Heredity (2012) 108, 124–133

www.nature.com/hdy

ORIGINAL ARTICLE

Worldwide genetic structure in 37 genes important in telomere biology

L Mirabello[1], M Yeager[2], S Chowdhury[2], L Qi[2], X Deng[2], Z Wang[2], A Hutchinson[2] and SA Savage[1]
[1]*Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA and* [2]*Core Genotyping Facility, National Cancer Institute, Division of Cancer Epidemiology and Genetics, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD, USA*

Telomeres form the ends of eukaryotic chromosomes and are vital in maintaining genetic integrity. Telomere dysfunction is associated with cancer and several chronic diseases. Patterns of genetic variation across individuals can provide keys to further understanding the evolutionary history of genes. We investigated patterns of differentiation and population structure of 37 telomere maintenance genes among 53 worldwide populations. Data from 898 unrelated individuals were obtained from the genome-wide scan of the Human Genome Diversity Panel (HGDP) and from 270 unrelated individuals from the International HapMap Project at 716 single-nucleotide polymorphism (SNP) loci. We additionally compared this gene set to HGDP data at 1396 SNPs in 174 innate immunity genes. The majority of the telomere biology genes had low to moderate haplotype diversity (45–85%), high ancestral allele frequencies (>60%) and low differentiation (F_{ST} <0.10). Heterozygosity and differentiation were significantly lower in telomere biology genes compared with the innate immunity genes. There was evidence of evolutionary selection in *ACD*, *TERF2IP*, *NOLA2*, *POT1* and *TNKS* in this data set, which was consistent in HapMap 3. *TERT* had higher than expected levels of haplotype diversity, likely attributable to a lack of linkage disequilibrium, and a potential cancer-associated SNP in this gene, rs2736100, varied substantially in genotype frequency across major continental regions. It is possible that the genes under selection could influence telomere biology diseases. As a group, there appears to be less diversity and differentiation in telomere biology genes than in genes with different functions, possibly due to their critical role in telomere maintenance and chromosomal stability.
Heredity (2012) **108,** 124–133; doi:10.1038/hdy.2011.55; published online 6 July 2011

Keywords: telomere; selection; single-nucleotide polymorphism; HGDP; HapMap

Introduction

Telomeres consist of $(TTAGGG)_n$ nucleotide repeats and an associated protein complex located at chromosome ends. They are essential for maintaining chromosomal integrity. Telomere-associated proteins include the telomerase reverse transcriptase (*TERT*) and its RNA component (*TERC*), plus an ordered protein complex, or shelterin, consisting of six proteins: *TERF1*, *TERF2*, *TINF2*, *TERF2IP*, *ACD* and *POT1* (Collins and Mitchell, 2002; de Lange, 2005). This telomere complex, and many other associated proteins, are responsible for preserving chromosome ends, and thus genomic stability, by protecting chromosomes from end-to-end fusion, atypical recombination and degradation (Moon and Jarstfer, 2007). Many of the components of the telomeric complex are highly conserved across species in comparative sequence and functional investigations (Nakamura and Cech, 1998; Li *et al*., 2000; Kanoh and Ishikawa, 2003; de Lange, 2004; Savage *et al*., 2005). It was also shown that seven of these genes (*TERT*, *POT1*, *TNKS*, *TERF1*, *TINF2*, *TERF2* and *TERF2IP*) had lower nucleotide diversity compared with other gene families; they were also highly conserved and the most common allele was ancestral (Savage *et al*., 2005).

Telomere nucleotide repeats progressively shorten with each cell division due to incomplete replication of the 3′ end by DNA polymerases. When they become critically short, cellular senescence or cellular crisis is induced in normal cells but in malignant cells this pathway is bypassed through the activation of telomerase or the alternative pathways (Gilley *et al*., 2005; Rodier *et al*., 2005). Short telomeres induce genetic instability and thereby promote the initiation and development of cancer (Blasco *et al*., 1997; Rudolph *et al*., 1999, 2001; Wu *et al*., 2003; Plentz *et al*., 2003, 2004). Telomere attrition has also been associated with aging, many diseases (including diabetes mellitus and cardiovascular disease), inflammation, oxidative stress, an unhealthy lifestyle and smoking (von Zglinicki, 2002; Wong and Collins, 2003; Morlá *et al*., 2006; Aubert and Lansdorp, 2008; Mirabello *et al*., 2009). Several disorders are associated with mutations in telomere biology genes (Crabbe *et al*., 2004; Blasco, 2007; Vulliamy *et al*., 2008; Armanios, 2009; Savage and Alter, 2009). Patients with dyskeratosis congenita, a heterogeneous inherited bone marrow failure and cancer predisposition syndrome, have extremely short telomeres and germline mutations in genes important in the maintenance of telomeres (*DKC1*, *TERC*,

Correspondence: Dr SA Savage, Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, NIH 6120 Executive Boulevard., EPS/7018, Rockville, MD 20892, USA.
E-mail: savagesh@mail.nih.gov
Received 22 February 2011; revised 10 May 2011; accepted 31 May 2011; published online 6 July 2011

TERT, *NOLA3* (alias *NOP10*), *TINF2* or *NOLA2* (alias *NHP2*)) (Crabbe *et al*., 2004; Vulliamy *et al*., 2008; Armanios, 2009; Savage and Alter, 2009). In addition, recent genome-wide association studies found that genetic variation at 5p15.33 (*TERT-CLPTM1L* locus) was associated with risk of glioma (Shete *et al*., 2009), basal cell carcinoma (Stacey *et al*., 2008, 2009), testicular cancer (Turnbull *et al*., 2010), pancreatic cancer (Petersen *et al*., 2010) and lung cancer (McKay *et al*., 2008; Jin *et al*., 2009; Landi *et al*., 2009); an association study of multiple tumor types suggest that this region may contain important markers of overall cancer risk (Rafnar *et al*., 2009).

The extent to which disease-associated alleles differ in frequency between populations and the evolutionary forces responsible for the observed degree of population differentiation may provide keys to further understanding disease pathogenesis. Allele frequencies for many genetic variants differ by geographical regions (Guthery *et al*., 2007; Lan *et al*., 2007; Myles *et al*., 2008), possibly the result of several factors including natural selection and neutral genetic drift. There may be functional consequences of a particular variant that leads to a more favorable response and thus certain variants may be under selective pressure. Searching for a signature of selection has the potential to identify functional and disease related variants (Bamshad and Wooding, 2003; Hurst, 2009).

We examined patterns of differentiation, allele frequencies and the haplotype structure of 37 genes important in telomere biology among 53 worldwide populations from Africa, the Middle East, Europe, Central/South (C/S) Asia, East Asia, Oceania and the Americas. Data from 1168 unrelated individuals were obtained from the genome-wide scan of the Human Genome Diversity Panel (HGDP-CEPH) (Cann *et al*., 2002; Li *et al*., 2008) and from the International HapMap Project (The International HapMap Consortium, 2003) at 716 single-nucleotide polymorphism (SNP) loci. We additionally compared our telomere gene set to HGDP-CEPH data that we had on 174 innate immunity genes at 1396 SNPs, this allowed us to determine if two sets of genes grouped by function have similar genetics. We hypothesized that genetic variation in telomere biology genes may be constrained because of both the high degree of sequence similarity previously observed across species and the critical roles their protein products have in chromosomal stability.

