Supplementary Information

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.

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

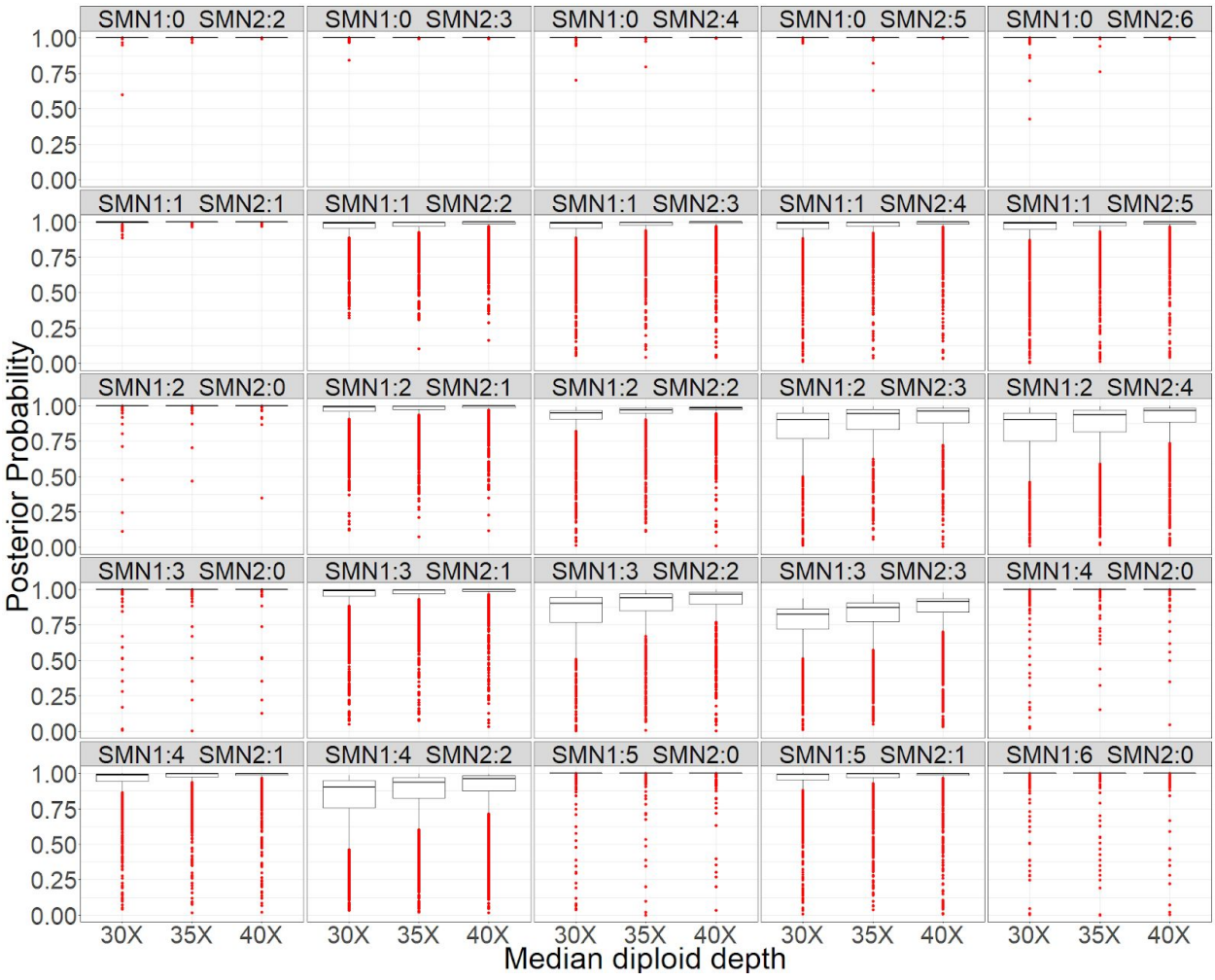


Figure S2. IGV snapshot of the *SMN2* region in a sample with the exon 7-8 deletion.

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.

