Journal of Computer Science and Cybernetics, V.35, N.2 (2019), 105–118
DOI 10.15625/1813-9663/35/2/13827

REVIEW

A COMPUTATIONAL FRAMEWORK TO ANALYZE HUMAN GENOMES

LE SY VINH
University of Engineering and Technology, Vietnam National University Hanoi
Vinhls@vnu.edu.vn

Abstract. The advent of genomic technologies has led to the current genomic era. Large-scale human genome projects have resulted in a huge amount of genomic data. Analyzing human genomes is a challenging task including a number of key steps from short read alignment, variant calling, and variant annotating. In this paper, the state-of-the-art computational methods and databases for each step will be analyzed to suggest a practical and efficient guideline for whole human genome analyses. This paper also discusses frameworks to combine variants from various genome analysis pipelines to obtain reliable variants. Finally, we will address advantages as well as discordances of widely-used variant annotation methods to evaluate the clinical significance of variants. The review will empower bioinformaticians to efficiently perform human genome analyses, and more importantly, help genetic consultants understand and properly interpret mutations for clinical purposes.

Keywords. Human Genome Sequencing; Bioinformatics; Human Genome Analyses; Single Nucleotide Variants; Structural Variants; Variant Annotation.

1. INTRODUCTION

The human genome is the blueprint of life. It comprises more than three billions of nucleotides and can be considered as a string of A, C, G, T characters representing for four nucleotides, i.e., Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The difference between genomes of two human individuals is about 1%, however, it results in two distinguishable persons from appearance, health conditions, and even characteristics. The genome of each person contains about 4 million nucleotide variants in comparison with the reference genome. The nucleotide variants might lead to the changes of protein sequences, structures, and functions, and consequently, affect health conditions or cause diseases.

The current cost of sequencing a whole human genome is only around 1000 US Dollars. This allows us to perform large-scale human genome projects to study genotype-phenotype relationships; personalized medicines; the evolutionary and diversity of human populations [1, 2, 3, 4]. Various human genome projects have been conducted for different Asian populations such as 100 Malay genomes project to detect low-frequency and rare variants [2]; 35 Korean genomes project to decipher the genetic architecture of the Korean population [5]; 90 Han Chinese genomes project to investigate Han Chinese human genomes [6]; Vietnamese genomes project to build Vietnamese variation database [7, 8].

Analyzing a human genome are separated into two phases: the wet-lab phase and the bioinformatics phase (see Figure 1). The wet-lab phase collects genomic DNA, then sequences the genome using next generation sequencing (NGS) technologies. The output of the wet-lab

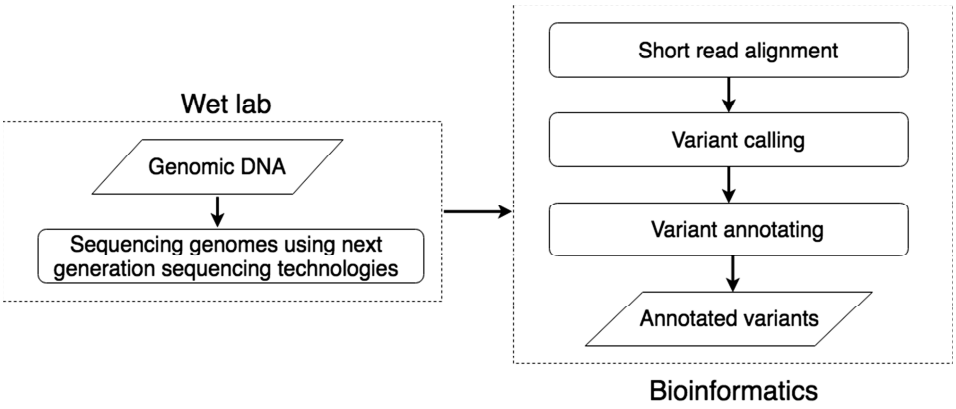

Figure 1. The workflow to sequence and analyze a human genome including wet-lab phase and bioinformatics phase

phase is a large dataset of DNA sequences that will play as the input for the bioinformatics phase. The bioinformatics phase will analyze the dataset to call variants and subsequently annotate them to determine the genotype-phenotype of the called variants. The bioinformatics phase is a computationally complex process including three main steps: short-read alignment, variant calling, and variant annotating. Different computational methods have been proposed for each step. A genome analysis pipeline is a combination of specific methods for the three steps.

