Telomere length measurement for longitudinal analysis: implications of assay precision

Daniel Nettle, Shahinaz M. Gadalla, Tsung-Po Lai, Ezra Susser, Melissa Bateson, Abraham Aviv

In press, *American Journal of Epidemiology*

Correspondence to Daniel Nettle, Newcastle University Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK, daniel.nettle@ncl.ac.uk

ABSTRACT

Researchers increasingly wish to test hypotheses concerning the impact of environmental or disease exposures on telomere length (TL), and use longitudinal study designs to do so. In population studies, TL is usually measured using a qPCR-based method, which has been validated by presenting a correlation with a gold standard method such as Southern blotting (SB) in cross-sectional datasets. However, in a cross-section, the range of true variation in TL is large, and measurement error is introduced only once. In a longitudinal study of an environmental effect, the target variation of interest is small, and measurement error is introduced both at baseline and follow-up. We present a small dataset (n = 20) where leukocyte TL was measured 6.6 years apart by both qPCR and SB. The cross-sectional correlations between qPCR and SB were high both at baseline (r = 0.90) and follow-up (r = 0.85), yet their correlation for TL change was poor (r = 0.48). Moreover, the qPCR but not SB data showed strong signatures of measurement error. Through simulation, we show that the statistical power gain from performing a longitudinal analysis is much greater for SB than qPCR. We discuss implications for optimal study design and analysis.

Key words: telomere length, longitudinal, measurement error, qPCR, Southern Blot, TRF, assay precision

Running head: Telomere length measurement for longitudinal analysis

INTRODUCTION

Recent decades have witnessed an explosion of telomere epidemiology research. The average telomere length (TL), usually measured in leukocytes (hence LTL), has been associated with a wide range of environmental exposures, diseases, and psychosocial parameters (see 1 for review). The first wave of such studies was almost entirely cross-sectional, but as the field has matured, attention has turned to longitudinal studies. As well as being more informative about possible causal relationships (see (2,3)), longitudinal studies are potentially more statistically powerful for testing hypotheses in telomere epidemiology (4). The key difference between a cross-sectional and a longitudinal study, from a purely analytic standpoint, is that in the former, the target parameter is average LTL, whereas in the latter, it is the average change in LTL within individuals over time. There is a large range of variation between individuals in LTL. This variation is stable over time in adulthood and substantially heritable (5,6), but it is effectively noise with respect to many of the hypotheses researchers wish to investigate. It is controlled for in longitudinal studies by making within-individual comparisons.

In telomere epidemiology, LTL is typically measured by a qPCR-based relative TL measurement technique (7). This is an inexpensive and high-throughput method allowing for large-scale studies. The method estimates the amount of the telomeric DNA sequence present in a sample (T), relative to the amount of a single-copy gene sequence (whose copy number in the genome does not vary; S). Validation of the qPCR method has been demonstrated by correlating T/S values for a set of samples with LTL measured by a 'gold standard' method, usually Southern blotting (SB). When performed in experienced laboratories, these correlations can be high ($r \geq 0.85$) (7–9). Researchers therefore reason that qPCR LTL measurement captures substantially the same variation as the gold standard. However, demonstrating that qPCR measurements are highly correlated to SB for a cross-section of individuals does not guarantee that the two methods will capture the same variation in a longitudinal analysis, nor that they are equally powerful for testing hypotheses about environmental and other effects on LTL. First, in a longitudinal analysis there are at least two LTL measurements, baseline and follow-up. Thus, the measurement error is introduced twice. As long as these errors are uncorrelated, the spurious variance introduced is twice as large as in a cross-sectional analysis. Second, in a cross-section of individuals, the range of true biological variation is very large: the standard deviation of LTL across individuals in adult humans is about 700 base pairs (bp), with the most extreme individuals differing by 3000-

4000 bp (10–12). However, much of this variation, reflecting individual differences in telomere length already evident at birth, is irrelevant to hypotheses about environmental or aging effects on telomere dynamics in adulthood. The change in LTL over time within adult individuals is only around 25-30 bp/year on average (10,12). This means that even an exposure that doubles the rate of telomere attrition will only change average LTL by a few tens of bp/year. This is a very small target relative to the range of variation in LTL. Accordingly, the effective precision of qPCR to detect effects of an exposure on telomere dynamics may be much lower than the high cross-sectional correlation with SB seems to imply.

In two-measurement longitudinal LTL studies where the effect of some exposure or treatment X is of interest, researchers have a number of options for data analysis strategy (13,14). One strategy is to simply test whether LTL at the final time point differs by X. This ignores the baseline information, in effect treating the longitudinal study as a cross-sectional one. We henceforth refer to it as the cross-section approach. A second strategy (henceforth difference score) is to calculate ΔLTL, the difference in LTL between baseline and follow-up, and test whether ΔLTL differs by X. Finally, the analysis of covariance (ANCOVA) strategy tests whether LTL at follow-up differs by X, with LTL at baseline included in the model as a covariate (though this approach produces biased estimates when baseline LTL is associated with X and there is measurement error; see (15,16) and Discussion).

