Patterns of Genome-Wide Nucleotide Diversity in the Gynodioecious Plant *Thymus vulgaris* Are Compatible with Recent Sweeps of Cytoplasmic Genes

Maeva Mollion[1], Bodil K. Ehlers[2], Emeric Figuet[3], Sylvain Santoni[4], Thomas Lenormand[5], Sandrine Maurice[3], Nicolas Galtier[3], and Thomas Bataillon[1,*]

[1]Bioinformatics Research Center, Aarhus University, C.F. Møllers Alle 8, Building 1110, 8000 Aarhus C, Denmark

[2]Department of Bioscience, Aarhus University, Vejlsøvej 25, 8600 Silkeborg, Denmark

[3]Institut des Sciences de l'Evolution, UMR5554 — Université de Montpellier — CNRS — IRD — EPHE, Place E. Bataillon — CC64, 34095 Montpellier, France

[4]Centre d'Ecologie Fonctionnelle et Evolutive (CEFE), CNRS, 1919, route de Mende 34293 Montpellier, France

[5]Centre d'Ecologie Fonctionnelle et Evolutive, CNRS, 1919, route de Mende 34293 Montpellier, France

*Corresponding author: E-mail: tbata@birc.au.dk.

Accepted: December 19, 2017

Data deposition: The set of contigs and summaries of the polymorphism data per contig are available at the Dryad repository (doi:10.5061/dryad.813mf). Reads were deposited in the SRA archive with all accession numbers available through the BioProject identifier PRJNA417241 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA417241).

Abstract

Gynodioecy is a sexual dimorphism where females coexist with hermaphrodite individuals. In most cases, this dimorphism involves the interaction of cytoplasmic male sterility (CMS) genes and nuclear restorer genes. Two scenarios can account for how these interactions maintain gynodioecy. Either CMS genes recurrently enter populations at low frequency via mutation or migration and go to fixation unimpeded (successive sweeps), or CMS genes maintain polymorphism over evolutionary time through interactions with a nuclear restorer allele (balanced polymorphism). To distinguish between these scenarios, we used transcriptome sequencing in gynodioecious *Thymus vulgaris* and surveyed genome-wide diversity in 18 naturally occurring individuals sampled from populations at a local geographic scale. We contrast the amount and patterns of nucleotide diversity in the nuclear and cytoplasmic genome, and find ample diversity at the nuclear level ($\pi = 0.019$ at synonymous sites) but reduced genetic diversity and an excess of rare polymorphisms in the cytoplasmic genome relative to the nuclear genome. Our finding is incompatible with the maintenance of gynodioecy via scenarios invoking long-term balancing selection, and instead suggests the recent fixation of CMS lineages in the populations studied.

Key words: balancing selection, cytoplasmic male sterility, single nucleotide polymorphism, selective sweeps, *Thymus vulgaris*.

Introduction

Gynodioecy is a reproductive system in which two sexual morphs (hermaphrodites and females) coexist. Richards (1997) states that 7% of angiosperm species are gynodioecious, but this figure is highly dependent on the flora studied, and other authors suggest a much lower frequency of gynodioecy (Charlesworth 2002). The most recent data survey comprising 449 plant families and 13,208 genera indicates that gynodioecy occurs in 2% of angiosperm genera and in 18% of all angiosperm families (Dufay et al., 2014). Together with monoecy (Renner, 2014), gynodioecy is viewed as a key pathway underlying the evolutionary transition from hermaphroditism to separate sexes (dioecy) during the diversification of flowering plants (Charlesworth and Charlesworth, 1978; Dufay et al., 2014). Yet rather than being a purely transitory stage, gynodioecy also appears to be a stable sexual system in several genera such as *Thymus, Saponaria, Dianthus*, and *Silene* (Desfeux et al., 1996;

Manicacci et al., 1998). As a consequence, the maintenance of gynodioecy (the coexistence of pure females and hermaphrodites) has attracted considerable theoretical attention (Frank, 1989; Hurst et al., 1996; Couvet et al., 1998; McCauley and Olson, 2008; Delph and Kelly, 2014).

One of the key factors influencing the maintenance of gynodioecy lies in its underlying genetics (Lewis, 1941). Sex determination in most gynodioecious species involves cytoplasmic genes that cause male sterility (cytoplasmic male sterility, hereafter CMS), which turn hermaphrodites into females, and nuclear restorer alleles that "restore" the male function (Couvet et al., 1990; Schnable and Wise, 1998). A female plant carries a CMS gene that blocks the development of functional anthers, whereas hermaphrodites either lack a CMS gene or carry one in combination with a nuclear allele for restoring the male function (Touzet and Meyer, 2014). Because CMS genes are maternally inherited, they gain fitness differently from nuclear genes (Hurst et al., 1996). In particular, CMS genes can invade a population of hermaphrodites provided the females have a small fertility advantage over hermaphrodites. However, in a population containing a CMS gene, any nuclear allele able to restore the male function will also be favored. It is expected that as the frequency of the CMS gene increases in the population, selection will more strongly favor the restorer alleles, which should subsequently spread in the population.

The maintenance of females in a population therefore depends either on a mechanism that prevents the fixation of nuclear restorer alleles, or on the recurring appearance of new CMS genes. This has led to two main scenarios for the maintenance of gynodioecy (Charlesworth, 2002; Delph and Kelly, 2014). In the first scenario, a new CMS type enters the population and starts spreading, creating the condition for a matching restorer allele to also sweep to fixation. Under this scenario, females can cooccur with hermaphrodites because new CMS lineages regularly enter the population, either by spontaneous mutation or via migrants carrying different CMS genes (Frank, 1989). This scenario, initially termed "an epidemic scenario" by Frank (1989), is hereafter labelled scenario A (see table 1 for a summary). Under the second scenario, usually termed "the balancing selection scenario" (hereafter scenario B, table 1), polymorphism is maintained by both the nuclear restorer and the CMS alleles by negative frequency-dependent selection (Charlesworth, 1981; Gouyon et al., 1991). This scenario implies a negative pleiotropic fitness effect (or "cost") associated with carrying a nuclear restorer allele in combination with a CMS gene that it does *not* restore. Once a CMS lineage and the corresponding nuclear restorer allele reach a high frequency in the population, restorer alleles of other CMS genes are counter-selected because they tend to cooccur with the frequent CMS gene that they do not restore. As a consequence, the frequency of females among the carriers of rare CMS genes will grow, ultimately leading to an increase in the frequency of these rare CMS lineages and a decrease in the common one. This negative frequency dependence leads to cyclical dynamics where the amplitude of the cycles depends mainly on the magnitude of the cost (Gouyon et al., 1991; Dufaÿ et al., 2007).

