Biomarker Discovery, High Performance and Cloud Computing: A Comprehensive Review

Jaine Blayney*, Valeriia Haberland†, Gaye Lightbody*, Fiona Browne‡
*School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast
†Tungsten Centre for Intelligent Data Analytics, Goldsmiths, University of London
‡ School of Computing and Mathematics, Ulster University
Email: f.browne@ulster.ac.uk

***Abstract*—The analysis of biological markers (biomarkers) have the ability to improve clinical outcomes of a disease through prediction and early detection. The constant improvement of next generation sequencing (NGS) technologies coupled with falling equipment price are driving research and application especially in the medical domain. NGS technology is being applied to cancer research promising greater understanding of carcinogenesis. However, as sequence capacity grows, algorithmic speed is becoming an important bottleneck. To understand these challenges we present a review on difficulties currently faced in discovering biomarkers. The review places the spotlight on sequencing technologies and the bottlenecks encountered in these pipelines. Cloud computing and high performance computing technologies such as Grid are summarized in tackling computational challenges presented by technologies and algorithms used in biomarker research.**

I. Introduction

Biological markers (biomarkers) are defined by the National Institute of Heath as: "A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes or pharmacologic responses to a therapeutic intervention". In practice, biomarkers include tools and technologies that can aid in our understanding of the cause, diagnosis, progression, outcome of treatment and prediction of disease [1].

Biomarkers in medicine could be a measurement or observation of an identified biological process which is used to indicate the presence of a specific disease [2]. For example, a protein, tissue, group of cells or fluids found in a human biopsy sample may be predictive or prognostic of a particular disease. A list of differentially expressed gene when clustered could stratify patients by prognosis or response to treatment [3]. Analysis from clinical data can be used to determine gene signatures with appropriate thresholds that can be used in the development of clinical tests, used for example, in treatment guidance.

A wide range of biomarkers exists including: biochemical and immunological molecular markers through to metabolites or processes such as apoptosis, angiogenesis and proliferation [2]. Biomarkers can be detected in plasma, serum, cerebral spinal fluid, urine, saliva and cyst fluids. Examples of disease biomarkers include: the diagnosis of diabetes based on the level of glucose in serum after 12 hours of fasting. The most sensitive indicator of a cardio-vascular event (including myocardial infarction) is an elevated level of cardiac troponin in serum [2]. The level of serum creatinine is the single most important indicator of renal function. The recent revolutionary advancements in molecular diagnostics and high-throughput DNA sequencing will likely provide many new biomarkers (including gene mutations, copy number variations, and/or single nucleotide polymorphisms) for predisposition to various diseases and prediction of therapeutic response to a treatment or its toxicity.

Biomarkers have improved identification, treatment and prevention of complex common diseases including cancer, diabetes and cardiovascular disease [4]. Moreover, they have been applied to identify diseases such as alzheimer's [5], breast cancer [6] and kidney disease [7] playing a central role in the identification of novel targeted and effective therapeutics. However, the current clinical implications of basic and translational research efforts are modest.

Much progress has been made in the development of high-throughput experimental and technical approaches which have been applied to identify and quantitatively measure biomarkers. The availability of cDNA, oligonucleotide microarrays [8] and Next Generation Sequencing (NGS) such as mRNA-Seq [9] which assesses the global transcript profiles from tissue and cell samples of clinical interest. Other techniques include mass spectrometry-based laser-capture microdissection [10] which are providing functional characteristics and size of protein extracts. These technologies are capable of producing vasts amount of biological data requiring the development and application of computational and statistical approaches to analyze and interpret the results. The reduction in cost of sequencing has been instrumental in prominent research projects such as the 1,000 Genomes project [11], whereby human genomes are sequenced and genetic variation cataloged. The sheer volume of data produced by these projects necessitates high performance and scalability computing solutions. For example, it has been estimated that raw data generated by the 1,000 genome project during the first six months of the project deposited in the NCBIs GenBank is two times more than all previous sequence deposits in the last 30 years [12]. Furthermore, research in areas such as cancer and microbiology including projects such as The

Cancer Genome Atlas [13] are undertaking sequencing in the order of magnitude larger than the 1,000 Genomes project.

