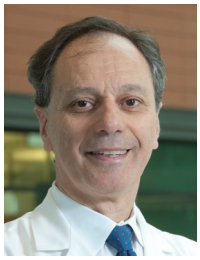

A.J. Marian, M.D.

SEQUENCING YOUR GENOME: WHAT DOES IT MEAN?

A.J. Marian, M.D.

The University of Texas Health Science Center, Houston, Texas; Texas Heart Institute, Houston, Texas

Abstract

The human genome contains approximately 3.2 billion nucleotides and about 23,500 genes. Each gene has protein-coding regions that are referred to as exons. The human genome contains about 180,000 exons, which are collectively called an exome. An exome comprises about 1% of the human genome and hence is about 30 million nucleotides in size. Today's technologies afford the opportunity to sequence all nucleotides in the human exome and even in the human genome. Given that more than three-quarters of the known disease-causing variants are located in the exome, and considering the cost and technical challenges in analyzing the whole genome sequence data, the focus of present research is primarily on whole exome sequencing (WES). While WES at the medical sequencing level is still expensive, it is becoming more affordable. Cost will not likely be a major barrier in the near future, and the data analysis is becoming less tedious. The most difficult challenge at the heart of medical sequencing is interpreting the findings. Each exome contains about 13,500 single nucleotide variants (SNVs) that affect the amino acid sequence, and a large number are expected to be functional variants. The daunting task is to distinguish the variants that are pathogenic from those that have minimal or no discernible clinical effects. While various algorithms exist, none are sufficiently robust. Thus, in-depth knowledge in genetics and medicine is essential for the proper interpretation of the WES findings. This review will discuss the potential applications of the WES data in the practice of cardiovascular medicine.

Basic Structure of the Human Genome

The human genome, a diploid genome, is comprised of 3.2 billion nucleotides that are packed into 23 pairs of chromosomes. It contains approximately 23,500 protein-coding genes. Each gene is comprised of the protein-coding segments, known as exons; the intervening sequences, known as introns; and the regulatory regions on each end of the gene (5′ and 3′ end regions). There are about 180,000 exons in each human genome that are collectively referred to as an exome. Since the exome occupies only about 1% of the genome, the size of an exome is roughly 30 million nucleotides; thus, approximately 99% of the human genome does not code for a protein. However, these regions by and large have biological functions that might affect gene expression and likely the clinical phenotypes. The current focus in medical sequencing is on the exome, as approximately three-quarters of the known pathogenic variants affect the protein-coding exons.

The Enabling Effects of "Disruptive" Sequencing Technologies

The high throughput DNA sequencing technologies have dramatically changed the landscape of genetic discoveries. The conventional technique of genetic linkage analysis in large families followed by sequencing, using the Sanger technique, of the genes residing at the mapped locus has all but been replaced with the new technologies, wherein millions of DNA fragments are sequenced simultaneously and in parallel. These high throughput sequencing (HTS) approaches have increased the output of a single sequencing reaction by several orders of magnitude, enabling sequencing of the entire human genome and a dozen exomes in a week. The enormity of these "disruptive" technologies is best illustrated by the fact that the initial sequencing of the human genome through the Human Genome Project took more than a decade, involved multiple sequencing centers, and cost approximately $3 billion.[1] Today, the entire human genome could be sequenced in a small laboratory at a cost of less than $10,000 and an exome at the cost of about $1,000. Despite these technical feats, the enormous size of the sequence readout and the complex genetic diversity of humans pose major challenges in applying whole genome sequencing and even whole exome sequencing (WES) at the bedside.

Genetic Diversity of Humans

Duplication of the genome is almost, but not entirely, perfect. The DNA polymerases and editing enzymes replicate the genome at a blazing speed with an amazing and near-perfect accuracy. The machinery that is responsible for genome duplication introduces one error for every 100 million nucleotides that it copies (10^{-8} error per base pair for the mammalian genome).[2-4] This error rate translates into approximately 30 new DNA variants in each offspring (de novo variants, as they are absent in the parents' genomes).[2-4] Given that the human species has evolved over 3.7 to 6.6 million years[5] and over billions of meiotic divisions (genome duplications), and in view of the introduction of approximately 30 de novo variants per meiosis, one might surmise the enormous diversity of the human genome.

Introduction of the new DNA sequence variants (DSVs)

throughout the evolution of humans has followed the population growth. The rapid expansion of the human population during the last 10,000 years, about 400 generations, has ultimately introduced a very large number of DSVs into the population genome.[6] Consequently, the vast majority of DSVs in the population genome are relatively new. These new variants, having had an inadequate time to spread among the population compared to older variants, are less common and often rare. Likewise, the new variants have not had adequate exposure to evolutionary selection pressure or a population drift; therefore, they generally are expected to exert larger biological effects. This is in contrast to ancient DSVs, which have had the chance to spread out and be subjected to selection pressure. Consequently, ancient variants are typically common and have small and often clinically negligible effects, as those with large effect sizes are typically eliminated over years of evolution.

