Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease

Jose Bras, Rita Guerreiro and John Hardy

Abstract | Over the past five years the field of neurogenetics has yielded a wealth of data that have facilitated a much greater understanding of the aetiology of many neurological diseases. Most of these advances are a result of improvements in technology that have allowed us to determine whole-genome structure and variation and to examine its impact on phenotype in an unprecedented manner. Genome-wide association studies have provided information on how common genetic variability imparts risk for the development of various complex diseases. Moreover, the identification of rare disease-causing mutations have led to the discovery of novel biochemical pathways that are involved in disease pathogensis. Here, we review these advances and discuss how they have changed the approaches being used to study neurological disorders.

Case–control studies
Studies in which genetic variability in genes of interest are compared between a group of cases (for example, patients) and a group of controls from the same population.

Mendelian genes
Genes in which mutations cause disease in a Mendelian manner. The disease can be recessive or dominant in its inheritance mode.

Obtaining a precise and complete understanding of the pathobiological mechanisms underlying neurological diseases has long been the focus of many genetic studies. Until recently, there were three main ways in which genetics was able to help clarify these mechanisms. First, genetic linkage studies could be performed when large families with multiple affected individuals were available and could be segregated according to the presence or absence of a particular phenotype. These studies aimed to identify genomic regions within which the causative defect that resulted in the phenotype would be present. Second, when genes that could carry disease-causing mutations had previously been identified, other populations could be screened for mutations in those genes in candidate gene association studies, yielding information on prevalence and overall clinical significance. Finally, case–control studies could be used when a gene was deemed to have a plausible role in disease onset.

The use of these three approaches resulted in considerable successes over the years, culminating in the identification of many Mendelian genes and some genetic risk factors (as well as the elucidation of their specific impact in different populations) for numerous neurological diseases. These findings can be seen, for example, by searching the Online Mendelian Inheritance in Man database. Nonetheless, these three approaches have several limitations. Linkage studies require large, multigenerational pedigrees within which both affected and unaffected individuals are needed for testing (and even in cases in which these individuals are available, this approach yields only regions of linkage and not the causative gene). Candidate gene association studies require an *a priori* hypothesis for the selection of the gene to be studied. Candidate gene association studies, in particular, were subject to large numbers of false-positive reports, in which associations that were published were not replicated by independent groups[1]. It is therefore clear that these approaches were not sufficient to identify all the genetic events underlying the disorders that were being examined.

Now, recent advances in technology have allowed the interrogation of very large numbers of markers dispersed throughout the genome in a highly rapid and inexpensive manner and the determination of the sequence of nucleotides for all the coding portions of all known genes. These two techniques, together with other technological advances, have provided an overview of both common and rare genetic variability across the whole genome that has improved our understanding of many neurological diseases. In addition, these new technologies have altered the way in which experiments can be designed, resulting in a move from mostly hypothesis-based approaches to studies that, by interrogating the entire genome, are largely hypothesis free. In this Review, we describe each of the major modes of inheritance of neurological diseases, briefly discuss classical methods

Reta Lilla Weston Laboratories and Department of Molecular Neuroscience, Institute of Neurology, University College London, Queen Square, London WC1N 3BG, UK.
Correspondence to J.H.
e-mail: j.hardy@ucl.ac.uk
doi:10.1038/nrn3271

© 2012 Macmillan Publishers Limited. All rights reserved

Sporadic diseases
Diseases that occur with no known genetic background and therefore no family history of occurence.

Sanger DNA sequencing
A method used to determine the nucleotides present in a fragment of DNA. It is based on the chain-terminator method developed by Frederick Sanger, but currently uses labelling of the chain-terminator dideoxynucleotides (ddNTPs), which allows sequencing in a single reaction.

Genotyping
A method used to determine the bases (genotypes) present at specific positions in the genome. This can be done for a small number of genotypes or for millions of genotypes spread throughout the genome (using whole-genome genotyping arrays).

Autosomal recessive diseases
Patterns of disease inheritance in which both alleles (one from each parent) need to present the genetic defect for the disease to manifest itself.

Consanguinity
This refers to individuals who are related by blood.

for gene identification and then consider how new genome-wide strategies can be used to identify disease-related genes and dissect genetic risk. We have opted, in some sections, to focus on experiments that have used whole-exome sequencing experiments rather than those using whole-genome sequencing, because a proportionally small number of studies use whole-genome sequencing; however, this is something that we anticipate will change in the future.

Methods for gene identification

Dissecting Mendelian diseases. Mendelian neurological diseases are those in which possession of one copy (for dominant genes) or two copies (for recessive genes) of the mutant gene leads inevitably to the development of the disease. Classical examples are Huntington's disease (a dominant Mendelian disease) and Friedreich's ataxia (a recessive Mendelian disease). Diseases with such simple inheritance patterns are generally rare and constitute a small proportion of all cases of neurological disease. However, these forms of disease have been the basis for our understanding of a large number of pathobiological events in other neurological disorders. The identification of faulty genes in Mendelian disorders indicates that certain pathways are involved in the pathogenesis of the disease; information that can be applied to more common sporadic diseases.

The way that genes that are involved in Mendelian diseases are identified has changed dramatically in recent years. Traditional approaches relied on linkage studies that necessitate large multi-generational pedigrees in which multiple affected and unaffected individuals are available for testing. When these were available, researchers looked for the presence of certain markers (for example, panels of genetically variable DNA sequences with known chromosomal locations) that could be used to determine which alleles were present only in the affected individuals. This provided information regarding the location of the causative gene, after which Sanger DNA sequencing would be performed to pinpoint the actual mutation. Although this is a powerful approach that has yielded many substantive findings (such as the cloning of the Parkinson disease (autosomal dominant) 8 (*PARK8*) locus for Parkinson's disease[2]), linkage studies have several drawbacks that limit their utility.

Large pedigrees are not available in many cases, particularly for late-onset diseases in which older generations have often died and descendants have not yet reached the age of disease onset. Additionally, although the markers used to perform linkage analysis were dispersed throughout the genome, their number was usually only in the hundreds. This meant that regions of linkage were large and generally contained tens to hundreds of genes, which affected the costs of follow-up sequencing. The use of more numerous and more informative panels of markers would still result in the identification of large regions of linkage, given that recombination rates, rather than the amount of markers, are the factors underlying size limitations in this type of analysis. Furthermore, the amount of time required to study a single family using this approach was considerable (usually several months to years).

Recently, the advent of genome-wide genotyping and second-generation sequencing (the ability to sequence millions of short fragments of DNA in parallel) has changed the way in which many Mendelian diseases are studied. For example, autosomal recessive diseases are well suited for autozygosity analysis by high-density genotyping. This process results in the identification of regions of the genome that are homozygous only in the affected individuals, thereby suggesting the presence of a homozygous mutation (FIG. 1). The procedure takes only 3 days of laboratory work to identify candidate regions, and needs DNA samples from as little as two affected and one unaffected individuals from the same kindred. As a comparison, the amount of laboratory work required to perform linkage analysis in such a kindred would be usually in the order of several weeks. As in traditional linkage analysis, the mapping of homozygous areas is followed by Sanger DNA sequencing to pinpoint the causal variation, which is the rate-limiting step for the entire procedure. This approach has been successfully applied to various neurological diseases, an example of which is the identification of mutations in phospholipase A2, group VI (*PLA2G6*) in early onset parkinsonism dystonia[3–6]. Even though autozygosity mapping is a straightforward and time-efficient approach, it does not always allow the identification of the genetic defect underlying the disease[7]. For example, a recent study that investigated a family from Israel presenting with early onset Alzheimer's disease in which there was extensive consanguinity failed to find the causal mutation[8].

