Genetics inMedicine

www.nature.com/gim

ARTICLE

# Methods and feasibility study for exome sequencing as a universal second-tier test in newborn screening

Nicole Ruiz-Schultz[1], David Sant[2], Stevie Norcross[1], Warunee Dansithong[1], Kim Hart[1], Bryce Asay[1], Jordan Little[2], Krystal Chung[2], Kelly F. Oakeson[1], Erin L. Young[1], Karen Eilbeck[2] and Andreas Rohrwasser[1]✉

**PURPOSE:** Newborn screening disorders increasingly require genetic variant analysis as part of second-tier or confirmatory testing. Sanger sequencing and gene-specific next-generation sequencing (NGS)-based tests, the current methods of choice, are costly and lack scalability when expanding to new conditions. We describe a scalable, exome sequencing–based NGS pipeline with a priori analysis restriction that can be universally applied to any NBS disorder.

**METHODS:** De-identified abnormal newborn screening specimens representing severe combined immune deficiency (SCID), cystic fibrosis (CF), VLCAD deficiency, metachromatic leukodystrophy (MLD), and in silico sequence read data sets were used to validate the pipeline. To support interpretation and clinical decision-making within the bioinformatics pipeline, variants from multiple databases were curated and validated.

**RESULTS:** *CFTR* variant panel analysis correctly identified all variants. Concordance compared with diagnostic testing results for targeted gene analysis was between 78.6% and 100%. Validation of the bioinformatics pipeline with in silico data sets revealed a 100% detection rate. Varying degrees of overlap were observed between ClinVar and other databases ranging from 3% to 65%. Data normalization revealed that 11% of variants across the databases required manual curation.

**CONCLUSION:** This pipeline allows for restriction of analysis to variants within a single gene or multiple genes, and can be readily expanded to full exome analysis if clinically indicated and parental consent is granted.

*Genetics in Medicine* (2021) 23:767–776; https://doi.org/10.1038/s41436-020-01058-w

## INTRODUCTION

Genomic sequencing has been eyed by the newborn screening community for many years as a means to validate and further understand biochemical and metabolic newborn screening (NBS) results.[1–3] In a 2015 study using genome sequencing (GS) of trios (proband plus parents) the method was shown to be a viable adjunct to traditional NBS. Results provided fewer false positives, were used to resolve inconclusive results, and could be deployed to detect a wider range of diseases than metabolic tests alone.[1] Sequence-based augmentation of the NBS workflow is of importance due to variable disease presentation, to aid interpretation of borderline results and for disorders that rely on variant analysis in second-tier screening and confirmatory diagnostics.[4–8] Variable presentation of clinical features is a key issue in the interpretation of NBS results. The range of variants across each gene may contribute differently to the phenotype, complicating traditional screening interpretation.[9–12] Resultant efforts across the globe have seen the use of next-generation sequencing (NGS) methodology to improve the clinical workflow to diagnose seriously ill neonates, to test the feasibility of deployment of exome panels for subsets of NBS testing, and to explore the use of sequence-based tests to disorders not amenable to biochemical diagnosis.[8–14] While Sanger sequencing and allele specific tests—most commonly used today—[13–16] are labor-intensive, time-intensive, and costly, neither method is scalable, and the process requires de novo method design and revalidation when expanding testing to additional variants or genes.[17]

We describe the development and validation of a universal second-tier sequencing-based testing method that can be expanded to additional disorders and gene sets. We show that this methodology is scalable and would not require extensive redesign and revalidation when expanded to additional disorders. The efforts comprised the establishment of both a laboratory framework and standardized variant curation and interpretation processes. We developed and validated a laboratory method using two 3.2-mm dried blood spot (DBS) punches. We show similar turnaround time and cost impact compared with Sanger and amplicon-based NGS tests and we show that in contrast to Sanger and amplicon sequencing this methodology is highly scalable. A key component of the genomics approach is the postsequencing analysis that enables us to provide competitive turnaround times. Within the bioinformatics pipeline, we have curated multiple variant resources to aid the clinical team in variant-impact interpretation. This curation work demonstrates the wide distribution of knowledge pertaining to disease-causing variants, and we provide a generic suite of tools that can be implemented in other disorder cases.

## MATERIALS AND METHODS

Detailed information for all methods is available in Supplementary materials and methods.

### Exome sequencing

Figure 1 summarizes the ES pipeline. Exome libraries were generated from DNA extracted from two 3.2-mm DBS punches using Illumina's Nextera DNA Flex Dried Blood Spot Extraction Protocol Guide and Nextera Flex for Enrichment kit. Libraries were sequenced on an Illumina NextSeq 550 Sequencing System as paired-end runs with 149 cycles per read (2 × 149)

[1]Utah Public Health Laboratory, Salt Lake City, UT, USA. [2]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA. ✉email: arohrwasser@utah.gov