In this review we will provide an overview on the current challenges faced in biomarker discover. We will focus on the technologies such as sequencing technologies, the bottlenecks encountered and the applications exploiting Cloud and high performance computing to improve power and speed, owing to the computational challenges presented by many of the technologies and algorithms that are used in biomarker research. Finally, we conclude on future challenges in this field.

II. Challenges in Biomarker Development

The past 20 years has seen major investments with the aim of discovering and validating cancer biomarkers. Three successful multi-gene signatures that have been incorporated into current clinical practice are MammaPrint [14], Oncotype DX [15] and Prosigma PAM50 [16] and have had a positive impact. However, no new major cancer biomarkers have been approved for clinical use for at least 25 years [2]. Furthermore, only 0.07% of published biomarkers have made their way into routine clinical use with very few biomarkers useful for cancer diagnosis and monitoring. [17].

A major challenge in biomarker discovery is due to the complexity in detecting and validating associations within large heterogeneous datasets. This is further complicated due to the lack of resources and multidisciplinary expertise within small academic teams. Validating potential biomarkers is yet another major obstacle. This has been recognized by the US Food and Drug Administration (FDA) who have established a Biomarker Qualification Program to support the Center for Drug Evaluation and Research work with external scientists and clinicians in developing biomarkers.

Some of the biomarker development limitations result from data issues resulting from tissue collection, non-standard protocols, technical/batch effects, bias, power, and lack of validation. Standardized approaches are needed to help eliminate variations in practice. This is of particular importance due to the vast amount of data sets generated as a result of testing multiple biomarkers. Development in standardized approaches is ongoing. For example the Cancer Institute's Cancer Human Biobank (caHUB) [18] has established stringent guidelines on collection, annotation, storage, and analysis of samples. In terms of algorithmic development there are further issues regarding inappropriate tests, lack of repeatability and general robustness.

Poste [17] suggests larger scale collaborative multidisciplinary teams bringing together industry, clinical practice and scientific research are necessary for biomarker development to reach its full potential. Yang et al. [19] highlight some of the challenges associated with developing prognostic indicators for breast cancer. Integration of information such as gene deletions, translocations, and locus amplification; biomarkers from high-throughput -omics technologies such as genomics, proteomics, and metabolomics; and long recognized outcome variables such as tumor size, histologic grade, axillary nodal status, and estrogen receptor (ER) status [19] can be used to provide a more tailored therapy. One of the key challenges is the integration of -omics data to provide a robust and effective solution. There is however current research interest in combining new NGS technologies with existing microarray technologies to enable a combining of multiple complementary resources.

III. Technology Applied to Identify and Validate Biomarkers

With the challenge issued by the FDAs Critical Path Initiative to make better use of genomic technologies, biomarkers are set to have an ever more important role in the drug development process. In the past decade, the use of nucleic acid sequencing has increased as the ability to sequence has become accessible to research and clinical labs all over the world. The demand for faster and more economical sequencing has led to the development of NGS. These technologies have allowed researchers to characterize the molecular landscape of diverse cancer types and led to dramatic advances in cancer genomic studies. NGS could have a central role in the discovery of new genomic biomarkers, owing to the many different types of experiment that can be performed on a single machine. Furthermore, researchers use a 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. NGS is one piece in the Bioinformatics puzzle and has the potential to augment or complement these existing technologies. The key computational challenges will be around developing data analysis tools that could simultaneously analyse across these vast data sets, looking for biomarker signatures [20].

A. *Common Applications for Next Generation Sequencing*

Virtually all areas of biological sciences now use sequencing. For example, the application of NGS has provided information on the complexity of cancer genomic alterations [20]. The comparison of these alterations to matched normal samples have enabled researchers to distinguish between two categories of variants: somatic and germ line [21]. There are a number of experimental processes in which NGS technologies can be applied. These include:

1) Variant detection: application to find genetic differences between the studied sample and the reference. These differences range from single nucleotide variants (SNVs) to large genomic deletions, insertions, or rearrangements.

2) RNA-seq: can be used to determine the expression level of annotated genes as well as to discover novel transcripts. Furthermore, it can detect alternative splicing, RNA editing and fusion transcripts.

3) ChIP-seq: is a method for genome-wide screening proteinDNA interactions. They can be used to measure DNA methylation change and histone modifications.

The combination of NGS technologies provides a high-resolution and global view of the cancer genome. Using powerful bioinformatics tools, researchers aim to decipher the huge amount of data generated to improve our understanding of cancer biology and to develop personalized treatment strategy.