Identifying causative genes in autosomal dominant diseases in small kindreds is considerably more complicated, as it is not possible to rely on homozygosity mapping to indicate the most plausible regions of linkage. The approach most commonly taken in these cases is to sequence all of the known coding portions of the genome using a second-generation sequencing approach commonly called exome sequencing[9,10] (FIG. 2) (for a review see REF. 11). In this approach, the coding portion of genomic DNA is selected from a pool of all DNA fragments by hybridization with labelled probes that are complementary to the amino-acid coding sequences. Following its selection, the coding DNA is sequenced using second-generation sequencing. This approach reveals a large number of variations in coding regions between affected and unaffected individuals, which may be further refined by sequencing additional family members to confirm segregation of the causal mutation. A typical experiment will identify ~20,000 variants per individual sequenced, some of which will be benign variants previously described in several healthy individuals. Depending on the number and relationships of additional family members, this number can be reduced to a more manageable amount of variants, which can then be examined further to eventually identify those variants that have a causative role in the disease. This approach is still somewhat time-consuming: sequencing takes a couple of weeks,

© 2012 Macmillan Publishers Limited. All rights reserved

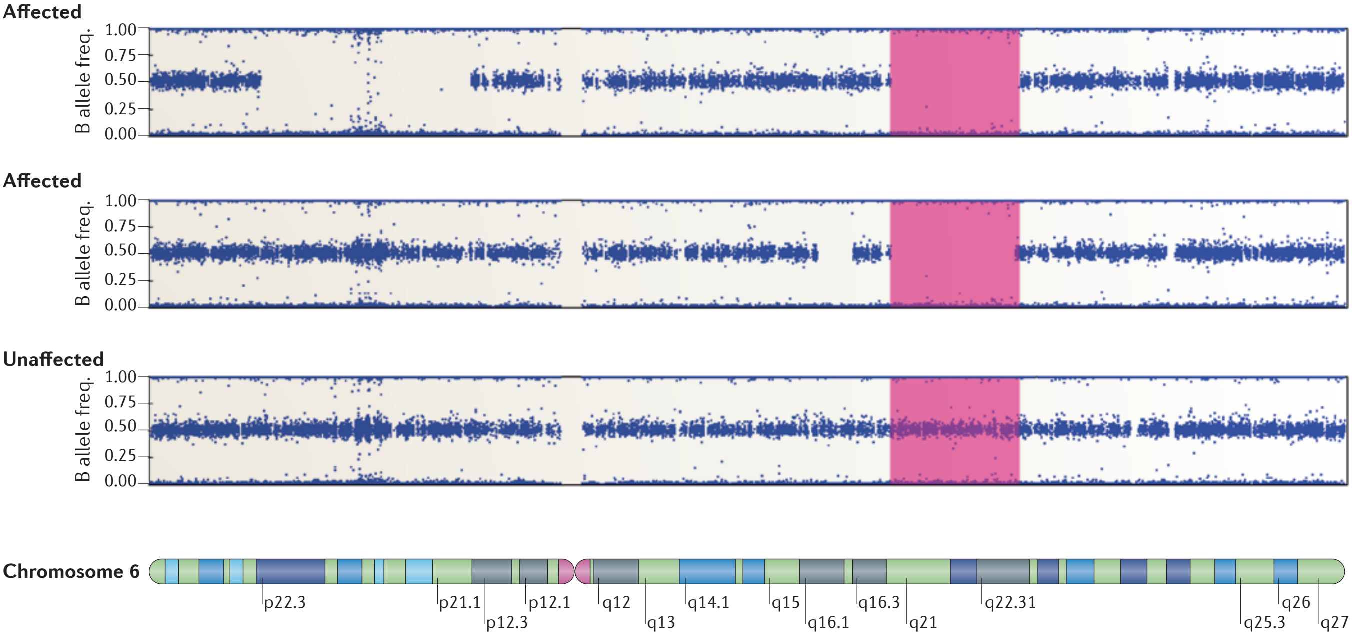


Figure 1 | **Homozygosity mapping using high-density arrays.** The figure shows homozygosity mapping for three individuals (two affected by a frontotemporal dementia-like disease and one unaffected). Each blue dot in the upper three panels represents one individual single-nucleotide polymorphism (SNP). For each SNP, a low B allele frequency indicates that the individual is a homozygote for the A allele; intermediate values mean they are a heterozygote and high B allele frequency means that they are a homozygote for the B allele. In this example, there is a region of chromosome 6 (shown in pink) in which the affected individuals (top two panels) present with a large homozygous region that is not shared by the unaffected family member (lower panel), thus suggesting that a disease-causing homozygous mutation may be present in this genomic region. The region of chromosome 6 studied is shown at the bottom of the figure.

bioinformatics analysis takes a few days, and testing segregation also takes a few weeks. Nevertheless, this still compares favourably to the considerable amount of time it would have taken to perform Sanger DNA sequencing of all the exons.

Although considerably more expensive, it is becoming increasingly frequent for groups to identify Mendelian mutations using whole-genome sequencing (FIG. 2). This approach attempts to sequence every base in the genome of a few individuals from a family presenting with a segregating disease. Not only is this approach more expensive than exome sequencing, it is also bioinformatically more challenging, as the volume of data produced is much larger and the ability to make sense of non-coding variability is still limited. Nonetheless, this is an approach that has already yielded some positive results[12,13] (FIG. 3).

These novel genome-wide methods have revolutionized the approaches being used to study Mendelian neurological diseases. They have a more rapid turnaround time and require smaller numbers of family members than traditional approaches, culminating in the identification of many new genetic loci[14–17].

Identifying common low-risk loci. On the opposite end of the spectrum to Mendelian disorders are common diseases that have multifactorial traits that contribute to the phenotype. In these cases, no single genetic defect causes the disease; instead, several genes impart risk for disease development. The common disease, common variant hypothesis postulates that, for these diseases, common genetic variability (that is, variability with high minor allele frequency that can therefore be seen in most individuals) would modulate risk of developing the disease[18,19]. Previously, such common low-risk loci could only be identified through candidate gene analysis, as described above. This approach became the basis for genome-wide association studies (GWASs), in which a large number of markers are examined in large numbers of case and control samples[20].

The GWAS approach has yielded some remarkably important results for several neurological diseases (a catalogue of all such studies is maintained by the US National Institutes of Health's National Human Genome Research Institute). For Parkinson's disease, for example, GWASs identified several genes and loci previously unknown to be connected to the disease and confirmed previous suggestions that genes involved in Mendelian forms of Parkinson's disease also modulate risk for the more common and sporadic form of this disease[21–23]. Similarly, in Alzheimer's disease, the single most significant risk factor, apolipoprotein E, was replicated in all the GWASs performed for this disease[24] and several previously unidentified genes involved in various biological processes were also found to exert an effect on disease risk[25,26]. Another neurological disease for which large-scale GWASs have been performed is multiple sclerosis. In this case, over 30 loci, most of which are involved in the immune response, were identified as modulating risk for the disease[27].

Minor allele frequency
For a single-nucleotide polymorphism this is the frequency of the less frequent allele in a population.

