**Title: The evolutionary dynamics of ploidy evolution in plants**

**1. Scientific Background**

This fellowship focuses on expanding our understanding of the evolutionary dynamics of polyploidy (whole genome duplication; WGD) and its relationship to the evolution of agriculturally important crop species. Analysis of early cytogenetic work revealed that polyploidy has occurred relatively recently in the evolutionary history of many of our important crop species [1, 2]. More recent genomic analyses have revealed that in fact all angiosperms have experienced at least one ancient WGD event and many lineages appear to have undergone repeated rounds of genome duplication [3-5]. In fact, some crops like wheat are characterized by diploid, tetraploid and even hexaploid cultivated varieties [6]. Polyploidy is particularly important in agriculture since it produces characteristics that can be leveraged to improve crops. For instance, polyploidy can increase the dosage of important genes, create permanent heterozygosity, or create duplicate gene “back-ups” that may buffer against mutations in genes of crucial functions [7]. Despite the widespread occurrence of polyploidy and the central role that it has played in the modification of crop species we have a limited understanding of its impact on the evolution of other traits, and its role in diversification is widely debated [8, 9].

Polyploidy has been suggested as a key explanation for the hyper diversity of several groups of angiosperms [10-12]. However, analyses using likelihood based evolutionary models suggest that polyploids actually have lower diversification rates [13]. Polyploidy also has profound impacts on the content of a genome; following WGD some gene families are preferentially retained as duplicates while others are quickly lost [14-17]. For example, several studies have suggested that genes involved in core cellular processes and housekeeping functions are "duplication resistant" being consistently restored to a single-copy status following independent WGD events [15, 17]. A number of probabilistic models, based on a birth-death process, have been developed to describe changes in gene-family size [18-24]. While these models allow the examination of changes in the size of gene families across a phylogeny in a comprehensive and unbiased manner, they do not integrate the history of WGD into the model of gene family evolution. Surprisingly, attempts to account for the occurrence of WGD within a probabilistic inference framework have been rare (but see [25] for non-phylogenetic approaches). The central problem that limits our understanding of how ploidy interacts with other traits is caused by a lack of probabilistic models that explicitly incorporate ploidy and other traits.

During the last several years, the Mayrose lab at Tel Aviv University has developed key tools for the investigation of ploidy evolution. The chromEvol probabilistic inference framework [26, 27] is currently

considered the state-of-the-art ,methodology to analyze patterns of chromosome number change along a phylogeny, allowing researchers to infer polyploidy and dysploidy (i.e., single chromosome number changes) event along specific lineages. Building upon my experience from my PhD research [28-32], and the basic model of ploidy evolution in chromEvol, during my post-doctoral studies I will further develop more realistic models of polyploidy evolution. I aim to specifically develop a model of the evolution of ploidy level (rather than choromsome numbers) and to integrate into the inference framework data from additional sources (e.g., the evolution of gene-family size) in addition to chromosome number that will increase the reliability of inference.

**2. First Year Research objectives**

**2.1** Develop a novel set of models for the evolution of ploidy levels. The models will be implemented within the state dependent speciation and extinction framework and will integrate other sources of ploidy inference beyond chromosome number. This will allow us to infer the ploidy level at each point along a phylogeny and to explore the long-term effect of polyploidy on speciation and extinction in a more robust fashion.

**2.2** Develop a novel model for gene-family size evolution that accounts for WGD. Using large-scale genomic data from multiple crop and wild species, we will identify gene families that exhibit distinct gene loss and retention patterns following independent WGD.

