A complete look with a magnifying glass of genomic and exome sequencing: what are we receiving in our clinical reports?

Abstract

The objective of this study was to review the ability of laboratory’s exome-sequencing test to detect known and novel sequence variants and identify the critical factors influencing the interpretation of a clinical exome test. Methods: There was select the guidelines papers and followed the validation strategy, essential considerations for sequencing analysis and how to detect known variants, interpretation-based approach that assessed relative ability to identify and interpret disease-causing variants, by analyzing and comparing the results and all informs. The study showed a detailed approach to exome analysis and reports. Conclusion: The analysis provide an assessment of critical areas that influence interpretation of an exome test, including comprehensive phenotype capture, assessment of clinical overlap, availability of parental data, and the addressing of limitations in database updates. This review can be used to inform improvements in phenotype driven interpretation of medical exomes in clinical and research.

Keywords: NGS, Genome, Exome, guidelines

Abbreviations: DNA, Deoxyribonucleic acid; NGS, Next generation sequencing; QC, Base call quality score; HGVS, Human genome variation society; NCBI, National Center for Biotechnology Information; LRG, Locus Reference Genomic; UTR, untranslated regions; ACMG, American College of Medical Genetics and Genomics; GUS, genes of uncertain significance; CAP, American College of Pathology; CCD, Centers for Disease Control and Prevention

Introduction

The exome is the part of the genome formed by the DNA fragments (exons) that are transcribed to give rise to the proteins. The study of exome is one of the most complete and complex ways to study our DNA. The exons are the coding regions that will provide the information for the synthesis of a protein, while the introns are non-coding regions, which are interspersed in the gene and have other functions. Clinical laboratories are rapidly implementing next generation sequencing (NGS)-based tests for the diagnosis of genetic disorders. While targeted, NGS-based gene panels are highly tuned to groups of genes that may directly explain the individual’s phenotype (Figure 1). Each laboratory should make a validation process for the NGS and establish the sensibility and specificity of the process.

When should sequencing be ordered?

Complete exomic sequencing should be ordered to confirm or exclude clinical diagnoses in complex diseases, more than three clinical suspicions, degenerative and neurodegenerative diseases, in chronic diseases with undetermined diagnosis, rare and orphan diseases and in cancer or family history of cancer.

Genomic sequencing identifies circulating tumor DNA in blood and allows an early identification of the cancer in addition to a pharmacogenomic study and identification of the response to treatment. Among the 20000 genes in the genome there are hundreds that can cause unwanted cell division, genetic diseases and cancer. The exomic sequencing of 4900 genes directly related to disease makes it possible to quickly identify which genes are altered and their specific mutations.

Use of WES as a diagnostic test changed the testing strategy from focusing on few genes known to cause a disorder or phenotype to sequencing all genes in the genome and focusing the analysis on those groups of genes that may directly explain the individual’s phenotype (Figure 1).

Essential considerations for sequencing analysis

According to the guidelines of the American College of Medical Genetics and Genomics, the following requirements must be fulfilled in order for a study to be qualified and valid:

- Metrics: The report must contain the essential elements including structural results, interpretation, references, methodology used and the appropriate renouncements of the test. The elements of the report...

© 2018 Beatriz et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
are emphasized in the regulations of the CAP and CLIA standards for next-generation sequencing. The metrics used in the study for analytical validation must be reported in the report.

**Base call quality score**. Q score. The Q score is a base call probability scale incorrectly and is inversely proportional to the probability that a single base in the sequence is correct. For example, a T with QC of 30 is considered probably correct with a confidence of P 0.001. Any base with a QC <20 should be considered low quality and any identified variant should be considered a false positive (Table 1).

**Table 1** Base call quality score

| Q Score | Incorrect base call probability |
|---------|---------------------------------|
| Q40     | 1 in 10000                      |
| Q30     | 1 in 1000                       |
| Q20     | 1 in 100                        |
| Q10     | 1 in 10                         |

**Reading depth**

The depth of reading or coverage is conventionally considered as a number followed by “x”. It is the number of independent readings with the alignment at a locus of interest. It is frequently expressed as an average or percentage. For example, the clinical report must say the percentage of coverage of the test: 150X with 95-98% coverage.

Low coverage, less than 70X, is at risk of losing variants and assigning incorrect allelic zygoty states, and decreasing the ability to filter artifacts. Laboratories should have a minimum coverage to detect variants based on diagnostic approaches that guarantee an adequate analytical performance for the report. The detection of germline heterozygous variants must have a minimum coverage of 100X for the proband. A coverage greater than 100X is required for the detection of mixed variants or in mosaicism (samples of somatic tumor, mitochondrial heteroplasmy or germinal mosaicism).

