The Plant Cell, Vol. 31: 1466–1487, July 2019, www.plantcell.org © 2019 ASPB.

LARGE-SCALE BIOLOGY ARTICLE

Breaking Free: The Genomics of Allopolyploidy-Facilitated Niche Expansion in White Clover[OPEN]

Andrew G. Griffiths,[a,1] Roger Moraga,[a] Marni Tausen,[b,c] Vikas Gupta,[b] Timothy P. Bilton,[d] Matthew A. Campbell,[e] Rachael Ashby,[d] Istvan Nagy,[f] Anar Khan,[d] Anna Larking,[a] Craig Anderson,[a] Benjamin Franzmayr,[a] Kerry Hancock,[a] Alicia Scott,[a] Nick W. Ellison,[a] Murray P. Cox,[e] Torben Asp,[f] Thomas Mailund,[c] Mikkel H. Schierup,[d,g] and Stig Uggerhøj Andersen[b,1]

[a] AgResearch, Grasslands Research Centre, Palmerston North 4442, New Zealand
[b] Department of Molecular Biology and Genetics, Aarhus University, 8000 Aarhus C, Denmark
[c] Bioinformatics Research Centre, Aarhus University, 8000 Aarhus C, Denmark
[d] AgResearch, Invermay Agricultural Centre, Mosgiel 9053, New Zealand
[e] Bioinformatics and Statistics Group, Institute of Fundamental Sciences, Massey University, Palmerston North 4410, New Zealand
[f] Department of Molecular Biology and Genetics, Aarhus University, 200 Slagelse, Denmark
[g] Department of Bioscience, Aarhus University, 8000 Aarhus C, Denmark

ORCID IDs: 0000-0002-0573-1668 (A.G.G.); 0000-0001-7806-1135 (R.M.); 0000-0003-0694-9199 (M.T.); 0000-0002-4590-4463 (V.G.); 0000-0001-5945-3766 (T.P.B.); 0000-0002-5826-0329 (M.A.C.); 0000-0002-1938-3027 (R.A.); 0000-0001-6549-9525 (I.N.); 0000-0003-0041-4679 (A.K.); 0000-0002-5871-0027 (A.L.); 0000-0003-3416-2027 (C.A.); 0000-0002-6595-7319 (B.F.); 0000-0001-8543-6310 (K.H.); 0000-0002-4755-0144 (A.S.); 0000-0001-5506-8342 (N.W.E.); 0000-0003-1936-0236 (M.P.C.); 0000-0002-6470-2410 (T.A.); 0000-0001-6206-9239 (T.M.); 0000-0002-5028-1790 (M.H.S.); 0000-0002-1096-1468 (S.U.A.)

The merging of distinct genomes, allopolyploidization, is a widespread phenomenon in plants. It generates adaptive potential through increased genetic diversity, but examples demonstrating its exploitation remain scarce. White clover (*Trifolium repens*) is a ubiquitous temperate allotetraploid forage crop derived from two European diploid progenitors confined to extreme coastal or alpine habitats. We sequenced and assembled the genomes and transcriptomes of this species complex to gain insight into the genesis of white clover and the consequences of allopolyploidization. Based on these data, we estimate that white clover originated ~15,000 to 28,000 years ago during the last glaciation when alpine and coastal progenitors were likely colocated in glacial refugia. We found evidence of progenitor diversity carryover through multiple hybridization events and show that the progenitor subgenomes have retained integrity and gene expression activity as they traveled within white clover from their original confined habitats to a global presence. At the transcriptional level, we observed remarkably stable subgenome expression ratios across tissues. Among the few genes that show tissue-specific switching between homeologous gene copies, we found flavonoid biosynthesis genes strongly overrepresented, suggesting an adaptive role of some allopolyploidy-associated transcriptional changes. Our results highlight white clover as an example of allopolyploidy-facilitated niche expansion, where two progenitor genomes, adapted and confined to disparate and highly specialized habitats, expanded to a ubiquitous global presence after glaciation-associated allopolyploidization.

INTRODUCTION

Polyploidy, where more than two genomes reside in a single nucleus, is widely distributed among eukaryotes and is particularly prevalent in angiosperms (Otto and Whitton, 2000; Mable, 2003; Wood et al., 2009), where it is considered a driver of speciation and biodiversity (Leitch and Leitch, 2008). Genome duplication leads to autopolyploidy, whereas interspecific hybridization followed by genome doubling or fusion of unreduced gametes results in allopolyploidy, where divergent homeologous subgenomes reside within the same nucleus. These duplications can be recent (<150 years; Ownbey, 1950; Soltis et al., 2004; Ainouche et al., 2009; Chester et al., 2012) or ancient, extending back to the origin of seed plants 350 million years ago (Mya; Jia and International Wheat Genome Sequencing Consortium et al., 2013; Li et al., 2015). In many lineages, such duplications are a recurring phenomenon (Soltis et al., 2016; Soltis and Soltis, 2016). Many neopolyploids likely become extinct shortly after formation, and there is much debate over the implications of polyploidization and whether it is an evolutionary jump-start or dead-end (Otto, 2007; Madlung, 2013). Polyploidy events can confer enhanced fitness, phenotypic plasticity, and adaptability, and have been correlated with survival in stressful environments, such as during the

[1] Address correspondence to: andrew.griffiths@agresearch.co.nz or sua@mbg.au.dk.
The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Andrew G. Griffiths (andrew.griffiths@agresearch.co.nz) and Stig Uggerhøj Andersen (sua@mbg.au.dk).
[OPEN]Articles can be viewed without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.18.00606

IN A NUTSHELL

Background: White clover (*Trifolium repens*) is found in lawns and grasslands around the world. It is used in agriculture for high quality animal feed, where its ability to fix atmospheric nitrogen in symbiosis with soil bacteria reduces the need for application of chemical fertilizers. White clover is an allopolyploid, which means that it harbors two distinct subgenomes as a result of the merging of genetic material (genomes) from two ancestor clovers. In contrast to white clover, these European ancestors are confined to specialized habitats. One is found in high alpine screes and the other on the coast within 100 meters of the shore.

Question: We wanted to understand how and when the two ancestors merged to form white clover and how the ancestral genomes responded to being joined within the same plant.

Findings: We confirmed the identity of white clover's ancestors using genomic data and found that white clover originated during the last European glaciation ~15–28,000 years ago. We determined that white clover arose from multiple genome merging events, enabling carry-over of variation from its ancestors into the genetically diverse white clover of today. We observed that the ancestral genomes had remained largely intact within white clover through the thousands of generations since they merged. This has left white clover with a very large number of active genes. We could show that white clover has exploited this genetic potential by switching from activating a gene from specific pathways in one subgenome to activating the matching gene in the other subgenome depending on the tissue. This enables white clover to fine-tune its complement of flavonoids and other secondary metabolites that are important for interactions with microbes and the environment. Our results highlight white clover as an example of allopolyploidy-facilitated habitat expansion, where two ancestor genomes, adapted and confined to very different and highly specialized environments, expanded to a global presence following merging during the last European glaciation.

Next steps: We are now investigating how white clover and its subgenomes respond to different environmental cues, and how these responses influence ecological success.

Cretaceous-Paleogene mass extinction (Gross et al., 2004; Leitch and Leitch, 2008; Fawcett et al., 2009; Vanneste et al., 2014a, 2014b; Selmecki et al., 2015). Consequences of genome duplication, particularly for allopolyploids, can include rapid gene loss and rearrangement (McClintock, 1984; Rapp et al., 2009; Grover et al., 2012; Tayalé and Parisod, 2013; Yoo et al., 2013; Garsmeur et al., 2014; Woodhouse et al., 2014) through a range of mechanisms in which transposable element activity is implicated (Woodhouse et al., 2014). Many studied allopolyploids show evidence of a genomic or transcriptomic response to accommodating divergent genomes within a single nucleus, and these processes can occur within very few generations (Chester et al., 2012; Grover et al., 2012; Tayalé and Parisod, 2013).

White clover (*Trifolium repens*; Tr) is an example of a relatively recent allopolyploid that likely arose during the last major glaciations 13,000ya to 130,000ya (Williams et al., 2012). A successful and highly variable perennial allotetraploid (AABB-type genome; $2n=4x=32$) legume exhibiting disomic inheritance (Williams et al., 1998), it has a relatively compact genome (1C = 1,093 megabases [Mb]; Bennett and Leitch, 2011) arranged in small similar-sized chromosomes with few obvious features to distinguish homeologs (Ansari et al., 1999). Furthermore, its outbreeding nature leads to significant sequence polymorphism and highly heterogeneous populations (Aasmo Finne et al., 2000; Zhang et al., 2010). White clover has a widespread natural range encompassing grasslands of Europe, Western Asia, and North Africa across latitudes and altitudes (Figure 1; Daday, 1958). Due to this ability to occupy a wide range of climatic conditions, white clover is used extensively as a companion forage in moist temperate agriculture, thereby extending its range globally (Figure 1; Daday, 1958; Abberton et al., 2006). Extant relatives of the paternal and maternal progenitors of white clover have been identified as western clover (*Trifolium occidentale*; To) and pale clover (*Trifolium pallescens*; Tp; Ellison et al., 2006; Williams et al., 2012). In contrast with white clover, these diploid ($2n=2x=16$) species have very limited and disparate extreme habitats at the margins of, but not overlapping with, the range of the white clover. The creeping *T. occidentale* is confined to within ~100 m of the shore in a maritime niche on the western coasts of Europe (Coombe, 1961), whereas the non-creeping *T. pallescens* is restricted to European alpine habitats at altitudes between 1,800 and 2,700 m (Figure 1; Raffl et al., 2008). The allopolyploidization event giving rise to white clover is thought to be relatively recent due to the lack of significant divergence of genic sequences from *T. occidentale* and its corresponding white clover subgenome, and is hypothesized to have occurred during the last major glaciations when environmental conditions brought alpine and maritime species in close proximity in glacial refugia (Williams et al., 2012). In contrast to white clover, which occupies markedly different environments to *T. occidentale* and *T. pallescens*, most allopolyploids with extant progenitors share similar habitats with either progenitor (Soltis et al., 2016).

Here, we sequenced and analyzed the full genomes and transcriptomes of white clover and extant relatives of its diploid progenitors to gain insight into the timing, process and consequences of the allopolyploidization event.

RESULTS

Assembly of White Clover and Progenitor Genomes

We sequenced inbred individuals of white clover and extant relatives of its diploid progenitors, *T. occidentale* and *T. pallescens*, using a range of Illumina sequencing libraries, including white clover TruSeq Synthetic Long-Reads (TSLRs; Supplemental Table 1). Extant progenitor genome size estimates derived from

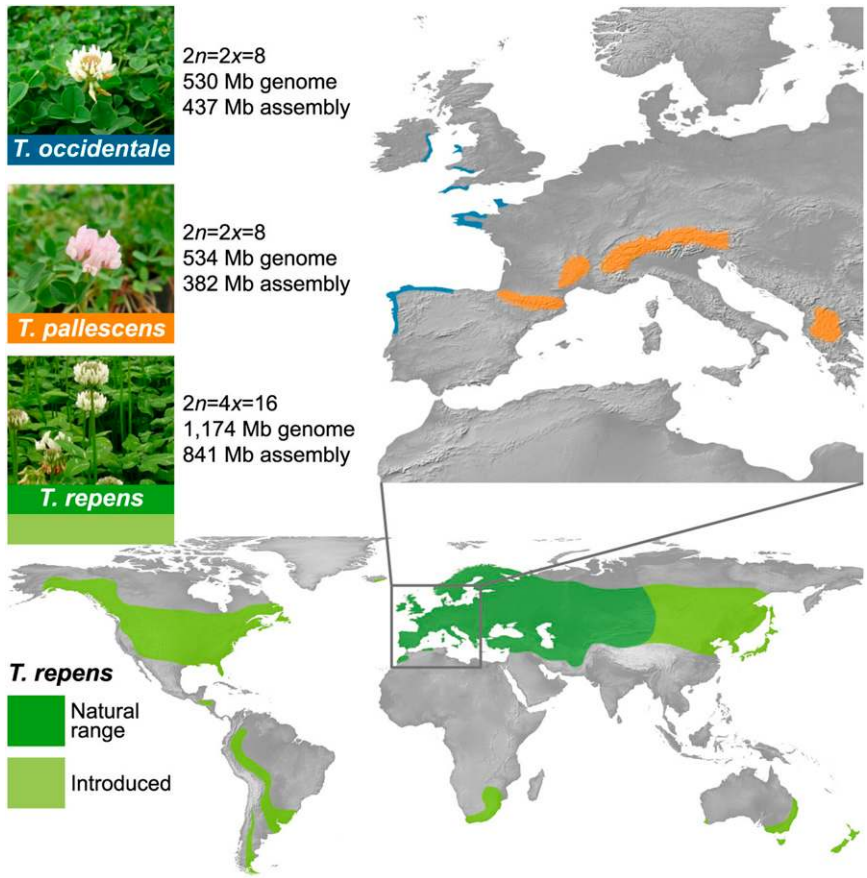

Figure 1. The Range of White Clover and Extant Relatives of its Progenitors.

The present-day ranges of white clover (*T. repens*, Daday, 1958; green), and extant relatives of its diploid progenitors *T. occidentale* (blue) and *T. pallescens* (orange). *T. occidentale* is found within ~100 m of the seashore while *T. pallescens* grows in alpine regions between 1,800 and 2,700 m.

the sequence data were similar (~530 Mb), together equaling approximately the 1,174-Mb size estimate of the white clover genome (Figure 1; Table 1; Supplemental Figure 1). *T. occidentale* and *T. pallescens* assemblies encompassed, respectively, 437 Mb (N_{50} = 192 kilobases [kb]) and 382 Mb (N_{50} = 173 kb) of sequence, accounting for 82% and 72% of the predicted genome sizes (Figure 1; Table 1). The white clover assembly spanned 841 Mb (N_{50} = 122 kb), and approximated the sum of the assemblies of the extant progenitors—suggesting that conflation of homeologous sequences had been successfully eliminated (Figure 1; Table 1). The *T. pallescens* assembly was more fragmented than the other species, which was expected due to the lower sequencing data input (Table 1; Supplemental Table 1).

Scaffold ordering, pseudomolecule construction, and assignment to subgenomes was guided by pairwise linkage disequilibrium (LD, calculated as r^2) between ~7,300 Genotyping-by-Sequencing (GBS) single nucleotide polymorphism (SNP) markers on 3,364 scaffolds (Figure 2A; Supplemental Figure 2) and alignment to the model forage legume *Medicago truncatula* genome (Mt4.0; Tang et al., 2014). Incorporation of single-locus homeolog-specific (SLHS) simple sequence repeat (SSR) markers (Supplemental Table 2) in the LD analysis enabled alignment of pseudomolecules and homeologs with a previous SSR-based genetic linkage map (Griffiths et al., 2013; Supplemental Figures 2C and 2D).

The white clover pseudomolecules were used to order the assemblies for *T. occidentale* and *T. pallescens*, and the resulting pseudomolecules were of similar size for the progenitors and their corresponding white clover subgenomes (Figure 2B; Supplemental Table 3). As an assessment of genome coverage in the assemblies, whole genome shotgun Illumina sequence reads (180-bp insert size; Supplemental Table 1) representing 70×, 80×, and 35× depth for *T. occidentale*, *T. pallescens*, and white clover, respectively, were mapped back to the corresponding assemblies, yielding a mapping rate of ~95% (Supplemental Table 4). This indicates the assemblies encompass a high proportion of their genomes. Alignment of the white clover *T. occidentale*- and *T. pallescens*-derived subgenomes (Tr_{To} and Tr_{Tp}, respectively) with the Mt4.0 genome (Tang et al., 2014) identified general macrosynteny with some major rearrangements relative to *M. truncatula* (Figure 2C), consistent with previous results (Griffiths et al., 2013).

