BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 5 2008, pages 655–661
doi:10.1093/bioinformatics/btn653

Databases and ontologies

SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms

Claude Chelala*, Arshad Khan and Nicholas R. Lemoine

Centre for Molecular Oncology and Imaging, Institute of Cancer & CR-UK Clinical Centre, Barts & The London School of Medicine (QMUL), Charterhouse Square, London EC1M 6BQ, UK

Received on September 29, 2008; revised on November 25, 2008; accepted on December 17, 2008

Advance Access publication December 19, 2008

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Design a new computational tool allowing scientists to functionally annotate newly discovered and public domain single nucleotide polymorphisms in order to help in prioritizing targets in further disease studies and large-scale genotyping projects.

Summary: SNPnexus database provides functional annotation for both novel and public SNPs. Possible effects on the transcriptome and proteome levels are characterized and reported from five major annotation systems providing the most extensive information on alternative splicing. Additional information on HapMap genotype and allele frequency, overlaps with potential regulatory elements or structural variations as well as related genetic diseases can be also retrieved. The SNPnexus database has a user-friendly web interface, providing single or batch query options using SNP identifiers from dbSNP as well as genomic location on clones, contigs or chromosomes. Therefore, SNPnexus is the only database currently providing a complete set of functional annotations of SNPs in public databases and newly detected from sequencing projects. Hence, we describe SNPnexus, provide details of the query options, the annotation categories as well as biological examples of use.

Availability: The SNPnexus database is freely available at http://www.snp-nexus.org.

Contact: claude.chelala@cancer.org.uk

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are simple and very common genetic variations in a single base of the four-bases genome sequence. They 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. It is well established that certain SNPs may predispose to diseases such as diabetes and cancer, as well as affecting disease progression. Based on linkage analysis results and/or knowledge about the disease pathways, genes are selected for sequencing and screening for novel SNPs. Variations that might be functionally relevant, such as those that lead to amino acids changes, affect splicing sites, are found in known regulatory elements or in conserved non-coding sequences are priority targets in disease studies and large-scale genotyping projects.

Functional annotation data for publicly known SNPs are scattered across many different databases, redundant and far from complete. They include a limited set of information and no options to allow functional annotations of newly detected SNP. When dealing with novel SNPs, researchers are faced with exploring many different data sources, deciding which one to use and manually assessing the functional relevance of their variants before selecting the subset to use in future disease genotyping studies. As a result, the task of making a complete assessment of the potential functional consequences of both novel and known SNPs from sequencing or genome-wide association studies is difficult and time consuming. The lack of suitable user-friendly tools for assessing the full impact of SNPs is evident from numerous publications where the potential functional role of SNPs was not properly explored. This could compromise the design and results of any further experimental investigation.

Here, we present SNPnexus, a one-stop solution for novel and publicly known SNPs functional annotation. SNPnexus allows single or batch queries using dbSNP identifiers or SNP genomic coordinates on clones, contigs as well as chromosomes. From our experience, this is particularly needed to complete public SNP annotations. Also, this is very helpful for researchers discovering novel SNPs by using clones or contigs as a reference sequence in their analysis of sequencing data.

Because it is well known that each gene prediction program has its own limitations that often lead to different results, SNPnexus computes a wide range of possible effects for SNPs through the main gene annotation systems: NCBI RefSeq (Pruitt *et al*., 2007), UCSC Known Genes (Hsu *et al*., 2006), Ensembl (Hubbard *et al*., 2007), the Vertebrate Genome Annotation (VEGA) (Ashurst *et al*., 2005) and AceView (Thierry-Mieg and Thierry-Mieg, 2006). While most other tools and studies mainly focused on the effects of SNPs based on NCBI RefSeq and/or Ensembl genes, SNPnexus offers for the first time the possibility of assessing additional possible roles for SNPs through VEGA, UCSC and AceView annotations. The NCBI RefSeq is a high-quality, manually curated gene set. While the Ensembl gene prediction dataset is computationally derived at the genome level, VEGA's annotations are manually curated on more limited regions. The UCSC set of predictions is moderately conservative, based on data from RefSeq, GenBank (Benson *et al*., 2005) and UniProt (Bairoch *et al*., 2005). The AceView gene dataset is produced from public experimental cDNA sequences from the same species. These gene annotation sets have different prediction methods and therefore show different information on alternative splicing. By integrating all these annotated datasets,

*To whom correspondence should be addressed.

