

The Renal Gene Ontology Annotation Initiative

Yasmin Alam-Faruque,^{1*} Emily C. Dimmer,¹ Rachael P. Huntley,¹ Claire O'Donovan,¹ Peter Scambler² and Rolf Apweiler¹

¹EMBL-European Bioinformatics Institute; Hinxton, Cambridge UK; ²Molecular Medicine Unit; UCL Institute of Child Health; London, UK

The Gene Ontology (GO) resource provides dynamic controlled vocabularies to aid in the description of the functional attributes and subcellular locations of gene products from all taxonomic groups (www.geneontology.org). A renal-focused curation initiative, funded by Kidney Research UK and supported by the GO Consortium, has started at the European Bioinformatics Institute and aims to provide a detailed GO resource for mammalian proteins implicated in renal development and function. This report outlines the aims of this initiative and explains how the renal community can become involved to help improve the availability, quality and quantity of GO terms and their association to specific proteins.

Introduction

Over the last decade the renal research community has embraced proteomic and genomic investigative methods to identify, quantify and characterize pathways and networks associated with the renal system.¹ For example, a number of recent proteomics analyses have identified several novel, potentially susceptible genes and proteins associated with various aspects of renal function, development and disease, whose role and mode of action within the renal system remain ambiguous.²⁻⁹ There is also a number of renal genome and proteome databases that exist, providing the scientific community with a range of central repositories of renal-related physiological data, published and unpublished mass spectrometry data and microarray data.¹⁰⁻¹³ Although these high-throughput resources are extremely powerful for investigating multi-factorial phenotypes such

as renal disease, these advances also mean that scientists must cope with the increasingly complex task of identifying, evaluating and managing the existing biological information for these large sequence sets. Hence, there is a need for effective bioinformatics tools as well as a supply of high-quality, detailed annotations that can support rapid evaluation of new experimental data and the generation of hypotheses for various biological insights.

The Renal GOA Initiative

The overall objective of the Renal Gene Ontology Annotation (GOA) Initiative is to provide a unique public resource of comprehensive functional annotations for proteins implicated in renal development, function and disease. The initiative aims to summarize the accumulated experimentally based knowledge for proteins using the popular structured Gene Ontology (GO) vocabulary by both improving the descriptiveness of terms describing renal processes as well as the number of associations of proteins involved in the renal system to information-rich GO terms. These efforts will ensure that the vast amount of published research on renal development and functional processes can be fully exploited by the renal research community to help guide future research towards alleviating renal disease.

The Gene Ontology and Annotation

Using structured controlled vocabulary terms, the GO project aims to fully describe three aspects of a gene product's attributes: the *molecular function(s)*, or activities that the sequence can directly

Key words: gene ontology, annotation, biocuration, kidney, renal

Abbreviations: GO, gene ontology; GOA, gene ontology annotation; DAG, directed acyclic graphs; GOC, gene ontology consortium; ID, identifier; UniProtKB, universal protein resource knowledgebase

Submitted: 10/23/09

Revised: 12/18/09

Accepted: 01/22/10

Previously published online:
www.landesbioscience.com/journals/organogenesis/article/11294

