Wageningen

Dating ancient polyploidy events in angiosperms, seed plants and land plants

MSc Biology Thesis

Feia Matthijssen, 921009548030

Chair group: Biosystematics

Supervisors: Lars Chatrou & Setareh Mohammadin

3/25/2016

Summary

Whole genome duplications have taken place multiple times during plant evolution. The oldest of the so far studied duplication events are before the radiation of angiosperms and before the radiation of seed plants. These duplications have been studied by Jiao et al.[1] using an autocorrelated clock model. Model choices are thought to have an important effect on the outcome of the calculations. Therefore, in comparison, in this study I used an uncorrelated clock model to recalculate the age of these duplications using Bayesian statistics.

Orthogroups of protein coding sequences used by Jiao et al. (2011) to estimate the age of these duplications now were extended to contain sequences of in total 22 genomes. After which each orthogroup was analysed using an uncorrelated clock model available from the BEAST v2 package[2]. The age of three distinct genome duplications was estimated, a duplication in all angiosperms, a seed plant duplication and a land plant duplication.

In total 240 orthogroups showed significant results. In total 69 angiosperm duplications were found with an average of 309 Ma, 39 seed plant duplications at 353 Ma and 149 land plant duplications at 497 Ma.

This study shows evidence for a possible land plant wide genome duplication. A phenomenon that would most certainly deserve further investigation.

Introduction

Gene duplications are thought to have an important role in evolution[3,4]. When a single gene holds an important function in an organism, natural selection does not often permit mutation of that gene. After duplication these genes can mutate more freely. Therefore gene duplications may contribute to rapid speciation and radiation. They may allow for diversification of the genes, contributing to an increase in complexity[4,5]. Changes may contribute to improvements such as the enhancement of plant defence[6,7]. One of the most important gene duplication events is polyploidy, or whole genome duplication (WGD). Although our understanding is increasing, there are still many uncertainties concerning the mechanism behind evolution after gene duplication[8].

Several events of WGD have been studied, indicating that most plants have at least gone through one such duplication event[8]. The oldest of these genome duplications, one before the radiation of angiosperms and one before seed plant evolution, have been indicated by Jiao et al. (2011)[1]. In their study the age of the duplications has been estimated around 192±3 (95% confidence interval) and 319±2 Ma (figure 1). In the same paper, with more species involved in the same analysis, they already came to older age estimation (234±9 and 349±3 Ma). This shows that results can vary greatly depending on the data and methods used.

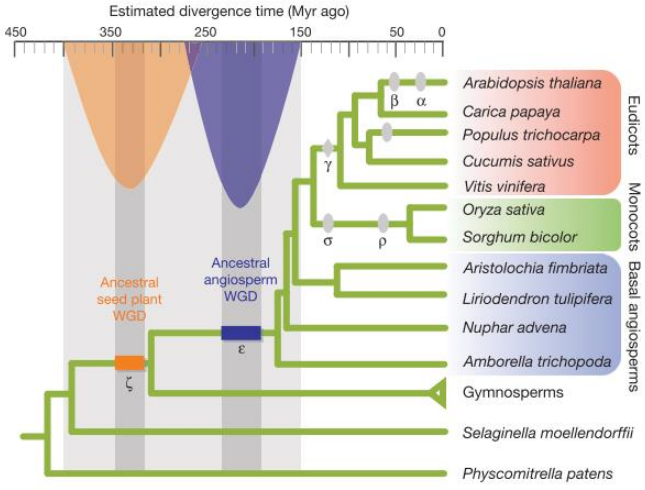

Figure 1. Several whole genome duplication events in seed plants (from Jiao et al. 2011)[1].

In this thesis molecular dating was used to date ancient gene duplications, of which the existence has been previously indicated by Jiao et al. (2011). To determine the timing of a proposed WGD, a molecular clock can be used. The molecular clock hypothesis, which was developed in the early 1960s[9,10], remains an important theory in molecular evolutionary biology. This theory states that the amount of substitutions is linear with time: the more two genes differ, the longer ago the divergence has taken place. This forms the basis of molecular dating. The theory became an important factor in the development of methods currently used in evolutionary studies[11]. The molecular clock hypothesis follows from the neutral theory, which states the importance of neutral substitutions and genetic drift for evolution[12]. The rate of change of the genetic code is relatively unaffected by the magnitude of selection, as

most substitutions that become fixed in the population are neutral.