SNPnexus gives the broadest view of the potential functional role of SNPs. The functional annotations through SNPnexus are detailed on the different isoforms and should be considered carefully in any SNP experimental functional assay. Furthermore, SNPnexus investigates any potential regulatory role by computing the overlaps with regulatory elements such as conserved transcription factor binding sites (TFBS) (Wingender *et al*., 2001), microRNA (miRNA) (Griffiths-Jones, 2004; Grimson *et al*., 2007; Lewis *et al*., 2003, 2005; Weber, 2005) and promoter predictions (Davuluri *et al*., 2001). SNPnexus summarizes any related information from genetic association studies of complex diseases and disorders available from the genetic association database (GAD) (Becker *et al*., 2004), overlaps with genomic structural variability (Conrad *et al*., 2006; Hinds *et al*., 2006; Iafrate *et al*., 2004; Locke *et al*., 2006; McCarroll *et al*., 2006; Redon *et al*., 2006; Sebat *et al*., 2004; Sharp *et al*., 2005; Tuzun *et al*., 2005), dbSNP (Sherry *et al*., 2001) as well as HapMap (Frazer *et al*., 2007) genotype and allele frequency.

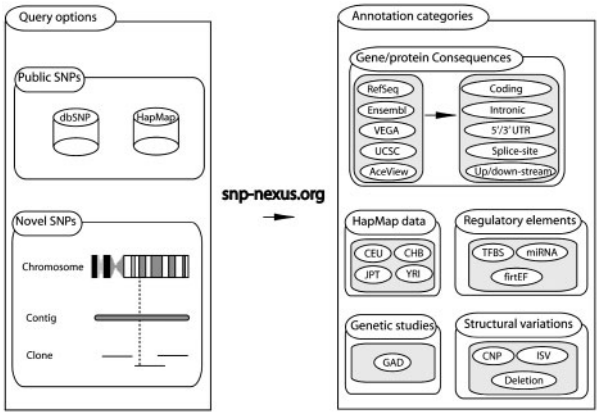

Fig. 1. SNPnexus query options and annotation categories.

2 METHODS

SNPnexus architecture has two layers: an access layer with a web server and a storage layer with a MySQL database server. We wrote a Perl annotation pipeline to connect the user data to the source database and perform all the calculations displayed in the web interface. Primary data sources are Ensembl and UCSC. We used the MySQL tables from the March 2006 GenBank freeze assembled by NCBI (hg18, Build 36, http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/) and EnsemblMart_48 (ftp://ftp.ensembl.org/pub/release-48/mysql/ensembl_mart_48). The Genetic Association Database team provided the link to download GAD data file. miRBase data were downloaded from the database FTP site (http://microrna.sanger.ac.uk/sequences/ftp.shtml) Structural variability data were derived from the Database of Genomic Variants (DGV) (http://projects.tcag.ca/variation/) via UCSC. SNPnexus will be updated on a regular monthly basis.

3 RESULTS

3.1 Queries

Queries can be made in both single and batch mode (10 000 SNPs). Users can annotate a novel SNP by providing its position on the genomic clone (clone, contig or chromosome), its alleles and strand. This is particularly useful when SNPs are detected by sequencing many DNA samples and aligning experimental results against a partial clone, contig or chromosome reference sequence. Users can also query SNPnexus for known SNPs by providing the rs# number. This could be used when dealing with reassessment of the functional role of HapMap and dbSNP variants, for example. SNPnexus query options and annotation categories are presented in Figure 1.