Gene annotation for all three species was enhanced by use of RNA-sequence (RNA-seq) assemblies derived from paired-end reads of transcripts expressed in floral, leaf, root, and stolon/shoot

Table 1. Genome Assembly Statistics for White Clover (*T. repens*) and Extant Relatives of its Diploid Progenitors, *T. occidentale* and *T. pallescens*

Summary Metrics	***T. occidentale***	***T. pallescens***	***T. repens***
Estimated Genome Size (bp)			
ALLPATHS-LG estimate	530,459,686	534,411,997	1,174,038,073
k-mer ($k = 17$)	581,040,135	591,946,893	1,105,102,999
1C genome size[a]	ND	ND	1,093,000,000
Total assembly length	436,797,749	382,382,469	841,400,916
Estimated genome coverage	82%	72%	72%
Chloroplast genome	133,780	130,394	132,086
Assembly Metrics			
Assembled scaffolds (>1,000 bp)	7,229	6,923	22,100
Longest scaffold (bp)	2,576,867	1,323,392	734,507
Average length (bp)	61,288	55,234	38,072
N_{10} (bp; *no. in category*)	599,981; *49*	571,804; *48*	318,557; *201*
N_{50} (bp; *no. in category*)	191,714; *624*	172,944; *586*	121,657; *2,016*
N_{90} (bp; *no. in category*)	41,953; *2,416*	34,989; *2,451*	19,736; *8,100*
Total undefined bases (Ns)	51,627,539	65,642,595	86,144,561
Gap ratio	11.8%	17.2%	10.2%
BUSCO[b] Scores of Assembly (based on 1,440 reference genes)			
Complete genes—total	1,353 (94%)	1,353 (94%)	1,321 (92%)
Complete genes—single copy	*1,203* (*84%*)	*1,197* (*83%*)	*494* (*34%*)
Complete genes—duplicated	*150* (*10%*)	*156* (*11%*)	*827* (*57%*)
Fragmented genes	37 (3%)	33 (2%)	39 (3%)
Missing genes	50 (3%)	54 (4%)	80 (6%)

[a]Flow cytometry (Bennett and Leitch, 2011); ND = Not Determined.
[b]BUSCO, Benchmarked Universal Single-Copy Orthologs (Simão et al. (2015).

tissue (Pooled data set) from each species, with an average of 77 million reads per sample (Supplemental Tables 4 and 5). Of the raw reads, ~88%, 96%, and 95% mapped to the *T. occidentale*, *T. pallescens*, and white clover genome sequences, respectively, indicating these assemblies encompass most of the gene space. The high read coverage facilitated gene annotation, with most gene models based directly on transcriptional evidence (Supplemental Table 5). The number of protein-coding genes was similar for the progenitors and the number of protein-coding genes in white clover was close to the sum of the contributing progenitors (Supplemental Table 6).

To and Tp Are the Progenitors of White Clover

Identification of the extant relatives of white clover's diploid progenitors (Figures 1 and 3A**)** had been based in part on similarities with discrete portions of the chloroplast genome as well as Internal Transcribed Spacer (ITS) sequences of the *T. pallescens* and *T. occidentale*, respectively (Ellison et al., 2006). Expanding this comparison with the complete chloroplast genomes of white clover and its extant progenitors, the alignment with *T. pallescens* revealed 99.5% identical bases and a mean coverage of 100% (Supplemental Table 7), confirming *T. pallescens* as the likely chloroplast donor (Supplemental Figure 3). Furthering this analysis, we surveyed the data generated from whole-genome resequencing (~49-fold coverage) of four outbred white clover individuals from different populations (Supplemental Table 8). Mapping white clover reads from each individual against the chloroplast genomes of white clover, *T. occidentale*, and *T. pallescens* identified high similarity with *T. pallescens* (mean = 54 SNP variants), whereas there were many differences relative to the *T. occidentale* chloroplast (mean = 563 SNP variants; Supplemental Table 9). In summary, sequencing of five white clover individuals (the S_9 and four resequenced individuals) confirm *T. pallescens* as the likely source of the white clover chloroplast and, therefore, the maternal progenitor of white clover.

The 18S-ITS1-5.8S-ITS2-26S rDNA region of the nuclear genome harboring the ITS sequences forms a gene cluster comprising many hundreds of tandem repeat arrays located in the nucleolar organizing region (NOR; Álvarez and Wendel, 2003). Previous analysis of PCR-amplified white clover ITS sequences had shown them to be either identical or contain a SNP variant to those of *T. occidentale*, and only the *T. occidentale* -NOR was detected (Ellison et al., 2006; Williams et al., 2012). In this study, we surveyed the 668,571 white clover TSLRs for ITS sequences and identified 774 full matches spanning the 623-bp ITS1-5.8S rRNA-ITS2 region. Alignment with extant progenitor ITS data identified 770 TSLRs containing SNPs diagnostic of *T. occidentale*-derived ITS and, additionally, four TSLRs diagnostic of *T. pallescens*-derived ITS, supporting confirmation of *T. occidentale* and *T. pallescens* as progenitors (Supplemental Figure 4). The prevalence of the *T. occidentale*-ITS sequence supports observations that the active NOR in white clover is *T. occidentale*-derived (Williams et al., 2012), whereas detection of *T. pallescens*-derived ITS is the first discovery of a remnant Tp-NOR. This is further evidence that *T. pallescens* contributed to the white clover nuclear genome in addition to the chloroplast genome. The reduction in the *T. pallescens*-NOR is consistent with nucleolar dominance observed in allopolyploids whereby the rDNA locus of one progenitor is often silenced and subsequently reduced or eliminated from the genome (Ansari et al., 2008; Kovarik et al., 2008).

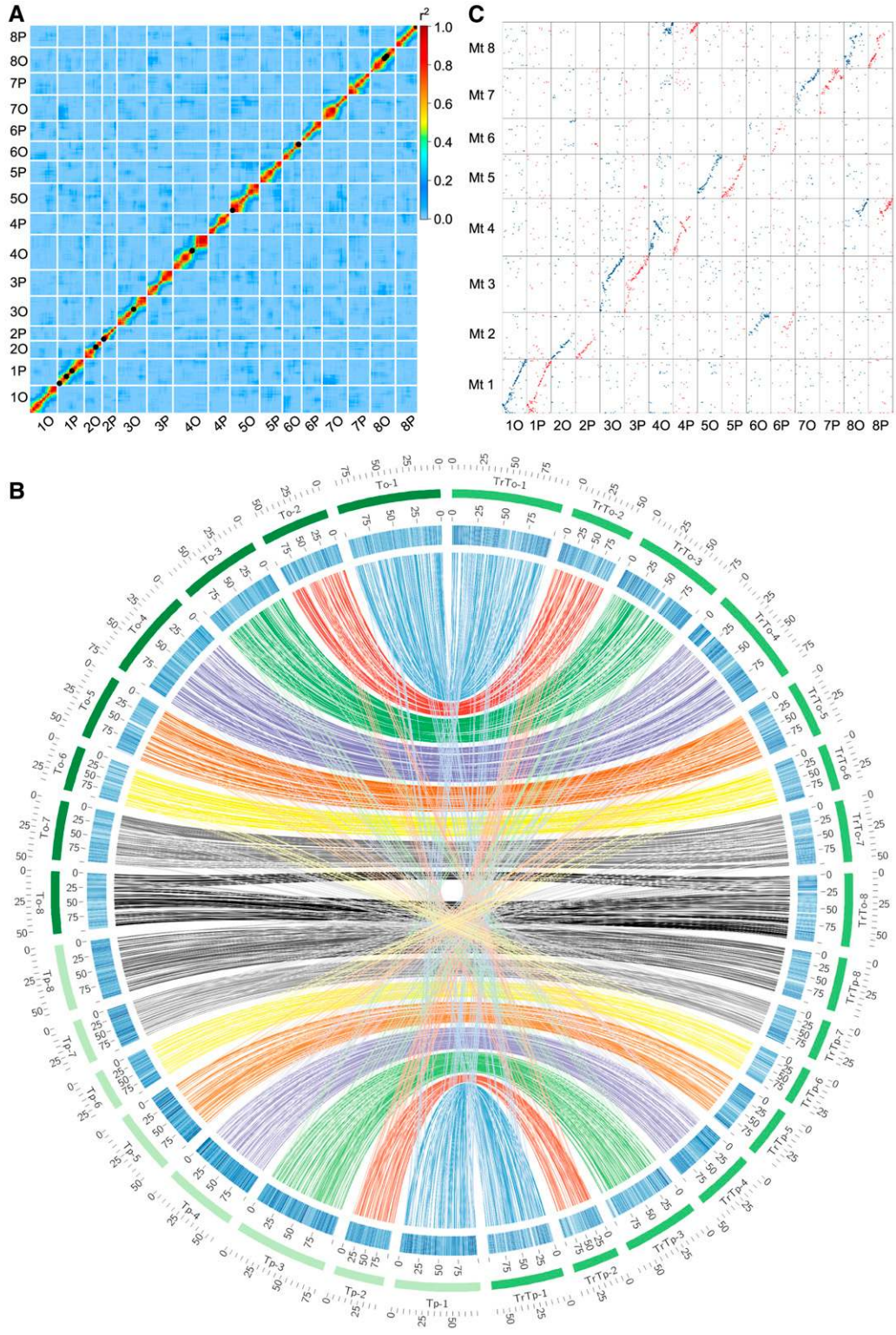

Figure 2. White Clover Genetic Map and Synteny with Extant Progenitor and Model Forage Legume Genomes.

(A) Genetic map based on LD analysis of ~7,300 scaffold-mapped markers from a F_1 biparental mapping population ($n = 93$). Red color indicates high level of pairwise LD (r^2) between markers. Black dots indicates SLHS microsatellite markers from an existing, colinear genetic linkage map (Griffiths et al., 2013; Supplemental Figure 2). Progenitor origin of LG homeologs/subgenomes is denoted by "O" (*Trifolium occidentale*) and "P" (*T. pallescens*).

Expanding comparisons beyond chloroplast and ITS evidence, we also assessed whole genome similarities between extant progenitors and their corresponding white clover subgenomes. We identified SNPs specific either to subgenome/progenitor pairs (Tr_{To}=To≠Tr_{Tp}=Tp) or to each progenitor genome (To≠Tr_{To}=Tr_{Tp}=Tp; Tp≠Tr_{To}=Tr_{Tp}=To) using the software Hybrid Lineage Transcriptome Explorer (HyLiTE; Duchemin et al., 2015), which determined genic SNP subgenome and genome identity based on transcriptomic and genomic data from the three species mapped to *T. occidentale* or *T. pallescens* reference gene models. We found ~8,000 SNPs specific to each progenitor in transcribed regions, whereas we identified ~50-fold more shared between each progenitor and its corresponding subgenome (Supplemental Table 10**)**. This indicates substantial conservation of SNPs between the progenitor/subgenome pairs, providing further evidence that *T. occidentale* and *T. pallescens* are the likely progenitors. In conclusion, the combined ITS, chloroplast, and broader genomic evidence confirmed that white clover is derived from *T. occidentale* and *T. pallescens* progenitors, with *T. pallescens* as the likely maternal progenitor.

White Clover Originated ¡15,000 to 28,000 Years Ago During the Last Glaciation

The analysis of genic SNPs specific to *T. occidentale* and *T. pallescens* using HyLiTE (Duchemin et al., 2015) detected relatively few and similar numbers of genic SNPs (~8,000) specific to each extant progenitor, suggesting similar and possibly recent divergence times between each progenitor and its corresponding white clover subgenome (Supplemental Table 10**)**. We investigated this further by estimating the timing of the white clover allopolyploidization event using an isolation model (Mailund et al., 2012) that analyzed pairwise alignments of progenitor genomes (*T. occidentale* and *T. pallescens*) and white clover subgenomes (Tr_{To} and Tr_{Tp}; Figure 3A). Whole pseudomolecule alignment of three genome pairs (To versus Tp; To versus Tr_{To}; and Tp versus Tr_{Tp}) produced 10k to 32k alignments with average lengths ranging from 7.2 to 8.0 kb for each comparison (Supplemental Figures 5 to 7**;** Supplemental Table 11). As the model assumes neutral polymorphic sites, we masked genes in the alignments to remove regions likely to be under selection. The mutation rate for white clover is unknown, but a base substitution rate of 6.5×10^{-9} per generation has been determined in Arabidopsis (*Arabidopsis thaliana*; Ossowski et al., 2010), and mutation rate appears to increase with genome size (Lynch, 2010). Based on these observations, we estimated the mutation rates of white clover and its progenitors to range between 1.1×10^{-8} and 1.8×10^{-8} per base per generation (Supplemental Figure 8). This aligns with the allotetraploid peanut (*Arachis hypogaea*)-estimated genome-wide mutation rate of 1.6×10^{-8} (Bertioli et al., 2016). As the divergence time scales with the mutation rate used, the mutation rate directly influences divergence time estimates. Based on a single generation per year, the Arabidopsis 6.5×10^{-9} mutation rate placed the white clover progenitors divergence ~530 thousand years ago (Kya), with the progenitor-subgenome divergence 43 Kya to 48 Kya (Supplemental Table 11). However, when using the proxy clover mutation rates, the estimated progenitor divergence was ~191 Kya to 313 Kya, and the progenitors and corresponding subgenomes divergence was 15 Kya to 28 Kya (Figures 3B and 3C; Supplemental Table 11). Assuming the timing of the allopolyploidization event equated to that of progenitor genome divergence from the white clover subgenomes (Figure 3A), the assessed mutation rates placed the allopolyploidization within the glaciation period that culminated in the last European Glacial Maximum ~20 Kya (Yokoyama et al., 2000; Mangerud et al., 2004), whereas the proxy clover rates coincided with the Glacial Maximum itself (Figure 3C). This was a period when the alpine and coastal progenitors were likely to be in close proximity in glacial refugia.

White Clover Arose from Multiple Hybridization Events

There is a range of allopolyploidization scenarios (Ramsey and Schemske, 1998; Mason and Pires, 2015) that may have given rise to white clover, resulting in different genomic diversity signatures. Common examples rely on unreduced gamete production and may include scenarios such as a single hybridization and chromosome doubling event (A×B=AB→AABB), or an interspecific hybridization of unreduced gametes (AA×BB=AABB). If the genesis of white clover resulted from a single hybridization event, this would be associated with a severe genetic bottleneck. We assessed this by analyzing whole-genome resequencing data (~49-fold coverage) from four outbred white clover individuals from different populations (Supplemental Table 8). These genomes were then evaluated with Pairwise Sequentially Markovian Coalescent (PSMC) models (Li and Durbin, 2011) using the mutation rates described above ranging from 6.5×10^{-9} to 1.8×10^{-8} per site per generation. The PSMC model estimated current and past effective population sizes (EPS or N_e) based on mutation rates of the four genomes and their subgenomes, which also infers EPS of the pre-allopolyploidization progenitor populations. For three of the four individuals, the model suggested only a mild genetic bottleneck 2k to 15k generations ago when the EPS dropped to ~20k from a pre-bottleneck level of ~48k (Figure 3D;

Figure 2. (continued).

(B) Circos (circos.ca) diagram showing inter-pseudomolecule relationships between the subgenomes of white clover (Tr_{To}; Tr_{Tp}) and their progenitors (To; Tp) in these assemblies. The outer ring (green-shaded) represents pseudomolecules in megabases (Mb), and the inner ring (blue) depicts gene density as a proportion along the pseudomolecules (%). Colored lines represent synteny blocks constructed by whole genome alignment using the program LASTZ (Harris, 2007) comprising matches >33 kb in length. Blocks within 100-kb windows were merged and represented by a single line. Cross-progenitor links indicate regions of high conservation between the subgenome and both progenitors, not putative recombination events.

(C) A matrix plot assessment of synteny between white clover subgenomes ("O" and "P") and the reference genome of *M. truncatula*, a model forage legume with eight pseudomolecules (Mt 1 to Mt 8). Synteny was based on alignment of the 3,364 LD-ordered white clover anchor scaffolds with *M. truncatula* genome Mt4.0 (Tang et al., 2014).

