BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 22 2005, pages 4181–4186
doi:10.1093/bioinformatics/bti682

Databases and ontologies

SNPselector: a web tool for selecting SNPs for genetic association studies

Hong Xu, Simon G. Gregory, Elizabeth R. Hauser, Judith E. Stenger, Margaret A. Pericak-Vance, Jeffery M. Vance, Stephan Züchner and Michael A. Hauser*

The Duke Center for Human Genetics, Duke University Medical Center, Durham, NC 27710, USA

Received on August 8, 2005; revised and accepted on September 18, 2005
Advance Access publication September 22, 2005

ABSTRACT

Summary: Single nucleotide polymorphisms (SNPs) are commonly used for association studies to find genes responsible for complex genetic diseases. With the recent advance of SNP technology, researchers are able to assay thousands of SNPs in a single experiment. But the process of manually choosing thousands of genotyping SNPs for tens or hundreds of genes is time consuming. We have developed a web-based program, SNPselector, to automate the process. SNPselector takes a list of gene names or a list of genomic regions as input and searches the Ensembl genes or genomic regions for available SNPs. It prioritizes these SNPs on their tagging for linkage disequilibrium, SNP allele frequencies and source, function, regulatory potential and repeat status. SNPselector outputs result in compressed Excel spreadsheet files for review by the user.

Availability: SNPselector is freely available at http://primer.duhs.duke.edu/

Contact: mike.hauser@duke.edu

INTRODUCTION

Single nucleotide polymorphism (SNP) is the most common form of polymorphism in the human genome. A variety of genotyping platforms are available for high-throughput assay of SNPs. They are widely used for human evolution research (Hammer *et al*., 2001; Underhill *et al*., 2000), association studies of complex diseases (Colomb *et al*., 2001; Martin *et al*., 2001) and studies of pharmacogenetics (Goldstein *et al*., 2003).

The amount of SNP data in public databases is increasing dramatically. The number of unique human SNPs in the current dbSNP release (build 123) is >10 million, approaching the theoretically expected number of SNPs in the human genome (Kruglyak and Nickerson, 2001). The SNP detection method varies, as does the reliability of the SNPs in dbSNP. Only 50% of the SNPs are validated and <20% of the validated SNPs have allele frequency information. In addition to the validation and allele frequency information, it is also important to select SNPs based on their genomic location and their proposed functional significance (coding, intronic, promoter, etc.). These annotation data are available in various resources, such as NCBI dbSNP (http://www.ncbi.nlm.nih.gov/SNP), UCSC Genome Browser (Karolchik *et al*., 2003) and Ensembl (Birney *et al*., 2004). They also provide data mining features to retrieve SNP information. There are several other bioinformatics tools developed to select SNPs based on various properties. PromoLign (Zhao *et al*., 2004) and PupaSNP Finder (Conde *et al*., 2004) are two web tools to find SNPs that may affect gene transcription levels. SNPper (Riva and Kohane, 2002) provides a web interface to retrieve SNP annotation by chromosome region or SNP names and to refine SNP selection with different filters (such as validation status or minor allele frequency).

Here we describe a new SNP selection program that combines many of these attributes. It has an easy-to-use web interface that provides a feature-rich result spreadsheet. It incorporates LD calculations into SNP selection to help reduce the number of SNPs required for a comprehensive analysis. Further, the output can be tailored to provide the required fields for commercial genotyping systems, such as the Illumina bead-based genotyping platform.

IMPLEMENTATION

SNPselector is implemented in object-oriented PERL language to search and analyze SNP data. Its core module can be run as a UNIX command-line application. To make it easy to use, a CGI wrapper is developed to provide the web interface between the users and the application.

