TITLE

Poison exon annotations improve the yield of clinically relevant variants in genomic diagnostic testing

AUTHORS

Stephanie A Felker[1], James MJ Lawlor[1], Susan M Hiatt[1], Michelle L Thompson[1], Donald R Latner[1], Candice R Finnila[1], Kevin M Bowling[2], Zachary T Bonnstetter[1], Katherine E Bonini[3], Nicole R Kelly[4], Whitley V Kelley[1], Anna CE Hurst[5], Melissa A Kelly[6], Ghunwa Nakouzi[6], Laura G Hendon[7], E Martina Bebin[8], Eimear E Kenny[3,9,10], Gregory M Cooper[1]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA 35806
[2]Washington University School of Medicine, Saint Louis, MO, USA 63110
[3]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai. New York, NY, USA 10029
[4]Department of Pediatrics, Division of Pediatric Genetic Medicine, Children's Hospital at Montefiore/Montefiore Medical Center/Albert Einstein College of Medicine, Bronx, NY, USA 10467
[5]University of Alabama in Birmingham, Birmingham, AL, USA 35294
[6]HudsonAlpha Clinical Services Lab, Huntsville, AL, USA
[7] University of Mississippi Medical Center, Jackson, MS, 39216
[8]Department of Neurology, University of Alabama at Birmingham, Birmingham, AL, USA 35294
[9]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA 10029
[10]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA 10029

CORRESPONDENCE

Gregory M Cooper, gcooper@hudsonalpha.org, 256-327-9490

KEY WORDS

Poison exon, alternative splicing, nonsense-mediated decay, neurodevelopmental disorders, voltage-gated sodium channels

ABSTRACT

Purpose:

Neurodevelopmental disorders (NDDs) often result from rare genetic variation, but genomic testing yield for NDDs remains around 50%, suggesting some clinically relevant rare variants may be missed by standard analyses. Here we analyze "poison exons" (PEs) which, while often absent from standard gene annotations, are alternative exons whose inclusion results in a premature termination codon. Variants that alter PE inclusion can lead to loss-of-function and may be highly penetrant contributors to disease.

Methods:

We curated published RNA-seq data from developing mouse cortex to define 1,937 PE regions conserved between humans and mice and potentially relevant to NDDs. We then analyzed variants found by genome sequencing in multiple NDD cohorts.

Results:

Across 2,999 probands, we found six clinically relevant variants in PE regions that were previously overlooked. Five of these variants are in genes that are part of the sodium voltage-gated channel alpha subunit family (*SCN1A*, *SCN2A*, and *SCN8A*), associated with epilepsies. One variant is in *SNRPB*, associated with Cerebrocostomandibular Syndrome. These variants have moderate to high computational impact assessments, are absent from population variant databases, and were observed in probands with features consistent with those reported for the associated gene.

Conclusion:

With only a minimal increase in variant analysis burden (most probands had zero or one candidate PE variants in a known NDD gene, with an average of 0.77 per proband), annotation of PEs can improve diagnostic yield for NDDs and likely other congenital conditions.

INTRODUCTION

Neurodevelopmental disorders (NDDs) affect 1-2% of children and display a wide phenotypic range that includes seizures, developmental delay, autism, and intellectual disability[1]. NDDs often result from highly penetrant genetic variation, and 1,493 genes are confidently associated with developmental disorders[2]. Exome and genome sequencing can lead to molecular diagnoses in NDDs, as they allow for detection of rare variants across many genes. In turn, a molecular diagnosis, particularly early in life, can provide many benefits to patients and their families[3,4]. However, despite rapid improvements in genomic testing, such tests frequently fail to reveal any clinically relevant variants; across most studies, the diagnostic yield of genomic testing for NDDs is at or below 50%[5].

The vast majority of variants found by diagnostic testing are in coding exons that are part of standard gene annotations. Non-coding intronic variants are often ignored due to lack of functional annotation. However, analyses of transcriptome data have revealed functionality in otherwise non-annotated intronic regions[6]. One such category of functional sequences includes “poison exons” (PEs). PEs are alternatively spliced exons that result in the creation of a premature termination codon (PTC), either by direct inclusion or by frameshifting the transcript. The inclusion of PTCs cause translation to stall, thereby triggering nonsense-mediated decay (NMD) of the transcript and reducing protein levels in the cell. Some PEs are spliced into transcripts at relatively high levels during mammalian brain development in early stages and are hypothesized to regulate protein levels of the genes in which they reside[7,8,9,10]. Crucially, since PE usage leads to NMD, and therefore reduction of protein, genetic variation that alters PE inclusion can

lead to a loss-of-function effect. Variant-induced PE inclusion is known to cause a small proportion of NDDs including *SCN1A*-associated epilepsies and Cerebrocostomandibular Syndrome (MIM 117650)[11,12].

Despite their biological and disease relevance, PEs are often absent from standard gene models and PE-associated variants may be dismissed as irrelevant deep intronic variants. To remedy this, there have been reannotation efforts resulting in the addition of some of these regions to gene models. For example, epilepsy-related PE transcripts were included in Human GENCODE release 33[13]. However, causal PE variants are likely still missed due to lack of routine implementation in the variant analysis process. This may, in part, reflect the concerns with analyst and clinician "alert fatigue"[14] due to the dilution of variant analysis with too many benign intronic variants.

We used published RNA-seq data from developing mouse cortices sampled from embryonic day 14.5 to 2 years old[15] to define a list of evolutionarily conserved PEs that are differentially expressed in the developing brain for application in variant analysis pipelines. We filtered these results to highlight PEs potentially relevant to NDD, and within these curated regions found PE variants suspected to contribute to NDD in six probands from a cohort of 2,999. These variants were missed in conventional genomic analyses. While these six probands only represent 0.2% of the total probands analyzed, there is minimal additional effort required to analyze PE variants, which compares favorably to the increase in the yield of clinically relevant variants. Further, beyond the potential diagnostic benefit to the individual probands, discovery of these alleles is of more general translational value, particularly in light of drug-discovery efforts targeted to PEs.

MATERIALS AND METHODS

Curation of Clinically Relevant Poison Exons

To identify potential PE regions of interest, we used publicly available results of deep mRNA sequencing of mouse cortexes during development at E14.5, E16.5, P0, P4, P7, P15, P30, 4 month, and 2-year timepoints[15]. Yan and colleagues (2015) reported 77,949 elements alternatively spliced in the mouse cortex over developmental time. To curate these elements for conserved PE in humans each element was filtered for the following criteria (Figure 1):

1) The AG/GU (+/- 1 or 2 canonical splice sites) are conserved between mm10 and hg19 (the reference assemblies used by Yan et al.,)
2) Alternatively-spliced (AS) element is determined to cause NMD upon inclusion ("NMD_in") per Yan et al.
3) The cassette must be included between two canonical exons in mm10, and its coordinates must successfully lift over from mm10 into hg38
4) The hypothetical PE must not overlap with a Matched Annotation from the NCBI and EMBL-EBI (MANE) or RefSeq Select transcript exon

Data utilized for Criteria 1 and 2 were provided by the "Non_redundant_cass_mm10_summary.txt" supplementary file provided in the Yan publication at https://zhanglab.c2b2.columbia.edu/index.php/Cortex_AS. Criteria 3 was determined by intersecting the conserved NMD_in elements with the mm10 NCBI RefSeq UCSC track to select elements with distal 5' and 3' intron coordinates that overlap with the canonical exon scaffold[16,17] using bedtools[18]. Criteria 4 evaluation was conducted by lifting PE cassettes into the hg38 genome using the UCSC Lift Genome

Annotations tool[19]. These elements were then intersected with the NCBI RefSeq Select and MANE (nbciRefSeqSelect) dataset table from the UCSC Table Browser using bedtools. Each gene was assigned a disease consequence as determined by Online Mendelian in Man (OMIM)[20]. These elements and cassettes are available in Supplementary Data 1. Genomic coordinates of the PE cassettes are available in Browser Extensible Data (BED) format as Supplementary Data 2, and genomic coordinates of the introns containing the PEs are available in BED format as Supplementary Data 3. Supplementary Data 1-3 and the code and URLs to the external data used to curate these regions can be found at https://github.com/HudsonAlpha/poison_exon_variant_analysis.

