TRANSLATIONAL GENETICS

Exome sequencing as a tool for Mendelian disease gene discovery

Michael J. Bamshad[*‡], *Sarah B. Ng*[‡], *Abigail W. Bigham*[*§], *Holly K. Tabor*[*||], *Mary J. Emond*[¶], *Deborah A. Nickerson*[‡] *and Jay Shendure*[‡]

Abstract | Exome sequencing — the targeted sequencing of the subset of the human genome that is protein coding — is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Over the past 2 years, experimental and analytical approaches relating to exome sequencing have established a rich framework for discovering the genes underlying unsolved Mendelian disorders. Additionally, exome sequencing is being adapted to explore the extent to which rare alleles explain the heritability of complex diseases and health-related traits. These advances also set the stage for applying exome and whole-genome sequencing to facilitate clinical diagnosis and personalized disease-risk profiling.

*Department of Pediatrics, University of Washington, Health Sciences Building RR349, 1959 N.E. Pacific Street, Seattle, Washington 98195-6320, USA.
‡Department of Genome Sciences, University of Washington, Foege Building, S-210 3720 15th Avenue N.E., Seattle, Washington 98195-5065, USA.
§Department of Anthropology, University of Michigan, 222C West Hall, 1085 S. University Avenue, Ann Arbor, Michigan 48014, USA.
||Treuman Katz Center for Pediatric Bioethics, Seattle Children's Research Institute, M/S C9S-6, 1900 Ninth Avenue, Seattle, Washington 98101, USA.
¶Department of Biostatistics, University of Washington, Health Sciences Building F-658, 1959 N.E. Pacific Street, Seattle, Washington 98195-6320, USA.
Correspondence to M.J.B. and J.S.
e-mails: mbamshad@u.washington.edu; shendure@uw.edu
doi:10.1038/nrg3031
Published online 27 September 2011

Elucidation of the genetic basis of human diseases and other health-related traits has commonly relied on the oversimplified but nevertheless useful dichotomy between 'monogenic, simple and rare' and 'multigenic, complex and common' diseases. Primarily through linkage mapping and candidate gene resequencing, loci underlying about one-half to one-third (~3,000) of all known or suspected Mendelian disorders (for example, cystic fibrosis and sickle cell anaemia) have been discovered[1,2]. However, there is a substantial gap in our knowledge about the genes that cause many rare Mendelian phenotypes. Several factors limit the power of traditional gene-discovery strategies[3]: for example, the availability of only a small number of cases or families to study, reduced penetrance, locus heterogeneity and substantially diminished reproductive fitness. At the other end of the spectrum, genome-wide association studies (GWASs) have identified large numbers of loci that contribute to the genetic basis of complex traits but, in almost all cases, these loci collectively account for only a small fraction of the observed heritability of the trait under study[4,5]. Furthermore, little is known about the extent to which rare alleles contribute to the heritability of complex traits[6].

Since 2005, next-generation DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by four orders of magnitude relative to Sanger sequencing[7]. The development of methods for coupling targeted capture and massively parallel DNA sequencing has made it possible to determine cost-effectively nearly all of the coding variation present in an individual human genome, a process termed 'exome sequencing' (REF. 8) (BOX 1). This technique has become a powerful new approach for identifying genes that underlie Mendelian disorders in circumstances in which conventional approaches have failed (Supplementary information 1 (table))[9–11]. Even where conventional approaches are eventually expected to succeed (for example, in homozygosity mapping), exome sequencing provides a means for accelerating discovery[12].

Our focus is to explain some of the experimental and analytical options for applying exome sequencing as a tool for disease gene discovery and to describe some of the key challenges in using this approach. We review how exome sequencing is being used to identify genes that underlie monogenic disorders using examples from several recent studies. Finally, we provide a brief overview of the application of exome sequencing in clinical diagnostics and describe some of the new manifestations of long-standing ethical issues arising from exome sequencing.

Why exome sequencing?

The exome as a source of rare disease-related variants. Despite the fundamental limitation that exome sequencing does not currently assess the impact of non-coding alleles, it is, for several reasons, a well-justified strategy for discovering rare alleles underlying Mendelian phenotypes and perhaps complex traits as well (BOX 2). First, positional cloning studies that are focused on protein-coding sequences have, when adequately powered, proved to be highly successful at identifying

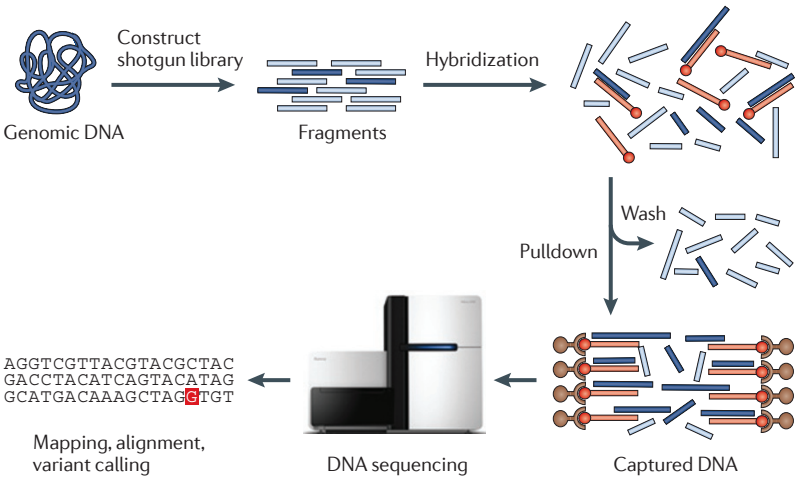
Box 1 | **Workflow for exome sequencing**

Since 2007, there has been tremendous progress in the development of diverse technologies for capturing arbitrary subsets of a mammalian genome at a scale commensurate with that of massively parallel sequencing[8,10,72–79]. To capture all protein-coding sequences, which constitute less than 2% of the human genome, the field has largely converged on the aqueous-phase, capture-by-hybridization approach described below.

The basic steps required for exome sequencing are shown in the figure. Genomic DNA is randomly sheared, and several micrograms are used to construct an *in vitro* shotgun library; the library fragments are flanked by adaptors (not shown). Next, the library is enriched for sequences corresponding to exons (dark blue fragments) by aqueous-phase hybridization capture: the fragments are hybridized to biotinylated DNA or RNA baits (orange fragments) in the presence of blocking oligonucleotides that are complementary to the adaptors (not shown). Recovery of the hybridized fragments by biotin–streptavidin-based pulldown is followed by amplification and massively parallel sequencing of the enriched, amplified library and the mapping and calling of candidate causal variants. Barcodes to allow sample indexing can potentially be introduced during the initial library construction or during post-capture amplification. Key performance parameters include the degree of enrichment, the uniformity with which targets are captured and the molecular complexity of the enriched library.

At least three vendors (Agilent, Illumina and Nimblegen) offer kitted reagents for exome capture. Although there are technical differences between them (for example, Agilent relies on RNA baits, whereas Illumina and Nimblegen use DNA baits — the kits vary in the definition of the exome), we find the performance of these kits to be largely equivalent, and each is generally scalable to 96-plex robotic automation. The fact that the costs of exome sequencing are not directly proportional to the fraction of the genome targeted is a consequence of several factors, including imperfect capture specificity, skewing in the uniformity of target coverage introduced by the capture step and the fixed or added costs that are associated with sample processing (for example, library construction and exome capture). This ratio will fall as the cost of whole-genome sequencing drops.

Although methods for calling single nucleotide substitutions are maturing[80], there is considerable room for improvement in detecting small insertion–deletions and especially copy number changes from short-read exome sequence data[81] (for example, detecting a heterozygous, single-exon deletion with breakpoints that fall within adjacent introns). Exome sequencing also needs improvements of a technical nature. First, input requirements (several micrograms of high-quality DNA) are such that many samples that have already been collected are inaccessible. Protocols using whole-genome amplification or transposase-based library construction offer a solution[82], but additional work is required to fully integrate and validate these methods. Second, as the minimum 'unit' of sequencing of massively parallel sequencing continues to increase, sample indexing with minimal performance loss and minimal crosstalk between samples will be required to lower the costs of exome sequencing. Third, a substantial fraction of the exome (~5–10%, depending on the kit) is poorly covered or altogether missed, largely owing to factors that are not specific to exome capture itself.



variants for monogenic diseases[3]. Second, most alleles that are known to underlie Mendelian disorders disrupt protein-coding sequences[13]. Third, a large fraction of rare, protein-altering variants, such as missense or nonsense single-base substitutions or small insertion–deletions (that is, indels), are predicted to have functional consequences and/or to be deleterious[14]. As such, the exome represents a highly enriched subset of the genome in which to search for variants with large effect sizes.