© 2012 Macmillan Publishers Limited. All rights reserved

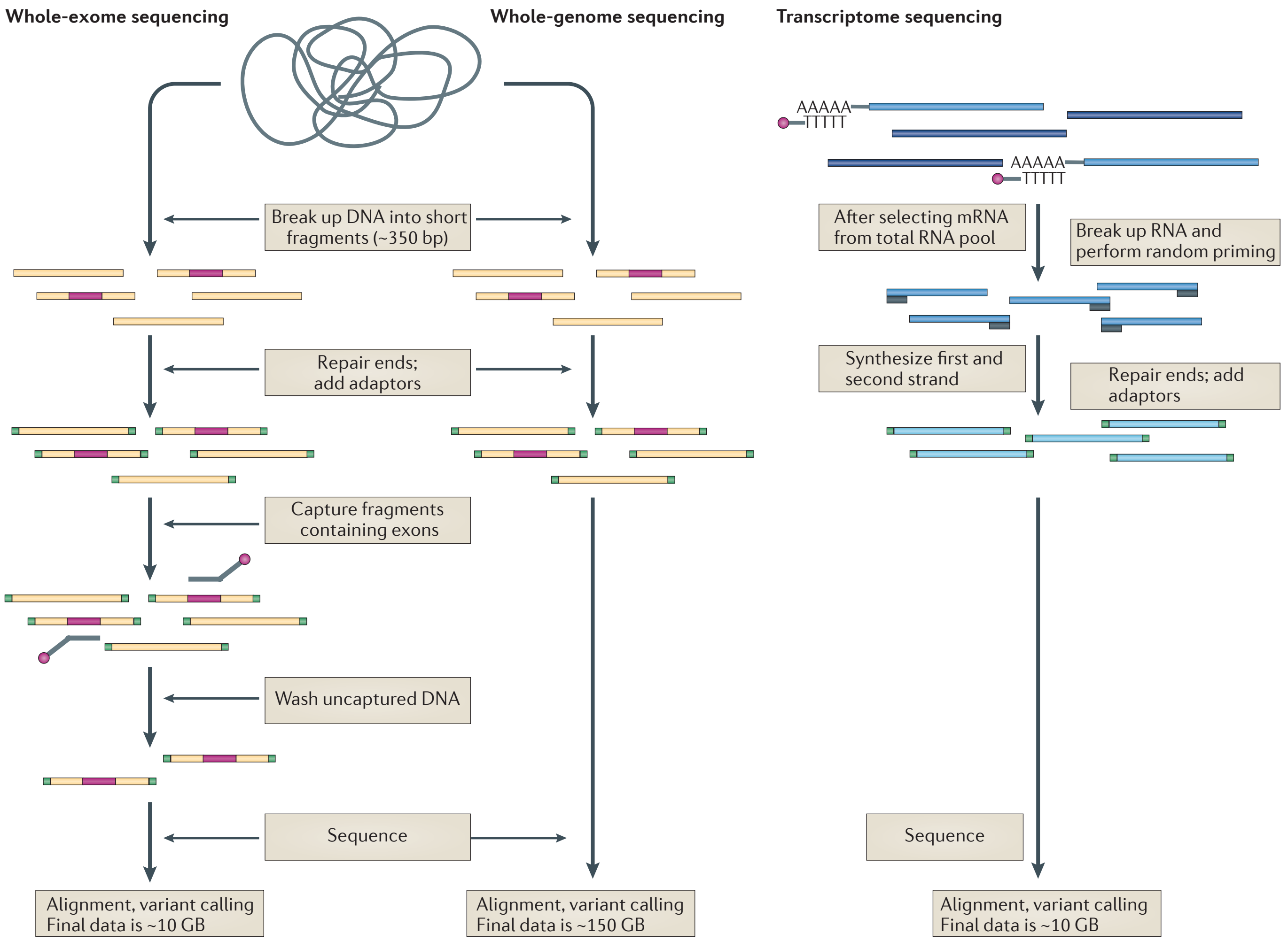

Whole-exome sequencing
Whole-genome sequencing
Transcriptome sequencing
AAAAA
TTTTT
AAAAA
TTTTT
Break up DNA into short fragments (~350 bp)
After selecting mRNA from total RNA pool
Break up RNA and perform random priming
Repair ends; add adaptors
Synthesize first and second strand
Repair ends; add adaptors
Capture fragments containing exons
Wash uncaptured DNA
Sequence
Sequence
Alignment, variant calling Final data is ~10 GB
Alignment, variant calling Final data is ~150 GB
Alignment, variant calling Final data is ~10 GB

Figure 2 | **Simplified workflows for whole-exome, whole-genome and transcriptome sequencing.** The initial sample preparation is identical for both whole-exome and whole-genome sequencing. Genomic DNA is broken up into small fragments and sequence adaptors, which allow each fragment to be hybridized to the flowcell where the sequencing occurs, are added. Whole-exome sequencing protocols proceed with the hybridization of the fragments to probes that are complimentary to all the known exons in the genome, which are then captured while the remaining DNA is washed away, leaving a pool of fragments containing exons. Whole-genome sequencing requires no extra steps following the addition of adaptors and the library is ready to be sequenced at that point. For transcriptome sequencing, the procedure is identical to the other two protocols, with the exception of the initial stages of sample preparation. Here, it is customary to start with a pool of total RNA, from which mRNA is captured and then sheared and finally cDNA is synthesized. At this step, the preparation of the library and sequencing follows the same general procedures as for the two other protocols.

Most published GWASs perform individual association tests for each of hundreds of thousands, or even millions, of genetic markers. Rarely is a GWAS published in which the authors have explicitly examined interactions between particular loci. For example, in Parkinson's disease both *SNCA* (the gene encoding α-synuclein) and microtubule-associated protein tau (*MAPT*) are known risk loci; however, whether their effects are additive or multiplicative is not yet known[21], and this is primarily because of a lack of statistical power to detect such events. As the number of tested markers increases, so does the potential for false-positive associations to be detected between markers. This problem arises as a result of the simultaneous comparison of many millions of possible pairs of genetic markers with disease. A GWAS typically tests about 5×10^5 primary markers for association with disease. This means that to declare statistical significance at the equivalent of $P = 0.01$ a nominal P value of 2×10^{-18} must be achieved. When studies look for epistatic interactions, they are thus testing $(5 \times 10^5)^2$ possible combinations for association with disease: an unfeasibly large number of statistical tests. Clearly, in order to detect epistatic interactions, the number of tests performed must be restricted, either by analysing loci known to be in one pathway or by testing only loci that have previously been shown to be independently associated with disease. However, it is possible that a proportion of as yet unaccounted for risk may result from epistatic interactions; that is, risk caused by gene combinations and not by variants acting on their own[28].

Epistatic interactions
Events that occur when the effects of one gene are modulated by one or several other independent genes.

© 2012 Macmillan Publishers Limited. All rights reserved

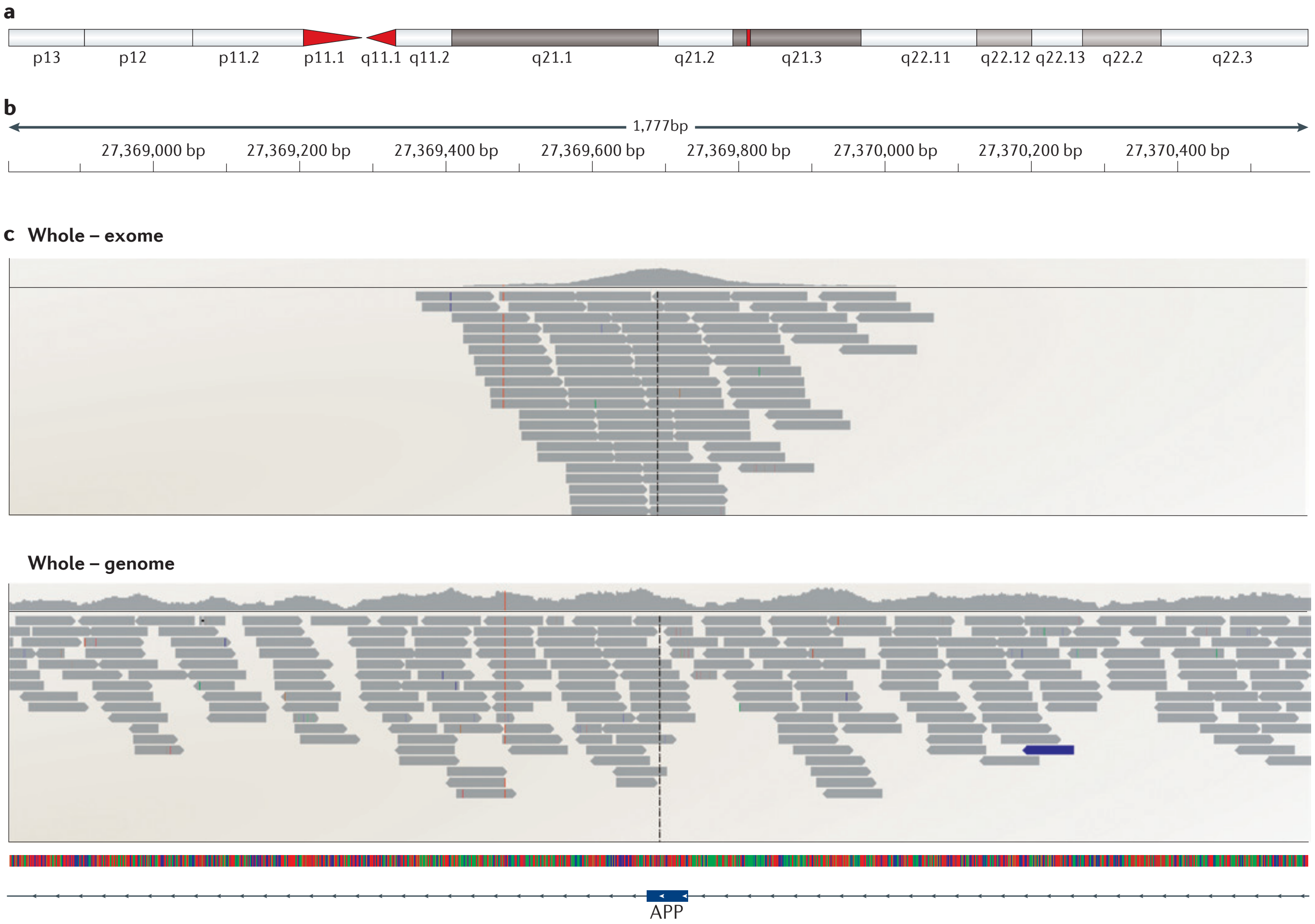


Figure 3 | **Comparison of whole-exome and whole-genome sequencing results.** Typical results from experiments of whole-exome sequencing and whole-genome sequencing are depicted, focusing on one exon of the amyloid precursor protein (*APP*) gene. **a** | The structure of chromosome 21 is represented, with the red vertical bar pinpointing the *APP* locus. **b** | The total size of this region of the chromosome shown is 1,777 bp and the positions of specific chromosomal base pairs within this region are shown. **c** | Each grey arrow represents an individual sequencing read (upper panel, whole-exome sequencing; lower panel, whole-genome sequencing) and shows the sequencing orientation. The blue arrow represents a read that has low mapping quality (that is, it could have mapped to other positions in the genome). The *APP* exon is identified as a blue block at the very bottom of the figure. For whole-exome sequencing it can easily be seen that sequencing reads cluster in the region of the exon and that there is very little coverage outside the exon. For whole-genome sequencing, reads are evenly dispersed throughout the whole region. Both experiments show the coverage (total number of reads) at every locus. At the bottom of the figure, the reference genome sequence for this region is shown. Individual bases are identified by colour (green represents adenine, red represents thymine, orange represents guanine and blue represents cytosine). For each of the reads in the upper two panels, those that agree with the human reference genome (and thus do not have variants) are coloured in grey; if a variant is present, that base is coloured in accordance with the base seen at that position. A non-coding variant base that is present in all individual reads can be seen just after the exon in both panels (shown in red, indicating that the sequencing identified a thymine at this position).

