Nucleic Acids Research, 2006, Vol. 34, Database issue ***D617–D621***
doi:10.1093/nar/gkj151

SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes

Bernice R. Packer[1,2,*], Meredith Yeager[1,2], Laura Burdett[1,2], Robert Welch[1,2], Michael Beerman[1,2], Liqun Qi[1,2], Hugues Sicotte[1,2], Brian Staats[1,2], Mekhala Acharya[3], Andrew Crenshaw[1,2], Andrew Eckert[1,2], Vinita Puri[1,2], Daniela S. Gerhard[4] and Stephen J. Chanock[2,5]

[1]Intramural Research Support Program, SAIC-Frederick, NCI-FCRDC, Frederick, MD, USA, [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, [3]Department of Bioinformatics, George Mason University, Manassas, VA, USA, [4]Office of Cancer Genomics, National Cancer Institute, Bethesda, MD, USA and [5]Section on Genomic Variation, Pediatric Oncology Branch, National Cancer Institute, National Institutes of Health, Gaithersburg, MD, USA

Received September 14, 2005; Revised and Accepted October 28, 2005

ABSTRACT

The SNP500Cancer database provides sequence and genotype assay information for candidate SNPs useful in mapping complex diseases, such as cancer. The database is an integral component of the NCI Cancer Genome Anatomy Project (http://cgap.nci.nih.gov). SNP500Cancer reports sequence analysis of anonymized control DNA samples (n = 102 Coriell samples representing four self-described ethnic groups: African/African-American, Caucasian, Hispanic and Pacific Rim). The website is searchable by gene, chromosome, gene ontology pathway, dbSNP ID and SNP500Cancer SNP ID. As of October 2005, the database contains >13 400 SNPs, 9124 of which have been sequenced in the SNP500Cancer population. For each analysed SNP, gene location and >200 bp of surrounding annotated sequence (including nearby SNPs) are provided, with frequency information in total and per subpopulation as well as calculation of Hardy–Weinberg equilibrium for each subpopulation. The website provides the conditions for validated sequencing and genotyping assays, as well as genotype results for the 102 samples, in both viewable and downloadable formats. A subset of sequence validated SNPs with minor allele frequency >5% are entered into a high-throughput pipeline for genotyping analysis to determine concordance for the same 102 samples. In addition, the results of genotype analysis for select validated SNP assays (defined as 100% concordance between sequence analysis and genotype results) are posted for an additional 280 samples drawn from the Human Diversity Panel (HDP). SNP500Cancer provides an invaluable resource for investigators to select SNPs for analysis, design genotyping assays using validated sequence data, choose selected assays already validated on one or more genotyping platforms, and select reference standards for genotyping assays. The SNP500Cancer database is freely accessible via the web page at http://snp500cancer.nci.nih.gov.

INTRODUCTION

SNP500Cancer is a component of the Cancer Genome Anatomy Project (CGAP) of the National Cancer Institute (NCI) and is specifically designed to generate resources for the identification and characterization of genetic variation in genes important in cancer (1). The database reports the validation of SNPs by sequence analysis and optimizes genotyping assays for SNPs of interest to molecular epidemiology studies in cancer.

CGAP is dedicated to the development of technology, including both assays and utilization of technical platforms, and to determining the gene expression profiles of normal, precancer and cancer cells (2). Accordingly, data pertaining to genes and their variation are available on the public website http://cgap.nci.nih.gov. SNP500Cancer represents one of several initiatives designed to characterize sequence variation and

*To whom correspondence should be addressed. Tel: +1 301 496 6019; Fax: +1 301 402 3134; Email: packerb@mail.nih.gov

is a resource for studying common germ-line genetic variation in the etiology of different cancers as well as related phenotypes. A validated SNP in the database has 100% concordance between sequence analysis and genotyping results on one or more platforms. Assays are developed and optimized in the Core Genotyping Facility (CGF) of the NCI for associations studies conducted in the Division of Cancer Epidemiology and Genetics (DCEG) and the Center for Cancer Research within the Intramural Program of the NCI. The primary focus of the DCEG's intramural studies is to conduct population-based research on environmental and genetic determinants of cancer.

