The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses

Gil Stelzer,[1,5] Naomi Rosen,[1,5] Inbar Plaschkes,[1,2] Shahar Zimmerman,[1] Michal Twik,[1] Simon Fishilevich,[1] Tsippi Iny Stein,[1] Ron Nudel,[1] Iris Lieder,[2] Yaron Mazor,[2] Sergey Kaplan,[2] Dvir Dahary,[2,4] David Warshawsky,[3] Yaron Guan-Golan,[3] Asher Kohn,[3] Noa Rappaport,[1] Marilyn Safran,[1] and Doron Lancet[1,6]

[1] Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel
[2] LifeMap Sciences Ltd., Tel Aviv, Israel
[3] LifeMap Sciences Inc., Marshfield, Massachusetts
[4] Toldot Genetics Ltd., Hod Hasharon, Israel
[5] These authors contributed equally to the paper
[6] Corresponding author

GeneCards, the human gene compendium, enables researchers to effectively navigate and inter-relate the wide universe of human genes, diseases, variants, proteins, cells, and biological pathways. Our recently launched Version 4 has a revamped infrastructure facilitating faster data updates, better-targeted data queries, and friendlier user experience. It also provides a stronger foundation for the GeneCards suite of companion databases and analysis tools. Improved data unification includes gene-disease links via MalaCards and merged biological pathways via PathCards, as well as drug information and proteome expression. VarElect, another suite member, is a phenotype prioritizer for next-generation sequencing, leveraging the GeneCards and MalaCards knowledgebase. It automatically infers direct and indirect scored associations between hundreds or even thousands of variant-containing genes and disease phenotype terms. VarElect's capabilities, either independently or within TGex, our comprehensive variant analysis pipeline, help prepare for the challenge of clinical projects that involve thousands of exome/genome NGS analyses. © 2016 by John Wiley & Sons, Inc.

Keywords: biological database • bioinformatics • diseases • GeneCards • gene prioritization • human genes • integrated information retrieval • next generation sequencing • VarElect

How to cite this article:
Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., and Lancet, D. 2016. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* 54:1.30.1-1.30.33.
doi: 10.1002/cpbi.5

INTRODUCTION

GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. It automatically integrates gene-centric data from ~125 sources, including genomic, transcriptomic,

Current Protocols in Bioinformatics 1.30.1-1.30.33, June 2016

doi: 10.1002/cpbi.5

proteomic, genetic, clinical, and functional information. GeneCards unifies genes with different names and from different sources based on their locations and on cross-referential annotation. In addition to GeneCards, the GeneCards Suite knowledgebase includes MalaCards (*UNIT 1.24*; Rappaport et al., 2014), the human disease database, and LifeMap Discovery, the cells and tissues database (Edgar et al., 2013). This knowledgebase empowers the suite's premium tools, which enable rapid biological interpretation of next-generation sequencing (NGS) data, including gene variant prioritization and RNAseq-derived gene set analysis. The premium tools include VarElect (Stelzer et al., in press), the NGS phenotyper, TGex, the Variants annotator and prioritizer, GeneAnalytics (Ben-Ari Fuchs et al., 2016), a comprehensive gene set analysis tool, and GeneALaCart, the GeneCards batch querying application. All enable streamlined analyses of NGS-derived data to deliver enhanced results and extract meaningful insights. GeneCards and its suite members can be accessed at *http://www.genecards.org/*.

GeneCards Generation and Versioning

Each GeneCards entry describes a gene, displayed in a computerized Web card. The information about each gene is organized into 17 main sections (Table 1.30.1). GeneCards are generated periodically by an automated process, which first compiles an integrated list of gene names from a variety of sources, and then unifies the annotative information from each source and associates it with the appropriate section(s). This process is performed for each version, which results in a new database 'build'. The current version number and version history can be obtained from *http://www.genecards.org/Guide/News*. More information about the build process and computed information can be obtained from *http://www.genecards.org/Guide/AboutGeneCards*, and is described in the Commentary section, below.

Table 1.30.1 GeneCards Section Names List

Section name	Step number[a]
Aliases	3
Summaries	4
Genomics	5
Proteins	6
Domains	7
Function	8
Localization	9
Pathways (and interactions)	10
Drugs (and compounds)	11
Transcripts	12
Expression	13
Orthologs	14
Paralogs	15
Variants	16
Disorders	17
Publications	18
Products	19
Sources	20

[a] The number of the step in Basic Protocol 2, where this section is described.

GeneCards Version 4 was introduced in June 2015, featuring a major facelift enabling a better user experience, including improved consolidation of data and product information, while preserving legacy content and functionality, as well as easy navigation to other suite sites. The infrastructure was revamped to help facilitate more timely incremental updates and has scalable capacity to accommodate growing traffic, as well as more robust Elasticsearch-based querying (Kononenko et al., 2014). An integrated knowledgebase centralizes all of the information, showcased in the various GeneCards suite members.

The GeneCards Suite Members

In addition to GeneCards, The GeneCards Suite is made up of several member databases and tools that focus on different facets of genomics and proteomics:

MalaCards: a database of human maladies and their annotations (Rappaport et al., 2013, 2014)
LifeMap Discovery: a compendium of embryonic development for stem cell research and regenerative medicine (Edgar et al., 2013)
PathCards: the integrated database of human pathways and their annotations (Belinky et al., 2015)
GeneAnalytics: a gene-set analysis tool for finding commonalities within next-generation sequencing, RNAseq, and microarray data (Ben-Ari Fuchs et al., 2016)
GeneALaCart: a batch querying application that allows retrieval of information about multiple genes
VarElect: the next generation sequencing phenotype interpreter for prioritizing a list of genes in relation to a phenotype (Stelzer et al., in press)
GenesLikeMe: a tool for finding related genes by using a similarity metric that is based on shared GeneCards descriptors (formerly GeneDecks Partner Hunter) (Stelzer et al., 2009)
GeneLoc: an integrated map for each human chromosome (Rosen et al., 2003)
TGex: Translational Genomics expert, which combines the phenotype interpretation power of VarElect at the end of a variant calling pipeline with flexible filtration capabilities.

GeneCards Begets VarElect

The advent of next-generation sequencing (NGS) has provided a crucial means for associating diseases with their causative genes. This necessitates the prioritization of thousands of variants, churned out from NGS machines, first by filtering criteria and subsequently by relating their corresponding genes to a desired disease/phenotype. VarElect evolved from a simple multigene "batch" query in GeneCards and grew to become an independent NGS analysis tool. VarElect is powered by the GeneCards search engine (see Search GeneCards), which in turn exploits abundant annotations available in the other suite members such as MalaCards and PathCards. VarElect finds relevant hits for a given phenotype based on the standard GeneCards keyword search, except that it addresses a smaller gene space, as instructed by the user-supplied gene list. The display of direct gene-phenotype relations greatly resembles typical GeneCards search results. The more sophisticated indirect feature infers "guilt by association" relationships between genes and phenotype terms using gene-to-gene relations from pathway, protein-protein interaction, textual mining, and paralogy information. These results are conveyed by detailing phenotype relations to "implicating genes," together with the evidence for connections of the latter to genes in the NGS input list.

The Protocols

GeneCards can be navigated in a variety of ways, as described in the protocols below. The search box on the home page is typically the initial starting point, where one can submit a search for a gene symbol or name. The search box in the banner, available on every page of the GeneCards site, allows one to submit free text as a query string, including Boolean expressions. The GeneCards advanced search allows querying GeneCards for more specific results by focusing on a specific type of search term and/or by limiting the search to specific fields within the card. A description of the search properties is available at *http://www.genecards.org/Guide/Search*. An earlier description of many aspects of GeneCards, including the various sections within a specific card, is available in Safran et al. (2010).

This unit gives a brief overview of the basic use cases of GeneCards and VarElect. Basic Protocol 1 describes how to perform searches and analyze result summaries, using the search facilities and the alphabetical index. Basic Protocol 2 describes how to scrutinize and navigate the different parts of the Web card for each gene. Basic Protocol 3 describes how to manage projects for prioritizing a gene list, originating from next-generation sequencing, with respect to member genes' relationships to specified phenotypes of interest.

BASIC PROTOCOL 1

SEARCHING AND BROWSING GENECARDS

GeneCards can be accessed on the Internet at *http://www.genecards.org/*. The homepage (Fig. 1.30.1) is a common entry point to the Web site, showcasing most of the features and tools, quick searches, and genome statistics. Links to version history, collaborators, GeneCards suite members, previous version, categories list, and a gene index are also found on the homepage, along with a short description of the Web site and the number of sources and genes. Major functionalities include: (1) 'Explore a Gene' box, which initially displays a sample gene, rich with information, that is changed for each new version; (2) random gene generator, which provides links to random genes after displaying the name of the random gene in the 'Explore a Gene' box; and (3) full search box on the top right of the page, which allows search by gene symbol, alias, or any keyword.

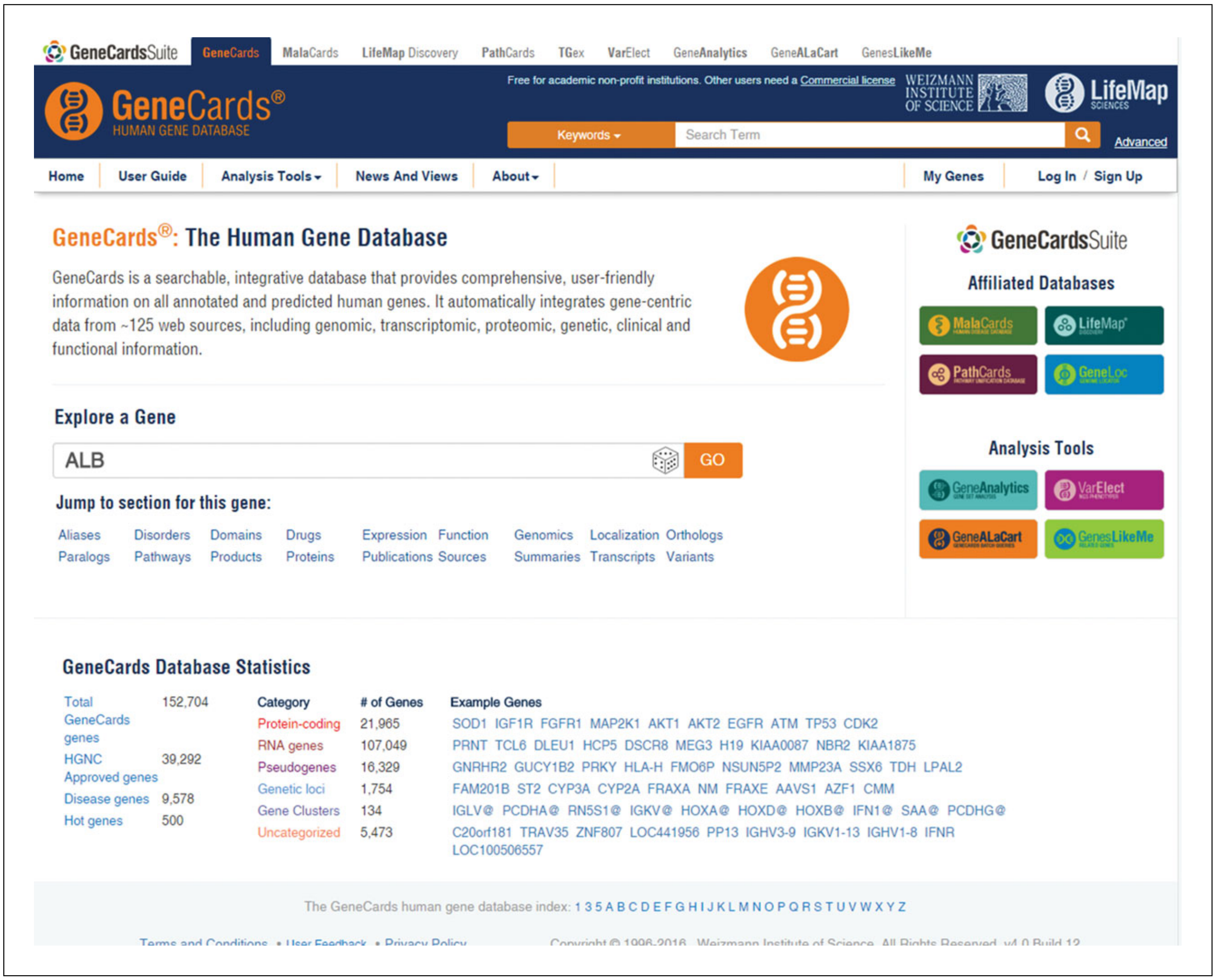
A "Jump to section" table allows rapid navigation to a specific section within the gene named in the Explore a Gene box. Below this are links to GeneCards statistics, including charts displaying the amount of information for various types of genes and also linked examples of different kinds of genes (e.g., protein-coding genes, RNA genes, pseudogenes). Following these links is a link to a list of disease genes, available when logged in. The disease gene page presents an alphabetical listing of disease gene symbols with their descriptions, loci, and a list of associated diseases, ordered and scored by their importance to the gene, and linked to MalaCards. Below the disease genes link, one can click on 'Hot genes' to retrieve a list of the most frequently viewed ("hot") genes in GeneCards. This page is updated frequently to reflect current usage. The Search box, which serves as a gateway to the gene universe within GeneCards, is at the top right corner of every page of the Web site. Next to it is a link to the Advanced Search, which allows complex queries in which each keyword can be restricted to a specific section of the GeneCard.

Various pages related to site documentation, the development team and institution, academic licensing, and companion site links are located in the main menu at the top of the homepage, and at the top of all other pages on the site.

A feedback link is available to allow users to pose questions, comments, and/or suggestions.

Necessary Resources

An up-to-date Web browser such as Google Chrome, Mozilla Firefox, Microsoft Edge, or Apple Safari



GeneCards®: The Human Gene Database

GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. It automatically integrates gene-centric data from ~125 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information.

Explore a Gene

ALB

GO

Jump to section for this gene:

Aliases Disorders Domains Drugs Expression Function Genomics Localization Orthologs
Paralogs Pathways Products Proteins Publications Sources Summaries Transcripts Variants

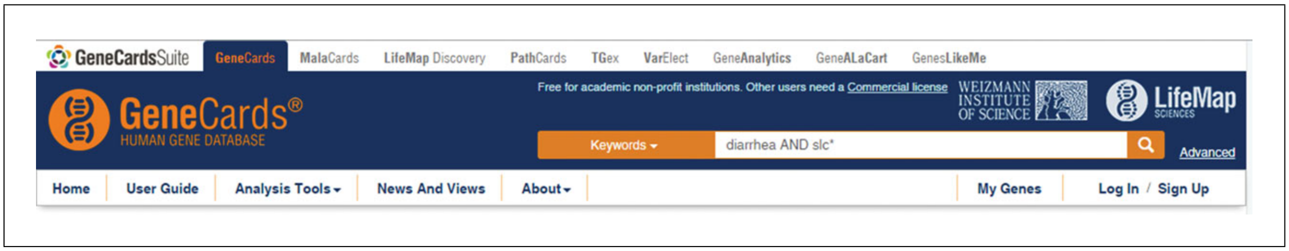
GeneCards Database Statistics

Total GeneCards genes	152,704
HGNC Approved genes	39,292
Disease genes	9,578
Hot genes	500

Category	# of Genes	Example Genes
Protein-coding	21,965	SOD1 IGF1R FGFR1 MAP2K1 AKT1 AKT2 EGFR ATM TP53 CDK2
RNA genes	107,049	PRNT TCL6 DLEU1 HCP5 DSCR8 MEG3 H19 KIAA0087 NBR2 KIAA1875
Pseudogenes	16,329	GNRHR2 GUCY1B2 PRKY HLA-H FMO6P NSUN5P2 MMP23A SSX6 TDH LPAL2
Genetic loci	1,754	FAM201B ST2 CYP3A CYP2A FRAXA NM FRAXE AAVS1 AZF1 CMM
Gene Clusters	134	IGLV@ PCDHA@ RN5S1@ IGKV@ HOXA@ HOXD@ HOXB@ IFN1@ SAA@ PCDHG@
Uncategorized	5,473	C20orf181 TRAV35 ZNF807 LOC441956 PP13 IGHV3-9 IGKV1-13 IGHV1-8 IFNR LOC100506557

The GeneCards human gene database index: 1 3 5 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Figure 1.30.1 The GeneCards homepage (*http://www.genecards.org/*) provides access to most of the features of the Web site through the keyword/symbol search box at the top of the page, the gene symbol box with jump to section links at the center of the page, and gene statistics, version history, main menu items, and more.



diarrhea AND slc*

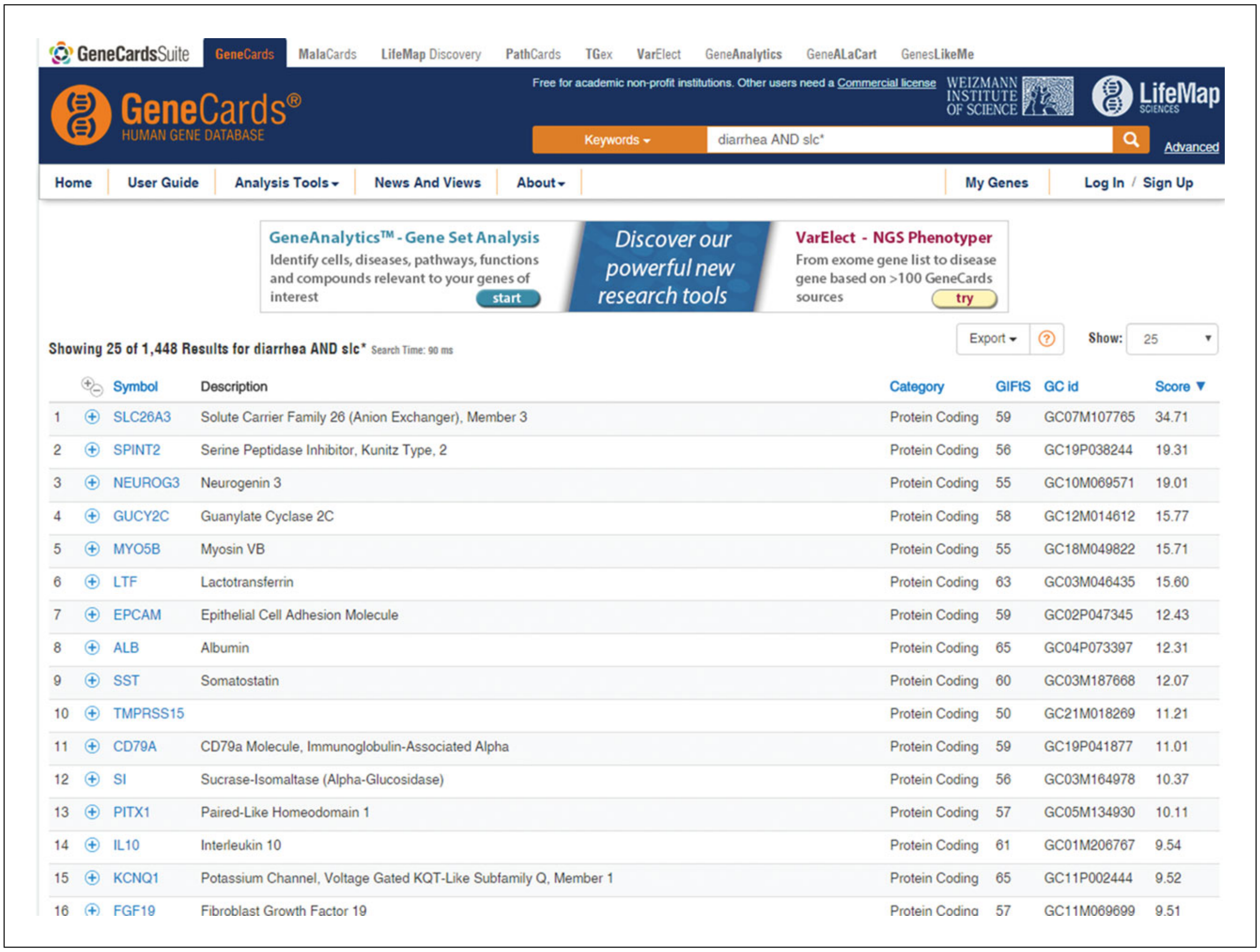
Figure 1.30.2 A GeneCards keyword query for `diarrhea AND slc*`.

Search GeneCards

1. Begin at the GeneCards home page (*http://www.genecards.org/*).

2. In the search box, at the top right corner of the page, enter `diarrhea AND slc*`, and click the magnifying glass icon to submit the query (Fig. 1.30.2).

