APPLICATIONS OF NEXT-GENERATION SEQUENCING

De novo mutations in human genetic disease

Joris A. Veltman and Han G. Brunner

Abstract | New mutations have long been known to cause genetic disease, but their true contribution to the disease burden can only now be determined using family-based whole-genome or whole-exome sequencing approaches. In this Review we discuss recent findings suggesting that *de novo* mutations play a prominent part in rare and common forms of neurodevelopmental diseases, including intellectual disability, autism and schizophrenia. *De novo* mutations provide a mechanism by which early-onset reproductively lethal diseases remain frequent in the population. These mutations, although individually rare, may capture a significant part of the heritability for complex genetic diseases that is not detectable by genome-wide association studies.

Single-nucleotide variants
Differences in the nucleotide composition at single positions in the DNA sequence. The most common form of variation in the human genome.

Indels
Small insertions or deletions of 1–1,000 nucleotides.

Copy number variants
Large insertions or deletions of more than 1,000 nucleotides.

Schinzel–Giedion syndrome
A rare genetic disorder that is characterized by congenital hydronephrosis, skeletal dysplasia and severe developmental retardation.

Department of Human Genetics, Nijmegen Centre for Molecular Life Sciences, Institute for Genetic and Metabolic disease, Radboud University Nijmegen Medical Center, PO Box 9101, Nijmegen, The Netherlands.
Correspondence to J.A.V.
e-mail: j.veltman@gen.umcn.nl

doi:10.1038/nrg3241

Over the past few decades, research in the field of medical genetics of disease has focused largely on inherited variation. This has resulted in great progress, through the application of family-based linkage studies in the case of Mendelian diseases and through genome-wide association studies for complex diseases. However, neither of these approaches is suitable for the study of genetic diseases that are caused by *de novo* mutations. Now, genomic microarrays and next-generation sequencing technologies enable us to overcome the limitations of traditional approaches to genetic disease research. With the advent of unbiased whole-genome and whole-exome sequencing approaches, we can now study, at single-nucleotide resolution, the mutational processes that occur in humans from generation to generation and from cell to cell[1,2]. The results provide basic insight into the mutational processes in humans and their impact on disease. As an example, family-based whole-genome sequencing studies have shown that, on average, 74 germline single-nucleotide variants (SNVs) occur *de novo* in an individual's genome[2], a number that is remarkably close to estimates from the pre-genome-sequencing era[3,4]. However, considerable technological improvements are required to reliably detect indels (small insertions or deletions) as well as larger copy number variants (CNVs), and therefore much less is known about the timing and frequency of these variants.

De novo mutations represent the most extreme form of rare genetic variation: they are more deleterious, on average, than inherited variation because they have been subjected to less stringent evolutionary selection[5,6]. This makes these mutations prime candidates for causing genetic diseases that occur sporadically. Indeed, recent whole-exome sequencing studies have revealed *de novo* germline SNVs in single genes as the major cause of rare sporadic malformation syndromes such as Schinzel–Giedion syndrome[7], Kabuki syndrome[8] and Bohring–Opitz syndrome[9].

As a result of these and similar studies, several features of *de novo* mutation have emerged. The overall rate of *de novo* germline mutations may be higher in individuals with genetic disease than in those without, and there seems to be a increased mutational load associated with higher paternal age. The phenotypic consequences of *de novo* mutations arise because these mutations affect specific genes and nucleotides. Mutations causing severe genetic diseases are often highly disruptive to gene function and tend to affect important domains of developmental genes. An open question is whether these mutations occur mainly in the germline, during embryogenesis or somatically. A number of studies have shown an apparent germline origin of mutations. In addition, exome sequencing of affected and unaffected tissues has recently revealed *de novo* somatic SNVs as the cause of overgrowth syndromes such as Proteus syndrome[10].

Because *de novo* mutations are not rare events collectively, it is possible that they are responsible for an important fraction of more commonly occurring diseases through disruption of any one of a large number of genes. Several pilot studies recently revealed that *de novo* mutations affecting many different genes in different individuals together might explain a proportion of common neurodevelopmental diseases such as intellectual disability (ID)[11], autistic-spectrum disorders (ASDs)[12–16] and schizophrenia[17,18]. This *de novo* model of complex neurodevelopmental genetic diseases essentially points to a monogenic basis of disease, with the mutation representing a single event of large effect. This contrasts with the multifactorial model, which invokes the interplay of many genetic and non-genetic factors of small effect in any individual patient. Thus, although it needs to be acknowledged that the phenotypic effect of any single mutation depends on the genetic background in which it occurs, the current overall picture is decidedly more monogenic than that envisaged just a few years ago[19]. However, it is of course possible that both models apply. The realization that *de novo* mutations are potentially important in complex genetic diseases has major implications for our thinking about the causes, mechanisms and preventive strategies for these diseases[20].

In this Review we highlight the insights obtained from recent studies on *de novo* mutations in humans and discuss the impact of this work for genetic disease studies in general, as well as for counselling individual families with sporadic disease. We discuss the risk factors that affect the mutation rate, such as increased paternal age, and evaluate methods for the improved prediction of the phenotypic consequences of *de novo* mutations. We mainly focus on the role of *de novo* germline mutations in sporadic genetic disorders and do not discuss somatic *de novo* mutations in cancer (for coverage of this topic, see recent reviews[21,22]). Before we discuss the impact of *de novo* mutations on human genetic disease, we summarize our current knowledge of the germline mutation rates in humans (see recent reviews[23,24]).

Germline mutation rates in the human population

Human germline mutations can range from alterations in the number of chromosomes down to mutations in single base pairs. Because germline mutations are so rare given the size of our genome, it has been stated that measuring the human per-generation mutation frequency is like measuring the frequency of needles in haystacks[25]. The rate at which these mutations occur differs for each class of mutation; most *de novo* mutations are SNVs, but considerable genomic variation also occurs at the level of indels and larger CNVs.

The rate of* de novo *SNVs. Considerable knowledge has been acquired about the rate of occurrence of SNVs. The studies that investigate these mutations were initially based on single genes[4,26,27], but more recently have been carried out at the level of entire genomes[2,28]. The current best estimate of the average human germline SNV mutation rate is 1.18×10^{-8} per position[2], which corresponds to ~74 novel SNVs per genome per generation. This mutation rate is remarkably close to estimates based on extrapolations from single-gene studies[4]. The mutation rate is known to vary considerably between nucleotide sites, depending on both the genomic location and the local sequence context. In particular, the rate of SNVs is elevated by an order of magnitude at CpG sites[2,4,28]. The availability of whole-genome sequencing data from parent–offspring trios allows us to look for variation in the mutation rate between individuals and to determine the parental origin of these mutations. The first such information was recently provided[2] from two apparently healthy families participating in the 1000 Genomes project. Remarkably, 92% of the 49 *de novo* SNVs identified in one family were from the paternal germline, whereas in the other family only 36% of the 35 mutations detected were paternal in origin. This is an intriguing result, as it indicates that individual mutation rates might vary considerably. Determining the true extent of variation in mutation rates between individuals will need much larger studies.

Indel and CNV mutation rates. The estimated mutation rates for indels and CNVs have not been established with as much confidence as the rate for SNVs, owing to complexities in the reliable identification of both forms of genomic variation. Both CNVs and indels seem to occur at much lower frequencies than SNVs, but owing to their larger size they collectively affect more base pairs. The indel mutation rate has been estimated to be approximately 4×10^{-10} per position, resulting in about three novel indels per genome per generation[29]. Small deletions are approximately three times as common as small insertions, and for both types of variation the mutation frequency declines with increasing fragment size. CNVs larger than 100 kb are estimated to occur *de novo* in approximately one out of every 50 individuals[30]; for CNVs smaller than 100 kb, no reliable numbers exist. Importantly, all of these rates are strongly influenced by factors such as parental sex, age and ethnicity (BOX 1). In addition, the presence of genetic risk factors — such as inversions, duplications, translocations and mutations in genes affecting DNA repair or recombination — may increase these mutation rates in certain individuals (BOX 1). Of note, estimates of *de novo* mutation rates have been based mostly on investigations in healthy parent–offspring trios, and these healthy individuals have been subjected to substantial prenatal selection. The true mutation rate is likely to be much higher if we could include all deleterious mutations and all stages of development.

De novo mutations in rare sporadic genetic disease

We are witnessing a rapid change in the emphasis of research into *de novo* mutations, from cytogenetically visible *de novo* chromosomal abnormalities via *de novo* CNVs to *de novo* SNVs. This change has been driven by the emergence of microarrays as a new and powerful technology at the turn of the century, followed by next-generation sequencing over the past 6 years. Consequently, the focus has changed from maternal age as the predominant risk factor for aneuploidies to paternal age for *de novo* CNVs and SNVs. Finally, the application of large-scale sequencing demonstrates that many previously enigmatic sporadic syndromes, malformations and diseases are due to *de novo* germline gene mutations, whereas others reflect somatic mosaicism for gene mutations.

