Supplemental information for:

Long-read sequencing identifies copy-specific markers of *SMN* gene conversion in spinal muscular atrophy

M.M. Zwartkruis[1,2], M.G. Elferink[2], D. Gommers[1,2], I. Signoria[1], L. Blasco-Pérez[3,4], M. Costa-Roger[3,4], J. van der Sel[1,2], I.J. Renkens[2,5,6], J.W. Green[1], J.V. Kortooms[1], C. Vermeulen[5,7], R. Straver[5,7], H.W.M. van Deutekom[2], J.H. Veldink[1], F. Asselman[1], E.F. Tizzano[3,4], R.I. Wadman[1], W.L. van der Pol[1], G.W. van Haaften[2*], E.J.N. Groen[1*]

Table of contents

Supplemental note 1: sequence variants in the *SMN* gene

Three missense, two synonymous and two untranslated region (UTR) variants were found in SMA patients (**Table 2**). Variant c.542A>G is described in the main text. Missense variant c.77G>A (p.Gly26Asp) is located in exon 1 of *SMN2*, near the Gemin2-binding domain of the SMN protein [1]. It is predicted to be pathogenic, and was found in *SMN2* in two patients with concordant and discordant (more severe) phenotype, making this variant a candidate negative modifier. A third missense variant, c.593C>T (p.Pro198Leu), was found in exon 4 in two out of four *SMN2* haplotypes of a discordant patient with SMA type 2b. Pathogenicity predictions from SIFT, PolyPhen and CADD and the relatively severe phenotype of this patient suggest that this variant might be a negative modifier. Synonymous variants c.84C>T and c.462A>G were not predicted to be pathogenic and were found in both concordant and discordant patients. 5' UTR variant c.-14C>T, intronic variant c.81+45C>T and 3' UTR variant c.*58G>A were found in the same haplotype of a concordant patient, and were not predicted to be pathogenic, whereas c.-14C>T and c.81+45C>T were previously reported in a different, severely affected, discordant patient [2]. Other intronic variants of unknown functional relevance, located in 100bp exon flanks are also reported (**Table 2**). No variants were found in the 100bp upstream of the translation start site, suggested to contain most critical promoter elements [3].

Supplemental figures

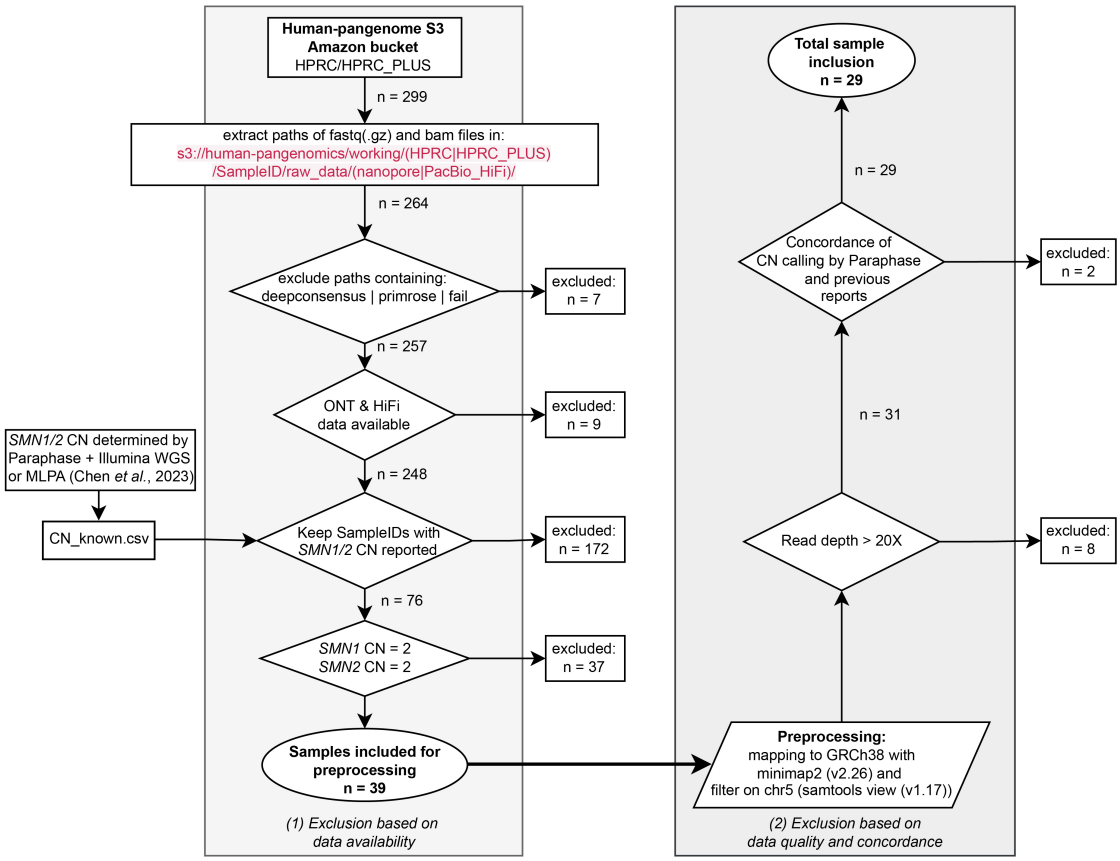

Figure S1: Flowchart of data inclusion of Human Pangenome Consortium (HPRC) working data bucket.

Sample inclusion was performed by exclusion based on data availability (1) and data quality and concordance (2). (1) All available paths to nanopore and PacBio HiFi BAM files were extracted, returning 264 out of 299 HPRC samples. File paths were excluded when they contained “deepconsensus”, “primrose” or “fail”, returning 257 unique samples. Both ONT and PacBio HiFi data was available for 248 samples. Next, samples were selected based on the availability *SMN1/2* CN previously determined by Paraphase and Illumina WGS and/or MLPA (n = 76) [4]. We selected samples with two *SMN1* and two *SMN2* copies (n = 39) for preprocessing. (2) ONT and PacBio HiFi BAM files were mapped to GRCh38 and filtered for chromosome 5. Samples were excluded if ONT or PacBio HiFi read depth was lower than 20x. Paraphase was run on the PacBio HiFi data and copy number calling was compared to MLPA results. Only concordant samples were included in the final sample set for further analysis with HapSMA (n = 29).