Finally, Couvet et al. (1998) showed that polymorphism for CMS types and their matching restorer alleles can also be maintained if populations are embedded in a metapopulation structure (Scenario C, table 1) where frequent founder events arise and new populations are created by a restricted number of originators from a migrant pool. Note that these different scenarios are not mutually exclusive. For instance, the metapopulation effect described by Couvet et al. (1998) could apply irrespective of the assumption regarding the cost of restoration (scenario B) or the frequency at which new CMS types or matching restorer types are introduced via mutation or migration (scenario A). Each of these scenarios is best viewed as a minimum set of assumptions needed to ensure the theoretical maintenance of gynodioecy.

Directly testing which of these processes actually governs the maintenance of gynodioecy in a given species is a daunting task, as it requires measuring whether nuclear restorer alleles carry a fitness cost, and documenting these costs in nature is challenging. Of all the studies conducted in gynodioecious species, very few have provided evidence for a cost of restoration (De Haan et al., 1997, in *Plantago lanceolata*; Bailey, 2002, in *Lobelia siphilitica*; Del Castillo and Trujillo, 2009, in *Phacelia dubia*). Furthermore, evidence for a cost of restoration makes scenario B credible without necessarily excluding scenario A.

Here, we argue that it is possible to distinguish between some of these scenarios using an indirect yet powerful source of information: specifically, the patterns of nucleotide diversity in the nuclear genome compared with the chloroplast and mitochondrial genomes (collectively referred to as the cytoplasmic genome). Given that no formal expectation for observed patterns of genome diversity has yet been derived for the scenarios in table 1, we first sketch why examining patterns of nucleotide diversity can be insightful. Scenarios A and B are expected to leave radically different footprints in the amount and patterns of nucleotide diversity in nuclear and cytoplasmic genomes. Under scenario A, we expect the cytoplasmic genomes to harbor patterns of genetic diversity typical of recent selective sweeps, whereas under scenario B, the cytoplasmic genomes should exhibit a signature of balancing selection (Ingvarsson and Taylor, 2002; Städler and Delph, 2002; Delph and Kelly, 2014; see fig. 1). Furthermore, mitochondrial and chloroplast genomes are uniparentally (maternally) inherited and hardly recombine (but see Barr et al., 2005, and Davila et al., 2011, for evidence of occasional double-strand breaks in plant mitochondrial genomes). The cytoplasmic genome is therefore expected to behave as a single "super locus." Thus, selection acting on any given CMS gene is expected to imprint the whole "super locus" by linkage, making the footprints of selection potentially

Table 1
Summary of Possible Scenarios for the Maintenance of Gynodioecy and Their Expected Impact on Patterns of Genome Diversity

Scenario	Expected Footprint on Genome Diversity	References
(A) Epidemic scenario		
New CMS genes and matching nuclear restorer alleles regularly enter the population (via mutation and/or migration) and sweep to fixation.	Strong recent sweep signal in the whole cytoplasmic genome Sweep signal at and in the immediate vicinity of restorer genes Neutrality in the nuclear genome except for demographic and sampling effects	Frank (1989)
(B) Balancing selection		
One or more CMS types and their matching nuclear restorer alleles carrying pleiotropic fitness costs reach a polymorphic equilibrium. This equilibrium goes from punctuated equilibria to cycles of varying amplitude and period depending on the parameter values.	Balancing selection, possibly complicated by the amplitude of the cycles in the whole cytoplasmic genome and at the nuclear restorer genes and their immediate genomic vicinity. Neutrality in the rest of the nuclear genome.	Charlesworth (1981) Gouyon et al. (1991) Bailey et al. (2003) Dufaÿ et al. (2007)
(C) Metapopulation effect		
A metapopulation with extinction–recolonization with two CMS genes. Strong founder effects with no subsequent migration and the fact that populations have different outputs depending on their local sex ratio are enough to maintain polymorphism of both CMS genes and their matching restorer alleles.	Balancing selection, possibly complicated by the metapopulation effect. Nuclear genome neutrality, but the effect of sampling and metapopulation functioning might skew expectations relative to a stable nonsubdivided population.	Couvet et al. (1998)

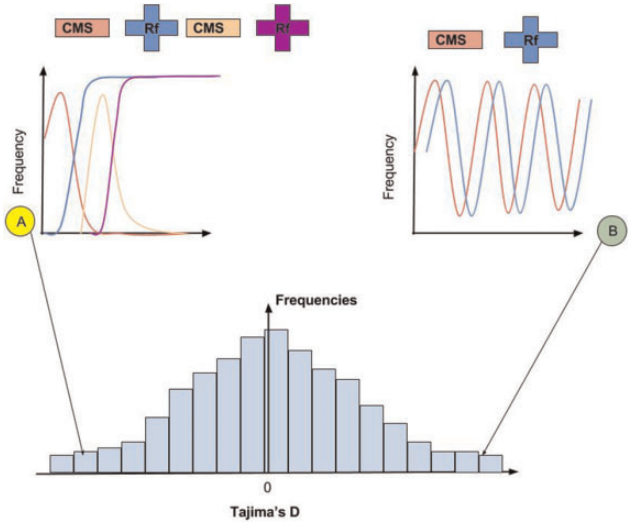

Fig. 1.—Contrasting expectations for Tajima's D statistic in cytoplasmic genomes relative to nuclear genomes under scenarios A and B. Scenario A involves recurrent selective sweeps of new CMS alleles (counteracted by recurrent selection of novel restorer [Rf] alleles that also periodically go to fixation). Scenario B involves balancing selection that maintains both CMS and restorer gene polymorphism over long time periods. If a recent sweep has affected CMS (scenario A), one expects fragments anchored in the cytoplasmic genome to exhibit negative *TD* values relative to the values in the nuclear genome. Conversely, if balancing selection affects CMS (scenario B), we expect both cytoplasmic organelles to exhibit positive *TD* values relative to the nuclear genome. The histogram represents the hypothetical distribution of expected *TD* values across nuclear fragments under neutrality. Further description of these scenarios is given in table 1.

conspicuous relative to those expected under neutrality (fig. 1). In contrast, selection affecting a nuclear gene involved in restoration will only affect patterns of polymorphism at nearby genomic regions, leaving the rest of the nuclear genome unaffected. Therefore, patterns of genetic diversity in the nuclear genome can be used as a reference that reflects sampling and demographic effects independent of gynodioecy-related selection.

