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*).

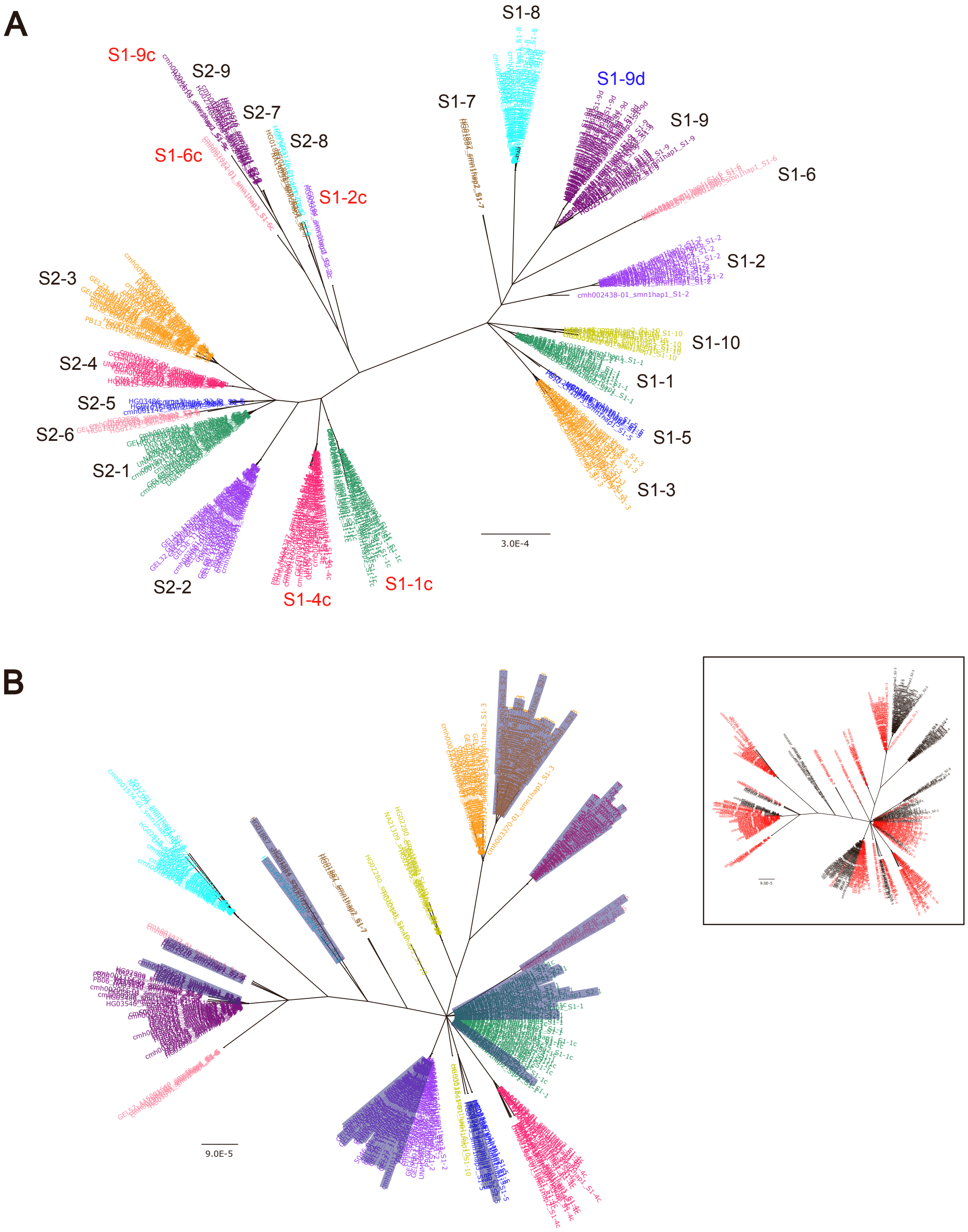


Figure S2. Discrepant PSV sites across populations.

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.

