RESEARCH ARTICLE

Characterization of population-based variation and putative functional elements for the multiple-cancer susceptibility loci at 5p15.33 [version 1; referees: 2 approved]

Lisa Mirabello[1], Charles C. Chung[1], Meredith Yeager[2], Sharon A Savage[1]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute,National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892, USA
[2]Cancer Genomics Research Laboratory, National Cancer Institute, Division of Cancer Epidemiology and Genetics, Leidos Biomedical Research, Inc., Frederick, MD 20877, USA

v1 **First published:** 02 Oct 2014, **3**:231 (doi: 10.12688/f1000research.5186.1)
Latest published: 02 Oct 2014, **3**:231 (doi: 10.12688/f1000research.5186.1)

Abstract

Background:
TERT encodes the telomerase reverse transcriptase, which is responsible for maintaining telomere ends by addition of $(TTAGGG)_n$ nucleotide repeats at the telomere. Recent genome-wide association studies have found common genetic variants at the *TERT-CLPTM1L* locus (5p15.33) associated with an increased risk of several cancers.

Results:
Data were acquired for 1627 variants in 1092 unrelated individuals from 14 populations within the 1000 Genomes Project. We assessed the population genetics of the 5p15.33 region, including recombination hotspots, diversity, heterozygosity, differentiation among populations, and potential functional impacts. There were significantly lower polymorphism rates, divergence, and heterozygosity for the coding variants, particularly for non-synonymous sites, compared with non-coding and silent changes. Many of the cancer-associated SNPs had differing genotype frequencies among ancestral groups and were associated with potential regulatory changes.

Conclusions:
Surrogate SNPs in linkage disequilibrium with the majority of cancer-associated SNPs were functional variants with a likely role in regulation of *TERT* and/or *CLPTM1L*. Our findings highlight several SNPs that future studies should prioritize for evaluation of functional consequences.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 02 Oct 2014	✓ report	✓ report

1 **Duncan Baird**, Cardiff University UK

2 **John L. Hopper**, University of Melbourne Australia, **Miroslav K. Kapuscinski**, University of Melbourne Australia

Discuss this article

Comments (0)

Corresponding author: Lisa Mirabello (mirabellol@mail.nih.gov)

How to cite this article: Mirabello L, Chung CC, Yeager M and Savage SA. **Characterization of population-based variation and putative functional elements for the multiple-cancer susceptibility loci at 5p15.33 [version 1; referees: 2 approved]** *F1000Research* 2014, **3**:231 (doi: 10.12688/f1000research.5186.1)

Grant information: 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 to M.Y. and C.C.C.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 02 Oct 2014, **3**:231 (doi: 10.12688/f1000research.5186.1)

Introduction

The 5p15.33 locus includes the *TERT* (human telomerase reverse transcriptase) and the *CLPTM1L* (alias CRR9; cleft lip and palate transmembrane 1 like) genes. Telomerase reverse transcriptase (TERT) is the essential catalytic component of the telomerase holoenzyme responsible for maintaining telomere ends. Telomerase compensates for DNA polymerase's inability to fully replicate the lagging DNA strand by adding hexanucleotide $(5'\text{-TTAGGG-}3')_n$ repeats to the 3' end of chromosomes using a template sequence within the RNA component (TERC) of the enzyme[1]. Telomeres, consisting of these hexanucleotide repeats and several associated proteins, are responsible for preserving chromosomal stability by protecting chromosomes from end-to-end fusion, atypical recombination, and degradation[2]. In normal differentiated cells, expression of telomerase is very low or absent and telomeres erode by 50 to 200 base pairs with each cell division[1]. When the telomeres become critically short, they act as a cellular clock and signal cellular senescence and apoptosis[3,4]. In contrast, telomerase activity has been detected in 90% of human cancers[5,6] and allows these malignant cells to continually divide by bypassing cellular crisis[7].

CLPTM1L is located approximately 23 kilobases (kb) centromeric of *TERT*. Little is known about the function of the *CLPTM1L* protein. It is a predicted transmembrane protein that is expressed in a range of normal and malignant tissues including skin, lung, breast, ovary and cervix, and has been shown to sensitize ovarian cancer cells to cisplatin-induced apoptosis[8].

The clinically related telomere biology disorders (TBDs), such as pulmonary fibrosis or aplastic anemia, are associated with germline mutations causing amino acid substitutions, additions, deletions, and frame shift mutations within *TERT*[9,10]. Patients with the more severe TBD, dyskeratosis congenita (DC) have very high risks of bone marrow failure and cancer, and have telomeres below the 1st percentile for their age[11]. DC represents the most clinically severe outcome of germline *TERT* mutations and often presents in childhood. Individuals with isolated aplastic anemia or pulmonary fibrosis due to *TERT* mutations tend to manifest clinical symptoms in adulthood.

Genome-wide association studies (GWAS) have found that common genetic variants, in the form of single nucleotide polymorphisms (SNPs), within the *TERT-CLPTM1L* locus (5p15.33) are associated with relatively low but highly statistically significant risks (odds ratios for risk alleles ranging between 1.05–1.6) of several cancers, including glioma[12,13], basal cell carcinoma[14,15], testicular[16], pancreatic[17], lung[18–20], bladder[21], colorectal[22], breast[23], and overall cancers[24] [reviewed in[25,26]].

Both *TERT* and *CLPTM1L* are evolutionarily conserved across diverse species, which suggests their functional importance[8,27,28]. *TERT* has low nucleotide diversity, and common SNPs in this gene region show low levels of differentiation among populations and high ancestral allele frequencies[28,29]; this pattern of low overall diversity suggests that *TERT* may be constrained[29].

The 1000 Genomes Project Consortium has reported that different populations have different profiles of rare and common variants; and, varying degrees of purifying selection at functionally relevant low-frequency sites which lead to substantial local population differentiation[30]. Large surveys of human genetic variation have described an excess of rare genetic variants as a result of a recent population expansion and weak purifying selection[31–33], particularly for variants in disease genes and for individuals of European ancestry[33].

In order to better understand the population genetics underlying the 5p13.3 locus associated with cancer, we conducted a detailed analysis of allele frequency patterns among ancestral group, levels of differentiation, and recombination at the 5p15.33 locus using 1000 Genomes Project[34] data. We retrieved data for the *TERT-CLPTM1L* genes and flanking regions for 1092 individuals from 14 populations. Analyses were focused on understanding how allele frequencies differ between populations, and evaluation of the cancer-associated SNPs and their surrogate markers for potential functional elements.

Materials and methods

Dataset

Data were retrieved for 1627 variants on 5p15.33 (hg19, chr5: 1,243,287–1,355,002) for all individuals in the 14 populations (1092 individuals) included in the 1000 Genomes project (2012 February release)[34]. Eighteen potentially related individuals were removed, which resulted in 1074 individuals. We also retrieved data for a flanking region, approximately 10kb upstream and downstream, in order to improve understanding of these gene regions [Data File 1].

