Chapter 2

Cloud Computing for Next-Generation Sequencing Data Analysis

Shanrong Zhao, Kirk Watrous, Chi Zhang and Baohong Zhang

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/66732

Abstract

High-throughput next-generation sequencing (NGS) technologies have evolved rapidly and are reshaping the scope of genomics research. The substantial decrease in the cost of NGS techniques in the past decade has led to its rapid adoption in biological research and drug development. Genomics studies of large populations are producing a huge amount of data, giving rise to computational issues around the storage, transfer, and analysis of the data. Fortunately, cloud computing has recently emerged as a viable option to quickly and easily acquire the computational resources for large-scale NGS data analyses. Some cloud-based applications and resources have been developed specifically to address the computational challenges of working with very large volumes of data generated by NGS technology. In this chapter, we will review some cloud-based systems and solutions for NGS data analysis, discuss the practical hurdles and limitations in cloud computing, including data transfer and security, and share the lessons we learned from the implementation of Rainbow, a cloud-based tool for large-scale genome sequencing data analysis.

Keywords: next-generation sequencing, cloud computing, data analysis, workflow, pipeline

1. Introduction

High-throughput next-generation sequencing (NGS) technologies have evolved rapidly and are reshaping the scope of genomics research [1, 2] and drug development [3, 4]. The significant advances in NGS technologies, and consequently, the exponential expansion of biological data have created a huge gap between the computer capabilities and sequencing throughput [5, 6]. Technical improvements have greatly decreased the sequencing costs

and, as a result, the size and number of datasets generated by large sequencing centers have increased dramatically. The lower cost also made the sequencing data more affordable to small and midsize research groups. As always, digging out the "treasure" from NGS data is the primary challenge in bioinformatics, which places unprecedented demands on big data storage and analysis. It is becoming increasingly daunting for small laboratories or even large institutions to establish and maintain their own computational infrastructures for large-scale NGS data analysis.

A promising solution to address this computational challenge is cloud computing [7–10], where CPU, memory, and storage are accessible in the form of virtual machines (VMs). In recent years, cloud computing has spread very rapidly for the supply of IT resources (hardware and software) of different nature, and is emerging as a viable option to quickly and easily acquire the computational resources for large-scale NGS data analyses. Cloud computing offers a wide selection of VMs with different hardware specifications and users can choose and configure these VMs to meet their computational demands. With the massive scale of users, cloud computing providers, such as Amazon, are continuously driving costs down, which in turn has led to the use of cloud computing for NGS data analyses attractive within the bioinformatics community. Despite the apparent benefits associated with cloud computing, there are also issues to be addressed. Data privacy and security are particularly important when managing sensitive data, such as the patients' information from clinical genomics studies [11].

The aim of this chapter is to describe the application of cloud computing in large-scale NGS data analysis and to help scientists to understand advantages and disadvantages of cloud computing, and to make an informed-choice on whether to perform NGS analysis on cloud services or to build the infrastructure themselves. It is organized as follows. First, we give a brief introduction to NGS technology, including DNA sequencing, RNA sequencing, and ChIP-sequencing. Secondly, we briefly introduce cloud computing and its services. Thirdly, we summarize and review publicly available cloud-based NGS tools and systems, with some particular emphasis on "Rainbow" [12], a cloud-based tool for large-scale whole-genome sequencing. Finally, we will discuss the challenges and remaining problems related to the full adoption of cloud computing in the NGS data analysis.

2. Next-generation sequencing

Next-generation sequencing [13] platforms allow researchers to ask virtually any question related to the genome, transcriptome, or epigenome of any organism. It has already profoundly changed the nature and scope of genomic research in the past few years. Sequencing methods differ primarily by how the DNA or RNA samples are obtained (e.g., organism, tissue type, normal vs. affected, experimental conditions) and by the data analysis options used. After the sequencing libraries are prepared, the actual sequencing processes are similar regardless of the method. There are a number of standard library preparation kits from different vendors that offer solutions for whole-genome sequencing (WGS), RNA sequencing (RNA-seq), targeted sequencing (such as exome sequencing, targeted RNA-seq or 16S

http://dx.doi.org/10.5772/66732

sequencing), and detection of DNA methylation and protein-DNA interactions. As the number of NGS methods is constantly growing, a brief overview covering the most common methods is presented below.

2.1. Genomics

A breakthrough in NGS in the last decade has provided an unprecedented opportunity to investigate the contribution of genetic variation to health and disease [14]. WGS and whole-exome capture sequencing (WES) have emerged as compelling paradigms for routine clinical diagnosis, genetic risk prediction, and rare diseases [15–18]. WGS of tumors [19] is an unbiased approach that provides extensive genomic information about a tumor at the single nucleotide level as well as structural variations such as large insertions, genomic rearrangements, gross deletions, and duplications. Using low-coverage WGS of many individuals from diverse human populations, the 1000 Genomes Project [20] has characterized common variations and a considerable proportion of rare variations present in human genomes. With falling costs, it is now possible to sequence genomes of many individuals for association studies and other genomic analyses [21].

The WGS workflow is depicted in **Figure 1**. A human genome is fragmented into many short pieces that are sequenced by a sequencer. The sequencing step typically generates billions of short reads. All short reads are mapped to a reference genome, and genetic and structural variants can be identified with respect to the reference genome sequence. Human DNA is comprised of approximately 3 billion base pairs. 30× coverage sequencing of a personal

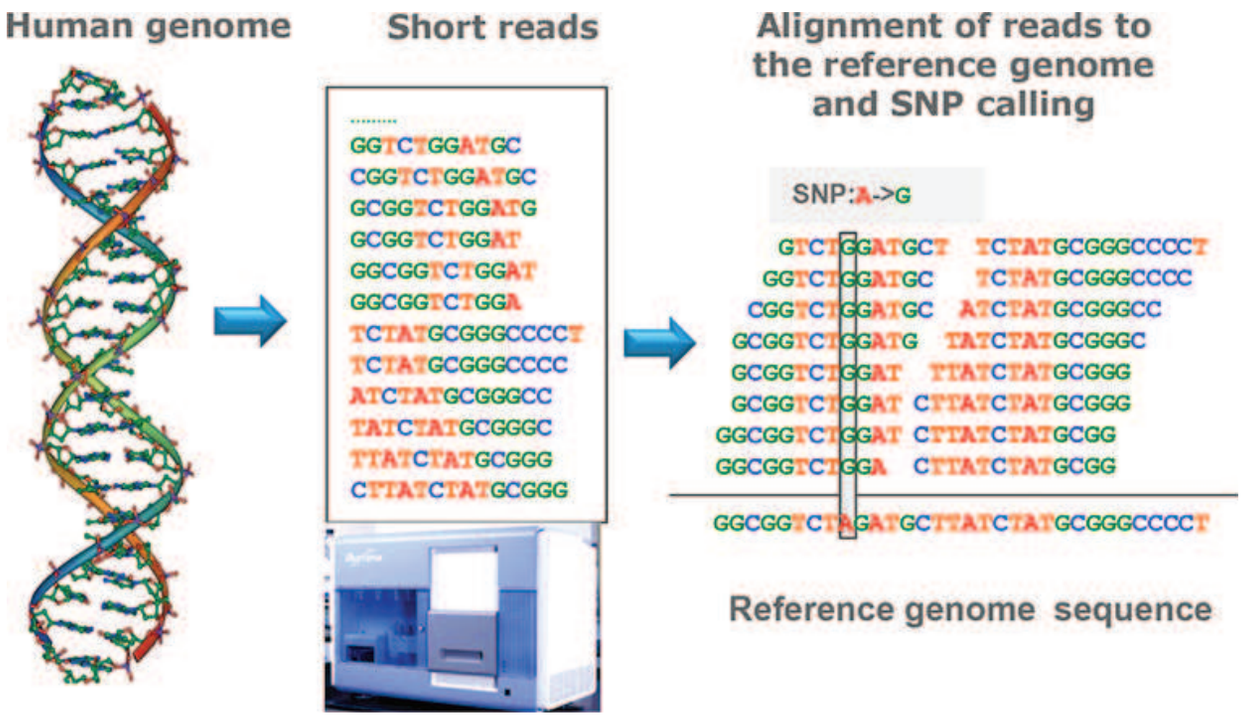

Figure 1. WGS workflow. A human genome is fragmented and sequenced, and billions of short reads are generated by a sequencer. All short reads are aligned to the reference genome and genetic variants are identified accordingly.

