Published online 28 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D711–D715***
doi:10.1093/nar/gkl962

SNP@Ethnos: a database of ethnically variant single-nucleotide polymorphisms

Jungsun Park[1,*], Sohyun Hwang[1,2], Yong Seok Lee[1,3], Sang-Cheol Kim[4] and Doheon Lee[2]

[1]Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea, [2]Department of BioSystems, KAIST, Daejeon 305-701, Korea, [3]Department of Biology, Kyungpook National University, Taegu 702-701, Korea and [4]Department of Applied Statistics, Yonsei University, 134, Shinchon-dong, Seodaemoon-gu, Seoul, Korea

Received August 15, 2006; Revised October 24, 2006; Accepted October 25, 2006

ABSTRACT

Inherited genetic variation plays a critical but largely uncharacterized role in human differentiation. The completion of the International HapMap Project makes it possible to identify loci that may cause human differentiation. We have devised an approach to find such ethnically variant single-nucleotide polymorphisms (ESNPs) from the genotype profile of the populations included in the International HapMap database. We selected ESNPs using the nearest shrunken centroid method (NSCM), and performed multiple tests for genetic heterogeneity and frequency spectrum on genes having ESNPs. The function and disease association of the selected SNPs were also annotated. This resulted in the identification of 100 736 SNPs that appeared uniquely in each ethnic group. Of these SNPs, 1009 were within disease-associated genes, and 85 were predicted as damaging using the Sorting Intolerant From Tolerant system. This study resulted in the creation of the SNP@Ethnos database, which is designed to make this type of detailed genetic variation approach available to a wider range of researchers. SNP@Ethnos is a public database of ESNPs with annotation information that currently contains 100 736 ESNPs from 10 138 genes, and can be accessed at http://variome.net and http://bioportal.net/ or directly at http://bioportal.kobic.re.kr/SNPatETHNIC/.

INTRODUCTION

Identifying genetic variations that give rise to human differences is one of the most interesting issues in human evolution. Many of the related natural-selection and selective-sweep studies have produced interesting findings (1–3), and the completion of the International HapMap Project (4) has increased the popularity of this type of study. According to the HapMap reports, candidate loci in which selection has occurred can be identified using long-range haplotype testing (5). However, previous studies did not measure nearly fixed variations despite evidence that rare variants with low minor-allele frequencies also contribute to observed variations in complex human traits (6,7). Therefore, in order to identify ethnically variant single-nucleotide polymorphisms (ESNPs), we devised a new systematic approach based on the nearest shrunken centroid method (NSCM) (8) that is not affected by the minor-allele frequency.

The present study compared the genotype profiles of three ethnic groups: Yoruba in Ibadan, Nigeria (YRI), a combination of Japanese in Tokyo (JPT) and Han Chinese (CHB) in Beijing (CHB+JPT), and Utah residents with ancestry from northern and western Europe (CEU). The study identified 100 736 SNPs that could classify the ethnic groups based on the NSCM (8). Of those SNPs, 5515 were in well-known loci of natural selection (e.g. *Duffy* and *lactase* genes) and disease-associated genes. Using the Sorting Intolerant From Tolerant system (9), 85 coding nonsynonymous ethnically variant SNPs (ESNPs) were predicted as damaging, indicating that these SNPs may be highly relevant in disease research. This study resulted in the creation of the SNP@Ethnos database that contains genetic-variation information for use in human differentiation studies.

DATABASE CONSTRUCTION

Data source

The International HapMap Phase I release #16.c genotype dataset was downloaded from the project web site (http://www.hapmap.org/index.html.en). Unrelated individuals were selected for examination, comprising 60 CEU, 45 CHB, 44 JPT and 60 YRI samples. Our analysis involved the examination of 3 565 483 common SNPs.

