RESEARCH ARTICLE OPEN ACCESS

Analysis of Structural Variants Previously Associated With ALS in Europeans Highlights Genomic Architectural Differences in Africans

Nomakhosazana R. Monnakgotla, MSc, Amokelani C. Mahungu, MSc, Jeannine M. Heckmann, MBChB, PhD, Gerrit Botha, MSc, Nicola J. Mulder, BSc, PhD, Gang Wu, PhD, Evadnie Rampersaud, BSc, PhD, Jason Myers, MS, Marka Van Blitterswijk, MD, PhD, Rosa Rademakers, PhD, J. Paul Taylor, MD, PhD, Joanne Wuu, ScM, Michael Benatar, MD, PhD, and Melissa Nel, MBChB, PhD

Correspondence
Dr. Nel
melissa.nel@uct.ac.za

Neurol Genet 2023;9:e200077. doi:10.1212/NXG.0000000000200077

Abstract

Background and Objectives

Amyotrophic lateral sclerosis (ALS) is a degenerative condition of the brain and spinal cord in which protein-coding variants in known ALS disease genes explain a minority of sporadic cases. There is a growing interest in the role of noncoding structural variants (SVs) as ALS risk variants or genetic modifiers of ALS phenotype. In small European samples, specific short SV alleles in noncoding regulatory regions of *SCAF4*, *SQSTM1*, and *STMN2* have been reported to be associated with ALS, and several groups have investigated the possible role of *SMN1*/*SMN2* gene copy numbers in ALS susceptibility and clinical severity.

Methods

Using short-read whole genome sequencing (WGS) data, we investigated putative ALS-susceptibility *SCAF4* (3′UTR poly-T repeat), *SQSTM1* (intron 5 AAAC insertion), and *STMN2* (intron 3 CA repeat) alleles in African ancestry patients with ALS and described the architecture of the *SMN1*/*SMN2* gene region. South African cases with ALS (n = 114) were compared with ancestry-matched controls (n = 150), 1000 Genomes Project samples (n = 2,336), and H3Africa Genotyping Chip Project samples (n = 347).

Results

There was no association with previously reported *SCAF4* poly-T repeat, *SQSTM1* AAAC insertion, and long *STMN2* CA alleles with ALS risk in South Africans ($p > 0.2$). Similarly, *SMN1* and *SMN2* gene copy numbers did not differ between South Africans with ALS and matched population controls ($p > 0.9$). Notably, 20% of the African samples in this study had no *SMN2* gene copies, which is a higher frequency than that reported in Europeans (approximately 7%).

Discussion

We did not replicate the reported association of *SCAF4*, *SQSTM1*, and *STMN2* short SVs with ALS in a small South African sample. In addition, we found no link between *SMN1* and *SMN2* copy numbers and susceptibility to ALS in this South African sample, which is similar to the conclusion of a recent meta-analysis of European studies. However, the *SMN* gene region findings in Africans replicate previous results from East and West Africa and highlight the importance of including diverse population groups in disease gene discovery efforts. The clinically relevant differences in the *SMN* gene architecture between African and non-African populations may affect the effectiveness of targeted *SMN2* gene therapy for related diseases such as spinal muscular atrophy.

From the Neurology Research Group (N.R.M., A.C.M., J.M.H., M.N.), Division of Neurology, Department of Medicine; Neuroscience Institute (N.R.M., A.C.M., J.M.H., M.N.); Computational Biology Division (G.B., N.J.M.), Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa; Center for Applied Bioinformatics (G.W., E.R., J.M.), St. Jude Children's Research Hospital, Memphis, TN; Department of Neuroscience (M.V.B.), Mayo Clinic, Jacksonville, FL; Center for Molecular Neurology (R.R.), University of Antwerp, Belgium; Department of Cell and Molecular Biology (J.P.T.), St. Jude Children's Research Hospital, Memphis, TN; and Department of Neurology (J.W., M.B.), University of Miami, FL.

Go to Neurology.org/NG for full disclosures. Funding information is provided at the end of the article.

The Article Processing Charge was funded by Carnegie Corporation of New York.

Glossary

ALS = amyotrophic lateral sclerosis; ***C9orf72*** = Chromosome 9 Open Reading Frame 72 gene; **CNVs** = copy number variants; **EGA** = European Genome-Phenome Archive; **GATK** = Genome Analysis Toolkit; **GWAS** = genome-wide association study; **IQR** = interquartile range; **LMN** = lower motor neuron; **MLPA** = multiplex ligation-dependent probe amplification; ***NEK1*** = NIMA-Related Kinase 1 gene; **PCA** = principal component analysis; **PCR** = polymerase chain reaction; **SAB** = South African Black; ***SAC*** = South African Coloured; ***SCAF4*** = SR-Related CTD-Associated Factor 4 gene; **SMA** = spinal muscular atrophy; ***SMN1*** = Survival of Motor Neuron 1 gene; ***SMN2*** = Survival of Motor Neuron 2 gene; **SNV** = single-nucleotide variant; ***SOD1*** = Superoxide Dismutase 1 gene; ***SQSTM1*** = Sequestosome 1 gene; **SSVs** = short structural variants; ***STMN2*** = Stathmin 2 gene; **SVs** = structural variants; **VCF** = variant call format; **WGS** = whole-genome sequencing.