Defining the exome. One particular challenge for applying exome sequencing has been how best to define the set of targets that constitute the exome. Considerable uncertainty remains regarding which sequences of the human genome are truly protein coding. When sequence capacity was more limiting, initial efforts at exome sequencing erred on the conservative side (for example, by targeting the high-confidence subset of genes identified by the Consensus Coding Sequence (CCDS) Project). Commercial kits now target, at a minimum, all of the RefSeq collection and an increasingly large number of hypothetical proteins. Nevertheless, all existing targets have limitations. First, our knowledge of all truly protein-coding exons in the genome is still incomplete, so current capture probes can only target exons that have been identified so far. Second, the efficiency of capture probes varies considerably, and some sequences fail to be targeted by capture probe design altogether (FIG. 1). Third, not all templates are sequenced with equal efficiency, and not all sequences can be aligned to the reference genome so as to allow base calling. Indeed, the effective coverage (for example, 50×) of exons using currently available commercial kits varies substantially. Finally, there is also the issue of whether sequences other than exons should be targeted (for example, microRNAs (miRNAs), promoters and ultra-conserved elements). These caveats aside, exome sequencing is rapidly proving to be a powerful new strategy for finding the cause of known or suspected Mendelian disorders for which the genetic basis has yet to be discovered.

Identifying causal alleles

A key challenge of using exome sequencing to find novel disease genes for either Mendelian or complex traits is how to identify disease-related alleles among the background of non-pathogenic polymorphism and sequencing errors. On average, exome sequencing identifies ~24,000 single nucleotide variants (SNVs) in African American samples and ~20,000 in European American samples (TABLE 1). More than 95% of these variants are already known as polymorphisms in human populations. Strategies for finding causal alleles against this background vary, as they do for traditional approaches to gene discovery, depending on factors such as: the mode of inheritance of a trait; the pedigree or population structure; whether a phenotype arises owing to *de novo* or inherited variants; and the extent of locus heterogeneity for a trait. Such factors also influence both the sample size needed to provide adequate power to detect trait-associated alleles and the selection of the most successful analytical framework.

Mendelian disorders
Phenotypes caused by a mutation (or mutations) in a single gene and inherited in a dominant, recessive or X-linked pattern.

Penetrance
The proportion of individuals with a specific phenotype among carriers of a particular genotype.

Locus heterogeneity
The appearance of phenotypically similar characteristics resulting from mutations at different genetic loci. Differences in effect size or in replication between studies and samples are often ascribed to different loci leading to the same disease.

Genome-wide association studies
(GWASs). Studies that search for a population association between a phenotype and a particular allele by screening loci (most commonly by genotyping SNPs) across the entire genome.

Complex traits
Traits that are influenced by the environment and/or through a combination of variants in at least several genes, each of which has a small effect.

Heritability
The proportion of the total phenotypic variation in a given characteristic that can be attributed to additive genetic effects.

Next-generation DNA sequencing
Highly parallelized DNA-sequencing technologies that produce many hundreds of thousands or millions of short reads (25–500 bp) for a low cost and in a short time.

Exome
The subset of a genome that is protein coding. In addition to the exome, commercially available capture probes target non-coding exons, sequences flanking exons and microRNAs.

Homozygosity mapping
Narrowing down the location of a gene underlying a trait by searching for regions of the genome in which both chromosomal segments are inherited identically-by-descent.

Box 2 | Exome sequencing to identify rare variants underlying complex traits

The systematic identification of rare alleles (that is, with a minor allele frequency (MAF) ≤ 1%) associated with common traits typically requires resequencing instead of genotyping[83] and has therefore been challenging. Such studies have largely been limited to assessing rare variants that have been found by the targeted sequencing of candidate genes or of genomic regions identified by linkage or genome-wide association studies (GWASs): the assumption is that rare variants that influence a trait colocalize with common variants that influence risk. Although it will eventually become standard to perform whole-genome sequencing on all subjects in a disease cohort, this approach is currently costly. In the meantime, exome sequencing provides an opportunity to capture nearly all of the rare and very rare (MAF < 0.1%) alleles in the protein-coding genes that are present in a sample, although the contribution of exome sequencing to our understanding of complex diseases has been much smaller than its contribution to our understanding of Mendelian traits.

Exome sequencing is often used in conjunction with two sampling strategies: family-based phenotypes (to exploit parent–child transmission patterns) and extreme phenotypes (to increase efficiency, FIG. 2d). In families in which multiple individuals are affected with a common trait, one approach is to sequence the most distally related individuals: the more distantly related the individuals, the fewer genetic variants they share. However, even distantly related individuals share many variants that require further stratification (for example, functional stratification) to identify a potentially causal allele. An alternative, family-based approach, which is used to identify *de novo* variants, involves sequencing parent–offspring trios in which only the offspring is affected. This strategy has been used to identify candidate genes for several complex traits[45,47] (Supplementary information 1 (table)).

In an extreme phenotype study design, individuals who are at both ends of a phenotype distribution are selected for sequencing[84]. For example, the US National Heart, Lung, and Blood Institute Exome Sequencing Project has sequenced the exomes of >7,000 individuals with extreme phenotypes to find genes that underlie common cardiovascular disease (for example, early-onset myocardial infarction and stroke) and lung disease (for example, chronic obstructive pulmonary disease). Because the frequency of alleles that contribute to the trait are enriched in the extremes of the distribution, sequencing even a modest sample size can potentially identify novel candidate alleles[85]. Fewer studies have used a case–control design, although reduced costs of exome sequencing coupled with increasing access to large numbers of publically available control exomes should increase the popularity of this approach.

A major challenge of relating rare variants to a trait is that even with very large sample sizes, the power to detect an association with a single rare variant is low. To address this issue, analysis strategies have been developed to assess the collective effects of rare variants across a gene or across multiple genes[86–89]. The assessment of association can be further aided by: incorporating prior evidence about variants (for example, functional class), genes and pathways; enhancing power using multivariate analyses of variants; or using quantitative rather than dichotomized phenotypes. A recent Review provides a more detailed description of these tests[90].

The main approaches for identifying causal variants in exome-sequencing data are discussed below.

Discrete filtering: approach and assumptions. In less than 2 years, exome sequencing has been used to identify causal alleles for several dozen Mendelian disorders[2,11,15–38] (Supplementary information 1 (table)). Most of these studies have, to varying extents, relied on comparisons with exome sequences and variants that are found in a small number of unrelated or closely related affected individuals to find rare alleles or novel alleles in the same gene shared among affected individuals (FIG. 2a). In these cases, novelty is assessed by filtering the variants against a set of polymorphisms that are available in public databases (for example, dbSNP and 1000 Genomes Project) and/or those found in a set of unaffected individuals (that is, controls). This 'discrete-filtering' step is used to eliminate candidate genes by assuming that any allele found in the 'filter set' cannot be causative. This approach is powerful in part because only a small fraction (~2% on average) of the SNVs identified in an individual by exome sequencing is novel (TABLE 1). The sequencing of only a modest number of affected individuals, and then applying discrete filtering to the data to reduce the number of candidate genes to a minimum number of high-priority candidates (if not a single one) is an important advantage that exome-sequencing approaches have over conventional approaches. In fact, this strategy alone can be exceptionally powerful for very rare Mendelian disorders[11].