3.2 Annotation categories and examples

3.2.1 Gene SNP consequences To the best of our knowledge, SNPnexus is the only tool that provides a comprehensive overview of potential functional consequences of SNPs on alternatively spliced genes by exploring five different transcriptome and proteome models: NCBI RefSeq (Pruitt *et al*., 2007), UCSC (Hsu *et al*., 2006), Ensembl (Hubbard *et al*., 2007), VEGA (Ashurst *et al*., 2005) and AceView (Thierry-Mieg and Thierry-Mieg, 2006). While the gene annotation from NCBI RefSeq and Ensembl are widely used gene reference systems, they do not properly reflect transcript variation due to alternative splicing. The number of UCSC transcripts is 56 722, which is almost equal to the number of Ensembl transcripts (56 155) and double of the number for RefSeq (25 078). VEGA has over three times more transcript variants than RefSeq (88 828). Gene model in AceView provides more extensive analysis of alternative splicing. Over the entire genome the number of AceView transcripts (258 618) exceeds Ensembl and UCSC by a factor of four, VEGA by a factor of three and RefSeq by a factor of 10. Potential effects of SNPs fall into seven categories: intronic, 5′UTR, 3′UTR, 5-upstream, 3-downstream, splicing site or coding (synonymous or non-synonymous). For intronic SNPs, the distance to the splicing site is reported. For coding SNPs, the coordinates of the position within the cDNA, coding sequence (CDS), amino acid as well as the peptide change produced (if any) are reported. For 5-upstream and 3-downstream, a distance of 2 kb from the UTR start or end is considered appropriate. As a first example, we explored the SNPs of DNA repair genes that were studied in the context of overall survival of pancreatic cancer patients after treatment (Li *et al*., 2007a). A strong effect of the polymerase β (POLB) gene variants on overall survival was reported. Also, a weak but significant effect on overall survival was demonstrated for 8-oxo-guanine DNA glycosylase (hOGG1), apurinic/apyrimidinic endonuclease I (APEX1) and X-ray repair cross-complementing protein 1 (XRCC1) polymorphisms. We considered all the variants in these genes studied in the article (Table 1). The authors reported that POLB A165G and T2133C variants were intronic with the dbSNP reference rs2272615 and rs2953993, respectively. Also, the first variant of hOGG1 G2657A was intronic (rs293794) and the second variant C315G was coding non-synonymous (S326C) or 3′UTR (rs1052133). The APEX1 variant D148E (rs3136820) as well as XRCC1 variants R194W (rs1799782) and Q399R (rs25487) were coding non-synonymous. Using SNPnexus, we were able to assess in details the functional consequences of these variants on each isoform from RefSeq, Ensembl, Vega, AceView and UCSC. First, we noted that hOGG1 variants overlap with isoforms of the neighboring gene calcium/calmodulin-dependent protein kinase I (CAMK1). Similarly, APEX1 variant rs3136820 (ID used in the original paper, currently merged to dbSNP rs1130409) could have a 5-upstream effect on the neighboring gene O-sialoglycoprotein endopeptidase (OSGEP) and 3-downstream effect on transmembrane protein 55B (TMEM55B). We also found many additional potential regulatory roles and coding non-synonymous changes leading to additional

Table 1. SNP functional consequences

Panel A: as reported in previous studies

dbSNP	Chrom	Gene	Functional consequences
rs293794	3	hOGG1	I
rs1052133	3	hOGG1	3′UTR, NS (S326C)
rs3136820	14	APEX1	NS (D148E)
rs2272615	8	POLB	I
rs2953993	8	POLB	I
rs1799782	19	XRCC1	NS (R194W)
rs25487	19	XRCC1	NS (Q399R)
rs2248690	3	AHSG	Promoter
rs4918	3	AHSG	Thr/Ser
rs1071592	3	AHSG	Thr/Thr

