Author/s:
Cowley, MJ;Liu, YC;Oliver, KL;Carvill, G;Myers, CT;Gayevskiy, V;Delatycki, M;Vlaskamp, DRM;Zhu, Y;Mefford, H;Buckley, MF;Bahlo, M;Scheffer, IE;Dinger, ME;Roscioli, T

Title:
Reanalysis and optimisation of bioinformatic pipelines is critical for mutation detection

Date:
2019-04-01

Citation:
Cowley, M. J., Liu, Y. C., Oliver, K. L., Carvill, G., Myers, C. T., Gayevskiy, V., Delatycki, M., Vlaskamp, D. R. M., Zhu, Y., Mefford, H., Buckley, M. F., Bahlo, M., Scheffer, I. E., Dinger, M. E. & Roscioli, T. (2019). Reanalysis and optimisation of bioinformatic pipelines is critical for mutation detection. Human Mutation, 40 (4), pp.374-379. https://doi.org/10.1002/humu.23699.

Persistent Link:
https://hdl.handle.net/11343/253377

Received: 28 June 2018 | Revised: 10 December 2018 | Accepted: 13 December 2018

DOI: 10.1002/humu.23699

BRIEF REPORT

WILEY HGVS HUMAN GENOME VARIATION SOCIETY

Reanalysis and optimisation of bioinformatic pipelines is critical for mutation detection

Mark J Cowley[1,2*] | Yu-Chi Liu[3,4,5] | Karen L. Oliver[3,4] | Gemma Carvill[6] | Candace T. Myers[7] | Velimir Gayevskiy[1] | Martin Delatycki[8] | Danique R.M. Vlaskamp[4] | Ying Zhu[9] | Heather Mefford[7] | Michael F. Buckley[10] | Melanie Bahlo[3,5] | Ingrid E. Scheffer[4,11,12] | Marcel E. Dinger[1,2] | Tony Roscioli[13,14,15]

[1]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia
[2]St Vincent's Clinical School, University of New South Wales, Darlinghurst, Australia
[3]Population Health and Immunity Division, Walter and Eliza Hall Institute, Melbourne, Australia
[4]Epilepsy Research Centre, Department of Medicine, University of Melbourne, Austin Health, Heidelberg, Australia
[5]Department of Medical Biology, The University of Melbourne, Parkville, Victoria, Australia
[6]Ken and Ruth Davee Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL
[7]Department of Pediatrics, University of Washington, Seattle, WA
[8]Austin Health, Melbourne, Australia
[9]Department of Medical Genetics, Royal North Shore Hospital, St Leonards, Australia
[10]NSW Health Pathology Randwick, Sydney, Australia
[11]Florey Institute, Melbourne, Australia
[12]Department of Paediatrics, University of Melbourne, Royal Children's Hospital, Parkville, Australia
[13]Centre for Clinical Genetics, Sydney Children's Hospital, Randwick, Australia
[14]Prince of Wales Clinical School, University of New South Wales, Sydney, Australia
[15]Neuroscience Research Australia, University of New South Wales, Randwick, Sydney, Australia

Correspondence
Dr Mark Cowley, Computational Biology Group Leader, Children's Cancer Institute, Lowy Cancer Centre, UNSW Sydney, Randwick, NSW, 2031, Australia.
Email: MCowley@ccia.org.au

*Present address: Children's Cancer Institute, University of New South Wales, Randwick, Sydney, Australia.

