BMC Bioinformatics

Methodology article

Genome comparison using Gene Ontology (GO) with statistical testing

Zhaotao Cai, Xizeng Mao, Songgang Li and Liping Wei*

Address: Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, P.R. China

Email: Zhaotao Cai - cait@mail.cbi.pku.edu.cn; Xizeng Mao - maoxz@mail.cbi.pku.edu.cn; Songgang Li - lsg@pku.edu.cn;
Liping Wei* - weilp@mail.cbi.pku.edu.cn

* Corresponding author

Published: 11 August 2006

BMC Bioinformatics 2006, **7**:374 doi:10.1186/1471-2105-7-374

Received: 24 February 2006
Accepted: 11 August 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/374

Abstract

Background: Automated comparison of complete sets of genes encoded in two genomes can provide insight on the genetic basis of differences in biological traits between species. Gene ontology (GO) is used as a common vocabulary to annotate genes for comparison. Current approaches calculate the fold of unweighted or weighted differences between two species at the high-level GO functional categories. However, to ensure the reliability of the differences detected, it is important to evaluate their statistical significance. It is also useful to search for differences at all levels of GO.

Results: We propose a statistical approach to find reliable differences between the complete sets of genes encoded in two genomes at all levels of GO. The genes are first assigned GO terms from BLAST searches against genes with known GO assignments, and for each GO term the abundance of genes in the two genomes is compared using a chi-squared test followed by false discovery rate (FDR) correction. We applied this method to find statistically significant differences between two cyanobacteria, *Synechocystis* sp. PCC6803 and *Anabaena* sp. PCC7120. We then studied how the set of identified differences vary when different BLAST cutoffs are used. We also studied how the results vary when only subsets of the genes were used in the comparison of human *vs.* mouse and that of *Saccharomyces cerevisiae* vs. *Schizosaccharomyces pombe*.

Conclusion: There is a surprising lack of statistical approaches for comparing complete genomes at all levels of GO. With the rapid increase of the number of sequenced genomes, we hope that the approach we proposed and tested can make valuable contribution to comparative genomics.

Background

Comparison of two completely sequenced genomes sheds lights on the genetic basis of differences in biological traits between species. Of particular interest is the comparison of complete sets of genes and gene products encoded in two genomes. Manual comparison is important but time-consuming and labor-intensive at the whole-genome scale and thus must be aided by automated approaches.

Unambiguous automated comparison requires that both genomes be annotated with the same structured, controlled vocabulary. Currently, the most common choice for such a vocabulary is gene ontology (GO) [1]. The November 15, 2005 version of GO contained 19,025 terms in three hierarchical structures—as Directed Acyclic Graphs (DAGs)—termed Biological Processes, Cellular Components, and Molecular Functions. Every branch in the graph represents a biological concept progressing from general to specialized with increasing graph depth. The depth of the branches in the graphs varies, with levels ranging from 2 to 15.

The GO web site currently lists 31 genomes that have been annotated with GO [2]. The annotations that are of the highest quality and updated most frequently are usually carried out by researchers who sequence and study a particular species; these annotations are primarily stored in species-specific databases such as SGD [3] for *Saccharomyces cerevisiae*, FlyBase [4] for *Drosophila melanogaster*, WormBase [5] for *Caenorhabditis elegans*, MGI [6] for *Mus musculus*, and TAIR [7] for *Arabidopsis thaliana*. Since these species-specific databases are located in different sites on the web, there is need for integrated, searchable databases that contain annotations for multiple species. The GO Consortium has developed such a resource, called AMIGO [8], that allows users to search and browse GO annotations integrated from many species-specific databases. Additionally, the European Bioinformatics Institute (EBI) has developed the Gene Ontology Annotation (GOA) database [9] that provides GO annotations for non-redundant proteins from many species in UniProt [10,11]. We compare these two resources in the Methods section. In addition to sequences annotated with GO, 15,754 functional domains in the InterPro domain database [12] have been linked to 2,627 GO terms [13].

Using the above-mentioned resources, there are two main types of methods developed to automatically annotate new gene products with GO terms: sequence similarity-based methods such as GOFigure [14], Goblet [15], OntoBlast [16], GOtcha [17], and Blast2GO [18], and sequence domain-based methods such as InterProScan [19] and GOTrees [20]. For genome-scale GO annotations the similarity-based, in particular BLAST-based methods have been the preferred choice [17,21-24]. BLAST is significantly faster than InterProScan and can annotate many more GO terms than InterProScan can. A recent evaluation showed that assigning GO terms of the top BLAST hit gave satisfactory results when compared with several more complex methods [25]. Thus we chose the BLAST approach in our work.

After the sets of genes encoded in the two genomes are annotated with GO, they can then be compared. The goal is to find functional categories that differ between the two genomes, which may explain differences in biological traits or suggest interesting families for further detailed investigation. The most common practice is to use tools such as GOslim [26,27] to tally the number of genes that fall within each functional category at the first level under Biological Processes, Cellular Components, and Molecular Functions, and then to compare between the two genomes. Because the two genomes usually differ in size, the absolute numbers of genes in each functional category need to be weighted before they are compared; they are often divided by the total number of genes in the respective genomes [28-30]. The results of the unweighted and weighted comparisons are usually presented as bar charts or fold changes.

The unweighted and weighted GO-based genome comparisons, although useful, have two drawbacks. First, focusing only on the high-level functional categories may miss differences that are detectable only at more refined levels. Second, bar charts or fold changes alone are not sufficient to separate true functional differences from those occurring by chance; thus, statistical testing of significance is necessary. Lessons can be learned from another, more extensively researched application of GO—the detection of significantly enriched GO categories in a set of co-expressed or differentially expressed genes in microarray experiments. Several tools have been developed to search complete GO trees (rather than just the high levels) and apply statistical testing of significance (e.g., Onto-Express [31]; FatiGO [32]; for an evaluation of these tools, see ref. [33]).

