Review

Cell PRESS

Feature Review

A *de novo* convergence of autism genetics and molecular neuroscience

Niklas Krumm[1], Brian J. O'Roak[1], Jay Shendure[1], and Evan E. Eichler[1,2]

[1] Department of Genome Sciences, University of Washington, Seattle, WA, USA
[2] Howard Hughes Medical Institute, University of Washington School of Medicine, Seattle, WA, USA

Autism spectrum disorder (ASD) and intellectual disability (ID) are neurodevelopmental disorders with large genetic components, but identification of pathogenic genes has proceeded slowly because hundreds of loci are involved. New exome sequencing technology has identified novel rare variants and has found that sporadic cases of ASD/ID are enriched for disruptive *de novo* mutations. Targeted large-scale resequencing studies have confirmed the significance of specific loci, including chromodomain helicase DNA binding protein 8 (CHD8), sodium channel, voltage-gated, type II, alpha subunit (SCN2A), dual specificity tyrosine-phosphorylation-regulated kinase 1A (DYRK1A), and catenin (cadherin-associated protein), beta 1, 88 kDa (CTNNB1, beta-catenin). We review recent studies and suggest that they have led to a convergence on three functional pathways: (i) chromatin remodeling; (ii) wnt signaling during development; and (iii) synaptic function. These pathways and genes significantly expand the neurobiological targets for study, and suggest a path for future genetic and functional studies.

Introduction

The identification of genes underlying ID and ASD has been most successful for syndromic Mendelian or monogenic disorders – for example, *FMR1* (Fragile-X syndrome, [1]), *MECP2* (Rett syndrome, [2]), or *UBE3A* (Angelman syndrome, [3]). Together, however, these syndromes are estimated to account for less than 10% of ASD/ID, suggesting the presence of additional genes and etiologies. Initial population-based studies failed to identify single genes of major effect and few major common risk variants have been replicated, despite the strong observed heritability of these diseases [4–7]. By contrast, targeted and genome-wide microarray studies revealed that large *de novo* copy number variants (CNVs) were significantly enriched among probands when compared to unaffected siblings and/or controls [8–14], a finding that echoed the earlier discovery of large chromosomal aberrations in ASD and ID. Both initial and subsequent higher-resolution studies estimate that 8% of sporadic ASD cases carried a *de novo* CNV, as compared with only 2% of unaffected siblings [11,12]. Furthermore, among children with general developmental delay (DD) and ID, rare large *de novo* CNVs are thought to account for up to 15% of disease burden [13]. Although individually rare, some of these CNVs were in fact recurrent mutations, mediated by locus-specific genomic instability [14], and many of these same recurrent CNVs observed initially in patients with ID [15] or ASD [16] have been identified in adults with epilepsy [17], bipolar disorder [18], or schizophrenia [19,20], suggesting overlap in the genetic etiology of these disorders.

The discovery of an aggregate burden of large *de novo* CNVs and the identification of recurrent events signaled a new paradigm for ASD and ID genetics. Although specific CNVs are individually rare, combined they account for a significant fraction of cases, indicating the presence of considerable locus heterogeneity of ASD and ID. The *de novo* nature of these CNVs, together with their absence in the general population, suggests they represent a class of highly deleterious and highly penetrant mutations. Their underlying genetic model does not explicitly fit a recessive model of disease because CNVs are primarily present as

Glossary

CNV (copy number variant): loss or insertion of DNA, typically larger than 50 bp and often up to several megabases.
Connected component: a set of connected nodes that are part of a PPI network and can represent a pathway, complex protein structure, or cellular function.
GC bias: the tendency for sequencing reactions to produce fewer reads in regions of the genome with a high fraction of GC base pairs.
Hidden species problem: a method for estimating an unknown number of classes (species) from a distribution of observed counts.
Indel (insertion/deletion): loss or insertion of DNA, between 1 and 50 bp in length.
Loss-of-function or truncating mutation: a nonsense, frameshift, or splice-site mutation that prevents complete translation of a functional protein.
Missense mutation: a mutation that alters the amino acid composition of a protein but does not prohibit its complete translation.
PPI network: a protein–protein interaction network that defines 'nodes' as proteins and 'edges' as interactions (which may be physical, expression-based, or computationally predicted).
Sequence coverage: the average or median number of sequence reads per genomic base pair in a sequencing experiment. Higher coverage enables more accurate discovery of variants.
SNP/SNV (single nucleotide polymorphism/variant): single-base changes in DNA. Typically, SNPs are higher frequency and refer to alleles observed to be segregating in a population.

Corresponding author: Eichler, E.E. (eee@gs.washington.edu).
Keywords: autism genetics; autism spectrum disorder; copy number variant; exome sequencing; intellectual disability; single nucleotide variant.

0166-2236/$ – see front matter
 http://dx.doi.org/10.1016/j.tins.2013.11.005

hemizygous deletions or duplications. These mutations alter the dosage of genes but do not completely abolish their presence. Collectively, these observations support a complex disease/rare variant model for ASD, in which a proportion of etiologic risk is conferred by very rare variants and *de novo* mutations.

The commoditization of next-generation or 'massively parallel' sequencing represents a turning point in human genetics and makes it possible to discover sequence-level variants across nearly all coding regions ('the exome') or the whole genome (Box 1). These methods were first applied to confirm point mutations underlying Mendelian disorders [21], and subsequent pilot studies demonstrated that family-based (trio) exome sequencing could discover pathogenic mutations in simplex ID [22] or ASD [23]. In the past year, this paradigm of *de novo* mutation discovery using exome sequencing of parent–child trios has been expanded to about 1000 ASD or ID families, resulting in the first detailed picture of how *de novo* coding mutations contribute to these disorders.

