ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

A Case Study of NGS Tools in Bioinformatics

Abhirami T, Adithi P, Hemalatha N

Dept of Bioinformatics, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India

ABSTRACT: New era of GENOME sequencing technology is developing rapidly day by day. It gives ample opportunity to study genomic landscapes as well as diagnosis for mutation therapy. Mainly the area like genome sequencing which uses NGS technology plays a main role in human genetics due to the minimal cost and is user friendly. Whole genome sequencing is developing many data being collected and this directly results in development of many tools. We did our survey based on five steps to check the quality, alignment, identification of variant, annotation of variant and visualization. We made a brief report on functionality, property and use of tools. In our work we have studied papers which used 15 programs dealing with variant identification, variant annotation and visualization. Further evaluation of these was subjected using data sets of patients with germ line mutations and cancer patients with somatic mutations. Finally with this whole survey we could conclude that NGS is the best tool to study human genetics of Mendelian disorders, complex diseases, cancers and mutated genes.

KEYWORDS*:* NGS, Genome sequencing.

I. INTRODUCTION

There is rapid growth in human genetics and research due to the advanced genome sequence technology. Now a days NGS is also being used in small laboratories. Complex diseases are easily found through methods of genome sequencing [1]. Whole genome sequencing technique is expensive; still this method is used to bridge the gap in between cost control and genome-wide comprehensiveness. Lot of technologies are being developed and in future there might be a tool which can sequence up to 10 G base pairs.

Whole exome sequencing is widely used to find genetic disorders and genetic diseases. It is also used to identify the coding regions present in data bases [2]. Whole genome sequencing is used to identify mendelian disorder which in turn help us to improve the functionality of human genomics [3]. Complex diseases and cancer are other interested fields in human genetics (NGS). NGS is also used in studies related to mutated genes and these studies shows that somatic mutations were first detected initially using NGS [4]. The main concept of whole genome and exome sequencing not only deals with sequencing of DNA but it also relays how the data management is structured and analysis of the computational experiments [5]. To obtain valid biological results each and every steps are carefully observed and appropriate tools are used. If we see the whole NGS process, it is complex, but it has multiple methods to handle enormous amounts of heterogeneous data. The NGS projects became successful and numerous tool were created to analyze the specific parts. Many articles were published which dealt with tools for particular application. The collected articles review about NGS data analysis like mapping, assembly and alignment, SNP algorithms of the sequence [6]. But the reviews related to analysis of individual steps have not been done. This type of reviews will be help the researches who are planning to do NGS. There is a chance that of missing data handling and tool compatibility when individual components are reviewed. Also in our paper, we had covered with a complete overview of typical human genetics using the tools which were used in mendelian disorder and cancer data set.

We did our work based on five steps to check the quality, alignment, identification of variant, annotation of variant and visualization. A brief report on functionality, property and uses of tools were also done. We have covered review papers which dealt with programs dealing with variant identification, variant annotation and visualization. To further evaluate data sets of patients with germ line mutations and cancer patients with somatic mutations were also considered.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

II. NGS IN HUMAN GENETICS

In this project we mainly took 3 areas to find the role of NGS in human genetics:
(a) gene that is causing mendelian disorders
(b) the role candidate genes in complex diseases
(c) genes involved in cancer and mutated genes

A. Mendelian Disorders

The main role of mendelian inheritance is to study the heriditory components. The aim to study mendelian disorder is based on positional cloning and linkage analysis [7]. The further modification of resulting genes is done by sanger technique. This is mainly used to study the family diseases and it doesn't hold good for *de novo* dominant mutations. Whole exome sequencing is a tool which helps identification of autosomal recessive disease genes in single as well as de novo dominant mutations [8]. It also identifies large number of variants. Whole genome sequence gives all the details about human genome which includes mutations and regulatory elements accompanied with hereditary diseases.