Amyotrophic lateral sclerosis (ALS) is a progressive degenerative disease of motor neurons and typically results in death within 2–5 years of symptom onset. Multiple evidence sources converge on a multistep process underlying ALS pathogenesis, involving genetic, environmental, and aging factors, which has been replicated across different population groups.[1,2] While a growing number of disease genes harboring pathogenic variants appear to be key drivers of this multistep model, such genetic factors do not explain all cases with ALS nor are they sufficient to cause ALS in all cases. Furthermore, ALS genes can also cause other diseases (pleiotropy) while patients with ALS harboring the same pathogenic gene variant can have clinically heterogeneous presentations ranging from ALS to the frontotemporal dementia phenotype in different individuals within the same family.[3] Genetic modifier variants, particularly structural variants in noncoding regulatory genomic regions, may contribute to ALS risk or modify ALS phenotype expression.[4,5]

Copy number variants (CNVs), as well as short structural variants (SSVs <50bp), such as insertions, deletions, or short tandem repeats, are much more common than single-nucleotide variants (SNV), and they are enriched on haplotypes identified by genome-wide association studies (GWAS), suggesting that they play an important role in complex diseases.[6] For SSVs, bioinformatics algorithms have been used to identify promising candidates for future study by annotating the genome-wide SSV catalog with available GWAS data.[7] This strategy has been applied to ALS by focusing on candidate SSVs in the putative regulatory regions of 2 known ALS disease genes (*SOD1* and *SQSTM1*),[8,9] as well as *STMN2*, a candidate disease gene with altered expression in ALS.[10] The association of various SSVs in these genes identified in European case control studies are summarized in Table 1. Only 1 independent study has investigated the reported ALS associated *STMN2* SSV and did not replicate an association with ALS, which highlights the need for replication studies in independent cohorts.[11]

A CNV that has been speculated to play a role in ALS, and for which a specialized bioinformatics tool makes it possible to study using whole-genome sequencing data, is the copy number of the *SMN* (survival of motor neuron) genes. These genes have been studied extensively in a mostly pediatric motor neuron disease called spinal muscular atrophy (SMA) where the homozygous deletion of *SMN1* causes SMA and the copy number of its paralog, *SMN2*, correlates with disease severity.[12] Due to the fact that lower motor neuron weakness is an essential clinical feature of both SMA and ALS and that *SMN2* gene therapy to increase SMN protein levels has entered clinical trials, the copy number state of *SMN1* and *SMN2* has been investigated as a possible risk factor and/or genetic modifier in ALS. Despite multiple studies reporting conflicting results regarding the association of *SMN1/SMN2*

Table 1 Previously Reported ALS-Associated Short Structural Variants (SSVs) Identified in European Cohorts

Gene	Variant Chromosome location (hg38) rsID	Associated allele	Sample size n individuals	Study
SCAF4	Poly-T repeat chr21:31671109-31671125 rs573116164	18 T (fALS[a])	190	Pytte, Flynn, et al., 2020
SQSTM1	AAAC insertion chr5:179830136-179830142 rs60327661	insAAAC (fALS[a])	196	Pytte, Anderton, et al., 2020
STMN2	CA repeat chr8:79641629-79641672 rs61386841	24 CA (sALS[b])	321	Theunissen et al., 2021

Abbreviations: fALS = familial amyotrophic lateral sclerosis, sALS = sporadic amyotrophic lateral sclerosis, rsID = reference SNV cluster ID.
[a] ALS cohort with a family history of disease and harboring pathogenic variants in ALS genes: *SOD1* (approximately 85%), *C9orf72* (approximately 5%), *TARDBP* (approximately 10%).
[b] ALS cohort without a family history of disease; pathogenic variant frequency unknown.

gene copy numbers and ALS,[13-16] a recent well-powered population-based meta-analysis pooling the results from various non-African ALS cohorts concluded that the *SMN1/SMN2* copy number state does not contribute to ALS susceptibility or severity.[17]

To strengthen the credibility of association studies, replicating the analysis of genotype-phenotype links in independent and diverse cohorts is necessary. In this study, we therefore sought to investigate putative European ALS-associated SSVs (in *SCAF4*, *SQSTM1*, and *STMN2*) and *SMN1/SMN2* copy numbers in African ancestry patients with ALS.

Methods

Patients With ALS

We included 114 African ancestry patients with ALS attending the ALS clinic at Groote Schuur Hospital, Cape Town, South Africa; 30 self-identified as South African Black (SAB) and 84 self-identified as South African Coloured (SAC) genetic ancestry, which was confirmed by ancestry principal component analysis (PCA), as previously described[18] (Table 2). Patients were diagnosed by a neurologist (JMH) as clinically probable or definite ALS according to the revised El Escorial criteria.[19] Thirty-four individuals in this ALS patient sample were enrolled in the CReATe Consortium's Phenotype-Genotype and Biomarker Study (PGB1).

Standard Protocol Approvals, Registrations, and Patient Consents

This study was approved by the University of Cape Town Faculty of Health Sciences Human Research Ethics Committee, and all patients provided informed written consent to participate.

Table 2 Characteristics of South African Patients With ALS Grouped According to Ancestry

Patients with ALS, n individuals	SAB 30	SAC 84
Male sex, n (%)	20 (67)	44 (52)
AOO in years, mean (IQR)	51 (45–59)	53 (43–64)
Family history of ALS	0 (0)	4[a] (5)
LMN-ALS, n (%)	0 (0)	10 (12)
Pathogenic variant frequency, n (%)		
***C9orf72*, n (%)**	1 (3)	6 (7)
***SOD1*, n (%)**	1 (3)	6[b] (7)

Abbreviations: ALS = amyotrophic lateral sclerosis; AOO = age at onset; *C9orf72* = chromosome 9 open reading frame 72 gene; IQR = interquartile range; SAB = South African Black; SAC = South African Coloured; *SOD1* = superoxide dismutase gene.
LMN-ALS refers to patients with predominantly lower motor neuron involvement (including those with flail arm/leg).
[a] Three related individuals from the same family.
[b] Two related individuals carried the same *SOD1* variant.

Population Control Data Sets

Whole-genome sequencing (WGS) data of 127 SAB and 23 SAC individuals were used as ancestry-matched controls for the ALS association analysis.[20-22] We further analyzed control samples of various ancestries from the phase 3 call set of the 1000 Genomes Project (n = 2,336)[6,23] and the H3Africa Genotyping Chip Project data set (n = 347).[24]

Whole-Genome Sequencing

DNA extraction was performed as previously described,[25] and sequencing libraries with read lengths of 100–150bp were generated using both PCR and PCR-free kits. Libraries were sequenced to a coverage of ≥30X on the BGI MGISEQ-2000 instrument or various Illumina sequencing platforms (see eMethods, links.lww.com/NXG/A612).