Contrary to the situation in microarray analysis, there is a surprising lack of statistical approaches for GO-based comparison of two genomes. Here we propose such a statistical approach to find reliable differences between the complete sets of genes encoded in two genomes at all levels of GO. For each GO term the abundance of genes in the two genomes is compared using chi-squared test followed by false discovery rate (FDR) correction. Furthermore, to analyze the reliability of the differences detected, we studied two important issues. First, when new sequences are assigned GO terms by similarity (as determined by BLAST) to other sequences having known GO assignments, the choice of BLAST cutoff may affect the results. We therefore analyzed the effects of employing a wide range of BLAST cutoffs. Second, we studied how the results vary when only subsets of the genes were used. To our knowledge, our work is the first to address all the aforementioned issues.

We used this statistical approach to compare two cyanobacterial genomes, *Synechocystis* sp. PCC6803 and *Anabaena* sp. PCC7120. Cyanobacteria (also called blue-green bacteria, blue-green algae, cyanophyceae, or cyanophytes) are important model organisms for the study of photosynthesis, nitrogen fixation, evolution of plant plastids, and survival in diverse environments [34-41]. Two of the most

Table 1: Comparison of AMIGO and GOA

GO Annotation Database	AMIGO[1]	GOA[2]
Curator	GO Consortium	European Bioinformatics Institute (EBI)
URL	http://www.godatabase.org/	http://www.ebi.ac.uk/GOA/
Total number of species	129,722	96,203
Total number of associations	7,745,168	7,600,805
Total number of non-redundant sequences	219,341	1,605,096
Total number of GO terms	10,916	9,258
Total number of other databases integrated	143	14

[1] AMIGO monthly release (November 1, 2005) downloaded from http://archive.godatabase.org/full/2005-11-01/go_200511-seqdb-data.gz
[2] GOA version 33.0 (October 25, 2005) downloaded from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz.

widely studied cyanobateria species are *Synechocystis* sp. PCC6803 and *Anabaena* sp. PCC7120. PCC6803 is a fresh water unicellular cyanobacterium incapable of nitrogen fixation [42]; PCC7120 is a filamentous, heterocyst-forming cyanobacterium that has long been used to study the genetics and physiology of cellular differentiation, pattern formation, and nitrogen fixation [43]. These interesting biological differences as well as the appropriate evolutionary distance between PCC6803 and PCC7120 make them a popular pair of species to compare and contrast[34,44-50]. We compared PCC6803 and PCC7120 genomes using our statistical method and evaluated the detected statistically significant differences against known biological differences. To analyze how results change when only subsets of the genes are used, a larger set of statistically significant differences is desirable and we used the comparison of human *vs.* mouse and that of *Saccharomyces cerevisiae vs. Schizosaccharomyces pombe* genomes.

Results

Whole-genome GO annotation

To annotate a new sequence, we used BLAST to compare it against a database of sequences with known GO annotations. Such a database should contain as many annotated sequences as possible from as many species as possible. AMIGO and GOA are two primary choices for such a database. We compared AMIGO and GOA, as shown in Table 1. Both databases have unique merit. AMIGO has been integrated to a greater extent with other databases and provides a better browsing function on the web, whereas GOA contains more sequences. For our purpose, it was attractive to have a larger collection of sequences for comparisons using BLAST, and thus we chose the GOA database. We set the default BLAST cutoff E-value to be 1E-20. With this method, a gene is assigned the GO terms of its top BLAST hit in GOA; it is also linked to all parent GO terms by propagating the DAG structures. Finally, the number of genes assigned to each GO term is tallied, representing the abundance of genes in each GO function within the genome.

We were able to annotate 2,224 genes in the PCC6803 genome to 1,933 GO terms, and 3,348 genes in the PCC7120 genome to 1,947 GO terms.

Testing the statistical significance of detected differences between genomes

For each GO category, we used the chi-squared test to determine whether the numbers of genes from the two genomes were statistically significantly different [51]. Since the total number of GO categories is large, a large number of tests is required. We adopted the widely used FDR correction (*q*-value cutoff = 0.01) to control the overall false positive rate [52]. We chose rather strict criteria to ensure reliability of the results; they can be set differently by other users.

We found seven terms in the GO Biological Process category that were statistically significantly different between the two genomes, including "transition metal ion transport" (GO:0000041, *q*-value 6.1E-6), "di-, trivalent inorganic cation transport" (GO:0015674, *q*-value 6.1E-6), "cobalt ion transport" (GO:0006824, *q*-value 7.3E-05), "metal ion transport" (GO:0030001, *q*-value 0.00056,), "protein amino acid phosphorylation" (GO:0006468, *q*-value 0.0021), "cellular biosynthesis" (GO:0044249, *q*-value 0.0022) and "nitrogen fixation" (GO:0009399, *q*-value 0.0094). These differences are shown in Figure 1 and discussed below. (The differences detected in the Molecular Function and Cellular Component categories are available in the Additional file 1)

The PCC7120 genome contains significantly more genes in "cobalt ion transport" (GO:0006824) compared with PCC6803, likely a consequence of the multicellular nature of PCC7120. Close inspection showed that the statistically significant difference in parent nodes "transition metal ion transport" (GO:0000041), "di-, trivalent inorganic cation transport" (GO:0015674), and "metal ion transport" (GO:0030001) is a consequence of the difference in the subfamily "cobalt ion transport"