genome will produce approximately 100 gigabytes (GB) of nucleotide bases, and its corresponding FASTQ file will be about 250 GB. For a WGS project consisting of 400 subjects, 100 terabytes of disk space is required to store the raw reads alone. Additional space is required for storing intermediate files generated during data analyses. Transferring and processing a dataset of such size would be extremely time-consuming and heavily computation-intensive and thus they pose huge practical challenges in data analyses.

2.2. Transcriptomics

RNA sequencing (RNA-seq) has emerged as a powerful technology for transcriptome profiling [22–25]. It allows both quantification of known or predefined RNA transcripts and the capability to detect and quantify rare and novel transcripts within a sample. Compared to microarray, RNA-seq has a broader dynamic range, which allows for the detection of more differentially expressed genes with higher fold-change [26]. It is also superior in detecting low abundance transcripts, differentiating biologically critical isoforms, and allowing the identification of genetic variants. Not only RNA-seq can detect underlying genomic alterations at single-nucleotide resolution within expressed regions of the genome, but also it can quantify expression levels and capture variation not detected at the genomic level, including the expression of alternative transcripts. In the past decade, RNA-seq has become one of the most versatile applications of NGS technology and has revolutionized the researches on transcriptome [27]. As in WGS, RNA-seq generates vast number of short reads that must be computationally aligned or assembled to quantify expression of hundreds of thousands of RNA transcripts. Similar to DNA sequencing, the enormous data from large-scale RNA-seq studies poses a fundamental challenge for data management and analysis in a local environment [28–30]. Consequently, limited access to computational infrastructure and high-quality bioinformatics tools, and the demand for personnel skilled in data analysis and interpretation, remains a serious bottleneck for most researchers.

2.3. Epigenomics and protein-DNA interactions

While genomics involves the study of heritable or acquired alterations in the DNA sequence, epigenetics is the study of heritable changes in gene activity caused by mechanisms other than DNA sequence changes [31, 32]. Mechanisms of epigenetic activity include DNA methylation, histone modification and more. A focus in epigenetics is the study of cytosine methylation (5-mC) states across specific areas of regulation such as promotors or heterochromatin. Cytosine methylation can significantly modify temporal and spatial gene expression and chromatin remodelling. Two methylation sequencing methods are widely used: whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS). With WGBS-seq, sodium bisulfite chemistry converts nonmethylated cytosines to uracils, which are then converted to thymines in the sequence reads. In RRBS-seq, DNA is digested with MspI—a restriction enzyme unaffected by methylation status. Fragments in the 100–150 bp size range are isolated to enrich CpG and promotor containing DNA regions. Sequencing libraries are then constructed using the standard NGS protocols.

ChIP-sequencing, also known as ChIP-seq [33, 34], is a method used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be

http://dx.doi.org/10.5772/66732

used for genome-wide mapping of transcription factor binding sites. Protein-DNA interactions have a significant impact on many biological processes and disease states. The sequence reads generated by ChIP-seq are massive and need to be aligned to reference genome first, and then the locations of protein-DNA interactions are inferred based on enrichment of sequence reads along the genome.

3. Cloud computing

"Cloud Computing," by definition, refers to the on-demand delivery of IT resources and applications via the Internet with pay-as-you-go pricing. Cloud computing is a model for enabling ubiquitous, on-demand access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services), which can be rapidly provisioned and released with minimal management effort. With cloud computing, you do not need to make large upfront investments in hardware and spend a lot of time on the heavy lifting of managing hardware. Instead, cloud computing providers such as Amazon Web Services own and maintain the network-connected hardware, and you can provision exactly the right type and size of computing resources you need. You can access as many resources as you need, almost instantly, and only pay for what you request and own. These computing resources include networks, servers, storage, applications, and services. There are several essential characteristics of the cloud computing model.

a. Rapid elasticity: you only allocate resources when you need them, and you are able to dynamically scale-up and -down your allocated resources as your needs change over time.

b. Pay-as-you-go: you only pay when you consume computing resources, and only pay for how much you consume.

c. On-demand self-service: the user can request and manage the computing resources without help from the service providers.

d. Cost-effective: classical computational infrastructure for data processing has become ineffective and difficult to easily scale-up and -down, and cloud computing is a viable and even a cheaper technology that enables large-scale data analysis.

Existing cloud-based services can be classified into four categories or layers (see **Figure 2**). The first one is Infrastructure as a Service (IaaS). This service model is offered in a computing infrastructure that includes servers (typically virtualized) with specific computational capability and/or storage. The user has full control on the operating system and applications that are deployed to, but with limited control, over the network settings. A good example is Amazon elastic compute cloud (EC2), which allows the user to request and manage virtual machines, and Amazon simple storage service (S3), which allows storing and accessing data. The second category of service is Platform as a Service (PaaS) in which the provider offers the customer the authority to create applications using developing tools supported by the provider. PaaS features rapid application development and good scalability, presenting usefulness in developing specific applications for big biological data analysis. Typically, the environment delivered by PaaS includes programming language environments, web servers,

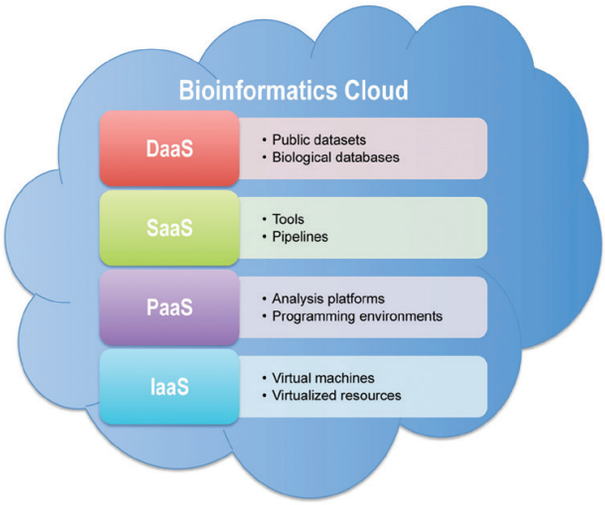

Figure 2. Illustration of cloud services [8]. Cloud-based services in are grouped into Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

and databases. The Amazon Web Services (AWS) software development kit (AWS SDK) and Google App Engine are good examples of this service.

The third service is Software as a Service (SaaS). SaaS eliminates the need for local installation and eases software maintenances and updates, providing up-to-date cloud-based services for data analysis. Customers do not manage the cloud infrastructure or network components, servers, operating systems, or storage, and can use the applications provided by the cloud provider. Most bioinformatics applications are open-source projects, and difficult to build, configure and maintain, primarily because they lack good documentation and have complex library dependencies. However, as all software applications are installed and configured within the VM, SaaS provides the perfect solution. The fourth layer is Data as a Service (DaaS). Bioinformatics clouds are heavily dependent on data, as data are fundamentally crucial for downstream analyses and knowledge discovery. Due to such unprecedented growth in biological data, delivering Data as a Service (DaaS) via the Internet is of utmost importance. DaaS enables dynamic data access and provides up-to-date data that are accessible by a wide range of devices that are connected over the web. AWS provides a centralized repository of public data sets, including GenBank [35], 1000 Genomes [20], encyclopedia of DNA elements [36], etc., and all public datasets are delivered as services in AWS and thus can be seamlessly integrated into cloud-based applications.