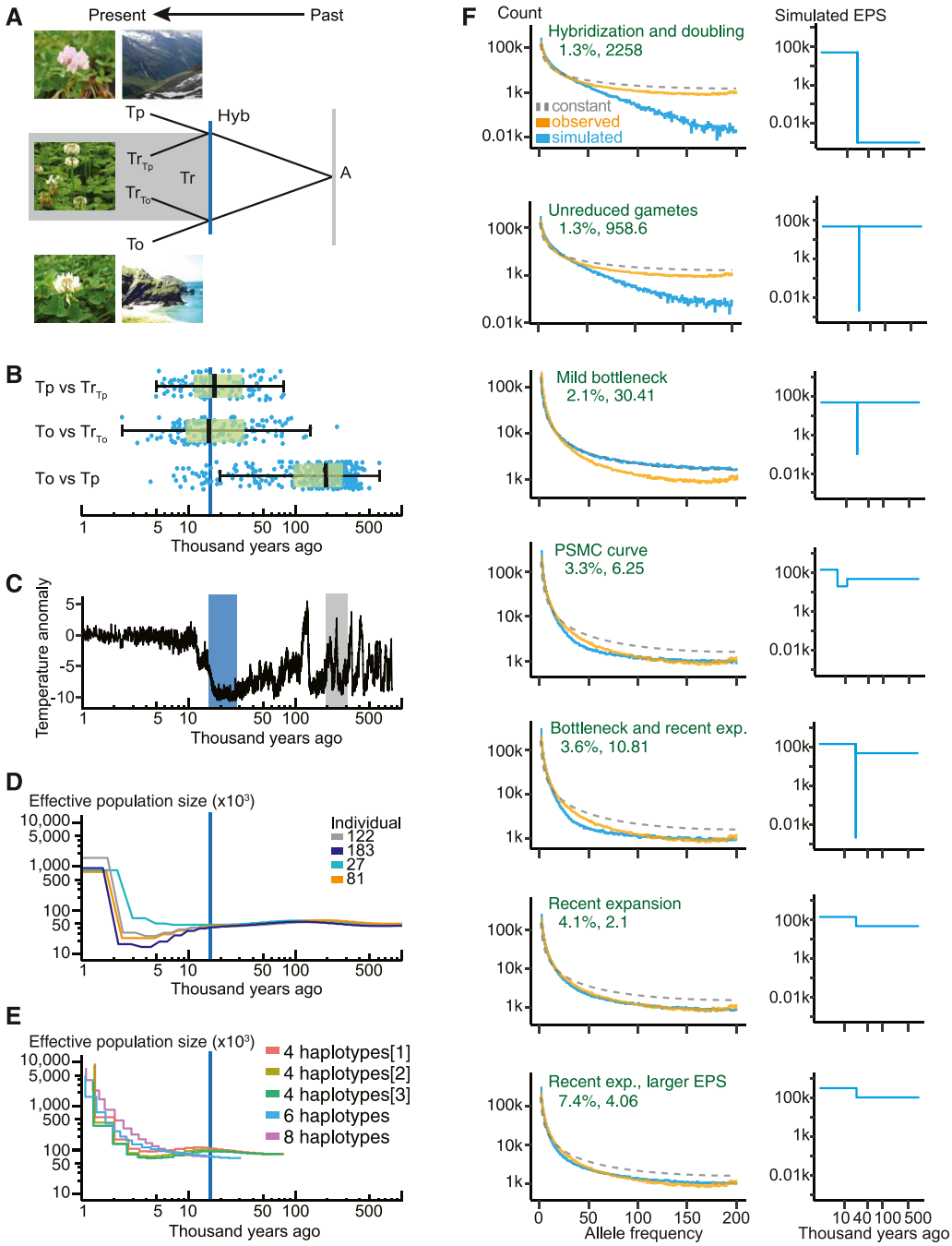

Figure 3. The White Clover Allopolyploidization Event.

(A) Schematic phylogenetic tree for white clover (Tr) and its progenitors *T. occidentale* (To) and *T. pallescens* (Tp). A common ancestor, "A," gave rise to To and Tp (gray line), which hybridized (Hyb, blue line) to give rise to the two subgenomes Tr_{To} and Tr_{Tp} in allopolyploid white clover.
(B) Blue dots indicate split time estimates based on sets of 100 alignment blocks derived from the pairwise genome comparisons indicated, e.g. To versus Tp. Dots are superimposed on box-and-whiskers plots, where the median is indicated by a black vertical line. Using a mutation rate of 1.8×10^{-8}, split time estimates suggest To and Tp diverged from a common ancestor ~192 Kya, whereas white clover Tr_{To} and Tr_{Tp} subgenomes split from their extant progenitors 15 Kya and 17 Kya, respectively, suggesting genesis of white clover ~16 Kya.
(C) White clover progenitor speciation (gray block) and hybridization giving rise to white clover (blue block) aligned with global temperature variation relative to present-day average temperature based on ice-core data (Jouzel et al., 2007). The blocks indicate the extent of possible divergence times using mutation rates in the range $1.1–1.8 \times 10^{-8}$ (Supplemental Table 9).

Supplemental Table 12). This was followed by a rapid recent expansion that lagged the warming period after the last Glacial Maximum (Figures 3C and 3D; Mangerud et al., 2004). The fourth individual (27) had a larger number of polymorphic sites than the other three, which could explain the difference in estimated population history (**S**upplemental Table 8). In addition to the mutation rate, the recombination rate also influences PSMC analysis, but we found the results stable over a wide range of recombination settings (Supplemental Figure 9). The 20-fold increase in the white clover EPS relative to the progenitor populations appears high and could be inflated as the PSMC model has limited power to accurately determine EPS in the recent past (<20 Kya; Li and Durbin, 2011). The PSMC analysis did not identify a severe population bottleneck coinciding with the hybridization event as would be expected if white clover arose from two or few parent plants, suggesting that diversity may have been carried over to white clover from its progenitors through multiple allopolyploidization events. These results were corroborated using Multiple Sequentially Markovian Coalescent (MSMC) analysis (Schiffels and Durbin, 2014), an extended version of PSMC where several individuals can be surveyed simultaneously. Incorporating more individuals (or haplotypes) improves EPS estimation in recent time periods (2 Kya to 30 Kya; Schiffels and Durbin, 2014), complementing the PSMC model. MSMC analysis identified a weak bottleneck, similar to the PSMC model, which became nonexistent—leaving only a recent expansion in population size as resolution improved with incremental addition of haplotypes (Figure 3E). As with the PSMC, the MSMC results were consistent over a wide range of recombination and mutation settings (Supplemental Figure 10).

These models, however, cannot reliably differentiate between a very severe bottleneck of short duration and a less pronounced reduction of EPS over a longer period (Li and Durbin, 2011; Schiffels and Durbin, 2014). To distinguish between these possibilities, we characterized diversity in extant white clover populations using GBS-derived SNPs identified in 200 individuals from 20 different clover populations (Supplemental Data Set 1). We found a high level of diversity, with polymorphism detected at 7.0% of all reliably sequenced sites with an average heterozygosity of 0.21, yielding an overall genome-wide nucleotide diversity of 0.015. This is similar to other outbreeding species such as teosinte (*Zea mays*; 0.012; Chen et al., 2017), *Arabidopsis lyrata*, and *Arabidopsis halleri* (0.024 and 0.020; Novikova et al., 2016), and higher than selfing species such as Arabidopsis (0.0060; Novikova et al., 2016) and *M. truncatula* (0.0043; Branca et al., 2011). This data set allowed us to generate a detailed allele or site frequency spectrum (SFS) of this white clover population sample against which we could compare simulated SFS produced under different demographic scenarios.

The most extreme bottleneck would result from a single hybridization and chromosome doubling event (A×B=AB→AABB), producing a single homozygous white clover founder individual with no polymorphic sites, from which all diversity would accumulate through mutations. This scenario exhibited a very strong bias against common alleles coupled with a low rate of polymorphism, and was inconsistent with the observed data (Figure 3F). A second possibility is that two unreduced gametes, one from each of the heterozygous diploid progenitors, fused to produce a single white clover hybrid (AA×BB=AABB), thereby carrying variation over to white clover from the two founder progenitors. This second scenario produced similar SFS to the unreduced gametes model, albeit with a less severe skew compared with the observed data (Figure 3F). These results suggested that a population bottleneck alone could not explain the observed data. Instead, we modeled combinations of various types of population bottlenecks with recent expansion of the population size up to 20-fold the size of the pre-allopolyploidization EPS (Supplemental Figure 11). Of all scenarios tested, a recent threefold expansion of the EPS in the absence of a population bottleneck came closest to matching the observed data (Figure 3F; Supplemental Figure 11; Supplemental Table 12). Simulations also matched the observed mutation density when the progenitor population size was increased assuming the PSMC analysis had underestimated the progenitor EPS, which at ~48k was lower than the usual 100k to 1,000k range for nuclear genomes of multicellular species (Figure 3F; Supplemental Table 12; Lynch, 2010).

White clover is thus unlikely to have experienced a severe genetic bottleneck at the time of allopolyploidization during the most recent glaciation ~15 Kya to 28 Kya, and must have arisen from multiple independently generated hybrids, which resulted in carryover of progenitor diversity into the new species. In addition, our PSMC and MSMC analyses and mutation accumulation simulation both indicated a rapid recent population expansion, which is consistent with white clover quickly filling available niches as the climate warmed to generate a larger EPS and occupy a broader habitat range than its progenitors had before allopolyploidization.

Figure 3. (continued).

(D) PSMC curve based on whole-genome resequencing data from each of four white clover individuals. Each curve indicates inferred population size history through time. The analysis depicted here was performed using a mutation rate of 1.8×10^{-8}. See Supplemental Table 11 for results with lower mutation rates.
(E) MSMC figure based on the same data as the PSMC figure. The number of haplotypes corresponds to the number of individuals included. Eight haplotypes comprise all four. Six haplotypes comprise individuals 81, 122, and 183—all of the individuals with the most pronounced bottlenecks. Four of the haplotypes have three combinations of two individuals: [1] includes 81 and 122; [2] includes 122 and 183; [3] includes 81 and 183. The results were scaled to a mutation rate of 1.8×10^{-8}. See Supplemental Figure 10 for MSMC using different mutation rates.
(F) SFS of simulations across 20k generations under the different demographic scenarios indicated on the right (Simulated Effective Population Size (EPS). The observed SFS is scaled to match total polymorphism count for the simulations. Dashed line represents the expected SFS under a constant EPS. Green numbers indicate simulated SNP densities (%) and goodness of fit between observed and simulated data. The goodness of fit was calculated by first dividing the simulated and observed values for each bin with the maximum value of the two and then using the formula: Σ(observed-simulated)^2/simulated. Lower values indicate better fit. The observed SNP density was 7.0%.

Progenitor Genome Integrity Has Been Retained in White Clover

The extant progenitors of white clover occupy restricted ranges in highly specialized niches (Figure 1). If the progenitor genomes have remained intact, their coexistence within white clover after allopolyploidization has enabled them to undergo global allopolyploidy-facilitated niche-expansion. To investigate subgenome integrity, we made further comparisons between the progenitor genomes and their corresponding white clover subgenome derivatives. At a whole-genome level, the sequence-derived estimated genome size of white clover approximated the combined estimates of the contributing progenitors, indicating that major chromosomal fragments had not been lost in white clover compared with the extant progenitors (Table 1; Supplemental Figure 1).

Interhomeologous recombination could lead to rapid subgenome alterations, but white clover is considered a functional diploid (Williams et al., 1998), and cytological evidence suggests that homeologous recombination events are very rare in white clover. To determine if the genomic data supported these observations of subgenome independence, we assessed the incidence of white clover Illumina TSLRs containing sequence from both progenitors. These long reads (total = 668,571; average length = 4,113 bp; Supplemental Table 1) were unlikely to contain chimeric artifacts as they were assembled from single DNA fragments. We identified only five high-confidence recombination events in contigs derived from 38 TSLR reads (0.006% of the TSLR pool), where TSLRs contained well-supported blocks arising from both progenitors (Supplemental Figure 12**)**. The very few signatures of homeologous recombination in white clover agree with the lack of observations of quadrivalent meiotic configurations in cytological studies (H. Ansari, AgResearch, personal communication). In conclusion, there appears to have been very limited interhomeologous recombination in white clover.

Gene loss after polyploidization is common (Grover et al., 2012), and could also compromise the integrity of the white clover subgenomes. To assess the level of gene loss, we mapped sequencing reads from white clover genomic DNA to the combined reference sequences of the extant progenitor genomes, requiring a unique match for each read. Of the ~39k and ~36k protein-coding genes found in both progenitor genomes (Supplemental Table 6), ~36k and ~35k were covered by uniquely mapping white clover reads in *T. occidentale* and *T. pallescens*, respectively, indicating no more than 5% gene loss.

Gene silencing is also commonly associated with genome hybridization or duplication, and often precedes gene loss. We identified 68,475 protein-coding genes in white clover with >99% based on direct transcriptional evidence (Supplemental Table 5), which approximated the sums of To and Tp protein-coding gene counts (Supplemental Table 6). This indicated that the subgenomes had retained a similar transcriptionally active gene complement to each other and to the corresponding extant progenitors, suggesting a limited level of allopolyploidy-associated gene silencing in white clover.

In summary, the white clover subgenome sizes and transcriptionally active gene counts reflect those of the contributing progenitors, indicating that the subgenomes retained in white clover have maintained independence and integrity since allopolyploidization.

Stable Subgenome Expression Ratios Across Tissues

Transcriptional consequences of allopolyploidization encompass homeolog expression bias and expression level dominance (genomic dominance), and have been well-documented in many polyploids (Grover et al., 2012). To better understand these consequences in white clover through characterization of subgenome transcription patterns, we first identified progenitor-specific SNPs by mapping genomic and Pooled transcriptomic data (Supplemental Table 13) from *T. occidentale* and *T. pallescens* onto *T. occidentale* gene models using HyLiTE (Duchemin et al., 2015). We then assigned subgenome identity to white clover RNA-seq reads based on these SNPs. A total of 720 million white clover RNA-seq reads derived from flowers, leaves, stolons/stems, and roots were processed, and 42% were assigned to a specific subgenome. Most of the remaining reads could not be assigned as they did not overlap subgenome-specific SNPs (Supplemental Table 14; Supplemental Data Set 2). The reads were assigned to a similar number of genes on each subgenome (Tr_{To}: 36,181 genes; Tr_{Tp}: 34,942 genes), indicating near equivalence in active gene count per subgenome under glasshouse conditions (Supplemental Table 14).

Previous studies in other allopolyploids have examined differential expression of homeologs across tissues, but as only few genes were examined or average tissue biases were calculated, it remains unclear how often the direction of bias changes between tissues in allopolyploids (Adams et al., 2003; Liu and Adams, 2007; Leach et al., 2014; Zhang et al., 2015b; Yang et al., 2016). For white clover, we found that 69% of the 19,954 genes, for which expression ratios could be reliably determined, exhibited the same direction of homeolog expression bias across the tissues assessed (Figure 4A). An example of consistent homeolog expression ratio (Tr_{To}/Tr_{Tp}) between tissues is shown (Figure 4B), and such homeolog ratio stabilities persisted even when the correlation between total gene expression levels ($Tr_{To} + Tr_{Tp}$) was low (Figure 4C).

Expression Ratio Outliers Were Enriched for Flavonoid Biosynthesis Genes

The high frequency of white clover genes showing stable cross-tissue homeolog expression ratios suggested that this may represent the default state after allopolyploidization. Genes deviating strongly from this pattern, could, therefore, represent instances of adaptive gene expression control in white clover driven by selection. To investigate this possibility, we selected genes with unusually large $\log(Tr_{To}/Tr_{Tp})$ expression ratio variance across tissues, identifying 1,263 transcripts as expression ratio outliers (Supplemental Data Set 3).

