Published in final edited form as:
Nat Neurosci. 2014 June ; 17(6): 764–772. doi:10.1038/nn.3703.

Prioritization of neurodevelopmental disease genes by discovery of new mutations

Alexander Hoischen[1], **Niklas Krumm**[2], and **Evan E. Eichler**[2,3]

[1]Department of Human Genetics, Nijmegen Center for Molecular Life Sciences, Institute for Genetic and Metabolic Disease, Radboud university medical center, Nijmegen, Netherlands [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA [3]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

Abstract

Advances in genome sequencing technologies have begun to revolutionize neurogenetics allowing the full spectrum of genetic variation to be better understood in relationship to disease. Exome sequencing of hundreds to thousands of samples from patients with autism spectrum disorder, intellectual disability, epilepsy, and schizophrenia provide 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 numerous factors including recurrence, prior evidence of overlap with pathogenic copy number variants, the position of the mutation within the protein, the mutational burden among healthy individuals, and membership of the candidate gene within disease-implicated protein networks. We discuss these emerging criteria for gene prioritization and the potential impact on the field of neuroscience.

INTRODUCTION

Recent exome (and genome) sequencing studies of families have aimed to comprehensively discover genetic variation in order to identify the most likely causal mutation in patients with disease. Sequencing studies of parent-proband trios 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 play an important role 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/fecundity of the given condition[12]. However, not only *de novo* events but

Correspondence to: Evan E. Eichler, Ph.D., Department of Genome Sciences, University of Washington School of Medicine, Foege S-413A, Box 355065, 3720 15th Ave NE, Seattle, WA 98195-5065, eee@gs.washington.edu; Alexander Hoischen, Ph.D., Department of Human Genetics, Radboud university medical center, Geert Grooteplein 10 (route 855), P. O. Box 9101, 6500 HB Nijmegen, alexander.hoischen@radboudumc.nl.

CONFLICT OF INTEREST STATEMENT

E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. and was an SAB member of Pacific Biosciences, Inc. (2009–2013) and SynapDx Corp. (2011–2013).

also rare inherited CNVs can have an effect on fecundity, their overall effect on fecundity is however still debated[13]. Biologically, 75–80% of *de novo* point mutations arise paternally[3,14] likely due to increasing numbers of cell divisions in the male germline lineage when compared to the female lineage. These findings are consistent with some epidemiological data which find advancing paternal age as 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 candidate gene identification. Deciding which genes to pursue, however, is not always self-evident since follow-up research and diagnostic studies are critical to understand the full contribution of a particular mutation to its respective phenotype.

In this review, we will discuss the prioritization of candidate genes, show emerging trends, and highlight potential strategies for subsequent functional characterization of these neurodevelopmental genes. We focus on lessons learned from eleven 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 (EE) 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 as ‘neurodevelopmental disorders’ but recognize that especially adult-onset diseases such as SCZ have etiologic components that are not neurodevelopmental in origin.

1) 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 as well as the absence of such mutations in healthy controls. This rule follows the precedent established for the discovery of pathogenic *de novo* copy number variants (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 are found, to date, exclusively in probands but never in controls. Simulation data suggest that at least two but certainly three or more recurrent *de novo* loss-of-function (LoF) events 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), diseases such simplex autism and SCZ are thought to arise from >500 genes, while studies of severe

ID and epileptic encephalopathies (infantile spasms and Lennox-Gastaux subtypes) suggest lower heterogeneity. Such estimates should be considered only rough approximations at this point since 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 it is interesting that several of these top-scoring candidate genes 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 the level of 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 DD (developmental delay) or ASD[11,18]. One study of SCZ, for example, highlighted four recurrently mutated genes, but maybe more remarkable the same study identified overlapping genes with 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 disease stems from a neurodevelopmental origin.

Perhaps, the most striking examples are recurrent identical *de novo* mutations within 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 significance of the majority of *de novo* mutations remains unclear. For most genes, only a single *de novo* mutation has been identified. Nevertheless, based on the observation that *de novo* LoF mutations occur 2–3 times more frequently in ASD probands when compared to 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-value = 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 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 (<$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].

2) Prior 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 significantly 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 of 'causality' of *de novo* LoF mutations (Fig. 2).

3) Position of the mutation within the protein

The bulk of *de novo* mutations discovered from exome sequencing projects are missense mutations, with more than 60% in probands and controls (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 either based on the protein domain they affect, the position in the gene (e.g., the amino- or carboxy-terminus of the resulting protein)[29], or their potential to modify the normal function of the protein. Examples for the latter 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/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 (NMD) and result in Hajdu-Cheney syndrome[36,37].

4) 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 has been used to define the load of LoF mutations per gene[2] and to prioritize >4,000 human genes that are most intolerant to variation[11]. Another approach uses 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 hone in on genes that are likely to be among the most penetrant because of the complete absence of disruptive variation in the general population (e.g., *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 5′ or 3′) in both cases and controls may suggest mis-annotation, the presence of a processed pseudogene, or an alternative nonfunctional splice form.

5) Pathway enrichment and links to cancer biology

Another popular approach to suss out the most important gene candidates for further 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 co-expression network may increase the likelihood of its association with disease. Numerous studies have reported significant enrichment of both *de novo* CNV and SNV mutations in particular pathways[3,4,42,43]. O'Roak *et al.*, for example, reported a significant enrichment of *de novo* disruptive autism mutations among proteins associated with chromatin remodeling, beta-catenin and WNT signaling—a finding that was replicated in a follow-up resequencing study of more than 2,400 probands. One instance, in which membership of a new candidate gene within a PPI network led to the discovery of an autism gene, is the recent example of *ADNP*. A single *ADNP* LoF mutation was initially observed from exome sequencing studies. Although the gene did not reach statistical significance when comparing cases and controls[20], it was strongly implicated in the PPI network originally defined by O'Roak *et al*. Targeted resequencing experiments combined with clinical exome sequencing identified several additional 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 beta-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 *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 degree of cases that also presented with comorbid ID remains to be determined.

