Supplementary Information

Supplementary methods

Samples and data processing

65 samples validated using digital PCR were procured from the Motor Neuron Diseases Research Laboratory (Nemours Alfred I. duPont Hospital for Children) collection and were generated from cell lines as described previously[1,2]. This cohort contained 37 SMA samples (14 type I SMA, 1 type I/II SMA, 14 type II SMA, 7 type III SMA and 1 SMA with unknown clinical grade), 10 non-SMA neuromuscular disease samples (including hereditary sensory and autonomic neuropathy 3, myotonic dystrophy type I, distal hereditary motor neuronopathy type I and Charcot-Marie-Tooth peripheral neuropathy type IA) and 18 non-SMA samples not known to have other diseases. 8 additional Coriell samples were also analyzed with digital PCR. 45 historical patient samples with known carrier or affected status for SMA (24 SMA, 20 carriers and 1 silent carrier, previously tested with MLPA) were obtained from Cambridge University. GS was performed with TruSeq DNA PCR-free sample preparation with 150bp paired reads sequenced on Illumina HiSeq X instruments. Genome build GRCh37 was used for read alignment.

For population studies, 13,343 individuals were queried from the NIHR BioResource Rare Diseases project (EGAS00001001012)[3], which performed GS on individuals with rare diseases and their close relatives. Additional individuals (n = 840) from the Next Generation Children (NGC) project (EGAD00001004357)[4], which performs diagnostic trio GS on patients and their parents from neonatal and pediatric intensive care units in the UK, were also investigated. GS in these studies was performed using the Illumina TruSeq DNA PCR-Free Sample Preparation kit with 100bp or 125bp paired reads sequenced on Illumina HiSeq 2500, or with 150bp paired reads sequenced on Illumina HiSeq X instrument, as previously described[3]. Genome build GRCh37 was used for read alignment. When doing our population analysis, we excluded related individuals and those of unknown ancestry, leaving 10,243 unrelated individuals.

For population studies, we also used the 1000 Genomes Project (1kGP) data, for which GS BAMs were downloaded from https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31736/. These BAMs were generated by sequencing 2x150bp reads on Illumina NovaSeq 6000 instruments from PCR-free libraries and aligning them to the human reference, hs38DH.

All of the samples used in this study were sequenced to an average depth of at least 30x.

Copy number calling for intact and truncated *SMN*

Our method calls the copy number of intact *SMN1* + *SMN2* (referred to as *SMN* hereafter) and truncated *SMN* (*SMN2*∆*7-8*) genes using the following steps.

- **Identify and count reads from *SMN1* and *SMN2*:** Read counts are calculated directly from the GS aligned BAM file using all reads mapped to either *SMN1* or *SMN2*, including those with a mapping quality of zero. Frequently reads align to these regions with a mapping quality of zero because the sequence is nearly identical between the two regions. Importantly, these two genes only share sequence with each other and not with other regions of the genome. Read counts in a 22.2kb region encompassing Exon 1 to Exon 6 are used to calculate the total *SMN* (*SMN1*, *SMN2* and *SMN2*∆*7-8*) CN, and read counts in the 6.3kb region including Exon 7 and Exon 8 are used to calculate the CN of intact *SMN* (*SMN1* and *SMN2*).

- **Calculate normalized depth of the *SMN* regions:** The read counts for the two regions described above are each normalized by region length and further normalized by dividing against the median depth of 3000 pre-selected 2kb regions across the genome. These 3000 normalization regions are randomly selected from the genome for stable coverage across population samples to infer the sequencing depth.

- **Convert normalized depth into copy numbers:** The normalized depth values are modeled with a one-dimensional mixture of 11 Gaussians with constrained means that center around each integer copy number value representing copy number states ranging from 0 to 10. Copy numbers of total *SMN* and intact *SMN* are called from the Gaussian mixture model (GMM) with a posterior probability threshold of 0.95.

- **Calculate the CN of the intact and truncated *SMN***: The intact *SMN* CN is defined as the CN of the 6.3kb region covering exons 7 and 8. The copy number of truncated *SMN* (*SMN2*∆*7-8*) is derived by subtracting the intact *SMN* CN from total *SMN* CN calculated from the 22.2kb region that includes exons 1-6.

Calling *SMN1* and *SMN2* CN

After calculating the total number of copies of *SMN* genes, we differentiated *SMN1* from *SMN2* using an algorithm described below. Since c.840C>T is the most important functional difference between *SMN1* and *SMN2*, the absolute copy number of these two genes can theoretically be derived using the ratio between the number of reads supporting *SMN1* and *SMN2* at this site. However, the read depth at one diploid position is typically 30-40X for a GS dataset and sometimes does not provide sufficient power to clearly differentiate between different CN states (see Figure S1). Therefore, we utilized additional base differences near c.840C>T so that information at these sites can be combined with c.840C>T when making a CN call. Because we wished to differentiate intact *SMN1* from *SMN2*, we only considered the variants that occur within the 6.3kb deletion. Excluding bases in homopolymers and short tandem repeats (STRs)

that may be more prone to errors, resulted in 16 base differences between *SMN1* and *SMN2* (Table S1).

For these 16 positions, we tested whether they were truly fixed in the population by comparing the CN call of the *SMN1* alleles for these positions with the CN call for the splice variant base *SMN1* c.840C. We identified eight positions, including c.840C>T, where the *SMN1* bases are fixed or close to being fixed in the population based on concordance with the splice variant base (see Results, Figure 3A).

To make a final CN call we required that either: 1) the *SMN1* CN calls agree across at least 5 out of 8 sites at a posterior probability cutoff of 0.8, or 2) at least 5 out of 8 sites (posterior probability > 0.6) agree with the CN call derived from all reads overlapping any of the 8 sites (posterior probability > 0.9). Otherwise a no-call is produced for both the *SMN1* and *SMN2* CNs. The thresholds were chosen based on our analysis of the concordance among sites in the population (See "Concordance between CN calls at the 8 selected base differences between *SMN1* and *SMN2*" section). SMA samples are identified as having zero copy of intact *SMN1* and carrier samples are identified as having one copy of intact *SMN1*.

At higher CN values, greater variability in read depth is expected, leading to less confident CN calls (with lower posterior probability) at individual sites and more disagreement between sites. As a result, no-calls are more likely to be made in samples with high *SMN1*/*SMN2* CNs, i.e. both values larger than or equal to two (see Figure S1). However, in such samples we can still confidently determine whether the *SMN1* CN is or is not 0 (SMA) or 1 (carrier), allowing us to call SMA/not SMA or carrier/not carrier. When the *SMN1* CN is a no-call, if at least seven of the *SMN1* CN calls are confidently greater than zero then the sample is called "not SMA". Similarly, if at least seven of the *SMN1* CNs are confidently greater than one, the sample is called "not carrier".

Additionally, when the intact *SMN* CN is a no-call, i.e. the sum of *SMN1*+*SMN2* is unknown, we also directly test for the absence of the c.840C allele that will be indicative of SMA. This is done by testing whether the number of reads supporting c.840C (the *SMN1* base) is more likely to derive from zero or one copy of *SMN1*. The likelihood is calculated based on a Poisson distribution with an expected value equal to the sequencing error (zero copy of *SMN1*) or the median haploid depth (one copy of *SMN1*).

Simulation for single site CN calling

We simulate the numbers of reads at one single site at a sample median depth of 30X, 35X and 40X based on a Poisson distribution, and sample *SMN1* supporting reads based on a binomial model with all possible combinations of *SMN1* CN and *SMN2* CN when the total *SMN* CN is between 2 and 6. With the number of supporting reads for *SMN1* and *SMN2*, we derive the posterior probability of our simulated *SMN1* CN (See Methods). The posterior probability is high (>0.9) when at least one value of *SMN1* or *SMN2* CN is low (smaller or equal to 1) (Figure S1).

