SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system

Alexander R. Pico[1,*], Ivan V. Smirnov[2], Jeffrey S. Chang[3], Ru-Fang Yeh[3], Joseph L. Wiemels[3], John K. Wiencke[2], Tarik Tihan[4], Bruce R. Conklin[1,5,6] and Margaret Wrensch[2]

[1]Gladstone Institute of Cardiovascular Disease, 1650 Owens Street, San Francisco, CA 94158, [2]Department of Neurological Surgery, Neuroepidemiology Division, [3]Department of Epidemiology and Biostatistics, [4]Department of Pathology, [5]Department of Medicine and [6]Cellular and Molecular Pharmacology, University of California, San Francisco, CA 94122, USA

ABSTRACT

SNPLogic (http://www.snplogic.org) brings together single nucleotide polymorphism (SNP) information from numerous sources to provide a comprehensive SNP selection, annotation and prioritization system for design and analysis of genotyping projects. SNPLogic integrates information about the genetic context of SNPs (gene, chromosomal region, functional location, haplotypes tags and overlap with transcription factor binding sites, splicing sites, miRNAs and evolutionarily conserved regions), genotypic data (allele frequencies per population and validation method), coverage of commercial arrays (ParAllele, Affymetrix and Illumina), functional predictions (modeled on structure and sequence) and connections or established associations (biological pathways, gene ontology terms and OMIM disease terms). The SNPLogic web interface facilitates construction and annotation of user-defined SNP lists that can be saved, shared and exported. Thus, SNPLogic can be used to identify and prioritize candidate SNPs, assess custom and commercial arrays panels and annotate new SNP data with publicly available information. We have found integration of SNP annotation in the context of pathway information and functional prediction scores to be a powerful approach to the analysis and interpretation of SNP-disease association data.

INTRODUCTION

Logical single nucleotide polymorphism (SNP) selection is a common challenge to designing a genotyping strategy for disease association studies. Analysis and interpretation of data from large-scale SNP arrays also pose major analytic challenges. Within the constraints of available microarray platforms, budget and analysis methods, one ideally targets the most informative set of polymorphic sites given what is known about the disease and the genes and processes likely to be involved. We developed the database and web site, SNPLogic, to facilitate and enhance the SNP selection process for gene–disease association studies targeting particular biological pathways and processes. SNPLogic combines pathway-based SNP selection with a prioritization scheme that lets users rank SNPs according to their associations, connections, annotations and functional predictions. The integrated annotations are also useful during the analysis of SNP-based association data. Furthermore, the use of biological pathways in SNP selection allows researchers to thoroughly investigate potential SNP interactions and provides mechanistic hypotheses to drive validation and future research.

In contrast to the other SNP-related databases, SNPLogic does *not* produce its own annotations of polymorphisms based on location, frequency, disease association or structural and functional effects, but rather aggregates and transforms annotations from such databases. The unique presentation of integrated SNP annotations allows researchers to logically build, assess and analyze candidate SNP lists (Figure 1).

*To whom correspondence should be addressed. Tel: +1 415 734 2741; Fax: +1 415 355 0960; Email: apico@gladstone.ucsf.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
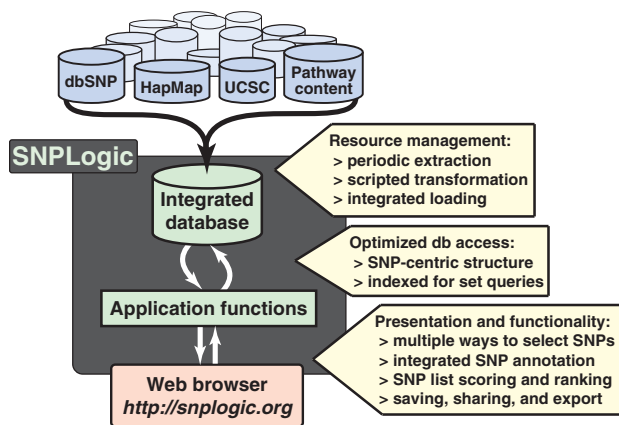
Figure 1. Overview of resource collection and integration into SNP-Logic for a unique selection, annotation and analysis tool. A semi-automatic process periodically extracts SNP, pathway and genomic information from various public resources (blue cylinders), which is then integrated into a single SNP annotation database. The SNPLogic web site provides access, unique functionality and presentation of the integrated database.

