Trends in Neurosciences

CellPress REVIEWS

Review

# The Role of *De Novo* Noncoding Regulatory Mutations in Neurodevelopmental Disorders

Tychele N. Turner [1] and Evan E. Eichler [1,2,*]

**Advances in sequencing technology have significantly expanded our understanding of the genetics of autism and neurodevelopmental disorders (NDDs). Continued technological improvements and cost reductions have now shifted the focus to investigations into the functional noncoding portions of the genome. There is a patient trend toward an excess of *de novo* and potentially disruptive mutations among conserved noncoding sequences implicated in the regulation of genes. The signals become stronger when restricted to genes already implicated in NDDs, but *de novo* mutation in such elements is estimated to account for <5% of patients. Larger sample sizes, improved variant detection, functional testing, and better approaches to classify noncoding variation will be required to identify specific pathogenic variants underlying disease.**

## Highlights

Recent sequencing advances have allowed whole-genome sequencing (WGS) of large numbers of individuals with autism.

WGS data are beginning to provide insight into the potential contribution of noncoding variants in neurodevelopmental disorders.

A trend has been observed for an excess of *de novo* variants in conserved noncoding regions among autism patients with no obvious genic cause.

The noncoding *de novo* mutation signature is stronger near genes already implicated in autism.

Increased sample size is necessary to further our understanding of these noncoding signals and to refine the picture of their action at the gene level.

Improved variant detection and functional classification of noncoding elements will increase our ability to detect pathogenic variants.

Ultimately, functional testing will be critical to understanding the effect of mutations on gene expression and their relevance to disease.

## Genetic Architecture and Genome Technology

Investigations into the genetic basis of autism and other neurodevelopmental disorders (NDDs) are limited by sample size and the scope and sensitivity of the genomic technology employed. Single-nucleotide polymorphism microarray data, for example, provided access to common variants under a genome-wide association study (GWAS) model as well as to large copy number variants (CNVs). Later, when **whole-exome sequencing (WES)** (see Glossary) became commonplace, the focus shifted to *de novo* and rare inherited variants in the protein-encoding portions of our genome.

Each technological advance provided unique insights into the **genetic architecture** of autism and other NDDs. While GWASs have identified only a few consistent variants or loci associated with autism [1,2], they did provide insights into its heritability, suggesting that common variants must contribute substantially to risk [3] or, at a minimum, to sensitizing an individual to develop autism. There is an emerging paradigm of polygenic risk scores as playing an important role, and it is likely that with larger sample sizes more definitive autism risk loci will become apparent [4]. The discovery of an excess of large CNVs among 8–15% patients with autism and NDD [5] was important because it suggested a genetic model where *de novo* and rare inherited mutations created gene-expression dosage imbalances early in development leading to the development of disease. WES extended this model confirming the importance of *de novo* [6–10] and rare inherited mutations [11] that disrupted gene function in an estimated 21% of individuals with autism [8] and 42% of those with NDD [12]. Importantly, the nature of the mutations discovered by WES provided the specificity required to identify new genes and pathways underlying NDD, leading to the discovery of 124 genes reaching exome-wide significance and 253 genes (5% FDR) with an excess of recurrent protein-damaging *de novo* mutations (DNMs) [13].

Combined rare/*de novo* coding variants and CNVs are now thought to contribute to approximately 20–30% [8,14] of individuals with autism. This means, of course, that for the majority of individuals with autism there is no obvious identified genetic cause. Given that the most recent

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

*Correspondence:
eee@gs.washington.edu (E.E. Eichler).

https://doi.org/10.1016/j.tins.2018.11.002

heritability estimates for autism are now ~80% [15], other genetic risk variants, both rare and common, await discovery and characterization. Of note, it is possible that both rare and common variants might contribute to disease in the same patients. For example, large-effect mutations such as CNVs might predispose to developmental delay, while the background common variants confer phenotypic specificity.

**Whole-genome sequencing (WGS)** data provide, in principle, access to the complete spectrum of human genetic variation in an individual irrespective of the class or frequency of a variant. Among patients where no obvious genic or CNV cause has been identified, there has been a shift in focus to investigations of the functional **noncoding** regions of the genome that are important in the regulation of gene expression [14,16–18]. This includes (but is not limited to) the untranslated regions (UTRs) of genes (3′ UTR, 5′ UTR), **enhancers**, **promoters**, noncoding RNAs (e.g., miRNA, piRNA), miRNA binding sites, and **topologically associating domain (TAD)** boundaries (see Figure I in Box 1). Variants identified in noncoding DNA have long been known to cause Mendelian diseases [19,20] and contribute to complex genetic traits [21–25]. For example, one of the most common causes of developmental delay and autism, fragile X syndrome, is due to the hyperexpansion of a CGG repeat sequence located in the 5′ UTR of the X chromosome gene *FMR1*. The expansion of the CGG repeat promotes hypermethylation of the promoter region leading to silencing of *FMR1* [26,27]. Despite this simple model of functional effect, almost all cases of fragile X syndrome are the result of CGG repeat expansion, with relatively few examples of loss-of-function mutations in the protein-coding portion of *FMR1* despite the sequencing of tens of thousands of individuals with idiopathic autism and NDD (see denovo-db [28] version 1.6.1).

While there is no question that noncoding variation will play a role in human NDDs, a major challenge has been defining the functional elements and interpreting mutational effects. Large-scale efforts like the ENCODE [29] and Roadmap Epigenomics [30] projects have attempted to systematically catalog the noncoding regions of the genome in different cell types and tissues. Notwithstanding these valiant efforts, a definitive set of functional noncoding elements (NCEs) does not yet exist. In this review, we consider the current evidence for the role of noncoding variants based on large-scale sequencing studies in autism and other NDDs and summarize the different approaches undertaken over the past few years to address this question. Our synthesis of the available data provides further support for specific NCE categories but also highlights potential pitfalls going forward. We lastly highlight some of the remaining questions in terms of classifying this variation, statistical testing, the application of newer, deep-learning-based approaches to further refine the NCEs, and the critical role of large-scale functional testing.

### Box 1. Types of Noncoding Variants

Functional noncoding regions include promoters, enhancers, repressors, insulators, UTRs, and noncoding RNA. Promoters (Figure IA) are 5′ proximal to the transcription start site and are the location at which the core transcription machinery binds to the DNA. Enhancers (shown in Figure IA bound by transcription factors and linked to the promoter) and repressors (not shown) are position-independent sequences involved in increasing and decreasing the transcriptional activity of genes, respectively. They can be close to or far from the transcriptional start site. UTRs (Figure IB) map to the 5′ and 3′ ends of genes and are part of the full-length transcript. The 5′ UTR contains the sequence for translation initiation and can also have other regulatory activity (exhibited as a hairpin here). The 3′ UTR (Figure IC) contains the sequence for the termination of translation and frequently harbors miRNA binding sites important for repression of translation. At a higher level, the genome is organized into TADs (Figure ID) that comprise genes and regulatory elements. These TADs are flanked by insulator elements that can either block enhancer activity on a gene or maintain the boundaries for a set of genes or regions contained within the TADs. In addition to these elements, there are a variety of noncoding genes, such as tRNA, rRNA, miRNA, siRNA, piRNA, snoRNA, and lncRNA among many others, that regulate the transcription, translation, and splicing of genes or play a role in chromatin organization (e.g., XIST and X chromosome inactivation).