When both values are larger than 2, i.e. in *SMN1:SMN2* combinations of 2:2, 2:3, 2:4, 3:2, 3:3, and 4:2, the posterior probability frequently becomes low and falls below 0.9. This results from the higher variability in read depth when the expected CN is higher. Therefore, in these scenarios it is less accurate to make *SMN1* and *SMN2* CN calls using one single site.

Identification of hybrid alleles between *SMN1* and *SMN2*

To estimate the frequencies of *SMN1* and *SMN2* alleles at the 16 base difference sites between *SMN1* and *SMN2*, we restricted ourselves to two simple CN states that allow easy identification of hybrid alleles (*SMN1*=CN2 and *SMN2*=CN0 or *SMN1*=CN2 and *SMN2*=CN1; Table S9, Figure S5). Specifically, in samples with either of the two CN states (*SMN1*=CN2 and *SMN2*=CN0 or *SMN1*=CN2 and *SMN2*=CN1), at an *SMN1/2* base difference site, if the *SMN1* CN is smaller than 2, this indicates the presence of *SMN2* alleles in the *SMN1* gene. In samples with *SMN1*=CN2 and *SMN2*=CN1, at an *SMN1/2* base difference site, if the *SMN1* CN is bigger than 2, this indicates the presence of *SMN1* alleles in the *SMN2* gene. Based on this analysis, we estimated that across the eight selected positions, up to 0.5% of the *SMN1* genes contain an *SMN2* allele. Conversely, up to 0.9% of the *SMN2* genes carry an *SMN1* allele. This may be the result of gene conversion or it could be that some sites are polymorphic in the population. A large portion of these hybrid alleles come from African populations (Table S9).

Concordance between CN calls at the 8 selected base differences between *SMN1* and *SMN2*

Compared with using c.840C alone, introduction of more base differences improved the ability to differentiate *SMN1* from *SMN2*. But because these sites are not truly invariant in the respective genes and CN calling at single sites can be subject to error, the likelihood that one of the individual calls will deviate from the true copy number state is increased. To make a final call, we required that the *SMN1* CN calls agreed with each other at 5 or more of 8 sites (see Supplementary Methods for a full description of the rules for CN calling).

With a posterior probability cutoff of 0.8, the majority of samples had consistent calls at at least 5 out of the 8 sites and only 1.4% of samples had fewer than 5 sites that agreed (Table S10). In 80% of those samples, a confident CN call was made based on the second consensus rule (requiring agreement with the CN call made by summing up reads at all 8 sites). The “non-agreeing” sites were more frequently no-calls due to a low posterior probability rather than discrepant calls, and only 15.3% of them were confident calls that disagreed with the consensus of the other sites. Again, a large portion of the disagreements come from African populations (Table S10).

Using fewer sites for the majority rule produced a larger number of no-calls and wrong calls compared with using eight sites (Table S11).

Discrepancies in validation samples

There was one sample, MB509, that was discrepant between our CN call and the digital PCR results. Upon further inspection, we found that this sample has two copies of *SMN2* and one copy of *SMN1* with a 1884bp deletion (chr5:70247145-70249029, hg19, Figure S6). While we cannot always trust read alignments in the *SMN1/2* region, careful analysis of the split reads shows that the reads or their mates overlap bases that are specific to *SMN1*. Thus we hypothesize that this deletion is correctly placed on *SMN1*. The deletion is small (does not change the depth significantly in the 6.3kb region used for determining the intact *SMN* CN) and has not been previously reported (nor found in the 1kGP samples, thus a very rare variant), so the caller was not designed to detect it. As a result, our caller called the total copy number of *SMN1*+*SMN2* as 3. The deletion is consistent with the CN calls we made in the 8 *SMN1*-*SMN2* difference sites, where the first 2 sites are not in the deletion and called at *SMN1* CN=1 and the next 6 sites are in the deletion and called at *SMN1* CN=0 (Figure S7A). Based on the majority rule, we called the *SMN1* copy number as 0, correctly identifying the sample as SMA. The *SMN2* copy number is calculated as the total copy number minus the *SMN1* copy number, so we called the *SMN2* copy number to be 3, overestimating it by 1.

Four other samples, MB231, MB367, MB383 and LP2101748, have discrepancies between our CN calls and results from either digital PCR or MLPA. Read counts and normalized depth values (read counts divided by haploid sample depth) at the 8 base difference sites support our CN calls (Figure S7A) and we speculate that the discrepancy is likely to be caused by errors in the orthogonal methods. In two samples, GS calls and digital PCR calls differ by a factor of two (MB231: GS-0,2, PCR-0,4 and MB383: GS-3,1, PCR-6,2). It is possible that there is a normalization issue with digital PCR, leading to an overestimation of copy number by two fold.

When comparing our CN calls with MLPA results in 1109 1kGP samples, we excluded one sample where we made a no-call for *SMN2*Δ*7-8* due to low posterior probability of the total *SMN* CN, as well as three samples where we made a no-call for *SMN1* and *SMN2* CN due to disagreements in the CN calls across the 8 selected sites that fail to meet our consensus rules (Figure S7B).

Detection of silent carriers

The c.*3+80T>G SNP is most strongly associated with two-copy *SMN1* alleles in Africans, where 84.5% of individuals with three copies of *SMN1* and 92.6% of individuals with four copies of *SMN1* have the c.*3+80T>G SNP (chi-squared test CN3 Africans vs. CN2 Africans, p-value <2.2e-16, Table 2). Calling the c.*3+80T>G SNP greatly increases the carrier detection rate in Africans as Africans have a higher frequency of alleles carrying two copies of *SMN1* (Table S8). However, 33% of individuals with two copies of *SMN1* also have the c.*3+80T>G SNP, suggesting that a significant portion of singleton *SMN1* alleles also carry this SNP in Africans. We calculated maximum likelihood estimates for the percentages of singleton and two-copy *SMN1* alleles that carry c.*3+80T>G (Table S7) and residual risks for the combination of CN and

SNP calling (Table S8). Our estimates are similar to previous studies[5–7], though there is considerable variability across all of these estimates. This variability is likely driven by population variability, e.g. Africans (this study) vs. African Americans (previous studies), and Northern Europeans (overrepresented in this study) vs. more diversely sampled Caucasians (previous studies).

Comparison between two aligners, BWA and Isaac

Our method analyzes reads permissively in both *SMN1* and *SMN2*, and thus is insensitive to how the aligner differentiates between the two genes. Therefore, using different aligners should produce similar results. The BAM data analyzed in this paper were generated using two different aligners: BWA for the 1kGP data and various versions of Isaac for the rest. The consistent *SMN1/2* CN distributions between 1kGP and NIHR (Table S5, Figure S4) samples suggests that our method is insensitive to the aligner. Additionally, we tested our method for consistency by aligning 117 samples with both BWA and Isaac, including 5 SMA samples and 3 carriers. All 117 samples produced the exact same calls (*SMN1/SMN2/SMN2Δ7-8* CN) with our method and the normalized depths for both Exon1-6 and Exon7-8 were virtually identical (Pearson's r>0.999, Figure S8).

Comparison between carrier calls by this study and Larson et al.

We compared the carrier calls made in the overlapping 1kGP samples in this study (N=37) to those reported by Larson et al.[8] (N=36), and found 26 overlapping calls (Table S12). MLPA results are available for 19 samples that are called as carriers by either method (16 called by this study and 14 by Larson et al., with 11 overlapping calls). MLPA calls agree with our calls in all of the 19 samples, indicating that Larson et al. made 3 false positives (FP) and 5 false negatives (FN) calls. Larson et al. identified carriers by determining whether the fraction of *SMN1* supporting reads was smaller than or equal to 1/3. That study used low depth sequencing data which would be expected to result in some errors but, more importantly, their approach is prone to error without calling the total copy number. For example, a sample with one copy of *SMN1* and one copy of *SMN2* will be called as a non-carrier (*SMN1* fraction 1/2), and a sample with two copies of *SMN1* and four copies of *SMN2* will be called as a carrier (*SMN1* fraction 1/3), resulting in false positive and false negatives (Table S12).