Materials and methods

Data set

We obtained SNP data for each gene, including 20 kbp upstream and 10 kbp downstream, from the HGDP-CEPH (Cann *et al*., 2002) genome-wide scan of 650 000 common SNPs (Li *et al*., 2008) and the HapMap Phase 2 (The International HapMap Consortium, 2003) public database. Genotype data were retrieved for 37 gene regions: *ACD*, *ATM*, *BLM*, *DDX1*, *DDX11*, *DKC1*, *MRE11A*, *NBN*, *NOLA1*, *NOLA2*, *NOLA3*, *PARP1*, *PARP2*, *PINX1*, *POT1*, *PRKDC*, *RAD50*, *RAD51AP1*, *RAD51C*, *RAD51L1*, *RAD51L3*, *RAD54L*, *RECQL*, *RECQL4*, *RECQL5*, *RTEL1*, *TEP1*, *TERC*, *TERF1*, *TERF2*, *TERF2IP*, *TERT*, *TINF2*, *TNKS*, *TNKS2*, *WRN* and *XRCC6*, as well as the region between *PARP2*/*TEP1*. Genes were chosen based on their involvement in telomere biology or presumed interaction with telomeres as reported in the literature. Telomerase complex genes include *TERC*, *TERT*, *DKC1*, *TEP1*, *NOLA1*, *NOLA2* and *NOLA3*; shelterin genes include *TERF1*, *TERF2*, *TERF2IP*, *POT1*, *TINF2* and *ACD*; DNA repair genes include *XRCC6*, *NBN*, *RAD50*, *ATM*, *RAD54L*, *RAD51L3*, *RAD51C*, *RAD51AP1*, *RAD51L1* and *MRE11A*; helicase genes include *WRN*, *BLM*, *RECQL*, *RECQL4*, *RECQL5*, *DDX1* and *DDX11*; and, other telomere-associated genes include *PRKDC*, *PARP1*, *PARP2*, *PINX1*, *TNKS*, *TNKS2* and *RTEL1*. All SNPs, regardless of minor allele frequency, were included in the analysis as for many of these genes there were only a few SNPs available. Data were retrieved for all individuals in the 52 populations (952 individuals) included in the HGDP-CEPH 952 panel and the four populations (270 individuals) of the HapMap project for the same 716 SNPs. Atypical and related individuals were removed (Rosenberg, 2006), which resulted in 898 individuals from the HGDP-CEPH panel and 270 from the HapMap project. The final data set included 1168 unrelated individuals from 53 unique populations. We did not limit the SNP data to SNPs only within exons, introns, promoters or 3′areas because the goal of this study was to understand the gene regions, including upstream and downstream regions.

We also obtained genotype data for 174 genes involved in innate immunity as a comparison set for our telomere biology genes. SNP data for each gene were acquired from the HGDP-CEPH (Cann *et al*., 2002) genome-wide scan for all individuals in the 52 populations, and cleaned as described above. The immune gene set was chosen as a comparison gene set as these genes are often highly variable. Additional comparisons were made with data reported in the literature. Supplementary Table 1 lists the 174 innate immune gene regions evaluated.

HapMap phase 3 (The International HapMap Consortium, 2003) SNP data for 11 populations (1115 individuals) were also retrieved for a subset of the telomere maintenance genes that were potential candidates for evolutionary selection (defined in results): *ACD*, *NOLA2*, *RECQL4*, *POT1*, *TERF2IP* and *TNKS*, as well as for *TERT*. Individuals in this phase do not overlap with HapMap phase 2 participants.

Data analysis

Haplotype and SNP frequencies were estimated using a Bayesian algorithm implemented in PHASE version 2.1 (Stephens *et al*., 2001; Stephens and Scheet, 2005). Haplotypes determined by PHASE were used as input for all other analyses. The package ARLEQUIN version 3.11 (Excoffier *et al*., 2005) was used to compute haplotype diversity, F_{ST} values, Mantel test, analysis of molecular variance (AMOVA) and heterozygosity. F_{ST} values based on allele frequencies were calculated as a measure of population differentiation and significance was estimated with 10 000 permutations. A Mantel test was used to test the significance of the regression of genetic distance on geographic distance between population pairs with 10 000 permutations. In order to apportion the fraction of the genetic variance due to differences between and within continental groups and infer the genetic structure of the populations, AMOVA

was performed with 10 000 permutations. Mega version 4.0 (Tamura *et al*., 2007) was used to construct a neighbor-joining tree based on genetic differentiation. Population structure was inferred by a Bayesian clustering analysis performed with structure version 2.2 (Pritchard *et al*., 2000; Falush *et al*., 2007) using the following settings: admixture model, correlated markers, K = 1–10, a length of 100 000 for the burn-in period, and 100 000 repetitions following the burn-in period. Haploview version 4.1 (Barrett *et al*., 2005) was used to determine the degree of linkage disequilibrium (LD) and minor allele frequencies (MAF). LD *P* values (with s.e.) were estimated by Monte Carlo approximation with 10 000 steps in the Markov Chain using ARLEQUIN. This LD calculation is an extension of Fisher exact probability test on contingency tables, and the results are given as a significance level of LD for each pair of loci with a small *P* value (<0.05) indicating high LD (Excoffier *et al*., 2005; Santos-Lopes *et al*., 2007). Differences between the telomere and immune gene set results were tested for significance with parametric (*t*-test) and non-parametric tests (Mann–Whitney *U*-test).

We retrieved ancestral (chimpanzee) alleles for 98.2% of the SNPs using the UCSC Genome Browser (March 2006 Assembly: http://genome.ucsc.edu/) and/or Ensembl (release 50, Jul 2008: http://www.ensembl.org/index.html). In cases where neither human allele corresponded to the chimpanzee allele or when the chimpanzee allele was unknown we excluded these SNPs from the analysis. Pairwise geographic distance between populations and distance from Addis Ababa, Ethiopia (the putative point of origin of modern humans (White *et al*., 2003)) was estimated in kilometers (km) following the likely colonization route (shortest path through landmasses) as in Prugnolle *et al*. (2005).

Selection was evaluated with the following analyses: (1) between population differentiation (F_{ST}), which can be inflated due to environmental pressures on populations causing local adaptation and allele frequency changes (positive directional selection), and negative or balancing selection can decrease the differentiation of selected loci (Akey *et al*., 2002; Nielsen, 2005; Sabeti *et al*., 2007); (2) genetic diversity, a significant decrease points to positive selection when a particular allele is favored, and increases could be balancing selection with more diversity being potentially adaptive; (3) LD across populations, selection can increase LD; and (4) MAF and derived allele frequency (DAF) tests (Walsh *et al*., 2006). The iHS for HapMap phase 2 data was retrieved with HAPLOTTER (The International HapMap Consortium, 2005; Voight *et al*., 2006).

Results

We analyzed SNP data from the HGDP-CEPH (Cann *et al*., 2002) genome-wide scan for 37 gene regions involved in telomere biology and from Phase 2 of the International HapMap project (The International HapMap Consortium, 2003). The telomere data set consisted of a total of 716 SNPs in 1168 individuals from 53 worldwide populations. Supplementary Tables 1 and 2 give summary statistics for the telomere biology and innate immune gene regions analyzed (for example, alleles, MAF, heterozygosity, variation components).