## Glossary

**Clustered regularly interspaced short palindromic repeats (CRISPR) technology:** molecular tool for editing the genome; can be used in a high-throughput manner.

***de novo* variant:** genetic variant present in a child that is not present in either parent.

**DNase I hypersensitive site (DHS):** location in the genome where DNase I is able to cleave DNA.

**Enhancer:** DNA sequences in the genome that raise the level of transcription of a gene.

**Genetic architecture:** the complete understanding of the genetic factors underlying a phenotype.

**Insertion/deletion (indel):** small genetic variant (1–49 bp in length) that either removes or adds bases to the genome.

**Machine learning:** computational approach that uses artificial intelligence to train on input data and make future predictions without being explicitly programmed.

**Massively parallel reporter assay:** high-throughput testing of thousands of DNA sequences for regulatory activity.

**Noncoding:** the part of the genome that does not encode proteins.

**Oligogenic:** intermediate between monogenic and polygenic models of disease where a few genes or loci of relatively large effect play a role in the resulting phenotype.

**Promoter:** DNA sequences in the genome that are close to and encompass the start site of transcription of a gene.

**Regulatory:** having an effect on the transcription of a gene.

**Short tandem repeat (STR):** DNA repeats with units that are typically 2–6 bp in length and vary polymorphically between individuals.

**Single-nucleotide variant (SNV):** DNA variant that changes one base to another.

**Structural variant (SV):** DNA variant that is greater than or equal to 50 bp in length and involves the deletion, duplication, inversion, or translocation of sequence.

**Topologically associating domain (TAD):** section of the genome that physically interacts 'only' with itself and typically varies in size from thousands to millions of base pairs; noncoding regulatory sequences in a

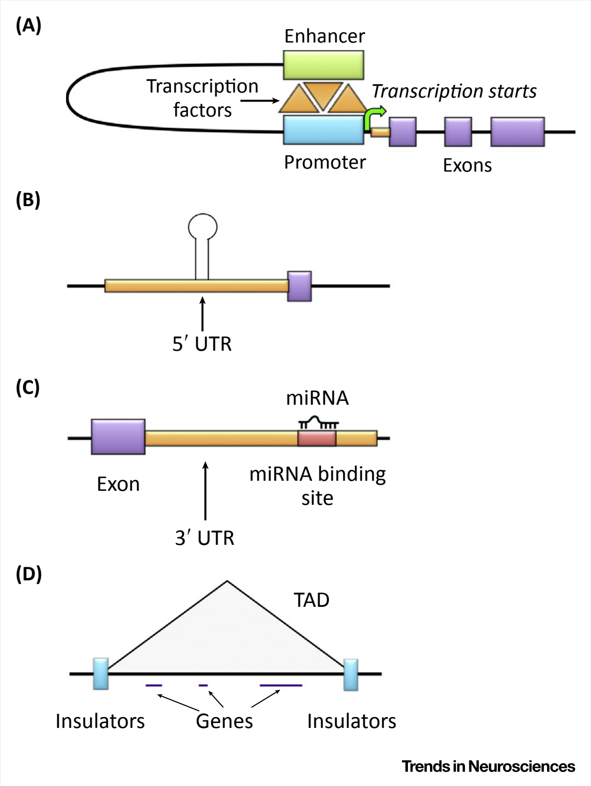


Figure I. Schematic of Noncoding Variants. (A) Diagram of the promoter region of a gene where the promoter is bound by transcription factors and linked to an enhancer. The green arrow indicates the location of the start of transcription. (B) Shown is an example 5′ untranslated region (UTR) with a hairpin indicating a site of regulatory activity. (C) An example 3′ UTR is exhibited here with an miRNA binding site. An miRNA is also shown bound to this site to exemplify a location of regulatory activity. (D) Diagram of a topologically associating domain (TAD) with genes inside it and insulators shown at the boundaries of the TAD.

TAD are thought to regulate only genes in the same TAD.

**Transcription factor binding site (TFBS):** location in the genome where transcription factors bind to DNA.

**Whole-exome sequencing (WES):** DNA sequencing approach that primarily targets and sequences the ~1.5% of the genome that is protein coding.

**Whole-genome sequencing (WGS):** DNA sequencing approach that assesses 'all' of the accessible portions of the genome; includes popular short-read sequencing approaches (e.g., Illumina) as well as longer-read sequencing technologies (e.g., Oxford Nanopore, Pacific Biosciences).

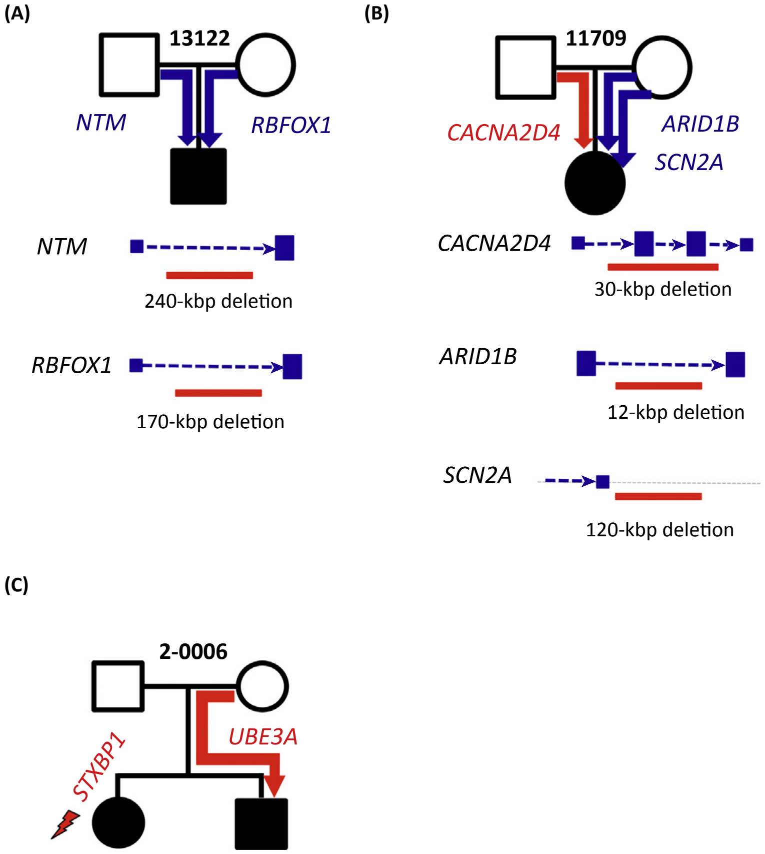
## Large-Scale Next-Generation Sequencing Studies

