W336–W340 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 25 April 2008*
doi:10.1093/nar/gkn214

GenoWatch: a disease gene mining browser for association study

Yan-Hau Chen[1], Chuan-Kun Liu[1], Shu-Chuan Chang[1], Yi-Jung Lin[1], Ming-Fang Tsai[1], Yuan-Tsong Chen[1,2] and Adam Yao[1,2,*]

[1]National Genotyping Center (NGC) and [2]Institute of Biomedical Sciences (IBMS), Academia Sinica, Taipei, Taiwan 11529, R.O.C.

Received January 28, 2008; Revised March 29, 2008; Accepted April 9, 2008

ABSTRACT

A human gene association study often involves several genomic markers such as single nucleotide polymorphisms (SNPs) or short tandem repeat polymorphisms, and many statistically significant markers may be identified during the study. GenoWatch can efficiently extract up-to-date information about multiple markers and their associated genes in batch mode from many relevant biological databases in real-time. The comprehensive gene information retrieved includes gene ontology, function, pathway, disease, related articles in PubMed and so on. Subsequent SNP functional impact analysis and primer design of a target gene for re-sequencing can also be done in a few clicks. The presentation of results has been carefully designed to be as intuitive as possible to all users.

The GenoWatch is available at the website http://genepipe.ngc.sinica.edu.tw/genowatch

INTRODUCTION

Human association studies often involve a large number of genomic markers on different chromosome regions. Researchers use these markers to locate candidate regions, and then go through a series of bioinformatic analyses of the regions to find disease-associated candidate genes. Frequently, these bioinformatic analyses require the manual gathering of information from many public websites involving several databases. Dealing with a huge volume of data from different sources is both time-consuming and error-prone, a problem that is easily magnified when many markers show significant association with a disease. Moreover, as genome assemblies and annotations are continually updated, and researchers may collect and download different versions of data over a period of months, there is a risk that some data may be outdated, invalid or inconsistent with others, especially if the major data sources are not primary sources. To alleviate this situation, we have developed GenoWatch, a real-time batch SNP (single nucleotide polymorphism) and STRP (short tandem repeat polymorphism) overview system, to effectively extract up-to-date information from public domain websites. Up to 100 markers can be processed in a batch so that researchers do not have to repetitively perform tedious information retrieval steps. GenoWatch utilizes real-time web integration to ensure that researchers obtain the same information as when manual browsing. The system greatly increases the throughput of candidate region analysis, avoids acquiring obsolete data following public database updates and reduces possible errors in manual operations.

GenoWatch is very suitable for disease candidate gene selection from candidate regions that are defined either by markers or by chromosome physical positions plus a flanking region. The system accepts two common types of genomic markers—SNPs and STRPs, and batch processes SNP inputs. An SNP marker name can be given in dbSNP RS ID or Affymetrix Probe ID. Once input, GenoWatch first locates targeted chromosome regions. Each target region may be defined and displayed by a single marker or a group of markers that are close to one another, within 1 Mbp range. Subsequently, GenoWatch extracts gene information from major public websites such as NCBI (1), UniProt (2), KEGG (3), GO (4), etc. The information available includes gene function, tissue specificity, disease, subcellular location, pathway, ontology and related PubMed articles from relevant journals (1). During processing, the system continuously reports the process status for every subtask. The system integrates extracted information from different databases into a carefully designed result page, which is displayed when all processes have been completed. For a batch task, the system positions all input markers on chromosomes and places these on an overview map called Genome View at the top of the result page, providing researchers a clear overall picture of their markers, which are colored according to impact risk (5), and of nearby genes in the whole human genome. When a marker on the Genome View is clicked,

*To whom correspondence should be addressed. Tel: +886 2 26523091; Fax: +886 27824066; Email: adam@ibms.sinica.edu.tw

a summary map, Gene View, displays not only the marker and its nearby genes but also the distance between them in the region. Markers and gene information including their positions on the current assembly, gene ontology from GO (4), pathway from KEGG (3), function and disease annotation from UniProt (2) and related articles in PubMed (1), can easily be accessed by a click. All extracted and displayed information can be linked to its original source page for verification. GenoWatch provides researchers an efficient and convenient way to analyze their markers or candidate regions in batch with associated gene annotation data and to perform downstream assays.