We study *Thymus vulgaris*, a gynodioecious species with nucleo-cytoplasmic sex inheritance. The genus *Thymus* contains many gynodioecious species but no record of dioecy. All gynodioecious *Thymus* species studied so far exhibit a high (>50% on average) and variable frequency of females (Manicacci et al., 1998; Thompson, 2002). In *T. vulgaris*, female frequency can vary from 10% to 85% among populations (Belhassen et al., 1989; Manicacci et al., 1998). This species has previously been used to study the genetics of sex inheritance (Charlesworth and Laporte, 1998; Ehlers et al., 2005) and the cost of restoration (Gigord et al., 1999). Both scenarios A and B and the metapopulation effect studied by Couvet et al. (1998) can potentially account for the variability in female frequency observed among populations (Belhassen et al., 1990).

We sequenced the transcriptome of 18 thyme plants that were grown from seeds collected in natural populations in southern France. We used a simple yet popular summary statistic, Tajimas's *D* (hereafter T*D*), to quantify patterns of nucleotide diversity and test selective neutrality. This statistic is often applied to summarize the effect of selection on the site frequency spectrum (SFS) of nucleotide polymorphism

(Tajima, 1989). It is particularly useful here as it is expected to behave differently under scenarios A and B: very negative estimates of *TD* are expected following a recent selective sweep (scenario A) and highly positive values under balancing selection (scenario B, see Conceptual fig. 1). *TD* values can also be influenced by demographic effects in the history of the sampled individuals. We therefore use the empirical distribution of *TD* in the nuclear genome to control for the effects of an unknown demography on the SFS.

Materials and Methods

Plant Material

Eighteen individual thyme plants were used for this study. All plants were obtained from seeds collected from four distinct populations around Saint Martin de Londres, 25 km north of Montpellier (an area comprising roughly 10 km North–South by 10 km East–West). All populations were at least 2 km apart. All plants were raised in the experimental garden of "LabEx CeMEB" in Montpellier. We collected young leaves from each individual plant in October at the time where plants begin to produce new leaves after the summer drought. As sampling occurred after the flowering season, the plants were not sex determined. However, an individual's sex was not expected to affect patterns of analyzed nucleotide diversity, which reflects population-level processes.

Previous work documented very different sex ratios among local populations in that region and uncovered functionally different CMS lineages via controlled crosses (Belhassen et al., 1991, 1993; Manicacci et al., 1996). Recent field observations (2012–2017) of the populations sampled in the present study confirm the variation in sex ratios between the sampling locations used (BKE, TB personal observations). Our regional sampling is thus at a meaningful scale to test scenarios underlying the maintenance of gynodioecy.

Preparation of RNA Samples

Leaf samples were ground in liquid nitrogen and total cellular RNA was extracted using a Sigma-Aldrich (St. Louis, MO) Spectrum™ Plant Total RNA kit with a DNase treatment followed by quantification with the Invitrogen (Carlsbad, CA) Quant-iT™ RiboGreen RNA reagent based on the manufacturer's protocol on a Tecan (Männedorf, SWISS) Genios spectrofluorometer. RNA quality was assessed by running 1 μl of each RNA sample on an RNA 6000 Pico chip on an Agilent Bioanalyzer 2100 (Santa Clara, CA). Samples with an RNA Integrity Number (RIN) value greater than eight were deemed acceptable according to the Illumina sequencing protocols.

Library Production

The Illumina, Inc. (San Diego, CA) TruSeq RNA Sample Preparation Kit v2 was used according to the manufacturer's protocol with the following modifications. In brief, poly-A containing mRNA molecules were purified from 2 μg of total RNA using poly-T oligo attached magnetic beads. The purified mRNA was cut by addition of the fragmentation buffer and heated at 94 °C in a thermocycler for 4 min to yield library fragments of 300–400 bp. First strand cDNA was synthesized using random primers to eliminate the general bias towards the 3′ end of the transcript. Second-strand cDNA synthesis, end repair, A-tailing, and adapter ligation was done in accordance with the protocols supplied by the manufacturer. Purified cDNA templates were enriched by 15 cycles of PCR for 10 s at 98 °C, 30 s at 65 °C, and 30 s at 72 °C using PE1.0 and PE2.0 primers and with Phusion High-Fidelity DNA polymerase (New England Biolabs, Ipswich, MA). Each indexed cDNA library was verified and quantified using a DNA 100 chip on a Bioanalyzer 2100 then equally mixed with nine (from different genotypes). The final library was then quantified by real time PCR with the Kapa Biosystems (Wilmington, MA) Library Quantification Kit for Illumina Sequencing Platforms adjusted to 10 nM in water and sent to the GeT core facility (Toulouse, FRANCE, http://get.genotoul.fr/) for sequencing.

Library Clustering and Sequencing Conditions

The final mixed cDNA library was sequenced using the Illumina mRNA-seq paired-end protocol on a HiSeq2000 sequencer for 2 × 100 cycles. The library was diluted to 2 nM with NaOH and 2.5 μl transferred into 497.5 μl HT1 to give a final concentration of 10 pM. Then 120 μl was transferred to a 200 μl strip tube and placed on ice before loading onto the Cluster Station mixed library from the ten individual indexed libraries being run on a single lane. The flow cells were clustered using the Paired-End Cluster Generation Kit v4 following the Illumina PE_amplification_Linearization_Blocking_PrimerHyb_v7 recipe. Following the clustering procedure, the flow cell was loaded onto the Illumina HiSeq 2000 instrument following manufacturer's instructions. The sequencing chemistry used was v4 (FC-104-4001, Illumina) using SCS 2.6 and RTA 1.6 software with the 2 × 100 cycle, paired-end, indexed protocol. The Illumina base call files were processed using the GERALD pipeline to produce paired sequence files containing reads for each sample in the Illumina FASTQ format.