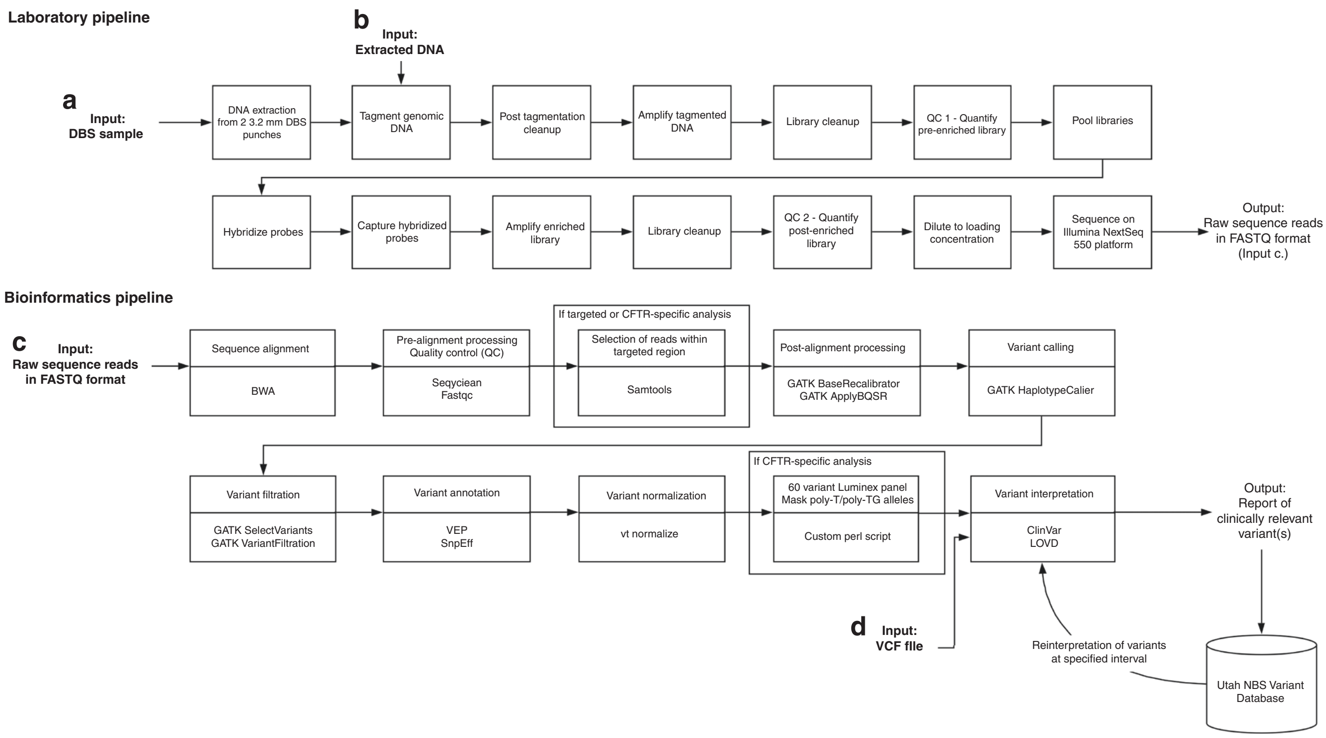
Laboratory pipeline
a
Input:
DBS sample
DNA extraction from 2 3.2 mm DBS punches
b
Input:
Extracted DNA
Tagment genomic DNA
Post tagmentation cleanup
Amplify tagmented DNA
Library cleanup
QC 1 - Quantify pre-enriched library
Pool libraries
Hybridize probes
Capture hybridized probes
Amplify enriched library
Library cleanup
QC 2 - Quantify post-enriched library
Dilute to loading concentration
Sequence on Illumina NextSeq 550 platform
Output:
Raw sequence reads in FASTQ format (Input c.)
Bioinformatics pipeline
c
Input:
Raw sequence reads in FASTQ format
Sequence alignment
BWA
Pre-alignment processing Quality control (QC)
Seqyciean Fastqc
If targeted or CFTR-specific analysis
Selection of reads within targeted region
Samtools
Post-alignment processing
GATK BaseRecalibrator GATK ApplyBQSR
Variant calling
GATK HaplotypeCalier
Variant filtration
GATK SelectVariants GATK VariantFiltration
Variant annotation
VEP SnpEff
Variant normalization
vt normalize
If CFTR-specific analysis
60 variant Luminex panel Mask poly-T/poly-TG alleles
Custom perl script
Variant interpretation
ClinVar LOVD
Output:
Report of clinically relevant variant(s)
d
Input:
VCF flle
Reinterpretation of variants at specified interval
Utah NBS Variant Database

**Fig. 1 Overview of the Utah Newborn Screening (NBS) Program exome sequencing and analysis pipeline.** The pipeline consists of two parts: a laboratory pipeline and a bioinformatics pipeline. The laboratory portion of the pipeline takes a dried blood spot (DBS) sample as input and uses two 3.2-mm DBS punches to generate an exome library. Exome library generation is performed using the Nextera Flex for Enrichment kit and is sequenced on the Illumina NextSeq 550 platform. FASTQ generation from BCL files is performed on the instrument. The bioinformatics pipeline is based on the GATK Best Practices pipeline. This pipeline can be run in full starting with (**a**) a DBS sample or can take input such as (**b**) extracted DNA and begin with exome library prep, (**c**) raw sequence reads in FASTQ format and begin with sequence analysis or (**d**) a VCF file and begin with variant interpretation.

and ten cycles per index read. A no-template control (NTC) and PhiX sample were included as a control and success metric.

## Variant database curation pipeline

Variants associated with genes and diseases are distributed across resources (ClinVar and smaller disease or gene-specific databases). We surveyed available curated data sources for variants associated with the disorder implicated genes and found ClinVar and several Leiden Open Variation Databases (LOVD) to contain relevant data. We did not include OMIM variants as they are subsumed by ClinVar, and while gnomAD provides frequency information for variants, it does not link variants with disease.[18–20] Figure 2 summarizes the variant database curation pipeline. Genomic variants from genes of interest were obtained from ClinVar and the LOVD sources. To obtain the variant annotations, a suite of tools was developed to extract the variant information from LOVD and ClinVar databases (https://github.com/eilbecklab/Utah-DOH-newborn-screening). Variants from each database were normalized using the biocommons hgvs python package and output in comma separated value (csv) format and imported into a MySQL database.[21] This pipeline can be run periodically to update variant information and additional databases can be included.

## Bioinformatics analysis pipeline

Custom bioinformatics pipelines for targeted, *CFTR*-specific, and exome analyses were used to analyze the sequencing data. These pipelines are based on the GATK Best Practices pipeline for germline variant discovery (Fig. 1).[22] All pipelines are contained within Snakemake workflow files (https://github.com/UtahNBS/WES-Secondary-Testing).[23]

## In silico validation and utilization of simulated read data sets

NEAT (NExt-generation sequencing Analysis Toolkit, version 2.0) was used to generate simulated paired-end 300 cycle reads representing exome data (https://github.com/UtahNBS/WES-Secondary-Testing).[24]

# RESULTS

The general validation design was defined by three broad case categories encompassing polygenic disorders, single-gene disorders, and an emerging disorder not yet included in the recommended uniform screening panel (RUSP).

1. Polygenic NBS disorders: severe combined immune deficiency (SCID). Three SCID samples were included for the targeted analysis of 39 genes.
2. Single-gene NBS disorders: cystic fibrosis (CF) and very long–chain acyl-CoA dehydrogenase (VLCAD) deficiency. In the case of CF, analysis can be limited to a set of common variants or include the entire coding sequence of CFTR. The Utah NBS Program currently uses the xTAG Luminex 60 variant assay for second-tier CF screening, which restricts the analysis to 60 common variants. Seven CF samples were subjected to analysis using an in silico panel containing the same 60 variants as well as analysis of the entire CFTR coding region. A VLCAD deficiency sample was included for targeted analysis of the *ACADVL* gene.
3. Emerging NBS disorders: Metachromatic leukodystrophy (MLD) is not included on any NBS panels in the United States; however, with emerging treatment opportunities, a screening assay has been developed in parallel.[25] While the actual study results are presented elsewhere, the application of this approach targeting screen-positive MLD cases illustrates the utility of this technology to emerging disorders and the opportunity of stepwise inclusion of additional loci based on clinical utility and investigator initiated requests. The Utah NBS Program collaborated with the University of Washington performing genotype analysis for biochemical screen-positive specimens. Genetic analysis

