Systematic reanalysis of genomic data improves quality of variant interpretation

Susan M. Hiatt[1],[a], **Michelle D. Amaral**[1],[a], **Kevin M. Bowling**[1], **Candice R. Finnila**[1], **Michelle L. Thompson**[1], **David E. Gray**[1], **James M.J. Lawlor**[1], **J. Nicholas Cochran**[1], **E. Martina Bebin**[2], **Kyle B. Brothers**[3], **Kelly M. East**[1], **Whitley V. Kelley**[1], **Neil E. Lamb**[1], **Shawn E. Levy**[1], **Edward J. Lose**[2], **Matthew B. Neu**[1], **Carla A. Rich**[3], **Shirley Simmons**[2], **Richard M. Myers**[1], **Gregory S. Barsh**[1], and **Gregory M. Cooper**[1],[*]

[1]HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35806, USA

[2]University of Alabama at Birmingham, Birmingham, AL, USA

[3]University of Louisville, Louisville, KY, USA

Abstract

As genomic sequencing expands, so does our knowledge of the link between genetic variation and disease. Deeper catalogs of variant frequencies improve identification of benign variants, while sequencing affected individuals reveals disease-associated variation. Accumulation of human genetic data thus makes reanalysis a means to maximize benefits of clinical sequencing. We implemented pipelines to systematically reassess sequencing data from 494 individuals with developmental disability. Reanalysis yielded pathogenic or likely pathogenic (P/LP) variants that were not initially reported in 23 individuals, 6 described here, comprising a 16% increase in P/LP yield. We also downgraded three LP and six variants of uncertain significance (VUS) due to updated population frequency data. The likelihood of identifying a new P/LP variant increased over time, as ~22% of individuals who did not receive a P/LP variant at their original analysis subsequently did after three years. We show here that reanalysis and data sharing increase the diagnostic yield and accuracy of clinical sequencing.

Graphical abstract

[*]Corresponding Author: gcooper@hudsonalpha.org.
[a]equal contributions
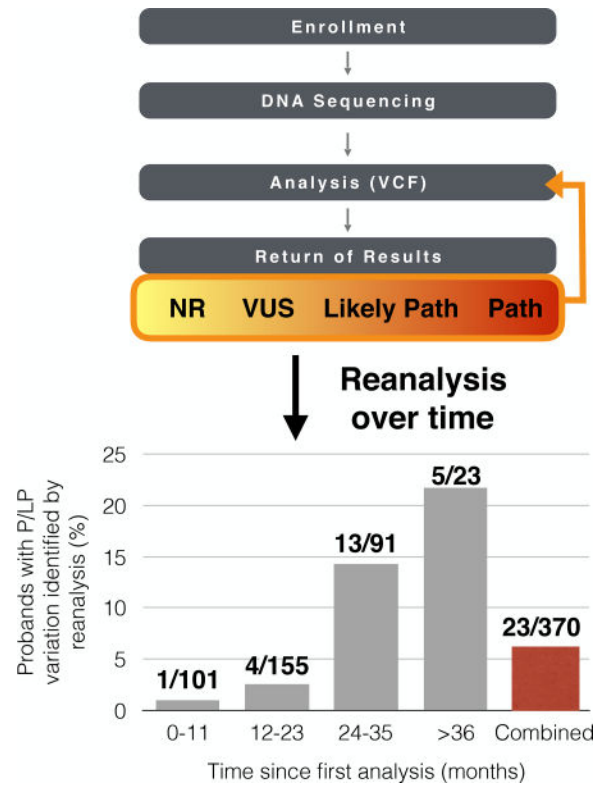
INTRODUCTION

Whole genome sequencing (WGS) and whole exome sequencing (WES) are increasingly used clinically, particularly for rare disease diagnosis. WGS/WES uncovers pathogenic or likely pathogenic (P/LP) variants in 20–60% of sequenced patients[1–4], leaving many patients without a relevant genetic finding. Although these individuals may have diseases that are non-genetic or result from complex genetic effects, incomplete knowledge of genetic variation likely prevents identification of many P/LP variants. Systematically reanalyzing data over time may prove useful, as accumulated knowledge—i.e., new publications, updated population frequencies, improved clinical variant databases, and data sharing among researchers—may facilitate new discoveries[5–7].