The query term may be a disease name, gene name, or any other keyword. Boolean operators (`AND`*/*`OR`*) can be used to query GeneCards, as can wildcards (*`*`*) when placed at the end of a word. Note that Boolean operators must be capitalized to yield expected results. The GeneCards search engine is based on Elasticsearch (https://www.elastic.co/products/elasticsearch) and supports the following features: (1) stemming in all of its searches, so that similar words will also be found rather than just exact matches; (2) a*


GeneCardsSuite
GeneCards
MalaCards
LifeMap Discovery
PathCards
TGex
VarElect
GeneAnalytics
GeneALaCart
GenesLikeMe
GeneCards®
HUMAN GENE DATABASE
Free for academic non-profit institutions. Other users need a Commercial license
WEIZMANN INSTITUTE OF SCIENCE
LifeMap SCIENCES
Keywords
Advanced
Home
User Guide
Analysis Tools
News And Views
About
My Genes
Log In / Sign Up
GeneAnalytics™ - Gene Set Analysis
Identify cells, diseases, pathways, functions and compounds relevant to your genes of interest
start
Discover our powerful new research tools
VarElect - NGS Phenotyper
From exome gene list to disease gene based on >100 GeneCards sources
try
Export
Show:
25

diarrhea AND slc*

Showing 25 of 1,448 Results for diarrhea AND slc* Search Time: 90 ms

		Symbol	Description	Category	GIFtS	GC id	Score ▼
1	⊕	SLC26A3	Solute Carrier Family 26 (Anion Exchanger), Member 3	Protein Coding	59	GC07M107765	34.71
2	⊕	SPINT2	Serine Peptidase Inhibitor, Kunitz Type, 2	Protein Coding	56	GC19P038244	19.31
3	⊕	NEUROG3	Neurogenin 3	Protein Coding	55	GC10M069571	19.01
4	⊕	GUCY2C	Guanylate Cyclase 2C	Protein Coding	58	GC12M014612	15.77
5	⊕	MYO5B	Myosin VB	Protein Coding	55	GC18M049822	15.71
6	⊕	LTF	Lactotransferrin	Protein Coding	63	GC03M046435	15.60
7	⊕	EPCAM	Epithelial Cell Adhesion Molecule	Protein Coding	59	GC02P047345	12.43
8	⊕	ALB	Albumin	Protein Coding	65	GC04P073397	12.31
9	⊕	SST	Somatostatin	Protein Coding	60	GC03M187668	12.07
10	⊕	TMPRSS15		Protein Coding	50	GC21M018269	11.21
11	⊕	CD79A	CD79a Molecule, Immunoglobulin-Associated Alpha	Protein Coding	59	GC19P041877	11.01
12	⊕	SI	Sucrase-Isomaltase (Alpha-Glucosidase)	Protein Coding	56	GC03M164978	10.37
13	⊕	PITX1	Paired-Like Homeodomain 1	Protein Coding	57	GC05M134930	10.11
14	⊕	IL10	Interleukin 10	Protein Coding	61	GC01M206767	9.54
15	⊕	KCNQ1	Potassium Channel, Voltage Gated KQT-Like Subfamily Q, Member 1	Protein Coding	65	GC11P002444	9.52
16	⊕	FGF19	Fibroblast Growth Factor 19	Protein Coding	57	GC11M069699	9.51

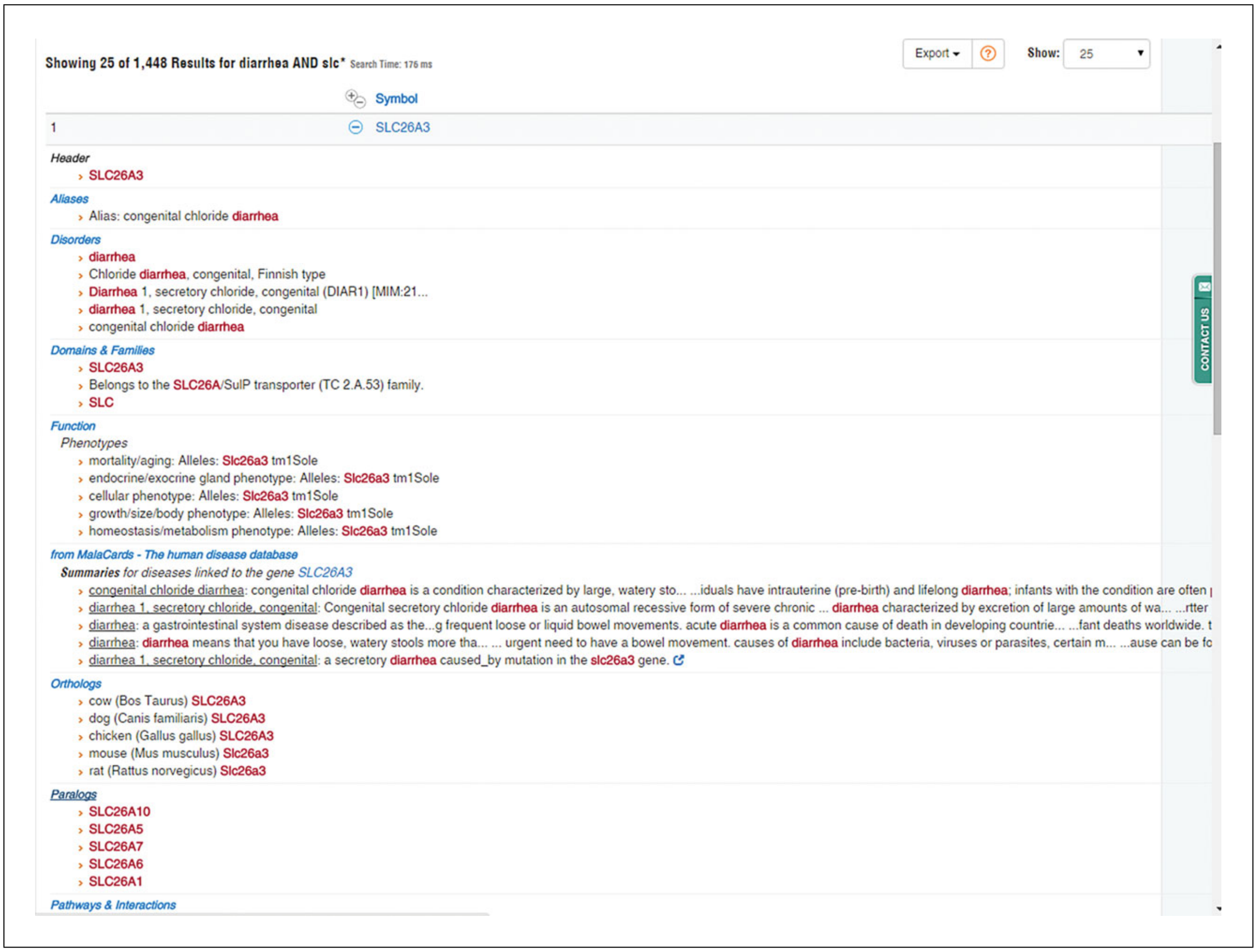
Figure 1.30.3 Results of the search processed in Basic Protocol 1. There are 1,448 human genes containing the searched string pattern, ranked by a relevance score. The "+" links open the highlighted hit context (minicard).

search for multiple words (e.g., Alzheimer disease) behaves as an AND *within the entire set of a gene's annotations; i.e., each of the words must exist at least once in each of the matched GeneCards. To search for an exact phrase, simply add quotes to your search. (3) Parentheses should be used in searches for complex Boolean strings in order to define precedence; otherwise* AND *operations will take precedence over* OR *operations. (4) Gene symbol, name/alias summary, function, and disorder fields are boosted, so searches for those terms will match GeneCards accordingly named with higher scores than GeneCards merely containing the strings in other fields such as publications.*

Note that, when relevant, an alternate "did you mean" search string option is offered to the user when spelling mistakes are suspected.

3. Running the `diarrhea AND slc*` query in February 2016 returned a list with 1448 items, shown in the search results title at the top of the results table (Fig. 1.30.3).

The GeneCards search results page includes a numbered list of genes, each of them displaying hit context information (minicards), description, gene category, GeneCards Inferred Functional Score (GIFtS) (Harel et al., 2009), GeneCards ID, and search specificity score given by Elasticsearch. The user can open (or close) all displayed minicards using the "+−" icon at the top of the minicards column. Each minicard highlights the hit within the specific entry; one can further click on the specific section and view the hit within the card. The GeneCards ID is a persistent identifier (see Commentary). All of the non-trivial information in all of the GeneCards is indexed, and provides fodder for the search engine.



Showing 25 of 1,448 Results for diarrhea AND slc* Search Time: 176 ms

1 SLC26A3

Header
- SLC26A3

Aliases
- Alias: congenital chloride diarrhea

Disorders
- diarrhea
- Chloride diarrhea, congenital, Finnish type
- Diarrhea 1, secretory chloride, congenital (DIAR1) [MIM:21...
- diarrhea 1, secretory chloride, congenital
- congenital chloride diarrhea

Domains & Families
- SLC26A3
- Belongs to the SLC26A/SulP transporter (TC 2.A.53) family.
- SLC

Function
Phenotypes
- mortality/aging: Alleles: Slc26a3 tm1Sole
- endocrine/exocrine gland phenotype: Alleles: Slc26a3 tm1Sole
- cellular phenotype: Alleles: Slc26a3 tm1Sole
- growth/size/body phenotype: Alleles: Slc26a3 tm1Sole
- homeostasis/metabolism phenotype: Alleles: Slc26a3 tm1Sole

from MalaCards - The human disease database
***Summaries** for diseases linked to the gene SLC26A3*
- congenital chloride diarrhea: congenital chloride diarrhea is a condition characterized by large, watery sto... ...iduals have intrauterine (pre-birth) and lifelong diarrhea; infants with the condition are often
- diarrhea 1, secretory chloride, congenital: Congenital secretory chloride diarrhea is an autosomal recessive form of severe chronic ... diarrhea characterized by excretion of large amounts of wa... ...rtter
- diarrhea: a gastrointestinal system disease described as the...g frequent loose or liquid bowel movements. acute diarrhea is a common cause of death in developing countrie... ...fant deaths worldwide. t
- diarrhea: diarrhea means that you have loose, watery stools more tha... ... urgent need to have a bowel movement. causes of diarrhea include bacteria, viruses or parasites, certain m... ...ause can be fo
- diarrhea 1, secretory chloride, congenital: a secretory diarrhea caused_by mutation in the slc26a3 gene.

Orthologs
- cow (Bos Taurus) SLC26A3
- dog (Canis familiaris) SLC26A3
- chicken (Gallus gallus) SLC26A3
- mouse (Mus musculus) Slc26a3
- rat (Rattus norvegicus) Slc26a3

Paralogs
- SLC26A10
- SLC26A5
- SLC26A7
- SLC26A6
- SLC26A1

Pathways & Interactions

Figure 1.30.4 Highlighted expanded search hit context (minicard) of the first hit of the search performed in Basic Protocol 1. The "+" sign to the left of the gene name opens and closes the minicard. The "+−" sign at the top of the minicards column opens all visible minicards.

4. Above the results on the right side, click the Export button to open a menu of export options. Using this feature, one can generate a copyable list of all of the symbols in the results. Using the other options in this feature, one can also save the symbols to an Excel file, or export them to GeneCards Suite analysis tools—VarElect, GeneAnalytics, or GeneALaCart.

5. Click the "+" icon on the first line to open the hit context information (minicard) for SLC26A3 (Fig. 1.30.4).

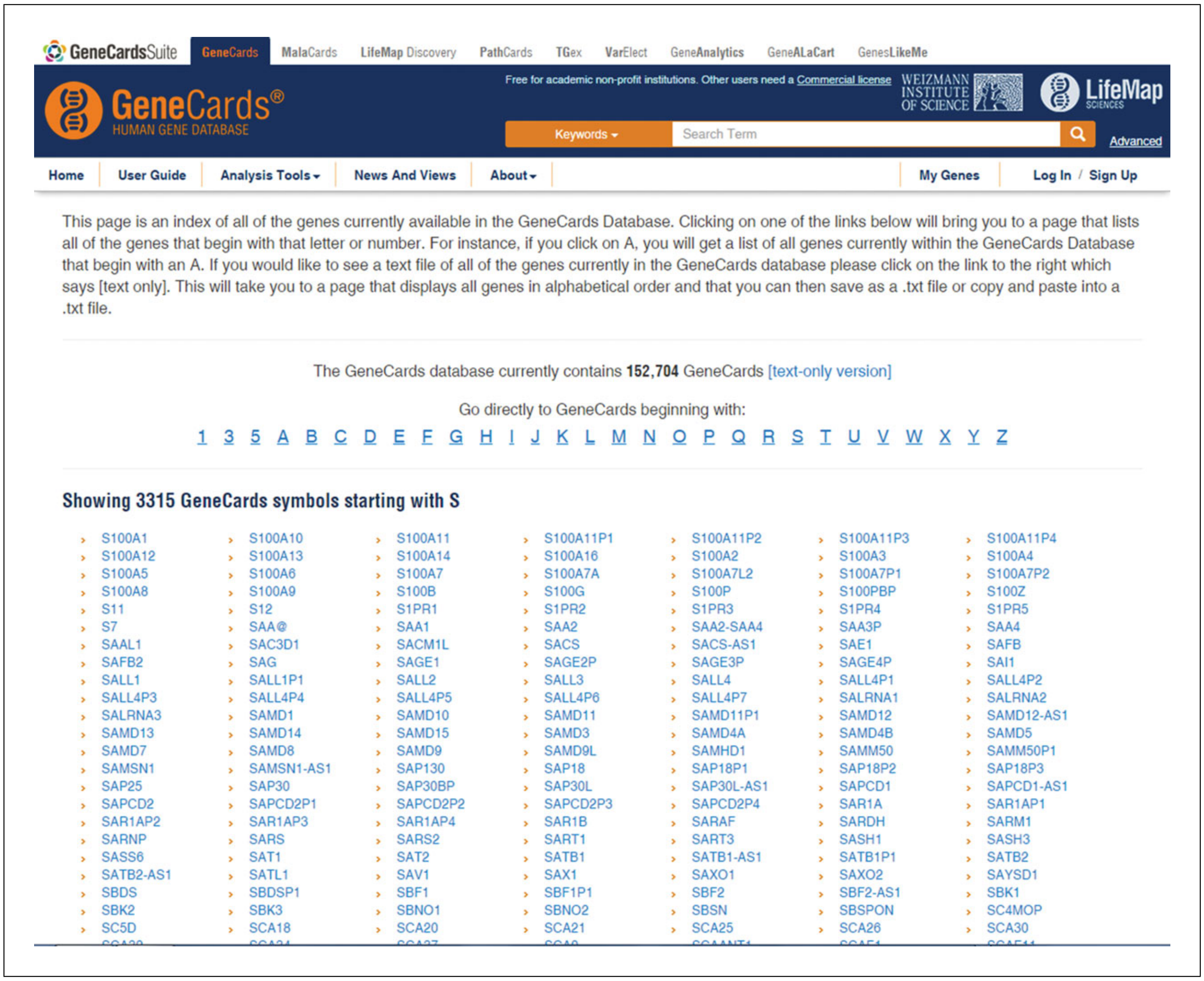
6. Click the "paralogs" hyperlink within the opened minicard, which directs to the paralogs section of the GeneCard.

Browse GeneCards by alphabetical index

7. To further study SLC genes, use the alphabetical index at the very bottom of each gene page. A page containing all of the genes for each letter is opened. Click the "S" hyperlink to view all genes beginning with the letter "S" (Fig. 1.30.5).

   The user can browse the alphabetical index to view all of the genes. The list displays the gene symbol, linked to the GeneCard.

8. Scroll down to SLC26A4 and click to view this gene.



This page is an index of all of the genes currently available in the GeneCards Database. Clicking on one of the links below will bring you to a page that lists all of the genes that begin with that letter or number. For instance, if you click on A, you will get a list of all genes currently within the GeneCards Database that begin with an A. If you would like to see a text file of all of the genes currently in the GeneCards database please click on the link to the right which says [text only]. This will take you to a page that displays all genes in alphabetical order and that you can then save as a .txt file or copy and paste into a .txt file.

The GeneCards database currently contains **152,704** GeneCards [text-only version]

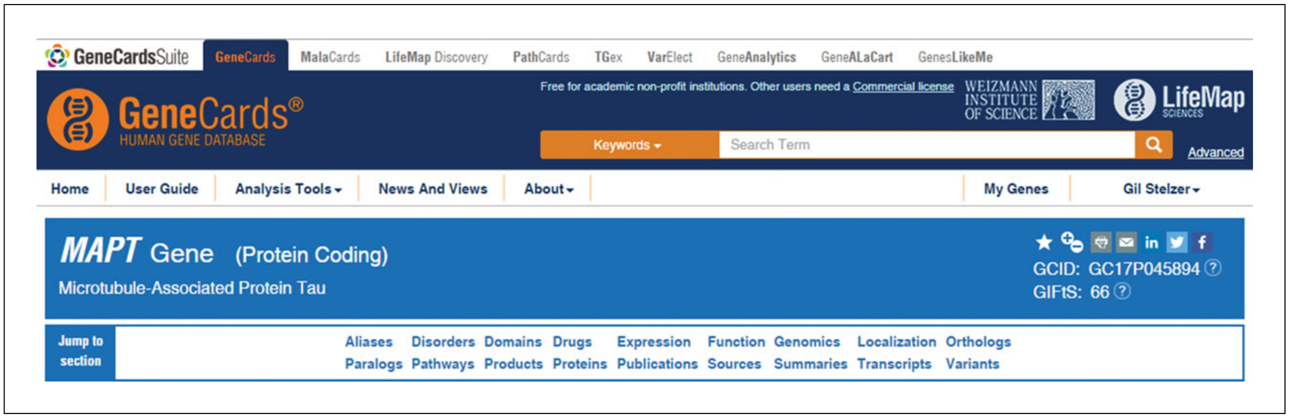
Go directly to GeneCards beginning with:

1 3 5 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Showing 3315 GeneCards symbols starting with S

S100A1	S100A10	S100A11	S100A11P1	S100A11P2	S100A11P3	S100A11P4
S100A12	S100A13	S100A14	S100A16	S100A2	S100A3	S100A4
S100A5	S100A6	S100A7	S100A7A	S100A7L2	S100A7P1	S100A7P2
S100A8	S100A9	S100B	S100G	S100P	S100PBP	S100Z
S11	S12	S1PR1	S1PR2	S1PR3	S1PR4	S1PR5
S7	SAA@	SAA1	SAA2	SAA2-SAA4	SAA3P	SAA4
SAAL1	SAC3D1	SACM1L	SACS	SACS-AS1	SAE1	SAFB
SAFB2	SAG	SAGE1	SAGE2P	SAGE3P	SAGE4P	SAI1
SALL1	SALL1P1	SALL2	SALL3	SALL4	SALL4P1	SALL4P2
SALL4P3	SALL4P4	SALL4P5	SALL4P6	SALL4P7	SALRNA1	SALRNA2
SALRNA3	SAMD1	SAMD10	SAMD11	SAMD11P1	SAMD12	SAMD12-AS1
SAMD13	SAMD14	SAMD15	SAMD3	SAMD4A	SAMD4B	SAMD5
SAMD7	SAMD8	SAMD9	SAMD9L	SAMHD1	SAMM50	SAMM50P1
SAMSN1	SAMSN1-AS1	SAP130	SAP18	SAP18P1	SAP18P2	SAP18P3
SAP25	SAP30	SAP30BP	SAP30L	SAP30L-AS1	SAPCD1	SAPCD1-AS1
SAPCD2	SAPCD2P1	SAPCD2P2	SAPCD2P3	SAPCD2P4	SAR1A	SAR1AP1
SAR1AP2	SAR1AP3	SAR1AP4	SAR1B	SARAF	SARDH	SARM1
SARNP	SARS	SARS2	SART1	SART3	SASH1	SASH3
SASS6	SAT1	SAT2	SATB1	SATB1-AS1	SATB1P1	SATB2
SATB2-AS1	SATL1	SAV1	SAX1	SAXO1	SAXO2	SAYSD1
SBDS	SBDSP1	SBF1	SBF1P1	SBF2	SBF2-AS1	SBK1
SBK2	SBK3	SBNO1	SBNO2	SBSN	SBSPON	SC4MOP
SC5D	SCA18	SCA20	SCA21	SCA25	SCA26	SCA30

Figure 1.30.5 The GeneCards alphabetical gene index for the letter ‘S’.



MAPT Gene (Protein Coding)

Microtubule-Associated Protein Tau

GCID: GC17P045894

GIFtS: 66

Figure 1.30.6 The GeneCards disease header, located at the top of each GeneCard, shown for MAPT.

BASIC PROTOCOL 2

EXPLORING A GENECARD

From the user perspective, the central entry of GeneCards is the gene page, referred to as the Web “card,” or simply GeneCard. This is where one can find all available information pertaining to a gene of interest. The information within a GeneCard is divided into 17 sections (plus a numbered sources section), with a “jump-to-section” component at the top of the page (Fig. 1.30.6), allowing navigation among the different sections. When scrolling, the “jump to section” also appears at the top of the page, with the current section highlighted, along with quick links to research products. A gray tab (or boxed arrow on a wide screen) on the bottom left of the page (following the user during scrolling) allows navigation to the top of the card. On the right side of the page, there is a 'Contact us' link that expands into a small box when clicked, allowing one to submit questions or comments without navigating away from the GeneCard. Documentation is accessible via hyperlinks, often context-specific, from within many parts of the GeneCard, by clicking on the question mark icon. Table 1.30.1 provides a comprehensive list of the GeneCards sections, along with the corresponding step number in the protocol below.

The card displays gene-specific information, and contains deep links to supporting sources, often with superscripts when multiple sources contain details about the datum. The sources section, at the bottom of the card, contains the list of all of the sources that contributed information to GeneCards, annotated with the same superscripts used in the subsections, with a hyperlink to each source's home page.

Necessary Resources

An up-to-date Web browser such as Google Chrome, Mozilla Firefox, Microsoft Edge, or Apple Safari

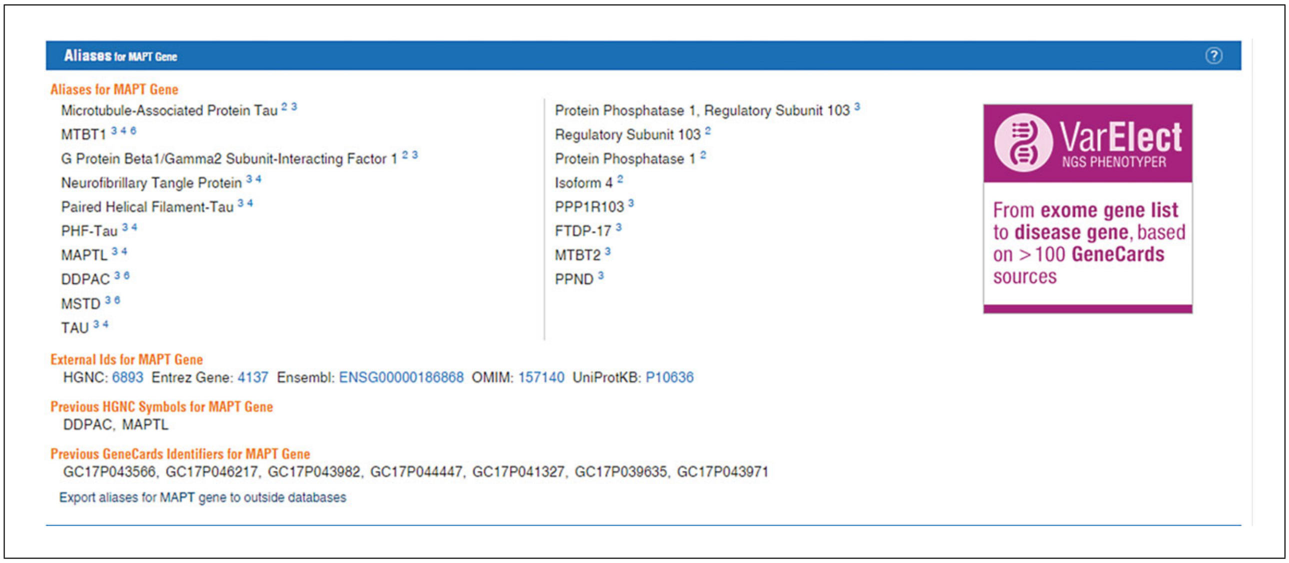
1. Access the homepage (*http://www.genecards.org/*). In the Explore a Gene box, enter `MAPT` and click the Go button.
2. Inspect the GeneCard header (Fig. 1.30.6).