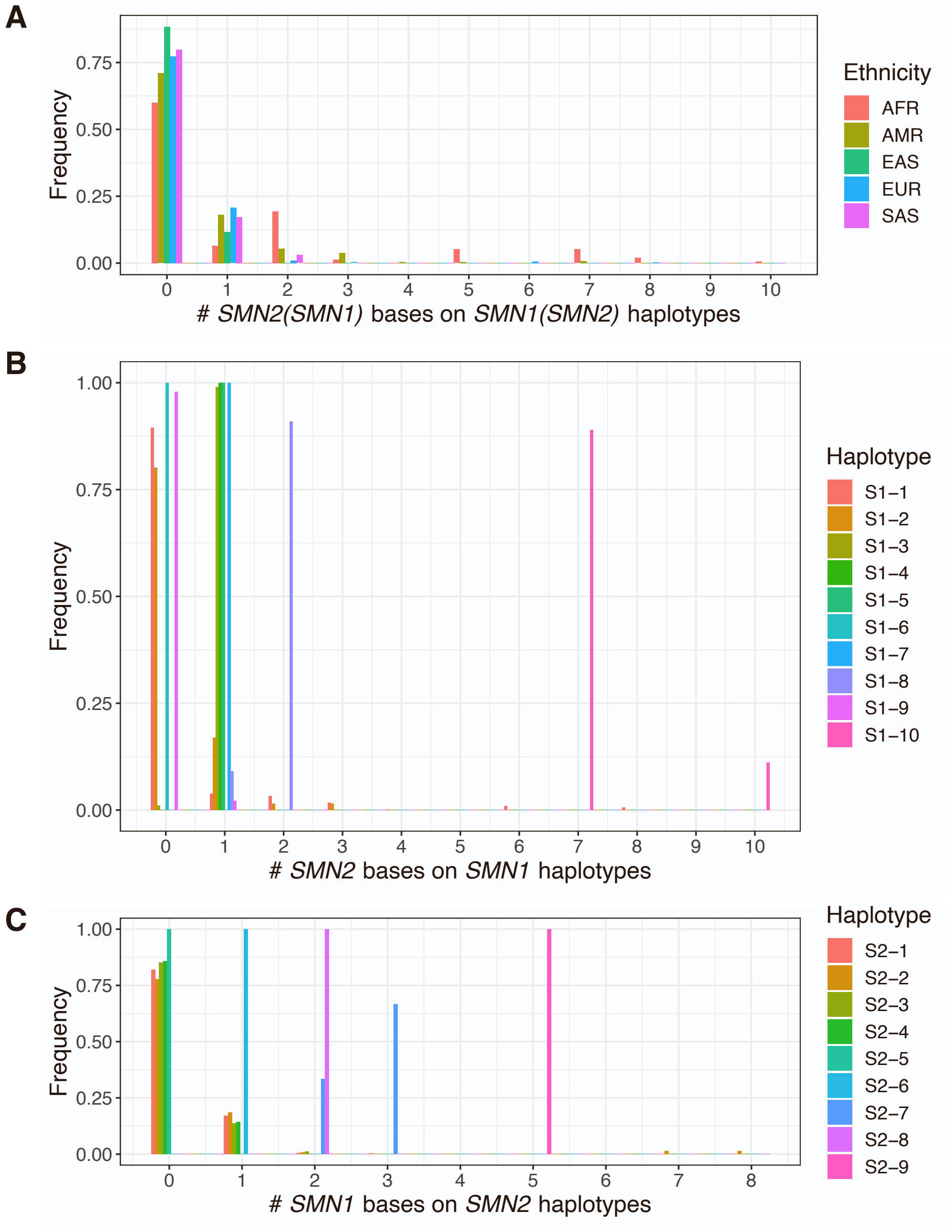


Figure S3. Sequence similarity between *SMN1* and *SMN2* haplogroups.

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.

