Informatics Approaches to Linking Mutations to Biological Pathways, Networks and Clinical Data

Arti Singh

Submitted to the faculty of the Bioinformatics Graduate Program
in partial fulfillment of the requirements
for the degree
Master of Science
in the School of Informatics,
Indiana University

August 2007

Accepted by the Faculty of Indiana University, in partial
fulfillment of the requirements for the degree of Master of Science

Sean Mooney, Ph.D. Chair

Jeesun Jung , Ph.D.

Pedro Romero , Ph.D.

For my mother Smt. Shobha Singh and my father Prof. Amerika Singh

whose love, dedication and strength taught me important lessons in life

You are always with me. I will always love and miss you.

ACKNOWLEDGEMENTS

I extend my gratitude and appreciation to the people who made this master's thesis possible. My foremost thanks go to my research advisor, Prof. Sean Mooney. I want to thank him for his generosity, insights and suggestions that helped me have a deeper understanding of the subject matter.

I am grateful to Prof. Jeesun Jung and Prof. Pedro Romero for the support, encouragement and invaluable guidance that I received from them. Special thanks are due to all my friends and colleagues at the Mooney lab and at the Center for Computational Biology and Bioinformatics, Indiana School of Medicine. Last but not the least I am greatly indebted to my family for always supporting me.

ABSTRACT

The information gained from sequencing of the human genome has begun to transform human biology and genetic medicine. The discovery of functionally important genetic variation lies at the heart of these endeavors, and there has been substantial progress in understanding the common patterns of single-nucleotide polymorphism (SNP) in humans- the most frequent type of variation in humans. Although more than 99% of human DNA sequences are the same across the population, variations in DNA sequence have a major impact on how we humans respond to disease; to environmental entities such as bacteria, viruses, toxins, and chemicals; and drugs and other therapies and thus studying differences between our genomes is vital. This makes SNPs as well other genetic variation data of great value for biomedical research and for developing pharmaceutical products or medical diagnostics.

The goal of the project is to link genetic variation data to biological pathways and networks data, and also to clinical data for creating a framework for translational and systems biology studies. The study of the interactions between the components of biological systems and biological pathways has become increasingly important. It is known and accepted by scientists that it as important to study different biological entities as interacting systems, as in isolation. This project has ideas rooted in this thinking aiming at the integration of a genetic variation dataset with biological pathways dataset. Annotating genetic variation data with standardized disease notation is a very difficult yet important endeavor. One of the goals of this research is to identify whether informatics approaches can be applied to automatically annotate genetic variation data with a classification of diseases.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The information gained from sequencing of the human genome has begun to transform human biology and genetic medicine. The discovery of functionally important genetic variation lies at the center of these endeavors, and there has been considerable progress in understanding the common patterns of single-nucleotide polymorphism (SNP) in humans- the most common type of variation in human (Stoneking, 2001). Knowledge about the functional effects of these DNA variations among the humans should lead to revolutionary new ways to diagnose, treat, and someday prevent the thousands of disorders that affect us in addition to providing clues to understanding human biology (Chakravarti, 2001; Taylor, Choi, Foster, & Chanock, 2001).

A SNP is stable substitution of a single base (A,T,C,or G) in DNA sequence with a frequency of more than 1% in at least one population. SNPs make up about 90% of all human genetic variation and thus in this text, the terms SNPs and genetic variation data are used interchangeably. SNPs occur every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement of cytosine (C) with thymine (T). SNPs can occur in both coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function, but others could predispose people to disease or influence their response to a drug (Taylor et al., 2001). A vast majority of the SNPs that do affect function have yet to be annotated (S. Mooney, 2005).

Although more than 99% of human DNA sequences are the same across the population, variations in DNA sequence can have a major impact on how humans respond to disease; environmental entities such as bacteria, viruses, toxins, and chemicals; and drugs and other therapies (Chakravarti, 2001) and thus studying differences between our genomes makes more sense. This makes SNPs of great value for biomedical research and for developing pharmaceutical products or medical diagnostics. SNPs are also evolutionarily stable --not changing much from generation to generation --making them easier to follow in population studies. Scientists believe SNP maps will help them identify the multiple genes associated with such complex diseases as cancer, diabetes, vascular disease, and some forms of mental illness. These associations are difficult to establish with conventional gene-hunting methods because a single altered gene may make only a small contribution to the disease (Risch, 2000). SNP maps are helping to identify thousands of additional markers along the genome, thus simplifying navigation of the much larger genome map generated by researchers in the Human Genome Project (HGP)(Venter et al., 2001).

SNPs can cause disease and can help determine the likelihood that someone will develop a particular disease. One of the genes associated with Alzheimer's(Wardell, Suckling, & Janus, 1982), apolipoprotein E or *ApoE*, is a good example of how SNPs affect disease development. This gene contains two SNPs that result in three possible alleles for this gene: E2, E3, and E4. Each allele differs by one DNA base, and the protein product of each gene differs by one amino acid.

Each individual inherits one maternal copy of *ApoE* and one paternal copy of *ApoE*. Research has shown that an individual who inherits at least one E4 allele will have a greater chance of getting Alzheimer's. Apparently, the change of one amino acid in the E4 protein alters its structure and function enough to make disease development more likely. Inheriting the E2 allele, on the other hand, seems to indicate that an individual is less likely to develop Alzheimer's.

Of course, SNPs are not absolute indicators of disease development. Someone who has inherited two E4 alleles may never develop Alzheimer's, while another who has inherited two E2 alleles may. *ApoE* is just one gene that has been linked to Alzheimer's. Another common example of such a gene would be BRCA1(Douglas et al., 2007).BRCA1 (breast cancer 1, early onset) is a human gene that belongs to a class of genes known as tumor suppressors, which regulate the cell cycle and prevent uncontrolled proliferation. The BRCA1 protein product of the gene is part of the DNA damage detection and repair system. Variation in the gene has been implicated in some cancers.

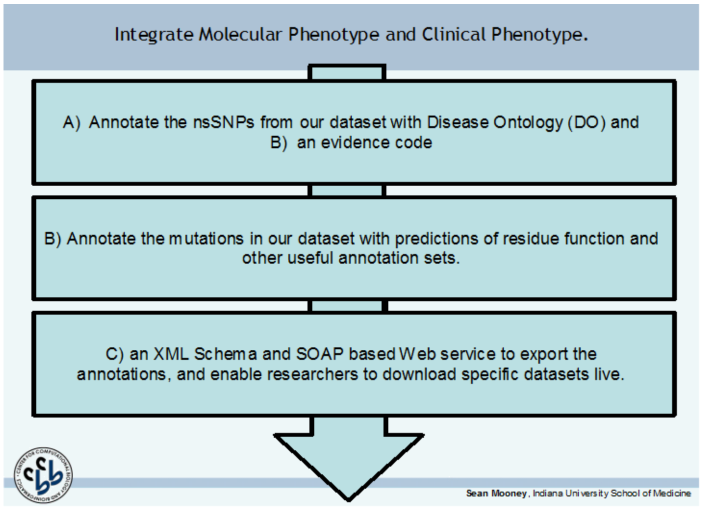
Fact Sheet for SNPs
• DNA sequence variations that occur when a single nucleotide in genome sequence is altered.
• For a variation to be considered a SNP, it must occur in at least 1% of the population.
• SNPs make up about 90% of all human genetic variation
• Occur every 100 to 300 bases along the 3-billion-base human genome.
• Can predispose people to disease or influence their response to a drug.
• Two of every three SNPs involve the replacement of cytosine (C) with thymine (T).
• SNPs can occur in both coding (gene) and noncoding regions of the genome.

Table 1: Fact Sheet for SNPs

The first part of this text describes the initial project to integrate biological pathways data with genetic variation data in an attempt to create a framework for systems biology studies. Systems biology is an emergent field that aims at system-level understanding of biological systems(Silver & Way, 2007). Unlike molecular biology which focus on molecules, such as sequence of nucleotide acids and proteins, systems biology focus on systems that are composed of molecular components. Although systems are composed of matter, the essence of system lies in dynamics and it cannot be described merely by enumerating components of the system.

The idea behind the concept of translational sciences is that scientific discoveries must be translated into practical applications primarily to benefit human health (Zerhouni, 2005). The scientific community is now realizing the gap between discoveries made in lab and benefits reaching patients. This is the goal of the second part of the project that tries to explore this new aspect of science by trying to relate genetic variation data with informatics methods for annotating disease data as well- in an attempt to create a framework for translational studies.

Thus current document and the project that this document explains essentially have these two logically related yet physically independent parts to it. On one hand, genetic variation dataset has been integrated with biological pathways dataset, on the other, possibilities to integrate genetic variation dataset with clinical data have been explored. Clinical dataset here means a standard notation or classification of clinical terminology, thus attempting to create a platform for studies in translational sciences.



Integrate Molecular Phenotype and Clinical Phenotype.

A) Annotate the nsSNPs from our dataset with Disease Ontology (DO) and
B) an evidence code

B) Annotate the mutations in our dataset with predictions of residue function and other useful annotation sets.

C) an XML Schema and SOAP based Web service to export the annotations, and enable researchers to download specific datasets live.

Sean Mooney, Indiana University School of Medicine

Figure 1: Visualizing ideas about how a framework can be build to annotate genetic variation data with disease ontology. (Visualized and created by Dr Sean Mooney).

BACKGROUND

SOA and web services

Service Oriented Architecture or SOA is an informatics approach for the development of loosely coupled distributed software applications. Service-oriented architecture is collection of many services in the network. These services communicate with each other and the communication involves data exchange and service coordination. Earlier SOA was based on the Distributed Component Object Model (DCOM) or Object Request Brokers (ORBs) (Chang, 2000). Currently SOA is based on the web services(Alexander Davis, 2002).

A service is a function or some processing logic that is well-defined, self-contained, and does not depend on the context or state of other services. A web service is defined as a piece of logic written in any computer programming language that can be used by other applications over the internet. If we go by component based software models then web service could be defined as a component over the internet. Web services are software components that interact with one another dynamically via standard Internet technologies, making it possible to build bridges between IT systems that otherwise would require extensive development efforts or is nearly impossible in some cases. There are various examples of web services. Google, a popular Internet search engine, provides the web service called the Google web API. The service enables users to develop software that accesses and manipulates a massive amount of web documents that are constantly refreshed.

Three major methods of producing and consuming web services are XML-RPC, REST and SOAP.

XML-RPC is a set of implementations that allow software running on disparate operating systems, running in different environments to make procedure calls over the Internet.

It's remote procedure calling using HTTP as the transport layer and XML as the encoding.

REST is an acronym standing for Representational State Transfer. This is the internet's underlying architectural style. One major difference between the REST- based and that of the SOAP-based web services architecture is that REST advocates view the internet as an information system in its own right and SOAP is part of ideology that advocates that other systems should be integrated into the information system through gateways. Further discussion about REST is beyond the scope of this text.

In SOAP based web services, SOAP over HTTP is used to communicate the between services. The basic web services platform elements are:

1. SOAP (Simple Object Access Protocol)

   The basic web services platform is XML plus HTTP. SOAP or Simple Object Access Protocol is a communication protocol. It is a format for sending messages. SOAP is designed to communicate via the internet. SOAP is platform and language independent. SOAP is based on XML and therefore it is simple and extensible. SOAP allows applications to get around firewalls. Finally, SOAP has been developed as a W3C standard

2. WSDL (web Services Description Language)

   WSDL or web Services Description Language is an XML-based language for describing web services and how to access them. WSDL is also used to locate web services. WSDL is not yet a W3C standard. It can be described as an XML format for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information. The operations and messages are described abstractly, and then bound to a concrete network protocol and message format to define an endpoint. Related concrete endpoints are combined into abstract endpoints (services).

3. UDDI (Universal Description, Discovery and Integration)

UDDI or Universal Description, Discovery and Integration is a directory service where businesses/organizations can register and search for web services. UDDI is a directory for storing information about web services. UDDI is a directory of web service interfaces described by WSDL. It communicates via SOAP.

Essential Features of SOAP-based web services
• Web services are application components
• Web services communicate using open protocols
• Web services are self-contained and self-describing
• Web services can be discovered using UDDI
• Web services can be used by other applications
• XML is the basis for web services

Table 2: Essential Features of SOAP-web services

Using web services, an application can publish its function or message to the rest of the world. The basic web services platform is XML and HTTP - one of the most used internet protocol. XML provides a language which can be used between different platforms and programming languages and still express complex messages and functions. A web service is a software application that can be accessed remotely using different XML-based languages. Normally, a web service is identified by a URL, just like any other web site. What makes web services different from ordinary web sites is the type of interaction that they can provide.

Most web sites are designed to provide a response to a request from a person. The person either types in the URL of the site or clicks on a hyperlink to create the request. This request takes the form of a text document that contains some fairly simple instructions for the server. These instructions are limited to the name of a document to be returned or a call to a server-side program, along with a few parameters. Figure2 shows this process graphically. A web service is similar in that it is accessed via a URL. The difference lies in the content of what is sent in the request from the client to the service. Web service clients send an XML document formatted in a special way in accordance with the rules of the SOAP specification. A SOAP message can contain a call to a method along with any parameters that might be needed. In addition, the message can contain a number of header items that further specify the intent of the client. These header items might designate what web services will get this method call after the current service finishes its work, or they might contain security information.

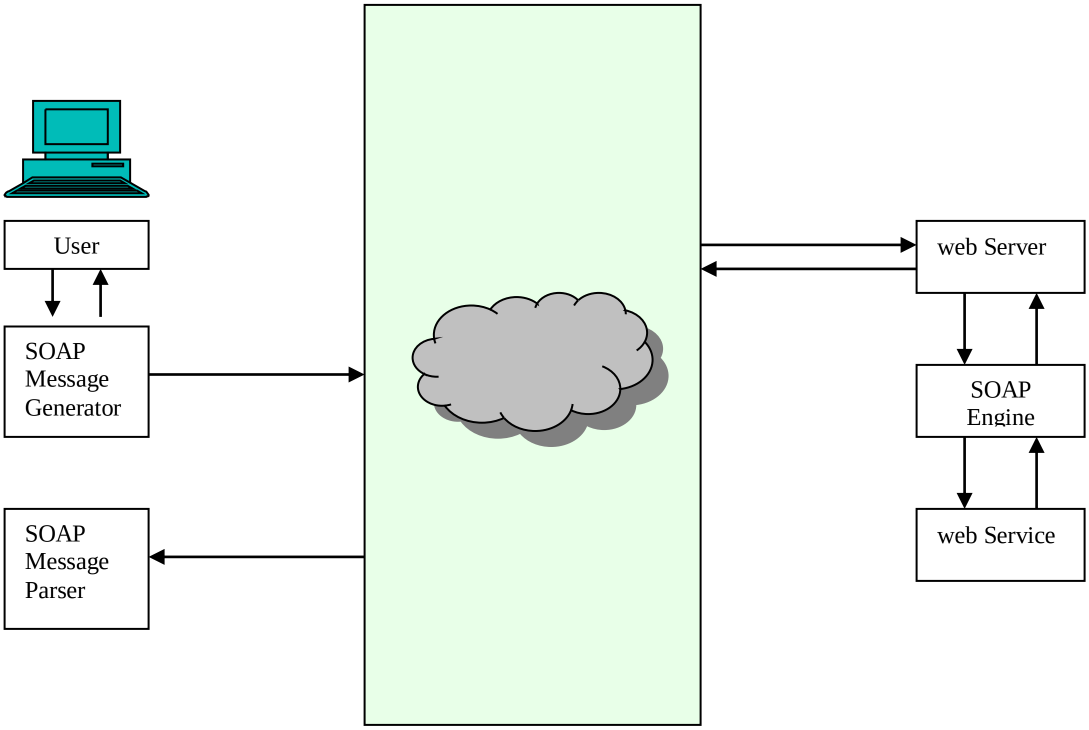


Figure 2: A client interacts with a web service via a web server such as Apache Tomcat or MS Internet Information Server.

Web services have two types of uses.

1. Reusable application components

Web services can offer application components such as currency conversion, weather reports or even language translation as services. Ideally, there will only be one type of each application component, and anyone can use it in their application.

2. Connecting existing software

Web services help solve the interoperability problem by giving disparate applications a standard way to link their data. Using web services one can exchange data between different applications and different platforms.

Every software application in the world can potentially talk to every other software application in the world using web service. This communication can take place across all the old boundaries of location, operating system, language, protocol, and so on.
In the application functionality of which is explained later in the document the concept of web services is used for both of these reasons. We have reused the methods provided by the pathway data source chosen and problem of interoperability and integration is solved in a reasonable way using web services. A pathway is a series of molecular interactions and reactions (or other biological relationships), often forming a network which is explained in further details in later sections.

Web Services in Bioinformatics

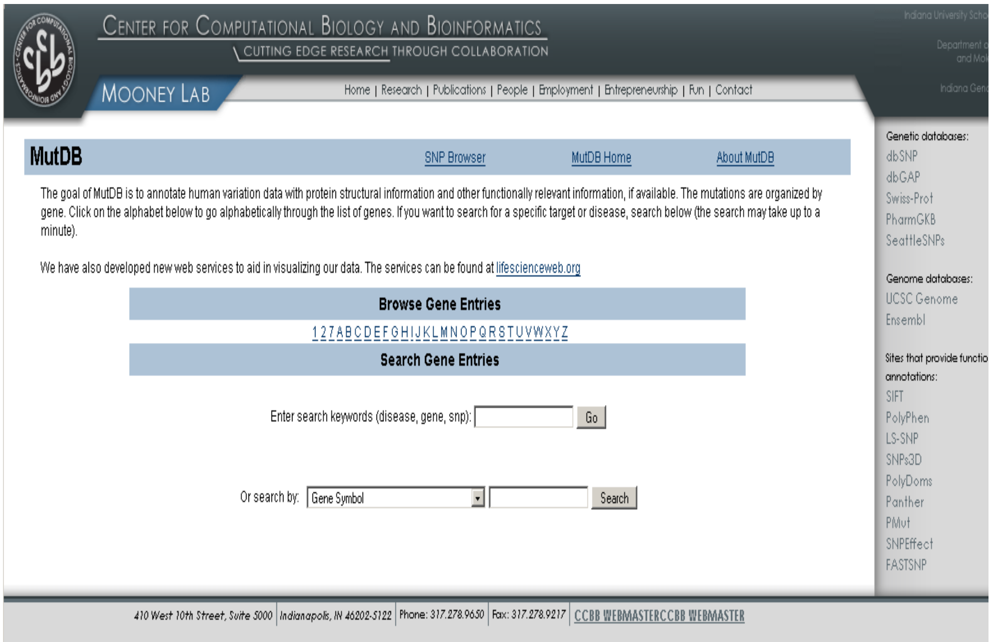
Though the concept of web services was developed basically for commercial purposes, the suitability of web services in academic bioinformatics was realized quite early (Stein, 2002). As of now there are numerous examples of web-services being used and provided in the field of bioinformatics. Some examples are a web service called DAS (distributed annotation system)(Dowell, Jokerst, Day, Eddy, & Stein, 2001), MutDB web services(Dantzer, Moad, Heiland, & Mooney, 2005), KEGG webservices (Wixon & Kell, 2000), European Bioinformatics Institute (EBI) web services (Pillai et al., 2005), etc.

