Population genomics of the facultatively asexual duckweed *Spirodela polyrhiza*

Eddie Ho*[1], Magdalena Bartkowska*[1,2], Stephen I. Wright†[1,3], Aneil Agrawal†[1,3]

1 Department of Ecology and Evolutionary Biology, University of Toronto 25 Willcocks St. Toronto ON M5S 3B2
2 Current address: Raven Telemetry Inc. Ottawa, ON
3 Center for Analysis of Genome Evolution and Function, University of Toronto 25 Willcocks St. Toronto ON M5S 3B2

* these authors contributed equally to the work

† corresponding authors contributed equally to the work; emails: stephen.wright@utoronto.ca, a.agrawal@utoronto.ca

Summary

- Clonal propagation allows some plant species to achieve massive population sizes quickly but also reduces the evolutionary independence of different sites in the genome.
- We examine genome-wide genetic diversity in *Spirodela polyrhiza*, a duckweed that reproduces primarily asexually.
- We find that this geographically widespread and numerically abundant species has very low levels of genetic diversity. Diversity at nonsynonymous sites relative to synonymous sites is high, suggesting that purifying selection is weak. A potential explanation for this observation is that a very low frequency of sex renders selection in effective. However, there is a pronounced decay in linkage disequilibrium over 40 kb, suggesting that though sex may be rare at the individual level it is not too infrequent at the population level. In addition, neutral diversity is affected by the physical proximity of selected sites, which would be unexpected if sex was exceedingly rare at the population level.
- The amount of genetic mixing as assessed by the decay in linkage disequilibrium is not dissimilar from selfing species such as *Arabidopsis thaliana*, yet selection appears to be much less effective in duckweed. We discuss alternative explanations for the signature of weak purifying selection.

Keywords: asexual reproduction, purifying selection, recombination, genetic diversity

44

Introduction

The majority of eukaryotes can reproduce through outcrossing, but uniparental reproduction through asexual reproduction or self-fertilization has arisen many times independently (Vrijenhoek, 1998; Barrett, 2002; Rice, 2002; Goodwillie *et al.*, 2005; Jarne & Auld, 2006; Whitton *et al.*, 2008). Given the genetic and ecological benefits of uniparental reproduction (Fisher, 1941; Maynard-Smith, 1978; Lloyd, 1979), the rarity of highly asexual and highly selfing species in nature is a long-standing problem in evolutionary biology (Otto, 2009). Reproductive systems can have dramatic consequences on the evolution of the genome and these consequences may be key to understanding the benefits of outcrossing over uniparental reproduction. One of the major consequences of selfing and asexuality is a reduction in the effective population size, N_e, because lower effective rates of recombination increase the strength of linked selection effects, such as background selection and selective sweeps (Golding & Strobeck, 1980; Hudson & Kaplan, 1995; Nordborg & Donnelly, 1997; Charlesworth *et al.*, 1997; Nordborg, 2000). In addition, species employing uniparental reproduction can colonize habitats with just a few individuals so population bottlenecks can further reduce local N_e.

While Asexuality and selfing are both modes of uniparental reproduction, they differ in their effects on allelic diversity within individuals. Selfing reduces diversity by homogenizing homologous alleles within individuals. Asexuality can increase diversity within diploid individuals because there is less opportunity for recombination and segregation. This effect of alleles diverging within individuals is called the 'Mesleson effect' and can appear as an excess in heterozygosity (Mark Welch & Meselson, 2000; Butlin, 2002; Ament-Velásquez *et al.*, 2016). However, asexual reproduction reduces genotypic diversity within populations (Balloux *et al.*, 2003), leading to fewer distinct genotypes despite the high allelic diversity. Furthermore, in the presence of mitotic gene conversion, the effect of increased coalescence time due to clonality can be weakened or reversed, driving a reduction in nucleotide diversity (Hartfield *et al.*, 2016). Background selection will also drive strong reductions in effective population size in highly asexual lineages, an effect that can be stronger than observed for highly selfing organisms (Agrawal & Hartfield, 2016).