References

1. Stabley DL, Harris AW, Holbrook J, et al. *SMN1* and *SMN2* copy numbers in cell lines derived from patients with spinal muscular atrophy as measured by array digital PCR. *Mol Genet Genomic Med*. 2015;3(4):248-257. doi:10.1002/mgg3.141
2. Stabley DL, Holbrook J, Harris AW, et al. Establishing a reference dataset for the authentication of spinal muscular atrophy cell lines using STR profiling and digital PCR. *Neuromuscul Disord NMD*. 2017;27(5):439-446. doi:10.1016/j.nmd.2017.02.002
3. BioResource TN, The 100 OBO, Project 000 Genomes. Whole-genome sequencing of rare

disease patients in a national healthcare system. *bioRxiv*. January 2019:507244. doi:10.1101/507244
4. French CE, Delon I, Dolling H, et al. Whole genome sequencing reveals that genetic conditions are frequent in intensively ill children. *Intensive Care Med*. 2019;45(5):627-636. doi:10.1007/s00134-019-05552-x
5. Luo M, Liu L, Peter I, et al. An Ashkenazi Jewish *SMN1* haplotype specific to duplication alleles improves pan-ethnic carrier screening for spinal muscular atrophy. *Genet Med Off J Am Coll Med Genet*. 2014;16(2):149-156. doi:10.1038/gim.2013.84
6. Feng Y, Ge X, Meng L, et al. The next generation of population-based spinal muscular atrophy carrier screening: comprehensive pan-ethnic *SMN1* copy-number and sequence variant analysis by massively parallel sequencing. *Genet Med Off J Am Coll Med Genet*. 2017;19(8):936-944. doi:10.1038/gim.2016.215
7. Alías L, Bernal S, Calucho M, et al. Utility of two *SMN1* variants to improve spinal muscular atrophy carrier diagnosis and genetic counselling. *Eur J Hum Genet*. 2018;26(10):1554. doi:10.1038/s41431-018-0193-4
8. Larson JL, Silver AJ, Chan D, Borroto C, Spurrier B, Silver LM. Validation of a high resolution NGS method for detecting spinal muscular atrophy carriers among phase 3 participants in the 1000 Genomes Project. *BMC Med Genet*. 2015;16:100. doi:10.1186/s12881-015-0246-2

Supplementary Figures

Figure S1. Distribution of posterior probability for simulated *SMN1* CN using a single site at different read depths and *SMN1:SMN2* CN combinations

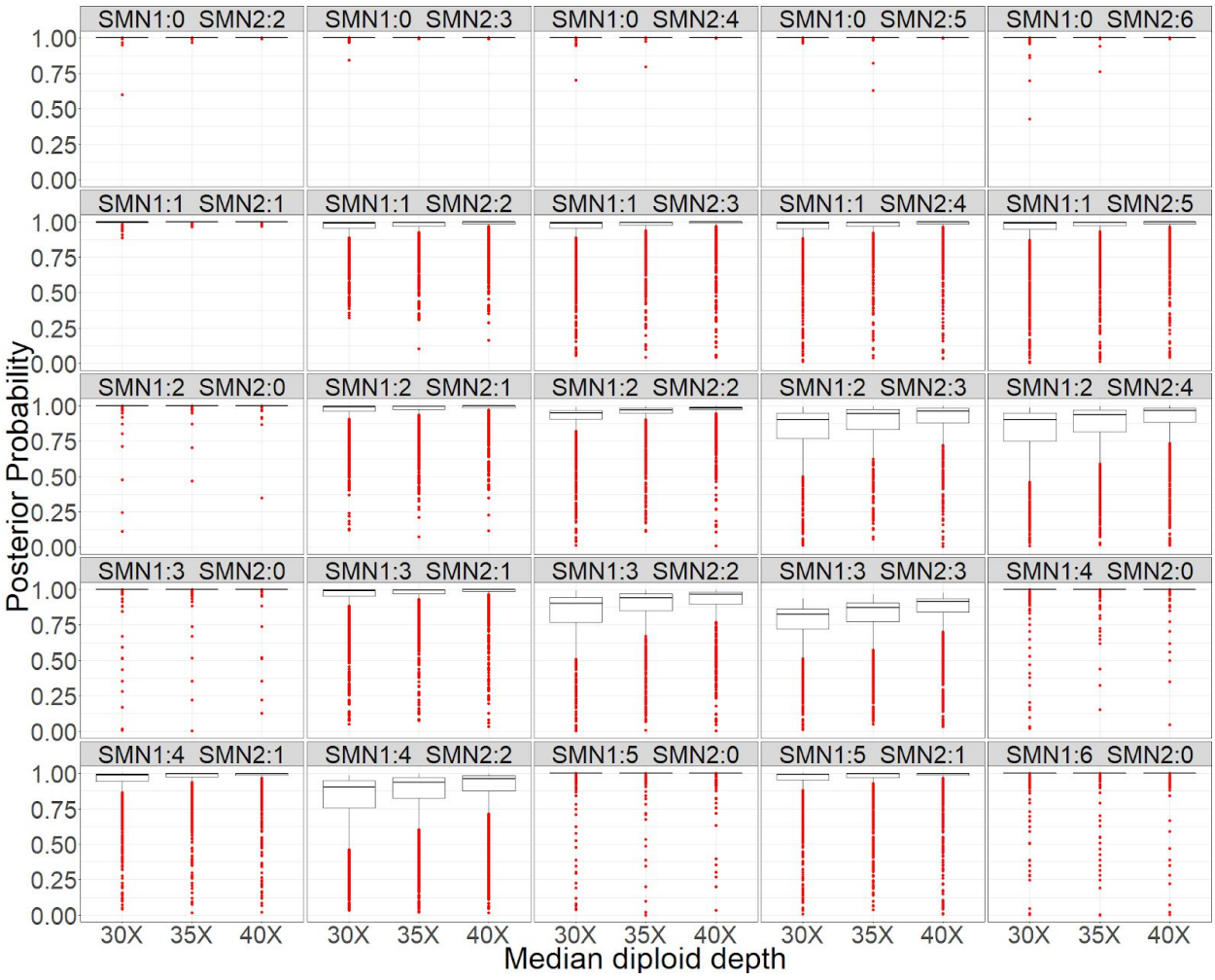


Figure S2. IGV snapshot of the *SMN2* region in a sample with the exon 7-8 deletion.

Horizontal lines join two reads in a pair in the center alignment track. BLAT results of two split reads spanning the breakpoint are shown in the bottom track, showing two segments of the same read aligning to either side of the deletion breakpoint.

