Published in final edited form as:

Birth Defects Res A Clin Mol Teratol. 2014 August ; 100(8): 642–646. doi:10.1002/bdra.23276.

Missing Genetic Risk in Neural Tube Defects: Can Exome Sequencing Yield an Insight?

Deidre R. Krupp¹, Karen L. Soldano², Melanie E. Garrett², Heidi Cope², Allison E. Ashley-Koch², and Simon G. Gregory^{*,1}

¹Duke Molecular Physiology Institute, DUMC, 300 North Duke Street, Durham, NC 27701

²Duke Center for Human Disease Modeling, Duke University, 300 North Duke Street, Durham, NC 27701

Abstract

Background—Neural tube defects (NTD) have a strong genetic component, with up to 70% of variance in human prevalence determined by heritable factors. Although the identification of causal DNA variants by sequencing candidate genes from functionally relevant pathways and model organisms has provided some success, alternative approaches are demanded.

Methods—Next generation sequencing platforms are facilitating the production of massive amounts of sequencing data, primarily from the protein coding regions of the genome, at a faster rate and cheaper cost than has previously been possible. These platforms are permitting the identification of variants (de novo, rare, and common) that are drivers of NTD etiology, and the cost of the approach allows for the screening of increased numbers of affected and unaffected individuals from NTD families and in simplex cases.

Conclusion—The next generation sequencing platforms represent a powerful tool in the armory of the genetics researcher to identify the causal genetic basis of NTDs.

Keywords

next generation exome sequencing; de novo; rare and common variation; compound heterozygosity

Introduction

Neural tube defects (NTD) have a strong genetic component, with up to 70% of variance in human prevalence determined by heritable factors (Leck, 1974). This evidence is based on the co-occurrence of NTDs with known genetic syndromes; by twin studies that demonstrate increased concordance for monozygotic versus dizygotic twins; and by segregation analyses that suggest dominant and recessive modes of inheritance (Demenais et al., 1982; Fineman et al., 1982). However, the observation that most NTD cases are simplex rather than multiplex within multigenerational NTD families, combined with their widespread

occurrence, suggests that NTDs have a multifactorial genetic etiology. To characterize these genetic underpinnings, significant efforts have been made toward identifying common, rare or de novo DNA variants that confer elevated NTD risk. Animal models represent an attractive starting point to identify causal variants in humans. For example, rare variants have been identified in Vang-like 1 (*VANGL1*) [MIM: 610132] and Vang-like 2 (*VANGL2*) [MIM: 600533] genes (Kibar et al., 2007, 2009; Lei et al., 2010). However, of the more than 200 potential candidate NTD genes identified using mouse models (Harris and Juriloff, 2010) none are significant drivers of genetic risk in NTDs. However, the majority of the candidate genes derived from mouse models have not been studied, or adequately studied, in humans. Furthermore, mimicking the genetic complexity of human NTDs in model organisms is challenging because disease etiology in humans may result from the cumulative effect of several loci including rare, highly penetrant, and damaging functional variants, or a primary loci with secondary modifiers. Steps toward identifying these variants have already been made (Kibar et al., 2007, 2009; Lei et al., 2010; Marini et al., 2011; Robinson et al., 2012); however, they represent targeted screening of functional candidates some of which have been suggested by animal models.

An alternative to targeted candidate gene profiling is to carry out a genome wide screening of genetic variation in NTD case and control datasets. Although studies have successfully used common variants in genome-wide association studies to identify genes contributing to the development of disorders, it is likely that common DNA variation only confers modest risk and likely accounts for only a small portion of total disease heritability. To date, no genome-wide association studies of NTDs has been published (Hindorff et al., 2009). This “missing” heritability may in part be explained by rare variants that are uncommon in the general population but contribute to NTD risk by producing larger genetic effects than common variants (Manolio et al., 2009). Rare variant discovery has previously proceeded on a gene by gene basis using traditional Sanger sequencing of candidate genes identified using a priori data from animal models or based on functional relevance; however, the process is slow, costly, and inherently biased. It is clear that alternative approaches are required—the next steps toward identifying the genetic component of NTDs are being driven by the evolution of sequencing technologies.

