Computational Approaches for Variant Identification

Diksha Garg
ME
PEC University of Technology,
Chandigarh, India- 160012

Ankita Jiwan
PhD Scholar
PEC University of Technology,
Chandigarh, India- 160012

Shailendra Singh, PhD
Professor
PEC University of Technology,
Chandigarh, India- 160012

ABSTRACT

Variant identification is a fundamental part in the analysis of genetic diseases. Variants are the alterations which occur in the arrangement of nucleotide in the DNA sequence. Genetic diseases are caused by variations occurring in genes which may cause change in protein, affecting the survival and adaptation of an individual. A number of computational techniques are applied to identify these variant. Precise diagnosis of genetic diseases is important for proper treatment of patients and to determine explicit prevention strategies. Introduction of next generation sequencing (NGS) techniques in the past have made large number of DNA sequences easily available. This has made variant identification using NGS data a area of interest. This paper briefly discussed the analysis steps followed for NGS data analysis. This paper later explains in detail a few approaches that are used for identifying variants such as Support vector machine based approach, Machine learning based approach, MOSAIK: hash-base approach, Bayesian statistical based approach, JointSLM based approach.

Keywords

Variants, Variations, Mutations, Genetic Disease, Variant Identification.

1. INTRODUCTION

Every human has a unique DNA sequence and no two individuals are genetically identical. The structure of DNA is double helix which is made up of four nucleotides i.e. adenine (A), cytosine (C) and guanine (G), thymine (T). Base pair is a unit having pair of complementary nucleotides i.e. A-T, G-C. Human consists of approx $3 * 10^9$ base pairs of DNA [1]. The DNA sequence of whole human genome is determined by DNA sequencing. DNA sequencing is a process of determining the arrangement of nucleotides. Any change in the arrangement of nucleotides in the DNA sequence causes alteration in genetic material of an individual which may leads to genetic disorders. Alteration in genetic material may result in abnormalities in the human. Evolution occurs due to multiple genetic variations in humans over time.

Human genetics includes the understanding of gene structure, gene expression, gene mapping, disease association studies and genetic variations which occurs in humans inheritably. Thus genetic variations are the alterations in DNA sequence of humans. Genetic disorder occurs due to the presence of variants in DNA sequence [2]. There are various factors due to which genetic disorder can occur i.e. hereditary, new mutations, and environmental causes. Each genetic disease is caused by a number of variants and thus variant are identification will help in better diagnosis and treatment of the disease. Identification of these variations which cause genetic disease is quite challenging task. There are various ways to identify genetic defects such as hereditary counselling and pre-birth testing (genetic testing) [3]. Genetic testing provides limited information about an inherited condition. It cannot determine the symptoms of disorder from which patient is suffering. Computational techniques were introduced to improve the identification of genes which are responsible for diseases. Various computational techniques used for variant identification are briefly discussed in this paper i.e. JointSLM based approach, support vector machine-based approach, MOSAIK hash-base approach, machine learning based approach, Bayesian statistical based approach. JointSLM based approach is joint distribution approach in which Hidden Markov model is used to detect Copy Number Variants (CNVs). Support vector machine based approach is variant selection strategy in which rare variants are weighted and collapsed representing positive or negative relationship with disease. MOSAIK is a hash-base approach utilizing sequence mapping approach in which short read alignments are used to determine the combination of genotypes. Machine learning based approach is a rule learning approach which uses subset of variant positions to detect the variations. Bayesian statistical based approach applies gene frequency estimation approach which estimate frequency from DNA pool to identify single nucleotide variants (SNVs).

2. VARIATION

Variations are the alterations which happen in the DNA sequence. Genetic variation is the change in the arrangement of nucleotides in the DNA sequence of an individual. In case of asexual organisms, occurrence of genetic variations is due to irregular mutations. Genetic variations in sexual organism occur due to the exchange of chromosome pairs. Genetic variation helps sexual organisms to adapt and survive in different climates and environments.