IMPLEMENTATION

GenoWatch, written in Java, takes advantage of Jakarta Struts framework technology to implement Model-View-Controller (MVC) architecture. To accommodate most users' needs at the GenoWatch input stage, JavaScript was used to implement a dynamic form with an interactive graph that provides various query options to designate a genome region.

To effectively extract data in real-time from different public sources, information-gathering processes are executed in parallel using multiple threads. An asynchronous process is implemented to offer users a request ID for retrieving results later, or to send email notification when the request is completed.

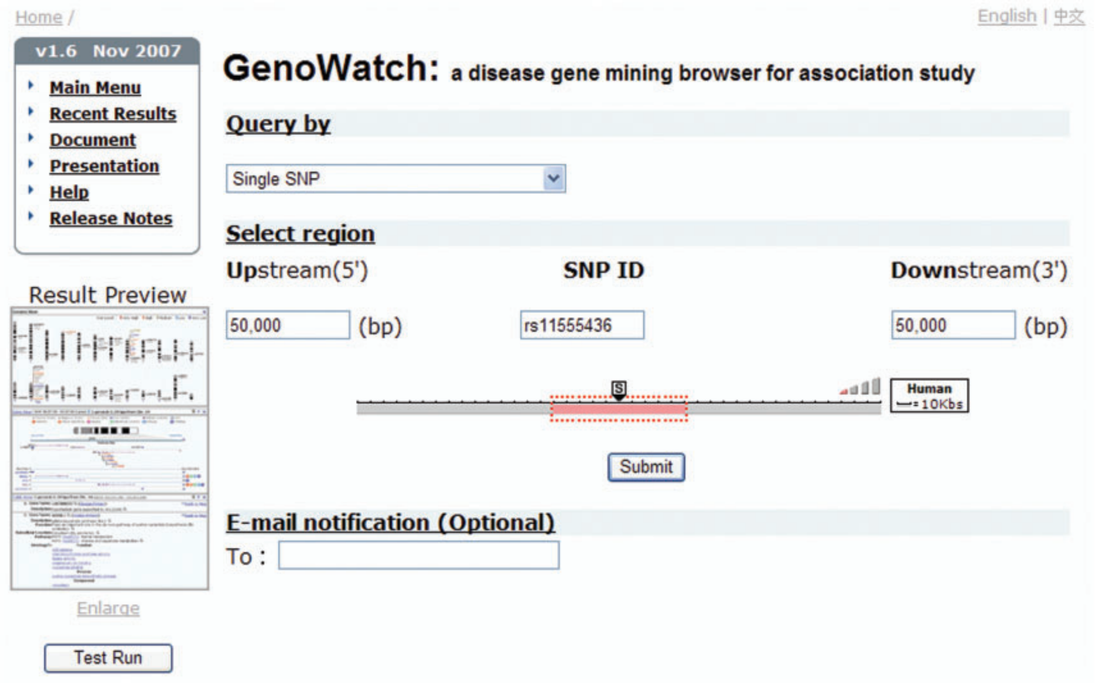
INPUT

GenoWatch can be accessed online with major web browsers. Users can opt to define a query region using chromosome positions, markers (SNP or STRP) or a batch file input. For example, they can use physical positions, a single marker (Figure 1), or a pair of markers plus downstream and upstream sequence to describe a chromosome region. When users select a query option, the content of the form will show corresponding fields with a graph illustrating the defined region for that option. The system also allows users to extend regions upstream and downstream by dragging the edges of the graph bar or by entering a specific position number in a text box. In addition, the system also accepts batch marker processing to simplify genome-wide candidate gene selection. A batch file has one marker ID or marker name per line for up to 100 markers in a text file. It should be noted that the maximum allowable length of a single candidate region is 4 Mbp, to avoid overloading the source websites. After data submission, the system will inform the user of task progress, and then display a results page after all retrieval processes are completed.

