Spinal muscular atrophy diagnosis and carrier screening from whole-genome sequencing data

Xiao Chen[1], Alba Sanchis-Juan[2,3], Courtney E French[4], Andrew J Connell[5], Aditi Chawla[1], Aaron L Halpern[1], Ryan J Taft[1], NIHR BioResource[3], David R Bentley[6], Matthew ER Butchbach[5,7,8,9], F Lucy Raymond[3,4], Michael A Eberle[1]

1. Illumina Inc., 5200 Illumina Way, San Diego, CA, USA
2. Department of Haematology, University of Cambridge, NHS Blood and Transplant Centre, Cambridge, CB2 0PT, UK
3. NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK
4. Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK
5. Center for Applied Clinical Genomics, Nemours Biomedical Research, Nemours Alfred I. duPont Hospital for Children, Wilmington, DE.
6. Illumina Cambridge Ltd., Illumina Centre 19 Granta Park, Great Abington, Cambridge, UK
7. Center for Pediatric Research, Nemours Biomedical Research, Nemours Alfred I. duPont Hospital for Children, Wilmington, DE, USA
8. Department of Pediatrics, Sidney Kimmel College of Medicine, Thomas Jefferson University, Philadelphia, PA, USA
9. Department of Biological Sciences, University of Delaware, Newark, DE, USA

Corresponding author: Michael Eberle, meberle@illumina.com

Abstract

Purpose

Spinal muscular atrophy (SMA), caused by loss of the functional *SMN1* gene, is a leading genetic cause of early childhood death. Due to the near identical sequences of *SMN1* and its paralog *SMN2*, analysis of this region is challenging. Population-wide SMA screening to quantify the *SMN1* copy number (CN) is recommended by the American College of Medical Genetics.

Methods

We developed an informatics method that accurately identifies the CN of *SMN1* and *SMN2* using whole-genome sequencing (WGS) data. This algorithm calculates the CNs of *SMN1* and *SMN2* using read depth and eight informative reference genome differences between *SMN1/2*.

Results

We characterized *SMN1/2* in 12,747 genomes across five ethnic populations and identified 251 (1317) samples with *SMN1* losses (gains) and 6241 (374) samples with *SMN2* losses (gains). We calculated a pan-ethnic carrier frequency of 2%, consistent with previous studies. Additionally, we validated our calls and all (48/48) *SMN1* and 98% (47/48) of *SMN2* CN calls agreed with digital PCR.

Conclusion

This WGS-based *SMN* copy number caller can be used to identify both carrier and affected status of SMA, enabling SMA testing to be offered as a comprehensive test in neonatal care and an accurate carrier screening tool in large-scale WGS sequencing projects.

Key words

Spinal muscular atrophy (SMA); carrier screening; copy number analysis; whole genome sequencing (WGS); bioinformatics

Introduction

Spinal muscular atrophy (SMA), an autosomal recessive neuromuscular disorder characterized by loss of alpha motor neurons, causes severe muscle weakness and atrophy presenting at or shortly after birth[1,2]. SMA is the leading genetic cause of infant death after cystic fibrosis[3]. The incidence of SMA is 1 in 6000-10,000 live births, and the carrier frequency is 1:40-80 among different ethnic groups[4–7]. The four clinical types of SMA are classified based on age of onset and severity of the disease[1]: very weak infants unable to sit unsupported (Type I), weak sitters but unable to stand (Type II), ambulant patients with weaker legs than arms (Type III) and adult onset SMA that is fairly benign (Type IV). Early detection of SMA can be critical for long term quality of life due to the availability of two early treatments, Nusinersen[8] and Zolgensma[9], which have received FDA approval for the amelioration of SMA symptoms.

The *SMN* region includes two paralogous genes: *SMN1* and *SMN2*. *SMN2* is 875kb away from *SMN1* on 5q and is created by an ancestral gene duplication that is unique to the human lineage[10,11]. The genomic region around *SMN1/2* is subject to unequal crossing-over and gene conversion, resulting in variable copy numbers (CNs) of *SMN1* and *SMN2*[7,12]. Importantly, *SMN2* has >99.9% sequence identity to *SMN1* and one of the base differences, NM_000344.3:c.840C>T in exon 7, has a critical functional consequence. By interrupting a splicing enhancer, c.840T promotes skipping of exon 7, resulting in the vast majority of *SMN2*-derived transcripts (70-85%, depending on tissue[13]) being unstable and not fully functional. Approximately 95% of SMA cases result from biallelic absence of the functional c.840C nucleotide[14] caused by either a deletion of *SMN1* or gene conversion to *SMN2* (c.840T).

In the remaining 5% of SMA cases, patients have other pathogenic variants in *SMN1 in trans* with an allele missing c.840C[15]. *SMN2* can produce a small amount of functional protein, and the number of *SMN2* copies in an individual modifies the disease severity and is highly correlated with the clinical types described above[16].

Due to the high incidence rate and disease severity, population-wide SMA screening is recommended by the American College of Medical Genetics[17]. The utility of population-wide carrier screening has been demonstrated in pilot studies[18]. The key to screening for SMA is: 1) determining the copy number of *SMN1* for SMA diagnosis and carrier testing and 2) determining the copy number of *SMN2* for clinical classification and prognosis. Traditionally, SMA testing and carrier testing are done with polymerase chain reaction (PCR) based assays, such as quantitative PCR (qPCR)[19], multiplex ligation-dependent probe amplification (MLPA)[20,21] and digital PCR[12,22]. These methods primarily determine the copy number of *SMN1* based on the c.840C>T site that differs between *SMN1* and *SMN2*.

With recent advances in next-generation sequencing (NGS), it is now possible to profile a large number of genes or even the entire genome at high throughput and in a clinically relevant timeframe. Driven by these advances, many countries are undertaking large scale population sequencing efforts[23–25] wherein testing for rare genetic disorders including carrier status will be one of the primary drivers. Demonstrating that WGS can meet or exceed the performance of PCR-based SMA tests would indicate that both current and future precision medicine initiatives could leverage genome data for population-level screening. Replicating the current SMA testing regime poses a problem for high throughput WGS due to the almost perfect sequence identity between *SMN1* and *SMN2*. Furthermore, it is thought that frequent gene conversion between *SMN1* and *SMN2* leads to hybrid genes. These challenges demand an informatics method specifically designed to overcome the difficulties of this region.

