The Evolutionary History and Polyploid Origin of Coast Redwood (*Sequoia sempervirens*)

By

Alison Dawn Scott

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Botany)

at the

UNIVERSITY OF WISCONSIN – MADISON

2017

Date of final oral examination: 05/05/2017

The dissertation is approved by the following members of the Final Oral Committee:
David A. Baum, Professor, Botany
Cécile Ané, Professor, Statistics & Botany
Kenneth M. Cameron, Professor, Botany
Kenneth J. Sytsma, Professor, Botany
Donald M. Waller, Professor, Botany

Dedication

This dissertation is dedicated to the women who came before me

& the women whose time is yet to come.

To Sydney, Natalia, and Kendall – you can have it your way, loves.

Acknowledgements

Money talks, and I am fortunate to have many generous supporters to acknowledge. Thank you to the National Science Foundation's Graduate Research Fellowship and Nordic Research opportunity for supporting me. Thanks also to the UW-Madison AOF program for their support. My research would not have been possible without Save-the-Redwoods-League taking a chance on a young researcher and awarding me my first grant. Last but certainly not least, thank you to the UW – Madison Department of Botany for financial support in so many forms, from fellowships to travel funding.

I owe massive thanks to my committee for their expertise and support over the years. Don Waller, your work on selfing in *Impatiens* is what motivated me to apply to this department. In a way you are responsible for this whole thing, and I thank you for that. Cécile Ané, I admire your patience and professionalism and I appreciate both immensely. Thank you for always being willing to sit down and help me understand things, and for never treating any question as too simple. Ken Cameron, thank you for all of the engaging discussions, good laughs, and your continued encouragement. Ken Sytsma, your enthusiasm and engagement are infectious. Thank you for being dedicated, inquisitive, and eager to not only teach, but to learn alongside your students.

David Baum, working with you has been a great privilege. Your unfailing patience and persistent curiosity are more than a student could hope for in an advisor. Your inspiration and support made this all possible. If I manage to project a shred of the diplomacy you regularly demonstrate, I'll be proud. Thank you for sharing so much with me over the years. Especially the stories about missed flights and forgotten computer chargers – those remind me that you are in fact human.

Special thanks to the J.F. Crow Institute for the Study of Evolution, for engaging seminars, journal clubs, and happy hours. In particular, thank you to Chris Todd Hittinger for allowing me to attend lab meetings while David was on sabbatical. Thank you to the Friends of Martin journal club (Martin Bontrager, Quinn Langdon, David Loehlin, and others) for many fruitful discussions, especially regarding segregation distortion and holocentric chromosomes.

Thank you to the Department of Botany staff and administrators who have patiently helped me out and answered my questions over the years. Mo Fayyaz, thank you for letting me grow giant redwoods in your greenhouse and for procuring odd conifers for me over the years. Thank you for letting us throw Halloween FAC in the greenhouse, and for never letting me live it down.

To my lab mates, old and new: thank you for your companionship, commiseration, and courage. Abby Mazie and Talline Martins, you took me under your wings and prepared me for so much. Thank you for your guidance, and for teaching me how to read David's moods. Nisa, I love that

you're just as temperamental as I am. Thank you for always leveling with me even when we're both having a terrible day. Mike, thanks for all the dry ice, your impressive impersonations during parties, and for teaching me how to properly dunk an Oreo. Melody, thank you for your directness and for your sense of humor.

A full keg's worth of gratitude to all FAC czars past, present, and future, but especially to the inimitable Andy Gardner. You're always FAC czar in my heart. Thank you for showing me the ropes, and for being a role model of How to Be a Good Grad Student & Nice Person.

Chloe, I appreciate you so much. I was worried I'd be all alone in Madison, and was lucky to instead find an incredible friend in you. Thank you for letting me sleep on your couch, letting me tag along on fieldwork, and sharing scotch with me. Sláinte!

Kate & Ken, co-founding members of B.A.S.H. (Botanists Against Sexual Harassment)-
Kate McCulloh, I knew from the moment I met you that we would be friends. Or at least I hoped we would be. Setting aside your horrible taste in candy, I admire you. Thank you for all the listening, the love, and the laughs. Thanks also to Matt for the chili, and T & C for so many hugs. KKR, it turns out all the Americans in Umeå WERE from the same city. Thanks for rising above 3.2 and keeping true to the spirit of independence.

Blanket thanks to my non-academic loved ones in Madison – you are legion, and I adore you. Thank you for reminding me I'm more than my job. Much gratitude in particular to my fellow Spirited Women for helping me blaze trails.

Greg, we are absolutely not friends but I love you a lot. Thank you for the rillettes, pep talks, and for not getting too jealous of Gunner. You make me want to be better at everything I do.

Special thanks to friends that are close to my heart but far away: Gustavo, my twin heart, multi-lingual physicist and all around gato – saudades. Alex Wartelle, you're the best sous chef a person could ever hope for. Brandi Jo Petronio and Amber Paasch, you two keep me afloat. Brandi, thank you for being my forever ride-or-die. Amber, thank you for helping me figure out how to be a person.

And to my family - Miss Paula, I know you wanted a grandbaby but this will have to do. You're the original independent woman in my life, and I thank you for showing me how strong a person can be. Preston, I still don't know why you opted in to all this mess but I'm sure glad you did. Thank you for teaching me how to read a map, and then showing me the world. Allen, John, & Todd – thanks to you three I grew up thinking I was tough and could punch above my weight. I still try. Dan, climate change is real and margaritas shouldn't come from a bottle but I love you anyway. Syd, Nat, and Kendall, it's all for you.

Table of Contents

Dissertation Summary …….....1

Chapter 1: On the Evolutionary History of the Coast Redwood ……4

Chapter 2: Whole genome duplication in coast redwood (*Sequoia sempervirens*) and its implications for explaining the rarity of polyploidy in conifers ……...24

Chapter 3: Polyploid Evolution in *Sequoia sempervirens*…….……..56

Dissertation Summary

The majestic coast redwood (*Sequoia sempervirens*) is a California endemic of economic, ecological, and cultural value. These stately trees are best known for their role in the emergence of the American conservation movement, their staggering height, and their longevity. Coast redwoods are also noteworthy due to their genomic composition: despite the vast majority of conifers being diploid, *Sequoia* is a hexaploid (2n = 6x = 66). Here, I use a combination of molecular sequence data and insights gained from the fossil record to shed light on the evolutionary history of this charismatic lineage

Chapter one reviews the fossil, cytogenetic, and phylogenetic literature on *Sequoia*, providing background for subsequent chapters. Despite the restricted current distribution of redwoods, the fossil record shows a variety of redwood-like lineages scattered across the Northern Hemisphere. This historical overlap allows that hybridization among redwood lineages could have contributed to polyploidy in coast redwood. Sequoioid fossils, which date back to the Jurassic, often include imprints of foliage, permitting the observation and measurement of epidermal morphological characters, including guard cell size. I interpret fossil guard cell data (a morphological proxy for genome size) as a proxy for historical ploidy in redwoods. Based on this line of evidence, we assume that *Sequoia* has been polyploid since at least 34Ma.

Chapter two reports on analyses based on *de novo* transcriptome sequence data for *Sequoia*, *Sequoiadendron*, *Metasequoia*, and an outgroup. We used Bayesian concordance analysis and K_s estimates to test hypotheses about the origin of polyploidy in *Sequoia*, assessing whether

hybridization with an extant lineage played a role. Phylogenetic analyses rule out the possibility that *Sequoia* resulted from hybridization outside the Californian redwood clade (*Sequoia* and *Sequoiadendron*) and further suggest that *Sequoia* is an autopolyploid. Quantitative analysis of sequence divergence between putatively homeologous sequences in *Sequoia* showed that the rate of synonymous substitutions per synonymous site is remarkably low within *Sequoia*. This finding seems to conflict with the fossil data, which suggests an ancient polyploidization event. However, these findings can be reconciled by suggesting that the *Sequoia* lineage has experienced extensive homeologous recombination since the time of whole genome triplication.

Although chapter two provided many new insights into the evolutionary history of *Sequoia*, the conclusions were confounded by the fact that the data come from transcriptomes. There is some possibility, therefore, that biased expression of the gene copies from different parents could distort the results, for example by obscuring one parental donor. In order to rule out this possibility, chapter three focused on genomic sequence data. This was accomplished using hybrid sequence capture to enrich a set of targeted genes identified from the transcriptome data, complemented by shotgun genome sequencing of haploid tissue for one *Sequoia* individual. Analyses of the resulting assemblies confirmed monophyly of *Sequoia* accessions, consistent with autohexaploidy, which is to say, the gametes that came together to generate the first hexaploid were all more closely related to one another than to *Sequoiadendron*. Divergence estimates suggest Sequoia and Sequoiadendron diverged approximately 51Ma.

Overall my research shows that *Sequoia* is an autohexaploid, yet one showing unexpectedly low sequence divergence between *Sequoia* homeologs. Combined with reports of mixed bivalent and

multivalent chromosome pairing my work shows that diploidization is as yet incomplete in the coast redwood. Autopolyploidy is known to have negative effects in the short-term, most notably a larger genome to be replicated and problems during meioisis, with benefits such as sub- and neo-functionalization of genes only beginning after diploidization. This poses the puzzle as to how the coast redwood has been able to persist for so long in the non-diploidized state. I propose that this apparent contradiction may be explained by life history traits of *Sequoia*, such as its extremely long lifespan and the ability to reproduce clonally, which allowed these magnificent trees to persist for at least 34 million years despite the genomic inefficiencies of autohexaploidy.

On the Evolutionary History of the Coast Redwood

Abstract

This chapter reviews the fossil, cytogenetic, and phylogenetic literature related to the coast redwood, *Sequoia sempervirens*, and lays the foundation for the genomic research described in the remainder of this thesis. Coast redwoods are best known for their long lifespan and extraordinary height. While their current distribution is restricted to a narrow band along the California coast, redwood fossils are found throughout the northern hemisphere, including range overlap among fossil redwood lineages. While the assignment of fossil sequoioids to different phylogenetic lineages has proven problematic, the fossils are helpful in suggesting that *S. sempervirens* has been a polyploidy since at least the late Eocene.

Introduction

Coast redwoods are long-lived trees (some over 2,000 years) and are among the very tallest species in the world (up to 115 meters)(Olson et al., 1990; Ishii *et al.*, 2014). Their geographical range is currently limited to a narrow strip of coast from central California to southern Oregon, dependent on fog. Best known for their great stature and role as an icon of the American conservation movement, coast redwoods are also distinctive as one of only two polyploid conifer species.

Sequoia is a monotypic genus whose closest extant relatives are the giant sequoia, *Sequoiadendron giganteum*, of the Californian Sierra Nevada, and the dawn redwood,

Metasequoia glyptostroboides, of China (Gadek *et al.*, 2000; Kusumi *et al.*, 2000). Cumulatively, *Sequoia, Sequoiadendron,* and *Metasequoia* are known as the redwood clade or Sequoioideae. This group also includes a number of fossil taxa, including the extinct southern hemisphere genus *Austrosequoia*.

Both *Sequoiadendron* and *Metasequoia* are diploid, whereas *Sequoia* is polyploid (Lawson, 1904; Stebbins, 1948). Diverse polyploidization mechanisms have been proposed, and some authors have argued that multiple sequoioid genera contributed to hexaploidy in *Sequoia* (Stebbins, 1948; Saylor & Simons, 1970; Ahuja & Neale, 2002). Despite the restricted present-day ranges of the redwoods, historical range overlap allows the possibility that ancient hybridization occurred among these lineages.

Historical range and fossil distribution

The fossil record of the redwoods has long been a source of confusion for paleobotanists. Due to the superficial similarity in *Metasequoia*, *Sequoia*, and *Taxodium* foliage (Fig. 1), fossils containing only shoot imprints have sometimes been misclassified. In fact, fossils of *Metasequoia* were described as distinct from *Sequoia* only in the 1940s, just a few years before the discovery of the extant species. Chaney (1950) published a lengthy and thorough revision of Taxodiaceous fossils in North America, confirming that many fossils initially attributed to *Sequoia* or *Taxodium* are indeed *Metasequoia*. This reassessment dramatically changed the number of fossil *Sequoia* reported in North America. However, some of Chaney's reassignments have since been re-evaluated resulting in further taxonomic changes. For example, Cridland

(1974) examined a large fossil deposit from the Paleogene of Alaska that Chaney had interpreted as a mixture of *Taxodium* and *Metasequoia* and concluded that a single *Sequoia* species was the most likely the source of all the fossils. This result illustrates the need for caution in interpreting the paleobotanical literature of the Sequoioideae.

The most ancient fossils believed to be redwoods are from both South Manchuria (present-day northeast China) and Boulogne-sur-Mer (northern France), and date back to the Jurassic (Fliche & Zeiller, 1904; Endô, 1951). This result suggests that the clade is at least 146 million years old. Whatever its center of origin, the full fossil record shows redwoods to be widespread (Fig. 2), with fossils reported in Europe, Asia, and North America (Miller, 1977). This historical range of *Sequoia* is shared by many other western North American plant taxa, collectively described by Chaney as arcto-tertiary flora (Chaney, 1947).

In addition to the Northern Hemisphere *Sequoia*-like fossils have been described from Australasia (Peters & Christophel, 1978). Hill *et al*. (1993) systematically assessed the similarity of *Austrosequoia tasmanica*, comprising an ovuliferous cones and associated foliage from the Oligocene Little Rapid River deposit in Northwestern Tasmania, to the genus *Sequoia*. They noted that the bract scales of fossil *A. tasmanica* are similar to those of the sequoioid conifers in that the bract scale and cone are fused together. This analysis also challenged the conclusion of Peters and Christophel (1978) that *Austrosequoia* and *Sequoia* might be closely related to *Athrotaxis*. Instead, Hill et al. (1993) concluded that if *Austrosequoia* were assigned to an extant genus, it should be incorporated into *Sequoia* itself, because only minor morphological distinction, consistent with species-level differentiation, could be found. Specifically, Hill et al.

(1993) felt that stomatal morphology of *Austrosequoia* suggested this specimen had a strong affinity to *Sequoia*.

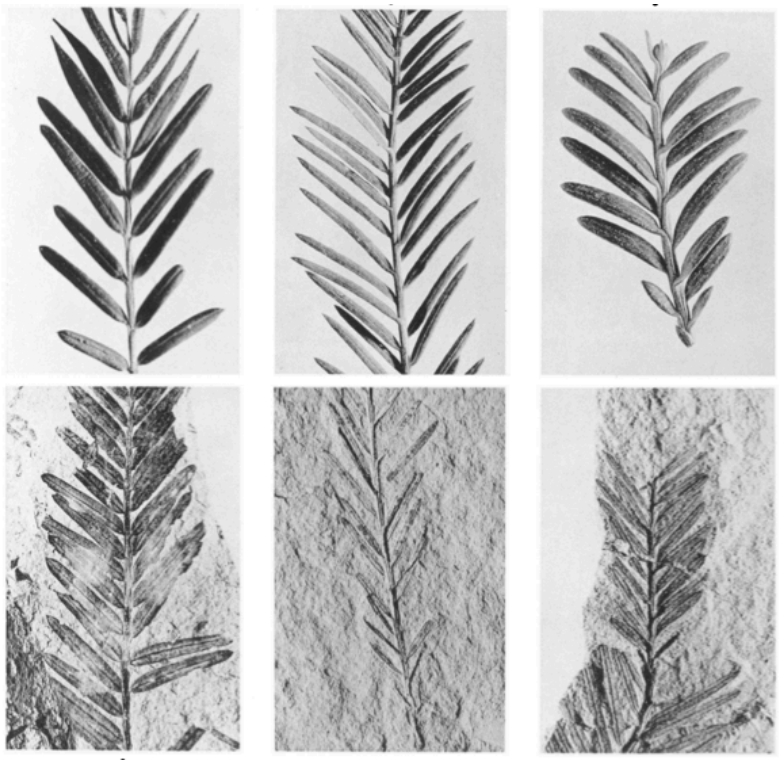

Figure 1: Comparison of extant and fossil foliage of *Metasequoia*, *Taxodium*, and *Sequoia*. Reproduced from Chaney, 1950.