The relative statistical powers of the three approaches depend on the correlation between baseline and follow-up LTL (14). If this correlation is close to zero, cross-section is more powerful than difference score, and as powerful as ANCOVA, whereas if the correlation is high, the power of the cross-section approach is much lower than the other two. The correlation between baseline and follow-up LTL is generally lower for qPCR than SB, exactly because the measurement error is greater (12,17). Thus, the relative power advantages of the different analysis strategies may be different for SB and qPCR data.
Here, we investigate for the first time the validity of qPCR measurement for capturing longitudinal LTL dynamics, and consider the potential implications of the findings for study design and analysis. We present a small longitudinal dataset (20 men whose LTL was measured twice, 6.6. years apart, by both techniques). We investigate the correlation between qPCR and SB not just for LTL at baseline and follow-up, but also for ΔLTL. Further, we compare the qPCR and SB data for known signatures of measurement error, namely a strong

apparent dependence of ΔLTL on baseline LTL due to regression to the mean (17,18), and a substantial fraction of individuals whose LTL appears to lengthen rather than shorten over time (10). We expect these signatures to be much more marked for the qPCR than the SB data. We then simulate datasets with the same cross-sectional correlations between qPCR and SB at baseline and follow-up as our empirical data have. This allows us to verify that the observed features of the qPCR estimates of TL dynamics are not quirks of one small dataset, but should be expected more generally. Finally, we use the simulated datasets to investigate the implications of qPCR's lower precision for statistical power to detect an effect on TL dynamics.

METHODS

Empirical dataset

We used samples from the Stony-Brook-University biorepository. These were collected on two occasions, baseline and follow-up, 6.6 ± 0.5 (SD) years apart, for twenty white males, aged 53.8 ± 4.3 years at follow-up. Donors consented for IRB-approved studies. LTL was measured both at baseline and follow-up by both qPCR and SB independently and blindly in different laboratories, qPCR at the National Cancer Institute Cancer Genomics Research laboratory, and SB at the laboratory of the Center of Human Development and Aging at Rutgers University. Each laboratory followed its standard TL measurement protocol (19,20).

LTL measurements

DNA was extracted by Gentra Puregene Blood Kit (Qiagen, Valencia, CA) and all samples passed DNA integrity tests (19). For SB, a cocktail of the restriction enzymes *HinfI* (10 U) and *RsaI* (10 U) was used to generate the terminal restriction fragments (TRFs). Measurements were carried out in duplicate and resolved on different gels. The intra-class correlation (ICC) for duplicates was 0.93 (95% CI 0.87-0.96). Digested DNA samples and DNA ladders were resolved on 0.5% agarose gels. After 16 h, the DNA was depurinated for 15 min in 0.25 N HCl, denatured 30 min in 0.5 M NaOH/1.5 M NaCl and neutralized for 30 min in 0.5 M Tris, pH 8/1.5 M NaCl. The DNA was transferred to a positively charged nylon membrane (Roche) for 1 h using a vacuum blotter. Membranes were hybridized at 65^{o} C with the DIG-labeled telomeric probe as previously described (19). The DIG-labeled probe was detected by DIG luminescence and exposure on X-ray film.

For qPCR, we used the monoplex method adopted from (21). Details have been described elsewhere (20). Briefly, PCR telomere primers were *Telo_FP* [5'-CGGTTT(GTTTGG)5GTT-3'] and *Telo_RP* [5'-GGCTTG(CCTTAC)5CCT-3']. Primers for the single-copy gene (*36B4*) were *36B4_FP* [5'-CAGCAAGTGGGAAGGTGTAATCC-3'] and *36B4_RP* [5'-CCCATTCTATCATCAACGGGTACAA-3']. The ratio of telomere signal concentration (T) to that of the single-copy gene (S; *36B4*) T/S ratios were normalized by the average T/S ratio obtained from the internal QC calibrator samples on the same plate. All telomere and *36B4* reactions were run in triplicate and the average of the measurements was used for all calculations. The intra-class correlation coefficient (ICC) of T/S ratios from the triplicates separately was 0.81 (95% confidence interval 0.70 – 0.89).

Simulations

We created simulation code to produce datasets that shared key properties with our empirical one. Simulation code runs in R (22) and is available via the Zenodo repository at: https://zenodo.org/record/3929509. In each simulated dataset, baseline (SB) LTL and ΔLTL were each drawn from distributions with the mean and standard deviation observed in the empirical SB data. We then generated qPCR values at each of baseline and follow-up whose correlations with SB were as observed in the empirical data.