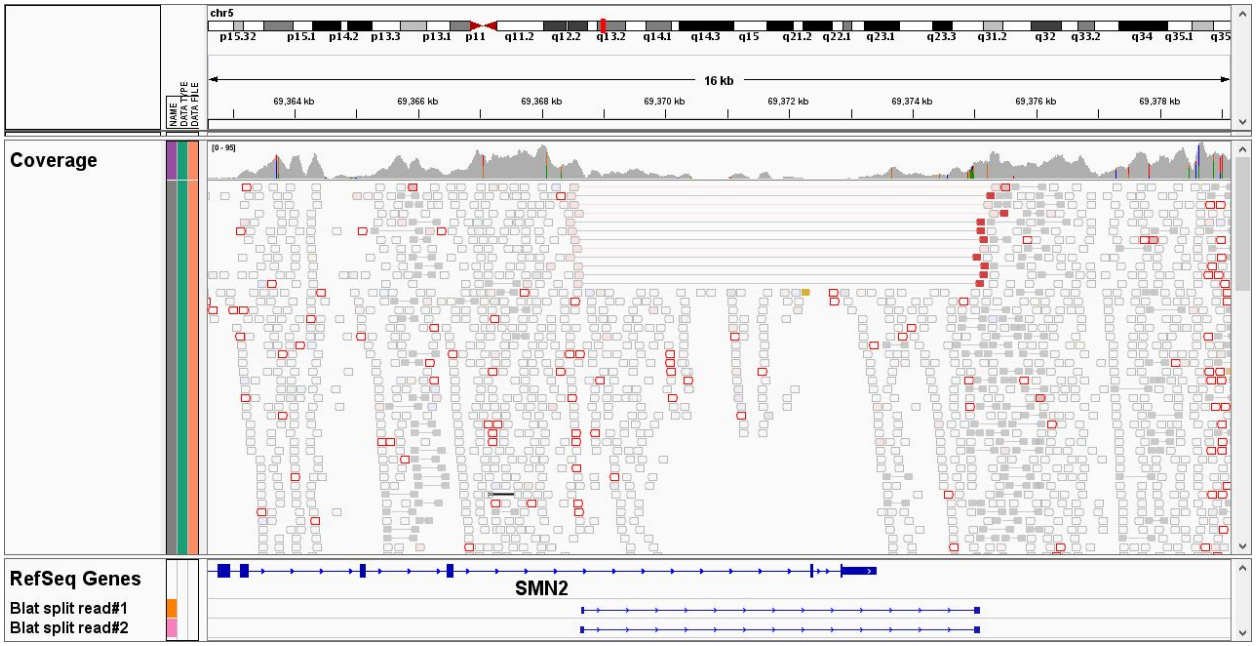


Figure S3. Correlation between raw *SMN1* CNs at 15 base differences near c840.C>T and raw *SMN1* CNs at the c840.C>T site.

The raw *SMN1* CN at each site is calculated as the CN of intact *SMN* times the fraction of *SMN1* supporting read counts out of *SMN1* + *SMN2* supporting read counts. Correlation coefficients are listed in the title of each plot.

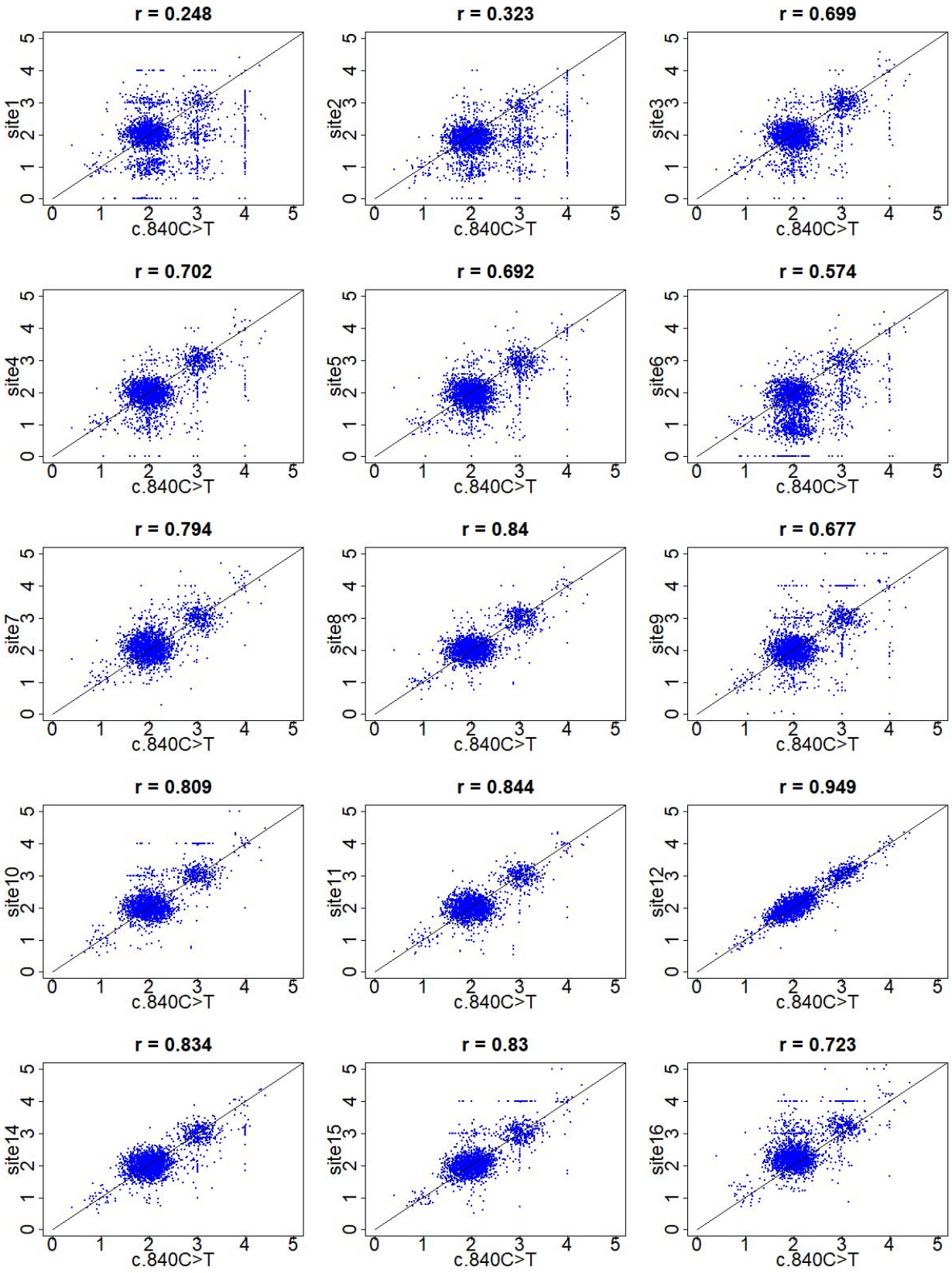

r = 0.248
site1
c.840C>T
r = 0.323
site2
c.840C>T
r = 0.699
site3
c.840C>T
r = 0.702
site4
c.840C>T
r = 0.692
site5
c.840C>T
r = 0.574
site6
c.840C>T
r = 0.794
site7
c.840C>T
r = 0.84
site8
c.840C>T
r = 0.677
site9
c.840C>T
r = 0.809
site10
c.840C>T
r = 0.844
site11
c.840C>T
r = 0.949
site12
c.840C>T
r = 0.834
site14
c.840C>T
r = 0.83
site15
c.840C>T
r = 0.723
site16
c.840C>T