Funding information
W.G.S. was funded by the Kinghorn Foundation. M.J.C. was supported by Cancer Institute NSW (13/ECF/1-46) and an NSW Health Early-Mid Career Fellowship. M.B. was supported by NHMRC Program grant (ID: 1054618) and NHMRC Senior Research Fellowship (ID: 1102971). I.E.S. is supported by NHMRC Program grant (1091593, 2016–2020) and Senior Practitioner Fellowship (1104831, 2016–2020). T.R. was supported through a project grant from the NHMRC (ID:AU/1/BA51117)

Communicated by Graham R. Taylor

Abstract

Rapid advances in genomic technologies have facilitated the identification pathogenic variants causing human disease. We report siblings with developmental and epileptic encephalopathy due to a novel, shared heterozygous pathogenic 13 bp duplication in *SYNGAP1* (c.435_447dup, p.(L150Vfs*6)) that was identified by whole genome sequencing (WGS). The pathogenic variant had escaped earlier detection via two methodologies: whole exome sequencing and high-depth targeted sequencing. Both technologies had produced reads carrying the variant, however, they were either not aligned due to the size of the insertion or aligned to multiple major histocompatibility complex (MHC) regions in the hg19 reference genome, making the critical reads unavailable for variant calling. The WGS pipeline followed different protocols, including alignment of reads to the GRCh37 reference genome, which lacks the additional MHC contigs. Our findings highlight the benefit of using orthogonal clinical bioinformatic pipelines and all relevant inheritance patterns to re-analyze genomic data in undiagnosed patients.

KEYWORDS
clinical bioinformatics, de novo, developmental and epileptic encephalopathy, whole genome sequencing

TABLE 1 Phenotypic features present in the proband and affected sibling

Clinical features	Proband	Sibling
Absence seizures with eyelid myoclonia (HP:0011149)	Yes	Yes
Myoclonic atonic seizures (HP:0011170)	Yes	Yes
Myoclonic seizures (HP:0002123)	Yes	Yes
EEG with generalized epileptiform discharges (HP:0011198)	Yes	Yes
EEG with photoparoxysmal response (HP:0010852)	Yes	Yes
Delayed developmental milestones (HP:0001263)	Yes	Yes
Developmental regression (HP:0002376)	Yes	No
Language impairment (HP:0002463)	Yes	Yes
Intellectual disability, moderate (HP:0002342)	Yes	Yes
Novel clinical features		
Autism spectrum disorder (HP:0000729)	No	Yes
Trichotillomania (HP:0012167)	Yes	No
Severe temper tantrums (HP:0025162)	Yes	No
Echolalia (HP:0010529)	No	Yes
Trouble sleeping (HP:0002360)	Yes	Yes
Pica (HP:011856)	Yes	Yes
Hypotonia (HP:0001290)	Yes	Yes
Ataxic gait (HP:0002066)	Yes	Yes
Hearing loss (HP:0000365)	Yes[a]	Yes[a]

These features were absent from both parents. Human Phenotype Ontology (HPO) terms are listed. EEG, electroencephalography.
[a]Unrelated to their inherited genetic condition.

Two sisters with developmental and epileptic encephalopathy (Scheffer et al., 2017) were referred for genomic testing. The 7-year-old proband was the first child to non-consanguineous parents, an English father and Australian mother, both of Ashkenazi Jewish origin. There was no family history of epilepsy or intellectual disability. Following an unremarkable perinatal history, her vocalization regressed at age 6 months and her subsequent development was delayed, walking at 22 months and speaking single words at 2.5 years. At age 7 years, she had a total of 100 single words, but lacked word combinations. She was assessed as having an intellectual disability, behavioral problems with tantrums, and trichotillomania. She also had some features consistent with autism spectrum disorder and obsessive behaviors.

Absence seizures with eyelid myoclonus began at 8 months. She then developed drop attacks secondary to myoclonic-atonic seizures. The seizures remained refractory to treatment with up to 50 absence seizures and eyelid myoclonus occurring per hour despite multiple antiepileptic medications. EEG studies showed frequent 3 Hz generalized spike-wave activity and prominent photosensitivity.

Her sister was 2 years younger and presented with seizures at age 12 months; she followed a similar clinical, but milder, course (Table 1). Her motor milestones were normal, but speech acquisition was delayed with single words at 22 months. At 5 years, she could combine words.

Both girls had hearing loss due to chronic ear infections. Their vision was normal. Overall, the sisters had an epilepsy phenotype that had features of two well-established epilepsy syndromes: epilepsy with myoclonic-atonic seizures and epilepsy with eyelid myoclonus.

