Title Page

Category: Biological Sciences / Genetics

Sungear: Interactive visualization, exploration and functional analysis of genomic datasets.

Chris Poultney, Rodrigo A. Gutiérrez*, Manpreet Katari*, Miriam L. Gifford*, W. Bradford Paley†, Gloria M. Coruzzi* and Dennis E. Shasha§

Courant Institute of Mathematical Sciences, New York University.

* Department of Biology, New York University.

† Digital Image Design, Inc. 170 Claremont, Suite 6. New York, NY. 10027.

‡ Corresponding author:

Dennis E. Shasha. Courant Institute of Mathematical Sciences, New York University.

Address: 251 Mercer Street, New York New York 10012. Phone: 212-998-3086. Fax: 212-254-7947.

Email: shasha@cs.nyu.edu

Keywords [comparative, genomics, microarray, visualization, Venn diagrams]

Abstract

Many software tools are available to analyze genomic data and other large data sets, but no existing tool supports a rapid, visually interactive and biologist-driven exploration of natural questions on many experiments at a genomic scale. Sungear is a tool built for just this purpose. Here we describe its use in several case studies. Simple operations executed using Sungear enabled us to identify genome-wide responses that are robust across a series of microarray experiments, while linked GO annotation features enabled us to develop a biological hypothesis about the processes that respond in these data. When applied to compare genomes, Sungear enabled us to quickly uncover the patterns of conservation of gene function across the tree of life. In addition to its biological applications, Sungear can be applied to any area involving comparisons of multiple large data sets, as shown by a case study for the analysis of baseball statistics. In biology, the Sungear tool enables researchers without informatics skills to visualize and integrate information in multiple large data sets to generate biological hypotheses based on the rapid comparative and statistical analysis of genomic data through an intuitive interface. By enabling the formation and rapid testing of biological hypotheses, the Sungear tool aims to amplify the intelligence of biological researchers in the post-genomic era.

Introduction

The analysis of large data sets has become a routine task for many researchers in the post-genome era. Genome sequences and microarray hybridizations are a common source of such data. Often, researchers want to see how different data sets relate to one another. For example, in comparative genomics, one might ask which genes are unique to a single species, which functionalities tend to be shared among related species, and which among remote species. Similarly, when analyzing gene expression changes under different conditions, one might ask how many and what types of genes controlling specific biological functions are regulated (based on some common statistical criterion) by some subset of the input conditions. For example, researchers interested in exploring and improving nitrogen use in plants would like to know which functionally related gene sets are regulated by nitrate in leaves of light-treated plants, but not in leaves of dark-treated plants. We sought to develop a tool that would enable biologists to perform such Boolean analysis of genomic data in a rapid and visually driven manner to enable them to test specific biological hypotheses or to develop new biological hypotheses.

There are several tools currently available to analyze and visualize genomic data (for example see (1-4)). One of the most popular such methods combines clustering and heatmaps, and was first introduced to biologists by Eisen and colleagues (3). Clustering typically is used to identify groups of genes that share expression profiles. A heatmap is a two-dimensional grid in which each position in the grid is a colored box that represents the gene expression value in several experiments (3). Such a display conveys an intuitive impression of which set of genes are similar to each other in expression. Other tools allow users to visualize modified heatmaps within the context of metabolic pathways (e.g. (4)), and yet others visualize expression data as network graphs (e.g. (1)). However, the simple task of visually exploring and analyzing overlaps (intersections, unions

and more complex logical operations, such as genes expressed in treatments A & B but not in C) among several large data sets with supporting annotation information (e.g. gene names, and functional terms) is not supported or is impractical with existing tools. Moreover, none of the tools have a rapid way to determine the biological context or significance of these boolean combinations of gene lists.

The most common visualization method used to show overlap relationships of up to three different data sources is the Venn diagram, introduced by John Venn in 1880 (5). In proper use, the Venn diagram consists of a collection of either two or three overlapping circles which help to visualize the intersections among experiments. However, the limit of visualization for a Venn diagram is three lists of data. As shown in Fig. 1, the overlaps among even just four experiments cannot be represented properly or intuitively. As shown in Fig. 1, there is no way to capture the intersection of E2 and E4, for example, without including either E1 or E3 (Fig. 1).

We have designed Sungear (named and visually modeled after the automotive overdrive mechanism and taking some features from a text analysis tool called "TextArc" (6)) to support the analysis and visualization of an arbitrary number of datasets and Boolean combinations within a framework that provides supporting annotation data (e.g. biological GO annotations) and a statistical ranking of GO terms which enable the data to be mined in the context of biological functions. As shown in the ensuing sections, the lists of entities may come from experiments within one species (Fig. 2), they may come from several species (Fig. 3), or even from domains as far flung as sporting statistics (Supplemental data). In addition to logical operations on lists, the GO annotation function of Sungear facilitates the interactive analysis and exploration of the functional attributes of the lists.

The Sungear tool empowers and enables biologists and other scientists not trained in informatics to mine and explore large data sets in an interactive way for the analysis and generation of biological hypotheses.

Results and Discussion

The Sungear interface

The Sungear interface presents four windows to the user: 1. the Sungear plot, 2. the Gene list, 3. the GO terms, and 4. the navigation/export controls (Fig. 2). These four windows are linked with one another, so that selections in one window will immediately be reflected in the other windows. Fig. 2A shows the use of Sungear to represent regulated genes from Arabidopsis microarray experiments carried out in three different laboratories (described in more detail below). The experiment names are listed around the polygonal vertices (hereinafter called “anchors”), and these names are linked to the list of genes in that experiment (e.g. regulated genes as discussed below). A “gear” or "vessel" is a circle within the polygon with arrows pointing to one or more anchors. The size of a vessel, for example pointing to anchors A1 and A2, is proportional to the number of genes in experiments A1 and A2 but not other anchors. The location of the vessel within the polygon is largely determined by the position of all anchors with which it is associated. The number of vessels that could be present corresponds to all possible subsets of the experiments, including the null subset. So, if there are X experiments (represented by an X-gon) there can potentially be 2^X vessels. However, the actual number of vessels that will be visualized depends on the dataset. For example, there may be fewer than 2^X vessels, because some possible intersections may contain no genes. Whereas the Venn diagram representation doesn't extend beyond X = 3, Sungear can represent an arbitrary number of experiments/lists, depending only on the researcher's willingness to understand a visual display having many anchors and associated vessels.