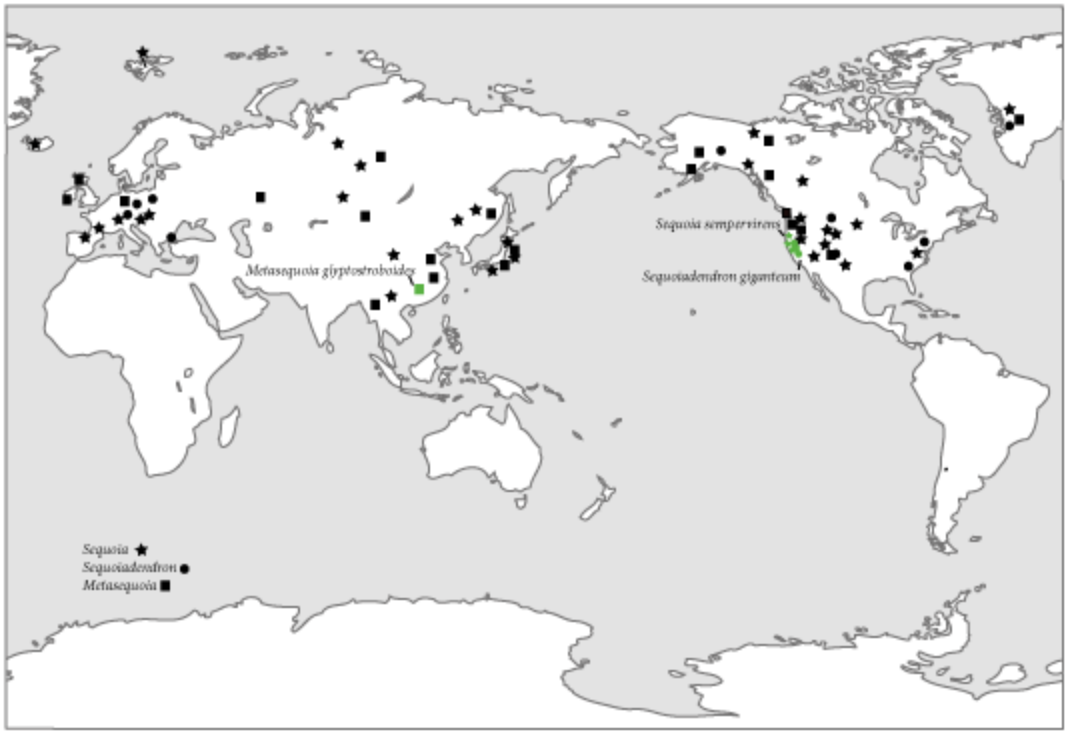

Figure 2: Distribution of sequoioid fossils (adapted from Ahuja, 2009). Current range shown in green.

The difficulty in assigning fossil foliage to a particular redwood lineage is confounded by heteromorphy within species. While *Metasequoia* has oppositely-arranged leaves which lie flat, leaves in both *Sequoia* and *Sequoiadendron* are usually spirally arranged and awl-like. However, a complication arises due to morphological variation within a *Sequoia* canopy, where *Sequoia* foliage at different canopy heights can differ in morphology (Figures 3-4). Specifically, leaves in the lower canopy (often called shade leaves) are broad and flat, while leaves in the upper canopy (sun leaves) are more densely packed and scale-like, resembling the foliage of *Sequoiadendron*. It has been shown that this dramatic variation reflects differences in water potential in a tall tree: low in the *Sequoia* canopy, there is sufficient water potential for cells to expand, resulting in the

large flattened leaves whereas higher up in the canopy, gravity and hydraulic path resistance decrease water potential which limits cellular expansion, resulting in foliage sprays with densely packed, scale-like leaves (e.g. Koch *et al.*, 2004). Indeed, Oldham et al. (2010) found that *Sequoia* mesophyll porosity, leaf shape, and the size of vascular tissues all varied according to tree height with 75% of the observed variation being explained by the hydrostatic gradient. The foliar plasticity is also manifested in leaf anatomy: leaf thickness and the total amount of transfusion tissue increase with canopy height, while xylem tracheid numbers decrease (Oldham *et al.*, 2010).

Figure 3: Photograph showing within-canopy variation in *Sequoia sempervirens* foliage. Photo credit: M.D. Vaden

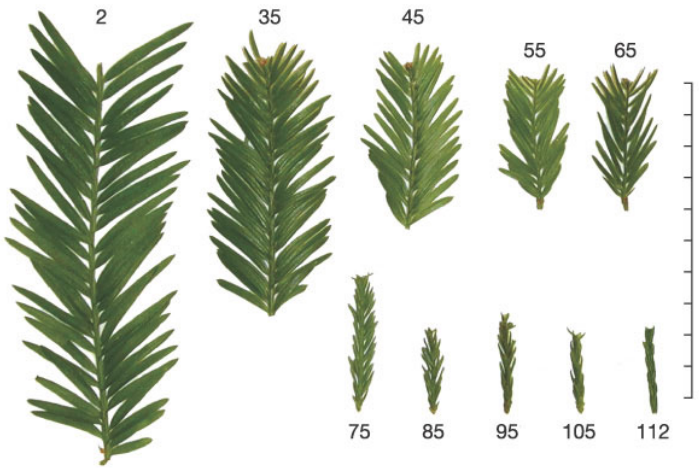

Figure 4: Canopy height variation in *Sequoia* foliage. Numbers indicate meters from ground. Reproduced from Koch et al. (2004).

Taken together the high degree of plasticity in leaf form within individual coast redwoods, which has not always been fully appreciated by paleontologists, meaning that many 20th Century taxonomic assignments for fossils should be treated with caution. Fortunately, however, recently discovered fossils have been assigned and named more carefully, thanks especially to modern techniques that allow imaging of epidermal features (SEM etc.). For example, Zhang *et al*. (2015) combined microscopic analysis of foliage and male cones to provide an unusually robust interpretation of a new fossil redwood from Southeast Yunnan, China, *Sequoia maguanensis*.

In addition to information about past distributions, the fossil record can help shed light on the history of polyploidy in the redwoods. This is possible because guard cell size is known to correlate with genome size (e.g. Beaulieu *et al.*, 2008). As early as 1951, Miki and Hikita compared the stomatal guard-cell sizes of fossil and extant *Sequoia* and *Metasequoia*, concluding that both living taxa have the same genome size as they did in the Pliocene (Miki & Hikita,

1951; Table 1). Taken at face value, this result suggests that *S. sempervirens* has been hexaploid since at least 2.5 million years ago.

Before further discussing the fossil record of polyploidy, it is important to mention that the relationship between guard cell length and genome size is not without its caveats. When grown in an enriched CO2 environment, some species develop epidermal cells that are larger than those on plants grown at ambient CO2 (Miller-Rushing *et al.*, 2009). Fortunately, however, Ogaya et al. (2011) studied the impact of increased CO2 on several tree species including *Metasequoia glyptostroboides* and *Sequoia sempervirens*, showing that these taxa showed little response to environmental conditions. Ogaya *et al*. (2011)concluded that the phenotypic effects of increased CO2 on redwoods are limited to other epidermal characters besides stomatal guard cells showing that in redwoods we can be reasonably confident in interpreting fossil guard cells as a good proxy for ploidy.

The close relationship between ploidy and guard cell size in redwoods allows us to extend the work of Miki and Hikita (1951) further back in time (Table 2). *Sequoia* fossils from China dated to both the Miocene (5.3-23.0 Ma), and even the Eocene (33.9-55.8 Ma), have stomatal guard-cell sizes that fall within the modern range (Ma *et al.*, 2005). This shows that polyploidy has been present in the *Sequoia* lineage for at least the last 34 million years. However, a Miocene fossil *Sequoia* from Iceland has guard-cell sizes that are smaller than previously documented in this genus, instead falling within the observed range of the diploid *Metasequoia* (Grímsson *et al.*, 2007). This could be explained most easily by either a misidentification of the Icelandic

specimen or the persistence of some diploid *Sequoia* populations for at least 18 Ma following the original polyploidization event.

Table 1: Guard cell sizes in fossil and extant redwoods. All measurements and species identifications are as reported from cited literature.

Genus	Origin	Stomata length (μM)	Epoch (age, Ma)	Reference
Sequoia	California, USA	59.8 ± 4.28	Present	Miki & Hikita, 1951
Sequoia	Kyoto, Japan	56.1 ± 3.63	Present	Miki & Hikita, 1951
Sequoia	Tokiguti, Japan	54.6 ± 2.37	Pliocene (2.6 – 5.3)	Miki & Hikita, 1951
Sequoia	Yunnan, China	72.9 ± 9.5	Late Miocene (5.3 – 11.6)	Zhang et al., 2015
Sequoia	Yunnan, China	56 – 62	Miocene (5.3 – 23.0)	Ma et al., 2005
Sequoia	Botn Valley, Iceland	39 – 44	Miocene (15)	Grímsson et al., 2007
Sequoia	Heilongjiang, China	60 – 72	Eocene (33.9 – 55.8)	Ma et al., 2005
Metasequoia	China	41.8 ± 3.12	Present	Miki & Hikita, 1951
Metasequoia	Noboritate, Japan	42.9 ± 2.60	Pliocene (2.5 – 5.3)	Miki & Hikita, 1951
Metasequoia	Tokiguti, Japan	31.3 ± 3.26	Pliocene (2.5 – 5.3)	Miki & Hikita, 1951
Sequoiadendron	California, USA	36.5 ± 3.13	Present	Scott, unpubl.

Karyotype

Prior to the advent of molecular sequencing, cytogenetic analyses allowed botanists to make evolutionary inferences about mechanisms of polyploidy. Numerous researchers have applied karyotypic techniques to redwoods over the last century. In 1904, Lawson first speculated that *S. sempervirens* may be a polyploid, reporting that haploid *Sequoia* tissue had approximately 16 chromosomes per nucleus. This number turned out to be a great underestimate. Dark (1932) reported ~50 chromosomes per diploid cell, Sax & Sax (1933)pinned the number somewhere

above 2n=40, and Buchholz (1939) estimated 2n=44. Despite these repeated efforts to confirm chromosome number in *Sequoia*, it was not until 1943 that Hirayoshi and Nakamura first reported what we now know to be the correct diploid number of 2n=66 (Hirayoshi & Nakamura, 1943). In the years that followed, several authors repeated this study and corroborated 2n=66 (Stebbins, 1948; Miki & Hikita, 1951; Fozdar & Libby, 1968; Saylor & Simons, 1970; Schlarbaum *et al.*, 1984; Ahuja & Neale, 2002). A few of these reports observed three pairs of accessory chromosomes, which can vary in size and occur in addition to the standard karyotype (Fozdar & Libby, 1968; Saylor & Simons, 1970). The origin of these accessory chromosomes is obscure as they are not found in either *Metasequoia* or *Sequoiadendron* (Ahuja & Neale, 2002).

Stebbins (1948) documented the arrangements of chromosome pairs during metaphase I in two *Sequoia* cells, observing a mixture of chromosomal groupings: pairs (bivalents), groups of four (quadrivalents), and chains or rings of six. Comparing his cytological observations to expected patterns of chromosome pairing, Stebbins (1948) concluded that both autopolyploidization and hybridization (perhaps with ancestors of *Metasequoia* or *Taxodium*) had been involved in polyploidization in *Sequoia*.

Saylor and Simons (1970) analyzed root tip squashes of seventeen *Sequoia* seedlings, consistently finding 66 chromosomes of similar size and shape. A majority of chromosomes were metacentric, but a few were submetacentric. Saylor & Simons (1970) sorted the 33 chromosome pairs into eleven triplets based on overall chromosome length and the presence or absence of an attached satellite chromosome (present in three of eleven triplets). In most *Sequoia* cells, Saylor and Simons (1970) observed four nucleoli of similar size, plus one or two much

smaller nucleoli. Based on this size distribution, the authors inferred that one subgenome is divergent from the other two subgenomes and, thus, favored an autoallopolyploid origin of *Sequoia*.

Ahuja and Neale (2002) reported that satellite chromosomes and nucleoli occur in the same numbers, although their relationship is unknown (2002). An important caveat noted by Ahuja and Neale (2002) is that cytological studies in *Sequoia* are limited to a few cells from a few individuals, so these conclusions may not capture the extent of intraspecific variation. Indeed, since *Sequoia* is highly variable based on other assessment methods it would be important to consider whether chromosome behavior (e.g. formation of bivalents vs multivalents) may differ among populations. If there were variation in chromosome pairing patterns across the *Sequoia* range, it could be attributed to either multiple polyploid origins within the group or (more likely given the fossil record) to post-polyploidy genome evolution.

Rogers (1997)studied inheritance patterns in *S. sempervirens* using isozymes to compare the genotypes of embryos with haploid maternal tissue, observing patterns that could not be explained by disomic inheritance. However, given the limited number of loci studied, she could not distinguish among alternative possibilities, such as entirely hexasomic inheritance or a mixture of multisomic inheritance at some loci, and disomic at others.

Pulling the available data together, Ahuja and Neale (2002) entertained three hypotheses for the polyploid origin of *Sequoia*: (1) that *Sequoia* is a partially diploidized autohexaploid, (2) that Sequoia is an autoallohexaploid perhaps due to hybridization among *Metasequoia* and

Sequoiadendron ancestors, and (3) that *Sequoia* is a segmental allopolyploid (a polyploid comprised of partially diverged subgenomes). The only way to clearly distinguish among these alternatives would be to conduct phylogenetic analyses that included all three homeologs (all six alleles) from *Sequoia* along with suitable outgroups.

Previous phylogenetic studies

Kusumi et al. (2000) estimated phylogenies using multiple chloroplast markers and recovered a species tree showing a sister group relationship between the two California redwoods, *Sequoia* and *Sequoiadendron*, with *Metasequoia* as the next closest relative. That same year, Gadek et al. (2000) found that combined analysis of the rbcL and matK loci also supported a California redwood clade (*Sequoia* + *Sequoiadendron*) sister to *Metasequoia*. Gadek et al. (2000) also showed that phylogenetic analysis of morphological data fails contradicted the monophyly of redwood genera. The latter result was contradicted, however, by a study of cone evolution in Cupressaceae, which yielded phylogenetic trees with a redwood tax, supported by the synapomorphy of the lack of distinct female cone scales (Schulz & Stützel, 2007). The same study also supported a sister-group relationship between *Sequoia* and *Sequoiadendron* based on archegonial position and the size of wood rays. Thus, both plastid and morphological data supported the same relationships among the three living redwood species.

Yang *et al*. (2012) studied phylogenetic relationships within Cupressaceae using two nuclear loci (*LFY* and *NLY*) along with chloroplast and mitochondrial regions (*matK* and *rps3*, respectively).

The resulting gene trees showed conflicting topologies with: the LFY tree showed that multiple sequences in *Sequoia* form a clade, with *Metasequoia* as the closest relative, while the NLY tree had *Sequoia* and *Sequoiadendron* in a clade with *Metasequoia* falling outside. Both the chloroplast and mitochondrial trees shared the same topology as the NLY tree, with a *Sequoia*/*Sequoiadendron* clade sister to *Metasequoia*. Based on the conflicting trees from nuclear loci, the authors concluded that hexaploid *Sequoia* is of hybrid origin, with *Metasequoia* and *Sequoiadendron* as parents, and, specifically, due to the paternal pattern of organellar inheritance in Cupressaceae, with *Sequoiadendron* as the paternal lineage. However, it should be borne in mind that the inference of allopolyploidy in Yang *et al*. (2012) comes from a single gene tree (*LFY*) and there is always a possibility that a phylogenetic artifact (e.g., long branch attraction) or a technical problem (such as chimeric PCR products) is confounding the results.

Two more papers that same year (Mao *et al.*, 2012; Leslie *et al.*, 2012) included dated Cupressaceae phylogenies, both supporting a California redwood clade, with *Metasequoia* as the nearest relative. Mao et al. (2012) estimated the divergence of *Sequoia* and *Sequoiadendron* at ~60 Ma, whereas Leslie et al. (2012) inferred a younger age of 40 Ma. Another study based on LFY and NLY showed *Sequoia* sister to *Metasequoia* (Lu *et al.*, 2014) and dated the divergence of *Sequoia* and *Metasequoia* at 80 Ma. However, the *Sequoia* + *Metasequoia* relationship was only weakly supported and as this topology occurs at only a single locus, it is unlikely to represent the true evolutionary history of the species.