4. NGS data analysis on cloud computing

In recent years, cloud computing offers an alternative approach to quickly and easily acquire computational resources for large-scale NGS data analysis. As a result, many cloud-based services and bioinformatics platforms (see **Table 1**), applications (see **Table 2**), and resources have been developed to address the specific challenges of working with the large volumes of data generated by NGS technology. Cloud computing has created new possibilities to analyze NGS data at reasonable costs, especially for laboratories lacking a dedicated bioinformatics

infrastructure. From the perspective of end users, there are three options to analyze NGS data on cloud computing (**Tables 1** and **2**). First, commercial systems such as DNAnexus and Seven Bridges can be used out of box to carry out the entire NGS data analysis. Second, commercial or open bioinformatics platforms are further customized to meet users' computational needs. Third, open-source tools (**Table 2**) can be deployed into the cloud for any customized data analysis.

4.1. Commercial services

Commercial services provide the users with well-established pipelines, user interfaces, and even application programming interfaces (APIs), and can reduce the time and effort required for setting up pipelines for NGS data analysis. For instance, DNAnexus and Seven Bridges offer various customizable NGS data analysis pipelines. In addition, DNAnexus also provides software that can directly upload the sequencing data produced. BaseSpace, launched by Illumina in collaboration with Amazon, is a genomics cloud computing platform that provides NGS data analysis services, such as mapping, de novo assembling, small RNA analysis, library quality control (QC), metagenomics analysis, and data storage. It is designed to bring simplified data management and analytical sequencing tools directly to researchers in a user-friendly manner. BaseSpace provides flexibility and convenience with an array of tools, significantly simplifying the process of yielding meaningful results from NGS data. Bina Technologies offers a service that is composed of a specialized hardware called Bina Box and a cloud service. Bina Box can employ accelerated BWA [54] and GATK [55] for data analyses.

Some commercial services also provide APIs with which the users can manage their jobs or build their own applications. Variant calling on datasets of hundreds or thousands of genomes is time-consuming, expensive, and not easily reproducible given the myriad components of a variant calling pipeline. To address these challenges, the Mercury [47] analysis pipeline was

Name	URL	Description
BaseSpace	http://basespace.illumina.com	Commercial services
Bina	http://www.bina.com/	Commercial services
DNAnexus	http://www.dnanexus.com	Commercial services
SevenBridges	http://www.sbgenomics.com	Commercial services
Eoulsan	http://transcriptome.ens.fr/eoulsan	Cloud-based platform
CloVR	http://clovr.org	Automated sequence analysis
Cloud BioLinux	http://cloudbiolinux.org	Virtual machine for bioinformatics cloud computing
CloudMan	https://wiki.galaxyproject.org/CloudMan	Cloud-scale Galaxy
Globus Genomics	https://www.globus.org/genomics	Cloud-based bioinformatics workflow for NGS analyses
GenomeCloud	http://www.genome-cloud.com/	Analyze genome data
COSMOS	http://cosmos.hms.harvard.edu/	Workflow management system

Table 1. Cloud computing services and platforms.

Software	URL	Description	Reference
Atlas2	http://atlas2cloud.sourceforge.net	Genome analysis	[37]
CloudAligner	http://cloudaligner.sourceforge.net	Reads mapping	[38]
CloudBurst	http://cloudburst-bio.sourceforge.net	Reads mapping	[39]
Crossbow	http://bowtie-bio.sourceforge.net/crossbow	Read mapping/SNP call	[40, 41]
FX	http://fx.gmi.ac.kr	RNA-seq	[42]
Myrna	http://bowtie-bio.sourceforge.net/myrna	RNA-seq	[43]
Stormbow	http://s3.amazonaws.com/jnj_stormbow/index.html	RNA-seq	[44]
STORMSeq	http://www.stormseq.org/	Read mapping	[45]
GenomeKey	https://github.com/LPM-HMS/GenomeKey	Whole genome analysis	[46]
Mercury	https://www.hgsc.bcm.edu/software/mercury	Workflow for genomic analysis	[47]
Rainbow	http://s3.amazonaws.com/jnj_rainbow/index.html	Whole genome analysis	[12]
PaekRanger	http://ranger.sourceforge.net	ChIP-seq	[48]
VAT	http://vat.gersteinlab.org	Variant annotation	[49]
YunBe	http://tinyurl.com/yunbedownload	Gene set analysis	[50]
BioVLAB-MMIA-NGS	https://sites.google.com/site/biovlab/	microRNA-mRNA integrated analysis	[51, 52]
SURPI	http://chiulab.ucsf.edu/surpi/	Pathogen identification	[53]

Table 2. Open-source tools for cloud computing.

developed on top of the DNAnexus platform. It integrates multiple sequence analysis components across various computational steps, from obtaining patient samples to providing a fully annotated list of variant sites for clinical applications. Mercury is an automated, flexible, and extensible analysis workflow that provides accurate and reproducible genomic results at scales ranging from individuals to large cohorts.

Although a number of cloud-based pipelines are available for analyses of sequencing data in massively parallel DNA sequencing, the majority of them can only identify variants within a single sample. While this approach has enough power for detecting variants in high-coverage sequencing, it performs worse than multiple-sample calling when applied to low-coverage sequencing data. To this end, another scalable DNAnexus-based pipeline for joint variant calling in large samples was developed and deployed to the Amazon cloud. Using this pipeline, Shringarpure et al. [21] identified 68.3 million variants in 2535 samples from Phase 3 of the 1000 Genomes Project. By performing the variant calling in a parallel manner, the data was processed within 5 days at a compute cost of just $7.33 per sample (a total cost of $18,590 for completed jobs and $21,805 for all jobs).

Despite their merits, these commercial services also have several disadvantages. First, the use of a commercial service requires extra expenses for the convenience of NGS data analysis and user-friendly interfaces. Second, compared to open-source tools on the cloud, the commercial

services are less customizable with respect to the use of the services and access to the cloud service. Although DNAnexus and Seven Bridges provide APIs to access and control their cloud services, their functionalities are restricted and the users have to request the service provider to set up new application software on their cloud services.

4.2. Bioinformatics platforms

CloudBioLinux [56] is a publicly accessible virtual machine (VM) that is based on an Ubuntu Linux distribution and is available to all Amazon EC2 users for free. It comes with a user-friendly graphical user interface (GUI), along with over 135 preinstalled bioinformatics packages. CloudBioLinux instances provide an excellent environment for users to become familiar with BioLinux and cloud computing. Galaxy is an open, web-based platform for data-intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share the entire data analyses. Galaxy Cloud [57], a cloud-based Galaxy platform for the analysis of data at a large scale, is the most used platform for bioinformatics. Unlike commercial software service solutions, users can customize their deployment and have complete control over their instances and associated data. Currently, a public Galaxy Cloud deployment, called CloudMan [57], is provided on the AWS cloud, enables bioinformatics researchers to easily deploy, customize, and share their cloud analysis environment, including data, tools, and configurations. By combining three platforms, CloudBioLinux, CloudMan, and Galaxy, into a cohesive unit, researchers can gain access to more than 135 preconfigured bioinformatics tools and gigabytes of reference genomes on top of the flexible cloud computing infrastructure [56].