Data analysis

The package ARLEQUIN version 3.5[35] was used to compute F_{ST} values, diversity, 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; and, these levels were compared to the genome-wide average for autosomal SNPs ($F_{ST} \approx 0.1$[36–39]). The population of African-Americans in the Southwestern United States (ASW) was grouped with the two populations of West African ancestry (Luhya in Kenya [LWK] and Yoruba in Nigeria [YRI]) since in our population level analyses they were found to be most closely related to these individuals of African ancestry, as previously observed[40]. In order to apportion the fraction of the genetic variance due to differences between and within ancestral groups (European, East Asian, West African, and American) and infer the genetic structure of the populations, AMOVA was performed with 10,000 permutations. HAPLOVIEW version 4.1[41] was used to determine the degree of linkage disequilibrium (LD) and minor allele frequency (MAF). The GLU genetics' ld.tagzilla module was used for the tag analysis with a LD pairwise r^2 threshold of 0.8. Pairwise LD was analyzed separately for the four ancestral groups and used to select tag SNPs for each region.

SNPs within *TERT* and *CLPTM1L* were grouped by functional category (*i.e.*, coding *vs.* non-coding, and synonymous *vs.* non-synonymous variants), and tested for significant differences in the normalized number of variant sites, allelic frequency divergence, heterozygosity, minor allele frequency (MAF), and levels of differentiation among populations; significant differences would suggest

that these functional categories of loci were not affected similarly, as expected under the assumption of neutrality. The allelic frequency divergence between ancestral groups was computed using: $d = 1\text{-}[(x_1y_1)^{1/2} + (x_2y_2)^{1/2}]$, where x_1 and y_1 are the frequencies of the first allele and x_2 and y_2 are the frequencies of the second allele[42]. The normalized number of variant sites was calculated as: $\theta\text{^} = K/\Sigma^{n-1}_{i=1}\ i^{-1}L$, where K is the number of variant sites, n is the number of chromosomes, and L is the total sequence length. Differences between the SNP functional categories were tested for significance with a two-tailed *t*-test. SIFT (**S**orts **I**ntolerant **F**rom **T**olerant) and Polyphen 2 (**Poly**morphism **Phen**otyping v**2**) were used to predict the potential impact of an amino acid substitution[43,44].

To identify recombination hotspots in this region, we used SequenceLDhot[45], a program that uses the approximate marginal likelihood method[46] and calculates likelihood ratio statistics at a set of possible hotspots. We used the four ancestral groups [European (EUR; n=379), East Asian (EA; n=286), American (AM; n=184), and African (AFR; n=246)] to calculate background recombination rates using PHASE v2.1[47,48]. The likelihood ratio statistics of 12 predicts the presence of a hotspot with a false-positive rate of 1 in 3,700 independent tests.

Putative functional elements were assessed using the UCSC genome browser (http://genome.ucsc.edu/), a publically available bioinformatics website, for ENCODE Regulation and Comparative Genomics tracks for all of the cancer-associated SNPs and their surrogates for each ancestral group. SNPs were considered surrogates for cancer-associated SNPs for each ancestral group if the $r^2 \geq 0.60$, the inter-marker distance ≤200kb, and the MAF ≥0.05. We assessed potential regions of open chromatin with DNase hypersensitivity; potential regulatory histone marks (H3K4Me1, H3K4Me3, H3K27Ac); protein binding sites; regulatory motifs; CpG islands; conserved mammalian microRNA regulatory binding sites; and evolutionary conservation among placental mammals using the phylop basewise conservation measurement[49]. Functional elements were also assessed using RegulomeDB, an integrated database that annotates SNPs with known or predicted regulatory DNA elements, including DNase hypersensitivity, transcription factor binging sites, and promoter regions that regulate transcription using data from GEO, ENCODE, and published literature[50]. RegulomeDB scores are a heuristic scoring system based on confidence that a variant is located in a functional region and likely results in a functional consequence, these are used to assist comparison among annotations[50]. Lower scores indicate increased evidence; category 2 scores are variants likely to affect binding, category 3 scores are less likely to affect binding; and 4, 5, or 6 scores are variants with minimal binding evidence.

Results

Dataset 1. Genotype data for 1627 variants on 5p15.33 (hg19, chr5: 1,243,287–1,355,002) for 1074 individuals from 14 populations

http://dx.doi.org/10.5256/f1000research.5186.d35521

Data were retrieved for 1627 variants on 5p15.33 (hg19, chr5: 1,243,287–1,355,002) for all individuals in the 14 populations (1092 individuals) included in the 1000 Genomes project (2012 February release). Eighteen potentially related individuals were removed, which resulted in 1074 individuals.

Allele frequency spectrum

There were 1627 variants in the *TERT-CLPTM1L* region among all individuals (N=1074): 167 were upstream of *TERT*, 563 in *TERT* (including UTR, intronic and exonic regions), 353 were between *TERT* and *CLPTM1L* (downstream of *TERT* and upstream of *CLPTM1L*), 412 in *CLPTM1L* (including UTR, intronic and exonic regions), and 132 downstream of *CLPTM1L*. A summary of the variation for the different functional categories of polymorphisms in *TERT* and *CLPTM1L* is given in Table 1. The majority of SNPs in *TERT* and *CLPTM1L* were in intronic regions (N=903), only 72 were exonic (49 in *TERT* and 18 in *CLPTM1L*). 46 of the exonic variants were synonymous changes (32 in *TERT* and 9 in *CLPTM1L*) and 26 were non-synonymous protein altering variants (PAV) (17 in *TERT* and 9 in *CLPTM1L*). The SNPs previously associated with cancer at 5p15.33[25] are all located in the intronic regions of *TERT* or *CLPTM1L* or intergenic between these genes, except for one which is a coding synonymous SNP in *TERT* (rs2736098; Table 2).

Since there were so few coding variants in the *TERT* and *CLPTM1L* loci, we combined them for the following analyses. The normalized number of variant sites, heterozygosity, and MAFs were significantly different by functional SNP category in *TERT* and *CLPTM1L* (*P* values <0.01; Table 1). Specifically, the non-coding SNPs (compared with coding SNPs) and synonymous SNPs (compared with non-synonymous SNPs) had significantly higher numbers of variant sites, heterozygosity, and MAFs (Table 1). These trends were consistent in all ancestral groups (Figure 1A). The most significant differences between coding and non-coding SNPs were in African populations (non-coding average MAF 9.8% *vs.* coding average

Table 1. Summary of variation for the different classes of polymorphisms for all individuals (n=1074).

Polymorphism type	bp screened	No. Polys	Frequency (SNP/bp)	θ^	Het.	MAF
Non-coding*	61,757	903	1/68	1.77E^{-03}	0.120	9.03%
Coding	7,126	72	1/99	1.22E^{-03}	0.036	2.14%
Synonymous		46	1/155	7.82E^{-04}	0.048	2.92%
Non-synonymous		26	1/274	4.42E^{-04}	0.014	0.69%

* includes intronic and 3' UTR SNPs; bp = base-pairs; Polys = polymorphisms; θ^ = normalized number of variant sites; Het. = heterozygosity; MAF = minor allele frequency; F_{ST} = level of differentiation among ancestral groups.

Table 2. Summary of the cancer-associated SNPs at the *TERT-CLPTM1L* locus.

