Erciyes Med J 2014; 36(4): 139-43 • DOI: 10.5152/etd.2014.7804

Exome Sequencing for The Identification of Mendelian Disease Genes

REVIEW

Gülsüm Kayman Kürekçi, Pervin Dinçer

ABSTRACT

Over the past several years, next-generation DNA sequencing technologies are used for the identification of genes responsible for Mendelian disorders and genetic variants related to common disorders. The development of exome sequencing and analysis approaches according to inheritance and pedigree information helps to overcome the majority of limitations encountered by traditional genetic mapping approaches. Different strategies used in previous studies constitute an important source for future studies on genetic disorders. In this review, exome sequencing approaches that are used to identify genetic causes of monogenic disorders and the pros and cons of conventional methods are presented.

Keywords: Exome sequencing, monogenic disorders, gene identification, genetic variant

INTRODUCTION

Many studies have been conducted for years to unravel the genetic causes of human diseases. According to the catalogue of rare monogenic disorders (OMIM- Online Mendelian Inheritance in Man), known as Mendelian disorders, more than 4,300 loci associated with single gene disorders were identified (1). Moreover, nearly 1,700 phenotypes with unknown molecular basis are described in OMIM. When nearly 1900 phenotypes that are suspected to have mendelian basis are added to this number, it can be expected that about 3,600 monogenic disorders still have to be identified. Additionally, investigation of genetic factors associated with common diseases has gained accelerated in recent years (2). Many genetic variants considered to affect the susceptibility to common diseases have been detected. The next-generation DNA sequencing platforms, being developed since 2005, help to overcome some factors that made the gene identification process difficult with traditional methods. The use of these new sequencing methods, particularly combined with targeted capture and enrichment techniques, rendered possible the detection of all coding sequences of the human genome easily. This approach is called exome sequencing. In this review, the exome sequencing approach, applied in the identification of genes responsible for single gene disorders, is presented.

Traditional gene identification approaches

The main method used to identify mutant genes responsible for single gene diseases is the positional cloning approach (3). This approach is based on the identification of the chromosomal location of the gene likely to be responsible for the disease. For this purpose, a candidate chromosomal region as narrow as possible is defined, and the candidate genes in this region are screened for a mutation. Genome-wide analysis of single-nucleotide polymorphisms (SNPs) is generally conducted for the identification of the candidate chromosomal region (4). Because the chromosomal position in the human genome of each polymorphism is defined, they are used to generate maps covering the whole genome. The genetic mapping approach used for single-gene diseases is linkage analysis (3). The linkage analysis is based on the calculation of the probability of a mutant disease allele to be inherited together with various genetic markers on the basis of genetic information obtained from family trees. By using the positions of genetic markers on chromosomes, most linked and closely related loci can be detected. Homozygosity mapping is the most common method used for the identification of mutant genes responsible for autosomal recessively inherited disorders (5). In this case, the candidate regions are restricted to homozygous regions in consanguineous families. Consequently, a mutational screen by DNA sequencing is conducted in candidate genes that are prioritized according to their association with the disease among all genes found in the identified chromosomal region.

The limitations of traditional approaches

Although most of the monogenic disorders have been elucidated by traditional approaches, there are cases for which these methods remain insufficient (2). First of all, the presence of too many genes in the defined candidate region is a limiting factor in a study with regard to sequencing cost. Furthermore, the genetic heterogeneity (the

Department of Medical Biology,
Hacettepe University Faculty of
Medicine, Ankara, Turkey

Submitted
11.06.2013

Accepted
16.09.2013

Correspondance
Gülsüm Kayman Kürekçi MD,
Department of Medical Biology,
Hacettepe University Faculty of
Medicine, Ankara, Turkey
Phone: +90 312 305 25 41
e.mail:
gulsumkymn@gmail.com

fact that similar disease phenotype can be cause by distinct mutant genes or alleles), which causes to deviate from mendelian inheritance and makes the correlation between genotype and phenotype difficult, and the presence of modifier genes changing the disease phenotype make the identification of the responsible locus difficult. Another common case is the presence of nuclear families formed by parents and their child or larger families with a few affected members (only 1 or 2). These families do that do not meet the criteria w to be efficiently analysed using traditional approaches. The genotype data extracted from these types of families generally remain statistically insufficient for classical analysis approaches.