In the article “Building a Bioinformatics Nation”(Nature 417, 119-120(9 May 2002)) Lincoln Stein has rightly pointed out the emerging importance of webservices in the field of Bioinformatics. The article asserts a web-services model will allow biological data to be fully exploited.

Genetic Variation Database

MutDB

Originally developed and designed at Stanford University by Sean Mooney, MutDB (Dantzer et al., 2005; S. D. Mooney & Altman, 2003) (http://mutdb.org/) is an online resource that integrates genetic variation from two public databases SWISS-PROT (Apweiler et al., 2004; Bairoch et al., 2005; Boeckmann et al., 2003; "The Universal Protein Resource (UniProt)," 2007)and dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) , and then annotates those variants with functionally relevant information. The SWISS-PROT protein knowledgebase connects amino acid sequences with the current knowledge in the life sciences.



The goal of MutDB is to annotate human variation data with protein structural information and other functionally relevant information, if available. The mutations are organized by gene. Click on the alphabet below to go alphabetically through the list of genes. If you want to search for a specific target or disease, search below (the search may take up to a minute).

We have also developed new web services to aid in visualizing our data. The services can be found at lifescienceweb.org

Enter search keywords (disease, gene, snp):

Or search by: Gene Symbol

410 West 10th Street, Suite 5000 | Indianapolis, IN 46202-5122 | Phone: 317.278.9650 | Fax: 317.278.9217

Figure 3: MutDB (http://mutdb.org/) is resource of genetic variation data.

Biological Pathway and Networks

A biological pathway is a series of events, molecular interactions and reactions maps for biological processes (or other biological relationships), often forming a network. . In reality, pathways are highly complicated with cross-talks among themselves and can occur a synchronously with other pathways in a biological system.

KEGG

The KEGG (Kanehisa & Goto, 2000; Wixon & Kell, 2000) , Kyoto Encyclopedia of Genes and Genotypes , is a resource of biological pathways and other biological data for several organisms, including human. It is an ongoing research database project operated jointly by the Bioinformatics Center, Institute for Chemical Research at Kyoto University and the Human Genome Center at the University of Tokyo.

KEGG is a computer representation of the biological system, consisting of building blocks and wiring diagrams, which can be utilized for modeling and simulation as well as for browsing and retrieval. The overall architecture of KEGG consists of four main databases: PATHWAY, GENES, LIGAND, BRITE and associated software. The databases are categorized as building blocks in the genomic space (Genes databases) and the chemical space (Ligand database), wiring diagrams in the network space (Pathway database) and ontologies for pathway reconstruction (Brite database).

KEGG is comprised of four independent databases namely Brite, Genes, Pathway, and Ligands. Besides the four main databases SSDB or sequence similarity database and various data retrieval techniques available are also discussed.

BRITE: The BRITE database stores functional hierarchies of biological systems. Information in this database is entered manually from already published materials.

BRITE differs from the PATHWAY database in that it includes relationships beyond molecular interactions and reactions. BRITE can be used as a supplement to PATHWAY to infer higher-order functions.

The mapping of genomic and molecular data to BRITE is done by the KEGG Orthology (KO) system. The KO system is how the genome becomes annotated. KEGG curators do this by consisting of decomposing all the genes in a complete genome into sets of likely orthologs.

There are three ways to navigate the relationships of the BRITE database. The first option is the search. Within the search mode, there are two more options. The bfind mode is a standard keyword search for organisms, such as human or bacteria. The second search type is the bget mode. This mode makes use of the DBGET information retrieval system. Information is accessed by using a database name in combination with an identifier. For example, if requesting information about the human gene for glucokinase, the search in bget mode would be hsa:gck. The name of the database is hsa (Homo sapiens) and the identifier is gck (glucokinase). This mode is only useful when the organism and gene name are known.

The second navigation type is browsing. KEGG BRITE contains links grouped by subjects such as genes and proteins, compounds and reactions, etc. There is also the option to navigate a hierarchy of organisms. It is here that the database name for each organism is obtained.

The third navigation type includes both searching and browsing, but is made easier through a downloadable desktop application called KegHier. The Java application works on multiple platforms, including Windows, Mac, and Linux. Only the hierarchies and relationships are contained within the application; ultimately, it links directly to pathways and structures contained in the database.

PATHWAY: PATHWAY database is a collection of manually drawn pathway maps representing current knowledge on the molecular interactions and reaction networks for metabolism, other cellular processes, and human diseases.

Biological systems are represented in KEGG by two types of graphs, called nested graphs and line graphs in theoretical computer science. The nested graph is a graph whose nodes can themselves be graphs. It is used for representing KEGG network hierarchy and for pathway reconstruction and functional inference. The line graph is a graph derived by interchanging nodes and edges of another graph. It represents the inherent complementarity of pathways, which can be viewed either as a network or genes (enzymes) or as a network of compounds, meaning that one can be generated from the other by the line graph transformation. The line graph is the basis for integrated analysis of genomic and chemical information.

Pathways in KEGG are organized in the following categories:

1. Metabolism
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Human Diseases

GENES: The GENES database is a collection of gene catalogs for all complete genomes and some partial genomes, generated from publicly available resources. The GENES database is the largest of the four KEGG databases. It is composed of over 1,164,610 genes derived from 35 eukaryotes, 353 bacteria, and 28 archaea. GENES is a collection of gene catalogs for all complete genomes as well as some partial genomes gathered from publicly available resources.

Because of the large number of entries in the database, the information architecture allows for categorization of the genes into high quality genomes, draft genomes (for eukaryotes only) and EST consensus contigs (for plants only). Some of the data was entered manually, while most was entered as part of a referenced data set.

There are four methods for searching for gene information in the database: Gene Catalogs, Organism Code, Gene Name Conversion, and Automatic Annotation. The Gene Catalog allows for locating a gene by its categorization, or the organism it was derived from. The Organism Code allows for genes to be located by a three letter KEGG organism code. The Gene Name Conversion allows for genes to be located by external references from other databases, such as UniProt and others. As well, genes can be located by cross-referencing genes by their KEGG Organism Codes, and external database references. The final method of locating a gene uses the KEGG orthology assignments and pathway mappings that are created by combining other KEGG databases with the GENES database.

LIGANDS: The Ligand database consists of chemical compounds and reactions and is designed to provide the linkage between chemical and biological aspects of life in light of enzymatic reactions. The database consists of six sections: Compound, Drug, Glycan, Reaction, RPair and Enzyme. The Compound database contains chemical structures of most know metabolic compounds and some pharmaceutical and environmental compounds. All chemical structures are manually entered, computationally verified and continuously updated. Currently the database contains 14,229 compounds, each of which is identified by the C number (accession number). The Drug database contains chemical structures of drugs, classification, therapeutic categories and target molecules. Currently the drug database contains 4103 drugs. The Glycan database contains carbohydrate structures. The Glycan pathway diagrams for metabolism of complex carbohydrates and metabolism of complex lipids linked to individual entries of carbohydrate structures. Each Glycan entry is identified by the G number (accession number) and the current total is 10,951 entries, among which only a few hundred were manually entered and linked to KEGG pathways. The rest represents unique structures derived from CarbBank. The Glycan database is maintained in a relational database with a structure drawing tool in Java. A database search is also made available based on newly developed algorithms for tree structure comparisons. The Reaction database contains reaction formulas for enzymic reactions, currently totaling 6,8ll. Each entry is identified by the R number (accession number) representing a unique reaction corresponding to sets of reactants and products represented by the C number in the Compound database or the G number in the Glycan database. This should be compared with the EC number, which may correspond to multiple reaction formulas. The EC number hierarchy is supposed to represent aspects of enzymatic reactions, but in reality it often contains aspects of enzyme molecules. Within the KEGG resource, these two aspects of EC numbers are clearly distinguished: R numbers for reactions and K numbers for molecules. KEGG is working to develop a new hierarchy, tentatively called RC (Reaction Classification), for understanding the chemistry of enzymic reactions. The Enzyme database contains enzyme nomenclature with numerous links to KEGG databases. It is generated semi-automatically from the enzyme nomenclature website (http://www.chem.qmul.ac.uk/iubmb/enzyme/). The role of this database within KEGG has diminished, but the EC number is still the simplest way to link to KEGG from outside resources.

SSDB: KEGG SSDB or Sequence Similarity Database contains the information about amino acid sequence similarities among all protein-coding genes in the complete genomes, which is computationally generated from the GENES database in KEGG. All possible pairwise genome comparisons are performed by the SSEARCH program, and the gene pairs with the Smith-Waterman similarity score of 100 or more are entered in SSDB, together with the information about best hits and bidirectional best hits (best-best hits). SSDB is thus a huge weighted, directed graph, which can be used for searching orthologs and paralogs, as well as conserved gene clusters with additional consideration of positional correlations on the chromosome. SSDB also contains precomputed motif patterns of Pfam and PROSITE for all protein coding genes.

Data Retrieval: DBGET is a simple database retrieval system for a diverse range of molecular biology databases. Here a database is considered as a set of entries, which may be stored in a single file or multiple files. Here definition of flat-file is not limited to text data; it also includes other types of data such as GIF images for KEGG pathways, Java graphics for genome maps and expression profiles, and 3D graphics for protein structures. This is accomplished by treating a collection of HTML files as a database. Data can also be downloaded as simple ftp commands from the KEGG website as tab-limited files. One of the unique features of KEGG is that it provides access to its data sources as web services.

The XML version of the pathway maps is available for both metabolic and regulatory pathways. These KEGG Markup Language (KGML) files provide graph information that can be used to computationally reproduce and manipulate KEGG pathway maps. KEGG is also accessible using SOAP-based web services.

Method	**Returned value**
get_elements_by_pathway(*string*:*pathway_id*)	ArrayOfPathwayElement • element_id; unique identifier of the object on the pathway (int) • type; type of the object ("gene", "enzyme" etc.) (string) • names; array of names of the object (ArrayOfstring) • components; array of element_ids of the group components (ArrayOfint)
get_element_relations_by_pathway(*string*:*pathway_id*)	ArrayOfPathwayElementRelation • element_id; unique identifier of the object on the pathway (int) • relation; kind of relation ("compound", "inhibition" etc.) (string) • type; type of relation ("+p", "--|" etc.) (string)

Table 3: Certain web-Methods provided by KEGG.

Ontology and Disease Ontology:

Ontology is the branch of metaphysics that deals with the nature of being. In other words, ontology is the centuries-old branch of philosophy that has as its subject the unchanging features of the universe. Ontology is the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality. For an information system, an ontology is a representation of some pre-existing domain of reality which reflects the properties of the objects within its domain in such a way that there obtains a systematic correlation between reality and the representation itself and is intelligible to a domain expert .Also it is formalized in a way that allows it to support automatic information processing An ontology in this sense is a thing made by a scientist or other domain expert. Thus, an ontology is a true-to-the-world representation of its domain. This stands in contrast to the more popular usages by many in the fields of information and computer science, which see an ontology as merely an ad-hoc model built for some specific purpose.

Computer-understandable ontologies are represented in logical languages, such as the W3C OWL (Ontology web Language) and the draft ISO standard, SCL (Simple Common Logic). However, logical languages are only a means to express content; they are themselves almost entirely devoid of information. This situation is much like how the natural language English relates to information expressed in English. It is the information being imparted in the words that drives how the individual words are selected and sequenced into sentences. It's not the language (or logic) that makes the difference, but how you use it. Ontology is one way to use language and logic more effectively.

From the point of view of information systems, Ontology can be considered both a means and the end; it is a modeling tool and the model.

Open Biomedical Ontologies (OBO): The OBO (Bada & Hunter, 2007; Moreira & Musen, 2007) flat file format is an ontology representation language. The concepts it models represent a subset of the concepts in the OWL description logic language(Moreira & Musen, 2007), with several extensions for meta-data modelling and the modelling of concepts that are not supported in DL languages. BioPax, another example of Data Exchange Format for Biological Pathways, uses OWL.

The format itself attempts to achieve the following goals:

- Human readability
- Ease of parsing
- Extensibility
- Minimal redundancy

Disease Ontology (http://diseaseontology.sourceforge.net/) is a controlled medical vocabulary. This ontology language is still in development phage. Disease Ontology is implemented as a directed acyclic graph (DAG) and utilizes the Unified Medical Language System (UMLS) as its immediate source vocabulary. Disease Ontology is stored in Open Biomedical Ontologies (OBO) format. It is being developed by Northwestern University.

METHODS

There are two logical segments in this section of the document. First segment deals with approaches and solutions to construction of a system to link genetic variation with biological pathways. In second segment of this section possibilities to integrating genetic variation dataset with clinical data have been explored and findings documented.

Construction of a system linking genetic variation with biological pathways

Problem Definition

Integrating data is one of the core problems that bioinformatics is dealing with today. Biologists are generating more and more amount of data everyday, yet there is not a standard mechanism or platform for doing this. Not having a standard platform is actually both good and bad. Good in the sense that there is an open flow of ideas and knowledge which is very important for the development of new technologies and standards themselves. But on the other hand it is often challenging when we try to integrate data from diverse platforms and domains. Below are three general ideas that we can use to deal with the challenges

Different Approaches to Solution to the Problem

Local Duplication

The most direct and perhaps the most common way to dealing with this problem is to duplicate data and to create a local copy of data set. This approach has some advantages as well as disadvantages. Keeping data local gives full control over the data, more freedom for designing an application and it is safe to that it will- with some exceptions though will be faster while accessing and manipulating. Most of the time biologist and bioinformaticians spend lot and most of their efforts in downloading, writing scripts to parse the downloaded data and saving and maintaining the local copy of data set. However, there is a pitfall. It is not only harder to maintain the data, there is absolutely no way of knowing if the data, and thus the research based on that, is still correct and up-to-date. The only way to make the data up-to-date is to import the data again which may create unnecessary work.

Standard Notation

This approach is analogous to having a common language of communication. There could always be an ideal situation where everything and everyone always works in way they should. But things rarely are like that in real world. Requirements come first and then standards. And standards are rarely limited to just one standard in one domain. Many cousins of standards are also born with same or overlapping domains. This is a very common scenario in the field of Bioinformatics. For example there are different standards with overlapping domains standards for representing systems biology data, pathways data and interactions data (Stromback & Lambrix, 2005). Some examples of such standards are BioPAX (http://www.biopax.org/)and SBML (http://sbml.org/index.psp) and PSI-MI (http://psidev.sourceforge.net/mi/xml/doc/user/). BioPAX is common exchange format for biological pathways data which is under development. SBML and PSI-MI are similar initiatives for systems biology and protein-protein interactions.

Actually having a well-defined, distinct standard should be seen as a first step towards designing solutions to data integration issues and not the complete solution it self. But this is an idealistic situation and we should always to striving towards it.

Component-based Software and web services

The background section has been used to talk about web services in detail. Inter-operability and seamless data exchange capability can very well be used for solving problem of data integration

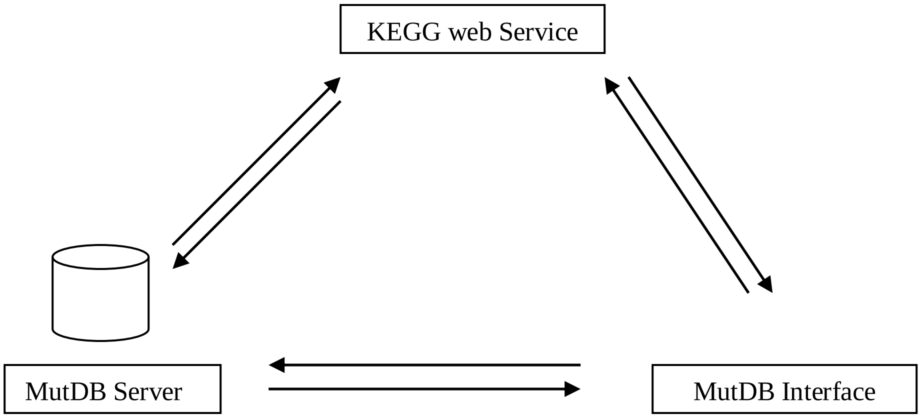


Figure 4: Using web services as a possible solution

Designing the solution

Different options as downloading dataset, using BioPax and using web services were thoroughly analyzed. Decision to use web services was made because of the following reasons.

- No need for updates. Data gets automatically updated.
- Less development time
- Less space utilization as the data resides on the remote server.

Perl was the language of choice because Perl is not just good programming language but also a good scripting language. CGI was used to communicate with web server.PHP was also studied as a replacement for CGI. But CGI/Perl was chosen over PHP due to former's versatility. CGI stands for common gateway interface and is free and one of the oldest web development as opposed to ASP and JSP that are proprietary and not exactly free.

Different parameters of genes and SNPs such as Entrez gene id and Entrez gene no as well Refseq IDs were considered to be the connecting parameter between MutDB and KEGG. Scripts were written in Perl to design a solution for integrating genetic variation dataset from MutDB to biological pathway dataset from KEGG. The local database storing genetic variation data is in mySQL database. To deal with performance issues as well as implement certain logic visualized for the application some pathways dataset was locally parsed and saved.

Construction of a system linking genetic variation to clinical data

Problem definition

There has been lot of research in field of genetics to try to associate some specific gene and their variants to known human diseases. There is lot of data generated over the years through ongoing research in science. Also, in recent years there have attempts to consolidate all the information together in one place. OMIM(Hamosh et al., 2002) is good example of a resource relating genetics data with specific disease phenotypes, but unfortunately it is not easy to use within informatics systems. Mostly these attempts have been made to only gene level and not gene variants level. The goal of this research is to figure out if there could be an informatics approach to associate standard notation of disease terms ie Disease Ontology to specific gene variants level phenotype dataset such as amino acid substitutions from Swissprot having disease related phenotype.