Read Alignment and Variant Calling

WGS FASTQ files were aligned to the NCBI GRCH38 reference genome with alt contigs using alt aware alignment, followed by joint variant calling according to the Genome Analysis Toolkit (GATK) best-practice guidelines[26] documented in github.com/grbot/varcall. *SCAF4* poly-T repeat length (region NC_000021.9:g. 31671109_31671125) and *SQSTM1* insAAAC (NC_000005.10:g.179830139_179830142dup) high-quality genotypes were extracted from the joint called VCF file after applying the following quality control filters: genotype quality ≥20, read depth ≥10, and an allele balance for the alternative allele of ≥0.2. A subset of these high-quality genotypes were visually inspected using the Integrative Genomics Viewer (IGV) to verify the accuracy of variant calling (see eMethods, links.lww.com/NXG/A612).

Determination of *STMN2* CA Repeat Alleles

ExpansionHunter v5[27] was used to determine *STMN2* CA repeat length (NC_000008.11:g.79641629_79641672) using a custom *STMN2* variant catalog.[11] *STMN2* CA repeat length genotypes with a 95% confidence interval for either allele spanning a range of repeat length sizes were excluded. A subset of these high-quality repeat length genotypes were visually inspected using the Illumina REViewer tool (github.com/Illumina/REViewer) (see eMethods, links.lww.com/NXG/A612).

Determination of *SMN* Copy Numbers

The Illumina *SMN* Copy Number Caller tool was used to determine the copy number states of *SMN1* and *SMN2*.[28] Only samples that passed the quality control metric (given by: PASS:Majority or PASS:AgreeWithSome labels) were included in the analysis. Samples that did not pass the quality control metric (given by: Ambiguous or FLCNnoCall labels) included those where 5 of 8 sites in the intron 6 to exon 8 region of both *SMN1* and *SMN2* were not found on the haplotype sequences required to make a call. *SMN* copy number calls were not validated by molecular assays such as multiplex ligation-dependent probe amplification (MLPA).

Statistical Analysis

Statistical analysis was performed using GraphPad Prism v9.4.1 (GraphPad software, San Diego, CA). The Fisher exact test was used to compare 2 categorical variables, and the χ^2 test was used to compare more than 2 categorical variables between groups.

Availability of Data and Materials

Anonymized whole-genome sequencing data from patients with ALS in this study form part of larger ongoing collaborative studies. Data generated by UCT Neurology will be released to *bone fide* researchers via the European Genome-Phenome Archive (EGA, ega-archive.org/) subject to data access committee approval after the completion of aggregate data analysis and the publication embargo period in accordance with H3Africa policy guidelines.[29] Whole-genome sequencing data from the 1000 Genomes Project Phase 3 set used in this study are available through internationalgenome.org/data-portal/datacollection/30x-grch38. The H3Africa Genotyping Chip Project data sets are available by request under the following EGA accession numbers: EGAD00001004220, EGAD00001004448, EGAD00001004393, EGAD00001004316, EGAD00001004533, EGAD00001004505, EGAD00001004334, EGAD00001004557, and EGAD00001005076. The AWI-Gen Phase 1 WGS data from 100 South Africans[22] (EGAD00001006418) and the Southern African Human Genome Programme data set[20] (EGAD00001003791) have been deposited in the EGA.

Results

Study Population

A summary of the demographic and clinical characteristics of the South African patients with ALS in this study is provided in Table 2. Ancestry principal component analysis of South African patients with ALS using genetic markers has been previously performed.[18] SAB patients cluster separately to the East and West African samples from the 1000 Genomes Project, while SAC patients are admixed with genetic contributions from Khoisan, Black African, European, and Asian individuals.[18] This highlights the necessity of analyzing SAB and SAC patients with ALS from South Africa as separate ancestry groups because variant allele frequencies differ between SAB and SAC groups, which could introduce a bias if these groups are analyzed as a combined African ancestry sample for association testing.

SCAF4, *SQSTM1*, and *STMN2* Short Structural Variants in South Africans With ALS

The *SCAF4* poly-T repeat genotypes for 21 patients with ALS and 15 ancestry-matched control samples did not pass the defined quality control filters and were excluded from the analysis (see methods). Six different alleles (14T–19T) were reported at this genomic location in this study. The *SCAF4* 18T allele, previously reported to be associated with ALS in a largely *SOD1* variant–positive cohort (Table 1), was not associated with ALS in South Africans ($p \geq 0.7$) (Table 3). No *SOD1* variants were found among the 10 SAC ALS 18T allele carriers, while 1 patient had a pathogenic *C9orf72* expansion.

Table 3 Comparison of *SCAF4* Poly-T Allele Frequencies in South African Patients With ALS and Ancestry-Matched South African Controls

	SAB, n alleles		SAC, n alleles	
***SCAF4* poly-T alleles, n (%)**	**ALS 54**	**Controls 238**	**ALS 132**	**Controls 32**
15T	0 (0)	0 (0)	1 (1)	0 (0)
16T	2 (4)	3 (1)	6 (5)	1 (3)
17T	52 (96)	232 (97)	115 (87)	30 (94)
18T	0 (0)	3 (1)	10 (8)	1 (3)
OR 18T (95% CI), *p*	0 (0–5.1), >0.9		2.5 (0.4–28.5), 0.7	

Abbreviations: ALS = amyotrophic lateral sclerosis; CI = confidence interval; OR = odds ratio; SAB = South African Black; SAC = South African Coloured; *SCAF4* = SR-Related CTD-Associated Factor 4 gene.
Controls refers to samples with matched South African ancestry. *p* refers to *p* value.

The *SQSTM1* insAAAC genotypes for 1 patient with ALS and 4 ancestry-matched control samples did not pass the defined quality control filters and were excluded from the analysis (see methods). In this study, we detected 3 alleles: reference (-), insAAAC (approximately 50% of all samples), and insACAAAAAAC. We did not replicate the previously reported *SQSTM1* insAAAC/insAAAC genotype association with ALS in our South African sample ($p \geq 0.4$) (Table 4).

The *STMN2* CA repeat genotypes for 4 patients with ALS and 7 ancestry-matched control samples did not pass the defined quality control filters and were excluded from the analysis (see methods). The *STMN2* alleles detected in this study ranged from 13-26 CA repeats. Long *STMN2* CA repeats were common in both cases with ALS and controls (approximately 50%) and were not associated with ALS ($p \geq 0.2$) (Table 5).

