OMICS A Journal of Integrative Biology
Volume 20, Number 3, 2016
Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2015.0168

Original Articles

GeneAnalytics: An Integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data

Shani Ben-Ari Fuchs,[1] Iris Lieder,[1] Gil Stelzer,[1,2] Yaron Mazor,[1] Ella Buzhor,[3] Sergey Kaplan,[1] Yoel Bogoch,[4] Inbar Plaschkes,[1] Alina Shitrit,[2] Noa Rappaport,[2] Asher Kohn,[5] Ron Edgar,[6] Liraz Shenhav,[1] Marilyn Safran,[2] Doron Lancet,[2] Yaron Guan-Golan,[5] David Warshawsky,[5] and Ronit Shtrichman[7]

Abstract

Postgenomics data are produced in large volumes by life sciences and clinical applications of novel omics diagnostics and therapeutics for precision medicine. To move from ''data-to-knowledge-to-innovation,'' a crucial missing step in the current era is, however, our limited understanding of biological and clinical contexts associated with data. Prominent among the emerging remedies to this challenge are the gene set enrichment tools. This study reports on GeneAnalytics™ (geneanalytics.genecards.org), a comprehensive and easy-to-apply gene set analysis tool for rapid contextualization of expression patterns and functional signatures embedded in the postgenomics Big Data domains, such as Next Generation Sequencing (NGS), RNAseq, and microarray experiments. GeneAnalytics' differentiating features include in-depth evidence-based scoring algorithms, an intuitive user interface and proprietary unified data. GeneAnalytics employs the LifeMap Science's GeneCards suite, including the *GeneCards*®—the human gene database; the *MalaCards*—the human diseases database; and the *PathCards*—the biological pathways database. Expression-based analysis in GeneAnalytics relies on the LifeMap Discovery®—the embryonic development and stem cells database, which includes manually curated expression data for normal and diseased tissues, enabling advanced matching algorithm for gene–tissue association. This assists in evaluating differentiation protocols and discovering biomarkers for tissues and cells. Results are directly linked to gene, disease, or cell ''cards'' in the GeneCards suite. Future developments aim to enhance the GeneAnalytics algorithm as well as visualizations, employing varied graphical display items. Such attributes make GeneAnalytics a broadly applicable postgenomics data analyses and interpretation tool for translation of data to knowledge-based innovation in various Big Data fields such as precision medicine, ecogenomics, nutrigenomics, pharmacogenomics, vaccinomics, and others yet to emerge on the postgenomics horizon.

Introduction

High throughput genomics technologies, such as next generation DNA/RNA sequencing or microarray analyses, are frequently used during biomedical research, as well as in diagnostic and therapeutic product development. These generate large quantities of Big Data that require advanced bioinformatics analysis and interpretation. The key step towards translating these results into meaningful scientific discoveries is deduction of biological and clinical contexts from the generated data. In this realm, several methods and tools have been developed to interpret large sets of genes or proteins, using information available in biological databases. Prominent among these are gene set enrichment tools.

In conventional examples, the Gene Ontology database is used for the functional study of large scale genomics or

[1]LifeMap Sciences Ltd., Tel Aviv, Israel.
[2]Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.
[3]Institute of Oncology, Sheba Medical Center, Tel Hashomer, Israel.
[4]Surgical Department, Sourasky Medical Centre, Tel Aviv, Israel.
[5]LifeMap Sciences, Inc., Marshfield, Massachusetts, USA.
[6]Venividi Solutions LLC, Rockville, Maryland, USA.
[7]Bonus BioGroup Ltd., Haifa, Israel.

transcriptomics data. Multiple applications such as GeneCodis, GOEAST, Gorilla, and Blast2GO (Conesa et al., 2005; Eden et al., 2009; Nogales-Cadenas et al., 2009; Zheng and Wang, 2008) can analyze and visualize statistical enrichment of GO terms in a given gene set. Other tools rely on popular data sources such as Kyoto Encyclopedia of Genes and Genomes (KEGG), TransPath, Online Mendelian Inheritance in Man (OMIM), and GeneCards to identify enriched pathways, diseases, and phenotypes (Backes et al., 2007; Huang da et al., 2009b; Safran et al., 2010; Sherman et al., 2007; Stelzer et al., 2009; Zhang et al., 2005). These analysis tools differ in several respects, including statistical methodology, supported organisms and gene identifiers, coverage of functional categories, source databases, and user interface. The common result is the identification of known functional biological descriptors that are significantly enriched within the experimentally-derived gene list.

Enrichment of biological descriptors for a given set of genes introduces three immediate challenges: The first is determining the statistical significance of enrichment of each descriptor. There are several approaches to calculating the statistics for a descriptor shared among genes, such as Gene Set Enrichment Analysis [GSEA (Maezawa and Yoshimura, 1991)] and Fisher's exact test [Database for Annotation, Visualization and Integrated Discovery—DAVID (Dennis et al., 2003)]. Some tools, such as the DAVID functional annotation tool, initially cluster the descriptors belonging to similar categories, and then present a score for an enriched group of terms.

The second challenge is judicious use of multiple data sources. It is a nontrivial task to integrate and model information derived from various origins. In an example, disease information could be derived from data sources such as OMIM (Hamosh et al., 2005), SwissProt/UniProt (Wu et al., 2006), and Orphanet (Maiella et al., 2013), and pathway information—from Reactome (Jupe et al., 2014; Matthews et al., 2009) and/or KEGG (Kanehisa et al., 2010). Therefore many analysis tools present separate enrichment results for each data source, while others perform consolidated analysis on source types.

A third challenge is optimal data presentation. Tools such as DAVID group enriched terms by biological categories in an attempt to provide a general sense of the biological processes involved in the experimental results. Other tools, such as MSigDB (GSEA) (Liberzon et al., 2011) and GeneDecks Set Distiller (Stelzer et al., 2009), interlace biological descriptors of various kinds, based on their statistical enrichment strength, thus emphasizing the individual significance of each in the context of the general enriched descriptor list. It would be optimal to give both a birds-eye view of grouped descriptors for a given set of genes, as well as display the descriptors in detail.

Multiple data sources are generally employed for both broad and in-depth depictions of enrichment. A related challenge is to develop a straightforward and easy-to-use application, with intuitive output results, rendering the tool accessible to inexperienced users, with little or no bioinformatics background.

We present GeneAnalytics™ (geneanalytics.genecards.org), designed to distill enriched descriptors for a given gene set, while optimally addressing the aforementioned challenges. It is empowered by the GeneCards Suite, embodied as LifeMap's integrated knowledgebase, which automatically mines data from more than 120 data sources. GeneAnalytics' broad descriptor categories enable users to focus on areas of interest, each rich with annotation and supporting evidence. The GeneAnalytics analyses provide gene associations with tissues and cells types from LifeMap Discovery (LMD, discovery.lifemapsc.com), diseases from MalaCards, (www.malacards.org), as well as GO terms, pathways, phenotypes, and drug/compounds from GeneCards (www.genecards.org), (Fig. 1). Navigation within such comprehensive information, as well as further scrutiny, is facilitated by GeneAnalytics categorization and filtration tools.

Methods

GeneAnalytics input

During the input stage (Fig. 2A), the relevant species, human or mouse, is selected. Then a gene list is typed, aided by an autocomplete feature to define the correct official gene symbol. Alternatively, a gene list may be pasted or uploaded as a text file. In the latter case, the gene list automatically undergoes gene symbol identification (''symbolization'') process yielding ''ready for analysis'' and ''unidentified genes'' lists (Fig. 2B, C). Each gene in the ''ready for analysis'' list is shown with its full name and all available aliases/synonyms, enabling review and approval of the input genes before analysis.

For the ''unidentified genes'' list, GeneAnalytics assists in manual symbol identification by directly linking to the gene search in GeneCards. To provide all relevant results for each gene symbol, GeneAnalytics unifies orthologs and paralogs into 'ortholog groups' based on the information available inc HomoloGene (www.ncbi.nlm.nih.gov/homologene), with minor adaptations (See Supplementary S1 Appendix; supplementary material is available online at www.liebertpub.com/omi).

Upon completion of the input stage, GeneAnalytics analysis produces results that are divided into the following categories: Tissues and Cells, Diseases, Pathways, GO terms, Phenotypes, and Compounds. Genes are associated with these categories either by their expression (''expression-based analysis'') or by their function (''function-based analysis'') (Table 1). All sections have a ''drill down'' capacity for performing subqueries, allowing users to focus only on genes from their original gene set, filtered by those that match the selected entity.

Tissues and cells

All gene expression data, including those that are manually collected, annotated, and integrated into LMD, are used to rank the GeneAnalytics matching results.

The gene expression data available in LMD are obtained from three types of sources:

a) Scientific peer-reviewed manuscripts and books (Edgar et al., 2013).
b) High Throughput (HT) gene expression comparisons available in the Gene Expression Omnibus (GEO) (Edgar et al., 2002). These are subject to various standardization and analyses methods. For this, we developed and fine-tuned an algorithm for extracting differentially expressed genes from GEO matrix files (normalized data, detailed in ''Differentially expressed genes identification algorithm'' in Supplementary S2 Appendix). Applying a uniform algorithm to the gene data increased the comparability of the resulting differentially expressed gene

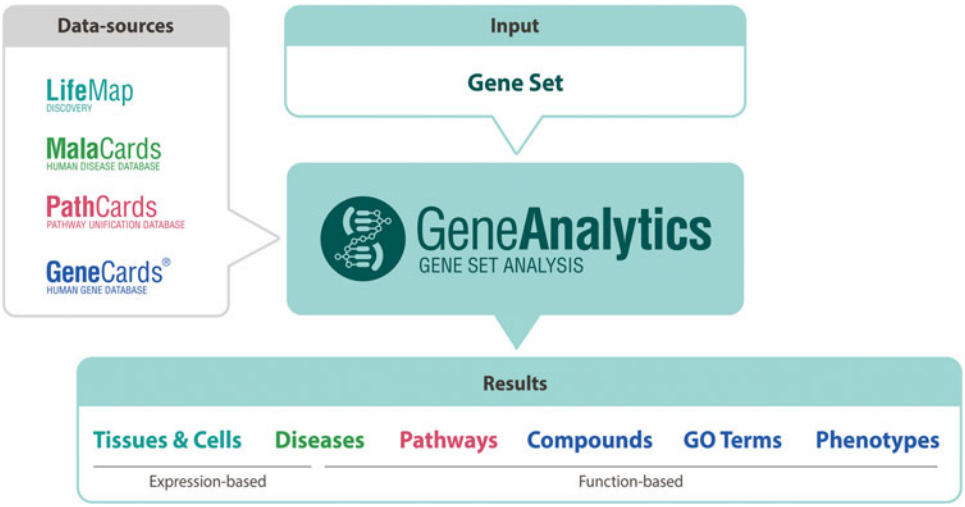

FIG. 1. GeneAnalytics structure. GeneAnalytics is powered by GeneCards, LifeMap Discovery, MalaCards, and PathCards, which integrate >100 data sources. These databases contain annotated gene lists for tissues and cells, diseases, pathways, compounds, and GO terms. GeneAnalytics compares the user's gene set to these compendia in search of the best matches. The output contains the best matched gene lists, scored and subdivided into their biological categories such as diseases or pathways. In the figure, each output category and its respective data source are marked with the same color.

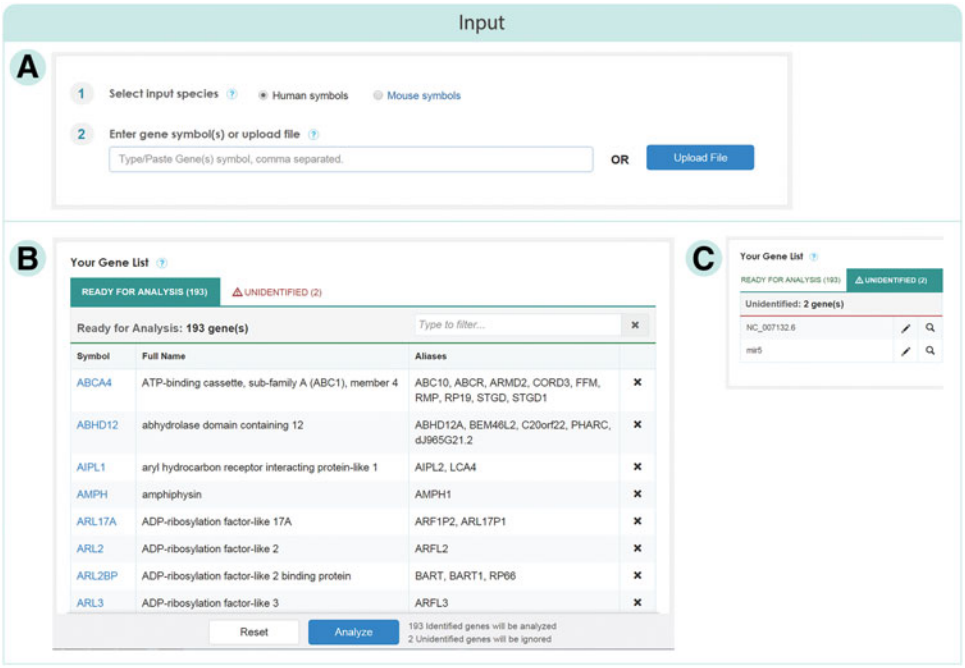

FIG. 2. The gene set input. **(A)** The input page is used to insert and identify the query gene list. 1) The identification process requires species indication in order to identify the gene symbols and their orthologs. GeneAnalytics identifies only official human and mouse gene symbols. 2) The genes can be inserted by typing/pasting gene symbols in the input window or by uploading a file containing the gene list. Typing a gene name in the search box initiates an autocomplete tool that includes only official gene symbols. The identification process yields two lists: **(B)** ''Ready for analysis'' gene list, which includes identified gene symbols, their full name, and all available aliases/synonyms, and **(C)** ''Unidentified genes'' list, which includes genes that were not recognized as official human or mouse gene symbols. These gene names can be manually corrected by running a search in GeneCards or by using the autocomplete option.

Table 1. GeneAnalytics Data Sources and Statistics

Results Category				
Analysis based on	*Entity type*	*Data sources*	*Total number of entities with associated genes*	*Total number of genes related to entities*
Expression	Normal tissues and cells	LifeMap Discovery	3,346	17,512
	Diseased tissues and cells*	LifeMap Discovery (via MalaCards)	96	6,963
Function	Disease	MalaCards	12,085	22,280
	Pathways	PathCards	1073 SuperPaths (unification of 3215 pathways)	11,479
	GO—biological process	GeneCards	9,436	14,907
	GO—molecular function		3,509	15,624
	Compounds		19,961 (unification of 44,942 compounds)	8,434

Data sources and statistics for each result category, based on the type of analysis.
*The expression data in diseased tissues and cells are available in the disease category.

list. For experiments that do not have normalized data deposited in a public repository, the differentially expressed gene lists, incorporated into the LMD database, are derived from the relevant article.

c) Large Scale Data Sets (LSDS): those obtained from wide-scope experiments that encompass multiple samples and require suitable standardization and analyses methods. This refers to data that obtained by *In situ* hybridization (ISH), immunostaining (IS), microarray, or RNA sequencing data sets. These data, retrieved from big-data repositories such as Mouse Genome Informatics (MGI) (Smith et al., 2014), Eurexpress (Geffers et al., 2012) or BioGPS (Wu et al., 2013), are filtered and analyzed in-house or obtained in analyzed form from projects that developed unique large-scale analysis methods such as Homer or Barcode.

The complete list of data sources is provided on the LMD webpage (discovery.lifemapsc.com/gene-expression-signals#ht-gene-expression).

In LMD, each anatomical entity has a unique card that contains a list of associated expressed genes [see (Edgar et al., 2013) for further details]. Organ and tissue cards include lists of genes expressed in whole tissue samples (e.g., RNA extracted from tissue homogenates). Genes reported to be expressed in a specific cell type (*in vivo* or *in vitro*) or in an anatomical compartment are listed in the relevant cards, which contain extensive manually curated information from the literature.

The High Throughput gene expression comparisons are described within 'experiment cards.' The top differentially expressed genes derived from these comparisons are linked into the highest resolution entity card possible (organs/tissues, anatomical compartments, or cells). Each card details the comparisons used in the experiment, listing the test and control samples comprising each comparison and supplying additional information for the experiment. The top differentially expressed genes (calculated as described in ''Differentially expressed genes identification algorithm'' in the Supplementary S2 Appendix) as well as links to LifeMap entities (tissues, compartments, etc.) may be viewed in the comparison cards associated with an experiment card.

Similarly, the lists of differentially expressed genes derived from Large Scale Data Sets are linked into entity cards, unless such a card is not available (for example, when the entity does not exist for a given release), in which case they are presented in Large Scale Data Sets cards. Thus the Tissues and Cells results are labeled by the four types of LMD entities shown in Table 2, with relevant links for further investigation (Fig. 3C).

The Tissues and Cells GeneAnalytics results contain useful filters that enable focus on specific subsets of the results (Fig. 3B). Each entity is classified into tissue(s) and/or system(s) in LMD, enabling results aggregation and filtration. This is done using higher anatomical hierarchy elements, tissues, and systems. For example, the *in vivo* cell Dopaminergic Progenitor Cells belongs to the anatomical compartment Substantia Nigra pars Compacta, which belongs to the tissue Brain, which is included in the system Nervous System.

The filtering into tissues or systems is associated with scores that reflect their matching quality to the query gene set (Fig. 3C, see next section). The Tissues and Cells results can also be used to filter *In vivo*/*In vitro* or Pre-natal/Post-natal entities (for further details, see ''Filters'' in Supplementary S2 Appendix). Further, GeneAnalytics allows user interaction for display of additional information. For example, for each entry in the Tissues and Cells table, we provide the type of entity, the expression type (expressed, selective marker, etc.), the number of genes matched to that entity (including the number of total genes expressed in the entity), and localization (within a popup).

When scoring after tissue/system filtering, during this aggregative filtering, a gene that appears in more than one entity will be represented only once at the tissue/system level, and will get the maximal score attributed to it in any of its associated entities. Once all of the genes are assembled for the tissue/system, the score is computed in the same manner as for every entity (shown in the detailed entity section, on the right).

The matching algorithm for this category aims to identify the anatomical entities most strongly associated with the query gene set. The algorithm is composed of two major stages:

a) Computation of a score for each gene associated with an entity. These pre-computed scores represent the importance of this gene in the specific entity as compared to its distribution in the entire entity landscape.

TABLE 2. LMD ENTITIES USED IN GENEANALYTICS MATCHING ANALYSIS IN TISSUES & CELLS CATEGORY

Entity type	*Icon*	*Data Origin*	*Example*	*Notes*
Organ Tissue		• High throughput gene expression comparisons • Large scale data sets	Heart	These entities contain a list of genes that have been found to be expressed in whole-tissue samples.
Anatomical compartment		• High throughput gene expression comparisons • Large scale data sets	Renal collecting duct system	These entities describe specific temporospatial regions within an organ/tissue.
In -vivo cell		• Data manually curated from the scientific literature • High throughput gene expression comparisons • Large Scale Data Sets	Inner cell Mass cells (ICM) Trabecular meshwork-derived mesenchymal stem cells	
In -vitro cell: cultured stem, progenitor and primary cell				
Protocol-derived cell				
Cell Family				
Large Scale Data Set sample cards		Large Scale Data Sets	GUDMAP: Ovary	These entities contain the gene list for each Large Scale Data Sets sample. These entities are only included in GeneAnalytics results if their gene list is not contained within any of the above entity types.

The entities available in the LMD database with gene expression information and an example for each.

b) Computation of the matching score, which is the similarity score between the user's query gene set and the genes associated with each of the entities, taking into account the differences in the expression information, both quantitative and qualitative, available for each entity.

The above is based on the fact that each gene associated with an entity is assigned one or more of the following specificity annotations: specific, enriched, selective, expressed, abundant, and/or low confidence (Edgar et al., 2013). The annotations are derived from the literature and/or from bioinformatic calculations. The calculations consider the source from which the gene–entity association was established and the distribution of the gene expression in LMD. Criteria include how rare is the gene in the database, how specific it is to a certain cell type or tissue, and whether there is extensive evidence for the expression of the gene in the tissue.

In addition, the gene score considers the entity type in which the expression is observed. Genes listed in organ/tissue, anatomical compartment or cell cards are ranked higher than genes with the same specificity annotations, which are listed in Large Scale Data Sets entities that are not linked to any of the above (tissue, compartments, etc.). Supplementary information elaborating on the determination of the gene annotation and the given scores, with additional details, is summarized in the Supplementary S1 Table.

After defining the gene scores, the gene set of each entity and the query gene set can be viewed as gene expression vectors. The entity gene–set vector holds defined scores for each of its genes and zero for all other genes, while the query gene–set vector is a binary vector that holds the value 1 for each of the query's genes and 0 for all other existing genes. The affinity between the query gene set and each of the entities is measured by the scalar product of the two vectors (i.e., the sum of the scores of the entity genes matched to the query gene set). The choice of normalization factor and the details of the score levels are described in ''Gene Scores, The matching score algorithm'' in Supplementary S2 Appendix.

The entity scores are divided into three levels, representing the strength of the results (high, medium, or low), which is indicated by the color of the score bar. This categorization is performed by a two-step procedure that runs automatically before each release. The first step is determining the threshold for medium and high scores for a group of query gene sets with varying sizes. The second step uses a linear regression between the various query sizes and their computed medium/high scores in order to create an equation from which the thresholds in the first step can be computed easily for any query gene size.

The first step of the automatic procedure uses a set of 50 test cases. From each test case, six gene lists of different sizes are generated (5 to 300 genes). The matching

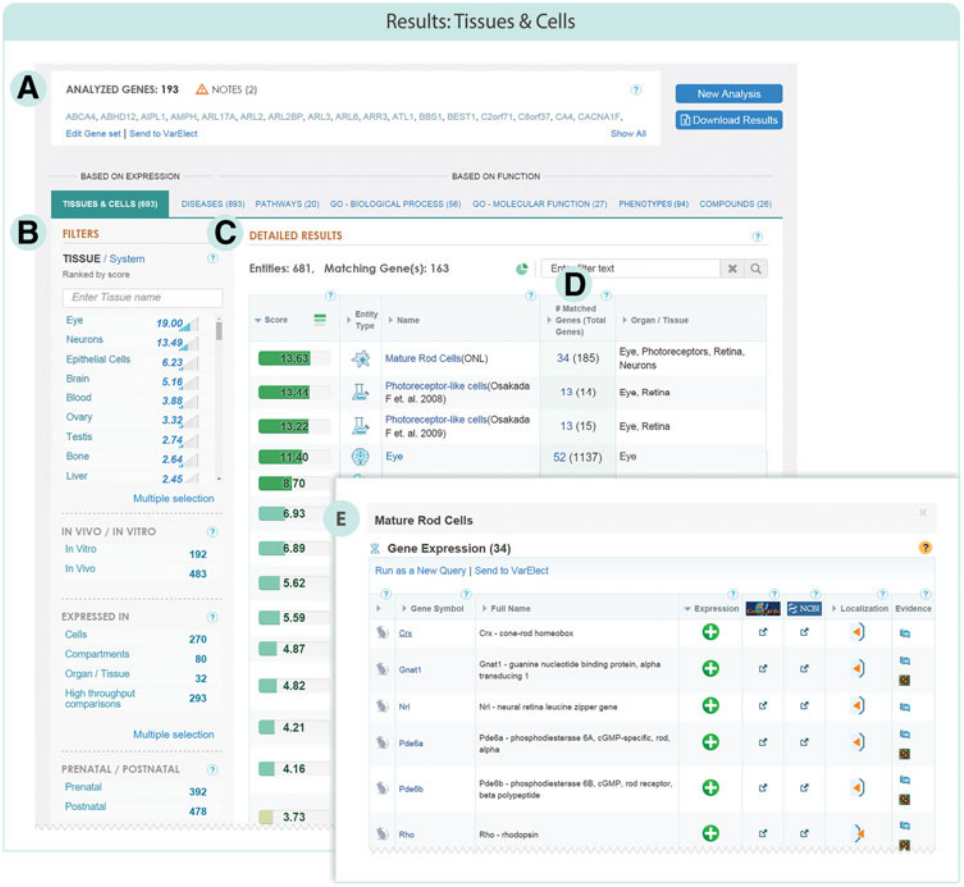

FIG. 3. Tissues and Cells results. **(A)** The Analyzed genes are the queried genes that were identified and included in the analysis. The ''Notes'' indicate genes in the query that were found to be abundant or defined as housekeeping genes in human. These genes get lower scores in the Tissue and Cells matching analysis. **(B)** The filters panel allows for filtering genes specifically expressed in Tissue/system, *In vivo/In vitro*, 'Expressed in' (cells, anatomical compartments, organs and tissues, and/or high throughput comparisons and large-scale dataset samples), Prenatal/Postnatal. **(C)** The detailed results table presents all entities in which at least one of the analyzed genes is expressed, along with links to their cards in LMD. **(D)** A link to the list of the matched genes and additional information for them (for example, ''Mature Rod Cells''). **(E)** The list of matched genes linked to the specific entity in LMD (connected to ''Mature Rod Cells'').

algorithm applied on these gene lists produces a range of typical scores for each query size. In order to obtain the high and medium threshold automatically, a preliminary analysis was performed on many control microarray experiments. Each experiment represents a known cell/compartment/tissue and therefore was expected to produce high scores for the highly relevant entities, medium scores for entities with modest relevance, and lower scores for weakly related entities.

By analyzing the distribution curves for all control sets, we established the percentiles of entities that produce medium and high scores. These determined percentiles enabled the high and medium boundaries in the aforementioned first step to be computed automatically. In the second step, a linear regression is applied between the various query sizes and their high or medium scores from which an equation for computing these boundaries in the general case is generated.

Diseases

Gene–disease relations in GeneAnalytics are divided into the following categories, indicated in the GeneAnalytics results (Fig. 4):

a) Gene associations along with their confidence classifications as derived from MalaCards data sources. Since each data source has its own annotation terminology, Table 3 categorizes all of the possible disease–gene associations in descending order according to their source-associated confidence, which is later transformed into a GeneAnalytics score.
b) Genes that are significantly up- or downregulated in disease tissues in comparison to their healthy counterparts. Differential gene expression profiles are derived from High Throughput experiments extracted from GEO or from the literature, and analyzed using LMD algorithms (Supplementary S3 Appendix).

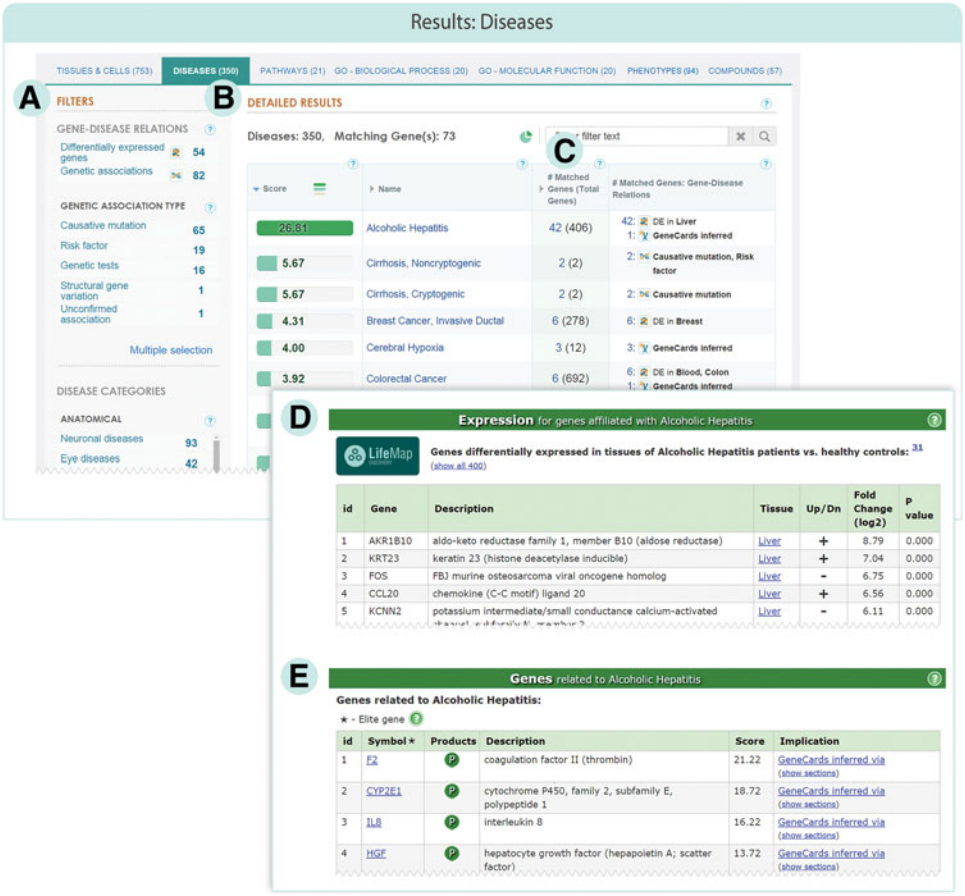

FIG. 4. GeneAnalytics Disease results. **(A)** The disease filter enables filtration of results by gene–disease associations and disease categories obtained from the MalaCards database. **(B)** The detailed results table presents diseases matched to the queried gene set. Each disease is linked to its card in MalaCards. **(C)** Clicking on the number of matched genes opens a list of the matched genes and associated information. **(D)** Differentially expressed genes ('expression'), and **(E)** disease-related genes in their respective sections in a disease card in MalaCards. Both sections serve as evidence for each matched disease in the GeneAnalytics disease category.

c) GeneCards inferred genes (i.e., genes with the disease name mentioned anywhere in the relevant GeneCards webcard, e.g., in the publication section). This is a somewhat weaker association, which often does not imply causality.

The disease matching score is calculated in three steps:

a) Each gene associated with each disease receives a score based on the gene–disease relations described in the disease data modeling section (Supplementary S2 Table):

TABLE 3. DISEASE–GENE ASSOCIATIONS FROM MANUALLY CURATED GENETIC SOURCES

Association category	*Source*
Causative mutation	ClinVar, OMIM, Orphanet
Risk factor	ClinVar, OMIM, Orphanet
Resistant factor	ClinVar, OMIM
Genetic tests	GeneTests
Drug response	ClinVar
Structural gene variation	OMIM, Orphanet
Unconfirmed association	OMIM, Orphanet

See the Supplementary S3 Appendix for additional details.

(i) Genes with a genetic association to the disease receive a score according to the association category described in Supplementary S3 Table. A gene linked with multiple filter categories is assigned the strongest association score among them.

(ii) Differentially expressed genes are binned and scored based on their rank in the list of differentially expressed genes in the diseased vs. normal tissue analysis (analyses were performed as per all High Throughput experiments, detailed in ''Differentially expressed genes identification algorithm'' in Supplementary S2 Appendix).

(iii) Genes with ''GeneCards inferred'' relations receive a score based on the number of sections in GeneCards in which the disease appears.

b) Each gene may have more than one type of relationship with the disease; the final gene score a disease receives is the highest among all of the possible scores mentioned in point a above.

c) The gene–disease matching score is calculated based on scores of each of the matched genes, the number of matched genes, and the total number of genes associated

with the disease in MalaCards (used for normalization). The scoring function is identical to the one used in the Tissues and Cells category (see ''The matching score algorithm'' in Supplementary S2 Appendix).

The disease results category in GeneAnalytics includes several filters that enable the user to focus on the results of interest (Fig. 4A).

a) Gene–disease relations. This enables the user to filter for gene–disease relation types, including differentially expressed genes and specific types of genetic associations. Selection of 'differentially expressed genes' (DE) or 'genetic association', will only show diseases for which their matched gene set includes at least one differentially expressed or genetically associated gene, respectively. This filtration caters to users who are interested in diagnostic disease markers, in the case of differentially expressed genes, or those with genetically associated variants for specific diseases. Importantly, the matching score for each disease category is recalculated following filtration, so the scoring algorithm considers only entities that contain at least one gene matching the requested filter terms.
b) Disease categories. This filter enables the user to focus on specific disease categories, as defined by MalaCards categorizations. MalaCards categorizes diseases into anatomical (e.g., eye, ear, liver, blood) and global (rare, fetal, genetic, cancer, and infectious) diseases. The categorization is based on either the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) (Organization, 1992) or on the MalaCards classification algorithm that utilizes category-specific keywords contained in the disease names and annotations, as well as textual heuristics. For example, if the disease name includes the words 'tumor' or 'malignant,' it is classified as a cancer disease (Rappaport et al., 2014). Further, a disease can be associated with more than one category.

Pathways

In GeneAnalytics, matched SuperPaths appear with their matching score and link to the relevant webcard in PathCards, as well as the list of matched genes and total number of genes associated with each SuperPath. The user can then expand each matched SuperPath to view the list of its clustered pathways with links to their original individual pathway sources and to the relevant genes in the user's query (Fig. 5).

The scoring algorithm in the pathways category is based on the algorithm used by the GeneDecks Set Distiller tool (Stelzer et al., 2009). Briefly, all genes in each SuperPath are given a similar weight in the analysis, and the matching score is based on the cumulative binomial distribution, which is used to test the null hypothesis that the queried genes are not over-represented within any SuperPath (see more details in Supplementary S4 Appendix). As in all sections, the score is represented by a colored score bar and classified by its quality (see details in the Tissues & Cells matching algorithm description).

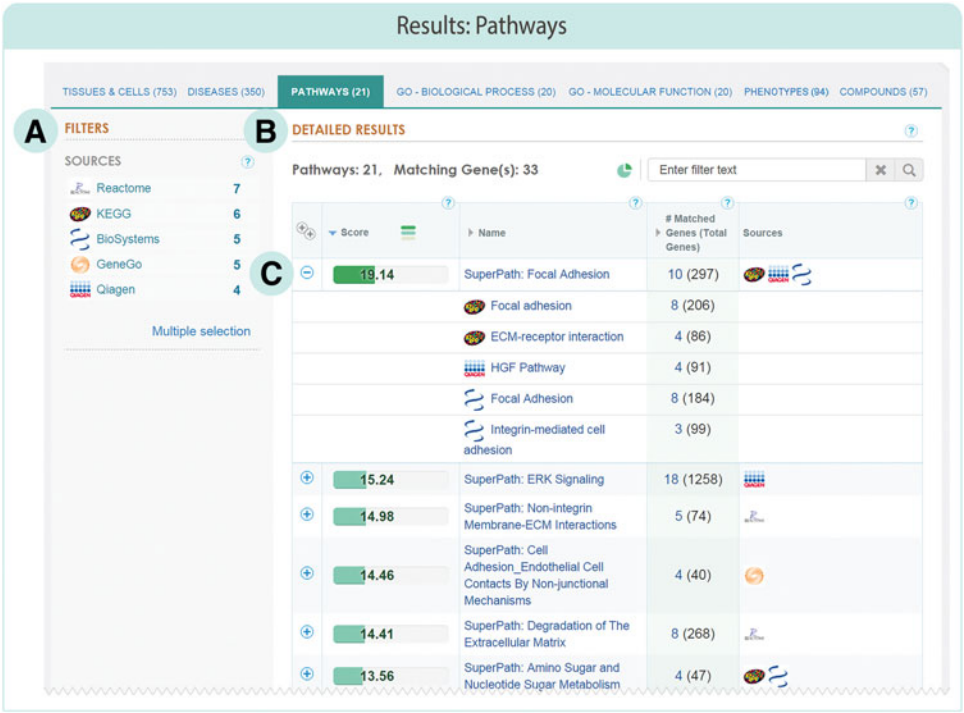

FIG. 5. GeneAnalytics Pathways results. **(A)** The pathway filters panel enables filtration of results according to their data sources. **(B)** The detailed results table includes all of the matched SuperPaths, presented in descending score and with links to the related card in PathCards. **(C)** Each SuperPath includes one or more pathways from different sources. Clicking on the plus sign exposes the names of the separate pathways that comprise the SuperPath, with links to the pathway page in the original data source.

Pathway unification is employed on all of the sources found in GeneCards. GeneAnalytics enables users to concentrate on as many sources as desired by applying a source filter.

Gene Ontology (GO) terms and phenotypes

The matching algorithm for both GO terms and phenotypes is based on the binomial distribution and is identical to that used in the pathways category (see Supplementary S3 Appendix).

Drugs and compounds

The GeneAnalytics compounds results category takes advantage of multiple sources that cover more than 83,000 compounds, approximately 45,000 of which are associated with genes. GeneAnalytics applies a unification process which reduces the number of compounds with associated genes by more than half, from ~45,000 to ~20,000 compounds (Table 1). This robust process saves time in reviewing identical compounds presented under various names by different data sources and enables massive aggregation of genes per compound, and is featured in GeneCards.

The compound unification process seeks out similar compounds described in different data sources, and is based on the following rules:

a) Unification of compounds with exact identical names (case/dash- insensitive).
b) Unification of compounds with identical identifiers, more specifically both a Chemical Abstracts Service (CAS) number (unique numerical identifier assigned to chemical substances) and a PubChem ID (PubChem is an NCBI database providing information on the biological activities of small molecules). Note that not all compounds have these identifiers, nor do all databases provide these identifiers for their compounds.
c) Unification of compounds with either an identical CAS number or PubChem ID and identical synonyms. Note that different compounds might have identical synonyms and therefore, only compounds with at least one identical identifier and one identical synonym are unified.
d) Metabolite unification based on metabolite family and gene sharing: several metabolite families contain thousands of compounds with almost identical names, many of which are associated with an identical list of genes. In GeneAnalytics, prevalent metabolite family subgroups belonging to Triglycerides, Diglycerides, Phosphatidylcholines, Phosphatidylethanolamines, have been unified based on identical lists of associated genes. These groups are described in the user guide (geneanalytics.genecards.org/user-guide#1628).

Unified compounds are shown with links to all supporting data sources, providing further information and its relevance to the evaluated genes, while the original compound name is shown near its data source. The matching algorithm is based on the binomial distribution and is identical to that used in the pathways category (see Supplementary S3 Appendix).

The compound category in GeneAnalytics provides the opportunity to explore relationships between compounds and gene sets, to define potential drugs and their mechanisms of action and to facilitate drug target discovery.

Results

Tissues and cells

GeneAnalytics provides novel and meaningful contextualization of various input gene sets. These include NGS-derived mutated genes, as well as differentially expressed genes identified by a microarray experiment, RNA sequencing, *in situ* hybridization or real-time PCR. In addition, lists of genes encoding protein targets of a specific drug or proteins known to be a part of a specific molecular pathway or biological process are recommended for this analysis.

The Tissues and Cells results category in GeneAnalytics leverages the extensive and high quality gene expression data available in LMD. Results include the following [cf. (Edgar et al., 2013)]: *in vivo* cells, *in vitro* cells, anatomical compartments, organs, and tissues (collectively referred to as ''anatomical entities''), whose reported gene expression profiles match the query gene set (Table 2 and Fig. 3). This category provides matching to cases in which given normal (healthy, wild type, untreated) tissues and cells show differential expression relative to control tissues. It excludes genes differentially expressed in diseased tissues and cells, as compared to healthy tissues, which are available in the disease category.

This distinction is essential for results interpretation and eliminates incorrect association of aberrant gene expression profiles with normal tissues and cells. Importantly, if members of a highlighted gene subset appear in both of the above scenarios, they will be shown in parallel in the two relevant categories.

Diseases

The diseases category of the GeneAnalytics results (Fig. 4) harnesses the broadly integrated information available in MalaCards (Rappaport et al., 2013; 2014). MalaCards mines and merges over 60 data sources and provides a comprehensive list of diseases, unified by various annotations, such as names, acronyms, and OMIM identifiers. Each disease entry has a webcard containing wide-ranging disease information, including related diseases, genes with relevant disease-implicating annotations, genetic tests, variations with pathogenic significance, expression information and more.

Pathways

The pathway category in GeneAnalytics is empowered by the information available in PathCards, the biological pathways database (Belinky et al., 2015). PathCards unifies 12 pathway sources into SuperPaths, generating an explicit list of the included pathways, as well as their associated genes. This unification is important because it integrates pathway information from various sources, thereby introducing novel gene–gene associations within the unified pathways.

The PathCards algorithm enables the unification of over 3000 pathways, obtained from all of its data sources, into a set of approximately 1000 pathway clusters called ''SuperPaths'' (see Table 1 for statistics). Each SuperPath encompasses up to 70 pathways and is presented in a webcard that includes an aggregated gene list and links to relevant pathway sources. The PathCards algorithm (Belinky et al., 2015) estimates pathway similarity by overlapping gene content, with the assumption that the gene content defines the pathway identity.

Thus, unifying pathways by names and/or hierarchal clustering, which significantly vary between different pathway sources, is simplified. The chosen SuperPath name is that of the most 'connected' pathway in the cluster, namely the pathway with the highest gene similarity to the other pathways in the SuperPath.

Gene ontology (GO) terms and phenotypes

Gene Ontology analysis in GeneAnalytics exploits the information available in the GO project (www.geneontology.org) and integrated in GeneCards (Safran et al., 2010). GO provides ontology of defined terms representing gene product properties. This project uses a set of structured, controlled vocabularies for the annotation of genes and gene products in an effort to standardize their attribute representation across species (Gene Ontology, 2010). GO consists of three hierarchically structured ontologies that describe gene products in terms of their associated biological processes, cellular components, and molecular functions.

Since its inception, many tools have been developed to explore, filter, and search the GO database (Conesa et al., 2005; Eden et al., 2009; Nogales-Cadenas et al., 2009; Zheng and Wang, 2008). One of the most common applications of the GO vocabulary is enrichment analysis (i.e., the identification of GO terms that are significantly over-represented in a given set of genes). GO terms can be either tissues or cells in the embryo and/or in the adult, pathways, diseases, or other biological functions. As such, GO enrichment analysis in GeneAnalytics provides supporting information about the functional roles of the query gene set.

Phenotype analysis in GeneAnalytics utilizes the Mouse Genome Informatics (MGI—www.informatics.jax.org) data that are presented in GeneCards. MGI phenotypes describe the outcome of either naturally occurring or induced mouse mutations. These aberrations are genetically characterized and portrayed with high resolution in a hierarchical phenotype tree (similar to the GO term tree). The observed outcome of genetic abnormalities in mice is a powerful instrument for inferring the functional influence of genes, therefore identifying enriched mouse phenotypes may give insight as to the biological processes involved in various experimental conditions.

Enriched GO terms results are divided into two categories, corresponding to two of the three ontologies, 'biological processes' and 'molecular function'. A third category displays mouse phenotype enrichment results. The terms in each category are ranked by their matching score, and appear with a direct link to the relevant webcard in either the AmiGO browser (www.geneontology.org) or MGI website (www.informatics.jax.org). Also shown are the list of the matched genes from within the input gene set and the total number of genes associated with each enriched GO term.

Drugs and Compounds

The Compounds analysis in GeneAnalytics derives its information from GeneCards, which associates genes with biochemical compounds and drugs. This information is extracted from several data sources, which contain extensive biochemical and pharmacological information about drugs, small molecules and metabolites, their mechanisms of action, and their targets. Gene–compound associations are determined either by direct binding between the compound and the gene product (e.g., enzyme, carrier, transporter), or by demonstration of a functional relationship (e.g., pharmacogenomics, genetic variants and drug pathways affecting drug activity).

The wide range of compounds in the Compound category naturally includes many in the realms of glycomics, metabolomics, and lipidomics. Consequently, GeneAnalytics has the capacity to portray various gene sets enriched in such compounds, hence point to relationships between genomics and other Omics domains. Such relations would include cases in which carbohydrates, metabolites, or lipids serve as ligands for receptors, substrates, or regulators for enzymes, as well as post-translational modifiers of proteins. In such cases, compounds such as N-acetylglucosamine, dopamine, or phosphatidylserine could readily be targets for enrichment in inputted gene sets.

Discussion

Advantages and applications

The GeneAnalytics expression-based analysis provides gene associations with both normal and diseased tissues and cells, leveraging the proprietary manually curated gene expression data available in LifeMap Discovery (LMD) (Edgar et al., 2013). The calculation of the matching score for gene expression in normal tissues and cells is based on a search for the highest similarity between expression vectors.

Several other approaches exist for identifying tissues, functions, and phenotypes maximally relevant to a gene set. Common approaches used in gene set enrichment analysis (GSEA), the DAVID functional annotation tool (Sherman et al., 2007), and the GeneDecks Set Distiller tool (Stelzer et al., 2009), statistically evaluates the over-representation of the query genes in association with a given descriptor as compared to the entire gene landscape. This can be achieved, for example, by applying a binomial or hypergeometric distribution, thereby giving all genes an identical weight, and ignoring the unique characteristics of particular genes in the entity.

In contrast, the GeneAnalytics matching algorithm considers many aspects of gene expression, such as tissue specificity of a gene, whether it serves as a cellular marker, as well as the type of entity in which it is expressed. As a result of this flexibility, the GeneAnalytics Tissues and Cells analyses can assist in the identification and characterization of tissue samples or cultured cells by their expressed genes and to validate their purity. Additional applications include evaluation of cell differentiation protocols, cell sorting quality assessment, or tissue dissection procedures. In addition, the results of these analyses can enhance the discovery of new selective markers for tissue and cells.

The disease results category in GeneAnalytics is powered by MalaCards, the human disease database (Rappaport et al., 2013; 2014), which integrates information from multiple sources. This includes manually curated information on disease expression data drawn from comparisons between diseased tissues and their matched normal control tissues. In addition to the high quality data presented in GeneAnalytics, its matching algorithm considers the specific type of evidence each gene has for its association with the disease, whereby stronger gene–disease associations augment the score.

The results provided in the disease category can be used to explore relationships between diseases and gene networks, as well as enhancing therapeutic discoveries and identification of causative genetic variations for a disease. Specifically, this

category can be used to identify relevant diseases for a list of genes, originating from prioritized variants generated by DNA sequencing experiments, or to identify the known variation-mapped genes for specific diseases rapidly. This category can also be used to identify the diseases most relevant to lists of differentially expressed genes generated under a wide range of experimental conditions.

The pathway category in GeneAnalytics is based on data extracted from PathCards (pathcards.genecards.org), which portray pathways (SuperPaths) consolidated from twelve sources (Belinky et al., 2015), which integrates numerous sources to provide consolidated information in the form of SuperPaths. This has a clear advantage in providing gene set analysis related to twelve different pathway sources, as opposed to several other relevant tools that typically base their analyses on one or just a few pathway sources [e.g., DAVID, (Huang da et al., 2009b)].

Further, GeneAnalytics benefits from the success of PathCards to decrease redundancy while enhancing the inference of gene-to-gene relations, necessary for pathway enrichment analysis. Scrutinizing enriched pathways helps uncover the underlying mechanisms involved in specific treatments, diseases or experimental conditions.

Comparison to other tools

GeneAnalytics offers the biomedical research community a useful tool for gene set analysis, with significant advantages over several comparable tools, such as the popular commercial product Ingenuity Pathway Analysis (IPA) and the free academic DAVID tool. These include:

a) *A simple-to-use interface*, providing immediate insight via results presented by biological categories. The effective interface enables interpretation and contextualization without the need for complex bioinformatics expertise or external tools. IPA has a very robust user interface with many visualization options, however there is steep learning curve required to acquaint oneself with this tool.
b) *A rich integrated data resource*, taking advantage of constantly updated knowledge-base within the GeneCards Suite of databases, namely MalaCards, PathCards, LifeMap discovery, and GeneCards, each possessing its own biological expertise. Using robust data collection procedures, they encompass information from over 150 data sources that are both automatically mined and also manually curated. IPA has a powerful, manually curated database with defined ontology classes that incorporate evidence from articles to generate a complex network that carries various types of annotations. DAVID has many sources of functional annotations that have not been updated recently (last update was on 2009, https://david.ncifcrf.gov/content.jsp?file=update.html).
c) *Proprietary gene expression data and unique data modeling*. GeneAnalytics leverages gene expression data from LifeMap Discovery, encompassing enrichment in organs/ tissues, anatomical compartments and cells, with relations to stem cell and developmental biology. In addition, these data are unique in their combination of manually curated and wide scope of sources interrogated for expression information. Both IPA and DAVID have only a weak representation of expression data in data repositories.
d) *Novel matching algorithm*. The expression and disease-based matching algorithms consider gene annotations including gene–disease association types and gene specificity, as well as enrichment or abundance in each specific tissue or cell. This reflects the added value of high quality data, such as that from manually curated sources, which is given extra weight in our algorithms. DAVID and IPA statistically evaluate [using Fisher's exact test, (Huang da et al., 2009a; Kramer et al., 2014)] the over-representation of descriptors within a gene set, but does not take into account qualitative information for gene-descriptor associations.
e) *Categorized output*. Enrichment results are classified into distinct term categories to enhance gene set interpretation. Both IPA and DAVID have categorized views of the data but special effort was made in GeneAnalytics to remove clutter and simplify information retrieval to accommodate novice users.
f) *Powerful filtration*. Filters enable the user to obtain a bird's eye view of shared descriptors and focus on specific subsets of the results. While many options exist in IPA for slicing and displaying information, this feature in GeneAnalytics is intuitive and flexible, quickly allowing one to grasp enrichment trends and easily narrow the criteria for displaying the data.
g) *Iterative analyses.* All sections are equipped with a ''drill down'' subquery mechanism based on the original set filtered by genes that match the selected entity. This feature that is also available in IPA but missing in DAVID hastens the workflow in finding pertinent shared descriptors and key genes.
h) *Supporting evidence links for matched biological terms.* All matches are directly linked to detailed cards in the knowledge base, providing further information about the specific entity, along with supporting evidence for the entity–gene relation. Direct links to the relevant external data sources are also available.

Future directions

Future developments aim to enhance the GeneAnalytics algorithm (e.g., by providing an option to employ a user-defined background gene list), as well as enhancing the interpretation and contextualization of the query gene set. The latter will be achieved by presenting detailed annotation of each gene, including score justification, and through improving filtration capacities by providing biological category and gene annotation filters.

In the Tissues and Cells category, all gene specificity annotations will be presented and will be filterable. This feature will further assist users in focusing on specific markers, or eliminating results obtained from abundant genes. Differential expression of proteins, based on the meta-analysis of several proteomic databases, will be integrated with transcriptomic data, thereby enhancing the information already available in GeneAnalytics (Simon Fishilevich et al., 2016).

In the Disease category, diseases belonging to the same MalaCards disease family will be clustered, specifically diseases that are lexically grouped together. Usually the disease family contains different disease subtypes, modes of

inheritance, or genetic basis. Future versions will present the best matched disease from each family, with an option to expand and view all matched diseases from this family. This presentation will avoid redundancy and enhance results interpretation.

Categorization filters will be applied to the pathways, GO terms, and compounds sections, thus enabling the filtration of matched results by their biological relations (e.g., relation to metabolism or diseases, or by biochemical and/or activity type). In the Drugs and Compounds category, additional clinical-related data sources will be integrated as they become incorporated into GeneCards and MalaCards.

A clustering feature that will enable grouping of matched entities from different results sections that share similar genes will be developed for future versions. These clusters could shed light on relations between different biological categories and highlight functionality of subsets of genes. GeneAnalytics will also enable users to save, manage, and share gene sets and analysis results via personalized accounts.

Enrichment categories will be improved or newly added. For example, the use of GO slim in the existing GO terms category, the addition of a regulatory features category that might include transcription factor binding sights and micro-RNA targets. Finally, we plan to enhance GeneAnalytics by visualizations, employing varied graphical display items of the results.

Such attributes make GeneAnalytics a broadly applicable postgenomics data analyses and interpretation tool for translation of data to knowledge-based innovation in various Big Data fields such as precision medicine, ecogenomics, nutrigenomics, pharmacogenomics, vaccinomics, and others yet to emerge on the postgenomics horizon.

Acknowledgments

Funding for this research was provided by LifeMap Sciences Inc. California (USA), the NHGRI grant U41HG003345, and the Crown Human Genome Center and the Nella and Leon Benoziyo Center for Neurosciences at the Weizmann Institute of Sciences. LifeMap Sciences was involved in the design of the databases and web tools, and in the interpretation of the data. We wish to thank Iris Lieder for fruitful discussions and Galia Duchovnaya for graphical design.

Author Disclosure Statement

GeneAnalytics utilizes data from GeneCards and other GeneCards suite members (www.genecards.org). No copyright issues exist since GeneAnalytics belongs to the GeneCards suite and is authored by the same team. The Weizmann Institute team receives a grant from LifeMap Sciences Inc. (http://www.lifemapsc.com/) under a research and licensing agreement, whereby scientific results may be incorporated into software products marketed by the company. This article was generated as a collaborative effort of the academic and corporate entities. The following authors are employed by LifeMap Sciences: GS, IP, YGG, AK, YM, SK. GS is also a consultant to DL at the academic institution.

References

Backes C, Keller A, Kuentzer J, et al. (2007). GeneTrail—Advanced gene set enrichment analysis. Nucleic Acids Res 35, W186–192.

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

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, and Robles M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, and Lempicki RA. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4, P3.

Eden E, Navon R, Steinfeld I, Lipson D, and Yakhini Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48.

Edgar R, Domrachev M, and Lash AE. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30, 207–210.

Edgar R, Mazor Y, Rinon A, et al. (2013). LifeMap Discovery: The embryonic development, stem cells, and regenerative medicine research portal. PLoS One 8, e66629.

Geffers L, Herrmann B, and Eichele G. (2012). Web-based digital gene expression atlases for the mouse. Mamm Genome 23, 525–538.

Gene Ontology C. (2010). The Gene Ontology in 2010: Extensions and refinements. Nucleic Acids Res 38, D331–335.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, and McKusick VA. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33, D514–517.

Huang DA W, Sherman BT, and Lempicki RA. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37, 1–13.

Huang DA W, Sherman BT, and Lempicki RA. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44–57.

Jupe S, Jassal B, Williams M, and Wu G. (2014). A controlled vocabulary for pathway entities and events. Database (Oxford) 2014.

Kanehisa M, Goto S, Furumichi M, Tanabe M, and Hirakawa M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38, D355–360.

Kramer A, Green J, Pollard J, JR, and Tugendreich S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics 30, 523–530.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, and Mesirov JP. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740.

Maezawa S, and Yoshimura T. (1991). Sequence of critical events involved in fusion of phospholipid vesicles induced by clathrin. Biochim Biophys Acta 1070, 429–436.

Maiella S, Rath A, Angin C, Mousson F, and Kremp O. (2013). [Orphanet and its consortium: Where to find expert-validated information on rare diseases]. Rev Neurol (Paris) 169, S3–8.

Matthews L, Gopinath G, Gillespie M, et al. (2009). Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37, D619–622.

Nogales-Cadenas R, Carmona-Saez P, Vazquez M, et al. (2009). GeneCodis: Interpreting gene lists through enrichment analysis and integration of diverse biological information. Nucleic Acids Res 37, W317–322.

Organization WH. (1992). The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines. World Health Organization, Geneva: World Health Organization.

Rappaport N, Nativ N, Stelzer G, et al. (2013). MalaCards: An integrated compendium for diseases and their annotation. Database (Oxford) 2013, bat018.

Rappaport N, Twik M, Nativ N, et al. (2014). MalaCards: A comprehensive automatically-mined database of human diseases. Curr Protoc Bioinformatics 47, 1.24.1–1.24.19.

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

Safran M, Dalah I, Alexander J, et al. (2010). GeneCards Version 3: The human gene integrator. Database (Oxford) 2010, baq020.

Sherman BT, Huang DA, Tan Q, et al. (2007). DAVID Knowledgebase: A gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics 8, 426.

Simon Fishilevich SZ, Kohn A, Iny-Stein T, Kolker E, Safran M, and Lancet D. (2016). Genic insights from integrated human proteomics in GeneCards. Database (Oxford).

Smith CM, Finger JH, Hayamizu TF, et al. (2014). The mouse Gene Expression Database (GXD): 2014 update. Nucleic Acids Res 42, D818–824.

Stelzer G, Inger A, Olender T, et al. (2009). GeneDecks: Paralog hunting and gene-set distillation with GeneCards annotation. OMICS 13, 477–487.

Wu C, MacLeod I, and Su AI.(2013). BioGPS and MyGene.info: Organizing online, gene-centric information. Nucleic Acids Res 41, D561–565.

Wu CH, Apweiler R, Bairoch A, et al. (2006). The Universal Protein Resource (UniProt): An expanding universe of protein information. Nucleic Acids Res 34, D187–191.

Zhang B, Kirov S, and Snoddy J. (2005). WebGestalt: An integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33, W741–748.

Zheng Q, and Wang XJ. (2008). GOEAST: A web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res 36, W358–363.

Address correspondence to:
Dr. Gil Stelzer
Molecular Genetics
Weizmann Institute of Science
Herzl St 234
Rehovot 7610001
Israel

E-mail: gil.stelzer@genecards.org

Abbreviations Used

CAS = Chemical Abstracts Service
DAVID = Database for Annotation, Visualization, and Integrated Discovery
DE = differentially expressed
GEO = Gene Expression Omnibus
GO = Gene Ontology
GSEA = Gene Set Enrichment Analysis
HT = High Throughput
KEGG = Kyoto Encyclopedia of Genes and Genomes
LMD = LifeMap Discovery
LSDS = Large Scale Data Sets
MGI = Mouse Genome Informatics
NGS = next generation sequencing
OMIM = Online Mendelian Inheritance in Man