Prioritization of neurodevelopmental disease genes by discovery of new mutations

Alexander Hoischen[1], Niklas Krumm[2] & Evan E Eichler[2,3]

Advances in genome sequencing technologies have begun to revolutionize neurogenetics, allowing the full spectrum of genetic variation to be better understood in relation to disease. Exome sequencing of hundreds to thousands of samples from patients with autism spectrum disorder, intellectual disability, epilepsy and schizophrenia provides strong evidence of the importance of *de novo* and gene-disruptive events. There are now several hundred new candidate genes and targeted resequencing technologies that allow screening of dozens of genes in tens of thousands of individuals with high specificity and sensitivity. The decision of which genes to pursue depends on many factors, including recurrence, previous evidence of overlap with pathogenic copy number variants, the position of the mutation in the protein, the mutational burden among healthy individuals and membership of the candidate gene in disease-implicated protein networks. We discuss these emerging criteria for gene prioritization and the potential impact on the field of neuroscience.

Recent exome (and genome) sequencing studies of families have aimed to comprehensively discover genetic variation to identify the most likely causal mutation in patients with disease. Sequencing studies of parent-proband trios for probands with intellectual disability (ID)[1,2], autism spectrum disorder (ASD)[3–7], schizophrenia (SCZ)[8–10] and epilepsy[11] have all suggested that *de novo* point mutations are important in pediatric and adult disorders of brain development (**Table 1**). The relative contribution of *de novo* mutations to each disorder remains to be determined but appears to correlate well with the degree of reduced fitness or fecundity of the given condition[12]. However, not only *de novo* events but also rare inherited copy number variants (CNVs) can have an effect on fecundity, though their overall effect on fecundity is still debated[13]. Biologically, 75–80% of *de novo* point mutations arise paternally[3,14], likely as a result of the greater number of cell divisions in the male germline lineage than in the female lineage. These findings are consistent with some epidemiological data that find advancing paternal age to be a significant predictor of ASD, ID and SCZ[15–17] and argue for the need to properly control for paternal age when comparing mutation rates between probands and siblings. The importance of *de novo* and private rare mutations is especially important clinically, as there are now reports of diagnostic yields ranging from 10–55% for select (usually the most severe) groups of patients with ID[1,2] and epilepsy[18], in addition to resolution of unsolved Mendelian disorders[19]. It is clear that next-generation sequencing approaches have provided powerful tools for identifying genes harboring potentially pathogenic mutations. Deciding which genes to pursue, however, is not always self-evident because follow-up research and diagnostic studies are critical to understanding the full contribution of a particular mutation to its respective phenotype.

In this Review, we will discuss the prioritization of candidate genes identified through sequencing studies, show emerging trends and highlight potential strategies for subsequent functional characterization of these neurodevelopmental genes. We focus on lessons learned from 11 recent studies that report 2,368 *de novo* mutations from a total of 2,358 probands and 600 *de novo* mutations from 731 controls (**Table 1**). The bulk of the data originate from sequencing studies of parents and probands with ASD, ID and epileptic encephalopathies, but more recent studies have also highlighted the importance of *de novo* mutations in SCZ. There is evidence that *de novo* mutations, particularly disruptive mutations, occur in the same genes despite the nosological distinction for these different diseases. For the purpose of this Review, we collectively term these diseases 'neurodevelopmental disorders' but recognize that some, especially adult-onset diseases such as SCZ, may have etiologic components that are not neurodevelopmental in origin.

Recurrently mutated genes

One of the frequently used concepts in considering possible 'new disease genes' responsible for a given neurodevelopmental phenotype is the recurrence of *de novo* mutations in the same gene, along with the absence of such mutations in healthy controls. This rule follows the precedent established for the discovery of pathogenic *de novo* CNVs during the last decade, with the highest priority given to recurrent mutations that lead to a complete loss of function of one of the parental copies of the gene. Up to ten independent reports of *de novo* mutations in *SCN2A* and nine independent reports of *de novo* mutations in *SCN1A* and *STXBP1* have been described (**Tables 2** and **3**). Strikingly, *de novo* mutations in those genes have so far been found exclusively in probands and never in controls. Simulation data suggest that at least two but certainly three or more recurrent *de novo* loss-of-function (LoF) events (that is, predicted nonsense, frameshift or canonical splice site mutations)

[1]Department of Human Genetics, Radboud Institute for Medical Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands. [2]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [3]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu) or A.H. (alexander.hoischen@radboudumc.nl).

Received 27 February; accepted 26 March; published online 27 May 2014; doi:10.1038/nn.3703

Table 1 Summary of 11 major (exome) sequencing studies

				Number of coding and splice site *de novo* point mutations in probands						Number of coding and splice site *de novo* point mutations in controls or siblings					
Study	Disorder	Number of probands	Number of controls (c) or siblings (s)	Total	LoF	Codon indels (in-frame)	Mis-sense	Syn-onymous	Nonsyn-onymous/ synonymous ratio	Total	LoF	Codon indels (in-frame)	Mis-sense	Syn-onymous	Nonsyn-onymous/ synonymous ratio
Rauch *et al.*[2]	ID	51	20 (c)	91	21	0	58	12	6.6 (79/12)	27	3	0	17	7	2.9 (20/7)
De Ligt *et al.*[1]	ID	100	0	79	15	0	48	16	3.9 (63/16)	NA	NA	NA	NA	NA	NA
O'Roak *et al.*[3]	ASD	209	50 (s)	260	38	0	154	68	2.8 (192/68)	50	3	0	31	16	2.1 (34/16)
Sanders *et al.*[4,a]	ASD	225	200 (s)	172	15	3	125	29	4.9 (143/29)	125	5	0	82	38	2.3 (87/38)
Neale *et al.*[7]	ASD	175	0	169	18	0	101	50	2.4 (119/50)	NA	NA	NA	NA	NA	NA
Iossifov *et al.*[5,b]	ASD	343	343 (s)	362	61	7	209	85	3.3 (277/85)	314	30	8	203	73	3.3 (241/73)
Jiang *et al.*[6]	ASD	32	0	42	5	0	28	9	3.7 (33/9)	NA	NA	NA	NA	NA	NA
Allen *et al.*[11]	EE	264	0	289	36	1	196	56	4.2 (233/56)	NA	NA	NA	NA	NA	NA
Gulsuner *et al.*[8]	SCZ	105	84 (c)	100	12	0	57	31	2.2 (69/31)	67	11	0	37	19	2.5 (48/19)
Xu *et al.*[9,c]	SCZ	231	34	164	20	4	115	25	5.6 (139/25)	17	1	0	11	5	2.4 (12/5)
Fromer *et al.*[10]	SCZ	623	0	640	65	9	410	156	3.1 (484/156)	NA	NA	NA	NA	NA	NA
Sums		2,358	731	2,368	306 (12.9%)	24 (1.0%)	1,501 (63.4%)	537 (22.7%)	3.4 (1,831/537)	600	53 (8.8%)	8 (1.3%)	381 (63.5%)	158 (26.3%)	2.8 (442/158)

Summary of *de novo* mutation discovery, including size of study, number of *de novo* mutations and severity for each group. EE, epileptic encephalopathies; ASD, autism spectrum disorder; ID, intellectual disability; SCZ, schizophrenia; NA, not applicable. Putative LoF includes nonsense, frameshift and canonical splice site mutations as based on gene annotation. For some studies the numbers deviate from the numbers given in the main publication; numbers considered here were retrieved from *de novo* mutation overviews from supplementary tables and reannotated using Seattle-Seq. When considering all studies together, the total number of *de novo* mutations identified in probands versus controls and siblings is different (Fisher's exact test, $P = 0.0012$); this may, however, instead reflect the technical differences among the studies. The most significant difference is seen for the number of LoF *de novo* mutations, which is significantly higher in probands than in controls and siblings (Fisher's exact test, $P = 0.0062$). Similarly, the total number of nonsynonymous versus synonymous variants is higher in probands than in controls and siblings; however, this does not reach significance (Fisher's exact test, $P = 0.066$).