Next Generation Sequencing

For decades the Sanger dideoxy-terminator sequencing provided the mainstay for genomic DNA profiling (Sanger and Coulson, 1975). Over time, improvements in safety and throughput have been made, however the technique is inherently limited in its capacity to generate large amounts of data. The advent of massively parallel sequencing in 2005 (Margulies et al., 2005) heralded a sea change by providing researchers with the ability to generate orders of magnitude more sequence in parallelized experiments. The initial platform was soon followed by others that were able to generate yet higher volumes of sequence data at cheaper per base costs (Stratton et al., 2009). Although these next generation sequencing (NGS) platforms generally generate shorter sequence reads than the Sanger sequencing method, they make up for their lack of read length by producing significantly greater sequence (or fold coverage) from parallelized polymerase chain reaction amplification of size fragmented template and sequence by synthesis. NGS does not

rely on the labeled dideoxy-terminators instead DNA is sequenced by incorporation of fluorescently tagged reversible dye-terminator nucleotides that are added to the synthetic molecule in a stepwise manner. Each A, C, G, and T nucleotide is individually labeled with a different fluorophore and, as they are incorporated, laser excitation and a charge-coupled device camera detect the order in which the nucleotides are synthesized against a genomic template (Bentley et al., 2008). More recent advances have moved away from charge-coupled device camera based approach toward detection of nucleotides by means of the release of a hydrogen ion, which is a by-product of nucleotide incorporation, and quantitated changes in pH using coupled silicon detector (Rothberg et al., 2001). This recent technology, while not matching the most often used NGS platform for production and application, promises faster and cheaper sequence generation in the future. Taken together, these and other NGS sequencing technologies (Mardis, 2013) are enabling NTD researchers to move away from solely relying on candidate gene identification by genetic linkage, association studies or model prediction, toward sequencing primarily the protein coding region of the genome, or “exome,” in affected individuals to identify common, but more likely rare variation, that contributes to NTD development.

Exome Sequencing

While NGS platforms have driven the precipitous drop in the per nucleotide cost of sequencing (<http://www.genome.gov/sequencingcosts/>), sequencing all 3.2 billion nucleotides of the human genome, which would provide the ultimate genetic profile of individuals affected by NTD, is currently not cost effective for most laboratories with per sample prices in the \$4000 to \$6000 range. Therefore exome sequencing, which targets 60 to 70 million nucleotides focused on mainly the protein coding regions of the genome, provides a cost effective alternative. The approach, which currently costs upward of \$700 per sample, uses complementary exon derived oligonucleotides designed to capture randomly fragmented DNA that has been preligated with NGS platform specific sequencing adapters. These adapters also contain “barcodes” that permit deconvolution of pooled exome sequences. Typically, sufficient exome sequence is generated to produce a minimum of 30-fold average coverage of all nucleotides of the captured targets (30×), however, because hybridization to oligonucleotide baits and polymerase chain reaction amplification is not uniform, there is significant variation in exon coverage (Fig. 1).

Exome sequencing has already identified causal DNA variants in genes associated with several disorders (Choi et al., 2012; Ng et al., 2009, 2010a, 2010b; Hoischen et al., 2010; Johnson et al., 2010; Krawitz et al., 2010; Otto et al., 2010; Pierce et al., 2010; Rios et al., 2010; Raffan et al., 2011), including severe brain malformations in cerebral cortical malformations (Bilguvar et al., 2010) and congenital heart defects. (Arrington et al., 2012) Furthermore, it is estimated that the protein coding region of the genome constitutes approximately 85% of disease-causing DNA variants (Choi et al., 2009). This suggests that exome sequencing represents an attractive option for identifying sequence variants (known or de novo variants) within affected individuals that segregate with NTD. However, the main challenge of the exome approach, which will only be magnified for whole genome sequencing, is the identification of variants, common, rare and de novo, that segregate with the disorder.

