nature medicine

LETTERS

https://doi.org/10.1038/s41591-020-0966-5

Check for updates

The role of exome sequencing in newborn screening for inborn errors of metabolism

Aashish N. Adhikari[1,2]✉, Renata C. Gallagher[2,3], Yaqiong Wang[1], Robert J. Currier[3], George Amatuni[3], Laia Bassaganyas[2], Flavia Chen[2,4], Kunal Kundu[1,5], Mark Kvale[2], Sean D. Mooney[6], Robert L. Nussbaum[2,7], Savanna S. Randi[8], Jeremy Sanford[8], Joseph T. Shieh[2,3], Rajgopal Srinivasan[5], Uma Sunderam[5], Hao Tang[9], Dedeepya Vaka[2], Yangyun Zou[1], Barbara A. Koenig[2,4], Pui-Yan Kwok[2,10,11], Neil Risch[2,12], Jennifer M. Puck[2,3,10,13,16]✉ and Steven E. Brenner[1,2,14,15,16]✉

Public health newborn screening (NBS) programs provide population-scale ascertainment of rare, treatable conditions that require urgent intervention. Tandem mass spectrometry (MS/MS) is currently used to screen newborns for a panel of rare inborn errors of metabolism (IEMs)[1–4]. The NBSeq project evaluated whole-exome sequencing (WES) as an innovative methodology for NBS. We obtained archived residual dried blood spots and data for nearly all IEM cases from the 4.5 million infants born in California between mid-2005 and 2013 and from some infants who screened positive by MS/MS, but were unaffected upon follow-up testing. WES had an overall sensitivity of 88% and specificity of 98.4%, compared to 99.0% and 99.8%, respectively for MS/MS, although effectiveness varied among individual IEMs. Thus, WES alone was insufficiently sensitive or specific to be a primary screen for most NBS IEMs. However, as a secondary test for infants with abnormal MS/MS screens, WES could reduce false-positive results, facilitate timely case resolution and in some instances even suggest more appropriate or specific diagnosis than that initially obtained. This study represents the largest, to date, sequencing effort of an entire population of IEM-affected cases, allowing unbiased assessment of current capabilities of WES as a tool for population screening.

Effective population-level NBS must rapidly identify the few individuals at risk of disease with extraordinary sensitivity, high specificity and limited manual review. In California, NBS for 48 different IEMs performed with MS/MS[1–4] achieved 99.0% sensitivity and >99.8% specificity[5]. For some disorders, MS/MS nonetheless has a low positive predictive value and results may be nonspecific (Fig. 1).

Genomic sequencing, now commonly used for diagnosis of rare disorders[6–10], has been recommended for nearly all seriously ill children in intensive care units[8–10], proposed for all newborns to personalize their medical care[11] and marketed for screening newborns[12]. Yet, population-scale studies to establish performance characteristics of sequencing for NBS have not been reported. NBS IEMs provide an ideal model for evaluating the role of sequencing in population screening because most are Mendelian disorders affecting well-understood biochemical pathways and many have been studied extensively. Moreover, sensitivity and specificity of sequence-based detection of IEMs can be directly compared to those of current MS/MS screening. Studying WES to identify IEMs already included in NBS can also suggest its potential utility for further treatable disorders not amenable to detection by MS/MS[13].

We quantified the performance of WES, were it to have been the primary NBS for IEMs in California. Between July 2005, and December 2013, the Genetic Disease Screening Program (GDSP) of the California Department of Public Health (CDPH) screened dried blood spots (DBSs) from nearly 4.5 million neonates for 48 IEMs using a multiplex MS/MS platform. We obtained a comprehensive set of 1,728 residual, de-identified, archived DBSs representing all cases with IEMs, as well as a select set that were screened as positive but later were determined to be unaffected. We performed WES for 1,416 DBSs[14] and determined that DBS exomes passing quality controls (Methods) were comparable to exomes from fresh blood (Extended Data Figs. 1–4). We analyzed 1,190 high-quality exomes from 805 IEM-affected individuals and 385 MS/MS false positives (Table 1); exomes were divided into validation and test sets, with 178 and 1,012 individuals, respectively (Table 1).

We analyzed variants within an 'exome slice'[15] of 78 genes associated with the 48 IEMs ascertained by NBS in California (Methods; Supplementary Table 1). We systematically explored pipeline parameters on the validation set to derive a customized, robust, sensitive pipeline for reporting potential disease-causing variants in IEM genes in a screening context (Fig. 2a; Methods). The automated pipeline had three arms. The curation arm reported all variants curated as pathogenic in existing disease databases (Human Gene

[1]Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA. [2]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA. [3]Department of Pediatrics, University of California San Francisco, San Francisco, CA, USA. [4]Program in Bioethics, University of California San Francisco, San Francisco, CA, USA. [5]Innovation Labs, Tata Consultancy Services, Hyderabad, India. [6]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA. [7]Invitae, San Francisco, CA, USA. [8]Department of Molecular, Cellular and Developmental Biology, Center for the Molecular Biology of RNA, UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA. [9]Genetic Disease Screening Program, California Department of Public Health, Richmond, CA, USA. [10]Cardiovascular Research Institute, University of California San Francisco, San Francisco, CA, USA. [11]Department of Dermatology, University of California San Francisco, San Francisco, CA, USA. [12]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA. [13]Division of Allergy, Immunology and Blood and Marrow Transplantation, UCSF Benioff Children's Hospital, San Francisco, CA, USA. [14]Center for Computational Biology, University of California Berkeley, Berkeley, CA, USA. [15]Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA. [16]These authors contributed equally and jointly supervised the work: Jennifer M. Puck, Steven E. Brenner. ✉e-mail: anadhikari@berkeley.edu; jennifer.puck@ucsf.edu; brenner@compbio.berkeley.edu

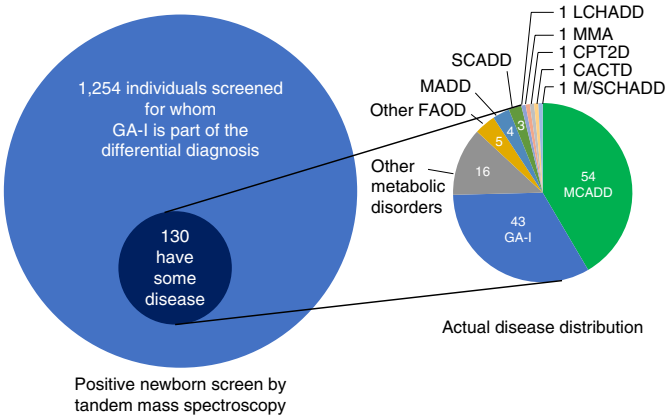


Fig. 1 | Low positive predictive value and complex differential diagnoses of MS/MS newborn screening for glutaric acidemia. Among 1,254 newborns with positive glutaric acidemia type I (GA-I) MS/MS screen (California, July 2005 through December 2013), only 130 were ultimately diagnosed with any IEM. Of these 130, only 43 actually had a diagnosis of GA-I, while the rest had other IEMs, including MCADD, LCHADD, MMA, carnitine palmitoyl transferase deficiency type II (CPT2D), medium/short-chain 3-hydroxyacyl-CoA dehydrogenase deficiency (M/SCHADD), MADD, short-chain acyl-CoA dehydrogenase deficiency (SCADD), carnitine-acylcarnitine translocase deficiency (CACTD) or another fatty acid oxidation disorder (FAOD). The GA-I MS/MS screen is based on elevations of glutarylcarnitine (C5-DC), along with informative ratios. During the early part of the study, a derivatized method was used, in which hydroxydecanoylcarnitine (C10-OH) had the same mass-to-charge ratio as C5-DC. After the methodology was switched to use un-derivatized metabolites, it became hydroxyhexanoylcarnitine (C6-OH) that was coincident with C5-DC.

Table 1 | Distribution of analyzed newborns by study set, disorder and race/ethnicity

	Total newborns ($n=1,190$)	Affected ($n=805$)	Unaffected false positives by MS/MS ($n=385$)
Study set			
Validation set	178	131	47
Test set	1,012	674	338
Categories of disorder			
Fatty acid oxidation disorders		334	
Organic acid disorders		237	
Amino acid disorders		234	
Race/ethnicity[a]			
Hispanic	454	330	124
Non-Hispanic white	371	226	145
East Asian	79	57	22
African American	63	42	21
Other	223	150	73

[a]Race/ethnicity was based on nursery-reported categories from the newborn DBS forms.

Mutation Database (HGMD)[16] or ClinVar[17]) with population minor allele frequency (MAF) < 0.1% and additional potentially pathogenic less-rare variants that we curated (see Methods; Supplementary Table 2). The second arm reported all rare (MAF < 0.5%) missense and nonsense variants and other rare variants predicted to be protein-altering or damaging. The third arm reported inferred copy-number variants (CNVs). Usually, phase could not be determined, so individuals with two reportable variants were designated 'exome positive' for their associated IEMs. Exome results were compared to the clinical data collected by GDSP[18].

As a primary screen, the pipeline correctly identified 571 of 674 IEM-affected infants in our test set as having a potentially pathogenic IEM genotype (Fig. 2b and Supplementary Table 3). Overall sensitivity, calculated as the ratio of exome-positives whose gene matched the reviewed diagnosis to IEM-affected individuals, weighted by the prevalence of each IEM in California[19], was 88%. For the clinically confident subset of individuals in the test set (Methods), the pipeline achieved 93.7% overall sensitivity. Our near-complete ascertainment of IEM-positive infants established this as a population-scale benchmark for WES sensitivity. Of the 571 infants, 360 (63%) had only pathogenic rare variants from databases[20,21] and 17 (3%) had at least one less-rare, curated variant. The remaining 194 (34%) carried at least one previously unannotated, rare, potentially pathogenic variant: 128 nonsynonymous missense; 51 nonsense, indel or canonical splice site; 12 splice-altering; and 3 CNVs. While 50 of the 103 affected infants not identified by the pipeline had a single reportable heterozygous variant, half had no reportable variants found in any gene associated with the clinical diagnosis.

Eleven infants were exome-positive for genes unrelated to their IEMs (Methods), thus, producing an overall specificity for WES of 98.4%. This would extrapolate to ~8,000 false positives among the half million annual births in California, far more than the actual 1,362 MS/MS false-positive cases in 2015. An alternative pipeline that reported only rare curated variants as pathogenic (Extended Data Fig. 5h) had high specificity, 99.4%, but unacceptably low sensitivity, 55%.

Sensitivity of our screening pipeline varied by disorder, generally performing better for less-rare disorders. Although sensitivity was 100% for nearly one-third of the NBS IEMs, statistical confidence, particularly for very rare IEMs, would require more data. (Fig. 2c and Supplementary Table 4).

Across all IEMs currently screened for by MS/MS, WES had insufficient sensitivity and specificity for sole, primary use as a replacement for MS/MS. Sensitivity could not be improved by tradeoffs against specificity, as the missed cases lacked any pair of rare missense or predicted damaging variants in the relevant genes.

Our study also illuminated the genetic landscape of IEMs in California. From the 571 exome true positive individuals in the test set, the pipeline reported a total of 1,157 variants, comprising 507 distinct variants: 343 missense (68%), 47 frameshifting indels (9%), 41 nonsense (8%), 32 splice site (6%), 19 predicted damaging intronic nucleotide substitutions near a splice site (4%) and 17 nonframeshifting indels (3%). The remaining eight (2%) were large deletions and curated variants (Extended Data Fig. 6a). Of these 507 variants, 195 (38%) were absent from ExAC[21] and 162 (32%) were absent from both HGMD[16] and ClinVar[17] (Supplementary Table 5); 384 of the 507 (76%) occurred in only one case.

We further investigated selected exome false negatives by performing whole-genome sequencing (WGS) for eight individuals lacking two reportable WES variants (Methods). Large CNVs cannot be identified reliably from WES; indeed, WGS revealed one individual with isovaleric acidemia (IVA) having a deletion of the first three exons of the *IVD* gene and another with a complex pattern of reduced *IVD* gene coverage (Extended Data Fig. 7). However, our targeted manual inspection of the read alignment and coverage of relevant genes in the remaining six failed to suggest a widespread role for CNVs in other IEMs (Methods), though complex structural rearrangements may remain undetected by WGS.

Additional WES false negatives could be due to limitations in variant interpretation. One individual with medium-chain acyl-CoA dehydrogenase deficiency (MCADD) had only one reportable variant in the *ACADM* gene (NP_000007.1:p.Tyr67His). A second variant located 14 bases 5' from the splice site (NM_000016.4:c.388-14A>G) had insufficient evidence for pathogenicity (Extended Data Fig. 8a). However, an experimental splicing reporter assay revealed that this variant switched splicing to a new 3' site, extending exon 6 by 13 nucleotides, creating a premature termination codon (Extended Data Fig. 8b–e; Methods).

We next considered WES as a reflex follow-up test for MS/MS-positive individuals before conducting further biochemical/clinical studies, to reduce the referral of unaffected infants. For six IEMs with challenges in MS/MS screening, we analyzed all IEM-affected and unaffected infants in validation and test sets with positive MS/MS, but with no reportable WES variants (Table 2 and Extended Data Fig. 9). Among all MS/MS-positive individuals, the negative predictive value (NPV) or proportion of unaffected individuals among those with no reportable WES variant, ranged from 33%, for multiple acyl-CoA dehydrogenase deficiency (MADD)/glutaric acidemia type II (GA-II) to 100% for long-chain 3-hydroxyacyl-CoA dehydrogenase deficiency (LCHADD) (Table 2). For very long-chain acyl-CoA dehydrogenase deficiency (VLCADD) and maple syrup urine disease (MSUD), one affected individual of each would have been wrongly excluded due to no reportable WES variant, whereas 48 and 16 false positives, respectively, would have avoided further follow-up, provided that WES could be performed rapidly.

In 12 individuals, the initial clinical diagnosis in the GDSP database was not consistent with the gene reported by WES, but the final disorder assignment from our subsequent clinical review was concordant with WES. Often, the disorders in question had overlapping MS/MS analyte profiles (Supplementary Table 6). For example, an individual clinically reported to have IVA, had a rare, homozygous, missense variant in the *ACADSB* gene encoding short/branched-chain acyl-CoA dehydrogenase (NM_001609.3:c.1165A>G; NP_001600.1:p.Met389Val (precursor protein), MAF 0.002%), suggesting 2-methylbutyryl-CoA dehydrogenase deficiency. This variant causes skipping of *ACADSB* exon 10 (ref. [22]), leading to elevated 2-methylbutyrylcarnitine, which has the same MS/MS signature as isovalerylcarnitine, an analyte elevated in IVA[13].

Another infant diagnosed with MADD/GA-II had a homozygous missense variant in *ETHE1*, encoding a mitochondrial sulfur dioxygenase (NM_014297.4:c.488G>A; NP_055112.2:p.Arg163Gln, MAF 0.003%). This variant causes ethylmalonic encephalopathy[23], which may be difficult to distinguish biochemically from MADD/GA-II (ref. [24]).

A third individual, MS/MS-positive for both VLCADD and MCADD, had been diagnosed as unaffected upon follow-up, with notes suggesting VLCADD carrier status. WES revealed two missense variants in the electron transfer flavoprotein dehydrogenase gene *ETFDH* (NM_004453.3:c.250G>A; NP_004444.2:p.Ala84Thr, MAF 0.02% and NM_004453.3:c.524G>A; p.Arg175His, not observed in ExAC). These variants, previously observed as compound heterozygous in patients with MADD/GA-II with onset of symptoms as teenagers or adults[25], reduce protein activity; sibling studies indicated that NP_004444.2:p.Ala84Thr is penetrant (Z.-Y. Wu, personal communication). Notably, all pertinent MS/MS analyte values for VLCADD carriers were consistent with MADD/GA-II (Supplementary Fig. 1).

The WES false positives in our study could have been due to lack of predictive accuracy for uncharacterized, rare, but benign variants, as well as colocalization of two reported variants on the same, rather than opposite, chromosomes. Long-read sequencing technologies could mitigate phasing limitations in the future.

