Inferring Causality and Functional Significance of Human Coding DNA Variants

Shamil R. Sunyaev, PhD

Genetics Division
Brigham & Women's Hospital
Harvard Medical School
Boston, Massachusetts

Society for Neuroscience

NOTES

Introduction

Sequencing technology enables the complete characterization of human genetic variation. Statistical genetics studies identify numerous loci linked or associated to phenotypes of direct medical interest. The major remaining challenge is to characterize functionally significant alleles that are causally implicated in the genetic basis of human traits. Here, I review three sources of evidence for the functional significance of human DNA variants in protein-coding genes. These include (1) statistical genetics considerations such as cosegregation with the phenotype, allele frequency in unaffected controls and recurrence; (2) *in vitro* functional assays and model organism experiments; and (3) computational methods for predicting the functional effect of amino acid substitutions. In spite of many successes of recent studies, functional characterization of human allelic variants remains problematic.

Large-scale sequencing projects have revealed the landscape of human genetic variation (1000 Genomes Project Consortium, 2010; Tennessen et al., 2012). Linkage and association studies identified a large number of loci involved in various human phenotypes. In spite of this spectacular progress, characterization of functionally significant human alleles causally involved in phenotypes (i.e., directly contributing to the biology of phenotypes) remains challenging.

The problem of establishing a causal relationship between a phenotype and a specific sequence variant arises at multiple levels (Table 1). It spans both Mendelian and complex trait genetics, even though many aspects of the problem and approaches to address them are different.

In the simplest case of a Mendelian monogenic trait unequivocally linked to a particular gene, the problem is in distinguishing between benign and pathogenic alleles in this gene. This creates a major bottleneck in clinical genetic diagnostics (Plon et al., 2008). Many allelic variants observed in genes of diagnostic importance remain classified as variants of unknown significance (VUSs).

For Mendelian phenotypes with unknown genetic background, sequencing studies now provide a powerful way to identify causal genes. Briefly, the strategy involves finding a gene where all or most patients carry functional variants that are not observed in multiple unaffected controls (Ng et al., 2010). Usually, all coding nonsynonymous variants and variants disrupting canonic splice sites are considered functional and other variants are ignored. Although this strategy generated many successes, it lacks power if sample sizes are small (only two or three patients available) or in the case of oligogenic phenotypes. Knowledge of functional significance of allelic variants would greatly empower sequencing studies aiming at mapping genes underlying Mendelian disorders.

Remarkable progress in sequencing technology now allows detecting *de novo* mutations using parent–child trio sequencing (Roach et al., 2010). This approach has been successfully applied to a number of Mendelian traits and to complex psychiatric phenotypes such as autism and schizophrenia (Roach et al., 2010; Xu et al., 2011; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). Relatively small numbers of *de novo* mutations facilitate the analysis. On average, humans carry on the order of 100 *de novo* point mutations with only few (on average 1) of them coding (Nachman and Crowell, 2000; Kondrashov, 2003; Kong et al., 2012; Sun et al., 2012). However, these mutations are typically unique to individual patients. Therefore, it is impossible to use statistical approaches to infer their involvement in phenotypes in case of whole-exome or whole-genome sequencing experiments.

Naturally, the magnitude of the problem is amplified when considering variants involved in complex traits. Genome-wide association studies (GWAS) identified a multitude of common SNPs (single nucleotide polymorphisms) associated with human complex traits. However, most of these SNPs are not causal and simply tag causal alleles due to linkage disequilibrium (LD). LD greatly facilitates mapping but equally complicates pinpointing causal variants by statistical means because association signals of many variants are confounded. In many cases, even the identity of a causal gene, rather than a specific allele, is not known. The problem is exacerbated because most of GWAS peaks are in noncoding regions. Moreover, it is possible that multiple causal variants give rise to a single GWAS peak. A number of sequencing projects aiming at finding causal variants underlying GWAS peaks are ongoing. The dominant hypothesis is that the variants responsible for the observed associations are common. Scenarios where associations of common SNPs are caused by low-frequency variants or even by multiple rare variants have also been proposed (Dickson et al., 2010), although subsequent work suggested that such scenarios do not explain many GWAS peaks (Wray et al., 2011).

NOTES

Table 1. Importance of the functional analysis in various types of human genetics studies

Analysis of Mendelian traits				Analysis of rare variants in complex traits			
Interpretation of variants in previously mapped genes		Mapping genes by whole-genome/whole-exome sequencing		Analysis of rare variants in candidate genes		Mapping genes by whole-genome/whole-exome sequencing	
Uncharacterized variants not known to be *de novo*	*De novo* mutations	Segregating variants	*De novo* mutations	Rare variants	*De novo* mutations	Rare variants	*De novo* mutations
Analysis of the functional effect and causality is essential	Usually regarded as sufficient evidence of functionality	Functional analysis is not essential to map genes but can potentially increase power	Functional analysis is essential for isolated mutations; recurrence may provide a statistical argument in favor of functionality	Functional analysis was shown to increase power	Likely a sufficient evidence of functionality	Functional analysis was hypothesized to increase power	Functional analysis is essential for isolated mutations; recurrence may provide a statistical argument in favor of functionality

Examples of the functional characterization of variants underlying GWAS signals are still rare. One early example includes demonstration that a common variant creating a transcription factor binding site for the CCAAT/enhancer-binding protein alters the hepatic expression of the *SORT1* gene. This variant explains the corresponding GWAS signal for association with LDL-cholesterol (Musunuru et al., 2010). Fine-mapping studies have been reported recently for strong association signals within the human leukocyte antigen region (International HIV Controllers Study et al., 2010; Raychaudhuri et al., 2012). For common noncoding variants, analysis of intermediate molecular phenotypes related to transcriptional regulation such as mRNA expression (Stranger et al., 2007) and chromatin accessibility (McDaniell et al., 2010; Degner et al., 2012; Maurano et al., 2012) offers a potential way forward. These early studies on functional effects of common noncoding variants are outside of scope of this review.

A number of successful candidate gene-sequencing studies discovered associations of multiple rare coding variants with complex phenotypes (Cohen et al., 2004; Ahituv et al., 2007; Ji et al., 2008; Romeo et al., 2009; Johansen et al., 2010; Momozawa et al., 2011; Rivas et al., 2011; Bonnefond et al., 2012; Jordan et al., 2012; Kiezun et al., 2012). Ongoing whole-exome sequencing studies attempt an unbiased search for genes harboring multiple rare variants collectively associated with complex traits (Price et al., 2010). In the simplest form, this analysis detects an excess of rare coding variants in cases versus controls. The association signal is provided by functional variants, whereas neutral alleles are a source of noise masking the association signal. Again, functional significance of individual rare variants cannot be inferred by statistical means. In contrast to common variants, LD does not confound the signal. However, the association test for individual rare variants lacks statistical power, given that they are observed a handful of times (or even once) in the sample. The ability to discriminate between functional and neutral alleles would dramatically increase the potential of sequencing studies focusing on rare variants in complex traits. Several published studies demonstrated that highlighting functional variants using experimental (Romeo et al., 2009; Bonnefond et al., 2012) or computational approaches (Ahituv et al., 2007; Ji et al., 2008; MacArthur et al., 2012) increases the power of these studies.

Understanding the functional significance of human alleles is also of great importance for evolutionary and population genetics. Accurate inference of functional consequences of human DNA variants would help characterizing the role of natural selection in shaping population genetic variation (Tennessen et al., 2012).

Overall, medical genetics is interested in finding "pathogenic" mutations that causally influence traits of medical interest. Population genetics focuses on "deleterious" alleles that evolve under purifying selection. In contrast, functional analysis is focused

NOTES

on the "damaging" effect on molecular function. The rationale for this approach is that the effects on phenotypes and fitness must be mediated by the effects on molecular function, even though the converse is not necessarily true. The existence of many common loss-of-function variants in humans (MacArthur et al., 2012) and events of adaptive pseudogenization (Wang et al., 2006) clearly show that damaging alleles may be neutral or beneficial rather than deleterious. It is also feasible that most of human alleles that are subject to purifying selection have no detectable effects on medically relevant phenotypes in the current environment. However, most studies implicitly assume the strong relationship between the effects on molecular function, fitness, and phenotypes. For example, many computational methods for predicting the functional effects of human alleles are based on the inference of purifying selection from comparative genomics data.

Here, I review current strategies to infer causality and functional significance of human protein–coding DNA variants, including variants involved in Mendelian human traits and rare coding variants involved in complex phenotypes.

Inferring the Functional Significance of Missense Mutations Involved in Mendelian Phenotypes

As noted earlier, the problem of assigning functional significance to variants involved in Mendelian phenotypes arises both in the context of gene discovery and in the context of interpreting VUSs in known genes. The overwhelming majority of sequence variants causing Mendelian traits are coding. Among coding variants, "missense" changes are the most difficult to interpret (most of synonymous changes are benign, and most of nonsense or splice-site changes are damaging). Three potential strategies to infer causality and functional significance could be employed: (1) the strategy based on statistical genetics, (2) *in vitro* or *in vivo* experimental analysis, and (3) computational predictions based on evolutionary and structural considerations.

Statistical arguments

In some cases, purely statistical arguments can be employed in favor of the causal relationship between DNA variants and Mendelian traits. Importantly, the arguments discussed below are specific to Mendelian genetics and, in the most part, cannot be applied to variants underlying complex phenotypes.

Analysis of cosegregation of the DNA variant with the phenotype is probably the most accurate method for establishing causality by statistical means. However, at least five informative meioses are needed to support causality (Jordan et al., 2011), and sufficiently large pedigrees are usually unavailable. In addition, segregation analysis may be misleading if more than one rare variant is present in the locus and cosegregate with the phenotype.

Another important consideration is the analysis of allele frequency in unaffected controls. This analysis has been dramatically facilitated by large-scale sequencing efforts such as the 1000 Genomes Project (1000 Genomes Project Consortium, 2010) and Exome Sequencing Project (ESP) (Tennessen et al., 2012). Presence in healthy controls at appreciable frequency may reveal whether the allelic variant is a benign polymorphism segregating in the population, which will exclude the possibility that this variant is involved in the disease phenotype with high penetrance (this approach is obviously noninformative for variance of incomplete penetrance unless larger case–control study is pursued). Although it is easy to infer that the variant is benign (or, at least, not of high penetrance) if it is seen in a number of unaffected individuals, it is much less clear if its absence in multiple controls may serve as a strong support for the pathogenicity. Most importantly, for some genes such as *BRCA1* and *BRCA2*, the number of sequenced cases vastly exceeds the number of sequenced controls, making the analysis of allele frequency in unaffected controls noninformative. Next, differences in global and even local ancestry may complicate conclusions because many rare variants are specific to individual human populations. Also, ESP contains data on individuals with various diseases, so not all sequenced individuals should be automatically assumed to be unaffected.

Even in the simplest possible case of a variant observed in a single patient with a dominant phenotype absent in a panel of ideally ancestry-matched control subjects, the number of control subjects should be very large.

At the first glance, population genetics supports the use of moderate numbers of controls. Under the standard model of a constant size population with no natural selection, the chance that a variant observed in a patient will not be seen in n normal controls is $1/(n + 1)$ (Mitchell et al., 2005) This suggests that if the variant is not found in 100 controls, then the chance for the mutation to have no phenotypic effect is < 1%. Therefore, absence in a moderate number of controls would support pathogenicity. The following

NOTES

factors suggest that this is a stark underestimate for human populations: (1) human population growth, which has resulted in an excess of rare alleles, (2) selection against moderately deleterious alleles, and (3) human migrations, which have resulted in rare alleles not seen in multiple controls (Sunyaev et al., 2000; Marth et al., 2004; Williamson et al., 2005; Kryukov et al., 2007, 2009; Boyko et al., 2008; Li et al., 2010). As seen from Figure 1, a more complex population genetics model incorporating population growth and natural selection (Kryukov et al., 2009) but not migration predicts that there is > 1% chance that a benign variant observed in a single patient would not be detected in as many as 10,000 controls. Taking into account the effects of migration would likely make this number even higher. Therefore, the sole observation of the absence in multiple unaffected controls is insufficient to convincingly imply functional significance of a sequence variant.

In some cases, the evidence for pathogenicity of specific mutations can be provided by the observation of recurrence. For example, independent occurrence of two exactly same mutations has been observed in Baraitser–Winter syndrome (Rivière et al., 2012a). Three different mutations in the same codon have been reported in the analysis of the Myhre syndrome (Le Goff et al., 2011), strongly suggesting the functional importance of this particular amino acid position.

A growing number of publications (Heinzen et al., 2012; Rivière et al., 2012a, b; Van Houdt et al., 2012) report *de novo* mutations as evident from parent–child trio sequencing. The observation of *de novo* mutation in a gene known to be involved in the phenotype (i.e., a gene under an independently reported linkage peak or a gene with multiple *de novo* mutations in other families) is highly informative

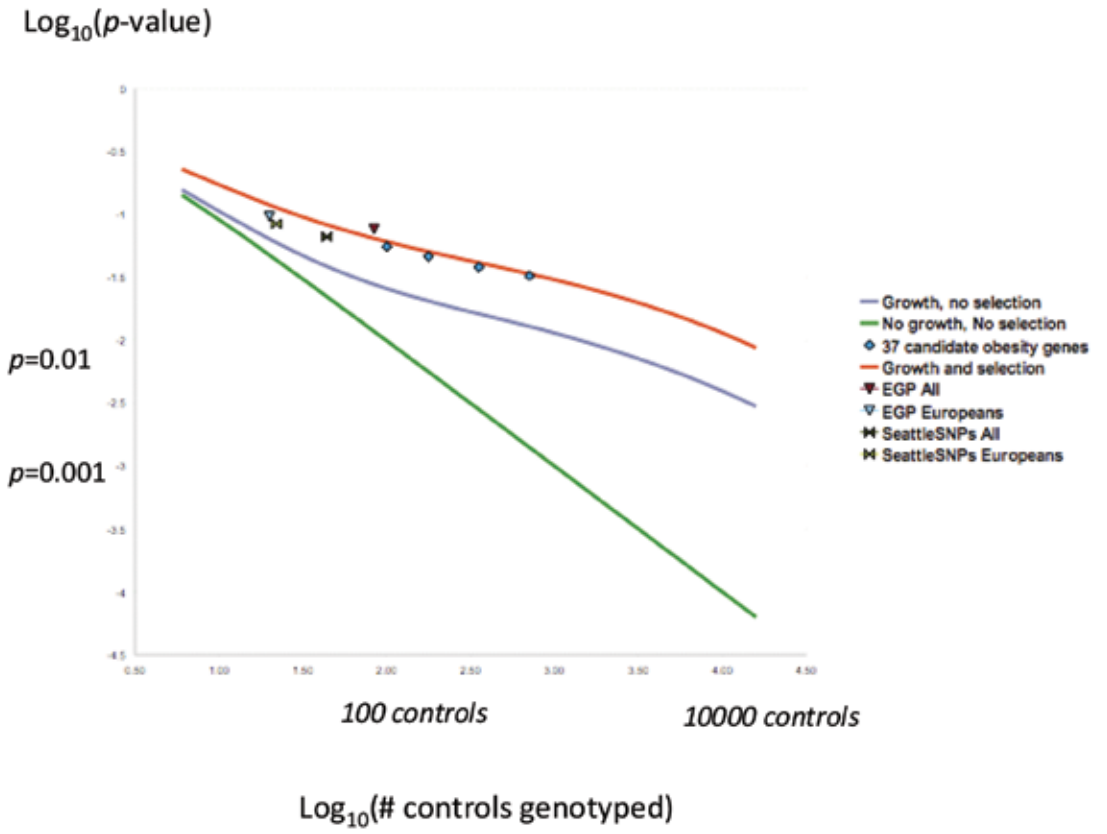

Figure 1. The probability that a nonpathogenic variant observed in a single patient would not be observed in multiple controls. Log–log scale plot is shown for theoretical model assuming constant population and no natural selection (Ahituv et al., 2007) (green line); a population genetic model assuming recent population growth and no natural selection (Kryukov et al., 2009) (blue line); and a population genetic model that incorporates both population growth and natural selection (Kryukov et al., 2009) (red line). Results of theoretical models are shown together with estimates based on real data obtained by resampling from three available systematic resequencing datasets: the Environmental Genome Project dataset, the Seattle SNP dataset (Livingston et al., 2004), and the Obesity Sequencing Study dataset (Ahituv et al., 2007).

NOTES

about the functional significance of the mutation. Indeed, the rate of point mutations in humans is on the order of 10^{-8} per nucleotide per generation and approximately 10^{-5} per protein-coding gene per generation (Nachman and Crowell, 2000; Kondrashov, 2003; Roach et al., 2010; Kong et al., 2012; Sun et al., 2012). Therefore, it is unlikely that a *de novo* mutation unrelated to the phenotype is observed in a known gene. The situation is different, however, in the analysis of whole-exome or whole-genome sequencing without the knowledge of causal genes. Although *de novo* mutations can be considered excellent candidates, especially for dominantly inherited traits, independent functional validation is usually required.

Experimental evidence

Direct experimental functional analysis is a highly laborious but a highly convincing method to study the effect of human allelic variants. Experimental approaches include the analysis of protein expression and localization, *in vitro* functional assays, and genetic manipulation on model organisms. The enthusiasm for direct experimental methods should be accompanied by a cautionary note that specific aspects of molecular function analyzed using *in vitro* assays in some cases may be unrelated to the phenotype, and the effects of mutations on model organisms sometimes may be uninformative about the human condition.

In many cases, missense mutations result in changes of protein expression and localization. Some recent studies relied on immunostaining to assess effects of individual human alleles (Boileau et al., 2012; Wortman et al., 2012). Testing other aspects of protein function requires development of specific functional assays. Phosphorylation assays can be applied for proteins involved in signaling. A recent study of implicated mutations in tyrosine kinase domain of the colony-stimulating factor 1 receptor (*CSF1R*) in hereditary diffuse leukoencephalopathy serves as an example (Rademakers et al., 2011). Autophosphorylation of CSF1R after stimulation with the colony-stimulating factor 1 (CSF1) was used to assay the function of human mutations. Phosphorylation of downstream targets was also examined in the study that identified mutations in *AKT3*, *PIK3R2*, and *PIK3CA*, causing a spectrum of related megalencephaly syndromes (Rivière et al., 2012b).

Changes in protein–protein interactions can be used to detect the effect of mutations on proteins involved in complexes. *In vitro* protein aggregation assay was used to test for the function of the co-chaperone DNAJB6 that was shown to cause limb-girdle muscular dystrophy (Sarparanta et al., 2012).

Functional assay to test lipid metabolism in incubated keranocytes was used in a recent study that linked *PNPLA1* to congenital ichthyosis (Grall et al., 2012). The same study used differentiation assay.

In some cases, mapping mutations on protein three-dimensional structure may provide a key insight into the functional mechanisms. For example, structural localization of *KLHL3* mutations causing familial hyperkalemic hypertension shows spatial clustering that helped to generate a biological hypothesis (Louis-Dit-Picard et al., 2012).

Model organisms amenable to genetic manipulation provide a possibility to test the phenotypic rather than molecular consequences of human allelic variants. The mammalian mouse model has been a model of choice for years to test the phenotypic effect of human genes. However, testing allelic series in the mouse is highly laborious. Therefore, zebrafish is being increasingly used to test the effect of human mutations because this vertebrate species is a powerful genetic model and a convenient system to screen for phenotypes. The approach involves knocking down the fish ortholog of the human gene and assaying the phenotypic effect. Next, if injecting human wild-type mRNA results in a phenotypic rescue, individual alleles can be tested for the potential to rescue the phenotype. Last, coinjection of wild-type and mutant mRNAs provides a test for dominant negative effects. Recent examples of the successful application of this approach include the analysis of mutations in co-chaperone DNAJB6 (Sarparanta et al., 2012) and mutations in the RNA exosome component causing pontocerebellar hypoplasia and spinal motor neuron degeneration (Wan et al., 2012). The zebrafish model was employed with the great success in characterizing multiple variants in several genes involved in ciliopathies (Zaghloul et al., 2011).

In some cases, much more distant model organisms appear helpful in interpreting human mutations. For example, a yeast system was successfully used to functionally characterize 84 human variants observed in patients with cystathionine-β-synthase deficiency (Mayfield et al., 2012).

Interestingly, dog is another species helping to establish the relationship between human mutations and phenotypes (Grall et al., 2012).

NOTES

Computational predictions

Additional supporting evidence for the functional significance of missense mutations can be provided by computational prediction algorithms. At this time, the accuracy of computational predictions is about 75–80% (Hicks et al., 2011), with the accuracy estimates dependent on datasets or databases that are used to define pathogenic and benign variants. Thus, the computational analysis is less informative than direct experimental evidence. However, given that the computational methods do not involve any additional labor and cost and can be applied to any gene, many studies rely, at least in part, on the application of computational methods. The accuracy of the methods can be higher for highly confident predictions (Jordan et al., 2010). If accompanied with rigorous accuracy estimation on a disease-specific dataset, bioinformatically derived prediction information may assist clinical decision-making. Current American College of Medical Genetics (ACMG) and International Agency for Research on Cancer (IARC) recommendations endorse the application of computational methods in genetic diagnostics but only in combination with other criteria (Plon et al., 2008; Richards et al., 2008). As more protein sequences and structures accompanied by training data (known disease-causing mutations and neutral polymorphisms) are available, the classification accuracy will improve. At the same time, principle difficulties in improving the accuracy of prediction methods remain and are discussed at the end of this section.

Despite the variety of approaches that exist (Table 2), these methods all rely heavily on two fundamental observations. First, the regions of proteins that are critical to function evolve under long-term negative selection; thus, when the sequence of a human protein is aligned for comparison to its homologs from other species, these sites will display only specific patterns of amino acid residue variation or complete conservation. The analysis of phylogenetic information in the form of multiple sequence alignment is a powerful source of information about the spectrum of residues allowed at a particular position of the protein of interest (Chasman et al., 2001; Ng et al., 2001; Sunyaev et al., 2001). Second, most pathogenic mutations affect protein stability (Yue et al., 2005; Potapov et al., 2009). In general, the prediction techniques based on protein spatial structure can be applied only if the structure has been resolved for the query protein or its close homolog, which is true only for a minor fraction of human proteins. However, even for the proteins with known spatial structure, structure-based methods work best only in addition to phylogeny-based approaches and provide only a slight increase in the accuracy of the methods (Kumar et al., 2009; Adzhubei et al., 2010).

Although existing methods all rely on evolutionary pattern and, sometimes, protein structure, they differ in algorithmic details. For example, SIFT (Sorting Intolerant from Tolerant) (Ng et al., 2001) and PolyPhen-2 (Adzhubei et al., 2010) estimate the probability that the mutant amino acid would fit the amino acid position given the observed substitution pattern. MAPP (Multivariate Analysis of Protein Polymorphism) (Stone and Sidow, 2005) analyzes conservation of physicochemical properties of amino acids, and LRT (Chun and Fay, 2009) and GERP (Cooper et al., 2010) estimate selective constraint. The methods based on multiple features also differ in the machine learning algorithms they employ. For example, MutationTaster (Schwarz et al., 2010) and PolyPhen-2 (Adzhubei et al. 2010) rely on the naive Bayes classifier, and SNAP (Bromberg et al., 2008) utilizes a neural network. Although different methods use essentially the same information, surprisingly, the methods are commonly discordant. This can be explained only in part by different threshold settings. This observation motivated the development of "umbrella" methods that combine predictions made by different algorithms such as Condel (González-Pérez and López-Bigas, 2011).

The accuracy of the methods could be potentially improved if the scope of the methods were narrower, specifically focused on a single phenotype and a group of genes involved in this phenotype. Such methods employ gene-specific training datasets, gene phylogeny, protein features, and classification rules optimized for a particular set of genes involved in a specific disease. Recently developed methods include a method focused on the *BRCA1* gene, involved in risk of breast and ovarian cancer (Karchin et al., 2007), and a method focused on genes encoding proteins of the heart sarcomere involved in hypertrophic cardiomyopathy (Jordan et al., 2011).

Two important basic effects hamper further development of new prediction methods of higher accuracy. First, the existing approaches may have intrinsic difficulties differentiating between mutations of large effect, important for genetic diagnostics, and slightly deleterious sequence variants in phylogenetically conserved positions, whose existence in genomes of apparently healthy humans is confirmed by numerous resequencing studies. Second, it was shown that human disease mutations are occasionally observed as wild-type alleles in vertebrate orthologs (Kondrashov et al., 2002). Most likely, this is the result of epistatic interactions.

NOTES

Table 2. A selection of online tools for predicting the functional effect of protein coding variants

AlignGVGD	Conservation of physicochemical properties	agvgd.iarc.fr
Condel	Prediction method based on combining other methods	bg.upf.edu/condel
MAPP (Multivariate Analysis of Protein Polymorphism)	Conservation of physicochemical properties	mendel.stanford.edu/sidowlab/downloads/MAPP
MutationTaster	Bayes classifier over multiple sequence features and conservation	mutationtaster.org
PMut	Evolutionary and structural features combined using a machine learning method	mmb2.pcb.ub.es:8080/PMut
PolyPhen-2	Evolutionary and structural features combined using naive Bayes classifier	genetics.bwh.harvard.edu/pph2
SIFT (Sorting Intolerant from Tolerant)	Evolutionary method based on position-specific scoring matrix	sift-dna.org
SNAP	Several evolutionary and structural features combined using a neural network	www.rostlab.org/services/snap
SNPs3D	Combination of a phylogenetic and a structural method; uses support vector machine	www.snps3d.org

Compensatory sequence changes enable amino acid changes corresponding to disease mutations in humans to be benign in a different genetic background. Current prediction methods analyze substitution patterns at individual positions and do not account for epistatic interactions. Compensatory changes should be taken into account to substantially increase the accuracy of computational approaches.

Functional Analysis of Rare Nonsynonymous Variants Involved in Complex Phenotypes

The analysis of complex traits presents a different set of issues. In this review, I limit the discussion to rare nonsynonymous variants in complex traits and leave out the discussion of functional effects of common variants identified by GWAS.

There is a growing interest in the role of rare variants in human complex traits. This interest, combined with the availability of next-generation sequencing technology, propels ongoing whole-exome sequencing studies (Do et al., 2012). For individual very rare variants, the phenotypic effect cannot be identified by the association test. Cosegregation is noninformative about variants involved in complex traits. The existing statistical approaches analyze rare variants collectively, grouping them by gene or pathway (Kiezun et al., 2012). In this approach, the statistical signal provided by functionally significant variants is frequently masked by noise due to benign alleles included in the same statistical test. Candidate gene–based studies showed that focusing on functionally significant alleles can increase statistical signal and, hence, the power to detect an association between the presence of rare variants and complex traits. The signal of association of rare variants in melatonin receptor 1B (*MTNR1B*) with type 2 diabetes increases only if variants that affect melatonin binding are considered (Bonnefond et al., 2012). *In vitro* experiments also helped to increase statistical signal of association in *ANGPL* genes with triglycerides (Romeo et al., 2009).

Statistical power of exome-sequencing studies is expected to be relatively low (Kryukov et al., 2009), so knowledge of functional variants would potentially help identify genes harboring rare variants associated with complex traits. Using experimental approaches at the whole-exome scale is not feasible. Some studies argued that computational methods for predicting the functional effect of human nonsynonymous alleles might be used to increase the power of sequencing studies (Ahituv et al., 2007; Price et al., 2010). Some statistical methods allow for weighting alleles based on potential functional effects. Likely, most sequencing studies would employ both tests weighted with predicted functional significance and tests grouping all nonsynonymous variants, disregarding predicted effect on function.

NOTES

Conclusion

Assigning functional significance to human alleles and inferring the causal relationship between DNA variants and phenotypes remains the central issue in human genetics. The most efficient way forward would combine statistical genetics considerations, *in vivo* and *in vitro* experimental studies, and computational approaches. Low throughput of current experimental methods and insufficient accuracy of computational predictions should be addressed to confidently annotate massive data on human genetic variation from the functional perspective.

An additional issue raising the problem to even a greater level of complexity is that, in many cases, the same functional variant can have different phenotypic consequences varying in both expressivity and penetrance depending on other genetic and environmental factors.

Acknowledgments

I am grateful to Gregory Kryukov for the analysis presented in Figure 1. I also acknowledge support from National Institutes of Health grants R01MH084676 and R01GM078598.

References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA (2007) Medical sequencing at the extremes of human body mass. Am J Hum Genet 80:779–791.

Boileau C, Guo DC, Hanna N, Regalado ES, Detaint D, Gong L, Varret M, Prakash SK, Li AH, d'Indy H, Braverman AC, Grandchamp B, Kwartler CS, Gouya L, Santos-Cortez RL, Abifadel M, Leal SM, Muti C, Shendure J, Gross MS, et al. (2012) *TGFB2* mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. Nat Genet 44:916–921.

Bonnefond A, Clément N, Fawcett K, Yengo L, Vaillant E, Guillaume JL, Dechaume A, Payne F, Roussel R, Czernichow S, Hercberg S, Hadjadj S, Balkau B, Marre M, Lantieri O, Langenberg C, Bouatia-Naji N; Meta-Analysis of Glucose and Insulin-Related Traits Consortium (MAGIC), Charpentier G, Vaxillaire M, et al. (2012) Rare *MTNR1B* variants impairing melatonin receptor 1B function contribute to type 2 diabetes. Nat Genet 44:297–301.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083.

Bromberg Y, Yachdav G, Rost B (2008) SNAP predicts effect of mutations on protein function. Bioinformatics 24:2397–2398.

Chasman D, Adams RM (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:683–706.

Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. Genome Res 19:1553–1561.

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305:869–872.

Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat Methods 7:250–251.

Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482:390–394.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294.

Do R, Kathiresan S, Abecasis G (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet 21:R1–R9.

NOTES

González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88:440–449.

Grall A, Guaguère E, Planchais S, Grond S, Bourrat E, Hausser I, Hitte C, Le Gallo M, Derbois C, Kim GJ, Lagoutte L, Degorce-Rubiales F, Radner FP, Thomas A, Küry S, Bensignor E, Fontaine J, Pin D, Zimmermann R, Zechner R, et al. (2012) *PNPLA1* mutations cause autosomal recessive congenital ichthyosis in golden retriever 4s and humans. Nat Genet 44:140–147.

Heinzen EL, Swoboda KJ, Hitomi Y, Gurrieri F, Nicole S, de Vries B, Tiziano FD, Fontaine B, Walley NM, Heavin S, Panagiotakaki E; European Alternating Hemiplegia of Childhood (AHC) Genetics Consortium; Biobanca e Registro Clinico per l'Emiplegia Alternante (I.B.AHC) Consortium; European Network for Research on Alternating Hemiplegia (ENRAH) for Small and Medium-sized Enterpriese (SMEs) Consortium, Fiori S, Abiusi E, Di Pietro L, Sweney MT, Newcomb TM, Viollet L, et al. (2012) De novo mutations in *ATP1A3* cause alternating hemiplegia of childhood. Nat Genet 44:1030–1034.

Hicks S, Wheeler DA, Plon SE, Kimmel M (2011) Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32:661–668.

International HIV Controllers Study, Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PI, Walker BD, Ripke S, Brumme CJ, Pulit SL, Carrington M, Kadie CM, Carlson JM, Heckerman D, Graham RR, Plenge RM, Deeks SG, Gianniny L, Crawford G, Sullivan J, et al. (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science 330:1551–1557.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40:592–599.

Johansen CT, Wang J, Lanktree MB, McIntyre AD, Ban MR, Martins RA, Kennedy BA, Hassell RG, Visser ME, Schwartz SM, Voight BF, Elosua R, Salomaa V, O'Donnell CJ, Dallinga-Thie GM, Anand SS, Yusuf S, Huff MW, Kathiresan S, Hegele RA (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 42:684–687.

Jordan CT, Cao L, Roberson ED, Duan S, Helms C.A, Nair RP, Duffin KC, Stuart PE, Goldgar D, Hayashi G, Olfson EH, Feng BJ, Pullinger CR, Kane JP, Wise CA, Goldbach-Mansky R, Lowes MA, Peddle L, Chandran V, Liao W, et al. (2012) Rare and common variants in *CARD14*, encoding an epidermal regulator of NF-kappaB, in psoriasis. Am J Hum Genet 90:796–808.

Jordan DM, Ramensky VE, Sunyaev SR (2010) Human allelic variation: perspective from protein function, structure, and evolution. Curr Opin Struct Biol 20:342–350.

Jordan DM, Kiezun A, Baxter SM, Agarwala V, Green RC, Murray MF, Pugh T, Lebo MS, Rehm HL, Funke BH, Sunyaev SR (2011) Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. Am J Hum Genet 88:183–192.

Karchin R, Monteiro AN, Tavtigian SV, Carvalho MA, Sali A (2007) Functional impact of missense variants in *BRCA1* predicted by supervised learning. PLoS Comput Biol 3:e26.

Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL (2012) Exome sequencing and the genetic basis of complex traits. Nat Genet 44:623–630.

Kondrashov AS (2003) Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21:12–27.

Kondrashov AS, Sunyaev S, Kondrashov FA (2002) Dobzhansky–Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 99:14878–14883.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, et al. (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. Nature 488:471–475.

Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. Am J Hum Genet 80:727–739.

Kryukov, GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR (2009) Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci USA 106:3871–3876.

NOTES

Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ (2009) Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. Genome Res 19:1562–1569.

Le Goff C, Mahaut C, Abhyankar A, Le Goff W, Serre V, Afenjar A, Destrée A, di Rocco M, Héron D, Jacquemont S, Marlin S, Simon M, Tolmie J, Verloes A, Casanova JL, Munnich A, Cormier-Daire V (2011) Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre syndrome. Nat Genet 44:85–88.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, Grarup N, Guo Y, Hellman I, Jin X, Li Q, Liu J, Liu X, Sparsø T, Tang M, Wu H, et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet 42:969–972.

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831.

Louis-Dit-Picard H, Barc J, Trujillano D, Miserey-Lenkei S, Bouatia-Naji N, Pylypenko O, Beaurain G, Bonnefond A, Sand O, Simian C, Vidal-Petiot E, Soukaseum C, Mandet C, Broux F, Chabre O, Delahousse M, Esnault V, Fiquet B, Houillier P, Bagnis CI, et al. (2012) *KLHL3* mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. Nat Genet 44:456–460.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, et al. (2012) A systematic survey of loss-of-function variants in human protein–coding genes. Science 335:823–828.

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166:351–372.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science 337:1190–1195.

Mayfield JA, Davies MW, Dimster-Denk D, Pleskac N, McCarthy S, Boydston EA, Fink L, Lin XX, Narain AS, Meighan M, Rine J (2012) Surrogate genetics and metabolic profiling for characterization of human disease alleles. Genetics 190:1309–1323.

McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, Erdos MR, Scott LJ, Morken MA, Kucera KS, Battenhouse A, Keefe D, Collins FS, Willard HF, Lieb JD, Furey TS, Crawford GE, Iyer VR, Birney E (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328:235–239.

Mitchell AA, Chakravarti A, Cutler DJ (2005) On the probability that a novel variant is a disease-causing mutation. Genome Res 15:960–966.

Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L, Cleynen I, Colombel JF, de Rijk P, Dewit O, Finkel Y, Gassull MA, Goossens D, Laukens D, Lémann M, Libioulle C, O'Morain C, Reenaers C, Rutgeerts P, Tysk C, et al. (2011) Resequencing of positional candidates identifies low frequency *IL23R* coding variants protecting against inflammatory bowel disease. Nat Genet 43:43–47.

Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, Pirruccello JP, Muchmore B, Prokunina-Olsson L, Hall JL, Schadt EE, Morales CR, Lund-Katz S, Phillips MC, Wong J, Cantley W, et al. (2010) From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. Nature 466:714–719.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156:297–304.

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, et al. (2012) Patterns and rates of exonic *de novo* mutations in autism spectrum disorders. Nature 485:242–245.

NOTES

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–874.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010) Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42:30–35.

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. Nature 485:246–50.

Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV; IARC Unclassified Genetic Variants Working Group (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 29:1282–1291.

Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 22:553–560.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838.

Rademakers R, Baker M, Nicholson AM, Rutherford NJ, Finch N, Soto-Ortolaza A, Lash J, Wider C, Wojtas A, DeJesus-Hernandez M, Adamson J, Kouri N, Sundal C, Shuster EA, Aasly J, MacKenzie J, Roeber S, Kretzschmar HA, Boeve BF, Knopman DS, et al. Mutations in the colony stimulating factor 1 receptor (*CSF1R*) gene cause hereditary diffuse leukoencephalopathy with spheroids. Nat Genet 44:200–205.

Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X, Alfredsson L, Padyukov L, Klareskog L, Worthington J, Siminovitch KA, Bae SC, Plenge RM, Gregersen PK, de Bakker PI (2012) Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat Genet 44:291–296.

Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE, Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med 10:294–300.

Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagacé C, Neale B, et al. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 43:1066–1073.

Rivière JB, Mirzaa GM, O'Roak BJ, Beddaoui M, Alcantara D, Conway RL, St-Onge J, Schwartzentruber JA, Gripp KW, Nikkel SM, Worthylake T, Sullivan CT, Ward TR, Butler HE, Kramer NA, Albrecht B, Armour CM, Armstrong L, Caluseriu O, Cytrynbaum C, et al. (2012a) *De novo* germline and postzygotic mutations in *AKT3*, *PIK3R2* and *PIK3CA* cause a spectrum of related megalencephaly syndromes. Nat Genet 44:934–940.

Rivière JB, van Bon BW, Hoischen A, Kholmanskikh SS, O'Roak BJ, Gilissen C, Gijsen S, Sullivan CT, Christian SL, Abdul-Rahman OA, Atkin JF, Chassaing N, Drouin-Garraud V, Fry AE, Fryns JP, Gripp KW, Kempers M, Kleefstra T, Mancini GM, Nowaczyk MJ, et al. (2012b) *De novo* mutations in the actin genes *ACTB* and *ACTG1* cause Baraitser–Winter syndrome. Nat Genet 44:440–444.

Roach JC, Glusman G, Smit AF, Huff C.D, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:67–639.

Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC (2009) Rare loss-of-function mutations in *ANGPTL* family members contribute to plasma triglyceride levels in humans. J Clin Invest 119:70–79.

NOTES

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, et al. (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485:237–241.

Sarparanta J, Jonson PH, Golzio C, Sandell S, Luque H, Screen M, McDonald K, Stajich JM, Mahjneh I, Vihola A, Raheem O, Penttilä S, Lehtinen S, Huovinen S, Palmio J, Tasca G, Ricci E, Hackman P, Hauser M, Katsanis N, et al. (2012) Mutations affecting the cytoplasmic functions of the co-chaperone DNAJB6 cause limb-girdle muscular dystrophy. Nat Genet 44:450–455.

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7:575–576.

Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15:978–986.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis E (2007) Population genomics of human gene expression. Nat Genet 39:1217–1224.

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, Stefansson K (2012) A direct characterization of human mutation based on microsatellites. Nat Genet 44:1161–1165.

Sunyaev SR, Ramensky VE, Koch I, Lathe W 3d, Kondrashov AS, Bork P (2001) Prediction of deleterious human alleles. Hum Mol Genet 15:591–597.

Sunyaev SR, Lathe WC 3d, Ramensky VE, Bork P (2000) SNP frequencies in human genes an excess of rare alleles and differing modes of selection. Trends Genet 16:335–337.

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337:64–69.

Van Houdt JK, Nowakowska BA, Sousa SB, van Schaik BD, Seuntjens E, Avonce N, Sifrim A, Abdul-Rahman OA, van den Boogaard MJ, Bottani A, Castori M, Cormier-Daire V, Deardorff MA, Filges I, Fryer A, Fryns JP, Gana S, Garavelli L, Gillessen-Kaesbach G, Hall BD, et al. (2012) Heterozygous missense mutations in *SMARCA2* cause Nicolaides-Baraitser syndrome. Nat Genet 44:445–449.

Wan J, Yourshaw M, Mamsa H, Rudnik-Schöneborn S, Menezes MP, Hong JE, Leong DW, Senderek J, Salman MS, Chitayat D, Seeman P, von Moers A, Graul-Neumann L, Kornberg AJ, Castro-Gago M, Sobrido MJ, Sanefuji M, Shieh PB, Salamon N, Kim RC, et al. (2012) Mutations in the RNA exosome component gene *EXOSC3* cause pontocerebellar hypoplasia and spinal motor neuron degeneration. Nat Genet 44:704–708.

Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS Biol 4:e52.

Williamson SH, Hernandez R, Feldel-Alon A, Zhe L, Nielsen R, Bustamante CD (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA 102:7882–7887.

Wortmann SB, Vaz FM, Gardeitchik T, Vissers LE, Renkema GH, Schuurs-Hoeijmakers JH, Kulik W, Lammens M, Christin C, Kluijtmans LA, Rodenburg RJ, Nijtmans LG, Grünewald A, Klein C, Gerhold JM, Kozicz T, van Hasselt PM, Harakalova M, Kloosterman W, Barić I, et al. (2012) Mutations in the phospholipid remodeling gene *SERAC1* impair mitochondrial function and intracellular cholesterol trafficking and cause dystonia and deafness. Nat Genet 44:797–802.

Wray NR, Purcell, SM, Visscher PM (2011) Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol 9:e1000579.

Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M (2011) Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. Nat Genet 43:864–868.

Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459–473.

Zaghloul NA, Katsanis N (2011) Zebrafish assays of ciliopathies. Methods Cell Biol 105:257–272.