As study sizes continue to increase, and novel statistical methods are developed, it is likely that these epistatic interactions and their effects will ultimately be discovered.

Two major factors govern the selection of markers for GWASs. First, they need to be common (that is, there must be a high minor allele frequency) and second, they need to mirror the variability in surrounding markers so that a smaller number of tests captures a larger portion of the variability (BOX 1). This has two major implications for data that are generated from GWASs: variants with a smaller frequency in the population are not assayed and results are usually translated into regions of association containing several genes (FIG. 4). Additionally, as testing is being performed on common variants, the odds ratios of these variants are low; even when several variants are combined in risk-prediction models they usually still yield only a modest effect. These two aspects of GWASs have led to the development, and continuous refinement, of imputation techniques[29], in which a reference panel of whole-genome sequenced samples is used to estimate the genotypes at positions not assayed in the study samples. Owing to the effect of linkage disequilibrium (BOX 1), it is possible to estimate what genotype

Odds ratios
Measures of effect size, defined as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. In the context of a genetic-association study, this might be the odds of major depression occurring in one genotype group against the odds of it occurring in another genotype group.

© 2012 Macmillan Publishers Limited. All rights reserved

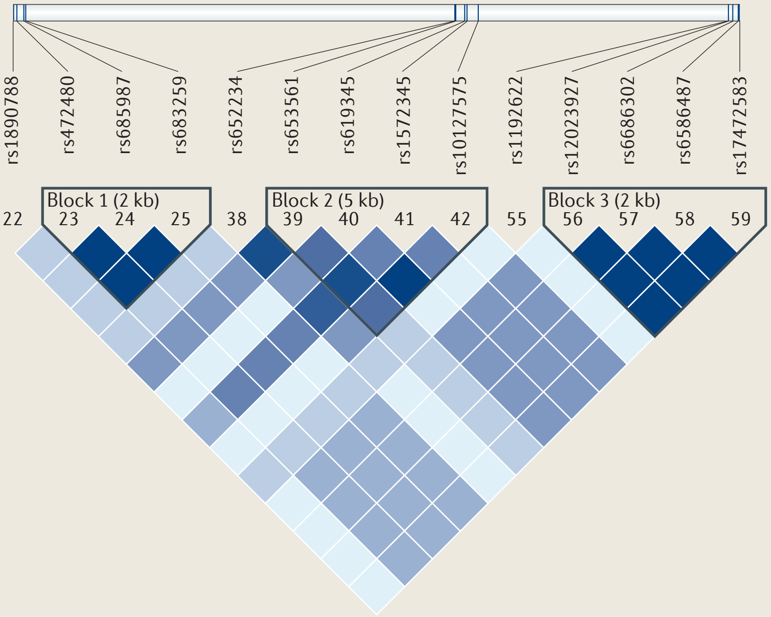
Box 1 | Linkage disequilibrium and imputation

Linkage disequilibrium occurs when genetic markers are inherited together with higher frequency than would be expected if they were inherited randomly. This generally occurs when the markers are located close together on a chromosome and are therefore less likely to be separated by meiotic recombination.

Linkage disequilibrium has had a great impact on genome-wide association study (GWAS) design and analysis. We can consider, for example, two markers — SNP 1 and SNP 2, each with two possible alleles, A or C for SNP 1 and T or G for SNP 2 — that are in complete linkage disequilibrium with each other. In this case, when SNP 1 presents the A allele, SNP 2 will present the G allele. This allows one to test only SNP 1 and automatically obtain information for SNP 2 without directly assessing it. This has clear implications for GWAS design because it greatly reduces the number of markers needed for such a study, while allowing the same amount of information to be obtained.

The figure illustrates another example of this principle. A region of chromosome 1 is shown at the top, with the relative positions of several genetic markers indicated. In the lower part of the figure the squares illustrate the degree to which pairs of markers are in linkage. Squares that are darker shades of blue represent complete linkage between each pair of markers, whereas lighter shades represent lower levels of linkage. Having information on marker 59 (corresponding to rs17472583), for example, will be informative for markers 56–58, as they are in complete linkage with each other.

A further application of this principle comes from imputation analysis in which a densely genotyped or deep-sequenced reference panel is used to increase the number of markers known in a study population. This is done by matching the markers that are in common between the two cohorts and then filling in the blanks of the missing (not genotyped) markers in the study cohort with the information from the reference panel. This approach relies heavily on the principle of linkage disequilibrium.



Genotyping arrays
These are a type of DNA microarray that are used to detect polymorphisms in DNA samples.

Single-nucleotide polymorphisms
(SNPs). The most common form of variation in human DNA sequences. They occur when a single nucleotide (for example, thymine) replaces one of the other three nucleotides (for example, cytosine).

will be present in positions surrounding assayed markers. This approach has been used in several studies with two main goals. First, when pooling the results of several studies performed on different platforms, imputation is usually carried out before the merging of the data sets to increase the number of overlapping markers. Second, when a region of interest is defined, imputation helps in fine-mapping the region to identify a larger number of markers. Two examples of this approach are the imputation of a large meta-analysis in Parkinson's disease that successfully performed fine-mapping of all the associated genomic regions and the imputation of the Wellcome Trust Case Control Consortium phase 1 data that yielded novel associations between the genes interleukin 2 receptor, alpha (*IL2RA*) and cyclin-dependent kinase inhibitor 2B (*CDKN2B*) and type 1 and type 2 diabetes, respectively[22,30].

Identifying rare, high-risk variants. Between rare, fully penetrant mutations causing Mendelian diseases and common variants that modulate risk (albeit with low impact) are rare, high-risk variants. These are variants that do not directly cause disease and have a low frequency in the population (minor allele frequencies of 0.5–3%), but have a higher impact on the risk for the development of disease than common variants. To study variability at this level of frequency, it is necessary to identify and assess these variants in large numbers of individuals, because the variants are, by definition, rare. Thus, the study of such variability and its impact on phenotype had largely been unfeasible until the recent advances in sequencing technology.

Now that a map of risk-conferring common variability has been established for most complex neurological disorders, what is lacking is a similar record of rarer but potentially higher risk variability. In addition to the two most common uses of imputation described above, imputed genotypes may also be used to identify higher risk loci from GWASs. Given that genotyping platforms only assay high-frequency markers, and therefore low-risk markers are missed, imputation allows for the *in silico* analysis of rarer genotypes that potentially confer higher risk; the only requirement being that these rarer markers be present in the reference population used to calculate the imputed genotypes.

Imputation does, however, have several limitations that ultimately prevent it from being a suitable approach to identify rare, high-risk variants. Most notably, imputed genotypes are a probable genotype rather than a certain genotype. Imputation works by comparing the sample (in which a subset of markers have been assessed) with a reference panel (in which a higher number of markers has been assessed) (BOX 1). Thus, genotypes obtained from imputation represent only the most likely genotype at that specific position based on what is known from the reference panel. Additionally, if the reference genotypes are not well matched with the sample, imputation will provide genotypes that are potentially wrong. An extreme example would be to impute a cohort of Caucasian samples using a reference panel comprising samples from individuals of an Asian background. Thus, to identify rare variants, imputation may not be the most appropriate procedure, particularly if a robust reference panel has not been created for the population in study. Currently, two main approaches are being used to address this issue. Custom genotyping arrays containing known rare variants are being designed and applied to different populations, usually as part of a larger array containing probes for both common and rarer variants (for example, the latest Illumina Infinium Beadchip contains ~4.3 million markers, with space available for an additional 500,000 custom single-nucleotide polymorphisms (SNPs)). In addition, exome sequencing studies have been widely applied to a number of diseases to