   The header contains the gene name, category, description, GeneCards ID, and GeneCards Inferred Functionality Score (GIFtS). For each gene, a unique internal ID is generated, composed of the letters GC, followed by the chromosome number (where '00' indicates unknown chromosome and 'MT' indicates the mitochondria), 'P' or 'M' for orientation (Plus or Minus strand), and approximate kilobase start coordinate. This ID is modified slightly for genes that are not currently placed with certainty on the reference sequence (http://www.genecards.org/Guide/GCids; Rosen et al., 2003). The gene category, such as protein coding or RNA gene, is determined by looking at the annotation from several different resources. When the information is unavailable, a gene is considered 'uncategorized'. In order to produce a direct quantification of the functionality of a gene, a GeneCards Inferred Functionality Score (GIFtS) is computed. The GIFtS defines the richness of information in each GeneCard (Harel et al., 2009), where a higher score is given to a more annotated card. See http://www.genecards.org/Guide/GeneCard#GIFtS for a full discussion of how GIFtS are calculated. Above the GC ID are several linked icons, including a star to add this gene to My Genes, a plus/minus to expand all tables, print and e-mail icons, and links to share in social media. The My Genes feature allows the user to mark genes of interest for future reference, and to follow and comment on them. Below the header is a "Jump to section" navigation box with links to each of the sections.

3. Go to the Aliases section by either scrolling down the page or using the jump menu on the top of the page (Fig. 1.30.7).

   This section includes the following subsections

   a. *Aliases & Descriptions subsection:* Displays synonyms and aliases for the relevant GeneCards gene, as extracted from a subset of the sources listed at the bottom of the page. Strongly similar aliases are included, for source inclusiveness and to match common expectations. Multi-word descriptions are listed in the left-hand column, while single-word aliases are listed in the right-hand column. The alias list is sorted by the count of contributing sources, sub-sorted by descending length.


VarElect
NGS PHENOTYPER
From exome gene list to disease gene, based on >100 GeneCards sources

Aliases for MAPT Gene

Aliases for MAPT Gene

Microtubule-Associated Protein Tau [2,3]
MTBT1 [3,4,6]
G Protein Beta1/Gamma2 Subunit-Interacting Factor 1 [2,3]
Neurofibrillary Tangle Protein [3,4]
Paired Helical Filament-Tau [3,4]
PHF-Tau [3,4]
MAPTL [3,4]
DDPAC [3,6]
MSTD [3,6]
TAU [3,4]
Protein Phosphatase 1, Regulatory Subunit 103 [3]
Regulatory Subunit 103 [2]
Protein Phosphatase 1 [2]
Isoform 4 [2]
PPP1R103 [3]
FTDP-17 [3]
MTBT2 [3]
PPND [3]

External Ids for MAPT Gene

HGNC: 6893 Entrez Gene: 4137 Ensembl: ENSG00000186868 OMIM: 157140 UniProtKB: P10636

Previous HGNC Symbols for MAPT Gene

DDPAC, MAPTL

Previous GeneCards Identifiers for MAPT Gene

GC17P043566, GC17P046217, GC17P043982, GC17P044447, GC17P041327, GC17P039635, GC17P043971

Export aliases for MAPT gene to outside databases

Figure 1.30.7 Portion of the GeneCard for MAPT displaying its aliases and descriptions, external IDs, and previous identifiers and symbols. Identical aliases are consolidated, with superscripts corresponding to relevant sources.

b. *External IDs subsection:* Displays external IDs, which are cross-references to IDs of external gene/protein databases/ontologies at HGNC, (*http://www.genenames.org*), NCBI's Entrez Gene (*http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene*), Ensembl (*http://www.ensembl.org/index.html*), UniProt (*http://www.uniprot.org*), OMIM (*http://www.omim.org;* Hamosh et al., 2005), and others. The external IDs are deep-linked to the relevant gene's page at the source's Web site.

c. *Previous HGNC Identifiers subsection:* Displays previous approved HGNC symbols.

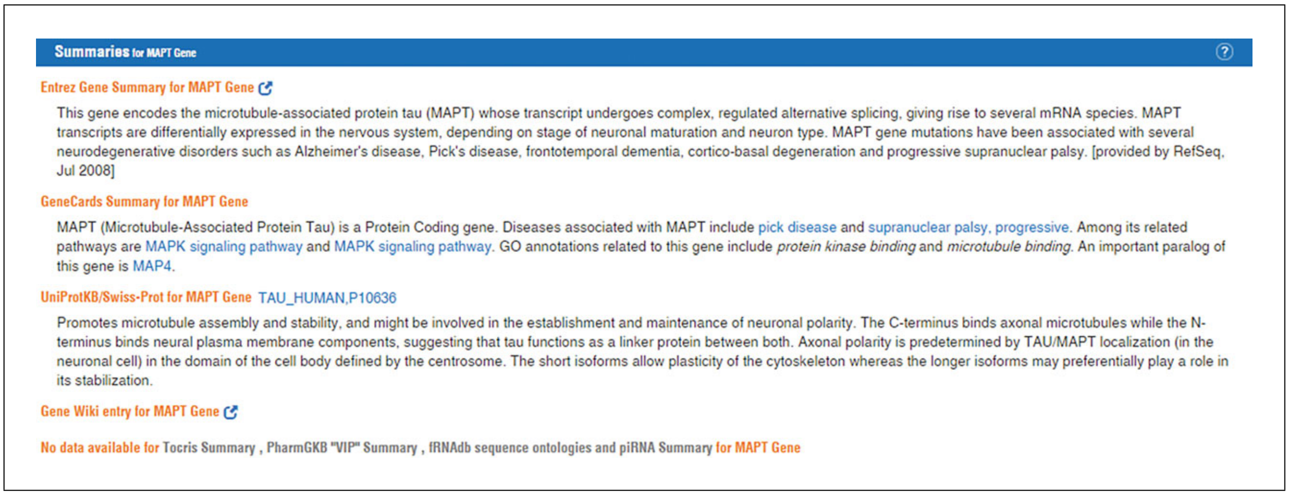
d. *Previous GeneCards Identifiers subsection:* Displays GC IDs from previous versions of GeneCards. While these IDs may change between versions, usually when the reported location of a gene is changed, the old ID is persistent and will remain associated with the gene.

e. *Export Aliases subsection:* This link displays a search box that allows the user to query outside databases (PubMed, OMIM, and NCBI Bookshelf) for information on the gene. The query can be simple or complex and allows the user to select one or more aliases and/or one or more diseases and/or enter a search string before clicking the search button to submit the search to the selected database.

In the GeneCard for MAPT, the user can see that different sources give different names for the same gene (e.g., Microtubule-Associated Protein Tau from HGNC and Entrez Gene, and Neurofibrillary Tangle Protein from Entrez Gene and UniProtKB), highlighting the major richness of GeneCards in the comprehensive naming arena.

4. Go to the Summaries section (Fig. 1.30.8).

This section displays descriptions of the gene from a variety of sources, as well as a GeneCards-generated summary. Summaries typically include the gene name, its identifier/protein (where applicable), protein function, and associated diseases. For example, the MAPT Entrez Gene Summary includes: "This gene encodes the microtubule-associated protein tau (MAPT) whose transcript undergoes complex, regulated alternative splicing, giving rise to several mRNA species". Summaries from some of the sources are fully displayed (e.g., Entrez Gene, UniProtKB), while for others, only a deep link is displayed (e.g., GeneWiki, PharmGKB). GeneCards-generated summaries highlight the gene's significant annotations (e.g., category, associated diseases, pathways, etc.). For MAPT, the GeneCards summary includes the following annotation: "Diseases associated with MAPT include pick disease and supranuclear palsy, progressive."



Summaries for MAPT Gene

Entrez Gene Summary for MAPT Gene

This gene encodes the microtubule-associated protein tau (MAPT) whose transcript undergoes complex, regulated alternative splicing, giving rise to several mRNA species. MAPT transcripts are differentially expressed in the nervous system, depending on stage of neuronal maturation and neuron type. MAPT gene mutations have been associated with several neurodegenerative disorders such as Alzheimer's disease, Pick's disease, frontotemporal dementia, cortico-basal degeneration and progressive supranuclear palsy. [provided by RefSeq, Jul 2008]

GeneCards Summary for MAPT Gene

MAPT (Microtubule-Associated Protein Tau) is a Protein Coding gene. Diseases associated with MAPT include pick disease and supranuclear palsy, progressive. Among its related pathways are MAPK signaling pathway and MAPK signaling pathway. GO annotations related to this gene include *protein kinase binding* and *microtubule binding*. An important paralog of this gene is MAP4.

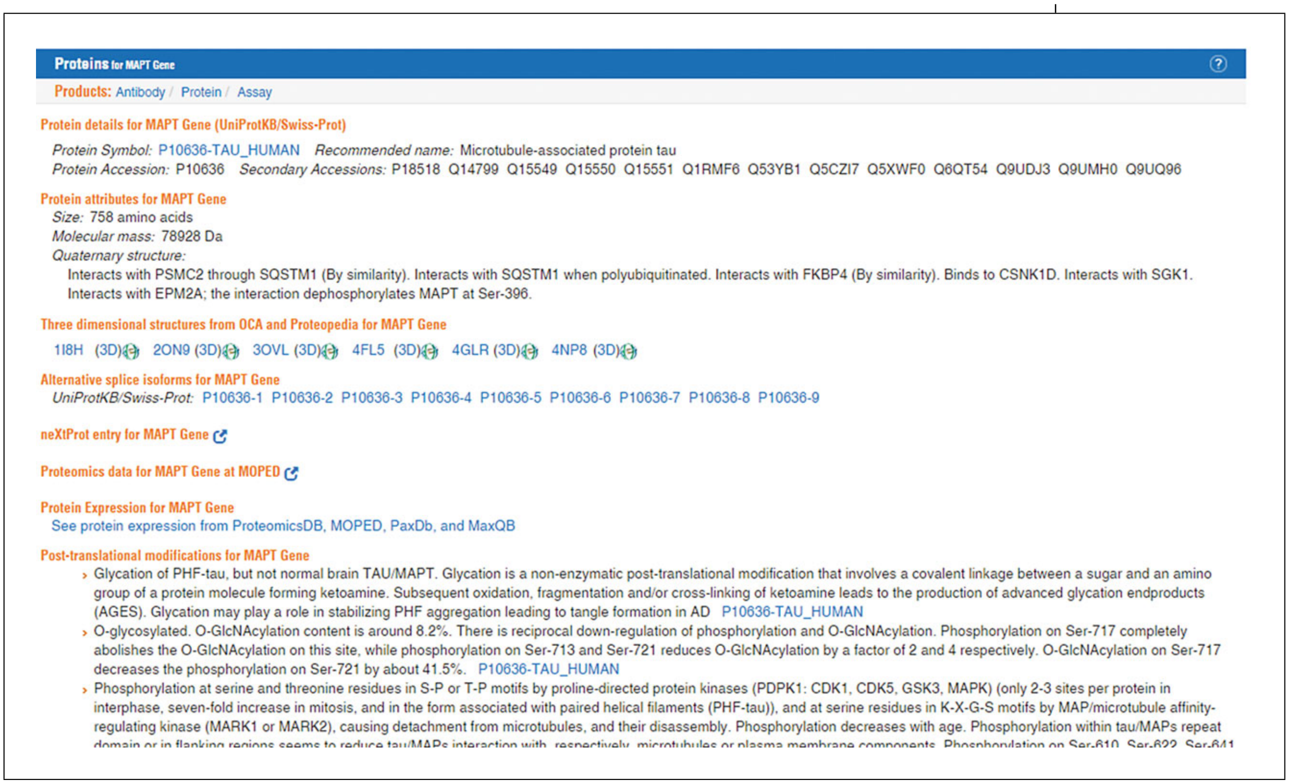
UniProtKB/Swiss-Prot for MAPT Gene TAU_HUMAN,P10636

Promotes microtubule assembly and stability, and might be involved in the establishment and maintenance of neuronal polarity. The C-terminus binds axonal microtubules while the N-terminus binds neural plasma membrane components, suggesting that tau functions as a linker protein between both. Axonal polarity is predetermined by TAU/MAPT localization (in the neuronal cell) in the domain of the cell body defined by the centrosome. The short isoforms allow plasticity of the cytoskeleton whereas the longer isoforms may preferentially play a role in its stabilization.

Gene Wiki entry for MAPT Gene

No data available for Tocris Summary , PharmGKB "VIP" Summary , fRNAdb sequence ontologies and piRNA Summary for MAPT Gene

Figure 1.30.8 Portion of the GeneCard for MAPT displaying a collection of detailed descriptions of the gene from a variety of sources. Links go to the summary or (in the case of GeneCards) relevant subsection at the source.



Proteins for MAPT Gene

Products: Antibody / Protein / Assay

Protein details for MAPT Gene (UniProtKB/Swiss-Prot)

Protein Symbol: P10636-TAU_HUMAN *Recommended name:* Microtubule-associated protein tau
Protein Accession: P10636 *Secondary Accessions:* P18518 Q14799 Q15549 Q15550 Q15551 Q1RMF6 Q53YB1 Q5CZI7 Q5XWF0 Q6QT54 Q9UDJ3 Q9UMH0 Q9UQ96

Protein attributes for MAPT Gene

Size: 758 amino acids
Molecular mass: 78928 Da
Quaternary structure:
Interacts with PSMC2 through SQSTM1 (By similarity). Interacts with SQSTM1 when polyubiquitinated. Interacts with FKBP4 (By similarity). Binds to CSNK1D. Interacts with SGK1. Interacts with EPM2A; the interaction dephosphorylates MAPT at Ser-396.

Three dimensional structures from OCA and Proteopedia for MAPT Gene

1I8H (3D) 2ON9 (3D) 3OVL (3D) 4FL5 (3D) 4GLR (3D) 4NP8 (3D)

Alternative splice isoforms for MAPT Gene

UniProtKB/Swiss-Prot: P10636-1 P10636-2 P10636-3 P10636-4 P10636-5 P10636-6 P10636-7 P10636-8 P10636-9

neXtProt entry for MAPT Gene

Proteomics data for MAPT Gene at MOPED

Protein Expression for MAPT Gene

See protein expression from ProteomicsDB, MOPED, PaxDb, and MaxQB

Post-translational modifications for MAPT Gene

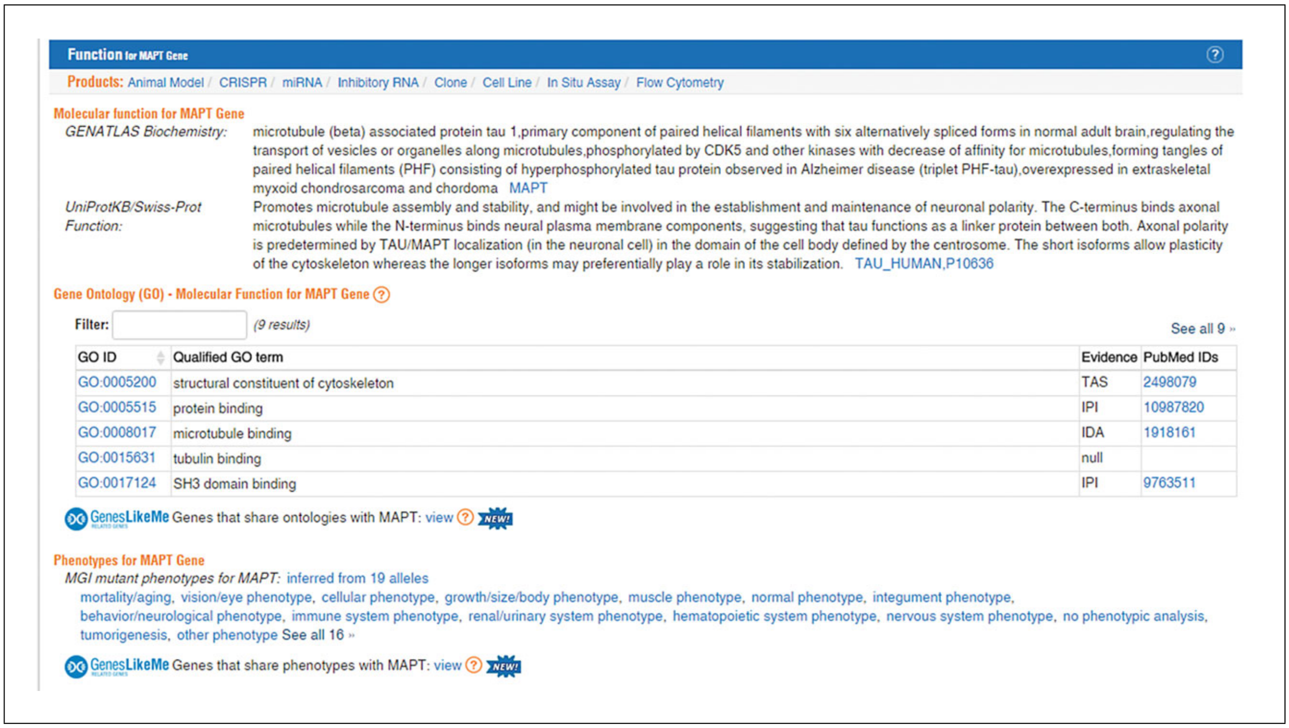
- Glycation of PHF-tau, but not normal brain TAU/MAPT. Glycation is a non-enzymatic post-translational modification that involves a covalent linkage between a sugar and an amino group of a protein molecule forming ketoamine. Subsequent oxidation, fragmentation and/or cross-linking of ketoamine leads to the production of advanced glycation endproducts (AGES). Glycation may play a role in stabilizing PHF aggregation leading to tangle formation in AD P10636-TAU_HUMAN
- O-glycosylated. O-GlcNAcylation content is around 8.2%. There is reciprocal down-regulation of phosphorylation and O-GlcNAcylation. Phosphorylation on Ser-717 completely abolishes the O-GlcNAcylation on this site, while phosphorylation on Ser-713 and Ser-721 reduces O-GlcNAcylation by a factor of 2 and 4 respectively. O-GlcNAcylation on Ser-717 decreases the phosphorylation on Ser-721 by about 41.5%. P10636-TAU_HUMAN
- Phosphorylation at serine and threonine residues in S-P or T-P motifs by proline-directed protein kinases (PDPK1: CDK1, CDK5, GSK3, MAPK) (only 2-3 sites per protein in interphase, seven-fold increase in mitosis, and in the form associated with paired helical filaments (PHF-tau)), and at serine residues in K-X-G-S motifs by MAP/microtubule affinity-regulating kinase (MARK1 or MARK2), causing detachment from microtubules, and their disassembly. Phosphorylation decreases with age. Phosphorylation within tau/MAPs repeat domain or in flanking regions seems to reduce tau/MAPs interaction with, respectively, microtubules or plasma membrane components. Phosphorylation on Ser-610, Ser-622, Ser-641

Figure 1.30.9 Portion of the GeneCard for MAPT displaying proteins associated with this gene, including links to three-dimensional structures. Product links in the subheading allow quick navigation to various protein products, such as antibodies and assays.

5. Go to the Genomics section.

   This section includes subsections covering regulatory elements and their products, epigenetics, genomic location and views, and RefSeq DNA sequences. The genomic view subsection contains a schematic diagram showing the gene's location on the chromosome (where applicable), as well as links to several external databases that show the gene's locations and the database's graphical viewer. Below the diagram are links to GeneLoc (http://genecards.weizmann.ac.il/geneloc/index.shtml), one of the members of the GeneCards Suite, which give further information about the gene's genomic context.

6. Go to the Proteins section (Fig. 1.30.9).


Function for MAPT Gene
Products: Animal Model / CRISPR / miRNA / Inhibitory RNA / Clone / Cell Line / In Situ Assay / Flow Cytometry

Molecular function for MAPT Gene

GENATLAS Biochemistry:	microtubule (beta) associated protein tau 1,primary component of paired helical filaments with six alternatively spliced forms in normal adult brain,regulating the transport of vesicles or organelles along microtubules,phosphorylated by CDK5 and other kinases with decrease of affinity for microtubules,forming tangles of paired helical filaments (PHF) consisting of hyperphosphorylated tau protein observed in Alzheimer disease (triplet PHF-tau),overexpressed in extraskeletal myxoid chondrosarcoma and chordoma MAPT
UniProtKB/Swiss-Prot Function:	Promotes microtubule assembly and stability, and might be involved in the establishment and maintenance of neuronal polarity. The C-terminus binds axonal microtubules while the N-terminus binds neural plasma membrane components, suggesting that tau functions as a linker protein between both. Axonal polarity is predetermined by TAU/MAPT localization (in the neuronal cell) in the domain of the cell body defined by the centrosome. The short isoforms allow plasticity of the cytoskeleton whereas the longer isoforms may preferentially play a role in its stabilization. TAU_HUMAN,P10636

Gene Ontology (GO) - Molecular Function for MAPT Gene

Filter: (9 results) See all 9 »

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0005200	structural constituent of cytoskeleton	TAS	2498079
GO:0005515	protein binding	IPI	10987820
GO:0008017	microtubule binding	IDA	1918161
GO:0015631	tubulin binding	null	
GO:0017124	SH3 domain binding	IPI	9763511

GenesLikeMe Genes that share ontologies with MAPT: view NEW!

Phenotypes for MAPT Gene

MGI mutant phenotypes for MAPT: inferred from 19 alleles

mortality/aging, vision/eye phenotype, cellular phenotype, growth/size/body phenotype, muscle phenotype, normal phenotype, integument phenotype, behavior/neurological phenotype, immune system phenotype, renal/urinary system phenotype, hematopoietic system phenotype, nervous system phenotype, no phenotypic analysis, tumorigenesis, other phenotype See all 16 »

GenesLikeMe Genes that share phenotypes with MAPT: view NEW!

Figure 1.30.10 Portion of the GeneCard for MAPT displaying gene function, including Gene Ontology Molecular Function terms. This section gives a quick overview of human and mouse phenotypes associated with the gene.

This section provides information and links about proteins associated with the gene. Subsections include protein details, such as names and accessions; protein attributes, such as size, weight, and structure; links to three-dimensional structures; and links to alternative splice isoforms. This section also lists post-translational modifications, and links to several external databases, including many with protein-related products including antibodies, recombinant proteins, and assays offered to researchers.