**Variant reads**

It is the number of sequences read that support the presence of a variant. Due to the NGS error rate, the call level of less than 5 readings per variant is considered a false positive.

**Terminology**: The identification of the variants and mutations must follow the classification of the five characteristics: pathogenic, probably pathogenic, benign, probably benign and variant of uncertain significance.

**Nomenclature**: It must follow the naming standard guide of the variants designation (http://www.hgvs.org/mutnomen) and the version of the Human Genome Variation Society (HGVS) and (https://mutalyzer.nl)\(^\text{14}\).

Clinical reports should include the reference sequence and avoid ambiguous naming of variants at the DNA level as well as providing the coding region and nomenclature of the protein to assist functional interpretations (example: “g.” Genomic sequence, “c. “DNA coding sequence,” p. “For mitochondrial protein,” m. “, Etc.). The reference sequence must be complete and using the NCBI RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/) with the version or number of the Locus Reference Genomic (LRG) database (http://www.lrg-sequence.org).

The genomic coordinates should be used and defined according to the standard genome (e.g. hg19) or to the reference genomic sequence that covers the entire gene (including the 5’ and 3’ untranslated regions (UTRs) and promoters).

A reference transcript should be used for each gene and provide a report of when the variants are described, frequently using LRG10, CCDS Database, Human Gene Mutation Database (http://www.hgmd.cf.ac.uk), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar) or locus-specific database.

The genomic coordinates should be used and defined according to the standard genome (e.g. hg19) or to the reference genomic sequence that covers the entire gene (including the 5’ and 3’ untranslated regions (UTRs) and promoters).

A reference transcript should be used for each gene and provide a report of when the variants are described, frequently using LRG10, CCDS Database, Human Gene Mutation Database (http://www.hgmd.cf.ac.uk), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar) or locus-specific database.

However, laboratories must evaluate the clinical impact of the variant and its transcripts to be interpreted clinically. Not all types of variants (eg complex variants) are covered by the recommendations of HGVS, the ACMG recommends three exceptions from the rules of the HGVS nomenclature:\(^\text{15,16}\)

“X” is considered acceptable for use in reporting non-sense variants “*” and “Ter”;

The exon number where the variant is reported is recommended; and

The pathogenic term is recommended instead of the term of affected function.

**Literature and databases**: The population frequency of the variant must be taken into account before reporting it (Table 2. Databases).

**Data storage and traceability of patient reports**

NGS generates a massive amount of data files with differing information contents and sizes. Laboratories should make explicit in their policies which files will be retained. We recommend that the laboratory consider a minimum of 2-year storage of a file type that would allow regeneration of the primary results as well as reanalysis.
with improved analytic pipelines (e.g., bam or fastq files with all reads retained). In addition, reinterpretation of variant significance may be done every year in order to verify the VUS variants.

**Table 2** Databases of the sequencing

| Population data base | Disease data base | Sequencing data base |
|----------------------|-------------------|----------------------|
| Exome Aggregation Consortium | Exome Variant Server | dbSNP |
| http://exac.broadinstitute.org/ | http://evs.gs.washington.edu/EVS | http://www.ncbi.nlm.nih.gov/snp |
| 1000 Genomes | 1000 Genomes | dbVar |
| http://browser.1000genomes.org | http://browser.1000genomes.org | http://www.ncbi.nlm.nih.gov/dbvar |
| ClinVar | Human Gene Mutation Database | ClinVar |
| http://www.ncbi.nlm.nih.gov/clinvar | http://www.arom.org | http://www.ncbi.nlm.nih.gov/clinvar |
| OMIM | | OMIM |
| http://www.ncbi.nlm.nih.gov/omim | | http://www.ncbi.nlm.nih.gov/omim |

**Computational analysis and prediction programs (In Silico):**

There is a variety of tools available for in silico analysis. The algorithms used may differ among themselves in the effect of the variant, the nucleotide sequence and amino acid levels including the effect in the variant or in the protein. Two categories include if the change is missense the result of the function of the protein may be damaged or the structure and predict its effect (Table 3).

The impact of the effect depends on the evolutionary conservation of the protein or the biochemical consequences. In general, most algorithms predict the relationship with the disease with a sensitivity of 65-80%. Among them PolyPhen, SIFT and Mutation Taster.

