Published online 28 April 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W65–W70***
doi:10.1093/nar/gks364

SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update)

Abu Z. Dayem Ullah, Nicholas R. Lemoine and Claude Chelala*

Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK

Received January 30, 2012; Revised March 26, 2012; Accepted April 10, 2012

ABSTRACT

Broader functional annotation of single nucleotide variations is a valuable mean for prioritizing targets in further disease studies and large-scale genotyping projects. We originally developed SNPnexus to assess the potential significance of known and novel SNPs on the major transcriptome, proteome, regulatory and structural variation models in order to identify the phenotypically important variants. Being committed to providing continuous support to the scientific community, we have substantially improved SNPnexus over time by incorporating a broader range of variations such as insertions/deletions, block substitutions, IUPAC codes submission and region-based analysis, expanding the query size limit, and most importantly including additional categories for the assessment of functional impact. SNPnexus provides a comprehensive set of annotations for genomic variation data by characterizing related functional consequences at the transcriptome/proteome levels of seven major annotation systems with in-depth analysis of potential deleterious effects, inferring physical and cytogenetic mapping, reporting information on HapMap genotype/allele data, finding overlaps with potential regulatory elements, structural variations and conserved elements, and retrieving links with previously reported genetic disease studies. SNPnexus has a user-friendly web interface with an improved query structure, enhanced functional annotation categories and flexible output presentation making it practically useful for biologists. SNPnexus is freely available at http://www.snp-nexus.org.

INTRODUCTION

The diversity, rapid evolution of the cutting-edge sequencing technologies and the amount of newly available sequencing data pose serious informatics challenges for screening, analysis and interpretation of genome variations and their phenotypic implications. Variations that might be functionally relevant, such as those leading to amino acid changes or splicing site alterations, found in known regulatory elements or conserved non-coding sequences are priority targets for further investigations. Single nucleotide polymorphisms (SNPs) are very common genetic variations and represent a valuable resource for investigating the genetic basis of diseases and are widely used for fine-scale genetic mapping and genome-wide association studies. The selection of an optimal number of tag SNPs based on both potential functional effect and haplotype information could increase the sensitivity and efficiency of large-scale genotyping projects (1). However, the interpretation of candidate SNP sites requires the integration of genome annotation data, such as gene(s)/protein(s) structure, and related information about the splicing isoforms with the sequence information. This is an essential step to enable and facilitate hypothesis generation for further experimentation and validation.

To meet this challenge, we have designed and implemented SNPnexus (2) as a robust functional annotation tool to assess the potential significance of candidate variants detected by sequencing and point to the gene/protein isoforms that might be phenotypically important. Since its inception, we have been receiving very positive feedback from the constantly growing, national and international SNPnexus user community. As the resource is only as comprehensive as its contents and as user-friendly as its query interface, we improved our system by taking all users' requests into consideration (Table 1) and intend to continue by accommodating new and advanced annotation features to facilitate scientific discoveries.

*To whom correspondence should be addressed. Tel: +44 207 8823570; Fax: +44 207 8823884; Email: c.chelala@qmul.ac.uk

Table 1. Summary of SNPnexus features

	Feature	SNPnexus 2008	Added in updated version
Input queries	Genome assembly	NCBI36/hg18	GRCh37/hg19
	Variations	dbSNP rs# (known), genomic position (novel)	Genomic region (known)
	Supported feature	1-base substitutions	Block substitutions, insertions/deletions (InDels), IUPAC code
Annotation categories	Query size	10 000	100 000
	Gene annotation systems	RefSeq, Ensembl, UCSC, Vega, AceView	CCDS, H-invitational
	Gene consequences	Coding, splice-site, 5′/3′-UTR, up/down-stream, intronic	Exonic, intronic, splice-site for non-coding transcripts
	Protein consequences	Synonymous, non-synonymous	Stop-gain/loss, frame-shift, peptide-shift
	Protein deleterious effects	–	SIFT predictions
	HapMap populations	CEU, YRI, CHB, JPT	ASW, CHD, GIH, LWK, MEX, MKK, TSI
	Regulatory elements	Transcription factor binding sites, first exon & promoter, miRNA & target sites, snoRNA/scaRNA	Vista enhancers, CpG islands
	Conservation scores	–	PHAST
	Disease studies	GAD	COSMIC, NHGRI-GWAS
	Structural variations	Copy number variations, inversions, deletions	Insertions, inversion breakpoints
Output	format	Html, tab-delimited text	Excel, graphical