Population structure and differentiation

Bayesian cluster analysis and a distance-based neighbor-joining tree segregated individuals into five genetic clusters: Africa, Eurasia (Middle East, Europe, Utah, C/S Asia), East Asia, Oceania and America (Supplementary Figure 1). The Utah, USA population clustered within Eurasia, with high genetic similarity to the European populations. We tested for isolation by distance using a Mantel test on F_{ST} estimates and found a significant positive correlation between the degree of genetic divergence and the pairwise geographical distance (correlation coefficient (r) = 0.64, $P<0.001$).

Table 1 shows the levels of differentiation (F_{ST}) by gene sorted in descending order by among regions differentiation. *PRKDC* and *POT1* have the lowest levels of differentiation, and nearly half (0.44) of the *POT1* SNPs have global F_{ST} estimates less than the 0.05 percentile of the overall F_{ST} distribution. *ACD* and *TERF2IP* had the highest levels of differentiation, and a large portion of their SNPs were above 0.95 (0.60 and 0.50, respectively) and 0.99 (0.40 and 0.38, respectively) percentiles. *ACD* had very high levels of differentiation observed between HapMap Yoruba and Utah populations and between Yoruba and Chinese/Japanese populations (0.52 and 0.75, respectively), and *TERF2IP* between Utah and Chinese/Japanese populations (0.56).

Overall, *TERT* had average levels of differentiation among regions and HapMap populations (Table 1 and Supplementary Table 3). Limiting the SNPs to only those within the *TERT* gene (introns, exons and UTRs, $n=4$ SNPs), the levels of differentiation were lower among regions ($F_{st}=0.072$) and within populations ($F_{st}=0.089$). We further evaluated the recently identified *TERT* SNP, rs2736100 (localized to intron 2: at chromosome 5p15.33, position 1 339 516) as it appears to be associated with risk of lung cancer, testicular cancer and glioma (McKay *et al*., 2008; Jin *et al*., 2009; Landi *et al*., 2009; Shete *et al*., 2009; Turnbull *et al*., 2010). rs2736100 had variable levels of differentiation among geographical regions. Its genotype frequencies varied among regions and levels of pairwise differentiation were particularly high among Oceania and all other regions, as well as among America and Eurasia, and low among Africa, East Asia and Eurasia (Supplementary Table 3 and Supplementary Figure 2).

As a gene set, the AMOVA partitions variation among the seven geographical regions similar to that observed for the entire 650k autosomal SNP panel (Li *et al*., 2008) with within-population variation accounting for the majority of the genetic diversity (Figure 1). There was some disparity in how variation is partitioned among individual genes. There was substantially higher among-regions variation observed in *ACD* and *TERF2IP* and the least in *POT1* (Figure 1). There was significantly lower differentiation observed among geographical regions for the telomere biology genes compared with the innate immune genes ($P=0.0002$) (Figure 1), and the distributions of F_{ST} values among all the HGDP populations showed a shift down towards lower F_{ST} for the telomere biology genes (Figure 2). AMOVA variation components and differentiation by locus are shown in Supplementary Tables 1 and 2 for the innate immune and telomere biology genes, respectively. Grouping the genes by function showed that telomerase complex genes had the lowest F_{ST} values (among regions = 0.07), followed by helicase genes, and other telomere associated genes

Table 1 Levels of differentiation (F_{ST}) by gene using HapMap 2 and HGDP data

Telomere biology gene	*AMOVA F_{ST}*[a]		*Pairwise F_{ST}*[b]			*SNPs with global F_{ST} values*[c]			
	AR	*WP*	*YRI vs CEU*	*YRI vs CHB+JPT*	*CEU vs CHB+JPT*	*<0.01 pct*	*<0.05 pct*	*>0.95 pct*	*>0.99 pct*
PRKDC	0.005*	0.082	0.001*	0.000*	0.000*	0 (0)	0 (0)	0 (0)	0 (0)
POT1	0.033	0.053	0.034	0.071	0.009*	0 (0)	10 (0.435)	0 (0)	0 (0)
MRE11A	0.039	0.064	0.077	0.049	0.035	0 (0)	1 (0.063)	0 (0)	0 (0)
PARP2/TEP1	0.040	0.066	0.109	0.104	0.0001*	0 (0)	1 (0.250)	0 (0)	0 (0)
NOLA3	0.059	0.076	0.102	0.057	0.075	0 (0)	0 (0)	0 (0)	0 (0)
TNKS	0.061	0.083	0.069	0.042	0.059	1 (0.019)	1 (0.019)	0 (0)	0 (0)
NOLA2	0.061	0.078	0.119	0.001*	0.101	0 (0)	0 (0)	0 (0)	0 (0)
RECQL5	0.065	0.083	0.123	0.071	0.079	0 (0)	0 (0)	0 (0)	0 (0)
TNKS2	0.072	0.104	0.105	0.082	0.073	0 (0)	0 (0)	0 (0)	0 (0)
WRN	0.075	0.096	0.071	0.166	0.061	0 (0)	1 (0.023)	0 (0)	0 (0)
TINF2	0.075	0.094	0.178	0.014	0.164	0 (0)	2 (0.400)	0 (0)	0 (0)
TEP1	0.076	0.099	0.147	0.188	0.040	0 (0)	3 (0.120)	0 (0)	0 (0)
RAD54L	0.078	0.101	0.087	0.166	0.045	0 (0)	0 (0)	0 (0)	0 (0)
RAD51L3	0.084	0.108	0.085	0.088	0.149	0 (0)	0 (0)	0 (0)	0 (0)
PINX1	0.089	0.118	0.104	0.217	0.083	1 (0.024)	2 (0.048)	0 (0)	0 (0)
ATM	0.091	0.117	0.087	0.094	0.031	0 (0)	1 (0.067)	0 (0)	0 (0)
DDX1	0.103	0.127	0.209	0.162	0.111	1 (0.063)	1 (0.063)	0 (0)	0 (0)
TERT	0.107	0.126	0.025	0.159	0.109	0 (0)	0 (0)	0 (0)	0 (0)
RECQL	0.108	0.135	0.155	0.265	0.043	0 (0)	2 (0.061)	2 (0.061)	0 (0)
NOLA1	0.109	0.131	0.238	0.146	0.052	0 (0)	0 (0)	0 (0)	0 (0)
BLM	0.110	0.124	0.140	0.198	0.056	0 (0)	2 (0.077)	0 (0)	0 (0)
NBN	0.113	0.132	0.233	0.152	0.079	0 (0)	0 (0)	2 (0.083)	0 (0)
RAD51C	0.115	0.140	0.124	0.107	0.182	0 (0)	0 (0)	0 (0)	0 (0)
RAD51L1	0.127	0.157	0.142	0.227	0.112	2 (0.011)	3 (0.017)	12 (0.068)	0 (0)
PARP2	0.130	0.146	0.318	0.297	0.014	0 (0)	1 (0.100)	1 (0.100)	0 (0)
TERF1	0.134	0.149	0.095	0.171	0.194	0 (0)	0 (0)	1 (0.056)	0 (0)
TERF2	0.140	0.167	0.041	0.186	0.131	0 (0)	0 (0)	0 (0)	0 (0)
RECQL4	0.142	0.165	0.295	0.312	0.000*	0 (0)	0 (0)	1 (0.250)	1 (0.250)
DDX11	0.151	0.167	0.175	0.062	0.255	0 (0)	0 (0)	0 (0)	0 (0)
PARP1	0.159	0.174	0.068	0.161	0.209	0 (0)	0 (0)	0 (0)	0 (0)
RAD50	0.180	0.191	0.323	0.294	0.070	3 (0.158)	3 (0.158)	2 (0.105)	1 (0.053)
DKC1[d]	0.194	0.246	0.292	0.490	0.067	0 (0)	0 (0)	0 (0)	0 (0)
XRCC6	0.204	0.221	0.248	0.226	0.039	0 (0)	0 (0)	1 (0.200)	0 (0)
RAD51AP1	0.213	0.236	0.169	0.334	0.091	0 (0)	0 (0)	3 (0.250)	1 (0.083)
TERC	0.235	0.236	0.139	0.486	0.292	0 (0)	0 (0)	0 (0)	0 (0)
RTEL1	0.250	0.277	0.137	0.368	0.208	0 (0)	0 (0)	4 (0.400)	0 (0)
ACD	0.362	0.387	0.519	0.749	0.139	0 (0)	0 (0)	3 (0.600)	2 (0.400)
TERF2IP	0.368	0.393	0.380	0.442	0.561	0 (0)	2 (0.250)	4 (0.500)	3 (0.375)

