Published online 29 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D707–D710***
doi:10.1093/nar/gkl969

Snap: an integrated SNP annotation platform

Shengting Li[1,2], Lijia Ma[2,3], Heng Li[1,2], Søren Vang[4], Yafeng Hu[2], Lars Bolund[1,2] and Jun Wang[1,2,5,*]

[1]The Institute of Human Genetics, University of Aarhus, DK-8000 Aarhus C. Denmark, [2]Beijing Genomics Institute (BGI), Chinese Academy of Sciences (CAS), Beijing Airport Industrial Zone B-6, Beijing 101300, China, [3]Graduate University of the Chinese Academy of Sciences, Yuquan Road 19A, Beijing 100049, China, [4]Research Unit for Molecular Medicine, Aarhus University Hospital and Faculty of Health Sciences, DK-8200 Aarhus N, Denmark and [5]College of Life Sciences, Peking University, Beijing 100871, China

Received August 14, 2006; Revised October 11, 2006; Accepted October 13, 2006

ABSTRACT

Snap (Single Nucleotide Polymorphism Annotation Platform) is a server designed to comprehensively analyze single genes and relationships between genes basing on SNPs in the human genome. The aim of the platform is to facilitate the study of SNP finding and analysis within the framework of medical research. Using a user-friendly web interface, genes can be searched by name, description, position, SNP ID or clone name. Several public databases are integrated, including gene information from Ensembl, protein features from Uniprot/SWISS-PROT, Pfam and DAS-CBS. Gene relationships are fetched from BIND, MINT, KEGG and are integrated with ortholog data from TreeFam to extend the current interaction networks. Integrated tools for primer-design and mis-splicing analysis have been developed to facilitate experimental analysis of individual genes with focus on their variation. Snap is available at http://snap.humgen.au.dk/ and at http://snap.genomics.org.cn/.

INTRODUCTION

The large amount of 'omics' data coming from the complete map of the human genome and downstream work such as transcriptomics, proteomics and variation analyses opens new avenues for decoding sequence data. A long-term strategy of our data management system is to integrate large scale 'omics' data with bio-medical focus into a practical setting that supports genetic research in complex human disease. The SNP Annotation Platform (Snap) server is produced to this end and establishes the foundation of an analytic system for single genes and relationships between genes with focus on effects produced by SNPs. Two individuals are 99.9% identical at the DNA level; however, the remaining 0.1% has high medical importance. They define the traits that make us unique and underlie our susceptibility to disease and changes in drug response.

Information from the public domains [Ensembl v38 (1), Uniprot 8.0 (2), Pfam (3), CBS-DAS (4), MINT (5), BIND (6), KEGG 0.6.1 (7)] has been combined with our database with ongoing work to keep the content current and relevant. Moreover, we have integrated our animal model platforms CVDB (8) and PigGIS (9), own comparative genomics platform TreeFam (10) and a protein interaction analysis system currently under construction.

For each gene in Snap, a SeqView entry describes basic genome information and SNPs information. Mapping of protein features to the DNA level, primer design for resequencing and RT–PCR and comparative mis-splicing analysis of both known and user-requested SNPs are available in the SeqView. In addition, a RelationView can be selected for a visual organization of gene networks centered on the selected gene. By integrating evolutionary connections from TreeFam, current interaction networks can be dramatically extended.

The purpose of Snap is to organize and integrate data of medical importance in a user-friendly manner and add a number of convenient tools to aid further analysis of genes and variations within them.

DATA SOURCE AND METHODS

Mapping protein features

The complete human gene and SNP sets from Ensembl (v38), annotated protein features from Swiss-Prot (r132) (11) and predicted features through the CBS-DAS protein annotation viewer were downloaded to a local server. Protein features were mapped to the human genome and SNPs were added by aligning Ensembl proteins and UniProt proteins using the BLAST program (12) with the assistance from the cross-references provided by Ensembl.