DNA SAMPLES

Sequence and genotype analysis are conducted in a set of 102 unique anonymized individuals of diverse geographic origin with self-described ethnic group affiliation information, chosen to represent four major ethnic groups in the USA; it includes 24 African/African-American, 31 Caucasian, 23 Hispanic and 24 Pacific Rim individuals. The 102 anonymized samples can be obtained from the Coriell Cell Repositories (Coriell Institute for Medical Research, Camden, NJ). The sets of individuals are not a random sampling of one or more human populations and thus the predictive value of the sequence and genotype data provided can vary for different population samples.

A subset of SNPs has been genotyped in 280 anonymized individuals drawn from the Human Diversity Panel (HDP) (3), using DNA that has been amplified by whole genome amplification (4). The ethnic distribution of the 280 individuals approximates the 102 of the SNP500Cancer set, namely, the four self-described ethnic groups listed here. It is notable that there are 49 Native Americans, 25 of Mayan background and 24 of Piman background.

SELECTION OF GENES AND SNPS

This database is biased towards SNPs that lie within or are situated close to 'candidate' genes. The selection of genes and SNPs for analysis has been drawn from the following sources: (i) genes that fit a plausible model for cancer studies (e.g. by pathway), (ii) review of the published literature on SNPs and cancer, (iii) SNPs reported in public databases with some associated non-*in silico* determined frequency and (iv) SNPs verified during re-sequence analysis of candidate genes (5). In addition, the database contains SNPs analysed in the Breast and Prostate Cohort Consortium (http://epi.grants.cancer.gov/BPC3/ and http://cgf.nci.nih.gov/cohort.cfm), which targets the genes of the sex steroid metabolism and insulin growth factor pathways ($n = 55$ in total).

As of October 2005, the database contains SNPs in 812 known genes. The average number of SNPs per gene is 14.1 and the median is 8 SNPs per gene; the range is between 1 and 237 SNPs per gene.

SEQUENCING PROTOCOL

A PCR amplicon of ~600 bp is generated for each SNP, which is localized to the center, creating flanking regions of ~300 bp in each direction. Additional putative SNPs (determined from dbSNP) are annotated on the sequence of the amplicon. Oligonucleotide primers are designed for bi-directional sequence analysis using Primer3 software (6) and ProbeITy primer design software (Celadon Laboratories, Hyattsville MD). Each oligonucleotide primer is extended with a universal sequencing primer, M13 forward (TGTAAAACGACGGCCAGT) or M13 reverse (CAGGAAACAGCTATGACC). The sequencing assay protocols and conditions are displayed on the SNP500Cancer website. Sequence tracings are analysed in Seqscape v2.5 (Applied Biosystems, Foster City, CA). A sequencing call is deemed acceptable if the Seqscape quality score is >20 for at least 98 of the 102 individuals. Genotype calls are determined for each of the 102 individuals. The genotype and allele frequencies are maintained in an Oracle database and displayed on the SNP500Cancer website.

