medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Detecting missed diagnoses of spinal muscular atrophy in genome, exome, and panel sequencing datasets

Ben Weisburd,[1,2,*] Rakshya Sharma,[1,3] Villem Pata,[4,5] Tiia Reimand,[4,6] Vijay S. Ganesh,[1,2,7,8] Christina Austin-Tse,[1,2] Ikeoluwa Osei-Owusu,[1] Emily O'Heir,[1] Melanie O'Leary,[1] Lynn Pais,[1,8] Seth A. Stafki,[9] Audrey L. Daugherty,[9] Carsten G. Bönnemann,[10] Sandra Donkervoort,[10] Göknur Haliloğlu,[10, 11] Peter B. Kang,[9] Gianina Ravenscroft,[12,13] Nigel Laing,[12,13] Hamish S. Scott,[14] Ana Töpf,[15] Volker Straub,[15] Sander Pajusalu,[4,6] Katrin Õunap,[4,6] Grace Tiao,[1] Heidi L. Rehm,[1,2] Anne O'Donnell-Luria[1,2,8]

1. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
2. Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
3. UC Santa Cruz Genomics Institute, UCSC, Santa Cruz, CA, USA
4. Department of Clinical Genetics, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia
5. Anesthesiology and Intensive Care Clinic, Tartu University Hospital, Tartu, Estonia
6. Genetics and Personalized Medicine Clinic, Tartu University Hospital, Tartu, Estonia
7. Department of Neurology, Brigham & Women's Hospital,Boston, MA, USA
8. Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
9. Greg Marzolf Jr. Muscular Dystrophy Center, Department of Neurology, and Institute for Translational Neuroscience, University of Minnesota, Minneapolis, MN, USA
10. Neuromuscular and Neurogenetic Disorders of Childhood Section, Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, NIH, Bethesda, MD, USA
11. Division of Pediatric Neurology, Department of Pediatrics, Hacettepe University Faculty of Medicine, Ankara, Turkey
12. Centre of Medical Research, The University of Western Australia, Perth, Western Australia, Australia
13. Harry Perkins Institute for Medical Research, Perth, Western Australia, Australia
14. Centre for Cancer Biology, An SA Pathology & UniSA Alliance, Adelaide, SA, Australia
15. John Walton Muscular Dystrophy Research Centre, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

* Corresponding author: Ben Weisburd, weisburd@broadinstitute.org

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Abstract

Spinal muscular atrophy (SMA) is a genetic disorder that causes progressive degeneration of lower motor neurons and the subsequent loss of muscle function throughout the body. It is the second most common recessive disorder in individuals of European descent and is present in all populations. Accurate tools exist for diagnosing SMA from short read and long read genome sequencing data. However, there are no publicly available tools for GRCh38-aligned data from panel or exome sequencing assays which continue to be used as first line tests for neuromuscular disorders. We therefore developed and extensively validated a new tool - SMA Finder - that can diagnose SMA not only in genome, but also exome and targeted sequencing samples aligned to GRCh37, GRCh38, or T2T-CHM13. It works by evaluating aligned reads that overlap the c.840 position of *SMN1* and *SMN2* in order to detect the most common molecular causes of SMA. We applied SMA Finder to 16,626 exomes and 3,911 genomes from heterogeneous rare disease cohorts sequenced at the Broad Institute Center for Mendelian Genomics as well as 1,157 exomes and 8,762 targeted sequencing samples from Tartu University Hospital. SMA Finder correctly identified all 16 known SMA cases and reported nine novel diagnoses which have since been confirmed by clinical testing, with another four novel diagnoses undergoing validation. Notably, out of the 29 total SMA positive cases, 21 had an initial clinical diagnosis of muscular dystrophy, congenital myasthenic syndrome, or congenital myopathy. This underscored the frequency with which SMA can be misdiagnosed as other neuromuscular disorders and confirmed the utility of using SMA Finder to reanalyze phenotypically diverse neuromuscular disease cohorts. Finally, we evaluated SMA Finder on 198,868 individuals that had both exome and genome sequencing data within the UK Biobank (UKBB) and found that SMA Finder's overall false positive rate was less than 1 / 200,000 exome samples, and its positive predictive value (PPV) was 96%**.** We also observed 100% concordance between UKBB exome and genome calls. This analysis showed that, even though it is located within a segmental duplication, the most common causal variant for SMA can be detected with comparable accuracy to monogenic disease variants in non-repetitive regions. Additionally, the high PPV demonstrated by SMA Finder, the existence of treatment options for SMA in which early diagnosis is imperative for therapeutic benefit, as well as widespread availability of clinical confirmatory testing for SMA, may warrant the addition of *SMN1* to the ACMG list of genes with reportable secondary findings after genome and exome sequencing.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Introduction

Spinal muscular atrophy (SMA) is a rare genetic condition characterized by progressive loss of muscle function due to the death of lower motor neurons. It is one of the most common recessive disorders, particularly in individuals of European descent where it affects between 1 in 6,100-11,500,[1] while in other populations it occurs at lower but appreciable rates.[2] In 2016, the US Food and Drug Administration (FDA) approved nusinersen, the first drug to slow disease progression, followed by the approval of onasemnogene abeparvovec-xioi in 2019, and risdiplam in 2020. The American Health Resources & Services Administration (HRSA) added SMA testing to the recommended uniform newborn screening panel (RUSP) in 2018, and it is now included in all 50 states. An increasing number of countries have also included SMA within their national screening programs. However, affected individuals may remain undiagnosed if they were born before testing became available in their region, if their parents declined newborn screening, or due to false-negative test reports.

The molecular etiology of SMA is linked to disruption of the SMN locus,[3] most commonly due to deletions or gene conversions resulting in *SMN1* deficiency. The SMN locus consists of two nearly identical paralogs, *SMN1* and *SMN2*. The 'C' nucleotide at the c.840 position of *SMN1* facilitates proper splicing, while the 'T' at that position in *SMN2* causes exon 7 skipping in 85% of *SMN2* transcripts (**Figure 1A**). As a result, *SMN2* alone does not produce a sufficient amount of functional SMN protein, causing neurons that have zero functional copies of *SMN1* to die prematurely. In the general population, individuals typically inherit two intact copies of each paralog, though genomes with between zero and five copies of each have been observed.[2] Since *SMN2* can still produce a small amount of SMN protein, disease severity and age of onset in affected individuals is influenced by the copy number of *SMN2*.

The near-perfect sequence homology between *SMN1* and *SMN2* causes read alignment algorithms like BWA[4] to either mismap or ambiguously map reads at this locus, often resulting in reads having a mapping quality of 0. This, in turn, confounds standard variant calling pipelines such as GATK[5], and prevents them from being directly useful for SMA diagnosis. To address these issues, multiple specialized tools have been developed to diagnose SMA from genome and long read sequencing data, including SMNCopyNumberCaller[2] and Paraphase.[6] Additionally, the MYO-SEQ study[7] published an SMA calling pipeline for exome data on GRCh37 and used it to identify 5 novel SMA diagnoses in a cohort of 1001 cases with limb-girdle weakness. Most recently, the Chameleolyser tool was described in a study involving GRCh37-aligned exome samples from a cohort of 17,650 undiagnosed patients.[8] There, Chameleolyser successfully identified 15 novel SMA cases. However, to our knowledge, no publicly available tool exists for detecting SMA in exome or targeted sequencing samples on GRCh38 or the new telomere-to-telomere (T2T-CHM13)[9] reference. We therefore created SMA Finder for use with exome, genome, or targeted sequencing data aligned to GRCh37, GRCh38, or T2T-CHM13, and applied it to diagnose unsolved cases within large rare disease cohorts from the Broad Institute Center for Mendelian Genomics (CMG) and Tartu University Hospital.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Materials and Methods

SMA Finder algorithm

SMA Finder works by computing two numbers from the read data, **r** and **N**, where **N** is the total number of aligned reads that overlap the c.840 position of *SMN1* and *SMN2*, and **r** is the subset of these reads that have a 'C' at the c.840 position. When there is sufficient read coverage (**N ≥** 14), SMA Finder interprets the lack of reads with a 'C' at c.840 as evidence that the individual has zero functional copies of *SMN1* (**Figure 1B**, **Supplementary Methods**), and so reports a positive call. **Figure 3** shows how **r** and **N** are used to distinguish between true positive and true negative samples.

