Estimating telomere length from whole genome sequence data

Zhihao Ding[1], Massimo Mangino[2], Abraham Aviv[3], UK10K Consortium, Tim Spector[2] and Richard Durbin[1,*]

[1]Genome Informatics, Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK, [2]Department of Twin Research and Genetic Epidemiology, King's College London, London, WC2R 2LS, UK and [3]The Center for Human Development and Aging, New Jersey Medical School, Rutgers University, Newark, NJ 07103, USA

ABSTRACT

Telomeres play a key role in replicative ageing and undergo age-dependent attrition *in vivo*. Here, we report a novel method, TelSeq, to measure average telomere length from whole genome or exome shotgun sequence data. In 260 leukocyte samples, we show that TelSeq results correlate with Southern blot measurements of the mean length of terminal restriction fragments (mTRFs) and display age-dependent attrition comparably well as mTRFs.

INTRODUCTION

Telomeres cap the ends of chromosomes and are critical for the maintenance of genome integrity. In humans, telomeres comprise sequences of 5–15 kb TTAGGG tandem repeats and their telomere binding proteins (1). In the absence of telomerase or the alternate pathway, telomeres undergo progressive attrition, which ultimately leads to replicative senescence or apoptosis. Thus, telomere length is an indicator of replicative history and replicative potential—two features of great importance to human health and disease (2).

Standard methods for telomere length measurement are generally classified into three categories: (i) Southern blot analysis of the terminal restriction fragments that measures the average length (mTRF) and length distribution of telomeres in a sample of cells (3); (ii) methods that examine variation in telomere length between chromosomes and cells, i.e. fluorescence *in situ* hybridization (FISH) techniques, including Q-FISH (4) and Flow-FISH (5); and (iii) quantitative polymerase chain reaction (qPCR)-based techniques that measure telomere deoxyribonucleic acid (DNA) content in relative units (compared to single gene DNA) (6).

Next-generation sequencing has now provided an opportunity to obtain genomic information computationally. Shotgun sequence data contains sequencing reads from the telomeres just as any other region of the genome. However, little information about the telomeres can be gained from standard alignments of these reads to the reference sequence. This is because the repetitive nature of the telomeric regions means that it is not possible to assign with confidence the exact origins of the reads, and also since in the human reference sequence (build GRCh37), the ends of most chromosomes are simply stretches of Ns, representing unknown nucleotides.

Instead, previous studies (7) have shown that information on telomere length is contained in the number of telomere motif copies (TTAGGG or CCCTAA) found in reads. Parker *et al.* (8) applied this idea to cancer samples. However, cancer samples typically suffer from aneuploidy, complicating the validation of their results by method such as qPCR (it relies on normalising against a unit copy region). This may be the reason why the measures in (8) only converge to a low resolution telomere status, defined as either gain, no change or loss relative to normal control. Additionally, the vast majority of the samples were paediatric with mean age 7.5 years, and they did not demonstrate a relationship between age and their sequence-based telomere length measurement.

Here, we further examine the relationship between reads containing telomere repeat sequence and telomere length, and describe software for estimating telomere length based on genome-wide sequence data. We demonstrate our method on 260 leukocyte samples (aged 27–74 years, mean age 51 years) from the TwinsUK cohort (9) that have both Illumina 100 bp paired-end whole genome sequence and telomere length measurements using Southern blot mTRFs. We also investigate 96 samples from the 1000 Genomes Project (10) that have both whole genome and exome data.

MATERIALS AND METHODS

We first examined the frequency of reads from the TwinUK dataset with different numbers of copies of TTAGGG and also each non-cyclical permutation of TTAGGG as a control. The frequencies of all non-TTAGGG hexamers showed a monotonic decay as the number of repeat

