MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases

Noa Rappaport,1 Michal Twik,1 Noam Nativ,1 Gil Stelzer,1 Iris Bahir,1 Tsippi Iny Stein,1 Marilyn Safran,1,2 and Doron Lancet1

1Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel
2Department of Biological Services, Weizmann Institute of Science, Rehovot, Israel

ABSTRACT

Systems medicine provides insights into mechanisms of human diseases, and expedites the development of better diagnostics and drugs. To facilitate such strategies, we initiated MalaCards, a compendium of human diseases and their annotations, integrating and often remodeling information from 64 data sources. MalaCards employs, among others, the proven automatic data-mining strategies established in the construction of GeneCards, our widely used compendium of human genes. The development of MalaCards poses many algorithmic challenges, such as disease name unification, integrated classification, gene-disease association, and disease-targeted expression analysis. MalaCards displays a Web card for each of >19,000 human diseases, with 17 sections, including textual summaries, related diseases, related genes, genetic variations and tests, and relevant publications. Also included are a powerful search engine and a variety of categorized disease lists. This unit describes two basic protocols to search and browse MalaCards effectively.

INTRODUCTION

The MalaCards human malady compendium aims to cover all known human diseases, integrating disease names and annotations from over 60 data sources. MalaCards includes all types of diseases, including genetic, infectious, and other types that have not elsewhere been integrated to date into one unifying resource. Diseases are meta-categorized and cross-referenced to multiple external identifiers. MalaCards can be accessed at http://www.malacards.org/. MalaCards employs proven data-mining strategies used by the human gene compendium GeneCards (Harel et al., 2009; Stelzer et al., 2009; Safran et al., 2010; Stelzer et al., 2011; Belinky et al., 2013).

MalaCards generation and versioning

Each entry of MalaCards describes a disease, displayed in a computerized Web card. The information about each disease is organized into 17 main sections (Table 1.24.1). MalaCards is generated twice a year by an automated process, which first compiles an integrated list of disease names from a variety of sources, and then unifies the annotative information from each source and associates it with the appropriate section. 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.malacards.org/pages/whatsnew. More information about the build process and computed information can be obtained from http://www.malacards.org/pages/info, and is described in the Commentary section below.
Overview of use of MalaCards

MalaCards 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 free text as a query string, including Boolean expressions. A description of the search properties is available at http://www.malacards.org/pages/searchguide. An earlier description of many aspects of MalaCards, including the various sections within a specific card, is available in Rappaport et al. (2013) and at http://www.malacards.org/pages/info. An advanced search mechanism is planned for a future version. MalaCards content is linked to GeneCards (http://www.genecards.org; Safran et al., 2010) wherever a gene symbol is mentioned.

This unit gives a brief overview of the basic use cases of MalaCards. Basic Protocol 1 describes how to perform searches and analyze result summaries, using the search capability, category browsing, and alphabetical index. Basic Protocol 2 describes how to scrutinize and navigate the different parts of the resulting Web card for each disease.

SEARCHING AND BROWSING MalaCards

MalaCards can be accessed on the Internet at http://www.malacards.org/. The homepage (Fig. 1.24.1) is a common entry point to the Web site, showcasing most of the features and tools, quick searches, and future plans. Links to version history, collaborators, GeneCards suite members, previous version, categories list, and a disease index are also found on the homepage, along with a short description of the Web site and the number of sources and diseases. Major functionalities include: (1) search box; (2) random malady generator, which provides links to random diseases, with a pull-down menu that enables refinement of the randomization to restrict results to diseases with associated genes or summaries; and (3) sample malady, a link to a specific disease, rich with information, that is changed

Figure 1.24.1 The MalaCards homepage (http://www.malacards.org/) provides access to most of the features of the Web site through the keyword search box at the center of the page, as well as via links to a sample malady, random malady, feedback form, version history, main menu items, and more.
for each new version. A pull-down menu allows navigation to a specific section within the sample malady. A MalaCards Information Score (MIFTS) is shown (next to the sample malady) to indicate to what extent the disease is rich in information (see Commentary).

An example disease network picture is placed on the homepage for illustration of MalaCards strategy pertaining to the related diseases section of a second featured malady. A feedback link is available to allow users to pose questions, comments, and/or suggestions. Various pages related to site documentation, the development team and institution, academic licensing, and companion site links and the feedback system are located in the main menu at the top of the homepage, and at the top of all other pages on the site.

The Search box serves as a gateway to the disease universe within MalaCards. It is centrally located on the homepage, as well as at the top right corner of every page comprising the Web site.

**Necessary Resources**

An up-to-date Web browser, such as Firefox, Google Chrome, Internet Explorer, or Safari

**Search MalaCards**

1. Begin at the MalaCards home page (http://www.malacards.org/).

2. In the Search MalaCards text box, enter the following, and click the magnifying glass icon to submit the query (Fig. 1.24.1): "Crigler-Najjar syndrome" AND cholest*.