*To whom correspondence should be addressed. Tel: +86 10 80481552; Fax: +86 10 80498676; Email: wangj@genomics.org.cn
*Correspondence may also be addressed to Lars Bolund. Tel: +45 89421675; Fax: +45 86123173; Email: bolund@humgen.au.dk

The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors

Predicting protein interactions

We have featured gene–gene relationships using protein–protein interactions. Information from the experimentally verified database MINT and the computational assistance-based database BIND are combined and data from KEGG are integrated to rank relationships between genes. Furthermore, orthologs defined by TreeFam are employed to transfer protein relationships of other species to the human system. An interaction between two human genes is established when one of the following conditions can be satisfied: (1) Gene A interacts with gene B in MINT or BIND; (2) Gene A and B are both involved in the same metabolic pathway provided by KEGG; (3) one ortholog of gene A interacts with one ortholog of gene B in other species.

DATABASE CONTENT AND ORGANIZATION

SNP index

The current version (v3) of Snap contains 23 710 human genes representing 48 218 individual transcripts and 3 480 292 SNPs from Ensembl v38, 123 884 of which lie in coding regions and 68 072 of these are nonsynonymous.

SNP annotation at transcripts level—prediction of protein features

The protein features from Swiss-Prot and Pfam numbering 2 115 643 are mapped to the DNA sequence to assign further biological meaning to each gene. We classify protein features referring to the category index of Swiss-Prot, and make use of Pfam and DAS-CBS data for complementarity. Five protein feature types with 39 sub-types from Swiss-Prot and two types of features with 11 sub-types from DAS-CBS are imported into Snap, covering 'protein sorting', 'post-translational modifications' and 'protein structure and function'. 'Amino-acid modifications', 'change indicators', 'regions', 'secondary structures' and 'others'—including several subunits separately—are features organized and provided in SeqView. 25 562 nonsynonymous SNPs (nsSNPs) are positionally co-localized in the sequence with protein features, comprising 37% of the nsSNPs. See supplementary data Table 1 for a list of all features.

Gene network—gene–gene relationship

Figuratively speaking, genes are spots and relations between them are roads that connect them. Snap presents 197 467 predicted interactions between human genes, of which 67 270 are contained in BIND, 47 826 in MINT, 2120 in KEGG and 80 251 are transferred from orthologous relationships in other species. To generate and show connections between genes, we have produced RelationView to describe networks centering on selected genes. Three formats are provided in Snap to show gene connections: RelationMap, RelationTree and RelationList.

RelationMap is generated by GraphViz (13) (Graph Visualization Software) and the different genes and connections are graphically distinctive by different types of borders and lines. Four levels of genes are shown; the root-gene is level-zero, this level connects with level-one genes, level-one genes connect with level-two genes, and so forth. All extensions for level-zero are always shown. To simplify the picture, extensions from each level-one and level-two gene are shown only if three or less exist. Additionally, the interaction network can be re-centered around any gene by clicking it. The RelationTree supports the RelationMap and presents all relations in the graphical tree detailing their data sources and levels of relationships. Both formats give hierarchical descriptions of gene connections. We also adopt a simple heuristic method, which accounts the number and quality of the supporting sources. All relations were rated and assigned a score reflecting their reliability. In the RelationList, the score was calculated for every two genes in a given map using an internal method based on their source quality. A higher 'score' reflects a shorter 'distance' between the genes. See supplementary data Table 2 for detailed algorithms.

Web service

(1) Primer design

An online service for PCR primer-design is provided to design primers for any region in the SeqView. Primers can be designed for resequencing individual SNPs, individual introns with flanking regions, individual exons with flanking regions, specific regions of interest or the entire sequence with or without introns. The service runs Primer3 (14) and individual primer pairs are checked for uniqueness using the UCSC *In Silico* PCR tool (15).

(2) Mis-splicing prediction

In recent years it has become clear that pre-translational regulation is complex and has shown vulnerable to sequence variation not only within the splice site consensus regions but also in a number of intronic and exonic *cis*-elements important for correct splice-site identification (16). The service 'Mis-splicing' estimates the degree of splicing defects resulting from a given nucleotide variation; either a SNP from the database or any base selected by the user. Six different tools are integrated in Snap [NNSplice (17), SpliceView (18), NetGene2 (19), ESEfinder (20), Rescue-ESE (21) and FAS-ESS (22)] to calculate various splicing parameters, and the results from both reference and SNP containing sequences are listed (see supplementary data Table 3 for details).

INTERFACE AND ACCESS

Snap is developed and maintained in a non-profit academic setting and can be directly accessed publicly (see Figure 1).