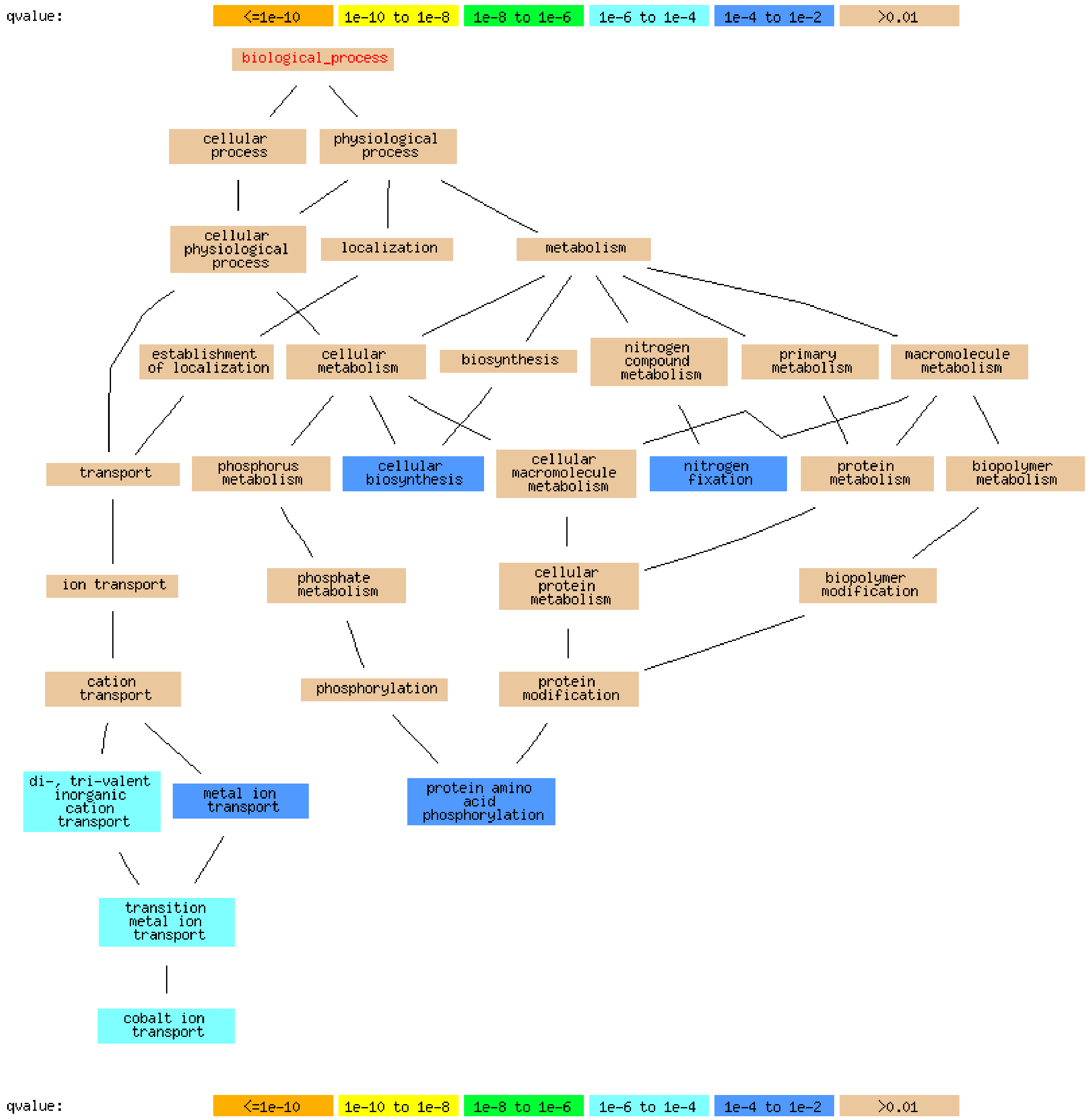

Figure 1
Comparison of PCC6803 and PCC7120 using our statistical approach. Comparison of PCC6803 and PCC7120 in the biological process category of GO, using the chi-squared test followed by FDR correction, with the *q*-value cutoff set to 0.01. The colors denote levels of statistical significance of differences between genomes, with the non-significant parent nodes of significant child nodes shown in tan color. (Results for the Molecular Function and Cellular Component categories are available in the Additional file 1)

(GO:0006824) rather than a cumulative effect of any other subfamilies. PCC7120 contains significantly more genes than PCC6803 in "protein amino acid phosphorylation" (GO:0006468). These genes are responsible for critical protein kinase functions in the multicellular PCC7120 [53-55]. The significantly greater number of genes in "nitrogen fixation" (GO:0009399) in PCC7120 is consistent with its ability to fix nitrogen, a function the simpler organism PCC6803 does not have. The "cellular biosynthesis" (GO:0044249) family differs from those above in that it is significantly more abundant in PCC6803 than in PCC7120. This result may be a consequence of PCC6803's rapid growth capability.

We compared the two genomes with regard to the GO molecular function category and obtained similar results. We then compared them with regard to the GO cellular component category and found three statistically significant differences: "cytoplasm" (GO:0005737), "integral to membrane" (GO:0016021), and "intrinsic to membrane" (GO:0031224), all of which are more abundant in PCC6803 than in PCC7120.

We compared our results with results from traditional GO-slim-based, weighted comparison. As shown in Figure 2, the fold difference in the GO-slim-based comparison ranged from 0.7 to 1.5. The fold difference gave only a rough indication of how much PCC6803 and PCC7120 differ in each high-level functional category. In addition, GO-slim-based approach compares two genomes at only the high level, as opposed to our approach that compares at every level and every node. Many important functional differences between two genomes may be detectable only at a finer level. For instance, GO-slim-based approach found little difference between the two cyanobacteria for GOslim term "metabolism, GO:0008152" in the Biological Process category (fold difference 1.03), whereas our approach found that the two species differ significantly in the sub-term "nitrogen fixation, GO:0009399", one of the most important known functional differences.

Effect of different BLAST cutoffs

We varied the BLAST E-value cutoff to study its effect on the number of statistically significant terms detected as well as the number of common terms between adjacent cutoffs. As shown in Figure 3, when the E-value cutoff is high (i.e., less strict, on the left end of the plot), the result is sensitive to the change in cutoff. The results stabilize around cutoff values of 1E-20 to 1E-40. We chose a default cutoff of 1E-20, which coincides with that chosen by GOblet [56].

