Received: 3 June 2020 | Revised: 30 July 2020 | Accepted: 10 February 2021

DOI: 10.1002/mgg3.1629

Molecular Genetics & Genomic Medicine Open Access WILEY

ORIGINAL ARTICLE

Worldwide variation of the *COL14A1* gene is shaped by genetic drift rather than selective pressure

Carla M. Calò[1] | Federico Onali[1] | Renato Robledo[2] | Laura Flore[1] | Myosotis Massidda[1] | Paolo Francalacci[1]

[1]Department of Life and Environment Sciences, University of Cagliari, Cagliari, Italy

[2]Department of Biomedical Sciences, University of Cagliari, Cagliari, Italy

Correspondence
Renato Robledo, Department of Biomedical Sciences, University of Cagliari, Cagliari, Italy.
Email: rrobledo@unica.it

Funding information
Università degli Studi di Cagliari

Abstract

Background: The aim of this study is to analyze the worldwide distribution of SNP rs4870723 in *COL14A1* gene to check if there are significant genetic differences among different populations and to test if the gene is a trait under selection.

Methods: Genomic DNA was extracted from 69 unrelated individuals from Sardinia and genotyped for SNP rs4870723. Data were compared with 26 different populations, clustered in 5 super-populations, from the public 1000 genomes database. Allele frequency and heterozygosity were calculated with Genepop. The Hardy–Weinberg equilibrium and pairwise population differentiation through analysis of molecular variance (AMOVA FST) were determined with Arlequin.

Results: Allele frequencies of *COL14A1* rs4870723 were compared in 27 populations clustered in 5 super-populations. All populations were in the Hardy–Weinberg equilibrium. In almost all populations, allele C was the most frequent allele, reaching the highest values in East Asia. The 27 populations showed an appreciable structure, with significant differences observed between European, African, and Asian populations.

Conclusion: Significant differences were observed in the rs4870723 SNP distribution among the populations studied. However, we found no evidence for a selective pressure. Rather, the differentiation among the populations is likely the result of founder effect, genetic drift, and cultural factors, all events known to establish and maintain genetic diversity between populations.

KEYWORDS
1000 genomes, collagen, Sardinia, selection, SNPs

1 | INTRODUCTION

Type-XIV collagen is a homotrimer belonging to a family of non-fibrillar collagens referred to as fibril associated collagens with interrupted triple helices (FACITs). It is able to form interfibrillar connections and may influence fibril and matrix density, suggesting that the type-XIV collagen may also be involved in fibrillogenesis (Young et al., 2002). Moreover, it has been demonstrated its role in the contraction of collagen gels, suggesting its ability to modulate tissue response to mechanical stress (Chiquet, 1999). Indeed, the type-XIV collagen is often present in areas of high mechanical stress, indicating a potential role in maintaining mechanical tissue or in affecting mechanical properties of a tissue (Berthod et al., 1997). For all these characteristics, its implication in onset of tendon pathologies has been suggested, hypothesizing that a single nucleotide polymorphism (SNP) may predispose to injuries due to either a reduction of the tensile strength of collagen or altered fibrillogenesis.

In a genome wide association study (GWAS) on twin siblings with an anterior cruciate ligament (ALC) rupture, single-nucleotide variants (SNVs) for *COL5A2*, *COL5A3*, *COL14A1*, and *COL15A1* genes showed a damaging disease impact profile (Caso et al., 2016). Moreover, positive association was found for pelvic organ collapse (Li et al., 2020) or ALC injuries (Massidda et al., 2018); however, no correlation was found when the rupture of Achilles tendon was investigated (September et al., 2008). All these considerations drove us to investigate the genetic variability of *COL14A1* (OMIM n. 120324) among human populations to assess its possible role on natural selection. Our choice to analyze *COL14A1*, over other *COL* genes, has been driven by the paucity of data in the literature, which also gave contradictory results.

