Bioinformatics, 32(11), 2016, 1697–1700
doi: 10.1093/bioinformatics/btw027
Advance Access Publication Date: 22 January 2016
Discovery Note

OXFORD

Genome analysis

Quantitative trait association study for mean telomere length in the South Asian genomes

Anna Hakobyan[1,2], Lilit Nersisyan[1,3] and Arsen Arakelyan[1,3,*]

[1]Bioinformatics Group, Institute of Molecular Biology NAS RA, Yerevan 0014, Armenia, [2]College of Science and Engineering, American University of Armenia, Yerevan 0019, Armenia and [3]Synopsys Inc., Yerevan 0026, Armenia

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on 27 May 2015; revised on 29 December 2015; accepted on 13 January 2016

Abstract

Motivation: Mean telomere length (MTL) is associated with cancers and age-related diseases, which necessitates identification of genomic and environmental factors that impact telomere length dynamics. Here, we present a pilot genome wide association (GWA) study for MTL in South Asian population using publicly available next generation whole genome sequences (WGS), both for MTL and genotype calculations.

Results: MTL in the studied population was not correlated with age, which is in accordance with previous reports. Further, we identified that individuals with Sikh religion had longer telomeres, which may be the result of complex interaction between genetic background and environmental factors. Finally, we identified 51 MTL-associated SNPs residing in five loci. The top ones were located in ADARB2 gene, which has previously been implicated with extreme old age.

Conclusion: Our results show that WGS data can be used in telomere length studies. In addition, we introduce novel loci implicated in MTL that may be worth considering in further telomere studies.

Contact: aarakelyan@sci.am

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Telomeres are repetitive stretches of DNA that cap the ends of chromosomes. In humans, telomeres are comprised of $(TTAGGG)_n$ tandem repeats and vary between 5 and 15 kb in length (Samassekou *et al.*, 2010). Among other functions, telomeres protect chromosome ends from being recognized as double-stranded DNA breaks. Each cell division results in shortening of telomeric sequences due to the end replication problem (Blackburn, 1991). In their study, Takubo *et al.* (2002) have calculated that in humans the annual loss of telomere length ranges between 20 and 60 bp. Telomere length can be restored in cells by telomerase enzyme, which is active in reproductive and stem cells, partially in leukocytes, as well as in certain types of tumors (Chen and Chen, 2011).

Significance of telomere length association with various conditions and diseases is clearly demonstrated through multiple studies. Shortening of telomeres is associated with aging, age-related diseases (Brouilette *et al.*, 2007) and certain cancer types (Sanders and Newman, 2013).

Identification of genetic and environmental factors impacting telomere length has been repeatedly addressed. From multiple loci implicated in association with telomere length, those including genes associated with telomerase were among the most validated ones. Codd *et al.* (2013) conducted a meta-analysis of 37 684 individuals and reported seven loci associated with leukocyte mean telomere length (MTL), including genes coding for telomerase RNA component (TERC), telomerase reverse transcriptase oligonucleotide/oligosaccharide-binding fold containing 1 protein (OBFC1), nuclear assembly factor 1 ribonucleoprotein (NAF1), regulator of telomere elongation helicase (RTEL1), which are involved in telomere biology, and two other loci including ACYP2 and ZNF208 genes. These genes, apart from ZNF208, were validated or had supportive evidence in a study within the COGS project (Pooley *et al.*, 2013).

