Exome and genome sequencing for diagnosing patients with suspected rare genetic disease

Go Hun Seo* and Hane Lee
Medical Genetics Division, 3billion, Inc., Seoul, Korea

Rare diseases, even though defined as fewer than 20,000 in South Korea, with over 8,000 rare Mendelian disorders having been identified, they collectively impact 6-8% of the global population. Many of the rare diseases pose significant challenges to patients, patients' families, and the healthcare system. The diagnostic journey for rare disease patients is often lengthy and arduous, hampered by the genetic diversity and phenotypic complexity of these conditions. With the advent of next-generation sequencing technology and clinical implementation of exome sequencing (ES) and genome sequencing (GS), the diagnostic rate for rare diseases is 25-50% depending on the disease category. It is also allowing more rapid new gene-disease association discovery and equipping us to practice precision medicine by offering tailored medical management plans, early intervention, family planning options. However, a substantial number of patients remain undiagnosed, and it could be due to several factors. Some may not have genetic disorders. Some may have disease-causing variants that are not detectable or interpretable by ES and GS. It’s also possible that some patient might have a disease-causing variant in a gene that hasn’t yet been linked to a disease. For patients who remain undiagnosed, reanalysis of existing data has shown promises in providing new molecular diagnoses achieved by new gene-disease associations, new variant discovery, and variant reclassification, leading to a 5-10% increase in the diagnostic rate. More advanced approach such as long-read sequencing, transcriptome sequencing and integration of multi-omics data may provide potential values in uncovering elusive genetic causes.

Key words: Rare diseases, Exome sequencing, Whole genome sequencing.

Introduction

The definition of a rare disease is set by each country. In the United States and South Korea, a rare disease is defined as one that affects fewer than 2,000 and 20,000 people, respectively [1-3]. Over 80% of the rare diseases are reported to have a genetic basis and are commonly referred to as rare Mendelian diseases [4]. To date, more than 8,000 Mendelian disorders have been documented [2,5]. While the prevalence of each of these diseases is rare, collectively, there are millions of individuals and families worldwide affected by rare diseases which equates to 6-8% of the global population. Many rare diseases do not have a cure and need long-term clinical care and management. This not only places a heavy personal burden on patients in terms of their medical, financial, and psychosocial well-being but also has a substantial public health and the economic impact [6].

Patients who are suspected of having a rare disease often undergo a series of testing, but often doesn’t receive a clear answer and end up in a long and arduous diagnostic journey. Although there are rare diseases with such clear phenotype and no or low
genetic heterogeneity, a single gene or a small gene panel testing is enough to find the molecular diagnosis, most rare diseases have high phenotypic complexity and genetic diversity, making them challenging to find a molecular diagnosis by conventional single/small panel genetic testing using Sanger sequencing method [7]. Diagnostic approach for rare disease have dramatically changed with the advent of next-generation sequencing (NGS), in particular exome sequencing (ES) and genome sequencing (GS) [8-10]. ES simultaneously sequences almost all protein-coding region of nearly all genes (20,000), while GS sequences all non-coding regions as well. ES and GS achieve a diagnostic rate of approximately 25% to 50%, although this rate varies depending on the disease category [8-13]. The unbiased assessment by ES or GS significantly speeds up the process of making an accurate diagnosis, especially for patients whose symptoms are not specific [14]. It has also led to a dramatic increase in the discovery of new disease-genes association with approximately 250 new disease genes identified each year over the past decade [4]. Finally, like any other diagnostic tests, patients receiving a clear diagnosis by ES and GS will be able to receive tailored medical managements such as initiation of precise treatments, better monitoring for additional symptoms, and receiving customized family planning [15,16]. The patient may also be eligible for clinical trials. There are studies reporting how early application of ES/GS could be overall cost-effective even though the cost for ES/GS is initially high, as cost for multiple diagnostic testing and medical intervention within the undiagnosed period is dropped [17,18]. This is consistent with the American College of Medical Genetics and Genomics (ACMG) recommending ES and GS as a first-line test for patients with congenital anomalies or intellectual disability [19]. Currently, ES is widely implemented in clinical practice and GS is being adopted in national projects and research areas for the diagnosis of rare diseases [9,12,13,19]. In this review, we discuss the overall workflow of ES and GS, as well as their features, limitations, and future directions.