The bioinformatic analysis of NGS designed to convert signals into data, interpret information and turn it into clinical application, is conceptualized as primary, secondary and tertiary analysis, as can be seen in Figure 2.

**Proposed criteria for interpretation of variants**

The following approach evaluates the evidence of primary Mendelian inheritance variants; its use is not for somatic, pharmacogenomic variations or associated with complex multigene phenomena. Variations of uncertain significance must be taken into special consideration and followed up (“genes of uncertain significance”, GUS) to identify new genes in the disease.

The interpretation for the pathogenicity determination is independent of being the cause or not of the disease. There are two criteria for classifying pathogenicity or probable pathogenicity (Table 4) and Benign or probably benign variants (Table 5).

The criteria are combined according to the scoring rules as shown in Table 6. Where flexibility is provided to the classification of the variant.

**Table 3** In silico predictive algorithms

| Prediction | Name | website |
|------------|------|---------|
| | ConSurf | http://bental.tau.ac.il/new_ConSurfDB/ |
| | FATHMM | http://fathmm.biocompute.org.uk/fathmmMKL.htm |
| | Mutation Assesor | http://www.ncbi.nlm.nih.gov/Man uter/page/missense-prediction-tools |
| | PANTHER | http://pantherdb.org/data/ |
| | PhD-SNP | http://snps.biofold.org/phd-snp/phd-snp.html |
| | SIFT | http://sift.jcvi.org/ |
| | SNPs&GO | http://snps.biofold.org/snps-and-go/snps-and-go.html |
| | AlignGVGD | http://p33.iarc.fr/AGVGDMethod.aspx |
| | Mutation Taster | http://www.mutationtaster.org/ |
| | PolyPhen 2 | http://genetics.bwh.harvard.edu/pph2/ |
| | Condel | https://omictools.com/consensus-deleteriousness-score-of-missense-snvs-tool |
| | CAAD | http://cadd.gs.washington.edu/ |
| | nsSNPAnalyzer | https://omictools.com/nssnpanalyzer-tool |
| | Provean | http://provean.jcvi.org/index.php |
| | GeneSplicer | http://www.ccbcb.umd.edu/software/GeneSplicer/gene_spl.shtml |
| | Human Splicing Finder | http://www.umd.be/HSF3/ |

**Citation:** Beatriz HAB, Olga LRC, Claudia PBH, et al. A complete look with a magnifying glass of genomic and exome sequencing: what are we receiving in our clinical reports? J Anesth Crit Care Open Access. 2018;10(2):52–58. DOI: 10.15406/jaccoa.2018.10.00358
Table 4 Classification criteria of pathogenic variants

| Very strong |  
|-------------|
| Null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease. |
| Caveats: |
| Beware of genes where LOF is not a known disease mechanism (e.g., GFAP, MYH7) |
| Use caution interpreting LOF variants at the extreme 3’ end of a gene |
| Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact |
| Use caution in the presence of multiple transcripts |
| Strong |
| Same amino acid change as a previously established pathogenic variant regardless of nucleotide change |
| Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level |
| PS1 Example: Val→Leu caused by either G>C or G>T in the same codon |
| Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level |
| PS2 De novo (both maternity and paternity confirmed) in a patient with the disease and no family history |
| Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo transfer, and so on, can contribute to non maternity. |
| PS3 Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product |
| Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical diagnostic laboratory setting are considered the most well established. |
| The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls |
| Note 1: Relative risk or OR, as obtained from case–control studies, is >5.0, and the confidence interval around the estimate of relative risk or OR does not include 1.0. See the article for detailed guidance. |
| Note 2: In instances of very rare variants where case–control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence. |
| Moderate |
| Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation. |
| PM1 Absent from controls (or at extremely low frequency if recessive) (Table 6) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium |
| Caveat: Population data for insertions/deletions may be poorly called by next-generation sequencing. |
| PM2 For recessive disorders, detected in trans with a pathogenic variant |
| Note: This requires testing of parents (or offspring) to determine phase. |
| PM3 Protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants |
| PM4 Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before |
| Example: Arg156His is pathogenic; now you observe Arg156Cys |
| Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level. |
| PM5 Assumed de novo, but without confirmation of paternity and maternity |
| PM6 Supporting |
| Co segregation with disease in multiple affected family members in a gene definitively known to cause the disease Note: May be used as stronger evidence with increasing segregation data |
| PP1 Missense variant in a gene that has a low rate of benign missense variation and in which missense variants are a common mechanism of disease |
| PP2 Multiple lines of computational evidence support a deleterious effect on the gene or gene product (conservation, evolutionary, splicing impact, etc.) |
| PP3 Caveat: Because many in-silico algorithms use the same or very similar input for their predictions, each algorithm should not be counted as an independent criterion. PP3 can be used only once in any evaluation of a variant. |
| PP4 Patient’s phenotype or family history is highly specific for a disease with a single genetic etiology |
| PP5 Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent evaluation |
| PP5 |
Table 5 Criteria for classifying benign variants