An extensive diagnostic workup did not identify an etiology. A SNP microarray detected a de novo 700 Kb duplication at chrXq27.1, including *SOX3*, which was not thought to be contributory, as well as two small regions of homozygosity on chromosomes 1 and 9. Mitochondrial sequencing for 22 mtDNA and three common *POLG* variants (Uusimaa et al., 2013) was normal. Methylation studies for Angelman Syndrome were normal. Biochemical studies including serum lactate, pyruvate, and CSF amino acids, glucose, neurotransmitters, and methyltetrahydrofolate were normal. Lysosomal enzymes, carnitine studies, serum vitamin D levels, selenium, red cell folate, active B12, transferrin isoforms, and iron levels were normal. A brain MRI performed on the proband was normal.

High throughput targeted sequencing of 65 epilepsy genes using molecular inversion probes (MIPs; Carvill et al., 2013), was performed on DNA from the family quartet (both girls and their parents) in 2014. Note that 100 bp paired end reads were aligned to a custom hg19 reference genome containing only chromosomes 1–22, X, Y, chrM, using bwa sampe (v0.5.9-r16; Li & Durbin, 2009), standard settings, and variant analysis and filtration as described (Carvill et al., 2013). No pathogenic variants were identified.

Whole exome sequencing (WES) was performed to ~50× depth on the quartet at the Australian Genome Research Facility in 2015. Reads were aligned to hg19 using bwa mem (v0.7.10; Li, 2013), with variant analysis and filtration as described (Y.-C. Liu et al., 2016). After in silico filtering for genes matching the patient phenotype, as well as for variants segregating in the affected individuals, no plausible candidate genes were identified.

WGS was performed to 28–40× depth using Illumina HiSeq X at the Kinghorn Centre for Clinical Genomics, Australia, on the quartet, in 2016. Genomic data were processed according to the GATK best practices guidelines (Van der Auwera et al., 2013), using GATK (v3.3; (DePristo et al., 2011; McKenna et al., 2010), as previously described (Mallawaarachchi et al., 2016). Reads were aligned to the b37d5 (1000 genomes + decoy) reference genome using bwa mem (v0.7.10; (Li, 2013)) followed by indel realignment and base quality recalibration. Single nucleotide variants and short insertions and deletions were joint-called using HaplotypeCaller in gVCF mode, with variant quality score recalibration. Variants were annotated using VEP (v79; McLaren et al., 2016), dbNSFP (v2.0; X. Liu, Jian, & Boerwinkle, 2013) and CADD (v1.0; Kircher et al., 2014), converted to a Gemini database (v0.11.0; Paila, Chapman, Kirchner, & Quinlan, 2013) and filtered using *Seave* (Gayevskiy, Roscioli, Dinger, & Cowley, 2018). The same heterozygous de novo frameshift variant in exon 5 of *SYNGAP1* was identified in both siblings using a shared de novo pattern of inheritance, which was consistent with gonadal mosaicism in one parent (Supporting Information Table S1). There were no additional candidate variants in relevant genes that segregated with the disease for an autosomal recessive, autosomal dominant or shared de novo model.

A heterozygous 13 bp duplication in *SYNGAP1* (NM_006772.2:c.435_447dup, NP_006763.2:p.(Leu150LysfsTer6), Supporting