The main goal of SNPLogic is to step beyond the 'single SNP' perspective in both the selection of candidate SNPs and the analysis of SNP effects. Using SNPLogic, researchers can generate SNP lists based on genes of interest, chromosomal regions, biological pathways, ontology terms and disease associations, thereby grouping SNPs in biologically meaningful ways. Researchers can merge or manually prune their lists to create custom SNP lists. All of the annotations compiled for SNPLogic are available for any custom list of SNPs, allowing disparate annotations to be considered and weighed side-by-side. Using SNPLogic, the researcher can then design scoring rules based on any of the available annotations (e.g. IF average heterozygosity $> X$, THEN score $= +1$ AND IF functional prediction $= Y$, THEN score $= +2$). Sets of scoring rules can be combined to generate a user-defined score, which is added to the annotation table and can be used to rank sort any custom list of SNPs.

Following data collection, the annotations, groups and scores associated with the custom lists of SNPs within SNPLogic can be used to inform analyses, such as random forests (RF) analysis (1). Thus, we envision the tool being used in at least two fashions: (i) as an SNP selection tool for low- to mid-density custom genotyping projects and (ii) as an analysis tool, applicable even to large pre-set platforms (e.g. Illumina, Affymetrix). Up until now, there has been a lack of resources for both of these use cases.

IMPLEMENTATION AND DESIGN

Overview

The SNPLogic backend database is built around NCBI's dbSNP collection (2). Annotations and gene associations from dbSNP are stored in a local MySQL 5 database, together with additional annotations and functional predictions from different souces. A meta-table stores information about each resource, including references to pre- and post-processing scripts to facilitate semi-automatic, periodic updates and integration of new resources as they are developed. Separate tables store the collections of genes defined by pathways, gene ontology (GO) terms and chromosomal bands and regions. In this way, links between pathways, genes, SNPs and SNP annotations can be queried rapidly and efficiently. The functionality of the front end web site for SNPLogic is written in PHP and dynamic HTML/Javascript.

Data sources

The latest version of the SNPLogic database has been expanded to cover all human SNPs, genes, pathways and diseases, and thus is potentially relevant to the design and execution of any SNP study. The SNPLogic database aggregates annotations from a wide variety of key resources (Table 1). The stable aggregation of data from multiple resources presents a nontrivial challenge. Our solution involves a modular, two-stage approach. First, we register each resource in a meta-table that stores all the information required to access, extract, transform and load its data into a local table. A master script references the resource meta-table and automatically updates local tables that have changed since the last update. When a resource moves or alters its data format, we have to manually edit its registered information, but otherwise, the process is automatic. In the second stage, an aggregate annotation table is built up by appending columns of data from the prepared local tables, using SNP rs numbers as a key. The simple structure allows individual annotations to be updated, added or removed without affecting the rest of the system. In this way, changes to primary annotation resources are isolated from the aggregate annotation table underlying SNPLogic.

Basic annotations for all human SNPs were extracted from dbSNP, including variation, validation, average heterozygosity, subSNP count, function class, associated gene ID and symbol (2). Information about allele frequency and haplotype tags and scores was derived from HapMap (3) and the Tagger tool (4). Annotation of the genomic context of SNPs was provided by Delta-MATCH (http://deltamatch.org), PupaSuite (http://pupasuite.bioinfo.cipf.es) and UCSC (http://genome.ucsc.edu). Delta-MATCH provides a difference Z-score for the change of predicted binding affinities at polymorphic transcription factor binding sites (TFBS) throughout the human genome. High scoring Delta-MATCH hits thus point to SNPs that may dramatically alter gene expression. PupaSuite annotates SNPs based on their overlap with splicing regulatory elements, TFBS, miRNA, triplex and conserved regions in the genome (5). UCSC provides genomic coordinates of highly conserved elements across multiple species, allowing for identification of SNPs overlapping evolutionarily conserved regions of the human genome (6–8). SNPLogic also identifies SNPs that are covered by the major Affymetrix (http://www.affymetrix.com/) and Ilumina (http://www.illumina.com/) SNP microarray probes. Mappings to protein