There are various type of variations observed in humans, briefly discussed in Table 1 which includes continuous variation, discontinuous variation, phylogenetic variation, environmental variation. Continuous variation is the combined effect of various genes and affected by environmental factors i.e. weather, comfort of surroundings etc. Discontinuous variation is controlled by small number of genes and environment has little effect on this type of variation.

Mutations are caused by the changes that occur in genomic sequence. These mutations maybe neutral or harmful or beneficial for humans. Mutations in the coding portion of DNA sequence cause genetic disorder which may alter the amino acid sequence of proteins [4]. Various types of mutations are discussed briefly in Table 2 which includes the brief description of small scale, large scale, germ-line, somatic, gain-of-function, loss-of-function, back or reversion.

Table 1 Difference between continuous and discontinuous variation

Attributes	**Type of Variations**	
	Continuous Variation	**Discontinuous Variation**
Description	Variations occur due to chance of segregation of chromosomes during gamete formation.	Characteristics that are determined by different allele at single locus.
Caused by	Environmental conditions	Diet, culture, lifestyle, climate
Effect of genes	Combined effect of many genes	Controlled by alleles of single gene
Feature measurement	Features can be easily measured across complete range	Features cannot be measured across complete range
Example	Length of hairs, height, shoes size	Hair colour, blood group, eye colour

Small scale mutations affect small genes in few nucleotides. Large scale mutations are mutations which changes from one generation to another which correspond to the repetition of triplets at DNA level (i.e. GCG, GAC). Germ-line mutation exists in parent germ cell which can be passed to future generations. A person having germ-line mutation will suffer variation in all cells of body. Somatic mutation is mutation which occurs in any cell of body spontaneously except germ cells (eggs and sperm) during person's lifetime. Gain of function mutations are active mutations which change gene product stronger in which new allele is formed which express the mutation as dominant phenotypes. Loss of function mutations is inactive mutations which does not change gene product where allele has complete loss of function which expresses that these mutations are recessive. Back or reverse mutations is point mutation (i.e. single nucleotide base substitution, insertion, or deletion of DNA) which restores original sequence by compensating the gene sequence change. Mutations in DNA may cause various changes in the aspects of life of an individual [5]. Mutations are studied on the basis of chromosomal, gene and Phylogenetic based mutations. In chromosomal mutations, genetic information gets mutated during fertilization stage which may cause negative effects. In gene mutations, permanent alteration in the sequence of DNA occurs. In Phylogenetic mutation is the moment of genes from one population to another and affected by environment i.e. if environments are unstable then population that are genetically variable can adapt to changing situations than those which do not contain this variation. If more base pairs of DNA are influenced by mutation then this has a large impact on the individual's lifestyle. Many analysts have started to estimate Dispersion of Mutational Impacts (DMEs) to analyze the effect of mutations in much better way. DMEs evaluate the effect of mutations on a number of variations occurring with its impact on natural system. Sometimes mutations combine with each other which changes the arrangement of base pairs, such phenomenon refers to Epistasis, because of this it becomes hard to get appropriate data about mutations [6].

Mutations are the major factor for the occurrence of genetic disorder. Study of genetic disorders is quite difficult because of different type of genetic disorders. Genetic disorders are divided into single-gene disorders; Multifactorial and polygenic disorders [7]. Single gene disorders include alterations in DNA sequence of single genes which is affected by heredity. Examples of single gene disorders are cystic fibrosis, galactosemia, and Huntington's illness etc. Multifactorial disorders involve alterations which occur in different gene which is affected by environmental conditions. Examples of Multifactorial disorders are Alzheimers disease, breast ovarian cancer, colon cancer etc. Polygenic disorders include the combined activity of alleles (i.e. alternate form of genes) having genetic patterns which are highly complicated i.e. such disorder is controlled by several genes at once. Examples of polygenic disorders are heart disease, diabetes, some cancers etc.