Designing the solution

The Disease Ontology (DO) used as a flat file and DO-Edit and DAG-viewer (DAG or directed acyclic graph) was used to visualize the whole structure. Figure 5 shows visualization of Disease Ontology using DAG viewer. We also obtained Gene to disease ontology terms mapped data from one of our collaborators, Predrag Radivojac. The mapping was based on exact string matching (OMIM to DO) whenever there was an exact match. If the matches were close but not exact, some manual validation was done. Still, there were a certain possible number of OMIM proteins which wouldn't have DO annotation because of mismatches in the datasets. Swiss-Prot was also used to map proteins to DO manually (for example first Swiss-Prot descriptions were looked and then an appropriate DO category was found). So, all diseases associated with every protein constitute a sub-DAG of a DO DAG. The graph structure from the DO file use "is_a" relationship). The disease ontology flat file was parsed and stored in a local mySQL database. Here a relational model for the database was implemented shown later in figure 7 and thus is-a relation ship is implemented as relational table and meaningful associations can be derived from it using simple SQL statements.

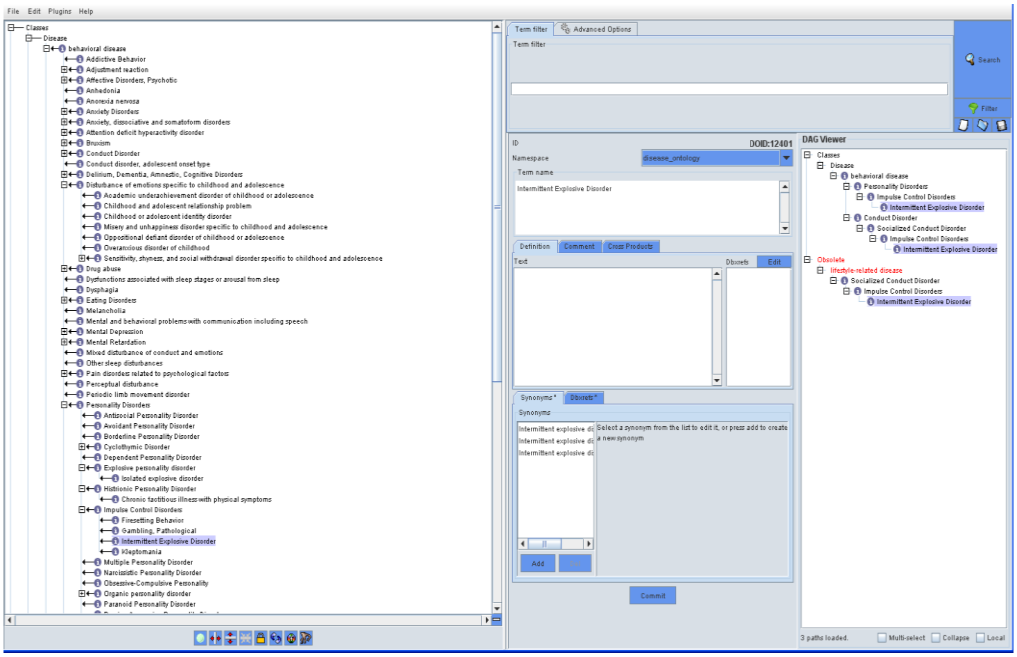
An informatics algorithm was designed to make the associations between the two datasets programmatically. It was hoped that Swissprot phenotype text which is actually a combination of an acronym and some descriptive text could be matched with either DO text or with DO synonyms text. If not, Pubmed title text could be search to match DO text or DO synonyms text. As an optional or remedial measure OMIM entry could be taken to be matched DO text or DO synonyms text. As an additional step a thesaurus could be created for matching acronyms to diseases.

There were several obstacles to implementing these steps in the algorithm, mainly accountable to the nature of diverse and richly verbose data. So decision was made to take different route. There was suggestion to make a ‘community-driven’ interface to the collaborator’s data mentioned above and consequently this whole system could be to build further associations of the two datasets to further enrich and verify the data we already have.

Software Tools Used/Referred or explored as an option to solve problem at hand

OBO-EDIT

OBO-Edit (Day-Richter, Harris, Haendel, & Lewis, 2007) is an open source ontology editor written in Java. OBO-Edit is optimized for the OBO biological ontology file format. It has an easy to use editing interface, a simple but fast reasoner, and powerful search capabilities. OBO-Edit is developed by the Berkeley Bioinformatics and Ontologies Project, and is funded by the Gene Ontology Consortium. This software was used to visualize disease ontology graph.


File
Edit
Plugins
Help
Classes
Disease
behavioral disease
Addictive Behavior
Adjustment reaction
Affective Disorders, Psychotic
Anhedonia
Anorexia nervosa
Anxiety Disorders
Anxiety, dissociative and somatoform disorders
Attention deficit hyperactivity disorder
Bruxism
Conduct Disorder
Conduct disorder, adolescent onset type
Delirium, Dementia, Amnestic, Cognitive Disorders
Disturbance of emotions specific to childhood and adolescence
Academic underachievement disorder of childhood or adolescence
Childhood and adolescent relationship problem
Childhood or adolescent identity disorder
Misery and unhappiness disorder specific to childhood and adolescence
Oppositional defiant disorder of childhood or adolescence
Overanxious disorder of childhood
Sensitivity, shyness, and social withdrawal disorder specific to childhood and adolescence
Drug abuse
Dysfunctions associated with sleep stages or arousal from sleep
Dysphagia
Eating Disorders
Melancholia
Mental and behavioral problems with communication including speech
Mental Depression
Mental Retardation
Mixed disturbance of conduct and emotions
Other sleep disturbances
Pain disorders related to psychological factors
Perceptual disturbance
Periodic limb movement disorder
Personality Disorders
Antisocial Personality Disorder
Avoidant Personality Disorder
Borderline Personality Disorder
Cyclothymic Disorder
Dependent Personality Disorder
Explosive personality disorder
Isolated explosive disorder
Histrionic Personality Disorder
Chronic factitious illness with physical symptoms
Impulse Control Disorders
Firesetting Behavior
Gambling, Pathological
Intermittent Explosive Disorder
Kleptomania
Multiple Personality Disorder
Narcissistic Personality Disorder
Obsessive-Compulsive Personality
Organic personality disorder
Paranoid Personality Disorder
Term filter
Advanced Options
Term filter
Search
Filter
Definition
Comment
Cross Products
Text
Dbxrefs
Edit
Synonyms *
Dbxrefs *
Synonyms
Add
Del
Commit
DAG Viewer
Classes
Disease
behavioral disease
Personality Disorders
Impulse Control Disorders
Intermittent Explosive Disorder
Conduct Disorder
Socialized Conduct Disorder
Impulse Control Disorders
Intermittent Explosive Disorder
Obsolete
lifestyle-related disease
Socialized Conduct Disorder
Impulse Control Disorders
Intermittent Explosive Disorder
☐ Multi-select
☐ Collapse
☐ Local

ID DOID:12401

Namespace disease_ontology

Term name

Intermittent Explosive Disorder

Intermittent explosive di

Intermittent explosive di

Intermittent explosive di

Select a synonym from the list to edit it, or press add to create a new synonym

3 paths loaded.

Figure 5: OBO Edit showing DO hierarchy

mySQL Administrator

MySQL Administrator is a powerful visual administration console. This tool was extensively used for different database related tasks such as data insertion, table creation etc.

Anjuta IDE

Anjuta IDE was used for software development in Perl. Anjuta is versatile and free Integrated Development Environment (IDE) software supporting diverse languages from Perl to C++ on GNU/Linux. It has been written for GTK/GNOME and features a number of advanced programming facilities. These include project management, application wizards, an on-board interactive debugger, and a powerful source editor with source browsing and syntax highlighting.

Software Techniques Specifications

The programming for the system was done in Perl 5.8. CGI (Common Gateway Interface) scripts were used to communicate with the web server. CGI is technique that a web client (commonly a web browser) uses to communicate to web server. SOAP based web services were used mostly to connect mutation and pathways datasets.

SQL queries were mostly written to manipulate data in different ways. There is scope for using triggers and procedure especially for the second part of the project (linking mutations to diseases) provided mySQL has a newer version.

The first part of the project (linking gene variation with biological pathways) uses a major part of service oriented architecture and web services. This is good implementation of creating a “data-mashup” in Bioinformatics.

Server Hardware and Software specifications

The application is hosted on Apache web server (Apache 2.0.46-67) and Database is housed in mySQL 3.23.58-16. The machine has 2-XEON 2.86 hz processor and 2 GB RAM. The Operating System on the machine is RedHat Enterprise Linux AS 3.

RESULTS AND CONCLUSIONS

The study of the interactions between the components of biological systems and biological pathways has become increasingly important. It is known and accepted by scientists that it as important to study genes/ or other biological entities as interacting systems, as in isolation.

The above work has ideas rooted in this thinking. It has resulted in integration of genetic variation dataset with pathways data set. One of the tangible results of this project is enhancement of MutDB server available at www.mutdb.org. Pathways can be browsed by getting to any gene page (using Entrez geneid or gene symbol etc as search keys) and clicking on link having text as “Visualize pathway”. This is a useful addition to MutDB, the genetic variation dataset. Once user clicks on the link, the next page displays all possible pathways in which the gene appears. Further a particular pathway can be viewed by clicking on the link displaying title of the pathway. This pathway page not only displays the pathway but also highlights all the genes on that pathway that are found to have associated disease-phenotypes as per Swiss-prot along with the list of all such genes and their EC numbers and disease associated phenotypes. In all about 3587 amino acid substitutions have been mapped to about 200 pathways from KEGG. One case study for the MAPK1 gene is attached in the appendix A. This system also serves the purpose of having an easy way to access different pathways that a gene may be involved in. Following figure shows an example of such a pathway.

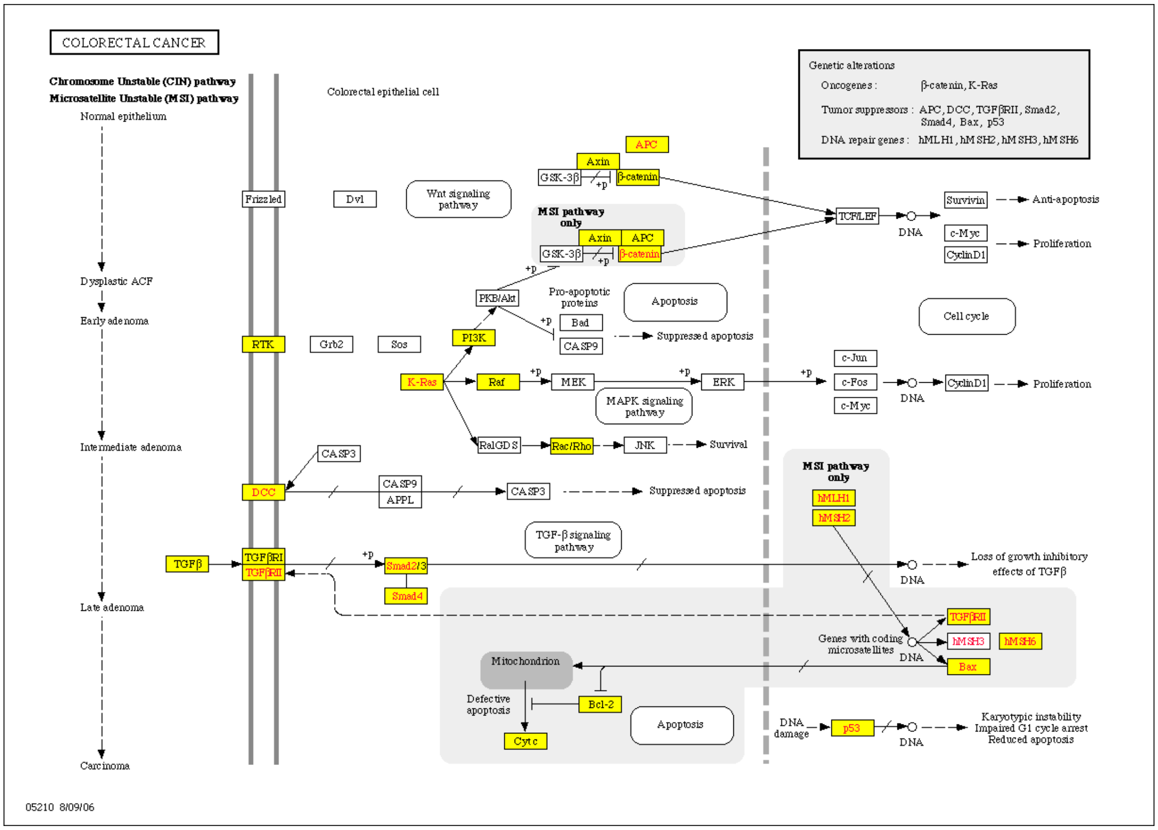


Figure 6: Figure displaying Colorectal Cancer pathway. This is one example of around 200 pathways that have been integrated with genetic variation data (from MutDB). The rectangles that are colored yellow are the genes that have some disease associated phenotype as per SwissProt.

As mentioned earlier the second part of the project attempts to connect mutations with disease ontology data which in itself not just a simple data integration but an attempt to build better system for building models for different studies for prediction of effects and exploration genetic variation data in altogether different ways. Other less tangible, yet important, result of the project is the construction of a local database of disease ontology data set. The original disease ontology file downloadable from the web site is in special format in which some ontology are described namely open biomedical ontology (or OBO) which is described in more detail in background section. This file is a specially formatted file. C# was used to parse the data in a relational format.

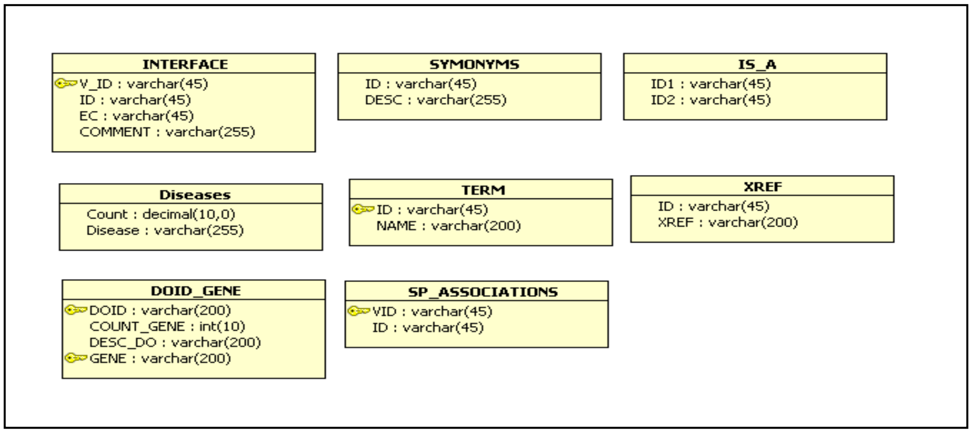


Figure 7: DO schema

Since user interface is still under consideration it is not yet available and can be seen as part of future work. Another outcome of this part of the project was a study that was done to determine a work flow for annotation of mutations with clinical data. The procedure is explained in more detail in the methods section of the current document. Result of the study in addition to a better understanding of the problem was the recognition of the need to have new options for achieving the mapping and the fact that this will not be a single step process. It has to be done iteratively and incrementally. This is discussed in more details in discussion and future section of the document. Ultimately this integration of data will enable computational prediction and creation of models of diseases.

DISCUSSION AND FUTURE WORK

Advances in genomic technologies and bioinformatics, combined with an enormous reduction in cost, have led to genome sequencing projects of different species. It is anticipated that the sequencing of individual human genomes will ultimately be required for a comprehensive genetic understanding of disease, but at present the cost of such efforts is prohibitive. Thus, for now, the discovery and study of functionally important genetic variation continues to be an important endeavor in the field of medicine.

This study is an effort in this direction. It is an initiative to understand and create a framework for associating genetic variation that may often relate to disease to biological pathways and standardized notation of human diseases.

A grand challenge in the undoubtedly post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction and creation of models of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Integrating mutation data with pathways is a step in this direction.

Wherever data are dealt with especially to base top-quality scientific research and discoveries, three issues relating data need to be addressed – consistency, completeness and need for data to be current or up-to-date. It can be said with good confidence that data is consistent and up-to-date as all data come from reliable international organizations such as KEGG and National Center for Biotechnology Information (NCBI) and use of web-services lets us to keep data current at least as far as pathways are concerned. But the issue of completeness of data is still needs to be addressed. The data can be brought closer to being complete when there are more than one pathways data sources integrated with mutation data.

This could be done by using different approaches for different data sources. For example some data sources can provide ways such as use of web services or XML based graphics to be able to display directly or after locally downloading pathway data. A more reasonable way of doing this could be developing and using a system of data interchange and language in that domain that automatically lets different applications communicate with one another. One such initiative is BioPAX. BioPAX is a collaborative effort to create a data exchange format for biological pathway data. So once we create an application that can read and understand data in BioPAX format it will be possible to use different data sources with ease provided they all provide data in same standard format namely BioPAX. This whole approach could still be made easier by using the component based approach of software architecture explained before in the document. Once this infrastructure is ready it will lead to seamless data integration.

Relating mutation data to a standardized notation and classification of diseases is a very difficult yet an important endeavor. There were different approaches that were studied in this study. Mapping of a system of classification of diseases with gene variants was attempted in many different ways such as development of an automated algorithm based on string matching of different terms. This research is ongoing and many more approaches need to considered. The gist of all efforts is that such a mapping can not be absolute. It has to be incremental. Also mapping has to be multi-dimensional ie more than one attributes of both sides should be considered while mapping. Also cumulative positive results (mapped term) should be given utmost significance. This is one of those areas of bioinformatics where lot of domain expert knowledge is absolutely needed. Experts can be of great help while doing the mapping using the incremental and iterative process for development of such mapping. To be more elaborate one option to approaching the problem could be having an interface build to whatever current data available. Then experts can help in doing new mappings as well as verifying original mappings using the interface. With these new mappings data could be designed to have automatic changes as per expert's suggestion. Once this system is in place along with other additional mapping efforts will lead to almost complete and accurate mapping of the two data sets in consideration.

We must think about more elaborate plans for annotating mutations with clinical phenotypes in a consistent and useful way. Development of such systems can be seen as the first step towards overall complete and better classified system for mutations where better and accurate models for systems biology and translational research can be created. From the information systems infrastructure point, this project study can be seen as stepping stone to building more sophisticated and novel annotation pipeline for genetic variation data. The idea of having a pipeline infrastructure in place will help in focusing efforts in discovery aspect of science. The research in this project as well as logic and code generated from the project can further be enhanced to create new web-services and/or software tools that will provide an infrastructure to genetic variations related studies. One example of such a web services can be developed where a user can upload list of disease gene and query against interaction data, pathway data and mutation data. Following figure demonstrates one such high-level design of implementing web services.

REFERENCES

Alexander Davis, D. Z. (2002). A Comparative Study of DCOM and SOAP. *IEEE Computer Society*

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res, 32*(Database issue), D115-119.