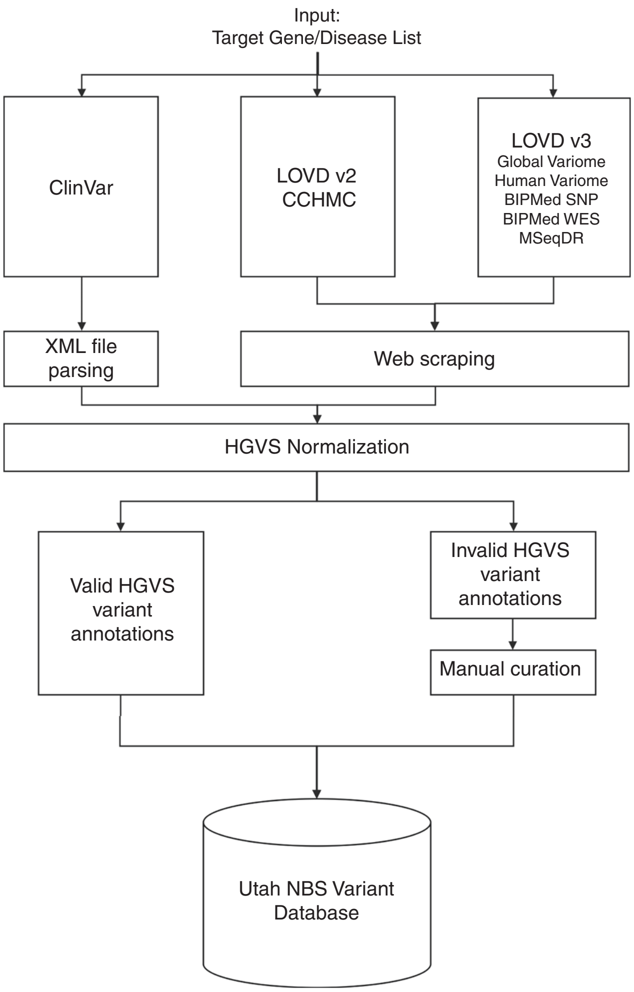

**Fig. 2 Variant database curation pipeline.** The pipeline begins with an input list containing genes associated with newborn screening (NBS) disorders which can be customized by the user. Genomic variant data within target genes was collected programmatically via parsers/python scripts. Individual parsers are required for ClinVar, Leiden Open Variation Database (LOVD) version 2 and LOVD version 3 databases due to differing data formats and data requirements. Variants are collected and Human Genome Variation Society (HGVS) annotations are normalized using the biocommons hgvs python package. Valid HGVS variant annotations are imported into the local variant database while invalid variant annotations are marked for manual curation before being imported into the variant database. This pipeline can be run at user specified intervals to keep the local variant database current with the remote variant databases.

was restricted to the *ARSA* gene. However, two additional loci, *PSAP* and *SUMF1*, are also associated with MLD and are included on clinical diagnostic testing panels. With permission and per request from the collaborator, the analysis was expanded to *PSAP*. The analysis of *SUMF1* was not requested.

## Validation of ES and bioinformatics pipelines

Eleven DBS samples from de-identified newborns with abnormal screening results for SCID ($n = 3$ cases), CF ($n = 7$ cases), and VLCAD deficiency ($n = 1$ case) were included in the validation. A positive control sample from a healthy adult volunteer and an NTC were also included. DNA extraction and exome library generation and sequencing were performed in three independent experiments on a high-throughput flow cell. To establish reproducibility between mid and high-throughput flow cells, a subset of these samples ($n = 5$ cases) were processed through the entire laboratory pipeline and sequenced on a mid-throughput flow cell in three independent experiments. A total of six experiments were performed and concordance between diagnostic testing results and NGS results was reported. Diagnostic testing refers to testing of an independently collected specimen, tested by a clinical reference laboratory employing a validated test, resulting in clinically actionable results.

## Polygenic NBS disorders validation: SCID

Three SCID samples were sequenced on a high-throughput flow cell with two of these samples also included on the mid-throughput validation sample set. Concordance rates of 100% ($n = 2$ variants) and 83.3% ($n = 6$ variants) were observed between diagnostic testing results and ES with in silico analysis restriction to 39 genes associated with SCID in mid- and high-throughput experiments respectively (Table 1).

One SCID case (SCID_3, Table 1) was hemizygous for a variant impacting the splice acceptor region of IL2RG. The pathogenicity of this variant is unknown since it has not previously been reported. Our ES with targeted analysis method detected this variant. Low read coverage (4× coverage) for this variant was observed in one mid-throughput experiment which would have resulted in the variant being filtered out of the results. SCID_1 was confirmed through diagnostic testing revealing a homozygous and pathogenic missense variant in the *ADA* gene.[26] This variant was detected on all mid and high-throughput experiments. SCID_2 was a complex case with four variants in various genes detected through diagnostic testing. Design based, our method could identify three of the four variants that were indeed identified in the study. These included a pathogenic duplication within the *LRBA* gene and two single-nucleotide variants (SNVs) of uncertain significance in *IL2RA* and *IRF8*. None of these variants have been reported in the literature to be associated with SCID. The variant that was not detected because it was outside the a priori specified and selected gene set was a 15q11.2 microdeletion that to our knowledge is associated with developmental disorders, psychiatric disorders, attention deficit disorders, and autism spectrum disorder (ASD) but has not been reported to be associated with SCID.[27] Additional benign variants were identified for these samples for both mid- and high-throughput experiments (data not shown) raising the possibility that (1) the microdeletion is not related to SCID, (2) the identified variants are causal, or (3) both contribute to clinical disease manifestation.

## Single-gene NBS disorders validation: CF and VLCAD deficiency

Seven CF samples and a VLCAD deficiency sample were sequenced on a high-throughput flow cell with two of the CF samples also being included in the mid-throughput validation. The VLCAD deficiency sample was subjected to targeted analysis of *ACADVL* while CF samples were analyzed using two modalities: (1) restricted analysis to 60 *CFTR* variants used by the Luminex assay; (2) restricted analysis of the entire coding portion of the *CFTR* gene with masking of poly-T/poly-TG alleles except in conjunction with c.350G>A (p.Arg117His) variant.[28] Studies have shown that this variant in combination with the 5T variant of the poly-T region is associated with CF as well as CBAVD.[29,30]

Two variants in the VLCAD deficiency sample were detected through diagnostic testing and through ES with targeted analysis (Table 1). For *CFTR*, there was 100% concordance between the Luminex assay and ES with restriction to the 60 Luminex variants for all samples in mid- or high-throughput validation experiments (Table 2). Variants not detected were not included in the panel or did not meet the condition for reporting (e.g., poly-T allele only reported if in conjunction with p.Arg117His). With regard to

**Table 1.** Concordance between diagnostic testing results and ES with in silico analysis restriction to target gene(s).