COL14A1 gene is located on chromosome 8 (8q24.12) and codes for alpha 1 chain of type-XIV collagen. Within the gene, 53,950 SNPs have been detected in dbSNP (National Centre for Biotechnology Information, NCBI). We applied a filter selecting for the exonic missense SNPs with a minimum allele frequency (MAF) >0.01. Among the three SNPs obtained, rs4870723, a mutation at position 90 of exon 14 (c.1952 C > A, p.H563 N) with the highest MAF was considered the most informative SNP.

The aim of this study is to analyze the worldwide distribution of SNP rs4870723 in the *COL14A1* gene, to determine if there are significant genetic differences among ethnic groups, and to check if such different distribution is due to selective pressure or it is the result of random genetic drift. Furthermore, we provide new data on rs4870723 for the Sardinian population (Italy). Sardinia is an interesting case study since evolutionary forces such as the balancing selection due to formerly endemic malaria, with the combined effect of isolation and inbreeding on genetic drift, shaped its genetic variation making the Sardinian population an outlier in the European context.

2 | MATERIALS AND METHODS

Ethical Compliance: The study was approved by Ethic Committees of Azienda Ospedaliera Universitaria (AOU) of Cagliari University (Italy), and written informed consent was obtained from each participant. A total of 69 individuals of both sexes (31 females and 38 males) from Sardinia (Italy) were analyzed. All selected individuals were unrelated, apparently healthy, born, and resident in the area for at least three generations. Genomic DNA was extracted from buccal swab, through salting out method, and amplified by standard PCR using the following primers:

Forward 5’-CTTTGCCAGAGTCACATGGT-3’. Reverse 5’-TGTCCCGGAACTTACCTCAT-3’. PCR was performed in 25 μL volumes containing 200 ng genomic DNA; 20 pmol of each primer; NZYTaq II 2×Master Mix (0.2 U/μL). Amplifications were conducted by denaturing at 94°C for 3 minutes, followed by 30 cycles at 94°C for 30 seconds, 54°C for 30 seconds, and 72°C for 1 minute, and a final extension at 72°C for 5 minutes. NcoI enzymatic digestion of amplified products yelded either a single band of 530 bp (A allele) or two bands of 465 bp and 65 bp (C allele). Fragments were separated by 2% agarose gel electrophoresis and stained with Sybr green. The worldwide variation of *COL14A1* (GenBank accession number: NG_033107.1) polymorphism A > C (SNP rs4870723, Chr. 8:121228679 Forward Strand in GRCh37) was analyzed with data obtained from the public database 1000 Genomes Phase 3 Browser (The, 1000 genomes project Consortium, 2015) plus Sardinia samples. The final dataset provided information on 27 different populations clustered in 5 super-populations (African, American, East Asia, European, and South Asia). Allele frequency and heterozygosity (gene diversity) for each population were calculated using the Genepop software (ver. 4.4). The Hardy–Weinberg equilibrium, population relationships through pairwise differences and hierarchical analyses of molecular variation (AMOVA FST) were determined with Arlequin v.3.5 (Excoffier et al., 2007) using all the 27 populations. Finally, global population relationships (same 27 populations) have been checked by means of Multidimensional Scaling (MDS) based on FST genetic distance through Statistica Programme (ver. 7) for rs4870723.

Selection signatures using the whole gene *COL14A1* were evaluated through the 1000 Genomes Selection Browser 1.0 (Pybus et al., 2014) comparing data from CEU (European, N = 97), YRI (Sub-Saharan, N = 88), and CHB (China, N = 85) for FST-rank scores (data from the integrated Phase 1 variant set), and through PopHuman (Casillas et al., 2018), which uses data generated by the 1000GP Phase III.