Next-generation DNA and exome sequencing

In cases where traditional approaches remained insufficient, the development of large-scale next-generation DNA sequencing technologies has greatly accelerated gene identification studies. The common goal of platforms having been developed since 2005 is the parallel sequencing of millions of DNA sequences at a time (6). These developments come to mean a speedy cost reduction while increasing the sequencing strength and accuracy. However, the cost of sequencing the whole human genome, which is complex and large, is still very high. Moreover, significant infrastructure is necessary to filter, interpret and store the large amount of data (7). Gene identification studies have been accelerated with the development of methods that provide the targeting and sequencing of only particular regions in the genome since 2008 (8). Exome sequencing has become especially prominent in research about Mendelian disorders. This method renders possible the capture and sequencing of the whole exome corresponding only to protein-coding sequences. Since its first application in 2009 exome sequencing has been used for the identification of hundreds of new genes that are responsible for monogenic disorders (9). Almost 57% of these disorders have an autosomal recessive inheritance. Moreover, in about 35% of these studies, the gene responsible for the disease have been defined by sequencing the exome of a single individual apart from the controls (10).

Exome sequencing to identify causes of monogenic disorders

In order for any genetic variation to be associated with a single-gene disorder, it is expected to be rare, highly penetrant the probability for an individual to exhibit the phenotype defined by a genotype it affects the function and structure of the protein encoded by the gene that it is found in, and it is generally found in protein-coding sequences (11). Although the noncoding regulatory regions are not covered, exome sequencing is an effective method to identify genes responsible for Mendelian disorders. First of all, the majority of variants identified by positional cloning are located in the protein coding sequences. In fact, almost 85% of alleles accounting for single-gene diseases are found in protein-coding regions (12, 13). It is thought that the variants in regulatory regions that do not encode proteins are usually harmless or have little effect on phenotype. The extent to which these changes affect monogenic diseases has not been revealed yet (12). Because rare variants with detrimental effects are generally found in exonic sequences, exome sequencing is successfully to identify hundreds of mutant genes responsible for single-gene diseases, particularly those displaying autosomal recessive inheritance (14).

Exome sequencing

Several technologies aiming to capture all protein-coding exons, accounting for 1% of the human genome, have been developed since 2007 (15). The most commonly used commercial kits were developed by three different companies, Agilent, Nimblegen, and Illumina (16, 17). The main steps of exome capturing and sequencing differ slightly. In the first step, the genomic DNA to be sequenced is randomly fragmented into small fragments, and a DNA library is formed. The exonic sequences in the DNA fragments are captured and enriched by hybridization with DNA or RNA templates. In solid phase hybridization exome capturing is realized by microchips while, DNA or RNA templates marked with biotin are used in the liquid-phase hybridization approach. After their hybridization to exonic sequences, they are captured and enriched by streptavidin-coated beads. Finally, after washing in order to remove unbound genomic fragments, the enriched exon library is amplified and then sequenced by one of the next-generation sequencing methods.

Variant filtration

After all protein-coding exons are sequenced, the large amount of data has to be filtered (18). Short sequences have to be compared to a reference genome sequence, and the differences between the reference genome and the sample have to be identified. More than 90% of approximately 20.000 to 24.000 single-nucleotide variants (SNVs) obtained from one sample constitute known polymorphisms (15). All variants acquired in the first stage are compared to common polymorphism databases (for instance, dbSNP, 1000 Genomes Project, and HapMap), and control individuals and known polymorphisms are eliminated (19-21). Then, nonsynonymous mutations are eliminated, since they are expected to be non pathogenic. Moreover, additional filters are performed depending on various criteria, such as interspecies conservation of variants and their possible detrimental effects to the gene products they are found in (22).

Analysis approaches with exome sequencing

Different analysis approaches allow to determine the causal variant that is associated with the disease, among those remaining after the filtration step (10, 15, 22, 23). Several approaches have been followed in the identification of genes responsible for single-gene diseases, depending on information, such as inheritance, family tree, and genetic heterogeneity.