To date, two NGS-based tests for SMA carrier detection have been reported[26,27]. Larson et al.[26] used a Bayesian hierarchical model to calculate the probability that the fraction of *SMN1*-derived reads is equal to or smaller than 1/3 at three base differences between *SMN1* and *SMN2*. This method can test for SMA; though since it does not perform copy number calling, it is not an ideal solution for carrier screening. Conversely, Feng et al.[27] developed a copy number caller for both *SMN1* and *SMN2* based on targeted sequencing data that closely mimics the current qPCR method. Their method is designed for targeted sequencing and thus requires specialized normalization that limits their method to one assay at one site. Their method derives the total copy number of *SMN* (including both *SMN1* and *SMN2*) from the read coverage in exon 7, and calculates the *SMN1*:*SMN2* ratio based on the numbers of *SMN1*- and *SMN2*-supporting reads at the c.840C>T site. Using the total coverage and *SMN1*:*SMN2* ratio, the method derives the absolute copy number of *SMN1* and *SMN2*. Because this method relies solely on a single locus, it is unreliable for WGS data where per-locus depth variability can be very high.

Compared with targeted sequencing, WGS provides a much more uniform coverage across the genome and provides a less-biased approach to detecting copy number variants (CNVs). In

addition, WGS offers an opportunity to comprehensively profile the spectrum of population variation in the *SMN* region, for which our understanding at the sequence level is poor. Here, we report a novel method that detects the CN of both *SMN1* and *SMN2* using WGS data. While most conventional assays only test for the absence of c.840C as a proxy for the "exon 7 deletion", here we describe a tool that can more fully characterize the variability in the region including: 1) DNA deletions, including both whole gene deletion/duplication and a partial deletion of a region that includes exons 7 and 8; and 2) small variant detection including the NM_000344.3:c.*3+80T>G (also referred to as g.27134T>G in literature) SNP that is correlated with "silent" carriers of SMA (two copies of *SMN1* on the same haplotype)[28]. To demonstrate the accuracy of this method, we compared CN calls using digital PCR with our WGS-based calls and showed a concordance of 100% (48/48) for *SMN1* and 98% (47/48) for *SMN2*. Additionally, we applied this method to 2,504 unrelated samples from the 1000 Genomes Project[29] and 10,243 unrelated samples from the NIHR BioResource Project[30] to report on the population distributions of *SMN1* and *SMN2* copy numbers. The carrier frequencies for SMA using this method agreed with those reported by previous PCR-based studies[5,6]. In addition to demonstrating the accuracy of our method to quantify variants in the *SMN* region, we also highlight the importance of using ethnically diverse populations when developing novel informatic methods to resolve difficult clinically relevant regions of the genome.

Materials and methods

Samples and data processing

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[12,31]. This cohort contained 29 SMA samples (14 type I SMA, 1 type I/II SMA, 10 type II SMA, 3 type III SMA and 1 SMA with unknown clinical grade), six 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 13 normal samples. WGS 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)[30], which performed WGS on individuals with rare diseases and their close relatives. Additional individuals (n = 840) from the Next Generation Children (NGC) project (EGAD00001004357)[32], which performs diagnostic trio WGS on patients and their parents from neonatal and pediatric intensive care units in the UK, were also investigated. WGS 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 HiSeq X instrument, as previously described[30]. 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 the 1000 Genomes Project (1kGP) data, WGS 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.

SMN copy number analysis by orthogonal methods

For the 48 validation samples, *SMN1* and *SMN2* CNs were measured with the QuantStudio 3D Digital PCR System (Life Technologies) using allele-specific exon 7 probes as described previously[12]. *SMN1* and *SMN2* copy numbers were normalized against those for *RPPH1* (*RNase P*).

Detected SMA samples in the Next Generation Children project were confirmed using standard MLPA (SALSA MLPA P060 SMA Carrier probemix, MRC-Holland).

Copy number calling for intact and truncated *SMN*

The *SMN1* and *SMN2* loci are affected by two common CNVs, the whole-gene CNV and a partial gene deletion of exons 7 and 8[33,34] (See Results). We named the truncated form of *SMN* with the deletion of exons 7 and 8, *SMN**. Our method calls the copy number of intact *SMN1* + *SMN2* (referred to as *SMN* hereafter) and truncated *SMN* (*SMN**) genes using the following steps.

- **Identify and count reads from *SMN1* and *SMN2*:** Read counts are calculated directly from the WGS 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 *SMN**) CN, and read counts in the 6kb 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.
- **Convert normalized depth into copy numbers:** The normalized depth values across the population 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* (*SMN**) is derived by subtracting the intact *SMN* CN from total *SMN* CN calculated from the 22.2kb region that includes exons 1-6.

Genotyping the copy number of alleles at single bases

We call the number of chromosomes carrying the *SMN1* and *SMN2* bases by combining the total *SMN* CN with the read counts supporting each of the gene-specific bases. Based on the called copy number of intact *SMN* at each position, the caller iterates through all possible combinations of *SMN1* and *SMN2* copy number and derives the combination that produces the highest posterior probability for the observed number of *SMN1* and *SMN2* supporting reads. In addition to calling the CN of bases that are specific to either *SMN1* or *SMN2*, this method can be applied to variant positions to identify the copy number of SNPs known to be specific to one of the two genes, e.g. c.*3+80T>G as described in Results.

Copy number of *SMN1* and *SMN2*

For the 16 positions (localized from intron 6 to exon 8) that are different between *SMN1* and *SMN2* in the reference genome, we tested whether these sites 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). It is likely that the remaining sites are polymorphic in the population and would be less reliable to use for CN calling.

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. 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 Supplementary Information and 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, 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*.

Results

Common CNVs affecting the *SMN1*/*SMN2* loci

The genes *SMN1* and *SMN2* reside in an ~2Mb region in the reference genome with a large number of complicated segmental and inverted segmental duplications. While existing PCR- or NGS-based methods focus primarily on the c.840C>T site, we adopted a copy number approach based on the sequencing data from the full genes. As is commonly used in the community, we defined the number of *SMN1* copies as the number of *SMN* genes that carry the c.840C allele, and defined the number of *SMN2* copies as the number of *SMN* genes with c.840T allele. We performed our sequence analysis using the high depth (>30x) WGS data from 2,504 samples from the 1000 Genomes Project (1kGP), as well as the 10,243 unrelated samples from the NIHR BioResource Project (See Methods).