Future directions

As discussed earlier, the available data from the fossil record and a limited number of phylogenetic studies leave open the possibility that *Sequoia* is an allopolyploidy. The diversity of redwood fossils from across the Northern Hemisphere suggests that redwood ancestors lived in sympatry, including around the time that polyploidy occurred (sometime between the *Sequoia-Sequoiadendron* split at 40-60 Ma and the oldest known polyploid fossils of 34-53 Ma). However, further dated phylogenetic analyses for multiple nuclear loci are needed to fully evaluate competing hypotheses. In the following two chapters, I use new sequence data along with fossil calibrated phylogenetics to further investigate the timing, mechanism, and consequences of polyploidy in *Sequoia*.

References

Ahuja MR. 2009. Genetic constitution and diversity in four narrow endemic redwoods from the family Cupressaceae. *Euphytica* 165: 5–19.

Ahuja MR, Neale DB. 2002. Origins of polyploidy in coast redwood (Sequoia sempervirens (D. Don) Endl.) and relationship of coast redwood to other genera of Taxodiaceae. *Silvae Genetica* 51: 93–100.

Beaulieu J, Leitch IJ, Patel S, Pendharkar A, Knight CA. 2008. Genome size is a strong predictor of cel size and stomatal density in angiosperms. *New Phytologist*: 975–986.

Buchholz JT. 1939. The embryogeny of Sequoia sempervirens with a comparison of the sequoias. *American Journal of Botany*: 248–257.

Chaney RW. 1947. Tertiary centers and migration routes. *Ecological Monographs* 17: 139–148.

Chaney RW. 1950. A revision of Fossil Sequoia and Taxodium in Western North America based on the recent discovery of Metasequoia. *Transactions of the American Philosophical Society, New Series* 40: 171–263.

Cridland AA. 1974. Presumed Taxodiaceous Fossils from the Paleogene of St. Lawrence Island, Alaska. *Taxon*: 347–351.

Dark SOS. 1932. Chromosomes of Taxus, Sequoia, Cryptomeria and Thuya. *Annals of Botany* 46: 965–977.

Endô S. 1936. New fossil species of Sequoia from the Far-East. *Proceedings of the Imperial Academy* 12: 172–175.

Endô S. 1951. A Record of Sequoia from the Jurassic of Manchuria. *Botanical Gazette* 113: 228–230.

Fliche P, Zeiller R. 1904. Note sur une florule portlandienne des environs de Boulogne-sur-Mer. *Bulletin de la Societe Geologique de France* 4: 787–812.

Fozdar BS, Libby WJ. 1968. Chromosomes of Sequoia sempervirens; 8--Hydroxy-Quinoline-Castor Oil Pretreatment for Improving Preparation. *Stain technology* 43: 97–100.

Gadek PA, Alpers DL, Heslewood MM, Quinn CJ. 2000. Relationships within Cupressaceae sensu lato: a combined morphological and molecular approach. *American journal of botany* 87: 1044–1057.

Grímsson F, Denk T, Símonarson LA. 2007. Middle Miocene floras of Iceland—the early colonization of an island? *Review of Palaeobotany and Palynology* 144: 181–219.

Hill RS, Jordan GJ, Carpenter RJ. 1993. Taxodiaceous macrofossils from Tertiary and Quaternary sediments in Tasmania. *Australian systematic botany* 6: 237–249.

Hirayoshi I, Nakamura Y. 1943. Chromosome number of Sequoia sempervirens. *Bot. Zool* 2: 73–75.

Ishii HR, Azuma W, Kuroda K, Sillett SC. 2014. Pushing the limits to tree height: could foliar water storage compensate for hydraulic constraints in Sequoia sempervirens? *Functional ecology* 28: 1087–1093.

Koch GW, Sillett SC, Jennings GM, Davis SD. 2004. The limits to tree height. *Nature* 428: 851–854.

Kusumi J, Tsumura Y, Yoshimaru H, Tachida H. 2000. Phylogenetic relationships in Taxodiaceae and Cupressaceae sensu stricto based on matK gene, chlL gene, trnL-trnF IGS region, and trnL intron sequences. *American Journal of Botany* 87: 1480–1488.

Lawson AA. 1904. The gametophytes, archegonia, fertilization, and embryo of Sequoia sempervirens. *Annals of Botany* 18: 1–28.

Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S. 2012. Hemisphere-scale differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 109: 16217–21.

Lu Y, Ran J-H, Guo D-M, Yang Z-Y, Wang X-Q. 2014. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PloS one* 9: e107679.

Ma QW, Li FL, Li C Sen. 2005. The coast redwoods (Sequoia, Taxodiaceae) from the Eocene of Heilongjiang and the Miocene of Yunnan, China. *Review of Palaeobotany and Palynology* 135: 117–129.

Mao K, Milne RI, Zhang L, Peng Y, Liu J, Thomas P, Mill RR, Renner SS. 2012. Distribution of living Cupressaceae reflects the breakup of Pangea. *Proceedings of the National Academy of Sciences of the United States of America* 109: 7793–8.

Miki S, Hikita S. 1951. Probable chromosome number of fossil Sequoia and Metasequoia found in Japan. *Science* 113: 3–4.

Miller CN. 1977. Mesozoic conifers. *The Botanical Review* 43: 217–280.

Miller-Rushing AJ, Primack RB, Templer PH, Rathbone S, Mukunda S. 2009. Long-term relationships among atmospheric CO2, stomata, and intrinsic water use efficiency in individual trees. *American Journal of Botany* 96: 1779–1786.

Ogaya R, Llorens L, Peñuelas J. 2011. Density and length of stomatal and epidermal cells in 'living fossil' trees grown under elevated CO2 and a polar light regime. *Acta Oecologica* 37: 381–385.

Oldham AR, Sillett SC, Tomescu AMF, Koch GW. 2010. The hydrostatic gradient, not light availability, drives height-related variation in Sequoia sempervirens (Cupressaceae) leaf anatomy. *American journal of botany* 97: 1087–1097.

Olson DF Jr, Roy DF, Walters GA. 1990. Sequoia sempervirens (D. Don) Endl. Redwood. In: Burns RM, Honkala BH, eds. Technical coordinators. Silvics of North America, volume 1, conifers, agriculture handbook 654. Washington, DC, USA: USDA Forest Service, 541–551.

Peters MD, Christophel DC. 1978. Austrosequoia wintonensis, a new taxodiaceous cone from Queensland, Australia. *Canadian Journal of Botany* 56: 3119–3128.

Rogers DL. 1997. Inheritance of allozymes from seed tissues of the hexaploid gymnosperm, Sequoia sempervirens (D. Don) Endl. (Coast redwood). *Heredity* 78: 166–175.

Sax K, Sax HJ. 1933. Chromosome number and morphology in the conifers. *Journal of the Arnold Arboretum* 14: 356–375.

Saylor LC, Simons HA. 1970. Karyology of Sequoia sempervirens: Karyotype and Accessory Chromosomes. *Cytologia* 35: 294–303.

Schlarbaum SE, Tsuchiya T, Johnson L. 1984. The chromosomes and relationships of Metasequoia and Sequoia (Taxodiaceae) an update. *Journal of the Arnold Arboretum* 65: 251–254.

Schulz C, Stützel T. 2007. Evolution of taxodiaceous Cupressaceae (Coniferopsida). *Organisms Diversity and Evolution* 7: 124–135.

Stebbins GL. 1948. The Chromosomes and Relationships of Metvsequoia and Sequoia. *Science* 108.

Yang Z, Ran J, Wang X. 2012. Molecular Phylogenetics and Evolution Three genome-based phylogeny of Cupressaceae s . l . : Further evidence for the evolution of gymnosperms and Southern Hemisphere biogeography.

Zhang JW, D'Rozario A, Adams JM, Li Y, Liang XQ, Jacques FM, Su T, Zhou ZK. 2015. Sequoia maguanensis, A new miocene relative of the coast redwood, Sequoia sempervirens, From china: Implications for paleogeography and paleoclimate. *American Journal of Botany* 102: 103–118.

Chapter 2:

Whole genome duplication in coast redwood (*Sequoia sempervirens*) and its implications for explaining the rarity of polyploidy in conifers*

*adapted from: Scott, A. D., Stenz, N. W., Ingvarsson, P. K., & Baum, D. A. (2016). Whole genome duplication in coast redwood (Sequoia sempervirens) and its implications for explaining the rarity of polyploidy in conifers. *New Phytologist*, *211*(1), 186-193.

Abstract

Whereas polyploidy is common and an important evolutionary factor in most land plant lineages, it is rare in gymnosperms. Coast redwood (*Sequoia sempervirens*) is one of just two polyploid conifer species and the only hexaploid. Evidence from fossil guard cell size suggests polyploidy in *Sequoia* dates to the Eocene. Numerous hypotheses about the mechanism of polyploidy and parental genome donors have been proposed, primarily based on morphological and cytological data, but it remains unclear how *Sequoia* became polyploid and why this lineage overcame an apparent gymnosperm barrier to whole-genome duplication (WGD). We sequenced transcriptomes and used phylogenetic inference, Bayesian concordance analysis, and paralog age distributions to resolve relationships among gene copies in hexaploid coast redwood and close relatives.

Our data show that hexaploidy in coast redwood is best explained by autopolyploidy or, if there was allopolyploidy, it happened within the Californian redwood clade. We found that duplicate genes have more similar sequences than expected, given the age of the inferred polyploidization. Conflict between molecular and fossil estimates of WGD can be explained if diploidization occurred very slowly following polyploidization. We extrapolate from this to suggest the rarity of polyploidy in gymnosperms may be due to slow diploidization in this clade.

Introduction

Polyploidy has profound long- and short-term genetic consequences (e.g., Adams & Wendel, 2005; Otto & Whitton, 2000), and facilitates adaptive evolution (e.g., Soltis et al., 2008). Studies of genome sequences, expressed genes, and cytogenetics suggest that all land plant lineages have experienced polyploidization in their evolutionary history, though clades differ in the extent of

recent whole genome duplication (neopolyploidization). While there are thousands of neopolyploid mosses, ferns and angiosperms, the phenomenon is relatively rare in gymnosperms, and especially conifers. The Norway spruce genome project found no evidence of whole genome duplication (Nystedt et al., 2013), though a recent reassessment of this dataset did infer polyploidy events early in conifer evolution (Li et al., 2015). Looking at extant taxa, there are only two naturally polyploid conifer species: alerce, *Fitzroya cupressoides* (4x), and coast redwood, *Sequoia sempervirens* (6x). Why is polyploidy so rare in conifers? Does it reflect rare formation of polyploid individuals, for example due to a lack of unreduced gametes, or another barrier to polyploid formation? Or, do polyploid taxa form in gymnosperms, but fail to give rise to successful clades? To shed light on these questions, we studied the evolutionary history of coast redwood with the goal of determining when polyploidy occurred and whether it entailed allopolyploidy.

Coast redwoods are long-lived trees, some individuals attaining 2,000 years (Olson et al., 1990), which thrive in the foggy coastal forests of central and northern California. Coast redwoods can achieve 115 meters in height (Ishii et al., 2014), making them among the world's tallest living trees. *Sequoia* is a monotypic genus whose closest relatives are the giant sequoia of the Californian Sierra Nevada (*Sequoiadendron giganteum*) and the Chinese dawn redwood (*Metasequoia glyptostroboides*). Though the three modern redwood species have distinct ranges without overlap, fossil data suggest that they were all more widely distributed across the Northern Hemisphere from the Cretaceous onwards (Miller, 1977). The oldest redwood fossils are from South Manchuria (present-day China) and Boulogne-sur-Mer (northern France) and

date back to the mid-to-late Jurassic, suggesting the redwood clade is at least 146 million years old (Fliche and Zeiller, 1904; Endô, 1951).

Sequoiadendron and *Metasequoia* are diploids with 2n=22 (Schlarbaum and Tshuchiya, 1984). Hirayoshi and Nakamura (1943) first determined the correct chromosome number of *Sequoia* and proved that it is a hexaploid with 2n=66. Hexaploidy in *Sequoia* was later corroborated by Stebbins (1948), Saylor and Simons (1970) and Ahuja and Neale (2002). Relying on the well-known correlation between guard cell size and genome size (e.g., Beaulieu et al., 2008), Miki and Hikita (1951) studied stomatal guard-cell size in Pliocene fossils of *Metasequoia* and *Sequoia*. As fossil guard cells were the same size as extant guard cells, Miki and Hikita concluded *Sequoia* has been hexaploid since at least the Pliocene (2.5-5 million years ago). This estimate was pushed back significantly by the results of Ma et al. (2005), who describe fossil *Sequoia* from the Eocene (34-53 million years ago) with guard cells indicative of polyploidy.

Long held hypotheses posit that *Sequoia* is an allopolyploid that arose from hybridization between extinct diploid *Sequoia* and ancestors of either *Metasequoia* (Stebbins, 1948) or *Sequoiadendron* (Doyle, 1945). Despite the distance among their modern ranges, the overlap in fossil distributions of *Sequoia*, *Sequoiadendron*, and *Metasequoia* make this hypothesis plausible. Another hypothesis is that an extinct member of the Taxodiaceae, perhaps a member of *Taxodium,* contributed to the hexaploid genome of *Sequoia* (Stebbins, 1948; Saylor and Simons, 1970). In addition to supporting these existing hypotheses, Ahuja and Neale (2002) also suggested that the "missing" parent of *Sequoia* may have been a member of the *Cryptomeria*, *Taiwania*, or *Athrotaxis* lineages.

Before the advent of molecular phylogenetics, auto- and allopolyploids were distinguished by observing chromosome behavior during meiosis. Autopolyploidy (generally interpreted as occurring within a single species) and allopolyploidy (involving hybridization among species) represent extremes of a spectrum. Autopolyploids have multiple sets of very similar homologous chromosomes, which tend to manifest cytogenetically as the formation of multivalents (e.g. groups of four or six chromosomes). Allopolyploids, in contrast, arise from the fusion of divergent genomes meaning that, in the cleanest cases, each homologous chromosome forms a bivalent as happens in diploid organisms. However, chromosome pairing at meiosis is rarely definitive. Allopolyploidy can result in multivalent formation among homeologs if hybridizing species are closely related, and bivalent formation is eventually reestablished following autopolyploidy by the process of diploidization (Ramsey and Schemske, 2002; Parisod et al., 2010).

In addition to cytogenetic lines of evidence, segregation patterns can be useful to distinguish auto- and allopolyploids. An autopolyploid forming multivalents at meiosis will typically show multisomic inheritance (hexasomic in the case of a hexaploid), producing equal frequencies of all possible allele combinations. Allopolyploids do not typically form multivalents at meiosis, resulting in disomic inheritance. Again, these are only the most extreme possibilities. Intermediate inheritance patterns have been observed as well, for example due to an incomplete diploidization process or segmental allopolyploidy (Stebbins, 1947) which involves hybridization among taxa with partially homologous chromosomes.

Studies of meiotic chromosome pairing in *S. sempervirens* reported a mixture of bivalents and multivalents (Stebbins, 1948; Schlarbaum and Tsuchiya, 1984; Ahuja and Neale, 2002). This led Stebbins (1948), Schlarbaum and Tsuchiya (1984), and Schlarbaum, Tsuchija, and Johnson (1984) to suggest that hexaploidy involved both auto- and allopolyploidy. A similar result was obtained by Rogers (1997), who used allozymes to study inheritance patterns in *Sequoia*. However, neither the pairing nor genetic data are sufficient to distinguish segmental allopolyploidy from autopolyploidy followed by partial diploidization. We set out to use modern genomic approaches to revisit the evolutionary history of polyploidy in *S. sempervirens* and see if, by doing so, we could also gain insights into the rarity of polyploidy in gymnosperms.