Fig. 2 | Whole-exome pipeline design and analysis. a, Diagram of pipeline for analysis. For each exome, the pipeline considered only variants with genotype quality (GQ) > 15 that impacted principal transcripts (per annotation of principal and alternative splice isoforms, APPRIS[35]) for 78 genes associated with currently screened IEMs. Variants were identified through any of three arms. The predicted impactarm (left) included variants with population MAF < 0.5% in both 1000 Genomes Project[20] and ExAC[21] databases and predicted protein alteration (stop gain or loss, frameshift insertion or deletion, alteration of canonical splice motif, nonsynonymous missense, in-frame insertion or deletion, start gain or loss) (i); combined annotation-dependent depletion (CADD)[36] score > 23 (ii); or predicted splicing impact (determined by database of splicing consensus single-nucleotide variant (dbscSNV) labeled random forest meta-estimator (RF))[37] with meta-prediction score >0.5 (iii). The curation arm (center) included variants with MAF < 0.1% annotated as disease mutation (DM) or questionable disease mutation (DM?) in HGMD[16] or as pathogenic/likely pathogenic by ClinVar[17], with at least one review star. Among HGMD DM/DM? or ClinVar pathogenic/likely pathogenic variants with MAF ≥ 0.1% ($n = 60$), 19 were considered reportable and 41 were excluded (Supplementary Table 2). The predicted CNV arm (right) applied XHMM[38] to IEM genes except three with common intragenic deletions (*ETFA*, *HCFC1* and *PRODH*). For variants in X chromosome genes(*), the MAF threshold was adjusted to 0.02%. For genes with ≥1 heterozygous variant, local phasing was performed when reads overlapped the multiple variant positions. Finally, variants annotated as benign in ClinVar with at least two review stars were excluded. From the list of resulting variants, the corresponding genes with ≥1 homozygous or ≥2 heterozygous variants were reported. Exceptions(**) were X-linked ornithine transcarbamylase, for which ≥1 flagged, heterozygous *OTC* gene variant was reported, as heterozygous females can display a clinical phenotype;[30] and methionine adenosyltransferase-1A, for which heterozygous variant *MAT1A* NP_000420.1:p.Arg264His causes autosomal dominant disease[39]. For each individual, a single gene was chosen as the likely disease-causing gene by a score incorporating IEM prevalence and variant severity (Methods). **b**, Contributions from components of the pipeline in identifying variants in the test set of exomes from 674 IEM-affected infants. Of the 571 cases correctly identified by the pipeline, 360 had only rare HGMD/ClinVar curated variants, whereas the rest required additional curation or predictions. Of the 103 individuals missed by exomes, the pipeline reported a single autosomal heterozygous variant in 53 individuals in a gene consistent with the disorder and no variants in 50. **c**, Sensitivity of exome pipeline by disorder for the 674 IEM-affected infants from the test set. For each IEM, the numbers of individuals correctly identified by exomes are shown by green bars and the number missed by exomes are shown by brown bars, with sensitivity for each IEM disorder shown in parentheses. Core IEMs screened by California (left); secondary/add-on IEMs (right). IEMs sharing a common causative gene were not distinguished by the exome predictions alone. These included trifunctional protein deficiency (TFP) and LCHADD (blue shading), PKU and hyperphenylalaninemia (pink shading) and the various MMA subtypes (yellow shading). ASAL, argininosuccinate lyase deficiency; BKT, beta-ketothiolase deficiency; 3HMG, 3-hydroxy-3-methylglutaryl-CoA lyase deficiency; PKU, phenylketonuria; CIT-I, citrullinemia type I; CTD/CUD, carnitine transporter deficiency/carnitine uptake defect; 3MCC, 3-methylcrotonyl-CoA carboxylase deficiency; MCD, multiple carboxylase deficiency; IBD/IBG, isobutyryl-CoA dehydrogenase deficiency; ARG, arginase deficiency; remethylation, remethylation defects; BIOPT-BS, biopterin disorders (co-factor biosynthesis); MGA type I, 3-methylglutaconyl-CoA hydratase deficiency; CPT-I and CPT-II, carnitine palmitoyl transferase deficiency type I and II, respectively; EE, ethylmalonic encephalopathy; Hyperphe, hyperphenylalaninemia; 2MBCD, 2-methylbutyryl-CoA dehydrogenase deficiency; OTC, ornithine transcarbamylase deficiency; MET/MAT, hypermethioninemia/methionine adenosyltransferase I/III deficiency; CATD/CACTD, carnitine-acylcarnitine translocase deficiency; MAL, malonic acidemia.

Of greater concern for screening, however, were the 12% exome false negatives. Upon manual inspection of all genes relevant to the IEM diagnosis in each individual in the validation set, observed variants were either common or not protein-altering, with low likelihood of pathogenicity. Inadequate exome coverage may have limited our ability to detect deletions, but might be improved by using optimized gene panels[26,27]. Areas of poor coverage included exon 1 of *MCCC2* (Extended Data Fig. 4), a locus of known pathogenic variants. Identifying pathogenic variants that are not clearly protein altering remains challenging.

Cases could also be missed owing to incomplete knowledge of genetic disease. We analyzed 78 proven IEM-associated genes, but other relevant genes may be yet undiscovered; 12 of 19 individuals with methylmalonic acidemia (MMA) had no WES variants detected (Fig. 2c), similar to a previous report[28]. Moreover, genetic mechanism may be dominant, epistatic, epigenetic or oligogenic with synergistic heterozygosity[29]. Finally, inherent biological factors limit disease prediction from sequence alone; for example, heterozygous OTC variants in females manifest variable penetrance depending on X chromosome inactivation patterns in hepatocytes[30].

Nonetheless, following identification of infants with abnormal analytes on MS/MS by NBS, the metabolic centers to which these infants are referred may find WES invaluable for suggesting

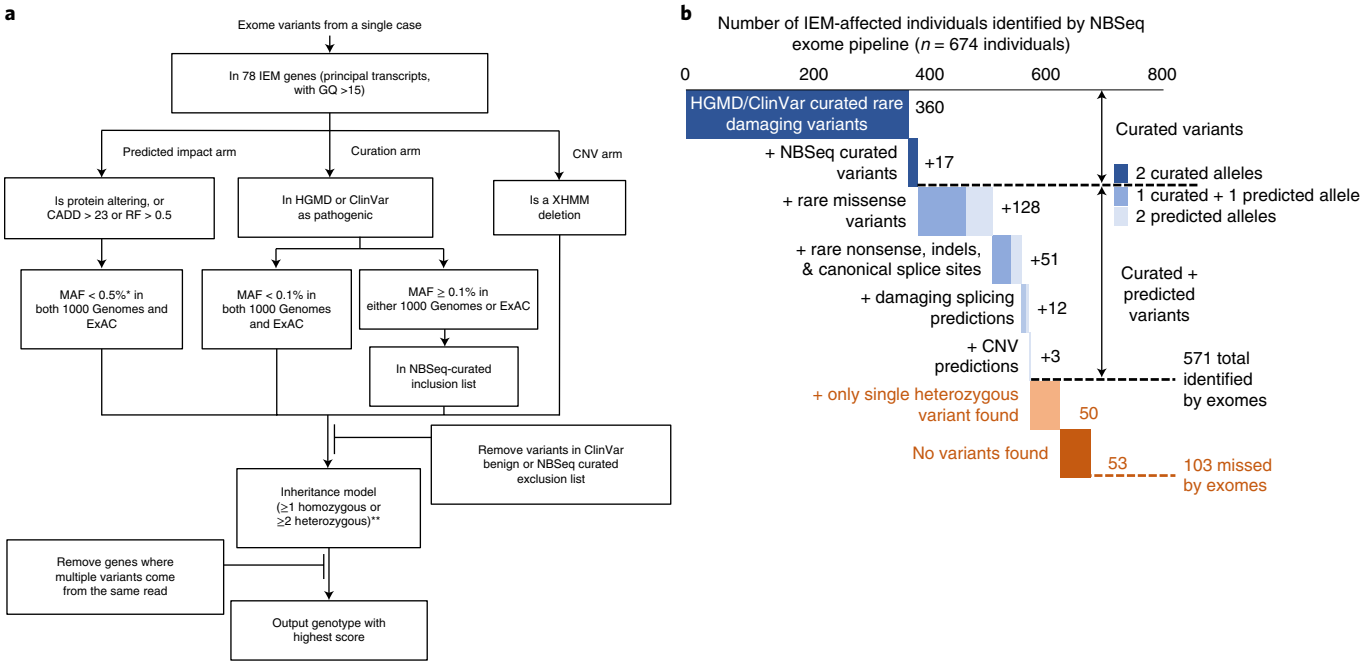


Table 2 | Performance of WES as a follow-up test after positive MS/MS for six selected IEMs, assuming an individual would not be referred for additional evaluation without at least one reportable variant identified for that IEM

Abnormal MS/MS screen result reported for	Number of MS/MS false positives	Number of exome false negatives (missed cases)	Number of exome true negatives	Specificity % (95% CI) (reduction in false positives)	NPV % (95% CI)
VLCADD	108	1	48	44.4 (34.9–54.3)	98.0 (89.1–100)
LCHADD	72	0	68	94.4 (86.4–98.5)	100.0 (94.7–100.9)
PKU	27	2	6	22.2 (8.6–42.3)	75.0 (34.9–96.8)
IVA	16	4	15	93.8 (69.8–99.8)	78.9 (54.3–93.9)
MSUD	16	1	16	100.0 (79.4–100)	94.1 (71.3–99.9)
GA-II	1	2	1	100.0 (2.5–100)	33.3 (0.8–90.6)
All of above	240	10	154	**64.2 (57.7–70.2)**	**93.9 (80.0–97.0)**

PKU, phenylketonuria. Two-sided Clopper–Pearson confidence interval (CI) was calculated using the 'exactci' function from R package PropCIs (https://github.com/shearer/PropCIs).

a definitive diagnosis by identifying pathogenic variants in IEM-associated genes. Our examples of discordance of recorded diagnoses and WES findings indicate that deep sequencing could facilitate rapid and precise clinical resolution for newborns with positive MS/MS on NBS, as previously suggested[31].

WES turnaround time and cost, which are critical concerns for NBS programs, were not addressed in our research study, in which archived batches of samples were processed together. Future WGS could be faster; diagnostic WGS for critically ill infants have ranged from 2 to 3 weeks[9] to <24 h[32,33]. The modest caseload of positive NBS screens for IEMs (0.3% of births) suggests that deep sequencing could become sufficiently economical and rapid to provide effective information as a secondary test after a positive MS/MS result. Clinical considerations for individual IEMs would dictate whether urgent referral after a positive MS/MS screen is required or whether referral could await sequencing results.

The high specificity of WES using curated variants prompts its consideration for well-characterized, treatable Mendelian disorders not amenable to MS/MS, but for which screening could be justified if a suitable DNA-based test were available. However, suitability of WES or WGS must be evaluated for each disorder. Though sensitivity of WES alone may be too low to meet standard criteria for NBS[34], sequencing could potentially identify many treatable conditions that presently go unrecognized until too late for optimal intervention due to lack of an alternative current NBS test. As a form of screening, sequencing would require weighing of benefits versus costs and societal implications, and might not be limited to the neonatal period. Our study also underscores the value of the California Biobank Program, which has provided unrivaled, diverse population representation for research to inform approaches to advance public health.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-0966-5.

Received: 23 December 2019; Accepted: 8 June 2020;
Published online: 10 August 2020

References

1. Hall, P. L. et al. Postanalytical tools improve performance of newborn screening by tandem mass spectrometry. *Genet. Med.* **16**, 889–895 (2014).
2. Mak, C. M., Lee, H. C., Chan, A. Y. & Lam, C. W. Inborn errors of metabolism and expanded newborn screening: review and update. *Crit. Rev. Clin. Lab. Sci.* **50**, 142–162 (2013).
3. McHugh, D. et al. Clinical validation of cutoff target ranges in newborn screening of metabolic disorders by tandem mass spectrometry: a worldwide collaborative project. *Genet. Med.* **13**, 230–254 (2011).
4. Wilcken, B., Wiley, V., Hammond, J. & Carpenter, K. Screening newborns for inborn errors of metabolism by tandem mass spectrometry. *N. Engl. J. Med.* **348**, 2304–2312 (2003).
5. Tang, H. et al. Damaged goods?: an empirical cohort study of blood specimens collected 12 to 23 hours after birth in newborn screening in California. *Genet. Med.* **18**, 259–264 (2016).
6. Adams, D. R. & Eng, C. M. Next-generation sequencing to diagnose suspected genetic disorders. *N. Engl. J. Med.* **379**, 1353–1362 (2018).
7. Biesecker, L. G. & Green, R. C. Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.* **371**, 1170 (2014).
8. Farnaes, L. et al. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. *NPJ Genom. Med.* **3**, 10 (2018).
9. French, C. E. et al. Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. *Intensive Care Med.* **45**, 627–636 (2019).
10. Friedman, J. M. et al. Genome-wide sequencing in acutely ill infants: genomic medicine's critical application? *Genet. Med.* **21**, 498–504 (2018).
11. Berg, J. S. et al. Newborn sequencing in genomic medicine and public health. *Pediatrics.* **139**, e20162252 (2017).
12. Regaldo, A. in *Technology Review* (2017).
13. Hoffmann, G. F. in *Inherited Metabolic Diseases:* A Clinical *Approach* (eds Hoffmann, G. F., Zschocke, J. & Nyhan, W. L.) 31–32 (Springer Berlin Heidelberg, 2017).
14. Bassaganyas, L. et al. Whole exome and whole genome sequencing with dried blood spot DNA without whole genome amplification. *Hum. Mutat.* **39**, 167–171 (2018).
15. Biesecker, L. G. Secondary findings in exome slices, virtual panels, and anticipatory sequencing. *Genet. Med.* **21**, 41–43 (2019).
16. Stenson, P. D. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
17. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
18. Feuchtbaum, L., Yang, J. & Currier, R. Follow-up status during the first 5 years of life for metabolic disorders on the federal recommended uniform screening panel. *Genet. Med.* **20**, 831–839 (2018).
19. Feuchtbaum, L., Carter, J., Dowray, S., Currier, R. J. & Lorey, F. Birth prevalence of disorders detectable through newborn screening by race/ethnicity. *Genet. Med.* **14**, 937–945 (2012).
20. Consortium, G. P. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
21. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
22. Matern, D. et al. Prospective diagnosis of 2-methylbutyryl-CoA dehydrogenase deficiency in the Hmong population by newborn screening using tandem mass spectrometry. *Pediatrics.* **112**, 74–78 (2003).
23. Tiranti, V. et al. Ethylmalonic encephalopathy is caused by mutations in ETHE1, a gene encoding a mitochondrial matrix protein. *Am. J. Hum. Genet.* **74**, 239–252 (2004).
24. Henriques, B. J. et al. Ethylmalonic encephalopathy ETHE1 R163W/R163Q mutations alter protein stability and redox properties of the iron centre. *PLoS ONE* **9**, e107157 (2014).
25. Wang, Z. Q., Chen, X. J., Murong, S. X., Wang, N. & Wu, Z. Y. Molecular analysis of 51 unrelated pedigrees with late-onset multiple acyl-CoA dehydrogenation deficiency (MADD) in southern China confirmed the most common ETFDH mutation and high carrier frequency of c.250G>A. *J. Mol. Med. (Berl.)* **89**, 569–576 (2011).