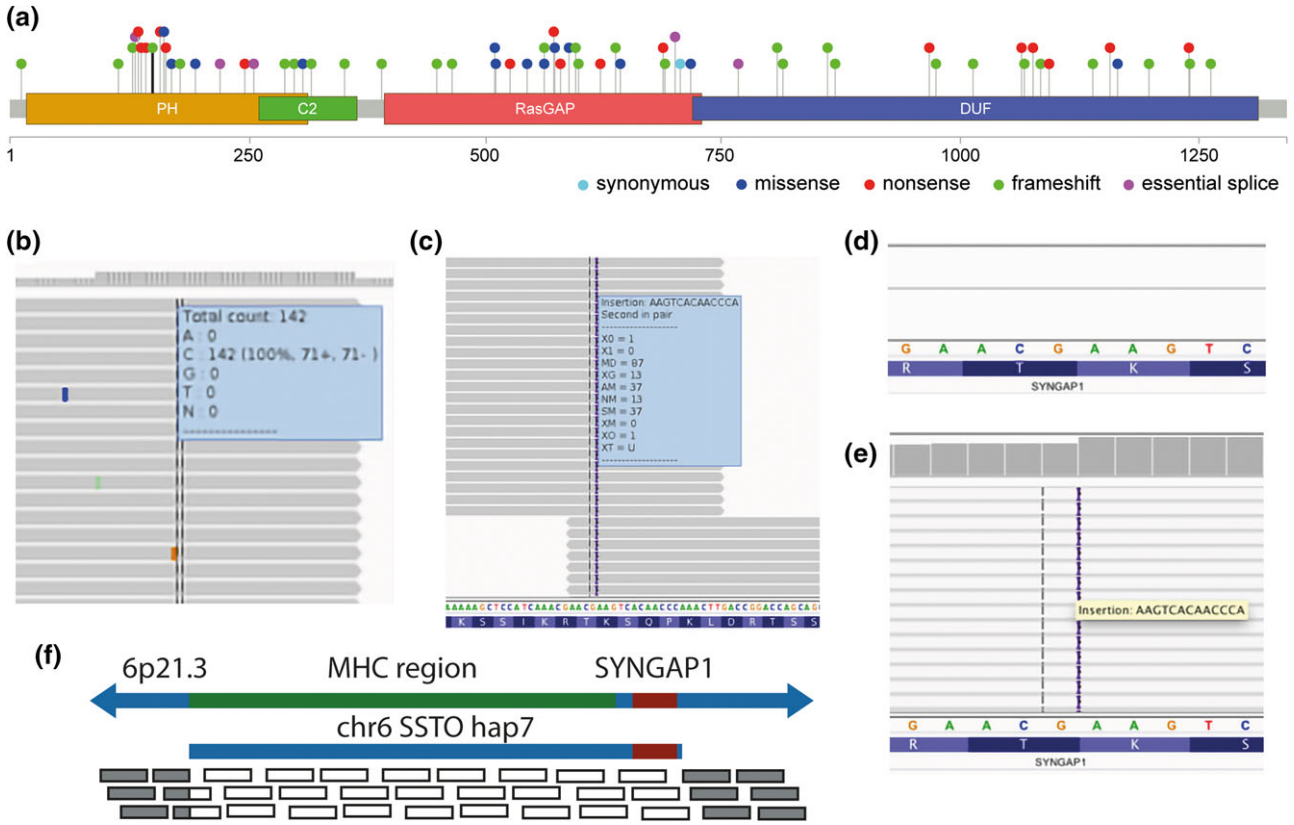

FIGURE 1 (a) Using whole genome sequencing (WGS), we identified a pathogenic SYNGAP1 p.L150Vfs*6 variant (black stalk within PH domain). All 65 pathogenic or likely pathogenic SYNGAP1 variants reported in ClinVar (accessed 1st April, 2018) are colored by mutation type. Figure created using lollipops, with domains from InterPro. PH: pleckstrin homology domain; C2: C2 domain; RasGAP: rho GTPase activation protein domain; DUF: domain of unknown function (DUF3498). (b) Standard analysis of targeted molecular inversion probes (MIPs) sequencing of epilepsy genes including SYNGAP1 identified 142 reads covering the same region, none of which carried the mutation. (c) After extending the maximum number of gap extensions (−e 20) during alignment, reads carrying the 13 bp duplication aligned correctly to SYNGAP1 exon 5. (d) Whole exome sequencing (WES) revealed zero reads covering SYNGAP1 with mapping quality >20. (e) Removing the mapping quality filter revealed an average of 91 reads, including 50 reads carrying the 13 bp duplication. (f) SYNGAP1 is distal to the major histocompatibility complex (MHC) on chr6p21.3. One of the MHC contigs included in the hg19 reference genome, chr6_ssto_hap7 includes SYNGAP1. The mapping quality score for reads that align to both chr6, and chr6_ssto_hap7 were penalized to zero (translucent reads) whereas reads that map only to chr6 have high mapping quality (grey reads)

