Genome Function Phylogenetics

by

Leila Fattel

## Introduction

Phylogenetics has gained importance in various areas of research in biology [1]. Its range of applications are as broad as studying the evolution of organisms [2], locating the source of a bacterial outbreak [3], providing evidence in criminal court [4], understanding socioeconomic effects in the evolution of different cultures [5], and much more. Usually, phylogenetic trees are built based on deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or protein sequences [6]. However, in this study, we will be creating phylogenetic trees based on gene ontologies (GO) in different plant genomes that were annotated using a pipeline developed by our lab called Gene Ontology Meta Annotator for Plants (GOMAP) that generates high-confidence, highly extensive and reproducible functional annotations by combining different prediction approaches [7].

Due to the exponential increase and complexity of biological data, the GO project was initiated to create a public resource for genes and their products using standardized vocabulary and defined relationships that biologists agree on when computationally annotating those genes [8]. The GO project encompassed three databases for three model organisms (fruit fly, yeast, and mice) initially, and has expanded to include many more species from different kingdoms (bacteria, plants, animals...) since then [9]. Gene ontologies are grouped under three key domains: 1) molecular function (describes activity at the molecular level, e.g. kinase activity) 2) biological process (describes what overall process does the product contribute to, e.g.

photosynthesis), and 3) cellular component (describes location, e.g. outer nuclear membrane). The ontologies are arranged in a hierarchal manner such that the terms are more general (ancestor terms) at the top and become more specific (child terms) as one moves down the chart (Figure 1).

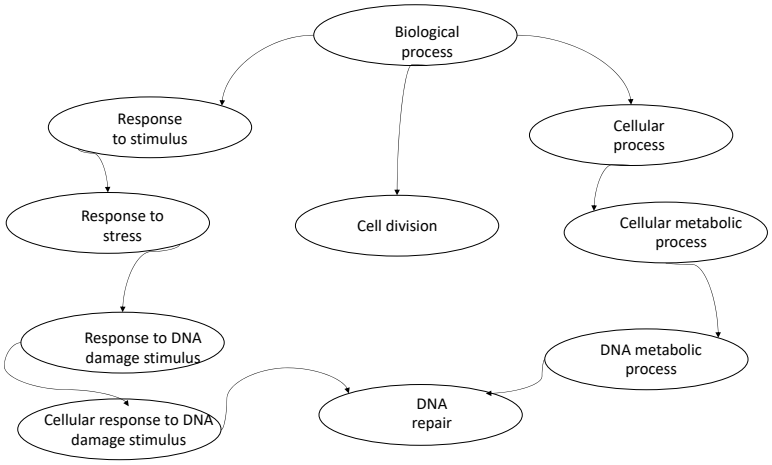



Figure 1: An example subtree of gene ontology

The significance of GO databases is that they allow researchers to characterize similar genes or gene products in different organisms, identify candidate genes involved in a specific process, or find genes involved in one specific process and exclude ones that are common to another [10].

The aim of this project is to continue a study initiated by my colleague in which he created phylogenetic trees based on GO terms for plant genomes annotated using the GOMAP pipeline developed by our lab. Here, more plant genomes have been added to the original study to build new phylogenetic trees, and the resulting trees were subject to bootstrapping analysis. This

statistical method was used to assess the confidence of each clade based on the proportion of bootstrap trees showing the same clade.

Methods

18 plant genomes were included in this project, including 3 plant genomes that were not previously part of the original species set: *Vigna unguiculate* (cowpea), *Cannabis sativa* (cannabis), and *Pinus lambertiana* (sugar pine). Their annotation data sets that were generated by GOMAP are available at https://dill-picl.org/projects/gomap/gomap-datasets/

Both distance-based and parsimony-based approaches were used to build the phylogenetic trees. The trees were constructed using the same method used by my colleague to ensure comparability (more details to reproduce the trees can be found in the appendix). Briefly, genome annotation datasets S are created such that any gene G is annotated with gene ontology term T, as well as all its ancestor terms A: S = $\bigcup_{i=1}^{\infty}(T_i \cup A_i)$. These functional annotation sets serve as a starting point of our analysis to be able to generate matrices containing all the GO annotations of the different plant species. For neighbor-joining, the matrix was created using Jaccard distance (1- $\frac{|S_a \cap S_b|}{|S_a \cup S_b|}$), which calculates how different any two sets are from each other, and was used as the input file. For parsimony, a binary matrix (presence/absence), also generated from the functional annotation sets, was used as an input instead. Trees were constructed using PHYLIP. *P. lambertiana*, a conifer, was included in this dataset to be used as an outgroup to the crop plant species to separate between the monocot and dicot plants. Dendroscope was used to visualize the trees using their Newick format, and root them at *P. lambertiana*.

Bootstrapping analysis was carried out for the parsimony tree. First, seqboot from PHYLIP was used to create new matrices by randomly selecting columns from the original matrix with replacement. Then, pars was used to repeatedly generate multiple trees using the output of the previous step as weight sets for the original binary matrix generated before. Next, the majority rule consensus tree was calculated with bootstrap values using consense. This procedure was repeated for different annotation sets having different combinations of plant species.

## Results

The phylogenetic trees created by neighbor-joining and parsimony methods are shown in Figures 2 and 3, respectively, both rooted at *P. lambertiana* (sugar pine).