Table 4 Comparison of *SQSTM1* AAAC Insertion Frequencies in South African Patients With ALS and Ancestry-Matched South African Controls

	SAB, n individuals		SAC, n individuals	
***SQSTM1* genotypes, n (%)**	**ALS 29**	**Controls 124**	**ALS 84**	**Controls 22**
-/-	9 (31)	42 (34)	17 (20)	7 (32)
-/insAAAC	15 (52)	64 (52)	40 (48)	9 (41)
insAAAC/insAAAC	5 (17)	18 (15)	27 (32)	5 (23)
insAAAC/insACAAAAAAC	0 (0)	0 (0)	0 (0)	1 (5)
OR insAAAC/insAAAC (95% CI), *p*	1.2 (0.5–3.5), 0.8		1.6 (0.5–4.3), 0.4	

Abbreviations: ALS = amyotrophic lateral sclerosis; CI = confidence interval; OR = odds ratio; SAB = South African Black; SAC = South African Coloured; *SQSTM1* = sequestosome 1 gene.
Controls refers to samples with matched South African ancestry and - refers to the reference allele. *p* refers to *p* value.

Table 5 Comparison of *STMN2 CA* Repeat Length Frequencies in South African Patients With ALS and Ancestry-Matched South African Controls

	SAB, n individuals		SAC, n individuals	
***STMN2* genotypes, n (%)**	**ALS 28**	**Controls 121**	**ALS 82**	**Controls 22**
Short/short (S/S)	3 (11)	12 (10)	6 (7)	3 (14)
Long/short (L/S)	14 (50)	44 (36)	31 (37)	9 (41)
Long/long (L/L)	11 (39)	65 (54)	45 (56)	10 (45)
L/L (with 24 CA)	5 (18)	33 (27)	22 (27)	6 (27)
L/L (without 24 CA)	6 (21)	32 (26)	23 (29)	4 (18)
L/L (95% CI), *p*	0.6 (0.3–1.2), 0.2		1.5 (0.6–3.7), 0.5	
L/L (with 24 CA) (95% CI), *p*	0.6 (0.2–1.7), 0.3		1.0 (0.3–2.9), >0.9	

Abbreviations: ALS = amyotrophic lateral sclerosis; CI = confidence interval; NS = not significant; OR = odds ratio; SAB = South African Black; SAC = South African Coloured; *STMN2* = stathmin 2 gene.
Controls refers to samples with matched South African ancestry. Short (S) refers to repeat lengths 13-18 CA, long (L) refers to repeats ≥19 CA. *p* refers to *p* value.

SMN1 and *SMN2* Copy Numbers in South Africans With ALS

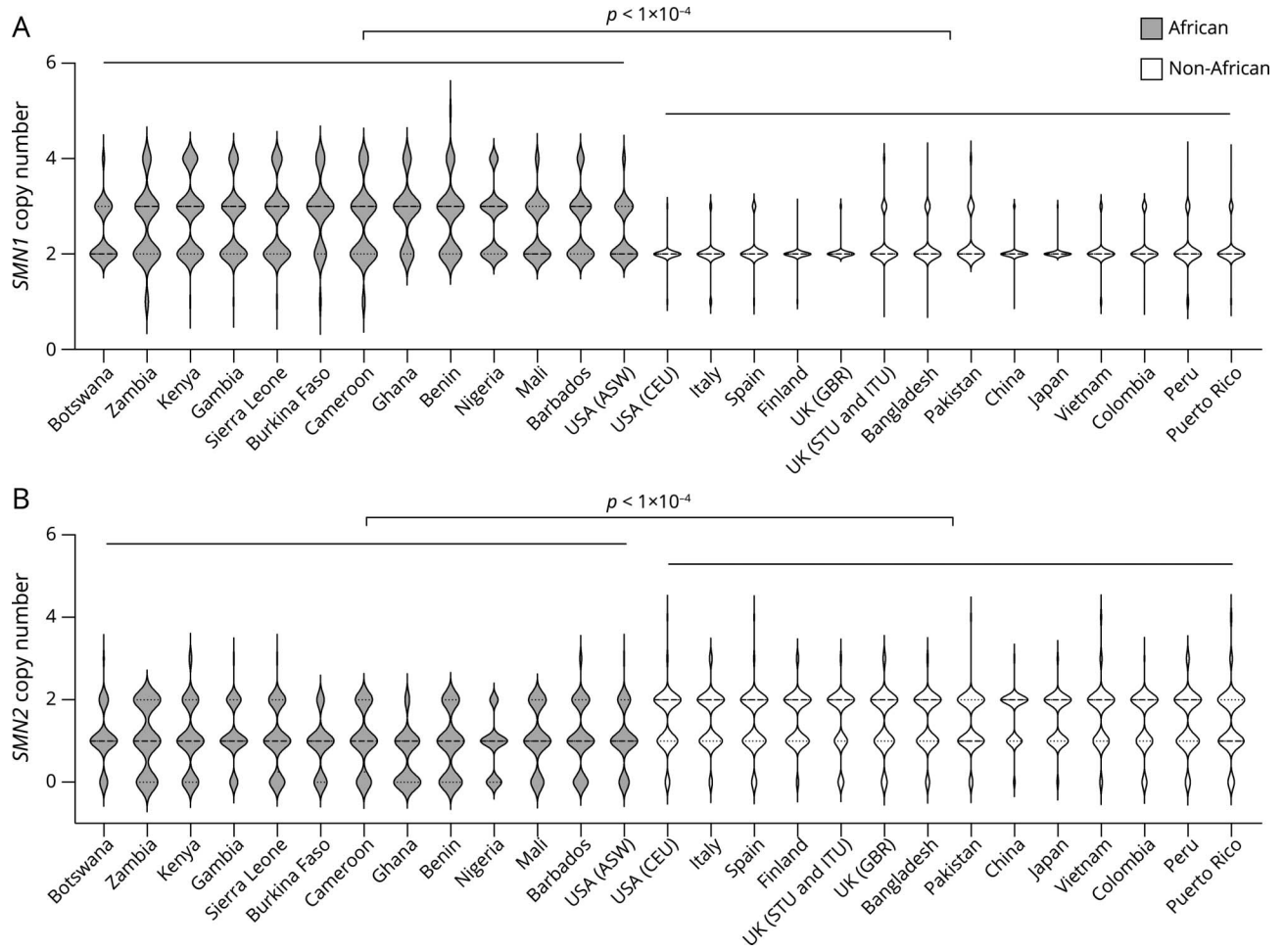
The carrier frequency (1 copy of *SMN1*) for the autosomal recessive disorder, spinal muscular atrophy (SMA), was approximately 2% in African ancestry individuals in this study (cases with ALS and matched African ancestry controls) (Table 6). Approximately 50% of SAB individuals have ≥3 copies of *SMN1*, and the overall frequency distribution of *SMN1* copy number states does not differ between cases with ALS and controls ($p \geq 0.5$) (Table 6). By contrast, *SMN1* copy number frequencies differed between SAC cases with ALS and ancestry-matched controls where cases with ALS had a higher frequency of *SMN1* copy number 2 (75%, $p = 0.03$) and a lower frequency of *SMN1* copy number 3 (17%, $p = 0.003$). Although the sample is too small for any conclusion, 2 out of 8 successfully genotyped SAC ALS patients with predominant lower motor neuron (LMN) involvement (LMN-ALS, Table 2) had ≥3 copies of *SMN1* (25%). Although the overall frequency distribution of *SMN2* copy number states did not differ between ALS cases and controls in both SAB and SAC groups ($p \geq 0.05$), individuals lacking the *SMN2* gene (12%–32% respectively) were not infrequent in this South African sample (Table 6). Total SMN protein levels were estimated using the Veldink formula (*SMN1* copy number + 0.2* *SMN2* copy number)[30] and were similar in patients with ALS and controls for both SAB and SAC groups. *SMN1* and *SMN2* copy number calls from the SMN Caller tool were absent or ambiguous in 22/264 (8%) of the combined African ancestry cases with ALS and matched ancestry controls in this study.