Table 2 Classification of various types of mutations on the basis of its effect on genes [8]

Mutations	**Description**	**Effects**
Small scale	Deletion/ insertion of small number of nucleotides in the chromosome regions.	Loss or gain of small genes in one or few nucleotides.
Large scale	Deletion of large number of nucleotides in the chromosomal regions	Loss of genes, loss of one allele
Germline	Changes in DNA sequence which occur in germ cell.	Genes transmit to next generation cause inherited genetic disease
Somatic	Changes in DNA sequence which occur in any cell except germ cell	Causes cancer or other disease
Gain-of-function	Change in DNA that result in the synthesis of protein with new function.	Enhanced activation in which new allele is created.
Loss-of-function	Gene product with zero or less functionality	The function for which allele encodes is lost
Back/ reversion	Alteration in DNA sequence to restore it back to the original DNA sequence	Altered gene product that act as wild-type allele

3. VARIANT DETECTION

Genetic diseases are caused by variants and to detect these variants is quite difficult task. Some sequencing techniques are used to analyze the significance of variants in the occurrence of genetic disease by sequencing the nucleotides of DNA**.** Whole genome sequencing and whole exome sequencing are sequencing techniques which disclose the significance of the genetic variants [9], [10]. Whole genome sequencing is a process of analysing DNA sequence of an individual at a single time. Whole exome sequencing is a technique of analysing the expressed genes in genome. Another technique used to detect variants is Next Generation Sequencing (NGS) is described in Figure 1 that includes the study of the order of nucleotides of DNA. NGS is a high throughput technique which divides a large DNA sequence into the small DNA fragments and arrange them in parallel [11], [12]. NGS techniques screen the successive expansion of nucleotides of immobilized DNA created from target tissue. NGS interrogates the whole genome to detect entire variations and disease causing genes. Such sequencing technique analyzes the reads for accurate diagnosis of disease [13]. NGS massively parallel technique allows million of reads to run simultaneously. Most reads come out as an output using mate-pair sequencing. Such analysis of reads gives sequential data as a result that can be further used for identifying structural rearrangements of DNA [14]. There are number of NGS platforms such as Roche 454 (introduced in 2005), Illumina (introduced in 2006) and ABI SOLiD (introduced in 2008). DNA sequencing is done by dissecting the signals radiated during the formation of DNA strand. Major difference among these platforms is in the generation of DNA strand [15].

4. ANALYSIS STEPS FOR VARIANT IDENTIFICATION USING NGS

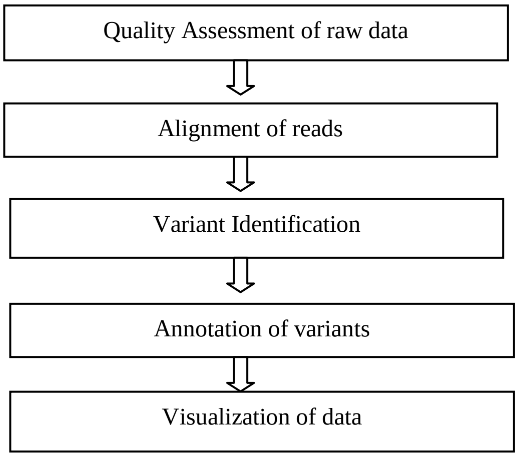

Fig: Workflow of variant calling using NGS

4.1. Quality Assessment of Raw Data

The raw NGS data which is obtained from sequencing platforms have many sequencing faults such as base calling errors, low quality reads, INDELs and adaptor contamination [16]. These sequencing faults affect the quality of sequencing reads which make further analysis inaccurate. To remove such faults, quality assessment is done which analyze the basic quality statistics of the sequence reads and trim or correct the reads. To represent the quality statistics, nucleotide appropriation trimming of reads and various other arrangement properties such as primer contamination, N content and GC bias of particular sequence is obtained [15].FASTQ file have text -based format which can store sequence letter and its quality scores. Tools used for quality assessment are FastQC, NGSQC Toolkit [16] and PRINSEQ [17] etc.