Effect of partial data

Using the GO-based comparison method, we compared the human and mouse genomes and found 458 statistically significantly different GO terms. We randomly sampled 90% from each of the input gene sets for 1,000 times and compared the statistically significantly different GO terms from each sampling with those from the whole data. As shown in Figure 4 (Hatched bars, "Common GO terms"), most of the GO terms occurred in the majority of the samplings; 298 of the 458 GO terms occurred 1,000 times in all sampling results, whereas at the lower extreme three GO terms occurred only 169 times. This analysis offers an additional measure of reliability of the significant terms detected. The more times a term occurs in the samples, the more reliable it may be. We plotted the distribution of the "unique GO terms"—significant terms detected in one or more of the samples but not in the whole data set—and found that they occurred in as few as one and as many as 247 samples (Figure 4, open bars). As shown in Figure 4, the histogram distributions of the common and unique GO terms overlap slightly. We sampled 60%, 70%, and 80% of the input genes, respectively, and observed similar patterns (see the supplementary figures in the Additional file 1). We performed analysis of the comparison of the two yeast genomes of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* and also observed similar patterns (Figure 5 and supplementary figures in the Additional file 1).

Discussion

BLAST and InterProScan are two most widely used automated GO annotation methods. BLAST is the preferred choice for genome-scale annotation because it runs much faster and, perhaps more importantly, can annotate many more GO terms than InterProScan can. We had used InterProScan to annotate and compare PCC6803 and PCC7120, and found that it missed some important differences including "nitrogen fixation, GO:0009399". However, BLAST has its own limitations. Accurate functional assignment is difficult in cases where the match is less well defined due to lower sequence similarity [57]. In future research we will investigate how to combine results from BLAST and InterProScan to improve annotation quality and use grid computing to reduce computation time.

We used BLAST E-value cutoff as the criteria in assigning GO terms. Local sequence alignment programs such as BLAST may prefer short strong matches to long weak matches and may cause inaccurate GO assignment. The strict E-value cutoff we chose in our analysis ensured the relatively high quality of the results. It was reported that a match between two sequences is most likely reliable if the alignment is at least 70 residues in length with at least 40% sequence identity [58]. We investigated the quality of the HSP (High scoring Segment Pair) in our BLAST results (detail provided in the Additional file 1). With E-value cutoff 1e-20, the minimum length of HSP was 64 and the minimum sequence identity was 68%. Thus the assignments in our results were reliable. It is possible that false negatives may occur with a strict cutoff. In our analysis we prefer accuracy to coverage. Others can use different criteria depending on their individual goals. The statistical testing we proposed in this paper is independent of the GO assignment method. We suggest doing the comparison and comparing the results using different E-value cut-

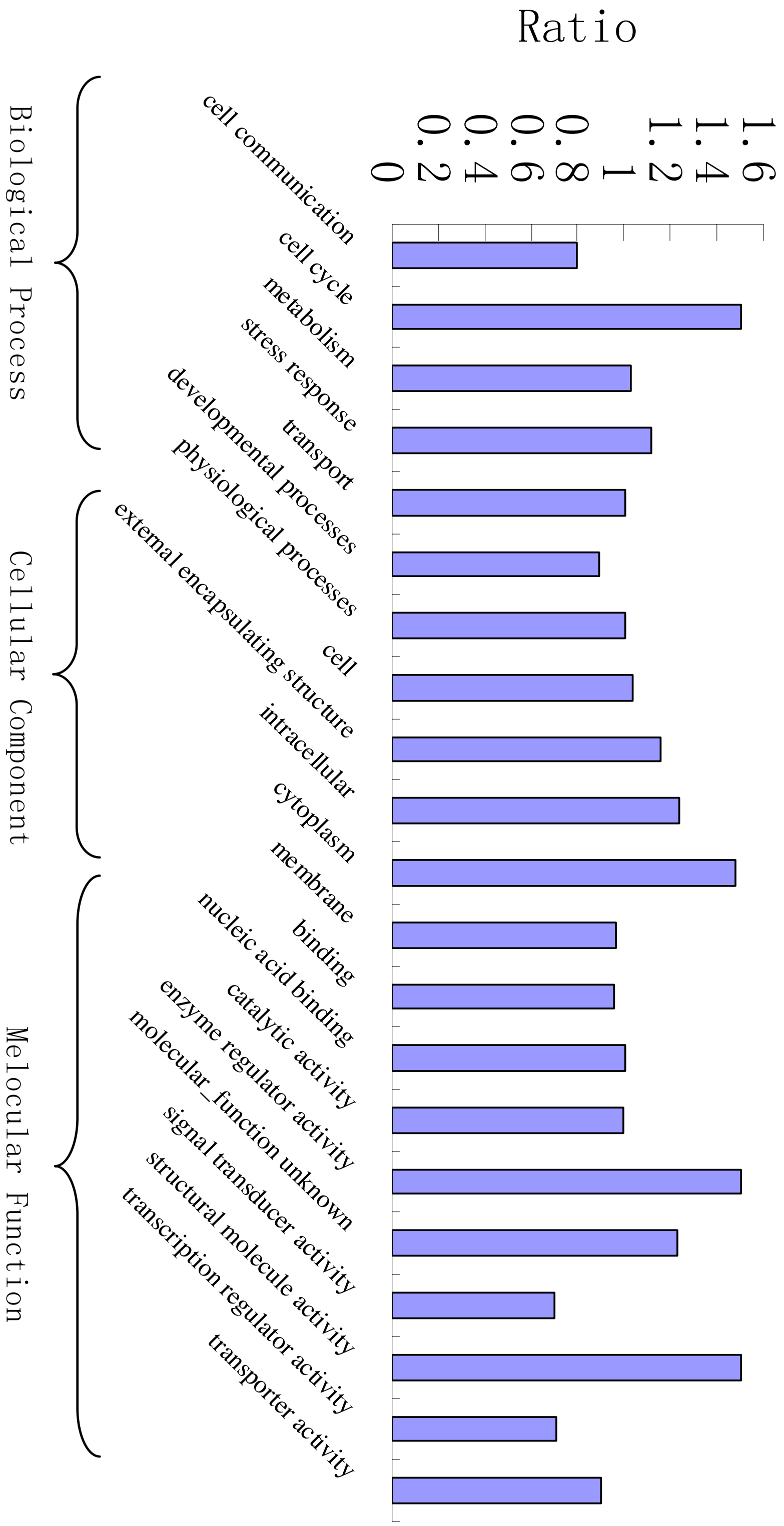

Figure 2

GOslim-based weighted comparison of PCC6803 and PCC7120. The bars show the fold difference between PCC6803 and PCC7120 in each GOslim functional catalogory, calculated by the weighted number of genes belonging to the functional category in PCC6803 divided by that in PCC7120. If there is no difference, the fold difference is equal to 1.

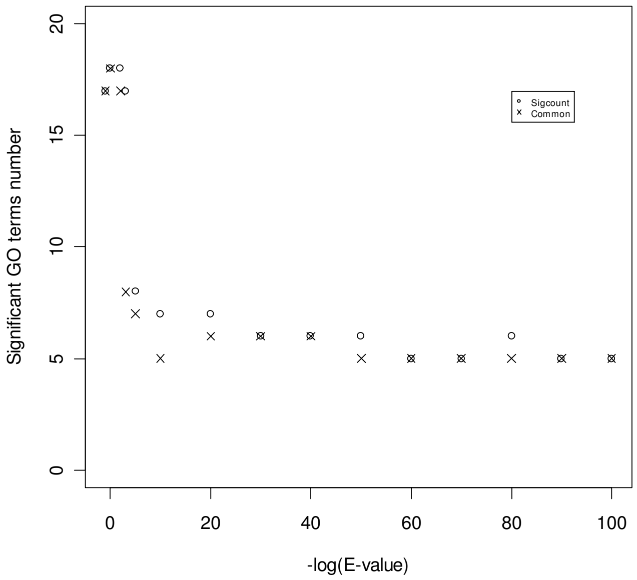

Figure 3
Effect of different BLAST cutoffs on GO results. This figure illustrates how much the result changes when the BLAST cutoff is changed. Circles show the number of significant GO terms ("Sigcount") at each cutoff. The symbol '× ' indicates the number of significant GO terms common between a given cutoff and its nearest right neighbor. The BLAST cutoff values range from 1E-100 to 10 (1E-100, 1E-90, 1E-80, 1E-70, 1E-60, 1E-50, 1E-40, 1E-30, 1E-20, 1E-10, 1E-5, 0.001, 0.01, 1, 10). The results stabilize around cutoff values of 1E-20 to 1E-40. We chose a default cutoff of 1E-20.

offs and different subsets of the input gene sets to identify the most reliable differences between two genomes.

In any GO analysis, the quality of the original GO annotation is critical. The GO annotation data are continuously expanded; however, the present data are incomplete and noisy [59], and the annotation quality is uneven, with a mix of literature-supported annotations and those inferred automatically. We did not modify the GO annotation data for our present study, but further research will consider the quality of the original GO annotations when assessing the reliability of the results. One limitation of our approach is that it only compared the number of genes in each functional category. It cannot capture differences in the level of gene expression. Another inherent limitation of GO is that it does not map directly to pathways. As a result GO-based comparison cannot detect differences at the pathway level. We have recently used the KEGG Orthology (KO) as an alternative controlled vocabulary in a KO-Based Annotation System (KOBAS) and demonstrated that KOBAS is effective in automated annotation and pathway identification [60]. In future research we will investigate KO-based comparison to compare two genomes at the pathway level.

Our goal is to achieve higher confidence in the differences detected between two genomes. Towards this end, we applied rigorous statistical testing followed by FDR correction instead of simply relying on fold changes. We also tested a wide range of BLAST cutoff values and different subsets of the input genes to provide additional measures of confidence in the results. If results beyond those having the highest confidence are required, then the cutoff values can be relaxed. The advantage of the statistical approach presented here is that, no matter what cutoff values are chosen, the resulting *p*-values, *q*-values, and sampling analysis can be used to assess the confidence in the results.

There are other procedures available to correct false positive rates resulting from multiple testing, including the Bonferroni correction, Sidak stepwise correction, Holm stepwise correction, Hochberg's stepwise correction, and others [61,62]. We chose the FDR correction because of its overall high quality and computational speed [63,64]. It is also the most common procedure used in GO-related and microarray analyses [62,65,66].

Conclusion

Contrary to the situation in microarray analysis, there is a surprising lack of statistical approaches used in GO-based comparison of two complete genomes. Our work is the first to propose and test a statistical approach to comparing the complete sets of genes in two whole genomes at all levels of GO and study the effect of varying BLAST cutoffs and using subset of the input gene sets. We believe that such an approach can provide a measure of confidence in the identified differences and help ensure the reliability of the results.

Methods

Supplementary materials and related programs for the paper are provided on-line [See Additional file 1].

Whole-genome GO annotation

We set the default BLAST cutoff E-value to be 1E-20. In Part 3 of results, we study the cutoff's effect on the final results. We parsed the BLAST result to obtain the GOA ID for the top hit and used the ID to query the GOA association database to retrieve the corresponding GO annotation and assign it to the query sequence. The result is written to a file in the format specified by the GO Consortium [67].

We parsed the gene ontology DAGs and stored the GO terms and their hierarchical relationships in a local data structure. The genes in a genome are linked to GO terms using the aforementioned approach; they are also linked

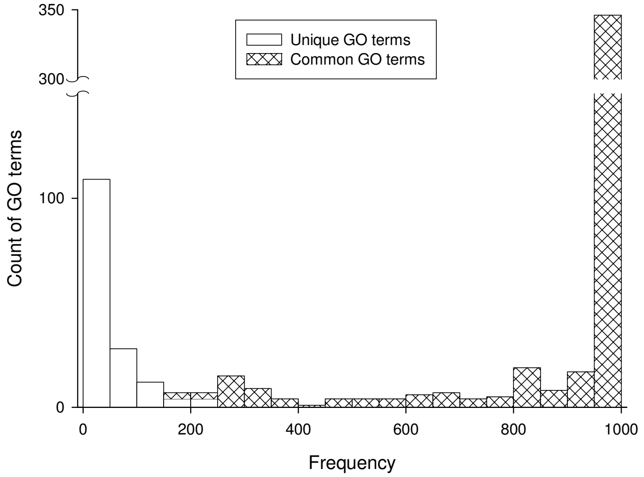

Figure 4

Histogram of sampling analysis results of the comparison between human and mouse genomes. The x-axis shows the number of samplings containing a significant GO term, grouped by 50. The y-axis shows the number of terms. The "Common GO terms" are those that occur both in the results from the complete data sets and in at least one sampling. The "Unique GO terms" are those that occur in the result from one or more samplings, but not in the results from the whole data set. For example, the right-most bar shows that 346 "Common GO terms" occurred in the results from 950 or more samples; the left-most bar shows that 109 "Unique GO terms" occurred in the results from less than fifty samples.

to all parent GO terms by propagating the DAG structures. If a gene has been assigned more than one GO terms that have a common parent GO term, the gene is counted only once in the parent GO term. Finally, the numbers of genes assigned to each GO term in the DAGs are tallied, representing the abundance of genes in each GO function within the genome.

The complete set of known and predicted genes in PCC6803 and PCC7120 genomes were downloaded from Cyanobase [68]. The PCC6803 genome contains 3,573,470 bp with 3,167 predicted ORFs; the PCC7120 genome contains 6,413,771 bp with 5,362 predicted ORFs.

Testing the statistical significance of detected differences between genomes

The goal is to identify all GO terms for which two genomes (A and B) are statistically significantly different. Define:

N = the total number of annotated genes in Genome A

n = the total number of annotated genes in Genome B

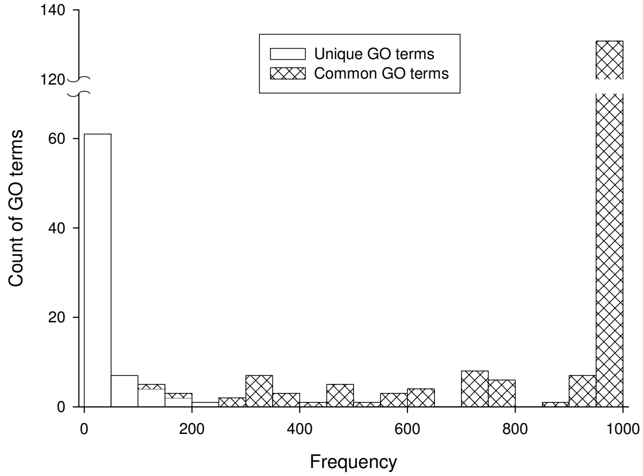

Figure 5

Histogram of sampling analysis results of the comparison between *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* genomes.

X = the number of genes in Genome A that are assigned the GO term currently under consideration

x = the number of genes in Genome B that are assigned the GO term currently under consideration

We used the chi-squared test to address whether the ratios, $p_0 = \frac{x}{n - x}$ and $p_1 = \frac{X}{N - X}$, come from the same distribution, either:

H_0: $p_0 = p_1$ or

H_1: $p_0 \neq p_1$

The *p*-value is calculated as the upper tail probability of the chi-squared distribution with one degree of freedom using the CPAN Statistics::Distributions modules [69].

Because the number of tests performed equals the number of GO terms, which may be thousands, multiple hypotheses testing is important to control the overall Type I error rate. We used the commonly applied FDR correction. For every test result that is considered statistically significant, the FDR correction calculates a *q*-value to measure the minimum FDR when calling that result significant. A *q*-value cutoff, α (alpha), guarantees that the expected proportion of false positives is α (alpha) among the set of significant features produced [52,66]. The default for α (alpha) was set to 0.01 in our study. The conservative FDR correction was implemented according to the GenTS package [70].

The statistically significantly different GO terms detected between two genomes are stored in text format, sorted by increasing *q*-value. We also modified the GO TermFinder package [71] to show the results graphically, with different colors showing different levels of significance.

All related programs are attached in Additional file 1

Effect of different BLAST cutoffs

We studied how the BLAST cutoff value can affect the comparison of results between two genomes of PCC6803 and PCC7120. We tested a wide range of BLAST E-value cutoffs, from 1E-100 to 10, and recorded the number of statistically significantly different GO terms between the two cyanobacterial genomes at each cutoff. We then recorded the number of common statistically significantly different GO terms between adjacent cutoffs to show how much the result changes when the cutoff is varied.

Effect of partial data

We performed the random sampling to study how the results are affected when only part of the data is used. For each sample, we randomly selected 90%, 80%, 70%, and 60% of the annotated genes in each genome, and recomputed the statistically significantly different GO terms. We then compared the result of each sampling with that for the complete data sets and counted the numbers of common and unique GO terms. Because comparison of the two cyanobacteria resulted in too few significant GO terms to make this analysis meaningful, we analyzed the comparison of human *vs.* mouse and *Saccharomyces cerevisiae vs. Schizosaccharomyces pombe*. The GO annotations for these four genomes were retrieved from the Gene Ontology Consortium web site.

Authors' contributions

ZC and LW conceived of the study; ZC carried out most of the implementation and analysis; XM and LW participated in the analysis; SL participated in the design of statistical tests. All authors participated in preparation of the manuscript.

Additional material

Additional File 1

Supplementary materials and related programs*. The compressed file contains supplementary materials and related programs for the paper, including the source codes and documents, the genome comparison results between PCC6803_PCC7120, Cerevisiae_Pombe and Human_Mouse, the figures for the effect of using different subsets of the input genes and the statistical analysis about the BLAST HSP (High scoring Segment Pair) length. Please unzip the file and read the "index.htm" for detail. Also, you can visit the website for the information (http://www.cbi.pku.edu.cn/cbird/GO/).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-374-S1.zip]

Acknowledgements

This work was supported by the China Ministry of Science and Technology "863" grants. We thank Drs. Arthur Grossman and Devaki Bhaya for insightful discussions. We thank the two anonymous reviewers for helpful suggestions.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
2. **Current annotated genomes in GO web site** [http://www.geneontology.org/GO.current.annotations.shtml]
3. **SGD** [http://www.yeastgenome.org/]
4. **FlyBase** [http://www.fruitfly.org/]
5. **WormBase** [http://www.wormbase.org/]
6. **MGI** [http://www.informatics.jax.org/]
7. **TAIR** [http://www.arabidopsis.org/]
8. **AMIGO** [http://www.godatabase.org]
9. **GOA** [ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/]
10. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Res* 2004, **32 Database issue:**D115-9.
11. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32 Database issue:**D262-6.
12. **InterPro** [http://www.ebi.ac.uk/interpro]
13. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM: **The InterPro Database, 2003 brings increased coverage and new features.** *Nucleic Acids Res* 2003, **31:**315-318.
14. Khan S, Situ G, Decker K, Schmidt CJ: **GoFigure: automated Gene Ontology annotation.** *Bioinformatics* 2003, **19:**2484-2485.
15. Groth D, Lehrach H, Hennig S: **GOblet: a platform for Gene Ontology annotation of anonymous sequence data.** *Nucleic Acids Res* 2004, **32:**W313-7.
16. Zehetner G: **OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms.** *Nucleic Acids Res* 2003, **31:**3799-3803.
17. Martin DM, Berriman M, Barton GJ: **GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.** *BMC Bioinformatics* 2004, **5:**178.

18. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.
19. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17:**847-848.
20. Hayete B, Bienkowska JR: **Gotrees: predicting go associations from protein domain composition using decision trees.** *Pac Symp Biocomput* 2005:127-138.
21. El-Sayed NM, Ghedin E, Song J, MacLeod A, Bringaud F, Larkin C, Wanless D, Peterson J, Hou L, Taylor S, Tweedie A, Biteau N, Khalak HG, Lin X, Mason T, Hannick L, Caler E, Blandin G, Bartholomeu D, Simpson AJ, Kaul S, Zhao H, Pai G, Van Aken S, Utterback T, Haas B, Koo HL, Umayam L, Suh B, Gerrard C, Leech V, Qi R, Zhou S, Schwartz D, Feldblyum T, Salzberg S, Tait A, Turner CM, Ullu E, White O, Melville S, Adams MD, Fraser CM, Donelson JE: **The sequence and analysis of Trypanosoma brucei chromosome II.** *Nucleic Acids Res* 2003, **31:**4856-4863.
22. Buell CR, Joardar V, Lindeberg M, Selengut J, Paulsen IT, Gwinn ML, Dodson RJ, Deboy RT, Durkin AS, Kolonay JF, Madupu R, Daugherty S, Brinkac L, Beanan MJ, Haft DH, Nelson WC, Davidsen T, Zafar N, Zhou L, Liu J, Yuan Q, Khouri H, Fedorova N, Tran B, Russell D, Berry K, Utterback T, Van Aken SE, Feldblyum TV, D'Ascenzo M, Deng WL, Ramos AR, Alfano JR, Cartinhour S, Chatterjee AK, Delaney TP, Lazarowitz SG, Martin GB, Schneider DJ, Tang X, Bender CL, White O, Fraser CM, Collmer A: **The complete genome sequence of the Arabidopsis and tomato pathogen Pseudomonas syringae pv. tomato DC3000.** *Proc Natl Acad Sci U S A* 2003, **100:**10181-10186.
23. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RKJ, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Res* 2003, **31:**5654-5666.
24. Wortman JR, Haas BJ, Hannick LI, Smith RKJ, Maiti R, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, White OR, Town CD: **Annotation of the Arabidopsis genome.** *Plant Physiol* 2003, **132:**461-468.
25. Jones CE, Baumann U, Brown AL: **Automated methods of predicting the function of biological sequences using GO and BLAST.** *BMC Bioinformatics* 2005, **6:**272.
26. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
27. **GOslim** [http://geneontology.org/GO.slims.shtml]
28. McCarter JP, Mitreva MD, Martin J, Dante M, Wylie T, Rao U, Pape D, Bowers Y, Theising B, Murphy CV, Kloek AP, Chiapelli BJ, Clifton SW, Bird DM, Waterston RH: **Analysis and functional classification of transcripts from the nematode Meloidogyne incognita.** *Genome Biol* 2003, **4:**R26.
29. Mitreva M, McCarter JP, Martin J, Dante M, Wylie T, Chiapelli B, Pape D, Clifton SW, Nutman TB, Waterston RH: **Comparative genomics of gene expression in the parasitic and free-living nematodes Strongyloides stercoralis and Caenorhabditis elegans.** *Genome Res* 2004, **14:**209-220.
30. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, Coulson A, D'Eustachio P, Fitch DH, Fulton LA, Fulton RE, Griffiths-Jones S, Harris TW, Hillier LW, Kamath R, Kuwabara PE, Mardis ER, Marra MA, Miner TL, Minx P, Mullikin JC, Plumb RW, Rogers J, Schein JE, Sohrmann M, Spieth J, Stajich JE, Wei C, Willey D, Wilson RK, Durbin R, Waterston RH: **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics.** *PLoS Biol* 2003, **1:**E45.
31. Khatri P, Draghici S, Ostermeier GC, Krawetz SA: **Profiling gene expression using onto-express.** *Genomics* 2002, **79:**266-270.
32. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.
33. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21:**3587-3595.
34. Raymond J, Blankenship RE: **The evolutionary development of the protein complement of photosystem 2.** *Biochim Biophys Acta* 2004, **1655:**133-139.
35. Koksharova OA, Wolk CP: **Genetic tools for cyanobacteria.** *Appl Microbiol Biotechnol* 2002, **58:**123-137.
36. Stewart WD, Rowell P, Rai AN: **Cyanobacteria-eukaryotic plant symbioses.** *Ann Microbiol (Paris)* 1983, **134B:**205-228.
37. Berman-Frank I, Lundgren P, Falkowski P: **Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria.** *Res Microbiol* 2003, **154:**157-164.
38. McFadden GI: **Endosymbiosis and evolution of the plant cell.** *Curr Opin Plant Biol* 1999, **2:**513-519.
39. Paerl HW, Pinckney JL, Steppe TF: **Cyanobacterial-bacterial mat consortia: examining the functional unit of microbial survival and growth in extreme environments.** *Environ Microbiol* 2000, **2:**11-26.
40. Thomas DN: **Photosynthetic microbes in freezing deserts.** *Trends Microbiol* 2005, **13:**87-88.
41. Bryant D: **The Molecular Biology of Cyanobacteria.** Netherlands, Kluwer Academic Publishers; 1994.
42. Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S: **Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions.** *DNA Res* 1996, **3:**109-136.
43. Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S: **Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120.** *DNA Res* 2001, **8:**205-13; 227-53.
44. Zhang CC, Gonzalez L, Phalip V: **Survey, analysis and genetic organization of genes encoding eukaryotic-like signaling proteins on a cyanobacterial genome.** *Nucleic Acids Res* 1998, **26:**3619-3625.
45. Knowles VL, Plaxton WC: **From genome to enzyme: analysis of key glycolytic and oxidative pentose-phosphate pathway enzymes in the cyanobacterium Synechocystis sp. PCC 6803.** *Plant Cell Physiol* 2003, **44:**758-763.

46. Kotani H, Tabata S: **Lessons from Sequencing of the Genome of a Unicellular Cyanobacterium, Synechocystis Sp. Pcc6803.** *Annu Rev Plant Physiol Plant Mol Biol* 1998, **49:**151-171.
47. Tamagnini P, Axelsson R, Lindberg P, Oxelfelt F, Wunschiers R, Lindblad P: **Hydrogenases and hydrogen metabolism of cyanobacteria.** *Microbiol Mol Biol Rev* 2002, **66:**1-20, table of contents.
48. Bhaya D, Dufresne A, Vaulot D, Grossman A: **Analysis of the hli gene family in marine and freshwater cyanobacteria.** *FEMS Microbiol Lett* 2002, **215:**209-219.
49. Su Z, Olman V, Mao F, Xu Y: **Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis.** *Nucleic Acids Res* 2005, **33:**5156-5171.
50. Martin KA, Siefert JL, Yerrapragada S, Lu Y, McNeill TZ, Moreno PA, Weinstock GM, Widger WR, Fox GE: **Cyanobacterial signature genes.** *Photosynth Res* 2003, **75:**211-221.
51. Man MZ, Wang X, Wang Y: **POWER_SAGE: comparing statistical tests for SAGE experiments.** *Bioinformatics* 2000, **16:**953-959.
52. Benjamini YYH: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Statist Soc B* 1995, **57:**289-300.
53. West AH, Stock AM: **Histidine kinases and response regulator proteins in two-component signaling systems.** *Trends Biochem Sci* 2001, **26:**369-376.
54. Foussard M, Cabantous S, Pedelacq J, Guillet V, Tranier S, Mourey L, Birck C, Samama J: **The molecular puzzle of two-component signaling cascades.** *Microbes Infect* 2001, **3:**417-424.
55. Wolanin PM, Thomason PA, Stock JB: **Histidine protein kinases: key signal transducers outside the animal kingdom.** *Genome Biol* 2002, **3:**REVIEWS3013.
56. Hennig S, Groth D, Lehrach H: **Automated Gene Ontology annotation for anonymous sequence data.** *Nucleic Acids Res* 2003, **31:**3712-3715.
57. Gerlt JA, Babbitt PC: **Can sequence determine function?** *Genome Biol* 2000, **1:**REVIEWS0005.
58. Brenner SE, Chothia C, Hubbard TJ: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.** *Proc Natl Acad Sci U S A* 1998, **95:**6073-6078.
59. Dolan ME, Ni L, Camon E, Blake JA: **A procedure for assessing GO annotation consistency.** *Bioinformatics* 2005, **21 Suppl 1:**i136-i143.
60. Mao X, Cai T, Olyarchuk JG, Wei L: **Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.** *Bioinformatics* 2005, **21:**3787-3793.
61. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics* 2004, **20:**3710-3715.
62. Nichols T, Hayasaka S: **Controlling the familywise error rate in functional neuroimaging: a comparative review.** *Stat Methods Med Res* 2003, **12:**419-446.
63. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures.** *Bioinformatics* 2003, **19:**368-375.
64. Qian HR, Huang S: **Comparison of false discovery rate methods in identifying genes with differential expression.** *Genomics* 2005, **86:**495-503.
65. Slonim DK: **From patterns to pathways: gene expression data analysis comes of age.** *Nat Genet* 2002, **32 Suppl:**502-508.
66. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100:**9440-9445.
67. **GO annotation file format** [http://www.geneontology.org/GO.annotation.html#file]
68. **Cyanobase** [http://www.kazusa.or.jp/cyanobase/]
69. **Statistics::Distributions modules** [http://search.cpan.org/~mikek/Statistics-Distributions-1.02/Distributions.pm]
70. **GenTS** [http://www.strimmerlab.org/software/genets/]
71. **GO TermFinder package** [http://search.cpan.org/~sherlock/GO-TermFinder-0.64/]