A decline in N_e is predicted to reduce neutral variation and lower the efficacy of selection such that deleterious mutations are more likely to accumulate and beneficial mutations are less likely to establish (Muller, 1932; Kimura, 1962; Heller & Maynard Smith, 1978; Charlesworth *et al.*, 1993); (Charlesworth & Charlesworth, 1997; Orr, 2000; Glémin & Ronfort, 2013; Kamran-Disfani & Agrawal, 2014) . Empirical tests of these predictions increasingly rely on obtaining

genomic sequences from selfing and asexual species (Wright *et al.*, 2002; Cutter, 2006; Paland & Lynch, 2006; Slotte *et al.*, 2010; Qiu *et al.*, 2011; Ness *et al.*, 2012; Brandvain *et al.*, 2013; Kamran-Disfani & Agrawal, 2014; Barraclough *et al.*, 2007; Escobar *et al.*, 2010; Tucker *et al.*, 2013; Arunkumar *et al.*, 2015; Laenen *et al.*, 2017). There is strong evidence that self-fertilization reduces neutral diversity (Schoen & Brown, 1991; Glémin *et al.*, 2006). Furthermore, highly selfing species suffer from a reduced efficacy of purifying selection as assessed from polymorphism and codon usage data, at least in some cases, though this effect is weak to non-existent when examining fixed differences between species (Glémin *et al.*, 2006; Wright *et al.*, 2002; Glémin & Galtier, 2012; Burgarella *et al.*, 2015). There may be multiple reasons for these contradicting results such as low sampling or the young age of highly selfing species (Wright *et al.*, 2002; Glémin & Galtier, 2012). In contrast, evidence for weakened purifying selection in asexual species is mostly derived from fixation data (Glémin & Muyle, 2014). Not only are there fewer studies testing the population genomic consequences of asexuality, most studies utilize mitochondrial sequence and few have used polymorphism data to test for relaxed selection (Hartfield, 2016). Given that approximately 80% of angiosperms undergo some form of asexual reproduction (Barrett, 2015; Hartfield, 2016), there is ample opportunity to study the consequences of asexuality.

Here, we analyze genomic sequences of the highly asexual aquatic duckweed, *Spirodela polyrhiza*. We are broadly interested in the population genomic consequences of the highly asexual lifestyle that characterize this duckweed. *S. polyrhiza*, the greater duckweed, is part of the Araceae family (subfamily Lemnoideae) containing 37 or 38 species of duckweed (Landolt, 1986; Les *et al.*, 2002). This aquatic angiosperm is cosmopolitan, excluding polar regions. It is thought to propagate almost exclusively through clonal reproduction (Landolt, 1986) though there are no robust estimates of sexual reproduction. Flowering was observed in 5% of samples in one survey (Hicks, 1932). We know little about *S. polyrhiza* population genomics in general. Many studies of duckweed genetics focus on species differentiation using different regions of the genome, such as AFLPs (Bog *et al.*, 2010), ISSR (Xue *et al.*, 2012) and chloroplast regions (Jordan *et al.*, 1996; Tang *et al.*, 2014; Xu *et al.*, 2015). However, most of these studies find little to no diversity within species. Crawford and Landolt (1993) examined 16 allozyme loci among 67 samples of *S. polyrhiza* isolated from Africa, Asia, Australia, Europe and North America. *S. polyrhiza* possessed a lower level of diversity than *S. intermedia* and *S. punctata*, which only had 21 and 43 isolates, respectively. They suggest that the low diversity in *S. polyrhiza* may be due to its lower frequency of seed production compared to the other two

Spirodela species. A new population genomic study parallel to ours reports very low levels of polymorphism within *S. polyrhiza* (Xu et al., 2018).

We examine whole genome short-read sequence data from 36 *S. polyrhiza* samples. We examine patterns of diversity at different site types, how diversity varies across the genome, and patterns of linkage disequilibrium. Diversity is exceptionally low in *S. polyrhiza*, purifying selection appears to be very weak, and linkage disequilibrium extends over long distances but shows clear signs of decay. The results are consistent with sex being very low at the individual level but not too infrequent at the species level.

Materials and Methods

Population samples

Populations of *S. polyrhiza* were haphazardly chosen to represent the species' genetic diversity across North America (Table S1). Of 38 samples, 26 were isolates collected from the wild across Canada and the United States. An additional 12 samples were obtained from the Rutgers University Stock Centre. Nine of the Rutgers isolates were derived from North American populations (seven from USA and two from Mexico) and the remaining three samples represent global diversity (one each from Colombia, India and France). Two of the Rutgers University samples (Texas-RU412 and Colombia-RU415) were excluded from all analyses because less than 55% of their sequenced reads mapped to the reference genome and both samples were extreme outliers in genotype distance to all other samples.