A search is simply done by inserting a gene ID, synonym name, SNP identifier or accession code from Ensembl Uniprot/SWISS-PROT. In SeqView, the primary result window, basic information about the gene is shown including description, position, transcripts, related diseases and polymorphism statistics with direct links to the information sources. The main part of SeqView is a gene map highlighting SNPs in combination with protein features from SWISS-PROT, Pfam and DAS-CBS selected by the user. The list of protein features can be reorganized by 'use', 'class' or 'source'. If a protein feature name is marked in red (e.g. Pfam), clicking the feature bar will highlight SNPs contained in regions positive for the feature. By clicking the 'redraw' button, the feature is shaded in yellow in the sequence map. The overall interface focuses at annotating features from

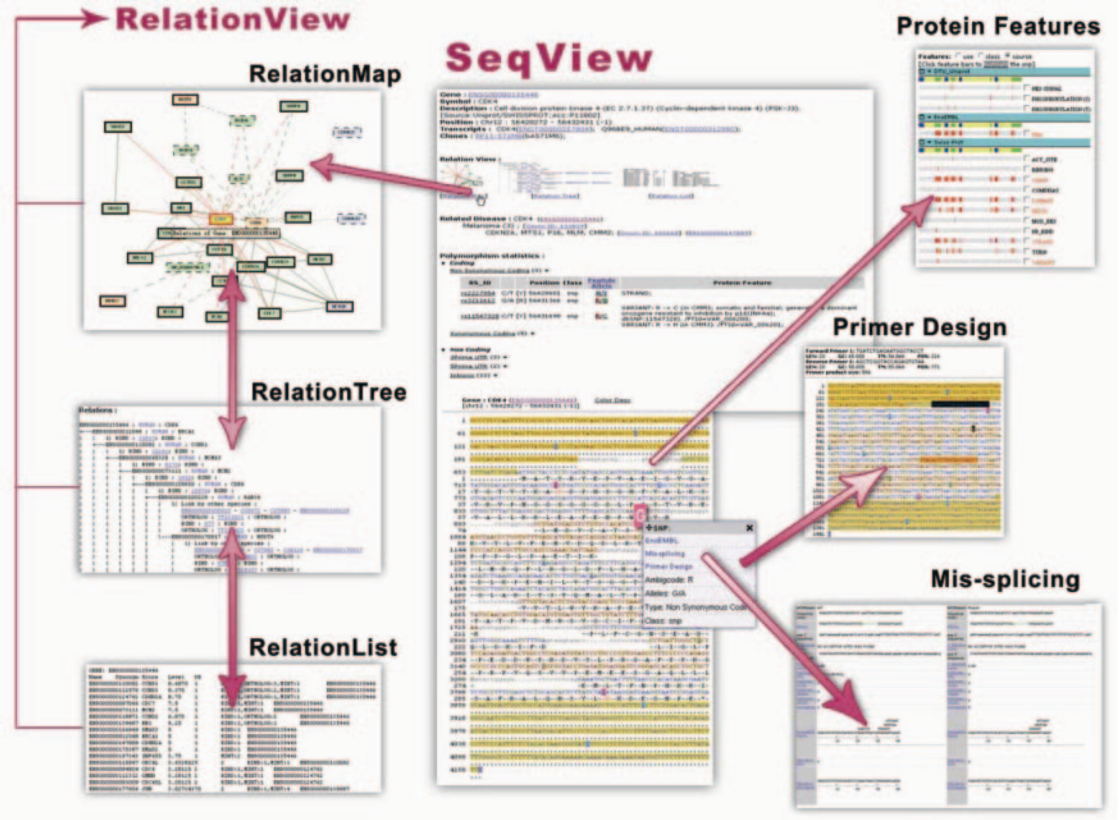

Figure 1. Main contents of Snap. From one central gene, SeqView and RelationView cover disease relevant aspects within the gene and within the gene's interaction network.

the transcript level to the DNA level and, most importantly, to the SNP level.

