International Journal of Applied Business and Economic Research

ISSN : 0972-7302

available at http: www.serialsjournals.com

Volume 15 • Number 22 • 2017

Next Generation Sequencing Data Analysis Software and Methods: A Survey

U. Vignesh[1] and R. Parvathi[1]

[1]*School of Computing Science and Engineering, VIT University, Chennai, Tamilnadu - 600 127, India*

ABSTRACT

Rapidly evolving high throughput technologies provides biological data on a very large scale. In this article, we will review the recent development of liner time algorithms and tools for big data analysis produced by next generation sequencing and the comparison of commonly used tools with different algorithms for NGS big data analysis based on their performance, input format and output format etc. Due to the diversity of contexts in which biological data analysis is performed, several problems are commonly studied. This survey paves a way to find provably better algorithmic tool to the underlying optimization problem. NGS resultant data generates short reads that address various biological questions. The methods to analyze NGS big data includes various algorithms on various tools which differs by their performance, complexity, cost etc. We have accumulated extensive experience in sample handling, variant detection and bioinformatics analysis on mentioned tools.

Keywords: Big data; Next Generation Sequencing; High Throughput Sequencing; Alignment Sequencing; Variant Detection; Polymorphism Detection.

1. INTRODUCTION

High throughput data sequencing and analysis are the major research area for the past decade to increase the speed. In HTS various tools are used with different algorithms. The methods used for analysis has various architecture with different platform incorporated on it. This sequencing technologies have been used in biomedical applications to identify the patterns present in a DNA which indicates certain diseases with symptoms, age of the sample given and drug target identification for a new discovery etc. It also plays a major role in epigenome, interactome and cytogenome (Wang et al., 2013). High throughput sequencing in genome has been decreased due to its cost and so the genomics become feasible. Thus, the big data analysis

including different groups of research provided framework becomes success with reducing time and cost by a factor of 1 million on personalized genomics (Costa et al., 2014). In this area data has grown tremendously that it can't be maintained by the usual software tools available for analysis, visualize etc. the large volume of biological data are resulted in HTS has become a challenge to analysis with information extraction and a transfer of useful information in an efficient manner with high security aspects. For example, Square Kilometer Array, which generates data exceptionally large 40 GB per second and another example, Flickr, which requires 3.6 TB storage for a single day (Wu et al., 2014). In the aspect of biomedical applications, to predict the disease in a patient's DNA and make remedy in an fast manner is must but not an option. Thus, the process has to maintain constant observation on improving performance, reducing dependency on hard disk, greater memory system where all data are maintained (Zhang et al., 2015). In order to perform biological big data management efficiently focus has to be done on following operations, viz.

1. **Indexes :** To maintain efficiency by avoiding memory intensive scan.
2. **Transaction management :** Accordance to the semaphore maintains many core systems.
3. **Data level parallelism :** To speed up the processes by performing multiple operations on single cycle.
4. **Query processing :** It register temporal locality and performs efficiency in time.
5. **Data layouts :** For the purpose of cache consciousness and space efficiency.
6. **Data overflow :** Hybrid systems with non-volatile memories produce speed for accessing the data to control data exceedance.

2. BIG DATA CHALLENGES IN NGS ANALYSIS METHODS

The process takes place in NGS is to unravel the ordered sequence of nucleic acids that group together to make DNA of the given sample. The big data from NGS becomes so big and the challenge of maintaining and analyzing data also increased in recent years. The first human genome sequencing took nearly 10 years with amount in number of billion USD, whereas now it takes only one week with 2000 USD in a single machine due to the different software tools available for big data analytics in an efficient manner that speeds up the processes. The factors concentrated to face the challenges in NGS analysis are listed, viz.

1. **Matching :** Heuristic algorithms developed to overcome approximate matching in an accurate level.
2. **Mapping :** Read alignment to find the structure, function, relation between multiple sequences, highly conserved regions are achieved through a cloud.
3. **Storing :** To provide a space for data for example when it increases from 5 to 5000 human genome, approximately it occupies 15 terabytes and also bandwidth maintenance.
4. **Questioning :** Depends on changes in the question, the search process has to be done on different genomes of database rather than on single or few genomes.

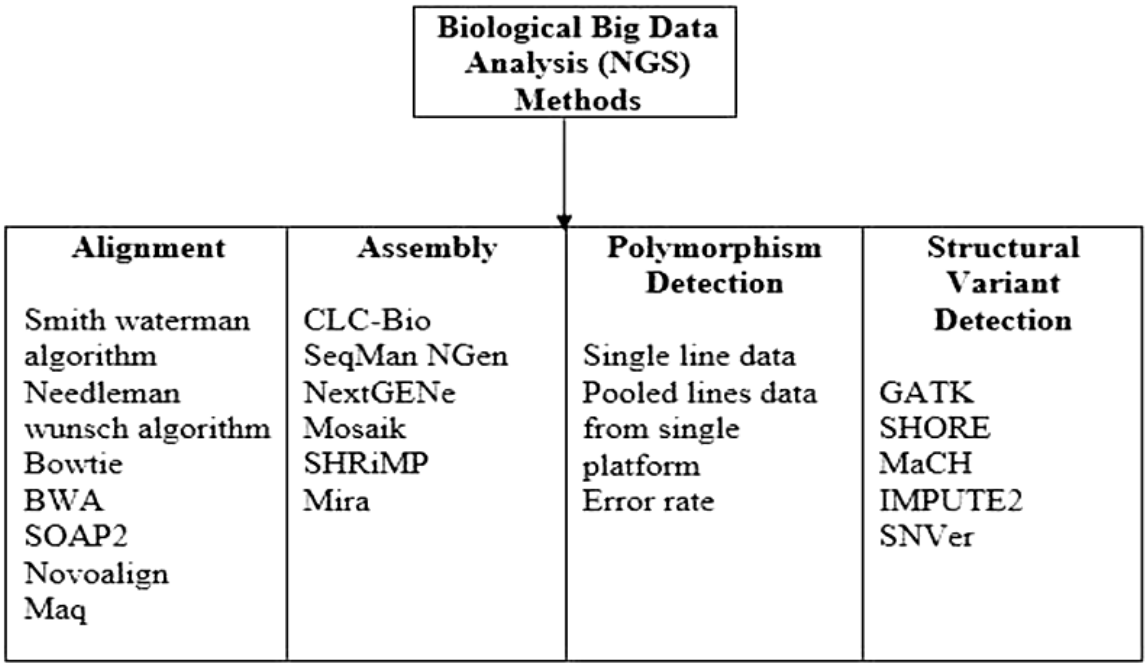

Figure 1: NGS data analysis methods

3. NGS DATA ANALYSIS

Next generation sequencing technologies has grown widely in the field of identifying patterns in epigenetic, genomic and transcriptome levels to predict the factors such as disease, age etc. this section gives the detailed overview of open source software's available for alignment programs, assembly and variant calling methodologies with the challenges to maintain these large giga base pairs big data Next generation sequencing data analysis methods involves four models incorporated in it as shown in Figure 1. NGS data has moved biology in a whole including molecular, micro, etc. into the big data era. For example, European Bioinformatics Institute doubles 2 petabyte of genomic data every 18 months out of total storage 20 petabytes and doubles biological data every 9 months (Bao et al., 2009). As the NGS generates data of volume hundreds of terabytes to petabytes, it is very difficult to maintain this big data with no loss of quality and by maintaining storage capacity and capabilities. Hence, the bioinformatics tools are discovered and they are refined with new efficient algorithms and technologies to manage the big data. Few public organization EBI, NCBI, NIH etc. has the capacity to overcome the drawbacks in storing, managing and information extraction on big data ease. The better way to process these big biological data are in the cloud by making use of cloud providers such as amazon etc. NGS data analysis always prefers a linear time algorithms in their software tools.

3.1. Big data problem in NGS

1. **Gene regulatory network (GRN) :** Gene regulatory network analysis in an expression is a complicated one by producing correlational possibilities and has a challenge to face an iterative problem. It requires a very large scale data analytics system for a regulatory identification in a affected or diseased network.

2. **Sequence and Protein protein interaction (PPI) :** Since from 1980, high volumes of data are managed in a database and their updates with new inventions also increases the volume, velocity and varied data formats. The process sequence analysis includes searches in the database, which has various dimensions of big data incorporated in it are not quite ease rather than use of big data technologies. PPI results voluminous data with changes in their nature provides a challenge for efficient and scalable framework to act in a fast and accurate PPI pattern generation.

3. **Microarray data :** The cost reduction aspect leads the growth in use of microarray data tremendously. It requires big data technologies for the speed to construct coexpression and regulatory networks using voluminous microarray data.

3.2. Alignment

Aligning sequences to read the references and functions of a pattern in a big data are done based on the indexes framework. The different types of alignment software's with various efficient algorithms are listed in table 1. Depending on index, there are three different categories for framing the algorithm routine (Li et al., 2010) viz.

1. **Hash tables :** The most suitable example for algorithm based on hash tables is BLAST algorithm, which was commonly in use for more than 2 decades with refinement by smith waterman algorithm.

2. **Merge sorting :** The maximum usage of information in manner of probability for read alignment based on algorithm merge sorting takes place in slider and sliderII software tool (Ivakhno et al., 2007).