We sought to systematically reanalyze WES/WGS data from probands with developmental delay and/or intellectual disability (DD/ID) enrolled in the Clinical Sequencing Exploratory Research (CSER) project at HudsonAlpha[8]. An initial reanalysis effort was described in an analysis of the first 371 affected probands[8], but an expanded and improved reanalysis pipeline, including new findings, approaches, and implications, is presented here.

MATERIALS AND METHODS

Study overview

Enrollment, sequencing, variant calling and Sanger confirmation were performed as previously described[8], although we have now included an additional 123 affected probands, increasing the cohort to 494 affected individuals (see Supplemental Table 1 and Supplemental Materials and Methods). For reanalysis, original joint-called VCFs were reannotated to include updated versions of ClinVar[5], ExAC/gnomAD[6], CADD[9], and DDG2P[10]. Filtering and curation were performed in light of new data. Reanalysis is performed on a rolling schedule so that all cases are reviewed at least once every 12 months. Genes of uncertain disease significance that harbored candidate variants were submitted to GeneMatcher[11]. Our study began before publication of the American College of Medical Genetics and Genomics (ACMG)[12] standards, but our original classification criteria were conceptually similar and are available on the ClinVar Submitters page (https://submit.ncbi.nlm.nih.gov/ft/byid/yR2NSzwW/HA_assertions_20161101.pdf). For reanalysis, however, ACMG criteria[12] were used and evidence codes for reinterpreted variants are provided. All returned variants were submitted to ClinVar[5]. Sequence data for consenting participants is available through dbGAP[13]. Supplemental Materials and Methods provide additional details.

Uniparental disomy (UPD) analysis

UPD was called using UPDio[14]. CNV calls (Supplemental Materials and Methods), were masked and not considered by UPDio.

SLC1A4 Analysis

DNA and RNA isolation, cDNA synthesis, PCR, and qPCR were conducted with standard protocols (Supplemental Materials and Methods). Graphpad Prism version 7.0c was used for graphing and statistics.

RESULTS

New Findings

Based on the success of our initial reanalysis efforts[8], we sought to improve and expand our strategy. We have subsequently identified P/LP variation in six additional probands (Table 1). Two of these "upgrades" resulted from recent publications and two from GeneMatcher[11] collaborations (Table 1). One of these variants (FGF12, NM_004113.5:c.145G>A, (p.Arg114His)) is a recurrent *de novo* variant now seen in two unrelated probands in our cohort[8,15]. In each of these four cases, the variation was in a gene not previously associated with disease and was originally classified as a non-returnable VUS. After data supporting gene-disease associations became available, application of ACMG criteria led to classification of these variants as P/LP (Table 1) and return to families.

In one case, we identified a single nucleotide variant (SNV) and a copy number variant (CNV) within *SLC1A4*, associated with an autosomal recessive neurodevelopmental disorder (MIM:616657)[16]. Initially (June 2015), no returnable variation was found in this

proband. However, upon reanalysis, a paternally-inherited pathogenic missense variant[16] was found using updated ClinVar information. A targeted search for a second variant in *SLC1A4* revealed a maternally-inherited 782-bp deletion (Table 1). We confirmed the presence of this deletion in genomic DNA from the proband and mother (Supplemental Figures 1, 2). While this variant does not change *SLC1A4* transcript levels, it does result in skipping of exon 6 (Supplemental Figures 1, 3) and is predicted to lead to an in-frame deletion of 65 amino acids encompassing two transmembrane domains. This CNV was only identified in the proband by one of the four algorithms in our pipeline and was completely missed in the mother; curation of unfiltered CNVs coupled to manual inspection of reads in both samples was required. Had we not found the previously reported missense variant, or lacked phenotype data suggesting the relevance of *SLC1A4*, this CNV would have been missed.

In addition to manual reanalysis in light of new data, we also implemented methods to detect uniparental disomy (UPD). While UPD, especially heterodisomy, often goes unnoticed, it can cause DD/ID when imprinted regions are affected[17–19]. We found two cases of disease-associated UPD, both affecting chromosome 15 (Table 1). Clinical methylation analyses confirmed one result. Clinical methylation analyses were recommended for the second but not performed.

