

## DataBiNS-Viz: A Web-Based Tool for Visualization of Non-Synonymous SNP Data

Fong Chun Chan<sup>1</sup>, Edward A. Kawas<sup>1</sup>, Mark D. Wilkinson<sup>1,2</sup> and Scott J. Tebbutt<sup>1,3\*</sup>

<sup>1</sup>The James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research

<sup>2</sup>Department of Medical Genetics

<sup>3</sup>Department of Medicine, Division of Respiratory Medicine; University of British Columbia, Providence Heart + Lung Institute, St. Paul's Hospital, Vancouver, BC, V6Z 1Y6, Canada

\*Corresponding author: Dr. Scott J. Tebbutt, The James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research, Tel: 604-682-2344 ext. 63051;

Fax: 604-806-9274; E-mail: stebbutt@mrl.ubc.ca

Received June 26, 2008; Accepted July 16, 2008; Published July 17, 2008

**Citation:** Fong CC, Edward AK, Mark DW, Scott JT (2008) DataBiNS-Viz: A Web-Based Tool for Visualization of Non-Synonymous SNP Data. *J Proteomics Bioinform* 1: 233-236. doi:10.4172/jpb.1000029

**Copyright:** © 2008 Fong CC, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

### Abstract

Here we describe DataBiNS-Viz – a visualization and exploration environment for non-synonymous coding single nucleotide polymorphisms (nsSNPs) data gathered by the BioMoby-based DataBiNS workflow. DataBiNS-Viz enables execution of the DataBiNS workflow on proteins described by KEGG, PubMed, or OMIM identifiers, followed by manual exploration of the integrated structure/function and pathway data for those proteins, with a particular focus on nsSNP data in-context. The tool can be freely accessed at <http://bioinfo.icapture.ubc.ca:8090/DataBiNS> (please use the Firefox or Safari web browsers). Examples of the retrieved data are given under the “Help on inputs” option. Detailed documentation can be accessed at <http://bioinfo.icapture.ubc.ca/mywiki/DataBiNS>.

**Keywords:** Bioinformatics; Web services; Data mining; Visualization; Genomics; Single nucleotide polymorphisms

### Introduction

Single nucleotide polymorphisms (SNPs) are single base mutations in a genomic sequence that occur at a frequency greater than 1% in a defined population. Codons are sets of three DNA bases in a gene sequence that code for a particular amino acid. Non-synonymous SNPs (nsSNPs) are SNPs that occur within codons and that change the encoded amino acid, sometimes ultimately affecting the protein that is constructed from the gene blueprint. nsSNPs are of great interest to researchers as they may be key to identifying and understanding various human disease susceptibilities, as well as disease and non-disease phenotypes in many other species.

*In silico* analysis of the potential biological impact of nsSNPs requires integration of data and knowledge from various Web-based resources, both databases and analytical tools. Manual retrieval and integration of this information is error-prone and tedious. This provided the motiva-

tion for the original DataBiNS - data-mining workflow (Song et al., 2007) for the BioMOBY (Wilkinson and Links, 2002) and Taverna (Oinn et al., 2004) environments which retrieved and integrated data relating to nsSNPs and the biological pathways affected by them. DataBiNS consumes Kyoto Encyclopedia of Genes and Genomes [KEGG] Pathway Identifiers (Kanehisa et al., 2006), and retrieves a list of publications, gene ontology annotations and nsSNP information for each gene involved in the pathway. Although the public DataBiNS workflow successfully retrieved and integrated these data, lack of a visualization tool for the output significantly limited its utility. We report here important extensions to the original DataBiNS workflow and environment, including retrieval of additional nsSNP data such as mapping of SNPs to their altered amino acids on a 3D protein structure, as well as easy to navigate web-based visualizations of the global DataBiNS output.