Table 6 Comparison of *SMN1* and *SMN2* Copy Number States in South African Patients With ALS and Ancestry-Matched South African Controls

	SAB, n individuals			SAC, n individuals		
	ALS 29	**Control 117**	**OR (95% CI), *p* Value**	**ALS 75**	**Control 21**	**OR (95% CI), *p* Value**
***SMN1* CN**						
1	1 (3)	4 (4)	1.0 (0.1–6.4), >0.9	1 (1)	0 (0)	∞, >0.9
2	15 (52)	57 (50)	1.1 (0.5–2.6), 0.8	56 (75)	10 (45)	3.2 (1.1–8.4), 0.03
3	10 (34)	47 (38)	0.8 (0.3–1.7), 0.7	13 (17)	11 (50)	0.2 (0.1–0.5), 0.003
4	3 (10)	8 (7)	1.6 (0.4–5.7), 0.5	3 (4)	0 (0)	∞, >0.9
5	0 (0)	1 (1)	0 (0–36), >0.9	2 (3)	0 (0)	∞, >0.9
***SMN2* CN**						
0	6 (21)	25 (21)	0.9 (0.4–2.6), >0.9	9 (12)	7 (32)	0.3 (0.1–0.8), 0.06
1	11 (38)	55 (47)	0.7 (0.3–1.5), 0.4	27 (36)	9 (41)	0.8 (0.3–1.9), 0.6
2	12 (41)	36 (30)	1.6 (0.7–3.7), 0.3	38 (51)	5 (27)	3.3 (1.0–8.7), 0.05
3	0 (0)	1 (1)	0 (0–36), >0.9	1 (1)	0 (0)	∞, >0.9
SMN protein levels (Veldink estimation)						
≤2.2	5 (17)	29 (25)	$p = 0.5$	21 (28)	5 (24)	$p = 0.8$
>2.2	24 (83)	88 (75)		54 (72)	16 (76)	

Abbreviations: ∞ = infinity; ALS = amyotrophic lateral sclerosis; CN = copy number; SAB = South African Black; SAC = South African Coloured; *SMN1* = survival of motor neuron 1 gene; *SMN2* = survival of motor neuron 2 gene.
Controls refers to samples with matched South African ancestry. *p* refers to *p* value.

Figure Distribution of *SMN1* and *SMN2* Gene Copy Numbers in Different Population Groups



SMN1 (A) and *SMN2* (B) gene copy numbers in different population groups grouped according to continental ancestry (African or non-African). UK = United Kingdom, USA = United States of America, ASW = African Ancestry in Southwest US, CEU = Utah residents with Northern and Western European ancestry, GBR = British in England and Scotland, STU = Sri Lankan Tamil in the UK, ITU = Indian Telugu in the UK. The overall frequency distribution of *SMN1* and *SMN2* copy numbers between Afr = can (n = 994) and non-Afr = can (n = 1,669) populations was compared using a χ^2 test.

The *SMN1*/*SMN2* gene architecture was further explored in samples from the 1000 Genomes as well as H3Africa Genotyping Chip Projects (Figure). Samples of African ancestry had higher *SMN1* copy numbers (approximately 40% have ≥3 copies) compared with non-African ancestry samples where most have 2 copies of *SMN1* ($p < 1 \times 10^{-4}$, Figure, A). The findings for the *SMN2* gene were reciprocal, where African ancestry samples had lower *SMN2* copy numbers (approximately 20% have no *SMN2* gene copies) compared with non-African ancestry samples (approximately 9%, $p < 1 \times 10^{-4}$, Figure, B).

Discussion

Previous work has shown that the genetic drivers of ALS may differ across geographies (i.e., an ALS associated allele may be relatively frequent in 1 ALS population but rare or absent in a different ALS population).[25,31] Nonetheless, new ALS gene discoveries such as the association of *NEK1* loss-of-function variants with ALS, first identified in a large European ALS cohort by unbiased gene burden analysis,[32] have subsequently been replicated in multiple smaller ALS cohorts of diverse ancestries.[33,34] This highlights the utility of trans-ancestry replication studies in confirming putative ALS genes. While inclusion of African cohorts are important, they are at present small and their high levels of genetic diversity and substructure require careful ancestry matching of control groups in association testing.