3 | RESULTS

Table 1 reports allele frequencies of *COL14A1* rs4870723 SNP in 27 populations distributed into five geographical areas. All populations fit the Hardy–Weinberg equilibrium (p > 0.05). With the exception of Great Britain (GBR), Finland (FIN), Northwestern Europe (CEU), and Sardinia (SAR), allele C shows the highest frequency in all populations, ranging from 0.514 in Puerto Rico (PUR) to 0.715 in Chinese Dai (CDX). When pairwise difference

TABLE 1 Frequencies of rs4870723 alleles in 27 populations clustered in five geographical areas

		allele	allele		
Population	**Code**	**freq. A**	**freq. C**	**H n.b.**	**N**
Africa					
Afrocaribbeans, Barbados	ACB	0.328	0.672	0.4444	97
Afroamerican, USA	ASW	0.443	0.557	0.4973	60
Esan, Nigeria	ESN	0.359	0.641	0.4623	99
Luhya, Kenya	LWK	0.394	0.606	0.4799	99
Gambian, Gambia	GWD	0.341	0.659	0.4512	113
Mende, Sierra Leone	MSL	0.406	0.594	0.4851	85
Yoruba, Nigeria	YRI	0.366	0.634	0.4661	108
America					
Colombians, Colombia	CLM	0.415	0.585	0.4881	94
Mexicans, USA	MXL	0.344	0.656	0.4547	64
Peruvians, Peru	PEL	0.424	0.576	0.4912	85
Puerto Rricans, Puerto Rico	PUR	0.486	0.514	0.5020	104
East Asia					
Dai, China	CDX	0.285	0.715	0.4097	93
N Han, China	CHB	0.330	0.670	0.4444	103
S Han, China	CHS	0.371	0.629	0.4692	105
Japanese, Japan	JPT	0.356	0.644	0.4606	104
Kinh, Vietnam	KHV	0.308	0.692	0.4285	99
South Asia					
Bengali, Bangla Desh	BEB	0.448	0.552	0.4974	86
Gujarati Indian, USA	GIH	0.417	0.583	0.4888	103
Indian Telogu, UK	ITU	0.338	0.662	0.4499	102
Punjabi, Pakistan	PJL	0.411	0.589	0.4869	96
Sri Lankan Tamil, UK	STU	0.436	0.564	0.4943	102
Europe					
NW Europeans, USA	CEU	0.535	0.465	0.5000	99
Finnish, Finland	FIN	0.646	0.354	0.4594	99
British, UK	GBR	0.544	0.456	0.4989	91
Iberians, Spain	IBS	0.453	0.547	0.4980	107
Tuscans, Italy	TSI	0.481	0.519	0.5016	107
Sardinians, Italy*	SAR	0.481	0.519	0.5016	107

Note: Data from: 1000 Genomes.

Abbreviations: H n.b., nonbiased expected heterozygosity; N, numbers of individuals.

*Present study.

analysis is carried out, significant p-values are observed when European populations are compared with the African (0.0233), East Asia (0), and South Asia populations (0.04080). It is noteworthy that Sardinia population appears significantly differentiated ($p < 0.05$) from all populations with the exception of Finland ($p = 0.54955$). The hierarchical structure of these groups, measured through AMOVA, emphasizes low degree of differentiation that is imputable mainly to European population. Indeed, of a total FST genetic variance of 3.51% ($p < 0.001$) when the five groups are compared, the variance attributable to differences among groups (FCT) accounts for 2.67% ($p < .001$). When Europe is eliminated from the analysis, the values of FST drastically decrease to 0.21%.

Possible traces of selection signatures in the whole *COL14A1* gene have been evaluated using FST rank scores for comparisons among CEU, YRI, and CHB. Significant probability values of FST-rank scores ($p < 0.01$) are found in the comparison CEU vs YRI and CEU vs CHB. In the first comparison, only five variants show significant values, but the corresponding area does not include the SNP under scrutiny. Instead, when CEU and CHB are compared, 29 variants, located within a 39.30 kb coding region in chr8:121223036–121262332, show significant values. This region includes the *COL14A1* SNP rs4870723 (Chr8: 121228679).

This result was verified with PopHuman (https://pophuman.uab.cat/), which showed a weak sign of negative selection for 3 out of the 10 parameters calculated for the detection of selective pressure (Tajima, Dos, and ka/ks), but in no case the region involved in the selection included the *COL14A1* SNP rs4870723 (Chr8: 121228679).