The paper reviews and discusses the state-of-the-art methods to create efficient and practical genome analysis pipelines for genomic studies as well as medical applications. It helps readers understand genome sequencing technologies, genome analyses, variant annotations, and especially pathological interpretations of variants for clinical purposes.

2. GENOME SEQUENCING

A human genome includes more than three billions of nucleotides/bases. Two popular and widely-used sequencing technologies are Sanger sequencing and next generation sequencing. The Sanger sequencing technology can obtain sequences of length from several hundred to a few thousand nucleotides. The technology costs billion US Dollars and several years to sequence a whole human genome. The Sanger sequencing technology is only suitable for sequencing targeted regions in the genome.

Currently, the NGS technologies are widely used to sequence whole human genomes. The NGS technologies will massively break the whole genome into small segments, called short reads. A large number of short reads are then sequenced in parallel to cover all positions in the genome. The depth coverage of a sequenced genome is the average number of short reads covering any position in the genome. For example, sequencing a genome with depth coverage of 30x means that on average each position in the genome is sequenced 30 times, and consequently covered by 30 short reads.

A number of NGS technologies have been developed. Illumina is recently the most popular NGS technology for whole human genome sequencing. The Illumina machine produces short reads of length from a few dozen to a few hundred nucleotides. Other NGS technologies

such as Pacific Biosciences (PacBio) or Oxford Nanopore can sequence longer reads with ten thousand nucleotides. However, these technologies produce a high sequencing error rate.

NGS machines independently estimate a quality score for each nucleotide (base) based on the standard Phred-quality score. Note that a Phred score of 20 corresponds to a 1% error rate in base calling. The estimation of base quality scores is subject to various sources of errors, i.e., technical errors, physics or chemistry errors, that result in inaccurate base quality scores. To solve the problem, the Phred quality scores estimated from the NGS machines are typically recalibrated based on empirical error models. For example, a model shows that three identical bases AAA rarely appear in a row, i.e., whenever two identical bases AA are already in a row, the next base A will actually have 1% higher error rate than estimated from sequencing machines. Thus, the model will decrease 1% from the quality score of any base A if it comes after two bases AA in a read. The base quality score recalibration improves the accuracy of base quality scores that will, in turn, improve the accuracy of other downstream analyses.

3. GENOME ANALYSIS

3.1. Variants

Two variant types when comparing a genome with the reference genome are single nucleotide variants and structural variants. The single nucleotide variants imply the change of one or some nucleotides. Three common types of single nucleotide variants are substitutions, insertions, and deletions. For example, there are 2 nucleotide variants when comparing a sequence S="ACTTG-TT" with the reference sequence R="ACATGATT" where '-' represents for a deletion in the sequence. Specifically, there is a nucleotide substitution at the third position (from A to T) and a nucleotide deletion at the 6^{th} position of sequence S (or an insertion of nucleotide A at the position 6^{th} of sequence R). As we cannot differentiate deletions and insertions, we call them indels.

The structural variants are the changes of large nucleotide segments in the genome. They include copy number variants (i.e., genomic segments with different number of copies in comparison to the reference genome), inversions and translocations that do not change the total DNA content of the genome. Although the number of structural variants is much less than the number of nucleotide variants, structural variants affect an order of magnitude more nucleotides than single nucleotide variants [9].

The human genome is diploid that means it has two alleles/nucleotides at each position with one allele inherited from the father and the other allele inherited from the mother. A haplotype is a group of alleles that are inherited together from either mother or father. A pair of alleles at a specific position is called a genotype. If a genotype consists of a non-reference allele, it indicates a variant at the position.

3.2. Short read alignment

A genome sequenced by the NGS technology is represented by a set of short reads. As sequencing machines do not provide information about exact positions of short reads in the genome, the first task in analyzing a genome is to determine the positions of short reads in the reference genome, called short read alignment/mapping step. A short read can be mapped

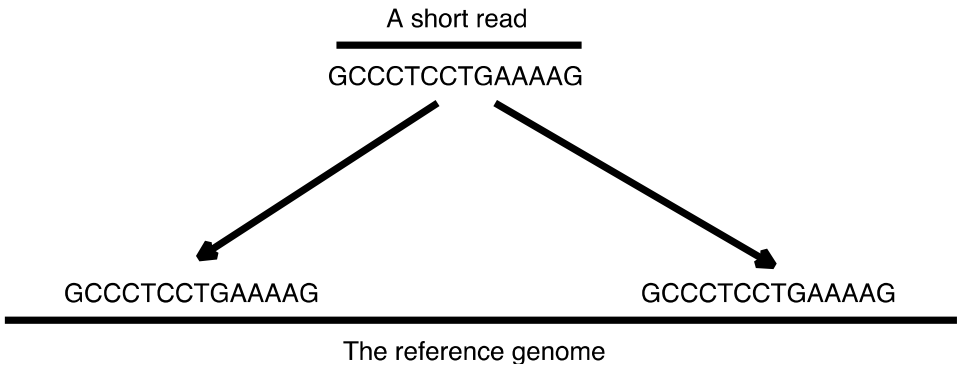

Figure 2. Exact alignment of a short read to the reference genome. The short read might occur at several positions in the genome

