Novel genetic determinants of telomere length from a trans-ethnic analysis of 109,122 whole genome sequences in TOPMed

Margaret A Taub [1], Matthew P Conomos [2], Rebecca Keener *[3], Kruthika R Iyer *[4], Joshua S Weinstock *[5, 6], Lisa R Yanek *[7], John Lane *[8], Tyne W Miller-Fleming *[9], Jennifer A Brody [10], Caitlin P McHugh [2], Deepti Jain [2], Stephanie Gogarten [2], Cecelia A Laurie [2], Ali Keramati [11], Marios Arvanitis [12], Albert V Smith [5], Benjamin Heavner [2], Lucas Barwick [13], Lewis C Becker [7], Joshua C Bis [10], John Blangero [14], Eugene R Bleecker [15, 16], Esteban G Burchard [17, 18], Juan C Celedon [19], Yen Pei C Chang [20], Brian Custer [21, 22], Dawood Darbar [23], Lisa de las Fuentes [24], Dawn L DeMeo [25, 26], Barry I Freedman [27], Melanie E Garrett [28, 29], Mark T Gladwin [30], Susan R Heckbert [31, 32], Bertha A Hidalgo [33], Marguerite R Irvin [34], Talat Islam [35], W Craig Johnson [36], Stefan Kaab [37, 38], Lenore Launer [39], Jiwon Lee [40], Simin Liu [41], Arden Moscati [42], Kari E North [43], Patricia A Peyser [44], Nicholas Rafaels [45], Laura M Raffield [46], Christine Seidman [47], Daniel E Weeks [48, 49], Fayun Wen[83], Marsha M Wheeler [50], L. Keoki Williams [51], Ivana V Yang [45], Wei Zhao [44], Stella Aslibekyan [33], Paul L Auer [52], Donald W Bowden [53], Brian E Cade [54, 26], Zhanghua Chen [35], Michael H Cho [25], L Adrienne Cupples [55, 56], Joanne E Curran [14], Michelle Daya [45], Ranjan Deka [57], Celeste Eng [17], Tasha Fingerlin [58, 59], Xiuqing Guo [60], Lifang Hou [61], Shih-Jen Hwang [62], Jill M Johnsen [63, 64], Eimear E Kenny [65, 42], Albert M Levin [66], Chunyu Liu [56, 67], Ryan L Minster [48], Take Naseri [68], Mehdi Nouraie [30], Muagututi'a Sefuiva Reupena [69], Ester C Sabino [70], Jennifer A Smith [44], Nicholas L Smith [31, 32], Jessica Lasky Su [25, 26], James G Taylor VI[83], Marilyn J Telen [28], Hemant K Tiwari [71], Russell P Tracy [72], Marquitta J White [17], Yingze Zhang [30], Kerri L Wiggins [10], Scott T Weiss [25, 26], Ramachandran S Vasan [73, 56], Kent D Taylor [60], Moritz F Sinner [37, 38], Edwin K Silverman [25], M. Benjamin Shoemaker [74], Wayne H-H Sheu [75], Frank Sciurba [76], David Schwartz [45], Jerome I Rotter [77], Daniel Roden [78], Susan Redline [54, 79], Benjamin A Raby [80, 81], Bruce M Psaty [82, 32], Juan M Peralta [14], Nicholette D Palmer [53], Sergei Nekhai [83], Courtney G Montgomery [84], Braxton D Mitchell [20, 85], Deborah A Meyers [15, 16], Stephen T McGarvey [86], Fernando D Martinez on behalf of the NHLBI CARE Network [87], Angel CY Mak [17], Ruth JF Loos [42, 88], Rajesh Kumar [89], Charles Kooperberg [90], Barbara A Konkle [63, 64], Shannon Kelly [21, 91], Sharon LR Kardia [44], Robert Kaplan [92], Jiang He [93], Hongsheng Gui [51], Frank D Gilliland [35], Bruce Gelb [94], Myriam Fornage [95, 96], Patrick T Ellinor [97], Mariza de Andrade [98], Adolfo Correa [99], Yii-Der Ida Chen [60], Eric Boerwinkle 100, Kathleen C Barnes 45, Allison E Ashley-Koch 28, 29, Donna K Arnett 101, Christine Albert [26, 102], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium [†], TOPMed Hematology and Hemostasis Working Group [#], TOPMed Structural Variation Working Group [§], Cathy C Laurie [2], Goncalo Abecasis [5, 103], Deborah A Nickerson [50], James G Wilson [104], Stephen S Rich [105], Daniel Levy [56, 67], Ingo Ruczinski [1], Abraham Aviv [106], Thomas W Blackwell [5, 6], Timothy Thornton [107], Jeff O'Connell [108, 109], Nancy J Cox [110], James A Perry [20], Mary Armanios [111], Alexis Battle [3], Nathan Pankratz [8], Alexander P Reiner [112, 90], Rasika A Mathias [7]

** These individuals contributed equally to this work.*

[†] https://www.nhlbiwgs.org/topmed-banner-authorship; Full banner author list is included in Supplementary Information.

[#] Full working group author list is included in Supplementary Information.

[§] Full working group author list is included in Supplementary Information.

Address Correspondence to: Rasika A Mathias, ScD, rmathias@jhmi.edu, 410-550-2487.

AFFILIATIONS

1 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD USA;
2 Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA;
3 Department of Biomedical Engineering, Johns Hopkins Whiting School of Engineering, Baltimore, MD, USA;
4 Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD USA;
5 Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA;
6 Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI, USA;
7 GeneSTAR Research Program, Department of Medicine, Johns Hopkins School of Medicine, Baltimore MD USA;
8 Department of Laboratory Medicine & Pathology, University of Minnesota, Minneapolis, MN, USA;
9 Department of Medicine, Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA;
10 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA;
11 Department of Cardiology, Johns Hopkins School of Medicine, Baltimore MD USA;
12 Department of Medicine, Division of Cardiology, Johns Hopkins School of Medicine, Baltimore MD USA;
13 LTRC Data Coordinating Center, The Emmes Company, LLC, Rockville, MD, USA;
14 Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA;
15 Department of Medicine, Division of Genetics, Genomics and Precision Medicine, University of Arizona, Tucson, AZ, USA;
16 Division of Pharmacogenomics, University of Arizona, Tucson, AZ, USA;
17 Department of Medicine, University of California San Francisco, San Francisco, CA, USA;
18 Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA;
19 Division of Pediatric Pulmonary Medicine, UPMC Children's Hospital of Pittsburgh, University of Pittsburgh, Pittsburgh, PA, USA;
20 Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA;
21 Vitalant Research Institute, San Francisco, CA, USA;
22 Department of Laboratory Medicine, University of California San Francisco, San Francisco, CA, USA;
23 Division of Cardiology, University of Illinois at Chicago, Chicago, IL, USA;
24 Cardiovascular Division, Department of Medicine, Washington University School of Medicine in St. Louis, St. Louis, MO, USA;
25 Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA;
26 Harvard Medical School, Boston, MA, USA;
27 Department of Internal Medicine, Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC, USA;
28 Department of Medicine, Duke University Medical Center, Durham, NC, USA;
29 Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC, USA;
30 Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA;

31 Cardiovascular Health Research Unit and Department of Epidemiology, University of Washington, Seattle, WA, USA;
32 Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA;
33 University of Alabama at Birmingham, Birmingham, AL, USA;
34 Department of Epidemiology, University of Alabama Birmingham, Birmingham, AL, USA;
35 Division of Environmental Health, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA;
36 Department of Biostatistics, Collaborative Health Studies Coordinating Center, University of Washington, Seattle, WA, USA;
37 Department of Medicine I, University Hospital Munich, Ludwig-Maximilian's University, Munich, Germany;
38 German Centre for Cardiovascular Research (DZHK); partner site: Munich Heart Alliance, Munich, Germany;
39 Laboratory of Epidemiology and Population Science, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA;
40 Department of Medicine, Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA;
41 Department of Epidemiology, Brown University, Providence, RI, USA;
42 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
43 Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA;
44 Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA;
45 Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO, USA;
46 Department of Genetics, University of North Carolina, Chapel Hill, NC, USA;
47 Genetics, Harvard Medical School, Boston, MA, USA;
48 Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA;
49 Department of Biostatistics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA;
50 Department of Genome Sciences, University of Washington, Seattle, WA, USA;
51 Center for Individualized and Genomic Medicine Research (CIGMA), Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA;
52 Zilber School of Public Health, University of Wisconsin Milwaukee, Milwaukee WI, USA;
53 Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA;
54 Division of Sleep Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA;
55 Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA;
56 The National Heart, Lung, and Blood Institute, Boston University's Framingham Heart Study, Framingham, MA, USA;
57 Department of Environmental and Public Health Sciences, University of Cincinnati, Cincinnati, OH, USA;
58 Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA;
59 Department of Biostatistics and Informatics, University of Colorado Denver, Aurora, Colorado, USA.;
60 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA;
61 Department of Preventive Medicine, Northwestern University, Chicago, IL, USA;
62 Population Sciences Branch, Division of Intramural Research, National Heart Lung and Blood Institute, National Institute of Health, Bethesda, MD, USA;
63 Bloodworks Northwest, Research Institute, Seattle, WA, USA;
64 University of Washington, Department of Medicine, Seattle, WA, USA;

65 Center for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
66 Department of Public Health Sciences, Henry Ford Health System, Detroit, MI, USA;
67 The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, MD, USA;
68 Ministry of Health, Government of Samoa, Apia, Samoa;
69 Lutia i Puava ae Mapu i Fagalele, Apia, Samoa;
70 Instituto de Medicina Tropical da Faculdade de Medicina da Uiversidade de Sao Paulo, Sao Paulo, Brazil;
71 Department of Biostatistics, University of Alabama Birmingham, Birmingham, AL, USA;
72 Departments of Pathology & Laboratory Medicine and Biochemistry , Larrner College of Medicine, University of Vermont , Colchester, VT, USA;
73 Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA;
74 Departments of Medicine, Pharmacology, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA;
75 Division of Endocrinology and Metabolism, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan;
76 Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA;
77 The Institute for Translational Genomics and Population Sciences, Departments of Pediatrics and Medicine, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA;
78 Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA;
79 Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA;
80 Division of Pulmonary and Critical Care, Brigham and Women's Hospital, Boston, MA, USA.;
81 Division of Pulmonary Medicine, Boston Children's Hospital, Boston, MA, USA;
82 Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology and Health Services, University of Washington, Seattle, WA, USA;
83 Center for Sickle Cell Disease and Department of Medicine, College of Medicine, Howard University, Washington, DC 20059 USA;
84 Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA;
85 Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD, USA;
86 Department of Epidemiology & International Health Institute, Brown University School of Public Health, Providence, USA;
87 Asthma & Airway Disease Research Center, University of Arizona, Tucson, USA;
88 The Mindich Child Health and Development Institute, Icahn Schoold of Medicine at Mount Sinai, New York, NY, USA;
89 Division of Allergy and Clinical Immunology, The Ann and Robert H. Lurie Children's Hospital of Chicago, and Department of Pediatrics Northwestern University, Chicago, IL, USA;
90 Division of Public Health Sciences, Fred Htutchinson Cancer Research Center, Seattle WA, USA;
91 UCSF Benioff Children's Hospital, Oakland, CA, USA;
92 Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA;
93 Department of Medicine, Tulane University School of Medicine, New Orleans, LA, USA;
94 Mindich Child Health and Development Institute, Departments of Pediatrics and Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
95 Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA;

96 Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA;
97 Cardiology Division, Department of Medicine, Massachusetts General Hospital, Boston, MA;
98 Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN,USA;
99 Jackson Heart Study and Departments of Medicine and Population Health Science, Jackson, MS, USA;
100 Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA;
101 College of Public Health, University of Kentucky, Lexington, KY, USA;
102 Division of Cardiovascular, Brigham and Women's Hospital, Boston, MA;
103 Regeneron Pharmaceuticals, Tarrytown, NY, USA;
104 Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MI, USA;
105 Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA;
106 Center of Human Development and Aging, Rutgers, New Jersey Medical School, Newark, NJ, USA;
107 Department of Biostatistics, University of Washington, Seattle, WA, USA;
108 Division of Endocrinology, Diabetes and Nutrition, Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA;
109 Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA;
110 Vanderbilt Genetics Institute and Division of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, USA;
111 Department of Oncology, Johns Hopkins School of Medicine, Baltimore, MD, USA;
112 Department of Epidemiology, University of Washington, Seattle, WA, USA

ABSTRACT

Telomeres shorten in replicating somatic cells, and telomere length (TL) is associated with age-related diseases [1,2]. To date, 17 genome-wide association studies (GWAS) have identified 25 loci for leukocyte TL [3-19], but were limited to European and Asian ancestry individuals and relied on laboratory assays of TL. In this study from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program, we used whole genome sequencing (WGS) of whole blood for variant genotype calling and the bioinformatic estimation of TL in n=109,122 trans-ethnic (European, African, Asian and Hispanic/Latino) individuals. We identified 59 sentinel variants (p-value $<5x10^{-9}$) from 36 loci (20 novel, 13 replicated in external datasets). There was little evidence of effect heterogeneity across populations, and 10 loci had >1 independent signal. Fine-mapping at *OBFC1* indicated the independent signals colocalized with cell-type specific eQTLs for *OBFC1* (*STN1*). We further identified two novel genes, *DCLRE1B* (*SNM1B*) and *PARN*, using a multi-variant gene-based approach.

RESULTS

The decreasing costs of high throughput sequencing have enabled WGS data generation at an unprecedented scale, and TOPMed data offer the opportunity to address sample size, population diversity, rare variant evaluation, and fine mapping limitations of prior TL GWAS. We selected TelSeq [20] to bioinformatically determine TL due to its computational efficiency and high correlation with Southern blot [21] and flowFISH [22] measurements (***Materials and Methods. Figures S1a-c***). We developed a novel principal components-based approach to remove technical artifacts arising from the sequencing process that affected TL estimation, which improved accuracy (***Materials and Methods, Figures S1d-e***). Pooled trans-ethnic association analysis was performed with n=109,122 subjects (including 51,654 of European ancestry, 29,260 of African ancestry, 18,019 Hispanic/Latinos, 5,683 of Asian ancestry, and 4,506 of

other, mixed, or uncertain ancestries, as determined by HARE [23], ***Materials and Methods***); 44% were male and age ranged from <1 to 98 years old (***Table S1***).

Genome-wide tests for association were performed across 163M variants. Using a series of single variant tests for association (primary to identify loci, iterative conditional by chromosome to identify additional independent variants, and joint tests including all independent variants to summarize effect sizes; see ***Materials and Methods***), we identified 59 independently associated variants mapping to 36 loci, each reaching a p-value $<5x10^{-9}$ (***Figure 1, Table 1, Table S2***); 16 known and 20 novel, as further described below.