B. Complex diseases

The candidate genes have helped us to know more about the genetics of complex phenotypes. It deals with pathophysiological considerations and this leads to the formation of phenotypes[9]. Later this method was discarded because of applying wrong statistical assumptions. Another to study candidate gene is related to genome wide associated studies. The main aim of the studies is linkage disequilibrium, the non-random association between alleles at different loci—at the population level. Study of complex phenotypes deals with common disease – common variant or common disease – rare variant hypotheses. GWAS tests the common disease – common variant hypothesis. This hypothesis says that multiple rare variants with large effect of heritability of the disease [34]. The field provides the study about frequency variations and which empowered by NGS and bioinformatics [10]. NGS can perform on complex diseases by the following (i) whole-genome, (ii) whole-exome and (iii) targeted subgenomic sequencing. These have been successfully used for identifying complex hereditary diseases [11], it also helps to testing both mentioned hypothesis. By re-sequencing candidate genes in a large number of patients can also identify trait loci and controls as demonstrated for Type 1 diabetes [12]. This will be a great method for whole exome sequencing in the near future and also helps the discovery of novel genes. Given the vast number of genetic and non-genetic etiological factors of complex diseases, the ultimate approach will require exploiting biological and clinical data, and integration of additional data sets including RNA sequencing data, proteomics data and metabolomics data.

C. Somatic mutations

There is a difference between constitutional and somatic mutation. Constitutional mutation that tells about mutation that are inherited from parents and are present all over the body cell. It also increase the sensitiveness of an individual to diagnosed with cancer [13]. Now new methods have been identified to predict genetic components which help in identifying cancer and also multiple locations of the cancer. It also identifies common alleles which help in predicting heritability of cancer [14]. These help in finding a small portion of risk of cancer.

Mainly 22 breast cancer loci have been reported out of that only 8% is heritability. According to hypothesis of common disease rare variant is high penetrance in the mutation may cause high risk of developing cancer. The whole exome sequencing has been identifying mutations i.e high-penetrance mutations in genes which are used only for small families [15]. Somatic mutation tool is used in identifying and detecting the disease. The main intention to give a drug for the target is for their genetic makeup. Example: imatinib an inhibitor of tyrosine kinase has a major therapeutic advance in patients [16]. Overall there is a huge demand for methods for detecting biomarkers.

III. NGS VARIANT ANALYSIS WORKFLOW

A. NGS platforms

Instruments of NGS give the output at great speed by finding millions of short DNA fragment [17]. At present three platforms are commonly used: Roche 454 (introduced in 2005), Illumina (launched in 2006) and ABI SOLiD (followed in 2008). All these platform sequence the DNA by analysing the signals that are released during second DNA strand

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

,and it depends on how the strands are generated. .In order to find the signals, the template DNA is fragmented into finite pieces, is amplified and immobilized on a slide before sequencing.
Roche implements pyrosequencing and measures the pyrophosphate released that help in analysing the fragments of hundred base pairs. The technique identifies number of nucleotide that is incorporated from signals. This mainly creates the problem when the homopolymer is stretched long up to 8bp long [18]. This will create a problem in identifying and deleting of small region. ABI SOLiD identifies DNA through ligating fluorescence di-base probes to the first strand if required multiple times. By the help of library preparation and sequencing technology it has been possible to read chromosomal distance. The paired-end reads the information that are required to enhance mapping accuracy and rearranging the structure [18].
Due to the nature of this approach, identified calls are not stored in nucleotide but in color space—a property that needs to be considered in downstream analyses. Depending on library preparation and sequencing technology, it is possible to sequence reads that are of a known chromosomal distance . These so-called paired-end or mate-pair reads provide additional information which can be used for enhancing mapping accuracy and identifying structural rearrangements
After the completion of laboratory work and sequencing, the researcher has come out with a huge amount of raw data. It mainly has five steps

(i) Checking the quality of raw data
(ii) read alignment for the reference genome,
(iii) identification of variant ,
(iv) annotation of the variants and
(v) visualization of data.

IV. QUALITY ASSESSMENT

After completing the sequencing, then the first step is to remove the repeats and make them correct. The sequencing methods produces the raw data and which may contain some common errors such as base calling errors, INDELs, poor quality reeds etc [19]. These can be fixed by performing filtering and trimming tasks. It includes visualization process of base quality scores and nucleotide distributions, N content and GC bias. There are different tools that are available to perform quality assessment. FASTQ and 454 files can be handled by stand-alone tools like NGSQC TOOLKIT and PRINSEQ [20] which produce summary reports. We can access almost all sequencing platforms, output summary graphs and tables by FASTQ. NGS data cross-sample contamination can be estimated by the tool ContEst [21]. Other tools are htSeqTools, SolexaQA, PIQA and TileQC.

