LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures

Michael Ryan\textsuperscript{1}, Mark Diekhans\textsuperscript{2}, Stephanie Lien\textsuperscript{3}, Yun Liu\textsuperscript{3} and Rachel Karchin\textsuperscript{3}\textsuperscript{*}

\textsuperscript{1}Department of Bioinformatics, George Mason University, Fairfax, Virginia USA
\textsuperscript{2}Department of Biomolecular Engineering, University of California, Santa Cruz, California USA
\textsuperscript{3}Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, 3400 N. Charles St., Baltimore, Maryland USA

\textbf{ABSTRACT}

Summary: LS-SNP/PDB is a new WWW resource for genome-wide annotation of human non-synonymous (amino-acid changing) SNPs. It serves high quality protein graphics rendered with UCSF Chimera molecular visualization software. The system is kept up-to-date by an automated, high-throughput build pipeline that systematically maps human nsSNPs onto Protein Data Bank structures and annotates several biologically relevant features.

Availability: LS-SNP/PDB is available at http://ls-snp.icm.jhu.edu/ls-snp-pdb and via links from the Protein Data Bank (PDB) Biology and Chemistry tabs, UCSC Genome Browser Gene Details and SNP Details pages and PharmGKB Gene Variants Downloads/Cross-References pages.

Contact: karchin@jhu.edu

Supplementary Information: Supplementary information is available at Bioinformatics online.

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common type of inter-individual human genetic variation (Goodstadt and Ponting, 2001), and the number of human SNPs catalogued in NCBI’s dbSNP database (Sherry et al., 2001) has grown exponentially over the past few years. Build 129 of dbSNP (10 June 2008) contains over 1.4 million uniquely mapped human nsSNPs, of which over 90,000 are non-synonymous. Because SNPs can impact an individual’s susceptibility to disease and sensitivity to drugs (Sunyaev, 2001; Evans, 1999), there is increasing interest in computational tools to identify SNPs that may affect molecular function. We have developed a resource to contribute to this effort, using genome-wide mapping of nsSNPs onto experimentally-determined protein structures.

LS-SNP/PDB annotates all human SNPs that produce an amino acid change in a protein structure in PDB (Deshpande, 2005), using features of their local structural environment, putative binding interactions and evolutionary conservation. For nsSNPs that are accessible to solvent, we provide two colorized views of the protein surface to represent: 1) evolutionary conservation among homologous proteins and 2) electrostatic potential. The presence of a nsSNP in a highly conserved surface patch or a charged surface patch suggests possible biological importance. These annotations allow users to quickly scan a large number of nsSNPs of interest and prioritize those with higher likelihood of impacting normal protein activities.

This work builds on an earlier version of LS-SNP (Karchin, 2005) that mapped human nsSNPs onto protein homology models. It features completely redesigned pipeline software, fully automatic builds, weekly incremental updates, new annotations, high-quality protein images for molecular visualization, and a new web interface.

2 SYSTEMS AND METHODS

The pipeline integrates data from a variety of sources and a new build is run for each new release of the UCSC Genome Browser (Karolchik, et al) SNP track as follows:

1. mapping of all human UniProtKB (Wu, 2006) protein sequences onto genomic DNA to identify the (unknown) codon that yielded each amino acid residue. Mapping is done by aligning the UniProtKB proteins to all human mRNAs in the Genome Browser and selecting the best match;
2. taking the genomic address of each SNP from dbSNP (currently from NCBI Build 36.1 / hg18) and finding SNPs that alter an amino acid residue;
3. using alignments of UniProtKB and PDB protein sequences to map each nsSNP onto a PDB structure;
4. annotating nsSNPs by structural location, secondary structure, solvent accessibility, evolutionary conservation, electrostatic potential, and proximity to small molecule ligands and domain interfaces.
5. Creating high-quality protein images for each nsSNP.

Users can access the database with a query form and select nsSNPs by dbSNP, HUGO gene, UniProtKB, KEGG Pathway ID, or PDB ID(s), genomic region or cytogenic band. Queries can be further restricted by: solvent accessibility (exposed,intermediate,buried); secondary structure (helix,strand,coil, proximity to ligands or domain interfaces; severity of amino acid...
several SNPs in intercellular adhesion molecule 1 to be significantly associated with lower soluble protein concentration in plasma (a vascular disease biomarker) (Pare, 2008). rs1799969 (glycine to arginine) is in an ICAM-1 domain that binds the integrin MAC-1 (Diamond, 1991). This domain has been shown to bind a charged surface on MAC-1 (Yang, 2007), and one study suggested that rs1799969 weakens this interaction (Ponthieux, 2003). LS-SNP/PDB electrostatic potential view of PDB ID 1lp3 shows that it is located on an acidic surface patch, thus partial neutralisation of the charged region by arginine could explain weakened binding. Alternatively, there may be a steric requirement for glycine at this position (a turn in the protein backbone). Relative importance of electrostatics and steric for MAC-1 binding could be experimentally tested by comparing stability and MAC-1 binding of wild-type ICAM-1 with R241 and E241 mutants.

5 CONCLUSION

Plans for LS-SNP include annotation of small deletions, insertions and inversions, integration of high-quality homology models, machine learning classification of variants, and handling of custom inputs, such as rare variants not in dbSNP. We will incorporate more detailed information about evolutionary conservation and electrostatic potential at each SNP site. In the near future, we will be applying the pipeline to somatic DNA sequence alterations found in the exons of tumor genomes.

REFERENCES


Fig. 1. Screenshot of rs11604921. a) Location in PDB 1ci4A. b) Protein surface rendering colored by conservation; c) colored by electrostatic potential.