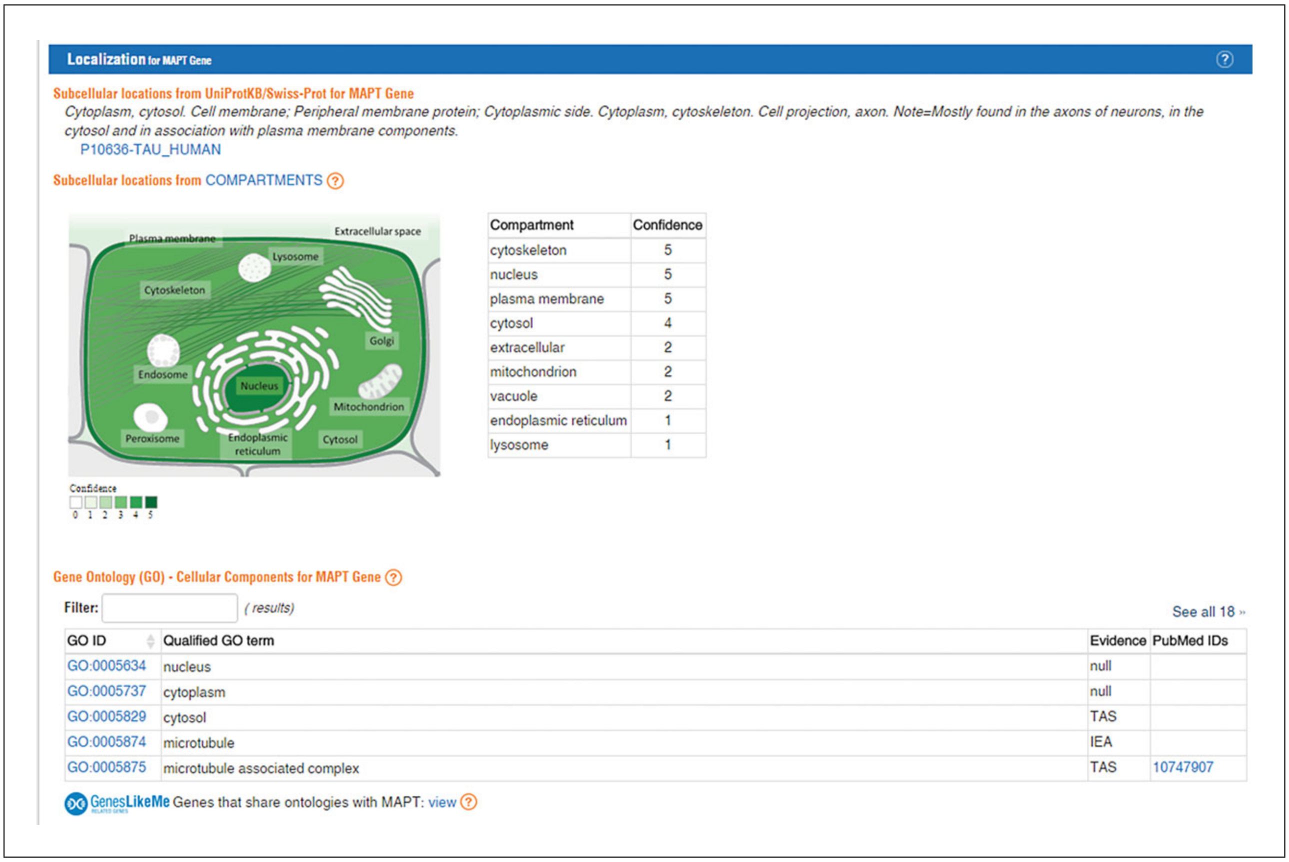
7. Go to the Domains section.

This section lists protein domains and gene families related to this gene, and links to the Web sites of databases that provide this information, including a link to a graphical view of the domain structure. This section further links to GenesLikeMe, a member of the GeneCards Suite, where the user can see other genes that share this gene's domains. For example, click the GenesLikeMe link for MAPT and submit the query to see the two genes that share domains with MAPT: MAP2 and MAP4.

8. Go to the Function section (Fig. 1.30.10).

This section includes the following subsections:

a. *Molecular function:* Displays descriptions of molecular function from external databases. There is also a table of Gene Ontology molecular functions that includes GO IDs, GO terms, Evidence, and PubMed IDs. This table can be expanded to show all results and can be refined, via the 'Filter' text box above it, to display only the results of interest. A link to GenesLikeMe allows the user to see other genes that share this gene's ontologies.
b. *Phenotypes and Animal Models:* Lists human and mouse phenotypes associated with this gene. A link to GenesLikeMe allows the user to see other genes that share these phenotypes. Also provided are links to animal models, such as mouse knockouts, that are available for this gene.
c. *Function-related products:* Links to function-related products, such as animal models, inhibitory RNA, clones, and cell lines.



Localization for MAPT Gene

Subcellular locations from UniProtKB/Swiss-Prot for MAPT Gene

Cytoplasm, cytosol. Cell membrane; Peripheral membrane protein; Cytoplasmic side. Cytoplasm, cytoskeleton. Cell projection, axon. Note=Mostly found in the axons of neurons, in the cytosol and in association with plasma membrane components.

P10636-TAU_HUMAN

Subcellular locations from COMPARTMENTS

Compartment	Confidence
cytoskeleton	5
nucleus	5
plasma membrane	5
cytosol	4
extracellular	2
mitochondrion	2
vacuole	2
endoplasmic reticulum	1
lysosome	1

Gene Ontology (GO) - Cellular Components for MAPT Gene

Filter: *(results)*

See all 18 »

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0005634	nucleus	null	
GO:0005737	cytoplasm	null	
GO:0005829	cytosol	TAS	
GO:0005874	microtubule	IEA	
GO:0005875	microtubule associated complex	TAS	10747907

GenesLikeMe Genes that share ontologies with MAPT: view

Figure 1.30.11 Portion of the GeneCard for MAPT displaying localization information, including subcellular location image from the COMPARTMENTS database.

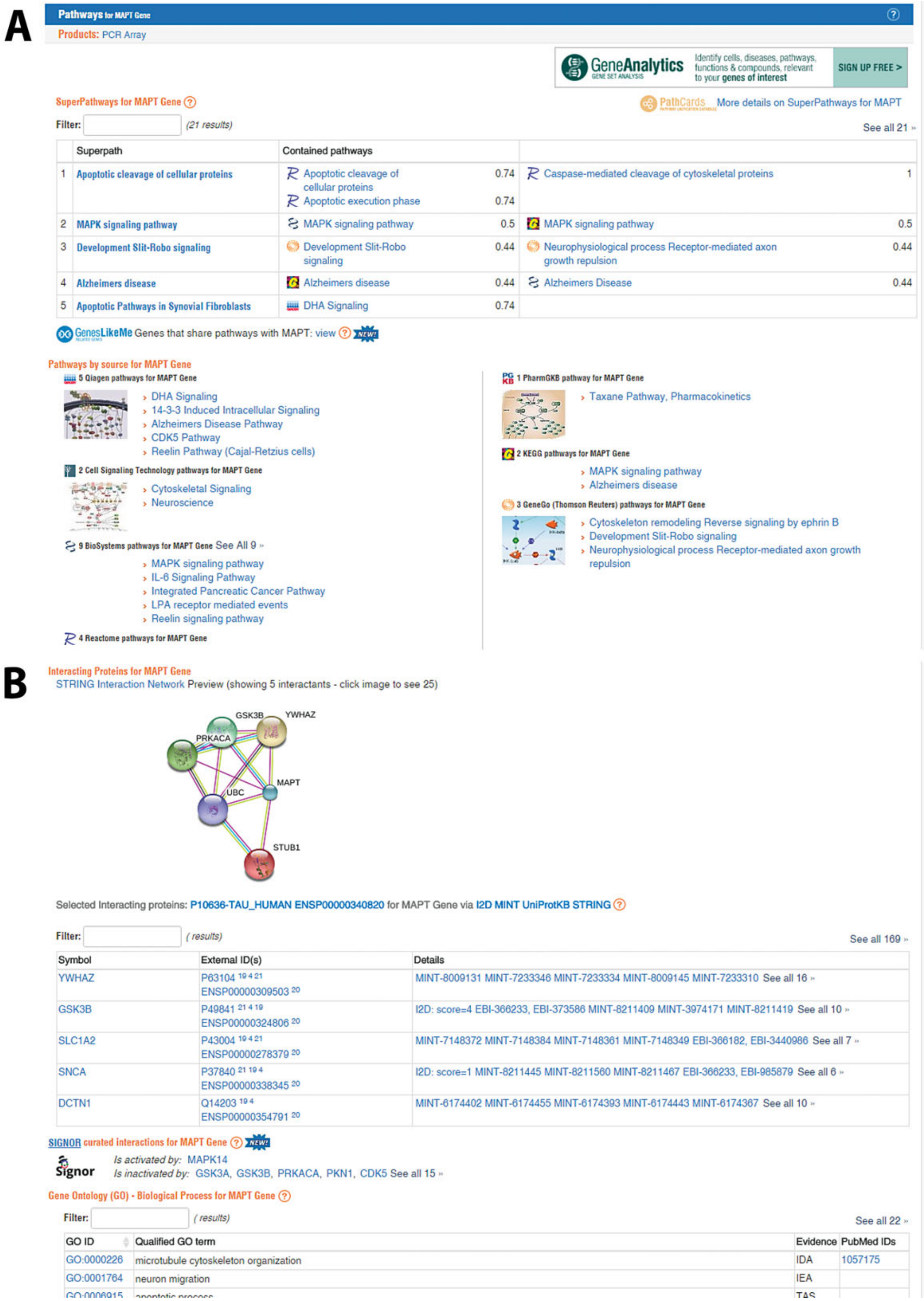
9. Go to the Localization section (Fig. 1.30.11).

   This section provides subcellular locations for the proteins associated with this gene, including a graphical representation from the COMPARTMENTS (http://compartments.jensenlab.org; Binder et al., 2014) database. There is also a table of Gene Ontology cellular compartments that includes GO IDs, GO terms, Evidence, and PubMed IDs. This table can be expanded to show all results and can be refined, via the 'Filter' text box above it, to display only the results of interest. A link to GenesLikeMe allows the user to see other genes that share this gene's ontologies.

10. Go to the Pathways and Interactions section (Fig. 1.30.12A).

    This section has the following subsections:

    a. *SuperPathways:* Displays a table listing clustered pathways (SuperPaths) from several sources (Belinky et al., 2015). Each row of the table lists the SuperPath and the pathways it contains, with source information and score indicating individual pathway similarity to the SuperPath. A link above the table allows the user to explore the SuperPaths at PathCards, a member of the GeneCards Suite. PathCards unifies 3,215 pathways from 12 sources into 1,073 SuperPaths, via judicious integration, reducing redundancy and optimizing the level of pathway-related informativeness for individual genes. This table can be expanded to show all results and can be refined, via the 'Filter' text box above it, to display only the results of interest. A link to GenesLikeMe allows the user to see other genes that share this gene's pathways. MAPT belongs to 21 biological SuperPaths, including several closely related to neurological processes, such as 'Alzheimers disease' and 'Neuroscience'.



Pathways for MAPT Gene

Products: PCR Array

GeneAnalytics GENE SET ANALYSIS Identify cells, diseases, pathways, functions & compounds, relevant to your **genes of interest**

SuperPathways for MAPT Gene

PathCards More details on SuperPathways for MAPT

Filter: ☐ *(21 results)*

See all 21 »

	Superpath	Contained pathways			
1	Apoptotic cleavage of cellular proteins	Apoptotic cleavage of cellular proteins Apoptotic execution phase	0.74 0.74	Caspase-mediated cleavage of cytoskeletal proteins	1
2	MAPK signaling pathway	MAPK signaling pathway	0.5	MAPK signaling pathway	0.5
3	Development Slit-Robo signaling	Development Slit-Robo signaling	0.44	Neurophysiological process Receptor-mediated axon growth repulsion	0.44
4	Alzheimers disease	Alzheimers disease	0.44	Alzheimers Disease	0.44
5	Apoptotic Pathways in Synovial Fibroblasts	DHA Signaling	0.74		

GenesLikeMe Genes that share pathways with MAPT: view

Pathways by source for MAPT Gene

5 Qiagen pathways for MAPT Gene

- DHA Signaling
- 14-3-3 Induced Intracellular Signaling
- Alzheimers Disease Pathway
- CDK5 Pathway
- Reelin Pathway (Cajal-Retzius cells)

2 Cell Signaling Technology pathways for MAPT Gene

- Cytoskeletal Signaling
- Neuroscience

9 BioSystems pathways for MAPT Gene See All 9 »

- MAPK signaling pathway
- IL-6 Signaling Pathway
- Integrated Pancreatic Cancer Pathway
- LPA receptor mediated events
- Reelin signaling pathway

4 Reactome pathways for MAPT Gene

1 PharmGKB pathway for MAPT Gene

- Taxane Pathway, Pharmacokinetics

2 KEGG pathways for MAPT Gene

- MAPK signaling pathway
- Alzheimers disease

3 GeneGo (Thomson Reuters) pathways for MAPT Gene

- Cytoskeleton remodeling Reverse signaling by ephrin B
- Development Slit-Robo signaling
- Neurophysiological process Receptor-mediated axon growth repulsion

Interacting Proteins for MAPT Gene

STRING Interaction Network Preview (showing 5 interactants - click image to see 25)

Selected Interacting proteins: P10636-TAU_HUMAN ENSP00000340820 for MAPT Gene via I2D MINT UniProtKB STRING

Filter: ☐ *(results)*

See all 169 »

Symbol	External ID(s)	Details
YWHAZ	P63104 $^{19\ 4\ 21}$ ENSP00000309503 ^{20}	MINT-8009131 MINT-7233346 MINT-7233334 MINT-8009145 MINT-7233310 See all 16 »
GSK3B	P49841 $^{21\ 4\ 19}$ ENSP00000324806 ^{20}	I2D: score=4 EBI-366233, EBI-373586 MINT-8211409 MINT-3974171 MINT-8211419 See all 10 »
SLC1A2	P43004 $^{19\ 4\ 21}$ ENSP00000278379 ^{20}	MINT-7148372 MINT-7148384 MINT-7148361 MINT-7148349 EBI-366182, EBI-3440986 See all 7 »
SNCA	P37840 $^{21\ 19\ 4}$ ENSP00000338345 ^{20}	I2D: score=1 MINT-8211445 MINT-8211560 MINT-8211467 EBI-366233, EBI-985879 See all 6 »
DCTN1	Q14203 $^{19\ 4}$ ENSP00000354791 ^{20}	MINT-6174402 MINT-6174455 MINT-6174393 MINT-6174443 MINT-6174367 See all 10 »

SIGNOR curated interactions for MAPT Gene

Signor *Is activated by:* MAPK14
Is inactivated by: GSK3A, GSK3B, PRKACA, PKN1, CDK5 See all 15 »

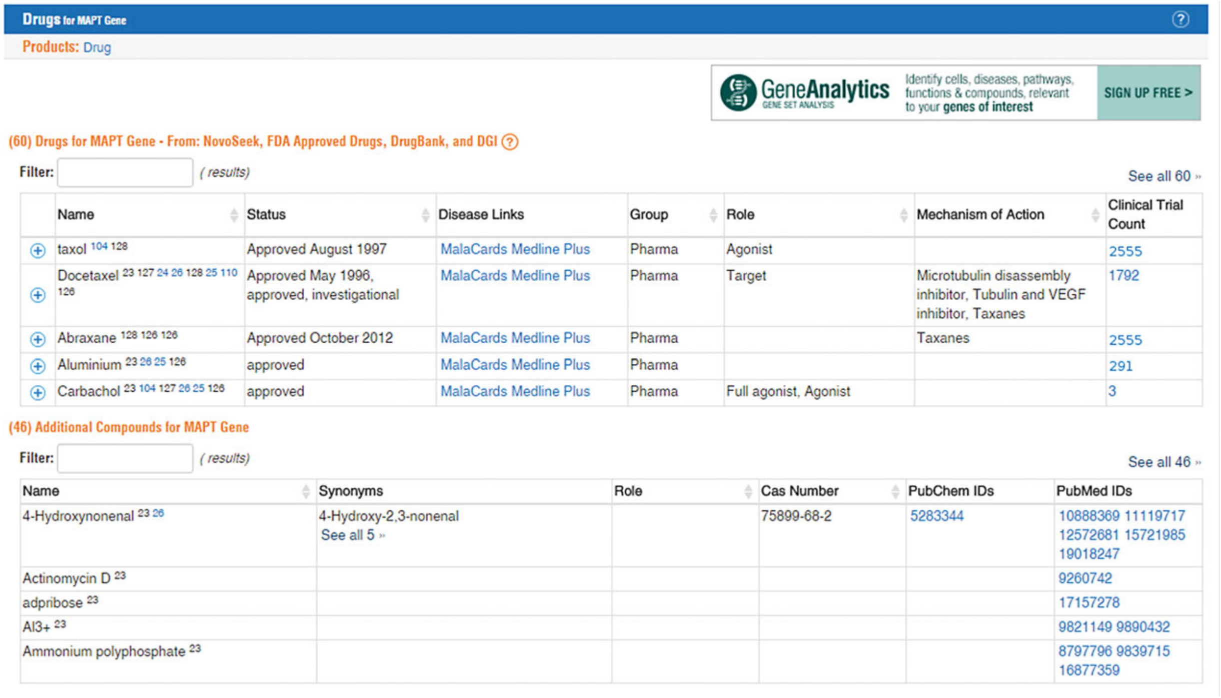
Gene Ontology (GO) - Biological Process for MAPT Gene

Filter: ☐ *(results)*

See all 22 »

GO ID	Qualified GO term	Evidence	PubMed IDs
GO:0000226	microtubule cytoskeleton organization	IDA	1057175
GO:0001764	neuron migration	IEA	
GO:0006915	apoptotic process	TAS	

Figure 1.30.12 (**A**) Portion of the GeneCard for MAPT displaying pathways in two subsections. The first depicts SuperPaths from PathCards in a table with the SuperPath name linked to PathCards in the first column, followed by a list of its member pathways, annotated with their sources and with scores depicting their overlap with the closest SuperPath neighbor, in the next two columns. Subsequently, the individual pathways are shown grouped by source. (**B**) Portion of the GeneCard for MAPT displaying interactions with this gene. A visual representation of the gene network, from STRING, is linked to a more comprehensive version of the figure.


GeneAnalytics
GENE SET ANALYSIS
Identify cells, diseases, pathways, functions & compounds, relevant to your genes of interest
SIGN UP FREE >

Drugs for MAPT Gene

Products: Drug

(60) Drugs for MAPT Gene - From: NovoSeek, FDA Approved Drugs, DrugBank, and DGI

Filter: (results)

See all 60 »

	Name	Status	Disease Links	Group	Role	Mechanism of Action	Clinical Trial Count
⊕	taxol [104 128]	Approved August 1997	MalaCards Medline Plus	Pharma	Agonist		2555
⊕	Docetaxel [23 127 24 26 128 25 110 126]	Approved May 1996, approved, investigational	MalaCards Medline Plus	Pharma	Target	Microtubulin disassembly inhibitor, Tubulin and VEGF inhibitor, Taxanes	1792
⊕	Abraxane [128 126 126]	Approved October 2012	MalaCards Medline Plus	Pharma		Taxanes	2555
⊕	Aluminium [23 26 25 126]	approved	MalaCards Medline Plus	Pharma			291
⊕	Carbachol [23 104 127 26 25 126]	approved	MalaCards Medline Plus	Pharma	Full agonist, Agonist		3

(46) Additional Compounds for MAPT Gene

Filter: (results)

See all 46 »

Name	Synonyms	Role	Cas Number	PubChem IDs	PubMed IDs
4-Hydroxynonenal [23 26]	4-Hydroxy-2,3-nonenal See all 5 »		75899-68-2	5283344	10888369 11119717 12572681 15721985 19018247
Actinomycin D [23]					9260742
adpribose [23]					17157278
Al3+ [23]					9821149 9890432
Ammonium polyphosphate [23]					8797796 9839715 16877359

Figure 1.30.13 Portion of the GeneCard for MAPT displaying drugs and other compounds related to this gene, mined from heterogeneous sources with names heuristically unified. Expandable tables organized by drug or compound give extensive information about each one at a glance.

b. *Pathways by source:* Lists pathways grouped by their source, with links to the pathways at their source's Web site. These lists initially show up to five pathways, with a link at the top to show all of the pathways for the desired source. Included among the pathways are PCR Array products.

c. *Interacting Proteins:* Shows a graphical representation from STRING (*http://string-db.org*) of a subset of the interaction network containing this gene (Fig. 1.30.12B), with a link to a diagram containing a more complex version of the network, with more interactants. It also shows a table listing interacting proteins. The table shows the gene symbol associated with each interacting protein, linked to its GeneCard; external IDs for the interacting proteins, linked to external databases; and links to the interactions at any of several protein-interaction Web sites. This table can be expanded to show all results and can be refined, via the 'Filter' text box above it, to display only the results of interest.

d. *SIGNOR-curated interactions:* Presents a deep link to the SIGnaling Network Open Resource (SIGNOR) (Perfetto et al., 2016), as well as a list of interacting genes, all linked to their GeneCards. The interactions are categorized as Activates, Inactivates, Is activated by, Is inactivated by, or Other effect.

e. *Gene Ontology Biological Processes:* Displays a table of Gene Ontology biological processes that includes GO IDs, GO terms, Evidence, and PubMed IDs for terms associated with the gene. This table can be expanded to show all results and can be refined, via the 'Filter' text box above it, to display only the results of interest. A link to GenesLikeMe allows the user to see other genes that share this gene's ontologies.

11. Go to the Drugs (and Compounds) section (Fig. 1.30.13).

This section has a unified table of drugs from high-quality sources, including DrugBank, ApexBio, DGIdb, FDA Approved Drugs, ClinicalTrials.gov, and PharmGKB. The table *is sorted by approval status, with drugs associated with the gene by the most curated sources taking precedence over those mapped via text mining. Following this table there is a unified ‘Additional Compounds’ table that displays compounds from IUPHAR, Novoseek, HMDB, Bitterdb, and Tocris that are not also found at the above drug sources. Each table can be expanded to show all results and can be refined, via the ‘Filter’ text box above it, to display only the results of interest. A link to GenesLikeMe allows the user to see other genes that share this gene’s drugs/compounds. The Drugs table lists the name of the drug, with superscripts corresponding to the sources that provide evidence of the gene-drug relationship. The superscripts are linked to the source database for all sources except Novoseek (whose data is frozen and no longer available externally). The tables further display the drug’s status (such as approved, experimental, investigational), links to associated diseases at MalaCards and Medline Plus, group (pharmaceutical or nutraceutical), role, mechanism of action, and the number of clinical trials. A plus in the first cell of each row allows the user to expand the display to show synonyms and accessions [such as chemical abstract (CAS) numbers, linked PubMed IDs, and linked PubChem IDs) for the drug and additional data when available, for example: active Ingredients, PharmGKB annotation, and a link to FDA Drug Label at DailyMed. The Additional Compounds table similarly displays the compound name, with linked superscripts, as well as its synonyms, role, CAS number, and linked PubChem and PubMed IDs. The bottom of the Compounds section provides compound-related product links.*

12. Go to the Transcripts section.

This section has the following subsections:

a. *mRNA/cDNA:* Displays RefSeq and other mRNA sequences from NCBI, cDNA sequences from AceView, and Ensembl transcripts, all with links to the source databases

b. *Unigene clusters:* Displays linked Unigene clusters for this gene, as well as their representative sequences.

c. *Product links:* Links to a variety of sequence-related products for this gene, such as inhibitory RNA and clone products.

d. *Alternative splicing:* Shows a schematic diagram of splice patterns using data from ASD (Alternative Splicing Database). Exons with alternative splice sites in different isoforms were broken into Exonic Units (ExUns). The letters indicate the order of the ExUns in the exon. The symbol ‘ ˆ ’ between ExUns indicates an intron, while ‘ ·’ indicates the junction of two ExUns. Mouse-overs on the dark blue squares show the Exun’s genomic coordinates, while mouse-overs on the light blue squares show its transcript coordinates. When showing ASD’s splice variants, GeneCards subtracts the 3000-bp flank that ASD adds to the transcript coordinates.

e. *External links:* Links to GeneLoc’s exon structure for this gene, and to alternative splicing isoforms from ECgene.