IV. Computational Bottlenecks in Biomarker Discovery

Genome sequencing is continuing to decrease in cost at a faster rate than the cost of storage [12]. The amount of data produced by NGS for example are in the orders of magnitude greater than that generated by earlier techniques such as Sanger. Analysis algorithms must therefore be optimized for speed and memory usage. Bottlenecks are observed across many tasks in the genomic processing pipeline from sample preparation through to the analysis of the data output from NGS platforms. A summary of these tasks are presented below.

A. Genomic Sequence Mapping

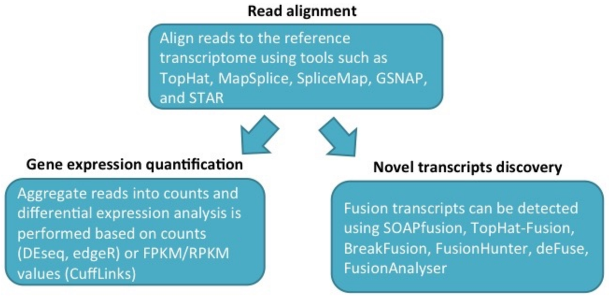
Reads produced by NGS technologies are relatively short in length and their throughput several magnitudes higher than traditional microarray technologies [22]. To use these reads for tasks such as transcriptome sequencing and gene structure identification the sequence reads must be aligned over intron boundaries [23]. When a reference genome assembly exists, alignment is performed. Taking the read, the alignment process determines the most likely source of the sequence within the specific species genome. Developed alignment algorithms are summarised in Table 1. These methods differ in terms of their ability to handle reads of different lengths, accuracy in mapping "noisy" reads, sensitivity and computational efficiency.

TABLE I
An Overview of Key Sequence Mapping Algorithms

Algorithm	Description	Reference
MAQ	Applied in human whole-genome alignments. Handles short reads and supports SNP calling on a diploid sample.	[24]
BWA	Integrates the BurrowsWheeler Transform algorithm with backward search to align short reads against a reference genome.Ability to handle gaps and mismatches.	[25]
BFAST	Creates whole genome indexes to map reads. Use of Smith-Waterman method supporting detection of small indel	[26]
SOAP3	Uses multi-processors in a graphic processing unit (GPU) along with an adapted BWA algorithm to map reads	[27]

For organisms without a sequenced reference genome, assembly algorithms can be applied for analysis. To assemble new genomes, benchmark assembled reference genomes in other species are used. These include "de novo assembly algorithms" which are based on graph theory such as the de Bruijn graph formulation [28].

Fig. 1. A summary of tasks in the RNA-Seq analysis pipeline


Read alignment
Align reads to the reference transcriptome using tools such as TopHat, MapSplice, SpliceMap, GSNAP, and STAR
Gene expression quantification
Aggregate reads into counts and differential expression analysis is performed based on counts (DEseq, edgeR) or FPKM/RPKM values (CuffLinks)
Novel transcripts discovery
Fusion transcripts can be detected using SOAPfusion, TopHat-Fusion, BreakFusion, FusionHunter, deFuse, FusionAnalyser

B. Genomic Sequencing Analysis

Analysis of the sequence data can include a variety of assessments such as genetic variant calling to detect SNPs, transcript expression level measurement through to identification of novel genes or regulatory elements. As outlined in Grada et al. [29] analysis can also include the identification of somatic and germline mutation events which may aid in the diagnosis of disease. A number of open source software solutions exist to perform these analysis including GATK [30], Samtools [31]. These platforms differing in terms of statistical approaches and sequencing depths.

C. RNA-Seq Analysis

RNA-Seq [32] technologies are providing transcriptome profiling using deep-sequencing technologies. As described by Metzher et al. [33] RNA-Seq are advantageous in identifying and quantifying rare transcripts where no prior knowledge of a gene exists. Furthermore, for identified genes, they can provide information on alternative splicing. Fig. 1 provides a summary of key tasks undertaken in RNA-Seq analysis.

