Novel genetic determinants of telomere length from a multi-ethnic analysis of 75,000 whole genome sequences in TOPMed

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

1 Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD USA;
2 Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI USA;
3 Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD USA;
4 GeneSTAR Research Program, Department of Medicine, Johns Hopkins School of Medicine, Baltimore MD USA;
5 Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA;
6 Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA;
7 Department of Cardiology, Johns Hopkins School of Medicine, Baltimore MD USA;
8 Department of Medicine, Division of Cardiology, Johns Hopkins School of Medicine, Baltimore MD USA;
9 Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI USA;
10 Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, Minneapolis, MN, USA;
11 Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX USA;

12 Department of Medicine, Division of Genetics, Genomics and Precision Medicine University of Arizona, Tucson, AZ, USA;
13 Division of Pharmacogenomics, University of Arizona, Tucson, AZ, USA;
14 Department of Medicine, University of California San Francisco, San Francisco, USA;
15 Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, USA;
16 Division of Pediatric Pulmonary Medicine, UPMC Children's Hospital of Pittsburgh, PIttsburgh, PA, USA;
17 Department of Medicine, University of Maryland School of Medicine, Baltimore, MD;
18 Vitalant Research Institute, San Francisco, CA, USA;
19 Department of Laboratory Medicine, UCSF, San Francisco, CA, USA;
20 Division of Cardiology, University of Illinois at Chicago, Chicago, IL, USA;
21 Cardiovascular Division, Department of Medicine, Washington University School of Medicine in St. Louis, St. Louis, MO, USA;
22 Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA;
23 Harvard Medical School, Boston, MA, USA;
24 Department of Internal Medicine, Section on Nephrology, Wake Forest School of Medicine; Winston-Salem, NC, USA;
25 Department of Medicine, Duke University Medical Center, Durham, NC, USA;
26 Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC, USA;
27 Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA;
28 Cardiovascular Health Research Unit and Department of Epidemiology, University of Washington, Seattle, WA, USA;
29 Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA;
30 University of Alabama at Birmingham, Birmingham, AL, USA;
31 Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA;
32 Dept. of Epidemiology, UAB, Birmingham, AL, USA;
33 Department of Biostatistics, Collaborative Health Studies Coordinating Center, University of Washisngton, Seattle, WA, USA;
34 Department of Medicine I, University Hospital Munich, Ludwig-Maximilian's University, Munich, Germany;
35 German Centre for Cardiovascular Research (DZHK); partner site: Munich Heart Alliance, Munich, Germany;
36 Laboratory of Epidemiology and Population Science, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA;
37 Department of Medicine, Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA;
38 Department of Epidemiology, Brown University, Providence, RI, USA;
39 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
40 Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA;
41 Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA;
42 Department of Medicine, University of Colorado Denver, Anschutz Medical Campus, Aurora, CO, USA;
43 Department of Genetics, University of North Carolina, Chapel Hill, NC, USA;
44 Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA;
45 Department of Biostatistics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA;
46 Genome Sciences, University of Washington, Seattle, WA, USA;

47 Center for Individualized and Genomic Medicine Research (CIGMA), Department of Internal Medicine, Henry Ford Health System, Detroit, MI, USA;
48 Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, USA;
49 Department of Oncology, Johns Hopkins School of Medicine, Baltimore, MD, USA;
50 Zilber School of Public Health, University of Wisconsin Milwaukee, Milwaukee WI, USA;
51 Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA;
52 Division of Sleep Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA;
53 The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA;
54 Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA;
55 The Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA, USA;
56 Department of Human Genetics and South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley School of Medicine, Brownsville, TX, USA;
57 Department of Environmental Health, University of Cincinnati, Cincinnati, OH, USA;
58 Department of Preventive Medicine, Northwestern University, Chicago, IL, USA;
59 Population Sciences Branch, Division of Intramural Research, National Heart Lung and Blood Institute, National Institute of Health, Bethesda, MD, USA;
60 Bloodworks Northwest, Research Institute, Seattle, WA, USA and University of Washington, Department of Medicine, Seattle, WA, USA;
61 University of Washington, Department of Medicine, Seattle, WA, USA;
62 Center for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
63 Department of Public Health Sciences, Henry Ford Health System, Detroit, MI, USA;
64 The National Heart, Lung, and Blood Institute, Boston University's Framingham Heart Study, Framingham, MA, USA;
65 The Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, Bethesda, MD, USA;
66 Instituto de Medicina Tropical da Faculdade de Medicina da Uiversidade de Sao Paulo, Sao Paulo, Brazil;
67 Dept. of Biostatistics, UAB, Birmingham, AL, USA;
68 Departments of Pathology & Laboratory Medicine and Biochemistry, Larrner College of Medicine, University of Vermont, Colchester, VT, USA;
69 Department of Medicine, University of California San Francisco, San Francisco, CA, USA;
70 Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston MA, USA;
71 Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA;
72 The Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, MA;
73 Departments of Medicine, Pharmacology, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA;
74 Division of Endocrinology and Metabolism, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan;
75 The Institute for Translational Genomics and Population Sciences, Departments of Pediatrics and Medicine, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA;
76 Division of Sleep Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA;
77 Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA;
78 Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology and Health Services, University of Washington, Seattle, WA, USA;

79 The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
80 Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA;
81 Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA;
82 Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD;
83 Department of Epidemiology & International Health Institute, Brown University School of Public Health, Providence, USA;
84 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;
85 Division of Public Health Sciences, Fred Htutchinson Cancer Research Center, Seattle WA, USA;
86 Bloodworks Northwest, Research Institute, Seattle, WA, USA;
87 Vitalant Research Institute, San Francisco, CA;
88 UCSF Benioff Children's Hospital, Oakland, CA, USA;
89 Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA;
90 Department of Medicine, Tulane University School of Medicine, New Orleans, LA, USA;
91 Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA;
92 Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA;
93 Cardiology Division, Department of Medicine, Massachusetts General Hospital, Boston, MA;
94 Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA;
95 Jackson Heart Study and Departments of Medicine and Population Health Science, Jackson, MS, USA;
96 Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX;
97 College of Public Health, University of Kentucky, Lexington, KY, USA;
98 Division of Cardiovascular, Brigham and Women's Hospital, Boston, MA;
99 Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA;
100 Center of Human Development and Aging, Rutgers, New Jersey Medical School, Newark, NJ, USA;
101 Department of Genome Sciences, University of Washington, Seattle, WA, USA;
102 Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MI, USA;
103 Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA;
104 Department of Biomedical Engineering, Whiting School of Engineering, Baltimore, MD, USA;
105 Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA;
106 Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, 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 Department of Epidemiology, University of Washington, Seattle, WA, USA.

** 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.

Abstract

Telomeres shorten in replicating somatic cells and with age; in human leukocytes, telomere length (TL) is associated with a host of aging-related diseases[1,2]. To date, 16 genome-wide association studies (GWAS) have identified twenty-three loci associated with leukocyte TL[3-18], but prior studies were primarily in individuals of European and Asian ancestry and relied on laboratory assays including Southern Blot and qPCR to quantify TL. Here, we estimated TL bioinformatically, leveraging whole genome sequencing (WGS) of whole blood from n=75,176 subjects in the Trans-Omics for Precision Medicine (TOPMed) Program. We performed the largest multi-ethnic and only WGS-based genome-wide association analysis of TL to date. We identified 22 associated loci (p-value $<5x10^{-8}$), including 10 novel loci. Three of the novel loci map to genes involved in telomere maintenance and/or DNA damage repair: *TERF2, RFWD3,* and *SAMHD1*. Many of the 99 pathways identified in gene set enrichment analysis for the 22 loci (multiple-testing corrected false discovery rate (FDR) <0.05) pertain to telomere biology, including the top five (FDR$<1x10^{-9}$). Importantly, several loci, including the recently identified *TINF2* and *ATM*[6] loci, showed strong ancestry-specific associations.

