Journal of Biomedical Informatics 46 (2013) 774–781

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

'Big data', Hadoop and cloud computing in genomics

Aisling O'Driscoll [a], Jurate Daugelaite [b], Roy D. Sleator [b,*]

[a] Department of Computing, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland
[b] Department of Biological Sciences, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland

ARTICLE INFO

Article history:
Received 16 January 2013
Accepted 8 July 2013
Available online 18 July 2013

Keywords:
Cloud computing
Bioinformatics
Big data
Genomics
Hadoop

ABSTRACT

Since the completion of the Human Genome project at the turn of the Century, there has been an unprecedented proliferation of genomic sequence data. A consequence of this is that the medical discoveries of the future will largely depend on our ability to process and analyse large genomic data sets, which continue to expand as the cost of sequencing decreases. Herein, we provide an overview of cloud computing and big data technologies, and discuss how such expertise can be used to deal with biology's big data sets. In particular, big data technologies such as the Apache Hadoop project, which provides distributed and parallelised data processing and analysis of petabyte (PB) scale data sets will be discussed, together with an overview of the current usage of Hadoop within the bioinformatics community.

1. Introduction

Advances in next generation sequencing technologies [1] has resulted in the generation of unprecedented levels of sequence data. Therefore, modern biology now presents new challenges in terms of data management, query and analysis. Human DNA is comprised of approximately 3 billion base pairs with a personal genome representing approximately 100 gigabytes (GB) of data, the equivalent of 102,400 photos. By the end of 2011, the global annual sequencing capacity was estimated to be 13 quadrillion bases and counting, enough data to fill a stack of DVDs two miles high [2].

Moore's Law describes a trend coined by Intel co-founder Gordon Moore which states that "the number of transistors that can be placed on an integrated circuit board is increasing exponentially, with a doubling time of roughly 18 months" [3]. Put more simply: computers double in speed and half in size every 18 months. Similar phenomena have been noted for the capacity of hard disks (Kryder's Law) [4] and network bandwidth (Nielsen's Law and Butter's Law) [5]. This trend has remained true for approximately 40 years, until the completion of the Human Genome project in 2003. Since then, a deluge of biological sequence data has been generated; a phenomenon largely spurred by the falling cost of sequencing [6]. Sequencing a human genome has decreased in cost from $1 million in 2007 to $1 thousand in 2012 [7]. As further evidence of this, the 1,000 Genomes project [8], which involves sequencing and cataloguing human genetic variation, has deposited two times more raw data into NCBI's GenBank during its first 6 months than all the previous sequences deposited in the last 30 years [9] and with mobile sequencing thumb drives on the horizon [10], it shows no sign of slowing. Over the coming years, the National Cancer Institute will sequence a million genomes to understand biological pathways and the genomic variation. Given that the whole genome of a tumour and a matching normal tissue sample consumes 1 TB of uncompressed data (this could be reduced by a factor of 10 if compressed); one million genomes will require 1 million TB, equivalent to 1000 petabyte (PB) or 1 Exabyte (EB) [11].

Until recent years, Moore's law managed to keep ahead of the genomic curve, slightly outpacing the generation of biological sequence data by its growth in storage and processing capacity. However, since 2008, genomics data is outpacing Moore's Law by a factor of 4 [12]. Biology's big data sets are now more expensive to store, process and analyse than they are to generate. This explosion of data is not exclusive to the life sciences and has also impacted on other diverse sectors; with the International Data Corporation (IDC) calculating worldwide data at 0.8ZB (a trillion GB) in 2009, with estimates that this will increase to 40ZB by 2020 [13]. Indeed, the big data problem has been at the forefront of the technology sector over the last 5–8 years as a result of the widespread rollout of high speed wide area network access and a proliferation of next generation applications, combined with the advent of social media [14]. A similar deluge of data is being experienced in diverse sectors such as finance, retail, smart sensor networks (the Internet of Things) and the physical sciences [15] where telescopes capture high quality images, from the 1.6 TB generated each day by NASA's Solar Observatory to the

* Corresponding author. Fax: +353 21 432 6851.
E-mail address: roy.sleator@cit.ie (R.D. Sleator).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.07.001

140 TB gathered daily from the Large Synoptic Survey Telescope [16]. The advent of such large datasets, has significant storage and, more importantly, computational implications. Given that processing and analysis of such datasets may be bursty in nature e.g., tasks such as read assembly is more computationally intensive than subsequent tasks, such high performance compute power may not be fully utilised over time. This has spurred the use of cloud or utility computing (also known as elastic computing) where users can hire infrastructure on a "pay as you go" basis, thereby avoiding large capital infrastructure and maintenance costs. Due to advances in virtualisation, such customised hardware and computational power can now be provisioned instantaneously using user friendly web interfaces.

However, the solution does not lie in cloud computing alone [17]. Big data presents problems in that it can deviate from traditional structured data (organised in rows and columns) and can be represented as semi-structured data such as XML, or unstructured data including flat files which are not compliant with traditional database methods. Furthermore, cloud computing alone does not address the challenge of big data analytics (known as Data as a Service) where large scale processing is required, particularly when the scale of the data exceeds a single machine. In this case it is necessary to develop applications that will execute in parallel on distributed data sets, a non-trivial task that has been the focus of grid computing research for many years. This explosion in big data and the need for the application of big data technologies has generated significant demand for "data scientists", computer scientists and mathematicians with expertise in big data, analytical techniques, statistics, data mining and computer programming. As evidence of this, the Harvard Business Review has heralded the Data Scientist as the "sexiest Job of the 21st Century" [18]. This hype is not without foundation. The market for big data is $70 billion and appears to be growing by 15% a year [19]. If this trend continues, it is estimated that by 2018 the US will need between 140,000–190,000 people with deep analytical skills and a knowledge of big data technologies [20] To address this, in recent months the US government has announced the "Big Data Research and Development Initiative", committing $200 million to big data research initiatives, including the National Institutes of Health (NIH) to "improve the tools and techniques needed to access, organise, and glean discoveries from huge volumes of digital data" [21].