Abbreviations: AMOVA, analysis of molecular variance; AR, among regions; SNP, single-nucleotide polymorphism; WP, within populations. Shaded region indicates the genome-wide autosomal SNP average level of differentiation (Akey *et al*., 2002; Shriver *et al*., 2004, 2005; Weir *et al*., 2005).

[a]AMOVA using five regions: Africa, Eurasia (Middle East, C/S Asia, Europe and USA), East Asia, Oceania and America.

[b]Among HapMap two populations: YRI, Yoruba; CEU, Utah; CHB, Han and JPT, Japanese.

[c]Number (proportion) of SNPs with global F_{ST} in the respective percentile (pct) of the full F_{ST} distribution.

[d]Located on the X chromosome, F_{ST} values are expected to be elevated as compared with markers on autosomes.

*P value >0.05, all other P values are <0.01.

compared with the genome-wide average for autosomal SNPs (0.10–0.15 (Akey *et al*., 2002; Shriver *et al*., 2004, 2005; Weir *et al*., 2005)) and the innate immune gene set, as well as among HapMap 2 populations in comparison to other gene sets (Table 2).

Haplotype diversity and LD

The number of haplotypes and diversity estimates by gene and region are shown in Table 3 and Supplementary Table 4. Overall, the haplotype diversity was highest in Africa (0.844) and lowest in Oceania (0.634; Table 3). The majority of genes had very low to moderate haplotype diversity (Supplementary Table 4). *DKC1*, *TERC* and *XRCC6* had very low haplotype diversity of less than 50%, and *TINF2*, *NOLA2* and *RECQL4* had low diversity estimates between 60–70%. Surprisingly, *TERT* had high haplotype diversity (ranging from 81 to 96%). Using only the SNPs within the *TERT* gene, the haplotype diversity was lower in East Asia, Oceania and America (ranging from 37 to 74%), and higher in Eurasia and Africa (88% and 90%, respectively) (Supplementary Figure 2). The mean haplotype diversity for all of the 53 populations was negatively correlated with geographic distance from Ethiopia ($r=-0.89$, slope $=-1.17e^{-5}$). Heterozygosity in the telomere gene set was significantly lower than in the immune gene set by geographical region ($P=0.014$; Table 3).

LD P values were estimated for all marker pairs for each gene and a summary of the proportion of SNPs with significant LD ($P<0.05$) is presented in Table 4. The proportion of marker pairs with significant LD varied among geographic regions and genes. A low proportion of marker pairs with LD was often observed in Oceania.

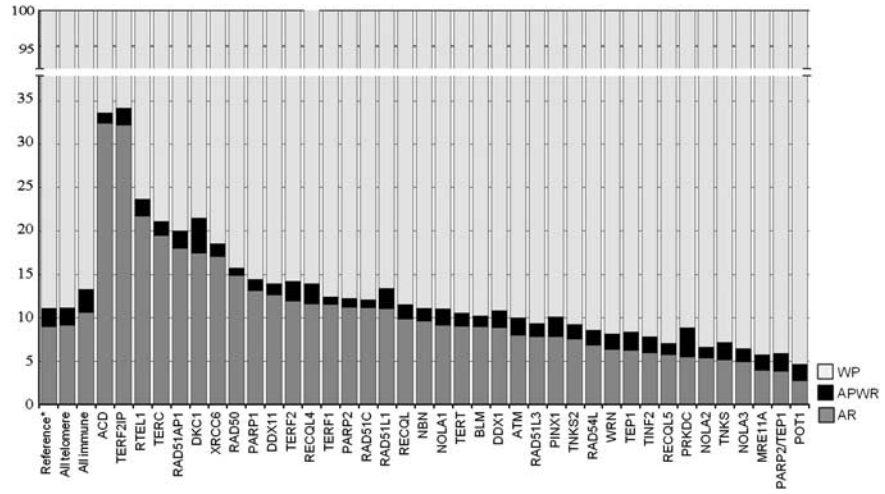

Figure 1 Analysis of molecular variance by gene using HapMap 2 and HGDP data. Partitioning variation into three components: within population (WP), among region (AR) and among-population-within-region (APWR). Populations are assigned to the seven main geographic regions from the HGDP-CEPH panel; *HGDP panel at 650k autosomal SNPs (Li *et al*., 2008).

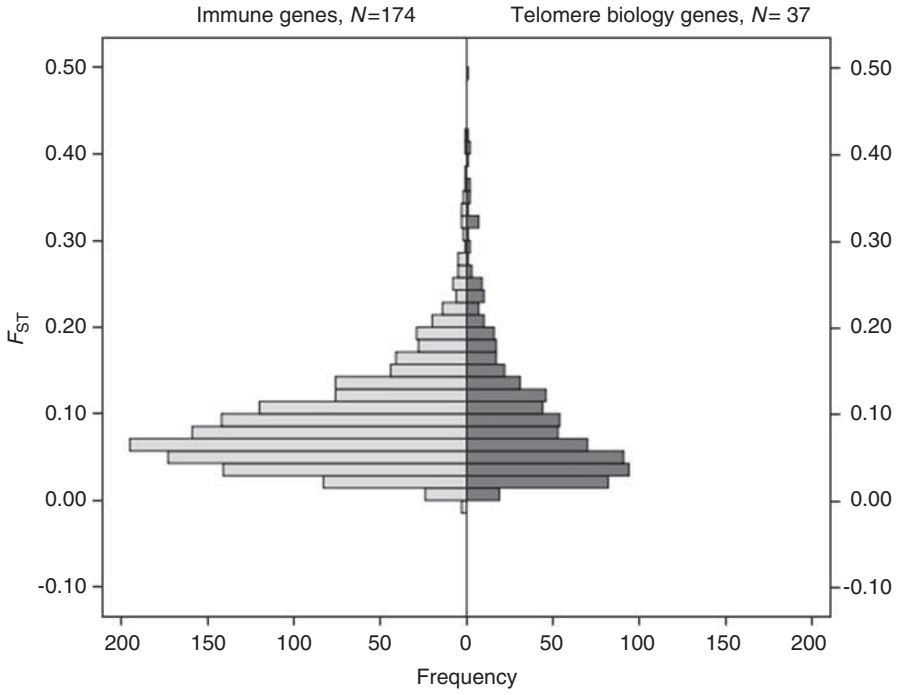

Figure 2 Differentiation among the HGDP populations for the telomere biology and innate immune gene sets.

