Chapter 4

Measuring telomere length from sequence data

Collaboration Note. *The method developed in this work was designed by Richard Durbin and implemented and evaluated by myself. The study uses data collected from the TwinsUK cohort.*

4.1 Overview

Telomeres cap the ends of chromosomes and are critical for the maintenance of genome integrity. In humans, telomeres comprise sequences of 5-15kb TTAGGG tandem repeats and their telomere binding proteins (Samassekou et al., 2010). In the absence of telomerase or the alternative lengthening pathways (Henson et al., 2002), 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 (Blasco, 2005).

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 (Kimura et al., 2010); (ii) methods that examine variation in telomere length between chromosomes and cells, i.e., fluorescence in situ hybridization (FISH) techniques, including Q-FISH (Martens et al., 1998) and Flow-FISH (Baerlocher et al., 2006); and (iii) quantitative PCR (qPCR)-based techniques that measure telomere DNA content in relative units (compared to single gene DNA) (Cawthon, 2009).

Next-generation sequencing has now provided an opportunity to obtain genomic information cost effectively in large scale. 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 because in the human reference sequence (build GRCh37) the ends of most chromosomes are simply stretches of Ns, representing unknown nucleotides.

Instead, previous studies (Castle et al., 2010) 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. (2012) 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 Parker et al. (2012) 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 pediatric 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 (Moayyeri et al., 2013b) that have both Illumina 100bp paired-end whole genome sequence and telomere length measurements using Southern blot mTRFs. We also investigate 96 samples from the 1000 Genomes Project (The 1000 Genomes Consortium, 2010) that have both whole genome and exome data.

4.2 Study samples and data

The 260 UK10K individuals investigated in this study were all female aged 27 - 74 years (mean age 51 years) from the TwinsUK cohort (Moayyeri et al., 2013b, http://www.twinsuk.ac.uk/). Except for 5 pairs of dizygous twins, the rest were all unrelated. Leukocyte telomere lengths of these individuals as mTRFs were measured using Southern blot. Whole genome sequencing was conducted using the Illumina HiSeq technology, yielding sequencing reads with coverage ranging from 4X to 16.6X (average 6.5X, pooled across lanes). Twelve individuals with a much higher read duplication rate (more than 3 fold that of other samples) were excluded from the rest of the analysis since they gave outlier results (Figure 4.1).

Sequence data are available from the European Genome-phenome Archive (EGA) study number EGAS00001000108, submitted by UK10K (http://www.uk10k.org). The 1000 Genomes Project sequence data were downloaded from http://www.1000genomes.org.

4.3 Estimating telomere length from whole genome sequence data.

4.3.1 Estimator

We first examined the frequency of reads from the TwinsUK 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 units increased, with none occurring in a read more than eleven times (Figure 4.2). 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 $k = 7$, values higher than which do not increase performance substantially. These can then be translated into an estimate of the physical length via a size factor s and a constant length c in $l = t_k g/(46s)$, where l is the length estimate, t_k is the number of telomeric reads at threshold k, g is the genome length and s is the total number of reads. The