SNPNEXUS

Since its inception, SNPnexus has established itself as a one-stop solution for novel and public domain SNP functional annotation allowing users to select from a broad range of annotation categories. While other tools and studies have mainly focused on transcript/protein level effects of SNPs based on NCBI RefSeq (3) and/or Ensembl (4) genes, SNPnexus offered the possibility of assessing additional possible roles for SNPs through VEGA (5), UCSC (6) and AceView (7) annotation systems. Furthermore, SNPnexus could investigate any potential regulatory role by computing the overlaps with regulatory elements such as conserved transcription factor binding sites (TFBS) (8), putative promoters (9), miRNAs (10), their predicted target sites (11) and snoRNAs/scaRNAs (12). SNPnexus could retrieve any related information from HapMap data and genetic association studies of complex diseases and disorders available from the genetic association database (GAD) (13), and detect any overlap with previously published genomic structural variants. Users could submit queries in single or batch mode (up to 10 000 SNPs) using dbSNP identifiers or SNP genomic coordinates on clones, contigs or chromosomes. This is very helpful for researchers who discover novel SNPs by using clones or contigs as references in sequencing data analysis. SNPnexus is the only tool allowing users to provide different formats of genomic mapping to annotate newly discovered SNPs.

Significant additions and improvements have been made to SNPnexus query system, annotation categories and output layers. The basic SNPnexus request–response architecture remains the same. An updated Perl annotation pipeline connects the variation data submitted through the front-end web interface to the back-end MySQL database, performs all the calculations on the fly and displays results back to the users. In addition to the sought annotation results, the tool provides detailed genomic locations for the submitted variations in terms of their cytogenetic and physical mapping on chromosomes/contigs. When novel SNPs are defined by genomic positions, the tool retrieves whether these are already in the public domain, in which case, it provides related links to dbSNP (14) and HapMap data (15). The enhanced version was made available in August 2011 and has been tested by different users to analyse over 4 million novel and publicly known variations. A snapshot of complete SNPnexus features are presented in Figure 1.

NEW DEVELOPMENTS

Improved query capabilities

The current SNPnexus web server accepts data submitted in three different forms: genomic position, chromosomal region or dbSNP identifier. Queries can be made in both single and batch mode. As with the previous version, users can annotate a novel variation by providing physical coordinates on the genomic clone (clone, contig or chromosome), its reference and observed alleles, and strand information.

While annotations for known SNPs can still be done by providing the dbSNP rs# number, an additional query feature has been added to annotate all known variants in a given genomic region by simply providing its start and end position on the chromosome. This can be useful when dealing with the reassessment of the functional role of HapMap and dbSNP variants within a given region, for example.

Alongside single base substitutions, SNPnexus has been upgraded to support multiple nucleotide substitutions, insertions and deletions (InDels) covering a wider range of