The Plethora of DSVs in an Individual Genome/Exome

Each genome contains approximately 3.2 billion nucleotides, of which approximately 4 million nucleotides are variants as compared to the reference genome. Therefore, each individual has a variant nucleotide for every 800 nucleotides in the genome. With the current population level, every nucleotide is expected to be polymorphic even though the vast majority of such variants are rare due to their modern origins.[6, 7] Since de novo variants are introduced in each offspring, no two individuals, with the exception of monozygotic twins, are genetically identical at the DNA sequence level. This diversity also extends to each individual: because of the error rate of the DNA replication machinery and replication of certain cells, the replicating cells in an individual are a genetic mosaic.

Of the approximately 4 million DSVs in each genome, about 3.5 million involve only a single nucleotide and hence are called single nucleotide variants (SNVs) or single nucleotide polymorphisms (SNPs).[8-12] The remainder of the variants involve multiple nucleotides that include small insertion/deletions (indels), which are the second most abundant variants in the genome. In addition, the human genome contains variations that are duplications, deletions, inversions, and rearrangements, all of which are referred to as structural variations (SVs).[13, 14] SVs might involve several thousand to millions of nucleotides, increasing or decreasing the two copies of the genes or chromosomal segments. Such SVs are referred to as copy number variants (CNVs). The genome of Nobel Laureate Dr. James Watson, who along with Francis Crick, Maurice Wilkins, and Rosalind Franklin described DNA as a double-stranded helix, typifies the abundance of DSVs in an individual genome.[11] Dr. Watson's genome has 3.5 million SNVs and large insertions and deletions including several that encompass up to 1.5 million nucleotides.

Each exome contains approximately 13,500 nonsynonymous (ns) SNVs, which by definition affect the amino acid sequence of the encoded proteins (Table 1).[8, 9, 11, 12] While all nsSNVs have the potential to exert biological effects, the vast majority of the nsSNVs are expected to be clinically inconsequential; only a handful of nsSNVs in each exome are expected to exert major functional and clinical effects. On average, there are approximately 50 to 100 variants in each exome that have been linked to inherited disorders, largely through association studies. Among the notable variants in each exome are those that practically neutralize function of the encoded proteins and hence are called loss-of-function (LoF) variants. Among the LoF variants, each exome contains about 25 to 30 heterozygous and 2 to 3 homozygous nonsense variants that lead to premature truncation of the proteins, which are typically unstable and are degraded. Likewise, frameshift variants that alter the sequence of the amino acids in the protein often lead to premature truncation of the proteins. Variants that affect exon–intron splicing might lead to deletion of one or more exons or incorporation of a new exon, affecting the protein structure and function. Collectively, there are about 100 to 120 LoF variants in each exome, of which approximately 20 are homozygous. This means that each individual lacks approximately 20 proteins.[15]

Nucleotides (base pairs)	3.2×10^9
Protein-coding genes	23,500
Number of exons	180,000
Size of exome (base pairs)	30×10^6
DNA sequence variants (DSVs)	4×10^6
Single nucleotide polymorphisms (SNPs)	3.5×10^6
Nonsynonymous SNPs (nsSNPs)	13,500
Loss-of-function (LoF) heterozygous variants	100-120
Variants associated with inherited diseases	50-100
De novo variants	30

Table 1. Abundance of DNA sequence variants in the human genome

Genetic Variants and Human Diseases

It is important to emphasize that the clinical phenotypes are multifactorial in etiology, as they result from complex, typically nonlinear, and often stochastic interactions among various factors that contribute to the phenotype. Therefore, DSVs are only partly responsible for the clinical phenotype, even in single-gene disorders. The magnitude of the contributions of DSVs to the clinical phenotypes follows a gradient ranging from negligible to large.[16] On one end of the spectrum are the single gene disorders, whereby a single variant in a single gene is sufficient to cause the clinical phenotype. Therefore, the causal variant's contribution to the phenotype is quite large. Typically, the causal variants are rare in the population, as is the disease prevalence. On the opposite end of the spectrum are typically common variants that have small or clinically negligible effects. Common diseases such as coronary artery disease and hypertension are caused, at the genetic level, by a combination of common and rare variants.[17]

Extracting Clinically Useful Information from WES Data

Inherent to all medical tests are the technical limitations, and genetic testing by WES is no exception. The WES techniques are not immune to false-positive and false-negative results. WES captures approximately 85% to 90% of the exons in the genome, which means WES does not adequately cover between 1,000 to 2,000 genes.[18] Therefore, a "negative" result should not be considered

an absolute finding. Clinical (phenotypic) information should be used to analyze the sequence output for an adequate coverage of the known and candidate genes, with the understanding that this approach is inherently restricted and insufficient for discoveries of novel genes and mutations. It might also be necessary to capture the exons using an alternative method in scenarios wherein a strong genetic etiology is anticipated.