Results

High throughput sequencing with decreasing sequencing cost per sample has enabled the generation of WGS data at an unprecedented scale, and the National Heart, Lung and Blood Institute's TOPMed Program offers the opportunity to address both sample size and population diversity limitations of prior TL GWAS. To optimize the computational task of estimating TL on the full set of TOPMed WGS samples, we compared two estimation methods, TelSeq [19] and Computel [20], on a subset of samples for which we had prior laboratory-based telomere length measurements from Southern blot. TelSeq and Computel estimates were highly correlated (Pearson correlation r=0.98, ***Supplementary Figure S1A***) and had similar correlation with Southern blot data (Pearson correlation r=0.57 and 0.55 for TelSeq and Computel, respectively, ***Supplementary Figure S1B***); this is similar to what has been reported previously [21]. We selected TelSeq due to its computational efficiency (***Supplementary Figure S1C***). Given the

sample heterogeneity and complexity of generating WGS across the large number of cohorts in the TOPMed program [22] (*Nature, submitted, 2019*), not unexpectedly, we observed cross-study and cross-sequencing center effects (***Supplementary Figures S2A*** and ***S2B***), and we chose a statistical approach to minimize them (see ***Materials and Methods*** and ***Supplementary Figures S2C and S2D***). The final sample set analyzed included 38,193 European ancestry, 21,179 African ancestry, 9,808 Hispanic/Latino, 4,754 Asian ancestry, and 1,242 Samoan individuals. 42% of participants were male and age ranged from <1 to 98, median 55 years (***Supplementary Tables S1A and S1B***).

Genome-wide tests for association across 93M variants (genotype calling pipeline, sample selection and sequencing details described under ***Materials and Methods***) were performed in multiple stages, reflecting different WGS freezes and the final analysis included a discovery set (n=46,458), a replication set (n=28,718), and a meta-analysis of both sets (n=75,176). We identified 22 loci reaching a meta-analysis p-value $<5x10^{-8}$ (***Figures 1 and 2, Table 1***), of which 12 loci met the threshold of $5x10^{-9}$ recently suggested for WGS-based GWAS analyses [23]. Of the 23 prior loci discovered through GWAS of TL assessed with laboratory assays, we confirmed twelve (*TERC, TERT, NAF1, RTEL1, OBFC1, DCAF4, ZNF676, ACYP2,* and the recently identified *TERF1, TINF2, POT1* and *ATM* loci) at a significance threshold of p-value $<5x10^{-8}$ ***(Table 1, Supplementary Tables S2A and S2B***). Nominal evidence (p-value < 0.05/23) was noted for an additional 5 prior known loci at the specific reported variants (*PARP1, NKX2-3, MPHOSPH6, TYMS,* and *ZNF208*; see ***Supplementary Table S3)***.

Among the 22 loci reaching traditional GWAS thresholds in the multi-ethnic TOPMed samples, we also identified 10 novel loci (***Table 1, Supplementary Tables S2A and S2B*****),** three of which include genes encoding proteins that have plausible roles in telomere biology (the index gene definition for each locus is described in ***Table 1***). RFWD3[24] plays a key role in DNA damage repair; TERF2 is a component of the telomere shelterin complex; and depletion of SAMHD1, which has reported roles in DNA resection and homology-directed repair, has been shown to lead to telomere breakage events in cells deprived of the

shelterin component TERF1[25], a recently reported GWAS locus that we also identify as a Tier 1 locus. Gene set enrichment analysis [26,27] including the index gene(s) for each of the 22 loci resulted in 99 sets with an FDR < 0.05 (***Supplementary Table S4***). The top 5 gene sets, all with an FDR $<1x10^{-9}$, were: regulation of telomere maintenance via telomere lengthening (GO:1904356), regulation of telomere maintenance (GO:0032204), negative regulation of telomere maintenance (GO:0032205), telomere maintenance (GO:0000723), and telomere organization (GO:0032200).

Each peak variant at a locus, henceforth referred to as the *sentinel* variant for that locus, accounts for a small proportion of phenotypic variation (***Table 1***), consistent with prior GWAS of telomere length. Prior GWAS SNPs cumulatively account for 2% - 3% of trait variance [28], with allelic effects ranging from ~ 49-117 base pairs. In the TOPMed data, effect sizes for common variants (minor allele frequency, MAF ≥5%) range from 22-71 bp/allele. Rare and low frequency variants (MAF <5%) show larger effects (152-631 bp/allele). Cumulatively, the 22 sentinel variants from the TOPMed WGS-based GWAS account for ~1.5% of phenotypic variance. Individually, *TERC, TERT* and *OBFC1* each account for the largest phenotypic variance (~0.2%) and have similar effect sizes (~60-70bp/allele).

In an attempt to look beyond the single variant approaches, gene-based tests identified five protein coding genes with deleterious rare and low frequency (MAF <5%, including singletons) coding variants associated with telomere length in the discovery samples (see ***Materials and Methods*** and ***Supplementary Figure S3A, Supplementary Table S5A***): *RTEL1, RTEL1-TNFRSF6B, ATM, KDELC2,* and *NAF1*. For each of these genes, a leave-one-out approach iterating over each deleterious variant identified one to three driver variants accounting for the association signal at the gene (***Supplementary Figures S3B-S3F)***. Testing for evidence at these specific driver variants in the independent replication sample provides confirmation for *RTEL1, RTEL1-TNFRSF6B* and *NAF1* (***Supplementary Table S5B***), with the same shared variants for *RTEL1* and *RTEL1-TNFRSF6B*. All three genes are noted to be index genes from the GWAS loci identified (***Table 1***). Linkage disequilibrium with the peak sentinel variant

supports overlap between the gene-based and single variant signals. Importantly, while the *KDELC2* driver variant was not replicated (while showing a consistent direction of effect between the discovery and replication), the minor allele (rs74911261/A) of the driver variant from the discovery sample has previously been shown to be associated with decreased risk of breast cancer [29] and increased risk of renal cell carcinoma [30] in European ancestry individuals. Notably, the A allele is associated with lower telomere length (-80.4 bp/allele) and supports the prior observation that shorter telomere length is strongly associated with increased risk for renal cancers [31].

An evaluation of the 22 loci by race/ethnicity demonstrates that many of these loci are associated with TL in multiple groups. As illustrated in ***Figure 3*** (see also ***Supplementary Figures S4A – S4V, Supplementary Table S2B***) the previously reported *TERC, TERT, RTEL1, TERF1, TINF2* and *OBFC1* loci have p-values $<10^{-5}$ among non-European populations. Among the novel loci identified, *RFWD3* and *TERF2* and have p-values $< 10^{-5}$ in non-European groups. Not surprisingly, most of the 22 loci had strong evidence of association in the European ancestry sample, which also had the largest sample size. In fact, there were several loci (*ATM,CHKB-AS1/MAPK8IP2, LINC01592/LOC100505739, OPRK1/ATP6V1H, RPN1* and *YY1P2,LRP1B*) where association was limited to the European ancestry sample (p-values$<10^{-5}$); no variant mapping to these loci reached a p-value < 0.0023 (0.05/22 for the total of 22 loci evaluated) in any other population. One notable exception was the *TINF2* locus, where the sentinel variant is highly differentiated between ancestral populations. The *TINF2* association was not observed in the European population, where the allele frequency for the alternate allele is extremely low (AAF=0.05%, p-value=0.04), as compared to the Asian (AAF=9%, p-value=1.3×10^{-5}), African (AAF=1%, p-value=2.6×10^{-5}) and Samoan (AAF=23%, p-value=1.3×10^{-7}) samples (***Supplementary Table S2B***).