Cohort summary

The Broad Institute Center for Mendelian Genomics (CMG) cohort consisted of 20,537 individuals from 10,754 families with a wide spectrum of medical conditions. Families sequenced through the CMG were enrolled at the Broad Institute by the Rare Genomes Project or enrolled in research studies with local regulatory approval through the participating collaborators, including for sharing de-identified samples for sequencing and analysis. This project was approved by the Mass General Brigham IRB (protocols #2016P001422 and #2013P001477). In this cohort, 12,045 (59%) individuals were affected with a variety of suspected monogenic disorders, while 8,401 (41%) were unaffected parents or relatives (**Figure 2A**). We performed ancestry inference for 20,205 individuals by computing principal components for high-quality bi-allelic autosomal SNVs using the gnomAD v2 method.[10] We found that 13,240 (66%) individuals were European and the remaining 34% came from various other populations (**Figure 2B**). Among the 12,045 affected individuals, 10,125 (84%) had phenotype descriptions encoded in Human Phenotype Ontology (HPO) terms.[11] Of these, 5,138 (51%) had at least one HPO term in the "Abnormality of the nervous system" category, and 4,035 (40%) had at least one HPO term in the "Abnormality of the musculoskeletal system" category (**Figure 2C**).

Genome sequencing used PCR-free library preparation and Illumina HiSeq X Ten v2 chemistry to generate 150bp paired-end reads. The mean target coverage was >30x. Exome sequencing used Illumina Nextera or Twist exome capture (~38 Mb target) and similarly generated 150bp paired-end reads. It aimed to cover >80% of targets at 20x and a mean target coverage of >60x. Both genome and exome data were processed using GATK best practices, starting with BWA alignment to GRCh38 and GATK base quality score recalibration (BQSR). The average number of reads overlapping the c.840 position of *SMN1* + *SMN2* was 77.9 across CMG genomes and 110.7 across CMG exomes. These numbers reflect the higher coverage in exomes compared to genomes at this locus, as well as the total copy number of SMN paralogs per individual.

The Tartu University Hospital cohort consisted of 9,919 pseudonymised samples sent for molecular diagnostics for various suspected medical conditions to Tartu University Hospital between 2015 - 2022. The cohort analysis was approved by the University of Tartu Research Ethics Committee (protocol # 374M-6). Out of the 9,919 samples, 1,157 (11.7%) were exomes

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

and 8,762 (88.3%) were targeted sequencing panels generated using the Illumina TruSight™ One or the Illumina TruSight™ One Expanded panel of 4,811 to 6,794 genes. Guaranteed mean coverage for both panels and exomes was >100x for targeted regions. The data was aligned to reference genome build GRCh37 using BWA. Samples were annotated with summary clinical characteristics, and de-pseudonymized data was accessed for positive SMA Finder results only.

The UK Biobank (UKBB)[12] dataset included de-identified genetic and medical records from 198,868 individuals as well as both exome and genome samples for each individual. These samples were aligned to GRCh38 using BWA. The UKBB is a large population-based prospective study that recruited UK residents between 40 and 69 years of age during the years 2006–2010. UKBB genotypes had a mean depth of coverage of 32.5x and a minimum of 23.5x.[13] For UKBB exomes, the depth of coverage, on average, exceeded 20x at 95% of targeted bases.[14] The average number of reads overlapping the c.840 position of *SMN1* + *SMN2* was 56.8 in UKBB genomes and 75.1 in UKBB exomes, again reflecting the higher coverage in exomes at this locus, as well as the total copy number of SMN paralogs per individual.

Our cohorts did not include any samples aligned to the T2T-CHM13 reference. Therefore, to test SMA Finder's performance on T2T-CHM13, we realigned the 21 confirmed-positive exome samples from the CMG cohort to the T2T-CHM13 reference using BWA MEM and ran SMA Finder on the realigned samples. We then repeated these steps on 21 randomly selected exomes from the CMG cohort that SMA Finder had previously called as SMA-negative. After realignment to T2T-CHM13, SMA Finder still reported the 21 positive samples as SMA-positive, and the 21 negative samples as SMA-negative, confirming its utility for T2T-aligned samples.

Results

We applied SMA Finder to 16,626 exomes and 3,911 genomes from phenotypically heterogeneous rare disease cohorts within the Broad CMG,[15] part of the GREGoR Consortium. SMA Finder identified all 13 known SMA-positive samples (10 exomes and 3 genomes), and flagged 10 previously undiagnosed exome samples as candidate SMA cases, of which six have now been validated by gold standard methods such as MLPA. For the remaining novel diagnoses (n=4), further confirmation testing is pending (**Table 1**). SMA Finder was negative for 8,401 unaffected samples (6,500 exomes and 1,901 genomes). SMA Finder reported insufficient read coverage to make a call in 112 exomes (0.7%) and 6 genomes (0.2%) due to these samples having fewer than 14 total reads overlapping c.840 in *SMN1* + *SMN2*. Strikingly, out of the 23 total SMA-positive cases in this heterogeneous rare disease cohort, 17 had an initial clinical diagnosis of muscular dystrophy, two had a clinical diagnosis of congenital myasthenic syndrome, and one had a diagnosis of congenital myopathy.

To evaluate its performance beyond the CMG dataset, we also ran SMA Finder on 9,919 samples (1,157 exomes and 8,762 targeted sequencing samples) from another heterogeneous rare disease cohort at the Tartu University Hospital. SMA Finder detected three previously known cases and three new cases which have since been confirmed by clinical testing (**Table 1**). Of the three new cases, two presented as SMA type III, and one had symptoms that did not include obvious SMA features and thus may be a presymptomatic case of SMA type IV.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Finally, to measure SMA Finder's positive predictive value (PPV) in a large population cohort as well as compare its performance to other tools and across sample types, we applied SMA Finder to 198,868 individuals from the UK Biobank (UKBB).[12] Given the availability of an exome and genome sample for each individual, we ran SMA Finder on all exomes and genomes, while also running SMNCopyNumberCaller[2] on the genomes. We found SMA Finder calls to be identical between exomes and genomes, with the only difference being that SMA Finder reported 85 out of 198,868 (0.042%) genome samples and 14 out of 198,868 (0.007%) exome samples (Fisher's exact $p=1.4x10^{-13}$) as having insufficient read coverage to make a call (**Table S3**). SMA Finder flagged two individuals (i1 and i2) as SMA-positive, with their exome and genome samples yielding the same result. It reported all other exomes and genomes that had sufficient coverage as SMA-negative (**Figure 3E, F**). Concordance between SMA Finder and SMNCopyNumberCaller was also nearly 100%. Like SMA Finder, SMNCopyNumberCaller (a genome-only tool) reported individual i1 to be SMA-positive, while reporting no-call for 35 genomes including i2 (**Table S2**).