26. Goldfeder, R. L. et al. Medical implications of technical accuracy in genome sequencing. *Genome Med.* **8**, 24 (2016).
27. Sulonen, A. M. et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* **12**, R94 (2011).
28. Peng, G. et al. Combining newborn metabolic and DNA analysis for second-tier testing of methylmalonic acidemia. *Genet. Med.* **21**, 896–903 (2019).
29. Vockley, J., Rinaldo, P., Bennett, M. J., Matern, D. & Vladutiu, G. D. Synergistic heterozygosity: disease resulting from multiple partial defects in one or more metabolic pathways. *Mol. Genet. Metab.* **71**, 10–18 (2000).
30. Batshaw, M. L., Msall, M., Beaudet, A. L. & Trojak, J. Risk of serious illness in heterozygotes for ornithine transcarbamylase deficiency. *J. Pediatr.* **108**, 236–241 (1986).
31. 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).
32. Clark, M. M. et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci. Transl. Med.* **11**, eaat6177 (2019).
33. Kingsmore, S. F. et al. A randomized, controlled trial of the analytic and diagnostic performance of singleton and trio, rapid genome and exome sequencing in Ill infants. *Am. J. Hum. Genet.* **105**, 719–733 (2019).
34. Calonge, N. et al. Committee report: method for evaluating conditions nominated for population-based screening of newborns and children. *Genet. Med.* **12**, 153–159 (2010).
35. Rodriguez, J. M. et al. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* **41**, D110–D117 (2013).
36. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
37. Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.* **42**, 13534–13544 (2014).
38. Fromer, M. et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
39. Chamberlin, M. E., Ubagai, T., Mudd, S. H., Levy, H. L. & Chou, J. Y. Dominant inheritance of isolated hypermethioninemia is associated with a mutation in the human methionine adenosyltransferase 1A gene. *Am. J. Hum. Genet.* **60**, 540–546 (1997).

Methods

NBSeq samples and study set. From the roughly 4.4 million children born in California from 7 July 2005 to 31 December 2013, the NBSeq project obtained two de-identified 3.2-mm punches from DBSs obtained for NBS from each of 1,728 newborns. The samples included those from all 1,325 infants screened by MS/MS who were subsequently confirmed to have an IEM, as well samples from 9 infants not identified by MS/MS screening, but diagnosed clinically. Also included were 394 MS/MS false-positive infants selected from 10,011 individuals, for whom MS/MS screening was positive, but who were ultimately diagnosed as unaffected. All MS/MS false positives from well-baby nurseries for three FAODs (VLCADD, LCHADD and MADD) and three other IEMs (PKU, MSUD and IVA) were requested to investigate the hypothesis that genetic variants could underlie the abnormal analytes leading to false-positive MS/MS results. False positives in neonatal intensive care units, where concomitant illness (for example, prematurity and liver abnormalities) and clinical interventions (for example, total parenteral nutrition) produce frequent alterations of MS/MS analytes, were excluded from our request for DBSs.

DNA was prepared from the DBSs and WES was performed as described[14]. We sequenced 1,416 exomes, which included all the DBS specimens from the CDPH Biobank, with the exception of those from a random subset of 312 non-Hispanic white and Hispanic individuals for SCADD, 3MCC and elevated phenylalanine, which were excluded due to budgetary limitations. This led to exclusion of specimens from 56 SCADD, 54 3MCC and 201 initial hyperphenylalaninemia by MS/MS (including 61 individuals with classic PKU, 45 with variant hyperphenylalaninemia, 38 with benign hyperphenylalaninemia and 58 individuals who were false-positive). From 1,190 exomes out of the 1,416 that passed quality control metrics and were phenotypically relevant, we developed and evaluated a pipeline to identify individuals for potential review based on variants reported in an IEM gene, as described later.

To obtain an unbiased estimate of the predictive accuracy of exomes, we first divided the 1,190 samples into a validation set (178 samples) in which we were unblinded to final diagnosis to explore the robustness of parameter choices in our exome analysis pipeline (data not shown) and a test set (1,012 samples), which was subjected to the final optimized pipeline only once and whose results are reported here. Of the 178 samples in the validation set, 129 were affected with an IEM diagnosed following positive NBS by MS/MS; two were affected, but had not been identified by MS/MS NBS but were diagnosed clinically; and 47 were unaffected MS/MS false positives (Table 1). Of the 1,012 samples in the test set, 674 were from IEM-affected individuals, 667 were diagnosed following MS/MS NBS and 7 were identified clinically with disease; there were also 338 MS/MS false positives.

Variant calling and quality control. *Variant calling and annotation.* DNA extraction, library construction and exome sequencing were performed as previously described[14]. The 1,416 samples were sequenced in three batches (batch 1, 188 samples; batch 2, 411 samples; batch 3, 817 samples) and the quality control (QC) metrics were grouped accordingly. Batches 1 and 2 had paired-end read lengths of 101 bp, whereas batch 3 had a paired-end read length of 151 bp. Raw sequences were mapped to the reference genome (v.37), using BWA mem algorithm (v.0.7.10)[40]. Resulting SAM files were converted to binary format, sorted and lane merged using Picard tools (v.1.81). Duplicates were marked in the alignment files with Picard tools v.1.81 (http://broadinstitute.github.io/picard/). Next, realignment around known indels and base quality score recalibration were performed using GATK toolkit (v.3.3)[41]. Variants were called using the GATK Haplotype Caller function and variant scores were recalibrated with the GATK VQSR function[42]. Combined calling was used on all samples. Variants were annotated using Varant (http://compbio.berkeley.edu/proj/varant/Home.html), a custom tool, as described[43], with the following public datasets: Gencode (v.19)[44], APPRIS (v.24)[35], 1000 Genomes Project (phase 3)[20], Exome Sequencing Project (ESP) (ESP6500SI-V2-SSA137)[45], Exome Aggregation Consortium (ExAC v.0.3.1)[21], CADD (v.1.3)[36], MetaSVM and MetaLR from dbNSFP v.3.1a[46] and dbscSNV v.1[37]. Variants previously associated with disease from HGMD (v.2014.1)[16] and those with star ratings for known deleterious variants from ClinVar[17] (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz) were added to the call set. CNV calls on all 1,190 NBSeq samples were made using XHMM[38].

Quality control metrics. We developed a battery of quality assessment metrics to ensure that exomes from DBS samples were suitable for variant interpretation[47]. The sequences were assessed for quality of read mapping, assessment of DNA damage and quality of the final called variants.

Mapping-related metrics. The percentage of unmapped reads was <1.7% in all the samples (median = 0.4%, first quartile = 0.3%, third quartile = 0.5%; Extended Data Fig. 1a). The nonduplicate, properly oriented paired reads with mapping quality ≥20 were classified as high quality. The percentage of high-quality read pairs was quantified (median = 61%; Extended Data Fig. 1b). Batch 3 contained a larger percentage of duplicates than the other batches (Extended Data Fig. 1c), but numbers of high-quality read pairs were comparable in all batches (Extended Data Fig. 1d,e). All batches had sufficiently large insert sizes for analysis, although batch 2 had slightly lower insert sizes for the same read length (Extended Data Fig. 1f). The proportion of read pairs with each end mapped to a different chromosome was small (median = 0.004).

Median coverage depth across the capture region was 59× (first quartile, 52×; third quartile, 68×). Comparison between the three batches showed that for median capture coverage, batch 1 and batch 2 were similar, whereas batch 3 had the highest median depth of coverage, likely because of longer reads (batch 1, median = 53; batch 2, median = 56; batch 3, median = 63; Extended Data Fig. 1g). Batch 2 had the lowest fraction of capture covered for all depths and batch 3 had the highest (Extended Data Fig. 1h). This could indicate lower uniformity of coverage in batch 2 compared to the other two batches. The coverage over the 78 genes transcripts associated with IEMs was examined for the same set as the metrics above and similar trends were observed (Extended Data Fig. 1i,j).

DNA damage-related metrics. The number of mismatches with respect to the reference genome was computed for all high-quality reads as an estimate of DNA damage caused by nucleotide mis-incorporations. The fraction of reads with 0 mismatches with the reference was slightly lower (median = 0.74) than typically observed in other datasets, which tended to be >0.8. A larger number of samples in batch 3 had fewer perfectly matched reads compared to the others (batch 1, median = 0.88; batch 2, median = 0.79; batch 3, median = 0.70; Extended Data Fig. 2a). The higher mismatch rate in batch 3 was due to their longer read length, with an increased rate toward the ends; when only the first 100 bases of each read were analyzed in batch 3, mismatch data were comparable to that for the other batches (Extended Data Fig. 2b).

The distribution of nucleotide mis-incorporations by base change was investigated by calculating the allele distribution at each position and from these distributions, calculating the frequency of a base change from the reference to every other base. The fraction for a change of, for example, A>C, was computed as the ratio of the number of times an A>C change was seen to the total number of As seen. Positions with variants reported after calling the alignments were excluded, as these are presumed to be real variations. The C>A and G>T distributions were higher in batch 2 for several of the samples compared to batches 1 and 3, with C>T and G>A also slightly higher (Extended Data Fig. 2c). However, when the nucleotide change fractions were analyzed using high-quality single-nucleotide variant (SNV) calls ('PASS' by the GATK VQSR algorithm and with GQ ≥ 30) this bias was absent, with all nucleotide change fractions in similar ranges for the three batches (Extended Data Fig. 2d).

Variant-related metrics. The fraction of sites with GQ ≥ 30, both reference and variant, was calculated from the total number of sites both for the entire exome capture as well as transcripts from the 78 IEM-associated genes. The fraction of sites with marked 'PASS' and with GQ ≥ 30 was within adequate range for the three batches (Extended Data Fig. 3a,b). Next, the high-quality variants were classified based on frequency in 1000 Genomes Project (defined as common, ≥0.1%; rare <0.1%). SNVs and indels were analyzed separately. Common and rare SNVs and indels occurred at similar frequencies in the three batches (Extended Data Fig. 3). Transition/transversion ratios were in the same range for the three batches for both common and rare SNVs and comparable to the range of 1000 Genomes Project samples (Extended Data Fig. 3).

Quality control criteria for sample exclusion. The samples that passed QC were required to satisfy the following quality criteria:

- Median coverage across Nimblegen capture ≥20×.
- Fraction of high-quality sites (GQ ≥ 30) ≥0.9 across the capture.
- Estimated contamination by VerifyBamID[48] (freemix) <3%.

Overall, 200 samples out of 1,416 failed at least one of the above criteria. Some areas of systematic decreased coverage were noted, as with the relatively poor coverage of exon 1 of the *MCCC2* gene, compared to *ACADM*, well covered throughout (Extended Data Fig. 4). Of 1,216 exomes that passed QC, 26 from infants with an IEM not within the core panel of 48 disorders were further excluded. The remaining 1,190 samples were analyzed in our study.

Exome analysis pipeline. Variant interpretation was guided by the principle that in current NBS, specificity is subordinate to maximal sensitivity because missed cases can have catastrophic consequences, whereas false positives can be resolved as unaffected following referral. Thus, like some NBS programs reporting DNA sequence information, our pipeline deliberately reported variants of uncertain significance (American College of Medical Genetics and Genomics guidelines)[49].

Pipeline development. Our goal was to evaluate exome sequencing as an NBS tool. In contrast to sequencing for diagnosis of individual patients already exhibiting clinical abnormal phenotypes, sequencing in a screening setting presents different challenges and requirements, some of which include:

- NBS has no a priori phenotypic information and is applied on a population-level to individuals, who are nearly all asymptomatic at the time of the screen.
- While it is accepted that diagnostic exomes will resolve only 25–60% of undiagnosed cases (depending on the types of disorders), much higher sensitivity is required for NBS. For example, MS/MS NBS for IEMs is typically >99% sensitive.

- Specificity also has to be high in NBS to avoid excessive cost, referrals to specialists and family anxiety.
- In diagnostic settings, DNA sequence analysis may take into account family history, sequence of family members in addition to the patient and customized, in-depth curation of variants; but NBS must be performed on single individuals in a high-volume, largely automated mode with limited manual review.

Given these differences, it was not obvious whether a diagnostic exome pipeline was appropriate for screening or how it should be modified. We therefore developed exome analysis pipelines taking NBS requirements in consideration (Extended Data Fig. 5). It was important to understand how altering some commonly used parameters in pipelines would influence the sensitivities and specificities of disease predictions from exome data. However, we lacked sufficient data to train a sophisticated model to learn these parameters automatically. Therefore, with the goal of identifying robust parameter choices, we devised a framework to iteratively perturb and tune pipeline parameters.

Systematic exploration of parameter perturbations formed the basis of the final exome analysis pipeline[50]. Briefly, using the 178-sample NBSeq validation set and starting from two initial reference pipelines that used parameter combinations favoring either high sensitivity or high specificity, we systematically explored the consequences of perturbing the following parameters on both sensitivity and specificity: size of IEM gene list, variant callers, choices of transcript models, choice of population databases, MAF thresholds in population databases, databases of predicted and curated variants and choice of inheritance models. We studied the impact of each parameter on overall performance by altering a single or a few parameters at a time. As an example, prediction results were surprisingly sensitive to the choice of population database to determine MAF. We studied the impact of MAF thresholds of 0.1%, 0.2%, 0.5%, 1%, 2% or 5% in three different databases: 1000 Genomes Project, ExAC and the NHLBI ESP. MAF thresholds using 1000 Genomes Project or ExAC yielded better sensitivities than in the ESP database alone, possibly due to technical artifacts from differing sequencing technologies and study designs used in the different databases. Informed by these observations, our final pipeline implemented population MAF thresholds that considered both the ExAC and 1000 Genomes Project databases. The impact of individual parameter choices on the 178-sample NBSeq validation set, evaluated in an iterative fashion, guided the design and parameter choices for the final pipeline (Fig. 2a) that was applied to the 1,012 samples (674 IEM-affect and 338 MS/MS false positives) in the NBSeq test set.

Architecture of the final exome analysis pipeline. Once finalized based on the NBSeq validation set, the automated exome analysis pipeline (Fig. 2a; code available at https://github.com/nbseq1200/NBSeq1200paper) was run once on each sample in the NBSeq test set while blinded to any phenotypic data, reflecting a typical application setting for a primary newborn screen. The analysis was restricted to the variants in the exon slice of the 78 known IEM-associated genes (Supplementary Table 1). The pipeline considered only variants with GQ > 15 and the impact of variants was annotated with respect to APPRIS principal transcripts. The pipeline reported variants through one of three primary arms:

- Predicted impact arm: all variants with MAF < 0.5% (in both ExAC and 1000 Genomes Project) that satisfied any of the following criteria:
- annotated as protein-altering by Varant (StopGain, StopLoss, FrameShiftInsert, FrameShiftDelete, SpliceDonor, SpliceAcceptor, NonSynonymous Missense, InFrameDelete, InFrameInsert, StartGain, StartLoss)
- CADD[36] score >23
- computational splicing-effect meta-prediction tool (dbscSNV)[37] score >0.5
- Curation arm: variants with population autosomal MAF < 0.1% (in both ExAC and 1000 Genomes Project databases) annotated as 'DM' or 'DM?' in HGMD or as 'pathogenic/likely pathogenic' by ClinVar with at least one review star. Our NBSeq team manually curated 60 variants with MAF ≥ 0.1% from these sources, finding 19 to be reportable as potentially pathogenic and excluding 41 (Supplementary Table 2). The NBSeq variant curation is described in detail below.
- CNV arm: XHMM[38], a tool to call CNVs from exomes, was run with default parameters on the full-sample BAM files. XHMM calls from three genes (*ETFA*, *HCFC1* and *PRODH*) were excluded because XHMM called CNVs with a frequency >1% in these genes in 600 exomes extracted from the 1000 Genomes Project. As XHMM does not output zygosity, we inspected the BAM files in the predicted deletion regions to make zygosity calls, the only manual step in the pipeline.

The variants reported through any of the above arms were considered further by the pipeline. For variants in X chromosome genes, the MAF threshold was adjusted (see below). To avoid spurious multiple, close-by variant calls when encountering a single deletion or insertion event, the pipeline reported only a single variant if multiple heterozygous variants appeared within 15 bases of each other. For heterozygous variants within 500 bases of each other, a local phasing procedure queried the reads overlapping both variant positions. If both variants appeared in the same read, only one was reported. Finally, any variants that were annotated as benign in ClinVar with at least two review stars were excluded.

