Letters in Applied NanoBioScience
Open-Access Journal (ISSN: 2284-6808)

Review
Volume 12, Issue 4, 2023, 130
https://doi.org/10.33263/LIANBS124.130

A Review on Bioinformatics Tools for Transcriptomics NGS Data Analysis

Mohini [1,*], Ashwani Dixit [2], Ravneet Kaur [3], Arpana Jayan [4], Satyabrat Dutta [5], Shafla Tharakkal [6]

1 Department of Botany,Hansraj College, University of Delhi; kajlamohini.mk5@gmail.com (M.);
2 Acharya Narendra Dev College, University of Delhi; ashwanidixit36@gmail.com (A.D.);
3 Department Of Botany, Panjab University; ravneetkdhaliwal6@gmail.com (R.K.);
4 Department of Botany, St. Albert's College, Autonomous Ernakulam; arpanajayan@gmail.com (A.J.);
5 Department of Veterinary Microbiology, Indian Veterinary Research Institute, Bareilly; satyabrat498@gmail.com (S.D.);
6 MSc Microbiology, Bharathidasan University; shafla.shafus@gmail.com (S.T.);
* Correspondence: kajlamohini.mk5@gmail.com (M);

Received: 28.06.2022; Accepted: 7.09.2022; Published: 18.09.2022

Abstract: With the advancement in technologies, there has been a great expansion of data generation in various fields and disciples, including genomics, pharmacogenomics, epigenomics, transcriptomics, metabolomics, and proteomics. Regarding sequencing technologies, the advent of next-generation sequencing methods has enabled researchers to analyze the entire genome or multiple genes simultaneously for mutation detection or gene expression studies. For this reason, the demand for appropriate bioinformatics tools and pipelines for NGS Data Analysis to conduct precision analysis with a high level of accuracy is very high. The field of transcriptomics requires analysis of the entire RNA transcripts that define the transcriptional state of the cells. The analysis of the entire transcriptome is challenging and is supplemented with the utilization of various bioinformatic tools. This review provides an overview of the - generation sequencing and analysis in various fields like genomics and epigenomics, with a special focus on transcriptome analysis.

Keywords: genomics; pharmacogenomics; epigenomic; transcriptomics; metabolomics, and proteomics.

1. Introduction

Next-generation sequencing (NGS) is a technology that enables massively parallel sequencing of entire genomes or sometimes the targeted regions in the genome. NGS has revolutionized research with its capabilities of screening multiple genes, mutations, transcripts, etc., at the same time. Presently, genetics is extremely important in clinics for providing diagnoses for various heterogenous ailments. This requires the ability to scan the genome at various levels, from single base alterations to chromosomal levels. Various tools and pipelines are based on the different platforms used for sequencing. Illumina GA, Roche 454, and ABI SOLiD [1] are only a few new next-generation sequencing technology platforms that have recently been developed [2].

NGS techniques have been utilized successfully in a wide range of applications to date, including RNA-sequencing [3], ChIP-sequencing[4], whole human genome sequencing [5], and genome-wide structural variation[6]. A study reviewed the most significant advancements in sequencing technology and techniques and the evolution of NGS. They looked at the existing state of NGS applications and forecasted future trends [7]. Numerous

studies looked at and examined diverse NGS systems. Using sequencing Escherichia coli isolates, another study evaluated the ability of three NGS systems from Roche, Illumina, and Life Technologies. They examined read length, throughput, read error rate and profile, and the quality and completeness of de novo assembly to compare the platforms' performance [8]. A study used a collection of four microbial genomes to evaluate three NGS systems (Pacific Biosciences RS, Torrents PGM, and Illumina MiSeq). They evaluated the sequencing data concerning accuracy, variant detection, GC distribution, bias, and coverage distribution [9]. In another investigation, three NGS technologies were examined, ABI SOLiD, Illumina GA, and Roche 454, to analyze and assess identical 260 kb human sequences[10]. Employing different NGS systems, researchers in a study gained substantial expertise in handling samples, sequencing samples, and bioinformatics analysis [11]. Another evaluation looked at several NGS methodologies and how the latest developments in the field impact the course of genetic research [2]. As revealed by numerous studies [12,13], the Illumina platform remains one of the most widely employed platforms on the market.

In this review, we introspect the various bioinformatic methods and strategies employed for analyzing the data generated from NGS platforms.

2. Bioinformatic Tools for Analysis for Genomics and Epigenomics

Various bioinformatics tools and pipelines are available to analyze the data generated from genomics, epigenomics, and transcriptomics (Table 1).

Table 1. Tools and pipeline for NGS data analysis.