| Criteria | Description |
|----------|-------------|
| BA1 | Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium |
| BS1 | Allele frequency is greater than expected for disorder |
| BS2 | Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age |
| BS3 | Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing |
| BS4 | Caveat: The presence of phenocopies for common phenotypes (i.e., cancer, epilepsy) can mimic lack of segregation among affected individuals. Also, families may have more than one pathogenic variant contributing to an autosomal dominant disorder, further confounding an apparent lack of segregation. |
| BP1 | Missense variant in a gene for which primarily truncating variants are known to cause disease |
| BP2 | Observed in trans with a pathogenic variant for a fully penetrant dominangene/disorder; or observed in cis with a pathogenic variant in any inheritance pattern |
| BP3 | In-frame deletions/insertions in a repetitive region without a known function |
| BP4 | Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc) |
| BP5 | Variant found in a case with an alternate molecular basis for disease |
| BP6 | Reputable source recently reports variant as benign but the evidence is not available to the laboratory to perform an independent evaluation |
| BP7 | A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved |

Table 6 Rules for combining criteria to classify sequence variants

| Status | Criteria |
|--------|----------|
| Pathogenic | 1. 1 Very Strong (PVS1) AND a. ≥1 Strong (PS1–PS4) OR b. ≥2 Moderate (PM1–PM6) OR c. 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) OR d. ≥2 Supporting (PP1–PP5) 2. ≥2 Strong (PS1–PS4) OR 3. 1 Strong (PS1–PS4) AND a. ≥3 Moderate (PM1–PM6) OR b. 2 Moderate (PM1–PM6) AND ≥2 Supporting (PP1–PP5) OR c. 1 Moderate (PM1–PM6) AND ≥4 Supporting (PP1–PP5) |
| Probably pathogenic | 1. 1 Very Strong (PVS1) AND 1 Moderate (PM1–PM6) OR 2. 1 Strong (PS1–PS4) AND 1–2 Moderate (PM1–PM6) OR 3. 1 Strong (PS1–PS4) AND ≥2 Supporting (PP1–PP5) OR 4. ≥3 Moderate (PM1–PM6) OR 5. 2 Moderate (PM1–PM6) AND ≥2 Supporting (PP1–PP5) OR 6. 1 Moderate (PM1–PM6) AND ≥4 Supporting (PP1–PP5) |
| Benign | 1. 1 Stand-Alone (BA1) OR 2. ≥2 Strong (BS1–BS4) |
Discussion and conclusion

Complete genomic and exomic sequencing is the ideal diagnostic test for complex diseases, it is cost-effective, and it requires an analysis with high quality parameters and applying diagnostic algorithms and tools to transform the data into valid clinical application.

Professional societies such as the American College of Pathology (CAP) and the US Centers for Disease Control and Prevention (CCD) and American College of Medical Genetics and Genomics (ACMG), as well as the Association of Molecular Pathology have established the regulations and guidelines for the realization of the test with high quality standards and for the analysis of data that must be fulfilled in all the institutions that carry out this type of studies.

The use of these tests allows for a specific treatment with greater benefit for the population.17 Knowing one’s own genetic information helps guide the distribution of economic resources according to the genomic profiles of health and disease.18,19

In Colombia, this type of studies is carried out with very good quality in few accredited laboratories where the quality parameters applied to each study are shown, the metrics used and also the complete bioinformatics that is available to the clinician for future analyzes. It is important to consider that population genomic studies are key to being able to find correct actions and medicines according to the ethnic group.

Acknowledgements

The authors thank American College of Pathology (CAP), the US Centers for Disease Control and Prevention (CCD) and American College of Medical Genetics and Genomics (ACMG), as well as the Association of Molecular Pathology for constructing the model and guidelines for this next generation sequencing technology.

Conflict of interest

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.
18. Bean LJ, Tinker SW, da Silva C, et al. Free the data: one laboratory’s approach to knowledge-based genomic variant classification and preparation for EMR integration of genomic data. Hum Mutat. 2013;34:1183–1188.

19. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141:210-217.