UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

A Review of Bioinformatics Model and Computational Software of Next Generation Sequencing

[1]Jalilah Arijah Mohd Kamarudin, [1]Afnizanfaizal Abdullah & [2]Roselina Sallehuddin
[1]Synthetic Biology Research Group
[2]Soft Computing Research Group
Faculty of Computing
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia
Email: [1]jalilaharijah@gmail.com, [1]afnizanfaizal@utm.my, [2]roselina@utm.my

Submitted: 12/01/2018. Revised edition: 29/10/2018 Accepted: 31/10/2018. Published online: 30/05/2019
DOI: https://doi.org/10.11113/ijic.v9n1.217

***Abstract*—In the past decade it has become increasingly the effort for researcher to surpass the bioinformatics challenges foremost in next generation sequencing (NGS). This review paper gives an overview of the computational software and bioinformatics model that has been used for next generation sequencing. In this paper, the description on functionalities, source type and website of the program or software are provided. These computational software and bioinformatics model are differentiating into three types of bioinformatics analysis stages including alignment, variant calling and filtering and annotation. Besides, we discuss the future work and the development for new bioinformatics tool to be advanced.**

***Keywords*—Bioinformatics model, computational software, next generation sequencing**

I. INTRODUCTION

The next-generation sequencing (NGS) methods give various opportunities for numerous applications in biomedical research and life sciences. There are several powerful generation sequencing platforms such as Illumina/Solexa, AB/SOLiD, Roche/454 and LIFE/Ion Torrent are indeed useful for researchers in high-throughput genetic analysis [1]. However, the computational challenges are still remains occur especially in bioinformatics. In genomic research the biggest challenge is large-scale of datasets and the efficiency of computational analysis to perform an accurate result. In this paper, we provide an overview of several computational software and bioinformatics model for genomic analysis and provide some future outlook on this area.

In bioinformatics analysis, the nucleotide base call is used as a final output regarding to different sequencing technologies use different raw data [1]. Furthermore, bioinformatics analysis consists of three common stages including alignment, variant calling and filtering and annotation. Alignment is the first step that being used to match each of the short reads onto positions on a reference genome [1]. The resulting sequences for this step are stores in a BAM (Binary alignment/map) or SAM (sequence alignment/map) file [2]. The next step is variant calling which functioning to compare between the aligned sequences and know sequences to identify the positions diverge from the reference position [1]. The results of positions list or calls in form of variant call format (VCF) file [3]. The final stage includes both filtering and annotation. Filtering is defined as method to reduce the large number of variants into smaller set. Meanwhile, annotation is process of collecting information of each variant which detected [1]. The final result from the analysis makes some compromises in predicting and identifying the single nucleotides polymorphism, the functional effect, the activity or function of the gene and also associated disease [1]. In this review, we discuss several of bioinformatics models or options for each of these three stages.

II. ALIGNMENT

This paper reviewed summary of the current computational and bioinformatics model that has been used for alignment process such as Bowtie, SOAP and SHRiMP. As describe in this paper, we state their functionalities, process running, input, output and their application type and also development language.

A. Bowtie

Langmead and group [4] introduces alignment program called Bowtie for aligning short DNA sequence reads into human genomes. Bowtie uses Burrows-Wheeler techniques with powerful backtracking algorithm which grants mis-matches [4]. Bowties have high possibly to provide better speed and memory usage but it may fail to gives high quality of read mapping with no precise match exists [5]. It is written in C++ and the SeqAn library is used [6]. Users are available and free to access through http://bowtie.cbcb.umd.edu [4].

B. SOAP

SOAP is a program that designed to handle the large amounts of short reads by using the new generation Illumina-Solexa sequencing technology [7]. SOAP was developed by Li and co-workers [7] makes compromises to provide effective gapped and un-gapped alignment for short oligonucleotides into reference sequences. SOAP was build-up as multifunctioning program for several applications such as re-sequencing of single-read or pair-read, mapping of mRNA tag sequence and discovery of small RNA [7]. In addition, SOAP also was designed as alignment algorithm which specifically for single nucleotides polymorphisms detection and genotyping [5].

SOAP accepts FASTA or FASTQ format for reference and also query read. SOAP is written in C++ language and use Macintosh or Linux/Unix system as a platform. It is freely available to access through http://soap.genomics.org.cn [7].

C. SHRiMP

SHRiMP or SHort Read Mapping Package is a set of algorithms and methods for mapping of short reads to a genome was developed by Rumble *et al*. [8]. SHRiMP was designed specifically for SOLiD colour-space reads mapping [5]. Moreover, it is also advances in sequence alignment include q-gram filters [9], spaced seeds [10], and Smith Waterman alignment algorithm [11]. SHRiMP is publicly available to access at http://compbio.cs.toronto.edu/shrimp [8].

III. VARIANT CALLING

After the alignment steps, the next steps in bioinformatics analysis is variant calling. We provide an overview of several of the software package for variant calling in this paper.