Transcriptome Assembly and SNP Calling

A set of predicted cDNA was obtained following the workflow of Cahais et al. (2012). Contigs shorter than 200 base pairs were discarded. Open reading frames (ORFs) were predicted using the script transcripts_to_best_scoring_ORFs.pl, which is part of the Trinity package (see a summary of the contig assembly in table 2). The reads were mapped to the predicted cDNA using BWA (Li and Durbin, 2009). More specifically, we used BWA ALN for read mapping with the

Table 2
Thymus vulgaris Transcriptome Assembly Characteristics

Contigs Set	*n*	N50 (bp)	Mean Length (bp)
Filtered contigs	1,11,942	1,004	681
Contigs containing ORFs	51,615	1,074	710
Homology to mitochondrion	131	756	562
Homology to chloroplast	75	852	597

ORF: open reading frames; N50: median size of contigs.

command line bwa aln -n 0.04 -o 1 -e -1 -d 16 -i 5 -k 1 -t 4 -M 10 -O 11 -E 4.

For each position of each contig and each individual, genotypes were called using the reads2snps method developed by Tsagkogeorga et al. (2012) and Gayral et al. (2013). This method includes a module for detecting hidden paralogy, which in the case of RNA-seq data manifests via an excess of heterozygous genotypes and positive across-individual correlation in read counts (Gayral et al., 2013). Positions detected as potentially paralogous by this method ($P < 0.01$) were coded as missing data and disregarded in further analysis.

Contig Annotation

We used a homology search to identify the contigs located in cytoplasmic genomes (i.e., chloroplast or mitochondrial genomes). We used two reference sequences from close relatives to thyme: the *Origanum vulgare* chloroplast genome (114 genes, NCBI accession number JX880022, Lukas and Novak, 2013) and *Salvia miltiorrhiza* mitochondrial genome (167 genes including 138 protein coding genes, NCBI accession number NC_023209.1, Qian et al., 2013). A blastn with default parameters and an identity threshold set to 95% was applied to the ORF-containing subset of our de novo assembled contigs using the two reference sequences as search databases. Contigs matching any of these reference sequences were annotated as cytoplasmic contigs and analyzed separately for SNP calling. This was done by using the reads2snps program (Gayral et al., 2013) with the default options except for the minimum required coverage to call a genotype, which was set to five instead of ten per position per individual because there was no risk of confusion between a sequencing error and heterozygosity, as the organelle genomes are haploid.

For similar reasons, we discarded cytoplasmic contigs for which the estimated F_{is} was below 0.95. As no heterozygosity is expected in organelle genes, contigs showing F_{is} values much below one must either correspond to wrongly annotated nuclear genes or to duplicated genes collapsed into a single contig during assembly. Filtering out undesired contigs based on heterozygosity allowed us to use a diploid genotyper in this analysis. Following Gayral et al. (2013), we discarded contigs with more than half the bases undetermined and with stop or frameshift codons >20 bases away from the start or end of the sequence. We then used custom-made scripts to merge the cytoplasmic contigs into a concatenated sequence, with an option to filter out contigs with traces of heterozygosity.

We also examined whether any contigs in our transcriptome displayed homology to the pentatricopeptide repeat gene family that contains known nuclear restorers (Kotchoni et al., 2010; Touzet and Meyer, 2014) but no such contigs were found (data not shown).

Blastn and blast2go (Conesa et al., 2005; Bethesda, 2008) were used on the fasta sequences of the identified candidate contigs to determine their putative functions, along with web searches on the returned results in Genbank.

Analysis of Nucleotide Diversity Patterns

We calculated the nucleotide diversity at nonsynonymous sites (pi_N), synonymous sites (pi_S) and T*D* (Tajima, 1989) on all nuclear contigs and the cytoplasmic super locus. The pi_N/pi_S ratio can be used as an integrative measure of the strength of purifying selection on a genome. The pi_N and pi_S measures were calculated for each contig using custom-written C ++ programs (Gayral et al., 2013). The T*D* statistics were obtained via a custom-made program, and the calculated values were checked against those obtained using DnaSP v5 (Librado and Rozas, 2009). For the contigs attributed to the cytoplasmic genome, we also computed the pi_N/pi_S ratio and the T*D* of the concatenated cytoplasmic "supercontig," as these are best viewed as a single nonrecombining unit.

We used both linear regression and robust locally weighted (lowess) regression (Cleveland, 1979) to investigate the covariation patterns between gene expression levels and nucleotide diversity. All statistical analyses were performed in R (R Core Team, 2013).

Results

Overview of the Contig Assembly and Amounts of Nucleotide Diversity

The contig assembly we obtained de novo from our RNA-seq data comprised 111,942 contigs after quality control filtering (table 2). Among these, 22,278 contigs included an ORF and were sufficiently covered for genotype and SNP calling.

Although the sample was obtained at a fairly local geographic scale, we detected ample nucleotide variation. Using pi_S to measure the amount of nucleotide polymorphism, we observed an average nucleotide diversity (mean $pi_S = 0.019$, median $pi_S = 0.016$; fig. 2) in the higher range of values recently reported in comparable population genomics studies based on transcriptome resequencing (Chen et al., 2017). The amount of synonymous diversity detected throughout the *T. vulgaris* genome therefore offers ample statistical power to detect and analyze footprints of selection. The nuclear contigs show a slight deficit of heterozygotes relative to Hardy–Weinberg expectations, as measured by F_{is} (table 3).

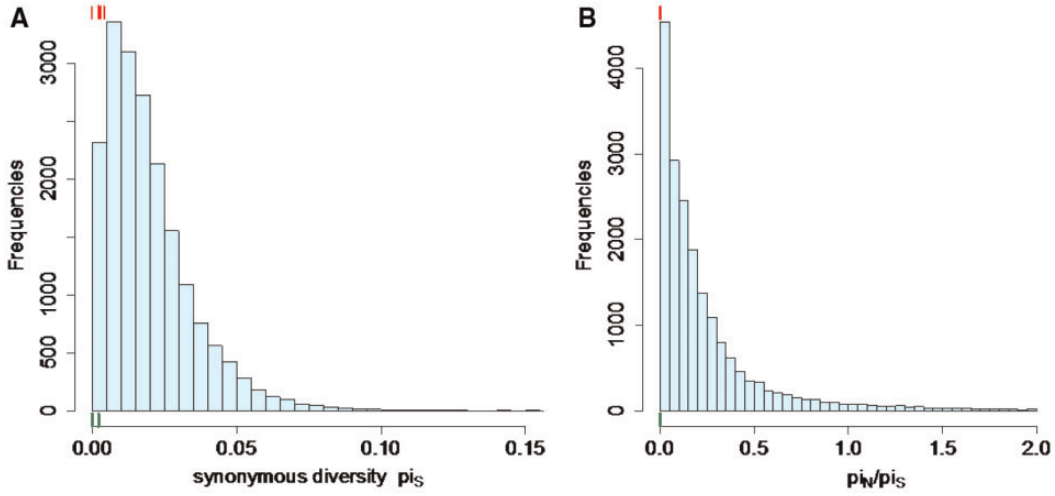

Fig. 2.—Genome-wide distribution of summary statistics of nucleotide diversity. Histograms denote genome-wide distributions of nuclear contigs. Individual ticks mark the values for contigs in the mitochondrial (top red) and chloroplast (bottom green) genomes. (*a*) Nucleotide diversity at synonymous positions (pi_S). (*b*) Distribution of the ratios of nucleotide diversity at nonsynonymous and synonymous sites (pi_N/pi_S).