In the healthcare sector, according to the McKinsey Global Institute, if big data is used effectively, the US healthcare sector could make $300 billion in savings per annum, reducing expenditure by 8% [22].

Therefore, while the life sciences are not the only sector experiencing big data overload and while the challenges in applying big data technologies are currently not trivial, particularly given their infancy, the benefit to mankind of deciphering such big biological data sets make it the ultimate use case.

2. Overview of cloud computing and big data technologies

Cloud computing provides a scalable and cost efficient solution to the big data challenge; although largely ill-defined and widely abused to represent anything that is 'online', the National Institute for Standards and Technology (NIST) defines Cloud computing as *"a pay-per-use model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction"*.

While some of the concepts behind cloud computing, such as distributed systems, grid computing and parallelised programming are not new, one of the primary enablers for cloud is virtualisation technology. This has facilitated the business model for clouds to evolve, enabling widespread rollout [23]. Using virtualisation technology, a single physical machine can often host multiple virtual machines (VMs), ensuring maximum utilisation of the hardware and capital investment. A virtual machine (VM) is a software application that emulates a physical computing environment in which an Operating System (OS) and associated applications can be run with multiple virtual machines installed on a single machine. A Hypervisor, a virtualisation management layer, translates the requests from the VM to the underlying hardware (CPU, memory, hard disks and network connectivity). The distinction between these is noted in Fig. 1.

Similarly, advances have been made in the robustness of distributed computing and parallelised programming [24]. Traditionally in grid computing, most effort was spent on maintaining the resilience and robustness of the cluster itself rather than on solving the actual problem. For example, if one processor was to fail or hang this could jeopardise the entire analysis until the processor recovers. Modern big data technologies have devised solutions to overcome such limitations and provide solutions to process and analyse massively parallelised data sets using scalable and cost effective techniques. The distinction between big data technologies and cloud computing is commonly overlooked as cloud computing is often used to facilitate the cost effective storage of such large datasets. Furthermore big data technologies are often offered as Platform as a Service (PaaS) within a cloud environment. It is however important to note that technologies, while often coinciding, are distinct and can operate mutually exclusively.

2.1. Infrastructure as a Service (IaaS)

IaaS is where Cloud Computing providers incur large capital expense to invest in establishing and maintaining High Performance Computing (HPC) infrastructure or data centres that are then made available to their customers on a "pay for what you need" basis, with associated fine grained billing. It means end users incur none

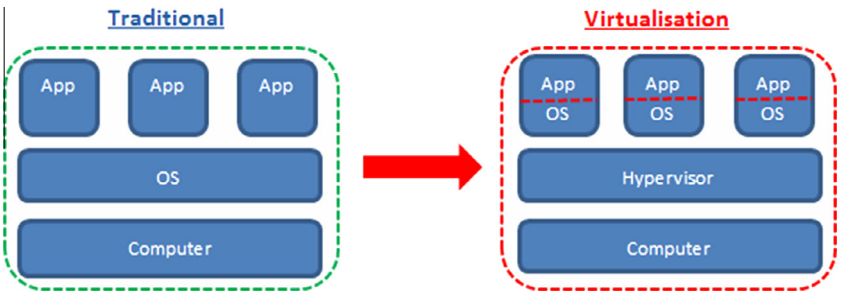

Fig. 1. Conceptualisation of the transition from traditional computing where applications interact with the hardware via one instance of the Operating System (OS), to virtualised environments where multiple OS images share the hardware resources (CPU, RAM, storage and networking), which are allocated and managed by virtualisation software known as a hypervisor or virtual machine monitor (VMM).

of the cost of building a HPC environment, especially given that use of such environments is often "bursty" in nature with the infrastructure often underutilised [25]. Users access this hardware using their broadband connection gaining access to server or storage infrastructure, and, similar to an electricity or gas bill, pay for what they use. As a result, IaaS is often referred to as utility computing or "elastic" computing, i.e., the ability to scale up or down on demand. IaaS providers employ virtualisation where users can create their own virtual machines, specifying the operating system and in some case the applications required [26].

Amazon Web Services (AWS), are the leaders in IaaS, with some estimates suggesting that AWS holds 70% of the total IaaS market share [27]. They have a number of offerings; the most popular being the Elastic Compute Cloud (EC2) and Simple Storage Service (S3). AWS EC2 enables users to rent servers, accessing these via a VM image that they have chosen. Users can either build their own Windows or Linux based VMs from a menu of configurations or choose a pre-built bioinformatics specific image. AWS S3 is an online storage service. By paying a nominal fee, users gain access to the world's largest data centre, accessing whatever type of infrastructure they require, for as long as they need, charged only for CPU usage, data storage and data transferred to and from Amazon. Furthermore, AWS offerings are extremely competitive, with instances of the EC2 service costing as little as 2 cents per hour for allocation of a VM (dependant on user requirements), while S3 pricing starts at 15 cents per GB stored per month. A more comprehensive review of the use of AWS in biomedical computing is provided in [28].

It should be noted that the aforementioned solutions involve a public cloud. In this context, a public cloud refers to resources (infrastructure, applications, platforms, etc.) made available to the general public, typically on a "pay as you go" basis, and accessible only over the Internet. However, other cloud deployment architectures exist, such as private or hybrid clouds. Private clouds refer to virtualised cloud infrastructure owned, housed and managed by a single organisation. Hybrid clouds refer to the connection of two or more types of cloud, e.g., private and public, typically via Virtual Private Networking (VPN) technology for the purpose of scalability and fault tolerance. There is a trend towards a fourth model referred to as Community Clouds, where organisations with a common purpose, e.g., public sector organisations, contribute financially towards a cloud infrastructure, typically managed by a third party.

The market is changing rapidly, with significant fluctuations in the IaaS landscape over the last few months; HP, Microsoft and Google are all vying for market supremacy. HP are now offering HPCloud based on OpenStack a cloud IaaS software originally developed by Rackspace and NASA. Microsoft, already established in the cloud platform field with the Windows Azure platform, is digressing into IaaS offerings, while Google have launched their IaaS offering called the Google Compute Engine. Other IaaS cloud providers of IaaS include Rackspace and Joyent.