Panel B: computed using SNPnexus

dbSNP	Gene	Source	Transcripts	Functional consequences [peptide shift (if any), number of transcripts]
rs293794	hOGG1	RefSeq	5	
		Ensembl	6	
		VEGA	6	I
		UCSC	5	
		AceView	8	
	CAMK1	RefSeq	1	I (1)
		Ensembl	2	I (2)
		VEGA	2	I (2)
		UCSC	3	I (2); 5up (1)
		AceView	6	I (4); 5up (1); 5′UTR (1)
rs1052133	hOGG1	RefSeq	8	NS (S326C, P332A, 2); 3′UTR (1); I (5)
		Ensembl	11	NS (S326C, P119A, P332A, 3); 3′UTR (1); I (6); 3down (1)
		VEGA	9	NS (S326C, P332A, 3); I (6)
		UCSC	11	NS (S326C, P332A, P97A, P41A, S91C, 5); I (5); 3'UTR (1)
		AceView	19	NS (S326C, S115C, S42C, P98A, P103A, P332A, 6); I (8); 3′UTR (4); 3down (1)
	CAMK1	RefSeq	1	
		Ensembl	2	
		VEGA	2	3down
		UCSC	3	
		AceView	8	
rs1130409	APEX1	RefSeq	3	NS (D148E, 3)
		Ensembl	2	NS (D148E, 2)
		VEGA	3	NS (D148E, 2); I (1)
		UCSC	4	NS (D148E, 4)
		AceView	22	NS (D148E, D124E, D130E, D131E, 13), I (4); 3′UTR (1); 5′UTR (1); 3down (3)
	NP	RefSeq	–	
		Ensembl	1	I (1)
		VEGA	–	
		UCSC	–	
		AceView	–	
	OSGEP	RefSeq	1	
		Ensembl	1	
		VEGA	1	5up
		UCSC	1	
		AceView	5	
	TMEM55B	RefSeq	2	
		Ensembl	2	
		VEGA	2	3down
		UCSC	2	
		AceView	6	
rs2272615	POLB	RefSeq	1	I (1)
		Ensembl	1	I (1)

(continued)

Table 1. Continued

Panel B: computed using SNPnexus

dbSNP	Gene	Source	Transcripts	Functional consequences [peptide shift (if any), number of transcripts]
		VEGA	5	I (5)
		UCSC	2	I (2)
		AceView	8	3′UTR (1); 3down (1); I (6)
rs2953993	POLB	RefSeq	1	I (1)
		Ensembl	1	I (1)
		VEGA	6	I (5); 5up (1)
		UCSC	2	I (2)
		AceView	8	I (8)
rs1799782	XRCC1	RefSeq	1	NS (R194W, 1)
		Ensembl	1	NS (R194W, 1)
		VEGA	5	NS (R194W, R157W, 5)
		UCSC	1	NS (R194W, 1)
		AceView	9	NS (R30W, R89W, R157W, R169W, R213W, R219W, R273W, 7); 5′UTR (2)
rs25487	XRCC1	RefSeq	1	NS (Q399R, 1)
		Ensembl	1	NS (Q399R, 1)
		VEGA	5	NS (Q399R, 1); 3down (4)
		UCSC	1	NS (Q399R, 1)
		AceView	9	NS (R126G, Q181R, Q235R, Q332R, Q422R, Q478R, 6); 3down (3)
rs2248690	AHSG	RefSeq	1	5up
		Ensembl	1	
		VEGA	2	
		UCSC	4	
		AceView	8	
rs4918	AHSG	RefSeq	1	NS (S256T, 1)
		Ensembl	1	NS (S256T, 1)
		VEGA	4	NS (S256T, S255T, S9T, 3); 5′UTR (1)
		UCSC	3	NS (S255T, S256T, S257T, 3)
		AceView	7	NS (S219T, S228T, S255T, S256T, S257T, 5); 3down (2)
rs1071592	AHSG	RefSeq	1	S (1)
		Ensembl	1	S (1)
		VEGA	4	NS (H6P, 1), S (2); 3down (1)
		UCSC	3	S (3)
		AceView	6	S (5); 3down (1)