Simplistically, there have been two different sequencing-based approaches for the assessment of noncoding variation: namely, WGS and targeted sequencing of affected individuals and/or their family members or other controls. Targeted sequencing includes WES studies because of its potential to recover **regulatory** variation within the UTR portions of genes but, of course, coding variation was the primary target of such efforts. Most published studies have focused on ***de novo* variants**, assuming a dominant model of disease where *de novo* or disruptive mutation in regulatory DNA might interfere with the normal expression of a gene. WGS is less biased because it puts no *a priori* knowledge on the initial experimental design compared with targeted sequencing, which preselects putative functional regions for testing.

### WGS Studies of Autism and NDD

The first WGS studies were limited in scope to 20–85 parent–child trios [16,31–33], were generated using a variety of sequencing platforms, and focused on establishing a framework for the discovery of new mutations and characterizing their patterns. The studies consistently reported increased detection sensitivity for DNM in protein-coding regions as well as smaller gene-disruptive CNVs increasing the diagnostic yield compared with WES platforms

[16,32,33]. Considerable genetic heterogeneity was noted even among families with multiple affected individuals where different autism-relevant mutations appear to be segregating (Figure 1) [33,34]. In ~10% of families, WGS identified more than one potential risk variant, suggesting a multifactorial rare-variant model of disease for some individuals with autism (Figure 1), although the number of families was limited and still underpowered [16]. Turner and colleagues reported nominal enrichment of *de novo* and private disruptive mutations in putative regulatory regions of the fetal brain as defined by DNase I hypersensitivity when comparing probands with their unaffected siblings. The fetal brain had been strongly implicated previously based on gene expression and protein–protein interaction network analysis of autism risk genes identified from WES data [35–37] and reviewed in [38].



Trends in Neurosciences

Figure 1. Genetic Heterogeneity and Multiple Disruptive Mutations in Autism Pedigrees. In ~10% of patients with autism, genome sequencing reveals multiple *de novo* and disruptive mutations in different genes and their regulatory DNA. For example, (A) a child from Simons Simplex Collection (SSC) family 13122 inherits two large deletions affecting putative regulatory regions of the autism-risk genes *NTM* and *RBFOX1*. (B) Similarly, a child in SSC family 11709 inherits three different deletions with two affecting putative regulatory regions of the autism-risk genes *ARID1B* and *SCN2A*. (C) Examination of families with multiple affected individuals frequently finds that a genetic risk variant segregates to only one of the two children or that different affected individuals each carry a different risk variant. For example, in MSSNG multiplex family 2-0006 [33], one autistic offspring carries a *de novo* loss-of-function event in *STXBP1* while the other affected male sibling carries a maternally inherited loss-of-function event in *UBE3A*. (A,B) adapted from [16]. MSSNG family 2-0006 was studied in [33].

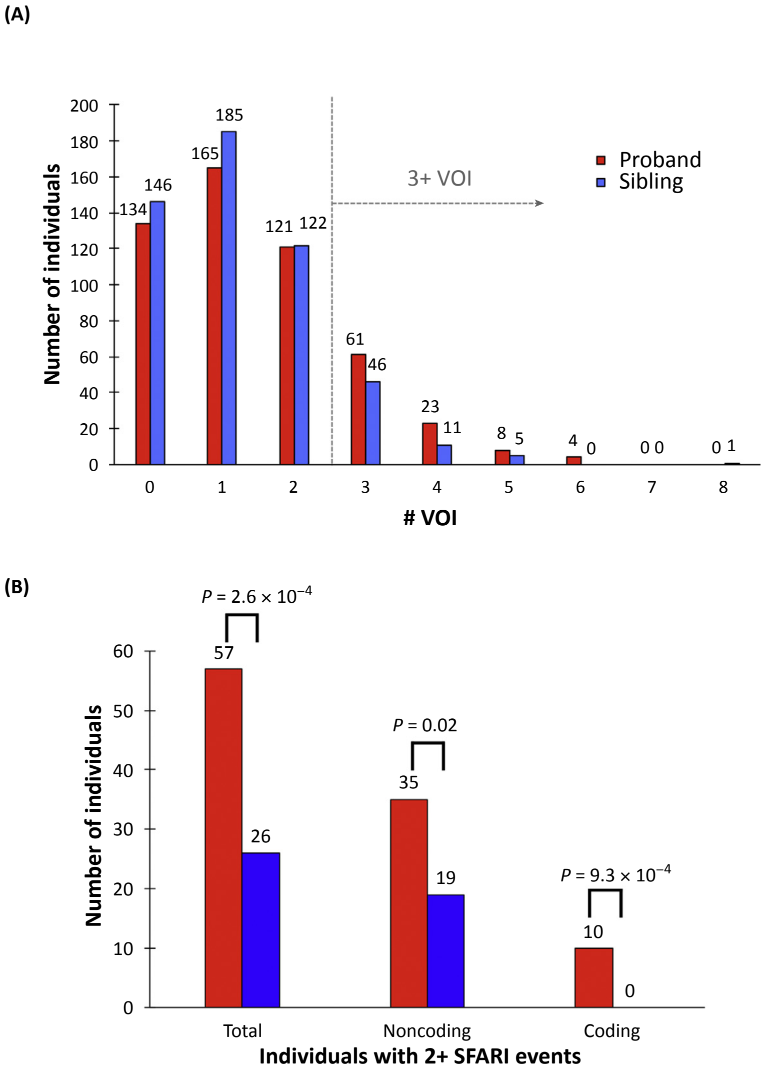
Later WGS studies were significantly larger (200–500 families), focusing almost exclusively on autism [14,17,18,39–42] with a subset emphasizing noncoding variation [14,17,18,40–42] (Table 1). In one study, for example, *de novo* variants were assessed in 200 parent–child families from the MSSNG autism cohort [40] and compared with *de novo* variants in a published control cohort (258 families) called Genome of the Netherlands (GoNL) [43]. The study reported DNM enrichment for noncoding variants in the conserved part of UTRs, variants that caused exon skipping, and **transcription factor binding sites (TFBSs)** located in **DNase I hypersensitive sites (DHSs)** mapping near genes. The experimental design was criticized, however, because the controls were sequenced at significantly lower sequence read depth (13-fold vs 32-fold) using a different sequencing platform as part of a different study, although steps were taken to minimize differences in sensitivity.