In addition to PPI networks, studies of co-expression have shown enrichment for specific spatio-temporal patterns of expression. A study of co-expressed 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 in early fetal life have

significant contribution to SCZ by *de novo* mutations. Similarly, Willsey *et al.*[46] working with a few high-confidence sets of 'ASD genes' as seeds reported a convergence of 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 *FMRP* 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 like the mutation of tumor suppressor genes, *PTEN* (Cowden syndrome) or *ARID1B* (Coffin-Siris syndrome), and neurodevelopmental disease have been extensively reviewed[48], more recent exome sequencing data from patients with neurodevelopmental disease suggests potentially 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*[32], *ASXL1*[49], and *EZH2*[50], as well as several genes of the RAS-MAP-kinase pathway associated with parental-age effect Mendelian disorders[51] (Table 5). It is important to stress that this is an observation at an individual gene level and should not be translated to an epidemiological link, i.e., 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 play a fundamental role in cell biology (e.g., cell proliferation and/or membership in multi-subunit complexes associated with chromatin remodeling). There is also the distinct possibility of pleiotropy; i.e., 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 within tumors, e.g., 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

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 many of the initial studies are likely enriching for the most severe cases since 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 of a particular gene's

involvement. In the past, gene discovery was usually driven by detailed description of a particular syndrome (e.g., 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*[53], for example, identical *de novo* mutations result in patients with a strikingly similar phenotypic outcome (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 but rather may be much more variable in their phenotypic presentation. For example, mutations in *ARID1B* can either lead to isolated ID[54] or syndromic forms 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 present with epilepsy in contrast to multiple recurrent missense mutations identified among epileptic encephalopathies, or more specifically infantile spasms or Lennox-Gastault 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.

1. Some of the classically defined neurodevelopmental clinically defined syndromes may present with broader (or milder) phenotypes as defined by initial clinical case reports[2]. There is evidence that the genetic background upon which these mutations occur significantly influences phenotypic outcome[56-58].
2. ‘Genotype-first’ approaches using current genomic technologies followed by ‘reverse phenotyping’ are beginning to define more subtle syndromes that are still opaque within large umbrella cohorts such as ASD or ID[13]. Some examples include macrocephalic subtypes of ASD/ID caused by mutations in *PTEN* and *CHD8*[20,59] and DD/ID and epilepsy caused by *de novo* mutations in *SCN2A*[1,2,18].
3. After discovery of potential causative mutations, more detailed and standardized phenotyping assessments are necessary to eliminate disease ascertainment biases. Since patients with a specific mutation will be individually rare, greater coordination, including patient recontact, will need to occur across clinical research centers.

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 often is provided by related pathologies in these model organisms and may be used to rapidly prioritize genes for further study as well as to provide additional 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 glial pathology associated with reduced cognitive function[62] and this phenotype was recognized in mouse models prior to the association in human disease. Similarly, heterozygous deletions or mutations of *DYRK1A* in humans[20,63], mice[64] and fruitflies[65] all show a phenotype of

reduced brain volume associated with microcephaly. In this regard, it is interesting that the *DYRK1A* LoF were the last to be documented with the models preexisting the discovery of human genetic diseases. With new resources like the Zebrafish Mutation Project[66] and the International Knockout Mouse Consortium[60], more systematic and high-throughput genome-wide approaches for model organisms may be achieved in particular for LoF mutations.

Limitations and future directions

Despite the great success of recent exome studies, it is important to note that most of the analyses, to date, have 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 currently ignored. Even though genome sequencing costs have reduced, 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 effects that may have a similar effect. Targeted resequencing of the entire genomic loci as well as full genome sequencing will undoubtedly discover 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 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 the 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 to call 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, i.e., mutations in somatic state, may play an important role in neurodevelopmental 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 likely occurred postzygotically with estimates ranging from 1–2% of new mutations based on analysis of DNA derived from blood. O'Roak *et al.*[3], for example, have shown that ~4% (9/209 cases)

of *de novo* mutations likely 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] are required to identify lower-level mosaics as well as access to clinically more relevant tissue types. For defined disorders with isolated neurologic involvement, this may never be possible if the mosaicism is restricted to neuronal subtypes in the brain.

Next to 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 (e.g., *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 lab cooperation and standardization have already been established for the exchange of CNV data, e.g., the ISCA (International Standards for Cytogenomic Arrays) Consortium, and there is momentum to do the same for exome and genome sequence datasets, e.g., the ICCG (International Collaboration for Clinical Genomics)[24]. The sheer size of the dataset (multiple Petabytes), ever-changing advances in sequencing technology, and the importance of standardized call sets, however, pose major challenges moving forward.

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 significantly 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 numerous 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 cases with disease. It is likely that in some cases a rare variant will be necessary but not sufficient to confer 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 burden of males and females differ significantly[42,72,88,89]. The evidence suggests that females are more likely to be carriers of deleterious mutations and, therefore, protected against such diseases perhaps sex dependent modifiers map genetic architecture to diagnostic boundaries differently between the sexes.

Since many of the genes linked to disease will likely have minimal functional annotation in the human genome, 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 iPSC cell lines and neuronal differentiation protocols[90], as well as massive screens using mass spectrometry to identify protein-interacting 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 since most of the known effects in humans occur in the heterozygous state and most knockout phenotypes are typically studied as homozygous LoF. Many phenotypic aspects of complex neuropsychiatric and neurobehavioral disease will not be amenable to model systems further limiting such functional approaches. Notwithstanding these challenges, it is a golden age of 'neurogene' discovery which promises to improve not only our understanding of disease but provide fundamental insight into the biology of human brain development.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We grateful to Tonia Brown and Christian Gilissen for assistance during manuscript preparation; Frank Kooy for early preprint access and Han G. Brunner for sharing patient photographs used for Figure 3. A.H. is supported by a ZonMW grant (916–12–095); E.E.E. is supported by a National Institute of Mental Health (NIMH) grant (1R01MH101221–01) and is an investigator of the Howard Hughes Medical Institute.