S.No.	Tool	Description & Function	References
1.	AzureBlast	A parallel BLAST executing on the Microsoft Azure cloud computing platform	[14]
2.	BlastReduce	BLAST-based on Hadoop	[15,16]
3.	CloudBLAST	Cloud-based BLAST implementation	[17]
4.	CloudBurst	MapReduce-based genomic sequence mapping using very sensitive short reads	[18]
5.	RSD	Comparative genomics (EC2 ortholog discovery using the reciprocal shortest distance approach)	[19]
6.	CloudAligner	Genomic sequence mapping (sequence mapping technology based on MapReduce that is full-featured and rapid)	[20]
7.	SEAL	Genomic sequence mapping (using Hadoop's duplication removal and short read pair mapping)	[21]
8.	Crossbow	Analysis of genomic sequences utilizing cloud computing, including read mapping and SNP calling	[22]
9.	Myrna	For RNA-Seq (Differential gene expression tool)	[23]
10.	Eoulsan	RNA sequencing analysis (a scalable, flexible framework built on the Hadoop platform)	[24]
11.	FX	For RNA-Seq analysis	[25]
12.	HadoopBAM	Management of sequence files (integration between analytic software and BAMfiles)	[26]
13.	SeqWare	Management of sequence files (query engine enabling information from databases)	[27,28]
14.	GATK	Next-generation resequencing data management for sequence files, a gene analysis toolset	[29,30]

2.1. Cloud computing.

The National Institute of Standards and Technology (NIST) established a centralized characterization or definition of cloud computing, which again received widespread support from several studies [31–33]. According to this definition, cloud computing is a coveted, self-servicing Internet platform that allows access to users for multiple computational

resources via the Internet at any time and from any location. It alters the way systems are seen in various domains and aids with challenges involving intense calculations [34–36].

2.1.1. Cloud types.

Studies have divided the cloud deployment models into three categories: hybrid, public and private [37]. Private cloud, On-Premise, and Off-Premise [33] are two private cloud types. Other developments, as revealed by many studies [32,38], created a new category called (community cloud). Below is a more in-depth explanation of four cloud categories:

2.1.1.1.Private cloud.

For the purpose of its functionality, management of data, and security, this form of cloud is operated by an individual user or an organization with several customers [32]. Many commercial and free software are capable of creating this cloud, of naming a few VMware Cloud, Open Nebula, Eucalyptus, OpenStack, and Terracotta. Private clouds include NASA's Nebula and Amazon's virtual private cloud (VPN).

2.1.1.2.Public cloud.

Third parties are generally in charge of these clouds. Web browsers are needed to access it. This cloud's security is rated lower than that of other [32] cloud categories. IBM's Blue Cloud, Sun Cloud, Microsoft Azure, Google Apps Engine, and Amazon's Elastic Compute Cloud (EC2) are all examples of public clouds.

2.1.1.3.Hybrid cloud.

This is a combination of several clouds (community, private or public) that provides a more secure approach to storing data. VMware vCloud and Salesforce are two examples of hybrid cloud Services [14].

2.1.1.4.Community cloud.

Several businesses collaborate to share and develop a cloud that a third party can manage. Some examples of community clouds include Microsoft Government Community Cloud and Google Apps for Government [14].

2.2. Pooling strategy for massive viral sequencing.

There can be a reduction in the cost of sequencing by employing the pooling strategy for multiple viral samples. This can be achieved by adding various barcodes that are sequence identifiers specific to each sample. The strategy includes mixing a designated pool of sample sets in a manner that the identity of each sample is encoded,the utilization of barcodes,and sample deconvolution, a viral variant assignment for individual samples from the pools [39].

2.3. Scaffolding algorithms.

De novo genome assembly is one of the most important next-generation sequencing analysis undertaking [40]. Though initial assemblers could only build tiny bacterial genomes, advances in quantity and quality of data, as well as state-of-the-art assembly algorithms in

addition to computer resources, have enabled the construction of more serpentine eukaryotic genomes assembly [41]. Numerous assembly algorithms were developed during the early years of next-generation sequencing, a few of which have kept up with advances in data production and algorithms' enhancements. In contrast, others have faded away [42].

The presently available assemblers give an output of DNA chucks, also known as contigs. These are used to tailor the genomes. The programs referred to as scaffolders use contigs to construct scaffolds and derive the connectivity information obtained from the NGS reads. SCARPA, MIP, SOPRA, OPERA, and SSPACE are some available scaffolding tools [43].

3. Computational Approaches in Next-Generation Sequencing Data

For the dry-lab analysis of NGS data, computational tools like SanGeniX, Galaxy, BWA, and Bowtie are employed.

3.1. Analysis for genome-wide DNA methylation studies.

The genome-wide DNA methylation analysis involves mapping sequence to methylation profile and further analyzing the differential methylation after quality assessment. MeDUSA (Methylated DNA Utility for Sequence Analysis) is one of the known pipelines used to analyze MeDIP-seq. The workflow includes mapping, filtering, quality assessment, coverage analysis, and plots to identify the differentially methylated regions in the genome using R Bioconductor [44].

3.2. Genomic variants detection and genotyping.

The widely analyzed form of genetic variation is single nucleotide polymorphism (SNP). However, there is a major challenge in analyzing these variants. One major factor includes the error rate towards the 3’ end of the reads. The variant calling is performed by aligning the reads to the analyzed genome and utilizing Bowtie2 and BWA tools [45].