Of 25 previously known loci, we identified 16 (*PARP1, ACYP2, TERC, NAF1, TERT, POT1, TERF1, OBFC1, ATM, TINF2, DCAF4, TERF2, RFWD3, MPHOSPH6, ZNF208/ZNF257/ZNF676,* and *RTEL1*) with a variant at a p-value $<5x10^{-9}$ ***(Table 1, Table S3***). Directionally consistent and nominal evidence for replication was noted for *CTC1* (rs3027234, p-value = $7.97x10^{-5}$) and *SENP7* (rs55749605, p-value = 0.023). A signal previously attributed to *PRRC2A* is located less than 200kb from our novel signal for *HSPA1A* but may be distinct given low linkage disequilibrium (r^2=0.26). We found no evidence of replication (all variants with p-value >0.05) for the remaining previously reported TL loci (*CXCR4, PXK, MOB1B, DKK2/PAPSS1, CARMIL1* and *CSNK2A2*, ***Table S3)***. Our comprehensive conditional analyses revealed that there was more than one independent sentinel variant at nine of the sixteen previously reported loci (***Table 1, Figure 2a***). The resolution possible with our trans-ethnic WGS data identified a sentinel variant different from the one previously reported by tagging-based GWAS for 11 of the 16 known loci. At known loci *RTEL1*, *RFWD3*, *POT1*, *ACYP2*, and *PARP1,* our WGS-based sentinels included a coding missense variant in genes *RTEL1*, *RFWD3*, *POT1*, *TSPYL6*, and *PARP1*, respectively. For the remaining known TL loci, many of the non-coding sentinel variants are annotated as having regulatory evidence (RegulomeDB score < 7, ***Table 1***), as illustrated further for *OBFC1* below.

Our 20 novel loci (***Table 1***) had a total of 22 independent sentinel variants, and we tested for replication at the 19 variants available in two prior published GWAS with non-overlapping subjects [18,19] (***Figure 3a***). Variants at ten of these loci (*BCL2L15, CXXC5, HSPA1A, NOC3L, NKX2-3, ATP8B4, CLEC18C, TYMS, SAMHD1,* and *TYMP*) had a Bonferroni-corrected $p<0.05/19=0.0026$, and an additional three had variants with $p<0.05$ (*TNP03, KBTBD7,* and *BANP*), as did a second variant at *TYMS.* The variant at *SAMHD1* was previously reported at an FDR < 0.05 (p-value = $1.41\text{x}10^{-7}$)[19] but here has genome-wide significance (p-value = $1.58\text{x}10^{-19}$). While qPCR and TelSeq quantify TL in different units (see ***Materials and Methods)***, there is high consistency in the effects at variants shared between our study and these prior studies (***Figure 3b-c***). Pearson correlations of effect sizes for all 19 shared variants were 0.83 (p-value = $7.0\text{x}10^{-5}$) for our study compared to Dorajoo et al. (n=23,096 Singaporean Chinese) and 0.73 (p-value = $3.7\text{x}10^{-4}$) for our study compared to Li et al. (n=78,592 European). The correlations were stronger (0.93 and 0.84, respectively) when restricted to variants with at least nominal significance in the prior studies (***Figure 3d-e***). The proteins encoded by two of these novel genes have strong biological connections to TL: CXXC5, which physically interacts with ATM and transcriptionally regulates p53 levels [24], two proteins implicated in telomere length regulation; and BANP (aka SMAR1) which forms a complex with p53 and functions as a tumor suppressor [25].

Each of the 59 sentinel variants individually accounted for a small percentage of phenotypic variation (***Table 1***), consistent with prior GWAS of TL, but cumulatively accounted for 4.35% of TL variance, compared to 2-3% from prior GWAS [3]. The 37 variants mapping to 16 known loci explained 3.38% of TL variability, with an additional 0.96% explained by the 22 variants mapping to our 20 novel loci; a sizable gain in explained variability for TL in this trans-ethnic sample. Prior GWAS report allelic effects ranging from ~ 49-120 base pairs [3,4,11,13]. In the TOPMed data, effect sizes for common variants (minor allele frequency, MAF ≥5%) ranged from 2-59 base pairs per allele. Rare and low frequency variants (MAF <5%) showed larger effects (40-1,063 base pairs per allele).

Stratified association analyses were performed in population groups with at least 5,000 samples to evaluate effect heterogeneity of the 59 variants (***Table S4***). Reduced sample sizes, coupled with variation in allele frequency, often limited our power to detect population-specific associations at GWAS thresholds in individual strata (***Table S4***); no additional loci were identified. A major advantage of our analysis was the ability to rely on the individual-level WGS data for the iterative conditional approach to identify the final set of independent sentinel variants at each trans-ethnic-identified locus. Our sentinel variants, identified without relying on tagging through linkage to measured marker variants like prior GWAS, reveal little evidence for heterogeneity across populations (***Table 1***). All Cochran's Q[26] p-values (***Table 1***) were above a Bonferroni correction threshold (p-value>0.001), and the five with nominal significance (0.001<p-value<0.05) appear to be primarily driven by differences in the (smallest) Asian stratum. An interesting illustration of a locus with strong allele frequency differences between groups is *TINF2*; the evidence at the peak variant (rs28372734) in the trans-ethnic analysis was driven by the smaller Hispanic/Latino and Asian groups (group-specific p-values 4.6×10^{-9} and 7.3×10^{-10}, respectively), and the secondary peak (rs8016076) was driven by the African group (group-specific p-value 1.7×10^{-10}, ***Table 1, Figure 2b***). No association is noted in the European group, where these variants are nearly monomorphic (***Figure 2c***).

Gene-based tests in the pooled trans-ethnic sample identified eight protein coding genes with deleterious rare and low frequency (MAF <1%, including singletons) variants associated with TL (p-value $<1.8 \times 10^{-6}$, see ***Materials and Methods, Figure S2***). Six of these genes support a role for rare variants in previously identified GWAS loci (*POT1, TERT, RTEL1, CTC1, SAMHD1,* and *ATM).* The two novel genes have strong biological plausibility: both *DCLRE1B* and *PARN* have been implicated in short telomere syndrome (STS) patients [27-29]. DCLRE1B protein localizes to the telomere via interaction with the protein of another previously implicated GWAS gene, *TERF2*, and contributes to telomere protection from DNA repair pathways [30,31]. Notably, two *PARN* loss-of-function variants included in our gene-based test were previously identified in STS patients [27]. Both rs878853260 and rs876661305 produce frame-shift

mutations; rs876661305 produces an early termination codon, truncating most of the nuclease domain [32]. For each of these eight genes, a leave-one-out approach iterating over each variant included in the aggregate test showed there were no detectible main driver variants and indicated that these gene-based association signals arise from cumulative signal across multiple rare deleterious variants (***Figure S2)***, with the possible exception of *ATM*. When conditioned on the 59 sentinel variants, all genes, except *POT1,* maintained or increased statistical significance (***Figure S2).*** For *POT1*, while the removal of the single variant identified in ***Table 1*** (rs202187871) and conditioning on all 59 sentinels resulted in a decrease in significance from 1.52×10^{-24} to 5.53×10^{-18}, it nonetheless remained strongly significant, meeting Bonferroni thresholds.

The identification of multiple independent sentinel variants for several loci offers the unique opportunity to evaluate the potential for distinct regulatory mechanisms (***Figure 2a***, ***Figure S3*****).** OBFC1 is part of a complex that binds single-stranded telomeric DNA [33] and is expressed across multiple tissues in GTEx [34] and in whole blood studies meta-analyzed in eQTLGen [35]. All four signals at the *OBFC1* locus are in the promoter and early introns of *OBFC1* (***Figure 4a-b***). Evidence for eQTL colocalization was detected at the primary, tertiary, and quaternary signals in various tissues (***Materials and Methods***). While all three signals colocalized with *OBFC1* eQTLs, the strongest colocalization evidence in each case was in a distinct tissue: sun exposed skin from the lower leg (posterior probability of shared signal, PPH4 = 98.0%) for the primary, skeletal muscle (PPH4 = 84.4%) for the tertiary, and whole blood (GTEx PPH4 = 75.5%, eQTLGen PPH4 = 75.5%) for the quaternary signal (***Figure 4c-e, Figure S4e, Table S5***). Data from the Roadmap Epigenomics Consortium[36] indicate that all four signals are consistent with promoter or enhancer regions across blood cells and skeletal muscle tissue (***Figure 4b***). We were unable to perform colocalization analysis on the secondary signal with data from either GTEx or eQTLGen as it is driven by rare variants only in the Hispanic/Latino and Asian individuals (rs111447985, ***Table S4***).

Using individual level data within the Vanderbilt University biobank BioVU, we performed a PheWAS (***Table S6)*** using 49 available sentinel variants individually. European and African specific effect sizes from the joint analysis from ***Table 1*** were also combined to create separate polygenic trait scores (PTS) for each population group to conduct PheWAS. PTS values were significantly higher in BioVU African Americans (AAs, mean=-217bp, sd=96bp) compared to European Americans (EAs, mean=-279bp, sd=96bp, p-value<0.05, Welch's two-sample t-test, ***Figure 5a***), offering evidence that previously observed differences in TL by ancestry (longer TL in individuals of African ancestry[1]) may be explained in part by TL genetics. The largest cumulative effect of the sentinel variants, as evidenced from the PTS, is for the category of neoplasms in the EAs, with higher PTS associated with increased risk to the individual phenotypes (11 of 14 significant results, ***Figure 5b, Table S6)***; associations were only nominal in the BioVU AAs, likely due to lower power from the smaller sample size. Single variant PheWAS (***Table S6)*** in the BioVU EAs are largely replicated within the UK Biobank (UKBB, ***Table S7)***, showing strong associations with neoplasms, and in general, demonstrating the alleles that increased TL also increased risk for these cancer related phenotypes. Additionally, both the UKBB and BioVU data revealed a strong association between the novel *HSPA1A* locus (rs1008438) and type I diabetes related endocrine/metabolism phenotypes; here the allele decreasing TL increased risk for these phenotypes. This agrees with prior associations between shorter TL and increased risk of type 1 diabetes [37], and between the protein product of *HSPA1A* (Hsp72) and diabetic ketoacidosis [38].

Leveraging WGS available through the NHLBI TOPMed program, we have illustrated the value of a large, trans-ethnic WGS study to generate a harmonized phenotype of broad interest (i.e. bioinformatically called TL), to confirm known TL GWAS loci, and to identify an additional set of novel loci that map to genes with strong biological plausibility for TL association. The well-powered study enabled identification of rare deleterious variants at known and novel loci with estimated effects larger than those of common variants. Utilizing WGS allowed us the unique opportunity to hone in on causal variants using fine-mapping approaches, and begin to identify tissue-specific genetic effects. We were

also able to establish that for most population groups, effects are highly consistent at sentinel variants, despite differences in association strength at loci like *TINF2* and *OBFC1*, where allele frequencies varied among populations. The ability to implement this phenotype assessment of TL in a large, trans-ethnic dataset with pre-existing WGS creates opportunities beyond the genetics of TL. It will expand our ability to evaluate the role of TL and genes determining TL in health and human disease as illustrated by the PheWAS in large biobanks where we document identifiable effects that differ between sentinel variants and the cumulative score across all loci, and start to dissect the genetic basis to TL differences across populations.

Acknowledgements:
Chen Li, Claudia Langernberg, and Veryan Codd for providing summary statistics from their TL GWAS for replication analysis.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Specific funding sources for each study and genomic center are given in the Supplementary Information. Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The full study specific acknowledgments as well as individual acknowledgements are detailed in the Supplementary Information.

The BioVU projects at Vanderbilt University Medical Center are supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10OD017985 and S10RR025141; CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, R01HD074711; and additional funding sources listed at https://victr.vumc.org/biovu-funding/.

Support for this work was provided by the National Institutes of Health, National Heart, Lung, and Blood Institute, through the BioData Catalyst program (award 1OT3HL142479-01, 1OT3HL142478-01, 1OT3HL142481-01, 1OT3HL142480-01, 1OT3HL147154). Any opinions expressed in this document are those of the authors and do not necessarily reflect the views of NHLBI, individual BioData Catalyst Consortium members, or affiliated organizations and institutions.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

Author contributions: M.A.T., R.A.M. conceived of and led the study. M.A.T., M.P.C., R.Keener, K.R.I., L.R.Y., C.P.M., D.J., S.G., C.A.L., A.K., M.Arvanitis, J.C.B., E.G.B., J.C.C., Y.C.C., L.M.R., M.H.C., J.E.C., M.D., B.M.P., C.C.L., D.L., I.R., T.W.B., J.A.P., M.Armanios, A.B., A.P.R., R.A.M. drafted the manuscript. M.A.T., M.P.C., R.Keener, J.S.W., J.A.B., D.J., A.K., C.C.L., G.A., D.A.N., J.G.W., S.S.R., D.L., I.R., A.A., T.W.B., T.T., J.O., N.J.C., J.A.P., M.Armanios, A.B., N.P., A.P.R., R.A.M. contributed substantive analytical guidance. M.A.T., M.P.C., R.Keener, K.R.I., J.S.W., L.R.Y., J.Lane, T.W.M., J.A.B., C.P.M., D.J., S.G., C.A.L., A.K., M.Arvanitis, A.V.S., B.H., T.T., N.J.C., M.Armanios, A.B., N.P., A.P.R., R.A.M. performed and led analysis. L.R.Y., L.B., L.C.B., J.C.B., J.B., E.R.B., E.G.B., J.C.C., Y.C.C., B.C., D.D., L.d., D.L.D., B.I.F., M.E.G., M.T.G., S.R.H., B.A.H., M.R.I., T.I., W.C.J., S.Kaab, L.L., J.Lee, S.L., A.M., K.E.N., P.A.P., N.R., L.M.R., C.S., D.E.W., M.M.W., L.W., I.V.Y., W.Z., S.A., P.L.A., D.W.B., B.E.C., Z.C., M.H.C., L.A.C., J.E.C., M.D., R.D., C.E., T.F., X.G., L.H., S.H., J.M.J., E.E.K., A.M.L., C.L., R.L.M., T.N., M.N., M.S.R., E.C.S., J.A.S., N.L.S., J.L.S., M.J.T., H.K.T., R.P.T., M.J.W., Y.Z., K.L.W., S.T.W., R.S.V., K.D.T., M.F.S., E.K.S., M.S., W.H.S., F.S., D.S., J.I.R., D.R., S.R., B.A.R., B.M.P., J.M.P., N.D.P., S.N., C.G.M., B.D.M., D.A.M., S.T.M., F.D.M., A.C.M., R.J.L., R.Kumar, C.K., B.A.K., S.Kelly, S.L.K., R.Kaplan, J.H., H.G., F.D.G., B.G., M.F., P.T.E., M.d., A.C., Y.I.C., E.B., K.C.B., A.E.A., D.K.A., C.A., N.J.C., M.Armanios were involved

in the guidance, collection and analysis for one or more of the studies which contributed data to this manuscript. All authors read and approved the final draft.