SNP	Position	Gene	Function	Ethnicity†	Cancer(s)	Alleles‡	RAF				F_{ST}
							AFR	EUR	AM	EA	
rs4246742	1267356	*TERT*	intron	Misc.	Lung	T:A	67.4%	83.5%	77.7%	60.7%	0.055
rs10069690	1279790	*TERT*	intron	EUR, AFR	Breast	C:T	62.7%	27.5%	25.1%	15.9%	0.17
rs2242652	1280028	*TERT*	intron	EUR	Prostate	G:A	14.4%	21.0%	18.1%	16.4%	0.003
rs13167280	1280477	*TERT*	intron	EUR	Bladder	G:A	2.8%	13.0%	13.8%	19.1%	0.036
rs2736100	1286516	*TERT*	intron	Misc, EUR, Asian	Lung, CNS, Bladder, Pancreas, Testis	A:C	43.8%	50.0%	44.6%	39.3%	0.009
rs2853676	1288547	*TERT*	intron	Misc.	CNS, Lung	C:T	21.2%	27.5%	26.8%	16.1%	0.016
rs2736098	1294086	*TERT*	coding, syn.	Misc.	Bladder, Lung	C:T	6.0%	23.4%	19.5%	32.9%	0.062
rs2736108	1297488	Intergenic		EUR	Breast	C:T	6.7%	27.5%	22.3%	25.9%	0.045
rs2853668	1300025	Intergenic		EUR, Misc.	Pancreas, Lung, Colon	G:T	52.6%	25.8%	30.8%	24.3%	0.069
rs2735845	1300584	Intergenic		Misc.	Lung	C:G	4.9%	20.1%	24.9%	30.1%	0.055
rs4635969	1308552	Intergenic		Misc., EUR	Lung, Pancreas, Testis	G:A	34.1%	19.3%	12.7%	12.1%	0.055
rs4975615	1315343	Intergenic		Misc.	Lung	A:G	49.4%	42.3%	28.3%	16.3%	0.088
rs4975616	1315660	Intergenic		Misc., EUR	Lung, Pancreas, Testis	A:G	72.1%	44.3%	31.9%	16.3%	0.201
rs1801075	1317949	Intergenic	near gene 3'	Misc.	Lung	T:C	14.0%	19.1%	15.8%	4.4%	0.035
rs451360	1319680	*CLPTM1L*	intron	Misc., EUR	Lung	C:A	2.6%	21.6%	14.1%	11.9%	0.053
rs380286	1320247	*CLPTM1L*	intron	Misc.	Lung	G:A	61.6%	45.4%	35.6%	13.6%	0.156
rs402710	1320722	*CLPTM1L*	intron	Misc., EUR, Asian	Bladder, Lung	C:T	46.8%	35.5%	32.8%	29.4%	0.017
rs401681	1322087	*CLPTM1L*	intron	Misc, EUR, Asian	Bladder, Prostate, Pancreas, BCC, Melanoma, SCC, Lung	C:T	58.6%	45.9%	42.7%	30.4%	0.048
rs465498	1325803	*CLPTM1L*	intron	Misc, Asian	Lung	A:G	57.9%	46.2%	35.0%	16.4%	0.124
rs452932	1330253	*CLPTM1L*	intron	Misc.	Lung	T:C	58.2%	46.2%	35.6%	15.7%	0.128
rs452384	1330840	*CLPTM1L*	intron	Misc.	Lung	T:C	58.2%	45.9%	35.6%	15.7%	0.128
rs467095	1336221	*CLPTM1L*	intron	Misc.	Lung	T:C	71.2%	46.3%	35.9%	15.9%	0.194
rs31489	1342714	*CLPTM1L*	intron	Misc., EUR, Asian	Lung, Pancreas, Testis	C:A	47.2%	43.1%	31.4%	15.7%	0.084

† Ethnicity as reported in Mocellin *et al.* (2012); ‡ major allele:minor allele, and the risk allele is underlined; syn. = synonymous change; RAF = risk allele frequency; F_{ST} = level of differentiation among ancestral groups; misc. = miscellany, indicating a mix of different races; AFR = African ancestry; EUR = European ancestry; AM = American ancestry; EA = East Asian ancestry.

MAF 0.9%); and, the most significant differences between synonymous (syn.) *versus* non-synonymous (non-syn.) SNPs were in East Asian populations (syn. average MAF 4.8% *vs.* non-syn. average MAF 0.2%) (Figure 1A). There were significantly different levels of differentiation among ancestral groups for coding *versus* non-coding and synonymous *versus* non-synonymous SNPs (Figure 1B).

Protein altering variation

All PAVs were present at a rare or low frequency (Figure 1C). European ancestry individuals had higher MAFs for many of the PAVs in *TERT* and *CLPTM1L*, and there were significant MAF differences among ancestral groups for rs35719940, rs61748181, rs33955038, and rs113203740 (Figure 1C). Nine (53%) of the 17 PAVs observed in *TERT* and three (33%) of the nine PAVs observed in *CLPTM1L* were reported to be damaging by Polyphen and/or SIFT (two *in silico* approaches; underlined in Figure 1C). Most of these potentially damaging variants were only observed in one individual. However, three possibly damaging variants in *TERT* were observed in multiple individuals [rs34094720 (N=3), rs61748181 (N=31), rs200843534 (N=5)] (Figure 1C).

Patterns of diversity and recombination among ancestral groups

A summary of the variation by ancestral group for this region is given in Table 3. There was low nucleotide diversity (average of $5.0E^{-4}$) by ancestral group and low differentiation among ancestral