[a]Study did not consider all indels for case-control comparisons. [b]Not all *de novo* mutations were validated by Sanger sequencing. [c]Study excluded splice site variants if not in canonical dinucleotide of splice site.

are unlikely to occur by chance, making such genes outstanding candidates (**Table 2**)[7].

The frequency of recurrence is dependent on the extent of locus heterogeneity associated with each disease. Leveraging observed recurrences of *de novo* mutations (**Fig. 1**), researchers believe that diseases such simplex autism and SCZ arise from mutations in >500 genes, while studies of severe ID and epileptic encephalopathies (infantile spasms and Lennox–Gastaut subtypes) suggest lower heterogeneity. Such estimates should be considered only rough approximations at this point because they are highly dependent upon the fraction of *de novo* mutations that are in fact pathogenic, as well as ascertainment biases in sample collection (**Fig. 1**). In this regard, several of the candidate genes with highest numbers of recurrent mutations have been observed in patients with epilepsy[11], ID[1,2] and ASD[3–7]. This may not be surprising in light of the comorbidity of these diseases and if one accepts that heterogeneity is lower for epilepsy-related disorders than for ASD, SCZ or ID. In such a scenario, sequencing of even a modest number of epilepsy cases delivers recurrent mutations more frequently than more broadly defined developmental delay or ASD[11,18]. One study of SCZ, for example, highlighted four recurrently mutated genes, but perhaps more remarkably, the same study identified overlapping genes with those implicated in ASD when focusing on prenatally expressed genes[9]. The stronger overlap between ASD, ID and epilepsy and yet limited overlap with SCZ (**Table 2**) could be largely because only a subset of the latter disease stems from a neurodevelopmental origin.

Perhaps the most striking examples are recurrent identical *de novo* mutations in the same gene. Across the various studies, such identical recurrences have already been observed for six genes (*ALG13*, *KCNQ3*, *SCN1A*, *CUX2*, *DUSP15* and *SCN2A*; **Table 4**). Such events are exceedingly unlikely, with estimates of identical recurrences in *ALG13* and *SCN2A* calculated at $P = 7.77 \times 10^{-12}$ and $P = 1.14 \times 10^{-9}$, respectively, in the case of epilepsy[11]. Most of these estimates of significance, however, assume a random mutation process. Yet mutational hotspots certainly exist, and recurrence of the same mutation cannot be taken as proof positive of an association.

The clinical significance of most *de novo* mutations discovered in patients remains unclear. For most genes, only a single *de novo* mutation has been identified. Nevertheless, on the basis of the observation that *de novo* LoF mutations occur two to three times more frequently in ASD probands than in unaffected siblings, it is now estimated that a large fraction of these singletons will be relevant to disease etiology. When considering all studies in aggregate, *de novo* LoF mutations are observed significantly more in cases than in controls (**Table 1**; Fisher's exact test, $P = 0.0062$). The interpretation of recurrent missense mutations, however, represents a greater challenge. Sanders *et al.*[4] estimated that four missense *de novo* mutations in the same genes would be required in simplex autism to exceed a chance finding; this was based on a cohort size of up to 2,000 with an estimated locus heterogeneity of 1,000 ASD risk loci. Given the extreme locus heterogeneity of diseases such as ASD and ID, other strategies have been adopted to prioritize likely causal genes. High-throughput targeted

Table 2 Recurrent and overlapping genes: *de novo* mutations in same genes observed among ID, ASD, EE and SCZ

Gene	Total observations	Mutation type		Neurodevelopmental disorder			
		LoF*	Missense	ASD	ID	EE	SCZ
SCN2A	11	7	4	4	4	2	1
SCN1A	9	4	5	1	0	8	0
STXBP1	9	2	7	1	3	5	0
GABRB3	5	0	5	1	0	4	0
TRIO	5	0	5	2	2	1	0
POGZ	4	3	1	2	0	0	2
MYH9	4	1	3	1	1	0	2
SYNGAP1	4	4	0	0	3	0	1

Genes reported with recurrent (≥4) nonsynonymous *de novo* mutations identified in 11 studies on four different neurodevelopmental phenotypes. EE, epileptic encephalopathies; ASD, autism spectrum disorder; ID, intellectual disability; SCZ, schizophrenia. *TTN* and *MUC5B* were excluded from this table owing to high variant load in controls and unlikely involvement in the phenotypes discussed. *Includes nonsense, frameshift and splice site mutations.

multiplexed resequencing technologies, such as molecular inversion probes, have been employed to screen ~50 candidate genes in thousands of patients and controls[3,18]. Such approaches are scalable, inexpensive (less than $1 per gene per sample), sensitive and specific, increasing by an order of magnitude the number of patients that can be screened. The strategy was particularly useful in discovering a burden of *de novo* LoF mutation of *CHD8* associated with ASD[20]. The relatively ease in detecting *de novo* mutations allows rapid identification of potential candidate genes; the number of cases that are required to make these findings statistically significant (**Fig. 1**) can be lower for the genes that are mutated exclusively in a large number of patients when compared to standard case-control studies[21].

Previous evidence of overlap with pathogenic CNVs

Another strategy has been to compare patterns of CNVs in patient and control populations to prioritize genes (**Fig. 2**)[22]. Extensive CNV morbidity maps have been developed for tens of thousands of children with autism, ID and epilepsy, helping to define pathogenic regions of dosage imbalance in the human genome[23–26]. Overlapping deletions in such collections occasionally refine the smallest region of overlap, highlighting a modest number of candidate genes. Recurrent *de novo* point mutations in a gene within such a region with CNV burden substantially increases the likelihood that LoF of the gene is responsible for a phenotype. O'Roak *et al.*[3] and Rauch *et al.*[2] each discovered, for example, LoF point mutations for *SETBP1*—a gene where a significant enrichment for deletion CNVs has been seen in patients with overlapping neurodevelopmental phenotypes but not in controls. Similar patterns have recently been observed for *DYRK1A* and *MBD5* (**Fig. 2**), including reports of balanced but gene-disrupting chromosomal translocations[27]. Such information has been used to compute haploinsufficiency scores[2,28] to strengthen the case for causality of *de novo* LoF mutations (**Fig. 2**).