*To whom correspondence should be addressed. Tel: +82 42 879 8531; Fax: +82 42 879 8519; Email: pj518@kribb.re.kr

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

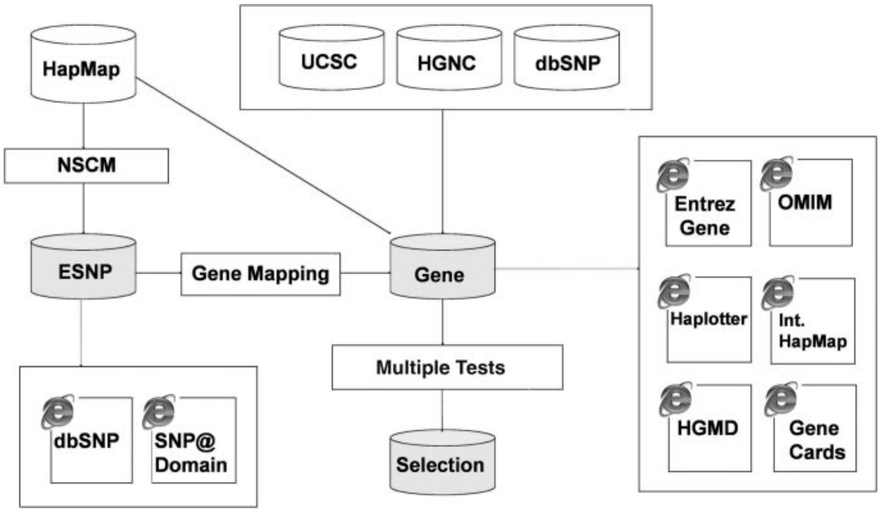


Figure 1. The data processing strategy for identifying ethnically variant SNPs and their functional annotations. Ethnically variant SNPs (ESNPs) were identified using the nearest shrunken centroid method (NSCM) of the R package pamr. Gene mapping was performed by combining three databases: the University of California, Santa Cruz, Genome Browser hg17, HUGO Gene Nomenclature Committee and dbSNP (build 125). Multiple tests were performed for genetic heterogeneity and frequency spectrum on genes having ESNPs. Links are provided for the following online databases: dbSNP, SNP@Domain, Entrez Gene, Online Mendelian Inheritance in Man (OMIM), Haplotter, International HapMap and Human Gene Mutation Database (HGMD).

Data processing

Pre-processing. Data pre-processing involved two steps: missing-allele imputation and the replacement of genotype features. We used the R package pamr, which does not allow for missing data and only allows numeric input data, so we had to impute the missing values and replace genotype features with numbers. For the missing-allele imputation, we replaced the missing values by the major allele of each ethnic group class (10): CEU, YRI and CHB+JPT. The proportion of missing values was 0.50%. For processing convenience, genotype features were coded using four numbers: (i) homo-reference allele; (ii) hetero allele; (iii) homo-other allele; and 0, missing value. The data processing is outlined in Figure 1.

ESNP selection. ESNPs were identified using the NSCM of the R package pamr. This method has been proposed as a suitable approach for solving the classification problem when there are a large number of features from which to predict classes and a relatively small number of cases, and it is important to identify which features contribute most to the classification (8).

A detailed mathematical explanation of NSCM is as follows. Let x_{ij} be the genotype for SNPs i $(= 1, 2, \ldots, p)$ and samples j $(= 1, 2, \ldots, n)$. We have classes $1, 2, \ldots, K$, and let C_k be indices of the n_k samples in class k. The i-th component of the centroid for class k is $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij}/n_k$, which gives the mean genotype value in class k for SNP i, and the i-th component of the overall centroid is $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$. In words, we shrink the class centroids toward the overall centroids after standardizing by the within-class SD for each SNP.

Let

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k(s_i + s_0)}, \qquad 1$$

where s_i is the pooled within-class SD for SNP i and $m_k = \sqrt{1/n_k + 1/n}$ makes $m_k \cdot s_i$ equal to the estimated standard error of the numerator in d_{ik}. In the denominator, s_0 is a positive constant equal to the median of the s_i values over the set of SNPs. Thus d_{ik} is a t statistic for SNP i that compares class k to the overall centroid. This method shrinks each d_{ik} toward zero, giving d'_{ik} and yielding shrunken centroids or prototypes