2.2. Software as a Service (SaaS)

SaaS refers to the process of availing of applications run on a remote cloud infrastructure, accessible via the Internet. SaaS represents a shift away from installing software locally, on individual PCs, towards a trend of "thin clients", where all the user requires is an Internet connection and a browser, such as Internet Explorer or Google Chrome. The user then connects to a desktop environment via a VM where all the software is installed. It should be noted that while the cloud service provider provides more functionality to the user in SaaS, the user has less control over their environment, as illustrated in Fig. 2. Therefore, the choice of option is very much dependent on the users' desired level of control. SaaS is analogous to the 1980s approach of operating "dumb terminals" connected to a mainframe containing all the required applications.

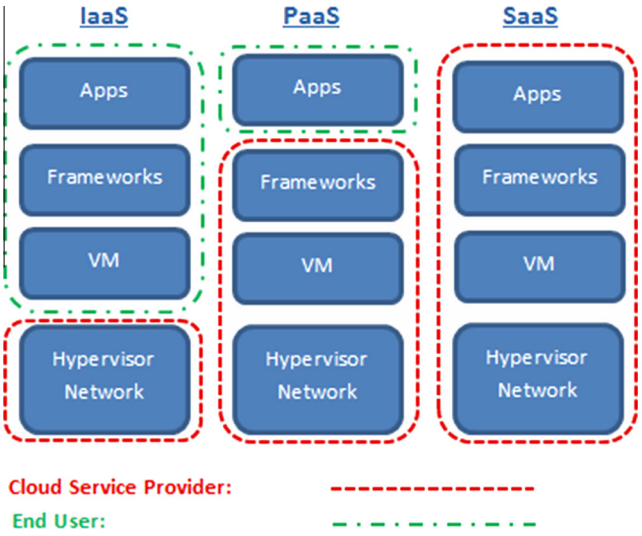

Fig. 2. Cloud Computing Models are largely categorised as Infrastructure, Platform or Software as a Service (IaaS, PaaS, SaaS). Each model differs in the level of functionality provided to the user by the cloud provider. The deciding factor in choosing a particular model is dependent on the user's particular requirements as well as the level of system control that is required.

Unfortunately, most bioinformatics based applications are difficult to build, configure and maintain, primarily because they are, for the most part, open source in nature, lacking good documentation and require many programming library dependencies [29]. As a result, this requires an advanced level of technical expertise on the behalf of the biologist and, as such, is a common bottleneck in the adoption of bioinformatics based applications. However, as all software applications are installed and configured within the VM, SaaS provides the perfect solution.

Cloud BioLinux [30], created by the J. Craig Venter Institute (JCVI) is an example of SaaS. It is a publicly accessible virtual machine that is stored at Amazon EC2, is freely available to EC2 users, and is based on an Ubuntu Linux distribution. It comes with a user friendly Graphical User Interface (GUI), along with over 100 pre-installed bioinformatics tools including Galaxy [31], BioPerl, BLAST, Bioconductor, Glimmer, GeneSpring, ClustalW and EMBOSS utilities, amongst others. While Linux based Bioinformatics distributions such DNALinux, BioSlax BioKnoppix, DebianMed, are not unusual, they are built to run on standalone local machines. SaaS initiatives such as BioLinux have been known to be referred to as Science as a Service (ScaaS). Another significant advantage of using such SaaS VM images on a public cloud, such as Amazon, is that Amazon provides access to several large genomic data sets including the 1000 Genome project, as well as NCBI, GenBank and Ensembl. CloVR [32] provides a similar image with pre-installed packages. Standalone bio/medical software applications/suites with a cloud backend include Taverna [33], FX [34], SeqWare [35], BioVLab [17] and commercial equivalents such as DNAnexus [36].

2.3. Platform as a Service (PaaS)

PaaS allows users to build software applications by building on software libraries or development platforms already developed by the cloud provider. Within the technology field, platforms include Google App Engine, Microsoft Azure and MapReduce/Hadoop amongst others. MapReduce/Hadoop is a data processing and analytics technology that has been revolutionary in the realm of com-

puter science and is one of the hottest technologies in the big data space [37].

MapReduce is Google's solution for processing big data [38] and was developed as large Internet search engine providers were the first to truly face the "big data tsunami", indexing billions of webpages in a quick and meaningful way. Map Reduce is a software framework, written in Java, designed to run over a cluster of machines in a distributed way. The data itself is split into smaller pieces and are distributed over thousands of computers, known as the Google File System (GFS) and a parallelised programming API called MapReduce is used to distribute the computations to where the data is located (Map) and to aggregate the results at the end (Reduce). Hadoop, an open source implementation of Google's solution, comprised of MapReduce and the Hadoop Distributed File System (HDFS), is used by leading technology companies such as Facebook, Amazon, Twitter and is based on a strategy of co-locating data and processing to significantly accelerate performance. In May 2009, Hadoop broke a world record, sorting a PB of data in 16.25 h and a TB of data in 62 s [39]. Hadoop clusters can be run on private infrastructure, however public offerings such as the Amazon Elastic MapReduce service (EMR) are proving popular, with EMR enabling users to easily and cost-effectively process large data sets and apply additional analytical techniques such as data mining, machine learning and statistical analysis. It should be noted however that programming Hadoop is not a trivial task; requiring significant expertise in Java to develop parallelised programs. As such, Hadoop has largely only been embraced in the technology sector.

Still evolving at an extremely rapid pace, the application of this technology is now being considered in order to make big data discoveries outside of the technology sector. Given that this platform has now evolved into a widely supported and powerful framework for parallelisation and distribution; two paradigms that are particularly applicable to large genomics and medical data sets, Hadoop has enormous potential for making medical discoveries, if and when applied to the life sciences. Furthermore, given that public clouds such as AWS are now offering DaaS by providing a repository of public data sets, including GenBank, Ensembl, 1000 Genomes, Model Organism Encyclopaedia of DNA Elements, Unigene, Influenza Virus, this potential is becoming an imminent reality [40].