Author Information

Data Deposition Statement: TOPMed genomic data and pre-existing parent study phenotypic data are made available to the scientific community in study-specific accessions in the database of Genotypes and Phenotypes (dbGaP) (https://www.ncbi.nlm.nih.gov/gap/?term=TOPMed). Telomere length calls were derived from the raw sequence data as described in the Online Methods, and the phenotype covariates of age, sex, and race/ethnicity are available through the study-specific dbGAP accession IDs as listed in the Supplementary Information.

Competing Interests: The authors declare the following competing interests:
J.C.C. has received research materials from GlaxoSmithKline and Merck (inhaled steroids) and Pharmavite (vitamin D and placebo capsules) to provide medications free of cost to participants in NIH-funded studies, unrelated to the current work.
B.I.F. is a consultant for AstraZeneca Pharmaceuticals and RenalytixAI
L.W. is on the advisory board for GlaxoSmithKline and receives grant funding from NIAID, NHLBI, and NIDDK, NIH
I.V.Y. is a consultant for ElevenP15
S.A. receives equity and salary from 23andMe, Inc.
M.H.C. receives grant support from GlaxoSmithKline
S.T.W. receives royalties from UpToDate
E.K.S. received grant support from GlaxoSmithKline and Bayer in the past three years.
B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson.
F.D.M. is supported by grants from NIH/NHLBI (HL139054,HL091889,HL132523,HL130045,HL098112,HL056177), the NIH/NIEHS (ES006614), the NIH/NIAID (AI126614), and the NIH/ Office of Director (OD023282). Vifor Pharmaceuticals provided medicine and additional funding to support recruitment for HL130045. Dr. Martinez is a council member for the Council for the Developing Child.
P.T.E. is supported by a grant from Bayer AG to the Broad Institute focused on the genetics and therapeutics of cardiovascular diseases. Dr. Ellinor has also served on advisory boards or consulted for Bayer AG, Quest Diagnostics, and Novartis.
K.C.B. receives royalties from UpToDate
G.A. is an employee of Regeneron Pharmaceuticals and owns stock and stock options for Regeneron Pharmaceuticals.

Correspondence and requests for materials should be addressed to *Rasika A Mathias, ScD, rmathias@jhmi.edu, 410-550-2487*

ALL TABLE LEGENDS

Table 1: 59 independently associated variants mapping to 36 loci from the whole genome sequencing of n=109,122 TOPMed individuals. Loci are labeled as novel if none of the sentinel variants in the locus was in LD ($r^2 < 0.7$) with any previously documented GWAS signal for telomere length. There are 5 variants marked with an * where the primary analysis did not meet our threshold of $p<5x10^{-9}$, however they reached significance after conditioning on significant variants mapping to the chromosome (detailed in ***Table S2***). Variants marked with ** are direct matches to prior reported sentinel variants. Percent of trait variation explained by each variant is provided from single-variant association tests. P-values and effect sizes (in base pairs) are reported from a joint model including all variants. P-values for effect heterogeneity across population groups were generated using Cochran's Q statistic. MAC is the minor allele count from the full combined sample. For all exonic variants, detailed annotation is provided, while for all non-coding variants the RegulomeDB score is given. See also ***Tables S2, S4***.

Table S1, Related to TOPMed study populations, Materials and Methods: Sample demographics summarized by each set of analysis performed: pooled trans-ethnic and four non-overlapping population groups (defined with HARE, using reported race/ethnicity and genetically inferred ancestry in combination; "Other/Uncertain" includes individuals with maximum stratum probability from HARE < 0.7, as well as Brazilians and Samoans who were excluded from the HARE analysis).

Table S2, Related to Table 1 and Figure 2: Results of the iterative conditional analysis using individual level data. Results are presented by chromosome providing a detailed overview of the conditional step at which each variant was identified as the peak signal. (Note that some variants had p-value$>5x10^{-9}$ in the primary analysis, but p-value$<5x10^{-9}$ in a conditional step). Within each chromosome, variants are ordered not on position, but in the sequence of identification through the conditional analysis showing the iterative process used: Primary and Rounds 1-6 of conditioning. Chromosomes varied in the number of analyses needed until no additional variants were included (maximum steps = 7 on chromosome 5).

Table S3, Related to Assessing novelty of identified loci and variants, Materials and Methods: TOPMed results from the primary pooled trans-ethnic analysis for all prior telomere length GWAS sentinel variants with reported p-value $<5x10^{-8}$ in prior published studies.

Table S4, Related to Table 1: Association results for the single variant association analysis for each of the 59 sentinel variants from ***Table 1***. Single variant results are shown for the pooled trans-ethnic analysis, and each population group. Variants with a minor allele count <5 were not included in the analysis and are listed as "-".

Table S5, Related to Figure 4: Iterative conditional analysis was repeated on chromosome 10 focusing exclusively on the *OBFC1* locus, defined as a 2Mb window around the original top sentinel, rs9420907. The sentinel for each signal was consistent with the iterative conditional analysis performed on the entirety of chromosome 10 (***Table S2***). The summary statistics from each iterative conditional analysis were used to perform colocalization analysis on the non-primary signals. Colocalization analysis was performed using coloc with all significant gene-tissue pairs in GTEx and with all genes in a 2Mb window around rs9420907 in eQTLGen. All results with PPH4 > 0.7 for each signal are reported.

Table S6, Related to Figure 5: Results of the PheWAS performed on 49 available sentinel variants and a polygenic trait score across these 49 sentinel variants using BioVU self-identified African Americans (AA, n=15,174) and BioVU self-identified European Americans (EA, n=70,439). The cumulative TL risk score for the BioVU AA and EA samples was derived from the African and European specific effects sizes from ***Table 1***, respectively. Results were evaluated at a Bonferroni threshold corrected for the number of informative phecodes for each variant.

Table S7, Related to Figure 5: Results of the PheWAS conducted using SAIGE, a method employing generalized linear mixed models and allowing for imbalance between case and control counts, adjusting for genetic relatedness, sex, birth year and the first 4 principal components of ancestry, on approximately 400,000 UK Biobank (UKBB) participants. Results were downloaded from files provided by the University of Michigan PheWeb server (http://pheweb.sph.umich.edu/SAIGE-UKB/about) on 47 available sentinel variants. Results were evaluated at a Bonferroni threshold corrected for the number of phecodes available for each variant (N=1,403).

ALL FIGURE LEGENDS

Figure 1: Genome-wide Manhattan plot. Trans-ethnic genome-wide tests for association using 163M sequence identified variants on n=109,122 samples with sequence generated telomere length from TOPMed. All loci had a peak $p<5x10^{-9}$ in the pooled trans-ethnic analysis. Prior known loci are indicated in red, and novel loci are indicated in blue.

Figure 2: LocusZoom plots for multi-hit loci and *TINF2*. [a] LocusZoom plots for all loci with >1 sentinel variant. Linkage disequilibrium (LD) was calculated from the set of samples used in the analysis with respect to the peak variant in the pooled trans-ethnic primary analysis, thereby reflecting LD patterns specific to the TOPMed samples. For each figure, the peak sentinel variant from the pooled trans-ethnic analysis is indexed and labeled in purple, and all independent variants identified through the iterative conditional approach are labeled in green and highlighted with green dotted lines. **[b]** LocusZoom plots for four population groups for the *TINF2* locus. **[c]** Forest plots displaying effect sizes and standard errors, as well as minor allele frequencies, by population group for the three sentinel variants in *TINF2*. See also ***Table S2***.

Figure 3: Replication by two prior studies. [a] Replication results for the 20 novel TOPMed loci, including 22 variants, pulled from two non-overlapping, prior published GWAS on telomere length using qPCR data. Threshold indicates the stronger level of significance between the two replication studies (Bonferroni significant: $p <0.0026$; nominally significant: $p <0.05$; non-significant: $p >0.05$). **[b-e]** Correlation between the estimated effect sizes for all novel loci (**b-c**) and replicated novel loci (**d-e**, $p<0.05$ in each dataset) between the present TOPMed pooled trans-ethnic analysis and two prior published GWAS studies.

Figure 4: Fine-mapping of multiple *OBFC1* signals. [a] LocusZoom plot of the *OBFC1* locus where green dotted lines indicate each independent signal, as in ***Figure 2***. **[b]** Roadmap Epigenomics Consortium data in hg19 coordinates for skeletal muscle tissue, Primary T $CD4^+$ memory cells from peripheral blood, and Primary T $CD8^+$ naïve cells from peripheral blood (Roadmap samples E108, E037, and E047 respectively; data was not available for sun exposed skin). ChromHMM state model is shown for the 18-state auxiliary model. The state model suggests the primary (rs9420907), secondary (rs111447985), and tertiary (rs112163720) signals are in the promoter region, while the quaternary signal (rs10883948) is in an enhancer region in all Roadmap blood cell types but is transcriptional for peripheral blood monocytes and $CD19^+$ B-cells. **[c-e]** GWAS and eQTL results for the primary, tertiary and quaternary signals. Top panels are the GWAS summary statistics from the primary, and iterative conditional analyses which were used to perform colocalization analysis (secondary signal was rare, and not available for colocalization). Bottom panels are eQTLs for *OBFC1* in the indicated tissue from GTEx. The GTEx eQTLs for these tissues do not colocalize with one another ($PPH4 < 4.4x10^{-7}$) and each signal did not significantly colocalize in the other tissues. LD was calculated from the pooled trans-ethnic samples with respect to the sentinel (black diamond). See also ***Figures S3, S4, Table S5***.

Figure 5: PheWAS results from Vanderbilt BioVU. Polygenic trait score (PTS) analysis across 49 available sentinel variants in BioVU. **[a]** Smoothed distributions of PTS values for European Americans (n=70,439) and African Americans (n=15,174) from BioVU biobank. **[b]** Overview of the PheWAS in the BioVU European Americans. ▲ = higher PTS is associated with the phenotype. ▼ = higher PTS is protective against the phenotype. See also ***Tables S6, S7***.

Figure S1, Related to Estimating telomere length for whole-genome sequencing (WGS) samples and Batch adjustment to correct for technical confounders, Materials and Methods: For 2,389 samples from the Jackson Heart Study (JHS), **[a]** scatter plot with Pearson correlation between TelSeq and Computel length estimates; [**b]** Comparison of computational times for TelSeq and Computel; **[c]** scatter

plots with Pearson correlations between TelSeq (left) and Computel (right) and Southern blot TL estimates; **[d]** scatter plot with Pearson correlation between TelSeq and Southern blot TL estimates after adjustment for the final set of 200 batch principal components (bPCs) used in our full analysis. Colors indicate on which plate samples were shipped to the sequencing center, in Panels A, C and D; and **[e]** scatter plot with Pearson correlation between bPC-adjusted TelSeq and flowFISH data on 19 samples from the GeneSTAR study.

Figure S2, Related to Gene-based coding variant tests - Tests for association, Materials and Methods: Eight genes were identified as passing the Bonferroni threshold based on number of genes tested (p-value < $0.05/27{,}558 = 1.8\text{x}10^{-6}$). For each gene, a leave-one-out analysis was performed iterating the SMMAT test and leaving one variant out at a time. The plots show the change in SMMAT p-value for each variant (orange line with marker) relative to the variant's allele frequency (blue bar), the overall gene-based test including all variants (dotted red line) and the single variant results for all variants with an MAC≥5 that were included in single variant tests for association (brown and green diamonds). For each gene, the number of rare and deleterious variants included in SMMAT is indicated. For any variant with a MAC≥5, a single variant test was also performed as part of the primary analysis. The count of these variants is indicated. In addition the SMMAT p-value for these genes when conditioning on the 59 sentinel variants is also given.

Figure S3, Related to Figure 4: [a] LocusZoom plots for four population groups for the *OBFC1* locus. Linkage disequilibrium (LD) was calculated from the set of samples used in the analysis with respect to the peak variant in the pooled trans-ethnic primary analysis, thereby reflecting LD patterns specific to the TOPMed samples. For each figure, the peak sentinel variant from the pooled trans-ethnic analysis is indexed and labeled in purple, and all independent variants identified through the iterative conditional approach are labeled in green and highlighted with green dotted lines. **[b]** Forest plots displaying effect sizes and standard errors, as well as minor allele frequencies, by population group for the four sentinel variants in *OBFC1*.

Figure S4, Related to Figure 4: [a-d] Credible set analysis and colocalization analysis in eQTLGen. Manhattan plots for each *OBFC1* signal are shown where p-values were taken from the appropriate conditional analysis output and LD was calculated with respect to the sentinel variant. Credible set variants are indicated with black diamonds; the sentinel variant is indicated as a black diamond with red outline. **[e]** Manhattan plot for colocalization analysis of the *OBFC1* quaternary signal with an *OBFC1* eQTL in eQTLGen. PPH3 and PPH4 for colocalization are indicated in the top right corner of the eQTL plot. LD was calculated with respect to the sentinel, indicated with a black diamond on each graph, and with pooled trans-ethnic analysis samples.

REFERENCES

1 Aviv, A. & Shay, J. W. Reflections on telomere dynamics and ageing-related diseases in humans. *Philos Trans R Soc Lond B Biol Sci* **373**, doi:10.1098/rstb.2016.0436 (2018).

2 McNally, E. J., Luncsford, P. J. & Armanios, M. Long telomeres and cancer risk: the price of cellular immortality. *J Clin Invest* **129**, 3474-3481, doi:10.1172/JCI120851 (2019).

3 Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet* **45**, 422-427, 427e421-422, doi:10.1038/ng.2528 (2013).

4 Codd, V. *et al.* Common variants near TERC are associated with mean telomere length. *Nat Genet* **42**, 197-199, doi:10.1038/ng.532 (2010).

5 Delgado, D. A. *et al.* Genome-wide association study of telomere length among South Asians identifies a second RTEL1 association signal. *J Med Genet* **55**, 64-71, doi:10.1136/jmedgenet-2017-104922 (2018).