Downgrades

In addition to searching for new P/LP variation, we also reanalyzed all previously returned variants, leading to downgrades of nine variants in seven individuals (Table 2). Six of these variants were originally classified as VUSs and three were considered LP. All downgrades resulted from addition of data in the ExAC/gnomAD databases[6]. Variants present at non-trivial frequencies in these databases are unlikely to be dominant, highly penetrant DD/ID variants[6]. Similarly, variation that is homozygous or hemizygous at non-trivial frequencies are unlikely to cause recessive or X-linked disease. Though reanalysis did uncover new P/LP variation in three of these seven probands, upgrade and downgrade decisions were made independently of one another.

Likelihood of Variant Upgrade over Time

We conducted WES on the first 127 probands of this study and switched to WGS for all subsequent probands (Supplemental Table 1). We measured P/LP rates separately for WES and WGS, both before and after reanalysis (Supplemental Table 2). Although the initial P/LP rate for WGS (26.2%) was greater than that of WES (22.0%), reanalysis yielded an 11.0% increase in the P/LP rate for WES and only 1.6% increase for WGS. To further refine this comparison, since all WES cases were trios, we also restricted calculations for WGS to trios, and found the P/LP rate in WGS-trios increased by 2.2% (28.1% to 30.3%). The larger relative gain in WES yield is likely due to time since initial analysis. WES was initially performed from November 2013 to May 2015, while WGS was performed from June 2015 onward. Indeed, while probands had only a 1% likelihood of upgrade in the first year following analysis, this rate increased over time. ~22% of cases over three years old with a VUS or no returnables were eventually found to have P/LP variation (Table 3, Supplemental Materials and Methods).

DISCUSSION

Accumulation of genetic knowledge suggests that reanalysis of sequencing data may lead to discovery of novel medically relevant variants and to refinement of initial variant interpretations. Our reanalysis efforts thus far have led to identification of 22 P/LP variants in 23 probands in a cohort of 494 total patients. These affect ~6% of probands who originally received either a VUS or no returnables, and represents a 16% increase in total P/LP yield. Other groups have also reported success with reanalysis, with upgrade rates from 10–36%[7,20].

The ACMG has published guidelines for the interpretation of sequence variants[12], and it is important to note that the changed interpretations in our study did not result from altered application of criteria, but rather new evidence supporting (or refuting) pathogenicity.

Three of the downgrades resulted in a change from LP to likely benign. The ACMG suggests that an LP designation represents pathogenicity with 90% confidence[12]. In our case, these three downgrades represent 9% of all variants initially determined to be LP, suggesting an empirical error rate similar to the conceptual target established by the ACMG. As population variant frequency databases expand, reanalysis of previously reported variants will continue to be necessary.

Based on our experience with reanalysis, we suggest the following framework:

1. If reanalysis has never been conducted on WES/WGS data, it should be performed, even in the absence of annotation or pipeline updates. New P/LP variation may be discovered simply by reviewing new literature during manual curation.

2. WES data lacking P/LP variation should be reanalyzed before performing WGS, especially for data over two years old.

3. If reanalysis of all data cannot be conducted regularly, an automated process to flag variants in genes recently linked to disease can be more easily implemented.

4. Improvements to bioinformatics pipelines are beneficial, especially updates from population and clinical genetic databases. Additionally, algorithms for detection of non-SNV/indel variants, such as CNVs, UPD, etc., are continually improving and worth updating. While any individual variant type may be a small fraction of P/LP variation, such additions can make a large cumulative difference.

5. Data sharing through GeneMatcher, ClinVar, and related resources is a key component of reanalysis. These platforms help to establish gene-disease relationships among research groups with small cohorts, in many cases well before formal publication (which can take months or years from the time that a robust disease association has been established). Most of our GeneMatcher submissions (45 of 52 probands) represented one gene of interest within one individual, with *de novo* variants present in ~50% of these. We expect the benefits of data sharing to increase, as 12% (43/350) of the individuals in our cohort that lack a P/LP variant harbor a variant within a gene that has been

submitted to GeneMatcher; 79% (34/43) of these genes have at least preliminary matches.

6. Time since initial analysis should inform the reanalysis strategy. If WGS/WES data was analyzed within the past year, reanalysis yield will be small – in our case, only 1%. However, datasets first analyzed over two years ago should be prioritized for reanalysis, consistent with observations from others[7,20]. We also note that timing of analysis is a major factor to consider when evaluating overall yield rates, particularly when focused on comparisons of technologies that have changed over time. For example, while we believe a number of factors support the benefits of WGS over WES for rare disease diagnosis (e.g., improved CNV detection, more uniform depth of coverage, etc.), the tendency for WGS-based analyses to be more recent than WES-based analyses must be accounted for when comparing P/LP variant yields.

