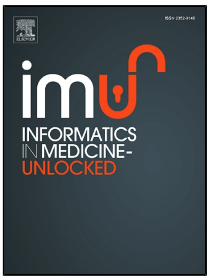

Comprehensive Comparison of Cloud-Based NGS Data Analysis and Alignment Tools

Qanita Bani Baker, Mahmoud Hammad, Wesam Al-Rashdan, Yaser Jararweh, Mohammad AL-Smadi, Mohammad Al-Zinati

PII:	S2352-9148(19)30367-3
DOI:	https://doi.org/10.1016/j.imu.2020.100296
Reference:	IMU 100296
To appear in:	*Informatics in Medicine Unlocked*
Received Date:	22 November 2019
Accepted Date:	17 January 2020

Please cite this article as: Qanita Bani Baker, Mahmoud Hammad, Wesam Al-Rashdan, Yaser Jararweh, Mohammad AL-Smadi, Mohammad Al-Zinati, Comprehensive Comparison of Cloud-Based NGS Data Analysis and Alignment Tools, *Informatics in Medicine Unlocked* (2020), https://doi.org/10.1016/j.imu.2020.100296

Comprehensive Comparison of Cloud-Based NGS Data Analysis and Alignment Tools

Qanita Bani Baker[a,*], Mahmoud Hammad[a], Wesam Al-Rashdan[a], Yaser Jararweh[a,b], Mohammad AL-Smadi[a], Mohammad Al-Zinati[a]

[a] *College of Computer and Information Technology Jordan University of Science and Technology, Irbid, Jordan 22110*
[b] *Duquesne University, Pittsburgh, PA, USA*

Abstract

Next-Generation Sequencing (NGS) is very helpful for conducting DeoxyriboNucleic Acid (DNA) Sequencing. DNA sequencing is the process for determining the order (sequence) of the main chemical bases in the DNA. Analyzing human DNA sequencing is important for determining the possibility that a person will develop certain diseases, and/or the ability to respond to medication. However, the NGS process is a complicated and resource-hungry technical process. To solve this dilemma, the majority of NGS software systems are deployed as cloud-based services distributed over cloud-based platforms. Cloud-based platforms provide promising solutions for the computationally intensive tasks required by the NGS data analysis. This work provides a comprehensive investigation of cloud-based NGS data analysis and alignment tools, both the commercial and the open-source tools. We also discuss in detail the main features and setup requirements for each tool, and then compare and contrast between them. Moreover, we extensively analyze and classify the studied NGS data analysis and alignment tools to help NGS biomedical researchers and clinicians in finding appropriate tools for their work, while understanding the similarities and the differences between them.

Keywords:
Next-Generation Sequencing (NGS), Sequence Alignment, Cloud Computing, Big Data, Bioinformatics

1. Introduction

DeoxyriboNucleic Acid (DNA) is a nucleic acid that contains all of the genetic instruction for an organism. Each molecule of DNA contains a chemical base which could be one of four types: adenine (A), cytosine (C), thymine (T), and guanine (G). These chemical bases or letters are the main building blocks of the DNA. A person has about 3 billion pairs of these letters, with the exact order or sequence being called genomic sequence. DNA sequencing is the process that enables scientists to read the exact order or sequence of all letters that make up the complete set of DNA, the genome. Thereafter, the DNA sequence is compared to a standardized code to identify the variance between the two sets of letters.

There are many benefits of DNA sequencing including determining the possibility of a person for developing certain diseases such as cancer, heart disease, or type II diabetes. It also can determine

*Corresponding author
Email address: qmbanibaker@just.edu.jo (Qanita Bani Baker)

the ability of a person to respond to certain medications, a technique known as *pharmacogenomics*. Moreover, some genomic disorders provide an indication that a person may develop rare cases of disease such as Huntington disease (a progressive brain disorder). The Next-Generation Sequencing (NGS) is defined as a massively parallelized sequencing technology that produces high-throughput DNA reading sequencing at a comparatively minimal cost [1]. NGS is considered a cutting-edge technologies in biological and biomedical research [2] [3].

NGS differs from Sanger sequencing, which is also called first-generation sequencing [4], in sequencing volume, in cost, in the velocity of sequencing, and in the amount of DNA data produced [5]. Compared with the third-generation sequencing techniques, NGS also differs [6] in many characteristics as discussed in [7, 8, 9, 10]. As presented in [11], Single-Molecule Sequencing (SMS), simple divergence from previous technologies, enabling a single molecule sequencing, and real-time sequencing are all characterizing as third-generation sequencing platforms. NGS technologies are evolved in clinical tests [12] and in revolutionary innovations to genomic studies [13]. Recently, NGS technologies have become routine procedures in Biotechnology research. DNA sequencing data analysis is a core procedure of the diagnostic test [14]. NGS provides considerable low-cost alternatives for several applications [15].

Advances in NGS technologies have resulted in an unprecedented proliferation of genomic sequence data. Therefore, several NGS data-related challenges are presented, especially in analytics and storage [16]. Therefore, to enhance the performance of the NGS tasks, the majority of NGS software systems are deployed as cloud-based services distributed over cloud-based platforms. Cloud computing is an emerging technology that provides a different infrastructure for tackling computational challenges in NGS data analysis [17, 18, 19]. Generally, the cloud computing service models are classified into three types: Platform as a Service (PaaS), Software as a Service (SaaS), and Infrastructure as a Service (IaaS) [20], [21]. In addition to these models, the bioinformatic cloud services that are mentioned in [4], [22], [23], and [24] added another model to the three models, called Data as a Service known (DaaS). Moreover, many studies also divide the types of the cloud into three main categories (private, public, and hybrid) [25]. In this paper, we extended our preliminary work in [26] and we have studied 47 cloud-based tools that are widely used in NGS data analysis, instead of only 20 NGS tools that have been studied elsewhere [26]. In addition, we also included 13 open-source and cloud-based NGS alignment tools that are widely used in the NGS data analysis. Moreover, in this paper, we described the functionality, features, and setup requirements for each NGS and alignment tool. We present a comparison between these tools in order to assist clinicians and researchers to choose the appropriate tools according to their working environments. To the best of our knowledge, this study is the first comprehensive study that investigates the widely used commercial and open-source cloud-based NGS and alignment tools.

The remainder of the paper includes the following sections: in Section 2, we discuss research efforts related to our work. Then, in Section 3, we present our methodology for conducting this study. In Section 4, we discuss the open, commercial, and alignment tools operating on various cloud services. Finally, in Section 6, we conclude the paper and suggest future directions.

2. Related Work

In this section, we present some of the previous works that have focused on studying NGS platforms and tools. We show the works that have investigated NGS platforms. Then, we highlight studies that have introduced the NGS data as a big data domain, and we refer to the studies that utilized cloud computing technology in NGS. At the end of this section, we introduce other studies that provide overviews of NGS tools in the cloud, and we discuss the shortcomings in these studies that motivated us in this work.

Recently, many Next-generation Sequencing platforms have emerged [13, 3] such as ABI SOLiD, Illumina GA, and Roche 454. However, there are great needs for more NGS tools, as stated in [5]. Till now, NGS technologies have been successfully applied to several applications, such as RNA-sequencing

[27], ChIP-sequencing [28], whole human genome sequencing [29], and genome-wide structural variation [30]. Many studies have evaluated and investigated NGS platforms in several applications [31], [32], [33], [34] [13]. For example, in [32], Loman et al. evaluated the performance of three different NGS platforms which are MiSeq (from Illumina), 454 GS Junior (from Roche), and Ion Torrent PGM (from Life Technologies). They compared these platforms using several criteria including quality, the performance of the platforms, read length, read error rate, and completeness. Based on many previous studies, Illumina platforms are the most used and well-known platforms in the market [35, 36, 37].