The MDS shows an appreciable population structure among the 27 populations (Figure 1). Asian populations are placed at the negative values of the first dimension, followed by African populations, while the European populations occupy the positive values. The Amerindian populations are rather scattered in the central part of the graph, possibly due to the heterogeneous composition of the sample. The Sardinian represents an outgroup in the graph, being at the extreme values in both dimensions.

4 | DISCUSSION

In this study, the worldwide distribution of rs4870723 A/C in the *COL14A1* gene was studied to determine if there are significant genetic differences among ethnically different populations and to understand if its distribution has been influenced by prehistoric and more recent demographic events or it is under a selective pressure.

Searching for Darwinian selection in natural populations has been the focus of a multitude of studies over the last decades. Different selection forces can negatively or positively select SNPs that are associated with disadvantageous or advantageous traits, respectively. For example, while negative selection tends to decrease the level of population differentiation, positive selection tends to increase it (Barreiro et al., 2008).

The data here reported do not show evidence of selective pressure since all the samples meet the Hardy–Weinberg equilibrium. Possible traces of selection signatures for the region including rs4870723 are found only for the comparison CEU vs CHB, two rather heterogeneous populations, but this result was not confirmed when the PopHuman browser was used.

The MDS points out to a population structure according to a model of isolation by distance with African populations in the central position and the Asian and European population structured in divergent direction. In particular, strong effect of the genetic drift can be revealed by the eccentric placement of the isolated population such as the Dai (a population located on the mountainous area of south China), the Finns (a separate northernmost European population) and, mostly, Sardinia (a central Mediterranean island). Finland and Sardinian are well known outgroups of European genetic