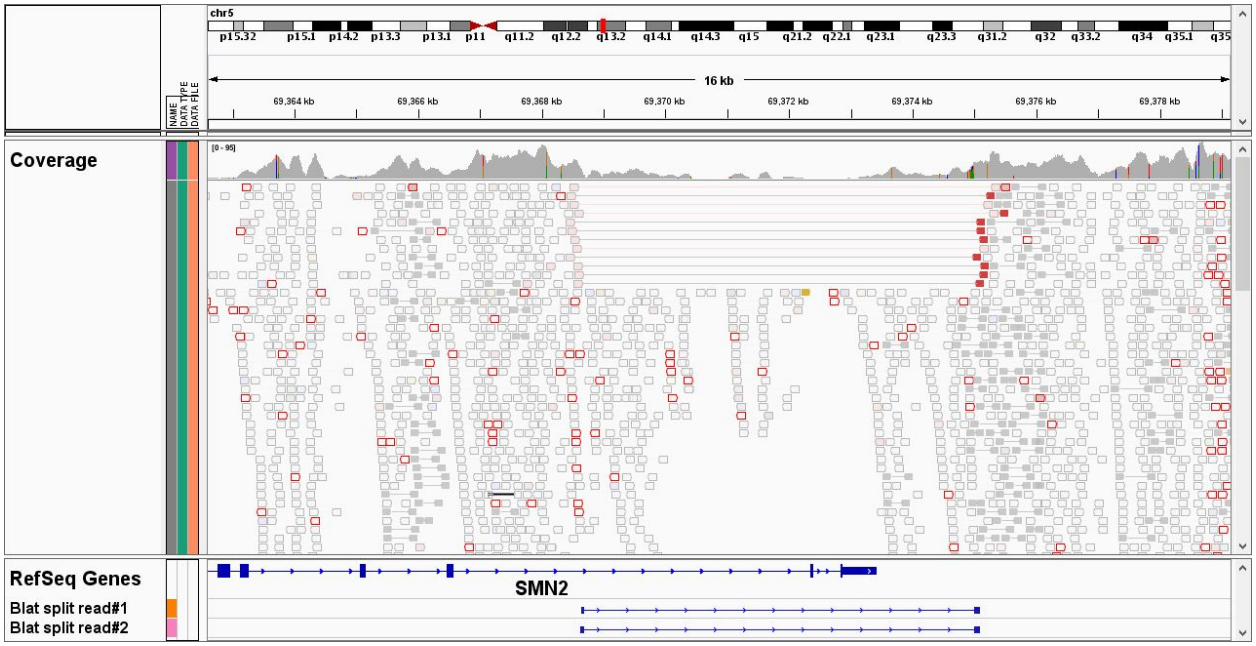


Figure S3. Correlation between raw *SMN1* CNs at 15 base differences near c840.C>T and raw *SMN1* CNs at the c840.C>T site.

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.