To increase the performance of the application, all the SNP data and related genome annotation data are stored in a local MySQL database (http://www.mysql.com/). SNP data, including SNP location, alleles, function and validation information were downloaded from UCSC Genome Browser server. Later two 100 bp flanking sequences for each SNP were extracted from the human genome (NCBI build 35) and added into the SNP table. SNP allele frequency and genotyping data were downloaded from the HapMap project (http://www.hapmap.org/), the SNP Consortium (http://snp.cshl.org/), JSNP (http://snp.ims.u-tokyo.ac.jp/), Affymetrix (http://www.affymetrix.com/) and Perlegen (Hinds *et al*., 2005). The SNPs with experimentally verified genotyping or allele frequency information are considered as 'high quality' SNPs. Ensembl gene structure information was obtained from the Ensembl project. Conserved region information was downloaded from the UCSC genome browser multi-genome alignments (Blanchette *et al*., 2004). CpG island, transcription factor binding site (TFBS), microRNA and simple repeat data were also downloaded from UCSC and stored in the local MySQL database.

The local database is updated whenever new public data are released.

*To whom correspondence should be addressed.

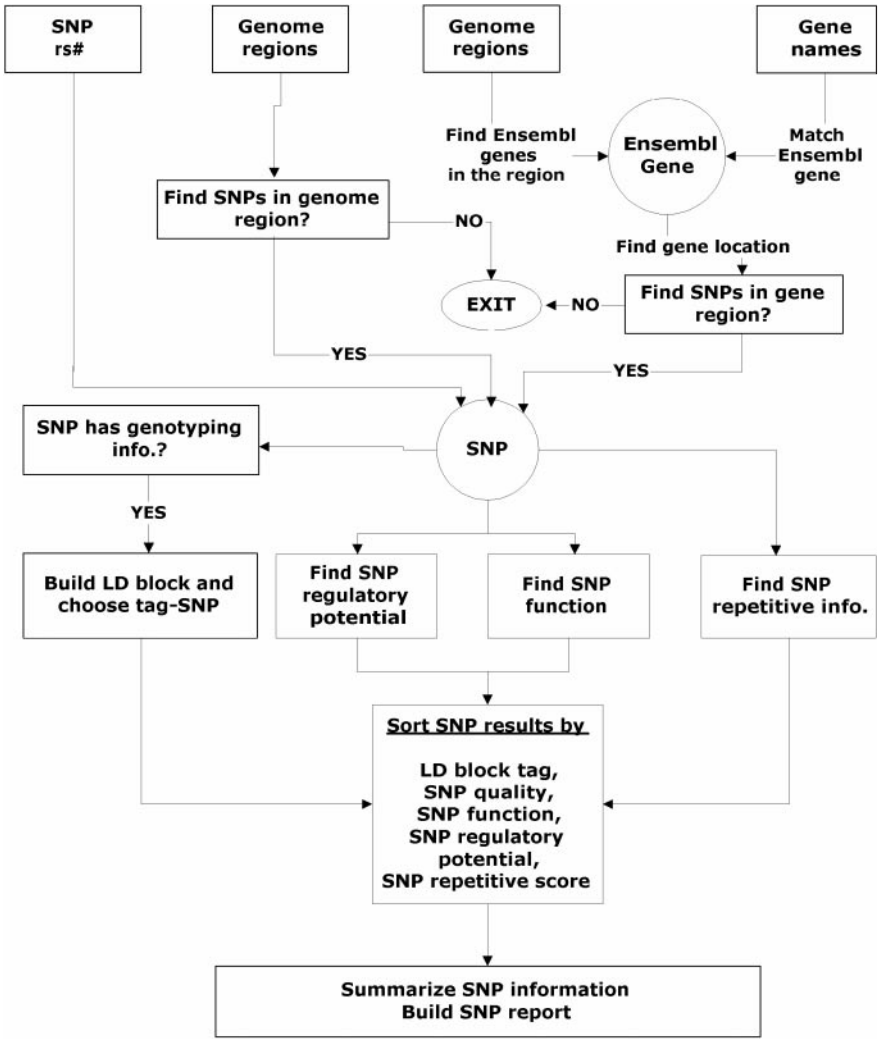


Fig. 1. Workflow of SNP selection process.