I = intronic, 5up = 5-upstream, 3down = 3-downstream, NS = coding non-synonymous, S = coding synonymous.

amino acid changes (Table 1). We noted that these effects were also missed in other papers where some of the variants were investigated in additional cancers (Huang *et al*., 2007; Sangrajrang *et al*., 2008). SNPnexus enabled discovery of new potential functions that should be considered in any further analysis. As a second example, we considered the 2-Heremans-Schmid glycoprotein (AHSG), a candidate for type 2 diabetes (T2D) susceptibility. Using direct sequencing of the AHSG promoter region and exons, a recent study identified nine common SNPs with a minor allele frequency (MAF) >5%, assessed their consequences using NCBI RefSeq AHSG mRNA sequence (NM_001622) and carried out a detailed genetic association study of their contribution to the genetic susceptibility of T2D in French Caucasians (Siddiq *et al*., 2005). Their results showed that the major allele of a synonymous coding SNP (rs1071592) presented significant evidence of association with T2D. Also, one promoter variant and one coding non-synonymous variant (rs2248690 and rs4918) in strong linkage disequilibrium (LD) with rs1071592 showed evidence-approaching significance. We reanalyzed the consequences of these three SNPs on all AHSG transcripts using SNPnexus. We confirmed the synonymous coding consequence of rs1071592. However, we found that this SNP can have additional 3-downstream and most importantly synonymous non-coding (H6P) consequences (Table 1). For rs4918, we confirmed the non-synonymous coding S256T and found additional potential effects. We confirmed the promoter location for the SNP rs2248690 in all transcripts annotations (Table 1). Again, these novel potential functional roles should be included in any experimental assay investigating the role of these variants in T2D.

3.2.2 HapMap population data The HapMap Project provides a resource of genotypic data of ∼4 million common SNPs in lymphoblastoid cell lines derived from four human populations: African YRI (from Yoruba in Ibadan, Nigeria), Japanese JPT (from Tokyo, Japan), Han Chinese CHB (from Beijing, China) and European CEU (from Utah, USA with ancestry from northern and western Europe) (Frazer *et al*., 2007). The HapMap genotypic data have proven to be a key resource for researchers investigating the genetic contribution to human diseases, variation in gene expression and drug response (Zhang *et al*., 2008). The HapMap data could

be used to estimate MAF of SNPs in the different populations. This is essential since common SNPs having a MAF value ≥5% or 1% are mainly targeted in genetic studies. This hypothesis has been challenged by studies reporting that rare variants may also influence common diseases (Cohen *et al*., 2004; Romeo *et al*., 2007). Rare SNPs might also be useful when the disease-causing variant is also rare. An estimation of the MAF in the different populations could also help in assessing the association between SNPs and a given disease. SNPnexus provides both genotype and allele information, count and frequency. If $f(11)$, $f(12)$ and $f(22)$ are the HapMap frequencies of the three genotypes at a given locus with two alleles (1 and 2), then the frequency p of allele 1 and the frequency q of allele 2 are: $p = f(11) + 0.5 \times f(12)$ and $q = f(22) + 0.5 \times f(12)$. MAF is simply the smaller of these two frequencies. As an example of use, we considered the association between the +1858C > T SNP (rs2476601) in PTPN22 and type 1 diabetes (T1D) that has been reported in separate studies among different populations (Chelala *et al*., 2007; Smyth *et al*., 2004). However, the +1858T allele was not found in a large Asian cohort (Kawasaki *et al*., 2006), which raises the possibility that rs2476601 is not associated with T1D in Asian populations. It is also possible that other potentially functional variants in PTPN22 may be responsible for the susceptibility to T1D in these populations or the PTPN22 locus contains another functional variant in LD with rs2476601. When querying HapMap data for rs2476601, SNPnexus confirmed that the +1858T allele was not present in the HapMap Asian populations but also showed that the same hypothesis could apply for the African population since the +1858T allele was not found in the HapMap Nigerian population.

