameworks for distributed computing on these clusters is *Apache Hadoop* [30]. Hadoop employs the *MapReduce* parallel programming framework which has been popularised by Google [31]. A broad description of this framework is as follows. It consists of the Map and Reduce stages, where input data is first split and presented as intermediate key-value pairs (mapping). Then, the pairs are sorted by their keys and the values are aggregated under their respective keys at the assigned reducer nodes [32]. Next the values are processed for each key (reducing), e.g. counting a frequency of name.

Kawalia et al. [33] describe a workflow for exome analysis which incorporates the MapReduce-like components for parallel calculations on the commodity HPC clusters. The authors' case study focuses on the clusters with a quite generic architecture which are shared by the researchers from the different disciplines, limiting their customization options. As in this case, it might not always be convenient to use Hadoop. The MapReduce concepts have been implemented in many other parallel solutions such as the *Genome Analysis Toolkit* (GATK) [23]; Hadoop-based set of tools, *SeqPig* [34]; parallel version of the well-known *BLAST* and *SOM* algorithms [35], etc. GATK framework helps the researchers to develop their own tools for the NGS data analysis, overcoming limitations of the existing problem-focused tools or complications of the general frameworks. This toolkit can be used for shared- or distributed-memory systems which enhances its applicability for the different types of clusters. SeqPig is a library and toolset, which is based on Apache Pig and aims to ease an analysis of sequencing data for researchers.

The most well-known parallel programming model is *Message Passing Interface* (MPI) [36], which is often compared to the MapReduce paradigm [37]. Chen et al. [37] briefly compare the MapReduce and MPI models as discussed below. MPI is generally argued to be more flexible in terms of passing more control to the user, but at the same time this flexibility complicates its usage. MapReduce is often praised for its fault tolerance as it can re-launch a task on another node if one of the nodes has failed. However, the applications which use MPI can create check-points in order to improve their fault tolerance. Generally, MapReduce is considered to be a more suitable solution for data independent rather than dependent tasks as MPI allows more control over data communication. There are also libraries which consider both paradigms, MapReduce and MPI, such as MR-MPI [38]. Apache also offers a powerful execution engine, known as *Spark* [39], which allows the applications to perform in-memory computations (e.g. iterative applications) which were traditionally disadvantaged by the MapReduce algorithm. Spark can be deployed on the cluster or in the cloud.

B. GPU Computing

Many researchers [29], [40], [41], compare *Graphics Processor Unit* (GPU)-based computing with a traditional CPU-based parallel computing. According to the price to performance ratio, parallel in nature GPUs are potentially more affordable and efficient as compared to sequential in nature CPUs [41], [40]. Hence, a GPU card can have thousands of cores, while more affordable workstations or servers usually have tens of CPU cores. A price for commodity GPUs also continues dropping driven by the expanding gaming industry [41]. In addition, one multi-core server or workstation takes much more space and energy than a GPU card with the approximately equivalent number of cores [42].

NVIDIA, a well-known GPU producer, offers a platform and model for GPU parallel programming which is *Compute Unified Device Architecture* (CUDA) [43]. The large number of CUDA-compatible tools have been developed in the past for NGS data processing and analysis such as *Cushaw* [44], *BarraCUDA* [45], *SOAP3* [46], *CUDASW++* [47], *SeqNFind* [42], etc. A large attention has been attributed to short reads sequence alignment on GPUs (e.g.[44], [45]), or CPUs and GPUs (e.g. [47]). SeqNFind is the set of tools for sequence analysis which can be applied to the NGS data. As for statistical data (e.g. gene expression levels) analysis and visualisation, the open-source R-environment [48] has gained a wide popularity among bioinformaticians in the past years. Hence, the number of packages has been developed for R in order to enable researchers accelerate their calculations using CPU, such as *Simple Network of Workstations* (snow) [49] or GPU parallel computing paradigms, such as *permGPU* [50], *gputools* [51], etc. Although GPU computing is a promising direction for bioinformatics, several bottlenecks arise from the side of GPU memory limitations; a possibly slow data exchange between CPU and GPU memories [52] and the lack of awareness and specialised knowledge among bioinformaticions [40].