Laboratory culturing and sequencing

All samples were established in laboratory cultures that were derived from single fronds. The cultures were grown for several months in axenic conditions at the University of Toronto. Prior to DNA extraction, fronds were collected, washed under tap water and flash frozen in liquid nitrogen. DNA was extracted from frond tissue using a modified CTAB protocol (Lutz *et al.*, 2011). Library preparation (Illumina TruSeq with PCR) and paired-end genomic sequencing were conducted at the Genome Quebec Innovation Centre at McGill University on the Illumina HiSeq2000 PE100 platform. Sequencing of the samples was conducted across three lanes generating paired-end 100 bp long reads.

Genotyping

We used the Stampy aligner version 1.0.22 with default settings to align genomic reads to the *Spirodela polyrhiza* reference genome (Wang *et al.*, 2014). Genotyping was then performed using the Genome Analysis Toolkit (GATK) v. 2.7 GATK HaploytpeCaller using default parameters (DePristo *et al.*, 2011). The median genotype quality across all samples and sites was 29 (ranged from 16-42), and the median individual depth across all sites ranged from 5-17. We excluded two samples (Texas-RU412 and Colombia-RU415) from all analyses and called SNPs using 36 rather than 38 samples because these two (RU412 and RU415) had the lowest proportion of mapped reads, the lowest coverage and the greatest sequence divergence from all other samples.

Hard-filtering based on sample and site quality and depth

First, we applied sample-specific filters. Sample genotypes at a site were considered missing if the sample depth was less than 5 or more than 1.5-fold the median sample depth. For variant sites, where GATK provides a genotype quality (GQ) score for individual genotypes, we excluded all genotypes with GQ score less than 20. Next, we applied a series of site-level filters. Invariant and variant sites were excluded if: 1) fewer than 20 samples had a depth between 5 and 40 reads, 2) the average sample depth exceeded 18 or was below 10, and 3) fewer than 2/3 (24) of samples passed the sample-specific filters. Filtering for high depth was performed to avoid regions with paralogous read mapping; SNPs at sites with high average sample depth were more likely to have fixed heterozygote sites (where all samples were heterozygous). Variant sites were retained if at least 20 samples had a genotype quality (GQ) score at least 20 (which corresponds to a genotyping accuracy of 99) and if the mapping quality (MQ) of the site was at least 90. Finally, we filtered entire regions of the genome. In particular, we removed sites that were identified as transposable elements or highly repetitive along with 100 bp on either side of the repetitive element. We used the repeat-masked assembly available on JGI (Spolyrhiza_290_v2.repeatmasked_assembly_v1.gff3), which masked retroelements, as well as RepeatModeler (v. 1.0.8) coupled with RepeatMasker (v. 4.0.5) to identify transposable elements and highly repetitive regions. Indels were removed along with 5 bp on either side of the site; for a deletion we also removed all sites spanning the length of the deletion. 20 kb windows were removed if fewer than 40% of sites within the window failed to pass all other filters. This filter eliminated regions that tended to have abnormally high number of poorly mapped reads and tended to have clusters of highly variable sites near poorly assembled regions of the genome. For gene-level analyses, we also removed 1595 genes (out of the

19623 annotated genes) that had sites where the average sample depth was greater than 18 reads, to further filter out paralogous genes.

Genetic distance

We calculated pairwise genetic distance using two methods. The pairwise 'genotypic distance' was calculated by summing the number of sites between two samples that differed in genotype. Pairs of sites that are homozygous for different alleles (homozygous differences) are weighted twice as much as pair of sites where one sample is homozygous and the other samples is heterozygous (heterozygous differences) (Table S2). The pairwise 'allelic distance' is the probability that a randomly selected allele from two samples at a given site will be different. The two distance metrics differ only in how they score the distance between samples that are both heterozygotes: the genotypic distance is zero but the allelic distance is ½, as it is for homozygote vs. heterozygote comparison (Table S3). Comparisons of the two metrics are useful when there are high rates of asexual reproduction as there may be low genotypic diversity despite a retention of allelic variation. We used these pairwise genetic distances to construct neighbour joining trees using the Ape package (Paradis *et al.*, 2004; Paradis, 2011) in R (R Core Team, N, 2016).

