2-2-2023

Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing.

Xiao Chen

John Harting

Emily G. Farrow
Children's Mercy Hospital

Isabelle Thiffault
Children's Mercy Hospital

Dalia Kasperaviciute

See next page for additional authors

Recommended Citation

Chen X, Harting J, Farrow E, et al. Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. Am J Hum Genet. 2023;110(2):240-250. doi:10.1016/j.ajhg.2023.01.001

Comprehensive *SMN1* and *SMN2* profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing

Authors

Xiao Chen, John Harting, Emily Farrow, ..., Christian Gilissen, Tomi Pastinen, Michael A. Eberle

Correspondence

meberle@pacificbiosciences.com

We developed Paraphase, an informatics method that, combined with highly accurate long reads, can resolve the highly homologous *SMN1/SMN2* genes involved in spinal muscular atrophy. We characterized *SMN1/SMN2* haplotypes across populations and identified new genetic markers for silent carriers (2+0) with both copies of *SMN1* on the same chromosome.

Chen et al., 2023, The American Journal of Human Genetics *110*, 240–250
February 2, 2023

https://doi.org/10.1016/j.ajhg.2023.01.001

ARTICLE

Comprehensive *SMN1* and *SMN2* profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing

Xiao Chen,[1] John Harting,[1] Emily Farrow,[2,3,4] Isabelle Thiffault,[2,3,5] Dalia Kasperaviciute,[6] Genomics England Research Consortium, Alexander Hoischen,[7,8,9,10] Christian Gilissen,[7,8] Tomi Pastinen,[2,3] and Michael A. Eberle[1,*]

Summary

Spinal muscular atrophy, a leading cause of early infant death, is caused by bi-allelic mutations of *SMN1*. Sequence analysis of *SMN1* is challenging due to high sequence similarity with its paralog *SMN2*. Both genes have variable copy numbers across populations. Furthermore, without pedigree information, it is currently not possible to identify silent carriers (2+0) with two copies of *SMN1* on one chromosome and zero copies on the other. We developed Paraphase, an informatics method that identifies full-length *SMN1* and *SMN2* haplotypes, determines the gene copy numbers, and calls phased variants using long-read PacBio HiFi data. The *SMN1* and *SMN2* copy-number calls by Paraphase are highly concordant with orthogonal methods (99.2% for *SMN1* and 100% for *SMN2*). We applied Paraphase to 438 samples across 5 ethnic populations to conduct a population-wide haplotype analysis of these highly homologous genes. We identified major *SMN1* and *SMN2* haplogroups and characterized their co-segregation through pedigree-based analyses. We identified two *SMN1* haplotypes that form a common two-copy *SMN1* allele in African populations. Testing positive for these two haplotypes in an individual with two copies of *SMN1* gives a silent carrier risk of 88.5%, which is significantly higher than the currently used marker (1.7%–3.0%). Extending beyond simple copy-number testing, Paraphase can detect pathogenic variants and enable potential haplotype-based screening of silent carriers through statistical phasing of haplotypes into alleles. Future analysis of larger population data will allow identification of more diverse haplotypes and genetic markers for silent carriers.

Introduction

Spinal muscular atrophy (SMA) is a neuromuscular disease caused in most cases by bi-allelic mutations of *SMN1* (MIM: 600354).[1–3] SMA is a leading cause of early infant death with an incidence of 1 in 6,000–10,000 live births and a carrier frequency of 1 in 40–80 across ethnic groups.[4–8] SMA can be classified into four clinical types (types I–IV [MIM: 253300, 253550, 253400, 271150]) that differ in age of onset and disease severity.[1]

SMN1 and its paralog *SMN2* (MIM: 601627) reside in a highly complex genomic region on chromosomal band 5q13 that is frequently subject to unequal crossing over and gene conversion, resulting in variable copy numbers (CNs) of *SMN1* and *SMN2*.[7,9] *SMN1* and *SMN2* are nearly identical in sequence with just one functionally different base (GenBank: NM_000344.3; c.840C>T). In *SMN2*, c.840T disrupts a splicing enhancer leading to skipping of exon 7[10] and, as a result, most *SMN2* transcripts are unstable and almost nonfunctional. Since *SMN2* can produce a small amount of functional protein, the CN of *SMN2* is a modifier of the SMA disease severity.[11] The majority (~96%) of 5q-linked SMA cases are caused by bi-allelic absence of *SMN1* c.840C through either large deletions or gene conversion to c.840T, while a smaller percentage (~4%) are caused by other small pathogenic variants in *SMN1* in *trans* with c.840C loss.[8,12–14]

Because of the high carrier frequency and severity of SMA, the American College of Medical Genetics and Genomics recommends population-wide SMA screening.[15] Conventional SMA screening tests use PCR-based methods, such as multiplex ligation-dependent probe amplification (MLPA)[16,17] and qPCR,[18] to determine the *SMN1* dosage (copy number) in exon 7, mostly targeting c.840C>T. To date, a few next-generation sequencing (NGS)-based *SMN1* callers have been reported.[19–22] These callers rely on short reads to identify copy-number variations and distinguish *SMN1* and *SMN2* based on a limited number of differentiating bases centered around c.840C>T. However, dosage testing fails to identify carriers with pathogenic variants other than c.840C>T, which represent ~1%–2% of all carriers.[5] In addition, detecting *SMN2* variants in individuals with SMA is also important for understanding the disease-modifying effect.[23] Both *SMN1* and

[1]PacBio, Menlo Park, CA, USA; [2]Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO, USA; [3]UMKC School of Medicine, University of Missouri Kansas City, Kansas City, MO, USA; [4]Department of Pediatrics, Children's Mercy Kansas City, Kansas City, MO, USA; [5]Department of Pathology and Laboratory Medicine, Children's Mercy Kansas City, Kansas City, MO, USA; [6]Genomics England Ltd., London, UK; [7]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands; [8]Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, the Netherlands; [9]Radboud Center for Infectious Diseases (RCI), Department of Internal Medicine, Radboud University Medical Center, Nijmegen, the Netherlands; [10]Radboud Expertise Center for Immunodeficiency and Autoinflammation and Radboud Center for Infectious Disease (RCI), Radboud University Medical Center, Nijmegen, the Netherlands
*Correspondence: meberle@pacificbiosciences.com
https://doi.org/10.1016/j.ajhg.2023.01.001.

SMN2 are ~28 kb long, and detailed sequence analysis of the complete genes is labor intensive for traditional Sanger sequencing and impossible for conventional short-read NGS methods due to the high sequence similarity between the two genes.

Furthermore, current tests (i.e., dosage testing) are unable to accurately phase alleles. Phasing is important to distinguish between individuals carrying the normal *SMN1* genes on both alleles (1+1) versus silent carriers (2+0) with two copies of *SMN1* on one chromosome and zero copies on the other. Silent carriers account for approximately 3%–9% of carriers in non-African populations and 27% of carriers in African populations.[5,6,21] Throughout this paper, we use the term "singleton *SMN1* allele" to refer to chromosomes with a single copy of *SMN1*, and "two-copy *SMN1* alleles" to refer to alleles with two copies of *SMN1* occurring on the same chromosome. Previous studies have identified the g.27134T>G SNP (GenBank: NG_008691.1; g.32134T>G; rs143838139; GenBank: NM_000344.3; c.*3+80T>G) as a marker of the two-copy *SMN1* allele[24] and this SNP is now commonly tested to modify the residual carrier risk, i.e., the probability that an individual with two copies of *SMN1* is a carrier. However, this SNP is rare and has low sensitivity in non-African populations. In Africans it is common but it is also present on almost 20% of singleton *SMN1* alleles,[21] so it does not have a high positive predictive value (PPV). When an African individual with two copies of *SMN1* tests positive for g.27134T>G, the residual risk of being a carrier, which is largely the silent carrier risk, is estimated to be just 1.7%–3.0%.[20,21,24] More population studies are needed to identify better markers to detect two-copy *SMN1* alleles, but again, short-read based methods suffer from the difficulty to differentiate *SMN1* from *SMN2* due to the high sequence similarity and thus are not ideal methods for identifying these markers.

To better facilitate SMA screening, there is an urgent need for a method that performs comprehensive full-gene *SMN1* and *SMN2* profiling. This method should ideally be able to (1) identify the CN of intact *SMN1* and *SMN2* based on c.840, (2) identify pathogenic variants in *SMN1* other than loss of c.840C, and (3) identify silent carriers. Accurate long-read sequencing is ideal for resolving regions with high sequence homology and the utility of long-read PacBio HiFi sequencing in *SMN1* was previously demonstrated in an amplicon-based study for a Chinese population,[25] though informatics methods are still lacking for shotgun HiFi sequencing, where high sequence homology results in ambiguous alignments. Here we describe a method, Paraphase, that accurately detects the CN, as well as variants throughout *SMN1* and *SMN2* using PacBio HiFi sequencing. We applied Paraphase to population samples from five ethnicities and performed a population-wide haplotype analysis of these genes. We identified major haplogroups for *SMN1* and *SMN2* and quantified their co-segregation patterns. Furthermore, we identified specific haplotypes forming two-copy *SMN1* alleles which could greatly improve the accuracy of silent carrier detection.

Material and methods

Paraphase: HiFi-based *SMN1* and *SMN2* caller

Paraphase extracts HiFi reads aligned to either *SMN1* or *SMN2* and realigns them to the *SMN1* region. It then identifies variant positions throughout the 44 kb long region of interest (chr5: 70,917,100–70,961,220, GRCh38), which includes the *SMN1* gene body plus upstream/downstream regions. Paraphase then assembles haplotypes by linking the phases of each variant site (Figure 1). Haplotypes are assigned to *SMN1* or *SMN2* based on the sequence at the c.840 site, i.e., C is *SMN1* and T is *SMN2*. In addition, Paraphase identifies the common truncated form, *SMNΔ7–8*, that has a 6.3 kb deletion of exons 7–8. Generally, the number of unique *SMN1* and *SMN2* haplotypes reflects *SMN1* and *SMN2* CNs. For samples with only one *SMN1* or *SMN2* haplotype identified, to rule out possible rare cases where two identical haplotypes exist, we calculate whether the depth at the c.840C (T) site is consistent with one or two copies of *SMN1* (*SMN2*). A no-call is reported when the read depth could not reliably distinguish CN1 vs. CN2. CN calls are also adjusted when the number of supporting reads of one haplotype suggests twice the CN of the other haplotypes. With the complete haplotypes resolved, Paraphase makes phased variant calls throughout the genes by calling differences from the reference. Paraphase also assigns haplotypes to haplogroups (see "assigning haplotypes to haplogroups" section below) to enable further haplotype-based analysis for identifying genetic markers. Paraphase works on both whole-genome sequencing (WGS) and hybrid capture-based enrichment data.

Validation of CN calls

To verify the accuracy of our CN calls, we included 107 Coriell samples, 7 from Genome in a Bottle (GIAB),[26] and 100 from the Human Pangenome Reference Center (HPRC).[27] For these samples, *SMN1* and *SMN2* CNs were previously called by a short-read WGS-based method[21] which has been shown to have 99.7% concordance against MLPA and digital PCR. Three of the 107 samples had MLPA calls that agree with short-read based calls.[28] We also included 9 carrier (1+0) samples from Genomic Answers for Kids (GA4K) at Children's Mercy Kansas City with MLPA results (SALSA MLPA P060 SMA Carrier probemix, MRC-Holland). Finally, we included an SMA trio from the 100,000 Genomes Project, where the *SMN1* CN of both parents is one and the proband has zero copies of *SMN1* (the *SMN2* CNs for these three samples are unknown). In total, we had 119 samples with *SMN1* CN information and 116 samples with *SMN2* CN information. Detailed validation sample information is summarized in Table S1.

Population samples

We included 341 pedigrees (26 duos, 308 trios and 7 quartets) from five ethnic populations to study co-segregation of *SMN1* and *SMN2* alleles (Tables S2 and S3). We collected these data from GIAB,[26] the Chinese Quartet project,[29] HPRC,[27] 1000 Genomes Project,[30] the 100,000 Genomes Project, Radboud University Medical Center, and GA4K. Among these pedigrees, 198 are of European (EUR) origin, 37 African (AFR), 35 admixed American (AMR), 26 South Asian (SAS), and 18 East Asian (EAS); 18 are of mixed ancestry and 9 are of unknown ethnicity. In addition, we included 67 samples without pedigree information from GA4K for other frequency calculations (Table S3).