1000G: 1000 Genomes; CN: copy number; HPRC: Human Pangenome Reference Consortium; MLPA: multiplex ligation-dependent probe amplification;

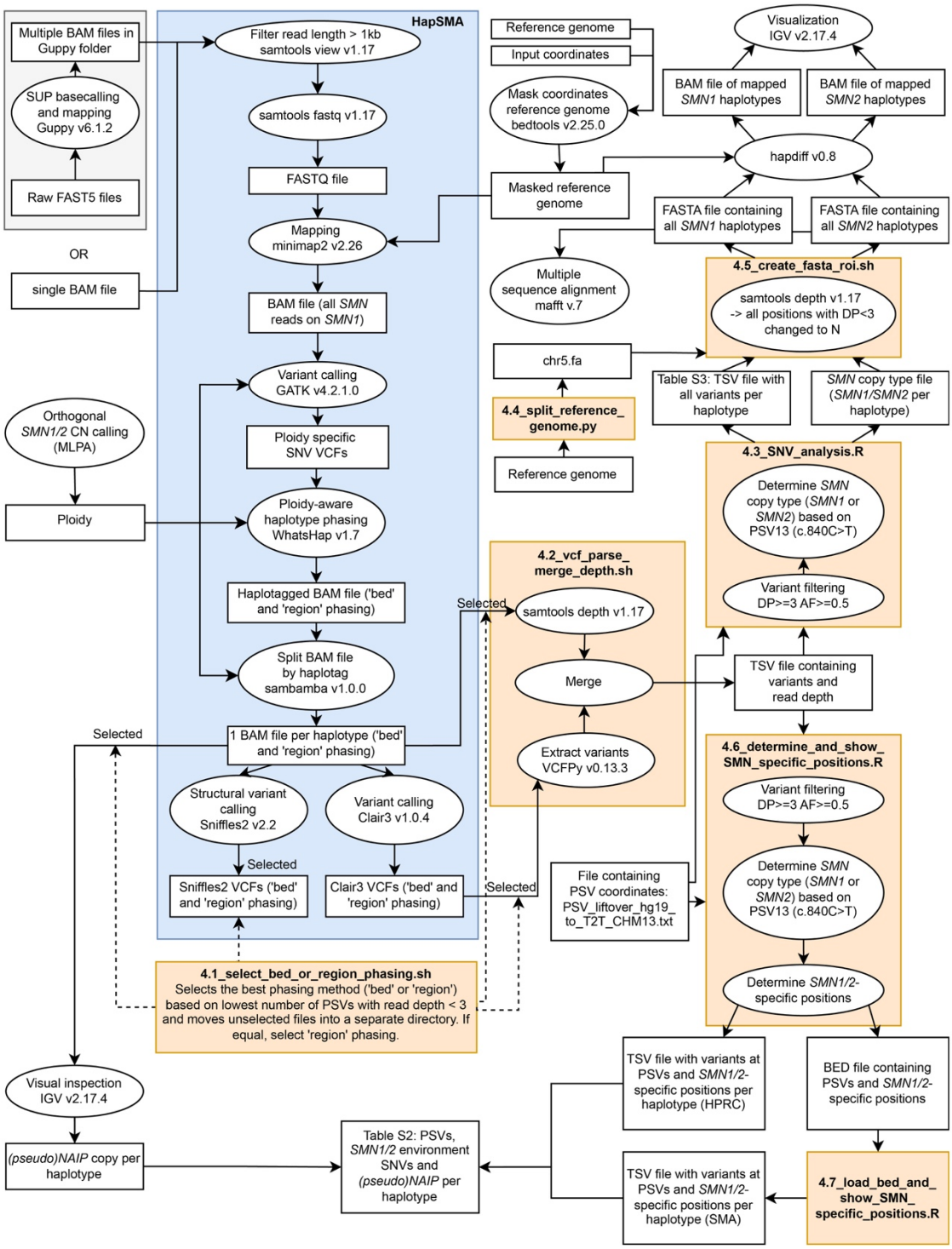

Figure S2: Flowchart of HapSMA and downstream workflow used in this study.
With the setting bam_single_remap, HapSMA (blue) maps a BAM file to a masked reference genome and performs haplotype phasing based on the ploidy of the sample (determined with orthogonal methods, such as MLPA). It outputs haplotagged and split BAM files and VCF files with two haplotype phasing methods: standard 'region' phasing and 'bed' phasing based on a manually curated bed file. Post-processing scripts (orange) were used to generate the results shown in this study. Selection of 'bed' or 'region' phasing approach, filtering of variants was performed after running HapSMA, and can therefore be easily customized. A more detailed

description of these scripts can be found on our Github repository (see **Availability of data and materials**).

CN: copy number; HPRC: Human Pangenome Reference Consortium; IGV: Integrative Genomics Viewer; MLPA: multiplex ligation-dependent probe amplification; PSV: paralogous sequence variant; SUP: super accuracy; TSV: tab-separated values.