Phase I of the Simons Simplex Collection (SSC) comprised ~520 families selected to be negative for 'known' disease-causing mutation events. The study design had the advantage that an unaffected sibling was sequenced from each family using the same sequencing platform and same sequence depth. The genome sequence from the unaffected child served as a genetic control for the pattern of DNM compared with the autism proband [44]. The WGS data were analyzed by three groups using different approaches and emphasizing different regions or classes of genetic variation [14,17,18]. In our own study, we applied multiple callers to identify *de novo* **single-nucleotide variants (SNVs)**, **insertions/deletions (indels)**, and **structural variants (SVs)**. We reported an excess of smaller deletions that disrupted genes and nominally significant genome-wide enrichment for *de novo* variants in UTRs and in putative regulatory regions [TFBSs in central nervous system (CNS) DHSs] that are the most likely to function as enhancers and promoters [14]. Although the study was criticized for not considering all possible categories of noncoding functional DNA [17,45], DNM signals became more significant if the analysis was restricted to autism-related genes. Interestingly, the study also found that patients are more likely to carry multiple coding and noncoding DNMs in different genes (Figure 2) and such genes with

Table 1. Autism and Intellectual Disability WGS Cohorts

| Cohort | Family type | Projected total number of families | Projected total number of individuals | Actual available samples as of July 2018 | Sequencer | Website[a] | Refs |
|---|---|---|---|---|---|---|---|
| SSC[b] | Simplex | 2412[c] | 9198[c] | 4675 | Illumina HiSeq X Ten | https://www.sfari.org/ | [14[d],16[d],17[d],18[d],41[d],72] |
| | | | | | | http://ccdg.rutgers.edu/ | |
| MSSNG | Simplex and multiplex | 2756[e] | 7187[e] | 7231 | Illumina HiSeq 2000 and some Complete Genomics | https://www.mss.ng/ | [33,39,40[d]] |
| AGRE | Multiplex | 1010[c] | 4551[c] | 2308 | Illumina HiSeq X Ten | http://www.ihart.org/ | [42][d] |
| NIMH | MZ twin pairs | 10 | 40 | 40 | Illumina HiSeq | https://ndar.nih.gov/study.html?id=322 | [31] |
| Intellectual disability patient cohort | Trios | 50 | 150 | 150 | Complete Genomics | https://www.ebi.ac.uk/ega/studies/EGAS00001000769 | [32] |
| Total | | 6238 | 21 126 | 14 404 | | | |

[a] Website URL as of July 2018.
[b] Families also being sequenced as part of the CCDG.
[c] From website as of July 2018.
[d] Indicates that the study had a major focus on noncoding somewhere in the paper.
[e] From readme file (mssngresearcherreadme_20171020.pdf) from website (July 2018).



Trends in Neurosciences

Figure 2. The Pattern of Multiple *De Novo* Mutations (DNMs) in Autism Genomes. (A) The distribution compares the number of *de novo* variants of interest (VOIs) between affected individuals (red) and their siblings (blue) for 516 families from the Simons Simplex Collection (SSC). VOIs are defined as severe DNMs that are likely to disrupt protein function and DNMs in putative regulatory DNA [promoter, 5′ untranslated region (UTR), 3′ UTR, and DNase I hypersensitive sites (DHSs)]. Autism genomes tend to carry a greater number of such mutations, creating a skewed distribution when compared with the genomes of their unaffected siblings. Multiple mutations would be expected if the individual mutations were pathogenic (increased probability of probands carrying multiple events based on a Poisson model) or if multiple mutations were necessary to reach a liability threshold of disease. (B) Restricting the analysis to known autism-risk genes [Simons Foundation Autism Research Initiative (SFARI)] shows a significant excess of two or more events in probands (red) compared with siblings (blue). The trend is observed when partitioning coding and noncoding portions of the genome, emphasizing the importance of whole-genome sequencing (WGS). Adapted from [14].

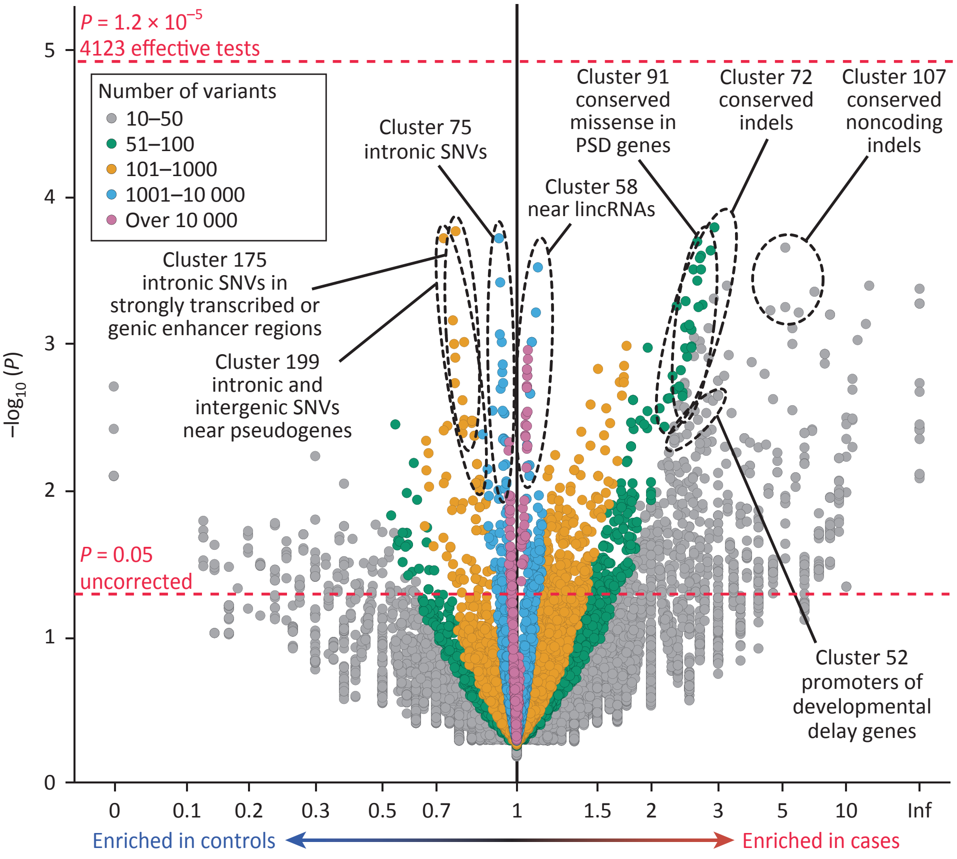
multiple hits are enriched for expression in striatal neurons. An excess of multiple DNMs in different noncoding DNA in patients would be consistent with this class of variation being pathogenic and/or support an **oligogenic** model of disease as has been suggested previously based on CNV studies [46,47].

