# *Genome Function Phylogenetics*

*Dennis Psaroudakis*

*April 28th 2019*

When we build phylogenetic trees, we often use the sequences of certain genes as the basis for our tree. They are the direct substrate of evolution so that makes sense, but practically often too abstract to be helpful. Let's say for example, that you have a patient infected with a novel pathogen. You can identify that pathogen and place it somewhere in the bacterial taxonomy, but unfortunately, none of the treatments that you usually use for closely related species show any effect against this one. How come? Just because two organisms are closely related does not *necessarily* mean that they are similar in their phenotype. So a tree that groups organisms based on their actual *in vivo* similarity would be much more helpful here. There are such trees built on phenotypic characteristics but unfortunately the choice of what characteristics to look at is not trivial and always situation-dependent. They can also be considered quite arbitrary[1]. What's needed would be something that is exhaustive, clearly defined, and reproducible, but also more meaningful than just a DNA sequence. And they did it [6]!

In this paper, I'm going to do something slightly similar. Not on bacteria, but on plants though. I know, disapointing, but I'll find a similarly fascinating reason why that's important soon. So, to summarize: I want to build a phylogenetic tree on plant species not based on their genetic sequence but on the *functions* that this plant is able to execute.

[1] 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.

## *The Gene Ontology*

Historically, the function/role that a gene plays in an organism has always 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, which has the 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 that 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:

"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)

- 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)

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:

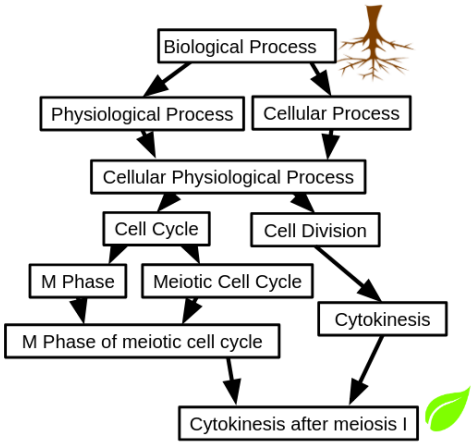



Figure 1: subtree of the *Biological Process* ontology. The terms are organized in such a way that more general terms are always true for any of their more specific child terms. For example, any protein that is part of *Cytokinesis after Meiosis I*, is also obviously part of Cytokinesis, the Cell Cycle etc. That way, a gene that has been annotated with the term *Cytokinesis after Meisosis 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 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?

The Gene Ontology is quite well established in the field, so you will find GO annotations for almost all relevant UniProt entries or 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 and predict the function of a given gene. Our lab has developed such a pipeline called 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 neighbor joining algorithm (provided by PHYLIP). I rooted the resulting tree manually outside of the grasses (maize, wheat, rice, barley).