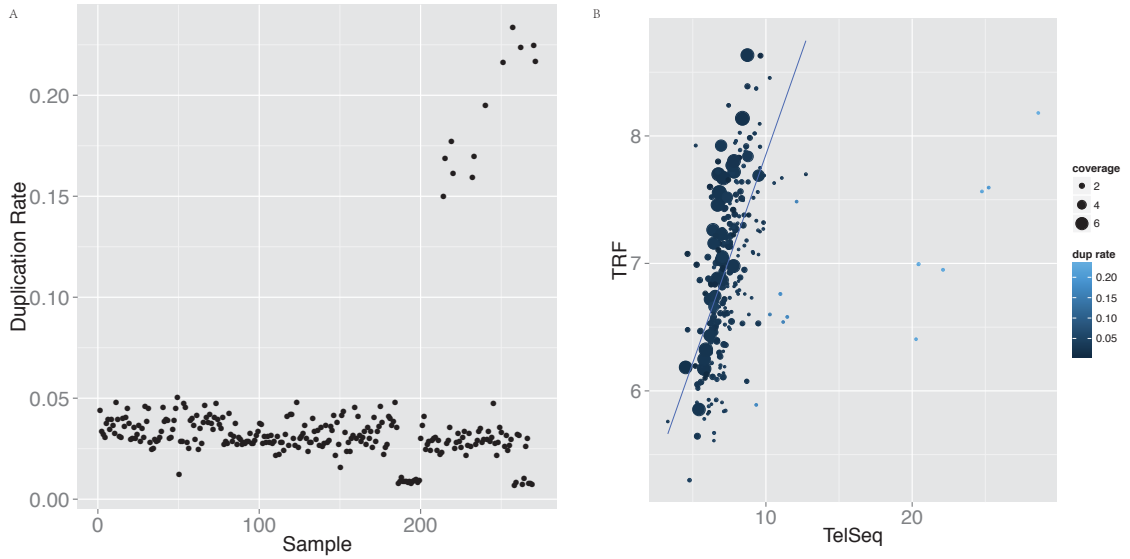

Figure 4.1: The effect of duplication rate and coverage to TelSeq performance. In essence, TelSeq relies on sampling of genomic regions from a sequencing library. Coverage and duplication thus affect the translation of a relative measure into an absolute one. Low coverage indicates insufficient sampling and thus results in high variation in estimation (Figure 4.4) while high duplication suggests over enrichment of certain genomic regions and thus changes the translation factor c. In whole genome sequencing high duplication rate indicates low library complexity and loss of information. Twelve of our samples were found to have an exceptionally high duplication rate (>3 fold greater than the rest, panel A), and were outliers when regressing against mTRF (panel B). We based our evaluation on samples with duplication rate below 10%, which is typically what is expected for whole genome sequencing.

factor of 46 corresponds to number of telomere ends 46(23×2).

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 favors more amplification in the PCR step (Dohm et al., 2008). This results in different representations of genome regions and makes defining s as the total read number not a good estimate. Instead, we define s as a fraction of all reads within a specific GC composition range, and similarly g as the length of genome for which 100bp segment lie within the same GC range. The range was chosen to be close to the telomeric GC composition, which is 50% at the TTAGGG dense regions (see Figure 4.3 for results for other GC composition ranges).

Considering the GC composition removed an important source of experimental error; and effectively increased the signal by nearly two-fold, as measured by the correlation between experimental estimates (Figure 4.3). This method is implemented in a program TelSeq which reads one or more BAM files and returns a report with one row per read group present in the input.

To calculate g we divide the reference sequence into 100bp consecutive bins and add 100bp to g if the GC composition of the bin is within the range.

Association to age and mTRF The Pearson's Correlation Coefficient was calculated using the *cor* function of the R language (Computing, R Foundation for Statistical Vienna, 2008, http://www.r-project.org/). The regression between age and TelSeq and between age and mTRF was calculated using the *lm* function of R in models *lm(age ~ telseq)* and *lm(age ~ mTRF)*. Two measures were also included in one model *lm(age~telseq + mTRF)* as two independent fixed effects. A t-test was done for each of the two regression coefficient (β) against null hypothesis $\beta = 0$, the results of which can be seen in the output of the summary function.

Calculating the variance explained To compute the proportion of variance of age explained, we used the *cor* function in R *cor(age, mTRF, method="pearson")^2*. To compute the additional variance that can be explained by mTRF while controlling

