ABSTRACT

Summary: We have developed a resource, MutDB (http://mutdb.org/), to aid in determining which single nucleotide polymorphisms (SNPs) are likely to alter the function of their associated protein product. MutDB contains protein structure annotations and comparative genomic annotations for 8000 disease-associated mutations and SNPs found in the UCSC Annotated Genome and the human RefSeq gene set. MutDB provides interactive mutation maps at the gene and protein levels, and allows for ranking of their predicted functional consequences based on conservation in multiple sequence alignments.

Availability: http://mutdb.org/

Contact: sdm@stanford.edu

Supplementary information: http://mutdb.org/about/about.html

TEXT

Single nucleotide polymorphisms (SNPs) are the most common form of human genomic variation. While many of the known SNPs are chiefly useful as markers, some of them affect the function of their associated protein product by altering transcription, translation, post-translational modification, or molecular function. We have developed a resource, MutDB (http://mutdb.org/), to aid in determining which SNPs are likely to alter the function of their associated protein product. MutDB currently contains protein structure annotations and comparative genomic annotations for 8000 disease-associated mutations and SNPs found in the UC Santa Cruz (UCSC) Annotated Genome and the human RefSeq gene set. We use multiple sequence alignments in combination with protein structure to highlight functional mutations and functionally important protein regions.

There are several ongoing public projects that store, characterize and present human genomic variation data (Collins et al., 1997). These data sets usually fall into two classes: (i) hand-curated, loci-specific databases that contain phenotypic annotations of rare variants and SNPs and (ii) large unannotated data sets associated with many genes. To understand the molecular basis of a phenotype, we require functional and molecular annotations of these data sets. Several research projects have developed methods for both predicting disease-associated mutations and estimating the number of disease-associated mutations. Early tools using phylogenetic and structural information have attempted to predict the functional consequences of mutations (Chasman and Adams, 2001). These methods predict that between 20 and 36% of non-synonymous SNPs alter the function of a gene’s protein product. In addition, evolutionary information was shown to be a useful component in determining whether a mutation is deleterious (Chasman and Adams, 2001). (Sunyaev et al., 2001). Ng and Henikoff have introduced SIFT a profile method for predicting functional SNPs from a database of unannotated polymorphisms (Ng and Henikoff, 2001, 2002) and Ramensky et al. have developed PolyPhen (Ramensky et al., 2002)[RBA1]. Simon et al. have described a phylogenetic method for inferring the functional regions of genes (Simon et al., 2002). Disease causing mutations are also likely to perturb structure at the protein level as well (Wang and Moult, 2001). Wacey et al. estimated the sequence and structural implications of disease-associated mutations in the p53 gene (Wacey et al., 1999); their results showed that substitution rates of disease-associated mutations correlate with changes in biophysical properties.

Our initial focus has been to annotate disease-associated mutations with protein structure and comparative genomics as demonstrated in Figure 1. Our next step is to broaden our annotation efforts to mapped common SNPs.

DATABASE DATA MODEL

Since functional mutations can occur anywhere in the genome, the data is typically indexed using its location within the human genomic sequence. Many functional mutations are likely to be near genes, so the data set is currently gene centric (i.e., all mutations are annotated and displayed with an associated gene). We use Refseq to extract the translated sequence, the cDNA/mRNA sequence, and journal references. The mapping between the human genome and Refseq is provided by UCSC Annotated Genome (http://genome.ucsc.edu), where Refseq is mapped to the human genome in the refGene.
MutDB: annotating human variation

Table. The annotated genome also provides many of the phenotypically unannotated polymorphisms stored in dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). Annotated mutations are provided by many sources, including loci specific databases, the HGMD (Krawczak and Cooper, 1997) and many coding mutations in SWISSPROT (Bairoch and Apweiler, 2000). In total, over 1,000,000 SNPs are mapped in the UCSC data and over 10,000 phenotypically annotated mutations are in SWISSPROT.

We first mapped mutations in SWISSPROT onto the UCSC annotated genome. This was performed by submitting the eight codons (24 bases) to the left and right of a mutation to BLAT (8) in a search of the human genome. Sequences that are unambiguously identified are considered mapped and the golden path position of the mutant codon, chromosome and strand are noted and stored in a MySQL (http://www.mysql.com) database.

STRUCTURAL ANNOTATIONS

The phenotypically annotated mutations that code for a protein are annotated with protein structure information. Initially, we matched the homologous structures in the PDB (Berman et al., 2000) with a query cDNA from Refseq, and then mapped the mutations falling within regions of structural homology (all alignments with a BLAST e-value of less than $10^{-9}$ are used). Each mutation can be visualized when mapped to a protein structure using the web browser plug-in, CHIME (http://www.mdlchime.com/chime/). Users can also view all the known mutations mapped to a single protein structure with only the most conserved mutations highlighted. This allows for rapid identification of functional regions and the distribution of functional mutations.

COMPARATIVE GENOMICS ANNOTATIONS

The phenotypically annotated mutations are also annotated with the results of a comparative analysis, aimed at identifying likely functional mutations. The method for analyzing mutations has been applied to androgen receptor (AR) (Mooney et al., 2003) gene mutations and broadly to many loci (Mooney and Klein, 2002). This automatic method relies on searching SWISSPROT using the BLAST algorithm and aligning the significant results with ClustalW (Thompson et al., 1994). Conservation is then quantified using a modified version of the Shannon entropy. Users can perform the analysis with a protein sequence compared against SWISSPROT, or with a nucleotide sequence compared against Genbank (http://www.ncbi.nlm.nih.gov/Genbank/).

Each mutation is then ranked by conservation, where mutations affecting more conserved locations are ranked more highly. Users can easily download the BLAST output file, ClustalW alignments and annotated mutation lists if further analysis is needed.

CONCLUSIONS AND FUTURE EFFORTS

Knowledge of the underlying function and distribution of disease-associated mutations is important for understanding of the molecular basis of disease. To improve identification of functionally important regions of genes, we have combined information about evolutionary conservation as well as three-dimensional structure information. These complementary annotations allow users to quickly identify likely functional mutations and the associated regions within their protein structures. Our resource, MutDB (http://mutdb.org/), therefore assists researchers in assigning a functional importance to mutations of interest.

ACKNOWLEDGEMENTS

S.M. is funded by an American Cancer Society Fellowship. We would like to gratefully acknowledge financial support from grants NIH LM06244 and NIH AR47720-01 (T.E. Klein, PI).
REFERENCES


