Published online 13 December 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D685–D689***
doi:10.1093/nar/gkl848

D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples

Koichiro Higasa, Katsuyuki Miyatake, Yoji Kukita, Tomoko Tahira and Kenshi Hayashi*

Division of Genome Analysis, Research Center for Genetic Information, Medical Institute of Bioregulation, Kyushu University, Fukuoka, Fukuoka 812-8582, Japan

Received August 15, 2006; Revised October 8, 2006; Accepted October 9, 2006

ABSTRACT

The Definitive Haplotype Database (D-HaploDB) is a web-accessible resource of genome-wide definitive haplotypes determined from a collection of Japanese complete hydatidiform moles (CHMs), each of which carries a genome derived from a single sperm. Currently, the database contains genotypes for 281 439 common SNPs from 74 CHMs which were determined by a high-throughput array-based oligonucleotide hybridization technique. The database also presents maps of haplotype blocks and linkage disequilibrium bins together with tagSNPs that might prove useful for association studies of disease genes. Cryptic relatedness among the samples in this study is unlikely, because the formation of a CHM is a maternal event of rare sporadic occurrence, and its genotype is that of the incoming sperm. This is demonstrated by the absence of long extended shared haplotypes (ESHs). The D-HaploDB is freely accessible via the Internet at http://orca.gen.kyushu-u.ac.jp

INTRODUCTION

There is a great interest in elucidating the relationship between common genetic variation and heritable risk for common diseases. Case-control association studies have been extensively carried out to identify the genetic variants that contribute to such human diseases by employing SNPs as markers. Many SNPs show correlated genotypes (haplotypes) because of their shared evolutionary history, and so, only a subset of SNPs, also known as tagSNPs (1), need to be genotyped in order to detect the haplotype that bears the disease mutation.

The International HapMap Project (IHMP) (2) has determined fundamental new information about the haplotype structures of four populations, namely, African (YRI), Caucasian (CEU), Chinese (CHB) and Japanese (JPT). However, in the IHMP, the haplotype structures were inferred computationally using genotypes obtained from diploid starting materials. Although various algorithms have been developed to estimate haplotypes from diploid genotype data, errors in the inference process remain unresolved (3). Especially, the haplotypes of Asians were inferred without family data, and were less accurate than those for Europeans or Africans, which were determined using trio data (genotype data of parents and a child).

Complete hydatidiform moles (CHMs) offer a unique opportunity to determine definitive haplotypes at a genome-wide level (4,5). The haplotypes are definitive because they are directly determined using haploid genomes, in contrast to those in the other databases, e.g. the HapMap database, in which the haplotypes are computationally inferred from diploid data. The D-HaploDB presented here contains a genome-wide map of haplotypes which was obtained by genotyping 281 439 common SNPs in 74 CHM samples that were collected throughout Japan. This project was carried out independent of the IHMP (2).

DATA SOURCE

CHM samples were collected on a nationwide scale, and the effort was supported by the Japan Association of Obstetricians & Gynecologists. Both the female donors of the CHM tissues and their male partners were Japanese. Use of these samples in the present work was approved by the Ethical Committee of Kyushu University. The purity of the CHM DNA samples was confirmed by genotyping with 17 highly variable microsatellite markers (6). These high-quality CHM samples served as materials for high-throughput genotyping using high-density oligonucleotide arrays. The SNPs have been chosen to represent all of the linkage disequilibrium (LD) bins in both the European and Han Chinese populations detected in a previous study (7). Appropriate

*To whom correspondence should be addressed. Tel: +81 92 642 6171; Fax: +81 92 632 2375; E-mail: khayashi@gen.kyushu-u.ac.jp