From the remaining variants, the pipeline identified the corresponding genes with ≥1 homozygous or ≥2 heterozygous variants, as the majority of these disorders are autosomal recessive. For the X-linked *OTC* gene, the pipeline reported ≥1 variant in either hemizygous males or heterozygous females, as heterozygous females can display a clinical phenotype[30,51]. Another exception was *MAT1A*, for which the pipeline reported a particular heterozygous variant NP_000420.1:p.Arg264His known to cause autosomal dominant disease[39]. Finally, for each sample, the pipeline reported at most a single prominent genotype based on the highest score (see below) combining disease prevalence and variant severity.

Minor allele frequency threshold adjustment for X-linked genes. In the exome analysis pipeline, the MAF for genes encoded on the X chromosome was adjusted as follows. Given the MAF threshold chosen in the pipeline for the autosomal chromosomes, f_A, the X chromosome MAF is adjusted to f_X, as given by the following relationship:

$$\frac{1}{2}f_X^2 + \frac{1}{2}f_X = f_A^2$$

When $f_X \ll 1$, we can approximate
$f_X \cong 2f_A^2$

NBSeq variant curation. The exome analysis pipeline excluded nonrare variants unless existing evidence suggested possible pathogenicity. In 31 IEM-associated genes (*ACAD8*, *ACADM*, *ACADS*, *ACADSB*, *ACADVL*, *ASL*, *ASS1*, *BCKDHA*, *CBS*, *CD320*, *CPT1A*, *CPT2*, *DBT*, *FAH*, *GCH1*, *HADHA*, *HPD*, *MCCC2*, *MCEE*, *MLYCD*, *MMAB*, *MTRR*, *MUT*, *OTC*, *PAH*, *PCCA*, *PCCB*, *PRODH*, *SLC22A5*, *SLC25A13*, *TAZ*), 60 variants with MAF > 0.1% in 1000 Genomes Project and ExAC databases were characterized as DM or DM? in HGMD and/or characterized as pathogenic or likely pathogenic in ClinVar. All evidence supporting the classifications was reviewed by the NBSeq team, which determined that missense variants should be considered potentially pathogenic, to beconsistent with the requirements of a screening test, including that specificity was generally subordinate to sensitivity. Variants were therefore considered reportable if there was any evidence that they could be disease-causing, even if at low penetrance. After NBSeq curation, 19 of the 60 variants reviewed were deemed 'reportable' and 41 were deemed 'not reportable' (Supplementary Table 2). The list of reportable variants were incorporated into the primary exome analysis pipeline, as the NBSeq inclusion list.

As an example, the NM_000098.3:p.Phe352Cys variant in *CPT2* is polymorphic, with MAF 4.7% in 1000 Genomes Project and 2.2% in ExAC, but was indicated 'DM?' in HGMD, while ClinVar status in the most recent version was likely benign, but with no stars. Literature review supported the assertions that this was a polymorphism, but not a totally benign one; published data suggested it was a risk factor for acute encephalopathy in the setting of serious viral illness in infancy[52]. In contrast, NM_000137.2:p.Arg341Trp in *FAH* was on the border of being polymorphic, with MAF 0.8% in 1000 Genomes Project and 1.7% in ExAC. This variant was classified as 'DM' in HGMD on the basis of literature[53] that, when carefully reviewed, did not support the assertion, while functional expression in yeast and cultured cells[54] showed the protein carrying the missense variant had activity equal to that of wild-type protein. This variant, therefore, was designated as 'not reportable.'

Selection of the most prominent genotype in a sample. A scoring scheme was designed to select one gene at most as the predicted cause of disease for each sample. For each reported variant allele a in a particular gene g for the sample, a variant score $S(a,g)$ was obtained as follows:

$$S(a,g) = \begin{cases} \frac{1}{3}*(M(a)+I(a)+D(a)) + P(g), & a \text{ is not a CNV} \\ 1 + P(g), & a \text{ is a CNV} \end{cases}$$

where $P(g)$ = unit normalized prevalence of the disorder corresponding to gene g (Supplementary Table 1)

$$M(a) = \begin{cases} 1, & \text{MAF(a)} < 0.05\% \\ 0.5, & 0.05\% \leq \text{MAF(a)} < 0.1\% \\ 0, & \text{otherwise} \end{cases}$$

$$D(a) = \begin{cases} 1, & \text{Curation(a)} \in \{\text{HGMD(DM)}, \text{Clinvar 4(aaa)}, \text{Clinvar 5(aaaa)}\} \\ 0.5, & \text{Curation(a)} \in \{\text{HGMD(DM)}\} \\ 0, & \text{otherwise} \end{cases}$$

$$I(a) = \begin{cases} 1, & \text{Consequence(a)} \in \text{SevereSet} \\ 0.5, & \text{Consequence(a)} \in \text{MildSet} \\ 0, & \text{otherwise} \end{cases}$$

SevereSet = {StopGain, StopLoss, FrameShiftInsert, FrameShiftDelete, SpliceDonor, SpliceAcceptor}

MildSet = {Missense, InFrameDelete, InFrameInsert, StartGain, StartLoss}.

MAF(a) was the maximum MAF of the allele a in 1000 Genomes Project and ExAC databases. Curation(a) determined the presence/absence of the allele

a in HGMD and ClinVar databases. Consequence(*a*) determined the impact of the variant allele. The *P(g)* values for the 78 genes were derived from the birth prevalence of the corresponding disorders in California[19].

For each gene that had either ≥1 homozygous or ≥2 heterozygous autosomal variants reported by the pipeline in a sample, the sum of the two highest scoring alleles in that gene was computed. Homozygous variants contributed two alleles and only one variant was required for X-linked genes. The gene with the highest sum for the sample was reported as the predicted disease gene for that sample.

Assessment of predictions from the exome analysis pipeline. The exome analysis pipeline reported 507 variants in genes associated with corresponding disorders in IEM-affected individuals in the NBSeq test set (Supplementary Table 5). The majority of these variants were nonsynonymous (Extended Data Fig. 6a). To ensure an objective and unbiased interpretation of the results, the exome analysis pipeline developers (A.N.A., Y.W. and S.E.B.) were blinded to the diagnoses of the individual cases in the NBSeq test set. The exome analysis pipeline was run on the NBSeq test set exomes and the resulting predictions from the exome analysis pipeline were submitted to the NBSeq clinical team (R.J.C., R.C.G., H.T. and M.K.), who then assessed those predictions independent of the pipeline development team.

The data used for assessment are shown in Supplementary Table 3, listing for each case the reviewed diagnosis key (deemed the gold standard) and the results reported by the primary exome analysis pipeline, as well as alternative pipelines. Numerical aspects of assessment were computed using software available at https://github.com/nbseq1200/NBSeq1200paper.

Incorporation of clinical evidence to generate a 'reviewed' diagnosis key. As part of the follow-up of positive NBS results, the GDSP database has recorded summaries of diagnostic testing and annual patient summaries for infants with IEMs. To assure that the exome predictions would be measured against the best current clinical disease assessments, for those cases for which the exome pipeline was discordant with the GDSP diagnosis, the clinical team generated a 'reviewed diagnosis' key for all NBSeq study cases based on the strength of evidence for the clinical diagnoses recorded in the GDSP database and recent literature. This reviewed diagnosis key was used to compare infant diagnoses with exome predictions.

There were three main types of discrepancies between exome predictions and clinical reports. The first group had a mismatch between the predicted gene by the pipeline and the gene implicated in the clinical diagnosis of an affected infant ($n=22$). The second group had gene predictions in MS/MS false-positive individuals, who had been initially reported as unaffected ($n=45$). The third group had a failure to predict a gene in an initially diagnosed infant ($n=188$). Members of NBSeq with access to the GDSP database (R.J.C. and H.T.) reviewed recorded clinical notes with a metabolic geneticist (R.C.G.) to assess explanations for the discrepancies.

The diagnoses of 12 of 22 infants in the first group (different exome identification versus clinical record) were resolved with establishment of a 'reviewed diagnosis.' In six infants, there may have been an error in selecting the disorder from the drop-down list for diagnosis in the online case report. In the remaining six, the exome identified the gene for an alternative in the differential diagnosis of the screening result, and the follow-up data were most consistent with that alternative. For evaluation of exome predictions, these 12 infants were re-classified as correctly identified by exome.

In the second group (exome positive, IEM negative), 26 of the 45 infants had been resolved as false-positive for PKU, but had two rare variants in the *PAH* gene. After levels of amino acids from follow-up testing were reviewed, 10 of the 26 diagnoses were re-classified as hyperphenylalaninemia. Almost all of the remaining infants in the exome-positive group were either clinically identified as VLCADD(het) or MS/MS false-positive for VLCADD, with one remaining unresolved due to death of the infant. In most of these cases, the exome revealed two variants in the *ACADVL* gene; in three, the exome identified variants in genes for MADD.

In a third group (exome-negative, IEM-positive) 78 infants were re-classified as unaffected. This included 40 individuals clinically identified as SCADD. The NBSeq curation team had reviewed and excluded from the exome pipeline the common variant in the *ACADS* gene (rs57443665, Supplementary Table 2), but infants with this variant often had been given an initial clinical diagnosis of SCADD [55–57]. These were re-classified as not affected. There were 23 individuals with carnitine uptake defect and 15 with VLCADD in this third group. Long-term follow-up data suggested that the majority of the infants were unaffected and the classification was updated to reflect this. The remainder of this group consisted of small numbers of each of the other metabolic disorders included in the study. For the majority of them, the follow-up data provided ample evidence of the correctness of the clinical diagnosis.

As a final step in the clinical review, the individuals for which the documentation was sufficient to be confident of the clinical diagnosis were designated as a 'clinically confident' subset. All diagnoses of SCADD, whether predicted by the exome or not, were excluded from the clinically confident subset due to published data indicating that the most commonly encountered variants may lead to biochemical abnormality on MS/MS screening, but are not associated with clinical disease[57].

Assessment of exome sequencing as a primary screen. Overall sensitivity and specificity. For evaluation of exomes as a screen for all NBS IEMs combined, we matched all test set predictions from the NBSeq exome analysis pipeline to their associated clinical diagnoses in the reviewed diagnosis key and calculated two main metrics: the overall sensitivity and overall specificity.

IEM-affected diagnoses, for which the gene predicted by the exome analysis pipeline matched the IEM in the reviewed diagnosis key, were considered correct. IEM-affected diagnoses, where no gene was predicted by the pipeline, were considered incorrect, as were instances where the pipeline predicted a gene inconsistent with the clinical diagnosis in the reviewed diagnosis key.

For each IEM, sensitivity was calculated as the number of infants in the test set correctly identified by the exome analysis pipeline divided by the total number of IEM-affected individuals in the test set. As some individuals with particular IEMs had been excluded from sequencing, the IEM-specific sensitivities were weighted by their prevalence in California[19] to obtain the overall sensitivity of the final exome analysis pipeline across all IEMs (values in Supplementary Table 1, those <0.04 per 10,000 were treated as 0.04). Among the 674 affected individuals in the test set, 571 were correctly identified by the exome pipeline (Extended Data Fig. 6b). The overall sensitivity of the final exome analysis pipeline when weighted by the disorder prevalence was 88%.

To calculate overall specificity, we considered situations where the pipeline predicted genes that did not match the IEM diagnosis according to the reviewed diagnosis key. For each IEM *d*, we calculated the disorder-level specificity as one minus the proportion of individuals not affected with *d*, where the exome analysis pipeline reported a gene associated with *d*. (Of note none of the false-positive cases were reported for a disorder on the same CDPH screen as the actual disease.) Using this strategy, overall specificity was 98.4%.

CDPH determines the sensitivity of MS/MS screening by evaluating whether an individual that screens positive is reported as diagnosed with any IEM in their screening panel (without regard to whether the diagnosed IEM is in the primary differential diagnosis of the screen). By measuring sensitivity using this approach for comparison with MS/MS performance, the primary exome analysis pipeline has a sensitivity of 90%.

Performance on data subsets. As the NBSeq study set included a large group of heterogeneous IEMs, we performed further ancillary assessments on various relevant slices of the data. The sensitivity and specificity of exome predictions across different IEMs were heterogeneous (Fig. 2c and Supplementary Table 4). The distribution of zygosity of the reportable variants also varied across different IEMs (Extended Data Fig. 10).

Some disorders were too rare to have sufficient samples to make a statistically robust estimation of sensitivity, reflected in their large 95% Clopper–Pearson CI ranges (Supplementary Table 4). Two-sided Clopper–Pearson CIs were calculated using the 'exactci' function from R package PropCIs (https://github.com/shearer/PropCIs). When grouping by conditions in the national Recommended Uniform Screening Panel, conditions (core conditions) versus secondary conditions, overall sensitivities were comparable (89% for core versus 88% for secondary conditions). We further classified the IEMs into four tranches based on their population prevalence in California[19]. The weighted sensitivity of the most common IEMs (>2.5 per 10,000) was 91%, whereas that of the rarest (<0.04 per 10,000) was 78%, indicating that the more common IEMs could be identified better from sequences than rarer ones. Finally, in the 'clinically confident' subset that included only individuals with sufficient clinical follow-up data for a highly confident clinical diagnosis, the overall sensitivity of the exome analysis predictions improved to 93.7%, but was still insufficient for use as a general primary screen across all IEMs.

Alternative specificity calculation. We also estimated the overall specificity of exome predictions using two additional, orthogonal approaches. The first approach used all 1,216 sequenced individuals who passed QC filters and compared exome predictions to MS/MS results to estimate specificity. Briefly, across all 1,216 individuals, from their raw MS/MS data, we used the CLIR analysis tools[1] to infer likelihood of each IEM. For each MS/MS screen, the CLIR tool produced a 4-point score for likelihood of a particular IEM: 1, negative; 2, possible condition; 3, probable condition; and 4, very probable condition. We therefore estimated exome false-positive calls by counting the number of cases where the primary exome analysis pipeline predicted a gene, but the associated CLIR likelihood score was '1' (Supplementary Table 7). The overall specificity of exome pipeline using this strategy was 97.45%.

A second approach involved running the exome analysis pipeline on 2,504 individuals from the 1000 Genomes Project phase 3 (ref. [20]). These were not directly comparable to our NBSeq data, primarily due to differences in sequencing technology (genomes versus exomes) and underlying cohort ancestry distribution. Nonetheless, the 1000 Genomes Project dataset represents a large cohort of individuals for whom the likelihood of newborn IEMs should be low. The whole-genome genotypes were obtained from: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ and annotated using Varant[43]. The specificity of the primary pipeline on the 1000 Genomes Project dataset was 95%, slightly lower than that compared to the value calculated based on NBSeq data. Estimates of overall specificity of exome predictions from these two approaches supported the

conclusion that our primary exome pipeline had insufficient specificity to be used alone as a primary screen.

Alternative pipelines and findings. To investigate the contributions of different components of our final exome analysis pipeline (Fig. 2a), we evaluated overall sensitivity and overall specificity of various alternate pipelines that either truncated or modified the pipeline in different ways (Extended Data Fig. 5b–i). Truncating the CNV arm altogether removed only three true-positive individuals and the overall sensitivity was 88% (Extended Data Fig. 5c). Truncation of the curation arm reduced the sensitivity to 87% (Extended Data Fig. 5d). More drastically, the truncation of the prediction arm, which considered all rare protein-altering variants, resulted in an overall sensitivity of 59% (Extended Data Fig. 5e). An alternative pipeline that used the predicted impact arm alone had a sensitivity of 86% (Extended Data Fig. 5f). A stringent pipeline that reported only rare ClinGen or HGMD high-confidence curated variants in the IEM genes had a far lower sensitivity of 55%, but had an extremely high specificity of 99.4%, the latter comparable to MS/MS (Extended Data Fig. 5h). Finally, even though the primary exome analysis pipeline allowed only a single gene call per sample, we evaluated an alternate pipeline that removed this restriction, allowing for multiple gene calls per sample if more than one gene contained two reportable variants. This pipeline had an overall sensitivity of 88.6% and reduced specificity of 94.7% (Extended Data Fig. 5i).