False positive calls are often more problematic and vary dramatically according to the sequencing platform used and the depth of coverage, i.e., how many times each nucleotide is sequenced. Various sequencing platforms have different false positive rates and some are not suitable for medical sequencing, wherein accuracy is of utmost importance. For medical sequencing, typically an average coverage of 100x or greater would be desirable. This relatively high coverage increases the cost of sequencing yet significantly reduces the burden of deciphering true from false allele calls. The significance of this point must be underscored, as even a very low false positive allele call of 1% is sufficient to introduce a large number of false calls to the readout and complicate clinical applications of the findings. Increasing the mean coverage rate reduces the false positive rate but does not totally eliminate them for various technical reasons. Accordingly, one has to merge the genetic data with the clinical information to discern the true causal variants from the false positive calls.

The biggest challenge with the medical application of WES is identifying the true pathogenic variants out of the 13,500 nsSNVs identified by WES. The following are various algorithms, including bioinformatics, used to restrict the number of putative pathogenic variants:

A. *Familial cosegregation.* Perhaps the most valuable information is segregation of the variant with inheritance of the phenotype in families. Each meiosis, with some caveats, reduces the number of the candidate pathogenic variants by 50%. Thus, the best way to restrict the number of putative causal variants is to include as many family members as possible. Likewise, sequencing of a trio, classically parents and the offspring, might increase the chance of identifying the causal variant. Therefore, the emphasis must be on families and less so on a single case or the proband. In a single individual, one at best could surmise the causal variant, although not with high certainty.

B. *nsSNVs.* By changing the amino acid sequence in the protein, the nsSNVs are more likely to be pathogenic than the synonymous variants, although there might be exceptions. The vast majority of the 13,500 nsSNVs in each exome are unlikely to be pathogenic.

C. *Rare variants.* Rare variants are more likely to be pathogenic than common variants for reasons that were discussed earlier. Common and uncommon variants that have a population frequency greater than the prevalence of the disease are unlikely to be pathogenic. A variant that is absent in the databases and therefore considered novel is a stronger candidate.

D. *LoF variants.* SNVs that result in premature truncation of the encoded protein, such as stop codon mutations and frameshift variants, are stronger pathogenic variants.

E. *De novo variants.* Variants that are present in the affected probands but absent in the healthy parents are strong candidates to be pathogenic.

F. *Functional rare variants in genes previously implicated in the pathogenesis of the phenotype.* Rare nsSNVs in known causal genes for the phenotype are likely but not definitively causal in an index case.

G. *Common and uncommon variants – nsSNV or otherwise – previously shown to be associated with a common phenotype.* The clinical impact of these variants is relatively modest.

Clinically Guided Classification of the DSVS

The algorithms used to identify the causal variants often do not result in a definitive isolation of the pathogenic variants but often provide a probable weight. To simplify the clinical decision making, we have classified the DSVs in the genome into five categories, as follows[17]:

A. *Disease-causing variants.* These variants are rare in each genome/exome and are typically responsible for single-gene diseases. They are typically missense, nonsense, or frameshift variants. When present, these variants cause the disease, although expressivity is variable and the major determinant of the severity of the phenotype. Several other genetic and nongenetic factors also contribute to the phenotypic expression of the disease.[19, 20] Identifying these variants through WES has the highest impact on the care of the individual and family.

B. *Likely disease-causing variants.* These variants are typically rare LoF variants that are absent in the general population. However, despite being functional and rare or even exclusive to an individual or family, these variants do not show a perfect cosegregation, partly because of low penetrance.[21, 22] These variants have the second largest effect sizes on the phenotype.

C. *Disease-associated variants.* This category encompasses those variants that have been associated with the clinical phenotype through allelic association studies, such as the genome-wide association studies. However, they typically are not the true pathogenic alleles but are in linkage disequilibrium with true causal alleles. These variants have relatively modest clinical impact and best guide the subsequent genetic studies to identify the true causal variants.

D. *Functional variants not linked to a disease.* These variants affect gene function and the encoded protein but have not been associated with a clinical phenotype. These variants comprise a large number of nsSNVs identified in each exome.

