BMC Plant Biology

BioMed Central
The Open Access Publisher

Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case

BMC Plant Biology 2014, **14**:329 doi:10.1186/s12870-014-0329-9

David Amar (ddam.am@gmail.com)
Itziar Frades (maria.iciar.frades@slu.se)
Agnieszka Danek (agnieszka.danek@polsl.pl)
Tatyana Goldberg (goldberg@rostlab.org)
Sanjeev K Sharma (Sanjeev.Sharma@hutton.ac.uk)
Pete E Hedley (Pete.Hedley@hutton.ac.uk)
Estelle Proux-Wera (estelle.proux@scilifelab.se)
Erik Andreasson (erik.andreasson@slu.se)
Ron Shamir (rshamir@tau.ac.il)
Oren Tzfadia (oren.tzfadia@weizmann.ac.il)
Erik Alexandersson (erik.alexandersson@slu.se)

ISSN 1471-2229

Article type Research article

Submission date 12 June 2014

Acceptance date 10 November 2014

Article URL http://www.biomedcentral.com/1471-2229/14/329

Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case

David Amar[1]
Email: ddam.am@gmail.com

Itziar Frades[2]
Email: maria.iciar.frades@slu.se

Agnieszka Danek[3]
Email: agnieszka.danek@polsl.pl

Tatyana Goldberg[4]
Email: goldberg@rostlab.org

Sanjeev K Sharma[5]
Email: Sanjeev.Sharma@hutton.ac.uk

Pete E Hedley[5]
Email: Pete.Hedley@hutton.ac.uk

Estelle Proux-Wera[2,6]
Email: estelle.proux@scilifelab.se

Erik Andreasson[2]
Email: erik.andreasson@slu.se

Ron Shamir[1]
Email: rshamir@tau.ac.il

Oren Tzfadia[7*]
[*] Corresponding author
Email: oren.tzfadia@weizmann.ac.il

Erik Alexandersson[2]
Email: erik.alexandersson@slu.se

[1] Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

[2] Deptartment of Plant Protection Biology, Swedish University of Agricultural Sciences, Alnarp, Sweden

[3] Institute of Informatics, Silesian University of Technology, Akademicka 2A, 44-100 Gliwice, Poland

[4] Department for Bioinformatics and Computational Biology, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany

[5] Cell and Molecular Sciences, The James Hutton Institute, Aberdeen, Scotland, UK

[6] Current affiliation: SciLifeLab, Stockholm University, Universitetsvägen 10, 114 18 Stockholm, Sweden

[7] Department of Plant Science, The Weizmann Institute of Science, Rehovot, Israel

Abstract

Background

For most organisms, even if their genome sequence is available, little functional information about individual genes or proteins exists. Several annotation pipelines have been developed for functional analysis based on sequence, 'omics', and literature data. However, researchers encounter little guidance on how well they perform. Here, we used the recently sequenced potato genome as a case study. The potato genome was selected since its genome is newly sequenced and it is a non-model plant even if there is relatively ample information on individual potato genes, and multiple gene expression profiles are available.

Results

We show that the automatic gene annotations of potato have low accuracy when compared to a "gold standard" based on experimentally validated potato genes. Furthermore, we evaluate six state-of-the-art annotation pipelines and show that their predictions are markedly dissimilar (Jaccard similarity coefficient of 0.27 between pipelines on average). To overcome this discrepancy, we introduce a simple GO structure-based algorithm that reconciles the predictions of the different pipelines. We show that the integrated annotation covers more genes, increases by over 50% the number of highly co-expressed GO processes, and obtains much higher agreement with the gold standard.

Conclusions

We find that different annotation pipelines produce different results, and show how to integrate them into a unified annotation that is of higher quality than each single pipeline. We offer an improved functional annotation of both PGSC and ITAG potato gene models, as well as tools that can be applied to additional pipelines and improve annotation in other organisms. This will greatly aid future functional analysis of '-omics' datasets from potato and other organisms with newly sequenced genomes. The new potato annotations are available with this paper.

Keywords

Functional annotation, Gene ontology, Gene co-expression, Potato, Genomics

Background

Potato (*Solanum tuberosum)* is the 3rd largest food crop in terms of human consumption [1]. It is therefore important for our food security, and understanding its genome is called for. Examples of major challenges in potato research are its sensitivity to drought stress and its lack of resistance to certain diseases, e.g., the oomycete *Phytopthora infestans*, which caused the Irish famine in the 1840's. Farmers need to use large amounts of fungicides to protect their potato crops, thereby increasing the cost of cultivation and threatening the environment. For example, the global cost of protection and yield loss due to *P. infestans* has been estimated at €4800 M annually [2].

Recently, the potato genome (*Solanum tuberosum* group Phureja) was sequenced by the Potato Genome Sequencing Consortium (PGSC). The PGSC analysis of the genome reported gene models for 39,031 representative transcripts, and 56,218 including splicing variants [3]. In a later effort, the International Tomato Annotation Group (ITAG) produced new gene models by jointly analyzing the tomato and potato genomes [4]. These new gene models covered 34,727 and 35,004 predicted protein-coding genes for the tomato and the potato genomes, respectively. Unfortunately, few experimentally validated genes (e.g., by fluorescent-tagged proteins, or gene knock-outs) are available in newly sequenced genomes in which, unlike established model organisms, few genes have verified functions such as the case is for potato. Comprehensive and accurate functional annotation of the genes in such recently sequenced genomes is a prerequisite to efficient exploitation of these genomic data.

A key tool for functional annotation is the Gene Ontology (GO), which provides a structured set of defined terms representing gene properties [5]. The structure of gene ontology is composed of three major domains: *cellular component* (CC), the parts of a cell or its extracellular environment; *molecular function* (MF), the elemental activities of a gene product at the molecular level; and *biological process* (BP), which describes a set of functionally related molecular events. Thus, the complete GO structure provides a unified vocabulary of biological terms, which can also be used to evaluate biological similarity of different terms [6]. Annotating a gene means placing it within some or all of the three gene ontology domains.

Recent advances in plant science are marked by the rapidly increasing availability and quality of high-throughput sequencing data. The most basic usage of these data is gene function prediction, wherein GO plays a pivotal part. There are several computational suites like EXPANDER [7], MapMan [8], Mercator [9] and AmiGO [10] that enable biologists to run GO enrichment analyses in several plant model systems. This is usually done by first identifying a group of genes that behave similarly in a given expression dataset, seeking ontology terms highly enriched in the group, and associating the highly enriched functions with unannotated genes that belong to the same group. This process is sometimes called "guilt by association". Automated gene function annotation is also relevant for well-investigated plant model organisms, such as *Arabidopsis thaliana*, tomato, *Brachypodium* and rice, wherein ~40% of the genes still do not have any known function [11].

In order to assign functional annotation to sequenced plant transcripts, researchers can use several sequence-based annotation pipelines. For a comprehensive summary of methods and principles behind automated functional annotation see [12]. Some recent efforts have been made to characterize the annotation quality of plant genomes. For example, Jaramillo-

Garzón, et al. [13] used sequence features and showed high predictability of MF and CC terms and lower predictability of BP terms. However, the analysis was limited to a small subset of the GO terms (GO-Slim). Ramsak, et al. [8] presented GOMapMan, a tool for visualization and analysis of gene annotation in plants. In potato, information from orthologous gene families across 26 sequenced plant genomes was analyzed in order to increase the number of potato genes associated with GO terms [14]. Still, a robust, automated approach to evaluate and compare genome-wide annotation pipelines is direly needed.

A typical genome-wide functional annotation of newly sequenced organisms starts by using a single 'default' pipeline. Here, we analyzed the two sets of potato gene models, from the ITAG and PGSC. We compared six annotation pipelines: Trinotate HMM, Trinotate BLAST [15], OrthoMCL-UniProt [16], BLAST2GO [17], Phytozome [18] and InterPro2GO provided in BioMart [19] (Figure 1). These pipelines were chosen because they seek to provide a comprehensive annotation of the whole genome. Some of these pipelines are based solely on sequence similarity (BLAST), others rely on specific domains and some are based on clustering of groups of orthologous gene families. As we shall show, one clear conclusion of this work is that functional annotations of genomes should rely on more than one annotation pipeline.

Figure 1 Overview of pipeline comparison, validation of accuracy and integration processes. (A) The PGSC and ITAG gene models were used as input for the six pipelines assessed. **(B)** The annotation from each pipeline was transformed into gene ID – GO term associations. **(C)** Annotations were compared by the number of annotated gene models, the number of GO terms associated per gene model, and GO similarity. **(D)** The quality and comprehensiveness of the annotation of each pipeline were calculated by comparing their predictions to experimentally validated annotation (gold standard). In addition, gene co-expression data were used to test if genes predicted to share the same GO processes are significantly co-expressed. **(E)** An integrated annotation using the ensemble of results of all pipelines was created and validated using the same criteria in D. Results of the ensemble annotations were compared to those of the individual pipelines.

By examining the GO terms generated by these pipelines, we demonstrate that they predict very dissimilar annotations (e.g., on average, less than 30% of the genes annotated by two pipelines are assigned with the same function). To evaluate the performance of the pipelines we first created a set of potato genes (hereafter referred to as "gold standard"), with known functional characterization, including genes from the well characterized biosynthetic Carotenoids pathway. We show that pipelines may have rather low accuracy compared to the gold standard. Since the size of the gold standard is rather modest (116 PGSC genes ids), we used an additional validation scheme based on gene expression data. Under the premise that genes participating in the same biological process should have more similar expression pattern than expected by chance, we evaluated the predictions of each pipeline based on its intra-process gene co-expression level. We show that while all pipelines provide much higher intra-process co-expression than expected by chance, there are large differences among the methods. We introduce a simple method to combine the results of the different pipelines into a single integrated annotation. Compared to the single pipelines, it improved gene coverage, prediction precision, and the overall co-expression of predicted GO processes. In addition to improved annotation of potato genes, our analysis provides generic tools that can be applied to improve the annotation of other newly sequenced plants.