A critical component of exome sequencing is not just the profiling of individuals with NTD, but combining the study design with parents of the affected (trios), unaffected sib(s), or even extended multiplex families. The purpose of the design is to facilitate the detection of de novo (only within the affected individual) or inherited (transmitted from parent [s]) variants. The simplest design is to exome sequence only affected individuals and identify de novo DNA variants that are not present in the public domain within databases, such as dbSNP (Smigielski et al., 2000), the exome variant server (Tennessen et al., 2012) or the 1000 Genomes Project (Genomes Project et al., 2012). However, there continues to be a significant amount of uncharacterized genetic variation in the human genome and, of the 3000 to 3500 variants per individual identified by exome sequencing, a large number will be novel, previously uncharacterized variation that is not related to the affection status of the individual with NTD. Therefore, targeted (of newly identified candidate variants) or exome sequencing in parents will assess de novo status of the variant, while sequencing unaffected sibs and family members will determine if the variant(s) segregates with affection status.

Applications of Exome Sequencing to NTD

Although there have been no whole exome sequencing studies published to date, a key to identifying NTD variants by exome sequencing, regardless of allele frequency, will be a well-established pipeline of read quality control, sequence assembly and alignment, and variant calling. By now the computational methods for these steps are relatively stable and are available as self-compiled or commercial pipelines (Li and Homer, 2010). Furthermore, the statistical methods of variant analysis, partly derived from traditional association approaches, are being resolved (Stitzel et al., 2011). However, it is important to note that the previous successes of exome sequencing are predominantly Mendelian phenotypes with highly penetrant alleles. As previously discussed, NTDs are likely to be caused by a combination of de novo, rare and common variation. Thus the analytical approach data has to include identification of de novo DNA variants in the affected individual and family members; homozygous variants within the affected individual inherited from heterozygous parents; compound heterozygous variants, i.e., each parent contributes independent variants that results in an offspring that is a composite of potential risk alleles; and variants with dominant inheritance with reduced penetrance (e.g., transmitted from an unaffected parent). Further filtering by the putative functional pathways associated with NTD development, such as folate metabolism, cellogenesis or planar cell polarity pathways, help refine which variant to validate in the affected individual by Sanger sequencing (to ensure the variant is not a polymerase chain reaction artifact) and then replication in the family or control cohort.

Regardless of frequency, establishing a functional relationship between the variant(s) and NTD development is critically important. The first step can be the use of in silico prediction programs, such as Spliceman (Lim et al., 2011), Cryp-Skip (Divina et al., 2009), and Human Splicing Finder (Desmet et al., 2009) for noncoding variants, or PROVEAN (Choi et al., 2012) and Polyphen2 (Adzhubei et al., 2013) for coding variants, to predict the impact of exome variants on gene function. Next, it is essential to address whether the variant affects the gene at an RNA and protein level, and furthermore, does the variant (s) influence in vitro and in vivo development. Validation of these latter variants could include functional testing at a cellular level or in model systems, such as mice or zebra-fish, using knockout, CRISPR

(Cong et al., 2013) or morpholino (Summerton, 1999) technologies. It should also be noted, however, that exactly the same DNA variant may not be identified within all NTD affecteds in a population, but rather the de novo, rare or common variants associated will occur within a common set of genes or pathways that are functionally relevant to neural tube development that will contain variants implicated in the etiology of NTDs.

Conclusions

It is clear that exome sequencing presents an attractive approach for identifying DNA variants that may be associated with the development of NTD. However, while exome sequencing has a myriad of potential advantages, it is important to recognize that there are several inherent limitations of the technology. First, it is possible that the protein-coding regions of the genome do not contain all variants that are responsible for the development of NTDs, including enhancer/repressors, promoter elements or transcription factor binding sites. Although some noncoding regions of the genome are captured by oligonucleotide baits, such as micro and long noncoding RNAs and untranslated regions, there is a significant proportion of the human genome that is functionally important but is not captured by these assays, including transcription factor binding sites and long noncoding RNAs (Harrow et al., 2012).