4. Tools for Next-generation Transcriptomics

4.1. Computational methods.

The generation of data is the first step in a standard transcriptomic study. Conventionally the mRNAs or total RNAs are transformed into a cDNA library subjected to subsequent fragmentation and sequencing to yield single-end short reads from one end or paired-end short reads from both ends [46]. The small reads are subsequently pieced together to construct sequences and gene and transcript models during transcript assembly. Intersections across reads, which indicate that they might have been sampled from the same location, might reveal information regarding local assembly. The paired-end read’s fragment size may be statistically described, and this data can be used throughout the assembly process to offer long-range connectivity.

Genome sequencing has revealed that the general assembly issue is fundamentally tough. Ambiguous reads corresponding to substantially comparable sections of the genome, as a consequence of repetitions or duplications, cause noteworthy snags. Despite the fact that transcripts tend to be short and remain quick to be assembled, they do

present significant challenges. For example, the transcripts number to be reconstructed is unknown a priori. Then, because RNA-seq has been demonstrated to acquire expression levels range, encompassing 5 orders of magnitude, every gene can be examined by hundreds of thousands or just a few reads. Due to bias in library preparation and sequencing, read coverage remains inconsistent even inside the same gene or transcript limits [47].

Moreover, every read might be attributed to multiple genes in a family with sequence similarity or different isoforms of the similar gene. Nested genes, overlapping genes, and trans-splicing make assembly even more challenging. At the same time, aberrations caused by intronic reads or inaccurate mapping can result in erroneous introns and exons. Finally, interpreting the vast amount of data generated by an RNA-seq experiment necessitates using efficient and scalable tools. Figure 1 represents the basic bioinformatic workflow for NGS data analysis.

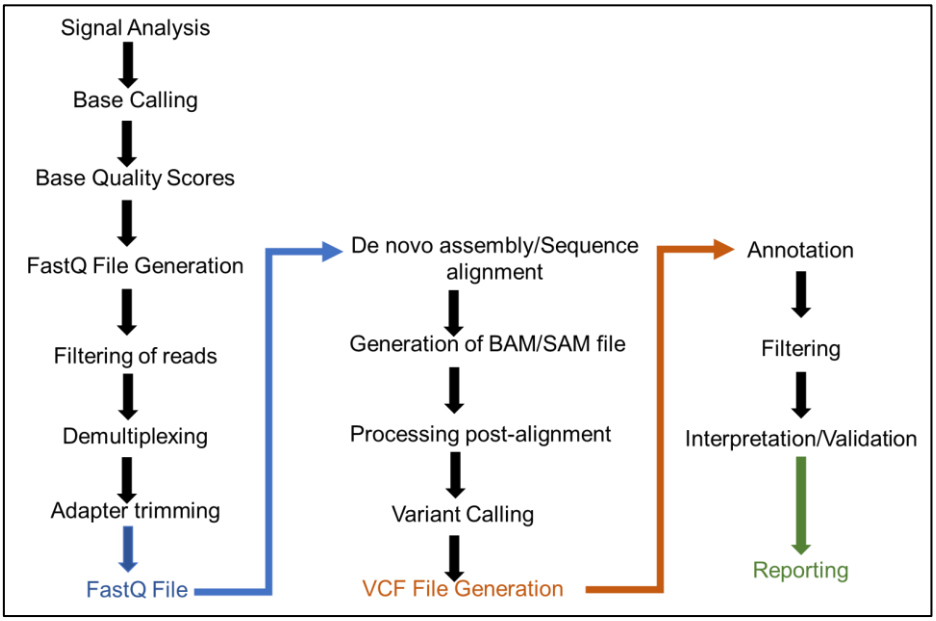

Figure 1. Basic bioinformatic workflow for NGS data analysis.

There are two types of methods regarding transcript assembly: de novo, which assembles reads exclusively based on sequence overlap, and the other one remains genome-based, which aligns reads to a reference genome first, and then the alignments that overlap are assembled. Methods based on the genome are more precise in general. De novo assembly, on the other hand, may be used to establish a representative collection of transcripts in the lack of a genome sequence or the case of a heavily fragmented genome; however, assembling is more difficult and prone to errors [48].

4.2. Tools for RNA-Seq assembly.

TopHat [49] is an open-source program for aligning RNA-Seq reads to a reference genome beyond utilizing or depending on the known splice sites. TopHat aligns RNA-Seq reads taken in FASTA or FASTQ format using a reference genome (as a Bowtie index), and SAM format alignments are produced. Even in the case of genes transcribed at paltry levels, TopHat can discover junctions when the default parameter values are used. TopHat 1.0.7 and later versions have been updated to take advantage of long paired reads and align them through splice junctions. Reads up to 75 base pairs or longer are divided into 3 or more segments of roughly identical size and are mapped separately [50,51]. For mapping the non-

junction reads to the reference genome, it employs Bowtie (http://bowtie-bio.sourceforge.net), a short-read mapping application [52,53]. The reads named initially unmapped reads (IUM reads) are the set aside reads that do not map to the genome. For every splice junction, Tophat scans the IUM reads for seeking reads that pan junctions. For consensus construction of the mapped areas, the assembly module MAQ is employed. TopHat is written in C++ and is available for Mac OS X and Linux. It includes MAQ and Bowtie and employs the SeqAn library[54].