Cohort data acquisition

Data from probands in eight different research cohorts were analyzed, all part of translational genome sequencing studies on probands suspected to have congenital disease (Table 1). Data were acquired and results were returned in these investigations in accordance with site-specific informed consent and IRB-approved protocols. SouthSeq and NYCKidSeq are projects within the Clinical Sequencing Evidence-Generating Research (CSER2) Consortium[28]. Genomic data from SouthSeq, HudsonAlpha Clinical Sequencing Exploratory Research Consortium (CSER1), Alabama Genomic Health Initiative (AGHI), University of Alabama in Birmingham treatment-resistant epilepsy cannabidiol response clinical trial (CBD), Children's of Alabama Genome Sequencing (COAGS), Alabama Pediatric Genomics Initiative (PGEN), and the University of Alabama in Birmingham Undiagnosed Disease Program (UDP)) were sequenced at HudsonAlpha Genomic Services Lab (GSL) or at the

HudsonAlpha Clinical Services Lab (CSL). Genomic data from the NYCKidSeq cohort was acquired via the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space[29] (AnVIL; see "NYCKidSeq Cohort Analysis on the NHGRI AnVIL"). Cohort-specific enrollment, sequencing, quality control, and analytic details are described in their respective publications (Table 1). Briefly, DNA was extracted from whole blood and sequencing libraries were constructed with or without PCR amplification. DNA library fragments were sequenced from both ends (paired) with a read length of 150 base pairs using the Illumina HiSeq or NovaSeq platforms to a target mean depth of 30x and target coverage of >80% of bases covered at 20x. Quality control included confirmation of each sample's expected biological sex based on counts of chrX heterozygous variants and chrY variants and, when applicable, expected family relationships using somalier v0.2.5[30] or KING v2.1.8[31].

For this analysis, all HudsonAlpha GSL/CSL samples were re-processed through a single sequence alignment and variant calling pipeline. Sequence reads were aligned to GRCh38.p12 using the Sentieon v201808.07 implementation[32] of BWA-MEM[33] and command line option -M -K 10000000. BWAKit was used for post-alt processing of the alternate contig alignments. Duplicate reads were marked, and base quality scores were recalibrated with Sentieon v201808.07 using dbsnp v.146 and Mills and 1000G gold standard indels as training data[32]. Variants were called on the hg38 primary contigs (chr1-chr22, chrX, chrY, chrM) using Strelka v2.9.10[34] in germline single-sample analysis mode. For related samples, Illumina's gvcfgenotyper v2019.02.26 (https://github.com/Illumina/gvcfgenotyper/) was used to merge the strelka genome VCF (gVCF) files into a multi-sample VCF to easily determine variant inheritance.

NYCKidSeq Cohort Analysis on the NHGRI AnVIL

NYCKidSeq genome VCFs were collected and analyzed at one of two clinical laboratories (the New York Genome Center or Rady Children's Institute for Genomic Medicine)[19,35]. Briefly, we collaborated with the CSER2 Data Coordination Center[36] to develop a Docker image containing common genomics tools including bcftools v1.9[37], htslib v1.9, and bedtools v2.30.0. We developed a workflow in Workflow Description Language (WDL) that uses bedtools and bcftools to quality-filter and region-filter VCFs equivalently to our local filtering process as described below in "Cohort Analysis with Clinically Relevant Regions." The Dockerfile, WDL, and URLs for the deployed image and AnVIL workflow are available at https://github.com/HudsonAlpha/poison_exon_variant_analysis.

Cohort Analysis with Clinically Relevant Regions

Variant call format (VCF) data from all cohorts was filtered for quality using the bcftools "filter" command requiring all of the following:

1. "PASS" variant calling filter status
2. Variant total read depth greater than 10 reads
3. Heterozygous genotypes having 20% to 80% of reads supporting the alternate allele
4. Genotype quality over 80

The resulting quality-filtered VCFs were region-filtered using bedtools intersect and a BED file containing the genomic coordinates of PE-containing introns detailed in Supplementary Data 1.

Annotations for the resulting quality-filtered and region-filtered variants were generated using the Ensembl VEP v102[38]. VEP plugins and custom annotations were used to add Combined Annotation Dependent Depletion scores v1.6 (CADD score)[39], Genomic Evolutionary Rate Profiling scores (GERP score)[38] scores, and variant frequencies from gnomAD v3.1.1[41] and TOPMed Freeze 8[42]. Finally, variants were filtered for biological relevance via a custom R script to require all the following:

1. Variant consequence includes a non-coding effect ("NMD_transcript_variant", "intron_variant", "non_coding_transcript_variant", or "non_coding_transcript_exon_variant")
2. Variant is predicted in the top 10% of deleterious variants by CADD (CADD score ≥ 10)
3. Variant position is predicted to be evolutionarily conserved by GERP (GERP score ≥ 0)
4. Variant is absent from, or appears at most with an allele count of 1, in the TOPMed and gnomAD databases

Filtered variants were curated based on proband phenotype and disease relevance and variants of interest were classified using American College of Medical Genetics and Genomics (ACMG) and Association of Molecular Pathology (AMP) guidelines[43,44,45]. Variant annotation settings are available at https://github.com/HudsonAlpha/poison_exon_variant_analysis. Poison exon variants are reported in Table 2 with GRCh38 coordinates and HGVS nomenclature, validated in VariantValidator[46].

RESULTS

Curation of Regions for Clinical Analysis

Yan and colleagues found 77,949 differentially spliced cassette exons in mice that resulted in alternative coding and NMD isoforms[15]. From these alternatively spliced exons, 47,918 had AG/GU splice sites conserved with human (hg19), and 2,730 of those resulted in NMD upon inclusion. 2,522 of these elements were included between two canonical exons in mm10 and successfully lifted over to hg38. PEs with overlap in the hg38 NCBI and EMBL-EBI MANE or RefSeq Select transcripts were removed to exclude human canonical exons resulting in a total of 1,937 candidate PEs (Supplemental Data 1). In this set, 571 are in genes that have one or more associated OMIM phenotypes. A BED file of introns and PEs defined for this analysis are available in Supplemental Data 2 and 3, respectively.

Poison Exon Variants

The regions described above were used to find PE variants in the eight cohorts described in Table 1. Individual proband VCFs were filtered to extract high-quality, rare, conserved, predicted-deleterious variants that lie within PE elements and are not annotated to disrupt a coding exon. Manual variant curation resulted in identification of six variants with potential relevance to proband phenotype (Table 2, Figure 2). Using ACMG guidelines, five were classified as Variants of Uncertain Significance and one as Pathogenic. These variants had moderate CADD scores (median 17.4, range 13.7-21.6), indicating they rank among the top 5% of most highly deleterious SNVs in the human reference genome; we note that, as all these variants are outside coding exons and PEs are not used in the CADD model[39], these scores may underestimate their

deleteriousness. These variants have moderate to high GERP scores (median 3.99, range 1.71-6.46[40]) and are within regions of significantly elevated mammalian sequence conservation (Figure 2), suggesting they affect positions that have been under selective constraint during mammalian evolution. All six variants are absent from gnomAD[41] and TOPMed[42].