Due to technological improvements the amount of genetic information is vastly increasing; new, faster and cheaper mechanisms of deciphering genetic material are being developed[13]. There is still much to discover around the mechanisms behind speciation, and the dating of such events. This vastly expanding pool of data allows for many more comprehensive studies to be done. However, even a large amount of data can give different conclusions, depending in the models used. This is for example clearly visible in a study by Pirie & Doyle (2012), where using a relaxed clock model one of their clades was estimated to be older than their outgroup[14].

Many models have been proposed to calculate the rate of evolution. This rate can be estimated as an interaction between time and evolution, often in the form of nucleotide substitutions. While the amount of substitutions can be measured, estimation of the rate of change remains a challenge. It is important to use a proper model to make a reliable estimation of the relation between substitution rate and elapsed time, as well as using proper model calibrations[15]. Two model types are used relatively often: uncorrelated and autocorrelated models[15]. In uncorrelated (UC) models it is assumed that rates can be drawn from a probability distribution, while autocorrelated (AC) models presume neighbouring branches to follow similar rates, and these rates can vary greatly between clades. More recently a random local clock (RLC) model has been proposed[16]. This model assumes a strict local clock that can vary among groups of branches.

Ho (2014) described sources of uncertainty in age estimations. They stated that with genomic analyses the age estimation largely depends on the used clock model and the model calibrations[17]. Jiao et al. (2011) assumed autocorrelation in substitution rates[1]. In a comparative study by Lepage et al. (2007) it was stated that autocorrelated models often outperform uncorrelated models[18]. Autocorrelation assumes that the primary determinant of rate variation is heritable[19]. As most mutations enter the population through reproduction, it is expected that the rate of evolution would be negatively correlated to generation length[20]. However, if taxa are only distantly related, the autocorrelation will be minimal. Therefore the distance between taxa may influence the effectiveness of AC models. More study is needed to fully understand this effect[21]. Uncorrelated models assume that substitution rates are randomly distributed. These assumptions may be violated in a scenario with for example a single woody clade amongst mostly herbaceous plants. In such a situation the assumption of rates being randomly distributed may have important effects on the outcome of the analysis. It was shown that the UC clock has a strong age bias when the assumption is violated that rates are randomly distributed[14,22,23]. Using a RLC this problem was not observed[22]. In calculations where the distance between taxa is large it may be better to use a method that uses both UC and AC models[21]. Unfortunately such a model is not yet available to general use.

Jiao et al.[1] used an AC clock model using the R8S[24] software package. Firstly they created phylogenetic trees using a maximum likelihood method using RAxML[25]. In these trees, calibration ages were assigned on five locations after which they were analysed using R8S to obtain age estimates for the gene duplication nodes. Either *Selaginella* or *Physcomitrella*, which diverged before the radiation of seed plants, was chosen as outgroup. As there is still much uncertainty around the assumption of autocorrelation, I have studied these duplications using an UC model. In my second analysis I have removed the assumption of *Selaginella* or *Physcomitrella* as outgroups, which allowed for results showing a duplication even before their divergence. For this an uncorrelated relaxed clock model using Bayesian statistics available in the BEAST v2 package[2] was used.

The aim of this study was to date the whole genome duplications at the base of seed plants and before the radiation of angiosperms. The results of this analysis may give more insight in these ancient duplications, as well as give more understanding on the effects of model choice.

Methods

Analysis 1

Rather than analysing the whole genome at once, WGD studies can benefit from the use of smaller genetic segments; partial sequence data can be used in a comparative approach[26]. In this study the genome

duplications were studied by analysing many gene duplications separately, and by combining these results.