Results and discussion

A compendium of the state-of-the-art annotation tools

In this study, we tested automatic annotation pipelines on the potato genome. We used six state-of-the-art tools for GO gene function prediction: (1) Trinotate HMM, (2) Trinotate BLAST [15], (3) OrthoMCL-UniProt [16], (4) BLAST2GO [17], (5) Phytozome [18], and (6) InterPro2GO [19]. See Materials and Methods and Additional file 1: Methods S1-4 for details. We note that every program has its own set of parameters and fitting the best parameter combination for a particular dataset is a substantial effort. The common practice in this area is to use published tools with the default parameter values (see e.g. [20,21]. If necessary, we then mapped its predicted functions to GO terms using automated mapping files such as Pfam2GO, and the genes and transcripts to protein identifiers. Thus, in our analysis a gene corresponds to either a transcript or a protein that appeared in the output of the pipelines. Next, the output of each pipeline was summarized as a set of predicted gene-GO term pairs. For each gene we then retained only the most "specific" GO terms. That is, in case a gene is associated with two GO terms A and B, but B is a generalization of A (i.e. an ancestor of A in the GO hierarchy), we excluded B. We call this step *ancestor removal*. Note that after filtering, many genes were still associated with more than one GO term, since a gene can have several associated annotations none of which is an ancestor of another. For the output of all pipelines, see Additional file 2: Table S1, Additional file 3: Table S2, Additional file 4: Table S3, Additional file 5: Table S4, Additional file 6: Table S5 and Additional file 7: Table S6 for PGSC, and Additional file 8: Table S7, Additional file 9: Table S8, Additional file 10: Table S9, Additional file 11: Table S10, Additional file 12: Table S11 and Additional file 13: Table S12 for ITAG. Although Gene Ontology has its limitations as it is biased towards what is already known, it is still a universal key tool for functional annotation inferring functionality based on sequence identity, domains and structure, and literature studies.

Disparity among pipelines

The output from each pipeline can be represented as a triplet (P, G, GO) where P is the set of all predicted gene-GO term pairs (after ancestor removal), G is the set of genes covered by P, and GO is the set of GO terms covered by P. We measured the pairwise similarity between the triplets obtained from the six pipelines used in the study. Three different ways were used to compare the output of two pipelines A = (P_A, G_A, GO_A) and B = (P_B, G_B, GO_B). First, we measured the overlap between the predictions of the pipelines P_A and P_B. This was done by calculating the ratio between the size of the intersection of P_A and P_B and the size of the union of P_A and P_B. This measure is called the *Jaccard* score [22,23]. Second, we measured the similarity between the covered gene sets G_A and G_B of the pipelines by calculating their Jaccard scores. These two scores are complementary: the first measures the overall similarity between A and B, whereas the second measures the tendency of A and B to cover the same genes. However, these scores ignore the GO structure and thus they are oblivious to the functional similarity among different GO terms. Therefore, we also used a similarity score based on the semantic similarity of GO terms [24]. Given a specific GO type GT (BP or MF), for each gene we measured the semantic similarity between its GO terms in A and its GO terms in B. We then took the average over all genes as the similarity of A and B in GT (see Methods for details). As this score uses the structure of the GO hierarchy, we call it *structure-based.*

An example of the structure-free similarity of the predictions is shown in Figure 2A. The figure shows the pairwise Jaccard score between the PGSC MF predictions of the pipelines. Overall the similarity is low, averaging 0.27. Nevertheless, local patterns can be observed. For example, InterPro2GO, Trinotate HMM, and Phytozome were more similar (average 0.46). Figure 2B shows the Jaccard similarity between the PGSC genes annotated by the different pipelines. The mean similarity was a higher 0.54, which is still quite low. This indicates that different pipelines tend to cover different genes and, even when covering the same genes, they often associate distinct annotations to them. Even when re-computing the structure-free similarity restricted only for the genes shared by each pair of pipelines (considering both MF and BP predictions), the average score was only 0.27.

Figure 2 Comparison of annotations of the PGSC genes by different pipelines. Each similarity matrix shows all pairwise similarities between the pipelines. **(A)** Structure-free Jaccard similarity of the MF predictions of the pipelines. **(B)** Jaccard similarity of the gene sets covered by each pipeline. **(C)** Structure-based similarity between the GO MF predictions of the pipelines. Unlike **(A)**, the calculation here used the GO hierarchy to quantify the similarity of the predictions (see Materials and Methods). **(D)** Structure-based similarity between the GO BP predictions of the pipelines.

The structure-based MF and BP similarity of PGSC genes is summarized in Figure 2C and 2D. Similar matrices on ITAG data are shown in Additional file 1: Figure S1. Again, pipelines tend to be very different, with average similarity of 0.29 in BP and 0.42 in MF. The scores are higher than for the structure-free approach because the structure-based approach assigns higher scores when predictions are different but biologically similar. Also, like in the structure-free scores in Figure 2A, InterPro2GO, Trinotate HMM, and Phytozome formed a cluster both in BP and in MF. Taken together, the discrepancies among pipelines show that pipelines differ in the sets of genes they cover, and the annotation of the same genes in different pipelines can be quite dissimilar.

Ensemble of pipelines

The marked disparity in gene annotation by different pipelines calls for an integration of the different predictions in order to provide a unified potato gene annotation. We developed a simple ensemble algorithm inspired by previous studies [25]. Our algorithm takes as input the predictions of all pipelines and for each gene merges its predictions into a vector of scores denoted as the gene's *combined profile* (Figure 3). Briefly, we first calculate the *pipeline-specific* gene profiles. For a specific pipeline that predicted the pair (G, t), where G is a gene and t is a GO term, the t-th position of the profile is 1 if G is associated with t or at least one of its descendants, and otherwise it is 0 (top right in Figure 3). The combined profile of each gene G is the sum of its pipeline-specific profiles (Figure 3 right). The value in the combined profile of a gene shows how many pipelines agree with each gene-GO term association. Given a threshold k, for each gene we report all GO terms with a combined score $\geq$ k. This process produces a list of GO terms for each gene. We call this variant *Ensemble-k.* Finally, we apply the ancestor removal filter described above. Thus, each value of k produces a different variant of the ensemble algorithm. Figure 3 shows a toy example of Ensemble-1 and 2. For clarity, in the next sections we use the name *annotation method* for both pipelines and variants of the ensemble algorithm. We also tested a more involved supervised ensemble method, which in addition ranks the pipelines by their average F-measure against a gold standard (see below), but this did not improve the results (see Additional file 1: Method S6).

Figure 3 A simple example of the ensemble algorithm. The input (top left) is a set of GO terms, the GO graph, and association between genes and GO terms. The example shows the ensemble process of a single gene G. First, the *pipeline-specific* gene profiles are calculated (top right). A GO term is assigned a value '1' in the profile if G is associated with it or with at least one of its descendants and '0' otherwise. Second, the combined profile of G is the sum of its pipeline-specific profiles. The scores in the combined profile show how many pipelines agree with each of G's GO term association. Given a threshold k, the GO terms with a combined score lower than k are removed to provide a final list of GO terms associated with G (bottom). Each different value of k constitutes a different variant of the algorithm.

We compared the annotation methods in terms of gene coverage and the average *n*umber of GO terms *p*er *g*ene, which we denote as NGPG. Ideally, gene coverage should be as high as possible, while NGPG should be low [26]. The results are shown in Figure 4A and 4B. One can observe marked differences between the different pipelines, and between ITAG and PGSC gene models. For example, based on PGSC data, InterPro2GO and OrthoMCL-UniProt have the highest gene coverage (29,445 and 26,371, respectively), and NGPG score (7 and 7.1, respectively). However, based on ITAG data, OrthoMCL-UniProt's results were similar to those for PGSC, while for InterPro2GO the number of genes dropped under 20,000 and the NGPG score increased to 8.1 (Figure 4B).

Figure 4 Gene coverage and mean number of GO terms per gene (NGPG). For each annotation method (i.e., a pipeline and a variant of the ensemble algorithm) the gene coverage **(A)** and NGPG **(B)** are shown both for PGSC and ITAG gene models.

Figure 4A and 4B also show the gene coverage and the NGPG of the ensemble algorithm. As expected, using either Ensemble-1 or 2 increased the gene coverage compared to the single pipelines using both ITAG and PGSC gene models. For example, based on PGSC the number of covered gene models (including splicing variants) was 41,668 (k = 1) and 29,495 (k = 2). Larger k values led to a sharp decrease in gene coverage, such that even single pipelines covered more genes. Using Ensemble-1, the NGPG score was similar to the highest score obtained by a single pipeline, reaching a score of 6.70 on PGSC data, and 8.15 on ITAG data. Ensemble-2 led to a sharp decrease in NGPG: 4.39 on PGSC, and 4.68 on ITAG.

In summary, our results show that the ensemble algorithm increases the gene coverage considerably without increasing the NGPG score. Ensemble-1 increased gene coverage by more than 5000 genes on both ITAG and PGSC data, while keeping the NGPG score similar to that of the highest single pipelines. Ensemble-2 increased the gene coverage only moderately compared to the single pipelines but the NGPG score declined sharply compared to all pipelines (except Phytozome, but the latter has low gene coverage), hence providing much more focused annotations. In the next sections we demonstrate that the aforementioned improvements were not achieved at the expense of precision.

Validation using the potato gold standard

To evaluate predictions of the different annotation methods we compiled a gold standard of 838 and 724 gene-GO term pairs based on PGSC and ITAG data, respectively, using manual annotation by experts (see Materials and Methods and Additional file 14: Table S13, Additional file 15: Table S14 and Additional file 16: Table S15). The number of genes included in the gold standard (43 with literature references, which are mapped to 116 PGSC

gene ids, see Additional file 14: Table S13), is small, but in an organism such as potato it still contains the majority of genes with experimental evidence. We evaluated the annotation methods by calculating their GO-based precision and recall. Use of the GO structure to calculate scores for gold standard validation has been previously suggested by [27]. The GO-based recall of a gene measures the extent to which its terms according to the gold standard are covered by its predicted GO terms. The GO-based precision of a gene measures the extent its predicted GO terms match the gold standard terms. For each pipeline we calculated the average precision and average recall (over the genes) and report the F-measure, which is the harmonic mean of the precision and the recall [28]. See Materials and Methods for a full description of these calculations.