Table 1. Annotation sources integrated into SNPLogic

Resource	Description of extracted annotations	SNPs annotated
dbSNP, build 128	SNPs rs IDs and basic annotations	11 833 664
HapMap	Allele frequencies from HapMap project	3 967 349
Tagger (HapMap)	Haplotype tags in CEU (0.8 r^2 cutoff)	*tags* 695 153
		tagged 2 009 725
Tagger (HapMap)	Haplotype tags in CHB (0.8 r^2 cutoff)	*tags* 580 509
		tagged 1 908 721
Tagger (HapMap)	Haplotype tags in JPT (0.8 r^2 cutoff)	*tags* 562 741
		tagged 1 883 580
Tagger (HapMap)	Haplotype tags in YRI (0.8 r^2 cutoff)	*tags* 1 282 451
		tagged 1 571 139
UCSC	PhastCons conserved elements, 28-way vertebrate	434 235
UCSC	PhastCons conserved elements, 28-way mammal	322 704
Delta-MATCH	Transcription factor binding sites, scored by ΔZ	2 456 473
PupaSuite	Transcription factor binding sites (Transfac)	81 293
PupaSuite	Transcription factor binding sites (JASPER)	60 082
PupaSuite	DNA triplex sequences	439 350
PupaSuite	Exonic splicing enhancers (ESE)	153 523
PupaSuite	Exonic splicing silencers (ESS)	22 926
PupaSuite	miRNA sequences	20 716
PupaSuite	New splice site formation	13 415
PupaSuite	Splice site disruption	1574
Affymetrix	Genome-Wide Human SNP Array 6.0 (+11 others)	924 216
Illumina	Human 1M BeadChip (+7 others)	1 126 075
Polyphen	Structure-based predictions	53 720
SNP3D	Structure-based predictions	4792
SNP3D	Sequence-based predictions	28 136
NCBI	Gene associations	4 676 589
OMIM	Disease associations via genes	3 169 680
Gene Ontology	*Molecular Function* term associations via genes	2 647 923
Gene Ontology	*Biological Process* term associations via genes	2 619 920
Gene Ontology	*Cellular Component* term associations via genes	1 550 319
KEGG	Pathway associations via genes	1 089 444
WikiPathways	Pathway associations via genes	705 200
BioCarta	Pathway associations via genes	386 817
BioCyc	Pathway associations via genes	92 675

A brief description of the annotations extracted from each resource is given, along with the number of SNPs annotated. Resources are referenced in the text.

structure models and functional predictions come from Polyphen (9) and SNP3D (10). Polyphen uses empirical rules to score SNP effects, while SNP3D uses homology profiling and support vector machine (SVM) stability assessments. LS-SNP is another structure-based prediction resource that will be included in the near future (11). Collections of genes and associated SNPs are defined by pathways, GO terms (12) and disease categories (OMIM) (13). Pathway resources include KEGG (14), WikiPathways (15), GenMAPP (16) (now at Wiki-Pathways), BioCarta (http://www.biocarta.com/) and BioCyc (17).

USER INTERFACE

Video tutorials demonstrating all the main user interfaces and user scenarios are available online at http://snplogic.org/tutorials/ (see Supplementary Material). The SNPLogic web site includes helpful tool tips displayed when you mouse-over any button or field on the page. And you can always find instructions and tips on the Help page, which is linked to from the top-right corner of every page on the site.

Users and groups

From the homepage at http://www.snplogic.org, anyone is free to login using the anonymous login information provided. This free, public account gives the user full access to the data and functionality of SNPLogic. To support collaborative efforts, we also provide private user and group accounts. As a registered user you can join an existing group, or create a new group, together with your colleagues and collaborators. This way you can control the building, scoring and sharing of annotated SNP lists. The viewing and editing of SNP lists can independently be kept private, shared within a group, or shared with the public. The SNP lists associated with your user account are saved for as long as you have an account, so you can continuously build, update and reference your lists.

SNP Selection

In SNPLogic, you can construct lists of SNPs in multiple ways using built-in browsers and tools. You can specify individual or multiple SNP at a time by typing, pasting or uploading standard dbSNP rs numbers. You can also select SNPs by specifying a gene name, where all SNPs associated with a given gene are returned. At an even

higher level, you can specify an entire pathway or GO term, and browse the SNPs associated with the component genes. You can also specify a chromosomal band or region and browse the contained SNPs. Each approach returns one or more SNPs, which you can annotate, filter and sort to help you review and ultimately select SNPs to be added to one or more of your lists.

For example, using the *Gene Browser* shown in Figure 2, you could enter 'ApoE' into the Gene Symbol field, the click Go (or hit enter). All the SNPs associated with the ApoE gene are then displayed as results. To review the results you can choose to view annotations for the SNP in the context of genes by clicking +Fields. Use the checkboxes to select individual annotations or check the top box to select all fields. Again, click Go (or hit enter). The selected annotations are now displayed next to each SNP. You can filter which SNPs are displayed using the Filter by form. Filters can be applied to any visible field. You can also export the results to review externally. In order to actually add an SNP to your list, you must check the checkbox next to each SNP. Checked SNPs are immediately added to the list displayed in the pull-down menu in the results table header. If you do not have an SNP list, the checkboxes will do nothing; you must first create an SNP list (see SNP list views section). You can select all SNPs in the current browser view by checking the checkbox next to the Total SNPs number above the results table. You can uncheck any SNP to remove it from your list.