To formulate our CN calling strategy, we first characterized two common CNVs that lead to DNA deletions. The primary CNV that we assessed involves the entire *SMN1*/*SMN2* gene region. We examined the read depth across the ~30kb homologous region harboring *SMN1* and *SMN2* genes. Figure 1A shows the normalized read depths, in 100bp sliding windows, in samples with different *SMN1*+*SMN2* CNs across this region (representing both *SMN1* and *SMN2*). The depth profile shows that this entire region is deleted or duplicated in these samples. The exact breakpoints of this CNV are expected to vary from sample to sample due to the extensive sequence homology within and beyond this region and can only be resolved in high resolution with long read sequencing. For SMA testing, we restricted our analysis to the (~30kb) regions that include the *SMN* genes (*SMN1* or *SMN2*).

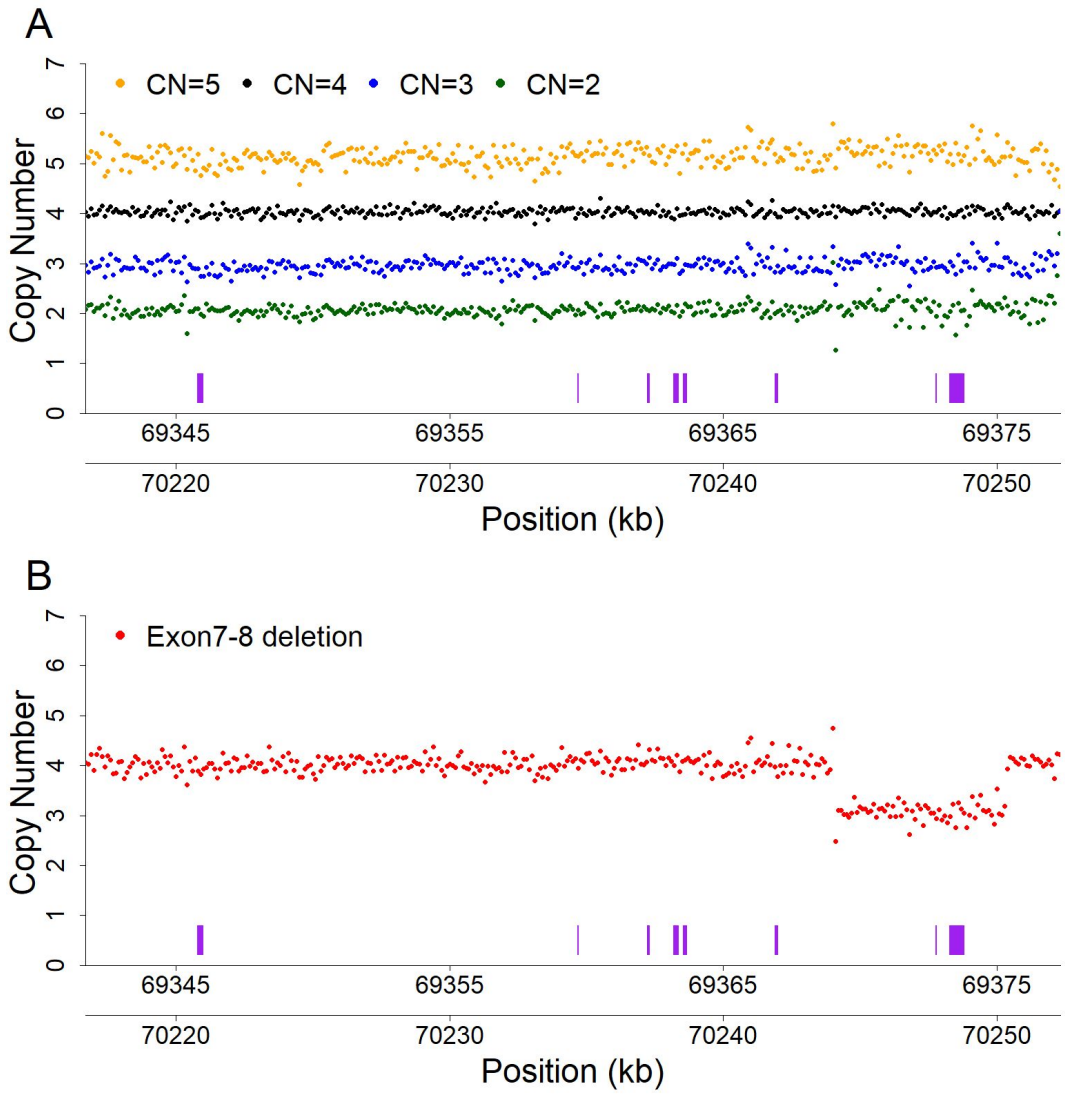

Figure 1. Common CNVs affecting the *SMN1*/*SMN2* loci.

A. Depth profiles across the *SMN1/SMN2* regions. Samples with a total *SMN1+SMN2* copy number of 2, 3, 4 and 5 are shown as green, blue, black, and orange dots, respectively. Depth from 50 samples are summed up for each CN category. Each dot represents normalized depth values in a 100bp window. Read counts are calculated in each 100bp window, summing up reads from both *SMN1* and *SMN2*, and normalized to the depth of wild-type samples (CN=4). The *SMN* exons are represented as purple boxes. The two x axes show coordinates in *SMN1* (bottom) and *SMN2* (upper). **B.** Depth profiles aggregated from 50 samples carrying a deletion of exons 7 and 8 are shown as red dots. Read depths are calculated in the same way as in (A).

In addition to whole gene CNVs, we also found a 6.3kb partial gene deletion encompassing both exons 7 and 8 (Figure 1B, Figure S2) that was recently described in another study[33]. The sequences at the breakpoint are identical between *SMN1* and *SMN2*, so this deletion occurs at either chr5: 70244114 - 70250420 in *SMN1* or chr5: 69368689 - 69375000 in *SMN2* (Figure S2, hg19). However, about 500bp downstream from the breakpoint that defines the end of this deletion there are three base differences between the *SMN1* and *SMN2* loci (70250881A>69375425C, 70250981A>69375525G, 70250991A>69375535G). Among the 1kGP samples that contain this deletion, we identified 245 read pairs from 237 samples where one spanned the breakpoint and the mate spanned at least two of the three SMN-differentiating bases. Analysis of these read pairs revealed that 100% were consistent with the deletion occurring on the *SMN2* sequence background. We named this truncated form of *SMN2*, "*SMN**" and since both exons 7 and 8 are deleted, *SMN** most likely has limited or no biological function. Therefore, *SMN** is an important variant that any *SMN* CN caller should take into account.