Another open-source C++ program, Cufflinks assembler [66] works on Mac OS X and Linux. It can detect whole new transcripts and allocate reads to isoforms deterministically. It also includes the Cuffdiff and Cuffcompare utilities. Corroboration of Cufflinks output is done by Cuff compare, along with transfrags (assembled transcript fragments), to annotated transcriptomes and detection of transfrags that are common across many assemblies. The testing of differential expression is then performed by Cuffdiff. The creation of this assembler was aimed to look at transcriptional and splicing regulation, as well as to determine the minimum quantity of transcripts that ’‘explain' the reads, which implies that each read must be accommodated in some transcript. Cufflinks accept cDNA fragment sequences as input, aligned to the genome by TopHat or others, that could align reads spanning splice junctions without depending on gene annotation to yield spliced alignments[55–57].

4.3. Differential gene expression.

Differential gene expression is referred to the change or differences in the expression of genes or gene counts between two conditions that have been challenged experimentally. For analysis, quantifying transcriptomes or genes is required. There are different tools and algorithms that are utilized for the quantification of transcriptome and performing differential gene expression analysis [58].

4.4. Transcriptome quantification and differential expression from NGS data.

From RNA-Seq data, GFOLD, a fold change algorithm, generates biologically relevant rankings of genes with differential expression. According to the posterior distribution of log-fold change, GFOLD offers trust worthy statistics for changes in expression. When implemented for single replicate data sets, the authors demonstrated that GFOLD surpasses other frequently used approaches[59].To determine the importance of change between samples, Cuffdiff employs a model, namely, the beta negative binomial distribution. The model takes into consideration both cross-replicate variability and read mapping ambiguity to account for uncertainty. Cuffdiff displays the fold change for gene expression and statistical significance [60].

Cufflinks also come with a companion tool Cuffdiff, which estimates expression in 2 or more samples and analyses the statistical significance of each change observed in expression. The statistical model utilized to assess changes presumes that the read number generated by each transcript remains proportionate to its abundance but that this varies due to technical variations throughout library preparation and sequencing and biological variability among the same experiment replicates. Cufflinks is a program that calculates transcript-level fragment counts. To model the expression of gene G, Cuffdiff employs an algorithm. It is possible to obtain expression distribution for G using this technique. Cuffdiff evaluates the mean of these distributions and calculates their log ratio in duplicate’s presence for the estimation of the

distribution of expression's log-fold changes for G under the null hypothesis. Thousands of times in both situations, the procedure is repeated. All the samples are sorted and counted; how many of them are more severe than that of the log-fold change they absolutely saw from the real data to determine a p-value for witnessing the actual log-fold change? The p-value is calculated by dividing this number by the total number of drawings [61].

The edgeR is a negative binomial distribution-based statistical approach for profiling differential gene expression. The edgeR is meant to function with replicates, but it may also be used with data sets that do not include replicates. As indicated in Reference 34, we used edgeR to count uniquely mapped reads. Both edgeR and DESeq, accessible as R/Bioconductor packages, are downstream count-based analysis tools. The edgeR may evaluate both replicate and nonreplicated data sets [62].

The empirical Bayes approach, which allows the estimate of biological variation specific to genes, even for investigations with modest biological replication levels, is a unique aspect of edgeR functionality. edgeR can detect differential expression levels at the tag level, transcript, exon, or gene level. In fact, every genetic characteristic may be used to aggregate read counts. Analysis with edgeR at the exon level may readily be expanded for the detection of isoform-specific differential expression or differential splicing [62,63].

edgeR and DESeq are two techniques and R packages for assessing quantitative readouts (counts) from experiments like RNA-seq, which are high-throughput. In RNA-Seq, reads are allocated to a post-alignment feature, with each feature representing a targeted transcript. The total number of reads in a feature is an essential summary statistic [64].

The methods for analyzing array-based data presume that the response variable is normally distributed and continuous. Response variables, however, for digital approaches like ChIP-seq and RNA-seq, on the other hand, are discrete counts. As a result, the negative binomial distributions are used in the DESeq and edgeR techniques. The shot noise (also known as Poisson noise, sampling noise, or counting noise) that occurs from the counting aspect of the data is separated from the variance introduced by other forms of noise in edgeR and DESeq using a model. The excess variance is then represented as either uniform for edgeR or quasi-correlated with reading depth for DESeq. DESeq implies that read depth and extra-Poisson noise are related, whereas edgeR thinks they are not [64].