REFERENCES

1. De Ligt J, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N. Engl. J. Med. 2012; 367:1921–9. [PubMed: 23033978]
2. Rauch A, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012; 380:1674–82. [PubMed: 23020937]
3. O'Roak BJ, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012; 485:246–50. [PubMed: 22495309]
4. Sanders SJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012; 485:237–41. [PubMed: 22495306]
5. Iossifov I, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012; 74:285–99. [PubMed: 22542183]
6. Jiang Y-H, et al. Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing. Am. J. Hum. Genet. 2013; 93:249–263. [PubMed: 23849776]
7. Neale BM, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012; 485:242–5. [PubMed: 22495311]
8. Gulsuner S, et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell. 2013; 154:518–29. [PubMed: 23911319]
9. Xu B, et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat. Genet. 2012; 44:1365–9. [PubMed: 23042115]
10. Fromer M, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014 doi:10.1038/nature12929.
11. Allen AS, et al. De novo mutations in epileptic encephalopathies. Nature. 2013; 501:217–21. [PubMed: 23934111]
12. Veltman, J. a; Brunner, HG. De novo mutations in human genetic disease. Nat. Rev. Genet. 2012; 13:565–75. [PubMed: 22805709]
13. Stefansson H, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature. 2014; 505:361–6. [PubMed: 24352232]
14. Kong A, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; 488:471–475. [PubMed: 22914163]

15. Malaspina D, et al. Advancing paternal age and the risk of schizophrenia. Arch. Gen. Psychiatry. 2001; 58:361–367. [PubMed: 11296097]
16. Hultman CM, Sandin S, Levine SZ, 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. 2011; 16:1203–1212. [PubMed: 21116277]
17. McGrath JJ, et al. A Comprehensive Assessment of Parental Age and Psychiatric Disorders. JAMA Psychiatry. 2014; 4072:1–9.
18. Carvill GL, et al. Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. Nat. Genet. 2013; 45:825–30. [PubMed: 23708187]
19. Yang Y, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N. Engl. J. Med. 2013 131002140031007. doi:10.1056/NEJMoa1306555.
20. O'Roak BJ, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science. 2012; 338:1619–22. [PubMed: 23160955]
21. O'Roak BJ, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science. 2012; 338:1619–22. [PubMed: 23160955]
22. O'Roak BJ, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat. Genet. 2011; 43:585–589. [PubMed: 21572417]
23. Cooper GM, et al. A copy number variation morbidity map of developmental delay. Nat. Genet. 2011; 43:838–46. [PubMed: 21841781]
24. Kaminsky EB, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet. Med. 2011; 13:777–84. [PubMed: 21844811]
25. Stefansson H, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008; 455:232–6. [PubMed: 18668039]
26. Vulto-van Silfhout AT, et al. Clinical significance of de novo and inherited copy number variation. Hum. Mutat. 2013 doi:10.1002/humu.22442.
27. Møller RS, et al. Truncation of the Down Syndrome Candidate Gene DYRK1A in Two Unrelated Patients with Microcephaly. 2008:1165–1170. doi:10.1016/j.ajhg.2008.03.001.
28. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. 2010; 6:e1001154. [PubMed: 20976243]
29. Van Bokhoven H, Brunner HG. Splitting p63. Am. J. Hum. Genet. 2002; 71:1–13. [PubMed: 12037717]
30. Bowen ME, et al. Loss-of-function mutations in PTPN11 cause metachondromatosis, but not Ollier disease or Maffucci syndrome. PLoS Genet. 2011; 7:e1002050. [PubMed: 21533187]
31. Tartaglia M, Gelb B. Noonan Syndrome and Related Disorders. Annu. Rev. Genomics Hum. Genet. 2005; 6:45–68. [PubMed: 16124853]
32. Hoischen A, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. 2010; 42:483–5. [PubMed: 20436468]
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. 2011; 48:117–22. [PubMed: 21037274]
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. 2012; 55:216–21. [PubMed: 22333924]
35. Kamath BM, et al. NOTCH2 mutations in Alagille syndrome. J. Med. Genet. 2012; 49:138–44. [PubMed: 22209762]
36. Isidor B, et al. Truncating mutations in the last exon of NOTCH2 cause a rare skeletal disorder with osteoporosis. Nat. Genet. 2011; 43:306–8. [PubMed: 21378989]
37. Simpson, M. a, et al. Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nat. Genet. 2011; 43:303–5. [PubMed: 21378985]
38. Kryukov GV, Pennacchio L. a, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. 2007; 80:727–39. [PubMed: 17357078]

39. Khurana E, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013; 342:1235587. [PubMed: 24092746]
40. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants.
41. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013; 14(Suppl 3):S3. [PubMed: 23819870]
42. Gilman SR, et al. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. 2011; 70:898–907. [PubMed: 21658583]
43. Parikshak NN, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell. in press.
44. Helsmoortel C, et al. A SWI/SNF-related autism syndrome caused by de novo mutations in ADNP. Nat. Genet. 2014 doi:10.1038/ng.2899.
45. Purcell SM, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014 doi:10.1038/nature12975.
46. Willsey AJ, et al. Co-expression networks implicate human mid-fetal deep cortical projection neurons in the pathogenesis of autism. Cell. in press.
47. Ronan JL, Wu W, Crabtree GR. From neural development to cognition: unexpected roles for chromatin. Nat. Rev. Genet. 2013; 14:347–59. [PubMed: 23568486]
48. Santen GWE, et al. Coffin-Siris Syndrome and the BAF Complex: Genotype-Phenotype Study in 63 Patients. Hum. Mutat. 2013 doi:10.1002/humu.22394.
49. Hoischen A, et al. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat. Genet. 2011; 43:729–31. [PubMed: 21706002]
50. Gibson WT, et al. Mutations in EZH2 cause Weaver syndrome. Am. J. Hum. Genet. 2012; 90:110–8. [PubMed: 22177091]
51. Goriely A, Wilkie AOM. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. Am. J. Hum. Genet. 2012; 90:175–200. [PubMed: 22325359]
52. Hudson TJ, et al. International network of cancer genome projects. Nature. 2010; 464:993–8. [PubMed: 20393554]
53. Schuurs-Hoeijmakers JHM, 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. 2012; 91:1122–7. [PubMed: 23159249]
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. 2012; 90:565–72. [PubMed: 22405089]
55. Santen GWE, et al. Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nat. Genet. 2012; 44:379–80. [PubMed: 22426309]
56. Girirajan S, et al. Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder. Am. J. Hum. Genet. 2013; 92:221–37. [PubMed: 23375656]
57. Girirajan S, et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. N. Engl. J. Med. 2012; 367:1321–31. [PubMed: 22970919]
58. Classen CF, et al. Dissecting the genotype in syndromic intellectual disability using whole exome sequencing in addition to genome-wide copy number analysis. Hum. Genet. 2013; 132:825–41. [PubMed: 23552953]
59. Zaidi S, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013; 498:220–3. [PubMed: 23665959]
60. Skarnes WC, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011; 474:337–42. [PubMed: 21677750]
61. Tweedie S, et al. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 2009; 37:D555–9. [PubMed: 18948289]