3.2.3 Regulatory elements SNPnexus can be queried against any overlap with conserved TFBS (Wingender *et al*., 2001), First-Exon promoter predictions (Davuluri *et al*., 2001) and miRNA (Griffiths-Jones, 2004; Grimson *et al*., 2007; Lewis *et al*., 2003, 2005; Weber, 2005). One could quickly investigate the SNP putative phenotypic effect and prioritize analysis of SNPs that disrupt TFBS relevant to the transcriptional pathways of the disease of interest or the specific tumor subtype. As the overlap with TFBSs is only putative, we made the prediction more likely to be of biological importance by restricting the Transfac Matrix Database (v7.0) TFBSs to those conserved in the human/mouse/rat alignment (http://genome.ucsc.edu/cgi-bin/hgTrackUi?g=tfbsConsSites). Promoter polymorphisms that could alter gene expression can be also investigated using SNPnexus. Overlaps are checked against the FirstEF (First-Exon Finder), a 5′-terminal exon and promoter prediction program. Also, one could investigate if a particular SNP of interest is a potential miRSNP occurring in putative miRNA target sites and hence could interfere in miRNA-mediated gene regulation. miRNAs bind to target sites in the 3′UTR region of mRNA and act as repressors of protein-coding genes or activators of RNA degradation. Aberrant expression of miRNAs may be involved in human diseases, including cancer. As an example, we chose the B-cell translocation gene-2 (BTG2), a potential target of cellular miRNAs with deregulated expression in breast cancer (Kawakubo *et al*., 2004; Yoon and De Micheli, 2005). We investigated if any SNP overlap with miRNAs that target BTG2 and thus could play a role in the aberrant regulation of BTG2 expression in breast carcinoma. We used SNPnexus with BTG2 SNPs as input and searched for overlaps with miRNA. SNPnexus prioritized BTG2 miRSNP rs1804734 showing that it overlaps with targetscan BTG2:miR-21. Added to the fact that miR-21 was reported to be upregulated in breast cancer (Iorio *et al*., 2005), the role of rs1804734 in breast cancer is worthy of investigation to see whether it could interfere with BTG2:miR-21 processing, expression and/or binding to target BTG2 mRNA.

3.2.4 Genetic association database GAD is an archive of published scientific papers on human genetic association studies of complex diseases and disorders (Becker *et al*., 2004). When investigating the role of variants, the user can mine GAD and extract any related information related to the gene of interest.