Bada, M., & Hunter, L. (2007). Enrichment of OBO ontologies. *J Biomed Inform, 40*(3), 300-315.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res, 33*(Database issue), D154-159.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res, 31*(1), 365-370.

Chakravarti, A. (2001). Single nucleotide polymorphisms: . . .to a future of genetic medicine. *Nature*.

Chang. (2000). A Large Scale Distributed Object Architecture - CORBA & COM for Real Time Systems. *IEEE Computer Society*

Dantzer, J., Moad, C., Heiland, R., & Mooney, S. (2005). MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res, 33*(Web Server issue), W311-314.

Day-Richter, J., Harris, M. A., Haendel, M., & Lewis, S. (2007). OBO-Edit - An Ontology Editor for Biologists. *Bioinformatics*.

Douglas, J. A., Levin, A. M., Zuhlke, K. A., Ray, A. M., Johnson, G. R., Lange, E. M., et al. (2007). Common Variation in the BRCA1 Gene and Prostate Cancer Risk. *Cancer Epidemiol Biomarkers Prev, 16*(7), 1510-1516.

Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R., & Stein, L. (2001). The distributed annotation system. *BMC Bioinformatics, 2*, 7.

Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., & McKusick, V. A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res, 30*(1), 52-55.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30.

Mooney, S. (2005). Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Briefings in Bioinformatics*.

Mooney, S. D., & Altman, R. B. (2003). MutDB: annotating human variation with functionally relevant data. *Bioinformatics, 19*(14), 1858-1860.

Moreira, D. A., & Musen, M. A. (2007). OBO to OWL: A Protege OWL Tab to Read/Save OBO Ontologies. *Bioinformatics*.

Pillai, S., Silventoinen, V., Kallio, K., Senger, M., Sobhany, S., Tate, J., et al. (2005). SOAP-based services provided by the European Bioinformatics Institute. *Nucleic Acids Res, 33*(Web Server issue), W25-28.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature, 405*(6788), 847-856.

Silver, P. A., & Way, J. C. (2007). Molecular Systems Biology in Drug Development. *Clin Pharmacol Ther*.

Stein, L. (2002). Creating a bioinformatics nation. *Nature, 417*(6885), 119-120.

Stoneking, M. (2001). Single nucleotide polymorphisms. From the evolutionary past. *Nature, 409*(6822), 821-822.

Stromback, L., & Lambrix, P. (2005). Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics, 21*(24), 4401-4407.

Taylor, J. G., Choi, E. H., Foster, C. B., & Chanock, S. J. (2001). Using genetic variation to study human disease. *Trends Mol Med, 7*(11), 507-512.

The Universal Protein Resource (UniProt). (2007). *Nucleic Acids Res, 35*(Database issue), D193-197.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351.

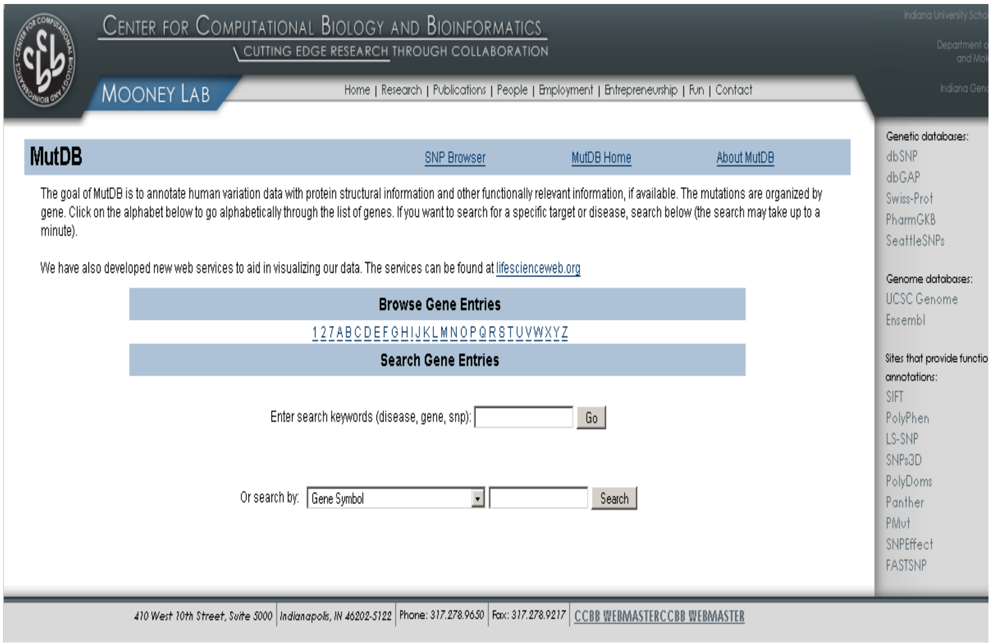
Wardell, M. R., Suckling, P. A., & Janus, E. D. (1982). Genetic variation in human apolipoprotein E. *J Lipid Res, 23*(8), 1174-1182.

Wixon, J., & Kell, D. (2000). The Kyoto encyclopedia of genes and genomes--KEGG. *Yeast, 17*(1), 48-55.

Zerhouni, E. A. (2005). Translational and clinical science--time for a new vision. *N Engl J Med, 353*(15), 1621-1623.

APPENDIX A

This appendix attempts to show how the system linking genetic variation to biological pathways works. In this example a gene MAPK1 is chosen as case study and some of the pathways associated with it are shown. In addition, the system also highlights all of the genes with some type of phenotype that can be related to disease associated with it from SwissProt (through MutDB) on pathway. With the list of all related genes that are hyperlinks, users can further explore other genes and their related pathways and disease phenotype associated with them. List of genes for each pathway is not complete.



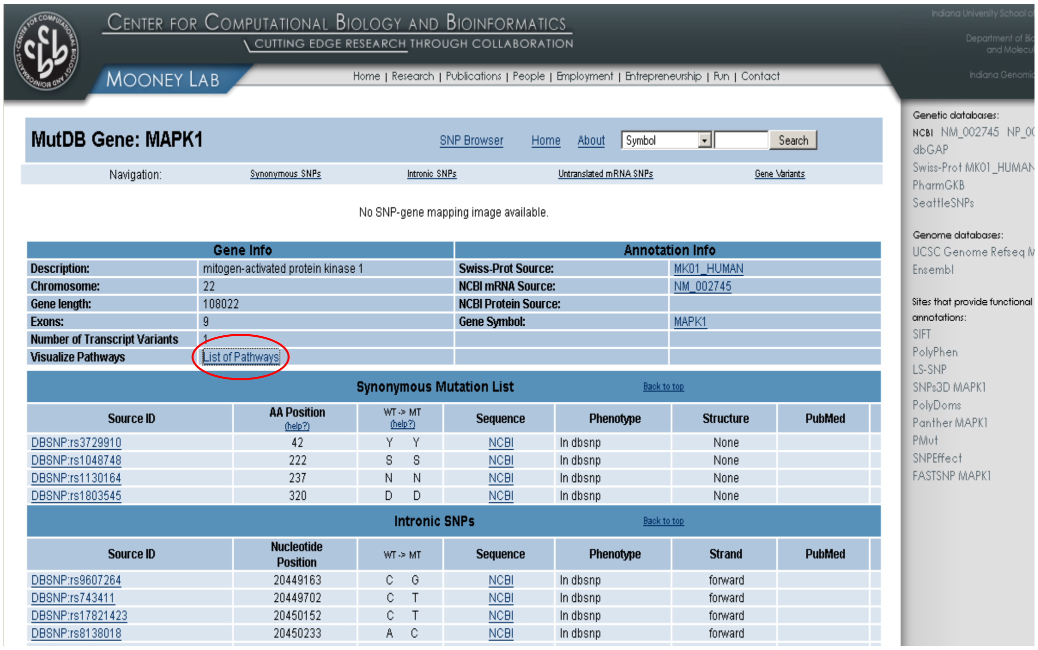
The goal of MutDB is to annotate human variation data with protein structural information and other functionally relevant information, if available. The mutations are organized by gene. Click on the alphabet below to go alphabetically through the list of genes. If you want to search for a specific target or disease, search below (the search may take up to a minute).

We have also developed new web services to aid in visualizing our data. The services can be found at lifescienceweb.org

Enter search keywords (disease, gene, snp):

Or search by: Gene Symbol

410 West 10th Street, Suite 5000 | Indianapolis, IN 46202-5122 | Phone: 317.278.9650 | Fax: 317.278.9217



MutDB Gene: MAPK1

No SNP-gene mapping image available.

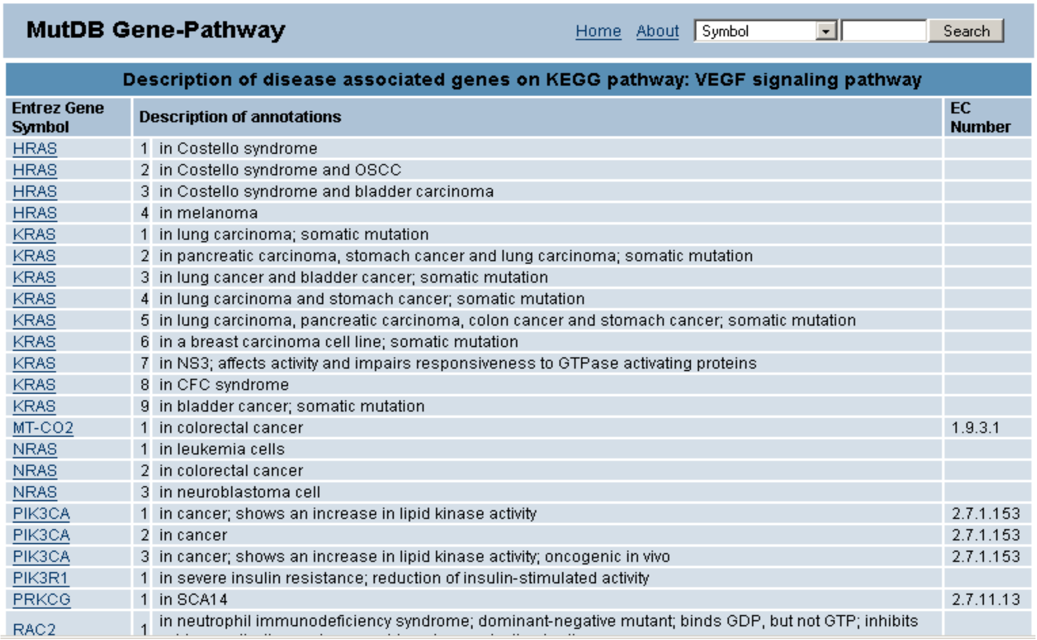
Gene Info		Annotation Info	
Description:	mitogen-activated protein kinase 1	Swiss-Prot Source:	MK01_HUMAN
Chromosome:	22	NCBI mRNA Source:	NM_002745
Gene length:	108022	NCBI Protein Source:	
Exons:	9	Gene Symbol:	MAPK1
Number of Transcript Variants	1		
Visualize Pathways	List of Pathways		

Synonymous Mutation List

Source ID	AA Position (help?)	WT -> MT (help?)	Sequence	Phenotype	Structure	PubMed
DBSNP:rs3729910	42	Y Y	NCBI	In dbsnp	None	
DBSNP:rs1048748	222	S S	NCBI	In dbsnp	None	
DBSNP:rs1130164	237	N N	NCBI	In dbsnp	None	
DBSNP:rs1803545	320	D D	NCBI	In dbsnp	None	

Intronic SNPs

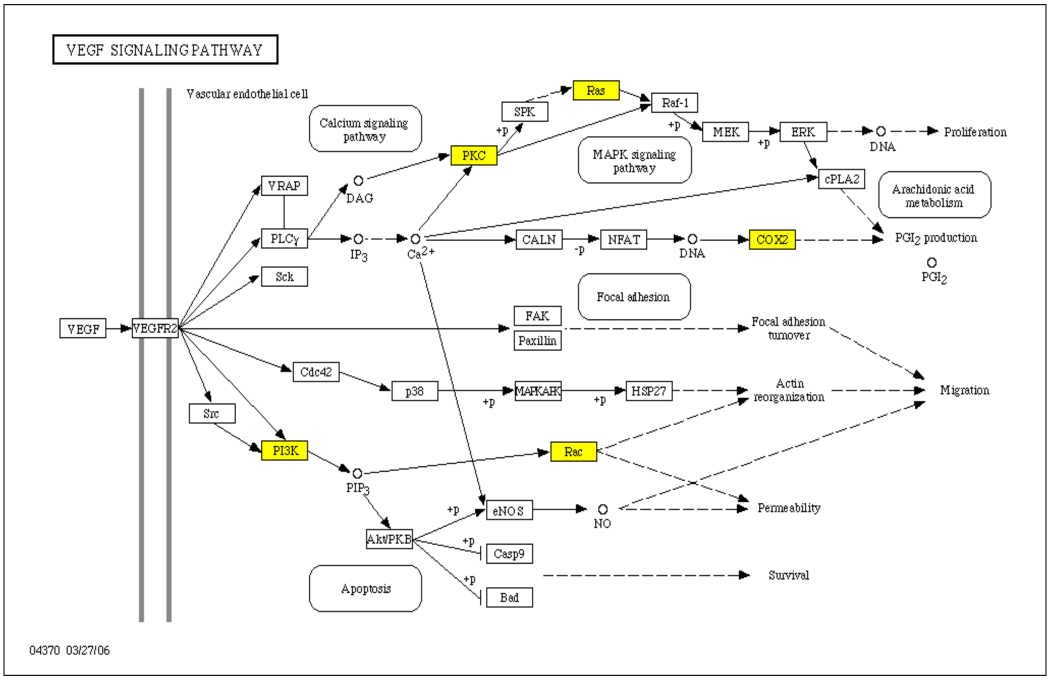
Source ID	Nucleotide Position	WT -> MT	Sequence	Phenotype	Strand	PubMed
DBSNP:rs9607264	20449163	C G	NCBI	In dbsnp	forward	
DBSNP:rs743411	20449702	C T	NCBI	In dbsnp	forward	
DBSNP:rs17821423	20450152	C T	NCBI	In dbsnp	forward	
DBSNP:rs8138018	20450233	A C	NCBI	In dbsnp	forward	



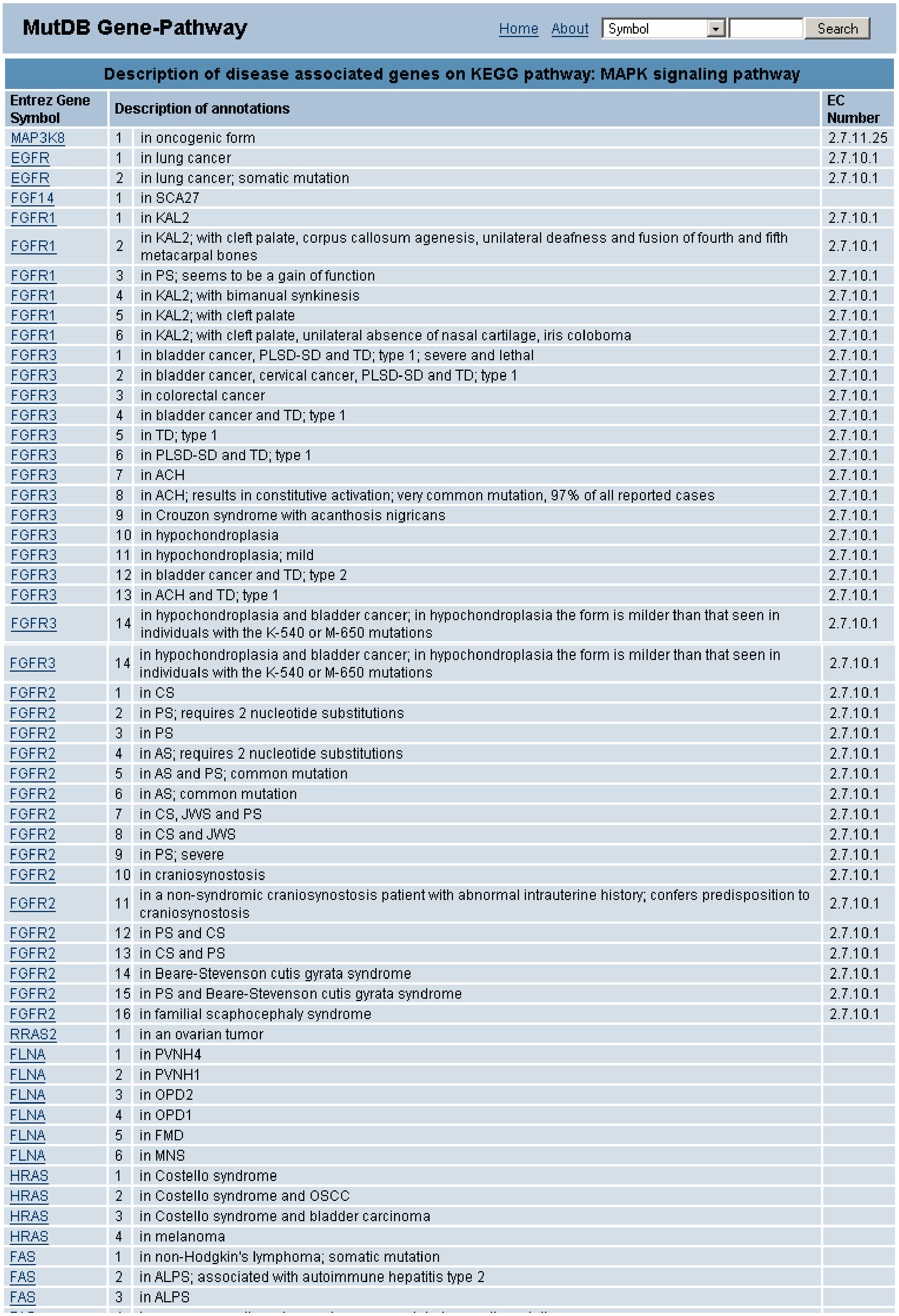
MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: VEGF signaling pathway