The big-data generated from NGS technologies causes challenges with storing, analyzing, and managing this data [17]. Elazhary [20] presented the opportunities and challenges of using cloud computing in processing Big Data. Elazhary [20] also presented computational biology applications as targeted fields. The size and complexity of NGS data have grown rapidly, and the extension of computing capabilities is becoming essential [17]. Cloud computing is considered as a replacement of the current on-premises solutions to address several issues in NGS data, as shown in [17, 23]. The importance of cloud computing to handle NGS data analysis is discussed in many works [16], [4], [38]. Dai et al. [22] reviewed cloud-based services in the bioinformatics domain, and they classified them into DaaS, PaaS, SaaS, and IaaS. Then they presented their perspectives on utilizing cloud computing in the bioinformatics arena. In their review, they did not focus on the NGS, and they reviewed some of the cloud-based resources in bioinformatics.

Several efforts were taken to study and provide overviews of NGS analysis tools. In [16], Celesti et al. provided a taxonomy of the NGS cloud-based tools according to cloud service levels. They presented a taxonomic tree of cloud-based systems by showing NGS applications. Thakur et al. [18] also presented cloud-based computing in biological systems, focusing on genomic informatics, comparative genomics, metagenomics, and SNP detection. Moreover, Zhao et al. [23] reviewed a part of cloud-based tools and systems used for NGS data analysis. They discussed the practical limitations and hurdles that can be found in cloud computing, focusing on security and data transfer. They also showed bioinformatics platforms and cloud-based services along with some applications. They classified these platforms into commercial systems, commercial or open bioinformatics platforms, and open-source tools. Geo et al. [4] presented an overview of Cloud Computing and they showed how cloud-computing services provide support for NGS data analysis. They provided a summary of some cloud-based resources used for NGS data analysis. In addition to these works, Kwon et al. also used two types to classify the tools for NGS i.e., commercial services and open-source tools [17].

None in previous works have addressed features and setup requirements for the NGS cloud-based tools. From investigating the previous works, we find it is essential to develop a study that addresses cloud-based tools for NGS analysis. Our study provides a comparison between the tools which help clinicians, scientists, and researchers to choose the proper tool according to working environments. To the best of our knowledge, this study is the first comprehensive study that investigates the commercial and open-source cloud-based NGS and alignment tools, and provides up-to-date investigation of them. Sixty NGS tools are reported in this study, divided as 41 open-source NGS tools, 6 commercial NGS tools, and 13 NGS alignment tools.

3. Methodology

In this section, we describe our methodology for collecting NGS data analysis tools and NGS alignment tools. We focus only on cloud-based NGS tools, due to their capability for handling big data generated from NGS technologies. We also focus on NGS alignment tools that also can be deployed on cloud services. Then, we describe the main features that biomedical researchers and clinicians consider for selecting NGS and alignment tools. Although these tools are cloud-based, most of them offer a desktop version for users to download and utilize.

3.1. Selected NGS Tools

We include widely used cloud-based NGS tools collected from previous studies [16], [4], [17], and [23]. The NGS tools that we have excluded are either non-cloud-based NGS tools such as the BioPerl tool [39] or cloud-based NGS tools but not active or supported anymore such as the Roundup tool [40] and CloudTSS tool [41]. As a result of these inclusion/exclusion criteria, we obtained 60 cloud-based NGS data analysis and alignment tools, divided into three categories as follows: 41 open-source NGS tools, 6 commercial NGS tools, and 13 NGS alignment tools.

Due to space limitations and to ensure readability, we list all of the included open-source NGS tools, commercial NGS tools, and NGS alignment tools in Table 3.1. However, we have made the description of all of these tools available online [42].

Open source NGS			Commercial NGS	Alignment NGS
Tool	**Tool**	**Tool**	**Tool**	**Tool**
Galaxy [43] [44]	CloudBurst [45]	Crossbow [46]	BaseSpace [47]	BWA [48]
SeqMapReduce [49]	DIYA [50]	GATK [51] [52]	Bina [53]	SAMtools [54]
Myrna [55]	Ergatis [56]	CloVR [57] [58] [58] [59]	DNAnexus [60]	MAQ [61] [62]
Cloudaligner [63]	RAPSearch2 [64]	Jnomics [65] [66]	LifeScope [67]	BLAT [68] [69]
PeakRanger [70]	ArrayExpressHTS [71]	SIMPLEX [72]	GeneSifter [73]	BLAST [74] [75] [76]
Rainbow [77]	MEGAN [78]	Stormbow [79]	SevenBridges [80]	MUMmer GPU 2.0 [81]
BioPig [82]	Eoulsan [83]	Atlas2 [84]		MUMmer [85] [86] [87] [88]
TREAT [89]	Cloud BioLinux [90]	HugeSeq [91]		SHRiMP [92] [93]
VAT [94]	FX [95]	YunBe [96]		Bowtie [97]
CloudMan [98] [99]	Hadoop-BAM [100]	SparkSeq [101]		Bowtie2 [102]
BioVLAB-MMIA-NGS [103]	Contrail [104]	Mercury [105]		SEAL [106]
STORMSeq [107]	SURPI [108]	SeqPig [109]		TopHat [110]
SNP2Structure [111]	Halvade [112]	CLUSTOM-CLOUD [113]		HISAT2 [114]
MG-RAST [115] [116]	MC-GenomeKey [117]			

Table 1: All of the included NGS tools

3.2. Descriptive Features

To describe and compare between the included NGS tools, we have selected a set of features that are crucial for identifying cloud-based NGS data analysis tools. These features are also important for biomedical researchers and clinicians to select NGS tools. For each category of NGS tool, recall that in Section 3.1, we have identified a set of features related to that category, as shown in Figure 1. We extracted these features in several ways, including studying the research paper for each tool, exploring the tools' websites, studying tools' manuals, and utilizing the OMICtools [118]. OMICtools is a website that contains a manually curated metadatabase of omic tools.

3.2.1. Features for open-source NGS tools

For this category, we have selected nine important features. Next, we describe these features in more detail:

1. Operating System: describes the operating system the desktop version of the open-source NGS tool requires to operate.
2. NGS technology: displays the NGS technology platform(s) or the hardware equipment the NGS tool is compatible with.
3. Cloud Service Model (CSM): shows the model of the cloud service for the NGS tool.
4. Cloud Type: displays the type of cloud service for the NGS tool (private, public, or hybrid).

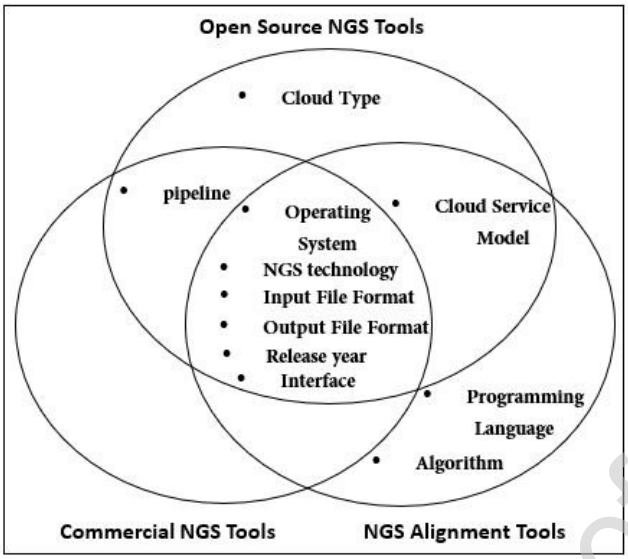

Figure 1: The descriptive features for all cloud-based NGS and alignment tools

5. Input File Format: displays the format for the accepted input file to the NGS tool.
6. Output File Format: displays the format for the output file generated by the NGS tool.
7. Release year: shows the released year for the NGS tool.
8. Interface: displays the type of the interface that the user can use to interact with the NGS tool. The possible interfaces for the NGS tools are: Web User Interface (WUI), Command Line Interface (CLI), Graphical User Interface (GUI), and Application Programming Interface (API).
9. Pipeline: this feature shows if the tool uses a pipeline technique or not. The Pipeline is defined as the steps used for analyzing the data when the output from one step is an input to the next step.