Based on gene annotations, the outliers appeared rich in enzyme-encoding genes, leading us to focus on investigating biosynthetic pathways. We assigned enzyme categories (ECs) using the MetaCyc database (Caspi et al., 2016) and found

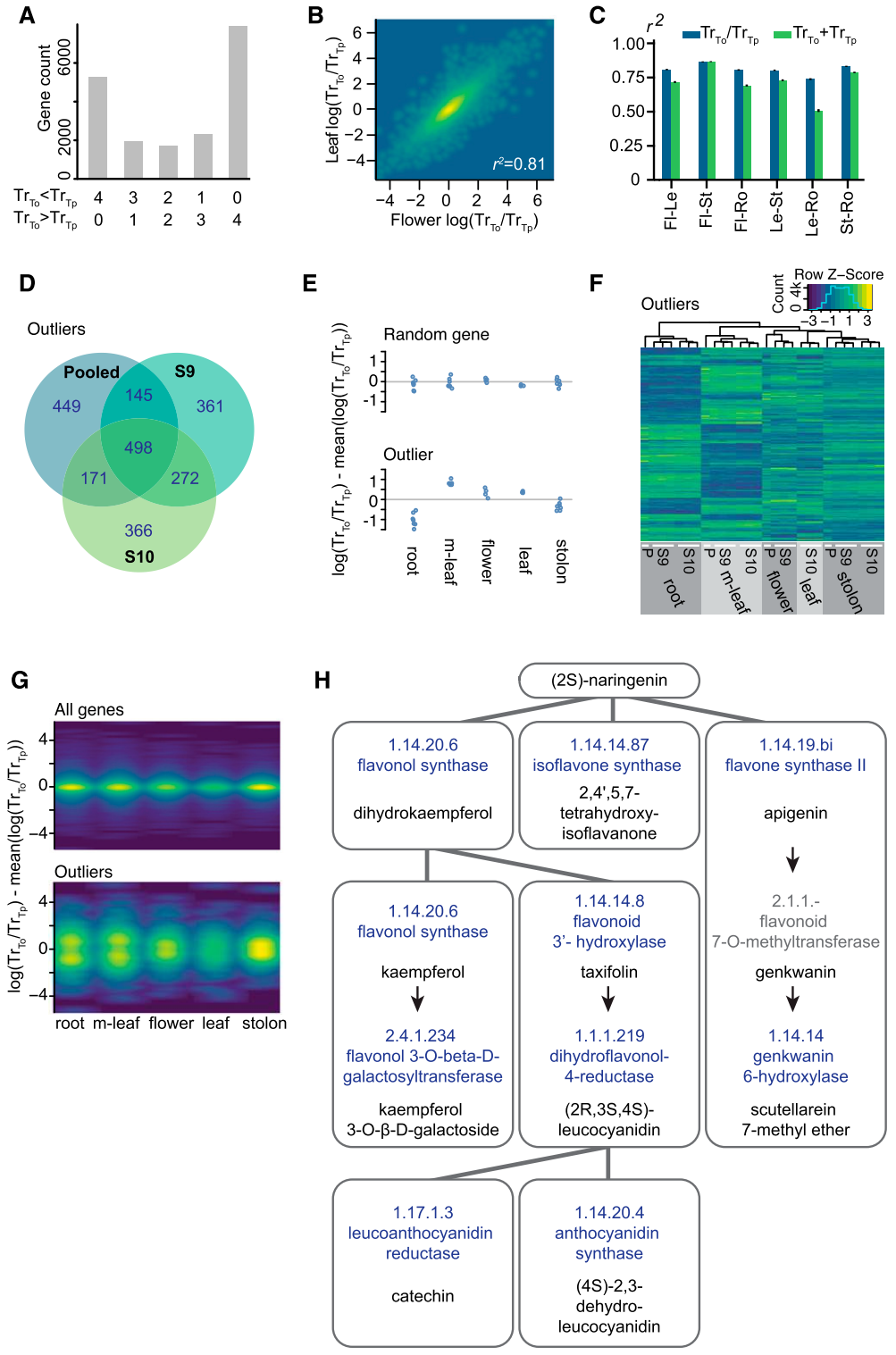

Figure 4. Homeolog-Specific Expression Analysis.

(A) Genes classified based on the number of tissues in which the majority of reads were derived from the Tr_{To} or the Tr_{Tp} homeolog. The numbers underneath the graph indicate the number of tissues for which the statement on the left is true. Most of the genes (69%) show the same direction of bias across all four tissues.

(B) Log(Tr_{To}/Tr_{Tp}) expression ratios in flowers versus leaves.

a striking enrichment for components of the phenylpropanoid derivatives biosynthesis pathway ($P = 4.6 \times 10^{-24}$ after Benjamini-Hochberg correction) and its subset, the flavonoid biosynthesis pathway ($P = 2.2 \times 10^{-10}$ after Benjamini-Hochberg correction), as well as an enrichment for biosynthesis of secondary metabolites ($P = 1.9 \times 10^{-15}$) and plant cell structural components ($P = 1.9 \times 10^{-10}$; Supplemental Table 15). The initial experiment had been performed using biological replicates pooled before sequencing. To examine if the expression ratio outliers could be reproducibly detected, we performed two additional experiments with three replicates of four tissues. These experiments were conducted using material from clones of the white clover S_9 inbred individual used for sequencing and its selfed progeny (S_{10}), respectively, and we refer to these three separate expression outlier detection experiments as "Pooled," "S9," and "S10" (Supplemental Tables 13 and 14). For the outliers in the S9 and S10 experiments, we found enrichment of the same pathways (Supplemental Table 15). As indicated by the pathway enrichment analysis, the 2,262 genes identified as expression ratio outliers showed a pronounced overlap among the three experiments (Figure 4D). Repeating the pathway enrichment analysis with genes identified as expression ratio outliers in at least two out of the three experiments again highlighted the same pathways (Supplemental Table 15). To increase stringency and take advantage of the replicate information, we performed analysis of variance (ANOVA) using all samples from the three experiments. In addition to a high mean expression variance across tissues we required that the means be significantly different in the ANOVA after Bonferroni correction. The ANOVA-filtered results again pointed to the same enriched pathways (Supplemental Table 15), and the identified genes showed consistent expression ratios within tissues across experiments and replicates, with particularly pronounced contrasts between expression ratios in roots and mature leaves (Figures 4E to 4G). The clustered heat map of the outlier expression ratios (Figure 4F) highlighted not only the contrast between root and leaves, but also a temporal and transgenerational similarity within tissues among the Pooled, S9, and S10 samples. The outlier genes included putative transcription factors, receptor-like kinases, and other enzymes, but the genes associated with the above-mentioned pathways made up the only clearly overrepresented sets (Supplemental Data Set 3).

To assess whether the pathway enrichment could occur by chance, we generated random gene sets meeting the same subgenome read assignment cutoff as the expression ratio outliers. We found that the white clover genome was not generally enriched for genes involved in the pathways associated with expression outliers, as only a single pathway (PWY-7214 baicalein degradation, $P < 1 \times 10^{-3}$) was enriched in one out of 10 random gene sets (Supplemental Table 15). Many of the identified flavonoid pathway enzymes appeared to cluster relatively closely in the biosynthetic pathway downstream of the metabolite naringenin, which is a key branch point for production of isoflavonoids, flavonols, anthocyanins, and condensed tannins (Figure 4H; Franzmayr et al., 2012).

Of 909 outlier genes detected in the ANOVA, 904 contained SNPs between the differentially expressed Tr_{To} and Tr_{Tp} homeologs that resulted in amino acid substitutions. The outliers did not, however, show above-average divergence at the protein level (Supplemental Figure 13).

Expression Ratio Outliers Diverged Mainly Post-Allopolyploidization

The deviations from stable subgenome expression ratios across tissues could be allopolyploidy-associated or result from parental inheritance of expression patterns that had already diverged between the progenitors. To distinguish between these possibilities, we compared the subgenome-specific white clover expression data to progenitor expression data. For the Pooled and S10 experiments, we also conducted RNA-seq for the progenitors (Supplemental Table 13), and analyzed data from both these experiments. After normalization of the expression data using Trimmed Mean of M-values (TMM; Robinson et al., 2010), Principal Component Analysis clustered the four tissues and showed tight grouping of the white clover subgenomes for all tissues, whereas the progenitors were more dispersed (Supplemental Figures 14A and 14E). If the subgenome expression ratio deviations were derived from parental inheritance, the outliers should also show extreme divergence in the interprogenitor comparison. We first tested this hypothesis by comparing Pearson correlation coefficients for logged expression values across the four tissues in pairwise genome comparisons. As expected, we found the outlier genes greatly skewed ($P < 10^{-27}$, Kolmogorov-Smirnov test) toward lower correlation coefficients for the inter-subgenome

Figure 4. (continued).

(C) Spearman correlation coefficients for cross-tissue comparisons of the total expression level for both homeologs ($Tr_{To}+Tr_{Tp}$) and for the homeolog expression ratios (Tr_{To}/Tr_{Tp}). Fl, flower; Le, leaf; St, stolon/shoot; Ro, root. Error bars = 95% confidence intervals calculated using the formula: tanh(arctanh [r^2] $\pm$ 1.96/sqrt[$n-3$]), where r^2 is the Pearson correlation estimate and $n = 19{,}954$ is the number of observations.

(D) Venn diagram showing the overlaps between the expression outlier genes detected in the "Pooled," "S9," and "S10" experiments.

(E) Log(Tr_{To}/Tr_{Tp}) difference from the mean values from all three experiments (Pooled, S9, and S10) plotted for a randomly selected gene and an expression ratio outlier. The gray horizontal line indicates the mean. m-leaf, mature leaf; leaf, emerging young leaf.

(F) Clustered heatmap showing row-normalized log(Tr_{To}/Tr_{Tp}) values for all 909 expression outliers detected using the ANOVA method for all three experiments (P [pooled], S9 and S10) across four tissues: m-leaf, mature leaf; leaf, emerging young leaf.

(G) Smoothed scatterplot showing the log(Tr_{To}/Tr_{Tp}) difference from the mean for all 19,954 genes and the 909 ANOVA outliers, respectively. m-leaf, mature leaf; leaf, emerging young leaf.

(H) Expression ratio outliers in flavonoid metabolism. Blue text highlights enzymes found as expression ratio outliers. Numbers above enzyme names are EC identifiers.

Tr_{To}-versus-Tr_{Tp} comparison. We did not, however, observe a similar skew for the interparental comparison *T. occidentale* versus *T. pallescens*, arguing against parental divergence and inheritance being causal for the white clover expression ratio outliers (Supplemental Figures 14B and 14F; Supplemental Data Set 4).

The correlation coefficient analysis captures the directions of change in expression levels between tissues, but does not consider the magnitude of the expression differences between the compared genomes. To examine these, we analyzed deviance defined as the sum of the squared differences between logged expression values in the pairwise genome comparisons. Again, we found the largest difference between the outliers and all genes in the inter-subgenome comparison, with the outliers skewed toward higher deviance. We also observed significant skews for all other comparisons, indicating that the outliers showed above average expression deviance in the progenitors and between progenitors and subgenomes ($P < 10^{-76}$, Kolmogorov-Smirnov test; Supplemental Figures 14C and 14G; Supplemental Data Set 4).

The expression outliers were selected based on variance in $\log(Tr_{To}/Tr_{Tp})$ ratios, making them clearly distinct from the overall gene average (Supplemental Figures 14D and 14H). However, we also detected a less pronounced, but statistically significant ($P < 10^{-93}$, Kolmogorov-Smirnov test), skew toward higher variance for the interprogenitor comparison (Supplemental Figure 14D; Supplemental Data Set 4). The correlation, deviance, and variance analyses produced very similar results for the Pooled and S10 experiments (Supplemental Figure 14).

Overall, the *T. occidentale*-versus-*T. pallescens* deviations appeared much less pronounced than the Tr_{To}-versus-Tr_{Tp} deviations for the outlier gene set, suggesting a limited contribution of parental inheritance to the outlier expression patterns. To quantify the effect, we classified genes as showing parental expression inheritance when they displayed higher correlation coefficients and lower deviance for the subgenome-progenitor than for the interprogenitor comparison(s) and showed subgenome-progenitor correlations coefficients >0.8. Only 104 and 239 genes met these criteria in the Pooled and S10 experiments, respectively, 22 and 43 of which were expression ratio outliers (Supplemental Data Set 4). A total of 19,954 genes and 909 outliers were included in the analysis, making the outliers significantly enriched in genes showing parental expression inheritance ($P < 10^{-4}$, Chi-squared test with Yates correction). However, parental expression inheritance only explained a few percent of the outliers detected, suggesting that outlier expression divergence occurred mainly post-allopolyploidization.

DISCUSSION

White clover is a successful allotetraploid naturalized globally in moist temperate grasslands and is an integral component of temperate pastoral agriculture. By contrast, its diploid extant European progenitor relatives remain in extreme specialized habitats with *T. occidentale* restricted to high salinity coastal niches and *T. pallescens* to alpine screes. Based on whole-genome data for both progenitors and white clover, our work has provided a clear example of allopolyploidization-enabled niche-expansion, which has facilitated global radiation of the previously confined specialist progenitor genomes. This occurred without compromising progenitor genome integrity in the past ~20,000 years since the allopolyploidization event, which took place in the glaciation that culminated in the last European Glacial Maximum. This was a period when alpine and coastal species were likely in proximity in glacial refugia. Such extreme conditions are conducive to allopolyploid formation, as the Arctic is one of the most polyploid-rich regions with post-deglaciation colonization dominated by polyploids (Brochmann et al., 2004). Furthermore, temperature extremes can increase production of unreduced male gametes (De Storme and Geelen, 2014), a pathway to repeated allopolyploid formation (Mason and Pires, 2015). This process also facilitates carryover of diversity from the progenitors to the allopolyploid, as we observed for white clover.

Genome rearrangements and loss of duplicate genes are common features of polyploids as they progress toward a diploid-like genome via fractionation. This process can occur at markedly different speeds depending on the polyploid system, and is preceded by changes in homeolog expression (Soltis et al., 2015, 2016). In white clover, this has occurred to a very limited degree, as the genomes and functional gene complement of its progenitors have been retained with little evidence of large-scale inter-homeologous recombination or unbalanced homeolog expression bias. As more polyploids are sequenced, it is clear there is a wide range of genomic responses to polyploidization events. Some very recent polyploids, and polyploids with similar genesis times to white clover, exhibit gene retention with little genome reduction and no evidence of biased fractionation or major genetic changes, such as common cordgrass (*Spartina anglica*, 130ya; Chelaifa et al., 2010; Ferreira de Carvalho et al., 2013), peanut (9.4 Kya; Bertioli et al., 2016), hexaploid wheat (*Triticum aestivum*, 9kya; IWGSC, 2014), Arabian coffee (*Coffea arabica*, 10 Kya to 50 Kya; Cenci et al., 2012; Combes et al., 2015), and shepherd's purse (*Capsella bursa-pastoris*, 100 Kya to 300 Kya; Douglas et al., 2015). Of these species, *C. arabica* (Bardil et al., 2011) and *S. anglica* (Chelaifa et al., 2010) exhibit expression level or genomic dominance implicated in the process of diploidization. The subgenomes of the allotetraploid cotton (*Gossypium hirsutum*, 1 Mya to 2 Mya) have little gene loss but show evidence of asymmetric evolution (Zhang et al., 2015b), as well as homeolog expression bias in certain tissues or conditions (Yoo et al., 2013; Zhang et al., 2015b). By contrast, some very recent species such as *Tragapogon* spp (80ya; Chester et al., 2012), synthetic wheat (Feldman et al., 1997), synthetic allotetraploid Arabidopsis (Wang et al., 2004), as well as ancient polyploids such as field mustard (*Brassica rapa*, 13 Mya to 17 Mya; Wang and Brassica rapa Genome Sequencing Project Consortium et al., 2011; Cheng et al., 2012) and maize (*Zea mays*, 5 Mya to 12 Mya; Schnable et al., 2011) have undergone significant genome and homeolog expression changes on the path to diploidization. It may be that factors including progenitor relatedness and genome size similarity influence the fate of the postpolyploid genomes.