After searching for anomalous read pairs, we did not identify any other common CNVs in the *SMN* region. Combining this information together, we called CNs of the *SMN* genes to specifically identify the number of intact and truncated forms by dividing the genes into two regions: the 6.3kb region that includes exons 7-8 and the 22.2kb region that includes exons 1-6. The CNs of these two regions were calculated from read depth as described in Methods. The CN calculated from the exons 7-8 region provided the number of intact *SMN* genes. Samples with *SMN** have a higher CN call from the exon 1-6 region compared to the CN call from the exon 7-8 region, and this difference represents the CN of *SMN**. Figure 2 shows the results of this calculation for 12,747 samples where we identified 2,144 instances of *SMN** including 140 samples with two copies of *SMN** and one sample with three copies of *SMN**.

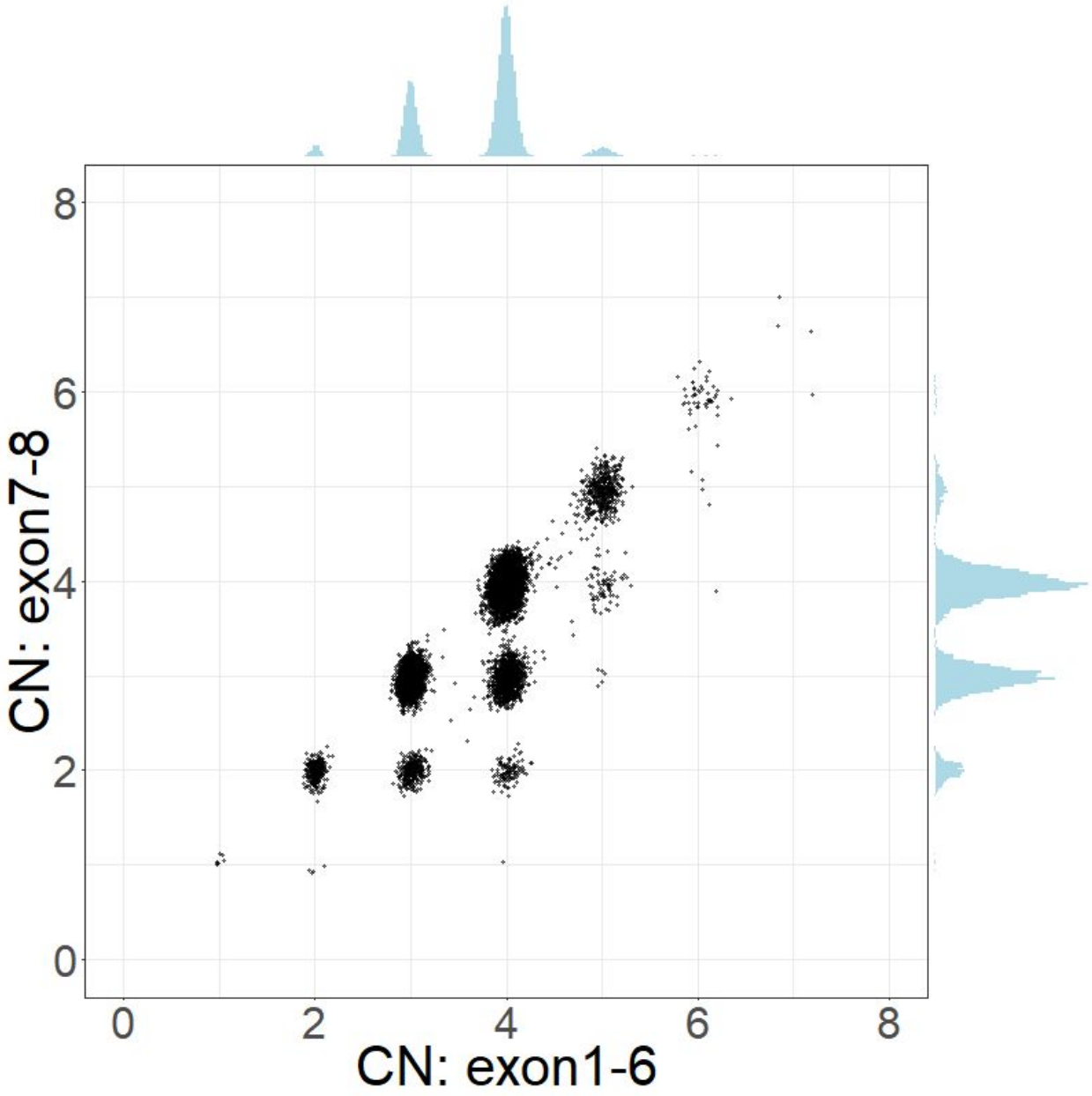

Figure 2. Scatterplot of total *SMN* (*SMN1*+*SMN2*) copy number (x axis, called by read depth in Exon 1-6) and intact *SMN* copy number (y axis, called by read depth in Exon 7-8).

Differentiating *SMN1* from *SMN2* CNs

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 WGS 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 SNPs 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 base differences, we independently called the CNs of the *SMN1* and *SMN2* alleles (see Methods) and compared the CN calls for each position with the CN calls at the splice variant site (Figure 3A, Figure S3). There was a notable difference between the concordance of calls in the African and non-African populations (Figure 3A). For the non-African samples, there were 13 sites that had high (>85%) CN concordance with the splice variant site. Conversely, for the African samples there were only seven sites that had high CN concordance with the splice variant site, and the concordance values were lower than in non-African populations. This is consistent with within-gene variation at many of these positions and higher frequencies for these non-reference alleles in the African population. We selected the splice variant site and the seven positions that were highly concordant with the splice variant site in both African and non-African populations to make CN calls on *SMN1* and *SMN2*. By restricting ourselves to two simple CN states that allow easy identification of hybrid alleles (*SMN1*=CN2 and *SMN2*=CN0 or *SMN1*=CN2 and *SMN2*=CN1), we were able to estimate the allele frequencies of these sites on the *SMN1* and *SMN2* genes (Table S2, Figure S4). Based on this analysis, we estimated that across these eight 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 S2).