## Workflow Initialization

To facilitate interoperability between the various Web resources, the workflow extensions we report here continue to be provided through the BioMoby Web Services framework. Rather than being limited to a single KEGG identifier, the new services allow for different types of identifiers to be used to initialize the workflow, including:

1. KEGG gene (<http://www.genome.jp/kegg/>)
2. PubMed (<http://www.ncbi.nlm.nih.gov/pubmed/>)
3. OMIM - Online Mendelian Inheritance in Man (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim>)
4. UniGene (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene>)
5. UniProt (<http://www.pir.uniprot.org/>)
6. GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>)
7. NCBI-GI

To initiate searches on multiple KEGG genes simultaneously, a comma can be placed between the different identifiers (*e.g.*, hsa:7097, hsa:7098)

## Extensions to Retrieved Data

Once the workflow has been initialized, the workflow first visits KEGG, PDB (<http://www.rcsb.org/pdb/home/home.do>), SwissProt (<http://www.expasy.ch/sprot/>), and Entrez (<http://www.ncbi.nlm.nih.gov/Entrez/>), to find the corresponding gene id(s) corresponding to the input identifier. Once retrieved, the LS-SNP (<http://alto.compbio.ucsf.edu/LS-SNP/>) database (Karchin et al., 2005) is initialized with the corresponding SwissProt id to find all the nsSNPs for the gene. The PDB id is then used on the coliSNP (<http://yayoi.kansai.jaea.go.jp/colisnp/>) database (Kono et al., 2008) to retrieve the 3D structure of the protein (if available). The various SNPs associated with this gene are already mapped onto this protein structure (within coliSNP), providing an efficient technique to analyze the location of SNPs on the protein. Supplementing the SNP information are frequency pie-charts of each SNP id from the HapMap (<http://www.hapmap.org/>) database (Thorisson et al., 2005). Detailed annotations about the gene are retrieved from the Gene Ontology (<http://www.geneontology.org/>) website, and finally the most recently relevant publications to the gene are retrieved from PubMed.

## Web-Based Visualization

Rather than being limited to the default Taverna nested-

folder browsing, or export of the data from Taverna as an Excel spreadsheet, both of which are problematic for manual exploration of these complex data networks, we have created a task-specific Web-based visualization and exploration environment for DataBiNS. The application is built using the Java Platform, Enterprise (J2EE) and is accessed by end-users through an intuitive Web page. The user simply enters an identifier of interest (*i.e.*, KEGG PATHWAY, OMIM, etc.), and then presses the “Execute Workflow” button. In the backend, the Taverna workflow execution engine is triggered to execute the modified DataBiNS workflow. The results of the workflow are then cached on the server to allow rapid browsing of results, and are browsable via the Web interface (Figure 1). There is an option on the front page to re-execute the workflow, where the tool will ignore any saved results and retrieve new, possibly updated data.

In addition to displaying all the results using standard Web technologies, two navigation tools/methods have been added to the web-application to help with the study of the data. First, a “search publication abstract” option allows for users to quickly search the retrieved publications for keywords. If a retrieved publication has the keyword, the publication will be highlighted allowing the user to focus on that publication. The PubCloud (<http://bioinfo.iculture.ubc.ca:8090/PubCloud/>) application has also been integrated into the web-application. The user can select a group of the retrieved publications and quickly use the PubCloud keyword tag-cloud visualization system to find possible correlations between the publications.

In a significant advance over prior exploration/browsing environments, the Web interface intuitively associates multiple inputs with their respective outputs. Thus the Web-application displays all data about each gene in a discrete section of the browser window; on a given results page there can be several genes and each gene will have its associated information clearly and intuitively organized and displayed. This approach eliminates the user’s need to backtrack through the results to correlate inputs to outputs, as was required in earlier versions of DataBiNS, thus allowing them to quickly analyze and use the retrieved data.

## Future

Though the framework we have developed to display the data is specific for the DataBiNS workflow, it can be generalized to accept any Taverna-based workflow, displaying the results as a browsable Web page and facilitating exploration of results from Taverna-based workflows. The modularized nature of workflows allows one to develop new

Identifiers searched

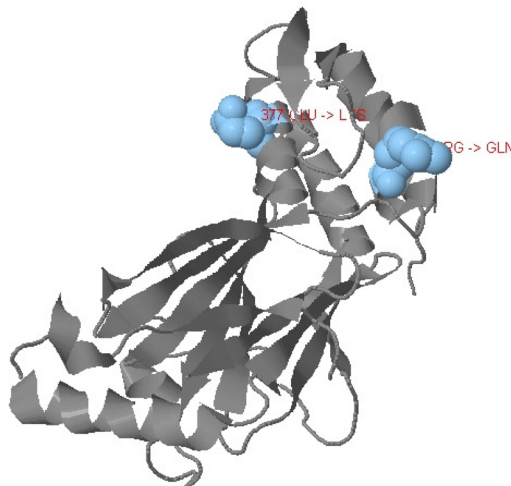
1. [hsa:3661](#)