PROGRAM WORKFLOW

SNPselector takes a list of gene names or genomic regions as input and finds all available SNPs in the genes or genomic regions. SNPselector finds tagging SNPs by calculating LD bins of genotyped SNPs (Carlson *et al*., 2004). It then finds SNP function based on whether the SNP may affect the gene transcript structure or the protein product. It checks the regulatory potential of the SNP based on SNP location at conserved site from multi-genome comparison, conserved TFBS, CpG island or microRNA gene. It also checks whether the SNP is in a repeat region. It scores and sorts SNPs on their LD tagging property, quality, function, regulatory potential and repeat status. Finally it exports SNP selection result into Excel files (Fig. 1).

SNP search

The SNPselector provides the user with four types of SNP searches:

(1) dbSNP accession ID (rs number).

(2) Gene names: Ensembl genes and their chromosomal locations are obtained. For each Ensembl gene, SNPselector searches all SNPs in the corresponding chromosomal region (plus the flanking sequence regions defined in the user input).

(3) Genomic regions (gene centric): SNPselector finds all Ensemble genes and their chromosome locations in that genomic region. For each Ensembl gene in the region, SNPselector searches all SNPs within the gene.

(4) Genomic regions: The program breaks each region into smaller 2 Mb regions if necessary. Then it will search all SNPs in each of the 2 Mb region.

SNP retrieval

After searching by one of these methods, SNPselector retrieves SNP information from the database by SNP rs numbers. The information includes SNP allele, allele frequency, chromosomal location, validation status, quality, predicted function and flanking sequence. When obtaining the SNP flanking sequence SNPselector annotates any neighboring SNPs within 100 bases of the target SNP using the IUPAC codes. This ensures that no assay will be designed over a neighboring SNP that can cause failure of many SNP genotyping

Table 1. SNP type and its function score

SNP type	Description	Score
Coding-non-synonymous	Change in peptide with respect to contig sequence	1.0
Splice-site	Variation in first two or last two bases of intron	1.0
Coding-synonymous	No change in peptide for allele with respect to contig sequence	0.9
Coding	Variation in coding region of gene, assigned if allele-specific class unknown	0.9
mRNA-UTR	Variation in transcript, but not in coding region interval	0.8
Locus-region	Variation in region of gene, but not in transcript (within 2 kb flanking sequence of the mRNA)	0.7
Intron	Variation in intron, but not in first two or last two bases of intron	0.6
Exception	Variation in coding region with exception raised on alignment. This occurs when protein with gap in sequence is aligned back to contig sequence	0.6
Reference	Allele observed in reference contig sequence	0.6
Unknown	No known functional classification	0.6

Table 2. SNP type and its function score

Region	Description	Score
Conserved region of 8-genome alignment	Human/chimp/mouse/rat/dog/chicken/fugu/zebra_fish conserved region from 8-genome multi-alignment	Use multiple alignment score[a]
Conserved TFBS	Contain the location and score of TFBSs conserved in the human/mouse/rat alignment	Use original raw score[b]
CpG island	CpG island may be potential regulatory region	Use CpG island expectation score[c]
microRNA gene	microRNA involves in gene regulation	Score[d] is normalized to the range between 0 and 1

[a]UCSC phastCons scores for multiple alignments (Siepel and Haussler, 2005).
[b]Penn State University Genome Alignment and Annotation Database (http://gala.cse.psu.edu/gala/downloads/hg17/tfbinding_sites/).
[c]UCSC human annotation database table for CpG island (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/cpgIslandExt.txt.gz).
[d]UCSC human annotation database table for microRNA (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/wgRna.txt.gz).

assays. SNPselector also compares the SNP location with Ensembl transcript structure to determine whether the SNP is intronic, exonic or intergenic and annotates at which exon or intron the SNP is located if it is not an intergenic SNP.