A dataset containing 799 orthogroups was used, these are groups of sequences derived from a single ancestral gene. This data was previously used and made available by Jiao et al (2011)[1,27]. These orthogroups were created using protein coding sequences from fully sequenced genomes of the following species: two monocots (*Oryza sativa* and *Sorghum bicolor*), five eudicots (*Arabidopsis thaliana*, *Carica papaya*, *Populus trichocarpa*, *Cucumis sativus* and *Vitis vinifera*), one lycophyte (*Selaginella moellendorffii*), and one moss (*Physcomitrella patens*). Each group was filtered to contain at least one monocot, eudicot and *Selaginella* and/or *Physcomitrella* sequence. Among the 7,470 created, Jiao et al found 799 orthogroups containing the proposed duplication that occurred before the γ triplication[1]. They then enriched these orthogroups with sequences from four basal angiosperms and 16 gymnosperms. I used this final data set in my studies.

In order to estimate the time at which the WGDs took place the age of the gene duplications in each orthogroup was estimated separately. For this I used an uncorrelated relaxed clock model available in the BEAST 2 software package[2,28]. *Selaginella* and *Physcomitrella* were chosen as outgroups, assuming monophyly of all seed plants. One calibration was used, at the divergence of the seed plant group, with a normal distribution with an average of 425 and a 95% confidence interval of 400-450. This is similar to the calibrations used by Jiao et al (2011)[1,29]. Substitution rate and shape were set to estimate and substitution model was set to GTR, while leaving other settings at default. The analysis was done using a chain length of 25.000.000 generations. Using a custom python script the analysis of these orthogroups has been automated.

The results were filtered based on statistical support using Tracer v1.6. Analyses with ESS of <200 in posterior, prior or likelihood have been omitted, leaving 593 of 799 orthogroups. The outcomes of the analyses have been combined in order to gain an average age for the proposed genome duplications.

Analysis 2

Several recently published genomes have been added to the previously used orthogroups, in order to improve species coverage. These genomes include the *Aquilegia caerulea v3.1, Fragaria vesca v1.1*[30], *Solanum tuberosum v3.4*[31], *Volvox carteri v2.1*[32], *Chlamydomonas reinhardtii v5.5*[33] and *Ostereococcus lucimarinus v2.0*[34] made available at Phytozome 11, as well as *Phoenix dactylifera*[35] available from *genbank, a*nd the *Pinus lambertiana v1.0* and *Pseudotsuga menziesii v1.0* available from Dendrome. With this addition the orthogroups consist of sequences of 22 genomes.

In order to add these genomes a consensus sequence with a conservation threshold of 50% was made for each of the existing orthogroups. Using BLASTN[36] these consensus sequences were compared with the protein coding sequences of the previously mentioned genomes. Sequences with at least 50% overlap, a length of 0.5x to 2x the original length and an e-value threshold of 10^{-7} were added to the orthogroups. The enriched orthogroups were aligned using MAFFT[37]. Afterwards the ends were trimmed using a custom python script.

These enriched orthogroups were calibrated and analysed using the same settings as in analysis 1 but without the monophyly constraint on seed plants. The calibration point was placed on the divergence of *Selaginella* or *Physcomitrella* when *Selaginella* was not available. Only orthogroups that converged in the previous analysis were used. An initial analysis was done without any calibration point, in order to analyse the tree topology. This analysis was done using the same settings as the previous analysis, without monophyly constraint. The created trees were used to determine at which nodes the duplication(s) and calibration were present. After this point three duplications were recognised: (1) before the radiation of angiosperms; (2) before the radiation of seed plants; (3) before the divergence of *Selaginella* and *Physcomitrella*, being before the radiation of land plants.

Only 342 of the 593 orthogroups showed at least one clear duplication point in addition to a clear calibration point. Common problems were: there was very low (<70%) support for key nodes; *Selaginella* and/or *Physcomitrella* was nested within an angiosperm group making proper calibration impossible; no relevant duplication was found. Afterwards a second analysis was done in BEAST 2 with these 342 orthogroups in order to estimate the age of each of the found duplications.

Results

In the first analysis 593 orthogroups out of the 799 orthogroups from Jiao et all (2011)[1] (ESS>200) had either a duplication before the radiation of angiosperms (492 orthogroups) or before the radiation of seed plants (101 orthogroups). The angiosperm duplication showed a mean of 339 Ma (95% CI 202 - 438) while the seed plant duplication showed a mean of 370 Ma (95% CI 276 - 440). Both curves show a very strong left skew rather than a Normal distribution, with a mode around 400 Ma (figure 2).