Using the simulation code, we first created 1000 datasets of the same sample size as the empirical data, to investigate the extent to which the empirical patterns should be expected to recur in other samples. Next, we simulated 1000 datasets at each of a range of sample sizes (20-1000), where a predictor variable X with a true effect on telomere attrition was applied to half the individuals. The scenario we have in mind is an experiment or randomized control trial. Thus, we assume that there was no association between X and baseline LTL, and that individuals differing in X are not different in any other systematic way relevant to their LTL dynamics. We then analyzed the simulated datasets using each of the three data analysis strategies described in the Introduction (cross-section, difference score, ANCOVA), to establish the power of each strategy to detect the effect of X at $P < 0.05$. This power analysis was applied both to the simulated SB values and the simulated qPCR values. The effect size of X was set to 0.5 standard deviations, a medium effect by conventional criteria (23). We repeated the simulations with smaller effect sizes, with no change to the qualitative conclusions.

RESULTS

Empirical dataset

In the empirical data, the correlations between SB and qPCR LTL were very high at both time points: r = 0.90 ($P < 0.001$) for baseline; r = 0.85 ($P < 0.001$) for follow-up (Fig 1a, Fig 1b). The correlation between SB and qPCR for ΔLTL was r = 0.48 ($P = 0.03$; Fig 1c). This was significantly lower than both the baseline (z = 2.77, $P < 0.01$) and follow-up (z = 2.25, $P = 0.02$) correlations.

The SB data showed a mean of 0.19 kilobase (kB) LTL shortening between baseline and follow-up, equivalent to 0.42 SD of the baseline LTL variation. This would be considered significant shortening by conventional criteria, even in this small sample (t-test against 0: t=-5.05, $P <0.001$). By contrast, the qPCR data showed average LTL shortening of 0.12 SD of the baseline LTL variation. This would be considered non-significant shortening by conventional criteria (t-test against 0: t = -1.19, $P = 0.25$).

In the qPCR data, ΔLTL depended negatively on baseline LTL (r = -0.64, $P = 0.03$, Fig 2a), whilst in the SB data, the correlation between ΔLTL and baseline LTL was weak and not significantly different from zero (r = -0.06, $P = 0.82$, Fig 2b). The difference between these two correlations was significant (z = 2.04, $P = 0.04$). In the qPCR data, those with relatively short LTL at baseline tended to show LTL lengthening, whilst only those with relatively long LTL at baseline showed shortening. Of those individuals whose baseline LTL by qPCR was below the mean, 7 of 11 showed apparent lengthening; whereas of those individuals whose baseline LTL by qPCR was above the mean, 8 of 9 showed apparent shortening. In the SB data, by contrast, seventeen individuals showed shortening and only three apparent lengthening, with those three having neither particularly long nor particularly short baseline LTL.

Simulated datasets

We created 1000 simulated datasets of n=20 with the correlation between qPCR and SB at both baseline and follow-up at 0.875 (the mean of the empirically observed baseline and follow-up values). We confirmed that in these datasets, the correlation between qPCR and SB for ΔLTL was always much lower than at either baseline or follow-up (median correlation

0.42, interquartile range 0.26 to 0.55). Similarly, the simulated datasets consistently showed marked signatures of measurement error in the qPCR but not SB data. There were consistently negative correlations between baseline LTL and ΔLTL for qPCR (median correlation -0.38, interquartile range -0.50 to -0.25) but not SB (median correlation 0.01, interquartile range -0.15 to 0.18). The percentage of apparent lengtheners was higher for qPCR than SB in 99.4% of simulated datasets (qPCR: median 50%, interquartile range 40% to 55%; SB: median 15%, interquartile range 10% to 25%). The properties of the empirical dataset fell well within the range of simulated datasets in all cases (Fig 3).

We then took a range of sample size from 20 to 1000, and simulated 1000 datasets at each one. In these simulations, there was a true effect of a predictor variable on TL attrition. We assumed an effect size of $d = 0.5$. This would be conventionally considered a medium effect (23), and would lead to an 85 bp difference in LTL on average at follow-up. We plotted the power to detect this true effect at $P < 0.05$ for each of the three analysis strategies, for the SB and qPCR data (figure 4). For SB, there was a dramatic power gain from incorporating the baseline information (compare the power of cross-section to difference score and ANCOVA). For qPCR this power gain was much more modest. Note that the power of qPCR and SB for cross-section was similar under these assumptions. However, the power for either of the longitudinal analyses was very much greater for SB than qPCR. In addition, for SB, the powers of difference score and ANCOVA were almost identical, whereas for qPCR, there was a small but consistent power advantage for ANCOVA over difference score.