Entrez Gene Symbol		Description of annotations	EC Number
HRAS	1	in Costello syndrome	
HRAS	2	in Costello syndrome and OSCC	
HRAS	3	in Costello syndrome and bladder carcinoma	
HRAS	4	in melanoma	
KRAS	1	in lung carcinoma; somatic mutation	
KRAS	2	in pancreatic carcinoma, stomach cancer and lung carcinoma; somatic mutation	
KRAS	3	in lung cancer and bladder cancer; somatic mutation	
KRAS	4	in lung carcinoma and stomach cancer; somatic mutation	
KRAS	5	in lung carcinoma, pancreatic carcinoma, colon cancer and stomach cancer; somatic mutation	
KRAS	6	in a breast carcinoma cell line; somatic mutation	
KRAS	7	in NS3; affects activity and impairs responsiveness to GTPase activating proteins	
KRAS	8	in CFC syndrome	
KRAS	9	in bladder cancer; somatic mutation	
MT-CO2	1	in colorectal cancer	1.9.3.1
NRAS	1	in leukemia cells	
NRAS	2	in colorectal cancer	
NRAS	3	in neuroblastoma cell	
PIK3CA	1	in cancer; shows an increase in lipid kinase activity	2.7.1.153
PIK3CA	2	in cancer	2.7.1.153
PIK3CA	3	in cancer; shows an increase in lipid kinase activity; oncogenic in vivo	2.7.1.153
PIK3R1	1	in severe insulin resistance; reduction of insulin-stimulated activity	
PRKCG	1	in SCA14	2.7.11.13
RAC2	1	in neutrophil immunodeficiency syndrome; dominant-negative mutant; binds GDP, but not GTP; inhibits	



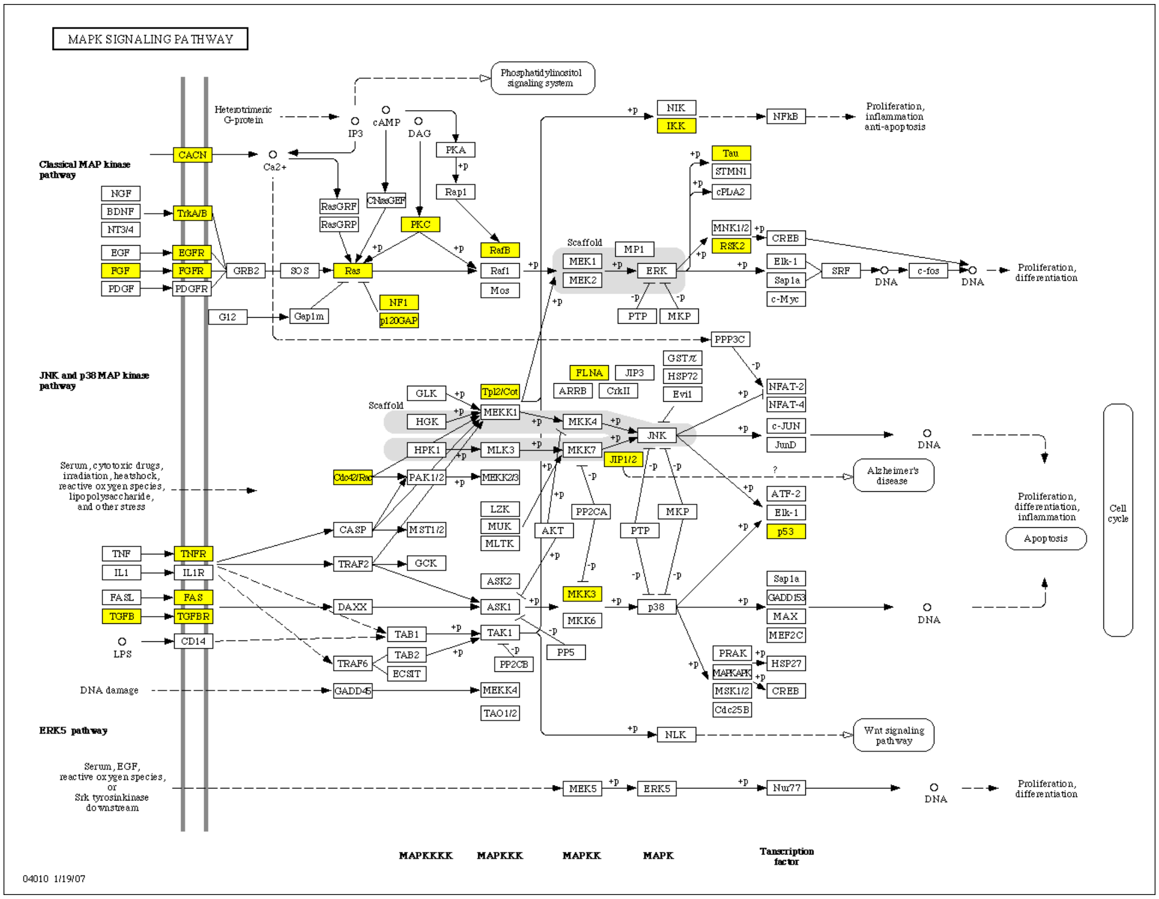
🟨 Genes having annotated mutations



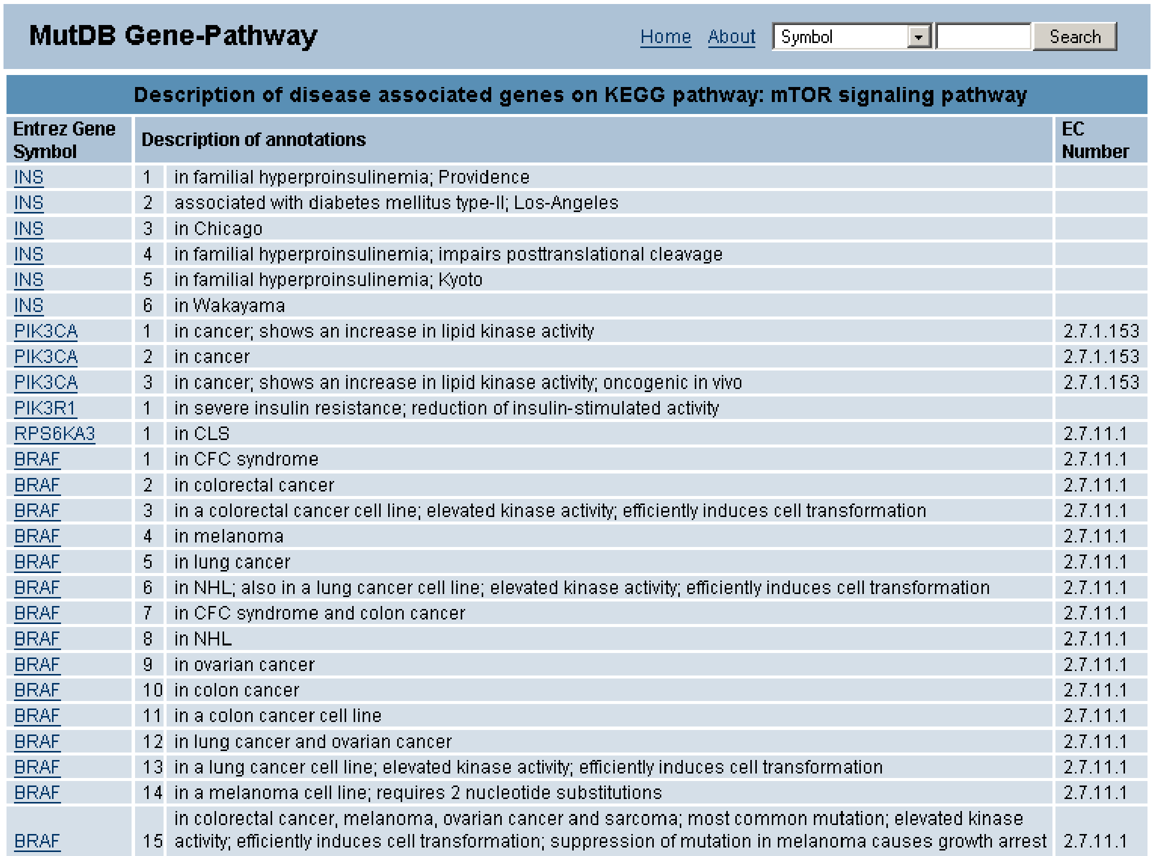
MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: MAPK signaling pathway

Entrez Gene Symbol	Description of annotations		EC Number
MAP3K8	1	in oncogenic form	2.7.11.25
EGFR	1	in lung cancer	2.7.10.1
EGFR	2	in lung cancer; somatic mutation	2.7.10.1
FGF14	1	in SCA27	
FGFR1	1	in KAL2	2.7.10.1
FGFR1	2	in KAL2; with cleft palate, corpus callosum agenesis, unilateral deafness and fusion of fourth and fifth metacarpal bones	2.7.10.1
FGFR1	3	in PS; seems to be a gain of function	2.7.10.1
FGFR1	4	in KAL2; with bimanual synkinesis	2.7.10.1
FGFR1	5	in KAL2; with cleft palate	2.7.10.1
FGFR1	6	in KAL2; with cleft palate, unilateral absence of nasal cartilage, iris coloboma	2.7.10.1
FGFR3	1	in bladder cancer, PLSD-SD and TD; type 1; severe and lethal	2.7.10.1
FGFR3	2	in bladder cancer, cervical cancer, PLSD-SD and TD; type 1	2.7.10.1
FGFR3	3	in colorectal cancer	2.7.10.1
FGFR3	4	in bladder cancer and TD; type 1	2.7.10.1
FGFR3	5	in TD; type 1	2.7.10.1
FGFR3	6	in PLSD-SD and TD; type 1	2.7.10.1
FGFR3	7	in ACH	2.7.10.1
FGFR3	8	in ACH; results in constitutive activation; very common mutation, 97% of all reported cases	2.7.10.1
FGFR3	9	in Crouzon syndrome with acanthosis nigricans	2.7.10.1
FGFR3	10	in hypochondroplasia	2.7.10.1
FGFR3	11	in hypochondroplasia; mild	2.7.10.1
FGFR3	12	in bladder cancer and TD; type 2	2.7.10.1
FGFR3	13	in ACH and TD; type 1	2.7.10.1
FGFR3	14	in hypochondroplasia and bladder cancer; in hypochondroplasia the form is milder than that seen in individuals with the K-540 or M-650 mutations	2.7.10.1
FGFR3	14	in hypochondroplasia and bladder cancer; in hypochondroplasia the form is milder than that seen in individuals with the K-540 or M-650 mutations	2.7.10.1
FGFR2	1	in CS	2.7.10.1
FGFR2	2	in PS; requires 2 nucleotide substitutions	2.7.10.1
FGFR2	3	in PS	2.7.10.1
FGFR2	4	in AS; requires 2 nucleotide substitutions	2.7.10.1
FGFR2	5	in AS and PS; common mutation	2.7.10.1
FGFR2	6	in AS; common mutation	2.7.10.1
FGFR2	7	in CS, JWS and PS	2.7.10.1
FGFR2	8	in CS and JWS	2.7.10.1
FGFR2	9	in PS; severe	2.7.10.1
FGFR2	10	in craniosynostosis	2.7.10.1
FGFR2	11	in a non-syndromic craniosynostosis patient with abnormal intrauterine history; confers predisposition to craniosynostosis	2.7.10.1
FGFR2	12	in PS and CS	2.7.10.1
FGFR2	13	in CS and PS	2.7.10.1
FGFR2	14	in Beare-Stevenson cutis gyrata syndrome	2.7.10.1
FGFR2	15	in PS and Beare-Stevenson cutis gyrata syndrome	2.7.10.1
FGFR2	16	in familial scaphocephaly syndrome	2.7.10.1
RRAS2	1	in an ovarian tumor	
FLNA	1	in PVNH4	
FLNA	2	in PVNH1	
FLNA	3	in OPD2	
FLNA	4	in OPD1	
FLNA	5	in FMD	
FLNA	6	in MNS	
HRAS	1	in Costello syndrome	
HRAS	2	in Costello syndrome and OSCC	
HRAS	3	in Costello syndrome and bladder carcinoma	
HRAS	4	in melanoma	
FAS	1	in non-Hodgkin's lymphoma; somatic mutation	
FAS	2	in ALPS; associated with autoimmune hepatitis type 2	
FAS	3	in ALPS	



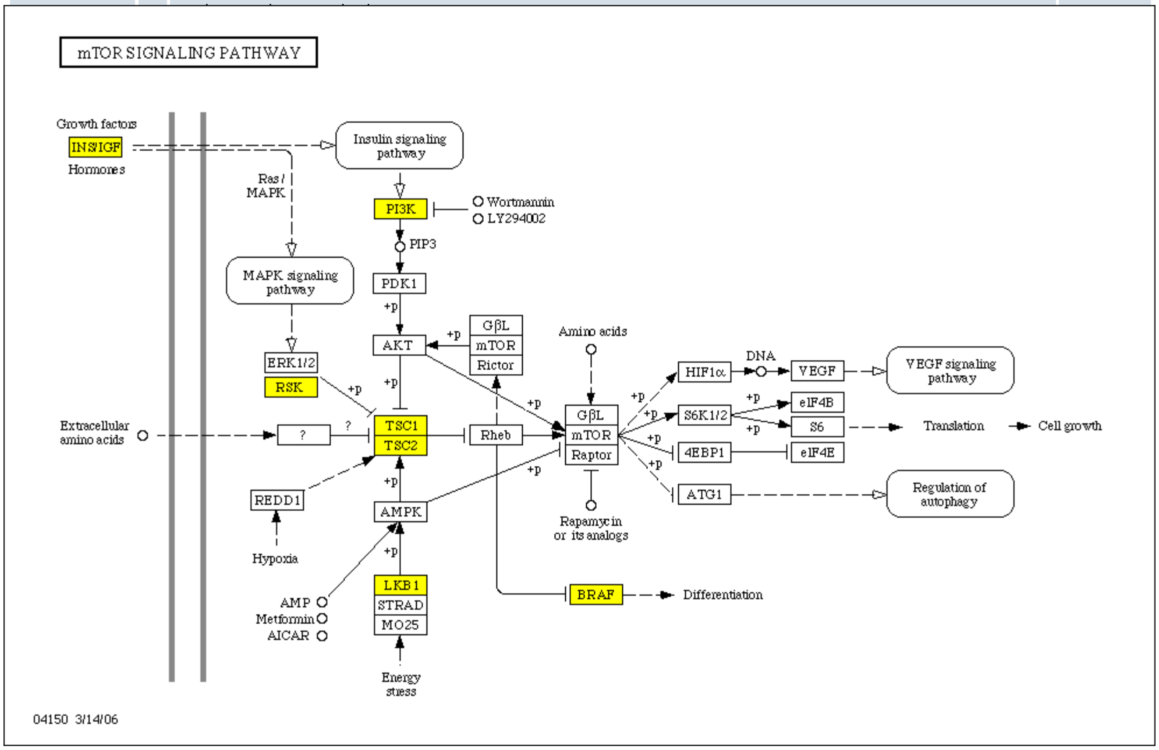
🟨 Genes having annotated mutations



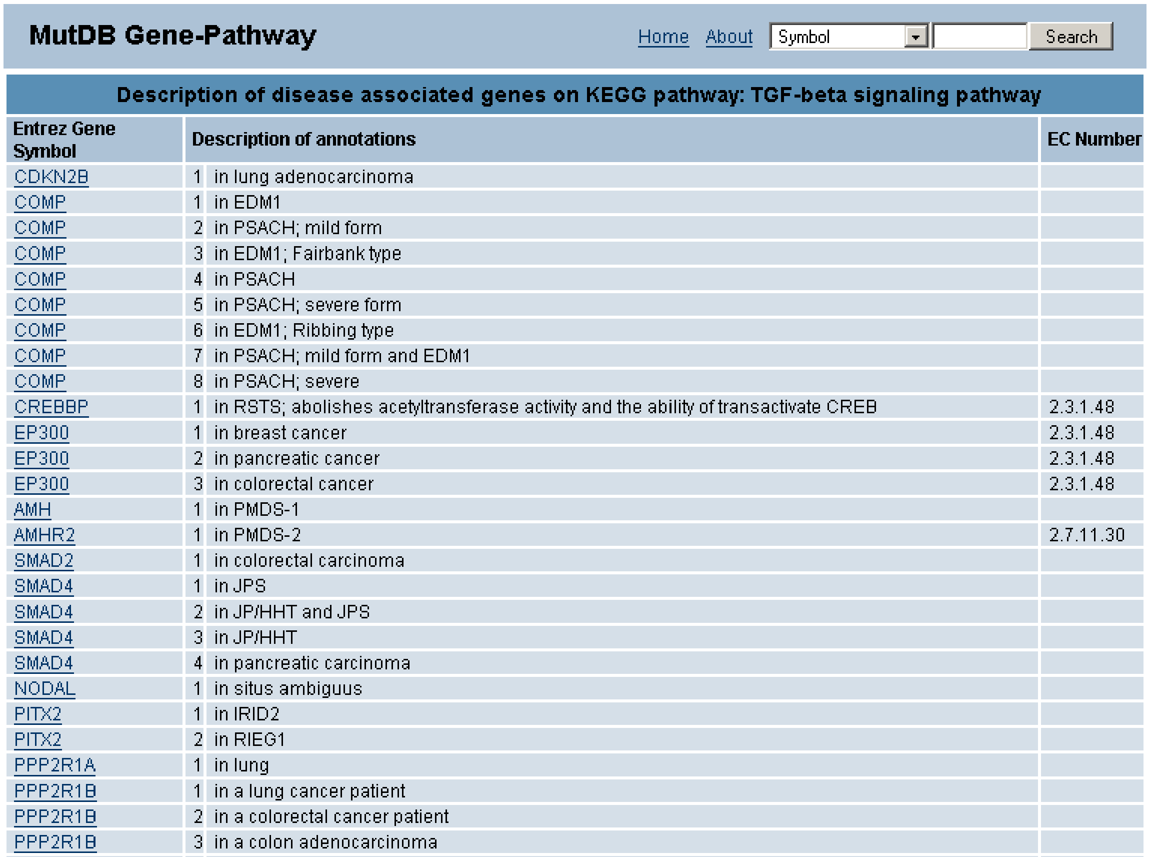
MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: mTOR signaling pathway