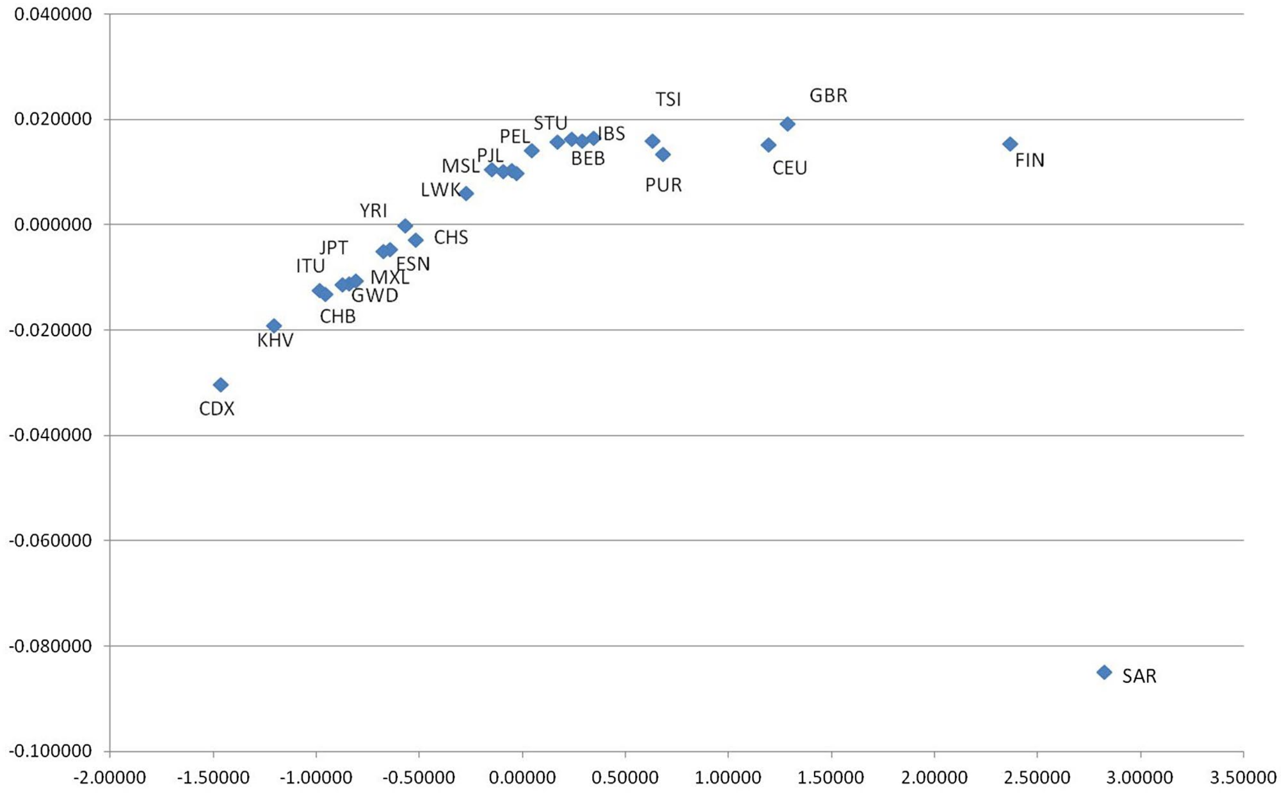

FIGURE 1 MDS of the rs4870723. Abbreviations as in Table 1

variation because of a founder effect and subsequent geographic isolation (Anagnostou et al., 2017; Francalacci et al., 2010; Francalacci & Sanna, 2008; Palo et al., 2009), and Dai turned to be an ethnic minority related to Lao-Thai people with peculiar genetic characteristics (Shi et al., 2010).

In conclusion, although there are significant differences between different populations for rs4870723, there is no evidence for the presence of a selective pressure and therefore the observed genetic variation is most probably due to random events such as genetic drift, founder effect and other demographic events that the populations have gone through. The different distribution of SNP rs4870723, despite correlating with tendinopathies, does not seem to affect population fitness, which implies different survival and reproduction ability.

Finally, despite having strengthened the idea that the European genetic variation is associated with geographic barriers (Novembre et al., 2008) and prehistoric demographic events (Ammerman & Cavalli-Sforza, 1984), our data highlighted the presence of some isolated population that increased the Eurasian variability (Anagnostou et al., 2017; Capocasa et al., 2014).

ACKNOWLEDGMENTS

This work was supported by the Italian Ministry of Research, MIUR (fund FIR to CCM and RR).

CONFLICT OF INTEREST

The authors state that they have no conflict of interest.

AUTHORS' CONTRIBUTION

Carla Maria Calò: Conceptualization. Federico Onali: data analysis. Renato Robledo: writing draft. Laura Flore: data analysis. Myosotis Massidda: writing review. Paolo Francalacci: critical revision.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Renato Robledo https://orcid.org/0000-0003-0661-9309

REFERENCES

Ammerman, A. J., & Cavalli-Sforza, L. L. (1984). *The Neolithic transition and the genetics of populations in Europe*. Princeton University Press.

Anagnostou, P., Dominici, V., Battaggia, C., Pagni, L., Vilar, M., Wells, S., Sarno, S., Boattini, A., Francalacci, P., Colonna, V., Vona, G., Calò, C., Destro, B. G., & Tofanelli, S. (2017). Overcoming the dichotomy between open and isolated populations using genomic data from a large European dataset. *Scientific Report*, *1*(7), 41614.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., & Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Geneics*, *40*, 340–345. https://doi.org/10.1038/ng.78

Berthod, F., Germain, L., Guignard, R., Lethias, C., Garrone, R., Damour, O., van der Rest, M., & Auger, F. A. (1997). Differential expression of collagens XII and XIV in human skin and in reconstructed skin. *Journal of Investigative Dermatology*, *108*, 737–742. https://doi.org/10.1111/1523-1747.ep12292122

Capocasa, M., Anagnostou, P., Bachis, V., Battaggia, C., Bertoncini, S., Biondi, G., Boattini, A., Boschi, I., Brisighelli, F., Calò, C. M., Carta, M., Coia, V., Corrias, L., Crivellaro, F., Dominici, V., Ferri, G., Francalacci, P., Franceschi, Z. A., Luiselli, D., … Destro, B. G. (2014). Linguistic, geographic and genetic isolation: a collaborative study on Italian populations. *Journal of Anthropological Sciences*, *92*, 201–231. https://doi.org/10.4436/jass92001