Overall, the process of building SNP lists is quick and easy. You do not have to worry about constructing the 'perfect list'; if you make a mistake you can simply start a new list.

SNP list views

After adding SNPs to your lists using one or more of the selection pages mentioned above, you can view and manage your lists on the *My SNPs* page. The first view

is a table of all of your SNP lists, including public SNP lists created by other users. You can start a new list by clicking on '+'. You can view and edit the details of a list (e.g., it's name, description and accessibility) by clicking on the title of the list. In order to review the SNPs in a list, just click on the number of SNPs in the list.

When you click on the title of an SNP list, you are presented with the list details. You can easily change the title or description of the list here. You can also set how the list is shared with other SNPLogic users: 'Share to View' defines who can view the list and 'Share to Update' defines who can make changes to the list. The settings are: 'Private' = only you, 'Group' = only users in your group and 'Public' = any user logged into SNPLogic. Click 'Update' to save any changes.

When you click on the number of SNPs in a list, you are presented with a table of all the SNPs in the list, regardless of how they were collected (e.g. by gene, by pathway, etc.). This is the heart of SNPLogic (Figure 3). Here, you can view and filter by annotations (just like in the selection pages) across a variety of selected SNPs. You can similarly export these lists of SNPs and annotations using the Export feature. You can also sort any field by clicking on the column label; click again to reverse sort. There are four unique fields in this table: score, cnt (count), logic and note. Score will be discussed in the Score and rank section. Count tells you by how many criteria a particular SNP was selected. Logic names the selection criteria (snp = SNP; gene = Gene; pw = Pathway; chr = Chromosome; my_snp = Upload). And, Note tells you the name of the gene, pathway or chromosome region used in the selection. These fields help you track how and why a particular SNP was included in your list.

Score and rank

One of the unique and key features at SNPLogic is the ability to rank your custom SNP lists by user-defined

Figure 2. Gene Browser interface for SNP selection. User input fields and functions are highlighted on this screenshot of the web interface.

scoring rules. The goal is to create and utilize project-specific scoring systems for SNP selection (during SNP list building) and *a priori* SNP function scores (during data analysis). The *My Score* page lists the sets of scoring rules you have created, including public sets shared by other users. Click on the title of a set to view and edit the details of the set (same settings as described for SNP lists). Click on the number of rules in a set to view and edit the actual scoring rules. To create a rule, simply select any of the SNPLogic annotation fields, apply a logical operator (e.g. equal, greater, etc.), fill in a value, and then assign a score. For example, to give a higher score to SNPs with an average heterozygosity greater than 0.05, pull-down 'Average Heterozygosity', then pull-down 'greater', then enter the value '0.05', then enter a score of say '10'. Mouse-over the column labeled 'Value' to view a sample or range of values for that annotation field. The score is arbitrary and completely relative to your scoring rule set. The score on a particular SNP will simply be the additive score considering all the rules in the set you apply to your SNP list.

To apply a set of scoring rules, go back to the SNP list review page (*My SNPs* > click on number of SNPs), select the set via pull-down 'Score using' and click 'Calculate'. The 'Score' column will then display the updated scores per SNP according the sum of scoring rules defined in the set. Now you can click on the column label score to sort by score; click again to reverse sort. While SNP scores based on user-defined rules are subjective, the application of a set of intelligently constructed, consistent scoring rules can facilitate selection and interpretation of genotyping results.

Storing and sharing

Your lists are automatically saved until you choose to delete them. Share your list with colleagues in your group by editing your list's accessibility setting (*My SNPs* > click on title of list and choose appropriate sharing option, public, group or private). Perhaps a colleague would want to add his or her own favorite SNPs to the shared list. Or, perhaps they might have other ideas for scoring criteria. SNPLogic allows you to work collaboratively on SNP selection. Remember, you can export your lists as tab-delimited text files at any time to work with them in Excel or your favorite spreadsheet application (*My SNPs* > click on number of SNPs > Export).