C. Cloud Computing

Some of the first adopters of big data in cloud computing are users that deployed Hadoop clusters in highly scalable and elastic computing environments provided by vendors, such as IBM, Microsoft, and Amazon. Cloud-based solutions are increasingly offered on the market such as *BIOVIA ScienceCloud* [24], *DNAnexus* [25] and *BaseSpace Sequence Hub* [53]. A major advantage of these solutions is that they provide scalable storage and performance. Hence, there is no necessity to deploy and maintain the in-house resources [24], [25], especially, that these resources might be required to scale-up towards the increasing amounts of data. They also offer the data and project management tools which facilitate collaborations, regulate access to the shared data, visualise and analyse the data, etc. It is important to mention the big players in Cloud provision: Amazon Elastic Compute Cloud [54],

Google Genomics [55] and Microsoft Azure [56], which all commit to the scalability, speed and data.

Although commercial Cloud solutions provide friendly interfaces and different tools for user convenience, they also have a few disadvantages compared to the open-source solutions. One of such disadvantages is a lack of flexibility on the public clouds as pointed out by Kwon et al. [57]. As an example, the authors suggest that a customization of services is usually limited to the provided functionality. Hence, a user may need to ask the cloud service provider to install additional software resulting in unnecessary waiting times. Another obvious disadvantage constitutes expenses for using a commercial solution [57]. There is a wide range of open-source platforms, pipelines and other tools available to researchers such as the read mapping algorithm for NGS data, *CloudBurst* [58]; the platform which combines virtual machine (VM) and cloud technologies, *CloVR* [26]; the automated pipeline, *Crossbow* [59].

However, the open-source solutions arguably require more time and effort from the user in order to set up and manage the system (e.g. CPU, memory) and the data analysis pipeline as compared to commercial solutions [57]. The system's set up and maintenance by themselves require substantial technical skills [26], [60]. Furthermore, various packages / applications often have to be integrated with each other for the different pipeline's stages [26]. More recently, some open-source software aims to reduce the user's burden, e.g. CloVR claims that they provide the pre-configured pipelines which can be easily installed as a part of VM.

V. Privacy and Data Management

Beyond the major challenges in developing and employing high performance computing platforms and infrastructures, there lies the critical area of data management and governance. Sequencing the human genome has led to challenges in how such huge datasets are created, handled, integrated, stored and shared. These challenges have been exacerbated by the increased complexity and size from HTS. Coupled with the inevitability of translational bioinformatics in which heterogeneous datasets are combined so to obtain a complete disease profile, the outcome becomes even more demanding. This section will discuss some of the key data privacy and management challenges that are encountered when gathering, sharing and analyzing such large repositories, and how cloud computing platforms among other solutions are addressing these concerns.

Existing data governance mechanisms (Table II) enable the relatively free sharing of de-identified data. Such mechanisms have guided genomic data sharing policies including the National Human Genome Research Institute (NHGRI) established the Electronic MEdical Records and GEnomics (eMERGE) Consortium to investigate the best approaches to achieve this [61]. The study by Reid et al. [62] provides a good overview of the challenges faced by the sheer *translational nature of modern genomic data.* In such cases to enable a complete overview of the many compounding factors that influence disease progression additional potentially identifiable data may need to be included. [63]. At this point the data may become traceable back to the subject and thus becomes protected health information (PHI) which is subject to a high degree of handling, storage and security compliance.

TABLE II. US and EU Organizations established for the protection of health and personal data

Legislation	Date	Description
Health Insurance Portability and Accountability Act (HIPPA) Health Information Technology for Economic and Clinical Health Act (HITECH)	1996 2009	HIPPA safeguards individuals' protected health information (PHI) [64]. Its privacy rules set guidelines on how health data can be disseminated through suitable de-identification. Two standards, (Safe Harbor and Expert Determination) may be used for the de-identification process [65]. HIPAA was later supplemented by the Health Information Technology for Economic and Clinical Health Act (HITECH).
Directive 95/46/EC of the European Parliament and Council of the European Union (EU) Directive (EU) 2016/680 Regulation (EU) 2016/679/	1995 With effect from 2018	This directive covers the protection of individuals with regard to the processing of personal data and on the free movement of such data. (Official Journal of the European Union L 281: 0031–0050.) This will be repealed and replaced by the regulation and directive on the protection of natural persons with regard to the processing of personal data [66].