Another important consideration is how to manage the significant volumes of data being generated by exome sequencing, let alone whole genome data when it becomes cheap enough to sequence affected and unaffected individuals. These issues not only relate to the storage and processing of the data using analysis pipelines, but also how to identify causal variants given that de novo, rare and common variation may contribute to a disorder like NTD. The statistical challenges inherent to these analyses are perhaps manifest in case control study designs in which 10's of thousands of samples may be required. The analytical approaches will need to address common and rare variant with (Cherkas et al., 2011; Li and Leal, 2008) and without (Wu and Zhi, 2013) rare variant collapsing which involves collapsing genotypes across variants and applying a univariate test (Do et al., 2012). The difficulties of case-control study designs may be somewhat addressed by resurrecting the NTD family samples that were all but ignored during the genome-wide association period. However, NTD family approaches are associated with their own caveats, i.e., are the variants identified from family based exome studies descriptive of the genetic risk of simplex NTD cases. Finally, how do de novo, rare or common variants interact individually or cumulatively with the significant number of environmental influences that are thought to contribute to the development of the disorder?

So, what is the future of next generation sequencing applications to NTD? The first and easiest to predict is the continued drop in cost and speed of sequencing. As previously described, cost of sequencing is decreasing at a much faster than predicted, and we are not far away from a genuine \$1000 genome. At these prices researchers will be able to afford at least exome sequencing of a significant number of individuals with NTD and their family members to identify segregation of putative disease related variants and NTDs. Perhaps emblematic of the speed of technological development, was the recent announcement that the platform that first heralded the new age of next generation sequencing (Margulies et al.,

2005) will no longer be sold, less than 10 years since it was introduced. However, new platforms (Mardis, 2013) are on the horizon and finally, the imminent reduction in cost and speed means that exome/genome sequencing will become a tangible reality in clinical setting (Biesecker et al., 2012).

Acknowledgments

The authors thank Dr. Christina Markunas for the descriptive figure used in the manuscript. This work was supported by: NIH, NS039818 (AAK/SGG) and a David H. Murdock research grant (SGG).

References

- Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*. 2013; Chapter 7(Unit7.20)
- Arrington CB, Bleyl SB, Matsunami N, et al. Exome analysis of a family with pleiotropic congenital heart disease. *Circ Cardiovasc Genet*. 2012; 5:175–182. [PubMed: 22337856]
- Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456:53–59. [PubMed: 18987734]
- Biesecker LG, Burke W, Kohane I, et al. Next-generation sequencing in the clinic: are we ready? *Nat Rev Genet*. 2012; 13:818–824. [PubMed: 23076269]
- Bilguvar K, Ozturk AK, Louvi A, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*. 2010; 427:207–210. [PubMed: 20729831]
- Cherkas Y, Raghavan N, Francke S, et al. Rare variant collapsing in conjunction with mean log p-value and gradient boosting approaches applied to Genetic Analysis Workshop 17 data. *BMC Proc*. 2011; 5(Suppl 9):S94. [PubMed: 22373203]
- Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*. 2009; 106:19096–19101.
- Choi Y, Sims GE, Murphy S, et al. Predicting the functional effect of amino acid substitutions and indels. *PloS One*. 2012; 7:e46688. [PubMed: 23056405]
- Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*. 2013; 339:819–823. [PubMed: 23287718]
- Demenais F, Le Merrer M, Briard ML, Elston RC. Neural tube defects in France: segregation analysis. *Am J Med Genet*. 1982; 11:287–298. [PubMed: 7081294]
- Desmet FO, Hamroun D, Lalande M, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*. 2009; 37:e67. [PubMed: 19339519]
- Divina P, Kvitkovicova A, Buratti E, Vorechovsky I. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. *Eur J Hum Genet*. 2009; 17:759–765. [PubMed: 19142208]
- Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet*. 2012; 21:R1–R9. [PubMed: 22983955]
- Fineman RM, Jorde LB, Martin RA, et al. Spinal dysraphia as an autosomal dominant defect in four families. *Am J Med Genet*. 1982; 12:457–464. [PubMed: 6751087]
- Genomes Project C. Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491:56–65. [PubMed: 23128226]
- Harris MJ, Juriloff DM. An update to the list of mouse mutants with neural tube closure defects and advances toward a complete genetic perspective of neural tube closure. *Birth defects Res A Clin Mol Teratol*. 2010; 88:653–669. [PubMed: 20740593]
- Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012; 22:1760–1774. [PubMed: 22955987]
- Hindorf LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009; 106:9362–9367. [PubMed: 19474294]