Probands A, B, and C have variants within intron 20 of *SCN1A* and each has seizure phenotypes that are consistent with *SCN1A*-associated epilepsy (see Case Reports in Supplemental Data 4). This intron harbors the 20N poison exon[9] variants in or near to 20N have been previously associated with epilepsy in both humans and mice[8,11]. The phenotype of Proband B also included developmental delay and spastic dystonic cerebral palsy with left triplegia. Proband C was reported to have intellectual disability and developmental delay. All three variants were inherited: Probands A and C inherited their variants from a parent affected with mild cognitive disabilities and seizures, respectively, and the variant found in proband B was inherited from an unaffected parent. Probands B and C had affected siblings that also shared their respective variants. Proband C has an additional deceased sibling who had seizures and whose variant status is unknown. The variants found in Probands A (chr2:166007299C>T) and B (chr2:166007176T>C) flank the 20N PE and are 7 and 54 base pairs away from the PE boundaries (chr2:166007230-166007293), respectively. The variant in Proband C (chr2:166003356A>G) is in a conserved region near the 5' splice site of intron 20.

Proband D has a variant 170 base pairs from the 3' boundary of the 17A PE (chr2:165358030C>A) of *SCN2A*. He exhibits developmental delay, intellectual

disability, and seizures, which are consistent with *SCN2A*-associated phenotypes. Proband D had four VUSs reported from a targeted gene panel and a VUS reported from genome sequencing, but none of these variants were considered to be compelling in terms of disease relevance (see Case Report in Supplemental Data 4).

Proband E has a variant in a ~1.5 kb conserved region containing both the 18N PE and the alternatively spliced exon 21 of *SCN8A*. Exon 21 is a canonical exon present in the MANE transcript of *SCN8A* but is skipped in 3 protein-coding transcripts (ENST00000668547.1, ENST00000545061.5, and ENST00000355133.7), and 18N is included in a truncated alternatively spliced isoform (ENST00000548086.3)[7]. She exhibits seizures and impaired awareness, consistent with Developmental and Epileptic Encephalopathy 13 (MIM 614558), associated with variation in *SCN8A*.

Proband F presented in the SouthSeq cohort with a phenotype consistent with Cerebrocostomandibular Syndrome (CCMS, MIM:182282), including rib anomalies, clinodactyly, ventricular and atrial septal defects, micrognathia, and facial dysmorphisms of deep-set eyes, low-set ears, and a prominent nasal bridge. The proband's variant (chr20:2467306C>G) is within a PE in *SNRPB* and has been submitted to ClinVar as pathogenic (VCV000183431.3). This variant has been reported in 10 unrelated, affected individuals with both *de novo* and inherited alleles with incomplete penetrance[12]. Application of ACMG evidence codes results in a classification of Pathogenic.

DISCUSSION

The diagnostic yield of genome sequencing for patients suspected to have congenital disease remains at or below 50%, and the incorporation of new genomic annotations may increase yield. However, additional annotations must be precise and biologically relevant so as to avoid diluting genome-wide analysis with an overwhelming number of inconsequential variants. Here, we curated a list of intronic regions of the human reference assembly that are: (1) associated with the PE-mediated regulation of gene expression in the developing mouse brain; (2) conserved between humans and mice; and (3) in genes associated with human neurodevelopmental and Mendelian diseases. We used these genomic regions to search for potentially disease-relevant variants that may affect PE inclusion in 2,999 probands from eight rare disease sequencing cohorts and discovered six variants of interest. These variants are all absent from the gnomAD and TOPMed variant frequency databases. The CADD scores for these variants are higher than 95-99% of human SNVs (range 13.74-21.6) but are at the low end of the range for variants associated with NDDs and other Mendelian diseases[47]. This may at least in part reflect the fact that these are deeply intronic variants in sparsely annotated regions and therefore lack most of the features used within the CADD model to infer deleteriousness. The primary annotation type supporting a deleterious effect is evolutionary sequence conservation. All these variants affect individual nucleotide positions with moderate to high degrees of mammalian conservation (GERP scores of 1.71-6.46) and are in regions of significantly elevated conservation[40] (Figure 2).

Based on ACMG scoring criteria, five of the six variants we describe are VUSs, with only one evidence code applied based on absence from population databases (PM2). These five variants still require additional evidence to be assigned a more definitive status. Experimental analyses of transcripts showing evidence of altered PE inclusion are key, as this could provide experimental confirmation of a loss of function effect. Such work has been conducted for PE variants in *SCN1A* 20N[8,11]. In that context, we note that the three VUSs in *SCN1A* that we have identified here are all within the same intron containing the 20N PE studied by Carvill and colleagues (2018), who found five probands with rare variants in or near to 20N. Several of these variants led to altered 20N splicing *in vitro*[11], and one was shown in mice to affect splicing of 20N *in vivo* and lead to phenotypes similar to those observed for *Scn1a* loss of function variants[8].

A recent set of recommendations for interpretation of non-coding variants[44] builds upon the previous recommendations[43,45], but PEs are not specifically covered in this assessment. The authors suggested use of PS3_Strong in one example case where inclusion of a “cryptic exon” with a PTC was confirmed at the RNA level in *CFTR*. We believe PEs comprise a distinct subset of “cryptic exons” that warrant specific consideration. They are abundant across human genomes, including being present within hundreds of disease-associated genes. As the PEs described here were identified as alternatively spliced in mouse brain tissue, they comprise “biologically relevant” transcripts that are found in a key disease tissue of neurodevelopmental genes. Their high degrees of evolutionary sequence conservation further support their

biological and disease relevance, and the potential value of defining PE-specific criteria for clinical variant assessment.

Given that loss of function is a well-established mechanism for many genetic conditions, and that by nature, PEs result in PTCs, this would suggest that application of the ACMG evidence code PVS1 may be appropriate. One previously reported variant near the *SCN1A* 20N PE, for example, was shown in heterozygous mice to lead to 50% reduction of SCN1A protein, as expected for a typical loss-of-function allele[8]. However, potentially weighing against usage of PVS1 is uncertainty in the magnitude of effect for a given PE variant and the relative dosage sensitivity of the affected genes. As noted by Ellingford et al., some PE variants may lead to only a partial effect wherein the variant allele produces both normal and truncated transcripts[44]. Further, the appropriateness of application of PP3 needs to be clarified for PE variants. For the five VUSs described here, we did not apply PP3 as the only available predictors included SpliceAI[48] and CADD[39], neither of which strongly support deleteriousness of these variants (SpliceAI maximum delta score of 0.00-0.31; CADD 13.74-21.6). This is likely due, at least in part, to the absence of PE annotations within the data used by variant impact prediction algorithms.

The pathogenic *SNRPB* variant in Proband F, while outside of standard coding exons, has been reported in at least 10 different patients, and has a confirmed loss-of-function consequence[12]. In the probands described here, it was initially filtered out as it has no coding effect annotation or other features that might protect it from removal. Its CADD score, for example, is only 13.82, which is lower than the vast majority of highly penetrant variants (and, as for other variants here, is likely deflated owing to the lack of

features typically seen for deleterious variants). That said, a “rescue” filter that retains variants with P/LP designations in ClinVar, regardless of their coding status or other annotations, would lead to retention of this variant. Such a rescue would also, however, lead to additional variants in each proband requiring curation, with the magnitude of that addition dependent on the specific ClinVar parameters used (e.g., number and quality of submissions, presence/absence of conflicting interpretations, etc.).