Leukocyte telomere length (LTL) is associated with mortality and aging-related diseases such as cancer[32], and genetic variants associated with LTL previously have been associated with risk of cancers as well as other non-neoplastic disease of aging[33]. We analyzed 1403 international classification of disease (ICD)-

based phenotypes in ~402,000 Europeans from the UK Biobank (***Supplementary Table S6, Supplementary Figures S5A-S5C***), and we noted that the sentinel variants at *TERT* and *TERC* each had multiple phenome-wide disease associations (PheWAS), including myeloproliferative neoplasms, cancers of skin and brain, and leiomyoma/benign neoplasms of the uterus (all p-values<1.8×10^{-6}). The associations with uterine leiomyomata are consistent with recently published GWAS which found that several telomere length-associated genes and variants (*TERT, TERC, OBFC1, ATM*) have genome-wide significant associations with uterine fibroids [34]. Notably, several of our TOPMed sentinel variants (*NAF1, TERF1, ZNF729, POT1, CHKB-AS1*) had uterine fibroid p-values in the range of 0.008 to 0.07 in our UK Biobank PheWAS analysis. Additionally, several of the sentinel telomere length variants or their proxies (*TERT, TERC, RFWD3, TCL1A, RPN1)* were associated with quantitative hematologic traits or myeloproliferative disorders and malignancies either in the UK Biobank or in recently published GWAS [30,35,36]. As a follow up to assess functional relevance, we used a set of 31,684 blood samples from eQTLGen [37] and found that 17 out of the 18 of our sentinel variants present in the data set were eQTLs for at least one local eGene. For many of these, the top eGene is the index gene we identified at the locus (***Supplementary Table S7***), but we recognize the limitation in the use of whole blood from adult samples as the sole tissue interrogated.

Leveraging WGS available through NHLBI's TOPMed program, we have illustrated the feasibility of generating high quality TL from WGS data. We were able to take advantage of the well-powered sample size and multi-ethnic nature of the sample to confirm known GWAS loci and identify an additional set of novel loci that map to genes with plausible biological validity. We also explored loci across populations of diverse ancestry. The ability to implement this phenotype assessment of TL in large, multi-ethnic datasets with pre-existing WGS creates opportunities beyond the genetics of TL; it will expand our ability to evaluate of the role of TL and genes determining TL in health and human disease.

Table 1: Sentinel SNPs from the 22 loci identified from the multi-ethnic TOPMed analysis of telomere length. All loci had a peak p-value $<5x10^{-8}$ in the combined meta-analysis (n=46,458 in the discovery and n=28,718 in replication dataset, total n=75,176), and are classified into three Tiers based on their evidence in the discovery dataset: **Tier 1** includes loci that were genome-wide significant in the discovery samples ($p<5x10^{-8}$); **Tier 2** includes loci that were suggestive in the discovery analysis ($1x10^{-5}<p<5x10^{-8}$); and **Tier 3** includes those loci that were nominal in the discovery analysis with $1x10^{-3}<p<1x10^{-5}$). Effect size in base pairs is with respect to the alternate (Alt) allele.

		Known					Discovery (n=46458)		Replication (n=28718)		Meta-Analysis			Effect Size	
Tier	Locus Name	vs Novel	Position (hg38)	rsID	Ref	Alt	AAF	P-Value	AAF	P-Value	AAF	P-Value	Direction	Variance	Base pairs
Tier 1	***TERT****		**Chr5:1285859**	**rs7705526**	**C**	**A**	**31%**	**3.3E-24**	**28%**	**1.3E-18**	**30%**	**4.6E-41**	**++**	**0.20%**	**60.0**
	TERC*		**Chr3:169764547**	**rs2293607**	**T**	**C**	**21%**	**6.7E-19**	**22%**	**1.2E-16**	**22%**	**1.1E-33**	**--**	**0.20%**	**-70.7**
	RTEL1*		**Chr20:63704906**	**rs6062497**	**C**	**T**	**70%**	**1.7E-13**	**71%**	**1.0E-07**	**71%**	**1.0E-19**	**--**	**0.10%**	**-42.0**
	SH3PXD2A,OBFC1(STN1),SLK*		**Chr10:103919583**	**rs2488002**	**C**	**T**	**69%**	**1.8E-19**	**66%**	**2.2E-18**	**68%**	**7.6E-36**	**--**	**0.20%**	**-64.3**
	RFWD3*	**Novel**	**Chr16:74643066**	**rs28616016**	**T**	**C**	**44%**	**4.1E-15**	**43%**	**3.6E-03**	**44%**	**1.6E-15**	**--**	**0.07%**	**-31.4**
	NAF1*		**Chr4:163127047**	**rs4691895**	**G**	**C**	**78%**	**1.8E-09**	**78%**	**1.3E-04**	**78%**	**1.3E-12**	**++**	**0.07%**	**39.5**
	ACYP2*		**Chr2:54268085**	**rs7579722**	**G**	**C**	**17%**	**2.4E-08**	**17%**	**2.0E-04**	**17%**	**2.4E-11**	**++**	**0.05%**	**37.5**
	TERF1*		**Chr8:73038324**	**rs12679652**	**G**	**A**	**58%**	**1.4E-07**	**54%**	**2.1E-03**	**56%**	**1.6E-09**	**--**	**0.05%**	**-28.8**
	LINC01592,LOC100505739	Novel	Chr8:69331466	rs144510686	G	T	0%	5.8E-09	0%	4.5E-02	0%	6.1E-09	--	0.03%	-382.3
Tier 2	***TINF2***		**Chr14:24242592**	**rs28372734**	**C**	**G**	**1%**	**1.1E-07**	**1%**	**4.3E-07**	**1%**	**2.8E-13**	**++**	**0.09%**	**152.1**
	SAMHD1	**Novel**	**Chr20:36950277**	**rs4810362**	**A**	**G**	**23%**	**7.6E-08**	**26%**	**1.1E-03**	**24%**	**4.1E-10**	**--**	**0.06%**	**-34.7**
	TERF2	**Novel**	**Chr16:69357811**	**rs9925619**	**C**	**G**	**31%**	**2.6E-06**	**30%**	**2.9E-04**	**31%**	**2.9E-09**	**++**	**0.04%**	**26.8**
	ZNF676,ZNF729		Chr19:22242195	rs281173	G	A	59%	4.5E-07	57%	7.3E-03	58%	1.9E-08	++	0.03%	22.3
	TCL1A	Novel	Chr14:95714348	rs11846938	T	G	34%	3.8E-06	37%	1.4E-03	35%	2.0E-08	++	0.04%	25.6
	YY1P2,LRP1B	Novel	Chr2:139954363	rs547680822	C	T	0%	3.7E-07	0%	9.7E-03	0%	2.2E-08	++	0.02%	631.4
	OPRK1,ATP6V1H	Novel	Chr8:53522200	rs188891454	T	C	0%	3.7E-06	0%	2.2E-03	0%	3.3E-08	--	0.03%	-234.0
	LINC01429	Novel	Chr20:51837445	rs6091385	C	T	14%	3.1E-06	15%	3.2E-03	15%	4.0E-08	++	0.04%	31.9
	RPN1	Novel	Chr3:128703333	rs60092972	A	T	26%	4.5E-06	23%	2.4E-03	25%	4.2E-08	++	0.03%	23.6
Tier 3	*DCAF4*		Chr14:72965392	rs78517833	A	T	10%	1.2E-05	10%	7.1E-05	10%	3.6E-09	++	0.03%	36.8
	POT1		**Chr7:124854807**	**rs10246424**	**A**	**G**	**21%**	**2.6E-04**	**19%**	**2.0E-05**	**20%**	**3.6E-08**	**--**	**0.03%**	**-28.8**
	ATM		**Chr11:108195635**	**rs4027719**	**CT**	**C**	**50%**	**2.2E-05**	**49%**	**6.0E-04**	**50%**	**4.9E-08**	**--**	**0.04%**	**-25.6**
	CHKB-AS1,MAPK8IP2	Novel	Chr22:50596441	rs131742	A	G	30%	9.5E-05	26%	1.2E-04	28%	5.0E-08	--	0.03%	-26.1