The results of the validation based on PGSC and ITAG data are illustrated in Figures 5 and Additional file 1: Figure S2, respectively. Figure 5A shows the F-measure for BP GO terms. Ensemble-1 and 2 reached F-measures of 0.8 and 0.77, respectively, while the top performing pipeline was InterPro2GO with only 0.61. Figure 5B shows the F-measure on the MF gold standard. Ensemble-1 and 2 reached F-measures of 0.84 and 0.83, respectively, whereas the top performing pipeline was InterPro2GO with an F-measure of only 0.71. Thus, the results are in agreement with the BP validation: Ensemble-1 and 2 performed best and improved upon the single pipelines. Taken together, our results indicate that Ensemble-1 and 2 provide a significant improvement in comparison to single pipelines.

Figure 5 Validation of annotations based on gold standard. For each annotation method (i.e., a pipeline and a variant of the ensemble algorithm) the F-measure of the gold standard validation is shown on PGSC gene models, see Materials and Methods for a full description of the scores. A score of 1 means perfect agreement between an annotation method and the gold standard. A score close to zero means poor concordance with the gold standard. **(A)** F-measure of the BP annotations. **(B)** F-measure of the MF annotations. The results show that both in BP and MF the ensemble algorithm improves the results considerably when used with k is 1 or 2.

Validation using gene expression data

An obvious disadvantage of any gold standard is that it is limited to experimentally validated genes and subject to the opinion of experts. Consequently, we added an additional validation based on gene co-expression analysis, where we measured the ability of pipelines to predict the same GO-term to highly co-expressed genes. Our co-expression analysis is based on the gene expression of 12,956 genes in 326 expression profiles from over 20 microarray studies. We used the Pearson correlation coefficient to measure co-expression between genes.

We used the gene pairwise co-expression scores to validate predicted GO BP terms. In order to reduce noise, we ignored terms with >500 genes, or with fewer than five genes. Given a set of genes predicted to be associated with the same GO term according to a specific annotation method, we tested if the level of co-expression among its genes is higher than expected by chance (see Materials and Methods for details). Thus, for each term in a specific annotation method we calculated a single p-value. To summarize these values when comparing methods we calculated two scores: (1) the number of GO terms with $p < 0.001$, and (2) the percentage of GO terms with $p < 0.001$ (out of all predicted terms with at least three genes). The former is a measure of coverage of significant GO terms, whereas the latter is a measure of quality of the predicted GO BP terms. Similarly to the gold standard, this analysis simply aimed to

compare pipelines. Future work can use similar approaches to select highly co-expressed GO terms from different pipelines for subsequent analyses.

The results of the gene co-expression validation based on PGSC data are shown in Figure 6. See Additional file 1: Figure S3 for results of ITAG. The top two pipelines in terms of the number of significant GO terms were InterPro2GO (n = 411) and BLAST2GO (n = 345). The top two pipelines in terms of the percentage of significant GO terms were InterPro2GO (35%) and Phytozome (30%). The ensemble algorithm markedly improved the number of significant GO terms: Ensemble-1 achieved 718, and Ensemble-2 achieved 650. However, the ensemble methods did not improve upon the single pipelines in terms of the percentage of significant GO terms: Ensemble-1 and 2 achieved 22% and 27%, respectively. Nevertheless, the score of Ensemble-2 was better than all pipelines except for InterPro2GO and Phytozome. Thus, the ensemble approach provided an improvement of at least 1.5-fold in the number of significant GO terms, at the expense of a drop of 8% in the percentage of significant GO terms compared to the best pipeline. Note that the co-expression and the GO analyses are complementary, since the gold standard genes do not manifest unusually high co-expression (see Additional file 1: Methods S7).

Figure 6 Validation of annotations based on co-expression. Given a set of PGSC genes linked to a biological process by a specific annotation method (i.e., the pipelines or a variant of the ensemble algorithm) the average co-expression of the genes was compared to that of random gene sets. For each annotation method the number of GO terms with $p < 0.001$ **(A)**, and the percentage of GO terms with $p < 0.001$ **(B)** are shown. Ensemble-2 has a lower percentage of significant GO terms compared to the best single pipeline (BioMart), but it has >1.5 fold more significant GO terms.

Merging the different merits using a rank-based comparison

Our analysis shows that the ensemble approach is beneficial according to most criteria. However, since we used multiple ways to score the methods, it is hard to decide which k value is best and which pipelines are better. To provide a clear unified view we used a non-parametric rank-based consolidation of the different scores [29]. In the previous sections, for each annotation method we calculated two F-measure scores in the gold standard analysis and two scores in the gene co-expression analysis. In addition, we compared the annotation methods by their gene coverage and NGPG. Note that when ranking methods by their NGPG score, lower scores are better. In contrast, when ranking methods by their gene coverage, higher scores are better. To consolidate these different scores, we used six rankings: by gene coverage and the NGPG score, by the two F-measures of the gold standard validation and by the two scores of the gene co-expression validation. We reversed the scores when necessary so that rank 1 was the best for each method, averaged the rankings and ranked the methods by their average rank. We call this score *rank-merge.*

Figure 7 displays the rank-merge results on PGSC (A) and ITAG (B) data. The top three methods are colored black. In both cases the top method was Ensemble-2, with an average rank of 1.66 in PGSC and 1.16 in ITAG. Among the different pipelines evaluated, Phytozome obtained the top score for PGSC data with an average rank of 3.66 while BLAST2GO obtained top score for ITAG data with an average rank of 3.50. Note that Ensemble-1, 2, and 3 were ranked consistently high in both tests. See also Additional file 17: Table S16 for PGSC and Additional file 18: Table S17 for ITAG. Thus, we conclude that the ensemble approach, especially with k = 2, is beneficial and can assist in integration of different gene

function prediction pipelines. See Additional file 1: Method S5 for details on reproducing the results and applying the pipeline to new genomes.

Figure 7 Rank-based consolidation of the different figures of merit. A non-parametric rank-based consolidation of the different scores of the annotation methods was used for a unified comparison. First, six rankings were calculated: by gene coverage, by NGPG, by the two F-measures of the gold standard validation, and by the two gene co-expression validations scores (i.e., the number and the percent of significant GO terms). To merge the different rankings we used the average rank. The results show that both for PGSC (panel **A**) and for ITAG (panel **B**), Ensemble-2 has the best average rank.

Note that using k = 1 is equivalent to assigning to each gene all its annotations from all pipelines (and their ancestors) and then performing ancestor removal. While this method is the most intuitive ensemble, we show here that varying the k parameter can improve the annotation of genomes.

A seemingly natural test case for our approach is to evaluate it in predicting function of Arabidopsis genes. However, it is not clear how this can be done in a rigorous and unbiased manner. Tools for functional annotation of genes in newly sequenced plants are heavily dependent on sequence similarity to genes in model species such as Arabidopsis. In order to test such tools in predicting Arabidopsis gene functions, one has to exclude all the annotations directly – or indirectly – derived from Arabidopsis. Doing so would entail tracing indirect annotation sources, which often are not recorded in the pipelines. Instead, we used the newly sequenced potato genome along with experimentally verified gene functions and rich gene expression data in our evaluation.

Conclusion

For recently sequenced, non-model organisms, automatic functional annotation of genes, which also mainly relies on sequence-based prediction, often suffers from low gene coverage and poor specificity. We confirmed that this is the case for the potato genome by analyzing six state of the art annotation pipelines.

We observed that the predictions of different pipelines for functional annotations of genes are markedly different, in spite of the fact that all pipelines are based on sequence analysis. We showed that combining predictions from several pipelines increases both the coverage and the accuracy of gene ontology predictions. The simple ensemble approach used here could be applied easily to other sequenced genomes and improve functional annotation by taking advantage of different GO prediction tools. However, a comparison of the consistency among pipelines is not enough when the goal is to either select the best pipeline or to integrate the different predictions. The pipelines should also be evaluated based on the precision of their predictions. The most intuitive way is to compare the pipelines to a set of known annotations. However, in newly sequenced organisms such as potato, known annotations are scarce in the main public databases. To overcome this, we compiled a gold standard of experimentally-validated gene-GO associations. Although this gold standard is relatively small, we have found it useful for comparing pipelines. Furthermore, to overcome the limited number of genes in the gold standard, we used a second validation method based on gene co-expression testing the ability of pipelines to predict co-expression of genes associated to the same GO-term.

Finally, we introduced an integrated annotation of the different pipelines that outperformed the single pipelines both in the gold standard validation and in the co-expression validation. Our integration approach depends on selecting a parameter k that corresponds to the stringency by which we filter out gene-GO associations. That is, when associating a gene to a GO term, at least k pipelines must agree with this association. Thus, we have implicitly assumed that each of the pipelines we used has meaningful predictions. Moreover, all pipelines have the same weight in the integration process. Future analyses can seek methods that give more weight to better pipelines, or add an initial step that filters out pipelines of exceptionally low prediction quality. The new functional annotations of the potato genome as well as for the probes on the JHI Solanum tuberosum microarray are available with this paper (Additional file 17: Table S16, Additional file 18: Table S17 and Additional file 19: Table S18). We also provide tools as open source R code for implementing the methodology with additional pipelines and for other sequenced organisms.

Methods

Executing the functional annotation pipelines

We defined a *pipeline* as an automated process that predicts association between genes and functions. The input to a pipeline can be DNA sequence, protein sequence, or protein domains. The output of a pipeline is a set of pairs in the form of (gene ID, GO term ID). We ran all pipelines for the ITAG (potato.Sotub.proteins.itag.v1.fasta) and PGSC (PGSC_DM_v3.4_pep_representative.fasta) gene models separately, using default settings as follows:

The OrthoMCL-UniProt pipeline

We ran the OrthoMCL [16] pipeline in two steps:

1. Building the clusters of homologs: We retrieved from Phytozome (v9.1) 16 plant proteomes, covering the whole plant phylogeny. Together with the proteomes predicted from the potato PGSC and ITAG gene models, we aligned the proteomes against each other using blastp [30]; (parameters: −e-value: 1e-05 -outfmt 6). We then used OrthoMCL v2 to build clusters of homologous proteins.
2. Annotating GO terms: To annotate every protein sequence of the 18 complete plant proteomes with GO terms we ran a blast search against the entire UniProt database (version 2013_08) [31] with an e-value cut-off of 1e-10. For every protein sequence we kept a ranked list of the ten best hits (i.e. hits with the lowest e-value). We associated the first hit in the list that had GO annotation in UniProt. An OrthoMCL cluster then inherits all GO terms associated with its proteins, and each PGSC (and ITAG) protein inherits the GO terms of its cluster.