3. **Suffix trees :** Suffix trees based algorithms work efficiently in matching patterns accurately and list the inexact matches that supports the alignment. Bowtie (Langmead et al., 2009) and Bowtie2 (Langmead et al., 2012) software uses bowtie algorithm which works by suffix tree framework.

Table 1
NGS Data Sequence Alignment Softwares

Name	*Input/Output*	*Supported platforms*	*Indexing Method/ Gapped Alignment*
Barra CUDA	FASTQ/SAM	Illumina	FM index (BWT)/yes
BFAST	FASTQ/SAM	Illumina, ABI Solid, 454	Multiple(hash, tree)/yes
Bowtie	FASTQ, FASTA/SAM	Illumina, ABI Solid	FM index(BWT)/no
Bowtie2	FASTQ, FASTA, QSEQ/SAM	Illumina, 454	FM index(BWT)/yes
BWA	FASTQ, FASTA/SAM	Illumina, ABI, Solid(1)	FM index(BWT)/yes
BWA-SW	FASTQ, FASTA/SAM	454	FM index(BWT)/yes
SliderII	PRB files/CSV	Illumina	Merge sorting
MAQ	FASTQ, FASTA/MAQ	Illumina	Hash based/yes
mrFAST	FASTQ, FASTA, SAM, DIVET	Illumina	Hash based/yes

Name	*Input/Output*	*Supported platforms*	*Indexing Method/ Gapped Alignment*
mrsFAST	FASTQ, FASTA/SAM, DIVET	Illumina	Hash based/no
SOAP2	FASTQ, FASTA/SOAP2	Illumina	FM index(BWT)/yes
SOAP3	FASTQ, FASTA/SAM	Illumina	FM index(BWT)/no
SSAHA2	FASTA/SAM, GFF	Illumina, ABI Solid, 454	Tree index/yes
Stampy	FASTQ, FASTA/SAM	Illumina, 454	FM index(BWT)

3.3. Assembly

This section gives the detailed description about the assembly software packages created since 2005 and revised specifically for the purpose of efficient assembly operations in next generation sequencing. Assembly operations follows a data structure of hierarchical type that points the data sequence to putative repetitive construction of the considered pattern target (Miller et al., 2010). The file format for assembly in common use is FASTA. The big data challenge in assembly is to elaborate the functions of short read lengths that are very much smaller compared to smallest genomes. Due to big data there will be a possibility of imperfect sequence alignments hence big data technologies are to be incorporated in software packages. The very first NGS assembly software packages used greedy algorithms i.e. an approximation result but it has been refined with greedy graph based algorithms. Solexa and Solid platforms uses de brujin graph approach to the short reads. The different types of NGS big data sequence assembly software's with their respective supporting platform and their web address are described in table 2.

Table 2
NGS Data Sequence Assembly Software's

Name	*Input/output*	*Supported platforms*	*Indexing Method/ Gapped Alignment*
ABySS	Solexa, Solid	2008/2014	http://www.bcgsc.ca/ platform/bioinfo/ software/abyss
ALLPATHS- LG	Solexa, Solid	2011	http: //www.broadinstitute. org/science/ programs/ genome-biology/crd
AMOS	Sanger, 454	2002/2011	http://sourceforge.net/apps/mediawiki/ amos/index.phptitle=AMOS
Arapan-M	All	2011/2012	http://sourceforge.net/ projects/ dnascissor/ files
Arapan-S	All	2011/2012	http://sourceforge.net/ projects/ dnascissor/ files
CABOG	Sanger, 454, Solexa	2004/2015	http://www.jcvi.org/cms/research/ projects/celera-assembler/overview/
CLC Assembly Cell	Sanger, 454, Solexa, SOLiD	2008/2010/2014	http://www.clcbio.com/ products/
Cortex	Solexa, Solid	2011	http://cortexassembler. sourceforge.net/

Name	*Input/output*	*Supported platforms*	*Indexing Method/ Gapped Alignment*
DNA Baser Assembler	Sanger, 454	2015	www.DnaBaser.com
DNA Dragon	Illumina, Solid, Complete Genomics, 454, Sanger	2011	https: //www.dna-dragon.com/
Edena	Illumina	2008/2013	http://www.genomic.ch/ edena.php
Euler	Sanger, 454	2001/2006	http: //nbcr.sdsc.edu/euler/
Euler-sr	454, Solexa	2008	http://euler-assembler. ucsd.edu/portal/
Fermi	Illumina	2012	https://github.com/lh3/ fermi
Forge	454, Solexa, Solid, Sanger	2010	http: //combiol.org/forge/
Geneious	Sanger, 454, Solexa, Ion Torrent, Complete Genomics, PacBio, Oxford Nanopore, Illumina	2009/2013	http://geneious.com/
IDBA	Sanger, 454, Solexa	2010	http://www.cs.hku.hk/~alse/idba/
LIGR Assembler	Sanger	2009/2012	http://sourceforge.net/ projects/ ligr-assembler/
MaSuRCA	Sanger, Illumina, 454	2012/2013	http://www.genome.umd. edu/masurca. html
MIRA	Sanger, 454, Solexa	1998/2014	http://sourceforge.net/ apps/mediawiki/ mira-assembler/
Nextgene	454, Solexa, Solid	2008	http://softgenetics.com/NextGENe.html
Newbler	454, Sanger	2009/2012	http://www.454.com/
PADENA	454, Sanger	2010	http: //bio.codeplex.com/
PASHA	Illumina	2011	http://sites.google. com/site/ yongchaosoftware/pasha
Phrap	Sanger, 454, Solexa	1994/2008	http://www.phrap.org/
Tigr Assembler	Sanger	1995/2003	ftp://ftp.jcvi.org/pub/ software/ assembler/
Ray	Illumina, mix of Illumina and 454, paired or not	2010	http://denovoassembler. sf.net/
SeqMan NGen	Illumina, ABI, Solid, Roche 454, Ion Torrent, Solexa, Sanger	2007/2014	http://www.dnastar.com/
SGA	Illumina, Sanger	2011/2012	https: //github.com/jts/sga
Sharcgs	Solexa	2007/2007	http://sharcgs.molgen. mpg.de/
SOPRA	Illumina, Solid,	2010/2011	http: //www.physics.rutgers. edu/~anirvans/SOPRA/
Sparse Assembler	Illumina, 454, Ion torrent	2012/2012	https: //sites.google.com/ site/ sparseassembler/

Name	*Input/output*	*Supported platforms*	*Indexing Method/ Gapped Alignment*
SSAKE	Solexa	2007/2014	http://www.bcgsc.ca/ platform/bioinfo/ software/ssake
SOAP denovo	Solexa	2009/2013	http://soap.genomics. org.cn/ soapdenovo.html
SPAdes	Illumina, Solexa, Sanger, 454, Ion Torrent, PacBio, Oxford Nanopore	2012/2015	http://bioinf.spbau.ru/ en/spades
Staden gap4 package	Sanger	1991/2008	http://staden. sourceforge.net/
Taipan	Illumina	2009/2009	http://sourceforge.net/ projects/taipan/
VCAKE	Solexa	2007/2009	http://sourceforge.net/ projects/vcake
Phusion	Sanger	2003/2006	http://www.sanger.ac. uk/Software/ production/ phusion/
QSRA	Sanger, Solexa	2009/2009	http://qsra.cgrb. oregonstate.edu/
Velvet	Sanger, 454, Solexa, Solid	2007/2011	http://www.ebi.ac.uk/~zerbino/velvet/

3.4. Variant analysis

3.4.1. CNV

Copy number variant is a type of secondary data analysis, rearrangements in the data process are carried out with copy number variant callers. The CNV identification takes the preprocess steps, i.e. data matrix formation involves the process easier in the software packages. In this way, the mutations in a DNA can also be pointed for big data analytics. The table 3 gives clear description about the copy number variant identification tools with their supporting platform, I/O format and last updates. Copy number variant is a type of secondary data analysis, rearrangements in the data process are carried out with copy number variant callers.

Table 3
CNV Identification Tools

Name	*Input Format*	*Output Format*	*Platforms*		*Last Updated*
			Illumina	*Solid*	
CNAseg	SAM/BAM	CSV	Yes	Yes	2010-09-14
CNVer	CSV, SAM/BAM	CNV files	Yes	Yes	2011-07-11
CNVnator	FASTA, SAM/BAM	CSV	Yes	Yes	2012-02-07
CNV-seq	SAM/BAM	CNV files	Yes	Yes	2011-07-15
CONTRA	2 BAM files, BED, SAM/BAM	VCF, CSV	Yes	Yes	2012-07-24
Copy Seq	SAM/BAM	CSV	Yes	Yes	2011-04-06
RDX plorer	FASTA, SAM/BAM	CSV	Yes	Yes	2012-01-13
read Depth	BED, R	Segmented CNVs, CSV	Yes	Yes	2011-04-15

3.4.2. Structural variant

Structural variants involves steps to sort the biological analysis of the known and novel variants by making use of the listed software packages. The table 4 gives detailed description about the structural variant identification tools with their supporting platform, I/O format and last updates. Structural variant identification in an genome sequencing plays a major role in identifying a specific disease that are associated with it. This process provides an relationship between the structural variant and the associated disease. polymorphism detection in the structural variant varies in their characteristics of mapping breakpoints by the assembly. The deletion and duplications identification in an structural variant can be done by using the read depth analysis. The gapped alignment in the visualization part shows the assembled sequence. Structural variant identification deals with the two read pairs viz.