Note: In **bold** are genes documented to play a role in telomere length or DNA damage repair. For the **Tier 1** loci (p-value $<5x10^{-8}$ in discovery sample), replication was evaluated in the TOPMed replication sample, and loci that have significant replication (see ***Supplementary Table S2A***) with SNPs within the locus having $p<0.0027$ (0.05/22) in the replication sample are indicated by *. Loci are labeled as Novel if they have not been a statistically significant locus in a prior GWAS. Index genes for each locus were selected based on (i) prior GWAS study definition for known loci, (ii) the gene to which the variant maps per annotation in ***Supplementary Table S2A*** for novel loci, and (iii) the exception of the *OBFC1* and *ATM* loci: For the *OBFC1* locus three index genes were selected *SH3PXD2A, OBFC1(STN1),* and *SLK* as all three had strong SNV signal not in LD with the sentinel variant (***Supplementary Figure S4D***); and for *ATM*, the sentinel variant mapped to *NPAT*, but was a peak eQTL for ATM (***Supplementary Figure S4U***).

Figure 1: Multiethnic genome-wide tests for association using 93M sequence identified variants on n=75,176 samples with sequence generated telomere length from TOPMed. All loci had a peak $p<5x10^{-8}$ in the combined meta-analysis, and are classified into three Tiers based on their evidence in the discovery dataset (n=46,458 samples): **Tier 1** includes loci that were genome-wide significant in the discovery samples ($p<5x10^{-8}$); **Tier 2** includes loci that were suggestive in the discovery analysis ($1x10^{-5}< p <5x10^{-8}$); and **Tier 3** includes those loci that were nominal in the discovery analysis ($1x10^{-3} < p <1x10^{-5}$).

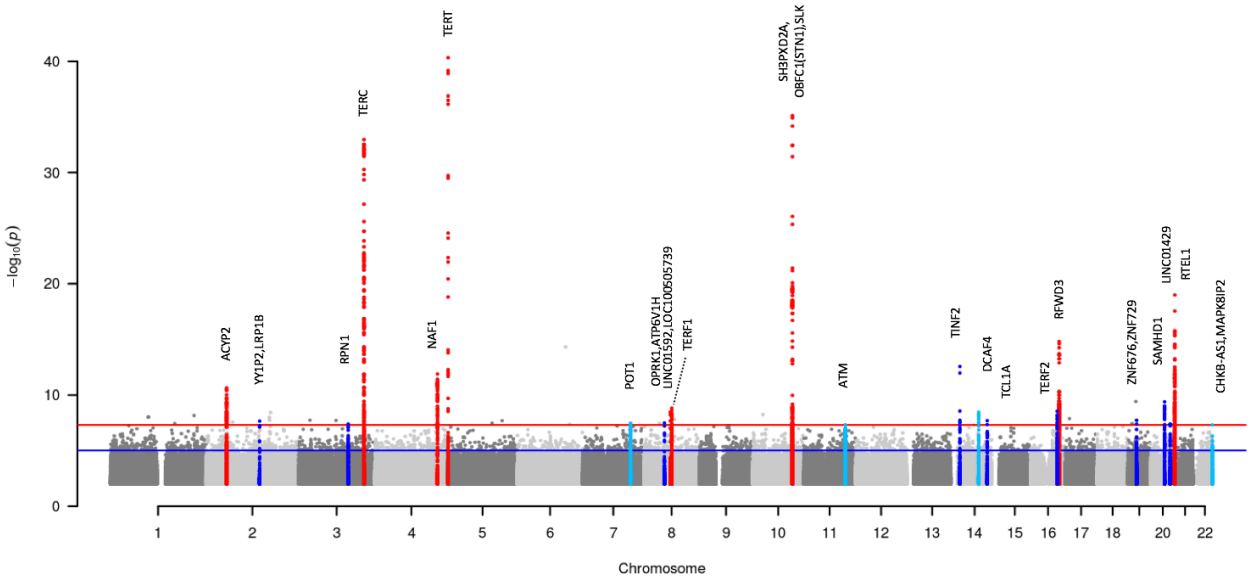


Figure 2: LocusZoom plots of the 22 loci identified from the TOPMed meta-analysis. SNPs with p<0.1 are plotted for all **Tier 1**, **Tier 2** and **Tier 3** loci. Linkage disequilibrium (LD) is with respect to the peak variant in the multi-ethnic analysis, and LD is calculated using the specific set of samples used in the analysis thereby reflecting LD patterns specific to the TOPMed samples. ***Supplementary Figures S4A-S4V*** has corresponding plots by race/ethnicity, with sample specific LD.