V. CPU & GPU-based Accelerated Computations for Biomarker Discovery

Moores law defined by Intel co-founder Gordon Moore states ''the number of transistors that can be placed on an integrated circuit board is increasing exponentially, with a doubling time of roughly 18 months" [34]. This law is currently being outpaced by generation of sequence data and the storage and processing power required to analyse them. In order to handle these large data, technological solutions are required to speed up the analysis of biological data. Technical solutions differ in terms of cost of resources and technical expertise required to implement them. Many research institutions may have limited budget for computational power. There are also concerns in respect of the data privacy and security, especially, when the external resources (such as public cloud services) are used as the data may contain personal information [35]. Hence, the technologies are discussed in this section, considering their financial and technical accessibility and security.

A. Commodity Clusters

Over the past decade, large-scale distributed systems of computing units have become a powerful resource to process large biological data in parallel. Commodity clusters have received significant attention in bioinformatics, because of their low-budget elements and scalability in respect of the user's requirements [36], [37]. Commodity clusters consist of regular computers (servers), connected through the network, compared to a supercomputer containing many processors. Parallel execution of tasks on distributed resources is an active area of research with one of the most famous standards: Message Passing Interface (MPI) [38]. MPI requires the users to specify explicitly parallelism in their algorithms, using MPI functions.

B. Apache Hadoop

The Apache Hadoop project [39] proposes an open-source software for distributed processing of the large-scale data on commodity computers. A comprehensive review of Hadoop-based software and its bioinformatics applications are discussed in the paper [40]. Apache Hadoop software focuses on the effective storage and processing of tasks, using the Hadoop Distributed File System (HDFS) [41] and MapReduce parallel programming framework [42] respectively. MapReduce programming framework consists of several stages. First, the input data is filtered and presented as a list of intermediate key and value pairs at the mapping stage. Then, the values with the same key are grouped together and such groups of values are sorted by their keys. Finally, the values for each key are merged together at the reducing stage. MapReduce programming framework has been widely used and extended in many parallel solutions, for example, the execution engine from Microsoft, Dryad [43]; the read mapping algorithm for the NGS data, CloudBurst [44]; the parallelised BLAST algorithm for a sequence alignment [45]. There are libraries that implement MapReduce with MPI such as MR-MPI by Plimpton and Devine [46].

C. Cloud

Clouds have been widely used in bioinformatics [47], [48], [49], [50], some of which have already been mentioned above. Cloud computing provides resources, data storage and processing as a service on a "pay as you go" basis, using a Service Level Agreement (SLA) [51] protocol. The providers of cloud services include but are not limited to Amazon, Microsoft and Google. Clouds allow users to not only use CPU capacity of their resources, but some also offer Graphics Processor Unit (GPU) acceleration. An example of this is the Amazon Elastic Compute Cloud (EC2) instance type, G2, [52].

D. Graphics Processor Unit

Melonakos [53] stated that GPU computing is a promising approach to biological data analysis, compared to a traditional parallel computing on CPUs for several reasons. First, the price to performance ratio is more favourable for GPU than CPU which was also earlier pointed out by Fan et al. [54]. Secondly, a GPU was originally designed to process data in parallel as compared to CPU, which processes data sequentially [53]. Hence, a GPU card can have thousands of GPU cores, which can process data in parallel, while more affordable workstations or servers usually have tens of CPU cores. The prices on commodity GPUs are also driven down by the large gaming industry [54]. One of the most popular GPU cards are produced by NVIDIA [55], and NVIDIA also offers a platform and model for parallel programming on GPUs, i.e. *Compute Unified Device Architecture* (CUDA) [56].

Potentially, GPU computing has promising prospects for bioinformaticians as it would reduce the cost of hardware, combining with a significant speed up in calculations. However, there is a bottleneck for GPU computing, concerned with developing tools and software, which can be easily utilised by researchers and developers [53]. Another concern in respect of GPU computing constitutes a possibly slow data exchange between GPU and CPU (especially, from GPU to CPU memory) [57]. Hence, the data exchange time between those memories might diminish any time surplus resulted from GPU acceleration. In the recent years, bioinformatics has paid a substantial attention to the GPU-based computations, and some of these examples include a GPU-accelerated alignment of DNA sequences [58], [59]; the statistical computations in R [60], which is widely used by biologists and bioinformaticians, such as permGPU package [61]; a molecular dynamics simulation for health research, using the Grid computing paradigm [62] and GPU computing, GPUGRID [63].

VI. Conclusion

The scope and use of omic technologies is vast. Key areas include biomarker and drug discovery, diagnostics and personal genomics.