Underlying this method is the assumption that the filter set contains no alleles from individuals with the phenotype being studied. This assumption can be problematic for two reasons. First, dbSNP is 'contaminated' with a small but appreciable number of pathogenic alleles. Second, as the number of sequenced exomes and genomes increases, the filtering of observed alleles in a manner that is independent of their minor allele frequency (MAF) runs the risk of eliminating truly pathogenic alleles that are segregating in the general population at low but appreciable frequencies. This risk is especially relevant for recessive disorders, in which carrier status will not result in a phenotype that might otherwise exclude an individual from a 'control' population (for example, 1000 Genomes Project)[11]. As shown in FIG. 3, analyses of recessive disorders in which one sets the maximum MAF to 1% are still well-powered. Recessive disorders for which carrier status is common (for example, cystic fibrosis) would still be missed using this filter, but most of the recessive disorders for which the genetic basis remains to be discovered are very rare. Alternatively, discrete filtering with a maximum MAF >1% can be carried out using a substantially larger sample size and/or in conjunction with pedigree-based approaches.

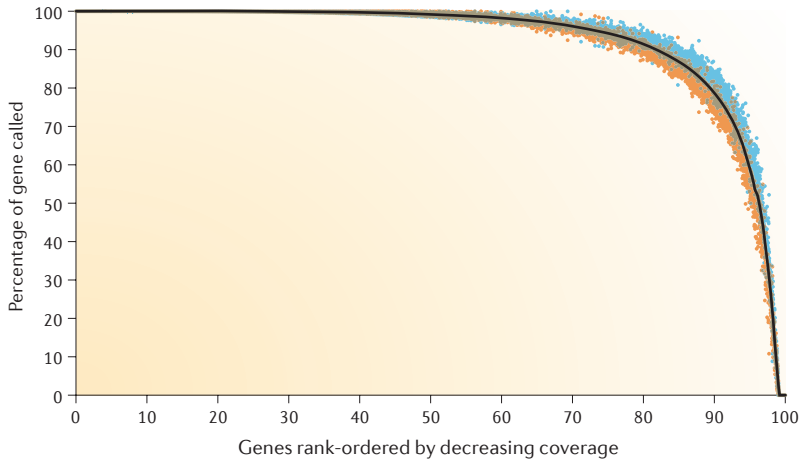


Figure 1 | **Exome coverage estimated as the percentage of bases called per gene.** Genes are rank-ordered by decreasing coverage, and the *x* axis is presented as the percentage of total genes. On average, 82% of genes have at least 90% of bases called ($n = 200$, black line). There is some variation by population. For African Americans (orange dots), $n = 100$, and for European Americans (blue dots), $n = 100$.

A lower MAF cutoff of 0.1% is helpful for dominant disorders, as the estimated prevalence of the disorder (generally well below 0.1%) provides an upper bound on the MAF. Additionally, the greater the number of novel variants with lower MAFs that are present in a sample population, the more difficult it will be to home in on the causal gene (or genes). This limitation underscores the importance of having access to control data that are derived from the same populations from which cases were sampled. Of note, these power analyses do not make use of dbSNP, suggesting that internally generated control data sets can be sufficiently deep, such that filtering against external databases to exclude common alleles is no longer required. In our experience, a large set of internally generated exome sequences also allows for the exclusion of systematic artefacts that are specific to the peculiarities of a production pipeline[23].

Stratifying candidates after discrete filtering. Candidate alleles can be further stratified on the basis of their predicted impact or deleteriousness. Alleles can be stratified by their functional class by giving greater weight to frameshifts, stop codons and disruptions of canonical splice sites than to missense variants. However, this is an oversimplification that is insensitive to causal alleles that do not directly alter protein-coding sequences or canonical splice sites. Additionally, candidate alleles can be stratified by existing biological or functional information about a gene: for example, its predicted role (or roles) in a biological pathway or its interactions with genes or proteins that are known to cause a similar phenotype.

Another approach for stratifying candidate alleles is to use quantitative estimates that have a functional impact, many of which exploit the observation that regions of genes and genomes in which mutations are deleterious tend to show high sequence conservation as a result of purifying selection. Sites that have experienced purifying selection can be identified by quantifying rates of mammalian evolution at the nucleotide level. Implementations of this strategy include phastCONS[39], phyloP and Genetic Evolutionary Rate Profiling (GERP)[40]. These annotation strategies can also be applied to predict the impact of potential causal alleles that are either coding or non-coding. Approaches that stratify non-synonymous alleles (for example, SIFT[41], Polymorphism Phenotyping v2 (PolyPhen2)[42] and Multivariate Analysis of Protein Polymorphism (MAPP)[43]) also explore the predicted changes in proteins caused by specific amino acid substitutions. All of these strategies enrich for functional sites at which observed variants are more likely to affect phenotype.

Filtering using tests of association. For identifying likely causal variants, an alternative strategy to discrete filtering is to apply tests of association. The use of two-sample tests that compare cases (that is, unrelated individuals with the same Mendelian phenotype) to a set of controls can either eliminate some of the problems of discrete filtering or provide estimates of the sample size needed for adequate power in the presence of complicating factors, such as genetic heterogeneity. For example, as long as false positives are equally probable both in cases and in controls, the expected number of variants in any gene will be the same both in cases and in controls under any distribution of mutations. When genetic heterogeneity is known to be present (as indicated, for example, by the presence of complementing groups of mutations) or suspected, then this information can be taken into account when performing power calculations to ensure that enough individuals are included in the study (BOX 3). Furthermore, the growing number of well-documented exome data sets available will allow for the use of thousands of control chromosomes, which can increase the power to detect causal alleles, even when the number of available cases is limited. It is noteworthy that the use of two-sample tests is the general rule in the search for variants underlying complex diseases; mathematically, the search for rare Mendelian variants and common variants underlying complex phenotypes is fundamentally indistinguishable. One modest caveat is that the use of parametric tests that are not 'exact' (for example, Fisher's exact test) generally require a permutation test to establish the correct significance level. Innovation in analytical methods for disease gene identification will be crucial for maximizing the success of exome sequencing in identifying genes for Mendelian disorders.

Effect of mode of inheritance on study design. The mode of inheritance of a monogenic disorder strongly influences both the experimental design (for example, the number of cases to sequence and selection of the most informative cases for sequencing in multiplex families) and the analytical approach. Intuitively, discrete filtering should be more efficient for recessive disorders (that is, they require sequencing of fewer cases) than for dominant disorders, because the genome of any given individual has around 50-fold fewer genes with two, rather than one, novel protein-altering alleles per gene (M.J.B., S.B.N., A.W.B., H.K.T., M.J.E., D.A.N. and

Sample indexing
Sequencing more than one sample in a single sequencing lane.

RefSeq
An open-access, annotated and curated collection of publicly available nucleotide sequences (DNA and RNA) and their protein translations.

Ultra-conserved elements
Subsequences of the genome that appear to be under extremely high levels of sequence constraint based on phylogenetic comparisons.

Purifying selection
Selection against a functionally deleterious allele.

Parametric tests
Statistical significance tests for which *P* values are based on models or assumed formulae for the distribution of the test statistic.

Permutation test
A statistical test in which the data are randomized many times to determine the statistical significance of the experimental outcome.

Multiplex families
Families in which two or more individuals are affected by the same disorder.

Table 1 | **Mean number of coding variants in two populations**

Variant type	Mean number of variants (± sd) in African Americans	Mean number of variants (± sd) in European Americans
Novel variants		
Missense	303 (± 32)	192 (± 21)
Nonsense	5 (± 2)	5 (± 2)
Synonymous	209 (± 26)	109 (± 16)
Splice	2 (± 1)	2 (± 1)
Total	520 (± 53)	307 (± 33)
Non-novel variants		
Missense	10,828 (± 342)	9,319 (± 233)
Nonsense	98 (± 8)	89 (± 6)
Synonymous	12,567 (± 416)	10,536 (± 280)
Splice	36 (± 4)	32 (± 3)
Total	23,529 (± 751)	19,976 (± 505)
Total variants		
Missense	11,131 (± 364)	9,511 (± 244)
Nonsense	103(± 8)	93 (± 6)
Synonymous	12,776 (± 434)	10,645 (± 286)
Splice	38 (± 5)	34 (± 4)
Total	24,049 (± 791)	20,283 (± 523)

The table lists the mean number (± standard deviation (sd)) of novel and non-novel coding single nucleotide variants from 100 sampled African Americans and 100 European Americans. Non-novel variants refer to those found in dbSNP131 or in 200 other control exomes. Capture was performed using the Nimblegen V2 target. The analysis pipeline consisted of: alignment using the Burrows–Wheeler alignment tool; recalibration; realignment around insertion–deletions and merging with the Genome Analysis Toolkit (GATK)[91]; and removal of duplicates with PICARD. Variants were called using the following parameters: quality score > 50, allele balance ratio < 0.75; homopolymer run > 3; and quality by depth < 8. Variants were called from a RefSeq37.2 target (35,804,408 bp).

