README batchAdjusted telomereLength

Dataset description

The datasets are investigator data products produced by the TOPMed Telomere Working Group from TOPMed Freeze 8 Whole Genome Sequencing. They are updates from a previous upload of telomere length and include telomere length and a batch-adjusted telomere length centered around 0 (residuals from regressing out 200 batch principal components). Each dataset is indexed by sample identifier ("SAMPLE_ID") and includes a variable identifying the TOPMed study abbreviation (for ease in combining datasets). The updated datasets also include additional samples in several studies as well as samples from additional studies. See "Further dataset description" below for more details.

Available datasets

The study datasets can be found within each study's TOPMed Exchange Area:

``Investigator_Data_Products/batchAdjusted_telomereLength_20200908_phsxxxxxx.tar.gz``.

Here the `phsxxxxxx` is the TOPMed dbGaP accession number. The tar.gz file contains the dataset (tab-delimited text file), the data dictionary (tab-delimited text file in dbGaP format), and a README.

The following table lists the TOPMed study name along with the corresponding Exchange Area directory for the studies included.

topmed_study	EA_dir
AFLMU	phs001543_TOPMed_CCDG_AFLMU
Amish	phs000956_TOPMed_WGS_Amish
ARIC	phs001211_TOPMed_WGS_ARIC
AustralianFamilialAF	phs001435_TOPMed_WGS_AustralianFamilialAF
BAGS	phs001143_TOPMed_WGS_Asthma_Barbados
BioMe	phs001644_TOPMed_CCDG_BioME
BioVU_AF	phs001624_TOPMed_CCDG_BioVU
CAMP	phs001726_TOPMed_WGS_CAMP
CARDIA	phs001612_TOPMed_WGS_CARDIA
CARE_BADGER	phs001728_TOPMed_WGS_CARE_BADGER
CARE_CLIC	phs001729_TOPMed_WGS_CARE_CLIC
CARE_PACT	phs001730_TOPMed_WGS_CARE_PACT
CARE_TREXA	phs001732_TOPMed_WGS_CARE_TREXA
CATHGEN	phs001600_TOPMed_WGS_CCDG_CATHGEN
CCAF	phs001189_TOPMed_WGS_Cleveland_AF
CFS	phs000954_TOPMed_WGS_CFS
ChildrensHS_GAP	phs001602_TOPMed_WGS_ChildrensHS_GAP
ChildrensHS_IGERA	phs001603_TOPMed_WGS_ChildrensHS_IGERA
ChildrensHS_MetaAir	phs001604_TOPMed_WGS_ChildrensHS_MetaAir
CHIRAH	phs001605_TOPMed_WGS_CHIRAH
CHS	phs001368_TOPMed_WGS_CHS_VTE
COPDGene	phs000951_TOPMed_WGS_COPDGene
CRA	phs000988_TOPMed_WGS_Asthma_CostaRica
DECAF	phs001546_TOPMed_WGS_DECAF
DHS	phs001412_TOPMed_WGS_DHS_AA_CAC
ECLIPSE	phs001472_TOPMed_WGS_ECLIPSE
EGCUT	phs001606_TOPMed_WGS_Estonia
EOCOPD	phs000946_TOPMed_WGS_Boston_EO_COPD
FHS	phs000974_TOPMed_WGS_Framingham

topmed_study	EA_dir
GALAI	phs001542_TOPMed_WGS_GALAI
GALAII	phs000920_TOPMed_WGS_GALAII
GCPD-A	phs001661_TOPMed_WGS_GCPD_A
GENAF	phs001547_TOPMed_WGS_CCDG_GENAF
GeneSTAR	phs001218_TOPMed_WGS_GeneSTAR
GENOA	phs001345_TOPMed_WGS_GENOA
GenSalt	phs001217_TOPMed_WGS_GenSalt
GGAF	phs001725_TOPMed_CCDG_GGAF
GOLDN	phs001359_TOPMed_WGS_GOLDN
HCHS_SOL	phs001395_TOPMed_WGS_HCHS_SOL
HVH	phs000993_TOPMed_WGS_HVH
HyperGEN	phs001293_TOPMed_WGS_HyperGEN
INSPIRE_AF	phs001545_TOPMed_CCDG_INSPIRE_AF
IPF	phs001607_TOPMed_WGS_IPF
JHS	phs000964_TOPMed_WGS_JHS
JHU_AF	phs001598_TOPMed_CCDG_JHU_AF
LTRC	phs001662_TOPMed_WGS_LTRC
Mayo_VTE	phs001402_TOPMed_WGS_Mayo_VTE
MESA	phs001416_TOPMed_WGS_MESA
MGH_AF	phs001062_TOPMed_WGS_MGH_AF
miRhythm	phs001434_TOPMed_WGS_miRhythm
MLOF	phs001515_TOPMed_WGS_MyLifeOurFuture_Hemophilia
MPP	phs001544_TOPMed_CCDG_MPP
OMG_SCD	phs001608_TOPMed_WGS_OMG_SCD
Partners	phs001024_TOPMed_WGS_PartnersBiobank
PCGC_CHD	phs001735_TOPMed_WGS_PCGC
PharmHU	phs001466_TOPMed_WGS_PharmHU
PIMA	phs001727_TOPMed_WGS_PIMA
PMBB_AF	phs001601_TOPMed_WGS_CCDG_UPenn
PUSH_SCD	phs001682_TOPMed_WGS_PUSH_SCD
REDS-III_Brazil	phs001468_TOPMed_WGS_REDSIII_BrazilSCD
SAFS	phs001215_TOPMed_WGS_SAFHS_CVD
SAGE	phs000921_TOPMed_WGS_SAGE
Samoan	phs000972_TOPMed_WGS_SamoansAdiposity
SAPPHIRE_asthma	phs001467_TOPMed_WGS_SAPPHIRE
Sarcoidosis	phs001207_TOPMed_WGS_AA_Sarcoidosis
SARP	phs001446_TOPMed_WGS_SARP
THRV	phs001387_TOPMed_WGS_THRV
UCSF_AF	phs001933_TOPMed_CCDG_UCSF_AF
VAFAR	phs000997_TOPMed_WGS_VAFAR
VU_AF	phs001032_TOPMed_WGS_Vanderbilt_AF
walk_PHaSST	phs001514_TOPMed_WGS_Walk_PHaSST
WGHS	phs001040_TOPMed_WGS_AF_Women
WHI	phs001237_TOPMed_WGS_WHI