The lowest LD was observed in *RAD51L1* ($\leqslant$0.4 of marker pairs in all populations) and the highest in *POT1* ($>$0.9). However, this analysis is limited by the small number of SNPs in many of these genes and the limited population sizes in Oceania and America.

Ancestral alleles

Comparing the ancestral allele frequency (AAF) spectrums among the HapMap 2 and HGDP populations, we found that populations in Africa had more SNPs with high AAFs and populations in America had the lowest (Figure 3). A steeper slope of SNP counts in the midrange of the distribution reflects more SNPs with high AAFs (Li *et al*., 2008). The slopes of SNP counts in the range of 0.2–0.8 AAF for all of the populations progressively declined moving away from Ethiopia (3.8–0.1; Figure 3b). This AAF pattern did not change after limiting the SNPs to only those in exons, introns and UTR regions. The average AAF was highest in African populations (0.735) and lowest in American populations (0.655) (data not shown). Average AAFs for the majority of our genes was high ($>$60%). For *TERT*, the average AAF was 64%, and for the *TERT* SNP, rs2736100, it was 49%. The AAF for rs2736100 (ancestral allele: T) was variable by geographic region, the highest

Table 2 Genetic differentiation (F_{ST}) among major continental groups and HapMap 2 populations in comparison to other data sets

Gene sets or SNP data sets	*Among HapMap 2 populations*			*Among regions*
	YRI vs CEU	*YRI vs CHB+JPT*	*CEU vs CHB+JPT*	
Telomere biology genes (37 genes, 716 SNPs, $N=1168$)[b]	0.136	0.189	0.095	0.095[a]
Shelterin genes	0.140	0.224	0.148	0.109
Telomerase complex genes	0.130	0.161	0.071	0.072
DNA repair genes	0.154	0.209	0.099	0.107
Helicase genes	0.139	0.196	0.071	0.085
Other telomere-associated genes	0.098	0.144	0.093	0.086
CVD genes (364 genes, 15 559 SNPs, $N=270$)[c]	0.139	0.158	0.095	—
Blood circulation and gas exchange genes	0.129	0.158	0.062	—
Lipoprotein metabolism genes	0.144	0.165	0.082	—
Insulin/IGF-mitogen-activated protein kinase kinase/ MAP kinase cascade genes	0.194	0.174	0.138	—
Disease associated SNPs (25 SNPs, $N=952$)[d]	—	—	—	0.10[a]
Innate immune genes (174 genes, 1396 SNPs, $N=917$)	—	—	—	0.11
Genome-wide autosomal SNP average[e]	—	—	—	0.10–0.15

Abbreviations: CEU, Utah; CHB, Han; CVD, cardiovascular disease; IGF, insulin-like growth factor; JPT, Japanese; MAP, mitogen-activated protein kinase; N = number of individuals; SNP, single-nucleotide polymorphism; YRI, Yoruba.
[a]Among the seven geographic regions represented in the CEPH-HGDP panel.
[b]See methods for a description of genes included in each gene set.
[c](Kullo and Ding, 2007)
[d](Myles *et al*., 2008)
[e](Akey *et al*., 2002; Shriver *et al*., 2004, 2005; Weir *et al*., 2005).

Table 3 Telomere biology genes average diversity of the major continental regions from the HapMap 2 project and HGDP-CEPH panel

Region	N	2N	*Average h_d*	*Het telomere gene set*	*Het immune gene set*	P[a]
Africa	185	370	84.37	0.299	0.286	0.059
Middle East	161	322	76.91	0.297	0.324	0.00005
Europe	156	312	75.04	0.309	0.330	0.0017
C/S Asia	198	396	73.04	0.300	0.324	0.0003
USA[b]	90	180	70.55	—	—	—
East Asia	316	632	72.19	0.295	0.303	0.284
Oceania	21	42	63.35	0.313	0.301	0.203
America[c]	41	82	65.51	0.287	0.264	0.0057

Abbreviations: h_d, haplotype diversity (%); *het*, heterozygosity in each region using HGDP-CEPH data; HGDP, Human Genome Diversity Panel; *N*, number of individuals; *2N*, number of chromosomes.
[a]for differences between the heterozygosity of the telomere and immune gene sets
[b]Includes individuals from Utah.
[c]Includes populations in Brazil, Colombia and Mexico.

AAF was observed in America (0.88) and the lowest in Oceania (0.045).

Test for selection

For evaluating F_{ST}, we concentrated on regions that show high or low values among multiple markers, as individual SNPs show considerable variation. According to the cut-points estimated by Akey *et al*. (Akey *et al*., 2002), *ACD* and *TERF2IP* had the highest proportion of SNPs with high F_{ST} ($\geqslant 0.45$) (0.6 and 0.5, respectively), and *TNKS*, *RAD51L1* and *RECQL* all had very low F_{ST} (two SNPs with an $F_{ST}=0$ and one SNP with an F_{ST} $\leqslant 0.005$). We also plotted the average F_{ST} versus the average heterozygosity by region. There were three outliers with high F_{ST} and low heterozygosity, *ACD*, *TERF2IP* and *TERF2*, and two outliers with low F_{ST} and high heterozygosity, *POT1* and *NOLA2* (data not shown). There were three outliers in the DAF test with a large amount of derived alleles (>80%): *ACD* (in CEU and Han populations), *NOLA2* (in CEU and America populations) and *RECQL4* (in Oceania) (data not shown). All of the loci in *TERC* ($n=3$) and *XRCC6* ($n=5$) had DAFs of <20% in all populations and in CEU and Han populations, respectively. The MAF test suggests *RECQL4* (Han), *POT1* (Han and CEU) and *RAD54L* (America) with an excess of SNPs with high MAF (>40%) and *ACD* (Han), *TERF2IP* (CEU), and *RAD51L3* (America) with an excess of SNPs with low MAF (<10%). The strong LD observed for *POT1* supports the existence of balancing selection. Overall, *TERT* did not show evidence of selection.