6 Gu, J. *et al.* A genome-wide association study identifies a locus on chromosome 14q21 as a predictor of leukocyte telomere length and as a marker of susceptibility for bladder cancer. *Cancer Prev Res (Phila)* **4**, 514-521, doi:10.1158/1940-6207.CAPR-11-0063 (2011).

7 Lee, J. H. *et al.* Genome wide association and linkage analyses identified three loci-4q25, 17q23.2, and 10q11.21-associated with variation in leukocyte telomere length: the Long Life Family Study. *Front Genet* **4**, 310, doi:10.3389/fgene.2013.00310 (2013).

8 Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc Natl Acad Sci U S A* **107**, 9293-9298, doi:10.1073/pnas.0911494107 (2010).

9 Liu, Y. *et al.* A genome-wide association study identifies a locus on TERT for mean telomere length in Han Chinese. *PLoS One* **9**, e85043, doi:10.1371/journal.pone.0085043 (2014).

10 Mangino, M. *et al.* DCAF4, a novel gene associated with leucocyte telomere length. *J Med Genet* **52**, 157-162, doi:10.1136/jmedgenet-2014-102681 (2015).

11 Mangino, M. *et al.* Genome-wide meta-analysis points to CTC1 and ZNF676 as genes regulating telomere homeostasis in humans. *Hum Mol Genet* **21**, 5385-5394, doi:10.1093/hmg/dds382 (2012).

12 Mangino, M. *et al.* A genome-wide association study identifies a novel locus on chromosome 18q12.2 influencing white cell telomere length. *J Med Genet* **46**, 451-454, doi:10.1136/jmg.2008.064956 (2009).

13 Pooley, K. A. *et al.* A genome-wide association scan (GWAS) for mean telomere length within the COGS project: identified loci show little association with hormone-related cancer risk. *Hum Mol Genet* **22**, 5056-5064, doi:10.1093/hmg/ddt355 (2013).

14 Prescott, J. *et al.* Genome-wide association study of relative telomere length. *PLoS One* **6**, e19635, doi:10.1371/journal.pone.0019635 (2011).

15 Saxena, R. *et al.* Genome-wide association study identifies variants in casein kinase II (CSNK2A2) to be associated with leukocyte telomere length in a Punjabi Sikh diabetic cohort. *Circ Cardiovasc Genet* **7**, 287-295, doi:10.1161/CIRCGENETICS.113.000412 (2014).

16 Walsh, K. M. *et al.* Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk. *Nat Genet* **46**, 731-735, doi:10.1038/ng.3004 (2014).

17 Zeiger, A. M. *et al.* Genetic Determinants of Telomere Length in African American Youth. *Sci Rep* **8**, 13265, doi:10.1038/s41598-018-31238-3 (2018).

18 Dorajoo, R. *et al.* Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies. *Nat Commun* **10**, 2491, doi:10.1038/s41467-019-10443-2 (2019).

19 Li, C. *et al.* Genome-wide Association Analysis in Humans Links Nucleotide Metabolism to Leukocyte Telomere Length. *Am J Hum Genet* **106**, 389-404, doi:10.1016/j.ajhg.2020.02.006 (2020).

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

21 Kimura, M. *et al.* Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat Protoc* **5**, 1596-1607, doi:10.1038/nprot.2010.124 (2010).

22 Alder, J. K. *et al.* Diagnostic utility of telomere length testing in a hospital-based setting. *Proc Natl Acad Sci U S A* **115**, E2358-E2365, doi:10.1073/pnas.1720427115 (2018).

23 Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet* **105**, 763-772, doi:10.1016/j.ajhg.2019.08.012 (2019).

24 Zhang, M. *et al.* The CXXC finger 5 protein is required for DNA damage-induced p53 activation. *Sci China C Life Sci* **52**, 528-538, doi:10.1007/s11427-009-0083-7 (2009).

25 Kaul, R. *et al.* Direct interaction with and activation of p53 by SMAR1 retards cell-cycle progression at G2/M phase and delays tumor growth in mice. *Int J Cancer* **103**, 606-615, doi:10.1002/ijc.10881 (2003).

26 Cochran, W. G. The combination of estimates from different experiments. *Biometrics* **10**, 101-129 (1954).

27 Stuart, B. D. *et al.* Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat Genet* **47**, 512-517, doi:10.1038/ng.3278 (2015).

28 Tummala, H. *et al.* Poly(A)-specific ribonuclease deficiency impacts telomere biology and causes dyskeratosis congenita. *J Clin Invest* **125**, 2151-2160, doi:10.1172/JCI78963 (2015).

29 Touzot, F. *et al.* Function of Apollo (SNM1B) at telomere highlighted by a splice variant identified in a patient with Hoyeraal-Hreidarsson syndrome. *Proc Natl Acad Sci U S A* **107**, 10097-10102, doi:10.1073/pnas.0914918107 (2010).

30 van Overbeek, M. & de Lange, T. Apollo, an Artemis-related nuclease, interacts with TRF2 and protects human telomeres in S phase. *Curr Biol* **16**, 1295-1302, doi:10.1016/j.cub.2006.05.022 (2006).

31 Lenain, C. *et al.* The Apollo 5' exonuclease functions together with TRF2 to protect telomeres from DNA repair. *Curr Biol* **16**, 1303-1310, doi:10.1016/j.cub.2006.05.021 (2006).

32 Wu, M. *et al.* Structural insight into poly(A) binding and catalytic mechanism of human PARN. *EMBO J* **24**, 4082-4093, doi:10.1038/sj.emboj.7600869 (2005).

33 Stewart, J. A., Wang, Y., Ackerson, S. M. & Schuck, P. L. Emerging roles of CST in maintaining genome stability and human disease. *Front Biosci (Landmark Ed)* **23**, 1564-1586, doi:10.2741/4661 (2018).

34 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).

35 Vosa, U. e. a. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* (2018).

36 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

37 Januszewski, A. S. *et al.* Shorter telomeres in adults with Type 1 diabetes correlate with diabetes duration, but only weakly with vascular function and risk factors. *Diabetes Res Clin Pract* **117**, 4-11, doi:10.1016/j.diabres.2016.04.040 (2016).

38 Oglesbee, M. J., Herdman, A. V., Passmore, G. G. & Hoffman, W. H. Diabetic ketoacidosis increases extracellular levels of the major inducible 70-kDa heat shock protein. *Clin Biochem* **38**, 900-904, doi:10.1016/j.clinbiochem.2005.05.011 (2005).

Table 1: 59 independently associated variants mapping to 36 loci from the whole genome sequencing of n=109,122 TOPMed individuals. Loci are labeled as novel if none of the sentinel variants in the locus was in LD ($r^2 < 0.7$) with any previously documented GWAS signal for telomere length. There are 5 variants marked with an * where the primary analysis did not meet our threshold of $p<5x10^{-9}$, however they reached significance after conditioning on significant variants mapping to the chromosome (detailed in ***Table S2***). Variants marked with ** are direct matches to prior reported sentinel variants. Percent of trait variation explained by each variant is provided from single-variant association tests. P-values and effect sizes (in base pairs) are reported from a joint model including all variants. P-values for effect heterogeneity across population groups were generated using Cochran's Q statistic. MAC is the minor allele count from the full combined sample. For all exonic variants, detailed annotation is provided, while for all non-coding variants the RegulomeDB score is given. See also ***Tables S2, S4***.

							Single variant anlaysis in pooled trans-ethnic sample		P-values from Joint Model					Effect Sizes from Joint Model					Cochran's Q (p-value)
Chr	Position	Locus	SNP	Novelty	Annotation	MAC	P-value	Percent variation explained	Trans-ethnic	European	African	Hispanic/ Latino	Asian	Trans-ethnic	European	African	Hispanic/ Latino	Asian	
1	35259602	ZMYM4	rs11581846	Novel	7	87500	3.04E-10	0.036%	1.74E-11	7.99E-05	8.80E-03	3.37E-06	4.13E-01	-19.9	-21.1	-15.0	-24.6	-8.7	0.44
1	113881455	BCL2L15	rs2296176	Novel	7	33274	2.84E-10	0.036%	1.52E-10	6.57E-08	2.38E-03	3.16E-01	2.07E-02	-19.5	-22.4	-22.0	-7.3	-26.9	0.28
1	226367601	PARP1	rs1136410		missense	36395	9.32E-20	0.076%	9.14E-22	6.10E-09	1.02E-02	7.51E-07	2.04E-04	-29.3	-26.0	-23.7	-29.5	-32.5	0.87
2	54255416	ACYP2	rs17189743		missense	3741	7.18E-12	0.043%	4.24E-11	2.70E-06	2.52E-01	5.20E-06	1.94E-01	-55.4	-50.7	-30.4	-82.4	-46.8	0.34
2	54263623		rs144980386		deletion	28773	1.32E-17	0.067%	1.99E-17	7.78E-13	7.32E-04	1.45E-04	6.78E-01	27.5	33.6	21.2	30.5	4.8	0.08
3	117584223	LINC00901	rs961617801	Novel	6	12	1.25E-11	0.042%	4.81E-11	2.82E-10	-	-	-	1009.6	1038.5	-	-	-	-
3	169769649	TERC	rs12637184		4	47452	1.30E-96	0.399%	1.58E-103	1.02E-50	1.84E-15	7.32E-30	2.33E-11	-59.8	-57.0	-65.8	-66.1	-57.3	0.51
3	169772313		rs9826466		N/A	4066	3.25E-17	0.065%	3.66E-21	2.04E-01	7.28E-19	2.96E-03	-	-77.0	253.9	-77.6	-77.8	-	0.25
3	190053412	P3H2	rs10937417	Novel	7	80209	6.89E-10	0.035%	1.89E-10	9.04E-05	1.65E-05	2.33E-03	6.30E-01	14.6	13.5	17.9	17.1	-4.4	0.16
4	9928595	SLC2A2	rs4235345	Novel	6	44302	3.82E-09	0.032%	1.88E-09	3.24E-07	4.39E-02	1.86E-02	5.70E-02	16.8	19.8	12.8	13.2	47.1	0.41
4	163126692		rs60735607 *		6	57418	0.00396382	0.008%	4.43E-12	4.45E-07	7.20E-03	5.33E-04	9.94E-01	-18.5	-20.0	-12.9	-22.4	-0.1	0.43
4	163144568	NAF1	rs113580095		7	290	4.72E-18	0.069%	1.63E-16	2.57E-08	7.12E-02	3.40E-07	-	-254.7	-231.9	-287.8	-257.9	-	0.89
4	163155406		rs1351222		7	50104	3.27E-26	0.103%	1.60E-32	9.04E-16	1.50E-07	6.80E-11	6.58E-03	32.7	32.7	28.9	41.2	28.9	0.49
5	1272383		rs192999400		5	2470	3.10E-15	0.057%	8.21E-23	2.15E-01	3.90E-17	3.42E-03	6.19E-02	101.3	88.7	97.4	96.9	108.4	1.00
5	1280823		rs6897196		5	102081	1.87E-83	0.344%	6.04E-13	7.24E-04	7.90E-08	1.87E-02	1.46E-01	20.8	16.2	25.2	16.8	23.4	0.55
5	1285859	TERT	rs7705526 **		7	64162	1.64E-92	0.382%	2.01E-18	1.16E-11	7.73E-04	1.46E-03	6.47E-02	30.0	35.5	22.5	27.1	34.1	0.47
5	1287079		rs2853677 **		5	80905	8.17E-65	0.265%	1.25E-19	9.27E-07	8.88E-10	1.33E-05	8.34E-02	-23.9	-17.7	-33.9	-28.9	-22.0	0.08
5	1292331		rs34052286		3a	6173	1.50E-12	0.046%	5.47E-22	3.86E-01	1.09E-14	2.39E-05	-	-66.0	-59.3	-61.6	-74.3	-	0.80
5	1292843		rs114616103 *		7	4527	2.39E-07	0.024%	2.03E-13	1.44E-09	2.09E-02	1.72E-02	4.93E-01	-57.2	-59.2	-37.9	-57.6	-71.9	0.73
5	139637905	CXXC5	rs75903170	Novel	2b	11895	5.69E-10	0.035%	7.01E-10	2.83E-04	2.58E-06	8.14E-02	4.31E-01	29.6	25.5	46.7	22.3	12.6	0.19
6	31815431	HSPA1A	rs1008438	Novel	2b	106908	3.42E-17	0.065%	6.64E-19	3.61E-09	1.40E-04	4.10E-07	1.71E-02	-20.3	-19.7	-18.4	-25.7	-20.5	0.73
7	124812616	POT1	rs720613		7	62325	1.27E-26	0.105%	1.37E-27	2.59E-18	5.87E-06	3.63E-05	2.51E-02	-26.3	-31.3	-20.4	-24.6	-21.0	0.26
7	124858989		rs202187871		missense	27	4.89E-12	0.044%	6.72E-13	6.42E-12	-	-	-	738.3	719.6	-	-	-	-
7	129041243	TNP03	rs7783384	Novel	5	92528	2.34E-12	0.045%	6.10E-12	1.22E-07	1.47E-03	2.54E-02	2.04E-02	-15.2	-17.8	-13.3	-11.4	-19.8	0.64
8	73004218	TERF1	rs183633026		7	1132	1.59E-10	0.038%	1.45E-10	3.48E-02	7.67E-01	1.51E-09	-	99.4	75.4	-22.0	109.6	-	0.18
8	73033303		rs73687065		5	1676	3.10E-12	0.045%	8.10E-12	7.74E-01	2.76E-10	5.71E-03	-	85.3	24.9	87.4	97.0	-	0.74
8	73046483		rs10112752		7	78968	4.59E-13	0.048%	9.83E-12	5.13E-09	5.69E-03	4.16E-02	2.83E-02	-15.8	-19.2	-12.6	-11.2	-26.2	0.40
10	94344908	NOC3L	rs3758526	Novel	missense	32511	6.80E-12	0.043%	5.77E-13	1.63E-05	2.70E-05	1.38E-02	1.42E-02	-22.0	-21.1	-22.6	-19.8	-23.1	0.99
10	99514276	NKX2-3	rs10883359	Novel	7	54905	3.60E-12	0.044%	9.34E-11	5.46E-05	5.14E-06	3.46E-02	1.70E-01	-16.5	-14.5	-28.1	-11.9	-11.9	0.19
10	103907794	OBFC1	rs10883948		7	94489	2.04E-34	0.137%	3.97E-12	1.20E-05	3.99E-04	2.18E-02	6.42E-04	-18.8	-15.7	-24.2	-13.9	-46.8	0.11
10	103915847		rs112163720 *		4	15559	0.44000458	0.001%	4.86E-16	9.57E-07	4.34E-07	3.69E-04	9.38E-04	37.1	48.1	39.1	35.3	54.9	0.66
10	103916707		rs9420907 **		3a	54838	3.90E-83	0.342%	6.80E-54	3.65E-18	6.94E-19	6.79E-13	9.41E-01	-49.2	-44.3	-52.4	-53.8	-2.4	0.30
10	103918153		rs111447985		2a	2391	2.29E-24	0.095%	3.03E-35	4.94E-03	2.24E-02	5.44E-22	2.81E-10	131.9	120.7	98.1	143.4	137.2	0.77
11	108158382	ATM	rs61380955		7	105969	2.47E-17	0.066%	1.11E-18	5.79E-14	2.97E-04	2.47E-03	9.19E-02	-19.6	-24.8	-15.6	-15.7	-14.6	0.24
13	41150640	KBTBD7	rs1411041	Novel	6	85572	6.29E-14	0.052%	6.65E-15	1.46E-08	6.04E-04	1.13E-03	1.77E-02	22.4	25.5	20.5	20.2	20.8	0.87
14	24242592	TINF2	rs28372734		4	2648	1.74E-27	0.108%	1.27E-30	4.91E-02	3.42E-06	4.59E-09	7.26E-10	112.6	120.9	103.9	132.1	94.8	0.59
14	24243052		rs8016076		2b	1977	1.80E-11	0.041%	4.46E-13	1.01E-01	1.70E-10	6.73E-03	-	83.8	374.8	80.9	87.5	-	0.43
14	24254544		rs41293824		5	1543	1.31E-09	0.034%	1.87E-10	7.43E-01	1.83E-07	2.58E-04	-	83.1	40.9	76.5	125.7	-	0.40
14	72959582	DCAF4	rs2572		5	20731	5.14E-12	0.044%	6.70E-14	2.00E-07	1.07E-04	3.42E-03	1.75E-03	28.0	27.7	36.5	25.5	33.8	0.80
15	50065546	ATP8B4	rs7172615	Novel	4	41027	4.31E-09	0.032%	3.53E-10	1.14E-07	3.77E-03	1.96E-01	1.90E-01	-17.8	-20.3	-21.4	-8.7	-12.0	0.41
16	69357811	TERF2	rs9925619		7	66224	3.01E-14	0.053%	7.84E-15	3.03E-04	1.01E-07	1.02E-06	5.10E-01	18.6	13.1	22.5	27.9	8.9	0.10
16	70193527	CLEC18C	rs62049363	Novel	7	61724	4.09E-10	0.036%	3.25E-11	5.24E-07	1.44E-01	4.80E-04	3.52E-01	-16.8	-16.7	-11.2	-19.4	-8.5	0.69
16	74630845	RFWD3	rs7193541		missense	93079	1.47E-16	0.063%	3.18E-17	3.39E-12	5.63E-05	1.66E-03	5.21E-01	-18.7	-22.9	-16.8	-16.9	-5.6	0.24
16	82166498	MPHOSPH6	rs2967355		6	34993	3.96E-19	0.073%	2.04E-20	2.69E-11	1.58E-04	4.86E-07	9.91E-01	-28.2	-26.2	-33.8	-36.8	-0.1	0.05
16	87961594	BANP	rs12934497	Novel	5	77109	9.15E-10	0.034%	8.16E-10	1.53E-05	1.62E-03	5.66E-03	1.15E-01	14.6	14.1	15.2	15.2	31.8	0.86
18	650764	TYMS	rs150119891 *		5	1320	2.27E-07	0.025%	1.92E-11	3.69E-10	2.80E-01	2.89E-03	-	-98.8	-104.9	-52.3	-160.9	-	0.32
18	666625		rs8088781	Novel	5	25774	2.49E-15	0.057%	8.91E-32	2.78E-15	1.75E-08	8.74E-09	8.66E-02	-50.6	-56.0	-40.8	-59.3	-160.9	0.21
18	676473		rs2612101 *		5	56194	0.40769627	0.001%	6.29E-16	7.76E-07	2.11E-04	5.13E-08	3.89E-02	26.0	29.4	18.4	35.1	171.4	0.05
18	44666476	SETBP1	rs2852770	Novel	7	46513	1.00E-11	0.042%	1.15E-12	2.32E-09	4.37E-02	1.23E-04	7.01E-01	-19.0	-25.1	-9.4	-24.3	-4.1	0.03
19	22032639	ZNF257/ZNF676	rs8105767 **		6	76591	3.59E-18	0.069%	1.52E-18	6.32E-09	1.14E-07	5.00E-03	5.14E-03	20.3	20.8	22.2	15.0	25.5	0.68
20	36922795	SAMHD1	rs2342113	Novel	6	51830	4.62E-18	0.069%	1.58E-19	2.50E-13	4.00E-04	1.34E-06	3.67E-01	-23.7	-33.8	-16.1	-27.4	-7.7	0.01
20	63661765	RTEL1	rs41308088		5	14105	1.58E-10	0.038%	8.42E-16	6.46E-15	6.24E-02	4.61E-02	2.05E-01	37.1	45.3	25.1	20.1	58.1	0.12
20	63676585		rs79981941		6	21592	1.46E-23	0.092%	1.37E-12	5.06E-06	4.73E-03	5.48E-06	9.09E-02	-26.7	-26.4	-18.6	-39.1	-47.5	0.25
20	63678201		rs41309367 **		5	71377	2.52E-33	0.133%	1.01E-42	5.92E-23	1.49E-15	9.58E-09	6.05E-01	-34.1	-37.0	-38.0	-33.0	-5.1	0.02
20	63689775		rs35640778		missense	2354	2.31E-28	0.112%	1.47E-38	3.77E-29	2.06E-07	1.11E-04	9.25E-01	-140.5	-141.8	-176.3	-116.7	13.8	0.42
20	63695521		rs181080831		synonymous	675	6.40E-17	0.064%	7.53E-19	8.00E-18	8.50E-01	1.39E-02	-	180.5	199.5	15.5	133.0	-	0.07
22	50532618	TYMP	rs361725	Novel	3a	101750	1.36E-10	0.038%	8.26E-12	3.97E-11	2.32E-01	4.11E-02	2.39E-02	-15.6	-22.1	-5.4	-10.6	-21.2	0.02
X	66015290	VSIG4	rs12394264	Novel	4	50912	9.67E-16	0.059%	5.36E-17	1.57E-07	1.94E-05	4.13E-05	7.28E-01	19.0	19.2	15.7	20.6	13.3	0.85
X	154720412	GAB3	rs2728723	Novel	5	63517	1.21E-12	0.046%	3.47E-12	4.45E-08	8.67E-02	7.96E-04	4.55E-02	13.2	15.7	6.0	14.4	17.0	0.17