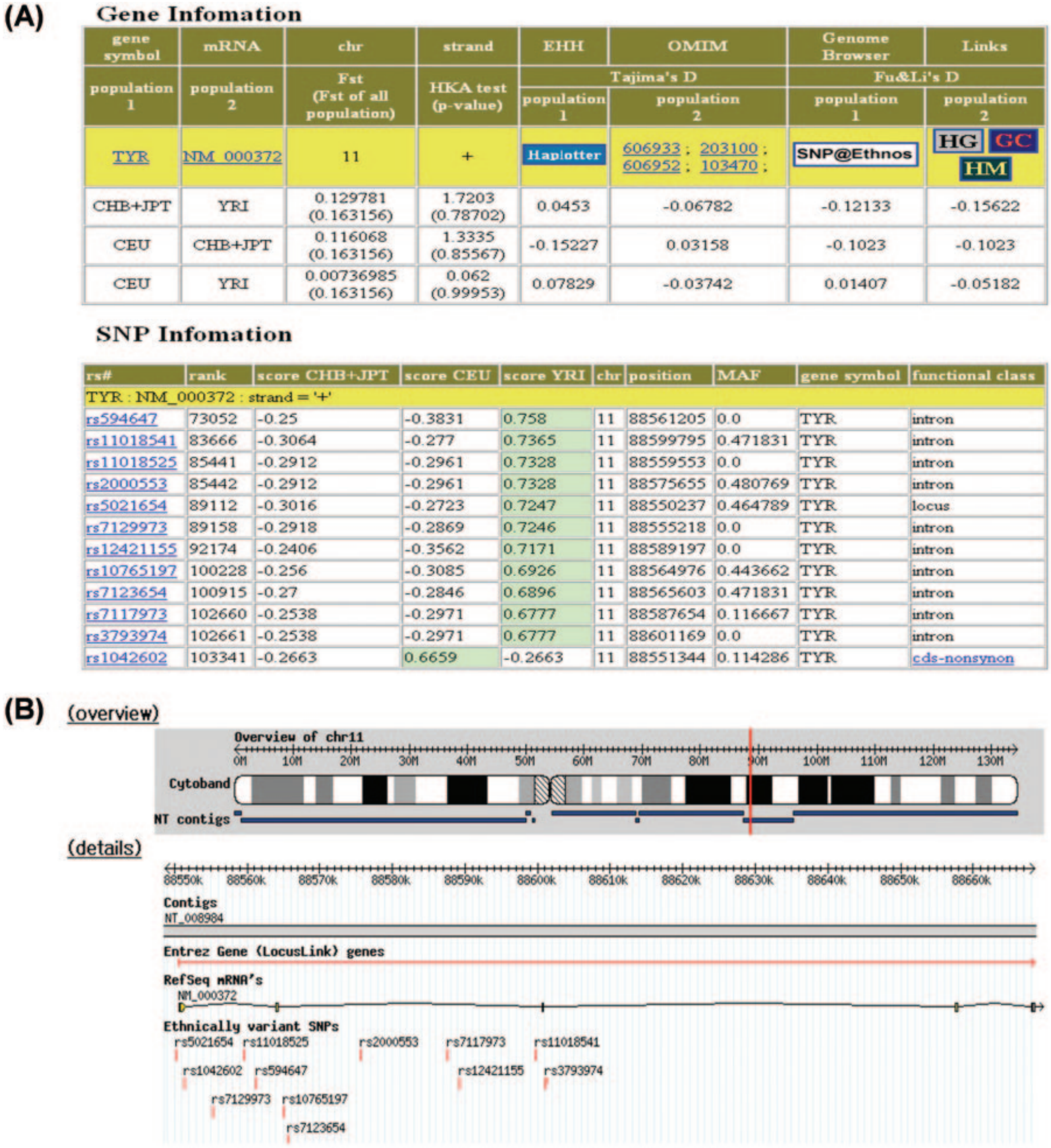
$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik}. \qquad 2$$

Specifically, if for a SNP i the value of d_{ik} is shrinks to zero for all classes k, then the centroid for SNP i is $\bar{x}_i$, and is the same for all classes. Thus SNP i does not contribute to the nearest-centroid computation.