3.2.2. Features for commercial NGS tools

Many features that we have identified for open source NGS tools cannot be obtained for commercial tools such as the Cloud Service Model and Cloud Type features. Therefore, we ended up with only several features for this category. These features are: (1) Operating System, (2) NGS technology, (3) Input File Format, (4) Output File Format, (5) Release year, and (6) Interface. The description of these features are the same as in the open-source NGS tools.

3.2.3. Features for NGS Alignment tools

For this category, we have identified nine features. These nine features are Operating System, NGS technology, Cloud Service Model (CSM), Input File Format, Output File Format, Release year, Interface, Programming Language (PL), and Algorithm. The description of the first seven features are as described in the open-source NGS tools, whereas the last two are described as follows:

1. Programming Language (PL): displays the programming language used to develop the NGS alignment tool.
2. Algorithm: displays the NGS alignment algorithm the tool performs to analyze and align the data.

4. Results

This section presents the results of our methodology for extracting the features of the commonly used NGS and alignment tools.

In this work, we investigated 41 open-source NGS tools obtained from applying inclusion/exclusion criteria shown in the Methodology section. It is difficult to fit all tools in one table, so we divided the NGS tools into dual five year time spans based on the tool's published year. Table 2 covers the tools published in the years from 2005 to 2012, while Table 3 shows the tools published in the years from 2013 to 2017. Both tables 2 and 3 depict the nine features, described in Section 3.2.1, of the 41 selected open-source NGS tools. As shown in Tables 2 and 3, all of the desktop versions of the open-source tools work with the Linux/Unix operating system. However, eight of them work with Windows and eight of them work with the MacOS operating system.

Regarding the NGS technology used, most of them are compatible with Illumina technology. In terms of the utilized Cloud Service Model, most of the tools are deployed as Software as a Service (SaaS). However, six of them are Infrastructure as a Service (IaaS) and eight of them are Platform as a Service (PaaS). As shown in the table, 30 of the selected NGS tools are hosted on the Amazon web services as a public cloud type.

Tables 2 and 3 show that most of the open-source NGS tools accept FASTA and FASTQ file format as input files. However, they generate files with different formats as an output. The table also shows that most of the tools have been released between the years 2009 and 2012. Thereafter, fewer tools have been developed.

Regarding the interface of the open-source NGS tools, most of them interact with the user via a Command Line Interface (CLI) in which the user needs to become familiar with the commands for that tool. Finally, as the table shows, most of the tools use pipelines to analyze the NGS data.

ID	NGS Tool	Operating System	NGS technology	CSM	Cloud Type	Input File	Output File	Release year	Interface	Pipeline
1	Galaxy	Windows, Unix, Linux	-	PasS	-	Various *	Various *	2005	WI	✓
2	Cloud-Burst	Unix, Linux	Illumina, Solexa	SaaS	Public (Amazon EC2)	FASTA	BED	2009	CLI	
3	Crossbow	Unix, Linux	Illumina	SaaS	Public (Amazon EC2)	FASTQ	Stream of SNP calls	2009	CLI	✓
4	Seq-Map-Reduce	Unix, Linux	Illumina, Solexa	SaaS	Public (Amazon EC2)	-	ELAND	2009	WI	
5	DIYA	Unix, Linux	Roche/454 Sequencing GS-FLX instruments	IaaS	-	FASTA	Various *	2009	CLI	✓
6	GATK	Unix, Linux, MacOS	Illumina/ HiSeq , Biosystems SOLiD System, 454 Life Sciences	SaaS, IaaS	-	BAM, SAM	VCF	2010, 2011	CLI	✓
7	Myrna	Unix, Linux	Illumina Genome Analyzer II	SaaS	Public (Amazon EC2)	FASTQ	-	2010	CLI	✓
8	ERGATIS	Unix, Linux	-	IaaS	-	FASTA, XML	XML	2010	CLI, WI	✓

9	CLOVR	Windows, Unix, Linux, MacOS	Roche/454, Illumina	IaaS	Public (Amazon EC2), DIAG cloud	SFF, FASTA, QUAL, FASTQ	FASTA, Genbank Flat files	2011	CLI , GUI	✓
10	Cloud-Aligner	Unix, Linux	Illumina HiSeq 2000	SaaS	Public (Amazon EC2)	FASTA, FASTQ, SAM	SAM, BED6	2011	CLI	
11	RAP-Search2	Unix, Linux	-	IaaS	Public (Amazon EC2)	FASTA	XML, ASN.1	2011	CLI	✓
12	Jnomics	Unix, Linux	Illumina	SaaS	-	BAM, SAM, FASTQ, BED	SAM	2011	CLI	✓
13	Peak-Ranger	Windows, Linux, MacOS	-	SaaS	Public (Amazon EC2)	Eland, Bowtie, SAM, BAM, BED	WIG	2011	CLI	
14	Array-Express HTS	Windows, Unix, Linux, MacOS	Illumina, Solexa	SaaS	-	FASTQ	HTML	2011	CLI	✓
15	SIMPLEX	Unix, Linux	Illumina, ABI SOLiD	-	Public (Amazon EC2)	FASTQ, FASTA	Pdf, BAM, VCF, TSV, PNG, xlsx	2012	CLI	✓
16	Eoulsan	Unix, Linux	Illumina	PaaS	Public (Amazon EC2)	FASTQ, FASTA	-	2012	CLI	
17	Atlas2	Unix, Linux	Roche/454, Illumina, SOLiD	SaaS	Public (Amazon EC2, S3), Genboree Workbench	BAM, FASTA	VCF, LFF	2012	CLI	✓
18	TREAT	Unix, Linux	-	-	Public (Amazon EC2)	FASTQ, BAM	Various *	2012	CLI	✓
19	Cloud Bio Linux	Windows, Unix, Linux, MacOS	-	PaaS, IaaS	Public (Amazon EC2)	BAM	-	2012	CLI	✓
20	HugeSeq	Unix,Linux	Illumina HiSeq	-	-	FASTQ, FASTA	VCF, GFF	2012	CLI	✓
21	VAT	Unix, Linux	-	SaaS	Public (Amazon EC2)	VCF	VCF	2012	CLI	✓
22	FX	Unix, Linux	Illumina Genome Analyzer IIx	SaaS	Public (Amazon EC2)	FASTQ	Various *	2012	CLI	✓

23	YunBe	Unix, Linux	BGI	SaaS	Public (Amazon EC2), BGI	-	-	2012	CLI	
24	CloudMan	Unix, Linux	-	PaaS	Public Amazon EC2, private (OpenStack and OpenNebula)	FASTQ	-	2012	CLI, WI	✓
25	Hadoop-BAM	Unix, Linux	-	SaaS	-	BAM, SAM, FASTQ, FASTA, QSEQ, BCF, VCF	BAM, SAM, FASTQ, FASTA, QSEQ, BCF, VCF	2012	CLI	

Table 2: Open-source cloud-based NGS tools developed between years (2005-2012)