A. Alignment

After the reads are processed to meet a standard, then they are aligned to an existing genome reference. At present there are two main process for genome assembly.
i) The University of Santa Cruz (UCSC) which is conducting the central repository to ENCODE data.
ii) The Genome Reference Consortium (GRC) which focuses on creating reference assemblies.
These two provide several versions of the human genome. UCSC offers versions hg18 and hg19 while GRC provides GRCh36 and GRCh37. All together these two are the most widely used reference genomes [22]. UCSC (hg)and GRC (GRCh) human assemblies are similar but only differ in their nomenclature.
Many alignment programs has been developed to process the short reads like Bowtie, BWA,MAQ, mrFAST , Novoalign SOAP , SSAHA2 , Stampy and YOABS. The sequencing technologies are pushing the lengths of generated reads. First-generation short-read aligners were found to optimize ungapped alignment. Nowadays programs deal with finding longer gap and longer read length. The long read alignment algorithm can be classified using hash table indexing like BALT. Most alignment algorithms mainly follow the seed-and-extend paradigm, where one or more seeds are searched which is followed by an extension to cover the whole query sequence which is added to the selection of the alignment program [23]. Firstly, to overcome the problem of ambivalence when mapping short reads to a genome reference, paired-end reads are highly recommended, Secondly, reads that mapped with many mismatches are not considered and mutations that are caused by such reads are removed for further analysis. And thirdly, the present NGS uses PCR steps to library preparation as current NGS technologies incorporate only multiple reads originated from one template. That is why they remove multiple copies of PCR after alignment in whole-genome sequencing.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

B. Variant identification

Identification of data variants plays an important role in NGS. Somatic mutation identification one of the applied strategy for genotype calling and is related to data usage. Mutations are an important factor in variant identification, and is supported by reads [24]. These tools can be grouped into four (i) germline callers, (ii) somatic callers, (iii) CNV identification and (iv) SV identification. Detection of mutations plays a key role in finding rare diseases and cancer studies. Structural modification identification tool divided into two can find CNVs and other SVs as inversions, translocations or large INDELs. In both whole genome and whole exome sequencing studies, they are detected by CNVs which are the only SVs using now. Therefore they can capture exome properties [25].

C. Variant annotation

The arrangement and storage of data produced by NGS experiments is becoming increasingly important. The different annotation tools are mainly focusing on the annotation of SNPs, and they can be easily identified [26]. There are some tools which covers INDELs and also CNVs are annotating structural variants. dbSNP is one of the common annotation method. Based on protein the tools available in different approaches, ranging simple sequence - based analysis over region based analysis and the functional analysis result classified into accepted and deleterious mutations. Many tools are accompanied with web applications, so no need to install and maintain a local copy. Web applications are easy to use and it is self-explanatory [27]. The disadvantage of this web applications are they do not support batch submission of variants which make them only viable for manual analysis and the data confidentiality arises some legal issues also. More over offline tools provides more flexibility.

D. Visualization

The most important and prominent step in NGS data analysis workflow is the validation and visualization of the results [28]. Visualization representation of data is tremendously helpful in obtaining results. Hence, NGS visualization tools must display aligned reads, mapping quality and identifiying mutations combined with annotations from various public resources. Visualization tools for genomic data are divided into three different types:.

(i) finishing tools that support interpretation of sequence data of de novo experiments.

(ii) genome browsers help the users to browse mapped experimental data with respect to different types of annotation and

(iii) comparative viewer which help in comparing sequences from multiple organisms.

Based on genome browsers, software suites have been published that help in identifiying CNVs and SVs. Genome browser has be divided into two types :

i) web-based application that runs on web server and

ii) stand-alone tool was used as GUI[29].

These are implemented in java and can be used in platforms like Windows, Mac and Linux system. The use of web-based genome browser is that it supports variety of annotation. Here the user can browse different genomic annotation derived from different databases.in future and there is no need of installing new application. Drawback of genome browser is the necessity of uploading data to a remote server which help in legal issues. Stand-alone genome browsers allows interactive browsing and zooming features, that might not been seen in some web-based genome browsers [30]. Therefore uploading the data to website is not required. Shortcomings are needed to download annotation files and to keep annotation up-to-date.

