GeNeViSTA

Exome Data Analysis for Clinicians: How & Why

Aneek Das Bhowmik and Ashwin Dalal

Diagnostics Division, Centre for DNA Fingerprinting and Diagnostics, Hyderabad

Email: adalal@cdfd.org.in

Introduction

Exome sequencing is a next generation sequencing (NGS) technology where only the coding regions of the genome (known as exome) are sequenced. The technology involves two major steps: capturing the exonic region of DNA that encodes proteins, constituting about 1% of the human genome (≈30 Mbps) and then sequencing these regions using high throughput massively parallel DNA sequencing technology. The idea is to identify the genetic variants that alter protein sequences at a much lower cost than whole genome sequencing (WGS).

Exome sequencing is especially effective in studying rare Mendelian disorders and/or single gene disorders (Bamshad et al., 2011). Single gene disorders are rare by themselves but collectively they are an important cause of morbidity and mortality. The identification of causal genetic variants for these disorders has important value in prenatal diagnosis and genetic counseling of affected families. These disorders are mostly caused by very rare genetic variants that are present in a small number of individuals, occurring sporadically or exhibiting phenotypic heterogeneity (Ng et al., 2009; Ng et al., 2010). Novel gene identification for rare diseases is difficult using classical methods like cytogenetic mapping, linkage analysis and homozygosity mapping. Different NGS strategies have made it possible to quickly identify the cause for single gene disorders using a few affected individuals, which can be used for identification of novel genes on a research basis and identification of mutations in known genes for single gene disorders in clinical practice. However, these sequencing strategies produce a large amount of data which needs to be interpreted using computational methods. Since most of the severe disease causing variants in single gene disorders are clustered within the exons or the protein coding regions of human genome (≈85%), sequencing just the exome significantly decreases the amount of sequencing to be done and the data to be interpreted, which in turn reduces the cost when compared to WGS and yet is effective in most of the single gene disorders for diagnosis and new gene discovery.

Exome sequencing data analysis

With availability of NGS technologies, the main challenge is in computational analysis to identify the causative variant and to differentiate between disease-causing variants and polymorphisms. A large quantity of data and sequence information is generated which requires a significant amount of data analysis. Various sequence technologies also have different error rates and generate various read-lengths which can pose challenges in comparing results from different sequencing platforms. Whole exome sequencing (WES) data analysis can be divided in these following major steps:

- Base-calling and image analysis - output as raw reads (FASTA/FASTQ files)
- Quality checking of FASTA/FASTQ files and data pre-processing
- Mapping and alignment with reference genome - output as SAM/BAM files
- Data processing and variant calling - output as VCF files
- Variant annotation - output as list of variant file usually in text or excel format

The first step is mostly done within the laboratories of the sequence service provider or contract research organizations (CROs). The trend of rapidly decreasing cost of exome sequencing has made it possible to quickly outsource the samples to different CROs for sequencing. Outsourcing can save valuable resources by assuring that time and money are not dissipated on unnecessary or failed work and wasted samples as the wet-lab (library preparation to sequencing) part of NGS needs specific expertise in molecular biology. After sequencing and performing the basic bioinformatic analysis of the sequences mentioned above, CROs will generally hand over the results or final outputs to the host clinics/laboratories. As per the request or at higher cost they will also share the raw files and metadata of sequencing. However, the challenge remains in correctly analyzing the data to identify the responsible disease-causing variants correlating with the clinical features of the patient. Hence it is always good and rather cost-effective to have at least some basic knowledge about the steps of exome data analysis and interpreting the test results correctly.

Table 1 Basic file formats of exome sequencing.