| Sample | Coding transcript HGVS annotation | Zygosity | Gene (disorder) | ClinVar interpretation | VEP predicted effect | SnpEff predicted effect | Was variant detected by diagnostic testing? | Was variant detected by ES with in silico restriction to target gene(s)? | Total coverage: validation run 1 | Total coverage: validation run 2 | Total coverage: validation run 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mid-output kit validation | | | | | | | | | | | |
| SCID_1 | c.911T>G (p. Leu304Arg) | Hom | *ADA* (SCID) | Likely pathogenic | missense_variant | missense_variant | Yes | Yes | 254 | 169 | 234 |
| SCID_3 | c.925-1G>A | Hemi | *IL2RG* (SCID) | Variant not reported in ClinVar | splice_acceptor_variant | splice_acceptor_variant&intron_variant | Yes | Yes | 4 | 24 | 34 |
| CF_1 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 70 | 121 | 91 |
| | c.489+1G>T | Het | *CFTR* (CF) | Pathogenic | splice_donor_variant | splice_donor_variant&intron_variant | Yes | Yes | 87 | 133 | 139 |
| CF_2 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 131 | 85 | 127 |
| | c.1753G>T (p. Glu585Ter) | Het | *CFTR* (CF) | Pathogenic | stop_gained | stop_gained | Yes | Yes | 110 | 69 | 105 |
| High-output kit validation | | | | | | | | | | | |
| SCID_1 | c.911T>G (p. Leu304Arg) | Hom | *ADA* (SCID) | Likely pathogenic | missense_variant | missense_variant | Yes | Yes | 202 | 198 | 234 |
| SCID_2 | c.4_16dup (p. Asn6delinsSerTer) | Het | *LRBA* (SCID) | Pathogenic | Transcript not annotated by VEP | frameshift_variant&stop_gained | Yes | Yes | 25 | 21 | 26 |
| | c.76G>C (p. Asp26His) | Het | *IL2RA* (SCID) | Uncertain significance | missense_variant | missense_variant | Yes | Yes | 57 | 81 | 100 |
| | c.602C>T (p. Ala201Val) | Het | *IRF8* (SCID) | Conflicting interpretations of pathogenicity | missense_variant | missense_variant | Yes | Yes | 90 | 118 | 46 |
| | 15q11.2 microdeletion | NA | *TUBGCP5*, *CYFIP1*, *NIPA1*, *NIPA2* | Uncertain significance | NA | NA | Yes | Not detectable by method | NA | NA | NA |
| SCID_3 | c.925-1G>A | Hemi | *IL2RG* (SCID) | Variant not reported in ClinVar | splice_acceptor_variant | splice_acceptor_variant&intron_variant | Yes | Yes | 33 | 26 | 34 |
| VLCADD_1 | c.1052C>T (p. Thr351Ile) | Het | *ACADVL* (VLCADD) | Uncertain significance | missense_variant | missense_variant | Yes | Yes | 170 | 191 | 201 |
| | c.1281G>C (p. Trp427Cys) | Het | *ACADVL* (VLCADD) | Uncertain significance | missense_variant | missense_variant | Yes | Yes | 105 | 135 | 148 |
| CF_1 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 84 | 76 | 91 |
| | c.489+1G>T | Het | *CFTR* (CF) | Pathogenic | splice_donor_variant | splice_donor_variant&intron_variant | Yes | Yes | 106 | 87 | 139 |
| CF_2 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 74 | 77 | 127 |
| | c.1753G>T (p. Glu585Ter) | Het | *CFTR* (CF) | Pathogenic | stop_gained | stop_gained | Yes | Yes | 49 | 52 | 105 |
| CF_3 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 85 | 71 | 87 |
| | c.1753G>T (p. Glu585Ter) | Het | *CFTR* (CF) | Pathogenic | stop_gained | stop_gained | Yes | Yes | 51 | 59 | 52 |
| CF_4 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 42 | 68 | 115 |

**Table 1** continued

| Sample | Coding transcript HGVS annotation | Zygosity | Gene (disorder) | ClinVar interpretation | VEP predicted effect | SnpEff predicted effect | Was variant detected by diagnostic testing? | Was variant detected by ES with in silico restriction to target gene(s)? | Total coverage: validation run 1 | Total coverage: validation run 2 | Total coverage: validation run 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | c.1210-33_1210-6GT[11]T[4] | Het | *CFTR* (CF) | Variant not reported in ClinVar | NA | NA | Yes | No | NA | NA | NA |
| | c.1210-33_1210-6GT[10]T[8] | Het | *CFTR* (CF) | Variant not reported in ClinVar | NA | NA | Yes | No | NA | NA | NA |
| CF_5 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 77 | 92 | 92 |
| | c.2490+1G>A | Het | *CFTR* (CF) | Pathogenic | splice_donor_variant | splice_donor_variant&intron_variant | Yes | Yes | 20 | 20 | 18 |
| CF_6 | c.224G>A (p. Arg75Gln) | Hom | *CFTR* (CF) | Conflicting interpretations of pathogenicity | missense_variant | missense_variant | Yes | Yes | 32 | 55 | 60 |
| CF_7 | c.1521_1523del (p. Phe508del) | Het | *CFTR* (CF) | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | 80 | 95 | 71 |
| | c.1210-12T[5] | Het | *CFTR* (CF) | Conflicting interpretations of pathogenicity | NA | NA | Yes | No | NA | NA | NA |

*IL2RG* coding transcript version = NM_000206.2. *ADA* coding transcript version = NM_000022.3; ADA protein reference transcript version = NP_000013.2. *CFTR* coding transcript version = NM_000492.3; CFTR protein transcript version = NP_000483.3. *LRBA* coding transcript version = NM_001364905.1; LRBA protein reference transcript version = NP_001351834.1. *IL2RA* coding transcript version = NM_000417.2; IL2RA protein reference transcript version = NP_000408.1. *IRF8* coding transcript version = NM_002163.4; IRF8 protein reference transcript version = NP_002154.1. *ACADVL* coding transcript version = NM_000018.4; ACADVL protein reference transcript version = NP_000009.1.

*CF* cystic fibrosis, *ES* exome sequencing, *Hemi* hemizygous, *Het* heterozygous, *Hom* homozygous, *SCID* severe combined immune deficiency, *VLCADD* VLCAD deficiency.

**Table 2.** Concordance between Luminex 60 variant *CFTR* assay and ES with in silico analysis restriction to Luminex 60 variant *CFTR* panel.