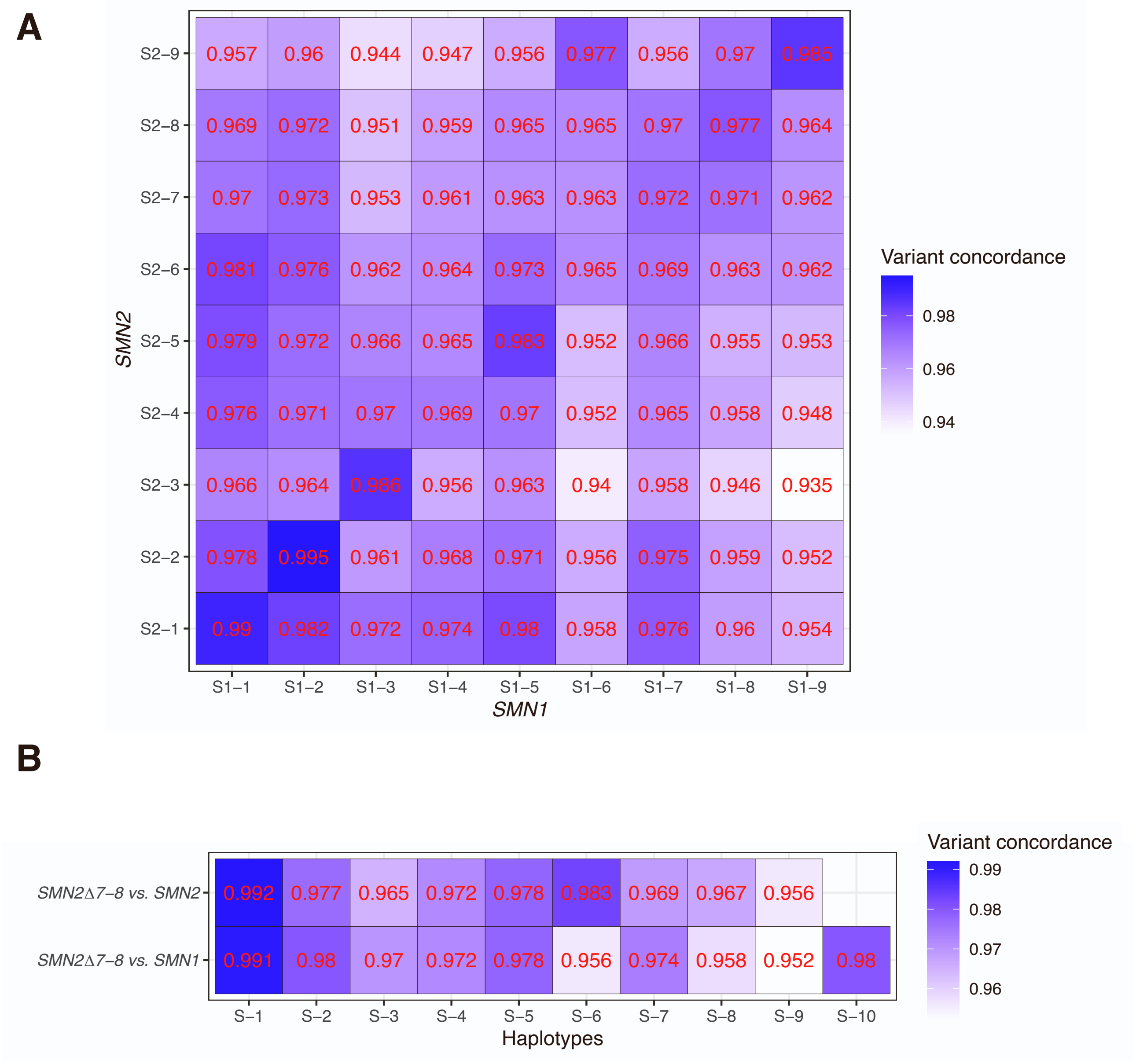


Figure S4. IGV snapshot of *SMN2* haplotypes with the downstream region similar to *SMN1.*

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).