1. **Paired-end:** The direction of sequencing is from backwards to the middle.
2. **Mate-pair:** The read pairs has the direction pointing outwards against the original fragments.

There are different read pair algorithms to do the processes in an efficient manner on the SV clusters. Clusters are categorized into two divisons viz. Less number of pairs with same signature; Large value of mean and standard deviation.

Structural variants rearranges the genomes and produces more than 50 base pairs in 0.99 variations among the genomes. several databses are maintained for genomic variants providing structural variants identification such as Database of genomic variants, dbvar, etc. Identification of structural variants in an single base pair can be achieved by using split reads. The best approach for identifying structural variants in the clinical application is said to be amplicon sequencing data technique. this technique incorporates different methods for the goal such as ONCOCNV, Principal component analysis, Segmentation and clustering approach etc.

Table 4
Structural variant identification tools

Name	*Input Format*	*Output Format*	*Platforms*		*Last Updated*
			Illumina	*Solid*	
Apolloh	CSV	CSV	Yes	Yes	2011-12-01
Break Dancer	BAM	BED	Yes	Yes	2011-02-21
Break pointer	BAM	GFF	Yes	Yes	2012-01-20
Breakway	BAM	CSV	Yes	Yes	2011-04-01
Clip Crop	SAM, FASA	BED	Yes	Yes	2012-01-27
Fusion Map	FASTQ	SAM	Yes	Yes	2012-04-17
GASV Pro	SAM/BAM	Clusters file	Yes	Yes	2012-06-14
PEMer	Several input files	Multiple output files	Yes	Yes	2009-02-02
Splitread	FASTA, SAM	BED	Yes	No	2011-10-13
SVDetect	BAM/SAM or ELAND or Bioscope output	Txt, BED, Circos	Yes	Yes	2011-07-12

3.4.3. Variant annotation

This process gives the functional annotation of the classified variant data, through which the identification of highly associated variants in a DNA are done. Variant annotation helps to predict the risk alleles in a given sample. The various genome browsers with the variant annotation indicating their I/O format are described in table 5.

Table 5
Genome browsers with variant annotations

Name	*Input/Output Formats*	*Explanation*
ABrowse	GFF, WIG	Shows tracks as large images similar to google maps;
Argo / Combo	FASTA, Genbank, GFF, BLAST, BED, Wiggle (WIG), Genscan files	Argo is a standalone genome viewer which integrates combo as a comparative genome browser.
Artemis	BCF, FASTA	Standalone tool where BAMView has been integrated;
Bambino	FASTA, UCSC, 2bit, nib	
Consed	Newbler, Cross_match, Phrap, MIRA, Velvet and PCAP	The standalone tool has been designed to display genome assemblies.
DiProGB	GenBank, FASTA, GFF PTT	Is able to display sequence graphs and a feature graphs;
EagleView	ACE, READS, EGL, MAP	A genome assembler viewer;
Ensembl	BED, BedGraph, GFF, GTF, PSL, WIG, BigWig	Web-based tool with a variety of reference genome and integrated annotations;
Gaggle	SQL, GFF	For systems biology;
Gap5	ACE, BAF	This standalone tool has been developed to facilitate the process of finishing assemblies.
GBrowse	GFF	This web-based tool is the precursor of JBrowse.
Geno Viewer	FASTA, GFF	It is a standalone genome viewer that is not developed or supported anymore.
Hawkeye	fastq, fastq	A genome assembler viewer;
Integrated Genome Browser (IGB)	DAS, wig	Standalone, Java tool with export feature into PDF, EPS, PNG, …;
Integrative Genomics Viewer (IGV)	(> 30 formats) TDF, CN, SNP, GCT, RES, GFF, GFF3, BED, GISTIC, LOH, MUT, GCT, SEG, CBS, IGV, TAB, WIG	Can be started locally or from websites; offers lots of customization features;
JalView	DAS	This tools is capable of performing multiple sequence alignment.
JBrowse	FASTA, BED, GFF, GFF3, WIG	It is a web based tool where tracks are rendered on the client side. Tracks need to be prepared by the user in advance.

Name	*Input/Output Formats*	*Explanation*
LookSeq	MAQ, CIGAR	Web based alignment viewer;
Magic Viewer	ACE	This tool is aimed at users who work with DNA methylation data.
MapView	MVF	
NGSView	XML, BED, BLAST, Eland, mapview processed MAQ, Corona, GFF	Sequence alignment editor;
Savant	FASTA, BED, GFF, WIG, any tab-delimited	Standalone, Java based genome viewer which allows users to create their own plug-ins;
UCSC Genome Browser	BED, bigBed, bedGraph, GFF, GTF, WIG, bigWig, MAF, BED, SNP, PSL	Web-based tool with a variety of public databases; It offers many customization features and allows the user to upload new tracks.
UTGB toolkit	FASTA, BED, WIG, DAS	The tool is web-based and uses a dedicated database and web-server. It offers flexible customization possibilities and tracks can hold private or public data.
VEGA	BED, bedGraph, BigBed, BigWig, GBrowse, GFF, GTF, PSL, WIG.	This application contains manually annotated genomes from different species. Large parts of the human genome are annotated.

4. HIGH THROUGHPUT SEQUENCING WORKFLOW SYSTEMS

Use High throughput sequencing workflow systems provide easy and cost reduced perspective to genome sequencing with timely detection of functions, accurate and fast solutions for big data in bioinformatics. The table 6 shows the detailed view of the different workflow systems that can support high throughput sequencing technologies which includes a big data incorporated in it for analysis, visualization and further process to extract the information. High throughput sequencing based platforms can be classified into two parts viz.

1. Template
2. Sequencing chemistry

Template HTS technologies includes major categories such as fragmentation, tagging and amplification process for the genomic sequencing to produce the result in an efficient manner and it reduces the costs of manpower and kit needed for the sequencing. Data tracking plas an activity to copy the data from local information technology infrastructure. These techniques are applicable for denovo assembly of a genome of unit strain, phlogenetic analysis etc. HTS technologies can be applied to metagenomic sequencing, new prospects for sequence profiling etc for proper sequencing methodology to produce major possibilities.

Table 6
NGS Workflow Systems

Name	*Requirement*	*Explanation*
Galaxy	Linux, MAC OS X	It is a web based platform for aligning, sequencing and visualizing NGS data. It incorporates different NGS tools as a package. It also provides user to perform their operation in specific package and add it to galaxy for sharing and performing the process globally.
Gene Pattern	Linux, MAC OS X, Windows	Genomic analytic tool with user friendly interface helps to automate, split the tough steps in the genomic analysis and which includes more than 150 tools in built to it.
KNIME	Linux, MAC OS X, Windows	Open platform for NGS data which provides fast, scalable and user-friendly software package with inclusion of analysis programs for the purpose of filtering aspects and exploit VCF files.
Taverna	Linux, MAC OS X, Windows	Open platform with large number of NGS tools integrated with it highlighting on programming paradigm but without sequence analysis tools that already existing

5. CONCLUSION

In this paper to explore the big data analysis technique in NGS biological research, we analyzed different software packages for computational modelling. This study also deals with the memory management to provide space for advances in bioinformatics to address big data problems. We have compared various software tools for the de novo NGS data analysis. Big genomes paves a way to give more complex repeated structures, which can be overcome by using best alignment and assembler software tools. This survey provides a technical review of NGS data analysis software tools, algorithms and workflow systems to date, which hopefully will be a useful resource for future next generation sequencing research aspects.

6. APPENDIX

We provide the proofs for egg dataset 1GHL.pdb downloaded from NCBI database with the data size less than 20 GB with the simulation results on tools bowtie2, tophat, etc on the NGS galaxy workflow system are described on the following figures.