| Sample | Coding transcript HGVS annotation | Zygosity | ClinVar interpretation | VEP predicted effect | SnpEff predicted effect | Was variant detected with Luminex assay? | Was variant detected by diagnostic testing? | Was variant detected by ES with CFTR panel filter? | Total coverage: validation run 1 | Total coverage: validation run 2 | Total coverage: validation run 3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mid-output kit validation | | | | | | | | | | | |
| CF_1 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 70 | 121 | 127 |
| | c.489+1G>T | Het | Pathogenic | splice_donor_variant | splice_donor_variant&intron_variant | Yes | Yes | Yes | 87 | 133 | 139 |
| CF_2 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 131 | 85 | 127 |
| | c.1753G>T (p. Glu585Ter) | Het | Pathogenic | NA | NA | Not included on panel | Yes | Not included on panel | NA | NA | NA |
| High-output kit validation | | | | | | High-output kit validation | | | | | |
| CF_1 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 84 | 76 | 91 |
| | c.489+1G>T | Het | Pathogenic | splice_donor_variant | splice_donor_variant&intron_variant | Yes | Yes | Yes | 106 | 87 | 139 |
| CF_2 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 74 | 77 | 127 |
| | c.1753G>T (p. Glu585Ter) | Het | Pathogenic | NA | NA | Not included on panel | Yes | Not included on panel | NA | NA | NA |
| CF_3 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 85 | 71 | 87 |
| | c.1753G>T (p. Glu585Ter) | Het | Pathogenic | NA | NA | Not included on panel | Yes | Not included on panel | NA | NA | NA |
| CF_4 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 42 | 68 | 115 |
| | c.1210-33_1210-6GT[11]T[4] | Het | Variant not reported in ClinVar | NA | NA | Not included on panel | Yes | Not included on panel | NA | NA | NA |
| | c.1210-33_1210-6GT[10]T[8] | Het | Variant not reported in ClinVar | NA | NA | Not included on panel | Yes | Not included on panel | NA | NA | NA |
| CF_5 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 77 | 92 | 92 |
| | c.2490+1G>A | Het | Pathogenic | NA | NA | Not included on panel | Yes | Not included on panel | NA | NA | NA |
| CF_6 | c.224G>A (p. Arg75Gln) | Hom | Conflicting interpretations of pathogenicity | NA | NA | Not included on panel | Yes | Not included on panel | NA | NA | NA |
| CF_7 | c.1521_1523del (p. Phe508del) | Het | Pathogenic | inframe_deletion | disruptive_inframe_deletion | Yes | Yes | Yes | 80 | 95 | 71 |
| | c.1210-12 T[5] | | Conflicting interpretations of pathogenicity | NA | NA | Only reported with p. Arg117His | Yes | Only reported with p. Arg117His | NA | NA | NA |

*CFTR* coding transcript version = NM_000492.3; CFTR protein transcript version = NP_000483.3.
*ES* exome sequencing, *Het* heterozygous, *Hom* homozygous.

concordance between diagnostic testing results and ES with analysis restricted to the full *CFTR* coding region, 4/4 (100%) and 11/14 (78.6%) variants detected by diagnostic testing were also discovered by our method in mid- and high-throughput results respectively. Two CF specimens had poly-T and poly-TG allele variants identified through diagnostic testing. These could not be validated by our pipeline. In one case, the variant was filtered out during the indel filtering step of the targeted analysis pipeline. It should be noted however that these variants would not be called by the pipeline since these samples do not have the c.350G>A (p. Arg117His) variant. Poly-T and poly-TG allele status will be confirmed through manual review only in samples with the c.350G>A (p.Arg117His) variant. One *CFTR* variant not included on the Luminex panel, c.1753G>T (p.Glu585Ter), was detected through analysis of the entire *CFTR* coding region.

### Emerging NBS disorders validation: MLD

A pilot biochemical newborn screening study for MLD was conducted screening for sulfatide accumulation in de-identified DBS.[25] MLD screening is complicated by the presence of pseudodeficiency alleles, whereby the structure or the expression of the protein is altered, but disease phenotype is not observed, or is subclinical. To validate the biochemical assay, samples with high sulfatide levels were submitted to an ARSA enzymatic activity assay identifying two samples with elevated sulfatides and deficient ARSA activity. These two DBS samples along with three screen negative samples were subjected to ES targeting *ARSA*, the gene most commonly affected in MLD. Very rare forms of MLD result from variation in *PSAP* or *SUMF1*. Following collaborators' requests, the analysis was expanded to include *PSAP* but not *SUMF1*. Sequencing results from this pipeline validated biochemical findings, with variants observed in *ARSA* in a compound heterozygous affected patient, and a heterozygous unaffected individual. Three unaffected individuals had no pathogenic variant, but two were heterozygous for known pseudovariants. This use case demonstrates the ability to stepwise expand this analysis to emerging disorders and genes and to rapidly expand the analysis to investigate additional genes at the request of the submitter.

### Validation of bioinformatics pipeline using simulated read data sets

To validate the established bioinformatics pipeline and to circumvent a lack of available biological reference resources, we generated variant-specific Variant Call Format (VCF) files. Twelve VCF files containing variants associated with CF, SCID, and Pompe disease were produced, which generated a total of 24 simulated read data sets at 20× and 60× mean exome coverage. We included Pompe disease to ready second-tier testing algorithms supporting biochemical screening beginning later this year. All variants were detected by the bioinformatics pipeline at both mean coverages (Table S1). Additionally, there was 100% agreement between the variant annotation tools VEP and SnpEff.

### Comparison of publicly available genomic variant databases

Interpretation of sequence variants relies in part on what has been observed and reported. Clinically actionable variants have been cataloged in multiple disparate places, including OMIM, which curates at the gene-level from literature reports, ClinVar, a National Institutes of Health (NIH) supported archive of variant-condition assertions from the testing and research communities, and smaller disease or gene focused specialty databases.[18,19] Many of these smaller databases use the same logical schema and supporting software, Leiden Open Variation Database (LOVD), which enables rapid deployment and interoperability between sites.[31] To provide our interpretation team with the most comprehensive assessments, we undertook a comparison and collation of the various databases assembling variants for our conditions of interest. The Human Genome Variation Society (HGVS) provides a structured nomenclature to define variants with regard to their position on the genome and their type (deletion or insertion). Tools like biocommons have been developed to parse and validate these descriptions.[21,32]

For the preliminary iteration of variant curation for our local NBS variant database, we focused on the use cases of polygenic (SCID), single-gene (a selection of metabolic disorders), and emerging NBS disorders (MLD). For SCID, we curated variants for 39 genes associated with the disorder previously included in a candidate gene panel by the New York NBS program.[33] Three genes known to be associated with MLD (*ARSA*, *PSAP*, *SUMF1*) and 13 genes associated with various metabolic disorders were included on the target gene list. The genes selected for MLD and metabolic disorder genes are known to be associated with their respective disorders and are included on diagnostic laboratory disorder panels.[34,35]

In the variant curation process it was necessary to assess the overlap and divergence between ClinVar and other variant databases using the LOVD schema. The total number of SCID variants in ClinVar was 14,113 and 6,865 in LOVDs. The percent overlap between ClinVar and LOVDs for the 39 SCID genes ranged from 3.13% to 31.21% (Fig. 3a). For metabolic disorders, 2,549 variants in ClinVar were associated with metabolic disorders while LOVDs contained 2,172 variants. The range of overlap between ClinVar and LOVDs for all genes associated with metabolic disorders was between 23.31% and 65.08% (Fig. 3b). For MLD, 632 relevant variants were found in ClinVar and 519 variants were found in LOVD databases. *ARSA*, the gene most commonly associated with MLD, had a total of 440 HGVS validated variants identified and aggregated from all databases. This gene also had the greatest percentage of overlap between ClinVar and LOVDs with 35.68% of variants reported in both databases (Fig. 3c). *PSAP* ($n = 279$) and *SUMF1* ($n = 208$) variants had 17.20% and 9.13% overlap respectively between both databases.