As in the above explanation, the present study involved a large number of features from 1 007 376 SNPs and a relatively small number of classes (three ethnic groups). The use of standard statistical methods may cause problems in multiple comparisons because of the huge number of SNPs (11). For example, if 10 000 SNPs are discovered using those methods with a significance level of 5%, it is likely that 500 of them will be false-positive errors. NSCM has the desirable property that many of the SNPs that do not contribute to the nearest-centroid computation are eliminated from the class prediction.

Gene mapping and multiple analyses. Gene mapping was performed by combining three databases: the University of California, Santa Cruz, Genome Browser hg17 (12), HUGO Gene Nomenclature Committee (13) and dbSNP (build 125) (14).

SNP sequence files were constructed for all genes using the gene mapping information and International HapMap genotype data for tests for genetic heterogeneity and frequency spectrum. The following tests were performed: Hudson, Kreitman and Aguade (HKA) (15) and Fst (16) for genetic heterogeneity and Tajima's D (17), Fu and Li's D (18) for



Gene Infomation

gene symbol	mRNA	chr	strand	EHH	OMIM	Genome Browser	Links
				Tajima's D		Fu&Li's D	
population 1	population 2	Fst (Fst of all population)	HKA test (p-value)	population 1	population 2	population 1	population 2
TYR	NM_000372	11	+	Haplotter	606933 ; 203100 ; 606952 ; 103470 ;	SNP@Ethnos	HG GC HM
CHB+JPT	YRI	0.129781 (0.163156)	1.7203 (0.78702)	0.0453	-0.06782	-0.12133	-0.15622
CEU	CHB+JPT	0.116068 (0.163156)	1.3335 (0.85567)	-0.15227	0.03158	-0.1023	-0.1023
CEU	YRI	0.00736985 (0.163156)	0.062 (0.99953)	0.07829	-0.03742	0.01407	-0.05182

SNP Infomation

rs#	rank	score CHB+JPT	score CEU	score YRI	chr	position	MAF	gene symbol	functional class
TYR : NM_000372 : strand = '+'									
rs594647	73052	-0.25	-0.3831	0.758	11	88561205	0.0	TYR	intron
rs11018541	83666	-0.3064	-0.277	0.7365	11	88599795	0.471831	TYR	intron
rs11018525	85441	-0.2912	-0.2961	0.7328	11	88559553	0.0	TYR	intron
rs2000553	85442	-0.2912	-0.2961	0.7328	11	88575655	0.480769	TYR	intron
rs5021654	89112	-0.3016	-0.2723	0.7247	11	88550237	0.464789	TYR	locus
rs7129973	89158	-0.2918	-0.2869	0.7246	11	88555218	0.0	TYR	intron
rs12421155	92174	-0.2406	-0.3562	0.7171	11	88589197	0.0	TYR	intron
rs10765197	100228	-0.256	-0.3085	0.6926	11	88564976	0.443662	TYR	intron
rs7123654	100915	-0.27	-0.2846	0.6896	11	88565603	0.471831	TYR	intron
rs7117973	102660	-0.2538	-0.2971	0.6777	11	88587654	0.116667	TYR	intron
rs3793974	102661	-0.2538	-0.2971	0.6777	11	88601169	0.0	TYR	intron
rs1042602	103341	-0.2663	0.6659	-0.2663	11	88551344	0.114286	TYR	cds-nonsynon

Figure 2. Example results of a SNP@Ethnos database search. (**A**) The gene information in the search results consists of statistical values for the neutrality test and annotation links to the OMIM database and the HGMD and genome browser. The SNP information in the search results consists of the NSCM score, minor-allele frequency, chromosomal location and the functional annotation. (**B**) The genome browser of SNP@Ethnos shows the location of ESNPs.

frequency spectrum. The glutamate receptor and iduronate 2-sulfatase (19) genes were used as reference natural-selection loci. It should be noted that these statistics are affected by natural selection and by the frequency spectrum associated with demographic processes in a population (e.g. population expansion).