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

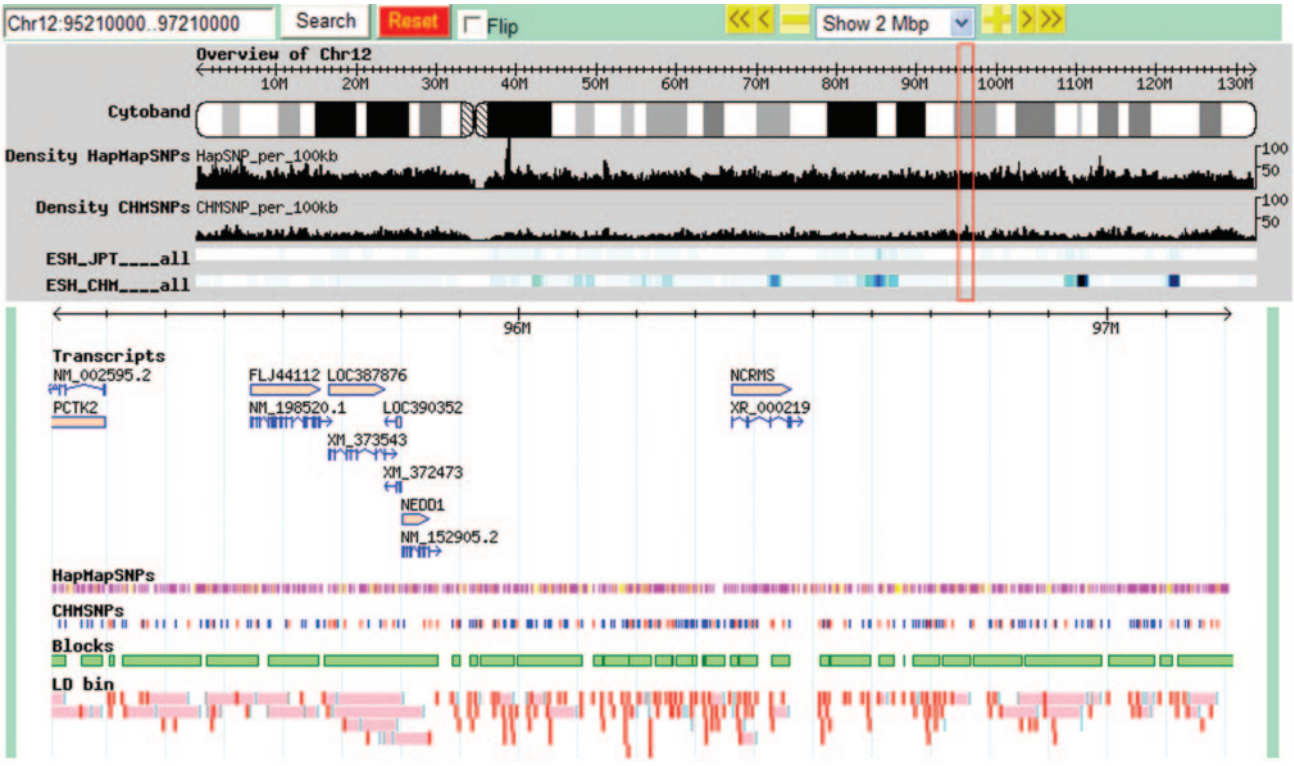

Figure 1. An example of the D-HaploDB web page. The region from 95 to 97 Mb of human chromosome 12 is shown.

filters were applied to control the genotyping quality as described previously (6). In brief, we initially typed 75 CHM samples using 358 550 SNPs. We then removed SNPs with a call rate of <80% and 'singleton' SNPs. The overall number of SNPs passed through these quality filters was 281 561. Of these SNPs, 281 439 (CHMSNPs) were mapped on the human reference sequence (NCBI Build 35). The data of one CHM sample were removed as its call rate (of SNPs) was low (71.6%). For the remaining 74 CHMs, the lowest and average call rates were 92.9 and 98.2% of the SNPs, respectively. We also genotyped 10 CHMs using an independent platform, the Affymetrix 100K Array, to evaluate the quality of the genotype data. The concordance rate between the two platforms was 99.91%. The average and the median distance between CHMSNPs were 10.0 and 5.5 kb, respectively. Some detailed statistics can be found in Supplementary Data.