File formats and terminology	Brief description
FASTA	The FASTA format, generally indicated with the suffix .fa or .fasta, is a straightforward, human readable format. Normally, each file consists of a set of sequences, where each sequence is represented by a one line header, starting with the '>' character, followed by the corresponding nucleotide sequence.
FASTQ	FASTQ is a text file format (human readable) that consists raw sequence reads and its corresponding quality information. It normally provides 4 lines of data per sequence: sequence identifier, the raw sequence, comments (optional), and quality values for the sequence. FASTQ format is commonly used to store sequencing reads, in particular from Illumina and Ion Torrent platforms. Paired-end reads may be stored either in one FASTQ file (alternating) or in two different FASTQ files.
SAM	SAM stands for Sequence Alignment MAP format. It is a tab delimited text format which stores the mapped or aligned (with reference human genome) sequences. It contains an optional header (typically starts with @) followed by alignment section which contains 12 columns with essential alignment information such as reference sequence name, mapping position, mapping quality, aligner specific information etc. and the aligned sequence reads.
BAM	BAM stands for Binary Alignment MAP format in which aligned sequences stored in a compressed, indexed and binary form. It provides the binary versions of most of the same data stored in SAM file and is designed to compress reasonably well. Hence the size of the BAM file becomes much less than the SAM file (easy to store).
VCF	The Variant Calling Format or VCF specifies the format of a text file where all the variant information of the sequences are stored in a compressed manner. VCF file starts with a header section which contains metadata describing the body of the file (denoted as starting with ##) followed by 9 tab-separated specific columns containing information about the variant position in the genome and other columns with genotype information on samples for each position.
BED	BED stands for Browser Extensible Display format. This format is used for describing genes and other features of DNA sequences. For exome sequencing, usually the genomic regions covered by different commercially available exome capture kits is provided in bed files. These files are freely available in the company websites (e.g. `https://earray.chem.agilent.com/suredesign/`). Upon request CROs also provide the bedfiles. These bedfiles are useful to restrict the exome data analysis only into the specific regions covered in exome sequencing, if provided in variant calling steps.

Table 2 Different widely used freely available analytical tools (popular) to perform exome data analysis.

Purpose	Softwares/Programs	Source
Data quality checking and trimming	FastQC (raw fastq & BAM files)	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
	NGSrich (BAM files only)	https://sourceforge.net/projects/ngsrich/files/
Sequence alignment (mapping)	BWA	http://bio-bwa.sourceforge.net/bwa.shtml
	Novoalign	http://www.novocraft.com/products/novoalign/
	Stampy	http://www.well.ox.ac.uk/stampy
	Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
	mrsFAST	http://sfu-compbio.github.io/mrsfast/
	NextGenMap	http://cibiv.github.io/NextGenMap/
Data processing	Picardtools	https://broadinstitute.github.io/picard/
	SAMtools	http://samtools.sourceforge.net/
Sequence visualization	IGV	http://software.broadinstitute.org/software/igv/
Variant calling (SNV/Indel)	GATK	https://software.broadinstitute.org/gatk/
	SAMtools	http://samtools.sourceforge.net/
	Platypus	http://www.well.ox.ac.uk/platypus
	Freebayes	https://github.com/ekg/freebayes
	SNVer	http://snver.sourceforge.net/
	VarScan 2	http://varscan.sourceforge.net/
Variant calling (CNV)	Conifer	http://conifer.sourceforge.net/
	ExomeCNV	https://secure.genome.ucla.edu/index.php/ExomeCNV_User_Guide
	CNVnator	https://github.com/abyzovlab/CNVnator
VCF & BED file processing (optional)	BCFtools	https://samtools.github.io/bcftools/bcftools.html
	VCFtools	http://vcftools.sourceforge.net/
	BEDtools	http://bedtools.readthedocs.io/en/latest/
Variant Annotation	Annovar	http://annovar.openbioinformatics.org/en/latest/
	wANNOVAR	http://wannovar.wglab.org/
	SeattleSeq Annotation	http://snp.gs.washington.edu/SeattleSeqAnnotation138/
	AnnTools	http://anntools.sourceforge.net/
	NGS-SNP	http://stothard.afns.ualberta.ca/downloads/NGS-SNP/
	SnpEff	http://snpeff.sourceforge.net/
	VARIANT	http://variant.bioinfo.cipf.es/
	VEP	http://asia.ensembl.org/info/docs/tools/vep/index.html

Exome data analysis for clinicians (Advanced)

As mentioned above, the first step of exome data analysis is mostly done within the laboratories of the sequence service providers, and only the latter four steps are usually performed after getting the raw data (usually in the form of FASTA/FASTQ file). Different file formats generated in NGS analysis are given in Table 1.