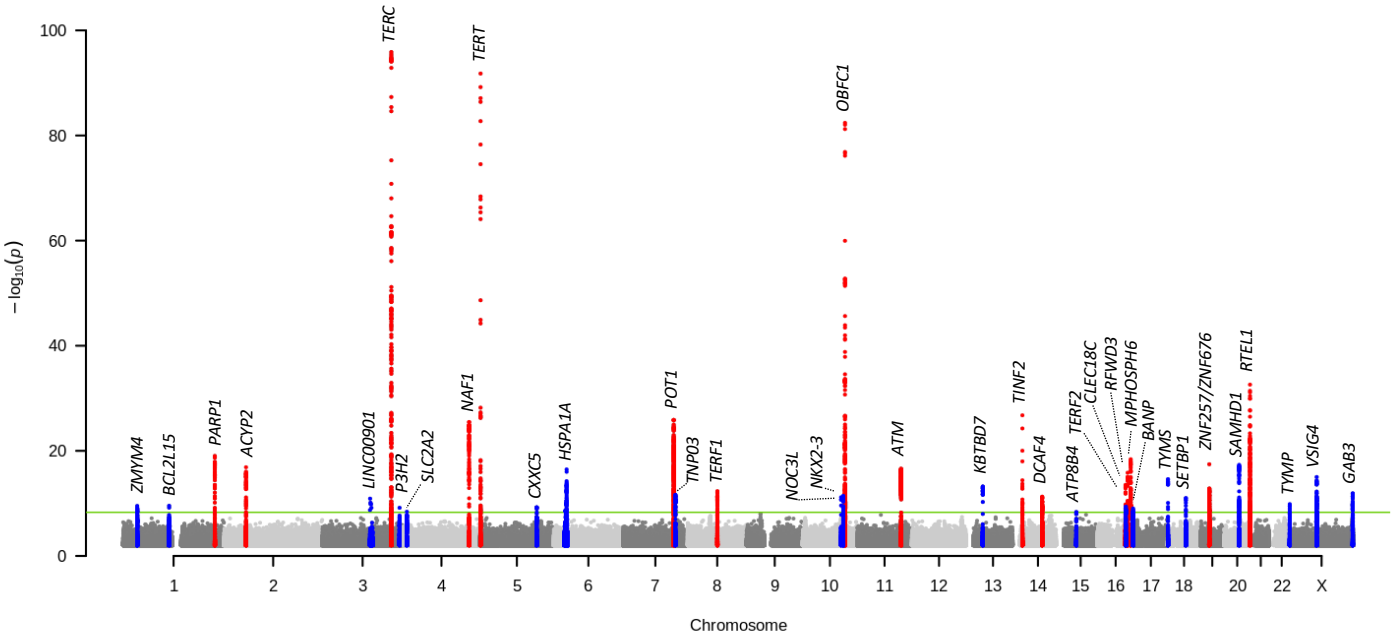

Figure 1: Genome-wide Manhattan plot. Trans-ethnic genome-wide tests for association using 163M sequence identified variants on n=109,122 samples with sequence generated telomere length from TOPMed. All loci had a peak $p<5x10^{-9}$ in the pooled trans-ethnic analysis. Prior known loci are indicated in red, and novel loci are indicated in blue.

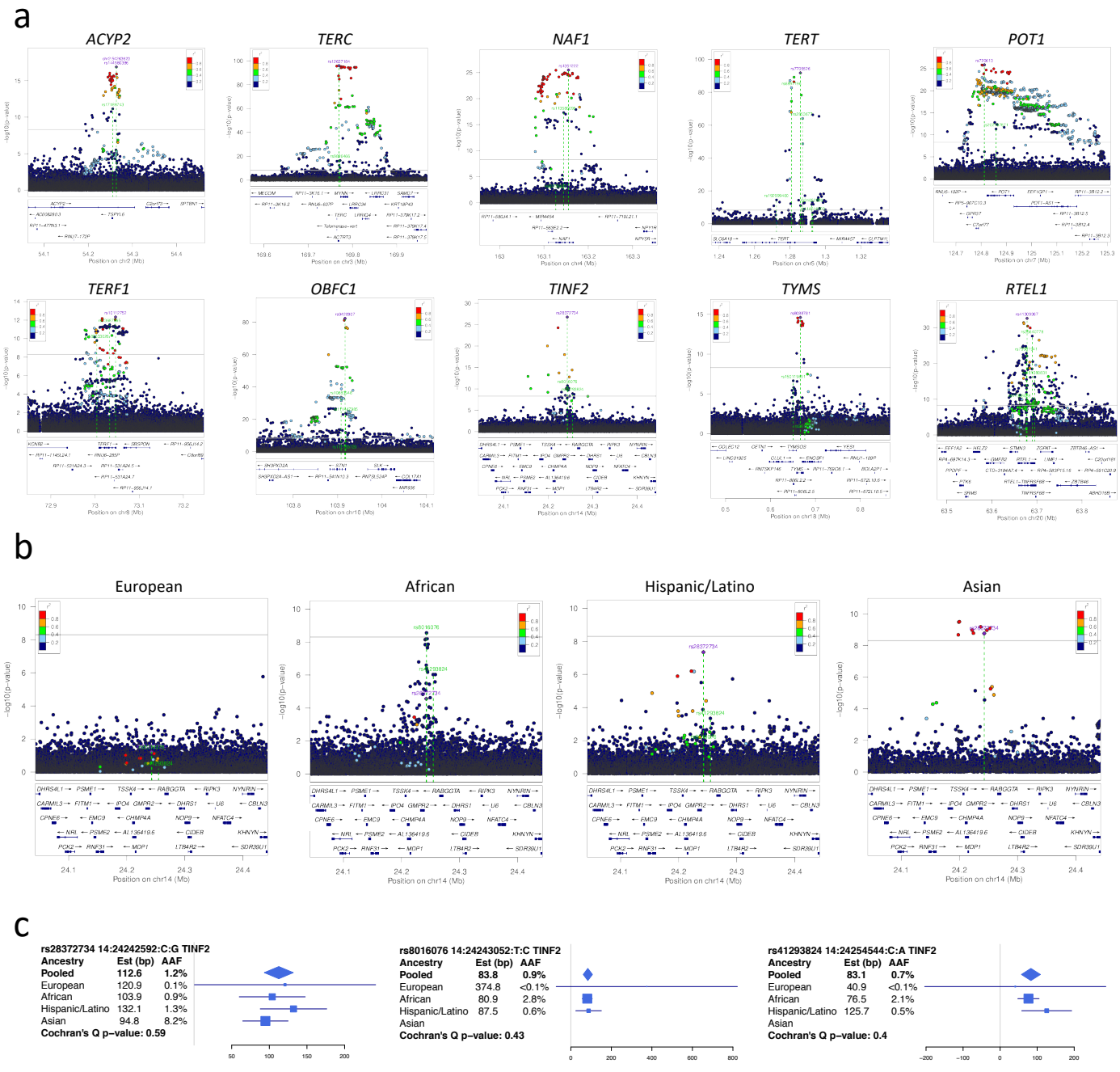

Figure 2: LocusZoom plots for multi-hit loci and *TINF2*. [a] LocusZoom plots for all loci with >1 sentinel variant. Linkage disequilibrium (LD) was calculated from the set of samples used in the analysis with respect to the peak variant in the pooled trans-ethnic primary analysis, thereby reflecting LD patterns specific to the TOPMed samples. For each figure, the peak sentinel variant from the pooled trans-ethnic analysis is indexed and labeled in purple, and all independent variants identified through the iterative conditional approach are labeled in green and highlighted with green dotted lines. **[b]** LocusZoom plots for four population groups for the *TINF2* locus. **[c]** Forest plots displaying effect sizes and standard errors, as well as minor allele frequencies, by population group for the three sentinel variants in *TINF2*. See also ***Table S2***.