From chromosomal to point mutations. The field of medical genetics traditionally emphasized the study of inherited forms of disease, both recessive and dominant.

Kabuki syndrome
A rare genetic condition that is characterized by distinctive facial features, skeletal abnormalities and intellectual disabilities.

Bohring–Opitz syndrome
A rare genetic disorder that is characterized by facial anomalies, multiple malformations, failure to thrive and severe intellectual disabilities.

Proteus syndrome
A rare syndrome that is characterized by patchy or mosaic overgrowth and hyperplasia of various tissues and organs.

CpG sites
Genomic regions of several hundred base pairs with a high GC content and many unmethylated CpG dinucleotides.

Somatic mosaicism
The presence of mutations in a proportion of the cells in the body but not in sperm and egg cells.

Box 1 | General factors that influence *de novo* mutation rates

De novo mutation frequencies vary between individuals and over time within an individual. Germline mutations show strong parent-of-origin biases as well as parental-age effects. The extra chromosome 21 in Down syndrome is mostly of maternal origin and occurs more frequently with increased maternal age[86]. On the other hand, *de novo* SNVs occur at higher rates in males than in females, and this difference increases with paternal age[5]. This male bias can be explained by the greater number of cell divisions in the male germline (compared with the female germline), during which replication mistakes can occur. Some gene mutations, however, show a paternal-age effect that is much stronger than expected and is driven by the mutation conferring a selective advantage during spermatogenesis, leading to clonal expansion in the testis[23,87,88]. The risk of passing on a rare monogenic condition such as achondroplasia, Apert syndrome, Crouzon syndrome and multiple endocrine neoplasia type 2 is increased by this mechanism by about tenfold in fathers older than 50 years. The collective burden for children born to older fathers may be considerable, and a large proportion of this extra risk is due to *de novo* mutations[89].

The *de novo* rate of CNVs is particularly sensitive to the local genomic architecture and to parent-of-origin effects. *De novo* CNVs linked to intellectual disability (ID) were recently found to be mostly of paternal origin and to be associated with increased paternal age. This was particularly evident for non-recurrent CNVs that arose by replication based mechanisms such as non-homologous end joining or microhomology-mediated break-induced repair[90]. By contrast, non-allelic homologous recombination (NAHR) mediated by segmental duplications can result in *de novo* CNVs during meiosis[32]. No parent-of-origin or parental-age effect has been demonstrated for this class of CNVs, which occur relatively frequently and recur because of the predisposing chromosomal architecture[90]. The number, location and orientation of these segmental duplications varies considerably between individuals, and this affects the risk of NAHR-mediated CNV generation. An instructive example of this is chromosome 17q21.31 microdeletion syndrome[91–93]. Each parent in whom the *de novo* 424 kb deletion originates carries a germline 900-kb chromosome 17q21.31 inversion polymorphism encompassing the deleted region[94]. Breakpoint sequencing showed that this inversion contains specific segmental duplications that are necessary for NAHR to occur[95]. Interestingly, this inversion is present in 20% of Europeans but is rare in other populations[96]. Thus, the likelihood of this as well as other genomic rearrangements may vary considerably between ethnic groups.

The mechanisms by which individual variation in germline mutation frequency arises remain largely to be elucidated. Some of this variability may be due to variation in specific genes such as that encoding PR domain-containing protein 9 (*PRDM9*). *PRDM9* is involved in mediating homologous recombination, and variation in this gene influences the use of meiotic recombination hot spots[97,98]. Allelic variation at the *PRDM9* locus affects the germline *de novo* mutation frequency at highly unstable minisatellites and at unstable genomic regions flanked by segmental duplications. This process affects, for example, the frequency of *de novo* CNVs at chromosome 17p11.2, causing Charcot–Marie–Tooth disease type 1a and hereditary liability to pressure palsies[99]. In addition, for trinucleotide repeat mutations occurring in myotonic dystrophy type 1, there is evidence to implicate genetic variation in DNA replication, repair and recombination in repeat expansion or contraction[100]. Complex mutational events that affect multiple independent chromosomal regions[101] represent examples of germline hypermutability for which a genetic mechanism remains to be uncovered. Systematic analysis of individuals with an increased occurrence of *de novo* mutations, and analysis of their parents, is essential to identify the genetic factors that influence the occurrence of *de novo* mutations[102].

Nonetheless, it is common knowledge that dominant *de novo* mutations are important and can cause rare genetic disease. A well-known example is Down syndrome, which is caused by a *de novo* trisomy of chromosome 21 (REF. 31). However, most sporadic diseases are not caused by microscopically visible chromosomal abnormalities, and the identification of their genetic cause had remained a major challenge. In fact, for many of these disorders, it long remained unclear whether there is a genetic cause at all.

Over the past decade, genomic microarrays have uncovered structural genomic variation in healthy people, which came as a great surprise and raised a question as to how much of this variability is due to mutation. Subsequent studies have shown that *de novo* CNVs can occur all over the genome and that they occur at higher frequency in individuals with a neurodevelopmental disorder than in individuals without such a disorder. Recurrent *de novo* microdeletions and microduplications are now recognized as a common cause of clinically defined malformation syndromes[32,33]. Several structural features of the genome have been recognized to increase the likelihood of *de novo* CNV generation at specific sites (BOX 1).

The use of microarrays has also allowed the identification of specific genes that underlie sporadic malformation syndromes. A *de novo* CNV at chromosome 8q12 led to the discovery of the chromodomain helicase DNA-binding protein 7 gene (*CHD7*) as the causal gene for the mostly sporadic CHARGE syndrome[34]. Since this discovery in 2004, *de novo* CNVs have been found to underlie several other monogenic sporadic diseases. However, for most patients with rare genetic diseases, the precise genetic cause remains to be defined. Unbiased whole-exome and whole-genome sequencing studies of patients and their unaffected parents now allow rapid screening for these *de novo* SNVs, although considerable technological and methodological challenges remain (BOX 2).

Exome sequencing is revolutionizing the detection of* de novo *mutations. Exome and genome sequencing have greatly facilitated the detection of *de novo* SNVs in rare genetic disease[35,36]. As a first example, whole-exome sequencing allowed the detection of *de novo* mutations in the gene encoding SET-binding protein 1 (*SETBP1*) in 12 out of 13 patients with Schinzel–Giedion syndrome[7]. Other recent successes of exome sequencing include the identification of *de novo* mutations in the mixed-lineage leukaemia 2 (*MLL2*) gene as a major cause of Kabuki syndrome[8], in the additional sex combs-like 1 (*ASXL1*) gene as a major cause of Bohring–Opitz syndrome[9] and in the ankyrin repeat domain 11 gene (*ANKRD11*) as a cause of KBG syndrome[37]. Because the intellectual disability associated with KBG syndrome can be mild, occasional transmission in families is possible. Indeed, a family with an inherited *ANKRD11* mutation was also identified in this study[37]. This last example illustrates the reciprocal relationship between the fitness effect of a mutation and the proportion of *de novo* mutations that is observed in a dominantly inherited disorder[22]. For reproductively lethal diseases, the frequency with which the disease occurs in the population is proportional to the chance of pathogenic *de novo* mutations affecting the causative gene. This in turn is largely determined by the size of the mutational target — that is, the cumulative size of the gene loci in which the pathogenic *de novo* mutations cluster. Note that this target can be a very small part of a single gene, as is the case for Schinzel–Giedion syndrome, in which all mutations occur in a stretch of just 11 nucleotides of the *SETBP1* gene[7]. By contrast, *de novo* mutations in the two genes actin beta (*ACTB*) and actin gamma 1 (*ACTG1*) can result

Achondroplasia
A common form of dwarfism that is inherited in an autosomal dominant manner.

Apert syndrome
An autosomal dominant disorder that is characterized by premature closing of cranial sutures and by fused fingers and toes.

Box 2 | Challenges in the detection of *de novo* mutations

Next-generation DNA sequencing technologies allow us to study the genome-wide frequency and distribution of *de novo* mutations in an unbiased manner. However, the study of *de novo* mutations poses specific challenges that need to be taken into account.