Materials and Methods

Transcriptome sequencing and assembly

Total RNA was extracted from foliage samples of *Sequoia sempervirens* (Lamb ex. D. Don) Endl., *Sequoiadendron giganteum* (Lindl.) J.Buchh., *Metasequoia glyptostroboides* Hu & W.C.Cheng, and the outgroup *Thuja occidentalis* L. with a CTAB/Chisam protocol followed by Qiagen RNeasy cleanup. Illumina TruSeq cDNA libraries were prepared and sequenced on an Illumina HiSeq 2000 with 100bp paired-end reads at either the UW Biotech Center (Madison, WI) or at the SciLife Laboratory (Stockholm, Sweden). We assembled raw reads *de novo* with Trinity vers. 2014-07-17 (Grabherr et al., 2011), with default settings and Trimmomatic processing.

Sequence analysis and alignment

After assembly, contigs were translated using TransDecoder vers. 2014-07-04 (Haas et al., 2013; http://transdecoder.github.io/) with a minimum protein length of 100 amino acids. Translated contigs were filtered using the Evigene pipeline vers. 2013.07.27 (http://arthropods.eugenes.org/EvidentialGene/about/EvidentialGene_trassembly_pipe.html). Ortholog clusters shared among *S. sempervirens*, *S. giganteum*, *M. glyptostroboides*, and *T. occidentalis* were identified using the translated transcriptome assemblies by ProteinOrtho ver. 5.11 (Lechner et al., 2011), using an algebraic connectivity cutoff of 0.25. Custom Perl scripts (available at https://github.com/nstenz/toca) were used to identify ortholog sets that contained a single copy in diploids (*S. giganteum*, *M. glyptostroboides*, and *T. occidentalis*) and between one and three copies in the hexaploid *S. sempervirens*. As these putatively single-copy protein-coding sequences show marked conservation among species, we assumed that allelic variants would generally be combined into a single contig. We used MUSCLE v. 3.8.13, 64bit (Edgar, 2004a,b), with default alignment settings to align the ortholog sets at the protein level before generating the corresponding nucleotide alignment.

Single-variant gene trees and concordance analyses

For each orthogroup that included only one sequence variant in *S. sempervirens* we estimated phylogenetic trees using MrBayes vers. 3.2.2 64bit (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) with the settings: nst = 6; rates = invgamma; ngen = 1.1 million; burnin = 100,000; samplefreq = 40; nruns = 4; nchains = 3; temp = 0.45; swapfreq = 10. BUCKy vers. 1.4.4 (Ané et al., 2007; Larget et al., 2010) was then used to estimate the proportion of genes that have each possible resolution in the redwood clade while taking account of uncertainty in individual gene trees. Post-burnin posterior distributions from MrBayes were combined in

BUCKy for 1 million generations with $\alpha = 1$. All trees were rooted on the outgroup, *Thuja occidentalis. Thuja* was chosen as an outgroup as a member of the clade Cupressoideae, none of whose members have been hypothesized as potential genome donors to *Sequoia*.

Density distribution of K_s estimates

To build an age distribution of K_s (the average number of synonymous substitutions per synonymous site) within each transcriptome we identified duplicate genes using custom Perl scripts (available at https://github.com/nstenz/plot-ks). Assembled contigs were translated using TransDecoder with a minimum protein length of 100 amino acids, as above. Duplicate genes were identified using BLAT (Kent, 2002) on translated contigs and then duplicate gene pairs were aligned and back translated into their corresponding nucleotide sequence. We estimated K_s on each pair of nucleotide alignments using K_aK_s calculator (model GY; Zhang et al., 2006). We excluded K_s values greater than 2 to avoid the effects of K_s saturation, and plotted the resulting K_s values in a density plot in R (R core team, 2015). To identify significant features of the K_s frequency distributions we used SiZer (Chaudhuri and Marron, 1999).

Multi-variant gene trees and tree-based K_s estimates

For alignments containing a single variant in diploid taxa and two or three variants in hexaploid *Sequoia*, we estimated phylogenetic trees with RAxML vers. 8.1.20 (Stamatakis, 2006) using 100 bootstrap replicates and GTRGAMMA. We then used the codeml function in PAML (Yang, 1997; Yang 2007) to obtain a tree-based estimate of K_s. PAML calculates branch lengths along the ML tree using a model that estimates the rate of synonymous and non-synonymous substitutions (K_s and K_n, respectively) separately for each branch. We imposed a

molecular clock assumption (clock=1) to obtain an ultrametric tree and used the F3x4 codon frequency setting. By multiplying the length of a branch by its K_s and summing over intervening branches between two tips, we could obtain an estimate of the patristic K_s distance between *Sequoia* homeologs and how this compares to the K_s between copies from different species.

In order to obtain an approximate date for gene duplication, we divided the depth of the gene duplication in K_s units by an average mutation rate for conifers of 0.68×10^{9} synonymous substitutions per synonymous site per year (Buschiazzo et al., 2012). *Sequoia* is hexaploid, so at least two whole genome duplications must have occurred in the past. As each whole genome duplication event is expected to yield a normal distribution of K_s values, we used EMMIX v.1.3 (Mclachlan et al., 1999) to fit a mixture model of normal distributions as a way to assign putative homeologs to each duplication event and estimate their ages. We allowed EMMIX to fit 1-2 normal distributions, with the optimal model selected based on AIC and BIC scores.

Data availability

Transcriptome assemblies, sequence alignments, and PAML output are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.7nb70. Scripts used in these analyses are available from https://github.com/nstenz

Results

We used phylogenetic analysis of transcriptome data to investigate the polyploid origin of *Sequoia sempervirens*. Our *de novo* transcriptome assemblies ranged from 70 to 101mbp in length (Supplementary Table 1). Assembled contigs per species ranged from 80,126 to 128,005.

Assuming synonymous substitutions happen at a constant rate over time, K_s can be used as a proxy for the age of duplicate genes. To estimate the distribution of pairwise K_s distance within each genome, we identified all duplicate genes, which numbered 33,544, 39,236, and 26,485, in *S. sempervirens*, *S. giganteum*, and *M. glyptostroboides*, respectively. Paralog age distribution plots for all three taxa revealed a peak at a $K_s \approx 1.5$, of which those for *S. sempervirens*, *S. giganteum* are shown in Fig. 1. This peak likely corresponds to the seed plant whole genome duplication previously dated at 319 million years ago (Jiao et al., 2011), which would imply a mutation rate of 0.47×10^{-8}, which is an order of magnitude higher than that reported by Buschiazzo et al. (2012). Despite the expectation that hexaploid *Sequoia* would have at least one other, much younger peak corresponding to a polyploidization event in perhaps the Eocene (Ma et al., 2005), this was not visible in the age distribution plots (Fig. 1). Results from SiZer also did not indicate any significant peak unique to the *Sequoia* K_s plot.

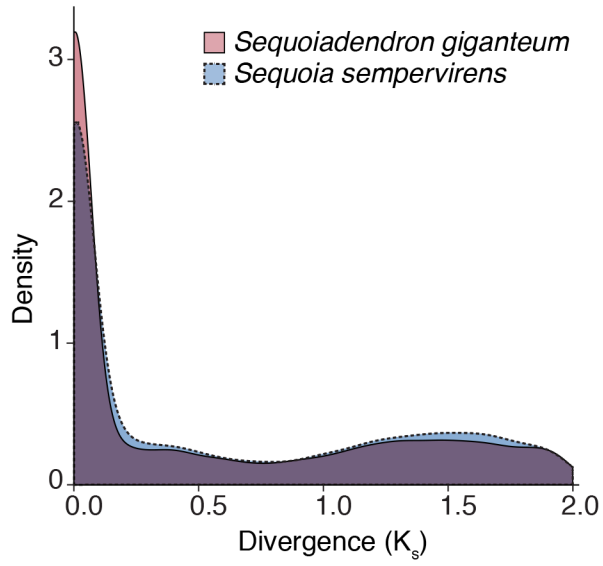

Figure 1: Density distribution of pairwise K_s between duplicate genes in *Sequoia* (blue) and *Sequoiadendron* (red).

To distinguish the evolutionary relationships among redwoods and look for evidence of ancestral hybridization, we used Bayesian concordance analysis and estimated genomic support for each of three possible topologies for an unrooted four-taxon tree. First we built individual gene trees from 7,819 ortholog groups that each had one sequence variant in each diploid species (*Sequoiadendron, Metasequoia, Thuja*) and one, two, or three sequence variants in the hexaploid, *Sequoia*. We built alignments for each ortholog group, comprised of contigs from our transcriptome assemblies. Alignment lengths in this set varied from 301-5,736 bp, with a median of 1,104. Of these alignments 7,602 included a single *Sequoia* copy, whereas 217 included two or three *Sequoia* sequence variants. Among the 7,602 alignments that included a single copy in *S. sempervirens* the most frequently supported topology placed *S. sempervirens* sister to *Sequoiadendron* (Fig. 2) with a concordance factor (CF; Baum, 2007) mean estimate of 0.79 and a 95% credibility interval of 0.78-0.80. The two minor topologies

(*Sequoia* + *Metasequoia*; *Metasequoia* + *Sequoiadendron*) had concordance factors of 0.10(0.09-0.11) and 0.11(0.10, 0.12), respectively (Fig. 2). The equality of the concordance factors of the two minor histories and the fact that both are well below the value of 0.33 (as expected for a donor of one third of the genome), we conclude that if *Sequoia* arose from allopolyploidy, it only involved genome donors within the Californian redwood clade (i.e., the clade that includes *S. sempervirens* and *Sequoiadendron*). However, autopolyploidy is also a possibility.

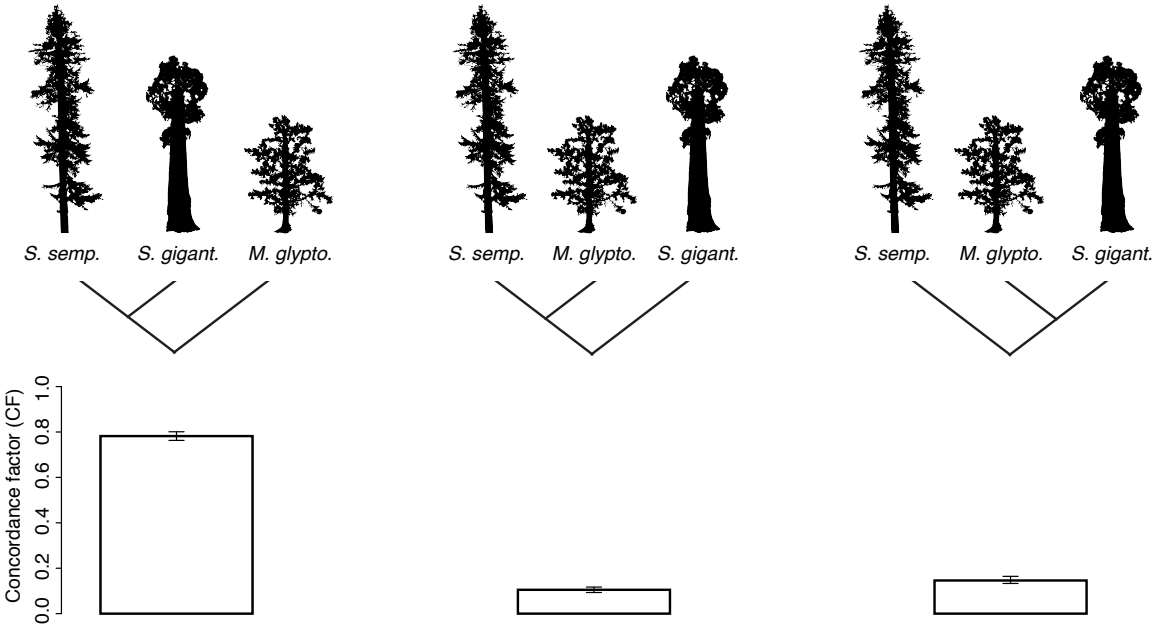

Figure 2: Bayesian concordance analysis of 7,602 gene trees. For each of three possible topologies, the concordance factor (proportion of loci in the sample having the clade) and its 95% credibility interval are shown.

In order to obtain estimates for the divergence of *Sequoia* duplicates relative to interspecies divergences and to evaluate allopolyploidy within the Californian redwood clade, we estimated phylogenetic trees for all genes with more than one sequence variant in *Sequoia*. A total of 217

genes were present in two or three copies in *S. sempervirens*. The optimal tree for 187 of these alignments (86.2%) showed monophyly of the *S. sempervirens* copies with *Sequoia* sister to *Sequoiadendron*, with 97% of these trees being well supported (i.e., having a bootstrap > 0.70). The remaining 30 genes (13.8%) either contradicted monophyly of *S. sempervirens* copies, supporting several other possible relationships, or lacked clear resolution of species relationships.

Based on ML estimates using a codon model in PAML, we could calculate the patristic K_a and K_s distances between each pair of tips on each gene tree. Doing this on the 176 well-supported gene trees that yielded a monophyletic *Sequoia*, the average phylogenetic K_s among *Sequoia* gene copies was 0.013 (Fig. 3). This would correspond to an age of 19.1 Ma using the conifer mutation rate of Buschiazzo et al. (2012) or 2.8 Ma if we calibrate instead using the rate implied if the $K_s \approx 1.5$ peak corresponds to an age of 319 Ma. Either way, the distance between *Sequoia* sequence variants was approximately one-third of that separating *Sequoia* sequences from other redwoods (Fig. 3).

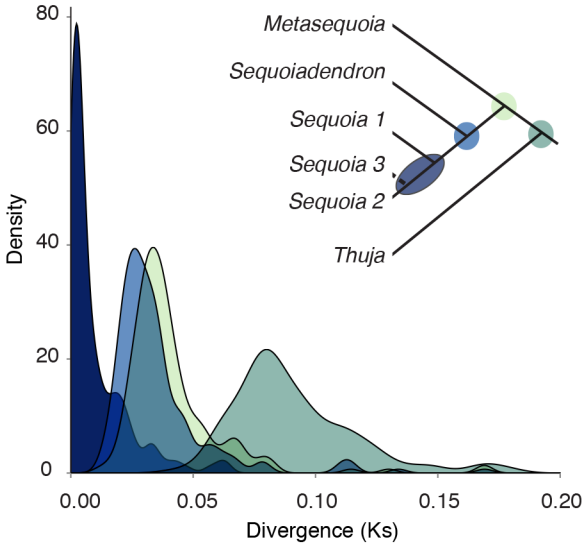

Figure 3: Density distribution of tree-based divergence estimates (in K_s). Four distributions are colored to indicate corresponding nodes on the tree.

We tested whether the patristic K_s estimates between *S. sempervirens* copies are sampled from one or two normal distributions. If hexaploidy arose from two sequential WGD events, there should be two, distinct normal distributions. We used EMMIX to fit a mixture model of normal distributions to the PAML K_s estimates. Based on AIC and BIC scores, the presence of two Gaussian distributions provides a better fit to the K_s distance data. Figure 4 shows the best fitting pair of distributions. Using an average mutation rate for conifers of 0.68×10^{-9} synonymous substitutions per site per year from Buschiazzo et al. (2012), the modes of these peaks correspond to ~3 Ma and ~10 Ma. Using the rate implied if the $K_s \approx 1.5$ peak corresponds to an age of 319 Ma, the two duplication events would be dated at ~0.4 and ~1.5 Ma.