In the linkage analysis-based approach, classically, a common haplotype is found in family members, and healthy individuals are used as controls (24, 25) (Figure 1a). In the case of unrelated sporadic individuals, common variants shared by the patients and associated with the disease are determined with an overlap strategy (26, 27) (Figure 1b). In this case, the assumption that there is no genetic heterogeneity in the disease has to be made. Moreover, as the number of patients whose exome is sequenced increases, this approach becomes more effective. In the de novo approach, the exomes of trios composed of an affected child and his parents are sequenced, and the variants found in the patient but not detected in the parents are determined (28, 29) (Figure 1c). Finally, a homozygosity-based approach can be used for small consanguineous an autosomal recessiv disorder. This approach assumes that the

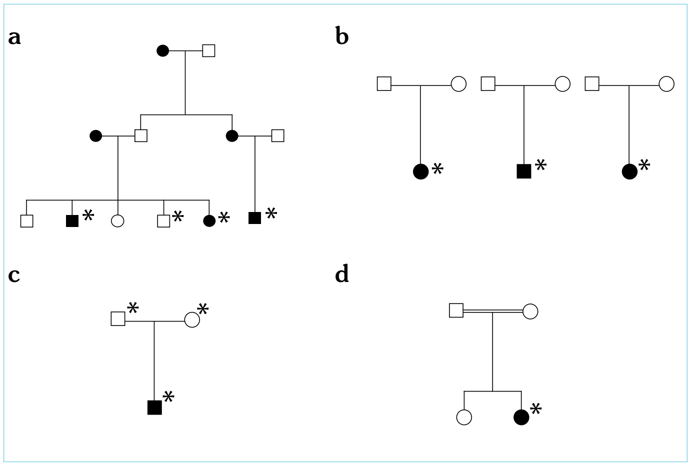


Figure 1. **Analysis approaches with exome sequencing. The individuals adequate for exome sequencing are marked with an asterisk. (a) Linkage analysis-based approach. (b) Overlap method. (c) De novo approach. (d) Homozygosity -based approach**

homozygote variant responsible for the disease together with homozygous markers is inherited through the parent from a common ancestor (30) (Figure 1d). After the sequencing of the exomes of the parent and the patient and the filtration of all variants obtained, homozygote variants located in large homozygous stretches are filtered and analysed.

Homozygosity mapping and exome sequencing

The rate of consanguineous marriage being approximately 21% in our country and rising up to 39% especially in the eastern and southeastern regions, increase the incidence of autosomal recessive disorders (31). In the case of autosomal recessive disorders, the consanguineous families have great importance. Homozygosity mapping is used for the identification of the gene responsible for the disease in these families (5). This approach is based on the assumption that the homozygous mutation is inherited from a common ancestor by the parental line (the alleles are identical by descent). Given that the probability of genetic markers around the homozygous mutation will be separated by recombination is low, the assumption that they are inherited together with the mutation is made (30). After scanning the whole genome of individuals with polymorphic markers, the responsible mutant gene is searched within the family-specific homozygote haplotypes. Homozygosity mapping is a method developed for families with consanguineous marriages and a few affected members. However, there are many small families and sporadic cases for whom the classical homozygosity mapping approach can not be applied. Moreover, large number of homozygous haplotypes can be identified or contain many candidate genes. This makes the determination of the mutation difficult.

In recent years, homozygosity mapping has been applied in combination with the exome sequencing approach in the identification of genes causing autosomal recessive disorders (32-35). Pippucci et al. (34), by using the approach they call 'Exome HOMozygosity,' identified the gene responsible for spastic paraplegia, displaying autosomal recessive inheritance and a leukodystrophy phenotype, by conducting homozygosity mapping using data obtained from exome sequencing. For this, after the exome of two affected brothers from a cousin marriage was sequenced and quality and polymorphism filtration of all variants was conducted, the remaining novel and rare single-nucleotide variants (SNVs) were detected (Figure 2). Homozygosis mapping has been conducted by combining the known single-nucleotide polymorphisms from SNP database (dbSNP130, and the new single-nucleotide variants detected by exome sequencing to form, forming a genetic map including, 135,035 genetic markers. As a result, 33 homozygous variants were detected after overlapping low-definition SNP genotyping data of the parents and children, loci obtained from data linkage analysis and regions from homozygosity mapping. After filtering the remaining variants depending on the type of mutation (missense, indel, nonsense, gain-of-function or loss-of-function mutations) possible detrimental effects and the expression level in the tissue affected in the disease; a single variant (NM_024306.2) responsible for the disease has been identified.