Incorporating the 1,937 PEs and their surrounding introns detailed in Supplementary Data 1 adds minimal burden for variant analysis. These regions contained a median of four variants per proband (maximum of 37) passing through the criteria of having a CADD score of 10 or greater, a GERP score of zero or greater, and an allele count of zero or one in gnomAD or TOPMed. Among these, typically zero or one (median zero, maximum eight) variants were in a gene that has been confidently associated with a NDD phenotype via the Development Disorder Genotype - Phenotype Database[2]. For comparison, if one were to consider variants in all introns, a median of 94 variants per proband (maximum 1,057) would result from the same filtration, with a median of 17 variants (maximum 186) in genes associated with an NDD phenotype. To include all regions detailed in Yan et al., a median of 57 variants per proband (maximum 602) would result from the same variant filtration, with a median of 10 variants (maximum 100) in an NDD gene. Thus, our PE curation process reduces the analytical burden by 10-20X and yields variants with more evidence for disease relevance than more generic intron analyses, making their inclusion in clinical assessment practical.

The work described here provides a list of introns containing PEs that may be relevant to NDDs. Inclusion of these regions in genomic analysis resulted in

identification of five novel VUSs; while uncertain, each of these variants is a compelling candidate for which functional follow-up may lead to reclassification. We also detected one Pathogenic variant that is likely to facilitate a clinical diagnosis. None of these six variants were previously detected and returned by previous genetic testing. The total increased yield among the 2,999 probands analyzed is thus ~0.2%. Further, five of six of these variants were detected in probands with no returned variants, indicating the yield is mostly additive to current processes and is higher than 0.2% among probands with initially negative genome sequencing results.

Beyond increasing diagnostic yield, inclusion of PEs in analysis pipelines may prove beneficial to basic biological research and potential therapeutic research. More comprehensive discovery of patient-associated variation may facilitate improvements in computational effect predictions for variants in or near to PEs. It is also likely, given the abundance of PEs across the genome[15], that new PE-disease associations remain to be discovered. Further, there are recent, ongoing efforts to use antisense oligonucleotide therapy (ASO) to target 20N inclusion in *Scn1a* via intracerebroventricular injection in mice at postnatal day 2[49]. These ASOs resulted in increased expression of *Scn1a* in brain tissue, and also a significant increase in survival of *Scn1a*-mutant mice. Phase 1 and 2 human clinical trials are being conducted for patients with Dravet syndrome (MIM 607208) using ASO therapy to target *SCN1A* mRNA containing the 20N poison exon (NCT04442295). A more comprehensive assessment of patients with PE variants is essential to understand the molecular and biological mechanisms of PE-associated disease, and thereby better understand the potential risks and benefits of PE-directed ASOs. Furthermore, it is possible if not

probable that PE-directed ASOs may be more effective in patients whose disease results from a variant that affects PE inclusion compared to patients whose disease results from non-PE variation; identifying affected individuals with PE variants would be essential to build cohorts for clinical trials to assess this possibility.

In sum, our results provide substantial justification for including PEs in standard clinical genomic analysis to provide both short-term diagnostic and long-term research benefits.

Table and Figure Legends

Table 1. Summary of Cohorts Screened. Probands were screened from previously published cohorts or clinical studies as indicated. All probands screened had suspected early-onset congenital NDD.

Table 2. Poison Exon Variants Discovered. Five VUS and 1 Pathogenic variant were discovered in this analysis. CADD and GERP scores are shown for each variant, as well as the RVIS score for each gene in which the variant resides, along with select clinical and inheritance information.

Figure 1. Extracting Clinically Relevant Poison Exons from Conserved Cassette Exons in Mouse Cortex. Graphic depiction of the methodology implemented to filter differentially spliced cassette exons in mouse cortex to find poison exons relevant to human neurodevelopmental disease. Poison exons are depicted in orange, and canonical exons are depicted in gray.

Figure 2. Variants found within introns containing poison exons in neurodevelopmental disease cohorts. Scale representations of the introns and canonical exons (gray), poison exons (orange), and alternatively spliced canonical exons (blue) in which variants were found. GERP scores are plotted below each and GERP conserved elements are noted (maroon bars).

DATA AVAILABILITY

Supplementary Data and Supplementary Methods

Poison exon data (Supplementary Data 1), hg38 PE cassette BED file (Supplementary Data 2), hg38 introns containing PE BED file (Supplementary Data 3), and variant extraction and supplementary methods including variant extraction, annotation, and filtering code can be found at:

https://github.com/HudsonAlpha/poison_exon_variant_analysis

Genome Sequencing Data

Genome sequencing and phenotype data are available for authorized access and hosted via dbGaP for the HudsonAlpha CSER cohort (phs001089.v4.p1). Genome and phenotype data will be available for authorized access via dbGaP and hosted on AnVIL for SouthSeq (phs002307.v1.p1) and NYCKidSeq (phs002337.v1.p1). Genome sequencing and phenotype data are not available for the AGHI, UDP, PGEN, UDP, or COAGS cohorts, as data sharing requirements were not incorporated into the funding mechanisms and/or consent processes of those cohorts.

ACKNOWLEDGEMENTS

We thank all the families who participated in the studies detailed in Table 1. SAF is supported by NIMH F31MH126628. CSER was supported by grants from the US National Human Genome Research Institute (NHGRI; UM1HG007301). The SouthSeq project (U01HG007301) and the NYCKidSeq project (1U01HG0096108) were supported by the Clinical Sequencing Evidence-Generating Research (CSER2) consortium, which is funded by the National Human Genome Research Institute with co-

funding from the National Institute on Minority Health and Health Disparities and the National Cancer Institute. AnVIL cloud compute credits were provided by grant support to the CSER2 Data Coordinating Center (U24HG007307). AGHI is conducted at the University of Alabama at Birmingham and the HudsonAlpha Institute for Biotechnology and funded by the state of Alabama. The Children's of Alabama Board of Trustees funded the Children's of Alabama Genome Sequencing (COAGS) study. The Alabama Pediatric Genomics Initiative (PGEN) cohort was funded by the Alabama Pediatric Genomics Initiative. We thank the University of Alabama at Birmingham Undiagnosed Diseases Program (UDP) and Dr. Bruce R. Korf. We thank the CSER2 Data Coordinating Center for their work facilitating data sharing and across the CSER Consortium and computation on the AnVIL, particularly Kathleen Ferar and Richard Green. We thank current and former employees of HudsonAlpha, HudsonAlpha Genomic Services Lab and the HudsonAlpha Clinical Services Lab who contributed to sequencing data acquisition and analysis, particularly J. Matthew Holt and David E. Gray.

AUTHOR CONTRIBUTIONS

Conceptualization: S.A.F., G.M.C; Data curation: S.A.F., J.M.J.L.; Formal analysis: S.A.F.; Funding acquisition: S.A.F., E.E.K., G.M.C.; Investigation: S.A.F.; Methodology: S.A.F, J.M.J.L.; Project administration: S.A.F., C.R.F., G.M.C., Resources: E.E.K., G.M.C; Software: S.A.F., J.M.J.L., Z.T.B.; Supervision: G.M.C.; Validation: S.M.H., M.L.T., D.R.L., K.M.B., N.R.K., W.V.K., M.A.K., G.N.; Visualization: S.A.F.; Writing-original draft: S.A.F., J.M.J.L., L.G.H., K.E.B., E.M.B, A.C.E.H.; Writing-review & editing: S.A.F., J.M.J.L., S.M.H., G.M.C.

ETHICS DECLARATION

All studies were conducted under the oversight of institutional review boards documented in Table 1. Informed consent was acquired from all probands and their families, and all information has been de-identified.

CONFLICT OF INTEREST

Disclosure: Dr. Kenny received personal fees from Illumina, 23andMe, and Regeneron Pharmaceuticals, and serves as a scientific advisory board member for Encompass Bio, Foresite Labs, and Galateo Bio. All other authors declare no competing interests.