Indication for Exome and Genome Sequencing

For patients with symptoms that strongly point to a specific disease, a single gene test is recommended as a first-tier test. Examples include patients likely to have Down syndrome, Duchenne muscular dystrophy or cystic fibrosis [20-22]. Then there are diseases such as cardiomyopathy or hearing loss with high genetic heterogeneity but still targeted panel sequencing test may suffice although there could be cases where the patient actually has a more complex syndromic disorder that cannot be detected by panels but is too young to have shown all symptoms or has symptoms mild or overlooked [23,24]. Therefore, an unbiased approach such as ES or GS may be more appropriate to find the molecular diagnosis more rapidly [25] and that’s why recently, ES and GS are being recommended as the first-line test for patients not only with pediatric neurodevelopmental delay and/or one or more congenital anomalies, but also with seemingly single system disorders [19,25-27].

Testing Workflow of Exome and Genome Sequencing

The NGS-based genomic sequencing test workflow can be divided into wet-lab part and dry-lab part and the dry-lab part is typically divided into three stages: primary, secondary, and tertiary [28-30]. Wet-lab workflow starts with sample accessioning, genomic DNA extraction and quality check, NGS library preparation including exome capture, and sequencing. The main difference between ES and GS is that ES captures and sequences only the protein-coding regions of almost all genes which make up about 1% to 2% of the genome, whereas GS sequences the entire genome including the non-coding regions. Exome capture step involves hybridizing the sample with capture probes. There are various versions of commercially available capture probe kits [31]. It is possible to enhance the exome performance by adding custom capture probes to augment the coverage of difficult-to-sequence regions, intronic regions with known disease-causing variants and mitochondrial genome. Once the sequencing is complete, the primary analysis starts with the base call files being converted to FASTQ files while demultiplexing the samples based on the index information that was attached to each DNA fragment during the library preparation step [29].

Secondary analysis starts with aligning all sequencing reads to the human reference genome. The current human reference genome version is GRCh38. However, there are still many laboratories using GRCh37 because switching the reference version is a major task. When variants are compared between the data that was mapped to GRCh37 and GRCh38 versions, GRCh38 is in general more accurate although not always [32]. After post-alignment fine tuning including potential PCR duplicate marking and base recalibration, variants are called. Multiple variant calling programs are employed to detect different types of variants, including single nucleotide variant (SNV), small insertion/deletion (INDEL), copy number variant (CNV), structural variant (SV), repeat expansion variant, and mobile element insertion variant [28,29,33]. In general, ES can detect SNVs, small INDELS,
and large CNVs that affect more than three consecutive exons within the protein-coding region, while GS has a broader scope and can detect almost all types of variants due to continuous sequencing data of the entire genome [28,34,35].

Tertiary analysis constitutes of interpreting variants including annotation, filtering, classification, prioritization of variants, and finally identification of variants that are most likely to explain patient’s phenotype [28,29]. Given the long list of variants generated by secondary analysis ranging from ~80,000 variants to ~5 million variants for ES and GS, respectively, it is critical to have an efficient yet consistent tertiary analysis algorithm with high sensitivity and specificity. All variants are classified into five groups: pathogenic (P), likely pathogenic (LP) or variants of uncertain significance (VUS), likely benign (LB), and benign (B), according to the ACMG guidelines released in 2015 and 2019 [36,37]. To assess variant pathogenicity, these guidelines take into consideration multiple factors such as population frequency of a given variant, functional impact the variant is reported to have on the protein or may have on the protein, segregation status, and how well the patient’s phenotype matches to the disease associated with the gene the variant is found in [36,37]. The first filtering step typically starts by removing variants commonly found in population databases as they cannot cause a rare disease and therefore classified as B/LB. This step removes more than ~90% of all variants and therefore it is the most effective filtering. However, as this filter solely relies on the population information, and variants that are common in sub-population that are underrepresented in public database would not be filtered out as effectively and could be misclassified as P/LP/VUS when it is actually benign. After the common variants are removed, the remaining variants are classified as P/LP/VUS based on various factors mentioned above. Then, the variants are prioritized by how similar the patient’s clinical symptoms are to the reported symptoms of the disease associated with the gene the variant occurred in. Since this process is labor-intensive and time-consuming, there are now variant recommendation algorithms being developed using artificial intelligence (AI) technology [38-40]. Top-k, the frequency the diagnostic variant is found in the top-k variants, is a commonly used measure to assess the performance of the algorithms. The higher the frequency is at a small number of k, the better the performance is. As the algorithms improve, variant interpretation efficiency and accuracy will be improved.