The HapMap SNP genotype data and their phased version (HapMap Public Release #16c.1) were obtained from the web site of the IHMP (http://www.hapmap.org/genotypes/2005-06_16c.1_phaseI/ and http://www.hapmap.org/downloads/phasing/2005-03_phaseI/, respectively) (15). Cytoband and transcript information using the coordinates of human reference sequence, NCBI Build 35.1, were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/cytoBandIdeo.txt.gz), and NCBI (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/mapview/seq_gene.md.gz), respectively (8,9). All of the data were converted into generic feature format (GFF), and then were loaded into the MySQL database using a script of GBrowse (bp_load_gff.pl) (More information about GBrowse follows).

DATABASE IMPLEMENTATION AND WEB INTERFACE

D-HaploDB has been developed and maintained on the Linux platform, and was designed for easy user access, which benefits from a platform-independent web-based application called generic genome browser (GBrowse v1.62) developed by Stein *et al*. (10) as a part of the Generic Model Organism System Database Project (GMOD) (http://www.gmod.org). The back end is a MySQL relational database, and the front end consists of a set of Perl scripts, all running in an Apache environment. The CGI scripts that produce the detailed information reports for CHMSNPs, blocks and LD bins were written in Perl. Data retrieval from the MySQL database is performed using the Perl DBI. The versions of the software are described in our website.

GBrowse supports various types of queries such as gene name, reference SNP ID and chromosome position (Figure 1). D-HaploDB also supports searches by the IDs of blocks and LD bins. Using checkboxes, users can select various tracks they wish to display or hide. Glyphs in each track are linked to relevant resources, such as the Entrez Gene and HapMap web sites. Users can select the tracks and dump the relevant data from the displayed region in a space-delimited text format, which can then be imported into, e.g. an Excel spreadsheet or other data analysis tools. Users can add annotations to the database by preparing GFF files describing the nature and position of the annotation and then uploading these files to GBrowse by the 'Upload a file' function.

The detailed information reports for the genotypes of CHMSNPs, haplotype blocks and LD bins can be opened

(A) CHM Genotyping Details

RS ID :	rs3808185
Perlegen ID :	afd0523930
Alleles :	T/C
Allele Count :	74 (T : 41, C : 28, N : 5)
Allele Frequency :	T : 0.594, C : 0.406
Assayed Sequence :	AAATTATGATAGCANCAAGTGGAATTAAT
Position :	Chr7 : 116808040 .. 116808040

Genotypes :	Sample Name	Genotype
	CHM001	C
	CHM002	T
	CHM003	C
	CHM005	T
	CHM006	C
	CHM007	C
	CHM008	T
	CHM009	T
	CHM010	C
	CHM011	T
	CHM012	C
	CHM013	T
	CHM015	C
	CHM016	C
	CHM018	C
	CHM019	C
	CHM020	T
	CHM021	T
	CHM022	T
	CHM023	C
	CHM024	C

(B) CHM LD bin Details

LD bin

LD bin ID :	LDbin12_2_82
No. of Marker :	5
LD bin Size :	140781
Position :	Chr12 : 96849980 .. 96990760
No. of TagSNP :	3

Haplotype frequency

Haplotype ID	Frequency	Haplotype
Freqency of Haplotype_1	0.676	AATTC
Freqency of Haplotype_2	0.265	GGGCT
Freqency of Haplotype_3	0.044	AATCC
Freqency of Haplotype_4	0.015	GATTC

SNP

RS ID :	rs7305809	rs7297769	rs11109338	rs2216021	rs11836620
Perlegen ID :	afd0911271	afd0911310	afd0911418	afd0911435	afd0911466
Position :	96849980	96895531	96966409	96975679	96990760
Allele :	A/G	A/G	T/G	C/T	C/T
tagSNP :		*	*		*