In summary, our data strongly support the benefits of systematic reanalysis of WES/WGS data. Such efforts lead to substantial increases in P/LP variant discoveries while simultaneously reducing false positive P/LP rates. Thus, reanalysis can substantially improve the accuracy and benefits of clinical sequencing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Ankala A, da Silva C, Gualandi F, et al. A comprehensive genomic approach for neuromuscular diseases gives a high diagnostic yield. Ann Neurol. 2015; 77(2):206–214. [PubMed: 25380242]

2. Yang Y, Muzny DM, Xia F, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA. 2014; 312(18):1870–1879. [PubMed: 25326635]

3. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015; 47(7):717–726. [PubMed: 25985138]

4. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015; 97(2):199–215. [PubMed: 26166479]

5. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016; 44(D1):D862–868. [PubMed: 26582918]

6. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–291. [PubMed: 27535533]

7. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med. 2017; 19(2):209–214. [PubMed: 27441994]

8. Bowling KM, Thompson ML, Amaral MD, et al. Genomic diagnosis for children with intellectual disability and/or developmental delay. Genome Med. 2017; 9(1):43. [PubMed: 28554332]

9. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46(3):310–315. [PubMed: 24487276]

10. Wellcome Trust Sanger Institute. [Accessed November 18, 2014] The Development Disorder Genotype - Phenotype Database (DDG2P). Dec 3, 2015 https://decipher.sanger.ac.uk/ddd - ddgenes

11. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat. 2015; 36(10):928–930. [PubMed: 26220891]

12. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17(5):405–424. [PubMed: 25741868]

13. Tryka KA, Hao L, Sturcke A, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014; 42(Database issue):D975–979. [PubMed: 24297256]

14. King DA, Fitzgerald TW, Miller R, et al. A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders. Genome Res. 2014; 24(4):673–687. [PubMed: 24356988]

15. Guella I, Huh L, McKenzie MB, et al. De novo FGF12 mutation in 2 patients with neonatal-onset epilepsy. Neurol Genet. 2016; 2(6):e120. [PubMed: 27872899]

16. Srour M, Hamdan FF, Gan-Or Z, et al. A homozygous mutation in SLC1A4 in siblings with severe intellectual disability and microcephaly. Clin Genet. 2015; 88(1):e1–4.

17. Driscoll, DJ., Miller, JL., Schwartz, S., Cassidy, SB. Prader-Willi Syndrome. In: Pagon, RA.Adam, MP.Ardinger, HH., et al., editors. GeneReviews(R). Seattle (WA): 1993.

18. Dagli, AI., Mueller, J., Williams, CA. Angelman Syndrome. In: Pagon, RA.Adam, MP.Ardinger, HH., et al., editors. GeneReviews(R). Seattle (WA): 1993.

19. Shuman, C., Beckwith, JB., Weksberg, R. Beckwith-Wiedemann Syndrome. In: Pagon, RA.Adam, MP.Ardinger, HH., et al., editors. GeneReviews(R). Seattle (WA): 1993.

20. Eldomery MK, Coban-Akdemir Z, Harel T, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. 2017; 9(1):26. [PubMed: 28327206]

21. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017; 542(7642):433–438. [PubMed: 28135719]

22. Hamdan FF, Myers CT, Cosette P, et al. High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies. Am J Hum Genet. 2017; 101(5):664–685. [PubMed: 29100083]

Table 1

Upgraded variants.