Finally, one or two variants that are most likely to have caused the patient’s symptoms are selected and reported to clinicians, who then will evaluate the reported variants in the clinical context to make a final diagnosis [28]. Additional phenotyping or further genetic testing may be required in case of a VUS report.

### Diagnostic Yield and Technical Limitation of Exome and Genome Sequencing

The diagnostic yield of ES is reported to range from 12% to 63% depending on the patient’s symptoms and onset of clinical presentation [8,10,11,13,41], while the diagnostic yield for GS ranges from 21% to 73%, also depending on the phenotypes and ages of the patients being studied [12,41,42]. A modest increase in the diagnostic yield observed with GS compared to ES is attributed to the detection of disease-causing variants that are non-coding, and complex SVs [12,41]. Even some of the coding variants may only be called by GS as there are genomic regions that are difficult to sequence by ES but not GS. These regions include GC-rich regions that suffer with low-coverage due to PCR bias [43]. Also, there could be CNVs affecting the coding regions but too small to be called by chromosomal microarray (CMA) or ES [34,35]. The resolution of CMA and ES with average depth-of-coverage of ~100x are 30–50 kb and 3 consecutive exons, respectively. Therefore, a CNV of a size less than 30kb or affecting fewer than 3 consecutive exons may be missed by both CMA and ES. Lastly, lower-level heteroplasmic mitochondrial genome variants may only be identified GS and not ES as GS typically has significantly higher mitochondrial genome coverage at the level of ~1000X while ES has much lower coverage (<100x) unless it’s specifically targeted [44,45].

However, GS still has its limitations. Lower mean depth-of-coverage of ~30 to 50x achieved for GS reduces its sensitivity to identify low-level of mosaic variants in the nuclear genome [46]. A substantial number of non-coding variants identified by GS remain as of uncertain significance as they require functional studies to determine their protein consequence such as altered splicing or abnormal expression pattern [47,48]. That is why the diagnostic rate increases when GS is complemented with transcriptome sequencing (RNAseq) [47-50]. However, RNAseq is mostly performed as research because not all genes are expressed in accessible tissues [50,51].

### Next Steps for Exome/Genome-Negative Patients

Despite the advances in comprehensive genomic testing, more than half of patients remain still undiagnosed [8-13,41,52]. There are several reasons for this. First, some patients may not
actually have a genetic disorder. Some patients may exhibit clinical features resembling Mendelian disease, but their conditions may have other underlying causes such as environmental factors, such as fetal alcohol syndrome or complications from preterm delivery. Secondly, due to technical limitations of short-read sequencing, certain variant types cannot be detected. There are a lot of repeat sequences in the human genome but it's difficult to align short sequence reads to these regions [53]. Also, ES and GS cannot detect methylation abnormalities, potentially leading to missed diagnoses [54]. Thirdly, interpretation of VUS with insufficient evidence to be clearly classified as either pathogenic or benign could make up to 50% of reports and patients receiving inconclusive results with VUS will remain undiagnosed [37,55]. They will have to wait until additional evidence is collected through studies such as transcriptome sequencing, functional assays, and segregation analysis in family members. At many times, performing more such analyses is not easily doable, requiring significant amount of time and resources [56-58]. Finally, it’s possible that the patient might have a disease-causing variant in a gene that hasn’t yet been linked to a disease. The total number of phenotypes and disease genes continues to grow and OMIM database grows by ~250 new gene-disease each year [4]. Novel gene discovery is one major reason how reanalysis of the existing ES or GS data resulting in a new diagnosis [51,58,59]. Studies have shown that reanalyzing the ES data can lead to 3-10% increase of the diagnostic rate [52,59-62]. This increase is attributed to several factors, including reclassification of variant, identification of new variant, and novel gene discoveries [52,59-62]. A VUS may be reclassified as pathogenic or LP if additional evidence such as new functional data or test results that were not available for the initial analysis becomes available [52,59-62]. For instance, a heterozygous VUS could be reclassified as LP if it is found to be in trans with a pathogenic variant for a recessive disorder, or if it turns out to be de novo for a dominant disorder after parental testing [36,37]. Advances in the bioinformatics pipeline that results in improvement in variant calling could identify a variant that was not identified before and result in a new diagnosis [52,59-62]. Patient’s clinical spectrum can evolve overtime and with new phenotypic information, a previously unclear variant-disease association could become diagnostic [52,59-62]. Finally, as mentioned above, the contribution of novel gene discoveries to diagnostic reanalysis is the largest among all factors resulting in successful reanalysis. The ACMG has published guidelines outlining how reassessment and reanalysis of ES/GS data should be performed [63]. Reference laboratories typically offer a complimentary reanalysis of the ES or GS data six months to a year after the initial testing is done [59,64] although more frequent free reanalysis at a shorter interval could be offered [52]. Achieving a new molecular diagnosis through reanalysis is a cost-effective approach for patients undiagnosed after ES or GS, and it is expected to become a routine clinical practice [17,64].