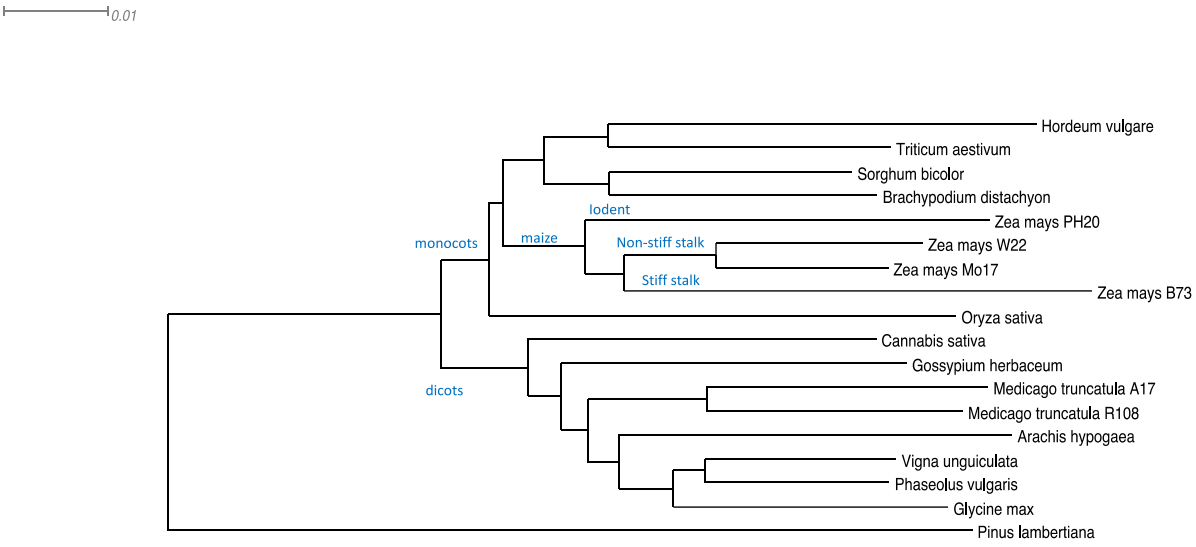



Figure 2: Neighbor-joining tree with 18 plant species

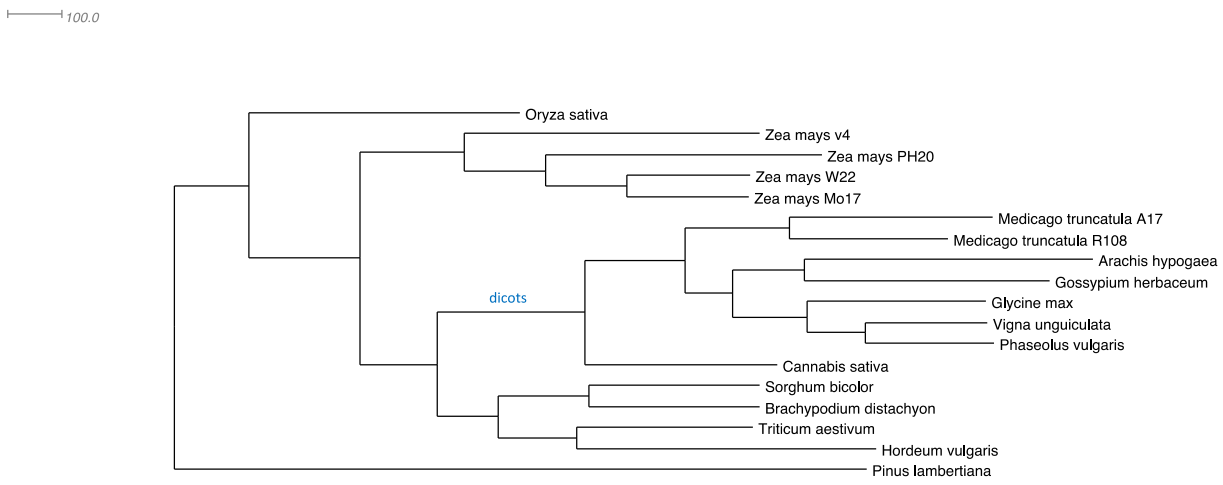



Figure 3: Parsimony tree with 18 plant species

We then proceeded to check the resolution of the parsimony tree by carrying out a bootstrap analysis to examine the confidence of each clade (Figure 4). The same procedure was repeated for the parsimony trees without *Medicago truncatula* (Figure 5), without *Brachypodium distachyon* (Figure 6), and without both species (Figure 7).