3.2.5 Structural variability Human DNA-level variation ranges from bi-allelic SNPs to large-scale structural variation [such as deletions, duplications, insertions, inversions and large-scale copy-number variants (CNVs) also called copy-number polymorphisms (CNPs)]. Structural variants can have a direct effect on gene dosage or an indirect effect on expression of a gene by changing its position on the genome (Feuk *et al*., 2006). They can predispose to additional damaging structural changes or function as susceptibility alleles in complex genetic diseases (Feuk *et al*., 2006). A recent study showed the importance of CNVs in complex phenotypes, and the need to assess it independently from SNPs (Stranger *et al*., 2007). Also, the analysis of SNPs within genes with CNV could be technically challenging. Furthermore, it is important to distinguish abnormal genetic lesions from normal CNP since some of the genetic differences in tumor genomes when analyzed in comparison to an unrelated normal genome could be due to germline CNPs. SNPnexus allows integration of both types of variation and assessment of the hypothesis cited above. This could be useful for genes and SNPs located within a CNV. In this case, both gene copy number and the genotyping of relevant SNPs should be investigated in order to understand the phenotypic impact of genome variation on the disease of interest. SNPnexus assesses overlaps with putative CNPs and sites of intermediate-sized structural variation (ISV) determined by various methods. This covers deletions from genotype analysis using the HapMap Phase I data, release 16a (McCarroll *et al*., 2006), HapMap Phase I data, release 16c.1, CEU and YRI samples (Conrad *et al*., 2006) and from haploid hybridization analysis in 24 unrelated individuals from the Polymorphism Discovery Resource, selected for SNP LD study (Hinds *et al*., 2006). This also include CNP from BAC microarray analysis in a first population of 55 individuals (Iafrate *et al*., 2004), a second population of 47 individuals (Sharp *et al*., 2005) as well as in the HapMap Phase II data (Redon *et al*., 2006). Further CNP regions are identified using array CGH in the HapMap populations (269 individuals) (Locke *et al*., 2006) and using representational oligonucleotide microarray analysis (ROMA) in a population of 20 normal individuals (Sebat *et al*., 2004). ISV sites are detected by mapping paired-end sequences from a human fosmid DNA library (Tuzun *et al*., 2005). As an example, we reused AHSG SNPs rs1071592, rs2248690 and rs4918 (Table 1). SNPnexus found an overlap with a CNV region thus suggesting that the genotypes of rs1071592 and rs4918 as well as the overlapping CNV should both be investigated in order to understand the phenotypic impact of these genome variations on T2D in French Caucasians. Similarly, rs2476601 in PTPN22 overlaps with a CNV that should be also studied in T1D.

4 DISCUSSION AND CONCLUSION

Both focused candidate-gene and whole-genome studies require the selection of an optimal number of tagSNPs. This is an important, challenging task that will affect the study outcome and cost, especially for complex diseases where the number of candidate genes is expected to be large. SNPnexus could help in the selection of tagSNPs with potential functional effect, which could be used to complement the selection based on the haplotype information and therefore increase the sensitivity and efficiency of large-scale genotyping projects (Jorgenson and Witte, 2006; Wiltshire *et al*., 2006).

SNPnexus is the most comprehensive resource currently available for assessing the functional consequences of novel and public domain SNPs on the major transcriptome, proteome, regulatory and structural variation models. We provide examples on how it allows researchers, through a user-friendly web interface, to work on novel and publicly available SNPs to extract important annotation and genotyping information. SNPnexus data should be considered in the design of experimental functional assays so as to prioritize the list of SNPs or CNVs to be investigated in further association studies.

To the best of our knowledge, there are no tools or databases that provide the same information as SNPnexus. Therefore, we can only review other tools to provide an evaluation of our system. SNPnexus has a gene/protein annotation category allowing the users to view and navigate through large sets of high-dimensional data from five major annotation systems (Ensembl, RefSeq, VEGA, AceView, UCSC). Out of these, only Ensembl is mainly used by other tools, such as PupaSNP (Conde *et al*., 2006) and SNAP (Li *et al*., 2007b). SNPnexus offers more query options than PolyPhen (Ramensky *et al*., 2002) and SNAP. While these tools in their current versions allow queries for one SNP at a time, SNPnexus allows single and batch queries. Most importantly, SNPnexus is the only tool providing annotations for novel and publicly known SNPs, while all other tools focus on already known variants from dbSNP. Thus, SNPnexus is complementary to existing tools, such as Polyphen, SNAP and PupaSNP by allowing researchers to connect novel annotated SNPs to these tools. One could get further functional annotation of non-synonymous SNPs using PolyPhen, SNPs that disrupt exon splicing enhancers using PupaSNPs or redesign PCR primers to re-sequence individual variants using SNAP.

The recent technological advances in next-generation sequencing enable comprehensive screening of genome variation. One of the main promises and challenges lies with the analysis of these variations and the selection of those who contribute to the phenotype. Although designed initially to annotate novel SNPs from candidate gene sequencing studies, SNPnexus query option for novel SNPs could also be used to assess the potential significance of candidate variants detected by next-generation sequencing, point to the gene/protein isoform that might be phenotypically important and guide future experimentation. SNPnexus is freely available and will be updated regularly. We welcome feedback from the user community on new annotation systems that could be added to SNPnexus.