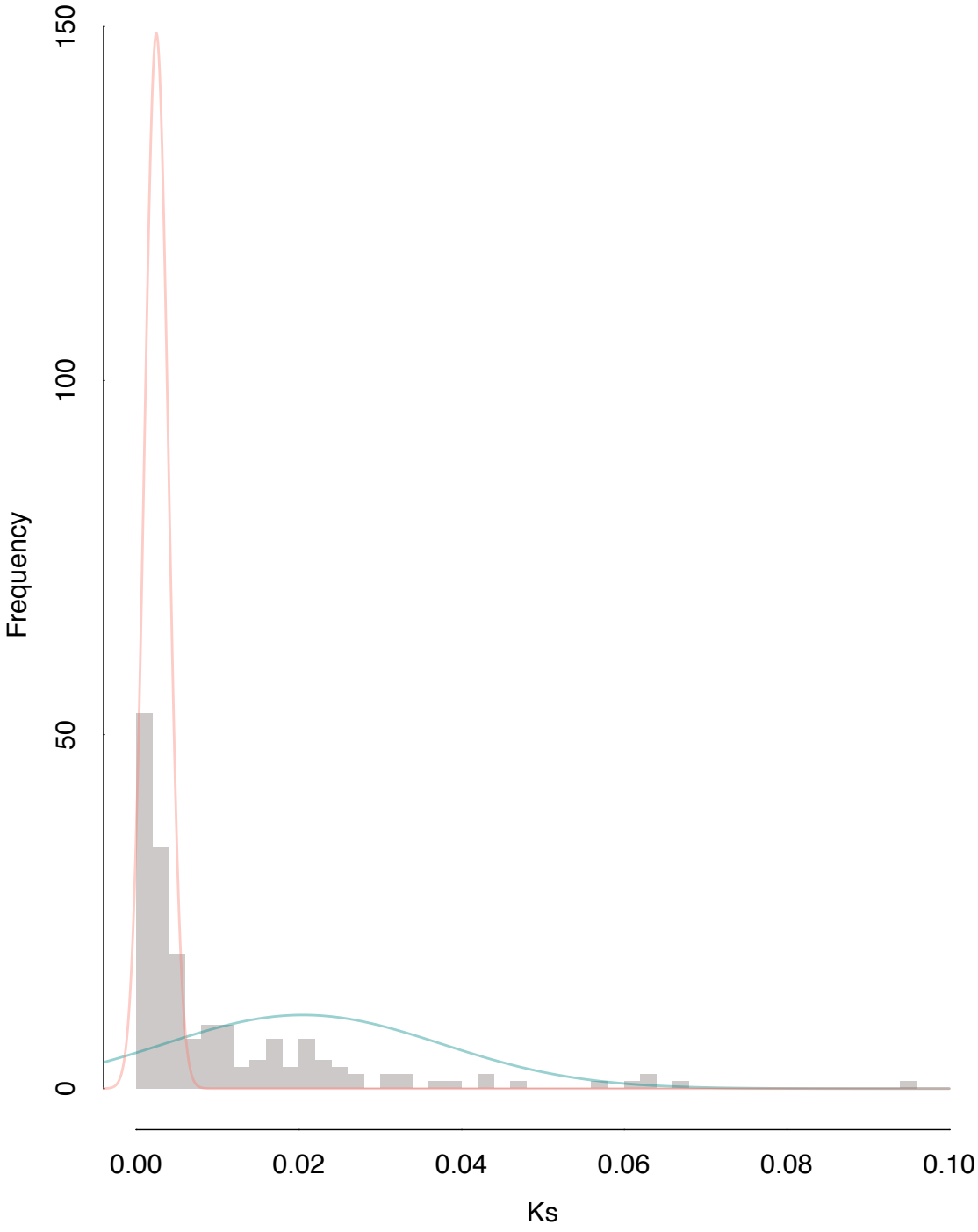

Figure 4: Age distribution of *Sequoia* variants. Colored lines denote the two normal distributions fit with EMMIX.

Discussion

Transcriptome sequencing in the redwoods supports a sister group relationship between *Sequoia* and *Sequoiadendron*.

Bayesian concordance analysis of single copy genes overwhelmingly supports *Sequoiadendron* as the closest relative of *Sequoia*. This conclusion is in agreement with decades of previous work based on morphology, karyotype, and sequence data (e.g. Brunsfield et al., 1994; Gadek et al., 2000; Kusumi et al., 2001; Leslie et al., 2012; Mao et al., 2012; Yang et al., 2012).

We found genes supporting two minor topologies, one with a *Sequoia-Metasequoia* clade and the other with a *Sequoiadendron–Metasequoia* clade. These discordant topologies could be due to incomplete lineage sorting (ILS), which arises when alleles persist between sequential splits in a population tree. In this case, the two minor trees have similar concordance factors, 0.10 and 0.11, and their associated credibility intervals overlap. This pattern is consistent with ILS, which predicts that the alternative minor topologies should have equal CFs (Baum, 2007). Furthermore, given a concordance factor of 0.80, coalescent theory would predict that *Sequoia-Sequoiadendron* clade is subtended by a population lineage whose duration was ~1.22 Ne generations, where Ne is the effective population size (Allman et al., 2011; Stenz et al., 2015). However, it is also possible that the internal branch is considerably longer and discordance is due to other factors such as mistaken orthology. The fact that the two minor histories have similar concordance factors tends to argue against introgression or hybridization as an important phenomenon in the group.

Hexaploidy in *Sequoia* did not involve hybridization among extant redwood lineages.

Our phylogenetic results support an autopolyploid origin for hexaploid *Sequoia*, with no evidence to support hybridization among modern redwood lineages. Single-copy trees convey strong support for *Sequoiadendron* as the closest relative of *Sequoia*, suggesting there was no genome contribution from *Metasequoia*. The lack of evidence that *Metasequoia* was involved with the polyploid origins of *Sequoia* puts some long-held hypotheses to rest (e.g. Stebbins, 1948; Saylor & Simons, 1970). Since these phylogenies include only one copy for hexaploid *Sequoia*, they could not distinguish between autopolyploidy within the *Sequoia* lineage or allopolyploidy within the *Sequoiadendron*-*Sequoia* clade. Single-copy trees may also be inconclusive due to extreme copy-specific expression or genome dominance, where genes from one parental genome are preferentially expressed (e.g. Woodhouse et al., 2014). Therefore, we sought additional evidence by studying orthogroups that included 2 or 3 distinct sequence variants, putatively homeologs, from *Sequoia.*

Phylogenetic analyses of orthogroups containing 2-3 *Sequoia* variants strongly supported monophyly of *Sequoia* homeologs. This result suggests that all gene copies in *Sequoia* originate from the same redwood lineage, implicating either autopolyploidy or hybridization among now extinct lineages that were more closely related to *Sequoia* than to *Sequoiadendron*. It is a theoretical possibility that an allopolyploid history could be masked by biased expression of homeologs from different parents (genome dominance) and/or biased inter-homeolog gene conversion. Although genomic sequences would be necessary to completely rule out this scenario, there is no *a priori* reason to expect that these biases would be strong enough to lead to the observed autopolyploid pattern if, in fact, allopolyploidy applies.

Polyploidy in *Sequoia* arose relatively recently.

The similarity of the K_s paralog age plots obtained from the polyploid *Sequoia* and the diploids *Sequoiadendron* and *Metasequoia* (Fig. 2), and specifically the lack of a recent peak restricted to *Sequoia*, is initially surprising, as these methods have been widely used to diagnose polyploidization events in numerous plant lineages (e.g. Barker et al., 2008; Jiao et al., 2011). This lack of a polyploidization peak might be expected if autopolyploidy had occurred very recently, such that the level of divergence among homeologs is not much different than that among alleles at a particular locus (Vanneste et al. 2013). However, a recent polyploidization event would be at odds with the fossil data, which suggests polyploidization by the end of the Eocene.

One possible explanation for the lack of a polyploidization peak is that only one homeolog is expressed in leaves. Such genome dominance has been observed in other polyploid species (e.g., Adams et al. 2004). However, the fact that we found many genes with two or three distinct copies in *Sequoia* but only one in each diploid argues against uniform silencing of all but one homeolog.

To further explore the history of gene duplication, we inferred trees for alignments that included one transcript in diploids and two or three from hexaploid *Sequoia*, then inferred the branch lengths of these trees in K_s units. We found that K_s estimates between even the most divergent *Sequoia* homeologs were very low (<0.10). One possible explanation is that *Sequoia* experienced a long period of multisomic inheritance following autopolyploidy during which time homeologs recombined repeatedly and variation was lost by genetic drift. This continued recombination could result in much lower K_s values than expected (described in Wolfe, 2001). These observations highlight some caveats of using paralog age distribution

graphs alone to infer recent polyploidization events, or to study ancient whole genome duplication events that were accompanied by extended periods of multisomic inheritance.

Fitting a mixture model of normal distributions to K_s estimates between homeologs yielded two distinct, but overlapping Gaussian distributions, implying two sequential whole genome duplication events. Using two different mutation rate calibrations for conifer K_s divergence, one from the literature and one obtained by assuming that a K_s of 1.5 corresponds to a duplication event dated at 319 Ma, we estimated the timing of the first whole genome duplication in *Sequoia* to have occurred 10 or 1.5 Ma, respectively, with the second occurring more recently, 3 or 0.4 Ma. These dates are in apparent contradiction to the discovery of *Sequoia* fossils in the Eocene (33-53 Ma) with guard cells of a size taken to indicate polyploidy (Ma et al., 2005).

One possible explanation for the discrepancy is that the Eocene fossils represent an independent instance of polyploidy in a closely related lineage that was misclassified as *Sequoia*. Some plant groups exhibit repeated formation of polyploid taxa, a possible case in point being the *Ephedra* lineage, which appears to have experienced multiple whole genome duplication events (Ickert-Bond, 2003; Ickert-Bond and Renner, 2016). Whereas measurements of guard cells in a larger number of different aged redwood fossils from different geographic locations would certainly be helpful, the rarity of polyploidy in conifers and the high quality of the redwood fossil record make it much more likely that the fossils in question represent true polyploid members of the lineage leading to *S. sempervirens*.

Assuming that polyploidy indeed occurred at least 34 Ma, the most plausible explanation for the low divergence of putative homeologs in *Sequoia* is that multisomic inheritance persisted for a long period of time following whole genome duplication, possibly even to the present for some loci. If that is so, the gene duplication events we dated do not correspond to the polyploidy event

per se but to more recent coalescent events at multisomically-inherited loci. This hypothesis is consistent with multivalent formation in modern *Sequoia*, and suggests a very slow diploidization process following whole genome duplication in *Sequoia*.

Implications for polyploidization patterns in gymnosperms.

Reviews of polyploid occurrence in gymnosperms have focused on a few potential reasons to explain their rarity, including a lack of self-fertilization (which facilitates polyploid establishment) and absence of endosperm-forming double fertilization in gymnosperms, (Khoshoo, 1959; Ahuja, 2005). Given what we know about polyploidy in *Sequoia*, what conclusions can we draw about patterns of polyploidization in gymnosperms overall?

With the exception of *Ephedra*, instances of polyploid gymnosperms are limited to monospecific genera (e.g. *Sequoia*, *Fitzroya*), or polyploid individuals within diploid species (e.g. *Juniperus x pftizeriana*; Ahuja, 2005). Notably, polyploidy in gymnosperms seems to be primarily due to autopolyploidization (again, with the exception of *Ephedra*). All cases of polyploid trees within diploid conifer species have been classified as autopolyploids (e.g. in *Larix decidua* and *Juniperus virginiana*; Khoshoo, 1959). Isozyme data suggest Fitzroya cupressoides is an autotetraploid with tetrasomic inheritance (Premoli et al., 2000).

The apparent mismatch between the inferred age of gene duplication and the timing of polyploidization as seen in the fossil record suggests an intriguing hypothesis to explain the paucity of polyploidy in gymnosperms. Perhaps diploidization happens more slowly in gymnosperms (except possibly *Ephedra*) than in angiosperms. The main long-term benefits of

polyploidy, namely potential sub- and neo-functionalization of genes, require divergence among homeologous chromosomes, which can only happen once loci are diploidized. Thus, continued multisomic inheritance precludes the emergence of any evolutionary advantage in polyploid lineages. *Ephedra* appears to be an exception as many polyploids in this genus are allopolyploid (Wu et al., 2016) and therefore might involve subgenomes that are sufficiently diverged that disomic inheritance occurs from the outset.

If polyploidy in gymnosperms is more burden than boon, the persistence of hexaploid *Sequoia* may reflect an ability to avoid extinction rather than superior fitness. In this regard it is perhaps noteworthy that *S. sempervirens* manifests some traits that might help stave off extinction, namely clonal reproduction, self-compatibility, and extreme longevity. In coast redwood populations, suckers often emerge from the base of adult trees, extending generation time (meiosis-to-meiosis) almost indefinitely. In addition to facilitating the spread of a polyploid genotype, vegetative reproduction could allow the persistence of multisomic inheritance by reducing selection for efficient meiosis. Furthermore, production of asexual stands may lead to abundant genetic selfing among clonal ramets, as coast redwoods are self-compatible (Burns & Honkala, 1990). This means that a spontaneous polyploid, perhaps gaining the transient advantage of fixed heterozygosity, could spread by a combination of asexual reproduction and selfing. It is conceivable, therefore, that even after the erosion of fixed heterozygosity the lineage could persist despite never gaining the long-term advantages typically associated with polyploidy, instead suffering the concomitant problem of enlarged genome size. And the same holds for the other natural polyploid conifer species, *Fitzroya cupressoides*, which is both long-lived and capable of clonal reproduction (Silla et al., 2002). Thus, while more work is needed to

evaluate the occurrence of multisomic inheritance in both polyploid species (e.g. *Sequoia, Fitzroya*) and polyploid individuals in diploid taxa (e.g. *Juniperus x pftizeriana*), our hypothesis both explains the rarity of neopolyploidy in gymnosperms and why *Sequoia* is an exception to this general rule.

Acknowledgements

We thank Cécile Ané, Matt Johnson, and Nisa Karimi for improving this manuscript through countless discussions. Support for this project was provided by a grant to ADS and DB from Save The Redwoods League. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0718123 to ADS. PKI was supported by a grant from the Knut and Alice Wallenberg foundation (KAW). NS was supported by a National Science Foundation to DB (DEB-1354793). The data analyses were partly performed using resources provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX.

Author Contributions

ADS, PKI, and DB designed the research. ADS collected the data. ADS and NS analyzed the data. ADS and DB wrote the manuscript.

References

Adams, K. L., Percifield, R., & Wendel, J. F. (2004). Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics*, *168*(4), 2217-2226.

Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current opinion in plant biology*, *8*(2), 135-141.

Ahuja, M. R. (2005). Polyploidy in gymnosperms: revisited. *Silvae Genetica*,*54*(2), 59-68.

Ahuja, M. R., & Neale, D. B. (2002). Origins of polyploidy in coast redwood (Sequoia sempervirens (D. Don) Endl.) and relationship of coast redwood to other genera of Taxodiaceae. *Silvae Genetica*, *51*(2-3), 93-99.

Allman, E. S., Degnan, J. H., & Rhodes, J. A. (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of mathematical biology*, *62*(6), 833-862.

Ané, C., Larget, B., Baum, D. A., Smith, S. D., & Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*,*24*(2), 412-426.

Barker, M. S., Kane, N. C., Matvienko, M., Kozik, A., Michelmore, R. W., Knapp, S. J., & Rieseberg, L. H. (2008). Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution*, *25*(11), 2445-2455.

Baum, D. A. (2007). Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*, 417-426.

Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A., & Knight, C. A. (2008). Genome size is a strong predictor of cell size and stomatal density in angiosperms. *New Phytologist*, *179*(4), 975-986.

Brunsfeld, S. J., Soltis, P. S., Soltis, D. E., Gadek, P. A., Quinn, C. J., Strenge, D. D., & Ranker, T. A. (1994). Phylogenetic relationships among the genera of Taxodiaceae and Cupressaceae: evidence from rbcL sequences. *Systematic Botany*, 19: 253-262.

Olson Jr., D.F, Roy, D.F, and Walters, G.A. (1990). *Sequoia sempervirens* (D. Don) Endl. Redwood. In: Burns R.M., Honkala B.H. (Technical coordinators), *Silvics of North America, Volume 1, Conifers*, Agriculture Handbook 654. USDA Forest Service, Washington, DC, pp. 541-551.

Buschiazzo, E., Ritland, C., Bohlmann, J., & Ritland, K. (2012). Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC evolutionary biology*, *12*(1), 8.

Chaudhuri, P., & Marron, J. S. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, *94*(447), 807-823.

Doyle, J. (1945). Naming of the redwoods. *Nature*, *155*, 254-257.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**(5):1792-1797

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, (**5**) 113

Endô, S. (1951). A record of Sequoia from the Jurassic of Manchuria. *Botanical Gazette*, 228-230.

Fliche, P. & Zeiller, R. (1904). Note sur une florule portlandienne des environs de Boulogne-sur-Mer. *Bulletin de la Société Géologique de France*,**4**:787-812.

Gadek, P. A., Alpers, D. L., Heslewood, M. M., & Quinn, C. J. (2000). Relationships within Cupressaceae sensu lato: a combined morphological and molecular approach. *American Journal of Botany*, *87*(7), 1044-1057.

Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29: 644-652.