E. Analytical pipelines and workflow systems

The review shows that the tool for NGS analysis has been accessed by the scientific community. The method for analysing biolological result is a challenging task. Alternative steps are used for complete analysis of raw sequences [31]. The analytical pipelines generally have a predefined order of analysis steps and it has built-in algorithms which help in modifying GUI and allows the user to modify complex workflow with little program.

V. TOOLS USED FOR EVALUATION

Three data analysis steps were done for evaluation (i) variant identification, (ii) variant annotation and (iii) visualization.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

A. Variant identification

The evaluated tools are divided into four groups,
germline callers(five tools), somatic callers (four tools), CNV identification tools (four tools) and SV identification tools (five tools).
Germline callers : For detecting common and rare variants we use CRISP and INDELs, it mainly identify SNPs and pooled NGS data [35]. 'GATK' is another software library which provides a suite of tools for working with human data, including depth of coverage analyzers, a quality score recalibrator, a local realigner and a SNP/INDEL caller. It can be used for identifying somatic mutation also. 'SAMtools' is a collection of tools for manipulating SAM and BAM files. 'SNVer' is an operating system independent of statistical tool for the identification of SNPs as well as INDELs, in both pooled and individual NGS data.
Somatic callers: A command-line application called Somatic Sniper is used to identify SNPs [35]. It is mainly used to find somatic score which correctly tells how normal genotype and tumor are different .Other tools like (SAM tools, SomaticSniper) were checked using whole-exome tumor data set.
Copy number variations identification: Copy number variations identification was done using 4 tools of whole exome data.
a. CNVnator: Tool used to identify whole-genome sequencing which does complete analysis related to mean shift.
b. CONTRA: Tool mainly calls copy gains and goes on degrading for spicing target.
c. ExomeCNV: Tool tells about degradation of heterozygosity.
d. RDXplorer: Tools to identify whole-genome sequencing data based on humans.
Structural variants identification: Inserting, deleting, converting, inter- and intra-chromosomal translocations, these types of structural variants are used in identifying Break Dancer. A command line tool called Break pointer is used to identify the breakpoints of intrachromosomal sequences. Commonly heuristic method is used mainly to locate potential intrachromosomal sequence breakpoints caused due to single end reads.
CLEVER is a tool which identifies SVs present in genomes from paired-end sequencing reads. Insert sizebased approach is used, which allows to take all reads into account. The tool has an intuitive script with default parameters which facilitate usability. The probabilistic verstion of GASV algorithm is represented buy GASVPro and it also detect SVs. SVMerge [112] is a software tool which combines results from different SV callers and performs subsequent validation and refinement of identified breakpoints. This software suite does not provided a ready-to-use virtual box or implementation of cloud, which enhances the usage.

B. Variant annotation

'ANNOVAR' is a command-line tool for up to date functional annotation of various genomes, supporting SNPs, INDELs, block substitutions as well as CNVs [36]. The tool provides a wide variety of different annotation techniques, organized in the categories gene-based, region-based and filter-based annotation. 'AnnTools' is a command-line tool for analyzing SNPs, INDELs and CNVs found in both coding and non-coding regions [36]. The program relies on 15 different widely used data sources such as dbSNP, which are regularly updated. A database update tool is provided to help keep the local database up-to-date. 'NGS–SNP' [36] is a collection of Perl scripts for the annotation of SNVs using the Ensembl database as a reference. The program uses the online version of the Ensembl database, which has the advantage that the reference database is always up-to-date. In comparison to other tools, NGS–SNP took several days to complete the annotation process, which is likely due to the latency of querying the online database during the tool's execution. 'Sequence variant analyzer (SVA)' [36] is a stand-alone tool with a GUI dedicated to annotating and visualizing variants identified by NGS experiments. The tool includes its own genome browser and supports annotation of SNPs, INDELs and CNVs. 'VARIANT' [37] can detect the functional properties of SNVs in coding as well as non-coding regions. The program can be used via a web interface, as a command-line tool or as a web service. Since the command-line tool also makes use of the remote VARIANT database, the tasks of maintaining and searching of databases are provided by the authors. Therefore, the command-line version of the tool is usable without profound IT expertise and can be executed on regular office PCs. The web interface features anonymous usage and allows creation of personal accounts, which enables users to view their uploaded input files and analysis results once they log in. 'Variant effect predictor (VEP)' [37] is Ensembl's own functional annotation tool, formerly known as SNP effect predictor. The tool can be used either by a web interface, as command-line tool or via a Perl API. The web interface version is aimed at users analyzing smaller sets of variants, as it is only capable of processing 750 variants per file.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

