Alignment-free telomere length estimation from whole genome NGS data

Szymon M. Kiełbasa[1], Jelle Goeman[2], Hein Putter[1], *GoNL Consortium*, Dorret Boomsma[3], Eline Slagboom[1], Kai Ye[4]

[1]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

[2]Department for Health Evidence, Radboud university medical center, Nijmegen, The Netherlands

[3]Netherlands Twin Register, Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands

[4]The Genome Institute, Washington University in St. Louis, St Louis, USA

Abstract

Telomeres are repetitive structures present at each end of a chromatid. They play role in maintenance of genome integrity. Due to the nature of the chromosome replication process, the telomeres shorten at each replication cycle. Consequently, with lifetime of the organism the average telomere length decreases and it may be used as a marker for organism's *biological age*.

Here we present a method for accurate estimation of telomere lengths from unaligned whole genome sequencing reads. We developed the method based on a dataset provided by The Genome of the Netherlands (GoNL) project which generated whole genome sequencing data for 754 samples of 248 Dutch families. For 381 of the samples telomere length measurements were available. These measurements were obtained without usage of next generation sequencing methods.

Our method contains two components: the read classifier and a linear model. The read classifier is a fast function for detection of repetitive sequences (in particular the telomeric motif TTAGGG) in read sequences. We apply this function to all reads of a sample and then we build a table of counts of reads with various repetitive motifs. Next, based on the read counts table and available telomere length measurements we train a linear predictor of telomere length.

We demonstrate that the simplest possible predictor, which only bases on frequency of reads with the telomeric motif TTAGGG, displays a strong sequencing batch bias. When frequencies of a few other repetitive motifs are incorporated to the model, its performance significantly improves.

Finally, we compare our predictions with predictions obtained from *telseq* algorithm. The *telseq* estimations show strong effect of sequencing batch. Moreover, we demonstrate that our method delivers estimations more strongly associated with individuals age.