Short Communication　　　　　　　　　　　　　　　　　　　　　　**Open Access**

Interpretation of Clinical Next-Generation Sequencing Data: A Hurdle to Jump Over

Wenhui Laura Li[1,2*]

[1]*Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, CA, USA*

[2]*Help Undiagnosed Children Foundation, Arcadia, CA, USA*

[*]**Corresponding author**: Wenhui Laura Li, Help Undiagnosed Children Foundation, Arcadia, CA, USA, Tel: +3104293165; E-mail: laura@hucf.org

Short Communication

Next-generation sequencing (NGS) has been revolutionary for the clinical diagnostics field. With its high throughput sequencing power and plummeting cost, it has been increasingly used in clinical labs. Instead of testing the candidate genes one at a time by Sanger sequencing, now a lab can test a group of candidate genes at the same time using the NGS method. For example, many clinical labs now offer epilepsy gene panel tests that usually sequence 100-500 genes that are known to be causal or have associations with different kinds of epilepsies. Gene panel tests are also offered for genetic heterogeneous diseases like neurodevelopmental disease, cardiomyopathy, immunodeficiency disease, etc. This approach dramatically increases diagnostic efficiency and helps clinicians to zoom in on the genetic cause for a certain disease in a timely manner. In addition, the NGS technology is also used to diagnose patients who have been through diagnostic odysseys. Exome sequencing is currently used for this purpose in clinical labs. Exome test sequences an exome that contains all the protein coding regions which comprises 1.5% of the genome but contains 80% of recognized disease-causing mutations. Numerous examples illustrated that the exome sequencing method can efficiently identify genetic causes for undiagnosed diseases, which not only helps clinicians to obtain accurate diagnoses but also guides clinicians in the personalized care and treatment of their patients [1,2].

Although NGS technology transforms the clinical diagnostics field, it is still not easy for routine clinical molecular labs to adopt this technology. The hurdle is the interpretation of clinical NGS data, which usually has gigabytes or terabytes of data. Normally, a clinical exome sequencing test detects 20,000-30,000 variants in protein-coding regions per patient, and identifying a disease-causing variant from this large number of variants poses a serious challenge for routine clinical molecular labs which usually only deal with one gene and several variants at a time. Without a bioinformatics team, which helps with the sequence alignment, variant calling, and variant filtration, it is impossible for the lab directors to make sense of the huge amount of NGS data, let alone to make clinical interpretations out of it. Commercial software is available; however, often there is an annual subscription fee and a fee for analysis per sample which a small clinical lab cannot afford.

In addition, even after bioinformatics data analysis and filtration, NGS data still needs manual interpretation of those identified genes and variants. Big academic labs such as Baylor Genetics Lab and UCLA Clinical Exome Sequencing Lab have an NGS data review board which includes lab directors, physicians, and researchers in the molecular biology and genetics fields to manually review exome sequencing data and make a final clinical interpretation. This manual review process makes sure that the genetic variant found can explain the clinical presentations of the patient and inheritance pattern of the disease if it is known. Recently bioinformatics pipelines were reported to be able to efficiently prioritize the genetic variants according to the phenotype information and inheritance pattern of disease causal genes [3-5].

The clinical labs that already offer clinical exome sequencing tests have shown the diagnostic yield of the clinical exome is just around 25% [6-9]. The remaining 75% of clinical cases still could not be diagnosed, preventing appropriate treatment of these patients. The explanation of this low diagnostic yield could be the following:

- Disease-causing variants are located outside of protein-coding regions, such as gene promoter, enhancer, deep intronic region, noncoding intergenic region, etc [10]. Not knowing much about these non-protein-coding regions prevents the interpretation of the functional impact of variations seen in them. For example, a gene promoter is important for gene expression, and it is well known that gene promoter mutations can lead to genetic diseases or cancer [11-14]. However, the pathogenicity of a promoter variant could not be easily inferred without doing functional analysis.

- Disease-causing genes or variants are novel or with few functional studies and associated clinical reports; therefore, it is difficult for clinical labs to perform clinical interpretations on those genes or variants. Often, exome-sequencing-identified variants that were not seen previously, the majority of which are missense mutations, and the pathogenicity of variants remains to be tested functionally. In addition, some novel genes are found in exomes that might be associated with human diseases (inferred from animal studies); however, they are not documented in the OMIM and HGMD databases so that clinical interpretation on the variants seen in these genes is impossible.

- Multigenic cause contributes to the disease onset. Currently the clinical labs only analyze genomic data according to the simple Mendelian inheritance pattern. However, the disease etiology could be multigenic.