Haas B, Papanicolaou A, Yassour M, Grabherr M, Blood P, Bowden J, Couger M, Eccles D, Li B, Lieber M et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8: 1494-1512.

Hirayoshi, I., & Nakamura, Y. (1943). Chromosome number of *Sequoia sempervirens*. *Bot. Zool*, *2*, 73-75.

Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754-755.

Ickert-Bond, S. M. 2003. Systematics of New World *Ephedra* L. (Ephedraceae): integrating morphological and molecular data. Ph.D. dissertation, Arizona State University, Tempe, Arizona, USA

Ickert-Bond, S. M., & Renner, S. S. (2016). The Gnetales: Recent insights on their morphology, reproductive biology, chromosome numbers, biogeography, and divergence times. *Journal of Systematics and Evolution*,*54*(1), 1-16.

Ishii, H. R., Azuma, W., Kuroda, K., & Sillett, S. C. 2014. Pushing the limits to tree height: could foliar water storage compensate for hydraulic constraints in Sequoia sempervirens? *Functional Ecology*, 28(5), 1087-1093.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., ... & Leebens-Mack, J. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, *473*(7345), 97-100.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome research*,*12*(4), 656-664.

Khoshoo, T. N. (1959). Polyploidy in gymnosperms. *Evolution*, 13: 24-39.

Kusumi, J., Tsumura, Y., Yoshimaru, H., & Tachida, H. (2000). Phylogenetic relationships in Taxodiaceae and Cupressaceae sensu stricto based on matK gene, chlL gene, trnL-trnF IGS region, and trnL intron sequences. *American Journal of Botany*, *87*(10), 1480-1488.

Larget, B. R., Kotha, S. K., Dewey, C. N., & Ané, C. (2010). BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, *26*(22), 2910-2911.

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (Co-) orthologs in large-scale analysis. *BMC bioinformatics*, *12*(1), 124.

Leslie, A. B., Beaulieu, J. M., Rai, H. S., Crane, P. R., Donoghue, M. J., & Mathews, S. (2012). Hemisphere-scale differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, *109*(40), 16217-16221.

Li Z, Baniaga A, Sessa E, Scascitelli M, Graham S, Rieseberg L, Barker M. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* 1: e1501084-e1501084.

Ma, Q. W., Li, F. L., & Li, C. S. (2005). The coast redwoods (Sequoia, Taxodiaceae) from the Eocene of Heilongjiang and the Miocene of Yunnan, China. *Review of Palaeobotany and Palynology*, *135*(3), 117-129.

Mao, K., Milne, R. I., Zhang, L., Peng, Y., Liu, J., Thomas, P., ... & Renner, S. S. (2012). Distribution of living Cupressaceae reflects the breakup of Pangea. *Proceedings of the National Academy of Sciences*, *109*(20), 7793-7798.

Mclachlan G, Peel D, Basford K, Adams P. 1999. The EMMIX software for the fitting of mixtures of normal and t- components. J Stat Softw. 4:2.

Miki, S., & Hikita, S. (1951). Probable chromosome number of fossil Sequoia and Metasequoia found in Japan. *Science*, *113*(2923), 3-4.

Miller, C. N. (1977). Mesozoic conifers. *The Botanical Review*, *43*(2), 217-280.

Nystedt B, Street N, Wetterbom A, Zuccolo A, Lin Y, Scofield D, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A et al. 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* 497: 579-584.

Otto, S. P., & Whitton, J. (2000). Polyploid incidence and evolution. *Annual review of genetics*, *34*(1), 401-437.

Parisod, C., Holderegger, R., & Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytologist*, *186*(1), 5-17.

Premoli, A. C., Kitzberger, T., & Veblen, T. T. (2000). Conservation genetics of the endangered conifer Fitzroya cupressoides in Chile and Argentina. *Conservation Genetics*, *1*(1), 57-66.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Version 3.2.1

Ramsey, J., & Schemske, D. W. (2002). Neopolyploidy in flowering plants. *Annual review of ecology and systematics*, 33: 589-639.

Rogers, D. L. (1997). Inheritance of allozymes from seed tissues of the hexaploid gymnosperm, Sequoia sempervirens (D. Don) Endl.(Coast redwood). *Heredity*, *78*(2), 166-175.

Ronquist, F. and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Saylor, L. C., &Simons, H. A. (1970). Karyology of Sequoia sempervirens: karyotype and accessory chromosomes. *Cytologia*, *35*(2), 294-303.

Schlarbaum, S. E., & Tsuchiya, T. (1984). Cytotaxonomy and phylogeny in certain species of Taxodiaceae. *Plant systematics and evolution*, *147*(1-2), 29-54.

Schlarbaum, S. E., Tsuchiya, T., & Johnson, L. C. (1984). The chromosomes and relationships of Metasequoia and Sequoia (Taxodiaceae): an update. *Journal of the Arnold Arboretum*, *65*(2), 251-254.

Silla, F., Fraver, S., Lara, A., Allnutt, T. R., & Newton, A. (2002). Regeneration and stand dynamics of Fitzroya cupressoides (Cupressaceae) forests of southern Chile's Central Depression. *Forest Ecology and Management*, *165*(1), 213-224.

Soltis, D. E., Bell, C. D., Kim, S., & Soltis, P. S. (2008). Origin and early evolution of angiosperms. *Annals of the New York Academy of Sciences*,*1133*(1), 3-25.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688-2690.

Stebbins, G. L. (1947). Types of polyploids: their classification and significance. *Advances in genetics*, *1*, 403-29.

Stebbins, G. L. (1948). The chromosomes and relationships of Metasequoia and Sequoia. *Science*, *108*(2796), 95-98.

Stenz, N. W., Larget, B., Baum, D. A., & Ané, C. (2015). Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species Arabidopsis thaliana (L.) Heynh. *Systematic biology*, 64:809-823.

Vanneste, K., Van de Peer, Y., & Maere, S. (2013). Inference of genome duplications from age distributions revisited. *Molecular biology and evolution*,*30*(1), 177-190.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, *2*(5), 333-341.

Woodhouse, M. R., Cheng, F., Pires, J. C., Lisch, D., Freeling, M., & Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proceedings of the National Academy of Sciences*,*111*(14), 5283-5288.

Wu, H., Ma, Z., Wang, M. M., Qin, A. L., Ran, J. H., & Wang, X. Q. (2016). A high frequency of allopolyploid speciation in the gymnospermous genus Ephedra and its possible association with some biological and ecological features. *Molecular Ecology*. **DOI:** 10.1111/mec.13538

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood Computer Applications in BioSciences 13:555-556.

Yang, Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum

likelihood. Molecular Biology and Evolution 24: 1586-1591

Yang, Z. Y., Ran, J. H., & Wang, X. Q. (2012). Three genome-based phylogeny of Cupressaceae sl: Further evidence for the evolution of gymnosperms and Southern Hemisphere biogeography. *Molecular phylogenetics and evolution*,*64*(3), 452-470.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J: KaK$_s$ Calculator: Calculating Ka and K$_s$ through model selection and model averaging. *Genomics Proteomics Bioinformatics* 2006, 4:259-263.

Genomic Analysis of Polyploid Evolution in *Sequoia sempervirens*

Introduction

Polyploidy and its consequences have been the focus of countless studies (e.g. Otto & Whitton, 2000; Adams & Wendel, 2005; Soltis *et al.*, 2008). Polyploidization can facilitate speciation via immediate reproductive isolation (e.g. Wood *et al.*, 2009) and provides genetic material for adaptive evolution via the sub- and neofunctionalization of genes (e.g. Levin, 1983). Cytogenetic and sequence data indicate that whole genome duplications (WGDs) have occurred in every land plant lineage, whether recently (neopolyploidy) or long ago (paleopolyploidy)(Jiao *et al.*, 2011). However, in contrast to their high frequency in ferns, mosses, and angiosperms, WGDs appear to be rare in gymnosperm lineages, and especially in conifers. Recent studies have inferred ancient WGD events in gymnosperms, including two paleopolyploidization events in conifers (Li *et al.*, 2015), and have identified a few polyploid individuals in otherwise diploid gymnosperm species (Šmarda *et al.*, 2016). However, there are only two neopolyploid conifer species, both in Cupressaceae, namely alerce, *Fitzroya cupressoides* (4x), and coast redwood, *Sequoia sempervirens* (6x).

Hexaploid coast redwoods are well known for their great height (over 100m) and longevity, with ages over 2,000 years reported (Olson et al., 1990). Although these colossal, ancient trees are currently restricted to the coastal forests of central and northern California and southwestern Oregon, the fossil record suggests a broader historical range across the Northern hemisphere. Though *S. sempervirens* was identified as a hexaploid long ago (Stebbins, 1948; Schlarbaum *et al.*, 1984), how it obtained six copies of each chromosome (three pairs) is not well understood.

Polyploids are typically categorized as either autopolyploids (formed via whole genome duplication within a single species) or allopolyploids (formed via whole genome duplication with hybridization among species). In reality, these two classes are best seen as two ends of a spectrum, ranging from pure autopolyploids, with completely homologous sets of chromosomes at the time of polyploidy, to pure allopolyploids whose chromosome complements are sufficiently diverged at the time of polyploidy that there is no homeologous pairing. Thus, soon after polyploidy autopolyploids are expected to demonstrate multivalent pairing in cytogenetic studies, whereas strict allopolyploids with divergent subgenomes will show only bivalent pairing. However, over evolutionary time it is known that autopolyploids often diploidize, meaning that different homeologs diverge, resulting in an increase in the consistency of bivalent pairings (e.g. Wolfe, 2001).

The mechanism of polyploidization in *Sequoia* has long been a source of speculation, especially the question of whether it involved hybridization or not. Cytogenetic work on *Sequoia* has revealed a mixture of bivalents and multivalents (Stebbins, 1948; Saylor & Simons, 1970; Schlarbaum *et al.*, 1984; Ahuja & Neale, 2002), which were seen to support a combination of both auto- and allopolyploidy (Stebbins, 1948; Schlarbaum *et al*., 1984), though strict autopolyploidy followed by partial diploidization cannot be ruled out (Stebbins, 1948).

Chromosome pairing patterns also impact the segregation of alleles. In diploids and pure allopolyploids, where chromosome pairing involves just bivalents (pairs), disomic inheritance of alleles is observed. Disomic inheritance entails classic Mendelian segregation of the two kinds of alleles at a given locus. In contrast, an autotetraploid with four sets of homologous chromosomes

that all able to pair up at meiosis can show tetrasomic inheritance: gametes can receive any possible combination of the four alleles in the maternal cell. Likewise, a hexaploid can show hexasomic inheritance.

Rogers (1997) studied inheritance patterns in *S. sempervirens* using isozymes. She observed patterns that could not be explained by disomic inheritance, instead suggesting hexasomic inheritance. However, due to limitations of the data, the author could not distinguish among alternative possibilities, such as multisomic inheritance at some loci, and disomic at others. Rogers (1997) concluded that her observations were consistent with existing hypotheses about *Sequoia*'s polyploid origin, namely that auto- and allopolyploidy could both be involved.

Despite several hypotheses suggesting that *Sequoia* is an allopolyploid resulting from hybridization between an extinct diploid ancestor of *Sequoia* and ancestors of either *Metasequoia* (Stebbins, 1948) or *Sequoiadendron* (Doyle, 1945), more recent studies have found little if any support for any hybridization. Instead, genomic evidence suggests *Sequoia* is likely an autopolyploid (Scott *et al.*, 2016). However, that inference was based on transcriptome data, which is potentially misleading because of the potential for copy-specific expression or genome dominance . Given this, Scott *et al*. (2016) could not rule out hybridization between the *Sequoia* lineage and the other California redwood lineage, which includes only the giant sequoia, *Sequoiadendron giganteum*. In this paper, we use a combination of targeted sequence capture of genomic DNA to better understand the mechanism, timing, and consequences of polyploidy in *Sequoia*. In addition, we incorporate fossils and sequence data to estimate divergence times in the redwood clade, providing a framework for understanding the means and timing of

polyploidization in *Sequoia*. The rich and diverse redwood fossil record, coupled with our divergence time estimates, suggest that multiple redwood lineages existed in sympatry in the past. We find further evidence that *Sequoia* is an autohexaploid, and that low divergence among individuals suggest it is at least partially undiploidized.

Methods

Bait design, DNA extraction, targeted sequence capture, and sequencing

Based on previously generated *de novo* transcriptomes (Scott et al., 2016), we designed baits for targeted sequence capture. We used BLAST to compare transcriptome data from *Sequoia*, *Sequoiadendron*, and *Metasequoia* to the Norway spruce reference genome (Nystedt *et al.*, 2013). We filtered potential targets by length, requiring all contigs to be at least 800 bp long. To ensure successful sequence capture among phylogenetically distant taxa, we also restricted our search to potential targets with at least 95% sequence identity among *Sequoia*, *Sequoiadendron*, and *Metasequoia*. After removal of repeats, the bait set of 907 targets had a cumulative length of 2,128,430 bases. Custom baits were synthesized by MYcroarray (Ann Arbor, MI).

We extracted genomic DNA from four accessions of *Sequoia sempervirens* and three accessions of *Sequoiadendron giganteum*, along with individual representatives of *Metasequoia glyptostroboides*, *Thuja plicata*, and *Fitzroya cupressoides* (Table 1). The UW Biotech Center prepared Illumina TruSeq libraries from the resulting genomic DNA (average insert size 435bp). MYcroarray performed hybridizations between Illumina libraries and our custom baits and the resulting enriched libraries were sequenced at the UW Biotech Center on an Illumina MiSeq with paired-end 300bp reads and sorted per-accession by barcode.

Sequence assembly and alignment

To assemble targeted genes from sequence data, we used HybPiper (Johnson *et al.*, 2016). Our bait file contained all exon-based targets, and each of the ten accessions was assembled independently. We only analyzed targets where no paralogs were detected. Contig sequences were extracted and aligned with MAFFT (Katoh, 2002), with default settings. Alignments were converted into Nexus format with Perl scripts, and then we estimated gene tree posteriors with MrBayes v3.2.6 64bit (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) with settings as follows: nst = 6; rates = invgamma; ngen = 1.1 million; burnin = 100,000; samplefreq = 40; nruns = 4; nchains = 3; temp = 0.45; swapfreq = 10. The posterior distributions were summarized using the mbsum function in BUCKy v.1.4.4 (Ané *et al.*, 2007, 2010; Larget *et al.*, 2010) and we then performed Bayesian concordance analysis with BUCKy (alpha settings ranging from 0.1 to 1 in increments of 0.1) to obtain a population tree (or "species tree") and a concordance tree with concordance factors (Baum, 2007).

For each gene we estimated phylogenetic trees with RAxML v.8.1.20 (Stamatakis, 2006) using the GTRGAMMA model. We used MACSE (Ranwez *et al.*, 2011) to generate in-frame exon alignments, and removed terminal stop codons. Then, using the RAxML trees and MACSE alignments as input, the CODEML function in the package Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang, 1997, 2007) was used to obtain a tree-based estimate of synonymous substitutions per synonymous site. PAML estimated branch lengths on the ML tree using a model that assumes one value of ω (dN/dS) for the entire tree. We imposed a molecular clock assumption (clock = 1) to obtain an ultrametric tree and used the F3x4 codon frequency setting.

We sorted the gene trees by topology using PhyloSort (Moustafa & Bhattacharya, 2008) to identify those gene trees matching the inferred population tree, that is with *Sequoia* accessions monophyletic, *Sequoia* and *Sequoiadendron* forming a clade, and *Metasequoia* sister to *Sequoia* + *Sequoiadendron*. We imported the topology-filtered trees, with their corresponding ω (dN/dS), into R (R Core Team, 2017) using the ape package (Paradis *et al.*, 2004). By replacing the length of each branch (reported by PAML as t), with the corresponding dS, we obtained trees with branch lengths in Ks. We then used the getMRCA function in ape to obtain estimates of the divergence times of all nodes in the tree and then plotted divergence times in Ks units using the ggplot2 package in R (Wickham, 2009).