Database contents and availability

SNP@Ethnos provides functional information of ESNPs, with natural-selection and disease-association annotation of genes in which ESNPs are placed. The SNP information in the search results consists of the NSCM score, minor-allele frequency, chromosomal location and the functional annotation. The NSCM score (d'_{ik}) is a discriminating value from Equation 2, which is small if there is little difference between classes or the variation of the SNP distribution is large. For example, three similar scores for CHB+JPT, CEU and YRI indicate that the SNP is not critical, whereas one score differing from the other two indicates that the SNP is specific to that population. The functional annotation of SNPs provides a link to the SNP@Domain database (20) when the SNP is on the coding region of a protein. The gene information in the search results consists of the statistical results from the tests for genetic heterogeneity and frequency spectrum and annotation links to the Online Mendelian Inheritance in

Man (OMIM) database (21), and the Human Gene Mutation Database (22) and genome browser. The results of the multiple tests include Fst, HKA test, Tajima's D and Fu and Li's D values. Some general guidelines for interpreting the statistics of the tests are shown in FAQ page of our web site. Example results from database searching are shown in Figure 2.

The database contains 100 736 ESNPs and 10 138 annotated genes, where 1009 of the latter have OMIM entries, while 436 SNPs are in protein domains. Some of the SNPs are found in disease-associated genes and cause functional protein defects. There are many reports on ethnic variations in genes associated with disease (23–27). Using SNP@Ethnos, the present study identified an interesting ESNP in a tyrosinase gene associated with albinism; this nonsynonymous ESNP may cause a functional defect, but this has yet to be shown. SNP@Ethnos appears to be useful for this type of genetic variation investigation.

Data characteristics

The identified ESNPs comprised 73.95% YRI-specific, 15.25% CEU-specific and 6.80% CHB+JPT-specific SNPs and 4.00% ethnically different SNPs. All ESNPs were evenly distributed across the chromosomes. However, when the boundary of ESNPs was fixed using top 1%, there were three times more SNPs in the X chromosome (152) than on any other chromosome. These findings are consistent with those of other recent studies showing that the extent of population differentiation is similar across the autosomes, but higher in the X chromosome (Fst = 0.21), and that the number of low-frequency alleles is smaller for CEU and CHB + JPT samples than for YRI samples (5). These patterns may be attributable to bottlenecks in the history of the non-YRI populations (5).

The Gene Ontology (GO) class was analyzed after the genes of the 100 736 selected SNPs were annotated with the GO database (28) using the OntoExpress (29) and FatiGO (30) programs. Searches using both OntoExpress and FatiGO resulted in some of the genes being assigned to biological processes (cell communications and cellular physiological processes) and cellular-component (membranes) classes. Comparison of ESNPs with non-ESNPs using the FatiGO program revealed significant correlations with biological processes (responses to biotic stimulus, localization and responses to external stimulus) and cellular components (membranes, voltage-gated calcium channel complexes and the extracellular matrix), providing population-specific adaptive polymorphisms. Detailed results and their probability values are given on the Statistics page of our web site.

There were 82 SNPs which could be used to perfectly classify the ethnic groups, of which three were nonsynonymous coding-region SNPs (3.7%), which is a very high percentage compared with the original SNP functional distribution (0.84%). This suggests that the ESNPs play an important role in protein function.

DATA ACCESS AND VISUALIZATION

The SNP@Ethnos database can be queried using gene symbols, RefSeq mRNA IDs, dbSNP rs numbers and lists containing multiple genes. Regardless of the query type, the results are displayed in the same format. For visualization, SNP@Ethnos offers a generic genome browser (31) that displays an overview of chromosomes, contigs, genes, mRNAs and ESNPs. This genome browser can be accessed via the gene region of the results page. Moreover, the database provides an open-architecture web page using a wiki interface for data access. A user id for accessing the system is available from the authors on request. After logging in, users can submit comments and feedback. The content of web pages can be edited by users who wish to contribute, correct or add information.