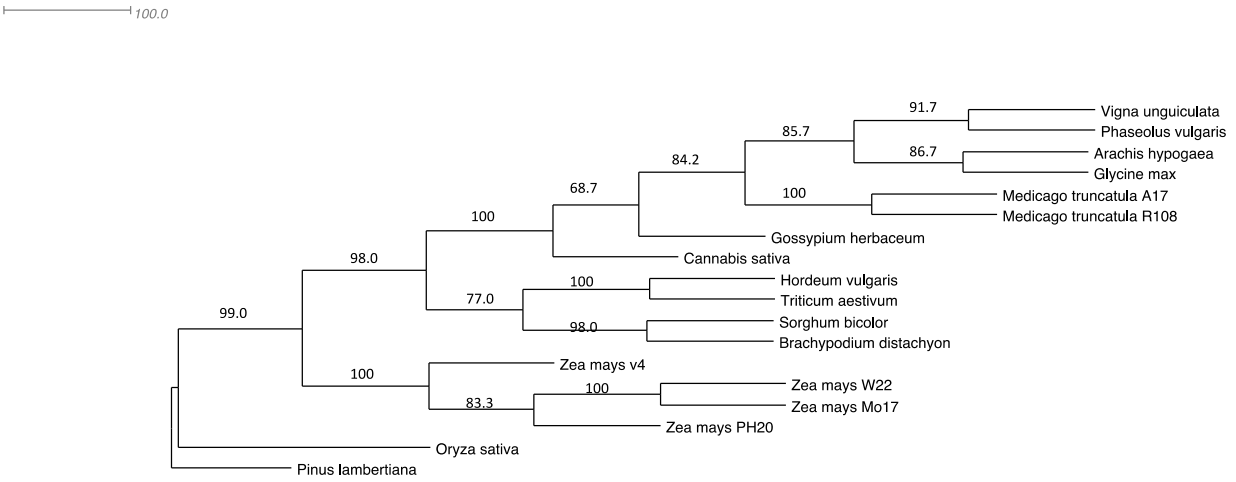



Figure 4: Bootstrap analysis of parsimony tree with all 18 plant species

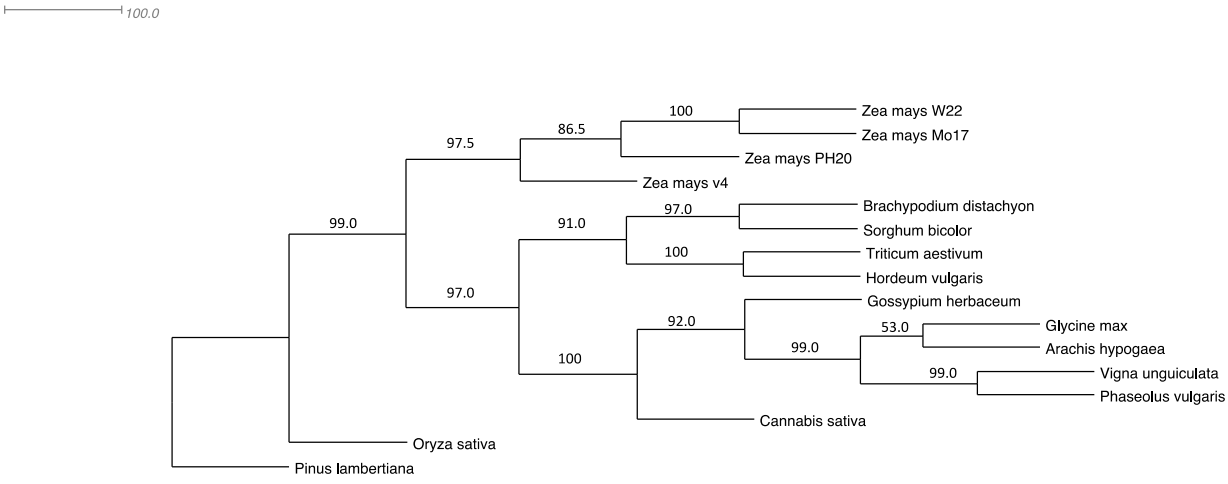



Figure 5: Bootstrap analysis of parsimony tree excluding *Medicago truncatula*

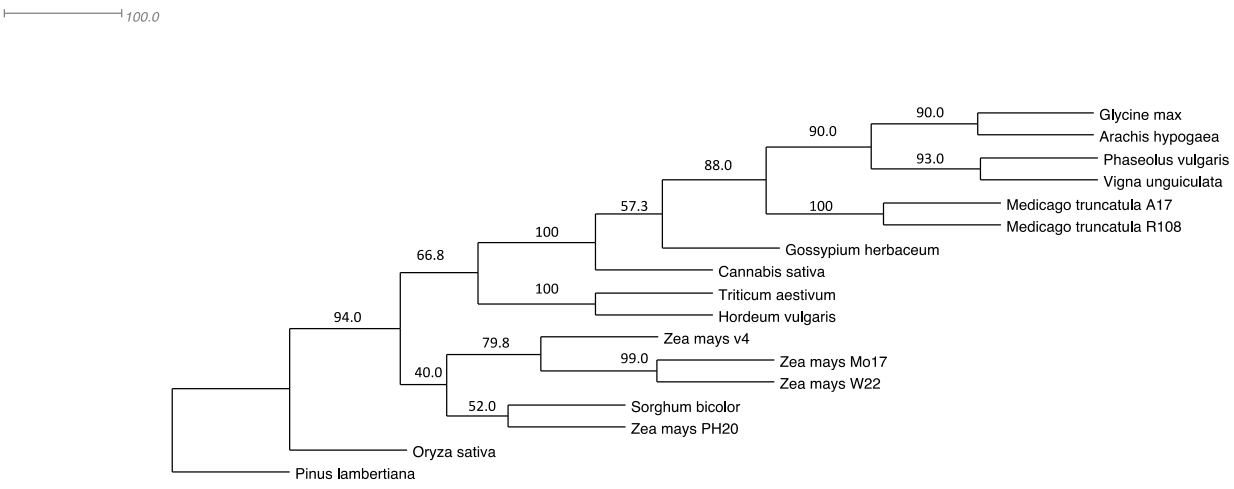



Figure 6: Bootstrap analysis of parsimony tree excluding *Brachypodium distachyon*

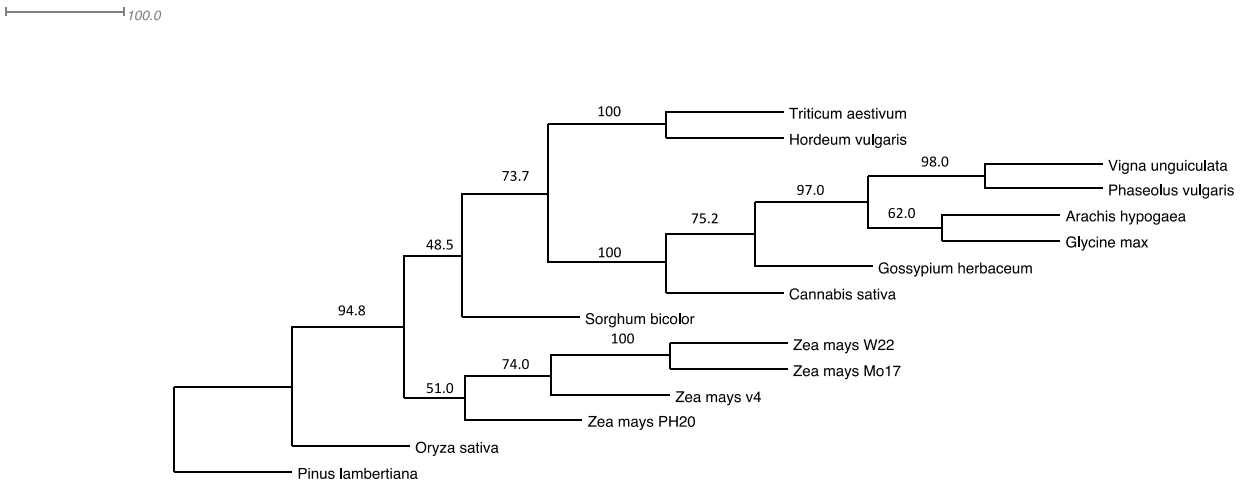



Figure 7: Bootstrap analysis of parsimony tree excluding both *Medicago truncatula* and *Brachypodium distachyon*

Discussion

In the previous study, 15 plant species were used to generate distance-based and parsimony-based phylogenetic trees. The resulting trees were very similar to the expected species tree (Figure 8), except for two main differences: *Sorghum* was not at the base of maize as expected, and *Arachis* should be closer to *Gossypium* than *Medicago* is. We hypothesized that this unexpected topology may be due to the relatively small size of the *Brachypodium* and *Medicago* genomes, and the high mutation rate of *Medicago* [14]. Also, it seemed that the two generated trees had different topologies.

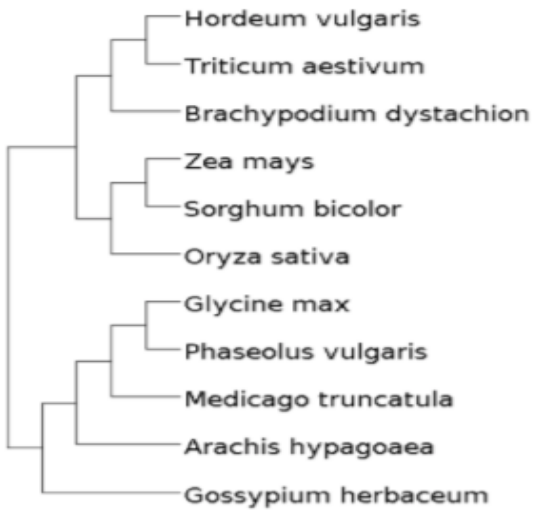



Figure 8: Expected taxonomic tree [11,12,13]

We decided to add three more genomes to our set of species and observe if the expected topology can be restored. Rooting the trees with *P. lambertiana*, a conifer, served as a good outgroup as it led to the clear resolution of monocots and dicots as seen in Figures 2 and 3. In addition, the distance-based tree and parsimony-based tree seem to have similar topology upon the addition of the new plant genomes, however, the separation of monocots and dicots was

clearer in the neighbor-joining tree than in parsimony tree. We hypothesized that this could be due to the parsimony method treating GO terms as equal regardless of their hierarchal position, unlike the distance-based method.