*Correspondence to: Yasmin Alam-Faruque;
Email: yalam@ebi.ac.uk

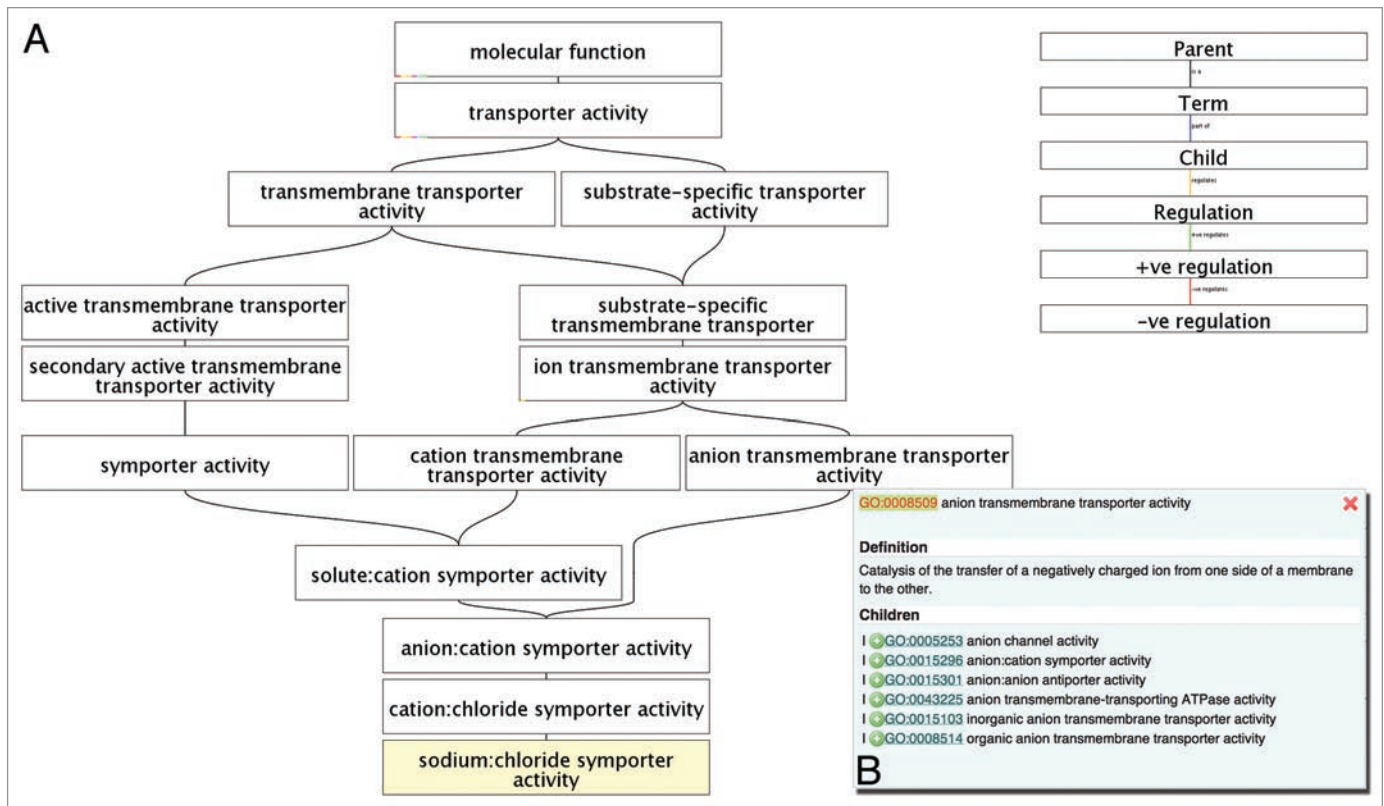


Figure 1. (A) A QuickGO view of a single Molecular Function GO term (i.e., GO:0015378 sodium:chloride symporter activity) showing its “parent” relations within the hierarchical DAG. (B) Clicking on a particular GO term box (i.e., Anion transmembrane transporter activity) brings up a window providing the GO term name, its stable GO ID, the definition of the term and its “child” terms.

perform; the *biological process(es)* it contributes to; and finally the subcellular locations (*cellular components*) in which it is located.^{14,15} For more informative and specific descriptions, new cross-product GO terms are created by combining existing GO terms with those from other ontologies such as the Cell Type ontology¹⁶ and the anatomy ontologies.¹⁷ Currently, over 29,800 GO terms exist, describing this wide range of concepts to differing levels of specificity.

GO terms are organized into Directed Acyclic Graphs (DAGs) which are hierarchical arrangements that allow a term to link to one or more general “parent” terms, as well as zero, one, or more specific “child” terms. For example, the GO term “alcohol catabolic process” (GO:0046164) has two parent terms, “catabolic process” (GO:0009056) and “alcohol metabolic process” (GO:0006066) and 14 child terms, including “phenol catabolic process” (GO:0019336) and “ethanolamine catabolic process” (GO:0046336). Each GO term has a unique, numerical, stable