Position of the mutation in the protein

More than 60% of *de novo* mutations discovered from exome sequencing projects are missense mutations (**Table 1**). Distinguishing pathogenic signal from the background of benign mutations is an active area of research. For genes for which previous CNVs or LoF mutations were described, 'severe' missense mutations are also likely to result in a dosage effect. However, there are examples, usually from clinically well-defined neurodevelopmental syndromes, showing that missense mutations result in different outcomes on the basis either of the protein domain they affect, the position in the gene (for example, the N or C terminus of the resulting protein)[29] or their potential to modify the normal function of the protein. Examples of this last include gain-of-function (GoF) mutations and LoF mutations in the same gene that result in different phenotypic outcomes[30,31]. *SETBP1* GoF in a degron sequence (ubiquitination motif) results in the rare but well-defined Schinzel-Giedion syndrome[32], while deletions or LoF mutations may result in a distinct and milder phenotype comprising ASD or ID with speech delay and other features[2,3,33,34]. Other examples include different phenotypic effects dependent on the location of the mutation: early truncating and missense mutations in *NOTCH2* are known to cause Alagille syndrome[35], while truncating events restricted to the last exon escape nonsense-mediated decay and result in Hajdu-Cheney syndrome[36,37].

Mutational burden among healthy individuals

One approach to prioritizing missense mutations leverages evolutionary conservation by assigning 'mutability scores' per gene or even at the base-pair level. O'Roak *et al.*[3], for example, established an evolutionary mutation score per human gene based on the human-chimpanzee divergence and the size of a gene. Similarly, the Epi4k Consortium used a gene-specific mutation rate based on a per-base score[38]; this score was, however, not based on human-chimpanzee evolution but made use of human-specific polymorphism data. A recent publication used rare variant data from healthy individuals and offers a new integrative annotation tool for noncoding variants[39]. The wealth of available control exome sequence data can also be used to estimate the (rare) variant load per gene (and distribution). For example, the analysis of data generated from sequencing 6,500 'control' exomes as part of the ESP6500 data release has been used to define the load of LoF mutations per gene[2] and to prioritize >4,000 human genes that are most intolerant of variation[11]; details on the respective samples can be obtained from the Exome Variant Server (http://evs.gs.washington.edu/EVS/). Another approach uses

Table 3 Details of recurrent *de novo* mutations in *SCN2A* identified in seven studies with three different neurodevelopmental phenotypes

Gene	Coding effect	Mutation (genomic DNA level)	Mutation (cDNA level)	Mutation (protein level)	Study	Disorder
SCN2A	Frameshift	Chr2(GRCh37):g.166170553_166170584del	NM_021007.2:c.1318_1349del	p.Glu440Argfs*20	Jiang *et al.*[6]	ASD
SCN2A	Frameshift	Chr2(GRCh37):g.166172105dup	NM_021007.2:c.1508dup	p.Asn503Lysfs*19	Rauch *et al.*[2]	ID
SCN2A	Frameshift	Chr2(GRCh37):g.166179825_166179826del	NM_021007.2:c.1831_1832del	p.Leu611Valfs*35	Rauch *et al.*[2]	ID
SCN2A	Missense	Chr2(GRCh37):g.166198975G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al.*[11]	EE
SCN2A	Missense	Chr2(GRCh37):g.166198975G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al.*[11]	EE
SCN2A	Missense	Chr2(GRCh37):g.166201311C>T	NM_021007.2:c.2809C>T	p.Arg937Cys	Rauch *et al.*[2]	ID
SCN2A	Nonsense	Chr2(GRCh37):g.166201379C>A	NM_021007.2:c.2877C>A	p.Cys959*	Sanders *et al.*[4]	ASD
SCN2A	Nonsense	Chr2(GRCh37):g.166210819G>T	NM_021007.2:c.3037G>T	p.Gly1013*	Sanders *et al.*[4]	ASD
SCN2A	Nonsense	Chr2(GRCh37):g.166231415G>A	NM_021007.2:c.4193G>A	p.Trp1398*	de Ligt *et al.*[1]	ID
SCN2A	Missense	Chr2(GRCh37):g.166234111C>T	NM_021007.2:c.4259C>T	p.Thr1420Met	Iossifov *et al.*[5]	ASD
SCN2A	Splice site	Chr2(GRCh37):g.166187838A>G	NM_001040142.1:c.2150–2A>G	p.?	Fromer *et al.*[10]	SCZ

EE, epileptic encephalopathies; ASD, autism spectrum disorder; ID, intellectual disability; SCZ, schizophrenia.

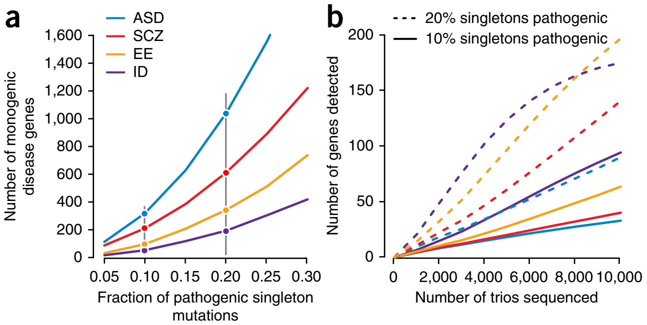


Figure 1 Genes with recurrent *de novo* mutations in four neurodevelopmental disorders. (**a**) We estimate the number of fully penetrant genes that can explain disease once mutated, based on a *de novo* model using the 'unseen species problem'. We consider all recurrent missense or LoF *de novo* mutations pathogenic, as well as a defined fraction of mutations in genes observed just once (because it is unlikely that all *de novo* mutations are pathogenic). The ratio between genes mutated recurrently and the rate of singleton mutations suggests an estimate for the true number of genes pathogenic when mutated. Including more singleton mutations increases the fraction of each disorder explained by single *de novo* SNVs at the cost of including more genes as pathogenic. Initial exome sequencing studies of epilepsy and ID focused on specific pediatric subtypes or the most severe cases; thus, the number of generalized epilepsy- or ID-associated genes is likely to be much higher. EE, epileptic encephalopathies; ASD, autism spectrum disorder; ID, intellectual disability; SCZ, schizophrenia. (**b**) Expected hit rate (or sensitivity) of true positive genes discovered using trio sequencing studies (under a family-wise error rate of 5%; that is, each gene passes exome-wide significance of 2.6 × 10^{-6}). We estimate the power of trio sequencing to detect statistically significant associations for disease-associated genes, under the assumption that 10% or 20% of singleton mutations could be fully penetrant (vertical bars in **a**). We assume the distribution of these genes is uniform within each disorder and that they do not differ significantly from all genes in terms of length and mutability, although these are taken into account when determining significance.

random mutation modeling[40] to calculate the likelihood that observed (*de novo*) mutations have a damaging effect. Similar prioritizations are provided by tools that score individual mutation severity (SIFT, PolyPhen2, MutationTaster, MutPred, CONDEL, etc.), some of which can be adapted to a gene-based prioritization score from genome-wide data[41]. These population data provide a powerful unbiased approach to home in on genes that are likely to be among the most penetrant because of the complete absence of disruptive variation in the general population (for example, *CHD8* or *DYRK1A*). A critical aspect of such analyses is the reliability of a particular gene model. Most human genes show evidence of alternative splice forms, many of which have no known function. Apparent hotspots of mutation for a particular exon (often exon-intron boundaries) in both cases and controls may suggest misannotation, the presence of a processed pseudogene or an alternative, nonfunctional splice form.