DISCUSSION

We have presented the first dataset in which LTL was measured at two time points in the same individuals by both SB and qPCR. The correlations between the two methods at both baseline and follow-up were high (0.90 and 0.85); similar correlations in the past have been taken as validating that qPCR has sufficient precision for population studies (7–9). However, the correlation between SB and qPCR for ΔLTL, the change in LTL over the 6.6 years of the study, was only 0.48. Estimating change in LTL involves detecting a smaller range of true biological variation than estimating LTL in a cross-section, and introduces twice as much measurement error. Thus, the correlation of a technique with substantial measurement error to the gold standard is bound to be much lower when the target parameter is LTL change rather than LTL.

We also found that the qPCR data showed known signatures of measurement error much more strongly than the SB data. By qPCR, apparent change in LTL depended negatively on LTL at baseline, and a substantial fraction of individuals showed apparent lengthening. These are predictable patterns due to regression to the mean in data sets containing measurement error (10,17,18). Again, simulations showed them to be consistently more marked in qPCR datasets where the cross-sectional correlations to SB are high but not perfect. Even with a small sample size (n=20), we were able to detect significant LTL shortening over the 6.6 years of the study by SB. Indeed, the estimated rate (190 bp over 6.6 years implies 29 bp per year) accords well with previous estimates (10,12). From the qPCR data, the researcher would have concluded telomeres had not shortened on average. Thus, although qPCR LTL estimates were highly correlated to SB estimates at both baseline and follow-up, relying on the qPCR data alone would have led to radically different conclusions about TL dynamics in this sample than the SB data. Validating qPCR against SB cross-sectionally is therefore insufficient to infer that its precision is adequate for a longitudinal study of effects on TL dynamics.

There are implications of our findings are as follows. First, since qPCR has not been validated as a measure of LTL for longitudinal use, only cross-sectionally, some skepticism may be in order about some published longitudinal findings by qPCR. Most obviously, apparently null effects, like the apparent non-shortening with age in our empirical dataset, may reflect false negatives due to low assay precision. However, there are also circumstances where the imprecision of qPCR carries substantial risks of false positive findings as well. An effect of measurement error in longitudinal data, as we have demonstrated, is to make apparent LTL change strongly dependent on initial LTL. Any group of individuals whose LTL at baseline is relatively short will appear to lengthen, often quite substantially. For example, one study concluded that gastric bypass surgery led to LTL lengthening in most cases (24). However, close examination shows that significant LTL lengthening was restricted to those individuals whose baseline LTL was the shortest (see (25)). Another study concluded that individuals with the greatest exposure to chronic stress showed the least LTL shortening over ten years (26). Again, these individuals were also the ones with shortest LTL at baseline. We emphasize that our empirical qPCR assays were performed by an experienced laboratory in the field, and the cross-sectional correlations with SB were at the high end of what has been previously published. There are grounds for believing that many published

qPCR datasets have much lower precision than observed here (see (12,17)). It is likely, therefore, that the issues we document are present and perhaps even more severe in the published literature.

Second, researchers using qPCR have a number of options to reduce the effective measurement error: increasing the number of technical replicates; specific corrections for plate and well position; consistency of sample storage and DNA extraction and integrity; and controlling for variable amplification efficiency. These have been discussed in detail elsewhere (7,17,27–30). Since the consequences of a small amount of imprecision are more dramatic for TL change than they are for TL itself, the point of diminishing returns in the application of these measures may actually be much higher than researchers appreciate. The fact that qPCR data correlate to a gold standard at r = 0.90 for a cross-section may appear 'good enough', and, indeed, that may be so in cross-sectional analyses. However, as we have shown, if the goal of the study is to detect subtle changes in TL over time, it may not be. In the simulations underlying Figure 4, raising the cross-sectional correlation to r = 0.95 has a quite dramatic effect on the power of qPCR for difference score and ANCOVA.

Third, our power simulations have implications for optimal choices of telomere measurement method and data analysis strategy. The cross-section analysis strategy used in our simulations makes no use of the baseline information, and so captures the statistical power consequences of performing a purely cross-sectional study. The other two strategies represent different ways of incorporating longitudinal information. Our simulations show that, for SB data, there is a dramatic statistical power advantage for performing longitudinal analysis. As argued elsewhere (4), for SB, within-individual, longitudinal comparison will provide much increased sensitivity for detecting factors that affect TL dynamics, even if the expense is considerable and the resulting sample size is smaller. For qPCR, the statistical power gain from longitudinal comparison, though still present, is much more modest, as our simulations show. The large gain from performing within-individual comparisons is partially offset by the loss of introducing a second set of measurement error. We note, of course, that increasing statistical power is not the only motive for choosing a longitudinal analysis. There are others, such as eliminating reverse causality (see (2)).