In terms of computational promise cloud-based solutions have many advantages, however there are still a challenges in respect of the security and privacy of the sensitive personal data uploaded to the clouds or other external sources [1], [67]. Hence, the researchers explore private clouds (e.g. in-house clouds) as well as a mixture of private and public clouds for the different data types [1]. Bendekgey [68] raises the point that looking back over the data breaches reported to the US government since 2009 that the fear of storing data on the cloud has been misplaced, highlighting that cloud based systems should be tailored to meet the high security demands for storing genomic PHI. In turn, the cloud-service providers put emphasis on their safety and privacy measures for the sensitive personal data which might include data encryption, a customisable level of data access, compliance with the health and clinical related data regulations. Amazon Web Services (AWS)[4] offers a suite of cloud computing solutions that endeavor to meet the stringent privacy and security rules, requirements for auditing, back-ups, and disaster recovery, established by the HIPAA among other certifications. The AWS DNAnexus [25] platform is one such example with the appropriate clinical and data governance certification. Some HTS-oriented platforms such as BC Platforms [69] can also be deployed in-house which might be an appealing solution in terms of security, cost and application. While Microsoft Azure [70] has moved to offer the option of a hybrid cloud combining both private and public clouds if necessary.

VI. Future

An important challenge in the Bioinformatics and Biomedical domain is bridging the gap between genomic and proteomic data production and the analytics required for the

[4] https://aws.amazon.com/compliance/

understanding of functional biology. Many countries are recognizing this challenge and are developing plans to integrate genomic and patient data to deliver personalized medicine and personal genomics. Such initiatives include the Genomics England led 100,000 Genome Project to sequence 100,000 of the genomes of UK patients and integrate with patient data, led by Genomics England. The PatientsLikeMe project aims to support the sharing of data and have shared agreements with the FDA and pharmaceutical agencies such as AstraZeneca. With the variety of high-throughput technologies, including transcriptomics using microarrays, genome-wide association studies (GWAS), metabolomics modeling, Yeast 2 Hybrid (Y2H) assays, proteomics, high-throughput chemistry screening and in-silico techniques. HTS is one piece in the Bioinformatics knowledge base but it has the potential to augment or complement these existing technologies. Integration of these diverse data in a clinical setting to understand disease is not without its challenges. The study by Xuan et al. [14] highlighted that whole genome sequencing can identify genomic variations between patients with a disease and without a disease, however moving to uncovering clinical useful information and validating genotype-phenotype. Translation of relevant prognostic markers identified by high-throughput techniques into clinical tools for personalized patient treatment has been slow due to issues such as the reproducibility and validity of findings across studies, unfocused study design and inappropriate selection and application of statistical techniques [71], [72]. What is required are standardized pipelines from sequencing until analysis are capable of generating reliable analytical results. To this end there are a number of standardization initiatives such as the Sequencing Quality Control (SEQC) project, a community-wide collaborative led by FDA and the HUPO Proteomics Standards Initiative. Interestingly, we can see that the next wave of sequencing technologies are moving away from high-throughput to small scale real time sequencing. Such devices include Oxford Nanopore's MinION, DNA sequencing sensors to integrate with devices for more bounded tasks such as pathogens surveillance [73].

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

[1] M. Baker, "Next-generation sequencing: adjusting to data overload.," *Nat. Methods*, vol. 7, no. 7, pp. 495–499, 2010.

[2] N. A. Miller, E. G. Farrow, M. Gibson, et *al.*, "A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases," *Genome Med.*, vol. 7, no. 1, p. 100, Sep. 2015.

[3] M. L. Metzker, "Sequencing technologies — the next generation," *Nat. Rev. Genet.*, vol. 11, no. 1, pp. 31–46, 2009.

[4] N. J. Loman, R. V Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen, "Performance comparison of benchtop high-throughput sequencing platforms.," *Nat. Biotechnol.*, vol. 30, no. 5, pp. 434–9, 2012.

[5] E. L. Van Dijk, H. Lè Ne Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends Genet.*, vol. 30, no. 9, pp. 418–426, 2014.

[6] S. N. Naccache, S. Federman, N. Veeeraraghavan, M. Zaharia, et *al.*, "A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples," *Genome Res.*, vol. 24, no. 7, pp. 1180–1192, 2014.

[7] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends Genet.*, vol. 24, no. 3, pp. 133–141, 2008.