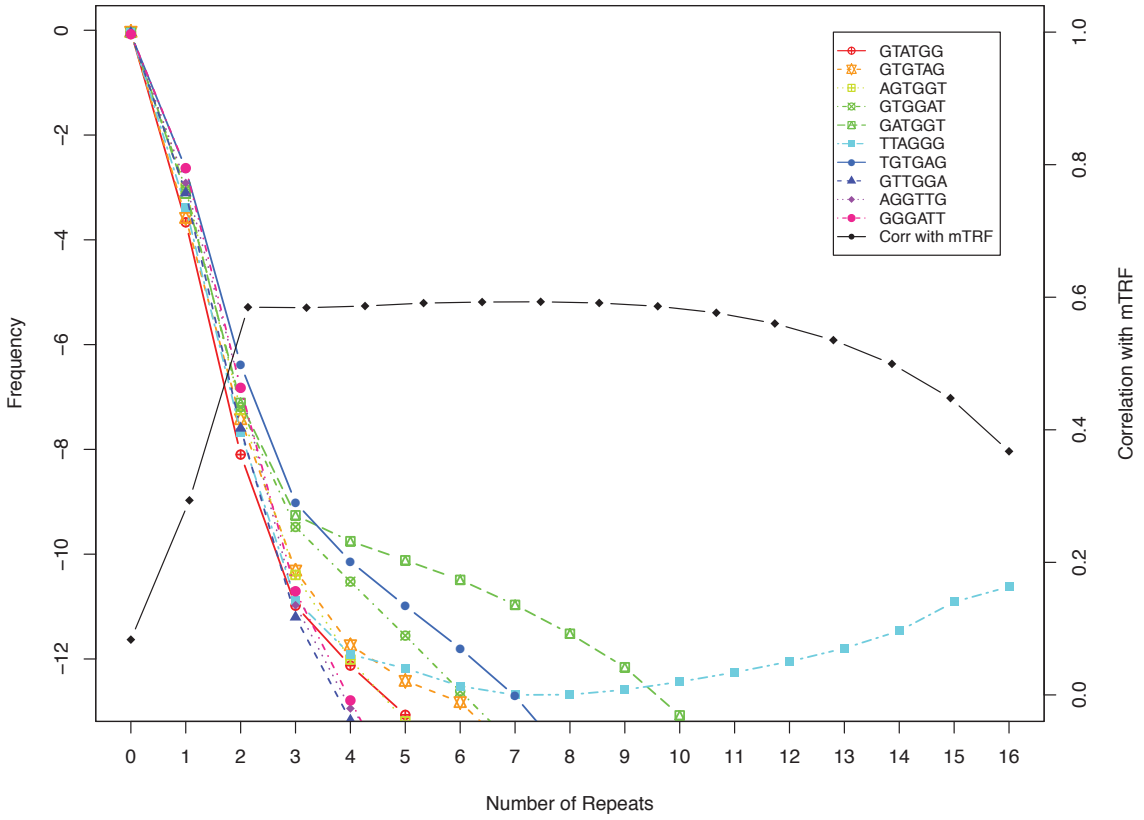

Figure 4.2: Identification of telomeric reads. In cyan the log scale frequencies of reads with different numbers of TTAGGG repeats averaged across the 260 TwinsUK samples, with corresponding plots for permutations of TTAGGG in other colours. In black the correlation of TelSeq to mTRF as a function of the threshold k for the number of repeats per read used in the TelSeq measurement.

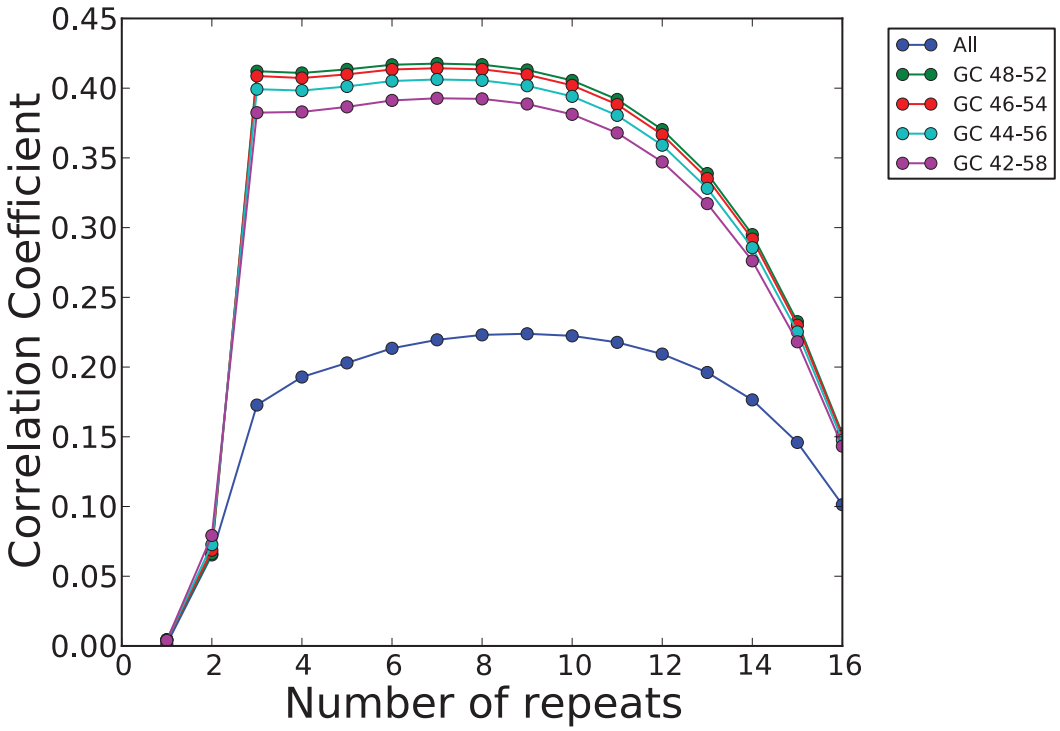

Figure 4.3: Normalising by reads with similar GC improves the performance of TelSeq. It is known that read abundance in a sequencing library is affected by the GC composition of a read, a bias primarily introduced in the PCR step where high GC reads get amplified more often due to their high molecular affinity. Thus, using reads with similar GC content as background accounts for this molecular property and reflects the signal to noise ratio more accurately. To demonstrate this we evaluated the performance of TelSeq, as measured by the correlation with mTRF, when normalised by reads from different GC groups, 42%-58% (purple), 44%-56% (light green), 46%-54% (red), 48%- 52% (dark green) as well as by all reads (blue). The result showed that there was a gradual increase to the correlation when GC range approaches 50%. And in all these cases, the correlation was much higher than that when all reads were used from a library. Here the analysis was done for the whole range of threshold k, the number of TTAGGG repeats in a read.

for TelSeq, we firstly obtained the residuals from a regression between age and TelSeq (*x<-lm(age~TelSeq)$residuals*); and then used the residuals to compute the additional variation explained (*cor(x,mTRF)^2*). The same procedure was done for TelSeq.

4.3.2 Simulation

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 30kb nucleotides (including unknown nucleotide Ns) removed from each end and replaced with telomere repeat sequences (TTAGGG) of the same length. We then simulated Illumina pair-end reads using the software SimSeq (https://github.com/jstjohn/SimSeq, parameters -1 100 -2 100 –insert_size 500 –insert_stdev 200) with sequencing coverage in individual BAMs varying from 0.2X (498,501 reads) to 10X(24,925,063 reads) in 0.2X increments (Figure 4.4). For each setting we repeated the simulation 5 times and generated 255 BAMs in total. We then applied TelSeq to estimate telomere lengths of these BAMs. TelSeq predicted a length of 29.4kb on average with 1.47kb standard deviation (5% of mean). Significant higher variation was seen when coverage was below 2.5X (F=10.5, P=2.2E-16 in the F test) when compared to results from the higher coverage BAMs (Figure 4.4). For BAMs with >2.5X coverage, TelSeq predicted telomere length to be 29.5kb with 0.71kb standard deviation (2.4% of mean).