Entrez Gene Symbol	Description of annotations		EC Number
INS	1	in familial hyperproinsulinemia; Providence	
INS	2	associated with diabetes mellitus type-II; Los-Angeles	
INS	3	in Chicago	
INS	4	in familial hyperproinsulinemia; impairs posttranslational cleavage	
INS	5	in familial hyperproinsulinemia; Kyoto	
INS	6	in Wakayama	
PIK3CA	1	in cancer; shows an increase in lipid kinase activity	2.7.1.153
PIK3CA	2	in cancer	2.7.1.153
PIK3CA	3	in cancer; shows an increase in lipid kinase activity; oncogenic in vivo	2.7.1.153
PIK3R1	1	in severe insulin resistance; reduction of insulin-stimulated activity	
RPS6KA3	1	in CLS	2.7.11.1
BRAF	1	in CFC syndrome	2.7.11.1
BRAF	2	in colorectal cancer	2.7.11.1
BRAF	3	in a colorectal cancer cell line; elevated kinase activity; efficiently induces cell transformation	2.7.11.1
BRAF	4	in melanoma	2.7.11.1
BRAF	5	in lung cancer	2.7.11.1
BRAF	6	in NHL; also in a lung cancer cell line; elevated kinase activity; efficiently induces cell transformation	2.7.11.1
BRAF	7	in CFC syndrome and colon cancer	2.7.11.1
BRAF	8	in NHL	2.7.11.1
BRAF	9	in ovarian cancer	2.7.11.1
BRAF	10	in colon cancer	2.7.11.1
BRAF	11	in a colon cancer cell line	2.7.11.1
BRAF	12	in lung cancer and ovarian cancer	2.7.11.1
BRAF	13	in a lung cancer cell line; elevated kinase activity; efficiently induces cell transformation	2.7.11.1
BRAF	14	in a melanoma cell line; requires 2 nucleotide substitutions	2.7.11.1
BRAF	15	in colorectal cancer, melanoma, ovarian cancer and sarcoma; most common mutation; elevated kinase activity; efficiently induces cell transformation; suppression of mutation in melanoma causes growth arrest	2.7.11.1



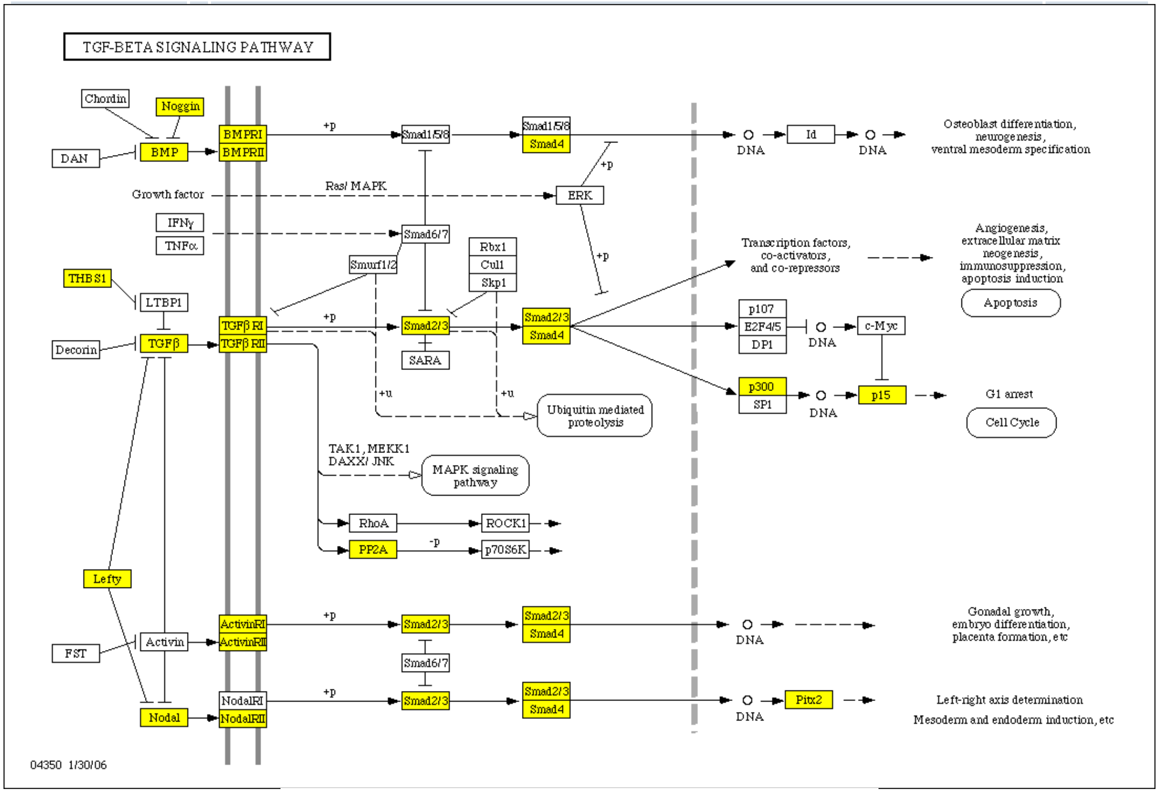
🟨 Genes having annotated mutations



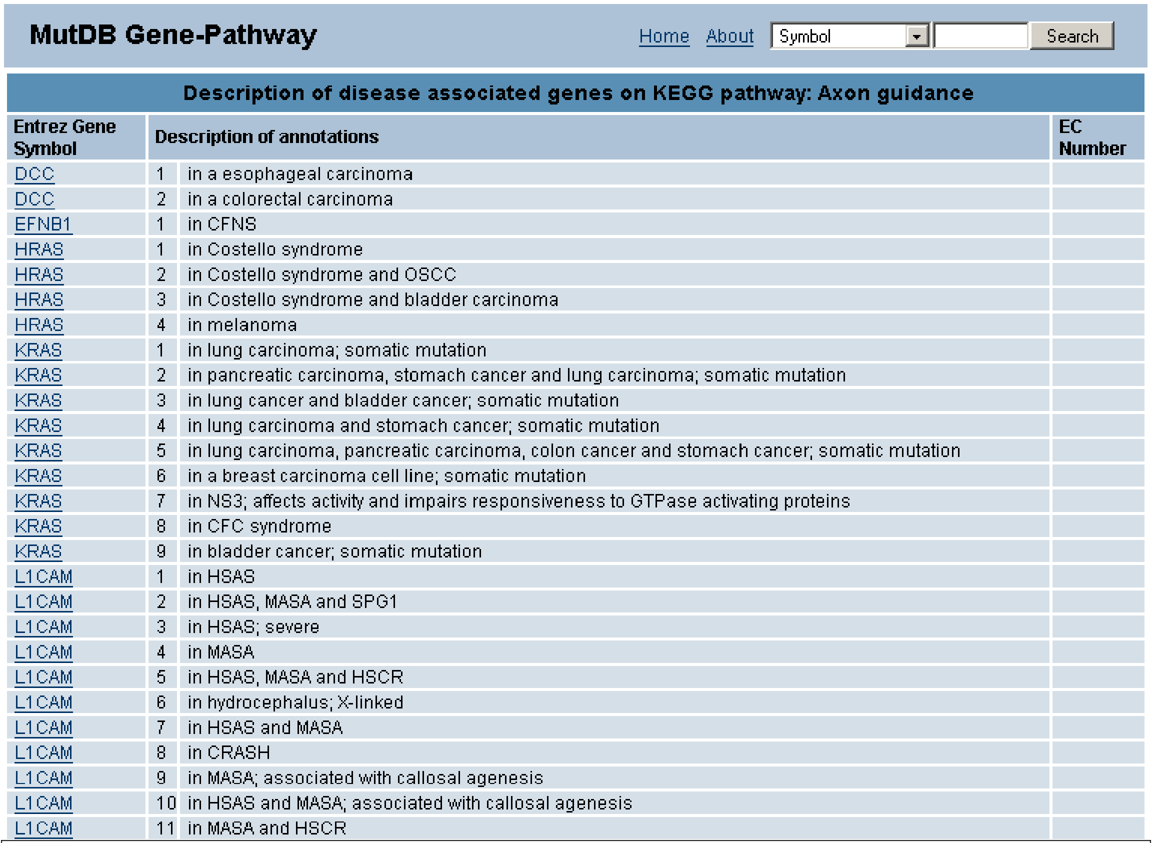
MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: TGF-beta signaling pathway

Entrez Gene Symbol	Description of annotations		EC Number
CDKN2B	1	in lung adenocarcinoma	
COMP	1	in EDM1	
COMP	2	in PSACH; mild form	
COMP	3	in EDM1; Fairbank type	
COMP	4	in PSACH	
COMP	5	in PSACH; severe form	
COMP	6	in EDM1; Ribbing type	
COMP	7	in PSACH; mild form and EDM1	
COMP	8	in PSACH; severe	
CREBBP	1	in RSTS; abolishes acetyltransferase activity and the ability of transactivate CREB	2.3.1.48
EP300	1	in breast cancer	2.3.1.48
EP300	2	in pancreatic cancer	2.3.1.48
EP300	3	in colorectal cancer	2.3.1.48
AMH	1	in PMDS-1	
AMHR2	1	in PMDS-2	2.7.11.30
SMAD2	1	in colorectal carcinoma	
SMAD4	1	in JPS	
SMAD4	2	in JP/HHT and JPS	
SMAD4	3	in JP/HHT	
SMAD4	4	in pancreatic carcinoma	
NODAL	1	in situs ambiguus	
PITX2	1	in IRID2	
PITX2	2	in RIEG1	
PPP2R1A	1	in lung	
PPP2R1B	1	in a lung cancer patient	
PPP2R1B	2	in a colorectal cancer patient	
PPP2R1B	3	in a colon adenocarcinoma	



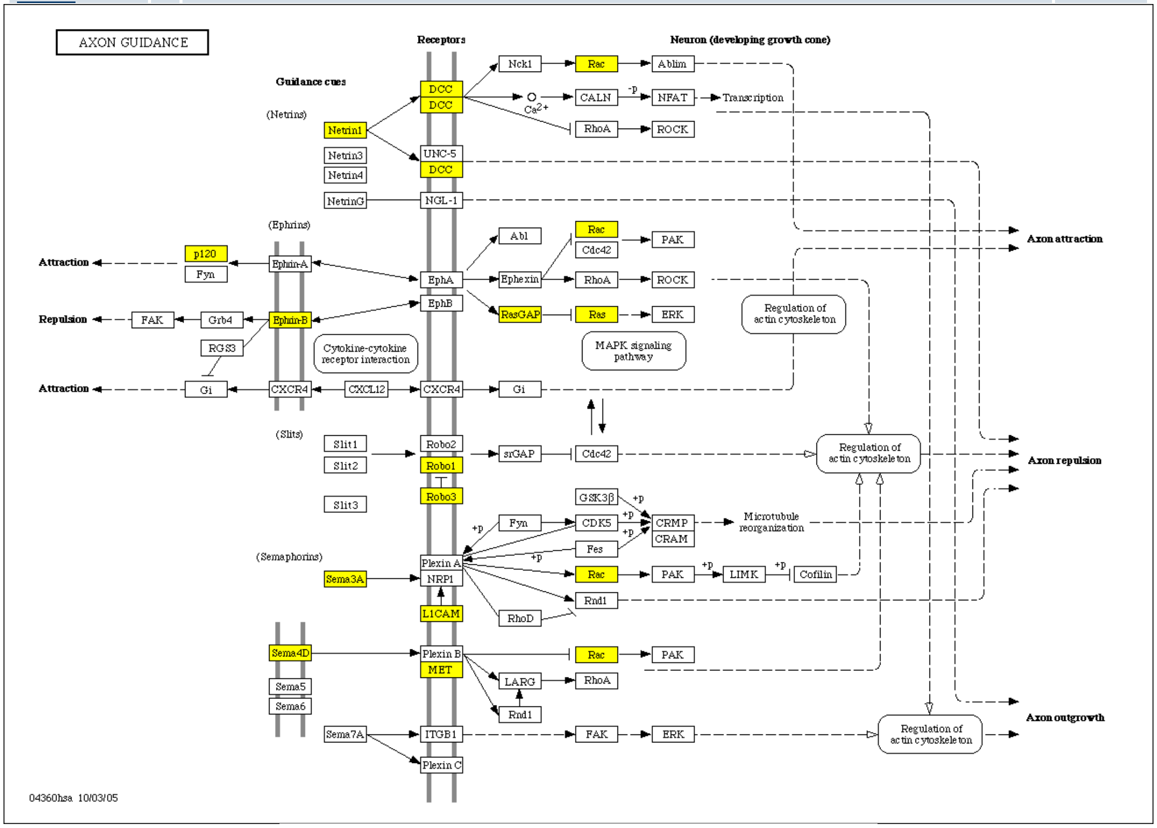
🟨 Genes having annotated mutations



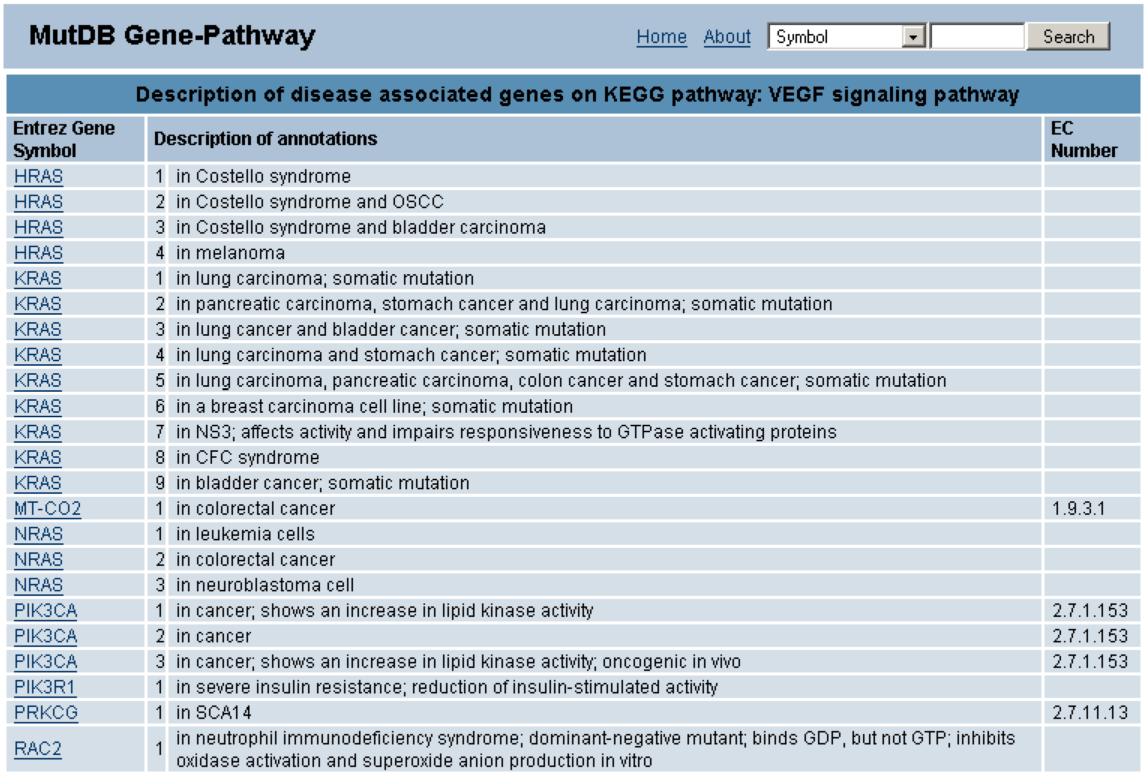
MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: Axon guidance

Entrez Gene Symbol	Description of annotations		EC Number
DCC	1	in a esophageal carcinoma	
DCC	2	in a colorectal carcinoma	
EFNB1	1	in CFNS	
HRAS	1	in Costello syndrome	
HRAS	2	in Costello syndrome and OSCC	
HRAS	3	in Costello syndrome and bladder carcinoma	
HRAS	4	in melanoma	
KRAS	1	in lung carcinoma; somatic mutation	
KRAS	2	in pancreatic carcinoma, stomach cancer and lung carcinoma; somatic mutation	
KRAS	3	in lung cancer and bladder cancer; somatic mutation	
KRAS	4	in lung carcinoma and stomach cancer; somatic mutation	
KRAS	5	in lung carcinoma, pancreatic carcinoma, colon cancer and stomach cancer; somatic mutation	
KRAS	6	in a breast carcinoma cell line; somatic mutation	
KRAS	7	in NS3; affects activity and impairs responsiveness to GTPase activating proteins	
KRAS	8	in CFC syndrome	
KRAS	9	in bladder cancer; somatic mutation	
L1CAM	1	in HSAS	
L1CAM	2	in HSAS, MASA and SPG1	
L1CAM	3	in HSAS; severe	
L1CAM	4	in MASA	
L1CAM	5	in HSAS, MASA and HSCR	
L1CAM	6	in hydrocephalus; X-linked	
L1CAM	7	in HSAS and MASA	
L1CAM	8	in CRASH	
L1CAM	9	in MASA; associated with callosal agenesis	
L1CAM	10	in HSAS and MASA; associated with callosal agenesis	
L1CAM	11	in MASA and HSCR	



