# The importance of dynamic re-analysis in diagnostic whole exome sequencing

Anna C Need,[1] Vandana Shashi,[2] Kelly Schoch,[2] Slavé Petrovski,[3,4] David B Goldstein[3]

Exome sequencing technologies are constantly evolving, with exome capture systems covering more coding bases, and the continual development of improved alignment and variant-calling programmes. At the same time, new genes are frequently being implicated in Mendelian genetic disease. In many cases, therefore, the generation of extra coverage, updating of alignment and variant calling tools and regular inspection for novel gene-disease associations emerging in the literature will yield a diagnosis that was not found in the initial analysis.

The rates of diagnosis with exome sequencing range from 25% to 40%.[1] The diagnosis rate depends on various factors including how patients are selected, the degree of genetic prescreening, the age and ancestry of the population and what is defined as a probable diagnosis. Some of those that remain undiagnosed will not, in fact, have a Mendelian genetic disorder, for example, those with disorders due to mutations in mitochondrial genes, somatic mutations and those with oligogenic or more complex genetic disorders. However, there are many ways that patients with a relevant *Mendelian* pathogenic genetic variant may not obtain a diagnosis in the initial analysis. These can be divided into two broad classes.

1. *The variant is not identified*. The simplest reason for patients remaining undiagnosed is that the pathogenic variant is not identified. This may be because it is in a region not included in the exome sequence, for example, intronic or intergenic variants, or because that site is just poorly covered in that individual due to fluctuations in coverage.[2] Other variant sites may be well covered but the variants themselves are not easily discoverable by current bioinformatic tools, for example, repeat polymorphisms and structural variants, or single nucleotide variants or small insertion/deletion polymorphisms in regions of local genomic complexity.
2. *The variant is not recognised as pathogenic*. This may be for a number of reasons. First, the variant itself may appear innocuous. We all contain in our genomes hundreds of gene-damaging variants,[3] and very rare variants that do not appear in any databases, so distinguishing between those that are and are not contributing to disease is the main hurdle in diagnostic exome sequencing. The obvious candidates are very rare, clearly damaging mutations such as nonsense or frameshift variants, or variants that affect splicing. Dominant disease is often much easier because many of the causal variants are de novo, whereas there can be a lot of candidate compound heterozygotes for autosomal recessives. Sometimes pathogenic genetic variants are not obviously damaging, for example, synonymous variants can sometimes cause disease by affecting splicing. But because most synonymous genetic variants are benign, they will often appear in disease genes and will largely be ignored when interpreting a genome for diagnostic purposes. Alternatively, a recognisably damaging pathogenic variant may be filtered out because it appears in unaffected parents or public databases of unaffected individuals due to reduced penetrance, or based on somatic variant calls.

Other variants that may not be recognised as pathogenic are those that are not in known disease genes. We are increasingly able to recognise genes that are likely to be pathogenic, using measures of their 'intolerance' to damaging variation such as the Residual Variation Intolerance Score (RVIS) score[4] and the probability of being Loss-of-function Intolerant (pLI) score in Exome Aggregation Consortium (ExAC), a publicly available database of variants from over 60 000 sequenced exomes.[3] However, if you only have a single patient with a damaging mutation in a gene previously unlinked to disease, it is very unlikely that patient would receive a genetic diagnosis based on this, however intolerant that gene is predicted to be. In these situations, the gene should be regularly investigated using databases such as GeneMatcher (https://genematcher.org/) and PhenomeCentral (https://phenomecentral.org/) available through the Matchmaker Exchange (http://www.matchmakerexchange.org/) to see if other patients have been reported with a similar phenotype with a variant in the same gene. A recent report indicated that 10% of patients with an initially negative whole exome sequence (WES) were subsequently diagnosed based just on inspection of novel disease-association literature.[5]

To illustrate this, we re-analysed the 6 unsolved trios from our 2012 study of 12 trios with unidentified presumed Mendelian disorders.[6] Of 12 trios with varied presentations who had already undergone thorough diagnostic workups, we originally identified a complete genetic diagnosis for 6, and a partial diagnosis (in which a gene mutation explains part but probably not all of the phenotype) for 1.

At the time of our original report, the sequence reads were aligned to genome build 36, and variants were called with SAMtools.[7] After realignment to build 37, and variant calling with Genome Analysis ToolKit (GATK),[8] two new diagnoses were made. In trio 8, a known pathogenic variant, R246C (rs122445105), was found in *ATRX* which causes the X linked recessive α-thalassaemia/mental retardation syndrome. The maternally inherited variant was hemizygous in the patient. The patient's phenotypic features of growth retardation, profound intellectual disability, hypospadias and dysmorphic facial features are a good phenotypic fit for *ATRX*, although he does not have anaemia, which would have increased clinical suspicion of this disorder and the bicoronal craniosynostosis that he was born with remains unexplained. In trio 12, a heterozygous de novo nonsense mutation (chr16:30748691C>T, R2444*) was identified in *SRCAP*, which fits the patient's clinical diagnosis of Floating-Harbor syndrome. Although this gene was specifically searched for pathogenic variants before realignment, nothing of interest was observed based on the earlier alignment and variant calling.

