Genome Medicine

RESEARCH

Open Access

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*†] and E. J. N. Groen[1*†]

Abstract

Background The complex 2 Mb *survival motor neuron* (*SMN*) locus on chromosome 5q13, including the spinal muscular atrophy (SMA)-causing gene *SMN1* and modifier *SMN2*, remains incompletely resolved due to numerous segmental duplications. Variation in *SMN2* copy number, presumably influenced by *SMN1* to *SMN2* gene conversion, affects disease severity, though *SMN2* copy number alone has insufficient prognostic value due to limited genotype–phenotype correlations. With advancements in newborn screening and *SMN*-targeted therapies, identifying genetic markers to predict disease progression and treatment response is crucial. Progress has thus far been limited by methodological constraints.

Methods To address this, we developed HapSMA, a method to perform polyploid phasing of the *SMN* locus to enable copy-specific analysis of *SMN* and its surrounding genes. We used HapSMA on publicly available Oxford Nanopore Technologies (ONT) sequencing data of 29 healthy controls and performed long-read, targeted ONT sequencing of the *SMN* locus of 31 patients with SMA.

Results In healthy controls, we identified single nucleotide variants (SNVs) specific to *SMN1* and *SMN2* haplotypes that could serve as gene conversion markers. Broad phasing including the *NAIP* gene allowed for a more complete view of *SMN* locus variation. Genetic variation in *SMN2* haplotypes was larger in SMA patients. Forty-two percent of *SMN2* haplotypes of SMA patients showed varying *SMN1* to *SMN2* gene conversion breakpoints, serving as direct evidence of gene conversion as a common genetic characteristic in SMA and highlighting the importance of inclusion of SMA patients when investigating the *SMN* locus.

Conclusions Our findings illustrate that both methodological advances and the analysis of patient samples are required to advance our understanding of complex genetic loci and address critical clinical challenges.

†G. W. van Haaften and E. J. N. Groen contributed equally to this work.

*Correspondence:
G. W. van Haaften
g.vanhaaften@umcutrecht.nl
E. J. N. Groen
e.j.n.groen-3@umcutrecht.nl
Full list of author information is available at the end of the article

Keywords Spinal muscular atrophy, Long-read sequencing, Gene conversion, Dark genomic regions, Segmental duplications

Background

The *survival motor neuron* (*SMN)* locus on chromosome 5q, containing the *SMN1* and *SMN2* genes, consists of ~ 2 Mb complex segmental duplications of a highly repetitive nature [1]. Homozygous loss-of-function of *SMN1* causes spinal muscular atrophy (SMA), a severe neuromuscular disease [2]. Paralog gene *SMN2* differs from *SMN1* at only 15 paralogous sequence variants (PSVs) [3–5]. PSV13 (c.840C > T) in exon 7 causes most mRNA to be alternatively spliced and translated into an unstable, truncated protein (SMNΔ7) [5, 6]. However, *SMN2* splicing also produces limited full-length mRNA that is translated into functional SMN protein, sufficient for survival but causing SMA in the absence of *SMN1* [7]. Although SMA is classified as a rare disease, its incidence of between 1:6000 and 8000 newborns suggests at any moment there may be up to 200,000 prevalent cases worldwide [8]. The severity of SMA ranges from prenatal onset and neonatal death (SMA type 0) to adult onset and a normal life expectancy (SMA type 4) [9].

Loss of *SMN1* is hypothesized to be caused by homozygous deletion in severe types of SMA or gene conversion of *SMN1* to *SMN2* in comparatively milder types of SMA [10]. However, there is no direct evidence supporting this hypothesis and the frequency of such events remains undetermined. The total number of *SMN2* copies in the diploid human genome is highly variable and inversely correlates with the severity of SMA: a higher *SMN2* copy number leads to higher SMN protein levels and a relatively milder disease phenotype [11, 12]. Yet, up to 40% of patients have a discordant disease phenotype relative to their *SMN2* copy number [11, 13]. Earlier studies have identified rare *SMN2* sequence variants—such as c.859G > C and c.835-44A > G—as explanation for discordancy [14, 15]. Copy number variation of surrounding genes such as *GTF2H2*, *(pseudo)NAIP* and *SERF1A/B*, has been associated with SMA disease severity in some reports, but this association remains inconclusive [16]. The 15 PSVs that discriminate *SMN1* from *SMN2* have been found to occur in different combinations as part of *SMN1/2* hybrid genes and suggest the presence of extensively varying *SMN* haplotypes. However, the repetitive nature of the *SMN* locus represents a challenge for bioinformatic analysis from short-read sequencing data, complicating the identification of explanations for genotype–phenotype discordances in SMA [16].

Recently, advancements in long-read sequencing have led to the assembly of novel telomere-to-telomere, pangenome reference sequences [17, 18]. This fills many gaps in our knowledge of the human genome [18] and includes the assembly of several novel *SMN* locus alleles [17, 19]. However, these studies also highlight the complexity and diversity of genetic variation in this locus. After long-read sequencing, 66% of *SMN* alleles remained unresolved and the composition of resolved alleles varied extensively in gene copy number and orientation [17, 19]. A recent PacBio HiFi long-read sequencing study on 438 samples, of which 11 known SMA carrier samples and one known SMA patient sample, identified a stretch of single nucleotide variants (SNVs) downstream of *SMN1* and *SMN2* that may be used to further characterize these haplotypes, but limited sequencing read length hampered analysis of larger flanks [20]. Many fundamental questions, including copy-specific *SMN2* sequence variation, the existence and frequency of *SMN1* to *SMN2* gene conversion events, the location and genetic environment of *SMN2* copies, and differences in the composition of the *SMN* locus between healthy controls and patients with SMA, are yet to be answered. This shows that, despite using advanced techniques, the *SMN* locus remains one of the most complex in the human genome.

In addition to enhancing our understanding of complex loci in the human genome, several clinically relevant issues in SMA urgently need to be addressed. First, the incomplete correlation between SMA severity and *SMN2* copy number makes patient and family counseling challenging as disease severity cannot be precisely predicted [9, 11]. Second, three gene-targeting therapies for SMA have been approved, two of which are *SMN2* splice-modifying treatments. There is significant variability in patient responses to these drugs [21, 22]. The underlying causes of this variability are unknown, and individual treatment outcomes remain unpredictable. It is unclear how personal genetic variants influence treatment outcomes, despite the direct interaction of splice-modifying drugs with *SMN2* pre-mRNA. This underscores the need for an improved understanding of the composition of and genetic variation in the *SMN* locus, also from a clinical perspective [23].

To address these issues, a more detailed study of the *SMN* locus in patients with SMA, supported by technological advances, is required. Here, we first developed HapSMA, a method to perform polyploid phasing of ~ 173 kb of the *SMN* locus to enable copy-specific

analysis of *SMN* and its surrounding genes, and performed long-read, targeted Oxford Nanopore Technologies (ONT) sequencing of the *SMN* locus of patients with SMA. Using HapSMA on publicly available healthy control data, we identified genetic variants that characterized the typical genetic environment of *SMN1* and *SMN2*, that could be used as markers for gene conversion. Using these variants in SMA patients, we identified highly variable haplotypes with varying *SMN1/2* gene conversion breakpoints in 42% of *SMN2* haplotypes. In summary, we here provide direct evidence of gene conversion as a common genetic characteristic in SMA and illustrate that both methodological advances and the analysis of patient samples are required to advance our understanding of complex genetic loci.

Methods

Study population

Thirty-one SMA patients with a variety of *SMN* copy numbers and SMA types (Table 1) were included from our single-center prevalence cohort study in the Netherlands. The study protocol (09307/NL29692.041.09) was approved by the Medical Ethical Committee of the University Medical Center Utrecht and registered at the Dutch registry for clinical studies and trials [24]. Written informed consent was obtained from all adult patients, and from patients and/or parents additionally in the case of children younger than 18 years old. *SMN1*, *SMN2*, and *NAIP* copy numbers were determined using multiplex ligation-dependent probe amplification (MLPA) (MRC Holland, SALSA MLPA Probemix P021 SMA Version B1) according to the manufacturer's protocol. Clinical SMA type was determined as described previously [25]. Patients were classified as concordant or discordant (less or more severe) based on the deviation of clinical SMA type from MLPA-determined *SMN2* copy number; concordant being SMA type 1 with two copies of *SMN2*, SMA type 2 with three copies of *SMN2*, SMA type 3 with four copies of *SMN2*, and SMA type 4 with five copies of *SMN2* [9, 11]. To represent the wide genotypic and phenotypic variety in the SMA population, we included patients with *SMN2* copy numbers ranging from two to five, and SMA types ranging from 1b to 4. Whole blood was obtained in EDTA blood tubes for DNA extraction and 3 mm dermal biopsies were obtained for generation of primary fibroblasts. The control dataset used in this study contains 29 samples from the 1000 Genomes project, produced by the Human Pangenome Reference Consortium (HPRC) (Table 1) [17, 26]. Ethnic origins were Bangladesh (SAS, $n=3$), Barbados (AFR, $n=3$), China (EAS, $n=2$), Colombia (AMR, $n=1$), Gambia (AFR, $n=2$), Pakistan (SAS, $n=4$), Peru (AMR, $n=3$), Puerto Rico (AMR, $n=3$), Sri Lanka (SAS, $n=2$) and Vietnam (EAS, $n=6$).