a

Chr	Position	Locus	SNP	TOPMed effect size in bp	Li et al. 2020 European (N=78,592) P-value	Beta	SE(Beta)	Dorajoo et al. 2019 Singaporean Chinese (N=23,096) P-value	Beta	SE(Beta)	Replication threshold
1	35259602	ZMYM4	rs11581846	-19.9	2.53E-01	-0.01	0.01	6.09E-01	-0.01	0.01	>0.05
1	**113881455**	**BCL2L15**	**rs2296176**	**-19.5**	**1.69E-04**	**-0.02**	**0.01**	**6.86E-05**	**-0.05**	**0.01**	**<0.0026**
3	117584223	LINC00901	rs961617801	1009.6	-	-	-	-	-	-	-
3	190053412	P3H2	rs10937417	14.6	2.49E-01	0.01	0.01	6.80E-01	0.00	0.01	>0.05
4	9928595	SLC2A2	rs4235345	16.8	5.62E-02	0.01	0.01	9.13E-01	0.00	0.03	>0.05
5	**139637905**	**CXXC5**	**rs75903170**	**29.6**	**3.15E-05**	**0.05**	**0.01**	**3.40E-04**	**0.06**	**0.02**	**<0.0026**
6	**31815431**	**HSPA1A**	**rs1008438**	**-20.3**	**9.38E-05**	**-0.02**	**0.01**	**4.47E-03**	**-0.03**	**0.01**	**<0.0026**
7	**129041243**	**TNP03**	**rs7783384**	**-15.2**	**8.29E-03**	**-0.01**	**0.01**	**2.80E-02**	**-0.02**	**0.01**	**<0.05**
10	**94344908**	**NOC3L**	**rs3758526**	**-22.0**	**6.31E-05**	**-0.03**	**0.01**	**1.68E-02**	**-0.02**	**0.01**	**<0.0026**
10	**99514276**	**NKX2-3**	**rs10883359**	**-16.5**	**1.50E-03**	**-0.02**	**0.01**	**2.18E-07**	**-0.05**	**0.01**	**<0.0026**
13	**41150640**	**KBTBD7**	**rs1411041**	**22.4**	**6.38E-03**	**0.02**	**0.01**	**5.85E-03**	**0.03**	**0.01**	**<0.05**
15	**50065546**	**ATP8B4**	**rs7172615**	**-17.8**	**2.88E-07**	**-0.03**	**0.01**	**2.10E-03**	**-0.03**	**0.01**	**<0.0026**
16	**70193527**	**CLEC18C**	**rs62049363**	**-16.8**	**4.87E-04**	**-0.03**	**0.01**	**1.38E-03**	**-0.03**	**0.01**	**<0.0026**
16	**87961594**	**BANP**	**rs12934497**	**14.6**	**3.84E-02**	**0.01**	**0.00**	**5.10E-01**	**0.02**	**0.02**	**<0.05**
18	650764		rs150119891	-98.8	2.81E-01	-0.03	0.03	-	-	-	>0.05
18	**666625**	**TYMS**	**rs8088781**	**-50.6**	**2.41E-07**	**-0.04**	**0.01**	**-**	**-**	**-**	**<0.0026**
18	**676473**		**rs2612101**	**26.0**	**3.16E-02**	**-0.01**	**0.01**	**-**	**-**	**-**	**<0.05**
18	44666476	SETBP1	rs2852770	-19.0	1.87E-01	-0.01	0.01	7.41E-01	0.00	0.01	>0.05
20	**36922795**	**SAMHD1**	**rs2342113**	**-23.7**	**2.78E-05**	**-0.04**	**0.01**	**1.57E-02**	**-0.02**	**0.01**	**<0.0026**
22	**50532618**	**TYMP**	**rs361725**	**-15.6**	**1.72E-03**	**-0.02**	**0.01**	**5.41E-05**	**-0.04**	**0.01**	**<0.0026**
X	66015290	VSIG4	rs12394264	19.0	-	-	-	-	-	-	-
X	154720412	GAB3	rs2728723	13.2	-	-	-	-	-	-	-

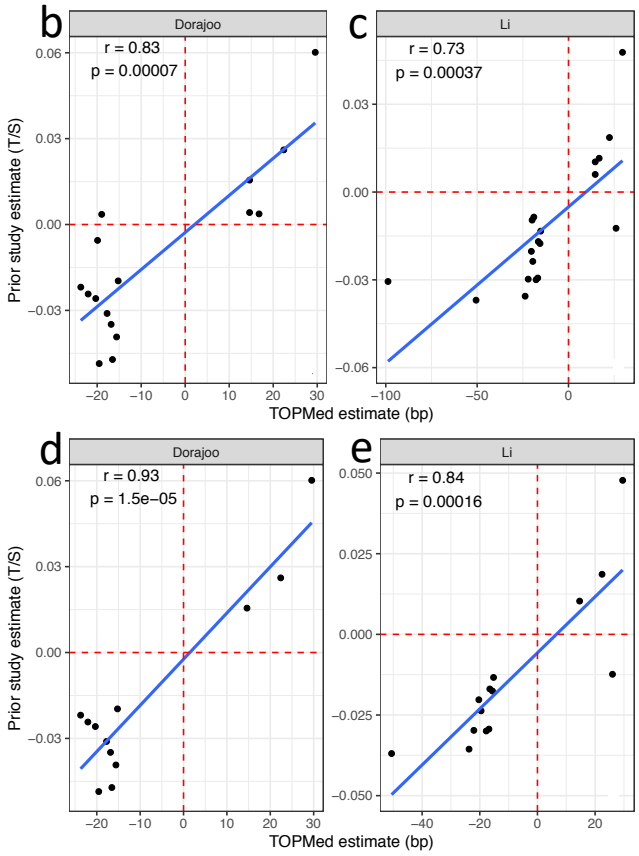

Figure 3: Replication by two prior studies. [a] Replication results for the 20 novel TOPMed loci, including 22 variants, pulled from two non-overlapping, prior published GWAS on telomere length using qPCR data. Threshold indicates the stronger level of significance between the two replication studies (Bonferroni significant: p <0.0026; nominally significant: p <0.05; non-significant: p >0.05). **[b-e]** Correlation between the estimated effect sizes for all novel loci (**b,c**) and replicated novel loci (**d,e**, p<0.05 in each dataset) between the present TOPMed pooled trans-ethnic analysis and two prior published GWAS studies.

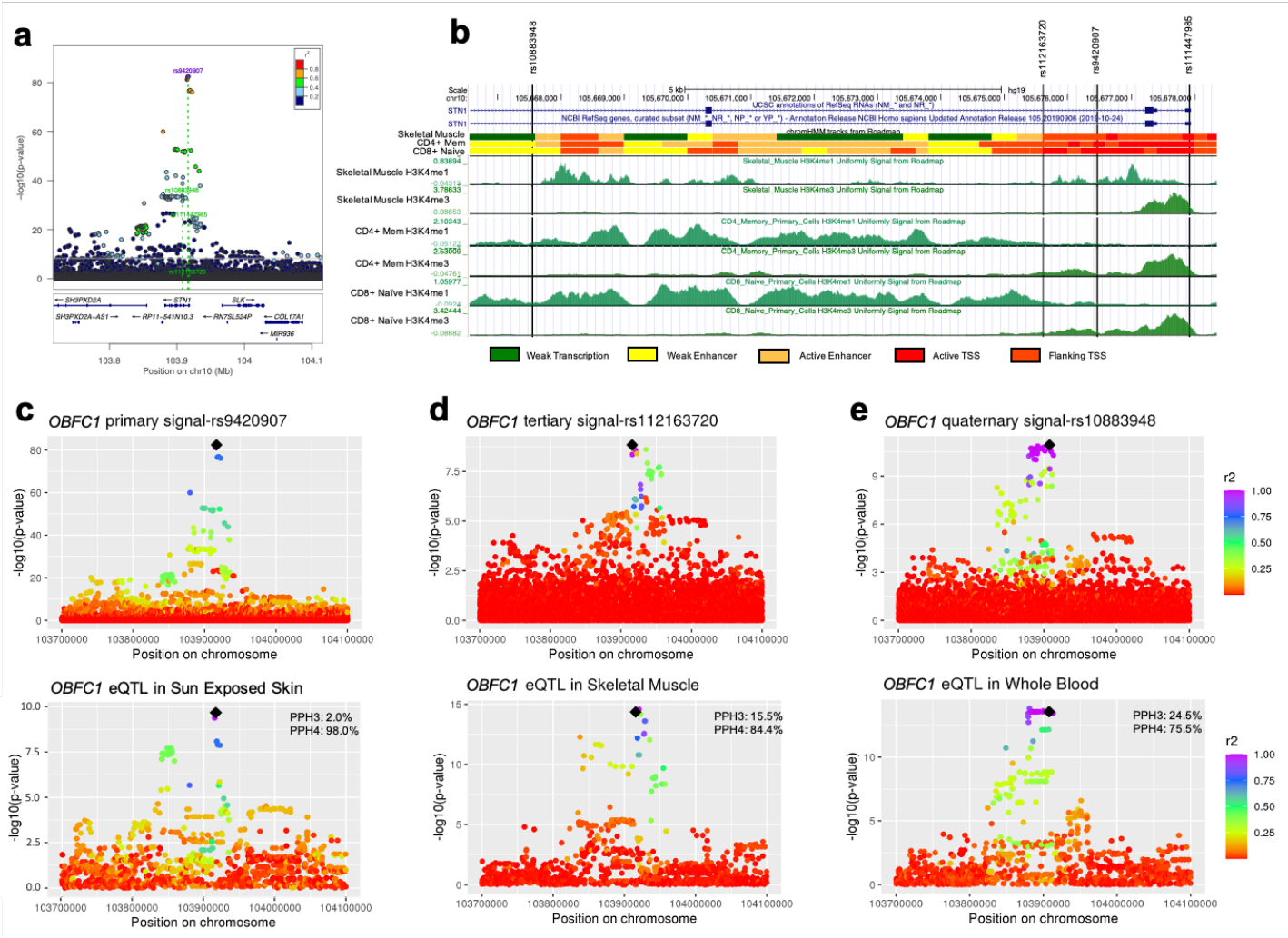

Figure 4: Fine-mapping of multiple *OBFC1* signals. [a] LocusZoom plot of the *OBFC1* locus where green dotted lines indicate each independent signal, as in ***Figure 2***. **[b]** Roadmap Epigenomics Consortium data in hg19 coordinates for skeletal muscle tissue, Primary T CD4^{+} memory cells from peripheral blood, and Primary T CD8^{+} naïve cells from peripheral blood (Roadmap samples E108, E037, and E047 respectively; data was not available for sun exposed skin). ChromHMM state model is shown for the 18-state auxiliary model. The state model suggests the primary (rs9420907), secondary (rs111447985), and tertiary (rs112163720) signals are in the promoter region, while the quaternary signal (rs10883948) is in an enhancer region in all Roadmap blood cell types but is transcriptional for peripheral blood monocytes and CD19^{+} B-cells. **[c-e]** GWAS and eQTL results for the primary, tertiary and quaternary signals. Top panels are the GWAS summary statistics from the primary, and iterative conditional analyses which were used to perform colocalization analysis (secondary signal was rare, and not available for colocalization). Bottom panels are eQTLs for *OBFC1* in the indicated tissue from GTEx. The GTEx eQTLs for these tissues do not colocalize with one another (PPH4 < 4.4x10^{-7}) and each signal did not significantly colocalize in the other tissues. LD was calculated from the pooled trans-ethnic samples with respect to the sentinel (black diamond). See also ***Figures S3,S4, Table S5***.

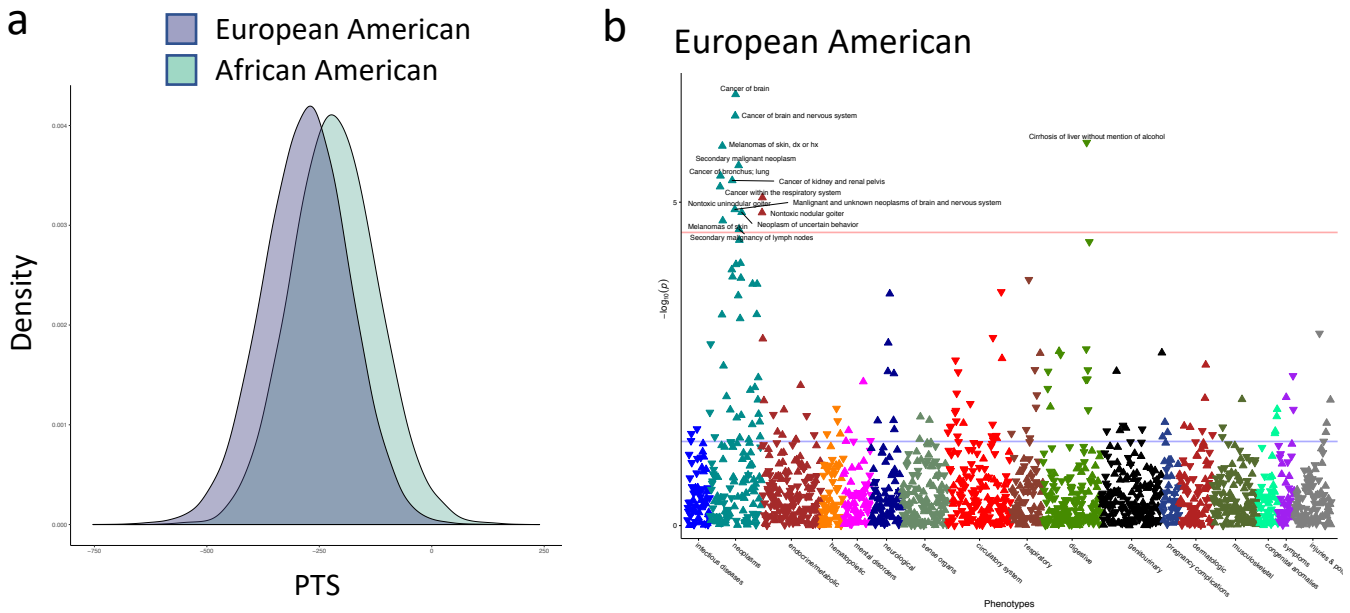

Figure 5: PheWAS results from Vanderbilt BioVU. Polygenic trait score (PTS) analysis across 49 available sentinel variants in BioVU. **[a]** Smoothed distributions of PTS values for European Americans (n=70,439) and African Americans (n=15,174) from BioVU biobank. **[b]** Overview of the PheWAS in the BioVU European Americans. ▲ = higher PTS is associated with the phenotype. ▼ = higher PTS is protective against the phenotype. See also ***Tables S6, S7***.

