Trio RNA sequencing in a cohort of medically complex children

Authors

Ashish R. Deshwar, Kyoko E. Yuki, Huayun Hou, ..., Lianna Kyriakopoulou, Gregory Costain, James J. Dowling

Correspondence

gregory.costain@sickkids.ca (G.C.), james.dowling@sickkids.ca (J.J.D.)

In this study we applied a trio RNA sequencing approach in a cohort of children with medical complexity who previously underwent genome sequencing. We find that trio analysis, while sometimes helpful in ruling out false-positive RNA-level aberrations, did not increase the diagnostic yield beyond singleton analysis.

Deshwar et al., 2023, The American Journal of Human Genetics *110*, 895–900
May 4, 2023
https://doi.org/10.1016/j.ajhg.2023.03.006

CellPress

Trio RNA sequencing in a cohort of medically complex children

Ashish R. Deshwar,[1,2,14] Kyoko E. Yuki,[2,14] Huayun Hou,[2,14] Yijing Liang,[3] Tayyaba Khan,[2] Alper Celik,[3] Arun Ramani,[3] Roberto Mendoza-Londono,[1,13] Christian R. Marshall,[4,5] Michael Brudno,[6] Adam Shlien,[2,5] M. Stephen Meyn,[7,8] Robin Z. Hayeems,[9] Brandon J. McKinlay,[10] Panagiota Klentrou,[10] Michael D. Wilson,[2,11] Lianna Kyriakopoulou,[4,5] Gregory Costain,[1,2,11,13,*] and James J. Dowling[2,11,12,13,*]

Summary

Genome sequencing (GS) is a powerful test for the diagnosis of rare genetic disorders. Although GS can enumerate most non-coding variation, determining which non-coding variants are disease-causing is challenging. RNA sequencing (RNA-seq) has emerged as an important tool to help address this issue, but its diagnostic utility remains understudied, and the added value of a trio design is unknown. We performed GS plus RNA-seq from blood using an automated clinical-grade high-throughput platform on 97 individuals from 39 families where the proband was a child with unexplained medical complexity. RNA-seq was an effective adjunct test when paired with GS. It enabled clarification of putative splice variants in three families, but it did not reveal variants not already identified by GS analysis. Trio RNA-seq decreased the number of candidates requiring manual review when filtering for *de novo* dominant disease-causing variants, allowing for the exclusion of 16% of gene-expression outliers and 27% of allele-specific-expression outliers. However, clear diagnostic benefit from the trio design was not observed. Blood-based RNA-seq can facilitate genome analysis in children with suspected undiagnosed genetic disease. In contrast to DNA sequencing, the clinical advantages of a trio RNA-seq design may be more limited.

Genome sequencing (GS) is currently the most comprehensive genetic test for the diagnosis of rare Mendelian disorders.[1] However, more than half of individuals with a suspected genetic condition remain undiagnosed after GS. One limitation of contemporary genome analysis is the difficulty of filtering, prioritizing, and interpreting relevant non-coding variants beyond those affecting canonical splice sites. RNA sequencing (RNA-seq) has emerged as a promising technology to help address this issue. Initial studies applying RNA-seq in select cohorts of individuals yielded promising results.[2–12] By contrast, the yield of RNA-seq as a direct complement to GS and with a family-based design has received limited study.

In this study we performed GS and RNA-seq from blood on 97 total individuals from 39 families: 2 quads, 22 trios, 8 duos, and 7 singletons. All probands were children with medical complexity who had previously undergone GS, and a subset of the GS results were reported previously.[13,14] Expression outliers, novel or missing splicing junctions, and putative splicing variants of uncertain significance found by GS were evaluated via RNA-seq in each affected individual. In addition, we utilized the paired GS and RNA-seq data to identify single-nucleotide variants (SNVs) with allele imbalance.

Please see supplemental methods for details on cohort recruitment, GS and RNA-seq methods, and filtering/analysis methods.[7,13,15–22] Written consent was provided by each proband's parents and/or guardians as well as the proband where appropriate. The study was approved by the Research Ethics Board at the Hospital for Sick Children.

We identified gene-expression outliers and aberrant splicing events using blood RNA-seq data of 97 individuals from this cohort and an additional 145 individuals from our internal cohorts (122 individuals with pediatric rare disease and 23 healthy children).[15,16] Our internal control cohort was used instead of GTEx given multiple technical differences (supplemental methods) and was expected to result in a lower false-positive rate. For example, when using over 200 selected high-quality blood RNA-seq datasets from GTEx, we identified a median of 3,096 genes with at least one aberrant splicing event, compared to 1,276 using our internal cohort.

Using OUTRIDER, we identified outlier genes with two different cutoffs based on either p value or Z score.[23]

[1]Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, ON, Canada; [2]Program in Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada; [3]Centre for Computational Medicine, The Hospital for Sick Children, Toronto, ON, Canada; [4]Division of Genome Diagnostics, The Hospital for Sick Children, Toronto, ON, Canada; [5]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; [6]Techna Institute for the Advancement of Technology for Health, University Health Network, Toronto, ON, Canada; [7]Center for Human Genomics and Precision Medicine, University of Wisconsin, Madison, WI, USA; [8]Department of Pediatrics, University of Wisconsin, Madison, WI, USA; [9]Child Health Evaluative Sciences, SickKids Research Institute, Toronto, ON, Canada; [10]Department of Kinesiology, Faculty of Applied Health Sciences, Brock University, St. Catharines, ON, Canada; [11]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada; [12]Division of Neurology, The Hospital for Sick Children, Toronto, ON, Canada; [13]Department of Paediatrics, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
[14]These authors contributed equally
*Correspondence: gregory.costain@sickkids.ca (G.C.), james.dowling@sickkids.ca (J.J.D.)
https://doi.org/10.1016/j.ajhg.2023.03.006.

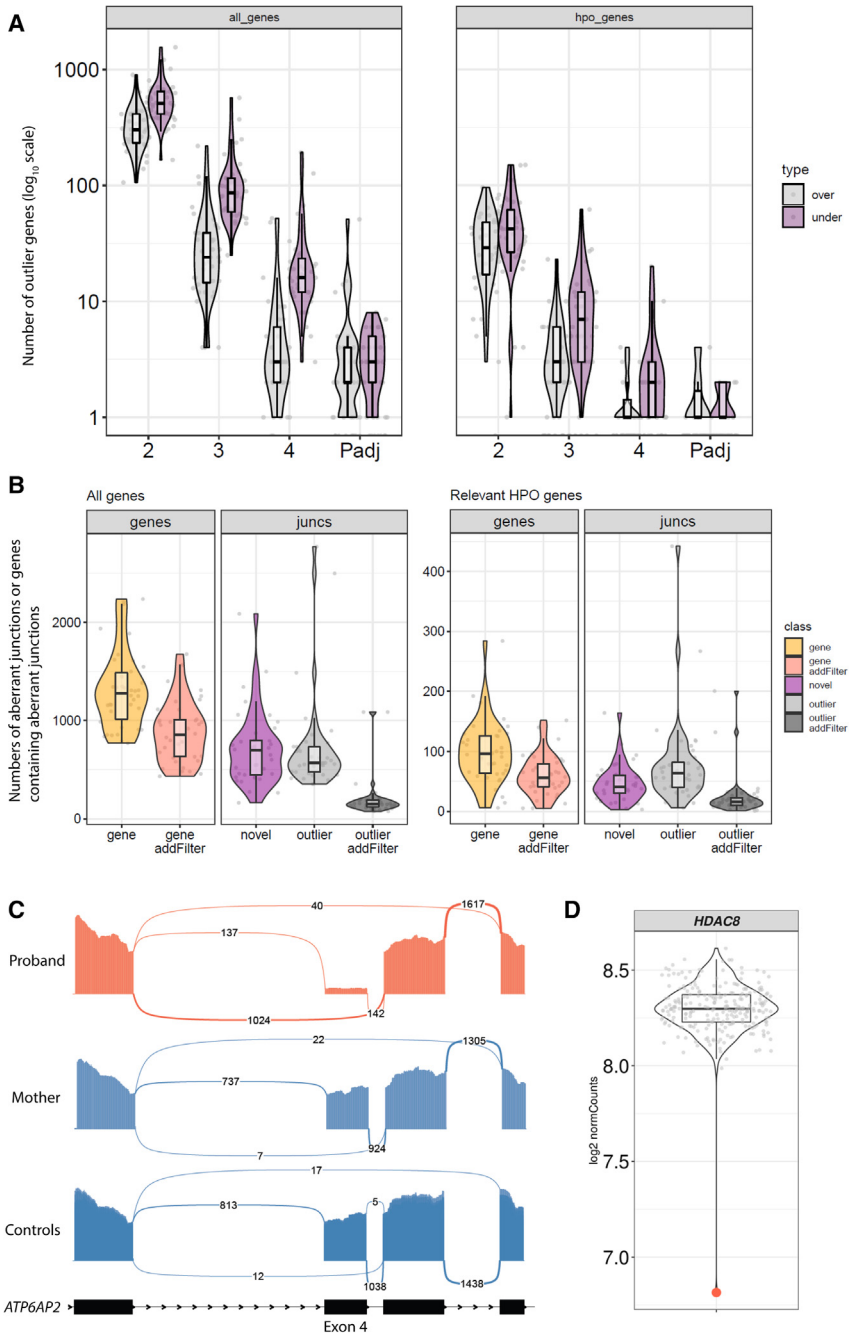

Figure 1. Identification of gene-expression outliers and aberrant junctions in a cohort of medically complex children