In the second analysis 240 of 342 orthogroups showed an ESS>200. Among these we found 69 angiosperm duplications (309 Ma (95% CI 177 – 428)), 39 seed plant duplications (353 Ma (95% CI 246 – 438)) and 149 land plant duplications (497 Ma (95% CI 417 – 674)) (figure 3). In only 3 of 149 of the orthogroups with a land plant duplication *Selaginella* and/or *Physcomitrella* is present in both of the duplication copies.

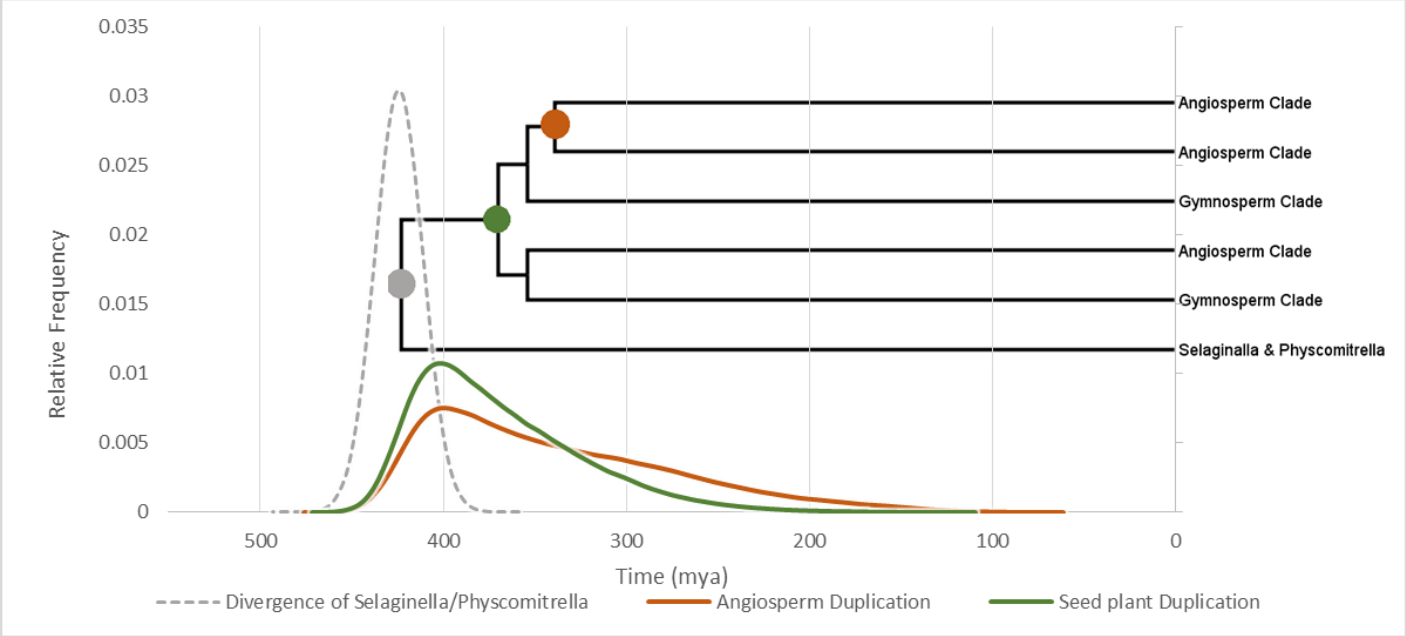

Figure 2. Relative frequency of age for the angiosperm duplication (492 orthogroups) and seed plant duplication (101 orthogroups), under the assumption of monophyly for seed plants. A drastically simplified example tree is shown, indicating the position of the duplication nodes.