exactly to the reference genome or mapped to the reference genome with some mismatches. Some short reads are not mapped to any position in the reference genome as they might be error reads or new nucleotide insertions.

Exact matching alignment

Given a reference genome R with length n and a short read S, the exact matching alignment will search the positions of the short read S in the reference genome R (see Figure 2). The short read S might be mapped exactly at different positions in R. Classical matching string algorithms such as Knuth-Morris-Pratt [10] can solve the exact mapping problem with the complexity proportional to n. The classical matching algorithms are not practical for mapping billion short reads to the human reference genome.

As the reference genome R is typically fixed, indexing techniques have been proposed to solve the problem. Generally, indexing techniques construct an auxiliary data structure (an index) of R such that it can quickly determine the occurrences of a short read without fully scanning the content of R. Classical indexing algorithms for the problem such as suffix tree and suffix array require a memory of $\mathrm{O}(n \log n)$ bits. They are not efficient enough to handle the whole human genome because it is technically impractical to store a too large index structure in the RAM for fast access.

Burrows-Wheeler transform (BWT) is currently considered as the most powerful indexing technique to handle the exact matching alignment problem for the whole human genome [11]. The space requirement of the BWT technique is only proportional to the length of R in compressed form. Theoretical concepts and engineering aspects of the BWT technique are fully discussed elsewhere [12, 13].

Approximate matching alignment

The variants in the genome in combination with sequencing errors resulted by the NGS machines make the short read alignment problem complicated. A large proportion of short reads cannot be mapped exactly to the reference genome. Different short read matching strategies have been developed to approximately map short reads to the reference genome. The approximate short read matching algorithms allow to map short reads to the reference genome with one or few nucleotide mismatches. The classical algorithm for approximately matching two sequences uses the dynamic programming paradigm [14] with a complexity of

$O(m, n)$ where m and n are the lengths of two sequences. Apparently, the classical algorithm is not practical for approximately matching short reads to the whole human reference genome.

Hashing based-methods have been proposed to solve the problem. The methods use a seed-and-extend strategy to determine the occurrences of short reads in the reference genome. The strategy assumes that a short read contains only few errors, therefore, a large part of a short read will be matched exactly without errors into the reference genome. The exact matching part is called the "seed". The occurrences of seeds in the reference genome are then extended to find approximate matching alignments for a short read. If the seed length is long enough, the number of seeds appearing in the reference genome is limited, and the alignment process is efficient for large datasets. Several hashing-based aligners have been developed and widely used such as SOAP [15], Stampy [16], or SHRiMP2 [17].

BWT-based techniques have been also developed for the approximate short read matching problem. The key idea is the use of FM-index [11] that allows implementing space-efficiency and fast string searching algorithms. BWT-based methods are an order of magnitude faster than hash-based methods [18]. The BWT-based aligners such as Bowtie2 [19] or BWA [20] are the most widely used in large scale whole human genome projects such as the 1000 Human Genomes (1KG) project.

Our studies on benchmark datasets showed no significant difference when calling variants based on alignments from Bowtie2 and BWA [21]. The approximate short read alignment algorithms are fully described by Stefan Canzar and Steven L. Salzberg [22].

3.3. Single nucleotide variant calling

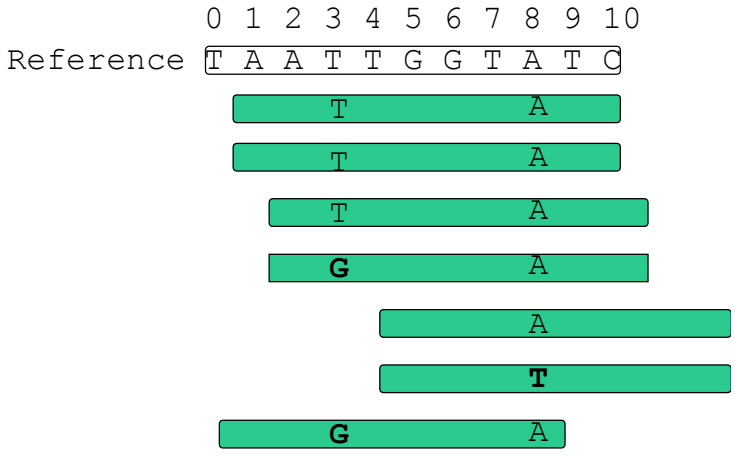

Figure 3. Short reads are mapped into the reference genome. Each position in the reference genome is covered by multiple short reads