The second study focused exclusively on SVs [18] (>100 bp in size) because of their greater likelihood of disrupting gene function and expression compared with SNVs. Although no difference was found for *de novo* SVs, the study did find a preferential transmission bias of *cis*-regulatory element SVs affecting promoters and the UTR of genes. They reported that these *cis*-regulatory element SVs were preferentially transmitted from fathers to their affected offspring in a study of 829 families [SSC and Relating Genes to Adolescent and Child Health (REACH)] that was subsequently replicated in a second study of 1771 autism families (MSSNG and SSC cohort). These findings contrast with previous reports that have observed a maternal transmission bias of private, putative truncating mutations in protein-coding sequence from mothers to their affected sons [11,48]. One possible explanation offered was that such paternal transmissions that affect regulatory mutations are less damaging than protein-coding mutations and that in the former cases multiple mutations (oligogenic or bilineal model) may be required to manifest as disease.

The final study to assess the SSC cohort [17] claimed a hypothesis-free approach where they assessed 51 801 annotation categories that reduced to 4123 correlated ones for the purpose of multiple testing correction. Unlike other approaches that focused on the most plausible functional NCEs (e.g., promoters, UTRs, enhancers), they considered many more categories, treating annotations such as long noncoding RNA (lncRNA) and pseudogenes as equivalent to promoters and enhancers. Dubbing their approach a category-wide association study (CWAS), they examined *de novo* and inherited SNVs and indels. They concluded that no category could achieve significance after multiple testing correction. Some interesting trends were noted, however, such as enrichment among autists of *de novo* SNVs and indels for promoters and UTRs, especially of developmental delay genes (Figure 3), confirming earlier studies. Predictably, less functional categories (pseudogenes) showed enrichment among unaffected siblings. Importantly, the authors' analytical framework established a statistical threshold on the order of $5 \times 10^{-6}$, estimating that >8000 families would be required to detect a genome-wide signal if all noncoding annotation categories are considered functionally equivalent.

## Targeted Sequencing of Autism and Other NDDs

Comparatively, there have been relatively few targeted studies focused on mutational burden in noncoding regulatory DNA, although one study [49] did examine available WES data in a small number of individuals with autism ($n = 48$) and their parents, reporting an excess of putative inherited regulatory variants in autism-risk genes, fetal development genes, and miRNA genes. In two recent studies, experiments were designed to target and sequence specific noncoding portions with the hypothesis that these NCEs would be more likely to exhibit a functional effect [50,51]. Doan *et al.* [50], for example, focused on human accelerated regions (HARs) – regions that have experienced a burst of mutation specifically in the human lineage and have been implicated in the regulation of genes important in human evolution, including neural genes. The authors found 6.5-times enrichment of rare *de novo* CNVs in HARs among individuals with autism compared with sibling-matched controls from the SSC cohort. Interestingly, in a consanguineous population, the authors reported a significant excess of rare, biallelic point mutations in these HARs, suggesting that compound heterozygotes could account for ~5% of individuals with autism among consanguineous families. In a second study, Short *et al.* [51] focused on 6139 putative regulatory elements corresponding to conserved noncoding elements as well as known enhancers from the VISTA browser and putative heart enhancers. Focusing on 6239 children with developmental delay who were negative for obvious pathogenic events by exome sequence, they reported nominal enrichment for DNMs in conserved NCEs. If they are restricted to those that are active in the fetal brain, the enrichment becomes significant ($P = 8.1 \times 10^{-4}$). They estimate that DNMs in this specific subset might contribute to 1.0–2.8% of 'exome-negative' patients.



Trends in Neurosciences

Figure 3. Category-wide Association Study for Noncoding Regulatory DNA. The volcano plot depicts the burden of *de novo* single-nucleotide variants (SNVs) and insertions/deletions (indels) from a genome-wide analysis of 519 autism and 519 unaffected siblings. It considers 13 704 annotation categories (points) and computes both case–control enrichments and significance, correcting for 4123 effective independent tests. No individual category survives Bonferroni correction (top horizontal red line) under this analytical framework, although biologically plausible categories are highlighted. PSD, postsynaptic density. Reproduced from [17].

## Concluding Remarks

The first genome-wide investigations into the role of noncoding regulatory mutation have highlighted both the potential and the challenges of this class of variation in helping to explain neurodevelopmental disease. Notwithstanding the fact that most work remains underpowered due to limited sample size, some common findings have begun to emerge from some of the early targeted sequencing and WGS studies. First, there is evidence for increased *de novo* or inherited disruptive mutation burden in putative regulatory regions in probands compared with controls [14,16,18,40,51]. Although evidence for the role of inherited rare variation is still emerging and statistically underpowered, a recent study of transmission in multiplex families based on WGS data provides additional support [42]. Second, these signals often become more significant when restricted to autism and NDD risk genes or to conserved regions active in the fetal brain (as determined by DNase I hypersensitivity) [14,16]. Third, the type of mutation is an important consideration, with larger and more disruptive mutations (e.g., SVs, CNVs) showing potentially larger effects [14,18,50]. Finally, early estimates suggest that such mutations account for a small fraction of patients (<5%), although these estimates are almost

## Outstanding Questions

What is the relative contribution of noncoding variants to NDDs?

What are the most important noncoding regions to assess for NDDs?

What percentage of pathogenic variation is missed by short-read WGS?

Are the effects of noncoding mutations less severe than those in protein-coding regions?

Are the genes affected by coding variants the same as those affected by noncoding variants? How can one assess the role of multiple rare variants (oligogenic) in contributing to disease outcome? What are the relative contributions of rare and common variants to autism risk?

What methods and assays could improve high-throughput functional testing of noncoding regulatory mutations?

How does the burden of mutation (both noncoding and coding) depend on the sex of the affected individual?

Are the same genes and biological pathways implicated by noncoding variants as in protein-coding-region variants?

What is the best clinical approach to apply WGS information when no 'known' coding event is identified?

What burden of proof is necessary to conclude pathogenicity for a noncoding variant? How will it fit into clinical standards (e.g., the American College of Medical Genetics and Genomics guidelines)?

Phenotypic data are often minimal or even lacking from population controls [e.g., the Genome Aggregation Database (http://gnomad.broadinstitute.org), the Trans-Omics for Precision Medicine Bravo Database (https://bravo.sph.umich.edu)]. In light of this, what are the best strategies for their use and should additional investment be made to create valid disease-specific controls?

certainly a lower bound in the absence of complete mutation ascertainment and a consideration of the potential additive effects of different classes of mutation [14]. Collectively, most of the available data point to a model where single and multiple disruptions of regulatory DNA contribute to NDD risk by leading to downregulation and misexpression of genes important in fetal brain development. Although these findings are tantalizing, important challenges remain, as discussed next.