Although Galaxy cloud provides a convenient platform for researchers, challenges remain in moving large amounts of data reliably and efficiently and in adding domain-specific tools for specific analyses. To address these challenges, Globus Genomics [58, 59] was developed at the Computation Institute (CI), a joint institute between the University of Chicago and Argonne National Laboratory. Globus Genomics is a cloud-based integrated solution for NGS data analysis. It extends the existing Galaxy workflow system by adding data management capabilities for transferring large quantities of data efficiently and reliably (via Globus Transfer), domain-specific analyses tools preconfigured for immediate use by researchers (via user-specific tools integration), automatic deployment on cloud for on-demand resource allocation and pay-as-you-go pricing (via Globus Provision), and a cloud provisioning tool for auto-scaling (via HTCondor scheduler). Genome sequencing is notoriously data-intensive, and Globus Transfer [59] is designed for fast and secure movement of large amounts of data. Setting up a production instance of Galaxy is a nontrivial task that involves a number of manual installation and configuration steps for both the platform and any dependent software packages—steps that can be both error-prone and time-consuming. Globus Provision addresses the above issues by providing on-demand cluster reconfiguration, user-specific node provisioning, and automatic instance deployment on Amazon EC2.

GenomeCloud (http://www.genome-cloud.com/) is another on-demand Galaxy cloud. It was built upon Galaxy, and is composed of g-Analysis, g-Cluster, g-Storage, and g-Insight services, providing convenient services to the researchers and other users. GenomeCloud is a complete and integrated platform for analyzing genome data to the interpretation of analysis

results. It combines the idea of cloud computing with bioinformatics to generate an integrated solution for data storage and sharing, database management, continuously updated computing and analysis tools, and security. GenomeCloud is designed to help researchers perform bioinformatics tasks more easily, as well as to support laboratories without the computational resources to conduct research without hurdles.

4.3. Open-source tools

The development of tools supporting NGS data analysis with cloud computing has recently become popular in the open-source community [8]. Currently, there are many pipelines and workflows that support cloud computing (**Table 2**). Despite their advantages in cost and flexibility, open-source tools on the cloud also have substantial drawbacks. The users are responsible for designing/setting up the entire analysis pipeline, the data management and hardware configuration, such as CPUs, memory, storage, and security. Quite often, the users have to overcome a laborious series of trial and error before setting up the proper configuration. Although several tools have been developed to date, in most cases, their cloud computing support is incomplete and their functionality is under-developed. Here, we will briefly report some existing bioinformatics tools and then describe Rainbow, a cloud-based tool for large-scale WGS data analysis, in detail in the next section.

CloudAligner [38] and CloudBurst [39] are parallel read mapping algorithms optimized for mapping short reads to human and other reference genomes and can produce alignments for a variety of downstream biological analyses including SNP discovery, genotyping, and personal genomics. Crossbow is a Hadoop- [60] based tool that combines the speed of the short read aligner bowtie [61], with the accuracy of the SNP caller SOAPsnp [62] to perform alignment and SNP detection from WGS data in parallel. Scalable tools for open-source read mapping (STORMseq) [45] is a graphical interface cloud computing solution that performs read mapping, read cleaning, variant calling, and annotation using personal genome data. Variant annotation tool (VAT) [49] has been developed to annotate variants from multiple personal genomes at the transcript level as well as to obtain summary statistics across multiple genes and individuals.

FX [42] is an RNA-seq analysis tool, which runs in parallel on cloud computing infrastructure, for the estimation of gene expression levels and genomic variant calling. Another cloud computing pipeline for calculating differential gene expression in large RNA-seq datasets is Myrna [43]. Myrna uses bowtie [61] for short read alignment and R/bioconductor for quantification, normalization, and statistical testing. These tools are combined in an automatic, parallel pipeline that runs in the cloud, exploiting the availability of multiple computers and processors wherever possible. Stormbow [44] is a scalable, cost-effective, and open-source-based tool for large-scale RNA-seq data analysis. Its performance has been tested by applying it to analyze 178 RNA-seq samples in the cloud. In the test, it took 6–8 h to process each RNA-seq sample with 100 million pair-ended reads in the m1.xlarge instance, and the average cost was only $3.50 per sample. BioVLAB-MMIA-NGS [51, 52] offers the integrated miRNA-mRNA analysis and can be used to identify the "many-to-many" relationship between miRNAs and target genes with high accuracy. PeakRanger [48]

http://dx.doi.org/10.5772/66732

is a software package for the analysis of ChIP-seq data. It can be run in a parallel cloud computing environment to obtain extremely high performance on large data sets. Unbiased NGS approaches enable comprehensive pathogen detection in the clinical microbiology laboratory [63] and have numerous applications for public health surveillance, outbreak investigation, and the diagnosis of infectious diseases. Sequence-based ultra rapid pathogen identification (SURPI™) [53] is a computational pipeline for pathogen identification from complex metagenomic NGS data generated.

4.4. Rainbow

Crossbow is a software tool that can detect SNPs in WGS data from a single subject; however, it has a number of limitations when applied to large-scale WGS projects. Rainbow [12] is a cloud-based software package that can assist in the automation of large-scale WGS data analyses. Rainbow was built upon Crossbow. By hiding the complexity of the Crossbow command-line options, Rainbow facilitates the application of Crossbow for large-scale WGS analysis in the cloud. Compared with Crossbow, the main improvements incorporated into Rainbow include the ability: (1) to handle BAM as well as FASTQ input files, (2) to split large sequence files for better load balance in downstream clusters, (3) to collect and track the running metrics of data processing and monitoring multiple Amazon EC2 instances, and (4) to merge SOAPsnp outputs from multiple individuals into a single file to facilitate downstream genome-wide association studies.

The workflow of Rainbow is shown in **Figure 3**. Multiple data drives are shipped to Amazon. After the BAM or FASTQ files have been uploaded to S3, large FASTQ files are split into smaller files in parallel. Then multiple clusters are launched in the cloud, with each cluster processing a single sample. Crossbow is responsible for mapping reads to the reference sequence and for SNP calling. The SNPs for all samples are then combined by a Perl script. When the analysis is complete, the results can either be downloaded directly or exported via Amazon Export. We applied Rainbow to analyze the 44 subjects, with 0.55–1 billion pair-ended 100 bp short reads per sample. The running environments were as follows. For step #1 in **Figure 3**, we chose the Amazon m1.large instance, which has two CPUs, 7.5 GB memory, and two 420 GB instance drives. For Crossbow run, each compute cluster has 40 c1.xlarge nodes as recommended by the Crossbow developers. Each c1.xlarge node has 8 CPUs, 7 GB memory, and 1690 GB instance storage.