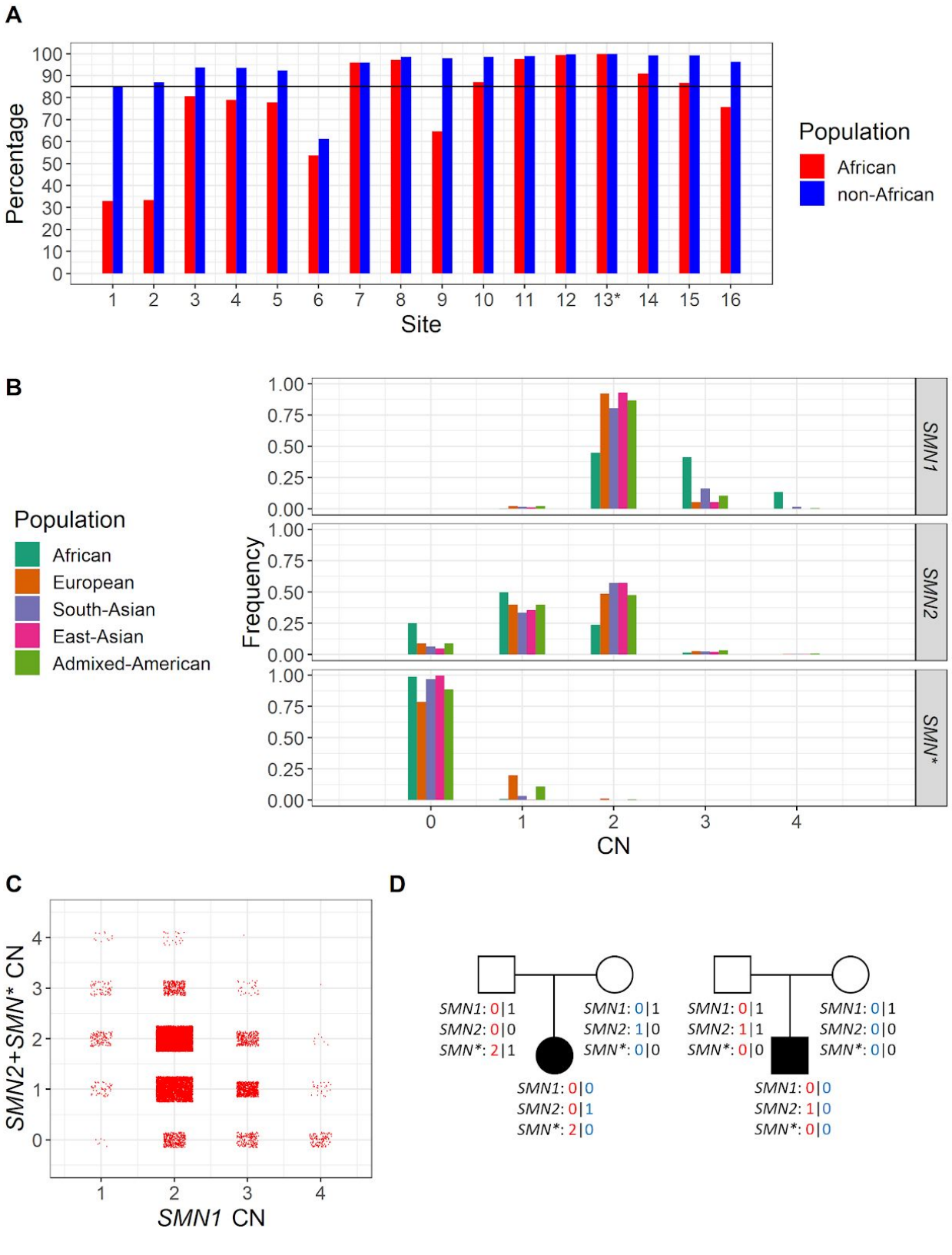

Figure 3. Distribution of *SMN1*/*SMN2*/*SMN** copy numbers in the population.

A. Percentage of samples showing CN call agreement with c.840C>T across 16 *SMN1-SMN2* base difference sites in African and non-African populations. Site 13* is the c.840C>T splice variant site. The black horizontal line denotes 85% concordance. **B.** Histogram of the distribution of *SMN1*, *SMN2* and *SMN** copy numbers across five populations in 1kGP and the NIHR BioResource cohort (numbers shown in Table S8). **C.** *SMN1* CN vs. total *SMN2* CN (intact *SMN2* + *SMN**). **D.** Two trios with an SMA proband detected by the caller and orthogonally confirmed in the Next Generation Children project cohort. CNs per allele of *SMN1*, *SMN2* and *SMN** are phased and labeled for each member of the trios.

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 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 S3). 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 S3). Using fewer sites for the majority rule produced a larger number of no-calls and wrong calls compared with using eight sites (Table S4).

Validation of the *SMN* copy number caller

To test this method, we sequenced 48 samples with known *SMN1* and *SMN2* CNs, including 29 SMA probands, 6 SMA carriers and 13 samples with an *SMN1* CN larger than 1. Our *SMN1* CN calls agreed with digital PCR results in all of the 48 cases, and the *SMN2* CN calls agreed in 47 (97.9%) of the 48 cases (Table 1, Table S5). In this single discrepant case (MB509), our caller called an *SMN2* CN of 3 while digital PCR showed an *SMN2* CN of 2 (Table S5). Upon closer examination, we found a 1884bp deletion in *SMN1* (chr5:70247145-70249029, hg19) in this sample (Figure S5). The deletion is small (does not change the depth significantly in the 6kb region used for determining the intact *SMN* CN) and has not been previously reported (nor

found in our population data), so the caller was not designed to detect it. As a result, we correctly identified this sample as SMA but overestimated the *SMN2* CN by one. 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.

We also analyzed the consistency of *SMN1/SMN2/SMN** CN calls in 258 trios from the Next Generation Children project cohort (see Methods). There is no Mendelian error in any of the calls (Table S6).