Single nucleotide variants are the most prevalent variants in the human genome. A number of methods have been proposed to call variants from short read alignments. Consider a specific position in a genome covered by p short reads, variant callers will examine nucleotides from the p short reads to determine the existence of any variant at the position. Figure 3 shows an example that five short reads cover position 3 of which three short reads contain

nucleotide T and two short reads contain nucleotide G. The data indicate a high possibility that variant G occurs at position 3 and T/G is the genotype at the position.

The counting approach

The most simple algorithm to detect a variant at a specific position is to count the number of short reads including the variant. A counting algorithm typically includes two steps:

Filtering step: This step filters out all short reads and/or nucleotides with quality scores smaller than a predefined threshold. A typical cutoff is $Q_{Phred} = 20$, i.e., the probability of an error base is 1%.

Variant calling step: The step counts the number of non-reference alleles/nucleotides that differ from the reference nucleotide. If the number is greater than a predefined cutoff, a variant is called at the position. Technically, the number of non-reference alleles should range from 20-80% of all bases.

The simple counting method might work well with high coverage ($>= 30x$) datasets. As the short reads contain errors, using predefined quality filtering or non-reference allele percentage cutoffs result in inaccurate variant calls.

The statistical approach

Statistical approaches have been developed to call variants from short read alignments. One major advantage of the statistical approaches is the use of prior probability of variants and the quality scores of short read alignments to infer a posterior probability for each variant [23, 24].

Let $D = \{d_1 \dots d_m\}$ be a set of m short reads covering a given position. We assume that these m short reads are independent, and the likelihood $P(D|G)$ of genotype G can be calculated by the product of $P(d_i|G)$ for all short reads. The probability $P(d_i|G)$ for a given genotype G can be estimated from the quality score of short read d_i. Note that the likelihood $P(d_i|G)$ can be improved by recalibrating quality scores of nucleotides using empirical error models.

The posterior probability of a genotype G given the data D can be calculated using the Bayes formula as following

$$P\left(G|D\right) = \frac{P\left(D|G\right)P\left(G\right)}{P\left(D\right)}$$

where P(G) is the prior probability of genotype G and P(D) is the probability of data D. The prior probability of genotype G can be estimated from databases such as the dbSNP or the 1KG project.

A number of statistical methods have been proposed to call variants, notably GATK [24, 25], Samtools [26], Platypus [27], or Freebayes [28]. Among them, GATK is the most popular and widely-used method to call variants from short read alignments.

It is well known that variants called from different methods can be discordant, therefore, large-scale human genome projects usually use multiple methods to call variants. Software packages such as Seqmule have been developed allowing users to call variants from various programs [29].

We have proposed a voting method, called Genomedics, to call and combine variants from several methods to determine consensus and reliable variants [21]. Genomedics includes two widely used aligners, i.e., BWA and Bowtie2, and three popular callers, i.e., GATK, Freebayes, and Platypus. The combinations of two aligners and three callers result in six

different pipelines (see Figure 4). A variant is considered reliable if it is called by at least two different pipelines. Experiments on benchmark datasets showed that Genomedics produced better results than single pipelines tested. Specifically, Genomedics has the F-score greater than 93.3% while the pipeline of BWA and GATK is the second best method (i.e, F-score = 92.7%). Genomedics has been used to analyze whole exomes and genomes in various human genome projects [8, 30, 31].

We note that large-scale human genome projects usually include dozens to thousands of genomes. The variants are typically called simultaneously from all genomes. This strategy allows us to efficiently call variants from low coverage genomes because low coverage data from a single genome might not provide enough evidence for calling variants. All current popular variant callers can handle multiple genomes with different depth coverages.