Sophisticated high end work stations (computer) and informatics tools are required to perform exome data analysis along with technical skills such as management and storage of huge amount of NGS data and databases. Moreover, the development of a streamlined and automated guidelines/pipelines for data analysis is very important for generating, annotating and analyzing sequence variants (D'Antonio et al., 2013). There are several bioinformatics workflows, personalized to particular NGS applications depending on the type of variation of interest and the technology employed. One universally recommended and widely used such workflow for variant discovery analysis is GATK (Genome Analysis Toolkit) best practices guidelines, developed by Broad Institute, USA (DePristo et al., 2011; Van der Auwera et al., 2013). However, these guidelines are focused largely on data from human whole-genome or whole-exome samples sequenced with Illumina technology, so working with different types of NGS platforms or experimental designs, requires adaptation to certain branches of the workflow, as well as certain parameter selections and values. Further details of these guidelines are available in the following link (`https://software.broadinstitute.org/gatk/best-practices/`). A list of different majorly used freely available analytical tools for different steps of exome data analysis is given in Table 2 (Pabinger et al., 2014). A general exome analysis workflow from raw reads (FASTA/FASTQ) to annotated list of variants (text/excel), is given in Figure 1.

Exome data analysis for clinicians (Basics)

A typical annotated list of variant file looks like the sample file shown in Figure 2 (usually provided by the CROs as the final output file). The main challenge of analyzing these variants in human diseases is to identify disease-related alleles (which may be known or novel) in a large number of non-pathogenic polymorphisms in the genome. Identification of disease-causing variants in rare Mendelian disorders through exome sequencing relies on different filtering steps to reduce the number of candidate genes. Initial filtering is usually done using different public databases like The International HapMap Consortium, 1000 Genomes Project, Exome Variant Server (EVS), Exome Aggregation Consortium (ExAC), Complete Genomics 69 (CG69), dbSNP etc. and in-house population specific databases (if available). Any variant present in these databases with minor allele frequency (MAF) greater than 0.01 can be excluded from further consideration for rare diseases. Only missense, nonsense, splice-site variants, and indels that are found to affect coding regions are used for clinical interpretation. Clinically relevant mutations are then annotated using published variants in literature and a set of variant databases including ClinVar, OMIM and Human Gene Mutation Database. Then on the basis of the mode of inheritance, for example, a recessive/dominant model, the list of candidate variants can be reduced further. An example of different filtering steps is given in Figure 3 (Das Bhowmik et al., 2015).

CROs will also generally provide a BAM (Binary Alignment Map) file (comes along with an index file as .bai) which is the comprehensive mapped raw data of exomes for sequence viewing in a high performance visualization tool like Integrative Genomics Viewer (IGV) and a VCF (Variant Call Format) file which contains exome sequence variations. This VCF file can be used for annotating the variants using various online variant annotation tools like wANNOVAR, SeattleSeq etc.

Utility of exome data analysis: Example case study

A six year old male child born to non-consanguineous parents, was diagnosed with an unexplained overgrowth syndrome. Clinical features were suggestive of Beckwith–Wiedemann syndrome (BWS). The patient was investigated extensively for BWS including karyotype, array comparative genomic hybridization, methylation analysis at IC1 locus and Sanger sequencing of *CDKN1C* gene. Since all the results were normal the patient was taken up for WES.

The sequences obtained after WES were analyzed following GATK best practices guidelines (D'Antonio et al., 2013; DePristo et al., 2011). Vari-