Table 1 Clinical characteristics and *SMN* copy number of SMA and HPRC samples

	Total SMA	SMA type 1	SMA type 2	SMA type 3	SMA type 4	HPRC
Sex (F:M)	13:18	3:2	5:8	5:5	0:2	13:16
Median age at onset in years (range)	1.0 (0.0–24.5)	0.4 (0–0.5)	0.9 (0.5–1.5)	7 (1–15.5)	22.5 (20.5–24.5)	NA
Median age at sampling in years (range)	9.7 (0.4–70.1)	4.6 (0.4–41.5)	8.0 (2.1–57.5)	18.1 (5.2–63.3)	60.3 (50.4–70.1)	NA
***SMN2* copy number (MLPA)**						
0 × *SMN2*						
1 × *SMN2*						
2 × *SMN2*	3	2		1[a]		29
3 × *SMN2*	16	3	10[b]	3		
4 × *SMN2*	11		3	5	2	
5 × *SMN2*	1			1		
***SMN1* copy number (MLPA)**						
0 × *SMN1*	29	5	12	9	2	
1 × *SMN1*	2		1[b]	1[a]		
2 × *SMN1*						29
3 × *SMN1*						
4 × *SMN1*						

SMA patients were classified as concordant in case of 2 × *SMN2* and type 1, 3 × *SMN2* and type 2 or 3 × *SMN2* and type 3; discordant (more severe) in case of 3 × *SMN2* and type 1, 4 × *SMN2* and type 2 or 5 × *SMN2* and type 3; discordant (less severe) in case of 2 × *SMN2* and type 3, 3 × *SMN2* and type 3 or 4 × *SMN2* and type 4. HPRC, Human Pangenome Reference Consortium; F, female; M, male; NA, not available; MLPA, multiplex ligation-dependent probe amplification

[a] Including patient with *SMN1* with c.542A > G mutation

[b] Including patient with *SMN1* with deletion of exon 1–6

Culture of primary fibroblasts

Patient-derived fibroblasts were cultured in Dulbecco's modified Eagle's medium (DMEM, Gibco, 41966–029) containing 4.5 g/L D-glucose, L-glutamine and pyruvate, supplemented with 10% heat-inactivated fetal bovine serum (Cytvia, SH30073.03) and penicillin–streptomycin (Sigma-Aldrich, P0781), in a humidified 5% CO_2 atmosphere, at 37 °C. At a confluency of 70–80%, fibroblasts were enzymatically detached with Accutase (Sigma Aldrich, A6964) and frozen at − 80 °C as pellets for RNA or protein extraction, or resuspended in DMEM containing 20% FBS and 6% dimethyl sulfoxide (DMSO) for high molecular weight (HMW) DNA extraction.

RNA and protein quantification using droplet digital PCR (ddPCR) and western blot

Previously generated RNA and protein expression data were available [27]. In short, *SMN1* full-length (*SMN1-FL*), *SMN2* full-length (*SMN2-FL*), *SMN2* lacking exon 7 (*SMN2Δ7*), and *TBP* mRNA levels were quantified by ddPCR in technical triplicates. Expression levels of *SMN1-FL*, *SMN2-FL*, and *SMN2Δ7* were normalized against *TBP* expression using QuantaSoft Software (Bio-Rad, 1864011). For SMN protein quantification, semi-quantitative western blotting was performed in technical triplicates. SMN protein level, determined with mouse-anti-SMN primary antibody (BD Bioscience, 610647; 1:1000) and donkey-anti-mouse secondary antibody IRDye 800 (Licor, 926–32212; 1:2500), was normalized against Revert 700 total protein stain (Licor, 926–11011). Results were normalized to an internal standard (SMN level from HEK293 cell lysates) to enable reliable comparison of quantifications obtained from different membranes [27].

HMW DNA extraction

For SMA patients, HMW DNA was extracted from fresh or frozen whole EDTA blood ($n=8$), cultured primary dermal fibroblasts ($n=20$), or both tissues ($n=3$) using the Monarch® HMW DNA Extraction Kit for Cells & Blood (New England Biolabs (NEB), T3050L) with lysis agitation at 1400 rpm. For three samples, additional ultra-high molecular weight (UHMW) DNA was extracted from fibroblasts with an adapted protocol as recommended by the manufacturer (including changes to lysis steps (lysis agitation speed at 700 rpm; performed optional RNase A step), elution steps (elution buffer-Triton Mix (EB+) instead of standard elution buffer) and dilution and resuspension steps [28]) according to Ultra-Long Sequencing Kit (Oxford Nanopore Technologies (ONT), SQK-ULK001) guidelines. DNA concentration was measured using the Qubit™ dsDNA Quantification Broad Range Assay (Invitrogen, Q32853) and purity was determined on a spectrophotometer (Nanodrop 2000, Thermo Scientific).

Segmental duplication analysis and masking

We used the recently published T2T-CHM13 genome [18] as the reference genome for our analyses, as it has been shown to be more suitable for analyzing complex regions than previous reference genomes [29]. Segmental duplication analysis of T2T-CHM13 chromosome 5 against itself was performed with the nucmer command within MUMmer 3.23 [30]. Nucmer output was filtered for segments with a sequence identity score of ≥ 95%, length of ≥ 10 kb, and coordinates between 70 and 72 Mb, visualized in R v4.4.0 [31] and converted into a bed file for visualization in Intergrative Genomics Viewer (IGV). Each segment was assigned a unique identifier from 1a to 59a, with corresponding duplicated counterparts labeled as 1b to 59b. Based on this analysis, segments 20a and 22a (chr5:70,772,138–70,944,284, containing *SMN2*) were masked with bedtools v2.25.0 maskfasta [32], to direct mapping of *SMN1* and *SMN2* sequencing reads to segments 20b and 22b. We defined this as our phasing ROI (chr5:71,274,893–71,447,410, containing *SMN1*). For GRCh38, we used previously published segmental duplication analyses available in the UCSC Genome Browser [33, 34]. We masked coordinates chr5:69,924,952–70,129,737 and defined chr5:70,800,001–71,005,160 as the phasing ROI.

Alignment of *NAIP* and *pseudoNAIP* genes

NAIP and *pseudoNAIP (NAIPP)* sequences with 10 kb flanks were extracted from the T2T-CHM13 reference genome: *NAIP* (NCBI RefSeq NM_004536.3), CAT/Liftoff *NAIPP1-201*, CAT/Liftoff *NAIPP2-201*, NCBI RefSeq *NAIPP3* and NCBI RefSeq *NAIPP4*. All *NAIPP* genes were aligned with the full-length *NAIP* gene with the megablast setting in NCBI Nucleotide BLAST [35].

Sequencing and rebasecalling

Per sequencing run, three library preps with 1.3 µg HMW DNA each were made with the ligation sequencing kit (ONT, SQK-LSK109). For the UHMW samples, the Ultra-Long Sequencing Kit (ONT, SQK-ULK001) was used. The library preps were sequenced on a FLO-MIN106 flow cell on a GridION (ONT) with MinKNOW v21.02.5–22.12.5 (ONT) and FAST basecalling for 72 h, with a nuclease flush and reloading a new library prep every 24 h. Adaptive sampling [36, 37] was used within MinKNOW, with a combined target FASTA file (see Additional file 1) of the 30 Mb region surrounding the *SMN* locus (GRCh38 chr5:55,000,000–85,000,000) and six resolved alleles of the *SMN* locus downloaded from [19, 38]. For SMA ONT data, rebasecalling of the raw sequencing data (FAST5

files) was performed using the SUP model in Guppy v6.1.2 [39], including mapping to GRCh38. Long-range PCR and Illumina sequencing were performed on 29 SMA samples as previously published [3].

Data download and processing of the HPRC dataset

The Human PanGenomics Project was accessed in November 2024 from [17, 26] to download unmapped ONT BAM files (to a whole-genome read depth ~ 30–60×) and PacBio HiFi BAM files (to a whole-genome read depth ~ 20–40×) of samples from the 1000 Genomes project, produced by the Human Pangenome Reference Consortium (HPRC). Only samples with two copies *SMN1* and two copies *SMN2* of which both ONT and PacBio HiFi data (at least 20× read depth, required for Paraphase [20]) was available were included in our study; Samples with discordant copy numbers between Paraphase and previous reports [20] were excluded (Additional file 2: Fig. S1), resulting in a final dataset of 29 samples (Table 1). Samples and file locations are listed in Additional file 3: Table S1. Unmapped BAM files were converted to FASTQ with samtools fastq v1.17 [40] and mapped to GRCh38 with minimap v2.26 [41, 42]. Reads mapped to chromosome 5 were extracted with samtools view v1.17.

Directed mapping, polyploid haplotype phasing, and variant calling

We summarized our bioinformatics pipeline in Additional file 2: Fig. S2. Mapped BAM files from the SMA patients [43] and HPRC samples were processed using the HapSMA workflow v1.0.0 [44] with option bam_remap (SMA data, using multiple BAM output files from Guppy) or bam_single_remap (HPRC data, using a single BAM file). In short, sequencing reads were re-mapped to the masked T2T-CHM13 or masked GRCh38 reference genome using minimap2 v2.26. T2T-CHM13 was primarily used for obtaining the results of this study; GRCh38 was only used for comparison with the Paraphase analysis of PacBio HiFi data and multiple sequence alignment (MSA). Ploidy-aware unguided variant calling was performed on the phasing ROI using GATK v4.2.1.0 [45] based on *SMN1/2* total copy number determined by MLPA. Only the SNVs (not INDELs) from the resulting VCFs were used for haplotype phasing with Whatshap v1.7 [46] (standard "region" phasing approach). The BAM file was then split for each haplotag (HP) of the phaseset (PS) covering *SMN1* using sambamba v1.0.0 [47], resulting in one BAM file per *SMN1/2* haplotype. BAM files were visualized in IGV v2.17.4 [48] with quick consensus mode and "hide small indels" (below 50 bp) options. Soft-clipped read segments were shown, unless mentioned otherwise. For each haplotype separately, variant calling was performed using Clair3 v1.0.4 [49], and structural variant calling was performed using Sniffles2 v2.2 [50].

For some samples with read depth (DP) < 3 in the *SMN* PSV region in one or more haplotypes, haplotype phasing was improved by manual curation: first, guided ploidy-aware variant calling and phasing was performed on all SMA samples on previously published SNV positions [20] using a bed file containing these positions. Next, the phasing region was expanded by selection of SNVs upstream and downstream of known SNVs on positions where all haplotypes had a minimum read depth of 4× and contained the variant in either 0% or 100% of the reads, allowing two errors in total and excluding variants located on homopolymer stretches of 5 bp or more. The selected SNV positions were added to the bed file of known SNV positions and used for the next iterations. After two iterations, the resulting bed file (see the "Availability of data and materials" section) was used for phasing ("bed" phasing approach) in seven SMA patients and eight HPRC samples (Additional file 4: Table S2). By default, HapSMA produces results with both standard "region" phasing and "bed" phasing (with the manually curated bed file). Selection between these phasing methods was performed after running HapSMA by assessing read depth per PSV position with samtools depth v1.17 (with script 4.1_select_bed_or_region_phasing.sh, Additional file 2: Fig. S2).