Figure S4. *SMN1/SMN2/SMN2Δ7-8* CNs in 1kGP and NIHR cohorts

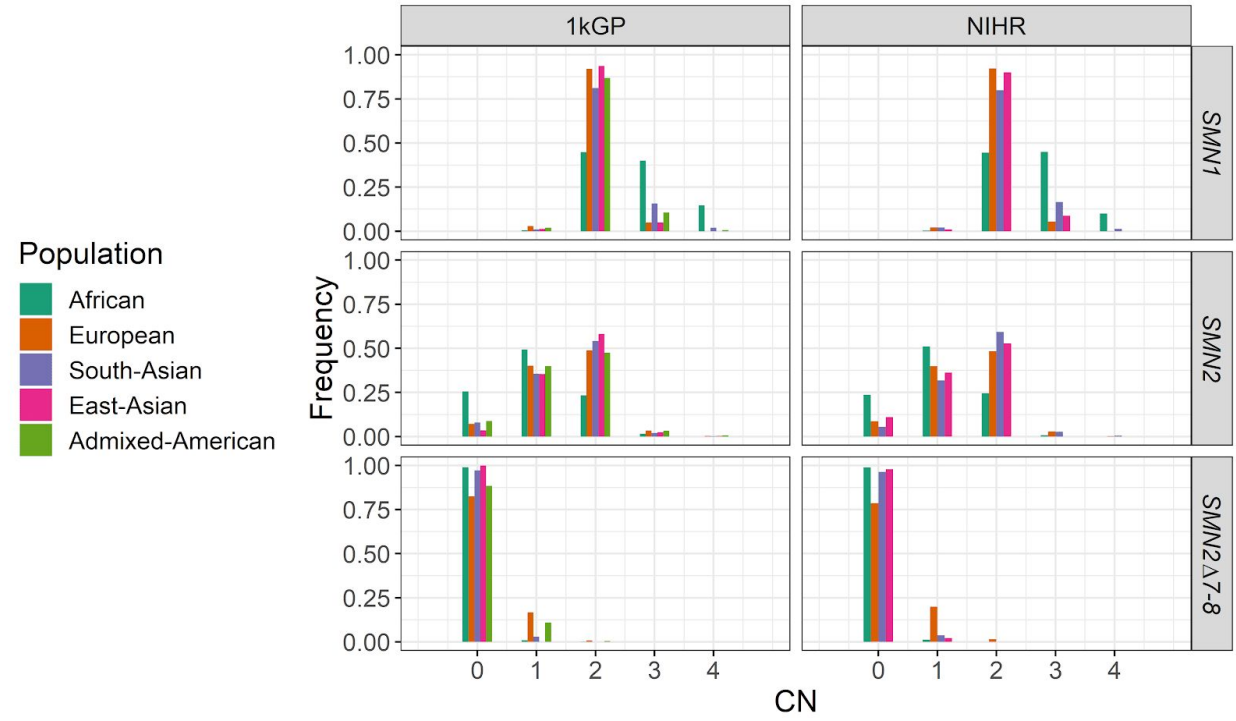


Figure S5. *SMN1*/*SMN2* haplotypes in samples with *SMN1*:2 *SMN2*:0 and *SMN1*:2 *SMN2*:1 in 1kGP.

The y axis shows the raw *SMN1* CNs as defined in Figure S3. The x axis shows the 16 sites whose indices are listed and explained in Table S1. Index #13 represents the c840.C>T site. Samples with *SMN1*:2 *SMN2*:0 are shown together in the upper left plot. Samples with *SMN1*:2 *SMN2*:1 are shown as 5 clusters. **A.** Non-Africans. **B.** Africans.

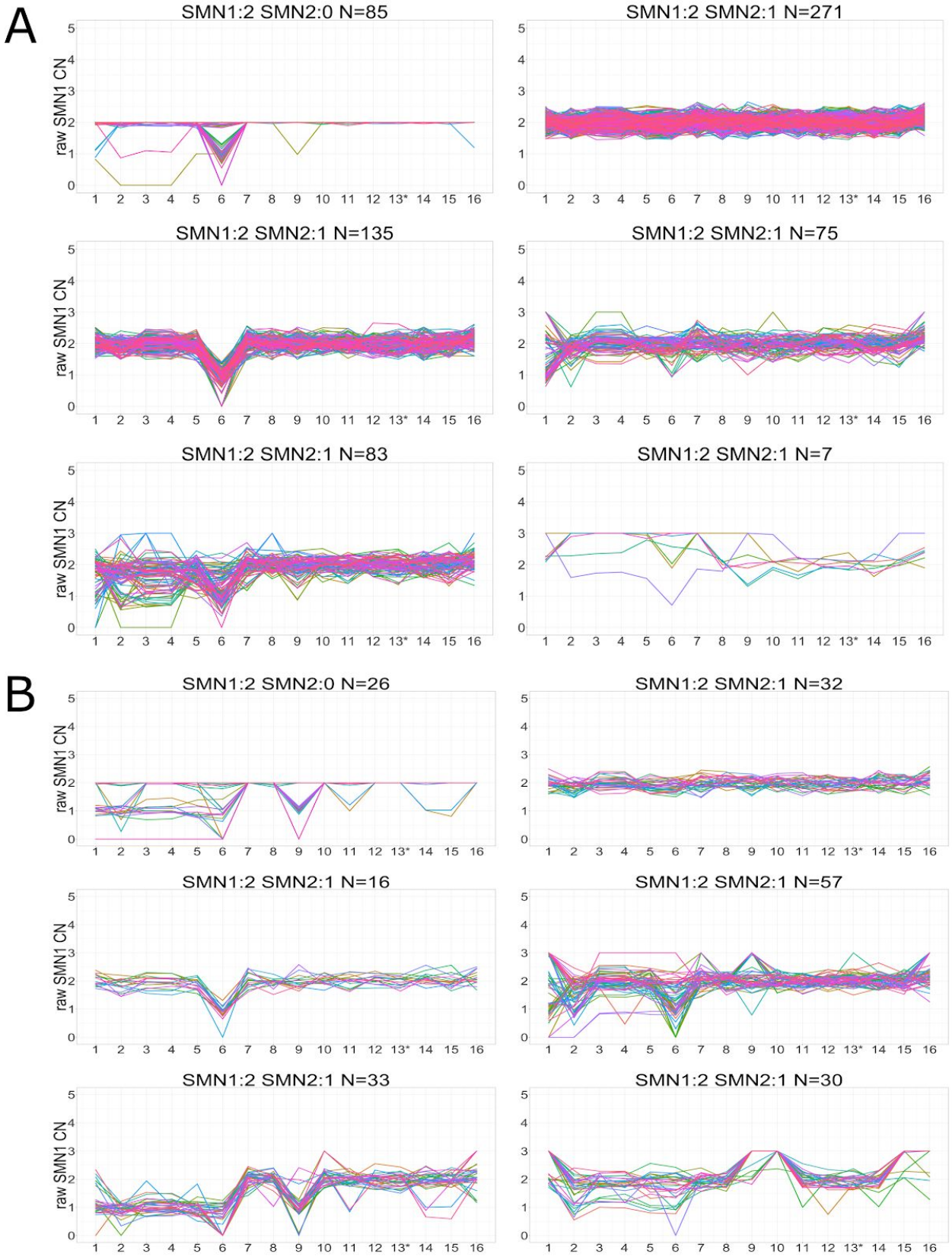

A
SMN1:2 SMN2:0 N=85
SMN1:2 SMN2:1 N=271
SMN1:2 SMN2:1 N=135
SMN1:2 SMN2:1 N=75
SMN1:2 SMN2:1 N=83
SMN1:2 SMN2:1 N=7
B
SMN1:2 SMN2:0 N=26
SMN1:2 SMN2:1 N=32
SMN1:2 SMN2:1 N=16
SMN1:2 SMN2:1 N=57
SMN1:2 SMN2:1 N=33
SMN1:2 SMN2:1 N=30
raw SMN1 CN
0 1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13* 14 15 16

Figure S6. 1.9kb deletion in *SMN1* in MB509

Analysis of split reads is consistent with the deletion occurring on *SMN1*. The deletion is not found in any of the 1kGP samples.