HapMap 3 data for select genes

HapMap 3 (The International HapMap Consortium, 2003) SNP data for 11 populations were retrieved for genes identified as potential evolutionary selection candidates in this study (based on at least two tests: *ACD*, *NOLA2*, *RECQL4*, *POT1 TERF2IP*), and previous studies (Savage *et al*., 2005 and the HapMap (The International HapMap Consortium, 2007): *TNKS*), to confirm our findings in an additional data set with a more dense SNP coverage. We also retrieved SNP data

Table 4 Proportion of marker pairs with significant linkage disequilibrium[a] using HapMap 2 and HGDP data

Telomere biology gene[b]	n	*Population*[c]				
		YRI (N = *110*)	*CEU* (N = *90*)	*CHB* (N = *89*)	*Oceania* (N = *21*)	*America* (N = *41*)
ATM	15	0.818	0.836	0.806	1.000	0.600
BLM	26	0.659	0.717	0.709	0.695	0.529
DDX1	16	0.733	0.697	0.581	0.308	0.552
MRE11A	16	0.783	0.825	1.000	0.982	0.582
NBN	24	0.593	0.575	0.721	0.577	0.559
NOLA3	13	0.745	0.615	0.641	0.462	0.731
PARP1	16	0.724	0.978	0.857	0.800	0.736
PARP2	10	0.714	0.600	0.750	0.619	0.667
PINX1	42	0.756	0.840	0.683	0.950	0.359
POT1	23	0.929	1.000	1.000	1.000	0.925
PRKDC	18	0.508	0.581	0.562	0.622	0.533
RAD50	19	0.634	0.756	0.864	0.652	0.527
RAD51AP1	12	0.848	0.911	0.889	0.933	0.622
RAD51L1	177	0.277	0.404	0.331	0.275	0.308
RAD51L3	11	0.444	0.929	0.489	0.286	0.571
RAD54L	10	0.714	0.917	0.750	NA	0.750
RECQL	33	0.679	0.750	0.939	0.570	0.772
RECQL5	7	0.714	1.000	1.000	0.533	0.762
RTEL1	10	0.528	0.778	0.821	0.528	0.578
TEP1	25	0.446	0.438	0.352	0.267	0.447
TERF1	18	0.713	0.895	0.758	0.706	0.721
TERF2	6	0.667	0.933	0.733	NA	NA
TERF2IP	8	0.667	0.429	0.571	0.467	NA
TERT	8	0.571	0.536	0.524	0.095	0.381
TNKS	52	0.813	0.797	0.879	0.703	0.606
TNKS2	11	0.855	0.867	0.822	0.639	0.556
WRN	43	0.680	0.769	0.710	0.492	0.690

Abbreviations: CEU, Utah; CHB, Han; HGDP, Human Genome Diversity Panel; *n*, number of SNPs; *N*, number of individuals; NA, ⩽5 polymorphic pairs of loci; YRI, Yoruba.
[a]See methods for a description of this analysis; significant LD refers to a significant *P* value for rejecting the null hypothesis of free recombination.
[b]Only genes with >5 polymorphic pairs of loci are shown due to the uncertainty of the estimation with so few loci.
[c]One population was chosen to represent Africa (YRI), Eurasia (CEU) and East Asia (CHB), due to limited sample sizes all of the populations in Oceania and America were combined.

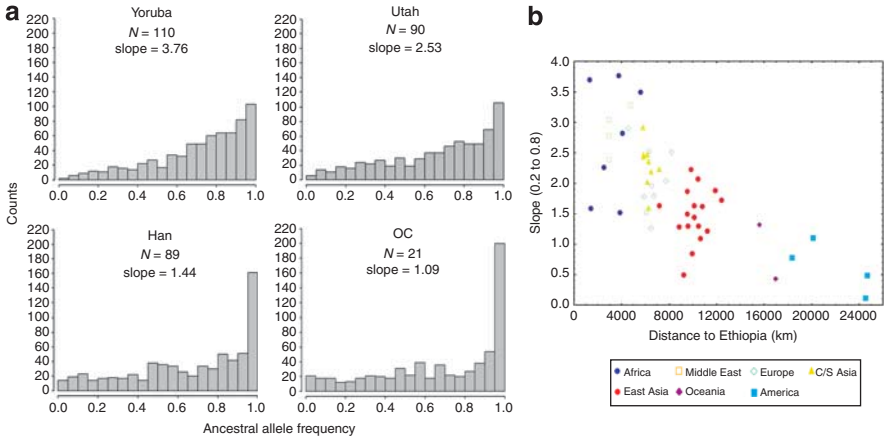

Figure 3 Ancestral allele frequency spectrum using HapMap 2 and HGDP data. (**a**) Histograms of AAFs for four populations: Yoruba, USA, Han and OC (the two populations in Oceania were combined due to small sample sizes). N is the number of individuals and the slope is for the SNP counts in the range of 0.2–0.8 AAF. (**b**) Slopes of AAFs between 0.2 and 0.8 for all of the 53 populations versus geographic distance from Ethiopia.

for *TERT*. Assigning the populations to the main geographic regions (identified by a distance-based neighbor-joining tree, Supplementary Figure 3), the AMOVA partitions the majority of the genetic diversity to within-population variation (91%); there is less variance attributable to among regions in *NOLA2*, *POT1* and *TNKS* (<5%), more variance among regions in *TERF2IP* and *ACD* (>20%) and average in *RECQL4* (14%), as observed in the HapMap 2 and HGDP data set.

TERT had high haplotype diversity (96–99%) and heterozygosity (0.26–0.35) in these 11 populations. There was average differentiation among geographical regions and within populations based on allele frequencies ($F_{ST} = 0.118$ and 0.138, respectively), similar to the

HapMap 2 and HGDP data set. However, population comparisons based on haplotype frequencies were all low, with $F_{ST} < 0.025$.

There was evidence of positive selection (high F_{ST}, low heterozygosity, high or low DAFs and low MAFs) in *ACD* and *TERF2IP*, and evidence of balancing selection (low F_{ST}, high heterozygosity and high MAFs) in *POT1*, *NOLA2* and regions of *TNKS* (Supplementary Figure 4). There were mixed signals in *RECQL4*, with a proportion of SNPs with extreme high and low F_{ST}, heterozygosity and MAFs. The patterns of variance attributable to among regions are also consistent, with extremely high values in *ACD* and *TERF2IP* and low in *POT1*, *NOLA2* and *TNKS*.

Discussion

In this study, we examined allele frequency distributions, diversity, differentiation, LD and population structure among 53 worldwide populations by combining HapMap 2 and HGDP-CEPH genome-wide scan data of 37 genes vital for telomere stability. This extensive data set allowed us to create a comprehensive catalog of worldwide genetic variation for these genes. Overall, most telomere biology genes had low to moderate diversity and less than average differentiation. There was significantly lower differentiation among HGDP populations and heterozygosity in the telomere biology genes compared with innate immunity genes. Differentiation among geographical regions in the telomere biology genes grouped by function showed the lowest values in the telomerase complex genes compared with other gene sets and the genome average. These genes are required for telomere elongation and maintaining chromosomal stability.

As a gene set, there is a specific population structure; cluster analyses segregated individuals into five genetic clusters, concordant with larger analyses with the HGDP-CEPH panel (Rosenberg *et al*., 2002; Jakobsson *et al*., 2008; Li *et al*., 2008). The significant positive correlation between the degree of genetic divergence and the pairwise geographical distance suggests that the observed genetic differentiation can be partially explained by isolation by geographic distance, which agrees with previous data (Ramachandran *et al*., 2005; Jakobsson *et al*., 2008). As expected, the mean haplotype diversity and AAFs were highest in Sub-Saharan Africa. For all populations, diversity and AAF slopes were negatively correlated with geographic distance from Addis Ababa, Ethiopia, consistent with a serial founder model during a spatial expansion from Africa (Ramachandran *et al*., 2005). The AAFs for the majority of telomere maintenance genes were high, with most having an average AAF $>60\%$, and the AAF slopes were much higher (range of 0.1–3.8) than observed by Li *et al*. (2008) (range of 0.001–0.004).