The performance of Rainbow is summarized in **Figure 4**. In a 320-CPU (=40 × 8) cluster, the alignment of billions of reads takes between 0.8 and 1.6 h. The linear relationship shown in **Figure 4** indicates that the sequence data blocks in the Hadoop distributed file system (HDFS) are physically local to the nodes that processed them, which reduces virtual I/O delays. The SOAPsnp running time ranges from 1 to 1.8 h, which takes a little longer than the alignment. All EC2 instances and clusters are terminated immediately after the jobs are finished. On average, it costs less than 120 US dollars to analyze each subject, and the total cost for analyzing those 44 subjects was 5800 US dollars, including data import. More important than the cost is the ability to scale Rainbow up or down, so that the analyses can be accomplished in a reasonably short amount time, regardless of sample size. No upfront investment in

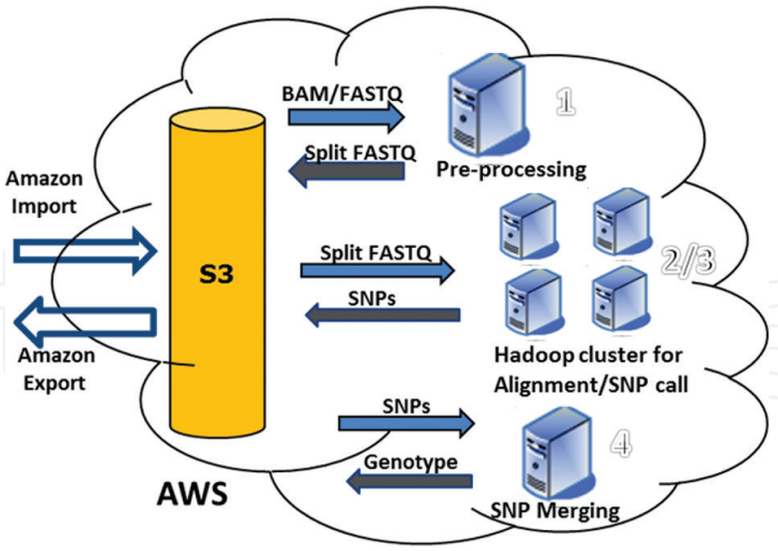

Figure 3. The Rainbow pipeline. S3 centralizes data storage, including inputs, intermediate results, and outputs. Alignment and SNP call are performed by Crossbow in a cluster with multiple nodes.

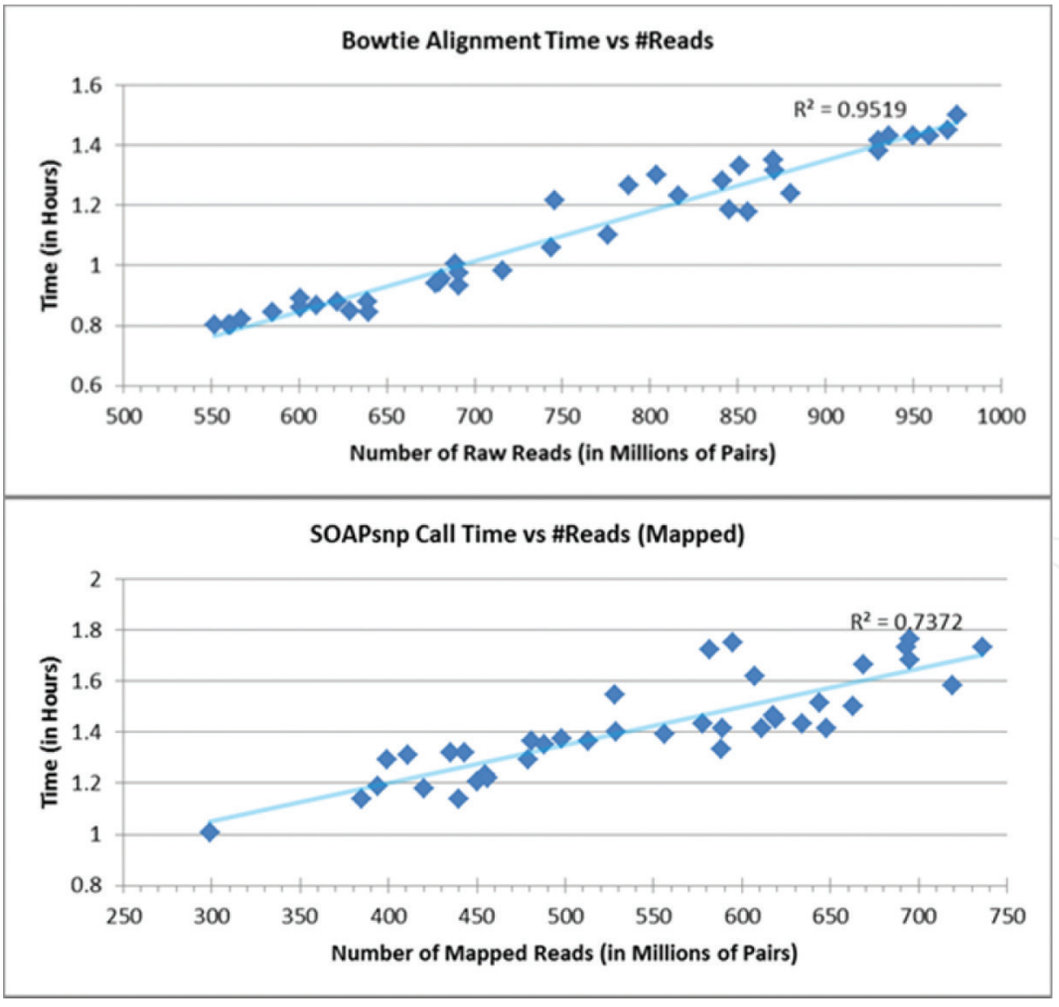

Figure 4. Top panel: running time of bowtie alignment vs. the number of paired sequence reads. Bottom panel: running time of SOAPsnp vs. the number of paired mapped reads. Note: each cluster consists of 40 c1.xlarge instances (8 CPUs per c1.xlarge instance).

http://dx.doi.org/10.5772/66732

infrastructure is required and there is no additional administrative costs involved using Amazon cloud. Rainbow is a scalable, cost-effective, and open-source tool for large-scale WGS data analysis. It is available for third-party implementation and use, and can be downloaded from the Rainbow website.

In order to access the Rainbow cloud pipeline, the user must first set up an AWS account (http://aws.amazon.com/). Once registered, the user needs to sign up for Amazon EC2, S3, EMR, and SES services. The user can then start an instance based on the public AMI: ami-0f1f9866 in US-East (N. Virginia); or ami-b6bc89f3 in US-West (N. California). All required software is already preinstalled and configured in the AMI. Then, the user can connect to the instance and configure EC2, EMR, and S3cmd command-line tools. After a successful connection to the instance has been established, the user needs to prepare a sample manifest file in order to run Rainbow. A master manifest file is a plain text file to describe all subjects in a WGS project. Each subject has a corresponding entry in the manifest file and each entry consists of three fields separated by spaces or tabs: (1) a unique identifier; (2) locations of the raw reads in S3; and (3) an output folder in S3. Each individual step in the Rainbow workflow uses this same manifest file as input, thus all output files are named and stored consistently. After the creation of the manifest file, the user just needs to run a couple of command lines and all the analyses will be done automatically in the cloud.

Analyzing large datasets in the cloud is different from performing the same analysis in a local environment. When developing Rainbow, we have learned the following lessons.

- It is not trivial to move large datasets around within the cloud. Users should be prepared to handle network congestion or failures, never assuming data transfer is always successful. If data move fails, it is a good practice to wait a couple of minutes and then try to transfer data again. Repeat the "wait-and-transfer" cycle till data transfer is successful.
- Boot time should be taken into account when new resources are starting up. It is recommended to give cloud providers 10–15 min grace period (waiting time) before attempting to use a newly requested resource, for instance, elastic block store (EBS).
- Cloud providers typically offer a variety of compute instances to choose from. To choose the best option, it is important to understand the bottleneck (CPU, I/O, or network) for the algorithm that is to be run.
- When large amounts of data are moved between cloud resources, it is essential to ensure that they are in the same location or data center.
- It is difficult to debug workflows in the cloud without heavy logging.

5. Cloud computing hurdles

Albeit relatively new, cloud computing holds great promise in effectively addressing big data storage and analysis problems in NGS data analysis. Despite the potential gains achieved, there are also several important issues that need to be addressed. Below, we present the main hurdles on the adoption of cloud computing.