   The query term may be a disease name, gene name, or any other keyword. Boolean operators (AND/OR) can be used to query MalaCards, as can wildcards (*) when placed at the end of a word. Note that Booleans must be capitalized to yield expected results. The MalaCards search engine is based on Lucene/Solr (https://lucene.apache.org/solr/features.html) 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 search for multiple words (e.g., “Alzheimer disease”) behaves as an AND within the entire document; i.e., each of the words must exist in at least one of the sections of the matched MalaCards. 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) Disease name and aliases fields are boosted, so searches for those terms will match MalaCards accordingly named with higher scores than MalaCards 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 "Crigler-Najjar syndrome" AND cholest* query in March 2014 returned a list with 82 items, shown in the search results title at the top of the results table (Fig. 1.24.2).

   The MalaCards search results page includes a numbered list of diseases, each of them displaying hit context information (minicards), family information where available, MalaCards internal ID, main disease name, MIFTS score, and search specificity score given by Solr. The user can open the top 500 minicards using the “++” icon at the top of the minicards column. Each minicard highlights the hit within the specific entry, and one can further click on the specific section and view the hit within the card. Family information (see more in Commentary) is denoted by either “P” for parent or “c” for child; the mouseover contains the name of the parent. Clicking either “P” or “c” directs the user to the family information subsection within the related diseases section of the relevant MalaCard, which shows all of the members in the family. The MalaCards internal ID is a persistent identifier (see Commentary). All of the nontrivial information in all of the MalaCards is indexed, and provides fodder for the search engine.
Figure 1.24.2  Results of the search processed in Basic Protocol 1. There are 82 human diseases containing the searched string pattern, ranked by a relevance score. The “+” links open the highlighted hit context (minicard).

Figure 1.24.3  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 disease name (marked with a red arrow) opens and closes the minicard. The “++” sign at the top of the minicards column opens the top 500 minicards.

4. Click the “+” icon on the first line to open the hit context information (minicard) for Crigler-Najjar syndrome (Fig. 1.24.3).

Hit-context information shows that Crigler-Najjar syndrome is linked to cholestasis by two publications (#104 and #132, Pubmed IDs: 6872808 and 8012512), both of whose titles indicate that cholestasis is an accompanying condition of this syndrome. When the titles are long (as in the second publication of this example), they are truncated for readability. Clicking on “publications” leads to the section in MalaCards that displays the entire title.

5. Click the “publications” hyperlink within the opened minicard, which directs to the publications section of the MalaCard. Click “show all 147” to view all publications, including those highlighted by the minicard.
Figure 1.24.4  The alphabetical index of diseases featured in MalaCards.

**Browse MalaCards by alphabetical index**

6. To further study the cholestasis phenomenon, use the alphabetical index at the very bottom of the disease page, prefixed by Browse MalaCards. A page containing all of the disease entries for each letter is opened (Fig. 1.24.4). Click the “C” hyperlink to view all diseases beginning with the letter “C.”

   The user can browse the alphabetical index to view all of the diseases. The list displays the disease name, MalaCards internal ID, and acronym, if available.

7. Scroll down to Cholestasis to view disease information.

   Disease-specific information is displayed in the card (see Basic Protocol 2 for more information about each section). The first section, “Summaries,” contains descriptive information on the condition.

**Browse MalaCards using category pages**

8. Scroll down to the Aliases & Classifications section to view the categorization of Cholestasis.

   There may be a few categorizations displayed for each disease, ranked by relevance (see Commentary).

   In this example, the user can see that Cholestasis is related to “bile duct obstruction” by ICD10, and belongs to the MalaCards “Gastrointestinal” category.

9. Click the link to the MalaCards category Gastrointestinal to view all of the diseases in this category.
The list of MalaCards categories names, and the number of diseases associated with each of them.

The category page displays all of the diseases belonging to a specific category. Relevance of the diseases to the category is ranked (see Commentary). Category pages can also be accessed through the homepage (Fig. 1.24.1) or through Tools → Disease Categories under the main menu, available from any page.

10. To display all MalaCards categories, click Tools → Disease Categories (Fig. 1.24.5).

The categories page displays all of the categories in MalaCards (see Commentary). The table is divided into “Anatomical categories” and “Global categories,” each displaying a list of categories accompanied by their respective disease count.

EXPLORING A MalaCard

From the user perspective, the central entry of MalaCards is the disease page, Web “card,” or simply MalaCard. This is where one can find all available information pertaining to a disease of interest. The information within a MalaCard is divided into 17 sections, with a “jump-to-section” component at the left of each section title bar (Fig. 1.24.6), allowing navigation among the different sections, as well as to the bottom and the top of the card. Documentation is accessible via hyperlinks, often context-specific, from within many parts of the MalaCard, to the right of the section, entitled: “about this section.” The sections are listed in Table 1.24.1 along with the corresponding step number in the protocol below.