4.3.3 Results

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 4.5A). 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,

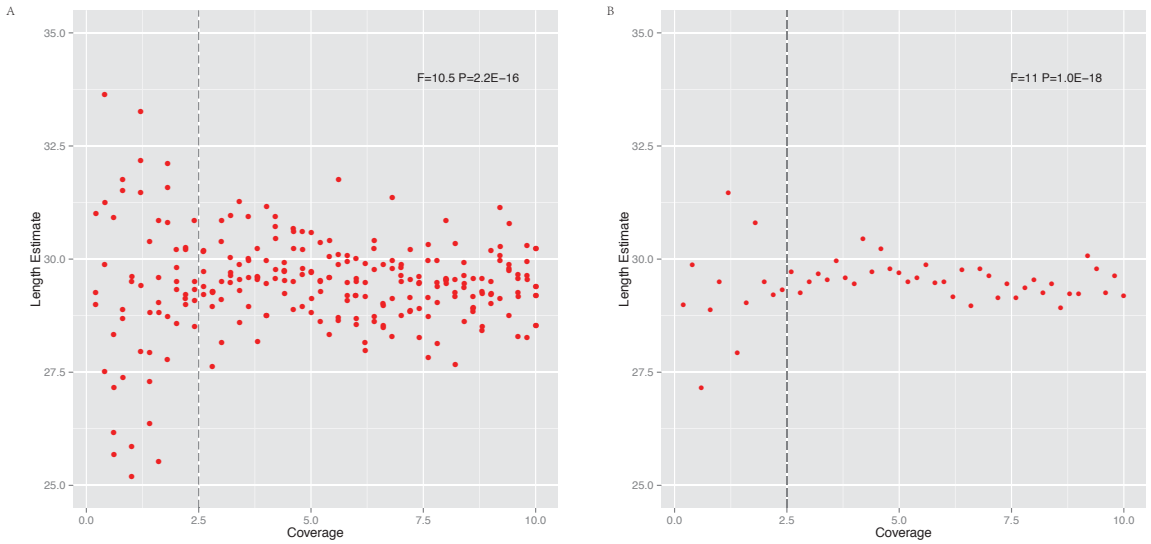

Figure 4.4: The effect of sequencing coverage on TelSeq measurement, assessed by simulation. A group of BAMs were simulated using software SimSeq (https://github.com/jstjohn/SimSeq). Sampling noise is substantially higher when the coverage is below 2.5X (mean=29.4kb, variation=5% of mean), compared to when coverage is above 2.5X (mean=29.5kb, variation=2.4% of mean) (**A**). The mean estimates are close to the true value 30kb independent of coverage. When using the weighted average of 5 BAMs for each coverage group (**B**), the variation is much smaller (1% of mean). This is justified theoretically by the relationship $X \sim N(\mu, \sigma^2)$, $\bar{X} \sim N(\mu, \sigma^2/n)$, where n is the sample size. In real experiments, ideally estimates should be obtained from multiple libraries across multiple lanes for a sample. The coefficient of variation across lanes per sample is on average 3.2% (Figure 4.7).

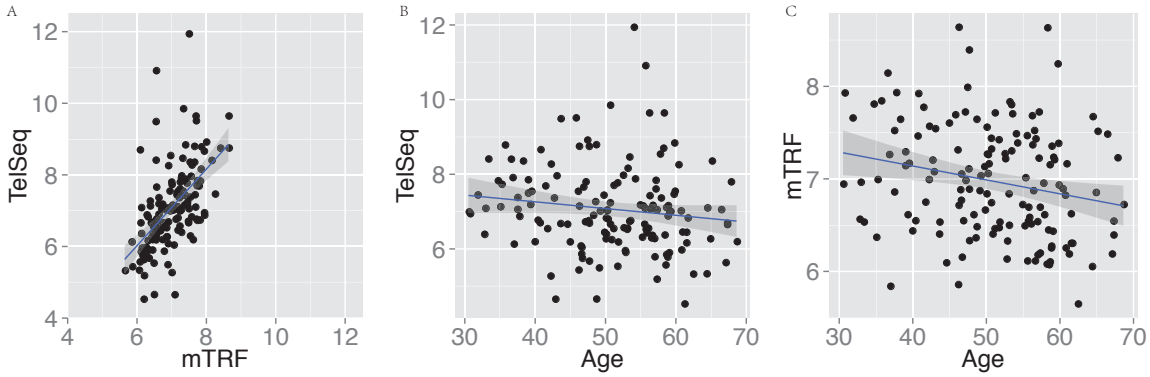

Figure 4.5: Comparison of TelSeq with experimental measure and age in TwinsUK samples. (A) TelSeq estimate of average telomere length plotted against mTRF estimate; TelSeq (B) and mTRF (C) estimates plotted against age. All average length estimates in kilobases and ages in years.

is usually modest (Valdes et al., 2005; Broer et al., 2013). 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 ρ=-0.24 (explaining 6.5% variance of age, Figure 4.5B) with age, comparable to that of mTRF (Figure 4.5C; ρ=-0.26, explaining 7.5% variance of age). The difference between -0.24 and -0.26 is not significant in a *t*-test using a standard deviation derived by bootstrapping (P=0.79, Figure 4.6). The coefficient of multiple correlation between age and both LTL and mTRF was higher than either individual correlation (ρ=- 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). This implies that neither TelSeq nor mTRF captured all the information available, and that TelSeq contains additional information independent from that provided by mTRF.

Comparing the correlation coefficients with age by the two methods To test whether the difference is significant in the strength of associations between age and each of two measures, ρ = -0.24 for TelSeq and ρ = -0.26 for mTRF, we con-

ducted bootstrapping using R (*sample(sample_index,sample_size,replace=TRUE)*) sampling our cohort 1000 times, from which we obtained an estimate for the standard deviation of ρ for mTRF (0.052) and TelSeq(0.056). We can then compute the t statistic $t = (\rho_{telseq} - \rho_{mTRF})/sqrt(s^2_{telseq} + s^2_{mTRF})$ for hypothesis testing (Figure 4.6).

Coefficient of variation A subset of our samples were 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 standard deviation (SD) to the mean across the technical replicates for each sample. We selected 110 samples that were sequenced on more than ten lanes to evaluate the CV and observed an average value of 3.17% with 0.98% standard deviation (4.7), comparable to or smaller than that from the experimental measurements (Kimura and Aviv, 2011).

Interestingly, when lanes analyzed separately and the telomere length estimate calculated as the mean across lanes, weighted by lane yield, the sampling error was further reduced and the correlation with mTRF was stronger (ρ=0.62 with mTRF when merged as opposed to ρ=0.60).

Difference in length estimates Notably, the TelSeq estimate of telomere length was consistently shorter than the mTRF estimate(mean 5.63kb compared to 6.97kb), and the mean rate of shortening per year was consistently greater (34.5bp/year against 19.8bp/year) (Figure 4.5B, Figure 4.5C). 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 and Aviv (2011) obtained a similar figure of around 1kb for the additional sub-telomeric 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 (Figure 4.8); although the correlation between TelSeq and mTRF remains similar across a range of values of k (Figure 4.2).

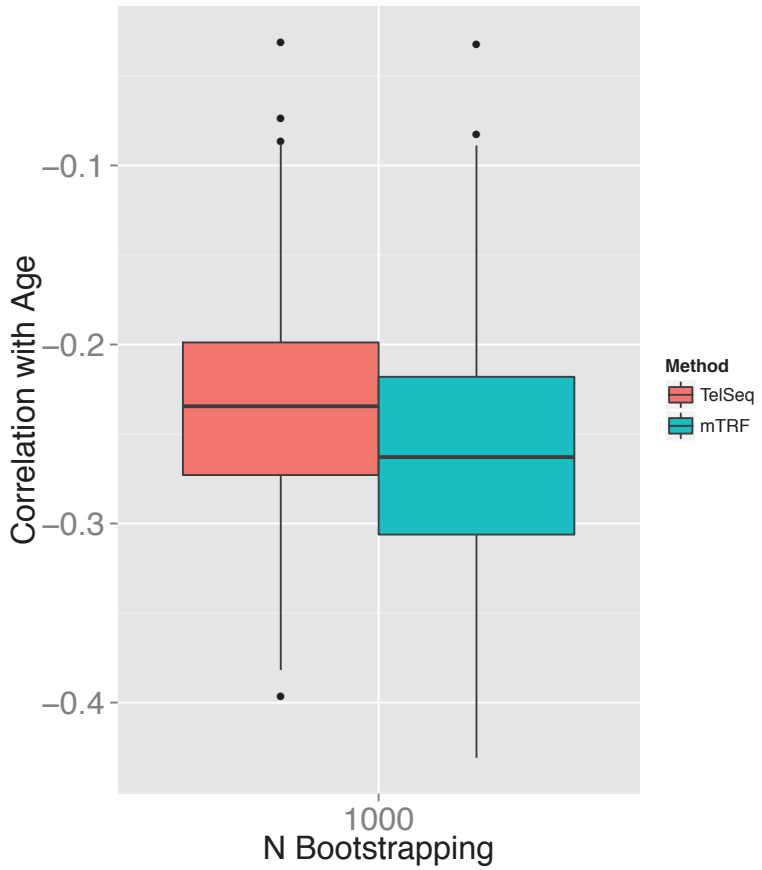

Figure 4.6: Compare correlation coefficient obtained from mTRF and TelSeq. To compare the correlation coefficients between age and telomere length estimates from TelSeq and mTRF, we conducted 1000 bootstraps with replacement from the data set to obtain an estimate of the standard deviations of the correlation estimates ρ. We can then perform a t-test for whether the difference between the observed values -0.24 and -0.26 is significant. The result gave t=0.26, P=0.79, which suggest no statistical difference between the coefficients obtained from the two measurements.