© 2012 Macmillan Publishers Limited. All rights reserved

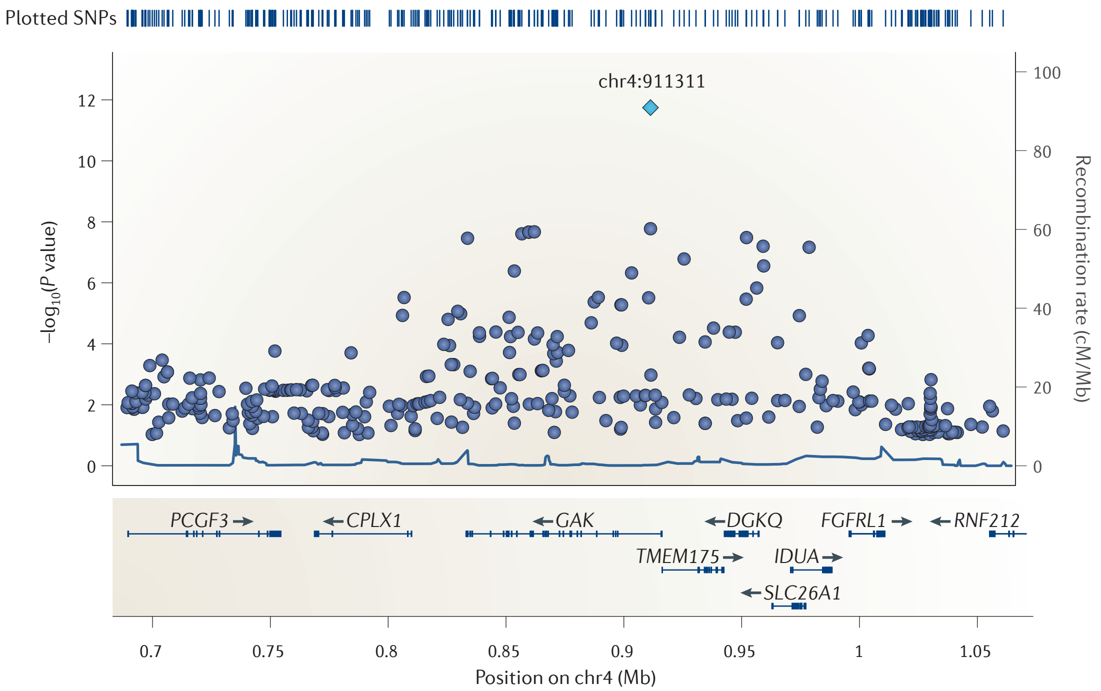


Figure 4 | **Many genes at a single genome-wide association study locus.** Genome-wide association study (GWAS) results showing a region of association encompassing several genes on chromosome 4 (chr4). Each dot represents an individual marker, plotted according to its position on the chromosome on the *x*-axis and the *P* value of its association with the disease on the *y*-axis. The recombination rate, based on the HapMap CEU population, is plotted underneath the markers, to aid in identifying blocks of linkage disequilibrium (recombination will be low within these blocks, but high in regions surrounding them). This example is from a recent study in Parkinson's disease[22], but is typical of many GWAS results. In addition, it shows how difficult it can be to pinpoint an association to a single gene based on GWAS alone. It is clear that significant associations are being detected not only at the cyclin G associated kinase (*GAK*) locus, but at surrounding genes as well. The top-associated marker at this locus is represented by a diamond-shaped symbol and is located in chromosome 4 at the genomic position 911,311 bp. *CPLX1*, complexin 1; *DGKQ*, diacylglycerol kinase theta 110 kDa; *FGFRL1*, fibroblast growth factor receptor-like 1; *IDUA*, iduronidase, alpha-L; *PCGF3*, polycomb group ring finger 3; *RNF212*, ring finger protein 212; *SLC26A1*, solute carrier family 26 (sulphate transporter), member 1; SNPs, single-nucleotide polymorphisms; *TMEM175*, transmembrane protein 175.

identify rare variants responsible for, or at least involved in, disease onset. The combination of whole-exome sequencing with custom arrays containing rare variants has the potential to lead to the creation of detailed maps of rare variability in large numbers of individuals in an affordable manner.

Exome sequencing has rapidly become the *de facto* approach used to study rare variation in different diseases, mainly because it is straightforward to perform and because costs have fallen dramatically in the short period of time since it became available. Nonetheless, exome sequencing is not without limitations. For instance, not all known genes are properly captured and sequenced by this approach: those that are excluded may harbour repetitive regions, have highly homologous sequences elsewhere in the genome or be located in GC-rich regions[31]. Repetitive regions and homologous sequences mean that mapping the resulting reads back to the genome is difficult, whereas GC-rich regions cause difficulties during sample preparation, which involves a polymerase chain reaction (PCR) step (FIG. 2). Similarly, exome sequencing can only capture what we know exists — the 'known knowns' — and does not capture the 'unknown unknowns'.

The consequences of not capturing all the known genes can be illustrated by using the examples of the genes *GBA* (glucosidase, beta, acid) and *CR1* (complement component receptor 1). Homozygous mutations in *GBA* are a known and well-described cause of the lysosomal storage disorder Gaucher's disease. Recently, it was found that the same mutations, when heterozygous, impart risk for developing Parkinson's disease[32]. Heterozygous *GBA* mutations are the strongest genetic risk factor for developing Parkinson's disease known so far, increasing the risk for developing the disease by about fivefold. Common variability in *CR1* was recently identified as a risk factor for Alzheimer's disease in a large GWAS[33]. These results suggest that performing sequencing in additional cohorts could identify additional *GBA* carriers (further elucidating the role of these mutations in Parkinson's disease) and clarify the association of *CR1* with Alzheimer's disease by fine mapping this genomic region. However, both genes pose difficulties for sequencing. *GBA* has a pseudogene with ~96% homology just a few kilobases downstream in chromosome 1; the existence of such a similar pseudogene makes it difficult to determine the source of the sequenced DNA fragments, particularly when dealing

© 2012 Macmillan Publishers Limited. All rights reserved

with short sequencing reads. *CR1* has a variable number of repeated exons, and so knowing the origin of a sequencing fragment is also a potential problem.

Selectively sequencing the exome — which is, to our knowledge, the most likely region of the genome to contain pathogenic mutations — also excludes non-coding regions. The genuine contribution of non-coding regions to disease mechanism is still to be determined. For example, an intronic hexanucleotide repeat in chromosome 9 open reading frame 72 was recently identified as the cause of amyotrophic lateral sclerosis and frontotemporal dementia[34,35]. This single mutation accounts for a considerable number of cases with these disorders, and it would probably have been missed if the sample was studied by exome sequencing alone: it lies in an intronic region, cannot be assayed by PCR because of its size and is also a repeat expansion (current sequencing technology relies on short fragments and so it is difficult to accurately map reads containing a low complexity of nucleotides, such as repeat sequences, back to the reference genome). Similarly, all polyglutamine-type diseases are also difficult to study by exome sequencing, as the underlying defect is an increased repeat expansion. Additionally, a recent finding in Usher syndrome indicates that a deep intronic mutation causes the activation of a pseudoexon, which is ultimately the cause of the disease[36]. How many more of these unusual mutations will be found to be disease-causing remains to be determined.

A third mutation mechanism that exome sequencing does not currently address is copy number variation. For example, we know that a fraction of cases of Parkinson's disease and Alzheimer's disease are caused by the multiplication or deletion of a number of genes or of portions of genes[37–40]. Exome sequencing is not the most adequate technology to detect germline copy number variation because it relies on PCR-based sample preparation methods. Copy number variants are very difficult to identify after a sample has undergone PCR-based amplification, as all targets will eventually achieve similar concentrations. It is therefore likely that many of the copy number variations discovered so far would have been missed if they were exome sequenced. It should be noted that an integrative approach using both genotyping arrays and sequencing would detect such variations, albeit with a considerable increase in cost.

Despite the limitations described above, whole-exome sequencing is a powerful technology, as illustrated by the large number of publications describing novel genes or mutations for several syndromes (TABLE 1; see Supplementary information S1 (table)). It has a high success rate for identifying disease-causing point mutations or short insertions or deletions (insertions or deletions that can fit within a read are usually detected with high confidence). One recent example of this success rate comes from the findings that recessive mutations in WD repeat domain 62 (*WDR62*) are a cause of a wide spectrum of severe cerebral cortical malformations, a study performed using small kindreds that would not be suitable for other approaches[41]. Similarly, mutations in colony stimulating factor 1 receptor (*CSF1R*) were recently identified as a common cause of hereditary diffuse leukoencephalopathy with spheroids[42]. These two examples show the power that exome sequencing has to identify pathogenic mutations in the presence of small kindreds and locus or disease heterogeneity.

As sequencing costs continue to decrease, a growing number of groups are using whole-genome sequencing instead of whole-exome sequencing. The main difference between the two approaches is that for whole-genome sequencing a capture step is not necessary (FIG. 2). This means that fewer biases are introduced into the sample (due to extra amplification steps, for example), but also that more data need to be generated in order to achieve adequate coverage of the genome. One of the most well-known results derived from whole-genome sequencing is the identification of compound heterozygous mutations in *SH3CT2* (SH3 domain and tetratricopeptide repeats 2) as a cause of Charcot–Marie–Tooth disease[12]. Whole-genome sequencing, therefore, seems to be the optimal approach for the discovery of rare, high-risk variants, as it provides more information than exome sequencing and it is probable that, in the near future, it will become the standard method for these studies. Nevertheless, it shares most of the benefits and limitations with exome sequencing, as they are based on the same sequencing technology, such as the use of short reads, slightly high error rates and difficulty in capturing mid-size insertions or deletions.