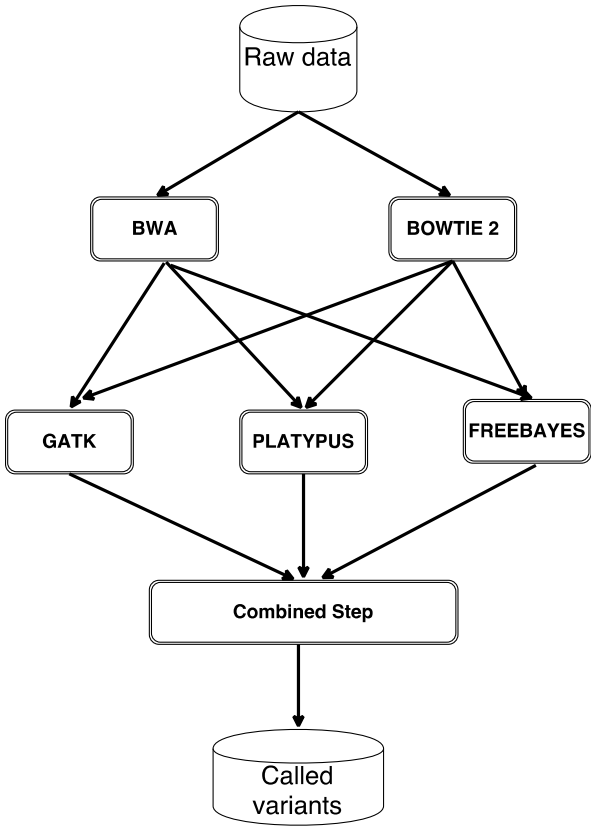

Figure 4. The framework of Genomedics to determine consensus variants from different pipelines [21]

3.4. Genotype imputation

Genotypes at different sites are called independently despite their genetic correlations. The non-random association of alleles at nearby sites known as linkage disequilibrium can be used to impute missing genotypes or improve the accuracy of called genotypes. In other words, alleles of nearby sites are genetically related and usually appear together.

Table 1. An example of genotype imputation

Position	1	2	3	4
Reference haplotypes	A	C	T	A
	C	G	A	C
	C	T	G	A
Observation genotypes	A/C	C/T	?/?	A/A
Prediction	A	C	T	A
	C	T	G	A

Given a list of genotypes at nearby sites in a region, genotype imputation methods use a panel of reference known haplotypes and linkage disequilibrium information between positions to select haplotypes that explain the genotypes at best. The selected haplotypes will be used to impute the missing genotypes. Table 1 shows an example of genotype imputation with three reference haplotypes ACTA, CGAT and CTGA at four nearby sites. If a genome has genotype A/C at the first site, C/T at the second site, unknown genotype at the third site, and A/A at the fourth site, the genome possibly has two haplotypes ACTA and CTGA, thus, the missing genotype at the third site is imputed as T/G.

Several methods have been proposed to impute missing genotypes. The current most popular methods include IMPUTE2 [32], Beagle5 [33], and SHAPEIT2 [34]. A genotype imputation method like IMPUTE2 includes two main steps: phasing step and imputation step. The phasing step infers haplotypes based on reference haplotypes from the population under the study using a Markov Chain Monte Carlo approach. The imputation step uses the hidden Markov model to impute the missing genotypes based on the haplotypes inferred from the phasing step. The phasing and imputation steps are iterated several times to maximize the posterior probabilities of the missing genotypes. Comprehensive experiments on genomic datasets from East Asian populations showed that current genotype imputation methods achieved the accuracy up to around 99% [35]. The genotype imputation methods play an important role in large-scale low or medium coverage human genome projects.

3.5. Structural variant calling

Large-scale human genome projects have revealed a large amount of structural variants [1, 36, 37]. The largest and most important component of structural variants is copy number variants (CNV), i.e., more than 99% of detected structural variants are CNVs.

Two main approaches to determine CNVs are whole genome sequencing approach and microarray-based approach. For whole genome sequencing, the CNVs can be detected by evaluating the decreases (deletions) or increases (amplifications) of short reads in the alignments. The approach is able to call small CNVs, however, they lack the sensitivity with only 10% [38] or 70% [39], and have very high (up to 89%) false positive rates [39, 40].

The CNVs can be also determined by the long reads sequenced by Pacific Biosciences (PacBio) or Oxford Nanopore machines. The long reads with length of up to million bases allow us to detect longer CNVs. However, these technologies have a very high sequencing error rate, i.e., currently around 10% to 15% for PacBio, and 5% to 20% for Oxford Nanopore sequencing machines [41]. The high error rates exceed the capabilities of most aligners and CNV detection algorithms. In short, current sequencing approaches are not reliable to detect

structural variants.