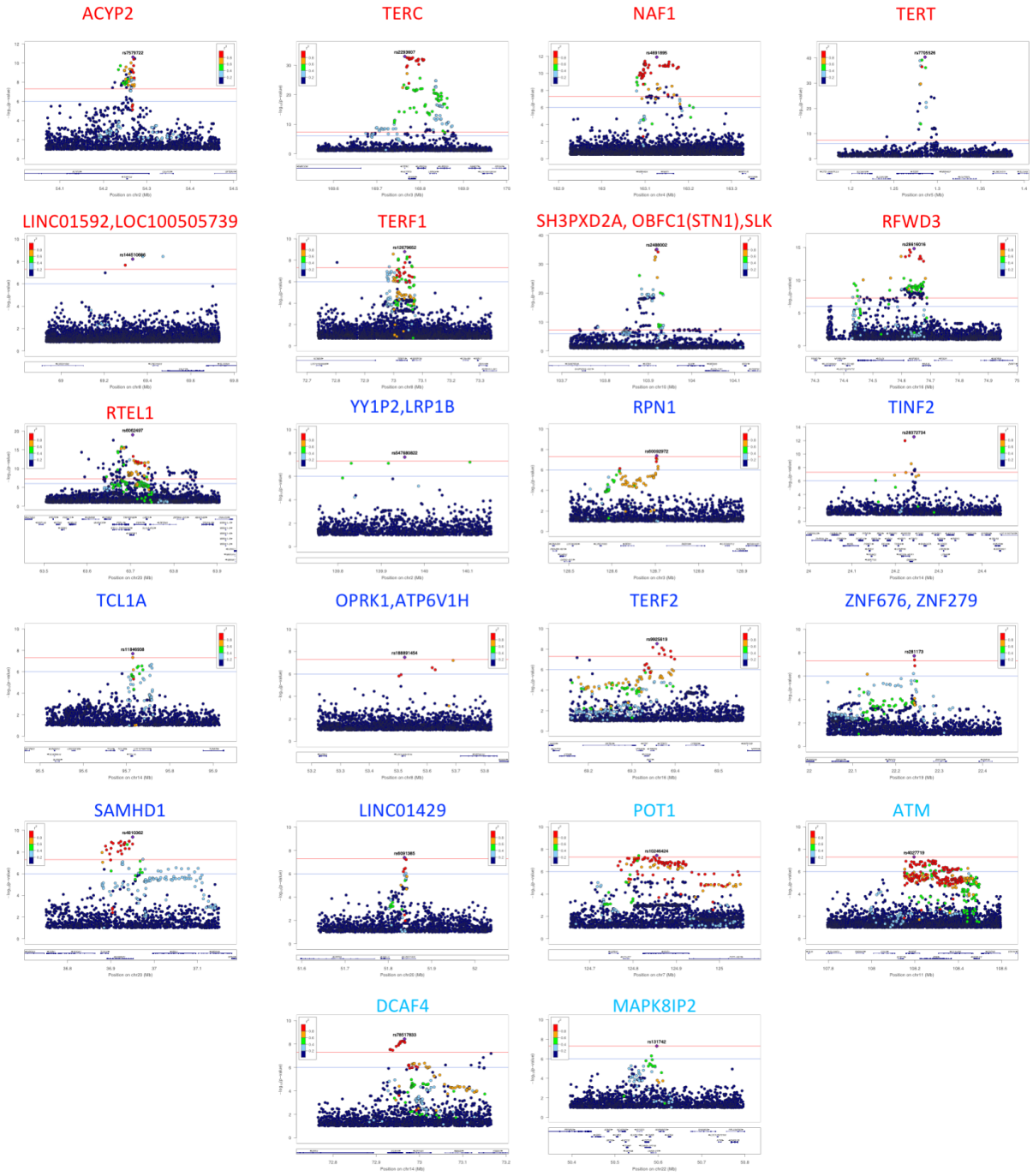
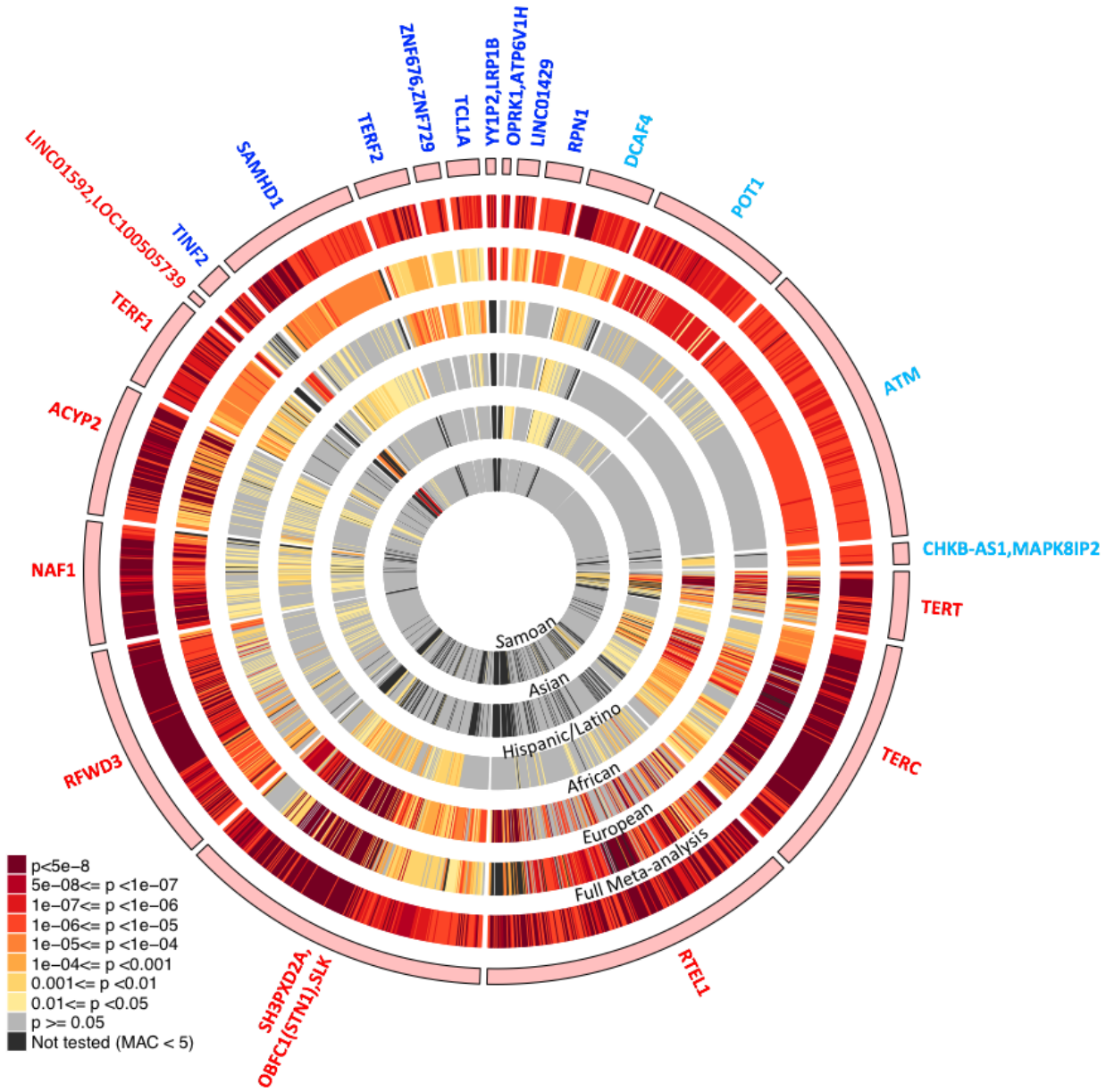


Figure 3: Ancestry specific signal at each variant with a p-value <1x10^{-5} in the meta-analysis mapping to the 22 loci. Locus names are colored by their Tier, i.e., statistical significance in the discovery sample. ***Supplementary Figures S4A-S4V*** show these variants in ancestry-specific LocusZoom plots. The results are based on 38,193 European ancestry, 21,179 African ancestry, 9,808 Hispanic/Latino, 4,754 Asian ancestry, and 1,242 Samoan individuals.



Supplementary Information is linked to the online version of the paper.