62. Vulih-shultzman I, et al. Activity-Dependent Neuroprotective Protein Snippet NAP Reduces Tau Hyperphosphorylation and Enhances Learning in a Novel Transgenic Mouse Model. 2007; 323:438–449.
63. Van Bon BWM, et al. Intragenic deletion in DYRK1A leads to mental retardation and primary microcephaly. Clin. Genet. 2011; 79:296–9. [PubMed: 21294719]
64. Fotaki V, et al. Dyrk1A Haploinsufficiency Affects Viability and Causes Developmental Delay and Abnormal Brain Morphology in Mice Dyrk1A Haploinsufficiency Affects Viability and Causes Developmental Delay and Abnormal Brain Morphology in Mice. 2002 doi:10.1128/MCB.22.18.6636.
65. Tejedor F, et al. minibrain: a new protein kinase family involved in postembryonic neurogenesis in Drosophila. Neuron. 1995; 14:287–301. [PubMed: 7857639]
66. Kettleborough RNW, et al. A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature. 2013; 496:494–7. [PubMed: 23594742]
67. Genovese G, et al. Using population admixture to help complete maps of the human genome. Nat. Genet. 2013; 45:406–14. 414e1–2. [PubMed: 23435088]
68. Sudmant PH, et al. Diversity of human copy number variation and multicopy genes. Science. 2010; 330:641–6. [PubMed: 21030649]
69. Beyer K, et al. New brain-specific beta-synuclein isoforms show expression ratio changes in Lewy body diseases. Neurogenetics. 2012; 13:61–72. [PubMed: 22205345]
70. Karakoc E, et al. Detection of structural variants and indels within exome data. Nat. Methods. 2012; 9:176–8. [PubMed: 22179552]
71. Fromer M, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. Am. J. Hum. Genet. 2012; 91:597–607. [PubMed: 23040492]
72. Krumm N, et al. Transmission Disequilibrium of Small CNVs in Simplex Autism. Am. J. Hum. Genet. 2013; 93:595–606. [PubMed: 24035194]
73. Lupski JR. Genetics. Genome mosaicism--one human, multiple genomes. Science. 2013; 341:358–9. [PubMed: 23888031]
74. Greenman C, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446:153–8. [PubMed: 17344846]
75. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–21. [PubMed: 23945592]
76. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674. [PubMed: 21376230]
77. Banka S, et al. MLL2 mosaic mutations and intragenic deletion-duplications in patients with Kabuki syndrome. Clin. Genet. 2013; 83:467–71. [PubMed: 22901312]
78. Huisman, S. a; Redeker, EJW.; Maas, SM.; Mannens, MM.; Hennekam, RCM. High rate of mosaicism in individuals with Cornelia de Lange syndrome. J. Med. Genet. 2013; 50:339–44. [PubMed: 23505322]
79. Rodríguez-Santiago B, et al. Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. Am. J. Hum. Genet. 2010; 87:129–38. [PubMed: 20598279]
80. Hiatt JB, Pritchard CC, Salipante SJ, O'Roak BJ, Shendure J. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. 2013; 23:843–54. [PubMed: 23382536]
81. Shiroguchi K, Jia TZ, Sims P. a, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc. Natl. Acad. Sci. U. S. A. 2012; 109:1347–52. [PubMed: 22232676]
82. Klei L, et al. Common genetic variants, acting additively, are a major source of risk for autism. Mol. Autism. 2012; 3:9. [PubMed: 23067556]
83. Lee SH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. 2013; 45:984–94. [PubMed: 23933821]
84. Yu TW, et al. Using whole-exome sequencing to identify inherited causes of autism. Neuron. 2013; 77:259–73. [PubMed: 23352163]

85. Morrow EM, et al. Identifying autism loci and genes by tracing recent shared ancestry. Science. 2008; 321:218–23. [PubMed: 18621663]

86. Najmabadi H, et al. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature. 2011; 478:57–63. [PubMed: 21937992]

87. He X, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013; 9:e1003671. [PubMed: 23966865]

88. Levy D, et al. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron. 2011; 70:886–97. [PubMed: 21658582]

89. Jacquemont S, et al. A Higher Mutational Burden in Females Supports a "'Female Protective Model'" in Neurodevelopmental Disorders. Am. J. Hum. Genet. 2014 in press.

90. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006; 126:663–76. [PubMed: 16904174]

91. Van Bon BWM, et al. The 2q23.1 microdeletion syndrome: clinical and behavioural phenotype. Eur. J. Hum. Genet. 2010; 18:163–70. [PubMed: 19809484]

92. Talkowski ME, 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. 2011; 89:551–63. [PubMed: 21981781]

93. Makishima H, et al. Somatic SETBP1 mutations in myeloid malignancies. Nat. Genet. 2013; 45:942–6. [PubMed: 23832012]

94. Piazza R, et al. Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. Nat. Genet. 2013; 45:18–24. [PubMed: 23222956]