REFERENCES

1. Ropers HH. Genetics of intellectual disability. Curr Opin Genet Dev. 2008;18(3):241-50. doi: 10.1016/j.gde.2008.07.008

2. DECIPHER (DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources) GRCh37, mapping the clinical genome [Internet] Cambridge: Wellcome Trust Sanger Institute; c2017. [cited 2022 Jan 18]. Available from: https://decipher.sanger.ac.uk

3. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K, Watkins L, Patterson KE, Reinier F, Blue E, Muzny D, Kircher M, Bilguvar K, López-Giráldez F, Sutton VR, Tabor HK, Leal SM, Gunel M, Mane S, Gibbs RA, Boerwinkle E, Hamosh A, Shendure J, Lupski JR, Lifton RP, Valle D, Nickerson DA; Centers for Mendelian Genomics, Bamshad MJ. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015 Aug 6;97(2):199-215. doi: 10.1016/j.ajhg.2015.06.009. Epub 2015 Jul 9. PMID: 26166479; PMCID: PMC4573249.

4. Childerhose JE, Rich C, East KM, Kelley WV, Simmons S, Finnila CR, Bowling K, Amaral M, Hiatt SM, Thompson M, Gray DE, Lawlor JMJ, Myers RM, Barsh GS, Lose EJ, Bebin ME, Cooper GM, Brothers KB. The Therapeutic Odyssey: Positioning Genomic Sequencing in the Search for a Child's Best Possible Life. AJOB Empir Bioeth. 2021;12(3):179-189. doi: 10.1080/23294515.2021.1907475

5. Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, Firth HV, Frazier T, Hansen RL, Prock L, Brunner H, Hoang N, Scherer SW, Sahin M, Miller DT; NDD Exome Scoping Review Work Group. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. Genet Med. 2019;21(11):2413-2421. doi: 10.1038/s41436-019-0554-6.

6. Lee H, Huang AY, Wang LK, Yoon AJ, Renteria G, Eskin A, Signer RH, Dorrani N, Nieves-Rodriguez S, Wan J, Douine ED, Woods JD, Dell'Angelica EC, Fogel BL, Martin MG, Butte MJ, Parker NH, Wang RT, Shieh PB, Wong DA, Gallant N, Singh KE, Tavyev Asher YJ, Sinsheimer JS, Krakow D, Loo SK, Allard P, Papp JC; Undiagnosed Diseases Network, Palmer CGS, Martinez-Agosto JA, Nelson SF. Diagnostic utility of transcriptome sequencing for rare Mendelian diseases. Genet Med. 2020;22(3):490-499. doi: 10.1038/s41436-019-0672-1

7. O'Brien JE, Drews VL, Jones JM, Dugas JC, Barres BA, Meisler MH. Rbfox proteins regulate alternative splicing of neuronal sodium channel SCN8A. Mol Cell Neurosci. 2012;49(2):120-6. doi: 10.1016/j.mcn.2011.10.005

8. Voskobiynyk Y, Battu G, Felker SA, Cochran JN, Newton MP, Lambert LJ, Kesterson RA, Myers RM, Cooper GM, Roberson ED, Barsh GS. Aberrant regulation of a poison exon caused by a non-coding variant in a mouse model of Scn1a-associated epileptic encephalopathy. PLoS Genet. 2021;17(1):e1009195. doi: 10.1371/journal.pgen.1009195.

9. Oh Y, Waxman SG. Novel splice variants of the voltage-sensitive sodium channel alpha subunit, Neuroreport. 1998; 9(7):1267-72. doi: 10.1097/00001756-199805110-00002.

10. Plummer NW, McBurney MW, Meisler MH. (1997) Alternative splicing of the sodium channel SCN8A predicts a truncated two-domain protein in fetal brain and non-neuronal cells. The Journal of Biological Chemistry, 272(38): 24008-15. https://doi.org/10.1074/jbc.272.38.24008. PMID: 9295353

11. Carvill GL, Engel KL, Ramamurthy A, Cochran JN, Roovers J, Stamberger H, Lim N, Schneider AL, Hollingsworth G, Holder DH, Regan BM, Lawlor J, Lagae L, Ceulemans B, Bebin EM, Nguyen J; EuroEPINOMICS Rare Epilepsy Syndrome, Myoclonic-Astatic Epilepsy, and Dravet Working Group, Barsh GS, Weckhuysen S, Meisler M, Berkovic SF, De Jonghe P, Scheffer IE, Myers RM, Cooper GM, Mefford HC. Aberrant Inclusion of a Poison Exon Causes Dravet Syndrome and Related SCN1A-Associated Genetic Epilepsies. Am J Hum Genet. 2018;103(6):1022-1029. doi: 10.1016/j.ajhg.2018.10.023

12. Lynch DC, Revil T, Schwartzentruber J, Bhoj EJ, Innes AM, Lamont RE, Lemire EG, Chodirker BN, Taylor JP, Zackai EH, McLeod DR, Kirk EP, Hoover-Fong J, Fleming L, Savarirayan R; Care4Rare Canada, Majewski J, Jerome-Majewska LA, Parboosingh JS, Bernier FP. Disrupted auto-regulation of the spliceosomal gene SNRPB causes cerebro-costo-mandibular syndrome. Nat Commun. 2014;5:4483. doi: 10.1038/ncomms5483

13. Steward CA, Roovers J, Suner MM, Gonzalez JM, Uszczynska-Ratajczak B, Pervouchine D, Fitzgerald S, Viola M, Stamberger H, Hamdan FF, Ceulemans B, Leroy

P, Nava C, Lepine A, Tapanari E, Keiller D, Abbs S, Sanchis-Juan A, Grozeva D, Rogers AS, Diekhans M, Guigó R, Petryszak R, Minassian BA, Cavalleri G, Vitsios D, Petrovski S, Harrow J, Flicek P, Lucy Raymond F, Lench NJ, Jonghe P, Mudge JM, Weckhuysen S, Sisodiya SM, Frankish A. Re-annotation of 191 developmental and epileptic encephalopathy-associated genes unmasks de novo variants in SCN1A. NPJ Genom Med. 2019;4:31. doi: 10.1038/s41525-019-0106-7

14. Manolio TA, Chisholm RL, Ozenberger B, Roden DM, Williams MS, Wilson R, Bick D, Bottinger EP, Brilliant MH, Eng C, Frazer KA, Korf B, Ledbetter DH, Lupski JR, Marsh C, Mrazek D, Murray MF, O'Donnell PH, Rader DJ, Relling MV, Shuldiner AR, Valle D, Weinshilboum R, Green ED, Ginsburg GS. Implementing genomic medicine in the clinic: the future is here. Genet Med. 2013;15(4):258-67. doi: 10.1038/gim.2012.157.