Microarray-based technologies provide the most robust approach for carrying out genome-wide scans to determine CNVs. To detect CNVs from a test genome, the technologies label DNA of the test genome and the reference genome with two different fluorophores of different colors. The labeled DNA from the test and reference genomes are hybridized on an array containing thousands to millions of known probes. A higher or lower intensity of the test genome color in a specific genomic region in comparison to the reference genome indicates a gain or loss of DNA at that region. Microarray-based methods are now a reliable genomic test to identify CNVs for clinical diagnostics [42, 43].

Large scale human genome projects have created a huge amount of structural variants. Several structural variant databases have been established such as Database of Genomic Variants [37] or Human Genomic Structural Variation Database (dbVar) of NCBI [44]. The Database of Genomic Variants (DGV) is a curated collection of more than five hundred thousand CNVs (70% losses and 30% gains) and a few thousand inversions from dozens of studies.

4. VARIANT ANNOTATIONS

A human genome consists of about 4 million variants in comparison with the reference genome. Although most of the variants do not affect health conditions of people, some might change protein structures and functions, consequently affect the health conditions or cause diseases [45]. We discuss two essential annotations for variants: variant allele frequency annotation and variant functional annotation.

4.1. Variant allele frequency annotation

The variant frequency information is essential to classify the clinical effects of variants. Particularly, variants with a frequency of $> 1\%$ are typically considered non-causing disease variants. Thus, the variant frequency information help to filter out non-causing disease variants from further clinical analyses.

Large scale human genome projects have been performed to create large human variation databases. The 1KG project built a database of total 88 million variants from 2504 healthy people [1]. The 1KG database contains almost all common variants with a frequency of 10% in all populations, however, more than 15% of low-frequency variants were only found in a single population. Analyzing variants in the Southeast Asia populations showed that the 1KG project did not have sufficient coverage of the human genetic diversity in this region [46]. Noticeably, a large number of variants that are rare in the global populations but common in a specific population.

Exome Aggregation Consortium combined a large number of genomes and exomes from a variety of large-scale sequencing projects [36]. Specifically, they combined 15,496 genomes and 123,136 exomes collecting from various disease-specific studies to create the gnomAD database. Note that the gnomAD database is only relevant for interpretations of severe pediatric diseases.

It is well known that the variant frequency varies among populations. Vietnam has about 95 million people, however, genetic studies for Vietnamese people mostly rely on genetic information from other populations. We created a Vietnamese single nucleotide

variant database from more than four hundred genomes and exomes of unrelated healthy Vietnamese people [8]. The database consists of more than 24 million variants including more than 99% of common variants with a frequency of greater than 1%. The Vietnamese database will play an important role in genetic studies and medical applications in Vietnam. The database is available at www.genomes.vn and free for academic purposes.

4.2. Variant functional annotation

The genome is separated into different regions, i.e., coding regions, regulatory regions or non-coding regions. The nucleotide variants in the coding or regulatory regions might alter contents, structures, and functions of corresponding proteins. Changing protein functions might cause health conditions or diseases.

The first annotation is the effects of nucleotide variants to corresponding amino acid sequences, i.e., synonymous variants, nonsynonymous variants, start gain/lost variants, stop gain/lost variants, frameshift variants. SnpEff [47] and Annovar [48] are two popular programs for the task. In addition, different methods have been developed to predict functional effects of nucleotide variants on their corresponding protein structures and functions such as SIFT [49], Polyphen2 [50], and MutationTaster [51]. The functional prediction methods evaluate various factors including amino acid changes, the evolutionary constraints of variant sites, protein structural disruptions, and other physics and chemistry characteristics of variants. We note that the Annovar software also integrates the results from different functional prediction methods (i.e., SIFT, Polyphen2, MutationTaster, etc.) in its results.

Finally, the pathological effects of variants can be obtained from general clinical databases such as Clinvar database [52], Human genome mutation database [53] or specific databases for particular diseases such as Sfari database for Autism spectrum disorders [54]. Noticeably, annotations from different databases might be discordant.

5. CONCLUSIONS

The sequencing cost of a human genome is now affordable for genomic studies and applications. Examining the whole exome/genome is becoming a routine genetic test for various health conditions and diseases. Determining and interpreting variants for clinical purposes is a challenging task for both bioinformaticians and genetic consultants.

This paper reviews the state-of-the-art technologies and computational methods to obtain and analyze whole human genomes. As the results from different variant callers might be discordant, we recommend to use at least two variant callers or voting frameworks like Genomedics to determine consensus and reliable variants. We also note that calling variants from multiple genomes, especially from low and medium coverage genomes, should be performed simultaneously to increase the sensitivity and specificity of the analyses.