Pathway enrichment and links to cancer biology

Another popular approach to discern the most important gene candidates for further disease association and characterization has been to identify specific biological networks of genes enriched in cases as compared to controls. Although this approach cannot be used unequivocally to define causality, membership of a specific gene in a particular protein-protein interaction (PPI) or coexpression network may increase the likelihood of its association with disease. Numerous studies have reported significant enrichment of both *de novo* CNV and single-nucleotide variant (SNV) mutations in particular pathways[3,4,42,43]. O'Roak *et al.*[3], for example, reported a significant enrichment of *de novo* disruptive autism mutations among proteins associated with chromatin remodeling and β-catenin and WNT signaling—a finding that was replicated in a follow-up resequencing study of more than 2,400 probands. One recent instance, in which membership of a new candidate gene in a PPI network led to the discovery of an autism-associated gene, is *ADNP*. A single *ADNP* LoF mutation was initially observed in exome sequencing studies. Although the observed mutation frequency in this gene did not reach statistical significance when cases and controls were compared[20], it was strongly implicated in the PPI network originally defined by O'Roak *et al.*[3] Targeted resequencing experiments combined with clinical exome sequencing identified several more cases with *de novo* mutations and remarkably similar phenotypes representing a new SWI-SNF–related autism syndrome (**Fig. 3**)[44]. Notably, many of the genes implicated in the β-catenin pathway have also been described as mutated in patients with ID[1] but not in patients with SCZ. Similarly, an enrichment of genes interacting with *FMR1* (also known as *FMRP*)—the gene responsible for fragile X syndrome—has been reported with *de novo* mutations in ASD[5], epilepsy[11] and, most recently, SCZ[10,45]. Whether this observation is due to the relative high incidence of cases that also presented with comorbid ID remains to be determined.

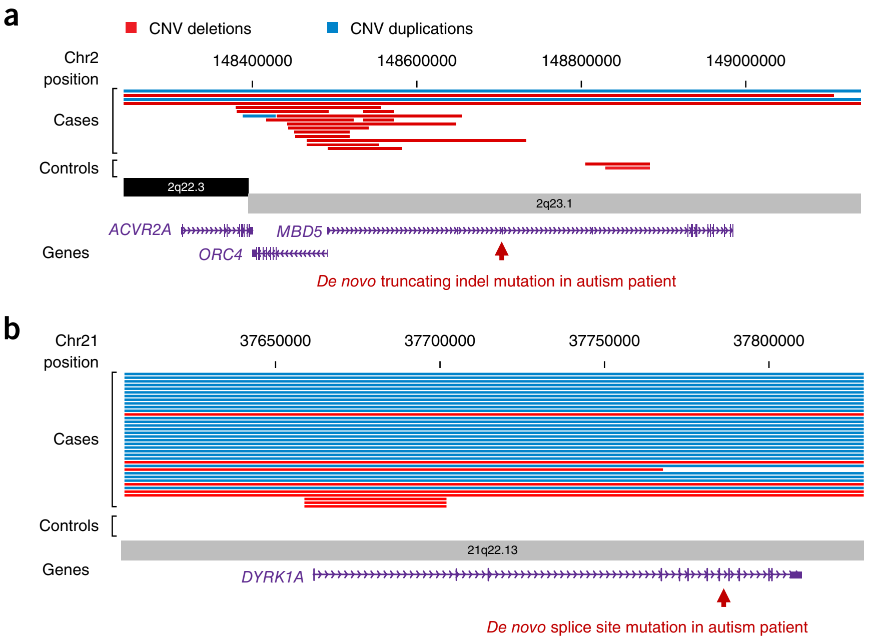
In addition to PPI networks, studies of coexpression have shown enrichment for specific spatio-temporal patterns of expression. A study of coexpressed genes affected by *de novo* mutations reported an enrichment in fetal prefrontal cortical network in SCZ[8], which is in line with the finding by Xu *et al.*[9] that genes with higher expression

Table 4 **Recurrent identical *de novo* mutations in 6 genes identified in 11 exome studies with different neurodevelopmental phenotypes**

Gene	Coding effect	Mutation (genomic DNA level)	Mutation (cDNA level)	Mutation (protein level)	Study	Disorder
ALG13	Missense	ChrX(GRCh37):g.110928268A>G	NM_001099922.2:c.320A>G	p.Asn107Ser	de Ligt *et al.*[1]	ID
ALG13	Missense	ChrX(GRCh37):g.110928268A>G	NM_001099922.2:c.320A>G	p.Asn107Ser	Allen *et al.*[11]	EE
ALG13	Missense	ChrX(GRCh37):g.110928268A>G	NM_001099922.2:c.320A>G	p.Asn107Ser	Allen *et al.*[11]	EE
KCNQ3	Missense	Chr8(GRCh37):g.133192493G>A	NM_001204824.1:c.328C>T	p.Arg110Cys	Rauch *et al.*[2]	ID
KCNQ3	Missense	Chr8(GRCh37):g.133192493G>A	NM_001204824.1:c.328C>T	p.Arg110Cys	Allen *et al.*[11]	EE
SCN1A	Splice donor	LRG_8:g.24003G>A	NM_006920.4:c.602+1G>A	p.?	Allen *et al.*[11]	EE
SCN1A	Splice donor	LRG_8:g.24003G>A	NM_006920.4:c.602+1G>A	p.?	Allen *et al.*[11]	EE
CUX2	Missense	Chr12(GRCh37):g.111748354G>A	NM_015267.3:c.1768G>A	p.Glu590Lys	Rauch *et al.*[2]	ID
CUX2	Missense	Chr12(GRCh37):g.111748354G>A	NM_015267.3:c.1768G>A	p.Glu590Lys	Allen *et al.*[11]	EE
SCN2A	Missense	Chr2(GRCh37):g.166198975G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al.*[11]	EE
SCN2A	Missense	Chr2(GRCh37):g.166198975G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al.*[11]	EE
DUSP15	Missense	Chr20(GRCh37):g.30450489G>A	NM_080611.2:c.320C>T	p.Thr107Met	Neale *et al.*[7]	ASD
DUSP15	Missense	Chr20(GRCh37):g.30450489G>A	NM_080611.2:c.320C>T	p.Thr107Met	Fromer *et al.*[10]	SCZ

EE, epileptic encephalopathies; ASD, autism spectrum disorder; ID, intellectual disability; SCZ, schizophrenia.

Figure 2 CNV and exome intersections define candidate genes. (**a**,**b**) Deletion (red) and duplication (blue) burden for developmental delay or ID cases and controls for two genes, *DYRK1A* (**a**) and *MBD5* (**b**), as compared to sporadic LoF mutations on the basis of exome sequencing of 209 autism simplex trios. *DYRK1A* is a strong candidate gene for cognitive deficits associated with Down syndrome; LoF mutations are associated with *minibrain* phenotype in *Drosophila*[65], autism-like behavior in mouse[64] and a deletion syndrome in humans[27,63]. *MBD5* has been implicated as the causal gene for the 2q23.1 deletion syndrome associated with epilepsy, autism and ID[91,92].



in early fetal life have substantial contribution to SCZ by *de novo* mutations. Similarly, Willsey *et al.*[46] working with a few high-confidence sets of ASD-associated genes as seeds reported a convergence of the expression of these genes in deep-layer cortical projection neurons (layers 5 and 6) in mid-fetal development. Another analysis using a larger set of ASD and ID risk genes suggested translational regulation by *FMR1* and an enrichment in superficial cortical layers[43]. Implicit in these types of analyses is the notion that, while more than 1,000 genes may be responsible for ASD or ID, in the end the genes will converge on a few highly enriched networks of related genes. It is possible that molecular therapies targeted to the network at a specific stage of development, as opposed to the individual gene, may be beneficial to specific groups of patients.