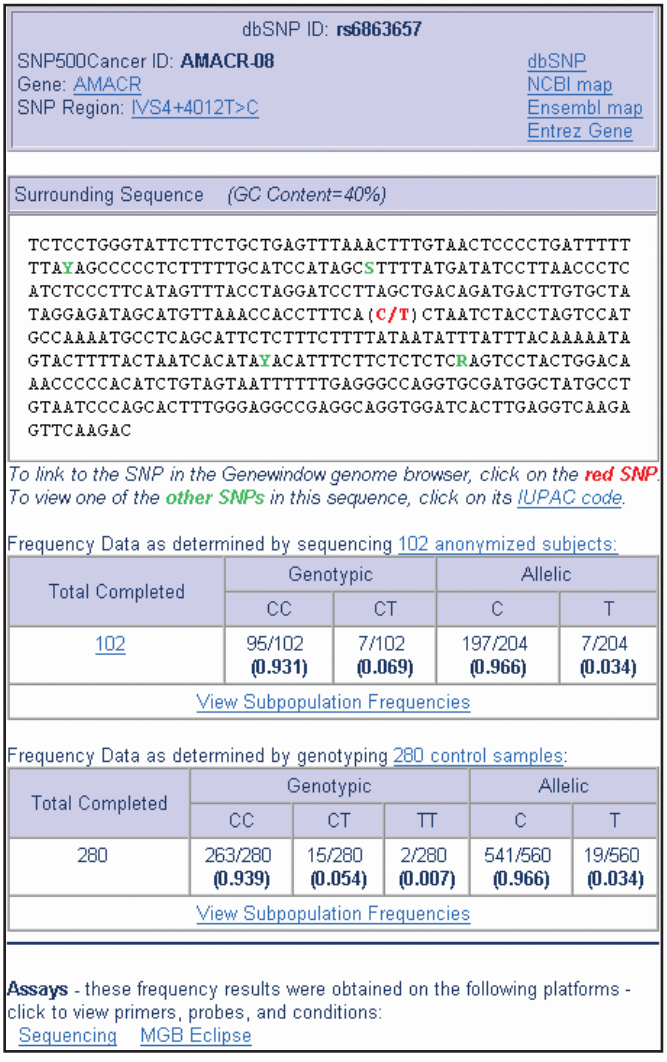
GENOTYPING PROTOCOLS AND VALIDATION

For SNPs that are determined to have >5% minor allele frequency (MAF) in at least one of the SNP500Cancer subpopulations, ~200 bp of DNA sequence surrounding each SNP is submitted for design on one of the CGF's genotyping platforms: (i) Applied Biosystems' TaqMan™ 'Assay by Design' service and (ii) EPOCH Biosciences' MGB Eclipse™ probes. The genotyping assay procedures and conditions are displayed on the SNP500Cancer website. Genotyping assays are validated if there is complete concordance between the genotype results and the primary sequence analysis for the 102 samples. In addition, the 280 HDP samples are genotyped for SNPs with >5% MAF. The frequencies found in the four HDP subpopulations are displayed on the SNP500Cancer website.

ANALYSIS OF ALLELE FREQUENCIES

For each validated SNP, allele and genotype frequencies are displayed for the 102 individuals overall (Figure 1) and for each SNP500Cancer subpopulation. The result of a test for Hardy–Weinberg equilibrium, χ^2 with one degree of freedom for two alleles (7) is posted for each subpopulation.

The observed allele frequencies in the SNP500Cancer population of $n = 102$ can provide a useful estimate overall, as well as for the four subgroups. In an analysis of 1164 SNPs genotyped in the additional 280 individuals compared with the data for the 102 which were both genotyped and sequenced, the coefficient of correlation (r^2) for each group was as follows: Caucasian 0.948, African-American 0.922, Pacific Rim 0.919 and Hispanic 0.742. The latter is notable because of the differences in the degree of admixture between the Hispanic subgroup of the 102 and the two Native American populations, of Maya and Pima heritage. Analysis of the SNP500Cancer dataset can be useful in estimating gene diversity. For instance, consistently reduced genetic diversity has been observed for SNP loci causing amino acid changes, especially radical shifts in protein structure (8); a similar trend was observed for 5′-untranslated regions (5′-UTRs). Moreover, the reduction of genetic diversity of nonsynonymous SNPs and those within the 5′-UTR are evidence that purifying selection could have acted at these sites, leading to reduced interpopulation divergence (9).