5. Conclusion

In conclusion, although many accomplishments were made in the past, there still needs to be advancements in the methods and pipelines. In the future, the sequencing platforms and the strategies for handling data are required to be improved. This will aid in reducing the rates of errors and increase the quality of data analysis. There is a rapid evolution of NGS consistently occurring to gain broader applicability. The future approach may require clinicians, scientists, and bioinformaticians to work in an ecosystem for better interpretation of data and arrive at a standard systemic level of analysis for deriving information that has diagnostic value in the clinical setting. Emerging developments in the field of artificial intelligence can further aid in improving the platforms and software for NGS analysis that will aid clinicians and scientists in resolving complex biological issues for developing novel therapeutic strategies.

Funding

This research received no external funding.

Acknowledgments

This research has no acknowledgment.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Hamed, T.; Saeed, M.A. Next Generation Sequencing Technologies: A Short Review. *Journal of Next Generation Sequencing & Applications***2015**, *01*, https://doi.org/10.4172/2469-9853.s1-006.
2. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of Age: Ten Years of next-Generation Sequencing Technologies. *Nature Reviews Genetics***2016**, *17*, 333–351, https://doi.org/10.1038/nrg.2016.49.
3. Auer, P.L.; Doerge, R.W. Statistical Design and Analysis of RNA Sequencing Data. *Genetics***2010**, *185*, 405–416, https://doi.org/10.1534/genetics.110.114983.
4. Nakato, R.; Sakata, T. Methods for ChIP-Seq Analysis: A Practical Workflow and Advanced Applications. *Methods***2021**, *187*, 44–53, https://doi.org/10.1016/j.ymeth.2020.03.005.
5. Wheeler, D.A.; Srinivasan, M.; Egholm, M.; Shen, Y.; Chen, L.; McGuire, A.; He, W.; Chen, Y.-J.; Makhijani, V.; Roth, G.T.; Gomes, X.; Tartaro, K.; Niazi, F.; Turcotte, C.L.; Irzyk, G.P.; Lupski, J.R.; Chinault, C.; Song, X.-z.; Liu, Y.; Yuan, Y.; Nazareth, L.; Qin, X.; Muzny, D.M.; Margulies, M.; Weinstock, G.M.; Gibbs, R.A.; Rothberg, J.M. The Complete Genome of an Individual by Massively Parallel DNA Sequencing. *Nature***2008**, *452*, 872–876, https://doi.org/10.1038/nature06884.
6. Liu, Z.; Roberts, R.; Mercer, T.R.; Xu, J.; Sedlazeck, F.J.; Tong, W. Towards Accurate and Reliable Resolution of Structural Variants for Clinical Diagnosis. *Genome Biol***2022**, *23*, https://doi.org/10.1186/s13059-022-02636-8.
7. Thermes, C. Ten Years of Next-Generation Sequencing Technology. *Trends in genetics : Trends in Genetics***2014**, *30*, 418–426, https://doi.org/10.1016/j.tig.2014.07.001.
8. Loman, N.J.; Misra, R. V.; Dallman, T.J.; Constantinidou, C.; Gharbia, S.E.; Wain, J.; Pallen, M.J. Performance Comparison of Benchtop High-Throughput Sequencing Platforms. *Nature Biotechnology***2012**, *30*, 434–439, https://doi.org/10.1038/nbt.2198.
9. Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A Tale of Three next Generation Sequencing Platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq Sequencers. *BMC Genomics***2012**, *13*, https://doi.org/10.1186/1471-2164-13-341.
10. Imelfort, M.; Batley, J.; Grimmond, S.; Edwards, D. Genome Sequencing Approaches and Successes. *Methods in Molecular Biology***2009**, *513*, 345–358, https://doi.org/10.1007/978-1-59745-427-8_18.
11. Liu, L.; Li, Y.; Li, S.; Hu, N.; He, Y.; Pong, R.; Lin, D.; Lu, L.; Law, M. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology***2012**, *2012*, https://doi.org/10.1155/2012/251364.
12. Escalona, M.; Rocha, S.; Posada, D. A Comparison of Tools for the Simulation of Genomic Next-Generation Sequencing Data. *Nature Reviews Genetics***2016**, *17*, 459–469, https://doi.org/10.1038/nrg.2016.57.
13. Kircher, M.; Sawyer, S.; Meyer, M. Double Indexing Overcomes Inaccuracies in Multiplex Sequencing on the Illumina Platform. *Nucleic Acids Research***2012**, *40*, https://doi.org/10.1093/nar/gkr771.
14. Baker, Q.B.; Al-Rashdan, W.; Jararweh, Y. Cloud-Based Tools for Next-Generation Sequencing Data Analysis. In: Proceedings of the 2018 5th International Conference on Social Networks Analysis, Management and Security, SNAMS 2018; **2018;** pp. 99–105,http://dx.doi.org/10.1109/SNAMS.2018.8554515.
15. Schatz, M. *BlastReduce: High Performance Short Read Mapping with MapReduce*.**2008.**
16. Mahamudin, A. *Bringing Genomics Research to the Cloud: A Qualitative Case Study of Genomics Cloud Computing Adoption.***2021**.
17. Matsunaga, A.; Tsugawa, M.; Fortes, J. CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications. In: Proceedings of the Proceedings - 4th IEEE International Conference on eScience. eScience 2008; **2008**; pp. 222–229.
18. Sakri, L.I.; Jagadeeshgowda, K.S. A Robust BSP Scheduler for Bioinformatics Application on Public Cloud. In:*Contemporary Issues in Communication, Cloud and Big Data Analytics*. Sarma, H.K.D.; Balas, V.E.;