(A) Summary of gene-expression outliers using different cut-offs. Scatterplot showing number of over-expressed (gray) or under-expressed (purple) outlier genes, as compared to our in-house control cohort, at absolute Z score ≤ 2, 3, or 4 or adjusted p value (adjusting across all genes in all samples; adjusted p values [Padj]) < 0.05. Each dot represents one sample. Violin and boxplots summarize the distribution of values in each group. In all boxplots the middle line is the median, the box edges are the 25th and 75th percentiles, and the whiskers represent 1.5× the interquartile range. y axis, numbers of outlier genes (log10 scale). Left panel, all genes; right panel, only genes associated with relevant HPO terms.

(B) Summary of aberrant junctions. Scatterplot showing number of reported genes (yellow), genes after additional filters (red), novel junctions (purple), reported outlier junctions (gray), and reported outlier junctions after additional filters (dark gray). Each dot represents one sample. Violin and boxplots summarize the distribution of values in each group. y axis, numbers of aberrant junctions or genes containing aberrant junctions. Left panel, all genes; right panel, only genes associated with relevant HPO terms.

(C) Sashimi plot of representative aberrant junctions in CMC 46 revealing a predominance of transcripts that skipped exon 4 in *ATP6AP2*. The skipping of exon 4 is evident in the proband (red) compared to his mother and to 10 randomly selected controls from the cohort (blue). y axis, number of aligned reads. The number of reads supporting each junction is shown between exons. The minimum number of reads to be drawn was set to 5 for this plot, for better visualization.

(D) Expression level of *HDAC8* in CMC 6 (red dot) showing decreased expression compared to the rest of the samples in the cohort (gray dots). y axis, log2-normalized counts.

With an adjusted p value < 0.05, we identified a median of three under-expressed and two over-expressed outlier genes per proband (Figure 1A). To increase the sensitivity of our assay, we expanded our search to genes with an absolute Z score ≥ 3 without a p value cutoff, resulting in a median of 86 under-expressed and 24 over-expressed outliers. Of these, only a median of six and three under- and over-expressed outlier genes, respectively, were associated with Human Phenotype Ontology (HPO) terms relevant to each individual's phenotypes (Figure 1A).

Using a set of relatively lenient cutoffs (see supplemental methods for more details), our bioinformatics pipeline reported a median of 1,276 genes per sample with novel or outlier junctions. This number was reduced to 856 (corresponding to a median of 676 novel junctions and 150 outlier junctions) after applying additional filters to prioritize outlier junctions (supplemental methods). After selecting genes associated with relevant HPO terms, we were left with a median of 16 outlier junctions and 41 novel junctions in 56 genes (Figure 1B).

Using an RNA-first approach (i.e., blind to GS findings) and filtering for splice junctions and expression outliers as described above, we were able to identify putative diagnostic RNA-level aberrations in 3 of the 39 probands (8%). In two individuals, aberrant splicing events were identified, whereas in the third individual, an expression outlier

was detected. For the two aberrant splicing events, RNA-seq provided functional supporting evidence facilitating reclassification of the corresponding DNA variants as likely pathogenic (supplemental note).

Several findings are of particular note. Proband CMC 27 had a history of global developmental delay, intellectual disability, and epilepsy. In our initial unbiased filtering approach, missing junctions were identified in *SMS* (MIM: 300105). Closer examination of the transcriptome revealed multiple splicing abnormalities, including skipping of both exon 4 alone and exons 3 and 4 as well as the inclusion of intron 3 (Figure S1). None of these aberrant splicing events were observed in the parents. GS had previously identified a *de novo* variant upstream of exon 4 (c.265−5T>A [GenBank: NM_004595.5]); however, there was no consensus between three different *in silico* splicing prediction tools (MaxEnt,[24] −57.6%; NNSPLICE,[25] −10.1%; SSF [http://www.umd.be/searchsplicesite.html], −100%), and the pathogenicity of the variant was initially unclear.

Proband CMC 46 had a history of global developmental delay, intellectual disability, epilepsy, and inflammatory bowel disease. Missing/outlier junctions were identified in *ATP6AP2* (MIM: 300556), with analysis of the transcript revealing a predominance of transcripts that skipped exon 4 (Figure 1C). A similar skew toward transcripts skipping exon 4 was not seen in either parent. GS had previously identified a *de novo* intronic indel (c.301−11_301−10delTT [GenBank: NM_005765.3]). There was no consensus between three different *in silico* splicing prediction tools (MaxEnt, −6.4%; NNSPLICE, −79.9%; SSF, −7.1%), and thus the variant was considered of unknown significance prior to RNA-seq analysis.

Proband CMC 6 had a history of bilateral choanal stenosis, bilateral dysplastic kidneys, dysmorphic features, microcephaly, global developmental delay, chronic lung disease, and intermittent pancytopenia. Gene expression analysis identified decreased *HDAC8* (MIM: 300269) expression with a fold change of 0.36, *Z* score of −10.15, and adjusted p value of 5.82e−12 (Figure 1D). GS had previously identified a *de novo* frameshift variant, c.134_137del (GenBank: NM_018486.3) (p.Ile45Lysfs*9).

Selected variants of uncertain diagnostic significance (VUSs) from the cohort were previously described.[13] These included five putative splicing variants in a total of four genes: *MED23* (MIM: 605042), *PLCB1* (MIM: 607120), *KIF1A* (MIM: 601255), and *JAM3* (MIM: 606871). *MED23* exhibits measurable expression in blood, thus allowing for the targeted analysis of a homozygous c.3939+5G>A (GenBank: NM_004830.4) variant. This variant was classified as "likely pathogenic" when detected by clinical exome sequencing in CMC 01 (ClinVar: SCV000681305.2). There was no consensus between three different *in silico* splicing prediction tools (MaxEnt, −55.9%; NNSPLICE, −40.7%; SSF, −13.8%). No aberrant splicing was observed across the transcript, nor any significant difference in overall or allele-specific expression (ASE). Altogether, this downgrades the variant to a VUS with conflicting evidence (Figure S2). *JAM3* is not expressed in blood, and *KIF1A* and *PLCB1* are not expressed in any of the typical clinically accessible tissues (lymphoblastoid cell lines, fibroblasts, or blood),[17,26] and thus the variants in these genes were not able to be further classified using blood-based RNA-seq.

We sought to determine if a family-based RNA-seq design would facilitate filtering and interpretation of results, as is the case for trio genome-wide sequencing compared with singleton testing.[27] We performed this filtering based on a *de novo* dominant model of inheritance where we wanted to prioritize novel events for further study. When examining gene-expression outliers, we found 186 statistically significant gene-expression outlier events (182 unique genes) in the probands with RNA-seq data from at least one parent. Of these, expression outlier events in 30 genes (in 17 probands) were seen in at least one other family member, compared to 14 genes (in 7 probands) that were seen in at least one other individual from the rest of the cohort, resulting in the ability to exclude 16% vs. 7.5% of identified outliers, respectively (Figures 2A and S3). This suggests that the familial samples may be important for prioritizing statistically significant expression outliers likely due to both the genetic and environmental similarities between probands and their family members. Using OUTRIDER-normalized read counts, the average Pearson correlation between a proband and their family members (median = 0.981) was slightly but significantly higher than that between the proband and the rest of the cohort (median = 0.977, p = 1.756e−06, one-sided paired Wilcoxon test) (Figure 2B). Although different RNA-seq normalization methods could have an impact on this analysis, our observation is consistent with a previous study suggesting familial similarity of the blood transcriptome.[28] When using a more lenient cut-off for expression outliers (absolute *Z* score $\geq$ 3), trio analysis was no longer effective: only 1%–16% of identified genes were seen in at least one other family member, whereas 24%–52% were seen in at least one other internal cohort member (Figure 2A).