The card is divided into two panels, the “source” panel on the left and the “content” panel on the right. The content panel displays disease-specific information, and contains deep links to supporting sources, often with superscripts when multiple sources contain details about the datum. The sources panel contains the list of the sources that contributed information in the relevant section of the card, with a hyperlink to the sources section in the bottom of the card displaying the sources contributing to all sections. The sources in the sources section are annotated with the same superscripts.

MalaCards highlights a curated set of anatomical terms, and supplies relevant deep links to the LifeMap Discovery database (Edgar et al., 2013) for further information. For example, for the term “amygdala,” the following link will be featured: http://discovery.lifemapsc.com/in-vivo-development/brain/amygdala.
Figure 1.24.6  Related Diseases section of the “Crigler-Najjar syndrome” MalaCard, displaying detailed information about diseases in its family, and those related via gene sharing and/or textual overlap. A “jump to section” component, located to the left of the section header, allows rapid navigation between sections.

Table 1.24.1  MalaCards Section Names List.

<table>
<thead>
<tr>
<th>Section Name</th>
<th>Step no.$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summaries</td>
<td>3</td>
</tr>
<tr>
<td>Aliases&amp;Classifications</td>
<td>6</td>
</tr>
<tr>
<td>Related Diseases</td>
<td>7</td>
</tr>
<tr>
<td>Clinical Features</td>
<td>8</td>
</tr>
<tr>
<td>Drugs &amp; Therapeutics</td>
<td>9</td>
</tr>
<tr>
<td>Genetic Tests</td>
<td>10</td>
</tr>
<tr>
<td>Anatomical Context</td>
<td>11</td>
</tr>
<tr>
<td>Animal Models</td>
<td>12</td>
</tr>
<tr>
<td>Publications</td>
<td>13</td>
</tr>
<tr>
<td>Genes</td>
<td>14</td>
</tr>
<tr>
<td>Genetic Variations</td>
<td>15</td>
</tr>
<tr>
<td>Expression</td>
<td>16</td>
</tr>
<tr>
<td>Pathways</td>
<td>17</td>
</tr>
<tr>
<td>Compounds</td>
<td>18</td>
</tr>
<tr>
<td>GO Terms</td>
<td>19</td>
</tr>
<tr>
<td>Products</td>
<td>20</td>
</tr>
<tr>
<td>Sources</td>
<td>21</td>
</tr>
</tbody>
</table>

$^a$The number of the step in Basic Protocol 2, where this section is described.
Different sections contain ranking and scoring of the elements, including genes in the Genes section, diseases in the “Related diseases” section, and pathways in the Pathways section. Information about the computation of the different types of scores is found in (Rappaport et al., 2013) and at http://www.malacards.org/pages/info/#scores.

**Necessary Resources**

An up-to-date Web browser, such as Firefox, Google Chrome, Internet Explorer, or Safari

**Reviewing disease-specific information in a MalaCard**

Access the homepage (http://www.malacards.org/).

1. Click the Sample Malady name

   *For version 1.05, the sample malady is “Crigler-Najjar syndrome”.*

2. Inspect the MalaCard header (Fig. 1.24.7).

   The header contains the malady name, internal ID, MIFTS score, and acronym (where available). Below the main name are associated category names, linked to the Aliases & Classifications section. A statistics bar at the bottom of the header provides counts related to selected information within the card, with links to the corresponding sections. The optional acronym is either supplied by the sources, where available, or chosen by MalaCards (the shortest alias for the disease containing at most 4 characters).

3. Scroll down to the Summaries section (Fig. 1.24.8).
This section displays descriptions of the disease, as well as a MalaCards-generated summary. Summaries typically include a short definition of the disease, organs involved, etiology, and main symptoms, e.g., in “Crigler-Najjar syndrome”: NIH Rare Diseases: Crigler najjar syndrome, type 1 is an inherited disorder in which bilirubin, a substance made by the liver, cannot be broken down. Summaries from some of the sources are fully displayed (e.g., NIH rare diseases, Disease Ontology), some are partially displayed (e.g., Wikipedia), and for others, only a deep link is displayed (e.g., GeneReviews and OMIM). MalaCards-generated summaries highlight the disease’s significant annotations (e.g., associated genes, drugs, pathways, etc.).

4. Click “Fully expand this MalaCard” to the left of the Summaries section to expand all of the subsection lists.

Long lists within the card sections are partially hidden by default (initially showing only the most relevant information for efficiency), with a “show all” option to display the complete list. Pressing “Fully expand this MalaCard” activates “see all” in all of the sections, and enables convenient searches within the card.

