Chapter 2
The GeneCards Suite

Marilyn Safran, Naomi Rosen, Michal Twik, Ruth BarShir, Tsippi Iny Stein, Dvir Dahary, Simon Fishilevich, and Doron Lancet

Abstract The GeneCards® database of human genes was launched in 1997 and has expanded since then to encompass gene-centric, disease-centric, and pathway-centric entities and relationships within the GeneCards Suite, effectively navigating the universe of human biological data—genes, proteins, cells, regulatory elements, biological pathways, and diseases—and the connections among them. The knowledgebase amalgamates information from >150 selected sources related to genes, proteins, ncRNAs, regulatory elements, chemical compounds, drugs, splice variants, SNPs, signaling molecules, differentiation protocols, biological pathways, stem cells, genetic tests, clinical trials, diseases, publications, and more and empowers the suite's Next Generation Sequencing (NGS), gene set, shared descriptors, and batch query analysis tools.

Keywords GeneCards · Bioinformatics · Biological database · Diseases · Gene prioritization · Integrated information retrieval · Next generation

2.1 Introduction

The GeneCards® database of human genes was launched in 1997 (Rebhan et al. 1997) and has expanded since then to encompass gene-centric, disease-centric, and pathway-centric entities and relationships within the GeneCards Suite, effectively navigating the universe of human biological data—genes, proteins, cells, regulatory elements, biological pathways, and diseases—and the connections among them. The suite's integrated biomedical knowledgebase includes GeneCards (Stelzer et al. 2016a), the integrated human gene database, MalaCards (Rappaport et al. 2017a),

M. Safran · N. Rosen · M. Twik · R. BarShir · T. I. Stein · S. Fishilevich · D. Lancet (✉)
Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot, Israel
e-mail: Doron.Lancet@weizmann.ac.il

D. Dahary
LifeMap Sciences Inc., Marshfield, MA, USA

I. Abugessaisa, T. Kasukawa (eds.), *Practical Guide to Life Science Databases*,
https://doi.org/10.1007/978-981-16-5812-9_2

the unified human disease database, PathCards (Belinky et al. 2015), the consolidated human pathways database, LifeMap Discovery (Edgar et al. 2013), the embryonic development and stem cell compendium, GeneLoc (Rosen et al. 2003), the human genomic neighborhood location-based database, and GeneHancer (Fishilevich et al. 2017), an innovative and growing regulatory element database with ~250,000 enhancer and promoter entries. The knowledgebase amalgamates information from >150 selected sources related to genes, proteins, ncRNAs, regulatory elements, chemical compounds, drugs, splice variants, SNPs, signaling molecules, differentiation protocols, biological pathways, stem cells, genetic tests, clinical trials, diseases, publications, and more, and empowers Next Generation Sequencing (NGS) analysis by highlighting associations between genes and phenotypes, providing supporting evidence for immediate evaluation via the suite's NGS analysis tools: VarElect (Stelzer et al. 2016b), the phenotype interpreter, receives a list of genes and phenotypes as input and computes prioritized direct (keyword-based) and indirect (inferred from gene-to-gene associations) gene/disease connections; TGex, the VCF-to-report clinical analyzer, incorporates VarElect's algorithms and automatically generates clinical case reports. Rounding out the suite are GeneAnalytics (Ben-Ari Fuchs et al. 2016), for gene set analysis, GenesLikeMe (Stelzer et al. 2009) for finding genes with shared descriptors, and GeneALaCart (Stelzer et al. 2016a) for batch queries.

The suite's websites, data dumps, APIs, publications, and collaborations are enjoyed by >3.5 million users, including research and applied scientists, doctors, geneticists, and lay-people, in >3000 institutions worldwide, encompassing academia, national patent offices, leading biopharma and diagnostic companies, and hospitals.

2.2 Database Overview

2.2.1 Importance and Current Status

Historically, users have characterized GeneCards as being their user-friendly "first port of call" to "orient their understanding" when coming across unfamiliar genes. Its popularity encouraged the expansion of the knowledgebase to provide the same functionality for diseases and pathways. Together with this growth came the realization that the depth and breadth of the data itself, while extremely useful in its own right, could be leveraged to solve problems. Today, there is increasing recognition by the scientific community that NGS is a pivotal technology for diagnosing the genetic cause of many human diseases; several large-scale projects implement NGS as a key instrument for elucidating the genetic components of rare diseases and cancer (Bamshad et al. 2012). Other clinical studies aimed at deciphering monogenic and complex diseases have also demonstrated the effectiveness of NGS approaches including whole genome, whole exome, and gene panel sequencing (van den Veyver and Eng 2015; Yang et al. 2013; Gilissen et al. 2014; Zheng et al. 2015; Stranneheim and Wedell 2016). Primary analysis of disease NGS results includes sequence read mapping and variant calling, with results stored in a Variant Call Format (VCF) file.

The VCF file typically contains ~20,000–50,000 positions that differ from the reference genome exome regions ("variant long list"). Subsequently, analysis pipelines sift these SNPs and indels by populating the VCF file with annotation data, such as segregation in affected families, genetic linkage information (Smith et al. 2011), population frequency (Ramos et al. 2012), and missense protein impact (Adzhubei et al. 2010; Sim et al. 2012; Hecht et al. 2015), all facilitating variant filtration (secondary analysis). This helps generate a "variant medium list" of typically dozens to a few hundred entries, depending on the assumed mode of inheritance and on the employed filtering cutoffs. In these analyses, variants are analyzed without regard to the disease phenotype of the sequenced individual. As a first step in introducing phenotype relationships, many pipelines use variant-disease relationships (e.g. from ClinVar (Landrum et al. 2014) and/or COSMIC (Forbes et al. 2015)) for further filtration of the sequence variants. But a typical gene can have a multitude of variants that have not yet been documented to have a relationship with a disease or a phenotype. In many cases, none of the annotated variant-disease relations appears relevant to the sequenced subject. The GeneCards suite's rich knowledgebase facilitates gene-based interpretation. The strategy entails finding disease or phenotype relationships for the gene itself, instead of only for the variant contained within it. VarElect (ve.genecards.org), the suite's web-based phenotype-dependent NGS variant prioritizer, leverages the wealth of information in GeneCards and its affiliated databases. VarElect's algorithm computes prioritized direct (keyword-based) and indirect (inferred from comprehensive gene-to-gene associations) gene/disease connections. The avalanche of variants residing in genomic non-coding "dark matter," available via whole genome sequencing (WGS), contributes three classes of functional genomic elements to variant analyses: promoters, enhancers, and ncRNAs, all central to tissue-related gene expression, with many underlying diseases. Together they amount to >20% of such "novel" DNA territories, unexplored in exome sequencing. Judiciously incorporated into the knowledgebase, the suite's GeneHancer and upgraded ncRNA data is leveraged by its WGS disease interpretation platform and provides a comprehensive route to clinical significance of coding and non-coding single nucleotide and structural genomic variations, often elucidating unsolved clinical cases.

2.2.2 *Future Update and Availability of the Database*

Major synchronized new versions of the suite sites are currently deployed every four months. This weighty effort involves regenerating the gene and diseases lists, updating data from all of the knowledgebase's sources, annotating each of the entities, re-computing the relationships, and quality assurance testing to ensure that all sites are in sync, that data integrity was maintained, and that nothing broke during the process due to changes in source formats and/or other pipeline technicalities. Further, new scientific features are provided by incorporating information from new and/or existing sources and developing/tweaking heuristics and algorithms when warranted. Minor revisions, providing incremental updates for a subset of the data and suite sites, are deployed as needed (typically within 1–2 months), for crucial time-dependent annotations like new publications, localized features, and hot bug fixes. We continue to work on increasing the frequency and content of our releases and expect significant speedup in 2019.

2.3 Content and Architecture of the Database

2.3.1 Main Database Features and Types of Data Stored

Figure 2.1 and Table 2.1 provide an overview of the major entities and relationships in GeneCards and MalaCards, in schematic and tabular forms, respectively. Some of the data include straightforward annotations (e.g. summary information about TP53 from NCBI's Entrez Gene database (Brown et al. 2015), the GeneCards Inferred Functionality Score (GIFtS) for APOA1, the KEGG pathway (Kanehisa et al. 2019) associated with Alzheimer's Disease, companies that provide antibody products for EGFR, publications associated with a gene or disease, and so on). Others reflect sophisticated behind-the-scenes data amalgamation: Compound groups, unified from 12 sources, with drug-specific and drug-gene annotations; GeneHancer (Fishilevich et al. 2017) regulatory element clusters, integrated from 7 sources based on location, with scored GeneHancer elements and GeneHancer-gene annotations; SuperPaths (Belinky et al. 2015), consolidated from 12 sources based on gene content, finding a balance between reducing pathway redundancies and optimizing pathway-related informativeness for individual genes; GeneCards genes (Safran et al. 2010), hierarchically choosing a symbol from HGNC (Yates et al. 2017), Entrez Gene (Brown et al. 2015), Ensembl (Zerbino et al. 2018), or GeneLoc (Rosen et al. 2003), and associating all relevant aliases, descriptions, and external identifiers; MalaCards diseases, canonicalizing, transforming, lexically manipulating, and unifying names from 10 primary and 5 secondary ranked sources (Rappaport et al. 2013).

Data collection methods: The GeneCards data collection process is a pipeline that starts with defining the full set of GeneCards genes, obtained from four primary sources as follows: First, the complete current snapshot of HGNC-approved symbols (Yates et al. 2017) is used as the core gene list. Second, human Entrez Gene (Brown et al. 2015) entries that are different from the HGNC genes are added. Next, human Ensembl (Zerbino et al. 2018) records are matched against the emerging gene list via GeneLoc's exon-based unification algorithm (Rosen et al. 2003); those that are not found to be equivalent to others in the set are included as novel Ensembl-based GeneCards gene entries. Finally, our RNA genes identification and unification facility ((Belinky et al. 2013) and work in progress) adds new ncRNAs not available in the other sources. These primary sources provide annotations for aliases, descriptions, previous symbols, gene category, location, summaries, paralogs, and ncRNA details. Once the gene list is in place with these significant annotations, over 150 data sources, including those noted above and others (Bateman et al. 2017; Gene Ontology Consortium 2015; Smith et al. 2018; Chalifa-Caspi et al. 2004) are mined for thousands of additional descriptors.

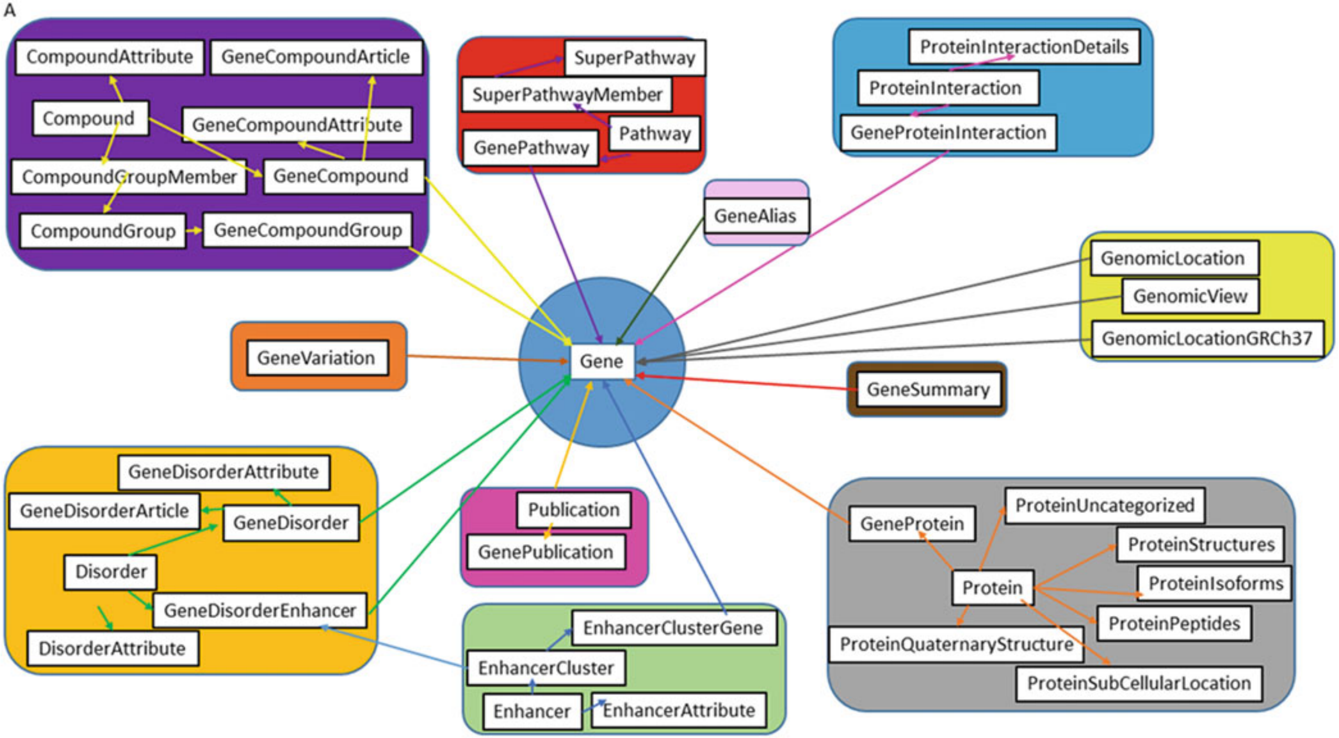


Fig. 2.1 Schematic representation of major GeneCards (**a**) and MalaCards (**b**) entities and relationships. Omitted GeneCards sections include domains, expression, function, localization, orthologs, paralogs, products, sources, and transcripts. Omitted MalaCards sections include summaries, genetic tests, anatomical context, expression, GO terms, and sources