15. Yan Q, Weyn-Vanhentenryck SM, Wu J, Sloan SA, Zhang Y, Chen K, Wu JQ, Barres BA, Zhang C. Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. Proc Natl Acad Sci U S A. 2015;112(11):3445-50. doi: 10.1073/pnas.1502849112

16. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(Database issue):D493-6. doi: 10.1093/nar/gkh103

17. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2014;30(7):1003-5. doi: 10.1093/bioinformatics/btt637

18. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2. doi: 10.1093/bioinformatics/btq033

19. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006;34(Database issue):D590-8. doi: 10.1093/nar/gkj144

20. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789-98. doi: 10.1093/nar/gku1205

21. Odgis JA, Gallagher KM, Suckiel SA, Donohue KE, Ramos MA, Kelly NR, Bertier G, Blackburn C, Brown K, Fielding L, Lopez J, Aguiniga KL, Maria E, Rodriguez JE, Sebastin M, Teitelman N, Watnick D, Yelton NM, Abhyankar A, Abul-Husn NS, Baum A, Bauman LJ, Beal JC, Bloom T, Cunningham-Rundles C, Diaz GA, Dolan S, Ferket BS, Jobanputra V, Kovatch P, McDonald TV, McGoldrick PE, Rhodes R, Rinke ML, Robinson M, Rubinstein A, Shulman LH, Stolte C, Wolf SM, Yozawitz E, Zinberg RE, Greally JM, Gelb BD, Horowitz CR, Wasserstein MP, Kenny EE. The NYCKidSeq project: study protocol for a randomized controlled trial incorporating genomics into the clinical care of diverse New York City children. Trials. 2021;22(1):56. doi: 10.1186/s13063-020-04953-4

22. Sebastin M, Odgis JA, Suckiel SA, Bonini KE, Di Biase M, Brown K, Marathe P, Kelly NR, Ramos MA, Rodriguez JE, Aguiñiga KL, Lopez J, Maria E, Rodriguez MA, Yelton NM, Cunningham-Rundles C, Gallagher K, McDonald TV, McGoldrick PE, Robinson M, Rubinstein A, Shulman LH, Wolf SM, Yozawitz E, Zinberg RE, Abul-Husn NS, Bauman LJ, Diaz GA, Ferket BS, Greally JM, Jobanputra V, Gelb BD, Horowitz CR, Kenny EE, Wasserstein MP (2021). The TeleKidSeq pilot study: incorporating telehealth into clinical care of children from diverse backgrounds undergoing whole genome sequencing. Under review at Pilot and Feasibility Studies.

23. Bowling KM, Thompson ML, Finnila CR, Hiatt SM, Latner DR, Amaral MD, Lawlor JMJ, East KM, Cochran ME, Greve V, Kelley WV, Gray DE, Felker SA, Meddaugh H, Cannon A, Luedecke A, Jackson KE, Hendon LG, Janani HM, Johnston M, Merin LA, Deans SL, Tuura C, Williams H, Laborde K, Neu MB, Patrick-Esteve J, Hurst ACE, Kandasamy J, Carlo W, Brothers KB, Kirmse BM, Savich R, Superneau D, Spedale SB, Knight SJ, Barsh GS, Korf BR, Cooper GM. Genome sequencing as a first-line diagnostic test for hospitalized infants. Genet Med. 2022;24(4):851-861. doi: 10.1016/j.gim.2021.11.020

24. Green RC, Goddard KAB, Jarvik GP, Amendola LM, Appelbaum PS, Berg JS, Bernhardt BA, Biesecker LG, Biswas S, Blout CL, Bowling KM, Brothers KB, Burke W, Caga-Anan CF, Chinnaiyan AM, Chung WK, Clayton EW, Cooper GM, East K, Evans JP, Fullerton SM, Garraway LA, Garrett JR, Gray SW, Henderson GE, Hindorff LA, Holm IA, Lewis MH, Hutter CM, Janne PA, Joffe S, Kaufman D, Knoppers BM, Koenig BA, Krantz ID, Manolio TA, McCullough L, McEwen J, McGuire A, Muzny D, Myers RM, Nickerson DA, Ou J, Parsons DW, Petersen GM, Plon SE, Rehm HL, Roberts JS,

Robinson D, Salama JS, Scollon S, Sharp RR, Shirts B, Spinner NB, Tabor HK, Tarczy-Hornoch P, Veenstra DL, Wagle N, Weck K, Wilfond BS, Wilhelmsen K, Wolf SM, Wynn J, Yu JH; CSER Consortium. Clinical Sequencing Exploratory Research Consortium: Accelerating Evidence-Based Practice of Genomic Medicine. Am J Hum Genet. 2016 Jun 2;98(6):1051-1066. doi: 10.1016/j.ajhg.2016.04.011. Epub 2016 May 12. Erratum in: Am J Hum Genet. 2016;99(1):246

25. Bowling KM, Thompson ML, Gray DE, Lawlor JMJ, Williams K, East KM, Kelley WV, Moss IP, Absher DM, Partridge EC, Hurst ACE, Edberg JC, Barsh GS, Korf BR, Cooper GM. Identifying rare, medically relevant variation via population-based genomic screening in Alabama: opportunities and pitfalls. Genet Med. 2021;23(2):280-288. doi: 10.1038/s41436-020-00976-z

26. Davis BH, Beasley TM, Amaral M, Szaflarski JP, Gaston T, Perry Grayson L, Standaert DG, Bebin EM, Limdi NA; UAB CBD Study Group (includes all the investigators involved in the UAB EAP CBD program). Pharmacogenetic Predictors of Cannabidiol Response and Tolerability in Treatment-Resistant Epilepsy. Clin Pharmacol Ther. 2021;110(5):1368-1380. doi: 10.1002/cpt.2408

27. Hayeems RZ, Luca S, Hurst ACE, Cochran M, Owens C, Hossain A, Chad L, Meyn MS, Pullenayegum E, Ungar WJ, Bick D. Applying the Clinician-reported Genetic testing Utility InDEx (C-GUIDE) to genome sequencing: further evidence of validity. Eur J Hum Genet. 2022;30(12):1423-1431. doi: 10.1038/s41431-022-01192-w

28. Amendola LM, Berg JS, Horowitz CR, Angelo F, Bensen JT, Biesecker BB, Biesecker LG, Cooper GM, East K, Filipski K, Fullerton SM, Gelb BD, Goddard KAB, Hailu B, Hart R, Hassmiller-Lich K, Joseph G, Kenny EE, Koenig BA, Knight S, Kwok

PY, Lewis KL, McGuire AL, Norton ME, Ou J, Parsons DW, Powell BC, Risch N, Robinson M, Rini C, Scollon S, Slavotinek AM, Veenstra DL, Wasserstein MP, Wilfond BS, Hindorff LA; CSER consortium, Plon SE, Jarvik GP. The Clinical Sequencing Evidence-Generating Research Consortium: Integrating Genomic Sequencing in Diverse and Medically Underserved Populations. Am J Hum Genet. 2018;103(3):319-327. doi: 10.1016/j.ajhg.2018.08.007

29. Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, Culotti A, Ellrott K, Goecks J, Grossman RL, Hall IM, Hansen KD, Lawson J, Leek JT, Luria AO, Mosher S, Morgan M, Nekrutenko A, O'Connor BD, Osborn K, Paten B, Patterson C, Tan FJ, Taylor CO, Vessio J, Waldron L, Wang T, Wuichet K. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. Cell Genom. 2022;2(1):100085. doi: 10.1016/j.xgen.2021.100085

30. Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, Bronner MP, Underhill HR, Quinlan AR. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. Genome Med. 2020;12(1):62. doi: 10.1186/s13073-020-00761-2

31. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73. doi: 10.1093/bioinformatics/btq559

32. Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, Hernaez M, Hudson ME, Kalmbach MT, Klee EW, Mattson NR, Ross CA, Taschuk M, Wieben ED, Wiepert M, Wildman DE, Mainzer LS. Sentieon DNASeq Variant Calling

Workflow Demonstrates Strong Computational Performance and Accuracy. Front Genet. 2019;10:736. doi: 10.3389/fgene.2019.00736

33. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).

34. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018;15(8):591-594. doi: 10.1038/s41592-018-0051-x

35. Odgis JA, Gallagher KM, Rehman AU, Marathe PN, Bonini KE, Sebastin M, Di Biase M, Brown K, Kelly NR, Ramos MA, Thomas-Wilson A, Guha S, Okur V, Ganapathi M, Elkhoury L, Edelmann L, Zinberg RE, Abul-Husn NS, Diaz GA, Greally JM, Suckiel SA, Jobanputra V, Horowitz CR, Kenny EE, Wasserstein MP, Gelb BD. (2022). Detection of mosaic variants using genome sequencing in a diverse pediatric cohort. In Press. AJMG Part A.