However, the rearrangement of *Sorghum* and *Medicago* was unresolved. So, we decided that the next step should be to check whether the resolution is real by bootstrapping analysis. In Figure 4, most of the support values seemed to be fairly high, except for a 77.0% at one of the branches that could support the hypothesis that *Sorghum* should not be clustered with *Brachypodium*, and a 68.7% at another branch that could also support that *Medicago* and *Arachis* should be switched. The removal of *Medicago* resulted in higher confidence values in general (Figure 5), except for the branch that splits into *Arachis* and *Glycine*, which could be an indicator that *Arachis* should be closer to *Gossypium* instead as seen in the expected taxonomic tree. We later removed *Brachypodium* from the species set to see its effect on the position of *Sorghum* (Figure 6). Low values were observed at the base of the sorghum-maize cluster, which is expected as sorghum should be at the base of the different maize plants, and not disrupting the clustering of maize together. Also, a 66.8% at one of the branches can be due to the placement of *Triticum* and *Hordeum* away from the other monocots, while the low value of 57.3% can again be due to the switching of *Medicago* and *Arachis*. Finally, a bootstrap analysis for a tree lacking both *Brachypodium* and *Medicago* was done (Figure 7). Low support values can be observed throughout, and this is most likely due to the lack of clustering of the monocots. Again, we notice a low value at the base of *Arachis* probably due to it not being as close to *Gossypium* as expected. As a conclusion, bootstrap analysis indicates that we were right to suspect the positioning of *Sorghum* and *Medicago*.

As an initial effort to understand the reason behind the unexpected clustering, we investigated the amount of GO terms that are unique to some species but not included in others. For example, we observed 916 (2.3%) GO terms found in both *Gossypium* and *Arachis*, but not in the *Medicago* species, while there were 847 (2.1%) GO terms to *Gossypium* and *Medicago*, but not *Arachis*. Meanwhile, *Sorghum* and *Brachypodium* had 1510 (2.6%) GO terms excluded in the maize species, while the later had 842 (1.4%) common GO terms with *Sorghum*, not found in *Brachypodium*. The difference in the amount of intersecting GO terms between *Brachypodium* and the maize species with *Sorghum* indicates that there is higher similarity in gene function between *Brachypodium* and *Sorghum* and could explain why *Sorghum* was clustered with *Brachypodium* instead of maize. However, the same reasoning cannot be applied for *Arachis* and *Medicago*, but this could be due to the similar percentage of respective common GO terms with *Gossypium*. Therefore, further analysis is needed to compare the functional annotation of the genes themselves to understand the source of such discrepancies, and the biological significance of the trees.

References


1. Soltis, D. E., & Soltis, P. S. (2003). The role of phylogenetics in comparative genetics. *Plant Physiology*, *132*(4), 1790-1800.

2. Bates, K. T., Mannion, P. D., Falkingham, P. L., Brusatte, S. L., Hutchinson, J. R., Otero, A., ... & Allen, V. (2016). Temporal and phylogenetic evolution of the sauropod dinosaur body plan. *Royal Society Open Science*, *3*(3), 150636.

3. Pightling, A. W., Pettengill, J. B., Luo, Y., Baugher, J. D., Rand, H., & Strain, E. (2018). Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. *Frontiers in microbiology*, *9*, 1482.

4. Sullivan, J., & Joyce, P. (2005). Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, *36*, 445-466.

5. Watts, J., Sheehan, O., Atkinson, Q. D., Bulbulia, J., & Gray, R. D. (2016). Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature*, *532*(7598), 228-231.

6. Munjal, G., Hanmandlu, M., & Srivastava, S. (2018). Phylogenetics Algorithms and Applications. *Ambient Communications and Computer Systems: RACCCS-2018*, *904*, 187–194.

7. Wimalanathan, K., Friedberg, I., Andorf, C. M., & Lawrence-Dill, C. J. (2018). Maize GO annotation—methods, evaluation, and review (maize-GAMER). *Plant Direct*, *2*(4), e00052.

8. Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research*, *45*(D1), D331-D338.

9. Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, *32*(suppl_1), D258-D261.

10. Gene Ontology Consortium. (2001). Creating the gene ontology resource: design and implementation. *Genome research*, *11*(8), 1425-1433.

11. Azani, N., Babineau, M., Bailey, C. D., Banks, H., Barbosa, A. R., Pinto, R. B., ... & Candido, E. (2017). A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny: The Legume Phylogeny Working Group (LPWG). Taxon, 66(1), 44-77.

12. Kellogg, E. A. (2001). Evolutionary history of the grasses. Plant physiology, 125(3), 1198-1205.

13. Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., ... & Donoghue, P. C. (2018). The interrelationships of land plants and the nature of the ancestral embryophyte. Current Biology, 28(5), 733-745.

14. Young, N. D., Debellé, F., Oldroyd, G. E., Geurts, R., Cannon, S. B., Udvardi, M. K., ... & Van de Peer, Y. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, *480*(7378), 520-524.

# Appendix

# *Genome Function Phylogenetics*

*Dennis Psaroudakis*

*April 28th 2019*

When we build phylogenetic trees, we often use the DNA or protein sequences of certain genes[1] as the basis for our analysis. They are the direct substance of evolution so that makes sense, especially when we need high resolution (i.e. a signal in very closely related organisms). In some cases, they don't give the answer we're looking for though: Let's say for example, that you're a doctor and one of your patients is infected with a novel pathogen. You can sequence that pathogen and place it somewhere in the bacterial taxonomy, and naturally you will start treating the patient based on what you know about closely related species. Unfortunately, none of the antibiotics you're using show any effect. How come? Just because two organisms are closely related does not *necessarily* mean that they are similar in their phenotype. Especially if you've only used genes that are suitable for regular phylogenetics but don't actually have anything to do with the organism's metabolism, membrane structure etc, you have no guarantee that close evolutionary relationship means similar response to environmental factors. In this case, a tree that can group your pathogen with organisms based on their actual *in vivo* similarity would be much more helpful here. There are trees built on phenotypic characteristics (nose color, ear length…) but the choice of what characteristics to look at is not trivial and has an incredible influence on the result. You will never be able to cover all relevant attributes and your choice can therefore be considered somewhat arbitrary[2]. A method would be needed that is exhaustive, clearly defined, and reproducible, but also more effective at finding the answer you're looking for than using the DNA sequence. And, unsurprisingly, there is [6] (it's a cool paper, check it out)!

In this paper, I'm going to do something less exciting. Instead of working with life-threatening bacteria and saving patients, I will focus on plants. The idea is somewhat similar though[3]: I want to build a phylogenetic tree of plant species, not based on their genetic sequence but on the processes, structures, and reactions that are present in it.

[1] with appropriate degree of conservedness depending on the evolutionary distance of the taxa, little probability of horizontal gene transfer, reliable for sequencing and alignment etc.

[2] Look at this (parodic) taxonomy of animals by Jorge Luis Borges: 1. those that belong to the Emperor, 2. embalmed ones, 3. those that are trained, 4. suckling pigs, 5. mermaids, 6. fabulous ones, 7. stray dogs, 8. those included in the present classification, 9. those that tremble as if they were mad, 10. innumerable ones, 11. those drawn with a very fine camelhair brush, 12. others, 13. those that have just broken a flower vase, 14. those that from a long way off look like flies.

[3] and I had the idea myself first and only found the paper afterwards

## *The Gene Ontology*

Historically, the function/role that a gene plays in an organism has been described in natural language, however the researcher characterizing that gene deemed best. While this is nice to read, it is not very useful if you want to do computation on it, as computers are (still) horrible at understanding natural language and determining the structure in meaning behind the words. Additionally, different people will describe the same thing with different words,

so there is always a potential for misunderstanding.