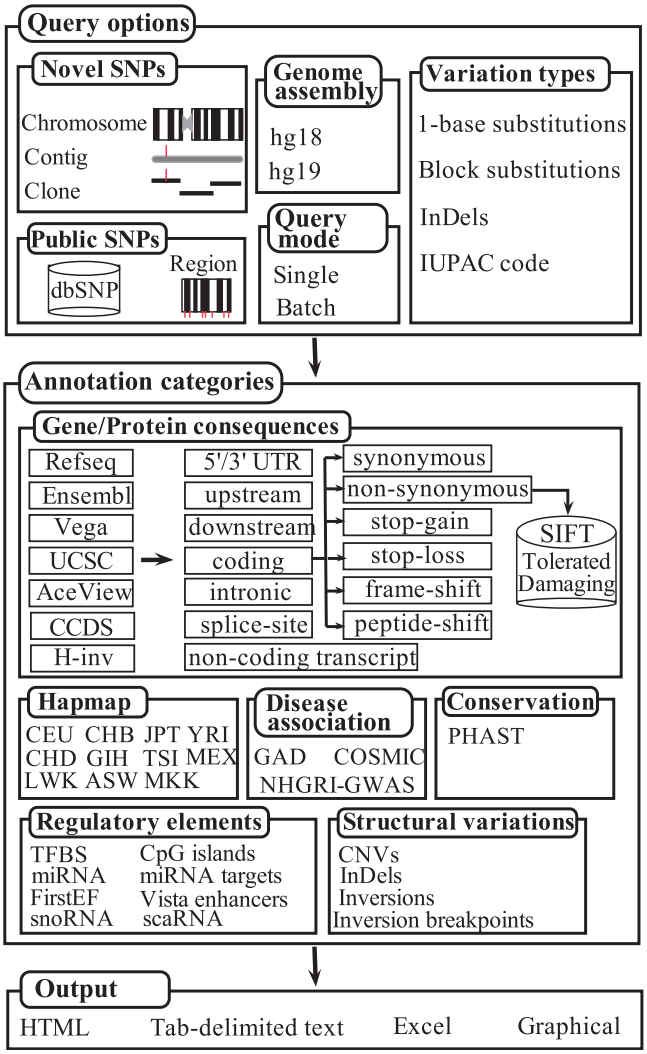

Figure 1. SNPnexus query options and annotation categories.

variation data. Users can also use the International Union for Pure and Applied Chemistry (IUPAC) code to denote ambiguous sites in a given DNA sequence motif, where a single character may represent more than one nucleotide. Allowing IUPAC code makes SNPnexus easy to use and complementary to the main genotype-calling algorithms widely used in sequencing projects.

SNPnexus accepts text-based batch query files, where each line corresponds to one genetic variant, either known or novel, including insertions, deletions or single block substitutions. It uses parallel processing to speed-up the calculation for larger queries up to 100 000 variants. In addition to the NCBI36/hg18 human genome assembly, SNPnexus now also supports the latest GRCh37/hg19 release.

With increasing interest in next-generation sequencing to rapidly discover novel variants and sub-select functionally relevant ones, allowing larger queries with improved capabilities through a web-based tool is timely for the research community.

Improved annotation categories

Enriched gene/protein consequences

In addition to the previously supported five major gene reference systems including NCBI RefSeq, Ensembl, Vega, UCSC and AceView, SNPnexus now enables users to compute functional consequences on H-Invitational (16) and CCDS (17) genes, thus providing the most extensive information on alternative splicing. For protein-coding transcripts, the predicted functional effect of a variant falls into one of the following categories: coding, splice site, 5′-UTR, 3′-UTR, upstream, downstream, intronic. In addition, SNPnexus now computes whether a variant falls into an exonic, intronic or splice-site region of a non-coding transcript.

For an intronic variant, the distance to the splicing site is reported. For a coding variant, SNPnexus reports the related base pair position within the cDNA and CDS, corresponding amino acid position in the peptide chain, the subsequent amino acid change and reference/altered protein sequences. On top of showing whether a coding variant is synonymous or non-synonymous, we report whether non-synonymous substitutions result in immediate stop-codon gain or loss. In case of an InDel or block substitution occurring within coding region, the occurrence is reported as peptide-shift or frame-shift.

For coding non-synonymous variations, SNPnexus provides the predicted deleterious effect on protein function based on SIFT predictions (18). Predictions are only shown for complete Ensembl proteins. No predictions are shown for non-synonymous substitutions resulting in stop-gain or stop-loss as these fundamentally change the protein sequence.

Updated HapMap population data

On top of the four HapMap populations available in the previous release of SNPnexus for hg18 assembly [African YRI (from Yoruba in Ibadan, Nigeria), Japanese JPT (from Tokyo, Japan), Han Chinese CHB (from Beijing, China) and European ancestry CEU (from Utah, USA)], the current version incorporates seven additional populations on hg19 assembly: African Ancestry ASW (from SouthWestern USA), Chinese Ancestry CHD (from Metropolitan Denver, USA), Gujarati Indians GIH (from Houston, USA), Luhya LWK (from Webuye, Kenya), Mexican Ancestry MEX (from Los Angeles, USA), Masai MKK (from Kinyawa, Kenya) and Toscani TSI (from Italy). SNPnexus provides both genotype and allele information, related count and frequency.

Updated regulatory data

In addition to annotations for conserved transcription factor binding sites, miRNAs, putative miRNA target sites, predicted 5′-terminal exons/promoters, SNPnexus adds information on two types of regulatory elements: CpG islands (19) and Vista Enhancers (20). CpG islands are DNA regions with high G+C content that can influence gene expression and modulate processes such as carcinogenesis (21). Vista Enhancers are the predicted non-coding distant-acting transcriptional enhancers in

the human genome identified as conserved in human, mouse and rat. With SNPnexus, one could quickly investigate whether a variant overlaps with Vista enhancers or CpG islands, therefore potentially altering the transcriptional and post-transcriptional regulation of gene expression.