3. Parallelised big data technologies and genomics

Applying big data platforms and analytics in the realm of natural science not only has the potential to change lives, but also to save them [41]. Medical/genomics research is thus the dream use case for big data technologies which, if unified, are likely to have a profoundly positive impact on mankind. Currently, a much bigger issue than simply storing big data is processing it in a timely manner, and subsequently analysing the data for meaningful deductions. It is recognised that applying leading technology big data solutions, such as Hadoop, is revolutionary and there have been some early adopters in this space. A categorisation of bioinformatics projects that utilise the Apache Hadoop platform are detailed in Table 1 with a more detailed review recently undertaken by Zou et al [42].

One of the first MapReduce projects applied in the biotechnology space resulted in the Genome Analysis Tool Kit (GATK) [43]. This was followed by several subsequent Hadoop enabled biological contributions [44–46]. CloudBurst was one of the first of these, developed by Michael Schatz et al. [76] when working at the University of Maryland. Schatz later developed Crossbow for SNP identification, using Hadoop's massive sort engine to order the alignments along the genome and then genotyping sample using SoapSNP. As input it aligns a mix of 3 billion paired-end and unpaired reads, equivalent to 110 GB of compressed sequence data, and as output it catalogues all the SNPs in the genome. According to Langmead et al. [46], CrossBow can genotype a human in approximately 3 hours on a 320 core cluster, discovering 3.7 million SNPs at >99% accuracy for $100 (including data transfer comprising an hour) using AWS EC2. However, although Crossbow is considered the state of the art in large scale cloud based SNP detection, it does not address the need to make the read mapping and SNP calling configurable, and only considers short reads of approximately 30 base pairs. Thus, Nguyen et al. [47] present CloudAligner as a user friendly interface application, which in comparison to local-based approaches show a higher quality performance, due to the partition and parallel processing of genome and reads. Langmead, funded via a grant from Amazon Web Service (AWS) and NIH, developed Myrna. Using the AWS cloud, Myrna calculated differential expression from 1.1 billion RNA sequence reads in less than 2 h, at cost of about $66 [48].

With the exception of these early adopted projects, applying these powerful big data technologies to biological problems has largely not been adopted into the mainstream. The technology, though extremely powerful, was built for those who are technically savvy and requires a high level of computational know-how, as applications built on top of the platform must be developed to be parallelised in nature. Few bioinformatics tools are designed to run in parallel, an issue which is not trivial. Until now, while the potential has been recognised in academic environments, little action has been taken for broader uptake and large scale development.

Notwithstanding, in recent months, serious steps have finally been taken to turn this significant potential into a reality outside of academia, with many leading technology multinationals announcing their intention to progress the application of big data technologies to biological problems. Firstly, Dell announced that it is donating unused server capacity to the first FDA-approved personalised medicine trial for pediatric cancer, specifically Neuroblastoma, fostering a specialised software system between the Translational Genomics Research Institute (TGen) and the Neuroblastoma and Medulloblastoma Translational Research Consortium (NMTRC) [49]. Specifically in the PaaS space, Intel announced that they are collaborating with NextBio to optimise the Hadoop Distributed File System, Hadoop and HBase for genomics, with any improvements contributed back to the open source community [50]. Importantly, Cloudera have teamed up with the Institute for Genomics and Multiscale Biology at the Mount Sinai School of Medicine, in a pioneering effort to aid researchers in applying big data technologies in the field of genomics and multi scale biology to "diagnose, understand and treat disease". Areas of research include analysis of human and bacterial genomes; study of the metabolic pathways of normal and disease states in the organism; structure and function of molecules used in treatment of disease, and more [51]. What is particularly noteworthy about this announcement is that both these organisations are committed to cutting edge research in their fields, and together form a formidable collaboration.

Cloudera are widely recognised as being the leading Apache Hadoop software and service provider in the big data landscapeAt the cutting edge of big data technology development, Cloudera contributes more than 50% of its engineering output into open source Apache licensed projects, e.g., Hive, Avro, HBase and towards the further development of the Hadoop framework. This is not surprising given that it was established by three leading engineers from Google, Yahoo and Facebook (namely Christophe Bisciglia, Amr Awadallah and Jeff Hammerbacher) in conjunction with Mike Olson, a former Oracle executive. They were later joined by Doug Cutting the founder of the Apache Hadoop project. Cloudera in

Table 1
Categorization of Hadoop-based bioinformatics implementations.