We next sought to filter splice junctions using the parental data, again based on a *de novo* dominant model of inheritance. When examining all junction outliers, parental data were not effective in filtering out non-diagnostic splicing events. A median of only 4% of aberrant junctions were seen in at least one family member compared to a median of 35% of aberrant junctions in at least one other individual in the entire cohort (Figures 2C, S4, and S5). This is partially because our pipeline was designed to be highly sensitive at the cost of a higher false-positive rate. In addition, the majority of these aberrant splicing events likely do not have a clear genetic cause and likely have limited biological significance. Indeed, a median of only 10 outlier junctions have at least one rare variant (gnomAD genome and exome allele frequency [AF] < 0.01) nearby (within 10 bp of either end of the junction). We found that 125 of a total of 326 such junctions identified were seen in at least one other family member, compared to 83 seen in other individuals in the cohort

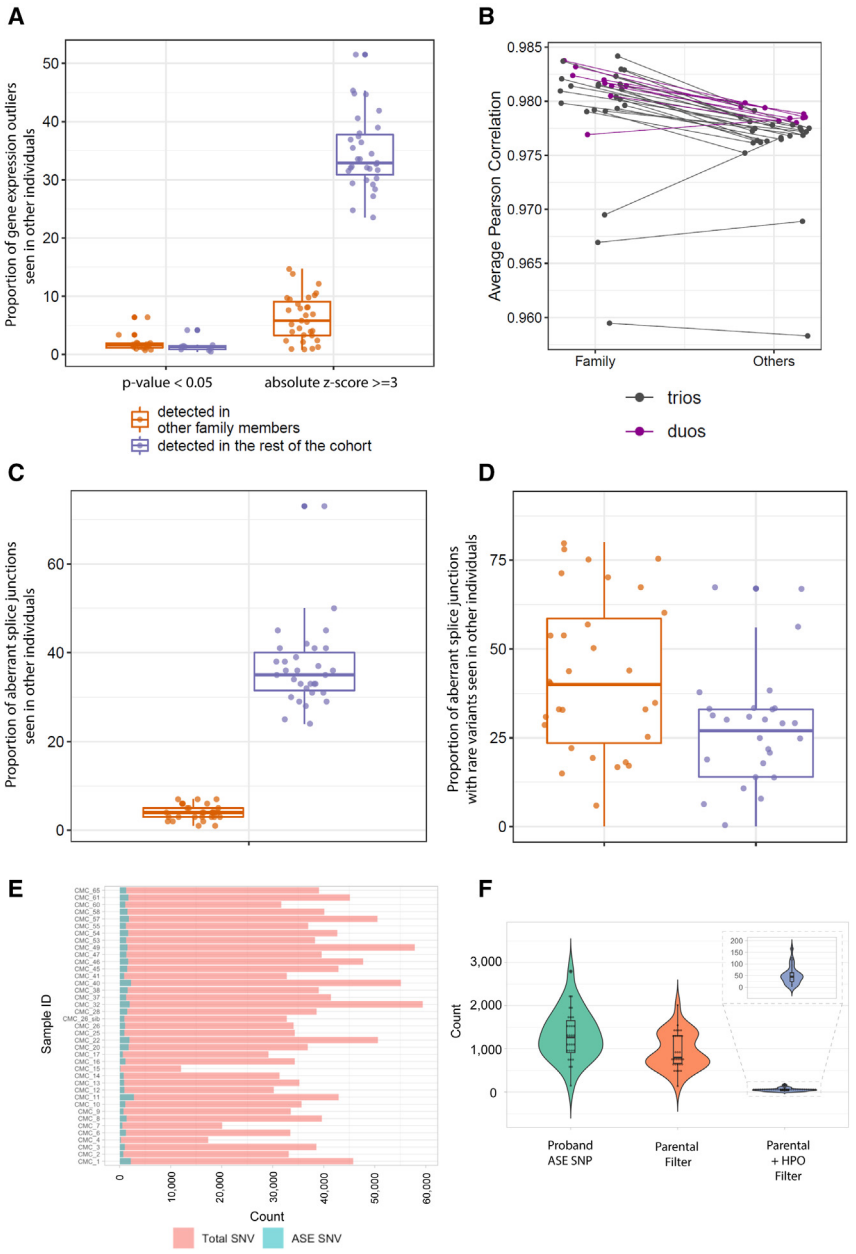

Figure 2. Trio vs. cohort analysis for expression outliers as well as aberrant junctions and allele-specific expression analysis in the cohort

(A) Proportions of gene-expression outliers also detected in other samples showing that trio analysis is effective in filtering out statistically significant expression outliers, but not when using a more lenient cut-off (absolute *Z* score). Each dot represents a proband with family data available. y axis, proportion of gene-expression outlier defined by statistical significance (adjusted p value < 0.05) or *Z* scores (absolute *Z* score ≥ 3) that are also detected in family members (orange) or the rest of the cohort (purple).

(B) Gene expression correlation is consistently higher between probands and their family members than between probands and the rest of the cohort. y axis, average Pearson correlation of gene expression. Each dot represents a proband. Lines connect the same proband in the two columns. Purple, duos; black, trios.

(C) Proportions of total aberrant splicing events also detected in other samples. Each dot represents a proband with family data available. y axis, proportions of genes containing aberrant junctions detected in other family members (orange) or the rest of the cohort (purple).

(D) Proportions of aberrant splicing events with at least one rare variant nearby also detected in other samples. Each dot represents a proband with family data available. y axis, proportions of genes containing aberrant junctions detected in other family members (orange) or the rest of the cohort (purple).

(E) Bar plot of the number of reported SNVs and ASE SNVs for all the affected individuals. Red, total number of rare SNVs; blue, number of ASE events.

(F) Violin plot of the distribution of ASE SNVs after parental filter and HPO term filter.

(Figures 2D and S6). Neither type of trio analysis resulted in the identification of any additional diagnostic variants.

We used our combined GS and RNA-seq datasets to assess each proband for ASE, where two alleles at the same locus are expressed differently (i.e., skew toward one of the alleles). To do this, we prioritized variants that were rare (gnomAD genome AF < 0.01) and exhibited an imbalanced expression between the two alleles. We reasoned that dominant disorders caused by reduced expression of one allele (i.e., haploinsufficiency) would be detected by our expression outlier analysis above, but wondered if we might be able to identify recessive disorders caused by skewing of expression toward a single pathogenic allele.

ASE analysis was performed on 38 probands and 1 affected sibling. Through our analysis pipeline, a median of 38,308 heterozygous SNV sites were reported per affected individual, and a median of 1,263 (mean of 3.4% of all heterozygous SNV sites) were identified as ASE sites after the filters and QC measures (supplemental methods; Figure 2E). Applying an additional parental filter (for probands with parental samples) that removed sites that were reported as having a significant imbalance in either of the parents lowered the number by 27% to 944 ASE sites. In combination with the relevant HPO term filter, the number is reduced to an average of 50 ASE sites per proband (Figure 2F). Manual review of all filtered ASE sites did not yield any additional diagnoses in our cohort.

Altogether, we find that RNA-seq in blood can facilitate the interpretation of GS data. RNA analysis allowed for the confirmation of two putative diagnostic DNA variants (not including one that was already classified as

pathogenic at the DNA level) and the exclusion of one other, thereby resulting in diagnostic utility in 8% (3/39) of the families. Trio RNA-seq analysis did not increase the yield of new diagnoses or candidate variants/genes. Altogether, our data provide further support for the use of singleton RNA-seq as an important diagnostic tool in rare disease.