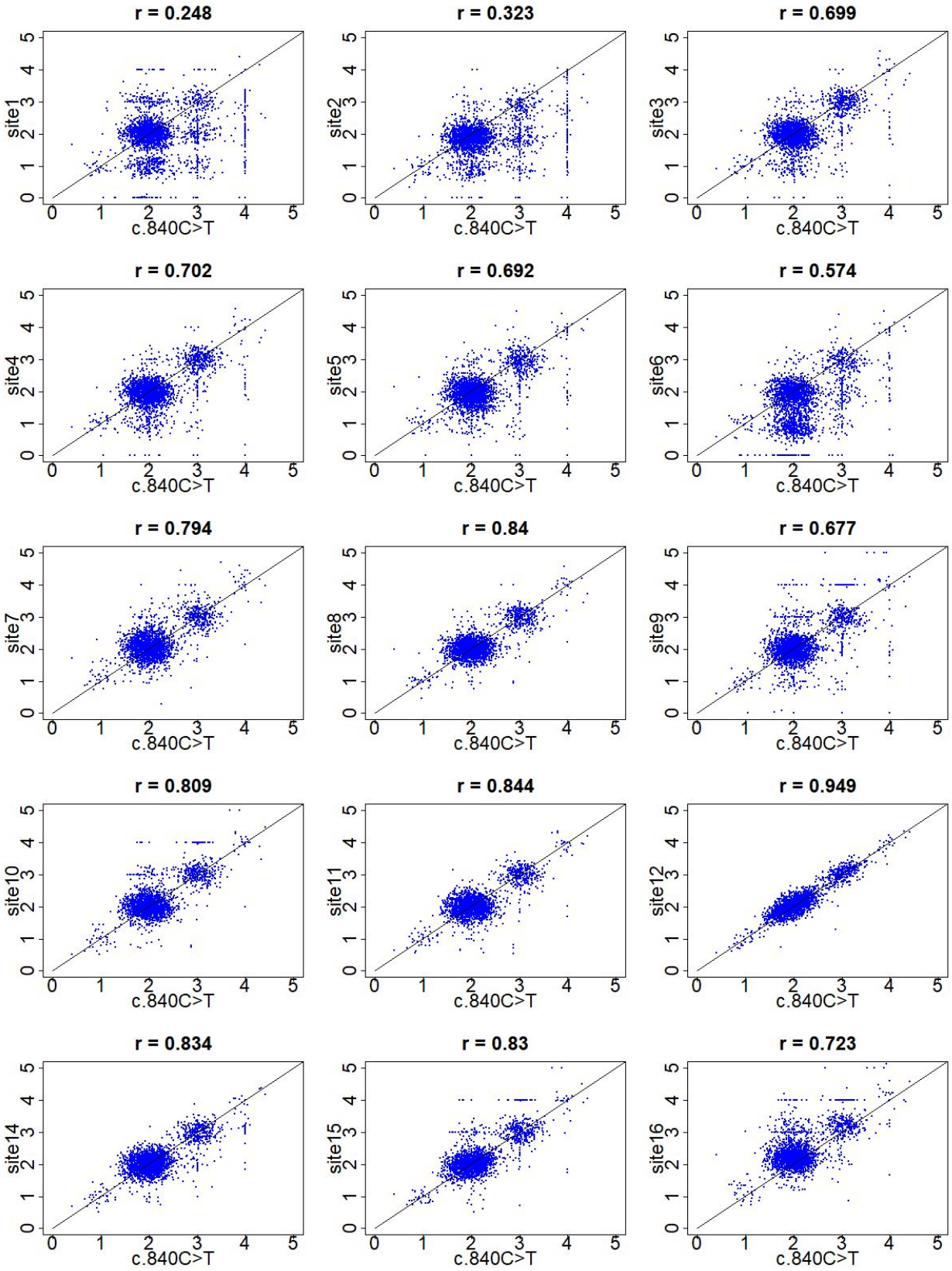

r = 0.248
r = 0.323
r = 0.699
r = 0.702
r = 0.692
r = 0.574
r = 0.794
r = 0.84
r = 0.677
r = 0.809
r = 0.844
r = 0.949
r = 0.834
r = 0.83
r = 0.723
site1
site2
site3
site4
site5
site6
site7
site8
site9
site10
site11
site12
site14
site15
site16
c.840C>T
0
1
2
3
4
5

Figure S4. *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.