USER SCENARIOS

We have found the SNPLogic approach to candidate SNP selection and annotation analysis critical to our own association studies of glioma biology. The SNPLogic database was, in fact, originally developed as an internal tool to meet our informatics needs. The importance of pathway-based analyses has been increasingly recognized in assessing the association between SNPs and diseases, as the occurrence of most complex diseases are likely to be influenced by multiple genes on the same biological pathway. Various analytical tools such as random forests and multifactor dimensionality reduction have been applied to pathway-based analysis (1,18). SNPLogic can serve as a useful pre-analytical tool for pathway-based analysis by grouping SNPs according to biological pathways based on information gathered from different pathway databases. In addition, SNPLogic can also provide important biological information post-analytically for SNPs that are found to be associated with the disease of interest. For example, we performed pathway-based analyses on adult glioma using a panel of 10,000 non-synonymous SNPs (1). Using SNPLogic, we selected SNPs from 10 biological pathways to conduct RF analyses, and found DNA repair to be a potentially important pathway associated with glioma risk. SNPLogic also provided information regarding the potential biological function of the top ranked SNPs in the analysis. For example, one SNP, rs12450550, was found to disrupt a potential binding site for the transcription factor MYB according to the Delta-MATCH information provided by SNPLogic. The web site was also critical to assessing commercial and custom genotyping microarray options for a follow-up study.

The following user scenarios highlight the stepwise logic comprising example workflows using SNPLogic from beginning to end.

Figure 3. Filtering and ranking SNP lists using the My SNPs interface. The main features of ranking are highlighted on this screenshot of the web interface.

User scenario #1: examine SNPs from pathways of interest

A researcher interested in a particular area of biology can often refer to one or more pathways to represent the components, interactions and processes they are investigating. Pathways also provide a powerful means of cataloging SNPs, and are thus useful in developing genotyping strategies. Pathways expand the search to biologically meaningful 'gene neighborhoods' and act as a guide for probing potential SNP interactions and generating testable hypotheses.

SNPLogic allows you to gather a broad range of annotations for SNPs associated with known, curated pathways. The associations between SNPs and pathways are made transitively, based simply on the SNP-to-gene links provided by dbSNP and the gene-to-pathway links provided by the various pathway resources (Table 1).

Login to SNPLogic. Click on *Pathway* in the top navigation bar. From the pull-down menus, select a pathway resource and then a particular pathway (GO and OMIM terms are also available). The complete list of SNPs associated with that pathway is displayed in a table. When available, a 'View' link is generated, allowing you to view the pathway at the primary resource. Click '+Fields', select all (top checkbox), and then hit 'Go' to view the available annotations for the genes and SNPs at the pathway level. These fields are helpful for sorting and filtering based on the gene, protein and SNP function information. Use the pull-down menu in the results table to select the SNP list that you would like to add SNPs to and then select individual SNPs in the table or the entire set (or filtered set) by selecting the 'Total SNPs' checkbox. After adding SNPs from the pathway, you can return to the pull-down menus near the top to select additional pathways of interest. The iterative process lets you build SNP lists by accumulation. Every time an SNP is added to a list, the source is recorded as the reason or logic for the addition. This record allows you to recall why and how an SNP was added and also provides another means of sorting, scoring and ranking.

Go to *My SNPs* via the top navigation bar and find your SNP list with the newly added pathway SNPs. Click on the number of SNPs to view the SNP table. Here, you can click '+Fields' and select from the complete list of annotation resources collected at SNPLogic for your SNPs of interest. If you added SNPs using more than one pathway, you can sort by count (descending) to view the SNPs that were added via a greater number of pathways. By defining your own scores at *My Score*, you can sort by any combination of annotations at SNPLogic. You can also combine sorting and filtering to further focus the SNP list. For example, add the 'function' field using the '+Fields' function, and then select 'function' from the 'Filter by' pull-down menu. Select 'like' and then type '%cds%' to indicate wildcard values before and after the term 'cds' in the 'function' field. This filter restricts the displayed SNPs to those that are within the coding sequence of a gene (i.e. excluding near, UTR and intron SNPs).

The combination of filtered and sorted SNPs presents a list of coding SNPs sorted by prevalence in pathways of interest. Such a list could be helpful in prioritizing SNPs to be included on a custom array panel or in assessing commercial arrays in terms of SNP coverage. SNPs collected into groups using pathway information are also useful for interaction analyses and mechanistic hypotheses following data collection.