With limited posthybridization alterations to the subgenomes, white clover appears to have resulted from a fortuitous combination of compatible progenitors, which allowed multiple allopolyploidization events to occur in the contact zones of the progenitors, quickly generating very diverse white clover

populations. Furthermore, the repeated allopolyploidization events suggest high compatibility between progenitor genomes. This may explain why we found a stable background of subgenome expression ratios across tissues on a per-homeologous gene pair basis, which in turn allowed us to identify genes with strongly aberrant expression patterns. Differential homeolog expression across tissues, during development and in response to stresses has been described in a number of species, including cotton, wheat, and *Tragopogon* (Adams et al., 2003; Bottley et al., 2006; Liu and Adams, 2007; Hovav et al., 2008; Buggs et al., 2010). It has also been investigated at the whole-transcriptome level, where differential regulation was confirmed in coffee, synthetic *Brassica* allotetraploids, cotton, wheat, and brown mustard (*Brassica juncea*; Combes et al., 2013; Liu et al., 2015; Zhang et al., 2015a;, 2015b; Yang et al., 2016). Our analysis strategy distinguishes itself from previous studies by first establishing a testable null hypothesis, in this case stable homeolog expression ratios across tissues, which allows quantitative identification of expression ratio outliers. This phenomenon appears to be a temporal and transgenerational feature as evidenced by the observed overlap in these outlier genes among the different experiments that comprised either the same individuals or progeny sampled at different times. Coupled with pathway enrichment analysis of the expression ratio outliers, we identified the phenylpropanoid and flavonoid biosynthesis pathways as putative targets of selection in white clover. This overrepresentation of the biosynthetic pathways among genes with deviating expression patterns argues against the outliers resulting from rare stochastic events, and instead suggests that selection has driven tissue-specific differential use of homeolog gene copies. The selective pressure in question could be internal, driven by incompatibilities between diverged homeologous proteins. However, we did not observe increased diversity at the protein level for the outliers, arguing against this possibility. Taken together with our results indicating that expression divergence occurred mainly post-allopolyploidization, the expression ratio outlier patterns could represent examples of adaptive, allopolyploidy-associated transcriptional changes. Such changes may be influenced by features such as the development of specialized tissue- or homeolog-specific transcription factors, or localized epigenetic characteristics including differential tissue-specific methylation among homeologs.

As more polyploids are explored in greater depth, other examples of allopolyploids exhibiting stable homeolog expression ratios are appearing. *Arabidopsis kamchatica*, an allotetraploid, arose at a similar time (~20 Kya; Akama et al., 2014) to white clover. Similarly, it has greater environmental plasticity than its extant progenitors, although with a significant overlap in distribution (Hoffmann, 2005). Interestingly, it exhibits stable homeologous expression ratios for the majority of homeologous pairs, with a very small proportion (~1%) altering expression ratios in response to abiotic stress (Akama et al., 2014; Paape et al., 2016). In contrast with white clover, however, it appears that *A. kamchatica* has combined the transcriptional patterns of both parental species (Paape et al., 2016). Another recent polyploid also exhibits homeolog expression features comparable to white clover. In allohexaploid wheat, ~72% homeologous gene triads showed stable expression ratios across 15 tissues (Ramírez-González and International Wheat Genome Sequencing Consortium et al., 2018), remarkably similar to the ~70% we observed in white clover. Furthermore, of the homeolog triads with variable expression ratios across tissues, the majority did not reflect expression ratios of the wheat progenitor species, indicating possible allopolyploidy-associated transcriptional change (Ramírez-González and International Wheat Genome Sequencing Consortium et al., 2018). These genes were enriched for defense and external stimuli responses involved in fitness (Ramírez-González and International Wheat Genome Sequencing Consortium et al., 2018), with a possible role in adaptation for this domesticated species.

In white clover, the overrepresentation of the phenylpropanoid, flavonoid, and secondary metabolite pathways suggests that secondary metabolites may be under particularly strong selective pressure. Indeed, flavonoids are widely distributed in plants and 9,000 different chemical structures have been characterized (Williams and Grayer, 2004). The pervasive nature and extreme diversity of flavonoids, along with variation at the population level (Cao et al., 2017), argues strongly for their role in plant adaptation, where they have been shown to facilitate tolerance to a wide range of biotic and abiotic stresses including protection against reactive oxygen species, pathogenic microbes, and insect predation, as well as in signaling and development (Mierziak et al., 2014; Mouradov and Spangenberg, 2014; Davies et al., 2018). It may be that white clover exploits homeologous isozymes, an option available only to allopolyploids, to fine-tune complex mixes of flavonoids and other secondary metabolites across tissues, for instance to balance responses to symbionts and pathogens in roots or to optimize herbivory deterrence and antifungal defense in leaves through differential regulation of naringenin and its derivatives (Figure 4H; Mierziak et al., 2014; Davies et al., 2018).

It will require further investigation to determine if allopolyploidy-associated transcriptional reprogramming of flavonoid biosynthesis genes contributed to the evolutionary success of white clover. The white clover species complex described here, with well-defined progenitors, limited interhomeolog recombination, limited subgenome gene loss, and clearly defined natural ranges, will facilitate such studies, and provides a powerful platform for understanding how allopolyploidy can underpin adaptation and range expansion. As a first step, this study has highlighted white clover as an example of allopolyploidy-facilitated niche-expansion, where two diverged genomes were reunited by a changing climate, and colonized the globe.

METHODS

Plant Material

Trifolium *Species*

Individuals of the three species used for genome assembly are detailed as follows. White clover (*Trifolium repens*): An inbred allopolyploid S_9 individual was selected for sequencing from the ninth sequential selfed generation derived by single seed descent from a self-fertile white clover cv 'Crau' derivative (Cousins and Woodfield, 2006). Western clover (*Trifolium occidentale*): An individual selected for sequencing was derived by controlled selfing from an individual from a wild population (OCD16; Margot

Forde Germplasm Centre). Pale clover (*Trifolium pallescens*): An individual was taken from a wild population (AZ1895; Margot Forde Germplasm Centre). For LD mapping, a population (PGen3×PGen9) of 93 F_1 full-sib progeny was developed from a hand-pollinated reciprocal pair cross of heterozygous individuals, PGen3 and PGen9. All plants were grown under glasshouse conditions before DNA extraction.

DNA Extraction and Sequencing

For the three *Trifolium* species in this study, nuclear DNA was extracted from young leaves as described in Anderson et al. (2018) and sequenced by Macrogen. Using a range of platforms (Roche 454-GLX; Illumina HiSeq 2000), libraries (shotgun, paired-end; mate-paired) and insert sizes (180 bp to 8,000 bp), sequence data were generated to an approximate coverage of 492×, 114×, and 230× for *T. occidentale*, *T. pallescens*, and *T. repens*, respectively (Supplemental Table 1). Furthermore, high-molecular–weight DNA was extracted (Anderson et al., 2018) to create six white clover libraries of Illumina TSLRs generated by the Illumina FastTrack Sequencing Services Long Reads pipeline to provide a 3× coverage (Supplemental Table 1). Original sequence data are deposited in the National Center for Biotechnology Information (NCBI) under Bioprojects PRJNA523044, PRJNA521254, and PRJNA523043 for white clover, *T. occidentale*, and *T. pallescens*, respectively, as detailed in "Accession Numbers."

RNA Extraction and Sequencing

Transcriptome analyses of the S_9 white clover individual and extant progenitors *T. occidentale* and *T. pallescens* were derived from Illumina RNA-seq data generated from four different tissue samples: mature fully expanded leaf (third node from stolon tip); tap root (*T. pallescens*) nodal root, or nodal root (white cloverand *T. occidentale*); young opened flowers from the middle whorls of the inflorescence; internodal shoot (*T. pallescens*); or stolon (white clover and *T. occidentale*). The tissues were sampled in 2013 as three biological replicates harvested at a single time point from plants grown in glasshouse conditions, and frozen immediately in liquid nitrogen. Total RNA was isolated from nonfloral tissue using an Isolate II RNA Plant Kit (Bioline). Due to the high phenolic content of white clover flowers, floral RNA was extracted as described in Jamalnasir et al. (2013) with the modifications including replacement of chloroform/isoamyl alcohol (24:1, v/v) with pure chloroform, followed by an additional purification step. The final precipitated RNA was resuspended in the Isolate II RNA Plant Kit (Bioline) extraction buffer before binding, washing, and eluting from an Isolate II RNA Plant Kit column (Bioline). Before sequencing, RNA from the biological replicates for each tissue were pooled, transported on dry ice, and sequenced as 500-bp paired-end libraries at Macrogen using the Illumina HiSeq 2000 2 × 100-bp paired-end chemistry (Supplemental Table 13). Reads were filtered using the software Kraken (Wood and Salzberg, 2014) to eliminate viral RNA, and filtered for low quality using the software SolexaQA (Cox et al., 2010). These data were used for gene annotation and homeolog expression analysis and referred to as the "Pooled" data.

To further explore stable homeolog expression across tissues of the white cloverS_9 individual, additional RNA-seq data were generated at Novogene from each of three cloned plants providing unpooled biological replicates per tissue sample (mature fully expanded leaf; root [nodal/tap]; stolon/shoot); and young open flowers from the middle whorls of the inflorescence). Tissues were harvested simultaneously in 2017 from each of three cloned plants grown in glasshouse conditions and snap-frozen in liquid nitrogen and extracted as described in the previous paragraph. RNA was transported frozen on dry ice to Novogene and 250 bp to 300 bp insert libraries were sequenced using the Illumina NovoSeq 6000 2 × 150-bp paired-end chemistry (Supplemental Table 13). Reads were filtered for quality and to remove viral RNA as described in the previous paragraph. This RNA-seq data set was used for replicated homeolog expression analysis and referred to as "S9."

To further investigate conservation of homeolog expression among extant progenitors and white clover plants of the same age, RNA-seq data were generated from tissue (emerging/young leaf; mature fully expanded leaf; stolon/shoot; root [nodal/tap]) from each of three cloned plants (biological replicates) of an individual selfed progeny each of the sequenced *T. occidentale* (selfed OCD16), *T. pallescens* (selfed AZ1895), and *T. repens* (S_{10}). These plants, colocated in the same glasshouse, were harvested simultaneously in 2018 at the same time of day (early afternoon) as previous tissue collections to minimize effects of diurnal variation on gene expression. RNA was extracted, sequenced, and filtered as described in the first paragraph of this subsection, and the RNA-seq data set referred to as "S10." Original sequence data are deposited in NCBI under Bioprojects PRJNA523044, PRJNA521254, and PRJNA523043 for white clover, *T. occidentale*, and *T. pallescens*, respectively, as detailed in the "Accession Numbers" subsection.

K-mer Analysis for Estimating Genome Size

Total and frequency of *k*-mers ($n = 17$) was counted in unassembled Illumina 180-bp insert paired-end sequence data for white clover and its progenitors using the software "Jellyfish" v2.2.0 (Marçais and Kingsford, 2011) default parameters. The data were plotted to determine maximum read depth for each specific 17-mer, and genome size was estimated as (total 17-mer number)/(peak depth).

Genome Assembly

Raw sequence data for all three *Trifolium* species underwent quality control with the software tool FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), was then trimmed using the DynamicTrim program from the SolexaQA package (Cox et al., 2010), and, in the case of mate-pair libraries, the high level of adapter contamination was reduced using the software tool Trimmomatic (Bolger et al., 2014).

Assembly of *T. pallescens* was accomplished exclusively with the use of ALLPATHS-LG (Gnerre et al., 2011) with default parameters. A low level of bacterial contamination was found in the DNA, and these were filtered with the use of a simple in-house script using the software Bowtie 2 (Langmead and Salzberg, 2012) for raw read mapping. The data set underwent a further error correction step by remapping pair-end data and searching for non-repeats with uneven read distribution.

Assembly of *T. occidentale* was separated in two stages. In the first stage, an assembly was produced using the program ALLPATHS-LG (Gnerre et al., 2011) with default parameters. All mate-pair libraries were used, but only the single 180-bp pair-end library was incorporated into the ALLPATHS-LG assembly. The other pair-end libraries were assembled separately using the algorithm package Velvet (Zerbino and Birney, 2008). The raw reads were first subjected to digital normalization using the software package khmer (https://github.com/dib-lab/khmer) using a *k*-mer size of 31, a maximum coverage of 50, and a minimum coverage of 3. The filtered data was assembled with Velvet using the following parameters: *k*-mer size of 55; expected coverage of 25; coverage cutoff of 7.5; maximum coverage of 45; maximum divergence of 0.05; minimum scaffolding pair count of 7.0. The resulting contigs were compared with the ALLPATHS-LG assembly, and used to extend and resolve gaps where long, high-quality overlaps were found. A final round of gap-filling using all pair-end data was done using an in-house script (https://github.com/Lanilen/SemHelpers). This resulted in a longer assembly with fewer ambiguities when compared with that of the similar-sized genome of *T. pallescens*.

Assembly of *T. repens* using ALLPATHS-LG yielded unsatisfactory results, with a reduced assembly of 600 Mb displaying a high degree of homeolog conflation. Neither ALLPATHS-LG nor MaSuRCA (Maryland

Super-Read Celera Assembler; Zimin et al., 2013) assemblers seemed capable of utilizing the Illumina TSLRs (both running out of memory while executing in a 64-core, 1-Tb RAM server), thus the final assembly was done using Velvet (Zerbino and Birney, 2008).

All pair-end libraries were subjected to khmer's (https://khmer.readthedocs.io/en/v2.1.1/) digital normalization to a maximum coverage of 50×, minimum coverage of 3×, and using a *k*-mer size of 31, independent of each other. The resulting filtered libraries were assembled together with all Illumina TSLRs with Velvet using the following parameters: *k*-mer size of 47; automatic expected coverage; minimum coverage of 10; fixed insert length of 350 for longer pair-end libraries and 0 for 180-bp pair-end library; maximum branch length of 500; maximum divergence of 0.05; minimum contig length of 1,000 (so as to match the filter used by ALLPATHS-LG and make assembly comparisons more meaningful); and turning the "conserveLong" flag on.

The resulting assembly was scaffolded using the mate-pair libraries with the use of the software tool "SSPACE-Basic 2.0" (Boetzer et al., 2011). Illumina TSLRs were split into 1,000-bp fragments, with a generous 750-bp overlap, and added to the mate-pair library collection to take advantage of the gap-filling option of SSPACE.

Once the assembly was completed, the Illumina TSLRs were used one last time for an error-correction step using in-house scripts (https://github.com/Lanilen/SemHelpers), by aligning them against the assembly and looking for small insertion-deletions (indels) in the alignment. When a gap or an insertion was well-supported by TSLRs, the reference genome was corrected with the consensus from the TSLRs.

Pseudomolecule Construction

White clover scaffolds were assigned to subgenomes and ordered into pseudomolecules using pairwise LD (calculated as r^2) between microsatellite (SSR) markers and GBS-derived biallelic SNPs segregating in F_1 progeny ($n = 93$) of a white clover biparental cross (PGen3×PGen9). To enable linkage group (LG) identification and distinguish homeologs in this population, and to align the sequence assembly pseudomolecules with previously published white clover linkage groups, DNA extracted from the parent and progeny plants (Anderson et al., 2018) was screened with 16 SLHS SSR markers (Supplemental Table 2) as described in Griffiths et al. (2013).

A denser marker set for LD mapping was then generated by preparing GBS libraries from each parent and 93 progeny as described in Elshire et al. (2011). Briefly, 100 ng of sample DNA was digested with *Ape*KI, followed by ligation of 99-ng barcoded and reverse adapters, then pooled and PCR-amplified. The pooled libraries were sequenced on two lanes of an Illumina HiSeq 2500, and samples demultiplexed using the tool package GBSX (Herten et al., 2015) with one mismatch allowed in the enzyme and barcode. Each sample was mapped back to the reference genome individually using the software Geneious 8 (Kearse et al., 2012) with the default "Fast" parameters and discarding nonunique hits. SNPs were called from the resulting Binary Alignment/Map (BAM) files using "ref_map" from the software pipeline STACKS (Catchen et al., 2013) with a minimum depth of three reads required to form a locus, with one mismatch allowed per locus. The "Populations" pipeline from STACKS was used to consolidate individuals and generate SNPs across all the samples.