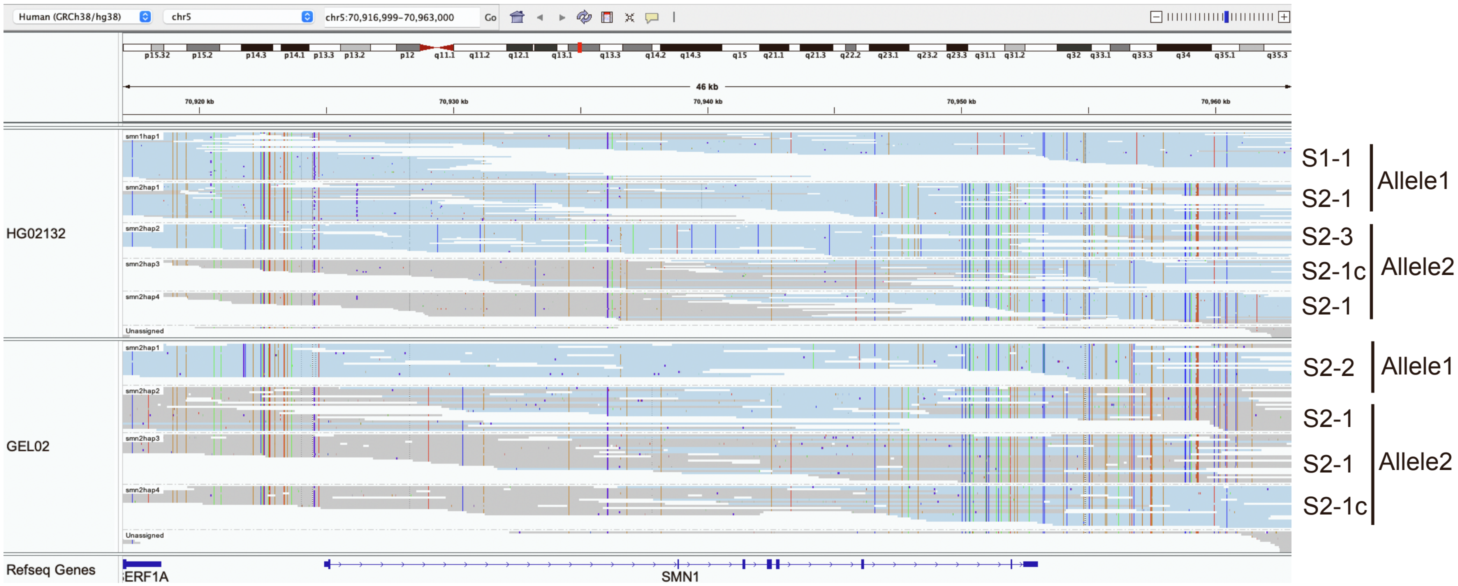
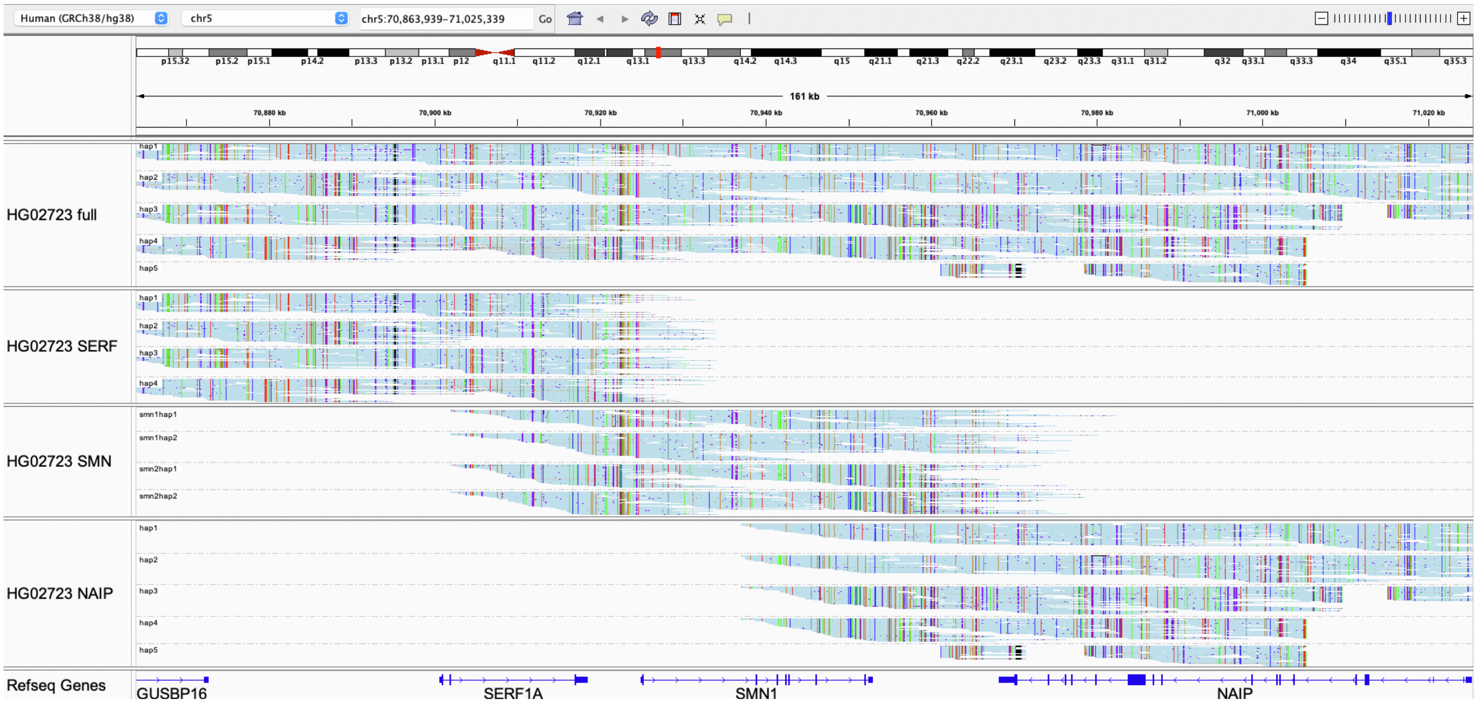