Recent developments have also facilitated the use of genotyping arrays that are directed to rarer variants. These are ordinary genotyping chips in which variants have been selected because they have a frequency in specific cohorts (a disease group, for example). After gathering a large number of these rare variants, observed in a particular group of samples, these are then assessed in very large numbers of cases. This approach allows a more thorough examination of the rarer variability, without the need to exome sequence large cohorts, which would be substantially more expensive and time consuming. It should be noted, however, that such an exome–chip approach will miss private mutations (extremely rare mutations occurring within a single family), and, in these cases, sequencing is still the optimal option.

In summary, current generation sequencing, in both the form of whole-genome and whole-exome sequencing, has revealed not only a significant number of novel genes involved in a variety of phenotypes, but also rare variability, the impact of which on a phenotype awaits for larger cohorts of samples to be gathered and genotyped. These discoveries show how powerful these approaches are in identifying genetic defects, and we should continue to see reports of rare mutations giving rise to either novel or atypical phenotypes.

Interpreting findings

Fully dissecting risk at a locus. One of the major problems associated with the discovery of genetic risk-associated markers for a disease is the interpretation of such risk in the context of disease pathobiology. Results from GWASs consist of single markers that have a different

Pseudoexon
A fragment of DNA that has characteristics of an exon, but plays no part in splicing events and thus does not code for a protein sequence.

Copy number variation
A change in the normal number of copies of a given gene/loci. Usually, there are two copies of each locus, but if, for example, duplications or triplications occur the number of copies will increase.

Point mutation
A change in one single nucleotide that occurs very rarely in the population.

© 2012 Macmillan Publishers Limited. All rights reserved

Table 1 | **Examples of genome-wide approaches used in neurological disorders**

Disease	Gene (locus)	Mutation(s)	Methodology	Refs
Parkinson's disease	*VPS35*	Heterozygous p.Asp620Asn	Exome sequencing	15,16
	SYT11, ACMSD, STK39, MCCC1/LAMP3; GAK, BST1, SNCA, HLA-DRB5, LRRK2, CCDC62/HIP1R, MAPT	NA	Genome-wide association study	22
	PARK16 (1q32), *STBD1* (4q21), *GPNMB* (7p15), *FGF20* (8p22), *STX1B* (16p11)	NA	Genome-wide association study	23
	SCARB2, SREBF1/RAI1	NA	Genome-wide association study	62
Alzheimer's disease	*ABCA7, MS4A6A/MS4A4E, EPHA1, CD33, CD2AP*	NA	Genome-wide association study	63,64
	CLU, CR1, PICALM	NA	Genome-wide association study	26,33
Stroke	*NINJ2–WNK1*	NA	Genome wide-association study	65
	HDAC9	NA	Genome wide-association study	66
Spinocerebellar ataxia 15	*ITPR1*	201 kb deletion	Large-scale mutagenesis combined with whole-genome genotyping and copy-number variants analysis	67
Autosomal-recessive cerebellar ataxia	*SYT14*	Homozygous p.Gly484Asp	Combination of homozygosity analysis and exome sequencing	68
Dystonia-parkinsonism disorder	*PRKRA*	Homozygous p.P222L	Autozygosity analysis followed by Sanger DNA sequencing	69

ABCA7, ATP-binding cassette, sub-family A (ABC1), member 7; *ACMSD*, aminocarboxymuconate semialdehyde decarboxylase; *BST1*, bone marrow stromal cell antigen 1; *CCDC62*, coiled–coil domain containing 62; *CD2AP*, CD2-associated protein; *CD33*, CD33 molecule; *CLU*, clusterin; *CR1*, complement component (3b/4b) receptor 1; *EPHA1*, EPH receptor A1; *FGF20*, fibroblast growth factor 20; *GAK*, cyclin G-associated kinase; *GPNMB*, glycoprotein (transmembrane) nmb; *HDAC9*, histone deacetylase 9; *HIP1R*, huntingtin interacting protein 1 related; *HLA-DRB5*, major histocompatibility complex, class II, DR beta 5; *ITPR1*, inositol 1,4,5-trisphosphate receptor 1; *LAMP3*, lysosomal-associated membrane protein 3; *LRRK2* leucine-rich repeat kinase 2; *MAPT*, microtubule-associated protein tau; *MCCC1*, methylcrotonoyl-CoA carboxylase 1; *MS4A4E*, membrane-spanning 4-domains, subfamily A, member 4E; *MS4A6A*, membrane-spanning 4-domains, subfamily A, member 6A; NA, not applicable; *NINJ2*, ninjurin 2; *PARK16*, Parkinson's disease 16; *PICALM*, phosphatidylinositol binding clathrin assembly protein; *PRKRA*, protein kinase, interferon-inducible double stranded RNA dependent activator; *RAI1*, retinoic acid-induced 1; *SCARB2*, scavenger receptor class B, member 2; *SNCA*, synuclein-α; *SREBF1*, sterol regulatory element binding transcription factor 1; *STBD1*, starch binding domain 1; *STK39*, serine threonine kinase 39; *STX1B*, syntaxin 1B; *SYT11*, synaptotagmin XI; *SYT14*, synaptotagmin XIV; *VPS35*, vacuolar protein sorting-associated protein 35; *WINK1*, WNK lysine deficient protein kinase 1.

allelic frequency in groups of cases and controls and these are usually located in areas of the genome for which the downstream effects of that variability are not obvious (such as intergenic regions). Thus, dissecting the biological meaning of the top-ranked results for a GWAS is usually not a trivial matter.

One possible approach to resolve this issue is to integrate genotyping, sequencing and expression data in such a manner that biological meaning can be inferred from the association between SNPs. This can be accomplished with relative ease if there is access to genome-wide expression and splicing databases that are derived from individuals for the tissue of interest, such as those previously published[43,44]. For diseases in which cell death is a major component of the pathobiological events, expression and splicing should ideally be assessed in samples that are derived from healthy individuals to avoid the detection of changes caused by cell loss. Having access to specific tissue regions instead of whole-brain homogenates is also important, as different brain regions have different expression patterns.

The integration of genotyping, sequencing and expression data has already been attempted in some cases and its impact on our understanding of disease mechanisms has been substantial. For instance, for Parkinson's disease, the risk identified at the *MAPT* locus was shown to be associated with an increase in the tau protein expression in frontal cortex samples[21]. This example highlights the straightforwardness of this approach: all that is required to identify associations between allele load and protein expression levels is for samples to be assessed in a genome-wide genotyping array and the tissue of interest obtained for the study of gene expression. Alleles showing an association between allele load and expression levels are commonly called expression quantitative trait loci (eQTL); when the association is between allele load and splicing events they are known as splicing QTL (sQTL).

Nonetheless, not all disease associations translate this clearly into expression changes, and at times no QTL can be determined from these studies. It is likely that in some of these cases the issue resides in the expression array used: what is tested is a subset of the known

© 2012 Macmillan Publishers Limited. All rights reserved

transcripts rather than all the present transcripts, and it is therefore possible that additional transcripts that are QTL for those associations have yet to be discovered. One way to address this issue would be to perform transcriptome sequencing (FIG. 2) and to integrate these data with genotyping results. Indeed, genotyping data is usually confined to the most common variants described in the Caucasian population, which suggests that if we are to identify QTL for a particular trait, whole-genome and transcriptome sequencing are the most suitable methods to use to achieve optimal results. However, it should also be noted that QTL might not only occur in tissue-specific or cell-specific manner, but also in a time-specific manner. This suggests that even if we can identify all the genetic variabilities and assess expression and splicing in a comprehensive manner, sample selection is of primary importance.

Our understanding of sample requirements has seen major developments over the past few years for all of the technologies described above, and is anticipated to continue to improve. Whole-genome genotyping is the most forgiving technique in terms of its DNA requirements. It requires a relatively low amount of total DNA (usually around 200–300 ng) and, although it is preferable that this is high-quality DNA, the sample preparation is robust enough for it to handle DNA that is of a less than ideal quality. Whole-exome and whole-genome sequencing not only require a larger amount of DNA (~1–10 µg), but also require that this DNA is of high molecular mass, which may be an issue when the source of DNA is fixed tissue. Similarly, studies performed on RNA, both for expression on arrays and for RNA sequencing, have identical requirements in terms of quality. Formalin-fixed paraffin-embedded tissue samples are a common source of RNA for this type of experiment and it is not trivial to obtain good-quality RNA from such a source. These sample requirements mean that for a proportion of the samples that have been stored for many years in laboratories, it will not be possible to assay them using these technologies. It is therefore essential that improvements to sample preparation methods continue to be developed so that these samples can eventually be studied.