Simply put, the statistical power of qPCR may be nearly as high as SB for a cross-sectional study. So, if using qPCR makes a larger sample size possible, it could represent a beneficial

decision. From the simulated data in Fig. 3, the power of a cross-sectional qPCR study with 500 individuals is better than the power of the equivalent SB study with 250 individuals. However, the researcher planning a longitudinal analysis might usefully consider using SB or another precise method, even if that will entail a much smaller achieved sample size. In our simulations, for either of the longitudinal analyses, the power of a qPCR study with 1000 individuals is still worse than the equivalent SB study with 250. Thus, optimal decisions about TL measurement method and sample size are linked to those about study design and objectives.

We also noted some differences between the statistical powers of the two longitudinal data analysis strategies, difference score and ANCOVA. For SB, the power of these two is almost identical. For this reason, researchers should choose difference score, since ANCOVA introduces collider bias if baseline LTL differs by the predictor variable, whereas difference score is not vulnerable to such bias (15). The similar power of the two approaches by SB accords with previous simulations, which show that the power of difference score and that of ANCOVA converge when the correlation of baseline and follow-up measurements is high (14). For qPCR, ANCOVA provides a small but consistent power advantage over difference score. Thus, ANCOVA should be used as long as the association between the predictor variable and baseline LTL is close to zero (i.e. the judgement should be made on the basis not of statistical significance, but the size of the parameter estimate). If there is any non-zero association, difference score should be selected, in order to avoid collider bias (and baseline LTL should not be included as an additional covariate, as is sometimes seen) (15). Note that researchers sometimes fit mixed models with LTL as the outcome variable, testing whether there is an interaction between time point and the predictor variable. This is equivalent to difference score from the standpoint of power and collider bias (15).

For our statistical power simulations, we treated SB as if it directly reflected the true TL. This makes sense in as much as we were using SB here as a gold standard against which qPCR could be validated. In reality, however, SB too involves some measurement error, and the true TLs are not known. Thus, our simulation results should be taken as reflecting the *relative* statistical powers of SB and qPCR for different analysis strategies (or, indeed, the statistical power of qPCR relative to any gold standard the researcher might use). The absolute values of the SB power given may be overestimates. Moreover, effects in telomere epidemiology

tend to be much weaker than that illustrated in Fig 4 (1). The simulation code provided in conjunction with this paper allows users to rerun the simulations for different effect sizes. In conclusion, demonstrating that a measurement method captures much of the same variation in TL as the gold standard method in a cross-section does not entail it will capture the same variation in TL change in a longitudinal one. We have shown empirically that, in a dataset where the cross-sectional correlations between qPCR and SB are high at both baseline and follow-up, the agreement between the two methods for TL change over time is fairly poor, and the qPCR data show stronger signatures of measurement error than the SB data. As we have shown, this has implications for researchers choosing how to allocate their research resources to a suitable combination of study design, measurement method, sample size, and data analysis strategy.

ACKNOWLEDGEMENTS

Author affiliations: Newcastle University Population Health Sciences Institute, Newcastle upon Tyne, UK (Daniel Nettle); Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA (Shahinaz Gaddalla); Mailman School of Public Health, Columbia University, New York, New York, USA and New York State Psychiatric Institute, New York, NY, USA (Ezra Susser); Newcastle University Biosciences Institute, Newcastle upon Tyne, UK (Melissa Bateson); Center of Human Development and Aging, New Jersey Medical School, Rutgers State University of New Jersey, Newark, New Jersey, USA (Tsung-Po Lai, Abraham Aviv). New York State Psychiatric Institute, New York, NY, USA

This was work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement AdG 666669, COMSTAR); by the intramural program of the National Cancer Institute at the National Institutes of Health; and by the National Institutes of Health (R01HL134840, U01AG066529) and the Norwegian Research Council (NFR; ES562296). The authors acknowledge the research contributions of the Cancer Genomics Research Laboratory funded with Federal funds from the National Cancer Institute, National Institutes of Health, under NCI Contract No. 75N910D00024. 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 U.S. Government.

Data and code relating to this study are freely available at: https://zenodo.org/record/3929509

The authors declare that they have no conflicts of interest.