Assessment of exome sequencing as a follow-up test. In addition, using the final pipeline results, we asked whether the exome slice could be used as a follow-up test to MS/MS, particularly to reduce false positives. That is, we assessed the ability of exome data to identify IEM-negative cases among MS/MS-positive cases (MS/MS false positives) without eliminating IEM positives. Hence, we determined the specificity (proportion of exome slice negatives among IEM negatives), which indicated how many true negatives would be removed from further consideration by eliminating cases with no gene variants on exome analysis after a positive MS/MS test. We also determined the NPV, the proportion of exome negatives among the MS/MS positives that were IEM negative (unaffected). The NPV must be very close to 1 to ensure that true IEM cases are not excluded, whereas the specificity should be high but not necessarily close to 1, as it simply indicates the effectiveness of the exome as a secondary screen. We defined 'exome negative' as individuals where no reportable variants were reported in genes for the disorder(s) corresponding to the MS/MS positive result.

For six MS/MS screened disorders for which we had originally requested all the MS/MS true positives and MS/MS false positives from the study period (VLCADD, PKU, LCHADD/TFP, IVA, MSUD and GA-II), we first quantified the number of alleles characterized by our pipeline as pathogenic for both IEM-affected and MS/MS false-positive individuals (Extended Data Fig. 9). Using the above exome-negative definition, we then calculated the specificity and NPV for each of the six MS/MS screens (Table 2).

Follow-up exploratory studies of selected exome false-negative individuals. In half of the cases for which exomes failed to predict the corresponding IEMs (exome false negative), the screening pipeline reported no variants, nor were any found upon manual review. An example was an individual affected with GA-I, for whom all variants in the associated *GCDH* gene were either common or in a deep intronic region. Exploration of the genetic data in a diagnostic or research context with experimental functional studies could be used to systematically explore why exome analysis failed to predict IEMs in 12% of the IEM-affected individuals. In a selected few such exome false-negative individuals, we performed follow-up studies to investigate potential causes for exome false negatives.

Whole-genome pilot study of eight exome false negatives. WGS was performed on samples from eight individuals from the NBSeq validation set, where the pipeline had failed to predict associated genes. The eight WGS samples were sequenced and jointly variant-called using the same protocol as described above for NBSeq exomes. As WGS typically provides more uniform coverage compared to exomes, we reanalyzed the exome slice of these eight individuals from their WGS data using the same primary analysis pipeline described above. The predictions from the exome analysis pipeline remained unchanged for these eight cases.

Using the IGV browser[58], we further manually inspected the read alignment, mapping quality and overall coverage in the associated genes in sample BAM files in every exome false negative case in the validation set. No anomalies in coverages were identified except in two of the eight individuals, where we noted poor coverage of an IEM gene, *IVD*. This led us to suspect large deletions, which were confirmed upon investigation of the genomic region around *IVD* in the WGS alignment files in these two individuals (Extended Data Fig. 7). The first infant had almost no coverage in the region spanning the first three exons of *IVD*. The second infant had almost no coverage of exon 12 of *IVD* along with low coverage across the whole gene. Both samples strongly indicated the presence of large deletions in the gene. To support this further, we searched for split reads (reported by the BWA mem algorithm) from the sample BAM files that also overlapped the *IVD* gene region. The first individual had 11 such split reads spanning the deleted region confirming the deletion event. We did not find such split spanning reads in the second infant.

Functional assay of an intronic variant in an exome false negative. IEM-affected individuals, where the pipeline reported only a heterozygous variant, could suggest a second variant in the relevant gene that the pipeline may have failed to interpret correctly. In an illustrative case of an MCADD-affected individual, the final analysis pipeline reported a heterozygous pathogenic variant in the *ACADM* gene (Y67H). A second intronic variant 14 bases from the splice site (NM_000016.4:c.388-14A>G) was observed but not reported by the pipeline (Extended Data Fig. 8a). (Note that the position −14 is just outside the range considered by the pipeline's splice impact predictor.) Given the proximity of the variant to the splice site, it was hypothesized that it could impact splicing. One possibility considered was that it could be a branchpoint A mutation; however, the branchpoint prediction tool SVM-BPfinder[59], which did predict six other As in the 50-bp region 5' to the exon, did not identify this site. Given the confidence of the clinical diagnosis and the proximity of the second variant to the splice site, we performed an experiment to determine the variant's impact on splicing of the *ACADM* gene.

To determine whether NM_000016.4:c388-14A>G influenced splicing of the ACADM pre-mRNA we generated a heterologous splicing reporter using the human *HBB* gene as a model as described[60]. Sequences corresponding to ACADM exon 6 and flanking introns were cloned into *HBB* intron 1. A mutant construct containing the A>G mutation at position −14 relative to the 3' splice site of intron 5 was also generated (Extended Data Fig. 8b). The following primer sequences were used:

ACADM exon 6 forward: 5' ccaatagaaactgggcatatgattagtgcatttcactatagtaga 3'; ACADM exon 6 reverse: 5' tgtctccacatgcccagatctattatgatactttcttggca 3'; ACADM mutant reverse: 5' cgaagaaaactgacattaaa 3'; ACADM mutant forward: 5' tttaatgtcagttttcttcg 3'; HBB exon 1 forward: 5' gcaacctcaaacagacacca 3'; and HBB exon 2 reverse: 5' agcttgtcacagtgcagctc 3'.

Plasmid DNA containing wild-type or mutant reporter was transfected into HEK293 cells and spliced reporter transcripts were analyzed by reverse transcription PCR (Extended Data Fig. 8c). As expected, the wild-type reporter construct exhibited constitutive inclusion of ACADM exon 6 in the HBB transcript. By contrast, the −14A>G mutation induced a new isoform with reduced electrophoretic mobility, suggesting activation of a cryptic 3' splice site, confirmed upon sequencing the amplicons derived from the wild-type or mutant reporter transcripts (Extended Data Fig. 8d). Assuming that the −14A>G mutation also activated a cryptic 3' splice site in the endogenous *ACADM* pre-mRNA, the 13 nucleotide extension of exon 6 in the *ACADM* mRNA would induce a premature termination codon and result in nonsense mediated decay of the aberrant message (Extended Data Fig. 8e). Taken together, our data indicate that the −14A>G mutation is sufficient to induce aberrant splicing of *ACADM* intron 5 and results in a defective mRNA isoform.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The de-identified residual DBS from the California Biobank for this project (SIS request number 496) were obtained with a waiver of consent from the Committee for the Protection of Human Subjects of the State of California, under project no. 14-07-1650 and in compliance with CDPH Biospecimen/Data Use and Confidentiality Agreement. California blood specimens and any data derived from the newborn screening program are confidential and subject to strict administrative, physical and technical protections. California law precludes any researcher from sharing blood specimens or uploading individual data derived from these blood specimens into any genomic data repository. Researchers desiring access to these data would need to make a separate application to the CPDH. Data in Fig. 2b,c and Extended Data Figs. 6, 9 and 10 can be found in Supplementary Table 3.

Code availability

Variant calling and annotation for the exome sequences were performed using previously published methods as described above. The code used for the screening analysis of exome data and subsequent assessments are deposited in GitHub (https://github.com/nbseq1200/NBSeq1200paper).

References

40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. 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**, 11.10.1–11.10.33 (2013).
42. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
43. Punwani, D. et al. Multisystem anomalies in severe combined immunodeficiency with mutant BCL11B. *N. Engl. J. Med.* **375**, 2165–2176 (2016).
44. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
45. Tabor, H. K. et al. Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. *Am. J. Hum. Genet.* **95**, 183–193 (2014).

46. Jian, X. & Liu, X. In silico prediction of deleteriousness for nonsynonymous and splice-altering single nucleotide variants in the human genome. *Methods Mol. Biol.* **1498**, 191–197 (2017).
47. Sunderam, U., et al. DNA from dried blood spots yields high quality sequences for exome analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.19.105304 (2020).
48. Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).
49. 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**, 405–424 (2015).
50. Wang, Y. et al. Perturbation robustness analyses reveal important parameters in variant interpretation pipelines. Preprint at *bioRxiv* https://doi.org/10.1101/2020.06.29.173815 (2020).
51. Yorifuji, T. et al. X-inactivation pattern in the liver of a manifesting female with ornithine transcarbamylase (OTC) deficiency. *Clin. Genet.* **54**, 349–353 (1998).
52. Hu, J. et al. Association of CPT II gene with risk of acute encephalitis in Chinese children. *Pediatr. Infect. Dis. J.* **33**, 1077–1082 (2014).
53. Bell, C. J. et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3**, 65ra64 (2011).
54. Bergeron, A., D'Astous, M., Timm, D. E. & Tanguay, R. M. Structural and functional analysis of missense mutations in fumarylacetoacetate hydrolase, the gene deficient in hereditary tyrosinemia type 1. *J. Biol. Chem.* **276**, 15225–15231 (2001).
55. Gallant, N. M. et al. Biochemical, molecular, and clinical characteristics of children with short chain acyl-CoA dehydrogenase deficiency detected by newborn screening in California. *Mol. Genet. Metab.* **106**, 55–61 (2012).
56. Jethva, R., Bennett, M. J. & Vockley, J. Short-chain acyl-coenzyme A dehydrogenase deficiency. *Mol. Genet. Metab.* **95**, 195–200 (2008).
57. Wolfe, L., et al. Short-chain acyl-CoA dehydrogenase deficiency. *GeneReviews* https://www.ncbi.nlm.nih.gov/books/NBK63582/ (2018).
58. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
59. Corvelo, A., Hallegger, M., Smith, C. W. & Eyras, E. Genome-wide association between branch point properties and alternative splicing. *PLoS Comput. Biol.* **6**, e1001016 (2010).
60. Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D. N. & Sanford, J. R. Loss of exon identity is a common mechanism of human inherited disease. *Genome Res.* **21**, 1563–1571 (2011).

Acknowledgements

The authors are grateful for expert technical and computational assistance from many diligent contributors, including W. Chan, J.-M. Chandonia, A. Chellappan, N. Dabbiru, B. Dispensa, A. Neumann, A. Nguyen, A. Rao, S. Rana and Z.-Y. Wu. The work was funded by the National Institutes of Health grant U19HD077627 as part of the NSIGHT project, a joint program between the National Human Genome Research Institute and the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health. This work was also supported by a research agreement with Tata Consultancy Services. The biospecimens and/or data used in this study were obtained from the California Biobank Program (SIS request no. 496). The CPDH is not responsible for the results or conclusions drawn by the authors of this publication.

Author contributions

R.J.C., R.L.N, B.A.K., P.-Y.K., N.R., J.M.P. and S.E.B. conceived and designed the study. A.N.A., R.J.C., G.A., L.B., F.C., M.K., S.S.R., J.S., U.S., H.T., D.V., P.-Y.K. and J.M.P. acquired data. A.N.A., R.C.G., Y.W., R.J.C., G.A., K.K., M.K., S.R., R.S., U.S., Y.Z., N.R., J.M.P. and S.E.B. analyzed data. A.N.A., R.C.G., Y.W., R.J.C., M.K., S.S.R., J.T.S, R.S., H.T., N.R., J.M.P. and S.E.B interpreted data. A.N.A., Y.W. and U.S. created software. A.N.A. wrote the first draft of the manuscript. R.C.G., Y.W., R.J.C., F.C., M.K., S.D.M., R.L.N., J.T.S., R.S., H.T., B.A.K., P.-Y.K., N.R., J.M.P. and S.E.B. provided critical revisions. All authors approved the final version of the manuscript.

Competing interests

A.A. is currently an employee of Illumina, Inc. K.K. was an employee of Tata Consultancy Services (TCS); U.S. and R.S. are employees of TCS. Y.Z. is currently an employee of Yikon Genomics Co., Ltd. R.N. is an employee of Invitae. J.P. is the spouse of R. Nussbaum, an employee of Invitae. S.E.B. receives support at the University of California Berkeley through a research agreement from TCS.

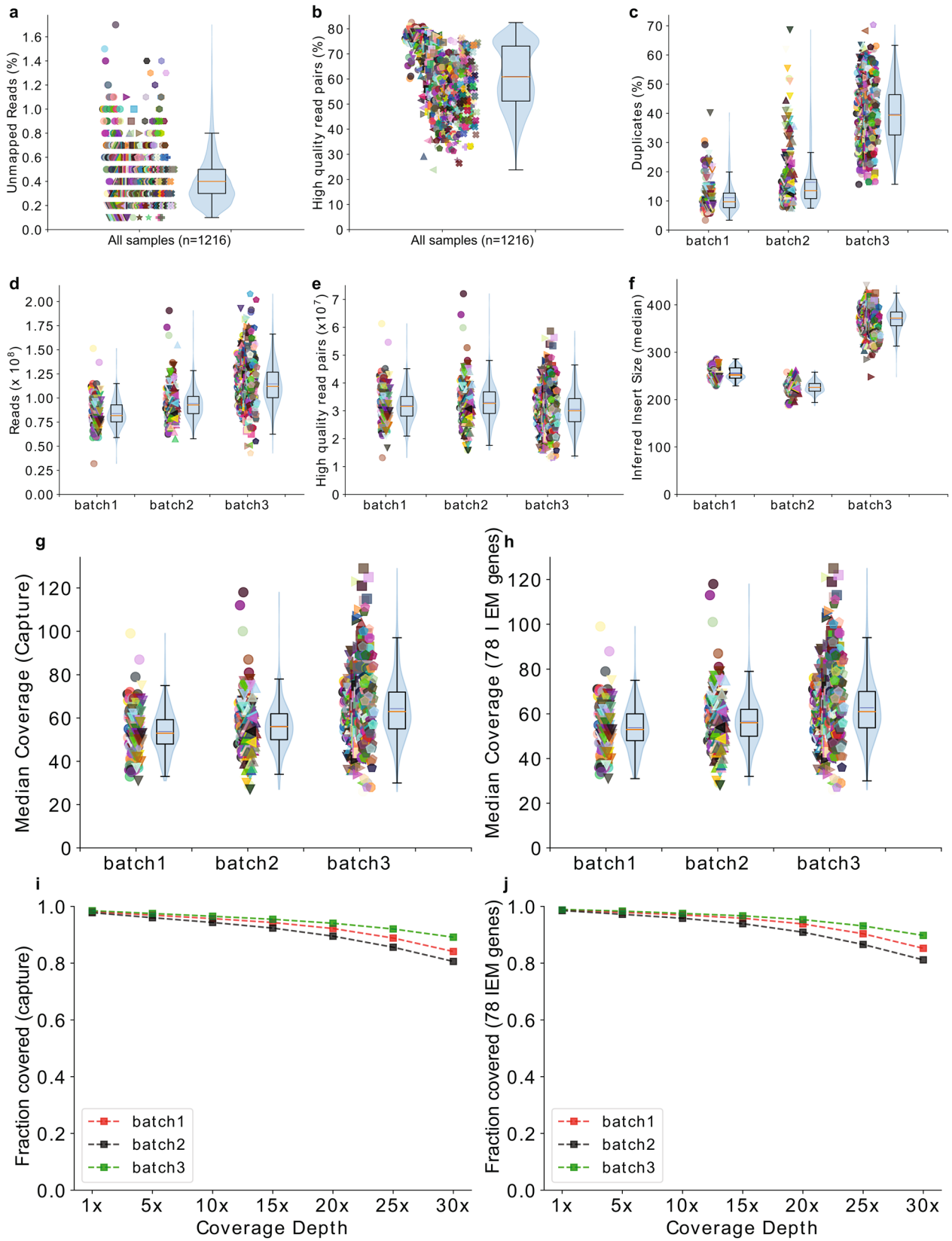
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-020-0966-5.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-0966-5.