Annotating and predicting genetic effects of variants are essential for interpreting clinical effects of variants to health conditions. We emphasize that the variant frequency information is a powerful tool to classify pathological effects of a variant. Remarkably, the frequency of a variant varies among populations, therefore, the population-specific variant frequency information is required to properly evaluate pathological effects of variants for the population under study.

Finally, we must be aware that annotations from different clinical databases might be conflicting. Thus, examining annotations from several reliable databases is necessary to confirm the pathological effects of variants. If a variant has conflicting interpretations of pathogenicity, additional analyses/studies must be performed to settle its clinical significance.

REFERENCES

[1] 1000 Genomes Project Consortium et al., "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[2] L.-P. Wong et al., "Deep whole-genome sequencing of 100 southeast Asian Malays," *Am. J. Hum. Genet.*, vol. 92, no. 1, pp. 52–66, Jan. 2013.

[3] R.M. Durbin et al., "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[4] D. I. Boomsma et al., "The Genome of the Netherlands: Design, and project goals," *Eur. J. Hum. Genet.*, vol. 22, no. 2, pp. 221–227, 2014.

[5] W. Zhang et al., "Whole genome sequencing of 35 individuals provides insights into the genetic architecture of Korean population," *BMC Bioinformatics*, vol. 15, no. Suppl 11, pp. S6–S6, Oct. 2014.

[6] T. Lan et al., "Deep whole-genome sequencing of 90 Han Chinese genomes," *Gigascience*, vol. 6, no. 9, pp. 1–7, 2017.

[7] D. T. Hai et al., "Whole genome analysis of a Vietnamese trio," *J. Biosci.*, vol. 40, no. 1, pp. 113–124, 2015.

[8] L. S. Vinh et al., "A Vietnamese Human Genetic Variation Database," *Hum. Mutat.*, p. Submitted, 2019.

[9] L. Tattini, R. D'Aurizio, and A. Magi, "Detection of Genomic Structural Variants from Next-Generation Sequencing Data," *Front. Bioeng. Biotechnol.*, 3:9, 2015.

[10] V. R. Knuth, Donald E and Morris, Jr, James H and Pratt, "Fast pattern matching in strings," *SIAM J. Comput.*, vol. 6, no. 2, pp. 323-350, 1977.

[11] P. Ferragina and G. Manzini, "Indexing compressed text," *J. ACM*, vol. 52, no. 4, pp. 552–581, 2005.

[12] G. Navarro and V. Mäkinen, "Compressed full-text indexes," *ACM Comput. Surv.*, vol. 39, no. 1, Article No. 2, 2007.

[13] P. Ferragina, R. González, G. Navarro, and R. Venturini, "Compressed text indexes," *J. Exp. Algorithmics*, vol. 13, Article No. 12, 2009.

[14] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, pp. 195-197, 1981.

[15] R. Li, Y. Li, K. Kristiansen, and J. Wang, "SOAP: Short oligonucleotide alignment program," *Bioinformatics*, vol. 24, no. 5, pp. 713–714, 2008.

[16] G. Lunter and M. Goodson, "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads," *Genome Res.*, vol. 21, no. 6, pp. 936–939, 2011.

[17] M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno, "SHRiMP2: Sensitive yet practical short read mapping," *Bioinformatics*, vol. 12, no. 7, pp. 1011–1012, 2011.

[18] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.

[19] Langmead B., Salberg S.L., Fast gapped-read aligment with Bowtie2," Nat. Methods, vol. 9 (4), 357-359, 2012.

[20] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[21] V. Le et al., "Genomedics: Whole exome analysis system for clinical studies," in *The 9th International Conference on Knowledge and Systems Engineering (KSE 2017)*, 2017, pp. 142–147.

[22] S. Canzar and S. L. Salzberg, "Short read mapping: An algorithmic tour," in *Proceedings of the IEEE*, vol. 105, no. 3, pp. 436–458, 2017. Doi: 10.1109/JPROC.2015.2455551

[23] H. Li, J. Ruan, and R. Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores," *Genome Res.*, vol. 18, no. 11, pp. 1851–1858, 2008.

[24] M. A. Depristo et al., "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nat. Genet.*, vol. 43, pp. 491–498, 2011.

[25] A. McKenna et al., "The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, vol. 20, no. 9, pp. 1297–1303, 2010.

[26] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, 2011.

[27] A. Rimmer *et al.*, "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications," *Nat. Genet.*, vol. 46, no. 8, pp. 912–918, 2014.