The UKBB phenotype records for i1 included ICD-10 code G121 "Other inherited spinal muscular atrophy", suggesting this to be a true positive call for both SMA Finder and SMNCopyNumberCaller. Additionally, i1's records indicated that they died between the age of 65 and 70, with the cause of death listed as "Spinal muscular atrophy". In contrast, i2 had few phenotype records, making it difficult to tell whether this was a false positive call by SMA Finder, or a missed call by SMNCopyNumberCaller. The fact that SMA Finder made the same call for both the exome and genome of i2 increases the likelihood that this is a true positive. However, even if we conservatively count i2 as a potential false positive call by SMA Finder based on the absence of phenotype records and the no-call result from SMNCopyNumberCaller, SMA Finder's overall false positive rate would still be < 1 / 200,000 exomes. Additionally, this would yield a PPV of 96% based on 26 true positive samples (19 from CMG cohorts, 6 from Tartu University Hospital, and 1 from UKBB) and at most 1 hypothetical false positive (counting sample i2 from UKBB).

Discussion

We developed and validated SMA Finder, a new tool for detecting SMA-positive samples within short read exome, genome, and panel sequencing data. After testing this tool on multiple heterogeneous rare disease cohorts, as well as nearly 200,000 individuals from the UKBB, we found its false positive rate to be less than 1 in 200,000. Our results showed that SMA Finder is a robust and accurate tool for detecting the most common molecular causes of SMA using exome, genome, and targeted sequencing samples.

However, the analysis had several limitations related to these cohorts. It was predominantly based on samples of European ancestry. A precise estimate of SMA Finder's accuracy in other populations will require testing on cohorts with larger numbers of non-European samples. Similarly, since most samples used DNA extracted from blood, further testing is needed to measure SMA Finder's accuracy for other sample sources such as saliva or dried blood spots.

The SMA Finder algorithm in its current form also has several important limitations. First, it only determines whether an individual has zero or more than zero functional copies of *SMN1*, and thus does not provide information on carrier status or *SMN2* copy number. In the future, we

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

may extend SMA Finder to estimate the exact copy numbers of *SMN1* and *SMN2*. Second, SMA Finder only tests for variants that disrupt the c.840 position of *SMN1*. Studies have shown that, depending on ancestry, between 6% and 49% of SMA cases[16] may be caused by other variants in regions of the *SMN1* gene that don't overlap the c.840 position. SMA Finder cannot currently detect these variants, and so would yield false negative results in those cases. Furthermore, SMA Finder was unable to make a call in 112 exomes (0.7%) and 6 genomes (0.2%) from the CMG as well as 14 exomes (0.007%) and 85 genomes (0.04%) from UKBB due to insufficient read coverage at the c.840 position. These no-call percentages may be useful for confirming sequencing data quality when applying SMA Finder to new cohorts where a significantly higher proportion of no-calls would indicate problems with target capture or sequencing at the SMN locus. Finally, SMA Finder was designed and tested on samples generated using Illumina sequencers and aligned using BWA.[4] Samples generated using a different sequencing technology, or aligned using a different, non-functionally-equivalent aligner may not work as expected, particularly when aligning to the GRCh38 reference where the ALT contigs include additional copies of *SMN1* and *SMN2* (**Supplementary Methods**).

Rare disease diagnostic pipelines based on short read sequencing data have, until recently, needed to exclude highly repetitive regions such as segmental duplications due to technical limitations of existing variant calling algorithms. However, our analysis adds to the growing body of research[2,8,17] which demonstrates that, with focused tool development, it is possible to detect variants in these regions with sufficient accuracy for rare disease diagnosis. Furthermore, at some loci such as the SMN c.840 position, our ability to accurately detect the causal variant from exome or targeted sequencing data can be comparable to the accuracy of variant calling in non-repetitive regions.

A striking observation from our cohorts was that 11 out of the 13 new SMA diagnoses were made in individuals carrying a clinical diagnosis of muscular dystrophy, congenital myasthenia, or congenital myopathy, and 12 of the 16 known SMA cases carried an initial clinical misdiagnosis of muscular dystrophy, reinforcing similar observations by other groups.[18] Although it is known that the differential diagnosis of an infant or adolescent with progressive weakness includes SMA in addition to muscular dystrophies, congenital myopathies, and congenital myasthenic syndromes[19], clinical gene panels for muscular dystrophy and congenital myopathy may not include SMA testing. Moreover, clinicians and researchers may be misled by the presence of variants of unknown significance (VUS) in other neuromuscular disease genes or by complementary diagnostic tools such as electromyography (EMG) and muscle MRI. Prior to the inclusion of SMA testing in newborn screening, there was commonly a diagnostic delay in SMA, particularly for the milder types II and III which can present with clinical features that are similar to later-onset Mendelian myopathies[20]. Now, with newly-available SMA treatments where early detection is imperative for the best therapeutic effect, there is increased urgency to address diagnostic delays. Our analysis demonstrates that the inclusion of an SMA detection tool within existing rare disease analysis pipelines - particularly for cohorts that include neuromuscular phenotypes - can successfully identify missed cases.

In 2021, the American College of Medical Genetics (ACMG) published an updated set of recommendations for reporting of secondary findings (SF) in clinical exome and genome sequencing.[21] In the section titled "Poor candidates for secondary findings, due to concerns about analytical validity", the authors listed reasons why genes with homologous sequences should be excluded from the SF list. These included concerns about the accuracy of variant

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

detection in these regions as well as challenges in orthogonal validation. Given the clinically overlooked diagnoses presented in this paper, the demonstrated PPV of SMA Finder, as well as the widespread availability of confirmatory molecular testing for SMA, we propose that *SMN1* should now be considered for inclusion in the ACMG SF list.

Declaration of interests

HLR receives research funding from Microsoft and previously received funding from Illumina to support rare disease gene discovery and diagnosis. AODL has consulted for Tome Biosciences and Ono Pharma USA Inc, and is member of the scientific advisory board for Congenica Inc and the Simons Foundation SPARK for Autism study. PBK has received research support from ML Bio and Sarepta Therapeutics, and has consulted for Lupin, Neurogene, NS Pharma, and Teneofour.

Acknowledgments

This paper is dedicated to Clara and Dmitry Bogomolny. We thank the many families who participate in these rare disease studies and the Broad CMG collaborators for sharing the samples and medical data. The Broad CMG sequencing and analysis was funded by the National Human Genome Research Institute (NHGRI), the National Eye Institute, the National Heart, Lung and Blood Institute grant UM1HG008900 and NHGRI grants U01HG011755 and R01HG009141. GH is supported by the GREGoR Consortium, and research in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U24HG011746. This publication has also been made possible in part by CZI grants 2019-19927, 2020-224274, and 2022-309464 https://doi.org/10.37921/236582yuakxy, from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (funder DOI 10.13039/100014989). The Tartu University Hospital cohort analysis was funded by the Estonian Research Council grants PSG774 and PRG471. VSG was supported by NIH NHGRI grant T32HG010464. Work in CGB's laboratory is supported by intramural funds of NINDS/NIH. This research has been conducted using the UK Biobank Resource under application number 48511. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the other organizations that have funded the work.

Data and code availability

SMA Finder is publicly available as a stand-alone tool at https://github.com/broadinstitute/sma_finder under the open-source MIT license.