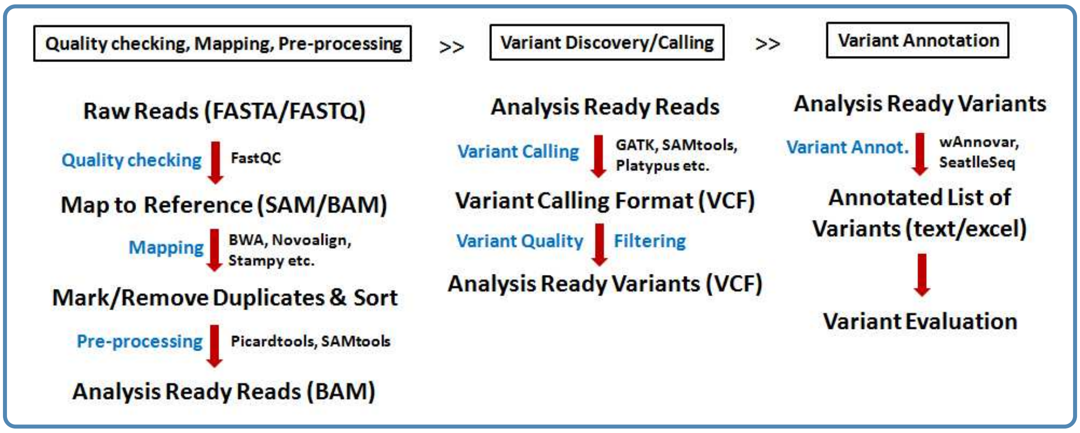
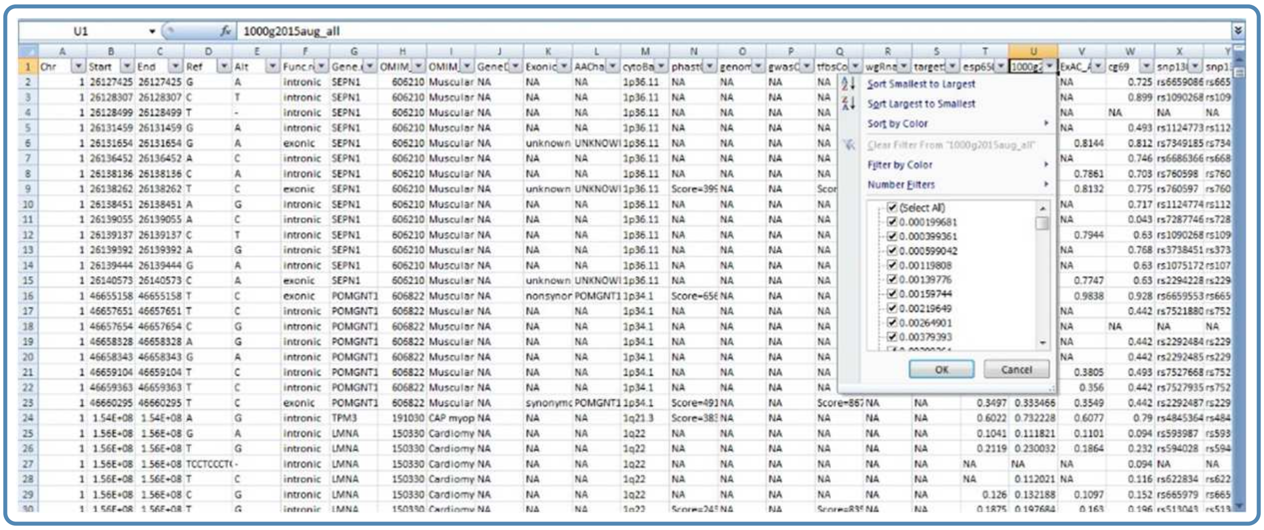

Quality checking, Mapping, Pre-processing
>>
Variant Discovery/Calling
>>
Variant Annotation
Raw Reads (FASTA/FASTQ)
Quality checking
FastQC
Map to Reference (SAM/BAM)
Mapping
BWA, Novoalign, Stampy etc.
Mark/Remove Duplicates & Sort
Pre-processing
Picardtools, SAMtools
Analysis Ready Reads (BAM)
Analysis Ready Reads
Variant Calling
GATK, SAMtools, Platypus etc.
Variant Calling Format (VCF)
Variant Quality
Filtering
Analysis Ready Variants (VCF)
Analysis Ready Variants
Variant Annot.
wAnnovar, SeatlleSeq
Annotated List of Variants (text/excel)
Variant Evaluation

Figure 1 A basic exome analysis workflow using different freely available analytical tools.