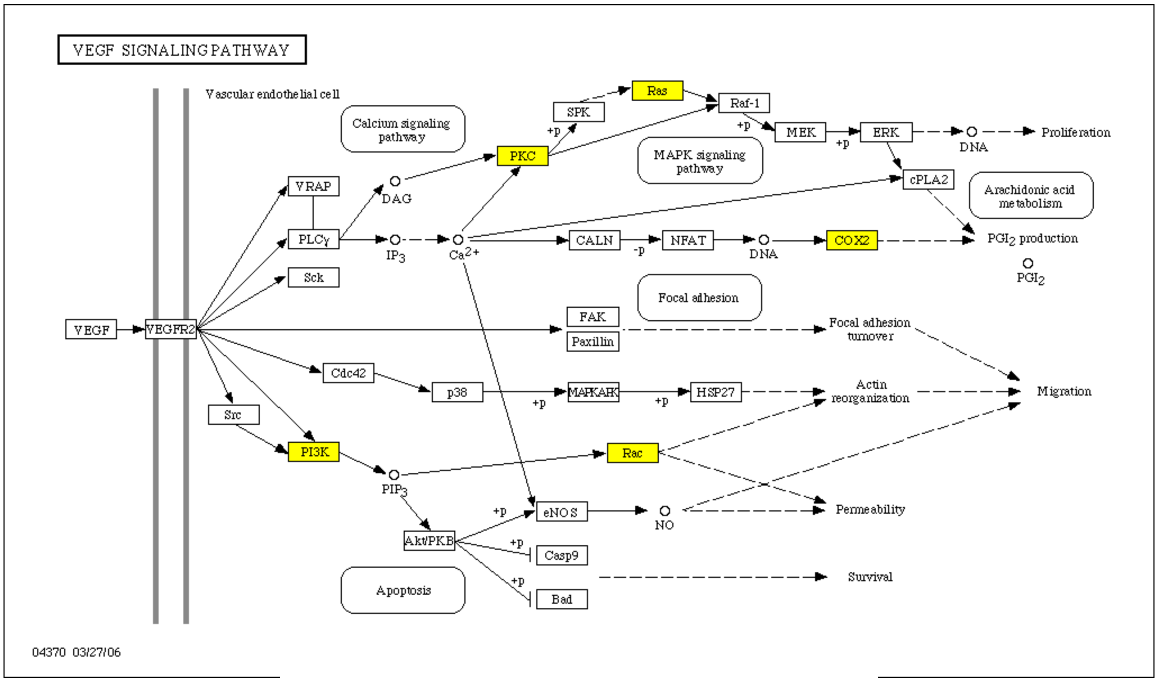
🟨 Genes having annotated mutations



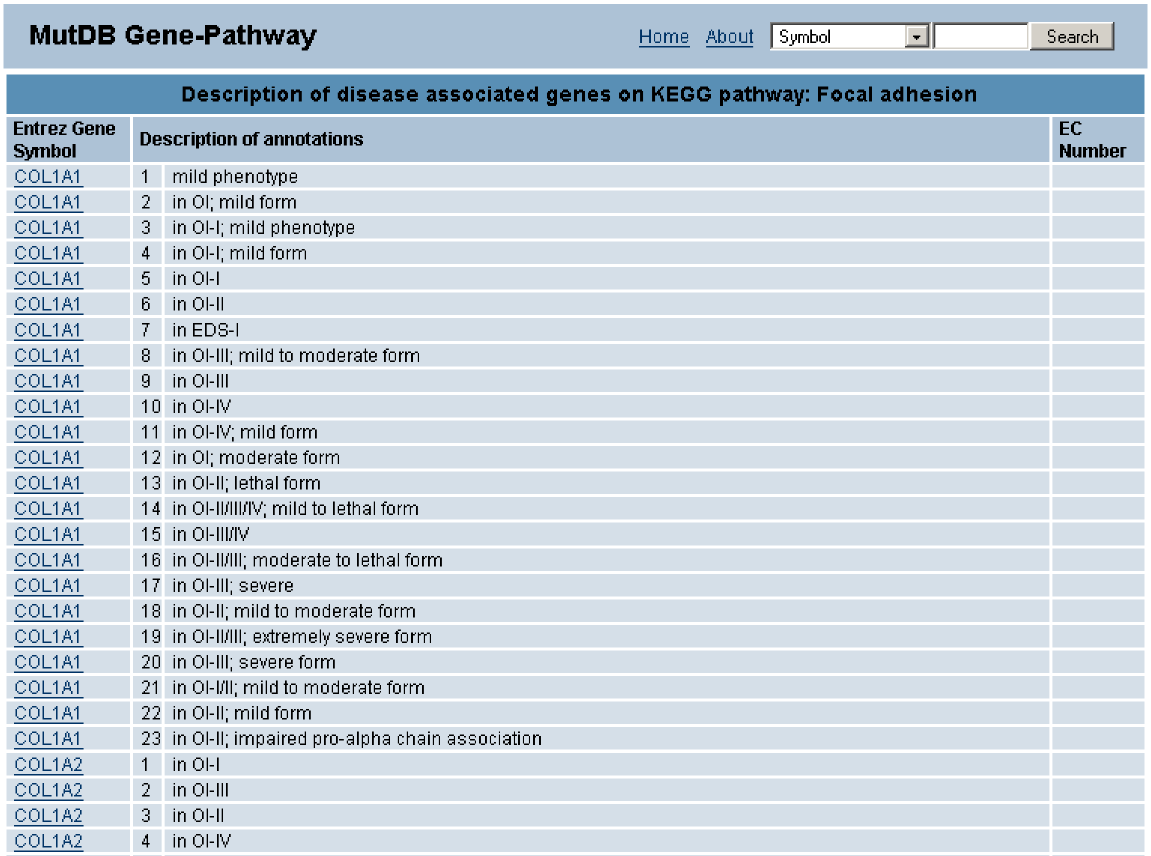
MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: VEGF signaling pathway

Entrez Gene Symbol	Description of annotations		EC Number
HRAS	1	in Costello syndrome	
HRAS	2	in Costello syndrome and OSCC	
HRAS	3	in Costello syndrome and bladder carcinoma	
HRAS	4	in melanoma	
KRAS	1	in lung carcinoma; somatic mutation	
KRAS	2	in pancreatic carcinoma, stomach cancer and lung carcinoma; somatic mutation	
KRAS	3	in lung cancer and bladder cancer; somatic mutation	
KRAS	4	in lung carcinoma and stomach cancer; somatic mutation	
KRAS	5	in lung carcinoma, pancreatic carcinoma, colon cancer and stomach cancer; somatic mutation	
KRAS	6	in a breast carcinoma cell line; somatic mutation	
KRAS	7	in NS3; affects activity and impairs responsiveness to GTPase activating proteins	
KRAS	8	in CFC syndrome	
KRAS	9	in bladder cancer; somatic mutation	
MT-CO2	1	in colorectal cancer	1.9.3.1
NRAS	1	in leukemia cells	
NRAS	2	in colorectal cancer	
NRAS	3	in neuroblastoma cell	
PIK3CA	1	in cancer; shows an increase in lipid kinase activity	2.7.1.153
PIK3CA	2	in cancer	2.7.1.153
PIK3CA	3	in cancer; shows an increase in lipid kinase activity; oncogenic in vivo	2.7.1.153
PIK3R1	1	in severe insulin resistance; reduction of insulin-stimulated activity	
PRKCG	1	in SCA14	2.7.11.13
RAC2	1	in neutrophil immunodeficiency syndrome; dominant-negative mutant; binds GDP, but not GTP; inhibits oxidase activation and superoxide anion production in vitro	



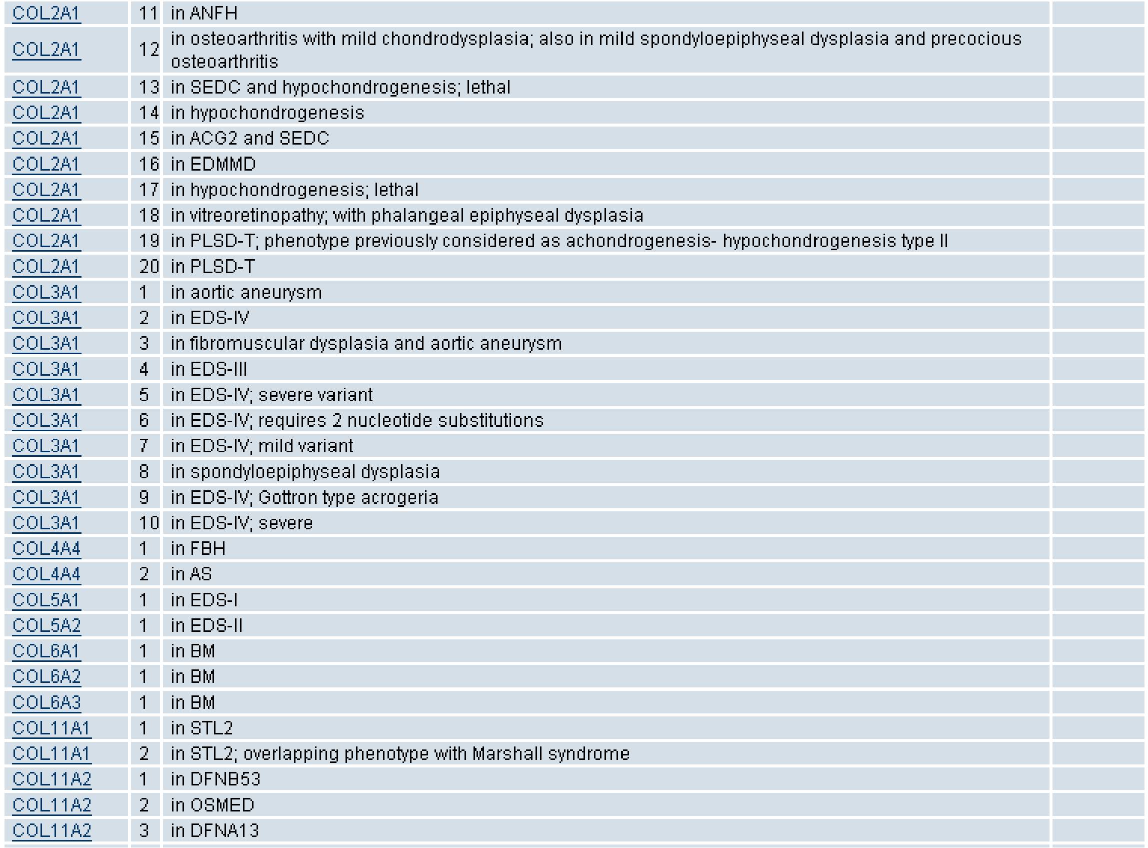
🟨 Genes having annotated mutations



MutDB Gene-Pathway

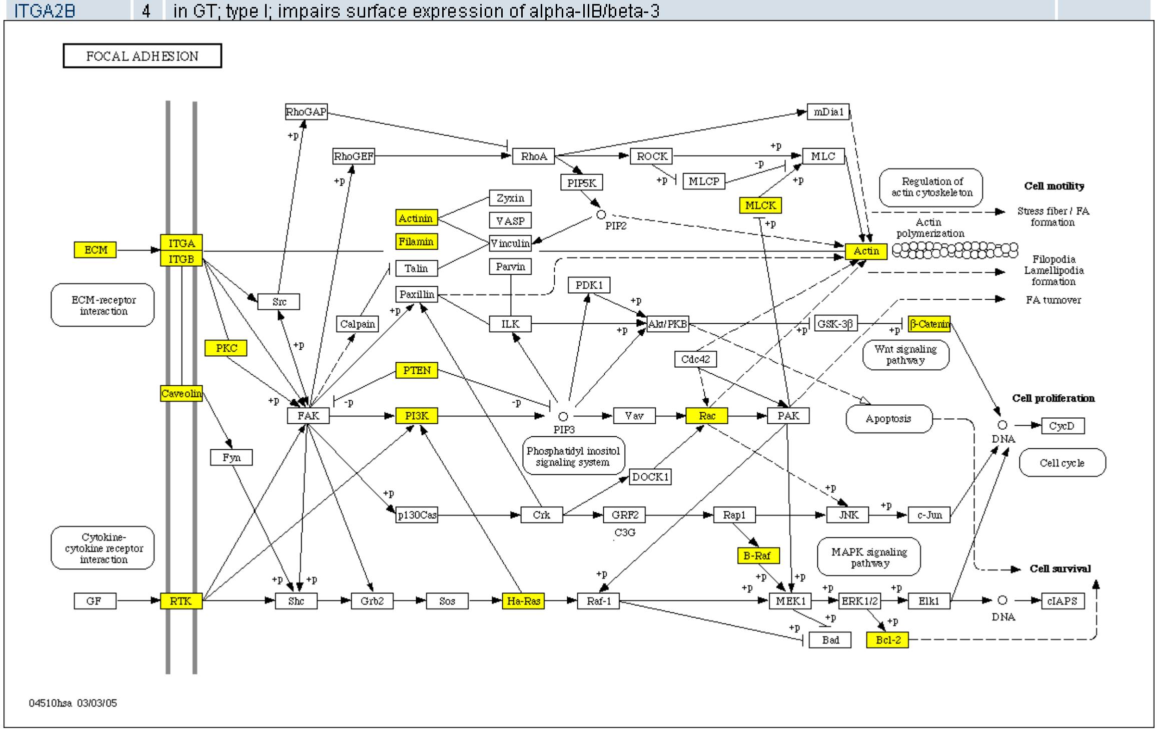
Description of disease associated genes on KEGG pathway: Focal adhesion

Entrez Gene Symbol	Description of annotations		EC Number
COL1A1	1	mild phenotype	
COL1A1	2	in OI; mild form	
COL1A1	3	in OI-I; mild phenotype	
COL1A1	4	in OI-I; mild form	
COL1A1	5	in OI-I	
COL1A1	6	in OI-II	
COL1A1	7	in EDS-I	
COL1A1	8	in OI-III; mild to moderate form	
COL1A1	9	in OI-III	
COL1A1	10	in OI-IV	
COL1A1	11	in OI-IV; mild form	
COL1A1	12	in OI; moderate form	
COL1A1	13	in OI-II; lethal form	
COL1A1	14	in OI-II/III/IV; mild to lethal form	
COL1A1	15	in OI-III/IV	
COL1A1	16	in OI-II/III; moderate to lethal form	
COL1A1	17	in OI-III; severe	
COL1A1	18	in OI-II; mild to moderate form	
COL1A1	19	in OI-II/III; extremely severe form	
COL1A1	20	in OI-III; severe form	
COL1A1	21	in OI-I/II; mild to moderate form	
COL1A1	22	in OI-II; mild form	
COL1A1	23	in OI-II; impaired pro-alpha chain association	
COL1A2	1	in OI-I	
COL1A2	2	in OI-III	
COL1A2	3	in OI-II	
COL1A2	4	in OI-IV	

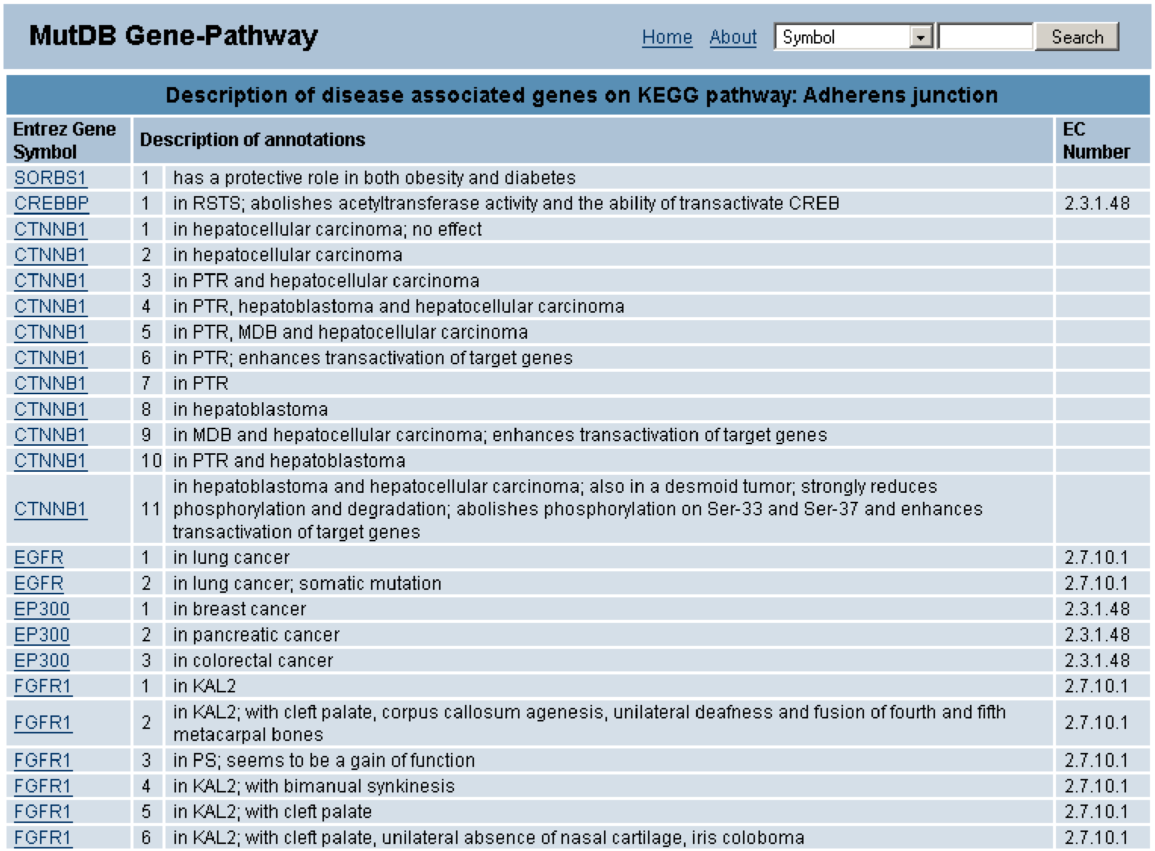


COL2A1	11	in ANFH	
COL2A1	12	in osteoarthritis with mild chondrodysplasia; also in mild spondyloepiphyseal dysplasia and precocious osteoarthritis	
COL2A1	13	in SEDC and hypochondrogenesis; lethal	
COL2A1	14	in hypochondrogenesis	
COL2A1	15	in ACG2 and SEDC	
COL2A1	16	in EDMMD	
COL2A1	17	in hypochondrogenesis; lethal	
COL2A1	18	in vitreoretinopathy; with phalangeal epiphyseal dysplasia	
COL2A1	19	in PLSD-T; phenotype previously considered as achondrogenesis- hypochondrogenesis type II	
COL2A1	20	in PLSD-T	
COL3A1	1	in aortic aneurysm	
COL3A1	2	in EDS-IV	
COL3A1	3	in fibromuscular dysplasia and aortic aneurysm	
COL3A1	4	in EDS-III	
COL3A1	5	in EDS-IV; severe variant	
COL3A1	6	in EDS-IV; requires 2 nucleotide substitutions	
COL3A1	7	in EDS-IV; mild variant	
COL3A1	8	in spondyloepiphyseal dysplasia	
COL3A1	9	in EDS-IV; Gottron type acrogeria	
COL3A1	10	in EDS-IV; severe	
COL4A4	1	in FBH	
COL4A4	2	in AS	
COL5A1	1	in EDS-I	
COL5A2	1	in EDS-II	
COL6A1	1	in BM	
COL6A2	1	in BM	
COL6A3	1	in BM	
COL11A1	1	in STL2	
COL11A1	2	in STL2; overlapping phenotype with Marshall syndrome	
COL11A2	1	in DFNB53	
COL11A2	2	in OSMED	
COL11A2	3	in DFNA13	

COMP	4	in PSACH	
COMP	5	in PSACH; severe form	
COMP	6	in EDM1; Ribbing type	
COMP	7	in PSACH; mild form and EDM1	
COMP	8	in PSACH; severe	
CTNNB1	1	in hepatocellular carcinoma; no effect	
CTNNB1	2	in hepatocellular carcinoma	
CTNNB1	3	in PTR and hepatocellular carcinoma	
CTNNB1	4	in PTR, hepatoblastoma and hepatocellular carcinoma	
CTNNB1	5	in PTR, MDB and hepatocellular carcinoma	
CTNNB1	6	in PTR; enhances transactivation of target genes	
CTNNB1	7	in PTR	
CTNNB1	8	in hepatoblastoma	
CTNNB1	9	in MDB and hepatocellular carcinoma; enhances transactivation of target genes	
CTNNB1	10	in PTR and hepatoblastoma	
CTNNB1	11	in hepatoblastoma and hepatocellular carcinoma; also in a desmoid tumor; strongly reduces phosphorylation and degradation; abolishes phosphorylation on Ser-33 and Ser-37 and enhances transactivation of target genes	
EGFR	1	in lung cancer	2.7.10.1
EGFR	2	in lung cancer; somatic mutation	2.7.10.1
FLNA	1	in PVNH4	
FLNA	2	in PVNH1	
FLNA	3	in OPD2	
FLNA	4	in OPD1	
FLNA	5	in FMD	
FLNA	6	in MNS	
HRAS	1	in Costello syndrome	
HRAS	2	in Costello syndrome and OSCC	
HRAS	3	in Costello syndrome and bladder carcinoma	
HRAS	4	in melanoma	
ITGA2B	1	in GT; impairs surface expression of alpha-IIB/beta-3 and abrogates ligand binding to the activated integrin	
ITGA2B	2	in GT; impairs surface expression of alpha-IIB/beta-3	
ITGA2B	3	in GT; alters the heterodimer conformation thus impairing their intracellular transport	
ITGA2B	4	in GT; type I; impairs surface expression of alpha-IIB/beta-3	