The advent of NGS technologies has resulted in the generation of biological data at a scale which is outpacing processing power and storage. Researchers need to reconsider traditional approaches to data analysis which typically involve the download of large data sets from resources such as the European Bioinformatics Institute EMBL database [64] onto locally maintained storage servers. Instead, technological solutions are required to address the computationally intensive processing pipelines. This is not a trivial task. Cloud and high performance computing although extremely powerful requires high levels of technical expertise. Bioinformatics applications are often open source and may be dependent on other programming libraries. Furthermore, documentation may be sparse resulting in difficulty in building, configuring and maintaining applications requiring technical expertise [65]. Algorithms developed will require attention in terms of speed. For example, read mapping with high accuracy will take longer to run than a procedure applying heuristics to limit errors and sequence polymorphism. New data types such as the interrupted read sequences from complete genomics will challenge the existing alignment algorithms, as will the increase in read length and experimental studies that focus on cancer genomes with multiple deletions, duplications and rearrangements.

Future application of biomarkers in the clinical setting will be advantageous in terms of disease prediction and outcome. New sequencing technologies promise to unravel complex cancer genomes quickly and at low cost resulting in massive amounts of data [66]. This clinical pipeline will require integration of diverse data into the patient pathway, and guide preventative and therapeutic options, both for diagnosis and personalized treatments. Current challenges will need addressed such as identifying clinically relevant genomic variation across the whole-human genome; potential error in both technical and computational analysis; the ability to manage and deploy the massive information sets arising from genomics, the availability of clinical interventions which can be informed by such genomic analysis. NGS technology will be a valuable tool to compliment traditional sequencing techniques in exploring the possible new biomarkers.

Acknowledgment

The funding for the publication comes from the Ulster University PVC Challenge Award.

References

[1] R. Mayeux, "Biomarkers: potential uses and limitations," *NeuroRx*, vol. 1, no. 2, pp. 182–188, 2004.

[2] E. P. Diamandis, "Cancer biomarkers: can we turn recent failures into success?" *Journal of the National Cancer Institute*, vol. 102, no. 19, pp. 1462–1467, 2010.

[3] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer*, vol. 14, no. 5, pp. 299–313, 2014.

[4] A. C. J. Janssens and C. M. van Duijn, "Genome-based prediction of common diseases: advances and prospects," *Human molecular genetics*, vol. 17, no. R2, pp. R166–R173, 2008.

[5] K. Blennow, H. Hampel, M. Weiner, and H. Zetterberg, "Cerebrospinal fluid and plasma biomarkers in alzheimer disease," *Nature Reviews Neurology*, vol. 6, no. 3, pp. 131–144, 2010.

[6] J. Li, Z. Zhang, J. Rosenzweig, Y. Y. Wang, and D. W. Chan, "Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer," *Clinical chemistry*, vol. 48, no. 8, pp. 1296–1304, 2002.

[7] S. Coca, R. Yalavarthy, J. Concato, and C. Parikh, "Biomarkers for the diagnosis and risk stratification of acute kidney injury: a systematic review," *Kidney international*, vol. 73, no. 9, pp. 1008–1016, 2008.

[8] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando *et al.*, "Microrna expression profiles classify human cancers," *nature*, vol. 435, no. 7043, pp. 834–838, 2005.

[9] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui *et al.*, "mrna-seq whole-transcriptome analysis of a single cell," *Nature methods*, vol. 6, no. 5, pp. 377–382, 2009.

[10] N. Q. Liu, R. B. Braakman, C. Stingl, T. M. Luider, J. W. Martens, J. A. Foekens, and A. Umar, "Proteomics pipeline for biomarker discovery of laser capture microdissected breast cancer tissue," *Journal of mammary gland biology and neoplasia*, vol. 17, no. 2, pp. 155–164, 2012.

[11] N. Siva, "1000 genomes project," *Nature biotechnology*, vol. 26, no. 3, pp. 256–256, 2008.

[12] L. D. Stein *et al.*, "The case for cloud computing in genome informatics," *Genome Biol*, vol. 11, no. 5, p. 207, 2010.

[13] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[14] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[15] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Cronin, F. L. Baehner, M. G. Walker, D. Watson, T. Park *et al.*, "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *New England Journal of Medicine*, vol. 351, no. 27, pp. 2817–2826, 2004.