Focusing on *de novo* mutations enriches for sequencing artefacts
As no sequencing technology is error-proof, the number of false-positive and false-negative variants increases with the size of the sequenced target (from gene to exome to genome). This is nicely illustrated in a recent comparative study in which a single genome was sequenced at high coverage (~150-fold) by two different sequencing platforms[103]. The concordance rate between these platforms was low for variation calling (88% for single-nucleotide variants (SNVs) and only 28% for indels (small insertions or deletions)), totalling more than half a million different calls. Sequencing artefacts are especially problematic for the detection of *de novo* mutations, as false-positive variants will appear as potential *de novo* mutations when they are observed in a child's genome or exome but not in the parental genomes. By the same token, a false-negative call in a parent may result in a *de novo* mutation being called in a child's genome or exome. The first family-based genome sequencing study[28] indeed identified thousands of such false-positive and false-negative candidate *de novo* SNVs for each true *de novo* SNV. Of note, SNVs are more reliably detected by next-generation sequencing technology than indels and copy number variants (CNVs). Finally, reliably detecting *de novo* somatic mutations is more complex than calling *de novo* germline mutations, because somatic mutations will vary between tissue types and may appear in percentages that are similar to current false-positive sequencing rates.

***De novo* mutations are induced during cell line creation and culturing**
Many genetic studies, such as the 1000 Genomes project, are carried out on DNA derived from lymphoblastoid cell lines (LCLs). The creation of these lines and subsequent cell culturing are known to introduce genomic changes that appear as *de novo* mutations when sequences derived from such cell lines are compared between parents and offspring. As part of the 1000 Genomes project, the genomes of two parent–offspring trios were sequenced using LCL-derived DNA[2]. In one of the trios, the authors identified and validated 643 *de novo* mutations in cell line DNA that were not observed in DNA derived from uncultured blood of these individuals. By contrast, only 35 *de novo* mutations were observed in both LCL and blood-derived DNA, demonstrating that the majority of these potential *de novo* mutations were in fact caused by cell line transformation and culturing, a finding that was recently confirmed[104]. The use of cell lines is therefore not recommended for *de novo* mutation studies, and independent validation on DNA from uncultured sources is essential. Of note, *de novo* CNVs are also generated at considerable frequencies during the reprogramming of somatic cells into induced pluripotent stem cells[105] and during the establishment and growth of embryonic stem cells[106].

Limited availability of parental samples in adult-onset diseases
The study of *de novo* mutations is limited by the availability of DNA from parent–offspring trios. This will be significantly more difficult to obtain for adult-onset diseases and requires an ongoing international collaborative effort to set up biobanks containing DNA and phenotypic information from multiple generations[107].

Crouzon syndrome
A rare genetic disorder that is characterized by premature fusion of the skull bones (craniosynostosis).

Multiple endocrine neoplasia type 2
Early-childhood thyroid cancer caused by mutations in the proto-oncogene *RET* that are inherited in an autosomal dominant manner.

Charcot–Marie–Tooth disease type 1a
A rare genetic neurological disorder that affects the peripheral nerves.

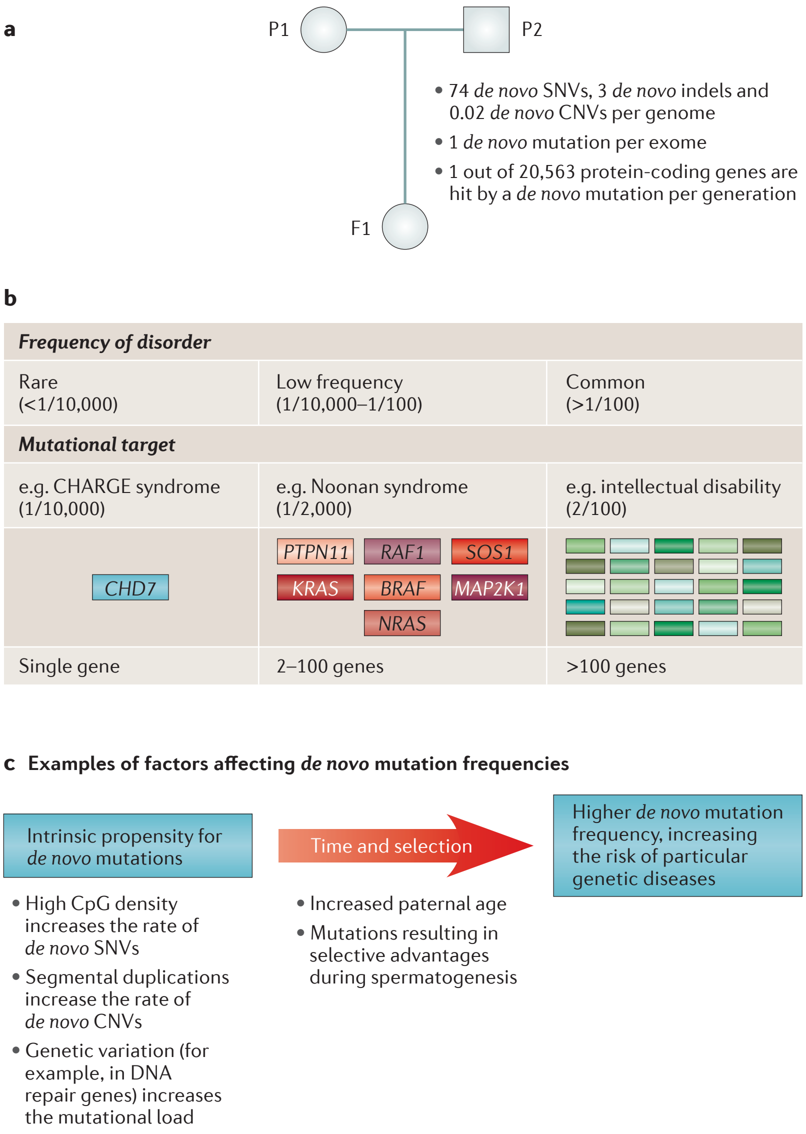
in Baraitser–Winter syndrome[38], and *de novo* mutations in each of six genes encoding SWI/SNF subunits were recently reported to cause Coffin–Siris syndrome[39,40]. FIGURE 1 illustrates the link between the mutational target size and the disease frequency in the human population. It is clear that diseases which are caused exclusively by *de novo* mutations in a single gene will occur at low population frequencies, whereas diseases that are caused by *de novo* mutations in any one of many different genes in different individuals could reach much higher population frequencies (see BOX 3 for a discussion of the burden of disease that is caused by *de novo* mutations).

From germline to somatic mutations. Somatic mosaicism for single-gene mutations is now increasingly documented in sporadic conditions of the skin, skeleton and blood vessels[10,41–43]. Ongoing mutational mechanisms in later life have also been documented; for example, spontaneous correction of genetic defects has been reported for some diseases of the skin[44]. Theoretical models predict that, in a typical individual, every gene mutates somatically many times. From this, one can predict that most cells in the body carry at least one somatic *de novo* mutation[45]. Documentation of such widespread mosaicism is still lacking and requires next-generation sequencing of single cells[46]. At the chromosome level, evidence for widespread somatic mutational events is already available for early human development. A striking conclusion from a number of recent studies is that at the cleavage and blastocyst stages most human embryos are mosaics of diploid and aneuploid cells[47,48]. Therefore, the generally diploid state of newborns must reflect selection rather than the absence of mutation.

Recently, exome sequencing was shown to also be useful in the detection of somatic mosaicism as a cause of rare sporadic disease, as the technique was used to identify the cause of Proteus syndrome[10]. Proteus syndrome does not recur in families, but it has been reported in discordant monozygotic twins, supporting the hypothesis that it is caused by somatic mutations which are lethal when present in all tissues. To find the genetic cause of this disorder, material from affected tissue with visible signs of overgrowth or vascular anomaly was biopsied and subjected to exome sequencing, and the resultant exome was compared to that of tissue without signs of overgrowth or vascular anomalies but from the same patient[10]. The authors detected a mutation in v-*akt* murine thymoma viral oncogene homologue 1 (*AKT1*) in one patient, and the predicted result of this mutation was a substitution of lysine for glutamine at amino acid 17. Affected tissues and cell lines from 25 other patients with Proteus syndrome carried this same mutation in 1–47% of alleles, demonstrating that the mutation is indeed of somatic origin. Importantly, Sanger sequencing showed that DNA from peripheral blood cells of these patients was negative for the *AKT1* mutation in all cases, strongly indicating that a genetic diagnosis should be carried out on biopsies from affected tissue. This study also shows that the detection of somatic mutations requires exome sequencing at a greater depth of coverage (>100-fold) than is required for the detection of germline mutations (~50-fold), and also needs researchers to follow up on more variants that occur in a small percentage of cells. It is important to note that this is likely to increase the number of false-positive variants (see also BOX 2).