OUTPUT

The results page is comprised of Genome View (Figure 2), Gene View (Figure 3) and Table View (Figure 4). Genome View, at the top of the page, displays an overview of input markers in chromosomes from a batch input. Each marker displayed in Genome View represents a chromosome region for nearby retrieved gene information. If the marker is a SNP, its color represents the risk level of functional impact on a gene. The risk analysis currently based on global SNP information therefore does not reveal differences among populations.

Clicking on a marker leads to Gene View, showing its physical location on a chromosome, the relative positions of neighboring markers and structured genes and a gene



GenoWatch: a disease gene mining browser for association study

Query by

Single SNP

Select region

Upstream(5'): 50,000 (bp)

SNP ID: rs11555436

Downstream(3'): 50,000 (bp)

E-mail notification (Optional)

To :

Figure 1. An example to locate a chromosome region by a single SNP with 50 kb flanking sequence.

list with different levels of gene annotation information. Genes shown in both forward and reverse strands are colored blue for a single isoform and purple for multiple isoforms. In addition, pseudogenes are shown in red (for 'unknown gene structure').

The annotation for each gene is accessed through colored squares listed below the summary map. The red letter 'F' square is for 'Gene Function' and gives a general description of functions of the gene or related proteins. The orange letter 'T' describes the 'Tissue-specific