Table 3

Mean and SEs of Nucleotide Diversity Statistics in Nuclear, Chloroplastic, and Mitochondrial Contigs

Genome	*n*	pi_N	pi_S	F_{is}	Tajima's *D*
Nuclear	18909	0.0031 (<0.0001)	0.0197 (<0.0001)	0.018 (<0.0001)	0.30 (0.009)
Mitochondrial	26	0.0005 (0.0007)	0.0005 (0.0002)	n.a.[a]	−1.93[b]
Chloroplastic	6	0.0007 (0.0004)	0.0004 (0.0002)	n.a.[a]	−1.80[b]

Note.—*n* denotes the number of polymorphic contigs per genome.
[a]F_{is} is only computed for nuclear contigs.
[b]we report a single value for the concatenated contigs of each organelle (no SE reported).

This finding is consistent with an outcrossing mating system, with a weak degree of genetic differentiation between the sample populations creating a deficit of heterozygotes.

Within our assembly, we detected 65 contigs homologous to the chloroplast genome reference and 90 homologous to the mitochondrial genome reference. Among these, 21 contigs displayed homology to both the chloroplast and mitochondrial reference genomes.

Although it might seem surprising that a sizeable fraction of contigs had homology to both the mitochondrial and chloroplast reference genomes, Veronico et al. (1996) and Cummings et al. (2003) report that there is insertion of chloroplast genes in mitochondrial genomes, even if the opposite case has not been observed. As a check, we downloaded the mitochondrial and chloroplast genomes for five plant species with well-annotated published genomes (*Arabidopsis thaliana*, *Medicago truncatula*, *Silene latifolia*, *Oryza sativa*, and *Zea mays*). Nucleotide blasts on the complete mitochondrial and chloroplast genomes revealed that shared paralogs between chloroplasts and mitochondria are not uncommon (table 4).

We used the four-gamete test (Hudson and Kaplan, 1985) and checked that there was no evidence of recombination in either the chloroplast or mitochondrial contigs (supplementary fig. 4, Supplementary Material online). Although this test might miss a fraction of the recombining events (Meyers and Griffith, 2003), these are unlikely to drastically affect the coalescent genealogies underlying the observed diversity patterns, otherwise they would have been revealed via the four-gamete test.

When comparing nucleotide diversity in cytoplasmic versus nuclear contigs (table 3), we note that nucleotide diversity is substantially reduced in mitochondrial and chloroplast contigs relative to nuclear ones (fig. 2*a* and table 3).

Footprints of Selection in Cytoplasmic Genomes Suggest a Recent Sweep in the Coalescence History of Our Sample

To set an expectation for T*D* under neutrality, we used the set of nuclear contigs containing a polymorphic coding sequence (CDS) as the neutral genomic background. This background distribution of T*D*, which incorporates possible biasing effects of demographic history, was slightly skewed relative to neutral expectations in a stable nonsubdivided population (mean = 0.17; skewness = 0.93; fig. 3).

We computed a single T*D* value per organelle by separately concatenating the sequence of contigs mapping to

Table 4
Homology Between Chloroplasts and Mitochondria for Five Plant Species

Species	Genbank Reference Chloroplast/Mitochondrion	Length Chloroplast/ Mitochondrion and Ratio	Chloroplast ⇒ Mitochondrion[a] (%)	Mitochondrion ⇒ Chloroplast[a] (%)
Arabidopsis thaliana	NC_000932/NC_001284.2	154478/366924 = 0.4	4–5	1
Medicago truncatula	NC_003119/NC_029641.1	124033/271618 = 0.45	1	<1
Silene latifolia	NC_016730.1/HM562727.1	151736/253413 = 0.59	1	<1
Oryza sativa japonica	KT289404.1/BA000029.3	134536/490520 = 0.27	18	6
Zea mays	NC_001666/NC_007982.1	140384/569630 = 0.24	29	4

[a]Reciprocal homology searches using mitochondrial or chloroplast sequences as Query/Database were performed to assess the amount of genes that shared exhibited shared homology between these two genomes.

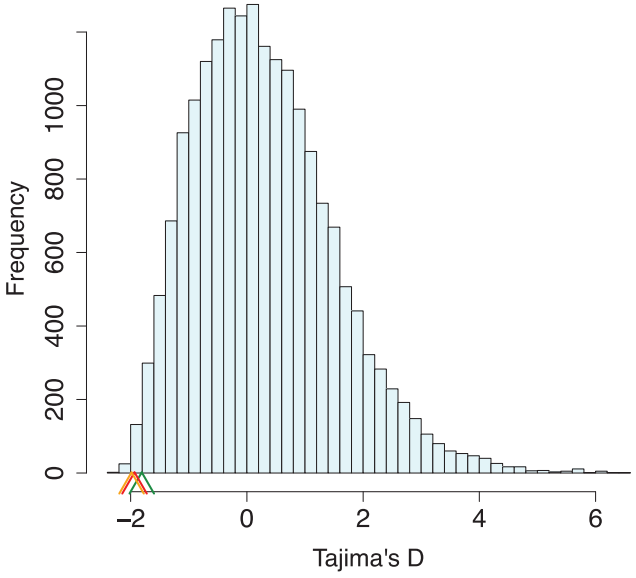

Fig. 3—Genome-wide distribution of Tajima's *D* statistic in the nuclear (blue histogram) genome versus cytoplasmic contigs (red: mitochondrial genome; green: cytoplasmic genome; orange: supercontig). Tajima's *D* for the chloroplasts lies in the 0.0089th quantile of most negative values observed in the nuclear genome. Tajima's *D* for the mitochondria lies in the 0.0027th quantile of most negative values.

mitochondria and to chloroplasts. We also computed T*D* on a "supercontig" containing all 32 contigs assigned to organelles, and eliminated redundancies between the two organelles. The chloroplast and mitochondrial genomes exhibited very negative T*D* values (values for concatenated contigs: mitochondria: −1.93; chloroplasts: −1.80; and supercontig: −1.98), which corresponded to the 0.8%, 0.27%, and 0.19% most extreme negative values, respectively, observed in the nuclear background distribution (fig. 3). The pi_N/pi_S ratios are markedly lower in the nuclear genome than in the cytoplasmic genome, but the sampling variance attached to the cytoplasmic estimates is considerable due to the low polymorphic contig numbers in these genomes (table 3). According to our predictions (table 1), these results support scenario A, which involves the occurrence of a recent selective sweep.