Furthermore, Pooley *et al.* (2013) found novel telomere length association at 3p14.4 (close to PXK), at 6p22.1 (ZNF311) and at 20q11.2 (BCL2L1) loci. Another study of families with exceptional longevity analyzed 4289 individuals, and reported two loci (17q23.2 and 10q11.21) containing novel candidate genes, as well as validated TERC, MYNN and OBFC1 (Lee *et al.*, 2013). Association of telomere length with SNPs is partially population specific, and until now there is no well-accepted genomic factor determining telomere length and telomere attrition rate. This may be partially attributed to population-specific genomic variations on one side, and lack of tools for measuring telomere length from whole genome sequencing (WGS) data, on the other. Recently, we have developed a methodology and software Computel that allows to calculate MTL from WGS data by capturing reads from telomeric regions via alignment to a special reference sequence. Computel has been proven to be valid with reference to existing experimental methods and to outperform other computational approaches (Nersisyan and Arakelyan, 2015). This gives the opportunity to considerably promote the research aimed at understanding how telomere length is linked to genomic context. In this study, we have performed a genome wide quantitative trait association study for MTL in South Asian population, using WGS data both for genotyping and MTL calculation.

2 Materials and methods

2.1 Dataset

The study was conducted on datasets produced by the South Asian Genome project (Zhi *et al.*, 2014), deposited in European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession number PRJEB5476. We used a dataset containing 2 × 101 bp WGS reads from Illumina $GAII_x$ at 4× coverage, as well as corresponding genotypes from whole blood samples of 168 individuals. The individuals were sampled from different age groups, religious backgrounds and language groups, predominantly from India (166 India, one Kenya, one East Africa). Information on age, sex, religion and language was available for all the study subjects.

2.2 MTL calculation

MTL was calculated from 4×x coverage WGS reads exploiting a novel method implemented in the software Computel developed by our group (Nersisyan and Arakelyan, 2015), using its default parameters. MTL data are presented as *mean* MTL ± SD throughout the text. MTL association with age, sex and religion was evaluated using linear regression. *P* values <0.05 were considered significant. Multiple correction was performed with false discovery rate estimation (FDR) with Benjamini–Hochberg (BH) method. Statistical calculations were performed in R environment.

2.3 Population stratification

We have used an R package *GenABEL* (Aulchenko *et al.*, 2007) to analyze the population structure. For dimensionality reduction, we have performed multidimensional scaling (MDS) with 10 components using the EIGENSTRAT algorithm (Price *et al.*, 2006).

2.4 Association analysis

The association analysis was performed with *Plink* toolset (Purcell *et al.*, 2007). SNPs with MAF < 0.1 and HWE significance threshold < 0.1 were excluded, which left us with 4, 106, 441 SNPs with genotyping rates among samples >99.93%. We used multiple linear regression model to assess additive effect of minor alleles (0, 1 or 2 copies) on MTL for each SNP, with adjustment for age, sex, religion and top four principal components derived from population stratification analysis.

3 Results

3.1 Population structure

Principal component analysis revealed that the first component was explaining 1.68% of the variability in the data, and top four components were amounting to only 5%. However, data projection over the first and the second principal components demonstrated two distinct clusters along the first component. The majority of samples were descending from India, but the exact region of birth was unknown, thus, we tried to explore the clustering based on available subject details, namely language and religion (Fig. 1, Supplementary Table 1). Cluster 1 contained mainly Christian, Hindu, Muslim subjects of different languages, while cluster 2 presented predominantly Punjabi speaking individuals whose religion was Hindu or Sikh, implying that these individuals were apparently originating from Punjab region. This assumption is in agreement with the sample description provided by the South Asian Genome project (Zhi *et al.*, 2014).