dbSNP ID: **rs6863657**

SNP500Cancer ID: **AMACR-08**
Gene: AMACR
SNP Region: IVS4+4012T>C

dbSNP
NCBI map
Ensembl map
Entrez Gene

Surrounding Sequence *(GC Content=40%)*

```
TCTCCTGGGTATTCTTCTGCTGAGTTTAAACTTTGTAACTCCCCTGATTTTT
TTAYAGCCCCCTCTTTTTGCATCCATAGCSTTTTATGATATCCTTAACCCTC
ATCTCCCTTCATAGTTTACCTAGGATCCTTAGCTGACAGATGACTTGTGCTA
TAGGAGATAGCATGTTAAACCACCTTTCA(C/T)CTAATCTACCTAGTCCAT
GCCAAAATGCCTCAGCATTCTCTTTCTTTTATAATATTTATTTACAAAAATA
GTACTTTTACTAATCACATAYACATTTCTTCTCTCTCRAGTCCTACTGGACA
AACCCCCACATCTGTAGTAATTTTTTGAGGGCCAGGTGCGATGGCTATGCCT
GTAATCCCAGCACTTTGGGAGGCCGAGGCAGGTGGATCACTTGAGGTCAAGA
GTTCAAGAC
```

To link to the SNP in the Genewindow genome browser, click on the ***red SNP****.*
To view one of the ***other SNPs*** *in this sequence, click on its IUPAC code.*

Frequency Data as determined by sequencing 102 anonymized subjects:

Total Completed	Genotypic		Allelic	
	CC	CT	C	T
102	95/102 **(0.931)**	7/102 **(0.069)**	197/204 **(0.966)**	7/204 **(0.034)**

View Subpopulation Frequencies

Frequency Data as determined by genotyping 280 control samples:

Total Completed	Genotypic			Allelic	
	CC	CT	TT	C	T
280	263/280 **(0.939)**	15/280 **(0.054)**	2/280 **(0.007)**	541/560 **(0.966)**	19/560 **(0.034)**

View Subpopulation Frequencies

Assays - these frequency results were obtained on the following platforms - click to view primers, probes, and conditions:
Sequencing MGB Eclipse

Figure 1. The SNP page for a typical SNP from the SNP500Cancer website, showing links to external resources, the SNP within flanking sequence, frequency data and links to assay information.

dbSNP SUBMISSION

All analysed SNPs from the SNP500Cancer project are submitted to dbSNP (10)—http://www.ncbi.nlm.nih.gov/SNP. This information includes flanking sequence, observed variation, assay primers, probes, and conditions, and frequency of the sequence variation among the SNP500Cancer total population and subpopulations. Figure 2 depicts the distribution of SNPs already in dbSNP and those newly observed in close proximity to target SNPs. It is also notable that ~10% of targeted SNPs are monoallelic in the 102 samples when compared with data in dbSNP build 124.

DATABASE AND WEB SERVER SPECIFICATIONS

The SNP500Cancer database is implemented using Oracle 9i (Oracle Corporation, Redwood Shores, CA). The web interface is written in ColdFusion (Macromedia, San Francisco, CA). The web server is a Compaq (Cupertino, CA) DL585 running Linux version 2.6.5. The database server is a Sun

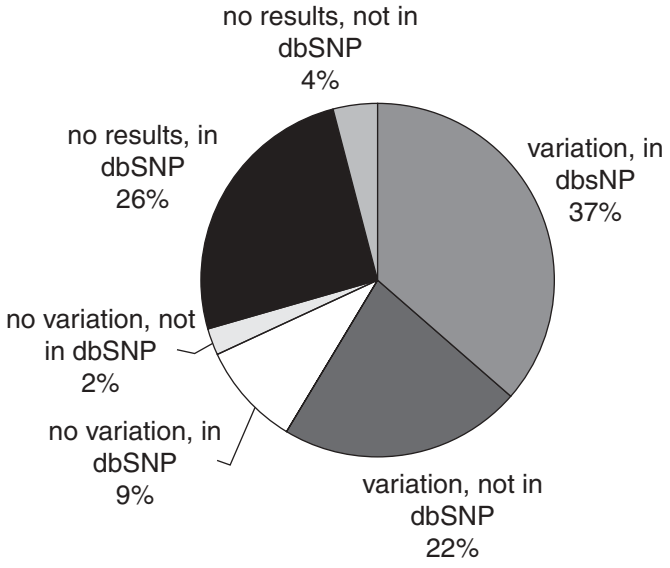


Figure 2. The distribution of 13.437 SNPs in SNP500Cancer—those already in dbSNP, those not yet in dbSNP (submitted for next build). Almost 10% of SNP500Cancer SNPs were found to have no variation in the set of 102 individuals of the SNP500Cancer population overall, even though they are in dbSNP.