Individual ID	Gene	Variant(s)	Reason for Update	Original Score[†]	Updated Score (with ACMG Criteria)
00231-C	DHX30	NM_138615.2:c.2353C>T, (p.Arg785Cys)	Publication[20,21]	VUS	Pathogenic (PS2, PS3, PS4, PM2, PP3)
00144-C	SLC1A4	NM_003038.4:c.766G>A, (p.Glu256Lys); NC_000002.11:g.6524824-6524560del, (p.Ser345Arg, Ser346_Leu410del)	Publication[16]	VUS; VUS	Pathogenic (PS3, PS4, PM2, PP1, PP4); Likely Pathogenic (PM2, PM3, PM4, PP4)
00012-C	FGF12	NM_021032.4:c.341G>A, (p.Arg114His)[‡]	GeneMatcher[15]	VUS	Pathogenic (PS2, PS4, PM2, PP3)
00303-C	NTRK2	NM_006180.4:c.1301A>G, (p.Tyr434Cys)	GeneMatcher[22]	VUS	Pathogenic (PS2, PS4, PM2, PP3)
00130-C	NA	Paternal isodisomy of chr15	UPD	NA	Pathogenic[§]
00346-C	NA	Maternal heterodisomy of chr15	UPD	NA	Pathogenic[§]

ACMG, American College of Medical Genetics and Genomics; VUS, Variant of Uncertain Significance; NA, not applicable; UPD, uniparental disomy.

[†] Original scores for SNV/CNVs based on retroactive scoring using ACMG criteria for a gene of uncertain significance (GUS). None of the variants in this table were returned to probands at time of the first analysis, nor was ACMG scoring or criteria explicitly used at that time.

[‡] Note that this variant was previously reported with our initial study findings[8], although the variant was recently found in one additional unrelated proband in our study, represented here.

[§] ACMG scoring criteria do not apply to UPD.

Table 2

Downgraded variants.

Individual ID	Gene	Suspected Mode of Inheritance	Variant	Original Score[†]	Updated Score (with ACMG criteria)	Allele Frequency at time of Original Analysis		Allele Frequency at time of Reanalysis (gnomAD)
						1KG	EVS	
00026-C	ULK4	Rec	NM_017886.3:c.2584C>T, (p.Arg862Ter)	VUS	Likely Benign (BS2)	0.05%	38/11762 (0 hom), 0.3%	777/276908 (4 hom), 0.002806%
00026-C	ULK4	Rec	NM_017886.3:c.2887G>A, (p.Val963Met)	VUS	Likely Benign (BS2)	0.32%	45/12217 (0 hom), 0.4%	942/276788 (4 hom), 0.003403%
00074-C	NRXN1	Rec	NM_001135659.2:c.3485-11013T>G	VUS	Likely Benign (BS2)	0.37%	NR	85/30792 (2 hom), 0.002760%
00074-C	NRXN1	Rec	NM_001135659.2:c.11405C>T, (p.Pro469Ser)	VUS	Likely Benign (BS2)	0.09%	34/12262 (0 hom), 0.3%	1048/273312 (6 hom), 0.003834%
00130-C	GRIA3	XLR	NM_007325.4:c.580G>A, (p.Gly194Arg)	Likely Pathogenic	Likely Benign (BS2)	0.06%	1/10562 (0 hemi), 0.0095%	12/200235 (3 hemi), 0.0000599%
00053-C	GRIA3	XLR	NM_007325.4:c.466T>C, (p.Tyr156His)	Likely Pathogenic	Likely Benign (BS2)	NR	NR	9/199897 (3 hemi), 0.0000450%
00059-C	PTPN11	Dom	NM_002834.4:c.1174G>A, (p.Ala392Thr)	Likely Pathogenic	Likely Benign (PS2, BS2)	NR	NR	8/246234 (0 hom), 0.00000324%
00111-C	CSMD1	Dom	NM_033225.5:c.1642A>G, (p.Thr548Ala)	VUS	Likely Benign (PS2, BS2)	NR	NR	7/240916 (0 hom), 0.0000290%
00007-C	SCN1A	Dom	NM_001165963.1:c.4547 C>T, (p.Ser1516Leu)	VUS	Likely Benign (BS2)	NR	NR	5/245496 (0 hom), 0.0000203%

1KG, 1000 Genomes Project; EVS, Exome Variant Server; Rec, Recessive; XLR, X-linked Recessive; Dom, Dominant; VUS, Variant of Uncertain Significance; hom, homozygotes; hemi, hemizygotes; NR, Not Reported

[†] ACMG criteria and scoring were not explicitly used at the time of first analysis.

Table 3

Likelihood of variant upgrade over time.

Time since first analysis (months)	Number of DD/ID-affected individuals with P/LP variation identified by reanalysis (%)
36	5/23 (21.7%)
24–35	12/91 (13.2%)
12–23	4/155 (2.6%)
0–11	1/101 (1.0%)
Overall	22/370 (5.9%)