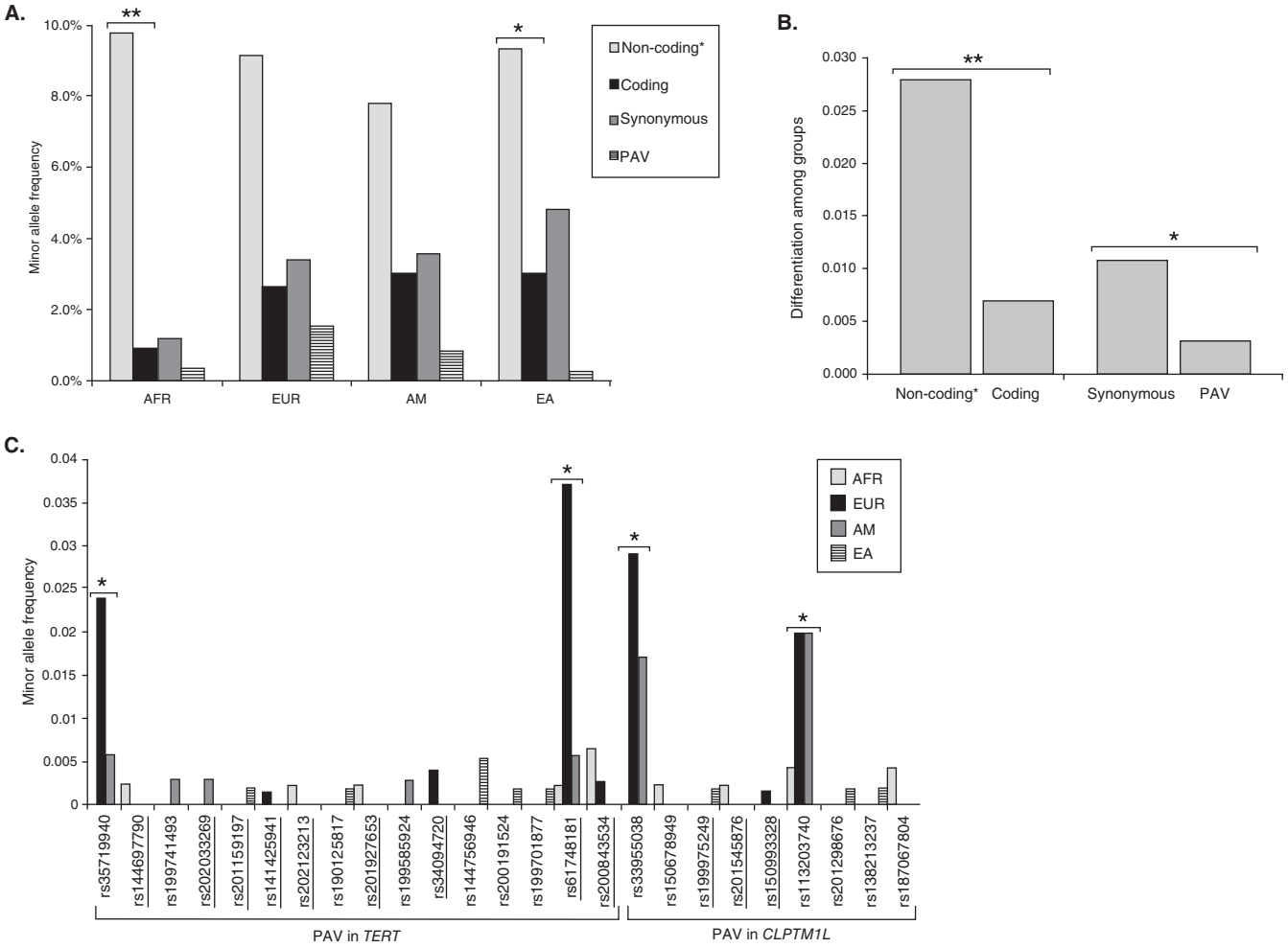

Figure 1. Variation in *TERT-CLPTM1L* by ancestral group. (**A.**) Average minor allele frequency of the polymorphisms by functional category for each group; (**B.**) average level of differentiation among ancestral groups (F_{ST}) for the polymorphisms by functional category; (**C.**) minor allele frequency of each protein-altering variant by ancestral group, the underlined variants are predicted to be potentially deleterious with SIFT and/or Poly-Phen. ** indicates a significant difference with a P <0.01, * P <0.05. PAV = non-synonymous protein-altering variation; AFR = African ancestry; EUR = European ancestry; AM = American ancestry; EA = East Asian ancestry.

Table 3. Summary of the diversity at 5p15.33 by ancestral group.

	African (AFR)	European (EUR)	American (AM)	East Asian (EA)
No. individuals	233	378	177	286
No. polymorphic loci	1009	732	808	503
Heterozygosity (SD)	0.120 (0.16)	0.127 (0.18)	0.111 (0.16)	0.129 (0.16)
Nucleotide diversity	6.5E^{-04}	5.0E^{-04}	4.9E^{-04}	3.8E^{-04}

SD = standard deviation.

groups (90.4% of loci in this region had low F_{ST} <0.10; median F_{ST} = 0.005) (data not shown). The median F_{ST} among ancestral groups (AG) and within populations (WP) for SNPs located within *TERT* and *CLPTM1L* were low (AG F_{ST} = 0.0039 and 0.0040, respectively; and, WP F_{ST} = 0.0078 and 0.0091, respectively). The greatest level of pairwise differentiation was among African and East Asian ancestry populations (pairwise F_{ST} = 0.208), and among European and East Asian ancestry populations (pairwise F_{ST} = 0.104) (Figure 2 and Supplementary Figure 1). The lowest level of pairwise differentiation was among European and American ancestry populations (pairwise F_{ST} = 0.01). The MAFs and heterozygosity estimates for SNPs in this region in European and American ancestry populations were highly correlated (r^2 = 0.95 and 0.965, respectively).

There was little to no LD in the *TERT* gene region but high LD was present in the *CLPTM1L* gene region (Figure 2 and Supplementary Figure 1). There were 4–5 main recombination hotspots in *TERT* and between *TERT* and *CLPTM1L*, there were no hotspots located within *CLPTM1L* (Supplementary Table 1). The greatest recombination was observed in individuals with African ancestry (5 recombination hotspots), and the lowest recombination in individuals with East Asian ancestry (4 recombination hotspots and lower likelihood ratio statistics) (Figure 2 and Supplementary Figure 1).