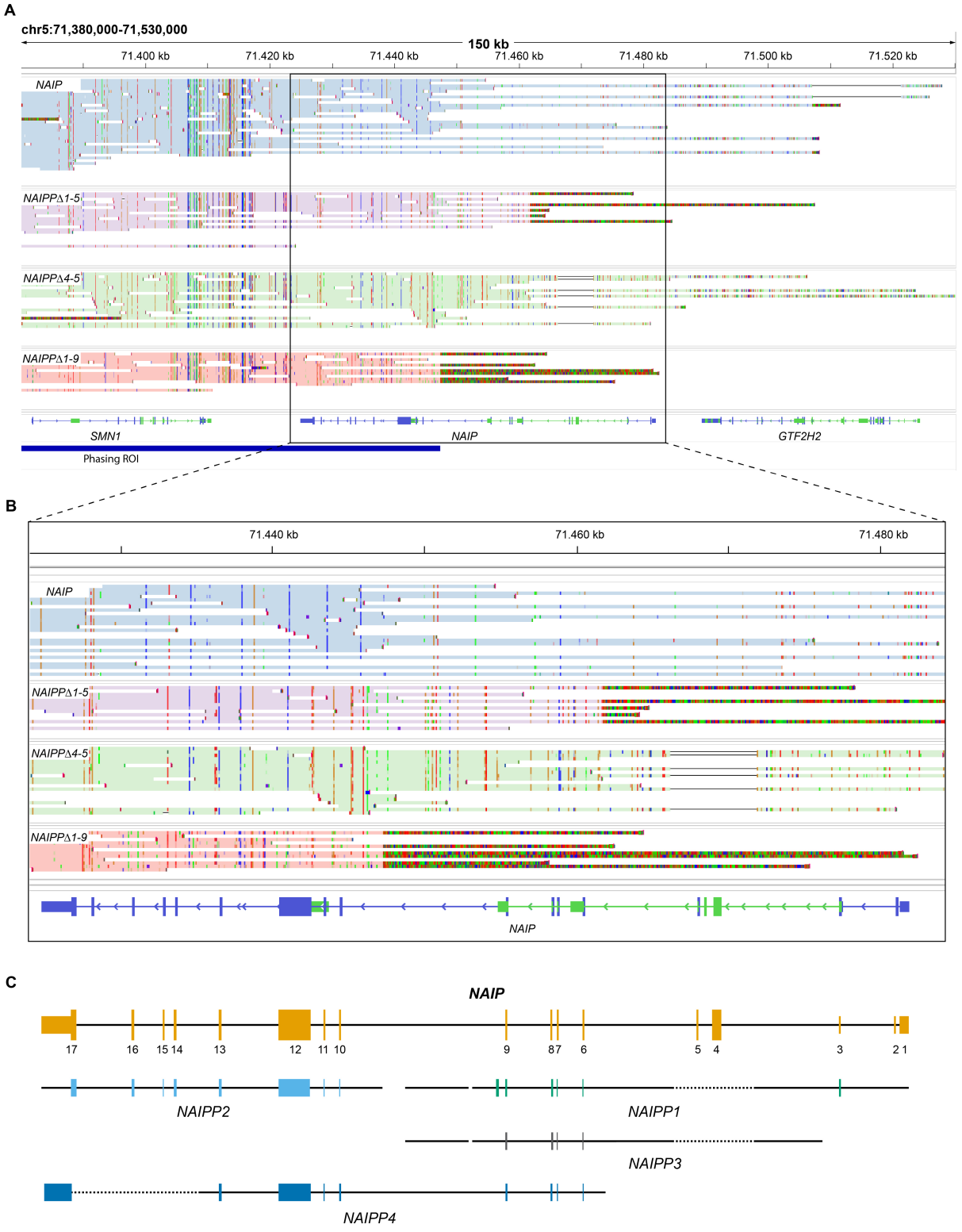

Figure S3: The *SMN* genes can be followed by different *(pseudo)NAIP* genes

(A) IGV snapshot of haplotypes in which different *NAIP* and *pseudoNAIP (NAIPP)* genes follow *SMN1/2*: full-length *NAIP*, with sequencing reads covering the entire gene; *NAIPPΔ4-5* missing exon 4-5; *NAIPPΔ1-5* missing exon 1-5; and *NAIPPΔ1-9* missing exon 1-9. Deletions are shown by thin black lines, soft-clipping is shown as intensely, non-uniformely colored read segments.

(B) Zoomed in image of (A).

(C) Different *NAIPP* genes aligned to full-length *NAIP*, at scale with (B). *NAIP* exons are numbered, and protein-coding exons are displayed wider than untranslated regions. Black lines represent introns and flanks of the *NAIPP* genes, that align to the *NAIP* gene. Dashed lines represent parts that are missing in the *NAIPP* gene relative to the *NAIP* gene. *NAIPP*Δ*4-5* from (B) resembles the *NAIPP1* deletion; *NAIPP*Δ*1-5* resembles the *NAIPP4* breakpoint; and *NAIPP*Δ*1-9* resembles the *NAIPP2* breakpoint.

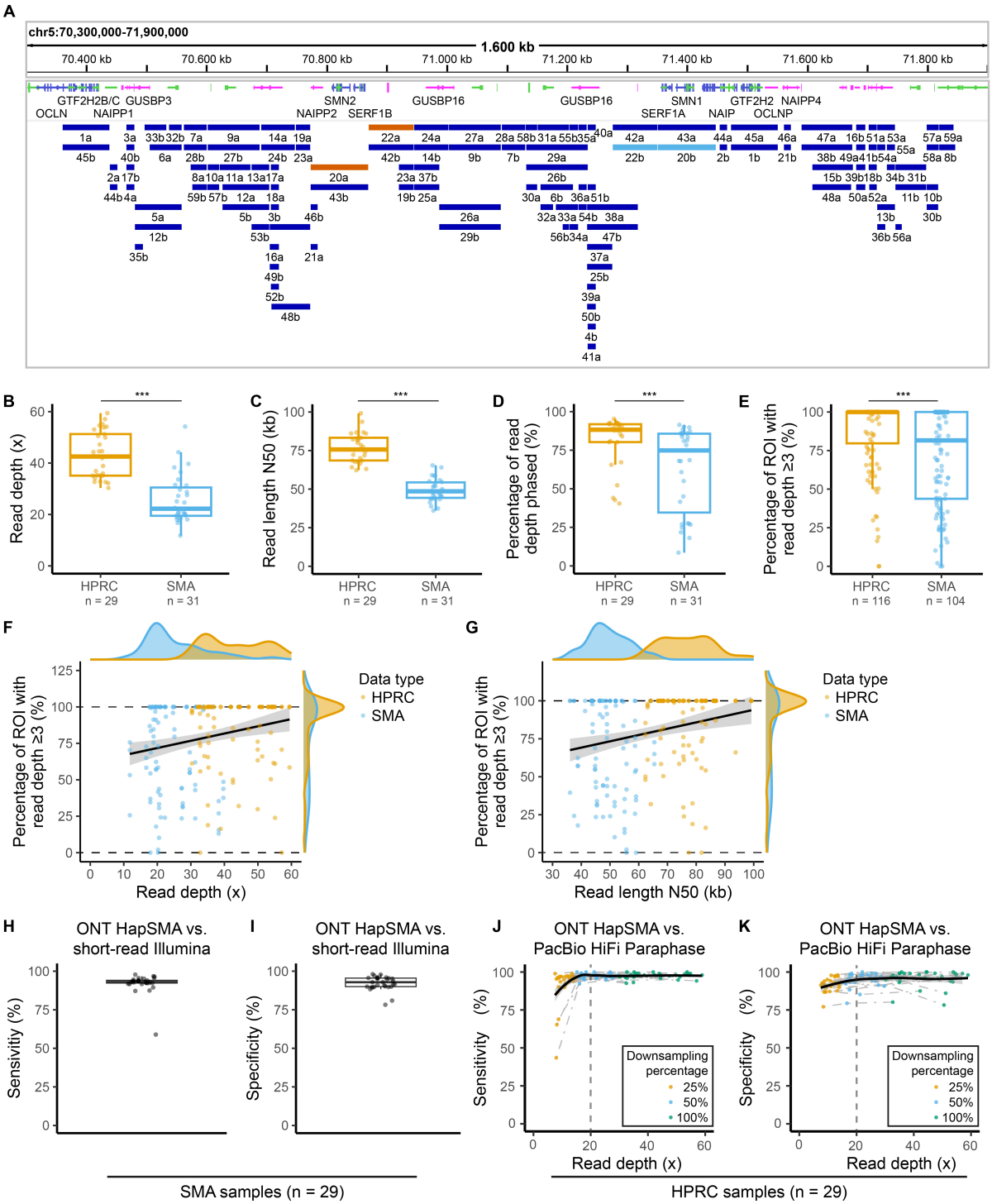

Figure S4: Performance of phasing and variant calling

(A) IGV snapshot of duplication segments on the *SMN* locus resulting from nucmer segmental duplication analysis of T2T-CHM13 chromosome 5 versus chromosome 5 (**Figure 1A**). Duplicated segments share the same number. Segments 20a and 22a (chr5:70,772,138-70,944,284) were masked, whereas segments 20b and 22b (chr5:71,274,893-71,447,410) were used as ROI for phasing.