**3. Detailed description of the proposed research**

**3.1 Ploidy evolution model**

The first step to an improved understanding of ploidy evolution is building a richer model for inference. The model that we will construct has two important advantages over existing approaches. First, we will explicitly model the probability of a lineage being either diploid or polyploid. Second, our model will allow the integration of evidence from sources other than chromosome number. Briefly, the basic model of chromosome number evolution as implemented in chromEvol [27] has three primary parameters $\rho$, $\lambda$, and $\mu$; denoting: rates of polyploidization, ascending and descending dysploidy, respectively. Here, I suggest a model that accounts also for the ploidy level of a lineage at each point in time. The suggested model has 6 parameters: $\rho$, $d$, $\lambda^1$, $\lambda^2$, $\mu^1$, and $\mu^2$; $p$ and $d$ parameters denote rates of polyploidization and diploidization (i.e., transition from a polyploid state to a diploid one); $\lambda$ and $\mu$ still represent ascending and descending dysploidy but we now allow them to take different values depending on the ploidy level of the lineage ($\lambda^1$ and $\mu^1$when a lineage is diploid and $\lambda^2$ and $\mu^2$ when a lineage is polyploid). We can then use a model comparison framework such as the likelihood ratio test to perform hypothesis testing. For example, comparing a model where we constrain $\mu^1$ and $\mu^2$ to be equal versus one where we allow $\mu^1$ and $\mu^2$ to vary

would allow us to test whether the rate of descending dysploidy is higher in a polyploid compared to a diploid background as has been suggested [33]. In addition, in the current chromEvol modeling approach, inference regarding the ploidy levels of tip taxa is not an integral part of the model, and can only be performed in an ad hoc manner [26]. In my post-doctoral studies I aim to develop a model that will map onto the phylogeny the duration of time a lineage was in a polyploid state and in a diploid state. Importantly, in the original chromEvol model, the phylogeny and the chromosome numbers of extant taxa are the only data input. With the suggested approach other sources of evidence regarding the ploidy level of extant taxa can be integrated (e.g., data regarding meiotic pairing behavior, or the age of gene duplicates), thus improving the quality of our inference. This model will allow us to calculate the posterior probability of being at each ploidy state at each tip and internal node of the phylogeny. Stochastic mapping can use these probabilities to reconstruct possible histories of ploidy [34]. These sampled histories of ancestral ploidy states can readily be integrated within a range of related phylogenetic models, allowing polyploidy-associated phenomena to be explicitly examined. For example, the effects of polyploidy on rates of sequence evolution can readily be investigated by integrating the suggested model within the traitRate model [35]. Finally, this approach offers a more robust approach to answering the long-standing questions of the long-term effect of whole genome duplication on diversification.

**3.2 Polyploidy and gene family evolution**

Polyploidy has profound impacts on the transcriptome, proteome, and metabolome, as well as on the rate of genome evolution [reviewed in 22,31,72]. During the last decade, studies in organisms as diverse as fungi, fish, and plants have revealed that following WGD, some gene families are preferentially retained as duplicates while others become extremely volatile [21,73–83], and these tendencies were found to be strongly associated with specific functions. For example, several studies have identified a set of "duplication resistant" gene families that have been consistently restored to a single-copy status following independent WGD events, with an increased tendency of these genes to be involved in core cellular processes and housekeeping functions [74–76]. Clearly, this subset of genes represents the extreme end of genes that are preferentially lost following WGD, and is the one most easily detected, as many other gene families are duplication resilient but their pre-duplication size is larger than one. To quantify the full repertoire of changes that WGDs have induced on particular gene families, the processes that govern the evolution of gene families should be incorporated within an explanatory evolutionary model and analyzed within a phylogenetic framework that would further account for the number and timing of different WGD events.