The Sungear display offers visual and quantitative information about the numbers of genes that are, for example, similarly regulated in any combination of the experiments analyzed. Thus, a cursory look at the Sungear window can quickly answer a common question posed by biologists when analyzing expression data. In addition, Sungear provides other views of the vessel contents, including gene lists (left side), a Gene Ontology (GO) (6) hierarchy (upper right), and a list of over-represented GO terms (lower right) (Fig. 2A). In Sungear, one or more vessels, genes, anchors and/or GO terms may be selected for analysis at any time. Fig. 2B shows the result of selecting one vessel that has three arrows pointing to anchors on the right-hand side of the hexagon. The selected vessel contains a collection of 65 genes called "group 1" (see top right of the anchors) (Controls window, Fig. 2B). The GO terms have been reordered in descending order of their degree of over-representation, as measured by their z-scores (lower panel in GO Terms window, Fig. 2B). A large positive z-score corresponds to a GO term that is likely to be statistically over-represented, even after correcting for multi-testing (7). We use the qualifying term "likely" because the best test for over-representation involves the hypergeometric distribution and a hierarchy-specific correction for multi-testing, a very slow computation. For the sake of speed, therefore, Sungear computes z-scores and suggests the rule of thumb that a z-score of 10 or more will survive the more rigorous tests. Thus, one can identify the functionality of genes that (in this case) are regulated by a specified set of experiments, answering another natural question posed by biologists.

Because of the associated GO annotations, Sungear can also be used to query for the regulation of specific processes. For example, Fig. 2C illustrates what happens when GO terms representing amino acid transporter activity are selected for analysis. The Sungear hexagon shows vessels whose outlines correspond to their pre-selection sizes, but that have only a smaller highlighted ball in their

interiors. The empty annulus represents genes that do not meet the criteria of the selection -- in this case, do not relate directly or indirectly to the selected GO term.

When querying, it is useful to distinguish Boolean "and-functionality" from "or-functionality". And-functionality arises, for example, when a researcher wants to select all those genes that are in vessel <X> and also satisfy GO term <Y>. By contrast, or-functionality would be used when, for example, seeking genes that are either in vessel <X> or satisfy GO term <Y>. Using a familiar combination of mouse clicks, shift and alt keys, Sungear can readily support these queries as well as many other logical operations including exclusion (i.e. experiment A and B, but not experiment C).

Sungear analysis case study 1: analyzing the robustness of genome-wide expression data

Microarray technology has provided us with the ability to measure genome-wide regulation of gene expression. Currently, several thousand microarray hybridizations are publicly available for several model organisms (8). The challenge ahead is to derive robust biological insights and to derive biological hypotheses from this vast amount of data. To illustrate an example of this application of Sungear, we analyzed published microarray studies that identified Arabidopsis genes regulated by transient treatments with the nutrients nitrogen (N) and or carbon (C) (9-11). In this example, six lists of genes containing N- or CN-regulated genes (I= induced; D= depressed) provide the anchors for Sungear (Fig. 2A-C). These experiments conducted by three different research groups all share the feature of transiently treating *Arabidopsis* seedlings with nitrogen or nitrogen plus carbon nutrients, and assaying gene responses using the ATH1 Affymetrix whole genome chips. However, these experiments vary with factors such as age of plant, growth, pre-treatment and treatment conditions, and also in the statistical analysis used to determine the regulated genes.

We used Sungear to determine whether a comparison of the gene lists in the shared vessels could provide any biological insights from the combined analysis of these three datasets, which could not be deduced from the study of any single dataset (Fig 2A). Sungear provided a rapid way to identify the vessels containing genes that are most biased towards a specific biological functionality, implemented in the "Find Cool" button of the navigation control window. "Find cool" ranks the vessels based on the number of functional GO terms that exceed a z-score of 10, suggestive of statistical over-representation of genes in a biological function. Using the "find cool" feature of Sungear, we identified a potentially interesting vessel containing 65 regulated genes. This vessel contains a set of statistically significant biologically related genes that are consistently induced by nitrogen treatments in all three studies (pink-colored vessel in Fig. 2C). Sungear revealed that the GO terms with the highest z-scores in this vessel were related to nitrate assimilation: 'nitrate assimilation' (z-score=35), nitrate reductase activity (z-score=28), 'ferredoxin-nitrate reductase activity' (z-score=20) (Fig. 2B); See Supplemental Table 1A for the top 10 z-score ranked GO terms for each vessel discussed. This Sungear function revealed that the genes that are consistently N-responsive across a wide range of treatment conditions, are genes whose encoded proteins are involved in very early steps of nitrate assimilation. By contrast, Sungear analysis revealed that N-regulated genes unique to each individual dataset (vessels in the periphery of the hexagon, Fig. 2A) encode proteins that are involved in processes downstream of primary nitrogen assimilation. Among the top z-scoring GO terms (with more than 1 gene) in each of these cases we found 'structural constituent of ribosome' (z-score=28, containing 81 genes), and 'catalytic activity' (z-score=10, containing 430 genes); See Supplemental Table 1A. In addition, we found genes involved in processes including: 'response to water', 'autophagy', 'defense response' and 'ethylene signaling'. These results suggest that the pathway involved in directly

metabolizing or perceiving the input signal (in this case nitrate), has a more robust and reproducible response to the nitrogen signal, while the nitrogen-regulated response of genes in downstream pathways or processes is heavily modulated by other environmental variables. This would partly explain the overall lack of reproducibility in microarray experiments carried out in different groups. This comparative analysis of these datasets enabled us to readily generate a biological hypothesis for nitrogen regulation of biological functions that are upstream *versus* downstream of the primary nitrogen assimilation pathway (Fig. 2D).

Sungear analysis case study 2: comparative genomics

As discussed above, Sungear is inherently well suited for handling several datasets each consisting of large amounts of data. In order to highlight this feature, we give an example in which Sungear is used to investigate the degree to which particular biological processes are shared across the tree of life. The Sungear plot in Fig. 3 shows the results of using BLASTP (E-value cutoff $\leq 10E^{-10}$) to compare genes encoding Arabidopsis proteins (hereinafter proteins) to genes encoding all the proteins in *C. elegans* (worm), *D. melanogaster* (fly), *H. sapiens* (human), *M. musculus* (mouse), *R. norvegicus* (rat), *S. cerevisiae* (yeast), *S. pombe* (fission yeast), and a collection of microbes including cyanobacteria, Archaea and bacterial proteomes, as described previously (12). In this analysis, the Arabidopsis proteins are the "background" used for annotation (Gene names and GO terms).