The high AAF, low diversity and differentiation in many of these genes and gene sets suggest that they may be constrained, possibly because of their essential roles in chromosomal stability. Several telomere maintenance genes have been previously shown to be highly conserved across species (Nakamura and Cech, 1998; Li *et al*., 2000; Kanoh and Ishikawa, 2003; de Lange, 2004; Savage *et al*., 2005). This conservation can be explained by a low mutation rate and/or negative selection, however, distinguishing the two is a difficult task as both result in little sequence change (Hurst, 2009). Savage *et al*., 2005 also found that seven of these genes had more synonymous compared with non-synonymous mutations per site. A plausible explanation for the lower levels of diversity and differentiation observed in many telomere maintenance genes is that negative selection acts to maintain the *status quo* of these essential genes. Perhaps these genes were highly conserved during evolution because of their important function and the accumulation of new mutations was not tolerable.

Negative, positive and balancing selection can each leave a specific signature on allele frequency patterns and LD (Walsh *et al*., 2006; Hurst, 2009). Using these patterns, we found evidence suggestive of positive selection in two separate data sets for *ACD* and *TERF2IP*, and evidence of balancing selection in *POT1*, *NOLA2* and *TNKS*. Regions of low recombination, and thus long-range LD, as observed in *POT1* and regions of *TNKS*, could be the result of balancing selection; alleles under balancing selection can drag linked alleles with them and cause increased LD (Hurst, 2009). Two additional studies also found evidence of selection in *POT1*, *TNKS* and *TERF2IP*. *POT1* and *TNKS* were found to have significantly positive Tajima's D (Tajima, 1989) using sequence data (Savage *et al*., 2005), *POT1* in non-Hispanic Caucasians and *TNKS* in individuals of Pacific Rim ancestry, suggestive of balancing selection. *POT1* (in Europeans), *TERF2IP* and *TNKS* (both in East Asian and African populations) were also identified as candidate regions for recent selection with the powerful long-range haplotype and iHS tests in the HapMap genome-wide study based on over 3.1 million SNPs (The International HapMap Consortium, 2007).

Allele-frequency-based tests are not considered the most powerful methods to detect a recent selective sweep (Hanchard *et al*., 2006) and there is no statistical significance associated with these results. However, they highlight regions that might justify further investigation. There is also the possibility of SNP ascertainment bias that may result in false positive signals. Some of the genome-wide platform-selected SNPs are chosen based on their location in and around specific genes as well as based on haplotype-tagging SNPs in the region. However, we did not limit our analyses to only SNPs within gene exons, introns and UTRs because the goal of the study was to understand the gene and its surrounding region. The HGDP data were generated based on common SNPs and the HapMap data are also skewed toward common alleles making it more difficult to detect an excess of rare or derived alleles near fixation. However, the identification of these genes as candidates for selection in other studies suggests that selection may indeed be present. It has been (Barreiro *et al*., 2008) suggested that positive selection has ensured the regional adaptation of human populations by increasing population differentiation in gene regions, and that these loci likely contribute to disease-related phenotypic diversity among these different human populations.

We further explored genetic variation in *TERT* because several studies have identified both SNPs and mutations in *TERT* as important in cancer and telomere biology disorders (Armanios, 2009; Rafnar *et al*., 2009; Savage and Alter, 2009). Others have observed a high degree of *TERT* sequence similarity across species, hence we hypothe-

sized that there would be limited genetic variation in these populations. However, we observed high haplotype diversity and heterozygosity in *TERT*. The level of *TERT* differentiation among populations was average or lower than average (genome-wide average for autosomal SNPs), which may reflect a lack of LD and likely a high recombination rate in this region. The cancer-associated SNP, rs2736100, varied substantially in genotype frequency across major continental regions, which could correlate to varying disease risk.

In conclusion, this study suggests that, as a group, telomere biology genes have less diversity and differentiation than genes with different functions. Data suggest that *TERT* may be an exception to this hypothesis. The identification of telomere biology genes under selection (for example, *ACD, TERF2IP, POT1* and *TNKS*) might provide clues to their roles in telomere and chromosomal stability. It is possible that higher levels of genetic variation may not be tolerated in these genes, possibly due to their critical role in telomere maintenance.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This project has been funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, and with federal funds from the National Cancer Institute, National Institutes of Health, under contract number HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government. We thank the staff of the NCI core genotyping facility, Dr Stephen Chanock, NCI and Elliott Richards, NCI, for valuable assistance and helpful discussions.

References

Akey J, Zhang G, Zhang K, Jin L, Shriver M (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: 1805–1814.

Armanios M (2009). Syndromes of telomere shortening. *Annu Rev Genomics Hum Genet* **10**: 45–61.

Aubert G, Lansdorp P (2008). Telomeres and aging. *Physiol Rev* **88**: 557–579.

Bamshad M, Wooding S (2003). Signatures of natural selection in the human genome. *Nat Rev Genet* **4**: 99–111.

Barreiro L, Laval G, Quach H, Patin E, Quintana-Murci L (2008). Natural selection has driven population differentiation in modern humans. *Nat Genet* **40**: 340–345.

Barrett J, Fry B, Maller J, Daly M (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**: 263–265.

Blasco M (2007). Telomere length, stem cells and aging. *Nat Chem Biol* **3**: 640–649.

Blasco M, Lee H, Hande M, Samper E, Lansdorp PM, DePinho RA *et al.* (1997). Telomere shortening and tumor formation by mouse cells lacking telomerase RNA. *Cell* **91**: 25–34.

Cann H, de Toma C, Cazes L, Legrand M, Morel V, Piouffre L *et al.* (2002). A human genome diversity cell line panel. *Science* **296**: 261–262.

Collins K, Mitchell J (2002). Telomerase in the human organism. *Oncogene* **21**: 564–579.

Crabbe L, Verdun R, Haggblom C, Karlseder J (2004). Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. *Science* **306**: 1951–1953.

de Lange T (2004). T-loops and the origin of telomeres. *Nat Rev Mol Cell Biol* **5**: 323–329.

de Lange T (2005). Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev* **19**: 2100–2110.

Excoffier L, Laval G, Schneider S (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol Bioinformatics Online* **1**: 47–50.

Falush D, Stephens M, Pritchard J (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* **7**: 574–578.

Gilley D, Tanaka H, Herbert B (2005). Telomere dysfunction in aging and cancer. *Int J Biochem Cell Biol* **37**: 1000–1013.

Guthery S, Salisbury B, Pungliya M, Stephens J, Bamshad M (2007). The structure of common genetic variation in United States populations. *Am J Hum Genet* **81**: 1221–1231.

Hanchard N, Rockett K, Spencer C, Coop G, Pinder M, Jallow M *et al.* (2006). Screening for recently selected alleles by analysis of human haplotype similarity. *Am J Hum Genet* **78**: 153–159.

Hurst L (2009). Genetics and the understanding of selection. *Nat Rev Genet* **10**: 83–93.

Jakobsson M, Scholz S, Scheet P, Gibbs J, VanLiere J, Fung H *et al.* (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998–1003.

Jin G, Xu L, Shu Y, Tian T, Liang J, Xu Y *et al.* (2009). Common genetic variants on 5p15.33 contribute to risk of lung adenocarcinoma in a Chinese population. *Carcinogenesis* **30**: 987–990.

Kanoh J, Ishikawa F (2003). Composition and conservation of the telomeric complex. *Cell Mol Life Sci* **60**: 2295–2302.