A. GATK

A computational program called Genome Analysis Toolkit (GATK) was designed to develop analysis tools for next-generation DNA sequence [12]. GATK is computational method that use alignment reads for variant calling [1]. This computational program was build-up using the functional programming model of MapReduce. GATK engine is efficiently in accessing the next-generation sequencing data by effectively handling the complexity. Besides that, it is improves the stability, correctness and efficiency of memory usage. GATK package has been applied for large-scale projects include 1000 Genome Project [13] and The Cancer Genome Atlas [14]. GATK is written in Java and uses sequence alignment/map (SAM) library [12].

B. VarScan

Koboldt *et al*. [15] presents an open source tool called VarScan for variant detection of short reads alignments. VarScan is compatible for insertions, deletions and SNPs detection and also for evaluating the frequency of massively parallel sequencing data. VarScan was designed to flexible with some read aligners such as BLAST, Newbler, cross_match, Bowtie [4] and Novoalign [16]. A full workflow for variant detection from alignments within next generation sequencing data is provided in VarScan package [15].

Additionally, VarScan was provides the accuracy of the sequence alignment based on the specificity and sensitivity in variant calls. VarScan also companionable in both individual and pooled samples which results for effectively variant calls data of several sequencing platform. This VarScan tool was supported on all platforms and implemented as a Perl package. The documentation and source code of VarScan is freely acess via http://genome.wustl.edu/tools/cancer-genomics [15].

C. Atlas2

Baylor Genome Center was develops Atlas2 for variant calling of aligned data [1]. This computational tool is compatible for various sequencing platform such as Illumina, Roche 454 and SOLiD which used to detect short range and SNPs. Apart from that, if Atlat2 is implemented on a computational cluster, it assists to running 92 exome from 64 processors within 4 hours. Atlas2 was applied for 1000 Genome Project with 92 samples of whole genome [17].

Atlas2 was enables to run on a windows platform and efficiency to analyze BAM file with 28 GB whole-exome within 2 hours [18]. Atlas2 is open source and available to download via http://sourceforge.net/projects/atlas2/.

The identified matching pairs represent the copy-pasted regions. The following steps are employed for the CMFD framework:

IV. FILTERING AND ANNOTATION

In this section, we review several computational tools that useful in identifying disease-causal variants among numerous candidates. There are three tools involves in filtering and annotations including SnpEff, ANNOVAR and SNPeffect Database.

A. SnfEff

The computer program for clustering the effects of variants in genome sequences is called as SnpEff [19]. This computational program was developing using Java-based program and effectively assembling SNP, MNP variants and indel within genomic sequences. It is open source and freely available for users through http://snpeff.sourceforge.net. SnpEff is a web-based platform and it have several main features includes high speed to make predict within per second and flexibility to add custom annotations and genomes [19]. Furthermore, based on alternative transcripts this program enables to discover multiple different functions for a single variant regarding to competing predictions [1].

B. ANNOVAR

Annotate Variation (ANNOVAR) was developed by Wang *et al*. [20] to annotate single nucleotides variants (SNPs) and insertions or deletions besides annotate the functional effects of variants genes [21]. Annotation is compatible in gene-based which users are able to select the gene definition system such as ENSEMBL, GENCODE, RefSeq and UCSC [1]. Another features of ANNOVAR including the capability to analyze genomic region-based annotations and makes a comparison of variants to existing variation databases.

ANNOVAR is useful as standalone application and it use text-based input files. ANNOVAR is freely accessible and open source which users are available to download at http://www.openbioinformatics.org/annovar/ [20].

C. SNPeffect

Another bioinformatics tool for prediction of the effect of protein coding SNPs towards the structural phenotype on proteins is known as SNPeffect database. SNPeffect has another functionality used for structural phenotyping by integrates aggregation prediction (TANGO), chaperone-binding prediction (LIMBO), amyloid prediction (WALTZ) and protein stability analysis (FoldX) [21].

This database contains 63 410 human unknown SNVs data and every 6 months SNPeffect database is updated from database of UniProt human variation. Using interface of SNPeffect database users are allowed to find SNVs via filtering based on gene name, disease, molecular phenotypic effects, mutation type, dbSNP identifier and UniProt identifier. SNPeffect is access via http://snpeffect.switchlab.org [21].

TABLE 1. Summary of Bioinformatics Model and Computational Software for Next Generation Sequencing

Program/ Software	**Description**	**Source Type**	**Website**
Alignment			
Bowtie	Aligning the short DNA reads to human genome	Open source	http://bowtie.cbcb.umd.edu
SOAP	Aligning the short oligonucleotides onto reference sequences	Open source	http://soap.genomics.org.cn
SHRiMP	Mapping method of short reads into a genome	Open source	http://compbio.cs.toronto.edu/shrimp
Variant Calling			
GATK	Method for variant calling which use alignment reads	Open source	-
VarScan	Tool for variant detection of short reads alignments	Open source	http://genome.wustl.edu/tools/cancer-genomics
Atlas2	Tool for short range and SNPs detection	Open source	http://sourceforge.net/projects/atlas2/
Filtering And Annotation			
SnpEff	Clustering the effects of variants in genome sequences	Open source	http://snpeff.sourceforge.net
ANNOVAR	Annotate SNPs, insertions or deletions and functional effects of variants genes	Open source	http://www.openbioinformatics.org/annovar/
SNPeffect	Tool of the effect of protein coding SNPs prediction towards the structural phenotype on proteins	Open source	http://snpeffect.switchlab.org

V. CONCLUSION

This paper has summarized the several type of computational software and bioinformatics model of next generation sequencing based on three stage of analysis including alignment, variant calling and filtering and annotation. Table 1 review the entire computational software or model that has been listed based on three stage of bioinformatics analysis respectively.