4.2. Alignment of Reads

After obtaining the quality statistics of reads, the sequences are aligned with reference genome. Reference genome refers to digital nucleic acid sequence which is assembled from sequencing of DNA from various donors. NGS sequence reads align with reference genome using fast short read aligner. Various alignment program used to process short reads efficiently are Bowtie/Bowtie2 [18] [19], BWA [20] [21], MAQ [22] , mrFAST [23]etc.

To select the alignment program some issues should be considered as follows:

- Overcome the issue of vagueness where short sequences are aligned with reference genome, where paired-end sequence are demonstrated as significant arrangement which should not be used for entire exome and genome sequencing. Paired-end sequences refer to two ends of the same DNA sequence. [24].
- Sequences which are mapped with criss-crosses sequences should not be considered and mutations which occur on such reads should be rejected [15].
- In PCR (polymerase chain reaction) method, DNA library is generated through random fragmentation of genomic DNA. When NGS technologies fuse with PCR steps during library arrangements, then multiple sequences from each template may be sequenced which complicate the identification of variants. So, it is necessary to delete the PCR copies after alignment in whole-genome and whole-exome sequence [15].

4.3. Variant Identification

Variant identification is very critical part in the analysis of NGS data. Disease gene identification deals with the study of identifying variants which are responsible for genetic disorder [25]. In the process of identification of variants, the detection of changes which are observed on many reads is necessary [26]. Variant identification techniques used to analyze the somatic mutations on the basis of raw data used [27]. Tools which are used for genome-wide variant identification consists of four classifications i.e. germline callers- CRISP [28], GATK [29], SAMtools [30], somatic callers- SomaticSniper [31], CNV identification- CNVnator [32], CONTRA [33] , SV identification- BreakDancer [34] and Breakpointer [35].

4.4. Annotation of Variant

Variant annotation enables to prioritize the disease causing variants. NGS platforms produce large data and analyzing the functional impact of variants is very crucial. Various tools such as ANNOVAR [36], NGS-SNP [37], SVA [38], VARIANT [39], etc. are used for variant annotation which determines the significance of the detected variants in the sequence. VARIANT can distinguish among the basic properties of SNVs in coding and non-coding regions. ANNOVAR, NGS- SNP, SVA tools are used to analyze the structural impact on proteins which are based on region-based analysis instead of sequence based analysis

4.5. Visualization of Data

Data validation and visualization is an important step in the process of variant identification. Visualization of distinguished CNVs and Structural Variants (SVs) represents the global image of genomic arrangements and also capable of representing the information about reference genome, transcriptome and aligned reads [40]. NGS visualization tools used to represent aligned sequence, quality of mapping and detect variations/ mutations with its annotations [15].Visualization tools which are used for NGS data are divided into three groups i.e. tools which support the knowledge about sequence information of de-novo or resequencing tests, genome browser that permit user to peruse the mapped data with various sort of annotations, comparative viewers which enhance the correlation of sequence from numerous people [41]. Genome browser is capable of displaying numerous 1D tracks. Two types of genome browser are web-based genome browser tool such as The Ensembl Genome Browser [42] and University of Santa Cruz (UCSC) Genome Browser [43]; and stand-alone genome browser tool such as Artemis [44], Integrative Genomics Viewer (IGV) [45].

5. COMPUTATIONAL APPROACHES FOR IDENTIFICATION OF VARIANTS OF GENETIC DISEASE

Accurate diagnosis of genetic disease can be performed by using computational approaches which detect various classes of variations such as Rare Variants (RVs), Somatic Variants, germ-line variants, SNPs, INDELs, Multi Nucleotide Polymorphism (MNPs). Various computational approaches relevant for next generation sequencing data and variant calling are discussed as follows:

5.1. JointSLM Based Approach

JointSLM approach is the joint distribution of equation that defines Hidden Markov Model (HMM) of order one [61]. This approach uses emission probability distribution f_k:

$$f_k(x) = \prod_{t=1}^{M} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\left(\frac{x-u}{\sigma}\right) \hat{}2\right]$$

Where u is associated to each state of markovian stochastic process; σ is the variance and x is that whose conditional probability has to be calculated.