(B-E) Overview of sequencing characteristics. **(E)** HPRC samples had median read depth of 42.6x (range 30.3-59.4x, IQR = 16.2x), significantly higher than SMA samples (22.3x (range 11.8-54.2x, IQR = 11.0x); Wilcoxon rank sum test, W = 816, p = 2.839e-09). (F) Read length

N50 was different between HPRC and SMA (75.7 kb (range 62.3-99.2 kb, IQR = 14.7 kb) and 48.6 kb (range 36.1-65.4 kb, IQR = 10.0 kb), respectively; two-sided t-test, t = 12.889, df = 58, p < 2.2e-16). (G) Percentage of read depth that was phased into a haplotype was different between HPRC (median = 88.3% (range 40.5-95.3%, IQR = 11.6%)) and SMA samples (median = 74.9% (range 8.6-91.5%, IQR = 51.1%); Wilcoxon rank sum test, W = 685, p = 0.0003616). (H) Percentage of phasing ROI with read depth ≥3 was higher for HPRC (median = 100.0% (range 0.0-100.0%, IQR = 20.3)) than for SMA (median = 81.6% (range 0.0-100.0%, IQR = 56.2%); Wilcoxon rank sum test, W = 7982, p = 1.1e-05).

(F-G) Correlation between read depth (F) or read length N50 (G) and performance of haplotype phasing, with one datapoint per haplotype (n = 116 for HPRC and n = 104 for SMA). There is a significant correlation between total read depth on 30Mb region surrounding the *SMN* locus and the percentage of the phasing ROI with read depth ≥3 (Spearman correlation test, Spearman's rho = 0.3048, p = 4.118e-06) and between read length N50 and the percentage of the phasing ROI with read depth ≥3 (Spearman correlation test, Spearman's rho = 0.2475, p = 0.0002092).

(H-I) Comparison of variant calling within the long-range PCR region by long-read sequencing (ONT) and short-read sequencing (Illumina), determined by intersecting the SNV variant positions between the two analysis methods for each sample. The overall median sensitivity was 93.0% (range 58.93-97.8%, IQR = 1.81) and the median specificity was 92.9% (range 78.3-98.1%, IQR = 5.52).

(J-K) Sensitivity and specificity of HapSMA compared to Paraphase reported per sample, retaining 100% of reads, downsampled to 50% of reads, or downsampled to 25% of reads before running HapSMA. **(J)** Median sensitivity is 98.0% (range 92.6-100%, IQR = 2.3%) when retaining all reads, 97.9% (range 92.1-100%, IQR = 2.9%) at 50% downsampling and 95.8% (range 43.5-99.0%, IQR = 4.6%) at 25% downsampling. **(K)** Median specificity is 97.9% (range 78.3-100.0%, IQR = 4.8%) when retaining all reads, 95.3% (range 79.4-100.0%, IQR = 5.0%) at 50% downsampling and 92.0% (range 77.1-98.1%, IQR = 4.2%) at 25% downsampling. In total, 386 unique SNV variant positions were called, of which 290 were called by both HapSMA and Paraphase. 21 SNV positions were not called by HapSMA and 75 SNV positions were not called by Paraphase. Sensitivity and specificity of HapSMA SNV calls decline below a read depth of 20x (grey vertical line).

HPRC: Human Pangenome Reference Consortium; Mb: megabases; ROI: region of interest; ONT: Oxford Nanopore Technologies.