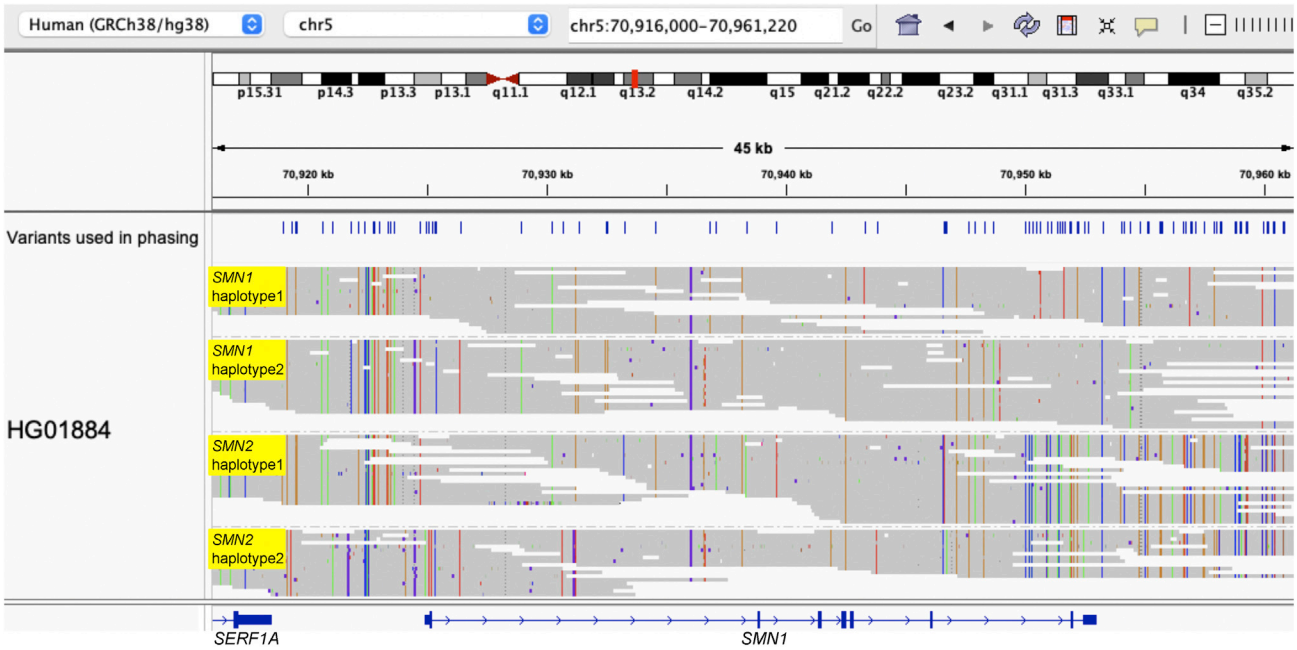


Figure 1. Visualization of phased *SMN1* and *SMN2* haplotypes, using HG01884 as an example
Paraphase produces haplotagged bamlets to facilitate examination of haplotypes with all relevant reads realigned to *SMN1*. Variant positions used in phasing are shown in the top panel and reads are grouped by their assigned haplotypes (IGV option: group by HP tag).

Assigning haplotypes to haplogroups

Multiple sequence alignment and a neighbor-joining tree for the *SMN1* and *SMN2* haplotypes identified across populations were produced by Mafft server[31] (v.7) with default parameters (https://mafft.cbrc.jp/alignment/server/). Haplogroups were identified by manually examining the tree for monophyletic groups. In Paraphase, a new haplotype is assigned a haplogroup by comparing the sequence similarity with representative sequences from each haplogroup and selecting the most similar haplogroup. A small number of haplotypes from each haplogroup were used to produce trees in Figure 2 and Figure S1, visualized with FigTree v.1.4.4 (https://github.com/rambaut/figtree/). Sequences of the same set of haplotypes were visualized in IGV in Figure 3.

Pedigree-based phasing of haplotypes into alleles

For this study, we use the term "haplotype" to refer to a set of phased variants (SNPs or indels) in one copy of a gene (*SMN1* or *SMN2*). Conversely, we use the term "allele" to refer to one or several haplotypes that are inherited on the same chromosome, e.g., co-segregation of two *SMN1* haplotypes or one *SMN1* and one *SMN2* haplotype. Phasing of haplotypes into alleles was done by comparing the haplotypes/haplogroups in parents and probands. Haplotypes were directly assigned haplogroups by Paraphase in samples with >20X HiFi WGS coverage. For parents with either Illumina short read data or low coverage HiFi data (Table S2), i.e., where phasing is not possible or accurate, representative variants for each haplogroup were queried in the parent data to identify the haplogroups in the parent. Haplogroups carried by the parents and the proband were compared to identify which haplotype(s) is inherited on each allele. In ambiguous cases, i.e., both parents have haplotypes of the same haplogroup, manual examination of data in IGV was conducted to find unique SNPs that distinguish these haplotypes and phase them into alleles.

Results

Validation of Paraphase copy-number calling

The *SMN1* and *SMN2* CN calls made by Paraphase were compared against orthogonal methods including short-read WGS-based CN calls, MLPA calls, and SMA trio-based inference (see material and methods). The CN call concordance is 99.2% for *SMN1* and 100% for *SMN2* (Table 1). We correctly called all SMA-affected individuals and carriers and did not make any false positive case or carrier calls. The *SMNΔ7–8* calls are also concordant with orthogonal methods.

We next applied Paraphase to our collection of population samples (see material and methods). While the sample sizes for non-European populations are small, among 259 unrelated European individuals, there are 6 (2.32%, all validated with MLPA) with one copy of *SMN1* (SMA 1+0 carriers), and 61 (23.6%) samples have *SMNΔ7–8*, agreeing with previous studies.[5,6,21]

SMN1 and *SMN2* haplotypes across populations

We performed a population-wide haplotype analysis of 925 *SMN1* haplotypes and 645 *SMN2* haplotypes (excluding *SMNΔ7–8*) and identified ten and nine major *SMN1* and *SMN2* haplogroups, respectively (Figure 2). Representative haplotype sequences from each haplogroup are shown in Figure 3, together with *SMNΔ7–8* sequences. Through pedigree-based analysis (see material and methods), we phased *SMN1* haplotypes into alleles and summarized their population frequencies (Table 2, *SMN2* allele frequencies are

Table 1. Validation against samples with known *SMN1*/*SMN2* copy numbers (CNs)

CN by orthogonal methods	Total	Concordant	Discordant	No-call	Agreement (excluding no-calls)
SMN1					
0	1	1	0	0	100%
1	12	12	0	0	100%
2	79	79	0	0	100%
>2	27	26	1[a]	0	96.3%
Total	119	118	1	0	99.2%
SMN2					
0	8	8	0	0	100%
1	43	42	0	1[b]	100%
2	63	63	0	0	100%
>2	2	2	0	0	100%
Total	116	115	0	1	100%
SMNΔ7–8					
0	104	104	0	0	100%
1	3	3	0	0	100%
Total	107	107	0	0	100%

[a]The discordant call was a CN3 miscalled as CN2, due to two of the three haplotypes being identical in sequence.
[b]The no-call was due to an ambiguous read depth that could not reliably distinguish CN1 vs. CN2 when only one haplotype was found.

listed in Table S4). A few *SMN1* haplotypes are labeled with suffix "c" to indicate that the downstream region of *SMN1* is similar to that of *SMN2* (Figure 3). For example, S1-1c is similar to its corresponding haplotype without the suffix, S1-1, in the gene body and is similar to *SMN2* downstream of the gene. These haplotypes form separate clades and group with *SMN2* haplotypes when sequences of the upstream and downstream regions are included in the phylogenetic analysis (Figure S1A). These haplotypes could have arisen through gene conversion.[32,33]

For single-copy *SMN1* alleles, S1-1 is the most common haplotype across all ethnicities, with a frequency ranging from 29.9% in Africans to 83.3% in East Asians. S1-2 and S1-3 are also common (10%–20%) in Europeans, South Asians, and Admixed Americans, while they are less common (<3%) in Africans and East Asians. Notably, it is not the most common haplotype, S1-1, but S1-2 that is represented by the reference genome (GRCh38). Additionally, we observed several African-specific haplogroups (S1-7, S1-8, S1-9, S1-9d, and S1-10). Out of all *SMN1* haplogroups, S1-10 is closest in sequence to *SMN2* (Figures 2, 3 and S2).

The sequence differences between *SMN1* and *SMN2* are mainly located in exon 7 and exon 8, as well as the downstream region (Figure 3). *SMNΔ7–8* is a truncated form with a 6.3 kb deletion of exons 7–8[21,23] (Figure 3). For all the *SMNΔ7–8* haplotypes found in our data, the downstream region is highly similar to that of *SMN2* (thus labeled as *SMN2Δ7–8*), confirming previous findings that this common truncated form likely derives from *SMN2*.[21] Note that as the sequence flanking the deletion breakpoint is identical between *SMN1* and *SMN2*, this deletion can possibly occur in *SMN1*[23] (a rare haplotype that we have not seen in our data), and the nonfunctional status would be the same as when it occurs in *SMN2*. Conversely, the upstream region and exons 1–6 are highly similar between *SMN1* and *SMN2* and there is not a single SNP that could distinguish *SMN1* from *SMN2* reliably in this region, i.e., there is not any SNP that is present in <10% of *SMN1* haplotypes and >90% of *SMN2* haplotypes, or vice versa. *SMN1* and *SMN2* haplotypes do not separate when only exons 1–6 sequences are included in the phylogenetic analysis (Figure S1B). As a result of the high similarity, read alignments are often ambiguous in this region, even for long reads.

In addition to small variants and the 6.3 kb known deletion in *SMN2*, we also found a previously unknown common structural variant in this region. A 3.6 kb (chr5: 70,917,700–70,921,260, GRCh38) deletion occurs upstream of *SMN1* in S1-9d, which is otherwise similar to S1-9.

Two-copy *SMN1* alleles

African individuals have more copies of *SMN1* than other populations, with about 45%–50% of the population carrying >2 copies of *SMN1* indicating the presence of two-copy *SMN1* alleles.[5,6,21] The higher frequency of two-copy *SMN1* alleles leads to a higher frequency (estimated at ~27% of all carriers) of 2+0 silent carriers where an individual has two copies of *SMN1* but both occur on the same chromosome. Without pedigree information, two-copy *SMN1* alleles are impossible to detect directly with current technologies. Through pedigree-based phasing of