13. To see in which tissues the gene is expressed, go to the Expression section.

This section has the following subsections:

a. *mRNA expression graph:* Provides normal tissue expression profiles for the gene, via experimental results from BioGPS (Wu et al., 2009), GTEx (Lonsdale et al., 2013[]), Illumina Body Map, and SAGE. The *y* axis represents the tissues; the *x* axis represents the expression level. Each column shows the expression level of the gene in 37 tissues from each of the sources.

b. *mRNA expression in embryonic tissues and stem cells:* Shows the expression level of the gene in specific tissues and cells, based on data from LifeMap Discovery (*http://discovery.lifemapsc.com/*) (Edgar et al., 2013), a member of the GeneCards Suite, with links to the source database.

c. *Differential expression:* Shows tissue specificity of genes, annotated based on cross-tissue gene expression vectors, highlighting a subset of tissues that overexpress a gene. Differential expression was calculated based on proteome data from HIPED (Fishilevich et al., 2016) and transcriptome data from GTEx (Lonsdale et al., 2013), resulting in two textual lists of tissues differentially expressing a given gene.

d. *Protein expression graph:* Provides graphical data on the expression of the protein associated with this gene in 69 different tissues. The expression levels are integrated from ProteomicsDB, PaxDb, MOPED, and MaxQB. Each tissue name has a mouse-over listing the expression level and the contributing sources. Below this table are links to expression at SOURCE and UniProt. A link to GenesLikeMe allows the user to see other genes that share this gene's expression.

e. *Expression partners:* Displays genes similar to the current gene with respect to protein and/or RNA expression, with associated scores.

f. *Expression-related products:* Links to expression-related products, including primers and in situ assay products.

MAPT (Microtubule-Associated Protein Tau) is a well-studied gene, with known relation to the nervous system. The UniProt summary (Fig. 1.30.8) for MAPT states that this gene promotes microtubule assembly and stability, and might be involved in the establishment and maintenance of neuronal polarity. The GeneCards expression section for MAPT reveals a strong connection to nervous system components, as can be seen in the expression bars images both for RNA and for protein (Fig. 1.30.14A and B). The numeric data behind the images is also converted to textual annotations of overexpressed tissues, and is thus further available for the search engine.

14. Go to the Orthologs section.

This section provides a table of orthologs to this gene. Each row gives the species common name and scientific name, its class, the ortholog gene name(s), and the similarity. The table further lists the type of orthology from Ensembl, based on Ensembl gene trees, and links to the ortholog in other databases. Below the table are links to gene trees at Ensembl and TreeFam.

15. Go to the Paralogs section.

This section lists other human genes that are similar to this gene. The sequence paralogs are based on data from HomoloGene, Ensembl, and SIMAP, with Genes linked to their GeneCards and proteins linked to SIMAP. There are also pseudogenes from psudogene.org. A link to GenesLikeMe allows the user to see other genes that share this gene's paralogs.

16. Go to the Variants section (Fig. 1.30.15).

This section contains the following subsections:

a. *Polymorphic variants:* Displays polymorphic variant information from UniProt.

b. *Sequence variations:* Table of SNPs based on data from dbSNP and Humsavar. The table includes SNP ID, clinical significance, location, sequence, and type. It has further links to more information about the amino acid substitutions, as well as allele frequency information.

c. *Structural variations:* Table of structural variations from Database of Genomic Variants (DGV). The table includes variant ID, type, subtype, and linked PubMed IDs.

d. *Variation tolerance:* Link to Residual Variation Intolerance Score (RVIS) at the Genic Intolerance database as well as Gene Damage Index Score (GDIS), with percentile, interpretation (tolerant, intolerant, or mediocre tolerance), displayed inline. Link to Gene Damage Index Score, with score, percentile and interpretation (tolerant, intolerant etc.) displayed inline.

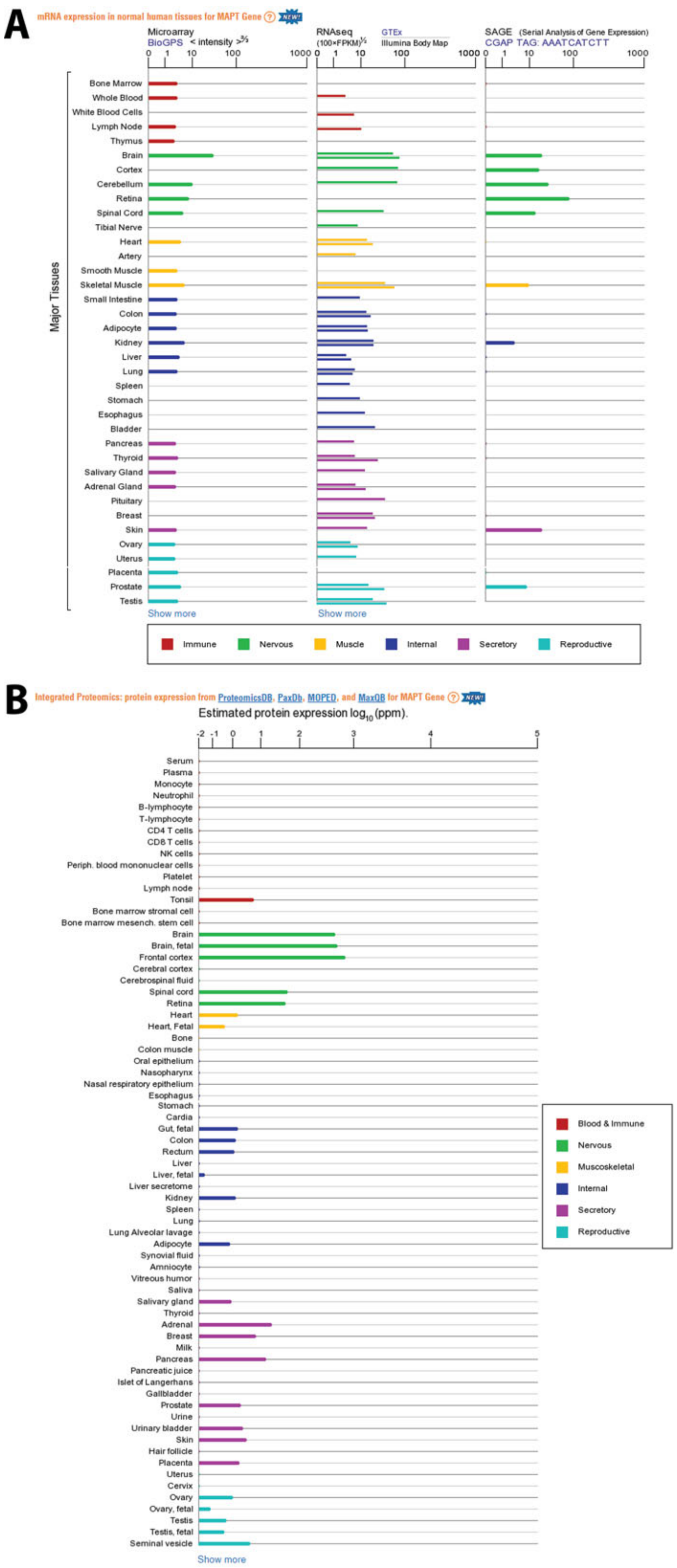
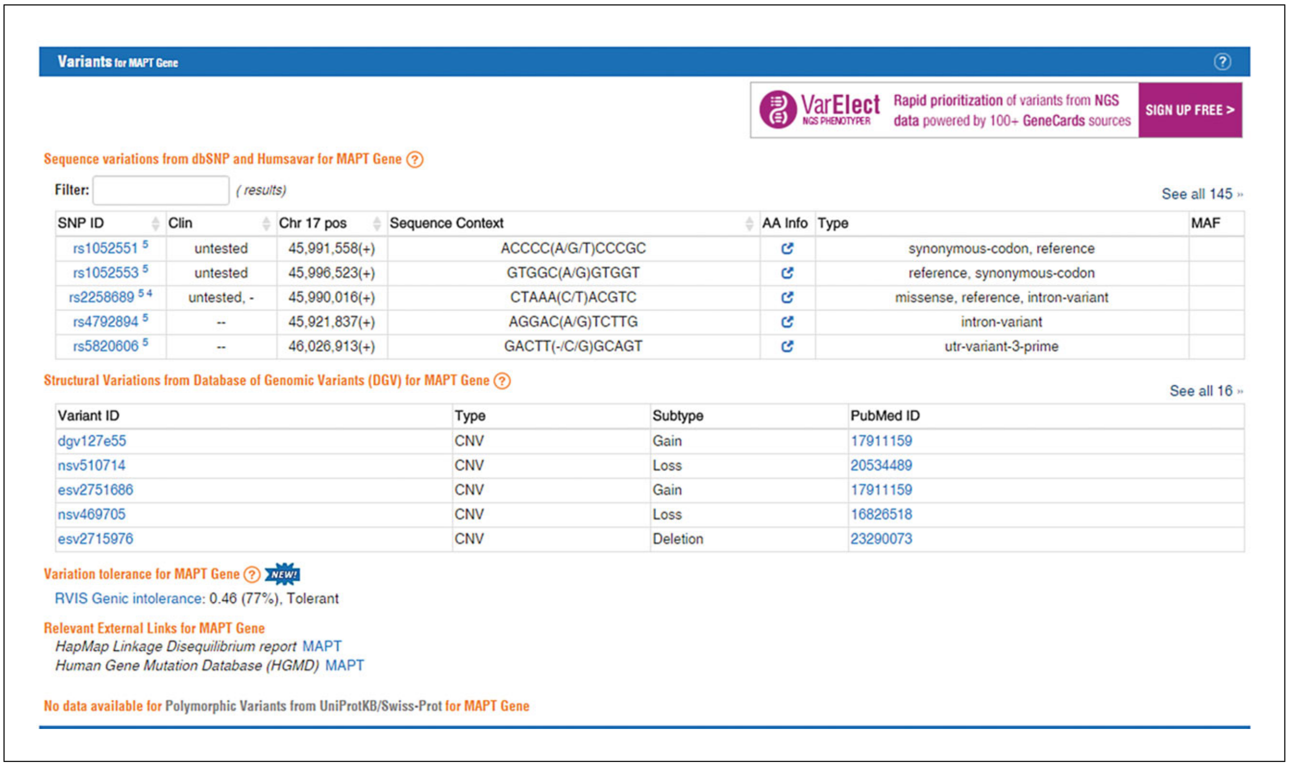


Figure 1.30.14 (**A**) RNA expression graph for MAPT. Color-coded bars indicate the expression level in different tissues, as reported by BioGPS, GTex, and SAGE. (**B**) Protein expression graph for MAPT. The protein expression is the integrated result of data from ProteomicsDB, PaxDb, MOPED, and MaxQB.



Sequence variations from dbSNP and Humsavar for MAPT Gene

SNP ID	Clin	Chr 17 pos	Sequence Context	AA Info	Type	MAF
rs1052551[5]	untested	45,991,558(+)	ACCCC(A/G/T)CCCGC	↗	synonymous-codon, reference	
rs1052553[5]	untested	45,996,523(+)	GTGGC(A/G)GTGGT	↗	reference, synonymous-codon	
rs2258689[5][4]	untested, -	45,990,016(+)	CTAAA(C/T)ACGTC	↗	missense, reference, intron-variant	
rs4792894[5]	--	45,921,837(+)	AGGAC(A/G)TCTTG	↗	intron-variant	
rs5820606[5]	--	46,026,913(+)	GACTT(-/C/G)GCAGT	↗	utr-variant-3-prime	

Structural Variations from Database of Genomic Variants (DGV) for MAPT Gene

Variant ID	Type	Subtype	PubMed ID
dgv127e55	CNV	Gain	17911159
nsv510714	CNV	Loss	20534489
esv2751686	CNV	Gain	17911159
nsv469705	CNV	Loss	16826518
esv2715976	CNV	Deletion	23290073

Variation tolerance for MAPT Gene NEW!

RVIS Genic intolerance: 0.46 (77%), Tolerant

Relevant External Links for MAPT Gene

HapMap Linkage Disequilibrium report MAPT

Human Gene Mutation Database (HGMD) MAPT

No data available for Polymorphic Variants from UniProtKB/Swiss-Prot for MAPT Gene

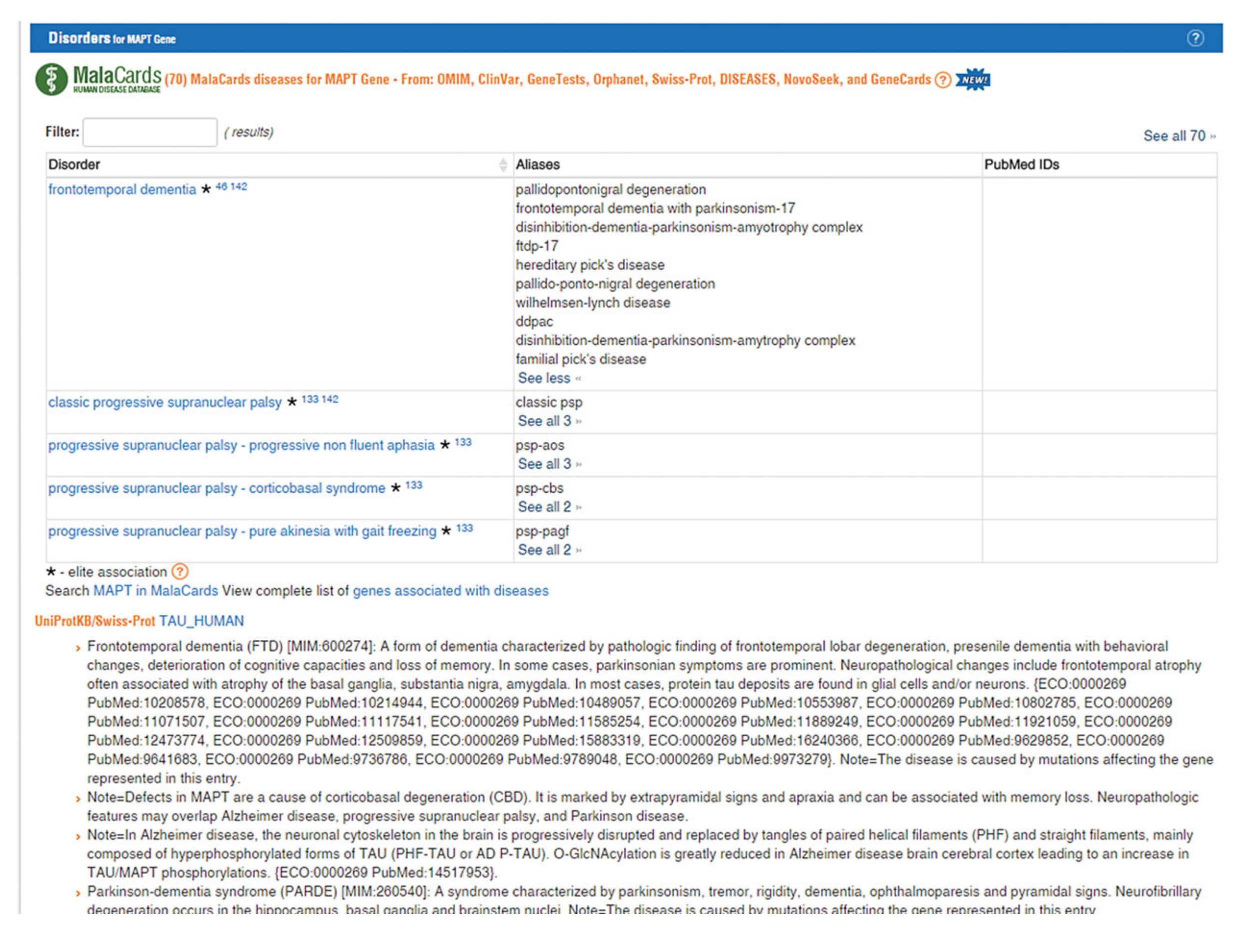
Figure 1.30.15 Portion of the GeneCard for MAPT displaying SNPs and other variants. This section displays variations in sequence and structure, as well as other variant-related data, such as variation tolerance.

e. *External links:* Links to linkage disequilibrium report and mutations from HapMap, Human Gene Mutation Database (HGMD), and Locus Specific Mutation Databases (LSDB).

17. Go to the Disorders section (Fig. 1.30.16).

This section provides a unified table of diseases associated with this gene by MalaCards, a member of the GeneCards Suite. The table, sorted by MalaCards gene-association score, shows the disease name, linked to MalaCards, with superscripts indicating the sources for the disease annotation and a mouse-over showing the MalaCards score. This score ranks diseases by how closely they are associated with the gene, factoring in the relative reliability of the sources that associate them. Elite associations are marked with an asterisk next to the disease name. The elite status is conferred when the gene-to-disease association is manually curated. Disease sources include OMIM, UniProt, GeneTests, ClinVar, Orphanet, the University of Copenhagen DISEASES database, Novoseek, and searches within GeneCards. The superscripts are linked to the source database for all sources except Novoseek (whose data is frozen and no longer available externally). The table further displays the most common alias for the disease, with a link to show all available aliases. A third column contains linked PubMed IDs associated with the disease. Below the disease table are links to MalaCards search results of the gene symbol, to the complete GeneCards disease genes list, to more in-depth information from UniProt and Genatlas, and to disease data at Atlas, GeneReviews, GAD, HuGENavigator, and TGDB. A link to GenesLikeMe allows the user to see other genes that share this gene's disorders. There is also an Export Disorders link that displays a search box that allows the user to query outside databases (PubMed, OMIM, and NCBI Bookshelf) for information on the gene. The query can be simple or complex and allows the user to select one or more aliases and/or one or more diseases and/or enter a search string before clicking the search button to submit the search to the selected database.

The majority of the 70 diseases related to MAPT via MalaCards are classified as neuronal, including Dementia, Pick disease, and Alzheimers disease.



Disorders for MAPT Gene

MalaCards HUMAN DISEASE DATABASE (70) MalaCards diseases for MAPT Gene - From: OMIM, ClinVar, GeneTests, Orphanet, Swiss-Prot, DISEASES, NovoSeek, and GeneCards ? NEW!

Filter: ☐ (results) See all 70 »

Disorder	Aliases	PubMed IDs
frontotemporal dementia ★ $^{46\ 142}$	pallidopontonigral degeneration frontotemporal dementia with parkinsonism-17 disinhibition-dementia-parkinsonism-amyotrophy complex ftdp-17 hereditary pick's disease pallido-ponto-nigral degeneration wilhelmsen-lynch disease ddpac disinhibition-dementia-parkinsonism-amytrophy complex familial pick's disease See less «	
classic progressive supranuclear palsy ★ $^{133\ 142}$	classic psp See all 3 »	
progressive supranuclear palsy - progressive non fluent aphasia ★ ^{133}	psp-aos See all 3 »	
progressive supranuclear palsy - corticobasal syndrome ★ ^{133}	psp-cbs See all 2 »	
progressive supranuclear palsy - pure akinesia with gait freezing ★ ^{133}	psp-pagf See all 2 »	

★ - elite association ?

Search MAPT in MalaCards View complete list of genes associated with diseases

UniProtKB/Swiss-Prot TAU_HUMAN

- Frontotemporal dementia (FTD) [MIM:600274]: A form of dementia characterized by pathologic finding of frontotemporal lobar degeneration, presenile dementia with behavioral changes, deterioration of cognitive capacities and loss of memory. In some cases, parkinsonian symptoms are prominent. Neuropathological changes include frontotemporal atrophy often associated with atrophy of the basal ganglia, substantia nigra, amygdala. In most cases, protein tau deposits are found in glial cells and/or neurons. {ECO:0000269 PubMed:10208578, ECO:0000269 PubMed:10214944, ECO:0000269 PubMed:10489057, ECO:0000269 PubMed:10553987, ECO:0000269 PubMed:10802785, ECO:0000269 PubMed:11071507, ECO:0000269 PubMed:11117541, ECO:0000269 PubMed:11585254, ECO:0000269 PubMed:11889249, ECO:0000269 PubMed:11921059, ECO:0000269 PubMed:12473774, ECO:0000269 PubMed:12509859, ECO:0000269 PubMed:15883319, ECO:0000269 PubMed:16240366, ECO:0000269 PubMed:9629852, ECO:0000269 PubMed:9641683, ECO:0000269 PubMed:9736786, ECO:0000269 PubMed:9789048, ECO:0000269 PubMed:9973279}. Note=The disease is caused by mutations affecting the gene represented in this entry.
- Note=Defects in MAPT are a cause of corticobasal degeneration (CBD). It is marked by extrapyramidal signs and apraxia and can be associated with memory loss. Neuropathologic features may overlap Alzheimer disease, progressive supranuclear palsy, and Parkinson disease.
- Note=In Alzheimer disease, the neuronal cytoskeleton in the brain is progressively disrupted and replaced by tangles of paired helical filaments (PHF) and straight filaments, mainly composed of hyperphosphorylated forms of TAU (PHF-TAU or AD P-TAU). O-GlcNAcylation is greatly reduced in Alzheimer disease brain cerebral cortex leading to an increase in TAU/MAPT phosphorylations. {ECO:0000269 PubMed:14517953}.
- Parkinson-dementia syndrome (PARDE) [MIM:260540]: A syndrome characterized by parkinsonism, tremor, rigidity, dementia, ophthalmoparesis and pyramidal signs. Neurofibrillary degeneration occurs in the hippocampus, basal ganglia and brainstem nuclei. Note=The disease is caused by mutations affecting the gene represented in this entry.