For SNPs that are queried by gene or genomic region, SNPselector calculates LD bins using the HapMap genotyping data and the 'ldSelect' program from University of Washington (Carlson *et al*., 2004). This helps to select the most informative SNPs and to avoid genotyping redundant SNPs. The Perlegen genotyping data of African American, Caucasian and Chinese were also added into the SNPselector database. Users can select one of the genotyping data sources or populations to do LD bin analysis.

SNP scoring and prioritization

SNP scoring After retrieving SNP information, SNPselector scores each SNP in multiple categories.

(1) LD score: If the SNP is a tagging SNP of an LD bin, its LD score is assigned as the number of SNPs in the LD bin. Otherwise, its LD score is zero. The LD score reflects how informative the tagging SNP is. The higher the LD score, the more SNPs in the LD bin, the more SNP information can be assayed by the tagging SNP.

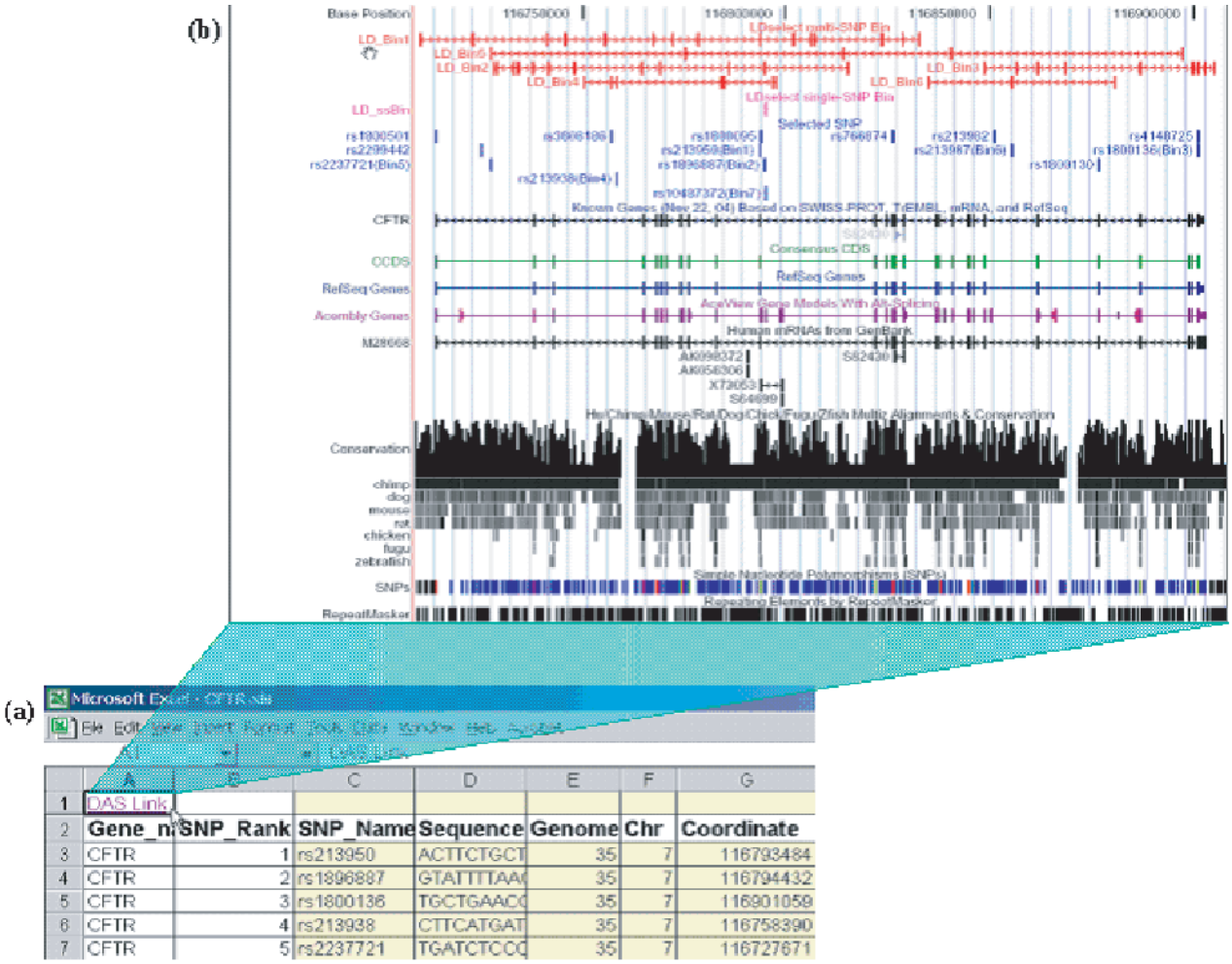
(2) Quality score: If the SNP has experimentally verified genotyping or allele frequency information, it is considered as a 'high-quality' SNP. Its quality score is 1. Otherwise, its quality score is 0.