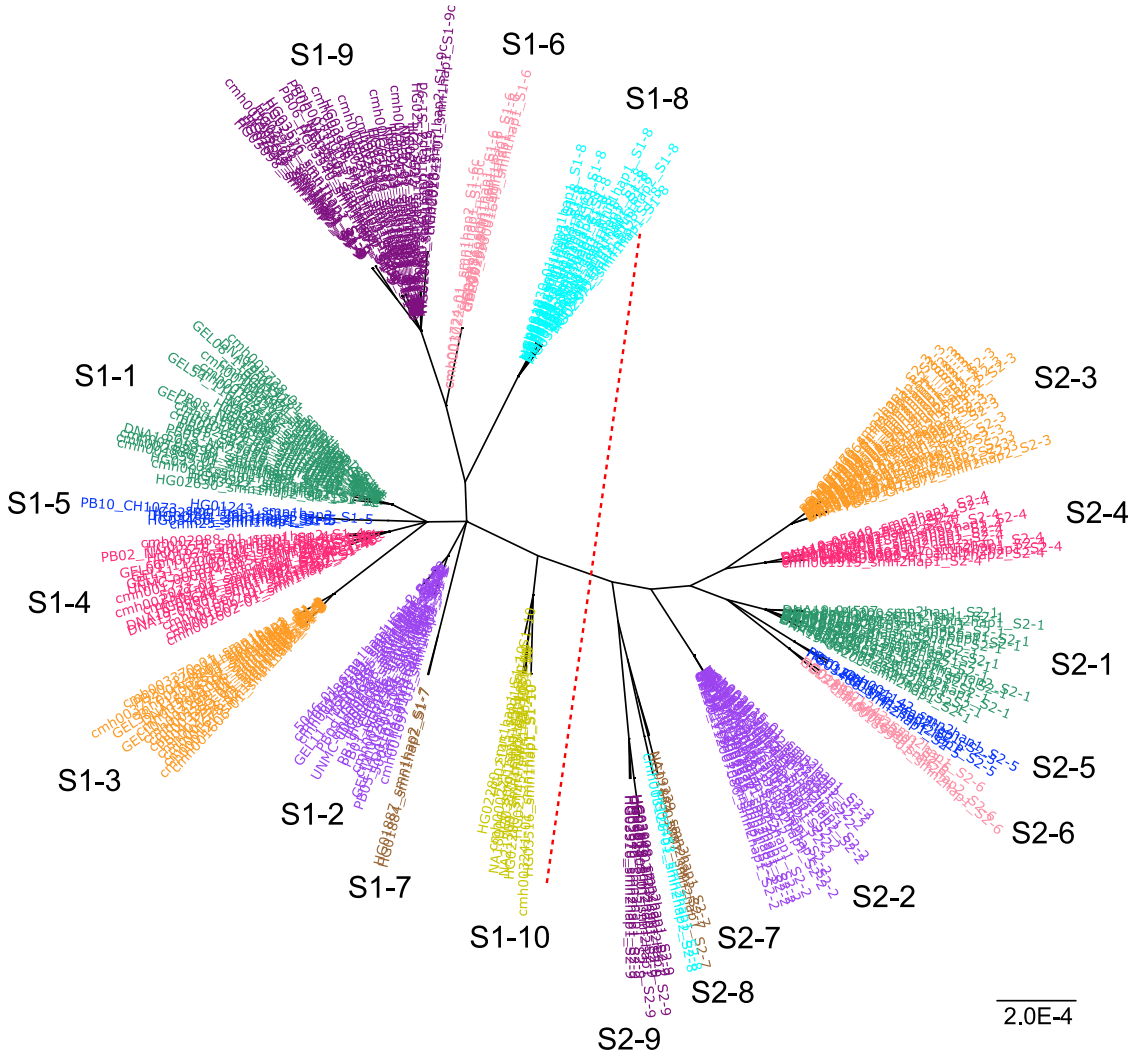


Figure 2. Population-wide haplotype analysis identified major *SMN1* and *SMN2* haplogroups
Representative haplotype sequences of the gene region from each *SMN1* and *SMN2* haplogroup were used to create an unrooted tree. The red dotted line in the middle separates *SMN1* (left) and *SMN2* (right). Figure S1 shows a tree of the same haplotypes created using the gene plus upstream/downstream regions, and a tree of the same haplotypes created using sequences of exons 1–6. The scale bar indicates the number of substitutions per site.

haplotypes into alleles, we studied two-copy *SMN1* alleles and their frequencies. In African individuals, there exist a few haplotypes (S1-8, S1-9c, and S1-9d) that are commonly found in two-copy *SMN1* alleles but not singleton *SMN1* alleles (Table 2) and these could serve as potential markers for two-copy *SMN1* alleles. In particular, we identified a common two-copy *SMN1* allele, S1-8+S1-9d, that comprises two-thirds (21 out of 31) of African two-copy *SMN1* alleles and 24.1% of total African alleles. These two *SMN1* haplotypes, S1-8 and S1-9d, are rarely present as singletons (both at 1.1%, Table 2). Taking the previous estimate of zero-copy *SMN1* allele frequency in Africans (0.68%[6]), if an African individual has two copies of *SMN1*, S1-8 and S1-9d, the likelihood of the two haplotypes being on the same chromosome, i.e., a silent carrier (2+0), is 7.7 times higher than the two haplotypes being present on different chromosomes, and thus the probability of being a silent carrier is 88.5%.

The SNP g.27134T>G in intron 7 of *SMN1* is commonly used as a marker of two-copy *SMN1* alleles.[24] In our data, this SNP is only found in haplogroups S1-8 (21.9%), S1-9 (100%), S1-9c (100%), and S1-9d (96.3%). Samples positive for g.27134T>G are mainly those carrying the two-copy alleles S1-8+S1-9d, S1-8+S1-9c, and S1-9 singletons. S1-9 is commonly found as singleton *SMN1* alleles in Africans (10.3% of all African alleles and 16.1% of singleton African alleles) and it differs from S1-9d only by the 3.6 kb deletion upstream of *SMN1* and differs from S1-9c only in the downstream region. Therefore, g.27134T>G is expected to be present on a high percentage of singleton *SMN1* alleles (16.1% in our data), consistent with previous maximum-likelihood estimates (18.4%),[21] and thus not an accurate marker for two-copy *SMN1* alleles. Conversely, using HiFi reads, Paraphase can accurately distinguish S1-9d or S1-9c from S1-9. In addition, being able to identify the other haplotype of the pair, S1-8, further improves

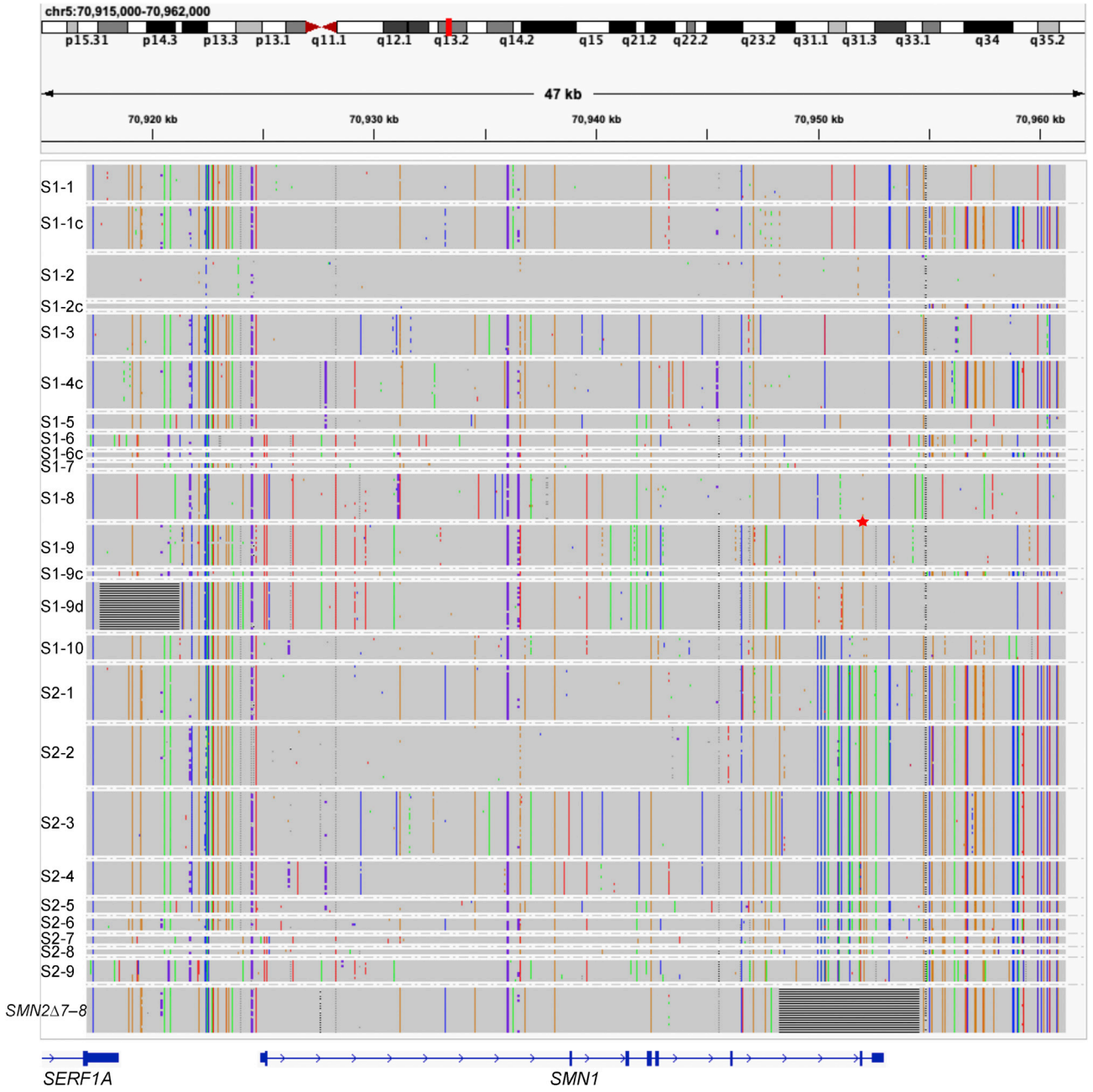


Figure 3. Representative haplotype sequences from each *SMN1* and *SMN2* haplogroup as well as *SMN2Δ7–8*

IGV snapshot showing the haplotype sequences used in the phylogenetic analysis in Figure 2. Sequences of the gene region plus upstream and downstream regions were included. *SMN2Δ7–8* has the 6.3 kb deletion of exons 7–8. S1-9d has a 3.6 kb deletion upstream of *SMN1*. The SNP g.27134T>G, commonly used in silent carrier screening, is marked with a red star symbol between S1-8 and S1-9.

Paraphase's accuracy of detecting the two-copy *SMN1* alleles.

For non-African populations, 57.1% (12 of 21) of two-copy *SMN1* alleles involve combinations of common *SMN1* haplotypes, i.e., S1-1+S1-1, S1-1+S1-2, and S1-1+S1-3 (Table 2). We also observed four two-copy *SMN1* alleles where one of the copies of *SMN1* includes the *SMN2* sequence in the downstream region (flagged with the "c" suffix), i.e., S1-1+S1-1c, S2-2+S2-2c, S1-4c+S1-4c, and S1-6+S1-6c. It is possible that gene conversion from *SMN2* to *SMN1* in exons 7–8 resulted in these two-copy *SMN1* alleles. Taking all non-African samples together, this pattern explains 8 out of 21 (38.1%) two-copy *SMN1* alleles, or 4 out of 8 (50%) distinct two-copy *SMN1* alleles. This is in line with the previous finding that paralog-specific variants (PSVs) between *SMN1* and *SMN2* downstream of the genes are overrepresented in signature variants enriched in two-copy *SMN1* alleles in a Chinese population.[25] However, as these "c" haplotypes are also present as singleton *SMN1* alleles (6.0% of all non-African singleton alleles) and the other

Table 2. *SMN1* allele frequencies across five ethnic populations

***SMN1* Alleles**	**European**		**East Asian**		**South Asian**		**Admix American**		**African**	
Zero-copy (no *SMN1*)	**5**	**1.2%**	**1**	**2.4%**	**1**	**1.9%**	**0**	**0.0%**	**0**	**0.0%**
Singleton *SMN1* alleles										
S1-1	233	55.9%	35	83.3%	27	51.9%	47	67.1%	26	29.9%
S1-1c	16	3.8%	1	2.4%	2	3.8%	2	2.9%	2	2.3%
S1-2	80	19.2%	2	4.8%	7	13.5%	6	8.6%	1	1.1%
S1-2c	0	0.0%	1	2.4%	1	1.9%	0	0.0%	0	0.0%
S1-3	65	15.6%	0	0.0%	7	13.5%	8	11.4%	1	1.1%
S1-4c	7	1.7%	0	0.0%	1	1.9%	1	1.4%	0	0.0%
S1-5	1	0.2%	0	0.0%	0	0.0%	1	1.4%	1	1.1%
S1-6	0	0.0%	0	0.0%	0	0.0%	1	1.4%	3	3.4%
S1-6c	0	0.0%	0	0.0%	0	0.0%	1	1.4%	1	1.1%
S1-7	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	2.3%
S1-8	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	1.1%
S1-9d	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	1.1%
S1-9	0	0.0%	0	0.0%	0	0.0%	0	0.0%	9	10.3%
S1-10	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	9.2%
Singleton total	**402**	**96.4%**	**39**	**92.9%**	**45**	**86.5%**	**67**	**95.7%**	**56**	**64.4%**
Two-copy *SMN1* alleles										
S1-1+S1-1	3	0.7%	0	0.0%	2	3.8%	0	0.0%	0	0.0%
S1-1+S1-2	1	0.2%	1	2.4%	1	1.9%	0	0.0%	0	0.0%
S1-1+S1-3	4	1.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
S1-1+S1-1c	0	0.0%	1	2.4%	3	5.8%	1	1.4%	0	0.0%
S1-2+S1-2c	1	0.2%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
S1-4c+S1-4c	1	0.2%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
S1-5+S1-5	0	0.0%	0	0.0%	0	0.0%	1	1.4%	0	0.0%
S1-6+S1-6c	0	0.0%	0	0.0%	0	0.0%	1	1.4%	1	1.1%
S1-8+S1-9d	0	0.0%	0	0.0%	0	0.0%	0	0.0%	21	24.1%
S1-8+S1-9c	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	2.3%
S1-9+S1-9	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	2.3%
S1-10+S1-10	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	2.3%
S1-1+S1-9d	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	2.3%
S1-1+S1-8	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	1.1%
Two-copy total	**10**	**2.4%**	**2**	**4.8%**	**6**	**11.5%**	**3**	**4.3%**	**31**	**35.6%**
Total alleles	417		42		52		70		87	

haplotype of the pair is often a highly common singleton allele such as S1-1 and S1-2, these haplotypes will frequently occur on two different chromosomes, so this "c" haplotype pattern as a marker does not have a high PPV as was observed for the S1-8+S1-9d allele in Africans.