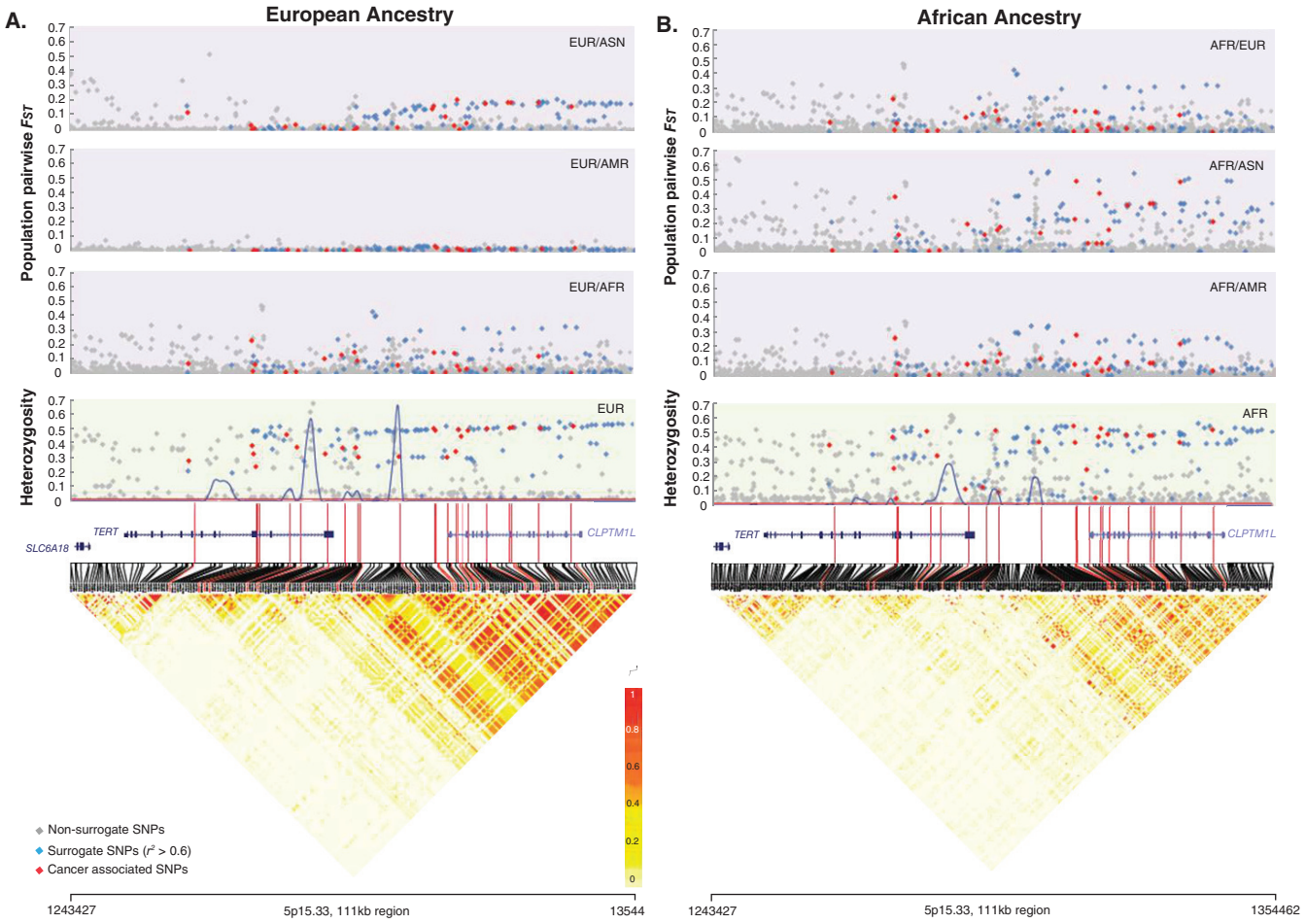

Figure 2. Summary of population genetics parameters in European (**A.**) and African (**B.**) ancestry individuals for 5p15.33. Linkage disequilibrium (LD), recombination hotspots, heterozygosity, and pairwise F_{st} values are shown for the cancer-associated SNPs (red dots), surrogate SNPs (blue dots), and non-surrogate SNPs (grey dots). LD pattern (see color legend) is shown for SNPs with a MAF ≥0.05. The red lines represent an extension of the location of the cancer-associated SNPs. The blue lines in the heterozygosity plot indicate the location of the recombination hotspots. For the pairwise F_{st} estimates, the populations are indicated in the top corner of each graph. AFR = African ancestry; EUR = European ancestry; AM = American ancestry; ASN = East Asian ancestry.

Cancer-associated SNPs

Twenty-three SNPs significantly associated with cancer at 5p15.33[25] were included in the analysis (Table 2). Many of the cancer associated SNPs in this region had differing allele frequencies and heterozygosity among ancestral groups and populations, and had F_{ST} values close to or greater than 0.1 (Table 2 and Supplementary Table 4). The risk allele was the rare allele at all of these SNPs, except at rs4246742 (associated with lung cancer; Table 2). Most of the cancer-associated SNPs in the *CLPTM1L* gene region are in regions of high LD, and therefore, have many surrogates (25–54 surrogate SNPs) with r^2 ≥0.6 (Table 4 and Supplementary Table 2). In contrast, most of the SNPs in the *TERT* gene region are in a region of low LD and have no or few surrogates (0–5 surrogate SNPs) with r^2 ≥0.6 (Table 4 and Supplementary Table 2). In East Asian ancestry individuals SNPs in the *CLPTM1L* gene region are particularly highly correlated, even some of the SNPs within *TERT* are in high LD in these individuals (*i.e.*, rs10069690, rs2242652, and rs13167280; Supplementary Figure 1).

Potential regulatory changes

All previously reported cancer-associated SNPs and all possible surrogates at r^2 ≥0.6 were assessed for the presence of potential regulatory elements and evolutionary conservation among mammalian species (summarized in Table 4 and Supplementary Table 3). Surprisingly, none of the cancer-associated SNP surrogates were located in the coding regions of *TERT* or *CLPTM1L*. Many of these SNPs are associated with open chromatin (DNase hypersensitivity) and/or regulatory histone marks (H3K4Me1, H3K4Me3, H3K27Ac) in multiple cell types, alter known regulatory motifs and/or protein binding sites. One of the surrogate SNPs in the putative promoter region of *TERT*, rs2853669, is a conserved binding site for POLR2A, as were six other surrogate SNPs located intergenic between *TERT* and *CLPTM1L*, within the *CLPTM1L* gene region, and in the putative promoter region of *CLPTM1L*. One of the cancer-associated SNPs, rs2736098, and three surrogate SNPs in the 5' region and putative promoter region of *TERT* were C>T SNPs located in the CpG island. Clusters of several surrogate SNPs

Table 4. Previously reported multiple-cancer susceptibility loci at 5q15.33 and their surrogates at an r^2 ≥0.6 and regulatory elements.

Locus			Surrogates†				H3K4 Me1	H3K4 Me3	H3K27 Ac	DNase	Regulatory motifs altered	Proteins bound	CpG island	Regulome DB score	Mammal Conserv.
			AFR	EUR	AM	EA									
rs4246742	1267356	*TERT*	0	0	1	1				• (3)				5	
rs10069690	1279790	*TERT*	2	1	0	2				• (19)				5	
rs2242652	1280028	*TERT*	3	1	1	1				• (17)	HEN1, ZFX, E2A, REST			5	
rs13167280	1280477	*TERT*	0	0	1	0				• (19)	NKX2			5	
rs2736100	1286516	*TERT*	3	0	8	9				• (4)				5	
rs2853676	1288547	*TERT*	0	0	1	1								5	
rs2736098	1294086	*TERT*	3	2	2	4	• (4)	• (4)		• (8)	NRSF, LRF		•	5	
rs2736108	1297488	Intergenic	3	2	3	3	• (3)			• (25)		EBF1		4	
rs2853668	1300025	Intergenic	0	0	1	1				• (2)				5	
rs2735845	1300584	Intergenic	0	2	2	3								—	•
rs4635969	1308552	Intergenic	13	4	3	45	• (2)		• (2)		FOXO1, SOX15			6	
rs4975615	1315343	Intergenic	24	48	48	54	• (8)	• (4)	• (4)		ZBTB3			5	
rs4975616	1315660	Intergenic	9	47	38	54	• (11)	• (5)	• (4)	• (8)				5	
rs1801075	1317949	Intergenic	2	6	6	0								—	
rs451360	1319680	*CLPTM1L*	0	7	4	52				• (4)	HIC1, OLF-1			5	
rs380286	1320247	*CLPTM1L*	18	47	47	47			• (3)	• (3)				5	
rs402710	1320722	*CLPTM1L*	20	8	0	0			• (3)		HEN1			5	
rs401681	1322087	*CLPTM1L*	25	46	21	0			• (3)	• (6)				5	
rs465498	1325803	*CLPTM1L*	27	47	46	54	• (3)		• (6)	• (9)				5	
rs452932	1330253	*CLPTM1L*	28	47	47	54	• (6)	• (5)	• (8)					6	
rs452384	1330840	*CLPTM1L*	28	47	47	54	• (5)	• (3)	• (7)	• (16)	MYC			5	
rs467095	1336221	*CLPTM1L*	8	47	46	54				• (2)		POLR2A, ETS1		4	
rs31489	1342714	*CLPTM1L*	31	47	47	54					MEF2			—	•

† r^2 ≥0.6, maximum inter-marker distance of 200kb and minimum MAF of 0.05;
AFR = African ancestry; EUR = European ancestry; AM = American ancestry; EA = East Asian ancestry;
Existence of a regulatory signature is indicated as dots (number of cell types this signature was observed, only indicated if occurring in ≥2 cell types);
RegulomeDB score indicates: 4 = TF binding + DNase peak, 5 = TF binding or DNase peak, 6 = motif hit, — = no data available;
Highlighted rows indicate that one or more surrogates for this SNP results in a likely functional consequence (RegulomeDB score of 2);
Mammal Conserv. = measurement of evolutionary placental mammal basewise conservation, the conserved sites are indicated.

F1000Research 2014, 3:231 Last updated: 05 OCT 2015

located within *CLPTM1L* and just 3' and 5' of *CLPTM1L* were associated with many histone marks and open chromatin, and/or altered regulatory motifs and protein binding sites. None of the cancer-associated SNPs or their surrogates were associated with microRNA binding sites.