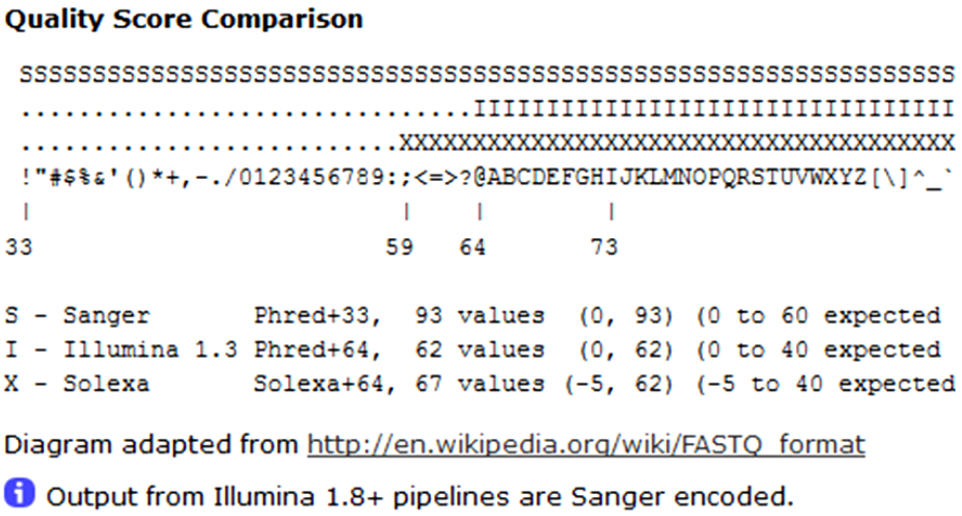

Figure 2: Galaxy – Quality Comparison

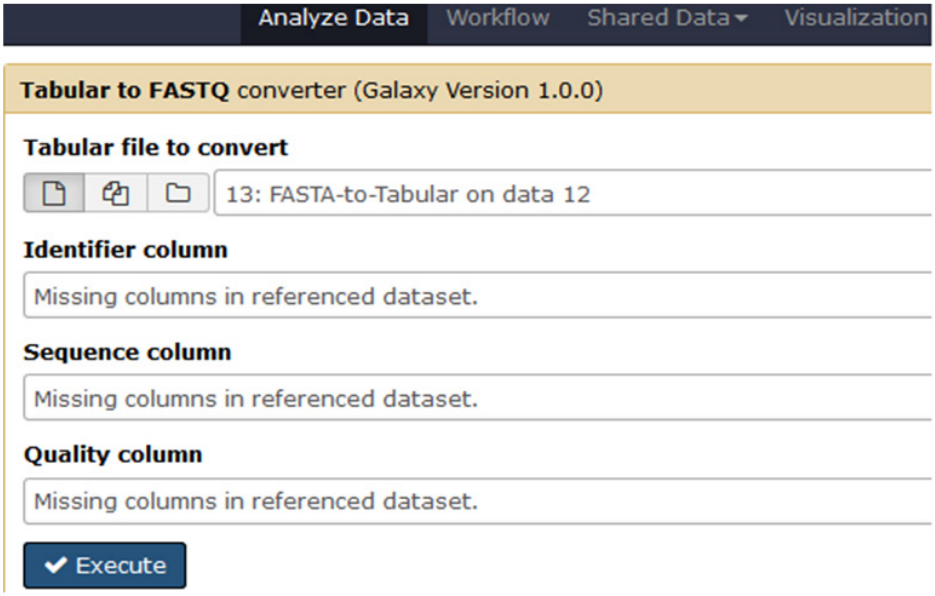

Figure 3: Galaxy – Format Conversion

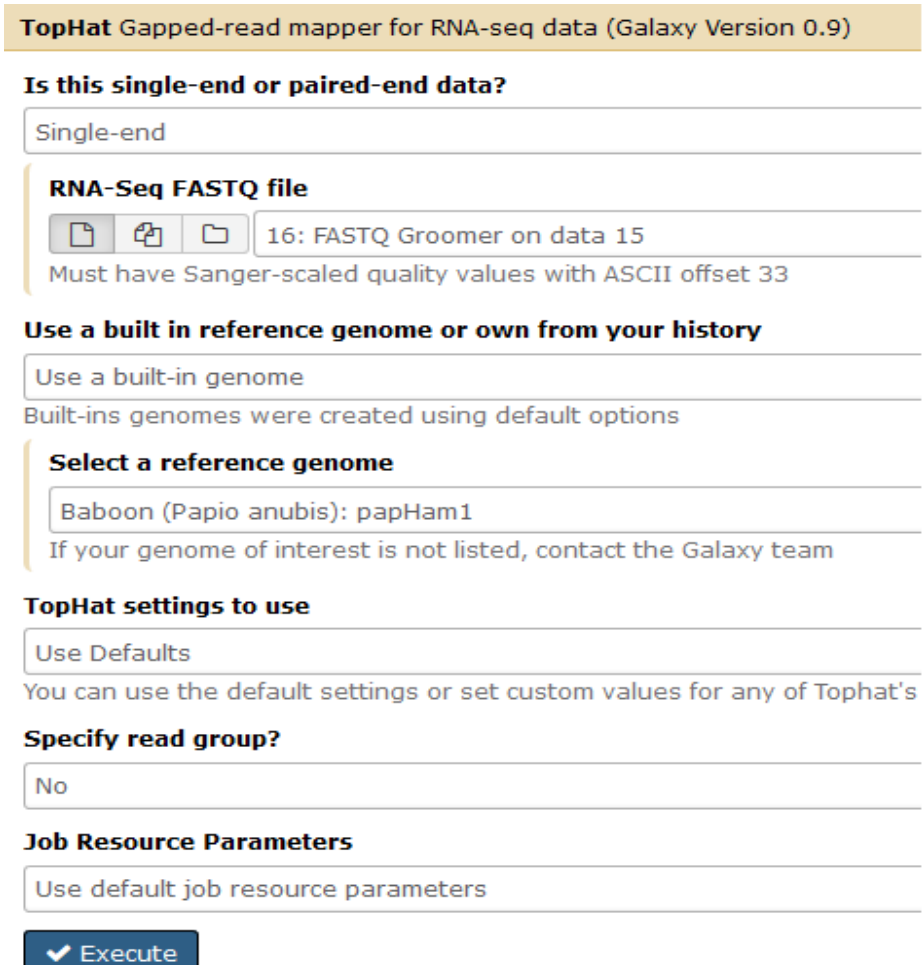

Figure 4: Galaxy – Data Pair Operations

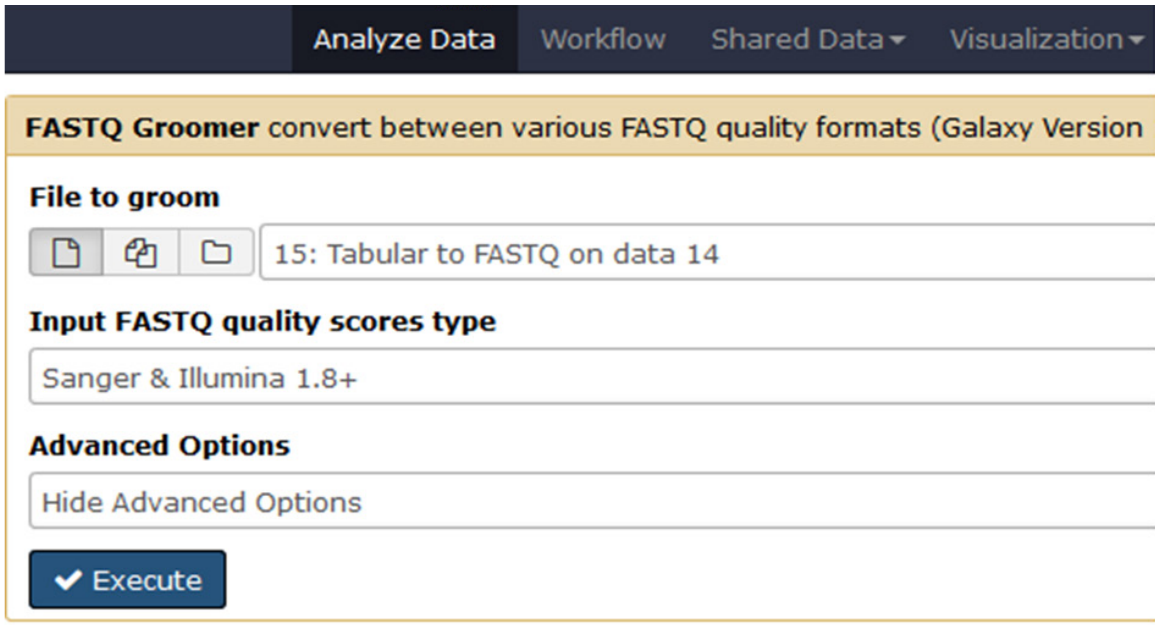

Figure 5: Galaxy – FASTQ Groomer Tool

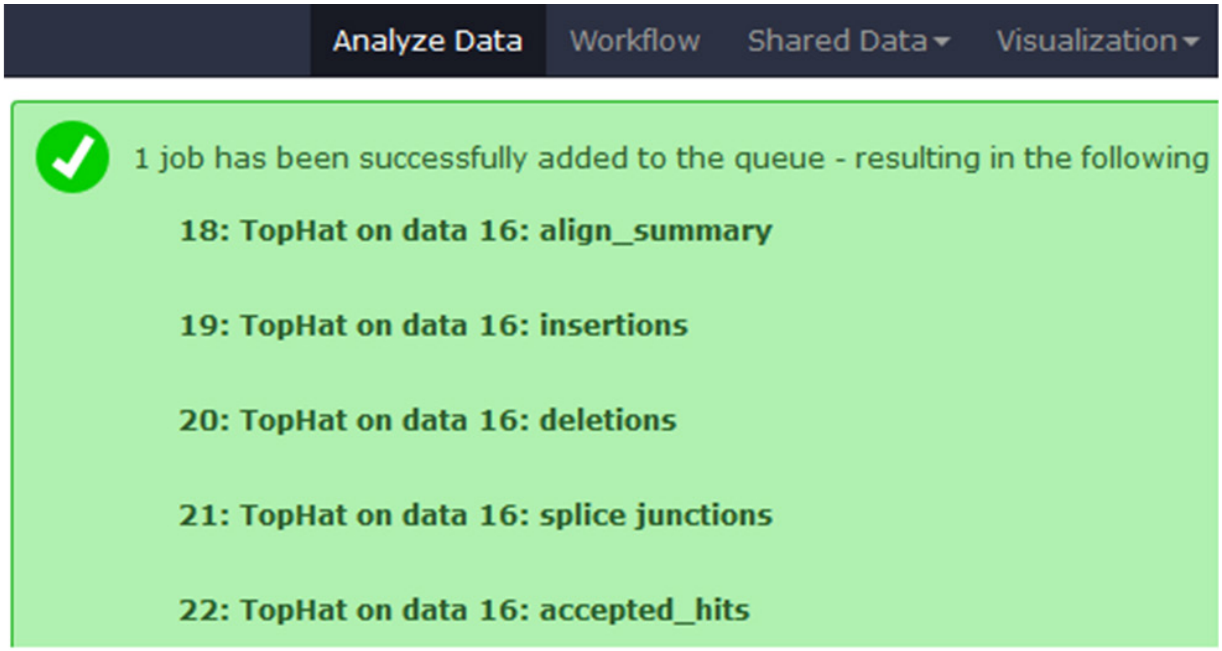

Figure 6: Galaxy – Data Results

References

Cibulskis K, McKenna A, Fennell T, et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 2011;27:2601–2602.

FastQC. http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc.

FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit.

Blankenberg D, Gordon A, Von Kuster G, et al. Manipulation of FASTQ data with Galaxy. *Bioinformatics* 2010;26:1783–1785.

Planet E, Attolini CS-O, Reina O, et al. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* 2012;28:589–590.

Dai M, Thompson RC, Maher C, et al. NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics* 2010;11 Suppl 4:S7.

Martínez-Alcántara A, Ballesteros E, Feng C, et al. PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* 2009;25:2438–2439.

Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;27:863–864.

Cox MP, Peterson DA, Biggs PJ. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 2010;11:485.

Schmieder R, Lim YW, Rohwer F, et al. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 2010;11:341.

Dolan PC, Denver DR. TileQC: a system for tile-based quality control of Solexa data. *BMC Bioinformatics* 2008;9:250.

Klus P, Lam S, Lyberg D, et al. BarraCUDA - a fast short read sequence aligner using graphics processing units. *BMC Res Notes* 2012;5:27.

Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE* 2009;4:e7767.

Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10:R25.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 2012;9:357–359.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589–595.

Cox AJ. ELAND: Efficient large-scale alignment of nucleotide databases. Illumina. 2007;

Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008;18:1851–1858.

Mosaik. http://code.google.com/p/mosaik-aligner/.

Alkan C, Kidd JM, Marques-Bonet T, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* 2009;41:1061–1067.

Hach F, Hormozdiari F, Alkan C, et al. mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* 2010;7:576–577.

Novoalign. http://novocraft.com.

Li R, Yu C, Li Y, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009;25:1966–1967.

Liu C-M, Wong T, Wu E, et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics* 2012;28:878–879.

Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome Res.* 2001;11:1725–1729.

Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 2011;21:936–939.

Galinsky VL. YOABS: yet other aligner of biological sequences--an efficient linearly scaling nucleotide aligner. *Bioinformatics* 2012;28:1070–1077.

Challis D, Yu J, Evani US, et al. An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* 2012;13:8.

Edmonson MN, Zhang J, Yan C, et al. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* 2011;27:865–866.

Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet.* 2009;85:847–861.

CoNAn-SNV. http://compbio.bccrc.ca/software/conan-snv/.

Iqbal Z, Caccamo M, Turner I, et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 2012;44:226–232.

Bansal V. A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* 2010;26:i318–324.

Albers CA, Lunter G, MacArthur DG, et al. Dindel: accurate indel calls from short-read data. *Genome Res.* 2011;21:961–973.

FreeBayes. http://bioinformatics.bc.edu/marthlab/FreeBayes.

DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 2011;43:491–498.

GSNP. http://jil.genomics.org.cn/index.php/en/software/gsnp.html.

Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 2009;5:e1000529.

Indelocator. https://confluence.broadinstitute.org/display/CGATools/Indelocator.

Ion Variant Hunter. https://github.com/iontorrent/Ion-Variant-Hunter.

Li Y, Willer CJ, Ding J, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 2010;34:816–834.

Lee S, Hormozdiari F, Alkan C, et al. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* 2009;6:473–474.

Chen K, McLellan MD, Ding L, et al. PolyScan: an automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res.* 2007;17:659–666.

Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* 2011;21:952–960.

realSFS. http://jil.genomics.org.cn/index.php/en/software/realsfs.html.

Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.

Malhis N, Jones SJM. High quality SNP calling using Illumina data at shallow coverage. *Bioinformatics* 2010;26:1029–1035.

Simola DF, Kim J. Sniper: improved SNP discovery by multiply mapping deep sequenced reads. *Genome Biol.* 2011;12:R55.

Wei Z, Wang W, Hu P, et al. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 2011;39:e132.