ACKNOWLEDGEMENTS

We thank Jong Bhak for editing the manuscript. This project was supported by the Korean Ministry of Science and Technology under grant number M10407010001-04N0701-00110 and by the MIC (Ministry of Information and Communication), Korea, under the KADO (Korea Agency Digital Opportunity and Promotion) support program (6-121). D.L. was supported by the National Research Laboratory Grant (2005-01450). Funding to pay the Open Access publication charges for this article was provided by KADO support program.

Conflict of interest statement. None declared.

REFERENCES

1. Carlson,C.S., Thomas,D.J., Eberle,M.A., Swanson,J.E., Livingston,R.J., Rieder,M.J. and Nickerson,D.A. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.*, **15**, 1553–1565.
2. Nielsen,R., Williamson,S., Kim,Y., Hubisz,M.J., Clark,A.G. and Bustamante,C. (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.*, **15**, 1566–1575.
3. Voight,B.F., Kudaravalli,S., Wen,X. and Pritchard,J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol.*, **4**, e72.
4. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
5. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
6. Hugot,J.P., Chamaillard,M., Zouali,H., Lesage,S., Cezard,J.P., Belaiche,J., Almer,S., Tysk,C., O'Morain,C.A., Gassull,M. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
7. Roses,A.D. (1997) A model for susceptibility polymorphisms for complex diseases: apolipoprotein E and Alzheimer disease. *Neurogenetics*, **1**, 3–11.
8. Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.
9. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
10. Kantardzic,M. (2003) *Data Mining Concepts, Models, Methods and Algorithms.* 1st edn. Wiley-Interscience, Hoboken.
11. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
12. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
13. Povey,S., Lovering,R., Bruford,E., Wright,M., Lush,M. and Wain,H. (2001) The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet.*, **109**, 678–680.
14. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.

15. Hudson,R.R., Kreitman,M. and Aguade,M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**, 153–159.
16. Wright,S. (1950) Genetical structure of populations. *Nature*, **166**, 247–249.
17. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
18. Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.
19. Sabeti,P.C., Reich,D.E., Higgins,J.M., Levine,H.Z., Richter,D.J., Schaffner,S.F., Gabriel,S.B., Platko,J.V., Patterson,N.J., McDonald,G.J. *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
20. Han,A., Kang,H.J., Cho,Y., Lee,S., Kim,Y.J. and Gong,S. (2006) SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. *Nucleic Acids Res*., **34**, W642–W644.
21. Hamosh,A., Scott,A.F., Amberger,J., Valle,D. and McKusick,V.A. (2000) Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat*., **15**, 57–61.
22. Cooper,D.N. and Krawczak,M. (1996) Human Gene Mutation Database. *Hum. Genet*., **98**, 629.
23. Berggren,P., Kumar,R., Steineck,G., Ichiba,M. and Hemminki,K. (2001) Ethnic variation in genotype frequencies of a p53 intron 7 polymorphism. *Mutagenesis*, **16**, 475–478.
24. Grassi,M.A., Fingert,J.H., Scheetz,T.E., Roos,B.R., Ritch,R., West,S.K., Kawase,K., Shire,A.M., Mullins,R.F. and Stone,E.M. (2006) Ethnic variation in AMD-associated complement factor H polymorphism p.Tyr402His. *Hum. Mutat*., **27**, 921–925.
25. Lohmueller,K.E., Wong,L.J., Mauney,M.M., Jiang,L., Felder,R.A., Jose,P.A. and Williams,S.M. (2006) Patterns of genetic variation in the hypertension candidate gene GRK4: ethnic variation and haplotype structure. *Ann. Hum. Genet*., **70**, 27–41.
26. Singh,P., Singh,M. and Mastana,S.S. (2002) Genetics of apolipoprotein H (beta2-glycoprotein I) polymorphism in India. *Ann. Hum. Biol*., **29**, 247–255.
27. Torkildsen,O., Utsi,E., Mellgren,S.I., Harbo,H.F., Vedeler,C.A. and Myhr,K.M. (2005) Ethnic variation of Fc gamma receptor polymorphism in Sami and Norwegian populations. *Immunology*, **115**, 416–421.
28. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
29. Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
30. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
31. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.