Published in final edited form as:

Ann Neurol. 2012 October ; 72(4): 628–629. doi:10.1002/ana.23660.

Exome sequencing for gene discovery: time does not stand still

Mark S. LeDoux, MD, PhD

Departments of Neurology and Anatomy and Neurobiology, University of Tennessee Health Science Center, Memphis, TN 38163, USA

The rapid evolution and continued refinement of whole-exome sequencing renders attempts to establish standards almost futile. Moreover, whole-genome sequencing may entirely replace whole-exome sequencing in the very near future. For all high-throughput DNA sequencing technologies, it should be emphasized that read accuracy is much more important than read depth. Even with deep reads, read errors are not uncommon. Theoretically, only shallow coverage would be needed if all reads were long, unbiased and 100% accurate. At this time, coverage depends on a variety of factors including method of exome capture, sequencing platform, average read length, definition of coverage, definition of the exome (CCDS +/- RefSeq, +/- GENCODE) and the number of distinct DNA samples analyzed per run. Ultra-deep coverage is prohibitively expensive in many situations and does not eliminate false positives and false negatives. Given the known limitations of next-generation sequencing technologies, Sanger sequencing (forward and reverse directions) of gDNA and cDNA was used to confidently exclude coding and non-coding (untranslated regions, introns and promoter) *TOR1A* mutations in our kindred with cervical dystonia.¹ Copy number variants were excluded with qPCR of gDNA.

A significant proportion of the variants reported in dbSNP, Exome Variant Server, and 1000 Genomes have not been validated with Sanger sequencing and are probably read errors. For example, Exome Variant Server reports 11 missense variants in *TOR1A* (a 5 exon gene associated with early-onset generalized dystonia) that are predicted to be possibly or probably damaging by PolyPhen and average sample read depth was > 80X for all of these variants. On the other hand, it would not be surprising to find potentially causal variants in large databases of putatively normal controls given the reduced penetrance and population prevalence of autosomal-dominant dystonia (*SGCE*, *GCH1*, *TOR1A*, *CIZ1* and *THAPI*).

The quality of data found in dbSNP is also questionable with some studies reporting false positive rates of 15 to 17%.² The dbSNP database is contaminated by single nucleotide differences which are artifacts due to the presence of paralogous sequences in the genome.³ False positive rates are probably much higher in databases derived from massively parallel sequencing (Exome Variant Server and 1000 Genomes). In brief, sequence variants identified with massively parallel sequencing should be validated with Sanger sequencing in the forward and reverse directions, correctly defined with Human Genome Variation Society (HGVS) nomenclature, characterized with molecular and cellular analyses, corroborated with carefully-designed case-control studies,⁴ and cautiously associated with neurological phenotypes.^{5,6}

*Corresponding author: Mark S. LeDoux, MD, PhD, University of Tennessee Health Science Center, Department of Neurology, 855 Monroe Avenue, Link Building, Suite 415, Memphis, TN 38163, USA, Phone: (901) 448-1662, Fax: (901) 448-7440, mledoux@uthsc.edu.

Acknowledgments

MSL is supported by the Dystonia Medical Research Foundation, Tyler's Hope for a Dystonia Cure, and NIH grants R01NS048458, R01NS060887, R01NS057722, R01NS058850, 5U01NS052592, 5U01AT000613 and U54NS065701.

References

1. Xiao J, Uitti RJ, Zhao Y, Vemula SR, et al. Mutations in *CIZ1* cause adult onset primary cervical dystonia. *Ann Neurol*. 2012; 71:458–469. [PubMed: 22447717]
2. Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics*. 2004; 20:1022–1032. [PubMed: 14764571]
3. Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum Mutat*. 2010; 31:67–73. [PubMed: 19877174]
4. Xiao J, Zhao Y, Bastian RW, et al. Thec-237_236GA>TT THAP1 sequence variant does not increase risk for primary dystonia. *Mov Disord*. 2011; 26:549–552. [PubMed: 21370264]
5. Doheny D, Danisi F, Smith C, et al. Clinical findings of a myoclonus-dystonia family with two distinct mutations. *Neurology*. 2002; 59:1244–1246. [PubMed: 12391355]
6. Seibler P, Djarmati A, Langpap B, et al. A heterozygous frameshift mutation in *PRKRA* (DYT16) associated with generalised dystonia in a German patient. *Lancet Neurol*. 2008; 7:380–381. [PubMed: 18420150]