E. *Variants with unknown biological function.* This category encompasses the vast majority of variants in the genome or exome. These variants typically have not been characterized but are not expected to be pathogenic. Characterization of these variants might change their classification.

In clinical medicine, the focus is on the disease-causing variants (category 1) and likely disease-causing variants (category 2). The clinical impact of disease-associated variants (category 3) is modest. The remaining two categories, which biologically might not be inconsequential, currently do not have direct clinical implications. The number of clinically relevant variants in each exome is expected to be less than 100. These candidate variants must be analyzed meticulously through various genetic and clinical algorithms to further restrict the number of putative causal variants and ultimately identify the one or two pathogenic variants. Upon further elucidation of the clinical significance of such variants, the information might be used for early preclinical detection of those who carry the risk variant, cascade screening of

the family members, and possibly even tailoring medical therapy, implementing preventive measures, and avoiding drug toxicity.

Conclusions

WES affords the opportunity to identify the vast majority of DSVs in the exome and is being increasing applied for medical purposes. Data interpretation is exceedingly challenging, and none of the bioinformatics algorithms alone or in combination are sufficiently robust to identify the pathogenic variants. Garnering medical information from the WES data requires knowledge of the genetic diversity of the humans, etiological complexity of the clinical phenotype, and phenotypic variability of the diseases. Similarly, applying the WES data to clinical practice requires in-depth knowledge of medical genetics as well as clinical medicine. There is considerable emphasis on cross training the physicians and the geneticist to enhance applications of the genetic information to the practice of medicine. Both a team approach and training the current and next generation of cardiovascular physicians are necessary to understand and apply the modern molecular genetic discoveries to the care of patients. A great physician, however, always treats the patient who has the disease and not the disease itself or the medical or genetic test, as advocated by Sir William Osler, the father of modern medicine.

Conflict of Interest Disclosure: The author has completed and submitted the *Methodist DeBakey Cardiovascular Journal* Conflict of Interest Statement and none were reported.

Funding/Support: The author receives research funding from the National Institutes of Health, the Roderick MacDonald Foundation, the TexGen Fund from the Greater Houston Community Foundation and the George and Mary Josephine Hamman Foundation.

Keywords: genome; exome; whole exome sequencing; single nucleotide variants

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921.
2. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061-73.
3. Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, et al. Estimating the human mutation rate using autozygosity in a founder population. Nat Genet. 2012 Nov;44(11):1277-81.
4. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012 Aug 23;488(7412):471-5.
5. 5Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nat Genet. 2012 Oct;44(10):1161-5.
6. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012 May 11;336(6082):740-3.
7. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012 May 17;337(6090):100-4.
8. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoSBiol. 2007 Sep 4;5(10):e254.
9. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, et al. Genetic variation in an individual human exome. PLoS Genet. 2008 Aug 15;4(8):e1000160.
10. Pennisi E. Genomics. 1000 Genomes Project gives new map of genetic diversity. Science. 2010 Oct 29;330(6004):574-5.
11. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008 Apr 17;452(7189):872-6.
12. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, et al. The diploid genome sequence of an Asian individual. Nature. 2008 Nov 6;456(7218):60-5.
13. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007 Oct 19;318(5849):420-6.
14. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008 May 1; 453(7191):56-64.
15. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823-8.
16. Marian AJ. Nature's genetic gradients and the clinical phenotype. Circ Cardiovasc Genet. 2009 Dec;2(6):537-9.
17. Marian AJ, Belmont J. Strategic approaches to unraveling genetic causes of cardiovascular diseases. Circ Res. 2011 May 13;108(10):1252-69.
18. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012 May 29;44(6):623-30.
19. Daw EW, Chen SN, Czernuszewicz G, Lombardi R, Lu Y, Ma J, et al. Genome-wide mapping of modifier chromosomal loci for human hypertrophic cardiomyopathy. Hum Mol Genet. 2007 Oct 15;16(20):2463-71.
20. Marian AJ. Molecular genetic studies of complex phenotypes. Transl Res. 2012 Feb;159(2):64-79.
21. Rodriguez G, Ueyama T, Ogata T, Czernuszewicz G, Tan Y, Dorn GW II, et al. Molecular genetic and functional characterization implicate muscle-restricted coiled-coil gene (MURC) as a causal gene for familial dilated cardiomyopathy. Circ Cardiovasc Genet. 2011 Aug 1;4(4):349-58.
22. Chen SN, Czernuszewicz G, Tan Y, Lombardi R, Jin J, Willerson JT, et al. Human molecular genetic and functional studies identify TRIM63, encoding muscle RING finger protein 1, as a novel gene for human hypertrophic cardiomyopathy. Circ Res. 2012 Sep 14;111(7):907-19.