Pathway-based analysis. A central goal of most genetic studies is to gain an understanding of the pathobiological mechanisms involved in disease onset or modulation. This is a difficult goal to achieve as the data that can currently be generated is far from comprehensive: genome-wide genotyping directly assesses only genetic markers and does not directly identify a single gene, whereas exome sequencing targets only the coding portion of the genome.

One potential way in which the results obtained with these methods can be integrated with the biological understanding of disease processes is by pathway-based analysis. In this approach, the aim is to identify multiple associated genes that affect one biological pathway, yielding information not only on that particular pathway's involvement in the disease, but also suggesting other potential risk-conferring genes[45]. For example, rather than focusing on individual loci that show an association, it is possible to rank all genes based on their association values and then interrogate whether a particular cluster of genes known to be involved in the same pathway is overrepresented in that list. Similar studies have been performed with expression data. In one such study that tested the expression levels of 22,000 genes individually, no single gene showed a statistically significant difference in expression after adjustment for multiple testing. However, when a pathway-based analysis was performed on the same data set, it was clear that a group of peroxisome proliferator-activated receptor-γ, coactivator 1 (*PGC1A*; also known as *PPARGC1A*)-responsive genes showed consistent changes in expression levels in muscle samples from subjects with diabetes[46].

This approach has been applied to several neurological diseases and perhaps the most notable result has come from the field of Alzheimer's disease. Using the two largest GWASs in Alzheimer's disease, it was shown that there is a considerable overrepresentation of disease-associated genes in pathways related to cholesterol metabolism and the immune response, thereby suggesting that these are pivotal pathways for this disease biology[47]. However, the findings obtained from such approaches should be considered with care, not only because of the limitations of the genetic data that can currently be produced, but also because the understanding of complete biological pathways and their interactions is still far from complete.

Conclusions and outlook

New genome-wide approaches have undoubtedly changed the field of genetics of human disease. Through the use of these techniques, detailed maps of genetic variation influencing disease have already been created for several diseases. In conjunction with these, maps of expression, splicing and methylation are also anticipated to be available soon. With the continuing decrease in sequencing cost, we can certainly expect that these maps of genomic variability will become even more detailed and complete. It is clear that research in human genetics will shift towards DNA sequencing; this is already evident in the increasing number of exome sequencing studies being published. Gradually, as the technology continues to evolve and costs continue to lower, the field is anticipated to move towards whole-genome sequencing, which is considered by many to be the holy grail of genetics. Nonetheless, to fully understand the effects of new genetic variability, a multi-pronged approach must be used that encompasses not only DNA sequencing, but also transcriptomics, proteomics and epigenomics (BOX 2). It is only with the simultaneous study of DNA, RNA, protein and their interactions with each other that the effects of genetic variability will become evident.

There are still significant limitations associated with the current generation of sequencing technology. Error rates are still high, which suggests that the direct introduction of sequencing to diagnostics may have to wait for improved approaches, as results need to be robust in this setting. Next-generation sequencing utilizes short reads, which are difficult to map back to the reference genome, cause problems when dealing with low complexity

© 2012 Macmillan Publishers Limited. All rights reserved

Genetic phase
Refers to the allelic combinations that an individual received from its parents. If two alleles originated from the same parent they are said to be in *cis* phase. If each allele originated from a different parent they are said to be in *trans* phase.

Box 2 | **Epigenetics**

A small number of studies have performed detailed analysis of epigenetic events and their role in neurological disease. Epigenetics broadly refers to any change in phenotype that is caused by mechanisms other than changes in the underlying DNA sequence. Most of these studies have examined methylation patterns and their role in gene expression and ultimately phenotype[49–51]. For example, one study evaluated how methylation changes are associated with chronological age in the human brain. Using four separate brain regions from nearly 400 donors, they identified several loci that showed a highly statistical significant and consistent correlation between DNA methylation and chronological age[50]. Another study showed that different levels of methylation of the ataxin 2 (*ATXN2*) gene promoter were associated with disease development in a family with spinocerebellar ataxia type 2[51]. There are, however, difficulties in performing such studies. Perhaps the most notable are that methods for analysis of genome-wide methylation are not optimal (the most commonly used approach involves the conversion of methylated cytosine residues using bisulphite and is therefore an indirect measure of methylation) and that an improved understanding of the transcriptome is needed in order to fully appreciate the impact of methylation[52–54].

Gene–environment interactions can also play a role in disease development[55,56]. These interactions are thought to be mediated by epigenetic modifications of the genome, and epigenetic changes of the genome often arise in response to changes in the environment[57,58]. This is a particularly difficult field of study, given the obvious problems posed by the study of various environmental factors in small numbers of individuals. It is plausible that the environment plays a role in modulating the phenotype for a range of neurological diseases (indeed, this has been shown clearly for Parkinson's disease[59–61]), but the study of such phenomena, in large enough numbers of samples for statistical significance to be achieved, is complex.

regions, and suggest that genetic phase cannot be readily established (that is, when two variants have been identified in one gene in an individual it is difficult to determine whether those two variants are in the same copy of the gene or on different copies of the gene). Furthermore, despite great improvements in sample preparation techniques and whole-genome sequencing, these processes are still moderately time-consuming.

When these limitations are addressed and whole-genome sequencing becomes a valid and widespread tool in clinical applications, a complete overhaul of how patients consent to testing will be required. This is because performing whole-genome sequencing will yield information not only about the medical problem at hand, but also about potential predisposition to conditions in the future (and this same potential predisposition in family members). Thus, the full implementation of whole-genome sequencing will necessitate a detailed understanding of genetics by health providers and ultimately by patients.

One of the pioneering projects that has attempted to bring these new technologies to a diagnostic space was established by the US National Institutes of Health. The Undiagnosed Diseases Program aimed to enrol individuals in whom traditional diagnostic approaches had failed to identify the underlying cause of the disease. In the first year of the project, 160 individuals were enrolled and a diagnosis was obtained for 39 of these. Of the entire cohort, 53% had a neurological disorder, which demonstrates the potential of these new technologies in this type of pathology but also grants high expectations for the results coming in the near future from this project[48].

It is clear that recent whole-genome strategies have greatly improved our knowledge and understanding of human disease. We are now in the enviable position of being able to assess how common variability plays a role in disease, detect Mendelian mutations in segregating families with ease and detect whether such variability regulates transcriptional events.

1. Hardy, J. The real problem in association studies. *Am. J. Med. Genet.* **114**, 253 (2002).
2. Paisan-Ruiz, C. *et al.* Cloning of the gene containing mutations that cause *PARK8*-linked Parkinson's disease. *Neuron* **44**, 595–600 (2004).
3. Paisan-Ruiz, C. *et al.* Characterization of *PLA2G6* as a locus for dystonia-parkinsonism. *Ann. Neurol.* **65**, 19–23 (2009).
4. Verkerk, A. J. *et al.* Mutation in the *AP4M1* gene provides a model for neuroaxonal injury in cerebral palsy. *Am. J. Hum. Genet.* **85**, 40–52 (2009).
5. Paisan-Ruiz, C. *et al.* Early-onset L-dopa-responsive parkinsonism with pyramidal signs due to *ATP13A2*, *PLA2G6*, *FBXO7* and spatacsin mutations. *Mov. Disord.* **25**, 1791–1800 (2010).
6. Gulsuner, S. *et al.* Homozygosity mapping and targeted genomic sequencing reveal the gene responsible for cerebellar hypoplasia and quadrupedal locomotion in a consanguineous kindred. *Genome Res.* **21**, 1995–2003 (2011).
7. Nalls, M. A. *et al.* Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10**, 183–190 (2009).
8. Clarimon, J. *et al.* Whole genome analysis in a consanguineous family with early onset Alzheimer's disease. *Neurobiol. Aging* **30**, 1986–1991 (2009).
9. Ng, S. B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nature Genet.* **42**, 790–793 (2010).
10. Pierce, S. B. *et al.* Mutations in the DBP-deficiency protein *HSD17B4* cause ovarian dysgenesis, hearing loss, and ataxia of Perrault syndrome. *Am. J. Hum. Genet.* **87**, 282–288 (2010).
11. Singleton, A. B. Exome sequencing: a transformative technology. *Lancet Neurol.* **10**, 942–946 (2011).
12. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
One of the first studies to use whole-genome sequencing to identify the cause of a neurological disease.
13. Veeramah, K. R. *et al. De novo* pathogenic *SCN8A* mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am. J. Hum. Genet.* **90**, 502–510 (2012).
14. Johnson, J. O. *et al.* Exome sequencing reveals *VCP* mutations as a cause of familial ALS. *Neuron* **68**, 857–864 (2010).
One of the first studies to show a now common finding in exome sequencing studies: genes known to cause one syndrome are often the cause of other unrelated diseases.
15. Zimprich, A. *et al.* A mutation in *VPS35*, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am. J. Hum. Genet.* **89**, 168–175 (2011).
16. Vilarino-Guell, C. *et al. VPS35* mutations in Parkinson disease. *Am. J. Hum. Genet.* **89**, 162–167 (2011).
17. Johnson, J. O., Gibbs, J. R., Van Maldergem, L., Houlden, H. & Singleton, A. B. Exome sequencing in Brown–Vialetto–van Laere syndrome. *Am. J. Hum. Genet.* **87**, 567–569; author reply 569–570 (2010).
18. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.***17**, 502–510 (2001).
19. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
20. Hardy, J. & Singleton, A. Genome-wide association studies and human disease. *N. Engl. J. Med.* **360**, 1759–1768 (2009).
21. Simon-Sanchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nature Genet.* **41**, 1308–1312 (2009).
The first large-scale GWAS in Parkinson's disease. It identified genes, previously known to cause Mendelian disease, that also conferred a risk for developing Parkinson's disease.