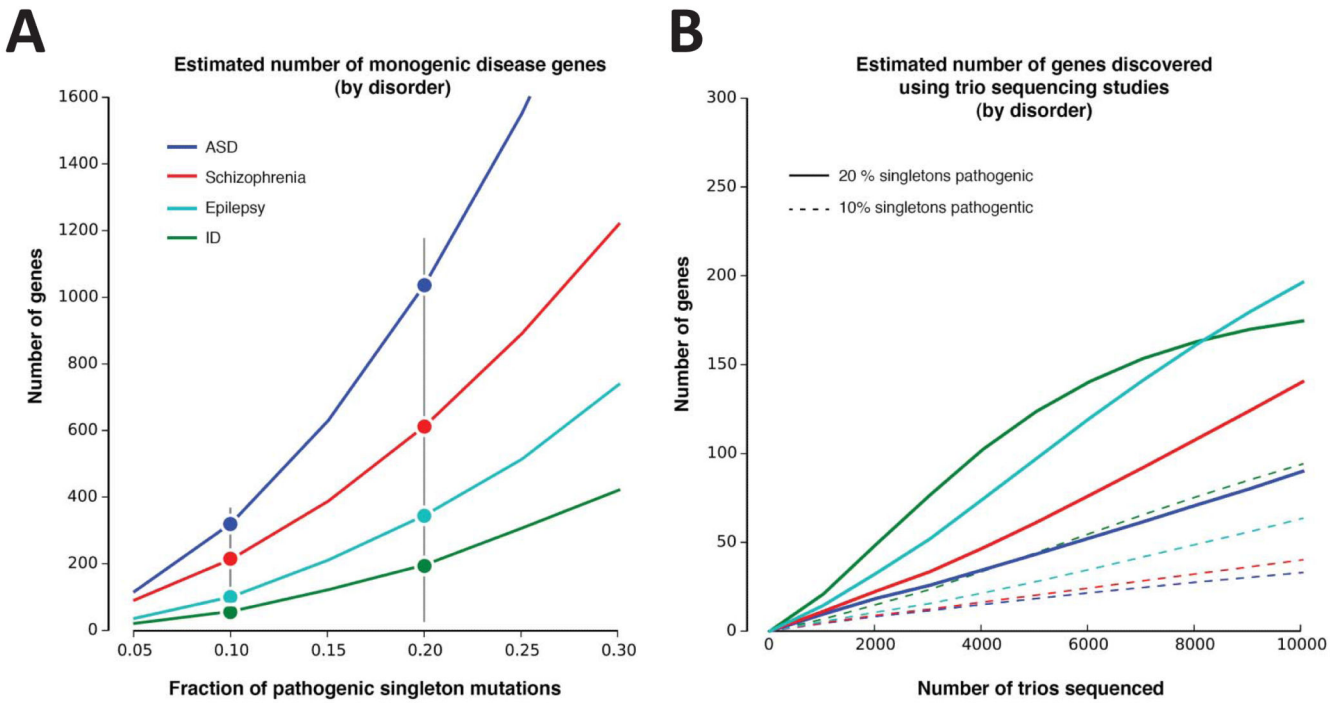

Figure 1. Genes with recurrent *de novo* mutations in four neurodevelopmental disorders

(A) We estimate the number of fully penetrant disease genes based on a *de novo* model using the "Unseen Species Problem". We consider all recurrent missense or loss-of-function *de novo* mutations pathogenic, as well as a defined fraction of mutations in genes observed just once (because all *de novo* mutations are unlikely to be pathogenic). The ratio between genes mutated recurrently and the rate of "singleton" mutations suggests an estimate for the "true" number of pathogenic genes. Including more singleton mutations increases the fraction of each disorder explained by single *de novo* SNVs at the "cost" of including more pathogenic genes. 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 genes is likely to be much higher. **(B)** Expected hit rate (or sensitivity) of true positive genes discovered using trio sequencing studies (under a family-wise error rate of 5%, i.e. each gene passes exome-wide significance of 2.6e–6). We estimate the power of trio sequencing to detect statistically significant associations for disease genes, under the assumption that 10% or 20% of singleton mutations could be fully penetrant genes (vertical black bar 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.

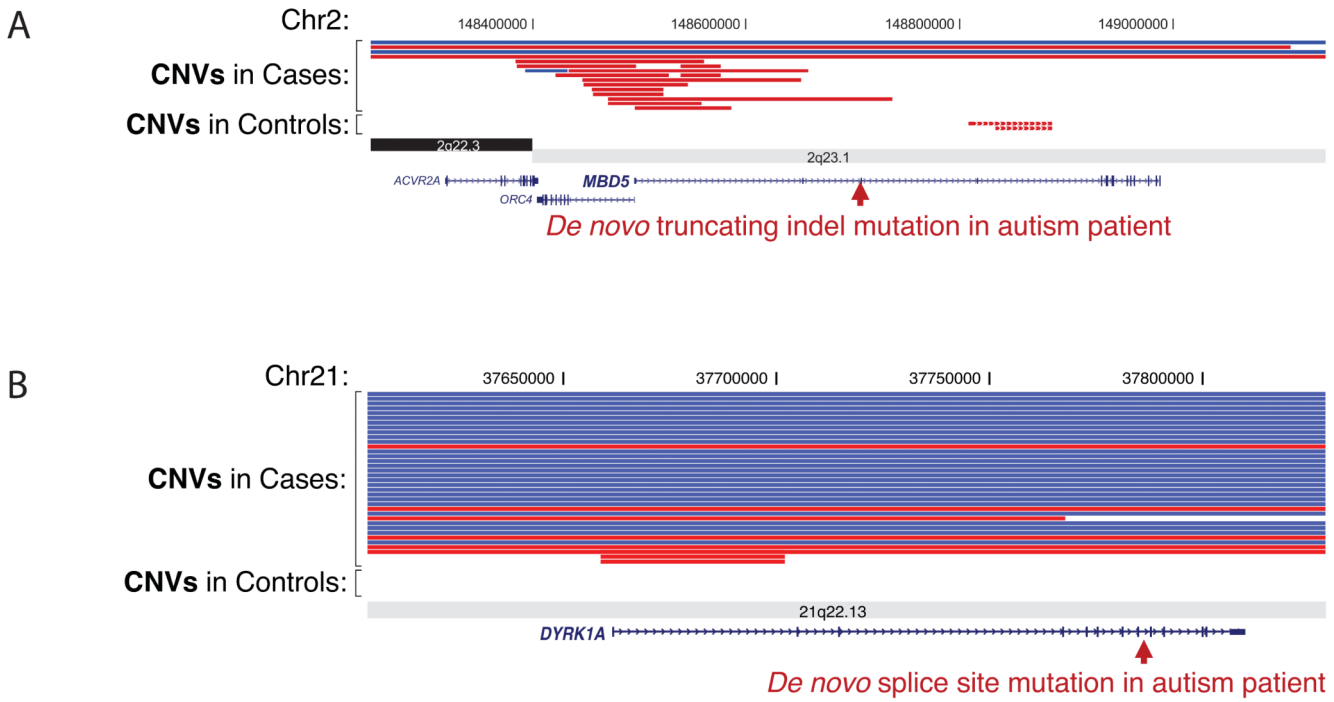

Figure 2. CNV and exome intersections define candidate genes

Deletion (red) and duplication (blue) burden for DD/ID cases and controls for two genes **(A)** *DYRK1A* and **(B)** *MBD5* as compared to sporadic LoF mutations based on 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 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].