In this approach global parameters (related to the similar family) of sequential process are analysed by using two step algorithm i.e. Baum-Welch algorithm and Viterbi algorithm used to find unknown parameters of HMM. This approach has ability to detect common shifts of various sizes and CNVs. The copy number of each DNA segment is analyzed by calculating the median of each segment. After calculating median of each segment, DNA copy number is estimated by rounding the median of each segment to nearest integer [62]. After estimating DNA copy number, comparison is done by joint model with known CNVs [63].

JointSLM approach analyzes its performance by Area Under Curve (AUC) which provides the direct way to diagnostic decision making process.

5.2. Support Vector Machine Based Approach

Backward Support Vector Machine (BSVM) use variant selection strategy which analyzes rare variants that causes the disease. The approach detects the occurrence of the rare variants while weighting all variants. In the selection method, rare variants are weighted and are collapsed on the basis of its positive or negative relationship with disease. Two major steps used in this approach are (a) giving way and pooling of various rare single nucleotide variants together. (b) using appropriate weight plan to enhance pooled RVs. Various RVs are pooled together having high recurrence and impact are determined which is connected with quality scores in different directions which reduces the measurable power [46]. All rare variants are classified on the basis of measuring power which is based on frequency. Variants due to which disease occurs are consider as up-weighted and variants which have less impact on disease consider as down-weighted [46]. These down-weighted rare variants which have unbiased impacts are expelled backwardly from the entire list to hold only informative rare variants.

Backward variable selection method detects rare variants on the basis of value of R^2 (coefficient of determination) which is estimation of goodness for variant selection calculated using following equation [47] :

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(t_k - \bar{y}_k)\hat{}2}{\sum_{k=1}^{n}(t_k - \bar{t})\hat{}2}$$

Where $\overline{y_k}$ is estimated y_k where y_k is hyperplane function which linear transformation function $\bar{t}$ is the average of t where t is estimated by y_k to identify factors associated with disease, and [48]. Hypothesis analysis is used for variant selection procedure [49], [50]. This methodology has been applied to Type 1 Diabetes Mellitus (TIDM) dataset [51].

5.3. MOSAIK: A Hash-Base Approach

MOSAIK is an open source program used for mapping sequence reads to a reference genome. MOSAIK uses hash clustering algorithm coupled with Smith-Waterman algorithm. In hash clustering approach, sequence reads are divided into sets of overlapping hashes and position of genes, where each hash can be queried from stored reference hash table [55]. Adelson-Velskii and Landis (AVL) tree [56], a self-balancing tree is used to cluster hash positions to form hash region. The hash clustering algorithm determines sequencing errors, SNPs and single base INDELs. After determining sequencing errors, alignment candidate region in AVL tree is determined. Alignment candidate region is that region in AVL tree which consolidate hashes. After identifying candidate region MOSAIK use Smith-Waterman algorithm to align sequence reads to candidate regions. This algorithm is used for aligning gapped sequence by analyzing all possible frames of alignment [57]. Smith-Waterman algorithm improves the performance by decreasing the runtime for aligning NGS data.

On the basis of MOSAIK alignments, SNPs and INDELs calls are generated. FREEBAYES, Bayesian variant calling software [58] is used to identify SNPs, INDELs, MNPs etc. It uses short read alignments to determine the combination of genotypes at each position in reference genome. This software uses set of variants as a source of prior information and copy number variant map as an input to analyze the variation.