For complete protocol details refer to the Additional file 1: Method S2.

The BLAST2GO pipeline

Using the BLAST2GO interface [17], we blasted the PGSC and ITAG protein sequences against the NCBI NR database (blastp parameters: −e-value: 1e-05 -max_target_seqs 20 -outfmt 5). We then loaded the blastp output files into Blast2GO (v2.6.6, with default

parameters) and assigned GO terms to the PGSC and ITAG sequences according to its output.

The trinotate pipeline

In the Trinotate suite [15] we used default settings for the NCBI-BLAST (SwissProt), HMMER [32], and Pfam [33]. For complete protocol details refer to the Additional file 1: Method S3.

The phytozome pipeline

We downloaded the potato annotation from Phytozome v9.1 [http://www.phytozome.net/potato.php; 18] (http://www.phytozome.net/potato.php). The gene annotation is *Solanum tuberosum* Group Phureja DM1-3 516R44 (CIP801092) *Genome Annotation v3.4 mapped to* pseudomolecule *sequence* (PGSC_DM_v3_2.1.10_pseudomolecules.fa).

InterPro2GO data from BioMart

We downloaded the potato data from (http://central.biomart.org/). GO terms in BioMart are derived from the semi-automated InterPro2GO [19].

Formatting pipelines

In order to compare pipelines, we mapped their predicted annotation to a set of common Gene Ontology (GO) terms. If the original pipeline output was not in GO term IDs it was mapped to GO IDs using the gene ontology consortium mapping files for GO terms. We applied this procedure to the pipelines Trinotate, InterPro2GO, BLAST2GO, Phytozome, and in mapping of orthologous and paralogous gene families in 18 sequenced plant species by OrthoMCL clustering.

Composing the potato ‘gold standard’

A ‘gold standard’ set of potato genes was constructed based on literature evidence from functional gene studies by wet-lab experiments in potato reported in PlantCyc [http://pmn.plantcyc.org/PLANT/organism-summary] and a few additional studies on potato [34-37]. In total a list of 43 potato genes/proteins was created (Additional file 14: Table S13). These protein names were searched for their corresponding identifiers published by the PGSC [3], resulting in 116 unique PGSC gene identifiers.

The aforementioned list of genes matched 1658 GO terms from all six tested pipelines. Each gene-GO term association was then manually scored with the help of literature searches in an unbiased manner, where the experts assigning scores to GO-associations did not know from which pipeline the annotation originated. Every GO term in the set was scored as ‘1’ (low evidence), ‘2’ (neutral or unknown) and ‘3’ (high evidence). In the final analysis only association scores of 3 were used for the gold standard, producing 838 annotations (Additional file 15: Table S14). To perform analyses on both gene models, PGSC genes were mapped to ITAG genes using BLAST (identity >95%, length >100 amino acids). This produced an ITAG gold standard with 724 annotations (Additional file 16: Table S15).

Comparing pipelines and gold-standard evaluation

Mathematical notations

In the Results section we sketched the calculations for comparing pipelines and evaluating pipelines against the gold standard. Here, we provide a full description of these calculations. For this purpose we start here with more detailed definitions.

Let G be the set of all genes in the tested organism and let T be the set of all GO terms. The output of a pipeline *P* is a set of pairs $P = \{p_1, \ldots, p_k\}$ where each *annotation pair* $p_i = (g_i,t_i)$ is an association between a gene g_i (in G) and a GO term t_i (in GO). Let BP(P) be the subset of P resulted from taking all pairs in P in which the term t is a biological process. Similarly, define MF(P) for molecular function, and CC(P) for cellular component. Below we define functions of pipelines. Note that by definition each of BP(P), MF(P), and CC(P) is a set of pairs. Thus, in the definitions below P is either the original output of a pipeline or the result of applying BP, MF, or CC on it.

We define *Genes(P)* as the set of genes covered by P and *Terms(P)* as the set of GO terms covered by *P*. We define *Genes(P,t)* as the set of genes associated with a GO term *t* according to *P*, and *Terms(P,g)* as the set of GO terms associated with a gene *g* according to *P*. Finally, we denote $Sem(t_i,t_j)$ as the semantic similarity between two GO terms t_i and t_j. Semantic similarity here is a measure that quantifies the closeness of two terms in the GO graph. There are several ways to calculate semantic similarity among GO terms. In this study we used Wang's method [6,24].

Jaccard coefficient between two pipelines

The Jaccard coefficient is a generic measure of similarity between two sets. It is defined as the ratio between the size of the intersection of the sets and the size of the union of the sets. For example, given two pipelines P_1 and P_2, denote $intersect(P_1,P_2)$ as the set of annotation pairs that are both in P_1 and in P_2, and let $union(P_1,P_2)$ be the set of annotation pairs that are either in P_1 or in P_2. The Jaccard coefficient $J_{pipeline}(P_1,P_2)$ is the ratio between the number of annotation pairs in $intersect(P_1,P_2)$ and the number of annotation pairs in $union(P_1,P_2)$. In addition, we calculate the Jaccard coefficient $J_{Genes}(P_1,P_2)$ between the gene sets $Genes(P_1)$ and $Genes(P_2)$ to measure the tendency of two pipelines to annotate the same genes.

Structure-based similarity between two pipelines

The Jaccard measure above is oblivious to the functional similarity among GO terms. Thus, we used semantic similarity as a means to define a structure-based similarity between two pipelines P_1 and P_2. We start by defining the similarity between the set of annotations of a single gene. Given a gene g our goal is to measure the semantic similarity between $Terms(P_1,g)$ and $Terms(P_2,g)$. As a first step we define the similarity between a single GO term *t* and a set of GO terms *T'* as:

$$Sim'(t,T') = \max_{t' \in T'} Sem(t,t')$$

This score is high only if T' contains *t* or similar GO terms. Next, we use this score to calculate the similarity between *Terms(P_1,g)* and *Terms(P_2,g)* using the running-max-average [6]:

$$rmaxa(P_1,P_2,g)=\frac{\sum_{t_i\in Terms\ P_1,g} Sim'(t_i,Terms(P_2,g))+\sum_{t_j\in Terms\ P_2,g} Sim'(t_j,Terms\ P_1,g\ )}{|Terms\ P_2,g\ |+|Terms\ P_1,g\ |}$$

This score will be high only if *Terms(P_1,g)* covers the biological functionalities of *Terms(P_2,g)* and vice versa. Finally, the overall similarity between P_1 and P_2 is the average gene-wise similarity:

$$Sim\ P_1,P_2\ =\frac{\sum_{g\in Genes\ P\ \cup Genes\ P_2} rmaxa\ P_1,P_2,g}{|Genes\ P\ \cup Genes\ P_2\ |}$$

GO-based precision and recall

The calculations above measure similarity among pipelines. Here we define a way to measure the precision and recall of a pipeline *P* compared to a gold standard *GS*. Similarly to *P, GS* is a set of annotation pairs *{gs_1, ..., gs_k}* where each pair $gs_i = (g_i,t_i)$ is an association between a gene g_i (in *G*) and a GO term t_i (in *T*). We first define the precision of a single gene *g*. The GO-based *precision* of pipeline *P* for gene *g* measures the extent by which *Terms(P,g)* is covered by *Terms(GS,g)*:

$$prec\ P,GS,g\ =\frac{\sum_{t_i\in Terms\ P,g} Sim'\ t_i,Terms\ GS,g}{|Terms\ P,g\ |}$$

The *precision* of *P* is defined as the average precision of the genes in *Genes(G)*:

$$prec\ P,GS\ =\frac{\sum_{g\in Genes\ GS} prec\ P,GS,g}{|Genes\ GS\ |}$$

The *GO-based recall* of pipeline *P* for gene *g* measures the extent by which *Terms(P,g)* covers *Terms(GS,g)*:

$$recall\ P,GS,g\ =\frac{\sum_{t_i\in Terms\ GS,g} Sim'\ t_i,Terms\ P,g}{|Terms\ GS,g\ |}$$

The *recall of P* is defined as the average recall of the genes in Genes (G):

$$recall\ P,GS\ =\frac{\sum_{g\in Genes\ GS} recall\ P,GS,g}{|Genes\ GS\ |}$$

Microarray data preprocessing and normalization

We have integrated potato gene expression data from over 20 studies based on the Agilent JHI Solanum tuberosum 60 k v1 microarray (ArrayExpress ID: E-MTAB-1655) processed at the James Hutton Institute using standard Agilent recommended methodologies [38]. The studies included 326 conditions derived from the following treatments: moderate heat-stress [38], short- and long-day growth regimes [39], bruising, phosphorous growth regimes, acidity, *Phytopthora infestans* infection [40], and phosphite [41], BABA [14], ABA, brassinosteroid, SA treatment. Varietal differences and tuber, stem and leaf tissues were included.

We applied quantile normalization using the Limma package [42] and subtracted the background intensity from the foreground intensity for each spot using the 'normexp' method [43]. Our normalized expression matrix contained 52,998 probes. In order to reduce statistical noise and to focus on genes with high variation we removed both probes with consistently low expression values across the samples and probes with low variance. Thresholds for probe removal were adjusted as proposed in [44], see Additional file 1: Method S4 for more details. 14,000 probes remained in the data. These probes were mapped to 12,956 genes, approximately the same amount of genes analyzed in Tzfadia, et al. [44].

Evaluating co-expression of predicted GO processes

Given a gene set U associated with a specific GO term, and a gene expression matrix X with genes as rows, we first calculate the Pearson correlation between all pair of genes in U using their expression profiles in X. To evaluate if the correlations in U tend to be higher than expected by chance we sample random gene pairs in X and calculate their correlation to get a distribution of random correlation scores. We used the Kolmogorov-Smirnov test to compare the real correlations scores of U to the random correlation scores. To improve robustness, we repeated this process 50 times for each gene set U and used the mean p-value over all repeats.