Variant filtering and visualization

The scripts for analyses subsequent to HapSMA are available at [51]. Clair3 VCFs were parsed into tab-separated files that were merged into one file for all samples. Read depth was called for all haplotypes on all variant positions with samtools depth v1.17 and added as an extra column to the variant files (4.2_vcf_parse_merge_depth.sh in Additional file 2: Fig. S2). The resulting file was loaded into R v4.4.0 for further analysis. Variants with read depth (DP) ≥ 3 and allele frequency (AF) ≥ 0.5 were categorized as alternative allele (ALT), DP ≥ 3 and AF < 0.5 (or not called by Clair3) as reference allele (REF), and DP < 3 as not available (NA). All ALT positions were summarized and split per haplotype. At this stage, haplotypes are defined as *SMN1* of *SMN2* based on PSV13 (c.840C > T) (4.3_SNV_analysis.R in Additional file 2: Fig. S2, results in Additional file 5: Table S3). Next, FASTA files were made of each haplotype, using Additional file 5: Table S3 and a minimum read depth of 3× on every position on the phasing ROI called by samtools depth v1.17. The resulting FASTA files were concatenated per *SMN* copy type, resulting in one FASTA file with all *SMN1* haplotypes and one FASTA file with all *SMN2* haplotypes (4.5_create_fasta_roi.sh in Additional file 2: Fig. S2). These were mapped to the masked

T2T-CHM13 reference genome using hapdiff v0.8 [52] for visualization in IGV v2.17.4 without quick consensus mode and with the "hide small indels" option disabled. Soft-clipped read segments were not shown. In HPRC samples, variant positions were classified as *SMN2*-specific when the alternative allele was present in ≥ 90% of resolved *SMN2* haplotypes and ≤ 10% of resolved *SMN1* haplotypes, based on at least 20 haplotypes each. Since PSV5 at position chr5:71,407,128 is a hybrid PSV in the T2T-CHM13 reference genome (cytosine instead of thymine, normally present in *SMN2*), it was classified as *SMN2*-specific when the alternative allele was present in ≤ 10% of resolved *SMN2* haplotypes and ≥ 90% of resolved *SMN1* haplotypes. At all *SMN2*-specific positions, REF alleles were classified as *SMN1* PSV or *SMN1* environment SNV, and ALT alleles were classified as *SMN2* PSV or *SMN2* environment SNV (vice versa for PSV5 and *SMN1*-specific positions; 4.6_determine_and_show_SMN_specific_positions.R and 4.7_load_bed_and_show_SMN_specific_positions.R in Additional file 2: Fig. S2). If an *SMN2* haplotype contained an *SMN1*-specific variant only at position chr5:71,407,825, previously reported as PSV8 [3], it was not considered a hybrid gene, because the genotypes at this position were highly variable in our patient population and was not consistently associated with other PSVs or *SMN1* environment markers in SMA haplotypes, in agreement with a recent study by Costa-Roger et al. [4]. To determine the downstream *SMN1* environment SNV ratio, the number of haplotypes with downstream *SMN1* environment SNVs was divided by the total number of haplotypes in a sample; a mixed environment was counted as 0.5. If not all haplotypes were resolved in this region due to low read depth, the ratio was estimated with the allele frequency on the *SMN2*-specific positions in the full sample (ALT frequency of ~ 0.67 results in *SMN1* environment ratio of 0.33). The presence of *NAIP* or truncated *NAIPP* gene in every haplotype was determined by visual inspection of read clipping as shown in Additional file 2: Fig. S3, with a minimum requirement of three supporting reads.

Paraphase

PacBio HiFi BAM files mapped to GRCh38 were analyzed with paraphase v3.1.1 [20] with -g smn1 –gene-1only options. Paraphase VCF files were split with bcftools + split v1.18 [40], filtered with bcftools view -i 'GT = "1"' v1.18 and indexed with GATK IndexFeatureFile v4.2.1.0. Haplotype FASTA files were made with GATK FastaAlternateReferenceMaker v4.2.1.0.

Multiple sequence alignment

MSA was performed with mafft v7.407 [53] with default parameters using all resolved *SMN1* and *SMN2* haplotype FASTA files generated with HapSMA variants, resolved *SMN1* and *SMN2* haplotypes generated with Paraphase variants and haplotypes generated with common variants of the *SMN* haplogroups reported by Chen et al. [20]. Similarly, MSA was run on HapSMA *SMN1* and *SMN2* haplotypes of SMA samples and *SMN* haplogroup FASTA files. Visualization of MSA was done as tree data in iTOL v6 [54]. All FASTA files used for MSA were made with GRCh38 as reference genome on coordinates chr5:70,917,100–70,961,220, to be compatible with previously published haplogroup FASTA files [20].

Comparison of ONT HapSMA and Illumina variant calling

Illumina FASTQ data was trimmed for adapter sequences using trimmomatic v0.39 [55] with the ILLUMINACLIP option and mapped to the masked T2T-CHM13 reference genome using BWAMEM2 v2.2.1 [56]. *SMN* ploidy-aware variant calling was performed by GATK v4.2.1.0 on the long-range PCR coordinates (T2T-CHM13 chr5:71,378,876–71,410,446) [3]. Variants were filtered for minimum depth (DP ≥ 10) and minimum allele frequency (AF ≥ 0.15, in which AF = (ALT-AD/DP)). The cutoff 0.15 was selected as this is just below the minimal expected allele frequency for the sample with the highest number of known *SMN* copies (5 copies, thus 0.20). Overlap between variants called by Illumina and ONT was considered per sample. A variant was considered present in both Illumina and ONT data when the variant was present in the Illumina GATK VCF (with DP ≥ 10 and AF ≥ 0.15) and at least one of the haplotype-specific ONT Clair3 VCFs (with DP ≥ 3 and AF ≥ 0.5). When defining Illumina sequencing as the gold standard, the sensitivity and specificity of HapSMA SNV calling was calculated as follows:

Sensitivity = (Called both methods)/((Called both methods) + (Called Illumina only)) * 100.

Specificity = (Called both methods)/((Called both methods) + (Called ONT only)) * 100.

Comparison of ONT HapSMA and PacBio HiFi Paraphase variant calling

HapSMA was evaluated against Paraphase by comparing the overlap of SNVs identified by both pipelines on GRCh38 coordinates chr5:70,917,100–70,961,220. For each HPRC sample, the Clair3 VCF files generated by HapSMA and the Paraphase VCF files were filtered to retain only SNPs using bcftools filter v1.18 with option -i 'TYPE = = "snp"'. The filtered files were then compared using bcftools isec v1.18 with option -c "some" to classify SNVs as unique to HapSMA, unique to Paraphase, or shared between both methods. When defining Paraphase

as the gold standard, the sensitivity and specificity of HapSMA SNV calling was calculated as follows:

Sensitivity = (Called both methods)/((Called both methods) + (Called Paraphase only)) * 100.

Specificity = (Called both methods)/((Called both methods) + (Called ONT only)) * 100.

To evaluate HapSMA performance across different sequencing depths, HPRC BAM files mapped to GRCh38 were downsampled with samtools view v1.17 retaining 25% and 50% of reads, respectively, and analyzed with HapSMA as described above. Sensitivity and specificity analyses were repeated on downsampled HapSMA analyses.

Annotation of possibly functional variants

For annotation of possibly functional variants, Clair3 VCFs of all SMA haplotypes were merged with vcftools v0.1.16 [57] and converted to GRCh38 coordinates with the bcftools v1.18 liftover plugin. The Ensembl Variant Effect Predictor v111 [58] was used to assess the impact of genetic variants on canonical *SMN1* transcript ENST00000380707.9 with the following tools: SIFT, PolyPhen, AlphaMissense, CADD (PHRED $\geq$ 15 considered possibly pathogenic) and SpliceAI (delta score $\geq$ 0.5 considered possibly pathogenic). Only Clair3 variants with DP $\geq$ 3 and AF $\geq$ 0.5, which were also found with Illumina sequencing, were kept. Exonic and flanking intronic variants that were not predicted to have any effect, are also shown for completeness.

Statistics

To test the difference between the two groups of datapoints, an unpaired two-sided *T*-test was used. If the normality assumption, tested with the Shapiro–Wilk test, was violated, the Wilcoxon rank sum test was used. To test linear relationships, simple linear regression was performed. To test correlations, the Spearman correlation test was performed. To test whether hybrid *SMN2* genes have a downstream *SMN1* environment more often than non-hybrid *SMN2* genes, a Fisher's exact test was used. To test the difference in *SMN1* environment SNV ratio between disease severity groups, the Kruskal–Wallis rank sum test was used. In boxplots, the box represents the interquartile range, including the median. Statistics were performed in R v4.4.0.

Results

Copy-specific analysis of *SMN* and surrounding genes by HapSMA: targeted mapping and polyploid phasing of ONT sequencing reads

We first developed HapSMA, a method to enable sequence analysis of individual *SMN* copies and flanking regions. The *SMN* locus contains many segmental duplications, complicating mapping of both short-read and long-read sequencing data. To determine the extent and location of these duplications, we performed segmental duplication analysis on chromosome 5 of the telomere-to-telomere CHM13 (T2T-CHM13) reference genome (Fig. 1A, Additional file 2: Fig. S4A). This analysis identified a ~ 173 kb region, including *SMN2* and flanking sequences, of high similarity to *SMN1* and flanking sequences. This region (containing *SMN2*) was masked to prevent ambiguous mapping and to direct mapping of reads to *SMN1* and flanking region (~ 173 kb), from here on called the phasing region of interest (ROI) (Fig. 1B). This allowed us to directly compare all *SMN1*- and *SMN2*-derived sequencing reads and identify variants that could reliably distinguish both genes and their environments. We performed Oxford Nanopore Technologies (ONT) sequencing of high molecular weight (HMW) DNA of 31 SMA patients (Table 1) and used adaptive sampling to enrich for a 30 Mb region surrounding the *SMN* locus (see the "Methods" section). Whole-genome ONT and PacBio HiFi sequencing data and previously determined *SMN1/2* copy number data of 29 healthy controls with two copies of *SMN1* and two copies

(See figure on next page.)