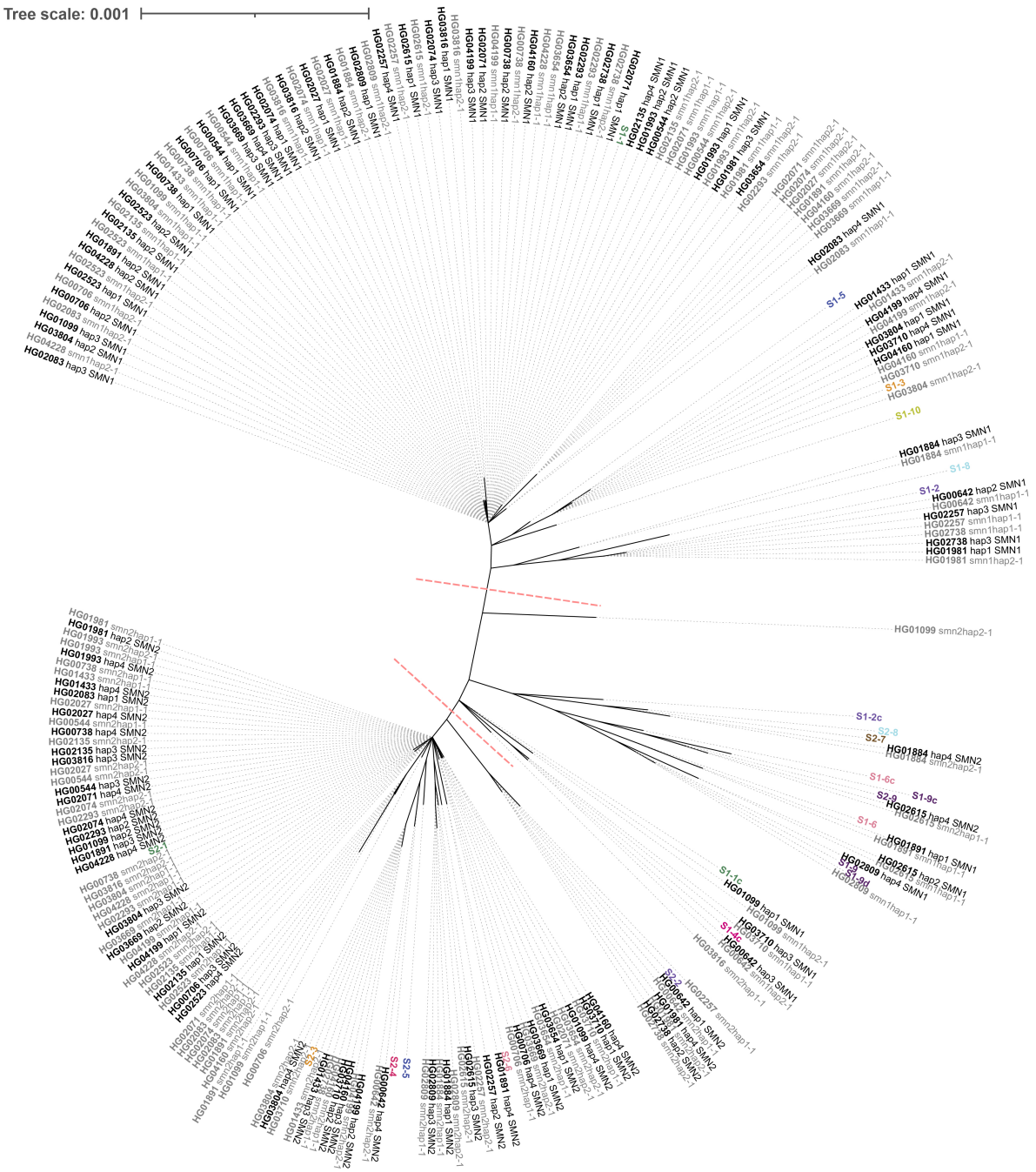

Figure S5: Phylogenetic tree of multiple sequence alignment (MSA) of HPRC haplotypes generated with HapSMA and Paraphase variants.

97 out of 116 HPRC haplotypes that could be called on all positions between coordinates chr5:70,917,100-70,961,220 (GRCh38) after HapSMA analysis were included in the MSA. Haplotypes generated with Paraphase are shown in grey, haplotypes generated with HapSMA are shown in black. Colored haplotypes represent the haplogroups as reported previously by Chen *et al.* [4]. The dashed red lines show the distinguished clades of *SMN1* and *SMN2*. 96.9% (94/97) of HapSMA haplotypes cluster within the same clade as the Paraphase haplotypes of the same sample and *SMN* copy type. The major haplogroups present in the HPRC dataset are S1-1 for *SMN1* haplotypes and S2-1 for *SMN2* haplotypes.

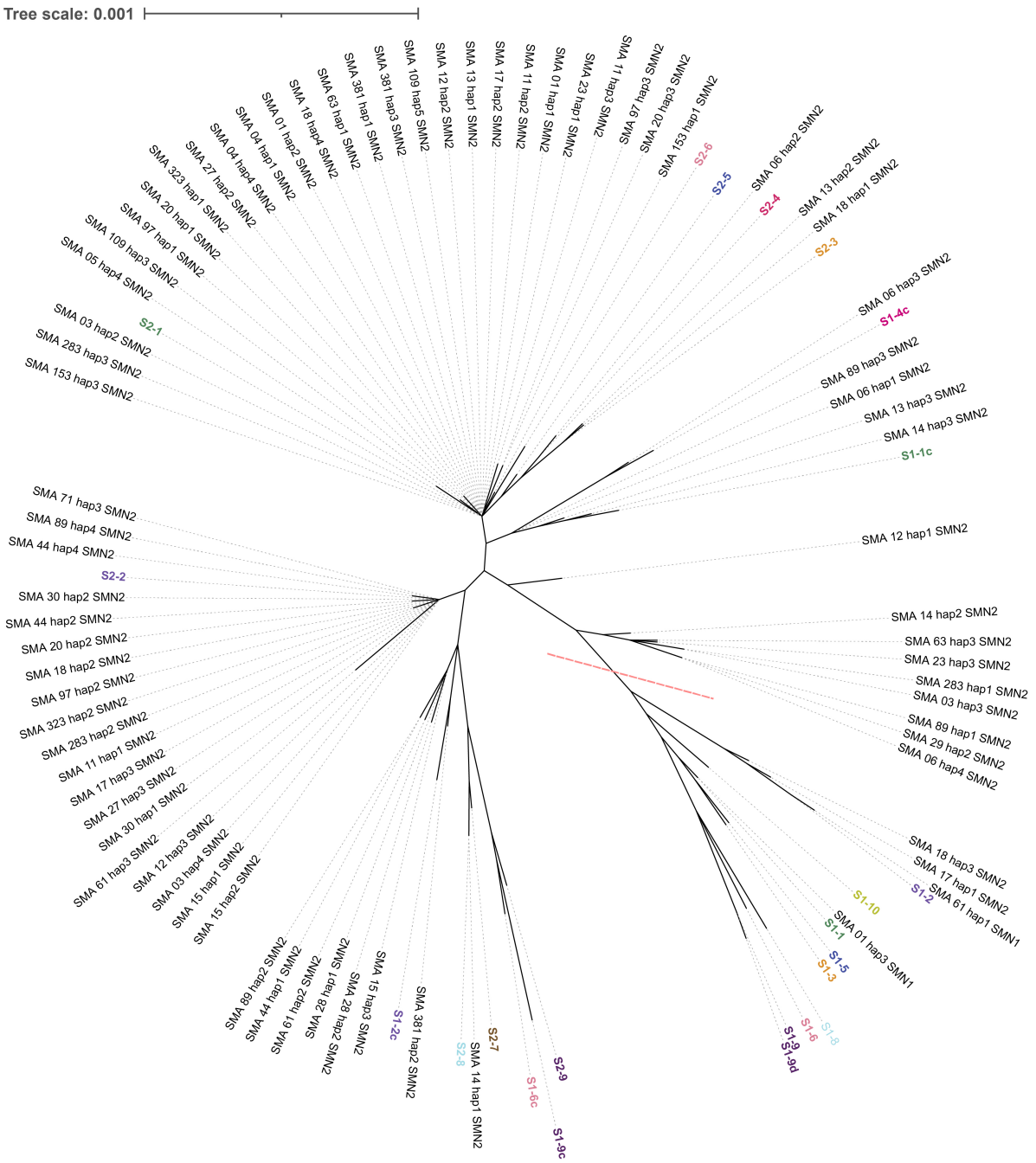

Figure S6: Phylogenetic tree of multiple sequence alignment (MSA) of SMA haplotypes generated with HapSMA variants.

75 out of 104 SMA haplotypes that could be called on all positions between coordinates chr5:70,917,100-70,961,220 (GRCh38) after HapSMA analysis were included in the MSA.