The resulting SNPs were filtered as follows: Individual heterozygous genotype calls where the ratio of reads of one allele relative to the other exceeded 9:1 were reset to "homozygous" for the most numerous allele read. Overall, ~1% of heterozygous genotypes were reset to "homozygous." SNPs were discarded that had a minor allele frequency <10%, or at least 50% missing genotype calls, or >75% heterozygous genotypes. The remaining SNPs were grouped according to parental genotype, provided that the read depth for both parents was 10 or more. Specifically, a SNP was classified as a *PGen3Het* SNP if parent PGen3 was heterozygous and parent PGen9 was homozygous; or a *PGen9Het* SNP if PGen3 was homozygous and PGen9 was heterozygous. These SNPs were discarded if the frequency of the GBS homozygous major genotype was either <0.25, or less than twice the frequency of the GBS homozygous minor genotype, as they may have been misclassified. After filtering, the number of SNPs was 22,663, comprising 10,824 *PGen3Het* and 11,839 *PGen9Het* SNPs.

LD analysis was performed using the *PGen3Het* and *PGen9Het* SNPs based on a pseudo-backcross approach where LD was only considered between SNPs with the same parental genotype configuration and along haplotypes inherited from the heterozygous parent. Estimates of pairwise r^2 values were computed using GUS-LD (Bilton et al., 2018), which accounts for uncertainty in GBS genotypes associated with low read depth, where the major allele frequencies were fixed at 0.5 and $\varepsilon = 0$. These LD estimates can be expressed in terms of recombination fraction, therefore equating the LD analysis to a two-point linkage analysis that accounts for read depth.

LGs representing parental chromosomes comprising either the *PGen3Het* or *PGen9Het* parental SNP data were assembled using the following procedure: For SNPs from GBS sequence tags mapped to the white clover assembly, an initial LG was formed by selecting a set of SNPs from the largest group of SNPs in LD based on a visual inspection of the matrix of r^2 values. These LGs were assigned to the assembly pseudomolecules according to where the SNP GBS tags mapped. The remaining SNPs aligned to a pseudomolecule were added to the corresponding LG if the mean of the largest 30 r^2 values between the unmapped SNP and the LG SNPs was >0.4. Each of the remaining unmapped SNPs were then assigned to an LG based on the greatest mean of the 10 largest r^2 values between the unmapped SNP and LG SNPs, provided the mean was >0.6. To ensure SNP placement in the LG was unequivocal, SNPs remained unassigned if the second largest r^2 mean was to another LG and was >66% of the highest r^2 mean. After assignment, only seven mapped SNPs were removed due to high LD estimates with SNPs in other LGs.

Two additional filtering steps were performed on the SNPs mapped to an LG. Firstly, to ensure only a single SNP was selected to represent a GBS sequence tag, SNPs were mapped to the reference genome and placed into bins if separated by <180 bp. From each bin, only the SNP with the highest mean r^2 value across the LG was retained for further analysis. The second filtering step removed SNPs that mapped to multiple loci in the genome. Briefly, SNPs were mapped to a version of the reference genome in which positions that had sufficient similarity that could result in tags being mapped to multiple locations, were masked. SNPs were only retained that mapped to the same place in both the masked and unmasked genomes. After filtering, parental LGs comprised 3,492 and 3,785 SNPs for parents PGen3 and PGen9, respectively.

The SLHS SSR data described above were also incorporated to align LGs with a previous genetic linkage map (Griffiths et al., 2013). Pairwise r^2 values between the SSRs and the mapped SNPs were estimated where the read depth for the SSR genotype calls were set to "infinite." Each SSR was then assigned to the LG group where there were pairwise r^2 values >0.5. Depending on the parental configuration, some of the SSRs were mapped in both parental LGs.

Ordering the SNPs and SSRs in each LG was achieved using the Sorting Points into Neighborhoods algorithm (Tsafrir et al., 2005) within the "R" package "seriation" (Hahsler et al., 2016). Adjustments to the implementation of the neighborhood algorithm within R were made, which consisted of solving the linear assignment problem using the Hungarian algorithm and running the algorithm using multiple starting orders to ensure the global maximum was obtained. The SD parameter for the weighting matrix was set to 15 for all LGs except two where the parameter was set at 10. Orientation of the ordered LGs was determined by SNP alignment to the reference genome and was confirmed by SSR position relative to the

genetic linkage map (Griffiths et al., 2013). The order of the SSR alleles among the 3,492 and 3,785 SNPs in the PGen3 and PGen9 parental maps, respectively, show conservation of position within the LGs relative to previous linkage maps (Supplemental Figure 2).

The GBS tags harboring the ordered SNPs from the biparental population were mapped to the white clover S_9 sequence data and used to group and order an anchor set of 3,364 of the 22,100 scaffolds. This anchor set of pseudomolecules was assigned to the appropriate Tr_{To} or Tr_{Tp} subgenome by Basic Local Alignment Search Tool (BLAST) alignment with the progenitor assemblies. To incorporate remaining scaffolds not linked to the anchor set by LD, the anchor scaffolds were BLAST-aligned to the model forage legume *Medicago truncatula* genome (Mt4.0; Tang et al., 2014), which has general macrosynteny with some major rearrangements relative to white clover (Griffiths et al., 2013). The nonlinked scaffolds were assigned a pseudomolecule and order relative to the anchor scaffolds based on proximity on the Mt4.0 genome. A small number (4%; Supplemental Table 3) of white clover scaffolds had no strong alignment to Mt4.0, and were assigned to "Chromosome 0." In summary, the white clover scaffolds were ordered based on LD of mapped GBS SNPs, assigned to subgenomes, and then supplemented with non-LD–linked scaffolds by alignment to the *M. truncatula* genome. The progenitor pseudomolecules were assembled based on synteny to white clover.

Gene Annotation

RNA-seq reads were mapped to three respective genomes using the softwares Bowtie v2.1.0 (Langmead and Salzberg, 2012) and TopHat v2.0.9 (Kim et al., 2013) with a maximum intron size of 30 kb. Genome-guided transcripts were then reconstructed using the software Cufflinks v2.1.1 (Trapnell et al., 2012) and those transcripts were used as primary source for final set of gene models (Sanggaard et al., 2014; Supplemental Table 5). Gene annotation was performed using a custom pipeline, which combined gene model evidence from experimental data and ab initio methods. Gene model evidence was obtained using five parallel approaches, as follows: (1) Cufflinks as described in a previous section; (2) evidence from the mapping of the de novo assembled transcripts using the GMAP v20/07/12 algorithm (Wu and Watanabe, 2005); (3) the software AUGUSTUS v2.6.1 (Stanke et al., 2006); (4) the software"GeneMark –E (Lomsadze et al., 2005), and (5) the software GlimmerHMM v3.0.1 (Majoros et al., 2004; Supplemental Table 5). The five layers of evidence were merged to create a nonredundant gene model set using a hierarchical filtering approach as described in Sanggaard et al. (2014). Cufflinks gene models were given highest priority, followed by de novo assembled transcripts (GMAP). Three ab initio predictors had lower priority in the following order: AUGUSTUS, GeneMark, and Glimmer. Glimmer appears to have been superseded by other prediction programs and contributed only very few gene models. AUGUSTUS was trained using protein-coding transcripts (>1,000 nucleotides) from Cufflinks. GeneMark uses ~10 Mbp of genome to fine-tune gene model prediction parameters whereas GlimmerHMM was trained with an Arabidopsis (*Arabidopsis thaliana*) gene set. The final set of gene models were then divided into "Protein coding" and "Unclassified." Gene models were labeled as "Protein coding" if gene model had a coding sequence length l > 300 or had a functional annotation based on the homology against a nonredundant plant database. The "Unclassified" category is the remainder of transcripts that did not fit the "Protein coding" category.

The annotation was further improved using the BRAKER2 pipeline tool (Hoff et al., 2017), which is based on AUGUSTUS. The RNA transcription data were mapped using Spliced Transcripts Alignment to a Reference (Dobin et al., 2013). The references for the white clover and the progenitors were soft-masked using the program RepeatMasker (Smit et al., 2015). The BRAKER2 pipeline produced an abundance of genes, due to the AUGUSTUS predictions. Many of these additional genes occurred in regions where we had no RNA-seq mapping. Using the gene models from the BRAKER2 annotation, we calculated confidence scores (0 to 1) based on the amount of overlap in exons found in the previous annotation. The final curated gene set with confidence scores >0.5 did not reduce Benchmarking Universal Single-Copy Orthologs (BUSCO v3.0.2; Simão et al., 2015) scores after filtering. BUSCO was run against the Embryophyta ODB9 database that was downloaded on 28 August 2017. Any genes not supported by the previous annotation were removed. Functional annotation was done by BLAST alignmentagainst all protein plant sequences from the NCBI database. The 'most likely match' category was determined as those that had not been labeled by the BLAST alignment as uncharacterized, hypothetical, or predicted proteins.

ITS and Chloroplast Analysis

ITS sequences (ITS1 and ITS2, together with the 5.8S rRNA) from *T. pallescens* (GenBank DQ312111) and *T. occidentale* (GenBank AF053168) were used to search the collection of white clover Illumina TSLRs to identify diagnostic SNPs. The *T. occidentale* chloroplast sequence (GenBank KJ788289), along with *T. pallescens* and white clover chloroplast sequences (current work), were aligned with the "mauveAligner" (Darling et al., 2004) and the "progressiveMauve" algorithm plugins from the software "Geneious 7" (Kearse et al., 2012), and the program "LASTZ" (Harris, 2007), using default parameters with inclusion of the "chain" parameter and a hit window of 10,000. A fourth chloroplast genome from *Trifolium hybridum* (GenBank KJ788286.1) was added to serve as an outgroup for comparison. Expanding the chloroplast analysis to encompass more white clover individuals, four resequenced clover individuals (described in the White Clover Resequencing section below), were mapped onto the chloroplast sequences using the software package "BWA" (Li and Durbin, 2011). Variants were called using the mapped reads using the Genome Analysis Tool Kit (DePristo et al., 2011) for all four individuals combined and separately. Reduced SNP variation among chloroplast comparisons indicated greater similarity, and this metric was used to determine the most likely source of the chloroplast sequence within each individual white clover.

Recombination Analysis

The assembled, unscaffolded white clover contigs were aligned to the assembled genomes of the progenitors using BLAST to identify contigs where the top high-scoring segment pairs alternated between *T. occidentale* and *T. pallescens* along the white clover contig. Subsequently, all mapping TSLRs were compared against the reference genomes of *T. occidentale* and *T. pallescens* using LASTZ (Harris, 2007), and the resulting alignments were filtered to find individual TSLRs having high scoring matches against alternate ancestors on each end of the read.

Whole-Genome Divergence Estimation

Whole-chromosome alignments were generated using LASTZ (v1.02.00; Harris, 2007) with the following settings: –hspthresh=12,500–strand=both–format=maf–ambiguous=iupac–chain. The resulting minor allele frequency alignments were then masked at genic positions, to focus on neutral polymorphic sites, and used as input for the IMCoalHMM isolation-model software run with the multichain Monte Carlo mode (Mailund et al., 2012) on sets of 100 alignments for each of which the median split time was recorded.

White Clover Resequencing, PSMC, and MSMC Analysis

Whole-genome resequencing was performed using full Illumina sequencing (150-bp paired-end libraries) to ~49× coverage for four

divergent individuals selected from among 20 cultivars (Supplemental Table 8). Each individual was mapped using BWA (Li and Durbin, 2011) and the number of total reads and mapped reads can be found in Supplemental Table 8. Reads not properly paired were removed. Diploid FASTA files were generated using mpileup, bcftools, and vcftools from SAMtools (http://samtools.sourceforge.net/). PSMC analysis (Li and Durbin, 2011) was conducted using the diploid FASTA files for each individual, and the results of the separate PSMCs were combined and then plotted. Additional PSMC analyses were performed to assess the influence of different recombination rates. With a mutation rate of 1.8×10^{-8}, the following recombination rates were used: (4+25*2+4+6) in which the first parameter spans four intervals, the next 25 parameters span two time intervals, the second to last parameter spans four intervals, and the last parameter spans six intervals (Li and Durbin, 2011); or (4+30*2+4+6+10) equivalent to a higher recombination rate in which the first parameter spans four intervals, the next 30 parameters span two time intervals, the third to last parameter spans four intervals, the second to last spans six intervals, and the last spans 10 intervals (Nadachowska-Brzyska et al., 2016). All scripts and settings can be found in the PSMC folder at https://github.com/MarniTausen/CloverAnalysisPipeline. MSMC analysis software (Schiffels and Durbin, 2014) was performed on the same individuals as PSMC. MSMC, in comparison to PSMC, allows the use of multiple individuals (haplotypes), which increases resolution of population estimation in more recent timeframes (2 Kya to 30 Kya). MSMC was run with eight haplotypes (four individuals), six haplotypes, four haplotypes, and two haplotypes. In the case of two haplotypes, MSMC is similar to PSMC. For each individual, all of the preprocessing was run per chromosome, as described in the MSMC manual (https://github.com/stschiff/msmc/blob/master/guide.md). The final results of the MSMC analysis contained unscaled time and EPS, which were then scaled using a generation time of "1" and the mutation rate of 1.8×10^{-8}. To compare the MSMC analyses with various recombination rate/mutation rate ratios compared with the default ratio (0.25), we assessed MSMC using increased and reduced ratios (0.35 and 0.15, respectively).

GBS Sequencing and Diversity Profiling

A panel of 200 white clover genotypes was sampled from 20 populations (Supplemental Data Set 1) and genotyped using GBS (Elshire et al., 2011). Briefly, 100 ng of DNA isolated using a cetyltrimethylammonium bromide-based method was digested with *Ape*KI, ligated to modified Illumina adaptors, pooled to create libraries (each consisting of up to 88 individually bar-coded DNA samples), then PCR-amplified as described in Elshire et al. (2011). Libraries were assessed for quality on an Agilent DNA 1000 Assay (Agilent Technologies) before sequencing each library on multiple lanes of an Illumina Hi-Seq flow cell to generate single-end sequences of 101 bp. Post-sequencing quality assessment included removal of adaptor contamination using the software scythe (https://github.com/vsbuffalo/scythe) with a prior contamination rate set to 0.10. The tool Sickle (https://github.com/najoshi/sickle) was used to trim reads when the average quality score in a sliding window (of 20 bp) fell below a *Phred* (Ewing and Green, 1998) score of 20. At this point, reads <40 bp were also discarded. The reads were demultiplexed using the tool sabre (https://github.com/najoshi/sabre), and all reads originating from the sample were concatenated. The software package BWA (Li and Durbin, 2011) was used to align the reads against the white clover reference and generate an alignment file in Sequence Alignment/Map (SAM; Li et al., 2009) format. Alignments were sorted by coordinate with Picard tools (http://broadinstitute.github.io/picard), and the Genome Analysis Tool Kit (GATK; DePristo et al., 2011) was used to generate a list of putative indels (RealignerTargetCreator), perform local realignment around these putative indels (IndelRealigner), and identify variants and call genotypes (UnifiedGenotyper). Only biallelic variant sites with a mean mapping quality of 30 were retained. The SNP density was calculated as the number of variants divided by the number of callable sites in bins of 100 kb. All scripts and commands used in the analysis are available on GitHub in the GBS and GBS visualization folders (https://github.com/MarniTausen/CloverAnalysisPipeline).

Simulation of Mutation Accumulation

Simulations were performed with the software "msprime" (Kelleher et al., 2016) using mutation rates based on Arabidopsis (6.5×10^{-9} [Ossowski et al., 2010]) or genome size (1.1×10^{-8} [progenitors], 1.8×10^{-8} (white clover [Lynch, 2010]; Supplemental Figure 8), a sample size of 400 (alleles), and a recombination rate of 1×10^{-7}. Multiple demographic models were run testing different hypotheses. The general structure was a fast bottleneck, modeling the start of the population with a rapid increase to full population size. The "Hybridizing and Doubling" hypothesis was tested by modeling growth of a population originating from a single *T. occidentale* × *T. pallescens* hybrid individual and assuming it attains the same EPS (N_e) as its progenitors (120,000) within 20 generations of exponential growth. The hypotheses of unreduced gametes and mild bottlenecks used the same model, assuming a constant population size and introducing a bottleneck followed by 20 generations with rapid growth. Each of the simulations produced a Variant Call file (VCF), where the SNP density was estimated in the same manner as the GBS data was analyzed. SFS were also produced and used to compare the simulated to the observed data. (The scripts used for the simulations and the precise settings can be found at https://github.com/MarniTausen/CloverAnalysisPipeline.)