Information Table S1) was identified by WGS in both girls. This variant was absent in DNA extracted from the peripheral blood from both parents, consistent with gonadal mosaicism as the most likely explanation. The variant was absent from ExAC (Lek et al., 2015), GnomAD, and 1000 genomes (1000 Genomes Project Consortium et al., 2015). While this variant has not been reported in the literature, ClinVar or OMIM, it is within the N-terminal Pleckstrin homology domain, close to many previously reported pathogenic loss of function variants (Figure 1a). *SYNGAP1* has a pLI score (Samocha et al., 2014) of 1.0, indicating it is intolerant of loss of function variation. The variant was validated by Sanger sequencing in a clinically accredited laboratory and reported as an ACMG class V pathogenic variant (Supporting Information Table S1).

Synaptic GTPase–activating protein 1 (*SYNGAP1*) encodes a brain-specific synaptic Ras GTPase activating protein that suppresses signaling pathways linked to NMDA receptor (NMDAR)-mediated synaptic plasticity and AMPA receptor (AMPAR) membrane insertion (McKusick, 2007). De novo truncating variants in *SYNGAP1* leading to haploinsufficiency were first identified in individuals with moderate-to-severe intellectual disability (Hamdan et al., 2009), and we identified it as a cause of developmental and epileptic encephalopathy (Carvill et al., 2013). De novo mutations in *SYNGAP1* are a relatively frequent cause of developmental delay (Deciphering Developmental Disorders Study et al., 2017). Forty of the 64 pathogenic or likely pathogenic *SYNGAP1* variants currently reported in ClinVar are associated with the OMIM disorder Mental Retardation, Autosomal Dominant 5 (MIM 612621).

WGS differed from previous genetic testing in several ways, including no potential for capture or design bias, longer read length (2 × 150 bp), the reference genome, and versions of alignment and variant calling software. After confirming that both the targeted MIPs panel and the WES targeted this region of *SYNGAP1*, we investigated which factors led to the variant being missed by previous testing.

In the MIPs targeted sequencing data, good gene coverage of *SYNGAP1* exon 5 was observed, with a read depth of >142× in the older sister (Figure 1b). Due to the use of an older read aligner (i.e., bwa sampe v0.5.9-r16), and the high ratio of 13 bp mutation to 100 bp read length, we hypothesized that the reads supporting the

duplication may have been misaligned. The reads were re-aligned to the reference genome, increasing the maximum number of gap extensions (–e 20), which resulted in the reads that carried the variant being appropriately aligned from two overlapping amplicons (Figure 1c). Updating the aligner to bwa mem v0.7.15 also correctly aligned the reads carrying the mutation using default settings (not shown).

In the WES dataset, we observed no reads across *SYNGAP1* with mapping quality >20 (Figure 1d). Removing the mapping quality filter revealed an average read depth of 91× across exon 5 (Figure 1e), consistent with read alignment to multiple regions of the reference genome. Importantly this also identified that there were reads carrying the heterozygous pathogenic variant.

The hg19 version of the reference genome from UCSC contains seven additional sequences at the 6p13 locus, to capture the extensive genetic variation in the major histocompatibility complex (MHC) (Lam, Tay, Wang, Xiao, & Ren, 2015). *SYNGAP1* is found centromeric to one of the common HLA haplotypes, A1-B8-DR3-DQ2 (Horton et al., 2008), represented by the chr6_ssto_hap7 contig (Figure 1F). Consequently, the sequencing reads from *SYNGAP1* mapped perfectly to both chr6p21.3 and chr6_ssto_hap7, and their mapping quality scores were set to zero (Figure 1f), making these reads invisible to variant identification tools. The pathogenic variant was identified using default detection settings once the reads were re-aligned to the GRCh37 reference genome that lacks the additional MHC contigs, or to GRCh38 with an 'alt-aware' read aligner, bwa mem (v0.7.12), which assigned the correct mapping quality scores to the reads.