De novo mutations in common genetic disease

CNV studies. Although the role of *de novo* mutations has been well established in rare genetic disease, this is not the case for more common genetic disorders, with the exception of *de novo* CNVs in neurodevelopmental disorders. The cytogenetics community has recognized the importance of *de novo* chromosomal abnormalities for many decades, and parental analysis is an important part of the procedure to substantiate or exclude


a
P1
P2
F1
• 74 de novo SNVs, 3 de novo indels and 0.02 de novo CNVs per genome
• 1 de novo mutation per exome
• 1 out of 20,563 protein-coding genes are hit by a de novo mutation per generation
b
c Examples of factors affecting de novo mutation frequencies
Intrinsic propensity for de novo mutations
Time and selection
Higher de novo mutation frequency, increasing the risk of particular genetic diseases
• High CpG density increases the rate of de novo SNVs
• Segmental duplications increase the rate of de novo CNVs
• Genetic variation (for example, in DNA repair genes) increases the mutational load
• Increased paternal age
• Mutations resulting in selective advantages during spermatogenesis

Frequency of disorder		
Rare (<1/10,000)	Low frequency (1/10,000–1/100)	Common (>1/100)
Mutational target		
e.g. CHARGE syndrome (1/10,000)	e.g. Noonan syndrome (1/2,000)	e.g. intellectual disability (2/100)
CHD7	*PTPN11* *RAF1* *SOS1* *KRAS* *BRAF* *MAP2K1* *NRAS*	🟩🟩🟩🟩🟩 🟩🟩🟩🟩🟩 🟩🟩🟩🟩🟩 🟩🟩🟩🟩🟩
Single gene	2–100 genes	>100 genes

Figure 1 | ***De novo* mutations and their impact on genetic disease.** **a** | Current estimates of the average mutation frequencies for the different types of *de novo* genomic variation observed per generation per genome. **b** | The general relationship between the mutational target size and the frequency of genetic diseases that are caused largely by *de novo* mutations. For disorders that are caused by particular mutations in single genes, the low probability of such a mutational event renders these disorders rare in the population. By contrast, disorders that can be caused by one (or a few) mutations in a large number of genes are relatively common. **c** | Factors increasing the frequency of *de novo* mutations. The occurrence of one or more of these factors can significantly affect the population frequency of certain genetic diseases.

CHARGE syndrome
A rare genetic disorder that arises during early fetal development and affects multiple organ systems, such as the eyes, heart and ears.

causality of rare chromosomal variants. The availability of high-resolution genomic microarrays in the past decade allowed the unbiased genome-wide analysis of *de novo* CNVs long before the same could be achieved for *de novo* SNVs and indels. Such analyses of CNVs have revealed the importance of this type of genomic variation in neurodevelopmental disorders such as ID, ASDs and schizophrenia (reviewed in REFS 49,50).

De novo CNVs larger than 100 kb are infrequent in the normal population, occurring in approximately one in 50 individuals[30]. By contrast, these large *de novo* CNVs occur in approximately 10% of all patients with sporadic ID[50,51], ASDs[52,53] or schizophrenia[54]. The use of high-resolution genomic microarrays to analyse the genetics of these disorders has resulted in the identification of many new recurrent microdeletion syndromes such as those caused by deletions affecting the chromosomal loci 1q21.1, 3q29, 15q13.3, 15q24, 17q12 and 17q21.31 (REF. 33). Many of these CNVs occur *de novo* in the patient and are rare or have never been observed in individuals without a neurodevelopmental phenotype, facilitating our assessment of their usefulness in a diagnostic setting. The observation that a particular CNV has occurred *de novo* in a patient with a sporadic disease is used in diagnostic decision making as an argument in favour of CNV pathogenicity[51], although it has been noted that inherited CNVs can be causative and *de novo* CNVs can be benign. Thus, *de novo* occurrence should not be the only criterion used when diagnosing disease[55]. Related to this, a two-hit CNV model was recently proposed for neurodevelopmental disease[56]; in this model, for patients with recurrent, mostly inherited CNVs, the presence of a second particular CNV elsewhere in the genome was associated with an increased penetrance and expressivity of disease.

In search of* de novo *SNVs: candidate gene studies. One factor that has hindered the detection of causative *de novo* SNVs in common genetic disorders has been the extreme genetic heterogeneity of these traits. This heterogeneity complicates both the detection and the functional interpretation of rare *de novo* mutations in common disease. Before exome sequencing became available, a set of imaginative studies of patients with ID and other neurodevelopmental disorders evaluated the contribution of *de novo* SNVs in genes encoding proteins that are known to have physiological roles at the synapse. Examples of such SNVs include *de novo* mutations in the gene encoding synaptic RAS GTPase-activating protein 1 (*SYNGAP1*) in non-syndromic forms of ID[57], and in the *SHANK3* (SH3 and multiple ankyrin repeat domains 3) gene in patients with schizophrenia[58]. In one particular study, a systematic investigation of *de novo* SNVs in 401 synapse-associated genes was carried out for 142 individuals with ASDs and 143 individuals with schizophrenia, mostly of sporadic origin[59]. Of note, all DNA sequencing in this study was carried out with conventional Sanger technology, which must have been a laborious undertaking. In total, 14 *de novo* SNVs were identified in blood samples from affected individuals but not in those from their parents. Eight of these mutations were non-synonymous and were predicted to substantially alter protein structure and/or function. Some of the mutations were in known disease genes such as *SHANK3*, *IL1RAPL1* (the interleukin-1 receptor accessory protein-like 1 gene) and *NRXN1* (the neurexin 1 gene).

A similar study that was carried out for non-syndromic ID was published in 2011 by the same group[60]. In this study, 197 candidate disease genes were sequenced by the Sanger method in 95 individuals with sporadic

Box 3 | The burden of disease that is due to *de novo* mutations

De novo mutations probably contribute to all human genetic diseases, but this contribution will vary greatly between diseases. The relative contribution of *de novo* mutations to a particular disease depends on both the frequency of *de novo* mutations causing this disease and the frequency of inherited and non-genetic factors that contribute to disease occurrence. This may be expressed as:

$$r_{DN} = \frac{n_{DN}}{n_E + n_M + n_{AD} + n_{AR} + n_{XL} + n_{DN}}$$

(in which r_{DN} is the relative contribution of *de novo* mutations to disease, n_{DN} is the number of patients with *de novo* mutations that cause this disease, n_E is the number of cases caused by environmental factors, n_M is the number of cases caused by multigenic inheritance, n_{AD} is the number of cases caused by autosomal dominant inheritance, n_{AR} is the number of cases caused by autosomal recessive inheritance and n_{XL} is the number of cases caused by X-linked inheritance).

The frequency of *de novo* mutations that cause a particular disease is largely determined by the size of the mutational target for this disease, which is roughly proportional to the genomic size of all genes and non-genic elements that can cause the disorder when mutated. As noted above, some genomic sites will have a relatively greater mutability, and some sites will have a strong paternal-age effect, but these will not be major factors for common diseases in which many genes and non-genic elements all over the genome are involved. We can deduce from this formula that the proportion of cases that is due to *de novo* mutation will be high if monogenic causes predominate, if the number of dominant disease-associated genes is high and if dominant mutations have strongly negative fitness effects, thereby reducing the role of inherited factors. Conversely, for conditions in which dominant inheritance has modest fitness effects, the number of inherited alleles will outnumber those that are due to *de novo* mutations. These conditions would seem to favour *de novo* mutations making a large contribution to neurodevelopmental conditions, given the large number of genes that are relevant for brain development and function, and the strongly reduced genetic fitness of affected individuals. There are undoubtedly many genes that contribute to autosomal and X-linked recessive forms of neurodevelopmental disorders[76]. The relative contribution of recessive alleles to neurodevelopmental disease will vary between populations and is not presently known. However, recessive inherited alleles are unlikely to explain most cases of these diseases, as the empirical sibling recurrence is much less than 25% for intellectual disability, autism and schizophrenia[19]. In agreement with this, large *de novo* copy number variants are observed in approximately 10% of all sporadic cases with intellectual disability, autism or schizophrenia. It not yet possible to determine the precise contribution of the other forms of *de novo* single-nucleotide variants or small insertions and deletions to any of the common neurodevelopmental disorders, but the first studies would seem to suggest that their joint contribution is of similar magnitude or greater than previously observed for *de novo* copy number variants[11–18].

forms of ID. Again, three truncating *de novo* mutations were identified in *SYNGAP1*, as well as a truncating mutation and a splice site mutation in the syntaxin-binding protein 1 gene (*STXBP1*), another gene that is known to be associated with ID. Both studies support the theory that severe *de novo* mutations in known disease genes have a role in these neurodevelopmental disorders. An unbiased analysis of *de novo* mutations in common disease, however, requires the use of next-generation sequencing technology.