ID	NGS Tool	Operating System	NGS technology	CSM	Cloud Type	Input File	Output File	Release year	Interface	Pipeline
26	Rainbow	Linux	Illumina HiSeq 2000, HiSeq 2500 platforms	SaaS	Public (Amazon EC2 and S3)	BAM, FASTQ	SOAP, SNP	2013	-	✓
27	MEGAN	Windows, Unix, Linux, MacOS	-	IaaS	-	Text tabular, XML), RapSearch2, SAM, RDP, NBC, QIIME, CSV.	-	2013	GUI	✓
28	Stormbow	Unix, Linux	Illumina HiSeq 2000	SaaS	Public (Amazon EC2 and S3)	FASTQ, FASTA	BAM	2013	CLI	✓
29	BioPig	Unix, Linux	-	PaaS	Public (Amazon EC2)	FASTQ, FASTA	FASTQ, FASTA	2013	CLI	
30	SparkSeq	Linux, MacOS	-	-	Public (Microsoft Azure)	BED, GTF	-	2014	CLI	
31	BioVLAB-MMIA-NGS	Windows, Unix, Linux	-	PaaS, SaaS	Public (Amazon EC2)	FASTQ	-	2014	CLI, WI	✓
32	Contrail	Unix, Linux	-	IaaS, PaaS	-	-	-	2014	CLI	✓
33	Mercury	Unix, Linux	Illumina HiSeq	-	Public (Amazon EC2, S3)	FASTQ	-	2014	CLI, WI	✓

34	STORMSeq	Unix, Linux	-	SaaS	Public (Amazon EC2, S3)	FASTQ, BAM	VCF	2014	CLI , GUI	✓
35	SURPI	Unix, Linux	Illumina	-	Public (Amazon EC2)	FASTQ	-	2014	CLI	✓
36	Seqpig	Unix, Linux	Illumina	PaaS	Public (Amazon S3, Elastic MapReduce)	BAM, SAM, FASTA, FASTQ, QSEQ	BAM, SAM, FastQ, Qseq	2014	CLI	✓
37	SNP2-Structure	Unix, Linux	-	SaaS	Public (Amazon EC2)	-	PDB	2015	WI	✓
38	Halvade	Unix, Linux	Illumina HiSeqNA12878	-	Public (Amazon EC2, S3)	FASTQ	VCF	2015	CLI	✓
39	CLUSTOM-CLOUD	Windows, Unix, Linux, MacOS	Roche/454 FLX Titanium	PaaS	Public (Amazon EC2)	FASTA, XML	FASTA	2016	CLI	
40	MG-RAST	Unix, Linux	454 reads Sanger sequences	IaaS	Public (Shock, AWE server), Amazon EC2	FASTA, FASTQ, SFF	FASTA, GFF3, GenBank	2016	WI	✓
41	MC-Genome Key	Unix, Linux	Illumina	-	Public (Amazon, Google, Azure), Private (OpenStack)	FASTQ, BAM	VCF	2017	CLI , WI	✓

Table 3: Open-source cloud-based NGS tools developed in the years 2013-2017.

Table 4 depicts the features, described in a prior section, of the selected commercial NGS tools. As shown in the table, all of the desktop versions of the commercial tools work with the Linux/Unix operating system except one tool that works on Windows and MacOS operating systems. Regarding the NGS technology used, the commercial NGS tools support different NGS technologies (Platforms), such as the Roche 454 GS FLX sequencer and Illumina.

In addition, Table 4 shows that three of the commercial tools accept the FASTA and FASTQ file format as input files. However, they produce files with different formats as an output. The table also shows that all of the tools have been released between years 2009 and 2012.

Regarding the interface of commercial NGS tools, all tools interact with the user via the Command Line Interface (CLI), except one tool which uses a Graphical User Interface (GUI). Finally, as the table shows, all of the tools use a pipeline to analyze the NGS data.

ID	NGS Tool	Operating System	NGS tech-nology	Input File	Output File	Release year	Interface	Pipeline
1	BaseSpace	Unix, Linux	Illumina	BCL	FASTQ	2011	CLI	✓
2	Bina	Windows, Unix, Linux, Mac OS	-	FASTQ	csv	2012	CLI	✓
3	DNAnexus	Unix, Linux	-	Various *	Various *	2009	CLI	✓
4	LifeScope	Unix, Linux	5500 Genetic Analyz-ers	XSQ	BAM, GFF3	2012	CLI, GUI	✓
5	GeneSifter	Unix, Linux	Roche 454 GS FLX Se-quencer	BAM, VCF, FASTA	xml	2010	GUI	✓
6	SevenBridges	Unix, Linux	-	FASTA	-	2009	CLI, WI, API	✓

Table 4: Commercial cloud-based NGS Tools.

Table 5 shows the 10 features, described in a prior section, of the 13 selected open-source NGS alignment tools. As shown in the table, all of the desktop versions of the open-source alignment tools work with the Linux/Unix operating system. However, four of them work with Windows, eight of them work with MacOS, and only one of them, which is the Bowtie NGS alignment tool [97], works with the Solaris operating system. In terms of the NGS technology used, all of the NGS alignment tools are compatible with the Illumina technology. Regarding the utilized Cloud Service Model, most of the tools are deployed as Infrastructure as a Service except one tool, the SEAL [106], which was deployed as a Software as a Service.

Table 5 also shows that all of the open-source alignment tools accept FASTA or FASTQ file format as input files. However, they generate files with different formats as an output. As shown in the table, some of the alignment tools have been available for some time. For example, the first version of the Blast NGS alignment tool was released in 1990. Similarly, the MUMmer NGS alignment tool was released in 1999.

Regarding the interface of the open-source alignment tools, all of them interact with the user via a Command Line Interface (CLI) except one tool which uses the Web User Interface (WUI). Additionally, Table 5 shows that all of the tools do not use a pipeline to analyze the NGS data. Moreover, the table shows that all of the open-source alignment tools are developed using C and C++ programming languages. However, they used different algorithms to align the NGS data.

ID	NGS Tool	OS	NGS technology	CSM	Input File	Output File	Release Year	Interface	PL	Algo.
1	BWA	Unix, Linux	Illumina, Oxford Nanopore, SOLiD, 454, Sanger reads, PacBio sequencer	IaaS	FASTA, FASTQ	SAM	2010	CLI	C, Java script	BWA-backtrack, BWA-SW, BWAMEM
2	SAMtools	Unix, Linux	Illumina GA, AB SOLiD	IaaS	SAM, BAM, CRAM, FASTA	SAM, BAM, VCF, CRAM	2009	CLI	C, Perl	-
3	MAQ	Windows, Unix, Linux, Mac OS	Illumina-Solexa 1G Genetic Analyzer	IaaS	FASTA	FASTQ, SNP, LOG	2008	CLI	C, C++, Perl	-
4	BLAT	Unix, Linux	-	IaaS	FASTA	HTML, PSL	2002	WUI, CLI	-	Graph and Dynamic algorithm
5	Blast	Windows, Linux, MacOS	-	IaaS	FASTA	XML, ASN.1	1990	WUI	C, C++	-
6	MUMmer GPU 2.0	Unix, Linux	Illumina, 454 Life Sciences, Applied Biosystems	IaaS	-	FIG, PDF	2009	CLI	C++	-
7	MUMmer	Unix,Linux, MacOS	-	IaaS	FASTA	FIG, PDF	1999, 2004, 2003, 2002	CLI	C, C++, Java, Perl, Python, Ruby	Suffix-Tree
8	SHRiMP	Unix,Linux, MacOS	Illumina-Solexa, Roche/454, AB SOLiD	IaaS	FASTA, FASTQ	SAM	2009, 2011	CLI	C C++	Smith-Waterman
9	Bowtie	Windows, MacOS, Linux, Solaris.	Illumina	IaaS	FASTA, FASTQ	SAM, FAI	2009	CLI	C, C++	Blockwise
10	Bowtie2	Unix, Linux, MacOS	Illumina, HiSeq 2000, Roche/454	-	FASTA, FASTQ	SAM ,SOAP	2012	CLI	C C++	Dynamic-Programming
11	SEAL	Unix, Linux	Illumina	SaaS	FASTQ, PRQ, QSeq	SAM	2011	CLI	-	-
12	TopHat	Unix, Linux, MacOS	Illumina SOLiD	-	FASTA, FASTQ	SAM, BAM	2009	CLI	C++, Python	TopHat-Fusion, Indel-finding
13	HISAT2	Windows, Unix, Linux, MacOS	-	-	-	SAM	2015	CLI	C, C++, Perl, Python, Bash	Two-pass

Table 5: Cloud-based NGS Alignment Tools.