U1 fx 1000g2015aug_all

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Chr	Start	End	Ref	Alt	Func.r	Gene.	OMIM_	OMIM_	GeneD	Exonic	AACha	cytoBa	phast	genom	gwasC	tfbsCo	wgRna	target	esp65	1000g2	ExAC_A	cg69	snp13	snp1
2	1	26127425	26127425	G	A	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.725	rs6659086	rs665
3	1	26128307	26128307	C	T	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.899	rs1090268	rs109
4	1	26128499	26128499	T	-	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	NA	NA	NA
5	1	26131459	26131459	G	A	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.493	rs1124773	rs112
6	1	26131654	26131654	G	A	exonic	SEPN1	606210	Muscular	NA	unknown	UNKNOWI	1p36.11	NA	NA	NA	NA					0.8144	0.812	rs7349185	rs734
7	1	26136452	26136452	A	C	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.746	rs6686366	rs668
8	1	26138136	26138136	C	A	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					0.7861	0.703	rs760598	rs760
9	1	26138262	26138262	T	C	exonic	SEPN1	606210	Muscular	NA	unknown	UNKNOWI	1p36.11	Score=39	NA	NA	Scor					0.8132	0.775	rs760597	rs760
10	1	26138451	26138451	A	G	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.717	rs1124774	rs112
11	1	26139055	26139055	A	C	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.043	rs7287746	rs728
12	1	26139137	26139137	C	T	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					0.7944	0.63	rs1090268	rs109
13	1	26139392	26139392	A	G	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.768	rs3738451	rs373
14	1	26139444	26139444	G	A	intronic	SEPN1	606210	Muscular	NA	NA	NA	1p36.11	NA	NA	NA	NA					NA	0.63	rs1075172	rs107
15	1	26140573	26140573	C	A	exonic	SEPN1	606210	Muscular	NA	unknown	UNKNOWI	1p36.11	NA	NA	NA	NA					0.7747	0.63	rs2294228	rs229
16	1	46655158	46655158	T	C	exonic	POMGNT1	606822	Muscular	NA	nonsynor	POMGNT1	1p34.1	Score=65	NA	NA	NA					0.9838	0.928	rs6659553	rs665
17	1	46657651	46657651	T	C	intronic	POMGNT1	606822	Muscular	NA	NA	NA	1p34.1	NA	NA	NA	NA					NA	0.442	rs7521880	rs752
18	1	46657654	46657654	C	G	intronic	POMGNT1	606822	Muscular	NA	NA	NA	1p34.1	NA	NA	NA	NA					NA	NA	NA	NA
19	1	46658328	46658328	A	G	intronic	POMGNT1	606822	Muscular	NA	NA	NA	1p34.1	NA	NA	NA	NA					NA	0.442	rs2292484	rs229
20	1	46658343	46658343	G	A	intronic	POMGNT1	606822	Muscular	NA	NA	NA	1p34.1	NA	NA	NA	NA					NA	0.442	rs2292485	rs229
21	1	46659104	46659104	T	C	intronic	POMGNT1	606822	Muscular	NA	NA	NA	1p34.1	NA	NA	NA	NA					0.3805	0.493	rs7527668	rs752
22	1	46659363	46659363	T	C	intronic	POMGNT1	606822	Muscular	NA	NA	NA	1p34.1	NA	NA	NA	NA					0.356	0.442	rs7527935	rs752
23	1	46660295	46660295	T	C	exonic	POMGNT1	606822	Muscular	NA	synonymc	POMGNT1	1p34.1	Score=491	NA	NA	Score=867	NA	NA	0.3497	0.333466	0.3549	0.442	rs2292487	rs229
24	1	1.54E+08	1.54E+08	A	G	intronic	TPM3	191030	CAP myop	NA	NA	NA	1q21.3	Score=38	NA	NA	NA	NA	NA	0.6022	0.732228	0.6077	0.79	rs4845364	rs484
25	1	1.56E+08	1.56E+08	G	A	intronic	LMNA	150330	Cardiomy	NA	NA	NA	1q22	NA	NA	NA	NA	NA	NA	0.1041	0.111821	0.1101	0.094	rs593987	rs593
26	1	1.56E+08	1.56E+08	T	G	intronic	LMNA	150330	Cardiomy	NA	NA	NA	1q22	NA	NA	NA	NA	NA	NA	0.2119	0.230032	0.1864	0.232	rs594028	rs594
27	1	1.56E+08	1.56E+08	TCTCCCT(	-	intronic	LMNA	150330	Cardiomy	NA	NA	NA	1q22	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.094	NA	NA
28	1	1.56E+08	1.56E+08	T	C	intronic	LMNA	150330	Cardiomy	NA	NA	NA	1q22	NA	NA	NA	NA	NA	NA	NA	0.112021	NA	0.116	rs622834	rs622
29	1	1.56E+08	1.56E+08	C	G	intronic	LMNA	150330	Cardiomy	NA	NA	NA	1q22	NA	NA	NA	NA	NA	NA	0.126	0.132188	0.1097	0.152	rs665979	rs665
30	1	1.56E+08	1.56E+08	T	G	intronic	LMNA	150330	Cardiomy	NA	NA	NA	1q22	Score=24	NA	NA	Score=83	NA	NA	0.1875	0.197684	0.163	0.196	rs513043	rs513