[8] J. W. Davey, P. A. Hohenlohe, P. D. Etter, et *al.*, "Genome-wide genetic marker discovery and genotyping using next-generation sequencing," *Nat. Publ. Gr.*, vol. 12, no. 7, pp. 499–510, 2011.

[9] L. Orlando, M. T. P. Gilbert, and E. Willerslev, "Reconstructing ancient genomes and epigenomes.," *Nat. Rev. Genet.*, vol. 16, no. 7, pp. 395–408, Jun. 2015.

[10] J. L. Vassy, D. M. Lautenbach, H. M. McLaughlin, S. W. Kong, et *al.*, "The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine.," *Trials*, vol. 15, p. 85, 2014.

[11] J. E. Lai-Cheong and J. A. McGrath, "Next-generation diagnostics for inherited skin disorders.," *J. Invest. Dermatol.*, vol. 131, no. 10, pp. 1971–1973, 2011.

[12] S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, et *al.*, "De novo mutations revealed by whole-exome sequencing are strongly associated with autism," *Nature*, vol. 485, no. 7397, pp. 237–241, 2012.

[13] M. Choi, U. I. Scholl, W. Ji, T. Liu, I et *al.*, "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 45, pp. 19096–101, 2009.

[14] J. Xuan, Y. Yu, T. Qing, L. Guo, and L. Shi, "Next-generation sequencing in the clinic: Promises and challenges," *Cancer Lett.*, vol. 340, no. 2, pp. 284–295, 2013.

[15] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, et *al.*, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nat. Methods*, vol. 4, no. 8, pp. 651–657, 2007.

[16] J. Lu, G. Getz, E. A. Miska, E. A. Saavedra, et *al.*, "MicroRNA expression profiles classify human cancers.," *Nature*, vol. 435, no. June, pp. 834–838, 2005.

[17] Cancer Genome Atlas Research Network, "The Molecular Taxonomy of Primary Prostate Cancer.," *Cell*, vol. 163, no. 4, pp. 1011–25, 2015.

[18] M. Quail, M. Smith, P. Coupland, T. D. Otto, et *al.*, "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.," *BMC Genomics*, vol. 13, no. 1, p. 341, Jan. 2012.

[19] A. J. Pinho and D. Pratas, "Mfcompress: A compression tool for fasta and multi-fasta data," *Bioinformatics*, vol. 30, no. 1, pp. 117–118, 2014.

[20] D. Qiao, W.-K. Yip, and C. Lange, "Handling the data management needs of high-throughput sequencing data: SpeedGene, a compression algorithm for the efficient storage of genetic data," *BMC Bioinformatics*, vol. 13, no. 1, p. 100, 2012.

[21] B. C.L. and A. Nair, "Benchmark dataset for Whole Genome sequence compression," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. PP, no. c, pp. 1–10, 2016.

[22] T. Ma and A. Zhang, "Omics Informatics: From Scattered Individual Software Tools to Integrated Workflow Management Systems.," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. PP, no. c, 2016.

[23] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, et *al.*, "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.," *Genome Res.*, vol. 20, no. 9, pp. 1297–303, Sep. 2010.

[24] "ScienceCloud, Dassault Systémes Biovia Corp." [Online]. Available: https://www.sciencecloud.com/. [Accessed: 12-Sep-2016].

[25] "DNAnexus." [Online]. Available: https://www.dnanexus.com/. [Accessed: 12-Sep-2016].

[26] S. V Angiuoli, M. Matalka, A. Gussman, K. Galens, et *al.*, "CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing," *BMC Bioinformatics*, vol. 12, no. 1, p. 356, 2011.

[27] T. E. Anderson, D. E. Culler, and D. A. Patterson, "Case for NOW (Networks of Workstations)," *IEEE Micro*, vol. 15, no. 1, pp. 54–64, 1995.

[28] A. Barak and O. La'adan, "The MOSIX multicomputer operating system for high performance cluster computing," *Futur. Gener. Comput. Syst.*, vol. 13, no. 4–5, pp. 361–372, 1998.

[29] J. Blayney, V. Haberland, G. Lightbody, and F. Browne, "Biomarker Discovery , High Performance and Cloud Computing : A

Comprehensive Review," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015, pp. 1514–1519.

[30] "Welcome to Apache™ Hadoop®!" [Online]. Available: http://hadoop.apache.org/. [Accessed: 12-Sep-2016].