## Improved Variant Discovery

With respect to regulatory effects, not all mutations are created equal; deletions, in particular, have been shown experimentally to be more disruptive than SNVs [52,53]. Most SNVs in unique regions of the genome are now readily detected using short-read sequencing platforms, but this is not the case for other forms of SV [54]. A recent comparison of genomes sequenced with both Illumina and long-read PacBio data showed that 50% of indels (10–49 bp in size), 51% of larger deletions (>50 bp), 83% of insertions (>50 bp), and nearly all inversions are not detected using short-read sequencing platforms [55]. Thus, there is the potential for large swaths of regulatory mutation to be missed even when genomes are sequenced deeply using short-read technologies. A major challenge will be to increase the detection sensitivity for these understudied classes of mutation, especially variable-number tandem repeat and **short tandem repeat (STR)** expansions (e.g., the *FMR1* CGG repeat), which already have a longstanding association with NDDs and neurodegenerative disorders. Notwithstanding these technological advances, it is likely that high-impact rare variants will be diagnosed in only a minority of cases. Another challenge will be the development of appropriate methods to integrate both rare coding and noncoding variants with the pattern of common variation associated with polygenic risk scores and environmental exposure. Such models are critical for understanding both phenotypic variability and an individual’s true risk of disease.

## Sample Size and Uniform Variant Calling

Most researchers agree that much larger sample sizes will be required to prove and replicate these early genome-wide observations. No specific loci or associated genes are even close to significance, although it is interesting that recurrent DNMs and damaging rare SVs have been identified among the regulatory DNA of known autism risk genes [42]. We estimate that large-scale efforts such as the Centers for Common Disease Genomics (CCDG), MSSNG, and SSC along with some of the first multiplex cohorts from the Autism Genetic Resource Exchange (AGRE) [42] will generate over 21 000 genomes from approximately 6200 autism and 50 intellectual disability families (Table 1) by the end of 2018 [see Outstanding Questions regarding additional genome data resources (e.g., Gnomad [56], Bravo)]. Caution should be exercised, however, in naively combining datasets or making comparisons with control genomes where different sequencing platforms, variant callers, or thresholds of coverage can affect sensitivity. For example, simply combining published *de novo* variant lists from the MSSNG and the SSC (available in denovo-db 1.6.1; 2204 autism probands and 521 unaffected siblings) would show, according to our estimates, significant enrichment of DNMs in CNS DHSs (10 526 proband variants vs 2637 unaffected variants; one-sided Fisher’s exact test $P$ value = $1.51 \times 10^{-38}$, odds ratio [OR] = 1.32). The enrichment increases if we restrict to variants in TFBSs in these regions (CNS DHS TFBSs) (739 in probands, 162 in unaffected siblings; one-sided Fisher’s exact test $P$ value = $8.51 \times 10^{-7}$, OR = 1.50). Although each of these tests would pass the category-wide significance threshold of $5 \times 10^{-6}$ proposed by Werling *et al.* [17], an examination of the data suggests that most of the signal originates from greater DNM variance and increased indel counts in the MSSNG dataset. By eliminating individuals sequenced by Complete Genomics technology and those with fewer than 50 DNMs, and by focusing on SNVs in regions of good mappability, the dataset drops to 1623 autism probands and 519 unaffected siblings. This reduction significantly reduces observed levels of significance, with CNS DHSs having 7024 proband variants and 2367 unaffected variants (one-sided Fisher's exact test $P$ value = 9.11 × $10^{-3}$, OR = 1.06) and CNS DHS TFBSs having 491 proband variants and 152 unaffected variants (one-sided Fisher's exact test $P$ value = 0.07, OR = 1.15). Ideally, cases and control WGS should be sequenced and processed identically to draw meaningful conclusions.

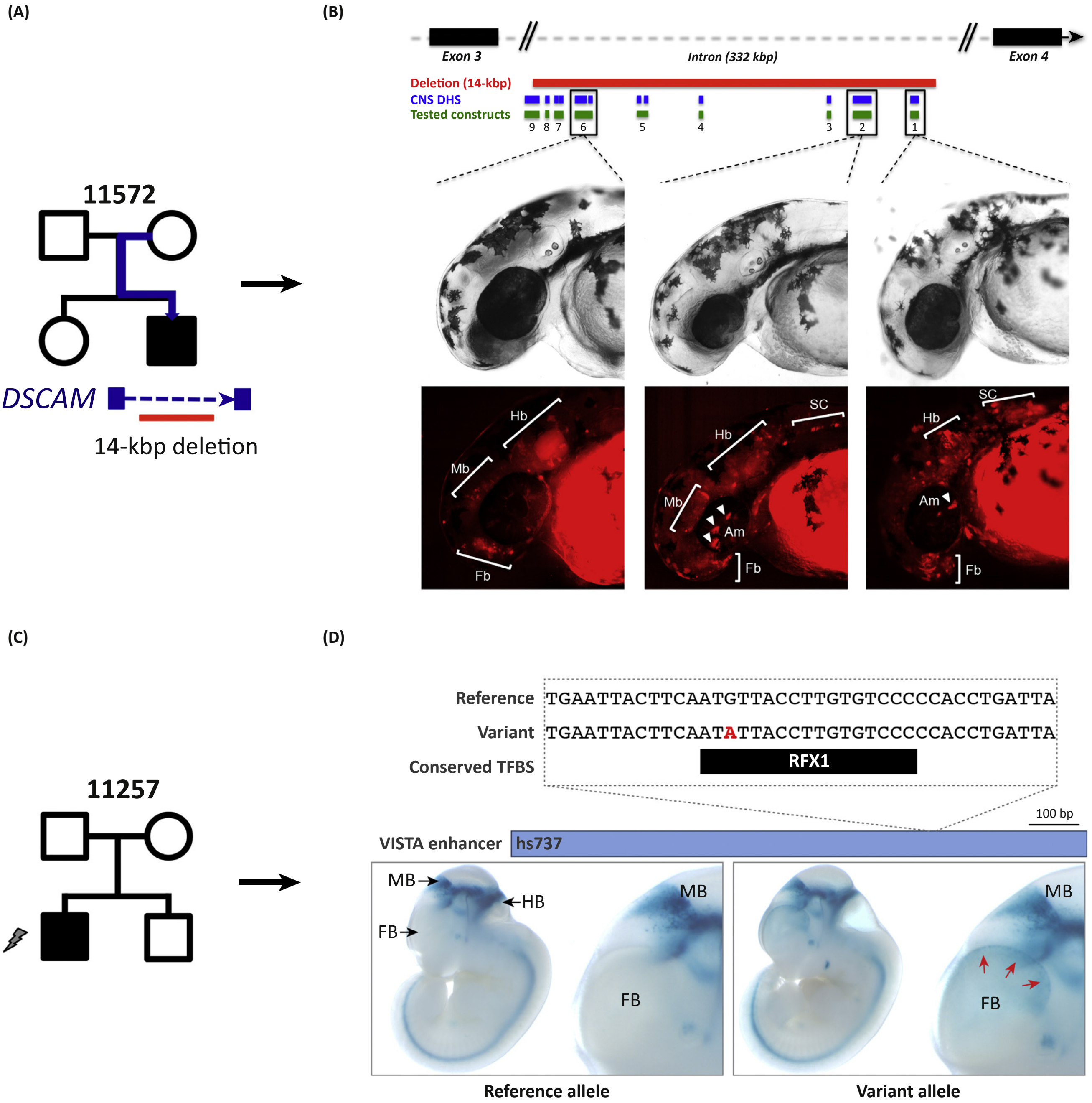
### Refinement of Functional Noncoding Regulatory DNA