Conservation analysis

This is a new addition to SNPnexus, which shows the estimated probability score that a particular variant belongs to a conserved genomic region, based on the multiple alignments of 44/46 vertebrate species using phastCons method from the PHAST package (22). Focusing on variants that fall in highly conserved genomic regions greatly helps prioritizing important candidate variants to be analysed for disease studies.

Enriched phenotype and disease association

SNPnexus allows users to establish connection with a rich collection of genetic association studies from three sources: the Genetic Association Database (GAD) (13)—an archive of human genetic association studies of complex diseases and disorders, the Catalogue Of Somatic Mutations In Cancer (COSMIC) (23)—an online database for somatic mutation information related to human cancers, and the NHGRI genome-wide association study (GWAS) catalogue (24)—a resource for mining published SNP-trait/disease associations. When investigating the role of variants, users can mine these databases and extract any information related to the gene(s)/variant(s) of interest.

Restructured and enriched structural variation data

SNPnexus has been restructured to locate variants in four types of structural variation regions from the Database of Genomic Variants (25): Copy number variations (CNVs), insertions/deletions (InDels), inversions and inversion breakpoints. The new version accommodates updated data from a large collection of peer-reviewed research studies.

Data processing from updated sources

SNPnexus is not merely a collection of annotated data sets, rather it utilizes primary annotation data sets from different sources to instantly calculate functional annotations. Primary data sets for most of the annotation categories are collected from UCSC. Currently SNPnexus maintains two separate databases for GRCh37/hg19 and NCBI36/hg18. The UCSC annotation data sets are built from MySQL tables available from the UCSC human genome annotation database (http://hgdownload.cse.ucsc.edu/downloads.html#human).
Reference human genome sequence and public domain SNP details are collected from BioMart release 63 for hg19 (ftp://ftp.ensembl.org/pub/release-63/mysql/) and release 54 for hg18 (ftp://ftp.ensembl.org/pub/release-54/mysql/). Pre-computed SIFT predictions for non-synonymous amino acid substitutions in Ensembl proteins are available from the SIFT Human Protein DB release 63 (ftp://ftp.jcvi.org/pub/data/sift/Human_db_37_ensembl_63/). The Genetic Association Database team provides the link to download GAD data file. COSMIC and miRBase data are downloaded from their corresponding FTP sites (ftp://ftp.sanger.ac.uk/pub/CGP/cosmic and http://mirbase.org/ftp.shtml). SNPnexus does not provide annotations for H-invitational gene/protein consequences and SIFT predictions on hg18 and predicted 5′ terminal exons/promoters on hg19 because of the unavailability of data in the corresponding primary data sources.

The ability to connect users' submitted queries to these data sources and compute a wide range of functional annotations on the fly makes SNPnexus a unique, timely and valuable tool.

Improved presentation of the results

The basic output remains the same with main focus on showing as detailed information as possible in the web page and providing links to the related web data sources, if available. For each selected output annotation category, results are shown in separate tables and available for download as tab-delimited text files. The new version also allows all results to be downloaded as an excel file composed of separate worksheets representing selected output annotations. From our experience with gene/protein consequences analysis, users are often interested to get not only the specific amino acid changes but also the altered protein sequences to be used for further investigation. Here, the downloadable text and excel files contain the reference and altered (if non-synonymous) protein sequences for coding variants (not available in the web page). For the gene consequences category, we have an additional graphical representation of the distribution of predicted functional consequences. This is particularly useful for variation analysis within a genomic region, where users could assess the relative functional importance of the region.

The user notification system has been improved as well. Users are no longer required to provide their email address with submitted queries. After submission, users are immediately notified of the current status of the query in the result page and can visit the page any time to check the query status. The results are accessible via the same page once the analysis is completed. Due to the huge amount of data processed every day, the results are kept and made available for 72 hours. If a user provides valid email address, then the notification of acceptance and completion are sent via email.