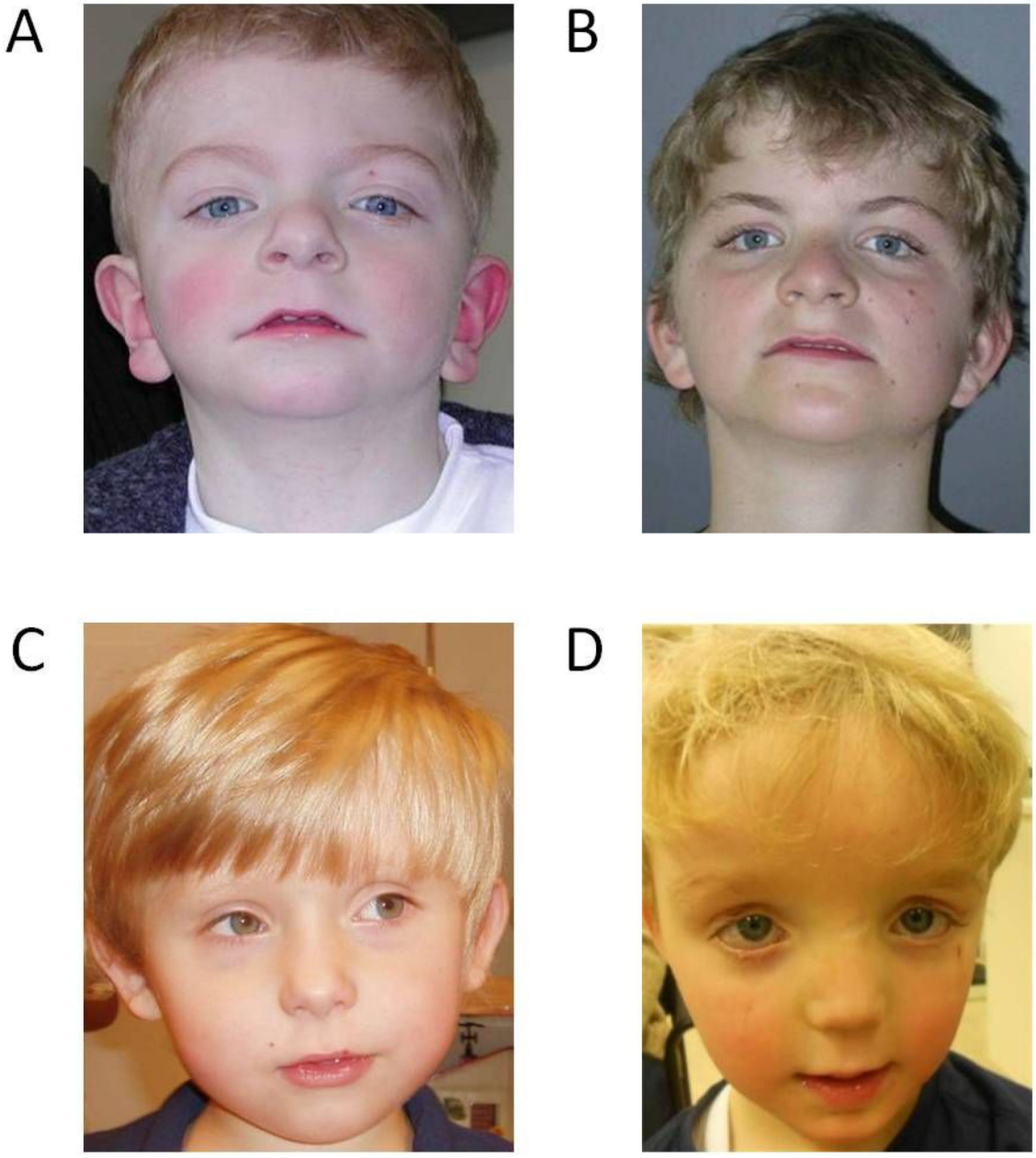

Figure 3. Phenotypic similarity of two patients with identical PACS1 *de novo* mutations and two patients with similar *ADNP* mutations

(A and **B)** These two unrelated patients show identical *de novo* point mutations (c.607C>T; p.(Arg203Trp)) mutation in *PACS1* (RefSeq NM_018026.2)[53]. The striking similarity in clinical phenotype include 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. **(C** and **D)** 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.