(3) Function score: The SNP function score is based on the SNP type annotation from dbSNP. A higher score is assigned to the SNPs that might affect gene transcript structure or protein product, such as coding non-synonymous SNPs or SNPs at a splicing site (Table 1).

(4) Regulatory potential score: For each SNP not in an exonic region, SNPselector calculates its potential regulatory score base on its location within human/chimp/mouse/rat/dog/chicken/fugu/zebra_fish conserved regions, conserved TFBS, CpG island or microRNA gene (Table 2). These scores are added to build a single regulatory potential score. Thus a high score suggests high regulatory potential.

(5) Repetitive score: If the SNP overlaps with a simple repeat region annotated by UCSC, its repetitive score is 1. Otherwise, its repetitive score is 0.

(6) Illumina pre-assay score: If SNP genotyping is to be performed with the Illumina bead platform, the user can upload an optional file containing the Illumina pre-assay score into the database. This score is calculated by the Illumina proprietary



	A	B	C	D	E	F	G
1	DAS Link						
2	Gene_n	SNP_Rank	SNP_Name	Sequence	Genome	Chr	Coordinate
3	CFTR	1	rs213950	ACTTCTGCT	35	7	116793484
4	CFTR	2	rs1896887	GTATTTTAA	35	7	116794432
5	CFTR	3	rs1800136	TGCTGAACC	35	7	116901059
6	CFTR	4	rs213938	CTTCATGAT	35	7	116758390
7	CFTR	5	rs2237721	TGATCTCCC	35	7	116727671

Fig. 2. Display LD bins and selected SNPs as custom tracks in UCSC Genome Browser. **(a)** The hyperlink—'DAS Link' in SNP report Excel spreadsheet. **(b)** LD bins and selected SNPs are displayed as custom tracks in UCSC genome browser. LD bins have two tracks—the 'red' track shows LD bins with multiple SNPs in one bin, the 'pink' tracks show LD bins with a single SNP in one bin. Selected SNPs are displayed in 'blue' track under tracks of LD bins.

algorithm to assess the success rate of genotyping the SNP with this platform.