[31] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *Sixth Symp. Oper. Syst. Des. Implement.*, 2004, vol. 51, no. 1, pp. 107–113.

[32] R. C. Taylor, "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics.," *BMC Bioinformatics*, vol. 11 Suppl 1, no. Suppl 12, p. S1, 2010.

[33] A. Kawalia, S. Motameny, S. Wonczak, H. Thiele, et *al.*, "Leveraging the Power of High Performance Computing for Next Generation Sequencing Data Analysis: Tricks and Twists from a High Throughput Exome Workflow," *PLoS One*, vol. 10, no. 5, p. e0126321, May 2015.

[34] A. Schumacher, L. Pireddu, M. Niemenmaa, A. Kallio, E. Korpelainen, G. Zanetti, and K. Heljanko, "SeqPig: Simple and scalable scripting for large sequencing data sets in hadoop," *Bioinformatics*, vol. 30, no. 1, pp. 119–120, 2014.

[35] S. J. Sul and A. Tovchigrechko, "Parallelizing BLAST and SOM algorithms with MapReduce-MPI library," *IEEE Int. Symp. Parallel Distrib. Process. Work. Phd Forum*, pp. 481–489, 2011.

[36] B. Barney, "Message Passing Interface (MPI)." [Online]. Available: https://computing.llnl.gov/tutorials/mpi/. [Accessed: 12-Sep-2016].

[37] W. Y. Chen, Y. Song, H. Bai, C. J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 568–586, 2011.

[38] S. J. Plimpton and K. D. Devine, "MapReduce in MPI for Large-scale graph algorithms," *Parallel Comput.*, vol. 37, no. 9, pp. 610–632, 2011.

[39] Apache, "Apache Spark™ - Lightning-Fast Cluster Computing." [Online]. Available: http://spark.apache.org/. [Accessed: 05-Nov-2016].

[40] J. Melanakos, "Parallel Computing on a Personal Computer | Biomedical Computation Review," *Biomedical Computation Review*, Jul-2008.

[41] Zhe Fan, Feng Qiu, A. Kaufman, and S. Yoakum-Stover, "GPU Cluster for High Performance Computing," in *Proceedings of the ACM/IEEE SC2004 Conference*, 2004, vol. 0, no. 1, pp. 47–47.

[42] D. A. Carr, C. Paszko, and D. Kolva, "SeqNFind®: A GPU Accelerated Sequence Analysis Toolset Facilitates Bioinformatics," *Nat. methods, Appl. notes*, pp. 1–4, 2011.

[43] "CUDA GPUs | NVIDIA Developer." [Online]. Available: https://developer.nvidia.com/cuda-gpus. [Accessed: 12-Sep-2016].

[44] Y. Liu, B. Schmidt, and D. L. Maskell, "Cushaw: A cuda compatible short read aligner to large genomes based on the Burrows-Wheeler transform," *Bioinformatics*, vol. 28, no. 14, pp. 1830–1837, 2012.

[45] P. Klus, S. Lam, D. Lyberg, M. Cheung, et *al.*, "BarraCUDA - a fast short read sequence aligner using graphics processing units," *BMC Res. Notes*, vol. 5, no. 1, p. 27, 2012.

[46] C. M. Liu, T. Wong, E. Wu, R. Luo, S. et *al.*, "SOAP3: Ultra-fast GPU-based parallel alignment tool for short reads," *Bioinformatics*, vol. 28, no. 6, pp. 878–879, 2012.

[47] Y. Liu, A. Wirawan, and B. Schmidt, "CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions," *BMC Bioinformatics*, vol. 14, no. 1, p. 117, 2013.

[48] "R: The R Project for Statistical Computing." [Online]. Available: https://www.r-project.org/. [Accessed: 12-Sep-2016].

[49] L. Tierney, "Simple Network of Workstations for R, Department of Statistics and Actuarial Science University of Iowa." [Online]. Available: http://homepage.stat.uiowa.edu/~luke/R/cluster/cluster.html. [Accessed: 12-Sep-2016].

[50] I. D. Shterev, S.-H. Jung, S. L. George, and K. Owzar, "permGPU: Using graphics processing units in RNA microarray association studies.," *BMC Bioinformatics*, vol. 11, p. 329, 2010.

[51] J. Buckner, J. Wilson, M. Seligman, B. Athey, S. Watson, and F. Meng, "The gputools package enables GPU computing in R," *Bioinformatics*, vol. 26, no. 1, pp. 134–135, 2009.