The diversity patterns we report here were obtained from transcriptome sequencing from a sample of genes that are biased towards those with moderate to strong expression levels in leaves. As biasing our sample towards more highly expressed genes might also bias the genome-wide distribution of T*D*, we addressed this concern by examining how much the expression levels of contigs covaried with the summary statistics we used. We detected a slight tendency for highly expressed genes to have stronger purifying selection as measured by pi_N/pi_S: a linear regression of log(expression) explains 5% of the total variation, but a local regression reveals little systematic trend (supplementary figs. 1 and 2, Supplementary Material online). We note that the differences in level of gene expression weakly covaried with T*D*: a linear regression model with log(expression) explained 4.8% of the total variation in T*D*, and a local robust regression suggested no discernable trend (supplementary fig. 3, Supplementary Material online). This observation and the strikingly similar T*D* values for mitochondria and chloroplasts (fig. 3, Supplementary Material online) support our hypothesis that both cytoplasmic genomes can be viewed as a single non-recombining entity.

Discussion

Diversity Patterns Are Compatible with a Recent Selective Sweep in Cytoplasmic Genomes

The levels of nucleotide diversity, as measured by pi_S, are comparable in the mitochondrial and chloroplast genomes and low compared with the diversity observed in nuclear contigs (fig. 2*a*). This pattern is consistent with selection affecting the cytoplasmic genomes and eliminating nucleotide diversity. This observation could also be explained by a systematically lower per-nucleotide mutation rate of cytoplasmic genes relative to nuclear ones. However, evidence for such a mechanism is indirect, and relies on the difference in substitution rates (Mower et al., 2007; Drouin et al., 2008); in some cases the inferred mutation rates in mitochondria are potentially high (Sloan et al., 2012).

Both cytoplasmic genomes also exhibit significantly negative T*D* compared with the nuclear genome values (fig. 3), which is incompatible with the positive T*D* predicted by

scenario B (maintenance of CMS alleles via balancing selection), but consistent with scenario A, in which a new CMS lineage has recently swept to fixation. Under "balancing selection" we expect at least two distinct CMS lineages to segregate at intermediate frequencies in both local and global (species-wide) populations. Using 20% as a threshold for the frequency of the least common CMS lineage maintained by balancing selection, the probability of missing a second CMS lineage in a sample of 18 individuals is only 0.818∼0.02. Our sampling scheme, and the fact that sufficient nucleotide diversity is available, ensures that we have adequate power to distinguish between scenarios.

Population subdivision, range expansion, and sampling all influence the SFS and *TD* values (Städler et al., 2009). Our sample probably does not comply with the assumption of a single nonsubdivided population. Indeed, the empirical distribution of *TD* values for the nuclear genome exhibits a slightly positive mean and a positive skew (fig. 3). A bottleneck followed by a population expansion generates an excess of rare variants (relative to a stable population), which will also yield more negative *TD* values than expected (Tajima, 1989; Städler et al., 2009). However, demographic history is unlikely to explain the negative *TD* in the cytoplasmic contigs, as it should affect nucleotide diversity in all contigs (both nuclear and cytoplasmic). This is clearly not the case as the mean *TD* in the nuclear contigs is slightly positive (table 3). Molecular variation patterns resulting from balancing selection should only be weakly affected by population structure; old alleles maintained by balancing selection can survive speciation and result in trans-species polymorphism (Delph and Kelly, 2014). Even in the presence of subdivided populations, scenario B would still predict positive, not negative, *TD* values in the cytoplasmic genomes. Therefore, only scenario A can account for the patterns in the data, perhaps complicated by an unknown regional population structure.

Relative to the nuclear genome, nucleotide diversity in both chloroplasts and mitochondria are more likely to be affected by linked selection (e.g., selective sweeps or background selection). Theoretical studies have shown that while these processes can reduce neutral diversity, *TD* is relatively insensitive to background selection (Charlesworth et al., 1995). Therefore, stronger background selection in cytoplasmic genes relative to nuclear genes might contribute to the differences in pi_S, but is unlikely to account for the strongly negative *TD* values observed in the cytoplasmic contigs relative to the nuclear genome. Few recent studies have extensively compared nucleotide diversity in nuclear versus cytoplasmic plant genomes, yet a study in two sister species of *Arabidopsis* found no such traces of cytoplasmic selection (Wright et al., 2008).

Evidence for Selective Sweeps versus Balancing Selection in Other Species

Although our data refute the scenario of long-term balancing selection in cytoplasmic genes, they cannot prove that the detected selective sweep signal is uniquely driven by the dynamics of CMS genes in maintaining gynodioecy and is not instead due to cytoplasmic genes involved in other types of adaptation. We hence review the empirical evidence for selective sweeps or balancing selection in other plant species where CMS is known.

The theory predicts that when a CMS gene sweeps to fixation together with the nuclear allele that restores male function, the population returns to being composed of hermaphrodites until a new CMS gene enters the populations via mutation or migration (Frank, 1989). These hermaphrodites are expected to carry "hidden" CMS genes (termed "hidden" because individuals harboring these CMS genes also carry a nuclear restorer), which can be uncovered in crosses with individuals having a different nuclear background. Empirical studies have shown examples of hidden CMS genes in a number of crop (e.g., maize, rice, and sunflower) and noncrop (e.g., *Nemophila menziesii*) species (Schnable and Wise, 1998; Barr, 2004; Chen and Liu, 2014). The expected footprint of the CMS genes in these examples is that of a selective sweep, similar to what we have documented.