Ontologies try to alleviate these problems by providing a strictly organized and controlled vocabulary and defined relationships between the terms, so that the same statement always means the same thing, no matter the context or the author. Additionally, ontologies can be understood by computers if all relationships and terms are clearly defined.

The Gene Ontology (GO) is such an ontology. It describes genes by the properties of their product. In our case these gene products are proteins, and they can be characterized in three different aspects:

- What biological processes is this protein part of? (e.g. photosynthesis or autophagy)
- What molecular functions does the protein carry out? (e.g. ethylene binding or RNA ligation)
- What cellular component is the protein active at? (e.g. outer membrane or nucleus)

"The mission of the GO Consortium is to develop an up-to-date, comprehensive, computational model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems."
— GO Consortium (geneontology.org)

Within each of these aspects, the Gene Ontology defines a huge number of terms (2,675,070 in total), that range from very general to very specific:

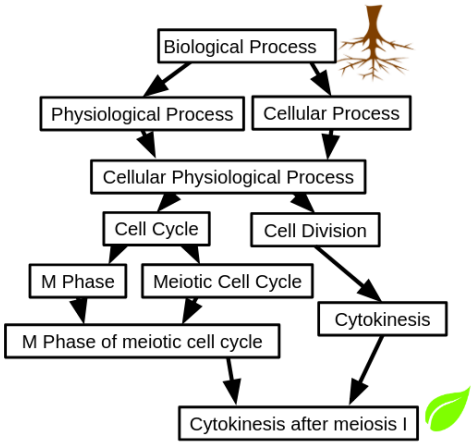



Figure 1: subtree of the *Biological Process* ontology. The terms are organized in such a way that more general terms are always true for any of their more specific child terms. For example, any protein that is part of *Cytokinesis after Meiosis I*, is also obviously part of Cytokinesis, the Cell Cycle etc. That way, a gene that has been annotated with the term *Cytokinesis after Meiosis I* (leaf term), has implicitly been annotated with all of that term's parent terms as well, all the way up to the root term.

When proteins are annotated with these terms instead of just natural language, we can now computationally answer some interesting questions, such as:

- How similar in function[4] is protein A to protein B? (One answer would be: How many steps in the GO graph do I need from term A to B? The fewer steps, the more similar the function)
- If protein A is involved in Biological Process XYZ, what other proteins are involved in that same process?

[4] *Function* refers to any aspect of the GO, not just Molecular Function.

The Gene Ontology is quite well established in the field, so you will find GO annotations all across UniProt entries or be able to use dedicated tools like AmiGO or QuickGO to examine a protein of interest.

## *Data*

Annotating genes with their functions can be done experimentally (e.g. by knocking out a certain gene and seeing what processes in the cell are affected), but that is a time-consuming and expensive process, so methods have been developed that try to predict the function of a given gene. Our lab has developed such a pipeline called Gene Ontology Meta Annotator for Plants (GOMAP) which combines different prediction approaches and is able to generate high-confidence and very extensive Functional Annotations in a reproducible manner [4]. We have been applying this pipeline to whole-genome assemblies of different plant species and generated functional annotations for every gene in each genome. These annotation sets are (or will be shortly) available from

`https://dill-picl.org/projects/gomap/gomap-datasets.`

## *Method*

I am using two different tree building approaches, one is a distanced based method and the other a parsimony one.

### *Distance Based*

Starting point of our analysis are the functional annotation sets, one for each genome, which annotate every gene in the genome with one or more GO terms. In more mathematical terms the genome annotation set is a list of tuples $(G, T)$ with $G \in$ Genes in that genome and $T \in$ Terms in the Gene Ontology.

We can use the hierarchical structure of the Gene Ontology to obtain the ancestors $A_i$ of any term $T_i$; in other words the gene $G_i$ is not just annotated with the term $T_i$ itself but also with all GO terms that are a more general statement of that term (e.g. any gene that is part of a metabolic process is thereby also part of a biological process). We do that for all terms $T$ in the dataset and combine all of the terms and their ancestors into one big genome-wide set $S$, irrespectively of the gene they were originally associated with: $S = \bigcup_{i=1}^{x}(T_i \cup A_i)$.

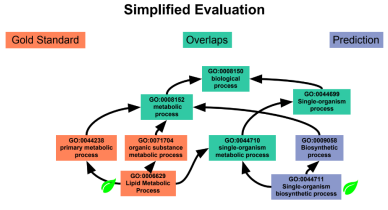
When this superset of annotations is created for each of the datasets, we can use the Jaccard Distance as a measure of how (dis-)similar any two sets are from each other, or in biological terms how different the two genomes are on a functional level:

$$\text{Jaccard Distance}(S_a, S_b) = 1 - \frac{|S_a \cap S_b|}{|S_a \cup S_b|}$$

Applying this formula to all pairwise combinations of the genomes we're looking at yields a $S \times S$ distance matrix that can then serve as the input for a

```
Gene           GO Term
Os01g0601625 GO:0050896
Os01g0601625 GO:0016021
Os01g0601625 GO:0016301
Os01g0601651 GO:0003677
Os01g0601651 GO:0009699
Os01g0601651 GO:0050790
Os01g0601651 GO:0050794
Os01g0601651 GO:0050896
Os01g0601675 GO:0007275
Os01g0601675 GO:0016310
Os01g0601675 GO:0050789
...
```

The general idea of using the Jaccard Distance in this context is to measure the overlap of two subtrees in the GO hierarchy. Say, for simplicity, that we're looking at two genomes (here called Gold Standard and Prediction) that each only contain one single GO term (marked by a leaf). First, we add all ancestors of that leaf term to each subtree. Then, we determine the overlap (which corresponds to $S_a \cap S_b$), and divide the number of nodes in this overlap by the number of nodes in either of the two subtrees ($S_a \cup S_b$).




In the case of this example, the Jaccard Distance of Gold Standard and Prediction would be $1 - \frac{4}{9} = \frac{5}{9}$

| Taxon | GO:0016021 | GO:0009699 | GO:0050794 | GO:0050789 | GO:0060739 | … |
|---|---|---|---|---|---|---|
| *G. max* | 1 | 0 | 1 | 0 | 1 | |
| *T. aestivum* | 1 | 1 | 0 | 1 | 1 | |
| … | | | | | | |

neighbor joining algorithm (provided by PHYLIP). I rooted the resulting tree manually outside of the grasses (maize, wheat, rice, barley).

### *Parsimony Based*

Like in the previous method, we again start by enriching all taxa sets $(G, T)$ with their ancestor terms and discarding the gene association: $S = \bigcup_{i=1}^{x}(T_i \cup A_i)$.

Instead of using these sets for distance calculation, we combine them into a big binary matrix that displays which terms are present in which set:

This matrix was then used as the input for `pars` from the PHYLIP package to find the maximally parsimonious tree.

## *Results*

The phylogram created by the distance based method is displayed in Figure 2, the maximum parsimony tree in figure 3.