Bhuyan, B.; Dutta, N. Eds.; Lecture Notes in Networks and Systems; Springer Singapore: Singapore, Volume 281,**2022**; pp. 161–171, https://doi.org/10.1007/978-981-16-4244-9_13.
19. Wall, D.P.; Kudtarkar, P.; Fusaro, V.A.; Pivovarov, R.; Patil, P.; Tonellato, P.J. Cloud Computing for Comparative Genomics. *BMC Bioinformatics***2010**, *11*, https://doi.org/10.1186/1471-2105-11-259.
20. Nguyen, T.; Shi, W.; Ruden, D. CloudAligner: A Fast and Full-Featured MapReduce Based Tool for Sequence Mapping. *BMC Research Notes***2011**, *4*, https://doi.org/10.1186/1756-0500-4-171.
21. Pireddu, L.; Leo, S.; Zanetti, G. Seal: A Distributed Short Read Mapping and Duplicate Removal Tool. *Bioinformatics***2011**, *27*, 2159–2160, https://doi.org/10.1093/bioinformatics/btr325.
22. Sharma, M.; Mondal, S.; Bhattacharjee, S.; Jabalia, N. Emerging Trends of Bioinformatics in Health Informatics. In:*Computational Intelligence in Healthcare*. Manocha, A.K.; Jain, S.; Singh, M.; Paul, S. Eds.; Health Information Science; Springer International Publishing: Cham, **2021**; pp. 343–367,http://dx.doi.org/10.1007/978-3-030-68723-6_19.
23. Langmead, B.; Hansen, K.D.; Leek, J.T. Cloud-Scale RNA-Sequencing Differential Expression Analysis with Myrna. *Genome biology***2010**, *11*, https://doi.org/10.1186/gb-2010-11-8-r83.
24. Lehmann, N.; Perrin, S.; Wallon, C.; Bauquet, X.; Deshaies, V.; Firmo, C.; Du, R.; Berthelier, C.; Hernandez, C.; Michaud, C.; Thieffry, D.; Le Crom, S.; Thomas-Chollier, M.; Jourdren, L. Eoulsan 2: an efficient workflow manager for reproducible bulk, long-read and single-cell transcriptomics analyses. *bioRxiv* **2021**,http://dx.doi.org/10.1101/2021.10.13.464219.
25. Hong, D.; Rhie, A.; Park, S.-S.; Lee, J.; Ju, Y.S.; Kim, S.; Yu, S.-B.; Bleazard, T.; Park, H.-S.; Rhee, H.; Chong, H.; Yang, K.-S.; Lee, Y.-S.; Kim, I.-H.; Lee, J.S.; Kim, J.-I.; Seo, J.-S. FX: an RNA-Seq analysis tool on the cloud. *Bioinformatics* **2012**, *28*, 721-723,https://doi.org/10.1093/bioinformatics/bts023.
26. Niemenmaa, M.; Kallio, A.; Schumacher, A.; Klemelä, P.; Korpelainen, E.; Heljanko, K. Hadoop-BAM: Directly Manipulating next Generation Sequencing Data in the Cloud. *Bioinformatics***2012**, *28*, 876–877, https://doi.org/10.1093/bioinformatics/bts054.
27. O'Connor, B.D.; Merriman, B.; Nelson, S.F. SeqWare Query Engine: Storing and Searching Sequence Data in the Cloud. *BMC Bioinformatics***2010**, *11*, https://doi.org/10.1186/1471-2105-11-S12-S2.
28. Wang, H.T. Contributions on Computational Intelligence in the Medical Sector. *Journal of Biomedical and Sustainable Healthcare Applications***2021**, *1*,1–8, https://doi.org/10.53759/0088/JBSHA202101001.
29. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; DePristo, M.A. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. *Genome Research***2010**, *20*, 1297–1303, https://doi.org/10.1101/gr.107524.110.
30. Brouard, J.S.; Bissonnette, N. Variant Calling from RNA-seq Data Using the GATK Joint Genotyping Workflow. *Methods in molecular biology (Clifton, N.J.)* **2022**, *2493*, 205-233, https://doi.org/10.1007/978-1-0716-2293-3_13.
31. Kwon, T.; Yoo, W.G.; Lee, W.J.; Kim, W.; Kim, D.W. Next-Generation Sequencing Data Analysis on Cloud Computing. *Genes and Genomics***2015**, *37*, 489–501, https://doi.org/10.1007/s13258-015-0280-7.
32. Jadeja, Y.; Modi, K. Cloud Computing - Concepts, Architecture and Challenges. In: Proceedings of the 2012 International Conference on Computing, Electronics and Electrical Technologies. ICCEET **2012**; pp. 877–880, https://doi.org/10.1109/ICCEET.2012.6203873.
33. Mell, P.; Grance, T. The NIST Definition of Cloud Computing. In:*Cloud Computing and Government: Background, Benefits, Risks*.**2011**; pp. 171–173,https://doi.org/10.6028/NIST.SP.800-145.
34. Celesti, A.; Fazio, M.; Celesti, F.; Sannino, G.; Campo, S.; Villari, M. New Trends in Biotechnology: The Point on NGS Cloud Computing Solutions. In: Proceedings of the Proceedings - IEEE Symposium on Computers and Communications; Volume 2016-August, **2016**; pp. 267–270, https://doi.org/10.1109/ISCC.2016.7543751.
35. Zhao, S.; Watrous, K.; Zhang, C.; Zhang, B. Cloud Computing for Next-Generation Sequencing Data Analysis. In:*Cloud Computing - Architecture and Applications*.**2017**; http://dx.doi.org/10.5772/66732.
36. Guo, X.; Yu, N.; Li, B.; Pan, Y. Cloud Computing for Next-Generation Sequencing Data Analysis. In:*Computational Methods for Next Generation Sequencing Data Analysis*.**2016**; pp. 1–24,https://doi.org/10.5772/66732.
37. Roy, P.; Kumar, R. A Hybrid Security Framework to Preserve Multilevel Security on Public Cloud Networks. In: Proceedings of the 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART). IEEE: MORADABAD, India, December 10,**2021**; pp. 336–340.
38. Calabrese, B.; Cannataro, M. Bioinformatics and Microarray Data Analysis on the Cloud. In:*Methods in Molecular Biology*. Volume 1375,**2016**; pp. 25–39, https://doi.org/10.1007/7651_2015_236.
39. Skums, P.; Artyomenko, A.; Glebova, O.; Ramachandran, S.; Mandoiu, I.; Campo, D.S.; Dimitrova, Z.; Zelikovsky, A.; Khudyakov, Y. Computational Framework for Next-Generation Sequencing of Heterogeneous Viral Populations Using Combinatorial Pooling. *Bioinformatics***2015**, *31*, 682–690, https://doi.org/10.1093/bioinformatics/btu726.
40. Edwards, D.; Batley, J. Plant Genome Sequencing: Applications for Crop Improvement. *Plant Biotechnology Journal***2010**, *8*, 2–9, https://doi.org/10.1111/j.1467-7652.2009.00459.x.