3.2 MTL association with gender, age, language and religion

The MTL was calculated for all 168 samples. The values were ranging from 3234 to 8738 bp with the *mean* MTL ± SD equal to 5401 ± 1153 bp (Supplementary Fig. 1). Further analyses were performed on 166 individuals born exclusively in India. Linear regression of MTL with adjustment of age, sex and religion revealed no significant association with age ($P = 0.23$) and sex ($P = 0.29$), while Sikh religion was the only factor significantly affecting MTL ($P = 0.00451$). Moreover, Sikhs had significantly longer MTLs compared with the rest of the samples (mean difference 816.5 bp, 95%CI = 337.3–1295.7 bp). We then tested whether any differences in age distribution in the populations could account for deviations in MTL and no significant difference was found (two-sample *t*-test of age for Sikh versus Hindu $P = 0.3711$, Sikh versus the rest of samples $P = 0.08165$). In order to account for distinct genetic background of Sikhs as compared with other groups, we included principal components of population stratification analysis into regression model. The results showed no significant association of MTL with Sikh religion, suggesting the religion as an indicator for genetic diversity in studied samples (Supplementary Table 2). Moreover, multinomial regression model showed significant association of religion with the most discriminative PC1 component (Supplementary Table 3). However, comparison of MTL in Cluster 2 revealed that Sikhs have significantly longer MTL (*mean* MTL ± SD in Sikhs: 5891 ± 1343 bp, in other samples: 5115 ± 978 bp, *t*-test $P = 0.001$), which suggests more complex nature of influencing factors beyond genetic background (Supplementary Fig. 2).

In order to evaluate the bias introduced by Sikh samples in MTL-age association, we excluded those and rerun the regression analysis. The resulting model ($MTL = -14.46 \times \text{Age} + 5676.76$, $R^2 = 0.025$, $P = 0.09$) became consistent with previously reported association found in South Asians, where MTL was measured with real-time PCR (Bhupatiraju *et al.*, 2012).

3.3 MTL associated loci

We analyzed 4, 106, 442 filtered SNPs for MTL association, with adjustment for age, sex, religion and top four principal components ($\lambda = 1.002528$, see Supplementary Fig. 3 for the QQ-plot of association values). The Manhattan plot of filtered SNPs with $P < 1 \times 10^{-2}$

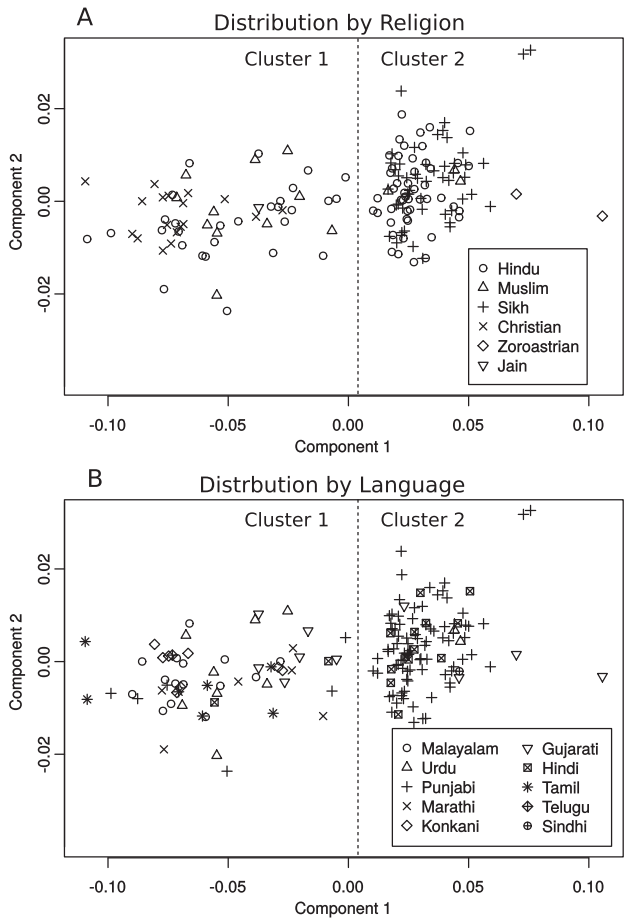

Fig. 1. MDS analysis of population stratification. Projection of samples over the first and the second principal components (four outliers excluded). (**A**) Distribution of religious groups. (**B**) Distribution of language groups. Clusters 1 and 2 depict the separation by PC1 component