Once the ploidy-evolution model (section 3.1) is implemented, it will allow me to develop a novel probabilistic model to test whether changes in particular gene families are associated with the ploidy level of a lineage. To this end, the model will accept as input the number of genes in each gene family for each of the species being analyzed. In the simplest scenario, changes in the numbers of genes in a family are governed by rate parameters $\lambda$ and $\delta$, representing the rates of single gene gains and losses, respectively. To allow rates of gene-family evolution to be dependent on the ploidy level, the proposed model accepts as an additional input the ploidy level of a lineage at each point along the phylogeny, as obtained through the stochastic mapping approach (section 3.1). This will allow several important modifications to previous models of gene-family evolution. First, a transition to a higher ploidy level instantly multiplies the number of members of each gene family. Second, since ploidy-level transitions may also induce changes in gene-family dynamics, we also allow rates of gene gain/loss to be determined by the ploidy state of a lineage. Thus, assuming two ploidy levels, the model represents a variant of a standard birth-death model and requires a total of four free parameters.

Given an input phylogeny, and a sample of ploidy-level mappings, the proposed model will be used to pinpoint gene families or functional classes most affected by polyploidy events involving different lineages. Of particular interest will be whether the same gene families in crop species show correlated responses to WGD events. These analyses will be performed on several large clades with substantial genomic data and chromosome-number information. Classifications of genes into gene families will be based on orthology assignments obtained through public databases [36-38]. I expect the use of the proposed method to become increasingly powerful as sequence data accumulates, which will further enable me to compare the consequences of polyploidy on gene-family dynamics across multiple plant clades.

**4. Summary**

This proposal describes exciting probabilistic phylogenetic methodologies that will provide the tools to understand plant evolution in a way that simply is not possible today. Here, I have described two achievable objectives for the first year of the fellowship. The second year of the fellowship will focus on extensions of the described methods and analyses of empirical data. All tools developed will be implemented in open-software environments, enhancing their accessibility to the research community and allowing a range of related questions to be investigated. I believe that current proposal offers an important path to a better understanding of central processes in the evolution of the genomes of crops and all plants.

Bibliography


1. Stebbins, G.L., *Chromosomal evolution in higher plants*. Chromosomal evolution in higher plants. 1971, London, UK: Edward Arnold.
2. Stebbins, G.L., *Variation and evolution in plants*. 1950, London, UK: Oxford University Press.
3. Cui, L., et al., *Widespread genome duplications throughout the history of flowering plants.* Genome research, 2006. **16**(6): p. 738-749.
4. Seoighe, C. and C. Gehring, *Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome.* Trends in Genetics, 2004. **20**(10): p. 461-464.
5. Tang, H., et al., *Angiosperm genome comparisons reveal early polyploidy in the monocot lineage.* Proceedings of the National Academy of Sciences, 2010. **107**(1): p. 472-477.
6. Marcussen, T., et al., *Ancient hybridizations among the ancestral genomes of bread wheat.* Science, 2014. **345**(6194): p. 1250092.
7. Udall, J.A. and J.F. Wendel, *Polyploidy and crop improvement.* Crop Science, 2006. **46**(Supplement_1): p. S-3-S-14.
8. Mayrose, I., et al., *Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al.(2014).* New Phytologist, 2015. **206**(1): p. 27-35.
9. Soltis, D.E., et al., *Are polyploids really evolutionary dead‐ends (again)? A critical reappraisal of Mayrose et al.(2011).* New Phytologist, 2014. **202**(4): p. 1105-1117.
10. De Bodt, S., S. Maere, and Y. Van de Peer, *Genome duplication and the origin of angiosperms.* Trends in Ecology & Evolution, 2005. **20**(11): p. 591-597.
11. Soltis, D.E., et al., *Polyploidy and angiosperm diversification.* American journal of botany, 2009. **96**(1): p. 336-348.
12. Tank, D.C., et al., *Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications.* New Phytologist, 2015. **207**(2): p. 454-467.
13. Mayrose, I., et al., *Recently formed polyploid plants diversify at lower rates.* Science, 2011. **333**(6047): p. 1257-1257.
14. Chapman, B.A., et al., *Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication.* Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(8): p. 2730-2735.
15. De Smet, R., et al., *Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.* Proceedings of the National Academy of Sciences, 2013. **110**(8): p. 2898-2903.
16. Guo, Y.L., *Gene family evolution in green plants with emphasis on the origination and evolution of Arabidopsis thaliana genes.* The Plant Journal, 2013. **73**(6): p. 941-951.
17. Paterson, A.H., et al., *Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon.* Trends in Genetics, 2006. **22**(11): p. 597-602.
18. Ames, R.M., et al., *Determining the evolutionary history of gene families.* Bioinformatics, 2012. **28**(1): p. 48-55.
19. De Bie, T., et al., *CAFE: a computational tool for the study of gene family evolution.* Bioinformatics, 2006. **22**(10): p. 1269-1271.