5. Discussion

This section discusses the main findings of our results regarding cloud-based NGS analysis and alignment tools.

Finding 1: NGS tools are not cross-platform tools. The majority of them work on the Linux/Unix operating system.

Finding 2: Most of the NGS tools interact with users via Command Line Interface (CLI)

For our first finding 1, as shown in Section 4, very few of the NGS tools are cross-platform tools, i.e., tools that work on different operating systems. The majority of the desktop versions of the NGS tools run on the Linux/Unix operating system. Such a restriction limits the number of users who can use these tools, since it requires special skills that most biomedical users lack. Therefore, we highly recommend developers of NGS tools to develop cross-platform tools to increase the usage of their tools.

Our second finding 2 is similar to the previous finding, i.e., most of the NGS tools interact with users via Command Line Interface (CLI). In general, users face many difficulties interacting with CLI interfaces including (1) they need to memorize various commands and they need to know how to use them; (2) in some cases, CLI interfaces require users to write scripts to execute various tasks; (3) comparing to the Web interfaces, CLI interfaces require more effort to execute the same task, especially from biomedical users.

Finding 3: Very few NGS tools have been developed during the last 5 years.

Figure 2 depicts the number of open-source NGS tools per year. The figure shows an important finding

Finding 4: There is no standard input/output file format for NGS tools. Hence, reducing the compatibility, portability, interoperability, and integration between the tools.

of our research, that is, most of the open-source NGS tools were developed between the years 2009 and 2012 and there are no recent tools developed in the past five years, Finding 3. This finding shows an industrial gap in developing NGS tools using recent technologies. Therefore, we recommend NGS tool developers to either develop new NGS tools, or upgrade existing NGS tools by adapting emerging technologies.

Although most of the tools accept FASTA and FASTAQ file format as input to their tools, other tools accept other formats such as XML and SSF formats. In addition, they generate files with different formats as output files. Therefore, there is no standard input format for NGS tools which reduces the compatibility, portability, interoperability, and integration between the NGS tools, Finding 4. As a recommendation, NGS tools need to standardize the format of their input and output files.

6. Conclusion

In this study, we investigated the most used cloud-based Next-Generation Sequencing (NGS) data analysis and alignment tools. We studied 60 tools divided into three categories: 41 open-source NGS tools, 6 commercial NGS tools, and 13 NGS alignment tools. For these cloud-based tools, we extracted and studied crucial features that biomedical researchers and clinicians consider for selecting the appropriate NGS tools according to the needs of their works. We present many findings that provide insights and recommendations for developers of NGS tools to improve them. In the future, we are planning to conduct an empirical study to measure various Quality of Service (QoS) attributes of these tools such as their performance, efficiency, security, power consumption, and reliability.

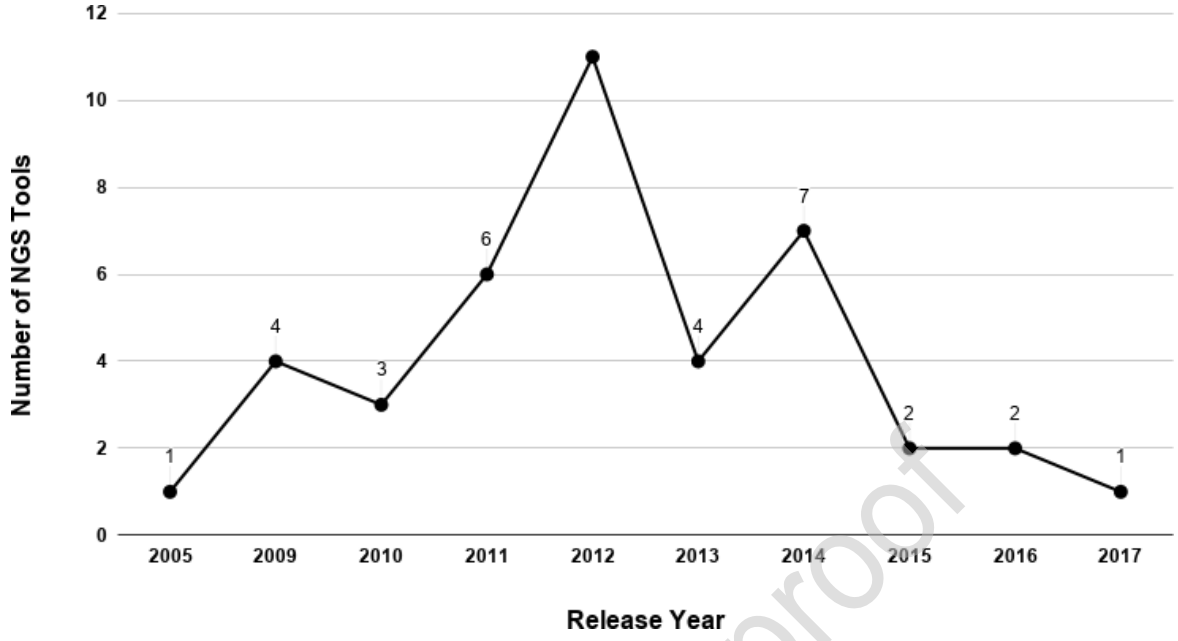

Figure 2: Number of open-source NGS tools per year.

7. Acknowledgements

We gratefully acknowledge Jordan University of Science and Technology for supporting this work under award number 20170030.

References

[1] K. S. Aggour, V. S. Kumar, D. P. Sangurdekar, L. A. Newberg, C. D. Kodira, A highly parallel next-generation dna sequencing data analysis pipeline in hadoop, in: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE, 2015, pp. 756–763.

[2] L. Shi, Z. Wang, W. Yu, X. Meng, A case study of tuning mapreduce for efficient bioinformatics in the cloud, Parallel Computing 61 (2017) 83–95.

[3] E. R. Mardis, Dna sequencing technologies: 2006–2016, Nature protocols 12 (2) (2017) 213–218.

[4] X. Guo, N. Yu, B. Li, Y. Pan, Cloud computing for next-generation sequencing data analysis, Computational Methods for Next Generation Sequencing Data Analysis (2016) 1–24.

[5] T. Attia, M. Saeed, Next generation sequencing technologies: A short review, Next Generat Sequenc & Applic S 1 (2016) 2.

[6] D. J. Munroe, T. J. Harris, Third-generation sequencing fireworks at marco island, Nature biotechnology 28 (5) (2010) 426–428.

[7] C. S. Pareek, R. Smoczynski, A. Tretyn, Sequencing technologies and genome sequencing, Journal of applied genetics 52 (4) (2011) 413–435.

[8] I. G. Gut, New sequencing technologies, Clinical and Translational Oncology 15 (11) (2013) 879–881.