Table 1. Validation against samples with known *SMN1*/*SMN2* CNs

	CN by digital PCR	Total	Concordant	Discordant	Agreement
SMN1	0	29	29	0	100.0%
	1	6	6	0	100.0%
	2	10	10	0	100.0%
	3	3	3	0	100.0%
	Total	48	48	0	100.0%
SMN2	0	1	1	0	100.0%
	1	4	4	0	100.0%
	2	29	28	1	96.6%
	3	11	11	0	100.0%
	4	3	3	0	100.0%
	Total	48	47	1	97.9%

Copy number of *SMN1*, *SMN2* and *SMN** by population

Given the high accuracy demonstrated by our validation against digital PCR results, we next applied this method to high depth (>30x) WGS data from 12,747 unrelated samples from the 1kGP and the NIHR BioResource Project (Table S7). We analyzed the CN distributions by

population (Europeans, Africans, East Asians, South Asians and admixed Americans consisting of Colombians, Mexican-Americans, Peruvians and Puerto Ricans). Figure 3B shows the histogram of the number of individuals with different CNs of intact *SMN1*, intact *SMN2* and *SMN**. The distributions are similar between the 1kGP samples and the NIHR BioResource samples (Figure S6). In general, individuals have more copies of *SMN1* than *SMN2*. The most common combinations of *SMN1/SMN2* copy number are 2/2 (44.9%) and 2/1 (33.4%). Excluding the Africans that show higher variability in both *SMN1* and *SMN2* CN, the variability of *SMN1* copy number is much lower than that of *SMN2* copy number. Conversely, 54.7% of Africans have three or more copies of *SMN1*, which is more than double what is observed in any of the other four populations (Figure 3B, Table 2). There is an inverse relationship between the copy number of *SMN1* and *SMN2*, where the CN of *SMN2* is lower with increasing CN for *SMN1* (Figure 3C, correlation coefficient -0.344, p-value < 2.2e-16). This observation is consistent with a mechanism where gene conversion occurs between *SMN1* and *SMN2*[35,36]. The observed higher *SMN1* CN relative to *SMN2* CN could be a result of a bias towards *SMN2*-to-*SMN1* conversion or selection against a low *SMN1* CN. Africans have significantly lower *SMN2* CN than the other populations.

The number of SMA carriers identified across populations is summarized in Table 2 and Table S8. In 12,683 individuals with confident *SMN1/SMN2* CN calls, Europeans have the highest carrier frequency at 2.2%, followed by admixed Americans (2.05%), South Asians (1.67%) and East Asians (1.35%). Africans have the lowest carrier frequency (0.44%). The CN frequency distributions observed in this study are consistent with previous studies of *SMN1/SMN2* CN distribution in the general population[5,6]. In addition, we also report the frequency of the exon 7-8 deletion (*SMN**) across populations. 21.2% of Europeans and 11.5% of admixed Americans have at least one copy of *SMN**, while the frequency is lower in South Asians (3.35%), Africans (1.1%) and East Asians (0.34%).

In the Next Generation Children project cohort (see Methods), we identified SMA in two neonatal probands from trio analysis, which were confirmed independently. *SMN1*, *SMN2* and *SMN** CNs are phased for each trio member (Figure 3D).

We compared the carrier calls made in the overlapping 1kGP samples in this study (N=37) to those reported by Larson et al. (N=36), and found 26 overlapping calls (Table S9). Presuming our calls are correct this means that they made 10 false positives (FP) and 11 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 S9).

Detection of "silent" carriers

The c.*3+80T>G SNP has been reported to be associated with the 2+0 SMA silent carrier status where one chromosome carries two copies of *SMN1* (either by *SMN1* duplication or gene conversion of *SMN2* to *SMN1*) and the other chromosome has no copies of *SMN1*[28]. Our method can also detect the presence of this SNP and thus can be used to screen for potential silent carriers. This 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 (Table 2). Calling this SNP greatly increases the carrier detection rate in Africans as Africans have a higher frequency of alleles carrying two copies of *SMN1* (Table S10 and Table S11). 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. We calculated maximum likelihood estimates for the percentages of singleton and two-copy *SMN1* alleles that carry c.*3+80T>G (Table S10) and residual risks for the combination of CN and SNP calling (Table S11). Our estimates are similar to previous studies[27,28,37], 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).

Table 2. *SMN1* CN and c.*3+80T>G frequency by population

Ethnicity	Total	*SMN1* CN=1		*SMN1* CN=2		*SMN1* CN=3		*SMN1* CN=4	
		Count	c.*3+80T>G+	Count	c.*3+80T>G+	Count	c.*3+80T>G+	Count	c.*3+80T>G+
African	902	4	0(0.0%)	404	134(33.17%)	373	315(84.45%)	121	112(92.56%)
European	9648	212	0(0.0%)	8899	4(0.04%)	524	22(4.2%)	13	2(15.38%)
South-Asian	1199	20	0(0.0%)	965	1(0.1%)	195	5(2.56%)	19	1(5.26%)
East-Asian	593	8	0(0.0%)	552	1(0.18%)	33	1(3.03%)	0	0(NA)

Admixed-American	341	7	0(0.0%)	296	7(2.36%)	36	9(25.0%)	2	1(50.0%)

Discussion

Due to the high sequence homology between *SMN1* and *SMN2*, the *SMN* region is difficult to resolve with both short and long read sequencing and thus far this important region has been excluded from standard WGS analysis. Here, we demonstrate an algorithm that can resolve the CNs of *SMN1* and *SMN2* independently using short-read WGS data, filling in an important gap in SMA diagnosis and carrier screening for precision medicine initiatives. Accurate measurement of *SMN1* and *SMN2* CNs is essential not only for the diagnosis of SMA but is also a prognostic indicator and the basis of therapeutic options[38]. *SMN2* CN has been used as a criterion for many clinical trials for SMA, including Nusinersen[8] and Zolgensma[9].