Figure 1.30.16 Portion of the GeneCard for MAPT displaying MalaCards diseases associated with this gene. The disease table provides a bird's-eye view of each disease and its data. * marks elite associations between the gene and disease.

18. Go to the Publications section.

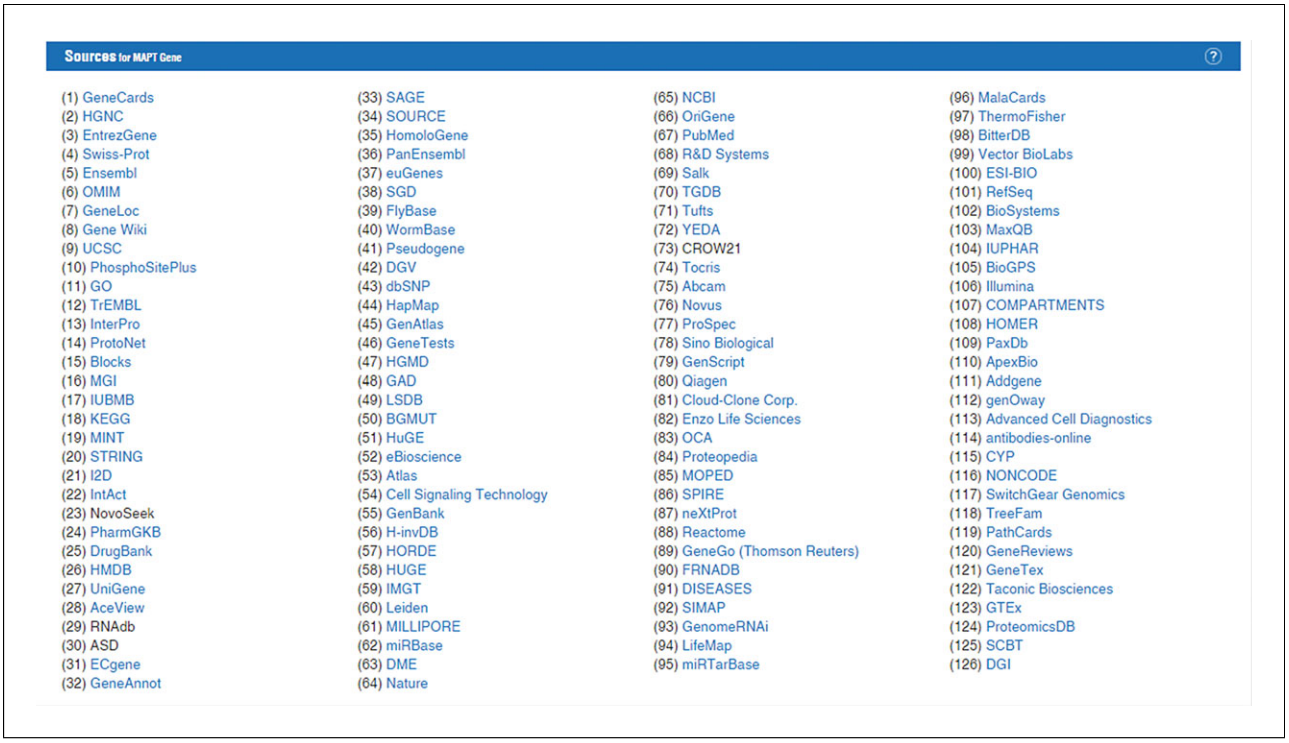
    This section lists the publications associated with this gene. For each article, the title, linked PubMed ID, first and last authors, journal name, and publication date are displayed, with superscripts indicating the sources that associate the gene with the publication. The publications are ranked by number of sources citing them in connection with this gene, and then by publication date (newest first). This list can be expanded to show all results and can be refined, via the 'Filter' text box above it, to display only the results of interest.

19. Go to the Products section.

    The Products section displays links to the home pages of all commercial companies supplying products within the sections above. Also, each commercial product, e.g., proteins or antibodies, is linked to a page listing products for the relevant gene.

20. Go to the Sources section (Fig. 1.30.17).

    This section provides numbered links to all of the GeneCards sources. The numbers are used as superscripts in data throughout the card, thereby accrediting the course of the cited information. A list of the sources with a short description for each can be found in http://www.genecards.org/Guide/Sources.



Sources for MAPT Gene

(1) GeneCards
(2) HGNC
(3) EntrezGene
(4) Swiss-Prot
(5) Ensembl
(6) OMIM
(7) GeneLoc
(8) Gene Wiki
(9) UCSC
(10) PhosphoSitePlus
(11) GO
(12) TrEMBL
(13) InterPro
(14) ProtoNet
(15) Blocks
(16) MGI
(17) IUBMB
(18) KEGG
(19) MINT
(20) STRING
(21) I2D
(22) IntAct
(23) NovoSeek
(24) PharmGKB
(25) DrugBank
(26) HMDB
(27) UniGene
(28) AceView
(29) RNAdb
(30) ASD
(31) ECgene
(32) GeneAnnot
(33) SAGE
(34) SOURCE
(35) HomoloGene
(36) PanEnsembl
(37) euGenes
(38) SGD
(39) FlyBase
(40) WormBase
(41) Pseudogene
(42) DGV
(43) dbSNP
(44) HapMap
(45) GenAtlas
(46) GeneTests
(47) HGMD
(48) GAD
(49) LSDB
(50) BGMUT
(51) HuGE
(52) eBioscience
(53) Atlas
(54) Cell Signaling Technology
(55) GenBank
(56) H-invDB
(57) HORDE
(58) HUGE
(59) IMGT
(60) Leiden
(61) MILLIPORE
(62) miRBase
(63) DME
(64) Nature
(65) NCBI
(66) OriGene
(67) PubMed
(68) R&D Systems
(69) Salk
(70) TGDB
(71) Tufts
(72) YEDA
(73) CROW21
(74) Tocris
(75) Abcam
(76) Novus
(77) ProSpec
(78) Sino Biological
(79) GenScript
(80) Qiagen
(81) Cloud-Clone Corp.
(82) Enzo Life Sciences
(83) OCA
(84) Proteopedia
(85) MOPED
(86) SPIRE
(87) neXtProt
(88) Reactome
(89) GeneGo (Thomson Reuters)
(90) FRNADB
(91) DISEASES
(92) SIMAP
(93) GenomeRNAi
(94) LifeMap
(95) miRTarBase
(96) MalaCards
(97) ThermoFisher
(98) BitterDB
(99) Vector BioLabs
(100) ESI-BIO
(101) RefSeq
(102) BioSystems
(103) MaxQB
(104) IUPHAR
(105) BioGPS
(106) Illumina
(107) COMPARTMENTS
(108) HOMER
(109) PaxDb
(110) ApexBio
(111) Addgene
(112) genOway
(113) Advanced Cell Diagnostics
(114) antibodies-online
(115) CYP
(116) NONCODE
(117) SwitchGear Genomics
(118) TreeFam
(119) PathCards
(120) GeneReviews
(121) GeneTex
(122) Taconic Biosciences
(123) GTEx
(124) ProteomicsDB
(125) SCBT
(126) DGI

Figure 1.30.17 GeneCards' 126 sources, linked to their Web sites. Numbers in parentheses correspond with superscripts used in the body of the GeneCard.

BASIC PROTOCOL 3

USING VarElect

VarElect may be accessed on the Internet via *https://ve.genecards.org/*. The homepage (Fig. 1.30.18) is the entry point for running new analyses and for re-analyzing saved project data with new parameters using the "My Projects" feature. The VarElect input page is simple in its design, displaying two text boxes that serve as the starting point. The first requirement is for a list of GeneCards gene symbols, such as those derived from NGS experiments and related to variants from exome sequencing, or differentially expressed genes from RNAseq (or other OMICs derived gene lists). The second compulsory input is a phenotype phrase expected to be found in connection with the analyzed genes. VarElect uses free-text searches for keywords appearing in any of its data sources, such as the Human Phenotype Ontology (HPO), and is not limited to a closed list of phenotypes. The top of the homepage displays two links for examples (diarrhea, "capillary leak"); when clicked, the necessary input in their designated text boxes are filled in for the user who can then peruse the results (Stelzer et al., in press).

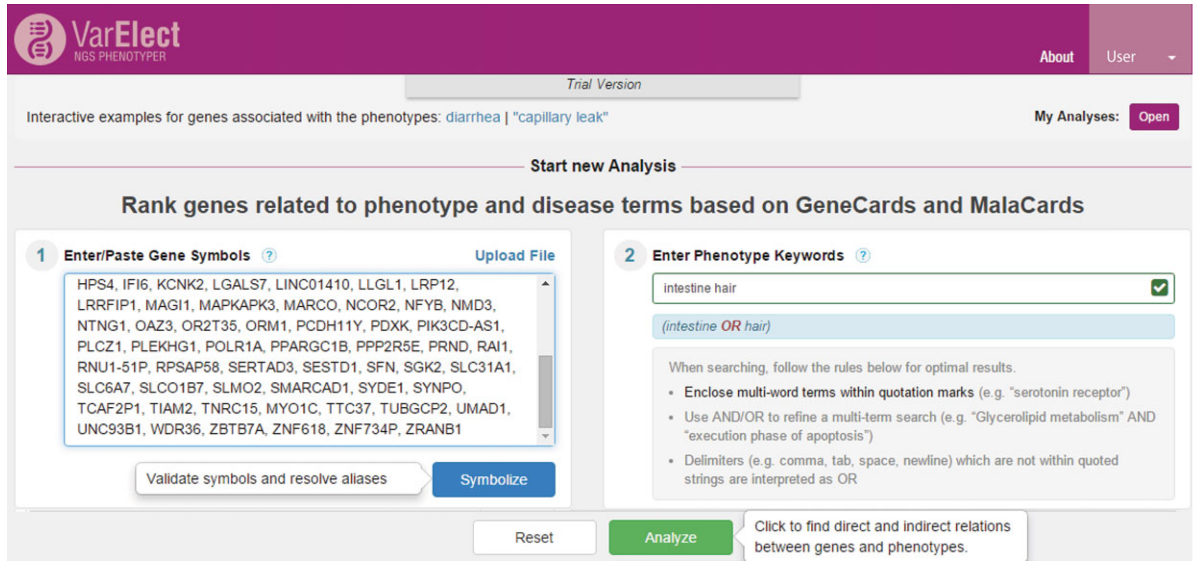
Necessary Resources

An up-to-date Web browser such as Google Chrome, Mozilla Firefox, Microsoft Edge, or Apple Safari

Gene list symbolization

1. Begin at the VarElect homepage (*https://ve.genecards.org/*).
2. In the Enter/Paste Gene Symbols text box, either paste the following gene list or upload it using the Upload File link (Fig. 1.30.18):

   ACTC1, ADAM17, ADCYAP1, ADPRHL1, AFM, AGAP7P, ANK1, ANKRD33B, ANKRD7, ANO8, ARHGAP18, ASAH2, ATMIN, ATP1B1, ATXN1, BRD3, BRSK2, C20orf26, CAMKK2, CD2AP, CDH26, CLEC18C, CLSPN, CNGA1, COL13A1, COPS3, CREB5, CTBP2, CUL5, DBF4, DHX8, EIF3E, EML4,



Interactive examples for genes associated with the phenotypes: diarrhea | "capillary leak"

Rank genes related to phenotype and disease terms based on GeneCards and MalaCards

HPS4, IFI6, KCNK2, LGALS7, LINC01410, LLGL1, LRP12, LRRFIP1, MAGI1, MAPKAPK3, MARCO, NCOR2, NFYB, NMD3, NTNG1, OAZ3, OR2T35, ORM1, PCDH11Y, PDXK, PIK3CD-AS1, PLCZ1, PLEKHG1, POLR1A, PPARGC1B, PPP2R5E, PRND, RAI1, RNU1-51P, RPSAP58, SERTAD3, SESTD1, SFN, SGK2, SLC31A1, SLC6A7, SLCO1B7, SLMO2, SMARCAD1, SYDE1, SYNPO, TCAF2P1, TIAM2, TNRC15, MYO1C, TTC37, TUBGCP2, UMAD1, UNC93B1, WDR36, ZBTB7A, ZNF618, ZNF734P, ZRANB1

intestine hair ☑

(intestine OR hair)

When searching, follow the rules below for optimal results.

- Enclose multi-word terms within quotation marks (e.g. "serotonin receptor")
- Use AND/OR to refine a multi-term search (e.g. "Glycerolipid metabolism" AND "execution phase of apoptosis")
- Delimiters (e.g. comma, tab, space, newline) which are not within quoted strings are interpreted as OR

Click to find direct and indirect relations between genes and phenotypes.

Figure 1.30.18 VarElect homepage with example input. A gene list and a phenotype phrase are entered into the Gene Symbols and Phenotype Keywords text boxes. Pressing the Symbolize button starts the symbol validation and aliases resolution processes. Phenotype phrase syntax is verified (indicated by the green check) and its interpretation is displayed in the blue box underneath the phenotype box. Pressing the Analyze button initiates analysis of the input, and Reset clears all fields. Example cases (diarrhea and "capillary leak") appear in the top of the homepage, and access to "My analyses" can be achieved via the Open button. More information about input may be viewed using the question mark icons.

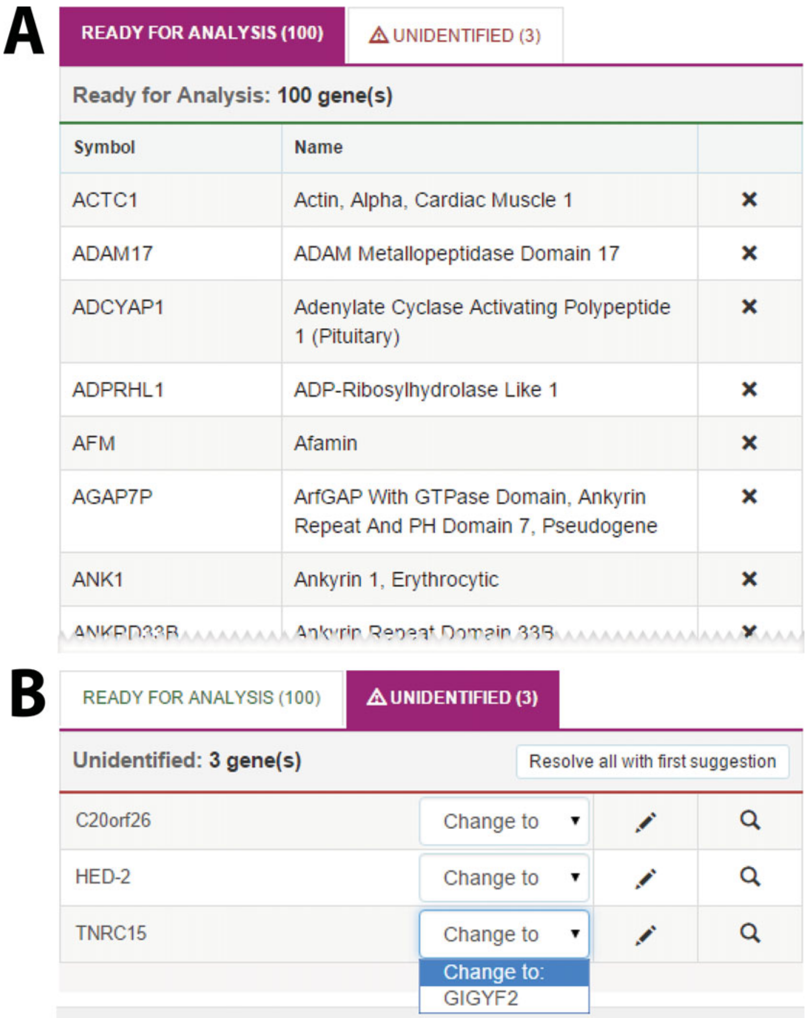
EP300, EPHA6, FEM1A, FOXD4L3, GABRB3, GADL1, GDF7, GIGYF1, GLG1, GLRA2, GOSR1, GRK1, H3F3A, HED-2, HLA-DQB1, HMCN1, HPS4, IFI6, KCNK2, LGALS7, LINC01410, LLGL1, LRP12, LRRFIP1, MAGI1, MAPKAPK3, MARCO, NCOR2, NFYB, NMD3, NTNG1, OAZ3, OR2T35, ORM1, PCDH11Y, PDXK, PIK3CD-AS1, PLCZ1, PLEKHG1, POLR1A, PPARGC1B, PPP2R5E, PRND, RAI1, RNU1-51P, RPSAP58, SERTAD3, SESTD1, SFN, SGK2, SLC31A1, SLC6A7, SLCO1B7, SLMO2, SMARCAD1, SYDE1, SYNPO, TCAF2P1, TIAM2, TNRC15, MYO1C, TTC37, TUBGCP2, UMAD1, UNC93B1, WDR36, ZBTB7A, ZNF618, ZNF734P, ZRANB1.

The list may be delimited by commas or any white space (spaces, tabs, new lines, etc.), and contain at most 4000 symbols.

The genes listed above were generated from the exome sequencing of a child with intestinal and hair abnormalities.

3. Click the "Symbolize" button to initiate the symbolization process. All validated symbols will appear in the Ready for Analysis tab (Fig. 1.30.19A) and any aliases or identifiers that need to be resolved will appear in the Unidentified tab (Fig. 1.30.19B), which contains several features that help with the identification resolution. Any of the validated or resolved symbols may be removed from the Ready for Analysis list using the "x" button available in each row.

4. For aliases that are mapped to several official gene symbols, one can disambiguate them by selecting the desired one from the "Change to" drop-down menu (Fig. 1.30.19B). The genes in this menu appear according to their GeneCards symbol/alias search score in decreasing order. The magnifying glass icon launches the GeneCards search in symbol/alias mode, with the original gene name as input, in order to provide additional information for deciding which of the suggested symbols should be used or if none of them seem correct. The pencil icon may be used for



Ready for Analysis: **100 gene(s)**

Symbol	Name	
ACTC1	Actin, Alpha, Cardiac Muscle 1	✖
ADAM17	ADAM Metallopeptidase Domain 17	✖
ADCYAP1	Adenylate Cyclase Activating Polypeptide 1 (Pituitary)	✖
ADPRHL1	ADP-Ribosylhydrolase Like 1	✖
AFM	Afamin	✖
AGAP7P	ArfGAP With GTPase Domain, Ankyrin Repeat And PH Domain 7, Pseudogene	✖
ANK1	Ankyrin 1, Erythrocytic	✖
ANKRD33B	Ankyrin Repeat Domain 33B	✖

Unidentified: **3 gene(s)**

C20orf26	Change to ▾	✎	🔍
HED-2	Change to ▾	✎	🔍
TNRC15	Change to ▾	✎	🔍

Figure 1.30.19 Symbolization of genes. (**A**) Input that was verified as official symbols appear in the Ready for Analysis tab. Gene symbols may be removed from the list using the “x” icons. (**B**) Identifiers that were not recognized as official symbols may be changed to suggested symbols using the “Change to” drop-down menu or manually typed in using the pencil icon. Magnifying glass icons initiate and open a GeneCards search page to aid in the resolution of unidentified supplied genes.

editing the gene name entered. Since the list of unresolved gene names may be extensive due to heterogeneous naming conventions implemented in various tools whose output serve as input for VarElect, the “Resolve all with first suggestion” button simplifies this procedure for those that can be mapped to at least one gene symbol.

Phenotype phrase validation

5. In the Enter Phenotype Keywords text box, enter (Fig. 1.30.18) `intestine hair`.

VarElect supports Boolean searches, so `AND/OR` may be used to refine a search. Searches aimed at finding gene connections to a phenotype with several synonyms or relevant keywords (e.g., `low stature`, `dwarfism`, `nanism`) should be searched using terms separated by `OR`. However, if several terms describe a single phenomenon or syndrome (e.g., microcephaly, deafness and intellectual disability), several keywords may be used in conjunction by placing an `AND` between them. Terms delimited by white space (spaces, tabs, new lines, etc.) are searched using

`OR` logic. Thus, the phenotype phrase listed above will be analyzed as `intestine OR hair`. Multi-string terms should be enclosed in quotation marks for them to be searched verbatim, e.g., `''serotonin receptor''`. Syntactical analysis of the phenotype phrase, which checks the Boolean logic, quotation marks, parentheses, etc., occurs in real time, and a verified search expression is indicated by a green check mark. The Boolean logic interpreted from the input phenotype phrase appears in the "Query output . . . " blue text box (Fig. 1.30.18).

6. Analyzing the data can commence only after supplying both verified (symbolized) input genes and a phenotype keyword phrase, which will enable the Analyze button.

7. Click the Analyze button to invoke VarElect's analysis of our example's gene list against `intestine OR hair`.

Scrutinizing VarElect direct results

8. The resulting scored phenotype associations table presents separate lists of genes according to their association type. VarElect first finds genes in the input gene symbols list that are directly related to the input phenotypes and displays those "hits" in the Directly Related tab. Running the aforementioned example (detailed in items 2 and 5) in February 2016 returned a list of 19 directly related genes (Fig. 1.30.20A).