Family-based exome sequencing in common neurodevelopmental disorders. The first application of a family-based exome sequencing approach was to investigate the role of germline *de novo* SNVs in ten patients with sporadic ID[11]. After filtering the exome data for potentially *de novo* variants that were not known to occur in the normal population and that were predicted to affect protein function, nine non-synonymous *de novo* SNVs were validated by Sanger sequencing in seven out of the ten individuals tested. All of these mutations affected different genes. In two patients, *de novo* nonsense mutations were found in known ID-associated genes: *RAB39B* (encoding a small GTPase) in one patient and *SYNGAP1* in the other. In addition, in a male patient in whom no *de novo* mutation was identified, a maternally inherited mutation was found in the lysine-specific demethylase 5C gene (*KDM5C*; also known as *JARID1C*), another well-known ID-associated gene located on chromosome X. Further analysis demonstrated that this mutation had occurred *de novo* in the proband's carrier mother. Whether the other seven *de novo* non-synonymous mutations are pathogenic or benign mutations is unknown at present. However, four of these mutations are likely to be detrimental to protein function, and they affect plausible candidate genes for brain structure and function. Although this was a small study, the data point to an important role for *de novo* SNVs in ID.

The impact of *de novo* SNVs in sporadic forms of ASDs has been evaluated in four recent large-scale exome sequencing studies that each reported on more than 100 patient–parent trios (and quartets, by including unaffected siblings)[13–16]. Different exome enrichment assays, sequencing methods and data-filtering steps were used in each study, which may explain why the number of validated *de novo* SNVs varied from 0.77 to 1.19 per patient with an ASD, but also between 0.63 and 1.00 per unaffected sibling (TABLE 1). These results demonstrate that detection of *de novo* SNVs is still imperfect, which makes it difficult to draw firm conclusions about the potential differences between the *de novo* SNV rates of patients with an ASD and their unaffected siblings. Importantly, the *de novo* mutation rate was consistently somewhat higher in patients with an ASD than in their siblings. This may be partly explained by the observations that *de novo* SNVs occurred predominantly on the paternal allele and that mutation was associated with increased paternal age[15,16]. However, no data were presented to determine whether the patients studied in this investigation were conceived later than their unaffected siblings, a factor that has been well documented for ASDs by meta-analysis of epidemiological studies[61].

The interpretation of the role of these rare *de novo* mutations in genetically and clinically heterogeneous disorders such as ASDs, ID and schizophrenia[17,18] is still in its infancy, and it requires a considerable effort to determine the phenotypic effect of each detected *de novo* SNV. Nonetheless, a number of interesting observations can be made from these studies.

Predicting phenotypic consequences

Now that the detection of *de novo* mutations is no longer the limiting factor in understanding the genetic basis of sporadic disease, the next pressing question is how to interpret any given *de novo* change in the context of a patient's phenotype. *De novo* mutations can be considered the most extreme form of rare genetic variation present in our human population, and many of the challenges that are faced when interpreting the effects of rare inherited variants are also valid for *de novo* mutations.

KBG syndrome
A rare genetic condition that is characterized by facial dysmorphisms, macrodontia, skeletal anomalies and developmental delay.

Lymphoblastoid cell lines
Cell lines that are created through *in vitro* infection (and thus immortalization) of B cells with Epstein–Barr virus.

Induced pluripotent stem cells
Adult cells that have been reprogrammed to stem cells, which can differentiate into different cell types.

Ariosa Exhibit 1219, p. 6

Table 1 | **Number of *de novo* mutations identified in studies of autistic spectrum disorders**

Cohort	Number of individuals tested	All *de novo* SNVs		*De novo* nonsense or splice site SNVs		*De novo* synonymous SNVs		*De novo* non-synonymous SNVs	
		Total in cohort	Per individual	Total in cohort	Per individual	Total in cohort	Per individual	Total in cohort	Per individual
Patients with an ASD, Sanders *et al.*[13]	200	154	0.77	15	0.08	29	0.15	110	0.55
Healthy siblings, Sanders *et al.*[13]	200	126	0.63	5	0.03	39	0.20	82	0.41
Patients with an ASD, O'Roak *et al.*[15]	189	225	1.19	19	0.10	61	0.32	145	0.77
Healthy siblings, O'Roak *et al.*[15]	50	50	1.00	3	0.06	16	0.32	31	0.62
Patients with an ASD, Iossifov *et al.*[16]	343	311	0.91	25	0.07	79	0.23	207	0.60
Healthy siblings, Iossifov *et al.*[16]	343	288	0.84	12	0.03	69	0.20	207	0.60
Patients with an ASD, Neale *et al.*[14]	175	161	0.92	10	0.06	50	0.29	101	0.58
All patients with an ASD[13–16]	907	851	0.94	69	0.08	219	0.24	563	0.62
All healthy siblings[13,15,16]	593	464	0.78	20	0.03	124	0.21	320	0.54

ASD, autistic-spectrum disorder; SNV, single-nucleotide variant.

For *de novo* coding mutations, a hierarchy of evidence is emerging to suggest that recurrence of the mutation in the same gene in another or several unrelated patients with a similar phenotype provides some support for interpreting the effects of a mutation, but the strongest support is provided when that recurrence is combined with an absence of similarly damaging mutations in the unaffected population (that is, in individuals without the phenotype in question). Information from mouse mutant phenotypes and protein function provides further support, together with information about the evolutionary conservation of the mutated nucleotide (see also FIG. 2).

Recurrently mutated genes. Clearly, one can learn much about the functional consequences of *de novo* mutations in a gene if these mutations occur in multiple patients with a similar phenotype. However, this is rarely the case for genetically heterogeneous diseases such as ASDs, ID and schizophrenia. When the data were combined[14] from three of the four large-scale ASD-associated exome sequencing studies (testing 564 families between them)[13–15], only 18 genes were found to be mutated *de novo* multiple times, a number that is not significantly different from simulated control data[14]. One study[15] decided to follow up six *de novo*-mutated candidate ASD-associated genes in ~2,500 patients with an ASD using targeted next-generation sequencing. However, only four additional *de novo* events were identified in this way: two in *GRIN2B* (the gene encoding glutamate (*N*-methyl D-aspartate) receptor subunit-ε2), one in *SCN1A* (the gene encoding sodium channel type I subunit-α) and one in *LAMC3* (the gene encoding laminin subunit-γ3). Detailed genotype–phenotype studies are required for each recurrently mutated gene to determine whether these mutations are reliably associated with ASDs or other neurodevelopmental phenotypes.

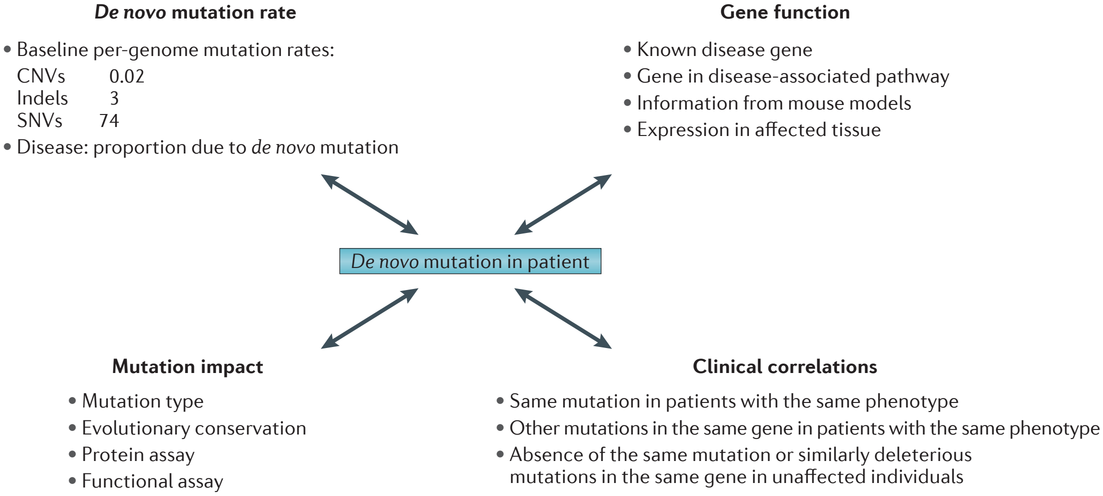
Gene function. A highly variable amount of information is available for different genes in order to guide statements about their likelihood of causing a specific human disease phenotype when mutated. The detection of *de novo* mutations in the exome can thus be regarded as a special case of the well-known candidate gene prioritization problem. All of the neurodevelopmental disease-associated exome sequencing studies discussed above include *in silico* evaluations of the cognate function of mutated genes in relation to the clinical characteristics of the disorder under study. There is an implicit hierarchy in the evidence, ranging (in order of decreasing strength) from a known gene responsible for the particular disorder, to a gene with a known function in the affected tissue, to mRNA expression patterns in relevant tissues, to various inferred functional attributes. The biological pathways in which a gene functions, together with data from model organisms and protein–protein interaction studies, can provide further supportive evidence. For example, mice that are heterozygous for *Yy1*, which encodes a transcriptional repressor, display growth retardation, neurulation defects and brain abnormalities[62], and these phenotypes may be relevant to the growth retardation and moderate ID that was seen in the first human patient found to have a *de novo* mutation in the human orthologue, *YY1* (REF. 11).