One caveat of our study is that whole blood was utilized for RNA-seq; this has well-recognized limitations when studying a heterogeneous cohort of individuals with suspected undiagnosed rare genetic diseases.[3] Another limitation of our study is that we utilized the trio RNA-seq data in a targeted way: for filtering out false-positive events based on a *de novo* dominant mode of inheritance. It remains possible that a trio design might facilitate the identification of disease-causing RNA-level aberrations with other modes of inheritance (e.g., inherited dominant-acting variants where there are ASE differences between parent and child causing apparent incomplete penetrance/highly variable expression). Trio RNA-seq may yet show utility if combined with additional novel analysis methods and/or if applied to a different or larger cohort of individuals. One additional caveat of our study is that our bioinformatics pipeline used genome build hg19 and Ensembl v.75, which have both since undergone additional improvements. However, the impact of newer builds on these results is expected to be minimal given that we focused on genes in which variants are known to cause disease, which are expected to already have been well annotated.

In total, these data highlight the value and limitations of blood-based RNA-seq as a clinical diagnostic test for rare genetic disease. Further study is required to better understand its diagnostic utility as a complement to GS in other cohorts of individuals with suspected genetic disorders.

Data and code availability

The data supporting the current study have not been deposited in a public repository to protect individual confidentiality but are available from the corresponding author on request.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.03.006.

Acknowledgments

We gratefully acknowledge all the individuals and their families who participated in this study. We thank the many healthcare providers involved in the diagnosis and care of these children with medical complexity. Special thanks to all staff affiliated with the Complex Care Program and The Centre for Applied Genomics. M.D.W. was supported by the Canada Research Chairs Program. Funding was provided by Genome Canada (OGI-158, M.D.W., A.S., and J.J.D.), the SickKids Centre for Genetic Medicine, the Sickkids Research Institute, and the University of Toronto McLaughlin Centre.

Declaration of interests

The authors declare no competing interests.

Received: November 28, 2022
Accepted: March 8, 2023
Published: March 28, 2023

References

1. Bick, D., Jones, M., Taylor, S.L., Taft, R.J., and Belmont, J. (2019). Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. J. Med. Genet. *56*, 783–791. https://doi.org/10.1136/jmedgenet-2019-106111.
2. Gonorazky, H.D., Naumenko, S., Ramani, A.K., Nelakuditi, V., Mashouri, P., Wang, P., Kao, D., Ohri, K., Viththiyapaskaran, S., Tarnopolsky, M.A., et al. (2019). Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. Am. J. Hum. Genet. *104*, 466–483. https://doi.org/10.1016/j.ajhg.2019.01.012.
3. Murdock, D.R., Dai, H., Burrage, L.C., Rosenfeld, J.A., Ketkar, S., Müller, M.F., Yépez, V.A., Gagneur, J., Liu, P., Chen, S., et al. (2021). Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing. J. Clin. Invest. *131*, e141500. https://doi.org/10.1172/JCI141500.
4. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824. https://doi.org/10.1038/ncomms15824.
5. Yépez, V.A., Gusic, M., Kopajtich, R., Mertes, C., Smith, N.H., Alston, C.L., Ban, R., Beblo, S., Berutti, R., Blessing, H., et al. (2022). Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. *14*, 38. https://doi.org/10.1186/s13073-022-01019-9.
6. Bournazos, A.M., Riley, L.G., Bommireddipalli, S., Ades, L., Akesson, L.S., Al-Shinnag, M., Alexander, S.I., Archibald, A.D., Balasubramaniam, S., Berman, Y., et al. (2022). Standardized practices for RNA diagnostics using clinically accessible specimens reclassifies 75% of putative splicing variants. Genet. Med. *24*, 130–145. https://doi.org/10.1016/j.gim.2021.09.001.
7. Frésard, L., Smail, C., Ferraro, N.M., Teran, N.A., Li, X., Smith, K.S., Bonner, D., Kernohan, K.D., Marwaha, S., Zappala, Z., et al. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. *25*, 911–919. https://doi.org/10.1038/s41591-019-0457-8.
8. Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al. (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*, eaal5209. https://doi.org/10.1126/scitranslmed.aal5209.
9. Maddirevula, S., Kuwahara, H., Ewida, N., Shamseldin, H.E., Patel, N., Alzahrani, F., AlSheddi, T., AlObeid, E., Alenazi, M., Alsaif, H.S., et al. (2020). Analysis of transcript-deleterious variants in Mendelian disorders: implications for RNA-based diagnostics. Genome Biol. *21*, 145. https://doi.org/10.1186/s13059-020-02053-9.
10. Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA

sequencing. Nat. Commun. *8*, 15824. https://doi.org/10.1038/ncomms15824.
11. Wai, H.A., Lord, J., Lyon, M., Gunning, A., Kelly, H., Cibin, P., Seaby, E.G., Spiers-Fitzgerald, K., Lye, J., Ellard, S., et al. (2020). Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. Genet. Med. *22*, 1005–1014. https://doi.org/10.1038/s41436-020-0766-9.
12. Truty, R., Ouyang, K., Rojahn, S., Garcia, S., Colavin, A., Hamlington, B., Freivogel, M., Nussbaum, R.L., Nykamp, K., and Aradhya, S. (2021). Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. Am. J. Hum. Genet. *108*, 696–708. https://doi.org/10.1016/j.ajhg.2021.03.006.
13. Costain, G., Walker, S., Marano, M., Veenma, D., Snell, M., Curtis, M., Luca, S., Buera, J., Arje, D., Reuter, M.S., et al. (2020). Genome sequencing as a diagnostic test in children with unexplained medical complexity. JAMA Netw. Open *3*, e2018109. https://doi.org/10.1001/jamanetworkopen.2020.18109.
14. Haque, B., Khan, T., Ushcatz, I., Curtis, M., Pan, A., Wu, W., Orkin, J., and Costain, G. (2023). Contemporary aetiologies of medical complexity in children: a cohort study. Arch. Dis. Child. *108*, 147–149. https://doi.org/10.1136/archdischild-2022-325094.
15. Lionel, A.C., Costain, G., Monfared, N., Walker, S., Reuter, M.S., Hosseini, S.M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., et al. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet. Med. *20*, 435–443. https://doi.org/10.1038/gim.2017.119.
16. Stavropoulos, D.J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., Nalpathamkalam, T., Pellecchia, G., Yuen, R.K.C., Szego, M.J., et al. (2016). Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. NPJ Genom. Med. *1*, 15012. https://doi.org/10.1038/npjgenmed.2015.12.
17. Walker, S., Lamoureux, S., Khan, T., Joynt, A.C.M., Bradley, M., Branson, H.M., Carter, M.T., Hayeems, R.Z., Jagiello, L., Marshall, C.R., et al. (2021). Genome sequencing for detection of pathogenic deep intronic variation: A clinical case report illustrating opportunities and challenges. Am. J. Med. Genet. *185*, 3129–3135. https://doi.org/10.1002/ajmg.a.62389.
18. Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202–2204. https://doi.org/10.1093/bioinformatics/btv112.
19. van de Geijn, B., Mcvicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods *12*, 1061–1063. https://doi.org/10.1038/nmeth.3582.
20. DePristo, M.A., Banks, E., Poplin, R., Garimella, K. v, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498. https://doi.org/10.1038/ng.806.
21. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.
22. Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol. *16*, 195. https://doi.org/10.1186/s13059-015-0762-6.
23. Brechtmann, F., Mertes, C., Matusevičiūtė, A., Yépez, V.A., Avsec, Ž., Herzog, M., Bader, D.M., Prokisch, H., and Gagneur, J. (2018). OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. Am. J. Hum. Genet. *103*, 907–917. https://doi.org/10.1016/j.ajhg.2018.10.025.
24. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394. https://doi.org/10.1089/1066527041410418.
25. Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. J. Comput. Biol. *4*, 311–323. https://doi.org/10.1089/cmb.1997.4.311.
26. Aicher, J.K., Jewell, P., Vaquero-Garcia, J., Barash, Y., and Bhoj, E.J. (2020). Mapping RNA splicing variations in clinically accessible and nonaccessible tissues to facilitate Mendelian disease diagnosis using RNA-seq. Genet. Med. *22*, 1181–1190. https://doi.org/10.1038/s41436-020-0780-y.
27. Manickam, K., McClain, M.R., Demmer, L.A., Biswas, S., Kearney, H.M., Malinowski, J., Massingham, L.J., Miller, D., Yu, T.W., Hisama, F.M.; and ACMG Board of Directors (2021). Exome and genome sequencing for pediatric patients with congenital anomalies or intellectual disability: an evidence-based clinical guideline of the American College of Medical Genetics and Genomics (ACMG). Genet. Med. *23*, 2029–2037. https://doi.org/10.1038/s41436-021-01242-6.
28. Tremblay, B.L., Guénard, F., Lamarche, B., Pérusse, L., and Vohl, M.-C. (2018). Familial resemblances in human whole blood transcriptome. BMC Genom. *19*, 300. https://doi.org/10.1186/s12864-018-4698-6.