[28] E. Garrison and M. Gabor, "Haplotype-based variant detection from short-read sequencing," *arXiv Prepr.*, p. arXiv:1207.3907, 2012.

[29] Y. Guo, X. Ding, Y. Shen, G. J. Lyon, and K. Wang, "SeqMule: Automated pipeline for analysis of human exome/genome sequencing data," *Sci. Rep.*, vol. 5, p. 14283, 2015.

[30] S. V. Le, P. H. T. Le, T. K. Van Le, T. T. Kieu Huynh, and T. T. Hang Do, "A mutation in GABRB3 associated with Dravet syndrome," *Am. J. Med. Genet. Part A*, vol. 173, no. 8, pp. 2126–2131, 2017.

[31] K. Tran, V. Le, C. Vu, and L. Nguyen, "A novel mutation in LAMA2 unraveled merosin deficient congenital muscular dystrophy type 1A: A case report from a Vietnamese with previously unknown diagnosis," *Submitted*, 2019.

[32] B. N. Howie, P. Donnelly, and J. Marchini, "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet.*, vol. 5, no. 6, p. e1000529, 2009.

[33] B. L. Browning and S. R. Browning, "Genotype Imputation with Millions of Reference Samples," *Am. J. Hum. Genet.*, vol. 98, no. 1, pp. 116–126, 2016.

[34] O. Delaneau, J.-F. Zagury, and J. Marchini, "Improved whole-chromosome phasing for disease and population genetic studies.," *Nat. Methods*, vol. 10, no. 1, pp. 5–6, 2013.

[35] S. Shi *et al.*, "Comprehensive assessment of genotype imputation peormance," *Hum. Hered.*, vol. 83, no. 3, pp. 107–116, 2019.

[36] M. Lek *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2016.

[37] J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer, "The database of genomic variants: A curated collection of structural variation in the human genome," *Nucleic Acids Res.*, vol. 42, no. D1, 2014.

[38] J. Huddleston *et al.*, "Discovery and genotyping of structural variation from long-read haploid genome sequence data," *Genome Res.*, vol. 27, no. 5, pp. 677–685, 2017.

[39] P. H. Sudmant *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, 2015.

[40] S. M. Teo, Y. Pawitan, C. S. Ku, K. S. Chia, and A. Salim, "Statistical challenges associated with detecting copy number variations with next-generation sequencing," *Bioinformatics*, vol. 28, no. 21, pp. 2711–2718, 2012.

[41] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: Ten years of next-generation sequencing technologies," *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, 2016.

[42] B. A. Bejjani and L. G. Shaffer, "Application of array-based comparative genomic hybridization to clinical diagnostics," *J. Mol. Diagnostics*, vol. 8, no. 5, pp. 528–533, 2006.

[43] D. T. Miller *et al.*, "Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies," *Am. J. Hum. Genet.*, vol. 86, no. 5, pp. 749–764, 2010.

[44] I. Lappalainen *et al.*, "DbVar and DGVa: Public archives for genomic structural variation," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D936-41, 2013.

[45] Y. Xue et al., "Deleterious- and disease-allele prevalence in healthy individuals: Insights from current predictions, mutation databases, and population-scale resequencing," *Am. J. Hum. Genet.*, vol. 91, no. 6, pp. 1022–1032, 2012.

[46] D. Lu and S. Xu, "Principal component analysis reveals the 1000 genomes project does not sufficiently cover the human genetic diversity in Asia," *Front. Genet.*, vol. 4, p. 127, 2013.

[47] P. Cingolani et al., "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 (Austin).*, vol. 6, no. 2, pp. 80–92, 2012.

[48] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, 2010.

[49] P. Cingolani et al., "Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift," *Front. Genet.*, vol. 3, p. 35, 2012.

[50] I. A. Adzhubei et al., "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[51] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow, "MutationTaster evaluates disease-causing potential of sequence alterations," *Nat. Methods*, vol. 7, no. 8, pp. 575–576, 2010.

[52] M. J. Landrum *et al.*, "ClinVar: Improving access to variant interpretations and supporting evidence," *Nucleic Acids Res.*, vol. 46, pp. D1062–D1067, 2018.

[53] P. Stenson, M. Mort, E. Ball, K. Shaw, A. Phillips, and D. Cooper, "The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine," *Hum. Genet.*, vol. 133, no. 1, pp. 1–9, 2014.

[54] B. S. Abrahams *et al.*, "SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs)," *Mol. Autism*, vol. 4, no. 1, p. 36, 2013.

Received on May 18, 2019
Revised on May 31, 2019