In previous studies of the nucleotide diversity of natural populations of *Beta vulgaris* and several species of *Silene* to infer the mechanisms underlying the maintenance of gynodioecy, some find evidence of balancing selection (reviewed in Delph and Kelly, 2014) but one reports evidence for recurrent sweeps in *Silene* (Ingvarsson and Taylor, 2002). Virtually all published studies use only a few (5–10) gene fragments from mitochondrial (e.g., Städler and Delph, 2002) or chloroplast genomes (Fénart et al., 2006), but never use nuclear diversity as a control. One exception is Lahiani et al. (2013), who reported the *TD* for four nuclear fragments, two mitochondrial gene fragments, and four chloroplast fragments in *Silene*. No consistent trend towards a negative *TD* was detected in the cytoplasmic fragments.

Connection with Previous Studies on Cytoplasmic Diversity in *Thymus vulgaris*

Cytoplasmic male sterility in natural *T. vulgaris* populations has been previously studied using plants originating from the same geographic region as those used in the present study. These studies relied on intra and interpopulation crosses and backcrosses, and the genotyping of individuals from these crosses using restriction fragment length polymorphism in mitochondrial DNA. These investigations document the existence of several different mitochondrial molecular types. Although this is not in itself evidence of an equally high number of functionally different CMS lineages, these results suggest the presence of multiple CMS lineages in the region, not just across but also within populations (Belhassen et al., 1991, 1993; Manicacci et al., 1996).

It is worth considering what patterns of nucleotide diversity would be expected if multiple CMS lineages were present. If two CMS lineages were present in the sample of 18 individuals and both coexisted for sufficient generations (in a coalescence timescale), the underlying genealogy would exhibit long internal branches separating the two CMS lineages. This situation would generate a sizeable amount of polymorphism at intermediate frequencies, generating a positive T*D*. Instead, our data is compatible with a recent sweep of cytoplasmic genomes.

An intriguing but theoretically plausible explanation is that because mutations in numerous cytoplasmic genes can cause CMS, different CMS lineages could be created on the backbone of a recently swept cytoplasmic lineage (scenario 7 in Frank, 1989). This idea remains speculative as it assumes a high cytoplasmic mutation rate for a new CMS. Cytoplasmic diversity data does not allow any inference of the age and strength of a sweep, so we cannot infer, even crudely, the number of CMS types that may have survived the selective sweep we have detected.

If truly distinct CMS lineages exist in the populations (Manicacci et al., 1996), we predict that many types should be recently derived via new mutations from a previous (possibly restored) type. Further studies should explore the extent to which this selective sweep pattern is also found on a wider geographic scale in *T. vulgaris* and in other species of the genus *Thymus*.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Xavier Vekemans for discussion and Matthew Hartfield, Susanne Renner, and two anonymous reviewers for comments on the manuscript. T.B. acknowledges financial support from the European Research Council under the European Union's Seventh Framework Program (FP7/20072013, ERC Grant 311341).

Author Contributions

T.B., B.K.E., and T.L. designed the research; SS collected the data; T.B., M.M., N.G., and E.F. performed the data analysis; T.B., B.K.E., T.L., S.M., and M.M. participated in data interpretation; and T.B., B.K.E., S.M., and M.M. wrote the manuscript with input from all coauthors.

Literature Cited

Bailey MF. 2002. A cost of restoration of male fertility in a gynodioecious species *Lobelia siphilitica*. Evolution 56(11):2178–2186.

Bailey MF, Delph LF, Lively CM. 2003. Modeling gynodioecy: novel scenarios for maintaining polymorphism. Am Nat. 161(5):762–776.

Barr CM. 2004. Hybridization and regional sex ratios in *Nemophila menziesii*. J Evol Biol. 17(4):786–794.

Barr CM, Neiman M, Taylor DR. 2005. Inheritance and recombination of mitochondrial genomes in plants, fungi and animals. New Phytol. 168(1):39–50.

Belhassen E, Trabaud L, Couvet D, Gouyon P-H. 1989. An example of nonequilibrium processes: the gynodioecy of *Thymus vulgaris* (L.) in burned habitats. Evolution 43(3):662–667.

Belhassen E, et al. 1990. Evolution des taux de femelles dans les populations naturelles de thym, *Thymus vulgaris* L: Deux hypothèses alternatives confirmées. CR Acad Sci Paris 310:371–375.

Belhassen E, et al. 1991. Complex determination of male sterility in *Thymus vulgaris* L.: genetic and molecular analysis. Theor Appl Genet. 82(2):137–143.

Belhassen E, Atlan A, Couvet D, Gouyon P-H, Quétier F. 1993. Mitochondrial genome of *Thymus vulgaris* L. (Labiate) is highly polymorphic between and among natural populations. Heredity 71(5):462–472.

Bethesda MD. 2008. BLAST Command Line Applications User Manual [Internet], National Center for Biotechnology Information (US); Available from: https://www.ncbi.nlm.nih.gov/books/NBK279690/.

Cahais V, et al. 2012. Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol Ecol Resour. 12(5):834–845.

Charlesworth D. 1981. A further study of the problem of the maintenance of females in gynodioecious species. Heredity 46(1):27–39.

Charlesworth B, Charlesworth D. 1978. A model for the evolution of dioecy and gynodioecy. Am Nat. 112(988):975–997.

Charlesworth D, Charlesworth B, Morgan MT. 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141(4):1619–1632.

Charlesworth D, Laporte V. 1998. The male-sterility polymorphism of *Silene vulgaris*: analysis of genetic data from two populations and comparison with *Thymus vulgaris*. Genetics 150(3):1267–1282.

Charlesworth D. 2002. What maintains male-sterility factors in plant populations?. Heredity 89(6):408–409.

Chen L, Liu YG. 2014. Male sterility and fertility restoration in crops. Annu Rev Plant Biol. 65:579–606.

Chen J, Glemin S, Lascoux M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. Mol Biol Evol. doi: 10.1093/molbev/msx088.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc. 74(368):829–836.

Conesa A, et al. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676.

Couvet D, et al. 1990. Co-evolution between two symbionts: the case of cytoplasmic male-sterility in higher plants. In: Futuyma, D., Antonovics, J, editors. Oxford surveys in evolutionary biology. Oxford: Oxford University Press. p. 225–247.

Couvet D, Ronce O, Gliddon C. 1998. Maintenance of nucleo-cytoplasmic polymorphism in a metapopulation: the case of gynodioecy. Am Nat. 152(1):59–70.