J.S., unpublished data). This conclusion is supported by simulation studies (FIG. 3) and by the greater rate at which exome sequencing is identifying genes for recessive disorders relative to dominant disorders (Supplementary information 1 (table)).

Use of pedigree information. For Mendelian phenotypes, the use of pedigree information can substantially narrow the genomic search space for candidate causal alleles (FIG. 2b). However, it is not necessary to perform exome sequencing on every individual in a pedigree, or even on every case, to take full advantage of the available information. Exactly which individuals are the most informative ones to sequence depends on the frequency of a disease-causing allele and the nature of the relationship between the individuals. For very rare alleles, the probability of identity-by-descent given identity-by-state is high even among distantly related individuals. For example, two first cousins share a rare allele that is identical-by-descent in approximately one-eighth of the genome. In the absence of mapping data, sequencing the two most distantly related individuals with the phenotype of interest can substantially restrict the genomic search space.

When mapping data are available, the most efficient strategy is to sequence a pair of affected individuals whose overlapping haplotype produces the smallest shared genomic region. If the haplotype shared by all affected individuals is sufficiently short that the candidate interval is unlikely to include multiple candidate causal alleles, then a single individual may be sequenced. For consanguineous pedigrees in which a recessive mode of inheritance is suspected, sequencing just the one person with the smallest region (or regions) of homozygosity, as determined by the genome-wide genotyping data, should be sufficient. In each of these instances, exome sequencing is merely used as a replacement for Sanger sequencing of all the genes in a crucial interval. This is often a cost-effective option and has been a popular application of exome sequencing (Supplementary information 1 (table)).

Exome sequencing of parent–child trios is a highly effective approach for identifying *de novo* coding mutations (FIG. 2c), as multiple *de novo* events occurring within a specific gene (or within a gene family or pathway) is an extremely unlikely event[44]. This study design may be particularly applicable to gene discovery in disorders for which most cases are sporadic (that is, the parents are unaffected) and when a dominant mode of inheritance is suspected (for example, when there are few instances of parent-to-child transmission) or substantial locus heterogeneity is expected. Sequencing of trios has been used to find *de novo* mutations that are responsible for rare Mendelian disorders (for example, Schinzel–Giedion syndrome[37]) as well as for genetically heterogeneous disorders, such as intellectual disabilities[45], schizophrenia[46] and autism[47]. Although identifying Mendelian inconsistencies in which the child has a variant that is not called in either parent is straightforward, more than 70% of these inconsistencies turn out to be false negatives that result from failure to call the corresponding germline variants in one or the other parent. This should become less of a problem as variant callers improve and as the number of exomes available for comparison to exclude rare variants increases.

Technical and analytical limitations. The most successful reports of the identification of a novel disease gene by exome sequencing have relied on discrete filtering, often with the aid of mapping data (Supplementary information 1 (table)). However, it is difficult to know how often this approach has failed, as negative results are rarely reported. Failure can result for many reasons, most of which can be broadly considered as either technical or analytical.

Technical failure can occur because: part or all of the causative gene is not in the target definition (for example, it is not known to be a gene or there is a failure in the design); there is inadequate coverage of the region that contains a causal variant (for example, because of poor capture or poor sequencing); the causal variant is covered but not accurately called (for example, in the presence of a small but complex indel); true novel variants in the same gene are repeatedly identified but only because of the large size of the gene; or false variants in a gene are called because of mismapped reads or errors in alignment. Improvements on current methods to overcome these weaknesses are being investigated (BOX 1), so technical failures are likely to diminish rapidly over the next few years.

Identity-by-descent
Alleles on different chromosomes that are identical because they are inherited from a shared common ancestor.

Identity-by-state
Alleles on different chromosomes that are identical but do not share a common ancestor with respect to a pedigree or population of interest.

Haplotype
A combination of alleles on a single chromosome.

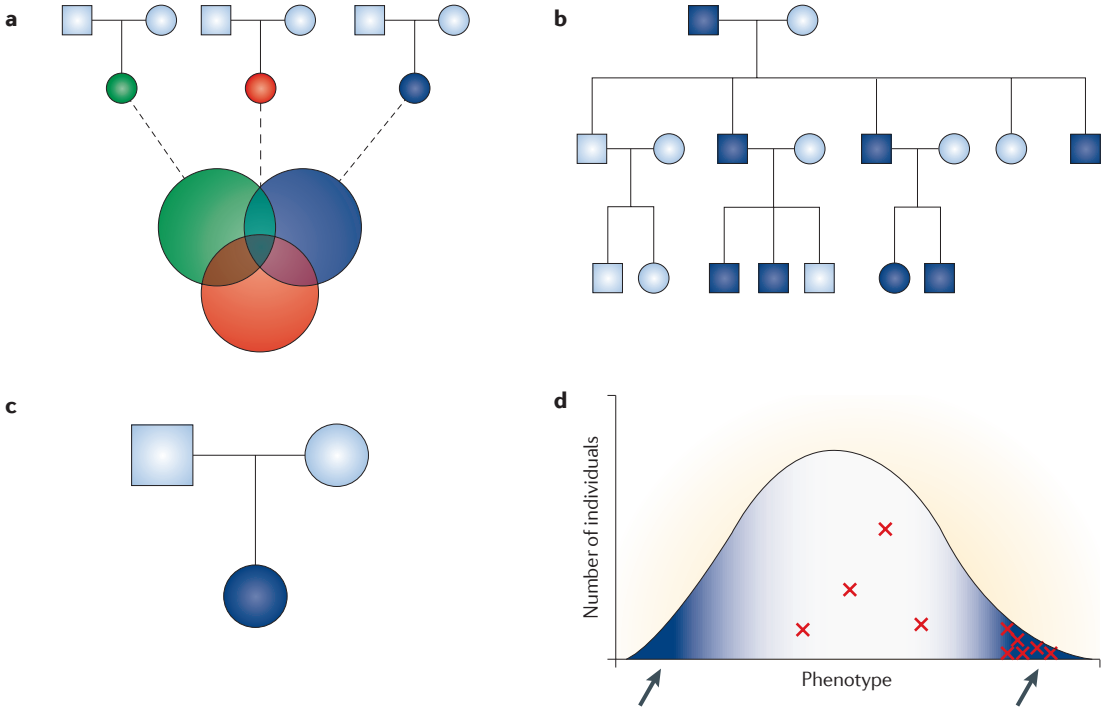


Figure 2 | **Strategies for finding disease-causing rare variants using exome sequencing.** Four main strategies are illustrated. **a** | Sequencing and filtering across multiple unrelated, affected individuals (indicated by the three coloured circles). This approach is used to identify novel variants in the same gene (or genes), as indicated by the shaded region that is shared by the three individuals in this example. **b** | Sequencing and filtering among multiple affected individuals from within a pedigree (shaded circles and squares) to identify a gene (or genes) with a novel variant in a shared region of the genome. **c** | Sequencing parent–child trios for identifying *de novo* mutations. **d** | Sampling and comparing the extremes of the distribution (arrows) for a quantitative phenotype. As shown in panel **d**, individuals with rare variants in the same gene (red crosses) are concentrated in one extreme of the distribution.