Co-segregation of *SMN1* and *SMN2* haplotypes

We next investigated the co-segregation of *SMN1* and *SMN2* haplotypes. Our results show that *SMN2* (including *SMN2Δ7–8*) is present on 85.3% of singleton *SMN1* alleles but only 26.9% of two-copy *SMN1* alleles. This indicates that gains of *SMN1* are often accompanied with losses of *SMN2*[21] and it is possible that many two-copy *SMN1* alleles were generated through gene conversion of *SMN2* into *SMN1*.[32]

For standard alleles with one copy of *SMN1* and one copy of full-length *SMN2*, i.e., excluding *SMN2Δ7–8*, we examined the types of *SMN1* and *SMN2* haplotypes on the

Table 3. *SMN1*-*SMN2* haplogroup co-segregation on alleles with one copy of full-length *SMN1* and one copy of full-length *SMN2*

SMN1 haplogroup	*SMN2* haplogroup	# co-segregated alleles	# *SMN1* haplogroups segregated with other *SMN2* haplogroups	# *SMN2* haplogroups segregated with other *SMN1* haplogroups	% co-segregation
S1-1/S1-1c	S2-1	297	8[a]	8[b]	94.9%
S1-2/S1-2c	S2-2	101	0	2	98.1%
S1-3	S2-3	70	5[b]	6[a]	86.4%
S1-4c	S2-4	8	2	0	80.0%
S1-5	S2-5	3	0	0	100.0%
S1-6/S1-6c	S2-6	4	1	0	80.0%
S1-7	S2-7	2	0	0	100.0%
S1-8	S2-8	1	0	0	100.0%
S1-9/S1-9d	S2-9	8	0	0	100.0%

[a]Among these alleles, 6 are S1-1 co-segregated with S2-3.
[b]Among these alleles, 5 are S1-3 co-segregated with S2-1.

same allele. We found that an *SMN1* haplogroup is usually segregated with a specific *SMN2* haplogroup (Table 3). This suggests that it is possible to probabilistically phase *SMN1* and *SMN2* together. For simplicity we named the *SMN2* haplogroups to match the corresponding *SMN1* haplogroups that they usually co-segregate with (e.g., S1-1 and S2-1 usually co-segregate). Interestingly, when we queried the sequence similarity between *SMN1* and *SMN2* haplogroups in exons 1–6 (exons 7–8 are not included as they are differentiated between *SMN1* and *SMN2*), *SMN1* haplogroups usually share the highest similarity with the co-segregating *SMN2* haplogroups (Figure S3A). This is true for the three most common haplogroups (S1-1, S1-2, and S1-3), as well as three out of the six less common haplogroups (S1-4 through S1-9; S1-10 is not included as none of S1-10 haplotypes occurs on the same allele as *SMN2*, see below). As a result, some of the co-segregating *SMN1* and *SMN2* haplogroups group together when exons 1–6 sequences were used to create the phylogeny (Figure S1B). For less common alleles, a larger sample size is needed to further confirm the co-segregation pattern and the sequence similarity, especially for S1-7 (n = 2) and S1-8 (n = 1).

We also examined co-segregation of alleles other than one copy of *SMN1* and one copy of full-length *SMN2*. First, S1-10 alleles always contain zero copy of *SMN2* (8 out of 8 alleles). Since S1-10 is closest in sequence to *SMN2* among all *SMN1* haplogroups (Figure 2) and S1-10 alleles never contain *SMN2*, S1-10 could be a hybrid gene between *SMN1* and *SMN2* created by a fusion deletion. Next, *SMN2Δ7–8* alleles segregate with S1-1 in 98% (51 out of 52) of cases. *SMN2Δ7–8* is most similar in sequence in exons 1–6 to S1-1 and S2-1 (Figure S3B). Both the co-segregation and the sequence similarity suggest that *SMN2Δ7–8* is most likely derived from S2-1. Finally, we summarized the frequency of *SMN1* (*SMN2*) haplotypes on alleles without *SMN2* (*SMN1*) (Table S5). Among our limited sample of four alleles without *SMN1* (zero-copy *SMN1* alleles), four contain more than one copy of *SMN2*. Among these four alleles, two of them carry an *SMN2* haplotype with the downstream region similar to *SMN1* (Figure S4), suggesting possible loss of *SMN1* through gene conversion from *SMN1* to *SMN2*.

Discussion

Here we provide the most comprehensive analysis of variation in one of the most difficult, clinically important regions of the human genome. Extending beyond copy-number testing based primarily on c.840C>T as is often done, Paraphase phases the region to provide a much richer level of information. Using the phasing information, Paraphase can detect other pathogenic variants and enable haplotype-based screening of silent carriers. Since Paraphase works mainly by phasing variant positions from long reads, it works for both WGS and hybrid capture-based enrichment data and can be adapted to work with amplicon sequencing data, when the full *SMN1*/*SMN2* regions are captured or amplified. Compared with short-read based methods, highly accurate HiFi reads can provide long-range haplotype information through entire genes and easily pick up large structural variants such as the 6.3 kb deletion in *SMNΔ7–8* and the 3.6 kb deletion in the *SMN1* haplotype S1-9d.

In this study we conducted a population-wide full-gene haplotype analysis of *SMN1* and *SMN2*. Combining our gene-level phasing with pedigree information, we identified haplotypes that form two-copy *SMN1* alleles. Most importantly, we identified a common two-copy *SMN1* allele that comprises 67.7% of two-copy *SMN1* alleles in Africans. The two individual haplotypes on this allele each occur very rarely as singleton *SMN1* alleles in the population. Based on our limited sample of 87 African alleles, we estimate that testing positive for these two haplotypes in an individual with two copies of *SMN1* gives a silent carrier risk

of 88.5%, which is significantly higher than the previously found marker SNP g.27134T>G (1.7%–3.0%).[20,21,24]

In addition, we found co-segregation patterns between *SMN1* and *SMN2* haplotypes. An *SMN1* haplogroup often co-segregates with the *SMN2* haplogroup that is most similar in sequence, suggesting that intrachromosomal gene conversion between *SMN1* and *SMN2* plays a significant role in the evolution of this region. With larger sample datasets enabling more accurate allele frequency calculations, it should be possible to build a probabilistic model to predict the most likely allele/genotype configurations based on the haplotypes seen in an individual. This would be very helpful for silent carrier detection. For example, an individual with S1-8, S1-9d, and S2-1 is very likely a silent carrier, as S1-8 and S1-9d rarely exist as singleton *SMN1* alleles and S2-1 rarely segregates with S1-8 or S1-9d. For an individual with these haplotypes, the most likely alleles are two copies of *SMN1* (S1-8+S1-9d) with no *SMN2* on one allele and one copy of *SMN2* (S2-1) with no *SMN1* on the other allele.

One limitation in this study is the relatively small number of samples (438) studied. To make more statistically powered findings out of the haplotype analysis, it is desirable to increase the sample size, particularly for non-European populations. Future analysis of large population data with Paraphase, using either HiFi WGS or possibly a hybrid capture based or other targeted long-read approaches, will allow a better characterization of variants in both genes, identification of more diverse haplotypes, analysis of alleles carrying two or even more copies of *SMN1* or *SMN2*, and discovery of more genetic markers for silent carrier detection.

Paraphase is designed to resolve single copies of *SMN1* or *SMN2* from both WGS and targeted sequence data and our study points to the utility of haplotype-based statistical phasing in predicting phasing information between multiple copies of *SMN1*/*SMN2*. While it would be useful to assemble and phase the complete region covering *SMN1* and *SMN2* (a 2–4 Mb highly complex region with many segmental duplications), we are limited by the high sequence homology throughout the region. The *SMN1*/*SMN2* region has long been known to have a high degree of variability in the population.[34] Recently, Vollger et al. investigated this region in CHM13 and a few more near-complete *de novo* assemblies and showed that there is a high amount of variation among different alleles and that this region could not be consistently resolved across samples.[35] Therefore, while we attempted to provide a preliminary analysis of the flanking genes to study the structure of the region (see Supplemental Note, Figures S5 and S6), complete resolution of the entire region will require a future study that utilizes carefully designed *de novo* assembly methods and high-quality pedigree data to QC assemblies.

The method employed in Paraphase can be applied to other segmental duplication regions with extremely high sequence similarity and frequent copy-number variations. We are currently extending this method to solve similar gene paralog problems such as *CYP21A2*, and we will apply this method to more clinically relevant genes in the future. The development of more targeted informatics solutions for difficult regions with HiFi data will bring us one step closer to consolidating the numerous genetic tests that are currently offered into a single test.

Data and code availability

Paraphase can be downloaded from https://github.com/PacificBiosciences/paraphase.

Bamlets for visualizing *SMN1* and *SMN2* haplotypes of GIAB and HPRC samples can be downloaded from https://github.com/xiao-chen-xc/SMN_phased_data.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2023.01.001.

Acknowledgments

We thank the Human Pangenome Reference Center (HPRC) for generating and releasing the HiFi WGS data. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK, and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. The GA4K HiFi sequencing data was made possible by the generous gifts to Children's Mercy Research Institute and the Genomic Answers for Kids program at Children's Mercy Kansas City.

Declaration of interests

X.C., J.H., and M.A.E. are employees of Pacific Biosciences.

Received: October 28, 2022
Accepted: December 20, 2022
Published: January 19, 2023

Web resources

Genbank, https://www.ncbi.nlm.nih.gov/genbank/
OMIM, https://www.omim.org/

References

1. Lunn, M.R., and Wang, C.H. (2008). Spinal muscular atrophy. Lancet *371*, 2120–2133. https://doi.org/10.1016/S0140-6736(08)60921-6.
2. Mercuri, E., Bertini, E., and Iannaccone, S.T. (2012). Childhood spinal muscular atrophy: controversies and challenges. Lancet Neurol. *11*, 443–452. https://doi.org/10.1016/S1474-4422(12)70061-3.