In this review, we synthesize the results of recent large-scale exome sequencing studies of ASD and ID [24–29] and summarize their implications for human neurodevelopmental genetics. There are three themes. (i) Exome sequencing of ASD/ID families has revealed a significant excess of *de novo* mutations in probands when compared to unaffected siblings and has identified novel candidate genes contributing to the neurological deficits. We note that the strongest effects are observed for *de novo* loss-of-function (or truncating) mutations (see Glossary), which prematurely truncate the protein due to frameshift and nonsense mutations. (ii) Both CNV and exome sequencing data suggest that no single gene will account for more than 1% of autism cases; rather, rare mutations in hundreds of genes may contribute to ASD or ID. (iii) Analyses of network connectivity further implicate potentially important neurodevelopmental and synaptic pathways in ASD and ID. Collectively, these studies represent a significant step forward for neurodevelopmental disorders providing a springboard for understanding their neurobiological underpinnings. We aim to focus on the molecular convergence revealed by these studies; for readers interested in other aspects of this topic, we suggest excellent reviews on ASD neurobiology [30], *de novo* mutation [31,32], and exome sequencing [33]. We emphasize that although this review is focused on the insights gained by considering a *de novo*/rare variant model of ASD and ID genetics, other genetic etiologies are implicated in ASD as well (for reviews, see [34,35]) and no single etiology is likely to be fully independent of other etiologies or of environmental

Box 1. Current sequencing technologies and their limitations

Whole-genome sequencing (WGS) provides the 'most complete' view of genomic variation and can detect SNVs, indels, and CNVs irrespective of frequency in a genome-wide fashion (although the power to detect events across the genome is not uniform based on local genomic composition). High costs and difficulty in the interpretation of nongenic variants are bottlenecks in wide-scale application of WGS.

Exome sequencing combines next-generation sequencing technologies with the targeted capture and amplification of exons (approximately 40 Mbp of the genome) in order to reduce the total amount of sequencing needed for the accurate determination of mutations in (or near) exons. However, the efficacy of the capture depends on the percentage GC nucleotide composition of the targeted sequence (also known as 'GC capture bias'). For exons with especially high GC content, exome sequencing can fail to produce enough coverage for accurate variant detection and calling. As shown in Figure I, several of the most GC-rich exons (blue bars) in the *SHANK3* gene (right) have inadequate coverage (black lines indicate mean coverage of a single family), preventing accurate assessment of variants in those exons. By contrast, the lower GC content (green histogram) of the *CHD8* gene (left) results in far higher and more uniform coverage of exons.

Molecular inversion probe (MIP) sequencing is a cost-effective targeted assay that uses custom oligonucleotides to efficiently capture target sequence. Unlike exome sequencing, however, MIPs use flanking primer arms and polymerase extension to capture the desired DNA target, thus reducing GC capture and other biases. In addition, MIP assays can be performed as highly multiplexed reactions, allowing for sequencing of targeted in genes in up to 192 individuals per sequencing reaction.

Genome-wide association studies (GWAS) are a widely used assay that leverage SNP markers (detected via hybridization to oligonucleotide arrays) to tag genomic regions and associate them with disease based on a comparison of cases and controls. Historically, GWAS studies have been limited to assaying common genetic variation and cannot detect novel or rare SNVs.

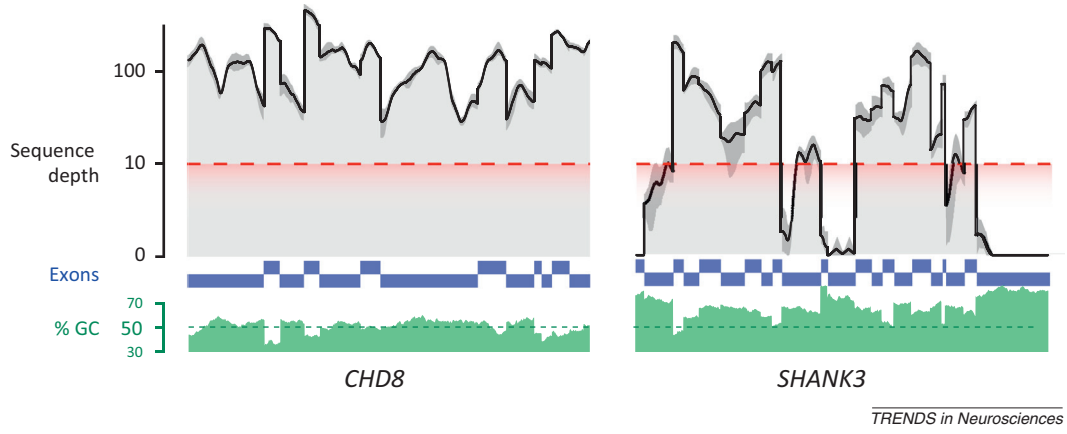

Figure I. High genomic GC nucleotide content (green histogram) hinders whole-exome sequencing for some genes, such as *SHANK3* (right). Individual coding exons are shown in blue with nongenic sequences removed. Black lines indicate mean sequence depth for a single ASD trio and dark gray intervals indicate maximum and minimum depth across the family. The red dashed line indicates the minimum threshold required for accurate variant detection.