Analytical failures can follow from the limitations and assumptions of discrete filtering. Perhaps the major limitation of discrete filtering is that its power is substantially reduced by genetic heterogeneity. For example, if alleles of one gene account for only a fraction of cases, no single gene will be found to have disease-causing alleles in all cases, and several other genes may carry neutral mutations in as many cases, depending on the sample size. In this scenario, it is impossible to separate the causal alleles from the non-causal alleles. From an analytical perspective, false-negative calls, the presence of disease-causing alleles in the comparative data set and reduced penetrance result in a reduced signal-to-noise ratio that is practically indistinguishable from genetic heterogeneity. False-positive calls will result in detection of candidate genes that cannot logically be eliminated by filtering alone. False-positive calls are frequently observed in segmental duplications and processed pseudogenes. Particularly notorious are processed pseudogenes that are not currently represented in the human genome (for example, *CDC27*; see REF. 11). Finally, quantifying the strength of the results of discrete filtering by significance testing (for example, by *P* value or by posterior probability) is problematic in the absence of a better understanding about the nature and distribution of variation across the exome.

Processed pseudogenes
Copies of the coding sequences of genes that lack promoters and introns, contain poly(A) tails and are flanked by target-site duplications.

Posterior probability
The probability of an event after combining prior knowledge of the event with the likelihood of that event given by observed data.

Application to clinical diagnostics

Discovery of variants that underlie both Mendelian and complex traits will naturally lead to a much deeper understanding of disease mechanisms that should, in turn, facilitate development of improved diagnostics, prevention strategies and targeted therapeutics[48,49]. For example, the finding that several families with dominantly inherited adult-onset arterial calcifications had mutations in *NT5E* — a gene that encodes a protein involved in adenosine metabolism — allowed for consideration of specific therapeutic interventions that would otherwise not have been considered[50]. Some of these improvements will soon be realized (for example, better diagnostics for Mendelian disorders and disorders of unknown aetiology), whereas others (for example, risk profiling for complex traits) are likely to be a more distant realization.

Diagnosis. One of the immediate applications of exome sequencing will be facilitating the accurate diagnosis of individuals with Mendelian disorders that: present with atypical manifestations; are difficult to confirm using clinical or laboratory criteria alone (for example, when symptoms are shared among multiple disorders); or require extensive or costly evaluation (for example, when there is a long list of possible candidate

genes, as is the situation for non-syndromic hearing loss, Charcot–Marie–Tooth syndrome or when several large genes need to be screened). For example, using exome sequencing, a novel homozygous missense (Asp652Asn) variant in solute carrier family 26, member 3 (*SLC26A3*) — a gene that is known to cause a congenital chloride-losing diarrhoea — was identified in a child originally suspected to have a different diagnosis of Bartter syndrome[51]. Similarly, exome sequencing was used to discover a novel Cys203Tyr variant in X-linked inhibitor of apoptosis (*XIAP*) in a young boy with severe inflammatory bowel disease in whom a definitive diagnosis was elusive, despite a comprehensive evaluation[52]. Mutations in *XIAP* are a known cause of X-linked lymphoproliferative syndrome type 2 (XLP2), but severe colitis is an unusual symptom of XLP2. Furthermore, the diagnosis of XLP2 suggested a specific course of treatment (namely, allogeneic haematopoietic progenitor cell transplant) that had not been considered previously and appears to have been, at least in the short term, successful. These examples and others[53,54] provide proof-of-concept that exome sequencing can be used as a clinical tool for evaluating patients with an undiagnosed, although not entirely unexpected, genetic illness.

A major challenge for clinicians is making a specific diagnosis in individuals with novel phenotypes or those with phenotypes that are difficult to differentiate into aetiologically distinct categories (for example, autism or global developmental delay). Recent applications of exome sequencing to identify *de novo* variants in children with idiopathic intellectual disabilities[45] and children with sporadic autism[47] suggest that such phenotypes could be tractable to genome-wide screening for protein-coding variants that are predicted to have deleterious effects.

Screening. Many genetic disorders are screened for before conception, before birth or in the newborn period. Approaches based on next-generation sequencing have proved to be capable of detecting fetal aneuploidies using free fetal DNA isolated from maternal plasma[55,56] and have proved to be useful for carrier screening of several hundred genes that are known to cause rare recessive Mendelian disorders[57]. The extent to which exome sequencing can offer even more comprehensive screening or risk profiling for common, complex diseases remains to be explored[58,59]. However, for some disorders (for example, phenylketonuria), biochemical or enzymatic assays are more highly correlated with clinical presentation or outcome than they are with the genotype[60]. Screening via exome or genome sequencing will therefore probably always have some limitations.

Challenges. Widespread, useful, convenient and cost-effective use of exome sequencing — and eventually whole-genome sequencing — for clinical diagnosis or screening will necessitate overcoming a number of major challenges that currently limit broad applicability[61]. These challenges can be divided into those that are related to technical considerations and those that pose challenges to implementation.

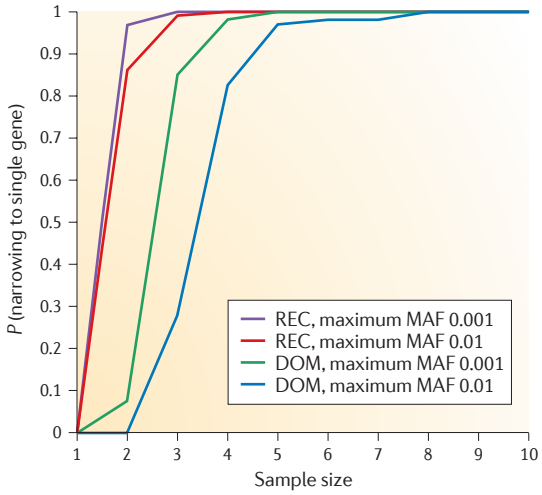
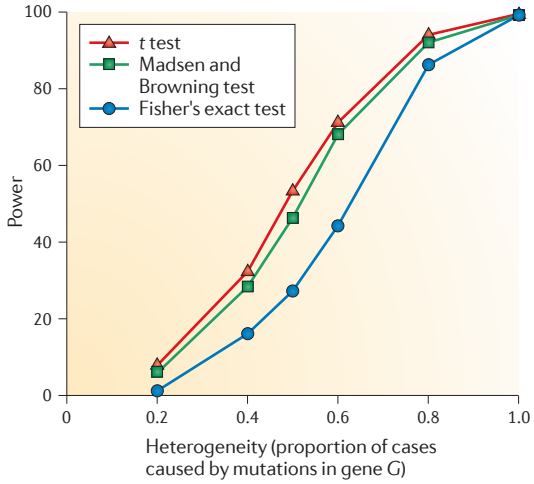


Figure 3 | **Estimated probability of identifying a single causal gene for a monogenic disorder under a discrete filtering framework.** The plot shows the probability (*P*; *y* axis) of identifying a single causal gene when exome sequencing is applied to a series of unrelated cases. Common variants are removed from consideration and causative variants are assumed to be protein-altering. Cases are sampled from 772 deeply sequenced exomes (from individuals of European ancestry) with 100 bootstrap replications per data point, and the remaining individuals are used as controls for defining the minor allele frequency (MAF) of individual variants. The graph shows how power increases with sample size. Estimates are shown for a maximum MAF = 0.001 versus a maximum MAF = 0.01 under a recessive (REC) or dominant (DOM) model.

There are several technical hurdles, but work that is aimed at overcoming them is proceeding at a rapid pace. First, sequencing and assembly will have to be highly accurate to avoid misdiagnosis. Second, algorithms for annotating variants will need to be automated, and approaches for characterizing the functional impact of rare and novel variants will have to be improved[62]. Relevance to disease-related risk will require comparison to a well-curated catalogue of variants that are known to influence risk of disease. General databases, such as the Human Gene Mutation Database (HGMD), and locus-specific databases are currently used with caution. Efforts are therefore underway to create comprehensive collections of validated associated variants (for example, the Human Variome Project)[63]. Nevertheless, it is clear that for the immediate future, a sizable fraction of the variants discovered in any given individual will need to be considered as variants of unknown importance[58,64]. Third, strategies for interpreting the use of variants (for example, clinical, reproductive or personal use) in a broad range of contexts (for example, for estimating the prior probabilities of disease based on exposure to environmental risk factors) need to be developed and tested[65]. Last, standards and guidelines for exome or whole-genome sequence testing and reporting in clinical laboratories will need to be established[66].