A different approach in which exome sequencing is used with homozygosity mapping enabled Özgül et al. (35) to detect a novel gene causing retinitis pigmentosa in a consanguineous family with with a single affected individual. In their approach, first of all, 250K SNP genotyping of family members (mother, father, the patient, and his healthy brothers/sisters) was realized.. In addition exome sequencing was conducted just for the affected family member. Although retinitis pigmentosa is genetically heterogeneous; according to the transmission of the disease through the family, and the autosomal recessive inheritance; 52 homozygous variants in 38 different genes were prioritized. By focusing only on the candidate genes found in the 9 homozygote haplotypes detected by homozygosity mapping among all genes, the number of candidate genes was reduced to 2, and the homozygous mutation responsible for the disease was detected in the MAK gene (35).

Limitations of exome sequencing

Recently, although the causes of many Mendelian disorders have been explored thanks to exome sequencing technology, cases where this approach fails to identify the responsible variant remain (15). Besides the advantages of the exome sequencing method, some limitations persist. The conditions that can lead to failure are the absence of the responsible gene in the regions targeted during exome capture or the presence of an unknown gene; low coverage of the locus, including the responsible variant (present platforms do not capture approximately 5%-10% of the known exons in the genome); failure to detect the signal (base calling) despite the responsible variant being covered; or presence of alignment errors with the reference sequence in particular regions such as those containing highly repetitive sequences. Moreover, the presence of pathogenic variants in the control set or in the polymorphism database during the analysis of data, false-positive results associated with processed pseudogenes and duplications, the presence of the responsible variant in non-exon regions (intronic or regulatory regions), or the existence of many candidate variants after filtration can make the identification of a single responsible variant difficult.

CONCLUSION

High-throughput next-generation DNA sequencing technologies have overcome the limitations of conventional gene identification approaches to a great extent. Because a serious infrastructure is to

1. Exome capturing and sequencing

2. Exome data analysis
   - Alignment to reference genome sequence
   - Polymorphism filtration according to SNP databases

SNVs	Patient 1	Patient 2
Total	20.967	22.651
Included in dbSNP	19.727	21.237
Not-included in dbSNP	1240	1414

3. Homozygosity mapping with exome data

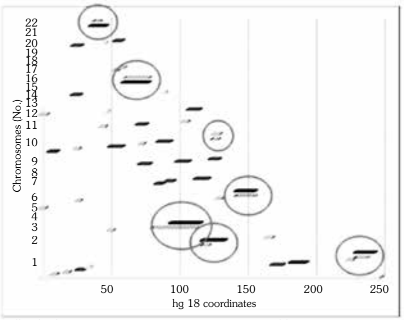
Known SNVs in dbSNP ↘ New SNVs detected in exome sequencing ↙

A genetic map consisting of 135,035 genetic markers

Homozygosity mapping

Homozygous haplotypes= ex-HOM regions

4. Comparison to SNP genotyping data



Full cylinders represent Ex-HOM regions. Dotted lines represent the regions with maximum LOD scores that were detected in linkage analysis.

↓

Common 33 homozygous SNVs

Figure 2. Homozygosis mapping and exome sequencing. "Exome HOMozygosity" approach

overcome for overcoming the financial and analytic burden of whole genome sequencing, researchers have focused on the variations in the whole exome for the last several years. In recently developed commercially available platforms, it is aimed to overcome some limitations of exome sequencing by providing to researchers the the opportunity to capture promoters, highly conserved sequences, microRNAs, and 5' and 3' untranslated regions, in addition to exonic sequences. Despite this, it is anticipated that by facilitating the analysis of the hundredfold data, whole-genome sequencing instead of whole-exome sequencing, in the next several years.

Peer-review: Externally peer-reviewed.

Authors' Contributions: Conceived and designed the experiments or case: GKK, PD. Performed the experiments or case: GKK, PD. Analyzed the data: GKK, PD. Wrote the paper: GKK, PD. All authors have read and approved the final manuscript.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study has received no financial support.

REFERENCES

1. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 2007; 80(4): 588-604. [CrossRef]
2. Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. Nat Rev Genet 2006; 7(4): 277-82. [CrossRef]
3. Collins FS. Positional cloning moves from perditional to traditional. Nat Genet 1995; (9)4: 347-50. [CrossRef]
4. Vink JM, Boomsma DI. Gene finding strategies. Biol Psychol 2002; 61(1-2): 53-71. [CrossRef]
5. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 1987; 236(4808): 1567-70. [CrossRef]
6. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010; 11(1): 31-46. [CrossRef]
7. Gullapalli RR, Lyons-Weiler M, Petrosko P, Dhir R, Becich MJ, LaFramboise WA. Clinical integration of next-generation sequencing technology. Clin Lab Med 2012; 32(4): 585-99. [CrossRef]
8. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nat Methods 2010; 7(2): 111-8. [CrossRef]
9. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009; 461(7261): 272-6. [CrossRef]
10. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet 2012; 57(10): 621-32. [CrossRef]
11. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010; 7(4): 248-9. [CrossRef]
12. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 2003; 33(Suppl): 228-37. [CrossRef]
13. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet 2011; 48(9): 580-9. [CrossRef]
14. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Unlocking Mendelian disease using exome sequencing. Genome Biol 2011; 12(9): 228. [CrossRef]
15. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011; 12(11): 745-55. [CrossRef]
16. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009; 27(2): 182-9. [CrossRef]
17. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007; 39(12): 1522-7. [CrossRef]
18. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova, M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 2014; 15(2); 259-78. [CrossRef]
19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001; 29(1): 308-11. [CrossRef]
20. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature 2010; 467(7319): 1061-73. [CrossRef]
21. International HapMap 3 Consortium, Altshuler DM, Gibbs RA, , Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010; 467(7311): 52-8. [CrossRef]

22. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet 2012; 20(5): 490-7.**[CrossRef]**
23. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. Hum Genet 2011; 129(4): 351-70. **[CrossRef]**
24. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 2010; 42(1): 30-5. **[CrossRef]**
25. Krawitz PM, Schweiger MR, Rödelsperger C, Marcelis C, Kölsch U, Meisel C, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet 2010; 42(10): 827-9. **[CrossRef]**
26. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 2010; 42(6): 483-5. **[CrossRef]**
27. Hoischen A, van Bon BW, Rodriguez-Santiago B, Gilissen C, Vissers LE, de Vries P, et al. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat Genet 2011; 43(8): 729-31. **[CrossRef]**
28. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 2011; 43(6): 585-9. **[CrossRef]**
29. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. Nat Genet 2011; 43(9): 860-3. **[CrossRef]**
30. Hildebrandt F, Heeringa SF, Rüschendorf F, Attanasio M, Nürnberg G, Becker C, et al. A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet 2009; 5(1): e1000353. **[CrossRef]**
31. Tunçbilek E. Clinical outcomes of consanguineous marriages in Turkey. Turk J Pediatr 2001; 43(4): 277-9.
32. Bilgüvar K, Öztürk AK, Louvi A, Kwan KY, Choi M, Tatli B, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 2010; 467(7312): 207-10. **[CrossRef]**
33. Becker J, Semler O, Gilissen C, Li Y, Bolz HJ, Giunta C, et al. Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. Am J Hum Genet 2011; 88(3): 362-71.**[CrossRef]**
34. Pippucci T, Benelli M, Magi A, Martelli PL, Magini P, Torricelli F, et al. EX-HOM (Exome HOMozygosity): a proof of principle. Hum Hered 2011; 72(1): 45-53. **[CrossRef]**
35. Ozgül RK, Siemiatkowska AM, Yücel D, Myers CA, Collin RW, Zonneveld MN, et al. Exome sequencing and cis-regulatory mapping identify mutations in MAK, a gene encoding a regulator of ciliary length, as a cause of retinitis pigmentosa. Am J Hum Genet 2011; 89(2): 253-64. **[CrossRef]**