---

*To whom correspondence should be addressed. Tel: +44 1223 496875; Fax: +44 1223 496826; Email: rd@sanger.ac.uk
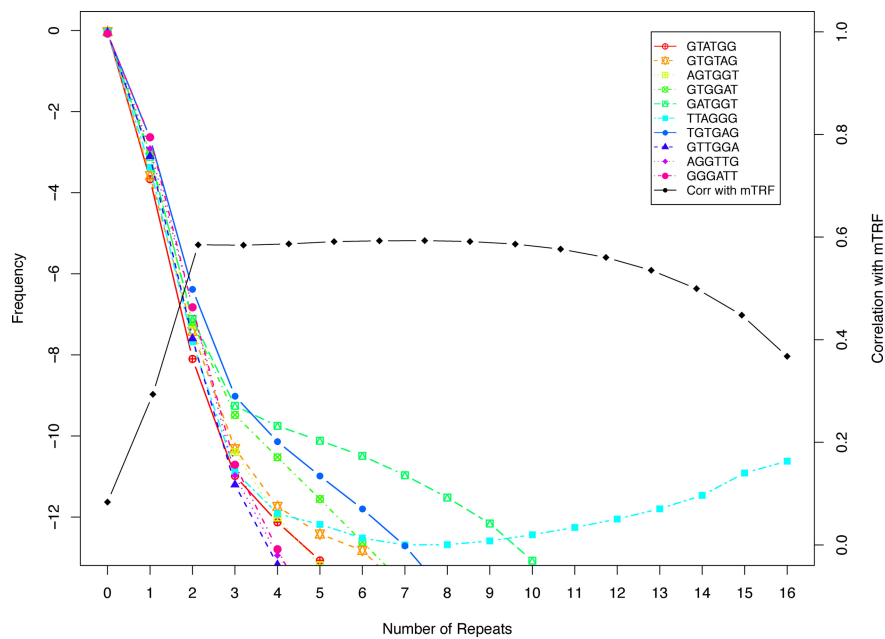
Figure 1. Identification of telomeric reads. In cyan, the log scale frequencies of reads with different numbers of TTAGGG repeats averaged across the 260 TwinsUK samples, with corresponding plots for permutations of TTAGGG in other colours. In black, the correlation of TelSeq to mTRF as a function of the threshold k for the number of repeats per read used in the TelSeq measurement.
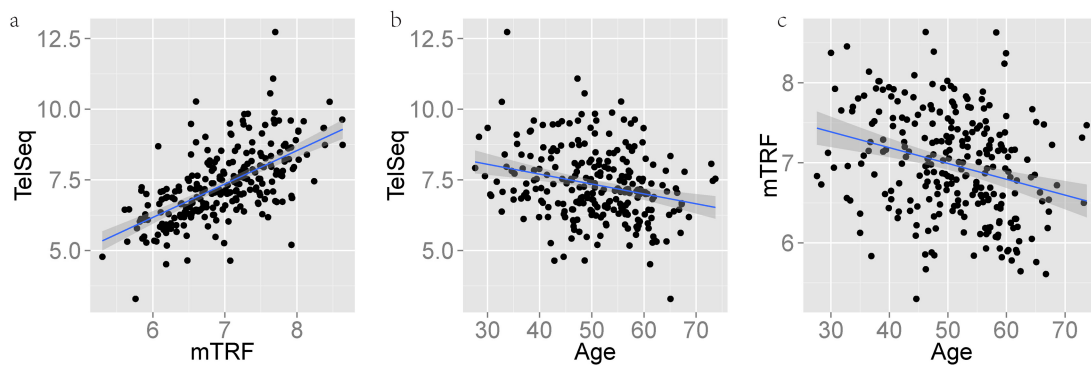

Figure 2. Comparison of TelSeq with experimental measure and age in TwinsUK samples. (a) TelSeq estimate of average telomere length plotted against mTRF estimate; TelSeq (b) and mTRF (c) estimates plotted against age. All average length estimates in kilobases and ages in years.

units increased, with none occurring in a read more than 11 times (Figure 1). In contrast, beyond seven repeats there was an increase in the number of reads containing TTAGGG. We defined reads as telomeric if they contained k or more TTAGGG repeats, with a default threshold value of $k = 7$. These can then be translated into an estimate of the physical length via a size factor s and a constant length c in $l = t_k s c$, where l is the length estimate, t_k is the abundance of telomeric reads at threshold k and c is a constant for genome length divided by number of telomere ends 46 (23×2). The total number of reads could be a good measure of sequence depth and thus a reasonable choice for s. However, studies have shown that DNA molecules in a sequencing library are not sampled and sequenced with equal probability, but instead are subject to biases due to different molecular properties such as GC composition—a high value of which favours more amplification in the PCR step (11). This results in different representations of genome regions and makes defin-

ing s as the total read number not a good estimate. Instead, we define s as a fraction of all reads with gas chromatography (GC) composition between 48 and 52%. The range was chosen to be close to the telomeric GC composition, which is 50% at the TTAGGG dense regions (see Supplementary Figure S1 for results for other GC composition ranges). This fraction is then converted to a mean telomere length estimate in kilobases by multiplying by the cumulative length of genomic regions with the same GC composition c. Considering the GC composition removed an important source of experimental error; and effectively increased the signal by nearly 2-fold, as measured by the correlation between experimental estimates (Supplementary Figure S1). This method is implemented in a program TelSeq which reads one or more Binary Alignment/Map (BAM) files files and returns a report with one row per read group present in the input.