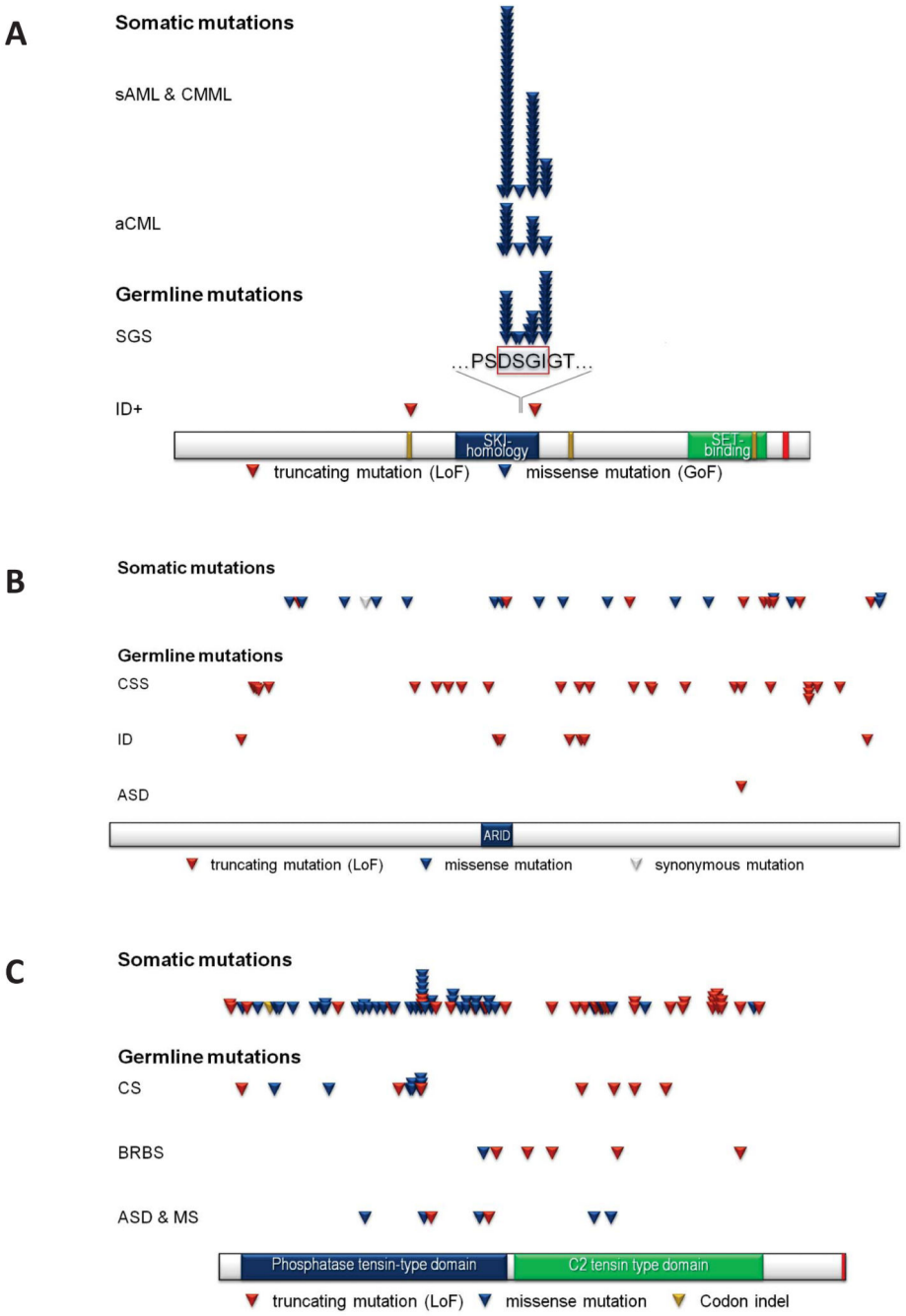

Figure 4. Coincidental *de novo* mutations in cancer and neurodevelopmental disorders

Examples: *SETBP1* (Figure 2A), *ARID1B* (Figure 2B) and *PTEN* (Figure 2C).

(A) Mutation spectrum of *SETBP1*. sAML & CMML = secondary acute myeloid leukemia & chronic myelomonocytic leukemia[93] [1× p.Asp868Tyr, 28× p.Asp868Asn, 1× p.Ser869Asn, 15× p.Gly870Ser, 5× p.Ile871Thr]; aCML = atypical chronic myeloid leukemia[94] [7× p.Asp868Asn, 1× p.Ser869Gly, 5× p.Gly870Ser, 2× p.Ile871Thr]; SGS = Schinzel-Giedion syndrome[32] [Hoischen *et al. unpublished*: 1× p.Asp868Ala, 7× p.Asp868Asn, 1× p.Ser869Arg, 1× p.Ser869Asn, 4× p.GGly870Ser, 2× p.Gly870Asn, 10×

p.Ile871Thr]; ID+ = intellectual disability with other features[2,3] [p.Leu592* & p. 906fs]. **(B)** Mutation spectrum of *ARID1B*. Somatic mutations retrieve from COSMIC database. Only 'somatic validated' and 'previously described' somatic mutations with 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 = intellectual disability[54] [p.Arg372Profs*163, p.Arg1102*, p.Lys1108Argfs*9, p.Gln1307*, p.Tyr1346*, p.Arg1338Argfs*76, p.Ser2155Leufs*33]; ASD = autism spectrum disorder[3] [p.Phe1798Leufs*52]; splice site mutations not considered. **(C)** Mutation spectrum of *PTEN*. Somatic mutations retrieved from COSMIC database. Only 'somatic validated' and 'previously described' somatic mutations with at least five independent entries are displayed. CS = Cowden syndrome; ASD & MS = autism spectrum disorder and macrocephaly syndrome; BRBS = Bannayan-Riley-Ruvalcana syndrome [based on OMIM entries]; splice site mutations not considered.

Table 1

A summary of major (exome) sequencing studies (n = 11). Summary of *de novo* mutation discovery, including: size of study, number of *de novo* mutations and severity for each group.

					Number of coding and splice site *de novo* point mutations in probands						Number of coding and splice site *de novo* point mutations in controls/siblings				
Study	**Phenotype**	**Number of patients**	**Number of controls (c)/ siblings (s)**	**Total**	**LoF**	**Codon indels (in-frame)**	**Missense**	**Synonymous**	**Nonsynonymous/synonymous ratio**	**Total**	**LoF**	**Codon indels (in-frame)**	**Missense**	**Synonymous**	**Nonsynonymous/synonymous ratio**
Rauch *et al*.	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.*	ID	100	0	79	15	0	48	16	3.9 (63/16)	NA	NA	NA	NA	NA	NA
O'Roak *et al.*	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.* *	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*.	ASD	175	0	169	18	0	101	50	2.4 (119/50)	NA	NA	NA	NA	NA	NA
Iossifov *et al.***	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*.	ASD	32	0	42	5	0	28	9	3.7 (33/9)	NA	NA	NA	NA	NA	NA
Allen *et al*.	EE	264	0	289	36	1	196	56	4.2 (233/56)	NA	NA	NA	NA	NA	NA
Gulsuner *et al.*	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.****	SCZ	231	34	164	20	4	115	25	5.6 (139/25)	17	1	0	11	5	2.4 (12/5)
Fromer *et al*.	SCZ	623	0	640	65	9	410	156	3.1 (484/156)	NA	NA	NA	NA	NA	NA
Sum:		**2358**	**731**	**2368**	**306 (12.9%)**	**24 (1.0%)**	**1501 (63.4%)**	**537 (22.7%)**	**3.4 (1831/537)**	**600**	**53 (8.8%)**	**8 (1.3%)**	**381 (63.5%)**	**158 (26.3%)**	**2.8 (442/158)**

Putative LoF includes nonsense, frameshift and canonical splice site mutations based on gene annotation

* study did not consider indels

** not all *de novo* mutations were validated by Sanger sequencing

*** excluded splice site variants if not in canonical di-nucleotide of splice site. 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 re-annotated using Seattle-Seq. When considering all studies together the total amount of *de novo* mutations identified in patients vs. controls is different (Fisher's exact test: 0.0012), this may however rather reflect the technical differences between the individual studies. The most significant difference is seen for the number of LoF *de*

novo mutations is significantly higher in patients than in controls (Fisher's exact test: 0.0062); similarly the total amount of nonsynonymous vs. synonymous variants is higher in cases than in controls, however this does not reach significance (Fisher's exact test: 0.066).

Table 2

Recurrent and overlapping genes—*de novo* mutations in same genes observed between ID, ASD, EE and SCZ. Genes reported with recurrent (≥4) nonsynonymous *de novo* mutations identified in eleven studies on four different neurodevelopmental phenotypes.

Gene	Total observations	Mutation type		Neurodevelopmental disorder			
		LoF (nonsense, fs, splice)	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

TTN and *MUC5B* were excluded from this table due to high variant load in controls and unlikely involvement in the phenotypes discussed.