Function	Algorithm	Description	Reference
Genomic sequence mapping	CloudAligner	A MapReduce based application for mapping short reads generated by next-generation sequencing	[47]
	CloudBurst	A parallel read-mapping algorithm used for mapping next-generation sequence data to the human genome and other genomes	[76]
	SEAL	A suite of distributed applications for aligning, manipulating and analyzing short DNA sequence reads	[77]
	BlastReduce	A parallel short DNA sequence read mapping algorithm optimised for aligning sequence data for use in SNP discovery, genotyping and personal genomics	[78]
Genomic sequencing analysis	Crossbow	A scalable software pipeline that combines Bowtie and SoapSNP for whole genome re-sequencing analysis	[46]
	Contrail	An algorithm for de novo assembly of large genomes from short sequencing reads. Contrail relies on the graph-theoretic framework of de Bruijin graphs	[79]
	CloudBrush	A distributed genome assembler based on string graphs	[80]
RNA sequence analysis	Myrna	A cloud computing pipeline for calculating differential gene expression in large RNA sequence datasets	[48]
	FX	RNA sequence analysis tool for the estimation of gene expression levels and genomic variant calling	[34]
	Eoulsan	An integrated and flexible solution for RNA sequence data analysis of differential expression	[81]
Sequence file management	Hadoop-BAM	A novel library for scalable manipulation of aligned next-generation sequencing data	[82]
	SeqWare	A tool set used for next generation genome sequencing technologies which includes a LIMS, Pipeline and Query Engine	[35]
	GATK	A gene analysis tool-kit for next-generation resequencing data	[43]
Phylogenetic analysis	MrsRF	A scalable, efficient multi-core algorithm that uses MapReduce to quickly calculate the all-to-all Robinson Foulds (RF) distance between large numbers of trees	[83]
	Nephele	A set of tools, which use the complete composition vector algorithm in order to group sequence clustering into genotypes based on a distance measure	[84]
GPU bioinformatics software	GPU-BLAST	An accelerated version of NCBI-BLAST which uses general purpose graphics processing unit (GPU), designed to rapidly manipulate and alter memory to accelerate overall algorithm processing	[85]
	SOAP3	Short sequence read alignment algorithm that uses the multi-processors in a graphic processing unit to achieve ultra-fast alignments	[86]
Search engine implementation	Hydra	A protein sequence database search engine specifically designed to run efficiently on the Hadoop MapReduce framework	[87]
	CloudBlast	Scalable BLAST in the cloud	[88]
Miscellaneous	BioDoop	A set of tools which modules for handling Fasta streams, wrappers for Blast, converting sequences to the different formats and so on	[89]
	BlueSNP	An algorithm for computationally intensive analyses, feasible for large genotype–phenotype datasets	[90]
	Quake	DNA sequence error detection and correction in sequence reads	[91]
	YunBe	A gene set analysis algorithm for biomarker identification in the cloud	[92]
	PeakRanger	A multi-purpose peak caller software package for detecting regions from chromatin immunoprecipitation (ChIP) sequence experiments	[93]

particular has also blossomed in pioneering cloud and big data technologies in the biological research and medical space. Jeff Hammerbacher, Cloudera's chief scientist and co-founder, has committed to dedicating 25% of his time to the initiative with Mount Sinai also heavily committed to the project with Eric Schadt (a visionary in the use of computational biology in genomics [52]) coordinating the initiative.

Thus, with leading pioneers in the fields of big data and computational biology as well as leading multinationals now committed to seeing progress in making medical discoveries by analysing large biological data sets, and with a projected annual growth for healthcare computing of 20.5% through 2017 [53], we are on the cusp of a technology uprising that is likely to have a profound impact on the diagnosis, understanding and treatment of disease.

4. Challenges

It must be emphasised that big data technologies are very much in their infancy and that although powerful, have a long way to go. Programming Hadoop requires a high level of Java expertise to develop parallelised programs. Efforts have been made to simplify this process, even in the technology sector, with developed software libraries such as Hive to add a "SQL" like interface that will generate parallelised Hadoop jobs in the background. Python streaming has also been made available to circumvent complex Java programming by wrapping it in Python, a more lightweight scripting language. Another point of consideration is that Hadoop MapReduce is designed by default to treat each line as an individual record. As many standard sequence formats involve multiple lines per sequence it is necessary to manipulate the data into one line formats, or to program custom Hadoop input formats and record readers – a less than trivial task.

Furthermore, there is a current trend towards further developing analytics and visualisation technologies on top of the Hadoop platform, to enable better standardisation of reporting and summarisation of results. This is a problem which is not adequately addressed in the technology sector and is vital if the technology is to be widely embraced by diverse Industry sectors. Hadoop, is currently still very much a "behind the scenes" technology with no front end visualisation, powerful only if in the right hands and still difficult to set up, use and maintain. There are concerted efforts being made towards adding developer friendly management interfaces or GUIs on top of Hadoop systems to move away from shell or command line interfaces. Recently, Schoenherr et al. [54] presented Cloudgene for this precise purpose. Cloudgene provides a standardised graphical execution environment for currently available and future MapReduce programs, which can all be integrated by using its plug-in interface.

There are also drawbacks associated with the utilisation of cloud computing. One of the most significant challenges, given the scale of the genomic data being generated, is that transmitting such data over the Internet or any other form of communication media takes prolonged periods of time, sometimes even in the region of weeks. Thus, the bottleneck is the rate of data transfer, i.e., getting data into and out of the cloud. As outlined in [55], in an interview with Vivien Bonazzi, program director for computational biology and bioinformatics at the National Human Genome Research Institute (NHGRI), "*putting data into a cloud cluster by way of the Internet can take many hours, even days, so cloud customers often resort to the "sneaker net": overnight shipment of data-laden hard drives*". In fact, BGI, one of the world's leading genomics research institutes produces 200 genomes a day, with disks transported manually via FedEx [56]. AWS are actively trying to overcome this by introducing a multi-part upload and companies such as Aspera are also designing a new layer to operate on top of the TCP (Transport Control Protocol) transport layer protocol in an attempt to alleviate this issue [57]. This high speed file transfer (expected to be between 10 and 100 times faster than traditional FTP and HTTP approaches) has already been integrated into the BGI EasyGenomics SaaS solution showcased at the Bio-IT World Conference & Expo in 2012. BGI also integrated EasyGenomics with the Hadoop platform. This is notable as Hadoop and other scale out big data technologies exhibit a distinct advantage over traditional data management approaches in that the computation is moved to the data. By utilising local commodity hardware, data is distributed across a cluster of machines each utilising local processing, storage and memory and is processed in parallel, negating the need to transfer the data across the network from its storage location to be processed, as is typically the case with traditional HPC solutions. Furthermore, the aforementioned upload challenges are typically only faced by the large sequencing centres or represent a once off challenge. In contrast, as recently noted by Andreas Sundquist, CEO DNAnexus, the upload of sequence data, produced in real time from a single modern sequencing instrument requires a lower bitrate than streaming a movie over the Internet [58].