Abbreviations

GO, gene ontology; PGSC, potato genome sequencing consortium; ITAG, international tomato annotation group; CC, cellular component; MF, molecular function; BP, biological process; NGPG, number of GO terms per gene.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

OT and E Alexandersson designed the study and DA, IF, AD, TG, SKS, EPW, OT and E Alexandersson performed the research. DA, IF and AD contributed new analytic computational tools. PH, E Andreasson and SKS provided the gene expression data sets. DA, IF, AD, TG, SKS, EPW OT and E Alexandersson analyzed the data. DA, IF, RS, OT and E Alexandersson wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank SURF-sara (https://www.surfsara.nl/) for hosting the hack-a-thon sessions and providing high performance computing services. This work is part of the Allbio initiative and was partially supported by grant number EU FP7; 289452; KBBE.2011.3.6-02). David Amar was supported in part by fellowships from the Azrieli foundation, and the Edmond J. Safra center for Bioinformatics at Tel Aviv University. Erik Alexandersson and Itziar Frades were supported by Crafoord grant (20120533) and the Swedish Foundation for Strategic Research (RB608-0006) and Estelle Proux-Wera by PlantLink. Agnieszka Danek was supported by POIG.02.03.01-24-099/13 grant: "GeCONiI - Upper Silesian Center for Computational Science and Engineering". We thank the MapMan team for assistance. We would also like to thank Ashfaq Ali, Kate Dreher and Paul Kersey for helpful discussions and input, and Efrat Weithorn for manuscript edit help.

References

1. Birch PR, Bryan G, Fenton B, Gilroy EM, Hein I, Jones JT, Prashar A, Taylor MA, Torrance L, Toth IK: **Crops that feed the world 8: Potato: are the trends of increased global production sustainable?** *Food Security* 2012, **4**(4)**:**477–508.

2. Haverkort A, Boonekamp P, Hutten R, Jacobsen E, Lotz L, Kessel G, Visser R, Van der Vossen E: **Societal costs of late blight in potato and prospects of durable resistance through cisgenic modification.** *Potato Res* 2008, **51**(1)**:**47–57.

3. Potato Genome Sequencing Consortium: **Genome sequence and analysis of the tuber crop potato.** *Nature* 2011, **475**(7355)**:**189–195.

4. Zouine M, Latché A, Rousseau C, Regad F, Pech J-C, Philippot M, Bouzayen M, Delalande C, Frasse P, Schiex T: **The tomato genome sequence provides insights into fleshy fruit evolution.** *Nature* 2012, **485:**635–641.

5. 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**(1)**:**25–29.

6. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF: **A new method to measure the semantic similarity of GO terms.** *Bioinformatics* 2007, **23**(10)**:**1274–1281.

7. Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, Tanay A, Sharan R, Shiloh Y, Shamir R: **Expander: from expression microarrays to networks and functions.** *Nat Protoc* 2010, **5**(2)**:**303–322.

8. Ramsak Z, Baebler S, Rotter A, Korbar M, Mozetic I, Usadel B, Gruden K: **GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology.** *Nucleic Acids Res* 2013, **42:**D1167–1175.

9. Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, Tohge T, Fernie AR, Stitt M, Usadel B: **Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data.** *Plant Cell Environ* 2014, **37**(5)**:**1250–1258.

10. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S: **AmiGO: online access to ontology and annotation data.** *Bioinformatics* 2009, **25**(2)**:**288–289.

11. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Res* 2012, **40**(Database issue)**:**D1202–D1210.

12. Promponas VJ, Ouzounis CA, Iliopoulos I: **Experimental evidence validating the computational inference of functional associations from gene fusion events: a critical survey.** *Brief Bioinform* 2012, **15**(3)**:**443–454.

13. Jaramillo-Garzón JA, Gallardo-Chacón JJ, Castellanos-Domínguez CG, Perera-Lluna A: **Predictability of gene ontology slim-terms from primary structure information in Embryophyta plant proteins.** *BMC Bioinformatics* 2013, **14**(1)**:**68.

14. Bengtsson T, Weighill D, Proux-Wera E, Levander F, Resjo S, Burra DD, Moushib LI, Hedley PE, Liljeroth E, Jacobson D, Alexandersson E, Andreasson E: **Proteomics and transcriptomics of the BABA-induced resistance response in potato using a novel functional annotation approach.** *BMC Genomics* 2014, **15**(1)**:**315.

15. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29**(7)**:**644–652.

16. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**(9)**:**2178–2189.

17. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**(18)**:**3674–3676.

18. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N: **Phytozome: a comparative platform for green plant genomics.** *Nucleic Acids Res* 2012, **40**(D1)**:**D1178–D1186.

19. Kasprzyk A: **BioMart: driving a paradigm change in biological data management.** *Database (Oxford)* 2011, **2011:**bar049.

20. Zhao K, Bartley LE: **Comparative genomic analysis of the R2R3 MYB secondary cell wall regulators of Arabidopsis, poplar, rice, maize, and switchgrass.** *BMC Plant Biol* 2014, **14**(1)**:**135.

21. Kim HA, Lim CJ, Kim S, Choe JK, Jo S-H, Baek N, Kwon S-Y: **High-throughput sequencing and De Novo Assembly of Brassica oleracea var. Capitata L. for transcriptome analysis.** *PLoS One* 2014, **9**(3)**:**e92087.

22. Jaccard P: *Etude comparative de la distribution florale dans une portion des Alpes et du Jura: Impr.* Corbaz; 1901.

23. Jaccard P: **The distribution of the flora in the alpine zone. 1.** *New Phytol* 1912, **11**(2)**:**37–50.

24. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.** *Bioinformatics* 2010, **26**(7)**:**976–978.

25. Khatri P, Done B, Rao A, Done A, Draghici S: **A semantic analysis of the annotations of the human genome.** *Bioinformatics* 2005, **21**(16)**:**3416–3421.

26. Klie S, Nikoloski Z: **The choice between mapman and gene ontology for automated gene function prediction in plant science.** *Front Genet* 2012, **3:**115.

27. Defoin-Platel M, Hindle M, Lysenko A, Powers S, Habash D, Rawlings C, Saqi M: **AIGO: Towards a unified framework for the Analysis and the Inter-comparison of GO functional annotations.** *BMC Bioinformatics* 2011, **12**(1)**:**431.

28. Powers D: **Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation.** *J Mach Learn Technol* 2011, **2**(1)**:**37–63.

29. Datta S, Pihur V: **An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data.** *BMC Bioinformatics* 2010, **11:**427.

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3)**:**403–410.

31. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data.** *Database (Oxford)* 2011, **2011:**bar009.

32. Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res* 2011, **39**(suppl 2)**:**W29–W37.

33. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40**(D1)**:**D290–D301.

34. Pasare SA, Ducreux LJ, Morris WL, Campbell R, Sharma SK, Roumeliotis E, Kohlen W, van der Krol S, Bramley PM, Roberts AG, Fraser PD, Taylor MA: **The role of the potato (Solanum tuberosum) CCD8 gene in stolon and tuber development.** *New Phytol* 2013, **198**(4)**:**1108–1120.

35. Sharma SK, Millam S, Hein I, Bryan GJ: **Cloning and molecular characterisation of a potato SERK gene transcriptionally induced during initiation of somatic embryogenesis.** *Planta* 2008, **228**(2)**:**319–330.

36. Navarro C, Abelenda JA, Cruz-Oro E, Cuellar CA, Tamaki S, Silva J, Shimamoto K, Prat S: **Control of flowering and storage organ formation in potato by FLOWERING LOCUS T.** *Nature* 2011, **478**(7367)**:**119–122.

37. Kloosterman B, Abelenda JA, Gomez Mdel M, Oortwijn M, de Boer JM, Kowitwanich K, Horvath BM, van Eck HJ, Smaczniak C, Prat S, Visser RG, Bachem CW: **Naturally occurring allele diversity allows potato cultivation in northern latitudes.** *Nature* 2013, **495**(7440)**:**246–250.

38. Hancock RD, Morris WL, Ducreux LJ, Morris JA, Usman M, Verrall SR, Fuller J, Simpson CG, Zhang R, Hedley PE, Taylor MA: **Physiological, biochemical and molecular responses of the potato (Solanum tuberosum L.) plant to moderately elevated temperature.** *Plant Cell Environ* 2014, **37**(2)**:**439–450.

39. Morris WL, Hancock RD, Ducreux LJM, Morris JA, Usman M, Verrall SR, Sharma SK, Bryan G, Mcnicol JW, Hedley PE: **Day length dependent restructuring of the leaf transcriptome and metabolome in potato genotypes with contrasting tuberization phenotypes.** *Plant Cell Environ* 2014, **37**(6)**:**1351–1363.