Also, it is available as a WDL workflow on Terra at https://portal.firecloud.org/?return=terra#methods/translational-genomics-group/sma-finder/4

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

References

1. Sarv, S. *et al.* The Birth Prevalence of Spinal Muscular Atrophy: A Population Specific Approach in Estonia. *Front. Genet.* **12**, 796862 (2021).
2. Chen, X. *et al.* Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet. Med.* **22**, 945–953 (2020).
3. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80**, 155–165 (1995).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
5. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178 (2018) doi:10.1101/201178.
6. Chen, X. *et al.* Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am. J. Hum. Genet.* **110**, 240–250 (2023).
7. Töpf, A. *et al.* Sequential targeted exome sequencing of 1001 patients affected by unexplained limb-girdle weakness. *Genet. Med.* **22**, 1478–1488 (2020).
8. Steyaert, W. *et al.* Systematic analysis of paralogous regions in 41,755 exomes uncovers clinically relevant variation. *Nat. Commun.* **14**, 6845 (2023).
9. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
10. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
11. Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610–615 (2008).
12. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

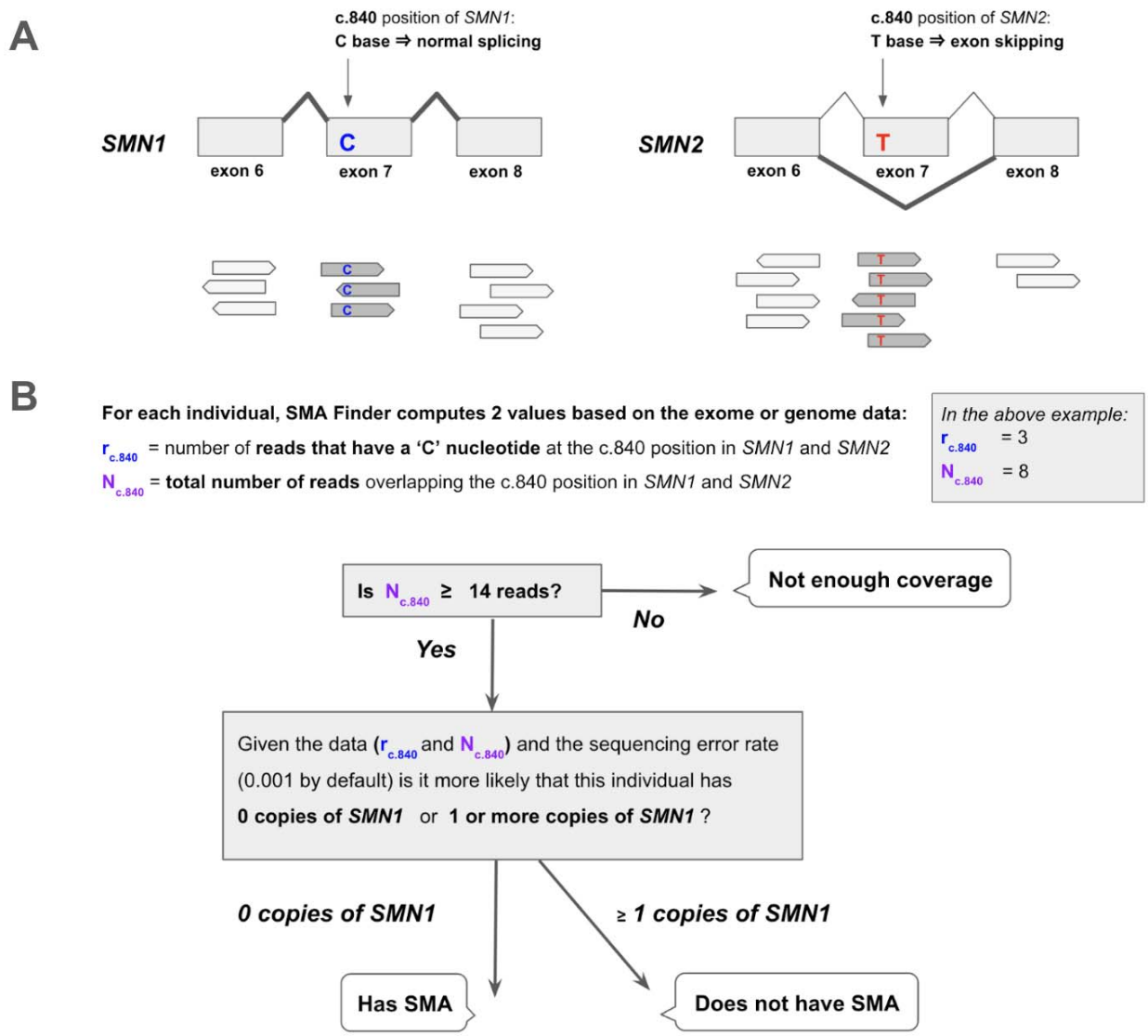
medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

13. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
14. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
15. Baxter, S. M. *et al.* Centers for Mendelian Genomics: A decade of facilitating gene discovery. *Genet. Med.* **24**, 784–797 (2022).
16. Vorster, E., Essop, F. B., Rodda, J. L. & Krause, A. Spinal Muscular Atrophy in the Black South African Population: A Matter of Rearrangement? *Front. Genet.* **11**, 54 (2020).
17. Peng, X. *et al.* Overcoming the Pitfalls of Next-Generation Sequencing-Based Molecular Diagnosis of Shwachman-Diamond Syndrome. *J. Mol. Diagn.* **24**, 1240–1253 (2022).
18. Krenn, M., Jengojan, S. & Grisold, W. Spinal muscular atrophy presenting with mild limb-girdle weakness in adulthood: Diagnostic pitfalls in the era of disease-modifying therapies. *J. Neurol. Sci.* **440**, 120347 (2022).
19. D'Amico, A., Mercuri, E., Tiziano, F. D. & Bertini, E. Spinal muscular atrophy. *Orphanet J. Rare Dis.* **6**, 71 (2011).
20. Lin, C.-W., Kalb, S. J. & Yeh, W.-S. Delay in Diagnosis of Spinal Muscular Atrophy: A Systematic Literature Review. *Pediatr. Neurol.* **53**, 293–300 (2015).
21. Miller, D. T. *et al.* Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1391–1398 (2021).