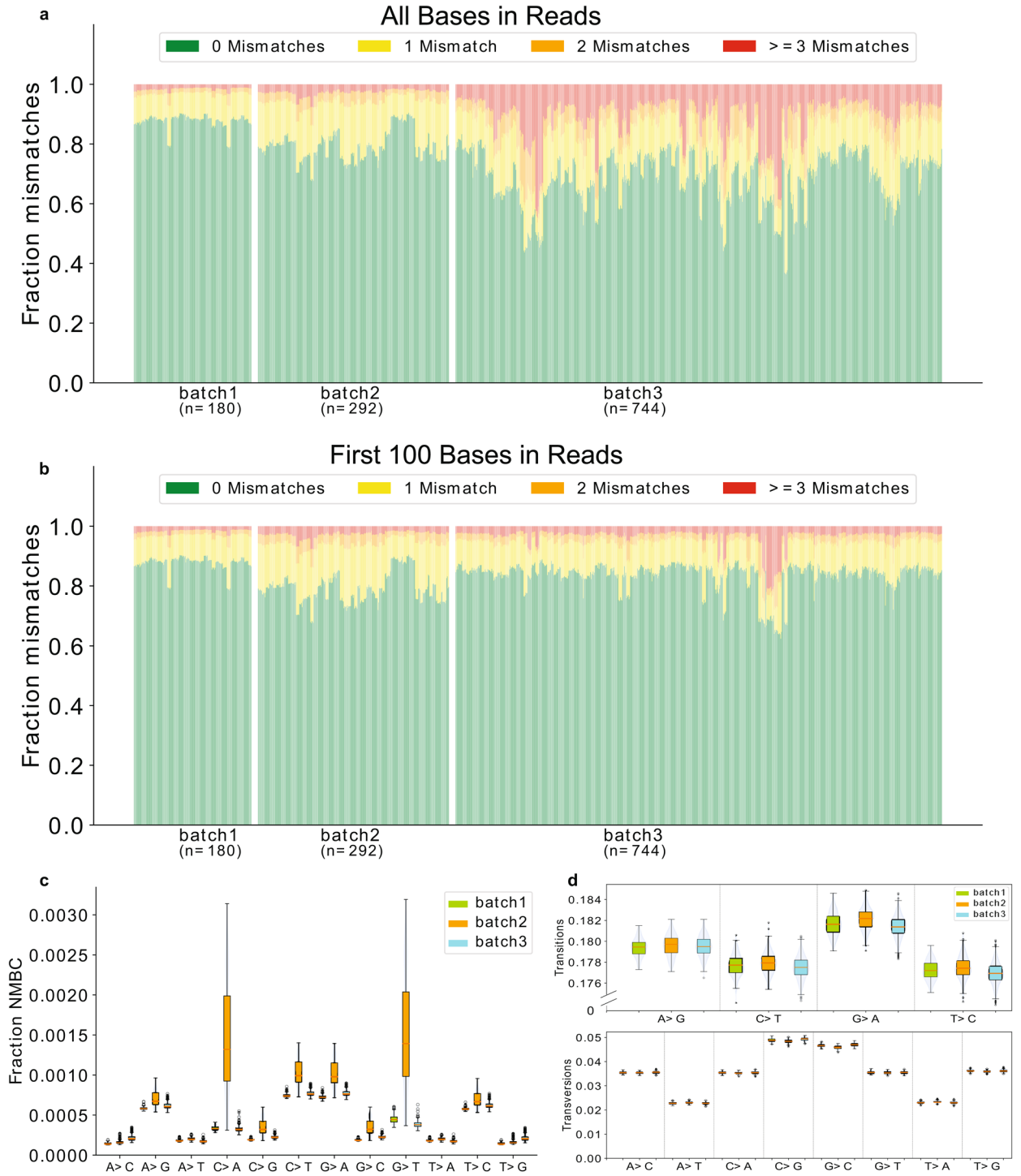
Correspondence and requests for materials should be addressed to A.N.A., J.M.P. or S.E.B.

Peer review information Kate Gao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

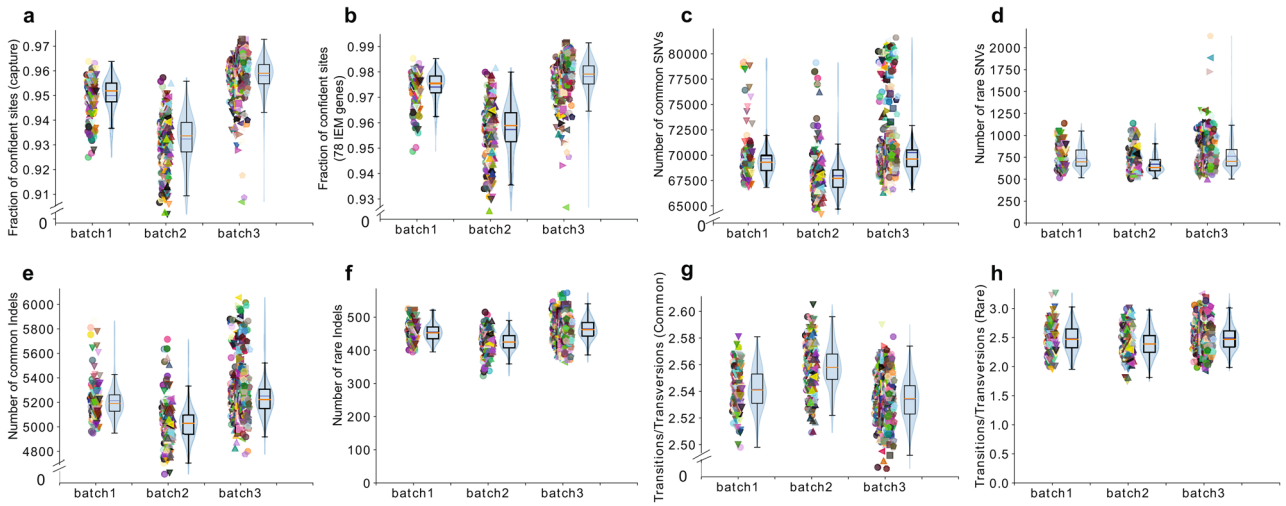


Extended Data Fig. 1 | See next page for caption.

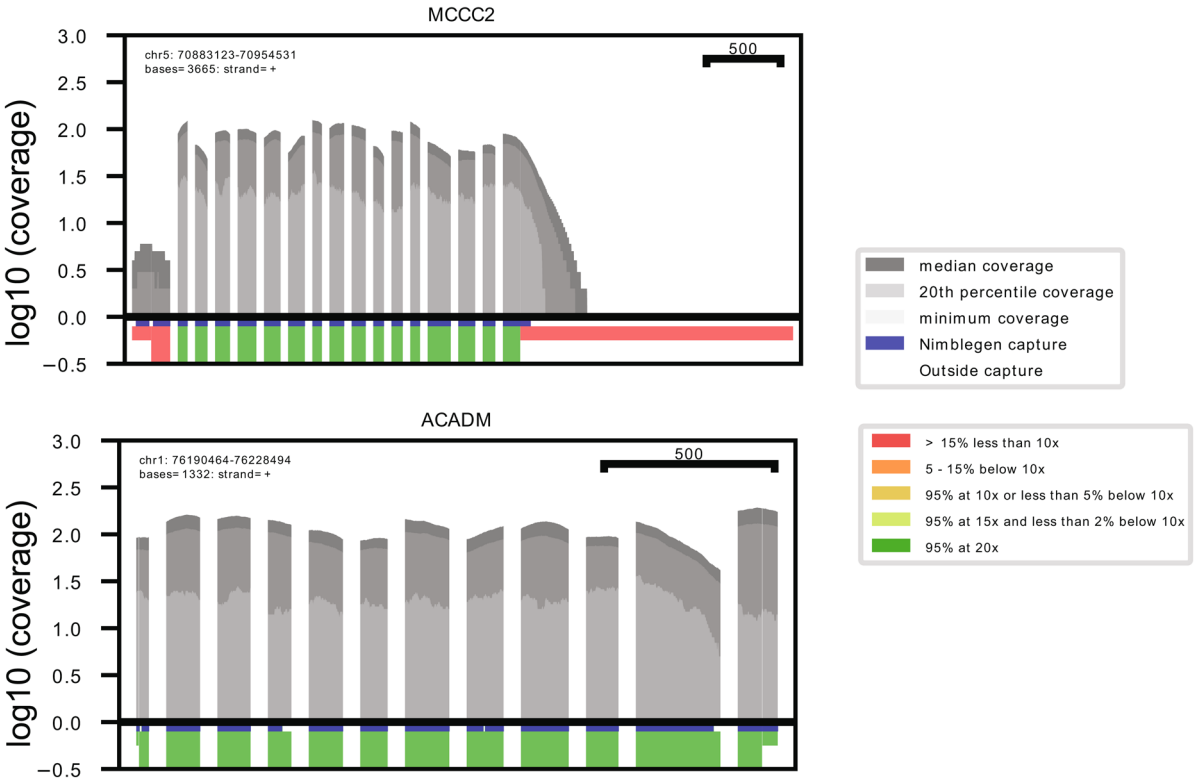
Extended Data Fig. 1 | Metrics for WES reads and coverage. a, Percentage of reads unmapped to the reference genome. **b**, Percentage of high quality read pairs (MQ > 20), without duplicates and properly paired. **c**, Percentage of duplicates in the reads across three sequencing batches **d-e**, Number of reads and high quality reads plotted batchwise. **f**, Inferred insert sizes plotted batchwise. **g**, Median coverage across Nimblegen capture region plotted batchwise **h**, Median coverage across 78 genes region plotted batchwise. **i**, Median fraction of capture covered at coverage depths of 1x to 30x plotted batchwise. **j**, Median fraction of 78 genes region covered at coverage depths of 1x to 30x plotted batchwise. In figures **a-f** and **i-j**, individual sample values are plotted, and adjacent box plots display the median (red) and interquartile ranges for the dataset, whiskers extend to the last data point within 1.5 times the interquartile range. The sample sizes for the boxplots in **a-h** were: batch1 (n = 180), batch2 (n = 292), batch3 (n = 744). Violin plots superimposed on the box plots show the data density and mean value (blue).



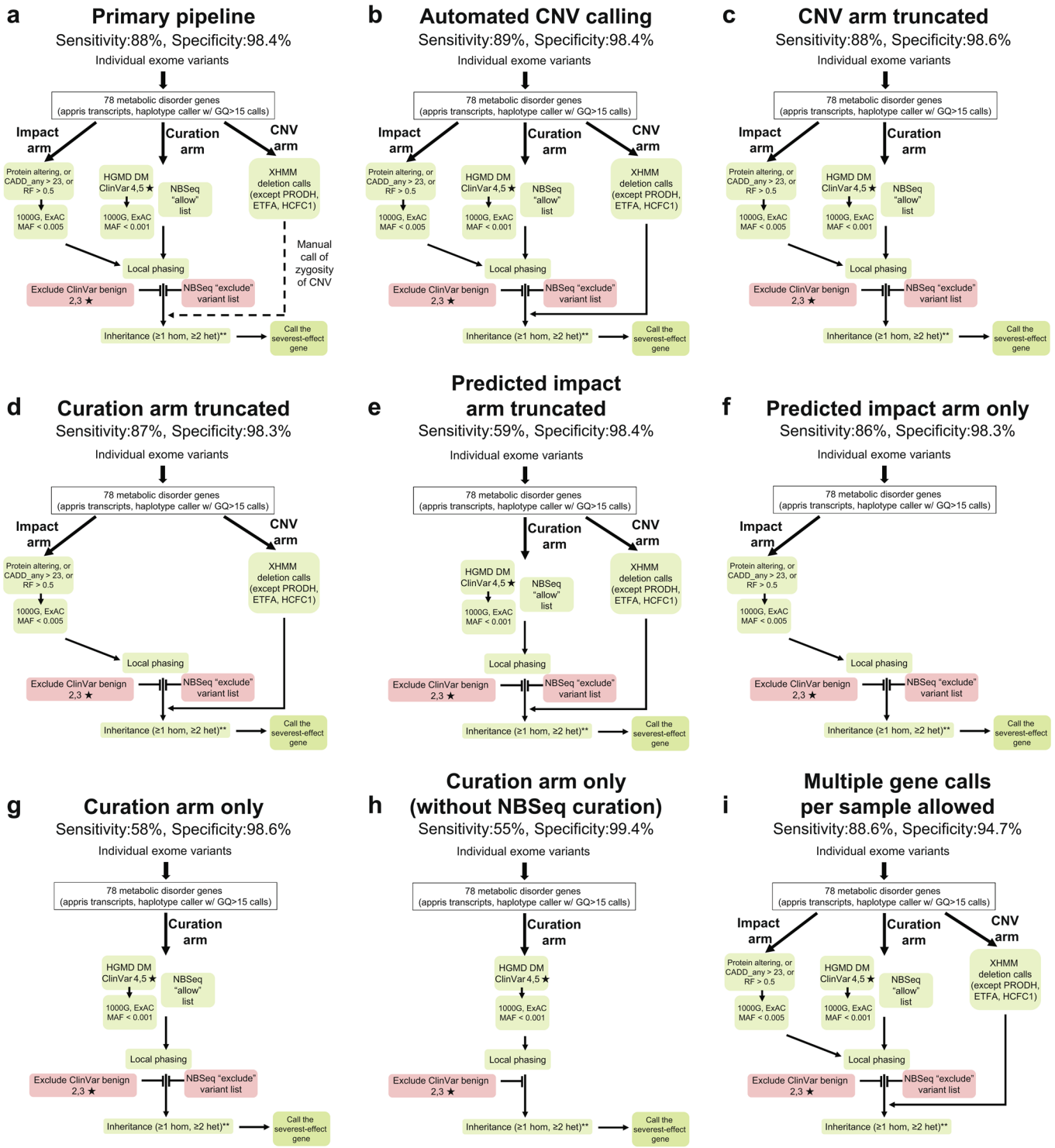
Extended Data Fig. 2 | DNA damage related metrics for the three sequencing batches. a, **b**, Fraction of reads with 0 (green), 1 (yellow), 2 (orange), and ≥3 (red) mismatches with reference genome considering (**a**) all bases of the reads and (**b**) first 100 bases of the reads. Batches 1 and 2 had read lengths of 101 bases and batch 3 had read length of 151 bases. All three batches had similar mismatch rates when only the first 100 bases were considered. **c**, Nucleotide mismatches by base change (NMBC) in the 1216 samples plotted batch wise. **d**, Frequencies of all single nucleotide changes by base type in high quality SNVs in the 1216 samples plotted batchwise. High quality SNVs from the VCF calls defined as marked PASS by GATK VQSR algorithm and with GQ ≥ 30. In both **c** and **d**, box plots display the median and inter quartile ranges for the dataset, whiskers extend to the last data point within 1.5 times the interquartile range and outliers beyond this are marked with circles. The sample sizes for the boxplots were batch1 (n = 180), batch2 (n = 292), batch3 (n = 744).