The American Journal of Human Genetics, Volume *110*

Supplemental information

Trio RNA sequencing in a cohort of medically complex children

Ashish R. Deshwar, Kyoko E. Yuki, Huayun Hou, Yijing Liang, Tayyaba Khan, Alper Celik, Arun Ramani, Roberto Mendoza-Londono, Christian R. Marshall, Michael Brudno, Adam Shlien, M. Stephen Meyn, Robin Z. Hayeems, Brandon J. McKinlay, Panagiota Klentrou, Michael D. Wilson, Lianna Kyriakopoulou, Gregory Costain, and James J. Dowling

Supplemental Note:

Interpretation and re-classification of DNA variants

For variant detected in CMC 27, the prior DNA classification was a VUS. Re-classification to likely pathogenic was achieved using PS2 (confirmed *de novo* with parental identity confirmed), PM2_Supporting (absent in gnomAD) and PS3_Supporting (RNA functional evidence detects mis-splicing).

For variant detected in CMC 46, the prior DNA classification was a VUS. Re-classification to likely pathogenic was achieved using PS2 (confirmed *de novo* with parental identity confirmed), PM2_Supporting (absent in gnomAD) and PS3_Supporting (RNA functional evidence detects mis-splicing).

Supplemental Figures and Legends:

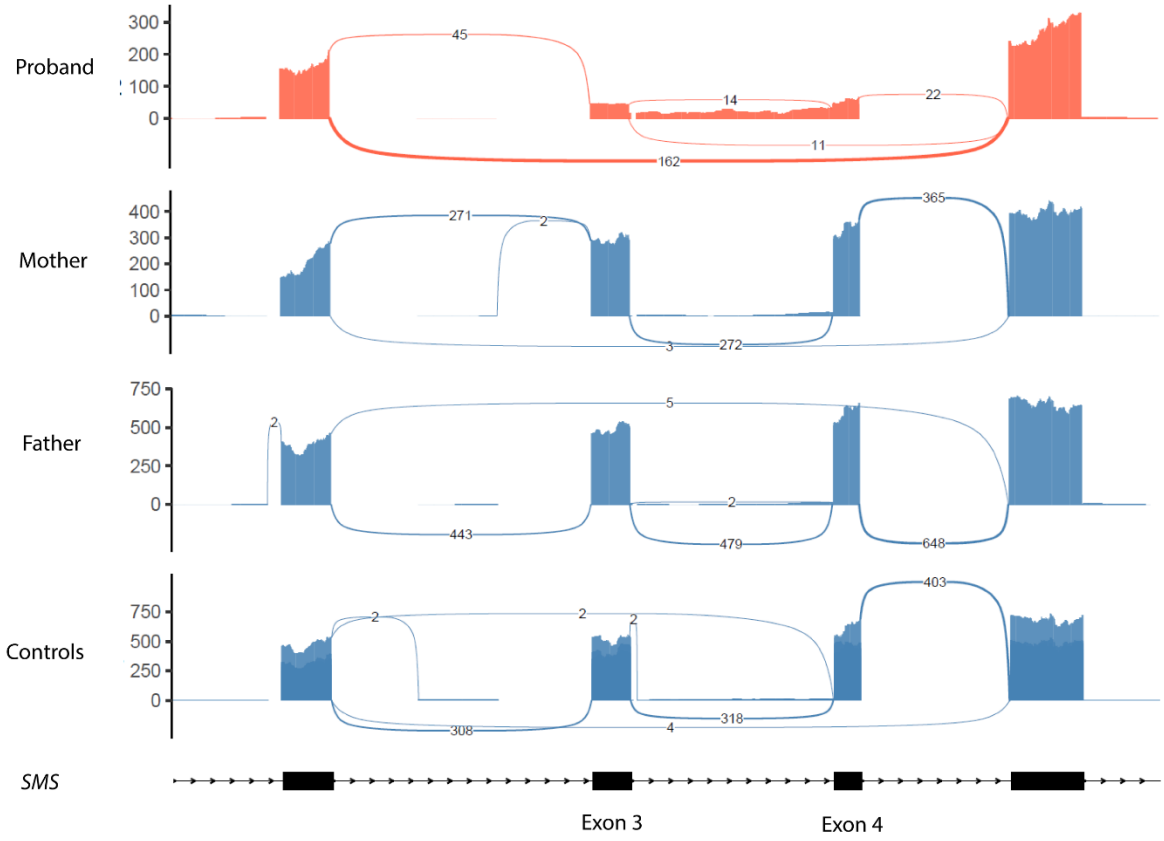

Figure S1 – Sashimi plot of aberrant junctions in CMC 27 in the gene *SMS*.

The skipping of exons 3-4 and retention of intron 3 is evident in the proband (red) compared to his two parents and to 10 randomly selected controls from the cohort (blue). Y axis: number of aligned reads. The number of reads supporting each junction is shown between exons.

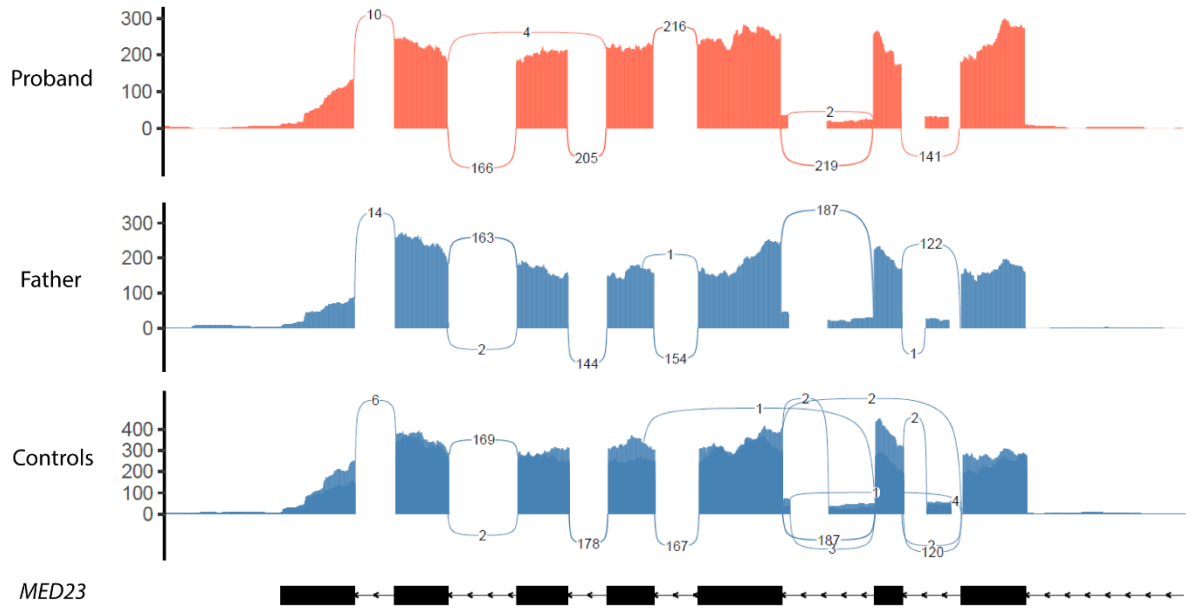

Figure S2 - Sashimi plot of relevant junctions in *MED23* for CMC 01.

No aberrant splicing was observed in the proband (homozygous for the NM_004830.4(MED23):c.3939+5G>A variant; red) when compared to his father (heterozygous for the *MED23* variant) and to 10 randomly selected controls from the cohort (blue). The minimum number of reads was set to 5 for this plot, for better visualization. Only exons near the variant are shown.