3. Prior, T.W. (2010). Perspectives and diagnostic considerations in spinal muscular atrophy. Genet. Med. *12*, 145–152. https://doi.org/10.1097/GIM.0b013e3181c5e713.
4. Ogino, S., Leonard, D.G.B., Rennert, H., Ewens, W.J., and Wilson, R.B. (2002). Genetic risk assessment in carrier testing for spinal muscular atrophy. Am. J. Med. Genet. *110*, 301–307. https://doi.org/10.1002/ajmg.10425.
5. Hendrickson, B.C., Donohoe, C., Akmaev, V.R., Sugarman, E.A., Labrousse, P., Boguslavskiy, L., Flynn, K., Rohlfs, E.M., Walker, A., Allitto, B., et al. (2009). Differences in *SMN1* allele frequencies among ethnic groups within North America. J. Med. Genet. *46*, 641–644. https://doi.org/10.1136/jmg.2009.066969.
6. Sugarman, E.A., Nagan, N., Zhu, H., Akmaev, V.R., Zhou, Z., Rohlfs, E.M., Flynn, K., Hendrickson, B.C., Scholl, T., Sirko-Osadsa, D.A., and Allitto, B.A. (2012). Pan-ethnic carrier screening and prenatal diagnosis for spinal muscular atrophy: clinical laboratory analysis of >72 400 specimens. Eur. J. Hum. Genet. *20*, 27–32. https://doi.org/10.1038/ejhg.2011.134.
7. MacDonald, W.K., Hamilton, D., and Kuhle, S. (2014). SMA carrier testing: a meta-analysis of differences in test performance by ethnic group. Prenat. Diagn. *34*, 1219–1226. https://doi.org/10.1002/pd.4459.
8. Verhaart, I.E.C., Robertson, A., Wilson, I.J., Aartsma-Rus, A., Cameron, S., Jones, C.C., Cook, S.F., and Lochmüller, H. (2017). Prevalence, incidence and carrier frequency of 5q-linked spinal muscular atrophy - a literature review. Orphanet J. Rare Dis. *12*, 124. https://doi.org/10.1186/s13023-017-0671-8.
9. Stabley, D.L., Harris, A.W., Holbrook, J., Chubbs, N.J., Lozo, K.W., Crawford, T.O., Swoboda, K.J., Funanage, V.L., Wang, W., Mackenzie, W., et al. (2015). *SMN1* and *SMN2* copy numbers in cell lines derived from patients with spinal muscular atrophy as measured by array digital PCR. Mol. Genet. Genomic Med. *3*, 248–257. https://doi.org/10.1002/mgg3.141.
10. Lorson, C.L., Hahnen, E., Androphy, E.J., and Wirth, B. (1999). A single nucleotide in the *SMN* gene regulates splicing and is responsible for spinal muscular atrophy. Proc. Natl. Acad. Sci. USA. *96*, 6307–6311. https://doi.org/10.1073/pnas.96.11.6307.
11. Butchbach, M.E.R. (2016). Copy number variations in the survival motor neuron genes: implications for spinal muscular atrophy and other neurodegenerative diseases. Front. Mol. Biosci. *3*, 7. https://doi.org/10.3389/fmolb.2016.00007.
12. Wirth, B., Herz, M., Wetter, A., Moskau, S., Hahnen, E., Rudnik-Schöneborn, S., Wienker, T., and Zerres, K. (1999). Quantitative analysis of survival motor neuron copies: identification of subtle *SMN1* mutations in patients with spinal muscular atrophy, genotype-phenotype correlation, and implications for genetic counseling. Am. J. Hum. Genet. *64*, 1340–1356. https://doi.org/10.1086/302369.
13. Wirth, B. (2000). An update of the mutation spectrum of the survival motor neuron gene (*SMN1*) in autosomal recessive spinal muscular atrophy (SMA). Hum. Mutat. *15*, 228–237. https://doi.org/10.1002/(SICI)1098-1004(200003)15:3<228::AID-HUMU3>3.0.CO;2-9.
14. Alías, L., Bernal, S., Fuentes-Prior, P., Barceló, M.J., Also, E., Martínez-Hernández, R., Rodríguez-Alvarez, F.J., Martín, Y., Aller, E., Grau, E., et al. (2009). Mutation update of spinal muscular atrophy in Spain: molecular characterization of 745 unrelated patients and identification of four novel mutations in the *SMN1* gene. Hum. Genet. *125*, 29–39. https://doi.org/10.1007/s00439-008-0598-1.
15. Prior, T.W.; and Professional Practice and Guidelines Committee (2008). Carrier screening for spinal muscular atrophy. Genet. Med. *10*, 840–842. https://doi.org/10.1097/GIM.0b013e318188d069.
16. Arkblad, E.L., Darin, N., Berg, K., Kimber, E., Brandberg, G., Lindberg, C., Holmberg, E., Tulinius, M., and Nordling, M. (2006). Multiplex ligation-dependent probe amplification improves diagnostics in spinal muscular atrophy. Neuromuscul. Disord. *16*, 830–838. https://doi.org/10.1016/j.nmd.2006.08.011.
17. Scarciolla, O., Stuppia, L., De Angelis, M.V., Murru, S., Palka, C., Giuliani, R., Pace, M., Di Muzio, A., Torrente, I., Morella, A., et al. (2006). Spinal muscular atrophy genotyping by gene dosage using multiple ligation-dependent probe amplification. Neurogenetics *7*, 269–276. https://doi.org/10.1007/s10048-006-0051-3.
18. Feldkötter, M., Schwarzer, V., Wirth, R., Wienker, T.F., and Wirth, B. (2002). Quantitative analyses of *SMN1* and *SMN2* based on real-time lightCycler PCR: fast and highly reliable carrier testing and prediction of severity of spinal muscular atrophy. Am. J. Hum. Genet. *70*, 358–368. https://doi.org/10.1086/338627.
19. Larson, J.L., Silver, A.J., Chan, D., Borroto, C., Spurrier, B., and Silver, L.M. (2015). Validation of a high resolution NGS method for detecting spinal muscular atrophy carriers among phase 3 participants in the 1000 Genomes Project. BMC Med. Genet. *16*, 100. https://doi.org/10.1186/s12881-015-0246-2.
20. Feng, Y., Ge, X., Meng, L., Scull, J., Li, J., Tian, X., Zhang, T., Jin, W., Cheng, H., Wang, X., et al. (2017). The next generation of population-based spinal muscular atrophy carrier screening: comprehensive pan-ethnic *SMN1* copy-number and sequence variant analysis by massively parallel sequencing. Genet. Med. *19*, 936–944. https://doi.org/10.1038/gim.2016.215.
21. Chen, X., Sanchis-Juan, A., French, C.E., Connell, A.J., Delon, I., Kingsbury, Z., Chawla, A., Halpern, A.L., Taft, R.J., et al.; NIHR BioResource (2020). Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. Genet. Med. *22*, 945–953. https://doi.org/10.1038/s41436-020-0754-0.
22. Lopez-Lopez, D., Loucera, C., Carmona, R., Aquino, V., Salgado, J., Pasalodos, S., Miranda, M., Alonso, Á., and Dopazo, J. (2020). *SMN1* copy-number and sequence variant analysis from next-generation sequencing data. Hum. Mutat. *41*, 2073–2077. https://doi.org/10.1002/humu.24120.
23. Ruhno, C., McGovern, V.L., Avenarius, M.R., Snyder, P.J., Prior, T.W., Nery, F.C., Muhtaseb, A., Roggenbuck, J.S., Kissel, J.T., Sansone, V.A., et al. (2019). Complete sequencing of the *SMN2* gene in SMA patients detects *SMN* gene deletion junctions and variants in *SMN2* that modify the SMA phenotype. Hum. Genet. *138*, 241–256. https://doi.org/10.1007/s00439-019-01983-0.
24. Luo, M., Liu, L., Peter, I., Zhu, J., Scott, S.A., Zhao, G., Eversley, C., Kornreich, R., Desnick, R.J., and Edelmann, L. (2014). An Ashkenazi Jewish *SMN1* haplotype specific to duplication alleles improves pan-ethnic carrier screening for spinal muscular atrophy. Genet. Med. *16*, 149–156. https://doi.org/10.1038/gim.2013.84.
25. Li, S., Han, X., Xu, Y., Chang, C., Gao, L., Li, J., Lu, Y., Mao, A., and Wang, Y. (2022). Comprehensive analysis of spinal

muscular atrophy: *SMN1* copy number, intragenic mutation, and 2 + 0 carrier analysis by third-generation sequencing. J. Mol. Diagn. *24*, 1009–1020. Published online May 31, 2022:S1525-1578. https://doi.org/10.1016/j.jmoldx.2022.05.001.

26. Zook, J.M., McDaniel, J., Olson, N.D., Wagner, J., Parikh, H., Heaton, H., Irvine, S.A., Trigg, L., Truty, R., McLean, C.Y., et al. (2019). An open resource for accurately benchmarking small variant and reference calls. Nat. Biotechnol. *37*, 561–566. https://doi.org/10.1038/s41587-019-0074-6.
27. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446. https://doi.org/10.1038/s41586-022-04601-8.
28. Vijzelaar, R., Snetselaar, R., Clausen, M., Mason, A.G., Rinsma, M., Zegers, M., Molleman, N., Boschloo, R., Yilmaz, R., Kuilboer, R., et al. (2019). The frequency of *SMN* gene variants lacking exon 7 and 8 is highly population dependent. PLoS One *14*, e0220211. https://doi.org/10.1371/journal.pone.0220211.
29. Ren L, Duan X, Dong L, et al. Quartet DNA reference materials and datasets for comprehensively evaluating germline variants calling performance. Published online September 28, 2022:2022.09.28.509844. doi:10.1101/2022.09.28.509844
30. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
31. Katoh, K., Rozewicki, J., and Yamada, K.D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. *20*, 1160–1166. https://doi.org/10.1093/bib/bbx108.
32. Chen, T.H., Tzeng, C.C., Wang, C.C., Wu, S.M., Chang, J.G., Yang, S.N., Hung, C.H., and Jong, Y.J. (2011). Identification of bidirectional gene conversion between *SMN1* and *SMN2* by simultaneous analysis of *SMN* dosage and hybrid genes in a Chinese population. J. Neurol. Sci. *308*, 83–87. https://doi.org/10.1016/j.jns.2011.06.002.
33. Vollger MR, DeWitt WS, Dishuck PC, Harvey, WT, Guitart, X, Goldberg, ME, Rozanski, A, Lucas, J, Asri, M, Munson, KM and Lewis, AP Increased mutation rate and interlocus gene conversion within human segmental duplications.Preprint at bioRxiv, Published online July 7, 2022:2022.07.06.498021. doi:10.1101/2022.07.06.498021
34. Campbell, L., Potter, A., Ignatius, J., Dubowitz, V., and Davies, K. (1997). Genomic variation and gene conversion in spinal muscular atrophy: implications for disease process and clinical phenotype. Am. J. Hum. Genet. *61*, 40–50. https://doi.org/10.1086/513886.
35. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental duplications and their variation in a complete human genome. Science *376*, eabj6965. https://doi.org/10.1126/science.abj6965.

The American Journal of Human Genetics, Volume *110*

Supplemental information

Comprehensive *SMN1* and *SMN2* profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing

Xiao Chen, John Harting, Emily Farrow, Isabelle Thiffault, Dalia Kasperaviciute, Genomics England Research Consortium, Alexander Hoischen, Christian Gilissen, Tomi Pastinen, and Michael A. Eberle

Supplemental Information

Variability of paralog specific variants (PSVs) between *SMN1* and *SMN2*

Previous studies[1,2] using short-read population data analyzed the paralog specific variants (PSVs) between *SMN1* and *SMN2* of the reference genome and found indirectly (i.e. without phasing) that they are much more variable in African populations than non-African populations. This calls for careful selection of PSVs for short read-based *SMN1/SMN2* copy number calculation[1,2]. Here we analyzed the variability of these PSVs on our phased *SMN1* and *SMN2* haplotypes. While PSVs are mostly fixed in non-African populations, African haplotypes show a much higher rate of sharing - i.e. *SMN2* bases in an *SMN1* haplotype, or *SMN1* bases in an *SMN2* haplotype (Figure S2A). Focusing on 15 reference-PSVs flanking c.840 in Intron 6-Exon 8, 33.5% of African haplotypes have at least 2 discrepant sites (13.3% have at least 5 discrepant sites), while most (96%) non-African haplotypes have zero or one (Figure S2A). The biggest contributors to the high PSV discrepancy in Africans are a few African-specific haplogroups (Figure S2B and Figure S2C), such as S1-10 (7 or more discrepant sites), and S2-9 (5 discrepant sites).