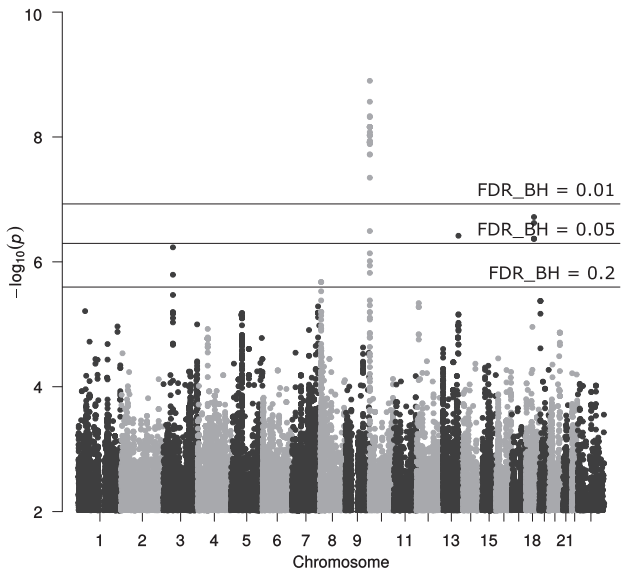

Fig. 2. Manhattan plot of SNPs with MTL association $P < 10^{-2}$. Horizontal lines represent FDR_BH thresholds for BH corrected *P* values

is presented in Figure 2. We used three thresholds for suggestive SNPs implicated in MTL (BH corrected $P < 0.01$, $P < 0.05$, $P < 0.2$), corresponding to 35, 42 and 51 SNPs, respectively (Supplementary Table 3). The top 35 SNPs were residing in the third intron of ADARB2 gene (10p15.3) with unadjusted $P = 4.49 \times 10^{-8}$. Alleles in this set were in strong linkage, with allele incidence correlation $R^2 = 0.87$ for all SNP pairs. The strongest MTL association was observed in the SNP *rs1500964* ($P = 1.26 \times 10^{-9}$). This SNP has two documented alleles (*C, T*) with *C* variant being associated with longer MTL. Heterozygous individuals (*CT*) had 812 bp longer MTL's than homozygous individuals with *T* allele ($P = 7.8 \times 10^{-6}$). Inclusion of second *C* allele was implying an MTL increase by 572 bp, though the *P* value did not reach significance level due to small number of individuals with CC genotype ($P = 0.079$). Other significant SNPs of ADARB2 gene had similar effects.

The second significance threshold of BH adjusted $P < 0.05$ extended the associated SNP list with another SNP from ADARB2 gene, and six SNPs from two additional regions. Five of these SNPs were located in 19p13.3 locus, in a close proximity to tumor necrosis factor (TNF) family and TNF ligand family proteins (TNFSF9, CD70, TNFSF14).

The broader list of SNPs with BH adjusted $P < 0.2$ was additionally encompassing SNPs from intergenic regions of 13q33.3, 8p23.1, loci (Supplementary Table 3) and *rs7643501* ($P = 4.937 \times 10^{-7}$) in an intronic region of CACNA2D3 gene (13p14.3), which is a putative tumor suppressor (Wong *et al.*, 2013).

Similar results were obtained when all 168 samples (including two samples out of India) were used in the analysis (Supplementary Text 1 and Supplementary Table 4).

The results of statistical power calculations for associated SNPs can be found in Supplementary data (Supplementary Text 2 and Supplementary Table 5). It should be noted that we did not observer significant differences in allele frequencies of SNPs associated with MTL in Sikh individuals compared with the rest of the study group (Supplementary Table 6).

4 Discussion

Our findings revealed that MTL in the studied population is not correlated with age, which is in agreement with previously published results on South Asian population, where quantitative PCR was used to measure relative telomere length (Bhupatiraju *et al.*, 2012). These results deviate from other studies on different populations (samples recruited from Austria, France, China and Denmark) that showed strong correlation of MTL with age (Benetos *et al.*, 2001; Hochstrasser *et al.*, 2012; Li *et al.*, 2014; Rode *et al.*, 2015), thus suggesting that the association might be population specific (Diez Roux *et al.*, 2009; Zhu *et al.*, 2011).