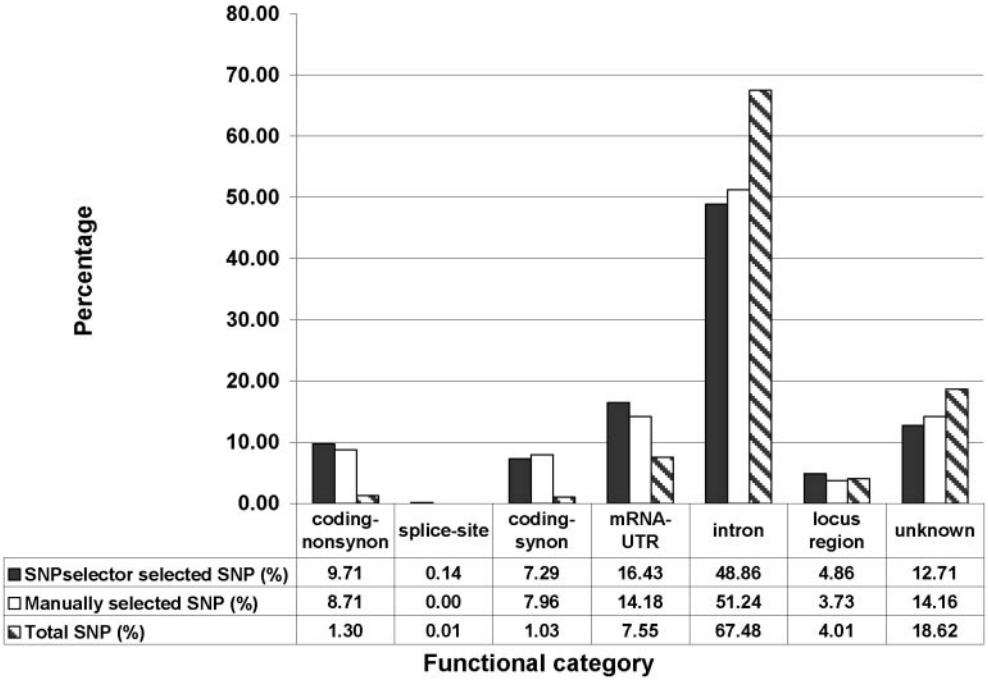
SNP prioritization After searching and assigning scores to SNPs, SNPselector sorts SNPs by LD score in descending order so that the tag SNPs with the larger LD bin will be at the top. Then SNPselector sorts SNPs by quality score in descending order, followed by functional score in descending order, so that SNPs with functional impact, such as non-synonymous coding SNPs, will have higher rank than those with unknown function. SNPselector also sorts SNPs by regulatory potential score so that SNPs in conserved regions, CpG islands or TFBSs will be ranked higher than those outside these regions. Finally SNPselector sorts SNPs by repetitive score in increasing order so that SNPs in non-repetitive regions will be ranked higher than those in repetitive regions. Since the output is an easily manipulated spreadsheet, the user can sort the SNPs to highlight different SNP features. For example, if the user wants to find SNPs that might affect gene expression, he/she may choose to sort SNPs by regulatory potential score before sorting SNPs by function score.

SNP selection and data report

For SNPs that are queried by SNP accession IDs, SNPselector selects and exports all the queried SNPs in one Excel spreadsheet file. For SNPs queried by gene names or gene locations, SNPselector exports top ranked SNPs at the user-specified number per gene into one Excel spreadsheet file. It also exports all SNPs available for each gene into a second gene-SNP spreadsheet. For genome-scan SNPs that are queried by genomic regions, SNPselector selects evenly distributed SNPs at the user-specified spacing (in base pairs) and puts the result into one Excel spreadsheet file. In each gene or genome SNP Excel file, SNPselector generates a hyperlink called 'DAS Link' at the first field of the first row. It links to the LD bins and selected SNPs are displayed as custom tracks in the UCSC genome browser (Fig. 2).

RESULTS AND DISCUSSION

SNPselector was used to select 5 SNPs for each of 140 candidate genes for human cardiovascular disease (Seo *et al*., 2004). The 140 genes were widely distributed across the human genome,



	coding-nonsynon	splice-site	coding-synon	mRNA-UTR	intron	locus region	unknown
■ SNPselector selected SNP (%)	9.71	0.14	7.29	16.43	48.86	4.86	12.71
□ Manually selected SNP (%)	8.71	0.00	7.96	14.18	51.24	3.73	14.16
◪ Total SNP (%)	1.30	0.01	1.03	7.55	67.48	4.01	18.62

Fig. 3. The distribution of selected SNPs and total SNPs in different functional categories.

and SNPs had previously been manually selected from these genes. Among the 700 SNPs selected by SNPselector, all were high quality SNPs with allele frequency or genotyping data, and 582 (83%) were LD tagging SNPs.