Besides periodic reanalysis, there are several next steps that can address some of the limitations GS and ES have. Long-read genome sequencing (LRGS) is one. LRGS generates sequence reads that are significantly longer (ranging from 1 kb to several mega bases) compared to the 150-300 base pair reads generated by short-read NGS [65]. It is easier to uniquely align long reads to genomic regions with repeat sequences, allowing variant detection within these regions that could have been missed when short-read sequencing was performed [65]. LRGS could also be useful for phasing two variants that are far apart from each other [66]. In addition, LRGS can detect CpG methylation to find variants dysregulating the epigenetic machinery [54,66,67]. LRGS comes with challenges on data management, storage, and analysis, and the application of LRGS in diagnosing undiagnosed patients is currently mostly performed as research, as it proves its utility in molecular diagnosis and the cost further decreases, it may become a clinical test in a near future [65]. Even though GS and LRGS can detect almost all types of variants, it is still impossible to interpret the non-coding variants. The AI-based in silico prediction tools such as SpliceAI or Pangolin are useful but they are still predictions and until a functional study is performed to show abnormal splicing, the non-coding variant cannot be determined as disease-causing with confidence [68,69]. Confirmation of aberrant splicing or expression levels for non-coding variants requires RNAseq, which could increase the diagnostic rate by 7.5% to 18% depending on phenotype and tissues sampled [47-50]. The biggest challenge of RNAseq is acquiring appropriate tissue to detect the transcript-level changes [47,50,51]. For instance, approximately 50% of the genes associated with neurodevelopmental disorders are reported to be expressed at sufficient level in whole blood, muscle, or skin-derived fibroblasts, three relatively accessible tissues. For the rest of the genes, a method to express them needs to be developed or a ultra-high coverage RNAseq would need to be performed. Lastly, the combination of multi-omics data, including metabolomics, epigenomics, and proteomics, can complement NGS and contribute to diagnosis and new disease gene discoveries [70,71].
Conclusion

ES and GS have proven to be extremely useful tools, providing comprehensive and unbiased search for a diagnostic variant. They have also led to a large number of novel gene discoveries and understanding the underlying molecular mechanisms of many diseases, which in turn enabled more patients to be diagnosed through reanalysis. However, continuous effort is needed to improve the diagnostic yield of both tests by more effectively prioritizing variants and functionally assessing the impact of each variant. The bioinformatics pipeline could also further improve to identify more variants within the complex and repetitive genomic regions. For patients who are still undiagnosed after ES or GS, periodic reanalysis of existing data in light of growing medical knowledge is essential and approaches using more advanced technologies such as LRGS, RNAseq and/or integration of multi-omics data could also be considered, particularly for patients undiagnosed even after reanalysis.

Acknowledgements

None.

Funding

No fundings to declare.

Authors’ Contributions

Conception and design: GHS. Drafting the article: GHS. Critical revision of the article: GHS, HL. Final approval of the version to be published: GHS.

References

3. National Health Insurance Service. The statistics of national health insurance: population with health insurance coverage. [https://nhiss.nh.or.kr/bd/ad/bdada013cv.do]


41. Chung CCY, Hue SPY, Ng NYT, Doong PHL, Chu ATW, Chung BHY. Meta-analysis of the diagnostic and clinical utility of exome and genome sequencing in pediatric and adult patients with rare diseases across diverse populations. Genet Med 2023;25:100896.

42. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseinii SM, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a