As recently noted by Schadt [59], the ability to protect medical and genomics data in the era of big data and a changing privacy landscape is a growing challenge. While cloud computing is championed as a method for handling such big data sets, it's perceived insecurity is viewed by many as a key inhibitor to its widespread adoption in the commercial life sciences sector. Indeed, this may explain why its employment has primarily been adopted by research and academic labs. However it should be noted at the outset that in many cases cloud solutions can provide equivalent, if not improved, security depending on the local security policy employed. Clinical sequencing, however, must meet rigorous regulatory requirements, primarily from the Health Insurance Portability and Accounting Act of 1996 (HIPAA) and thus cloud computing is being cautiously considered in such cases. HIPAA standards require compliance across service, platform and infrastructure layers, i.e., across IaaS/PaaS and SaaS. As this is difficult to enforce and validate in a public cloud, with third party contributors, Amazon is not HIPAA compliant. Amazon has however released a paper that will allow customers to develop healthcare applications that comply with a subset of HIPAA's regulations [60]. Hybrid clouds are thus considered a more secure approach with hybrid FDA and HIPAA compliant clouds used as part of the collaboration discussed in the previous section between Dell and TGen to support the world's first personalised medicine trial for paediatric cancer [61] and commercial solutions such as those provided by GenomeQuest and DNANexus. However, fundamental aspects of data security will need to be addressed before widespread adoption of cloud-based clinical sequencing can occur. Some of the key issues include encryption mechanisms (particularly key management), the vulnerabilities of Internet based customer access interfaces, replication in the case of disaster recovery, along with inadvertent data access *via* incorrect data 'deletion' i.e. reassignment of virtual resources allowing customer access to other customers' 'deleted data'. This will not be an overnight solution and with increasingly advanced decryption and de-anonymisation techniques, the privacy of "anonymised" sequence data or Electronic Health Records (HER) may be extremely difficult to definitively guarantee. In the case of highly sensitive data, when all the technical precautions are provided, the weakest link in the chain may, as has traditionally been the case, be the human one. Nonetheless, the increasing impetus to utilise such technologies in order to exploit their economic benefits has highlighted the need for increased legislation in this area [62]. Data tenancy is another perceived challenge, particularly with public cloud usage i.e., availability of data should a commercial cloud service provider cease trading. This was evidenced when Google discontinued Google Health in early 2012, giving users a year within which to make alternative arrangements for their data. Furthermore, should this occur or if another motivating factor causes a user to decide to move their data to another provider, the ease with which this transition can occur largely depends on the interoperability of the initial cloud service.

Unfortunately, most cloud infrastructures provide very little capability on data, application, and service interoperability. This makes it difficult for a customer to migrate from one provider to another, or move data and services back to an in-house IT environment. Finally, a further challenge relates to data privacy legislation (e.g., data in the EU cannot be stored in a US region) as well as legal ownership and responsibility pertaining to data stored between international zones (e.g., the 1000 Genome project exists only in US zone, not the EU zone) [63].

5. Future outlook

Applications such as those outlined in Section 3 illustrate that cloud computing and big data technologies have a significant future role in the life sciences; facilitating high throughput analytics that allow users to rapidly interrogate vast data sets. It must be noted however that Hadoop is not a "catch all" technology but rather is best suited to batch processing applications, as opposed to real time ad-hoc queries. The application of this technology to suitable high impact areas, such as metagenomics [64], personalised medicine, systems biology and protein function and structure prediction [65] has the potential for killer applications.

As the genomics revolution gives way to metagenomics – the functional and sequence based analysis of the collective microbial genomes (microbiome) in a particular environment or environmental niche – Biology's big data sets are about to become several orders of magnitude bigger [64]. When one considers that the total number of bacterial cells on earth is estimated to be $\sim 4–6 \times 10^{30}$, and that the majority of these are uncharacterised; this diversity represents a vast, and as yet largely untapped genetic bounty that can be exploited for the discovery of novel genes, entire metabolic pathways and potentially medically valuable end-products. These novel genes encode new and as yet uncharacterised proteins whose structure and function will have to be elucidated [66,67]. This is likely to put significant stain on existing *in silico* based solutions, such as Folding@Home [68] and Foldit [69]; distributed computing projects which employ a network of idle home PCs to resolve protein 3D structure and infer function. Replacing internet-connected home PCs with cloud based VMs will ultimately facilitate improved protein function prediction.

Furthermore, such computational power, as represented for example by CrossBow which can genotype a human genome in

~3 h at a cost of just $100, opens the very real possibility of personalised medicine; the ability to treat individuals on a case by case basis, tailored to their specific genomic blueprint [70]. This is the holy grail of medicine, which ultimately has the potential to do away with traditional chemotherapeutic regimens (often more damaging than the disease being treated) in favour of a more benign targeted approach.

The big data processing capabilities of cloud computing facilitating the analysis of all the variables at once is a significant enabler of the new area of systems biology [71–73] – a holistic approach which helps to visualise the 'omics', as an interacting network, creating a paradigm shift – allowing us to move from hypothesis-driven to hypothesis-generating research. An example of such an approach is the recent publication of a whole-cell computational model of the uropathogenic bacterium *Mycoplasma genitalium* [74]. This bacterial avatar represents the first truly integrated effort to simulate the complete workings of a free-living microbe *in silico* [75]. The overall model is based on over 900 peer reviewed publications and includes more than 1900 experimentally observed parameters. Model training and parameter reconciliation was achieved by recreating 128 different *M. genitalium* culture simulations – each predicting both molecular and cellular properties of the *in silico* cell – recapitulating the key features of the training data. Model validation was achieved using data sets not used in the construction of the model and which encompass multiple biological functions (from transcriptomics to metabolomics) and scales (from single cells to microbial populations). Systems and synthetic biology based projects, like the one described above, represent an obvious application for next generation cloud based computational biology and has the potential to revolutionise the life sciences.

Acknowledgments

Jurate Daugelaite is funded under the Embark Initiative by an Irish Research Council (IRC) grant RS/2012/122. Aisling O' Driscoll and Dr. Roy D. Sleator are Principal Investigators on ClouDx-i an FP7-PEOPLE-2012-IAPP project.