© 2012 Macmillan Publishers Limited. All rights reserved

22. Nalls, M. A. *et al.* Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* **377**, 641–649 (2011).
23. Plagnol, V. *et al.* A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet.* **7**, e1002142 (2011).
24. Coon, K. D. *et al.* A high-density whole-genome association study reveals that *APOE* is the major susceptibility gene for sporadic late-onset Alzheimer's disease. *J. Clin. Psychiatry* **68**, 613–618 (2007).
25. Wijsman, E. M. *et al.* Genome-wide association of familial late-onset Alzheimer's disease replicates *BIN1* and *CLU* and nominates *CUGBP2* in interaction with *APOE*. *PLoS Genet.* **7**, e1001308 (2011).
26. Harold, D. *et al.* Genome-wide association study identifies variants at *CLU* and *PICALM* associated with Alzheimer's disease. *Nature Genet.* **41**, 1088–1093 (2009).
27. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
28. Clayton, D. G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet.* **5**, e1000540 (2009).
29. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Rev. Genet.* **11**, 499–511 (2010).
A thorough review of imputation and its application to GWASs.
30. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 data. *Eur. J. Hum. Genet.* 1 Feb 2012 (doi:10.1038/ejhg.2012.3).
31. Sulonen, A. M. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* **12**, R94 (2011).
32. Sidransky, E. *et al.* Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N. Engl. J. Med.* **361**, 1651–1661 (2009).
A study that unequivocally showed that mutations in *GBA*, when heterozygous, confer a risk for developing Parkinson's disease.
33. Lambert, J. C. *et al.* Genome-wide association study identifies variants at *CLU* and *CR1* associated with Alzheimer's disease. *Nature Genet.* **41**, 1094–1099 (2009).
34. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS–FTD. *Neuron* **72**, 257–268 (2011).
35. Dejesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
36. Vache, C. *et al.* Usher syndrome type 2 caused by activation of an *USH2A* pseudoexon: implications for diagnosis and therapy. *Hum. Mutat.* **33**, 104–108 (2012).
37. Abbas, N. *et al.* A wide variety of mutations in the parkin gene are responsible for autosomal recessive parkinsonism in Europe. French Parkinson's disease genetics study group and the European Consortium on genetic susceptibility in Parkinson's disease. *Hum. Mol. Genet.* **8**, 567–574 (1999).
38. Chartier-Harlin, M. C. *et al.* α-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* **364**, 1167–1169 (2004).
39. Singleton, A. B. *et al.* α-synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
40. Rovelet-Lecrux, A. *et al. APP* locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nature Genet.* **38**, 24–26 (2006).
41. Bilguvar, K. *et al.* Whole-exome sequencing identifies recessive *WDR62* mutations in severe brain malformations. *Nature* **467**, 207–210 (2010).
42. Rademakers, R. *et al.* Mutations in the colony stimulating factor 1 receptor (*CSF1R*) gene cause hereditary diffuse leukoencephalopathy with spheroids. *Nature Genet.* **44**, 200–205 (2012).
43. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
A good example of how eQTLs can be assessed in a genome-wide manner to shed light on gene expression patterns in a tissue of interest.
44. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nature Genet.* **39**, 1494–1499 (2007).
45. Holmans, P. *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.* **85**, 13–24 (2009).
46. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34**, 267–273 (2003).
47. Jones, L. *et al.* Genetic evidence implicates the immune system and cholesterol metabolism in the aetiology of Alzheimer's disease. *PLoS ONE* **5**, e13950 (2010).
48. Gahl, W. A. *et al.* The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet.Med.* **14**, 51–59 (2012).
49. Dedeurwaerder, S. *et al.* DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol. Med.* **3**, 726–741 (2011).
50. Hernandez, D. G. *et al.* Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Hum. Mol. Genet.* **20**, 1164–1172 (2011).
These data reveal methylation patterns in the brain that are correlated with age, thus suggesting that this epigenetic mechanism may be involved in healthy ageing and in disease.
51. Laffita-Mesa, J. M. *et al.* Epigenetics DNA methylation in the core ataxin-2 gene promoter: novel physiological and pathological implications. *Hum. Genet.* **131**, 625–638 (2012).
52. Martin-Subero, J. I. & Esteller, M. Profiling epigenetic alterations in disease. *Adv. Exp. Med. Biol.* **711**, 162–177 (2011).
53. Shen, L. & Waterland, R. A. Methods of DNA methylation analysis. *Curr. Opin. Clin. Nutr. Metabol. Care* **10**, 576–581 (2007).
54. Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nature Rev. Genet.* **12**, 641–655 (2011).
55. Tsuang, M. T., Bar, J. L., Stone, W. S. & Faraone, S. V. Gene–environment interactions in mental disorders. *World Psychiatry* **3**, 73–83 (2004).
56. Vance, J. M., Ali, S., Bradley, W. G., Singer, C. & Di Monte, D. A. Gene–environment interactions in Parkinson's disease and other forms of parkinsonism. *Neurotoxicology* **31**, 598–602 (2010).
57. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genet.* **33**, 245–254 (2003).
58. Jirtle, R. L. & Skinner, M. K. Environmental epigenomics and disease susceptibility. *Nature Rev. Genet.* **8**, 253–262 (2007).
59. Wang, A. *et al.* Parkinson's disease risk from ambient exposure to pesticides. *Eur. J. Epidemiol.* **26**, 547–555 (2011).
60. Langston, J. W., Ballard, P., Tetrud, J. W. & Irwin, I. Chronic parkinsonism in humans due to a product of meperidine-analog synthesis. *Science* **219**, 979–980 (1983).
61. Liou, H. H. *et al.* Environmental risk factors and Parkinson's disease: a case–control study in Taiwan. *Neurology* **48**, 1583–1588 (1997).
62. Do, C. B. *et al.* Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet.* **7**, e1002141 (2011).
63. Hollingworth, P. *et al.* Common variants at *ABCA7*, *MS4A6A*/*MS4A4E*, *EPHA1*, *CD33* and *CD2AP* are associated with Alzheimer's disease. *Nature Genet.* **43**, 429–435 (2011).
64. Naj, A. C. *et al.* Common variants at *MS4A4*/*MS4A6E*, *CD2AP*, *CD33* and *EPHA1* are associated with late-onset Alzheimer's disease. *Nature Genet.* **43**, 436–441 (2011).
65. Ikram, M. A. *et al.* Genomewide association studies of stroke. *N. Engl. J. Med.* **360**, 1718–1728 (2009).
66. Bellenguez, C. *et al.* Genome-wide association study identifies a variant in *HDAC9* associated with large vessel ischemic stroke. *Nature Genet.* **44**, 328–333 (2012).
67. van de Leemput, J. *et al.* Deletion at *ITPR1* underlies ataxia in mice and spinocerebellar ataxia 15 in humans. *PLoS Genet.* **3**, e108 (2007).
68. Doi, H. *et al.* Exome sequencing reveals a homozygous *SYT14* mutation in adult-onset, autosomal-recessive spinocerebellar ataxia with psychomotor retardation. *Am. J. Hum. Genet.* **89**, 320–327 (2011).
69. Camargos, S. *et al.* DYT16, a novel young-onset dystonia-parkinsonism disorder: identification of a segregating mutation in the stress-response protein *PRKRA*. *Lancet Neurol.* **7**, 207–215 (2008).

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

John Hardy's homepage: http://www.ucl.ac.uk/ukpdc/principal-investigators/john-hardy

DATABASES

Online Mendelian Inheritance in Man database: http://www.ncbi.nlm.nih.gov/omim

Catalogue of published genome-wide association studies: http://www.genome.gov/gwastudies

SUPPLEMENTARY INFORMATION

See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2012 Macmillan Publishers Limited. All rights reserved