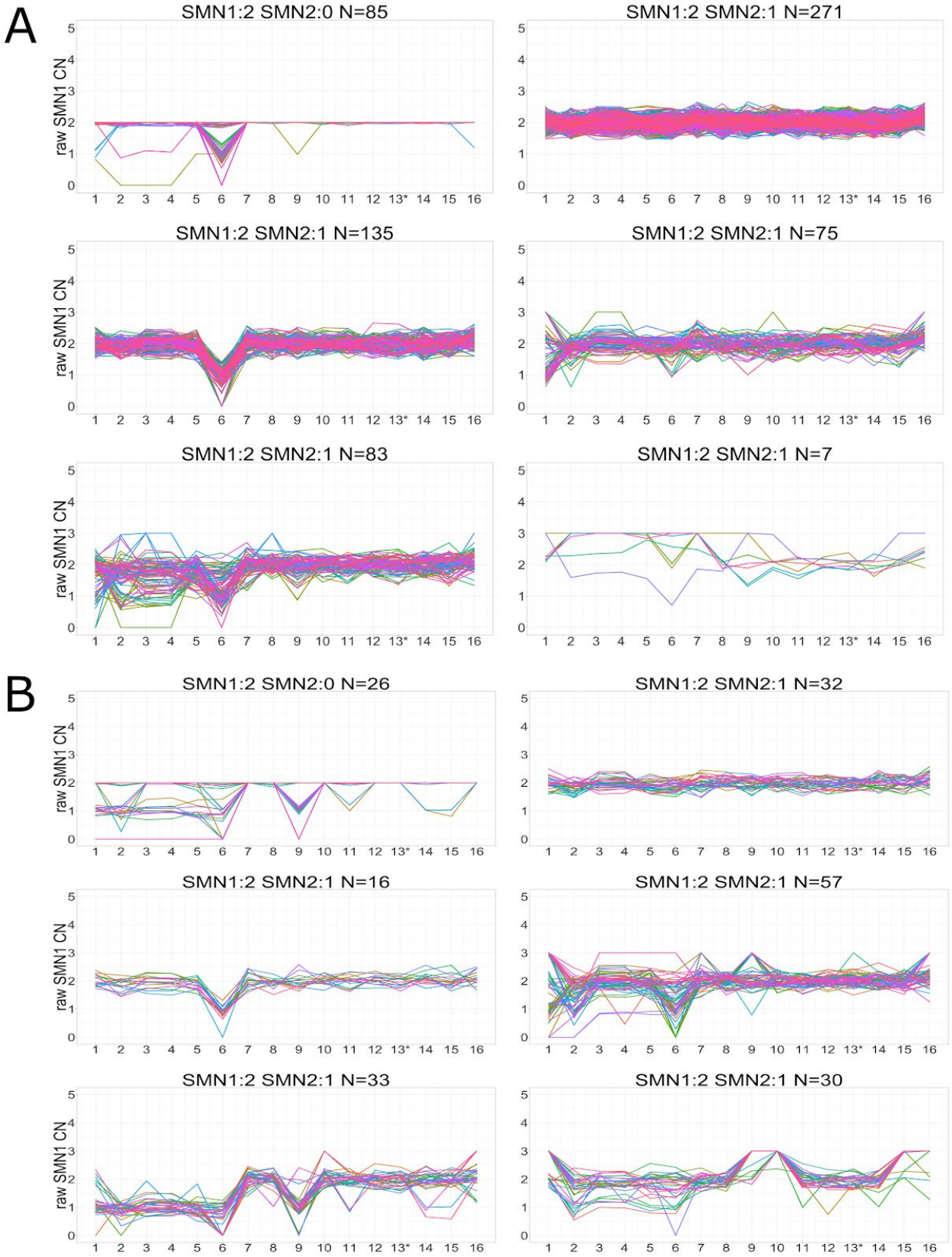

A
SMN1:2 SMN2:0 N=85
SMN1:2 SMN2:1 N=271
SMN1:2 SMN2:1 N=135
SMN1:2 SMN2:1 N=75
SMN1:2 SMN2:1 N=83
SMN1:2 SMN2:1 N=7
B
SMN1:2 SMN2:0 N=26
SMN1:2 SMN2:1 N=32
SMN1:2 SMN2:1 N=16
SMN1:2 SMN2:1 N=57
SMN1:2 SMN2:1 N=33
SMN1:2 SMN2:1 N=30
raw SMN1 CN
0 1 2 3 4 5
1 2 3 4 5 6 7 8 9 10 11 12 13* 14 15 16