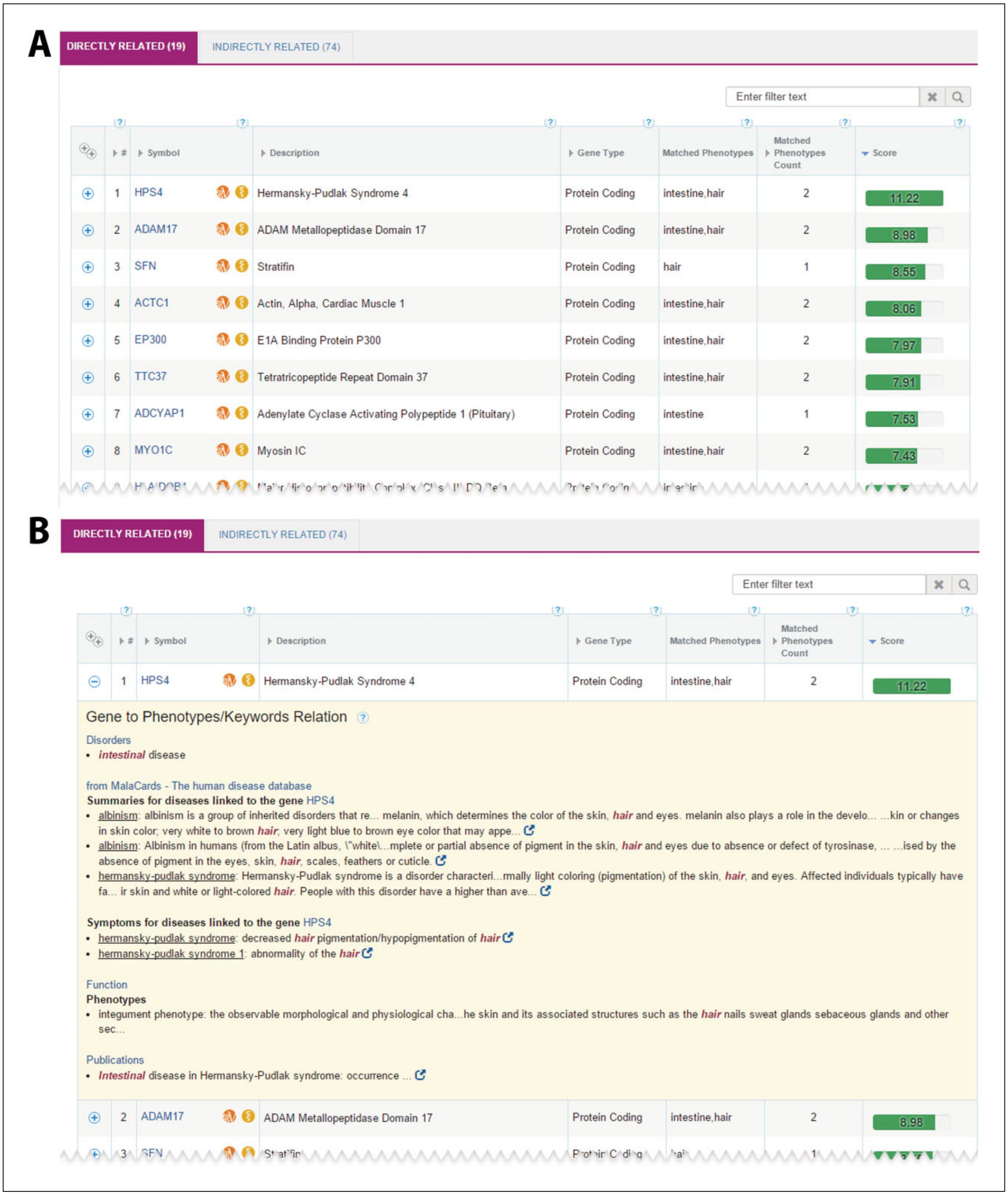
   The results reflect phenotype connections within a search space limited to a gene list rather than the entire gene landscape. Therefore, few genes may be directly related, but additional gene-to-phenotype connections may appear in the indirectly related genes tab (see item 9). The gene results with clear phenotype associations have a similar structure to those in the GeneCards search, with two additions. The first appears in the score column, where the relative size of the green bar in the background of the score indicates the strength of a gene-phenotype connection compared to the top-scoring gene in the same search. Second, in the cases of multi-term phenotype phrases, two additional columns are provided to indicate the count and identity of the phenotypes that are associated with each gene in the list. The results table may be sorted by any of its columns as well as filtered, thereby giving maximum flexibility for judicious analysis of the data.

9. Click the + icon in the first row to open the hit context information for HPS4, in "minicard" format, to the top-scoring gene (Fig. 1.30.20B). This gene has connections with intestine and hair via the publications and function sections of GeneCards, respectively. Since the search mechanism directly interrogates MalaCards data, its summaries and detailed symptoms increase the likelihood of establishing gene-disease connections, as in this case, in addition to relationships based on disease name or synonyms. Moreover, differential expression information for tens of tissues and cell types at both RNA and protein levels may attest to gene up-regulation in tissues of interest. Search hits are equipped with a link to the external source evidence for further scrutiny.

Scrutinizing VarElect indirect results

10. The symbols of the genes that were not found to be directly related to the input phenotypes are queried to determine indirect connections to the phenotypes via intermediate genes, and may be viewed in the Indirectly Related tab. The aforementioned (detailed in items 2 and 5) in February 2016 returned a list of 74 indirectly related genes (Fig. 1.30.21A).

11. Click the + icon in the first row to view up to five implicating genes for the top-scoring gene LRRFIP1 with the entered phenotype phrase (Fig. 1.30.21B, yellow background). This gene, carrying a suspected variant as determined by Next Generation Sequencing, which is not directly connected to the phenotype of interest, shares a pathway with an implicating gene which is directly connected to the phenotype;



	#	Symbol	Description	Gene Type	Matched Phenotypes	Matched Phenotypes Count	Score
⊕	1	HPS4	Hermansky-Pudlak Syndrome 4	Protein Coding	intestine,hair	2	11.22
⊕	2	ADAM17	ADAM Metallopeptidase Domain 17	Protein Coding	intestine,hair	2	8.98
⊕	3	SFN	Stratifin	Protein Coding	hair	1	8.55
⊕	4	ACTC1	Actin, Alpha, Cardiac Muscle 1	Protein Coding	intestine,hair	2	8.06
⊕	5	EP300	E1A Binding Protein P300	Protein Coding	intestine,hair	2	7.97
⊕	6	TTC37	Tetratricopeptide Repeat Domain 37	Protein Coding	intestine,hair	2	7.91
⊕	7	ADCYAP1	Adenylate Cyclase Activating Polypeptide 1 (Pituitary)	Protein Coding	intestine	1	7.53
⊕	8	MYO1C	Myosin IC	Protein Coding	intestine,hair	2	7.43

	#	Symbol	Description	Gene Type	Matched Phenotypes	Matched Phenotypes Count	Score
⊖	1	HPS4	Hermansky-Pudlak Syndrome 4	Protein Coding	intestine,hair	2	11.22

Gene to Phenotypes/Keywords Relation

Disorders
- ***intestinal*** disease

from MalaCards - The human disease database
Summaries for diseases linked to the gene HPS4
- albinism: albinism is a group of inherited disorders that re... melanin, which determines the color of the skin, ***hair*** and eyes. melanin also plays a role in the develo... ...kin or changes in skin color; very white to brown ***hair***, very light blue to brown eye color that may appe...
- albinism: Albinism in humans (from the Latin albus, \"white\...mplete or partial absence of pigment in the skin, ***hair*** and eyes due to absence or defect of tyrosinase, ...ised by the absence of pigment in the eyes, skin, ***hair***, scales, feathers or cuticle.
- hermansky-pudlak syndrome: Hermansky-Pudlak syndrome is a disorder characteri...rmally light coloring (pigmentation) of the skin, ***hair***, and eyes. Affected individuals typically have fa... ir skin and white or light-colored ***hair***. People with this disorder have a higher than ave...

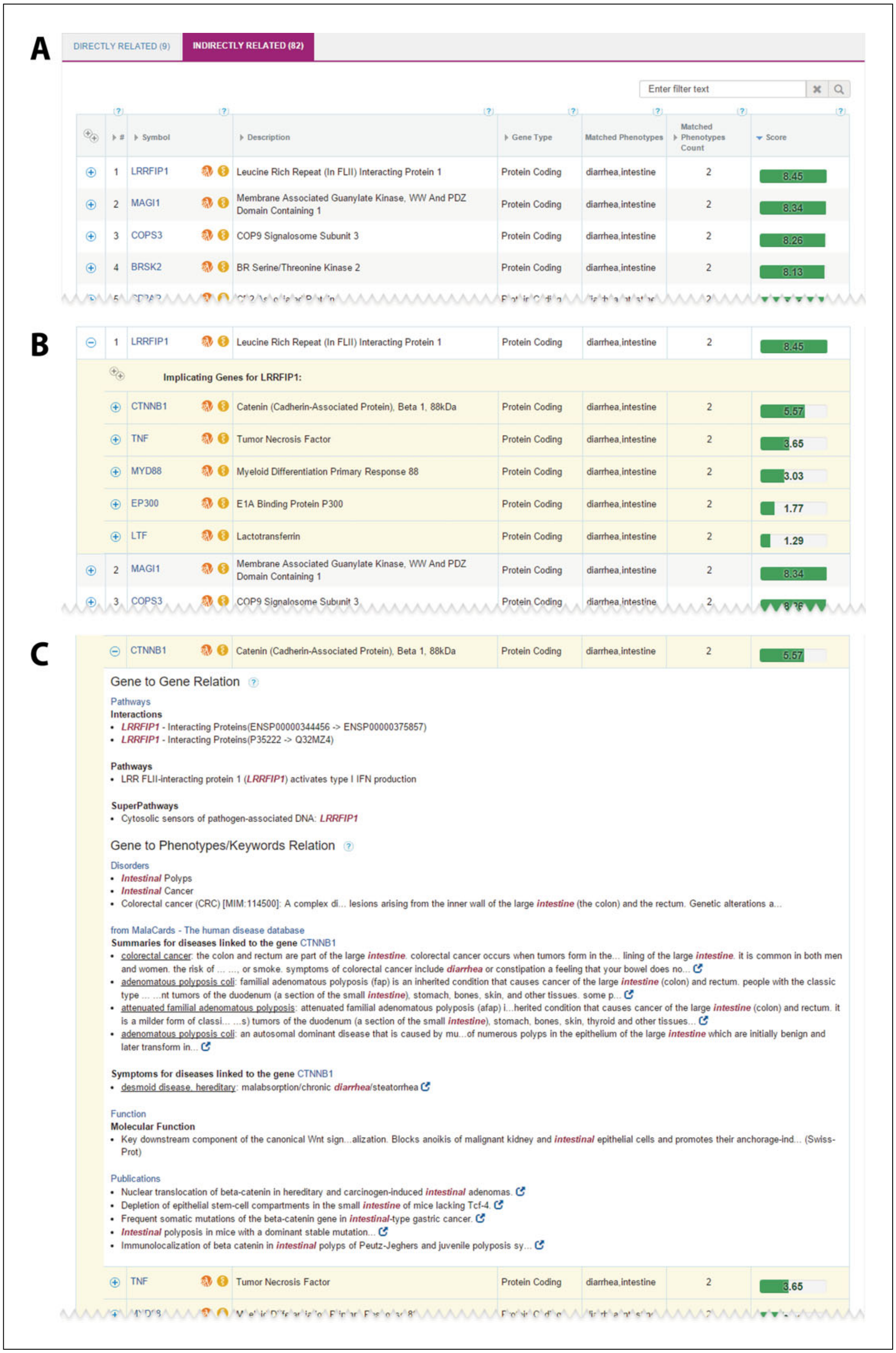
Symptoms for diseases linked to the gene HPS4
- hermansky-pudlak syndrome: decreased ***hair*** pigmentation/hypopigmentation of ***hair***
- hermansky-pudlak syndrome 1: abnormality of the ***hair***

Function
Phenotypes
- integument phenotype: the observable morphological and physiological cha...he skin and its associated structures such as the ***hair*** nails sweat glands sebaceous glands and other sec...

Publications
- ***Intestinal*** disease in Hermansky-Pudlak syndrome: occurrence ...

	#	Symbol	Description	Gene Type	Matched Phenotypes	Matched Phenotypes Count	Score
⊕	2	ADAM17	ADAM Metallopeptidase Domain 17	Protein Coding	intestine,hair	2	8.98

Figure 1.30.20 VarElect direct results. (**A**) The directly related tab displays genes with direct connection to phenotype keywords. Genes are sorted by relevance scores by default, and may be reordered using any column in the table, such as Matched Phenotype Count. Each gene links to GeneCards or MalaCards (orange and yellow icons in the symbol column). The table may be filtered by entering keywords in the filter box. (**B**) Each row may be expanded using the + icon to display a minicard with evidence associating the gene with phenotype keywords. Clicking the external links enables further scrutiny of evidence in its reporting data source.



	#	Symbol	Description	Gene Type	Matched Phenotypes	Matched Phenotypes Count	Score
⊕	1	LRRFIP1	Leucine Rich Repeat (In FLII) Interacting Protein 1	Protein Coding	diarrhea,intestine	2	8.45
⊕	2	MAGI1	Membrane Associated Guanylate Kinase, WW And PDZ Domain Containing 1	Protein Coding	diarrhea,intestine	2	8.34
⊕	3	COPS3	COP9 Signalosome Subunit 3	Protein Coding	diarrhea,intestine	2	8.26
⊕	4	BRSK2	BR Serine/Threonine Kinase 2	Protein Coding	diarrhea,intestine	2	8.13

	#	Symbol	Description	Gene Type	Matched Phenotypes	Matched Phenotypes Count	Score
⊖	1	LRRFIP1	Leucine Rich Repeat (In FLII) Interacting Protein 1	Protein Coding	diarrhea,intestine	2	8.45
		Implicating Genes for LRRFIP1:					
	⊕	CTNNB1	Catenin (Cadherin-Associated Protein), Beta 1, 88kDa	Protein Coding	diarrhea,intestine	2	5.57
	⊕	TNF	Tumor Necrosis Factor	Protein Coding	diarrhea,intestine	2	3.65
	⊕	MYD88	Myeloid Differentiation Primary Response 88	Protein Coding	diarrhea,intestine	2	3.03
	⊕	EP300	E1A Binding Protein P300	Protein Coding	diarrhea,intestine	2	1.77
	⊕	LTF	Lactotransferrin	Protein Coding	diarrhea,intestine	2	1.29
⊕	2	MAGI1	Membrane Associated Guanylate Kinase, WW And PDZ Domain Containing 1	Protein Coding	diarrhea,intestine	2	8.34
⊕	3	COPS3	COP9 Signalosome Subunit 3	Protein Coding	diarrhea,intestine	2	

	Symbol	Description	Gene Type	Matched Phenotypes	Matched Phenotypes Count	Score
⊖	CTNNB1	Catenin (Cadherin-Associated Protein), Beta 1, 88kDa	Protein Coding	diarrhea,intestine	2	5.57

Gene to Gene Relation

Pathways
Interactions
- *LRRFIP1* - Interacting Proteins(ENSP00000344456 -> ENSP00000375857)
- *LRRFIP1* - Interacting Proteins(P35222 -> Q32MZ4)

Pathways
- LRR FLII-interacting protein 1 (*LRRFIP1*) activates type I IFN production

SuperPathways
- Cytosolic sensors of pathogen-associated DNA: *LRRFIP1*

Gene to Phenotypes/Keywords Relation

Disorders
- *Intestinal* Polyps
- *Intestinal* Cancer
- Colorectal cancer (CRC) [MIM:114500]: A complex di... lesions arising from the inner wall of the large *intestine* (the colon) and the rectum. Genetic alterations a...

from MalaCards - The human disease database
Summaries for diseases linked to the gene CTNNB1
- colorectal cancer: the colon and rectum are part of the large *intestine*. colorectal cancer occurs when tumors form in the... lining of the large *intestine*. it is common in both men and women. the risk of ... ..., or smoke. symptoms of colorectal cancer include *diarrhea* or constipation a feeling that your bowel does no...
- adenomatous polyposis coli: familial adenomatous polyposis (fap) is an inherited condition that causes cancer of the large *intestine* (colon) and rectum. people with the classic type ... ...nt tumors of the duodenum (a section of the small *intestine*), stomach, bones, skin, and other tissues. some p...
- attenuated familial adenomatous polyposis: attenuated familial adenomatous polyposis (afap) i...herited condition that causes cancer of the large *intestine* (colon) and rectum. it is a milder form of classi... ...s) tumors of the duodenum (a section of the small *intestine*), stomach, bones, skin, thyroid and other tissues...
- adenomatous polyposis coli: an autosomal dominant disease that is caused by mu...of numerous polyps in the epithelium of the large *intestine* which are initially benign and later transform in...

Symptoms for diseases linked to the gene CTNNB1
- desmoid disease, hereditary: malabsorption/chronic *diarrhea*/steatorrhea

Function
Molecular Function
- Key downstream component of the canonical Wnt sign...alization. Blocks anoikis of malignant kidney and *intestinal* epithelial cells and promotes their anchorage-ind... (Swiss-Prot)

Publications
- Nuclear translocation of beta-catenin in hereditary and carcinogen-induced *intestinal* adenomas.
- Depletion of epithelial stem-cell compartments in the small *intestine* of mice lacking Tcf-4.
- Frequent somatic mutations of the beta-catenin gene in *intestinal*-type gastric cancer.
- *Intestinal* polyposis in mice with a dominant stable mutation...
- Immunolocalization of beta catenin in *intestinal* polyps of Peutz-Jeghers and juvenile polyposis sy...

	Symbol	Description	Gene Type	Matched Phenotypes	Matched Phenotypes Count	Score
⊕	TNF	Tumor Necrosis Factor	Protein Coding	diarrhea,intestine	2	3.65

Figure 1.30.21 VarElect indirect results. (**A**) The "indirectly related" tab displays genes with connections to phenotype keywords via intermediate genes. Implicated genes are sorted by relevance scores by default, and the table has features similar to those in the directly related tab (see Fig. 1.30.20A). (**B**) Each row may be expanded using the + icon to display up to five implicating genes that associate an implicating gene with phenotype keywords. (**C**) The row of each implicating gene may be expanded using the + icon to display the evidence that the implicating gene with both implicated gene and phenotype keywords in the top and bottom of the minicard respectively.

this establishes an indirect connection. Scores are given for implicating genes in relation to both an implicated gene (from the NGS experiment) and the phenotype.

12. Click the + icon in the first row to open the minicard displaying the evidence that connects the intermediary gene CTNNB1 with both the NGS-derived gene as well as the phenotype phrase (Fig. 1.30.21C). The gene products of CTNNB1 and LRRFIP1 interact, in addition to these genes sharing a pathway. The intestine and hair, on the other hand, are associated with CTNNB1 through the function, expression, and publications sections. MalaCards provides additional evidence for the aforementioned gene-disease connections in malady summaries and explicit naming of intestinal diseases.

    Gene-gene relationships appear at the top of the implicating gene minicard, thereby showcasing protein-protein interactions or shared biological pathways. Gene similarity may also be inferred through paralogy or textual associations, as from the summary or publications sections. Gene-phenotype associations may be supported by evidence from any GeneCards section and appear at the bottom of the minicard.

Managing VarElect projects

13. In order to save the results of the current analysis, click the Save button (Fig. 1.30.22A, top). Next, click Create New Project, enter `Patient 1` as its name, for example, and click Create (Fig. 1.30.22B, C). Enter `intestine hair` as the analysis name (Fig. 1.30.22D).

    VarElect supports the creation of projects for storing several analyses that may differ in their sets of genes, phenotypes, or both. This is useful for re-analyzing data at different timelines, sharing the analysis with colleagues, or presentation of additional phenotypes by patients. Projects may be created before an analysis is run or after it is completed, and require a project name. In order to associate an analysis that was run with a project, the appropriate project should be selected and the analysis must be named and then saved.

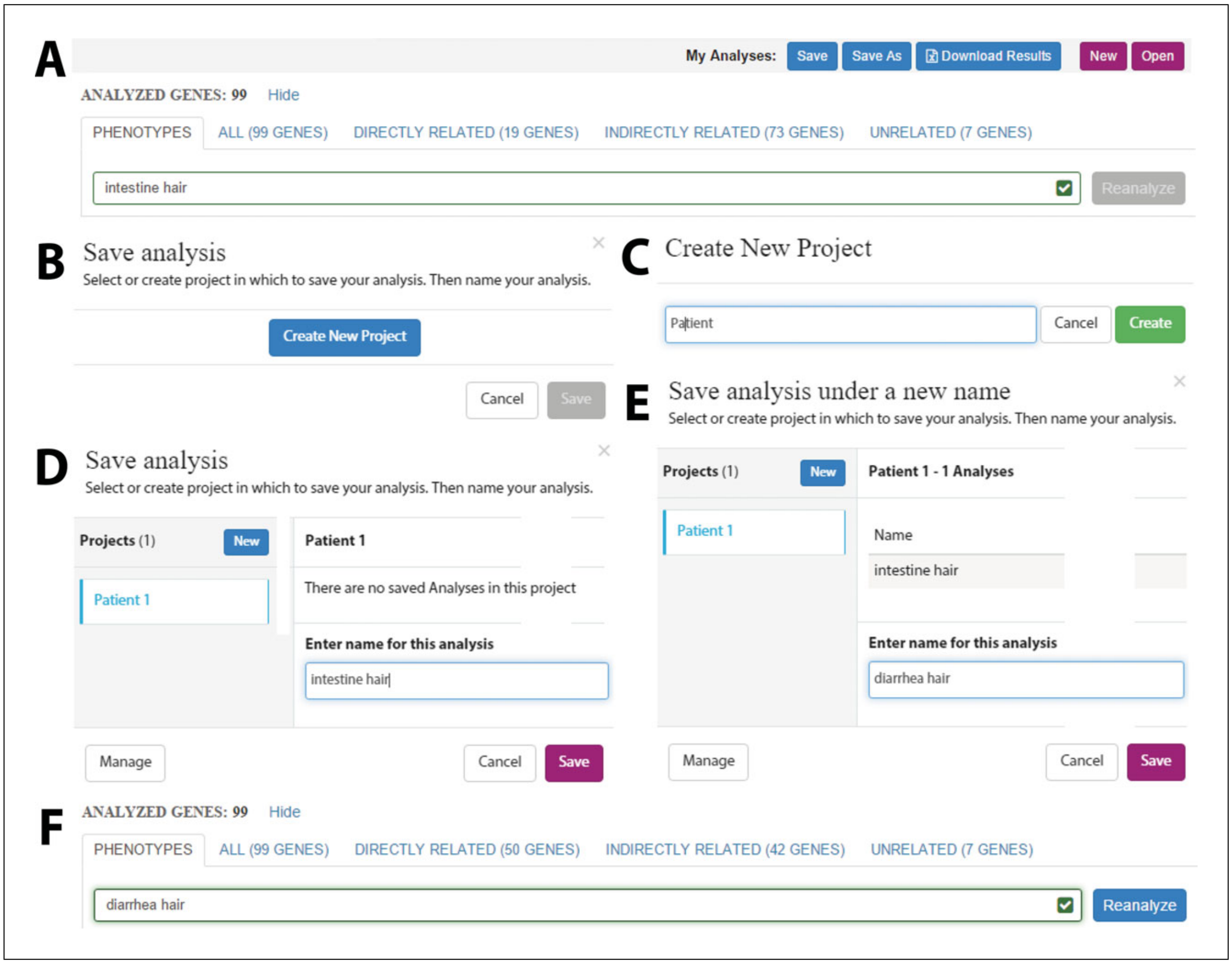
14. Run a new search by modifying the phenotype phrase in the query information section (Fig. 1.30.22F), which appears above the scored phenotype-associations section. In the phenotypes tab, change the phenotype to `diarrhea hair` and click the Reanalyze button. Re-analyzing the same gene list with the modified phenotype phrase decreases the directly related genes to 17, and causes the score to drop after the 5th gene, as opposed to dropping after the 11th gene in the previous phenotype phrase search. This could indicate results with increased specificity.