References

[1] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 2012;13:341.
[2] Pollack A. DNA sequencing: caught in the deluge of data. NY: New York Times; 2011.
[3] Moore G. Cramming more components into integrated circuits. Electronics 1965;38:4–7.
[4] Walter C. Kryder's Law Sci Am 2005;293(August):32–3.
[5] As We May Communicate. <http://www.tmcnet.com/articles/comsol/0100/0100pubout.htm>.
[6] Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 2012;10:599–606.
[7] Davies K. The $1,000 genome: the revolution in DNA sequencing and the new era of personalized medicine [hardcover]. New York: Free Press; 2010.
[8] Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. Nucl Acids Res 2002;30:4103–17.
[9] Stein L. The case for cloud computing in genome informatics. Rev J: Genome Biol 2010;11(May):207.
[10] Mason CE, Elemento O. Faster sequencers, larger datasets, new challenges. Genome Biol 2012;13:314.
[11] Managing and Analysing 1,000,000 Genomes. <http://rgrossman.com/2012/09/18/million-genomes-challeng/>.
[12] Genomics Takes Flight. . ..To the Cloud. <https://idc-insights-community.com/health/life-sciences/genomics-takes-flight-to-the-cloud>.
[13] Gantz J, Reinsel, D. The Digital Universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. In: IDC iView: IDC Analyze the, Future; 2012.
[14] Social Media And The Big Data Explosion. <http://www.forbes.com/sites/onmarketing/2012/06/28/social-media-and-the-big-data-explosion/>.
[15] Big Data Offers Big Opportunities for Retail, Financial, Web Companies. <http://www.eweek.com/enterprise-apps/big-data-offers-big-opportunities-for-retail-financial-web-companies/>.
[16] Data Deluge and the Human Microbiome Project. <http://www.issues.org/28.4/sagoff.html>.
[17] Chae H, Jung I, Lee H, Marru S, Lee S-W, Kim S. Bio and health informatics meets cloud: BioVLab as an example. Health Inform Sci Syst 2013;1:6.
[18] Davenport THP. D. J. Data scientist: the sexiest job of the 21st century. Harward Business 2012;90(70–6):128.
[19] EMC Sitting In Sweet Spot Of $70 Billion Big Data Industry. <http://www.forbes.com/sites/greatspeculations/2011/11/18/emc-sitting-in-sweet-spot-of-70-billion-big-data-industry/>.
[20] Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. Genome Res 2001;11:803–16.
[21] Obama Administration Unveils "Big Data" Initiative: Announces $200 Million In New R&D Investments. <http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf>.
[22] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute; 2011.
[23] The Benefits Of Data Center Virtualization For Businesses. <http://www.cloudtweaks.com/2012/03/the-benefits-of-data-center-virtualization-for-businesses/>.
[24] Big Data, Meet the Huge Data That Will Shape Your Future. <http://www.information-management.com/news/big-data-meet-the-huge-data-that-will-shape-your-future-10023324-1.html>.
[25] Bridging the gap between HPC and IaaS clouds. <http://datasys.cs.iit.edu/seminar/2012-03-06_bogdan-nicolae.html>.
[26] Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biology Direct 2012;7:43 [discussion 43].
[27] What will happen to Amazon's massive cloud business? <http://tech.fortune.cnn.com/2012/05/22/aws/>.
[28] Fusaro V, Patil P, Gafni E, Wall D, Tonellato P. Biomedical cloud computing with amazon web services. PLOS J 2011 [August].
[29] Shachak A, Shuval K, Fine S. Barriers and enablers to the acceptance of bioinformatics tools: a qualitative study. J Med Libr Assoc 2007;95:454–8.
[30] Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, et al.. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinform; 2012;13:42. <calendar:T1:13:42>.
[31] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 2005;15:1451–5.
[32] Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, et al. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinform 2011:12.
[33] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004;20:3045–54.
[34] Hong D, Rhie A, Park SS, Lee J, Ju YS, Kim S, et al. FX: an RNA-Seq analysis tool on the cloud. Bioinformatics 2012;28:721–3.
[35] O'Connor BD, Merriman B, Nelson SF. SeqWare Query Engine: storing and searching sequence data in the cloud. BMC Bioinform 2010;11.
[36] Available at https://dnanexus.com/.
[37] How Hadoop Makes Short Work of Big Data. <http://www.forbes.com/sites/netapp/2012/09/24/hadoop-big-data/>.
[38] Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinform 2010;11(Suppl. 12):S1.
[39] Hadoop Sorts a Petabyte in 16.25 Hours and a Terabyte in 62 Seconds. <http://developer.yahoo.com/blogs/hadoop/posts/2009/05/hadoop_sorts_a_petabyte_in_162/>.
[40] Dai L, Gao X, Guo Y, Xiao JF, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct 2012;7.
[41] Cloudera and Mount Sinai: The structure of a Big Data Revolution? <http://www.zdnet.com/cloudera-and-mount-sinai-the-structure-of-a-big-data-revolution-7000000354/>.
[42] Zou Q, Li XB, Jiang WR, Lin ZY, Li GL, Chen K. Survey of MapReduce frame operation in bioinformatics. Brief Bioinform 2013.
[43] McKenna A et al. The genome analysis toolkit: a MapReduce framework for analysing next-generation DNA sequencing data. Genome Res 2010;20(July):1297–303.
[44] Gurtowski J, Schatz MC, Langmead B. Genotyping in the cloud with Crossbow. Current Protocol Bioinform 2012 [chapter 15, Unit 153].
[45] Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. Nat Biotechnol 2010;28:691–3.
[46] Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. Genome Biol 2009;10:R134.
[47] Nguyen T, Shi W, Ruden D. CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes 2011;4:171.
[48] Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 2010;11.
[49] Helping accelerate treatment for pediatric cancer with Dell cloud technology. <http://content.dell.com/us/en/corp/d/corp-comm/pediatric-cancer>.
[50] NextBio, Intel to collaborate on improving Hadoop Stack for Genomic Data Analysis. <http://www.genomeweb.com/informatics/nextbio-intel-collaborate-improving-hadoop-stack-genomic-data-analysis>.