Goya R, Sun MGF, Morin RD, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* 2010;26:730–736.

SOAPindel. http://soap.genomics.org.cn/soapindel.html.

Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* 2011;43:1066–1073.

Dalca AV, Rumble SM, Levy S, et al. VARiD: a variation detection framework for color-space and letter-space platforms. *Bioinformatics* 2010;26:i343–349.

Altmann A, Weber P, Quast C, et al. vipR: variant identification in pooled DNA using R. *Bioinformatics* 2011;27:i77–84.

Ding J, Bashashati A, Roth A, et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* 2012;28:167–175.

MuTect. https://confluence.broadinstitute.org/display/CGATools/MuTect.

SomaticCall. http://www.broadinstitute.org/science/programs/genome-biology/computational-rd/somaticcall-manual.

Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 2012;28:311–317.

Vallania FLM, Druley TE, Ramos E, et al. High-throughput discovery of rare insertions and deletions in large cohorts. *Genome Res.* 2010;20:1711–1718.

Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22:568–576.

Ivakhno S, Royce T, Cox AJ, et al. CNAseg--a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* 2010;26:3051–3058.

Medvedev P, Fiume M, Dzamba M, et al. Detecting copy number variation with mated short reads. *Genome Res.* 2010;20:1613–1622.

cnvHMM. http://genome.wustl.edu/software/cnvhmm.

Abyzov A, Urban AE, Snyder M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 2011;21:974–984.

Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 2009;10:80.

Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012;28:1307–1313.

Waszak SM, Hasin Y, Zichner T, et al. Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput. Biol.* 2010;6:e1000988.

Sathirapongsasuti JF, Lee H, Horst BAJ, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011;27:2648–2654.

Yoon S, Xuan Z, Makarov V, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 2009;19:1586–1592.

Miller CA, Hampton O, Coarfa C, et al. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS ONE* 2011;6:e16327.

Chiang DY, Getz G, Jaffe DB, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 2009;6:99–103.

Ha G, Roth A, Lai D, et al. Integrative analysis of genome-wide loss of heterozygosity and mono-allelic expression at nucleotide resolution reveals disrupted pathways in triple negative breast cancer. *Genome research* 2012;

Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 2009;6:677–681.

Sun R, Love MI, Zemojtel T, et al. Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. *Bioinformatics* 2012;28:1024–1025.

Lam HYK, Mu XJ, Stütz AM, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat. Biotechnol.* 2010;28:47–55.

Clark MJ, Homer N, O'Connor BD, et al. U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet.* 2010;6:e1000832.

Marschall T, Costa I, Canzar S, et al. CLEVER: Clique-Enumerating Variant Finder. *arXiv:1203.0937* 2012;

Suzuki S, Yasuda T, Shiraishi Y, et al. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* 2011;12 Suppl 14:S7.

Wang J, Mullighan CG, Easton J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* 2011;8:652–654.

Ge H, Liu K, Juan T, et al. FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* 2011;27:1922–1928.

Sindi SS, Onal S, Peng L, et al. An integrative probabilistic model for identification of structural variation in sequencing data. *Genome biology* 2012;13:R22.

Quinlan AR, Clark RA, Sokolova S, et al. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res.* 2010;20:623–635.

Korbel JO, Abyzov A, Mu XJ, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* 2009;10:R23.

Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;25:2865–2871.

Karakoc E, Alkan C, O'Roak BJ, et al. Detection of structural variants and indels within exome data. *Nat. Methods* 2012;9:176–178.

Zeitouni B, Boeva V, Janoueix-Lerosey I, et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* 2010;26:1895–1896.

Wong K, Keane TM, Stalker J, et al. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol.* 2010;11:R128.

Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;470:59–65.

Hormozdiari F, Hajirasouliha I, Dao P, et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 2010;26:i350–357.

Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 2012;30:413–421.

Mathe E, Olivier M, Kato S, et al. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res.* 2006;34:1317–1325.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38:e164.

Makarov V, O'Grady T, Cai G, et al. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics* 2012;28:724–725.

Masso M, Vaisman II. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. *Protein Eng. Des. Sel.* 2010;23:683–687.

Schmitt AO, Assmus J, Bortfeldt RH, et al. CandiSNPer: a web tool for the identification of candidate SNPs for causal variants. *Bioinformatics* 2010;26:969–970.