Acknowledgements: 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 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., K.R.I., L.R.Y., M.P.C., A.K., M.Arvanitis, Y.C.C., L.M.R., M.Armanios, M.H.C., M.D., D.L., A.B., T.W.B., I.R., J.A.P., A.P.R., R.A.M. drafted the manuscript. M.A.T., J.S.W., M.P.C., J.A.B., A.K., C.C.L., G.A., A.A., D.A.N., J.G.W., S.S.R., D.L., A.B., T.W.B., I.R., T.T., J.O., J.A.P., N.P., A.P.R., R.A.M. contributed substantive analytical guidance. M.A.T., J.S.W., K.R.I., L.R.Y., M.P.C., J.A.B., A.K., C.A.L., M.Arvanitis, A.V.S., J.Lane, A.P.R., R.A.M. performed and led analysis. L.R.Y., 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., C.I., M.R.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., D.E.W., M.M.W., L.W., W.Z., M.Armanios, S.A., P.L.A., D.W.B., B.E.C., I.Y.C., M.H.C., L.A.C., J.E.C., M.D., R.D., X.G., L.H., S.H., J.M.J., E.E.K., A.M.L., C.L., R.L.M., M.N., 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., J.I.R., S.R., B.M.P., J.M.P., N.D.P., R.J.L., C.G.M., B.D.M., D.A.M., S.T.M., A.C.M., R.Kumar, C.K., B.A.K., S.Kelly, S.L.K., R.Kaplan, J.H., H.G., M.F., P.T.E., M.d., A.C., E.B., K.C.B., A.E.A., D.K.A., C.A., A.A., J.G.W., S.S.R., D.L., J.O., A.P.R., R.A.M. 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 GSK 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 Ionis and AstraZeneca Pharmaceuticals.
L.W. is on the advisory board for GSK and receives grant funding from NIAID, NHLBI, and NIDDK, NIH
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 and travel support from GlaxoSmithKline in the past three years.
B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson.
P.T.E. is supported by a grant from Bayer AG to the Broad Institute focused on the genetics and therapeutics of cardiovascular diseases; has served on advisory boards or consulted for Bayer AG, Quest Diagnostics, and Novartis.
K.C.B. receives royalties from UpToDate

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

Online Methods

TOPMed study populations: To perform this multi-ethnic genome-wide association study of telomere length, we leveraged the whole genome sequence samples available through the NHLBI's Trans Omics for Precision Medicine (TOPMed) [22] (Taliun, *Nature, Submitted, 2019*) program. The program currently consists of more than 80 participating studies [38] across a range of study designs as described in Taliun et al [22] (*Nature, submitted, 2019*). These participants are mainly U.S. residents with diverse ancestry and ethnicity (European, African, Hispanic/Latino, Asian, and other). Details on the specific samples included for telomere length analysis are outlined below, summarized in ***Supplementary Tables S1A and S1B***, and described by TOPMed [38].

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. [22] (*Nature, submitted, 2019*). Briefly, variant discovery and genotype calling was performed jointly, across all the available TOPMed Freeze 5b (September 2017) and Freeze 6a studies (August 2018), using the GotCloud [39] 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 [40]. 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 [19], 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 [20] 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 3362 samples from the Jackson Heart Study (JHS) and found them to be highly correlated with one another (Pearson correlation r=0.98, ***Supplementary Figure S1A***). We also compared computational time to generate the telomere length estimates on these samples and show that Computel is an order of magnitude more time-consuming (***Supplementary Figure S1C***). 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.

As a further comparison to orthogonally measured telomere length values, we used data from 2429 samples from JHS with Southern blot[41] telomere length estimates [42]. 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=0.57 and 0.55 for TelSeq and Computel, respectively, ***Supplementary Figure S1B***). We do note some technical sources of variability in our data, both within a study (colors in ***Supplementary Figures S1A*** and ***S1B*** indicate grouping by shared 96-well plate for shipment to sequencing center for these JHS samples) and across studies (***Supplementary Figures S2A and S2B***). Cross-study differences are accounted for in our modeling process (see ***Supplementary Figures S2C and S2D,*** and ***Single variant tests for association***, below).

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 performed on a set of 93,219 samples whose CRAM-files were available for analysis at the TOPMed IRC at the time of analysis.

Samples included in genetic analysis: Samples with telomere length estimated from the WGS data from the TOPMed Studies described above were included in either a *discovery* or *replication* dataset (***Supplementary Table S1A and S1B***) based primarily on their release as part of the TOPMed WGS data processing "Freezes" (Taliun, *Nature, Submitted, 2019*) [22]. The *discovery* dataset (n=46,458, ***Supplementary Table S1A***) is comprised of samples that were included in the TOPMed freeze5b data set (Taliun, *Nature, Submitted, 2019*) [22], released in September 2017, passing sample-level quality control (QC) checks as determined by the TOPMed Data Coordinating Center (DCC) (e.g. concordance of annotated and genetic sex, comparisons of genetically inferred and pedigree reported relatedness, and concordance of WGS genotype calls with prior array data), and with consent groups that allowed for genetic analysis of telomere length. Only samples with sequencing read lengths of 151 or 152 basepairs and having age at blood draw and reported race/ethnicity data available were included. For the set of samples that were part of a duplicate pair (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. Relying on the same set of criteria, samples were included in the *replication* dataset (n=28,718, ***Supplementary Table S1B***) if they were available as additional samples in the freeze6a TOPMed data release available in August 2018.

Race/ethnicity classifications as presented in ***Supplementary Table S1*** were harmonized by the TOPMed DCC across studies based on study-specific self-reported questionnaire data. We included samples belonging to the following five race/ethnicity categories for our analysis: African ancestry, Asian ancestry, European ancestry, Hispanic/Latino and Samoan. For inclusion within the final set of samples described above, the minimum sample size for any study-race-sequencing center stratum had to be n=50. Samples belonging to a smaller stratum were not included in any analyses.

Single variant tests for association: The genome-wide tests for association were performed on the Analysis Commons[43]. Variants with minor allele count (MAC) of at least 5 and passing IRC quality filters were included for single variant analyses. Individual genotype calls with a read depth less than 10 at a particular variant were considered "missing" and were imputed using the sample allele frequency.

A two stage procedure[8] was performed to test for association genome-wide in the *discovery* dataset; the steps were as follows:

1. Telomere length was regressed on age and sex separately within each study-race/ethnicity-sequencing center stratum for the n=46,458 discovery samples. Within each stratum, the regression residuals were then inverse-normal transformed and subsequently scaled by their original variances. This rescaling returns the within-stratum variance back to its original value, allowing for clearer interpretation of estimated genotype effect sizes (see ***Supplementary Figure S6***). These inverse-normalized and scaled residuals were then combined across all strata for the discovery dataset, and tests for association were performed as follows.
2. Given the large sample size of the discovery dataset, a mega-analysis including all n=46,458 samples was performed in two steps:
    a. All genetic loci were tested for association with the inverse-normalized residuals using a standard additive linear model again adjusting for age, sex, and study.

b. All loci with p-values for association between genotype and outcome < 0.01 from this standard additive linear model were then re-analyzed using a linear-mixed model (described below) that included a genetic relationship matrix (GRM) estimated using MMAP [44] to account for ancestry differences as well as within and between study relatedness among individuals, included age, sex, and study as model covariates, and allowed for heterogeneous residual variances across sample groups defined by study.
c. The final reported p-value for association is the value from b, if available, and is otherwise the value from a.

A two stage procedure similar to that used for the discovery dataset described above was performed to test for association genome-wide in the *replication* dataset:

1. Residuals from a linear model of telomere length regressed on age, sex, and 11 principal components (PCs) of ancestry were calculated within each study-race/ethnicity-sequencing center stratum for the n=28,718 replication samples. Within each stratum, the residuals were then inverse-normal transformed, and subsequently scaled by their original variances to return the within-stratum variance back to its original value.
2. A mega-analysis including all n=28,718 samples was performed using a linear-mixed model (described below) that included an empirical kinship matrix to account for all relatedness among individuals, included sex, age, 11 PCs of ancestry, and study as model covariates, and allowed for heterogeneous residual variances across sample groups defined by study.