5.1. Big data transfer

To analyze the NGS data in the cloud, data have to be transferred across the wired network and uploaded onto AWS. The volume and complexity of NGS data have exponentially increased, giving rise to issues related to data analysis, management, and transfer to the cloud [64–66]. For example, WGS of 400 subjects at 30× coverage will generate approximately 100 TB raw sequence reads in FASTQ format. In the future, more and more sequencing projects would generate ultra-large volumes of biological data and thus require bioinformatics clouds for big data storage, sharing, and analysis. One of the most challenging issues of cloud computing is data transfer. Transferring vast amounts of biological data to the cloud is a significant bottleneck in cloud computing.

The speed of data transfer is usually slow and at present there are not many solutions available for moving the huge amount of data to cloud. Therefore, we need more efficient data transfer technologies in cloud computing. According to CloudHarmoy's report on download speed relative to the year 2010 (http://blog.cloudharmony.com/2010/02/cloud-speed-test-results.html), the download speed from Amazon AWS EC2 in North Virginia (U.S.) was 2.95 Mb/s, which corresponds to downloading a 10 GB file in 29,116 s (≈8 h). Therefore, data transfer is a serious bottleneck in NGS data analysis on cloud service. To deal with the data transfer issue, Aspera (http://www.asperasoft.com/) has developed the fast and secure protocol (FASP) for data transfer with a speed of up to 5 GB/s. Ideally, using FASP, the user can download a 10 GB file in 17.2 s, which is a revolutionary improvement. But still it cannot transfer data at the TB scale. Alternatively, sequencing service providers such as BGI and Illumina offer a service in which they deliver a hard disk drive (HDD) containing the sequencing data. However, the shipping has limitations aside from the time taken up by travel, and once a package (of hard drives) is stamped and sealed, researchers cannot control when (and sometimes if) the parcel arrived or what condition it arrived. This service cannot guarantee safe transfer of data, since the HDD can get lost, stolen, or physical errors can occur.

5.2. Most bioinformatics tools are not cloud-aware

Most bioinformatics software tools are written for desktop (rather than cloud) applications and are therefore not provided as cloud-based web services accessible via the Web, making it infeasible to perform complex bioinformatics tasks in the cloud. For instance, bowtie [61] is one of the most popular mapping algorithms, but it requires that input files are stored on local disk when mapping reads and is not compatible with Amazon S3. Even if you run bowtie in an EC2 instance, the raw FASTQ files have to be fetched to an elastic block store (EBS) volume that is attached to the EC2 instance. In other words, bowtie does not have built-in support for S3. Spliced transcripts alignment to a reference (STAR) [67–69] is a popular RNA-seq mapper that performs highly accurate spliced sequence alignment at an ultrafast speed. However, it is not cloud-friendly either. Like bowtie, STAR does not take advantage of AWS cloud services, and cannot work with S3 either. Unfortunately, the majority of bioinformatics tools are developed without native support for cloud computing.

MapReduce [70, 71], developed by Google, is an easy-to-use and general-purpose parallel programming model that is suitable for large data set analysis on a commodity hardware

http://dx.doi.org/10.5772/66732

cluster. MapReduce is a software framework, written in Java, designed to run over a cluster of machines in a distributed way. A MapReduce program is composed of a user-defined map function and a reduce function. When a program that is implemented with the map and reduce functions has been launched, the map function processes each key/value pair and produces a list of intermediate key/value pairs, while the reduce function aggregates all the intermediate values with the same keys. MapReduce is an important advancement in cloud computing because it can process huge data sets quickly and safely using commodity hardware.

Hadoop, comprised of MapReduce and the Hadoop distributed file system (HDFS), is based on a strategy of colocating data and processing to significantly accelerate computing performance [60]. Hadoop allows for the distributed processing of large datasets across multiple computer nodes, supports big data scaling, and enables fault-tolerant parallel analysis. The Hadoop framework has been recently deemed as the most suitable method for handling bioinformatics data [70]. Unfortunately, many traditional bioinformatics tools and algorithms have to be redesigned and implemented in order to support and benefit from Hadoop MapReduce infrastructure. Even with the help of the corresponding developers, it will take a while for most bioinformatics tools currently available to add this feature.

Apache Spark™ (https://spark.apache.org/) is a fast and general engine for large-scale data processing, natively supported in Amazon EMR. Apache Spark supports a variety of languages, including Java, Scala, and Python, for developers to build applications. Hadoop and Apache Spark are both big data frameworks, but they do not really serve the same purposes. Hadoop is essentially a distributed data infrastructure. It distributes massive data collections across multiple nodes within a cluster of commodity servers, Spark, on the other hand, is a data-processing tool that operates on those distributed data collections; it does not do distributed storage. To study the utility of Apache Spark in the genomic context, SparkSeq was created [72]. It is a general-purpose, flexible, and easily extendable library for genomic cloud computing, and can be used to build genomic analysis pipelines in Scala and run them in an interactive way. Recently, SparkBWA [73] was introduced; a new tool that exploits Spark to boost the performance of one of the most widely adopted sequence aligner, the Burrows-Wheeler Aligner (BWA). It is hoped more Apache Spark-based bioinformatics algorithms will be developed for large-scale genomic data analysis in the future.

5.3. Open clouds for bioinformatics

Currently, the largest cloud computing provider is Amazon, which provides commercial clouds for processing big data. Additionally, Google also provides a cloud platform to allow users to develop and host applications, and to store and analyze data. However, commercial clouds are not yet able to provide ample data and software for bioinformatics analysis. By placing public biological database and software into the cloud and delivering them as services, data and software can be seamlessly and easily integrated into the cloud. AWS hosts a variety of public data sets for free access (https://aws.amazon.com/public-data-sets/). All public datasets in AWS are delivered as services. Previously, large data sets, such as the mapping of the human genome, required hours or days to locate, download, customize, and analyze.

Now, anyone can access these data sets via the AWS centralized data repository from any Amazon EC2 instance or Amazon EMR cluster. Google Genomics also helps the life science community organize the world's genomic data and make them accessible and useful.

In the era of big data, however, only a tiny amount of biological data is accessible in the cloud at present (only AWS, including GenBank [35], Ensembl [74], 1000 Genomes [20], etc.) and the vast majority of data are still deposited in conventional biological databases. It is difficult for commercial clouds to keep pace with the emerging needs from academic research, opening up the demand for specific open clouds for bioinformatics studies. Needless to say, open access and public availability of data and software are of great significance to biological science. To satisfy the need for big data storage, sharing, and analysis with lower cost and higher efficiency, it is essential that a large number of biological data as well as a wide variety of bioinformatics tools should be publicly accessible in the cloud and delivered as services. Therefore, future efforts should be devoted to building open bioinformatics clouds for the bioinformatics community. GenomeSpace [75] is a cloud-based, cooperative community resource that currently supports the streamlined interaction of 20 bioinformatics tools and data resources. To facilitate integrative analysis by nonprogrammers, it offers a growing set of 'recipes', short workflows to guide investigators through high-utility analysis tasks. The potential benefits of open bioinformatics clouds include maximizing the scope for data sharing, easing large-scale data integration, and harnessing collective intelligence for knowledge discovery.

5.4. Security and privacy

The many characteristics of cloud computing have made the long-dreamed vision of "computing as a utility" a reality. The cloud computing offers scalable and competitively priced computing resources for the analysis and storage of data from large-scale genomics studies, but it must also ensure that genetic data coming from human subjects are hosted in a context that is both secure and compliant with regulations [76]. When deciding whether to move the analyses into the cloud or not, potential cloud users need to weigh all the factors including system performance, service availability, cost, and most importantly, data security. Genomics data extracted from clinical samples are sensitive data and present unprecedented requirements of privacy and security [11, 77]. In general, there are concerns that genomics and clinical data managed through a cloud are susceptible to loss, leakage, theft, unauthorized access, and attacks. The centralized storage and shared tenancy of physical storage space means the cloud users are at higher risk of disclosure of their sensitive data to unwanted parties. A secure protection scheme will be necessary to protect the sensitive information from medical records. There is considerable amount of work to enforce data protection against security attacks.