In summary, we identified a shared heterozygous *SYNGAP1* p.(L150Vfs*6) variant (ACMG class V, pathogenic) in two siblings with developmental and epileptic encephalopathy through an improved WGS bioinformatics pipeline, and consideration of multiple modes of inheritance. Despite having greater than 40 high quality reads supporting this variant, it was missed by both high-depth targeted MIPs sequencing and WES, due to two different technical reasons, principally based on read alignment issues. Even as many groups converge upon similar BWA-GATK best practice pipelines (Van der Auwera et al., 2013), there are still many variables including choice of reference genome, base and variant quality recalibration settings, variant annotation and filtration tools, and versions of software that influence variant detection. The differences between the reference genome versions have been recently summarized (Li, 2017).

This case report focused only on *SYNGAP1*, therefore, we investigated which additional genes may be affected by the same mapping issues. There are 245 genes represented by at least one overlapping MHC contig, including 31 genes associated with human diseases (Supporting Information Table S2). Among these, *COL11A2* is associated with Autosomal Dominant Stickler Syndrome (MIM 604841), Marshall Syndrome (MIM 154780), and Autosomal recessive fibrochondrogenesis (MIM 228520). Neuraminidase 1 (*NEU1*) is associated with autosomal recessive sialidosis (MIM 256550), a lysosomal storage disease affecting the nervous system. These, and others reviewed in Supporting Information Table S2 are similarly affected by the read mapping issues reported here and so should have appropriate coverage analysis performed. We note that the duplication of sequences in the MHC region, or other 'alt' contigs are distinct from other truly duplicated regions of the genome, including highly homologous pseudogenes, or the pseudoautosomal regions. Innovative approaches are beginning to resolve some regions of the genome previously classified as inaccessible to short-read sequencing, e.g., *SMN1* and *SMN2* (Feng et al., 2017).

Updating analysis pipelines has been shown to increase the diagnostic yield on systematic retrospective re-analyses (Wright et al., 2018). In a recent multi-laboratory study of challenging variants, bioinformatic errors were a major cause of considerable inter-laboratory discordance, even among clinical laboratories (Lincoln et al.,). Our results suggest that updating the reference genome and aligner versions should be considered in any retrospective re-analyses of undiagnosed patient genome data. Additionally, alignment-free (Ostrander et al., 2018), or deep-learning based variant calling methods (Poplin et al., 2018) may be considered as maximally orthogonal approaches for re-analyzing data. Initiatives such as the Broad Institute's "Functional Equivalence" specification. PrecisionFDA (Petrone, 2016), Genome in a Bottle (Zook, Catoe et al., 2016; Zook, Chapman et al., 2014) and the DREAM challenges (Boutros et al., 2014; Zook, Catoe et al., 2016; Zook, Chapman et al., 2014) provide objective feedback as to the performance of bioinformatic pipelines and help labs know if their pipelines may be underperforming, and warrant updating.

This case report suggests that lack of diagnosis with different genomic technologies may occur due to technical limitations and that clinical genomic re-analysis including all potential inheritance patterns and the use of orthogonal and updated bioinformatic pipelines may identify previously undetected pathogenic variants. Comprehensive assessment of read coverage across all disease-relevant genes should be performed in parallel with variant detection pipelines to highlight poorly covered genes with potential pathogenic variation.

ACKNOWLEDGMENTS

The authors thank the family for their participation in this study. The authors thank the Kinghorn Centre for Clinical Genomics for assistance with production and processing of whole genome sequencing data. The authors thank Marie-Jo Brion and Bronwyn Terrill for helpful suggestions for this manuscript.