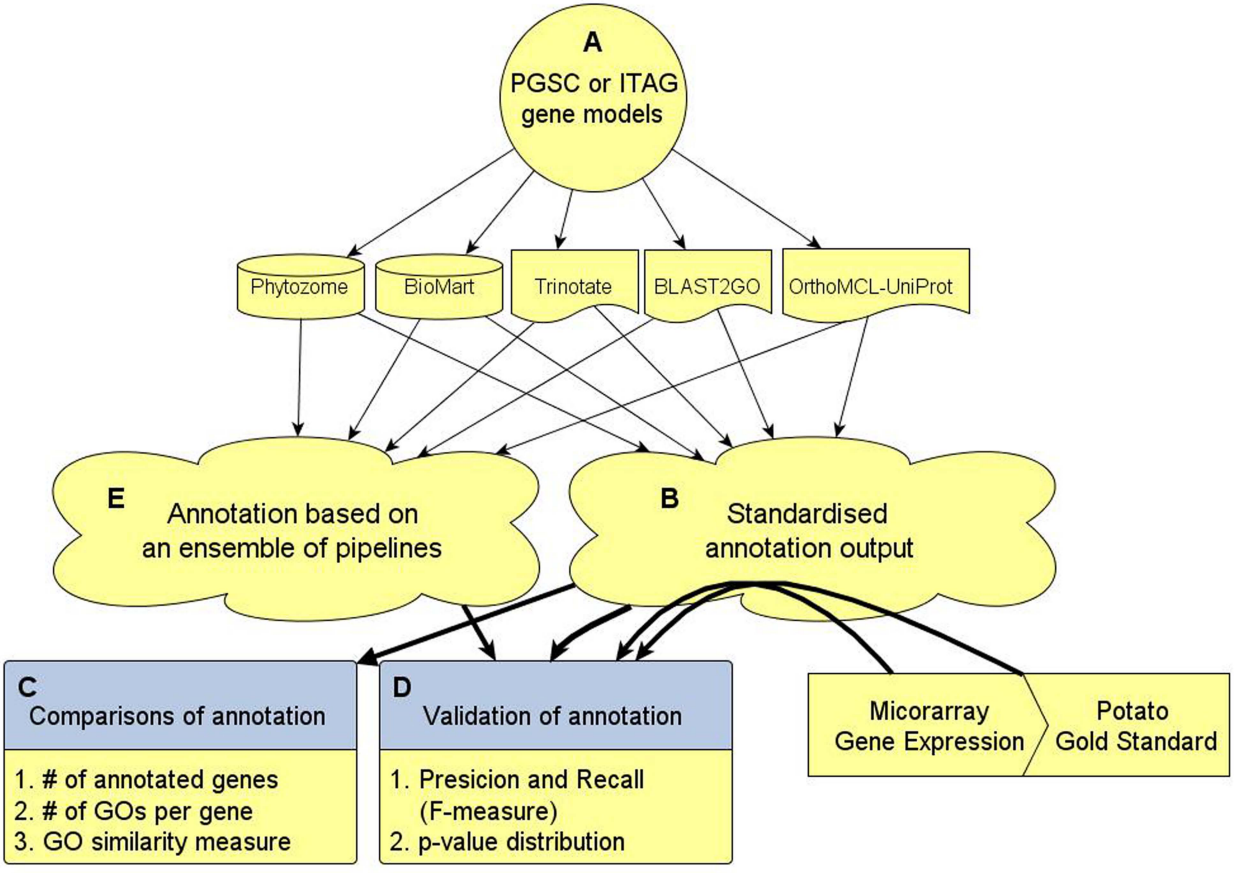
40. Ali A, Alexandersson E, Sandin M, Resjö S, Lenman M, Hedley P, Levander F, Andreasson E: **Quantitative proteomics and transcriptomics of potato in response to Phytophthora infestans in compatible and incompatible interactions.** *BMC Genomics* 2014, **15**(1)**:**497.

41. Burra DD, Berkowitz O, Hedley PE, Morris J, Resjö S, Levander F, Liljeroth E, Andreasson E, Alexandersson E: **Phosphite-induced changes of the transcriptome and secretome in Solanum tuberosum leading to resistance against Phytophthora infestans.** *BMC Plant Biol* 2014, **14**(1)**:**254.

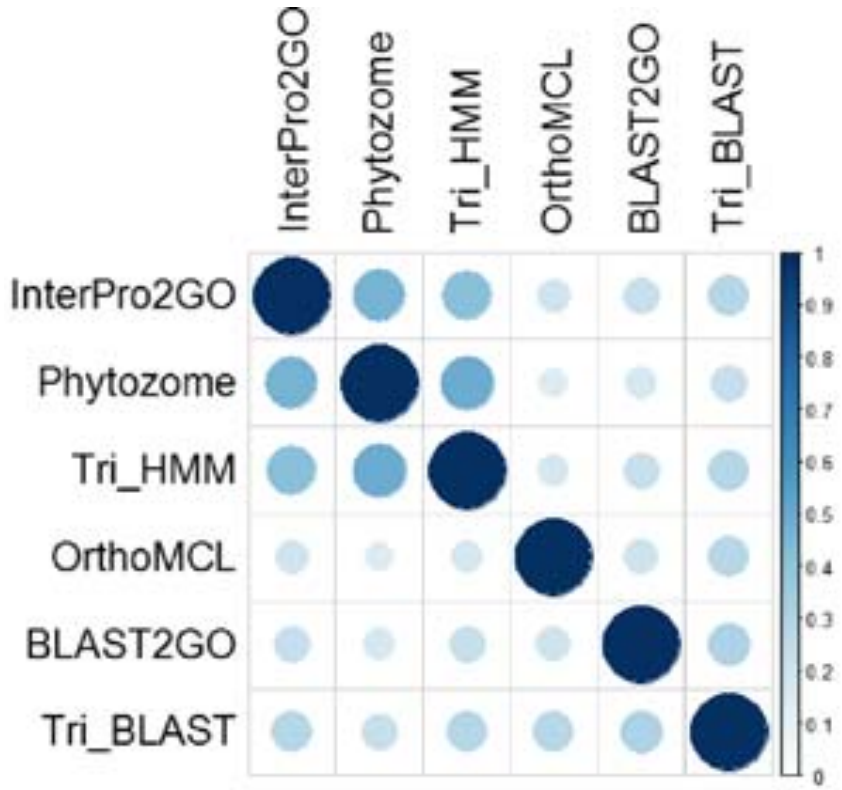
42. Smyth GK: L: **Linear Models for Microarray Data**. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Edited by Gentleman R, Carey V, Dudoit S, R Irizarry WH. New York: Springer; 2005:397–420.

43. Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK: **A comparison of background correction methods for two-colour microarrays.** *Bioinformatics* 2007, **23**(20)**:**2700–2707.

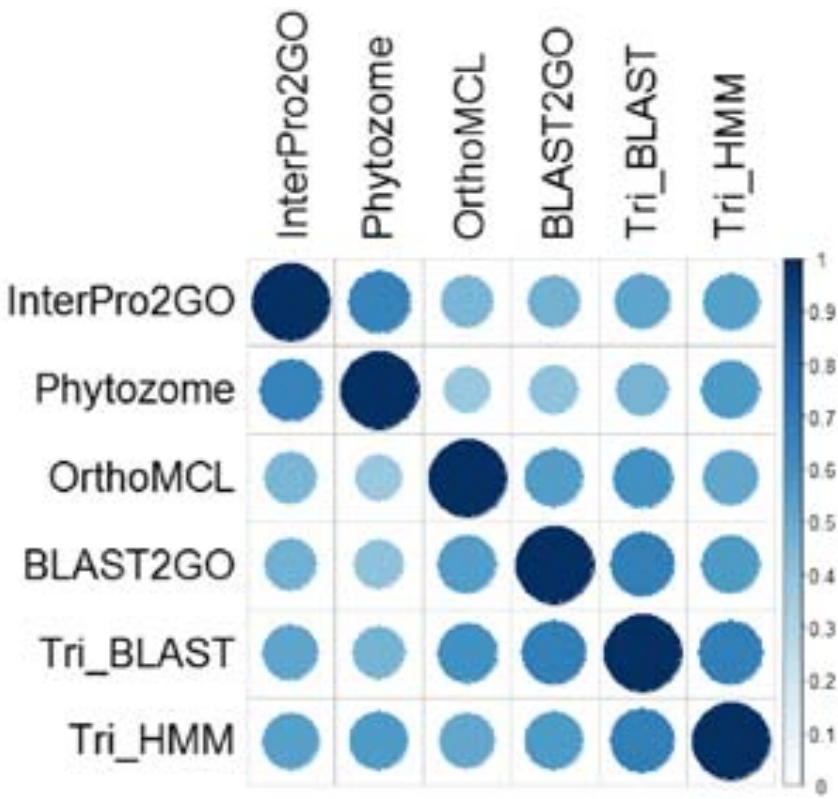
44. Tzfadia O, Amar D, Bradbury LM, Wurtzel ET, Shamir R: **The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways.** *Plant cell* 2012, **24**(11)**:**4389–4406.


A
PGSC or ITAG gene models
Phytozome
BioMart
Trinotate
BLAST2GO
OrthoMCL-UniProt
E
Annotation based on an ensemble of pipelines
B
Standardised annotation output
C
Comparisons of annotation
1. # of annotated genes
2. # of GOs per gene
3. GO similarity measure
D
Validation of annotation
1. Presicion and Recall (F-measure)
2. p-value distribution
Micorarray Gene Expression
Potato Gold Standard