Wong WC, Kim D, Carter H, et al. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* 2011;27:2147–2148.

Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* 2006;34:W239–242.

Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 2011;32:894–899.

McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010;26:2069–2070.

Cartegni L, Wang J, Zhu Z, et al. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* 2003;31:3568–3571.

Fairbrother WG, Yeh R-F, Sharp PA, et al. Predictive identification of exonic splicing enhancers in human genes. *Science* 2002;297:1007–1013.

Liu C-K, Chen Y-H, Tang C-Y, et al. Functional analysis of novel SNPs and mutations in human and mouse genomes. *BMC Bioinformatics* 2008;9 Suppl 12:S10.

Yuan H-Y, Chiou J-J, Tseng W-H, et al. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.* 2006;34:W635–641.

Kang HJ, Choi KO, Kim B-D, et al. FESD: a Functional Element SNPs Database in human. *Nucleic Acids Res.* 2005;33:D518–522.

Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 2002;320:369–387.

Lee PH, Shatkay H. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.* 2008;36:D820–824.

Davydov EV, Goode DL, Sirota M, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 2010;6:e1001025.

Mermel CH, Schumacher SE, Hill B, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 2011;12:R41.

Venselaar H, Te Beek TAH, Kuipers RKP, et al. Protein structure analysis of mutations causing inheritable diseases. An *e*-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* 2010;11:548.

Desmet F-O, Hamroun D, Lalande M, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 2009;37:e67.

Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 2005;33:W306–310.

Karchin R, Diekhans M, Kelly L, et al. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* 2005;21:2814–2820.

Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* 2005;15:978–986.

Wainreb G, Ashkenazy H, Bromberg Y, et al. MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural data. *Nucleic Acids Res.* 2010;38:W523–528.

Stoyanovich J, Pe'er I. MutaGeneSys: estimating individual disease susceptibility based on genome-wide SNP array data. *Bioinformatics* 2008;24:440–442.

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011;39:e118.

Schwarz JM, Rödelsperger C, Schuelke M, et al. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* 2010;7:575–576.

Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 2009;25:2744–2750.

Banerji S, Cibulskis K, Rangel-Escareno C, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* 2012;486:405–409.

Grant JR, Arantes AS, Liao X, et al. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* 2011;27:2300–2301.

Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.* 2005;33:W480–482.

Oncotator. http://www.broadinstitute.org/oncotator/.

Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 2003;13:2129–2141.

Tian J, Wu N, Guo X, et al. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics* 2007;8:450.

Zhang XH-F, Kangsamaksin T, Chao MSP, et al. Exon inclusion is dependent on predictable exonic splicing enhancers. *Mol. Cell. Biol.* 2005;25:7323–7332.

Wang J, Ronaghi M, Chong SS, et al. pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. *Hum. Mutat.* 2011;32:19–24.

Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinformatics* 2011;12:41–51.

Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 2006;22:2729–2734.

Ferrer-Costa C, Gelpí JL, Zamakola L, et al. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 2005;21:3176–3178.

Jegga AG, Gowrisankar S, Chen J, et al. PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. *Nucleic Acids Res.* 2007;35:D700–706.

Freimuth RR, Stormo GD, McLeod HL. PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. *Hum. Mutat.* 2005;25:110–117.

Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat. Methods* 2010;7:248–249.

Conde L, Vaquerizas JM, Santoyo J, et al. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.* 2004;32:W242–248.

Grover D, Woodfield AS, Verma R, et al. QuickSNP: an automated web server for selection of tagSNPs. *Nucleic Acids Res.* 2007;35:W115–120.

Yeo G, Hoon S, Venkatesh B, et al. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl. Acad. Sci. U.S.A.* 2004;101:15700–15705.

Ye Z-Q, Zhao S-Q, Gao G, et al. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics* 2007;23:1444–1450.

Gamazon ER, Zhang W, Konkashbaev A, et al. SCAN: SNP and copy number annotation. *Bioinformatics* 2010;26:259–262.

Asthana S, Roytberg M, Stamatoyannopoulos J, et al. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.* 2007;3:e254.

Seattle Seq Annotation. http://snp.gs.washington.edu/SeattleSeqAnnotation/.

Shetty AC, Athri P, Mondal K, et al. SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics* 2010;11:471.

Capriotti E, Arbiza L, Casadio R, et al. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Hum. Mutat.* 2008;29:198–204.

Ge D, Ruzzo EK, Shianna KV, et al. SVA: software for annotating and visualizing sequenced human genomes. *Bioinformatics* 2011;27:1998–2000.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;4:1073–1081.

Hu J, Ng PC. Predicting the effects of frameshifting indels. *Genome biology* 2012;13:R9.

Garber M, Guttman M, Clamp M, et al. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 2009;25:i54–62.

Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* 2007;35:3823–3835.

Wang P, Dai M, Xuan W, et al. SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics* 2006;22:e523–529.

Han A, Kang HJ, Cho Y, et al. SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. *Nucleic Acids Res.* 2006;34:W642–644.

Schaefer C, Meier A, Rost B, et al. SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics* 2012;28:601–602.

De Baets G, Van Durme J, Reumers J, et al. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res.* 2012;40:D935–939.

Wang L, Liu S, Niu T, et al. SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. *BMC Bioinformatics* 2005;6:60.

Chelala C, Khan A, Lemoine NR. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* 2009;25:655–661.

Riva A, Kohane IS. A SNP-centric database for the investigation of the human genome. *BMC Bioinformatics* 2004;5:33.

Calabrese R, Capriotti E, Fariselli P, et al. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* 2009;30:1237–1244.

Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 2006;7:166.

SNPseek. http://snp.wustl.edu/cgi-bin/SNPseek/index.cgi.

Xu H, Gregory SG, Hauser ER, et al. SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics* 2005;21:4181–4186.

Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front Genet* 2012;3:35.

Saccone SF, Bolze R, Thomas P, et al. SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. *Nucleic Acids Res.* 2010;38:W201–209.

Uzun A, Leslin CM, Abyzov A, et al. Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. *Nucleic Acids Res.* 2007;35:W384–392.

Hemminger BM, Saelim B, Sullivan PF. TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics* 2006;22:626–627.

Stitziel NO, Binkowski TA, Tseng YY, et al. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.* 2004;32:D520–522.

Medina I, De Maria A, Bleda M, et al. VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing. *Nucleic Acids Res.* 2012;40:W54–58.

Kong L, Wang J, Zhao S, et al. ABrowse--a customizable next-generation genome browser framework. *BMC Bioinformatics* 2012;13:2.

Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008;133:523–536.

Lee E, Harris N, Gibson M, et al. Apollo: a community resource for genome annotation editing. *Bioinformatics* 2009;25:1836–1837.

Engels R, Yu T, Burge C, et al. Combo: a whole genome comparative browser. *Bioinformatics* 2006;22:1782–1783.

Carver T, Harris SR, Berriman M, et al. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* 2012;28:464–469.

Carver T, Harris SR, Otto TD, et al. BamView: visualizing and interpretation of next-generation sequencing read alignments. *Briefings in bioinformatics* 2012;

Gordon D. Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinformatics* 2003;Chapter 11:Unit11.2.

Friedel M, Nikolajewa S, Sühnel J, et al. DiProGB: the dinucleotide properties genome browser. *Bioinformatics* 2009;25:2603–2604.

Huang W, Marth G. EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res.* 2008;18:1538–1543.

Spudich GM, Fernández-Suárez XM. Touring Ensembl: a practical guide to genome browsing. *BMC Genomics* 2010;11:295.

Bare JC, Koide T, Reiss DJ, et al. Integration and visualization of systems biology data in context of the genome. *BMC Bioinformatics* 2010;11:382.

Bonfield JK, Whitwham A. Gap5--editing the billion fragment sequence assembly. *Bioinformatics* 2010;26:1699–1703.

Donlin MJ. Using the Generic Genome Browser (GBrowse). *Curr Protoc Bioinformatics* 2009;Chapter 9:Unit 9.9.

Kawahara Y, Sakate R, Matsuya A, et al. G-compass: a web-based comparative genome browser between human and other vertebrate genomes. *Bioinformatics* 2009;25:3321–3322.

Huntley D, Tang YA, Nesterova TB, et al. Genome Environment Browser (GEB): a dynamic browser for visualising high-throughput experimental data in the context of genome features. *BMC Bioinformatics* 2008;9:501.

Abeel T, Van Parys T, Saeys Y, et al. GenomeView: a next-generation genome browser. *Nucleic Acids Res.* 2012;40:e12.

Schatz MC, Phillippy AM, Sommer DD, et al. Hawkeye and AMOS: visualizing and assessing the quality of genome assemblies. *Briefings in bioinformatics* 2011;

Nicol JW, Helt GA, Blanchard SG Jr, et al. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics* 2009;25:2730–2731.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* 2012;

Waterhouse AM, Procter JB, Martin DMA, et al. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;25:1189–1191.

Westesson O, Skinner M, Holmes I. Visualizing next-generation sequencing data with JBrowse. *Briefings in bioinformatics* 2012;

Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. *Genome Res.* 2009;19:2125–2132.