Silent carrier risk calculation of S1-8+S1-9d in Africans

We took the frequency of S1-8+S1-9d (21/31) out of two-copy *SMN1* alleles, as well as the frequency of S1-8 (1/56) and S1-9d (1/56) out of singleton *SMN1* alleles from our data. We took the frequency of zero-copy (0.68%), singleton (71.79%) and two-copy (27.51%) *SMN1* alleles from Sugarman et al[3]. The probability of S1-8/S1-9d is 2*(71.79%*1/56)*(71.79%*1/56). The probability of -/S1-8+S1-9d is 2*0.68%*(27.51%*21/31). The silent carrier risk is calculated as the weighted probability of -/S1-8+S1-9d.

SMN1/*SMN2* variant calls

The *SMN1*/*SMN2* gene is 27.9kb long and consists of 8 exons, among which Exon 1 is far away from the rest of the exons (13.7kb away from Exon 2). Due to the distance and the fact that *SMN1* and *SMN2* are highly similar in sequence in Exons 1-6, it could be more challenging to phase haplotypes through Exon 1 than Exons 2-8. Among the haplotypes resolved by Paraphase, 98.5% of them cover Exons 2-8, and

88.4% of them cover Exons 1-8. Note that Exon 1 encodes 27 amino acids and currently there is only one pathogenic/likely pathogenic variant in Exon 1 with more than one star in ClinVar (ClinVar ID:9168) (ClinVar last accessed on Oct 12, 2022).

Small variants were called in each phased haplotype. Among the protein changing variants in *SMN1*, we identified two missense variants and one in-frame insertion. They are:
S4G, 70925113A>G, not in ClinVar
G6S, 70925119G>A, not in ClinVar
G7GSGGGV, 70925123G>GCAGTGGTGGCGGCGT, not in ClinVar
K93T, 70942362A>C, ClinVar ID:638580, uncertain significance

Among the protein changing variants in *SMN2*, we identified three missense variants. They are:
G26D, 70049762G>A, not in ClinVar
G106S, 70066976G>A, not in ClinVar
G287R, 70076545G>C, called in four samples. This variant was previously shown to be a positive modifier of SMA[4].
Interestingly, G106S is reported for *SMN1* in ClinVar (ID:634938, uncertain significance), and G26D has been reported by a previous study[5] where they identified the variant but could not map it to *SMN1* or *SMN2*. It is possible that these variants can occur on either *SMN1* or *SMN2*, or these are *SMN2*-specific variants that were mapped to *SMN1* by mistake in the case of G106S.

Phasing *SMN1*/*SMN2* with nearby genes

SMN1 resides in a segmental duplication (SD) that is present in variable copy numbers (CNs) on each chromosome (most often two copies). This SD contains *SMN1* and two other flanking genes, *SERF1A* and *NAIP*, and the other copy of the SD contains *SMN2*, *SERF1B* and *NAIP* pseudogene. In order to understand the structure of the region and the mechanisms leading to CN changes, we sought to phase a bigger region containing these three gene families (Figure S5, top panel). We were limited by the read length and spacing of variants, so we were not able to get complete haplotypes throughout the ~160kb region in most samples. Instead, we individually phased the *SERF1A*/*SERF1B* region and the *NAIP* region, and these haplotypes can be compared against the *SMN1*/*SMN2* haplotypes where they overlap (Figure S5, bottom three panels). Additional copies of partial *NAIP* (fifth haplotype, Figure S5) were also phased, but they occur elsewhere in the genome and are not directly connected to *SMN1*/*SMN2*.

To understand the structure of the region when there are CN changes, we first compared the total CN of *SMN1*+*SMN2* (including *SMN2*Δ*7–8*) against the total CN of *SERF1A*+*SERF1B*, as well as the total CN of *NAIP* genes+pseudogenes (only considering those copies connected to *SMN1*/*SMN2*). In samples where we could resolve *SERF1A/SERF1B,* 65 samples have *SMN1*+*SMN2* CN loss and 8 samples have *SMN1*+*SMN2* CN gain, and all of them have a total *SERF1A*+*SERF1B* CN equal to the total CN of *SMN1*+*SMN2*. In samples where we could resolve the *NAIP* region*,* 73 samples have *SMN1*+*SMN2* CN loss and 14 samples have *SMN1*+*SMN2* CN gain, and all of them have a total *NAIP* gene+pseudogene CN equal to the total CN of *SMN1*+*SMN2*. This suggests that CN changes involve a bigger region than *SERF1A*/*SERF1B* and *NAIP*.

Next, we looked into the relative position of genes as evidence of gene conversion. In the example HG02723 (Figure S5), the *NAIP* copy downstream of *SMN1* is intact on both alleles, while the *NAIP* copy downstream of *SMN2* is truncated on both alleles, i.e. pseudogenes, one with a deletion of Exons 4-5 and the other missing Exons 1-5. We examined whether the intact/truncated *NAIP* could serve as a proxy for "*SMN1*/*SMN2* location". We examined samples where both alleles each contain one copy of *SMN1* and one copy of *SMN2* and they do not contain the "c" haplotypes. 177 (96.7%) out of 183 *SMN1* copies with successful phasing to *NAIP* are upstream of an intact *NAIP*, while 192 (99.5%) out of 193 *SMN2* copies with successful phasing to *NAIP* are upstream of a truncated *NAIP*. This suggests that in the majority of cases, we could define the "*SMN1*/*SMN2* location" as relative to intact/truncated *NAIP*. Note that we do not have information about the exact physical location of the genes and this relative location does not always hold true as *SMN1* and *SMN2* could possibly swap their downstream *NAIP* copies via processes such as inversion if they are in reverse orientation.

We checked the "gene location" of interesting *SMN1* haplotypes. First, 18 (94.7%) out of 19 *SMN1* "c" haplotypes appear to be in the "*SMN1* location" (next to intact *NAIP*), suggesting that they arose by *SMN1* converted to be similar to *SMN2* in the downstream region (Figure S6, top panel). Next, we examined two-copy *SMN1* alleles. For two-copy *SMN1* alleles that do not have any *SMN2*, one of the two *SMN1* copies appears to be in the "*SMN2* location" (next to truncated *NAIP*) in 28 (96.6%) out of 29 alleles, suggesting conversion of the original *SMN2* into *SMN1* (Figure S6, middle panel). For two-copy *SMN1* alleles that do have *SMN2*, both *SMN1* copies appear to be in the "*SMN1* location" in 7 out of 7 alleles, suggesting that the extra copy of *SMN1* arose from duplication of the SD (Figure S6, bottom panel). This analysis provides evidence for two possible mechanisms of getting two-copy *SMN1* alleles, conversion and duplication. This also indicates that phasing with truncated *NAIP* may serve as an

additional marker for two-copy *SMN1* alleles (those that lack *SMN2*) - an individual with two copies of *SMN1*, one of which is next to a truncated *NAIP*, has an increased risk of being a silent carrier.

While this analysis provides some preliminary insights into the structure of the region, it was conducted in a small number of samples where we were able to phase *SMN1*/*SMN2* and nearby genes. Complete resolution of the SD region is beyond the scope of this study and will require a future study that utilizes carefully designed de novo assembly methods and high quality pedigree data to QC assemblies.

Supplemental figures

Figure S1. Trees of the same set of haplotypes used in Figure 2 created with gene sequences plus upstream/downstream regions (A) and Exons 1-6 only (B).

Haplogroups are colored in the same way as in Figure 2. In Panel B, shaded nodes indicate *SMN2* haplogroups. Some *SMN1* and *SMN2* haplogroups of the same color (co-segregating haplogroups) group together (green, purple, blue, magenta and orange, etc.). The inset shows the same tree reduced to two colors (red: *SMN1*; black: *SMN2*).

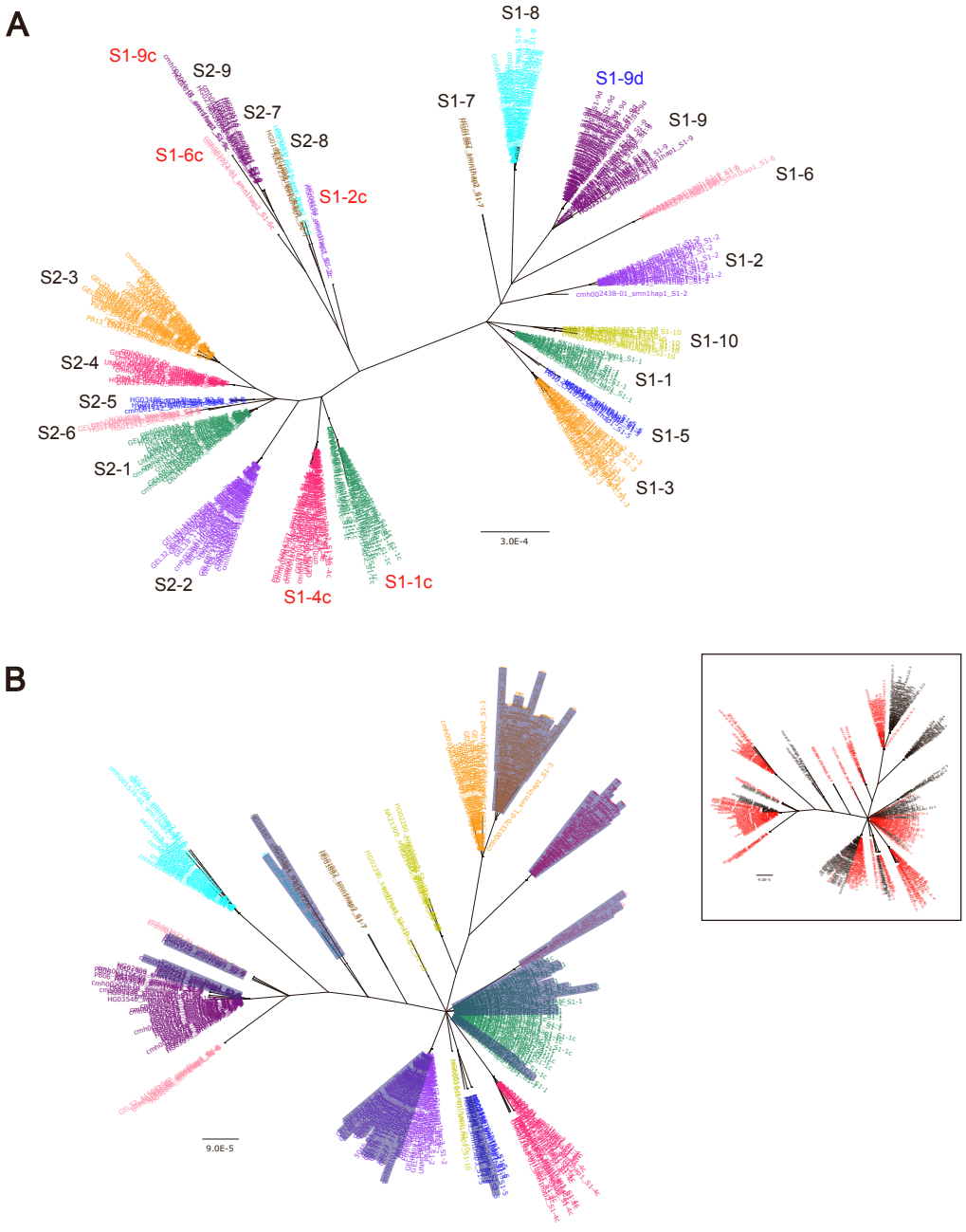


Figure S2. Discrepant PSV sites across populations.

A. Frequency of haplotypes carrying discrepant sites across populations. The x axis shows the number of discrepant PSV sites, i.e. *SMN2* bases on *SMN1* haplotypes or *SMN1* bases on *SMN2* haplotypes, out of 15 reference-PSVs flanking c.840C, taken from Chen et al. 2020[1]. **B.** Frequency of haplotypes carrying discrepant sites across *SMN1* haplotypes. The "c" and "d" haplotypes are identical to their corresponding haplotypes in the gene body, so they are considered as their corresponding haplotypes, e.g. S1-1c considered as S1-1. **C.** Frequency of haplotypes carrying discrepant sites across *SMN2* haplotypes.