Fig. 1 Copy-specific analysis of *SMN* and surrounding genes by targeted mapping and polyploid phasing. **A** Dot plot resulting from segmental duplication analysis of T2T-CHM13 chromosome 5 against itself (left panel), zoomed in on the *SMN* locus (right panel). Red lines indicate segments of at least 95% similarity and at least 10 kb. The structure of the *SMN* locus as shown in B is shown at scale on both the *x*- and *y*-axes. **B** Structure of the *SMN* locus on the T2T-CHM13 reference genome. Genes are indicated by colored arrows. Genome coordinates chr5:70,772,138–70,944,284 were masked, and its segmental duplication counterpart chr5:71,274,893–71,447,410 was the main region of interest (ROI) for haplotype phasing. **C** Overview of sequencing and bioinformatics approaches in this study. ONT sequencing with adaptive sampling was performed on the HMW DNA of SMA patients. Raw sequencing data was basecalled with the Guppy SUP model and mapped to GRCh38. Within the HapSMA workflow, reads were remapped to the masked T2T-CHM13 reference genome as indicated in **B**. Polyploid variant calling and haplotype phasing were performed, followed by variant calling per haplotype. Figure created with BioRender.com. **D** Example of haplotype phasing of an SMA sample with three *SMN2* copies across *SMN1/2* and surrounding genes. Sequencing reads are colored by haplotype. Soft-clipping is not shown. **E** Comparison of SNVs called by Paraphase and HapSMA. Panel a: SNVs called by Paraphase but not by HapSMA (in at least one sample). Panel b: SNVs called by both Paraphase and HapSMA (in at least one sample). Panel c: SNVs called by HapSMA but not by Paraphase (in at least one sample). With HapSMA, SNVs are called in a larger region (~ 500 kb) than with Paraphase (~ 44 kb). Panel d: Phasing ROI as shown in **B**. Panel e: gene annotation. HMW, high molecular weight; HPRC, Human Pangenome Reference Consortium; kb, kilobases; Mb, megabases; ROI, region of interest

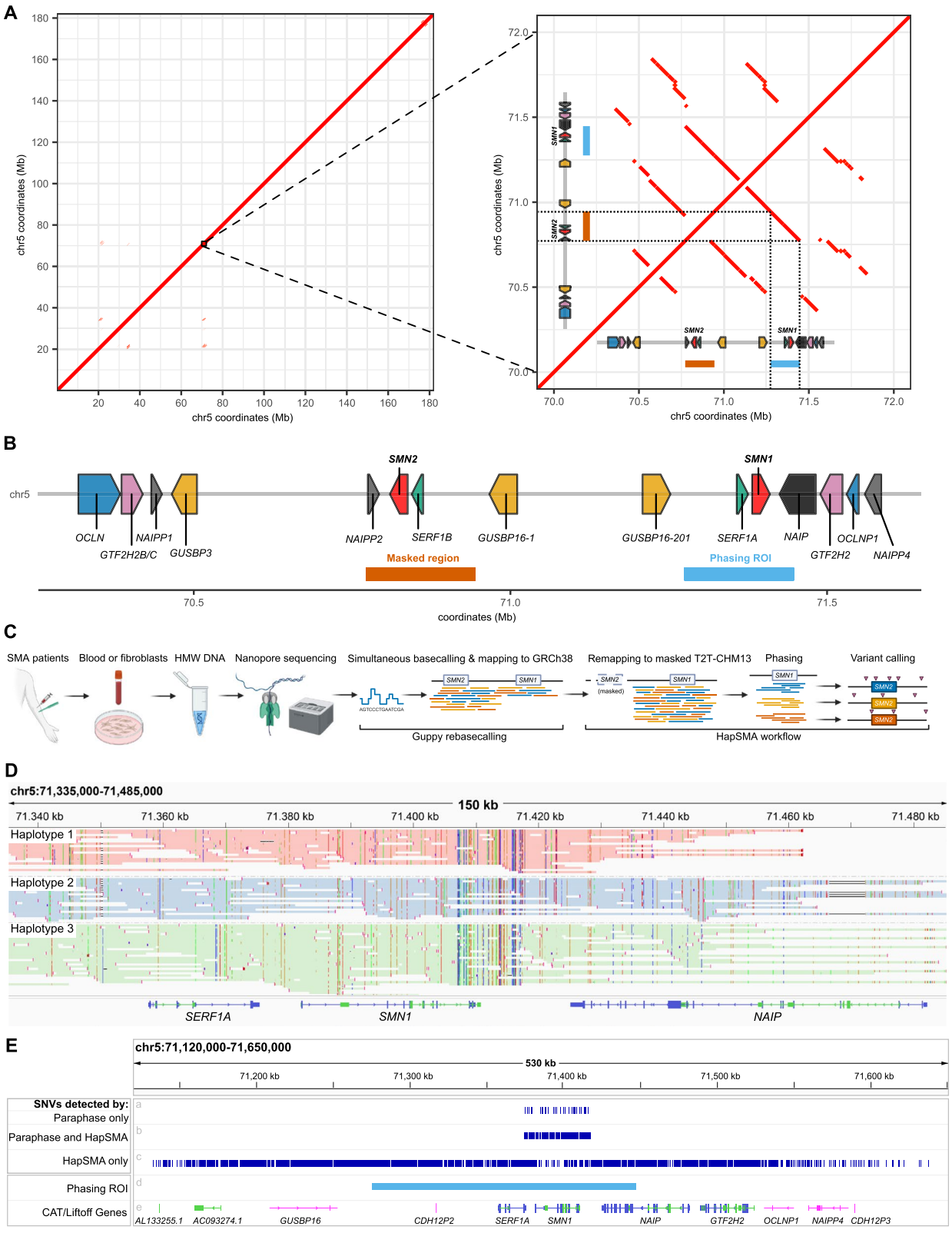

Fig. 1 (See legend on previous page.)

of *SMN2* were available from the Human Pangenome Reference Consortium (HPRC) [17, 20]. Applying our HapSMA bioinformatic workflow, sequencing reads were mapped to the masked T2T-CHM13 reference genome and phased into two to six haplotypes (Fig. 1C). To determine the total *SMN1/2* copy number for the HapSMA ploidy input parameter, use of orthogonal methods is required, such as MLPA. An example of haplotype phasing across *SMN* and surrounding genes for one sample is shown in Fig. 1D. Median read depth on the 30 Mb region surrounding the *SMN* locus was 42.6×for HPRC and 22.3×for SMA (Additional file 2: Fig. S4B). Read length N50 was different between HPRC (75.7 kb) and SMA (48.6 kb) samples (Additional file 2: Fig. S4C). Percentage of read depth on the phasing ROI that was successfully phased into haplotypes (i.e., not left unassigned to any haplotype) differed between HPRC (median = 88.3%) and SMA datasets (median = 74.9%; Additional file 2: Fig. S4D). Minimum read depth for calling variants was set at 3x. Median per-haplotype coverage of the phasing ROI with a minimum read depth of 3×was 100.0% for HPRC and 81.6% for SMA (Additional file 2: Fig. S4E), which was associated with total read depth and read length N50 per sample (Additional file 2: Fig. S4F–G). Across HPRC and SMA datasets, we identified a median of 679 variants per sample (range 162–1416) using HapSMA. In total, we identified 4586 unique variants, of which 3498 SNVs and 1088 insertions/deletions (INDELs). Although HapSMA does output INDEL calls, HapSMA does not use INDELs for phasing and it is not recommended to use HapSMA for reporting INDELs as the performance of ONT sequencing for INDEL calling is limited [59].

To assess the performance of HapSMA, we determined SNVs within the *SMN* gene using long-range PCR and Illumina sequencing in 29 of our SMA patients [3], and found a high level of concordance of HapSMA compared to Illumina (median sensitivity of 93.0% and median specificity of 92.9% in the SMA dataset, Additional file 2: Fig. S4H–I). In addition, we compared the SNV calling performance of HapSMA to Paraphase SNV calling on PacBio HiFi data (median read depth 34.1×(range 23.3–39.4×), read length N50 19.3 kb (range 16.8–23.9 kb)). When comparing HapSMA to Paraphase, the median sensitivity of HapSMA is 98.0% and specificity is 97.9% in the HPRC dataset. Through downsampling of the ONT data used for HapSMA, we found that sensitivity and specificity are stable above a read depth of 20×, which we recommend as minimum read depth (Additional file 2: Fig. S4J–K). Moreover, the main advantage of HapSMA is that it calls variants in a > 10×wider window (~ 500 kb) than Paraphase (~ 44 kb; Fig. 1E). This allows assessment of flanking genes and non-coding regions for each *SMN1/2* haplotype separately. Thus, haplotype phasing and copy-specific analysis of *SMN* genes including large flanks can be achieved with ONT sequencing and HapSMA.

Specific SNVs and *NAIP* variants mark *SMN1* and *SMN2* environments in healthy controls

Our current knowledge of *SMN1*- or *SMN2*-specific variants is mostly limited to the presence of 15 intragenic paralogous sequence variants (PSVs) that distinguish *SMN1* from *SMN2* [3–5]. To expand the range of variants capable of differentiating *SMN1* and *SMN2* and their broader genetic environments, we first used publicly available ONT sequencing data of 29 healthy control samples from the HPRC [17] (Table 1) and applied HapSMA to determine 116 copy-specific *SMN1* and *SMN2* haplotypes (Fig. 2A). Haplotypes were classified as *SMN1* or *SMN2* based on PSV13 (c.840C > T variant). When performing multiple sequence alignment (MSA), HapSMA *SMN1* and *SMN2* haplotypes form separate clusters. In addition, 96.9% of resolved HapSMA haplotypes cluster in the same clade as the Paraphase haplotypes of the same sample, suggesting a high similarity between haplotypes generated with both methods (Additional file 2: Fig. S5). Genetic variants were classified as *SMN2*-specific, when at least 90% of *SMN2* haplotypes and a maximum of 10% of *SMN1* haplotypes contained the variant. At these cutoffs, 13 out of 15 known PSVs were detected as *SMN2*-specific (Additional file 6: Table S4). PSV5 and PSV16 were not detected as *SMN2*-specific, likely because these PSVs are hybrid in a high number of haplotypes (Fig. 2B and Additional file 4: Table S2). In addition, we identified 26 *SMN2*-specific SNVs and one *SMN1*-specific SNV downstream of *SMN1/2* (Fig. 2A, Additional file 6: Table S4). The *SMN1*-specific SNV located at position chr5:71,411,916 (G) in T2T-CHM13 differs from GRCh38 position chr5:70,955,164 (A) when lifted over, and would therefore be considered *SMN2*-specific on GRCh38 (similar to PSV5, see the "Methods" section). When the alternative allele was present on an *SMN2*-specific variant position, it was reported as an *SMN2* environment SNV and when the reference allele was present, as an *SMN1* environment SNV; and vice versa for *SMN1*-specific variant positions. Twenty-one of these downstream SNVs have been reported as downstream *SMN2* region before (present in all "c" haplogroups) using PacBio sequencing and the bioinformatic tool Paraphase [20], confirming the validity of our results.