REFERENCES

1. Pepper G V, Bateson M, Nettle D. Telomeres as integrative markers of exposure to stress and adversity: A systematic review and meta-analysis. *R. Soc. Open Sci.* 2018;5:180744.
2. Bateson M, Nettle D. Why are there associations between telomere length and behaviour? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 2018;373:20160438.
3. Bateson M, Aviv A, Bendix L, et al. Smoking does not accelerate leucocyte telomere attrition: A meta-analysis of 18 longitudinal cohorts. *R. Soc. Open Sci.* 2019;6:190420.
4. Aviv A, Valdes AM, Spector TD. Human telomere biology: Pitfalls of moving from the laboratory to epidemiology. *Int. J. Epidemiol.* 2006;35(6):1424–1429.
5. Benetos A, Kark JD, Susser E, et al. Tracking and fixed ranking of leukocyte telomere length across the adult life course. *Aging Cell*. 2013;12(4):615–621.
6. Broer L, Codd V, Nyholt DR, et al. Meta-analysis of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect. *Eur. J. Hum. Genet.* 2013;21(10):1163–1168.
7. Cawthon RM. Telomere measurement by quantitative PCR. *Nucleic Acids Res.* 2002;30(10):1–6.
8. Aviv A, Hunt SC, Lin J, et al. Impartial comparative analysis of measurement of leukocyte telomere length/DNA content by Southern blots and qPCR. *Nucleic Acids Res.* 2011;39(20):3–7.
9. Martin-Ruiz CM, Baird D, Roger L, et al. Reproducibility of telomere length assessment: An international collaborative study. *Int. J. Epidemiol.* 2015;44(5):1673–1683.
10. Steenstrup T, Hjelmborg JVB, Kark JD, et al. The telomere lengthening conundrum--artifact or biology? *Nucleic Acids Res.* 2013;41(13):e131.
11. Aviv A, Valdes AM, Spector TD. Human telomere biology: Pitfalls of moving from the laboratory to epidemiology. *Int. J. Epidemiol.* 2006;35(6):1424–1429.
12. Bateson M, Nettle D. The telomere lengthening conundrum - it could be biology. *Aging Cell*. 2017;16:312–9.
13. Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. *Br. Med. J.* 2001;323(7321):1123–1124.
14. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled

trial is statistically inefficient: A simulation study. *BMC Med. Res. Methodol.* 2001;1:1–4.

15. Bateson M, Eisenberg DTA, Nettle D. Controlling for baseline telomere length biases estimates of the effect of smoking on leukocyte telomere attrition in longitudinal studies. *R. Soc. Open Sci.* 2019;6:190937.
16. Oakes JM, Feldman HA. Statistical power for nonequivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Eval. Rev.* 2001;25(1):3–28.
17. Nettle D, Seeker L, Nussey D, et al. Consequences of measurement error in qPCR telomere data: A simulation study. *PLoS One*. 2019;14(5).
18. Verhulst S, Aviv A, Benetos A, et al. Do leukocyte telomere length dynamics depend on baseline telomere length? An analysis that corrects for "regression to the mean." *Eur. J. Epidemiol.* 2013;28:859–866.
19. Kimura M, Stone RC, Hunt SC, et al. Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat. Protoc.* 2010;5:1596–1607.
20. Gadalla SM, Wang T, Dagnall C, et al. Effect of recipient age and stem cell source on the association between donor telomere length and survival after allogeneic unrelated hematopoietic cell transplantation for severe aplastic anemia. *Biol. Blood Marrow Transplant.* 2016;22:2276–2282.
21. Callicott RJ, Womack JE. Real-time PCR assay for measurement of mouse telomeres. *Comp. Med.* [electronic article]. 2006;56(1):17—22. (http://europepmc.org/abstract/MED/16521855)
22. R Core Development Team. R: A Language and Environment for Statistical Computing. 2018;
23. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd editio. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
24. Dershem R, Chu X, Wood GC, et al. Changes in telomere length 3-5 years after gastric bypass surgery. *Int. J. Obes.* 2017;41(11):1718–1720.
25. Smith DL, Thomas DM, Siu CO, et al. Regression to the mean, apparent data errors and biologically extraordinary results: Letter regarding "changes in telomere length 3-5 years after gastric bypass surgery." *Int. J. Obes.* 2018;42(4):949–950.
26. Meier HCS, Hussein M, Needham B, et al. Cellular response to chronic psychosocial stress: Ten-year longitudinal changes in telomere length in the Multi-Ethnic Study of

Atherosclerosis. *Psychoneuroendocrinology* [electronic article]. 2019;107(February):70–81. (https://doi.org/10.1016/j.psyneuen.2019.04.018)

27. Seeker LA, Holland R, Underwood S, et al. Method specific calibration corrects for DNA extraction method effects on relative telomere length measurements by quantitative PCR. *PLoS One* [electronic article]. 2016;11(10):1–15. (http://dx.doi.org/10.1371/journal.pone.0164046)
28. Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 2001;29:16–21.
29. Dagnall CL, Hicks B, Teshome K, et al. Effect of pre-analytic variables on the reproducibility of qPCR relative telomere length measurement. *PLoS One*. 2017;12(9):1–10.
30. Nussey DH, Baird D, Barrett E, et al. Measuring telomere length and telomere dynamics in evolutionary biology and ecology. *Methods Ecol. Evol.* 2014;5(4):299–310.