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Figures

Figure 1. Detecting SMA using reads aligned to the *SMN1* and *SMN2* paralogs

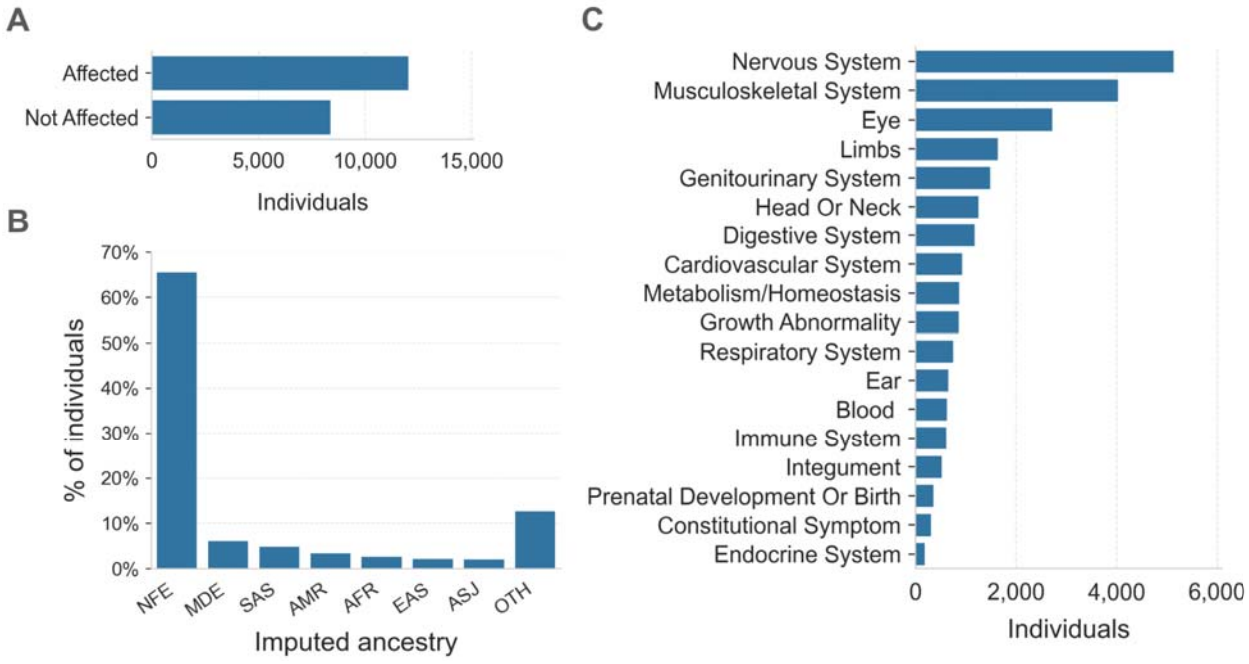


A. The *SMN1* and *SMN2* paralogs are 99.9% identical. One of the few differences between them occurs at their c.840 position. The 'C' at this position in *SMN1* leads to proper splicing, while the 'T' in *SMN2* leads to skipping of exon 7 in most *SMN2* transcripts. Individuals that have zero functional copies of *SMN1* develop spinal muscular atrophy (SMA), and the severity of their disease is inversely proportional to the number of copies of *SMN2* in their genome since each copy of *SMN2* can produce a small amount of SMN protein.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

B. SMA Finder works by counting all aligned reads that overlap the c.840 position in both *SMN1* and *SMN2* and then computing the fraction of reads that have a 'C' at that position. This fraction is interpreted as the fraction of intact *SMN1* copies in the individual's genome. When it is near zero, it implies the absence of any functional copies of *SMN1*, and therefore suggests that the sample is positive and the individual has a diagnosis of SMA.

Figure 2. Overview of the CMG rare disease cohort



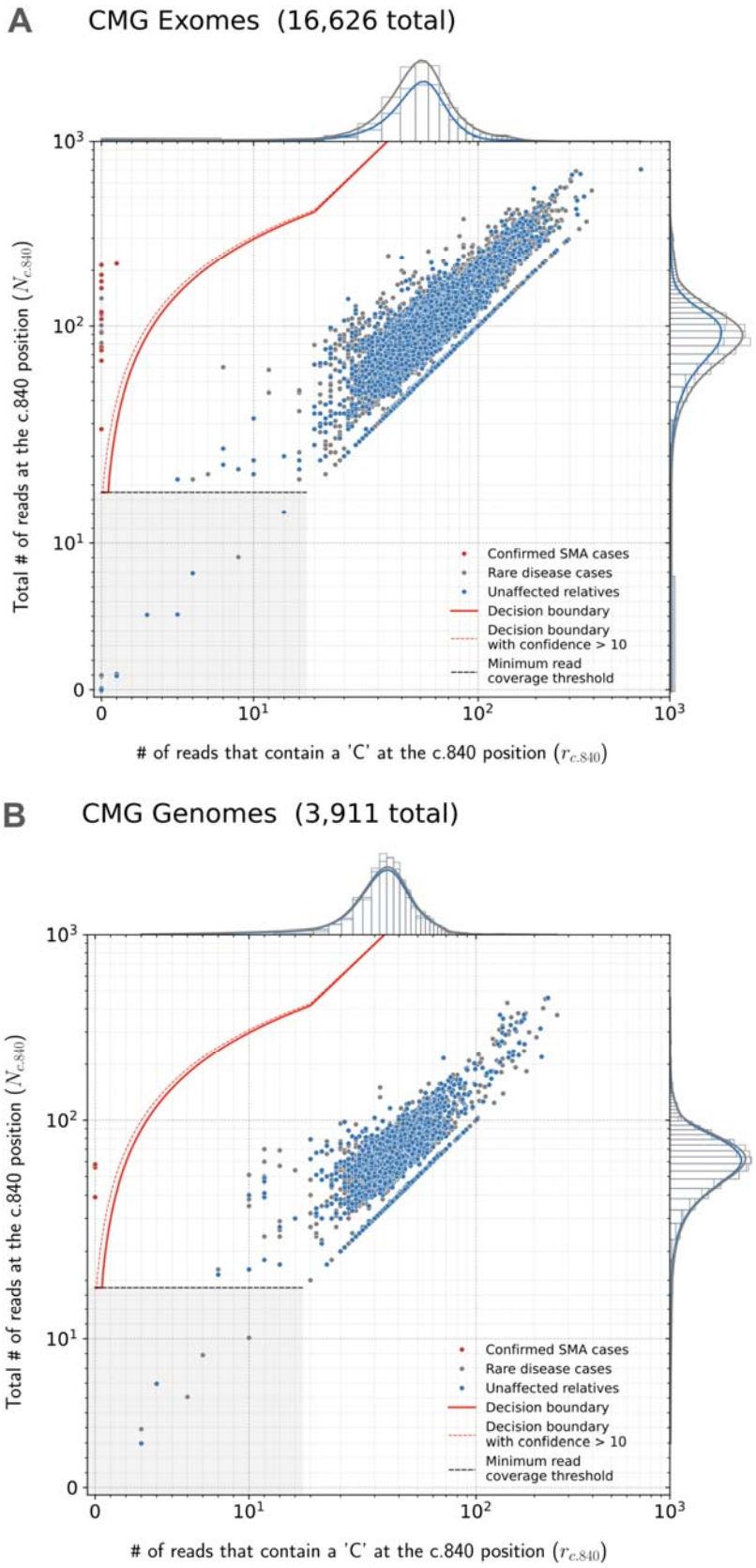
A. The affected status of individuals in the CMG cohort is shown on the y-axis. 12,045 individuals are in the Affected category, 8,401 are Not Affected, and 91 individuals have unknown affected status. Here "Affected" means that the individual was enrolled in a rare disease cohort due to having a disease considered to be rare and most likely genetic in origin.

B. Inferred ancestry of individuals within the CMG cohort is shown on the x-axis: NFE (Non-Finnish Europeans), MDE (Middle Eastern), SAS (South Asian), AMR (Admixed American), AFR (African/African American), EAS (East Asian), and ASJ (Ashkenazi Jewish), OTH (other ancestries).

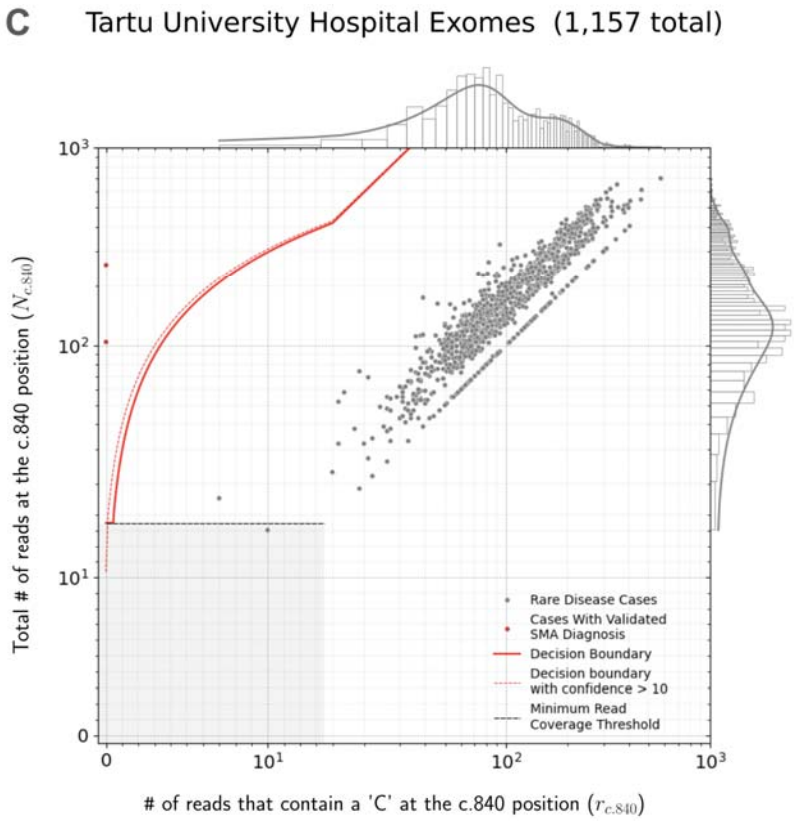
C. The top-level categories from the Human Phenotype Ontology (HPO) are shown on the y-axis. Any individual with multiple HPO terms was counted only once in each category but may be counted more than once across categories.