Next, we investigated the presence of the *(pseudo) NAIP* gene downstream of *SMN* in each haplotype. In most previously published haploid reference alleles (with one copy *SMN1* and one copy *SMN2*), *NAIP* is

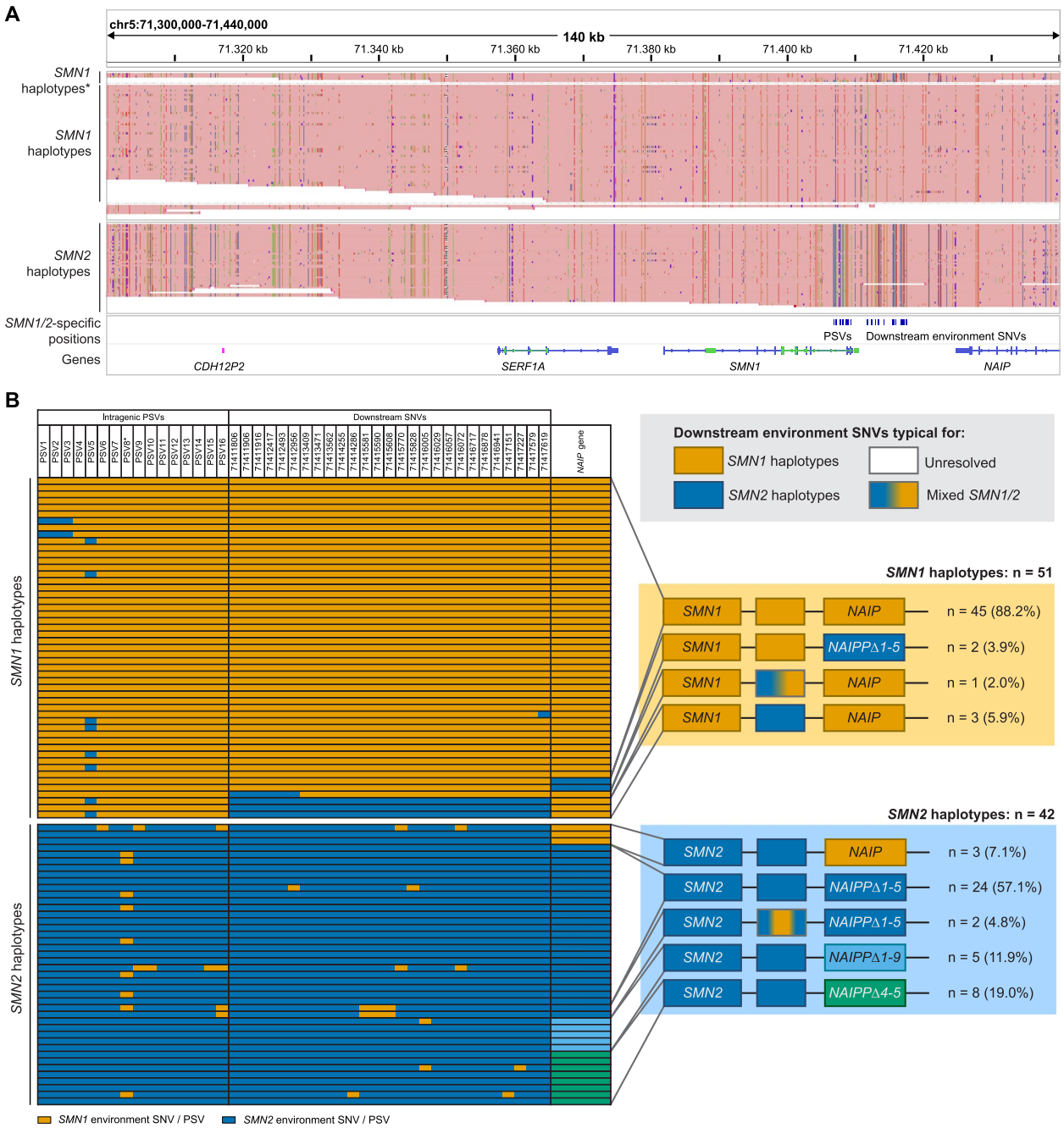

Fig. 2 In healthy controls, specific SNVs and *NAIP* variants characterize the downstream environment of *SMN1/2*. **A** IGV overview of *SMN1* and *SMN2* haplotypes (divided based on PSV13 (c.840C > T) in exon 7 (see the "Methods" section)) from HPRC healthy control samples, mapped to the masked T2T-CHM13 reference genome. Each "read" represents one haplotype from one sample. *SMN2*-specific variant positions (present in ≥ 90% of *SMN2* haplotypes and ≤ 10% of *SMN1* haplotypes) and *SMN1*-specific variant positions (present in ≥ 90% of *SMN1* haplotypes and ≤ 10% of *SMN2* haplotypes), are indicated by blue lines above the genes, including PSVs and downstream *SMN1/2* environment SNVs. **SMN1* haplotypes with downstream *SMN2* environment SNVs. **B** Schematic representation of PSVs, *SMN1/2* environment SNVs, and presence of the *(pseudo)NAIP* gene per haplotype. Only haplotypes with complete phasing between PSV13 (c.840C > T) and *(pseudo)NAIP* are shown. In the right panel, downstream haplotype frequencies are shown schematically. Downstream environment other than the "expected" environment was called when 3 or more consecutive *SMN1/2* environment SNVs were present. Full-length *NAIP* was characterized as *SMN1* environment, whereas truncated *NAIPPΔ1–5*, *NAIPPΔ1–9* or *NAIPPΔ4–5* was characterized as *SMN2* environment [20]. *PSV8 (5 bp insertion at position chr5:71,407,825) is currently not considered a PSV, but a common variant [4]. IGV, Integrative Genomics Viewer; *NAIPP*, *pseudoNAIP*; PSV, paralogous sequence variant; SNV, single nucleotide variant

adjacent to *SMN1*, and a truncated *pseudoNAIP* is adjacent to *SMN2* [20]. Therefore, we defined full-length *NAIP* as *SMN1* environment. In addition to full-length *NAIP*, we observed different types of truncated *pseudoNAIP* (*NAIPP*) genes downstream of *SMN1/2*: *NAIPPΔ4–5* lacking exon 4–5, *NAIPPΔ1–5* lacking exon 1–5, and *NAIPPΔ1–9* lacking exon 1–9 (Additional file 2: Fig. S3). We defined *NAIPPΔ1–5*, *NAIPPΔ1–9*, and *NAIPPΔ4–5* as *SMN2* environment. For 93 of the 116 HPRC haplotypes (80.2%), the *(pseudo)NAIP* gene was phased successfully. These haplotypes are shown in Fig. 2B; with a more detailed list in Additional file 4: Table S2. In 45 of 51 *SMN1* haplotypes (88.2%), only elements of a downstream *SMN1* environment (downstream *SMN1* environment SNVs and full-length *NAIP*) were present, whereas six haplotypes (11.8%) contained an element of downstream *SMN2* environment (downstream *SMN2* environment SNVs, *NAIPPΔ1–5*, *NAIPPΔ1–9*, *NAIPPΔ4–5* or a combination). Thirty-seven out of 42 *SMN2* haplotypes (88.1%) had a full downstream *SMN2* environment, and five (11.9%) a (partial) downstream *SMN1* environment (Fig. 2B). In summary, our results show typical genomic environments for *SMN1* and *SMN2* in healthy controls, based on the analysis of SNVs and *(pseudo)NAIP* variants.

Markers for gene conversion are abundant in SMA patients and the underlying recombination events can occur at different breakpoints downstream of *SMN2*

Next, we determined the presence of the identified markers of *SMN1* and *SMN2* environments in SMA patient haplotypes (Fig. 3A). When performing MSA, HapSMA *SMN1* and *SMN2* haplotypes from SMA samples form separate clusters. Haplogroups S2-1 and S2-2 are the most common in the SMA dataset (Additional file 2: Fig. S6). For 54 of 104 SMA haplotypes (51.9%), the *(pseudo)NAIP* gene was phased completely. We included two patients with pathogenic variants in *SMN1* and found that, as expected, both *SMN1* haplotypes had a downstream *SMN1* environment. Amongst *SMN2* haplotypes, different types of downstream *SMN1* environments indicating gene conversion were present: (1) all downstream *SMN1* environment SNVs and *NAIP*, (2) three or more consecutive downstream *SMN1* environment SNVs and *NAIP*, or (3) fewer than three consecutive downstream *SMN1* environment SNVs and *NAIP*. This indicates that the recombination events underlying gene conversion in SMA occur at different genomic locations (Fig. 3B). Out of 39 non-hybrid *SMN2* haplotypes, 24 haplotypes (61.5%) had a full downstream *SMN2* environment and 15 haplotypes (38.5%) had a full or partial downstream *SMN1* environment.

Previously, hybrid genes have been noted as markers of gene conversion events [60], suggesting they are more likely to have a downstream *SMN1* environment. Here, we found 10 different *SMN* hybrid gene structures across 21 haplotypes, five of which were novel (Fig. 3B, two lower left panels). From 13 completely resolved hybrid *SMN2* haplotypes, six (46.2%) contained a downstream *SMN2* environment and seven (53.8%) contained a downstream *SMN1* environment. The percentage of downstream *SMN1* environment in hybrid haplotypes (53.8%) was not significantly higher than in non-hybrid *SMN2* haplotypes (38.5%; Fisher's exact test, $p = 0.353$). We next hypothesized that the downstream environment of *SMN2* may affect its activity through regulatory elements such as enhancers. Exploratory analyses comparing *SMN* gene environments to SMA type, and *SMN2-Δ7/FL* mRNA and SMN protein expression illustrate substantial variability that warrants further study in a larger cohort (Additional file 2: Fig. S7). In summary,

(See figure on next page.)

Fig. 3 Markers of *the SMN1* environment are abundant and highly variable in *SMN2* haplotypes of SMA patients. **A** IGV overview of *SMN1* and *SMN2* haplotypes (divided based on PSV13 (c.840C > T) in exon 7 (see the "Methods" section)) from SMA patients, mapped to the masked T2T-CHM13 reference genome. Each "read" represents one haplotype from one sample. *SMN2*-specific variant positions (present in ≥ 90% of *SMN2* haplotypes and ≤ 10% of *SMN1* haplotypes) and *SMN1*-specific variant positions (present in ≥ 90% of *SMN1* haplotypes and ≤ 10% of *SMN2* haplotypes) as determined in Fig. 2A, are indicated by blue lines above the genes, including PSVs and downstream *SMN1/2* environment SNVs. **SMN2* haplotypes with downstream *SMN1* environment SNVs. ***SMN2* haplotypes with an incompletely resolved downstream *SMN1/2* environment. **B** Schematic representation of PSVs, *SMN1/2* environment SNVs, and presence of the *(pseudo)NAIP* gene per haplotype. Of non-hybrid *SMN2* haplotypes, only haplotypes with complete phasing between PSV13 (c.840C > T) and *(pseudo)NAIP* are shown. In the two lower left panels, haplotypes with hybrid *SMN2* genes are shown, of which five hybrid structures are novel: PSV1–4, 6–9, and 16 (a); PSV1–7 and 14–16 (b); PSV1–11 (c); PSV1 (d); PSV1–9 (e). In the right panel, downstream haplotype frequencies are shown schematically. Downstream environment other than the "expected" environment was called when 3 or more consecutive *SMN1/2* environment SNVs were present. Full-length *NAIP* was characterized as *SMN1* environment, whereas truncated *NAIPPΔ1–5*, *NAIPPΔ1–9*, or *NAIPPΔ4–5* was characterized as *SMN2* environment [20]. The percentage of downstream *SMN1* environment in hybrid haplotypes (53.8%) was not significantly higher than in non-hybrid *SMN2* haplotypes (38.5%; Fisher's exact test, $p = 0.353$). *PSV8 (5 bp insertion at position chr5:71,407,825) is currently not considered a PSV, but a common variant [4]. IGV, Intergrative Genomics Viewer; *NAIPP*, *pseudoNAIP*; PSV, paralogous sequence variant; SNV, single nucleotide variant