AUTHOR CONTRIBUTIONS

M.J.C., Y.C.L., K.L.O., and M.B. performed clinical bioinformatics; C.T.M. was associated with library preparation; M.J.C., Y.C.L., K.L.O., G.L.C., C.T.M., V.G., Y.Z., M.B., and T.R. performed genomic analysis and variant analysis; M.D., D.R.M.V., and I.E.S. were associated with phenotyping; M.F.B. performed genomic pathology; M.J.C., I.E.S., M.E.D., and T.R. wrote the manuscript and led the project.

ETHICAL COMPLIANCE

Genetic studies were approved by local ethics committees and written informed consent was obtained for molecular genetic analysis.

ORCID

Mark J Cowley https://orcid.org/0000-0002-9519-5714
Karen L. Oliver https://orcid.org/0000-0001-5188-6153

Gemma Carvill https://orcid.org/0000-0003-4945-3628
Michael F. Buckley https://orcid.org/0000-0002-8298-8758
Melanie Bahlo https://orcid.org/0000-0001-5132-0774
Ingrid E. Scheffer https://orcid.org/0000-0002-2311-2174
Marcel E. Dinger https://orcid.org/0000-0003-4423-934X

REFERENCES

Boutros, P. C., Ewing, A. D., Ellrott, K., Norman, T. C., Dang, K. K., Hu, Y., … Stuart, J. M. (2014). Global optimization of somatic variant identification in cancer genomes with a global community challenge. *Nature Genetics*, *46*(4), 318–319. https://doi.org/10.1038/ng.2932

Carvill, G. L., Heavin, S. B., Yendle, S. C., McMahon, J. M., O'Roak, B. J., Cook, J., … Mefford, H. C. (2013). Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nature Genetics*, *45*(7), 825–830. https://doi.org/10.1038/ng.2646

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Deciphering Developmental Disorders Study, McRae, J. F., Clayton, S., Fitzgerald, T. W., Kaplanis, J., Prigmore, E., … Hurles, M. E. (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, *542*, 433. https://doi.org/10.1038/nature21062

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Feng, Y., Ge, X., Meng, L., Scull, J., Li, J., Tian, X., … Zhang, J. (2017). The next generation of population-based spinal muscular atrophy carrier screening: Comprehensive pan-ethnic SMN1 copy-number and sequence variant analysis by massively parallel sequencing. *Genetics in Medicine*, *19*(8), 936–944. https://doi.org/10.1038/gim.2016.215

Gayevskiy, V., Roscioli, T., Dinger, M. E., & Cowley, M. J. (2018). Seave: A comprehensive web platform for storing and interrogating human genomic variation. *Bioinformatics*, *35*(1), 122–125. https://doi.org/10.1093/bioinformatics/bty540

Hamdan, F. F., Gauthier, J., Spiegelman, D., Noreau, A., Yang, Y., Pellerin, S., … Michaud, J. L. (2009). Mutations in SYNGAP1in autosomal nonsyndromic mental retardation. *The New England Journal of Medicine*, *360*(6), 599–605. https://doi.org/10.1056/NEJMoa0805392

Horton, R., Gibson, R., Coggill, P., Miretti, M., Allcock, R. J., Almeida, J., … Beck, S. (2008). Variation analysis and gene annotation of eight MHC haplotypes: The MHC haplotype project. *Immunogenetics*, *60*(1), 1–18. https://doi.org/10.1007/s00251-007-0262-2

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*, 310–315. https://doi.org/10.1038/ng.2892

Lam, T. H., Tay, M. Z., Wang, B., Xiao, Z., & Ren, E. C. (2015). Intra-haplotypic Variants Differentiate Complex Linkage Disequilibrium within Human MHC Haplotypes. *Scientific Reports*, *5*, 16972. https://doi.org/10.1038/srep16972

Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Lek, M., … MacArthur, D. G. (2015). Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*. https://doi.org/10.1101/030338