New technology and algorithms gives future improvements especially increase the length of sequencing reads, complete more genomes and provides better populated for annotation database. Additionally, by using sequencing technology the process of alignment will become more accurate even though facing the longer reads [1]. The process of variant calling will have an advantages from large databases of completed genome and lastly the filtering and annotation will be increase the functional prediction regarding to improvement of more data and databases [1].

In future works, researchers need to focus on emerging tools to increase the effectiveness of analyzing samples not only as a homogenous whole. Besides that, some tools needs to be integrated for high throughput modalities which advances the interpretation of proteomics into genomic [1]. As a conclusion, researchers have been provided a useful guide on bioinformatics and computational framework of next generation sequencing in this review.

ACKNOWLEDGMENT

We would like to acknowledge Malaysian Ministry of Higher Education and UTM for sponsoring our work in Research University Grant (GUP) vot number Q.J130000.2528.16H57 and Research Management Centre, Universiti Teknologi Malaysia for managing our grant.

REFERENCES

[1] Dolled-Filhart, M. P., Lee, M., Ou-yang, C. W., Haraksingh, R. R., & Lin, J. C. H. (2013). Computational and Bioinformatics Frameworks for Next-generation Whole Exome and Genome Sequencing. *The Scientific World Journal*, 2013.

[2] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & Durbin, R. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, 25(16), 2078-2079.

[3] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., & McVean, G. (2011). The variant Call Format and VCFtools. *Bioinformatics*, 27(15), 2156-2158.

[4] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and Memory-efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biology*, 10(3), R25.

[5] Ruffalo, M., LaFramboise, T., & Koyutürk, M. (2011). Comparative Analysis of Algorithms for Next-generation Sequencing Read Alignment. *Bioinformatics*, 27(20), 2790-2796.

[6] Döring, A., Weese, D., Rausch, T., & Reinert, K. (2008). SeqAn an Efficient, Generic C++ Library for Sequence Analysis. *BMC Bioinformatics*, 9(1), 11.

[7] Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: Short Oligonucleotide Alignment Program. *Bioinformatics*, 24(5), 713-714.

[8] Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., & Brudno, M. (2009). SHRiMP: Accurate Mapping of Short Color-space Reads. *PLoS Computational Biology*, 5(5), e1000386.

[9] Rasmussen, K. R., Stoye, J., & Myers, E. W. (2006). Efficient q-gram Filters for Finding all ε-matches over a Given Length. *Journal of Computational Biology*, 13(2), 296-308.

[10] Califano, A., & Rigoutsos, I. (1993, June). FLASH: A Fast Look-up Algorithm for String Homology. *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*. IEEE, 353-359.

[11] Waterman, M. S. (1981). Identification of Common Molecular Subsequence. *Mol. Biol*, 147, 195-197.

[12] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce Framework for Analyzing Next-generation DNA Sequencing Data. *Genome Research*.

[13] 1000 Genomes Project Consortium. (2010). A Map of Human Genome Variation from Population-scale Sequencing. *Nature*, 467(7319), 1061.

[14] Cancer Genome Atlas Network. (2012). Comprehensive Molecular Portraits of Human Breast Tumours. *Nature*, 490(7418), 61.

[15] Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., & Ding, L. (2009). VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples. *Bioinformatics*, 25(17), 2283-2285.

[16] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., & Mende, D. R. (2010). A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing. *Nature*, 464(7285), 59.

[17] Challis, D., Yu, J., Evani, U. S., Jackson, A. R., Paithankar, S., Coarfa, C., & Yu, F. (2012). An Integrative Variant Analysis Suite for Whole Exome Next-generation Sequencing Data. *BMC Bioinformatics*, 13(1), 8.

[18] Ji, H. P. (2012). Improving Bioinformatic Pipelines for Exome Variant Calling. *Genome Medicine*, 4(1), 7.

[19] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., & Ruden, D. M. (2012). A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain w1118; iso-2; iso-3. *Fly*, 6(2), 80-92.

[20] Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-throughput Sequencing Data. *Nucleic Acids Research*, 38(16), e164-e164.

[21] De Baets, G., Van Durme, J., Reumers, J., Maurer-Stroh, S., Vanhee, P., Dopazo, J., & Rousseau, F. (2011). SNPeffect 4.0: On-line Prediction of Molecular and Structural Effects of Protein-coding Variants. *Nucleic Acids Research*, 40(D1), D935-D939.