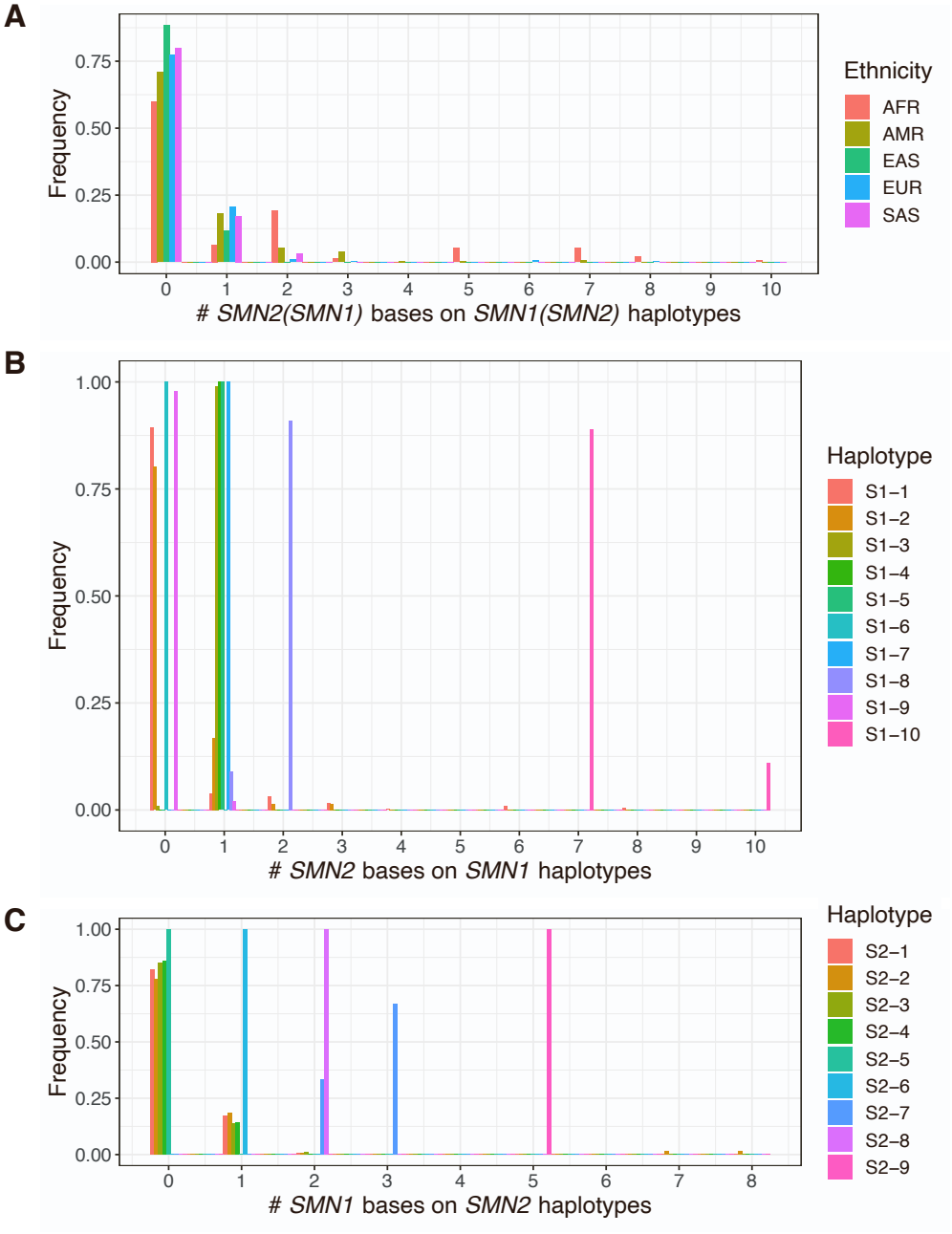


Figure S3. Sequence similarity between *SMN1* and *SMN2* haplogroups.

A. *SMN1* haplotypes are compared against *SMN2* haplotypes and the weighted average similarity between each haplogroup is plotted. For each pairwise comparison, variant concordance is calculated as the fraction of concordant bases out of 444 total sites where variants occur across populations in Exons 1-6. The “c” and “d” haplotypes are identical to their corresponding haplotypes in Exons 1-6, so they are considered as their corresponding haplotypes, e.g. S1-1c considered as S1-1. **B.** *SMN2Δ7–8* haplotypes are compared against *SMN1* and *SMN2* haplotypes among the same set of 444 total variant sites in Exons 1-6. Variant concordance calculation is the same as in A.

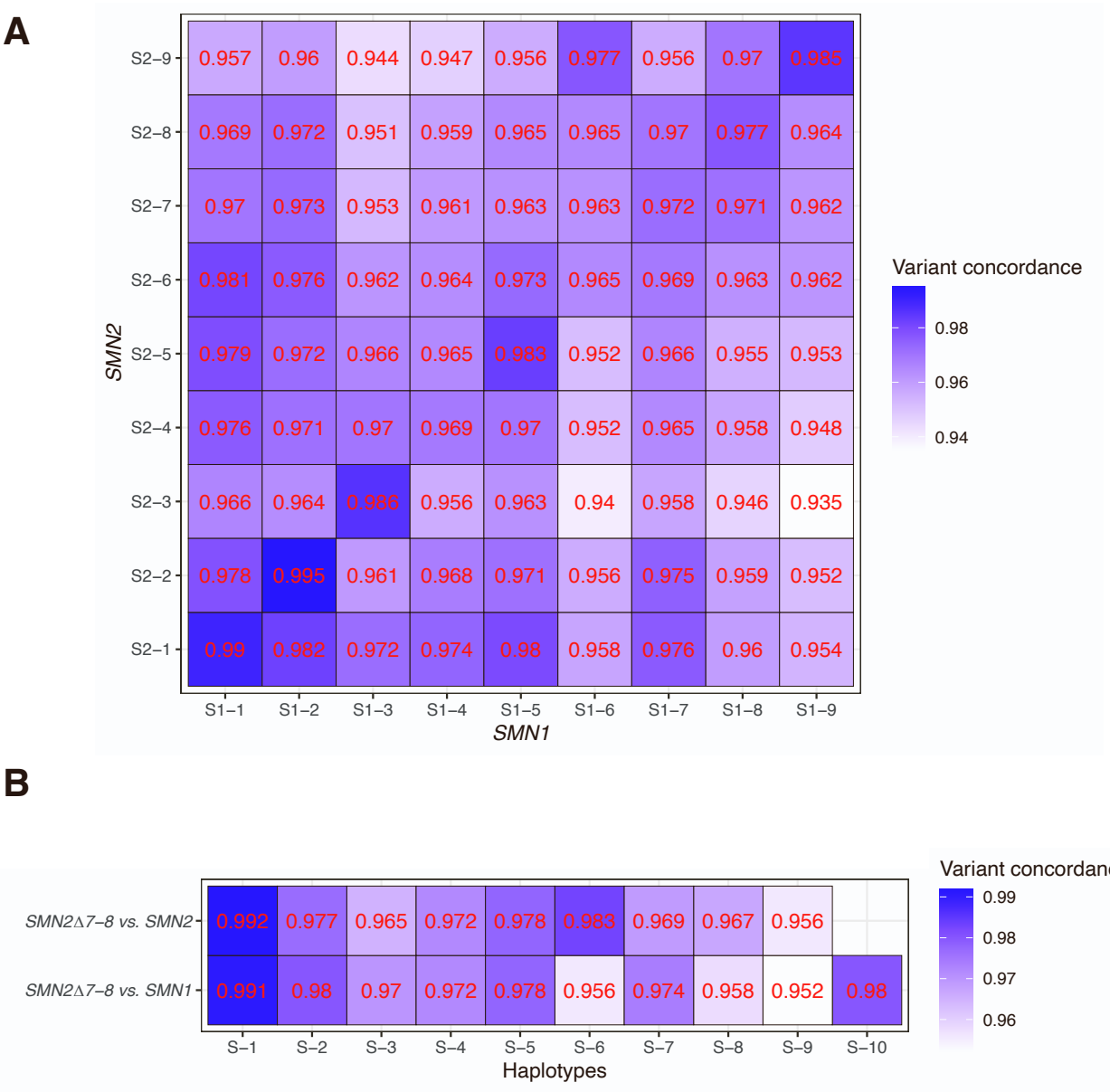


Figure S4. IGV snapshot of *SMN2* haplotypes with the downstream region similar to *SMN1.*

In HG02132, the downstream region of *SMN2* haplotype 3 is similar to *SMN1*. In GEL02, the downstream region of *SMN2* haplotype 4 is similar to *SMN1*. Reads in blue are uniquely assigned to a haplotype, while reads in gray can be assigned to more than one possible haplotype and a random one is selected (this happens when haplotypes are identical over a region).

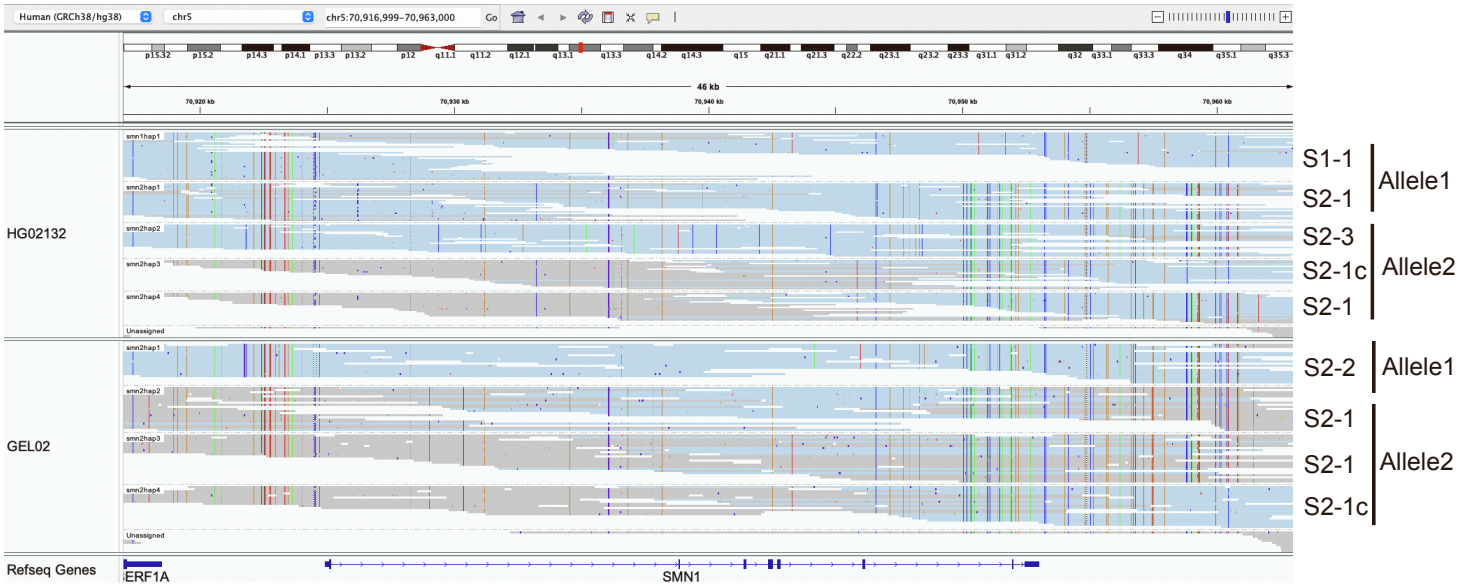


Figure S5. Phasing *SMN1*/*SMN2* with nearby genes: *SERF1A*/*SERF1B* and *NAIP.*

Phasing through a 161kb region containing *SERF1A*/*SERF1B*, *SMN1*/*SMN2* and *NAIP* (or its pseudogene) (top panel) is limited by read length in most samples. In order to study the structure of the bigger region, phasing of individual genes was conducted instead for *SERF1A*/*SERF1B* (second panel), *SMN1*/*SMN2* (third panel) and *NAIP* (last panel) so that these haplotypes could be analyzed and pieced together to understand the bigger region. All copies of the segmental duplications are shown, including those that contain *SERF1A/SMN1/NAIP* and those that contain *SERF1B/SMN2/NAIP* pseudogene. Reads clipped at the same position (clipped sequences are hidden) indicate structural variants, e.g. deletions or translocations.