Figure 2 A representative snapshot of a list of variant file in excel format.

ant annotation was performed using Annovar for location and predicted function (Wang et al., 2010). Gross filtering was done using 1000 Genomes (≤0.01 MAF), EVS (≤0.01 MAF), ExAC (≤0.01 MAF) and dbSNP databases. Clinically relevant mutations were annotated using published variants in literature and a set of variant databases including ClinVar, OMIM and HGMD. Only non-synonymous, splice site, nonsense and frameshift variants found in the coding regions were used for clinical interpretation. Silent variations that do not result in any change in amino acid in the coding region were excluded.

Exome sequencing resulted in a total number of 26,612 variants. Figure 3 illustrates the filtering strategy used, resulting in a total of 896 exonic non-synonymous, splice-site and frameshift variants of which 34 variants were homozygous/hemizygous. Since none of the heterozygous variants were related to the phenotype of the patient, only homozygous/hemizygous variants were considered for further analysis. Among these, 9 variants were present in our in-house exome database and excluded from the study. After this, 25 variants were left among which 24 variants were reported in dbSNP with no clinical significance and hence excluded from the study. Finally only one variant was left, which was also relevant to clinical indication, an unreported hemizygous single base pair deletion in exon 8 of *GPC3* gene (chrX:132670203delA) in the patient. Mutations in *GPC3* are known to cause Simpson-Golabi-Behmel syndrome. Further

Filtering criteria	Number of variants
Total number of variants	26,612
Variants after filtering for 1000 Genomes (≤0.01)	3026
Variants after filtering for Exome Variant Sewer (≤50.01)	2692
Variants after filtering for Exome Aggregation Consortium (≤50.01)	1760
Variants remaining after filtering for intergenic, intronic and synonymous variants (retained exonic nonsynonymous, splice-site variants and indels causing frameshift)	896
Homozygous and hemizygous variants	34
Variants remaining after excluding Indian polymorphisms from in house data	25

Figure 3 An example of variant filtering strategy followed in whole exome (adapted from Das Bhowmik et al., 2015).

in silico analysis revealed that this mutation results in a frameshift and is likely to create a new stop codon at 62 amino acids downstream to codon 564 (c.1692delT; p.Leu565SerfsTer63) of the protein. Thus, WES helped in this case to establish the diagnosis in a patient with unexplained overgrowth syndrome as Simpson-Golabi-Behmel syndrome.

Conclusion

With the trend of gradually decreasing cost of exome sequencing, the technology has become imperative in the molecular diagnosis of rare Mendelian disorders. In this era of NGS, it is important for everyone related to this field to have at least some basic knowledge of exome analysis to correctly interpret the results which will ultimately help to carry out the appropriate pretest and post-test counseling of the patients. Also, it is always good to stay in tune with the continuous flow of updates of exome analysis since the technology is still evolving and so also the analytical methods.

It is believed that because of our poor understanding of non-coding genetic variation, the analytical components of most of the whole genome studies have inconsistently depended on variation within the exome. However, if the cost of sequencing continues to fall at this pace, it is possible that the field will gradually move from whole exome to whole genome sequencing. However, taking advantage of the more compact data of exome for disease gene discovery and molecular diagnostics in patients crucially depends on the development of analytical strategies for our understanding of non-coding variation. This is as much an opportunity as it is a challenge.

References

1. Bamshad MJ, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011; 12: 745-755.
2. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009; 461: 272-276.
3. Ng SB, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010; 42: 30-35.
4. D'Antonio M, et al. WEP: a high-performance analysis pipeline for whole-exome data. BMC Bioinformatics. 2013; 14 Suppl 7: S11.
5. DePristo M, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011; 43: 491-498.
6. Van der Auwera GA, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013; 43: 11.10.1-33.
7. Pabinger S, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 2014; 15: 256-278.
8. Das Bhowmik A, et al. Whole exome sequencing identifies a novel frameshift mutation in GPC3 gene in a patient with overgrowth syndrome. Gene 2015; 572: 303-306.
9. Wang K, et al. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38: e164.