Related to this, it is intriguing that several recurring genes and pathways that have been implicated in neurodevelopmental disease have also been associated with different forms of cancer (**Fig. 4**)[47]. While clear-cut examples such as the mutation of the tumor suppressor genes *PTEN* (Cowden syndrome) or *ARID1B* (Coffin-Siris syndrome) in neurodevelopmental disease have been extensively reviewed[48], more recent exome sequencing data from patients with neurodevelopmental disease suggest new links. The most striking observation here is the identical point mutations reported to cause cancer when mutated somatically and severe neurodevelopmental syndromes when mutated in the germline. Examples include the identical mutations in *SETBP1* (ref. 32), *ASXL1* (ref. 49) and *EZH2* (ref. 50), as well as several genes of the RAS–MAP kinase pathway associated with parental-age-effect Mendelian disorders[51] (**Supplementary Table 1**). It is important to stress that this is an observation at an individual gene level and should not be translated to an epidemiological link: that is, this cannot be generalized to speculate that patients with neurodevelopmental disorders in these specific genes will all be at a higher risk for certain cancer types. Instead, it is likely that this convergence represents a selection of genes that are fundamental to cell biology (for example, cell proliferation and/or membership in multi-subunit complexes associated with chromatin remodeling). There is also the distinct possibility of pleiotropy; that is, the genes and pathways have completely unrelated functions, explaining developmental defects and cancer independently. Therefore, *de novo* mutations in those genes can result in different outcomes depending on timing, genetic background and cellular context. Nevertheless, there may be advantages to integrating sequence data from patients with neurodevelopmental disease and massive sequencing programs devoted to the discovery of somatic mutations in tumors—for example, the International Cancer Genome Project[52]. It is possible that these intersections will help to further prioritize genes important in both cellular development and neurodevelopment.

Phenotypic similarity of recurrent *de novo* mutations

Although essential, statistical support of recurrent mutations is not the sole arbiter in determining pathogenicity of particular mutations and genes. In particular, it is important to consider the phenotypic presentation and overlap of the individuals with the same presumptive underlying genetic lesion. In this regard, we note that

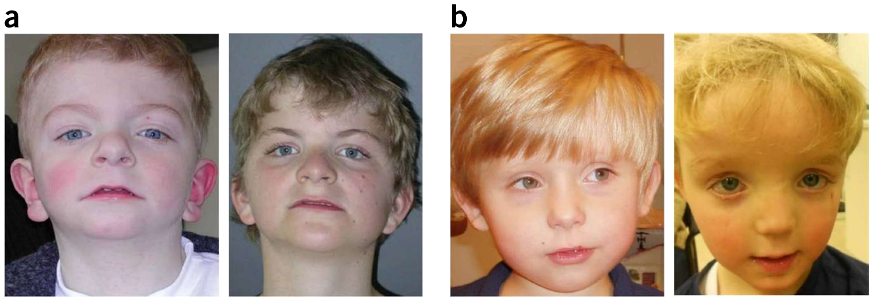


Figure 3 Phenotypic similarity of two patients with identical *PACS1 de novo* mutations and two patients with similar *ADNP* mutations. (**a**) These two unrelated patients show identical *de novo* point mutations (c.607C>T; p.Arg203Trp) in *PACS1* (RefSeq NM_018026.3)[53]. The striking similarity in phenotype includes low anterior hairline, highly arched eyebrows, synophrys, hypertelorism with downslanted palpebral fissures, long eyelashes, a bulbous nasal tip, a flat philtrum with a thin upper lip, downturned corners of the mouth and low-set ears. Reprinted from ref. 53, Copyright (2012), with permission from The American Society of Human Genetics. (**b**) These two unrelated patients both show LoF mutations in *ADNP* (c.2496_2499delTAAA; p.Asp832Lysfs*80 and c.2157C>G; p.Tyr719*)[44] resulting in a new SWI-SNF–related autism syndrome. Patients present with clinical similarities, including a prominent forehead, a thin upper lip and a broad nasal bridge. Reprinted from ref. 44.