identifier e.g., GO:0070634, a term name, e.g., “transepithelial ammonium transport,” and a definition (Fig. 1).¹⁸ A number of different, distinct relationships can exist between terms, which capture how a term relates to others in the ontology. The two standard relationships are “**is a**” and “**part of**.” All terms are linked by the “is a” relation which describe “subclasses” of concepts, e.g., the term “mitochondrion” (GO:0005739) *is a* “intracellular organelle” (GO:0043229), which in turn, *is a* “organelle” (GO:0043226). The “part of” relationship is used to represent part-whole relationships between terms, e.g., the “replication fork” (GO:0005657) is *part of* the “chromosome” (GO:0005694). Newer relationships are being added to the ontologies to further increase the descriptiveness, and the “**regulates**” (where one process directly affects the manifestation of another process or quality), “**negatively regulates**” and “**positively regulates**” relationships have been included in the GO since March 2008. This structure

of related terms being joined together by multiple, defined relationships provides users with the powerful ability to fully manipulate the ontology. It allows for expansion of an area of the GO to see the most detailed descriptions for specific functions, or to group together terms using specific relationships to gain an overview of the functions that different associated gene products share.

A wide range of model organism and cross-species database groups are involved in the GO Consortium (GOC), that apply automated prediction and manual curation methods to generate associations or “annotations” between specific GO terms and gene products.^{19–22} These GO annotations additionally include: (1) a reference to indicate the source of data used to support an annotation (a PubMed identifier or a publicly available description of an electronic annotation method) and (2) an evidence code; a three-letter acronym describing the type of investigation applied in the cited reference to the veracity of an annotation.^{23,24}

1	2	3	4	5	6	7	8	9	10		
ID	Symbol	Taxon	Qualifier	GO ID	GO Term name	Reference	Ev	With	A Date	From	
Process											
Q9H267	VPS33B	9606		GO:0006904	vesicle docking during exocytosis	interpro	IEA	IPR001619	P	20091001	UniProtKB
Q9H267	VPS33B	9606		GO:0016192	vesicle-mediated transport	interpro	IEA	IPR001619	P	20091001	UniProtKB
Q9H267	VPS33B	9606		GO:0006810	transport	spkw	IEA	KW-0813	P	20091001	UniProtKB
Q9H267	VPS33B	9606		GO:0015031	protein transport	spkw	IEA	KW-0653	P	20091001	UniProtKB
Q9H267	VPS33B	9606		GO:0070889	platelet alpha granule organization	16123220	IMP		P	20090902	UniProtKB
Q9H267	VPS33B	9606		GO:0015031	protein transport	15052268	IMP		P	20090609	UniProtKB
Q9H267	VPS33B	9606		GO:0006944	membrane fusion	15052268	IMP		P	20090609	UniProtKB
Q9H267	VPS33B	9606		GO:0032400	melanosome localization	15790593	IDA		P	20090827	UniProtKB
Q9H267	VPS33B	9606		GO:0032418	lysosome localization	15790593	IDA		P	20090827	UniProtKB
Q9H267	VPS33B	9606		GO:0016192	vesicle-mediated transport	15790593	IDA		P	20060414	MGI
Function											
Q9H267	VPS33B	9606		GO:0005515	protein binding	19109425	IPI	C14orf133 (H. sapiens)	F	20090826	UniProtKB
Component											
Q9H267	VPS33B	9606		GO:0005768	endosome	spkw	IEA	KW-0967	C	20091001	UniProtKB
Q9H267	VPS33B	9606		GO:0005764	lysosome	spkw	IEA	KW-0458	C	20091001	UniProtKB
Q9H267	VPS33B	9606		GO:0016020	membrane	spkw	IEA	KW-0472	C	20091001	UniProtKB
Q9H267	VPS33B	9606		GO:0031091	platelet alpha granule	16123220	IDA		C	20090827	UniProtKB
Q9H267	VPS33B	9606	colocalizes_with	GO:0030897	HOPS complex	19109425	IDA		C	20090826	UniProtKB
Q9H267	VPS33B	9606		GO:0005770	late endosome	15052268	IDA		C	20090609	UniProtKB
Q9H267	VPS33B	9606		GO:0005764	lysosome	15052268	IDA		C	20090609	UniProtKB
Q9H267	VPS33B	9606		GO:0005737	cytoplasm	15052268	IDA		C	20090609	UniProtKB
Q9H267	VPS33B	9606		GO:0048471	perinuclear region of cytoplasm	19109425	IDA		C	20090826	UniProtKB
ID	Symbol	Taxon	Qualifier	GO ID	GO Term name	Reference	Ev	With	A Date	From	