However, the question of security in cloud computing is intrinsically complicated. Cloud computing is built on the top of existing architectures and techniques such as SaaS and distributed computing. When combining all the benefits of these architectures and techniques, cloud computing also inherits almost all of their security issues at various levels of the system stack. When cloud users move their applications from within their enterprise/organization boundary into the open cloud, they will lose physical control over their data, and traditional security protection mechanisms such as firewalls are no longer applicable to cloud applications. As a

result, cloud users have to heavily rely upon the service providers for data privacy and security protection. In cloud computing, the data and applications from different customers reside on the same physical computing resources. This fact will inevitably bring forth more security risks in the sense that any intentional or inadvertent misbehavior by one cloud user would make other coresidences victims.

6. Conclusion

Pharmacogenomics is an important branch of genomics that studies the impact of SNP on drug response in patients, the toxicity or efficacy of a drug, and the development of diseases [78, 79]. Pharmacogenomics aims to improve drug therapy with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects, and is at the basis of the idea of "precision medicine" where drugs are chosen or optimized to meet the genetic profile of each patient. Thus, the presence (or the absence) of specific SNPs may be used as a clinical marker to predict drug effectiveness, and to foresee the drug responses of individuals with different SNPs. Precision medicine also opens up the possibility to treat diseases based on the genetic makeup of the patient. Identification of the genetic defect underlying early-onset diabetes is important for determining the specific diabetes subtype, providing personalized treatment. The study by Artuso et al. [14] has revealed that NGS provides a highly sensitive method for identification of variants in a new-set of "driver genes" causing diabetes. NGS can be used to analyze the comprehensive landscape of genetic alterations, including known disease-causing gene fusions in transcripts, which brings new insights to study diseases with a highly complex and heterogeneous genetic composition such as cancer [19]. Therefore, NGS facilitates precision medicine and changes the paradigm of cancer therapy, and holds expanded promise for its diagnostic, prognostic, and therapeutic applicability in various diseases.

The substantial decrease in the cost of NGS techniques in the past decade has dramatically reshaped the biomedical research and has led to its rapid adoption in biological research and drug development [3, 4]. Nowadays, massive amount of data, targeting a variety of biological questions, can be generated quickly using NGS platforms. These data range from the function and regulation of genes, the clinical diagnosis and treatment of diseases, to the omics profiling of individual patients for precision medicine. To better understand the association between SNPs and diseases, and to gain deeper insights into the relation between drug response and genetic variations, large-scale sequencing projects are continuously being initiated in research institutes and pharmaceutical companies. The availability of NGS and the genomics studies of large populations are producing an increasing amount of data. However, the storage, preprocessing, and analysis of NGS data are becoming the main bottleneck in the analysis pipeline. With the exponential increase in volume and complexity of NGS data, cluster or high performance computing (HPC) systems are essential for the analysis of large amounts of NGS data. But the associated costs with the infrastructure itself and the maintenance personnel will likely be prohibitive for small institutions or laboratories, and even too much to swallow for big institutions and pharmaceutical companies.

This chapter provides a useful perspective on cloud computing and helps researchers, who plan to analyze their NGS data on cloud services, to gain an understanding of the basic concept. As discussed earlier, cloud computing drives down infrastructure costs both up-front and on an on-going basis. It offers many operational advantages, such as fast completion of massive computational projects and infrastructures can be set up in minutes rather than months. There are growing demands for cloud-based NGS data analysis, which constitutes a good alternative for researchers with little interest in investing in a cluster system. Cloud computing is becoming a technology mature enough for its use in genomic researches. The use of large datasets, the demanding analysis algorithms, and the urgent need for computational resources, make large-scale sequencing projects an attractive test-case for cloud computing. It is likely that in the future, researchers will be able to analyze their NGS data on cloud services for a cost low enough to eliminate the need to introduce clusters into their laboratories. However, we must emphasize that there are a number of open issues and problems associated with the use of cloud, such as privacy and security, especially when managing patients' clinical data. Data transfer of large amounts of NGS data to the cloud remains a challenge and bottleneck in NGS data analysis. Furthermore, many traditional NGS tools are not designed to use in cloud environment, and thus cannot harness the benefits of cloud services. With the ever-increasing demand, commercial companies and bioinformaticians will develop more NGS analysis tools, using APIs offered by service providers, in the foreseeable future, making NGS data analysis in the cloud more efficient, friendly, and secure.

Author details

Shanrong Zhao[1]*, Kirk Watrous[2], Chi Zhang[1] and Baohong Zhang[1]

*Address all correspondence to: shanrong.zhao@pfizer.com

1 Early Clinical Development, Pfizer Worldwide Research and Development, Cambridge, MA, USA

2 Business Technology, Pfizer Worldwide Research and Development, Groton, Connecticut, USA

References

[1] Bahassi el M, Stambrook PJ. Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis. 2014;29(5):303-10.

[2] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333-51.

[3] Woollard PM, Mehta NA, Vamathevan JJ, Van Horn S, Bonde BK, Dow DJ. The application of next-generation sequencing technologies to drug discovery and development. Drug Discov Today. 2011;16(11-12):512-9.

[4] Yadav NK, Shukla P, Omer A, Pareek S, Srivastava AK, Bansode FW, et al. Next generation sequencing: potential and application in drug discovery. Sci World J. 2014;2014:802437.

[5] Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. Nat Biotechnol. 2010;28(7):691-3.

[6] Baker M. Next-generation sequencing: adjusting to data overload. Nat Meth. 2010;7(7):495-9.

[7] Calabrese B, Cannataro M. Bioinformatics and microarray data analysis on the cloud. Meth Mol Biol. 2016;1375:25-39.

[8] Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct. 2012;7:43; discussion

[9] O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform. 2013;46(5):774-81.

[10] Kwon T, Yoo WG, Lee W-J, Kim W, Kim D-W. Next-generation sequencing data analysis on cloud computing. Genes & Genomics. 2015;37(6):489-501.

[11] Datta S, Bettinger K, Snyder M. Secure cloud computing for genomic data. Nat Biotechnol. 2016;34(6):588-91.

[12] Zhao S, Prenger K, Smith L, Messina T, Fan H, Jaeger E, et al. Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing. BMC Genomics. 2013;14:425.

[13] Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586-97.

[14] Artuso R, Provenzano A, Mazzinghi B, Giunti L, Palazzo V, Andreucci E, et al. Therapeutic implications of novel mutations of the RFX6 gene associated with early-onset diabetes. Pharmacogenomics J. 2015;15(1):49-54.

[15] Allard MW. The future of whole-genome sequencing for public health and the clinic. J Clin Microbiol. 2016;54(8):1946-8.

[16] Edwards D, Batley J, Snowdon RJ. Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet. 2013;126(1):1-11.

[17] Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52(4):413-35.

[18] Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.

[19] Xue Y, Wilcox WR. Changing paradigm of cancer therapy: precision medicine by next-generation sequencing. Cancer Biol Med. 2016;13(1):12-8.

[20] Zheng-Bradley X, Flicek P. Applications of the 1000 genomes project resources. Brief Funct Genomics. 2016; pii: elw027. [Epub ahead of print]

[21] Shringarpure SS, Carroll A, De La Vega FM, Bustamante CD. Inexpensive and highly reproducible cloud-based variant calling of 2,535 human genomes. PLoS One. 2015;10(6):e0129277.