Identifiers that have nsSNPs (No. nsSNPs)

1. [hsa:3661](#) (2)

Identifiers

Kegg ID	<a href="#">hsa:3661</a>
Swiss Prot ID	<a href="#">Q14653</a>
Entrez Gene ID	<a href="#">83596</a>
Protein Data Bank (PDB) ID	<a href="#">1J2F</a>
Protein Structure	<a href="#">View 3D Protein Structure with SNPs</a>




SNPs

nsSNP ID

<b>rs968457: 96 Arg -&gt; Gln</b>	
nsSNP Strand	-
nsSNP Amino Acid Position	96
nsSNP Amino Acid Variant	RQ
nsSNP Allele	A/G
nsSNP Codon Position	2
<b>rs7251: 427 Ser -&gt; Thr</b>	

SNP Frequencies

nsSNP ID

<b>rs968457</b>
Frequency - CEU: G/G: 1: 58; A/G: 0: 0; A/A: 0: 0; 58; G: 1: 116; A: 0: 0; 116 - CHB: G/G: 1: 45; A/G: 0: 0; A/A: 0: 0; 44; G: 1: 88; A: 0: 0; 88 - YRI: G/G: 0.831: 49; A/G: 0.169: 10; A/A: 0: 0; 59
Frequency Image

<b>rs 7251</b>

- Toll-like receptor 3-mediated activation of NF-kappaB and IRF3 diverges at Toll-IL-1 receptor domain-containing adapter inducing IFN-beta.
- Oxidized low density lipoprotein blocks lipopolysaccharide-induced interferon beta synthesis in human macrophages by interfering with IRF3 activation.
- The host response to West Nile Virus infection limits viral spread through the activation of the interferon regulatory factor 3 pathway.

Summarize using PubCloud

**PubCloud** Publication (PDF) (HTML) Legend Print Close

PubMed Query: , **15107417, 15220448**  
 Cloud Type: **abstract**  
 Cloud Option: **Most relevant 100 articles**  
 Date range from -- to --  
[Show/Hide all 2 PubMed ID\(s\)](#)

able absence accounts **activate** analyses antiviral asses  
beta biochemical birds blocked blood-derived c-Jun cbp cell-to-ce  
coactivator constrain cultured cycle delayed demonstrated density  
downstream drive effect efficiently embryo essential evades even  
expressed factors fail fibroblasts function genes ho

**Figure 1:** DataBiNS-Viz Outputs. Examples of retrieved data and visualizations from a DataBiNS-Viz workflow search of nsSNP information related to the KEGG gene hsa:3661 (interferon regulatory factor 3 from *H. sapiens*).

BioMoby services to add to DataBiNS in order to expand the information retrieved and displayed.

One lingering question is always the validity of the data being retrieved. The workflow is designed to retrieve information from a specified group of web resources. The validity of the information obtained from the web resources is not checked by the workflow and thus there is currently no way to verify the integrity of the information across the different resources, without a great deal of manual data inspection. Future developments may lead to automation of such processes with electronic flags highlighting inconsistencies in data between different web resources.

## Acknowledgements

This research was supported by the National Sanitarium Association (Canada), AllerGen NCE, and the Michael Smith Foundation for Health Research. EK is supported by an award to MDW from Genome Alberta, in part through Genome Canada.

*Conflict of Interest:* none declared.

## References

1. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34: D354-357. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
2. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, et al. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* 21: 2814-2820. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
3. Kono H, Yuasa T, Nishiue S, Yura K (2008) coliSNP database server mapping nsSNPs on protein structures *Nucleic Acids Res* 36: D409-413. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
4. Oinn T, Addis M, Ferris J, Marvin D, Senger M, et al. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20: 3045-3054. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
5. Song YC, Kawas E, Good BM, Wilkinson MD, Tebbutt, SJ (2007) DataBiNS: a BioMoby-based data-mining workflow for biological pathways and non-synonymous SNPs. *Bioinformatics* 23: 780-782. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
6. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The International HapMap Project Web site *Genome Res* 15: 1592-1593. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)
7. Wilkinson MD, Links M (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform* 3: 331-341. » [CrossRef](#) » [Pubmed](#) » [Google Scholar](#)