[9] E. E. Schadt, S. Turner, A. Kasarskis, A window into third-generation sequencing, Human molecular genetics 19 (R2) (2010) R227–R240.

[10] T. P. Niedringhaus, D. Milanova, M. B. Kerby, M. P. Snyder, A. E. Barron, Landscape of next-generation sequencing technologies, Analytical chemistry 83 (12) (2011) 4327–4341.

[11] J. M. Heather, B. Chain, The sequence of sequencers: The history of sequencing dna, Genomics 107 (1) (2016) 1–8.

[12] A. S. Gargis, L. Kalman, D. P. Bick, C. Da Silva, D. P. Dimmock, B. H. Funke, S. Gowrisankar, M. R. Hegde, S. Kulkarni, C. E. Mason, et al., Good laboratory practice for clinical next-generation sequencing informatics pipelines, Nature biotechnology 33 (7) (2015) 689.
[13] S. Goodwin, J. D. McPherson, W. R. McCombie, Coming of age: ten years of next-generation sequencing technologies, Nature Reviews Genetics 17 (6) (2016) 333.
[14] J. S. Black, M. Salto-Tellez, K. I. Mills, M. A. Catherwood, The impact of next generation sequencing technologies on haematological research–a review, Pathogenesis 2 (1-2) (2015) 9–16.
[15] K. J. van Nimwegen, R. A. van Soest, J. A. Veltman, M. R. Nelen, G. J. van der Wilt, L. E. Vissers, J. P. Grutters, Is the $1000 genome as near as we think? a cost analysis of next-generation sequencing, Clinical chemistry 62 (11) (2016) 1458–1464.
[16] A. Celesti, M. Fazio, F. Celesti, G. Sannino, S. Campo, M. Villari, New trends in biotechnology: The point on ngs cloud computing solutions, in: 2016 IEEE Symposium on Computers and Communication (ISCC), IEEE, 2016, pp. 267–270.
[17] T. Kwon, W. G. Yoo, W.-J. Lee, W. Kim, D.-W. Kim, Next-generation sequencing data analysis on cloud computing, Genes & Genomics 37 (6) (2015) 489–501.
[18] R. S. Thakur, R. Bandopadhyay, B. Chaudhary, S. Chatterjee, Now and next-generation sequencing techniques: future of sequence analysis using cloud computing, Frontiers in genetics 3 (2012) 280.
[19] G. Onsongo, J. Erdmann, M. D. Spears, J. Chilton, K. B. Beckman, A. Hauge, S. Yohe, M. Schomaker, M. Bower, K. A. Silverstein, et al., Implementation of cloud based next generation sequencing data analysis in a clinical laboratory, BMC research notes 7 (1) (2014) 314.
[20] H. Elazhary, Cloud computing for big data, MAGNT Res Rep 2 (4) (2014) 135–144.
[21] Y. Jadeja, K. Modi, Cloud computing-concepts, architecture and challenges, in: 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), IEEE, 2012, pp. 877–880.
[22] L. Dai, X. Gao, Y. Guo, J. Xiao, Z. Zhang, Bioinformatics clouds for big data manipulation, Biology direct 7 (1) (2012) 43.
[23] S. Zhao, K. Watrous, C. Zhang, B. Zhang, Cloud computing for next-generation sequencing data analysis, Cloud Computing-Architecture and Applications, InTech, Rijeka (2017) 29–51.
[24] B. Calabrese, M. Cannataro, Bioinformatics and microarray data analysis on the cloud, in: Microarray Data Analysis, Springer, 2015, pp. 25–39.
[25] L. Qian, Z. Luo, Y. Du, L. Guo, Cloud computing: An overview, Springer, 2009, pp. 626–631.
[26] Q. B. Baker, W. Al-Rashdan, Y. Jararweh, Cloud-based tools for next-generation sequencing data analysis, in: 2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), IEEE, 2018, pp. 99–105.
[27] P. L. Auer, R. Doerge, Statistical design and analysis of rna sequencing data, Genetics 185 (2) (2010) 405–416.
[28] A. Barski, K. Zhao, Genomic location analysis by chip-seq, Journal of cellular biochemistry 107 (1) (2009) 11–18.
[29] D. A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G. T. Roth, et al., The complete genome of an individual by massively parallel dna sequencing, nature 452 (7189) (2008) 872.
[30] J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, M. L. Blaxter, Genome-wide genetic marker discovery and genotyping using next-generation sequencing, Nature Reviews Genetics 12 (7) (2011) 499.
[31] O. Harismendy, P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, N. J. Schork, S. S. Murray, E. J. Topol, S. Levy, et al., Evaluation of next generation sequencing platforms for population targeted sequencing studies, Genome biology 10 (3) (2009) R32.
[32] N. J. Loman, R. V. Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, M. J. Pallen, Performance comparison of benchtop high-throughput sequencing platforms, Nature biotechnology 30 (5) (2012) 434.
[33] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, Y. Gu, A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers, BMC genomics 13 (1) (2012) 341.
[34] E. L. Van Dijk, H. Auger, Y. Jaszczyszyn, C. Thermes, Ten years of next-generation sequencing technology, Trends in genetics 30 (9) (2014) 418–426.
[35] M. Escalona, S. Rocha, D. Posada, A comparison of tools for the simulation of genomic next-generation sequencing data, Nature Reviews Genetics 17 (8) (2016) 459.
[36] S. Pattnaik, S. Gupta, A. A. Rao, B. Panda, Sinc: an accurate and fast error-model based simulator for snps, indels and cnvs coupled with a read generator for short-read sequence data, BMC bioinformatics 15 (1) (2014) 40.
[37] M. Kircher, S. Sawyer, M. Meyer, Double indexing overcomes inaccuracies in multiplex sequencing on the illumina platform, Nucleic acids research 40 (1) (2011) e3–e3.
[38] M. Alberich, A. Artetxe, E. Santamar´ıa-Navarro, A. Nonell-Canals, G. Maclair, Genesis—cloud-based system for next generation sequencing analysis: A proof of concept, in: Innovation in Medicine and Healthcare 2016, Springer, 2016, pp. 291–300.
[39] J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, et al., The bioperl toolkit: Perl modules for the life sciences, Genome research 12 (10) (2002) 1611–1618.