- Genetic cause is not enough for the disease onset; environmental insult also plays an important role in disease manifestation. Up to date, how the genetic and environment factors interact with each other to initiate disease onset is not clear yet; however, active research is being done for diseases like autism, allergy, and cancer [15-18]. The knowledge gleaned from this research undoubtedly helps us understand these complicated diseases.

In conclusion, with increasing adoption of genomic sequencing technology in clinical labs, the main challenge is to interpret the clinical genomic data accurately in a timely manner so that the lab can give back helpful clinical reports to physicians. The main obstacle for the interpretation of genomic data is due to the unreadiness of

processing huge amounts of data, lack of knowledge of genetic variation in normal populations, insufficient clinical and research studies on important disease genes and variants, and unknown functionalities of non-protein-coding regions in the genome.

Currently, genomic data processing is getting more efficient due to a lot of research effort has been put in, and thus more advanced software tools have been made available for free usage. Such software tools can be adapted for clinical usage with the mind of patient privacy, and automatic clinical report upload to patients' medical record.

In addition, for clinical testing, each patient's clinical features have to be considered for data analysis too. With more and more applications of whole genome sequencing, transcriptome sequencing, and epigenomic sequencing, cutting-edge bioinformatics tools are also being created to integrate various genomic data so that clinical lab directors can process complete genomic profile of patients before final clinical interpretations.

In addition, big population sequencing projects, sequencing projects on specific patient cohorts and functional studies on non-coding regions of the genome will allow us, in the near future, to discover the meaning of genomic codes that we do not yet understand, and unveil the secrets hidden in our genome. With the advancement of our tools and knowledge in genomics, we will accumulate unprecedented capabilities to interpret our genome, and thus unleash the full potential of NGS technology.

References

1. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, et al. (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med 13: 255-262.

2. Rabbani B, Tekin M, Mahdieh N (2014) The promise of whole-exome sequencing in medical genetics. J Hum Genet 59: 5-15.

3. Smedley D, Jacobsen JO, Jager M, Kohler S, Holtgrewe M, et al. (2015) Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat protoc 10: 2004-2015.

4. Smedley D, Robinson PN (2015) Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Med 7: 81.

5. Yang H, Robinson PN, Wang K (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods 12: 841-843.

6. Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, et al. (2014) Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA 312: 1880-1887.

7. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369: 1502-1511.

8. Gahl WA, Tifft CJ (2011) The NIH Undiagnosed Diseases Program: lessons learned. JAMA 305: 1904-1905.

9. Farwell KD, Shahmirzadi L, El-Khechen D, Powis Z, Chao EC, et al. (2015) Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet Med 2015 17: 578-586.

10. Diederichs S, Bartsch L, Berkmann JC, Frose K, Heitmann J, et al. (2016) The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. EMBO Mol Med 8: 442-457.

11. Theuns J, Brouwers N, Engelborghs S, Sleegers K, Bogaerts V, et al. (2006) Promoter mutations that increase amyloid precursor-protein expression are associated with Alzheimer disease. Am J Hum Genet 78: 936-946.

12. Richter S, Vandezande K, Chen N, Zhang K, Sutherland J, et al. (2003) Sensitive and efficient detection of RB1 gene mutations enhances care for families with retinoblastoma. Am J Hum Genet 72: 253-269.

13. Sagi M, Frenkel A, Eilat A, Weinberg N, Frenkel S, et al. (2015) Genetic screening in patients with Retinoblastoma in Israel. Fam Cancer 14: 471-480.

14. Cowell JK, Bia B, Akoulitchev A (1996) A novel mutation in the promotor region in a family with a mild form of retinoblastoma indicates the location of a new regulatory domain for the RB1 gene. Oncogene 12: 431-436.

15. Koufaris C, Sismani C (2015) Modulation of the genome and epigenome of individuals susceptible to autism by environmental risk factors. Int J Mol Sci 16: 8699-8718.

16. Vijayakumar NT, Judy MV (2016) Autism spectrum disorders: Integration of the genome, transcriptome and the environment. J Neurol Sci 364: 167-176.

17. Campbell DE, Boyle RJ, Thornton CA, Prescott SL (2015) Mechanisms of allergic disease - environmental and genetic determinants for the development of allergy. Clin Exp Allergy 45: 844-858.

18. Simonds NI, Ghazarian AA, Pimentel CB, Schully SD, Ellison GL, et al. (2016) Review of the Gene-Environment Interaction Literature in Cancer: What Do We Know? Genet Epidemiol 40: 356-365.