Cummings MP, Nugent JM, Olmstead RG, Palmer JD. 2003. Phylogenetic analysis reveals five independent transfers of the chloroplast gene rbcL to the mitochondrial genome in angiosperms. Curr Genet. 43(2):131–138.

Davila JI, et al. 2011. Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biology 9:64.

De Haan AA, Hundscheid MP, van Hinsberg A. 1997. Effects of CMS types and restorer alleles on plant performance in *Plantago lanceolata* L.: an indication for the cost of restoration. J Evol Biol. 10(5):803–820.

Del Castillo RF, Trujillo S. 2009. Evidence of restoration cost in the annual gynodioecious *Phacelia dubia*. J Evol Biol. 22(2):306–313.
Delph LF, Kelly JK. 2014. On the importance of balancing selection in plants. New Phytol. 201(1):45–56.
Desfeux C, Maurice S, Henry JP, Lejeune B, Gouyon PH. 1996. Evolution of reproductive systems in the genus *Silene*. Proc Biol Sci. 263(1369):409–414.
Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol. 49(3):827–831.
Dufaÿ M, Touzet P, Maurice S, Cugen J. 2007. Modelling the maintenance of male-fertile cytoplasm in a gynodioecious population. Heredity 99(3):349–356.
Dufay M, et al. 2014. An angiosperm-wide analysis of the gynodioecy-dioecy pathway. Ann Bot. 114(3):539–548.
Ehlers BK, Maurice S, Bataillon T. 2005. Sex inheritance in gynodioecious species: a polygenic view. ProcBiol Sci. 272(1574):1795–1802.
Frank SA. 1989. The evolutionary dynamics of cytoplasmic male sterility. Am Nat. 133(3):345–376.
Fénart S, Touzet P, Arnaud J-F, Cuguen J. 2006. Emergence of gynodioecy in wild beet (*Beta vulgaris ssp. maritima* L.): a genealogical approach using chloroplastic nucleotide sequences. Proc Biol Sci. 273(1592):1391–1398.
Gayral P, et al. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet. 9(4):e1003457.
Gigord L, Lavigne C, Shykoff JA, Atlan A. 1999. Evidence for effects of restorer genes on male and female reproductive functions of hermaphrodites in the gynodioecious species *Thymus vulgaris* L. J Evol Biol. 12:596–604.
Gouyon PH, Vichot F, Van Damme JMM. 1991. Nuclear-cytoplasmic male sterility: single-point equilibria versus limit cycles. Am Nat. 137(4):498–514.
Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111(1):147–164.
Hurst LD, Atlan A, Bengtsson BO. 1996. Genetic conflicts. Q Rev Biol. 71(3):317–364.
Ingvarsson PK, Taylor DR. 2002. Genealogical evidence for epidemics of selfish genes. Proc Natl Acad Sci U S A. 99(17):1265–11269.
Kotchoni SO, Jimenez-Lopez JC, Gachomo EW, Seufferheld MJ, Pastore A. 2010. A new and unified nomenclature for male fertility restorer (RF) proteins in higher plants. PLoS One 5(12):e15906.
Lahiani E, et al. 2013. Disentangling the effects of mating systems and mutation rates on cytoplasmic diversity in gynodioecious *Silene nutans* and dioecious *Silene otites*. Heredity 111(2):157–164.
Lewis D. 1941. Male sterility in natural populations of hermaphrodite plants the equilibrium between females and hermaphrodites to be expected with different types of inheritance. New Phytol. 40(1):56–63.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.
Librado P, Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25(11):1451–1452.
Lukas B, Novak J. 2013. The complete chloroplast genome of *Origanum vulgare* L. (Lamiaceae). Gene 528(2):163–169.
Manicacci D, Couvet D, Belhassen E, Gouyon PH, Atlan A. 1996. Founder effects and sex ratio in the gynodioecious *Thymus vulgaris* L. Mol Ecol. 5(1):63–72.
Manicacci D, Atlan A, Rossello JAE, Couvet D. 1998. Gynodioecy and reproductive trait variation in three *Thymus* species (Lamiaceae). Int J Plant Sci. 159(6):948–957.
McCauley DE, Olson MS. 2008. Do recent findings in plant mitochondrial molecular and population genetics have implications for the study of gynodioecy and cytonuclear conflict?. Evolution 62(5):1013–1025.
Meyers S, Griffith R. 2003. Bounds on the minimum number of recombination events in a sample history. Genetics 163(1):375–394.
Mower JP, et al. 2007. Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol. 7:135.
Qian J, et al. 2013. The complete chloroplast genome sequence of the medicinal plant *Salvia miltiorrhiza*. PLoS One 8(2):e57607.
R Core Team. 2013. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: http://www.R-project.org/.
Renner SS. 2014. The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. Am. J Bot. 101(10):1588–1596.
Richards AJ. 1997. Plant breeding systems. London: Chapman and Hall.
Schnable PS, Wise RP. 1998. The molecular basis of cytoplasmic male sterility and fertility restoration. Trends Plant Sci. 3(5):175–180.
Sloan DB, et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol. 10(1):e1001241.
Städler T, Delph LF. 2002. Ancient mitochondrial haplotypes and evidence for intragenic recombination in a gynodioecious plant. Proc Natl Acad Sci U S A. 99:11730–11735.
Städler T, Haubold B, Merino C, Stephan W, Pfaffelhuber P. 2009. The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. Genetics 182(1):205–216.
Tajima F. 1989. Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595.
Thompson JD. 2002. Population structure and the spatial dynamics of genetic polymorphism in Thyme. In: Stahl-Biskup E, Saez F, editors. Thyme: the genus Thymus. London: Taylor Francis.
Touzet P, Meyer EH. 2014. Cytoplasmic male sterility and mitochondrial metabolism in plants. Mitochondrion 19:166–171.
Tsagkogeorga G, Cahais V, Galtier G. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint and molecular adaptation in the tunicate *Ciona intestinalis*. Genome Biol Evol. 4(8):852–861.
Veronico P, Gallerani R, Ceci LR. 1996. Compilation and classification of higher plant mitochondrial tRNA genes. Nucleic Acids Research 24(12):2199–2203.
Wright SI, Nano N, Foxe JP, Dar V-UN. 2008. Effective population size and tests of neutrality at cytoplasmic genes in *Arabidopsis*. Genet Res. 90(1):119–128.

Associate editor: Susanne Renner