Variant types found in all databases included substitutions, deletions, duplications, insertions, indels, and inversions. Substitutions were the most frequent variant type across ClinVar and LOVDs (Fig. S1). Overall, ClinVar and LOVDs appear to contain proportional amounts of variant types regardless of the disorder.

Variants that could not be annotated were binned into seven categories and require further manual curation. Detailed information regarding these categories is summarized in Supplementary materials and methods. In ClinVar, the main reasons variants failed HGVS validation were due to missing variant information and complex HGVS annotations whereas invalid variants in LOVDs lacked the correct reference bases or were complex HGVS annotations (Fig. S2). ClinVar variant annotations are processed through a quality control (QC) pipeline to validate the annotation before upload into the database. LOVD variant databases lack these uniform processing standards and require validation and mapping to an updated reference sequence prior to use.

## DISCUSSION

We developed an NGS-based ES pipeline for second-tier testing in NBS that is disorder and gene agnostic. ES with a priori analysis restriction to one or multiple genes allows initially limited analyses to gene-specific variants and allows expansion to the entire gene-specific coding region(s) if the variant analysis would remain inconclusive. If candidate gene analyses would remain inconclusive, the analysis could be further expanded to additional genes or the entire exome, following parental consent and clinical indication. We have implemented the laboratory methodology using two 3.2-mm DBS punches to generate reliably high-quality sequence data. Data analysis is performed using a custom bioinformatics pipeline. In silico restriction of the analysis is limited to a priori defined genes. As part of the sequencing

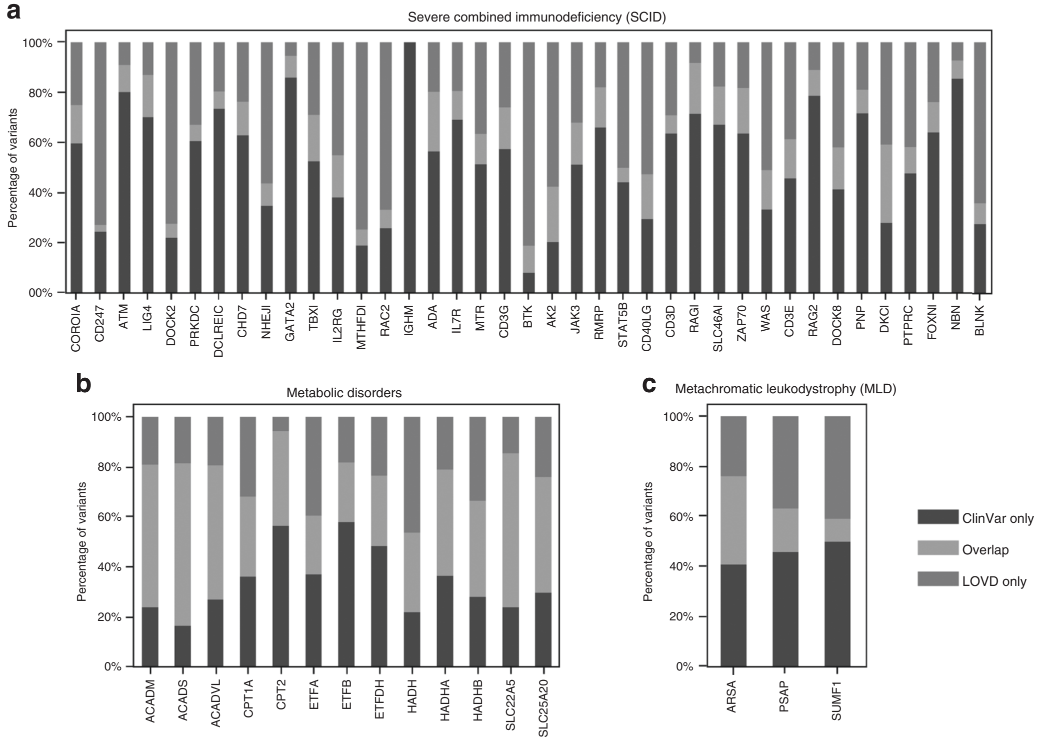

**Fig. 3 Genomic variation overlap between ClinVar and Leiden Open Variation Databases (LOVDs).** The percentage of overlap of valid Human Genome Variation Society (HGVS) annotated variants between ClinVar and LOVDs is shown for (**a**) severe combined immune deficiency (SCID), (**b**) metabolic disorders (MCAD deficiency, SCAD deficiency, VLCAD deficiency, CPT 1 deficiency, CPT 2 deficiency, glutaric acidemia type II, SCHAD deficiency, LCHAD deficiency, primary carnitine deficiency and carnitine–acylcarnitine translocase deficiency), and (**c**) MLD.

pipeline, a local variant database resource was generated and populated with data from an automated pipeline, curating genomic variants from multiple publicly available variant databases. In theory, this method can be applied as a second-tier test to any NBS disorder.

One of the strengths of our method is the multiple entry points for analysis (Fig. 1). While we developed the pipeline for second or third-tier testing from DBS, the analysis can also be performed using already extracted DNA. We demonstrated this for MLD specimens we analyzed using crude DNA extracts.[25] Analysis can also be initiated using raw sequence files (FASTQ) or limited to interpretation using VCF files. Considering the importance of validation of second or third-tier testing methodologies, executing analyses using VCF files is a key strength of initial validation as well as ensuring ongoing accuracy and precision assessments.

The developed pipeline also allows analysis expansion to secondary genes or all coding sequences if no variant information is found in selected genes. While such expanded analysis reduces genetic odysseys, we would require secondary consent by parents or guardians prior to expanded analysis. Such consent must be documented in the patient's electronic health record (EHR) as well. The expanded analysis approach was demonstrated in the analysis of suspected MLD samples, where the *ARSA* gene was included in the primary analysis with *PSAP* included in a secondary analysis at the request of the submitting investigator. In cases where analyses need to be expanded to multiple genes or in cases of diagnostic odysseys, ES analysis can be performed with parental consent and education or counseling strategies.

Limitations to the ES analysis pipeline include restriction to only the coding portions of the genome, limited coverage in exon/intron boundaries, and limited ability to detect large structural variations. ES also does not allow for the identification of variants in deep intronic or in regulatory regions. We observed selection bias present in the exome capture process that can result in high read coverage for some genes while other genes are at or below expected coverage. Omitting the exome capture step and running experiments in full GS mode, however, can detect variants in regulatory and deep intronic regions. As a proof of concept to determine feasibility, the positive control was subjected to GS on a high-throughput flow cell. When comparing coverage for select genes in our ES and GS experiments, some of the selection bias is removed in GS (Table S2). Our current criteria for accepting a variant call is 30× variant coverage with manual review. This cutoff parameter will continue to evolve as we include additional disorders for second-tier ES analysis.