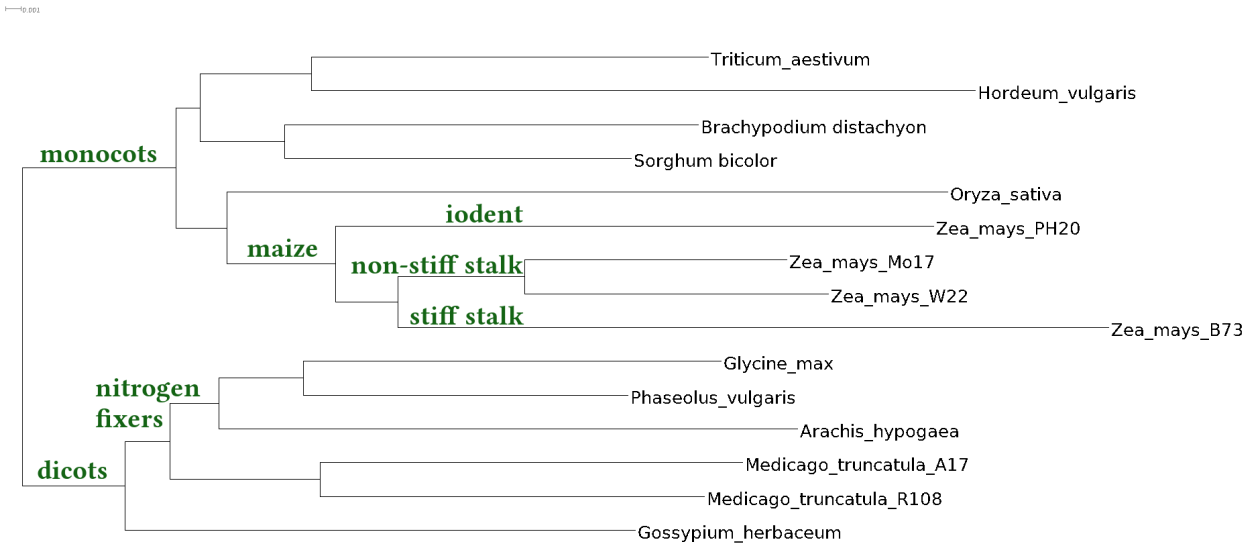



Figure 2: Phylogram built on the Jaccard distance matrix with Neighbor Joining. Manually rooted between monocots and dicots and text in green added.

Both trees were combined into a tanglegram (see figure 4).

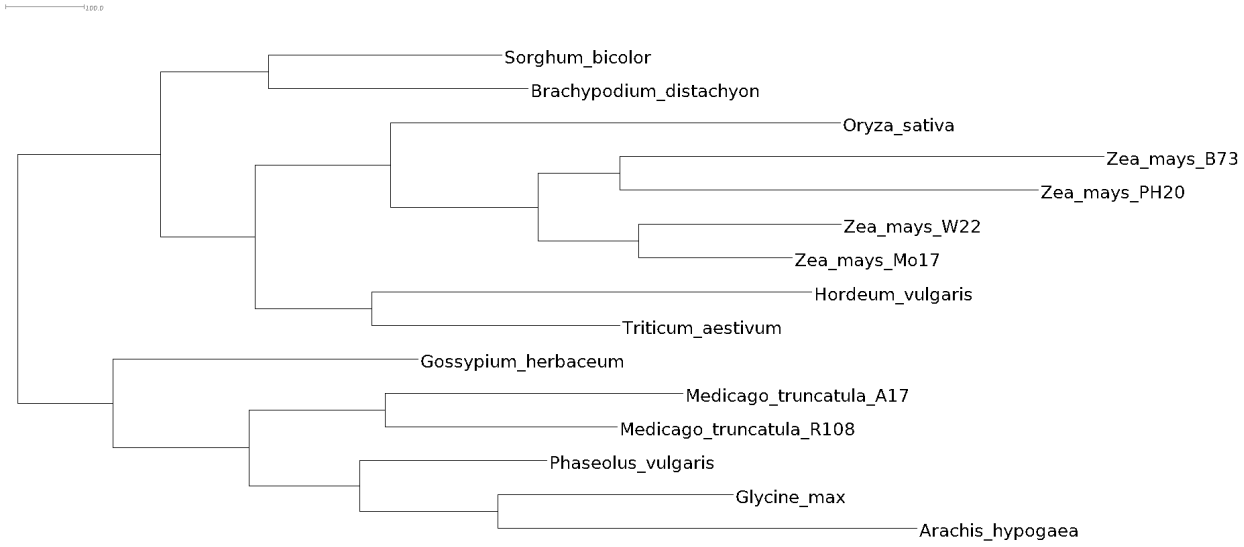



Figure 3: Phylogram built by looking for the maximally parsimonious tree (total of 7780 changes). Manually rooted between monocots and dicots.

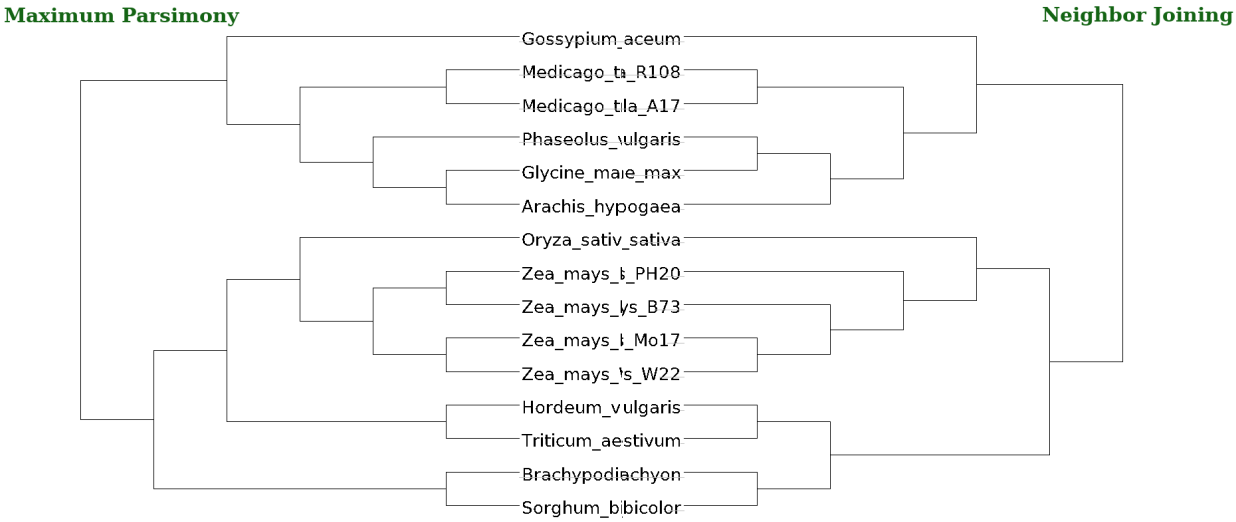



Figure 4: Tanglegram of the distance based and maximum parsimony tree (both manually rooted).