5.4. Machine Learning Based Approach

Machine learning is rule learning approach which is also used to identify variants of genetic disease. SNOOPER is an example of machine learning approach which uses a Leo Breiman RF classifier [52] is analyse the training information of substantial datasets [52]. This approach is categorized into three stages to identify the variants of genetic disease i.e. feature extraction; preparation stage and calling stage. The approach uses subset of variant positions as an input, obtained from the output of DNA sequencing which helps in detection of variations or sequencing error. For every variant position, a set of features are used for the identification of somatic variations. Feature extraction is performed on each somatic variant which is passed in quality filters. After extraction of features, components are standardized which are used for comparing the median value obtained from the subsets of variants [53]. Components are positioned and chosen by measuring Information Gain (IG) [54]. In preparing stage, detected variants are separated into two classes: false positive class (errors) and genuine positive class (variations). In calling stage, Fisher's exact test is used to analyse the sequence reads having reference and option allele between ordinary and disease related samples. SNOOPER has both ordinary and disease related files having two extra filters with BED format documents with germ-line (somatic) dataset. In calling stage, new disease related model and coordinated ordinary mpileup documents (document describing the base-pair information at each chromosomal position) are used to identify somatic variants. Output of calling stage a VCF (Virtual Contact File) document having somatic p-value calculated using the approach [53].

5.5. Bayesian Statistical Based Approach

Bayesian model is a graphical model having various nodes which signifies different parameters. Model parameters are u_0, a global non-reference read rate which captures error rate across all position; M_0, global precision which extracts false rate of variation across position in sequence; M_j, local precision that extracts the false rate of variation at j position across different replicates. Bayesian statistical model is used to estimate Non-Reference Allele Frequency (NRAF) i.e estimating the gene frequency from DNA pool and identify SNVs [59]. In such approach, variational-Expectation Maximization (EM) algorithm is used to detect the rare SNVs. The performance of variational EM algorithm is represented by Receiver-Operating Characteristic curves (ROCs) for NRAFs. ROC provides graphical representation to diagnose decision making process. In this algorithm a non-conjugate variational inference algorithm is developed to approximate posterior distribution using following equation i.e.

$$p(u,\theta|r,n;\emptyset) = \frac{p(r,u,\theta|n;\emptyset)}{p(r|n;\emptyset)}$$

where p is the posterior distribution, r is the number of reads with non-reference base in experimental replicate position; u and θ are latent variables; n is the total number of reads at experimental replicate location and $\emptyset$ is maximization parameter. Variational EM algorithm has been applied to MTH1 gene (negative regulator of glucose-sensing signal transduction) to detect the mutations [60].

6. CONCLUSION

The development of next generation sequencing technique has made available large volumes of raw DNA sequence data. This has revolutionized the variant identification process. The analysis process of NGS data is complex and consumes tremendous amount of effort and time. But NGS data is more efficient than any other sequencing data. Further development for variant identification of genetic disease requires extensive knowledge about genome variations. Advancements in sequencing technology and analytical tools are also necessary for accurate diagnosis of genetic disease. A number of tools are used in next-generation sequencing analysis and some of them have been listed. Various computational approaches are available to detect variants from NGS data and some of them are briefly explained in the paper. Such methods use different strategies such as feature extraction, variant selection and calling of variants are performed to determine the existence of disease at early stage. These procedures can detect various size, classes, position and frequency estimation of variations. These approaches help pathologists to identify different classes of variants. The power of NGS technology can be applied in various research areas including species identification, MZ twins study etc. It can be further help in forensic where DNA samples are limited. Most variant identification approaches focus on detecting one or two classes of variations. New and improved technique need to be developed which can detect all types of mutations simultaneously. Improvements can be made in genome amplification in which genome is amplified from nanogram quantity of DNA to microgram quantity. There is a need to develop advance tools that can collect phenotypic characteristics from image data.

7. REFERENCES

[1] Lodish, H., Baltimore, D., Berk, A., Zipursky, S. L., Matsudaira, P. and Darnell, A. 1995. Molecular cell biology. New York: Scientific American Books.