LD table

	r^2				
rs7305809		0.87	0.87	0.76	0.87
rs7297769	1.00		1.00	0.82	1.00
rs11109338	1.00	1.00		0.81	1.00
rs2216021	1.00	1.00	1.00		0.81
rs11836620	1.00	1.00	1.00	1.00	
	D'				

(C) CHM Block Details

Block

Block ID :	Block1815
No. of Marker :	6
Block Size :	43452
Position :	Chr7 : 116428814 .. 116472265
No. of TagSNP :	2
Unambiguous Haplotype :	67

Common Haplotype (> 5 %)

Haplotype ID	Frequency	Haplotype
Common Haplotype_1	0.358	CTTAAC
Common Haplotype_2	0.328	CTTGAT
Common Haplotype_3	0.134	TTTGAT

SNP

RS ID :	rs38892	rs10808186	rs193586	rs3757812	rs6954357	rs38895
Perlegen ID :	afd0523493	afd0523495	afd0523498	afd0523505	afd0523526	afd0523540
Position :	116428814	116431561	116442145	116445690	116460762	116472265
Allele :	C/T	T/G	T/G	G/A	C/A	T/C
tagset0001 :	*			*		
tagset0002 :	*					*

by clicking the glyphs in each track. These reports include individual genotypes, common haplotypes and tagSNPs, as shown in Figure 2.

DATABASE PRESENTATION

GBrowse (10) is used to visualize genetic and genomic data. Currently 11 overview tracks and six detail tracks are available (Figure 1).

SNP genotypes of CHM samples

CHMSNPs indicate SNPs used for genotyping CHM samples. SNPs that have been genotyped only in CHMs are indicated in red, while those genotyped both in CHMs and HapMap JPT are shown in blue. The glyph for each SNP is linked to a table page, which contains individual genotypes and allele counts (Figure 2A). The density of genotyped SNPs (SNPs per 100 kb) is viewable in an overview track.

Haplotype blocks and LD bins

In D-HaploDB, haplotype blocks were partitioned using the methods of Zhang *et al.* (HapBlock v30) (11). We also created the map of LD bins using the algorithm of 'ldSelect', which is based on the pairwise LD measurements (12), because of their direct relevance to association studies. In this method, no block structures are assumed and the identified tagSNPs can serve as markers in a pooled DNA analysis (13). We have modified the ldSelect program (12) to accept the haploid data and then used it to identify bins of common SNPs (MAF ≥ 0.1) that are in strong LD ($r^2 \geq 0.8$). We identified a total of 176 152 LD bins using the CHMSNPs, and the results were integrated into D-HaploDB. In the LD bin track of the database, the tagSNPs and best-tagSNPs (the tagSNP that showed the highest average r^2 value for the remaining members within the bin) are highlighted in light blue and red, respectively. We found these best-tagSNPs to be efficient markers in a genome-wide association study (Miyatake, K., manuscript in preparation). Detailed information, such as common haplotypes and tagSNPs, can be viewed by clicking each bin (Figure 2B).

Simulated region

The PHASE program (3) was adopted in the IHMP to infer haplotypes from diploid data, as the program is known to be one of the most accurate software to estimate the haplotypes. The genome-wide definitive haplotype data obtained here provides an excellent chance of assessing the accuracy of the inference process. To assess the accuracy, we selected 134 non-overlapping regions each containing 50 SNPs, and then the process of phasing was simulated using 100 sets of pseudo-individuals that were randomly paired definitive haplotypes (6). We found some difference between the block structures deduced from the CHM haplotype set and those from the inferred haplotype sets that are visually comparable in a separate window, which can be opened by

Figure 2. Reports exported from D-HaploDB. CHM genotyping detail report (**A**), LD bin detail report (**B**) and block detail report (**C**) are shown. The asterisks indicate the tagSNPs. The best-tagSNPs are highlighted in red.