41. Li, Z.; Xu, Y. Bulk Segregation Analysis in the NGS Era: A Review of Its Teenage Years. *The Plant Journal***2022**, *109*, 1355–1374, https://doi.org/10.1111/tpj.15646.
42. Lee, H.C.; Lai, K.; Lorenc, M.T.; Imelfort, M.; Duran, C.; Edwards, D. Bioinformatics Tools and Databases for Analysis of Next-Generation Sequence Data. *Briefings in Functional Genomics***2012**, *11*, 12–24, https://doi.org/10.1093/bfgp/elr037.
43. Luo, J.; Wei, Y.; Lyu, M.; Wu, Z.; Liu, X.; Luo, H.; Yan, C. A Comprehensive Review of Scaffolding Methods in Genome Assembly. *Briefings in Bioinformatics***2021**, *22*, https://doi.org/10.1093/bib/bbab033.
44. Wilson, G.A.; Dhami, P.; Feber, A.; Cortázar, D.; Suzuki, Y.; Schulz, R.; Schär, P.; Beck, S. Resources for Methylome Analysis Suitable for Gene Knockout Studies of Potential Epigenome Modifiers. *GigaScience***2012**, *16*, https://doi.org/10.1186/2047-217X-1-3.
45. Nielsen, R.; Paul, J.S.; Albrechtsen, A.; Song, Y.S. Genotype and SNP Calling from Next-Generation Sequencing Data. *Nature Reviews Genetics***2011**, *12*, 443–451, https://doi.org/10.1038/nrg2986.
46. Kukurba, K.R.; Montgomery, S.B. RNA Sequencing and Analysis. *Cold Spring Harbor Protocols***2015**, *2015*, 951–969, https://doi.org/10.1101/pdb.top084970.
47. Hansen, K.D.; Brenner, S.E.; Dudoit, S. Biases in Illumina Transcriptome Sequencing Caused by Random Hexamer Priming. *Nucleic Acids Research***2010**, *38*, https://doi.org/10.1093/nar/gkq224.
48. Huang, X.; Chen, X.G.; Armbruster, P.A. Comparative Performance of Transcriptome Assembly Methods for Non-Model Organisms. *BMC Genomics***2016**, *17*, https://doi.org/10.1186/s12864-016-2923-8.
49. Trapnell, C.; Pachter, L.; Salzberg, S.L. TopHat: Discovering Splice Junctions with RNA-Seq. *Bioinformatics***2009**, *25*, 1105–1111, https://doi.org/10.1093/bioinformatics/btp120.
50. Bhati, J.; Avashthi, H.; Kumar, A.; Majumdar, S.G.; Budhlakoti, N.; Mishra, D.C. Protocol for Identification and Annotation of Differentially Expressed Genes Using Reference-Based Transcriptomic Approach. In:*Genomics of Cereal Crops*. Wani, S.H.; Kumar, A. Eds.; Springer Protocols Handbooks; Springer US: New York, NY, **2022**; pp. 175–193,https://doi.org/10.1007/978-1-0716-2533-0_7.
51. Lamanchai, K.; Salmon, D.L.; Smirnoff, N.; Sutthinon, P.; Roytrakul, S.; Leetanasaksakul, K.; Kittisenachai, S.; Jantasuriyarat, C. OsVTC1-1 RNAi Mutant with Reduction of Ascorbic Acid Synthesis Alters Cell Wall Sugar Composition and Cell Wall-Associated Proteins. *Agronomy***2022**, *12*, https://doi.org/10.3390/agronomy12061272.
52. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biology***2009**, *10*, https://doi.org/10.1186/gb-2009-10-3-r25.
53. Garg, V.; Varshney, R.K. Analysis of Small RNA Sequencing Data in Plants. In *Plant Bioinformatics*; Edwards, D., Ed.; Methods in Molecular Biology; Springer US: New York, NY, Volume 2443, **2022**; pp. 497–509,https://doi.org/10.1007/978-1-0716-2067-0_26.