Haplotypes generated with HapSMA are shown in black. Colored haplotypes represent the haplogroups as reported previously by Chen *et al.* [4]. The dashed red line shows the distinguished clades of *SMN1* and *SMN2*. All *SMN2* haplotypes cluster in the *SMN2* clade and the *SMN1* haplotypes cluster within the *SMN1* clade. Haplogroups S2-1 and S2-2 are the most similar to the majority of haplotypes of the SMA sample set.

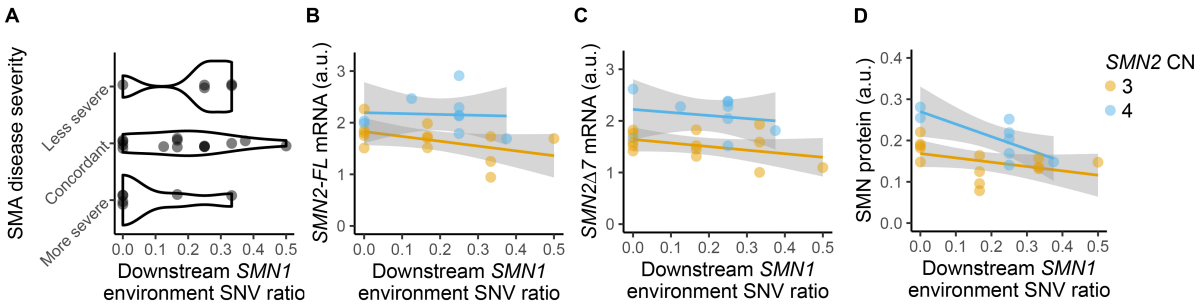

Figure S7: Link between downstream *SMN1* environment SNV ratio and disease severity, mRNA expression and protein expression in patients with homozygous *SMN1* deletion and 3 or 4 *SMN2* copies

Downstream *SMN1* environment SNV ratio was calculated as the number of haplotypes with downstream *SMN1* environment SNVs divided by total number of haplotypes.

(A) Comparison between disease severities. *SMN1* environment SNV ratio did not significantly differ between less severe (n = 5, median = 0.25, range = 0.0-0.33, IQR = 0.083), concordant (n = 14, median = 0.17, range = 0.0-0.5, IQR = 0.22), and more severe (n = 6, median = 0.0, range = 0.0-0.33, IQR =0.13) patients (Kruskal-Wallis rank sum test, $\chi^2(2) = 3.0944$, p = 0.2128).

(B) *SMN2-FL* mRNA expression differed between copy groups but was not associated with downstream *SMN1* environment SNV ratio (simple linear regression, used model: expression = 0.2641 + 0.5091 * (copy group) - 0.7211 * (ratio), adjusted R-squared: 0.3561, F(2,19) = 6.808, p = 0.005896 for complete model; p = 0.00236 for copy group and p = 0.15782 for ratio).

(C) *SMN2Δ7* mRNA expression differed between copy groups but was not associated with downstream *SMN1* environment SNV ratio (simple linear regression, used model: expression = -0.1635 + 0.6000 * (copy group) - 0.6632 * (ratio), adjusted R-squared: 0.4742, F(2,19) = 10.47, p = 0.00086 for complete model; p = 0.000271 for copy group and p = 0.161303 for ratio).

(D) SMN protein expression differed between copy groups and was associated with downstream *SMN1* environment SNV ratio (simple linear regression, used model: expression = -0.01258 + 0.06330 * (copy group) - 0.15822 * (ratio), adjusted R-squared: 0.4588, F(2,18) = 9.479, p = 0.001542 for complete model; p = 0.00183 for copy group and p = 0.01208 for ratio).

CN: copy number; *SMN1-FL*: full-length *SMN1*; *SMN2-FL*: full-length *SMN2*; *SMN2Δ7*: *SMN2* lacking exon 7.