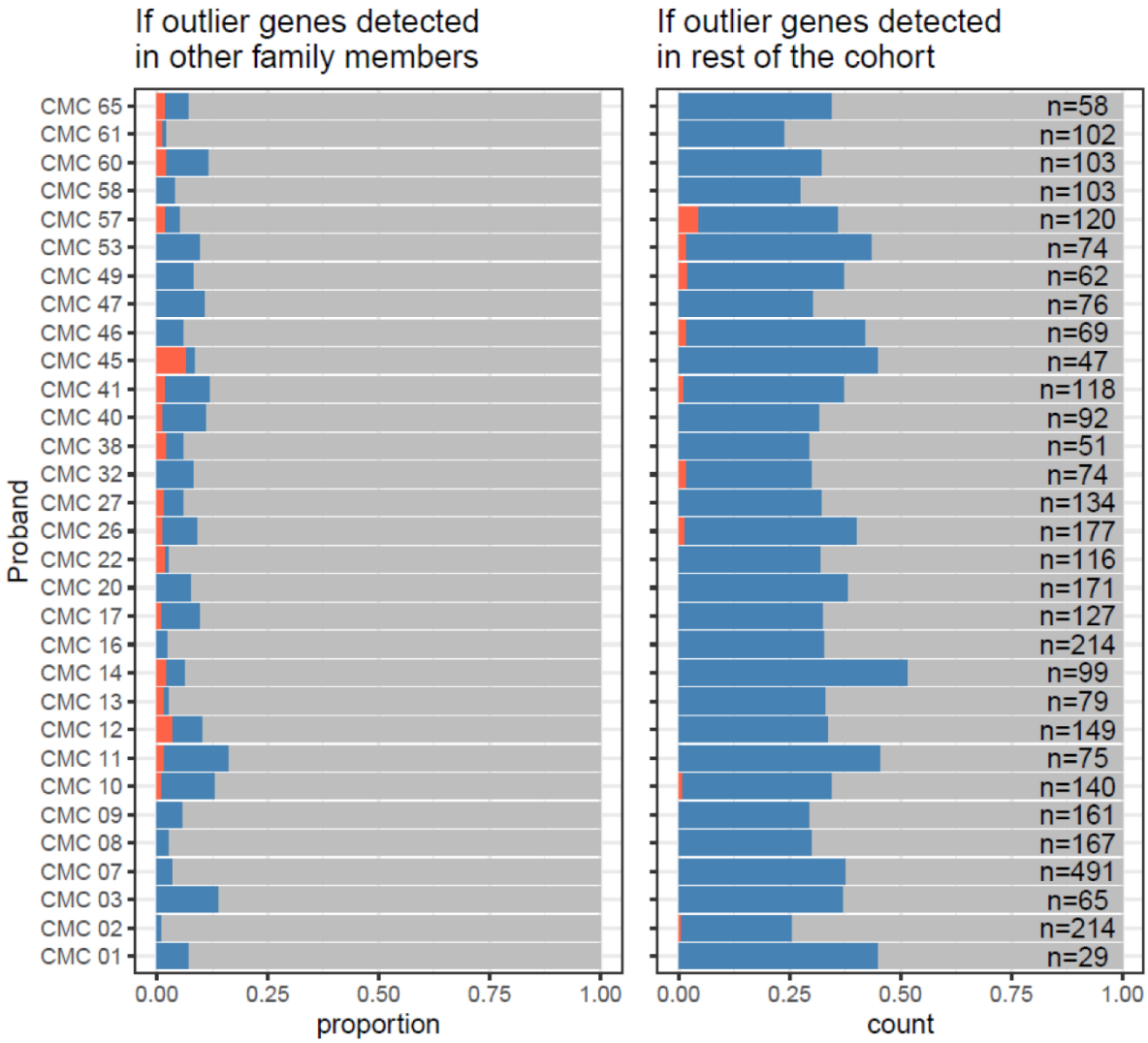

Figure S3 - Individual data for trio vs cohort analysis for expression outliers.

Detection of gene expression outliers in other samples. Each row represents a proband from a different family. Bar plots show the proportion of gene expression outlier defined by statistical significance (adjusted p-value < 0.05, red) or z-scores (absolute z-score >=3, blue) that are also detected in family members (left panel) or in the rest of the cohort (right panel). Total numbers of outlier genes are labelled in the right panel.

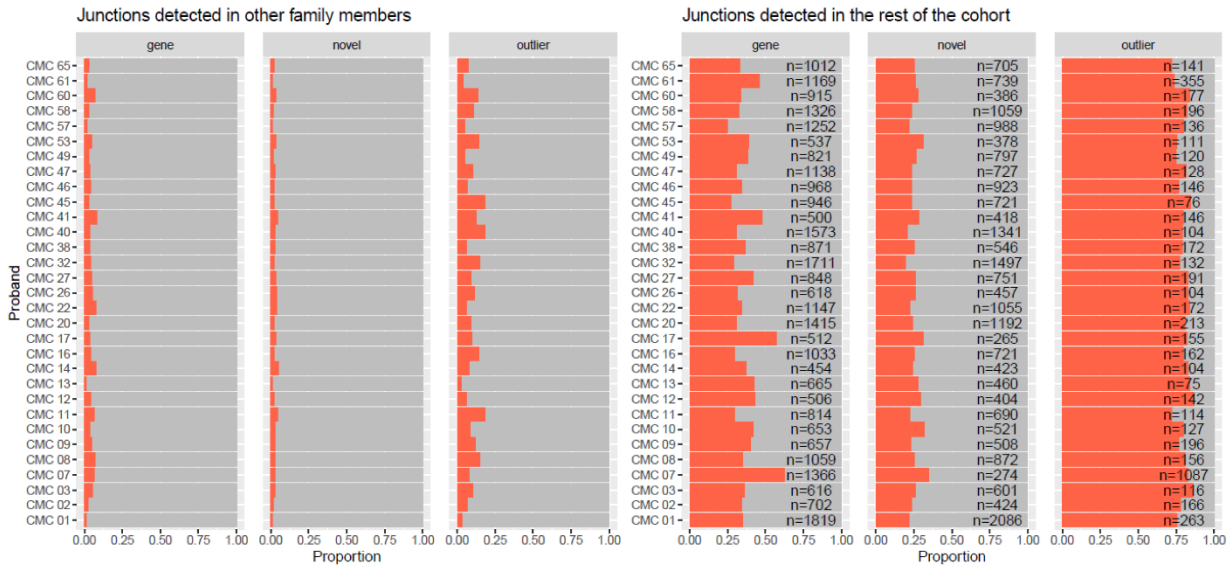

Figure S4 – Individual data for trio vs cohort analysis for aberrant junctions

Detection of aberrant splicing in other samples. Each row represents a proband from a different family. Bar plots show the proportion of novel junctions, outlier junctions, or genes containing aberrant junctions that are detected in family members (left panels) or in the rest of the cohort (right panels). Total numbers of genes/junctions in each proband are labelled in the right panels.

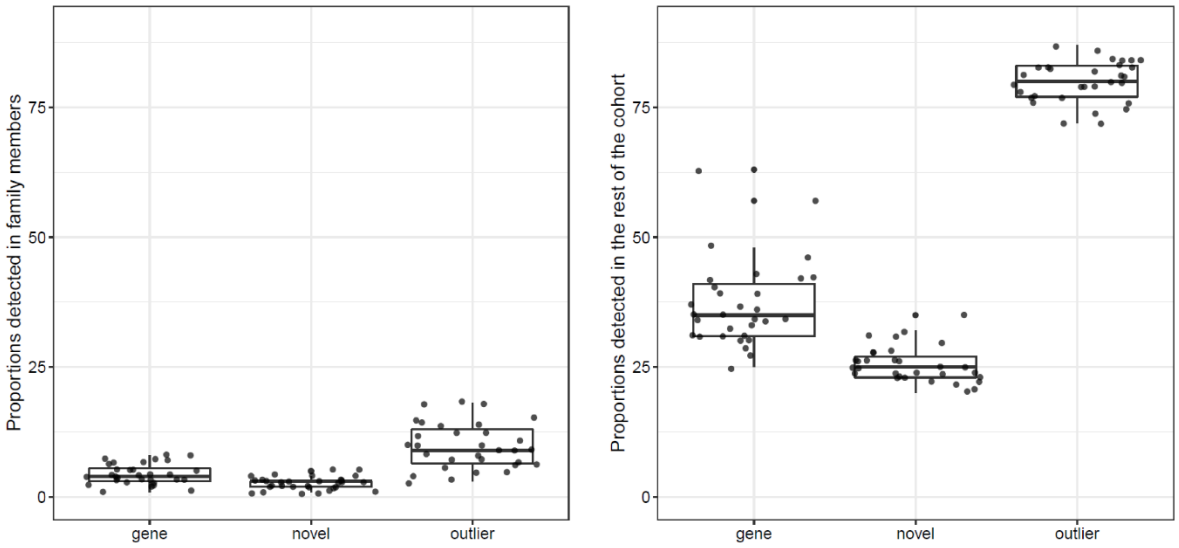

Figure S5 – Proportions of aberrant splicing events detected in other samples by category

Proportions of aberrant splicing events also detected in other samples. Each dot represents a proband with family data available. Y axis: proportions of genes containing aberrant junctions, novel junctions, or outlier junctions detected in other family members (left panel) or the rest of the cohort (right panel).