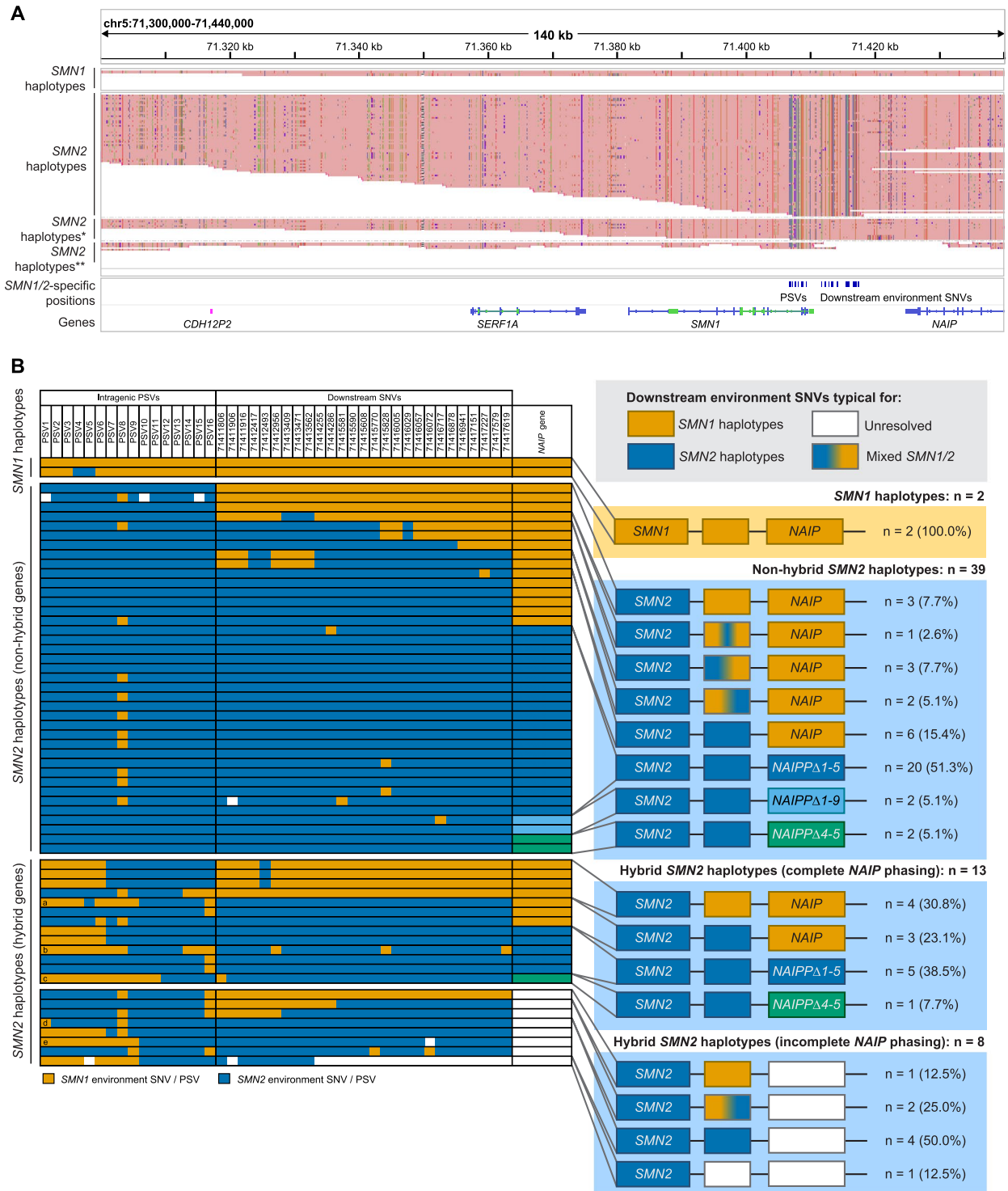

Fig. 3 (See legend on previous page.)

we provide evidence for gene conversion in 42% (22 out of 52) of phased *SMN2* haplotypes and show that the underlying recombination events can occur at different breakpoints downstream of *SMN2*.

Copy-specific sequence and structural variants are common in patients with SMA

Because HapSMA allows for copy-specific analysis, we next assessed the presence of copy-specific exonic and

structural variants. We detected several types of exonic variants in SMA patients: three missense, two synonymous, and two untranslated region (UTR) variants (Table 2). The pathogenic missense variant c.542A > G (p.Asp181Gly), which was reported previously [13], was confirmed to be present in exon 4 of *SMN1* in one patient. It is predicted to create a new splice-donor site within exon 4 of *SMN1* leading to a truncated transcript by introducing a preliminary stop codon [13]. The mild phenotype of this patient (SMA type 3a) with respect to their *SMN2* copy number (two) suggests residual activity of this *SMN1* copy. Indeed, a small amount of *SMN1-FL* mRNA could still be detected (Additional file 2: Fig. S8A). Other known positive modifying variants, such as c.859G > C and c.835-44A > G [14, 15], were not detected in this patient. Two missense variants in *SMN2*, c.77G > A present in two patients and c.593C > T in one patient, were predicted to be pathogenic and thus candidate negative modifiers (Supplemental Note 1). We found no clear link between the abovementioned exonic variants and *SMN* mRNA or SMN protein level in patients' derived fibroblasts with and without the variant (Additional file 2: Fig. S8B–D).

In addition to small sequence variants, we also identified several structural variants in *SMN* and its surrounding genes. First, in one SMA patient, we detected four copies of *SMN* exon 7–8—one of which derived from *SMN1*—and three copies of *SMN* exon 1–6, indicating either an exon 1–6 deletion or *SMN1* exon 7–8 insertion. Either situation indicates an incomplete *SMN1* copy containing only exon 7 and 8, which corresponded with the absence of detectable *SMN1* mRNA in this patient (Additional file 2: Fig. S8A, Additional file 2: Fig. S9). Further structural variants were identified in *(pseudo) NAIP* (Additional file 2: Fig. S10), and *SERF1A/B* (Additional file 2: Fig. S11). Unfortunately, an incomplete understanding of the functionality of *SERF1* and *NAIP* (pseudo)genes prevents us from further speculating on the functional relevance of these structural variants. In summary, HapSMA allows for the detection of sequence and structural variants in a haplotype-aware manner, which provides useful information about the localization of different variants on the same or different copies of *SMN* and surrounding genes, improving opportunities to interpret the possible functional effect of *SMN* variants.

Discussion

A complete understanding of genetic variation in the complex *SMN* locus requires both methodological advances and its study in the genomes of SMA patients. Therefore, we performed copy-specific haplotype phasing of the ~ 173 kb genomic environment surrounding *SMN* gene copies in a large cohort of SMA patients using targeted long-read sequencing. We identified downstream *SMN1* environment SNVs serving as markers of gene conversion, and the downstream presence of varying *NAIP* (pseudo) genes as markers of genomic location. These markers provide direct evidence of *SMN1* to *SMN2* gene conversion as a common genetic characteristic of SMA. We found that broad phasing including the *NAIP* gene allowed for a more complete view of *SMN* locus variation, increasing the number of identified downstream environment types from six to nine compared to when only downstream SNVs would have been considered [20]. Moreover, both the number of different haplotypes and the genetic variation in *SMN2* haplotypes were larger in patients with SMA, highlighting the importance of the inclusion of SMA patients when investigating the *SMN* locus.

HapSMA analysis of ONT data showed a high sensitivity and specificity for SNV detection (both ~ 98%) when compared to Paraphase analysis of PacBio HiFi long-read sequencing data of HPRC samples, and HapSMA haplotypes showed high similarity to Paraphase haplotypes of the same sample. Through downsampling, we found a recommended read depth of at least 20× for running HapSMA. The sensitivity and specificity of HapSMA when compared to long-range PCR and Illumina short-read sequencing in SMA samples was lower (~ 93%), but this might be due to lower read depth in the SMA dataset (one sample < 15×, nine samples 15–20×) and/or lower read length N50 than in the HPRC dataset. For finding *SMN2*-specific variants, we used strict cutoffs (present in ≥ 90% of *SMN2* haplotypes and ≤ 10% of *SMN1* haplotypes; vice versa for *SMN1*-specific variants) to generate robust results. With these cutoffs, PSV5 and PSV16 were not detected as *SMN2*-specific, likely because they are variable in the currently used HPRC dataset. These cutoffs can be easily adapted in the analysis steps downstream of HapSMA. For discovery purposes, less stringent thresholds (e.g., 80%/20%) can be applied to find additional (upstream and downstream) *SMN2*-specific variants. In summary, for investigating the sequence of the *SMN* genes with PacBio HiFi data, Paraphase is recommended; for investigation of *SMN* haplotypes including ~ 500 kb surrounding genes and non-coding regions, HapSMA is recommended.