Microsystems (Santa Clara, CA) Sunfire 4800 running Oracle version 9.2.0.6.0. Both servers are supported by NCI Computer Services (Rockville, MD).

SNP500Cancer data are available through the caGRID, part of the Cancer Biomedical Informatics Grid (caBIG), a common interface providing data sharing among cancer researchers worldwide (http://cabig.nci.nih.gov) (11).

USING THE SNP500CANCER WEBSITE

Searching for genes

The SNP500Cancer website provides capabilities for searching for genes in several ways: (i) gene name or alias (including wild card searches), (ii) chromosome location and (iii) Gene Ontology (GO) pathway (12)—numeric or text. The gene is displayed with a list of SNPs that have been validated, and those that were not found to occur in the SNP500Cancer population.

Searching for SNPs

SNPs can be searched for using the dbSNP ID (rs cluster number, e.g. rs799917), or the internal SNP500Cancer polymorphism ID (gene symbol followed by a sequence number, e.g. BRCA1-02). The SNP is displayed in the center of surrounding sequence, and other SNPs in the sequence are annotated with IUPAC codes (Figure 1). Clicking on the SNP variation links to Genewindow (13), a publicly available genome browser developed at the CGF.

Viewing genotypic and allelic frequencies

For the SNP of interest, genotypic and allelic frequencies for the entire SNP500Cancer population of 102 individuals are displayed. The 'view subpopulation frequencies' link displays

a page with genotypic and allelic frequencies for each subpopulation—African/African-American, Caucasian, Hispanic and Pacific Rim. Each subpopulation link leads to a list of individual genotypes for the samples within that subpopulation.

Viewing haplotype and htSNP data

For the gene of interest, haplotypes and haplotype-tagging SNPs (htSNP) data can be displayed for the entire control population or by subpopulation. Two different programs are available for determination of htSNPs: TagSNPs (14) and Tagger (http://www.broad.mit.edu/mpg/tagger/). The haplotype block structure can be visualized using Haploview (15).

Displaying assay conditions

For the SNP of interest, links are displayed for all validated assays (sequencing, MGB Eclipse™, TaqMan™). The link displays a page with detailed information on primers, probes, temperature and procedural steps.

Downloading information via FTP

FTP files are posted for SNPs by gene (including identifiers, genomic location and amino acid change), assays (identifier, primers, probes and conditions), frequencies and genotypes for each SNP with variation, as well as genotypes for all SNPs with >5% allelic frequency. These files can be found at ftp://ftp-snp500cancer.nci.nih.gov.

Connecting to other information sources

Each gene and SNP page on the SNP500Cancer website includes links to external resources; for genes: Entrez (16) and GO database; for SNPs: dbSNP, NCBI MapViewer (17) and Ensembl (18).

FUTURE DIRECTIONS

SNP500Cancer is a genetic resource designed to provide data on SNPs with neighboring sequence, for design and optimization of genotyping assays to be used in molecular epidemiology studies of cancer. A major goal of SNP500Cancer is to increase the breadth of genes and specifically develop sets of genes within a common pathway (e.g. DNA repair or telomere stability) (19). To this end, the project will utilize public resources to identify and validate SNPs across each gene at a greater density (e.g. one SNP every 1–3 kb) (20). The additional SNPs needed to densely cover the genes of interest will be selected based on their location, as well as suitability for inclusion as an htSNP in one or more population. The project will also add SNPs of great importance to cancer based on high quality citations in the published literature (21). There will be a special emphasis on SNPs drawn from coding and regulatory regions (including microRNAs) as well as those SNPs that have been shown to be associated with cancer susceptibility or outcome, particularly in pharmacogenomics.

SNPs with validated sequencing performed in other public projects will be imported, and assays optimized based on the concordance between sequence and genotype analysis. For instance, current plans are underway to optimize and validate assays for SNPs identified in the NIEHS Environmental Genome Project (NIEHS SNPs. NIEHS Environmental Genome Project, University of Washington, Seattle, WA, http://egp.gs.washington.edu/).