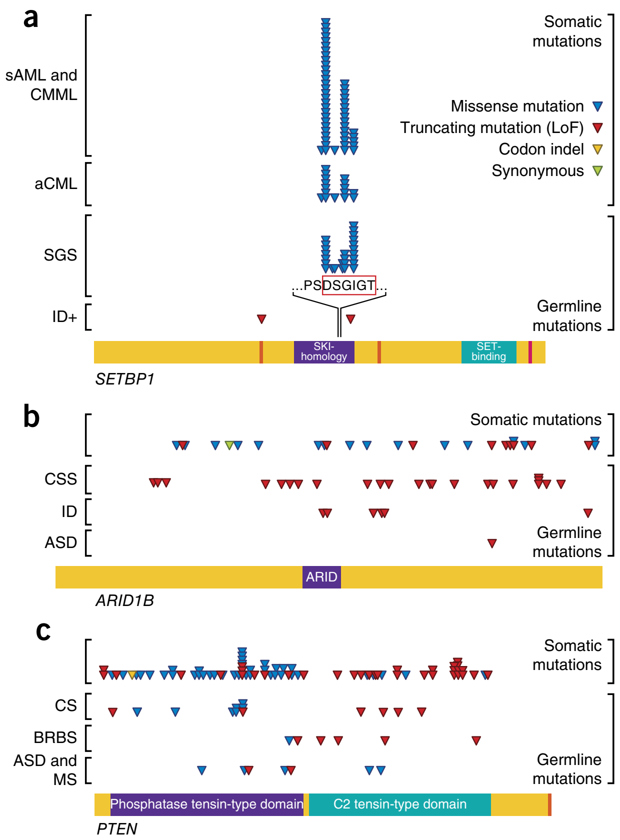


Figure 4 Examples of coincidental *de novo* mutations in cancer and neurodevelopmental disorders. (**a**) Mutation spectrum of *SETBP1*. sAML and CMML, secondary acute myeloid leukemia and chronic myelomonocytic leukemia[93] (p.Asp868Tyr, 1 mutation event identified; p.Asp868Asn, 28 events; p.Ser869Asn, 1 event; p.Gly870Ser, 15 events; p.Ile871Thr, 5 events); aCML, atypical chronic myeloid leukemia[94] (p.Asp868Asn, 7 events; p.Ser869Gly, 1 event; p.Gly870Ser, 5 events; p.Ile871Thr, 2 events); SGS, Schinzel-Giedion syndrome[32] (A.H., unpublished data: p.Asp868Ala, 1 event; p.Asp868Asn, 7 events; p.Ser869Arg, 1 event; p.Ser869Asn, 1 event; p.GGly870Ser, 4 events; p.Gly870Asn, 2 events; p.Ile871Thr, 10 events); ID+, intellectual disability with other features[2,3] (p.Leu592* and p.906fs, 1 event each). The red box shows the amino acids that represent the degron sequence, four of which are frequently mutated. (**b**) Mutation spectrum of *ARID1B*. Somatic mutations retrieved from COSMIC database. Only ‘somatic validated’ and ‘previously described’ somatic mutations with a PubMed entry were considered. CSS, Coffin-Siris syndrome[48,55] (p.Gln408Profs*127, p.Ser413Valfs*122, p.Asn420Lysfs*115, p.Pro449Argfs*53, p.Tyr867Thrfs*47, p.Met935Asnfs*7, p.Ser959Argfs*9, p.Ala1000Argfs*5, p.Arg1075*, p.Gly1283Trpfs*38, p.Arg1337*, p.Tyr1366*, p.Pro1489Leufs*10, p.Tyr1540*, p.Gln1541Argfs*35, p.Trp1637Cysfs*6, p.Lys1777*, p.Phe1798Leufs*52, p.Asp1879Thrfs*95, p.Arg1990*, p.Arg1990*, p.Arg1990*, p.Trp2013*, p.Pro2078Leufs*21); ID[54] (p.Arg372Profs*163, p.Arg1102*, p.Lys1108Argfs*9, p.Gln1307*, p.Tyr1346*, p.Arg1338Argfs*76, p.Ser2155Leufs*33); ASD[3] (p.Phe1798Leufs*52); splice site mutations not considered. ARID, AT-rich interactive domain. (**c**) Mutation spectrum of *PTEN*. Somatic mutations retrieved from the COSMIC database. Only ‘somatic validated’ and ‘previously described’ somatic mutations with at least five independent entries are displayed. CS, Cowden syndrome; ASD and MS, autism spectrum disorder and macrocephaly syndrome; BRBS, Bannayan-Riley-Ruvalcana syndrome (based on OMIM entries); splice site mutations not considered.

many of the initial studies are likely to be enriching for the most severe cases because ascertainment is clinical as opposed to population based. As a result, initial estimates of penetrance may be overestimated and phenotypic heterogeneity underestimated. Nevertheless, identification of clinically recognizable syndromes or sets of phenotypic features has historically been used to strengthen the case for a particular gene's involvement. In the past, gene discovery was usually driven by detailed description of a particular syndrome (for example, fragile X or Rett syndrome), followed by a systematic hunt for the mutated gene. Recognition of clinical subtypes, however, is now beginning to occur after mutation and gene discovery. In the case of *PACS1* (ref. 53), for example, identical *de novo* mutations result in patients with a strikingly similar phenotypic outcomes (**Fig. 3**). Such clinical discernment, now more than ever, requires the expertise of the clinician.

It should be noted, however, that not all genes when mutated will show a phenotypic convergence; they rather may be much more variable in their phenotypic presentation. For example, mutations in *ARID1B* can either lead to isolated ID[54] or syndromic form of ID with a recognizable phenotype known as Coffin-Siris syndrome[48,55]. The type of mutation may be critical in this regard. It is noteworthy that patients with LoF mutations of *SCN2A* described in autism cohorts[4] do not show epilepsy, in contrast to multiple recurrent missense mutations identified among epileptic encephalopathies or, more specifically, the infantile-spasms or Lennox–Gastaut subtype. Similarly, *de novo* missense mutations in *CHD2*, *SETD5* and *SLC6A1* have been reported in patients with ASD, yet frameshift mutations in the same genes are seen in patients with ID but without ASD features[2]. There are several considerations regarding genotype-phenotype correlations.

First, some of the classically defined neurodevelopmental syndromes may present with broader (or milder) phenotypes as defined by initial clinical case reports[2]. There is evidence that the genetic background on which these mutations occur substantially influences phenotypic outcome[56–58].

Second, genotype-first approaches using current genomic technologies followed by ‘reverse phenotyping’ are beginning to define more subtle syndromes that are still opaque in large umbrella cohorts such as ASD or ID[13]. Some examples include macrocephalic subtypes of ASD or ID caused by mutations in *PTEN* and *CHD8* (refs. 20,59) and developmental delay/ID and epilepsy caused by *de novo* mutations in *SCN2A* (refs. 1,2,18).

Third, after discovery of potential causative mutations, more detailed and standardized phenotyping assessments are necessary to eliminate disease ascertainment biases. As patients with a specific mutation will individually be rare, greater coordination, including patient recontact, will need to occur across clinical research centers.

Biological impact in model organisms

Detailed phenotypic characterization of patients is an important first step in modeling mutations in other organisms. Indeed, additional support for a gene's involvement in disease is often provided by related pathologies in these model organisms and may be used to rapidly prioritize genes for further study, as well as to provide further insight into function. In many cases, mouse models[60] or *Drosophila* mutant lines[61] already exist and neurologic phenotypes have been, at least partially, documented. For example, recurrent LoF mutations of *ADNP* (activity-dependent neuroprotective peptide) were recently described in patients with autism and ID[3]. Heterozygous knockout mice show a neuronal and glial pathology associated with reduced cognitive function[62], and this phenotype was recognized in mouse models before the association with human disease. Similarly, heterozygous deletions or mutations of *DYRK1A* in humans[20,63], mice[64] and fruit flies[65] all show a phenotype of reduced brain volume associated with microcephaly. In this regard, it is interesting that the *DYRK1A* LoF mutations were the last to be documented, with the models predating the discovery of human genetic diseases. With new resources such as the Zebrafish Mutation Project[66] and the International Knockout Mouse Consortium[60], we may achieve more systematic and high-throughput genome-wide approaches for model organisms, in particular for LoF mutations.

Limitations and future directions

Despite the great success of recent exome studies, most analyses have so far been restricted to the protein-coding portion of the genome—a very small fraction (1.5%) of all human genetic variation. Furthermore, the definition of the protein-coding portion is far from perfect. Portions of the reference genome[67,68] and gene annotation[69] are incomplete, especially in relation to isoforms specifically expressed in the brain. Regulatory variation and its impact are as yet ignored. Even though genome sequencing costs have decreased, discovery and interpretation of genetic variation remain significant hurdles. Unlike protein-coding sequencing, defining the functional regions and the type of mutations that will abrogate such function remain active areas of research. Nevertheless, the genes where dosage imbalance have been found to strongly associate with disease represent a logical starting point to begin to interrogate regulatory mutations, as well as epigenetic factors that may have a similar effect. Targeted resequencing of the entire genomic loci of interest (that is, including noncoding parts of the genes), as well as full genome sequencing, will undoubtedly find new mutations and further improve our understanding of the phenotype-genotype relationship. Despite the recent emphasis on *de novo* mutations, their contribution to disease can only be understood in the context of the full spectrum of genetic variation of each individual[25,57].

Even for the protein-coding component, sequence discovery is incomplete, with 5–10% of the exons either being missed or insufficiently captured to confidently call genetic variation. The bias is particularly pronounced for genes mapping to high-GC-content regions of the genome, where as many as 20–30% of exons may be insufficiently covered. The sequencing technology also introduces biases against certain types and classes of mutation. The discovery of indels is largely regarded as incomplete because of difficulties associated with mapping short sequence reads in low-complexity regions[70]. Although there have been recent methods for calling smaller CNVs, validation experiments indicating specificity and sensitivity are still far from ideal[71,72]. The development of new sequencing chemistries and platforms that can cheaply and in a high-throughput manner access these regions of the genome should remain a high priority.