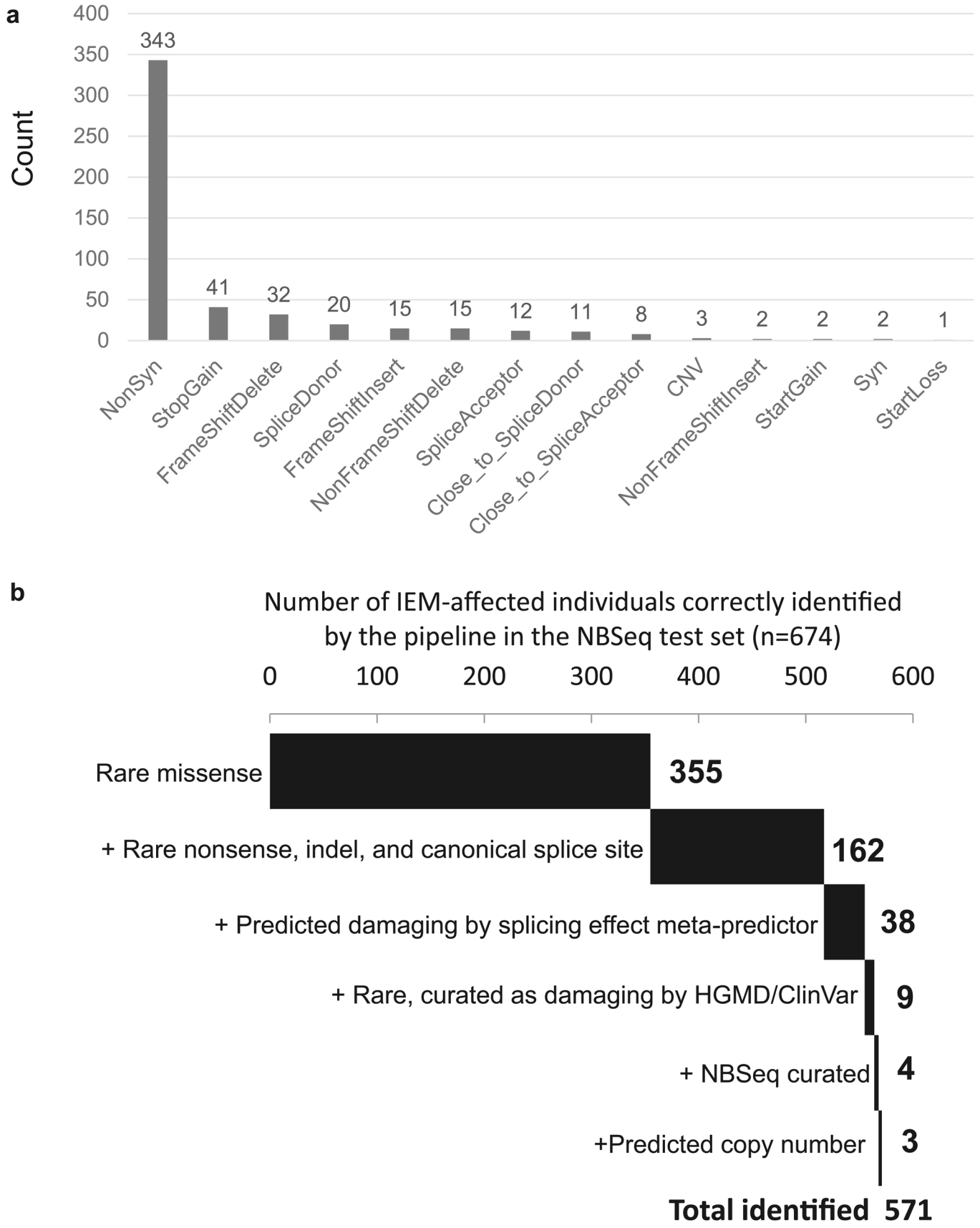
Extended Data Fig. 3 | Variant related quality metrics for 1,216 samples plotted batch wise. a, Confident sites across capture (from the GVCF file) **b**, Confident sites across 78 genes (from the GVCF file) **c**, Common high quality SNVs **d**, Rare high quality SNVs **e**, Common high quality indels **f**, Rare high quality indels **g**, Transition/Transversion ratios for high quality common SNVs **h**, Transition/Transition ratios for high quality rare SNVs. High quality variants are those marked as PASS by GATK VQSR and have GQ ≥ 30. Common variants have a frequency greater than 0.001 in 1000 Genomes Project phase 3 database and rare variants have a frequency less than 0.001 in the database. Individual sample values are plotted and adjacent box plots display the median (red) and interquartile ranges for the dataset, whiskers extend to the last data point within 1.5 times the interquartile range. Violin plots superimposed on the box plots show the data density and mean value (blue). The sample sizes for the boxplots were batch1 (n = 180), batch2 (n = 292), batch3 (n = 744).



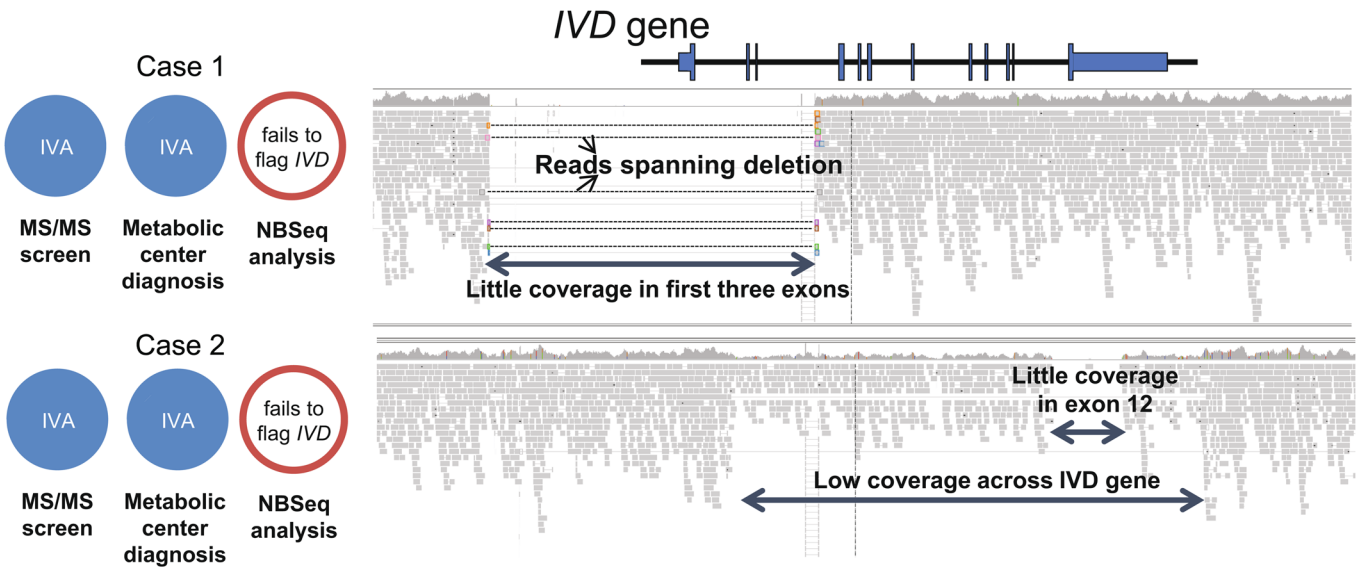
Extended Data Fig. 4 | Example showing variability of gene coverage in two IEM genes in the study across 1,216 samples. *MCCC2*, top, has poor coverage in the first exon across all samples. In contrast, *ACADM*, bottom, has good coverage across the gene. The blue vertical lines indicate positions with known pathogenic variants in HGMD and ClinVar. Plot of $\log_{10}$ of the median, 20th percentile and minimum coverage for each coding exon across all samples for a given sample set. Dark grey: Median coverage, medium grey: 20th percentile coverage, light grey: minimum coverage for each position. Coverage quality of each exon is indicated by colored blocks beneath the exon. Coverage quality of each exon is indicated by colored blocks beneath the coverage plot. Red: Greater than 15% of exon has less than 10x median coverage; green: 95% of the exon has minimum 20x coverage. UTRs that are part of the coding exons have a smaller indicator thickness. Regions of the exon that overlap with the capture array are indicated in blue just below the coverage plot. Exon scale in bases is shown in each plot.



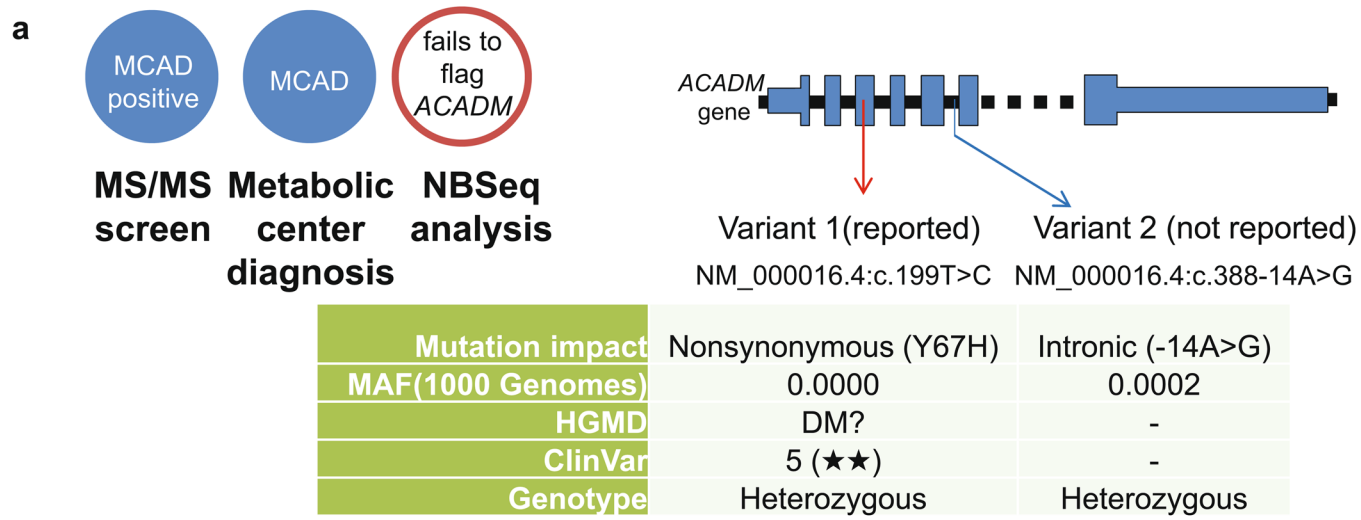
Extended Data Fig. 5 | Alternative pipelines derived from the final exome analysis pipeline to explore sensitivity-specificity tradeoffs. We created several alternate pipelines, altering or truncating different parts of the final exome analysis pipeline to probe contributions to overall sensitivity and specificity from various components of the pipeline. For each pipeline, the overall sensitivity and specificity on the NBSeq test set are shown. **a**, Final exome analysis pipeline **b-i)** Alternatives: **b**) Altering final pipeline by considering every CNV call homozygous **c-e**) Truncating the CNV arm, curation arm and predicted impact arm, respectively. **f-g**, Retaining the predicted impact arm or curation arm only, respectively **h**) Retaining only the rare pathogenic HGMD & ClinVar databases **i)** Allowing multiple gene calls for each sample if more than one gene predicted.



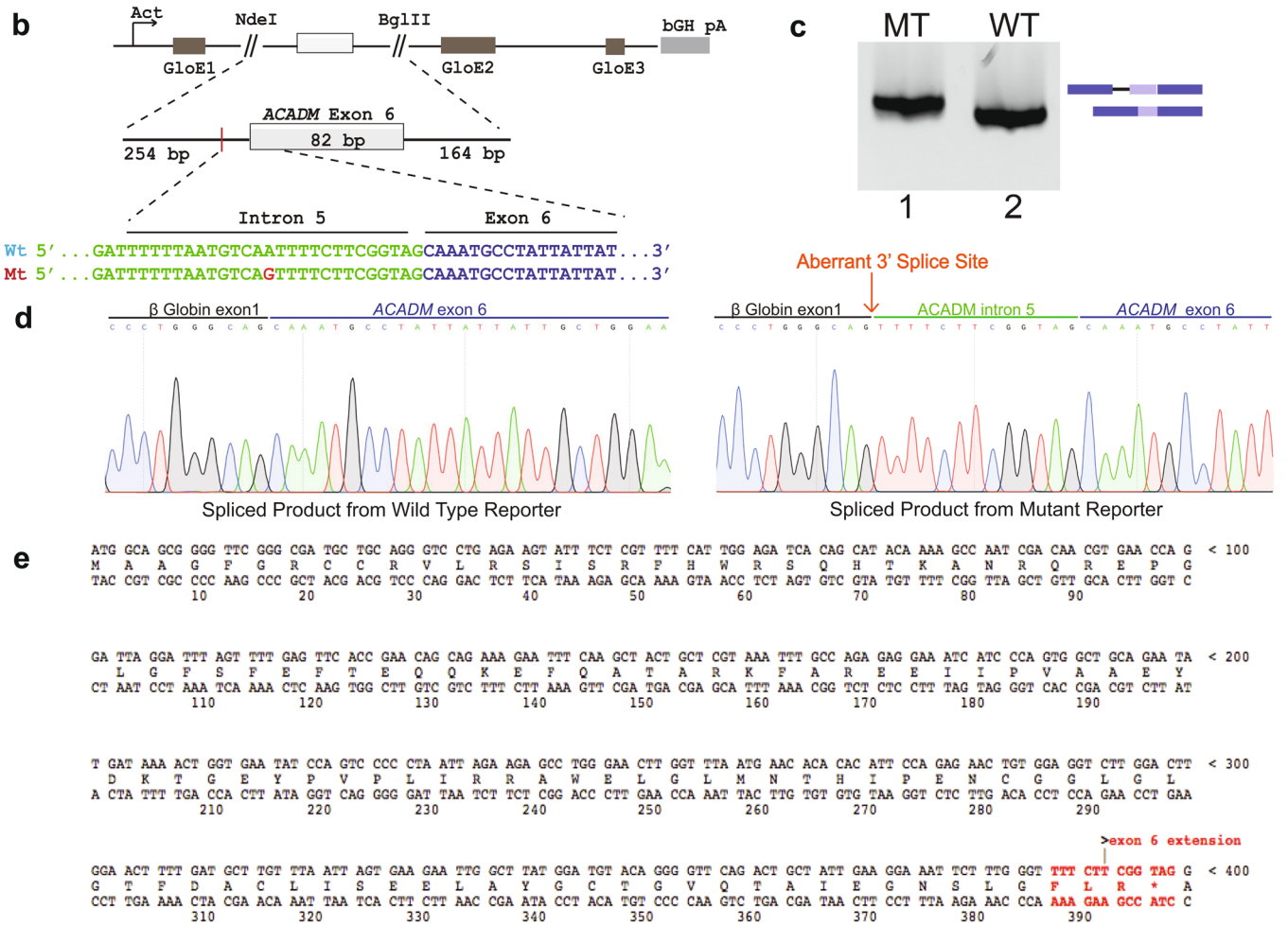
Extended Data Fig. 6 | Distribution of variants reported by the exome analysis pipeline in the NBSeq test set. a, Number of different variant types reported by the pipeline in IEM-affected individuals in genes associated with their IEMs the NBSeq test set (n = 674 individuals). **b**, Distribution of the types of variants responsible for the predictions of disease status in the 571 affected individuals correctly identified by the exome analysis pipeline.