36. Muenzen KD, Amendola LM, Kauffman TL, Mittendorf KF, Bensen JT, Chen F, Green R, Powell BC, Kvale M, Angelo F, Farnan L, Fullerton SM, Robinson JO, Li T, Murali P, Lawlor JMJ, Ou J, Hindorff LA, Jarvik GP, Crosslin DR. Lessons learned and recommendations for data coordination in collaborative research: The CSER consortium experience. HGG Adv. 2022;3(3):100120. doi: 10.1016/j.xhgg.2022.100120

37. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008. doi: 10.1093/gigascience/giab008

38. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122. doi: 10.1186/s13059-016-0974-4

39. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13(1):31. doi: 10.1186/s13073-021-00835-9

40. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12):e1001025. doi: 10.1371/journal.pcbi.1001025

41. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME; Genome Aggregation Database Consortium, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-443. doi: 10.1038/s41586-020-2308-7

42. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, Pitsillides AN, LeFaive J, Lee SB, Tian X, Browning

BL, Das S, Emde AK, Clarke WE, Loesch DP, Shetty AC, Blackwell TW, Smith AV, Wong Q, Liu X, Conomos MP, Bobo DM, Aguet F, Albert C, Alonso A, Ardlie KG, Arking DE, Aslibekyan S, Auer PL, Barnard J, Barr RG, Barwick L, Becker LC, Beer RL, Benjamin EJ, Bielak LF, Blangero J, Boehnke M, Bowden DW, Brody JA, Burchard EG, Cade BE, Casella JF, Chalazan B, Chasman DI, Chen YI, Cho MH, Choi SH, Chung MK, Clish CB, Correa A, Curran JE, Custer B, Darbar D, Daya M, de Andrade M, DeMeo DL, Dutcher SK, Ellinor PT, Emery LS, Eng C, Fatkin D, Fingerlin T, Forer L, Fornage M, Franceschini N, Fuchsberger C, Fullerton SM, Germer S, Gladwin MT, Gottlieb DJ, Guo X, Hall ME, He J, Heard-Costa NL, Heckbert SR, Irvin MR, Johnsen JM, Johnson AD, Kaplan R, Kardia SLR, Kelly T, Kelly S, Kenny EE, Kiel DP, Klemmer R, Konkle BA, Kooperberg C, Köttgen A, Lange LA, Lasky-Su J, Levy D, Lin X, Lin KH, Liu C, Loos RJF, Garman L, Gerszten R, Lubitz SA, Lunetta KL, Mak ACY, Manichaikul A, Manning AK, Mathias RA, McManus DD, McGarvey ST, Meigs JB, Meyers DA, Mikulla JL, Minear MA, Mitchell BD, Mohanty S, Montasser ME, Montgomery C, Morrison AC, Murabito JM, Natale A, Natarajan P, Nelson SC, North KE, O'Connell JR, Palmer ND, Pankratz N, Peloso GM, Peyser PA, Pleiness J, Post WS, Psaty BM, Rao DC, Redline S, Reiner AP, Roden D, Rotter JI, Ruczinski I, Sarnowski C, Schoenherr S, Schwartz DA, Seo JS, Seshadri S, Sheehan VA, Sheu WH, Shoemaker MB, Smith NL, Smith JA, Sotoodehnia N, Stilp AM, Tang W, Taylor KD, Telen M, Thornton TA, Tracy RP, Van Den Berg DJ, Vasan RS, Viaud-Martinez KA, Vrieze S, Weeks DE, Weir BS, Weiss ST, Weng LC, Willer CJ, Zhang Y, Zhao X, Arnett DK, Ashley-Koch AE, Barnes KC, Boerwinkle E, Gabriel S, Gibbs R, Rice KM, Rich SS, Silverman EK, Qasba P, Gan W; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Papanicolaou

GJ, Nickerson DA, Browning SR, Zody MC, Zöllner S, Wilson JG, Cupples LA, Laurie CC, Jaquish CE, Hernandez RD, O'Connor TD, Abecasis GR. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590(7845):290-299. doi: 10.1038/s41586-021-03205-y

43. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24. doi: 10.1038/gim.2015.30

44. Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, Campbell C, Downes K, Ellard S, Duff-Farrier C, FitzPatrick DR, Greally JM, Ingles J, Krishnan N, Lord J, Martin HC, Newman WG, O'Donnell-Luria A, Ramsden SC, Rehm HL, Richardson E, Singer-Berk M, Taylor JC, Williams M, Wood JC, Wright CF, Harrison SM, Whiffin N. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Med. 2022;14(1):73. doi: 10.1186/s13073-022-01073-3

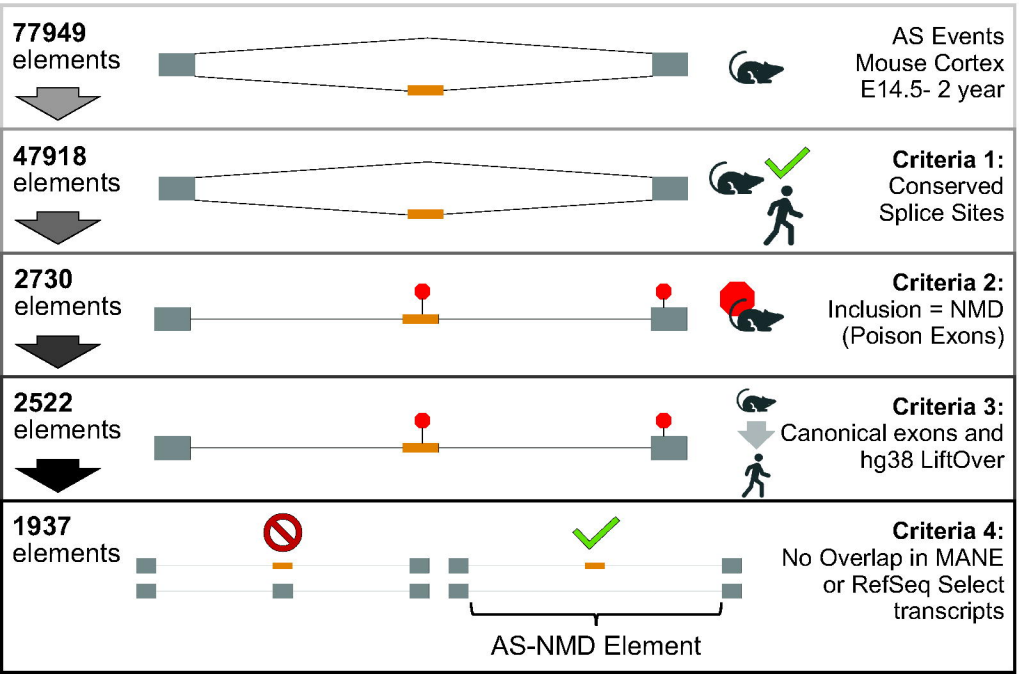
45. Abou Tayoun AN, Pesaran T, DiStefano MT, Oza A, Rehm HL, Biesecker LG, Harrison SM; ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI). Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum Mutat. 2018;39(11):1517-1524. doi: 10.1002/humu.23626

46. Freeman, PJ, Hart, RK, Gretton, LJ, Brookes, AJ, Dalgleish, R. VariantValidator: Accurate validation, mapping and formatting of sequence variation descriptions. *Human Mutation*. 2018; 39: 61- 68. doi: 10.1002/humu.23348

47. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-5. doi: 10.1038/ng.2892

48. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ, Farh KK. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-548.e24. doi: 10.1016/j.cell.2018.12.015

