European Journal of Human Genetics (2004) 12, 1015–1023
© 2004 Nature Publishing Group All rights reserved 1018-4813/04 $30.00
www.nature.com/ejhg

ARTICLE

New insights on the evolution of the *SMN1* and *SMN2* region: simulation and meta-analysis for allele and haplotype frequency calculations

Shuji Ogino*[,1,2,3], Robert B Wilson[4] and Bert Gold[5]

[1]*Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA;* [2]*Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA;* [3]*Harvard Medical School, Boston, MA, USA;* [4]*Department of Pathology and Laboratory Medicine, University of Pennsylvania Medical Center, Philadelphia, PA, USA;* [5]*Human Genetics Section, Laboratory of Genomic Diversity, National Cancer Institute—Frederick, Frederick, MD, USA*

Most spinal muscular atrophy patients lack both copies of *SMN1*. Loss of *SMN1* ('0-copy alleles') can occur by gene deletion or *SMN1*-to-*SMN2* gene conversion. Despite worldwide efforts to map the segmental duplications within the *SMN* region, most assemblies do not correctly delineate these genes. A near pericentromeric location provides impetus for the strong evidence that *SMN1* and *SMN2* arose from a primate-specific paralogous gene duplication. Here we meta-analyzed our recent laboratory results together with available published data, in order to calculate new mutation rates and allele/haplotype frequencies in this recalcitrant and highly unstable region of the human genome. Based on our tested assumption of compliance with Hardy–Weinberg equilibrium, we conclude that the *SMN1* allele frequencies are: '0-copy disease alleles,' 0.013; '1-copy normal alleles,' 0.95; '2-copy normal alleles (ie, two copies of *SMN1* on one chromosome),' 0.038; and '1^D disease alleles (*SMN1* with a small intragenic mutation),' 0.00024. The *SMN1* haplotype ['(*SMN1* copy number)-(*SMN2* copy number)'] frequencies are: '0-0,' 0.00048; '0-1,' 0.0086; '0-2,' 0.0042; '1-0,' 0.27; '1-1,' 0.66; '1-2,' 0.015; '2-0,' 0.027; and '2-1,' 0.012. Paternal and maternal *de novo* mutation rates are 2.1×10^{-4} and 4.2×10^{-5}, respectively. Our data provide the basis for the most accurate genetic risk calculations, as well as new insights on the evolution of the *SMN* region, with evidence that nucleotide position 840 (where a transition 840C>T functionally distinguishes *SMN2* from *SMN1*) constitutes a mutation hotspot. Our data also suggest selection of the 1-1 haplotype and the presence of rare chromosomes with three copies of *SMN1*.

European Journal of Human Genetics (2004) **12,** 1015–1023. doi:10.1038/sj.ejhg.5201288
Published online 6 October 2004

Keywords: *SMN1*; SMN; spinal muscular atrophy; copy number; genetic counseling; risk assessment

Introduction

Spinal muscular atrophy (SMA; type I, OMIM# 253300; type II, OMIM# 253550; type III, OMIM# 253400), an autosomal recessive disorder characterized by loss of motor neurons in the anterior horn of the spinal cord, affects ~1 in 10 000 live births.[1] All three types of SMA are caused by mutations in *SMN1* (OMIM# 600354) on chromosome 5q13.[2] The coding regions of *SMN1* and its homologue, *SMN2* (OMIM# 601627) differ in only one base.[3] Although translationally silent, this transition 840C>T in *SMN2* exon 7 alters the splicing pattern of *SMN2* transcripts, resulting in a lower level of full-length SMN mRNA from *SMN2* than from *SMN1*.[4,5] Recent evidence suggests that 840C>T in *SMN2* activates an exonic splicing silencer,

*Correspondence: Dr S Ogino, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St., Boston, MA 02115, USA. Tel: +1-617-632-3978; Fax: +1-617-277-9015; E-mail: shuji_ogino@dfci.harvard.edu
Received 27 December 2003; revised 16 July 2004; accepted 28 July 2004

which functions as a binding site for the known repressor protein hnRNP A1.[6] Approximately 94% of clinically typical SMA patients lack both copies of *SMN1* by either gene deletion or *SMN1*-to-*SMN2* gene conversion (ie, 840C>T).[7] Approximately 30 small intragenic mutations have also been identified.[8]

By *SMN* gene dosage analyses to identify SMA carriers with only one copy of *SMN1*, three copies of *SMN1* have been identified in normal individuals, implying the presence of two copies of *SMN1* on a single chromosome 5.[9,10] The presence of this '2-copy-*SMN1* allele' is associated with a decreased *SMN2* copy number.[11] Importantly, SMA carriers may have two copies of *SMN1* on one chromosome 5 and a deletion/conversion mutation of *SMN1* on the other chromosome 5 (the '2+0' *SMN1* genotype).[10] By dosage analysis alone, such carriers, as well as carriers of small, intragenic mutations, are indistinguishable from normal individuals with one copy of *SMN1* on each chromosome 5, unless a monosomal analysis technique[12] is utilized. Adding to the complexity of SMA genetics is a high *de novo SMN1* mutation rate.[13] The pericentromeric location[14] and presence of the paralogous segmental duplication[15] likely predispose the region to recombination events, leading to the high observed *de novo* deletion mutation rate.

Because of the genetic complexity of SMA and the high carrier frequency, genetic counseling and risk assessment are essential components of genetic testing for SMA patients and their families.[16] However, previous risk-assessment studies were based on data derived from a relatively limited number of studies and subjects.[1,16,17] For the most accurate possible risk calculations, we meta-analyzed all available and reliable data including our own new data to determine the *SMN1* allele frequencies and *de novo SMN1* mutation rates. To better understand the molecular evolution of the *SMN* region, we also estimated the various *SMN1-SMN2* haplotype frequencies by haplotype reconstruction and meta-analysis.

Symbols used

Throughout the manuscript, a plus (+) symbol separates the *SMN1* copy number on each chromosome; a colon (:) separates the total *SMN1* copy number and the total *SMN2* copy number; and a hyphen (-) separates the *SMN1* copy number and the *SMN2* copy number on the same chromosome (*SMN1-SMN2* haplotype). An *SMN1* genotype is indicated by '(*SMN1* allele on one chromosome 5)+(*SMN1* allele on the other chromosome 5).' For example, a '2+0' genotype denotes two copies of *SMN1* on one chromosome 5 ('2-copy allele'), and zero *SMN1* copies on the other chromosome 5 ('0-copy allele').

Methods

Subject samples and *SMN* gene dosage analysis

SMN gene dosage analysis to determine the copy numbers of *SMN1* and *SMN2* was performed on genomic DNA samples from 176 unaffected individuals without a family history of SMA in the Molecular Pathology Laboratory, Hospital of the University of Pennsylvania, on a clinical basis. These were new cases in addition to those analyzed for our previous studies.[11,18] Results were anonymized and used for this study. There was no evidence for the presence of any ethnic group with skewed *SMN1* or *SMN2* copy numbers. Genomic DNA was extracted from peripheral blood leukocytes using Puregene reagents (Gentra Systems, Minneapolis, MN, USA). *SMN* gene dosage analysis was developed and validated as described previously.[10,19,20] *SMN1* and *SMN2* copy numbers were determined by quantification of the PCR products after *Dra*I digestion to differentiate *SMN1* from *SMN2*, and normalizations to genomic standards, internal standards, and two-*SMN1*-copy controls. All samples were assayed in duplicate.

Literature used for data collection

We previously described methods of meta-analysis to determine *SMN1* allele frequencies.[1] We improved our methods by using *SMN1* copy number data in the general population for the calculations of disease allele frequencies. At the time of our previous study, copy number data were insufficient. For the calculation of *SMN1* disease allele frequencies, SMA population incidence data, which we used previously, may overestimate the disease allele frequency due to confounders such as consanguinity. Population data might also lead to underestimation of the disease allele frequency, because SMA patients with very severe disease might not survive to birth. In addition, SMA diagnoses in population studies are, in general, based on variable clinical criteria, rather than genetic testing results. To calculate *SMN1* allele frequencies as accurately as possible, and in the absence of a single large study, all reliable data available in the literature should be used. We included the following data: the frequencies of individuals with one, two, three, and four copies of *SMN1* in the general population[9,18,21–24] (and this study) (Table 1); the fraction of individuals for each SMA type who lack both copies of *SMN1* exon 7 among those with identifiable mutations in both *SMN1* alleles;[25] and the frequency of patients completely lacking *SMN1* who received a *de novo SMN1* deletion/conversion mutation.[1,13,18,23,25] We excluded data from other studies for reasons described previously.[1] The data of Cusin *et al*[23] include those of Gerard *et al*[26] (V Cusin, personal communication). To determine the various *SMN1-SMN2* haplotype frequencies, data were collected from studies of the distribution of *SMN2* copy numbers in SMA patients who lack both copies of *SMN1*[11,21,26–28] (Table 2), and the distribution of the *SMN1* and *SMN2* copy numbers in the general

Table 1 *SMN1* copy number distributions among individuals without a family history of SMA in the general population

Authors	*Country*	*SMN1 Copy Number*				*Total*
		1	*2*	*3*	*4*	
Anhuf *et al*[24]	Germany	4	95	1	0	100
Corcia *et al*[22]	France	3	161	3	0	167
Cusin *et al*[23]	France	11	313	48	3	375
Feldkötter *et al*[21]	Germany	4	132	3	1	140
McAndrew *et al*[9]	USA	1	50	3	0	54
Ogino *et al*[18]	USA	3	174	12	2	191
This study	USA	4	155	16	1	176
Total		30	1080	86	7	1203

Table 2 *SMN2* copy number distributions in SMA patients lacking *SMN1*

		Type I				*Type* II				*Type III*				
Authors	*Country*	*One copy*	*Two copies*	*Three copies*	*Subtotal*	*Two copies*	*Three copies*	*Four copies*	*Subtotal*	*Two copies*	*Three copies*	*Four copies*	*Subtotal*	*Total*
Feldkötter *et al*[21]	Germany	13	138	37	188	12	90	8	110	3	39	35	77	375
Gérard *et al*[26]	France	0	5	5	10	0	4	0	4	0	1	2	3	17
Mailman *et al*[27]	USA	7	43	2	52	—	—	—	—	0	70	20	90	142
Ogino *et al*[11]	USA	3	13	0	16	1	6	1	8	0	3	1	4	28
Vitali *et al*[28]	Italy	0	3	2	5	0	7	0	7	0	1	4	5	17
Total		23	202	46	271	13	107	9	129	3	114	62	179	579

Table 3 Combined *SMN1* and *SMN2* copy number distributions in the general population

		SMN Genotype designated as '(SMN1 copy number) : (SMN2 copy number)'																
Authors	*Country*	*4:0*	*4:1*	*4:2*	*3:0*	*3:1*	*3:2*	*3:3*	*2:0*	*2:1*	*2:2*	*2:3*	*2:4*	*1:0*	*1:1*	*1:2*	*1:3*	*Total*
Anhuf *et al*[24] and personal communication	Germany	0	0	0	0	1	0	0	9	36	46	4	0	0	1	2	1	100
Cusin *et al*[23] and personal communication	France	2	0	1	5	28	14	1	24	129	152	6	2	1	3	5	2	375
McAndrew *et al*[9]	USA	0	0	0	0	3	0	0	4	22	23	1	0	0	0	0	1	54
Ogino *et al*[11]	USA	1	1	0	0	9	2	0	10	61	90	3	0	1	1	1	0	180
This study	USA	0	1	0	3	7	3	1	4	49	57	3	0	0	1	2	1	132
Total		3	2	1	8	48	19	2	51	297	368	17	2	2	6	10	5	841

population[9,11,23,24] (with V Cusin, personal communication; D Anhuf, personal communication; and this study) (Table 3).

Haplotype reconstruction and statistical analysis

Data contained in Tables 2 and 3 were inspected and systematically rewritten in terms of the haplotypes illustrated in Figure 1. Disease haplotypes in Table 2 were reconstructed with the simplest explanatory haplotypes among disease alleles. Normal haplotypes in Table 3, were reconstructed from among the simplest explanatory normal alleles. Where haplotype reconstructions were equivocal, both haplotypes were tabulated as equal probabilities.

Reconstructed haplotypes were formatted for presentation using Tools for Population Genetic Analysis TFPGA [Mark P Miller, 1997. Tools for population genetic analysis (TFPGA) 1.3: A windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by the author.] TFPGA provided both χ^2 and Fisher's exact statistics. Some frequencies were recalculated using SPSS for Windows, version 11.0.1. The reconstructed haplotypes were also compiled through submission to PHASE, version 2.0.2, written by Stephens *et al.*[29]

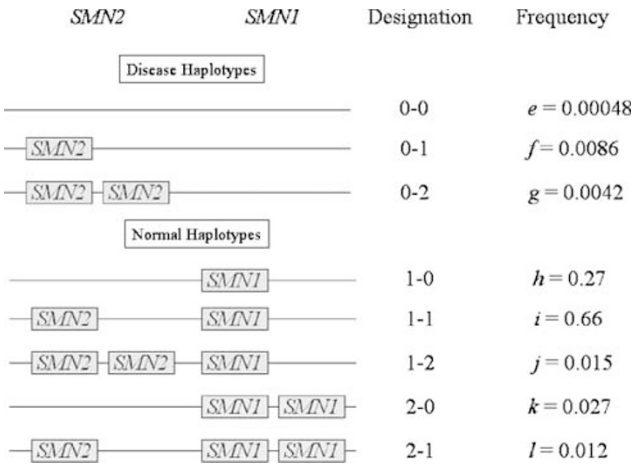

Figure 1 *SMN1-SMN2* haplotypes and frequencies.

Results

SMN1 allele frequencies

SMN1 alleles were designated as follows: most *SMN1* disease alleles are '0-copy alleles' (*SMN1* gene deletions or *SMN1*-to-*SMN2* gene conversions). A normal chromosome 5 usually has one copy of *SMN1* ('1-copy allele'). However, two copies of *SMN1* are occasionally present on a single normal chromosome 5 ('2-copy allele'). *SMN1* disease alleles with a small, intragenic mutation are referred to as '1^D alleles' (standing for '1-copy-Disease'); 1^D alleles are indistinguishable from normal 1-copy alleles by dosage analysis. The allele frequencies are designated as follows:

$a = 0$-copy (disease) allele frequency
$b = 1$-copy (normal) allele frequency
$c = 2$-copy (normal) allele frequency
$d = 1^D$ (disease) allele frequency

Asymptomatic individuals who lack both copies of *SMN1*, all of whom were relatives of affected individuals, have been described.[9,30–35] However, all 0-copy alleles are regarded as disease alleles for reasons described in detail elsewhere.[1]

Two asymptomatic individuals without a family history who have a polymorphism (9T) at the polythymidine tract (8T) in *SMN1* intron 6 (IVS6-24dupT) have been reported (one[18] and the other in this study). Thus, an observed frequency of the 9T allele among all normal *SMN1* alleles is $2/2813 = \sim 7.1 \times 10^{-4}$. Since there is no evidence indicating that 9T alleles are anything other than polymorphisms without functional consequence, they are regarded as normal alleles.

We previously calculated SMA incidence as $\sim 1.0 \times 10^{-4}$ using several large population studies.[1] However, as described in Methods, using SMA population incidence data to calculate the *SMN1* disease allele frequency is problematic, because of variable diagnostic criteria, consanguinity, and potential embryonic lethality of some severely affected fetuses. Instead, we calculated the frequencies of the various *SMN1* alleles (except for the 1^D alleles), using data on the frequencies of individuals with one, two, three, and four copies of *SMN1* in the general population[9,18,21–24] (and this study) (Table 1). Assuming Hardy-Weinberg equilibrium,

$$c^2 = \sim 7/1203 \quad \text{(for the 4}-SMN1-\text{copy individuals)} \tag{1}$$

$$2bc + 2cd = \sim 86/1203 \quad \text{(for the 3}-SMN1-\text{copy individuals)} \tag{2}$$

$$2ac + b^2 + 2bd = \sim 1080/1203 \quad \text{(for the 2}-SMN1-\text{copy individuals)} \tag{3}$$

$$2ab = \sim 30/1203 \quad \text{(for the 1}-SMN1-\text{copy individuals)} \tag{4}$$

$$\text{Because } a, c \text{ and } d \text{ are small, assume } 2ac = 2bd = 2cd = 0 \text{ Then, from (3), } b = 0.947. \tag{5}$$

$$\text{From (2), (4) and (5), } c = 0.0377, a = 0.0132, \text{ and } 2ac = 0.000993. \tag{6}$$

From (3) and (6), $b = 0.947$. After reiterative calculations, $a = 0.0132$, $b = 0.947$, $c = 0.0377$.

Among patients with SMA of all three types who had identifiable mutations in both alleles of *SMN1*, Wirth *et al*[25] found that 18 of 501 individuals (with a total of 1002 alleles) had one 0-copy allele and one 1^D allele. These data imply that the ratio of a to d is approximately 984/18. Thus, $d = 0.0132 \times 18/984 = 0.000241$. If a, b, c, and d are proportionally adjusted so that their sum equals 1, then, a (the 0-copy disease allele frequency) $= 0.0132$, b (the 1-copy normal allele frequency) $= 0.949$, c (the 2-copy normal allele frequency) $= 0.0378$, and d (the 1^D disease allele frequency) $= 0.000241$ (Table 4). Studies to determine *SMN1* copy numbers among carriers showed that a total of 19 of 494 carriers with a deletion/conversion mutation of *SMN1* on one chromosome have two copies of *SMN1*,[18,21,23,36] implying that 19 (3.8%) of 494 normal alleles are 2-copy alleles, which exactly matches the above *SMN1* allele frequencies $\{c/(b+c) = 3.8\%\}$. The disease frequency derived from our data $\{(a+d)^2 = 1.8 \times 10^{-4}\}$ is considerably higher than the disease frequency ($\sim 1.0 \times 10^{-4}$) derived from large population studies[1] (also, see On-line Supplemental Material). This may be due in part to variable diagnostic criteria for inclusion of SMA cases and embryonic lethality in some severely affected fetuses.

When an asymptomatic family member of individuals affected with clinically typical SMA of known type has two or three copies of *SMN1*, but no genetic testing has been

Table 4 *SMN1* allele and genotype frequencies

SMN1 allele			*Designation*	*General population*	*Type I*[a]	*Type II*[a]	*Type III*[a]
'0-copy' (disease)			*a*	1.32×10^{-2}	1.33×10^{-2}	1.32×10^{-2}	1.30×10^{-2}
'1-copy' (normal)			*b*	9.49×10^{-1}	9.49×10^{-1}	9.49×10^{-1}	9.49×10^{-1}
'2-copy' (normal)			*c*	3.78×10^{-2}	3.78×10^{-2}	3.78×10^{-2}	3.78×10^{-2}
'1^D-copy' (disease)			*d*	2.41×10^{-4}	1.52×10^{-4}	2.78×10^{-4}	4.09×10^{-4}
Status	*Genotype*[b]	*Copy Number*	*Designation*	*General population*	*Type I*[a]	*Type II*[a]	*Type III*[a]
Non-carrier	'2+2'	4	c^2	1.43×10^{-3}	1.43×10^{-3}	1.43×10^{-3}	1.43×10^{-3}
	'2+1'	3	*2bc*	7.18×10^{-2}	7.18×10^{-2}	7.18×10^{-2}	7.18×10^{-2}
	'1+1'	2	b^2	9.00×10^{-1}	9.00×10^{-1}	9.00×10^{-1}	9.00×10^{-1}
Carrier	'2+1^D'	3	*2cd*	1.83×10^{-5}	1.15×10^{-5}	2.10×10^{-5}	3.09×10^{-5}
	'2+0'	2	*2ac*	9.98×10^{-4}	1.00×10^{-3}	9.95×10^{-4}	9.85×10^{-4}
	'1+1^D'	2	*2bd*	4.58×10^{-4}	2.89×10^{-4}	5.27×10^{-4}	7.76×10^{-4}
	'1+0'	1	*2ab*	2.50×10^{-2}	2.52×10^{-2}	2.50×10^{-2}	2.47×10^{-2}
Affected	'1^D+1^D'	2	d^2	5.82×10^{-8}	2.31×10^{-8}	7.70×10^{-8}	1.67×10^{-7}
	'1^D+0'	1	*2ad*	6.37×10^{-6}	4.04×10^{-6}	7.30×10^{-6}	1.06×10^{-5}
	'0+0'	0	a^2	1.74×10^{-4}	1.76×10^{-4}	1.73×10^{-4}	1.70×10^{-4}

[a]Relative allele frequencies for each SMA type, which should be used when the risk of a 1^D allele is present due to an index case affected with clinically typical SMA of a given type.
[b]Genotype is indicated by '(*SMN1* allele on one chromosome 5)+(*SMN1* allele on the other chromosome 5).'

performed on the affected individual, the type of SMA in the patient has an impact on risk assessment.[1] This is because the fraction of 1^D alleles among all disease alleles changes according to the type of SMA.[25] Thus, using the data of Wirth *et al*,[25] we calculated relative 1^D allele and 0-copy allele frequencies for each SMA type (Table 4), but only for the purpose of risk assessment, and not for the calculations of the true *SMN1* allele frequencies in the general population.

The observed frequency of 4-copy individuals in the general population, $7/1203 = 5.81 \times 10^{-3}$, is significantly higher than the expected frequency of $c^2 = 1.43 \times 10^{-3}$ ($P < 0.005$ by χ^2-test), suggesting that '3-copy alleles' (ie, one chromosome with 3 copies of *SMN1*) might exist. Assuming c' = the '3-copy allele' frequency, $c^2 + 2bc' = 5.81 \times 10^{-3}$. Thus, $c' = 2.30 \times 10^{-3}$, which is approximately one order of magnitude lower than $\boldsymbol{c}$. Because the existence of the '3-copy allele' remains unproven, and because, even if it exists, its frequency would be very small relative to other normal alleles, the *SMN1* allele frequencies in Table 4 were determined on the assumption that '3-copy alleles' do not exist.

SMN1 de novo mutation rates

A total of 12 independent *SMN1 de novo* mutations (among 494 patients lacking *SMN1*) have been described. Wirth *et al*[13] reported six paternal *de novo* deletions and one maternal *de novo* conversion[25] among 340 patients lacking *SMN1*. Ogino *et al*[1] reported two paternal *de novo* deletions among 53 patients. Cusin *et al*[23] reported one paternal *de novo* deletion, one paternal *de novo* conversion, and one maternal *de novo* deletion among 101 patients. Using the methods of Wirth *et al*,[13] we estimated paternal and maternal *de novo* mutation rates to be $v = 2.11 \times 10^{-4}$ and $\mu = 4.15 \times 10^{-5}$, respectively. Estimated rates of paternal and maternal *de novo* conversion mutations (based on one event for each) are the same: 2.07×10^{-5}. These figures are based on small numbers of patients with *de novo* mutations; further studies are necessary to calculate *de novo* mutation rates more accurately.

SMN1-SMN2 haplotype frequencies

SMN1-SMN2 haplotypes are designated as '(*SMN1* copy number)-(*SMN2* copy number)' (Figure 1). We did not take into account *SMN1-SMN2* haplotypes with small intragenic *SMN1* mutations, which are rare, due to insufficient published data. To determine *SMN1-SMN2* haplotype frequencies, we used published data from several studies, and our own data, on *SMN1* and *SMN2* copy number distributions in SMA patients and individuals in the general population (Tables 2 and 3). Typically, haplotypes are determined by linkage studies. However, it is often difficult to determine the distribution of *SMN1* and *SMN2* copies between the two chromosome 5s in a given individual, even after linkage analysis. Instead, to estimate the *SMN1-SMN2* haplotype frequencies, we assumed Hardy–Weinberg equilibrium and constructed mathematical formulae using the observed *SMN1* and *SMN2* copy number distributions. The haplotype frequencies are designated as follows ($a = e + f + g$, $b = h + i + j$, and $c = k + l$):

e = 0-0 (disease) haplotype frequency
f = 0-1 (disease) haplotype frequency
g = 0-2 (disease) haplotype frequency
h = 1-0 (normal) haplotype frequency

$i = 1\text{-}1$ (normal) haplotype frequency
$j = 1\text{-}2$ (normal) haplotype frequency
$k = 2\text{-}0$ (normal) haplotype frequency
$l = 2\text{-}1$ (normal) haplotype frequency

To decrease errors associated with small numbers of examples, we calculated the disease haplotype frequencies (e, f, and g) using the *SMN1:SMN2* copy number distribution data in patients (Table 2), rather than the *SMN1:SMN2* copy number distribution data in normal populations (Table 3). Using the data in the On-line Supplemental Material, it is estimated that type I, type II, and type III comprise ~57.5, ~29.2, and ~13.2%, respectively, of all typical SMA cases. The data in Table 2 do not take into account the incidence of each SMA type. Hence, to calculate e, f, and g in the general population, the data in Table 2 need to be adjusted according to the incidence of type II and type III SMA relative to that of type I. For type II SMA, the total number of patients should be adjusted to $271 \times (0.292/0.575) = 137.6$, with 13.9 two-*SMN2*-copy individuals, 114.1 three-*SMN2*-copy individuals, and 9.6 four-*SMN2*-copy individuals. For type III, the total number of patients should be adjusted to $271 \times (0.132/0.575) = 62.2$, with 1.04 two-copy individuals, 39.6 three-copy individuals, and 21.5 four-copy individuals. Then, the adjusted number (fraction) of SMA patients (total $N = 470.8$) with one, two, three, and four copies of *SMN2* are 23 (0.0488), 216.9 (0.461), 199.8 (0.424), and 31.2 (0.0662), respectively. The incidence per live birth of all types of SMA is defined as m ($= \sim 1.80 \times 10^{-4}$). Assuming Hardy–Weinberg equilibrium [a genotype is designated as '(total *SMN1* copy number) : (total *SMN2* copy number)'],

$$2ef = \sim 0.0488\,m \quad \text{(for the 0 : 1 genotype)} \tag{7}$$

$$2eg + f^2 = \sim 0.461\,m \quad \text{(for the 0 : 2 genotype)} \tag{8}$$

$$2fg = \sim 0.424\,m \quad \text{(for the 0 : 3 genotype)} \tag{9}$$

$$g^2 = \sim 0.0662\,m \quad \text{(for the 0 : 4 genotype)} \tag{10}$$

From (10), $g = 3.45 \times 10^{-3}$. Then, from (9), $f = 1.11 \times 10^{-2}$, and from (7), $e = 3.96 \times 10^{-4}$. Because this value for g is based on a relatively small number of individuals, the above value for f also has a substantial error. Therefore, the formula (8) should be used to calculate f; then, from (8), $f = 8.96 \times 10^{-3}$. Then, from (9), $g = 4.26 \times 10^{-3}$. Then, from (7), $e = 4.90 \times 10^{-4}$. After iterative calculations; $f = 8.87 \times 10^{-3}$, $g = 4.30 \times 10^{-3}$, and $e = 4.95 \times 10^{-4}$. e, f, and g are proportionally adjusted so that their sum equals a; f (the 0-1 haplotype frequency) $= 8.6 \times 10^{-3}$, g (the 0-2 haplotype frequency) $= 4.2 \times 10^{-3}$, and e (the 0-0 haplotype frequency) $= 4.8 \times 10^{-4}$. These results indicate that the 0-1 haplotype, the 0-2 haplotype, and the 0-0 haplotype comprise ~65, ~31, and ~3.6%, respectively, of disease haplotypes with 0 copies of *SMN1*.

Because a knockout of the murine *SMN* gene is associated with embryonic lethality,[37] and because humans with the 0:0 genotype have never been observed, the expression of some full-length SMN protein seems to be essential for survival. Hence, by starting with data on individuals affected with SMA for our calculations, we are potentially introducing an ascertainment bias against the 0-0 haplotype, since we would never expect to see the 0:0 genotype among living individuals. However, the 0-0 haplotype is sufficiently rare that the existence of the 0:0 genotype can be ignored, and the far more common 0:1 genotype can be used, to estimate e. Using the data in Table 3, and assuming Hardy-Weinberg equilibrium,

$$2eh = \sim 2/841 \quad \text{(for the 1 : 0 genotype)} \tag{11}$$

$$2ei + 2fh = \sim 6/841 \quad \text{(for the 1 : 1 genotype)} \tag{12}$$

$$2ej + 2fi + 2gh = \sim 10/841 \quad \text{(for the 1 : 2 genotype)} \tag{13}$$

$$2fj + 2gi = \sim 5/841 \quad \text{(for the 1 : 3 genotype)} \tag{14}$$

$$2gj = \sim 0/841 \quad \text{(for the 1 : 4 genotype)} \tag{15}$$

$$h^2 + 2ek = \sim 51/841 \quad \text{(for the 2 : 0 genotype)} \tag{16}$$

$$2hi + 2el + 2fk = \sim 297/841 \quad \text{(for the 2 : 1 genotype)} \tag{17}$$

$$i^2 + 2hj + 2fl + 2gk = \sim 368/841 \quad \text{(for the 2 : 2 genotype)} \tag{18}$$

$$2ij + 2gl = \sim 17/841 \quad \text{(for the 2 : 3 genotype)} \tag{19}$$

$$j^2 = \sim 2/841 \quad \text{(for the 2 : 4 genotype)} \tag{20}$$

$$2hk = \sim 8/841 \quad \text{(for the 3 : 0 genotype)} \tag{21}$$

$$2ik + 2hl = \sim 48/841 \quad \text{(for the 3 : 1 genotype)} \tag{22}$$

$$2il + 2jk = \sim 19/841 \quad \text{(for the 3 : 2 genotype)} \tag{23}$$

$$2jl = \sim 2/841 \quad \text{(for the 3 : 3 genotype)} \tag{24}$$

$$k^2 = \sim 3/841 \quad \text{(for the 4 : 0 genotype)} \tag{25}$$

$$2kl = \sim 2/841 \quad \text{(fo the 4 : 1 genotype)} \tag{26}$$

$$l^2 = \sim 1/841 \quad \text{(for the 4 : 2 genotype)} \tag{27}$$

To decrease errors associated with small numbers of subjects, we used equations containing larger numerators (ie, formulae (16) through (19) and (21) through (23)). Likewise, to calculate e, f, and g, we used population data based on large numbers of SMA patients (ie, formulae (7) through (10)), rather than small numbers of carrier individuals (ie, formulae (11) through (15)). The potential

existence of a haplotype with three copies of *SMN1* might compromise the accuracy of estimates based primarily on formulae (25) through (27), and, to lesser degrees, on formulae (21) through (24). Thus, we did not primarily use formulae (25) through (27).

$$\text{Because } j, k, \text{ and } l \text{ are small (from formulae (20), (25) and (27)), and } e, f, \text{ and } g \text{ are small (see above), we assumed that any products of two of these are close to 0.} \quad (28)$$

$$\text{From (16) and (28)}, h^2 = 51/841.\ h = 0.246 \quad (29)$$

$$\text{From (17), (28) and (29)}, i = 0.717 \quad (30)$$

$$\text{From (23), (28) and (30)}, l = 0.0157 \quad (31)$$

$$\text{From (22), (30), (29) and (31)}, k = 0.0344 \quad (32)$$

$$\text{From (9), (19), (30) and (31)}, j = 0.0140 \quad (33)$$

With the above values for *j*, *k* and *l*, the values for *h*, *i*, *j*, *k* and *l* can be refined further.

$$\text{From (8), (9), (18), (29), (31), (32) and (33)}, i = 0.656 \quad (34)$$

$$\text{From (8), (10), (17), (31), (32) and (34)}, h = 0.269 \quad (35)$$

$$\text{From (23), (32), (33) and (34)}, l = 0.0165 \quad (36)$$

$$\text{From (22), (34), (35) and (36)}, k = 0.0368 \quad (37)$$

$$\text{From (9), (19), (34) and (36)}, j = 0.0153 \quad (38)$$

Because $b = h + i + j$, we adjusted the values for *h*, *i* and *j* proportionately: *h* (the 1-0 haplotype frequency) = 0.27, *i* (the 1-1 haplotype frequency) = 0.66, and *j* (the 1-2 haplotype frequency) = 0.015. Likewise, because $c = k + l$, we adjusted the values for *k* and *l* proportionately: *k* (the 2-0 haplotype frequency) = 0.027, and *l* (the 2-1 haplotype frequency) = 0.012. Using the above haplotype frequency estimates, and assuming Hardy–Weinberg equilibrium, the expected number of individuals with each genotype among the 841 individuals is in agreement with the raw population data (Table 3). The slightly higher-than-expected frequency of individuals with four copies of *SMN1* may be due to the presence of '3-copy-*SMN1* alleles,' which we disregarded in our calculations of *SMN1*-allele and *SMN*-haplotype frequencies.

To verify our allele-frequency and haplotype calculations, we composed a simulation that reconstructed the haplotypes in the affected and normal population using the data presented in Tables 2 and 3, as described in the Methods section. We used both a χ^2 and Fisher's exact tests to determine compliance of the simulation data set with Hardy–Weinberg equilibrium. This simulation provided the insight that three populations[9,24] (and this study) of the six normal populations examined for *SMN* haplotype analysis in Table 3 were consistent with a Hardy-Weinberg distribution. The remainder of the six normal populations showed increased numbers of the 1-1 haplotype, causing a distortion from Hardy–Weinberg equilibrium. Analysis of the haplotype reconstructions and simulations from the disease allele data set provided evidence for a relative two-fold increase in the frequency of the 0-2 haplotype among affected individuals as compared with calculations assuming Hardy–Weinberg equlibrium carried out above.

Discussion

To test our assumption of Hardy–Weinberg equilibrium in calculating allele and haplotype frequencies, we carried out a simulation as described in Methods and Results. Overall, the simulation demonstrated some evidence for distortion from Hardy–Weinberg equilibrium among the normal alleles. The distortions from Hardy–Weinberg equilibrium could derive from the heterogeneity of populations examined or because of selective advantage for some alleles. We believe that our simulations support the additional evidence delineated below for selection at this complex locus.

Although functional evidence for selection of the 1-1 haplotype is lacking, it is the most common haplotype in *Homo sapiens*. Because the paralogous duplication giving rise to *SMN2* occurred after human–primate divergence,[41] it will be difficult to provide phylogenetic proof of selection. However, the young age of this haplotype structure, accompanied by elevated haplotype frequency, suggests selection. Our simulation data reinforces over-representation of the 1-1 haplotype in the human gene pool. The availability of a HapMap should enable coalescent modeling studies using polymorphisms in the region and testing numerous human populations for elevated haplotype frequencies.[38] This type of analysis should provide definitive evidence for or against selection of the 1-1 haplotype in human populations.

It is likely that all of the 0-0 haplotypes, and most, if not all, of the 0-1 haplotypes arose by *de novo SMN1* deletion mutations. On the other hand, it is likely that most of the 0-2 haplotypes, and perhaps some of the 0-1 haplotypes, arose by *de novo SMN1*-to-*SMN2* gene conversion mutations. Although some SMA type III patients reproduce, most of the SMA disease haplotypes, including the 0-2 haplotype, are genetically lethal when homozygous, or when heterozygous with another disease haplotype. Thus, disease haplotypes must be replenished constantly in the general population by *de novo* mutations. The available data on SMA patients who received *de novo* mutations (among all SMA patients) are limited, comprising only 12 such patients. However, two of these patients received *de*

novo SMN1-to-*SMN2* conversion mutations.[23,25] Other small intragenic *SMN1* mutations (present in 1.8% of all disease alleles) are relatively rare. Hence, the 31% frequency of the 0-2 haplotype among all disease haplotypes implies that *SMN1*-to-*SMN2* gene conversion (i.e., a single nucleotide substitution at position 840) occurs at a much higher frequency (approximately 10-fold higher) than all other small intragenic mutations combined. Thus, position 840 might be a mutation hotspot. A mutation rate of 1.2×10^{-9} mutations/site per year has been suggested for *Homo sapiens*.[39,40] This number agrees well with a more-recent direct estimate of the frequency of all kinds of small mutations (1.8×10^{-8}/nucleotide per generation[42]) as well as an indirect, evolutionary estimate (2.5×10^{-8}/nucleotide per generation[43]). Single nucleotide substitutions are $\sim$25 times more common than all other small mutations.[41] The observed frequency of *SMN1*-to-*SMN2* gene conversion is 2.1×10^{-5} per generation for both men and women, which is approximately three orders of magnitude higher than the average mutation frequency estimates above. Rochette *et al*[41] suggested that gene conversion between *SMN1* and *SMN2* is facilitated by the 99.8% sequence identity between the two genes, and the fact that they are confined to a region of about 1 Mb. Alternative hypotheses for the relatively high frequency of the 0-2 haplotype include a selective advantage for the 0-2 haplotype over other disease haplotypes, that is, some mildly affected (eg, so-called type IV) individuals homozygous for the 0-2 haplotype may reproduce. However, the vast majority of SMA patients with four copies of *SMN2* are affected by type III SMA, which is generally pediatric in onset and genetically lethal. Type IV SMA is too rare to account fully for the high frequency of the 0-2 haplotype.

The *SMN1* disease allele frequency estimated from compiled large population incidence data (in On-line Supplemental Material) is $\sim$0.010, implying that the SMA carrier frequency is $\sim$1/50 under Hardy–Weinberg equilibrium. However, the carrier frequency estimated from our allele frequency data is $\sim$1/38 $\{=2\ (a+d)\ (b+c)\}$. This discrepancy may be attributed to embryonic lethality of some severely affected fetuses, exclusion of some SMA cases with atypical features, and statistical variation.

We recommend that genetic counsellors use our allele and haplotype frequency estimates for risk calculations, as our estimates are the best available at this time. The take home message of this article is that the *SMN* region in 5q is unstable and in active flux (based on the high *de novo* mutation rate and the recent evolutionary basis for the duplication), and that geneticists and genetic counsellors must keep informed of changes in genomic sequence information (because the assembly is still inaccurate), *de novo* mutation rates (due to the small numbers of cases on which our estimates depend), and allele and haplotype frequencies (because the numbers of cases on which our estimates depend will increase over time, thereby improving accuracy).

In conclusion, the *SMN1* allele frequencies (Table 4) and *de novo* mutation rates are based on all available and reliable data, and can be used for genetic counseling and accurate risk assessments for SMA patients and their families. An unusually high frequency of *de novo* deletion/conversion mutations occur in the *SMN1* gene, a substantial fraction of which seem to be *de novo* gene conversions from *SMN1* to *SMN2* (ie, *de novo* 840C>T mutation), suggesting that the nucleotide position 840 might be a mutation hotspot. Additional studies will further shed light on the molecular evolution of the *SMN1* and *SMN2* genes.

Acknowledgements

We thank H Rennert, M Dean and C O'hUigin for helpful discussions; and DGB Leonard, V Van Deerlin, and CS Fuchs for supporting various aspects of the work. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This work has been funded in whole or in part with Federal Funds from the National Cancer Institute and the National Institutes of Health.

References

1 Ogino S, Leonard DG, Rennert H, Ewens WJ, Wilson RB: Genetic risk assessment in carrier testing for spinal muscular atrophy. *Am J Med Genet* 2002; **110**: 301–307.
2 Lefebvre S, Burglen L, Reboullet S *et al*: Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 1995; **80**: 155–165.
3 Burglen L, Lefebvre S, Clermont O *et al*: Structure and organization of the human survival motor neurone (*SMN*) gene. *Genomics* 1996; **32**: 479–482.
4 Monani UR, Lorson CL, Parsons DW *et al*: A single nucleotide difference that alters splicing patterns distinguishes the SMA gene *SMN1* from the copy gene *SMN2*. *Hum Mol Genet* 1999; **8**: 1177–1183.
5 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 USA* 1999; **96**: 6307–6311.
6 Kashima T, Manley JL: A negative element in *SMN2* exon 7 inhibits splicing in spinal muscular atrophy. *Nat Genet* 2003; **34**: 460–463.
7 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**: 228–237.
8 Ogino S, Wilson RB: Spinal muscular atrophy: molecular genetics and diagnostics. *Expert Rev Mol Diagn* 2004; **4**: 15–29.
9 McAndrew PE, Parsons DW, Simard LR *et al*: Identification of proximal spinal muscular atrophy carriers and patients by analysis of SMNT and SMNC gene copy number. *Am J Hum Genet* 1997; **60**: 1411–1422.
10 Chen KL, Wang YL, Rennert H *et al*: Duplications and de novo deletions of the SMNt gene demonstrated by fluorescence-based carrier testing for spinal muscular atrophy. *Am J Med Genet* 1999; **85**: 463–469.
11 Ogino S, Gao S, Leonard DG, 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**: 275–277, (see Addendum in Vol 11; 723).

12 Mailman MD, Hemingway T, Darsey RL *et al*: Hybrids monosomal for human chromosome 5 reveal the presence of a spinal muscular atrophy (SMA) carrier with two *SMN1* copies on one chromosome. *Hum Genet* 2001; **108**: 109–115.
13 Wirth B, Schmidt T, Hahnen E *et al*: *De novo* rearrangements found in 2% of index patients with spinal muscular atrophy: mutational mechanisms, parental origin, mutation rate, and implications for genetic counseling. *Am J Hum Genet* 1997; **61**: 1102–1111.
14 Eichler EE, Sankoff D: Structural dynamics of eukaryotic chromosome evolution. *Science* 2003; **301**: 793–797.
15 Cheung J, Estivill X, Khaja R *et al*: Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol* 2003; **4**: R25.
16 Ogino S, Wilson RB: Genetic testing and risk assessment for spinal muscular atrophy (SMA). *Hum Genet* 2002; **111**: 477–500.
17 Ogino S, Wilson RB: SMN dosage analysis and risk assessment for spinal muscular atrophy. *Am J Hum Genet* 2002; **70**: 1596–1598; discussion 1598–1599.
18 Ogino S, Leonard DG, Rennert H, Wilson RB: Spinal muscular atrophy genetic testing experience at an academic medical center. *J Mol Diagn* 2002; **4**: 53–58.
19 Ogino S, Wilson RB: Quantification of PCR bias caused by a single nucleotide polymorphism in SMN gene dosage analysis. *J Mol Diagn* 2002; **4**: 185–190.
20 Ogino S, Leonard DG, Rennert H, Gao S, Wilson RB: Heteroduplex formation in SMN gene dosage analysis. *J Mol Diagn* 2001; **3**: 150–157.
21 Feldkotter 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**: 358–368.
22 Corcia P, Mayeux-Portas V, Khoris J *et al*: Abnormal SMN1 gene copy number is a susceptibility factor for amyotrophic lateral sclerosis. *Ann Neurol* 2002; **51**: 243–246.
23 Cusin V, Clermont O, Gerard B, Chantereau D, Elion J: Prevalence of SMN1 deletion and duplication in carrier and normal populations: implication for genetic counselling. *J Med Genet* 2003; **40**: E39.
24 Anhuf D, Eggermann T, Rudnik-Schoneborn S, Zerres K: Determination of *SMN1* and *SMN2* copy number using TaqMan™ technology. *Hum Mutat* 2003; **22**: 74–78.
25 Wirth B, Herz M, Wetter A *et al*: Quantitative analysis of survival motor neuron copies: identification of subtle *SMN1* mutations in patients with spinal muscular atrophy, genotype–phenotype correlation, and implications for genetic counseling. *Am J Hum Genet* 1999; **64**: 1340–1356.
26 Gerard B, Ginet N, Matthijs G *et al*: Genotype determination at the survival motor neuron locus in a normal population and SMA carriers using competitive PCR and primer extension. *Hum Mutat* 2000; **16**: 253–263.
27 Mailman MD, Heinz JW, Papp AC *et al*: Molecular analysis of spinal muscular atrophy and modification of the phenotype by *SMN2*. *Genet Med* 2002; **4**: 20–26.
28 Vitali T, Sossi V, Tiziano F *et al*: Detection of the survival motor neuron (*SMN*) genes by FISH: further evidence for a role for *SMN2* in the modulation of disease severity in SMA patients. *Hum Mol Genet* 1999; **8**: 2525–2532.
29 Stephens M, Smith NJ, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.
30 Brahe C, Servidei S, Zappata S *et al*: Genetic homogeneity between childhood-onset and adult-onset autosomal recessive spinal muscular atrophy. *Lancet* 1995; **346**: 741–742.
31 Cobben JM, van der Steege G, Grootscholten P *et al*: Deletions of the survival motor neuron gene in unaffected siblings of patients with spinal muscular atrophy. *Am J Hum Genet* 1995; **57**: 805–808.
32 DiDonato CJ, Ingraham SE, Mendell JR *et al*: Deletion and conversion in spinal muscular atrophy patients: is there a relationship to severity? *Ann Neurol* 1997; **41**: 230–237.
33 Hahnen E, Forkert R, Marke C *et al*: Molecular analysis of candidate genes on chromosome 5q13 in autosomal recessive spinal muscular atrophy: evidence of homozygous deletions of the SMN gene in unaffected individuals. *Hum Mol Genet* 1995; **4**: 1927–1933.
34 Somerville MJ, Hunter AG, Aubry HL *et al*: Clinical application of the molecular diagnosis of spinal muscular atrophy: deletions of neuronal apoptosis inhibitor protein and survival motor neuron genes. *Am J Med Genet* 1997; **69**: 159–165.
35 Wang CH, Xu J, Carter TA *et al*: Characterization of survival motor neuron (*SMNT*) gene deletions in asymptomatic carriers of spinal muscular atrophy. *Hum Mol Genet* 1996; **5**: 359–365.
36 Cusco II, Barcelo MJ, Baiget M, Tizzano EF: Implementation of SMA carrier testing in genetic laboratories: comparison of two methods for quantifying the *SMN1* gene. *Hum Mutat* 2002; **20**: 452–459.
37 Monani UR, Coovert DD, Burghes AH: Animal models of spinal muscular atrophy. *Hum Mol Genet* 2000; **9**: 2451–2457.
38 Sabeti PC, Reich DE, Higgins JM *et al*: Detecting recent positive selection in the human genome from haplotype structure. *Nature* 2002; **419**: 832–837.
39 Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D: Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol* 1996; **5**: 182–187.
40 Gibbons A: When it comes to evolution, humans are in the slow class. *Science* 1995; **267**: 1907–1908.
41 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**: 255–266.
42 Kondrashov AS: Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* 2003; **21**: 12–27.
43 Nachman MW, Crowell SL: Estimate of the mutation rate per nucleotide in humans. *Genetics* 2000; **156**: 297–304.

(Supplementary information accompanies the paper on European Journal of Human Genetics website http://www.nature.com/ejhg).