Although some have argued that a proper statistical framework to assess the effect of NCE mutation should consider all possible annotation categories [17,45], an alternative approach would be to refine the subset to the regions of largest biological effect. The ENCODE [29] and Roadmap Epigenomics [30] projects, in this regard, were an important first step that served to enrich for functional elements in specific cell types and tissues. Currently, the field continues to refine these maps to even greater resolution, driving down to the single-cell level [57] with methods such as ATAC-seq [57,58] that will help to define regulatory regions in specific cell types of the developing fetal brain [59]. A focus on defining fetal brain enhancers during cortical development using ATAC-seq/Hi-C methods [60], including those that have been gained in the human lineage [61], will be particularly powerful in refining the noncoding search space to the most functionally relevant portions for neurodevelopment. Others have applied deep-**machine-learning** algorithms (reviewed elsewhere [62]) to better delineate and predict the potential effect of noncoding mutations. A recent preprint posted on *bioRxiv* [41], for example, applied such a deep-learning-based framework to 1790 autism families and showed that ASD probands harbor transcriptional and post-transcriptional regulation-disrupting mutations of significantly higher functional impact than unaffected siblings. A third approach involves expanding the known list of autism- and NDD-risk genes [12,13,63], especially those associated with haploinsufficiency, and then systematically characterizing all long-range (by Hi-C) and short-range regulatory DNA associated with those high-impact targets.

### High-Throughput Functional Assays

The ultimate litmus test for the relevance of specific NCE mutations to the etiology of NDDs is demonstrating that they have biological consequences. Historically, researchers have utilized relatively low-throughput assays (e.g., luciferase or transgenic models) to assay function. One version of the transgenic assays utilizes a Tol-transposon in zebrafish and can relatively quickly provide a visual readout of the enhancer activity of a DNA sequence [64]. This has been utilized to test the spatial and temporal location of enhancer activity driven by NCE in an intron of *DSCAM* (Figure 4). The other major transgenic approach is a gold-standard experiment in the mouse that assays the potential enhancer activity of a DNA sequence utilizing a lacZ reporter gene [65,66]. Although these methods are powerful in providing insight into the effect of specific variants on the spatial and temporal dynamics of enhancers (Figure 4), the scale of discovery of thousands of NCE mutations demands technological advances in throughput.

Over the past 5 years, researchers have developed methods to rapidly assay noncoding variants using **massively parallel reporter assays** that leverage high-throughput sequencing (e.g., MPRA [67], STARR-seq [68], STAP-seq [69]) to investigate variant effects on enhancers and promoters. While these assays are quantitative, such reporter construct assays suffer from relatively high false-positive rates and are not by themselves definitive. Additional methods such as **clustered regularly interspaced short palindromic repeats (CRISPR)**–Cas genome-editing technologies that systematically introduce mutations into their native regulatory context and measure their effects on expression or cellular/organismal phenotype are being envisioned at a massive scale (reviewed in [70]). It is likely that various levels of high-throughput functional



Trends in Neurosciences

Figure 4. Functional Assessment of Putative Regulatory Regions. (A) A 14-kbp *de novo* deletion of the intron of the autism-risk gene *DSCAM* deletes multiple putative regulatory elements [DNase I hypersensitive sites (DHSs)] in an autism patient [Simons Simplex Collection (SSC) family 11572]. (B) Independent testing of various elements in a zebrafish enhancer–reporter assay shows that nine elements mapping in the deleted region drive expression to different parts of the central nervous system (CNS) of the developing embryo. Shown are three of these elements. (C) Discovery of a *de novo* variant in an autism proband (SSC family 11257) mapping to a functionally assessed enhancer (VISTA) conserved between mouse and human. The mutation maps to a transcription factor binding site (TFBS) and a fetal brain DHS. (D) A mouse lacZ reporter assay shows that the single-base-pair mutation causes a gain of function where novel expression is identified in the forebrain in addition to the expected expression in the midbrain and hindbrain. Adapted from [14,16].

assay triage, followed by gold-standard transgenic assays, will need to be employed to systematically identify NCEs of the largest clinical (see Outstanding Questions; see also the ACMG guidelines [71]) and biological effect.

## Acknowledgments

We thank Jay Shendure, Daniela Witten, Jessica Chong, and members of the Eichler laboratory for helpful discussions regarding the contents of this review. We thank Bernie Devlin, Michael Talkowski, and Stephan Sanders for letting us reproduce a figure from their *Nature Genetics* paper in this review. We thank Bradley Coe and Mark Chaisson for early access to data included in this review. We also thank Tonia Brown for edits. The authors' research is supported, in part, by grants from the US National Institutes of Health (NIH R01MH101221) to E.E.E. and the National Institute of Mental Health to T.N.T. (1K99MH117165). We are grateful to all of the families who participated in research studies to understand the genetic risk factors underlying autism and NDDs. E.E.E. is an investigator of the Howard Hughes Medical Institute.

## Disclaimer Statement

E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc.