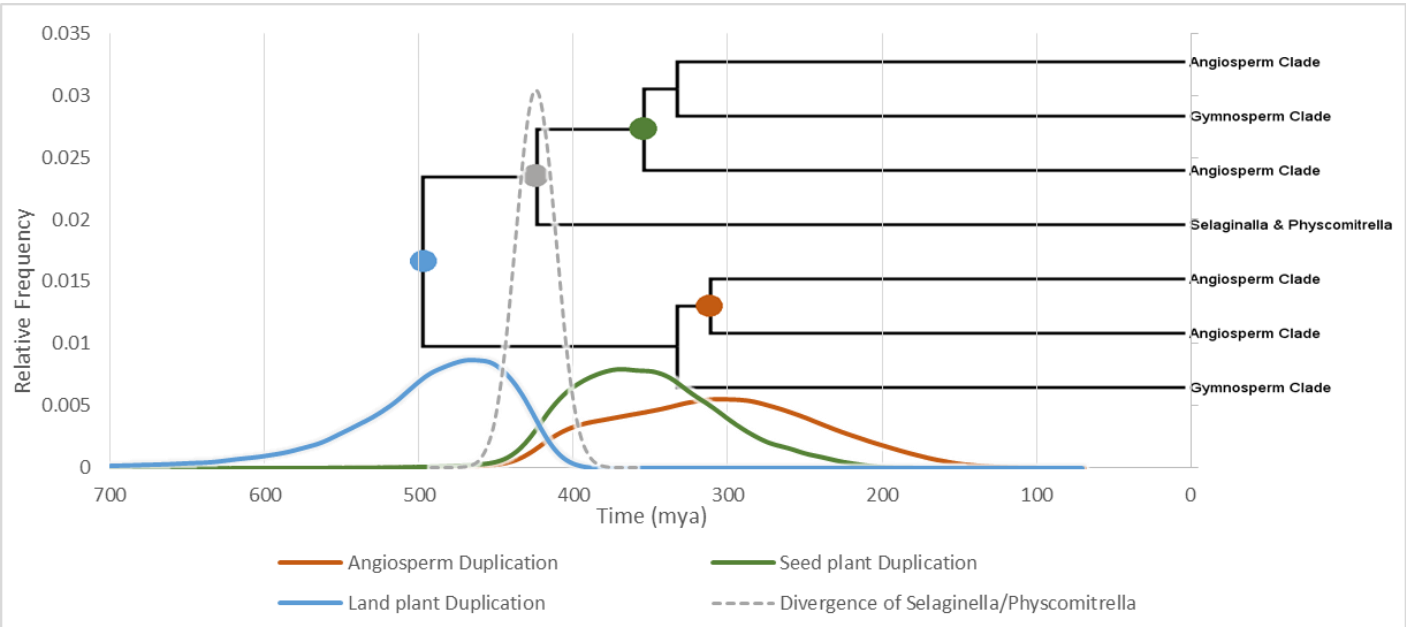

Figure 3. Relative frequency of age for the angiosperm duplication (69 orthogroups), seed plant duplication (39 orthogroups) and land plant duplication (149 orthogroups). A drastically simplified example tree is shown, indicating the position of the duplication nodes.

Discussion

The two duplications previously found by Jiao et al. were clearly present in the final results of this study. The average age of these duplications (309 for the angiosperm duplication and 353 for the seed plant duplication) was slightly different from the age that was found by Jiao et al. (234 and 349 respectively). The difference in age for the angiosperm duplication may be explained by the fact that using my methods it was not possible to separate seed plant and angiosperm duplications in orthogroups where seed plants were not present.

In analysis 1 the data showed a strong skew rather than a normal distribution. This may have been the result of assuming monophyly over all seed plants. The results of analysis 2, where a third duplication was found, also show that this initial assumption may have been incorrect. In the second analysis 149 of 240 orthogroups showed a duplication before the divergence of *Selaginella* and *Physcomitrella*. Assuming monophyly over seed plants, in these groups the nested *Selaginella* and *Physcomitrella* sequences are forcefully placed as outgroups. This results in the average difference between outgroup and duplication copy to become smaller. Resulting in a duplication age relatively close to the divergence of the outgroups.

Jiao et al. assumed *Selaginella* and *Physcomitrella* as outgroups to the data and did not find this skew in their age distribution. The third duplication point was also not found in their studies. This difference can be explained by several factors. First they assigned only one sequence as outgroup, while multiple *Selaginella* and/or *Physcomitrella* sequences can be present in one orthogroup. As a result only one of these is placed as outgroup, while the other sequences may still be placed elsewhere in the tree. This was observed at least several times. This drastically reduces the effect of the age of the node shifting towards its ancestral node. It also makes some trees ambiguous to interpret, as some of the *Selaginella* and *Physcomitrella* are still nested in the seed plant group while they are assumed outgroup. This phenomenon was not described in their paper[1]. A second possible explanation is that the trees used in their analysis were made prior to the age calculations. Afterwards the R8S software assumes the tree topology correct, and assigns rates and ages to the tree, using many different calibration points. Due to this it may be more likely to gain a normal distribution of ages of the duplication node.