We used the RegulomeDB scoring system to compare and prioritize potential functional consequences of these SNPs. The cancer-associated SNPs in the 5' region of *TERT*, most of the intergenic cancer-associated SNPs, and all the cancer-associated SNPs within *CLPTM1L* had surrogates with a likely functional consequence of affecting binding, indicated by a category 2 score (highlighted in Table 4 and Supplementary Table 3). None of the SNPs were identified to be associated with changes in expression of these genes.

Discussion

Data from the 1000 Genomes Project[34] on 1627 variants at 5p15.33 for 1074 unrelated individuals were used to describe the population genetic patterns in this region. We evaluated differentiation among ancestral groups, allele frequency patterns, and the cancer-associated SNPs and surrogates for potential regulatory elements. We have previously shown that there is low nucleotide diversity and differentiation among populations in *TERT* and suggested that *TERT* may be constrained[28,29]; however, our previous population genetics study focused on telomere genes as a gene set and was limited to only four SNPs located within the *TERT* gene[29]. In this study with better coverage of the *TERT-CLPTM1L* region, we determined that there is low nucleotide diversity across the 5p15.33 region in all ancestral groups and low differentiation among groups. As expected, African populations had more diversity, specifically at non-coding SNPs, compared to the other ancestral groups. However, East Asian populations had greater diversity at synonymous SNPs, and Europeans had the greatest frequency of non-synonymous changes. European and American ancestry individuals had very similar allele frequency patterns, as others have observed[51].

The significantly reduced normalized number of variant sites, heterozygosity, and MAFs, and low differentiation among ancestral groups for the coding sites, particularly for non-synonymous sites, compared with non-coding and silent changes suggests purifying selection in *TERT* and *CLPTM1*. African ancestry individuals had the greatest difference between the frequencies of non-coding *vs*. coding variants, consistent with stronger purifying selection; in contrast, European ancestry individuals had an excess of potentially deleterious non-synonymous SNPs. These observations are consistent with reports of genes important in cancer and complex disease[42,52–54] and recent genomic reports[30–33]. European ancestry individuals have been reported to have an excess of recently arisen potentially deleterious variants in disease genes[33]. American and East Asian ancestry individuals also had an excess of coding variants compared to African ancestry individuals, suggesting weaker purifying selection in these populations as well. East Asian individuals had a particular excess of synonymous variants and very few non-synonymous variants. For the cancer-associated SNPs in this region, the risk allele was primarily the rare allele which additionally provides support for the hypothesis of constraint in this region. This evidence of purifying selection supports the importance of *TERT* and *CLPTM1* in disease, and the variation by ancestry suggests the level of selection differs by geographic region.

We found that several of the 23 SNPs that have been significantly associated with cancer at 5p15.33 [Reviewed in 25] had differing MAFs and heterozygosity among ancestral groups. Europeans and Americans had the most similar MAFs and heterozygosity estimates, which suggests significant admixture. These differences, reflected in the high F_{ST} values, may correlate to varying disease incidence rates among ancestral groups. For example, the breast cancer associated SNP, rs10069690[23], had significantly different minor allele frequencies among ancestral groups; the homozygous risk allele genotype was significantly more common in African ancestry individuals (genotype frequency of 40% *vs*. 2.4% in East Asian, 6.8% in American, and 8.4% in European ancestry individuals) and less common in East Asian ancestry individuals. This difference may be associated with the higher incidence of breast cancer in African ancestry individuals (particularly for estrogen receptor-negative breast cancer) and lower incidence in East Asian individuals.

Many of the cancer-associated SNPs and surrogate SNPs were associated with potential regulatory elements, including histone marks, open chromatin, transcription factor binding sites, and/or regulatory motifs. There were only a few surrogates for the SNPs located within *TERT* and just 5' of *TERT* due to the low levels of LD in these regions; and, there were a large number of surrogates for the SNPs located close to and within *CLPTM1L* where LD was strong and recombination low, most of these surrogates were shared among the cancer-associated SNPs in this region. Many of the surrogate markers were located in the putative promoter regions of *TERT* and *CLPTM1L* and may affect gene regulation. The RegulomeDB scoring approach allowed us to classify variants based on all of the regulatory information. This approach determined that surrogate SNPs for many of the cancer-associated SNPs are functional variants with a likely role in regulation; these should be prioritized for functional assays.

Conclusions

Our analysis of diversity in this important cancer-associated region of 5p15.33 provides background information for understanding variation in the general population. The functional impact of common variation in this region needs to be examined experimentally, but we could speculate that the diversity of coding variants among different ethnicities could have mild effects on the phenotype disparity observed among these populations. Many of the cancer-associated SNPs and/or surrogates at 5p15.33 are associated with regulatory changes and candidates for evolutionary selection. Evidence of purifying selection in *TERT* and *CLPTM1L* highlights their functional importance and associations with complex disease. We have identified SNPs in this region that are likely involved in regulation of the *TERT* and/or *CLPTM1* genes. Future studies of the functional consequences of the 5p15.33 variants will be required to understand their contribution to cancer etiology.

Data availability

F1000Research: Dataset 1. Genotype data for 1627 variants on 5p15.33 (hg19, chr5: 1,243,287–1,355,002) for 1074 individuals from 14 populations, 10.5256/f1000research.5186.d35521[55]

Author contributions

Project design was carried out by S.A.S., L.M., and M.Y.

Genotyping data were retrieved by C.C.C.

Analyses were performed by L.M.

The manuscript was written by L.M. and S.A.S., and reviewed by all co-authors.

Competing interests

No competing interests were disclosed.

Grant information

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 to M.Y. and C.C.C.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary Table 1. Recombination hotspot inference summary. Click here to access the data.

Supplementary Table 2. All possible surrogate markers by ancestral group and their rank for the 23 cancer-associated SNPs based on a $R^2 \geq 0.60$, maximum inter-marker distance of 200kb, and minimum MAF of 0.05. Click here to access the data.

Supplementary Table 3. Previously reported multiple-cancer susceptibility loci at 5q15.33 and their surrogates at an $r^2 \geq 0.6$ and regulatory elements. Click here to access the data.

Supplementary Table 4. Risk allele frequencies of the cancer-associated SNPs at the *TERT-CLPTM1L* locus by population. Click here to access the data.

Supplementary Figure 1. Summary of population genetics parameters in East Asian (**A.**) and American (**B.**) ancestry individuals for 5p15.33. Click here to access the data.

Linkage disequilibrium (LD), recombination hotspots, heterozygosity, and pairwise *F*st values are shown for the cancer-associated SNPs (red dots), surrogate SNPs (blue dots), and non-surrogate SNPs (grey dots). LD pattern (see color legend) is shown for SNPs with a MAF ≥0.05. The red lines represent an extension of the location of the cancer-associated SNPs. The blue lines in the heterozygosity plot indicate the location of the recombination hotspots. For the pairwise *F*st estimates, the populations are indicated in the top corner of each graph. AFR = African ancestry; EUR = European ancestry; AM = American ancestry; ASN = East Asian ancestry.