Fossil calibrated phylogenetics

To estimate divergence times among the redwoods, we took advantage of their rich fossil record and estimated calibrated phylogenies in BEAST 2 (Bouckaert *et al.*, 2014) . We used MAFFT (Katoh et al., 2002) to align 20 target loci for seven taxa: two accessions of *Sequoia sempervirens*, two accessions of *Sequoiadendron giganteum*, and one each of *Metasequoia glyptostroboides, Thuja plicata,* and *Fitzroya cupressoides*. Alignment files were prepared for BEAST using BEAUTi v2.4.5 (Drummond *et al.*, 2012) with settings: GTR+Γ substitution model, relaxed uncorrelated clock using a lognormal distribution, Yule speciation process, and 50,000,000 MCMC generations. We used four calibration points. (1) A primary calibration that provides a minimum age for the divergence of *Metasequoia* from *Sequoia* + *Sequoiadendron* using a whole plant reconstruction of *Metasequoia occidentalis* from the Paleogene of China (Liu *et al.*, 1999) Following Leslie *et al*. (2012), this was assigned a 20 million year lognormal

prior, with a mean of 65 (95% HPD 55,75). (2) A primary calibration for the age for the divergence between *Sequoia* and *Sequoiadendron* based on the earliest *Sequoia* specimen with guard cells whose size indicated polyploidy (Ma *et al.*, 2005) . As this fossil was dated to the Eocene, we assigned a uniform prior bounded by 33 Ma, corresponding to the end of the Eocene, and 85Ma, which provides a reasonable maximum age for this node. (3) A secondary calibration at the most basal node of the tree (i.e. the stem of Sequoioideae), using a uniform prior of 150-215 Ma, as suggested by Mao et al. (2012), who used twelve fossil calibrations to date divergences within the Cupressaceae. (4) A secondary calibration to date the divergence of *Fitzroya* and *Thuja* (i.e. Callitroideae and Cupressoideae), using a uniform prior of 124-183 Ma, as estimated by Mao et al. (2012).

Results

We used targeted sequence capture data to explore the polyploid origin of hexaploid *Sequoia* and investigate relative divergence times of *Sequoia* sequence variants compared to deeper nodes in the sequoioid clade. Of the targeted 906 genes, sequence capture yielded 436 targets with sufficient coverage for assembly in all ten accessions. HybPiper reports the single-best contig for each target locus, but reports the possibility of paralogs if the second-best contig covers 85% of the reference (Johnson et al., 2016). Of the targets with sufficient coverage, HybPiper reported potential paralogs in at least one taxon for 12 targets, leaving 424 targets that yielded only a single sequence variant.

Assembled contigs for the 424 single variant targets ranged (across genes and accessions) from 102 to 9,240bp. There was no clear difference in the frequency of paralog detection among

species. Specifically, we did not see additional targets with paralog warnings in either the hexaploid *Sequoia* or the tetraploid *Fitzroya*. This fact provides prima facie evidence that these genomes do not contain divergent homeologs.

To investigate the evolutionary relationships both within redwood species and among redwood taxa, we used Bayesian concordance analysis. We first estimated gene trees from the 416 single variant loci with contigs of at least 1000bp long (1,040 to 9,084bp). Concordance analysis of these 416 loci yielded strong support for a population and concordance tree that shows monophyly of *Sequoia* accessions (Fig. 1), with a sample-wide concordance factor mean estimate of 0.972 and a 95% credibility interval (CI) of 0.971 – 0.974. Results of concordance analyses did not demonstrate sensitivity to values of alpha ranging from 0.01 – 1. Within the *Sequoia* clade, relationships among the four accessions were less clear, with various competing topologies supported by low concordance factors with overlapping credibility intervals.

Monophyly of *Sequoiadendron* accessions was also supported by a high concordance factor with a mean estimate of 0.995 (95% CI = 0.995-0.995). The three possible relationships among *Sequoiadendron* accessions were each supported by equal mean concordance factors. The sister group relationship between *Sequoia* and *Sequoiadendron* had a mean concordance factor of 0.741 (95% CI = 0.719 - 0.764). The separation of the sequoioid clade (*Sequoia* + *Sequoiadendron* + *Metasequoia*) from the outgroups (*Thuja* and *Fitzroya*) was strongly supported, with a mean concordance factor of 0.984 (95% CI =0.983 - 0.986) Overall these results provide no evidence for ancient hybridization among these lineages, suggesting allopolyploidy played no role in the evolution of *Sequoia*.

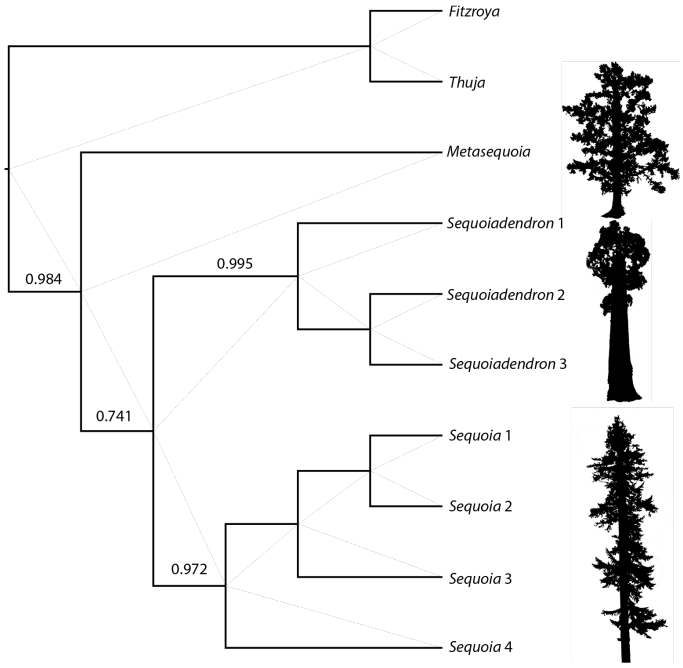

Figure 1: Bayesian concordance analysis of 416 genes. Branch labels are concordance factors (CF). Only concordance factors that are significantly higher (based on overlap of CI) than all conflicting resolutions are given.

To assess the timing of divergence among *Sequoia* accessions relative to divergence among redwood species, we estimated individual phylogenetic trees for each of 416 genes. Of these, 245 genes yielded a best tree that matched the concordance tree, with *Sequoia* accessions monophyletic, *Sequoiadendron* accessions monophyletic, *Sequoia* and *Sequoiadendron* forming a clade, and a sequoioid clade of *Metasequoia* + *Sequoia* + *Sequoiadendron*. Four example phylograms are shown in Figure 2. The other 117 genes included polytomies or supported contradictory topologies.

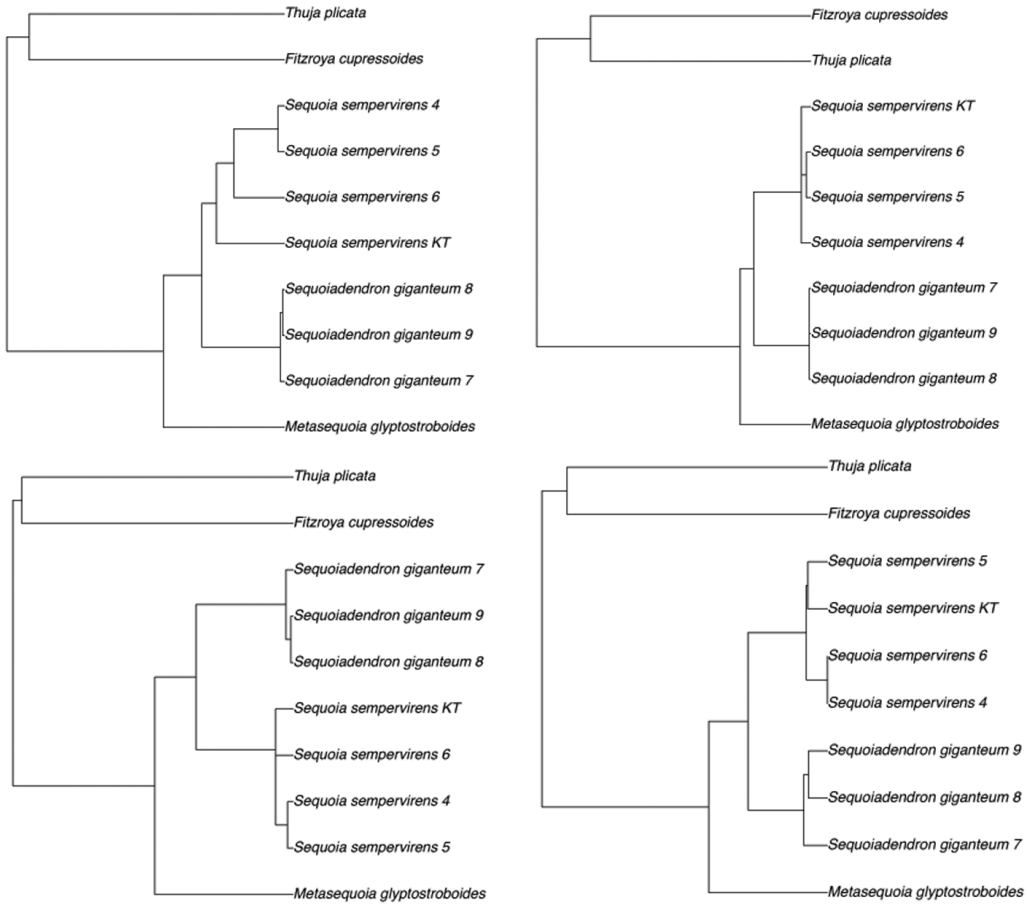

Figure 2: Example trees from analysis of 245 gene trees. Branch lengths are drawn proportional to the maximum likelihood estimate of the average number of substitutions per site.

Using the codeml codon model implemented in PAML, we estimated ω (dN/dS), the ratio of nonsynonymous substitutions per synonymous site to synonymous substitutions per synonymous site for each coding region. Using ω to transform branch lengths, we obtained estimates of branch length in units of Ks, which provides a crude proxy for time. As summarized in Figure 3, the average age of the MRCA of all *Sequoia* accessions was 0.004 (0.000 – 0.063) synonymous substitutions per synonymous site. In comparison, the average divergence between *Sequoia* and

Sequoiadendron was 0.021 (0.004 – 0.064) and the average divergence of *Metasequoia* from *Sequoia* + *Sequoiadendron* was 0.031 (0.005-0.077). To provide a crude conversion to absolute time, if the *Sequoia* – *Sequoiadendron-Metasequoia* split is assigned an age of 140Ma (Lu *et al.*, 2014), the *Sequoia* accessions all coalesce ~18Ma.

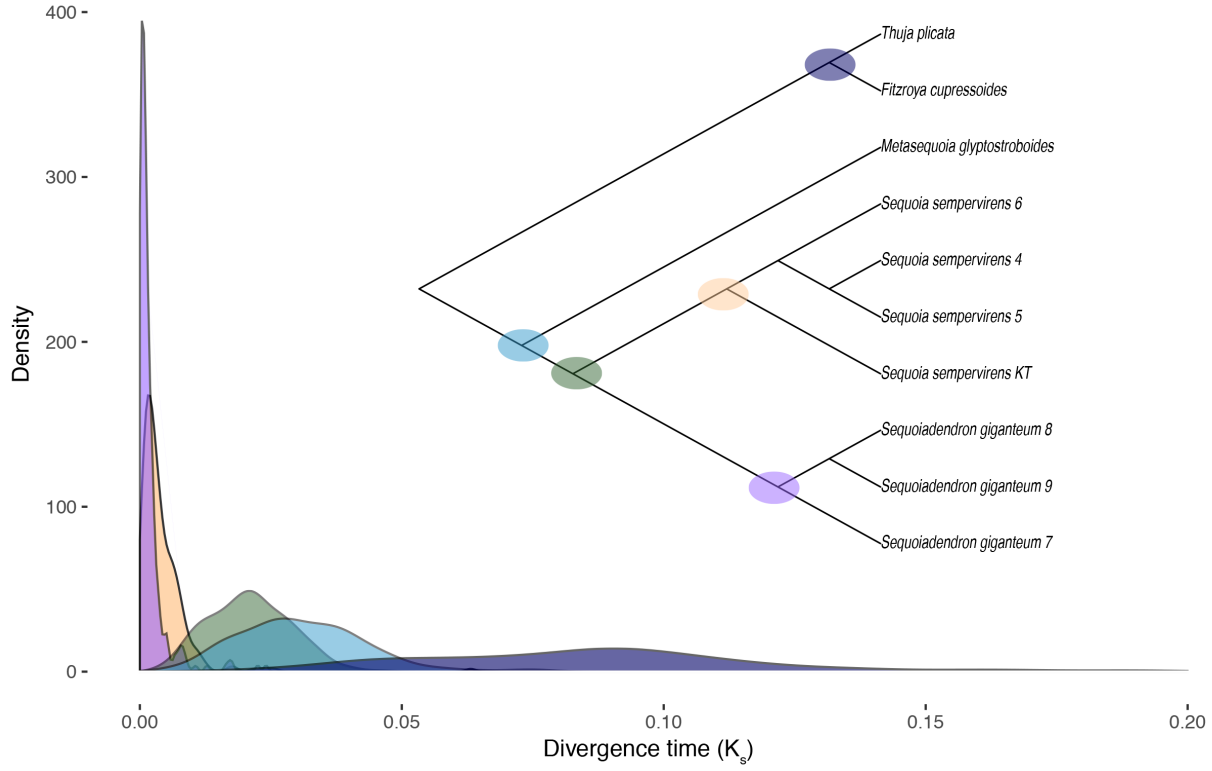

Figure 3: Density plot of tree-based divergence times (in units of Ks)

To help convert the divergence in Ks units into absolute time we generated fossil-calibrated phylogenies for 20 genes and combined their posterior distributions (Table 1). This analysis indicates *Sequoia* and *Sequoiadendron* last shared a common ancestor in the Eocene, as this node had an average age estimate of 50.7Ma. This inferred date of divergence between *Sequoia* and *Sequoiadendron* calls into doubt the identification of Jurassic fossils assigned to *Sequoia* (Endô, 1951), suggesting that these are more likely stem-group sequoioids.