There is another level of reduced sensitivity related to the timing of *de novo* mutations. It is increasingly recognized that postzygotic *de novo* mutations—that is, mutations in somatic state—may be important in many more disorders[73]. While the importance of somatic *de novo* mutations has been recognized for many years in the field of cancer genetics[52,74–76], we are only starting to appreciate its prevalence in neurodevelopmental disease[77,78]. Several exome studies report that individual *de novo* mutations are likely to have occurred postzygotically, with estimates ranging from 1 to 2% of new mutations based on analysis of DNA derived from blood. O'Roak *et al.*[3], for example, have shown that ~4% (9 of 209 cases) of *de novo* mutations were likely to have occurred postzygotically. Mosaic mutations have been observed as a more general theme for CNVs but not yet linked systematically to disease[79]. More sensitive technologies[80,81], as well as access to more clinically relevant tissue types, are required to identify lower level mosaicism. For defined disorders with isolated neurologic involvement, this may never be possible if the mosaicism is restricted to neuronal subtypes in the brain.

Besides technical hurdles, there is the daunting prospect of the extreme locus heterogeneity of these diseases. This raises the distinct possibility that a recurrent *de novo* mutation in a second patient will never be seen again in the same clinic. This can be partially overcome by developing new models for data sharing (for example, *de novo* variant databases) and generating larger sample collections of patients (>50,000) that may be screened in follow-up targeted resequencing experiments. This requires a shift toward a more integrated and collaborative model of clinical and basic research. Successful models of clinical laboratory cooperation and standardization have already been established for the exchange of CNV data—for example, the International Standards for Cytogenomic Arrays (ISCA) Consortium—and there is momentum for doing the same for exome and genome sequence data sets—for example, the International Collaboration for Clinical Genomics (ICCG)[24]. The sheer size of the data set (petabytes), ever-changing advances in sequencing technology and the importance of standardized call sets, however, pose major challenges.

Although sporadic mutations have been the focus of this Review, the importance of inherited mutations should not be underestimated. There is, in fact, compelling evidence that such variation contributes substantially to these diseases[45,72,82,83]. While specific gene effects are much more difficult to tease apart in the general population owing to the genetic heterogeneity of these diseases[45], other approaches, such as studies of consanguineous families, have identified many candidate risk genes under a recessive disease model[84–86]. It should also be noted that the effect size and penetrance for many of the recurrent *de novo* mutations is not yet known. For autism, *de novo* mutations have been thought to collectively increase risk 10- to 20-fold for up to 20% of patients with disease. It is likely that in some cases a rare variant will be necessary but not sufficient to confer the phenotype, requiring that both inherited and *de novo* mutations be jointly considered in order to understand their impact, as has been noted for some CNV risk variants[57,87]. Understanding the gender bias, which is particularly pronounced for ASD and ID, will require integrating inherited and *de novo* mutations from both the X chromosome and autosomes. Data from CNVs as well as SNVs suggest that the carrier burdens of males and females differ significantly[42,72,88,89]. The evidence suggests that healthy females are more likely to be carriers of deleterious mutations and, therefore, protected against such diseases. Perhaps sex-dependent modifiers are responsible for slightly different diagnostic boundaries for neurodevelopmental disorders between male and female patients.

Because the physiological function of many of the genes linked to disease are yet unknown, it will be necessary to perform systematic studies to understand their specific role in neurodevelopment. The sheer volume of high-impact genes will probably necessitate large-scale model organism knockouts in *Drosophila*, zebrafish and mouse[60,66], industrial-level development of induced pluripotent stem cell lines and neuronal differentiation protocols[90], as well as massive screens using mass spectrometry to identify protein interaction partners. All of these approaches have their own limitations. For example, it is an open question how well knockouts will model neurodevelopmental diseases such as ASD or ID because most of the known effects in humans occur in the heterozygous state and most knockout phenotypes are typically studied as homozygous LoF mutations. Many phenotypic aspects of complex neuropsychiatric and neurobehavioral disease will not be amenable to model systems, further limiting such functional approaches. Notwithstanding these challenges, we are in a golden age of 'neurogene' discovery that promises not only to improve our understanding of disease but to provide fundamental insight into the biology of human brain development.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS
We are grateful to T. Brown and C. Gilissen for assistance during manuscript preparation, F. Kooy for early preprint access and H.G. Brunner for sharing patient photographs used in **Figure 3**. A.H. is supported by a ZonMW grant (916–12–095); E.E.E. is supported by a US National Institute of Mental Health grant (1R01MH101221–01) and is an investigator of the Howard Hughes Medical Institute.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
2. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
3. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
4. Sanders, S.J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
5. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
6. Jiang, Y.-H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013).
7. Neale, B.M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
8. Gulsuner, S. *et al.* Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013).
9. Xu, B. *et al. De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat. Genet.* **44**, 1365–1369 (2012).
10. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* doi:10.1038/nature12929 (2014).
11. Allen, A.S. *et al. De novo* mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
12. Veltman, J.A. & Brunner, H.G. *De novo* mutations in human genetic disease. *Nat. Rev. Genet.* **13**, 565–575 (2012).
13. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
14. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
15. Malaspina, D. *et al.* Advancing paternal age and the risk of schizophrenia. *Arch. Gen. Psychiatry* **58**, 361–367 (2001).
16. Hultman, C.M., Sandin, S., Levine, S.Z., Lichtenstein, P. & Reichenberg, A. Advancing paternal age and risk of autism: new evidence from a population-based study and a meta-analysis of epidemiological studies. *Mol. Psychiatry* **16**, 1203–1212 (2011).
17. McGrath, J.J. *et al.* A comprehensive assessment of parental age and psychiatric disorders. *JAMA Psychiatry* **71**, 301–309 (2014).
18. Carvill, G.L. *et al.* Targeted resequencing in epileptic encephalopathies identifies *de novo* mutations in *CHD2* and *SYNGAP1*. *Nat. Genet.* **45**, 825–830 (2013).
19. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
20. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
21. O'Roak, B.J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**, 1619–1622 (2012).
22. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nat. Genet.* **43**, 585–589 (2011).
23. Cooper, G.M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
24. Kaminsky, E.B. *et al.* An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet. Med.* **13**, 777–784 (2011).
25. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
26. Vulto-van Silfhout, A.T. *et al.* Clinical significance of de novo and inherited copy number variation. *Hum. Mutat.* **34**, 1679–1687 (2013).
27. Møller, R.S. *et al.* Truncation of the Down syndrome candidate gene *DYRK1A* in two unrelated patients with microcephaly. *Am. J. Hum. Genet.* 1165–1170 (2008).
28. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
29. van Bokhoven, H. & Brunner, H.G. Splitting p63. *Am. J. Hum. Genet.* **71**, 1–13 (2002).
30. Bowen, M.E. *et al.* Loss-of-function mutations in *PTPN11* cause metachondromatosis, but not Ollier disease or Maffucci syndrome. *PLoS Genet.* **7**, e1002050 (2011).
31. Tartaglia, M. & Gelb, B. Noonan syndrome and related disorders. *Annu. Rev. Genomics Hum. Genet.* **6**, 45–68 (2005).
32. Hoischen, A. *et al. De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).
33. Filges, I. *et al.* Reduced expression by *SETBP1* haploinsufficiency causes developmental and expressive language delay indicating a phenotype distinct from Schinzel-Giedion syndrome. *J. Med. Genet.* **48**, 117–122 (2011).
34. Marseglia, G. *et al.* 372 kb microdeletion in 18q12.3 causing *SETBP1* haploinsufficiency associated with mild mental retardation and expressive speech impairment. *Eur. J. Med. Genet.* **55**, 216–221 (2012).
35. Kamath, B.M. *et al. NOTCH2* mutations in Alagille syndrome. *J. Med. Genet.* **49**, 138–144 (2012).
36. Isidor, B. *et al.* Truncating mutations in the last exon of *NOTCH2* cause a rare skeletal disorder with osteoporosis. *Nat. Genet.* **43**, 306–308 (2011).
37. Simpson, M.A. *et al.* Mutations in *NOTCH2* cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. *Nat. Genet.* **43**, 303–305 (2011).
38. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
39. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
40. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
41. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14** (suppl. 3), S3 (2013).
42. Gilman, S.R. *et al.* Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70**, 898–907 (2011).
43. Parikshak, N.N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* (in the press).
44. Helsmoortel, C. *et al.* A SWI/SNF-related autism syndrome caused by *de novo* mutations in *ADNP*. *Nat. Genet.* **46**, 380–384 (2014).
45. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
46. Willsey, A.J. *et al.* Co-expression networks implicate human mid-fetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).
47. Ronan, J.L., Wu, W. & Crabtree, G.R. From neural development to cognition: unexpected roles for chromatin. *Nat. Rev. Genet.* **14**, 347–359 (2013).
48. Santen, G.W.E. *et al.* Coffin-Siris syndrome and the BAF complex: genotype-phenotype study in 63 patients. *Hum. Mutat.* doi:10.1002/humu.22394 (2013).
49. Hoischen, A. *et al. De novo* nonsense mutations in *ASXL1* cause Bohring-Opitz syndrome. *Nat. Genet.* **43**, 729–731 (2011).
50. Gibson, W.T. *et al.* Mutations in EZH2 cause Weaver syndrome. *Am. J. Hum. Genet.* **90**, 110–118 (2012).
51. Goriely, A. & Wilkie, A.O.M. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90**, 175–200 (2012).
52. Hudson, T.J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
53. Schuurs-Hoeijmakers, J.H.M. *et al.* Recurrent *de novo* mutations in *PACS1* cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *Am. J. Hum. Genet.* **91**, 1122–1127 (2012).
54. Hoyer, J. *et al.* Haploinsufficiency of *ARID1B*, a member of the SWI/SNF-a chromatin-remodeling complex, is a frequent cause of intellectual disability. *Am. J. Hum. Genet.* **90**, 565–572 (2012).
55. Santen, G.W.E. *et al.* Mutations in SWI/SNF chromatin remodeling complex gene *ARID1B* cause Coffin-Siris syndrome. *Nat. Genet.* **44**, 379–380 (2012).
56. Girirajan, S. *et al.* Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. *Am. J. Hum. Genet.* **92**, 221–237 (2013).
57. Girirajan, S. *et al.* Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N. Engl. J. Med.* **367**, 1321–1331 (2012).
58. Classen, C.F. *et al.* Dissecting the genotype in syndromic intellectual disability using whole exome sequencing in addition to genome-wide copy number analysis. *Hum. Genet.* **132**, 825–841 (2013).
59. Zaidi, S. *et al. De novo* mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220–223 (2013).