References

1. Collins K, Mitchell JR: **Telomerase in the human organism.** *Oncogene.* 2002; **21**(4): 564–579.
PubMed Abstract | Publisher Full Text
2. Moon IK, Jarstfer MB: **The human telomere and its relationship to human disease, therapy, and tissue engineering.** *Front Biosci.* 2007; **12**: 4595–4620.
PubMed Abstract | Publisher Full Text
3. Gilley D, Tanaka H, Herbert BS: **Telomere dysfunction in aging and cancer.** *Int J Biochem Cell Biol.* 2005; **37**(5): 1000–13.
PubMed Abstract | Publisher Full Text
4. Maser RS, DePinho RA: **Connecting chromosomes, crisis, and cancer.** *Science.* 2002; **297**(5581): 565–569.
PubMed Abstract | Publisher Full Text
5. Shay JW, Bacchetti S: **A survey of telomerase activity in human cancer.** *Eur J Cancer.* 1997; **33**(5): 787–791.
PubMed Abstract | Publisher Full Text
6. Shay JW, Roninson IB: **Hallmarks of senescence in carcinogenesis and cancer therapy.** *Oncogene.* 2004; **23**(16): 2919–2933.
PubMed Abstract | Publisher Full Text
7. Gilley D, Tanaka H, Herbert BS: **Telomere dysfunction in aging and cancer.** *Int J Biochem Cell Biol.* 2005; **37**(5): 1000–1013.
PubMed Abstract | Publisher Full Text
8. Yamamoto K, Okamoto A, Isonishi S, *et al.*: **A novel gene, CRR9, which was up-regulated in CDDP-resistant ovarian tumor cell line, was associated with apoptosis.** *Biochem Biophys Res Commun.* 2001; **280**(4): 1148–1154.
PubMed Abstract | Publisher Full Text
9. Savage SA, Bertuch AA: **The genetics and clinical manifestations of telomere biology disorders.** *Genet Med.* 2010; **12**(12): 753–764.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Armanios M: **Syndromes of telomere shortening.** *Annu Rev Genomics Hum Genet.* 2009; **10**: 45–61.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Alter BP, Rosenberg PS, Giri N, *et al.*: **Telomere length is associated with disease severity and declines with age in dyskeratosis congenita.** *Haematologica.* 2012; **97**(3): 353–359.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Rajaraman P, Melin BS, Wang Z, *et al.*: **Genome-wide association study of glioma and meta-analysis.** *Hum Genet.* 2012; **131**(12): 1877–1888.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Shete S, Hosking FJ, Robertson LB, *et al.*: **Genome-wide association study identifies five susceptibility loci for glioma.** *Nat Genet.* 2009; **41**(8): 899–904.
PubMed Abstract | Publisher Full Text
14. Stacey S, Sulem P, Masson G, *et al.*: **New common variants affecting susceptibility to basal cell carcinoma.** *Nat Genet.* 2009; **41**(8): 909–914.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Stacey SN, Gudbjartsson DF, Sulem P, *et al.*: **Common variants on 1p36 and 1q42 are associated with cutaneous basal cell carcinoma but not with melanoma or pigmentation traits.** *Nat Genet.* 2008; **40**(11): 1313–1318.
PubMed Abstract | Publisher Full Text
16. Turnbull C, Rapley E, Seal S, *et al.*: UK Testicular Cancer Collaboration; **Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer.** *Nat Genet.* 2010; **42**(7): 604–607.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Petersen GM, Amundadottir L, Fuchs CS, *et al.*: **A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33.** *Nat Genet.* 2010; **42**(3): 224–228.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Jin G, Xu L, Shu Y, *et al.*: **Common genetic variants on 5p15.33 contribute to risk of lung adenocarcinoma in a Chinese population.** *Carcinogenesis.* 2009; **30**(6): 987–990.
PubMed Abstract | Publisher Full Text

19. McKay JD, Hung RJ, Gaborieau V, *et al.*: **Lung cancer susceptibility locus at 5p15.33.** *Nat Genet.* 2008; **40**(12): 1404–1406.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Landi MT, Chatterjee N, Yu K, *et al.*: **A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma.** *Am J Hum Genet.* 2009; **85**(5): 679–691.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Rothman N, Garcia-Closas M, Chatterjee N, *et al.*: **A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci.** *Nat Genet.* 2010; **42**(11): 978–984.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Peters U, Hutter CM, Hsu L, *et al.*: **Meta-analysis of new genome-wide association studies of colorectal cancer risk.** *Hum Genet.* 2012; **131**(2): 217–234.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Haiman CA, Chen GK, Vachon CM, *et al.*: **A common variant at the *TERT-CLPTM1L* locus is associated with estrogen receptor-negative breast cancer.** *Nat Genet.* 2011; **43**(12): 1210–1214.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Rafnar T, Sulem P, Stacey SN, *et al.*: **Sequence variants at the TERT-CLPTM1L locus associate with many cancer types.** *Nat Genet.* 2009; **41**(2): 221–227.
PubMed Abstract | Publisher Full Text

25. Mocellin S, Verdi D, Pooley KA, *et al.*: **Telomerase reverse transcriptase locus polymorphisms and cancer risk: a field synopsis and meta-analysis.** *J Natl Cancer Inst.* 2012; **104**(11): 840–854.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Chung CC, Chanock SJ: **Current status of genome-wide association studies in cancer.** *Hum Genet.* 2011; **130**(1): 59–78.
PubMed Abstract | Publisher Full Text

27. Nakamura TM, Cech TR: **Reversing time: origin of telomerase.** *Cell.* 1998; **92**(5): 587–590.
PubMed Abstract | Publisher Full Text

28. Savage S, Stewart B, Eckert A, *et al.*: **Genetic variation, nucleotide diversity, and linkage disequilibrium in seven telomere stability genes suggest that these genes may be under constraint.** *Hum Mutat.* 2005; **26**(4): 343–350.
PubMed Abstract | Publisher Full Text

29. Mirabello L, Yeager M, Chowdhury S, *et al.*: **Worldwide genetic structure in 37 genes important in telomere biology.** *Heredity (Edinb).* 2012; **108**(2): 124–33.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Abecasis GR, Auton A, Brooks LD, *et al.*: 1000 Genomes Project Consortium, **An integrated map of genetic variation from 1,092 human genomes.** *Nature.* 2012; **491**(7422): 56–65.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Keinan A, Clark AG: **Recent explosive human population growth has resulted in an excess of rare genetic variants.** *Science.* 2012; **336**(6082): 740–743.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Tennessen JA, Bigham AW, O'Connor TD, *et al.*: NHLBI Exome Sequencing Project: **Evolution and functional impact of rare coding variation from deep sequencing of human exomes.** *Science.* 2012; **337**(6090): 64–69.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Fu W, O'Connor TD, Jun G, *et al.*: NHLBI Exome Sequencing Project, Akey JM: **Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants.** *Nature.* 2013; **493**(7431): 216–220.
PubMed Abstract | Publisher Full Text | Free Full Text

34. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, *et al.*: **A map of human genome variation from population-scale sequencing.** *Nature.* 2010; **467**(7319): 1061–1073.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Excoffier L, Lischer HE: **Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.** *Mol Ecol Resour.* 2010; **10**(3): 564–567.
PubMed Abstract | Publisher Full Text

36. Shriver MD, Mei R, Parra EJ, *et al.*: **Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation.** *Hum Genomics.* 2005; **2**(2): 81–89.
PubMed Abstract | Free Full Text

37. Akey JM, Zhang G, Zhang K, *et al.*: **Interrogating a high-density SNP map for signatures of natural selection.** *Genome Res.* 2002; **12**(12): 1805–1814.
PubMed Abstract | Publisher Full Text | Free Full Text