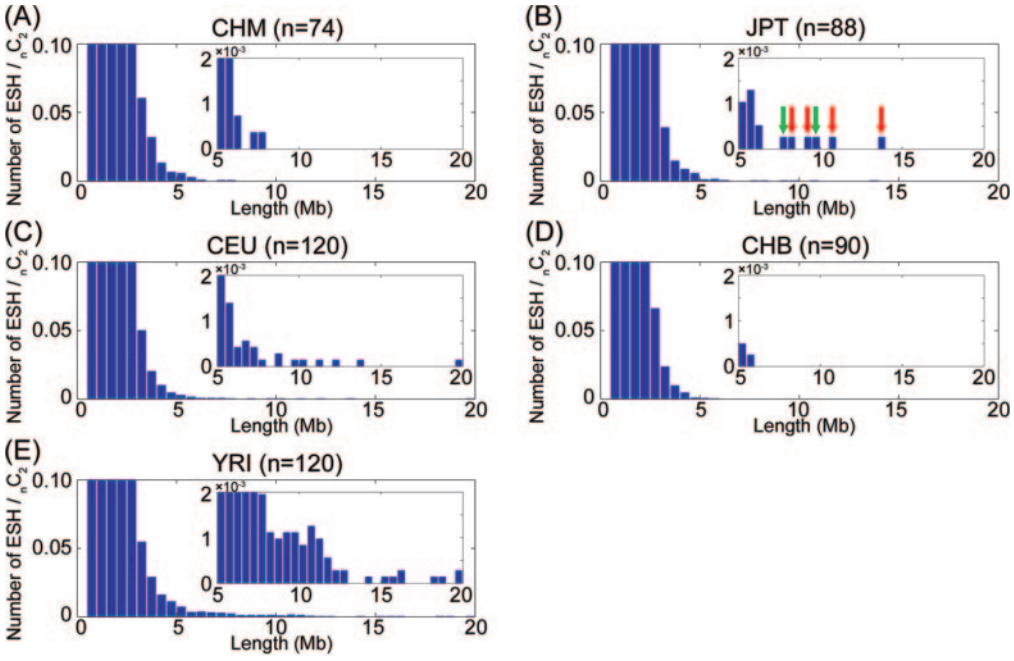

Figure 3. The length distribution of the ESHs. The number of ESHs, normalized by the number of combinations of chromosomes, in the populations of CHM (**A**), JPT (**B**), CEU (**C**), CHB (**D**) and YRI (**E**) are shown. ESHs longer than 500 kb were counted. The SNP intervals longer than 100 kb were treated as breaks in the ESHs. It is noteworthy that unusually long haplotypes were observed in the JPT samples, while no such long ESHs were detected in the CHM samples. The long ESHs originating from the same individuals in JPT are indicated by the green and red arrows.

clicking a glyph of the simulation track. We are currently studying how these differences can affect the power of the association study.

Extended shared haplotype

The extended shared haplotype (ESH) analysis is an effective statistical method for the detection of recent positive selection events in the human genome. We determined the ESH intervals for all combinations of the CHM haplotypes. The ESH densities, calculated as the number of overlapping ESHs per 100 kb window, are presented in the ESH overview track. The ESH densities for other populations have also been calculated using the phased data from HapMap and are presented in the additional overview tracks for comparative purposes.

It has long been the concern that association studies may suffer from high rates of false positives if there is an unrecognized population structure. It is perhaps less widely appreciated that so-called 'cryptic relatedness' (i.e. kinship among the samples that is not known to the investigator) (14) might also potentially inflate the false positive rate. One of the important aspects of our database is the ESH map, which provides not only the opportunity to identify candidate selection regions (6), but it also sheds some light on cryptic relatedness among the collected samples. Figure 3 shows the length distribution of the ESH intervals. As is evident from the figure, significant numbers of long ESH are observed in all HapMap samples except for CHB. Considering the fact that the ESHs in the HapMap samples have a greater chance to be broken than those in CHM because of the three times higher density of SNPs (and also possible phasing errors in JPT and CHB) in that database, such long ESHs are indicative of the cryptic relatedness in some of the CEU, YRI and JPT samples of HapMap. The presence of cryptic relatedness in the HapMap samples was also mentioned in a recent publication (15).