By examining *SMN1* environment SNVs in *SMN2* haplotypes, we determined that the recombination events underlying gene conversion can happen at different breakpoints between *SMN* and *NAIP*. Occasionally, more than one recombination breakpoint is present within one haplotype. The high variability of *SMN2* haplotypes in SMA is a confirmation of the high variability in the *SMN* locus between individuals, as reported before [19]. Previously, *SMN2* hybrid genes have been used as markers for gene conversion [60], mostly based on PSV16 in exon 8. In this study, we show that downstream markers of *SMN1* environment are also present in resolved haplotypes

Table 2 Exonic variants in *SMN1* and *SMN2* identified by ONT sequencing

Gene with variant	HGVSc	HGVSp	Variant consequence	SIFT	PolyPhen	Alpha Missense	CADD PHRED	Number of haplotypes	Number of patients	Patient phenotypes	Reported before in SMA
SMN1[a]	c.542 A > G	p.Asp181Gly	Missense	Deleterious	Benign	Likely benign	22.2	1/90	1	Less severe	[13]
SMN1[a]	c.835 − 96A > G	-	Intron	-	-	-	14.87	1/96	1	Less severe	This study
SMN2	c.−14 C > T	-	5′ UTR	-	-	-	4.697	1/82	1	Concordant	[13]
SMN2	c.77 G > A	p.Gly26Asp	Missense	Deleterious	Possibly damaging	Likely pathogenic	24.5	2/82	2	More severe, concordant	This study
SMN2	c.81 + 45C > T	-	Intron	-	-	-	5.642	1/81	1	Concordant	[13]
SMN2	c.84 C > T	p.Ser28Ser	Splice region, synonymous	-	-	-	11.69	3/88	3	More severe, concordant, less severe	[13, 71]
SMN1[a]/*SMN2*	c.462 A > G	p.Gln154Gln	Synonymous	-	-	-	8.282	55/90	24	-	[13, 71]
SMN2	c.593 C > T	p.Pro198Leu	Missense	Deleterious	Probably damaging	Likely benign	19.94	2/90	1	More severe	This study
SMN2	c.723 + 59G > C	-	Intron	-	-	-	5.842	5/92	5	Less severe (3), concordant (1), more severe (1)	[13]
SMN2	c.724 − 54C > T	-	Intron	-	-	-	7.545	10/95	8	Less severe (2), concordant (6)	[13]
SMN2	c.*58 G > A	-	3′ UTR	-	-	-	10.33	1/97	1	Concordant	This study

Annotation and variant calls were based on *SMN1* transcript ENST00000380707.9. Genomic locations of variants are listed in Additional file 7: Table S5

[a] *SMN1* with a known pathogenic variant

[*] notation indicates that this genetic variant is present 58 nucleotides downstream of the stop codon

containing a hybrid *SMN2* gene, but not significantly more often than in non-hybrid haplotypes. In addition, we also identified hybrid gene structures in patients without the PSV16 variant, indicating that the absence of this variant cannot be used to exclude the presence of gene hybrids on its own. Twenty-one of 27 downstream *SMN2* environment SNVs identified in our study were previously also identified in all *SMN1* "c" haplotypes by PacBio HiFi sequencing [20]. Interestingly, similar to our findings, *SMN2* haplotypes with a full downstream *SMN1* environment were not detected in non-SMA individuals [20].

We showed that different *(pseudo)NAIP* genes can be present downstream of *SMN1/2*. A previous study showed that on "normal" alleles with one copy of *SMN1* and one copy of *SMN2*, the vast majority of *SMN1* copies were upstream of full-length *NAIP* and *SMN2* copies were upstream of truncated *pseudoNAIP* [20]. In HPRC control data, we similarly observed that *SMN1* was most often followed by *NAIP*, whereas *SMN2* was most often followed by truncated *NAIPPΔ1–5*, *NAIPPΔ1–9* or *NAIPPΔ1–5*. No *SMN1/2*-environment-specific variants were found in the *SERF1A/B* gene, suggesting that *SERF1A* and *SERF1B* are identical. The consistent co-phasing of *SERF1A/B* with *SMN* and the absence of clear structural variation breakpoints confirm that most genomic rearrangements causing deletions associated with SMA include the *SERF1A/B* gene [20, 61].

Our newly identified markers of gene conversion and genomic location provide a foundation for further studies into their possible prognostic potential in larger cohorts. One hypothesis is that variation in a genetic environment of *SMN* may affect RNA and protein levels by influencing local regulatory elements such as enhancers or DNA methylation, which is supported by exploratory analyses of SMN protein and mRNA analyses included in our study. However, data on chromatin accessibility (H3K27ac ChIP-Seq and ATAC-seq), 3D DNA interactions (Hi-C), and DNA methylation are currently limited within the *SMN* locus, as this type of data is largely based on short-read sequencing and thus, does not map well on the *SMN* locus [62–64] or has not yet been studied in significant detail [65]. However, long-read methods for chromatin accessibility and interactions are emerging and could prove useful in identifying regulatory regions within the *SMN* locus [66–68]. Similarly, methods to obtain information on base modifications such as methylation from long-read sequencing reads are increasingly sensitive and reliable [69, 70], opening possibilities to obtain this information concurrently with sequence and structural variation. These developments may also be useful in estimating the potential functional effects of intergenic structural variants such as the Alu insertions and deletions identified upstream of *SERF1A/B*. Furthermore, HapSMA enables the identification of the specific *SMN* copy on which a pathogenic variant is present, which supports the prediction of remaining SMN functionality in case of multiple pathogenic variants.

Conclusions

While our understanding of the complexity of the *SMN* locus has advanced considerably, it remains one of the most elusive regions in the human genome. This study, along with advancements in technology and the increasing cost-effectiveness of long-read sequencing, holds the promise of obtaining a more nuanced understanding of this locus and its relationship to SMA. Our findings reveal a heightened complexity in the *SMN* locus of SMA patients when compared to healthy controls, underscoring the necessity of including SMA patients in analyses of *SMN* locus composition and variation. The broad heterogeneity we observed suggests that extensive de novo assembly of both control and SMA patient samples will be essential to achieve a comprehensive understanding of the *SMN* locus. This will not only be vital for enhancing our basic understanding of complex genetic loci but also for addressing critical clinical challenges in SMA.

Abbreviations

GTF2H2	General transcription factor IIH subunit 2
HMW	High molecular weight
HPRC	Human Pangenome Reference Consortium
IQR	Interquartile range
Mb	Megabases
MLPA	Multiplex Ligation-dependent Probe Amplification
MSA	Multiple sequence alignment
NAIP	Neuronal apoptosis inhibitory protein
ONT	Oxford Nanopore Technologies
PSV	Paralogous sequence variant
SERF1A	Small EDRK-rich factor 1A
SMA	Spinal muscular atrophy
SMN	Survival motor neuron
SNV	Single nucleotide variant
T2T	Telomere-to-telomere

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-025-01448-2.

Additional file 1. Target FASTA file for adaptive sampling.

Additional file 2. Includes Supplemental note 1, Figures S1-S11 and Supplemental references.

Additional file 3. Includes Table S1, an overview of paths to files downloaded for analysis of the 29 HPRC samples in this study.

Additional file 4. Includes Table S2, an overview of *SMN1/2*-specific variants per haplotype determined on T2T-CHM13.

Additional file 5. Includes Table S3, an overview of all identified variants per haplotype determined on T2T-CHM13.

Additional file 6. Includes Table S4, an overview of PSVs and *SMN1/SMN2*-specific variant positions, including frequencies and ratios of variants in *SMN1* and *SMN2* haplotypes determined on T2T-CHM13.

Additional file 7. Includes Table S5, an overview of T2T-CHM13 and GRCh38 coordinates of exonic (and flanking) variants reported in Table 2.

Acknowledgements
We thank the patients and their families for their participation in this study. We thank the Human Pangenome Reference Consortium for generating and releasing the ONT and PacBio HiFi WGS data (https://humanpangenome.org/); we would like to acknowledge the National Genome Research Institute (NHGRI) for funding the following grants which are in support of creating the human pangenome reference: 1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973, 1U01HG010963. We would also like to thank the 1000 Genomes Project ONT Sequencing Consortium (1KGP-ONT) for generating and releasing ONT WGS data, from which 25 samples were included in the initial preprint of this manuscript. We thank B.P.C. Koeleman for critically reading this manuscript and providing extensive feedback.

Authors' contributions
Conceptualization, MMZ, MGE, DG, EJNG, GWvH; experimental studies, MMZ, IS, LBP, MCR, IRJG, JK; bioinformatics, MMZ, MGE, DG, JvdS, CV, RS; clinical data & patient material, FA, RIW, WLP; writing–original draft, MMZ, EJNG, GWvH; writing–review, all; resources, EFT, WLvdP, EJNG; supervision, WLvdP, EJNG, GWvH; funding acquisition, EJNG, GWvH, WLvdP, EFT. All authors read and approved the final manuscript.

Funding
This work was supported by grants from Stichting Spieren voor Spieren (to WLvdP), the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie grant (H2020 Marie Skłodowska-Curie Actions) agreement no. 956185 (SMABEYOND ITN to WLvdP, EJNG, EFT) and Prinses Beatrix Spierfonds (W.OB21-01 to EJNG). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n° 772376—EScORIAL). This work was partially funded by grants to EFT from Biogen (ESP-SMG-17–11256), Roche, GaliciAME, and the Spanish Instituto de Salud Carlos III, Fondo de Investigaciones Sanitarias and co-funded with ERDF funds (grant no. FIS PI18/000687). We acknowledge the Utrecht Sequencing Facility (USEQ) for providing sequencing service and data. USEQ is subsidized by the University Medical Center Utrecht and The Netherlands X-omics Initiative (NWO project 184.034.019).

Data availability
The paths to HPRC datasets used in this study are available in Additional file 3: Table S1. The adaptive sampling target file, variant calls at *SMN1/2*-specific positions and an overview of all Clair3 variant calls per haplotype are available in the supplemental files. The code for HapSMA is available at: https://github.com/UMCUGenetics/HapSMA [44] (v1.0.0 was used for analyses in this study, v1.1.0 contains extra support for different types of data input). The code for analyses subsequent to HapSMA and input files used in these analyses are available at: https://github.com/UMCUGenetics/ManuscriptSMNGeneConversion [51]. Long-read nanopore sequencing data generated for SMA patients included in this study, including the clinical metadata, cannot be made publicly available due to legal restrictions and patient confidentiality. However, academic request to reuse the data will be granted if the research question aligns with the original informed consent and IRB approval. Detailed metadata, variable descriptions, and the procedure to obtain access to these via a data sharing agreement are available via DataverseNL (https://doi.org/10.34894/G7YG0V [43]). Requests can be made by contacting the corresponding authors.

Declarations

Ethics approval and consent to participate
The study protocol (09307/NL29692.041.09) was approved by the Medical Ethical Committee of the University Medical Center Utrecht and registered at the Dutch registry for clinical studies and trials [24]. Written informed consent was obtained from all adult patients, and from patients and/or parents additionally in the case of children younger than 18 years old.

Consent for publication
Not applicable.

Competing interests
JHV reports to have sponsored research agreements with Biogen and Astra Zeneca. The remaining authors declare that they do not have any competing interests.

Author details
[1]Department of Neurology and Neurosurgery, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, the Netherlands. [2]Department of Genetics, University Medical Center Utrecht, Utrecht, the Netherlands. [3]Medicine Genetics Group, Vall d'Hebron Research Institute (VHIR), Barcelona, Spain. [4]Department of Clinical and Molecular Genetics, Hospital Vall d'Hebron, Barcelona, Spain. [5]Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands. [6]Utrecht Sequencing Facility, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands. [7]Oncode Institute, Utrecht, the Netherlands.