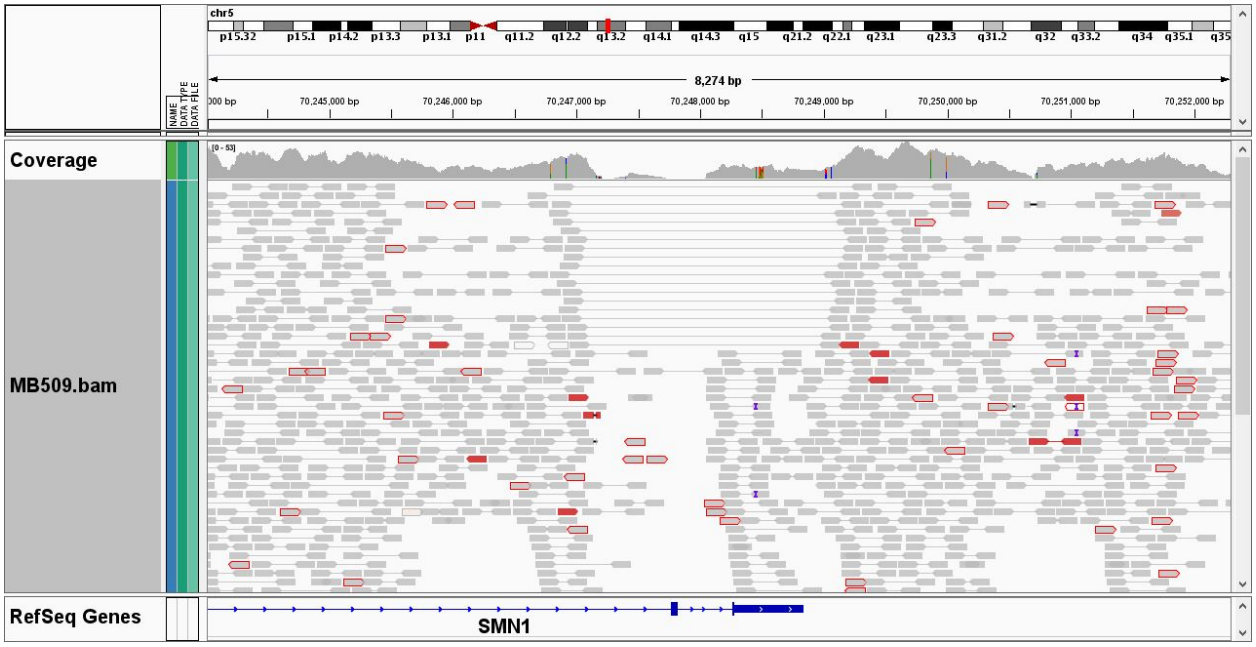


Figure S7. Discrepancies and no-calls in validation samples

A. Five samples with discrepancies between GS calls and digital PCR or MLPA results. The x axis shows the 16 sites whose indices are listed and explained in Table S1. Index #13 represents the c840.C>T site. The left y axis for the bars shows the supporting read counts for *SMN1* and *SMN2*. The right y axis for the lines shows the normalized read depth, a proxy for copy number, for *SMN1* and *SMN2* (read counts divided by haploid depth). The title of each panel shows the GS and digital PCR/MLPA calls for each sample, for *SMN1* and *SMN2*, separated by a comma. **B.** Three 1kGP validation samples where the *SMN* caller made no-calls on *SMN1* and *SMN2* CN due to disagreements among *SMN1/SMN2* base difference sites. The eight sites used for our consensus rules are #7-8 and #10-15. The y axis shows the raw *SMN1* CNs as defined in Figure S3.

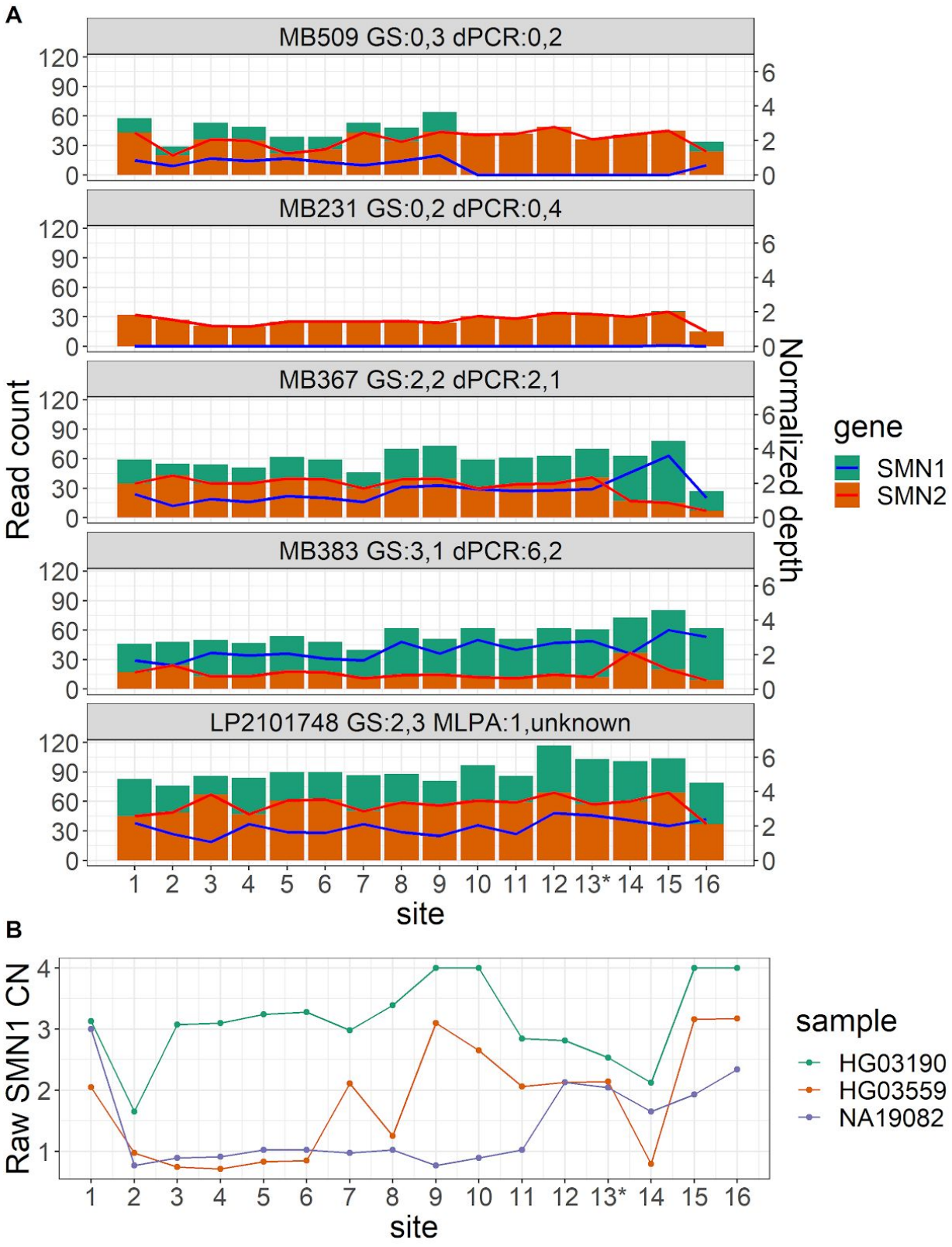
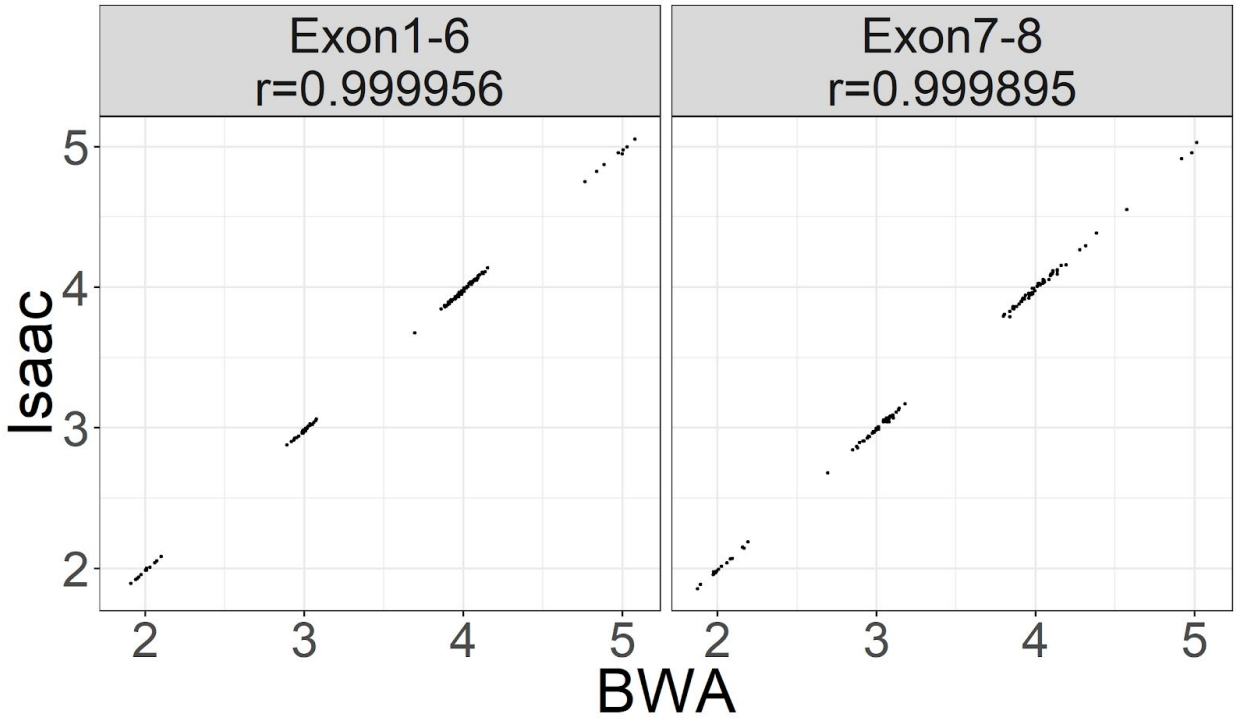