DISCUSSION

A number of web-based applications are available for SNP functional annotation such as SNAP (26), Pupasuite (27), FASTSNP (28), SCAN (29), but none of those tools independently support as elaborate annotation categories as SNPnexus. We allow users to view and navigate through large sets of high-dimensional data from seven major gene/protein annotation systems. Out of these, only Ensembl and Refseq are mainly used by other tools. Most importantly, SNPnexus is the only tool providing annotations for both novel and publicly

known variants, whereas all the other tools only focus on already known SNPs and do not support InDels or block substitutions. FASTSNP does provide an option for novel SNP annotations with single base substitutions, but one has to provide the whole sequence of interest and understandably the results are unrelated to any known genes. To the best of our knowledge, ANNOVAR (30) is the only tool that provides a similar broader support for functional annotations of both novel and known variants. However, unlike SNPnexus which is an online tool designed to be used by biologists, ANNOVAR is a command-line software tool that requires users to have enough programming expertise to prepare the underlying annotation databases, install/run the tool on users' own machine and parse the output text files to comprehend large annotation results. ANNOVAR accepts variation data submission only in chromosomal coordinates format whereas SNPnexus allows additional contig or clone coordinates for novel variants, dbSNP identifiers for known SNPs and IUPAC code. Unlike ANNOVAR that provides outputs only in text files, SNPnexus offers maximum flexibility to the users with additional html tables and excel files for quick inspection of the results. SNPnexus is therefore the only available tool that can handle a wider variety of functional annotations for known and novel genetic variants in a comprehensive online manner alleviating the need for bioinformatics expertise and limitations due to available memory/disk space on users' machines.

The amount of sequence data and detected variants continues to grow. Functional annotation of newly detected variants is needed to make sense of generated data and address related functional impact. There is an increased demand for bioinformatics tools and methods to maximize research outcome. With time, more researchers will move ahead with large sequencing projects and making sense of variation data will become dominant in the coming decade. SNPnexus has already been proven as a pioneer in that respect by providing the most elaborate set of annotation options for diverse range of variation data. The underlying system architecture is modularized making it easily extendable to include additional functional annotations. We welcome feedback from the user community on new annotation systems that could be added to the future releases of SNPnexus and are happy to support researchers in using the new version.

ACKNOWLEDGEMENTS

The authors would like to thank the Ensembl and COSMIC teams for helpful advice.

FUNDING

Funding for open access charge: Cancer Research UK [programme grant reference 15310].

Conflict of interest statement. None declared.

REFERENCES

1. Wiltshire,S., de Bakker,P.I. and Daly,M.J. (2006) The value of gene-based selection of tag SNPs in genome-wide association studies. *Eur. J. Hum. Genet.*, **14**, 1209–1214.
2. Chelala,C., Khan,A. and Lemoine,N.R. (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain Single Nucleotide Polymorphisms. *Bioinformatics*, **25**, 655–661.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
4. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
5. Wilming,L.G., Gilbert,J.G., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
6. Hsu,F., Kent,W.J., Clawson,H., Kuhn,R.M., Diekhans,M. and Haussler,D. (2006) The UCSC known genes. *Bioinformatics*, **22**, 1036–1046.
7. Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl. 1)**, S12.
8. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhäuser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
9. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.
10. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
11. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
12. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
13. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
14. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
15. The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
16. Yamasaki,C., Murakami,K., Takeda,J., Sato,Y., Noda,A., Sakate,R., Habara,T., Nakaoka,H., Todokoro,F., Matsuya,A. *et al.* (2010) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res.*, **38**, D626–D632.
17. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
18. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
19. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
20. Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
21. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
22. 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.

23. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
24. Hindorff,L.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
25. 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.
26. Li,S., Ma,L., Li,H., Vang,S., Hu,Y., Bolund,L. and Wang,J. (2007) Snap: an integrated SNP annotation platform. *Nucleic Acids Res.*, **35**, D707–D710.
27. 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.
28. Yuan,H.Y., Chiou,J.J., Tseng,W.H., Liu,C.H., Liu,C.K., Lin,Y.J., Wang,H.H., Yao,A., Chen,Y.T. and Hsu,C.N. (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Res.*, **34**, W635–W641.
29. Gamazon,E.R., Zhang,W., Konkashbaev,A., Duan,S., Kistner,E.O., Nicolae,D.L., Dolan,M.E. and Cox,N.J. (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.
30. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Res.*, **38**, e164.