    When replacing the term `intestine` *with* `diarrhea`*, which is more indicative of this example's sequenced patient, the top-scoring gene, TTC37, exhibits strong connections with diarrhea. This gene is associated with trichohepatoenteric syndrome, and could explain liver malfunctions that were also observed in the patient. Upon validation of the mutation in this gene, the patient was confirmed as having atypical syndromic diarrhea (Oz-Levi et al., 2015). This case demonstrates the importance of term selection in the phenotype phrase.*

15. Save the new name for this analysis `diarrhea hair` (Fig. 1.30.22E).

16. Display the gene list directly related to the phenotype phrase by clicking the "Directly related" tab in the query information section (Fig. 1.30.22A).

    The results show the phenotype phrase that was used for the search, the lists of genes that were connected to the phenotype either directly or indirectly, in addition to those not connected at all, as well as the entire verified list of genes. Each gene in these lists links to the webcard in GeneCards, and all symbols may be copied for further use.


A
My Analyses: Save Save As Download Results New Open
ANALYZED GENES: 99 Hide
PHENOTYPES ALL (99 GENES) DIRECTLY RELATED (19 GENES) INDIRECTLY RELATED (73 GENES) UNRELATED (7 GENES)
Reanalyze
B
Create New Project
Cancel Save
C
Cancel Create
D
Projects (1) New
Manage Cancel Save
E
Projects (1) New
Manage Cancel Save
F
ANALYZED GENES: 99 Hide
PHENOTYPES ALL (99 GENES) DIRECTLY RELATED (50 GENES) INDIRECTLY RELATED (42 GENES) UNRELATED (7 GENES)
Reanalyze

A: intestine hair

B: Save analysis
Select or create project in which to save your analysis. Then name your analysis.

C: Create New Project
Patient

D: Save analysis
Select or create project in which to save your analysis. Then name your analysis.
Patient 1
Patient 1
There are no saved Analyses in this project
Enter name for this analysis: intestine hair

E: Save analysis under a new name
Select or create project in which to save your analysis. Then name your analysis.
Patient 1
Patient 1 - 1 Analyses
Name
intestine hair
Enter name for this analysis: diarrhea hair

F: diarrhea hair

Figure 1.30.22 Managing VarElect projects. (**A**) The query information section displays the phenotype phrase used for the VarElect search, in addition to four tabs with various gene lists (all, directly, indirectly, and unrelated). My Analyses buttons may be used for saving or downloading current results, as well as for beginning a new VarElect search or loading a previously saved one. (**B**) In order to save results, one must first “Create New Project”. (**C**) Only when the project is named does Create become available. (**D**) Only when the analysis is named does Save become available. (**E**) Several analyses may be saved in the same project. (**F**) A gene list may be re-analyzed by changing the phenotype phrase in the Phenotypes tab.

ACCESSIBILITY

The GeneCards database is freely available for educational and research purposes by nonprofit institutions at *http://www.genecards.org*. Commercial usage requires a license from LifeMap Sciences Inc. (Version 4.0 launch date—June 2015).

GUIDELINES FOR UNDERSTANDING RESULTS

In understanding the content and search output of an automatically generated database like GeneCards, it is important to put into perspective the complexity of gene-specific information. Genes often have multiple names and multiple definitions. Further, version inconsistency may exist between GeneCards and its sources. GeneCards traces every bit of the displayed information to its source. Some of the sources use manual curation, which is ongoing; some use text mining of published papers; and some include information submitted by users. Regardless of the source, however, different techniques, algorithms, and policies are used for data analysis and curation, creating a complex variety of data types that must be consolidated and ranked for relevance by GeneCards via a mostly automatic generic process. The user should be aware of the various techniques used to populate the different sections, and give the appropriate credence to manually curated information, which is more accurate but less comprehensive, as opposed to text-mining-based information (often labeled as inferred), which might be less accurate but more inclusive. All of these caveats must be kept in mind when analyzing the content of the cards. The purpose of GeneCards is to be a unified source of information, as well as to provide hints to facilitate the study of novel research questions. Further, search results scores obtained at different searching instances, whether in GeneCards or VarElect searches, cannot be compared. Within a given search, scores may be used to estimate the association strength between a given search term and the genes linked to it.

COMMENTARY

Background Information

GeneCards integrates functional information about human genes to facilitate exploring their biological data. The strength of GeneCards lies in the fact that all information is efficiently integrated into one easily searchable resource, with deep links and cross referencing to all sources and interlinked ontologies.

GeneCards Gene List

An offline process is responsible for generating the comprehensive integrated list of genes by mining heterogeneous, partially overlapping sources (the list of sources is available at *http://www.genecards.org/Guide/Sources*), unifying names and acronyms, and organizing characterizations. Gene unification is achieved by comparing gene locations and annotation from two sources (Ensembl and NCBI's Entrez Gene) and combining genes with overlapping locations (Coordinators, 2015; Cunningham et al., 2015; Rosen et al., 2003). When genes from the two sources are merged, all annotations from both sources are then associated with the merged gene. The gene's aliases are used to identify it in GeneCards' more than 100 sources, and their annotations, including additional aliases, are displayed in the card.

GeneCards Search and Scoring

The GeneCards search platform is Elasticsearch (Kononenko et al., 2014) Version 1.4.2. The Elasticsearch relevance scoring takes into account the frequency of the term in the document (which raises the score), the frequency of the term in documents across the site (which lowers the score), and the size of the subfield containing the term (if the term appears in a smaller field, such as gene name, the score increases).

Elasticsearch allows further customization of the score calculation, by weighting specific sections of the resulting GeneCard, in a process called boosting. GeneCards boosts the following annotation sections: (1) The Symbol; (2) Aliases and Descriptions; (3) Accessions for the major bioinformatics databases (NCBI, Ensembl, SwissProt); (4) Gene Summaries; (5) Disorders; and (6) Functions.

VarElect uses Elasticsearch with modifications that give precedence for concomitant hits of several keywords in a phenotype phrase rather than a strong association identified for one search term out of several entered. In addition, scores calculated in the indirect mode reflect the robustness of an association between gene A (implicated) and gene B (implicating), as well as the strength of affiliation between gene B and phenotype P. The first relationship is scored using a pre-computed gene-to-gene matrix that depicts all connections between genes in the GeneCards database. The search engine is queried using all gene symbols, thus scoring the propensity of a symbol to appear within the webcard of additional genes. This may occur when genes share a pathway or if their gene products interact; there are also textual appearances in various sections, such as the function or publication sections, thus conferring a phenotype by proxy. Pathway sharing is based on our recently developed algorithm which unifies pathways from various data sources into SuperPaths (Belinky et al., 2015). This process clusters pathways based on their gene content, thus greatly reducing redundancy by integrating 3,215 pathways from 12 sources into 1,073, while keeping them maximally informative. An immediate outcome of SuperPath generation is the inference of new gene connections that were not reported via any pre-unification pathway. The deduction of new gene associations (that also share publications and protein interactions; see Belinky et al., 2015) via SuperPaths has great potential to link a gene with established phenotype annotations with new genes.

An important challenge is to show the right mix between genes directly related to phenotypes of interest and "guilt-by-association" genes. There are cases whereby a gene with several strong indirect links to a phenotype via connecting genes could weigh more than a gene with a single moderate connection to the same phenotype. To this end, a normalization factor was introduced that enables the generation of a unified (interlaced) metric for direct and indirect genes (Stelzer et al., in press). Thus, the combined presentation of both direct and indirect gene-to-phenotype links provides a broad estimation of the robustness each type of evidence exhibits for the reported associations. In the near future, a third tab will be added to VarElect which will display an integrated rank of the direct and indirect candidate genes.

Data Collection and Integration

The data collection for each version of GeneCards begins by defining the full set of GeneCards genes, mined from three primary sources: first, the complete current list of HGNC-approved symbols (Gray et al., 2015) is used as the core gene set. Human Entrez Gene (Brown et al., 2015) entries that are different from the HGNC genes are then added. Finally, human Ensembl (Cunningham et al., 2015) records are compared to this gene list via our GeneLoc's exon-based unification algorithm (Rosen et al., 2003); those not found to be equivalent to others in the list are included as novel Ensembl-based GeneCards gene entries. Additional RNA genes are added using data from sources including fRNAdb, HinvDB, and Rfam (Belinky et al., 2013). These primary sources provide annotations for aliases, descriptions, previous symbols, gene category, location, summaries, paralogs, and ncRNA details. Once the gene set is in place with these significant annotations, over 125 data sources, including those noted above and others, are mined for thousands of additional descriptors.

Consistency Across Versions

GeneCards attempts to be as consistent as possible across versions to allow for usage and linking. Consequently, while gene symbols and IDs may change, the GeneCards Identifier (GC ID) is persistent. If the primary location of a gene changes significantly, it will receive a new GC ID, but the old one will always remain associated with that gene. When a user searches for a GC ID that is no longer current, such as GC04P070083, the gene to which that GC ID was assigned, in this case ALB, will be in the search results, with the GC ID shown as a previous identifier under the aliases section of the minicard.

GeneCards Suite

The tools and databases in the GeneCards Suite can be used together to maximize data retrieval. Each suite member gives the user in-depth information on a different facet of biological research. GeneCards is gene-centered, and provides a wealth of data specifically related to the gene of interest. The user who seeks more information in certain areas can effectively zoom into these areas using the other suite members. MalaCards focuses on diseases and disorders, presenting an individual card for each malady, with annotations and links including symptoms, drugs, articles, genes, clinical trials, related diseases/disorders, and more. LifeMap Discovery concentrates on gene expression, providing data on the developmental ontology of organ/tissues, anatomical compartments, and cells. It also presents manually curated gene expression at all developmental stages, as well as data extracted from high-throughput experiments and large-scale in situ databases. The user seeking pathways data will find it in PathCards, an integrated database of human biological pathways and their annotations, in which each PathCard provides information on one SuperPath that represents one or more human pathways. GeneLoc consolidates the genes from GeneCards' sources, merging them by location and assigning each GeneCards gene a unique GeneCards Identifier. The GeneLoc site provides a tabular view of a gene's genomic context, including neighboring genes, EST cluster, and markers. GenesLikeMe measures how genes are related, based on shared characteristics, such as expression, ontologies, or disorders. Using a gene set from the results of a GeneCards search, or any set of genes of interest, the user can extract GeneCards annotations for all genes in the set using GeneALaCart. The set can be further analyzed using GeneAnalytics, which can identify cell types, diseases, pathways, and functions related to the gene set and provides tools for further in-depth analysis of all genes in the set. VarElect identifies and prioritizes genes and variants according to their relevance to diseases and phenotypes of interest and allows the user to explore relationships between genes and gene variants and selected diseases, phenotypes, or any desired biological term via relevant pathways, interaction networks, and publications.

Critical Parameters and Troubleshooting

When interpreting information displayed in GeneCards, caution must be applied. GeneCards data is automatically generated from dozens of data sources (note that not all of our sources are publically available), and therefore is only as accurate as the information on which it is based.

Our implemented workflow analyzes over 100 sources, but some of the data may not be completely up-to-date. With the release of version 4, we have been increasing the frequency of incremental updates in order to keep the data more in sync with our sources. We are always open to comments and suggestions through our feedback form.

The first place to check regarding questions on how the data is generated is the User Guide pages, which are accessible from the main menu at the top of every page, and also from the "?" hyperlink on the right side of each section header. If any questions are not answered, users are welcome to contact us for further information and clarification.

Suggestions for Further Analysis

NGS analysis platform

TGex is a knowledge-driven NGS analysis platform (available at *http://tgex.genecards.org/*) that integrates the following basic components:

1. *VCF upload.* This first component is a management system where the variant list is uploaded (vcf format; the standard output of NGS). In addition, general and clinical information is entered by the user.
2. *Variant annotation.* At this stage, the variants of the uploaded sample go through extensive annotation, adding dozens of attributes per variant, including genetic data, effect and impact on the protein, frequency in control populations, and others.
3. *Variant browsing and filtration.* Here, the user can browse the genetic variants, modify filters on each of the data attributes, and select the relevant candidate variants. TGex provides a consolidated view of all the relevant data and leverages the strength of VarElect by ranking the variants according to their potential association with the given phenotypes.
4. *Reporting.* This is where all the information about the selected variants is automatically summarized into a comprehensive report, using GeneCards and MalaCards together with other data sources.

Comparison of VarElect to other tools

This section describes an essential activity for further development of VarElect—benchmark comparisons. We performed such analyses as described in Stelzer et al., (in press). These comparisons are not straightforward, as each of the compared tools has its own unique functionalities. Therefore, custom adaptations were made in each specific analysis pipeline with the goal of facilitating the comparisons, as described below.

Due to confidentiality constraints, it is difficult to obtain appropriate datasets in the public domain. Therefore, we performed much of the analyses without using private exome data. This is done by spiking of the causal gene/variant (based on either literature information or on in-house experimental results) into a background gene list (for VarElect, Phenolyzer) or into a variant call format (vcf) file. For gene list input we used 499 randomly selected background genes from the Illumina TruSight One gene panel *http://www.illumina.com/products/trusight-one-sequencing-panel.html*.

For vcf file input we use healthy volunteer data from *http://www.sanger.ac.uk/resources/software/exomiser/submit/resources/Pfeiffer.vcf*. Here, for each run, an extra vcf line was added for the causal variant with synthetically added high-quality calls and correct genotype information (homozygote or heterozygote).

Details on the specific analysis for each compared tool (also see Stelzer et al., in press) are as follows:

1. *Phenolyzer.* A list of gene symbols and relevant phenotype terms are entered. The phenotype terms undergo modifications as per a closed phenotype dictionary based among others on the human phenotype ontology (HPO). Phenolyzer default parameters are used, with the addition of "Word Cloud" and selection of "Mentha Protein Interaction Database" as an "Addon Gene Relations". The tool is then run and the rank of the spiked gene recorded.
2. *Exomiser.* We used the command line mode (V 7.1), that takes as input a vcf file and phenotypes. We used the tool's example vcf file; Since the tool accepts only HPO codes or OMIM disease identifiers, we converted phenotype entries using Phenomizer (*http://compbio.charite.de/phenomizer*). Analyses were performed using the hiPHIVE algorithm for cross-species phenotype scrutiny [23], with the parameters: prioritiser = hiPHIVE -F 2 -Q 30 –P true. The comparative phenotype interpretation analysis in VarElect acts upon 715 variants from 607 genes that pass Exomiser filtration.
3. *Ingenuity Variant Analysis.* The same spiked vcf files were analyzed as single patient rare genetic disease analysis with the default parameters of the software. The biological context filter was applied at first

without hops (comparable to VarElect's direct gene-phenotype relation). If the spiked variant was eliminated by the tool's filtration, then the "within one hop relation" option was applied (comparable to VarElect's indirect gene-phenotype relation). The comparative phenotype interpretation analysis in VarElect acts upon the genes that pass Ingenuity's filtration pipeline that precedes the "Biological context" filter which is the last one to be applied.

4. *Phevor— Omicia.* We use Omicia Opal 4.16.1, applying Phevor2 phenotype prioritization to a vcf file as described for Exomizer. This is followed by filtering with Omicia's tool VAAST3. If the spiked variant was eliminated by the tool's filtration, the relevant gene was considered "not found" in the rank comparison.

Acknowledgements

This work was supported by LifeMap Sciences Inc. (Massachusetts, U.S.A.), and the Crown Human Genome Center at the Weizmann Institute of Science (W.I.S.). The authors thank the W.I.S. GeneCards, Lancet Lab, and LifeMap Sciences teams for fruitful collaborations and comments, and the GeneCards Suite collaborators and users for their valuable feedback and support.

Literature Cited

Belinky, F., Bahir, I., Stelzer, G., Zimmerman, S., Rosen, N., Nativ, N., Dalah, I., Iny Stein, T., Rappaport, N., Mituyama, T., Safran, M., and Lancet, D. 2013. Non-redundant compendium of human ncRNA genes in GeneCards. *Bioinformatics* 29:255-261.

Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., and Lancet, D. 2015. PathCards: Multi-source consolidation of human biological pathways. *Database (Oxford)* 2015:bav006. doi: 10.1093/database/bav006.

Ben-Ari Fuchs, S., Lieder, I., Stelzer, G., Mazor, Y., Buzhor, E., Kaplan, S., Bogoch, Y., Plaschkes, I., Shitrit, A., Rappaport, N., Kohn, A., Edgar, R., Shenhav, L., Safran, M., Lancet, D., Guan-Golan, Y., Warshawsky, D., and Strichman, R. 2016. GeneAnalytics: An integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data. *OMICS* 20:139-151.

Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S.I., Schneider, R., and Jensen, L.J. 2014. COMPARTMENTS: Unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014:bau012. doi: 10.1093/database/bau012.

Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R., and Murphy, T.D. 2015. Gene: A gene-centered information resource at NCBI. *Nucleic Acids Res*. 43:D36-D42. doi: 10.1093/nar/gku1055.

Coordinators, N.R. 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 43:D6-D17. doi: 10.1093/nar/gku1130.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Giron, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Johnson, N., Juettemann, T., Kahari, A.K., Keenan, S., Martin, F.J., Maurel, T., McLaren, W., Murphy, D.N., Nag, R., Overduin, B., Parker, A., Patricio, M., Perry, E., Pignatelli, M., Riat, H.S., Sheppard, D., Taylor, K., Thormann, A., Vullo, A., Wilder, S.P., Zadissa, A., Aken, B.L., Birney, E., Harrow, J., Kinsella, R., Muffato, M., Ruffier, M., Searle, S.M., Spudich, G., Trevanion, S.J., Yates, A., Zerbino, D.R., and Flicek, P. 2015. Ensembl 2015. *Nucleic Acids Res*. 43:D662-D669. doi: 10.1093/nar/gku1010.

Edgar, R., Mazor, Y., Rinon, A., Blumenthal, J., Golan, Y., Buzhor, E., Livnat, I., Ben-Ari, S., Lieder, I., Shitrit, A., Gilboa, Y., Ben-Yehudah, A., Edri, O., Shraga, N., Bogoch, Y., Leshansky, L., Aharoni, S., West, M.D., Warshawsky, D., and Shtrichman, R. 2013. LifeMap Discovery: The embryonic development, stem cells, and regenerative medicine research portal. *PLoS One* 8:e66629. doi: 10.1371/journal.pone.0066629.

Fishilevich, S., Zimmerman, S., Kohn, A., Iny-Stein, T., Olender, T., Kolker, E., Safran, M., and Doron, L. 2016. Genic insights from integrated human proteomics in GeneCards. *Database (Oxford)*.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A. 2015. Genenames.org: The HGNC resources in 2015. *Nucleic Acids Res.* 43:D1079-D1085. doi: 10.1093/nar/gku1071.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 33:D514-D517. doi: 10.1093/nar/gki033.

Harel, A., Inger, A., Stelzer, G., Strichman-Almashanu, L., Dalah, I., Safran, M., and Lancet, D. 2009. GIFtS: Annotation landscape analysis with GeneCards. *BMC Bioinformatics* 10:348. doi: 10.1186/1471-2105-10-348.

Kononenko, O., Baysal, O., Holmes, R., and Godfrey, M.W. 2014. Mining modern repositories with elasticsearch. Proceedings of the 11th Working Conference on Mining Software Repositories, pp. 328-331. Hyderabad, India.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D, Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B, Sawyer, S., Lockhart, N.C, Demchok, J., Moore, H.F., 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45:580-585. doi: 10.1038/ng.2653.

Oz-Levi, D., Weiss, B., Lahad, A., Greenberger, S., Pode-Shakked, B., Somech, R., Olender, T., Tatarsky, P., Marek-Yagel, D., Pras, E., Anikster, Y., and Lancet, D. 2015. Exome sequencing as a differential diagnosis tool: Resolving mild trichohepatoenteric syndrome. *Clin. Genet.* 87:602-603. doi: 10.1111/cge.12494.

Perfetto, L., Briganti, L., Calderone, A., Perpetuini, A.C., Iannuccelli, M., Langone, F., Licata, L., Marinkovic, M., Mattioni, A., Pavlidou, T., Peluso, D., Petrilli, L.L., Pirro, S., Posca, D., Santonico, E., Silvestri, A., Spada, F., Castagnoli, L., and Cesareni, G. 2016. SIGNOR: A database of causal relationships between biological entities. *Nucleic Acids Res.* 44:D548-D554. doi: 10.1093/nar/gkv1048.

Rappaport, N., Twik, M., Nativ, N., Stelzer, G., Bahir, I., Stein, T.I., Safran, M., and Lancet, D. 2014. MalaCards: A comprehensive automatically-mined database of human diseases. *Curr. Protoc. Bioinform.* 47:1.24.1-1.24.19.

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Stein, T.I., Bahir, I., Belinky, F., Morrey, C.P., Safran, M., and Lancet, D. 2013. MalaCards: An integrated compendium for diseases and their annotation. *Database (Oxford)* 2013:bat018. doi: 10.1093/database/bat018.

Rosen, N., Chalifa-Caspi, V., Shmueli, O., Adato, A., Lapidot, M., Stampnitzky, J., Safran, M., and Lancet, D. 2003. GeneLoc: Exon-based integration of human genome maps. *Bioinformatics* 19 Suppl 1:i222-i224. doi: 10.1093/bioinformatics/btg1030.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. 2010. GeneCards Version 3: The human gene integrator. *Database (Oxford)* 2010:baq020. doi: 10.1093/database/baq020.

Stelzer, G., Inger, A., Olender, T., Iny-Stein, T., Dalah, I., Harel, A., Safran, M., and Lancet, D. 2009. GeneDecks: Paralog hunting and gene-set distillation with GeneCards annotation. *OMICS* 13:477-487. doi: 10.1089/omi.2009.0069.

Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., Twik, M., Belinky, F., Fishilevich, S., Nudel, R., Guan-Golan, Y., Warshawsky, D., Dahary, D., Kohn, A., Mazor, Y., Kaplan, S., Iny Stein, T., Baris, H.N., Rappaport, N., Safran, M., and Lancet, D. 2016. VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics*. in press.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., 3rd, and Su, A.I. 2009. BioGPS: An extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* 10:R130. doi: 10.1186/gb-2009-10-11-r130.

Internet Resources

http://www.genecards.org/

GeneCards is a compendium of human genes that provides genomic, proteomic, transcriptomic, genetic, and functional information on all known and predicted human genes. It is developed and maintained by the Lancet lab in the Department of Molecular Genetics at the Weizmann Institute of Science.

http://varelect.genecards.org

VarElect is a cutting-edge Variant Election application for disease/phenotype-dependent gene variant prioritization. It is an effective and user friendly tool for analyzing genes with variants following Next-Generation Sequencing ("NGS") experiments. VarElect can rapidly prioritize genes that have been found to have variants according to selected disease/phenotype-gene associations.