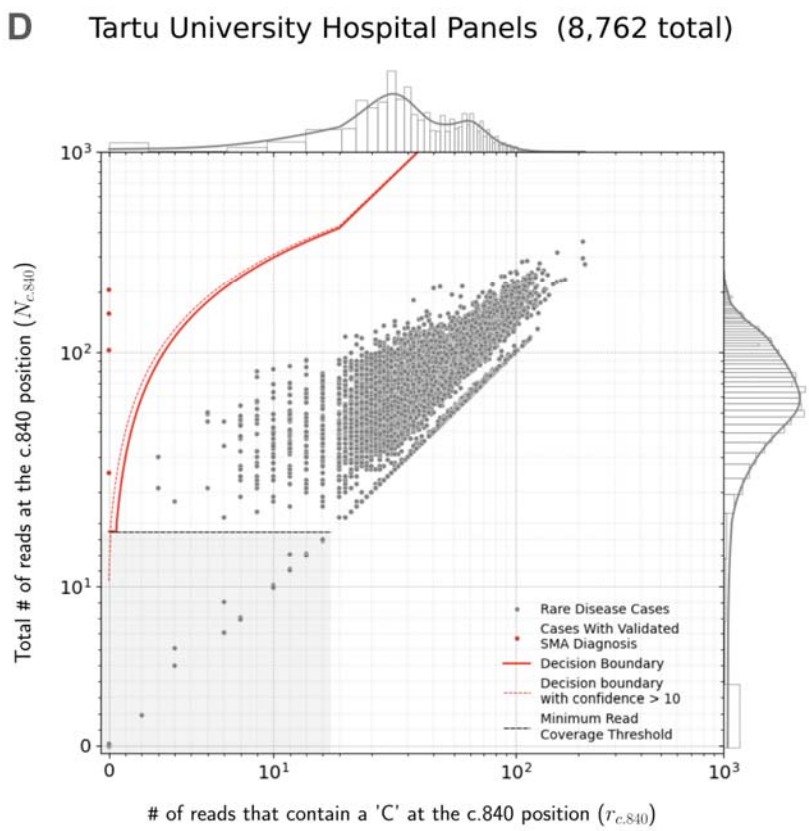
medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Fig 3. SMA Finder results

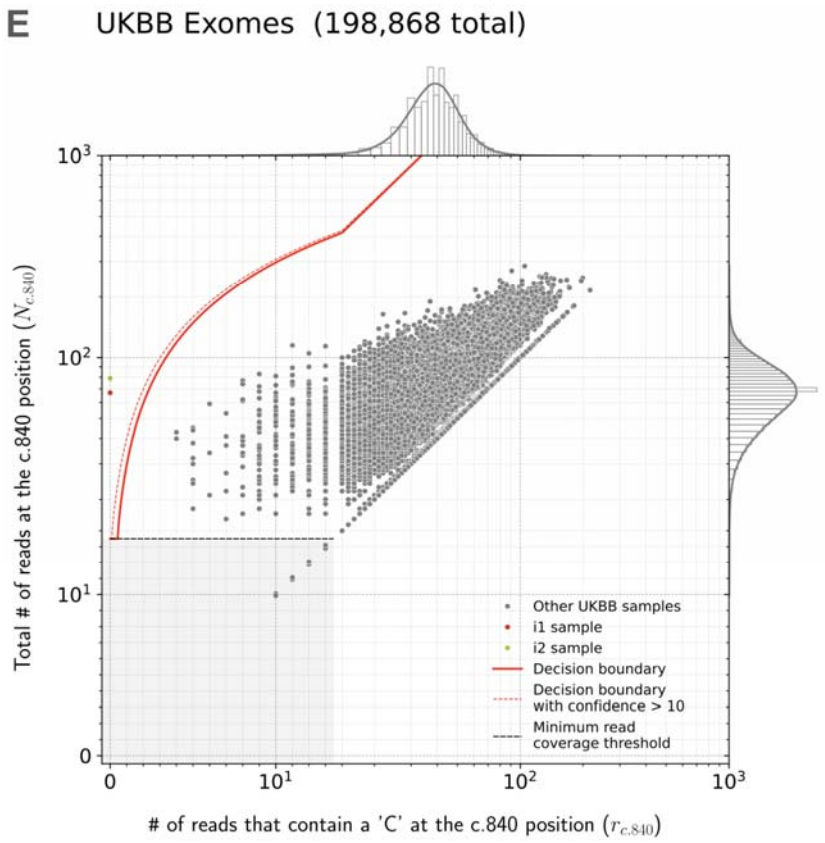


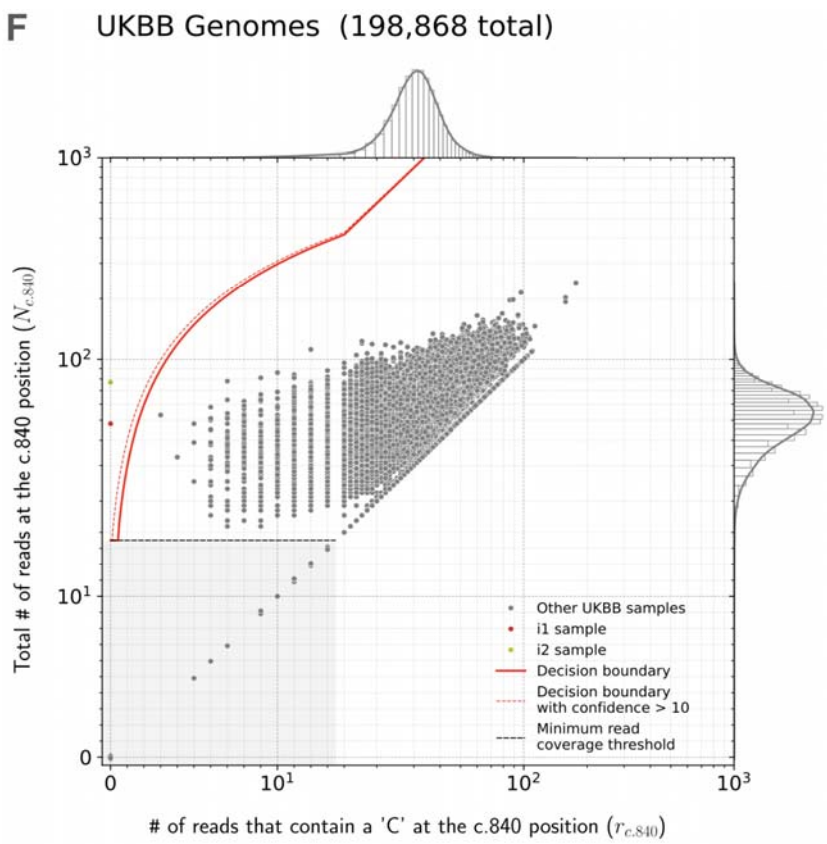
medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .





medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .





medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Read counts measured by SMA Finder in exome (**A**) and genome (**B**) samples from CMG cohorts, as well exomes (**C**) and panel sequencing samples (**D**) from Tartu University Hospital. Each dot represents a sample. The red line represents the decision boundary used by SMA Finder which reports samples to the left of the boundary as SMA-positive. Samples in the gray box where $y \leq 14$ are reported as having insufficient read coverage to make a call. The red dots represent previously known SMA diagnoses, the gray dots are rare disease cases (including the new SMA diagnoses), and the blue dots are unaffected relatives. To clearly show points across a large range of read count values, the x and y axes use a symmetrical log scale that is linear in the range $0 \leq x \leq 14$ and $0 \leq y \leq 14$ before switching to a logarithmic scale for x or $y > 14$. This choice of scale causes part of the decision boundary to appear curved even though it is linear in standard Cartesian coordinates. **E** and **F** show SMA Finder read counts for 198,868 UKBB exomes and genomes respectively. The red dot represents UKBB sample i1 which had phenotype records consistent with an SMA diagnosis and was called positive by both SMA Finder and SMNCopyNumberCaller. The yellow dot represents i2 which was only called positive by SMA Finder and was a no-call from SMNCopyNumberCaller. Marginal histograms show the density of scatter plot points along each axis, with the histogram along the vertical axis showing a distribution of read counts overlapping the c.840 position in *SMN1* + *SMN2*, while the histogram along the horizontal axis shows the number of reads with a 'C' at the c.840 position. **NOTE:** The exome, genome, and targeted sequencing panel samples in **A** and **B**, as well as in **C** and **D** are largely from non-overlapping sets of individuals, while the exomes and genomes in **E** and **F** are alternative samples from the same set of 198,868 individuals in UKBB.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table 1: Positive cases identified by SMA Finder in the rare disease cohorts

	Sample Id	Cohort	Type	Original Clinical Diagnosis	SMA Confirmation Status	Confidence Score	Reads With 'C' at c.840	Total Reads at c.840	Imputed Ancestry
G1	RGP_1153_3	CMG	genome	limb-girdle muscular dystrophy	Positive control	56	0	58	NFE
G2	RGP_823_3	CMG	genome	muscular dystrophy	Positive control	37	0	39	NFE
G3	RGP_916_3	CMG	genome	spinal muscular atrophy	Positive control	54	0	56	NFE
E1	MAAC057	CMG	exome	limb-girdle muscular dystrophy	Positive control	74	0	77	NFE
E2	MANT033	CMG	exome	limb-girdle muscular dystrophy	Positive control	207	0	214	NFE
E3	MCAI001	CMG	exome	limb-girdle muscular dystrophy	Positive control	168	0	174	MDE
E4	MNOT001	CMG	exome	limb-girdle muscular dystrophy	Positive control	182	0	189	NFE
E5	MPAU019	CMG	exome	limb-girdle muscular dystrophy	Positive control	90	0	94	NFE
E6	MLEU047	CMG	exome	limb-girdle muscular dystrophy	Positive control	154	0	160	OTH
E7	SCO_PED004D_GA0014_D1	CMG	exome	neonatal death	Positive control	27	0	28	OTH
E8	MLEU044	CMG	exome	limb-girdle muscular dystrophy	Positive control	105	0	109	NFE
E9	MPAU042	CMG	exome	limb-girdle muscle weakness	Positive control	71	0	74	OTH
E10	MGLA019	CMG	exome	limb-girdle muscular dystrophy	Positive control	97	0	101	NFE
E11	RD6237	Tartu	exome	spinal muscular atrophy I, epilepsy	Positive control	247	0	256	NFE
T1	IT334	Tartu	targeted	spinal muscular atrophy I	Positive control	24	0	25	NFE
T2	NV919	Tartu	targeted	limb-girdle muscular dystrophy	Positive control	197	0	204	NFE
E12	MBEL164	CMG	exome	limb-girdle muscular dystrophy	New diagnosis: confirmed	113	0	117	NFE
E13	MTAM002	CMG	exome	limb-girdle muscle weakness	New diagnosis: confirmed	112	0	116	MDE
E14	MBRN009	CMG	exome	limb-girdle muscular dystrophy	New diagnosis: confirmed	89	0	93	NFE
E15	MWAR098	CMG	exome	limb-girdle muscular dystrophy	New diagnosis: confirmed	182	1	218	NFE
E16	BON_B22_41_1_D1	CMG	exome	muscular dystrophy	New diagnosis: confirmed	115	0	119	NFE
E17	UWA_LAI8985_LAID13-1363_1	CMG	exome	congenital myopathy	New diagnosis: confirmed	62	0	65	OTH
E18	YE8324	Tartu	exome	mild proximal muscle weakness	New diagnosis: confirmed	100	0	104	NFE
T3	KT6879	Tartu	targeted	moderate intellectual disability	New diagnosis: confirmed	149	0	155	NFE
T4	VH4169	Tartu	targeted	myopathy since 3yo	New diagnosis: confirmed	98	0	102	NFE
E19	MTAM032	CMG	exome	limb-girdle muscular dystrophy	Validation in progress	78	0	81	OTH
E20	K0004_01_1	CMG	exome	limb-girdle muscular dystrophy	Validation in progress	136	0	141	AFR
E21	TOP_N799366_1_1	CMG	exome	congenital myasthenic syndrome	Validation in progress	97	0	101	SAS
E22	D304-13N-TB220379	CMG	exome	congenital myasthenic syndrome	Validation in progress	217	0	225	NFE
						24	**0**	**25**	**= Min**
						112		**118**	**= Average**
						247	**1**	**256**	**= Max**

Table 1 lists all samples called SMA-positive by SMA Finder within the Broad Institute Center for Mendelian Genomics (CMG) and Tartu University Hospital cohorts. The Original Clinical Diagnosis column lists the clinical diagnosis prior to SMA testing. The SMA Confirmation Status column shows “Positive control” if the SMA diagnosis was already known at the time that the SMA Finder tool was applied to these cohorts, “New diagnosis: confirmed” if this was a novel diagnosis made via SMA Finder and subsequently confirmed by clinical laboratory testing, or “Validation in progress” if the SMA Finder diagnosis has not yet completed orthogonal confirmation via clinical laboratory testing. The next three columns display the confidence score, **r**, and **N** counts computed by SMA Finder (see Figure1, Methods), while the last column shows the imputed ancestry as NFE (Non-Finnish European), MDE (Middle Eastern), AFR (African/African American), SAS (South Asian), or OTH (other ancestries).