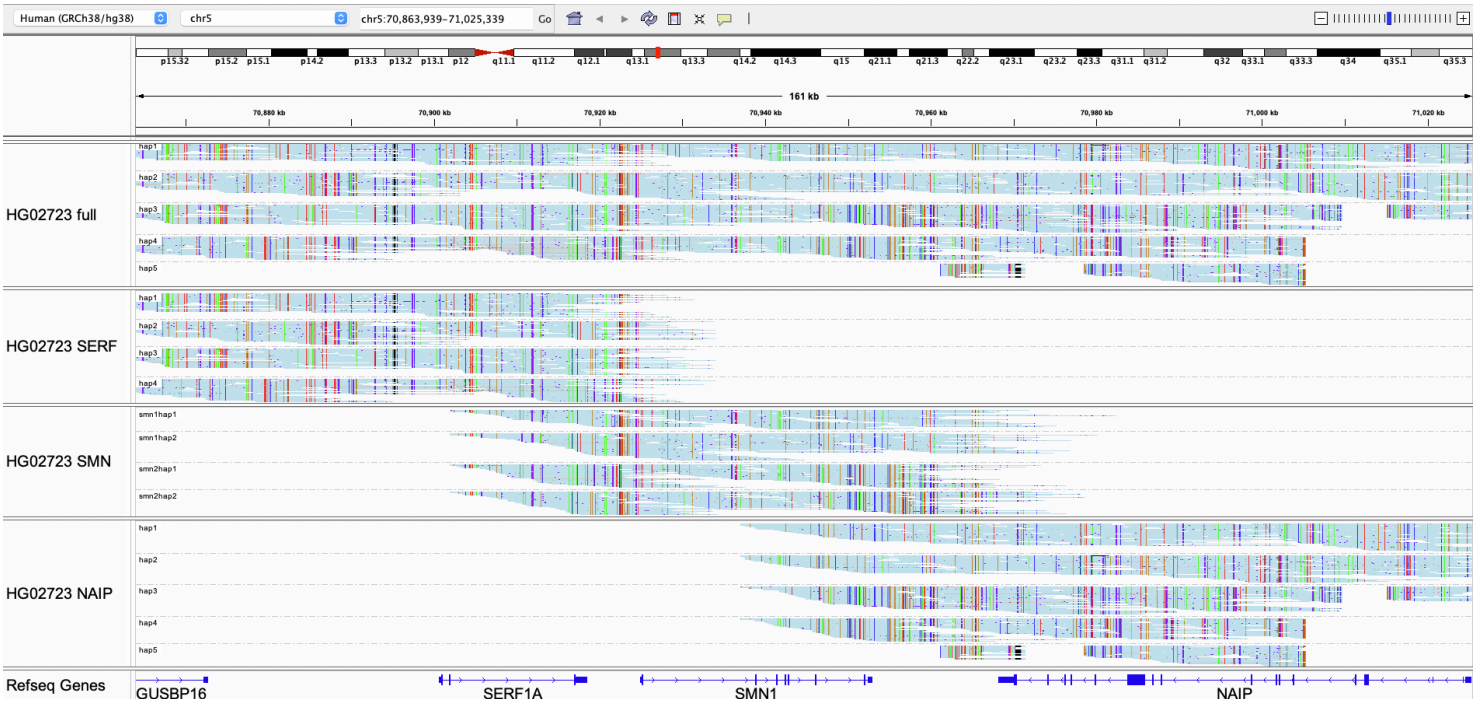
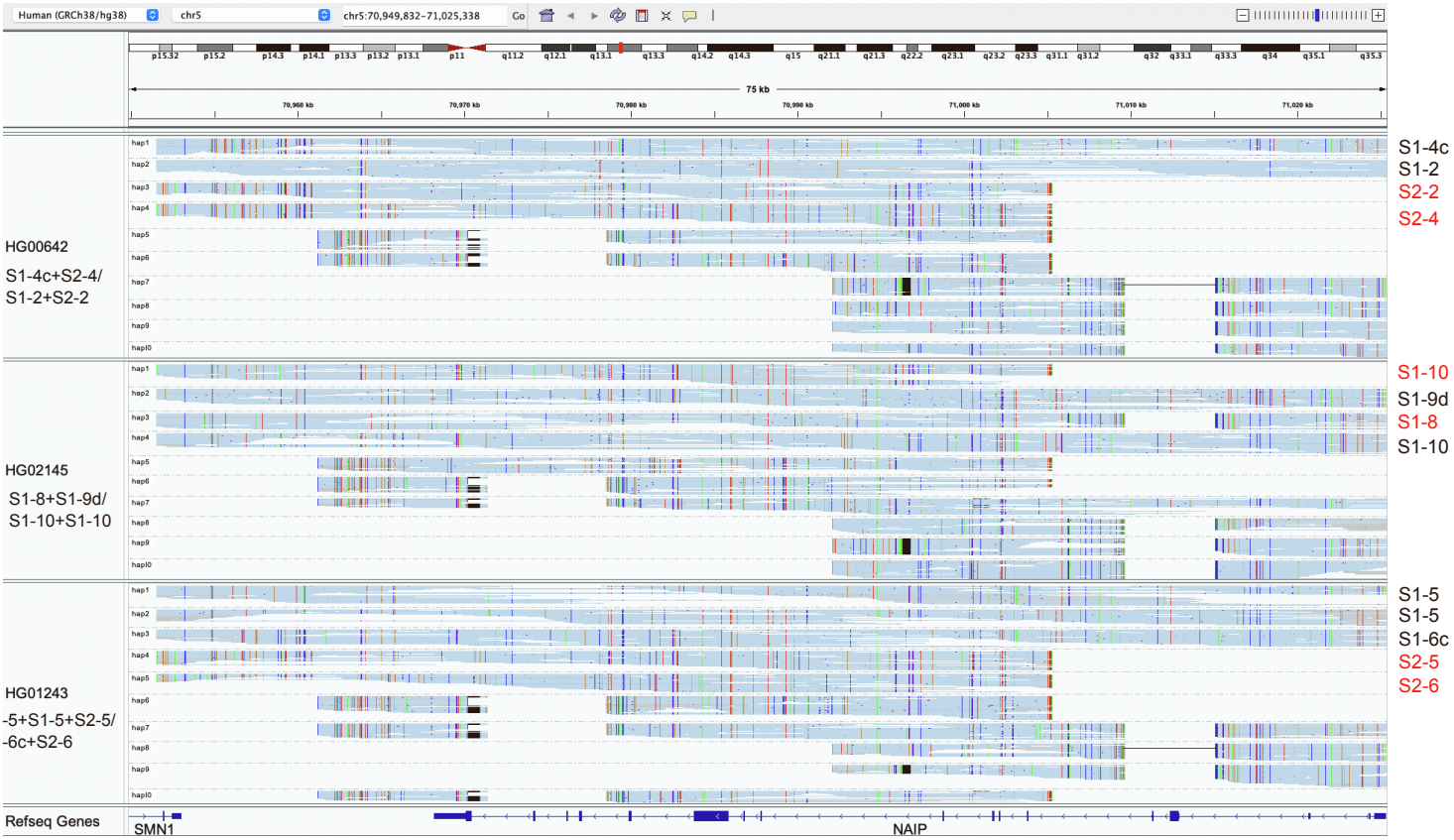


Figure S6. Interesting *SMN1* alleles and their "gene locations" suggested by intact or truncated *NAIP.*

Examples of samples with interesting *SMN1* alleles. Top: HG00642 contains a *SMN1* "c" haplotype (S1-4c) that is located in the "*SMN1* location" (intact *NAIP*). Middle: HG02145 has two two-copy *SMN1* alleles without *SMN2* (S1-8+S1-9d, S1-10+S1-10), each of which contains an *SMN1* copy that is located in the "*SMN2* location" (truncated *NAIP*). Bottom: HG01243 has a two-copy *SMN1* allele that contains *SMN2* (S1-5+S1-5+S2-5) and both *SMN1* copies are located in the "*SMN1* location" (intact *NAIP*). Haplotypes marked in red indicate those that are in the "*SMN2* location" (truncated *NAIP*).



Supplemental tables

Table S1. Validation sample details. (Excel Spreadsheet)

Table S2. Pedigree information.

Data source	EUR	AFR	EAS	SAS	AMR	unknown	mixed ancestry	notes
RadboudUMC (Kucuk et al. in review[6])	8	0	0	0	0	0	0	30X HiFi WGS for all samples
100,000 Genomes Project (GEL)	1	0	0	0	0	0	0	30X HiFi WGS for all samples
GIAB	1	0	1	0	0	0	0	30X HiFi WGS for all samples
ChineseQuartet	0	0	1	0	0	0	0	30X HiFi WGS for all samples
HPRC/1kGP	0	29	16	24	28	0	0	30X HiFi WGS genomes for the proband and 30X short read WGS data for the parents
GA4K	188	8	0	2	7	9	18	20-30X HiFi WGS genomes for the proband and 5-10X HiFi genomes for the parents
Total	198	37	18	26	35	9	18	

Table S3. Population sample results. (Excel Spreadsheet)

Table S4. *SMN2* allele frequencies across five ethnic populations.

	EUR		EAS		SAS		AMR		AFR	
no *SMN2*	54	12.9%	5	11.9%	13	25.0%	12	17.1%	43	49.4%
S2-1	163	39.1%	33	78.6%	25	48.1%	38	54.3%	27	31.0%
S2-2	80	19.2%	3	7.1%	7	13.5%	5	7.1%	1	1.1%
S2-3	61	14.6%	0	0.0%	7	13.5%	4	5.7%	1	1.1%
S2-4	7	1.7%	0	0.0%	0	0.0%	1	1.4%	0	0.0%
S2-5	1	0.2%	0	0.0%	0	0.0%	2	2.9%	1	1.1%
S2-6	0	0.0%	0	0.0%	0	0.0%	2	2.9%	2	2.3%
S2-7	0	0.0%	0	0.0%	0	0.0%	0	0.0%	2	2.3%
S2-8	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	1.1%
S2-9	0	0.0%	0	0.0%	0	0.0%	0	0.0%	8	9.2%
*SMN2*Δ*7–8*	44	10.6%	0	0.0%	0	0.0%	5	7.1%	0	0.0%
more than one copy of *SMN2*	7	1.7%	1	2.4%	0	0.0%	1	1.4%	1	1.1%
Total	417		42		52		70		87	

Table S5. Pan-ethnic frequencies of *SMN1 (SMN2)* haplotypes on alleles without *SMN2 (SMN1)*.

SMN1	*SMN2*	# alleles	percentage
S1-1	no *SMN2*	64	47.1%
S1-2		11	8.1%
S1-3		11	8.1%
S1-6		1	0.7%
S1-9		3	2.2%
S1-10		8	5.9%
two copies of *SMN1*		38	27.9%
Total		136	
no *SMN1*	S2-1	1	11.1%
	S2-2	4	44.4%
	S2-2+S2-2	1	11.1%
	*SMN2*Δ7–8+S2-2	1	11.1%
	S2-1+S2-1+S2-1c	1	11.1%
	S2-3+S2-1+S2-1c	1	11.1%
	Total	9	

Table S6. Variants shared within each haplogroup. (Excel spreadsheet)

Supplemental references

1. Chen X, Sanchis-Juan A, French CE, et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet Med*. 2020;22(5):945-953. doi:10.1038/s41436-020-0754-0

2. Prodanov T, Bansal V. Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing. *Nat Commun*. 2022;13(1):3221. doi:10.1038/s41467-022-30930-3

3. Sugarman EA, Nagan N, Zhu H, et al. Pan-ethnic carrier screening and prenatal diagnosis for spinal muscular atrophy: clinical laboratory analysis of >72 400 specimens. *Eur J Hum Genet*. 2012;20(1):27-32. doi:10.1038/ejhg.2011.134

4. Prior TW, Krainer AR, Hua Y, et al. A Positive Modifier of Spinal Muscular Atrophy in the *SMN2* Gene. *Am J Hum Genet*. 2009;85(3):408-413. doi:10.1016/j.ajhg.2009.08.002

5. Blauw HM, Barnes CP, van Vught PWJ, et al. *SMN1* gene duplications are associated with sporadic ALS. *Neurology*. 2012;78(11):776-780. doi:10.1212/WNL.0b013e318249f697

6. Kucuk et al. Comprehensive de novo mutation discovery with HiFi long-read sequencing. 2022. In review.