Figure S5. Phasing *SMN1*/*SMN2* with nearby genes: *SERF1A*/*SERF1B* and *NAIP.*

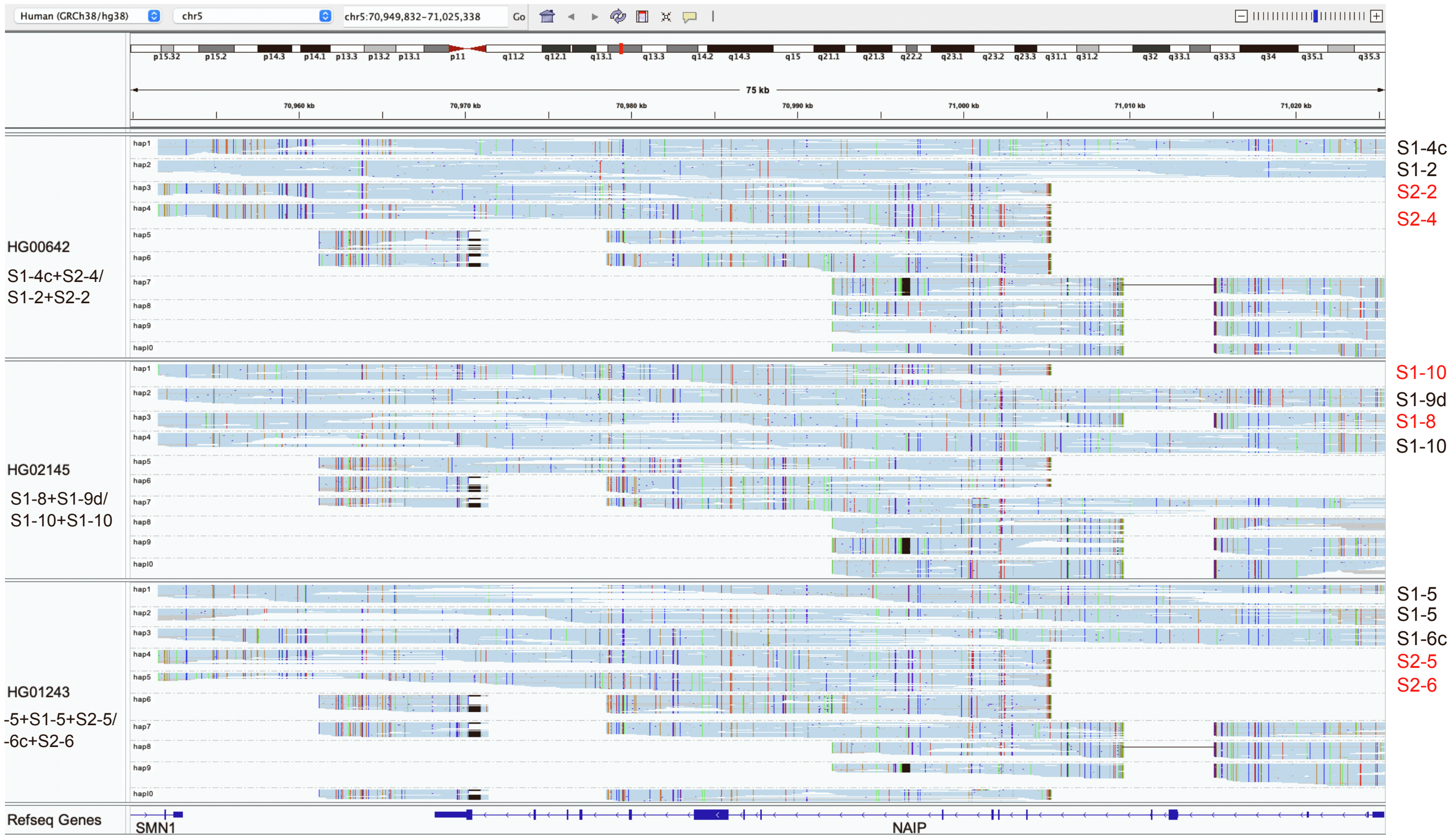
Phasing through a 161kb region containing *SERF1A*/*SERF1B*, *SMN1*/*SMN2* and *NAIP* (or its pseudogene) (top panel) is limited by read length in most samples. In order to study the structure of the bigger region, phasing of individual genes was conducted instead for *SERF1A*/*SERF1B* (second panel), *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.



Human (GRCh38/hg38) | chr5 | chr5:70,863,939-71,025,339 | Go

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∆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.