Figures

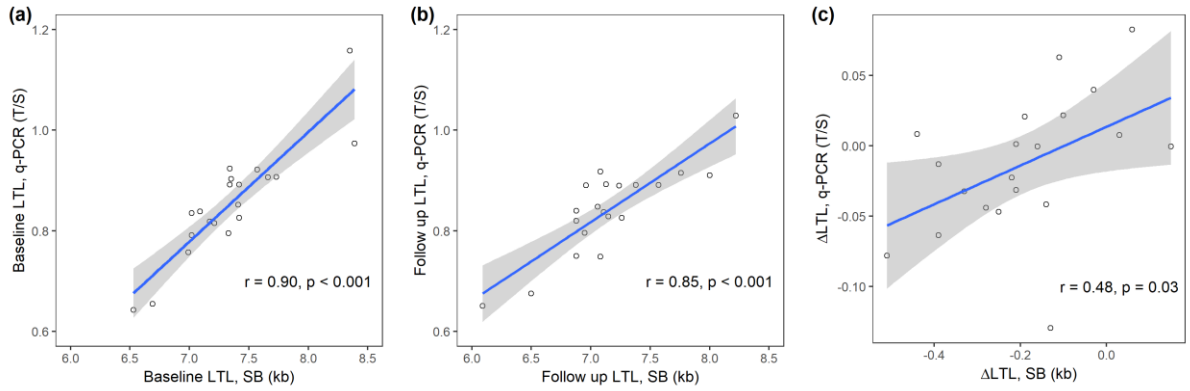

Figure 1. Relationships between leukocyte telomere length (LTL) parameters as measured by qPCR and SB in the empirical dataset: (a) Cross-sectional correlation at baseline. (b) Cross-sectional correlation at follow-up. (c) Change in LTL from baseline to follow-up.

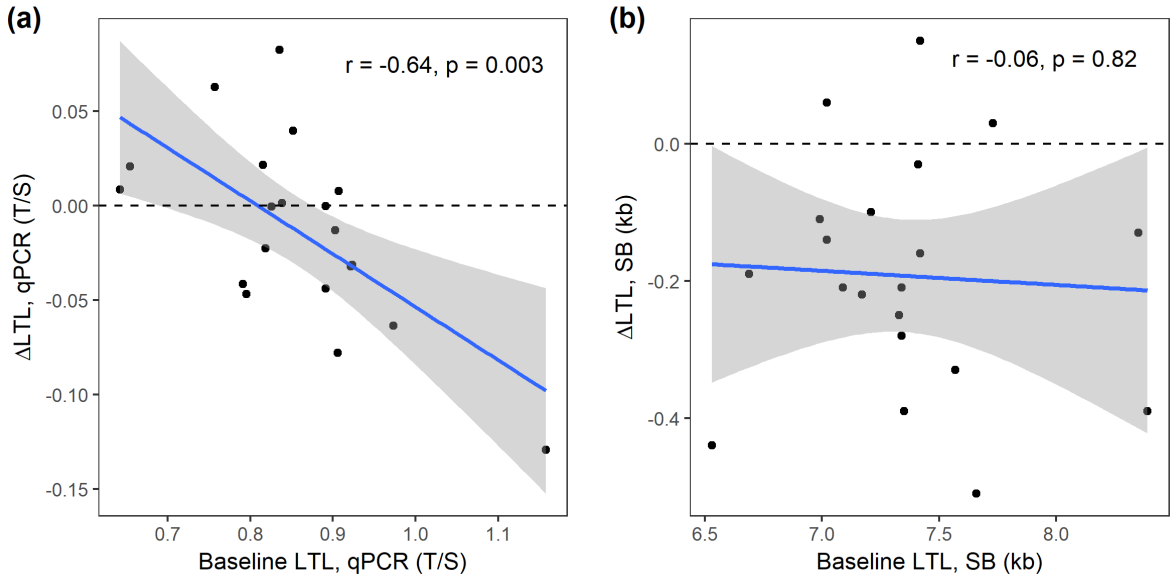

Figure 2. Relationship between change in leukocyte telomere length (ΔLTL) and LTL at baseline in the empirical dataset: (a) qPCR data (b) SB data. In each case, the horizontal dashed line indicates the boundary between shortening (below the line) and lengthening (above the line).