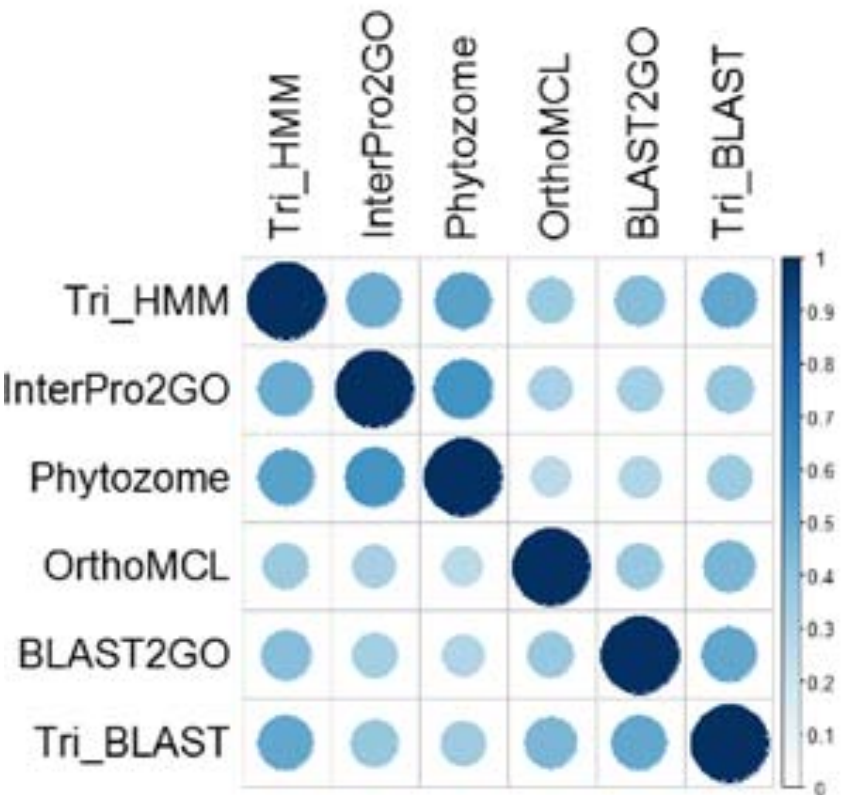
A) Structure-free similarity: MF



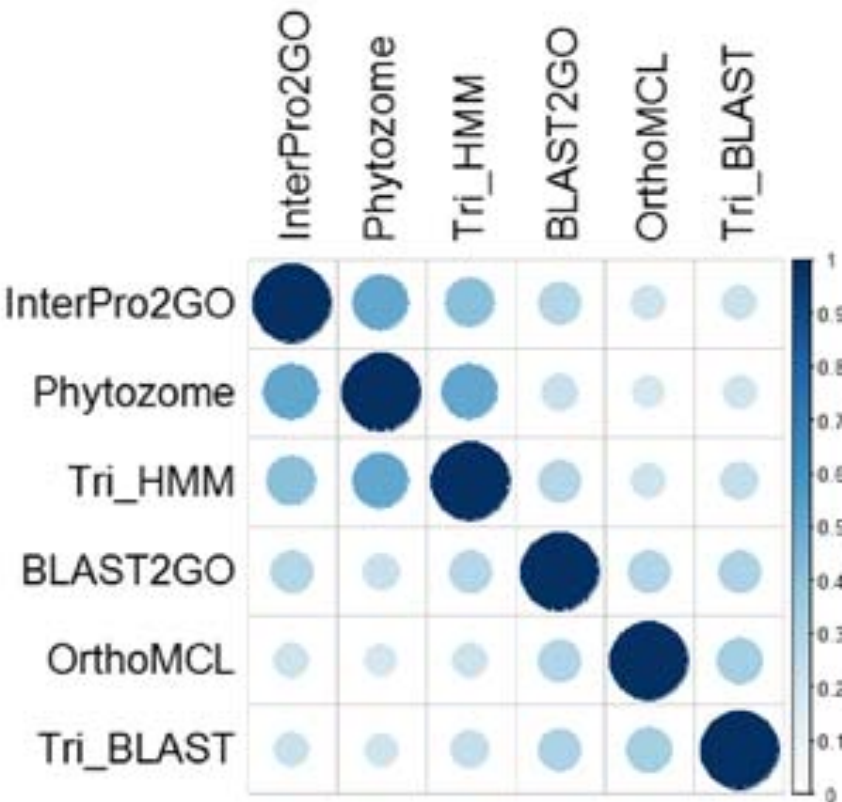
B) Similarity of gene sets

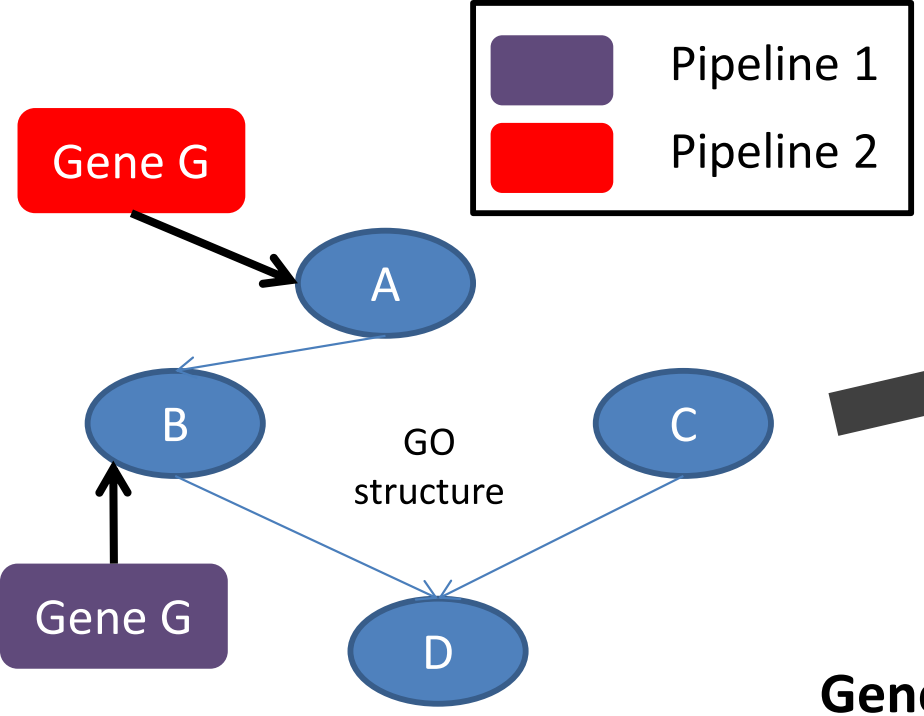


C) Structure–based similarity: MF



C) Structure–based similarity: BP





Gene G: profile in pipeline 1

A	B	C	D
1	1	0	0

Gene G: profile in pipeline 2

A	B	C	D
1	0	0	0

Gene G: combined profile

A	B	C	D
2	1	0	0

K=1

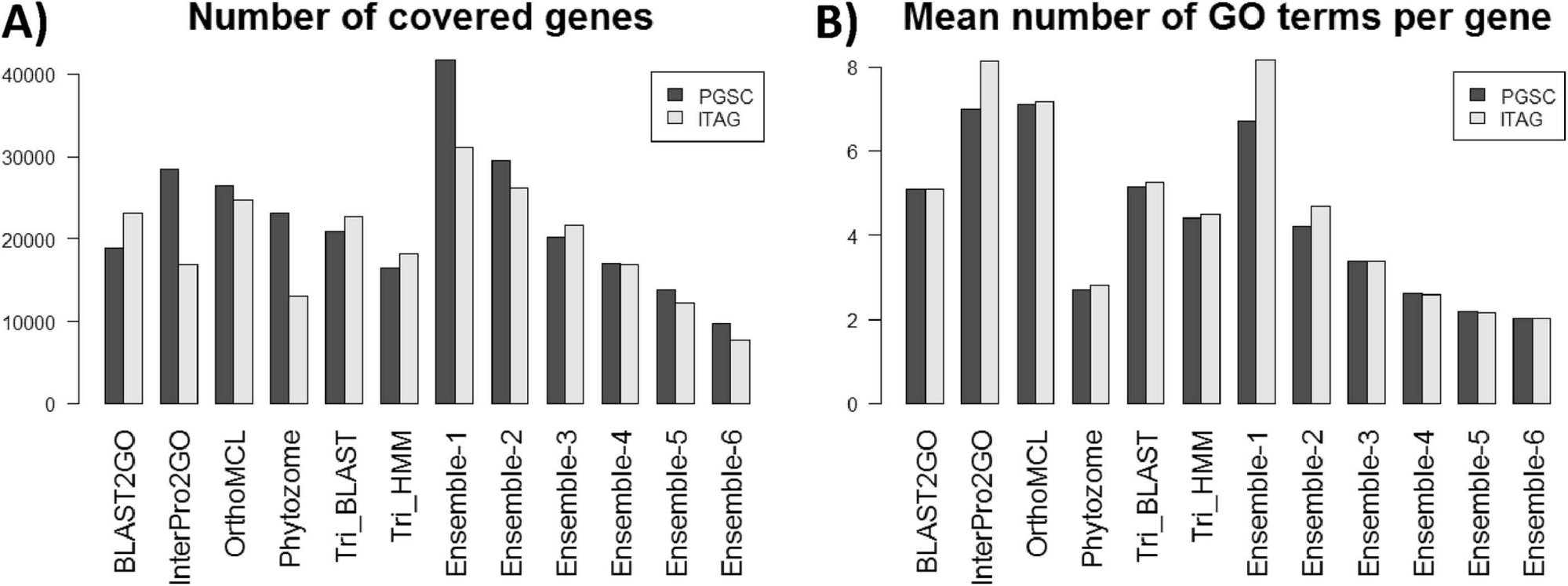
K=2

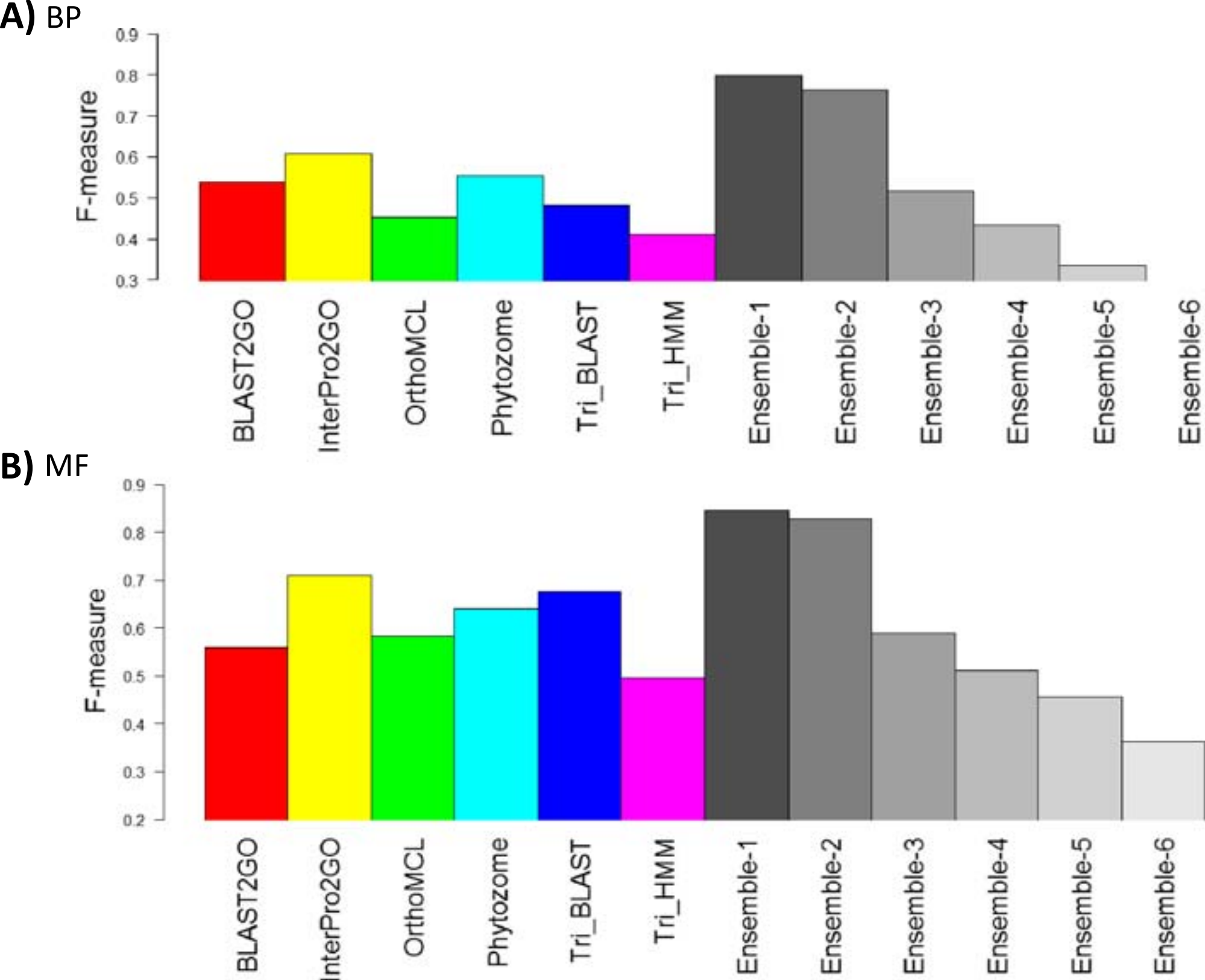
A	**B**	C	D
2	**1**	0	0

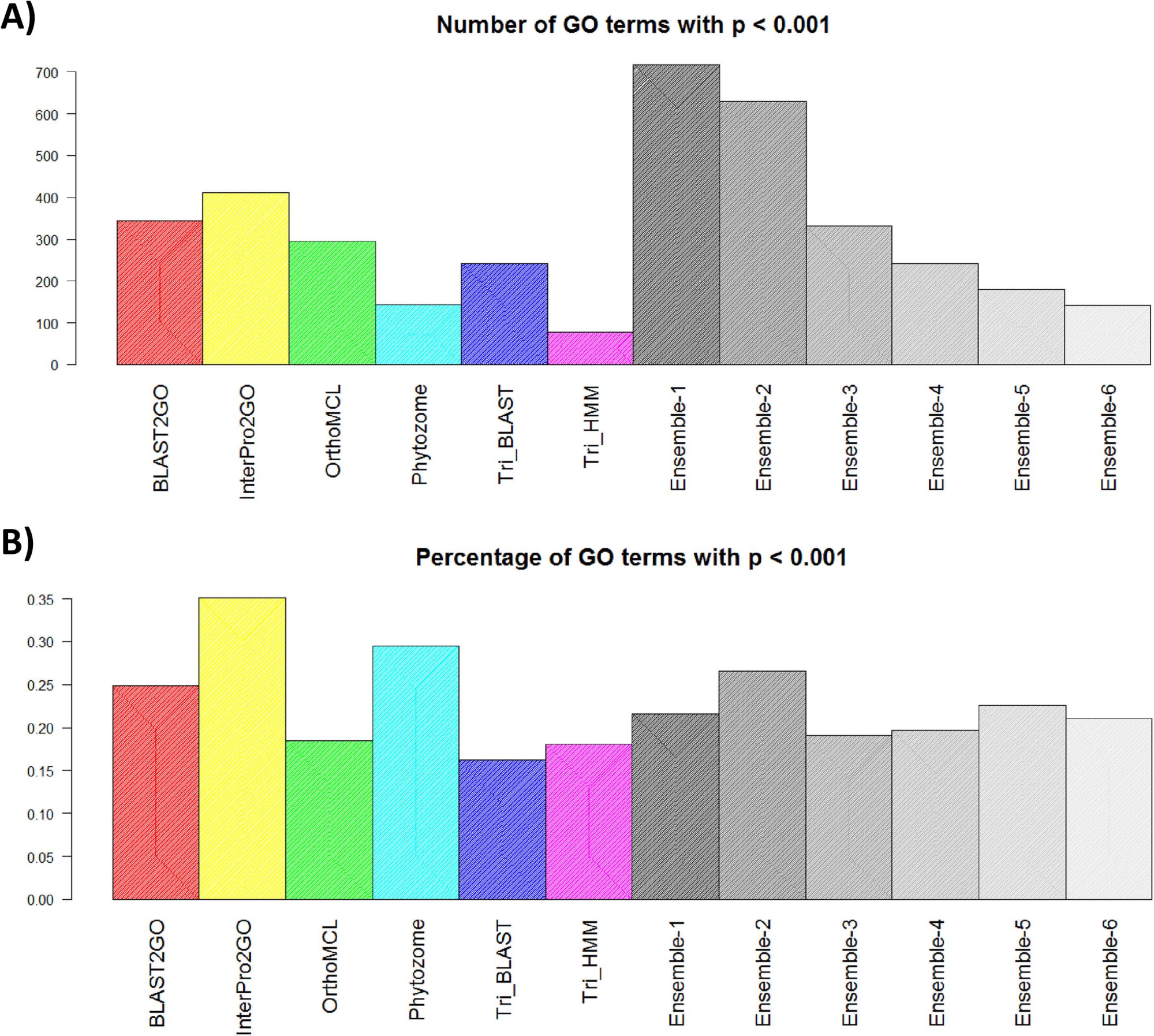
G's most specific GO terms

A	B	C	D
2	0	0	0

Conservative: all methods in agreement


A)
Number of covered genes
40000
30000
20000
10000
0
PGSC
ITAG
BLAST2GO
InterPro2GO
OrthoMCL
Phytozome
Tri_BLAST
Tri_HMM
Ensemble-1
Ensemble-2
Ensemble-3
Ensemble-4
Ensemble-5
Ensemble-6
B)
Mean number of GO terms per gene
8
6
4
2
0
PGSC
ITAG
BLAST2GO
InterPro2GO
OrthoMCL
Phytozome
Tri_BLAST
Tri_HMM
Ensemble-1
Ensemble-2
Ensemble-3
Ensemble-4
Ensemble-5
Ensemble-6


A) BP
F-measure
0.9
0.8
0.7
0.6
0.5
0.4
0.3
BLAST2GO
InterPro2GO
OrthoMCL
Phytozome
Tri_BLAST
Tri_HMM
Ensemble-1
Ensemble-2
Ensemble-3
Ensemble-4
Ensemble-5
Ensemble-6
B) MF
F-measure
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
BLAST2GO
InterPro2GO
OrthoMCL
Phytozome
Tri_BLAST
Tri_HMM
Ensemble-1
Ensemble-2
Ensemble-3
Ensemble-4
Ensemble-5
Ensemble-6


A)
Number of GO terms with p < 0.001
700
600
500
400
300
200
100
0
BLAST2GO
InterPro2GO
OrthoMCL
Phytozome
Tri_BLAST
Tri_HMM
Ensemble-1
Ensemble-2
Ensemble-3
Ensemble-4
Ensemble-5
Ensemble-6
B)
Percentage of GO terms with p < 0.001
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
BLAST2GO
InterPro2GO
OrthoMCL
Phytozome
Tri_BLAST
Tri_HMM
Ensemble-1
Ensemble-2
Ensemble-3
Ensemble-4
Ensemble-5
Ensemble-6

A) PGSC

Aggregated rank
0
2
4
6
8
10
Ensemble-2
Ensemble-1
Phytozome
InterPro2GO
Ensemble-3
BLAST2GO
Ensemble-4
Tri_BLAST
Ensemble-5
OrthoMCL
Tri_HMM

B) ITAG

Aggregated rank
0
2
4
6
8
Ensemble-2
BLAST2GO
Ensemble-1
Ensemble-3
Ensemble-4
OrthoMCL
InterPro2GO
Tri_BLAST
Ensemble-5
Phytozome
Tri_HMM

Addtional files provided with this submission:

Additional file 1. Figure S1 ITAG pipeline similarity, Figure S2 ITAG gold standard validation, Figure S3 ITAG gene expression validation and Methods S1-5 (840k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s1.pdf

Additional file 2: Table S1. InterPro2GO PGSC pipeline output (3372k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s2.xls

Additional file 3: Table S2. BLAST2GO PGSC pipeline output (3350k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s3.xls

Additional file 4: Table S3. OrthoMCL-UniProt PGSC pipeline output (5680k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s4.xls

Additional file 5: Table S4. Phytozome PGSC pipeline output (2005k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s5.xls

Additional file 6: Table S5. Tri_BLAST PGSC pipeline output (3588k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s6.xls

Additional file 7: Table S6. Tri_HMM PGSC pipeline output (1315k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s7.xls

Additional file 8: Table S7. BioMart ITAG pipeline output (2036k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s8.xls

Additional file 9: Table S8. BLAST2GO ITAG pipeline output (3852k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s9.xls

Additional file 10: Table S9. OrthoMCL-UniProt ITAG pipeline output (4978k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s10.xls

Additional file 11: Table S10. Phytozome ITAG pipeline output (1112k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s11.xls

Additional file 12: Table S11. Tri_BLAST ITAG pipeline output (3741k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s12.xls

Additional file 13: Table S12. Tri_HMM ITAG pipeline output (1386k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s13.xls

Additional file 14: Table S13. Potato gold standard genes with literature references (17k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s14.xlsx

Additional file 15: Table S14. PGSC gold standard (20k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s15.xlsx

Additional file 16: Table S15. ITAG gold standard (18k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s16.xlsx

Additional file 17: Table S16. PGSC ensemble output with k = 2 (4371k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s17.xls

Additional file 18: Table S17. ITAG ensemble output with k = 2 (4074k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s18.xls

Additional file 19. GO annotation file based on ensemble k = 2 for JHI Solanum tuberosum 60 k v1 microarray (ArrayExpress ID: E-MTAB-1655) (3527k)
http://www.biomedcentral.com/content/supplementary/s12870-014-0329-9-s19.xlsx