The varying degree of overlap between variant databases points toward the requirement of frequently updating curation. It also highlights the requirement of repeat variant analysis and updating "clinical reports" when interpretations change. The requirements of amended reports and the impact on clinical management challenges newborn screening follow-up systems, requiring long-term follow-up structures and the maintenance of accurate demographic and provider information. Variant database upgrades might also require revalidation of the pipeline. To deal with the issue of approximately 10% of variants failing validation, we binned such variants based on "failure mechanisms," marking them for manual assessment at a later time or when diagnostically needed.

While ClinVar is becoming the industry standard archive for variant annotation and is heavily used as a source of reference during clinical variant interpretation, we have demonstrated varying degrees of overlap between the current content of ClinVar and other curated boutique databases. We had expected that the databases would include a larger proportion of the same variants, and the differences would be at the level of clinical significance. The disjunctive union between databases has multiple causes. While some conditions have relatively common variants, such as deltaPhe(508)-*CFTR* in cystic fibrosis, there are many other rare or private variants that cause disease that have yet to propagate into the large variant resources due to very low frequency in the population. Another reason is that variants of uncertain significance and known benign variants may not propagate as rapidly to the large databases. Similar results have been observed during sequence-based NBS, where a significant proportion of detected variants were not present in existing databases.[1] Here the authors showed that for commonly screened disorders, between 13% and 38% of the observed variants were not annotated in ClinVar. Our findings build upon this research, and provide a reminder to those performing genomic interpretation that a single catalog of genomic variation for NBS genes has yet to be achieved. Another source of information vital to interpretation is variant frequency from databases such as gnomAD.[20] We believe that automated methods such as those we have developed can be used to supplement the detailed curation of clinical domain working groups such as those working via the ClinGen Initiative, and provide clinical genetics providers a single source of variant annotations to aid with their interpretation activities.[36] There are multiple clinical domain working groups in the area of inborn errors of metabolism and this described pipeline is a clear adjunct to those activities.[37] A detailed and comprehensive catalogue of collated NBS variant interpretations is another tool to aid those charged with making clinical diagnoses.

Biologic variability potential at every nucleotide position measured by sequencing-based tests challenges the validation standards and requirements of laboratory and diagnostic medicine.[38–41] While traditional biochemical tests measure one analyte, the validation of the actual test measuring the single analyte is straightforward and in general universally agreed upon. By definition, applying such biochemical validation standards to NGS based tests would require performance characterization at every nucleotide position, a task that is impossible based on the number of theoretical variations and the lack of biological reference material. While the laboratory component of the test can be straightforwardly controlled through extraction controls and traditional control steps, we developed simulated, in silico reads to measure and standardize analysis performance. Such control materials can be developed based on a variable frequency ranging from common to rare variants. These resources can be analyzed through the pipeline in a quality control assurance step prior to any patient analysis, proficiency testing, or to fulfill revalidation requirements after periodic variant database upgrades. Furthermore, such simulated “material” can be readily shared with auditors and collaborators to compare performance across programs and laboratories.

Many times an initial newborn screening is inconclusive due to the presence of an intermediate phenotype.[42] Given time, comprehensive population screening of intermediate phenotypes in combination with the genetic variant assessment will result in a more thorough and comprehensive understanding of the variant space and consequences. We advocate that the community must focus on comprehensiveness of annotation and curation of observed variants irrespective of agreement on interpretation or discourse. As such analyses are adopted globally, community knowledge will advance understanding of natural history as well as establish any underlying phenotype–genotype relationships between marker and trait. While there are national efforts to collect and curate variant–phenotype pairs, the NBS community is the first responder to new variants and in a position to greatly impact and improve the community of knowledge.[19,43]

We chose ES for second-tier testing from a cost/benefit standpoint. Our current turnaround time for ES with targeted analysis of five DBS samples is four days with variable costs at $600 per sample. In the future, methods such as rapid GS or long read methodologies may also be considered as they eliminate selection bias and have significantly faster turnaround times. Using the Illumina NextSeq platform, sequencing one genome of one sample on a high-throughput flow cell we observed an average coverage of 30×. If GS was the method of choice, this platform would not be sufficient for a production environment.

While this NGS method is not replacing biochemical NBS, it aims to expand second-tier testing aiding in clinical decision support. To maximize these benefits, screening programs must seek consensus with the medical care teams regarding the utility of the test. If testing is performed on the same dried blood specimen are the results clinically actionable? Or should testing be performed on an independent new specimen? Likewise, considering potentially long turnaround times, should such testing be only performed under the umbrella of the diagnostic testing framework? Expanded analyses can result in incidental findings or the identification of disease-causing variants with unrelated disorders or disease manifestations. Such unintended consequences have to be part of the consenting process and must be clearly explained.

## DATA AVAILABILITY

Sequence analysis pipelines and read simulation data sets are available at https://github.com/UtahNBS/WES-Secondary-Testing.

## CODE AVAILABILITY

Code developed for the variant database curation pipeline is available at https://github.com/eilbecklab/Utah-DOH-newborn-screening.

Received: 7 August 2020; Revised: 24 November 2020; Accepted: 25 November 2020;
Published online: 13 January 2021

## REFERENCES

1. Bodian, D. L. et al. Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. *Genet. Med.* **18**, 221–230 (2016).
2. Johnston, J. et al. Sequencing newborns: a call for nuanced use of genomic technologies. *Hastings Cent. Rep.* **48**(Suppl 2), S2–S6 (2018).
3. Landau, Y. E., Lichter-Konecki, U. & Levy, H. L. Genomics in newborn screening. *J. Pediatr.* **164**, 14–19 (2014).
4. Bodamer, O. A., Scott, C. R. & Giugliani, R., Pompe Disease Newborn Screening Working Group. Newborn screening for Pompe disease. *Pediatrics.* **140**(Suppl 1), S4–S13 (2017).
5. Burton, B. K., Kronn, D. F., Hwu, W. L. & Kishnani, P. S. The initial evaluation of patients after positive newborn screening: recommended algorithms leading to a confirmed diagnosis of Pompe disease. *Pediatrics.* **140**(Suppl 1), S14–S23 (2017).
6. Arnold, W. D., Kassar, D. & Kissel, J. T. Spinal muscular atrophy: diagnosis and management in a new therapeutic era. *Muscle Nerve* **51**, 157–167 (2015).
7. Clarke, L. A. et al. Mucopolysaccharidosis type I newborn screening: best practices for diagnosis and management. *J. Pediatr.* **182**, 363–370 (2017).
8. Kemper, A. R. et al. Newborn screening for X-linked adrenoleukodystrophy: evidence summary and advisory committee recommendation. *Genet. Med.* **19**, 121–126 (2017).
9. Schram, C. A. Atypical cystic fibrosis: identification in the primary care setting. *Can. Fam. Physician* **58**, 1341–1345 (2012).
10. Harrington, M. et al. Insights into the natural history of metachromatic leukodystrophy from interviews with caregivers. *Orphanet J. Rare Dis.* **14**, 89 (2019).
11. Oksenhendler, E. et al. Infections in 252 patients with common variable immunodeficiency. *Clin. Infect. Dis.* **46**, 1547–1554 (2008).
12. Shchelochkov, O., Wong, L. J., Shaibani, A. & Shinawi, M. Atypical presentation of VLCAD deficiency associated with a novel ACADVL splicing mutation. *Muscle Nerve* **39**, 374–382 (2009).
13. Baker, M. W. et al. Improving newborn screening for cystic fibrosis using next-generation sequencing technology: a technical feasibility study. *Genet. Med.* **18**, 231–238 (2016).
14. Kharrazi, M. et al. Newborn screening for cystic fibrosis in California. *Pediatrics.* **136**, 1062–1072 (2015).
15. Taylor, J. L. et al. The North Carolina experience with mucopolysaccharidosis type I newborn screening. *J. Pediatr.* **211**, e192 (2019).
16. Lee, S. et al. Evaluation of X-linked adrenoleukodystrophy newborn screening in North Carolina. *JAMA Netw. Open* **3**, e1920356 (2020).