In only few (3 of 149) of the orthogroups with a land plant duplication *Selaginella* or *Physcomitrella* is present in both of the duplication copies. It is quite common that only part of the species is present in both duplication copies, but even that would not explain the very low number found. It is possible however that *Selaginella* and *Physcomitrella* genes were lost more often than other genes after duplication. Therefore the addition of more non-seed plant genomes may give more insight in this phenomenon. It may also be interesting to investigate which species retained genes more often than other species.

In a case with Annonaceae there are two distinct clades, where there is one clade which showed a much higher mutation rate than its neighbouring clade[14]. When this group is analysed using an uncorrelated relaxed clock the group with the long branches was placed even older than the outgroup. To exclude the possibility of this effect the trees made using the uncorrelated relaxed clock were compared to the trees published by Jiao et al., where this model was not used. There was no clear difference found in the branch lengths between these methods.

It occurred multiple times that the gymnosperm clade was placed as sister to *Selaginella* and *Physcomitrella*, rather than as sister to angiosperms. This resulted in a clade containing *Selaginella, Physcomitrella* and Gymnosperms, as a sister to angiosperms. This may be explained by the very old age of the common ancestor of land plants, making the evolutionary relation difficult to reconstruct. This phenomenon may have had an effect on the estimated age of the land plant duplications, as proper calibration of the divergence of *Selaginella* and *Physcomitrella* is made more difficult. Gymnosperms were only available in part of the orthogroups, while Angiosperms were available in all orthogroups. In order to maintain consistency in the calibration of the divergence of *Selaginella* and *Physcomitrella* the calibration has been placed at their most recent common ancestor with angiosperms rather than with gymnosperms.

The addition of the three green algae genomes (*Volvox carteri, Chlamydomonas reinhardtii* and *Ostereococcus lucimarinus*) did not prove useful in this study. In many orthogroups they were not present. If they were present they were often nested in other groups, making their evolutionary meaning ambiguous.

This study has provided results indicating a whole genome duplication for all or most land plants. To give more resolution to the dating of the duplications, and at what point in evolution it has taken place, a future study may benefit greatly from the addition of more genomes. Addition of non-seed plants may increase the understanding of the position of *Selaginella* and *Physcomitrella* in these duplication events. Addition of Gymnosperms may ease the distinction between the seed plant duplication and the angiosperm duplication. It may also be beneficial to investigate these duplications using other methods to cross validate the existence of this land plant wide genome duplication.

References

1 Jiao, Y. *et al.* Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97-100, doi:10.1038/nature09916 (2011).

2 Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* **10**, e1003537, doi:10.1371/journal.pcbi.1003537 (2014).

3 Ohno, S. *Evolution by Gene Duplication*. (Springer Berlin Heidelberg, 1970).

4 Zhang, J. Evolution by gene duplication: an update. *Trends in ecology & evolution* **18**, 292-298, doi:http://dx.doi.org/10.1016/S0169-5347(03)00033-8 (2003).

5 Edger, P. & Pires, J. C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research* **17**, 699-717, doi:10.1007/s10577-009-9055-9 (2009).

6 Hofberger, J. A., Lyons, E., Edger, P. P., Chris Pires, J. & Eric Schranz, M. Whole Genome and Tandem Duplicate Retention Facilitated Glucosinolate Pathway Diversification in the Mustard Family. *Genome Biology and Evolution* **5**, 2155-2173, doi:10.1093/gbe/evt162 (2013).

7 Hofberger, J. A., Nsibo, D. L., Govers, F., Bouwmeester, K. & Schranz, M. E. A Complex Interplay of Tandem- and Whole-Genome Duplication Drives Expansion of the L-Type Lectin Receptor Kinase Gene Family in the Brassicaceae. *Genome Biology and Evolution* **7**, 720-734, doi:10.1093/gbe/evv020 (2015).