Figure 2. View of a combination of manual and electronic annotations for protein Q9H267, displayed by the QuickGO browser (<http://www.ebi.ac.uk/ego/GProtein?ac=Q9H267>). Annotations include information on (1) the sequence accession number from a named database (e.g., UniProtKB), (2) gene symbol, (3) species taxon ID, (5) GO term stable ID (6) GO term name, (7) the reference ID of the source of the manual or electronic annotation, (8) a three-letter acronym of the GO evidence code, (10) a date stamp of when the annotation was made/updated. The “Qualifier” column (4) is an optional addition to an annotation that acts to modify the interpretation of the associated GO term; similarly the ‘With’ column (9) provides further functional information such as the name/accession number of an interacting partner of the annotated protein (manual annotation) or the accession number from the underlying mapped source of the electronic annotation.

©2010 Landes Bioscience.

Electronic annotation prediction pipelines particularly benefit users of GO for non-model organisms or non-characterized sequences, since with conservative usage they can rapidly produce large numbers of annotations either from sequence data or by “translating” annotations made to external controlled vocabularies.^{19,24} The GOC currently uses a limited number of electronic pipelines; the most widely used applies protein signatures from the InterPro resource to predict functional attributes. For example, the protein signature “IPR000438: Acetyl-CoA carboxylase carboxyl transferase beta subunit” is used to identify a functionally similar protein set, all of whose members are also assigned the GO term “GO:0003989 Acetyl-CoA carboxylase activity.”²⁵ Another electronic method providing an improved level of consistency between manually annotated orthologs uses the Ensembl Compara orthology resource to transfer manual GO annotations between 1:1 orthologs in over 45 different species (http://www.ebi.ac.uk/GOA/compara_go_annotations.html).

Therefore, whilst electronic annotation can produce many millions of valid

annotations in a short space of time, these methods are limited. Automatic annotations are often only able to predict to high-level, less-detailed GO terms and rely on manual annotation activities in external groups to ensure correctness. Electronic annotation methods cannot capture the details of new, highly valuable experimental results that are found in peer-reviewed publications. Hence, manual annotation is employed, which requires highly trained curators to read and evaluate the available evidence in published literature in order to associate appropriate GO terms to proteins and to choose the most appropriate evidence code to apply to the annotation, thus resulting in a detailed summary of the knowledge about a protein (Fig. 2).¹⁸ Undoubtedly, manual annotation is a labor-intensive process; however, it does produce more annotations *per* protein, and uses GO terms which are far more informative and accurate than can be achieved by the current electronic pipelines.²² Manual methods also allow the curators to monitor electronic predictions of annotations to specific protein families and, when necessary, improve or correct them.²⁴

Using GO to Ascertain Biological Significance

To annotate the human genome comprehensively using GO is an arduous task and although several approaches are currently being used to achieve this,¹⁹⁻²² more manual annotation is essential. The biological insights provided by large scale genetic, genomic and proteomic studies can be difficult to ascertain and largely depend on computational analyses that incorporate functional annotation datasets. In certain cases, the current annotation datasets restrict the interpretation of these large-scale results, since the quality and quantity of the GO annotations is highly variable between different gene products.²⁶ The GO annotation dataset provided by the GOC is one of the most widely used resources in secondary biomedical data analysis, assisting researchers in interpreting, validating and forming hypotheses for their data. For example, one recent investigation by RamachandraRao et al.²⁷ has analysed GO annotations applied to protein-protein interactions to suggest that the antifibrotic effects of Pirfenidone may regulate

RNA processing and is renoprotective in diabetic kidney disease.

The Need for a Renal GOA Initiative

Currently, the number of GO terms describing kidney development or renal-related processes such as fluid volume regulation and detoxification is very limited. Therefore, the aim of the Renal GOA Initiative is not only to generate detailed manual GO annotation, but also to develop and improve the terms in the Gene Ontology to ensure that the whole of renal biology is well represented. We believe it will be of great benefit to the entire renal research community if this central information resource is improved, generating an annotation dataset which renal biologists can use with confidence. This resource could also be very useful for the many existing renal genome and proteome databases, whereby showing the relevant GO annotations for their renal-specific datasets could provide consistency and a useful link between each one, enabling visibility and promotion in other high profile databases which, at present, seems to be lacking.

How to participate in renal GOA. For the Renal GOA Initiative to have a large impact in the area of renal biology, it is important that experts from the renal community be consulted to ensure that the current accumulated knowledge has been comprehensively reviewed and correctly summarized by the dedicated curation team. Consequently, an international scientific advisory panel exists for consultation (<http://www.ebi.ac.uk/GOA/kidney/>) and a range of on-line facilities have been made available to encourage renal scientists to review and comment on the annotations or renal-related GO terms and to suggest publications or proteins for curation:

(1) Gene-specific web pages containing further information about the Renal GOA Initiative with links to GO annotations and newsletters can be viewed at: <http://www.ebi.ac.uk/GOA/kidney/> and <http://www.geneontology.org/GO.renal>.

(2) A renal interest group mailing list has been set up (<http://www.geneontology.org/GO.list.renal>), so that registered users

can be kept up-to-date on new developments and participate in discussions about the annotation of specific genes (accessible via the above websites). Interested researchers can contribute by either simply supplying the curators with details of key experimental publications which require curation, or by reviewing particular annotation sets. Information on gene products from any species is very welcome.

(3) For discussions about improvements and expansion to renal development-specific GO terms, a kidney development wiki page exists at: http://wiki.geneontology.org/index.php/Kidney_Development.

(4) A simple web form can also be used for feedback and is available at: <http://www.ebi.ac.uk/GOA/contactus.html>.

(5) Current GO annotations for the prioritized renal-related gene list can be viewed within the QuickGO browser by selecting the acronym KRUK (www.ebi.ac.uk/QuickGO/GAnnotation?protein=KRUK).

Although final annotation decisions are made by the professional curators, individual researchers contributing to the Renal GOA Initiative may do so purely to ensure their gene(s) of interest are well-curated or ensure that data from their own publications are annotated (which would hence be promoted in several highly visible databases). All contributions will be much appreciated by the GO curation teams and, when requested, a record will be maintained of those contributing so that their participation can be publicly acknowledged.

Current Activities of the Renal GOA Initiative

Collaborations have been initiated with a number of external and internal groups; work with the Genitourinary Development Molecular Anatomy Project team (GUDMAP—<http://www.gudmap.org/>) has begun to review the state of renal GO terms that currently exist in the ontology in relation to nephrogenesis, and has led to the creation of additional development terms in-line with the GUDMAP anatomy ontology (<http://www.gudmap.org/Resources/Ontologies.html>).

An association with the Reactome group at the EBI has led to the addition

of further members of the solute-carrier transmembrane transporter protein superfamily to the Reactome database (<http://www.reactome.org/>). These proteins, as well as others (ion channels, proton pump and aquaporins), can be selectively viewed using the keyword “kidney,” and hence the reaction pathways in which these gene products are involved can be analyzed.

Collaborations within UniProt and with other model organism databases, including FlyBase and AgBase, have begun to improve annotation of renal-related proteins for non-mammalian organisms, which will also ensure accurate description of excretory and osmoregulatory systems in these species. Such work should also highlight biological similarities and differences of the orthologous gene products in distinct species.

Acknowledgements

The Renal GOA Initiative is funded by the Kidney Research UK Project Grant RP26/2008. The GOA project is supported by the National Institutes of Health grant R01HG02273-02, the British Heart Foundation grant SP:07/007/23671 and EMBL.

We would like to thank the GO editorial team, Bijay Jassal from Reactome and the members of the Edinburgh team of the GUDMAP Consortium.

References

1. Janech MG, Raymond JR, Arthur JM. Proteomics in renal research. *Am J Physiol Renal Physiol* 2007; 292:501-12.
2. Nowik M, Lecca MR, Velic A, Rehrauer H, Brändli AW, Wagner CA. Genome-wide gene expression profiling reveals renal genes regulated during metabolic acidosis. *Physiol Genomics* 2008; 32:322-34.
3. Yasuda Y, Cohen CD, Henger A, Kretzler M. European Renal cDNA Bank (ERCB) Consortium. Gene expression profiling analysis in nephrology: towards molecular definition of renal disease. *Clin Exp Nephrol* 2006; 10:91-8.
4. Brunskill EW, Aronow BJ, Georgas K, Rumballe B, Valerius MT, Aronow J, et al. Atlas of gene expression in the developing kidney at microanatomic resolution. *Dev Cell* 2008; 15:781-91.
5. Siu KW, DeSouza LV, Scorilas A, Romaschin AD, Honey RJ, Stewart R, et al. Differential protein expressions in renal cell carcinoma: new biomarker discovery by mass spectrometry. *J Proteome Res* 2009; 8:3797-807.
6. Tilton RG, Haidacher SJ, Lejeune WS, Zhang X, Zhao Y, Kurosky A, et al. Diabetes-induced changes in the renal cortical proteome assessed with two-dimensional gel electrophoresis and mass spectrometry. *Proteomics* 2007; 7:1729-42.
7. Chabardes-Garonne D, Mejean A, Aude JC, Cheval L, Di Stefano A, Gaillard MC, et al. A panoramic view of gene expression in the human kidney. *Proc Natl Acad Sci* 2003; 100:13710-5.

8. Martinez G, Georgas K, Challen GA, Rumballe B, Davies MJ, Taylor D, et al. Definition and spatial annotation of the dynamic secretome during early kidney development. *Dev Dyn* 2006; 235:1709-19.
9. Jia L, Zhang L, Shao C, Song E, Sun W, Li M, Gao Y. An attempt to understand kidney's protein handling function by comparing plasma and urine proteomes. *PLoS One* 2009; 4:5146.
10. Gonzales PA, Pisitkun T, Hoffert JD, Tchapyjnikov D, Star RA, Kleta R, et al. Large-scale proteomics and phosphoproteomics of urinary exosomes. *J Am Soc Nephrol* 2009; 20:363-79.
11. Hoffert JD, Wang G, Pisitkun T, Shen RF, Knepper MA. An automated platform for analysis of phosphoproteomic datasets: application to kidney collecting duct phosphoproteins. *J Proteome Res* 2007; 6:3501-8.
12. Legato J, Knepper MA, Star RA, Mejia R. Database for renal collecting duct regulatory and transporter protein. *Physiol Genomics* 2003; 13:179-81.
13. Harris PJ, Buyya R, Chu X, Kobialka T, Kazmierczak E, Moss R, et al. The Virtual Kidney: an eScience interface and Grid portal. *Philos Transact A Math Phys Eng Sci* 2009; 367:2141-9.
14. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004; 32:161-258.
15. Blake JA, Harris MA. The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. *Current protocols in bioinformatics* 2008; 7:72.
16. Bard J, Rhee SY, Ashburner M. An ontology for cell types. *Genome Biol* 2005; 6:21.
17. Burger A, Davidson D, Baldock R. Anatomy ontologies for bioinformatics: Principles and practice, Springer 2008.
18. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 2009.
19. Smith CM, Finger JH, Hayamizu TF, McCright JJ, Eppig JT, Kadin JA, et al. The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res* 2007; 35:618-23.
20. Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol* 2009; 5:1000431.
21. The Gene Ontology Consortium, The Gene Ontology Project in 2008. *Nucleic Acids Res* 2008; 36:440-4.
22. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic Acids Res* 2009; 37:396-403.
23. <http://www.geneontology.org/GO.evidence.shtml> has further information on the evidence codes used by the GOC.
24. Dimmer E, Berardini TZ, Barrell D, Camon E. Methods for gene ontology annotation. *Methods Mol Biol* 2007; 406:495-520.
25. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009; 37:211-5.
26. Dimmer EC, Huntley RP, Barrell DG, Binns D, Draghici S, Camon EB, et al. The Gene Ontology—Providing a Functional Role in Proteomic Studies. *Proteomics* 2008.
27. RamachandraRao SP, Zhu Y, Ravasi T, McGowan TA, Toh I, Dunn SR, et al. Pirfenidone is renoprotective in diabetic kidney disease. *J Am Soc Nephrol* 2009; 20:1765-75.

©2010 Landes Bioscience.
Do not distribute.