Implementation of the Linear Mixed Model used for association tests: The tests for association were conducted using linear mixed models as implemented in the GENESIS ["Genetic association testing using the GENESIS R/Bioconductor package", Gogarten et al., Bioinformatics, in press] application on the Analysis Commons. For both the discovery and replication analyses, the genesis_nullmodel app (versions v0.3 for discovery and v1.0.5 for replication) was used to fit the linear mixed model under the

null hypothesis of no genetic association (i.e. without any individual genotype terms in the model), where the transformed residuals from step 1 above were used as the outcome, and the model was specified as described above. The output from the null model analysis was then used to perform single variant score tests of association with the genesis_tests app (versions genesis_dscan_single for discovery, genesis_tests_v.1.3.2 for replication). In the discovery analysis, the GRM used to account for both individual ancestry differences and relatedness was computed using MMAP [44]. In the replication analysis, ancestry-representative PCs generated using PC-AiR [45] were included in the two steps of analysis to adjust for individual ancestry differences, and an empirical kinship matrix generated using PC-Relate [46] was used in step 2 to account for relatedness among individuals. The switch from a GRM to a kinship matrix for the TOPMed wide sample set on the Analysis Commons was done to accommodate the increased sample size in freeze 6a relative to freeze5b.

Meta-analysis: Meta-analysis was performed genome-wide combining the Discovery and Replication association results using the sample size weighted approach implemented in METAL (version 2018-08-28) [47].

Assessing significance and defining genetic loci: All variants with meta-analysis p-value $< 5x10^{-8}$ were considered as significant in the meta-analysis. All variants passing this threshold were examined in BRAVO [48] to assess quality, and a set of 154 variants were filtered out due to variant call quality issues. Using the remaining significant variants, we determined which belonged to a "locus" (and were not just one-off singleton variants) by taking each peak variant and identifying if there were additional variants with a linkage disequilibrium (LD) $r^2 > 0.5$ with this variant (across all samples) that also achieved a level of significance $< 5x10^{-8}$ in the meta-analysis. From each set of variants at a locus, the sentinel variant was determined by selecting the position which was present in both the discovery and replication analysis (i.e., had minor allele count > 5 in both data sets) and which showed the smallest meta-analysis p-value of any variants falling in that locus. Index genes for each locus were selected based on (i) prior GWAS study

definition for known loci, (ii) the specific gene annotation for each variant mapping directly to a gene in ***Supplementary Table S2A*** for novel loci, and (iii) the exception of the *OBFC1* and *ATM* loci: For the *OBFC1* locus three index genes were selected *SH3PXD2A, OBFC1(STN1), SLK* as all three had strong SNV signal not in LD with the sentinel variant (***Supplementary Figure S4D***); and for *ATM*, the sentinel variant mapped to *NPAT*, but was a peak eQTL for ATM (***Supplementary Figure S4U***).

Estimation of ancestry-specific p-values: Single variant tests for association were performed as described above for each of the five race/ethnicity subgroups within the discovery and replication data sets, splitting the samples after the first step (i.e., after calculating, inverse-normal transforming and rescaling residuals). Meta-analysis to combine the discovery and replication results within a race/ethnicity group was also performed as described above.

Estimation of effect sizes and percent of variance explained: To estimate the effect size and percent variance explained for individual variants, we performed the same two stage procedure as described for association testing with the replication dataset, but with two differences: we used the full set of 75,176 samples, and we only computed score test statistics for the 22 associated variants identified through the meta-analysis. Estimates of the additive effect size per copy of the alternate allele for each variant were approximated from the score test statistics using the approach illustrated in Zhou et al. [49] (i.e. $\hat{\beta} = U_\beta / V_\beta$, where U_β is the covariate-adjusted score for testing the variant, and V_β is its variance). Despite using inverse-normalized residuals as the outcome variable, we expect these effect size estimates to be approximately on the original trait scale (i.e. number of basepairs) because the distribution of residuals pre-inverse-normalization was not too far from Normal (***Supplementary Figure S6***), and we re-scaled the variance back to its original value[50]. To estimate the percent of phenotypic variance explained (PVE) by each individual variant, we used the formula $PVE = 1 - RSS_1/RSS_0$, where RSS_0 and RSS_1 are the

residual sums of squares computed from the null model, and the model including the variant of interest, respectively. Following the idea of Zhou et al., we derived a similar approximation for PVE using only estimates from the null model: $\widehat{PVE} = U_\beta^2/(RSS_0 V_\beta)$.

Gene-based coding variant tests - Variant annotation: For their use in the gene-based tests for association, variant annotation was performed using WGSA7[51] and dbNSFP [52]. Variants were annotated as exonic, splicing, ncRNA, UTR5, UTR3, intronic, upstream, downstream, or intergenic. Exonic variants were further annotated as frameshift insertion, frameshift deletion, frameshift block substitution, stopgain, stoploss, nonframeshift insertion, nonframeshift deletion, nonframeshift block substitution, nonsynonymous variant, synonymous variant, or unknown. Additional scores available included REVEL [53], MCAP [54] or CADD [55] effect prediction algorithms.

Gene-based coding variant tests - Tests for association: Gene-based analysis was performed on the discovery samples only (n=46,458). To improve the power of identifying rare variant associations in coding regions, we aggregated deleterious rare coding variants in 19,387 protein-coding genes and then tested for association with telomere length. To enrich for functional variants, only variants with a "deleterious" consequence for its corresponding gene or genes [56], were included. For each protein-coding gene, a set of rare coding variants (MAF < 0.05, including singletons where MAC=1) was constructed, which was composed of all stop-gain, stop-loss, and frameshift variants, as well as the exonic missense variants that fulfilled one of these criteria: 1) REVEL score > 0.5, 2) M_CAP score was "Deleterious", or 3) CADD score > 20. We applied the Sequence Kernel Association Test (SKAT) [57] as implemented in GENESIS, using the genesis_tests app on the Analysis Commons, with minor allele frequency based variant weights given by a beta-distribution with parameters of 1 and 25, as proposed by Wu et al [57], using the same null model products/objects used in single variant analysis. Significance was evaluated after a Bonferroni correction for multiple testing ($0.05 / 19387 = 2.58 \times 10^{-6}$).

Next, we sought to determine which rare deleterious variants in each significant gene were driving the association signal. We iterated through the variants, removing one variant at a time (*Leave-one-out* approach, *LOO*) [58], and repeated the SKAT 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.

Mining association analysis results. The "Omics Analysis, Search and Information System" (OASIS) [59] is a web-based application for transforming the massive volumes of association results, such as those generated by investigators in the Trans-Omics for Precision Medicine program (TOPMed) Telomere Length Working Group, into biological discovery. OASIS is a one-of-a-kind application that enables fast, efficient data mining integrated with a broad spectrum of functional annotation, online resources (e.g. dbSNP[60], gnomAD [*Genome Aggregation Database (gnomAD)* [61]], GTEx [62], Open Targets Genetics[63], UK Biobank[64] and user-provided "known loci" lists to facilitate identification of novel genetic discoveries. Real-time analysis tools include linkage disequilibrium (LD) calculations, on-demand visualizations (e.g. boxplots, bar charts, histograms, Haploview [65] and LocusZoom [66] plots) and direct integration of selected variants with the UCSC Genome Browser [67] to visualize their proximity to functional regions (e.g. binding sites, Dnase hypersensitivity sites, enhancer/promoter regions). For the telomere length research, OASIS provided customized LD calculations based on genotypes for the actual TOPMed subjects with telomere length phenotypes and for multiple ancestry-based subsets. OASIS automatically fed the customized LD calculations directly to LocusZoom and thus provided an efficient method for producing multiple LocusZoom visualizations for inspection and comparison.