8 Adams, K. L. & Wendel, J. F. Polyploidy and genome evolution in plants. *Current opinion in plant biology* **8**, 135-141, doi:http://dx.doi.org/10.1016/j.pbi.2005.01.001 (2005).

9 Zuckerkandl, E. & Pauling, L. in *Horizons in Biochemistry* 189-225 (Academic Press, 1962).

10 Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* **97**, 97-166 (1965).

11 Morgan, G. Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959–1965. *Journal of the History of Biology* **31**, 155-178, doi:10.1023/a:1004394418084 (1998).

12 Kimura, M. Evolutionary Rate at the Molecular Level. **217**, 624-626 (1968).

13 Metzker, M. L. Sequencing technologies - the next generation. **11**, 31-46 (2010).

14 Pirie, M. D. & Doyle, J. A. Dating clades with fossils and molecules: the case of Annonaceae. *Botanical Journal of the Linnean Society* **169**, 84-116, doi:10.1111/j.1095-8339.2012.01234.x (2012).

15 Duchene, S., Lanfear, R. & Ho, S. Y. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Molecular phylogenetics and evolution* **78**, 277-289, doi:10.1016/j.ympev.2014.05.032 (2014).

16 Drummond, A. & Suchard, M. Bayesian random local clocks, or one rate to rule them all. *BMC Biology* **8**, 114 (2010).

17 Ho, S. Y. The changing face of the molecular evolutionary clock. *Trends in ecology & evolution* **29**, 496-503, doi:10.1016/j.tree.2014.07.004 (2014).

18 Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. A General Comparison of Relaxed Molecular Clock Models. *Molecular Biology and Evolution* **24**, 2669-2680, doi:10.1093/molbev/msm193 (2007).

19 Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol* **4**, e88, doi:10.1371/journal.pbio.0040088 (2006).

20 Smith, S. A. & Donoghue, M. J. Rates of Molecular Evolution Are Linked to Life History in Flowering Plants. *Science* **322**, 86-89, doi:10.1126/science.1163197 (2008).

21 Ho, S. Y. W. An examination of phylogenetic models of substitution rate variation among lineages. *Biology Letters* **5**, 421-424, doi:10.1098/rsbl.2008.0729 (2009).

22 Crisp, M. D., Hardy, N. B. & Cook, L. G. Clock model makes a large difference to age estimates of long-stemmed clades with no internal calibration: a test using Australian grasstrees. *BMC evolutionary biology* **14**, 263 (2014).

23 Beaulieu, J. M., O'Meara, B., Crane, P. & Donoghue, M. J. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. (2015).
24 Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics (Oxford, England)* **19**, 301-302, doi:10.1093/bioinformatics/19.2.301 (2003).
25 Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics (Oxford, England)*, doi:10.1093/bioinformatics/btu033 (2014).
26 Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**, 433-438, doi:http://www.nature.com/nature/journal/v422/n6930/suppinfo/nature01521_S1.html (2003).
27 Jiao, Y. *et al.* (Dryad Data Repository, 2011).
28 Drummond, A. J. & Bouckaert, R. R. *Bayesian evolutionary analysis with BEAST 2*. (Cambridge University Press, 2015).
29 Rensing, S. A. *et al.* The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. *Science* **319**, 64-69, doi:10.1126/science.1150646 (2008).
30 Shulaev, V. *et al.* The genome of woodland strawberry (Fragaria vesca). *Nature genetics* **43**, 109-116, doi:10.1038/ng.740 (2011).
31 Xu, X. *et al.* Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-195, doi:10.1038/nature10158 (2011).
32 Prochnik, S. E. *et al.* Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* **329**, 223-226, doi:10.1126/science.1188800 (2010).
33 Merchant, S. S. *et al.* The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* **318**, 245-250, doi:10.1126/science.1143609 (2007).
34 Palenik, B. *et al.* The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 7705-7710, doi:10.1073/pnas.0611046104 (2007).
35 Al-Mssallem, I. S. *et al.* Genome sequence of the date palm Phoenix dactylifera L. *Nat Commun* **4**, doi:10.1038/ncomms3274 (2013).
36 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).
37 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).