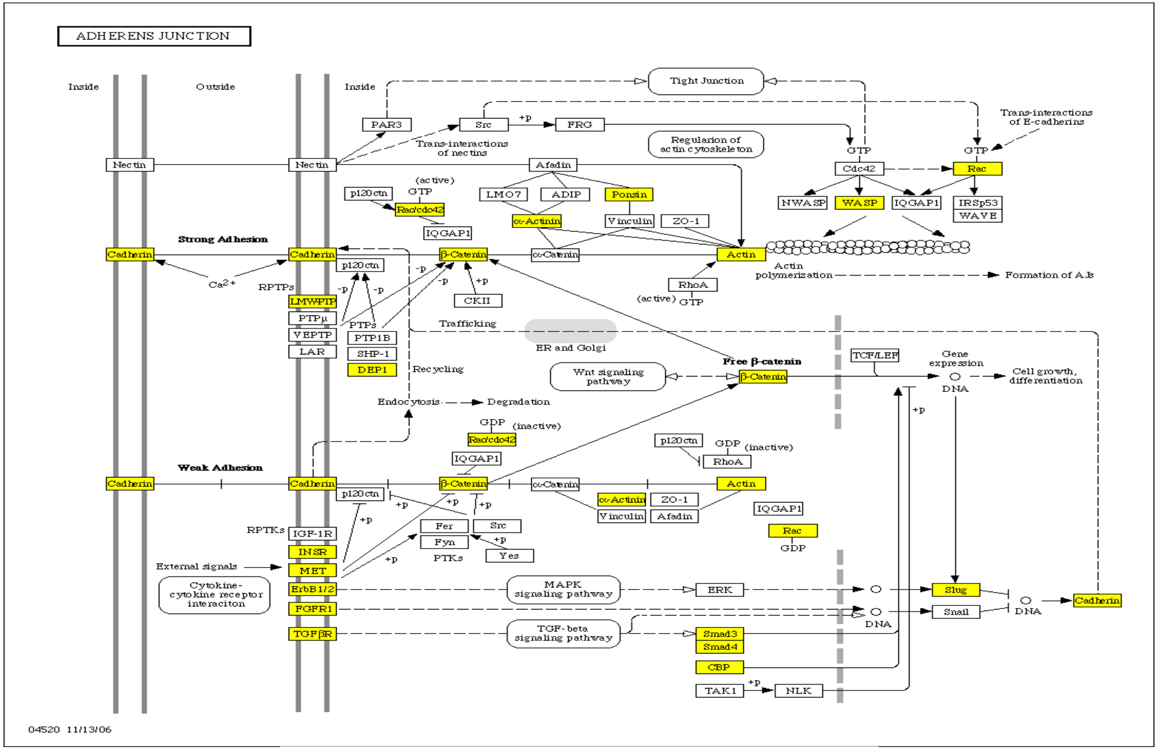
🟨 Genes having annotated mutations



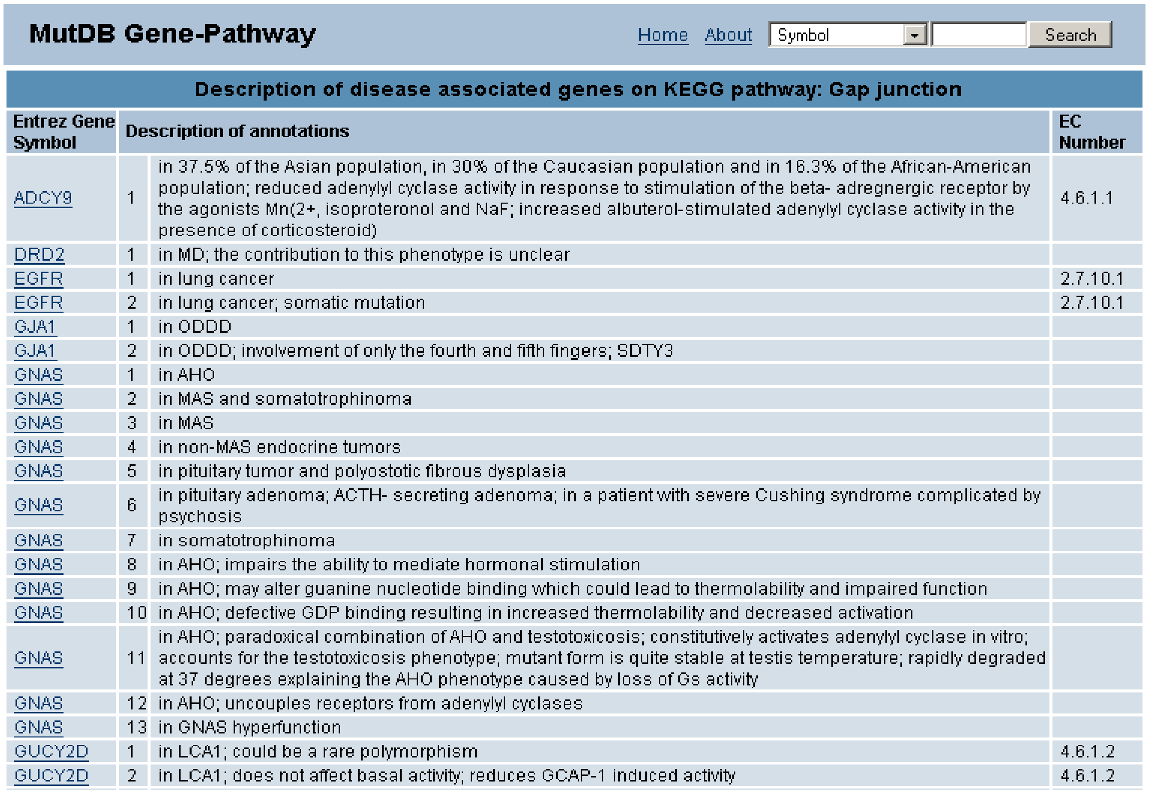
MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: Adherens junction

Entrez Gene Symbol	Description of annotations		EC Number
SORBS1	1	has a protective role in both obesity and diabetes	
CREBBP	1	in RSTS; abolishes acetyltransferase activity and the ability of transactivate CREB	2.3.1.48
CTNNB1	1	in hepatocellular carcinoma; no effect	
CTNNB1	2	in hepatocellular carcinoma	
CTNNB1	3	in PTR and hepatocellular carcinoma	
CTNNB1	4	in PTR, hepatoblastoma and hepatocellular carcinoma	
CTNNB1	5	in PTR, MDB and hepatocellular carcinoma	
CTNNB1	6	in PTR; enhances transactivation of target genes	
CTNNB1	7	in PTR	
CTNNB1	8	in hepatoblastoma	
CTNNB1	9	in MDB and hepatocellular carcinoma; enhances transactivation of target genes	
CTNNB1	10	in PTR and hepatoblastoma	
CTNNB1	11	in hepatoblastoma and hepatocellular carcinoma; also in a desmoid tumor; strongly reduces phosphorylation and degradation; abolishes phosphorylation on Ser-33 and Ser-37 and enhances transactivation of target genes	
EGFR	1	in lung cancer	2.7.10.1
EGFR	2	in lung cancer; somatic mutation	2.7.10.1
EP300	1	in breast cancer	2.3.1.48
EP300	2	in pancreatic cancer	2.3.1.48
EP300	3	in colorectal cancer	2.3.1.48
FGFR1	1	in KAL2	2.7.10.1
FGFR1	2	in KAL2; with cleft palate, corpus callosum agenesis, unilateral deafness and fusion of fourth and fifth metacarpal bones	2.7.10.1
FGFR1	3	in PS; seems to be a gain of function	2.7.10.1
FGFR1	4	in KAL2; with bimanual synkinesis	2.7.10.1
FGFR1	5	in KAL2; with cleft palate	2.7.10.1
FGFR1	6	in KAL2; with cleft palate, unilateral absence of nasal cartilage, iris coloboma	2.7.10.1



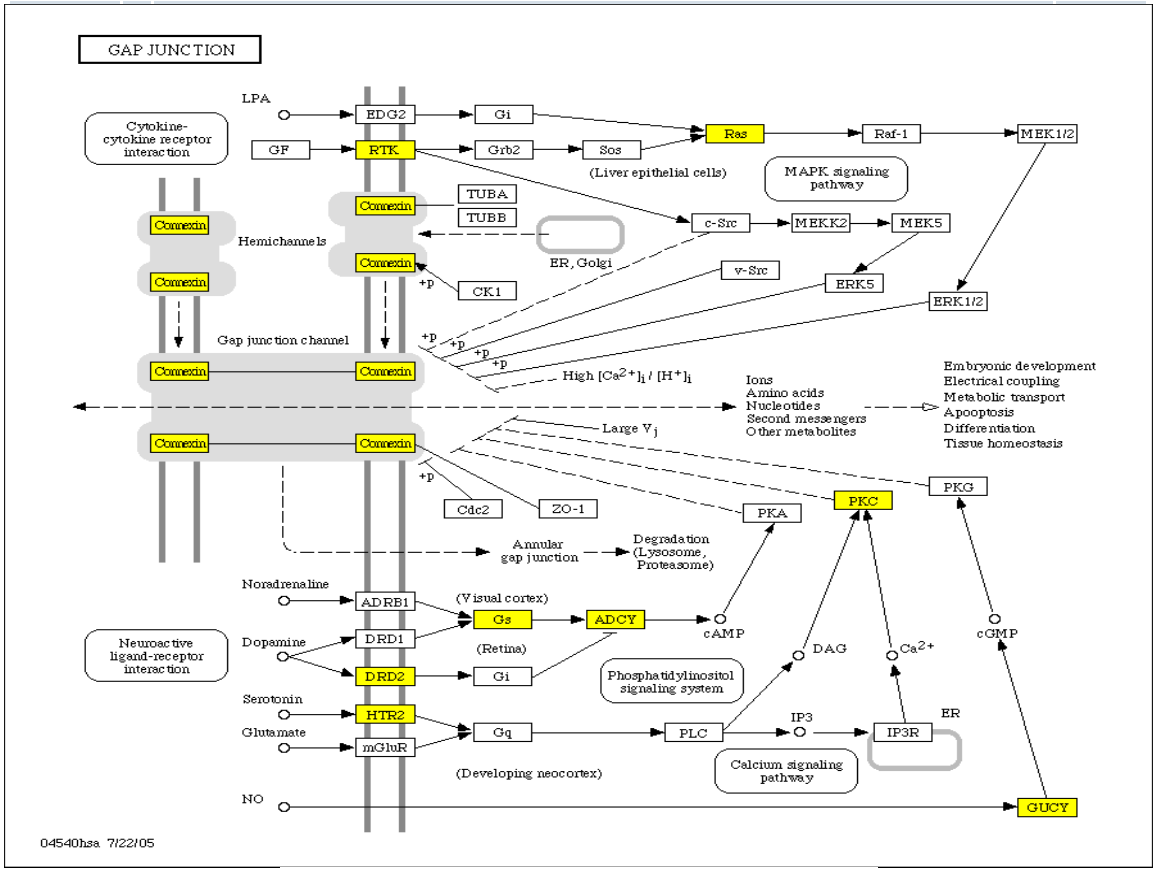
🟨 Genes having annotated mutations



MutDB Gene-Pathway

Description of disease associated genes on KEGG pathway: Gap junction

Entrez Gene Symbol		Description of annotations	EC Number
ADCY9	1	in 37.5% of the Asian population, in 30% of the Caucasian population and in 16.3% of the African-American population; reduced adenylyl cyclase activity in response to stimulation of the beta- adregnergic receptor by the agonists Mn(2+, isoproteronol and NaF; increased albuterol-stimulated adenylyl cyclase activity in the presence of corticosteroid)	4.6.1.1
DRD2	1	in MD; the contribution to this phenotype is unclear	
EGFR	1	in lung cancer	2.7.10.1
EGFR	2	in lung cancer; somatic mutation	2.7.10.1
GJA1	1	in ODDD	
GJA1	2	in ODDD; involvement of only the fourth and fifth fingers; SDTY3	
GNAS	1	in AHO	
GNAS	2	in MAS and somatotrophinoma	
GNAS	3	in MAS	
GNAS	4	in non-MAS endocrine tumors	
GNAS	5	in pituitary tumor and polyostotic fibrous dysplasia	
GNAS	6	in pituitary adenoma; ACTH- secreting adenoma; in a patient with severe Cushing syndrome complicated by psychosis	
GNAS	7	in somatotrophinoma	
GNAS	8	in AHO; impairs the ability to mediate hormonal stimulation	
GNAS	9	in AHO; may alter guanine nucleotide binding which could lead to thermolability and impaired function	
GNAS	10	in AHO; defective GDP binding resulting in increased thermolability and decreased activation	
GNAS	11	in AHO; paradoxical combination of AHO and testotoxicosis; constitutively activates adenylyl cyclase in vitro; accounts for the testotoxicosis phenotype; mutant form is quite stable at testis temperature; rapidly degraded at 37 degrees explaining the AHO phenotype caused by loss of Gs activity	
GNAS	12	in AHO; uncouples receptors from adenylyl cyclases	
GNAS	13	in GNAS hyperfunction	
GUCY2D	1	in LCA1; could be a rare polymorphism	4.6.1.2
GUCY2D	2	in LCA1; does not affect basal activity; reduces GCAP-1 induced activity	4.6.1.2



🟨 Genes having annotated mutations

APPENDIX B

CURRICULUM VITAE

Arti Singh
410 W 10th Street Ste 5000
Indianapolis, IN 46202
Cell: 504-339-0827
Email: arsingh@iupui.edu, reach2arti@gmail.com

OBJECTIVE

Exploring and working in intriguing areas of bioinformatics and computational biology.

WORK EXPERIENCE

- **Research Assistant, Center for Computational Biology and Bioinformatics**, **School of Medicine, Indiana University**, Indianapolis: Developing software and working on genetic mutation and biological pathways databases (Sept 2006-present).

- **Summer Intern, Dow Agrochemicals** (Parent Company: Dow Chemicals), Indianapolis: Development of Bioinformatics related software application and database (May 2006 to Aug 2006).

- **Software Developer**, **Systems Department, Indian Oil Corporation Ltd (IOCL)**, New Delhi - A Fortune 500 Company: Analyzed, designed, developed and implemented software systems for IOCL. Worked on Oracle, ASP.Net (Dec 2003 to Nov 2004).

EDUCATION

- **MS Bioinformatics**, GPA: 3.92/4.00, School of Informatics, Indiana University, IN, Aug 2005- Aug 2007

- **Bachelor of Information Technology**, IGNOU, New Delhi, India, Aug 2000-Aug 2003

SELECTED PROJECTS

- **Summer Project at Dow Agrochemicals:** Developed an automated workflow for Expressed Sequence Tag (EST) clustering and assembly. Used PipeLine Pilot, Perl and mySQL to design an automated workflow.
- **Web Application Development:** Designed and developed integrated IT solutions for the company for Safety and Environment Protection Division of IOCL. Used ASP.Net and Oracle.

INFORMATION TECHNOLOGY CERTIFICATIONS

- **RHCE** (Completed RedHat Certified Linux Course), Taught by RedHat Faculty, New Delhi, India February 2005.
- **GNIIT,** Graduate *of* National Institute of Information Technology (NIIT), New Delhi, India, Aug 2000-Nov 2003

COMPUTER AND TECHNICAL SKILLS

- **Computer Language**: Perl , VB.Net , C#.Net , Java (Core), XML, Matlab
- **Operating System**: Windows, Red Hat Linux
- **DBMS**: MS SQL Server 2000 , Oracle 8i , mySQL, Aquadata
- **Others**: ASP.Net , web services on .Net Platform , PipeLine Pilot , .Net Components COM+ , XML spy
- **Bioinformatics Software**: Phrap, Cross_match for sequence analysis

COURSEWORK

- Object Oriented Programming, Undergraduate
- Databases, Undergraduate and Graduate
- Software Engineering, Undergraduate
- Data Structures & Algorithms, Undergraduate
- Data Analysis & Database Design, Undergraduate
- Quality Management Principles, Undergraduate
- Introduction to Bioinformatics (Completed a project on Microarray Data Clustering), Graduate
- Bioinformatics Tools and Techniques, Graduate (Worked on a project studying relationship between protein network hubs, drug targets, essential genes and disease specific genes)
- Principles of Molecular Biology, Graduate
- Project Management, Graduate
- Scientific Applications of XML, Graduate
- Structural Bioinformatics, Graduate

PUBLICATIONS AND CONFERENCES

- *Singh Arti*, Olowoyeye Adebayo, Baenziger Peter H., Dantzer Jessica, Kann Maricel G., Radivojac Predrag, Heiland Randy, Mooney Sean D., **MutDB: Update on development of tools for the biochemical analysis of genetic variation**, Nucleic Acids Research, Database Issue, Submitted.
- Amy Schmidbauer, Arvind Kumar, *Arti Singh*, Mallika Mahoui, Ignacio Larrinua, Neil Kirby **Automation of Large–Scale EST Assembly & Annotation Using Pipeline Pilot®**, Poster Session, 4th Indiana Bioinformatics Conference, Indianapolis, IN, May 31-June 2, 2007.
- *Singh Arti*, Olowoyeye Adebayo, Baenziger Peter, Dantzer Jessica, Kann G Maricel, Radivojac Predrag, Mooney D Sean, **MutDB: Comprehensive tools for functional analysis of genetic variation**, Poster Session, Pacific Symposium on Biocomputing, Maui, Hawaii, January 3-7, 2007.
- *Singh Arti* and Chen Y Jake., **Integrated Analysis of Essential Genes and Network Hubs as Potential Druggable Targets**, Poster Session, 3rd Indiana Bioinformatics Conference, Indianapolis, IN, May 20-21, 2006.

PROFESSIONAL ASSOCIATIONS

- Student Member, The Institute of Electrical and Electronics Engineers (IEEE)
- Student Member, Association for Computing Machinery (ACM)