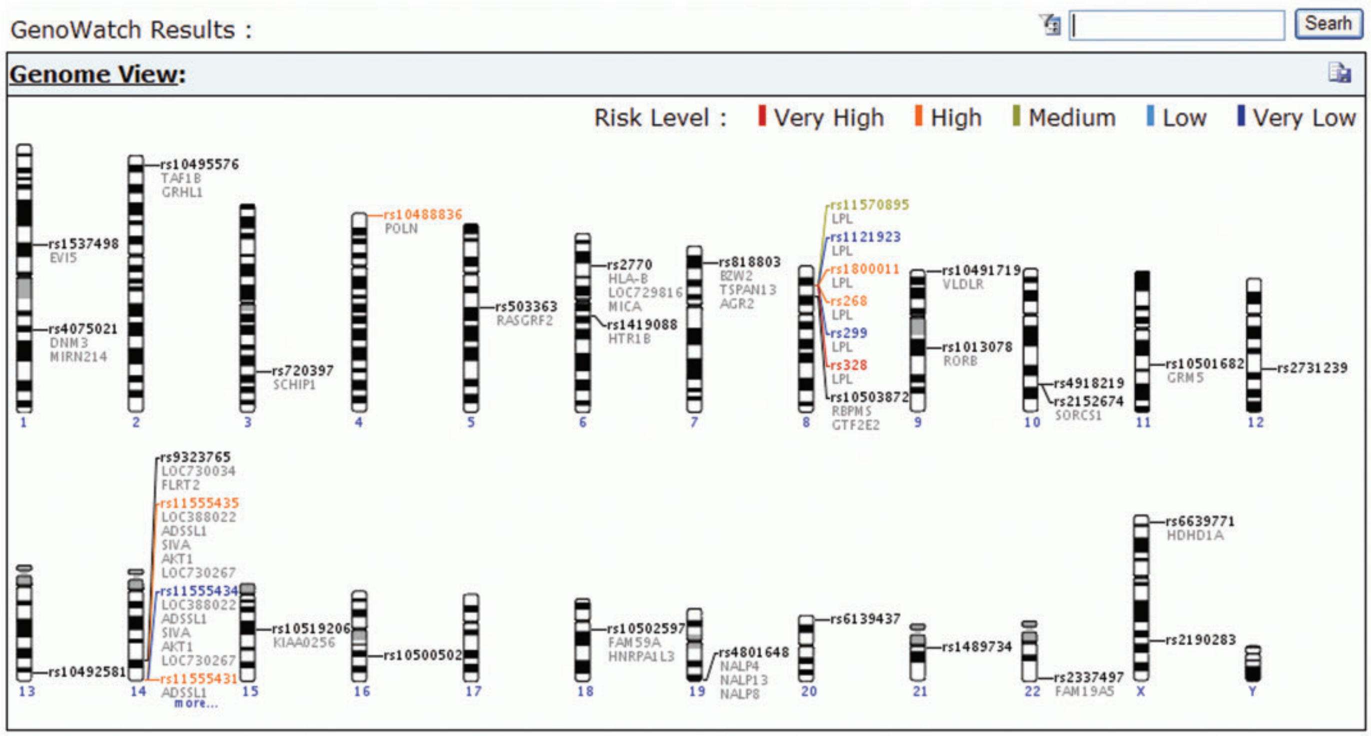
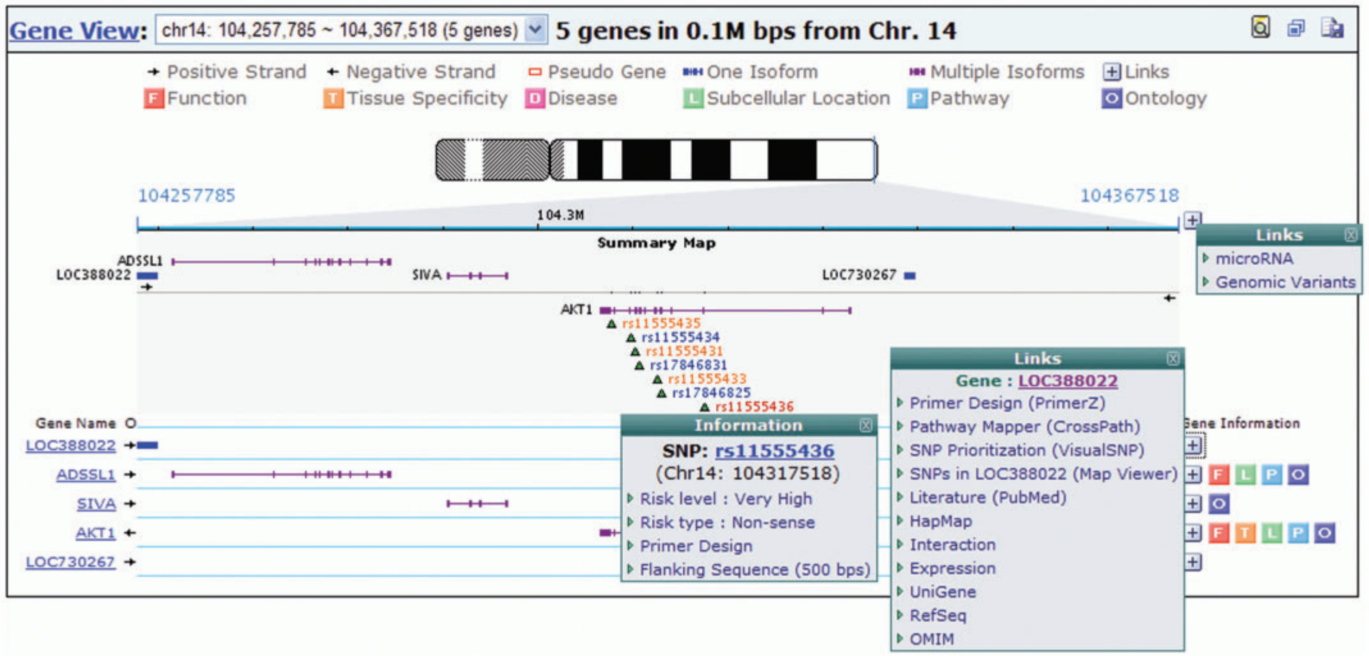


Figure 2. An overview of all queried regions in chromosomes from a batch SNP input in Genome View.



Information

SNP: rs11555436
(Chr14: 104317518)

- Risk level : Very High
- Risk type : Non-sense
- Primer Design
- Flanking Sequence (500 bps)

Links

Gene : LOC388022

- Primer Design (PrimerZ)
- Pathway Mapper (CrossPath)
- SNP Prioritization (VisualSNP)
- SNPs in LOC388022 (Map Viewer)
- Literature (PubMed)
- HapMap
- Interaction
- Expression
- UniGene
- RefSeq
- OMIM

Figure 3. Gene View displays the distribution of SNPs and genes in a candidate region with SNP functional impact risk level, gene annotation and useful links to other online systems for further analysis or primer design.