Homeolog-Specific Transcriptional Profiling

Raw genomic DNA- and RNA-seq reads for all tissues were assessed for quality with SolexaQA++ v3.14 (Cox et al., 2010). The largest contiguous sections of reads with *Phred* (Ewing and Green, 1998) quality scores >13 were retained and reads with <25 bp discarded.

We applied the software HyLiTE v1.6.2 (Duchemin et al., 2015) to obtain read count matrices for differential expression analysis. As input to HyLiTE, we provided SAM (Li et al., 2009) files. These SAM files were generated by mapping of mRNA reads both paired and unpaired with Bowtie 2 (Langmead et al., 2009; Langmead and Salzberg, 2012) against a set of 36,638 *T. occidentale* reference gene models. For each tissue type a separate set of SAM files was created for each species of clover. To create a combined analysis, these separate SAM files were merged across all tissues. For increasing coverage for SNP calling, genomic DNA sequence data were mapped under the same conditions to the reference gene models. From a pileup file generated with SAMtools (Li et al., 2009), HyLiTE then identified polymorphisms indicative of parental origin for RNA-seq reads and classified reads to parental origin in the hybrid species. The resulting output from HyLiTE comprises tables with read counts for parent species to gene models (counts for orthologs) and read counts of unambiguously assignable reads for the hybrid for each homeolog. (The scripts and the workflow for running HyLiTE are available at https://github.com/MarniTausen/CloverAnalysisPipeline#hylite-pipeline.)

Analysis of Subgenome Expression Ratio Outliers

Expression ratios were first normalized across all progenitor and white clover tissue samples, using TMM (Robinson et al., 2010), which we have previously shown is well suited for correcting for compositional biases in cross-tissue comparisons (Munch et al., 2018). For white clover, the sum of the reads assigned to subgenomes was normalized, and the normalized sum was then redistributed on the white clover subgenomes, maintaining the original subgenome ratio, and then multiplied by two to maintain normalization with respect to the progenitor samples. Principal Component Analysis was performed using the "prcomp" *R* function and results

were plotted using the program ggplot2 (https://ggplot2.tidyverse.org/). Pearson correlation coefficients, deviance, and variance were calculated using standard *R* (v3.4.3) functions. To identify outliers, a subset of 19,954 genes were used, which showed an average of >1 Transcript per Million per sample after TMM normalization and for which at least 40% of the reads were informative for subgenome assignment in the HyLiTE analysis. The variance in $\log_e(Tr_{To}/Tr_{Tp})$ subgenome expression across tissues was calculated. Expression ratio outliers were identified for each of the three experiments (Pooled, S9, and S10) as the genes with a $\log_e(Tr_{To}/Tr_{Tp})$ variance across tissues >1.5 SD above the mean. ANOVA analysis comparing groups of root, leaf, and stolon samples from all experiments, flower samples from Pooled and S9 experiments, and young leaf samples from the S10 experiment, was performed using *R* (v3.4.3). Ten control gene sets were generated by random sampling from the set of 19,954 genes that fulfilled all read count and assignment requirements. (All scripts used for the analysis can be found in Supplemental Data Set 3.) EC numbers were assigned using the software Pathway Tools (http://bioinformatics.ai.sri.com/ptools/) with the MetaCyc database (v2.2.6; http://metacyc.org/) by searching with the acquired MetaCyc gene IDs. Pathway enrichment was evaluated for matches to the MetaCyc database with e-value < 10e-10 using the "Genes enriched for pathways" function of the MetaCyc website, with the enrichment setting, Fisher's exact test, and Benjamini-Hochberg correction for multiple testing.

The amount of diversity of the outlier genes between subgenomes was estimated using the SNPs called through the HyLiTE pipeline. The gene sequences were "mutated" according to the SNPs and translated into protein, if the amino acid change had occurred in the sequence before mutating and after did not match.

Accession Numbers

Original sequence data from this article can be found in the EMBL/GenBank data libraries with the following accession numbers, which are also detailed in Supplemental Tables 16 to 18:

White clover (Bioproject PRJNA523044)

Genome sequence: SRR8670776, SRR8670777, SRR8670778, SRR8670779, SRR8670780, SRR8670781, SRR8670782, SRR8670783, SRR8670784, SRR8670785, SRR8670786, SRR8670787, SRR8670788, SRR8670789, SRR8670790, SRR8670791, SRR8670792, SRR8670793, SRR8670794, SRR8691041.
RNA-seq Pooled: SRR8691037, SRR8691038, SRR8691039, SRR8691040.
RNA-seq S9: SRR8691836, SRR8691837, SRR8691838, SRR8691839, SRR8691840, SRR8691841, SRR8691842, SRR8691843, SRR8691844, SRR8691845, SRR8691846, SRR8691847.
RNA-seq S10: SRR8693957, SRR8693958, SRR8693959, SRR8693960, SRR8693961, SRR8693962, SRR8693963, SRR8693964, SRR8693965, SRR8693966, SRR8693967, SRR8693968.

T. occidentale *(Bioproject PRJNA521254)*

Genome sequence: SRR8593467, SRR8593468, SRR8593469, SRR8593470, SRR8593471, SRR8593472, SRR8593473, SRR8593474, SRR8593475.
RNA-seq Pooled: SRR8691453, SRR8692350, SRR8692351, SRR8692352.
RNA-seq S10: SRR8692338, SRR8692339, SRR8692340, SRR8692341, SRR8692342, SRR8692343, SRR8692344, SRR8692345, SRR8692346, SRR8692347, SRR8692348, SRR8692349.

T. pallescens *(Bioproject PRJNA523043)*

Genome sequence: SRR8617465, SRR8617466, SRR8617467.
RNA-seq Pooled: SRR8675752, SRR8675753, SRR8675754, SRR8675755.
RNA-seq S10: SRR8692353, SRR8692354, SRR8692355, SRR8692356, SRR8692357, SRR8692358, SRR8692359, SRR8692360, SRR8692361, SRR8692362, SRR8692363, SRR8692364.

Supplemental Data

Supplemental Figure 1. *k*-mer–based genome size estimates.

Supplemental Figure 2. White clover assembly ordering and alignment to a genetic linkage map.

Supplemental Figure 3. Chloroplast genome alignments of white clover and extant progenitors.

Supplemental Figure 4. ITS2 alignment for *T. occidentale* and *T. pallescens* compared with Illumina TSLR sequences from white clover.

Supplemental Figure 5. To-versus-Tp chromosome alignment.

Supplemental Figure 6. To-versus-white clover Tr_{To} subgenome chromosome alignment.

Supplemental Figure 7. Tp-versus-white clover Tr_{Tp} subgenome chromosome alignment.

Supplemental Figure 8. Estimation of mutation rates.

Supplemental Figure 9. PSMC analysis of four sequenced white clover individuals using different recombination rates.

Supplemental Figure 10. MSMC analysis of four sequenced white clover individuals using different settings.

Supplemental Figure 11. SFS and SNP density simulations of white clover with different EPS (N_e) expansions.

Supplemental Figure 12. Putative interhomeologous recombination breakpoints identified in white clover.

Supplemental Figure 13. Protein level divergence among progenitors and white clover subgenomes for gene expression ratio outliers.

Supplemental Figure 14. Comparison of extant progenitor and white clover expression patterns.

Supplemental Table 1. Sequencing library statistics.

Supplemental Table 2. Mapped single-locus, homeolog-specific microsatellite markers.

Supplemental Table 3. Pseudomolecule statistics.

Supplemental Table 4. RNA-seq data used in gene annotation and genomic 180-bp library statistics.

Supplemental Table 5. Gene annotation sources.

Supplemental Table 6. Gene annotation summary.

Supplemental Table 7. Chloroplast alignment summary statistics.

Supplemental Table 8. Read statistics for resequenced white clover individuals.

Supplemental Table 9. Chloroplast mapping summary statistics for the four fully sequenced white clover individuals.

Supplemental Table 10. Private progenitor and subgenome diagnostic SNPs.

Supplemental Table 11. Clover genome alignment statistics and estimated divergence times.

Supplemental Table 12. PSMC analysis and mutation accumulation simulation using a range of mutation rates.

Supplemental Table 13. Summary of RNA-seq libraries from the Pooled, S9, and S10 experiments.

Supplemental Table 14. HyLiTE RNA-seq read assignment summary.

Supplemental Table 15. Pathways enriched for outlier genes.

Supplemental Table 16. *T. occidentale* NCBI accession numbers.

Supplemental Table 17. *T. pallescens* NCBI accession numbers.

Supplemental Table 18. *T. repens* NCBI accession numbers.

Supplemental Data Set 1. GBS summary for white clover individuals.

Supplemental Data Set 2. HyLiTE RNA-seq read counts.

Supplemental Data Set 3. Expression ratio outliers.

Supplemental Data Set 4. "R" scripts used for expression analysis.

ACKNOWLEDGMENTS

This work was supported by Pastoral Genomics (a joint venture cofunded by DairyNZ, Beef+Lamb New Zealand, Dairy Australia, AgResearch Ltd, New Zealand Agriseeds Ltd, Grasslands Innovation Ltd, DEEResearch, and the Ministry of Business, Innovation and Employment, New Zealand); Ministry of Business, Innovation and Employment, New Zealand (contract C10X1306, 'Genomics for Production & Security in a Biological Economy'); Innovation Fund Denmark (grant 4105-00007A to S.U.A.); and The Danish Council for Independent Research/Technology and Production Sciences (grant 10-081677 to S.U.A.). We thank Greig Cousins (AgResearch) for the white clover S_8 inbred cv 'Crau' derivative;Cindy Lawley and Karl Sluis of Illumina Inc for early access to the TSLR sequencing technology; and Charles Hefer (AgResearch) for uploading the raw and assembled sequence data to NCBI repositories.

AUTHOR CONTRIBUTIONS

A.G.G. initiated the white clover genome sequencing project, codeveloped the sequencing strategy, and selected the germplasm for sequencing; R.M. codeveloped the sequencing strategy, performed genome assembly, synteny, and progenitor validation analyses; M.T. performed analysis of GBS and resequencing data and carried out PSMC and MSMC analysis, SFS simulations, gene annotation, and HyLiTE analysis; V.G. performed gene annotation; R.M. and V.G. performed subgenome recombination analysis; T.P.B. carried out LD analysis and prepared the genetic map; M.A.C. performed HyLiTE analysis of subgenome specific expression; R.A. performed GBS SNP discovery for the white clover biparental population; I.N. prepared GBS libraries for 200 clover accessions, extracted DNA for resequencing, and analyzed data; A.K. codeveloped the sequencing strategy; A.L. optimized RNA extraction methodology and prepared RNA for sequencing; C.A. and B.F. produced the white clover S9 germplasm, extracted DNA for genome sequencing and from the biparental population, and carried out SSR genotyping; K.H. prepared RNA for sequencing; A.S. developed the white clover biparental population; N.W.E. assembled the chloroplast genomes; M.P.C. supervised HyLiTE analysis; T.A. designed the GBS protocol used for the 200 accession diversity panel and analyzed data; T.M. prepared scripts for divergence time analysis; M.H.S. conceived and supervised mutation accumulation simulations; S.U.A. performed divergence time and expression ratio outlier analysis; A.G.G. and S.U.A. designed and supervised the study, interpreted results, and wrote the article with input from all authors.

Received August 16, 2018; revised March 15, 2019; accepted April 22, 2019; published April 25, 2019.

REFERENCES

Aasmo Finne, M., Rognli, O.A., and Schjelderup, I. (2000). Genetic variation in a Norwegian germplasm collection of white clover (*Trifolium repens* L.). Euphytica **112:** 45–56.

Abberton, M.T., Fothergill, M., Collins, R.P., and Marshall, A.H. (2006). Breeding forage legumes for sustainable and profitable farming systems. Asp. Appl. Biol. **80:** 81–87.

Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc. Natl. Acad. Sci. USA **100:** 4649–4654.

Ainouche, M.L., Fortune, P.M., Salmon, A., Parisod, C., Grandbastien, M.-A., Fukunaga, K., Ricou, M., and Misset, M.-T. (2009). Hybridization, polyploidy and invasion: Lessons from *Spartina* (Poaceae). Biol. Invasions **11:** 1159–1173.

Akama, S., Shimizu-Inatsugi, R., Shimizu, K.K., and Sese, J. (2014). Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. Nucleic Acids Res. **42:** e46.

Álvarez, I., and Wendel, J.F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. **29:** 417–434.

Anderson, C.B., Franzmayr, B.K., Hong, S.W., Larking, A.C., van Stijn, T.C., Tan, R., Moraga, R., Faville, M.J., and Griffiths, A.G. (2018). Protocol: A versatile, inexpensive, high-throughput plant genomic DNA extraction method suitable for genotyping-by-sequencing. Plant Methods **14:** 75.

Ansari, H.A., Ellison, N.W., Reader, S.M., Badaeva, E.D., Friebe, B., Miller, T.E., and Williams, W.M. (1999). Molecular cytogenetic organization of 5S and 18S–26S rDNA loci in white clover (*Trifolium repens* L.) and related species. Ann. Bot. **83:** 199–206.

Ansari, H.A., Ellison, N.W., and Williams, W.M. (2008). Molecular and cytogenetic evidence for an allotetraploid origin of *Trifolium dubium* (Leguminosae). Chromosoma **117:** 159–167.

Bardil, A., de Almeida, J.D., Combes, M.C., Lashermes, P., and Bertrand, B. (2011). Genomic expression dominance in the natural allopolyploid *Coffea arabica* is massively affected by growth temperature. New Phytol. **192:** 760–774.

Bennett, M.D., and Leitch, I.J. (2011). Nuclear DNA amounts in angiosperms: Targets, trends and tomorrow. Ann. Bot. **107:** 467–590.

Bertioli, D.J., et al. (2016) The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat. Genet. **48:** 438–446.

Bilton, T.P., McEwan, J.C., Clarke, S.M., Brauning, R., van Stijn, T.C., Rowe, S.J., and Dodds, K.G. (2018). Linkage disequilibrium estimation in low coverage high-throughput sequencing data. Genetics **209:** 389–400.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics **27:** 578–579.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30:** 2114–2120.

Bottley, A., Xia, G.M., and Koebner, R.M.D. (2006). Homoeologous gene silencing in hexaploid wheat. Plant J. **47:** 897–906.

Branca, A., et al. (2011) Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proc. Natl. Acad. Sci. USA **108:** E864–E870.

Brochmann, C., Brysting, A.K., Alsos, I.G., Borgen, L., Grundt, H.H., Scheen, A.C., and Elven, R. (2004). Polyploidy in arctic plants. Biol. J. Linn. Soc. Lond. **82:** 521–536.

Buggs, R.J.A., Elliott, N.M., Zhang, L., Koh, J., Viccini, L.F., Soltis, D.E., and Soltis, P.S. (2010). Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. New Phytol. **186:** 175–183.

Cao, M., Fraser, K., Jones, C., Stewart, A., Lyons, T., Faville, M., and Barrett, B. (2017). Untargeted metabotyping *Lolium perenne* reveals population-level variation in plant flavonoids and alkaloids. Front. Plant Sci. **8:** 133.

Caspi, R., et al. (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. **44** (D1): D471–D480.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. (2013). STACKS: An analysis tool set for population genomics. Mol. Ecol. **22:** 3124–3140.

Cenci, A., Combes, M.C., and Lashermes, P. (2012). Genome evolution in diploid and tetraploid *Coffea* species as revealed by comparative analysis of orthologous genome segments. Plant Mol. Biol. **78:** 135–145.

Chelaifa, H., Monnier, A., and Ainouche, M. (2010). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species *Spartina* × *townsendii* and *Spartina anglica* (Poaceae). New Phytol. **186:** 161–174.

Chen, J., Glémin, S., and Lascoux, M. (2017). Genetic diversity and the efficacy of purifying selection across plant and animal species. Mol. Biol. Evol. **34:** 1417–1428.

Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G., and Wang, X. (2012). Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. PLoS One **7:** e36442.

Chester, M., Gallagher, J.P., Symonds, V.V., Cruz da Silva, A.V., Mavrodiev, E.V., Leitch, A.R., Soltis, P.S., and Soltis, D.E. (2012). Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragopogon miscellus* (Asteraceae). Proc. Natl. Acad. Sci. USA **109:** 1176–1181.

Combes, M.-C., Dereeper, A., Severac, D., Bertrand, B., and Lashermes, P. (2013). Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. New Phytol. **200:** 251–260.

Combes, M.C., Hueber, Y., Dereeper, A., Rialle, S., Herrera, J.C., and Lashermes, P. (2015). Regulatory divergence between parental alleles determines gene expression patterns in hybrids. Genome Biol. Evol. **7:** 1110–1121.

Coombe, D. (1961). *Trifolium occidentale*, a new species related to *T. repens* L. Watsonia **5:** 68–87.

Cousins, G., and Woodfield, D.R. (2006). Effect of inbreeding on growth of white clover. *In* 13th Australasian Plant Breeding Conference, C.F. Mercer, ed. New Zealand Grassland Association, Christchurch, New Zealand, pp 568–572.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics **11:** 485.

Daday, H. (1958). Gene frequencies in wild populations of *Trifolium repens* L. III. World distribution. Heredity **12:** 169–184.

Darling, A.C.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. **14:** 1394–1403.

Davies, K.M., Albert, N.W., Zhou, Y., and Schwinn, K.E. (2018). Functions of flavonoid and betalain pigments in abiotic stress tolerance in plants. Ann. Plant Rev. Online **1:** 1–41.

DePristo, M.A., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. **43:** 491–498.

De Storme, N., and Geelen, D. (2014). The impact of environmental stress on male reproductive development in plants: Biological processes and molecular mechanisms. Plant Cell Environ. **37:** 1–18.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics **29:** 15–21.

Douglas, G.M., et al. (2015) Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. Proc. Natl. Acad. Sci. USA **112:** 2806–2811.

Duchemin, W., Dupont, P.Y., Campbell, M.A., Ganley, A.R., and Cox, M.P. (2015). HyLiTE: Accurate and flexible analysis of gene expression in hybrid and allopolyploid species. BMC Bioinformatics **16:** 8.

Ellison, N.W., Liston, A., Steiner, J.J., Williams, W.M., and Taylor, N.L. (2006). Molecular phylogenetics of the clover genus (*Trifolium*–Leguminosae). Mol. Phylogenet. Evol. **39:** 688–705.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6:** e19379.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using *phred*. II. Error probabilities. Genome Res. **8:** 186–194.

Fawcett, J.A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc. Natl. Acad. Sci. USA **106:** 5737–5742.

Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, A.A., and Vega, J.M. (1997). Rapid elimination of low-copy DNA sequences in polyploid wheat: A possible mechanism for differentiation of homoeologous chromosomes. Genetics **147:** 1381–1387.

Ferreira de Carvalho, J., Poulain, J., Da Silva, C., Wincker, P., Michon-Coudouel, S., Dheilly, A., Naquin, D., Boutte, J., Salmon, A., and Ainouche, M. (2013). Transcriptome de novo assembly from next-generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). Heredity **110:** 181–193.

Franzmayr, B.K., Rasmussen, S., Fraser, K.M., and Jameson, P.E. (2012). Expression and functional characterization of a white clover isoflavone synthase in tobacco. Ann. Bot. **110:** 1291–1301.

Garsmeur, O., Schnable, J.C., Almeida, A., Jourda, C., D'Hont, A., and Freeling, M. (2014). Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. **31:** 448–454.

Gnerre, S., et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA **108:** 1513–1518.

Griffiths, A.G., Barrett, B.A., Simon, D., Khan, A.K., Bickerstaff, P., Anderson, C.B., Franzmayr, B.K., Hancock, K.R., and Jones, C.S. (2013). An integrated genetic linkage map for white clover (*Trifolium repens* L.) with alignment to *Medicago*. BMC Genomics **14:** 388.

Gross, B.L., Kane, N.C., Lexer, C., Ludwig, F., Rosenthal, D.M., Donovan, L.A., and Rieseberg, L.H. (2004). Reconstructing the origin of *Helianthus deserticola*: Survival and selection on the desert floor. Am. Nat. **164:** 145–156.

Grover, C.E., Gallagher, J.P., Szadkowski, E.P., Yoo, M.J., Flagel, L.E., and Wendel, J.F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol. **196:** 966–971.

Hahsler, M, Buchta, C, and Hornik, K. (2016). Infrastructure for seriation. R Package Version 1.2-1. https://cran.r-project.org/web/packages/seriation/index.html.

Harris, R.S. (2007). Improved pairwise alignment of genomic DNA. PhD dissertation, College of Engineering, The Pennsylvania State University, p 84.

Herten, K., Hestand, M.S., Vermeesch, J.R., and Van Houdt, J.K. (2015). GBSX: A toolkit for experimental design and demultiplexing genotyping by sequencing experiments. BMC Bioinformatics **16:** 73.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2017). BRAKER2 (http://github.com/Gaius-Augustus/BRAKER). Downloaded September 28, 2018.

Hoffmann, M.H. (2005). Evolution of the realized climatic niche in the genus Arabidopsis (Brassicaceae). Evolution **59:** 1425–1436.

Hovav, R., Udall, J.A., Chaudhary, B., Rapp, R., Flagel, L., and Wendel, J.F. (2008). Partitioned expression of duplicated genes during development and evolution of a single cell in a polyploid plant. Proc. Natl. Acad. Sci. USA **105:** 6191–6195.

IWGSC (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science **345:** 1251788.

Jamalnasir, H., Wagiran, A., Shaharuddin, N.A., and Samad, A.A. (2013). Isolation of high quality RNA from plant rich in flavonoids, *Melastoma decemfidum* Roxb ex. Jack. Aust. J. Crop Sci. **7:** 911–916.

Jia, J., et al.; International Wheat Genome Sequencing Consortium (2013). *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature **496:** 91–95.

Jouzel, J., Masson-Delmotte, V., Cattani, O., Dreyfus, G., Falourd, S., Hoffmann, G., Minster, B., Nouet, J., Barnola, J.M., Chappellaz, J., Fischer, H., and Gallet, J.C., et al. (2007) Orbital and millennial Antarctic climate variability over the past 800,000 years. Science **317:** 793–796.

Kearse, M., et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28:** 1647–1649.

Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLOS Comput. Biol. **12:** e1004842.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14:** R36.

Kovarik, A., Dadejova, M., Lim, Y.K., Chase, M.W., Clarkson, J.J., Knapp, S., and Leitch, A.R. (2008). Evolution of rDNA in *Nicotiana* allopolyploids: A potential link between rDNA homogenization and epigenetics. Ann. Bot. **101:** 815–823.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods **9:** 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:** R25.

Leach, L.J., Belfield, E.J., Jiang, C., Brown, C., Mithani, A., and Harberd, N.P. (2014). Patterns of homoeologous gene expression shown by RNA sequencing in hexaploid bread wheat. BMC Genomics **15:** 276.

Leitch, A.R., and Leitch, I.J. (2008). Genomic plasticity and the diversity of polyploid plants. Science **320:** 481–483.

Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature **475:** 493–496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:** 2078–2079.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., and Barker, M.S. (2015). Early genome duplications in conifers and other seed plants. Sci. Adv. **1:** e1501084.

Liu, Z., and Adams, K.L. (2007). Expression partitioning between genes duplicated by polyploidy under abiotic stress and during organ development. Curr. Biol. **17:** 1669–1674.

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., and Sun, Q. (2015). Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.). BMC Plant Biol. **15:** 152.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. **33:** 6494–6506.

Lynch, M. (2010). Evolution of the mutation rate. Trends Genet. **26:** 345–352.

Mable, B.K. (2003). Breaking down taxonomic barriers in polyploidy research. Trends Plant Sci. **8:** 582–590.

Madlung, A. (2013). Polyploidy and its effect on evolutionary success: Old questions revisited with new tools. Heredity **110:** 99–104.

Mailund, T., Halager, A.E., Westergaard, M., Dutheil, J.Y., Munch, K., Andersen, L.N., Lunter, G., Prüfer, K., Scally, A., Hobolth, A., and Schierup, M.H. (2012). A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet. **8:** e1003125.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics **20:** 2878–2879.

Mangerud, J., et al. (2004) Ice-dammed lakes and rerouting of the drainage of northern Eurasia during the Last Glaciation. Quat. Sci. Rev. **23:** 1313–1332.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics **27:** 764–770.

Mason, A.S., and Pires, J.C. (2015). Unreduced gametes: Meiotic mishap or evolutionary mechanism? Trends Genet. **31:** 5–10.

McClintock, B. (1984). The significance of responses of the genome to challenge. Science **226:** 792–801.

Mierziak, J., Kostyn, K., and Kulma, A. (2014). Flavonoids as important molecules of plant interactions with the environment. Molecules **19:** 16240–16265.

Mouradov, A., and Spangenberg, G. (2014). Flavonoids: A metabolic network mediating plants adaptation to their real estate. Front. Plant Sci. **5:** 620.

Munch, D., Gupta, V., Bachmann, A., Busch, W., Kelly, S., Mun, T., and Andersen, S.U. (2018). The Brassicaceae family displays divergent, shoot-skewed NLR resistance gene expression. Plant Physiol. **176:** 1598–1609.

Nadachowska-Brzyska, K., Burri, R., Smeds, L., and Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. Mol. Ecol. **25:** 1058–1072.

Novikova, P.Y., et al. (2016) Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Genet. **48:** 1077–1082.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science **327:** 92–94.

Otto, S.P. (2007). The evolutionary consequences of polyploidy. Cell **131:** 452–462.

Otto, S.P., and Whitton, J. (2000). Polyploid incidence and evolution. Annu. Rev. Genet. **34:** 401–437.

Ownbey, M. (1950). Natural hybridization and amphiploidy in the genus Tragopogon. Am. J. Bot. **37:** 487–499.

Paape, T., Hatakeyama, M., Shimizu-Inatsugi, R., Cereghetti, T., Onda, Y., Kenta, T., Sese, J., and Shimizu, K.K. (2016). Conserved but attenuated parental gene expression in allopolyploids: Constitutive zinc hyperaccumulation in the allotetraploid *Arabidopsis kamchatica*. Mol. Biol. Evol. **33:** 2781–2800.

Raffl, C., Holderegger, R., Parson, W., and Erschbamer, B. (2008). Patterns in genetic diversity of *Trifolium pallescens* populations do not reflect chronosequence on alpine glacier forelands. Heredity **100:** 526–532.

Ramírez-González, R.H., et al.; International Wheat Genome Sequencing Consortium (2018). The transcriptional landscape of polyploid wheat. Science **361:** 6403.

Ramsey, J., and Schemske, D.W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst. **29:** 467–501.

Rapp, R.A., Udall, J.A., and Wendel, J.F. (2009). Genomic expression dominance in allopolyploids. BMC Biol. **7:** 18.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26:** 139–140.

Sanggaard, K.W., et al. (2014) Spider genomes provide insight into composition and evolution of venom and silk. Nat. Commun. **5:** 3765.

Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. Nat. Genet. **46:** 919–925.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA **108:** 4069–4074.

Selmecki, A.M., Maruvka, Y.E., Richmond, P.A., Guillet, M., Shoresh, N., Sorenson, A.L., De, S., Kishony, R., Michor, F., Dowell, R., and Pellman, D. (2015). Polyploidy can drive rapid adaptation in yeast. Nature **519:** 349–352.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31:** 3210–3212.

Smit, A.F.A., Hubley, R., and Green, P. (2015). RepeatMasker Open-4.0 (http://www.repeatmasker.org). Downloaded December 2018.

Soltis, P.S., and Soltis, D.E. (2016). Ancient WGD events as drivers of key innovations in angiosperms. Curr. Opin. Plant Biol. **30:** 159–165.

Soltis, D.E., Soltis, P.S., Pires, J.C., Kovarik, A., Tate, J.A., and Mavrodiev, E. (2004). Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): Cytogenetic, genomic and genetic comparisons. Biol. J. Linn. Soc. Lond. **82:** 485–501.

Soltis, D.E., Visger, C.J., Marchant, D.B., and Soltis, P.S. (2016). Polyploidy: Pitfalls and paths to a paradigm. Am. J. Bot. **103:** 1146–1166.

Soltis, P.S., Marchant, D.B., Van de Peer, Y., and Soltis, D.E. (2015). Polyploidy and genome evolution in plants. Curr. Opin. Genet. Dev. **35:** 119–125.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. Nucleic Acids Res. **34:** W435–W439.

Tang, H., et al. (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics **15:** 312.

Tayalé, A., and Parisod, C. (2013). Natural pathways to polyploidy in plants and consequences for genome reorganization. Cytogenet. Genome Res. **140:** 79–96.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. **7:** 562–578.

Tsafrir, D., Tsafrir, I., Ein-Dor, L., Zuk, O., Notterman, D.A., and Domany, E. (2005). Sorting points into neighborhoods (SPIN): Data analysis and visualization by ordering distance matrices. Bioinformatics **21:** 2301–2308.

Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014b). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. **24:** 1334–1347.

Vanneste, K., Maere, S., and Van de Peer, Y. (2014a). Tangled up in two: A burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. **369:** 20130353.

Wang, X., et al.; Brassica rapa Genome Sequencing Project Consortium (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. **43:** 1035–1039.

Wang, J., Tian, L., Madlung, A., Lee, H.S., Chen, M., Lee, J.J., Watson, B., Kagochi, T., Comai, L., and Chen, Z.J. (2004). Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. Genetics **167:** 1961–1973.

Williams, C.A., and Grayer, R.J. (2004). Anthocyanins and other flavonoids. Nat. Prod. Rep. **21:** 539–573.

Williams, W.M., Mason, K.M., and Williamson, M.L. (1998). Genetic analysis of shikimate dehydrogenase allozymes in *Trifolium repens* L. Theor. Appl. Genet. **96:** 859–868.

Williams, W.M., Ellison, N.W., Ansari, H.A., Verry, I.M., and Hussain, S.W. (2012). Experimental evidence for the ancestry of allotetraploid *Trifolium repens* and creation of synthetic forms with value for plant breeding. BMC Plant Biol. **12:** 55.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. **15:** R46.

Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., and Rieseberg, L.H. (2009). The frequency of polyploid speciation in vascular plants. Proc. Natl. Acad. Sci. USA **106:** 13875–13879.

Woodhouse, M.R., Cheng, F., Pires, J.C., Lisch, D., Freeling, M., and Wang, X. (2014). Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc. Natl. Acad. Sci. USA **111:** 5283–5288.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21:** 1859–1875.

Yang, J., et al. (2016) The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nat. Genet. **48:** 1225–1232.

Yokoyama, Y., Lambeck, K., De Dekker, P., Johnston, P., and Fifield, L.K. (2000). Timing of the Last Glacial Maximum from observed sea-level minima. Nature **406:** 713–716.

Yoo, M.J., Szadkowski, E., and Wendel, J.F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity **110:** 171–180.

Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:** 821–829.

Zhang, T., et al. (2015b) Sequencing of allotetraploid cotton (*Gossypium hirsutum* L. acc. TM-1) provides a resource for fiber improvement. Nat. Biotechnol. **33:** 531–537.

Zhang, D., Pan, Q., Cui, C., Tan, C., Ge, X., Shao, Y., and Li, Z. (2015a). Genome-specific differential gene expressions in resynthesized *Brassica* allotetraploids from pair-wise crosses of three cultivated diploids revealed by RNA-seq. Front. Plant Sci. **6:** 957.

Zhang, X., Zhang, Y., Yan, R., Han, J., Fuzeng, H., Wang, J., and Cao, K. (2010). Genetic variation of white clover (*Trifolium repens* L.) collections from China detected by morphological traits, RAPD and SSR. Afr. J. Biotechnol. **9:** 3033–3041.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics **29:** 2669–2677.