[40] P. Kudtarkar, T. F. DeLuca, V. A. Fusaro, P. J. Tonellato, D. P. Wall, Cost-effective cloud computing: a case study using the comparative genomics tool, roundup, Evolutionary Bioinformatics 6 (2010) EBO–S6259.
[41] C.-L. Hung, Y.-L. Lin, G.-J. Hua, Y.-C. Hu, Cloudtss: a tagsnp selection approach on cloud computing, in: International Conference on Grid and Distributed Computing, Springer, 2011, pp. 525–534.
[42] Ngs survey, http://tiny.cc/ngs19, nov. 2019.
[43] E. Afgan, D. Baker, B. Batut, M. Van Den Beek, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, B. A. Grüning, et al., The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic acids research 46 (W1) (2018) W537–W544.
[44] D. Blankenberg, J. Hillman-Jackson, Analysis of next-generation sequencing data using galaxy, in: Stem Cell Transcriptional Networks, Springer, 2014, pp. 21–43.
[45] M. C. Schatz, Cloudburst: highly sensitive read mapping with mapreduce, Vol. 25, Oxford University Press, 2009, pp. 1363–1369.
[46] B. Langmead, M. C. Schatz, J. Lin, M. Pop, S. L. Salzberg, Searching for snps with cloud computing, Vol. 10, BioMed Central, 2009, p. R134.
[47] Basespace, https://developer.basespace.illumina.com/docs/content/documentation/getting-started/overview, accessed: 19 Aug. 2019.
[48] H. Li, R. Durbin, Fast and accurate long-read alignment with burrows–wheeler transform, Vol. 26, Oxford University Press, 2010, pp. 589–595.
[49] Y. Li, S. Zhong, Seqmapreduce: software and web service for accelerating sequence mapping, Critical Assessment of Massive Data Anaysis (CAMDA) 2009.
[50] A. C. Stewart, B. Osborne, T. D. Read, Diya: a bacterial annotation pipeline for any genomics lab, Bioinformatics 25 (7) (2009) 962–963.
[51] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al., The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data, Vol. 20, Cold Spring Harbor Lab, 2010, pp. 1297–1303.
[52] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas, M. Hanna, et al., A framework for variation discovery and genotyping using next-generation dna sequencing data, Vol. 43, Nature Research, 2011, pp. 491–498.
[53] Bina careers, http://www.bina.com/careers, accessed: 19 Aug. 2019.
[54] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and samtools, Vol. 25, Oxford University Press, 2009, pp. 2078–2079.
[55] B. Langmead, K. D. Hansen, J. T. Leek, Cloud-scale rna-sequencing differential expression analysis with myrna, Vol. 11, BioMed Central, 2010, p. R83.
[56] J. Orvis, J. Crabtree, K. Galens, A. Gussman, J. M. Inman, E. Lee, S. Nampally, D. Riley, J. P. Sundaram, V. Felix, et al., Ergatis: a web interface and scalable software system for bioinformatics workflows, Vol. 26, Oxford University Press, 2010, pp. 1488–1492.
[57] S. V. Angiuoli, M. Matalka, A. Gussman, K. Galens, M. Vangala, D. R. Riley, C. Arze, J. R. White, O. White, W. F. Fricke, Clovr: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing, Vol. 12, BioMed Central, 2011, p. 356.
[58] J. R. White, C. Arze, M. Matalka, et al., Clovr-16s: Phylogenetic microbial community composition analysis based on 16s ribosomal rna amplicon sequencing–standard operating procedure, version 1.0, Nature Publishing Group, 2011.
[59] K. Galens, J. White, C. Arze, M. Matalka, M. G. Giglio, T. C. Team, S. Angiuoli, W. F. Fricke, Clovr-microbe: Assembly, gene finding and functional annotation of raw sequence data from single microbial genome projects–standard operating procedure, version 1.0, Nature Publishing Group, 2011, pp. 1–1.
[60] Dnanexus, https://www.dnanexus.com/contact, accessed: 19 Aug. 2019.
[61] Manual reference pages - maq (1), http://maq.sourceforge.net/maq-manpage.shtml, accessed: 19 Aug. 2019.
[62] H. Li, J. Ruan, R. Durbin, Mapping short dna sequencing reads and calling variants using mapping quality scores, Vol. 18, Cold Spring Harbor Lab, 2008, pp. 1851–1858.
[63] T. Nguyen, W. Shi, D. Ruden, Cloudaligner: A fast and full-featured mapreduce based tool for sequence mapping, Vol. 4, BioMed Central, 2011, p. 171.
[64] Y. Zhao, H. Tang, Y. Ye, Rapsearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data, Vol. 28, Oxford University Press, 2011, pp. 125–126.
[65] Jnomics, https://sourceforge.net/projects/jnomics/, accessed: 19 Aug. 2019.
[66] Jnomics—a cloud-scale sequence analysis suite, http://schatzlab.cshl.edu/publications/posters/2011.GenomeInformatics.Jnomics.pdf, accessed: 19 Aug. 2019.
[67] Thermofisher scientific, https://www.thermofisher.com/jo/en/home.html, accessed: 19 Aug. 2019.
[68] M. Bhagwat, L. Young, R. R. Robison, Using blat to find sequence similarity in closely related genomes, Wiley Online Library, 2012, pp. 10–8.
[69] W. J. Kent, Blat—the blast-like alignment tool, Vol. 12, Cold Spring Harbor Lab, 2002, pp. 656–664.

[70] X. Feng, R. Grossman, L. Stein, Peakranger: a cloud-enabled peak caller for chip-seq data, Vol. 12, BioMed Central, 2011, p. 139.
[71] A. Goncalves, A. Tikhonov, A. Brazma, M. Kapushesky, A pipeline for rna-seq data processing and quality assessment, Vol. 27, Oxford University Press, 2011, pp. 867–869.
[72] M. Fischer, R. Snajder, S. Pabinger, A. Dander, A. Schossig, J. Zschocke, Z. Trajanoski, G. Stocker, Simplex: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data, Vol. 7, Public Library of Science, 2012, p. e41948.
[73] Perkinelmer informatics support news, http://www.genesifter.net/, accessed: 19 Aug. 2019.
[74] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, T. L. Madden, Ncbi blast: a better web interface, Vol. 36, Oxford University Press, 2008, pp. W5–W9.
[75] T. Madden, The blast sequence analysis tool, in: The NCBI Handbook [Internet]. 2nd edition, National Center for Biotechnology Information (US), 2013.
[76] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L. Madden, Blast+: architecture and applications, Vol. 10, BioMed Central, 2009, p. 421.
[77] S. Zhao, K. Prenger, L. Smith, T. Messina, H. Fan, E. Jaeger, S. Stephens, Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing, Vol. 14, BioMed Central, 2013, p. 425.
[78] D. H. Huson, N. Weber, Microbial community analysis using megan., Vol. 531, 2013, pp. 465–485.
[79] S. Zhao, K. Prenger, L. Smith, Stormbow: a cloud-based tool for reads mapping and expression quantification in large-scale rna-seq studies, Vol. 2013, Hindawi Publishing Corporation, 2013, pp. 1–8.
[80] Sevenbridges, https://www.sevenbridges.com/, accessed: 19 Aug. 2019.
[81] C. Trapnell, M. C. Schatz, Optimizing data intensive gpgpu computations for dna sequence alignment, Vol. 35, Elsevier, 2009, pp. 429–440.
[82] H. Nordberg, K. Bhatia, K. Wang, Z. Wang, Biopig: a hadoop-based analytic toolkit for large-scale sequence data, Vol. 29, Oxford University Press, 2013, pp. 3014–3019.
[83] L. Jourdren, M. Bernard, M.-A. Dillies, S. Le Crom, Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses, Vol. 28, Oxford University Press, 2012, pp. 1542–1543.
[84] U. S. Evani, D. Challis, J. Yu, A. R. Jackson, S. Paithankar, M. N. Bainbridge, A. Jakkamsetti, P. Pham, C. Coarfa, A. Milosavljevic, et al., Atlas2 cloud: a framework for personal genome analysis in the cloud, Vol. 13, BioMed Central, 2012, p. S19.
[85] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, S. L. Salzberg, Alignment of whole genomes, Vol. 27, Oxford University Press, 1999, pp. 2369–2376.
[86] A. L. Delcher, A. Phillippy, J. Carlton, S. L. Salzberg, Fast algorithms for large-scale genome alignment and comparison, Vol. 30, Oxford University Press, 2002, pp. 2478–2483.
[87] A. L. Delcher, S. L. Salzberg, A. M. Phillippy, Using mummer to identify similar regions in large sequence sets, Wiley Online Library, 2003, pp. 10–3.
[88] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes, Vol. 5, BioMed Central, 2004, p. R12.
[89] Y. W. Asmann, S. Middha, A. Hossain, S. Baheti, Y. Li, H.-S. Chai, Z. Sun, P. H. Duffy, A. A. Hadad, A. Nair, et al., Treat: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data, Vol. 28, Oxford University Press, 2011, pp. 277–278.
[90] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, K. E. Nelson, Cloud biolinux: pre-configured and on-demand bioinformatics computing for the genomics community, Vol. 13, BioMed Central, 2012, p. 42.
[91] H. Y. Lam, C. Pan, M. J. Clark, P. Lacroute, R. Chen, R. Haraksingh, M. O'huallachain, M. B. Gerstein, J. M. Kidd, C. D. Bustamante, et al., Detecting and annotating genetic variations using the hugeseq pipeline, Vol. 30, Nature Research, 2012, pp. 226–229.
[92] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, M. Brudno, Shrimp: accurate mapping of short color-space reads, Vol. 5, Public Library of Science, 2009, p. e1000386.
[93] M. David, M. Dzamba, D. Lister, L. Ilie, M. Brudno, Shrimp2: sensitive yet practical short read mapping, Vol. 27, Oxford University Press, 2011, pp. 1011–1012.
[94] L. Habegger, S. Balasubramanian, D. Z. Chen, E. Khurana, A. Sboner, A. Harmanci, J. Rozowsky, D. Clarke, M. Snyder, M. Gerstein, Vat: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment, Vol. 28, Oxford University Press, 2012, pp. 2267–2269.
[95] D. Hong, A. Rhie, S.-S. Park, J. Lee, Y. S. Ju, S. Kim, S.-B. Yu, T. Bleazard, H.-S. Park, H. Rhee, et al., Fx: an rna-seq analysis tool on the cloud, Vol. 28, Oxford University Press, 2012, pp. 721–723.
[96] L. Zhang, S. Gu, Y. Liu, B. Wang, F. Azuaje, Gene set analysis in the cloud, Vol. 28, Oxford University Press, 2011, pp. 294–295.
[97] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short dna sequences to the human genome, Vol. 10, BioMed Central, 2009, p. R25.
[98] E. Afgan, B. Chapman, J. Taylor, Cloudman as a platform for tool, data, and analysis distribution, BMC