49. Han Z, Chen C, Christiansen A, Ji S, Lin Q, Anumonwo C, Liu C, Leiser SC, Meena, Aznarez I, Liau G, Isom LL. Antisense oligonucleotides increase Scn1a expression and reduce seizures and SUDEP incidence in a mouse model of Dravet syndrome. Sci Transl Med. 2020;12(558):eaaz6100. doi: 10.1126/scitranslmed.aaz6100



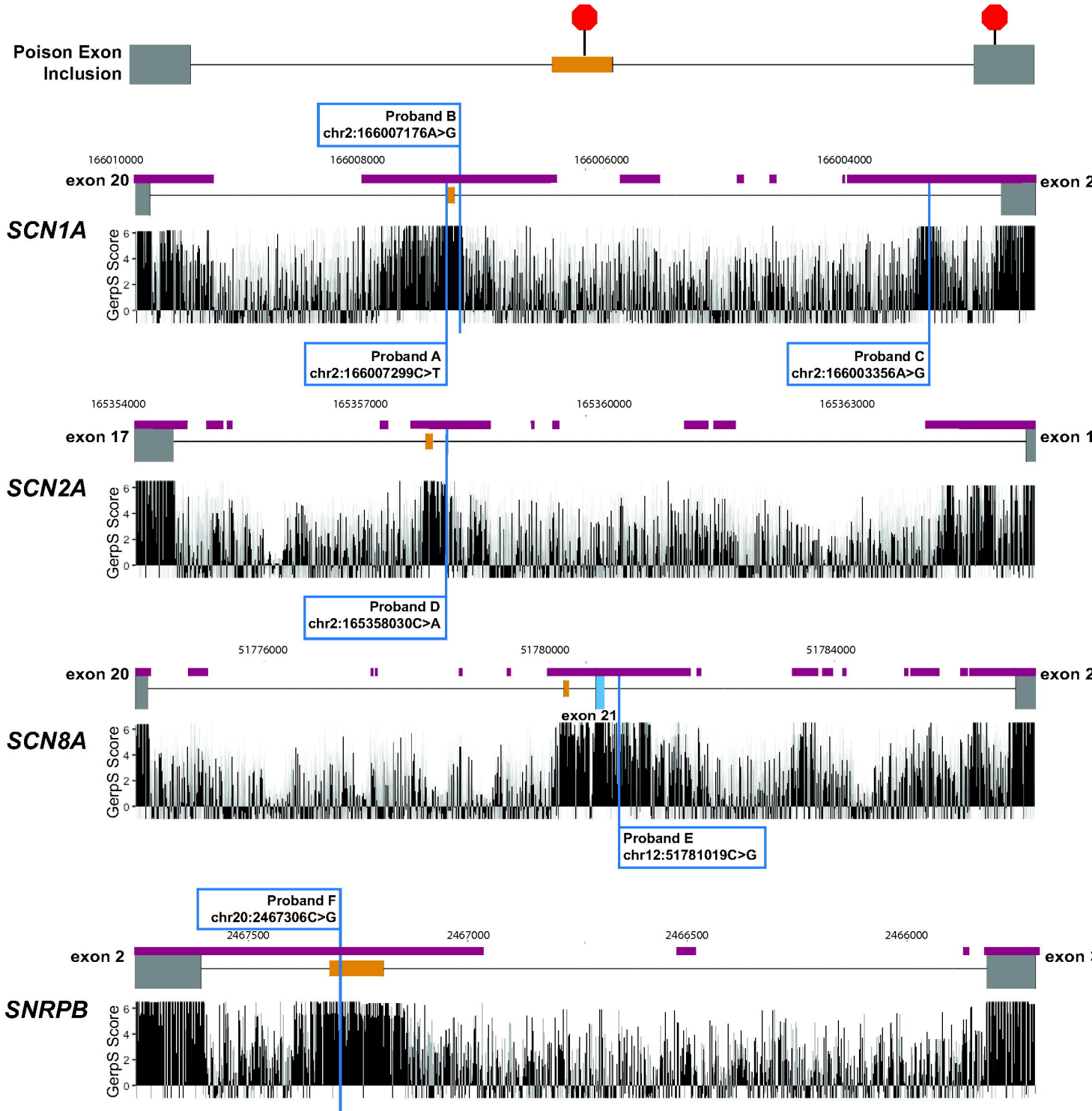


Table 1. Summary of Cohorts Screened

Cohort Name	Reference	IRB Number	IRB Number
NYCKidSeq	21,22	Icahn School of Medicine at Mount Sinai IRB: (STUDY-17-00780/STUDY-20- 01353); Albert Einstein College of Medicine/Montefiore Medical Center IRB: (STUDY-17-00780/STUDY-20-01353)	1032[a]
SouthSeq	23	University of Alabama in Birmingham IRB: IRB-300000328	636
CSER1	24	University of Washington IRB: 44776	532
AGHI	25	University of Alabama in Birmingham IRB: IRB-170303004	484
CBD	26	University of Alabama at Birmingham IRB: IRB-140905010/IRB-140826007	113
Children's of Alabama	27	University of Alabama at Birmingham IRB: IRB 17-0314004	98[b]
PGEN	N/A (clinical study)	Western IRB: 00071	78
UDP	N/A (clinical study)	University of Alabama at Birmingham IRB: IRB 170303004	26
		Total	2999

[a]Total number of probands enrolled in study is 1052
[b]Total number of probands enrolled at time of writing is 146

Table 2. Poison Exon Variants Discovered

	Proband A	Proband B	Proband C	Proband D	Proband E	Proband F
Proband Age at Enrollment (years)	12	3	1	6	9	0
Proband Sex	M	M	F	M	F	M
Phenotype Summary	Developmental Delay; Intellectual Disability; Seizures	Developmental Delay; Seizures; Spastic Dystonic Cerebral Palsy with Left Triplegia	Seizures, impaired awareness	Developmental Delay; Intellectual Disability; Seizures	Seizures; Impaired awareness	Rib Anomalies; Clinodactyly of 4th Toe; Low Set Ears; Atrial Septal Defect; Ventricular Septal Defect
Gene (RVIS Score)	SCN1A (4.03%)	SCN1A (4.03%)	SCN1A (4.03%)	SCN2A (1.77%)	SCN8A (2.34%)	SNRPB (33.97%)
Variant (hg38)	chr2:166007299C>T	chr2:166007176T>C	chr2:166003356A>G	chr2:165358030C>A	chr12:51781019C>G	chr20:2467306C>G
HGVS	NM_001165963.4: c.4002+2420G>A	NM_001165963.4: c.4002+2543A>G	NM_001165963.4: c.4003-603T>C	NM_001040142.2: c.3399+3359C>A	NM_001330260.2: c.3942+248C>G	NM_003091.4: c.155+301G>C
Zygosity	Heterozygous	Heterozygous	Heterozygous	Heterozygous	Heterozygous	Heterozygous
AS-NMD Element	20N poison exon	20N poison exon	20N poison exon	17A poison exon	18N poison exon or exon 21	alternative exon 3 poison exon
ACMG Classification	VUS	VUS	VUS	VUS	VUS	Pathogenic
CADD Score	16.92	20.2	21.6	13.74	17.95	13.82
GerpS Score	3.08	5.37	6.46	1.71	4.45	3.52
Cohort	COAGS	AGHI	AGHI	NYCKidSeq	NYCKidSeq	SouthSeq
Inheritance	Inherited from mother with mild cognitive disabilities	Inherited from unaffected parent	Inherited from affected parent	Unknown	Unknown	Unknown
Notes		Also present in affected sibling	Also present in affected sibling. Another sibling affected, but genotype is unknown.			This variant has been described in similar probands in Lynch et al., 2014