A cursory look at the data shows that the largest vessel in the Sungear window contains 13,562 proteins, and these correspond to Arabidopsis proteins that are not shared with any of the other organisms in this analysis. This vessel is named "Arabidopsis specific" and because it lacks intersection with any other anchor, it is located outside of the polygon (Fig. 3A). Among the Arabidopsis-specific proteins, we find a high over-representation of "unknown" proteins for which

the molecular function (z-score=46), biological process (z-score=41), and cellular component (z-score=17) are not annotated (see Supplemental Table 1B). In addition, this Arabidopsis-specific vessel also contains an over-representation of proteins annotated to 'transcription factor activity' (z-score=14). This is consistent with the idea that divergence of factors that control gene expression is an important driver in plant evolution (13) and also reflects the fact that transcription factor families have undergone a large expansion in plants (14). The second largest vessel contains 2,228 proteins shared between all species in this analysis. Interestingly, this vessel was also the 'coolest' vessel based on our "find cool" function (described above). Within this vessel, the top three z-scoring GO terms are 'kinase activity', 'catalytic activity' and 'ATPase activity' (z-scores= 57, 48 and 43 respectively. See Supplemental Table 1B). This supports the hypothesis that phosphate-transfer is a fundamental process common to all forms of life.

To discover which processes are shared at basal levels of the evolutionary tree, we searched for proteins shared between Arabidopsis, all the other eukaryotes and Archaea, but not shared with bacteria and cyanobacteria. This analysis yielded a vessel containing 312 proteins, amongst which the core elements of protein synthesis are over-represented. The top three z-scoring GO terms are 'structural constituent of ribosome', 'ribosome' and 'small ribosomal subunit'; these terms have z-scores of 56, 53 and 39 respectively. Within this vessel there are many 60S and 40S ribosomal subunit proteins. This type of ribosome is distinct to eukaryotes, and the fact that this analysis suggests that such proteins are shared between eukaryotes and Archaea, is consistent with the hypothesis that Archaea are closer to eukaryotes than bacteria (20).

We next moved up the tree to analyze proteins common to all multicellular eukaryotes. For this, we used Sungear to select the vessel with 560 Arabidopsis proteins that are shared with human, mouse, rat, fly and worm. The top three z-scoring GO terms in this vessel are

‘sulfotransferase activity’ (z-score=27), ‘cyclic nucleotide binding’ (z-score=25), and ‘ion channel activity’ (z-score=21). Little is known about plant sulfotransferases (SOTs), although they are structurally similar to mammalian SOTs, and catalyze reactions involved in cell-cell signaling (21). Cyclic nucleotides are found in animal and plant cells and play key roles in cell-cell signaling (22). Furthermore, this vessel includes ion channel activity: a process also inherently associated with having more than one cell.

In the next step of this case study, we considered Arabidopsis proteins shared with either fungi or animals. We first analyzed the 325 proteins that are shared between Arabidopsis and either of the yeast species, but not shared with animals (Fig. 3B). The top three shared processes are 1,3-beta-glucan synthase activity (z-score=28), 1,3-beta-glucan synthase complex (z-score=28) and beta-1,3 glucan biosynthesis (z-score=28). 1,3-beta-D-glucan synthase is a multi-enzyme complex that catalyzes the synthesis of 1,3-beta-linked glucan, a component of both the yeast and plant cell wall. We then selected a group including 213 proteins that are shared between either of the invertebrate species (fly and worm) and Arabidopsis, but that are not found in vertebrates. The top three GO terms are response to pathogen (z-score=16), amino acid permease activity (z-score=15), and amino acid transport (z-score=15). These processes have to do with stress responses, signal perception and cell-cell transport. Finally, comparing Arabidopsis to vertebrates, yields a group of 645 proteins shared exclusively between Arabidopsis and any of the vertebrates in this study (human, mouse and rat). The top three z-scoring GO terms are zinc ion binding (z-score=16), transferase activity, transferring glycosyl groups (z-score=11) and protein binding (z-score=10). These functions are involved in protein modification and protein interactions. Supplemental Table 1B contains the complete list of GO-terms and their z-scores for the groups discussed above.

The use of Sungear to perform the above Boolean analysis of genome contents, enabled us to identify biological processes that have evolved across the tree of life. The observation that cell-cell signaling functions are linked to multicellularity is proof that Sungear can highlight known biological principles. Our finding that protein modification is a process shared between plants and vertebrates, and not in microbes suggests protein modification is a relatively derived trait. This analysis provided the hypotheses that can be used to guide future research.

Sungear case study 3: baseball scores

The flexible nature of Sungear is illustrated by the fact it can be used to visualize data coming from fields other than biology or even science, for example, baseball. As explained in Methods, changing the application domain of Sungear for this application is easy. In the baseball example, "gene lists" are replaced by "baseball players", and the "GO hierarchy" is replaced by a "league-team hierarchy". For our example, we used publicly available baseball team information and player statistics (23). In this example, each anchor corresponds to one of four player performance measures during the 2004 season: (1) batting average of .250 or better (avg >= .250), (2) 20 or more home runs (HR >= 20), (3) 50 or more runs batted in (RBI >= 50), and (4) 10 or more stolen bases during the season (SB >= 10). All players active for some part of the 2004 season are shown, but not all of them meet our criteria above: the largest vessel, located in the upper left outside the polygon, represents the set of players meeting none of the criteria (pink-colored vessel in Suppl. Fig. 1A). We deem players to be "remarkable" if they meet all four criteria. Restricting our analysis to remarkable players, we can quickly see that the American League is over-represented, with nearly twice as many remarkable players as the National League (Suppl. Fig. 1B). A quick intersection between the two leagues shows that one of these players, Carlos Beltran, was active in both leagues during the season (Suppl. Fig. 1C). By viewing and exploring the GO term window

one can find that his respective teams for the American and National leagues were the Kansas City Royals and the Houston Astros.

Conclusion

Sungear is a tool that generalizes Venn diagrams to view multiple collections of genes, relates those collections to functional categories, and permits visual real-time, statistically-based data exploration. After minutes of training, users without any computer skills can learn to be comfortable navigating Sungear to explore and compare datasets. For the moderately sophisticated user, Sungear permits various data selection capabilities including "and-functionality", "or-functionality", and range selection. Sungear also provides support for the discovery of over-representation using any directed acyclic graph, such as the gene ontology. Sungear can be easily integrated with other visualization tools such as Cytoscape (1), used for modeling gene networks, through an overarching framework (Gutiérrez et al. in preparation). In an attempt to make the Sungear tool easily accessible to biologists working on different species, we have created supporting Annotation and Gene Ontology files for the major model plant, animal and microbe species under study. This enables users to upload their own data lists and also to create Sungear plots for their species of interest. More sophisticated users can upload other datasets for exploration from other genomes, from proteomes, or from non-biological disciplines.