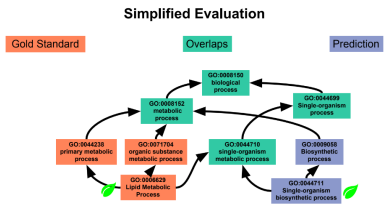
### *Parsimony Based*

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:

```
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 | |
| … | | | | | | |

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.

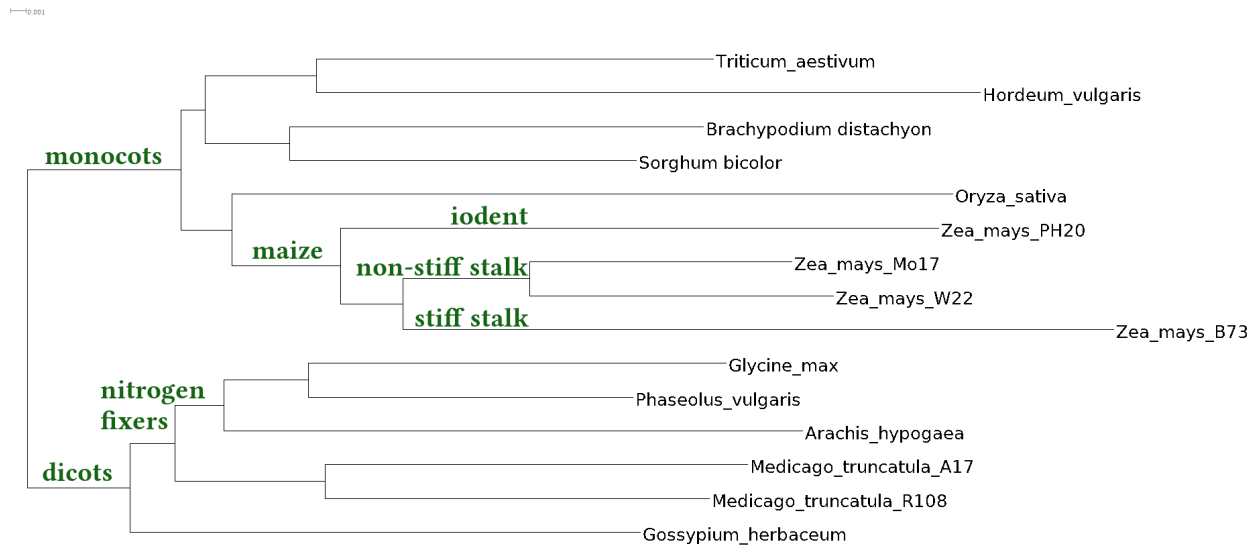



Figure 2: Phylogram built on the Jaccard distance matrix with Neighbor Joining. Manually rooted between monocots and dicots and text in green added.

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

## *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.

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

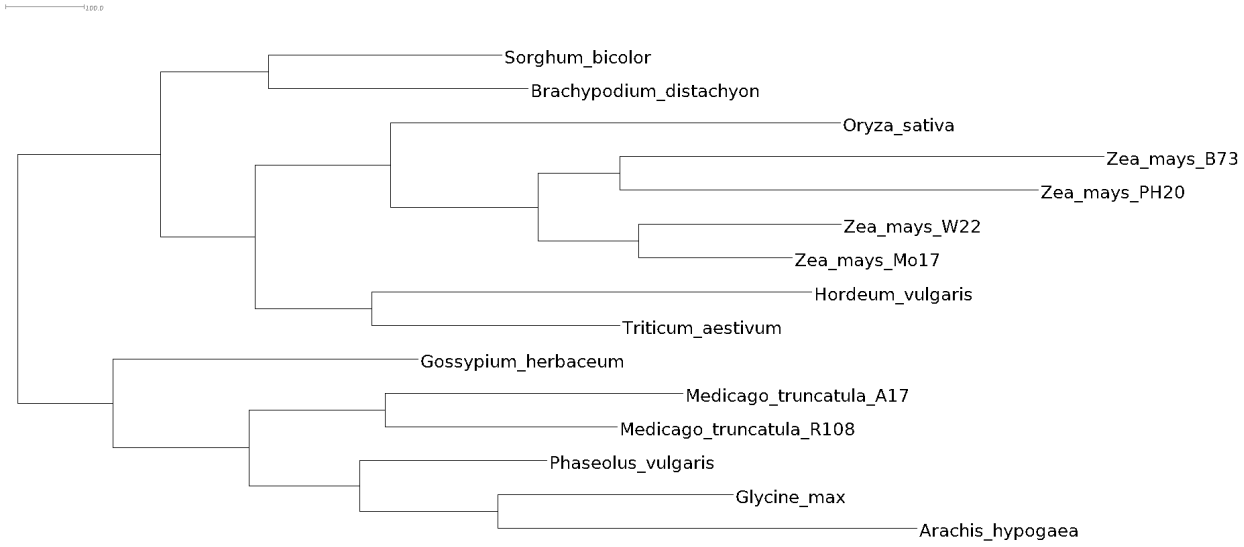



Figure 3: Phylogram built by looking for the maximally parsimonious tree (total of 7780 changes). Manually rooted between monocots and dicots.

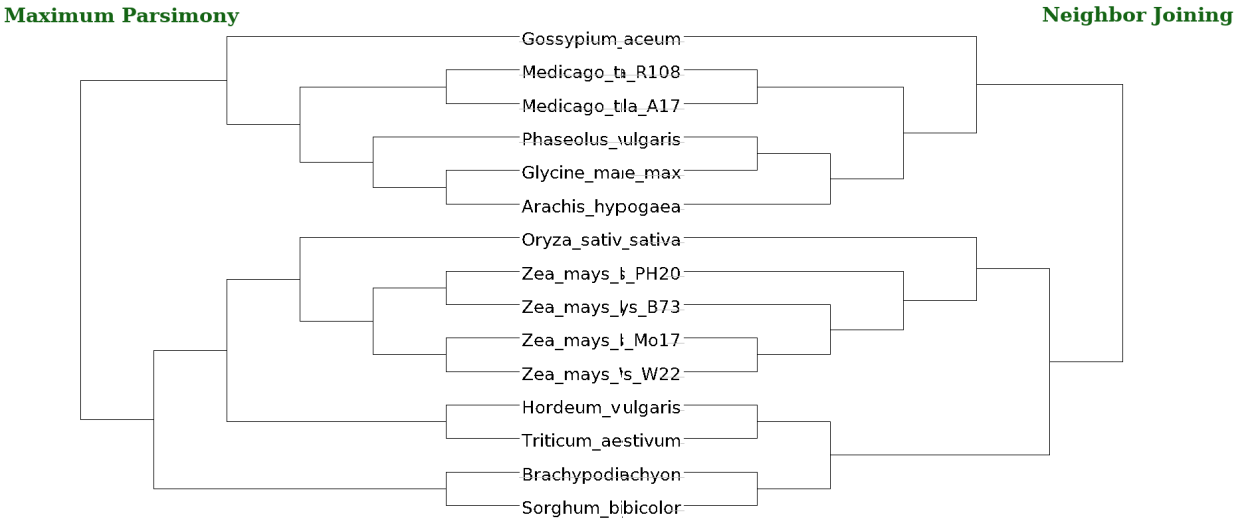



Figure 4: Tanglegram of the distance based and maximum parsimony tree (both manually rooted).

*Arachis* should switch their places. We are currently investigating the reason for this strange phenomenon and the best explanatioi we've come up with so far is that *Brachy* and *Medicago* are not actually representative 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].

It is difficult to come to a clear evolutionary tree within the same species (I am talking about maize[2]) 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.

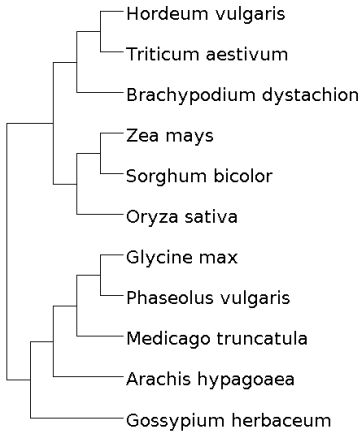



Figure 5: Expected cladogram according to sequence based phylogenetics. [2, 1, 3]

[2] 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.

### *Tracing Back the Signal*

Another big question is why does the tree actually look the way it does? Is it truly a biologically meaningful display of differences and similarities in function or is it just an artifactual tree, mainly 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?[3] 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:

[3] 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.

### *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`.