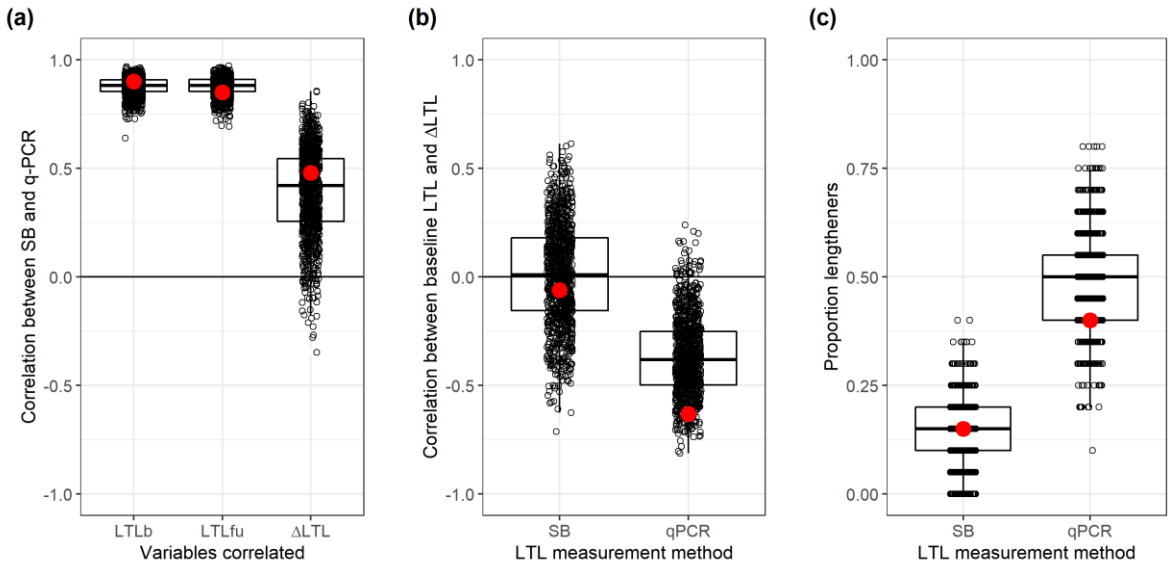

Figure 3. Properties of 1000 simulated datasets (n=20 for each one) drawn from distributions whose parameters matched those inferred from the empirical data. Boxplots indicate means and interquartile ranges. Small symbols indicate individual simulated datasets. Larger red symbols indicate observed values from the empirical data. (a) Correlations between qPCR and SB for baseline LTL (LTLb), follow-up LTL (LTLfu), and LTL change (ΔLTL). (b) Correlations between LTLb and ΔLTL for SB and qPCR. (c) Proportion of apparent lengtheners for SB and qPCR.

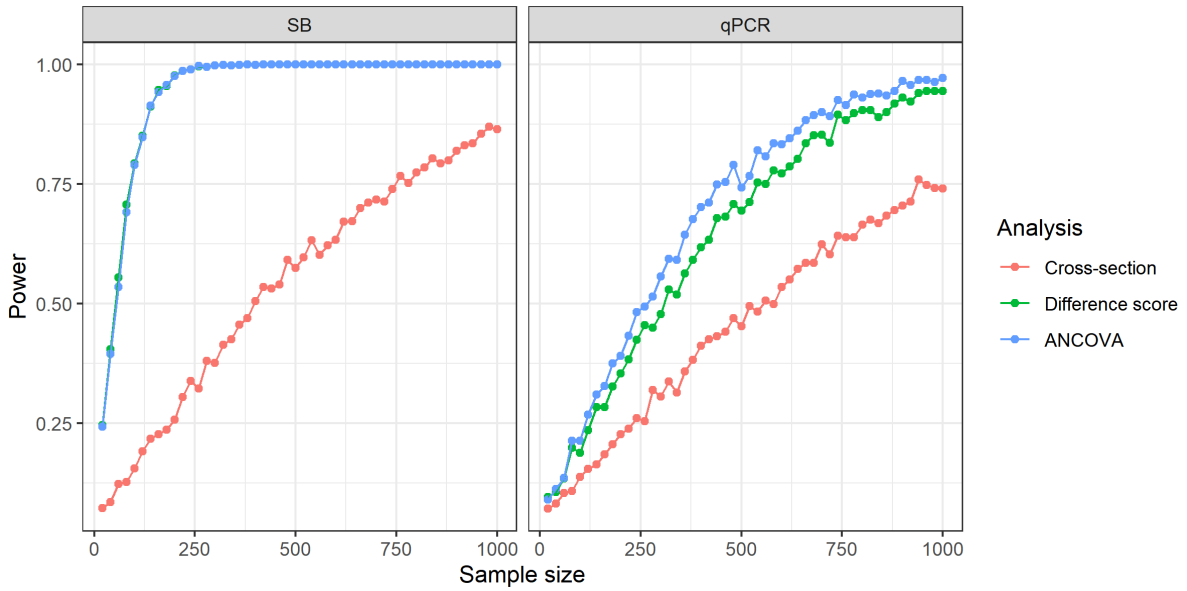

Figure 4. Statistical power to detect a true effect on telomere attrition (effect size $d = 0.5$) at $P < 0.05$ in SB and qPCR data, by sample size and analysis approach. Cross-section ignores the baseline information and performs a cross-sectional analysis at the follow-up time point. Difference score treats ΔLTL, the difference in LTL between baseline and follow-up, as the outcome variable, whereas ANCOVA treats LTL at follow-up as the outcome variable and includes baseline LTL as a covariate.