bioinformatics 13 (1) (2012) 315.

[99] E. Afgan, D. Baker, N. Coraor, H. Goto, I. M. Paul, K. D. Makova, A. Nekrutenko, J. Taylor, Harnessing cloud computing with galaxy cloud, Nature biotechnology 29 (11) (2011) 972.

[100] M. Niemenmaa, A. Kallio, A. Schumacher, P. Klemelä, E. Korpelainen, K. Heljanko, Hadoop-bam: directly manipulating next generation sequencing data in the cloud, Vol. 28, Oxford University Press, 2012, pp. 876–877.

[101] M. S. Wiewiórka, A. Messina, A. Pacholewska, S. Maffioletti, P. Gawrysiak, M. J. Okoniewski, Sparkseq: fast, scalable and cloud-ready tool for the interactive genomic data analysis with nucleotide precision, Vol. 30, Oxford University Press, 2014, pp. 2652–2653.

[102] B. Langmead, S. L. Salzberg, Fast gapped-read alignment with bowtie 2, Vol. 9, Nature Publishing Group, 2012, p. 357.

[103] H. Chae, S. Rhee, K. P. Nephew, S. Kim, Biovlab-mmia-ngs: microrna–mrna integrated analysis using high-throughput sequencing data, Vol. 31, Oxford University Press, 2014, pp. 265–267.

[104] Contrail, https://omictools.com/contrail-tool, accessed: 19 Aug. 2019.

[105] J. G. Reid, A. Carroll, N. Veeraraghavan, M. Dahdouli, A. Sundquist, A. English, M. Bainbridge, S. White, W. Salerno, C. Buhay, et al., Launching genomics into the cloud: deployment of mercury, a next generation sequence analysis pipeline, Vol. 15, BioMed Central, 2014, p. 30.

[106] L. Pireddu, S. Leo, G. Zanetti, Seal: a distributed short read mapping and duplicate removal tool, Vol. 27, Oxford University Press, 2011, pp. 2159–2160.

[107] K. J. Karczewski, G. H. Fernald, A. R. Martin, M. Snyder, N. P. Tatonetti, J. T. Dudley, Stormseq: an open-source, user-friendly pipeline for processing personal genomics data in the cloud, Vol. 9, Public Library of Science, 2014, p. e84860.

[108] S. N. Naccache, S. Federman, N. Veeraraghavan, M. Zaharia, D. Lee, E. Samayoa, J. Bouquet, A. L. Greninger, K.-C. Luk, B. Enge, et al., A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples, Vol. 24, Cold Spring Harbor Lab, 2014, pp. 1180–1192.

[109] A. Schumacher, L. Pireddu, M. Niemenmaa, A. Kallio, E. Korpelainen, G. Zanetti, K. Heljanko, Seqpig: simple and scalable scripting for large sequencing data sets in hadoop, Vol. 30, Oxford University Press, 2013, pp. 119–120.

[110] C. Trapnell, L. Pachter, S. L. Salzberg, Tophat: discovering splice junctions with rna-seq, Vol. 25, Oxford University Press, 2009, pp. 1105–1111.

[111] D. Wang, L. Song, V. Singh, S. Rao, L. An, S. Madhavan, Snp2structure: a public and versatile resource for mapping and three-dimensional modeling of missense snps on human protein structures, Vol. 13, Elsevier, 2015, pp. 514–519.

[112] D. Decap, J. Reumers, C. Herzeel, P. Costanza, J. Fostier, Halvade: scalable sequence analysis with mapreduce, Vol. 31, Oxford University Press, 2015, pp. 2482–2488.

[113] J. Oh, C.-H. Choi, M.-K. Park, B. K. Kim, K. Hwang, S.-H. Lee, S. G. Hong, A. Nasir, W.-S. Cho, K. M. Kim, Clustom-cloud: In-memory data grid-based software for clustering 16s rrna sequence data in the cloud environment, Vol. 11, Public Library of Science, 2016, p. e0151064.

[114] D. Kim, B. Langmead, S. L. Salzberg, Hisat: a fast spliced aligner with low memory requirements, Vol. 12, Nature Publishing Group, 2015, p. 357.

[115] W. Tang, J. Bischof, N. Desai, K. Mahadik, W. Gerlach, T. Harrison, A. Wilke, F. Meyer, Workload characterization for mg-rast metagenomic data analytics service in the cloud, in: Big Data (Big Data), 2014 IEEE International Conference on, IEEE, 2014, pp. 56–63.

[116] A. Wilke, W. Gerlach, T. Harrison, T. Paczian, W. L. Trimble, F. Meyer, Mg-rast manual for version 4, revision 3, 2017.

[117] H. Elshazly, Y. Souilmi, P. J. Tonellato, D. P. Wall, M. Abouelhoda, Mc-genomekey: a multicloud system for the detection and annotation of genomic variants, Vol. 18, BioMed Central, 2017, p. 49.

[118] H. Perrin, M. Denorme, J. Grosjean, E. Dynomant, V. J. Henry, F. Pichon, S. Darmoni, A. Desfeux, B. J. Gonzalez, et al., Omictools: a community-driven search engine for biological data analysis, 2017.

- **Conflict of Interest**

  No conflict of interest exists. We wish to confirm that there are no known conflicts of interest associated with this publication.

- **Data Availability Statement**

  All relevant data are within the manuscript and its Supporting Information files.

- **Financial Disclosure**

We thank Jordan University of Science and Technology for support of this publication.

Ethical Statement

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript

Author's name	Affiliation
Qanita Bani Baker	Jordan University of Science and Technology
Mahmoud M. Hammad	Jordan University of Science and Technology
Wesam Al-Rashdan	Jordan University of Science and Technology
Yaser Jararweh	Duquesne University\ Jordan University of Science and Technology
Mohammad AL-Smadi	Jordan University of Science and Technology
Mohammad Al-Zinati	Jordan University of Science and Technology