5. Click “Export this MalaCard” to the left of the “Summaries” to download the full content of the card.

The user can download the contents of a MalaCard using this option to facilitate computational analysis. The results file is in a parsable Excel format, containing all information including deep links. “###” is a section separator; “##” is a title marker; and “#” precedes table column names. Data for scientific collaborations can also be requested by filling out the “academic licensing” form accessed through the main menu under “About MalaCards”.

6. Scroll down to the Aliases & Classifications section (Fig. 1.24.9).

This section includes the following subsections:

a. **Classifications subsection**: Displays classifications, currently from the International Classification of Diseases (ICD10; World-Health-Organization, 1992) and MalaCards itself (see Commentary). Mapping to ICD10 was done using the Orphanet mapping file (can be found at http://www.orpha.net/cgi-bin/index.php/) and via our original heuristic name-comparison algorithm. The full tree of ICD10 classification “above” the disease is displayed, with each branch linked to the corresponding place in the ICD10 hierarchy browser. Additional classification types will be added in the future. The MalaCards classification algorithm is described in the Commentary.

b. **Characteristics subsection**: (Orphanet epidemiological data) provides Orphanet data (where available) on mode of inheritance, age of onset, age of death, and prevalence of the disease.

c. **Aliases & Descriptions subsection**: Displays synonyms and aliases for the relevant MalaCards malady, as extracted from a subset of the sources listed on the left side of the section. Strongly similar aliases, even if trivially different, are included, to match common expectations and to facilitate searches. The disease name appears first, with its own associated source-indicating superscripts. The alias list is sorted by the count of contributing sources, sub-sorted by descending length. The main malady name is shown in bold.

d. **External IDs subsection**: Displays external IDs, which are cross-references to IDs of external databases/ontologies. The external IDs are searchable. Some mappings, as noted, are performed by MalaCards, and some are supplied by other sources.
For “Crigler-Najjar syndrome,” the user can see that different sources give different names for the same condition, highlighting the major benefit of MalaCards in this respect. While Orphanet calls the condition “hereditary unconjugated hyperbilirubinemia type 1” and does not mention “Crigler-Najjar syndrome” as an alias, Genetic Home Reference, Novoseek, Disease Ontology, and other sources use the more common name “Crigler-Najjar syndrome,” which is chosen to be the main disease name in MalaCards. MalaCards is thus able to unify all of these names and annotations. Also, if one wishes to use external IDs for diagnosis coding or cross-linking, a variety of these are mapped to the disease. This disease is classified as Genetic, Metabolic, and related to the Liver, as expected.

7. Scroll down to the “Related diseases” section to view which diseases are related to “Crigler-Najjar syndrome,” to suggest similar symptoms or treatments (Fig. 1.24.6).

This section includes the following subsections:

a. Family information subsection: Lists diseases that are lexically grouped with the current disease (see Commentary). Usually the family contains disease types, same diseases with different mode of inheritance, same diseases with different genetic basis, and so on.

b. Related diseases subsection: Displays a unified scored list of diseases obtained in two ways: first, by GeneDecks set analysis (Stelzer et al., 2009), whereby other diseases computed to have significant shared descriptors for the target disease’s affiliated genes are collected (see Commentary); second, as matched by MalaCards searches. All obtained related diseases are sorted by the MalaCards composite relevance score, described in Rappaport et al. (2013), and at http://www.malacards.org/pages/info/#scores.
c. **Graphical network subsection:** Displays images generated using the Gephi toolkit (Bastian and Jacomy, 2009). An image on a MalaCard encompasses the top 20 scored related diseases. Each related disease is a node, while edges represent a connection to the disease of interest, as well as the top 20 interconnections between the related diseases themselves, where available. Images are not generated for diseases having fewer than five connections. Edge thickness is proportional to the connection strength. Direct links from the current MalaCard’s disease are colored in red. The most related conditions to “Crigler-Najjar syndrome” are, as expected, family member “Crigler Najjar syndrome, type 2” as well as “Gilbert syndrome,” which is also characterized by high bilirubin levels (Koiwai et al., 1995).

8. Scroll down to the “Clinical Features” section.

This section provides information and links about symptoms and other clinical attributes of the disease. If available, deep links to “clinical features” and “clinical synopsis” data from OMIM are provided. Also, a symptoms list is taken from Orphanet. Symptoms typically represent changes from normal function, sensation, or appearance, but may also be other MalaCards maladies with their own cards.

9. Scroll down to the Drugs & Therapeutics section.

This section includes the following subsections:

a. Approved drugs—deep link for search in CenterWatch for newly approved drugs.

b. Clinical trials:
   i. Deep link for search in ClinicalTrials.gov for clinical trials.
   ii. Deep link for search in NIH Clinical Center for clinical research studies.
   iii. Deep link for search in CenterWatch (http://www.centerwatch.com/) for clinical trials. For a few maladies, sample drug profiles have been pulled from CenterWatch’s proprietary Drugs in Clinical Trials Database, a subscriber-only service offering access to more than 4,000 detailed profiles on drugs in the pipeline to monitor drug performance, track competitor activity, and find study opportunities.

c. Inferred drug relations via UMLS/NDF-RT, where available—combined information from the Unified Medical Language System (UMLS) and the National Drug File-Reference Terminology (NDF-RT). Initially, a MalaCards name is mapped to a UMLS concept representing a disease by utilizing the MetaMap system. Subsequently, the NDF-RT terminology within UMLS is used to provide a link of such disease concepts to drug(s) via the ‘may be treated by’ relationship. This work was done in collaboration with C. Paul Morrey (further details in Rappaport et al., 2013). See example in http://malacards.org/card/huntingtons_disease#therapeutic.

d. Cell-based therapeutics approaches from LifeMap Discovery:
   i. Link to the disease page in LifeMap discovery, where available.
   ii. Stem-cell-based therapeutic approaches.
   iii. Embryonic/adult cultured cells which are candidate therapeutic approaches, with the corresponding supporting Pubmed IDs.

*LifeMap data indeed associates stem cells therapy approaches using hepatocytes to “Crigler Najjar syndrome.”*

10. Scroll down to the Genetic Tests section.

This section provides descriptions of genetic testing, specialized cytogenetic testing, and biochemical testing for inherited disorders. Genetic tests are extracted from both LifeMap Discovery...
GeneTests and the Genetic Testing Registry (GTR). Where available, the tested loci are noted, with links to GeneCards.

*Genetic tests for “Crigler-Najjar syndrome” are available, and the user can follow the links to find out which labs perform the diagnostic tests.*

11. Scroll down to the Anatomical Context section.

This section provides descriptions on cells, compartments, and organs relevant to the disease. This section may contain the following subsections, where available:

a. The MalaCards organs/tissues related to the disease—obtained by searching MalaCards for a set of predefined tissues; the score is given by counting the hits in the various sections, where some section hit counts are boosted. The organs are sorted by score, and the score is displayed in the mouseover.

b. Foundational Model of Anatomy (FMA) ontology data interconnections are extracted via the Disease Ontology (Schriml et al., 2012; see example in http://www.malacards.org/card/avian_influenza#anatomical_context).

c. LifeMap cells/anatomical compartments in embryo or adult related to the disease, with deep links to the anatomical part and the relevance type.

*As expected, “Crigler-Najjar syndrome” is related to the liver, as well as to the brain, which can be damaged by a buildup of bilirubin (Al-Shurafa et al., 2001).*

12. Scroll down to the Animal Models section.

This section provides orthologous mouse phenotypes, which are obtained by being contextually related to the key disease using the GeneDecks mechanism described in the Commentary, applied to the set of affiliated genes (see step 14). Phenotypes are scored according to their relevance, and the table also contains MGI accessions and affiliating genes. The section also displays a link to inGenious Targeting Laboratory, which can supply customizable mouse models.

*“Crigler-Najjar syndrome” does not have related mouse phenotypes; for an example of a malady that does, see http://www.malacards.org/card/alzheimers_disease#phenotypes.*

13. Scroll down to the Publications section.

*This section provides publications associated with the disease, obtained by pre-searching the Pubmed database for the main disease name using E-Utilities (http://www.ncbi.nlm.nih.gov/books/NBK25500/). For each publication, the title and link to the PubMed article are supplied, as well as the first and last author names, and year. The articles are ranked by date of publication.*

14. Scroll down to the Genes section (Fig. 1.24.10).

This section provides a list of associated genes. Note that association does not necessarily mean causality, and could just as well be a negative association. The genes list is composed by taking into account:

a. The results of a search of the main disease name in GeneCards, yielding genes whose cards mention the disease name.

b. Genetic testing resources supplying specific genetic tests for the disease.

c. Genetic variations resources supplying specific causative variations in genes for the disease (see “Genetic variations” section in step 15).

d. Resources that manually curate the association of the disease with genes (e.g., Orphanet, OMIM).

*A prioritizing algorithm is applied to generate the genes list. The table includes gene symbols (linked to GeneCards), commercial products info, gene description, and*
relevance scores. The relevance score is computed by factoring in the importance of the different resources that associate the gene with the disease. For each gene, where available, the “products” column contains an icon that opens up a products page relevant to this gene—currently including proteins, lysates, antibodies, and knockout mice. Below the genes table, a company-based products table summarizes the relevant products for all of the disease genes. In version 1.05, a new Implication column was added to the genes table, containing the supporting evidence for the association between the gene and the disease, with deep links to the sources supplying the association support.

For “Crigler-Najjar syndrome,” the first gene UGT1A1 contains the most support for its association with the disease, as shown in the implication column: “Molecular basis known,” “Causative germline mutation,” “Causative variation,” and Genetic Tests, as well as support from GeneCards. Clicking “show sections” under the GeneCards Inferred annotation opens a list of relevant GeneCards sections, with deep links to the location of the hit of the disease name search within the GeneCard. UGT1A1 is indeed known to be the causative gene for the disease (Rossi et al., 2005), as stated in the summary from NIH Rare diseases in the Summaries section.

15. Click the superscript next to the “Causative variation” annotation in the Implication column of the first gene UGT1A1.

Clicking takes the user to the Genetic Variations section. This section provides known causative variations for the disease currently extracted from the UniProt Humsavar project (http://www.uniprot.org/docs/humsavar). The table shows the gene symbol...
(linked to GeneCards), amino acid change, link to the variation details and source, and SNP ID where available.

Indeed, all causative variations listed in this section for “Crigler-Najjar syndrome” are in the disease gene UGT1A1.

16. To see in which tissues the associated genes for the disease are mostly expressed, scroll down to the Expression section (Fig. 1.24.11).

This section provides normal tissue expression profiles for genes affiliated with the disease, via experimental results from BioGPS (Wu et al., 2009). The expression plots display the tissue-specific gene expression levels typifying the disease. The y axis represents the genes ranked by expression levels; the x axis represents the tissues. Each column shows up to 100 of the most highly expressed disease-associated genes, ranked by their expression level. The size of the bar is determined by the number of genes in the tissue that have the top 20% expression level of the disease-associated genes. The corresponding gene squares are colored according to the log10 of their expression levels (to the 2/3 power). The color scale is common to all diseases in all tissues. The closer the color to red, the higher the expression for the specific gene. Higher bars for specific tissues represent tissues in which the fraction of highly expressed genes of the disease gene set is higher than other tissues for that disease.

In addition, a link to the Gene Expression Omnibus (GEO) database (Edgar et al., 2002) is featured to direct the user to a search of the disease name within its datasets for expression profiling by array and high-throughput sequencing experiments that contain human disease-states data.

In “Crigler-Najjar syndrome,” the bar in the liver column is the highest, suggesting that this tissue is dominant for the genes associated with the disease, also in normal conditions.

17. Scroll down to the Pathways section.
This section provides super pathways (see http://www.genecards.org/info.shtml#pathways_interactions) related to the disease, obtained by being contextually related to the key disease using the GeneDecks mechanism described in the Commentary, applied to the set of affiliated genes. The pathways are extracted from a subset of the sources listed on the left panels. Entries are scored according to their relevance. The table displays the super pathways, with their member pathways indented, the relevance score, and the affiliating genes that led to the association of the specific pathway to the disease.

In “Crigler-Najjar syndrome,” indeed, the first related pathways are involved in liver metabolism and clearance.

18. Scroll down to the Compounds section.

This section provides relationships between MalaCards diseases and chemical compounds, obtained by being contextually related to the key disease using GeneDecks Set Distiller as described in the Commentary, applied to the set of affiliated genes. Compounds are extracted from a subset of the sources listed in the left panel. Entries are scored according to their relevance.

In “Crigler-Najjar syndrome,” the top compounds are indeed related to the top disease gene product activity; these include bilirubin, diglucuronide, and aldosterone 18-glucuronide, which assist in the excretion of toxic substances such as bilirubin in the liver (Neims, 1982).

19. Scroll down to the GO Terms section.

This section provides cellular component ontologies, biological process ontologies, and molecular function ontologies enriched in the set of genes affiliated with the disease. The table displays the name of the relevant ontology, the GO ID, which is the identifier used by Gene Ontology and linked to the GO entry, and the genes related to the disease, as well as to the specific ontology using the GeneDecks mechanism described in the Commentary. The entries are scored according to their relevance.

In “Crigler-Najjar syndrome,” the top biological process GO terms are related to flavone metabolism, which is the global process in which the top disease associated genes are involved by modifying toxic compounds in order to enable their allocation. “Glucuronosyltransferase activity” is indeed the specific reaction catalyzed by the disease gene product UGT1A1 (Al-Shurafa et al., 2001).

20. Scroll down 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 genes related to the disease.

21. Scroll down to the Sources section.

This section provides links to all of the MalaCards sources. A list of the sources with a short description for each can be found in http://www.malacards.org/pages/info/#sources.

GUIDELINES FOR UNDERSTANDING RESULTS

In understanding the content and searches output of an automatically generated database like MalaCards, it is important to put into perspective the complexity of disease-specific information. Diseases often have multiple names, multiple definitions, and subtyping. Further, version inconsistency may exist between MalaCards and its sources. MalaCards 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
MalaCards by a mostly automatic generic process. The user should be aware of the way the different sections are populated, and give the appropriate credence to manually curated information, which is more accurate but less comprehensive, as opposed to text-mining-based information, which is less accurate but more inclusive. Moreover, some of the section data, such as pathways, compounds, GO terms, and others, are based on shared annotations of the associated genes, and although very informative and suggestive, may not always be completely accurate. All of these caveats must be kept in mind when analyzing the content of the disease cards. The purpose of MalaCards is to be a unified source of information, as well as to provide hints to facilitate the study of novel research questions and therapeutic strategies.

COMMENTARY

Background Information

MalaCards integrates functional information about human diseases to facilitate exploring their biological data. The strength of MalaCards 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.

MalaCards disease list

An offline process is responsible for generating the comprehensive integrated list of diseases by mining heterogeneous, partially overlapping sources (the list of sources is available at http://www.malacards.org/pages/info/id:sources#sources), unifying names and acronyms, and organizing characterizations. Disease name unification is performed by transforming each name to a canonical form. The canonical form is constructed by lower-casing, lexical sorting of words, removing special characters and common words like “disease” and “syndrome,” as well as merging equivalent words (like “juvenile” and “childhood”). This canonical form is then hashed and used for comparison against transformed new names. The sources are observed in a predefined hierarchical order. The diseases in the first source are added to the disease list, and each subsequent source names are canonicalized and compared to the canonical form of the existing list. If a name exists, its aliases and source annotations are added to the existing entry; otherwise a new entry is added.

For each malady, a unique internal ID is generated, composed of the first letter of its name, followed by the next two consonants, followed by a serial number. For example, the ID generated for “rett syndrome” is RTT001.

MalaCard. An MIFTS is assigned to each disease by summing the base 10 logarithms of the counts of its populated annotations. This score currently ranges from 1 to 101, with the MalaCard scored 101 being the most annotated card. The MIFTS is displayed in the card header, as well as in the search results table.

MalaCards annotations

MalaCards annotation schemes are described in detail in Rappaport et al. (2013). Briefly, MalaCards employs four different annotation schemes, as follows:

1. **Source mining.** Data sources containing disease-specific information are mined in order to populate relevant sections of a MalaCard. To this end, we define two types of sources: primary sources are those which are used to derive both disease names and annotations; secondary sources are those from which only annotations are derived. Direct source mining provides information for the Aliases and Classifications, Summaries, Clinical Features, Drugs and Therapeutics, Genetic Tests, and Anatomical Context sections. When appropriate, in-house analysis is performed, in order to link annotations to diseases or to integrate and display disease-specific data. For example, we have developed a process that employs UMLS concepts to map a disease to drugs used for its treatment.

2. **GeneCards search.** One central annotation source for MalaCards is an automated use of the GeneCards search engine, including section-specific advanced searches. For example, the gene set affiliated with a disease is obtained by using the disease name as a search string, which enables the generation of the related genes section in MalaCards. Importantly, gene association does not imply causality between the gene and the disease. Associations sometimes include annotations like ‘unaffected’; these should therefore be verified using the “GeneCards section context” link.
Similarly, the publications associated with a disease are obtained via a search for its name in all of the publication titles within GeneCards.

3. GeneCards set analyses (GeneDecks). MalaCards implements a strategy in which gene-disease relationships within GeneCards are used to create disease-specific content. For this, we leverage the GeneCards GeneDecks tool, in its Set Distiller mode [more information can be found in (Stelzer et al., 2009) and http://www.genecards.org/index.php?path=/GeneDecks/aboutGeneDecks#setDistiller].

The disease-associated gene set (generated as described above) is forwarded to GeneDecks, which distills statistically significant descriptors enriched in this set. This process also assigns a relevance score for every hit, and is employed to populate the Related Diseases, Phenotypes, Pathways, Compounds, Animal Models, and GO terms sections. In these sections, the relevant tables display the top affiliating genes, linked to their respective contexts within GeneCards.

4. MalaCards search. We use MalaCards searches to populate additional sections, including elucidating new relations among diseases in the Related Diseases section, and associating tissues in the Anatomical Context section.

Redundancy and integration

Constructing an inclusive, universal list of diseases is the main challenge of MalaCards. A major difficulty stems from the fact that MalaCards automatically extracts its data from different databases, each having different granularities and disease-unification policies. For example, one database may have only one card for Diabetes Mellitus, unifying all known information on this condition, while another database may have two separate entries for Type 1 Diabetes Mellitus and Type 2 Diabetes Mellitus. MalaCards policy is to have a separate Web card for each subtype, but to also have a coherent and efficient mechanism to associate related entries.