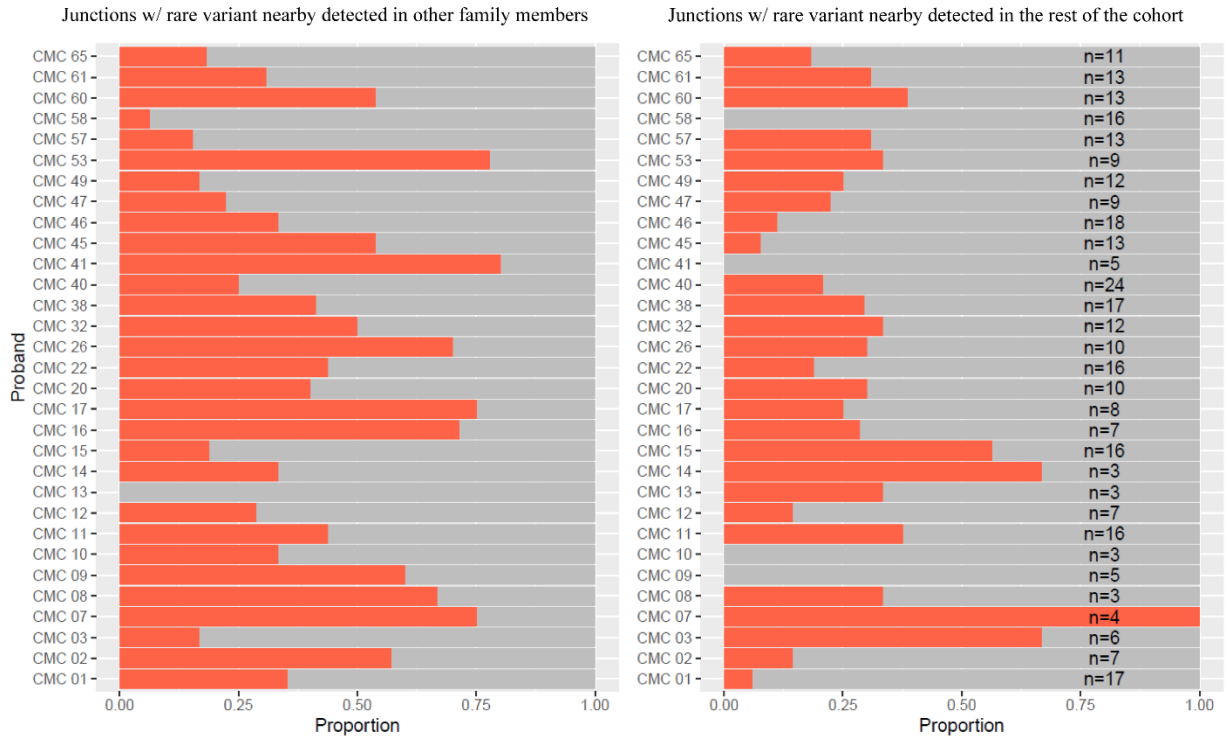

Figure S6 – Individual data for trio vs cohort analysis for aberrant junctions with a rare variant nearby

Detection of aberrant splicing in other samples when a rare variant is within 10bp of either end of a given junction. Each row represents a proband from a different family. Bar plots show the proportion of aberrant junctions with a rare variant nearby detected in family members (left panel) or in the rest of the cohort (right panel).

Supplemental Tables:

Study ID	Sex	Selected Features (HPO terms)	DNA Diagnosis	Gene	Variant Details (Zygosity) [Transcript]	Origin
CMC 27	M	GDD, ID, seizures	Yes	*SMS*	c.265-5T>A (hem) [NM_004595.4]	dn
CMC 53	M	Abnormal autonomic nervous system physiology, gastroparesis, seizures	No			
CMC 54	M	Abnormal heart morphology, tracheal atresia, facial dysmorphism, arachnoid cyst	No			
CMC 55	F	Abnormal heart morphology, pulmonary arterial hypertension, chronic lung disease	Yes	*TBX4*	c.1018C>T; p.(Arg340*) (het) [NM_018488.3]	uk
CMC 57	F	Cloacal exstrophy, abnormality of the kidney	No			
CMC 58	F	GDD, facial dysmorphism, cleft palate	No			
CMC 60	F	Cloacal exstrophy, abnormality of the kidney, anorectal anomaly	No			
CMC 61	M	GDD, spastic tetraplegia, hemophagocytic lymphohistiocytosis	Yes	*ATP7A*	c.4110_4115del; p.(Pro1371_Ile1372del) (hem) [NM_000052.7]	mat
CMC 65	F	GDD, seizures, cortical visual impairment, abnormal muscle tone	Yes	*KCNQ2*	c.634G>T; p.(Asp212Tyr) (het) [NM_004518.6]	dn

Table S1 – Clinical and molecular details for individuals not previously reported.

GDD: global developmental delay, ID: intellectual disability, hem: hemizygous

Supplemental Methods:

Cohort recruitment and phenotyping

Individuals were recruited from a structured Complex Care Program at a tertiary care pediatric hospital (The Hospital for Sick Children)[1]. All individuals were under the age of 18 years at the time of recruitment and were eligible for the study if an underlying genetic condition was suspected but had not yet been established by prior genetic testing. Additional selection criteria for this cohort have been previously described[1]. Phenotypic data were extracted from the electronic medical record and coded in PhenoTips using terms from Human Phenotype Ontology (HPO), a standardized vocabulary for each phenotypic feature.

Genome Sequencing (GS) methods and results

GS was performed using high-quality DNA extracted from blood at The Centre for Applied Genomics (Toronto, Canada), using established methods[2,3]. Genome data filtering and analysis was as previously described[4]. Phenotypic and molecular details of individuals not already described in Costain et al.[1] are described in Table S1. Candidate variants that were deemed relevant to the primary phenotype using established laboratory reporting criteria were discussed with the clinical team and designated as diagnostic by consensus. A genetic diagnosis was attained for 15 of the 39 probands (38 percent) included in this study from GS alone.

RNA extraction and sequencing

Whole blood was collected in PAXGene Blood RNA tubes (BD Biosciences) and total RNA was extracted using the PAXGene Blood RNA Kit (Qiagen). RNA quality and quantity was determined with TapeStation RNA ScreenTape (Agilent). 250 ng of total RNA was spiked with SIRV Set 3 (Lexogen) and was enriched for poly(A). Libraries were prepared using an automated NEBNext Poly(A) mRNA Magnetic Isolation Module and NEBNext Ultra II Directional RNA Library Prep kit by Illumina (New England Biolabs) on the NGS Workstation (Agilent). Libraries were analyzed for quality using TapeStation DNA High Sensitivity ScreenTape (Agilent) and quantified with KAPA library quantification (Roche) prior to sequencing on a NovaSeq6000 (Illumina) with paired-end 150bp runs. While we had originally obtained samples from 40 families, one RNAseq library from one proband (CMC 18) only obtained < 1 million reads and was subsequently

excluded from all analyses. RNAseq data from CMC 15 was of poor quality (very low number of genes detected). This sample was included in all analyses but excluded from gene expression and junction outlier summary plots.

Bioinformatics methods

We developed a customized RNAseq processing pipeline including read alignment, quality control (QC), identification of expression and splicing aberrations, and variant calling. First, raw sequencing reads were aligned to a hybrid genome of hg19 (1000 genomes reference genome, hs37d5) and the spike-in sequences (SIRVome, SIRV set3) using STAR (v2.6.1c)[5]. Gene annotation was obtained from Ensembl (v75) and combined with SIRVome transcript annotation. Fastqc (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, v0.11.5), RNA-seQC (v2.3.5)[6], picard (http://broadinstitute.github.io/picard/, v2.18.0) Markduplicates and CollectRnaSeqMetrics were used to collect various quality metrics, including sequencing quality, duplication rate, percent ribosomal reads, 5'-3' bias, and genomic distribution of the reads. Gene and transcript expression level quantification was performed with RSEM (v1.2.22)[7]. HPO and Orphanet terms associated with each gene were obtained from https://hpo.jax.org/ and https://www.orpha.net/ respectively. OMIM (Online Mendelian Inheritance in Man) disease associations for each gene were obtained from Ensembl v75.