54. Zeni, A.; Di Donato, G.W.; Di Tucci, L.; Rabozzi, M.; Santambrogio, M.D. The Importance of Being X-Drop: High Performance Genome Alignment on Reconfigurable Hardware. In: Proceedings of the 2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM); IEEE: Orlando, FL, USA, May **2021**; pp. 133–141, https://doi.org/10.1109/FCCM51124.2021.00023.
55. Kang, D.S.; Kim, S.; Cotten, M.A.; Sim, C. Transcript Assembly and Quantification by RNA-Seq Reveals Significant Differences in Gene Expression and Genetic Variants in Mosquitoes of the Culex Pipiens (Diptera: Culicidae) Complex. *Journal of Medical Entomology***2020**, *58*, 139-145,https://doi.org/10.1093/jme/tjaa167.
56. Dong, W.; Yang, J.; Zhang, Y.; Liu, S.; Ning, C.; Ding, X.; Wang, W.; Zhang, Y.; Zhang, Q.; Jiang, L. Integrative Analysis of Genome-wide DNA Methylation and Gene Expression Profiles Reveals Important Epigenetic Genes Related to Milk Production Traits in Dairy Cattle. *J Anim Breed Genet***2021**, *138*, 562–573, https://doi.org/10.1111/jbg.12530.
57. Barral, A.; Pozo, G.; Ducrot, L.; Papadopoulos, G.L.; Sauzet, S.; Oldfield, A.J.; Cavalli, G.; Déjardin, J. SETDB1/NSD-Dependent H3K9me3/H3K36me3 Dual Heterochromatin Maintains Gene Expression Profiles by Bookmarking Poised Enhancers. *Molecular Cell***2022**, *82*, 816-832.e12, https://doi.org/10.1016/j.molcel.2021.12.037.
58. McDermaid, A.; Monier, B.; Zhao, J.; Liu, B.; Ma, Q. Interpretation of Differential Gene Expression Results of RNA-Seq Data: Review and Integration. *Briefings in Bioinformatics***2019**, *20*, 2044–2054, https://doi.org/10.1093/bib/bby067.
59. Feng, J.; Meyer, C.A.; Wang, Q.; Liu, J.S.; Liu, X.S.; Zhang, Y. GFOLD: A Generalized Fold Change for Ranking Differentially Expressed Genes from RNA-Seq Data. *Bioinformatics***2012**, *28*, 2782–2788, https://doi.org/10.1093/bioinformatics/bts515.
60. Trapnell, C.; Hendrickson, D.G.; Sauvageau, M.; Goff, L.; Rinn, J.L.; Pachter, L. Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq. *Nature Biotechnology***2013**, *31*, 46–53, https://doi.org/10.1038/nbt.2450.
61. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; Van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and

Isoform Switching during Cell Differentiation. *Nature Biotechnology***2010**, *28*, 511–515, https://doi.org/10.1038/nbt.1621.
62. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics***2009**, *26*, 139–140, https://doi.org/10.1093/bioinformatics/btp616.
63. Sustr, F.; Machackova, T.; Pesl, M.; Trachtova, K.; Svacinova, J.; Starek, Z.; Spinarova, L.; Kianicka, B.; Slaby, O.; Novak, J. MicroRNAs as the Predictors for Atrial Fibrillation Recurrence after Catheter Ablation: Next-Generation Sequencing Study. *EP Europace***2022**, *24*, https://doi.org/10.1093/europace/euac053.272.
64. Anders, S.; Huber, W. Differential Expression Analysis for Sequence Count Data. *Genome Biology***2010**, *11*, https://doi.org/10.1186/gb-2010-11-10-r106.