We employed simulated datasets to investigate the effect of sequencing coverage. This was also to discover

the minimum amount of sequence required for reasonable length estimation. We chose the reference sequence (GRCh37) of human chromosome 1 as the sequence source, but with 30 kb nucleotides (including unknown nucleotide Ns) removed from each end and replaced with telomere repeat sequences of the same length. We then simulated 255 synthetic BAMs using the software SimSeq (https://github.com/jstjohn/SimSeq) with sequencing coverage in individual BAMs varying from $0.2X$ to $10X$ in $0.2X$ increments (Supplementary Methods, Supplementary Figure S2). When applied to all BAMs, TelSeq predicted telomere length to be on average 29.4 kb with 1.47 kb standard deviation (SD) (5% of mean). Significant higher variation was seen when coverage was below $2.5X$ ($F = 10.5$, $P = 2.2E−16$ in F test) when compared to results from the higher coverage BAMs (Supplementary Figure S2). For BAMs with $>2.5X$ coverage, TelSeq predicted telomere length to be 29.5 kb with 0.71 kb SD (2.4% of mean).

RESULTS AND DISCUSSION

When TelSeq was applied to the TwinsUK data, the estimates of leukocyte telomere length (LTL) correlated well with the mTRFs measurements across a range of choices of k, with correlation $\rho = 0.60$ at the default threshold $k = 7$ ($P < 10E−16$; Figure 2a). We next examined the relationship between the TelSeq-based LTLs and age of the donors. Given the wide inter-individual variation in LTLs for persons of the same age and the impact of environmental factors on this parameter, the correlation between LTL measurements and age in cross-sectional studies, including TwinsUK, is usually modest (12,13). Nevertheless, since the relationship between measurement and donor age depends on the true LTL value, the correlation provides a means for independent assessment of the informativeness of different experimental techniques for estimating LTL. The TelSeq measurement displayed correlation of $\rho = −0.24$ (explaining 6.5% variance of age, Figure 2b) with age, comparable to that of mTRF (Figure 2c; $\rho = −0.26$, explaining 7.5% variance of age) (Supplementary Method). The difference between $−0.24$ and $−0.26$ is not significant in a t-test using a SD derived by bootstrapping ($P = 0.79$, Supplementary Methods, Supplementary Figure S4). The coefficient of multiple correlation between age and both LTL and mTRF was higher than either individual correlation ($\rho = −0.34$, explaining 9% variance of age); both measurements contributed significantly to the underlying linear regression model ($P = 0.016$, t-test for the TelSeq term; $P = 0.009$, t-test for the mTRF term, Supplementary Methods). This implies that neither TelSeq nor mTRF captured all the information available, and that TelSeq contains additional information independent from that provided by mTRF.

A subset of our samples was sequenced on multiple lanes in separate runs. They can be considered as technical replicates and used to assess the variability of TelSeq measures. The coefficient of variation (CV) was computed as the ratio of the SD to the mean across the technical replicates for each sample. We selected 110 samples that were sequenced on more than 10 lanes to evaluate the CV and observed an average value of 3.17% with 0.98% SD (Figure 3), compara-
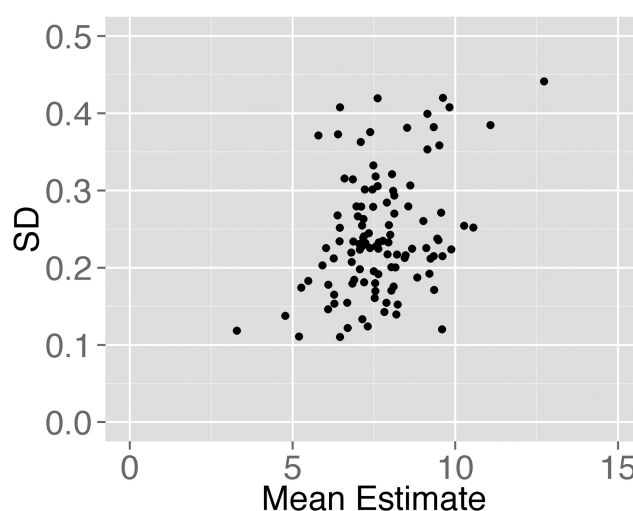

Figure 3. Sequencing lane variation in TelSeq measures. For each sample that was sequenced on more than 10 lanes, the SD of the length estimates across lanes is plotted against the mean length estimate. The CV, defined as the ratio of the SD to the mean, varies between 1.3 and 6.4%, with mean 3.17% and SD 0.98%.

ble to or smaller than that from the experimental measurements (14).

Notably, the TelSeq estimate of telomere length was consistently shorter than the mTRF estimate (mean 5.63 kb compared to 6.97 kb), and the mean rate of shortening per year was consistently greater (34.5 bp/year against 19.8 bp/year) (Figure 2b and c). The mTRF measurements reflect the average distance from a restriction enzyme site (HinfI/RsaI or HphI/MnlI) to the end of a chromosome, and hence overestimate the canonical region of the telomeres of TTAGGG repeats only. Kimura *et al.* (14) obtained a similar figure of around 1 kb for the additional subtelomeric length included in an mTRF measurement. The difference between the TelSeq and mTRF estimates changes as the TelSeq threshold k changes, reflecting inclusion of different amounts of subtelomeric sequence (Supplementary Figure S5); although the correlation between TelSeq and mTRF remains similar across a range of values of k (Figure 1).