Surprisingly, we have observed that Sikhs have significantly longer telomeres compared with the rest of the samples. The results of regression analysis with adjustment for genetic background suggested about possibility of complex influence of genetic background and environmental factors. It is worth noting that smoking, drug taking and using tobacco are banned in Sikhism, alcohol is rarely consumed and many Sikhs are lifelong vegetarians. Unfortunately, we could not assess the effects of these factors due to lack of details regarding lifestyle of sampled individuals. However, the impact of oxidative stress, including smoking, and lifestyle on telomere length was repeatedly investigated, indicating that telomere length is negatively affected by smoking, while the impact of healthy lifestyle is positive (Cassidy *et al.*, 2010). Additional investigations are needed to assess whether long telomeres are characteristic to Sikh population, and are preserved due to high heritable nature of telomere lengths (Hjelmborg *et al.*, 2015).

The genome wide association (GWA) scan of the South Asian population revealed 51 SNPs associated with MTL. Among these, the

most significant ones were residing in ADARB2 gene. This is an RNA editing gene of double-stranded RNA adenosine deaminase family. Sebastiani *et al.* (2009) detected 10 SNPs in ADARB2 gene, strongly associated with extreme longevity . Two of those SNPs *rs10903420* ($P = 1.199 \times 10^{-8}$) and *rs1007147* ($p = 6.908 \times 10^{-9}$) were among the most implicated SNPs in MTL, reported in our study. Additional three SNPs (*rs2805562*, *rs884949* and *rs2805535*) had small association *P* values, but did not reach the significance threshold. In total, five SNPs of ADARB2, associated with extreme old age, are also supposedly associated with MTL. These findings call for further investigation aimed at understanding molecular mechanisms through which ADARB2 is involved in telomere length regulation and longevity.

At this point of discussion it is worthwhile to identify several limitations of our study. The first and the most important limitation in GWA analysis was the small sample size, which we had access to. This dramatically reduced the GWA power to detect associated SNPs, thus, many functionally relevant SNPs did not reach the significance threshold after multiple correction. Further, we had limited information about the samples: BMI, smoking habits and other lifestyle details could serve for adjustment in the association analysis phase and give some important insights over the MTL aberrations. The next limitation was the inhomogeneity of the studied population. Even though sampled individuals were descending from South Asia, the genetic footprints of different religious groups were considerably diverse.

5 Conclusion

Here, we have presented a pilot study of GWA for MTL in South Asians, where we exploited WGS data for SNP information and for MTL calculation. Concordance of certain findings with previously published results validated our approach and demonstrated that the usage of WGS data can be extended to be utilized in telomere studies. This eliminates the necessity of conducting additional experiments for telomere length measurement, greatly facilitating further research. Moreover, there is large amount of already existing WGS data focused on age-related diseases and cancers, where telomeres play an important role, and our approach can be used to exploit available datasets to enrich the results with telomere length data. Our study showed that Sikhs are distinguished with longer telomeres compared with other religious groups from South Asia. This phenomenon needs further investigation in order to assess the involvement of genetic and/or environmental factors on telomere length dynamics. Moreover, our results suggest that not only telomere length, but also its association with age can be affected by ethnicity.

Finally, we have identified that ADARB2 gene highly impacts telomere length in South Asians. Longevity-related nature of this gene is characterized in other populations, thus combination of this information with our results calls for more investigation to understand the role of this gene in telomere regulation and ageing in general.

Acknowledgements

The authors express their acknowledgment to the South Asian Genome Project contributors for making their data publicly available and The Bioinformatics Centre of University of Copenhagen for providing computational resources for calculations.

Conflict of Interest: none declared.