[16] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *Journal of clinical oncology*, vol. 27, no. 8, pp. 1160–1167, 2009.

[17] G. Poste, "Bring on the biomarkers," *Nature*, vol. 469, no. 7329, pp. 156–157, 2011.

[18] H. A. Massett, N. L. Atkinson, D. Weber, R. Myles, C. Ryan, M. Grady, and C. Compton, "Assessing the need for a standardized cancer human biobank (cahub): findings from a national survey with cancer researchers." *Journal of the National Cancer Institute. Monographs*, vol. 2011, no. 42, pp. 8–15, 2010.

[19] X. Yang, X. Ai, and J. M. Cunningham, "Computational prognostic indicators for breast cancer," *Cancer management and research*, vol. 6, p. 301, 2014.

[20] P. M. Woollard, N. A. Mehta, J. J. Vamathevan, S. Van Horn, B. K. Bonde, and D. J. Dow, "The application of next-generation sequencing technologies to drug discovery and development," *Drug discovery today*, vol. 16, no. 11, pp. 512–519, 2011.

[21] D. Shyr and Q. Liu, "Next generation sequencing in cancer research and clinical application," *Biol Proced Online*, vol. 15, no. 4, 2013.

[22] C. S. Pareek, R. Smoczynski, and A. Tretyn, "Sequencing technologies and genome sequencing," *Journal of applied genetics*, vol. 52, no. 4, pp. 413–435, 2011.

[23] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch, "Optimal spliced alignments of short sequence reads," *BMC Bioinformatics*, vol. 9, no. Suppl 10, p. O7, 2008.

[24] H. Li, J. Ruan, and R. Durbin, "Mapping short dna sequencing reads and calling variants using mapping quality scores," *Genome research*, vol. 18, no. 11, pp. 1851–1858, 2008.

[25] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[26] N. Homer, B. Merriman, and S. F. Nelson, "Bfast: an alignment tool for large scale genome resequencing," *PloS one*, vol. 4, no. 11, p. e7767, 2009.

[27] C.-M. Liu, T. Wong, E. Wu, R. Luo, S.-M. Yiu, Y. Li, B. Wang, C. Yu, X. Chu, K. Zhao *et al.*, "Soap3: ultra-fast gpu-based parallel alignment tool for short reads," *Bioinformatics*, vol. 28, no. 6, pp. 878–879, 2012.

[28] W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen, "A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies," *PloS one*, vol. 6, no. 3, p. e17915, 2011.

[29] A. Grada and K. Weinbrecht, "Next-generation sequencing: methodology and application," *Journal of Investigative Dermatology*, vol. 133, no. 8, p. e11, 2013.

[30] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly *et al.*, "The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data," *Genome research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[31] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin *et al.*, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[32] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[33] M. L. Metzker, "Sequencing technologies?the next generation," *Nature reviews genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[34] R. R. Schaller, "Moore's law: past, present and future," *Spectrum, IEEE*, vol. 34, no. 6, pp. 52–59, 1997.

[35] M. C. Schatz, B. Langmead, and S. L. Salzberg, "Cloud computing and the DNA data race," *Nature Biotechnology*, vol. 28, no. 7, pp. 691–693, Jul. 2010.

[36] T. E. Anderson, D. E. Culler, and D. Patterson, "A case for NOW (networks of workstations)," *Micro, IEEE*, vol. 15, no. 1, pp. 54–64, 1995.

[37] A. Barak and O. La'adan, "The MOSIX multicomputer operating system for high performance cluster computing," *Future Generation Computer Systems*, vol. 13, no. 4-5, pp. 361 – 372, 1998.

[38] B. Barney. Message Passing Interface (MPI). Lawrence Livermore National Laboratory (Last Modified: 2015).

[39] Apache Hadoop. Apache Hadoop - Available from: http://hadoop.apache.org/, Visited on 07 September 2015.

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

[41] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, May 2010, pp. 1–10.

[42] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[43] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: Distributed data-parallel programs from sequential building blocks," *SIGOPS Oper. Syst. Rev.*, vol. 41, no. 3, pp. 59–72, Mar. 2007.

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

[45] S.-J. Sul and A. Tovchigrechko, "Parallelizing BLAST and SOM algorithms with MapReduce-MPI library," in *2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW),*, 2011, pp. 481–489.