User scenario #2: annotate custom and commercial SNP panels

A researcher interested in collecting data on SNPs included on custom or commercial arrays can use SNPLogic to annotate relevant lists of SNPs. There are in fact SNP lists already prepared and publicly accessible at SNPLogic for the ParAllele/Affymetrix 10k SNP array and the Illumina 370k array. We constructed these lists using the current annotations and tools at SNPLogic and then made the lists public on the *My SNPs* page. Any user can access these and any other public SNP lists. If you think a particular SNP list you have created would be of general utility to other users of SNPLogic, you can set the access rights to share viewing with the public. In addition to sharing, the viewing and editing of SNP lists with specific collaborators, the public access right facilitates the collection of useful lists for the entire research community.

Login to SNPLogic. Click on *My SNPs* in the top navigation bar to view the set of public SNP lists, along with any private or group lists you have access to. The public set will include 'Affy_ParAllele' and 'Illumina_370k'. Click on either of these titles to view the properties of the list. Click on the number of SNPs to view the SNP table. Alternatively, if you are working with a custom array panel, you can use the *Upload* function to build a custom SNP list and then select it from the set to get to the SNP table view. From here, you can click '+Fields' and select from the complete list of annotation resources collected at SNPLogic for your SNPs of interest. For example, selecting fields like 'function', 'MAF_CEU', 'tag_CEU' and 'tagged_CEU' provides information about an SNPs location, frequency and linkage disequilibrium that you could directly compare with your own experimental data. Selecting fields like 'polyphen', 'dm_difz', 'PupaSuite_TFBS', 'SNPs3D' and 'HCE28-Mammal' provide information about the how disruptive

Table 2. Video tutorials for SNPLogic available at http://snplogic. org/tutorials/

Title	Description	Video filename
Selection by gene	How to select SNPs by gene name	select_by_gene.mov
Selection by chromosome	How to select SNPs by chromosomal band or region	select_by_chr.mov
Selection by pathway	How to select SNPs by pathway source and name	select_by_pathway.mov
SNP list view	Working with SNP lists	snp_list_view.mov
Scoring and ranking	How to define scores and apply them to SNP lists for prioritization	scoring_ranking.mov
User Scenario #1	Examining pathway SNPs	user_scenario_1.mov
User Scenario #2	Annotating SNP panels	user_scenario_2.mov

an SNP may be to protein structure, transcription factor binding and generic sequence conservation.

Like with any SNP list, you can sort, filter, score and rank based on any of the annotation fields. You can also export the list in tab-delimited format to view in a spreadsheet application. Collecting a broad variety of public SNP data and annotations into a single table view can provide a powerful context for evaluating and interpreting your own experiment data.

Online tutorial videos demonstrate each of the main functions of SNPLogic along with a complete demonstration of the two user scenarios described above (Table 2 and Supplementary Material).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

FUNDING

REFERENCES

1. Chang,J.S., Yeh,R.F., Wiencke,J.K., Wiemels,J.L., Smirnov,I., Pico,A.R., Tihan,T., Patoka,J., Miike,R., Sison,J.D. *et al.* (2008) Pathway analysis of single-nucleotide polymorphisms potentially associated with glioblastoma multiforme susceptibility using random forests. *Cancer Epidemiol. Biomarkers Prev.*, **17**, 1368–1373.
2. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
3. Thorisson,G.A., Smith,A.V., Krishnan,L. and Stein,L.D. (2005) The international HapMap project web site. *Genome Res.*, **15**, 1592–1593.
4. de Bakker,P.I., Yelensky,R., Pe'er,I., Gabriel,S.B., Daly,M.J. and Altshuler,D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.
5. Conde,L., Vaquerizas,J.M., Dopazo,H., Arbiza,L., Reumers,J., Rousseau,F., Schymkowitz,J. and Dopazo,J. (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res.*, **34**, W621–W625.
6. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC genome browser database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
7. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
8. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.F., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
9. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
10. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
11. Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.
12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
13. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
14. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
15. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
16. Salomonis,N., Hanspers,K., Zambon,A.C., Vranizan,K., Lawlor,S.C., Dahlquist,K.D., Doniger,S.W., Stuart,J., Conklin,B.R. and Pico,A.R. (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.
17. Caspi,R. and Karp,P.D. (2007) Using the MetaCyc pathway database and the BioCyc database collection. *Curr. Protoc. Bioinformatics* Chapter 1, Unit1 1.7.
18. Duell,E.J., Bracci,P.M., Moore,J.H., Burk,R.D., Kelsey,K.T. and Holly,E.A. (2008) Detecting pathway-based gene-gene and gene-environment interactions in pancreatic cancer. *Cancer Epidemiol. Biomarkers Prev.*, **17**, 1470–1479.