A
MB509 GS:0,3 dPCR:0,2
MB231 GS:0,2 dPCR:0,4
MB367 GS:2,2 dPCR:2,1
MB383 GS:3,1 dPCR:6,2
LP2101748 GS:2,3 MLPA:1,unknown
Read count
Normalized depth
site
gene
SMN1
SMN2
B
Raw SMN1 CN
site
sample
HG03190
HG03559
NA19082

Figure S8. CN calls derived from BWA and Isaac BAMs



Supplementary Tables

Table S1. Genome coordinates of base differences between *SMN1* and *SMN2*

Index	Location	Selected	*SMN1*		*SMN2*	
			Position, hg19	Base	Position, hg19	Base
1	Intron 6		70244142	A	69368717	G
2	Intron 6		70245876	T	69370451	C
3	Intron 6		70246016	G	69370591	A
4	Intron 6		70246019	T	69370594	C

5	Intron 6		70246156	G	69370731	A
6	Intron 6		70246167	T	69370742	C
7	Intron 6	yes	70246320	G	69370895	A
8	Intron 6	yes	70246793	G	69371368	A
9	Intron 6		70246919	A	69371499	C
10	Intron 6	yes	70247219	G	69371799	A
11	Intron 6	yes	70247290	T	69371870	C
12	Intron 6	yes	70247724	G	69372304	A
13	Exon 7 (c.840C>T)	yes	70247773	C	69372353	T
14	Intron 7	yes	70247921	A	69372501	G
15	Intron 7	yes	70248036	A	69372616	G
16	Exon 8		70248501	G	69373081	A

Table S2. Validation samples (Excel file)

Table S3. *SMN1*, *SMN2* and *SMN2Δ7-8* CN calls for 258 trios in the Next Generation Children project cohort

SMN1					*SMN2*					*SMN2Δ7-8*				
Number of families	Father	Mother	Proband1	Proband2	Number of families	Father	Mother	Proband1	Proband2	Number of families	Father	Mother	Proband1	Proband2
207	2	2	2		53	2	2	2		174	0	0	0	
8	2	2	2	2	29	2	1	1		20	0	1	0	
8	2	3	3		27	1	2	2		15	0	1	1	
8	3	2	2		23	1	2	1		15	1	0	0	
7	3	2	3		23	2	1	2		9	1	0	1	
4	2	3	2		17	1	1	1		6	0	0	0	0
3	1	2	1		12	2	0	1		4	1	1	1	
3	1	2	2		11	1	1	2		3	1	1	0	
2	1	1	0		9	1	1	0		2	0	2	1	
2	2	2	1		7	0	1	1		2	1	0	1	0
2	2	3	2	3	6	0	2	1		2	1	0	1	1
2	3	3	3		4	1	0	1		2	2	1	1	

1	2	1	1		3	0	0	0		1	0	2	2	
1	2	2	3		3	2	2	1		1	1	1	2	
					2	1	2	1	1	1	2	0	1	
					2	1	2	2	2	1	3	0	2	
					2	1	3	1						
					2	2	1	1	2					
					2	2	1	3						
					2	2	2	1	3					
					2	2	2	2	2					
					2	2	2	3						
					2	2	3	3						
					2	3	2	3						
					1	0	1	0						
					1	1	0	0						
					1	1	3	2						
					1	1	4	3						

					1	2	3	2						
					1	2	4	4						
					1	3	0	1						
					1	3	1	2						
					1	3	2	2						
					1	3	2	4						
					1	4	1	2						

Table S4. Number of samples by population in 1kGP and NIHR BioResource cohorts

Ethnicity	1kGP	NIHR BioResource, unrelated (including NGC)	NIHR BioResource, total (including NGC)
African	661	253	295
European	503	9186	11652
South Asian	489	713	1012
East Asian	504	91	97
Admixed-American	347	0	0

Other	0	0	1127
Total	2504	10243	14183

Table S5. *SMN1/SMN2/SMN2Δ7-8* CNs in 1kGP and NIHR cohorts

	1kGP					NIHR					P-value, Kolmogorov-Smirnov test
	Total	SMN CN=1	2	3	4	Total	SMN1 CN=1	2	3	4	
EUR	503	15 (2.98%)	463 (92.05%)	25 (4.97%)	0 (0.0%)	9145	197 (2.15%)	8436 (92.25%)	499 (5.46%)	13 (0.14%)	1
EAS	502	7 (1.39%)	470 (93.63%)	25 (4.98%)	0 (0.0%)	91	1 (1.1%)	82 (90.11%)	8 (8.79%)	0 (0.0%)	0.9999
AFR	653	3 (0.46%)	293 (44.87%)	261 (39.97%)	96 (14.7%)	249	1 (0.4%)	111 (44.58%)	112 (44.98%)	25 (10.04%)	0.8284
SAS	489	5 (1.02%)	397 (81.19%)	77 (15.75%)	10 (2.04%)	710	15 (2.11%)	568 (80.0%)	118 (16.62%)	9 (1.27%)	1

Table S6. *SMN1*, *SMN2* and *SMN2Δ7-8* copy number frequencies by population

Ethnicity	Total	*SMN1*				*SMN2*					*SMN2Δ7-8*		
		1	2	3	4	0	1	2	3	4	0	1	2
African	902	4 (0.44%)	404 (44.79%)	373 (41.35%)	121 (13.41%)	226 (25.06%)	449 (49.78%)	214 (23.73%)	13 (1.44%)	0 (0.0%)	892 (98.89%)	9 (1.0%)	1 (0.11%)
European	9648	212 (2.2%)	8899 (92.24%)	524 (5.43%)	13 (0.13%)	833 (8.63%)	3850 (39.9%)	4667 (48.37%)	279 (2.89%)	19 (0.2%)	7591 (78.74%)	1912 (19.83%)	137 (1.42%)
South-Asian	1199	20 (1.67%)	965 (80.48%)	195 (16.26%)	19 (1.58%)	78 (6.51%)	400 (33.39%)	686 (57.26%)	29 (2.42%)	5 (0.42%)	1155 (96.65%)	40 (3.35%)	0 (0.0%)
East-Asian	593	8 (1.35%)	552 (93.09%)	33 (5.56%)	0 (0.0%)	28 (4.72%)	211 (35.58%)	340 (57.34%)	12 (2.02%)	2 (0.34%)	591 (99.66%)	2 (0.34%)	0 (0.0%)
Admixed-American	341	7 (2.05%)	296 (86.8%)	36 (10.56%)	2 (0.59%)	30 (8.8%)	136 (39.88%)	162 (47.51%)	11 (3.23%)	2 (0.59%)	302 (88.56%)	37 (10.85%)	2 (0.59%)

Table S7. Maximum likelihood estimates for percentage of singleton and two-copy *SMN1* alleles carrying c.*3+80T>G

Ethnicity	Singleton *SMN1* allele	two-copy *SMN1* allele
African	18.4%	78.5%
European	0.02% *(1kGP European: 0.11%)	4.35% *(1kGP European: 10.0%)

South Asian	0.05%	2.54%
East Asian	0.09%	2.94%
Admixed-American	1.2%	24.5%

*The NIHR BioResource cohort, which takes up the majority of the European population analyzed in this study due to its large sample size, consists of Northern European samples that carry a lower frequency of c.*3+80T>G SNP than the more diverse European samples from the 1000 Genomes project.