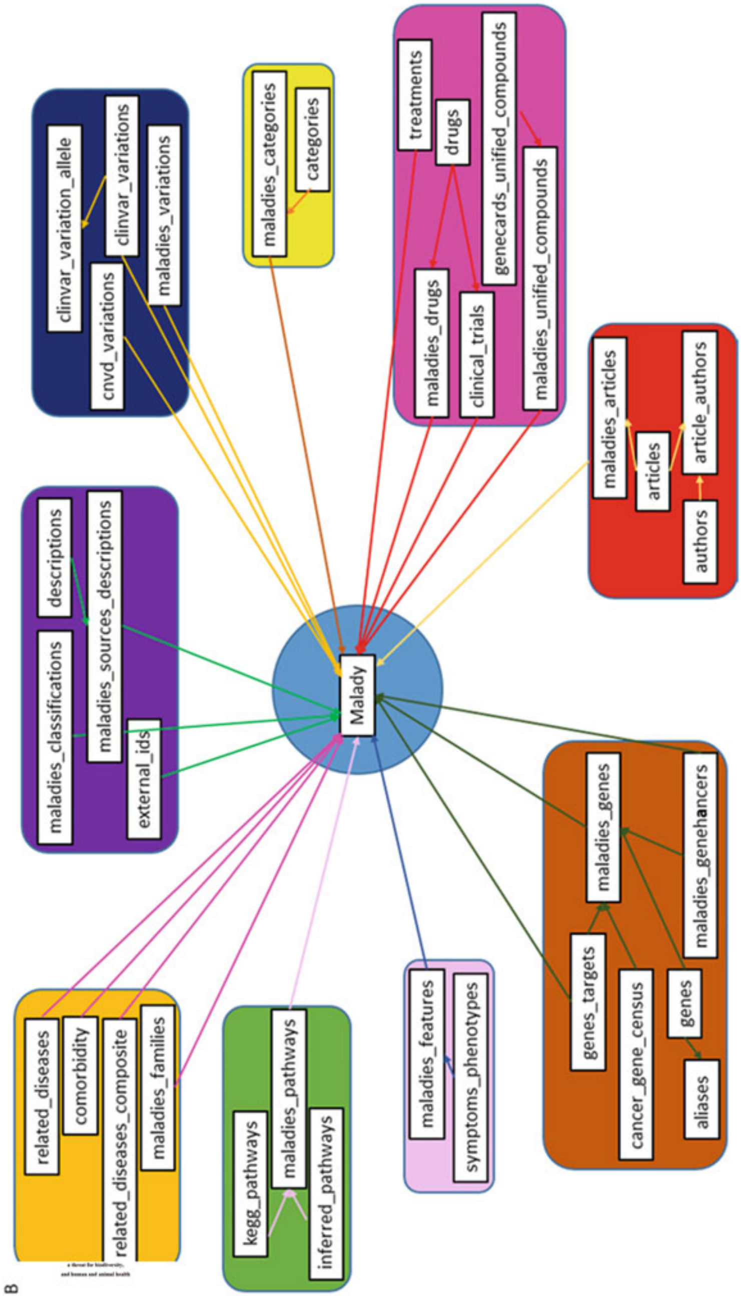


Fig. 2.1 (continued)

Table 2.1 GeneCards and MalaCards entity and relationship tables: (**a**) subset of major entities' tables and their fields; (**b**) types and quantities of tables

(a)

Entity	Table	Fields
Compounds	Compound	id
		sourceId
		sourceAccession
		Name
	CompoundAttribute	id
		compoundId
		type
		Value
	CompoundGroup	id
		Name
	CompoundGroupMember	id
		groupId
		compoundId
Superpathways	SuperPathway	id
		name
		sourceId
		sourceAccession
	SuperPathwayMember	id
		score
		superPathId
		pathwayId
	Pathway	id
		name
		sourceAccession
		sourceID
Enhancers	Enhancer	id
		clusterId
		sourceId
		Chromosome
		Start
		End
		Identifier
		FriendlyName
		HasLink
		Version
		Classification
Maladies	Maladies	id
		symbol
		acronymId
		descId
		slug

(continued)

Table 2.1 (continued)

(a)		
Entity	Table	Fields

(b)	
Category	**Number of tables**
Major entities	57
Relationships	67
Annotation tables	23

MalaCards builds its comprehensive-integrated list of diseases by hierarchically mining heterogeneous, partially overlapping naming sources (15 primary and 29 secondary), unifying disease names and acronyms, initially transforming each name to a canonical form while simultaneously retaining original strings for the alias list. This canonical form is constructed by a series of steps (conversion to lowercase; removal of words like “disease,” “syndrome,” “deficiency,” “failure,” “type,” as well as conjunctions, articles, and prepositions); merging equivalent words (e.g. “juvenile” and “childhood,” “kidney,” and “renal”); handling of different number formats (Roman versus Indian/Arabic), and of plurals and possessives; word stemming, using the porter stemming algorithm (Porter 2006), and others (Rappaport et al. 2013) to enable textual comparison. Diseases with names that are identical except for type specification (e.g. “Alzheimer disease type 3”) are grouped into parent/child families. Once the disease list is in place with these significant annotations, over 70 data sources, including those noted above, and others including GeneCards, MalaCards, and the suite’s gene set analysis capabilities (Ben-Ari Fuchs et al. 2016; Stelzer et al. 2009) are interrogated to yield thousands of additional descriptors and relationships.

2.3.2 *Data Collection and Curation Methods*

The knowledgebase, for the most part, is automatically generated. Our data sources range from those that are manually curated, (e.g. UniProt/SwissProtKB (Bateman et al. 2017)) to those that rely on text mining algorithms (e.g. DISEASES (Pletscher-Frankild et al. 2015)). Our generation software and portals rank the information in the various sections accordingly, giving greater weight to curated over inferred annotations. If the QA process (see below) and/or user feedback uncovers anomalies which cannot immediately be addressed by the relevant sources, we edit the data or use a “cheat list” of corrections to compensate.

2.3.3 *Dataset Indexing/Accession Number/Identification*

Alphabetical human gene and disease database indices appear at the footer of each respective GeneCards and MalaCards page, providing linked lists of symbols/

disease names. Clicking on a letter in the index, say D, brings up a page that lists all genes/diseases that start with "D," each linked to the relevant GeneCard or MalaCard.

GeneCards gene symbols, used in accessing the GeneCards pages for particular genes, are derived from HGNC (Yates et al. 2017), Entrez Gene (Brown et al. 2015), Ensembl (Zerbino et al. 2018), and GeneCards identifiers (GCIDs) (Rosen et al. 2003). GCIDs are unique, informative, and stable, provided by the GeneLoc Algorithm (see http://www.genecards.org/Images/Guide/GeneLocAlgorithm.jpg) as follows.

- The id begins with GC, which is 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.

  For example: OXA1L, with GC id **GC14P022766** is on chromosome **14** on the **plus** strand, starting at **22766** kilobases.
- Genes that are currently placed on a specific chromosome, but whose exact location on the chromosome is not yet known, receive a modified GC id, consisting of the chromosome and strand information, followed by a number, which indicates uncertain location, followed by a letter representing the specific contig containing the gene, and the gene's kilobase position on that contig.

  For example: ENSG00000278198, with GC id **GC07P9O0173** is on chromosome **7** on the **plus** strand of **contig GL000195.1**, starting at **173** kilobases.
- Genes located on the alternative reference sequences (haplotypes—see NCBI (https://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html#assembly) for a full explanation) have a special GC id made up of the chromosome and strand information, followed by a letter, and the gene's approximate kilobase start coordinate.

  For example: KIR2DS5, with GC id **GC19MA00037** is on chromosome **19** on the **minus** strand of **ALT_REF_LOCI_18**, starting at **37** kilobases.
- Genes whose positional information includes only the chromosome need a further modified GC id, which includes the chromosome number, followed by "U9," indicating lack of strand and positional information, followed by five digits, assigned sequentially.

  For example: GUK2, with GC id **GC01U990078** is on chromosome **1**. Its **strand** and **position** are currently unknown.

  If an id needs to change in future versions because the previously reported position is refined, the superseded id remains associated with the gene, along with the new one, so it cannot be assigned to any other gene, and so that users can still find the gene by that id.

MalaCards identifiers, used in its URLs, are its *main disease names* supplied by primary sources (Rappaport et al. 2013) (e.g. Pick Disease) converted to lowercase, with spaces replaced by underscores (pick_disease for this example). To be as consistent as possible across versions, all such URLs are preserved, even if the disease name has changed or the disease was merged with another. In situations like these, old URLs are redirected to new ones. If a disease was removed completely from MalaCards, the old link is redirected to the search results page generated by querying the old disease name. In addition, a unique internal MCID is generated for each malady, composed of the first letter of its name, followed by the next two consonants, followed by a sequence number. For example, the MCID for "rett syndrome" is RTT001.

PathCards SuperPath identifiers, used in its URLs, are the names of the SuperPaths (e.g. glucose metabolism) converted to lowercase with spaces replaced by underscores (glucose_metabolism for this example).

2.3.4 Quality Control Methods

Before releasing a version of the knowledgebase, the system undergoes a semi-automated QA process. An in-house tool verifies the integrity of the GeneCards database by comparing it with that of the previous version, and it highlights inconsistencies and extreme results. The anomalies are then manually reviewed. Web cards and their links for a sample set of genes and diseases are manually checked by our QA professionals and a medical doctor consultant. As our heuristics are still evolving, problematic disease names (e.g. "Interferon" or "memory") are entered into a "cheat list" and removed from the system. VarElect and GeneHancer have their own set of automated QA scenarios, wherein deviations from expected results are reported and followed up by manual scrutiny. Test scenarios, bugs, and suggestions for improvements are all ticketed in our JIRA tracking system (https://www.atlassian.com/software/jira) and mapped to target releases.

2.3.5 Database Update and Maintenance Strategy

The knowledgebase is regenerated from scratch for each major version. For incremental updates, source-specific generation modules are rerun using the latest data. In both situations, the search index is regenerated for the benefit of the database portals themselves, as well as for usage by VarElect and TGex.

2.4 Database Access and Mining Methods

2.4.1 Tools and Techniques to Access, Discover, and Mine the Content of the Database

Gene-centric, disease-centric, location-centric, and pathway-centric information are, respectively, available and searchable from the GeneCards, MalaCards, GeneLoc, and PathCards portals, each with their own entity-specific web "card" and powerful search engine. GeneHancer data is incorporated in the knowledgebase, and in GeneCards, MalaCards, GeneLoc, VarElect, and TGex. The extensive knowledgebase (Ben-Ari Fuchs et al. 2016) is exploited to provide NGS interpretation and gene set analysis solutions as follows:

2.4.1.1 VarElect: The NGS Phenotyper of the GeneCards Suite

A key challenge in the interpretation of NGS in genetic disease studies is to effectively associate the identified variant-containing genes with a patient's disease phenotypes. This is addressed by VarElect (Stelzer et al. 2016b), the GeneCards Suite powered NGS interpretation tool, leveraging the broad knowledgebase for gene prioritization. VarElect is a comprehensive search tool that helps to effectively and rapidly identify and prioritize direct and indirect associations between genes and user-supplied disease terms, joined with providing extensive evidence for such associations.

Typical NGS analyses of a patient discover tens of thousands non-reference coding single nucleotide variants (SNVs), but only one or very few are expected to be significant for the relevant disease. In a filtering stage, various approaches, such as family segregation, frequency in the population, predicted protein impact, and evolutionary conservation are combined to shorten the variant list. A major challenge is the interpretation of the remaining (typically) few hundred genes, aiming to further focus on the most viable disease-causing candidate genes.