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

[47] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781, 2013.

[48] X. Qiu, J. Ekanayake, S. Beason, T. Gunarathne, G. Fox, R. Barga, and D. Gannon, "Cloud technologies for bioinformatics applications," in *Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers*, ser. MTAGS '09. New York, NY, USA: ACM, 2009, pp. 6:1–6:10.

[49] G. Fox, X. Qiu, S. Beason, J. Choi, J. Ekanayake, T. Gunarathne, M. Rho, H. Tang, N. Devadasan, and G. Liu, "Biomedical case studies in data intensive computing," in *Cloud Computing*, ser. Lecture Notes in Computer Science, M. Jaatun, G. Zhao, and C. Rong, Eds. Springer Berlin Heidelberg, 2009, vol. 5931, pp. 2–18.

[50] L. Dai, X. Gao, Y. Guo, J. Xiao, Z. Zhang *et al.*, "Bioinformatics clouds for big data manipulation," *Biology direct*, vol. 7, no. 1, p. 43, 2012.

[51] M. Ahronovitz, D. Amrhein, P. Anderson, A. de Andrade, J. Armstrong, E. A. B, J. Bartlett, R. Bruklis, K. Cameron, M. Carlson, R. Cohen, and et al., *Cloud computing use cases*, white paper - version 4.0 ed., IBM, 2010.

[52] Amazon Web Services. Amazon EC2 GPU - Available from: https://aws.amazon.com/about-aws/whats-new/2013/11/04/announcing-new-amazon-ec2-gpu-instance-type/, Posted on 4 November 2013.

[53] J. Melonakos, "Parallel computing on a personal computer," *Biomedical Computation Review*, p. 29, 2008.

[54] Z. Fan, F. Qiu, A. Kaufman, and S. Yoakum-Stover, "GPU cluster for high performance computing," in *Proceedings of the 2004 ACM/IEEE conference on Supercomputing*. IEEE Computer Society, 2004, p. 47.

[55] CUDA GPUs - Available from: https://developer.nvidia.com/cuda-gpus, Visited on 10 September 2015.

[56] M. Ebersole. What is CUDA? - Available from: http://blogs.nvidia.com/blog/2012/09/10/what-is-cuda-2/, Posted on 10 September 2012.

[57] V. Starostenkov. Hadoop + GPU: Boost performance of your big data project by 50x-200x? - Available from: http://www.networkworld.com/article/2167576/tech-primers/hadoop—gpu–boost-performance-of-your-big-data-project-by-50x-200x-.html, Posted on 24 June 2013.

[58] H. Khaled, E. R. Gohary, N. L. Badr, and H. M. Faheem, "Accelerating pairwise DNA sequence alignment using the CUDA compatible GPU," *International Journal of Computer Applications*, vol. 84, no. 1, pp. 25–31, 2013.

[59] C. Trapnell and M. C. Schatz, "Optimizing data intensive gpgpu computations for dna sequence alignment," *Parallel Computing*, vol. 35, no. 8-9, pp. 429 – 440, 2009.

[60] The R project for statistical computing - Available from: https://www.r-project.org/, Visited on 10 of september 2015.

[61] I. D. Shterev, S. L. G. S.-H. Jung, and K. Owzar, "permGPU: Using graphics processing units in RNA microarray association studies," *BMC Bioinformatics*, vol. 11, no. 329, 2010.

[62] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid: Enabling scalable virtual organizations," *Int. J. High Perform. Comput. Appl.*, vol. 15, no. 3, pp. 200–222, 2001.

[63] GPUGRID - Available from: https://www.gpugrid.net/, Visited on 7 of September 2015.

[64] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez, "A new bioinformatics analysis tools framework at embl–ebi," *Nucleic acids research*, vol. 38, no. suppl 2, pp. W695–W699, 2010.

[65] A. Shachak, K. Shuval, and S. Fine, "Barriers and enablers to the acceptance of bioinformatics tools: a qualitative study," *Journal of the Medical Library Association: JMLA*, vol. 95, no. 4, p. 454, 2007.

[66] R. Y. Kim, H. Xu, S. Myllykangas, and H. Ji, "Genetic-based biomarkers and next-generation sequencing: the future of personalized care in colorectal cancer," *Personalized medicine*, vol. 8, no. 3, pp. 331–345, 2011.