In addition to whole genome sequence data, a large number of samples have exome sequence data collected by enrichment of whole genome shotgun sequencing libraries using capture reagents. In theory, if the exome capture works perfectly, it would not be possible to use these data for our method. However, in practice with current technology, a typical exome sequencing output contains some fraction (typically 10–50%) of sequence that is off-target, i.e. not exonic. This fraction represents information on the rest of the genome and can be used to estimate relative telomere length by our method. To test this approach, we selected 96 samples from the 1000 Genomes Project pilot that have matched whole genome and exome sequence and applied TelSeq to both datasets. We found that when we classify telomeric reads as those containing more than three TTAGGG hexamers, estimates of telomere length from the two datasets started to be tightly correlated (Supplementary Figure S6). Using our default threshold of $k = 7$, the two measures have

a Spearman Rank correlation coefficient 0.78. This result suggests that TelSeq can effectively work with exome data, which substantially extends its potential applications.

In conclusion, we have demonstrated an approach for measuring telomere length using whole genome or exome sequencing data. This is the first study to our knowledge to evaluate in detail the relationship between the frequency of telomere repeats and telomere length; and also to validate extensively with experimental measurements in a representative large sample cohort with a wide range of ages. We have implemented our methods in a software package that has been made freely available under the GPL open software licence (https://github.com/zd1/telseq). This allows any cohort with existing genomewide sequence data, including increasingly many cancer genomics and epidemiological cohort studies, to produce a validated measure of telomere length at effectively no cost, with no need for the further sample collection and experimental procedures required by other methods of ascertaining telomere length.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online, including Supplementary References (15).

REFERENCES

1. Samassekou,O., Gadji,M., Drouin,R. and Yan,J. (2010) Sizing the ends: normal length of human telomeres. *Ann. Anat.*, **192**, 284–291.
2. Blasco,M.A. (2005) Telomeres and human disease: ageing, cancer and beyond. *Nat. Rev. Genet.*, **6**, 611–622.
3. Kimura,M., Stone,R.C., Hunt,S.C., Skurnick,J., Lu,X., Cao,X., Harley,C.B. and Aviv,A. (2010) Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat. Protoc.*, **5**, 1596–1607.
4. Martens,U. and Zijlmans,J. (1998) Short telomeres on human chromosome 17p. *Nat. Genet.*, **18**, 76–80.
5. Baerlocher,G.M., Vulto,I., de Jong,G. and Lansdorp,P.M. (2006) Flow cytometry and FISH to measure the average length of telomeres (flow FISH). *Nat. Protoc.*, **1**, 2365–2376.
6. Cawthon,R.M. (2009) Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res.*, **37**, e21.
7. Castle,J.C., Biery,M., Bouzek,H., Xie,T., Chen,R., Misura,K., Jackson,S., Armour,C.D., Johnson,J.M., Rohl,C.A. *et al.* (2010) DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. *BMC Genomics*, **11**, 244.
8. Parker,M., Chen,X., Bahrami,A., Dalton,J., Rusch,M., Wu,G., Easton,J., Cheung,N.-K., Dyer,M., Mardis,E.R. *et al.* (2012) Assessing telomeric DNA content in pediatric cancers using whole-genome sequencing data. *Genome Biol.*, **13**, R113.
9. Moayyeri,A., Hammond,C.J., Valdes,A.M. and Spector,T.D. (2013) Cohort Profile: TwinsUK and healthy ageing twin study. *Int. J. Epidemiol.*, **42**, 76–85.
10. Durbin,R.M., Altshuler,D.L., Abecasis,G.R., Bentley,D.R., Chakravarti,A., Clark,A.G., Collins,F.S., De La Vega,F.M., Donnelly,P., Egholm,M. *et al.* (2010) A map of human genome variation from population scale sequencing. *Nature*, **467**, 1061–1073.
11. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
12. Valdes,A.M., Andrew,T., Gardner,J.P., Kimura,M., Oelsner,E., Cherkas,L.F., Aviv,A. and Spector,T.D. (2005) Obesity, cigarette smoking, and telomere length in women. *Lancet*, **366**, 662–664.
13. Broer,L., Codd,V., Nyholt,D.R., Deelen,J., Mangino,M., Willemsen,G., Albrecht,E., Amin,N., Beekman,M., de Geus,E.J.C. *et al.* (2013) Meta-analysis of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect. *Eur. J. Hum. Genet.*, **21**, 1–6.
14. Kimura,M. and Aviv,A. (2011) Measurement of telomere DNA content by dot blot analysis. *Nucleic Acids Res.*, **39**, e84.