To cope with genes that have no direct association to the phenotype terms on their own, VarElect infers indirect (or "guilt by association") relationships between genes and phenotype keywords exploiting the GeneCards Suite diverse gene-to-gene relationships. Gene-to-gene relationships are generated using the GeneCards search engine, by searching gene symbols in selected GeneCards sections. The integrated pathway information from PathCards is a major contribution to the gene-to-gene relationships.

2.4.1.2 TGex: The Knowledge-Driven Clinical Genetics Analysis Platform of the GeneCards Suite

Clinical genetics analysis of thousands of variants requires a user interface that will enable browsing, viewing, filtering, and interpretation interactively. To this aim, TGex, the GeneCards Suite Knowledge-Driven Clinical Genetics Analysis platform, combines VarElect strength with comprehensive variant annotation and filtering capabilities in a consolidated view, which enables the genetic analyst to quickly pinpoint the strongest candidates. The comprehensive reporting system of TGex leverages the capabilities of VarElect and the vast amount of structured data available in the GeneCards Suite to automatically generate a full clinical report. TGex supports comprehensive data scrutiny, from raw patient genetic data (a VCF file), through intermediate annotations and interpretations, to detailed final reports.

2.4.1.3 Analysis of Genomic Structural Variants (SVs) Enabled by GeneHancer

A major source of pathogenic genomic alterations are structural variants (SVs), comprising both balanced modifications (inversions and translocations) and unbalanced variations—copy number variants (CNVs), including deletions, duplications, and insertions (Hurles et al. 2008; Weischenfeldt et al. 2013). Evaluation of the impact of SVs with respect to phenotype or disease relies on the genomic functional units associated with the SVs. Disease-related functional consequences of SVs involve changes in gene expression, which might occur when the SV encompasses the gene territory, either completely or partially. In this vein, the GeneCards Suite tools are useful for SVs interpretation, by helping to identify and prioritize SVs using the potential disease-causing genes damaged in each SV.

Often SVs do not overlap the coding regions of the disease-associated gene. SVs might influence genes over large distances by altering non-coding functional components such as regulatory elements and non-coding RNA genes. Tackling variations in non-coding regulatory elements to decipher the genetic underpinnings of human diseases is a great challenge in the analysis of both SNVs and SVs. Addressing this challenge necessitates the ability to map variants to regulatory elements such as promoters and enhancers. The mapping program requires access to a comprehensive database of regulatory elements. Since the biomedical knowledge directly linking regulatory elements to a disease/phenotype is obscure, the variant mapping step needs to be complemented by annotative information regarding a relationship between such an element and its target gene, for which a phenotype relationship is already known.

These capabilities are the core of GeneHancer, the GeneCards Suite database of regulatory elements and their gene targets. GeneHancer's comprehensive-integrated and scored set of regulatory elements and their gene-associations enables translating the finding of a WGS variant in a non-coding region into a variant-to-gene annotation, along with a confidence indication. Thus, integrating GeneHancer into the WGS annotation and filtering functions of VarElect and TGex assists in the mapping of non-coding variants to regulatory elements and via the gene targets forms a basis for variant-phenotype interpretation of whole genome sequences in health and disease.

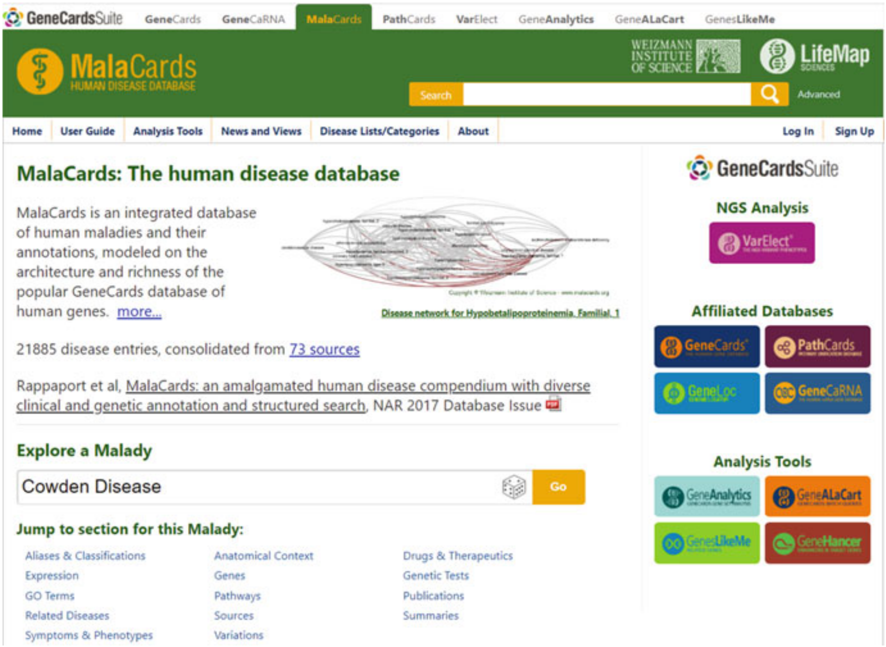
2.4.1.4 Gene Set Enrichment Analysis

GeneAnalytics (Ben-Ari Fuchs et al. 2016) is an analysis tool for finding commonalities within gene sets resulting from NGS, RNAseq, and microarray experiments. Using in-depth evidence-based scoring algorithms and taking advantage of the

GeneCards Suite knowledgebase, GeneAnalytics identifies cell types, diseases, pathways, and functions related to the gene set and provides supporting evidence links for matched biological terms in the GeneCards Suite.

2.4.2 *How to Explore and Browse the Database*

We illustrate exploring and browsing of the various suite sites by describing the MalaCards (Rappaport et al. 2014) compendium of human diseases portal (www.malacards.org), which features ~22,000 human diseases, with annotations integrated from 73 sources and shown in 14 sections. The homepage (Fig. 2.2a) is a common entry point to the Web site, showcasing most of the features and tools including exploring a particular (sample, random, or specified) malady, jumping to a particular section within it, quick searches, a disease index, statistics, a menu bar with links to documentation and disease list/category pages, and links to the other GeneCards



MalaCards: The human disease database

MalaCards is an integrated database of human maladies and their annotations, modeled on the architecture and richness of the popular GeneCards database of human genes. more...

21885 disease entries, consolidated from 73 sources

Rappaport et al, MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search, NAR 2017 Database Issue

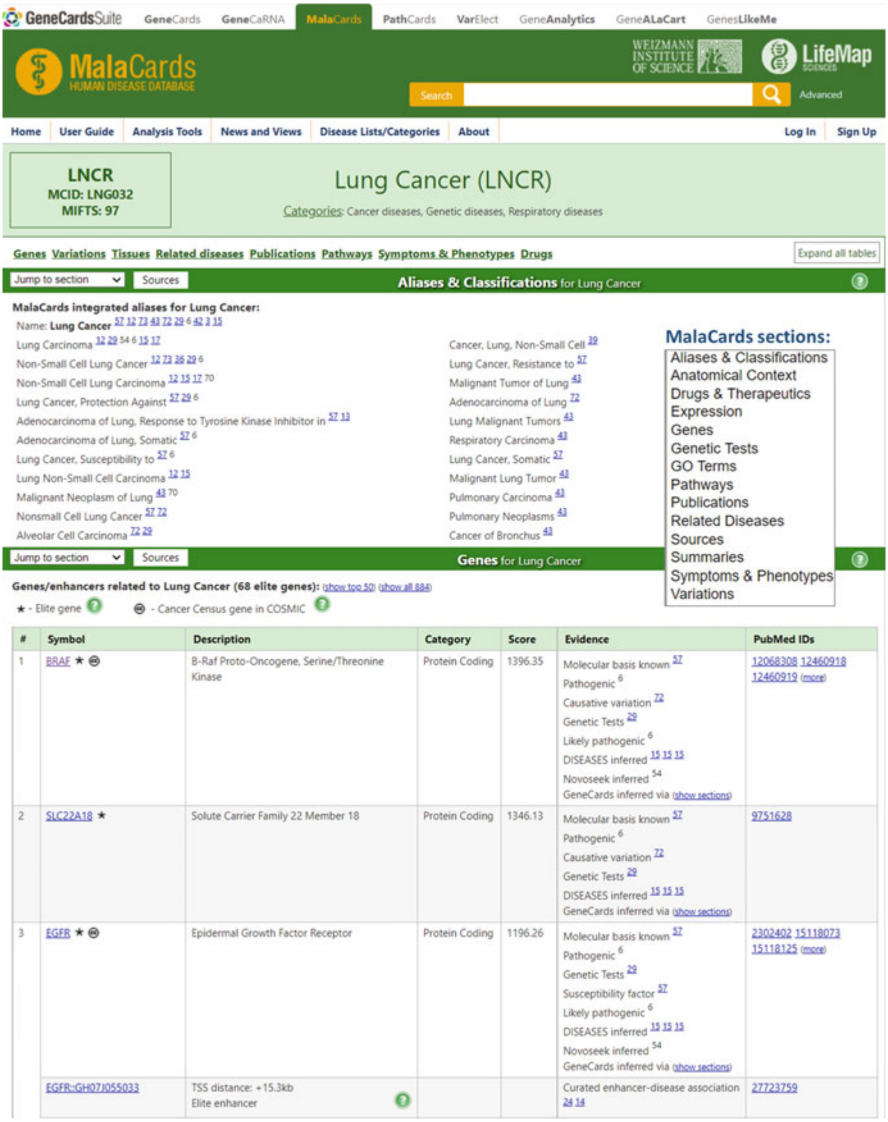
Explore a Malady

Cowden Disease

Jump to section for this Malady:

Aliases & Classifications
Expression
GO Terms
Related Diseases
Symptoms & Phenotypes
Anatomical Context
Genes
Pathways
Sources
Variations
Drugs & Therapeutics
Genetic Tests
Publications
Summaries

Fig. 2.2 (**a**) The homepage of MalaCards, the human disease database. (**b**) The MalaCard for Lung Cancer includes the Genes section, which provides the list of the affiliated genes and enhancers found to be associated with the disease. MalaCards "elite" genes (marked with *) are those likely to be associated with causing the disease, since their gene-disease associations are supported by manually curated and trustworthy sources. The cancer COSMIC Gene Census list is an ongoing effort to catalog those genes for which mutations have been causally implicated in cancer. Cancer census gene list genes are marked with a CC icon



LNCR
MCID: LNG032
MIFTS: 97

Lung Cancer (LNCR)

Categories: Cancer diseases, Genetic diseases, Respiratory diseases

Aliases & Classifications for Lung Cancer

MalaCards integrated aliases for Lung Cancer:

Name: **Lung Cancer** [57, 12, 73, 43, 72, 29, 6, 42, 3, 15]

Lung Carcinoma [12, 29, 54, 6, 15, 17]
Non-Small Cell Lung Cancer [12, 73, 36, 29, 6]
Non-Small Cell Lung Carcinoma [12, 15, 17, 70]
Lung Cancer, Protection Against [57, 29, 6]
Adenocarcinoma of Lung, Response to Tyrosine Kinase Inhibitor in [57, 13]
Adenocarcinoma of Lung, Somatic [57, 6]
Lung Cancer, Susceptibility to [57, 6]
Lung Non-Small Cell Carcinoma [12, 15]
Malignant Neoplasm of Lung [43, 70]
Nonsmall Cell Lung Cancer [57, 72]
Alveolar Cell Carcinoma [72, 29]
Cancer, Lung, Non-Small Cell [39]
Lung Cancer, Resistance to [57]
Malignant Tumor of Lung [43]
Adenocarcinoma of Lung [72]
Lung Malignant Tumors [43]
Respiratory Carcinoma [43]
Lung Cancer, Somatic [57]
Malignant Lung Tumor [43]
Pulmonary Carcinoma [43]
Pulmonary Neoplasms [43]
Cancer of Bronchus [43]

Genes for Lung Cancer

Genes/enhancers related to Lung Cancer (68 elite genes): (show top 50) (show all 884)

★ - Elite gene   ㊙ - Cancer Census gene in COSMIC

#	Symbol	Description	Category	Score	Evidence	PubMed IDs
1	BRAF ★ ㊙	B-Raf Proto-Oncogene, Serine/Threonine Kinase	Protein Coding	1396.35	Molecular basis known [57] Pathogenic [6] Causative variation [72] Genetic Tests [29] Likely pathogenic [6] DISEASES inferred [15, 15, 15] Novoseek inferred [54] GeneCards inferred via (show sections)	12068308 12460918 12460919 (more)
2	SLC22A18 ★	Solute Carrier Family 22 Member 18	Protein Coding	1346.13	Molecular basis known [57] Pathogenic [6] Causative variation [72] Genetic Tests [29] DISEASES inferred [15, 15, 15] GeneCards inferred via (show sections)	9751628
3	EGFR ★ ㊙	Epidermal Growth Factor Receptor	Protein Coding	1196.26	Molecular basis known [57] Pathogenic [6] Genetic Tests [29] Susceptibility factor [57] Likely pathogenic [6] DISEASES inferred [15, 15, 15] Novoseek inferred [54] GeneCards inferred via (show sections)	2302402 15118073 15118125 (more)
	EGFR::GH07J055033	TSS distance: +15.3kb Elite enhancer			Curated enhancer-disease association [24, 14]	27723759

Fig. 2.2 (continued)

Suite members. MalaCards can be navigated in a variety of ways. The search box is typically the initial starting point, where one can submit free text as a query string, including Boolean expressions. It is centrally located on the homepage, as well as at the top right corner of every page comprising the Web site.

A MalaCards disease page (Web "card" or simply MalaCard) is where one can find all available information pertaining to a disease of interest. The information within a MalaCard is divided into 14 sections: Aliases and Classifications, Summaries, Related Diseases, Symptoms and Phenotypes, Drugs and Therapeutics, Genetic Tests, Anatomical Context, Publications, Genes, Variations, Expression, Pathways, GO Terms, and Sources. Documentation is accessible via hyperlinks, often context-specific, from within many parts of the MalaCard, to the right of the section, by clicking on the question mark icon. Each section displays disease-specific information and contains deep links to supporting sources, often with superscripts when multiple sources contain details about the datum. 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. Figure 2.2b shows portions of the MalaCard for Lung Cancer, including the Genes section, which provides the list of the affiliated genes and enhancers found to be associated with the disease. MalaCards “elite” genes (marked with *) are those likely to be associated with causing the disease, since their gene-disease associations are supported by manually curated and trustworthy sources. The cancer Gene Census list from COSMIC is an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer. Genes listed in the cancer census gene list are marked with a CC icon. When relevant, shown GeneHancers are genomic regulatory elements-gene-disease associations provided by GeneHancer. Initially, at least 10 affiliated genes are shown (all of the elite genes are always shown), with an option to see the complete list.

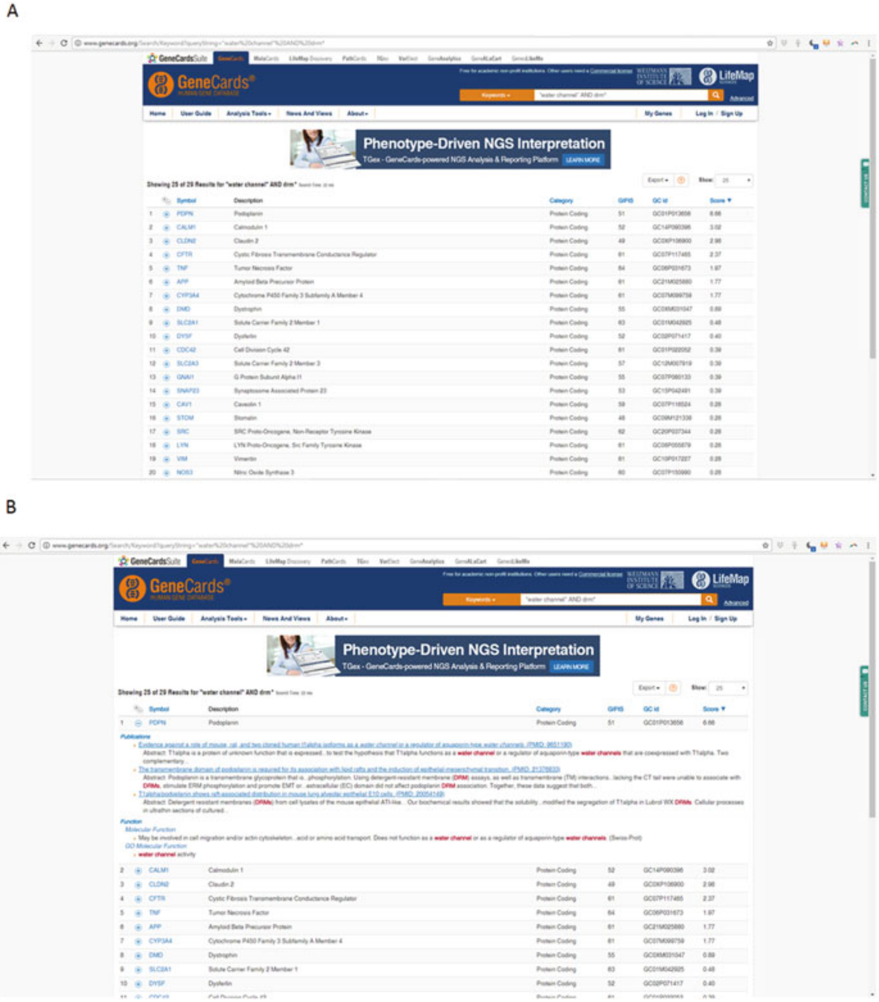
The ranked genes list is composed by taking into account: (1) genetic testing resources supplying specific genetic tests for the disease: (2) genetic variations resources supplying specific causative variations in genes for the disease; (3) resources that manually curate the association of the disease with genes; (4) searches within GeneCards, providing inferred associations.

The section’s genes table shows gene symbols, descriptions, category, relevance scores, the context according to which the gene is related to the disease, and Pubmed ids. The relevance score is computed by factoring in the importance of the different resources associating the gene with the disease.

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 “Expand all tables” activates “see all” in all of the sections and enables convenient searches within the card.

2.4.3 How to Query the Database

We illustrate the search capabilities of the various suite sites by describing GeneCards searches. In the top right corner of the GeneCards banner on each of its pages, enter your search terms into the search box and click the magnifying glass icon to submit the query. 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



Showing 25 of 29 Results for "water channel" AND drm*

	Symbol	Description	Category	GIFtS	GC id	Score ▼
1	PDPN	Podoplanin	Protein Coding	51	GC01P013656	6.66
2	CALM1	Calmodulin 1	Protein Coding	52	GC14P090396	3.02
3	CLDN2	Claudin 2	Protein Coding	49	GC0XP106900	2.98
4	CFTR	Cystic Fibrosis Transmembrane Conductance Regulator	Protein Coding	61	GC07P117465	2.37
5	TNF	Tumor Necrosis Factor	Protein Coding	64	GC06P031673	1.97
6	APP	Amyloid Beta Precursor Protein	Protein Coding	61	GC21M025880	1.77
7	CYP3A4	Cytochrome P450 Family 3 Subfamily A Member 4	Protein Coding	61	GC07M099758	1.77
8	DMD	Dystrophin	Protein Coding	55	GC0XM031047	0.89
9	SLC2A1	Solute Carrier Family 2 Member 1	Protein Coding	63	GC01M042925	0.48
10	DYSF	Dysferlin	Protein Coding	52	GC02P071417	0.40
11	CDC42	Cell Division Cycle 42	Protein Coding	61	GC01P022052	0.39
12	SLC2A3	Solute Carrier Family 2 Member 3	Protein Coding	57	GC12M007919	0.39
13	GNAI1	G Protein Subunit Alpha I1	Protein Coding	55	GC07P080133	0.39
14	SNAP23	Synaptosome Associated Protein 23	Protein Coding	53	GC15P042491	0.39
15	CAV1	Caveolin 1	Protein Coding	58	GC07P116524	0.28
16	STOM	Stomatin	Protein Coding	48	GC09M121338	0.28
17	SRC	SRC Proto-Oncogene, Non-Receptor Tyrosine Kinase	Protein Coding	62	GC20P037344	0.28
18	LYN	LYN Proto-Oncogene, Src Family Tyrosine Kinase	Protein Coding	61	GC08P055879	0.28
19	VIM	Vimentin	Protein Coding	61	GC10P017227	0.28
20	NOS3	Nitric Oxide Synthase 3	Protein Coding	60	GC07P150990	0.28

Showing 25 of 29 Results for "water channel" AND drm*

	Symbol	Description	Category	GIFtS	GC id	Score ▼
1	PDPN	Podoplanin	Protein Coding	51	GC01P013656	6.66

Publications

- Evidence against a role of mouse, rat, and two cloned human T1alpha isoforms as a water channel or a regulator of aquaporin-type water channels. (PMID: 9651190)
  Abstract: T1alpha is a protein of unknown function that is expressed...to test the hypothesis that T1alpha functions as a water channel or a regulator of aquaporin-type water channels that are coexpressed with T1alpha. Two complementary...
- The transmembrane domain of podoplanin is required for its association with lipid rafts and the induction of epithelial-mesenchymal transition. (PMID: 21376833)
  Abstract: Podoplanin is a transmembrane glycoprotein that is...phosphorylation. Using detergent-resistant membrane (DRM) assays, as well as transmembrane (TM) interactions...lacking the CT tail were unable to associate with DRMs, stimulate ERM phosphorylation and promote EMT or...extracellular (EC) domain did not affect podoplanin DRM association. Together, these data suggest that both...
- T1alpha/podoplanin shows raft-associated distribution in mouse lung alveolar epithelial E10 cells. (PMID: 20054149)
  Abstract: Detergent resistant membranes (DRMs) from cell lysates of the mouse epithelial ATI-like... Our biochemical results showed that the solubility...modified the segregation of T1alpha in Lubrol WX DRMs. Cellular processes in ultrathin sections of cultured...

Function

Molecular Function

- May be involved in cell migration and/or actin cytoskeleton...acid or amino acid transport. Does not function as a water channel or as a regulator of aquaporin-type water channels. (Swiss-Prot)

GO Molecular Function

- water channel activity

	Symbol	Description	Category	GIFtS	GC id	Score ▼
2	CALM1	Calmodulin 1	Protein Coding	52	GC14P090396	3.02
3	CLDN2	Claudin 2	Protein Coding	49	GC0XP106900	2.98
4	CFTR	Cystic Fibrosis Transmembrane Conductance Regulator	Protein Coding	61	GC07P117465	2.37
5	TNF	Tumor Necrosis Factor	Protein Coding	64	GC06P031673	1.97
6	APP	Amyloid Beta Precursor Protein	Protein Coding	61	GC21M025880	1.77
7	CYP3A4	Cytochrome P450 Family 3 Subfamily A Member 4	Protein Coding	61	GC07M099758	1.77
8	DMD	Dystrophin	Protein Coding	55	GC0XM031047	0.89
9	SLC2A1	Solute Carrier Family 2 Member 1	Protein Coding	63	GC01M042925	0.48
10	DYSF	Dysferlin	Protein Coding	52	GC02P071417	0.40

Fig. 2.3 MalaCards search results: (**a**) sorted, scored gene hits. (**b**) with Minicards including hit context

must be capitalized to yield expected results: For example, specifying "*water channel" AND drm** yields 29 results.

Searches result in a list of genes, each with its description, category, GeneCards Inferred Functionality Score (GIFtS) (Harel et al. 2009), and GeneCards identifier (Rosen et al. 2003), sorted by Elastic search relevance score (Fig. 2.3a). Clicking the plus to the left of the symbol opens a "MiniCard," which shows the hit context of the search terms (Fig. 2.3b). Clicking on the symbol opens the gene's card.

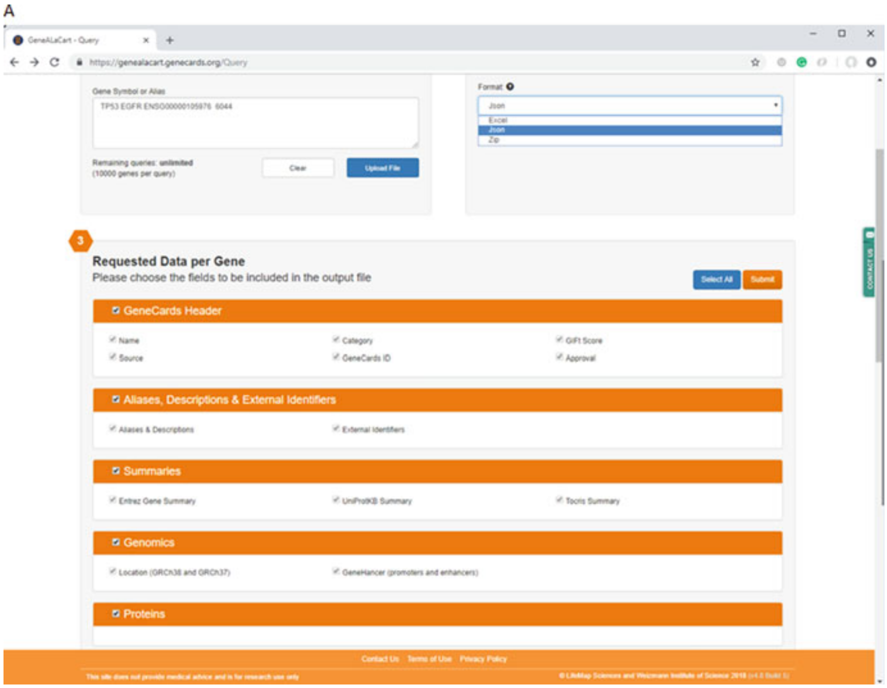
GeneCards can also be searched for a specific symbol, using the search dropdown (choose "Symbols"). When searching for a symbol that might not be the gene's official symbol (from a paper, for example), and when using a gene identifier from another database, the other dropdown options should be used ("Symbols/Aliases" and "Symbols/Aliases/Identifiers," respectively).

To use the GeneCards advanced search, click on the "Advanced" link to the right of the search box. The advanced search allows complex queries in which each keyword can be restricted to a specific section of the GeneCard.

MalaCards and PathCards have similar querying facilities.

2.4.4 *How to Upload/Download Data*

Registered users have a variety of download facilities. GeneALaCart (https://genealacart.genecards.org/), the GeneCards batch query portal generates a file of GeneCards annotations associated with input gene lists. For each query, one supplies the "batch" of gene symbols or identifiers and selects the annotations of interest (Fig. 2.4a); GeneALaCart then extracts the information from the knowledgebase and produces a customized results file in Excel [Fig. 2.4b] or JSON format [Fig. 2.4c].


A
GeneALaCart - Query
https://genealacart.genecards.org/Query
Clear
Upload File
Select All
Submit
CONTACT US
Contact Us Terms of Use Privacy Policy
This site does not provide medical advice and is for research use only
© LifeMap Sciences and Weizmann Institute of Science 2018

Gene Symbol or Alias: TP53 EGFR ENSG00000105976 6044

Remaining queries: **unlimited**
(10000 genes per query)

Format: Json
- Excel
- Json
- Zip

3 **Requested Data per Gene**
Please choose the fields to be included in the output file

☑ GeneCards Header
- ✓ Name
- ✓ Category
- ✓ GIFt Score
- ✓ Source
- ✓ GeneCards ID
- ✓ Approval

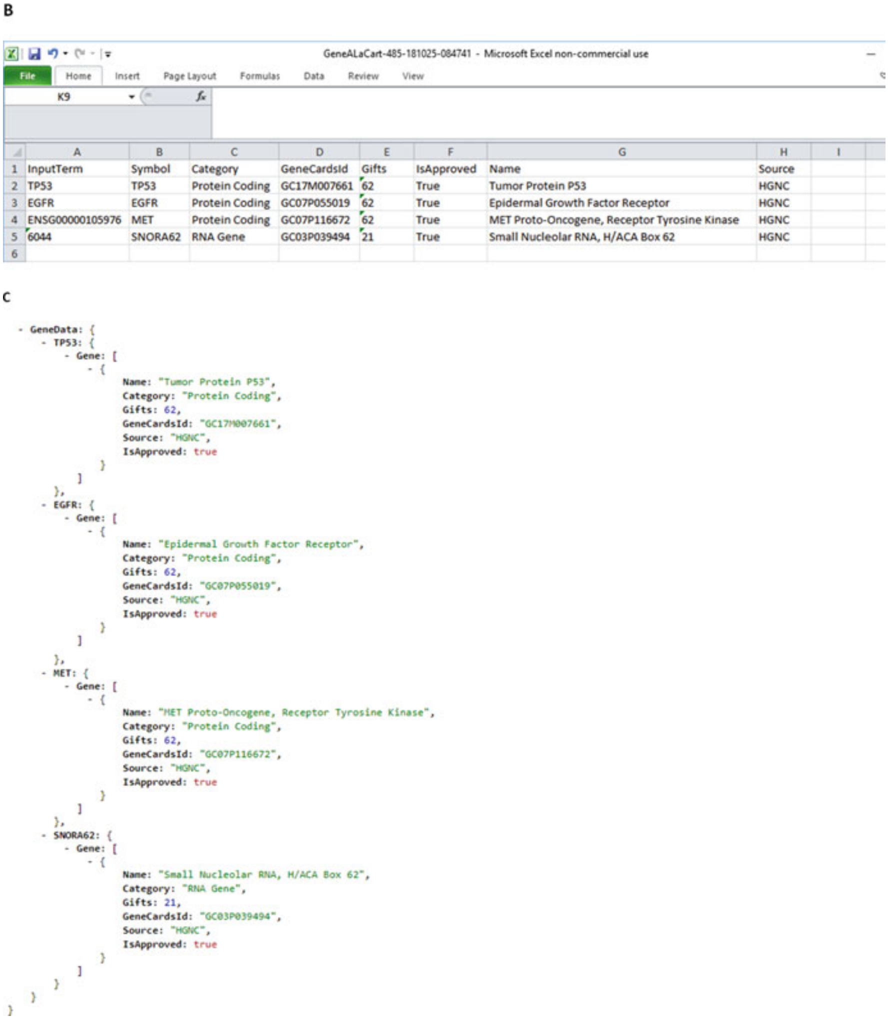
☑ Aliases, Descriptions & External Identifiers
- ✓ Aliases & Descriptions
- ✓ External Identifiers

☑ Summaries
- ✓ Entrez Gene Summary
- ✓ UniProtKB Summary
- ✓ Tocris Summary

☑ Genomics
- ✓ Location (GRCh38 and GRCh37)
- ✓ GeneHancer (promoters and enhancers)

☑ Proteins

Fig. 2.4 GeneAlaCart input and output: (**a**) user inputs genes/identifiers of choice, selected annotations, and output file format; (**b**) sample Excel sheet output; (**c**) sample JSON output


B
GeneALaCart-485-181025-084741 - Microsoft Excel non-commercial use
File
Home
Insert
Page Layout
Formulas
Data
Review
View
K9
C

	A	B	C	D	E	F	G	H	I
1	InputTerm	Symbol	Category	GeneCardsId	Gifts	IsApproved	Name	Source	
2	TP53	TP53	Protein Coding	GC17M007661	62	True	Tumor Protein P53	HGNC	
3	EGFR	EGFR	Protein Coding	GC07P055019	62	True	Epidermal Growth Factor Receptor	HGNC	
4	ENSG00000105976	MET	Protein Coding	GC07P116672	62	True	MET Proto-Oncogene, Receptor Tyrosine Kinase	HGNC	
5	6044	SNORA62	RNA Gene	GC03P039494	21	True	Small Nucleolar RNA, H/ACA Box 62	HGNC	
6									

```
 - GeneData: {
     - TP53: {
         - Gene: [
             - {
                     Name: "Tumor Protein P53",
                     Category: "Protein Coding",
                     Gifts: 62,
                     GeneCardsId: "GC17M007661",
                     Source: "HGNC",
                     IsApproved: true
                 }
             ]
         },
     - EGFR: {
         - Gene: [
             - {
                     Name: "Epidermal Growth Factor Receptor",
                     Category: "Protein Coding",
                     Gifts: 62,
                     GeneCardsId: "GC07P055019",
                     Source: "HGNC",
                     IsApproved: true
                 }
             ]
         },
     - MET: {
         - Gene: [
             - {
                     Name: "MET Proto-Oncogene, Receptor Tyrosine Kinase",
                     Category: "Protein Coding",
                     Gifts: 62,
                     GeneCardsId: "GC07P116672",
                     Source: "HGNC",
                     IsApproved: true
                 }
             ]
         },
     - SNORA62: {
         - Gene: [
             - {
                     Name: "Small Nucleolar RNA, H/ACA Box 62",
                     Category: "RNA Gene",
                     Gifts: 21,
                     GeneCardsId: "GC03P039494",
                     Source: "HGNC",
                     IsApproved: true
                 }
             ]
         }
     }
}
```

Fig. 2.4 (continued)

Other download capabilities within the suite sites include exporting GeneCards search results, details about MalaCards diseases, GeneLikeMe functional partners with evidence, GeneHancer details, VarElect prioritized results, GeneAnalytics enriched gene sets, and TGex annotated reports. Facilities for database acquisition for the purposes of further analyses and integration include a variety of knowledgebase dumps and APIs. For more details, please contact the authors.

2.5 Use Cases

As noted above, discovery within the GeneCards Suite is exemplified by how VarElect and TGex leverage the extensive knowledgebase to provide NGS interpretation. The following use cases illustrate this.

2.5.1 Interpretation of Single Nucleotide Variants (SNVs)

VarElect is useful for variant interpretation in genetic disease studies by helping to identify and prioritize associations between variant-containing genes and phenotype keywords. VarElect helped us solve clinical cases in our own laboratory (Alkelai et al. 2016, 2017; Oz-Levi et al. 2015; Heimer et al. 2016, 2018) and was further used in numerous studies worldwide (Yang et al. 2017; Einhorn et al. 2017; Ekhilevitch et al. 2016; Jia et al. 2017; Bafunno et al. 2018; Zhang et al. 2016; Azim et al. 2019; Carneiro et al. 2018; Feliubadalo et al. 2017; Syama et al. 2018). VarElect exploits the GeneCards Suite diverse gene-to-gene relationships to pinpoint the relevance of genes that have no direct association to the phenotype keywords on their own (using the indirect, or "guilt by association" mode). The indirect approach proved crucial to solving a case of systemic capillary leak syndrome (Stelzer et al. 2016b). Figure 2.5a depicts an example of another VarElect case solved in our group (Rappaport et al. 2017b). In this example, the genome of a 6 year old boy, who suffered from atypical epilepsy combined with retinitis pigmentosa, was sequenced. Eighty-one rare homozygous variants, which were heterozygous in both parents, were identified in the patient. The list of 63 variant-containing genes was submitted to VarElect, along with the phenotype search terms; "epilepsy OR macular OR retinitis." VarElect's top scoring gene was *CLN6*. The patient had a homozygous missense variation (V148D) in this gene with zero population frequency and a high predicted protein damage impact. Following this discovery, the patient was clinically diagnosed with accuracy, enabling appropriate genetic counseling and preimplantation diagnosis for the family in the event of future pregnancies.

VarElect can be used stand-alone as described above, or within TGex, the GeneCards Suite Knowledge-Driven Clinical Genetics Analysis platform. TGex requires two inputs: (Rebhan et al. 1997) A VCF file; (Stelzer et al. 2016a) disease/phenotype/symptom terms for VarElect gene-phenotype interpretation. With TGex (Fig. 2.5b), thousands of variants within the uploaded patient VCF file are analyzed in an interactive web-based interface, allowing the user to browse, view, and filter input variants. Those capacities are combined with VarElect's gene-phenotype interpretation strength, allowing one to effectively identify disease-causing candidates. Top candidate variants, along with disease association evidence, are automatically pulled into the detailed clinical report.

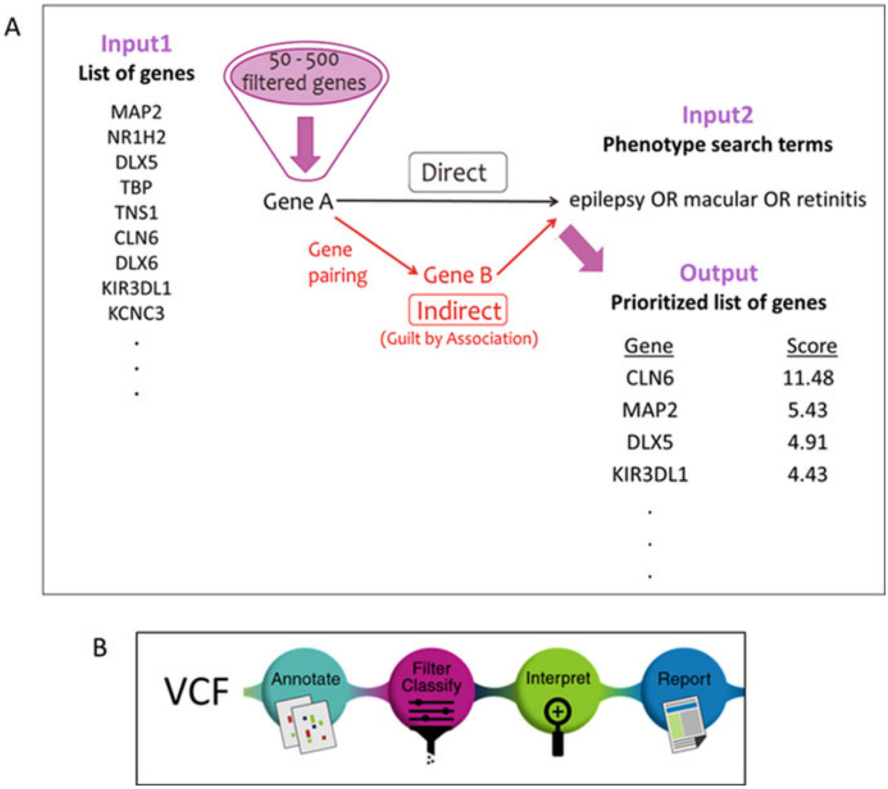
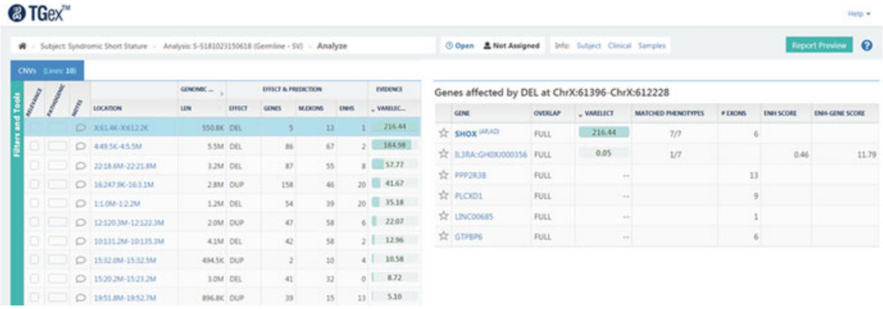