There are three links in the interface of SeqView that can lead to the RelationView, which is another part of Snap created from the perspective of gene–gene interactions. In the RelationMap, a web interface developed to visualize the network in a user-friendly way, three border patterns represent the number of outside links from the gene: (1) no outside links, (2) appropriate number of outside links and (3) too many outside links (threshold: 250). Relations between human genes are joined by bold lines, whereas those based on ortholog data are joined by dashed lines. Genes can be graphically grouped according to their ortholog data to clarify dependence on information from other species or all ortholog data can be disabled and only human gene relations are shown. Five types of layouts, three formats of figures and three levels of relationships are available to honor users' requirements. Gene relations from BIND and orthologs are displayed by default. In the RelationTree, we employed a plus sign '+' to stand for the existing of more relationships, while a backslash '\' to represent no more extended connections. The detailed underlying data can be seen in the RelationList views also.

The primer-design and mis-splicing services are accessible by clicking on the individual SNPs in 'Polymorphisms Statistics' or in the gene map in the SeqView. Furthermore, any base can be selected from the sequence and submitted from the panel to the right. For each design, two pairs of primers are given satisfying the parameters of primer size, TM value and product size; all parameters can be changed freely. A list of primer pairs for resequencing can be calculated for an entire mRNA or genomic sequence overlapping one by one, and covering the whole region of interest. The overlap is 100 bp by default. For RT–PCR use, primers are designed on mRNA and span introns specified by the user. Each primer pair must be at least 700 bp apart on the mRNA level and 1500 bp on the genome level. In addition, primers surrounding specified introns including flanking regions can be requested also.

CONCLUDING REMARKS

Complex diseases involve complex interactions of many aspects such as genetically coded variations, epigenetic modification and environmental influences. We need to define heterogeneous patterns of gene variations and their genetic modifiers to fully describe the genetic background of a common disease. The biggest challenge in achieving this aim is to organize the individual genes into one immense network. The three types of RelationView presented in Snap is a, still ongoing, attempt in this field.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Pia P. Madsen and Brage S. Andresen for discussions on splicing. This project was supported by Chinese Academy of Sciences (KSCX2-YW-N-023; GJHZ0518), Ministry of Science and Technology under high-tech program 863, Ministry of Education (XXBKYHT2006001), National Natural Science Foundation of China (90608010; 90208019; 90403130; 30221004; 90612019; 30392130), and China National Grid. The work was further supported by the Danish Platform for Integrative Biology, the Ole Rømer grant from the Danish Natural Science Research Council and the Danish Medical Research Council. Funding to pay the Open Access publication charges for this article was provided by the Danish National Research Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.
2. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.
3. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.
4. Olason,P.I. (2005) Integrating protein annotation resources through the Distributed Annotation System. *Nucleic Acids Res*., **33**, W468–W470.
5. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
6. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al*. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*., **33**, D418–D424.
7. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.
8. Wang,J., He,X., Ruan,J., Dai,M., Chen,J., Zhang,Y., Hu,Y., Ye,C., Li,S., Cong,L. *et al*. (2005) ChickVD: a sequence variation database for the chicken genome. *Nucleic Acids Res*., **33**, D438–D441.
9. Ruan,J., Guo,Y., Li,H., Hu,Y., Wang,J. and Bolund,L. PigGIS: Pig Genomic Informatics System. *Nucleic Acids Res*,. in press.
10. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Heriche,J.K., Osmotherly,L., Li,R., Liu,T., Zhang,Z., Bolund,L. *et al*. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*., **34**, D572–D580.
11. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
12. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
13. Gansner,E.R. and North,S.C. (1999) An open graph visualation system and its applications to software engineering. *Software Practice and Experience*, **30**, 1209–1233.
14. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.
15. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al*. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*., **34**, D590–D598.
16. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Rev. Genet*., **3**, 285–298.
17. Reese,M.G., Eeckman,F.H., Kulp,D. and Haussler,D. (1997) Improved splice site detection in Genie. *J. Comput. Biol*., **4**, 311–323.
18. Rogozin,I.B. and Milanesi,L. (1997) Analysis of donor splice sites in different eukaryotic organisms. *J. Mol. Evol*., **45**, 50–59.
19. Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouze,P. and Brunak,S. (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Res*., **24**, 3439–3452.
20. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*., **31**, 3568–3571.
21. Fairbrother,W.G., Yeh,R.F., Sharp,P.A. and Burge,C.B. (2002) Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**, 1007–1013.
22. Wang,Z., Rolish,M.E., Yeo,G., Tung,V., Mawson,M. and Burge,C.B. (2004) Systematic identification and analysis of exonic splicing silencers. *Cell*, **119**, 831–845.