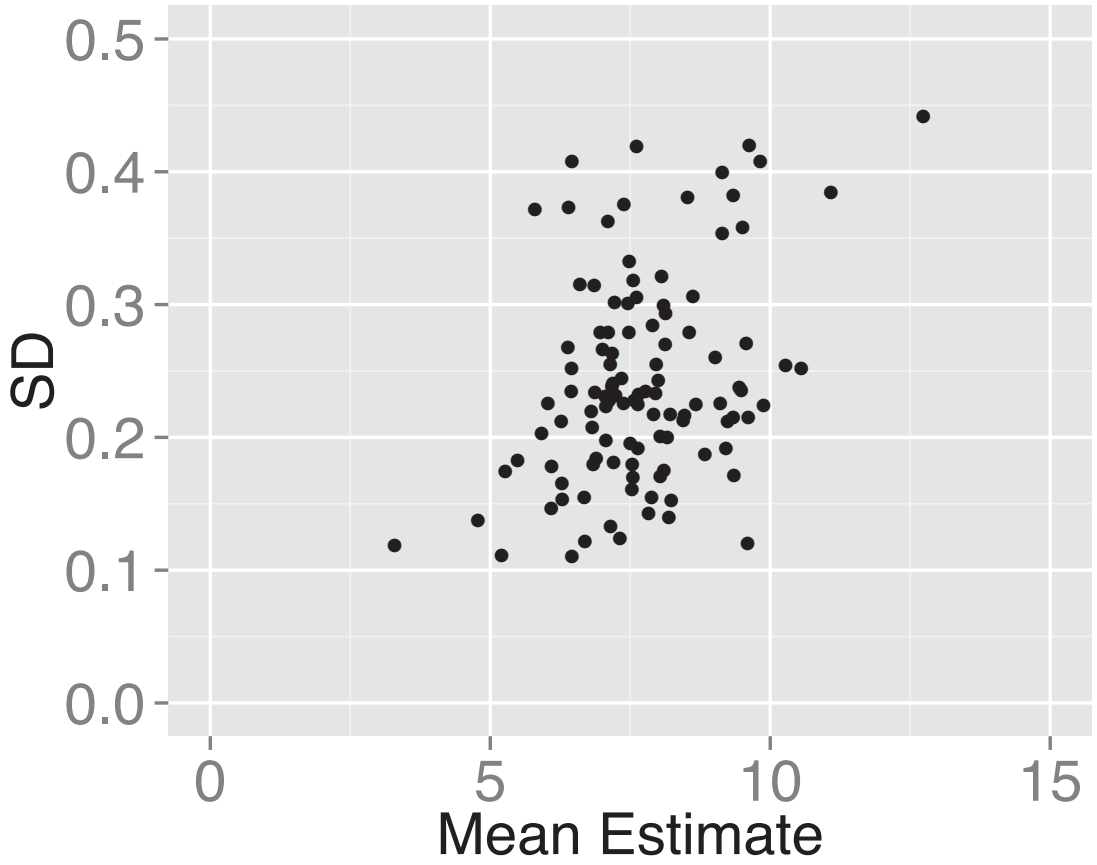

Figure 4.7: Sequencing lane variation in TelSeq measures. For each sample that was sequenced on more than ten lanes, the standard deviation of the length estimates across lanes is plotted against the mean length estimate. The coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, varies between 1.3% and 6.4%, with mean 3.17% and standard deviation 0.98%.

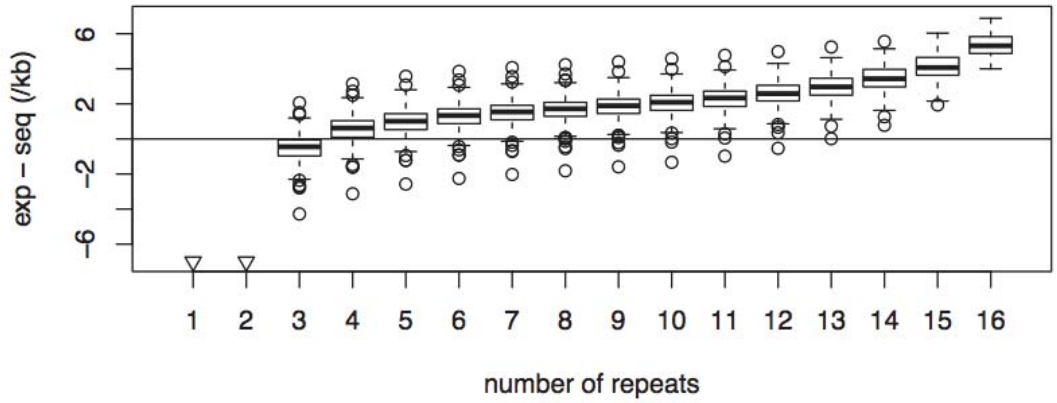

Figure 4.8: The mTRF measurement is longer than TelSeq estimates across a range of values for the choices of TelSeq threshold (k). The difference between mTRF and TelSeq is 1.49kb at k=7, and 5.34kb at k=16. The difference reflects the fact that mTRF measures the average distance from subtelomeic regions, where the excision sites of restriction enzymes exist, to the chromosome ends, while TelSeq approaches include only the ends when choosing a large k. Measurements of two methods correlate with age similarly, suggesting they both capture the information of telomere shortening with age.

4.4 Estimating telomere length from exome sequence data.

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 data sets. We found that when we classify telomeric reads as those containing more than three TTAGGG hexamers, estimates of telomere length from the two data sets started to be tightly correlated (Figure 4.9). Using our default threshold of k=7, the two measures have a Spearman's Rank correlation coefficient 0.78. This result suggests that TelSeq can effectively work with exome data, which substantially extends its potential applications.

4.5 Applications of the method

Mutations in POT1 gene predispose to melanoma Robles-Espinoza et al. (2014) performed exome sequencing on pedigrees recruited in the UK, Netherlands and Australia with melanoma cases looking for variants that are explanatory to the disease. Four loss of function variants in the protection of telomeres 1 gene (POT1) were identified as cosegregating with melanoma cases in family UF20 (See Figure 4.10A for the pedigree with melanoma cases (arrowed) and missense mutations in POT1 at p.Tyr89Cys). The mutation disrupts the interaction between POT1 and single-stranded DNA and led to elongated telomere length (Robles-Espinoza et al., 2014). Telomere length information is thus an important phenotype to this study.

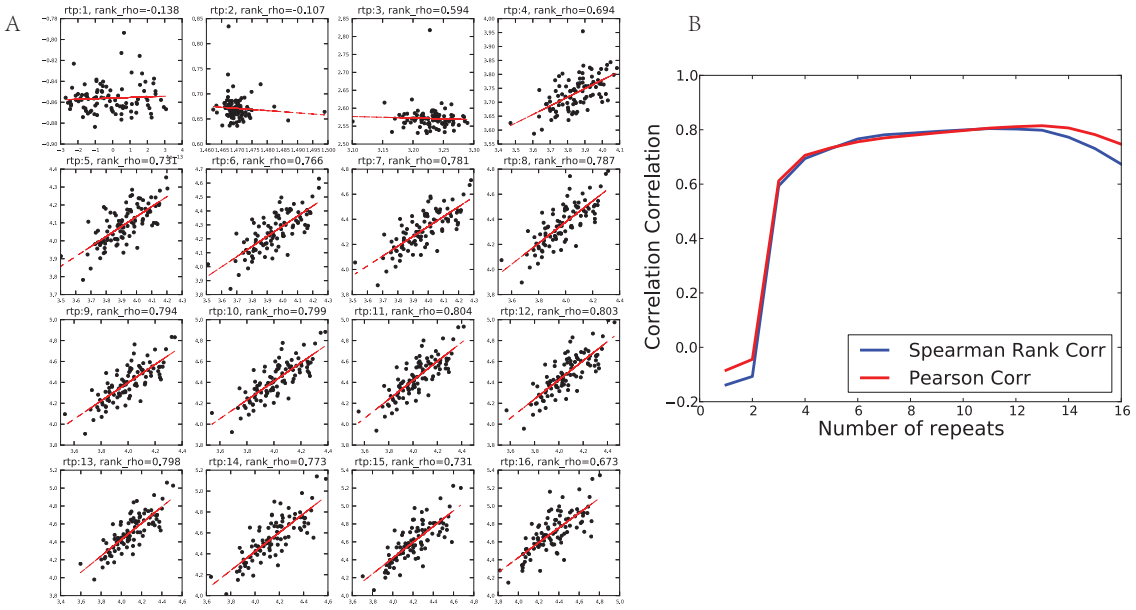

Figure 4.9: TelSeq estimates from exome data are highly correlated with those from whole genome data in 96 samples from the 1000 Genomes Project with matched whole genome sequences and exome sequence data. A. Scatter plots for TelSeq estimates from matched whole genome sequence and exome sequence at different thresholds of k, the amount of TTAGGG repeats in a read. Panels are organised from left to right, top to bottom as k increases from 1 to 16, where in each plot X axis is the estimates from the whole genome sequences and y axis is the estimates for the matched exome sequences. A correlation coefficient is calculated for each panel and plotted in B. The two measurements start becoming tightly correlated with each other when $k >= 3$.

Telomere lengths of the cases along with 38 controls that have wild type POT1 gene were measured using the qPCR method (Figure 4.10B) and Telseq (Figure 4.10C). Two methods show consistent signal that the cases with mutated POT1 gene have much longer telomere than the controls ($P < 0.00019$).