17. van Nimwegen, K. J. et al. Is the $1000 genome as near as we think? A cost analysis of next-generation sequencing. *Clin. Chem* **62**, 1458–1464 (2016).
18. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res.* **47**, D1038–D1043 (2019).
19. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. **46**, D1062–D1067 (2018).
20. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* **581**, 434–443 (2020).
21. Wang, M. et al. hgvs: a Python package for manipulating sequence variants using HGVS nomenclature: 2018 update. *Hum. Mutat.* **39**, 1803–1813 (2018).
22. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.11–33 (2013).
23. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* **28**, 2520–2522 (2012).
24. Stephens, Z. D. et al. Simulating next-generation sequencing data sets from empirical mutation and sequencing models. *PLoS One* **11**, e0167047 (2016).
25. Hong, X. et al. Toward newborn screening of metachromatic leukodystrophy: results from analysis of over 27,000 newborn dried blood spots. *Genet Med.* https://doi.org/10.1038/s41436-020-01017-5 (2020).
26. Hirschhorn, R. Identification of two new missense mutations (R156C and S291L) in two ADA- SCID patients unusual for response to therapy with partial exchange transfusions. *Hum. Mutat.* **1**, 166–168 (1992).
27. Cox, D. M. & Butler, M. G. The 15q11.2 BP1–BP2 microdeletion syndrome: a review. *Int. J. Mol. Sci.* **16**, 4068–4082 (2015).
28. Deignan, J. L. et al. CFTR variant testing: a technical standard of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med*. https://doi.org/10.1038/s41436-020-0822-5 (2020).
29. Thauvin-Robinet, C. et al. CFTR p.Arg117His associated with CBAVD and other CFTR-related disorders. *J. Med. Genet.* **50**, 220–227 (2013).
30. Bienvenu, T., Beldjord, C., Adjiman, M. & Kaplan, J. C. Male infertility as the only presenting sign of cystic fibrosis when homozygous for the mild mutation R117H. *J. Med. Genet.* **30**, 797 (1993).
31. Fokkema, I. F. et al. LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat*. **32**, 557–563 (2011).
32. den Dunnen, J. T. et al. HGVS recommendations for the description of sequence variants: 2016 update. *Hum. Mutat.* **37**, 564–569 (2016).
33. Stevens, C. Next Generation Sequencing in the New York State Newborn Screening Molecular Lab. https://www.aphl.org/programs/newborn_screening/Documents/2017%20Gene%20Sequencing%20Meeting/Stevens_Second%20Tier%20and%20Future%20Applications.pdf. Accessed 7 Aug 2020.
34. van Rappard, D. F., Boelens, J. J. & Wolf, N. I. Metachromatic leukodystrophy: disease spectrum and approaches for treatment. *Best. Pract. Res. Clin. Endocrinol. Metab.* **29**, 261–273 (2015).
35. Rice, G. M. & Steiner, R. D. Inborn errors of metabolism (metabolic disorders). *Pediatr. Rev.* **37**, 3–15 (2016).
36. Rehm, H. L. et al. ClinGen—the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
37. Zastrow, D. B. et al. Unique aspects of sequence variant interpretation for inborn errors of metabolism (IEM): the ClinGen IEM Working Group and the phenylalanine hydroxylase gene. *Hum. Mutat.* **39**, 1569–1580 (2018).
38. Aziz, N. et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. *Arch. Pathol. Lab. Med.* **139**, 481–493 (2015).
39. Richards, 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.* **17**, 405–423 (2015).
40. Luh, F. & Yen, Y. FDA guidance for next generation sequencing-based testing: balancing regulation and innovation in precision medicine. *NPJ Genom. Med.* **3**, 28 (2018).
41. Roy, S. et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **20**, 4–27 (2018).
42. Drendel, H. M. et al. Intermediate MCAD deficiency associated with a novel mutation of the ACADM gene: c.1052C>T. *Case Rep. Genet*. **2015**, 532090 (2015).
43. Rivera-Muñoz, E. A. et al. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum. Mutat.* **39**, 1614–1622 (2018).

## ACKNOWLEDGEMENTS

This work was supported by the Association of Public Health Laboratories through the Health Resources and Services Administration (HRSA), US Department of Health and Human Services (HHS UG9MC30369). This information or content and conclusions are those of the author and should not be construed as the official position or policy of, nor should any endorsements be inferred by HRSA, HHS, or the US Government. Funding was also provided by the National Institutes of Health, National Human Genome Research Institute (5U41HG009650 to K.E.) and National Library of Medicine (T15LM007124 to D.S.).

## AUTHOR CONTRIBUTIONS

Conceptualization: A.R., K.E. Curation: D.S., J.L., K.C., N.R.-S. Formal analysis: D.S., J.L., K.C., N.R.-S. Funding acquisition: A.R., K.E. Investigation: A.R., D.S., J.L., K.C., K.E., K.H., N.R.-S., S.N., W.D. Methodology: A.R., D.S., E.L.Y., K.E., K.F.O., N.R.-S. Software: B.A., D.S., J.L., K.C., N.R.-S. Visualization: D.S., N.R.-S. Writing—original draft: A.R., D.S., K.E., N.R.-S., S.N. Writing—review & editing: A.R., B.A., D.S., E.L.Y., K.E., K.F.O., K.H., N.R.-S., S.N., W.D.

## ETHICS DECLARATION

Institutional review board (IRB) approval for the analysis of MLD screen-positive and screen-negative samples was granted by the Washington state IRB Project B-062702-H07.23. The Utah Department of Health IRB determined that IRB approval was not required for work utilizing de-identified CF, SCID, and VLCAD deficiency specimens, as these were analyzed as a part of a validation and process improvement project.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-020-01058-w) contains supplementary material, which is available to authorized users.

**Correspondence** and requests for materials should be addressed to A.R.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.