Bootstrap
A type of statistical analysis that is generally used for measuring the reliability of a sample estimate. It proceeds by the repeated sampling, with replacement, of the original data set. In the application described here, bootstrapping is used to assess the probability of identifying the causal variant for a genetic condition in a population.

Box 3 | **Power of two-sample tests for detecting novel variants in the same gene**



Two-sample tests are useful in cases in which locus heterogeneity and/or non-negligible false-negative rates are present. In these situations, it can be impossible for discrete filtering to narrow down the variants to a single candidate. Two-sample tests that are applicable are Fisher's exact test and the two-sample *t* test. The tests are applied on a per-gene basis to a sample of cases and a sample of controls, counting either the proportion of subjects having a deleterious variant in the tested gene (for Fisher's exact test) or the total number of rare deleterious variants in that gene for each sample (for the *t* test). Except for Fisher's exact test, permutation testing is used to derive *P* values, given the non-normal distribution of the data and small sample sizes; correction for testing multiple genes must also be done.

We determined the power of Fisher's exact test (blue line) and the *t* test (red line) to detect a causal gene, *G*, under locus heterogeneity with sample sizes of 10 cases and 50 controls (controls can be obtained from publicly available exomes from individuals whose ethnicity is matched to that of cases). We also compared results for the test of Madsen and Browning that was developed for complex disease settings and that modestly down-weights variants with higher frequency (green line). The simulation assumed a rare dominant disease with prevalence of 1 in 100,000 and a penetrance of 0.80. Rare, non-disease-causing variants occur with a frequency equal to the disease-causing variants. The proportion of cases caused by variants in *G* was then varied to assess the effect of locus heterogeneity on power. The critical value of the power was set at 5×10^{-5}, allowing one false positive in 20,000 tests. The results show: a degradation in power as heterogeneity increases; that the Fisher's exact test is considerably less powerful than the *t* test; and that the weighting of the Madsen and Browning test provides no power gains. For example, power is approximately 75% for the *t* test when 60% of cases are caused by the tested gene. Note that this situation can also be interpreted as a false-negative rate of 40%. (Power is 99% in this situation when 30 cases and 100 controls are used; results, not shown, from M.J.B., S.B.N., A.W.B., H.K.T., M.J.E., D.A.N. and J.S.).

The speed at which the challenges related to implementation are overcome will probably determine the pace at which genomic sequence information is used for personalized care and the breadth of its use. However, application in a stepwise fashion that focuses first on testing for all known Mendelian disorders, for example, will facilitate the introduction of sequence information for diagnostic use and should be instructive for its application to common diseases. Perhaps the greatest challenge will be to determine the specific scenarios in which personal exome or whole-genome data provide benefits to prevention or clinical management (for example, diagnosis and treatment) that are substantial enough to justify the costs. A second challenge is the need to train health-care providers to incorporate genomic information into their practice. On this same point, given that there are only about 3,000 genetic counsellors in the United States and Canada (K. Dent, personal communication, US National Society of Genetic Counselors), it is unclear how the volume and scope of the results from exome and whole-genome sequencing will be effectively communicated to an individual. Furthermore, the interpretation of information will change over time as new risks are reported and others are refuted, as the magnitude of risks change and as interactions among variants and interactions with environmental factors are discovered. Despite all of these challenges, we are optimistic that exome sequence information, and eventually whole-genome data, will eventually become part of the routine clinical evaluation of all persons suspected of having a genetic disorder and may eventually be used to provide personalized health-care profiling.

Ethical considerations

The use of exome sequencing for disease gene discovery generates new manifestations of several long-standing ethical issues in human genetics research. Two areas of research ethics that require consideration in particular are the limitations of the current consent process and management of individual research results. There are several other important ethical issues that should be considered in the context of exome-sequencing studies, but these are beyond the scope of this Review. They include issues surrounding data sharing, the return of test results to individuals over time and the necessity of Clinical Laboratory Improvement Amendments (CLIA) validation of exome-sequencing protocols.

Informed consent for exome sequencing. There are several important ethical challenges in exome-sequencing research related to informed consent. Many studies that incorporate exome sequencing may use banked samples collected using consent documents that did not specifically anticipate, let alone describe, exome sequencing. This raises questions about what type of information is needed to make informed decisions about participation in exome-sequencing research and whether and how this information differs from standard information about the

risks and benefits of genetic research. In many cases, the answers are complex and contextual. Furthermore, in many ways, the goals of exome sequencing are similar to the targeted sequencing approaches already applied in genetic analysis. However, there are possible risks that can be considered.

First, exome-sequencing approaches increase the chance of uncovering clinically useful results that are unrelated to the primary aim of the study (for example, identification of a disease gene). The need to describe the increased chance of returning results to the tested individual will need to be balanced with the desire to avoid unrealistic participant expectations and potential therapeutic misconception about possible benefits of participation. Second, the risks of sharing individual-level genotype or raw sequence data from exome-sequencing studies in databases such as dbGaP may differ from GWAS data. These risks should be assessed, as such information is essential for the development of a set of appropriate data-sharing policies and protections.

Return and management of results. Researchers and policy makers continue to struggle to develop a framework and guidelines for the return of results from genetic studies. Although there is not a clear consensus, several practices have emerged that generally minimize the need to return results unless: they are identified in the course of routine research analysis; they have been validated; and they are determined to be clinically useful and to be actionable. The details of how the terms 'clinically useful' and 'actionable' are defined, and by whom, remain under dispute and are generally approached on a case-by-case basis[67,68].

Exome sequencing also challenges existing assumptions for the return of results, particularly regarding the nature of so-called 'incidental findings' (REF. 69). As a major aim of exome sequencing — unlike more targeted approaches — is to identify all variants in an exome, it cannot be assumed that few, if any, clinically important or actionable results will be identified in the course of routine research. Instead, exome sequencing identifies clinically useful results that could currently be identified through targeted genetic testing. Therefore, it is no longer a question of whether clinically useful results will be found in any research participant, but rather how many such results will be identified in each participant.

Given the new realities of exome sequencing, researchers who consider returning results will have to contend with several major issues. First, they will need to identify 'known' variants that are associated with health-related traits and interpret their clinical importance. Second, they will need to consider what kinds of health-related results (for example, carrier status, cancer predisposition or drug response) to return to participants. Researchers should place highest priority on results that fall under the category of 'duty to warn' surrounding health and disease risk[65]. Third, researchers will need to consider participant expectations about re-contact and develop an ethically appropriate, context-specific plan for the return of results. Finally, whereas the return of results by conventional means (for example, face-to-face genetic counselling) is the gold standard, it is also expensive, especially given the added cost imposed by returning a larger number of results.

Although exome sequencing will identify a much larger number of clinically useful results than other genetic research approaches, it does not follow that there ought to be a mandatory review and return of all such results to participants in all studies. The decision about whether and how to return results must take into account factors such as the commitments made at the time of informed consent and the resources available both to analyse variant data and to confirm and return results responsibly on a large scale[70].

Future directions

Because of our poor ability to make sense of non-coding variation, the analytical components of most 'whole-genome' studies have disproportionately focused on variation within the exome. As the cost of sequencing continues to fall, the field will probably gradually move from exome to whole-genome sequencing[71]. However, taking advantage of these more comprehensive data for disease gene discovery and molecular diagnostics in patients crucially depends on the development of analytical strategies for making sense of non-coding variation. This is as much an opportunity as it is a challenge.

There are several specific areas in which focused efforts are likely to advance the field substantially. These include, first, the proper curation of phenotypes, particularly in the context of Mendelian disorders. To this end, there are hundreds, perhaps thousands, of poorly defined familial phenotypes that are rare or unique. Development of repositories in which descriptive information about such phenotypes and an accompanying DNA sample could be banked by clinicians from anywhere in the world would facilitate both delineation of new Mendelian disorders and discovery of the underlying genes. Second, we need improved technical, statistical and bioinformatic methods for: reducing the rate of false-positive and false-negative variant calls; calling indels; prioritizing candidate causal variants; and predicting and annotating the potential functional impact for disease gene discovery or molecular diagnostics. Third, to realize fully the potential of sequencing for clinical diagnostics and personal genomic profiling, we need to address the challenges posed by ethics and policy issues. Nevertheless, exome and even genome sequencing are likely to be introduced in the clinical setting before these challenges are fully resolved owing in part to their ability to facilitate diagnosis and inform therapy[52].

In the immediate future, and at a small fraction of the cost per disorder compared to conventional discovery strategies, the power of existing approaches should enable the identification of the genes underlying a large fraction of all known Mendelian disorders that are currently unsolved. Identifying a candidate gene (or genes) for every recessive disorder using as few as one affected individual per disorder is a realistic goal at present. Solving all dominant disorders will be more difficult, but is increasingly tractable with technical and analytical

Incidental findings
Findings that are not explicitly related to the original research hypotheses (that is, primary findings).

improvements that should take place over the next several years. From our perspective, solving all Mendelian disorders should be an imperative for both the human genetics community and funding agencies, as such discoveries will be of enormous service to families while also providing novel entry points for the investigation of the mechanisms underlying the development of disease. Solving the remaining several thousand Mendelian disorders by these new methods will require an unprecedented degree of cooperation and coordination in the field of medical genetics. However, efforts are underway at multiple centres throughout the world to establish the collaborative framework and physical infrastructure required. Accordingly, we can realistically look towards a future in which the genetic basis of all Mendelian traits is known, and emphasis shifts further towards understanding disease mechanisms and genotype–phenotype relationships, developing improved therapeutics and translating knowledge of the exome to the improvement of human health.

1. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007).
2. Kaiser, J. Human genetics. Affordable 'exomes' fill gaps in a catalogue of rare diseases. *Science* **330**, 903 (2010).
3. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nature Rev. Genet.* **7**, 277–282 (2006).
4. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).
5. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
6. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
7. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
8. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**, 111–118 (2010).
9. Biesecker, L. G. Exome sequencing makes medical genomics a reality. *Nature Genet.* **42**, 13–14 (2010).
10. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
This was the first study to show the feasibility of using exome sequencing to identify disease-causing variants.
11. Ng, S. B. *et al.* Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
This was the first study to use exome sequencing to discover the genetic basis of a monogenic disorder.
12. Bilguvar, K. *et al.* Whole-exome sequencing identifies recessive *WDR62* mutations in severe brain malformations. *Nature* **467**, 207–210 (2010).
This is an outstanding paper demonstrating the narrowing to a single candidate gene that is made possible by exome sequencing a single case in the context of a consanguineous pedigree and a recessive phenotype.
13. Stenson, P. D. *et al.* The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum. Genomics* **4**, 69–72 (2009).
14. 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).
15. Simpson, M. A. *et al.* Mutations in *NOTCH2* cause Hajdu–Cheney syndrome, a disorder of severe and progressive bone loss. *Nature Genet.* **43**, 303–305 (2011).
16. Krawitz, P. M. *et al.* Identity-by-descent filtering of exome sequence data identifies *PIGV* mutations in hyperphosphatasia mental retardation syndrome. *Nature Genet.* **42**, 827–829 (2010).
17. Tsurusaki, Y. *et al.* Rapid detection of a mutation causing X-linked leucoencephalopathy by exome sequencing. *J. Med. Genet.* **48**, 606–609 (2011).
18. Liu, Y. *et al.* Confirmation by exome sequencing of the pathogenic role of *NCSTN* mutations in acne inversa (hidradenitis suppurativa). *J. Invest. Dermatol.* **131**, 1570–1572 (2011).
19. Yamaguchi, T. *et al.* Exome resequencing combined with linkage analysis identifies novel *PTH1R* variants in primary failure of tooth eruption in Japanese. *J. Bone Miner. Res.* **26**, 1655–1661 (2011).
20. Zuchner, S. *et al.* Whole-exome sequencing links a variant in *DHDDS* to retinitis pigmentosa. *Am. J. Hum. Genet.* **88**, 201–206 (2011).
21. Otto, E. A. *et al.* Candidate exome capture identifies mutation of *SDCCAG8* as the cause of a retinal–renal ciliopathy. *Nature Genet.* **42**, 840–850 (2010).
22. Haack, T. B. *et al.* Exome sequencing identifies *ACAD9* mutations as a cause of complex I deficiency. *Nature Genet.* **42**, 1131–1134 (2010).
23. Ng, S. B. *et al.* Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nature Genet.* **42**, 790–793 (2010).
24. Al Badr, W. *et al.* Exome capture and massively parallel sequencing identifies a novel *HPSE2* mutation in a Saudi Arabian child with Ochoa (urofacial) syndrome. *J. Pediatr. Urol.* 28 Mar 2011 (doi:10.1016/j.jpurol.2011.02.034).
25. Bolze, A. *et al.* Whole-exome-sequencing-based discovery of human FADD deficiency. *Am. J. Hum. Genet.* **87**, 873–881 (2010).
26. Caliskan, M. *et al.* Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the *TECR* gene on chromosome 19p13. *Hum. Mol. Genet.* **20**, 1285–1289 (2011).
27. Glazov, E. A. *et al.* Whole-exome re-sequencing in a family quartet identifies *POP1* mutations as the cause of a novel skeletal dysplasia. *PLoS Genet.* **7**, e1002027 (2011).
28. Walsh, T. *et al.* Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am. J. Hum. Genet.* **87**, 90–94 (2010).
29. Johnston, J. J. *et al.* Massively parallel sequencing of exons on the X chromosome identifies *RBM10* as the gene that causes a syndromic form of cleft palate. *Am. J. Hum. Genet.* **86**, 743–748 (2010).
30. Norton, N. *et al.* Genome-wide studies of copy number variation and exome sequencing identify rare variants in *BAG3* as a cause of dilated cardiomyopathy. *Am. J. Hum. Genet.* **88**, 273–282 (2011).
31. Musunuru, K. *et al.* Exome sequencing, *ANGPTL3* mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* **363**, 2220–2227 (2010).
32. Johnson, J. O. *et al.* Exome sequencing reveals *VCP* mutations as a cause of familial ALS. *Neuron* **68**, 857–864 (2010).
33. Wang, J. L. *et al. TGM6* identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain* **133**, 3510–3518 (2010).
34. Gilissen, C. *et al.* Exome sequencing identifies *WDR35* variants involved in Sensenbrenner syndrome. *Am. J. Hum. Genet.* **87**, 418–423 (2010).
35. Lalonde, E. *et al.* Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Hum. Mutat.* **31**, 918–923 (2010).
36. Sirmaci, A. *et al. MASP1* mutations in patients with facial, umbilical, coccygeal, and auditory findings of Carnevale, Malpuech, OSA, and Michels syndromes. *Am. J. Hum. Genet.* **87**, 679–686 (2010).
37. Hoischen, A. *et al. De novo* mutations of *SETBP1* cause Schinzel–Giedion syndrome. *Nature Genet.* **42**, 483–485 (2010).
38. Kalay, E. *et al.* CEP152 is a genome maintenance protein disrupted in Seckel syndrome. *Nature Genet.* **43**, 23–26 (2011).
39. Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief. Bioinf.* **12**, 41–51 (2011).
40. Cooper, G. M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods* **7**, 250–251 (2010).
41. Kumar, P. H. & S. Ng, P. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protoc.* **4**, 1073–1081 (2009).
42. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
43. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).
44. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304 (2000).
45. Vissers, L. E. *et al.* A *de novo* paradigm for mental retardation. *Nature Genet.* **42**, 1109–1112 (2010).
This was the first study to use exome sequencing of parent–child trios of affected offspring and their unaffected parents to identify *de novo* variants and thus candidate genes for a complex trait characterized by substantial locus heterogeneity.
46. Girard, S. L. *et al.* Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nature Genet.* **43**, 860–863 (2011).
47. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43**, 585–589 (2011).
48. Blakemore, A. I. & Froguel, P. Investigation of Mendelian forms of obesity holds out the prospect of personalized medicine. *Ann. N.Y. Acad. Sci.* **1214**, 180–189 (2010).
49. Dietz, H. C. New therapeutic approaches to Mendelian disorders. *N. Engl. J. Med.* **363**, 852–863 (2010).
50. St. Hilaire, C. *et al. NT5E* mutations and arterial calcifications. *N. Engl. J. Med.* **364**, 432–42 (2011).
51. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* **106**, 19096–19101 (2009).
This paper provides the first example of applying exome sequencing to make an unanticipated diagnosis in a clinical setting.
52. Worthey, E. A. *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.* **13**, 255–262 (2011).
This is an outstanding example of the clinical diagnosis of a rare disorder by exome sequencing leading to a subsequent, life-saving change in treatment.
53. Bonnefond, A. *et al.* Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS ONE* **5**, e13630 (2010).
54. Montenegro, G. *et al.* Exome sequencing allows for rapid gene identification in a Charcot–Marie–Tooth family. *Ann. Neurol.* **69**, 464–470 (2011).
55. Chiu, R. W. *et al.* Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc. Natl Acad. Sci. USA* **105**, 20458–20463 (2008).
56. Chiu, R. W. & Lo, Y. M. Non-invasive prenatal diagnosis by fetal nucleic acid analysis in maternal plasma: the coming of age. *Semin. Fetal Neonatal Med.* **16**, 88–93 (2011).
57. Bell, C. J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra4 (2011).
This work reports on efforts to implement pre-conception carrier screening for over 400 recessive disorders by hybrid capture and next-generation sequencing.

58. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010). **This paper illustrates both the promise and challenges we face in the clinical interpretation of exome or genome sequences of individual patients.**
59. Kingsmore, S. F. & Saunders, C. J. Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Sci. Transl. Med.* **3**, 87ps23 (2011).
60. Scriver, C. R. The *PAH* gene, phenylketonuria, and a paradigm shift. *Hum. Mutat.* **28**, 831–845 (2007).
61. Ormond, K. E. *et al.* Challenges in the clinical application of whole-genome sequencing. *Lancet* **375**, 1749–1751 (2010).
62. Tong, M. Y., Cassa, C. A. & Kohane, I. S. Automated validation of genetic variants from large databases: ensuring that variant references refer to the same genomic locations. *Bioinformatics* **27**, 891–893 (2011).
63. Kohonen-Corish, M. R. *et al.* How to catch all those mutations—the report of the third Human Variome Project Meeting, UNESCO Paris, May 2010. *Hum. Mutat.* **31**, 1374–1381 (2010).
64. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010). **This was the first study to report the sequencing of the entire genome for each member of a family with a Mendelian disorder.**
65. Beskow, L. M. & Burke, W. Offering individual genetic research results: context matters. *Sci. Transl. Med.* **2**, 38cm20 (2010).
66. Richards, C. S. *et al.* ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. *Genet. Med.* **10**, 294–300 (2008).
67. Fabsitz, R. R. *et al.* Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. *Circ. Cardiovasc. Genet.* **3**, 574–580 (2011).
68. Caulfield, T. *et al.* Research ethics recommendations for whole-genome research: consensus statement. *PLoS Biol.* **6**, e73 (2008).
69. Wolf, S. M. *et al.* Managing incidental findings in human subjects research: analysis and recommendations. *J. Law Med. Ethics* **36**, 219–248 (2008).
70. Ravitsky, V. & Wilfond, B. S. Disclosing individual genetic results to research participants. *Am. J. Bioeth.* **6**, 8–17 (2006).
71. Green, E. D. & Guyer, M. S. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
72. Dahl, F. *et al.* Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl Acad. Sci. USA* **104**, 9387–9392 (2007).
73. Fredriksson, S. *et al.* Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* **35**, e47 (2007).
74. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotech.* **27**, 182–189 (2009).
75. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4**, 907–909 (2007).
76. Porreca, G. J. *et al.* Multiplex amplification of large sets of human exons. *Nature Methods* **4**, 931–936 (2007).
77. Albert, T. J. *et al.* Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**, 903–905 (2007).
78. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
79. Turner, E. H., Ng, S. B., Nickerson, D. A. & Shendure, J. Methods for genomic partitioning. *Annu. Rev. Genomics Hum. Genet.* **10**, 263–284 (2009).
80. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Rev. Genet.* **12**, 443–451 (2011).
81. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Rev. Genet.* **12**, 363–376 (2011).
82. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
83. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Rev. Genet.* **11**, 415–425 (2010).
84. Lanktree, M. B., Hegele, R. A., Schork, N. J. & Spence, J. D. Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease. *Circ. Cardiovasc. Genet.* **3**, 215–221 (2010).
85. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004). **This was an important study that demonstrated the effectiveness of sequencing candidate genes at the extremes of a phenotype to find rare alleles influencing risk for a complex trait.**
86. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
87. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
88. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
89. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
90. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nature Rev. Genet.* **11**, 773–785 (2010).
91. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).

Acknowledgements

We thank the US National Institutes of Health (NIH)/National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (Lung Grand Opportunity (GO) Sequencing Project (HL-102923 to M.J.B.), the US Women's Health Initiative (WHI) GO Sequencing Project (HL-102924), the Heart GO Sequencing Project (HL-103010), the Broad GO Sequencing Project (HL-102925) and the Seattle GO Sequencing Project (HL-102926 to D.A.N. and J.S.) for early data release that proved useful for demonstrating filtering strategies. Our work was supported in part by grants from the NIH/NHLBI (5R01HL094976 to D.A.N. and J.S.), the NIH/National Human Genome Research Institute (5R21HG004749 to J.S., 1RC2HG005608 to M.J.B., D.A.N. and J.S., and 5RO1HG004316 to H.K.T.), NIH/National Institute of Environmental Health Sciences (HHSN273200800010C to D.N.), the Life Sciences Discovery Fund (2065508 and 0905001), the Washington Research Foundation and the NIH/National Institute of Child Health and Human Development (1R01HD048895 to M.J.B.). S.B.N. is supported by the Agency for Science, Technology and Research, Singapore. A.W.B. is supported by a training fellowship from the NIH/National Human Genome Research Institute (T32HG00035).

Competing interests statement

The authors declare no competing financial interests.

DATABASES

1000 Genomes Project: http://www.1000genomes.org
dbSNP: http://www.ncbi.nlm.nih.gov/snp
EntrezGene: http://www.ncbi.nlm.nih.gov/gene
GeneReviews: http://www.ncbi.nlm.nih.gov/sites/GeneTests/review
Human Gene Mutation Database (HGMD): http://www.hgmd.cf.ac.uk/ac/index.php
miRBase: http://mirbase.org
OMIM: http://www.ncbi.nlm.nih.gov/omim
arterial calcifications | Bartter syndrome | congenital chloride-losing diarrhoea | Charcot–Marie–Tooth syndrome | cystic fibrosis | Schinzel–Giedion syndrome | sickle cell anaemia | X-linked lymphoproliferative syndrome type 2
RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq

FURTHER INFORMATION

Burrows–Wheeler alignment tool: http://bio-bwa.sourceforge.net
Consensus coding sequence (CCDS) project: http://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi
Exome Variant Server: http://snp.gs.washington.edu/EVS
Generic exome analysis plan: http://genome.sph.umich.edu/wiki/Generic_Exome_Analysis_Plan
Genetic Evolutionary Rate Profiling (GERP): http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html
Genome Analysis Toolkit (GATK): http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
Human Variome Project: http://www.humanvariomeproject.org
Multivariate Analysis of Protein Polymorphism (MAPP): http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html
***Nature Reviews Genetics* series on Applications of Next-Generation Sequencing:** http://www.nature.com/nrg/series/nextgeneration/index.html
***Nature Reviews Genetics* series on Translational Genetics:** http://www.nature.com/nrg/series/translational/index.html
***Nature Reviews Genetics* series on Study Designs:** http://www.nature.com/nrg/series/studydesigns/index.html
NHLBI Grand Opportunity Exome Sequencing Project: https://esp.gs.washington.edu/drupal
Phylogenetic analysis with space/time models (PHAST): http://compgen.bscb.cornell.edu/phast
phyloP: http://compgen.bscb.cornell.edu/phast/help-pages/phyloP.txt
Polymorphism Phenotyping v2 (Polyphen2): http://genetics.bwh.harvard.edu/pph2
Seqanswers: http://www.seqanswers.com
SIFT: http://sift.jcvi.org

SUPPLEMENTARY INFORMATION

See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF