GeneDecks: Paralog Hunting and Gene-Set Distillation with GeneCards Annotation

Gil Stelzer,1 Aron Inger,1 Tsviya Olender,1 Tsippi Iny-Stein,1 Irina Dalah,1 Arye Harel,1 Marilyn Safran,1,2 and Doron Lancet1

Abstract

Sophisticated genomic navigation strongly benefits from a capacity to establish a similarity metric among genes. GeneDecks is a novel analysis tool that provides such a metric by highlighting shared descriptors between pairs of genes, based on the rich annotation within the GeneCards compendium of human genes. The current implementation addresses information about pathways, protein domains, Gene Ontology (GO) terms, mouse phenotypes, mRNA expression patterns, disorders, drug relationships, and sequence-based paralogy. GeneDecks has two modes: (1) Paralog Hunter, which seeks functional paralogs based on combinatorial similarity of attributes; and (2) Set Distiller, which ranks descriptors by their degree of sharing within a given gene set. GeneDecks enables the elucidation of unsuspected putative functional paralogs, and a refined scrutiny of various gene-sets (e.g., from high-throughput experiments) for discovering relevant biological patterns.

Introduction

Finding shared attributes among genes is becoming very important in biology and Systems Biology. The general problem of annotation transfer entails the use of documented information about genes in order to uncover functional similarities. Annotation transfer tools based on local sequence similarity, such as Goblet (Groth et al., 2004) and AutoFACT (Koski et al., 2005), are powerful because they increase the inference for unknown and unstudied genes. Annotation transfer is justified by sequence similarity between a known gene and a new or unstudied one; however, the specific similarity threshold, which is legitimate for this transfer, is up for debate (Rost, 2002). Other approaches in the same vein include sharing of sequence motifs (Sigrist et al., 2002), 3D structure (Whistock et al., 2003), and/or 3D motifs (Wallace et al., 1996). The closest functional paralog can be found for any gene of interest, when pairwise comparison is done over the entire genome. For large gene sets, such as those obtained from microarray experiments, one can attempt to uncover underlying attributes shared by some or all these genes. The main efforts invested here are in the development of utilities that tackle the enrichment of certain descriptors (specific annotations for a given biological feature), within such sets (Dennis et al., 2003; Khatri et al., 2002).

The importance of functional paralog finding is exemplified by several research challenges. One is gene-targeting experiments, where in many cases no obvious phenotype can be observed. This may be due to the existence of functional paralog(s) not always discoverable by a sequence similarity search. A related advantage may ensue in studies of synthetic lethality (Kaelin, 2005; Kennedy and D’Andrea, 2006), whereby a broader search for functional partners may complement sequence-based searches. In addition, general function-based gene-similarity predictions may help expand the search for new drug targets (Ofran et al., 2005). Although experimental strategies such as complementation studies are often useful (Brenner, 1974; Gonzalez-Santos et al., 2005; Grallert and Nurse, 1997), a computational tool that could systematically and comprehensively enhance the search capacity would also be beneficial.

Because many disorders and other biological phenotypes are orchestrated by multiple genes, there is a growing need for gaining system-wide insights (Williams, 2006). Collecting data from numerous gene sets is an obvious means for obtaining a broad perspective on cellular events. The consequent high-throughput experiments, ranging from two-dimensional gels to whole transcriptome scrutiny, commonly result in sets of tens to hundreds of genes. It is then necessary to wisely interrogate the wealth of gene annotation to shed light on the biological significance of such sets (Gat-Viks et al., 2003; Gibbons and Roth, 2002). The core scheme is to characterize a given gene set by seeking the most enriched descriptors maximally shared by its members, noting their enrichment compared to the entire gene. Examples are tools based on shared Gene Ontology (GO) terms (Groth et al., 2004; Khatri

1Departments of Molecular Genetics and 2Biological services, Weizmann Institute of Science, Rehovot 76100, Israel.
et al. 2002), as well as on a broader range of attributes (Dennis et al., 2003).

We introduce GeneDecks, a research tool aimed at discovering gene-pair and gene-set relations, exploiting the rich data found within GeneCards, the human gene compendium (www.genecards.org), which collects and unifies data from over 80 sources, for over 53,000 genes, among which over 22,000 encode proteins (Rebhan et al., 1997; Safran et al., 2003). The first GeneDecks mode, coined Paralog Hunter, attempts to find functional paralogs based on various shared annotations rather than sequence similarity alone. It features a variety of attributes, including pathways, domains, and gene to compounds or disorder relationships. The second GeneDecks mode, Set Distiller, produces a ranked list of shared GeneCards descriptors for a given gene set. It thus addresses the needs to quantify and organize voluminous high throughput results, for example, from expression and SNP microarrays, chromatin immunoprecipitation (ChIP) experiments, and mass spectrometry proteomics analyses. GeneDecks shares features with other gene-set oriented tools such as David (Dennis et al., 2003), NetAffx (Cheng et al., 2004; Liu et al., 2003), and Genomica (Segal et al., 2003). We provide comparisons to one of these (David), and highlight Set Distiller’s advantages.

Methods

GeneDecks currently uses 18 sources from the GeneCards (Safran et al., 2003) Version 3 MySQL database that include almost 35,000 descriptors (available on request) associated with 10 attributes listed in Table 1 (view at the end of the article). (Our analyses are based on version 2.38.1 of GeneCards.) A descriptor is a specific instance of a given attribute, for example, Oestrogen_recept is a specific domain, and methotrexate is a specific compound.

Paralog Hunter algorithm

Paralog Hunter, which is implemented in Java and available online (www.genecards.org/v3/genegecks), calculates similarity scores between each query gene and all remaining candidate genes in the GeneCards database, for each of the 10 attributes that appear in Table 1. For all attributes except Gene Ontology, sequence paralogy, and expression, the similarity score between a query gene and a candidate gene is calculated in the following manner: each descriptor score (DS) is the result of dividing its rank by \( \log_{10} \) of its frequency in the database

\[
DS = \frac{\text{Rank}}{\log_{10}(\text{Dscr.Freq})}.
\]

Descriptor ranks are each assigned the value of 1, except for those associated with the GO attribute, which are assigned the descriptor’s evidence code (Buza et al., 2008); for example, Inferred from Direct Assay (IDA) will receive a descriptor score of 5

\[
\text{Rank} = \left\{ \begin{array}{ll}
5 & \geq \text{Evid.Code} \geq 1 \\
1 & \text{otherwise}
\end{array} \right. \]

The assignment of ranks to descriptors is an attempt to discriminate between automatic versus human-ascribed descriptors to genes; this is currently available only for GO terms. The attribute score (AS) is the sum of the descriptor scores for those descriptors shared by both the query gene and the candidate gene, divided by the sum of the descriptor scores for all descriptors associated with the query gene

\[
AS = \frac{\sum DS_{CQ}}{\sum DS}.
\]

For the sequence paralogy attribute, if a paralog candidate is also identified as a sequence paralog (SP), then it is assigned a value of 1 for this attribute and 0 otherwise

\[
AS = \left\{ \begin{array}{ll}
1 & \text{SP} \\
0 & \text{otherwise}
\end{array} \right. \]

For genes with mapped Affymetrix HG-U95 probe-sets, we used quantile expression patterns presented in GeneNote (Yanai et al., 2005). The similarity score is the mean Pearson correlation (P.Corr) between all gene quantile vectors for the query and candidate gene

\[
AS = \left\{ \begin{array}{ll}
P.Corr & \text{P.Corr} > 0 \\
0 & \text{otherwise}
\end{array} \right. \]

The attribute score is then multiplied by the weight given for the attribute and all attribute scores are then summed to give the Paralog Hunter score (PHS)

\[
PHS = \sum_{i=1}^{10} w_i \cdot AS_i.
\]

Set Distiller algorithm

Set Distiller, implemented in Java and available online (www.genecards.org/v3/genegecks), employs descriptors from 8 out of the possible 10 attributes that appear in Table 1, for user-defined query gene sets. For each descriptor, a \( p \)-value is calculated from the binomial distribution, testing the null hypothesis that the frequency of the descriptor in the query set is not significantly different from what is expected with a random sampling of genes, given the frequency of the descriptor in the set of all genes. Descriptors are sorted by increasing \( p \)-value and then by decreasing occurrence counts within the gene set. Bonferroni correction was used to correct for multiple testing and only descriptors with \( p \)-value >0.05 are displayed.

Generation of descriptors either randomly or by David

For every descriptor in the list generated by Set Distiller, random replacement descriptors were selected by matching attribute type and gene association count. The descriptor belonging to the same attribute and associated with a number of genes closest to the number associated with the descriptor being replaced.

The David API was polled, with an entire set of gene symbols and desired attributes, with their specific sources (Table 2, Supplementary), using the provided specifications (http://david.abcc.ncifcrf.gov/content.jsp?file=DAVID_API.html). Then, descriptors, attributes, and \( p \)-values were ex-
PubMed hit count and ID comparison

The Entrez Programming ESearch Utility was polled, by a PHP program (available on request), with 2 queries using the provided specifications (http://www.ncbi.nlm.nih.gov/entrez/query/static/esearch_help.html). The queries test the conjunction of either: (1) the query gene symbol and Paralog Hunter candidate symbol, or (2) the set name and a descriptor from the list generated either by Set Distiller, David, or random descriptors. For example, the terms “nucleolus” (set name) and “rRNA processing” (descriptor) were sent together to the online Entrez ESearch system as a Web query (using the required url format). Thereupon, the hits, corresponding to the number of publications found for the search terms were extracted from the resulting ESearch XML response. Identical set names and descriptors (such as for keyword-derived gene sets) were ignored. We normalized by word count when looking for conjunctions between set name and descriptor, because both may consist of more than one word, and searching in any database for multiple words is correlated with lower hits. The ESearch utility was employed for PubMed ID retrieval for keywords with fewer than 100,000 IDs (resulting in 36 of the 50 keyword terms used for this analysis) and genes with less than 2000 IDs associated with them.

Blast and FunSimMat comparison

The Blast algorithm was run for protein sequences against the RefSeq database and either a cutoff of E-score = 10 or the top 300 results were collected. E-score was plotted against the top 100 Paralog Hunter results. Similarly, when comparing to FunSimMat (Schlicker and Albrecht, 2008) (release 3.1, database built using: UniProKB 15.3, GOA release May 2009, Pfam 23, SMART from InterPro release 20, and OMIM), the 300 best GO classes found functionally similar to the query proteins were translated into gene symbols (via Uniprot IDs) and scores were plotted against the top 100 paralogs inferred by Paralog Hunter.

Descriptor lists were generated for 65 gene sets both by David and Set Distiller; the latter was subsequently used for random descriptor list creation. Mean values were calculated for the sets, but because they varied exponentially, we calculated the geometric mean over all set averages for Set Distiller, random descriptor, or David list hit counts. When comparing the results from Set Distiller, David, and random descriptors, we always limited the length of the descriptor list according to the shortest of the three parallel descriptor lists. The p-values reflect a t-test conducted between the geometric means of descriptor lists testing their statistical divergence. All analyses were conducted using Matlab.

Results

The GeneDecks system, established as a member of the GeneCards suite, strives to analyze relations among genes, including within gene sets, based on combinatorial annotation. Users can upload their data utilizing a new GeneDecks interface and choose either of 2 modes, Paralog Hunter or Set Distiller (Fig. 1).

The GeneDecks functional similarity metric

Paralog Hunter

In its Paralog Hunter mode, GeneDecks provides a numerical functional similarity score (Paralog Hunter score) for any pair of genes, computed as a sum of weighted normalized scores for 10 attributes (Table 3, view at the end of the article). For each attribute (e.g., domain or disorder)
normalization is effected by computing the fraction of the constituent descriptors (e.g., SH2, ZNF_C2H2; or hepatitis, asthma) shared by the two genes (see Methods).

The most widely used means of identifying potential relationships among genes in a given species is sequence-based paralogy. To assess the sequence-independent contribution of Paralog Hunter scores to similarity inference, we routinely set the weight of sequence paralogy attribute to 0 (the weights of all other attributes were set to 1) in the Paralog Hunter score calculation. Importantly, based only on the remaining nine attributes, GeneDecks considerably broadens the list of genes showing similarity to the query, with a clear trend for the sequence-based paralogs to appear high on the list of ranked Paralog Hunter scores (Fig. 2A). Paralog Hunter scores are particularly useful in the 24% of the cases where no sequence-based paralog is known. A specific case that highlights the utility of Paralog Hunter is the gene ATM, which has no sequence-based paralogs. Paralog Hunter identifies a functional paralog, ATR, known to act in parallel along a similar cellular pathway; this would be missed by checking sequence data alone.

It is suggested here that Paralog Hunter scores reflect the degree of functional paralogy between the listed genes and the query gene, independent of sequence similarity. It may be argued that although sequence similarity is routinely eliminated from the Paralog Hunter scores, many other annotations are performed in ways that indirectly bring in sequence similarity, as exemplified by protein domains and GO terms. We therefore attempted to evaluate Paralog Hunter using what may be considered as an independent gold standard for functional similarity between certain gene pairs. For this we used co-occurrence of query gene and paralog symbols as a pair of search terms in PubMed publications (Fig. 2B), which indicated a Spearman correlation $< -0.7$ between shared PubMed hit count and functional paralog rank. A majority (87%) of the functional paralogs that scored low in Paralog

---

**FIG. 2.** Paralog Hunter paralogs compared to sequence-based paralogs. Paralog Hunter score was computed using nine attributes, excluding sequence-based paralogy and the top 100 scoring paralog candidates for 50 highly annotated query genes were plotted by their rank (Paralog Hunter score). Query genes and their Paralog Hunter paralog appear in Table 9 (Supplementary). (A) Paralogs are colored by their Paralog Hunter score (bar on right) and columns are arranged by their best Paralog Hunter score. Genes that constitute sequence-based paralogs are marked with a black dot. In 24% of the genes tested, no sequence-based paralogs were found in the top 100 candidates, and in an additional 26% only one such gene was found. (B) Mean PubMed hit counts for the top 100 paralogs with identical rank was calculated for the conjunction of the query gene with each paralog found. A decreasing number of hits can be observed for paralogs with decreasing rank, suggesting decreasing relevance, as indicated by the trend line. (C) Blast E-scores ($-\log_{10}$) for the top 300 matches were plotted against Paralog Hunter scores for the top 100 paralogs found. A correlation (Pearson correlation = 0.25, $p = 3 \times 10^{-4}$) can be observed between the two scoring methods as also seen by the trend line. (D) FunSimMat scores for the top 300 functionally similar GO classes and their corresponding gene symbols were plotted against Paralog Hunter scores, showing a significant correspondence (Pearson correlation = 0.163, $p < 10^{-6}$).
Hunter (four points with lowest abscissa value in Fig. 2B) had a zero PubMed publications conjunction score, suggesting scant experimental evidence for gene pair association, even though most of the paralogs were highly annotated (>70%, Fig. 3 and Table 4, Supplementary). As a specific case in point, for the gene MAPK1, 50% of the top 100 functional paralogs had strong experimental support (PubMed score) for connecting them to their query gene while 14% had no publication association (Fig. 3). We also observed a correspondence, although weak, between pairwise Paralog Hunter scores and Blast E-scores (Pearson correlation = 0.24, p < 1 x 10^-6, Fig. 2C). Of the genes and their Paralog Hunter targets shown in Figure 2A, ~2.5% only were bona fide paralogs, and another ~13% were found to have protein sequence similarity by Blast (Fig. 2C). Interestingly, some of the functional paralogs identified by Paralog Hunter, for example, the genes ABCC6 and ABCB10 for the query gene ABCB1, did not show a significant Blast score, but were functionally interesting because they are related to the ATP-binding cassette. Finally, we compared the Paralog Hunter outcome to that of the FunSimMat program (Schlicker and Albrecht, 2008) that relates functionally similar proteins to a protein of interest via GO terms only, showing weak but significant correlation (Fig. 2D).

As seen in Figure 4A, the typical Paralog Hunter score is around 2 for highly annotated genes (specifically, those having a high GeneCards Inferred Functionality Score) (GIFtS, Harel et al., 2009; http://www.genecards.org/info.shtml# GIFtS, and Fig. 5), and 1 for genes with lower levels of annotation. However, as seen in Figure 6, the maximal scores for genes in the upper part of the GIFtS annotation scale is considerably higher. The average number of attributes that participate in the generation of their Paralog Hunter scores is 3.74 (Fig. 4B), suggesting a true combinatorial contribution to GeneDecks’ assessment of functional similarity. The most frequent contributors to the Paralog Hunter score are GO terms, expression patterns and disorders (Fig. 4C). The latter is significant, because it indicates that disease annotations are important contributors to connections among genes.

For most attributes, contributions to the Paralog Hunter scores is unweighted, that is, the descriptor ranks are equal to 1. In contrast, for the GO attribute, we have introduced a variable descriptor rank between 1 and 5 (see Methods), based on evidence code (Buza et al., 2008). Selecting this model may seem arbitrary; however, we in fact examined a series of descriptor rank models of the form r^z, where r is the base rank and z is an exponent between 0 and 10. We attempted to select the model that maximizes the correlation in analyses akin to that of Figure 2B. The results (Table 5, Supplementary) indicate that using no variable descriptor ranking (z = 0, hence, rank = 1 for all GO terms for all genes) is inferior to ranking (z > 0), with an optimum at z = 0.45.

Figure 7A shows the distribution of Paralog Hunter scores according to paralog rank for 50 randomly selected query genes (with high GIFtS), showing all ~10^4 target genes annotated by at least one attribute among the nine used for

**FIG. 3.** Paralogs are colored by the PubMed hit counts for the conjunction of the query gene (high GIFtS) and paralog candidate (bar on right) and columns are arranged by their best Paralog Hunter score (as in Fig. 2A and B). Paralogs with no publication linkage (scored 0) appear in white. *Indicates the MAPK1 query gene.
generating the Paralog Hunter score. A roughly exponential decay is seen as a function of score. There is a clear correlation between the best Paralog Hunter score (X-axis intercept) and the scores at much higher ranks (e.g., 1,000). This may indicate an inherent property of certain annotation-rich query genes to be functionally linked to numerous other genes. This is reflected in the partial tendency of genes with high GIFtS to be disposed on the right part of the distribution in Figure 7A, and also in the trend shown in Figure 7B for low GIFtS genes.

The GeneDecks gene-set tool: Set Distiller

Set Distiller is the GeneDecks mode that enables distilling a gene set’s essence by comparing the frequency of each descriptor in the set to its frequency in the entire gene population. To demonstrate Set Distiller’s capacities, we tested 65 gene sets belonging to two categories, microarray and keyword-based (Table 6, view at the end of the article). Fifteen
microarray-based gene sets originated from the ArrayExpress database (http://www.ebi.ac.uk/microarray-as/ae and appearing in Chandran et al., 2007; Greenawalt et al., 2007; Gustafsson et al., 2008; Haslett et al., 2002; Hummel et al., 2006; Lee et al., 2005; Saetre et al., 2007; Sood et al., 2006; Taylor et al., 2007; Woodruff et al., 2007) (Table 7, Supplementary); the other 50 were search results from GeneCards for various keywords that retrieve between 112 and 241 genes (Table 8, Supplementary). Even though the keyword-derived sets were of the same approximate size, the most shared descriptor varied greatly in its frequency (43–219) (Fig. 8A). Because descriptor frequency does not appear to be an appropriate yardstick, Set Distiller analyses sort the results by increasing p-value for the significance of enrichment. For objective assessment, we should ideally use a gene test set where the best terms for such a set of genes are known a priori. As an approximation of this scenario, and to obtain statistical power, we analyzed the entire collection of 65 gene sets, using PubMed search hits as a benchmark. For each of the up to 30 top descriptors in each Set Distiller list, the gene-set average number of hits was recorded for the publication conjunction of the set name (e.g., nucleolus) and each of the descriptors (e.g., rRNA processing). This process was repeated for a
matched list of random descriptors (Fig. 8B and C). A highly significant conjunction enhancement was seen for both the keyword-derived gene sets ($X_{6.3}, p = 9.8 \times 10^{-5}$) and for the microarray-derived gene sets ($X_{3.7}, p = 0.002$). Therefore, Set Distiller’s descriptors characterize the gene sets much better than random descriptors.

As a further sanity check, we reanalyzed 36 of our 50 gene sets generated by GeneCards searches (see Methods). For this we have added a new PubMed-based comparison tool, which evaluates the validity of the relationship between the set-generating words and the genes contained in the sets (Fig. 9). We compared the intersection of all PubMed IDs found for a set-generating keyword with the union of PubMed IDs related to the resulting gene set. Gene sets generated from the same keyword used for PubMed ID retrieval shared, on average, a greater fraction of IDs ($X_{10}, p = 9.24 \times 10^{-61}$) than a random gene set.

Subsequently, Set Distiller’s performance was compared to that of NCBI’s David (Fig. 8) (Dennis et al., 2003). For the keyword-derived sets (Fig. 8B) David scored significantly higher than random ($X_{4.9}, p = 1.44 \times 10^{-8}$), but it fared 1.3× worse than Set Distiller ($p = 0.033$). Moreover, for the microarray-derived gene sets David’s scores were 3.7× lower than Set Distiller ($p = 0.0063$), and were not significantly higher than the random lists score (Fig. 8C). It should be noted that in this comparison we even employed modifications of attribute selection to counteract certain inherent disadvantages of the David program. Gene specific nomenclature identifiers, such as “main accessions,” were excluded because they do not seem to be relevant when characterizing gene sets. We also omitted expression pattern descriptors for David because they were long (e.g., “mammary gland_invasive breast cancer ER+, PR+, Her2−, grade II, 3rd”); when included, they decreased David results to almost zero for the microarray gene sets. Normalizing the PubMed search hit results according to word count aided David the most, because its descriptors tend to be longer in general. Thus, of the six attributes compared, five were common to Set Distiller and David while one descriptor was unique to Set Distiller, and an additional one was unique to David. The protein interactions attribute is currently exclusive to David; this information appears in the GeneCards database and will be added to GeneDecks in the near future. David compounds appear under their pathways attribute, whereas Set Distiller’s are separate and pronounced.

The differences in results between the two gene-set analysis tools may be partly related to the differences in attribute propensity (Fig. 10); GeneCards has a much stronger representation of disorders and compounds, whereas David has a predominance of GO terms contributions. The attribute biases in the GeneCards results stems in part from a higher database-wide representation (Fig. 10B).

**Discussion**

The GeneDecks tool described here provides a means for analyzing the resemblance among genes. This includes a report of pairwise similarity among genes, as well as a capacity to find descriptor commonalities within gene sets. Both are based on the rich human gene annotation within GeneCards, which in turn, feeds upon nearly 80 world-wide digital resources, that summarize genome-wide knowledge. Both GeneDecks features, respectively embodied in the Paralog Hunter and Set Distiller modes, are important for systems biology endeavors. Although at present only 10 attributes, derived from 18 sources, have been fully implemented, the near future will see added attributes, among which our priority list includes protein interactions and publications.
We propose that annotation-based pairwise similarity may be interpreted as a generalization of gene paralogy. The commonly used means of paralog identification is based on a sequence similarity metric, like the one used in homology-based annotation transfer (Whisstock and Lesk, 2003). A key concept of Paralog Hunter is that one can obtain a more broadly disposed definition of gene similarity, coined herein “functional paralogy,” by seeking shared annotation descriptors stored within whole-genome databases, exemplified by GeneCards. In cases where no sequence paralogs were found, a functional paralog identified by Paralog Hunter and also literature based (Hurley and Bunz, 2007), such as ATR identified for the ATM gene, could prove essential.

In Paralog Hunter we made sure that sequence-based paralogy was weighted out, and when possible, such as for GO terms, we used descriptor ranks (Buza et al., 2008) to discriminate between sequence-based computer generated gene-descriptor associations and manually curated ones. It is reassuring that in 46% of the 50 cases tested, the first functional paralog was also a sequence-based paralog, and in 70% there is at least one sequence-based paralog within the top 20 Paralog Hunter scores (Fig. 2A). On the other hand, the remaining instances, where only one or no sequence-based paralogs were found in the top 20 scoring candidate genes suggest an advantage for Paralog Hunter over the other method. Obviously, in routine runs of Paralog Hunter the user can select whether or not to include sequence-based paralogy, serving as a minor but not insignificant factor among the 10 possible attributes used to identify functional paralogs.

Our GeneDecks/Set Distiller tool shows a high efficacy in analyzing gene sets, portraying the descriptors that best characterize the entire set. A major advantage of the described method is its reliance on GeneCards, one of the richest Web-wide resources of human gene-specific information. In the absence of a true gold standard, we performed a systematic comparison to David, a popular tool that operates along similar lines, and obtained indications that Set Distiller has a certain edge. This was done based on a specific criterion of keyword conjunction in published articles. We ascribe the differences between Set Distiller and David in part to unique attributes of the former, namely, mouse mutant phenotypes and compounds as explicit descriptors (David has a lower representation of compounds, and they are embedded within the pathways attribute). GeneDecks/Set Distiller’s advantage will necessarily depend on future incorporation of additional attributes, for example, expression data already implemented in Paralog Hunter, so as to remain useful in the analysis of sets of genes.

We used the PubMed search conjunction as a yardstick for the success of both Paralog Hunter and Set Distiller. The main advantage of this approach is its objectivity, because it does not rely on human judgment of gene relationships or descriptor relevance. Its disadvantages, on the other hand, are potentially related to a bias of the Medline database toward specific terms, for example, medically oriented, such as disorders. This might provide an advantage to a tool based on GeneCards, having a highly enriched disorder descriptor list, in comparison to other tools (such as David) with a lower disorder representation (Fig. 10A).

We are well aware of the fact that there is a certain level of circularity using PubMed as a benchmark. Although PubMed articles, as obtained directly from the publications section of GeneCards, do not contribute to any GeneDecks score, there are circuitous contributions. This is the case for Alma Knowledge Server (AKS, now renamed NovoSeek), which mines PubMed for connecting compounds or disorders to genes. However, this concern is alleviated to a great degree when considering the indirect nature of such relationships. Thus, in Paralog Hunter, AKS connects two genes via the conjunction of appearance in publications of each of these genes with a particular descriptor (e.g., a specific drug/compound or disorder), and the PubMed testing entails reporting the conjunction of the genes with each other, irrespective of the descriptor. Likewise, in Set Distiller, the indirect nature of PubMed contribution is seen as follows: a descriptor comes up for a gene set, because in PubMed testing, a connection is sought between the set name and a descriptor, whereas the actual Set Distiller signal is due to the conjunction of each member of the gene set to the descriptor. Moreover, the PubMed-related descriptors in both GeneDecks modes constitute only a relatively minor part (<25%) of the overall contribution to the score; this is true for only two out of nine attributes.

Independently of the previous evaluation of Paralog Hunter results by PubMed publications, a correspondence could be observed between Blast similarity for query and paralog proteins sequences and Paralog Hunter scores (Fig. 2C), strengthening the relevance of paralogs found by Gene-Decks. An additional test compared Paralog Hunter to FunSimMat, which operates similarly by matching functionally analogous proteins to a query protein (Schlicker and Albrecht, 2008) with the difference of being based solely on GO term linkage, and having a somewhat less straightforward user interface. The difference between the two methods observed could be ascribed to the use of several additional attributes in Paralog Hunter.

It is obvious that at present, although the process of deciphering the human genome sequences is still ongoing, the annotation depth of different genes is widely disparate. Our analyses use the GeneCards Inferred Functionality Score (GIFIS) for estimating annotation level. They show differences between genes with high and low levels of annotation. High GIFIS genes tend to have higher Paralog Hunter scores, and include a larger number of contributing attributes. Obviously, well annotated paralogs are more likely to make it to the top of the list for a given query gene, and as a result will be found to be functional paralogs using our method. In the future, it is expected that the significant population of very low annotation genes will become more highly annotated. Such future data accumulation will likely constitute a tradeoff between two trends: (1) increasing the number of novel genes by automated annotation, which will typically introduce GO or expression data and generate low GIFIS values; and (2) increasing the annotation level of more well-documented genes, often by manual curation, with a resulting enhancement of GIFIS values. Our results seem to indicate a preference for the second process for gene pairwise analyses, and prescribe a potential need for weighting algorithms to filter potential noise caused by poorly annotated genes.

The GeneDecks/Paralog Hunter algorithm may need to be improved in the future. The current scoring method is stringent, utilizing a binary yes/no outcome for exact descriptor matches. Some of the attributes, such as GO terms
and phenotypes, could be evaluated by graded distance, for example within GO or phenotype trees (Ashburner et al., 2000; Smith et al., 2005). For the phenotype descriptor tree, we currently use only the first level descriptors; future improvement will allow flexibility in this aspect as well. For other descriptors, such as compounds and disorders, tree or clustering methods could be used as well to measure descriptor distance.

One method for graded scoring, namely, evidence-based ranking in GO (Buza et al., 2008), has already been implemented in the current GeneDecks/Paralog Hunter version. We also introduced a user-controlled rank exponent to assess the optimal leverage that the rank should hold (0.1; Table 5, Supplementary) and set it as the default exponent. Similar ranking could also be employed in the future for compounds and disorders, for example, with the AKS scores present in GeneCards. Also, for pathways, one may use the network pairwise distance for the gene products distance metric.

Set Distiller currently has a built-in thresholding device in the form of p-value cutoff. Paralog Hunter, on the other hand, reports very long lists of pairwise similarity events (see Fig. 7). For many real-world applications a user-controlled threshold may be advisable for Paralog Hunter as well. We are considering implementing this, for example, by consulting a distribution of Paralog Hunter scores and reporting results that have a minimal Z-score computed from the mean and standard deviation of that distribution. In addition, it is possible to count the absolute number of shared descriptors, in parallel to the relevant fraction.

We have recently applied these concepts to a specific example with biological significance. This relates to the SYNLET project consortium (http://synlet.ibz.uni-leipzig.de/), which attempts to identify synthetically lethal partners for genes involved in tumorigenesis (Kaelin, 2005; Kennedy and D’Andrea, 2006) and in the resistance of tumors to drug therapy. The basic notion is that resistant tumors may harbor a mutation, whereby a second mutation or knockdown, which shows synthetic lethality to the first, will specifically kill the tumor. Paralog Hunter was used extensively to identify functional paralogs of tumor genes implicated in gene expression experiments, and some of the implicated synthetically lethal partner genes are currently tested experimentally.

GeneDecks performs, in essence, two types of mining operations on whole-genome annotation data. Such data may be envisioned as constituting a GeneDecks matrix in which rows represent genes and columns represent descriptors. Paralog Hunter seeks rows that are similar to a query row, whereas Set Distiller finds highly populated columns in a submatrix that has the rows representing a gene set. In future applications, GeneDecks may be envisioned as providing tools for the clustering of genes (rows) into groups with similar annotation content.

If applied genome-wide, our approach can result in a new type of GeneCards-based gene interaction network, akin to such obtained by other similar methods (Platzer et al., 2007). In parallel, one could ask about similarity of columns, allowing one to find relationships among descriptors based on gene content columns in the GeneDecks matrix. In addition, an outcome of future analysis could be the discovery of possible knowledge gaps, stemming from the paucity of experimental results or from database shortcomings. Thus, our method could supply experimental or curation targets.

Conclusions
The GeneDecks version 3 system enables sophisticated genome-wide comparisons. Its Paralog Hunter characterizes genes as functionally equivalent, based on their degree of weighted shared annotations. Its Set Distiller emphasizes qualitative characteristics that best elucidate sets of genes. The exploitation of the GeneCards database’s plethora of diverse annotations relieves GeneDecks of the reliance on sequence homology alone for finding potential paralogs, and broadens the descriptor range when scrutinizing a gene set without the need to rely solely on GO. GeneDecks has already been found to be useful in some research efforts, providing competitive or superior results to existing similar Web-based facilities.

Acknowledgments
We thank Dr. Yehuda Salzberg for his scientific insight and stimulating discussions. We acknowledge Shany Ron, whose initial work contributed to the GeneDecks project. Funding was provided by Xennex Inc., Boston, MA; the EU Specific Targeted Research Project consortium “Regulatory Control Networks Synthetic Lethality” (SYNLET, grant 043312); and the Weizmann Institute’s Gurwin Fund and Crown Human Genome Center.

Author Disclosure Statement
No competing financial interests exist.

Supplementary Data
Supplementary tables are at http://www.genecards.org/GenDecks_OMICS09

References