Hou H, Zhao F, Zhou L, et al. MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. *Nucleic Acids Res.* 2010;38:W732–736.

Bao H, Guo H, Wang J, et al. MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics* 2009;25:1554–1555.

Arner E, Hayashizaki Y, Daub CO. NGSView: an extensible open source editor for next-generation sequencing data. *Bioinformatics* 2010;26:125–126.

Popendorf K, Sakakibara Y. SAMSCOPE: an OpenGL-based real-time interactive scale-free SAM viewer. *Bioinformatics* 2012;28:1276–1277.

Fiume M, Williams V, Brook A, et al. Savant: genome browser for high-throughput sequencing data. *Bioinformatics* 2010;26:1938–1944.

Ganesan H, Rakitianskaia AS, Davenport CF, et al. The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* 2008;9:333.

Jung K, Park J, Choi J, et al. SNUGB: a versatile genome browser supporting comparative and functional fungal genomics. *BMC Genomics* 2008;9:586.

Milne I, Bayer M, Cardle L, et al. Tablet--next generation sequence assembly visualization. *Bioinformatics* 2010;26:401–402.

Sanborn JZ, Benz SC, Craft B, et al. The UCSC Cancer Genomics Browser: update 2011. *Nucleic Acids Res.* 2011;39:D951–959.

Dreszer TR, Karolchik D, Zweig AS, et al. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* 2012;40:D918–923.

Saito TL, Yoshimura J, Sasaki S, et al. UTGB toolkit for personalized genome browsers. *Bioinformatics* 2009;25:1856–1861.

Loveland J. VEGA, the genome browser with a difference. *Brief. Bioinformatics* 2005;6:189–193.

Dubchak I. Comparative analysis and visualization of genomic sequences using VISTA browser and associated computational tools. *Methods Mol. Biol.* 2007;395:3–16.

Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009; 19:1639–1645.

O'Brien TM, Ritz AM, Raphael BJ, et al. Gremlin: an interactive visualization model for analyzing genomic rearrangements. *IEEE Trans Vis Comput Graph* 2010;16:918–926.

Bcbio-nextgen. https://github.com/chapmanb/bcbb/blob/master/nextgen/README.md.

Langmead B, Schatz MC, Lin J, et al. Searching for SNPs with cloud computing. *Genome Biol.* 2009;10:R134.

Sana ME, Iascone M, Marchetti D, et al. GAMES identifies and annotates mutations in next-generation sequencing projects. *Bioinformatics* 2011;27:9–13.

Lam HYK, Pan C, Clark MJ, et al. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat. Biotechnol.* 2012;30:226–229.

Qi J, Zhao F, Buboltz A, et al. inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics* 2010;26:127–129.

MutationTaster. http://www.mutationtaster.org/NextGenerationSequencing.html.

Blanca JM, Pascual L, Ziarsolo P, et al. ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using next generation sequence. *BMC Genomics* 2011;12:285.

RTG. http://www.realtimegenomics.com/.

Deng X. SeqGene: a comprehensive software solution for mining exome- and transcriptome- sequencing data. *BMC Bioinformatics* 2011;12:267.

Ossowski S, Schneeberger K, Clark RM, et al. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.* 2008;18:2024–2033.

Fischer M, Snajder R, Pabinger S, et al. SIMPLEX: Cloud-Enabled Pipeline for the Comprehensive Analysis of Exome Sequencing Data. *PLoS ONE* 2012;7:e41948.

Asmann YW, Middha S, Hossain A, et al. TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. *Bioinformatics* 2012;28:277–278.

Orvis J, Crabtree J, Galens K, et al. Ergatis: a web interface and scalable software system for bioinformatics workflows. *Bioinformatics* 2010;26:1488–1492.

Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010;11:R86.

Genboree. www.genboree.org/.

Reich M, Liefeld T, Gould J, et al. GenePattern 2.0. *Nat. Genet.* 2006;38:500–501.

Halbritter F, Vaidya HJ, Tomlinson SR. GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods* 2012;9:7–8.

Kepler. https://kepler-project.org/.

Jagla B, Wiswedel B, Coppée J-Y. Extending KNIME for next-generation sequencing data analysis. *Bioinformatics* 2011;27:2907–2909.

Rex DE, Ma JQ, Toga AW. The LONI Pipeline Processing Environment. *Neuroimage* 2003;19:1033–1048.

Moa. http://mfiers.github.com/Moa/.

Abouelhoda M, Issa SA, Ghanem M. Tavaxy: Integrating Taverna and Galaxy workflows with cloud computing support. *BMC bioinformatics* 2012;13:77.

Hull D, Wolstencroft K, Stevens R, et al. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.* 2006;34:W729–732.

Hunter AA, Macgregor AB, Szabo TO, et al. Yabi: An online research environment for grid, high performance and cloud computing. *Source Code Biol Med* 2012;7:1.

Altschul S, Gish W, Miller W, Myers E, Lipman D: Basic local alignment search tool. J Mol Biol 1990, 215(3):403-410.

Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, Lipman D: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389-3402.

NCBI C toolkit [http://www.ncbi.nlm.nih.gov/IEB/ToolBox/SDK DOCS/INDEX.HTML]

Zhang Z, Schäffer A, Miller W, Madden T, Lipman D, Koonin E, Altschul S: Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res 1998, 26(17):3986-3990.

Schäffer A, Wolf Y, Ponting C, Koonin E, Aravind L, Altschul S: IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 1999, 15(12):1000-1011.

Schäffer A, Aravind L, Madden T, Shavirin S, Spouge J, Wolf Y, Koonin E, Altschul S: Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001, 29(14):2994-3005.

Zhang Z, Schwartz S, Wagner L, Miller W: A greedy algorithm for aligning DNA sequences. J Comput Biol 7(1-2):203-214.

Werner, T. (1999) Models for prediction and recognition of eukaryotic promoters. Mamm. Genome 10, 168–175

Frisch, M., Frech, K., Klingenhoff, A., Cartharius, K., Liebich, I., and Werner, T. (2002) In silico prediction of scaffold/matrix attachment regions in large genomic sequences. Genome Res. 12, 349–354

Robertson, K. D. (2002) DNA methylation and chromatin— unraveling the tangled web. Oncogene 21, 5361–5379

Fessele, S., Maier, H., Zischek, C., Nelson, P. J., and Werner, T. (2002) Regulatory context is a crucial part of gene function. Trends Genet. 18, 60–63

Werner, T. (2001) The promoter connection. Nat. Genet. 29, 105–106

Chen, Q. K., Hertz, G. Z., and Stormo, G. D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput. Appl. Biosci. 11, 563–566

Kel, A. E., Kondrakhin, Y. V., Kolpakov Ph, A., Kel, O. V., Romashenko, A. G., Wingender, E., Milanesi, L., and Kolchanov, N. A. (1995) Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 3, 197–205

Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., and Wingender, E. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. Nucleic Acids Res. 27, 318–322

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 23, 4878–4884

Bullard JH, Purdom E, Hansen KD, Dudoit S: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinforma 2010, 11:94.

Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biol 2010, 11:R106.

Robinson MD, Oshlack A: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 2010, 11:R25.

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F: A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 2012. doi:10.1093/bib/bbs046. epub ahead of print.

Smyth GK: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004, 3:Article 3.

Auer PL: Srivastava S. Doerge RW: Differential expression - the next generation and beyond. Brief Funct Genomics; 2011.

Robinson MD, Smyth GK: Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 2008, 9:321–332.

Auer PL, Doerge RW: A two-stage poisson model for testing RNA-seq data. Stat Appl Gen Mol Biol 2011, 10:Article 26.

Hardcastle TJ, Kelly KA: baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinforma 2010, 11:422.

Di Y, Schafer DW, Cumbie JS, Chang JH: The NBP negative binomial model for assessing differential gene expression from RNA-seq. Stat Appl Genet Mol Biol 2011, 10:Article 24.

Zhou Y-H, Xia K, Wright FA: A powerful and flexible approach to the analysis of RNA sequence count data. Bioinformatics 2011, 27(19):2672–2678.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008, 18(9):1509–1517.

Bancroft T, Nettleton D: Estimation of false discovery rate using permutation p-values with different discrete null distributions. Technical Report: Iowa State University; 2009 [www.stat.iastate.edu/preprint/articles/2009-05.pdf]

Bottomly D, Walter NA, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzermann R: Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PLoS One 2011, 6(3):e17820.

Robinson MD, Smyth GK: Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007, 23:2881–2887.

Rue H, Martino S, Chopin N: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Statist Soc B 2009, 71(2):319–392.

Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y: Sex-specific and lineage-specific alternative splicing in primates. Genome Res 2010, 20(2):180–189.

G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: a system for large-scale graph processing," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010, pp. 135–146.

K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE, 2010, pp. 1–10.

M. Ovsiannikov, S. Rus, D. Reeves, P. Sutter, S. Rao, and J. Kelly, "The quantcast file system," Proceedings of the VLDB Endowment, vol. 6, no. 11, pp. 1092–1101, 2013.

S. Owen, R. Anil, T. Dunning, and E. Friedman, Mahout in action. Manning, 2011.

J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, and G. Fox, "Twister: a runtime for iterative mapreduce," in Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, 2010, pp. 810–818.