Received: 18 September 2024 Accepted: 7 March 2025
Published online: 21 March 2025

References

1. Schmutz J, Martin J, Terry A, Couronne O, Grimwood J, Lowry S, et al. The DNA sequence and comparative analysis of human chromosome 5. Nature. 2004;431:268–74.
2. Lefebvre S, Bürglen L, Reboullet S, Clermont O, Burlet P, Viollet L, et al. Identification and characterization of a spinal muscular atrophy-determining gene. Cell. 1995;80:155–65.
3. Blasco-Pérez L, Paramonov I, Leno J, Bernal S, Alias L, Fuentes-Prior P, et al. Beyond copy number: A new, rapid, and versatile method for sequencing the entire *SMN2* gene in SMA patients. Hum Mutat. 2021;42:787–95.
4. Costa-Roger M, Blasco-Pérez L, Gerin L, Codina-Solà M, Leno-Colorado J, Gómez-García De la Banda M, et al. Complex *SMN* Hybrids Detected in a Cohort of 31 Patients With Spinal Muscular Atrophy. Neurol Genet. 2024;10:e200175.
5. Monani UR, Lorson CL, Parsons DW, Prior TW, Androphy EJ, Burghes AHM, et al. A Single Nucleotide Difference That Alters Splicing Patterns Distinguishes the SMA Gene *SMN1* From the Copy Gene *SMN2*. Hum Mol Genet. 1999;8:1177–83.
6. Lorson CL, Hahnen E, Androphy EJ, Wirth B. A single nucleotide in the *SMN* gene regulates splicing and is responsible for spinal muscular atrophy. Proc Natl Acad Sci. 1999;96:6307–11.
7. Lefebvre S, Burlet P, Liu Q, Bertrandy S, Clermont O, Munnich A, et al. Correlation between severity and SMN protein level in spinal muscular atrophy. Nat Genet. 1997;16:265–9.
8. Verhaart IEC, Robertson A, Wilson IJ, Aartsma-Rus A, Cameron S, Jones CC, et al. Prevalence, incidence and carrier frequency of 5q–linked spinal muscular atrophy – a literature review. Orphanet J Rare Dis. 2017;12:124.
9. Wadman RI, Stam M, Gijzen M, Lemmink HH, Snoeck IN, Wijngaarde CA, et al. Association of motor milestones, *SMN2* copy and outcome in spinal muscular atrophy types 0–4. J Neurol Neurosurg Psychiatry. 2017;88:365–7.
10. Burghes AHM. When Is a Deletion Not a Deletion? When It Is Converted. Am J Hum Genet. 1997;61:9–15.
11. Calucho M, Bernal S, Alías L, March F, Venceslá A, Rodríguez-Álvarez FJ, et al. Correlation between SMA type and *SMN2* copy number revisited: An analysis of 625 unrelated Spanish patients and a compilation of 2834 reported cases. Neuromuscul Disord. 2018;28:208–15.
12. Wirth B. Spinal Muscular Atrophy: In the Challenge Lies a Solution. Trends Neurosci. 2021;44:306–22.
13. 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.
14. Prior TW, Krainer AR, Hua Y, Swoboda KJ, Snyder PC, Bridgeman SJ, et al. A Positive Modifier of Spinal Muscular Atrophy in the *SMN2* Gene. Am J Hum Genet. 2009;85:408–13.
15. Wu X, Wang S-H, Sun J, Krainer AR, Hua Y, Prior TW. A-44G transition in *SMN2* intron 6 protects patients with spinal muscular atrophy. Hum Mol Genet. 2017;26:2768–80.

16. Costa-Roger M, Blasco-Pérez L, Cuscó I, Tizzano EF. The Importance of digging into the genetics of *SMN* genes in the therapeutic scenario of spinal muscular atrophy. Int J Mol Sci. 2021;22:9029.
17. Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617:312–24.
18. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.
19. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. Science. 2022;376: eabj6965.
20. 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.
21. Mercuri E, Sumner CJ, Muntoni F, Darras BT, Finkel RS. Spinal muscular atrophy. Nat Rev Dis Primer. 2022;8:1–16.
22. Scheijmans FEV, Cuppen I, van Eijk RPA, Wijngaarde CA, Schoenmakers MAGC, van der Woude DR, et al. Population-based assessment of nusinersen efficacy in children with spinal muscular atrophy: a 3-year follow-up study. Brain Commun. 2022;4:fcac269.
23. Cuscó I, Bernal S, Blasco-Pérez L, Calucho M, Alias L, Fuentes-Prior P, et al. Practical guidelines to manage discordant situations of *SMN2* copy number in patients with spinal muscular atrophy. Neurol Genet. 2020;6:e530.
24. Centrale Commissie Mensgebonden Onderzoek. Ministerie van Volksgezondheid, Welzijn en Sport; 2013. Available from: https://www.ccmo.nl/.
25. Wijngaarde CA, Stam M, Otto LAM, van Eijk RPA, Cuppen I, Veldhoen ES, et al. Population-based analysis of survival in spinal muscular atrophy. Neurology. 2020;94:e1634–44.
26. Human PanGenomics Project - Registry of open data on AWS. Available from: https://registry.opendata.aws/hpgp-data/.
27. Signoria I, Zwartkruis MM, Geerlofs L, Perenthaler E, Faller KME, James R, et al. Patient-specific responses to *SMN2* splice-modifying treatments in spinal muscular atrophy fibroblasts. Mol Ther - Methods Clin Dev. 2024;32:101379.
28. Protocol Guidance for Extraction of Ultra-High Molecular Weight (UHMW) Genomic DNA for Ultra-Long (UL) Read NGS Sequencing applications in Oxford Nanopore Technologies® workflows. 2022. Available from: https://www.neb.com/en/protocols/2022/02/17/protocol-guidance-for-extraction-of-high-molecular-weight-uhmw-genomic-dna-for-ultra-long-ul-read-ngs-sequencing-applications-in-oxford-nanopore-technologies-workflows.
29. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, et al. A complete reference genome improves analysis of human genetic variation. Science. 2022;376:eabl3533.
30. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5: R12.
31. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/.
32. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
33. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. Genome Res. 2001;11:1005–17.
34. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent Segmental Duplications in the Human Genome. Science. 2002;297:1003–7.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
36. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat Methods. 2016;13:751–4.
37. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021;39:442–50.
38. Vollger MR. Data files for: segmental duplications and their variation in a complete human genome. 2021; Available from: https://zenodo.org/records/5501736.
39. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019;20:129.
40. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10: giab008.
41. Li H. New strategies to improve minimap2 alignment accuracy. Alkan C, editor. Bioinformatics. 2021;37:4572–4.
42. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34:3094–100.
43. Groen E. Metadata for: 'Long-read sequencing identifies copy-specific markers of *SMN* gene conversion in spinal muscular atrophy. DataverseNL; 2025. Available from: https://dataverse.nl/dataset.xhtml?persistentId=10.34894/G7YG0V.
44. Elferink MG, Ernst RF, Zwartkruis MM, Gommers D. HapSMA. GitHub; 2025. Available from: https://github.com/UMCUGenetics/HapSMA.
45. Van der Auwera GA, O'Connor BD. Genomics in the cloud: using docker, GATK, and WDL in Terra. 1st Edition. O'Reilly Media; 2020. Available from: https://www.oreilly.com/library/view/genomics-in-the/9781491975183/.
46. Martin M, Ebert P, Marschall T. Read-Based Phasing and Analysis of Phased Variants with WhatsHap. Methods Mol Biol Clifton NJ. 2023;2590:127–38.
47. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.
48. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.
49. Zheng Z, Li S, Su J, Leung AW-S, Lam T-W, Luo R. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. Nat Comput Sci. 2022;2:797–803.
50. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol. 2024;42:1571–80.
51. Zwartkruis MM, Gommers D, Elferink MG. ManuscriptSMNGeneConversion. GitHub; 2025. Available from: https://github.com/UMCUGenetics/ManuscriptSMNGeneConversion.
52. Kolmogorov M. Hapdiff. Kolmogorov lab at NCI; 2025. Available from: https://github.com/KolmogorovLab/hapdiff.
53. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80.
54. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. 2024;52:W78-82.
55. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
56. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009;25:1754–60.
57. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
58. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.
59. Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, et al. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. Genome Res. 2023;33:2029–40.
60. Butchbach MER. Genomic Variability in the *Survival Motor Neuron* Genes (*SMN1* and *SMN2*): Implications for Spinal Muscular Atrophy Phenotype and Therapeutics Development. Int J Mol Sci. 2021;22: 7896.
61. Scharf JM, Endrizzi MG, Wetter A, Huang S, Thompson TG, Zerres K, et al. Identification of a candidate modifying gene for spinal muscular atrophy by comparative genomics. Nat Genet. 1998;20:83–6.
62. Gao T, He B, Liu S, Zhu H, Tan K, Qian J. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics. 2016;32:3543–51.
63. Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Res. 2019;47:D106–12.
64. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. 2018;19:151.
65. Hauke J, Riessland M, Lunke S, Eyüpoglu IY, Blümcke I, El-Osta A, et al. *Survival motor neuron gene 2* silencing by DNA methylation correlates with spinal muscular atrophy disease severity and can be bypassed by histone deacetylase inhibition. Hum Mol Genet. 2009;18:304–17.
66. Akbari V, Leelakumari S, Jones SJM. Profiling chromatin accessibility in humans using adenine methylation and long-read sequencing. bioRxiv; 2023. p. 2023.10.05.561129. Available from: https://www.biorxiv.org/content/10.1101/2023.10.05.561129v1. Cited 2024 May 29.

67. Hu Y, Jiang Z, Chen K, Zhou Z, Zhou X, Wang Y, et al. scNanoATAC-seq: a long-read single-cell ATAC sequencing method to detect chromatin accessibility and genetic variants simultaneously within an individual cell. Cell Res. 2023;33:83–6.
68. Xie Y, Ruan F, Li Y, Luo M, Zhang C, Chen Z, et al. Spatial chromatin accessibility sequencing resolves high-order spatial interactions of epigenomic markers. Shi X, Weigel D, editors. eLife. 2024;12:RP87868.
69. Ahsan MU, Gouru A, Chan J, Zhou W, Wang K. A signal processing and deep learning framework for methylation detection using Oxford Nanopore sequencing. Nat Commun. 2024;15:1448.
70. Yuen ZW-S, Srivastava A, Daniel R, McNevin D, Jack C, Eyras E. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat Commun. 2021;12:3438.
71. Ruhno C, McGovern VL, Avenarius MR, Snyder PJ, Prior TW, Nery FC, et al. 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. 2019;138:241–56.

Publisher's Note