In this study, we did not replicate the previously reported association of various SSVs (*SCAF4*, *SQSTM1* and *STMN2*) with ALS in a small sample of African ancestry patients. In contrast to previous findings[8,9] where the *SCAF4* and *SQSTM1* variant associations were reported in European familial ALS cohorts with multiple individuals from the same family and a high frequency of pathogenic *SOD1* variants (approximately 65% with *SOD1* A5V variant, formerly known

as A4V), our study investigated a largely sporadic African ancestry ALS cohort with only 2 related individuals and a 6% frequency of pathogenic *SOD1* variants (none of which were A5V). While sampling bias (population-based, family-based, and variant-based) may have contributed to the detection of *SCAF4* and *SQSTM1* variant associations with ALS in previous reports, the fact that we did not replicate these associations in our small African ALS sample does not provide conclusive evidence against their role in disease. Indeed, *SCAF4* and *SQSTM1* ALS associated alleles were more frequent in SAC patients with ALS, who may have approximately 20% European admixture,[35] although their association with ALS (OR 2.5 and 1.6 respectively) was not statistically significant. For the *SCAF4* analysis, our study had 96% power to detect an odds ratio of 5, while the *SQSTM1* analysis was underpowered (55% power to detect an odds ratio of 2).

STMN2 long genotypes with 24 CA repeats, previously reported to be associated with ALS in Europeans (where the control carrier frequency was approximately 5%), were common in both subpopulations of African patients with ALS and controls (approximately 25%). We did not replicate the association of long *STMN2* CA repeats (with or without 24 CA) with ALS in our adequately powered African ancestry sample (80% chance of detecting an odds ratio >2), which is similar to the findings reported in a recent replication study in Europeans.[11]

For *SMN1*/*SMN2* copy number calling, we used the Illumina *SMN* Copy Number Caller, which performs similarly (>99% accuracy) to the gold standard diagnostic detection of *SMN* copy numbers (digital PCR and MLPA).[28] Although the tool has been designed for use on WGS data from diverse ancestries, due to the incorporation of 8 SNV sites that are nearly fixed in all populations (see Methods), we were not able to perform *SMN1*/*SMN2* copy number calling in 8% of African ancestry individuals (owing to ambiguity at these SNV sites) which is greater than the previously reported 1% of no calls in African ancestry individuals.[28] This highlights the uniqueness of Southern African genetic variation, which is not represented in the large-scale publicly available population data sets, used to train computational models and tools.

Consistent with the findings of the recent meta-analysis examining *SMN1*/*SMN2* copy number states and ALS,[17] *SMN1* and *SMN2* copy numbers did not differ between ALS cases and ancestry-matched controls for SAB individuals. Although the SAC patients with ALS had fewer copies of *SMN1* compared with ancestry-matched controls, it is worth noting that the SAC ALS copy number distribution for *SMN1* is more similar to non-African *SMN1* gene architecture (Figure, A), and together with few ancestry-matched controls in this admixed group, the results could merely reflect statistically different proportions of non-African ancestry between the SAC ALS case and control groups.

Our broader analysis of *SMN1*/*SMN2* copy numbers in diverse ancestries confirms previous reports that the architecture of these genes differs between African and non-African populations where Africans have higher copies of *SMN1* and lower copies of *SMN2*[36,37] (Figure). The overall SMA carrier frequency (2% with 1 copy of *SMN1*) in this study (Table 6) is similar to reported frequencies in Europeans (2%–5%).[37]

While most individuals of non-African ancestry have 2 copies of *SMN2*, in this study, we confirm that most individuals with African ancestry have ≤1 copy of *SMN2*. This might be clinically relevant because SMA results from complete deficiency of SMN protein, and a gene-based therapy acting on *SMN2* to increase SMN protein levels for therapeutic benefit has been established.[38]

This study highlights the usefulness of investigating putative ALS associated variants in independent population groups, particularly where the ancestry differs from the discovery cohort and underscores the importance of including African ancestry patients in gene discovery efforts. This will ensure that gene-based therapies, developed for ALS and other disorders, will benefit Africans in future.

Acknowledgment

The authors thank the CReATe consortium and Paul Taylor's laboratory at St Jude's funded by the Amyotrophic Lateral Sclerosis Association (ALSA) and the St Jude American Lebanese Syrian Associated Charities (ALSAC) for their support and for funding the WGS data generation on 34 ALS cases. The Clinical Research in ALS and related disorders for Therapeutic Development (CReATe) Consortium (U54NS092091) is part of Rare Diseases Clinical Research Network (RDCRN), an initiative of the Office of Rare Diseases Research (ORDR), National Center for Advancing Translational Sciences (NCATS). This consortium is funded through collaboration between NCATS and the NINDS. They acknowledge the support of the UCT Division of Computational Biology (N Mulder) who funded the whole-genome sequencing of 25 cases with ALS (Human Genome Research Institute: U24HG006941) and provided bioinformatics support (G Botha). The SAHGP dataset was generated by the Southern African Human Genome Program, a national initiative funded by the Department of Science and Technology of South Africa. This study makes use of data generated by H3Africa. A full list of the investigators who contributed to the generation of the data is available from h3africa.org. The funding for this project comes through the Human Heredity and Health in Africa (H3Africa) Initiative, which is funded by the National Institutes of Health and the Wellcome Trust through SFA Foundation. The authors acknowledge the use of the Ilifu cloud computing facility (ilifu.ac.za), a partnership between the University of Cape Town, the University of the Western Cape, the University of Stellenbosch, Sol Plaatje University, the Cape Peninsula University of Technology, and the South African Radio Astronomy Observatory. The Ilifu facility is supported by contributions from the Inter-University Institute for Data

Intensive Astronomy (IDIA—a partnership between the University of Cape Town, the University of Pretoria, and the University of the Western Cape), the Computational Biology division at UCT, and the Data Intensive Research Initiative of South Africa (DIRISA). This (publication) was made possible (in part) by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the authors.