- Hoischen A, van Bon BW, Gilissen C, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet.* 2010; 42:483–485. [PubMed: 20436468]
- Johnson JO, Mandrioli J, Benatar M, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron.* 2010; 68:857–864. [PubMed: 21145000]
- Kibar Z, Bosoi CM, Kooistra M, et al. Novel mutations in VANGL1 in neural tube defects. *Hum Mutat.* 2009; 30:E706–E715. [PubMed: 19319979]
- Kibar Z, Torban E, McDermid JR, et al. Mutations in VANGL1 associated with neural-tube defects. *N Engl J Med.* 2007; 356:1432–1437. [PubMed: 17409324]
- Krawitz PM, Schweiger MR, Rodelsperger C, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet.* 2010; 42:827–829. [PubMed: 20802478]
- Leck I. Causation of neural tube defects: clues from epidemiology. *Br Med Bull.* 1974; 30:158–163. [PubMed: 4619574]
- Lei YP, Zhang T, Li H, et al. VANGL2 mutations in human cranial neural-tube defects. *N Engl J Med.* 2010; 362:2232–2235. [PubMed: 20558380]
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83:311–321. [PubMed: 18691683]
- Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform.* 2010; 11:473–483. [PubMed: 20460430]
- Lim KH, Ferraris L, Filloux ME, et al. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc Natl Acad Sci U S A.* 2011; 108:11093–11098. [PubMed: 21685335]
- Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009; 461:747–753. [PubMed: 19812666]
- Mardis ER. Next-generation sequencing platforms. *Annu Rev Anal Chem.* 2013; 6:287–303.
- Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005; 437:376–380. [PubMed: 16056220]
- Marini NJ, Hoffmann TJ, Lammer EJ, et al. A genetic signature of spina bifida risk from pathway-informed comprehensive gene-variant analysis. *PloS One.* 2011; 6:e28408. [PubMed: 22140583]
- Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet.* 2010a; 42:790–793. [PubMed: 20711175]
- Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet.* 2010b; 42:30–35. [PubMed: 19915526]
- Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009; 461:272–276. [PubMed: 19684571]
- Otto EA, Hurd TW, Airik R, et al. Candidate exome capture identifies mutation of SDCCAG8 as the cause of a retinal-renal ciliopathy. *Nat Genet.* 2010; 42:840–850. [PubMed: 20835237]
- Pierce SB, Walsh T, Chisholm KM, et al. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. *Am J Hum Genet.* 2010; 87:282–288. [PubMed: 20673864]
- Raffan E, Hurst LA, Turki SA, et al. Early diagnosis of Werner's syndrome using exome-wide sequencing in a single, atypical patient. *Front Endocrinol (Lausanne).* 2011; 2:8. [PubMed: 22654791]
- Rios J, Stein E, Shendure J, et al. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum Mol Genet.* 2010; 19:4313–4318. [PubMed: 20719861]
- Robinson A, Escuin S, Doudney K, et al. Mutations in the planar cell polarity genes CELSR1 and SCRIB are associated with the severe neural tube defect craniorachischisis. *Hum Mutat.* 2012; 33:440–447. [PubMed: 22095531]
- Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature.* 2011; 475:348–352. [PubMed: 21776081]
- Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol.* 1975; 94:441–448. [PubMed: 1100841]

- Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* 2000; 28:352–355. [PubMed: 10592272]
- Stitzel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* 2011; 12:227. [PubMed: 21920052]
- Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature.* 2009; 458:719–724. [PubMed: 19360079]
- Summerton J. Morpholino antisense oligomers: the case for an RNase H-independent structural type. *Biochimica et biophysica acta.* 1999; 1489:141–158. [PubMed: 10807004]
- Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science.* 2012; 337:64–69. [PubMed: 22604720]
- Wu G, Zhi D. Pathway-based approaches for sequencing-based genome-wide association studies. *Genet Epidemiol.* 2013; 37:478–494. [PubMed: 23650134]

Figure 1.

A representation of exome sequencing of a target gene, aligned reads, and read depth. Called variants are listed as differently colored vertical bars in the depth track.