## References

1. Wang, K. *et al.* (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 459, 528–533
2. Weiss, L.A. *et al.* (2009) A genome-wide linkage and association scan reveals novel loci for autism. *Nature* 461, 802–808
3. Gaugler, T. *et al.* (2014) Most genetic risk for autism resides with common variation. *Nat. Genet.* 46, 881–885
4. Grove, J. *et al.* (2017) Common risk variants identified in autism spectrum disorder. *bioRxiv* Published online November 27, 2017. http://dx.doi.org/10.1101/224774
5. Cooper, G.M. *et al.* (2011) A copy number variation morbidity map of developmental delay. *Nat. Genet.* 43, 838–846
6. O'Roak, B.J. *et al.* (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat. Genet.* 43, 585–589
7. O'Roak, B.J. *et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* 485, 246–250
8. Iossifov, I. *et al.* (2014) The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* 515, 216–221
9. Iossifov, I. *et al.* (2012) *De novo* gene disruptions in children on the autistic spectrum. *Neuron* 74, 285–299
10. Sanders, S.J. *et al.* (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241
11. Krumm, N. *et al.* (2015) Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* 47, 582–588
12. Deciphering Developmental Disorders Study (2017) Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature* 542, 433–438
13. Coe B.P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and CNV morbidity. *Nat. Genet.* (in press)
14. Turner, T.N. *et al.* (2017) Genomic patterns of *de novo* mutation in simplex autism. *Cell* 171, 710–722.e12
15. Sandin, S. *et al.* (2017) The heritability of autism spectrum disorder. *JAMA* 318, 1182–1184
16. Turner, T.N. *et al.* (2016) Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* 98, 58–74
17. Werling, D.M. *et al.* (2018) An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* 50, 727–736
18. Brandler, W.M. *et al.* (2018) Paternally inherited *cis*-regulatory structural variants are associated with autism. *Science* 360, 327–331
19. Lettice, L.A. *et al.* (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* 12, 1725–1735
20. de Kok, Y.J. *et al.* (1995) A duplication/paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the POU3F4 gene. *Hum. Mol. Genet.* 4, 2145–2150
21. Grant, S.F. *et al.* (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* 38, 320–323
22. Sladek, R. *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445, 881–885
23. Scott, L.J. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316, 1341–1345
24. Emison, E.S. *et al.* (2010) Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. *Am. J. Hum. Genet.* 87, 60–74
25. Chatterjee, S. *et al.* (2016) Enhancer variants synergistically drive dysfunction of a gene regulatory network in Hirschsprung disease. *Cell* 167, 355–368.e10
26. Fu, Y.H. *et al.* (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* 67, 1047–1058
27. Pieretti, M. *et al.* (1991) Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* 66, 817–822
28. Turner, T.N. *et al.* (2016) denovo-db: a compendium of human *de novo* variants. *Nucleic Acids Res.* 45, D804–D811
29. ENCODE (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640
30. Kundaje, A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330
31. Michaelson, J.J. *et al.* (2012) Whole-genome sequencing in autism identifies hot spots for *de novo* germline mutation. *Cell* 151, 1431–1442
32. Gilissen, C. *et al.* (2014) Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511, 344–347
33. Yuen, R.K. *et al.* (2015) Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* 21, 185–191
34. Guo, H. *et al.* Genome sequencing identifies multiple deleterious variants in autism patients with more severe phenotypes. *Genet. Med.* Published online December 3, 2018. https://doi.org/10.1038/s41436-018-0380-2
35. Willsey, A.J. *et al.* (2013) Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* 155, 997–1007
36. Parikshak, N.N. *et al.* (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155, 1008–1021

37. Hormozdiari, F. *et al.* (2015) The discovery of integrated gene networks for autism and related disorders. *Genome Res.* 25, 142–154

38. Sanders, S.J. (2015) First glimpses of the neurobiology of autism spectrum disorder. *Curr. Opin. Genet. Dev.* 33, 80–92

39. Yuen, R.K. *et al.* (2017) Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* 20, 602–611

40. Yuen, R.K. *et al.* (2016) Genome-wide characteristics of *de novo* mutations in autism. *NPJ Genom. Med.* 1, 16027

41. Zhou, J. *et al.* (2018) Whole-genome deep learning analysis reveals causal role of noncoding mutations in autism. *bioRxiv* Published online May 11, 2018. http://dx.doi.org/10.1101/319681

42. Ruzzo, E.K. *et al.* (2018) Whole genome sequencing in multiplex families reveals novel inherited and *de novo* genetic risk in autism. *bioRxiv* Published online June 6, 2018. http://dx.doi.org/10.1101/338855

43. Genome of the Netherlands Consortium (2014) Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* 46, 818–825

44. Fischbach, G.D. and Lord, C. (2010) The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* 68, 192–195

45. Wray, N.R. and Gratten, J. (2018) Sizing up whole-genome sequencing studies of common diseases. *Nat. Genet.* 50, 635–637

46. Girirajan, S. *et al.* (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* 42, 203–209

47. Duyzend, M.H. *et al.* (2016) Maternal modifiers and parent-of-origin bias of the autism-associated 16p11.2 CNV. *Am. J. Hum. Genet.* 98, 45–57

48. Iossifov, I. *et al.* (2015) Low load for disruptive mutations in autism genes and their biased transmission. *Proc. Natl. Acad. Sci. U. S. A.* 112, E5600–E5607

49. Williams, S.M. *et al.* (2018) An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder. *Mol. Psychiatry* Published online April 27, 2018. http://dx.doi.org/10.1038/s41380-018-0049-x

50. Doan, R.N. *et al.* (2016) Mutations in human accelerated regions disrupt cognition and social behavior. *Cell* 167, 341–354.e12

51. Short, P.J. *et al.* (2018) *De novo* mutations in regulatory elements in neurodevelopmental disorders. *Nature* 555, 611–616

52. Osterwalder, M. *et al.* (2018) Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* 554, 239–243

53. Dickel, D.E. *et al.* (2018) Ultraconserved enhancers are required for normal development. *Cell* 172, 491–499.e15

54. Huddleston, J. and Eichler, E.E. (2016) An incomplete understanding of human genetic variation. *Genetics* 202, 1251–1254

55. Chaisson, M.J.P. *et al.* (2017) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv*

56. Lek, M. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291

57. Cusanovich, D.A. *et al.* (2018) A single-cell atlas of *in vivo* mammalian chromatin accessibility. *Cell* 1309–1324.e18

58. Cusanovich, D.A. *et al.* (2015) Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914

59. Nowakowski, T.J. *et al.* (2017) Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* 358, 1318–1323

60. de la Torre-Ubieta, L. *et al.* (2018) The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell* 172, 289–304.e18

61. Reilly, S.K. *et al.* (2015) Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* 347, 1155–1159

62. Khurana, E. *et al.* (2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* 17, 93–108

63. Stessman, H.A. *et al.* (2017) Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* 49, 515–526

64. Fisher, S. *et al.* (2006) Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nat. Protoc.* 1, 1297–1305

65. Pennacchio, L.A. *et al.* (2006) *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* 444, 499–502

66. Visel, A. *et al.* (2007) VISTA Enhancer Browser – a database of tissue-specific human enhancers. *Nucleic Acids Res.* 35, D88–D92

67. Melnikov, A. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30, 271–277

68. Arnold, C.D. *et al.* (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077

69. Arnold, C.D. *et al.* (2017) Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat. Biotechnol.* 35, 136–144

70. Montalbano, A. *et al.* (2017) High-throughput approaches to pinpoint function within the noncoding genome. *Mol. Cell* 68, 44–59

71. Richards, S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424

72. Brandler, W.M. *et al.* (2016) Frequency and complexity of *de novo* structural mutation in autism. *Am. J. Hum. Genet.* 98, 667–679