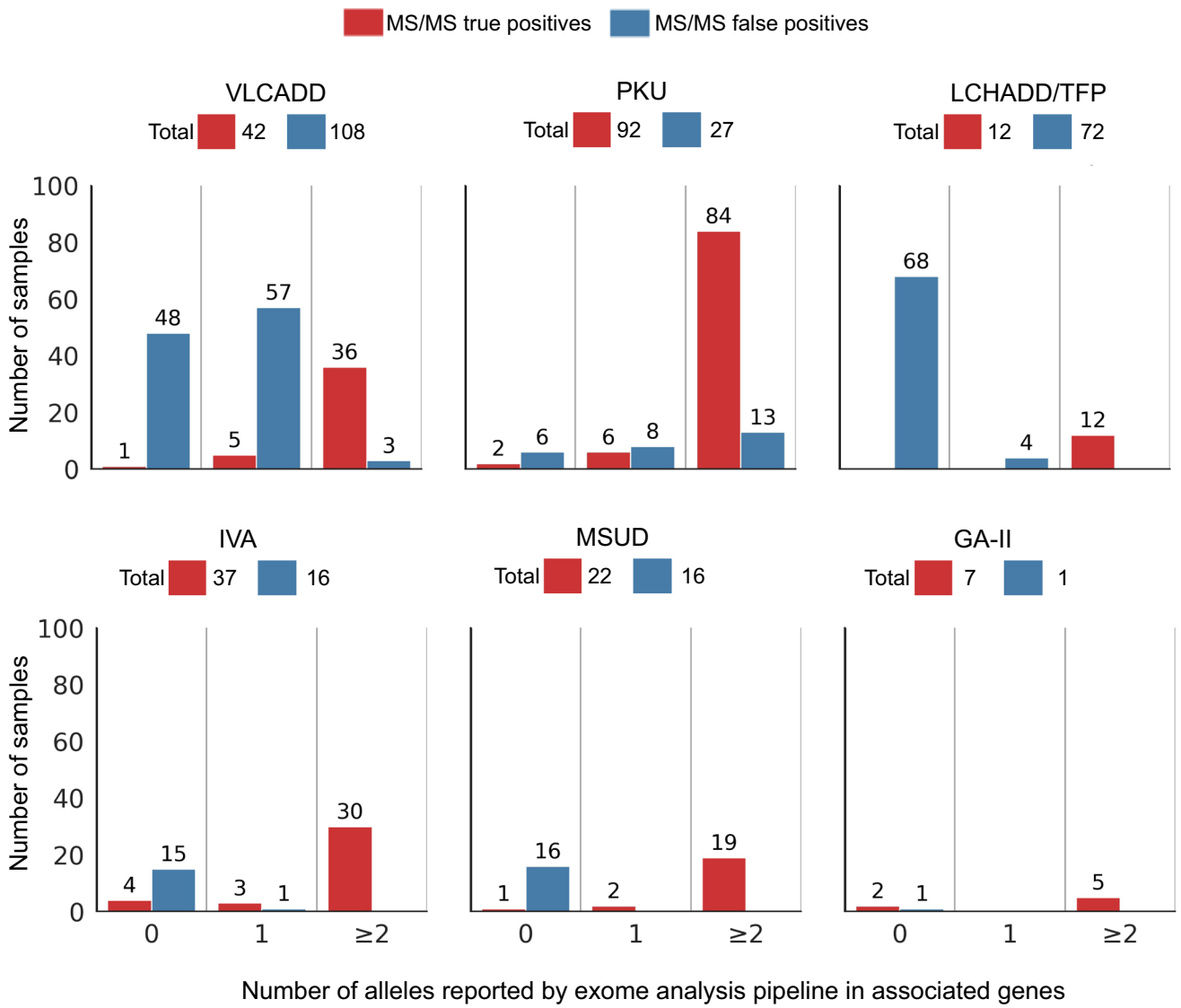
Extended Data Fig. 7 | Whole genome sequencing confirms potential *IVD* deletions in two individuals diagnosed with isovaleric acidemia initially missed in exome. In two cases where we performed WGS upon follow up of an exome false negative, we identified large deletions in the associated *IVD* gene. The WGS read alignments in the genomic region spanning the *IVD* is shown on the right for the two cases. The first case had almost no coverage in the region spanning the first three exons of *IVD*. The second case had almost no coverage of exon 12 of *IVD* along with low coverage across the whole gene. The first case had 11 split reads spanning the deleted region confirming the deletion event of the first three exons.



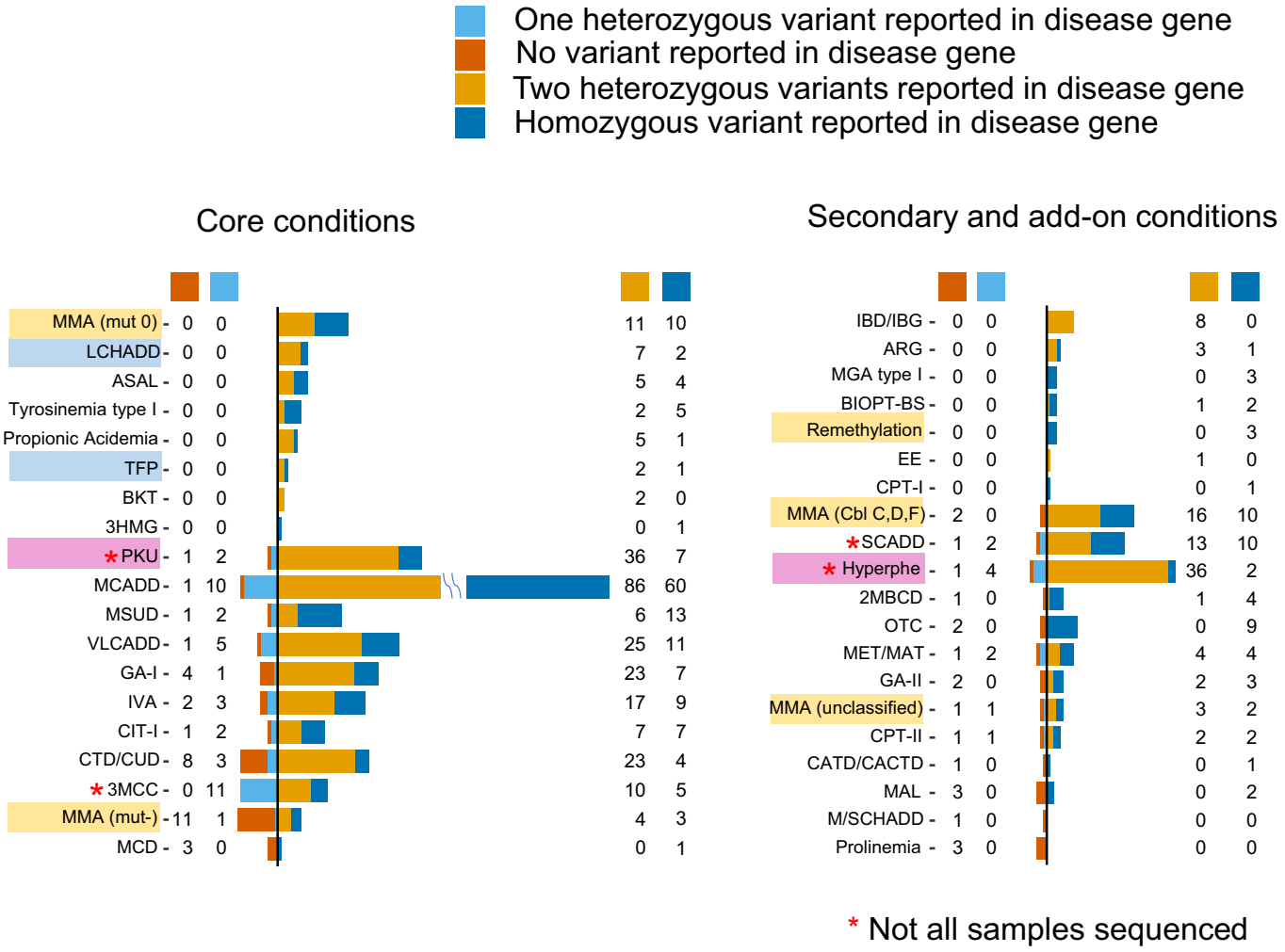
Mutation impact	Nonsynonymous (Y67H)	Intronic (-14A>G)
MAF(1000 Genomes)	0.0000	0.0002
HGMD	DM?	-
ClinVar	5 (★★)	-
Genotype	Heterozygous	Heterozygous



Extended Data Fig. 8 | Experimental splicing assay of a potentially pathogenic intronic variant in an exome false negative case. a, In an individual affected with MCADD, the exome analysis pipeline reported only a single rare nonsynonymous variant. A second rare intronic variant 14 bases from the splice site (NM_000016.4:c.388-14A>G) was a suspected pathogenic modification of the branchpoint A nucleotide. **b**, Diagram of the heterologous HBB splicing reporter construct containing the wild type ACADM sequence or the c.388-14A>G variant. **c**, RT-PCR analysis of reporter transcripts from wild type or mutant (lanes 1 and 2, respectively) reporter plasmids expressed in HEK293T cells (amplicons resolved by 12% PAGE and stained with SYBR Gold). The two spliced products are shown to the right of the gel image. The experiments were performed three times independently with similar results. **d**, Chromatograms corresponding to the sequence spliced junctions between HBB exon 1 and the wild type or mutant ACADM exon 6 constructs (left and right panel, respectively). **e**, Open reading frame of aberrant ACADM mRNA containing a 13 nt extension of exon 6 (red), resulting in a premature termination codon (PTC, *). Top, DNA sense strand; middle, predicted polypeptide; bottom, DNA reverse complement.



Extended Data Fig. 9 | Stratification of IEM-affected and MS/MS false positives by alleles reported by the exome analysis pipeline for NPV estimation of NPV of exome as a follow-up test after a positive MS/MS screen. For six MS/MS screens (VLCADD, PKU, LCHADD/TFP, IVA, MSUD, and GA-II), IEM-affected and MS/MS false positive cases in the NBSeq test set are stratified by the number of alleles reported by the exome analysis pipeline in the genes associated with those screens.



Extended Data Fig. 10 | Zygosity distribution of variants reported by the pipeline in relevant gene(s). For each IEM, bars show the zygosity distribution of the variants in relevant genes reported by the exome pipeline for the 674 IEM-affected cases from the test set. The numbers of cases correctly identified by the pipeline are broken down into those that had homozygous variants in relevant gene(s) (dark blue) and those that had two heterozygous variants in relevant genes(s) (orange). The number of cases that failed to be identified by the pipeline are broken down into those that had one heterozygous variant in relevant gene(s) (light blue) and those that had no reported variants in the relevant gene(s) (dark red). Left, core IEMs screened by California; right, secondary/add-on IEMs. IEMs sharing a common causative gene were not distinguished by the exome predictions alone. These included TFP and LCHADD (blue shading), PKU and hyperphenylalaninemia (pink shading), and the various MMA subtypes (yellow shading).

nature research

Corresponding author(s): Jennifer M. Puck, Steven E. Brenner and Aashish N. Adhikari

Last updated by author(s): 18 May 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Residual, de-identified newborn dried blood spot samples were obtained from the California Genetic Disease Screening Program (GDSP) and used to prepare DNA for exome sequencing. The GDSP Screening Information System (SIS) is a custom database of the California Department of Public Health Information Technology Services Division. It houses all case data for the newborn screening and follow-up programs. Information available at GDSPAdmins@cdph.ca.gov
Data analysis	The 1,416 samples were sequenced in three batches (Batch 1: 188 samples, Batch 2: 411 samples, Batch 3: 817 samples) and the QC metrics were grouped accordingly. Batches 1 and 2 had paired-end read lengths of 101 bp whereas batch 3 had paired-end read length of 151 bp. Raw sequences were mapped to the reference genome (v37), using BWA mem algorithm (v0.7.10). Resulting SAM files were converted to binary format, sorted and lane merged using Picard tools (v1.81). Duplicates were marked in the alignment files with Picard tools v1.81 (http://broadinstitute.github.io/picard/). Next, realignment around known indels and base quality score recalibration were performed using GATK toolkit (v3.3). Variants were called using the GATK Haplotype Caller function and the variant scores were recalibrated with GATK VQSR function. Combined calling was used on all samples. Variants were annotated using Varant (http://compbio.berkeley.edu/proj/varant/Home.html), a custom tool, as described, with the following public datasets: Gencode (v19), APPRIS (v24), 1000 Genomes (phase 3), ESP Project (ESP6500SI-V2-SSA137), Exome Aggregation Consortium (ExAC v0.3.1), Combined Annotation Dependent Depletion (CADD) (v1.3), MetaSVM and MetaLR from dbNSFP v3.1a, and dbscSNVv1. Variants previously associated with disease from HGMD (v2014.1), and those with star ratings for known deleterious variants from ClinVar (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz, accessed September 2017) were added to the call set. Copy number variant (CNV) calls on all 1,190 NBSeq samples were made using XHMM. Using the IGV browser, we further manually inspected the read alignment, mapping quality and overall coverage in the associated genes in sample BAM files in every exome false negative case in the validation set. The code and scripts used for the screening analysis of exome data and subsequent assessments are deposited in GitHub (https://github.com/nbseq1200/NBSeq1200paper)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The de-identified DBS from the California Biobank for this project (SIS request number 496) were obtained with a waiver of consent from the Committee for the Protection of Human Subjects of the State of California, under project number 14-07-1650, and in compliance with CDPH Biospecimen/Data Use and Confidentiality Agreement. California blood specimens and any data derived from the newborn screening program are confidential and subject to strict administrative, physical and technical protections. California law precludes any researcher from sharing blood specimens or uploading individual data derived from these blood specimens into any genomic data repository. If other researchers desire access to these data, they would need to make a separate application to the California Department of Public Health. Data in figures 2b and 2c and Extended Data Figs 6,9,10 can be found in Supplemental Table 3.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. The sample size reflects the total number of affected cases in California in the study period, all false positive cases for a select set of disorders from the well-baby nursery. From the roughly 4.4 million children born in California July 7, 2005, to December 31, 2013, we obtained two de-identified 3.2 mm punches from dried blood spots (DBS) obtained for newborn screening (NBS) from each of 1,728 newborns. The samples included those from all 1,325 infants screened by tandem mass spectroscopy (MS/MS) who were subsequently confirmed to have an inborn error of metabolism (IEM), as well samples from 9 cases not identified by MS/MS screening, but diagnosed clinically. Also included were 394 MS/MS false positive cases selected from 10,011 individuals, for whom MS/MS screening was positive, but who were ultimately diagnosed as unaffected. All MS/MS false positives from well-baby nurseries for the long-chain fatty acid oxidation disorders (VLCAD, LCHAD, MADD), and three other IEMs (PKU, MSUD, IVA) were requested to investigate the hypothesis that genetic variants could underlie the abnormal analytes leading to false positive MS/MS results. We sequenced 1,416 exomes and analyzed 1,190 exomes out of the 1,416 that passed quality control metrics and were phenotypically relevant. This is the largest study to date addressing the question of exome sequencing for IEMs, and the study was highly powered to provide accurate estimates of sensitivity and specificity in the test set.
Data exclusions	From the 1,728 de-identified dried blood spots obtained, a random subset of 312 non-Hispanic white and Hispanic cases for SCADD, 3MCC, and elevated phenylalanine, were excluded due to budget limitations. Of the 1,416 samples sequenced, 200 were further excluded as they did not meet quality thresholds set for the study, and 26 were excluded as they were affected with disorders not under evaluation in the current study. Exclusion criteria for the random subset of 312 cases were not pre-established. The quality metrics that resulted in the exclusion of 200 cases were pre-established. The 26 cases excluded due to the underlying disorder were excluded based on pre-established criteria.
Replication	A validation set was used to design the automated exome analysis pipeline. The automated exome analysis pipeline (Fig. 2a, main paper; code available at https://github.com/nbseq1200/NBSeq1200paper) was run once on each sample in the NBSeq test set while blinded to any phenotypic data, reflecting a typical application setting for a primary newborn screen. The clinical team generated a “reviewed diagnosis” key for all NBSeq study cases based on the strength of evidence for the clinical diagnoses recorded in the GDSP database and recent literature. This reviewed diagnosis key was used to compare clinical diagnoses with exome predictions. The test set is a replication study of the validation set results – and the sensitivity and specificity derived from the test set replicated what was found in the validation set. That was the only replication study performed or needed.
Randomization	We first arbitrarily divided the 1,190 samples into a validation set (178 samples) in which we were unblinded to final diagnosis so as to explore the robustness of parameter choices in our exome analysis pipeline (data not shown), and a test set (1,012 samples), which was subjected to the final optimized pipeline only once and whose results are reported here.
Blinding	The team developing the analysis pipeline (A.N.A., Y. W. and S.E.B.) were blinded to the test set during pipeline development and assessment. S.E.B. remains blinded to permit future analyses on this unique dataset.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	This study employed residual DBS derived from individuals affected by IEMs during an 8.5yr period in the State of California, as well as all of those who screened positive for 6 IEMs. A random subset of 312 non-Hispanic white and Hispanic cases for SCADD, 3MCC, and elevated phenylalanine were excluded due to budget limitations
Recruitment	The NBSeq project used deidentified newborn dried blood spot samples obtained from the California Biobank Program (SIS Biobank request ID number 496) in compliance with CBP Biospecimen/ Data Use and Confidentiality Agreements. The CBP contains biospecimen and data resources of the California Department of Public Health's Genetic Disease Screening Program (GDSP), which administers the Newborn Screening Program in the state of California. Samples and data are made available to researchers for approved purposes, including the development and evaluation of screening tests and strategies.
Ethics oversight	This study was approved by the California Committee for the Protection of Human Subjects (CPHS), under project number 14-07-1650.

Note that full information on the approval of the study protocol must also be provided in the manuscript.