References

Aulchenko,Y.S. *et al.* (2007) Genabel: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

Benetos,A. *et al.* (2001) Telomere length as an indicator of biological aging the gender effect and relation with pulse pressure and pulse wave velocity. *Hypertension*, **37**(2 Pt 2), 381–385.

Bhupatiraju,C. *et al.* (2012) Association of shorter telomere length with essential hypertension in Indian population. *Am. J. Hum. Biol.*, **24**, 573–578.

Blackburn,E. (1991) Structure and function of telomeres. *Nature*, **350**, 569–573.

Brouilette,S.W. *et al.* (2007) Telomere length, risk of coronary heart disease, and statin treatment in the west of Scotland primary prevention study: a nested case–control study. *Lancet*, **369**, 107–114.

Cassidy,A. *et al.* (2010) Associations between diet, lifestyle factors, and telomere length in women. *Am. J. Clin. Nutr.*, **91**, 1273–1280.

Chen,C.H. and Chen,R.J. (2011) Prevalence of telomerase activity in human cancer. *J. Formosan Med. Assoc.*, **110**, 275–289.

Codd,V. *et al.* (2013) Identification of seven loci affecting mean telomere length and their association with disease. *Nat. Genet.*, **45**, 422–427.

Diez Roux,A.V. *et al.* (2009) Race/ethnicity and telomere length in the multiethnic study of atherosclerosis. *Aging Cell*, **8**, 251–257.

Hjelmborg,J.B. *et al.* (2015) The heritability of leukocyte telomere length dynamics. *J. Med. Genetics*, **52**, 297–302.

Hochstrasser,T. *et al.* (2012) Telomere length is age-dependent and reduced in monocytes of Alzheimer patients. *Exp. Gerontol.*, **47**, 160–163.

Lee,J.H. *et al.* (2013) Genome wide association and linkage analyses identified three loci-4q25, 17q23. 2, and 10q11.21-associated with variation in leukocyte telomere length: the long life family study. *Front. Genet.*, **4**. 310.

Li,Z. *et al.* (2014) Shorter telomere length in peripheral blood leukocytes is associated with childhood autism. *Sci. Rep.*, **4**, 7073.

Nersisyan,L. and Arakelyan,A. (2015) Computel: computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS One*, **10**, e0125201.

Pooley,K.A. *et al.* (2013) A genome-wide association scan (GWAS) for mean telomere length within the cogs project: identified loci show little association with hormone-related cancer risk. *Hum. Mol. Genet.*, **22**, 5056–5064.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rode,L. *et al.* (2015) Peripheral blood leukocyte telomere length and mortality among 64 637 individuals from the general population. *J. Nat. Cancer Inst.*, **107**, djv074.

Samassekou,O. *et al.* (2010) Sizing the ends: normal length of human telomeres. *Ann. Anat.*, **192**, 284–291.

Sanders,J.L. and Newman,A.B. (2013) Telomere length in epidemiology: a biomarker of aging, age-related disease, both, or neither? *Epidemiol. Rev.* **35**, 112–131.

Sebastiani,P. *et al.* (2009) Rna editing genes associated with extreme old age in humans and with lifespan in *C. elegans*. *PLoS One*, **4**, e8210.

Takubo,K. *et al.* (2002) Telomere lengths are characteristic in each human individual. *Exp. Gerontol.*, **37**, 523–531.

Wong,A.M.G. *et al.* (2013) Characterization of cacna2d3 as a putative tumor suppressor gene in the development and progression of nasopharyngeal carcinoma. *Int. J. Cancer*, **133**, 2284–2295.

Zhi,D. *et al.* (2014) The South Asian genome. *PLoS One*, **9**, e102645.

Zhu,H. *et al.* (2011) Leukocyte telomere length in healthy Caucasian and African-American adolescents: relationships with race, sex, adiposity, adipokines, and physical activity. *J. Pediatr.*, **158**, 215–220.