38. Shriver MD, Kennedy GC, Parra EJ, *et al.*: **The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs.** *Hum Genomics.* 2004; **1**(4): 274–286.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Weir BS, Cardon LR, Anderson AD, *et al.*: **Measures of human population structure show heterogeneity among genomic regions.** *Genome Res.* 2005; **15**(11): 1468–1476.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Tian C, Hinds DA, Shigeta R, *et al.*: **A genomewide single-nucleotide–polymorphism panel with high ancestry information for African American admixture mapping.** *Am J Hum Genet.* 2006; **79**(4): 640–649.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Barrett JC, Fry B, Maller J, *et al.*: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics.* 2005; **21**(2): 263–265.
PubMed Abstract | Publisher Full Text

42. Hughes AL, Packer B, Welch R, *et al.*: **Effects of natural selection on interpopulation divergence at polymorphic sites in human protein-coding loci.** *Genetics.* 2005; **170**(3): 1181–1187.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Adzhubei I, Schmidt S, Peshkin L, *et al.*: **A method and server for predicting damaging missense mutations.** *Nat Methods.* 2010; **7**(4): 248–249.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Res.* 2001; **11**(5): 863–874.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Fearnhead P: **SequenceLDhot: detecting recombination hotspots.** *Bioinformatics.* 2006; **22**(24): 3061–3066.
PubMed Abstract | Publisher Full Text

46. Fearnhead P, Donnelly P: **Approximate likelihood methods for estimating local recombination rates.** *J Royal Statistical Society Series B-Statistical Methodology.* 2002; **64**(4): 657–680.
Publisher Full Text

47. Crawford DC, Bhangale T, Li N, *et al.*: **Evidence for substantial fine-scale variation in recombination rates across the human genome.** *Nat Genet.* 2004; **36**(7): 700–706.
PubMed Abstract | Publisher Full Text

48. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, *et al.*: **A map of human genome variation from population-scale sequencing.** *Nature.* 2010; **467**(7319): 1061–1073.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Cooper GM, Stone EA, Asimenos G, *et al.*: **Distribution and intensity of constraint in mammalian genomic sequence.** *Genome Res.* 2005; **15**(7): 901–913.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Boyle AP, Hong EL, Hariharan M, *et al.*: **Annotation of functional variation in personal genomes using RegulomeDB.** *Genome Res.* 2012; **22**(9): 1790–1797.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Wang QY, Song J, Gibbs RA, *et al.*: **Characterizing polymorphisms and allelic diversity of von Willebrand factor gene in the 1000 Genomes.** *J Thromb Haemost.* 2013; **11**(2): 261–269.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Hughes AL, Packer B, Welch R, *et al.*: **Widespread purifying selection at polymorphic sites in human protein-coding loci.** *Proc Natl Acad Sci U S A.* 2003; **100**(26): 15754–15757.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Freudenberg-Hua Y, Freudenberg J, Kluck N, *et al.*: **Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population.** *Genome Res.* 2003; **13**(10): 2271–2276.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Halushka MK, Fan JB, Bentley K, *et al.*: **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.** *Nat Genet.* 1999; **22**(3): 239–247.
PubMed Abstract | Publisher Full Text

55. Mirabello L, Chung CC, Yeager M, *et al.*: **Dataset 1. Genotype data for 1627 variants on 5p15.33 (hg19, chr5: 1,243,287–1,355,002) for 1074 individuals from 14 populations.** *F1000Research.* 2014.
Data Source

Open Peer Review

Current Referee Status:  

Version 1

Referee Report 16 June 2015

doi:10.5256/f1000research.5532.r9055

John L. Hopper[1], **Miroslav K. Kapuscinski**[2]
[1] Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Carlton, Vic, Australia
[2] Centre for Epidemiology and Biostatistics Melbourne School of Population and Global Health, University of Melbourne, Carlton, Vic, Australia

Mirabello *et al* present here a comprehensive bioinformatics investigation of genetic variation at the telomerase-containing locus (5p15.33) that has been associated with a range of malignancies. Given high biological plausibility of telomerase involvement in cancer pathology, this is an important study that could assist in further research on this putative susceptibility locus.

The research strategy described in this well written paper should be applauded as it can be easily applied to other genomic regions of interest and provides an excellent example of extracting more useful information from existing data. In particular, the use of 1000 Genomes data provides an opportunity to examine the distribution of a wider range of variants in detail not possible using GWAS genotyping alone.

As the authors point out, highly significant associations of a number of SNP variants are paralleled by rather small phenotypic associations with these variants. The most common protein altering variant (rs61748181) identified in the available data appears to have modest associations. This is not a unique situation and it makes choosing variants for functional characterization difficult considering the investment required for such comprehensive studies. It should be stressed that direct identification of causal variants from GWAS data has not been very successful. The present report demonstrates the need for well-designed analytical approach based on the sequence information (1000 Genomes) together with other data (ENCODE) to reveal credible causal candidates and narrow the choices for subsequent experimental verification. The authors acknowledge the key role of future functional work in this discovery process.

As the data from 1000 Genomes Consortium comes from unaffected people inclusion of other information in the analytical pipeline that allows comparison of germline and tumour sequence information (e.g. The Cancer Genome Atlas, eQTLs) might allow further refinement of variant evaluation with different mechanisms evident in different cancers (e.g. relevance of promoter mutations - Lindner *et al.,* 2015 and Spiegl-Kreinecker *et al.,* 2015).

The evidence for purifying selection in TERT-CLPTM1L region points to the importance of maintaining the structural integrity of this locus but also suggests that mechanisms other than protein altering mutations may play significant role such as interactions with other genes such as MYC (Koh *et al.*, 2015) or miR-34a

(Xu *et al.*, 2015).

The rationale for setting the threshold for marker surrogacy at $r^2 = 0.6$ (p7) while using $r^2 = 0.8$ for LD calculations (p3) should be explained.

In summary, this is well designed and presented study that demonstrates the potential of using high throughput sequencing data together with growing resources such as ENCODE to enhance understanding of traditional genome-wide genotyping experiments. The title reflects well the contents, the abstract is appropriate and occlusions are justified and balanced.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 10 October 2014

doi:10.5256/f1000research.5532.r6299

Duncan Baird
Institute of Cancer and Genetics, Cardiff University, Cardiff, UK

Numerous studies have identified variation at the TERT-CLPTM1L locus in conferring an increased risk of many different cancer types.

Here the authors have examined the genetic architecture of the TERT-CLPTM1L locus using sequence data from the 1000 genomes project. Given the potential significance of this locus, this type of work is important as it has the potential to identify functional variants that might not have been uncovered with the various GWAS undertaken to identify risk variants. Thus far none of the risk variants identified at this locus with GWAS results in non-synonymous protein changes, however this study provides data to indicate that some of these variants may be associated with regulatory sequences and chromatin marks. This study also identified 26 variants that result in non-synonymous protein changes in the hTERT or the CLPTM1L genes.

This is a well written manuscript and the conclusions are appropriately backed up by the data provided. The title is appropriate and the abstract adequately summarises the article. Overall this manuscript provides useful information that that will underpin future work to establish the importance of this locus in conferring cancer risk.

I have no major criticisms of this work; however I recommend that a more rigorous statistical review, than I am able to provide, is undertaken of this manuscript.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.