## *Discussion*

### *Fulfilling expectancies?*

In a perfect world, the evolution of organisms could be just as easily retraced by the evolution of their functions as by their sequence; after all the selective pressure is on the mechinisms and activities of an organism and not on its DNA sequence and what differentiates two species from each other are the differences in their *in vivo* pheontype.

Although you may get lower resolution, because silent mutations or such that don't alter the function of a protein would be missed.

The expected taxonimic tree is displayed in figure 5 and indeed the tree produced by the distance based method is quite similar (much more similar than I had expected). There are only two notable differences: *Sorghum* should be at the base of maize and not grouped with *Brachypodium* and *Medicago* and *Arachis* should switch their places. We are currently investigating the reason for this phenomenon and the best explanation we've come up with so far is that *Brachy* and *Medicago* are not actually good representations of their respective group; they have been chosen as model organisms because they are easy to sequence but that's mainly because their genome is much smaller and less complex than that of actual crop plants. *Medicago* is also well known for having an unusally high rate of genomic evolution, gene births and gene deaths [5].

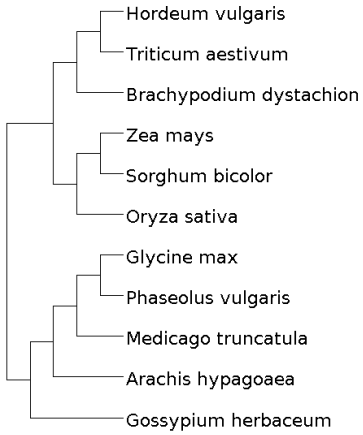



Figure 5: Expected cladogram. [2, 1, 3]

It is difficult to come up with a clear evolutionary tree within the same species (I am talking about maize[5]) but it is encouraging to see that in the generated tree the 3 different classes of maize (stiff stalk, non-stiff stalk, and iodent) are differentiated from each other in a reasonable way.

[5] Many of the maize lines have more than one ancestor and there is an immense amount of horizontal gene transfer with maize, even to closely related species.

What about the parsimony tree? Surprisingly, the tree built on the parsimony method is not identical with the distance based one, and it is even less similar to the expected tree. Our first thought was that since the parsimony method does not search the complete tree space, it might simply not have found the distance based tree even though it is the most parsimonious one. Unfortunately that does not seem to be the case: The tree it found has a score of 7780, the distance based tree 7821. When manually switching *Medicago* and *Arachis* to get closer to the taxonomic tree, the score improved to 7812, but putting *Sorghum* at the place it was expected increased the score again to 7866. So it seems the answer is not that simple and we will have to reflect more on our methods to understand where the difference comes from and which is the more meaningful result.

### *Tracing Back the Signal*

Another big question is why the tree actually looks the way it does. Is it truly a biologically meaningful display of differences and similarities in function or is it just an artifactual tree, caused by a bias in the method. Given the number of taxa it seems unlikely that the similarity to the expected tree is just random

noise, so there must be some systematic reason behind it. To make a first step answering this question, I investigated what actually causes species to cluster the way they do. I asked the following question: Which terms are common to all nitrogen fixing plants but do not occur in any other plant in the tree?[6] If these terms are actually meaningful to the process of nitrogen fixation (or some other characteristic that clearly differentiates this group of plants from the others) that would indicate that the phylogenetic signal actually comes from biologically plausible differences in the functional annotations. So, here is the answer:

[6] in mathy words this is the intersection of terms in the nitrogen fixing species minus the union of terms in all other species

- GO:0080184 - response to phenylpropanoid
- GO:0033800 - isoflavone 7-O-methyltransferase activity
- GO:0042577 - lipid phosphatase activity
- GO:0031174 - lifelong otolith mineralization
- GO:0045299 - otolith mineralization
- GO:0006742 - NADP catabolic process
- GO:0019364 - pyridine nucleotide catabolic process
- GO:0019677 - NAD catabolic process
- GO:0070823 - HDA1 complex

Three of these terms do seem plausible: The Nod Factor molecule that is secreted by the plant as a signal to Rhizobia bacteria is a phenylpropanoid, isoflavone 7-O-methyltransferase activity is involved in Nod Factor synthesis, and the nodule has a complex lipid membrane system – and Nod Factor has a lipid component (which could explain lipid phosphatase activity). The last 4 terms are possibly plausible, but less certain, but the two otolith terms seem very out of place, since they describe small oval calcareous bodies in the inner ear of vertebrates, not something expected to be found in legumes. There might still be some biological significance to it but we're not sure yet what that might be. So for now it seems that some of the signal clearly comes from biological differences while some other part might be due to the method.[7]

[7] I looked at other, similar constellations as well and the answer was the same (e.g. what differentiates non-stiff stalk maize lines from the others).

### *What's next?*

Since this is part of a publication I am working on, a lot more thought will be put into, for example, the following questions:

- How similar is this tree to the taxonomic tree?
- What are the reasons for the differences, what are the reasons for the similarities?
- What biases in the construction process need we be aware of?
- What does the tree depict/how do we interpret it?
- In what way could it be scientifically valuable?
- Are there any other taxa that would be good to have in the tree?
- Is there any further analysis that would be good to do? (e.g. branch support, looking at each GO aspect separately, a more sophisticated measure

of GAF similarity...)
- Is rooting the tree there reasonable?

### *Reproducing the Tree*

To reproduce this tree, follow these steps on a Linux machine that has PHYLIP available:

```
git clone https://github.com/Thyra/EEOB563.git paper_dennis # Clone the repository
cd paper_dennis/final_project # change into the directory

# Distance based method
bin/build_distance_matrix annotation_sets/*tree.json > distance_matrix.phy
module load phylip # (if you're on HPC)
neighbor # (use standard options)

# Parsimony based Method
bin/binary_table annotation_sets/*.tree.json > binary.csv
bin/binary2phylip binary.csv > binary.phy
module load phylip
pars # (again, standard options)

# If you're adding more GAFs
bin/gaf2json <new_gaf>
# Then repeat distance or parsimony as above

# To replace GAF filenames with species names in newick tree
bin/rename_taxa <tree.newick> resources/taxa_name_mapping.csv
```

If any of the binaries fails without a message that says otherwise, you probably need to install some more libraries. If that doesn't help you may have to recompile them for your processor type (they're written in Crystal, source code is in `src/`)

### *References*

[1] Nasim Azani, Marielle Babineau, C. Donovan Bailey, Hannah Banks, ArianeR. Barbosa, Rafael Barbosa Pinto, JamesS. Boatwright, LeonardoM. Borges, GillianK. Brown, Anne Bruneau, Elisa Candido, Domingos Cardoso, Kuo-Fang Chung, RuthP. Clark, Adilva deS. Conceição, Michael Crisp, Paloma Cubas, Alfonso Delgado-Salinas, KyleG. Dexter, JeffJ. Doyle, Jérôme Duminil, AshleyN. Egan, Manuel De La Estrella, MarcusJ. Falcão, DmitryA. Filatov, Ana Paula Fortuna-Perez, RenéeH. Fortunato, Edeline Gagnon, Peter Gasson, Juliana Gastaldello Rando, Ana Maria Goulart de Azevedo Tozzi, Bee Gunn, David Harris, Elspeth