[51] Cloudera Chief Scientist Jeff Hammerbacher Teams with Mount Sinai School of Medicine to Solve Medical Challenges Using Big Data. <http://www.marketwire.com/press-release/Cloudera-Chief-Scientist-Jeff-Hammerbacher-Teams-With-Mount-Sinai-School-Medicine-1676135.htm>.
[52] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet 2010;11:647–57.
[53] Healthcare Cloud Computing (Clinical, EMR, SaaS, Private, Public, Hybrid) Market – Global Trends, Challenges, Opportunities & Forecasts (2012–2017). <http://www.reportlinker.com/p0924631-summary/Healthcare-Cloud-Computing-Clinical-EMR-SaaS-Private-Public-Hybrid-Market-Global-Trends-Challenges-Opportunities-Forecasts-.html.
[54] Schoenherr S, Forer L, Weissensteiner H, Specht G, Kronenberg F, Kloss-Brandstaetter A. Cloudgene: a graphical execution platform for MapReduce programs on private and public clouds. BMC Bioinform 2012;13(1):200.
[55] Pennisi E. Will computers crash genomics? Science 2011;331:666–8.
[56] 1,000 Genomes in the Cloud and NCBI Experiences. <https://respond.niaid.nih.gov/conferences/bioinformatics2012/Festival%20Proceedings/Preuss_1000_Genomes.pdf>.
[57] Available at http://asperasoft.com/.
[58] How "Cloud" Services Democratize DNA Sequencing. <http://techonomy.com/2012/08/how-cloud-services-democratize-dna-sequencing/>.
[59] Schadt EE. The changing privacy landscape in the era of big data. Mol Syst Biol 2012;8:612.
[60] Creating HIPAA-Compliant Medical Data Applications With AWS. <http://aws.amazon.com/about-aws/whats-new/2009/04/06/whitepaper-hipaa/>.
[61] Managing data in the Cloud Age. <http://www.dddmag.com/articles/2012/10/managing-data-cloud-age>.
[62] Robertson JA. The $1000 genome: ethical and legal issues in whole genome sequencing of individuals. Am J Bioeth 2003;3. W-IF1.
[63] Klein CA. Cloudy confidentiality: clinical and legal implications of cloud computing in health care. J Am Acad Psychiatry Law 2011;39:571–8.
[64] Sleator RD, Shortall C, Hill C. Metagenomics. Lett Appl Microbiol 2008;47:361–6.
[65] Sleator RD. An overview of the processes shaping protein evolution. Sci Prog 2010;93:1–6.
[66] Sleator RD. Prediction of protein functions. Methods Mol Biol 2012;815:15–24.
[67] Sleator RD. Proteins: form and function. Bioeng Bugs 2012;3:80–5.
[68] Marianayagam NJ, Fawzi NL, Head-Gordon T. Protein folding by distributed computing and the denatured state ensemble. Proc Natl Acad Sci USA 2005;102:16684–9.
[69] Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, et al. Predicting protein structures with a multiplayer online game. Nature 2010;466:756–60.
[70] Murray JF. Personalized medicine: been there, done that, always needs work! Am J Respir Crit Care Med 2012;185:1251–2.
[71] Furusawa C, Kaneko K. A dynamical-systems view of stem cell biology. Science 2012;338:215–7.
[72] Sleator RD. The human superorganism – of microbes and men. Med Hypotheses 2010;74:214–5.
[73] O'Driscoll A, Sleator RD. Synthetic DNA: the next generation of big data storage. Bioengineered 2013:4.
[74] Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, et al. A whole-cell computational model predicts phenotype from genotype. Cell 2012;150:389–401.
[75] Sleator RD. Digital biology: a new era has begun. Bioengineered 2012;3:311–2.
[76] Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 2009;25:1363–9.
[77] Pireddu L, Leo S, Zanetti G. SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics 2011;27:2159–60.
[78] Blastreduce: high performance short read mapping with mapreduce. <http://www.cbcb.umd.edu/software/blastreduce/>.
[79] Schatz M, Sommer D, Kelley D, Pop M. De Novo assembly of large genomes with cloud computing. Biology of genomes, vol. 10. NY: Cold Spring Harbor; 2010.
[80] Chang YJ, Chen CC, Chen CL, Ho JM. A de novo next generation genomic sequence assembler based on string graph and MapReduce cloud computing framework. BMC Genomics 2012;13(Suppl. 7):S28.
[81] Jourdren L, Bernard M, Dillies MA, Le Crom S. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics 2012;28:1542–3.
[82] Niemenmaa M, Kallio A, Schumacher A, Klemela P, Korpelainen E, Heljanko K. Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. Bioinformatics 2012;28:876–7.
[83] Matthews SJ, Williams TL. MrsRF: an efficient MapReduce algorithm for analyzing large collections of evolutionary trees. BMC Bioinform 2010;11(Suppl. 1):S15.
[84] Colosimo ME, Peterson MW, Mardis S, Hirschman L. Nephele: genotyping via complete composition vectors and MapReduce. Source Code Biol Med 2011;6:13.
[85] Vouzis PD, Sahinidis NV. GPU-BLAST: using graphics processors to accelerate protein sequence alignment. Bioinformatics 2011;27:182–8.
[86] Liu CM, Wong T, Wu E, Luo RB, Yiu SM, Li YR, et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics 2012;28:878–9.
[87] Lewis S, Csordas A, Killcoyne S, Hermjakob H, Hoopmann MR, Moritz RL, et al. Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework. BMC Bioinform 2012;13.
[88] Matsunaga A, Tsugawa M, and Fortes J. CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications. IEEE Fourth International Conference on eScience, Indiana, USA, 2008 222-229.
[89] Leo S, Santoni F, Zanetti G. Biodoop: bioinformatics on hadoop. In: Parallel processing workshops, 2009. ICPPW '09. International Conference on; 2009. p. 415–22.
[90] Huang HL, Tata S, Prill RJ. BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters. Bioinformatics 2013;29:135–6.
[91] Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biol 2010;11:R116.
[92] Zhang L, Gu S, Liu Y, Wang B, Azuaje F. Gene set analysis in the cloud. Bioinformatics 2012;28:294–5.
[93] Feng X, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinform 2011;12:139.