FUTURE PERSPECTIVE

D-HaploDB is a work in progress and we are presently genotyping another set of SNPs (Affymetrix 500k Array) using 100 CHM samples, which is expected to further improve the resolution and quality of haplotype map of Japanese population. In addition, efforts are also underway to re-select tagSNPs, based on an improved LD measure mapping method (16).

BULK DOWNLOAD OF THE DATA

In the download section, all of the D-HaploDB data set is available as a series of text files at http://orca.gen.kyushu-u.ac.jp/download.html. This includes raw genotypes and analytical results such as blocks, LD bins, common haplotypes and tagSNPs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports,

Science and Technology of Japan. The authors wish to acknowledge the genotyping work of people in Perlegen Sciences, Inc., USA (Drs Renee Stokowski, David Hinds, Krishna Pant and David Cox). CHM samples were collected by Drs Toshio Hirakawa, Hidenori Kato, Takao Matsuda and Norio Wake of Kyushu University in collaboration with the Japan Association of Obstetricians & Gynecologists. Part of the computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo. Some of the data included in this article are from The International HapMap Project web site. We also thank Dr Todd Taylor (RIKEN Genomic Sciences Center) for critical reading and comments of this manuscript. Funding to pay the Open Access publication charges for this article was provided by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas ''Applied Genomics'' from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Johnson,G.C., Esposito,L., Barratt,B.J., Smith,A.N., Heward,J., Di Genova,G., Ueda,H., Cordell,H.J., Eaves,I.A., Dudbridge,F. *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nature Genet*., **29**, 233–237.
2. The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
3. Stephens,M. and Scheet,P. (2005) Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet*., **76**, 449–462.
4. Taillon-Miller,P., Bauer-Sardina,I., Zakeri,H., Hillier,L., Mutch,D.G. and Kwok,P.Y. (1997) The homozygous complete hydatidiform mole: a unique resource for genome studies. *Genomics*, **46**, 307–310.
5. Fan,J.B., Surti,U., Taillon-Miller,P., Hsie,L., Kennedy,G.C., Hoffner,L., Ryder,T., Mutch,D.G. and Kwok,P.Y. (2002) Paternal origins of complete hydatidiform moles proven by whole genome single-nucleotide polymorphism haplotyping. *Genomics*, **79**, 58–62.
6. Kukita,Y., Miyatake,K., Stokowski,R., Hinds,D., Higasa,K., Wake,N., Hirakawa,T., Kato,H., Matsuda,T., Pant,K. *et al*. (2005) Genome-wide definitive haplotypes determined using a collection of complete hydatidiform moles. *Genome Res*., **15**, 1511–1518.
7. Hinds,D.A., Stuve,L.L., Nilsen,G.B., Halperin,E., Eskin,E., Ballinger,D.G., Frazer,K.A. and Cox,D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
8. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
9. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
10. 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.
11. Zhang,K., Deng,M., Chen,T., Waterman,M.S. and Sun,F. (2002) A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl Acad. Sci. USA*, **99**, 7335–7339.
12. Carlson,C.S., Eberle,M.A., Rieder,M.J., Yi,Q., Kruglyak,L. and Nickerson,D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet*., **74**, 106–120.
13. Sham,P., Bader,J.S., Craig,I., O'Donovan,M. and Owen,M. (2002) DNA Pooling: a tool for large-scale association studies. *Nature Rev. Genet*., **3**, 862–871.
14. Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
15. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
16. Qin,Z.S., Gopalakrishnan,S. and Abecasis,G.R. (2006) An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. *Bioinformatics*, **22**, 220–225.