20. Han, M.V., et al., *Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3.* Molecular biology and evolution, 2013. **30**(8): p. 1987-1997.
21. Huson, D.H. and M. Steel, *Phylogenetic trees based on gene content.* Bioinformatics, 2004. **20**(13): p. 2044-2049.
22. Karev, G.P., Y.I. Wolf, and E.V. Koonin, *Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve?* Bioinformatics, 2003. **19**(15): p. 1889-1900.
23. Librado, P., F. Vieira, and J. Rozas, *BadiRate: estimating family turnover rates by likelihood-based methods.* Bioinformatics, 2012. **28**(2): p. 279-281.
24. Liu, L., et al., *A Bayesian model for gene family evolution.* BMC bioinformatics, 2011. **12**(1): p. 426.
25. Maere, S., et al., *Modeling gene and genome duplications in eukaryotes.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(15): p. 5454-5459.
26. Glick, L. and I. Mayrose, *ChromEvol: assessing the pattern of chromosome number evolution and the inference of polyploidy along a phylogeny.* Molecular biology and evolution, 2014. **31**(7): p. 1914-1922.
27. Mayrose, I., M.S. Barker, and S.P. Otto, *Probabilistic models of chromosome number evolution and the inference of polyploidy.* Systematic Biology, 2010. **59**(2): p. 132-144.
28. Blackmon, H. and J.P. Demuth, *The fragile Y hypothesis: Y chromosome aneuploidy as a selective pressure in sex chromosome and meiotic mechanism evolution.* BioEssays, 2015. **37**(9): p. 942-950.
29. Blackmon, H. and J.P. Demuth, *Coleoptera Karyotype Database.* The Coleopterists Bulletin, 2015. **69**(1): p. 174-175.
30. Blackmon, H. and J.P. Demuth, *Genomic origins of insect sex chromosomes.* Current Opinion in Insect Science, 2015. **7**: p. 45-50.
31. Blackmon, H., N.B. Hardy, and L. Ross, *The evolutionary dynamics of haplodiploidy: Genome architecture and haploid viability.* Evolution, 2015. **69**(11): p. 2971-2978.
32. Ross, L., et al., *Recombination, chromosome number and eusociality in the Hymenoptera.* Journal of evolutionary biology, 2015. **28**(1): p. 105-116.
33. Lysak, M.A., *Live and let die: centromere loss during evolution of plant chromosomes.* New Phytologist, 2014. **203**(4): p. 1082-1089.
34. Huelsenbeck, J.P., R. Nielsen, and J.P. Bollback, *Stochastic mapping of morphological characters.* Systematic Biology, 2003. **52**(2): p. 131-158.
35. Mayrose, I. and S.P. Otto, *A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution.* Molecular biology and evolution, 2010: p. msq263.
36. Goodstein, D.M., et al., *Phytozome: a comparative platform for green plant genomics.* Nucleic acids research, 2012. **40**(D1): p. D1178-D1186.
37. Matasci, N., et al., *Data access for the 1,000 Plants (1KP) project.* GigaScience, 2014. **3**(1): p. 1-10.
38. Proost, S., et al., *PLAZA: a comparative genomics resource to study gene and genome evolution in plants.* The Plant Cell, 2009. **21**(12): p. 3718-3731.