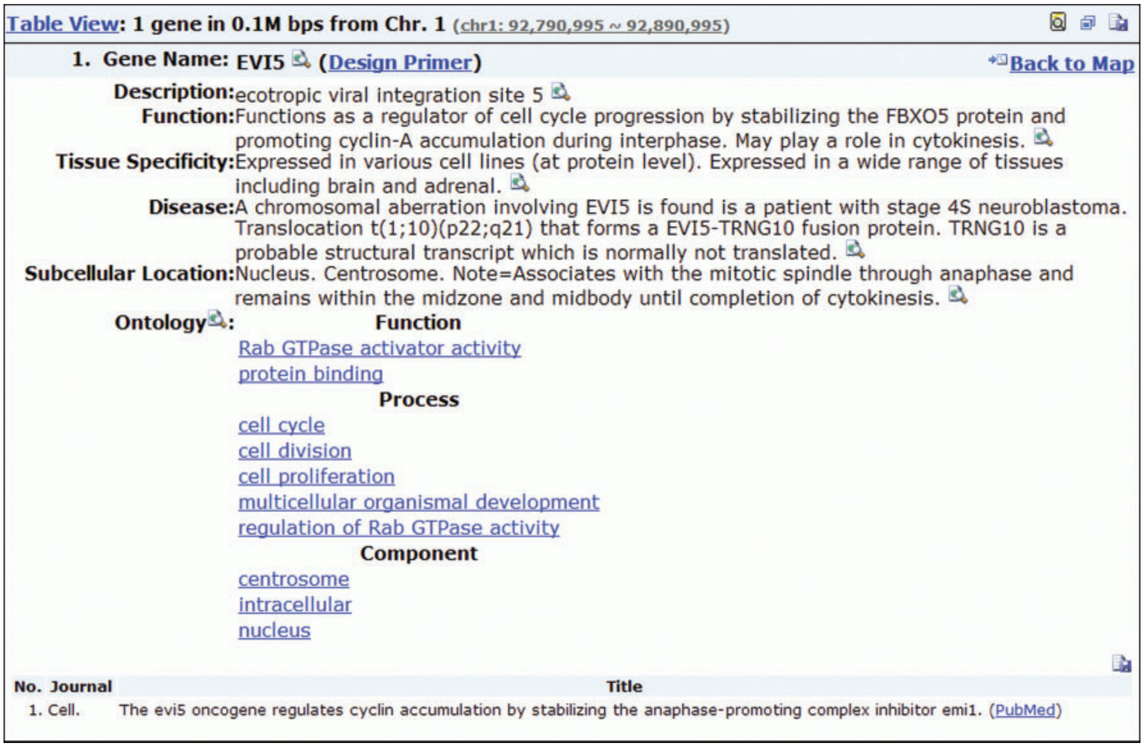


Table View: 1 gene in 0.1M bps from Chr. 1 (chr1: 92,790,995 ~ 92,890,995)

1. Gene Name: EVI5 (Design Primer) **Back to Map**

Description: ecotropic viral integration site 5

Function: Functions as a regulator of cell cycle progression by stabilizing the FBXO5 protein and promoting cyclin-A accumulation during interphase. May play a role in cytokinesis.

Tissue Specificity: Expressed in various cell lines (at protein level). Expressed in a wide range of tissues including brain and adrenal.

Disease: A chromosomal aberration involving EVI5 is found is a patient with stage 4S neuroblastoma. Translocation t(1;10)(p22;q21) that forms a EVI5-TRNG10 fusion protein. TRNG10 is a probable structural transcript which is normally not translated.

Subcellular Location: Nucleus. Centrosome. Note=Associates with the mitotic spindle through anaphase and remains within the midzone and midbody until completion of cytokinesis.