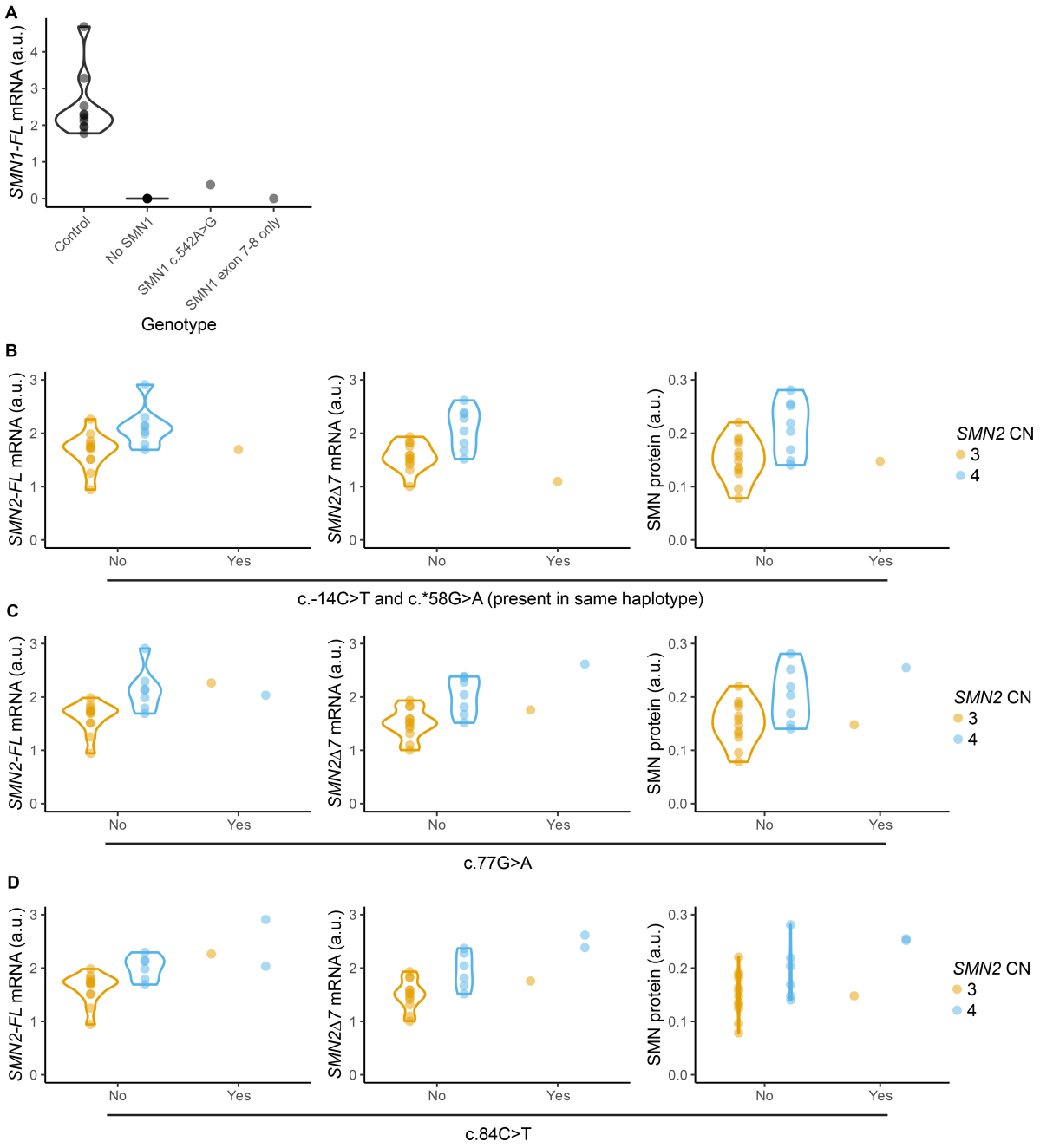

Figure S8: Link between exonic variants and *SMN* mRNA and protein expression
Each datapoint represents one individual. Only the variants for which expression data was available, are shown. Statistics were not performed due to small sample size per group. **(A)** *SMN1-FL* mRNA expression in fibroblasts of healthy controls, SMA patients with homozygous deletion of *SMN1*, one SMA patient with the c.542A>G variant in *SMN1* and one patient with only exon 7 and 8 of *SMN1*.
(B-D) mRNA and protein expression in fibroblasts of patients with exonic variants in SMN2. Variants c.-14C>T and c.*58G>A (B), c.77G>A (C) and c.84C>T (D) were present in only one haplotype per patient.
CN: copy number; *SMN1-FL*: full-length *SMN1*; *SMN2-FL*: full-length *SMN2*; *SMN2Δ7*: *SMN2* lacking exon 7.

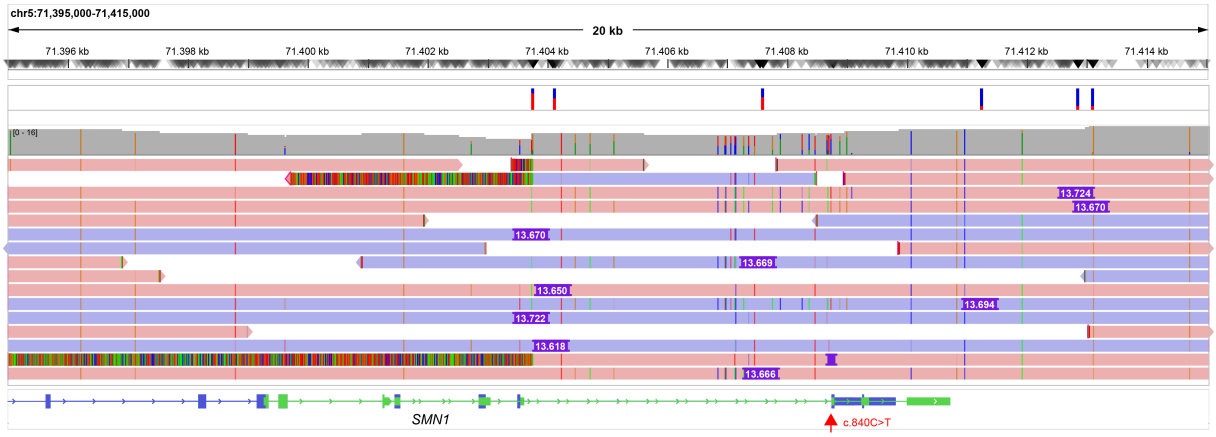

Figure S9: Structural variant (SV) in the *SMN* gene

The panel contains from top to bottom SV calls made with Sniffles2, read depth plot and sequencing reads colored by read strand. Sequencing reads of one haplotype from a sample with decreased copy number of exon 1-6 in MLPA are shown. In most ONT reads, this SV was detected as a ~13.7kb insertion containing exon 7 and 8 of *SMN1* or *SMN2*, depending on the location of the insertion upstream or downstream the c.840C>T variant (PSV13) (marked by a red arrow). This indicates that *SMN1* exon 7 and 8 are inserted downstream of *SMN2*. However, it is possible that this structural variant is a deletion of exon 1-6 of *SMN1*, as found previously [5,6], extending to the downstream area of *SMN2,* if both genes were originally in the same orientation. Both situations result in an incomplete *SMN1* copy containing only exon 7 and 8. The insertion/deletion breakpoint, as can be seen from soft-clipping of three reads, is located on Alu element AluYd8, which could be the mediator for this structural rearrangement.