Attempts to develop more objective and statistically robust means of functional enrichment were provided first by studies focusing on the role of *de novo* CNVs in neurodevelopmental disease[63,64]. One group noted that the regions encompassed by these CNVs contain a significant enrichment of genes that give specific nervous system phenotypes when disrupted in the mouse[63]. Another group used a similar functional-enrichment mapping approach on the genes that were disrupted by rare CNVs (many of which occurred *de novo*) in patients with ASDs[64]. Such a method can be used to place single-gene effects into biologically meaningful groups[65]. This second

Penetrance
The proportion of patients with a specific phenotype among all carriers of a specific genotype.

Expressivity
The severity of the disease in individuals who have both the risk variant and the disease.

Genetic heterogeneity
The phenomenon by which mutations in different genes can cause a similar phenotype.



***De novo* mutation rate**

- Baseline per-genome mutation rates:
  CNVs 0.02
  Indels 3
  SNVs 74
- Disease: proportion due to *de novo* mutation

Gene function

- Known disease gene
- Gene in disease-associated pathway
- Information from mouse models
- Expression in affected tissue

De novo mutation in patient

Mutation impact

- Mutation type
- Evolutionary conservation
- Protein assay
- Functional assay

Clinical correlations

- Same mutation in patients with the same phenotype
- Other mutations in the same gene in patients with the same phenotype
- Absence of the same mutation or similarly deleterious mutations in the same gene in unaffected individuals

Figure 2 | **Information used to establish the pathogenicity of *de novo* mutations.** As for inherited mutations, an important challenge for interpreting *de novo* mutations is to identify which mutation (or mutations) is causal for a particular disease. This is not a trivial exercise, as the irrelevant *de novo* mutations in a patient are likely to outnumber those that might be disease causal. However, various lines of evidence can increase the confidence that a particular *de novo* mutation is causal. First, it is important to establish the population frequency of *de novo* mutations for each type of genomic variation. Copy number variants (CNVs) larger than 100 kb, for example, are rare in the general population, and therefore *de novo* occurrence itself is already an indication for pathogenicity; this is less true for the more commonly occurring *de novo* single-nucleotide variants (SNVs) and indels (small insertions or deletions) (top left). Next, it is important to evaluate the function of the gene (or genes) affected by the *de novo* mutation (top right). In addition, the type of mutation (nonsense, frameshift, splice site, non-synonymous, synonymous and non-coding) can be assessed for the likelihood that it results in deleterious consequences (such as disruption of a protein product or an alteration in a *cis*-regulatory region). This can be achieved either experimentally or computationally (bottom left). Finally, strong evidence of a causal role is provided by *de novo* mutations of interest being present in a gene in multiple affected individuals but absent from healthy controls (bottom right).

study showed that the regions affected by these rare CNVs were enriched for gene sets involved in neuronal development and function and in RAS and GTPase signalling. A third group applied degree-aware disease gene prioritization (DADA)[66] to link genes containing severe *de novo* mutations in patients with ASDs to a highly interconnected β-catenin and chromatin-remodelling protein network[15]. The problem of weighting functional evidence is conceptually similar to that involved in candidate disease gene prioritization for Sanger sequencing, for which numerous solutions have been developed. Many of these programs use multiple sources of gene function data and apply network approaches to derive a final likelihood score[67]. We expect in the near future to see a rapid development of algorithms that are targeted to the analysis of *de novo* mutations in the context of exome studies, possibly incorporating several of the elements outlined in FIG. 2.

Mutation type. Sites in the genome that have experienced purifying selection are considered important for normal function and are more likely to result in disease when mutated. Similarly, mutations that have an impact at the protein level are better candidates for pathogenicity than those without a functional impact. The ASD studies indeed show that nonsense and splice site *de novo* mutations in particular are enriched in patients with ASDs as compared to controls, indicating that severe disruptive mutations do play an important part in ASDs (TABLE 1). By contrast, no such enrichment is observed for *de novo* synonymous mutations that only rarely affect normal function. Most abundant, but also most difficult to interpret, are *de novo* missense mutations. All the studies described above[11–18] assessed the evolutionary conservation of the affected nucleotide by using either the phyloP[68] or similar Genomic Evolutionary Rate Profiling (GERP) conservation score[69]. A comparison of conservation scores for benign and pathogenic variants (mutations derived from dbSNP and The Human Gene Mutation Database (HGMD), respectively) showed that most pathogenic missense variants have a phyloP score of >2; this indicates a greater degree of conservation than the majority of common SNPs, which have scores of <2 (REF. 11). One study compared the distribution of phyloP scores between *de novo* and privately inherited variants in sporadic cases of schizophrenia and noted a statistically significant shift to higher phyloP scores for *de novo* mutations[18].

Non-synonymous missense mutations are often scored for their functional impact using the 'Grantham difference score' (REF. 70) or the 'polymorphism phenotyping' (PolyPhen-2) classification[71]. Although the average Grantham score is higher for pathogenic than for neutral SNVs[72], no significant difference was observed between the Grantham scores for *de novo* mutations and for rare inherited variants in one of the first studies on

Purifying selection
The conservation of functional genetic features during evolution because of selection against deleterious mutations.

Privately inherited
Pertaining to a genetic variant: confined to a single individual, family or population.

Grantham difference score
A score that predicts the effect of non-synonymous mutations based on the chemical properties of the substituted amino acids.

schizophrenia[14]. Another study combined the Grantham and phyloP scores to derive a probability score for each *de novo* mutation being observed in HGMD (that is, displaying characteristics of pathogenic mutations) or in dbSNP (having the characteristics of benign variants)[11]. In this small study, all the *de novo* mutations that were identified in genes with a functional link to ID showed a higher probability of being observed in HGMD than in dbSNP. The opposite applied to *de novo* mutations in genes without a functional link to ID. By contrast, another investigation did not observe a difference in the PolyPhen-2 classification of 101 non-synonymous *de novo* mutations identified in patients with an ASD as compared to random simulations[14].

As argued above, much more exome data is needed in order to develop validated prediction algorithms that are useful in the diagnosis of *de novo* mutations in individual patients. Intuitively, the combination of evidence at the gene function level and at the mutation level should make the best case for pathogenicity. As many *de novo* mutations will cause disease through a loss-of-function mechanism, it is important to know whether a gene is dosage sensitive or not. Although the dosage sensitivity of genes cannot be predicted with great confidence currently, some evidence indicates that this parameter could be better predicted in the future[73]. In addition, the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER[74]) and the Database of Genomic Variants (DGV[75]) provide insight into genomic regions that are sensitive and insensitive to CNVs, respectively. Nonetheless, the clinical relevance of a *de novo* mutation remains questionable as long as there are no additional patients with mutations in the same gene. In the end, the human phenotype decides.

Genetic counselling for *de novo* mutations

For parents who have had a child with a genetic disease caused by a *de novo* mutation, what should we tell them about the probable risks to existing or future siblings of the affected child? Clearly, the recurrence risk would be negligible if *de novo* mutations occurred exclusively in germ cells. But this is not always so. New mutations can and do occur at any stage of gametogenesis and, indeed, of development[76]. This creates two kinds of mosaicism, each of which carries different consequences for the risk of recurrence in siblings.

If the patient carries a *de novo* mutation in mosaic form, the most likely scenario is that the mutation arose postzygotically, and therefore the recurrence risk would be essentially zero. Somatic mosaicism may be a frequent event. For instance, somatic mosaicism for a mutation in the adenomatous polyposis coli gene (*APC*) was detectable in 11% of patients with polyposis coli but without a prior family history[77].

A related situation but with different counselling consequences would be a *de novo* mutation in an affected child for whom one of the clinically unaffected parents carries the mutation in a percentage of somatic and germline cells[78]. Clearly, somatic mosaicism in one of the parents increases the likelihood of recurrence of the condition in future offspring. In those instances in which one of the parents has confirmed somatic mosaicism, the recurrence risk after the birth of an affected child may be as high as 50%. There is now a wide range of disorders for which the occurrence of parental germline mosaicism has been reported[78]. Parental germline mosaicism has been documented in up to 5% of mothers of patients with Duchenne muscular dystrophy[81], and in 11% of mothers of patients with haemophilia A[80]. In addition, somatic mosaicism was demonstrated in either of the parents in 11% of tested families containing a child with craniofrontonasal dysplasia[82]. In a study of adrenoleukodystrophy in families with a single affected child carrying an apparently *de novo* mutation in the ATP-binding cassette gene *ABCD1*, the risk of recurrence was estimated to be at least 13%[83].