4.6 Conclusion

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 repeat sequence in shotgun sequence data and telomere length, and also to validate extensively with experimental measurements in a representative large sample cohort with a wide range of ages. There are some limitations to TelSeq, such as it is not able to obtain individual telomere length for chromosome arms. Nevertheless, Telseq allows any cohort with existing genomewide sequence data, including increasingly many cancer genomics and epidemiological cohort studies, to produce a validated measure of the average 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.

4.7 Software implementation

Telseq is implemented in C++. It uses BamTools (Barnett et al., 2011) to read BAM files. The source code is licensed under GNU General Public License Version 3 and is freely available online (https://github.com/zd1/telseq). To compile, a recent version of GNU Compiler Collection (GCC) is recommended (Version 4.8 or above).

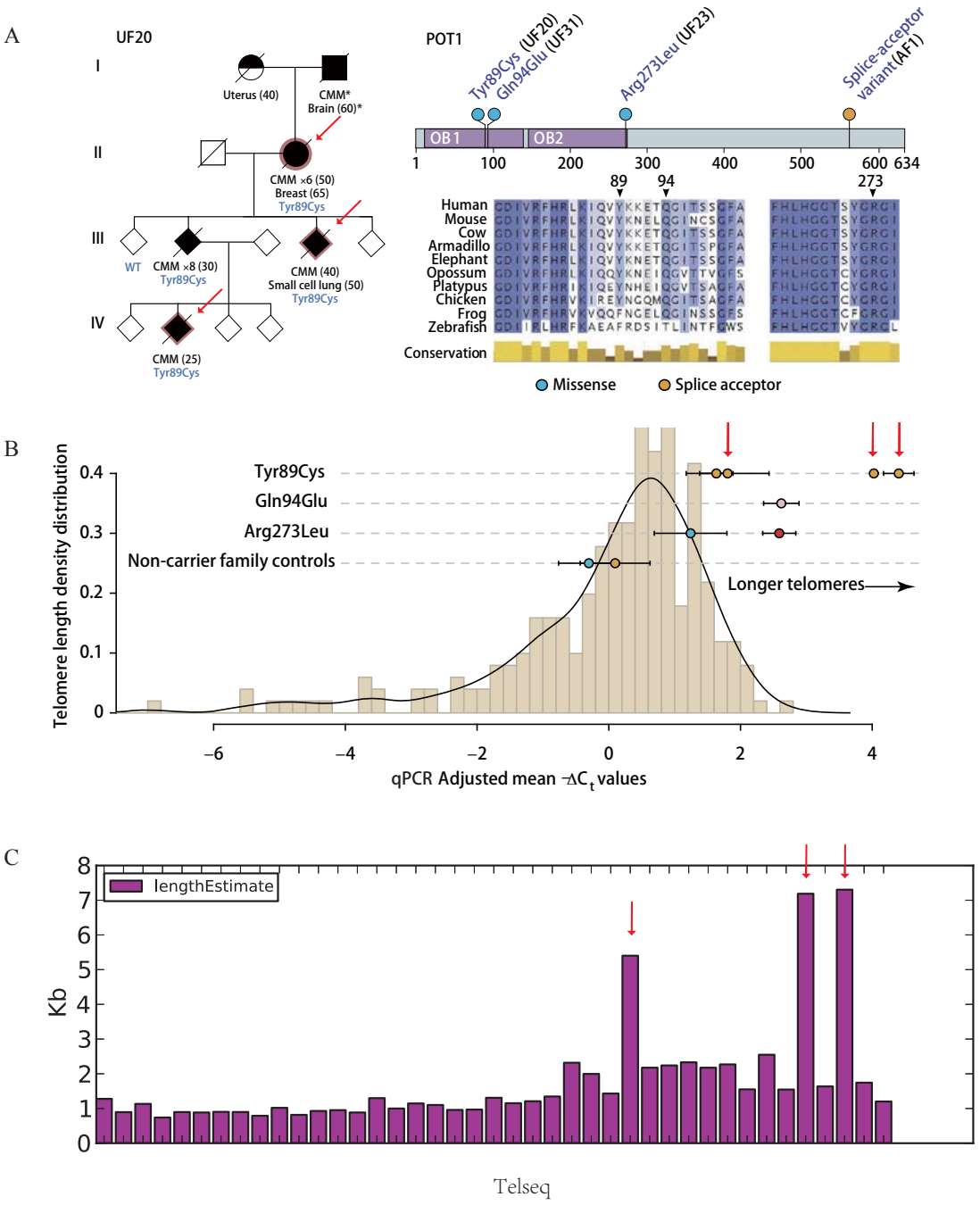

Figure 4.10: Measuring telomere lengths in melanoma cases. Mutations in the Protection of Telomeres 1 gene (POT1) were found transmitted in melanoma cases in pedigree UF20 (A). The telomere length estimates were obtained independently using a qPCR approach and Telseq (B and C). The cases that red-arrowed and compared against controls that have wild type POT1. Both methods indicate longer telomeres in the three cases. Panel A and B are adapted from Figure 2 in Robles-Espinoza et al., 2014.