Table 1: Divergence times and 95% HPDs (in parentheses) across 20 loci

Locus	Root	*Fitzroya* + *Thuja*	*Metasequoia* + *Sequoia/Sequoiadendron*	*Sequoia*-*Sequoiadendron*	MRCA *Sequoia*	MRCA *Sequoiadendron*
1	172.25 (150, 202.49)	144.98 (124, 171.98)	67.02 (56.39, 77.46)	42.07 (33, 52.49)	3.43 (0.84, 6.8)	0.94 (0.01, 2.36)
2	178.22 (150, 208.11)	152.09 (124.09, 178.45)	63.86 (53.23, 74.97)	56.88 (39.64, 72.2)	10.61 (1.17, 25.57)	6.7 (0.44, 17.35)
3	171.38 (150.01, 201.02)	152.82 (126.64, 179.65)	65.45 (54.94, 76.4)	50.31 (33.45, 64.55)	8.62 (1.51, 18.79)	3.93 (0.4, 9.39)
4	168.16 (150, 197.89)	147.08 (124, 173.48)	67.77 (56.52, 79.77)	48.32 (33.37, 61.32)	26.95 (13.56, 41.87)	10.71 (1.62, 23.97)
5	176.24 (150.01, 206.3)	149.61 (124.01, 176.68)	64.66 (54.13, 75.52)	54.61 (40.34, 67.92)	9.22 (2.68, 17.29)	8.63 (2.82, 15.6)
6	167.11 (150, 195.97)	144.23 (124.01, 170.62)	68.89 (57.58, 81.02)	46.03 (33, 59)	8.03 (2.2, 15.66)	1.04 (0.04, 2.59)
7	176.75 (150, 206.67)	152.74 (126.1, 180.51)	64.6 (53.97, 75.43)	43.86 (33, 56.62)	4.02 (0.75, 8.56)	0.59 (0, 1.9)
8	172.42 (150, 203.1)	144 (124, 171.65)	67.45 (56.39, 78.42)	42.95 (33, 54.25)	27.95 (16.31, 40.79)	0.73 (0, 2.3)
9	172.91 (150.01, 204.34)	144.49 (124, 172.07)	66.52 (55.5, 77.86)	59.25 (45.42, 73.33)	10.42 (3.14, 19.65)	9.27 (2.77, 17.29)
10	174.74 (150, 205.95)	146.54 (124, 174.74)	66.31 (55.41, 77.7)	45.4 (33.01, 58.78)	22.99 (8.65, 37.58)	4.27 (0.29, 10.69)
11	175.05 (150, 207.84)	138.45 (124, 163.93)	69.35 (57.7, 81.71)	61.56 (47.89, 75.49)	11.81 (0.21, 31.58)	6.37 (0.1, 16.52)
12	185.98 (158.77, 214.99)	145.48 (124.01, 172.25)	63.89 (54.13, 73.86)	43.54 (33, 53.52)	23.14 (13.85, 33.08)	1.37 (0.02, 3.37)
13	173.67 (150, 201.04)	149.32 (124.01, 176.05)	65.03 (55.15, 75.28)	54.24 (38, 68.05)	14.98 (2.29, 54.64)	6.71 (1.24, 21.87)
14	173.36 (150, 203.7)	144.94 (124, 172.31)	66.23 (55.57, 76.87)	48.6 (34.59, 61.28)	9.19 (3.12, 16.37)	1.85 (0.02, 4.64)
15	183.19 (155.54, 213.87)	150.62 (124.03, 176.83)	63.22 (53.5, 73.51)	53.95 (42.14, 66.17)	6.46 (1.91, 11.83)	9.18 (3.78, 15.24)
16	177.44 (150.01, 207.75)	149.43 (124.01, 176.64)	65.25 (54.61, 76.33)	43.96 (33, 56.79)	3.92 (0.77, 8.59)	2.02 (0.21, 4.72)
17	183.45 (154.94, 214.85)	147.36 (124, 175.94)	63.85 (53.4, 74.68)	52.46 (37.49, 66.73)	13.5 (4.63, 23.57)	3.61 (0.42, 7.78)
18	183.38 (150.02, 211.44)	153.21 (125.05, 180.72)	63.25 (50.93, 77.24)	55.17 (33.1, 69.26)	10.74 (0.17, 37.36)	5.73 (0, 29.88)
19	178.25 (150.01, 206.83)	146.99 (124, 174.31)	64.72 (54.73, 75.06)	46.74 (34.4, 58.44)	1.1 (0, 3.33)	12.15 (5.46, 19.52)
20	179.98 (150.2, 209.12)	145.81 (124, 173.85)	64.65 (54.44, 75.21)	58.8 (45.84, 71.43)	34.69 (20.07, 49.72)	4.43 (0.56, 9.39)

The two *Sequoia* accessions, tentatively interpreted as alternate haplotypes, coalesce on average ~13Ma, during the Miocene, but with a wide range across the 20 genes (1-35Ma). All but one of the sampled genes yields a coalescence age for the *Sequioa* clade that is much younger than the inference, based on fossil guard cell data, that *Sequoia* was already hexaploid by the end of the Eocene (33mya; Ma et al., 2005). The node representing the MRCA of *Sequoiadendron* accessions was estimated at 3Ma (1-12 Ma), considerably younger than the MRCA of accessions in its sister genus *Sequoia*.

Discussion

Polyploidy in *Sequoia* did not involve hybridization with *Sequoiadendron*

A sister-group relationship between *Sequoia* and *Sequoiadendron* is supported by Bayesian concordance analysis of 416 loci, which yielded a concordance factor of 0.972. This result is in accord with species relationships reported previously (Brunsfeld *et al.*, 1994; Gadek *et al.*, 2000; Kusumi *et al.*, 2000; Yang *et al.*, 2012; Mao *et al.*, 2012; Leslie *et al.*, 2012; Scott *et al.*, 2016). If hybridization with an extant taxon had played a role in *Sequoia*'s hexaploid evolution, we would not expect a *Sequoia* clade with a 97.2% concordance factor. For example, if there had been ancient hybridization with an ancestor of *Sequoiadendron*, we would expect one-third or two-thirds of *Sequoia* accessions to appear to be more closely related to *Sequoiadendron* than to other *Sequoia* alleles. Thus, analyses of genomic sequences obtained using Hyb-Seq confirm previous conclusions based on transcriptomic analysis.

Sequoia is an undiploidized autopolyploid

To investigate the consequences of whole genome duplication in *Sequoia*, we built phylogenetic trees and estimated divergence times of all nodes using branch lengths in Ks units. As our analyses included multiple accessions of *Sequoia*, we compared intraspecific divergence times to divergence times among species. We found that the mean Ks divergence between the most distance *Sequoia* accessions for a given gene (corresponding to the MRCA of sampled *Sequoia*) was remarkably low (~0.004). Assuming the MRCA of *Sequoia* / *Sequoiadendron* and *Metasequoia* lived 140 Ma (Lu *et al.*, 2014), the *Sequoia* alleles coalesce ~18Ma.

Divergence estimates within Sequoioideae

Using primary fossil calibration, accompanied by secondary calibration points, we estimated divergence times among and within redwood species. We found that the two California redwood genera (*Sequoia* and *Sequoiadendron*) last shared a common ancestor in the mid-Eocene approximately 51Ma (±6Ma). In addition, we found that sampled *Sequoia* sequences coalesce in the Miocene (~13Ma), whereas accessions of *Sequoiadendron* appear to coalesce at the end of the Pliocene (~5 Ma). The former value is similar to, but somewhat younger than the age estimate obtained using a simple, linear calibration of Ks, which yielded an age of ~18 Ma.

The coalescence of *Sequoia* accessions in the Miocene, corresponds to a period with a good sequoioid fossil record. Many fossils from this period can be assigned to the genus *Sequoia*, confirming that the coalescence among sequence variants post dates the group's divergence from the *Sequoiadendron* lineage. For example, Endô (1936) described *Sequoia* fossil specimens from the Miocene of Japan and Korea based on cone and foliar morphology, whereas Zhang et al.

(2015) and Grímsson *et al.* (2007) confirmed the existence of *Sequoia* species other than *S. sempervirens.*

Evidence of homeologous recombination in *Sequoia*

The four sequence variants identified for each gene in *Sequoia*, appear to coalescence approximately 12-18 Ma, much later than the minimum age of the *Sequoia* polyploidization event, which has been dated using measurement of guard cell size to at least 34 Ma (Ma et al., 2005). As discussed by Scott et al. (2016), this finding suggests a long period of multisomic inheritance after polyploidization. Considering the low Ks divergence, cytogenetic observations of multivalent and bivalent pairing, and direct evidence of non-disomic inheritance (Rogers, 1997) it is likely that multisomic inheritance has continued to the present day, at least for some loci. Extended periods of multisomic inheritance allow homeologs to continuously recombine (whether as consistent multivalents or bivalents involving different homeologs each generation). This recombination, combined with genetic drift, would result in a loss of variation among homeologs and young coalescence times. This conclusion highlights an important caveat of using Ks divergence estimates to identify or date WGD events. Because sequence analysis can only estimate the time of homeolog divergence, they provide limited information on when polyploidy became established. Indeed, sequence coalescence can greatly underestimate (in the case of extreme allopolyploidy) or overestimate (in the case of autopolyploidy followed by multisomic inheritance) the age of the WGD event itself.

Consequences of autopolyploidy in *Sequoia*

Whether a polyploid individual is able to persist and ultimately establish as a polyploid lineage depends on a number of factors related to reproduction. Of particular relevance, by multivalent formation is expected to impair fitness due to disrupted segregation and the potential for aneuploidy (Le Comber *et al.*, 2010). As a result, the transition from a polyploid with multisomic inheritance to a diploid with consistent segregation of chromosome pairs and disomic inheritance represents a critical barrier for the establishment of new autopolyploid taxa. Given this, it seems remarkable that *Sequoia* has persisted for at least 34 Ma with extensive multivalent formation and multisomic inheritance.

Continued multisomic inheritance has a number of significant evolutionary consequences. For example, the effective population size at multisomic loci is elevated, resulting in an expected lower rate of genetic drift. At the same time, the presence of additional gene copies might be expected to result in a relaxation of purifying selection and, thus, an elevated rate of molecular evolution (Otto & Whitton, 2000; Hileman & Baum, 2003). Additionally, the constant homogenization of alleles by recombination would tend to slow down or prevent gene copies from undergoing subfunctionalization or neofunctionalization.

Resolving the extent of multisomic inheritance in *Sequoia* is an important prerequisite for future population-level genetic studies, including the design of breeding programs and conservation genetics. Meirmans and Van Tienderen (2013) demonstrated how assumptions about inheritance patterns in polyploids can bias estimates of both genetic diversity and divergence among

populations. In addition, there may be variation in multisomic inheritance among loci when an autopolyploid is not completely rediploidized (e.g. Gaut & Doebley, 1997). As a result, future genetic work is needed to explore the extent of disomic, tetrasomic, and hexasomic inheritance in *Sequoia*.

To maintain *ex situ* collections that reflect the genetic diversity present in natural populations, we have to understand how genetic inheritance works. In the context of redwoods, long a focus for conservation biologists, continued studies of the diploidization process is essential to future conservation genetic research. Conservation efforts are underway to preserve the diversity of coast redwood from throughout its range, with an eye to the threat posed by climatic changes. For example, in coastal California there has been a decrease in fog frequency, a key correlate of coast redwood ecology, by approximately 33% since the start of the 20th century (Johnstone & Dawson, 2010). Likewise, climate projections suggest that the current range of coast redwoods will be both warmer and drier than at present, potentially imposing drought stress that could have devastating consequences. In light of these very real threats, additional studies of coast redwood genetics and genomics should be a priority so as to minimize the chances of losing this iconic and magnificent tree to extinction.

Acknowledgements

Support for this project was provided by a grant to A.D.S. and D.A.B. from Save the Redwoods League. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE-0718123 to A.D.S. This research was

performed using the compute resources and assistance of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

References

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* 8: 135–141.

Ahuja MR, Neale DB. 2002. Origins of polyploidy in coast redwood (Sequoia sempervirens (D. Don) Endl.) and relationship of coast redwood to other genera of Taxodiaceae. *Silvae Genetica* 51: 93–100.

Ané C, Dewey CN, Kotha SK, Larget BR. 2010. Gene tree reconciliation: new developments in Bayesian concordance analysis with BUCKy.

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution* 24: 412–426.

Baum DA. 2007. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon* 56: 417–426.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology* 10.

Brunsfeld SJ, Soltis PS, Soltis DE, Gadek PA, Quinn CJ, Strenge DD, Ranker TA. 1994. Phylogenetic Relationships Among the Genera of Taxodiaceae and Cupressaceae: Evidence from

rbcL Sequences. *Systematic Botany* 19: 253–262.

Le Comber SC, Ainouche ML, Kovarik A, Leitch AR. 2010. Making a functional diploid: From polysomic to disomic inheritance. *New Phytologist* 186: 113–122.

Doyle J. 1945. Naming of the redwoods. *Nature* 155: 254–257.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.

Endô S. 1936. New fossil species of Sequoia from the Far-East. *Proceedings of the Imperial Academy* 12: 172–175.

Endô S. 1951. A Record of Sequoia from the Jurassic of Manchuria. *Botanical Gazette* 113: 228–230.

Gadek PA, Alpers DL, Heslewood MM, Quinn CJ. 2000. Relationships within Cupressaceae sensu lato: a combined morphological and molecular approach. *American journal of botany* 87: 1044–1057.

Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences of the United States of America* 94: 6809–14.

Grímsson F, Denk T, Símonarson LA. 2007. Middle Miocene floras of Iceland—the early colonization of an island? *Review of Palaeobotany and Palynology* 144: 181–219.

Hileman LC, Baum DA. 2003. Why do paralogs persist? Molecular evolution of CYCLOIDEA and related floral symmetry genes in Antirrhineae (Veronicaceae). *Molecular Biology and Evolution* 20: 591–600.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754–755.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ. 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-Throughput Sequencing Reads Using Target Enrichment. *Applications in Plant Sciences* 4: 1600016.

Johnstone J a, Dawson TE. 2010. Climatic context and ecological implications of summer fog decline in the coast redwood region. *Proceedings of the National Academy of Sciences of the United States of America* 107: 4533–8.

Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066.

Kusumi J, Tsumura Y, Yoshimaru H, Tachida H. 2000. Phylogenetic relationships in Taxodiaceae and Cupressaceae sensu stricto based on matK gene, chlL gene, trnL-trnF IGS region, and trnL intron sequences. *American Journal of Botany* 87: 1480–1488.

Larget BR, Kotha SK, Dewey CN, Ané C. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26: 2910–2911.

Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S. 2012. Hemisphere-scale differences in conifer evolutionary dynamics. *Proceedings of the National Academy of Sciences of the United States of America* 109: 16217–21.

Levin DA. 1983. Polyploidy and novelty in flowering plants. *American Naturalist* 122: 1–25.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Science Advances* 1: e1501084–e1501084.

Liu Y-J, Li C-S, Wang Y-F. 1999. Studies on fossil Metasequoia from north-east China and their taxonomic implications. *Botanical Journal of the Linnean Society* 130: 267–297.

Lu Y, Ran J-H, Guo D-M, Yang Z-Y, Wang X-Q. 2014. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PloS one* 9: e107679.

Ma QW, Li FL, Li C Sen. 2005. The coast redwoods (Sequoia, Taxodiaceae) from the Eocene of Heilongjiang and the Miocene of Yunnan, China. *Review of Palaeobotany and Palynology* 135: 117–129.

Mao K, Milne RI, Zhang L, Peng Y, Liu J, Thomas P, Mill RR, Renner SS. 2012. Distribution of living Cupressaceae reflects the breakup of Pangea. *Proceedings of the National Academy of Sciences of the United States of America* 109: 7793–8.

Meirmans PG, Van Tienderen PH. 2013. The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity* 110: 131–137.

Moustafa A, Bhattacharya D. 2008. PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of Chlamydomonas. *BMC Evolutionary Biology* 8: 6.

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, *et al.* 2013. The Norway spruce genome sequence and conifer genome evolution. *Nature* 497: 579–584.

Otto SP, Whitton J. 2000. Polyploid Incidence and Evolution. *Annual Review of Genetics* 34: 401–437.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria* 0: {ISBN} 3–900051–07–0.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. *PLoS ONE* 6.

Rogers DL. 1997. Inheritance of allozymes from seed tissues of the hexaploid gymnosperm, Sequoia sempervirens (D. Don) Endl. (Coast redwood). *Heredity* 78: 166–175.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Saylor LC, Simons HA. 1970. Karyology of Sequoia sempervirens: Karyotype and Accessory Chromosomes. *Cytologia* 35: 294–303.

Schlarbaum SE, Tsuchiya T, Johnson L. 1984. The chromosomes and relationships of Metasequoia and Sequoia (Taxodiaceae) an update. *Journal of the Arnold Arboretum* 65: 251–

254.

Scott AD, Stenz NWM, Ingvarsson PK, Baum DA. 2016. Whole genome duplication in coast redwood (Sequoia sempervirens) and its implications for explaining the rarity of polyploidy in conifers. *New Phytologist* 211: 186–193.

Šmarda P, Vesely P, Šmerda J, Bureš P, Knápek O, Chytrá M. 2016. Polyploidy in a 'living fossil'Ginkgo biloba. *New Phytologist* 212: 11–14.

Soltis DE, Bell CD, Kim S, Soltis PS. 2008. Origin and early evolution of angiosperms. *Annals of the New York Academy of Sciences* 1133: 3–25.

Stamatakis A. 2006. The RAxML 7.0.4 Manual. *Bioinformatics* 22(21): 2688–2690.

Stebbins GL. 1948. The Chromosomes and Relationships of Metasequoia and Sequoia. *Science* 108.

Wickham H. 2009. *ggplot2*: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* 2: 333–341.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The

frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106: 13875–13879.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences: CABIOS* 13: 555–556.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.

Yang Z, Ran J, Wang X. 2012. Molecular Phylogenetics and Evolution Three genome-based phylogeny of Cupressaceae s.l.: Further evidence for the evolution of gymnosperms and Southern Hemisphere biogeography.

Zhang JW, D'Rozario A, Adams JM, Li Y, Liang XQ, Jacques FM, Su T, Zhou ZK. 2015. Sequoia maguanensis, A new miocene relative of the coast redwood, Sequoia sempervirens, From china: Implications for paleogeography and paleoclimate. *American Journal of Botany* 102: 103–118.