Internal Cohort Selection

For both gene expression and splicing analyses, our large internal cohort of pediatric blood samples was used as a comparison cohort instead of publicly available databases such as the Genotype-Tissue Expression (GTEx). Our internal cohort consists of individuals with pediatric rare disease recruited through the SickKids Genome Clinic (122 individuals with pediatric rare disease) as well as 23 healthy children for a total of 145 individuals[NO_PRINTED_FORM]. The pediatric rare disease cohort consist of 68 male and 54 female children from 0-18 years of age presenting with diverse and complex phenotypes including epilepsy, global developmental delay and multiple congenital anomalies. The healthy children cohort samples were pre-experiment blood samples collected from 10 male and 13 female adolescent athletes who participated in two clinical trials on protein and dairy supplementation as previously described[8,9]. Participants were 12-16 years of age,

free of injuries and any medical conditions that prevented them from participating in their respective trials and were not taking medications or nutritional supplements.

We utilized our own internal cohort for our study given several experimental advantages: 1) we used a clinically validated, standardized experimental protocol with automated library preparation which minimizes technical variations; and 2) our internal cohort contained age-matched individuals which should increase test sensitivity. In addition, the median sequencing depth of our libraries was 113 million reads, which is higher than the targeted 50 million reads of GTEx samples.

Identification of gene expression outliers

Gene expression outliers were identified by comparing each sample to the rest of the samples in the cohort using the R package OUTRIDER (v 1.8.0). An internal cohort of pediatric blood samples were included in the analysis as controls resulting in a total of 243 samples. Gene read counts estimated by RSEM were first filtered for low expressed genes (only genes with >= 10 reads mapped in >= one third of the samples were used) before being used as the input for OUTRIDER analysis. For each sample, genes with either an adjusted p-value < 0.05 or an absolute z-score >= 3 were then extracted as expression outliers. When parental data was available, genes were further prioritized if gene expression in the parents was not considered an outlier. Remaining genes were prioritized based on associations with HPO terms and known OMIM and Orphanet diseases. We reasoned that filtering based on HPO terms would have a higher yield of genes associated with the individual's phenotype.

Identification of aberrant splicing events

To identify aberrant splicing events, we adopted the approach described by Fresard *et al*[10]. Briefly, we used junction quantification by STAR (*SJ.out file). Only junctions that have more than 5 uniquely mapped reads were considered in the analysis. Next, we calculated a junction coverage score, defined as the number of reads mapping to a junction of interest divided by the total number of reads mapping to other junctions that share a splicing donor or acceptor site with the junction of interest. We then calculated the z-score of junction coverage score for each junction within the study cohort. Junctions with an absolute z-score >= 2 were further examined. Aberrant junctions

were identified by first looking for missing or outlier junctions defined as any junction detected in >80 percent of the control cohort with a junction coverage score of >0.6 but with a junction coverage score of <0.75 in the proband. Junctions were further filtered by removing any junction called as aberrant in >5 samples. Novel junctions in genes with a missing or outlier junction were then pulled in for further analysis. Each junction was manually inspected with the Integrative Genomics Viewer (IGV). When parental data was available, aberrant junctions were prioritized if the missing/outlier junction was not called aberrant in the parent or a novel junction was prioritized if the junction was not detected in the parent. Remaining junctions were prioritized based on associations with HPO terms and known OMIM and Orphanet diseases.

Allele specific expression analysis

The genome VCF file was first decomposed using vt (v0.57721)[11] and then extracted for all the heterozygous single nucleotide variants (SNVs). RNAseq reads were aligned to (hg19/GRCh37) using STAR (v2.6.1c) in two-pass mode with --waspOutputMode flag activated for allelic mapping bias correction. This flag enables the WASP algorithm that was introduced in van de Geijn *et al*[12]. The reads that passed the WASP filter were collected and duplicated reads were removed. SNV level allele expression data was generated using the GATK (v3.7.0) ASEReadcounter tool[13,14]. SNV sites that had less than 10 mapped reads, fell within low-mappability regions (UCSC ENCODE 100-mer mappability < 1), had more than 5% of the reads mapped to allele other than the REF and ALT, or where REF/ATL counts from the RNAseq did not support their heterozygosity (FDR < 1%), were excluded in the downstream analysis (as described in Castel *et al*.[15]). For the remaining SNV sites, the binomial p-value was calculated with the expected null ratio and corrected for multiple hypothesis with FDR < 5%. An alternative allele expression ratio was generated by dividing the reads mapped to the reference allele by the total reads at a particular site. GnomAD v2.1.1 allele frequency data was used to annotate each SNV site. Allele specific expression outliers were filtered for SNVs that had an alt ratio ≥0.7 and a gnomAD genome allele frequency <0.01. SNV sites in the proband that showed allele imbalanced expression in either of the parents were further filtered out from the analysis. The remaining SNV sites were prioritized based on associations with HPO terms and known OMIM and Orphanet diseases.

Supplemental References:

1. Costain, G., Walker, S., Marano, M., Veenma, D., Snell, M., Curtis, M., Luca, S., Buera, J., Arje, D., Reuter, M.S., et al. (2020). Genome Sequencing as a Diagnostic Test in Children with Unexplained Medical Complexity. JAMA Netw Open *3*. 10.1001/jamanetworkopen.2020.18109.
2. Stavropoulos, D.J., Merico, D., Jobling, R., Bowdin, S., Monfared, N., Thiruvahindrapuram, B., Nalpathamkalam, T., Pellecchia, G., Yuen, R.K.C., Szego, M.J., et al. (2016). Whole Genome Sequencing Expands Diagnostic Utility and Improves Clinical Management in Pediatric Medicine. NPJ Genom Med *1*. 10.1038/npjgenmed.2015.12.
3. Lionel, A.C., Costain, G., Monfared, N., Walker, S., Reuter, M.S., Hosseini, S.M., Thiruvahindrapuram, B., Merico, D., Jobling, R., Nalpathamkalam, T., et al. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med *20*, 435–443. 10.1038/gim.2017.119.
4. Walker, S., Lamoureux, S., Khan, T., Joynt, A.C.M., Bradley, M., Branson, H.M., Carter, M.T., Hayeems, R.Z., Jagiello, L., Marshall, C.R., et al. (2021). Genome sequencing for detection of pathogenic deep intronic variation: A clinical case report illustrating opportunities and challenges. Am J Med Genet A *185*, 3129–3135. 10.1002/ajmg.a.62389.
5. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. 10.1093/bioinformatics/bts635.
6. DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics *28*, 1530–1532. 10.1093/bioinformatics/bts196.
7. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323. 10.1186/1471-2105-12-323.
8. McKinlay, B.J., Wallace, P.J., Olansky, S., Woods, S., Kurgan, N., Roy, B.D., Josse, A.R., Falk, B., and Klentrou, P. (2022). Intensified training in adolescent female athletes: a crossover study of Greek yogurt effects on indices of recovery. J Int Soc Sports Nutr *19*, 17–33. 10.1080/15502783.2022.2044732.
9. McKinlay, B.J., Theocharidis, A., Adebero, T., Kurgan, N., Fajardo, V.A., Roy, B.D., Josse, A.R., M Logan-Sprenger, H., Falk, B., and Klentrou, P. (2020). Effects of Post-Exercise Whey Protein Consumption on Recovery Indices in Adolescent Swimmers. Int J Environ Res Public Health *17*. 10.3390/ijerph17217761.
10. Frésard, L., Smail, C., Ferraro, N.M., Teran, N.A., Li, X., Smith, K.S., Bonner, D., Kernohan, K.D., Marwaha, S., Zappala, Z., et al. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat Med *25*, 911–919. 10.1038/s41591-019-0457-8.
11. Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202–2204. 10.1093/bioinformatics/btv112.

12. van de Geijn, B., Mcvicker, G., Gilad, Y., and Pritchard, J.K. (2015). WASP: Allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods *12*, 1061–1063. 10.1038/nmeth.3582.
13. DePristo, M.A., Banks, E., Poplin, R., Garimella, K. v, Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491–498. 10.1038/ng.806.
14. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297–1303. 10.1101/gr.107524.110.
15. Castel, S.E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. Genome Biol *16*. 10.1186/s13059-015-0762-6.