Casillas, S., Mulet, R., Villegas-Mirón, P., Hervas, S., Sanz, E., Velasco, D., Bertranpetit, J., Laayouni, H., & Barbadilla, A. (2018). Pophuman: the human population genomics browser. *Nucleic Acids Research*, *46*, D1003–D1010. https://doi.org/10.1093/nar/gkx943

Caso, E., Maestro, A., Sabiers, C. C., Godino, M., Caracuel, Z., Pons, J., Gonzalez, F. J., Bautista, R., Gonzalo, C. M., Caso-Onzain, J., Viejo-Allende, E., Giannoudis, V., Alvarez, S., Maietta, P., & Guerado, E. (2016). Whole-exome sequencing analysis in twin sibling males with an anterior cruciate ligament rupture. *Injury*, *47*(Suppl 3), S41–S50. https://doi.org/10.1016/S0020-1383(16)30605-2

Chiquet, M. (1999). Regulation of extracellular matrix gene expression by mechanical stress. *Matrix Biology*, *18*, 417–426. https://doi.org/10.1016/s0945-053x(99)00039-6

Excoffier, L., Laval, G., & Schneider, S. (2007). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, *1*, 47–50

Francalacci, P., Morelli, L., Useli, A., & Sanna, D. (2010). The history and geography of the Y chromosome SNPs in Europe: an update. *Journal of Anthropological Sciences*, *88*, 207–214.

Francalacci, P., & Sanna, D. (2008). History and geography of human Y-chromosome in Europe: a SNP perspective. *Journal of Anthropological Sciences*, *86*, 59–89.

Li, L., Sun, Z., Chen, J., Zhang, Y., Shi, H., & Zhu, L. (2020). Genetic polymorphisms in collagen-related genes are associated with pelvic organ prolapse. *Menopause*, *27*, 223–229. https://doi.org/10.1097/GME.0000000000001448

Massidda M., Kikuchi N., Calò C., Pushkarev V., Cieszczyk P., & Salvi M. (2018). COL14A1 rs4870723 and Anterior Cruciate Ligament Rupture in physically active people from four different countries. 35th World Congress of Sports Medicine, 12–15 September Rio de Janeiro, Brasil.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., Carlos, D., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*, 98–101. https://doi.org/10.1038/nature07331

Palo, J. U., Ulmanen, I., Lukka, M., Ellonem, P., & Sajantila, A. (2009). Genetic markers and population history: Finland revisited. *European Journal of Human Genetics*, *17*, 1336–1346. https://doi.org/10.1038/ejhg.2009.53

Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J., & Engelken, J. (2014). 1000 Genomes Selection Browser 1.0: A genome browser dedicated to signatures of natural selection in modern humans.

Nucleic Acids Research, *42*, D903–D909. https://doi.org/10.1093/nar/gkt1188

September, A. V., Posthumus, M., van der Merwe, L., Schwellnus, M., Noakes, T. D., & Collins, M. (2008). The COL12A1 and COL14A1 genes and Achilles tendon injuries. *International Journal of Sports Medicine*, *29*, 257–263. https://doi.org/10.1055/s-2007-965127

Shi, L., Yao, Y. F., Shi, L., Matsushita, M., Yu, L., Lin, Q. K., Tao, Y. F., Oka, T., Chu, J. Y., & Tokunaga, K. (2010). HLA alleles and haplotypes distribution in Dai population in Yunnan province, Southwest China. *Tissue Antigens*, *75*, 159–165. https://doi.org/10.1111/j.1399-0039.2009.01407.x

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74. https://doi.org/10.1038/nature15393

Young, B. B., Zhang, G., Koch, M., & Birk, D. E. (2002). The roles of types XII and XIV collagen in fibrillogenesis and matrix assembly in the developing cornea. *Journal of Cellular Biochemistry*, *87*, 208–220. https://doi.org/10.1002/jcb.10290

How to cite this article: Calò CM, Onali F, Robledo R, Flore L, Massidda M, Francalacci P. Worldwide variation of the *COL14A1* gene is shaped by genetic drift rather than selective pressure. *Mol Genet Genomic Med*. 2021;00:e1629. https://doi.org/10.1002/mgg3.1629