Funding: Cancer Research UK programme (C355/A6253).

Conflict of Interest: none declared.

REFERENCES

Ashurst,J.L. *et al*. (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.*, **33**, D459–D465.

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Becker,K.G. *et al*. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.

Benson,D.A. *et al*. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.

Chelala,C. *et al*. (2007) PTPN22 R620W functional variant in type 1 diabetes and autoimmunity related traits. *Diabetes*, **56**, 522–526.

Cohen,J.C. *et al*. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872.

Conde,L. *et al*. (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res.*, **34**, W621–W625.

Conrad,D.F. *et al*. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.

Davuluri,R.V. *et al*. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.

Feuk,L. *et al*. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Frazer,K.A. *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.

Grimson,A. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Hinds,D.A. *et al*. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.*, **38**, 82–85.

Hsu,F. *et al*. (2006) The UCSC Known Genes. *Bioinformatics*, **22**, 1036–1046.

Huang,M. *et al*. (2007) High-order interactions among genetic variants in DNA base excision repair pathway genes and smoking in bladder cancer susceptibility. *Cancer Epidemiol. Biomarkers Prev.*, **16**, 84–91.

Hubbard,T.J. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Iafrate,A.J. *et al*. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Iorio,M.V. *et al*. (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.*, **65**, 7065–7070.

Jorgenson,E. and Witte,J.S. (2006) A gene-centric approach to genome-wide association studies. *Nat. Rev. Genet.*, **7**, 885–891.

Kawakubo,H. *et al*. (2004) Expression of the NF-kappaB-responsive gene BTG2 is aberrantly regulated in breast cancer. *Oncogene*, **23**, 8310–8319.

Kawasaki,E. *et al*. (2006) Systematic search for single nucleotide polymorphisms in a lymphoid tyrosine phosphatase gene (PTPN22): association between a promoter polymorphism and type 1 diabetes in Asian populations. *Am. J. Med. Genet. A*, **140**, 586–593.

Lewis,B.P. *et al*. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Li,D. *et al*. (2007a) Effects of base excision repair gene polymorphisms on pancreatic cancer survival. *Int. J. Cancer*, **120**, 1748–1754.

Li,S. *et al*. (2007b) Snap: an integrated SNP annotation platform. *Nucleic Acids Res.*, **35**, D707–D710.

Locke,D.P. *et al*. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, **79**, 275–290.

McCarroll,S.A. *et al*. (2006) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Ramensky,V. *et al*. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Romeo,S. *et al*. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513–516.

Sangrajrang,S. *et al*. (2008) Polymorphisms in three base excision repair genes and breast cancer risk in Thai women. *Breast Cancer Res. Treat.*, **111**, 279–288.

Sebat,J. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.

Sharp,A.J. *et al*. (2005) Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.*, **77**, 78–88.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Siddiq,A. *et al*. (2005) A synonymous coding polymorphism in the alpha2-Heremans-schmid glycoprotein gene is associated with type 2 diabetes in French Caucasians. *Diabetes*, **54**, 2477–2481.

Smyth,D. *et al*. (2004) Replication of an association between the lymphoid tyrosine phosphatase locus (LYP/PTPN22) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes*, **53**, 3020–3023.

Stranger,B.E. *et al*. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7** (Suppl. 1), S12 1–14.

Tuzun,E. *et al*. (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.

Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.

Wiltshire,S. *et al*. (2006) The value of gene-based selection of tag SNPs in genome-wide association studies. *Eur. J. Hum. Genet.*, **14**, 1209–1214.

Wingender,E. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.

Yoon,S. and De Micheli,G. (2005) Prediction of regulatory modules comprising microRNAs and target genes. *Bioinformatics*, **21** (Suppl. 2), ii93–ii100.

Zhang,W. *et al*. (2008) The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform. Biol. Insights*, **2**, 15–23.