[22] Nagalakshmi U, Waern K, Snyder M. RNA-seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol (edited by Frederick M Ausubel [et al]). 2010;Chapter 4:Unit 4.11.1-3.

[23] Mutz KO, Heilkenbrinker A, Lonne M, Walter JG, Stahl F. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol. 2013;24(1):22-30.

[24] Zhao S, Zhang B, Zhang Y, Gordon W, Du S, Paradis T, et al. Bioinformatics for RNA-Seq Data Analysis. In: Abdurakhmonov I, editor. Bioinformatics—Updated Features and Applications: InTech; 2016. pp. 125-49.

[25] Picelli S. Single-cell RNA-sequencing: the future of genome biology is now. RNA Biol.2016;21:1-14.

[26] Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9(1):e78644.

[27] Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17(5):257-71.

[28] Zhao S, Xi L, Quan J, Xi H, Zhang Y, von Schack D, et al. QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. BMC Genomics. 2016;17:39.

[29] Hoeijmakers WA, Bartfai R, Stunnenberg HG. Transcriptome analysis using RNA-seq. Methods Mol Biol. 2013;923:221-39.

[30] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133-45.

[31] Jeong HM, Lee S, Chae H, Kim R, Kwon MJ, Oh E, et al. Efficiency of methylated DNA immunoprecipitation bisulphite sequencing for whole-genome DNA methylation analysis. Epigenomics. 2016;8(8):1061-77.

[32] Li N, Ye M, Li Y, Yan Z, Butcher LM, Sun J, et al. Whole genome DNA methylation analysis based on high throughput sequencing technology. Methods. 2010;52(3):203-12.

[33] Massie CE, Mills IG. Mapping protein-DNA interactions using ChIP-sequencing. Meth Mol Biol. 2012;809:157-73.

[34] Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Brief Bioinformatics. 2016; pii: bbw023. [Epub ahead of print]

[35] Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016;44(D1):D67-72.

[36] Qu H, Fang X. A brief review on the human encyclopedia of DNA elements (ENCODE) project. Genomics, Proteomics Bioinformatics. 2013;11(3):135-41.

[37] Evani US, Challis D, Yu J, Jackson AR, Paithankar S, Bainbridge MN, et al. Atlas2 Cloud: a framework for personal genome analysis in the cloud. BMC Genomics. 2012;13(Suppl 6):S19.

[38] Nguyen T, Shi W, Ruden D. CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes. 2011;4:171.

[39] Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics. 2009;25(11):1363-9.

[40] Gurtowski J, Schatz MC, Langmead B. Genotyping in the cloud with crossbow. Genome Biol. 2009;10.

[41] Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Genotyping in the cloud with crossbow. Curr Protoc Bioinformatics (edited by Andreas D Baxevanis [et al]. 2012;Chapter 15:Unit15.3.

[42] 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(5):721-3.

[43] Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 2010;11(8):R83.

[44] Zhao S, Prenger K, Smith L. Stormbow: a cloud-based tool for reads mapping and expression quantification in large-scale RNA-seq studies. ISRN Bioinformatics. 2013;2013:481545.

[45] Karczewski KJ, Fernald GH, Martin AR, Snyder M, Tatonetti NP, Dudley JT. STORMSeq: an open-source, user-friendly pipeline for processing personal genomics data in the cloud. PLoS One. 2014;9(1):e84860.

[46] Souilmi Y, Lancaster AK, Jung JY, Rizzo E, Hawkins JB, Powles R, et al. Scalable and cost-effective NGS genotyping in the cloud. BMC Med Genomics. 2015;8:64.

[47] Reid JG, Carroll A, Veeraraghavan N, Dahdouli M, Sundquist A, English A, et al. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics. 2014;15:30.

[48] Feng X, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics. 2011;12:139.

[49] Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A, et al. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics. 2012;28(17):2267-9.

[50] Zhang L, Gu S, Liu Y, Wang B, Azuaje F. Gene set analysis in the cloud. Bioinformatics. 2012;28(2):294-5.

[51] Chae H, Rhee S, Nephew KP, Kim S. BioVLAB-MMIA-NGS: microRNA-mRNA integrated analysis using high-throughput sequencing data. Bioinformatics. 2015;31(2):265-7.

[52] Lee H, Yang Y, Chae H, Nam S, Choi D, Tangchaisin P, et al. BioVLAB-MMIA: a cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2. IEEE Trans Nanobioscience. 2012;11(3):266-72.

[53] Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. 2014;24(7):1180-92.

[54] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.

[55] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.

[56] Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, et al. CloudBioLinux: preconfigured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics. 2012;13:42.

[57] Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J. Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics. 2010;11(Suppl 12):S4.

[58] Bhuvaneshwar K, Sulakhe D, Gauba R, Rodriguez A, Madduri R, Dave U, et al. A case study for cloud based high throughput analysis of NGS data using the globus genomics system. Comput Struct Biotechnol J. 2015;13:64-74.

[59] Madduri RK, Sulakhe D, Lacinski L, Liu B, Rodriguez A, Chard K, et al. Experiences building globus genomics: a next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services. Concurr Comput. 2014;26(13):2266-79.

[60] Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics. 2010;11(Suppl 12):S1.

[61] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.

[62] Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19(6):1124-32.

[63] Bragg L, Tyson GW. Metagenomics using next-generation sequencing. Meth Mol Biol. 2014;1096:183-201.

[64] Marx V. Biology: the big challenges of big data. Nature. 2013;498(7453):255-60.

[65] Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1(2):293-314.

[66] Mardis ER. The challenges of big data. Dis Models Mech. 2016;9(5):483-5.

[67] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

[68] Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics (edited by Andreas D Baxevanis [et al]). 2015;51:11.4.1-9.

[69] Dobin A, Gingeras TR. Optimizing RNA-seq mapping with STAR. Meth Molecular Biol. 2016;1415:245-62.

[70] Zou Q, Li XB, Jiang WR, Lin ZY, Li GL, Chen K. Survey of MapReduce frame operation in bioinformatics. Brief Bioinformatics. 2014;15(4):637-47.

[71] Mohammed EA, Far BH, Naugler C. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. BioData Min. 2014;7:22.

[72] Wiewiorka MS, Messina A, Pacholewska A, Maffioletti S, Gawrysiak P, Okoniewski MJ. SparkSeq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision. Bioinformatics. 2014;30(18):2652-3.

[73] Abuin JM, Pichel JC, Pena TF, Amigo J. SparkBWA: speeding up the alignment of high-throughput DNA sequencing data. PLoS One. 2016;11(5):e0155461.

[74] Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44(D1):D710-6.

[75] Qu K, Garamszegi S, Wu F, Thorvaldsdottir H, Liefeld T, Ocana M, et al. Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. Nat Methods. 2016;13(3):245-7.

[76] Aldeen YA, Salleh M, Aljeroudi Y. An innovative privacy preserving technique for incremental datasets on cloud computing. J Biomed Inform. 2016; 62:107-16.

[77] Dove ES, Joly Y, Tasse AM, Knoppers BM. Genomic cloud computing: legal and ethical points to consider. Eur J Hum Genetics. 2015;23(10):1271-8.

[78] Ortega VE, Meyers DA. Pharmacogenetics: implications of race and ethnicity on defining genetic profiles for personalized medicine. J Allergy Clin Immunol. 2014;133(1):16-26.

[79] Rabbani B, Nakaoka H, Akhondzadeh S, Tekin M, Mahdieh N. Next generation sequencing: implications in personalized medicine and pharmacogenomics. Mol Biosyst. 2016;12(6):1818-30.

IntechOpen