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org, q-bio.GN*. https://doi.org/10.6084/m9.figshare.963153

Li, H. (2017). Which human reference genome to use? Retrieved from http://lh3.github.io/2017/11/13/which-human-reference-genome-to-use

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lincoln, S. E., Zook, J. M., Chowdhury, S., Mahamdallie, S., Fellowes, A., Klee, E. W., … Shirts, B. H. (2018). An interlaboratory study of complex variant detection. *bioRxiv* 218529. Retrieved from https://www.biorxiv.org/content/biorxiv/early/2017/11/23/218529.full.pdf

Liu, X., Jian, X., & Boerwinkle, E. (2013). dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. *Human Mutation*, *34*(9), E2393–E2402. https://doi.org/10.1002/humu.22376

Liu, Y.-C., Lee, J. W. A., Bellows, S. T., Damiano, J. A., Mullen, S. A., Berkovic, S. F., … Group, C. (2016). Evaluation of non-coding variation in GLUT1 deficiency. *Developmental Medicine & Child Neurology*, *58*(12), 1295–1302. https://doi.org/10.1016/j.ymgme.2012.01.011

Mallawaarachchi, A. C., Hort, Y., Cowley, M., McCabe, M. J., Minoche, A., Dinger, M. E., … Furlong, T. J. (2016). Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. *European Journal of Human Genetics*, *24*(11), 1584–1590. https://doi.org/10.1038/ejhg.2016.48

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKusick, V. A. (2007). Mendelian inheritance in man and its online version, OMIM. *American Journal of Human Genetics*, *80*(4), 588–604. https://doi.org/10.1086/514346

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl variant effect predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Ostrander, B. E. P., Butterfield, R. J., Pedersen, B. S., Farrell, A. J., Layer, R. M., Ward, A., … Quinlan, A. R. (2018). Whole-genome analysis for effective clinical diagnosis and gene discovery in early infantile epileptic encephalopathy. *NPJ Genomic Medicine*, *3*, 22. https://doi.org/10.1038/s41525-018-0061-8

Paila, U., Chapman, B. A., Kirchner, R., & Quinlan, A. R. (2013). GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS Computational Biology*, *9*(7), e1003153. https://doi.org/10.1371/journal.pcbi.1003153

Petrone, J. (2016). FDA wades into sequencing-based diagnostics regulation. *Nature Biotechnology*, *34*(7), 681–682. https://doi.org/10.1038/nbt0214-111

Poplin, R., Chang, P. C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., … DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., … Daly, M. J. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics*, *46*(9), 944–950. https://doi.org/10.1038/ng.3050

Scheffer, I. E., Berkovic, S., Capovilla, G., Connolly, M. B., French, J., Guilhoto, L., … Zuberi, S. M. (2017). ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia*, *58*(4), 512–521. https://doi.org/10.1111/epi.13709

Uusimaa, J., Gowda, V., McShane, A., Smith, C., Evans, J., Shrier, A., … Poulton, J. (2013). Prospective study of POLG mutations presenting in

children with intractable epilepsy: Prevalence and clinical features. *Epilepsia*, *54*(6), 1002–1011. https://doi.org/10.1002/humu.20824

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 11.10.1–11.10.33

Wright, C. F., McRae, J. F., Clayton, S., Gallone, G., Aitken, S., FitzGerald, T. W., … Study, D. D. D. (2018). Making new genetic diagnoses with old data: Iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genetics in Medicine*, *20*(10), 1216–1223. https://doi.org/10.1038/gim.2017.246

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*, 160025. https://doi.org/10.1038/sdata.2016.25

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, *32*(3), 246–251. https://doi.org/10.1038/nbt.2835

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Cowley MJ, Liu Y-C, Oliver KL, et al. Reanalysis and optimisation of bioinformatic pipelines is critical for mutation detection. *Human Mutation*. 2019;40:374–379. https://doi.org/10.1002/humu.23699