Study Funding

N.R. Monnakgotla received funding from the University of Cape Town (fellowships administered by the Neurology Research Group, Department of Medicine and UCT). M. Nel is the recipient of a CReATe scholar award and a Carnegie Developing Emerging Academic Leaders (DEAL) award. This publication was made possible (in part) by grants from the Carnegie Corporation of New York (M. Nel), the Joost van der Westhuizen Centre for Neurodegeneration (donated by Aspen Pharmacare) and the South African National Research Foundation (J.M. Heckmann and A.C. Mahungu). The statements made and views expressed are solely the responsibility of the authors.

Disclosure

J.P. Taylor is a consultant for Nido Biosciences; M. Benatar serves on the ALS Association Board of Trustees and holds grants from NIH (R01-NS105479, U01-NS107027, U54-NS092091) and the Muscular Dystrophy Association (645863), intellectual property from the University of Miami licensed to Biogen (IP-142A), a provisional patent related to determining the onset of amyotrophic lateral sclerosis and consults for Alector, Annexon, Arrowhead, Biogen, Denali, Novartis, Orphazyme, Roche, Sanofi and UniQure; J.M. Heckmann holds a seed grant from the ALSA (23-SGP-626), serves on the scientific advisory committee for the International ALS MND Alliance and consults for Merck. The other authors declare that they have no competing interests. Go to Neurology.org/NG for full disclosures.

Publication History

Received by *Neurology: Genetics* February 14, 2023. Accepted in final form April 3, 2023. Submitted and externally peer reviewed. The handling editor was Associate Editor Raymond P. Roos, MD, FAAN.

Appendix Authors

Name	Location	Contribution
Nomakhosazana R. Monnakgotla, MSc	Neurology Research Group, Division of Neurology, Department of Medicine; Neuroscience Institute, University of Cape Town, South Africa	Drafting/revision of the article for content, including medical writing for content; analysis or interpretation of data
Amokelani C. Mahungu, MSc	Neurology Research Group, Division of Neurology, Department of Medicine; Neuroscience Institute, University of Cape Town, South Africa	Analysis or interpretation of data
Jeannine M. Heckmann, MBChB, PhD	Neurology Research Group, Division of Neurology, Department of Medicine; Neuroscience Institute, University of Cape Town, South Africa	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data
Gerrit Botha, MSc	Computational Biology Division, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa	Major role in the acquisition of data; analysis or interpretation of data
Nicola J. Mulder, BSc, PhD	Computational Biology Division, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, South Africa	Major role in the acquisition of data; analysis or interpretation of data
Gang Wu, PhD	Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Evadnie Rampersaud, BSc, PhD	Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN	Drafting/revision of the article for content, including medical writing for content
Jason Myers, MS	Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN	Drafting/revision of the article for content, including medical writing for content
Marka Van Blitterswijk, MD, PhD	Department of Neuroscience, Mayo Clinic, Jacksonville, Florida	Drafting/revision of the article for content, including medical writing for content
Rosa Rademakers, PhD	Center for Molecular Neurology, University of Antwerp, Belgium	Drafting/revision of the article for content, including medical writing for content
J. Paul Taylor, MD, PhD	Department of Cell and Molecular Biology, St. Jude Children's Research Hospital, Memphis, TN	Drafting/revision of the article for content, including medical writing for content
Joanne Wuu, ScM	Department of Neurology, University of Miami, Florida	Drafting/revision of the article for content, including medical writing for content
Michael Benatar, MD, PhD	Department of Neurology, University of Miami, Florida	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Melissa Nel, MBChB, PhD	Neurology Research Group, Division of Neurology, Department of Medicine; Neuroscience Institute, University of Cape Town, South Africa	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data