Many of these studies also have datasets available for “age_at_dna_blood_draw_wgs” by subject identifier “SUBJECT_ID”. To use “age_at_dna_blood_draw_wgs” in conjunction with the telomere datasets one would need to first match “SUBJECT_ID” (in the age dataset) with each study’s TOPMed sample-subject mapping subject identifier (or match subject identifier AND study name variables within a given TOPMed freeze-specific sample annotation) and then match the result by sample identifiers with the telomere dataset. When using a freeze-wide sample annotation, note that subject identifiers are unique within a study but not unique across study (hence the need to match subject identifier AND study). Please note that sample-subject mappings and freeze-specific sample annotations can change as sample swaps and other sample identity issues are identified during QC.

In addition, the telomere data is available for TOPMed control samples (HapMap and 1000 Genomes) in the dbGaP Exchange Area `Combined_Study_Data`:

```
`Investigator_Data_Products/batchAdjusted_telomereLength_20200908_Controls.tar.gz`.
```

The tar.gz file contains the dataset (tab-delimited text file), the data dictionary (tab-delimited text file in dbGaP format), and a README. Please note that this dataset includes the additional variables of "SUBJECT_ID" (subject identifier) and "age_at_dna_blood_draw_wgs".

Further dataset description

Estimated telomere length from whole-genome sequencing (WGS) samples The TelSeq method was used to perform telomere length estimation on the TOPMed WGS data. Final telomere length (TL) estimation was performed on a set of 128,901 samples from Freeze 8 whose sequencing reads were available for analysis at the TOPMed IRC at the time of analysis. Details of the TelSeq method are found in Ding et al.(1). TelSeq calculates an estimate of individual TL using counts of sequencing reads containing a fixed number of repeats of the telomeric nucleotide motif TTAGGG. Given that 98% of the TOPMed data was sequenced using read lengths of 151 or 152, we chose to use a repeat number of 12. These read counts are then normalized according to the number of reads in the individual with between 48% and 52% GC content, to adjust for potential technical artifacts related to GC content.

NOTE: If your read lengths are not 151 or 152, these TelSeq estimates were not calculated with the correct number of motif repeats and we recommend caution in using them for analysis, in particular in combination with results of other read lengths.

Batch adjustment to correct for technical confounders To account for technical sources of variability in our telomere length estimates, both within a study and across studies, we developed a method to estimate components of technical variability in our samples. We estimated these covariates using the sequencing data itself, similar to methods developed for other multivariate genomics data types (SVA or PEER factors), using aligned sequencing reads and relying on the fact that genomic coverage patterns of aligned reads can reflect technical variation.

We computed average sequencing depth for every 1,000 bp genomic region ("bin") genome-wide using mosdepth. We removed bins known to be problematic: those containing repetitive DNA sequence with difficulty mapping (mappability<1.0 using 50bp k-mers in GEMTools v1.759 15) or that overlap the list of known problematic SVs 16 or overlap known CNVs in the Database of Genomic Variants. To avoid overcorrecting for sex, bins were limited to autosomes. After normalizing the approximately 150,000 remaining bin counts within sample, we performed Randomized Singular Value Decomposition (rSVD), a scalable alternative to principal components analysis, to generate batch principal components (bPCs). We then calculated batch-adjusted telomere length estimates by regressing out 200 bPCs (i.e., taking residuals from the linear model with the original telomere length estimates as the outcome variable and the 200 bPCs as predictors). These adjusted telomere length estimates have had the mean subtracted out; if you want to rescale them back to the original scale, **the intercept that should be added to them is 3.311832.**

1. Ding, Z. et al. Estimating telomere length from whole genome sequence data. Nucleic Acids Res 42, e75, doi:10.1093/nar/gku181 (2014).