Sungear enables logical operations such as intersections, unions, and negations, provides supporting data and organizes data based on statistical criteria. This combination of features facilitates rapid, visually intuitive, and statistically driven exploration of data sets for functionality that can suggest new avenues for deeper analysis and experimentation. By enabling the formation and rapid testing of biological hypotheses in the post-genomic era, Sungear amplifies the intelligence of biological researchers.

Methods

Sungear software implementation

Sungear is implemented in Java using Sun's J2SE v1.4.2. It can be run in a web browser as an applet, or as a stand-alone application. The default Sungear interface presents four windows: the gene list, the Sungear plot, the GO terms, and the navigation/export controls. These windows are linked together, so that selections in one window will immediately be reflected in the other windows. This linking behavior is achieved by making genes the "common currency" of Sungear: all windows provide a representation of a single master copy of the currently selected set of genes. A selection operation in one window (e.g., clicking on a GO term in the GO term hierarchy or list) will change the master copy; the other windows will then be sent messages indicating that the selected set has changed, and each window will update itself based on the new list. All other Sungear operations, such as loading a new set of experiments or narrowing down to a smaller set of genes within an experiment set, use the same principle of gene lists and messaging. Since rapid response to messages is critical to Sungear's real-time user interaction, each window also performs some pre-processing of data when a new experiment set is loaded to ensure quick data handling.

Sungear is designed to be able to incorporate new species, new data types, and new display components. Preparing Sungear to display a new species requires a one-time process to create two input files Sungear relies upon to describe each species. One of these files gives the gene IDs and annotations, and the other gives the associations between genes and GO terms, including the distributional information used to calculate z-scores (see below). Once these files have been created and placed in appropriate directories, any experiment set using those gene IDs can be viewed.

Sungear is not limited to exploring data consisting of gene sets. Rather, sets of genes are just a paradigmatic example of the much larger universe of data sets consisting of items and categories (corresponding to genes and GO terms, respectively). As mentioned, Sungear can be used to display baseball statistics, where the "gene lists" are players and the hierarchical "GO categories" are teams and leagues. The "experiments" are then sets of players that meet some criteria (for example, thresholds for the number of home runs or bases stolen).

In the Polygonal display, the gene collection names are placed in the order of columns of the input file. The position of the center of a vessel that points to some set of anchors v1, ..., vk is roughly the mean of the x-positions of the anchors and the mean of the y-positions of the anchors. Since the anchors are placed so that they form a convex polygon, the mean of any non-null subset of the anchors must lie within the polygon. In practice, the anchors used for vessel placement are moved slightly inward (toward the center of the polygon) from the displayed anchor positions, guaranteeing that not just the vessel center but the entire vessel will reside within the polygon. The only exception is the vessel that represents the null subset, which if it exists is placed at a standard location outside the polygon. Vessel positions are then relaxed using a spring algorithm that attempts to keep vessels close to their desired positions while preventing vessel overlaps.

Setting up and running Sungear

Written in Java and usually run as an applet, each Sungear installation allows a set of users to upload data files representing multiple experiments, or, in general, datasets of any kind, in a simple format (see Methods). Once the data is loaded into an executing instance of Sungear, users interact with Sungear over the Web by selecting one or more collections, one or more GO terms, one or more intersections among collections, or any combination of the above. Selections reorder

GO terms in descending order of their over-representation, often suggesting insights that are not obvious to the designer of a single experiment. The interface normally responds within seconds thus allowing very fast data exploration.

Data analysis

The bottom right-hand corner of the Sungear display has GO terms ranked in descending order by z-score. The goal of the ranking is to give an approximate indication of which GO terms are likely to be significantly over-represented after correcting for multi-testing. The procedure for calculating the z-scores is the following:

1. Offline (i.e. just once every time the master list of GO terms is changed), find the number of genes associated directly or indirectly with each go term t:

$$p\_t = \frac{number\ of\ genes\ associated\ with\ t}{total\ number\ of\ genes\ in\ genome}$$

$$std(t) = \sqrt{p\_t \times (1 - p\_t)}$$

2. When viewing the data after some selection has taken place, there are a smaller number N genes remaining.

$$Let\ f\_t = \frac{number\ associated\ with\ t}{N}$$

3. Finally, the z-score :

$$z\_score(t) = \frac{(f\_t - p\_t) \times \sqrt{N}}{std(t)}$$

Acknowledgments

This work was funded by grants from the National Science Foundation (IIS-9988345, IIS-0414763, DBI-0445666) and (0115586) to D.E.S.; grants from the National Science Foundation – N2010 (IBN0115586) and (DBI-0445666) to G.M.C.; grant from the National Science Foundation (DBI-0445666) to R.A.G.; EMBO postdoctoral fellowship ALTF107-2005 to M.L.G.

References

1. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003) *Genome Res* **13,** 2498-2504.
2. Breitkreutz, B.-J., Stark, C. & Tyers, M. (2003) *Genome Biology* **4,** r22.21-r22.24.
3. Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998) *Proc Natl Acad Sci USA* **95,** 14863 - 14868.
4. Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L. A., Rhee, S. Y. & Stitt, M. (2004) *Plant J* **37,** 914-939.
5. Venn, J. (1880) *Philosophical Magazine and Journal of Science* **July**.
6. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000) *Nature Genetics* **25,** 25-29.
7. Dudoit, S., Van Der Laan, M. J. & Pollard, K. S. (2004) *Statistical Applications in Genetics and Molecular Biology* **13,** Article 13.
8. Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G. G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S. & Brazma, A. (2005) *Nucleic Acids Res* **33,** D553-555.
9. Price, J., Laxmi, A., St. Martin, S. K. & Jang, J.-C. (2004) *Plant Cell* **16,** 2128-2150.

10. Scheible, W.-R., Morcuende, R., Czechowski, T., Fritz, C., Osuna, D., Palacios-Rojas, N., Schindelasch, D., Thimm, O., Udvardi, M. K. & Stitt, M. (2004) *Plant Physiol.* **136,** 2483-2499.

11. Wang, R., Tischner, R., Gutierrez, R. A., Hoffman, M., Xing, X., Chen, M., Coruzzi, G. & Crawford, N. M. (2004) *Plant Physiol* **136,** 2512-2522.

12. Gutierrez, R. A., Green, P. J., Keegstra, K. & Ohlrogge, J. B. (2004) *Genome Biol* **5,** R53.

13. Doebley, J. & Lukens, L. (1998) *Plant Cell* **10,** 1075-1082.

14. Shiu, S. H., Shih, M. C. & Li, W. H. (2005) *Plant Physiol* **139,** 18-26.

15. Richmond, T. (2000) *Genome Biol* **1,** REVIEWS3001.

16. Von Schweinichen, C. & Buttner, M. (2005) *Plant Biol (Stuttg)* **7,** 469-475.

17. Heyer, A. G., Raap, M., Schroeer, B., Marty, B. & Willmitzer, L. (2004) *Plant J* **39,** 161-169.

18. Mitsuhashi, W., Sasaki, S., Kanazawa, A., Yang, Y. Y., Kamiya, Y. & Toyomasu, T. (2004) *Biosci Biotechnol Biochem* **68,** 602-608.

19. Marshall, S. D., Putterill, J. J., Plummer, K. M. & Newcomb, R. D. (2003) *J Mol Evol* **57,** 487-500.

20. Brown, J. R. & Doolittle, W. F. (1997) *Microbiol Mol Biol.Rev.* **61,** 456-502.

21. Klein, M. & Papenbrock, J. (2004) *J Exp Bot* **55,** 1809-1820.

22. Bridges, D., Fraser, M. E. & Moorhead, G. B. (2005) *BMC Bioinformatics* **6,** 6.

23. The baseball archive (2005) http://www.baseball1.com/statistics/

Figure Legends

Figure 1. (A) The Venn diagram, a common way to compare the overlap between lists of data. **(B)** Beyond three lists of entities the Venn diagram can no longer visualize all possible intersections e.g. the intersect of E2 and E4).

Figure 2. Use of Sungear to analyze microarray experiments. **(A)** Sungear comparison of genes called to be N/CN-regulated from three published datasets (9-11). Scheible(I) – genes induced (I) by criteria described in Scheible *et al*. 2004; Scheible(D) – as previous for depressed (D) genes. Similarly for Wang(I), Wang(D), Price(I) and Price(D). **(B)** Use of the Sungear 'Find Cool' button yields a vessel containing induced genes shared between the Price(I), Scheible(I) and Wang(I) datasets (pink-shaded vessel). This vessel contains 65 genes and includes a statistically significant number that are associated with nitrate assimilation and nitrate reductase activity (see GO term window sorted by z-scores). These comprise nitrate reductase and glutamate synthase. **(C)** Sungear showing selection of a specific GO term, amino acid transporter activity. Internal vessel structure alters after this selection. **(D)** Overview of what has been found by using Sungear; z-scores for each process and number of genes are in parentheses (z-score, No genes). The nitrogen assimilation/uptake pathway is consistently found to be N-regulated across datasets from multiple labs. The N-regulated processes which are unique to single-lab datasets are in contrast other cellular functions operating further downstream of the incorporation of N into amino acids (Gln, glutamine; Glu, glutamate). The only significant process according to z-score which is associated with more than one gene is 'structural constituent of ribosome' (see Supplemental Table 1A). This process was found to be significant in the Scheible et al Induced dataset.

Figure 3. Visual representation of data from a whole-genome comparison with Sungear. **(A)** Species anchors contain proteins shared between human, mouse, rat, fly, worm, fission yeast, yeast,

archaea, bacteria, or cyanobacteria with Arabidopsis. **(B)** A comparison of the number of Arabidopsis proteins shared with yeast or fission yeast, but no other species. Together the three vessels comprise a group of 325 proteins. The top three z-scoring processes within this group are 1,3-beta-glucan synthase activity, 1,3-beta-glucan synthase complex and beta-1,3 glucan biosynthesis. **(C)** Schematic showing the relative positions in the tree of life of the species being investigated in this comparative genomics analysis (species within taxa are shown in bold). Diagram adapted from the Tree of Life Web Project (Maddison, D.R. and Schulz, K.-S. (ed.) 2004; URL: http://tolweb.org). Dotted lines denote intermediary tree-structure between the taxa being connected (to simplify the figure, only portions of the tree are shown). Boxes around the tree show the top three most over-represented biological processes that have found to be shared between selected species. The number of proteins within each group investigated are shown above the GO terms, and z-scores for each GO term and the numbers of proteins associated to that GO term are in parentheses: (No proteins, z-score). Not shown are the top three over-represented processes in the Arabidopsis-alone vessel that contains 13,562 proteins. These are all unknown processes/functions/components (see main text). The next three most highly z-scoring terms are Transcription Factor Activity (1253,14), Transcription Regulator Activity (1314,13) and Regulation of Transcription (874,11).

Table 1.
Input format for Sungear.
Input files for Sungear consist of rows and column formatted as shown below. The first row contains the column names which will be used for the anchor labels. The last column contain the gene names. Each value in the matrix indicate whether a given gene is present in a set or not, with "1" = present and "0" = absent. Each entry in this matrix is separated by a <space>|<space> delimiter. Please note that the final data matrix must be sorted according to set membership prior to opening with Sungear (genes in each vessel must be contiguous in the file).

C	N	L	O	gene
0	0	0	1	At3g50260
0	0	1	0	At3g61890
0	0	1	1	At4g17695
0	0	1	1	At1g26800
0	1	1	0	At1g77920
1	0	1	0	At2g20030

Figure 1

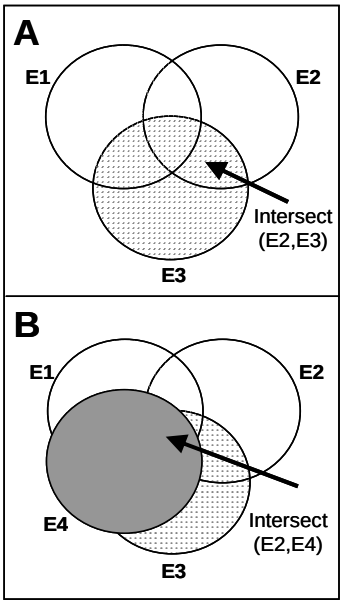
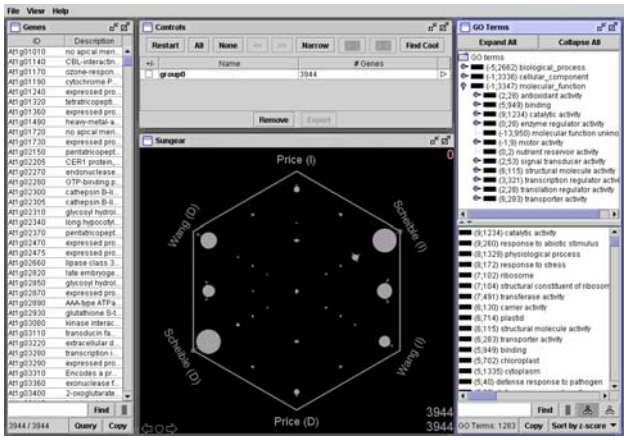
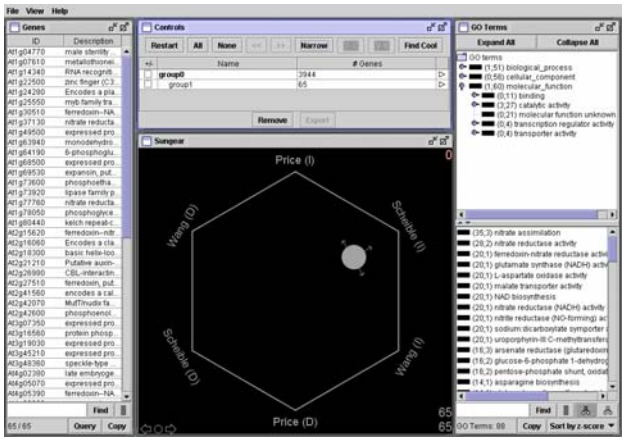


Figure 2

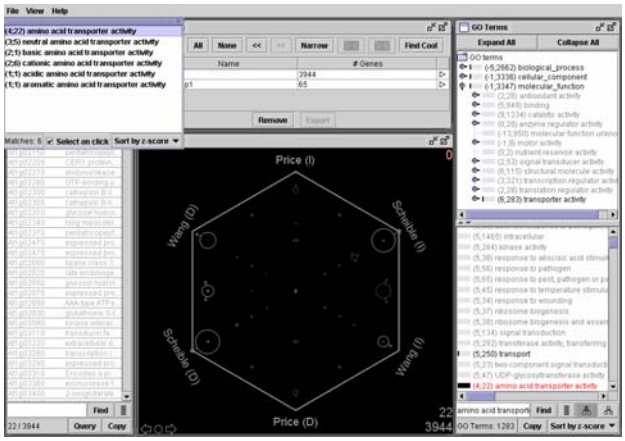
A



B



C



D

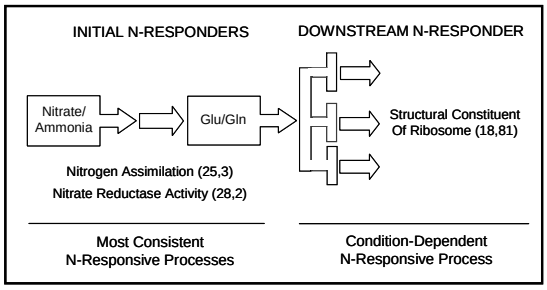
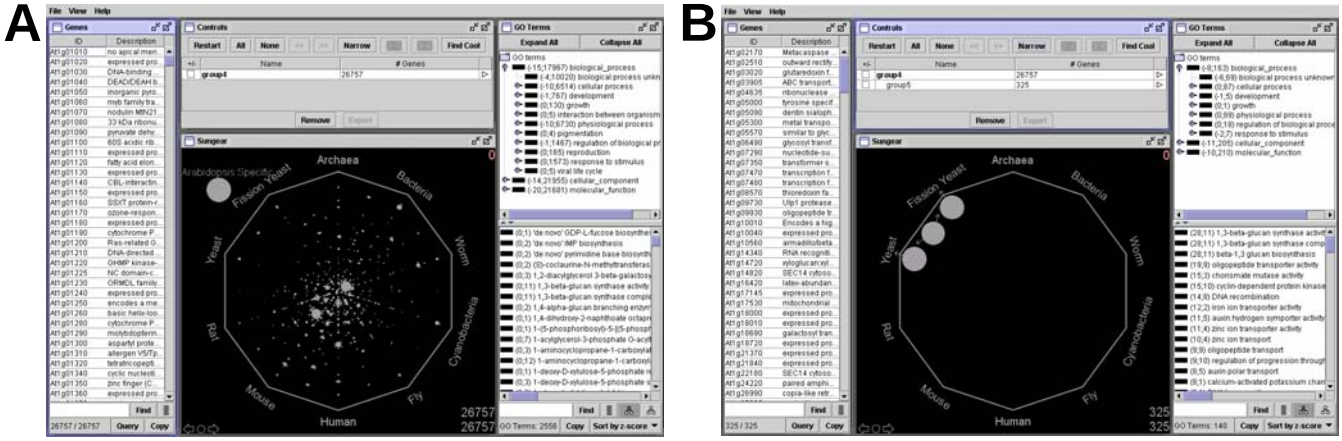
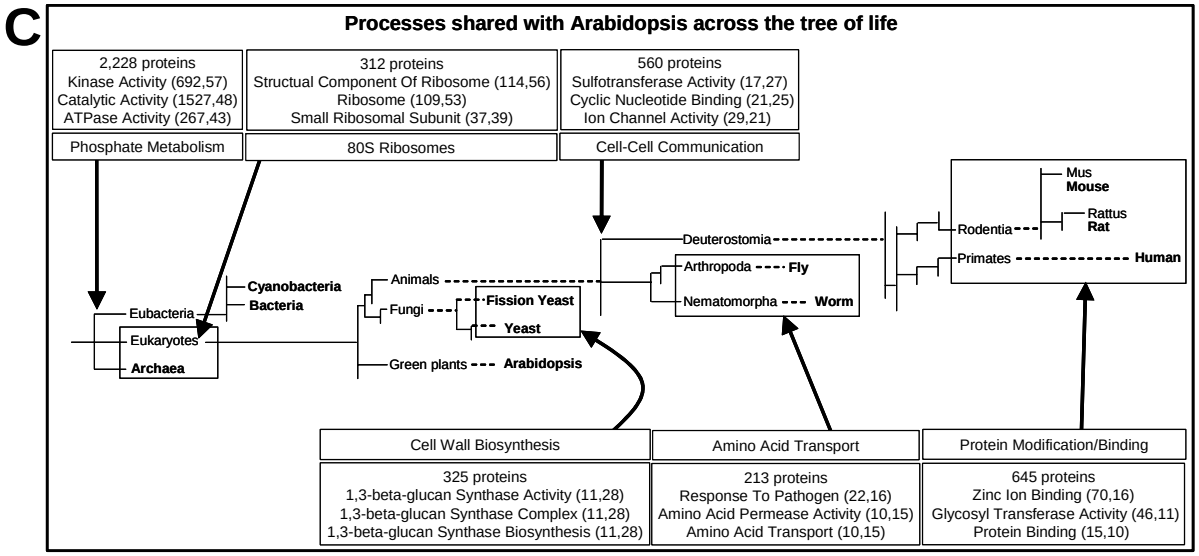
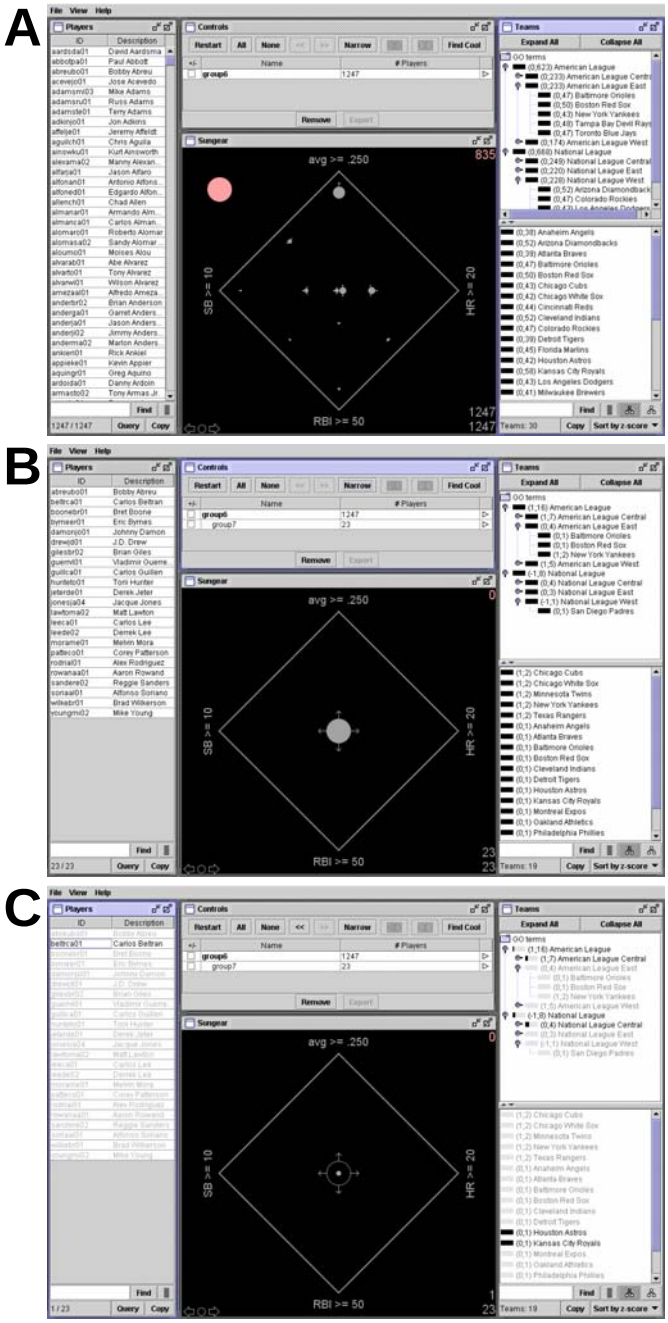


Figure 3


A
B
Archaea
Bacteria
Worm
Cyanobacteria
Fly
Human
Mouse
Rat
Yeast
Fission Yeast


C
Processes shared with Arabidopsis across the tree of life
2,228 proteins
Kinase Activity (692,57)
Catalytic Activity (1527,48)
ATPase Activity (267,43)
Phosphate Metabolism
312 proteins
Structual Component Of Ribosome (114,56)
Ribosome (109,53)
Small Ribosomal Subunit (37,39)
80S Ribosomes
560 proteins
Sulfotransferase Activity (17,27)
Cyclic Nucleotide Binding (21,25)
Ion Channel Activity (29,21)
Cell-Cell Communication
Eubacteria
Cyanobacteria
Bacteria
Eukaryotes
Archaea
Animals
Fungi
Fission Yeast
Yeast
Green plants
Arabidopsis
Deuterostomia
Arthropoda
Fly
Nematomorpha
Worm
Rodentia
Mus
Mouse
Rattus
Rat
Primates
Human
Cell Wall Biosynthesis
325 proteins
1,3-beta-glucan Synthase Activity (11,28)
1,3-beta-glucan Synthase Complex (11,28)
1,3-beta-glucan Synthase Biosynthesis (11,28)
Amino Acid Transport
213 proteins
Response To Pathogen (22,16)
Amino Acid Permease Activity (10,15)
Amino Acid Transport (10,15)
Protein Modification/Binding
645 proteins
Zinc Ion Binding (70,16)
Glycosyl Transferase Activity (46,11)
Protein Binding (15,10)

Supplemental Figure 1



Supplemental Figure 1. Use of Sungear to analyze non-scientific data. **(A)** Baseball players are the 'genes' in this example, with the players included whose performance during the 2004 season met the criterion of either: (1) batting average of .250 or better (avg >= .250), (2) 20 or more home runs (HR >= 20), (3) 50 or more runs batted in (RBI >= 50), and (4) 10 or more stolen bases during the season (SB >= 10). **(B).** 23 players meet all four criteria (central vessel). **(C).** Narrowing on the selection of players associated with both the National and American leagues (the equivalent of GO terms) yields a single player, Carlos Beltran (see name in bold on the left).

Supplemental Table 1. Summary showing the top ten GO terms ranked by z-score (in parentheses), and the numbers of genes or proteins within the vessels discussed in case studies 1 (A) and 2 (B).

(A) Case Study 1 - Robustness of Genomic Data

Group (number of genes)	Top 10 GO terms (Z-score; No of genes) GO term name
Induced in all labs (65)	(35;3) nitrate assimilation
	(28;2) nitrate reductase activity
	(20;1) ferredoxin-nitrate reductase activity
	(20;1) glutamate synthase (NADH) activity
	(20;1) L-aspartate oxidase activity
	(20;1) malate transporter activity
	(20;1) NAD biosynthesis
	(20;1) nitrate reductase (NADH) activity
	(20;1) nitrite reductase (NO-forming) activity
	(20;1) sodiumdicarboxylate symporter activity
Induced in Wang et al. alone (237)	(10;1) allene oxide synthase activity
	(10;1) basic amino acid transporter activity
	(10;1) dihydroflavonol(thiole) lyase activity
	(10;1) dihydrokaempferol 4-reductase activity
	(10;1) epoxygenase P450 pathway
	(10;1) glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity
	(10;1) GTP cyclohydrolase I activity
	(10;1) response to arsenic
	(10;1) sulfate reduction, APS pathway
	(10;1) tRNA adenylyltransferase activity
Induced in Price et al. alone (62)	(20;1) octadecanal decarbonylase activity
	(20;1) regulation of membrane potential
	(20;1) somatic cell DNA recombination
	(14;1) recombinase activity
	(12;2) inward rectifier potassium channel activity
	(11;3) cation channel activity
	(11;3) cyclic nucleotide binding
	(11;1) double-strand break repair via homologous recombination
	(11;1) female meiosis
	(11;1) inositol oxygenase activity

Induced in Scheible et al. alone (1101)	(18;81) structural constituent of ribosome
	(17;79) ribosome
	(16;83) structural molecule activity
	(11;27) ribosome biogenesis
	(11;28) ribosome biogenesis and assembly
	(9;470) cytoplasm
	(9;507) intracellular
	(7;17) ATP-dependent helicase activity
	(7;144) biosynthesis
	(7;249) chloroplast
Depressed in Wang et al. alone (510)	(9;13) response to water
	(8;2) cellular response to water deprivation
	(8;4) negative regulation of ethylene mediated signaling pathway
	(8;3) trypsin inhibitor activity
	(7;1) alpha-ketoacid dehydrogenase activity
	(7;1) alpha-N-arabinofuranosidase activity
	(7;1) aluminum ion transport
	(7;1) cinnamic acid biosynthesis
	(7;1) dihydrolipoamide branched chain acyltransferase activity
	(7;8) ethylene mediated signaling pathway
Depressed in Price et al. alone (45)	(10;1) cyclin binding
	(9;1) ammonium transporter activity
	(9;1) cyclin-dependent protein kinase inhibitor activity
	(9;1) negative regulation of cyclin dependent protein kinase activity
	(9;1) voltage-gated ion-selective channel activity
	(7;1) histidine phosphotransfer kinase activity
	(6;1) fucosyltransferase activity
	(6;1) mitochondrial outer membrane
	(4;1) anion transport
	(3;1) cell wall loosening (sensu Magnoliophyta)

Depressed in Scheible et al. alone (1116)	(10;430) catalytic activity (9;9) autophagy (9;188) transferase activity (8;115) kinase activity (8;122) transferase activity, transferring phosphorus-containing groups (6;2) acidic amino acid transport (6;2) cytokinin receptor activity (6;44) defense response (6;3) protein tyrosine phosphatase activity (6;4) protein tyrosine/serine/threonine phosphatase activity

(B) Case Study 2 - Comparative Genomics

Group (number of proteins)	**Top 10 GO terms (Z-score; No of proteins) GO term name**
Arabidopsis-specific (13562)	(46;7257) molecular function unknown (41;7596) biological process unknown (17;5191) cellular component unknown (14;1253) transcription factor activity (13;1314) transcription regulator activity (11;874) regulation of transcription (10;1403) DNA binding (9;876) transcription (7;202) anchored to membrane (6;92) lipid transport
All organisms (2228)	(57;692) kinase activity (48;1527) catalytic activity (43;267) ATPase activity (40;271) pyrophosphatase activity (34;237) protein kinase activity (33;129) ATPase activity, coupled to transmembrane movement of substances (29;157) protein amino acid phosphorylation (28;163) phosphorylation (27;188) ATP binding (27;104) helicase activity

Multicellular eukaryotes (560)	(27;17) sulfotransferase activity (25;21) cyclic nucleotide binding (21;29) ion channel activity (17;16) actin binding (16;8) calcium-dependent phospholipid binding (15;5) extracellular matrix (sensu Metazoa) (15;99) protein binding (13;7) microsporogenesis (12;6) diacylglycerol kinase activity (12;6) protein kinase C activation
Eukaryotes & Archaea (312)	(56;114) structural constituent of ribosome (53;109) ribosome (39;37) small ribosomal subunit (38;20) proteasome core complex (sensu Eukaryota) (37;33) cytosolic ribosome (sensu Eukaryota) (37;31) cytosolic small ribosomal subunit (sensu Eukaryota) (31;125) protein biosynthesis (27;9) protein phosphatase type 1 activity (26;22) large ribosomal subunit (25;8) DNA replication initiation
Cyanobacteria & Arabidopsis (349)	(30;61) carboxylic ester hydrolase activity (24;9) beta-fructofuranosidase activity (24;18) cellulose synthase activity (24;34) thylakoid (21;21) thylakoid lumen (sensu Viridiplantae) (20;8) oxygen evolving complex (19;182) chloroplast (19;25) lipase activity (19;18) polysaccharide biosynthesis (18;8) cellulose biosynthesis
Arabidopsis & either yeast or fission yeast (325)	(28;11) 1,3-beta-glucan synthase activity (28;11) 1,3-beta-glucan synthase complex (28;11) beta-1,3 glucan biosynthesis (19;9) oligopeptide transporter activity (15;3) chorismate mutase activity (15;10) cyclin-dependent protein kinase activity (14;8) DNA recombination (12;2) iron ion transporter activity (11;5) auxin:hydrogen symporter activity (11;4) zinc ion transporter activity

Arabidopsis & either worm or fly (213)	(16;22) response to pathogen
	(15;10) amino acid permease activity
	(15;10) amino acid transport
	(13;10) amino acid-polyamine transporter activity
	(11;1) euchromatin
	(11;1) sphingosine transporter activity
	(10;13) UDP-glycosyltransferase activity
	(7;1) acidic amino acid transporter activity
	(7;1) carboxypeptidase A activity
	(7;1) molybdenum incorporation into molybdenum-molybdopterin complex
Arabidopsis & either human, mouse, or rat (645)	(16;70) zinc ion binding
	(11;46) transferase activity, transferring glycosyl groups
	(10;82) protein binding
	(10;14) response to pathogenic bacteria
	(10;3) sialyltransferase activity
	(8;6) acetylglucosaminyltransferase activity
	(8;20) UDP-glycosyltransferase activity
	(7;20) carbohydrate biosynthesis
	(6;1) 1-phosphatidylinositol-3-phosphate 5-kinase activity
	(6;1) amino-terminal vacuolar sorting propeptide binding