Adrenoleukodystrophy
A rare genetic disorder that results in progressive brain damage, failure of adrenal glands and, eventually, death.

As more sensitive mutation detection methods are applied, the frequency of somatic and germline mosaicism may turn out to be considerably higher than we now appreciate. We expect that sensitive detection and quantification of mutations using next-generation sequencing technology will soon become a standard tool for the accurate estimation of recurrence risks in families for which a genetic disease is caused by an apparently *de novo* disease gene mutation event.

Towards prenatal screening for *de novo* SNVs?

In many countries, special screening programmes have been established for pregnant women above a certain age. The risk of having a child with a *de novo* chromosomal abnormality increases substantially with maternal age, and these chromosomal abnormalities can be easily detected by conventional chromosome analysis. Now that we are better able to identify *de novo* SNVs, indels and CNVs, one may wonder whether in the future prenatal screening should be extended to conditions that are caused by these types of *de novo* mutation. Clearly the focus should no longer be on older women, as most types of *de novo* mutation are in fact correlated with increased paternal age. Next-generation sequencing, perhaps using the free fetal DNA in maternal plasma[84,85], will soon allow us to offer screening to every pregnant mother, first for chromosomal abnormalities and later for smaller genetic defects. The interpretation of these rare *de novo* events will clearly be extremely challenging in a prenatal setting, especially because many of these mutations have variable penetrance and no phenotype information is available to guide interpretation. The application of trio sequencing for *de novo* mutations in prenatal diagnosis will need careful consideration of all the issues involved.

Outlook

The widespread availability of next-generation sequencing technology has boosted the study of *de novo* mutations in health and disease. Pilot studies in early-onset neurodevelopmental disease have indicated that *de novo* mutations may play a much more important part than was previously assumed, explaining in part why these diseases with severely reduced fitness remain frequent in the human population. We expect that *de novo* mutations are relevant to many other common diseases. A focus on *de novo* mutations is also an attractive analytical strategy for whole-genome sequences obtained in cases of sporadic disease. In contrast to the millions of inherited variants per genome, the number of *de novo* germline variants will be only about 100 per genome. As these variants will be mostly unique, it will remain challenging for a long time to distinguish benign from pathogenic *de novo* mutations outside of coding regions. The systematic collection and international sharing of these mutation data, together with associated phenotypic and additional functional information, may prove to be crucial for furthering our understanding of the non-coding part of our genome.

1. Raychaudhuri, S. Mapping rare and common causal alleles for complex human diseases. *Cell* **147**, 57–69 (2011).
2. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43**, 712–714 (2011).
3. Vogel, F. & Rathenberg, R. Spontaneous mutation in man. *Adv. Hum. Genet.* **5**, 223–318 (1975).
4. Kondrashov, A. S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.* **21**, 12–27 (2002).
These authors accurately estimate the *de novo* mutation rate of SNVs long before the availability of whole-genome and whole-exome sequencing technologies.
5. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1**, 40–47 (2000).
6. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Rev. Genet.* **8**, 610–618 (2007).
7. Hoischen, A. *et al. De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nature Genet.* **42**, 483–485 (2010).
The first demonstration of exome sequencing being used to identify *de novo* mutations in a rare clinical syndrome.
8. Ng, S. B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nature Genet.* **42**, 790–793 (2010).
9. Hoischen, A. *et al. De novo* nonsense mutations in *ASXL1* cause Bohring-Opitz syndrome. *Nature Genet.* **43**, 729–731 (2011).
10. Lindhurst, M. J. *et al.* A mosaic activating mutation in *AKT1* associated with the Proteus syndrome. *N. Engl. J. Med.* **365**, 611–619 (2011).
The first application of exome sequencing to discover somatic *de novo* mutations as the cause of a genetic disorder.
11. Vissers, L. E. L. M. *et al.* A *de novo* paradigm for mental retardation. *Nature Genet.* **42**, 1109–1112 (2010).
The first study to use exome sequencing of patient–parent trios to identify *de novo* mutations in a complex trait that is characterized by extreme genetic heterogeneity.
12. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43**, 585–589 (2011).
13. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
14. Neale, B. M. *et al.* Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
15. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* **485**, 246–250 (2012).
16. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
The largest-scale exome sequencing study carried out to date. The authors study 343 quartets, each consisting of a patient with an ASD, an unaffected sibling and their unaffected parents, to evaluate the frequency and type of *de novo* mutations in affected and unaffected siblings.
17. Girard, S. L. *et al.* Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nature Genet.* **43**, 860–863 (2011).
18. Xu, B. *et al.* Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. *Nature Genet.* **43**, 864–868 (2011).
19. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
20. McClellan, J. & King, M. C. Genomic analysis of mental illness: a changing landscape. *JAMA* **303**, 2523–2524 (2010).
21. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nature Rev. Genet.* **11**, 685–696 (2010).
22. Ding, L., Wendl, M. C., Koboldt, D. C. & Mardis, E. R. Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum. Mol. Genet.* **19**, R188–R196 (2010).
23. Arnheim, N. & Calabrese, P. Understanding what determines the frequency and pattern of human germline mutations. *Nature Rev. Genet.* **10**, 478–488 (2009).
24. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nature Rev. Genet.* **12**, 756–766 (2011).
25. Loewe, L. & Hill, W. G. The population genetics of mutations: good, bad and indifferent. *Phil. Trans. R. Soc. B.* **365**, 1153–1167 (2010).
26. Haldane, J. B. S. The rate of spontaneous mutation of a human gene. *J. Genet.* **31**, 317–326 (1935).
27. Kondrashov, A. S. & Crow, J. F. A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.* **2**, 229–234 (1993).
28. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
The first family-based whole-genome sequencing study in which all potential *de novo* mutations are independently validated to provide accurate *de novo* mutation rates.
29. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
30. Itsara, A. *et al. De novo* rates and selection of large copy number variation. *Genome Res.* **20**, 1469–1481 (2010).
31. Allen, G. Aetiology of Down's syndrome inferred by Waardenburg in 1932. *Nature* **250**, 436–437 (1974).
32. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).
33. Girirajan, S., Campbell, C. D. & Eichler, E. E. Human copy number variation and complex genetic disease. *Annu. Rev. Genet.* **45**, 203–226 (2011).
34. Vissers, L. E. L. M. *et al.* Mutations in a novel member of the chromodomain gene family cause CHARGE syndrome. *Nature Genet.* **36**, 955–957 (2004).
35. Bamshad, M. J., *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Rev. Genet.* **12**, 745–755 (2011).
A valuable review that explains the experimental and analytical options for applying exome sequencing in studies of disease genes. The key challenges in using this approach are also discussed.
36. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol.* **12**, 228 (2011).
37. Sirmaci, A., *et al.* Mutations in *ANKRD11* cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia. *Am. J. Hum. Genet.* **89**, 289–294 (2011).
38. Rivière, J. B., *et al. De novo* mutations in the actin genes *ACTB* and *ACTG1* cause Baraitser-Winter syndrome. *Nature Genet.* **44**, 440–444 (2012).
39. Tsurusaki, Y., *et al.* Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nature Genet.* **44**, 376–378 (2012).
40. Santen, G. W., *et al.* Mutations in SWI/SNF chromatin remodeling complex gene *ARID1B* cause Coffin-Siris syndrome. *Nature Genet.* **44**, 379–380 (2012).
41. Pansuriya, T. C., *et al.* Somatic mosaic *IDH1* and *IDH2* mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. *Nature Genet.* **43**, 1256–1261 (2011).
42. Vissers, L. E. L. M. *et al.* Whole-exome sequencing detects somatic mutations of *IDH1* in metaphyseal chondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA). *Am. J. Med. Genet. A* **155**, 2609–2616 (2011).
43. Limaye, N., Boon, L. M. & Vikkula, M. From germline towards somatic mutations in the pathophysiology of vascular anomalies. *Hum. Mol. Genet.* **18**, R65–R74 (2009).
44. Pasmooij, A. M., Pas, H. H., Bolling, M. C. & Jonkman, M. F. Revertant mosaicism in junctional epidermolysis bullosa due to multiple correcting second-site mutations in *LAMB3*. *J. Clin. Invest.* **117**, 1240–1248 (2007).
45. Frank, S. A. Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proc. Natl Acad. Sci. USA* **107** (Suppl. 1), 1725–1730 (2010).
46. Carlson, C. A. *et al.* Decoding cell lineage from acquired mutations using arbitrary deep sequencing. *Nature Methods* **9**, 78–80 (2011).
47. Vanneste, E. *et al.* Chromosome instability is common in human cleavage-stage embryos. *Nature Med.* **15**, 577–583 (2009).
A remarkable study which shows that *de novo* chromosomal aberrations are common in early embryogenesis. This identifies postzygotic chromosome instability as a major cause of constitutional genomic disorders.
48. van Echten-Arends, J. *et al.* Chromosomal mosaicism in human preimplantation embryos: a systematic review. *Hum. Reprod. Update* **17**, 620–627 (2011).
49. Cook, E. H. Jr & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919–923 (2008).
50. Vissers, L. E. L. M., de Vries, B. B. & Veltman, J. A. Genomic microarrays in mental retardation: from CNV to gene, from research to diagnosis. *J. Med. Genet.* **47**, 289–297 (2010).
51. Koolen, D. A. *et al.* Genomic microarrays in mental retardation: a practical workflow for diagnostic applications. *Hum. Mutat.* **30**, 283–292 (2009).
52. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
One of the first large-scale studies to highlight the important role of *de novo* CNVs in sporadic forms of ASDs, establishing *de novo* germline mutation as a more significant risk factor for ASDs than was previously recognized.
53. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
54. Xu, B. *et al.* Strong association of *de novo* copy number mutations with sporadic schizophrenia. *Nature Genet.* **40**, 880–885 (2008).
55. Vermeesch, J. R., Balikova, I., Schrander-Stumpel, C., Fryns, J. P. & Devriendt, K. The causality of *de novo* copy number variants is overestimated. *Eur. J. Hum. Genet.* **19**, 1112–1113 (2011).
56. Girirajan, S. *et al.* A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nature Genet.* **42**, 203–209 (2010).
57. Hamdan, F. F. *et al.* Mutations in *SYNGAP1* in autosomal nonsyndromic mental retardation. *N. Engl. J. Med.* **360**, 599–605 (2009).
58. Gauthier, J. *et al. De novo* mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. *Proc. Natl Acad. Sci. USA* **107**, 7863–7868 (2010).