MATERIALS AND METHODS

TOPMed study populations: To perform this trans-ethnic genome-wide association study of telomere length, we leveraged the whole genome sequence samples available through the NHLBI Trans Omics for Precision Medicine (TOPMed) program. The program currently consists of more than 80 participating studies [1] ,with a range of study designs as described in Taliun et al [2] (*Nature, submitted, 2019*). Participants are mainly U.S. residents with diverse ancestry, race, and ethnicity (European, African, Hispanic/Latino, Asian, and Other). Smaller representation comes from non-US populations including Samoan, Brazilian, and Asian studies. Details on the specific samples included for telomere length analysis are outlined below, summarized in ***Table S1***, and described by TOPMed [1].

TOPMed whole genome sequencing (WGS): WGS was performed to an average depth of 38X using DNA isolated from blood, PCR-free library construction, and Illumina HiSeq X technology. Details for variant calling and quality control are described in Taliun et al. [2] (*Nature, submitted, 2019*). Briefly, variant discovery and genotype calling was performed jointly, across all the available TOPMed Freeze 8 studies, using the GotCloud [3] pipeline resulting in a single multi-study genotype call set.

Estimating telomere length for whole-genome sequencing (WGS) samples: A variety of computational tools exist that leverage WGS data to generate an estimate of telomere length [4]. Here, we performed a thorough comparison of two leading methods for estimating telomere length from WGS data to choose the preferred scalable method for performing the estimation on all available samples from TOPMed. The first method, TelSeq [5], calculates an estimate of individual telomere length using counts of sequencing reads containing a fixed number of repeats of the telomeric nucleotide motif TTAGGG. Given that 98% of our data was sequenced using read lengths of 151 or 152 (as confirmed from the SEQ field in the analyzed CRAM files), we chose to use a repeat number of 12. These read counts are then normalized according to the number of reads in the individual WGS data set with between 48% and 52% GC content to adjust for potential technical artifacts related to GC content. The second method, Computel [6] uses an

alignment-based method to realign all sequenced reads from an individual to a "telomeric reference sequence". Reads aligning to this reference sequence are considered to be telomeric and are included in the estimate of telomere length. Because Computel performs a complete realignment, additional computational steps are involved compared to those needed for TelSeq.

To compare the results and scalability from these two methods, we first directly compared estimates obtained from TelSeq and Computel on 2,389 samples from the Jackson Heart Study (JHS) and found them to be highly correlated with one another (Pearson correlation r=0.98, ***Figure S1a***). We also compared computational time to generate the telomere length estimates on these samples and show that Computel is around ten times more time-consuming (***Figure S1b***). This is in part due to the fact that Computel requires CRAM-formatted files (as the WGS data are currently stored) to first be converted back to Fastq format (while TelSeq requires a CRAM to BAM conversion), but also due to the computationally expensive step of realignment to the telomeric reference genome that the Computel algorithm employs.

Telseq generates an estimate of TL in bp similar to laboratory assays such as Southern blot [7] and flowFISH [8]; in contrast qPCR approaches are represented as T/S ratios[9,10]. As a further comparison to orthogonally measured telomere length values, we used data on the same 2,389 samples from JHS with Southern blot[7] telomere length estimates [11]. For these samples, the Southern blot assay was performed on the same source DNA sample that was used to generate the WGS in TOPMed. The Pearson correlation values between the TelSeq and Computel estimates and the Southern blot estimates did not differ (r=0.58 and 0.56 for TelSeq and Computel, respectively, ***Figure S1c***). Based on our observation that both Computel and TelSeq showed similar correlation to the Southern blot estimates and high correlation with each other, and that TelSeq was an order of magnitude more computationally efficient, we chose to use TelSeq to perform telomere length estimation on our data. Final telomere length estimation was

53 performed on a set of 128,901 samples whose CRAM-files were available for analysis at the TOPMed
54 IRC at the time of analysis.
55
56 **Batch adjustment to correct for technical confounders:** To account for technical sources of variability
57 in our telomere length estimates, both within a study (see, for example, colors in ***Figures S1a*** and ***S1b***
58 which indicate grouping by shared 96-well plate for shipment to the sequencing center) and across
59 studies, we developed a method to estimate components of technical variability in our samples. We
60 estimated these covariates using the sequencing data itself, similar to methods developed for other
61 multivariate genomics data types (SVA or PEER factors [12,13]), using aligned sequencing reads and relying
62 on the fact that genomic coverage patterns of aligned reads can reflect technical variation.
63
64 We computed average sequencing depth for every 1,000 bp genomic region ("bin") genome-wide using
65 mosdepth [14]. We removed bins known to be problematic: those containing repetitive DNA sequence with
66 difficulty mapping (mappability<1.0 using 50bp k-mers in GEMTools v1.759 [15]) or that overlap the list of
67 known problematic SVs [16] or overlap known CNVs in the Database of Genomic Variants. To avoid
68 overcorrecting for sex, bins were limited to autosomes. After normalizing the approximately 150,000
69 remaining bin counts within sample, we performed Randomized Singular Value Decomposition [17]
70 (rSVD), a scalable alternative to principal components analysis, to generate batch principal components
71 (bPCs). We included increasing numbers of bPCs in a linear regression model predicting TelSeq TL, and
72 computed the correlation of the resulting residuals with external data measurements, including Southern
73 blot measurements for JHS (n=2,389) and the Women's Health Initiative (WHI; n=596) and age at blood
74 draw (JHS n=3,294; WHI n=10,708). Based on the observed correlation, the final decision was to include
75 the first 200 bPCs across all samples. Using the n=2,389 JHS samples described above, we compared TL
76 estimates before and after batch correction. The percent of variance in TL explained by sequencing plate
77 reduced from 21.9% (baseline) to 10.5% (200 bPCs), and the variance explained by age increased from
78 8.0% (baseline) to 10.3% (200 bPCs), evidence that the signal-to-noise ratio was improved. Overall, the

correlation between the bPC corrected TL and Southern blot data improved from r=0.58 to 0.68 (***Figure S1d***) in the JHS data and from r=0.54 to 0.72 for the WHI data. Further, we compared TelSeq estimates of 19 samples within a single sequencing batch from the GeneSTAR study to the clinical gold standard of flowFISH [8] (***Figure S1e***) and observed a correlation of 0.80 in both granulocytes and lymphocytes. Therefore, our data show that we are able to reduce the sequencing artifacts stemming from batch variability to attain correlation of TelSeq to Southern blot similar to the correlation of TelSeq to flowFISH.

Samples included in genetic analysis: All samples with telomere length estimated from the WGS data from TOPMed Freeze 8 were considered for inclusion, provided they had consent that allowed for genetic analysis of telomere length. Only samples with sequencing read lengths of 151 or 152 base pairs and having age at blood draw data available were included. For the set of samples that were part of a duplicate pair/group (either part of the intended duplicates designed by TOPMed, or a duplicate identified across the studies through sample QC) only one sample from each duplicated pair/group was retained. The final counts and demographic summary statistics for subjects grouped by TOPMed study for all 54 studies included in our analysis are shown in ***Table S1.***

While reported race/ethnicity data are available in TOPMed, these data have limitations for analysis that include individuals with missing information or non-specific responses (e.g., 'other' or 'multiple') and high variability in genetically inferred measures of ancestry among individuals with the same reported race/ethnicity. To overcome these limitations, we used a computational method called HARE (harmonized ancestry and race/ethnicity), a newly developed machine learning approach for jointly leveraging reported and genetic data in the definition of population strata for GWAS [18]. HARE uses provided race/ethnicity labels and genetic ancestry principal component (PC) values to compute probability estimates for each individual's membership in each race/ethnicity stratum. For our HARE analysis, we used provided race (Asian, Black, White) or Hispanic ethnicity group (Central American,

Costa Rican, Cuban, Dominican, Mexican, Puerto Rican, South American) as input labels to define population strata, and we used 11 PCs computed with PC-AiR using 638,486 LD-pruned ($r^2 < 0.1$) autosomal variants with minor allele frequency > 1% to represent genetic ancestry. Genetic outliers for population strata were identified as individuals for whom their maximum stratum probability was more than 5 times greater than their reported stratum probability. Stratum values for genetic outliers and individuals with missing or non-specific race/ethnicity were imputed as the stratum for which they had the highest membership probability.

Our primary analysis allowed for heterogeneous residual variance (see **Primary single variant tests for association** for details) among groups defined jointly by study and HARE-based population stratum assignment, with minor study-specific modifications to account for small strata. We required at least 30 individuals within a study-HARE grouping and collapsed individuals into merged HARE groups within a study as necessary to retain everyone for analysis. For our population-specific analyses, we used HARE assignment to stratify individuals into the following population groups: African (corresponding to the Black HARE stratum), Asian (Asian), European (White), and Hispanic/Latino (Central American, Costa Rican, Cuban, Dominican, Mexican, Puerto Rican, and South American). To better preserve genetic ancestry similarity among individuals in population-specific stratified analyses, we restricted to individuals for whom their HARE population stratum membership probability was at least 0.7; the population stratum counts in ***Table S1*** reflect the counts in the stratified analyses, where individuals not meeting this criterion are labeled as "Other/Uncertain".

Samoan individuals from the Samoan Adiposity Study and Brazilian individuals from the Reds-III Brazil study were excluded from the HARE analyses due to their unique ancestry in the TOPMed dataset; these studies were treated as their own population groups for analyses.

Primary single variant tests for association:

Genome-wide tests for association were performed using the R Bioconductor package GENESIS [19]. The primary analysis included all available trans-ethnic TOPMed samples (n=109,122). A secondary analysis was performed for all population groups with n>5,000, which included European (n=51,654), African (n=29,260), Hispanic/Latino (n=18,019) and Asian (n=5,683) groups as defined above using HARE. Prior to genetic modeling, we generated residuals from a linear regression model on all 109,122 samples with 200 batch principal components (bPCs), as described above; for clarity we call these residuals TL_{bPC} below. For the pooled trans-ethnic analysis, we used a fully-adjusted two-stage model, as described in the next two bullets[20]. For each population-specific analysis, the same approach was used, limited to samples within that population group.

- **Stage 1:** We fit a linear mixed model (LMM) on n=109,122 samples, using TL_{bPC} as the outcome; adjusting for age, sex, study, sequencing center, and 11 PC-AiR [21] PCs of ancestry as fixed effect covariates; including a random effect with covariance matrix proportional to a sparse empirical kinship matrix computed with PC-Relate[22] to account for genetic relatedness among samples; and allowing for heteroskedasticity of residual variance across study-HARE groups as defined above. The marginal residuals from this Stage 1 model were then inverse-normalized and rescaled by their original standard deviation. This rescaling restores values to the original trait scale, providing more meaningful effect size estimates from subsequent association tests [23].
- **Stage 2:** We fit a second LMM on all n=109,122 samples, using the inverse-normalized and rescaled residuals from Stage 1 as the outcome; all other aspects of the model including fixed effects adjustment, random effects, and residual variance structure were identical to the model in Stage 1. This two-stage covariate adjustment has been shown to be most effective at controlling for false-positives and increasing statistical power in this setting[20]. The output of this Stage 2 model was then used to perform both single variant and gene-based tests for association.

Single variant tests for association: We used the output of the two-stage LMM to perform score tests of association for each variant with minor allele count (MAC) ≥ 5 that passed TOPMed Informatics Research Center (IRC) at the University of Michigan quality filters [2] and which had <10% of samples with read depth <10. Genotype effect size estimates and percent of variability explained (PVE) were approximated from the score test results [24].

Assessing significance, performing conditional analysis to identify independent variants, and defining genetic loci: A p-value cutoff of $5x10^{-9}$ was used to determine genome-wide significance in the primary trans-ethnic analysis. We identified our set of independent significant variants (as reported in ***Table 1***) through an iterative conditioning process within each chromosome. For a given chromosome, if at least one variant from the primary analysis crossed the genome-wide significance cutoff, this peak variant was included as an additional fixed-effect covariate in a new two-stage LMM (see Stages 1 and 2 described above), and score test results were examined to see if any remaining variants crossed the $5x10^{-9}$ threshold. If so, we performed a second round of conditioning, including both the original peak variant and the new conditional peak variant as fixed-effect covariates in another two-stage LMM; and so on, adding conditional peak variants for additional rounds (***Table S2***). For each chromosome, the conditioning procedure was completed when no additional variants crossed the genome-wide threshold ($p < 5x10^{-9}$) on that chromosome. At each step, all variants passing the $p < 5x10^{-9}$ threshold were examined in BRAVO [25] to assess quality, and 334 variants were filtered out due to variant call quality issues. In the case where a current peak variant was flagged for quality, the next most significant variant, provided its p-value was below the $5x10^{-9}$ cutoff, was considered the peak variant instead. Variants were grouped into loci based on physical distance and an examination of linkage disequilibrium (LD) patterns, and locus names were determined using a combination of previous literature, known telomere biology, and physical location.

Cumulative Percent of Variability Explained (PVE)

Through the iterative conditional approach, we identified a total of 59 variants (***Table 1***) that met our genome-wide significance threshold of $p < 5x10^{-9}$. The cumulative PVE values for this full set of 59 variants (4.35%), the set of 37 variants mapping to known loci (3.38%), and the set of 22 variants mapping to novel loci (0.96%, see **Assessing novelty of identified loci and variants** below for definition of novel variants) were each estimated jointly using approximations from multi-parameter score tests. This joint PVE approximation is similar to the single variant PVE approximation described above, except that the set of variants is tested jointly, accounting for covariance among the genotypes. This approach avoids inadvertently double counting any partially shared signal among the set of identified variants.

Joint tests for association and testing for heterogeneity across population groups. We then performed joint association analyses for the full multi-ethnic sample (n=109,122), as well as each of the four population groups with n>5000, to determine effect sizes and p-values when all 59 variants were considered together. Using the inverse-normalized and rescaled residuals from the primary analysis Stage 1 LMM as the outcome, we fit a new Stage 2 LMM that was the same as described above, except with the additional inclusion of the genotypes for these 59 variants as additive genetic fixed effects. Given this joint modeling framework, the variant effect size estimates are all adjusted for one another. These estimates were used as input for calculation of a polygenic trait score used for the PheWAS described below. Finally, we tested for heterogeneity of effect sizes from these analyses among the population groups by adapting Cochran's Q statistic and its p-value [26], commonly used to test for effect heterogeneity in meta-analysis (***Table 1***). For each variant, the effect size estimates and standard errors from each population group analysis were used to calculate Q, and a Bonferroni threshold of 0.001 (0.05/59) was used to assess significance.

Assessing novelty of identified loci and variants. For each of the 59 variants identified, we examined the linkage disequilibrium (LD) with previously reported sentinel variants from 17 published GWAS.

Only sentinel variants with $p < 5x10^{-8}$ in their published study were considered, which included a total of 56 variants (***Table S3***). If one of our variants had LD≥0.7 with a published variant, it was labeled as a known variant/part of a known locus; otherwise it was labeled as novel in ***Table 1***. Within a locus, we then compared each independent variant to the prior GWAS reported sentinel variant. If they were identical, the variant was labeled as a known sentinel variant in ***Table 1***. Additionally, locus names for the final set of independent variants were selected based on (i) prior GWAS study definition for known loci, and (ii) the specific gene annotation for each variant mapping directly to a gene for novel loci.

Replication of novel results with published GWAS: To determine whether our novel results are supported by findings from prior studies, we considered the two largest most recent studies of telomere genetics in European [27] (Li et al., n=78,592) and Asian [28] (Dorajoo et al., n=26,875) ancestry individuals. These studies both used telomere length as measured by qPCR. For all novel variants in ***Table 1***, we pulled the effect size estimates, standard errors, and p-values, where available (***Figure 3a***). These results were available in at least one of the two studies for 19 of our 22 novel variants, so we considered a p-value cutoff of 0.05/19 = 0.0026 to be replicated, after multiple testing correction. We also labeled variants where at least one study reported p<0.05 as suggestive.

Gene-based coding variant tests - Variant annotation: For its use in gene-based tests for association, annotation based variant filtering and GENCODE v28 gene model-based [29] aggregation was performed using the TOPMed freeze 8 WGSA Google BigQuery-based variant annotation database on the BioData Catalyst powered by Seven Bridges platform (http://doi.org/10.5281/zenodo.3822858). The annotation database was built using variant annotations for TOPMed freeze 8 variants gathered by Whole Genome Sequence Annotator (WGSA) version v0.8[30] and formatted by WGSAParsr version 6.3.8 (https://github.com/UW-GAC/wgsaparsr). Variants were annotated as exonic, splicing, transcript ablation/amplification, ncRNA, UTR5, UTR3, intronic, upstream, downstream, or intergenic using Ensembl Variant effect predictor (VEP) [31]. Exonic variants were further annotated as frameshift insertion,

frameshift deletion, frameshift block substitution, stop-gain, stop-loss, start-loss, non-frameshift insertion, non-frameshift deletion, non-frameshift block substitution, nonsynonymous variant, synonymous variant, or unknown. Additional scores used included REVEL [32], MCAP [33] or CADD [34] effect prediction algorithms.

Gene-based coding variant tests - Tests for association: Gene-based association testing was performed on the pooled trans-ethnic dataset (n=109,122). To improve the power to identify rare variant associations in coding regions, we aggregated deleterious rare coding variants in all protein-coding genes and then tested for association with telomere length. To enrich for likely functional variants, only variants with a "deleterious" consequence for its corresponding gene or genes [35], were included. For each protein-coding gene, a set of rare coding variants (MAF < 0.01, including singletons where MAC=1, restricted to variants which passed IRC quality filters [2] and which had <10% of samples with read depth <10) was constructed, which was composed of all stop-gain, stop-loss, start-loss, transcript ablation, transcript amplification, splice acceptor variants, splice donor variants and frameshift variants, as well as the exonic missense variants that fulfilled one of these criteria: 1) REVEL score > 0.5, 2) predicted M_CAP value was "Damaging", or 3) CADD PHRED-scaled score > 30. We applied the variant Set Mixed Model Association Test (SMMAT) [36] as implemented in GENESIS, using the genesis_tests app on the Analysis Commons [37], with MAF based variant weights given by a beta-distribution with parameters of 1 and 25, as proposed by Wu et al. [38], and using the same two-stage LMM output as used in the primary single variant analysis. Only genes with a cumulative MAC ≥ 5 over all variants were evaluated, leaving a total of 27,558 genes, and significance was evaluated after a Bonferroni correction for multiple testing ($p < 0.05 / 27{,}558 = 1.815 \times 10^{-6}$) (***Figure S2***).

Next, we sought to determine the influence of each rare deleterious variant in each significant gene on the association signal. We iterated through the variants, removing one variant at a time (leave-one-out approach) [39], and repeated the SMMAT analysis. If a variant made a large contribution to the original

association signal, one would expect the signal to be significantly weakened with the removal of the variant from the set ***(Figure S2***).

Finally, we further tested for independence of the gene-based and single variant signals by performing a conditional SMMAT analysis that included the 59 genome-wide significant variants from our primary analysis as fixed-effect covariates in the two-stage LMM. These 59 variants were also removed from the aggregated set of rare variants for a gene if they had been previously included (e.g. rs202187871 in *POT1*). All other analysis parameters were the same as described above (***Figure S2***).

Colocalization analysis of *OBFC1* signals using GTEx [40] and eQTLGen[41]: Iterative conditional analysis was repeated for chromosome 10 focusing on a 2Mb window centered on the primary signal near *OBFC1* (rs10883948). The original pooled GWAS results (n = 109,122) were used for colocalization analysis with the primary signal while the appropriate round of conditional analysis was used for each subsequent signal (e.g., the output of the second round of conditional analysis was used for colocalization analysis with the tertiary signal). Credible set analysis was performed using CAVIAR on primary signal data and the output of each conditional analysis each with a single assumed causal variant [42]. For each independent *OBFC1* signal, the credible set contained the top sentinel variant (***Figure S4a-d***). Colocalization analysis was performed using coloc, a Bayesian posterior probability method that estimates the probability of shared signal across testing modalities at each variant [43]. We report the posterior probability that the two signals are independent (PPH3) and the posterior probability that the two signals overlap (PPH4). The sentinel variants from each signal were assayed as expression quantitative trait loci (eQTLs) in both GTEx[40] and eQTLGen [41] datasets. For each sentinel, significant gene-tissue pairs for that sentinel were identified from GTEx v8 (FDR < 0.05) and assayed for colocalization comparing the beta and standard error of the beta from our GWAS results and the eQTL results. For colocalization analysis in the eQTLGen dataset, all eGenes within a 2Mb window of the sentinel were identified and assayed for colocalization comparing the MAF, p-value, and number of

284 observations. MAF was estimated for eQTLGen data using the TOPMed MAF. Colocalization analysis
285 was not possible for the *OBFC1* secondary signal as that variant is absent in both datasets and a
286 representative proxy variant was not available. Roadmap [44] data was accessed July, 2020 using the hg19
287 (February, 2009 release) UCSC genome browser[45] track data hubs [46,47].
288
289 **Phenome-wide association tests (PheWAS):** Using individual level data within the Vanderbilt
290 University biobank BioVU, PheWAS [48] (tests for association between genotype and phenotype) were
291 performed using the 49 (of 59) sentinel variants available in the multi-ethnic genotyping array (MEGA)
292 chip results imputed to the Haplotype Reference Consortium [49]. Single variant tests using SNP dosage
293 values were performed for all available phecodes (number of cases at least 20), including the covariates
294 age, sex, genotype batch and the first ten ancestry principal components. Analysis was performed
295 separately in BioVU self-identified African Americans (AA, n=15,174) and BioVU self-identified
296 European Americans (EA, n=70,439)**.** In addition, European and African specific effect sizes from the
297 joint analysis from ***Table 1*** were combined to create separate polygenic trait scores (PTS) for each
298 population group which were then tested for association with available phecodes, again including the
299 covariates age, sex, genotype batch and the first ten ancestry principal components. Results were
300 evaluated at a Bonferroni threshold corrected for the number of informative phecodes for each variant
301 (range n=1,114-1,361) or the PTS (n=1,704) (***Table S6***). Analysis was performed using the PheWAS R
302 package [50].
303
304 We queried United Kingdom Biobank (UKBB) GWAS results using the University of Michigan PheWeb
305 web interface (http://pheweb.sph.umich.edu/SAIGE-UKB/). The UKBB PheWeb interface contains
306 results from a SAIGE [51] genetic analysis of 1,403 ICD-based traits of 408,961 UKBB participants of
307 European ancestry. PheWeb is a publicly accessible database that allows querying genome-wide
308 association results for 28 million imputed genetic variants. 47 out of our 59 sentinel variants were present

in PheWeb. We report all hits passing a Bonferroni correction for the number of tests performed for each variant ($0.05/1403 = 3.6 \times 10^{-5}$, ***Table S7***).

References:

1 NHLBI Trans-Omics for Precision Medicine. TOPMed Projects and their Parent Studies. Available at: https://www.nhlbiwgs.org/group/project-studies.

2 Taliun, D. e. a. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* https://doi.org/10.1101/563866 (2019).

3 Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25**, 918-925, doi:10.1101/gr.176552.114 (2015).

4 Lee, M. *et al.* Comparative analysis of whole genome sequencing-based telomere length measurement techniques. *Methods* **114**, 4-15, doi:10.1016/j.ymeth.2016.08.008 (2017).

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

6 Nersisyan, L. & Arakelyan, A. Computel: computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS One* **10**, e0125201, doi:10.1371/journal.pone.0125201 (2015).

7 Kimura, M. *et al.* Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat Protoc* **5**, 1596-1607, doi:10.1038/nprot.2010.124 (2010).

8 Alder, J. K. *et al.* Diagnostic utility of telomere length testing in a hospital-based setting. *Proc Natl Acad Sci U S A* **115**, E2358-E2365, doi:10.1073/pnas.1720427115 (2018).

9 Aviv, A. *et al.* Impartial comparative analysis of measurement of leukocyte telomere length/DNA content by Southern blots and qPCR. *Nucleic Acids Res* **39**, e134, doi:10.1093/nar/gkr634 (2011).

10 O'Callaghan, N. J. & Fenech, M. A quantitative PCR method for measuring absolute telomere length. *Biol Proced Online* **13**, 3, doi:10.1186/1480-9222-13-3 (2011).

11 Mwasongwe, S. *et al.* Leukocyte telomere length and cardiovascular disease in African Americans: The Jackson Heart Study. *Atherosclerosis* **266**, 41-47, doi:10.1016/j.atherosclerosis.2017.09.016 (2017).

12 Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-1735, doi:10.1371/journal.pgen.0030161 (2007).

13 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**, 500-507, doi:10.1038/nprot.2011.457 (2012).

14 Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867-868, doi:10.1093/bioinformatics/btx699 (2018).

15 Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7**, e30377, doi:10.1371/journal.pone.0030377 (2012).

16 <http://cf.10xgenomics.com/supp/genome/GRCh38/sv_blacklist.bed>

17 Halko, N., Martinsson, P. G. & Tropp, J. A. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Rev* **2**, 217–288.

18 Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *Am J Hum Genet* **105**, 763-772, doi:10.1016/j.ajhg.2019.08.012 (2019).

19 Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346-5348, doi:10.1093/bioinformatics/btz567 (2019).

20 Sofer, T. *et al.* A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet Epidemiol* **43**, 263-275, doi:10.1002/gepi.22188 (2019).

21 Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**, 276-293, doi:10.1002/gepi.21896 (2015).

22 Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**, 127-148, doi:10.1016/j.ajhg.2015.11.022 (2016).

23 Tang, Z. Z. & Lin, D. Y. Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs. *Am J Hum Genet* **97**, 35-53, doi:10.1016/j.ajhg.2015.05.001 (2015).

24 Zhou, B., Shi, J. & Whittemore, A. S. Optimal methods for meta-analysis of genome-wide association studies. *Genet Epidemiol* **35**, 581-591, doi:10.1002/gepi.20603 (2011).

25 *The NHLBI Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Program. BRAVO variant browser: University of Michigan and NHLBI; 2018.*, <https://bravo.sph.umich.edu/freeze5/hg38/>

26 Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics, 10(1), 101-129.* **10**, 101-129 (1954).

27 Li, C. *et al.* Genome-wide Association Analysis in Humans Links Nucleotide Metabolism to Leukocyte Telomere Length. *Am J Hum Genet* **106**, 389-404, doi:10.1016/j.ajhg.2020.02.006 (2020).

28 Dorajoo, R. *et al.* Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies. *Nat Commun* **10**, 2491, doi:10.1038/s41467-019-10443-2 (2019).

29 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773, doi:10.1093/nar/gky955 (2019).

30 Liu, X. *et al.* WGSA: an annotation pipeline for human genome sequencing studies. *J Med Genet* **53**, 111-112, doi:10.1136/jmedgenet-2015-103423 (2016).

31 Ahn, D. H. *et al.* Whole-exome tumor sequencing study in biliary cancer patients with a response to MEK inhibitors. *Oncotarget* **7**, 5306-5312, doi:10.18632/oncotarget.6632 (2016).

32 Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885, doi:10.1016/j.ajhg.2016.08.016 (2016).

33 Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581-1586, doi:10.1038/ng.3703 (2016).

34 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 (2014).

35 Graham, G. Disparities in cardiovascular disease risk in the United States. *Curr Cardiol Rev* **11**, 238-245 (2015).

36 Chen, H. *et al.* Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. *Am J Hum Genet* **104**, 260-274, doi:10.1016/j.ajhg.2018.12.012 (2019).

37 Brody, J. A. *et al.* Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* **49**, 1560-1563, doi:10.1038/ng.3968 (2017).

38 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93, doi:10.1016/j.ajhg.2011.05.029 (2011).

39 Keramati, A. R. *et al.* Targeted deep sequencing of the PEAR1 locus for platelet aggregation in European and African American families. *Platelets* **30**, 380-386, doi:10.1080/09537104.2018.1447659 (2019).

40 Aguet, F. e. a. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *BioRxiv*, doi:https://doi.org/10.1101/787903 (2019).

41 Vosa, U. e. a. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* (2018).

42 Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508, doi:10.1534/genetics.114.167908 (2014).

43 Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383, doi:10.1371/journal.pgen.1004383 (2014).

44 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

45 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006, doi:10.1101/gr.229102 (2002).

46 Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU Epigenome Browser update 2019. *Nucleic Acids Res* **47**, W158-W165, doi:10.1093/nar/gkz348 (2019).

47 Raney, B. J. *et al.* Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003-1005, doi:10.1093/bioinformatics/btt637 (2014).

48 Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205-1210, doi:10.1093/bioinformatics/btq126 (2010).

49 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

50 Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375-2376, doi:10.1093/bioinformatics/btu197 (2014).

51 Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* **101**, 37-49, doi:10.1016/j.ajhg.2017.05.014 (2017).