As a demonstration of this algorithm, we made CN calls for *SMN1* and *SMN2* using sequencing data from 12,747 samples covering five distinct subpopulations. We identified a total of 251 samples with *SMN1* losses (less than two copies) and 1317 with *SMN1* gains (more than two copies); 6241 samples with *SMN2* losses and 1274 with *SMN2* gains; 2144 samples carrying one or more copies of the truncated form *SMN**. We cannot quantify the role that deletions, duplications or gene conversion play to drive the CN changes in this region but we see evidence supporting all three mechanisms including: 1) 3853 samples with total (*SMN1*+*SMN2*) CN<4 (deletions), 2) 670 samples with total CN>4 (duplications) and 3) a strong inverse correlation between the *SMN1* and *SMN2* CN (gene conversion, Figure 3C). Additionally, we identified a carrier frequency between 1:42 and 1:101 depending on ancestral population (Table 2). Comparing the CN frequencies by population shows that they are highly different and our per-population results agree with previous population studies[5,6]. While this provides qualitative support for the accuracy of our method, we also directly assessed its accuracy by comparing our CN calls against the results from digital PCR. In this direct comparison, all (48/48) of our *SMN1* and 98% (47/48) of our *SMN2* CN calls agreed with the digital PCR-based results. The one disagreement was due to a 2kb deletion that was not targeted by our method and, importantly, our method properly identified the SMA status of this sample.

In this study, we optimized our CN calling to work for individuals of any ancestry and thus limited *SMN1/2* differentiation to the functionally important splice variant plus seven sites in high concordance with the splice variant across all populations (Figure 3A). By quantifying the concordance between all of the reference differences and the splice variant, we were able to identify variations in these fixed differences that, if not accounted for properly (i.e. removed from our analysis) could lead to errors in our CN calls. This would be especially problematic in analyzing Africans because they harbor more diverse haplotypes. Future population genetic studies, possibly including using long read sequencing, will help profile the haplotypic diversity

across populations more directly and identify new variant sites that could further improve the accuracy of *SMN1*/*SMN2* differentiation.

An important area for improvement is the detection of “silent” carriers. One type of “silent” carrier occurs when an individual has two copies of the *SMN1* gene but they are both on the same haplotype. A SNP (c.*3+80T>G) has been used to identify individuals that are at an increased risk of being carriers when *SMN1* CN is two but the risk associated with this SNP can vary greatly between studies and populations (Table S11). When an individual has just one copy of *SMN1* they can be definitively identified as a carrier, but this variant only indicates a 2-8% chance of being a carrier when *SMN1* CN is two. With WGS, it would be possible to catalog the different variants that occur with different CN combinations of *SMN1* and *SMN2* and possibly identify additional markers that could be used to improve our ability to identify these “silent” carriers. In addition, the loss of the c.840C>T splice variant currently explains around 95% of SMA cases and the remaining cases include other pathogenic variants. These other pathogenic variants represent another type of “silent” carrier and as more of them are identified, we will extend this software to directly genotype these as part of the testing process, further improving the ability to detect SMA carriers and cases.

While there exist difficult regions in the genome where normal WGS pipelines do not deliver variant calls, here we demonstrate the ability to apply WGS paired with a targeted informatics approach to solve one such difficult region. So far, this targeted strategy (WGS + specialized informatics) has been applied successfully to a number of difficult variants, such as repeat expansions[39] and CYP2D6[40]. Traditionally, it is not cost effective to perform all of the known genetic tests and carrier screening on every individual, so candidates for specific genetic testing are identified using information such as the carrier rate and family history. However, this process means that many people without a family history who would benefit from knowing their SMA status do not routinely have access to this data. Once WGS analysis can detect all SNVs and CNVs in all clinically relevant genes accurately then a more general and population-wide genetic testing strategy will be feasible with a single test. Improving WGS to become economical as a substitute for one current genetic test will help facilitate the integration of more genetic tests and carrier screens into WGS, allowing more general access to genetic testing population-wide. WGS provides a valuable opportunity to assess the entire genome for genetic variation and the continued development of more targeted informatics solutions for difficult regions with WGS data will help bring the promise of personalized medicine one step closer to a reality.

Software and data availability

The *SMN* copy number caller described here can be downloaded from: https://github.com/Illumina/SMNCopyNumberCaller.

The 1kGP data can be downloaded from https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31736/. Data from the NIHR BioResource participants have been deposited in European Genome-phenome Archive (EGA) at the EMBL European Bioinformatics Institute (accession codes available at Ref 30[30]). Those participants from the NIHR BioResource who enrolled for the 100,000 Genomes Project–Rare Diseases Pilot can be accessed by seeking access via Genomics England Limited following the procedure outlined at: https://www.genomicsengland.co.uk/about-gecip/joining-research-community. The Bam files from the NGC individuals have been deposited in EGA under accession number EGAD00001004357.

Ethics approval and consent

For the 13,343 individuals from the NIHR BioResource Rare Diseases project, participants were recruited through NHS Cambridge University Hospitals Foundation Trust under Cambridge South Research Ethics Committee approval 13/EE/0325.

For the 48 validation samples from Motor Neuron Diseases Research Laboratory, patient-derived DNA samples were isolated from established fibroblast or lymphoblastoid cell lines. For those cell lines obtained from non-commercial sources, biospecimens were obtained after written consent or assent and parental permission. This study was approved by the Nemours/Alfred I. duPont Hospital for Children Institutional Review Board. These samples were de-identified so that no protected health information is known for these lines.

Acknowledgements

This work was supported by the Cambridge Biomedical Research Centre and the National Institute for Health Research (NIHR) for the NIHR BioResource (grant number RG65966), the National Institute of General Medical Sciences of the National Institutes of Health (P30GM114736 and P20GM103446; to MERB) and the Nemours Foundation (to MERB).
We thank the New York Genome Center (supported by NHGRI Grant 3UM1HG008901-03S1), and the Coriell Institute for Medical Research for generating and releasing the 1kGP WGS data. We thank NIHR BioResource volunteers for their participation, and gratefully acknowledge NIHR BioResource centres, NHS Trusts and staff for their contribution. We thank the National Institute for Health Research and NHS Blood and Transplant. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

References

1. Lunn MR, Wang CH. Spinal muscular atrophy. *Lancet Lond Engl*. 2008;371(9630):2120-2133. doi:10.1016/S0140-6736(08)60921-6
2. Mercuri E, Bertini E, Iannaccone ST. Childhood spinal muscular atrophy: controversies and challenges. *Lancet Neurol*. 2012;11(5):443-452. doi:10.1016/S1474-4422(12)70061-3
3. Prior TW. Perspectives and diagnostic considerations in spinal muscular atrophy. *Genet Med Off J Am Coll Med Genet*. 2010;12(3):145-152. doi:10.1097/GIM.0b013e3181c5e713
4. Ogino S, Leonard DGB, Rennert H, Ewens WJ, Wilson RB. Genetic risk assessment in carrier testing for spinal muscular atrophy. *Am J Med Genet*. 2002;110(4):301-307. doi:10.1002/ajmg.10425
5. Hendrickson BC, Donohoe C, Akmaev VR, et al. Differences in SMN1 allele frequencies among ethnic groups within North America. *J Med Genet*. 2009;46(9):641-644. doi:10.1136/jmg.2009.066969
6. Sugarman EA, Nagan N, Zhu H, et al. Pan-ethnic carrier screening and prenatal diagnosis for spinal muscular atrophy: clinical laboratory analysis of >72 400 specimens. *Eur J Hum Genet*. 2012;20(1):27-32. doi:10.1038/ejhg.2011.134
7. MacDonald WK, Hamilton D, Kuhle S. SMA carrier testing: a meta-analysis of differences in test performance by ethnic group. *Prenat Diagn*. 2014;34(12):1219-1226. doi:10.1002/pd.4459
8. Finkel RS, Chiriboga CA, Vajsar J, et al. Treatment of infantile-onset spinal muscular atrophy with nusinersen: a phase 2, open-label, dose-escalation study. *Lancet Lond Engl*. 2016;388(10063):3017-3026. doi:10.1016/S0140-6736(16)31408-8
9. Mendell JR, Al-Zaidy S, Shell R, et al. Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy. *N Engl J Med*. 2017;377(18):1713-1722. doi:10.1056/NEJMoa1706198
10. Lefebvre S, Bürglen L, Reboullet S, et al. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell*. 1995;80(1):155-165. doi:10.1016/0092-8674(95)90460-3
11. Rochette CF, Gilbert N, Simard LR. SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. *Hum Genet*. 2001;108(3):255-266.
12. 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
13. 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 U S A*. 1999;96(11):6307-6311.
14. Wirth B. An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinal muscular atrophy (SMA). *Hum Mutat*. 2000;15(3):228-237. doi:10.1002/(SICI)1098-1004(200003)15:3<228::AID-HUMU3>3.0.CO;2-9
15. Burghes AHM, Beattie CE. Spinal Muscular Atrophy: Why do low levels of SMN make motor neurons sick? *Nat Rev Neurosci*. 2009;10(8):597-609. doi:10.1038/nrn2670
16. Butchbach MER. Copy Number Variations in the Survival Motor Neuron Genes: Implications for Spinal Muscular Atrophy and Other Neurodegenerative Diseases. *Front Mol Biosci*. 2016;3. doi:10.3389/fmolb.2016.00007
17. Prior TW. Carrier screening for spinal muscular atrophy. *Genet Med*. 2008;10(11):840-842. doi:10.1097/GIM.0b013e318188d069

18. Kraszewski JN, Kay DM, Stevens CF, et al. Pilot study of population-based newborn screening for spinal muscular atrophy in New York state. *Genet Med*. 2018;20(6):608-613. doi:10.1038/gim.2017.152
19. Feldkötter M, Schwarzer V, Wirth R, Wienker TF, Wirth B. Quantitative analyses of SMN1 and SMN2 based on real-time lightCycler PCR: fast and highly reliable carrier testing and prediction of severity of spinal muscular atrophy. *Am J Hum Genet*. 2002;70(2):358-368. doi:10.1086/338627
20. Arkblad EL, Darin N, Berg K, et al. Multiplex ligation-dependent probe amplification improves diagnostics in spinal muscular atrophy. *Neuromuscul Disord NMD*. 2006;16(12):830-838. doi:10.1016/j.nmd.2006.08.011
21. Scarciolla O, Stuppia L, De Angelis MV, et al. Spinal muscular atrophy genotyping by gene dosage using multiple ligation-dependent probe amplification. *Neurogenetics*. 2006;7(4):269-276. doi:10.1007/s10048-006-0051-3
22. Zhong Q, Bhattacharya S, Kotsopoulos S, et al. Multiplex digital PCR: breaking the one target per color barrier of quantitative PCR. *Lab Chip*. 2011;11(13):2167-2174. doi:10.1039/c1lc20126c
23. Ashley EA. The Precision Medicine Initiative: A New National Effort. *JAMA*. 2015;313(21):2119-2120. doi:10.1001/jama.2015.3595
24. The Genome of the Netherlands Consortium, Francioli LC, Menelaou A, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46(8):818-825. doi:10.1038/ng.3021
25. Turnbull C, Scott RH, Thomas E, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*. 2018;361:k1687. doi:10.1136/bmj.k1687
26. 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
27. 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
28. 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
29. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
30. 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
31. 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
32. 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
33. Ruhno C, McGovern VL, Avenarius MR, 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(3):241-256. doi:10.1007/s00439-019-01983-0

34. Vijzelaar R, Snetselaar R, Clausen M, et al. The frequency of SMN gene variants lacking exon 7 and 8 is highly population dependent. *PloS One*. 2019;14(7):e0220211. doi:10.1371/journal.pone.0220211
35. Ogino S, Gao S, Leonard DGB, Paessler M, Wilson RB. Inverse correlation between SMN1 and SMN2 copy numbers: evidence for gene conversion from SMN2 to SMN1. *Eur J Hum Genet*. 2003;11(3):275. doi:10.1038/sj.ejhg.5200957
36. Chen T-H, Tzeng C-C, Wang C-C, et al. Identification of bidirectional gene conversion between SMN1 and SMN2 by simultaneous analysis of SMN dosage and hybrid genes in a Chinese population. *J Neurol Sci*. 2011;308(1):83-87. doi:10.1016/j.jns.2011.06.002
37. 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
38. Mercuri E, Finkel RS, Muntoni F, et al. Diagnosis and management of spinal muscular atrophy: Part 1: Recommendations for diagnosis, rehabilitation, orthopedic and nutritional care. *Neuromuscul Disord NMD*. 2018;28(2):103-115. doi:10.1016/j.nmd.2017.11.005
39. Dolzhenko E, van Vugt JJFA, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res*. 2017;27(11):1895-1903. doi:10.1101/gr.225672.117
40. Lee S, Wheeler MM, Patterson K, et al. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet Med*. 2019;21(2):361. doi:10.1038/s41436-018-0054-0