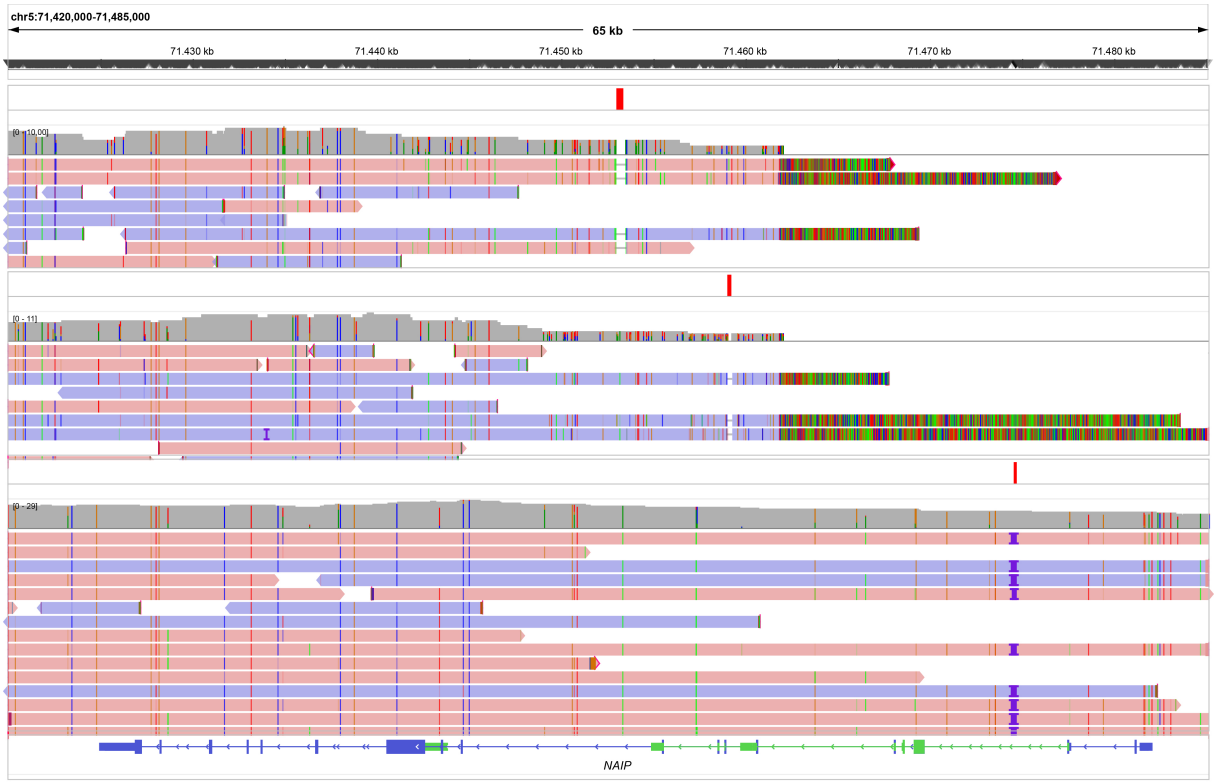

Figure S10: SVs in the *(pseudo)NAIP* genes other than the truncations shown in Figure S2

Each panel contains from top to bottom SV calls made with Sniffles2, read depth plot and sequencing reads colored by read strand. Upper panel: a 519bp deletion was found in intron 9 of *NAIP*Δ*1-5* in one SMA haplotype and of *NAIP*Δ*4-5* in one HPRC haplotype. Middle panel: a 219bp deletion was found in intron 6 of *NAIP*Δ*1-5* in three SMA haplotypes. Lower panel: in one HPRC sample, an AluYa5 insertion of 337bp was found in *NAIP* intron 3.

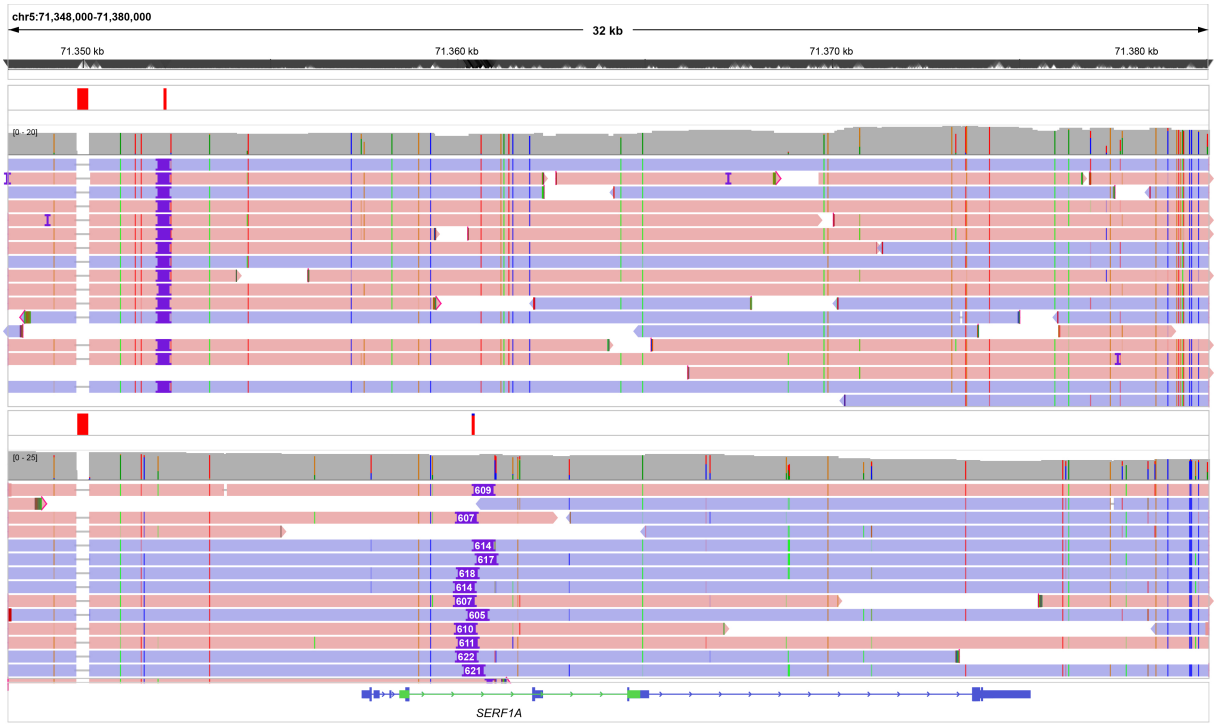

Figure S11: SVs in and near the *SERF1A/B* gene

Each panel contains from top to bottom SV calls made with Sniffles2, read depth plot and sequencing reads colored by read strand. Upper panel: Upstream of *SERF1A/B*, a 346bp AluYb8 deletion was found in 58 SMA and 56 HPRC haplotypes, and a ~330bp AluYa5 insertion in 13 SMA and four HPRC haplotypes. The insertion was only found in haplotypes where the deletion was also present. Lower panel: in intron 2 of *SERF1A/B*, two HPRC haplotypes of one sample had a ~614bp insertion.

Supplemental references

1. Detering NT, Schüning T, Hensel N, Claus P. The phospho-landscape of the survival of motoneuron protein (SMN) protein: relevance for spinal muscular atrophy (SMA). Cell Mol Life Sci. 2022;79:497.

2. Wadman RI, Jansen MD, Stam M, Wijngaarde CA, Curial CAD, Medic J, et al. Intragenic and structural variation in the SMN locus and clinical variability in spinal muscular atrophy. Brain Commun. 2020;2:fcaa075.

3. Ottesen EW, Seo J, Luo D, Singh NN, Singh RN. A super minigene with a short promoter and truncated introns recapitulates essential features of transcription and splicing regulation of the SMN1 and SMN2 genes. Nucleic Acids Res. 2024;gkad1259.

4. Chen X, Harting J, Farrow E, Thiffault I, Kasperaviciute D, Hoischen A, et al. Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. Am J Hum Genet. 2023;110:240–50.

5. Arkblad EL, Darin N, Berg K, Kimber E, Brandberg G, Lindberg C, et al. Multiplex ligation-dependent probe amplification improves diagnostics in spinal muscular atrophy. Neuromuscul Disord. 2006;16:830–8.

6. Thauvin-Robinet C, Drunat S, Saugier Veber P, Chantereau D, Cossée M, Cassini C, et al. Homozygous SMN1 exons 1–6 deletion: Pitfalls in genetic counseling and general recommendations for spinal muscular atrophy molecular diagnosis. Am J Med Genet A. 2012;158A:1735–41.