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Supplementary Materials

Installing and Running SMA Finder

The SMA Finder tool is implemented in python v3.7+ and can be installed by running

```
python3 -m pip install sma_finder
```

The --help option can be used to print out SMA Finder's command line arguments:

```
python3 -m sma_finder --help

positional arguments:
  cram_or_bam_path      One or more CRAM or BAM file paths

optional arguments:
  -h, --help            show this help message and exit
  --hg37-reference-fasta HG37_REFERENCE_FASTA
                        HG37 reference genome FASTA path. This should
be specified if the input bam or cram is aligned to HG37.
  --hg38-reference-fasta HG38_REFERENCE_FASTA
                        HG38 reference genome FASTA path. This should
be specified if the input bam or cram is aligned to HG38.
  --t2t-reference-fasta T2T_REFERENCE_FASTA
                        T2T reference genome FASTA path. This should
be specified if the input bam or cram is aligned to the CHM13
telomere-to-telomere benchmark.
  -o OUTPUT_TSV, --output-tsv OUTPUT_TSV
                        Optional output tsv file path
  -v, --verbose         Whether to print extra details during the run
```

SMA Finder has two required inputs:

1) a reference genome file provided via the appropriate `--*-reference-fasta` option
2) the path(s) of one or more aligned read data files in bam or cram format

It outputs a tab-separated table with one row per input file with the following columns:

filename_prefix = input read data filename without the ".bam" or ".cram" suffix
file_type = "bam" or "cram"
genome_version = "hg37", "hg38", or "t2t"
sample_id = sample id parsed from the read data file header
sma_status = "does not have SMA", "has SMA", or "not enough coverage at SMN c.840 position"

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

confidence_score = estimates the degree of confidence in the "sma_status" value

c840_reads_with_smn1_base_C = number of reads with a "C" nucleotide at the c.840 position in *SMN1* + *SMN2*

c840_total_reads = total number of reads that overlap the c.840 position in *SMN1* + *SMN2*

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Table S1. Example SMA Finder output table

filename_prefix	file_type	genome_version	sample_id	sma_status	confidence_score	c840_reads_with_smn1_base_C	c840_total_reads
sample1	bam	hg38	s5	does not have SMA	2524	85	146
sample2	cram	hg38	E18	has SMA	182	1	218
sample3	cram	hg37	s12421	not enough coverage at SMN c.840 position	0	2	8

Table S2. Concordance between SMA Finder exome calls vs SMNCopyNumberCaller genome calls in 198,868 UKBB participants

```
198,818 individuals      negative call by both SMA Finder & SMNCopyNumberCaller
      1 individual (i1)  positive call by both SMA Finder & SMNCopyNumberCaller
      1 individual (i2)  positive call by SMA Finder, no-call by SMNCopyNumberCaller
     34 individuals      negative call by SMA Finder, no-call by SMNCopyNumberCaller
     14 individuals            no-call by SMA Finder, negative call by SMNCopyNumberCaller
```

Table S3. Concordance between SMA Finder calls in UKBB exomes vs genomes

```
198,772 individuals           negative call for both their exome and genome
      2 individuals (i1, i2)  positive call for both their exome and genome
      5 individuals           no-call for both their exome and genome
     80 individuals           negative call for their exome, no-call for their genome
      9 individuals           no-call for their exome, negative call for their genome
```

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

SMN sequences within GRCh38 ALT contigs

In addition to the standard *SMN1* and *SMN2* gene annotations on chr5, the GRCh38 reference contains three more copies of SMN within its ALT contigs. Specifically, a copy of *SMN1* can be found at chr5_KI270897v1_alt:473491-501447, a copy of *SMN2* occurs at chr5_KI270897v1_alt:274951-303848, and a reverse complement sequence of *SMN1* occurs at chr5_GL339449v2_alt:456849-485731. Hypothetically, during read alignment, these extra copies could absorb informative reads and distort the chr5 read counts that SMA Finder relies on to distinguish positive from negative samples. In practice, the BWA MEM aligner performs ALT-aware alignment so that, if a read has comparable alignments to both chromosomal and ALT contig sequence(s), BWA will report the chromosome alignment as the primary one, while any ALT contig alignments will be marked as secondary [†]. Since SMA Finder does not count secondary alignments, we assumed that it would be safe to ignore these extra SMN sequences on ALT contigs. To test the validity of this assumption, we reanalyzed all exome and genome samples within the CMG cohort to confirm that there were no primary alignments overlapping the c.840 positions of these three extra SMN copies on ALT contigs. Specifically, we used a modified version SMA Finder to count reads at the following genomic positions which correspond to c.840 of the three alternative SMN sequences: chr5_KI270897v1_alt:500378 (*SMN1*), chr5_KI270897v1_alt:301867 (*SMN2*), and chr5_GL339449v2_alt:458845 (reverse complement *SMN1*). As expected, in all CMG samples there were zero primary alignments overlapping these three positions, confirming that SMA Finder could ignore the GRCh38 ALT contigs as long as its input read data was aligned using BWA MEM or a functionally-equivalent aligner.

† See https://github.com/lh3/bwa/blob/master/README-alt.md#step-1-bwa-mem-mapping for a detailed description of BWA's approach to ALT contigs.

Detailed description of the SMA Finder algorithm

SMA Finder starts by retrieving aligned reads that overlap the c.840 position in *SMN1* and *SMN2*. Then, it computes two read counts, **r** and **N**, which are defined as follows:

N = total number of reads that overlap the c.840 position in *SMN1* + *SMN2*
r = the number of reads that have a 'C' base at the c.840 position of *SMN1* or *SMN2* and therefore support the presence of at least one intact copy of *SMN1*

These counts include reads whose mapping quality equals zero, but exclude reads whose base quality score at the c.840 position is < 13.

When **N** ≥ 14, SMA Finder uses maximum likelihood estimation to determine whether, given these counts, it is more likely that the sample has zero functional copies of *SMN1*, or that it has more than zero copies. Otherwise, if **N** < 14, SMA Finder reports that the sample has insufficient read coverage to make a call.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.11.24302646; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

When computing likelihoods, SMA Finder makes the following assumptions:

Assumption 1: The probability that a sequencing error caused a particular base within a given read to be incorrectly called a 'T' when the true base was a 'C' is = 0.001 / 3. This is similar to calculations used within GATK HaplotypeCaller's reference confidence model.[5] We do not rely on base quality scores for this parameter since their accuracy can depend on the sequencing technology used and whether algorithms like GATK Base Quality Score Recalibration were applied prior to running SMA Finder. Instead, we uniformly set **p_error = 3.3e-4**.

Assumption 2: Let's say **SMN1_cn** and **SMN2_cn** represent the copy numbers of *SMN1* and *SMN2* in a given sample. If an individual has at least one functional copy of *SMN1*, then the probability of the observed read counts **r** and **N** can be modeled by the binomial distribution B(**r, N, p)** where **p** = **SMN1_cn** / (**SMN1_cn** + **SMN2_cn**). On the other hand, if an individual has zero intact *SMN1* copies due to a mutation disrupting the *SMN1* c.840 position, any observed reads with a C at that position are assumed to be sequencing errors, and so the read counts are instead modeled by B(**r, N, p_error**)**.**

Assumption 3: We assume that, in any given sample, the true value of **SMN1_cn** + **SMN2_cn** **≤ 5** based on the data in Chen et al.[2] which found that genomes rarely contain more than 5 total copies of these paralogs.

Given these assumptions, the algorithm computes the likelihood of observing counts **r** and **N** in each of six scenarios: **SMN1_cn = 0, 1, 2, 3, 4,** or **5**, while keeping **SMN1_cn + SMN2_cn = 5**. Then, it concludes that a sample has zero functional copies of *SMN1* if the likelihood for **SMN1_cn** = 0 is greater than the likelihood for any other value of **SMN1_cn** between 1 and 5.

Since we don't know the true value of **SMN1_cn + CMN2_cn** in a given sample, we always set **SMN1_cn + SMN2_cn** to 5 when computing likelihoods because the resulting model is the least likely to produce a false positive result compared to if we had chosen a smaller value for **SMN1_cn + SMN2_cn**.

SMA Finder also computes a Phred-scaled confidence score by subtracting the log likelihood for **SMN1_cn** = 0 from the maximum log likelihood for **SMN1_cn** = i when i is between 1 and 5:

confidence score = 10 * abs(log10(B(r, N | **SMN1_cn = 0**)) - log10(max({ B(r, N | **SMN1_cn = i**) : i = 1, 2, 3, 4, 5 }))).