Figure 3 shows the distribution of selected SNPs and total SNPs of the 140 genes in different functional categories. The majority of the SNPs were intronic SNPs. Through the SNP prioritizing rules, SNPselector decreased the percentage of intronic SNPs from 67.48% in the total available SNPs to 48.86% in the final selected SNPs. It also enriched SNPs that might have an effect on gene function. These included coding-non-synonymous SNPs (enriched 7 times), splice-site SNPs (enriched 14 times), coding-synonymous SNPs (enriched 7 times) and mRNA-UTR SNPs (enriched 2 times). This enrichment of functional SNPs by SNPselector was similar to the SNPs selected by the manual SNP selection process. In some categories, such as splice-site and mRNA-UTR, SNPselector did even better than the manual SNP selection. SNPselector prioritizes splice-site SNPs at the same level as coding-non-synonymous SNPs and chooses UTR SNPs located in conserved genomic region.

There are a few limitations to SNPselector as it is currently configured. The software requires SNP genotyping data to calculate tagging SNPs with ldSelect. To provide the richest possible dataset, we have merged HapMap genotyping data with other genotyping information, such as Perlegen genotyping data (Hinds *et al*., 2005). This approach generates a large number of LD bins when there is little overlap between the genotyped datasets. This will become less of an issue as the HapMap genotyping progresses. SNPselector infers each SNP's impact on gene function based on its location (e.g. coding region, promoter site or UTR). However, SNPs located in these regions may not affect the gene function. We are working to add more detailed functional annotation from other resources (Karchin *et al*., 2005).

In summary, SNPselector is a powerful tool for the identification of SNPs for large-scale genetic association studies. This software's output is comparable to that obtained from manual selection, but can be produced in a fraction of the time. The detailed descriptive output can be formatted for submission to a variety of commercial genotyping systems. SNPselector will be a valuable addition to many high-throughput SNP genotyping applications.

ACKNOWLEDGEMENTS

We would like to thank Carrie Browning, Liyong Wang, Jason Rose and others for their helpful suggestions. This work was supported by the following grants P01 HL73042 (NHLBI); R01 AG021547, R01 NS36768 and R01 NS31153 (NINDS); R01 AG19085 (NIA) and R01 EY12012 and R01 EY13315 (NEI).

Conflict of Interest: none declared.

REFERENCES

Birney,E. *et al.* (2004) An overview of ensembl. *Genome Res.*, **14**, 925–928.

Blanchette,M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.

Carlson,C.S. *et al.* (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.

Colomb,E. *et al.* (2001) Association of a single nucleotide polymorphism in the TIGR/MYOCILIN gene promoter with the severity of primary open-angle glaucoma. *Clin. Genet.*, **60**, 220–225.

Conde,L. *et al.* (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.*, **32**, W242–W248.

Goldstein,D.B. *et al.* (2003) Pharmacogenetics goes genomic. *Nat. Rev. Genet.*, **4**, 937–947.

Hammer,M.F. *et al.* (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol. Biol. Evol.*, **18**, 1189–1203.

Hinds,D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.

Karchin,R. *et al*. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.

Karolchik,D. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.

Kruglyak,L. and Nickerson,D.A. (2001) Variation is the spice of life. *Nat. Genet*., **27**, 234–236.

Martin,E.R. *et al*. (2001) Association of single-nucleotide polymorphisms of the tau gene with late-onset Parkinson disease. *J. Am. Med. Assoc*., **286**, 2245–2250.

Riva,A. and Kohane,I.S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics*, **18**, 1681–1685.

Seo,D. *et al*. (2004) Gene expression phenotypes of atherosclerosis. *Arterioscler. Thromb. Vasc. Biol*., **24**, 1922–1927.

Siepel,A. and Haussler,D. (2005) Phylogenetic hidden Markov models. In Nielsen,R. (ed.), Statistical Methods in Molecular Evolution. Springer, New York, pp. 325–351.

Underhill,P.A. *et al*. (2000) Y chromosome sequence variation and the history of human populations. *Nat. Genet*., **26**, 358–361.

Zhao,T. *et al*. (2004) PromoLign: a database for upstream region analysis and SNPs. *Hum. Mutat*. **23**, 534–539.