Relationship between MalaCards entries

MalaCards’ concept of relatedness connects entries using various functional and textual criteria. MalaCards links diseases using two strategies:

1. Through significant gene sharing of the top associated genes.

2. Through textual search, specifically by requiring the disease name of one malady to appear in the card of another (e.g., in the “Summaries”, “Publications” or other significant sections).

The scores of both algorithms are combined in order to create a unified composite list. In the future, we plan to expand the connectivity factors to include other functional characteristics such as symptoms and drugs.

Another type of disease connection is “family association”; this is done textually by removing descriptive, high-frequency words and looking for relatedness of the base name. This usually unites types, inheritance modes, and so on. In version 1.05, there are 1276 families, grouping 5574 diseases. The disease “parent” is chosen to be the disease containing the most information in its card (the one with the highest MIFTS).

MalaCards categories

MalaCards categorizes diseases into two types of categories: global (fetal, genetic, cancer, and infectious) and anatomical (eye, ear, liver, blood, etc.). A disease can belong to up to four anatomical categories, in addition to a variety of global categories. For example, Chronic Myeloid Leukemia is both a cancer and a blood-related disease. Categories are associated with diseases in the following ways: first, by the mapping of MalaCards categories to existing and widely used categorization systems like ICD10 and Orphanet; second, by an algorithm using a choice of category-specific keywords contained in disease names and annotations. The algorithm gives precedence to the presence of specific keywords in the main name (e.g., carcinoma for cancer, hepato* for liver etc.), but also searches for a threshold frequency in disease descriptions. Each category may, in addition, have excluded keywords which should not appear in conjunction with its keywords (such as “bone marrow” for the “bone” category). Lastly, heuristics are applied to associate diseases into categories; for example, the existence of causative variations or a genetic test puts a disease into the genetic category.

In order to rank the relevance of multiple categories within the same specific disease, a composite score is compiled. This score takes into account both the number of keyword hits in the different subsections of the disease based on significance of the subsections, as well as a mapping of the disease to an external category system (such as ICD10 or Orphanet).

Within the category pages, the list of diseases is ranked as well, to reflect the level of support for the association of each disease to the category. The score is composed in the
following way: \( S_m = N - r + 1 + I/I_{\text{max}} \), where: (1) \( N \) is the total number of categories, (2) \( r \) is the rank of the current category within the current disease, (3) \( I \) is the internal score of the specific category within the specific disease (see above), and (4) \( I_{\text{max}} \) is the maximal score of all categories within all diseases.

Consistency across versions
MalaCards attempts to be as consistent as possible across versions to allow for usage and linking. Consequently, URLs are preserved, even if the disease name has changed or the disease was merged with another. In situations like these, each of the old URLs will be redirected to the new one. If a disease was removed completely from MalaCards, the old link will be redirected to the search results page found by querying the old disease name.

Critical Parameters and Troubleshooting
When interpreting information displayed in MalaCards, caution must be applied. MalaCards data is automatically generated from publicly available data, and therefore is only as accurate as the information on which it is based. Different disease granularity, as well as diverse conventions and medical definitions, may cause redundancy and reduce integration levels. MalaCards’ policy is to have all disease subtypes in separate cards, but to connect them through the related diseases section, family information, and, consequently, search results. Sometimes, a gene or organ noted as being related to a disease is in fact not connected based on a real association, but via a textual artifact of the system’s automated textual searches within GeneCards and MalaCards.

Our implemented workflow analyzes over 60 sources, but since currently MalaCards is typically updated about twice a year, some of the data may not be completely up to date. 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 Help pages, which are accessible from the main menu at the top of every page, and also from the “about” hyperlink to the right of each section. If any questions are not answered, users are welcome to contact us for further information and clarification.

Acknowledgements
This work was supported by LifeMap Sciences Inc. (California, U.S.A.), the Crown Human Genome Center at the Weizmann Institute of Science (W.I.S.), and the SysKid EU FP7 project (number 241544). The authors thank C. Paul Morrey, Moran Gershoni, and the W.I.S. GeneCards and LifeMap Sciences teams for fruitful collaborations and comments. The authors would also like to thank MalaCards users for their feedback and support.

Literature Cited


**Internet Resources**

http://www.genecards.org/

GeneCards (http://www.genecards.org/) is a compendium of human genes that provides genomic, proteomic, transcriptomic, genetic and functional information on all known and predicted human genes. It is being developed and maintained by the Crown Human Genome Center at the Weizmann Institute of Science.

http://www.genecards.org/index.php?path=Gene Decks

GeneDecks is a novel analysis tool which provides a similarity metric by highlighting shared descriptors between genes, based on the rich annotation within the GeneCards compendium of human genes. GeneDecks features Partner Hunter and Set Distiller.