Kullo I, Ding K (2007). Patterns of population differentiation of candidate genes for cardiovascular disease. *BMC Genetics* **8**: 48.

Lan Q, Shen M, Garcia-Rossi D, Chanock S, Zheng T, Berndt S *et al.* (2007). Genotype frequency and FST analysis of polymorphisms in immunoregulatory genes in Chinese and Caucasian populations. *Immunogenetics* **59**: 839–852.

Landi M, Chatterjee N, Yu K, Goldin L, Goldstein A, Rotunno M *et al.* (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* **85**: 679–691.

Li B, Oestreich S, de Lange T (2000). Identification of human Rap1: implications for telomere evolution. *Cell* **101**: 471–483.

Li J, Absher D, Tang H, Southwick A, Casto A, Ramachandran S *et al.* (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–1104.

McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G *et al.* (2008). Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**: 1404–1406.

Mirabello L, Huang W, Wong J, Chatterjee N, Reding D, Crawford E *et al.* (2009). The association between leukocyte telomere length and cigarette smoking, dietary and physical variables, and risk of prostate cancer. *Aging Cell* **8**: 405–413.

Moon I, Jarstfer M (2007). The human telomere and its relationship to human disease, therapy, and tissue engineering. *Front Biosci* **12**: 4595–4620.

Morlá M, Busquets X, Pons J, Sauleda J, MacNee W, Agustí A (2006). Telomere shortening in smokers with and without COPD. *Eur Respir J* **27**: 525–528.

Myles S, Davison D, Barrett J, Stoneking M, Timpson N (2008). Worldwide population differentiation at disease-associated SNPs. *BMC Med Genomics* **1**: 22.

Nakamura T, Cech T (1998). Reversing time: Origin of telomerase. *Cell* **92**: 587–590.

Nielsen R (2005). Molecular signatures of natural selection. *Annu Rev Genet* **39**: 197–218.

Petersen G, Amundadottir L, Fuchs C, Kraft P, Stolzenberg-Solomon R, Jacobs K *et al.* (2010). A genome-wide association

study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* **42**: 224–228.

Plentz R, Caselitz M, Bleck J, Gebel M, Flemming P, Kubicka S *et al.* (2004). Hepatocellular telomere shortening correlates with chromosomal instability and the development of human hepatoma. *Hepatology* **40**: 80–86.

Plentz R, Wiemann S, Flemming P, Meier P, Kubicka S, Kreipe H *et al.* (2003). Telomere shortening of epithelial cells characterises the adenoma-carcinoma transition of human colorectal cancer. *Gut* **52**: 1304–1307.

Pritchard J, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Prugnolle F, Manica A, Balloux F (2005). Geography predicts neutral genetic diversity of human populations. *Curr Biol* **15**: R159–R160.

Rafnar T, Sulem P, Stacey SN, Geller F, Gudmundsson J, Sigurdsson A *et al.* (2009). Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet* **41**: 221–227.

Ramachandran S, Deshpande O, Roseman C, Rosenberg N, Feldman M, Cavalli-Sforza L (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* **102**: 15942–15947.

Rodier F, Kim S, Nijjar T, Yaswen P, Campisi J (2005). Cancer and aging: the importance of telomeres in genome maintenance. *Int J Biochem Cell Biol* **37**: 977–990.

Rosenberg N (2006). Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann Hum Genet* **70**: 841–847.

Rosenberg N, Pritchard J, Weber J, Cann H, Kidd K, Zhivotovsky L *et al.* (2002). Genetic structure of human populations. *Science* **298**: 2381–2385.

Rudolph K, Chang S, Lee H, Blasco M, Gottlieb GJ, Greider C *et al.* (1999). Longevity, stress response, and cancer in aging telomerase-deficient mice. *Cell* **96**: 701–712.

Rudolph K, Millard M, Bosenberg M, DePinho R (2001). Telomere dysfunction and evolution of intestinal carcinoma in mice and humans. *Nat Genet* **28**: 155–159.

Sabeti P, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C *et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913–918.

Santos-Lopes SS, Pereira RW, Wilson IJ, Pena SDJ (2007). A worldwide phylogeography for the human X chromosome. *PLoS ONE* **2**: e557.

Savage S, Alter B (2009). Dyskeratosis congenita. *Hematol Oncol Clin North Am* **23**: 215–231.

Savage S, Stewart B, Eckert A, Kiley M, Liao J, Chanock S (2005). Genetic variation, nucleotide diversity, and linkage disequilibrium in seven telomere stability genes suggest that these genes may be under constraint. *Hum Mutat* **26**: 343–350.

Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B *et al.* (2009). Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* **41**: 899–904.

Shriver M, Kennedy G, Parra E, Lawson H, Sonpar V, Huang J *et al.* (2004). The genomic distribution of population substructure in four populations using 8525 autosomal SNPs. *Hum Genomics* **1**: 274–286.

Shriver M, Mei R, Parra E, Sonpar V, Halder I, Tishkoff S *et al.* (2005). Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Hum Genomics* **2**: 81–89.

Stacey S, Gudbjartsson D, Sulem P, Bergthorsson J, Kumar R, Thorleifsson G *et al.* (2008). Common variants on 1p36 and 1q42 are associated with cutaneous basal cell carcinoma but not with melanoma or pigmentation traits. *Nat Genet* **40**: 1313–1318.

Stacey S, Sulem P, Masson G, Gudjonsson S, Thorleifsson G, Jakobsdottir M *et al.* (2009). New common variants affecting susceptibility to basal cell carcinoma. *Nat Genet* **41**: 909–914.

Stephens M, Scheet P (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *Am J Hum Genet* **76**: 449–462.

Stephens M, Smith N, Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978–989.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**: 789–796.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* **437**: 1299–1320.

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–862.

Turnbull C, Rapley E, Seal S, Pernet D, Renwick A, Hughes D *et al.* (2010). Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nat Genet* **42**: 604–607.

Voight B, Kudaravalli S, Wen X, Pritchard J (2006). A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72.

von Zglinicki T (2002). Oxidative stress shortens telomeres. *Trends Biochem Sci* **27**: 339–344.

Vulliamy T, Beswick R, Kirwan M, Marrone A, Digweed M, Walne A *et al.* (2008). Mutations in the telomerase component NHP2 cause the premature ageing syndrome dyskeratosis congenita. *Proc Natl Acad Sci USA* **105**: 8073–8078.

Walsh E, Sabeti P, Hutcheson H, Fry B, Schaffner S, de Bakker P *et al.* (2006). Searching for signals of evolutionary selection in 168 genes related to immune function. *Hum Genet* **119**: 92–102.

Weir B, Cardon L, Anderson A, Nielsen D, Hill W (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Res* **15**: 1468–1476.

White T, Asfaw B, DeGusta D, Gilbert H, Richards G, Suwa G *et al.* (2003). Pleistocene omo sapiens from Middle Awash, thiopia. *Nature* **423**: 742–747.

Wong J, Collins K (2003). Telomere maintenance and disease. *Lancet* **362**: 983–988.

Wu X, Amos CI, Zhu Y, Zhao H, Grossman BH, Shay JW *et al.* (2003). Telomere dysfunction: a potential cancer predisposition factor. *J Natl Cancer Inst* **95**: 1211–1218.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)