Adding coverage for the remaining unsolved trios revealed that patient 10 had a heterozygous de novo frameshift variant (chr6:157454179CAAAG>C, R798TfsTer46) in *ARID1B*, a mutation

[1]Division of Brain Sciences, Department of Medicine, Imperial College London, London, UK; [2]Department of Pediatrics, Division of Medical Genetics, Duke University School of Medicine, Durham, North Carolina, USA; [3]Institute for Genomic Medicine, Columbia University, New York, New York, USA; [4]Department of Medicine, The University of Melbourne, Austin Health and Royal Melbourne Hospital, Melbourne, Victoria, Australia

**Correspondence to** Dr Anna C Need, Division of Brain Sciences, Department of Medicine, Imperial College London, Hammersmith Campus, 7th floor Commonwealth Building, Du Cane Road, London W12 0NN, UK; a.need@imperial.ac.uk

that would be expected to result in Coffin-Siris syndrome, a good clinical fit. The variant was not initially identified because the WES coverage and alignment at this site was poor.

Inspection of recent literature[9] indicates that a de novo splice-acceptor mutation in HNRNPU in patient 6 that was classified as an 'interesting finding' in the original report can now be considered to be a likely cause of (at least) the patient's epilepsy and intellectual disability.

Finally, we note that in our original report we identified compound heterozygous loss-of-function variants in *NGLY1* in a patient who had a phenotype resembling a congenital disorder of glycosylation. The gene was not a known disease gene, but because the clinical phenotype was biologically consistent with NGLY1 dysfunction and the patient had a near-absence of NGLY1 protein expression, the finding was reported back to the family as likely causal. Since then (largely as a result of efforts by the parents), NGLY1 deficiency has become a recognised genetic disorder and many other patients have been diagnosed.[10] In WES diagnostic pipelines that focus only on known disease genes, this finding would have been missed, emphasising the value of careful interpretation in the absence of known disease associations.

Of interest, the only patient who remained without a diagnosis was the only patient of African ancestry. Patient 9 has a number of new, damaging genotypes in genes that are intolerant to genetic variation but not yet associated with disease, including *FAM134C* and *MSI1*, which may yet prove to be causal. Because WES control databases often include relatively small numbers of individuals from populations of non-European ancestry, it is harder, during diagnostic sequencing of patients from these populations, to separate the pathogenic variants from the rare benign background genetic variation. This results in patients of non-European ancestry having longer, less accurate lists of candidate variants, creating potential healthcare disparities.[11 12]

This re-analysis demonstrates that with periodic assimilation and analysis of new data, rates of genetic diagnosis with WES can be substantially >25%–40%, and we suggest that a multifaceted approach to re-analysing the WES data should be a standard part of clinical diagnostic paradigms. We recognise that our diagnostic rate of 11/12, with re-analyses over time, is much higher than one would anticipate. Our extremely high rate is likely to be because we carefully selected cases whose clinical features were strongly suggestive of Mendelian disorders and excluded patients with potential non-genetic contributors to disease. Current clinical referrals for WES likely include patients whose features are not as strongly indicative of Mendelian disorders, and may not have as clearly ruled out other possible non-genetic factors. Nonetheless, we believe that with time and analysis of new data, rates of diagnosis with WES will continue to increase within most cohorts.

**Acknowledgements** We thank the patients and their families who gave us the opportunity to carry out this work.

**Contributors** VS and KS helped to interpret the exome data and contributed to the writing of the paper, SP, DBG and ACN contributed to the re-analysis of the exome data and the writing of the paper.

**Competing interests** None declared.

**Ethics approval** Duke IRB.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**To cite** Need AC, Shashi V, Schoch K, *et al*. *J Med Genet* 2017;**54**:155–156.

Received 14 September 2016
Revised 1 November 2016
Accepted 2 November 2016
Published Online First 29 November 2016

*J Med Genet* 2017;**54**:155–156.
doi:10.1136/jmedgenet-2016-104306

## REFERENCES

1 Need AC, Goldstein DB. Neuropsychiatric genomics in precision medicine: diagnostics, gene discovery, and translation. *Dialogues Clin Neurosci* 2016;18:237–52.
2 Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* 2014;15:121–32.
3 Lek M, Karczewski KJ, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285–91.
4 Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013;9:e1003709.
5 Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet Med* Published Online First 21 Jul 2016. http://dx.doi.org/10.1038/gim.2016.88
6 Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, Meisler MH, Goldstein DB. Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet* 2012;49:353–61.
7 Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27:2987–93.
8 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303.
9 de Kovel CG, Brilstra EH, van Kempen MJ, Van't Slot R, Nijman IJ, Afawi Z, De Jonghe P, Djémié T, Guerrini R, Hardies K, Helbig I, Hendrickx R, Kanaan M, Kramer U, Lehesjoki AE, Lemke JR, Marini C, Mei D, Møller RS, Pendziwiat M, Stamberger H, Suls A, Weckhuysen S. EuroEPINOMICS RES Consortium, Koeleman BP. Targeted sequencing of 351 candidate genes for epileptic encephalopathy in a large cohort of patients. *Mol Genet Genomic Med* 2016;4:568–80.
10 Enns GM, Shashi V, Bainbridge M, Gambello MJ, Zahir FR, Bast T, Crimian R, Schoch K, Platt J, Cox R, Bernstein JA, Scavina M, Walter RS, Bibb A, Jones M, Hegde M, Graham BH, Need AC, Oviedo A, Schaaf CP, Boyle S, Butte AJ, Chen R, Chen R, Clark MJ, Haraksingh R, FORGE Canada Consortium, Cowan TM, He P, Langlois S, Zoghbi HY, Snyder M, Gibbs RA, Freeze HH, Goldstein DB. Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genet Med* 2014;16:751–8.
11 Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. *Trends Genet* 2009;25:489–94.
12 Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome Biol* 2016;17:157.