Fig. 2.5 The GeneCards Suite NGS analysis tools VarElect and TGex. (**a**) Example of a VarElect case solved in our group (Rappaport et al. 2017b); (**b**) NGS data analysis with TGex. TGex allows data scrutiny and analysis, starting from raw patient genetic data (a VCF file) to a detailed report. Variants are annotated using information from the GeneCards knowledgebase, allowing interactive filtering. These variant annotation and filtering steps are strengthened by gene-phenotype interpretation using VarElect. Hence, TGex allows the examination of variants using both variant-based annotations and variant-containing-genes-based interpretation, presenting this information for optimal candidate variant selection for the clinical report

2.5.2 Interpretation of Genomic Structural Variants (SVs)

VarElect is useful for structural variants interpretation by the identification and prioritization of SVs via the potential disease-causing genes damaged by each SV. In this workflow, the gene list submitted to VarElect includes genes residing (completely or partially) within the detected SVs. This mode of analysis using VarElect helped solve a number of cases (Homma et al. 2018; Fidalgo et al. 2016). One study aimed to diagnose recurrent CNVs associated with syndromic short stature of unknown cause (Homma et al. 2018). Two hundred and twenty-nine patients were genotyped by chromosomal microarray analysis, leading to identification of candidate CNVs. The gene content of those CNVs was submitted to VarElect



Subject: Syndromic Short Stature › Analysis: S-S181023150618 (Germline - SV) › Analyze

CNVs (Lines: 10)

RELEVANCE	PATHOGENIC	NOTES		GENOMIC ...		EFFECT & PREDICTION			EVIDENCE
			LOCATION	LEN	EFFECT	GENES	M.EXONS	ENHS	VARELEC...
			X:61.4K-X:612.2K	550.8K	DEL	5	13	1	216.44
□	□		4:49.5K-4:5.5M	5.5M	DEL	86	67	2	184.98
□	□		22:18.6M-22:21.8M	3.2M	DEL	87	55	8	57.77
□	□		16:247.9K-16:3.1M	2.8M	DUP	158	46	20	41.67
□	□		1:1.0M-1:2.2M	1.2M	DEL	54	39	20	35.18
□	□		12:120.3M-12:122.3M	2.0M	DUP	47	58	6	22.07
□	□		10:131.2M-10:135.3M	4.1M	DEL	42	58	2	12.96
□	□		15:32.0M-15:32.5M	494.5K	DUP	2	10	4	10.58
□	□		15:20.2M-15:23.2M	3.0M	DEL	41	32	0	8.72
□	□		19:51.8M-19:52.7M	896.8K	DUP	39	15	13	5.10

Genes affected by DEL at ChrX:61396-ChrX:612228

	GENE	OVERLAP	VARELECT	MATCHED PHENOTYPES	# EXONS	ENH SCORE	ENH-GENE SCORE
☆	SHOX (AR,AD)	FULL	216.44	7/7	6		
☆	IL3RA:GH0XJ000356	FULL	0.05	1/7		0.46	11.79
☆	PPP2R3B	FULL	--		13		
☆	PLCXD1	FULL	--		9		
☆	LINC00685	FULL	--		1		
☆	GTPBP6	FULL	--		6		

Fig. 2.6 SV analysis with TGex

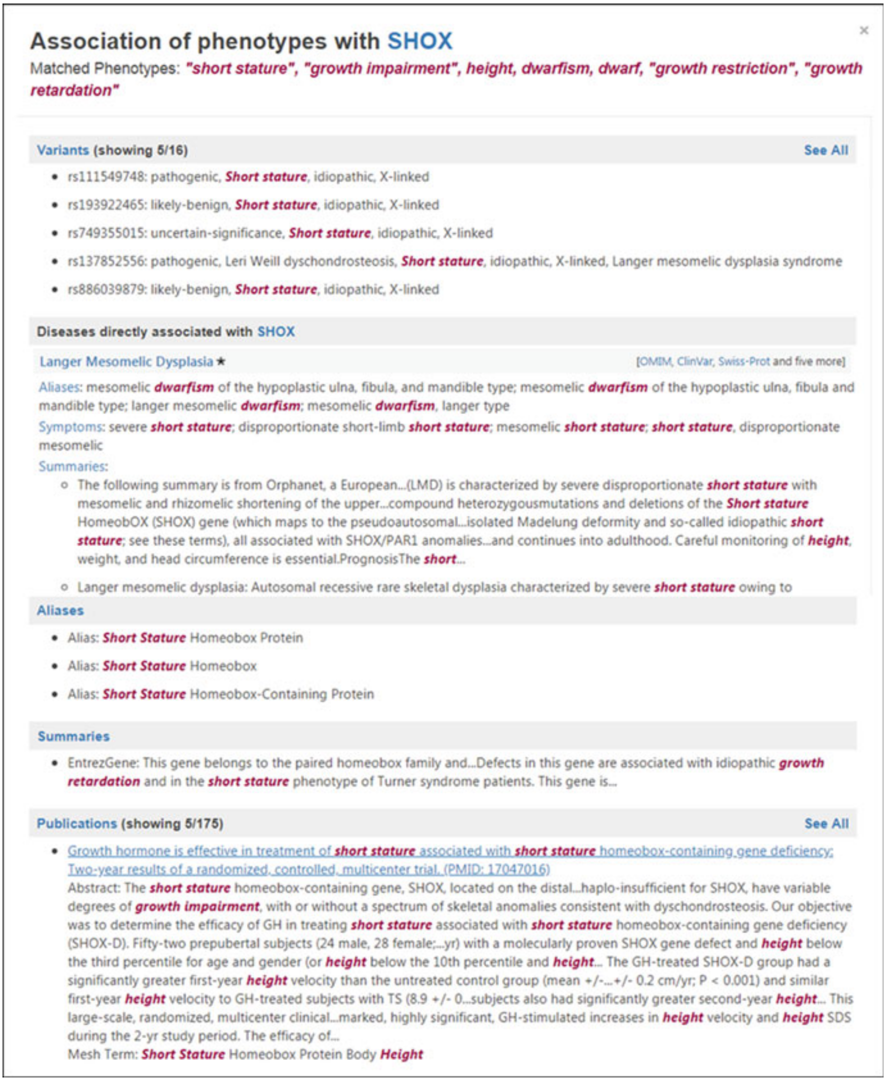
to find and prioritize phenotype related genes, leading to identification of pathogenic CNVs. We demonstrate this workflow using the TGex SVs module (Fig. 2.6).

The user inputs to TGex are: (Rebhan et al. 1997) a list of SVs; (Stelzer et al. 2016a) disease/phenotype/symptom terms. The analysis screen allows the user to browse and interpret the SVs. The list of entered SVs (Fig. 2.6, left pane) is presented along with annotations, such as the genomic location and length, SV type, number of genes in the region, and more. Those annotations are amplified with the VarElect score, which is also used as the default sort column for the SVs list. The value in this column is the highest VarElect phenotype score of the gene pool in each SV gene list. In this analysis the highest scoring SV is a 550kb deletion on chromosome X, overlapping 5 genes and one enhancer element.

The user can click on any of the SVs in the list (left pane) for the detailed view of each SV. In this view (Fig. 2.6, right pane) functional genomic elements in overlap with the SV region are shown (including not only protein coding genes, but also ncRNA genes, enhancers, and promoters), with annotations such as the overlap type (full/partial), the number of exons in overlap (for genes), and GeneHancer confidence scores for regulatory elements (see below). For the selected SV, the gene *SHOX* (Short Stature Homeobox) is the VarElect top scoring gene for the submitted keyword list (“short stature” OR “growth impairment” OR height OR dwarfism OR dwarf OR “growth restriction” OR “growth retardation”). Clicking on the VarElect score opens the “MiniCard,” which shows the hit context of the search terms within different sections of the *SHOX* gene in GeneCards, and diseases related to *SHOX* in MalaCards (Fig. 2.7).

2.5.3 GeneHancer-Powered Interpretation of SVs

GeneHancer, the GeneCards Suite database of regulatory elements and their gene targets, has been used by the community as an annotation standard for enhancers and promoters in the human genome, as well as for the associations of those elements with their gene targets (Quigley et al. 2018; Zhang et al. 2018; Holzinger et al. 2017;



Association of phenotypes with SHOX

×

Matched Phenotypes: ***"short stature", "growth impairment", height, dwarfism, dwarf, "growth restriction", "growth retardation"***

Variants **(showing 5/16)** See All

- rs111549748: pathogenic, ***Short stature***, idiopathic, X-linked
- rs193922465: likely-benign, ***Short stature***, idiopathic, X-linked
- rs749355015: uncertain-significance, ***Short stature***, idiopathic, X-linked
- rs137852556: pathogenic, Leri Weill dyschondrosteosis, ***Short stature***, idiopathic, X-linked, Langer mesomelic dysplasia syndrome
- rs886039879: likely-benign, ***Short stature***, idiopathic, X-linked

Diseases directly associated with SHOX

Langer Mesomelic Dysplasia ★ [OMIM, ClinVar, Swiss-Prot and five more]

Aliases: mesomelic ***dwarfism*** of the hypoplastic ulna, fibula, and mandible type; mesomelic ***dwarfism*** of the hypoplastic ulna, fibula and mandible type; langer mesomelic ***dwarfism***; mesomelic ***dwarfism***, langer type

Symptoms: severe ***short stature***; disproportionate short-limb ***short stature***; mesomelic ***short stature***; ***short stature***, disproportionate mesomelic

Summaries:

- The following summary is from Orphanet, a European...(LMD) is characterized by severe disproportionate ***short stature*** with mesomelic and rhizomelic shortening of the upper...compound heterozygousmutations and deletions of the ***Short stature*** HomeobOX (SHOX) gene (which maps to the pseudoautosomal...isolated Madelung deformity and so-called idiopathic ***short stature***; see these terms), all associated with SHOX/PAR1 anomalies...and continues into adulthood. Careful monitoring of ***height***, weight, and head circumference is essential.PrognosisThe ***short***...
- Langer mesomelic dysplasia: Autosomal recessive rare skeletal dysplasia characterized by severe ***short stature*** owing to

Aliases

- Alias: ***Short Stature*** Homeobox Protein
- Alias: ***Short Stature*** Homeobox
- Alias: ***Short Stature*** Homeobox-Containing Protein

Summaries

- EntrezGene: This gene belongs to the paired homeobox family and...Defects in this gene are associated with idiopathic ***growth retardation*** and in the ***short stature*** phenotype of Turner syndrome patients. This gene is...

Publications **(showing 5/175)** See All

- Growth hormone is effective in treatment of ***short stature*** associated with ***short stature*** homeobox-containing gene deficiency: Two-year results of a randomized, controlled, multicenter trial. (PMID: 17047016)
  Abstract: The ***short stature*** homeobox-containing gene, SHOX, located on the distal...haplo-insufficient for SHOX, have variable degrees of ***growth impairment***, with or without a spectrum of skeletal anomalies consistent with dyschondrosteosis. Our objective was to determine the efficacy of GH in treating ***short stature*** associated with ***short stature*** homeobox-containing gene deficiency (SHOX-D). Fifty-two prepubertal subjects (24 male, 28 female;...yr) with a molecularly proven SHOX gene defect and ***height*** below the third percentile for age and gender (or ***height*** below the 10th percentile and ***height***... The GH-treated SHOX-D group had a significantly greater first-year ***height*** velocity than the untreated control group (mean +/-...+/- 0.2 cm/yr; P < 0.001) and similar first-year ***height*** velocity to GH-treated subjects with TS (8.9 +/- 0...subjects also had significantly greater second-year ***height***... This large-scale, randomized, multicenter clinical...marked, highly significant, GH-stimulated increases in ***height*** velocity and ***height*** SDS during the 2-yr study period. The efficacy of...
  Mesh Term: ***Short Stature*** Homeobox Protein Body ***Height***

Fig. 2.7 MiniCards—evidence for gene-phenotype associations. This figure shows selected parts of the MiniCard for the gene SHOX and the phenotypes used in the short stature study. A list of matched phenotypes is shown in red in the top part. This is followed by several gene-centric evidence for queried phenotype association, e.g. from the GeneCards Variants, Aliases, Summaries, and Publication sections. This evidence is combined by MalaCards-based evidence, showing queried phenotype associations in diseases associated with the gene SHOX, from various MalaCards sections, e.g. Aliases, Symptoms, and Summaries. For all sections, only partial evidence list is shown here

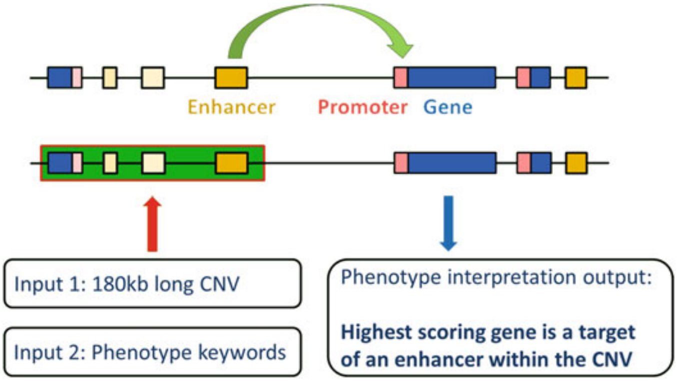

Enhancer
Promoter
Gene
Input 1: 180kb long CNV
Input 2: Phenotype keywords
Phenotype interpretation output:
Highest scoring gene is a target of an enhancer within the CNV

Fig. 2.8 Solving a genetic disease with GeneCards Suite NGS tools SV analysis capacities. GeneHancer enriches the GeneCards Suite NGS tools VarElect and TGex, providing the ability to map SVs to non-coding functional regulatory elements such as enhancers and promoters. This mapping, combined with GeneHancer's information on the association of those elements with target genes, enables pinpointing variant-phenotype relationships that otherwise might be undiscovered, increasing the potential solve rate of genetic disease studies

Singh et al. 2018; Huang et al. 2018; Yang et al. 2018; Bermejo et al. 2019; Erlangsen et al. 2020; Nikulin et al. 2018; Slater et al. 2018). With the growing understanding of the importance of non-coding variants for NGS interpretation, GeneHancer-enriched VarElect and TGex offer novel modes of analysis for tackling this challenge.

First, we augmented VarElect to be able to process GeneHancer element identifiers. For a given element, VarElect performs gene-phenotype prioritization for its GeneHancer gene targets. The phenotype prioritization in this workflow is performed by combining the VarElect gene-phenotype score with the GeneHancer element and gene-association confidence scores. This mode of analysis allows users to perform phenotype interpretation of mixed lists of genes and regulatory elements after both SNV and SV primary data analysis steps.

Second, we enhanced TGex to include regulatory elements in SV interpretation. User-submitted SVs are mapped to both genes and regulatory elements, followed by VarElect interpretation of the mixed list of genes and enhancers/promoters. This mode of analysis helped our lab solve a genetic disease study (Fig. 2.8). In this case, a family with a rare congenital autosomal dominant genetic skin disease was genotyped, leading to the identification of a CNV shared by all affected individuals. Phenotype interpretation of this CNV discovered that it overlaps an enhancer, whose gene target, albeit not residing within the CNV, is extremely relevant for the studied phenotype.

2.5.4 Other VarElect Use Cases

While interpretation of genetic disease NGS analyses was the focus of our described use cases, VarElect is also a potent tool for supporting the interpretation of other experimental results. In such scenarios, VarElect is utilized to analyze gene lists retrieved from various methodologies, helping to focus on more affordable candidate gene lists based on gene-phenotype information. Scenarios benefitting from the gene-phenotype prioritization capacities of VarElect include gene expression (RNAseq/Microarrays), protein expression (mass spectrometry), and other multi-OMICS downstream analyses (Hulst et al. 2017; Yang et al. 2016; Biro et al. 2017; Voisey et al. 2017; Amorim et al. 2017; Fonseca et al. 2018); genome-wide association studies (Luzon-Toro et al. 2015); Quantitative Trait Locus (QTL) gene targets downstream analysis (Martinez-Montes et al. 2018); and others (Chen et al. 2016; Alvarez-Castelao et al. 2017; Butler et al. 2016; Makler and Narayanan 2017; Hashemi et al. 2017).

2.6 Summary and Future Development of the Database

The tools and databases in the GeneCards Suite synergistically work in concert to provide information, elucidate relationships, and facilitate solving clinical cases. Each suite member provides deep insights about particular facets of biological research. Specifically, GeneCards is gene-centered, the one-stop-shop for comprehensive details related to genes of interest. MalaCards focuses on diseases and disorders, presenting a detailed view of 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. Users who want to explore human pathways data will find it in PathCards, an integrated database of human biological pathways and their annotations, wherein each record presents a SuperPath that represents one or more human pathways, their gene content, and relationships within member pathways. GeneLoc consolidates genes from major worldwide 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. GeneHancer, an innovative and growing regulatory element database, focuses on enhancers and promoters, central to tissue-related gene expression, with many known strong connections to diseases. GenesLikeMe measures how genes are related to a target gene, based on shared characteristics, including expression, ontologies, or disorders. Using a gene set from the results of a GeneCards search, or any set of genes of interest, one can extract GeneCards annotations for all genes in the set using GeneALaCart, the suite's batch query facility. The set can be further analyzed using GeneAnalytics, which can identify cell types, diseases, pathways, and functions enriched in the gene set, and provides tools for further in-depth analysis of all of the genes in the set. VarElect identifies and prioritizes genes and variants according to their relevance to diseases and phenotypes of interest and allows one to explore relationships between genes and gene variants and selected diseases, phenotypes, or any pertinent biological term via relevant pathways, interaction networks, and publications. TGex, the suite's end-to-end NGS solution, is a VCF-to-report clinical analyzer which incorporates VarElect's algorithms.

Future plans include continuing to build on the efforts of the last twenty years, ensuring that information from current sources is kept up-to-date, relevant, and provided in a user-friendly manner, in parallel with continuing to innovate in the "dark matter" arena of regulatory elements and RNA genes. The GeneCards Suite's extensive KnowledgeBase and disease interpretation platform fortifies its capacities to relate diseases to non-coding variants identified by WGS, towards providing a comprehensive route to clinical significance of coding and non-coding single nucleotide and structural genomic variations, in order to elucidate unsolved clinical cases and enable accurate clinical diagnosis and comprehensive genetic counseling.

Acknowledgements **Funding**: This work was supported by a research and licensing grant from LifeMap Sciences, Inc. (California, USA) and by the Crown Human Genome Center at the Weizmann Institute of Science (Rehovot, Israel).

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249

Alkelai A, Olender T, Haffner-Krausz R, Tsoory MM, Boyko V, Tatarskyy P, Gross-Isseroff R, Milgrom R, Shushan S, Blau I, Cohn E, Beeri R, Levy-Lahad E, Pras E, Lancet D (2016) A role for TENM1 mutations in congenital general anosmia. Clin Genet 90(3):211–219

Alkelai A, Olender T, Dode C, Shushan S, Tatarskyy P, Furman-Haran E, Boyko V, Gross-Isseroff R, Halvorsen M, Greenbaum L, Milgrom R, Yamada K, Haneishi A, Blau I, Lancet D (2017) Next-generation sequencing of patients with congenital anosmia. Eur J Hum Genet 25(12):1377–1387

Alvarez-Castelao B, Schanzenbacher CT, Hanus C, Glock C, Tom Dieck S, Dorrbaum AR, Bartnik I, Nassim-Assir B, Ciirdaeva E, Mueller A, Dieterich DC, Tirrell DA, Langer JD, Schuman EM (2017) Cell-type-specific metabolic labeling of nascent proteomes in vivo. Nat Biotechnol 35(12):1196–1201

Amorim IS, Graham LC, Carter RN, Morton NM, Hammachi F, Kunath T, Pennetta G, Carpanini SM, Manson JC, Lamont DJ, Wishart TM, Gillingwater TH (2017) Sideroflexin 3 is an alpha-synuclein-dependent mitochondrial protein that regulates synaptic morphology. J Cell Sci 130(2):325–331

Azim MK, Mehnaz A, Ahmed JZ, Mujtaba G (2019) Exome sequencing identifies a novel frameshift variant causing hypomagnesemia with secondary hypocalcemia. CEN Case Rep 8(1):42–47

Bafunno V, Firinu D, D'Apolito M, Cordisco G, Loffredo S, Leccese A, Bova M, Barca MP, Santacroce R, Cicardi M, Del Giacco S, Margaglione M (2018) Mutation of the angiopoietin-1 gene (ANGPT1) associates with a new type of hereditary angioedema. J Allergy Clin Immunol 141(3):1009–1017

Bamshad MJ, Shendure JA, Valle D, Hamosh A, Lupski JR, Gibbs RA, Boerwinkle E, Lifton RP, Gerstein M, Gunel M, Mane S, Nickerson DA, Centers for Mendelian Genomics (2012) The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am J Med Genet A 158A(7):1523–1525

Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Bye-A-Jee H, Cowley A, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Castro LG, Figueira L, Garmiri P et al (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169

Belinky F, Bahir I, Stelzer G, Zimmerman S, Rosen N, Nativ N, Dalah I, Iny Stein T, Rappaport N, Mituyama T, Safran M, Lancet D (2013) Non-redundant compendium of human ncRNA genes in GeneCards. Bioinformatics 29(2):255–261

Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, Lancet D (2015) PathCards: multi-source consolidation of human biological pathways. Database (Oxford) 2015: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, Shtrichman R (2016) GeneAnalytics: an integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data. OMICS 20(3):139–151

Bermejo JL, Huang G, Manoochehri M, Mesa KG, Schick M, Silos RG, Ko Y-D, Bruning T, Brauch H, Lo W-Y, Hoheisel JD, Hamann U (2019) Long intergenic noncoding RNA 299 methylation in peripheral blood is a biomarker for triple-negative breast cancer. Epigenomics 11(1):81–93

Biro O, Nagy B, Rigo J Jr (2017) Identifying miRNA regulatory mechanisms in preeclampsia by systems biology approaches. Hypertens Pregnancy 36(1):90–99

Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD (2015) Gene: a gene-centered information resource at NCBI. Nucleic Acids Res 43(D1):D36–D42

Butler MG, McGuire AB, Masoud H, Manzardo AM (2016) Currently recognized genes for schizophrenia: high-resolution chromosome ideogram representation. Am J Med Genet B Neuropsychiatr Genet 171B(2):181–202

Carneiro TN, Krepischi AC, Costa SS, da Silva IT, Vianna-Morgante AM, Valieris R, Ezquina SA, Bertola DR, Otto PA, Rosenberg C (2018) Utility of trio-based exome sequencing in the elucidation of the genetic basis of isolated syndromic intellectual disability: illustrative cases. Appl Clin Genet 11:93–98

Chalifa-Caspi V, Yanai I, Ophir R, Rosen N, Shmoish M, Benjamin-Rodrig H, Shklar M, Stein TI, Shmueli O, Safran M, Lancet D (2004) GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. Bioinformatics 20(9):1457–1458

Chen P, Mancini M, Sonis ST, Fernandez-Martinez J, Liu J, Cohen EE, Toback FG (2016) A novel peptide for simultaneously enhanced treatment of head and neck cancer and mitigation of oral mucositis. PLoS One 11(4):e0152995

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 MD, Warshawsky D, Shtrichman R (2013) LifeMap discovery: the embryonic development, stem cells, and regenerative medicine research portal. PLoS One 8(7):e66629

Einhorn Y, Weissglas-Volkov D, Carmi S, Ostrer H, Friedman E, Shomron N (2017) Differential analysis of mutations in the Jewish population and their implications for diseases. Genet Res 99: e3

Ekhilevitch N, Kurolap A, Oz-Levi D, Mory A, Hershkovitz T, Ast G, Mandel H, Baris HN (2016) Expanding the MYBPC1 phenotypic spectrum: a novel homozygous mutation causes arthrogryposis multiplex congenita. Clin Genet 90(1):84–89

Erlangsen A, Appadurai V, Wang Y, Turecki G, Mors O, Werge T, Mortensen PB, Starnawska A, Borglum AD, Schork A, Nudel R, Baekvad-Hansen M, Bybjerg-Grauholm J, Hougaard DM, Thompson WK, Nordentoft M, Agerbo E (2020) Genetics of suicide attempts in individuals with and without mental disorders: a population-based genome-wide association study. Mol Psychiatry 25(10):2410–2421

Feliubadalo L, Tonda R, Gausachs M, Trotta JR, Castellanos E, Lopez-Doriga A, Teule A, Tornero E, del Valle J, Gel B, Gut M, Pineda M, Gonzalez S, Menendez M, Navarro M, Capella G, Gut I, Serra E, Brunet J, Beltran S et al (2017) Benchmarking of whole exome sequencing and Ad Hoc designed panels for genetic testing of hereditary cancer. Sci Rep 7: 37984

Fidalgo F, Rodrigues TC, Silva AG, Facure L, de Sa BC, Duprat JP, Achatz MI, Rosenberg C, Carraro DM, Krepischi AC (2016) Role of rare germline copy number variation in melanoma-prone patients. Future Oncol 12(11):1345–1357

Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, Lancet D, Cohen D (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) 2017:bax028

Fonseca PAS, Id-Lahoucine S, Reverter A, Medrano JF, Fortes MS, Casellas J, Miglior F, Brito L, Carvalho MRS, Schenkel FS, Nguyen LT, Porto-Neto LR, Thomas MG, Canovas A (2018) Combining multi-OMICs information to identify key-regulator genes for pleiotropic effect on fertility and production traits in beef cattle. PLoS One 13(10):e0205295

Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding MJ, Bamford S, Cole C, Ward S, Kok CY, Jia MM, De TS, Teague JW, Stratton MR, McDermott U, Campbell PJ (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 43(D1):D805–D811

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43(Database issue):D1049–D1056

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, Tearle R, Bo T, Pfundt R, Yntema HG, de Vries BBA, Kleefstra T, Brunner HG, Vissers LELM et al (2014) Genome sequencing identifies major causes of severe intellectual disability. Nature 511(7509):344–347

Harel A, Inger A, Stelzer G, Strichman-Almashanu L, Dalah I, Safran M, Lancet D (2009) GIFtS: annotation landscape analysis with GeneCards. BMC Bioinformatics 10:348

Hashemi S, Fernandez Martinez JL, Saligan L, Sonis S (2017) Exploring genetic attributions underlying radiotherapy-induced fatigue in prostate cancer patients. J Pain Symptom Manage 54(3):326–339

Hecht M, Bromberg Y, Rost B (2015) Better prediction of functional effects for sequence variants. BMC Genomics 16:S1

Heimer G, Oz-Levi D, Eyal E, Edvardson S, Nissenkorn A, Ruzzo EK, Szeinberg A, Maayan C, Mai-Zahav M, Efrati O, Pras E, Reznik-Wolf H, Lancet D, Goldstein DB, Anikster Y, Shalev SA, Elpeleg O, Ben Zeev B (2016) TECPR2 mutations cause a new subtype of familial dysautonomia like hereditary sensory autonomic neuropathy with intellectual disability. Eur J Paediatr Neurol 20(1):69–79

Heimer G, Eyal E, Zhu X, Ruzzo EK, Marek-Yagel D, Sagiv D, Anikster Y, Reznik-Wolf H, Pras E, Oz Levi D, Lancet D, Ben-Zeev B, Nissenkorn A (2018) Mutations in AIFM1 cause an X-linked childhood cerebellar ataxia partially responsive to riboflavin. Eur J Paediatr Neurol 22(1):93–101

Holzinger ER, Li Q, Parker MM, Hetmanski JB, Marazita ML, Mangold E, Ludwig KU, Taub MA, Begum F, Murray JC, Albacha-Hejazi H, Alqosayer K, Al-Souki G, Albasha Hejazi A, Scott AF, Beaty TH, Bailey-Wilson JE (2017) Analysis of sequence data to identify potential risk variants for oral clefts in multiplex families. Mol Genet Genomic Med 5(5):570–579

Homma TK, Krepischi ACV, Furuya TK, Honjo RS, Malaquias AC, Bertola DR, Costa SS, Canton AP, Roela RA, Freire BL, Kim CA, Rosenberg C, Jorge AAL (2018) Recurrent Copy Number Variants Associated with Syndromic Short Stature of Unknown Cause. Horm Res Paediatr 89(1):13–21

Huang H, Zhang C, Wang B, Wang F, Pei B, Cheng C, Yang W, Zhao Z (2018) Transduction with lentiviral vectors altered the expression profile of host microRNAs. J Virol 92(18):e00503-18

Hulst M, Jansman A, Wijers I, Hoekman A, Vastenhouw S, van Krimpen M, Smits M, Schokker D (2017) Enrichment of in vivo transcription data from dietary intervention studies with in vitro data provides improved insight into gene regulation mechanisms in the intestinal mucosa. Genes Nutr 12:11

Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. Trends Genet 24(5):238–245

Jia Z, Mao FB, Wang L, Li MZ, Shi YY, Zhang BR, Gao GL (2017) Whole-exome sequencing identifies a de novo mutation in TRPM4 involved in pleiotropic ventricular septal defect. Int J Clin Exp Pathol 10(5):5092–5104

Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M (2019) New approach for understanding genome variations in KEGG. Nucleic Acids Res 47(D1):D590–D595

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980–D985

Luzon-Toro B, Bleda M, Navarro E, Garcia-Alonso L, Ruiz-Ferrer M, Medina I, Martin-Sanchez M, Gonzalez CY, Fernandez RM, Torroglosa A, Antinolo G, Dopazo J, Borrego S (2015) Identification of epistatic interactions through genome-wide association studies in sporadic medullary and juvenile papillary thyroid carcinomas. BMC Med Genomics 8:83

Makler A, Narayanan R (2017) Mining exosomal genes for pancreatic cancer targets. Cancer Genomics Proteomics 14(3):161–172

Martinez-Montes AM, Fernandez A, Munoz M, Noguera JL, Folch JM, Fernandez AI (2018) Using genome wide association studies to identify common QTL regions in three different genetic backgrounds based on Iberian pig breed. Plos One 13(3):e0190184

Nikulin SV, Knyazev EN, Poloznikov AA, Shilin SA, Gazizov IN, Zakharova GS, Gerasimenko TN (2018) Expression of SLC30A10 and SLC23A3 transporter mRNAs in Caco-2 cells correlates with an increase in the area of the apical membrane. Mol Biol 52(4):577–582

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, Lancet D (2015) Exome sequencing as a differential diagnosis tool: resolving mild trichohepatoenteric syndrome. Clin Genet 87(6):602–603

Pletscher-Frankild S, Palleja A, Tsafou K, Binder JX, Jensen LJ (2015) DISEASES: text mining and data integration of disease-gene associations. Methods 74:83–89

Porter MF (2006) An algorithm for suffix stripping. Program-Electronic Library and Information Systems 40(3):211–218

Quigley DA, Dang HX, Zhao SG, Lloyd P, Aggarwal R, Alumkal JJ, Foye A, Kothari V, Perry MD, Bailey AM, Playdle D, Barnard TJ, Zhang L, Zhang J, Youngren JF, Cieslik MP, Parolia A, Beer TM, Thomas G, Chi KN et al (2018) Genomic hallmarks and structural variation in metastatic prostate cancer. Cell 174(3):758–769. e9

Ramos E, Levinson BT, Chasnoff S, Hughes A, Young AL, Thornton K, Li AL, Vallania FLM, Province M, Druley TE (2012) Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing. BMC Genomics 13:683

Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D (2013) MalaCards: an integrated compendium for diseases and their annotation. Database (Oxford) 2013:bat018

Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein TI, Safran M, Lancet D (2014) MalaCards: a comprehensive automatically-mined database of human diseases. Curr Protoc Bioinformatics 47:1.24.1–1.24.19

Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M, Lancet D (2017a) MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res 45(D1):D877–D887

Rappaport N, Fishilevich S, Nudel R, Twik M, Belinky F, Plaschkes I, Stein TI, Cohen D, Oz-Levi D, Safran M, Lancet D (2017b) Rational confederation of genes and diseases: NGS interpretation via GeneCards. MalaCards and VarElect Biomed Eng Online 16(Suppl 1):72

Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13(4):163

Rosen N, Chalifa-Caspi V, Shmueli O, Adato A, Lapidot M, Stampnitzky J, Safran M, Lancet D (2003) GeneLoc: exon-based integration of human genome maps. Bioinformatics 19(Suppl 1): i222–i224

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, Lancet D (2010) GeneCards Version 3: the human gene integrator. Database (Oxford) 2010:baq020

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40(Web Server issue):W452–W457

Singh G, Bhat B, Jayadev MSK, Madhusudhan C, Singh A (2018) mutTCPdb: a comprehensive database for genomic variants of a tropical country neglected disease-tropical calcific pancreatitis. Database (Oxford) 2018:bay043

Slater SC, Jover E, Martello A, Mitic T, Rodriguez-Arabaolaza I, Vono R, Alvino VV, Satchell SC, Spinetti G, Caporali A, Madeddu P (2018) MicroRNA-532-5p regulates pericyte function by targeting the transcription regulator BACH1 and angiopoietin-1. Mol Ther 26(12):2823–2837

Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M (2011) Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol 12(9):R85

Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Mouse Genome G (2018) Database, Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res 46(D1):D836–D842

Stelzer G, Inger A, Olender T, Iny-Stein T, Dalah I, Harel A, Safran M, Lancet D (2009) GeneDecks: paralog hunting and gene-set distillation with GeneCards annotation. OMICS 13(6):477–487

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan-Golan Y, Kohn A, Rappaport N, Safran M, Lancet D (2016a) The GeneCards Suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 54:1.30.1–1.30.33

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 HN, Rappaport N, Safran M et al (2016b) VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics 17(Suppl 2):444

Stranneheim H, Wedell A (2016) Exome and genome sequencing: a revolution for the discovery and diagnosis of monogenic disorders. J Intern Med 279(1):3–15

Syama A, Sen S, Kota LN, Viswanath B, Purushottam M, Varghese M, Jain S, Panicker MM, Mukherjee O (2018) Mutation burden profile in familial Alzheimer's disease cases from India. Neurobiol Aging 64:158 e7–158 e13

van den Veyver IB, Eng CM (2015) Genome-wide sequencing for prenatal detection of fetal single-gene disorders. Cold Spring Harb Perspect Med 5(10):a023077

Voisey J, Mehta D, McLeay R, Morris CP, Wockner LF, Noble EP, Lawford BR, Young RM (2017) Clinically proven drug targets differentially expressed in the prefrontal cortex of schizophrenia patients. Brain Behav Immun 61:259–265

Weischenfeldt J, Symmons O, Spitz F, Korbel JO (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14(2):125–138

Yang YP, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu ZY, Hardison M, Person R, Bekheirnia MR, Leduc MS, Kirby A, Pham P, Scull J, Wang M, Ding Y, Plon SE et al (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369(16):1502–1511

Yang WE, Suchindran S, Nicholson BP, McClain MT, Burke T, Ginsburg GS, Harro CD, Chakraborty S, Sack DA, Woods CW, Tsalik EL (2016) Transcriptomic analysis of the host response and innate resilience to enterotoxigenic Escherichia coli infection in humans. J Infect Dis 213(9):1495–1504

Yang C, Xu Y, Yu M, Lee D, Alharti S, Hellen N, Ahmad Shaik N, Banaganapalli B, Sheikh Ali Mohamoud H, Elango R, Przyborski S, Tenin G, Williams S, O'Sullivan J, Al-Radi OO, Atta J, Harding SE, Keavney B, Lako M, Armstrong L (2017) Induced pluripotent stem cell modelling of HLHS underlines the contribution of dysfunctional NOTCH signalling to impaired cardiogenesis. Hum Mol Genet 26(16):3031–3045

Yang C, Lim W, Bazer FW, Song G (2018) Avobenzone suppresses proliferative activity of human trophoblast cells and induces apoptosis mediated by mitochondrial disruption. Reprod Toxicol 81:50–57

Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA (2017) Genenames.org: the HGNC and VGNC resources in. Nucleic Acids Res 45(D1):D619–D625

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, J.K. To, Laird MR et al (2018) Ensembl 2018. Nucleic Acids Res 46(D1):D754–D761

Zhang L, Jia Z, Mao F, Shi Y, Bu RF, Zhang B (2016) Whole-exome sequencing identifies a somatic missense mutation of NBN in clear cell sarcoma of the salivary gland. Oncol Rep 35(6):3349–3356

Zhang W, Bojorquez-Gomez A, Velez DO, Xu G, Sanchez KS, Shen JP, Chen K, Licon K, Melton C, Olson KM, Yu MK, Huang JK, Carter H, Farley EK, Snyder M, Fraley SI, Kreisberg JF, Ideker T (2018) A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. Nat Genet 50(4):613–620

Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, Dahia CL, Park-Min KH, Tobias JH, Kooperberg C, Kleinman A, Styrkarsdottir U, Liu CT, Uggla C, Evans DS, Nielson CM, Walter K, Pettersson-Kymmer U, McCarthy S, Eriksson J et al (2015) Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526(7571):112–117