C. Visualization

All of the evaluated genome browsers have in common that they are capable of displaying numerous 1D tracks which contain information about the reference genome, the transcriptome, aligned reads, found mutations, annotations collected from public data sources or other data types important for the correct interpretation of results [37]. The two types of genome browsers, namely web-based applications and stand-alone tools, as well as CNV/SV visualization tools are used.

VI. CONCLUSION

In our report, we provide a comprehensive survey of tools available for the analysis of whole-genome/ whole-exome data covering all analytical steps: quality assessment, alignment, variant identification, variant annotation and visualization. The information provideD represents a valuable guideline for both an expert in the field and a less-experienced user to select the appropriate tools for a specific application and assemble an optimal analytical pipeline. The analysis of NGS data is a fast moving field and recommendations which tools to use might quickly change. Nevertheless, we make the following general recommendations. First, tools for quality assessment and alignment matured to a great extent and the choice is rather straightforward. Second, for variant identification, we suggest a consensus approach, e.g. running CRISP, GATK and SAM tools on the same data set. Third, the choice of an annotation tool is largely dependent on the desired selection of variant annotations. Fourth, visualization tools are usually easy to install and it might therefore be a valid approach to test different software suites.

REFERNCES

1. Gonzaga-Jauregui C, Lupski JR and Gibbs RA, 'Human genome sequencing in health and disease,' Annu Rev Med, Vol. 63, pp. 35–61, 2012.
2. Ng PC, Levy S, Huang J, et al., 'Genetic variation in an individual human exome', PLoS Genet, Vol. 4, 2008.
3. Bamshad MJ, Shendure JA, Valle D, et al., 'The centers for Mendelian genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions' ,AmJMed Genet, pp. 1523–1525, 2012.
4. Castle JC, Kreiter S, Diekmann J, et al., 'Exploiting the mutanome for tumor vaccination,'Cancer Res, Vol. 72, pp. 1081–91.
5. Schadt EE, Linderman MD, Sorenson J, et al. 'Computational solutions to large-scale data management and analysis,' Nat Rev Genet, Vol 11,pp. 647–57, 2010
6. Li H and Homer N, 'A survey of sequence alignment algorithms for next-generation sequencing,' Brief Bioinformatics, Vol. 11, pp. 473–83, 2010
7. Botstein D, Risch N. 'Discovering genotypes underlying human phenotypes: past uccesses for mendelian disease, future approaches for complex disease,' Nat Genet, Vol. 33, pp. 228–37, 2003.
8. Lalonde E, Albrecht S, Ha KCH, et al. 'Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing,' HumMutat, Vol.31, pp. 918–923, 2010
9. Marian AJ. 'Molecular genetic studies of complex phenotypes', Transl Re, Vol 159, pp. 64-79. 2012
10. Kathiresan S, Srivastava D. 'Genetics of human cardiovascular disease,' Cell, Vol.148, pp. 1242–57, 2012.
11. Norton N, Li D, Rieder MJ, et al.'Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy,' AmJ Hum Genet, Vol. 88, pp. 273–82, 2011.
12. Nejentsev S, Walker N, Riches D, et al. 'Rare variants of IFIH1, a gene implicated in Antiviral responses, protect against type 1 diabetes,' Science,; Vol. 324, pp. 387–389, 2009.
13. Foulkes WD,' Inherited susceptibility to common cancer,'. NEnglJMed, Vol. 359, pp. 2143–2153, 2008.
15 Meindl A, Hellebrand H, Wiek C, et al., 'Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gen', Nat Genet, Vol. 42, pp. 410–414, 2010.
16. Walther A, Johnstone E, Swanton C, et al. 'Genetic prognostic and predictive markers in colorectal cancer,' Nat Rev Cancer, Vol. 9, pp. 489–499, 2009.
17. Mardis ER, 'Next-generation DNA sequencing methods,' Annu Rev Genomics Hum Genet, Vol. 9, pp. 387–402, 2008.
18. Margulies M, Egholm M, Altman WE, et al., 'Genome sequencing in microfabricated high-density picolitre reactors,' Nature, Vol. 437, pp. 376–380, 2005.
19. Medvedev P, Stanciu M and Brudno M, 'Computational methods for discovering structural variation with next-generation sequencing,' NatMethods, Vol. 6, S1, pp. 3–20, 2009.
20. Schmieder R and Edwards R, 'Quality control and preprocessing of metagenomic datasets,' Bioinformatics, Vol. 27, pp 863–864, 2011.
21. Cibulskis K, McKenna A, Fennell T, et al. 'ContEst: estimating cross-contamination of human samples in next generation sequencing data,' Bioinformatics, Vol 27, pp. 2601–2602, 2011.
.22. Planet E, Attolini CS-O, Reina O, et al. 'htSeqTools:high-throughput sequencing quality control, processing and visualization in R,' Bioinformatics, Vol 28, pp. 589–590, 2012.
23. Dolan PC and Denver DR., 'TileQC: a system for tile-based quality control of Solexa data,' BMC Bioinformatics, Vol. 9, pp. 250, 2008.
24. Neuman JA, Isakov O and Shomron N, 'Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection,' Brief Bioinformatics, Vol. 14, pp. 46–55, 2013.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