Ariosa Exhibit 1219, p. 10

59. Awadalla, P. *et al.* Direct measure of the *de novo* mutation rate in autism and schizophrenia cohorts. *Am. J. Hum. Genet.* **87**, 316–324 (2010).
These authors carry out the first systematic large-scale sequencing study to evaluate the role of *de novo* mutations in candidate genes for ASDs and schizophrenia.
60. Hamdan, F. F. *et al.* Excess of *de novo* deleterious mutations in genes associated with glutamatergic systems in nonsyndromic intellectual disability. *Am. J. Hum. Genet.* **88**, 306–316 (2011).
61. Hultman, C. M. *et al.* 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).
62. He, Y. & Casaccia-Bonnefil, P. The Yin and Yang of YY1 in the nervous system. *J. Neurochem.* **106**, 1493–1502 (2008).
63. Webber, C. *et al.* Forging links between human mental retardation-associated CNVs and mouse gene knockout models. *PLoS Genet.* **5**, e1000531 (2009).
64. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
65. O'Dushlaine, C. *et al.* Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol. Psychiatry* **16**, 286–292 (2011).
66. Erten, S. *et al.* DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Min.* **4**, 19 (2011).
67. Chen, Y. *et al.* *In silico* gene prioritization by integrating multiple data sources. *PLoS ONE* **6**, e21137 (2011).
68. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
69. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
70. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
71. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
72. 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).
73. Huang, N. *et al.* Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet.* **6**, e1001154 (2010).
74. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
75. Zhang, J. *et al.* Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* **115**, 205–214 (2006).
76. Najmabadi, H. *et al.* Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478**, 57–63 (2011).
77. Vadlamudi, L. *et al.* Timing of *de novo* mutagenesis — a twin study of sodium-channel mutations. *N. Engl. J. Med.* **363**, 1335–1340 (2010).
78. Aretz, S. *et al.* Somatic APC mosaicism: a frequent cause of familial adenomatous polyposis (FAP). *Hum. Mutat.* **10**, 985–992 (2007).
79. Goriely, A. *et al.* Germline and somatic mosaicism for *FGFR2* mutation in the mother of a child with Crouzon syndrome: implications for genetic testing in "paternal age-effect" syndromes. *Am. J. Med. Genet. A* **152A**, 2067–2073 (2010).
80. Erickson, R. P. Somatic gene mutation and human disease other than cancer: an update. *Mutat. Res.* **705**, 96–106 (2010).
81. Helderman-van den Enden, A. T. *et al.* Recurrence risk due to germ line mosaicism: Duchenne and Becker muscular dystrophy. *Clin. Genet.* **75**, 465–472 (2009).
82. Twigg, S. R. *et al.* The origin of *EFNB1* mutations in craniofrontonasal syndrome: frequent somatic mosaicism and explanation of the paucity of carrier males. *Am. J. Hum. Genet.* **78**, 999–1010 (2006).
83. Wang, Y. *et al.* X-linked adrenoleukodystrophy: *ABCD1 de novo* mutations and mosaicism. *Mol. Genet. Metab.* **104**, 160–166 (2011).
84. Lo, Y. M. Fetal nucleic acids in maternal plasma. *Ann. NY Acad. Sci.* **1137**, 140–143 (2008).
85. Lo, Y. M. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2**, 61ra91 (2010).
86. Fragouli, E., Wells, D. & Delhanty, J. D. Chromosome abnormalities in the human oocyte. *Cytogenet. Genome Res.* **133**, 107–118 (2011).
87. Goriely, A. & Wilkie, A. O. Missing heritability: paternal age effect mutations and selfish spermatogonia. *Nature Rev. Genet.* **11**, 589 (2010).
88. Goriely, A. & Wilkie, A. O. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90**, 175–200 (2012).
89. Toriello, H. V., Meck, J. M. & Professional Practice and Guidelines Committee. Statement on guidance for genetic counseling in advanced paternal age. *Genet. Med.* **10**, 457–460 (2008).
90. Hehir-Kwa, J. Y. *et al.* *De novo* copy number variants associated with intellectual disability have a paternal origin and age bias. *J. Med. Genet.* **48**, 776–778 (2011).
91. Koolen, D. A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature Genet.* **38**, 999–1001 (2006).
92. Sharp, A. J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genet.* **38**, 1038–1042 (2006).
93. Shaw-Smith, C. *et al.* Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nature Genet.* **38**, 1032–1037 (2006).
94. Koolen, D. A. *et al.* Clinical and molecular delineation of the 17q21.31 microdeletion syndrome. *J. Med. Genet.* **45**, 710–720 (2008).
95. Itsara, A. *et al.* Resolving the breakpoints of the 17q21.31 microdeletion syndrome with next-generation sequencing. *Am. J. Hum. Genet.* **90**, 599–613 (2012).
96. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
97. Baudat, F. *et al.* PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* **327**, 836–840 (2010).
98. Berg, I. L. *et al.* Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc. Natl Acad. Sci. USA* **108**, 12378–12383 (2011).
99. Berg, I. L. *et al.* PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nature Genet.* **42**, 859–863 (2010).
100. Tomé, S. *et al.* Maternal germline-specific effect of DNA ligase I on CTG/CAG instability. *Hum. Mol. Genet.* **20**, 2131–2143 (2011).
101. Liu, P. *et al.* Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements, *Cell* **146**, 889–903 (2011).
102. Bunyan, D. J. & Robinson, D. O. Multiple *de novo* mutations in the *MECP2* gene. *Genet. Test.* **12**, 373–375 (2008).
103. Lam, H. Y. *et al.* Performance comparison of whole-genome sequencing platforms. *Nature Biotech.* **30**, 78–82 (2011).
104. Londin, E. R. *et al.* Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics* **12**, 464 (2011).
105. Hussein, S. M. *et al.* Copy number variation and selection during reprogramming to pluripotency. *Nature* **471**, 58–62 (2011).
106. Liang, Q., Conte, N., Skarnes, W. C. & Bradley, A. Extensive genomic copy number variation in embryonic stem cells. *Proc. Natl Acad. Sci. USA* **105**, 17453–17456 (2008).
107. Pamphlett, R., Morahan, J. M. & Yu, B. Using case-parent trios to look for rare *de novo* genetic variants in adult-onset neurodegenerative diseases. *J. Neurosci. Methods* **197**, 297–301 (2011).

Acknowledgements

Joris A. Veltman is supported by personal grants from the Netherlands Organization for Health Research and Development (917-66-363) and the European Research Council (281964).

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Genomic Disorders Nijmegen: http://www.genomicdisorders.nl
NRG article series on the Applications of Next-Generation Sequencing: http://www.nature.com/nrg/series/nextgeneration/index.html
1000 Genomes: http://www.1000genomes.org
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/
DECIPHER: http://decipher.sanger.ac.uk/
DGV: http://projects.tcag.ca/variation/project.html
GERP: http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html
HGMD: http://www.hgmd.org
phyloP: http://compgen.bscb.cornell.edu/phast/help-pages/phyloP.txt
PolyPhen-2: http://genetics.bwh.harvard.edu/pph2/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF