doi:10.1038/nature13127

Factors to consider in assessing candidate pathogenic mutations in presumed monogenic conditions

The questions itemized below expand upon the definitions in Table 1 and are provided to the reader interested in a more in-depth assessment of the evidence base implicating a gene or variant in disease. None of these factors is intended to represent definitive or necessary evidence for implication; instead, investigators and reviewers should consider the overall weight of the evidence supporting causality across the areas described below, and where possible assess their probability using formal statistical approaches as described in the manuscript text.

Importantly, we do not discriminate between previously reported/published genes or variants and those being newly implicated. In both cases we encourage investigators to consider the full spectrum of evidence – both previously published and newly generated – for and against implication.

Gene-level evidence for implication:

- Causal variants in the gene have been confidently implicated in multiple independent families with the same or closely related disease phenotypes
- The probability of the distribution of variants seen in the candidate gene in disease patients in low in a large, appropriately matched reference panel, with an appropriately low reported *P* value
- The overall evidence for segregation of rare missense and loss-of-function variants with disease status within affected families is strong (LOD score >3)
- The background frequency of the class of variation observed in the affected individual in this gene (for instance "compound heterozygosity for two rare missense mutations") is low, with an appropriately low reported P value

- For a novel candidate gene with proposed recessive inheritance, no unaffected individuals have been observed with homozygosity for high-confidence loss of function variants in the candidate gene
- The gene is expressed in tissues relevant to the disease
- The overall expression profile of the gene resembles that of other genes mutated in this disease or phenotypically similar diseases
- The protein encoded by the gene interacts with proteins encoded by other genes mutated in this disease or phenotypically similar diseases
- For expression and protein-protein interaction analyses, formal statistical significance has been assessed, and where possible placed within a genome-wide context (i.e. compared to the distribution of the same metric for randomly selected genes)

<u>Variant-level evidence for implication</u>

- Previous reports of the same variant:
    - The variant previously been reported as a causal variant in this disease or a phenotypically similar disease
    - The patient phenotype is consistent with the phenotype observed in other manifesting variant carriers
    - The overall weight of evidence supporting causality in published reports is clearly described, so that the true level of evidence for pathogenicity can be accurately assessed
- Segregation:
    - The variant segregates appropriately in members of the examined pedigrees, given the proposed penetrance and mode of inheritance
    - The formal LOD score associated with segregation has been calculated, as has the maximum possible LOD score given the size of the pedigree
    - All available family members – both affected and unaffected – been surveyed for the variant

  - If the variant is reported to be a *de novo* mutation, the parents have been assessed with sufficient confidence to rule them out as carriers, as has the possibility of non-paternity
  - If an individual has been reported to be compound heterozygous for two pathogenic mutations, parental genotyping or a direct experimental approach has been used to confirm that the variants are in *trans* (i.e. on separate haplotypes)
- Frequency:
  - Appropriate public databases (see Resources) been used to examine variant frequency across multiple populations and demonstrated it to be low or absent
  - Variant frequency been estimated using an appropriately large sample (preferably >5,000 chromosomes) of individuals drawn from the same population as the patient and demonstrated to be low or absent
  - If the variant is seen in unaffected individuals, it is found at a frequency consistent with the proposed mode of inheritance and the known incidence of the disease in any examined population
  - The same genotype observed in the affected individual has not been observed in an unaffected individual
- Other variants in affected individuals:
  - Sequencing has been performed in the patient on other genes known to be associated with this disease and has failed to identify potentially causal variants
  - A sufficiently high (and quantitated) fraction of the bases in these known genes, and of known mutations associated with this disease, were sequenced sufficiently well to confidently rule out the presence of variation
  - If the mutated gene has previously been reported to show recessive inheritance, both copies carry mutations in the affected individual

  - The haplotype carrying the variant does not carry other variants predicted to alter its functional impact (such as another SNV in the same codon that alters the effect on protein sequence, or a second indel predicted to restore the reading frame following a frameshift variant)
- Functional annotation:
  - For missense substitutions or in-frame indels: multiple conservation-based metrics support the potential deleteriousness of the variant
  - For predicted truncating mutations: the variant found upstream of the last 50 bases of the penultimate exon and therefore most likely to cause nonsense-mediated decay, and/or the truncated portion of the gene is highly conserved or known to be functionally important
  - For predicted splice-disrupting or splice-creating variants: multiple splice prediction algorithms been assessed and all support disruption, and is there is little possibility of in-frame rescue by nearby cryptic splice sites
  - The affected exon or transcript (in the event of alternatively spliced genes) is expressed in the tissue(s) relevant to the disease
- Experimental support:
  - Disruption of the affected gene has been experimentally demonstrated in primary tissue or cell lines from affected individuals
  - Disruptive potential of the observed sequence changes has been demonstrated in artificial tissue culture or animal models
  - The assays used to assess functionality provide a sensible analogue for the intact biological system of interest
  - The results are unusual in a genome-wide context: in other words, a similar result is unlikely to have been obtained if the same assay was performed on a randomly selected genes or variant

Reporting evidence for implication

We propose that all variants reported to be implicated in a severe monogenic disease satisfy the following criteria wherever possible:

- The report includes the chromosomal coordinates (relative to a specified version of the human reference sequence), and the predicted nucleotide and protein changes in HGVS nomenclature
- All of the genetic, informatic and experimental methods used to assess implication have been described in sufficient detail for external reviewers to assess the overall support for implication
- The variant has been assigned a confidence level for pathogenicity using a 5 tiered scale (Box 2)
- The variant has been submitted, with appropriate supporting evidence, to a central mutation repository such as ClinVar
- The raw sequence data for the individuals assessed in the study has been deposited in a sequence repository such as dbGaP

Table S1. Resources for variant interpretation

Databases of reported disease-causing mutations		
Human Gene Mutation Database (HGMD)	http://www.hgmd.cf.ac.uk/	Catalogue of published disease variants; recent content requires subscription fee
Online Mendelian Inheritance in Man (OMIM)	http://www.omim.org/	Non-comprehensive sampling of published disease variants with detailed associated information
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/	National Center for Biotechnology Information database of annotated human variation
DECIPHER	http://decipher.sanger.ac.uk/	Protected-access database of genomic deletions and duplications seen in clinical samples
Locus-specific databases	Various locations, many listed at http://www.hgvs.org/dblist/glsdb.html	Thousands of databases hosted either using Leiden Open Variation Database structure or custom systems to annotate variants for individual genes or disease-centric sets of genes
Databases of genetic variation		
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/	Central database of reported single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels); contains data submitted from many different sources, often lacking detailed frequency data
HapMap	http://hapmap.ncbi.nlm.nih.gov/	Database with a focus on common variants (>5% frequency) genotyped across multiple populations
1000 Genomes	http://www.1000genomes.org/	Ongoing project applying low-coverage whole-genome sequenced and targeted exome sequencing to 2,500 individuals with diverse ancestries; raw individual-level variant data available
dbVar, DGV, DGVa	http://www.ncbi.nlm.nih.gov/dbvar/ http://projects.tcag.ca/variation/ http://www.ebi.ac.uk/dgva/	Resources for analysis of large-scale structural genomic variants
Exome Variant Server	http://evs.gs.washington.edu/EVS/	Public database of SNP/indel frequencies from over 6,500 European and African-American individuals from the NHLBI Exome Sequencing Project.
General tools for annotation of sequence variants		
Variant Effect Predictor (Ensembl)	http://www.ensembl.org/info/docs/variation/vep/index.html	
Variant Annotation Tool	http://vat.gersteinlab.org/	
ANNOVAR	http://www.openbioinformatics.org/annovar/	
snpEff	http://snpeff.sourceforge.net/	
Deleteriousness prediction algorithms		
PolyPhen2[1]	http://genetics.bwh.harvard.edu/pph2/	Predictions based on eight sequence and three structure based features
SIFT[2]	http://sift.jcvi.org/	Conservation based predictions
MutationTaster[3]	http://www.mutationtaster.org/	Predictions based on conservation, splice site changes, and alterations in protein and mRNA
PhD-SNP	http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html	
PhyloP	http://compgen.bscb.cornell.edu/phast/index.php	Provides conservation scores for both protein-coding and non-coding regions.
GERP[4]	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	Conservation based scores for both protein-coding and non-coding regions.
ConDel[5]	http://bg.upf.edu/condel/home	Method for combining deleteriousness scores across methods (e.g. PolyPhen2, SIFT, Mutation Taster, etc.)
Logit model [6]		Like ConDel, a method for combining deleteriousness scores across methods.
Splice prediction algorithms		
NNSplice	http://www.fruitfly.org/seq_tools/	

	splice.html
NetGene2	http://www.cbs.dtu.dk/services/ NetGene2/
Human Splicing Finder	http://www.umd.be/HSF/

References

1. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010).
2. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucl. Acids Res.* **31**, 3812–3814 (2003).
3. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods* **7**, 575–576 (2010).
4. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
5. González-Pérez, A. & López-Bigas, N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *Am J Hum Genet* **88**, 440–449 (2011).
6. Li, M.-X. *et al.* Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. *PLoS Genet* **9**, e1003143 (2013).