[2] Milunsky, A. and Milunsky, J. 2015. Genetic disorders and the fetus: diagnosis, prevention, and treatment. John Wiley & Sons.

[3] Renkema, K. Y., Stokman, M. F., Giles, R. and Knoers, A. 2014. Next-generation sequencing for research and diagnostics in kidney disease. Nature Reviews Nephrology. 433-444.

[4] Mahdieh, N. and Rabbani, A. 2013. An overview of mutation detection methods in genetic disorders. Iranian journal of pediatrics. 23(4), 375.

[5] Drake, J. W., Charlesworth, B., Charlesworth, J. D. and Crow. 1998. Rates of spontaneous mutation. Genetics. 1667-1686.

[6] Eyre-Walker, A. and Keightley, A. 2007. The distribution of fitness effects of new mutations. Nature Reviews Genetics. 610-618.

[7] Wei, X., Ju, X., Yi, X., Zhu, Q., Qu, N., Liu, T., Chen, Y., Jiang, H., Yang, G. and Zhen, R., 2011. Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing. PloS one. 6(12).

[8] Mili, A., Charfeddine, I. B., Mamaï, O., Cherif, W., Adala, L., Amara, A., Pagliarani, S., Lucchiari, S., Ayadi, A. and Tebib, N.2012. Molecular and biochemical characterization of Tunisian patients with glycogen storage disease type III. Journal of human genetics. 170-175.

[9] Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W. and Nickerson, D. 2010. Exome sequencing identifies the cause of a mendelian disorder. Nature genetics. 42(1), 30-35.

[10] Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., Beck, A. E., Tabor, H. K., Cooper, G. M. and Mefford, H. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics. 42(9), 790-793.

[11] Metzker and M. L. 2010. Sequencing technologies—the next generation. Nature reviews genetics. 11(1), 31-46.

[12] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, A. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437(7057), 376-380.

[13] Li, H. and Homer, A. A survey of sequence alignment algorithms for next-generation sequencing. 2010. Briefings in bioinformatics. 11(5), 473-483.

[14] Medvedev, P., Stanciu, M. and Brudno, A. 2009. Computational methods for discovering structural variation with next-generation sequencing. Nature methods. 6, (S13-S20).

[15] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R. and Zschocke, J. 2014. A survey of tools for variant analysis of next-generation genome sequencing data.. Briefings in bioinformatics. 15(2), 256-278.

[16] Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., Omenn, G. and Meng, A. 2010. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC genomics. 11(4), 7.

[17] Schmieder, R., Edwards, A. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27(6), 863-864.

[18] Langmead, B., Trapnell, C., Pop, M. and Salzberg, A. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 10(3), 25.

[19] Langmead, B. and Salzberg, A. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods. 9(4), 357-359.

[20] Li, H. and Durbin, A. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25(14), 1754-1760.

[21] Li, H. and Durbin, A. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform Bioinformatics. 26(5), 589-595.

[22] Li, H., Ruan, J. and Durbin, A. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research. 18(11), 1851-1858.

[23] Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O. and Sahinalp, A. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nature genetics. 41(10), 1061-1067.

[24] Lee, H. and Schatz, A. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics. 28(16), 2097-2105.

[25] Camilleri, M., Carlson, P., McKinzie, S., Grudell, A., Busciglio, I., Burton, D., Baxter, K., Ryks, M. and Zinsmeister, A. 2008. Genetic variation in endocannabinoid metabolism, gastrointestinal motility, and sensation. American Journal of Physiology-Gastrointestinal and Liver Physiology. 294(1), G13-G19.

[26] Neuman, J. A., Isakov, O. a nd Shomron, A. 2013. Analysis of insertion–deletion from deep-sequencing data: software evaluation for optimal detection. Briefings in Bioinformatics. 14(1), 46-55.

[27] Kim, S. Y., Li, Y., Guo, Y., Li, R., Holmkvist, J., Hansen, T., Pedersen, O., Wang, J. and Nielsen, A. 2010. Design of association studies with pooled or un-pooled next-generation sequencing data. Genetic epidemiology. 34(5), 479-491.

[28] B. V. 2010. A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics. 26, (318-24).

[29] DePristo, M A. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43, (491-8).

[30] Li, H. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25, (2078-9).

[31] Larson, D E. 2012. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 28, (311-7).

[32] Abyzov, A. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21, (974-84).

[33] Li, J. 2012. CONTRA: copy number analysis for targeted resequencing. Bioinformatics. 28, (1307-13).

[34] Chen, K. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods. 6, (677-81).

[35] Sun, R. 2012. Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics. 28, (1024-5).

[36] Wang, K. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, 164.

[37] Grant, J R. 2011. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics. 27 , (2300-1).

[38] Ge, D. 2011. SVA: software for annotating and visualizing sequenced human genomes. Bioinformatics. 27, (1998-2000).

[39] Medina, I. 2012. VARIANT: command line, web service and web interface for fast and accurate functional characterization of variants found by next-generation sequencing. Nucleic Acids Res. 40, (54-8).

[40] Loraine, A E. 2002. Visualizing the genome: techniques for presenting human genome data and annotations. BMC Bioinformatics. 3 , 19.

[41] Nielsen, R., Paul, J. S., Albrechtsen, A. and Song, A. 2011. Genotype and SNP calling from next-generation sequencing data.. Nature Reviews Genetics. 12(6), 443-451.

[42] Spudich, G M. 2010. Touring Ensembl: a practical guide to genome browsing. BMC Genomics. 11 , 295.

[43] Dreszer, T R. 2012. The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res. 40, (918-23).

[44] Carver, T. 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 28, (464-9).

[45] Thorvaldsdóttir, H. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinformatics. 14, (178-92).

[46] Fang, Y. and Chiu, A. 2013. A novel support vector machine-based approach for rare variant detection. PloS one. 8(8), 71114.

[47] Tax, D. and Duin, A. 2004. Support vector data description. Machine learning. 54(1), 45-66.

[48] Spiess, A. and Neumeyer, A. 2010. An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC pharmacology. 10(1), 6.

[49] Li, L., Jiang, W., Li, X., Moser, K. L., Guo, Z., Du, L., Wang, Q., Topol, E. J., Wang, Q. and Rao, A. 2005. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics. 85(1), 16-23.

[50] Malhotra, R. and Chug, A. 2012. Software maintainability prediction using machine learning algorithms. Software Engineering: An International Journal (SEIJ). 19-36.

[51] Wu, C., Walsh, K. M., DeWan, A. T., Hoh, J. and Wang, A. 2011,November. Disease risk prediction with rare and common variants. BMC proceedings. 5, (S61).

[52] Breiman, L. 2001. Random forests. Machine learning. 45(1), 5-32.

[53] Spinella, J. F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., Ouimet, M., Healy, J. and Sinnett, A. 2016. SNooPer: a machine learning-based method for somatic variant identification from low-pass next-generation sequencing. BMC genomics. 17(1), 912.

[54] Kullback, S. 1959. Information theory and statistics. New York: wiley.

[55] Lee, G. 2014. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PloS one. 9(3), 90581.

[56] Vel'skii, A. 1962. An algorithm for the organization of information. Sov Math Dok. 3, (263–266).

[57] Smith, W.M.1981. Indentification of common molecular subsequences. J Mol Biol. 147, (195–197).

[58] Garrison, G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv. 9.

[59] Zhang, F. P. 2016. Variational inference for rare variant detection in deep, heterogeneous next-generation sequencing data. arXiv preprint arXiv. 1604, 04280.

[60] Kvitek, S. G. 2013. Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment.. PLoS Genet. 9(11), 1003972.

[61] Magi, T. F. 2011. Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. Nucleic acids research. 068.

[62] Yoon, S. J. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome. 19, 1586-1592.

[63] McCarroll, M. K. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 40, 1166-1174.