[52] V. Starostenkov, "Hadoop + GPU: Boost performance of your big data project by 50x-200x? | Network World," *Network World*, 2013. [Online]. Available: http://www.networkworld.com/article/2167576/tech-primers/hadoop---gpu--boost-performance-of-your-big-data-project-by-50x-200x-.html. [Accessed: 12-Sep-2016].

[53] "BaseSpace Hub NGS Data Analysis | Cloud and onsite bioinformatics analysis." [Online]. Available: http://www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace.html. [Accessed: 12-Sep-2016].

[54] "Elastic Compute Cloud (EC2) Cloud Server & Hosting." [Online]. Available: https://aws.amazon.com/ec2/. [Accessed: 12-Sep-2016].

[55] "Google Genomics - Store, process, explore and share | Google Cloud Platform." [Online]. Available: https://cloud.google.com/genomics/. [Accessed: 12-Sep-2016].

[56] Microsoft, "Microsoft Azure: Cloud Computing Platform and Services." [Online]. Available: https://azure.microsoft.com/en-us/. [Accessed: 12-Sep-2016].

[57] T. Kwon, W. G. Yoo, W.-J. Lee, W. Kim, and D.-W. Kim, "Next-generation sequencing data analysis on cloud computing," *Genes Genomics*, vol. 37, no. 6, pp. 489–501, Jun. 2015.

[58] M. C. Schatz, "CloudBurst: Highly sensitive read mapping with MapReduce," *Bioinformatics*, vol. 25, no. 11, pp. 1363–1369, 2009.

[59] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg, "Searching for SNPs with cloud computing.," *Genome Biol.*, vol. 10, no. 11, p. R134, 2009.

[60] D. Field, B. Tiwari, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston, "Open software for biologists: from famine to feast.," *Nat. Biotechnol.*, vol. 24, no. 7, pp. 801–803, 2006.

[61] A. L. Mcguire, M. Basford, L. G. Dressler, A. L. Mcguire, et *al.* "Ethical and practical challenges of sharing data from genome-wide association studies : The eMERGE Consortium experience," pp. 1001–1007, 2011.

[62] J. G. Reid, A. Carroll, N. Veeraraghavan, M. Dahdouli, et *al.*, "Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline.," *BMC Bioinformatics*, vol. 15, no. 1, p. 30, 2014.

[63] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Inf. Syst.*, vol. 47, pp. 98–115, 2015.

[64] U.S. Government, "Health Insurance Portability and Accountability Act of 1996," 1996. [Online]. Available: https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm. [Accessed: 12-Sep-2016].

[65] "Methods for De-identification of PHI | HHS.gov." [Online]. Available: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. [Accessed: 12-Sep-2016].

[66] European Commission, "Reform of EU data protection rules - European Commission." [Online]. Available: http://ec.europa.eu/justice/data-protection/reform/index_en.htm. [Accessed: 05-Nov-2016].

[67] M. Schatz, B. Langmead, and S. Salzberg, "Cloud computing and the DNA data race," *Nat. Biotechnol.*, vol. 28, no. 7, pp. 691–693, 2010.

[68] L. Bendekgey, "Cloud computing reduces HIPAA compliance risk in managing genomic data | Healthcare IT News," *Healthcare IT News*, 2013. [Online]. Available: http://www.healthcareitnews.com/blog/cloud-computing-reduces-hipaa-compliance-risk-managing-genomic-data. [Accessed: 12-Sep-2016].

[69] "BC Platforms - Software platforms for next-generation sequencing." [Online]. Available: http://bcplatforms.com/. [Accessed: 12-Sep-2016].

[70] "Big Compute: HPC and Batch | Microsoft Azure." [Online]. Available: https://azure.microsoft.com/en-gb/solutions/big-compute/. [Accessed: 12-Sep-2016].

[71] A. Dupuy and R. M. Simon, "Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting M ethods," *J. Natl. Cancer Inst.*, vol. 99, no. 2, pp. 147–157, 2007.

[72] R. Simon, "Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers," *J. Clin. Oncol.*, vol. 23, no. 29, pp. 7332–7341, 2005.

[73] Y. Erlich, "A vision for ubiquitous sequencing," *Genome Res.*, vol. 25, no. 10, pp. 1411–1416, 2015.