Figure S5. 1.9kb deletion in *SMN1* in MB509

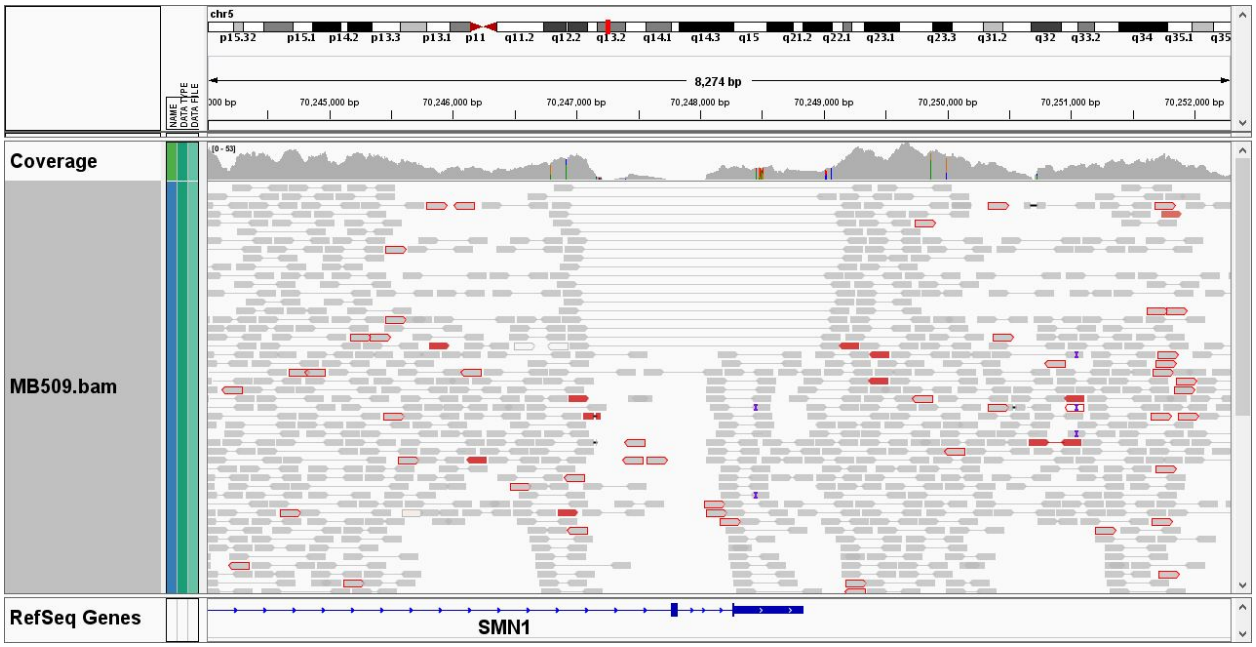
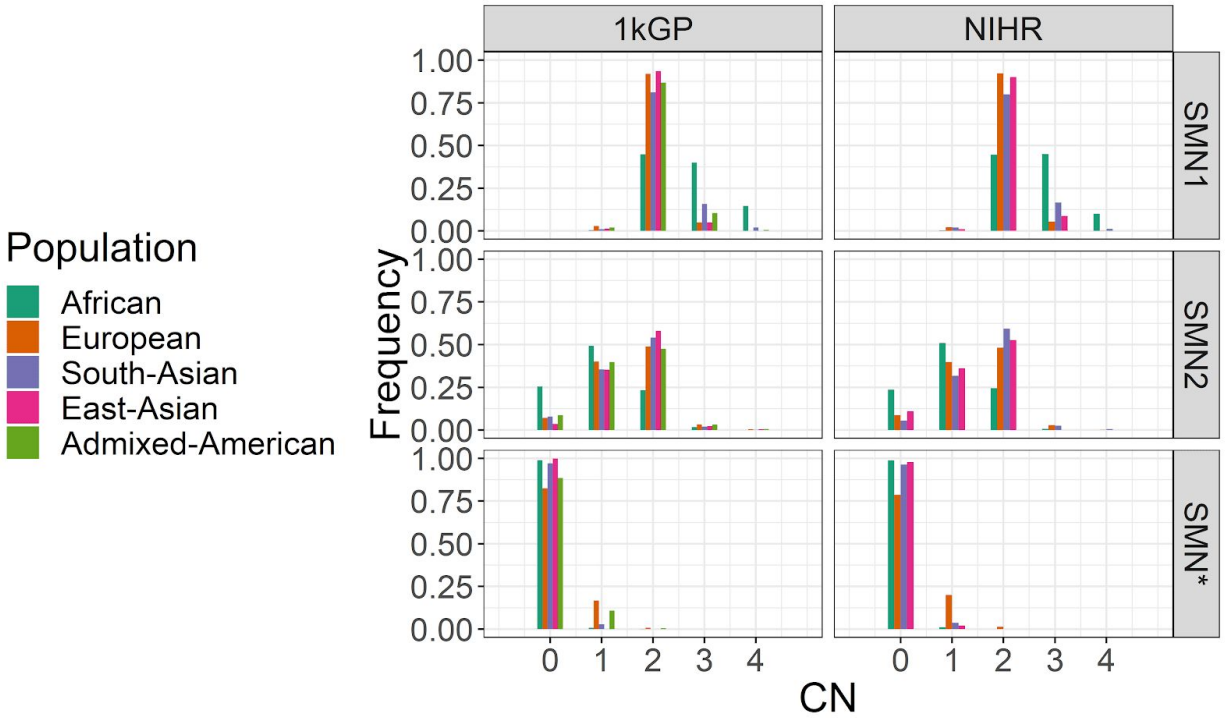


Figure S6. *SMN1/SMN2*/SMN* CNs in 1kGP and NIHR cohorts



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. 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 S3. 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 Methods).