Haston, JulieA. Hawkins, PatrickS. Herendeen, ColinE. Hughes, JoãoR.V. Iganci, Firouzeh Javadi, Sheku Alfred Kanu, Shahrokh Kazempour-Osaloo, GeoffreyC. Kite, BenteB. Klitgaard, FábioJ. Kochanovski, ErikJ.M. Koenen, Lynsey Kovar, Matt Lavin, Marianne le Roux, GwilymP. Lewis, HaroldoC. de Lima, Maria Cristina López-Roberts, Barbara Mackinder, Vitor Hugo Maia, Valéry Malécot, VidalF. Mansano, Brigitte Marazzi, Sawai Mattapha, JosephT. Miller, Chika Mitsuyuki, Tania Moura, DanielJ. Murphy, Madhugiri Nageswara-Rao, Bruno Nevado, Danilo Neves, DarioI. Ojeda, R. Toby Pennington, DariénE. Prado, Gerhard Prenner, Luciano Paganucci de Queiroz, Gustavo Ramos, FabianaL. Ranzato Filardi, PétalaG. Ribeiro, María de Lourdes Rico-Arce, MichaelJ. Sanderson, Juliana Santos-Silva, WallaceM.B. São-Mateus, MarcosJ.S. Silva, MarceloF. Simon, Carole Sinou, Cristiane Snak, ÉlviaR. de Souza, Janet Sprent, KellyP. Steele, JuliaE. Steier, Royce Steeves, CharlesH. Stirton, Shuichiro Tagane, BenjaminM. Torke, Hironori Toyama, Daiane Trabuco da Cruz, Mohammad Vatanparast, JanJ. Wieringa, Michael Wink, MartinF. Wojciechowski, Tetsukazu Yahara, Tingshuang Yi, and Erin Zimmerman. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny – The Legume Phylogeny Working Group (LPWG). *Taxon*, 66(1):44–77, feb 2017. ISSN 0040-0262. DOI: 10.12705/661.3. URL http://doi.wiley.com/10.12705/661.3.

[2] E A Kellogg. Evolutionary history of the grasses. *Plant physiology*, 125(3):1198–205, mar 2001. ISSN 0032-0889. DOI: 10.1104/PP.125.3.1198. URL http://www.ncbi.nlm.nih.gov/pubmed/11244101http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1539375.

[3] Mark N. Puttick, Jennifer L. Morris, Tom A. Williams, Cymon J. Cox, Dianne Edwards, Paul Kenrick, Silvia Pressel, Charles H. Wellman, Harald Schneider, Davide Pisani, and Philip C.J. Donoghue. The Interrelationships of Land Plants and the Nature of the Ancestral Embryophyte. *Current Biology*, 28(5):733–745.e2, mar 2018. ISSN 0960-9822. DOI: 10.1016/J.CUB.2018.01.063. URL https://www.sciencedirect.com/science/article/pii/S0960982218300964.

[4] Kokulapalan Wimalanathan, Iddo Friedberg, Carson M. Andorf, and Carolyn J. Lawrence-Dill. Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER). *Plant Direct*, 2(4):e00052, apr 2018. ISSN 24754455. DOI: 10.1002/pld3.52. URL http://doi.wiley.com/10.1002/pld3.52.

[5] Nevin D. Young, Frédéric Debellé, Giles E. D. Oldroyd, Rene Geurts, Steven B. Cannon, Michael K. Udvardi, Vagner A. Benedito, Klaus F. X.

Mayer, Jérôme Gouzy, Heiko Schoof, Yves Van de Peer, Sebastian Proost, Douglas R. Cook, Blake C. Meyers, Manuel Spannagl, Foo Cheung, Stéphane De Mita, Vivek Krishnakumar, Heidrun Gundlach, Shiguo Zhou, Joann Mudge, Arvind K. Bharti, Jeremy D. Murray, Marina A. Naoumkina, Benjamin Rosen, Kevin A. T. Silverstein, Haibao Tang, Stephane Rombauts, Patrick X. Zhao, Peng Zhou, Valérie Barbe, Philippe Bardou, Michael Bechner, Arnaud Bellec, Anne Berger, Hélène Bergès, Shelby Bidwell, Ton Bisseling, Nathalie Choisne, Arnaud Couloux, Roxanne Denny, Shweta Deshpande, Xinbin Dai, Jeff J. Doyle, Anne-Marie Dudez, Andrew D. Farmer, Stéphanie Fouteau, Carolien Franken, Chrystel Gibelin, John Gish, Steven Goldstein, Alvaro J. González, Pamela J. Green, Asis Hallab, Marijke Hartog, Axin Hua, Sean J. Humphray, Dong-Hoon Jeong, Yi Jing, Anika Jöcker, Steve M. Kenton, Dong-Jin Kim, Kathrin Klee, Hongshing Lai, Chunting Lang, Shaoping Lin, Simone L. Macmil, Ghislaine Magdelenat, Lucy Matthews, Jamison McCorrison, Erin L. Monaghan, Jeong-Hwan Mun, Fares Z. Najar, Christine Nicholson, Céline Noirot, Majesta O'Bleness, Charles R. Paule, Julie Poulain, Florent Prion, Baifang Qin, Chunmei Qu, Ernest F. Retzel, Claire Riddle, Erika Sallet, Sylvie Samain, Nicolas Samson, Iryna Sanders, Olivier Saurat, Claude Scarpelli, Thomas Schiex, Béatrice Segurens, Andrew J. Severin, D. Janine Sherrier, Ruihua Shi, Sarah Sims, Susan R. Singer, Senjuti Sinharoy, Lieven Sterck, Agnès Viollet, Bing-Bing Wang, Keqin Wang, Mingyi Wang, Xiaohong Wang, Jens Warfsmann, Jean Weissenbach, Doug D. White, Jim D. White, Graham B. Wiley, Patrick Wincker, Yanbo Xing, Limei Yang, Ziyun Yao, Fu Ying, Jixian Zhai, Liping Zhou, Antoine Zuber, Jean Dénarié, Richard A. Dixon, Gregory D. May, David C. Schwartz, Jane Rogers, Francis Quétier, Christopher D. Town, and Bruce A. Roe. The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, 480(7378):520–524, dec 2011. ISSN 0028-0836. DOI: 10.1038/nature10625. URL `http://www.nature.com/articles/nature10625`.

[6] Chengsheng Zhu, Tom O. Delmont, Timothy M. Vogel, and Yana Bromberg. Functional Basis of Microorganism Classification. *PLOS Computational Biology*, 11(8):e1004472, aug 2015. ISSN 1553-7358. DOI: 10.1371/journal.pcbi.1004472. URL `https://dx.plos.org/10.1371/journal.pcbi.1004472`.