Gene-set enrichment analysis: Gene set enrichment for indexed gene(s) mapping to the 22 GWAS loci was performed using PANTHER [26,27] . Gene set over-representation was evaluated against the GO Ontology Database for all genes in the *Homo sapiens* database using the FISHER test and all sets with an

FDR <0.05 are listed. Input genes were: *TERT, TERC, RTEL1, SH3PXD2A, OBFC1(STN1), SLK, RFWD3, NAF1, ACYP2, TERF1, LINC01592, LOC100505739, TINF2, SAMHD1, TERF2, ZNF676, ZNF729, TCL1A, YY1P2, OPRK1, LRP1B, LINC01429, ATP6V1H, RPN1, DCAF4, POT1, ATM, CHKB-AS1, MAPK8IP2*. There were six unmapped IDs: *TERC, LINC01592, LOC100505739, YY1P2, LINC01429* and *CHKB-AS1*. Index genes were selected based on (i) prior GWAS study definition for known loci, (ii) annotation for each variant mapping directly to a gene in ***Supplementary Table S2A*** for novel loci, and (iii) the exception of the *OBFC1* and *ATM* locus: For the *OBFC1* locus three index genes were selected (*SH3PXD2A, OBFC1(STN1)* and *SLK*) as all three had strong SNV signal not in LD with the sentinel variant (***Supplementary Figure S4D***); and for *ATM*, the sentinel variant mapped to *NPAT*, but was a peak eQTL for *ATM* (***Supplementary Figure S4U***).

Phenome-wide association tests (PheWAS): We queried United Kingdom Biobank (UKBB) GWAS results using the University of Michigan PheWeb web interface (http://pheweb.sph.umich.edu/SAIGE-UKB/). The UKBB PheWeb interface contains results from a SAIGE [68] genetic analysis of 1403 ICD-based traits of 408,961 UKBB participants of European ancestry. PheWeb is a publicly accessible database that allows querying genome-wide association results for 28 million imputed genetic variants. 20 out of our 22 sentinel variants were present in PheWeb. We report all hits passing a Bonferroni correction for the number of tests performed (0.05/(20*1403) = 1.8x10-6).

Expression quantitative trait locus (eQTL) analysis using eQTLGen: The sentinel variants from the meta-analysis results were assessed for their role as eQTLs using the eQTLGen [37] data set, which includes eQTLs found in blood from a set of n=31,684 individuals. For all sentinel variants which were present in eQTLGen, we report all eGenes associated with these variants, as well as the most significant eGene and its FDR-corrected eQTL p-value.

References:

1 Greider, C. W. Telomeres and senescence: the history, the experiment, the future. *Curr Biol* **8**, R178-181 (1998).

2 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).

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 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).

7 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).

8 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).

9 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).

10 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).

11 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).

12 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).

13 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).

14 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).

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

16 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).
17 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).
18 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).
19 Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. *Nucleic Acids Res* **42**, e75, doi:10.1093/nar/gku181 (2014).
20 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).
21 Farmery, J. H. R., Smith, M. L., Diseases, N. B.-R. & Lynch, A. G. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci Rep* **8**, 1300, doi:10.1038/s41598-017-14403-y (2018).
22 Taliun, D. e. a. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv* **https://doi.org/10.1101/563866** (2019).
23 Lin, D. Y. A simple and accurate method to determine genomewide significance for association tests in sequencing studies. *Genet Epidemiol* **43**, 365-372, doi:10.1002/gepi.22183 (2019).
24 Gong, Z. & Chen, J. E3 ligase RFWD3 participates in replication checkpoint control. *J Biol Chem* **286**, 22308-22313, doi:10.1074/jbc.M111.222869 (2011).
25 Majerska, J., Feretzaki, M., Glousker, G. & Lingner, J. Transformation-induced stress at telomeres is counteracted through changes in the telomeric proteome including SAMHD1. *Life Sci Alliance* **1**, e201800121, doi:10.26508/lsa.201800121 (2018).
26 Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703-721, doi:10.1038/s41596-019-0128-8 (2019).
27 Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **8**, 1551-1566, doi:10.1038/nprot.2013.092 (2013).
28 Hansen, M. E. *et al.* Shorter telomere length in Europeans than in Africans due to polygenetic adaptation. *Hum Mol Genet* **25**, 2324-2330, doi:10.1093/hmg/ddw070 (2016).
29 Milne, R. L. *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet* **49**, 1767-1778, doi:10.1038/ng.3785 (2017).
30 Scelo, G. *et al.* Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nat Commun* **8**, 15724, doi:10.1038/ncomms15724 (2017).
31 Wentzensen, I. M., Mirabello, L., Pfeiffer, R. M. & Savage, S. A. The association of telomere length and cancer: a meta-analysis. *Cancer Epidemiol Biomarkers Prev* **20**, 1238-1250, doi:10.1158/1055-9965.EPI-11-0005 (2011).

32 Blackburn, E. H., Epel, E. S. & Lin, J. Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. *Science* **350**, 1193-1198, doi:10.1126/science.aab3389 (2015).
33 Telomeres Mendelian Randomization, C. *et al.* Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases: A Mendelian Randomization Study. *JAMA Oncol* **3**, 636-651, doi:10.1001/jamaoncol.2016.5945 (2017).
34 Rafnar, T. *et al.* Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-related traits. *Nat Commun* **9**, 3636, doi:10.1038/s41467-018-05428-6 (2018).
35 Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429 e1419, doi:10.1016/j.cell.2016.10.042 (2016).
36 Law, P. J. *et al.* Genome-wide association analysis of chronic lymphocytic leukaemia, Hodgkin lymphoma and multiple myeloma identifies pleiotropic risk loci. *Sci Rep* **7**, 41071, doi:10.1038/srep41071 (2017).
37 Vosa, U. e. a. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* (2018).
38 *NHLBI Trans-Omics for Precision Medicine. TOPMed Projects and their Parent Studies. Available at: https://www.nhlbiwgs.org/group/project-studies.*
39 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).
40 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).
41 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).
42 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).
43 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).
44 *MMAP*, <https://github.com/MMAP>
45 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).
46 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).
47 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).

48 *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/>
49 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).
50 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).
51 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).
52 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241, doi:10.1002/humu.22932 (2016).
53 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).
54 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).
55 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).
56 *Ensembl Variation - Calculated variant consequences*, <http://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html>
57 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).
58 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).
59 OASIS Resources, Video Library and Contact Information.
60 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311, doi:10.1093/nar/29.1.308 (2001).
61 Genome Aggregation Database (gnomAD.
62 Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank* **13**, 307-308, doi:10.1089/bio.2015.29031.hmm (2015).
63 Carvalho-Silva, D. *et al.* Open Targets Platform: new developments and updates two years on. *Nucleic Acids Res* **47**, D1056-D1065, doi:10.1093/nar/gky1133 (2019).
64 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).
65 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-265, doi:10.1093/bioinformatics/bth457 (2005).
66 Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337, doi:10.1093/bioinformatics/btq419 (2010).

67 Casper, J. *et al.* The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46**, D762-D769, doi:10.1093/nar/gkx1020 (2018).
68 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).