25. Sathirapongsasuti JF, Lee H, Horst BAJ, et al. 'Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV,' Bioinformatics, ;Vol 27, pp. 2648–2654, 2011.
26. The Genome Reference Consortium. http://www.ncbi. nlm.nih.gov/projects/genome/assembly/grc.
27. Genome Bioinformatics Group (UCSC). Comparison of UCSC and NCBI human assemblies. http://genome.ucsc. edu/FAQ/FAQreleases.html#release4.
28. Langmead B, Trapnell C, Pop M, et al., 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,' Genome Biol,Vol. 10:R25, 2009.
29. Yu X, Guda K, Willis J, et al. 'How do alignment programs perform on sequencing data with varying qualities and from repetitive regions?', BioDataMining, Vol 5,pp. 6, 2012.
30. Li H and Durbin R, 'Fast and accurate long-read alignment with Burrows-Wheeler transform,' Bioinformatics, Vol. 26, pp. 589–95,2010.
31. Li H, Ruan J and Durbin R, 'Mapping short DNA sequencing reads and calling variants using mapping quality scores,' Genome Res, Vol. 18, pp. 1851–1858, 2008.
32. Schmieder R, Lim YW, Rohwer F, et al. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets,' BMC Bioinformatics, Vol 11, pp. 341, 2010.
33. Raney BJ, Cline MS, Rosenbloom KR, et al. 'ENCODE whole-genome data in the UCSC genome browser (2011 update),' Nucleic Acids Res, Vol. 39, pp. 871–875, 2011.
34. Alkan C, Kidd JM, Marques-Bonet T, et al. 'Personalized copy number and segmental duplication maps using next-generation sequencing,' Nat Genet, Vol. 41, pp. 1061–1067, 2009.
35. Li R, Yu C, Li Y, et al.,'SOAP2: an improved ultrafast tool for short read alignment,' Bioinformatics,Vol. 25, pp.1966–1967, 2009.
36. Galinsky VL ,'YOABS: yet other aligner of biological sequences— an efficient linearly scaling nucleotide aligner,' Bioinformatics, Vol. 28, pp. 1070–1077, 2012.
37. Ruffalo M, LaFramboise T and Koyutu«rk M, 'Comparative analysis of algorithms for next-generation sequencing read alignment,'Bioinformatics,Vol. 27, pp. 2790–2796, 2011.
38. Kim SY, Li Y, Guo Y, et al. 'Design of association studies with pooled or un-pooled next-generation sequencing data,' Genet Epidemiol, Vol. 34, pp. 479–491, 2010.
39. Nielsen CB, Cantor M, Dubchak I, et al. 'Visualizing genomes: techniques and challenges,' Nat Methods, Vol. 7, S5, pp 15, 2010.
40. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics,' GenomeRes, Vol. 19, pp 1639–1645, 2009.