Ontology:

Function

Rab GTPase activator activity
protein binding

Process

cell cycle
cell division
cell proliferation
multicellular organismal development
regulation of Rab GTPase activity

Component

centrosome
intracellular
nucleus

No. Journal	Title
1. Cell.	The evi5 oncogene regulates cyclin accumulation by stabilizing the anaphase-promoting complex inhibitor emi1. (PubMed)

Figure 4. Table View has full text annotation and descriptions of genes.

expression' of mRNA and protein from the gene. The pink letter 'D' lists 'Diseases' associated with the gene or a deficiency of a protein from the gene. The green 'L' gives the 'Subcellular Location' of mature proteins related to the gene. The blue 'P' for 'Pathway' describes the metabolic pathway with which the gene is associated. The purple 'O' ('Gene Ontology') provides information from the GO database (4).

These colored squares provide an overview of all gene annotation in a particular region. In front of these colored squares, there is a special cross-reference square that provides links to allow the user to submit gene data directly to other online query systems, such as PrimerZ (6) for primer design of a single gene, CrossPath (7) for pathway mapping of a group of genes, VisualSNP (http://genepipe.ngc.sinica.edu.tw/visualsnp) for SNP prioritization of SNP markers or genes, HapMap (8) for haplotype analysis, and other sites for microRNA (9), genomic variants (10) searches and so on. Furthermore, clicking on any of the markers, genes or squares on the maps will lead to Table View for additional detailed information.

Table View tabulates full descriptions of genes shown in Gene View, in categories of function, tissue specificity, disease, subcellular location, pathway, ontology and related articles. These descriptions are extracted and integrated from public domain databases (Table 1) including NCBI (dbSNP, UniSTS, Gene and PubMed databases for SNP, STRP, gene and literature information, respectively) (1), UniProt (2), KEGG (3) and GO (4). In addition, a list of related articles from PubMed (1) for each gene is available in the Table View. The article titles or abstracts are searched for the gene name as a keyword in 40 prominent research journals. The journals are first selected according to their research relevance, followed by their impact factor as cited in ISI Journal Citation Report. The articles are then ranked by their score, with each gene name appearance in a title scoring 5 points and in an abstract, 1 point. All article abstracts can be directly retrieved from the system.

Table 1. All external resources used in GenoWatch

Reference information	Public websites
Global SNP marker information	NCBI dbSNP
STRP marker information	NCBI UniSTS
Gene location and structure	NCBI Entrez Gene
Gene Annotation	GO, UniProt
Literature search	NCBI PubMed
Pathway	KEGG
SNP risk analysis	VisualSNP

All result data can also be conveniently exported in CSV (comma-separated value) file format for further editing and analysis by clicking the 'download' icon on the upper right corner of each view.

CONCLUSION

GenoWatch performs multiple real-time queries of independent public databases, and presents well-organized annotated information, both known and predicted, about many genes at once. It greatly simplifies a traditionally time-consuming task by integrating batch marker input, associated gene annotation collection, SNP prioritization and primer design in a pipeline style. Future work will focus on improving data extraction efficiency by maintaining main public databases locally so that the restrictions on marker number and region length can be relaxed. We expect this system will substantially facilitate the process of finding and analyzing disease candidate genes.

ACKNOWLEDGEMENTS

Special thanks to Dr Jer-Yuarn Wu for his support of this work and to Dr Harry Wilson for editing the article. This study was funded by Academia Sinica Life Sciences (Grant No. 40-19); National Science Council (Grant no. NSC96-3112-B-001-011). Funding to pay the Open Access publication charges for this article was provided by National Science Council, Taiwan, under Grant no. NSC96-3112-B-001-011 (National Genotyping Center).

Conflict of interest statement. None declared.

REFERENCES

1. 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*. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **35**, D5–D12.
2. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
3. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
4. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D266.
5. Tabor,H.K., Risch,N.J. and Myers,R.M. (2002) Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet*., **3**, 391–397.
6. Tsai,M.F., Lin,Y.J., Cheng,Y.C., Lee,K.H., Huang,C.C., Chen,Y.T. and Yao,A. (2007) PrimerZ: streamlined primer design for promoters, exons and human SNPs. *Nucleic Acids Res*., **35**, W63–W65.
7. Chen,Y.A. and Hwang,P.I. (2006) CrossPath – mapping pathways in KEGG or Biocarta with functional genomics data. *J. Genet. Mol. Biol*., **17**, 151.
8. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
9. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, **34**, D140–D144.
10. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet*., **36**, 949–951.