References

1. Al-Chalabi A, Calvo A, Chio A, et al. Analysis of amyotrophic lateral sclerosis as a multistep process: a population-based modelling study. *Lancet Neurol.* 2014;13(11):1108-1113. doi:10.1016/S1474-4422(14)70219-4
2. Vucic S, Higashihara M, Sobue G, et al. ALS is a multistep process in South Korean, Japanese, and Australian patients. *Neurology*. 2020;94(15):E1657-E1663. doi:10.1212/WNL.0000000000009015
3. Lowry JL, Ryan ÉB, Esengul YT, Siddique N, Siddique T. Intricacies of aetiology in intrafamilial degenerative disease. *Brain Commun.* 2020;2(2). doi:10.1093/braincomms/fcaa120
4. Theunissen F, Flynn LL, Anderton RS, et al. Structural variants may be a source of missing heritability in sALS. *Front Neurosci.* 2020;14:1-11. doi:10.3389/fnins.2020.00047
5. Al Khleifat A, Iacoangeli A, van Vugt JJFA, et al. Structural variation analysis of 6,500 whole genome sequences in amyotrophic lateral sclerosis. *NPJ Genom Med.* 2022;7(1):8. doi:10.1038/s41525-021-00267-9
6. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81. doi:10.1038/nature15394
7. Saul R, Lutz MW, Burns DK, Roses AD, Chiba-Falek O. The SSV evaluation system: a tool to prioritize short structural variants for studies of possible regulatory and causal variants. *Hum Mutat.* 2016;37(9):877-883. doi:10.1002/humu.23023
8. Pytte J, Flynn LL, Anderton RS, et al. Disease-modifying effects of an *SCAF4* structural variant in a predominantly *SOD1* ALS cohort. *Neurol Genet.* 2020;6(4):e470. doi:10.1212/NXG.0000000000000470
9. Pytte J, Anderton RS, Flynn LL, et al. Association of a structural variant within the *SQSTM1* gene with amyotrophic lateral sclerosis. *Neurol Genet.* 2020;6(2):e406. doi:10.1212/NXG.0000000000000406
10. Theunissen F, Anderton RS, Mastaglia FL, et al. Novel *STMN2* variant linked to amyotrophic lateral sclerosis risk and clinical phenotype. *Front Aging Neurosci.* 2021;13:658226. doi:10.3389/fnagi.2021.658226
11. Ross JP, Akçimen F, Liao C, et al. Questioning the association of the *STMN2* dinucleotide repeat with amyotrophic lateral sclerosis. *Neurol Genet.* 2022;8(4):e678. doi:10.1212/NXG.0000000000000678
12. Bowerman M, Becker CG, Yáñez-Muñoz RJ, et al. Therapeutic strategies for spinal muscular atrophy: SMN and beyond. *Dis Model Mech.* 2017;10(8):943-954. doi:10.1242/dmm.030148
13. Veldink JH, van den Berg LH, Cobben JM, et al. Homozygous deletion of the survival motor neuron 2 gene is a prognostic factor in sporadic ALS. *Neurology*. 2001;56(6):749-752. doi:10.1212/WNL.56.6.749
14. Corcia P, Camu W, Halimi J-M, et al. *SMN1* gene, but not *SMN2*, is a risk factor for sporadic ALS. *Neurology*. 2006;67(7):1147-1150. doi:10.1212/01.wnl.0000233830.85206.1e
15. Wang X-B, Cui N-H, Gao J-J, Qiu X-P, Zheng F. *SMN1* duplications contribute to sporadic amyotrophic lateral sclerosis susceptibility: evidence from a meta-analysis. *J Neurol Sci.* 2014;340(1-2):63-68. doi:10.1016/j.jns.2014.02.026
16. Sangare M, Dicko I, Guinto CO, et al. Does the survival motor neuron copy number variation play a role in the onset and severity of sporadic amyotrophic lateral sclerosis in Malians? *eNeurologicalSci*. 2016;3:17-20. doi:10.1016/j.ensci.2015.12.001
17. Moisse M, Zwamborn RAJ, van Vugt J, et al. The effect of *SMN* gene dosage on als risk and disease severity. *Ann Neurol.* 2021;89(4):686-697. doi:10.1002/ana.26009
18. Nel M, Mahungu AC, Monnakgotla N, et al. Revealing the mutational spectrum in Southern Africans with amyotrophic lateral sclerosis. *Neurol Genet.* 2022;8(1):e654. doi:10.1212/NXG.0000000000000654
19. Brooks BR, Miller RG, Swash M, Munsat TL. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph Lateral Scler.* 2000;1(5):293-299. doi:10.1080/146608200300079536
20. Choudhury A, Ramsay M, Hazelhurst S, et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat Commun.* 2017;8(1):1-12. doi:10.1038/s41467-017-00663-9
21. Nel M, Mulder N, Europa TA, Heckmann JM. Using whole genome sequencing in an African subphenotype of myasthenia gravis to generate a pathogenetic hypothesis. *Front Genet.* 2019;10:136. doi:10.3389/fgene.2019.00136
22. Sengupta D, Choudhury A, Fortes-Lima C, et al. Genetic substructure and complex demographic history of South African Bantu speakers. *Nat Commun.* 2021;12(1):2080. doi:10.1038/s41467-021-22207-y
23. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
24. Choudhury A, Aron S, Botigué LR, et al. High-depth African genomes inform human migration and health. *Nature*. 2020;586(7831):741-748. doi:10.1038/s41586-020-2859-7
25. Nel M, Agenbag GM, Henning F, Cross HM, Esterhuizen A, Heckmann JM. *C9orf72* repeat expansions in South Africans with amyotrophic lateral sclerosis. *J Neurol Sci.* 2019;401:51-54. doi:10.1016/j.jns.2019.04.026
26. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-1303. doi:10.1101/gr.107524.110
27. 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
28. Chen X, Sanchis-Juan A, French CE, et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet Med.* 2020;22(5):945-953. doi:10.1038/s41436-020-0754-0
29. de Vries J, Tindana P, Littler K, et al. The H3Africa policy framework: negotiating fairness in genomics. *Trends Genet.* 2015;31(3):117-119. doi:10.1016/j.tig.2014.11.004
30. Veldink JH, Kalmijn S, Van der Hout AH, et al. *SMN* genotypes producing less SMN protein increase susceptibility to and severity of sporadic ALS. *Neurology*. 2005;65(6):820-825. doi:10.1212/01.wnl.0000174472.03292.dd
31. Mejzini R, Flynn LL, Pitout IL, Fletcher S, Wilton SD, Akkari PA. ALS genetics, mechanisms, and therapeutics: where are we now? *Front Neurosci.* 2019;13:1-27. doi:10.3389/fnins.2019.01310
32. Kenna KP, Van Doormaal PTC, Dekker AM, et al. *NEK1* variants confer susceptibility to amyotrophic lateral sclerosis. *Nat Genet.* 2016;48(9):1037-1042. doi:10.1038/ng.3626
33. Gratten J, Zhao Q, Benyamin B, et al. Whole-exome sequencing in amyotrophic lateral sclerosis suggests *NEK1* is a risk gene in Chinese. *Genome Med.* 2017;9(1):1-9. doi:10.1186/s13073-017-0487-0
34. Naruse H, Ishiura H, Mitsui J, et al. Loss-of-function variants in *NEK1* are associated with an increased risk of sporadic ALS in the Japanese population. *J Hum Genet.* 2021;66(3):237-241. doi:10.1038/s10038-020-00830-9
35. De Wit E, Delport W, Rugamika CE, et al. Genome-wide analysis of the structure of the South African Coloured population in the Western Cape. *Hum Genet.* 2010;128(2):145-153. doi:10.1007/s00439-010-0836-1
36. Vorster E, Essop FB, Rodda JL, Krause A. Spinal muscular atrophy in the black South African population: a matter of rearrangement? *Front Genet.* 2020;11:1-15. doi:10.3389/fgene.2020.00054
37. Sangaré M, Hendrickson B, Sango HA, et al. Genetics of low spinal muscular atrophy carrier frequency in sub-Saharan Africa. *Ann Neurol.* 2014;75(4):525-532. doi:10.1002/ana.24114
38. Mercuri E, Darras BT, Chiriboga CA, et al. Nusinersen versus Sham control in later-onset spinal muscular atrophy. *N Engl J Med.* 2018;378(7):625-635. doi:10.1056/NEJMoa1710504