Table S8. SMA carrier detection and residual risk estimates

Ethnicity	Carrier frequency[a]	Detection rate (CN)[a]	Residual risk (CN=2)	Detection rate (CN+c.*3+80 T>G SNP)	This study		Luo et al		Feng et al.		Alias et al.	
					Residual risk (CN=2, SNP-)	Residual risk (CN=2, SNP+)	Residual risk (CN=2, SNP-)	Residual risk (CN=2, SNP+)	Residual risk (CN=2, SNP-)	Residual risk (CN=2, SNP+)	Residual risk (CN=2, SNP-)	Residual risk (CN=2, SNP+)
African	1 in 72	70.5%	1 in 129	91.8%	1 in 346	1 in 58	1 in 396 (African American)	1 in 34	1 in 375 (African American)	1 in 39	NA	NA
European	1 in 47	94.8%	1 in 790	95.0%	1 in 814 (1kGP European 1 in 846)	1 in 12 (1kGP European 1 in 27)	1 in 770	1 in 29	1 in 921	1 in 69	1 in 888 (Spanish)	~1

Asian[b]	1 in 59	93.3%	1 in 767	93.4%	1 in 779	1 in 57	1 in 702	~1	1 in 907	1 in 61	NA	NA
Admixed-American	1 in 68	90.0%	1 in 559	91.9%	1 in 674	1 in 71	1 in 1762 (Hispanic)	1 in 140	1 in 906 (Hispanic)	1 in 99	NA	NA

[a]Numbers and *SMN1* allele frequencies for residual risk calculation taken from Sugarman et al.

[b]Includes East and South Asians

Table S9. Frequencies of *SMN1* haplotypes with *SMN2* allele and *SMN2* haplotypes with *SMN1* allele in two simple CN states (*SMN1*=CN2 and *SMN2*=CN0 or *SMN1*=CN2 and *SMN2*=CN1). Numbers in parentheses indicate those contributed by African populations.

Site index	# *SMN1* haplotypes with confident CN call	# *SMN1* haplotypes with *SMN2* allele	Percentage	# *SMN2* haplotypes with confident CN call	# *SMN2* haplotypes with *SMN1* allele	Percentage
1	12292	490 (71)	4	5041	101 (34)	2
2	9372	542 (79)	5.8	3669	46 (0)	1.3
3	11784	187 (48)	1.6	4788	48 (1)	1
4	11056	205 (51)	1.9	4428	43 (1)	1
5	10212	312 (51)	3.1	4087	34 (1)	0.8
6	9974	1787 (111)	17.9	3946	28 (1)	0.7

7	11956	58 (0)	0.5	4874	45 (3)	0.9
8	12218	15 (1)	0.1	5005	8 (0)	0.2
9	11872	79 (47)	0.7	4831	56 (35)	1.2
10	12484	2 (0)	0	5137	39 (29)	0.8
11	11964	19 (5)	0.2	4880	1 (0)	0
12	12506	1 (1)	0	5148	0 (0)	0
13	12836	0 (0)	0	5313	0 (0)	0
14	12386	9 (6)	0.1	5088	0 (0)	0
15	12544	9 (4)	0.1	5167	33 (24)	0.6
16	12336	12 (3)	0.1	5063	76 (41)	1.5

Table S10. Number of samples with different number of agreeing sites across the 8 base difference sites. Numbers in parentheses indicate those contributed by African populations.

SNP agreement	*SMN1* CN=1	CN=2	CN=3	CN=4	CN=no-call	Total	Percentage of sites that disagree
8	163	6325	594	111	0	7193 (475)	0 (0)
7	52	3141	285	28	0	3506 (199)	11.3 (1.6)
6	25	1197	150	9	0	1381 (137)	16.3 (6)

5	9	356	86	6	1	458 (74)	21.1 (10)
<5	2*	92*	44*	1*	36	175 (26)	19.6 (6.9)

*Calls are made in these samples based on the second consensus rule (See Supplementary Methods).

Table S11. Number of no-calls due to disagreement among sites and discrepant calls made with reduced number of sites.

# sites for majority rule	8 (Require 5 to agree)	6 (4)	4 (3)	2 (2)	1 (c.840C) (1)
# no-calls due to disagreement	175	298	766	1149	700
# calls different from those made with using 8 sites	0	0	1	6	41

Table S12. Comparison of carrier calls made in the 1kGP samples by this study and Larson et al.

Sample ID	Ethnicity	*SMN1* CN	*SMN2* CN	*SMN2Δ7-8* CN	Called as carrier in Larson et al.	Carrier probability, adj, by Larson et al.	GS calls validated by MLPA
HG03583	AFR	1	1	0	yes	0.645	yes
HG01205	AMR	1	1	0	yes	0.756	
HG01892	AMR	1	1	0	yes	0.902	yes
HG01801	EAS	1	1	0	yes	0.541	
NA11932	EUR	1	1	0	yes	0.716	
NA20760	EUR	1	1	0	yes	0.638	yes
NA20896	SAS	1	1	0	yes	0.514	yes
HG01948	AMR	1	2	0	yes	0.678	yes
HG02265	AMR	1	2	0	yes	0.982	
HG01085	AMR	1	2	0	yes	1	
NA20812	EUR	1	2	0	yes	0.999	yes
NA20764	EUR	1	2	0	yes	0.982	yes
HG00324	EUR	1	2	0	yes	0.997	yes

NA12383	EUR	1	2	0	yes	1	
HG03953	SAS	1	2	0	yes	0.972	
HG02771	AFR	1	3	0	yes	0.997	
HG01893	AMR	1	3	0	yes	1	
HG02079	EAS	1	3	0	yes	0.976	
NA20814	EUR	1	3	0	yes	1	
HG00281	EUR	1	3	0	yes	1	yes
HG00346	EUR	1	3	0	yes	1	yes
HG03740	SAS	1	3	0	yes	0.874	
HG02087	EAS	1	4	0	yes	1	
HG02134	EAS	1	4	0	yes	1	
NA12778	EUR	1	4	0	yes	1	
HG01773	EUR	1	4	0	yes	1	yes
HG01492	AMR	2	2	0	yes	0.914	
NA19723	AMR	2	2	0	yes	0.681	
NA18542	EAS	2	2	0	yes	0.633	

HG00525	EAS	2	2	0	yes	0.763	yes
NA20792	EUR	2	2	0	yes	0.671	yes
NA11843	EUR	2	2	0	yes	0.509	
NA19711	AFR	2	3	0	yes	0.943	
NA19346	AFR	2	3	0	yes	0.52	yes
HG01248	AMR	2	4	0	yes	0.935	
HG01094	AMR	2	4	0	yes	0.738	
HG02156	EAS	1	0	0	no	2.36E-33	
HG02180	EAS	1	1	0	no	7.26E-05	
NA20790	EUR	1	1	0	no	0.489	yes
NA20787	EUR	1	1	1	no	0.322	yes
HG01686	EUR	1	1	1	no	0.00119	yes
NA19456	AFR	1	2	0	no	0.278	
HG01455	AMR	1	2	0	no	0.176	
HG01863	EAS	1	2	0	no	0.42	
HG01612	EUR	1	2	0	no	1.20E-07	yes

NA20845	SAS	1	2	0	no	0.398	
HG03928	SAS	1	2	0	no	0.442	yes

Table S13. *SMN1*, *SMN2* and *SMN2*Δ*7-8* CN calls for all population samples analyzed (Excel file)