Table S4. 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 S5. Validation samples

Sample ID	*SMN* copy number caller		Digital PCR	
	intact *SMN1* CN	intact *SMN2* CN	*SMN1* CN	*SMN2* CN
NA03813	0	3	0	3
NA09677	0	3	0	3
NA23689	0	3	0	3
NA00232	0	2	0	2
NA10684	0	2	0	2
NA23687	1	2	1	2
NA23688	1	2	1	2
NA03815	1	1	1	1
MB122	2	0	2	0
MB226	2	1	2	1
MB119	3	1	3	1

MB370	3	1	3	1
MB489	0	2	0	2
MB364	0	2	0	2
MB691	0	2	0	2
MB488	0	2	0	2
MB219	0	2	0	2
MB228	0	2	0	2
MB501	0	2	0	2
MB362	0	2	0	2
MB692	0	2	0	2
MB234	0	2	0	2
MB693	0	2	0	2
MB510	0	2	0	2
MB114	0	2	0	2
MB116	1	2	1	2
MB115	1	2	1	2
MB104	2	2	2	2
MB384	2	2	2	2
MB338	2	2	2	2
MB344	2	2	2	2
MB345	2	2	2	2
MB349	2	2	2	2
MB113	2	2	2	2
MB366	2	2	2	2
MB351	3	2	3	2
MB355	0	3	0	3

MB361	0	3	0	3
MB378	0	3	0	3
MB232	0	3	0	3
MB106	0	3	0	3
MB222	0	3	0	3
MB509	0	3	0	2
MB112	0	3	0	3
MB339	1	3	1	3
MB377	0	4	0	4
MB356	0	4	0	4
MB503	0	4	0	4

Table S6. *SMN1*, *SMN2* and *SMN** CN calls for 258 trios in the Next Generation Children project cohort

SMN1					*SMN2*					*SMN**				
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 S7. 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 S8. *SMN1*, *SMN2* and *SMN** copy number frequencies by population

Ethnicity	Total	*SMN1*				*SMN2*					*SMN**		
		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 S9. Comparison of carrier calls made in the 1kGP samples by this study and Larson et al.

Sample ID	Ethnicity	*SMN1* CN	*SMN2* CN	*SMN** CN	Called as carrier in Larson et al.	Carrier probability, adj, by Larson et al.
HG03583	AFR	1	1	0	yes	0.645
HG01205	AMR	1	1	0	yes	0.756
HG01892	AMR	1	1	0	yes	0.902
HG01801	EAS	1	1	0	yes	0.541
NA11932	EUR	1	1	0	yes	0.716
NA20760	EUR	1	1	0	yes	0.638
NA20896	SAS	1	1	0	yes	0.514

HG01948	AMR	1	2	0	yes	0.678
HG02265	AMR	1	2	0	yes	0.982
HG01085	AMR	1	2	0	yes	1
NA20812	EUR	1	2	0	yes	0.999
NA20764	EUR	1	2	0	yes	0.982
HG00324	EUR	1	2	0	yes	0.997
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
HG00346	EUR	1	3	0	yes	1
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
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
NA20792	EUR	2	2	0	yes	0.671

NA11843	EUR	2	2	0	yes	0.509
NA19711	AFR	2	3	0	yes	0.943
NA19346	AFR	2	3	0	yes	0.52
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
NA20787	EUR	1	1	1	no	0.322
HG01686	EUR	1	1	1	no	0.00119
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
NA20845	SAS	1	2	0	no	0.398
HG03928	SAS	1	2	0	no	0.442

Table S10. 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 S11. 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 S12. *SMN1*, *SMN2* and *SMN** CN calls for all population samples analyzed (Excel file)