60. Skarnes, W.C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337–342 (2011).
61. Tweedie, S. *et al.* FlyBase: enhancing *Drosophila* Gene Ontology annotations. *Nucleic Acids Res.* **37**, D555–D559 (2009).
62. Vulih-Shultzman, I. *et al.* Activity-dependent neuroprotective protein snippet NAP reduces tau hyperphosphorylation and enhances learning in a novel transgenic mouse model. *J. Pharmacol. Exp. Ther.* **323**, 438–449 (2007).
63. van Bon, B.W.M. *et al.* Intragenic deletion in *DYRK1A* leads to mental retardation and primary microcephaly. *Clin. Genet.* **79**, 296–299 (2011).
64. Fotaki, V. *et al. Dyrk1A* haploinsufficiency affects viability and causes developmental delay and abnormal brain morphology in mice. *Mol. Cell. Biol.* **22**, 6636–6647 (2002).
65. Tejedor, F. *et al.* Minibrain: a new protein kinase family involved in postembryonic neurogenesis in *Drosophila*. *Neuron* **14**, 287–301 (1995).
66. Kettleborough, R.N.W. *et al.* A systematic genome-wide analysis of zebrafish protein-coding gene function. *Nature* **496**, 494–497 (2013).
67. Genovese, G. *et al.* Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406–414 (2013).
68. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
69. Beyer, K. *et al.* New brain-specific beta-synuclein isoforms show expression ratio changes in Lewy body diseases. *Neurogenetics* **13**, 61–72 (2012).
70. Karakoc, E. *et al.* Detection of structural variants and indels within exome data. *Nat. Methods* **9**, 176–178 (2012).
71. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
72. Krumm, N. *et al.* Transmission disequilibrium of small CNVs in simplex autism. *Am. J. Hum. Genet.* **93**, 595–606 (2013).
73. Lupski, J.R. Genetics. Genome mosaicism–one human, multiple genomes. *Science* **341**, 358–359 (2013).
74. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
75. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
76. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
77. Banka, S. *et al.* MLL2 mosaic mutations and intragenic deletion-duplications in patients with Kabuki syndrome. *Clin. Genet.* **83**, 467–471 (2013).
78. Huisman, S.A., Redeker, E.J.W., Maas, S.M., Mannens, M.M. & Hennekam, R.C.M. High rate of mosaicism in individuals with Cornelia de Lange syndrome. *J. Med. Genet.* **50**, 339–344 (2013).
79. Rodríguez-Santiago, B. *et al.* Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am. J. Hum. Genet.* **87**, 129–138 (2010).
80. Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J. & Shendure, J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* **23**, 843–854 (2013).
81. Shiroguchi, K., Jia, T.Z., Sims, P.A. & Xie, X.S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci. USA* **109**, 1347–1352 (2012).
82. Klei, L. *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism* **3**, 9 (2012).
83. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
84. Yu, T.W. *et al.* Using whole-exome sequencing to identify inherited causes of autism. *Neuron* **77**, 259–273 (2013).
85. Morrow, E.M. *et al.* Identifying autism loci and genes by tracing recent shared ancestry. *Science* **321**, 218–223 (2008).
86. Najmabadi, H. *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478**, 57–63 (2011).
87. He, X. *et al.* Integrated model of *de novo* and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).
88. Levy, D. *et al.* Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886–897 (2011).
89. Jacquemont, S. *et al.* A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94**, 415–425 (2014).
90. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
91. van Bon, B.W.M. *et al.* The 2q23.1 microdeletion syndrome: clinical and behavioural phenotype. *Eur. J. Hum. Genet.* **18**, 163–170 (2010).
92. Talkowski, M.E. *et al.* Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am. J. Hum. Genet.* **89**, 551–563 (2011).
93. Makishima, H. *et al.* Somatic *SETBP1* mutations in myeloid malignancies. *Nat. Genet.* **45**, 942–946 (2013).
94. Piazza, R. *et al.* Recurrent *SETBP1* mutations in atypical chronic myeloid leukemia. *Nat. Genet.* **45**, 18–24 (2013).