Though the primary goal of the SNP500Cancer database has been to focus on SNPs in and around known genes, future plans include the results of sequence analysis around small RNA sequences and their possible targets in cancer. Lastly, SNP500Cancer will make available results of dense, whole genome SNP scans using the 102 samples.

ACKNOWLEDGEMENTS

This research was supported (in part) by the Intramural Research Program of the National Cancer Institute, NIH, and Office of Cancer Genomics, NCI, NIH. Funding to pay the Open Access publication charges for this article was provided by the National Cancer Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Packer,B.R., Yeager,M., Staats,B., Welch,R., Crenshaw,A., Kiley,M., Eckert,A., Beerman,M., Miller,E., Bergen,A. *et al*. (2004) SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes. *Nucleic Acids Res*., **32**, 528–532.
2. Strausberg,R.L., Simpson,A.J.G. and Wooster,R. (2003) Sequence-based cancer genomics: progress, lessons, and opportunities. *Nat. Rev. Genet*., **4**, 409–418.
3. Cann,H.M., de Toma,C., Cazes,L., Legrand,M., Morel,V., Piouffre,L., Bodmer,J., Bodmer,W., Bonne-Tamir,B., Cambon-Thomsen,A. *et al*. (2002) A human genome diversity cell line panel. *Science*, **296**, 261–262.
4. Bergen,A.W., Haque,K.A., Qi,Y., Beerman,M.B., Garcia-Closas,M., Rothman,N. and Chanock,S.J. (2005) Comparison of yield and genotyping performance of multiple displacement amplification and Omniplex™ whole genome amplified DNA generated from multiple DNA sources. *Hum. Mutat*., **26**, 262–270.
5. Bernig,T., Taylor,J.G., Foster,C., Staats,B., Yeager,M. and Chanock,S. (2004) Sequence analysis of the mannose-binding lectin (MBL2) gene reveals a high degree of heterozygosity with evidence of selection. *Genes Immun*., **5**, 461–476.
6. Rozen,S. and Skaletsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.
7. Weir,B.S. (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetics Data*. Sinauer, Sunderland, MA.
8. Hughes,A.L., Packer,B.R., Welch,R., Bergen,A.W., Chanock,S.J. and Yeager,M. (2005) Effects of natural selection on inter-population divergence at polymorphic sites in human protein-coding loci. *Genetics*, **170**, 1181–1187.
9. Hughes,A.L., Packer,B.R., Welch,R., Bergen,A.W., Chanock,S.J. and Yeager,M. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc. Natl Acad. Sci. USA*, **100**, 15754–15757.
10. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
11. Buetow,K.H. (2005) Cyberinfrastructure: empowering a ''third way'' in biomedical research. *Science*, **308**, 821–824.
12. Gene Ontology Consortium (2000), Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.
13. Staats,B.S., Qi,L., Beerman,M., Sicotte,H., Burdett,L.A., Packer,B., Chanock,S.J. and Yeager,M. (2005) Genewindow: an interactive tool for visualization of genomic variation. *Nat. Genet*., **37**, 109–110.
14. Stram,D.O., Pearce,C.L., Bretsky,P., Freedman,M., Hirschhorn,J.N., Altshuler,D., Kolonel,L.N., Henderson,B.E. and Thomas,D.C. (2003) Modeling and E-M estimation of haplotype-specific relative risks from

genotype data for a case-control study of unrelated individuals. *Hum. Hered.*, **55**, 179–190.
15. Barrett,J.C., Fry,B., Maller,J. and Daly,M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
16. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
17. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
18. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
19. Savage,S.A., Stewart,B.J., Eckert,A., Kiley,M., Liao,J.S. and Chanock,S.J. (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.
20. Carlson,C.S., Eberle,M.A., Rieder,M.J., Yi,Q., Kruglyak,L. and Nickerson,D.A. (2004) Selecting a maximally informative set of SNPs for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.
21. International HapMap Consortium. (2005), The International HapMap Project. *Nature*, **437**, 1299–1320.