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*.	ASD
SCN2A	frame shift	Chr2(GRCh37):g.166172105dup	NM_021007.2:c.1508dup	p.Asn503Lysfs*19	Rauch *et al.*	ID
SCN2A	frameshift	Chr2(GRCh37):g.166179825_166179826del	NM_021007.2:c.1831_1832del	p.Leu611Valfs*35	Rauch *et al.*	ID
SCN2A	missense	Chr2 (GRCh3 7):g.166198975G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al*.	EPI
SCN2A	missense	Chr2(GRCh37):g.166198975G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al*.	EPI
SCN2A	missense	Chr2(GRCh37):g.166201311C>T	NM_021007.2:c.2809C>T	p.Arg937Cys	Rauch *et al.*	ID
SCN2A	nonsense	Chr2(GRCh3 7):g.166201379C>A	NM_021007.2:c.2877C>A	p.Cys959*	Sanders *et al.*	ASD
SCN2A	nonsense	Chr2(GRCh3 7):g.166210819G>T	NM_021007.2:c.3037G>T	p.Gly1013*	Sanders *et al.*	ASD
SCN2A	nonsense	Chr2(GRCh37):g.166231415G>A	NM_021007.2:c.4193G>A	p.Trp1398*	de Ligt *et al.*	ID
SCN2A	missense	Chr2(GRCh37):166234111C>T	NM_021007.2:c.4259C>T	p.Thr1420Met	Iossifov *et al*.	ASD
SCN2A	splice site	Chr2(GRCh37):g.166187838A>G	NM_001040142.1:c.2150–2A>G	p.?	Fromer *et al.*	SCZ

Table 4

Recurrent identical *de novo* mutations in six genes identified in eleven published 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*.	ID
ALG13	Missense	ChrX(GRCh37):g.110928268A>G	NM_001099922.2:c.320A>G	p.Asn107Ser	Allen *et al.*	EE
ALG13	Missense	ChrX(GRCh37):g.110928268A>G	NM_001099922.2:c.320A>G	p.Asn107Ser	Allen *et al.*	EE
KCNQ3	Missense	Chr 8(GRCh3 7):g.133192493 G> A	NM_001204824.1:c.328C>T	p.Arg110Cys	Rauch *et al.*	ID
KCNQ3	Missense	Chr 8(GRCh3 7):g.133192493 G> A	NM_001204824.1:c.328C>T	p.Arg110Cys	Allen *et al.*	EE
SCN1A	splice donor	LRG_8:g.24003G>A	NM_006920.4:c.602+1G>A	p.?	Allen *et al.*	EE
SCN1A	splice donor	LRG_8:g.24003G>A	NM_006920.4:c.602+1G>A	p.?	Allen *et al.*	EE
CUX2	Missense	Chr 12(GRCh3 7):g.111748354G> A	NM_015267.3:c.1768G>A	p.Glu590Lys	Rauch *et al.*	ID
CUX2	Missense	Chr 12 (GRCh3 7):g.111748354G>A	NM_015267.3:c.1768G>A	p.Glu590Lys	Allen *et al.*	EE
SCN2A	Missense	Chr2(GRCh37):g.166198975 G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al.*	EE
SCN2A	Missense	Chr2(GRCh37):g.166198975 G>A	NM_021007.2:c.2558G>A	p.Arg853Gln	Allen *et al.*	EE
DUSP15	Missense	Chr20(GRCh37):g.30450489G>A	NM_080611.2:c.320C>T	p.Thr107Met	Neale *et al.*	ASD
DUSP15	Missense	Chr20(GRCh37):g.30450489G>A	NM_080611.2:c.320C>T	p.Thr107Met	Fromer *et al.*	SCZ

Table 5

Genes implicated in cancer (somatic mutations) and developmental disorders (germline mutations)

Gene name	Neurodevelopmental disorder	PubMed ID	Mutation types	Cancer types/malignancies	PubMed ID	Mutation types
Isolated neurodevelopmental phenotypes and cancer						
PTEN	ASD, Cowden S.*	23160955, 9259288	LoF, missense	Many types	21252315	missense, LoF
CTCF	ID	23746550	LoF	Breast cancer, Prostate cancer	9591631	deletion
ARID1B (SWI/SNF complex)	ID, ASD, Coffin-Siris S.	22426309, 22426308	LoF	Many types	22426308	LoF
Clinically defined neurodevelopmental syndromes and cancer						
MED12	Lujan-Fryns S., Ohdo S., Opitz-Kaveggia S.	17369503, 23395478,	missense	Prostate cancer	22610119	missense
MLL2 (KMT2D)	Kabuki S.	20711175	LoF, CNV, missense	Many types	21796119, 21163964	LoF, missense
CREBBP	Rubinstein-Taybi S.; epilepsy	7630403, 11331617	LoF, CNV, missense	ALL	21390130	LoF, missense
ATRX	X-linked α-thalassaemia/mental retardation S.	7697714	Missense, LoF	PanNET, Glioblastoma	21252315, 22286061	missense, LoF
Identical mutations in neurodevelopmental phenotypes and cancer						
ASXL1	Bohring Opitz S.	21706002	LoF	Myeloid malignancies	19388938	LoF
SETBP1	Schinzel Giedion S.	20436468	GoF	Leukemia	23222956	GoF
EZH2	Weaver S.*	22177091	GoF	B-cell lymphoma	20081860	GoF
'Paternal age effect disorders' (FGFR2, FGFR3, HRAS, PTPN11, BRAF, MAP2K1)	Apert S., Crouzon/Pfeiffer S., achondroplasia, Muenke S., Costello S.*, Noonan S.*, cardio-facio-cutaneous S.	22325359	GoF	Many types	22325359	GoF
Different mutation types observed in neurodevelopmental phenotypes and cancer						
CTNNB1	ID/ASD	23033978	LoF	Many types	12060769	activating?, LoF? duplications, CNV,
CHD7	CHARGE S.	15300250	CNVs, LoF, missense	Small-cell lung cancer	20016488	amplifications, translocations
MYCN	Feingold S.	15821734	LoF, missense	Neuroblastoma	6197179	amplification
ABCC9	Cantu S.	22610116, 22608503	GoF	Endometrial cancer	23104009	missense

Abbreviations: PanNET – Pancreatic neuroendocrine tumors; ALL – acute lymphoblastic leukemia; AML – Acute myeloid leukemia; S. – Syndrome

* Patients have increased risk of cancer;