M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, 2012, pp. 2–2.

Y. Low, J. E. Gonzalez, A. Kyrola, D. Bickson, C. E. Guestrin, and J. Hellerstein, "Graphlab: A new framework for parallel machine learning," arXiv preprint arXiv:1408.2041, 2014.

W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A highperformance, portable implementation of the mpi message passing interface standard," Parallel computing, vol. 22, no. 6, pp. 789– 828, 1996.

C. M. Bishop et al., Pattern recognition and machine learning. springer New York, 2006, vol. 4, no. 4.

F. Nei, Y. Huang, X. Wang, and H. Huang, "New primal svm solver with linear computational cost for big data classifications," in Proceedings of the 31st international conference on Machine Learning. JMLR, 2014, pp. 1–9.

S. Haller, S. Badoud, D. Nguyen, V. Garibotto, K. Lovblad, and P. Burkhard, "Individual detection of patients with parkinson disease using support vector machine analysis of diffusion tensor imaging data: initial results," American Journal of Neuroradiology, vol. 33, no. 11, pp. 2123–2128, 2012.

D. Giveki, H. Salimi, G. Bahmanyar, and Y. Khademian, "Automatic detection of diabetes diagnosis using feature weighted support vector machines based on mutual information and modified cuckoo search," arXiv preprint arXiv:1201.2173, 2012.

S. Bhatia, P. Prakash, and G. Pillai, "Svm based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features," in Proceedings of the World Congress on Engineering and Computer Science, WCECS, 2008, pp. 22–24.

Y.-J. Son, H.-G. Kim, E.-H. Kim, S. Choi, and S.-K. Lee, "Application of support vector machine for prediction of medication adherence in heart failure patients," Healthcare informatics research, vol. 16, no. 4, pp. 253–259, 2010.

J. Ye, J.-H. Chow, J. Chen, and Z. Zheng, "Stochastic gradient boosted distributed decision trees," in Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009, pp. 2061–2064.

D. Borthakur, "The hadoop distributed file system: Architecture and design," Hadoop Project Website, vol. 11, no. 2007, p. 21, 2007.

R. Calaway, L. Edlefsen, L. Gong, and S. Fast, "Big data decision trees with r," Revolution.

L. O. Hall, N. Chawla, and K. W. Bowyer, "Decision tree learning on very large data sets," in Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on, vol. 3. IEEE, 1998, pp. 2579–2584.

Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009, 10, 57-63.

Park, P.J.. ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 2009, 10, 669-680.

Chiang, D.Y.; Getz, G; Jaffe, D.B.; O'Kelly, M.J.T.; Zhao, X; Carter,S.L.; Russ, C.; Nusbaum, C.; Meyerson, M.; Lander, E.S. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat. Methods 2009, 6, 99-103.

Alkan, C.; Kidd, J.M.; Marques-Bonet, T.; Aksay, G.; Antonacci, F.; Hormozdiari, F.; Kitzman, J.O.; Baker, C.; Malig, M.; Mutlu, O.; et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat. Genet. 2009, 41, 1061-1067.

Campbell, P.J.; Stephens, P.J.; Pleasance, E.D.; O'Meara, S.; Li, H.; Santarius, T.; Stebbings, L.A.; Leroy, C.; Edkins, S.; et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat. Genet. 2008, 40, 722-729.

Hyman, E.D. A new method of sequencing DNA. Anal. Biochem. 1988, 174, 423–436.

454 Home Page. http://www.454.com/indecx.asp (accessed on 27 August 2010).

Fedurco, M.; Romieu, A.; Williams, S.; Lawrence, I.; Turcatti, G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 2006, 34, e22.

Turcatti, G.; Romieu, A.; Fedurco, M.; Tairi, A.P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res. 2008, 36, e25.

Adessi, C.; Matton, G.; Ayala, G.; Turcatti, G.; Mermod, J.J.; Mayer, P.; Kawashima, E. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res. 2000, 28, e87.

Solexa Home Page. http://www.solexa.com/ (accessed on 27 August 2010).

Shendure, J.; Porreca, G.J.; Reppas, N.B.; Lin, X.; McCutcheon, J.P.; Rosenbaum, A.M.; Wang, M.D.; Zhang, K.; Mitra, R.D.; Church, G.M. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005, 309, 1728–1732.

McKernan, K.; Blanchard, A.; Kotler, L.; Costa, G. Reagents, methods, and libraries for beadbased sequencing. US patent application 20080003571 2006.

Applied Biosystems Home Page. www3.appliedbiosystems.com/index.htm (accessed on 27 August 2010).

Jett, J.H.; Keller, R.A.; Martin, J.C.; Marrone, B.L.; Moyzis, R.K.; Ratliff, R.L.; Seitzinger, N.K.; Shera, E.B.; Stewart. C.C. High-speed DNA sequencing: an approach based upon fluorescence detection of single molecules. J. Biomol. Struct. Dyn. 1989, 7, 301-309.

Helicos Home Page. http://www.helicosbio.com/ (accessed on 27 August 2010).

Pushkarev, D.; Neff, N.F.; Quake, S.R. Single-molecule sequencing of an individual human genome. Nat. Biotechnol. 2009, 27, 847–852.

Metzker, M.L. Sequencing technologies – the next generation. Nat. Rev. Genet. 2010, 11, 31–46.

Kent, W.J. BLAT--the BLAST-like alignment tool. Genome Res. 2002, 4, 656-664.

Li, H.; Ruan, J.; Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008, 11, 1851-1858.

Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009, 3, R25.

Ning, Z.; Cox, A.J.; Mullikin, J.C. SSAHA: a fast search method for large DNA databases. Genome Res. 2001, 11, 1725–1729.

Li H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010, 5, 589-595.

Salama RA, Stekel DJ (2010) Inclusion of neighboring base interdependencies substantially improves genome-wide prokaryotic transcription factor binding site prediction. Nucleic acids research 38: e135.

Mehta P, Schwab DJ, Sengupta AM (2011) Statistical Mechanics of Transcription-Factor Binding Site Discovery Using Hidden Markov Models. Journal of statistical physics 142: 1187–1205.

Marinescu VD, Kohane IS, Riva A (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC bioinformatics 6: 79.

Marinescu VD, Kohane IS, Riva A (2005) The MAPPER database: a multigenome catalog of putative transcription factor binding sites. Nucleic acids research 33: D91–7.

Raman R, Overton GC (1994) Application of hidden Markov modeling to the characterization of transcription factor binding sites. In: Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences HICSS94. IEEE Comput. Soc. Press, pp. 275–283.

Durbin R, Edddy S, Krogh A, Mitchison G (1998) Biological sequence analysis Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.

Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, et al. (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics (Oxford, England) 21: 2657–66.

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein- DNA interactions. Science (New York, NY) 316: 1497–502.

ENCODE (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science (New York, NY) 306: 636–40.

Segal E, Fondufe-Mittendorf Y, Chen L, Tha˚ stro¨m A, Field Y, et al. (2006) A genomic code for nucleosome positioning. Nature 442: 772–8.

Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic acids research 18: 6097–6100.

Schuster-Bo¨ckler B, Schultz J, Rahmann S (2004) HMM Logos for visualization of protein families. BMC bioinformatics 5: 7.

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, pp. 28– 36.

Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics (Oxford, England) 14: 48–54.

Wilbanks EG, Facciotti MT (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PloS one 5: e11471.

Wilcoxon F (1945) Individual Comparisons by Ranking Methods. Biometrics Bulletin 1: 80–83.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289–300.

Benos PV, Lapedes AS, Stormo GD (2002) Probabilistic Code for DNA Recognition by Proteins of the EGR Family. Journal of Molecular Biology 323: 701–727.

Benos PV, Lapedes AS, Fields D, Stormo GD (2001) SAMIE: Statistical algorithm for modeling interaction energies. In: Pacific Symposium on Biocomputing. volume 126, pp. 6:115–126.

Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein-DNA interactions: how good an approximation is it? Nucleic acids research 30: 4442–51.

Spearman C (1904) The proof and measurement of association between two things. American Journal of Psychology : 72–101.

Li, R.; Yu, C.; Li, Y.; Lam, T.; Yiu, S.; Kristiansen, K.; Wang, J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 2009, 15, 1966-1967.

Zerbino, D.R.; Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome. Res. 2008, 5, 821-829.

Li, R.; Li, Y.; Kristiansen, K.; Wang, J. SOAP: short oligonucleotide alignment program. Bioinformatics 2008, 5, 713-714.

Pevzner, P.A.; Borodovsky, M.Y.; Mironov, A.A. Linguistics of nucleotide sequences. II: Stationary words in genetic texts and the zonal structure of DNA. J. Biomol. Struct. Dyn. 1989, 6, 1027–1038.

Marth, G.T.; Korf, I.; Yandell, M.D.; Yeh, R.T.; Gu, Z.; Zakeri, H.; Stitziel, N.O.; Hillier, L.; Kwok, P.Y.; Gish W.R. A general approach to single-nucleotide polymorphism discovery. Nat. Genet. 1999, 23, 452–456.

Laczik M, Tukacs E, Uzonyi B, et al. Geno viewer, a SAM/BAM viewer tool. *Bioinformation* 2012;8:107–109.