Grouping samples into genets

Because *S. polyrhiza* reproduces asexually, it is very likely that samples from the same or nearby pond may be clones (i.e., descended from a common ancestor via only clonal reproduction, possibly over many generations). For most analyses, we grouped samples that were highly genotypically similar into one genet that was genetically distinct from other genets. Samples were grouped into genets as follows. For each pair of samples, we calculated the number of sites where one sample is heterozygous and the other is homozygous ("heterozygous differences") and the number of sites where one samples is homozygous for one allele and the other is homozygous for the other allele ("homozygous differences"). Clonal samples will not necessarily be completely identical because of genotyping error (though bioinformatics filtering should minimize this) and because of mutation and gene conversion. Nonetheless, clonal samples should be very similar. If two samples are related exclusively through clonal propagation, heterozygous differences can occur simply due to a point mutation in one of the lineages. With clonal reproduction, homozygous differences will be even rarer because they require two rare events, e.g., two point mutations occurring at the same site or two lineages to be separated by both mutation and a gene conversion event.

Based on an inspection of the data (see Results and Discussion), we clustered samples together if each pair within the cluster had ≤ 0.01% of sites with homozygous differences and ≤ 2% of sites with heterozygous differences, as we found these thresholds to form very distinct groups. Using this threshold, we found nine genotypes composed of multiple samples and three genotypes each represented by a single sample (Table S5). For each of the 12 genotypes (hereafter, genets) we created a 'consensus genotype' whereby the genotype of each site was randomly chosen among the samples that form the group. Pairwise genotypic distance between samples within genets was, on average, 35 times smaller than genotypic distances between samples from different genets (Table S4). We confirmed our grouping using k-means clustering (Adengenet package in R; Jombart & Ahmed, 2011). This procedure consists of running successive K-means with an increasing number of clusters (k), after transforming the data using a principal component analysis (PCA).

These analyses found that one genet (consisting of RU448 and RU99) was very divergent from the other genotypes and possessed much higher heterozygosity (below). For subsequent analyses, we excluded this genet and only included the remaining 11 genets for all subsequent analyses.

Estimating genetic diversity

Gene annotations for *S. polyrhiza* were obtained from the Joint Genome Institute's Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Spolyrhiza). We used this annotation to obtain genetic diversity estimates for different site types. After filtering, 15310 of the 19623 annotated genes in the *S. polyrhiza* reference genome remained. We categorized these genes based on their level of expression in fronds and turions (characterized by Wang *et al.* 2014 , FKPM values provided by Joachim Messing, Rutgers University) and homology to three monocot species (*Sorghum bicolor*, *Zea mays*, and *Oryza sativa*). Of these, 1058 genes had no detectable level of expression (FPKM was 0 in both fronds and turions), leaving 14252 genes in the expression groups. We separated these remaining genes into four categories of expression, each with 3562 genes (low, mid-low, mid-high and high).

We further split these genes based on three levels of evolutionary constraint by using Blastx to assess sequence similarity between *S. polyrhiza* genes and three other species (*Sorghum bicolor*, *Zea mays*, *Oryza sativa*). We found genes with blast hits to all three species. For each *S. polyrhiza* gene we summed the bit scores for each of the three blast hits and categorized genes into 3 levels of constraint (high, mid and low).

Linkage disequilibrium

The pattern of LD decay across all genotypes should reflect historical recombination events between the disparate genotypes. We calculated LD as the covariance across diploid genotypes using the script provided by Rogers and Huff (2009) to estimate the covariance between two vectors of genotypes. We modified this script to calculate LD across all loci where the minor allele frequency was greater than 0.05. Within each scaffold we calculated LD between pairs of sites up to 500 kb. We also estimated the average r^2 between pairs of scaffolds to establish genome-wide background levels of LD. For this, we randomly sampled one site for each of 32 scaffolds and calculated all pairwise r^2 among these 32 sites. We repeated this 10000 times to get an estimate of r^2 across scaffolds; we got similar results if we only repeated the sampling 500 or 1000 times.

Recombination hotspots

We used the 'rhomap' function within LDhat (McVean *et al.*, 2004) to estimate $\rho = 4N_e r$ for each scaffold separately, where r is the recombination rate per generation. We ran the program for 9,900,000 iterations sampling every 1000 iterations after a burn-in of 100,000 iterations with a block penalty of 5. To search for putative recombination hotspots, we used the output of LDhat to calculate the average recombination rate of 1kb non-overlapping window within each scaffold. We then putatively define recombination hotspots as windows that have 10 times the recombination rate of the scaffold average; adjacent windows that pass this threshold were merged into one putative hotspot.

To examine whether the putative recombination hotspots possessed unique characteristics relative to the rest of the genome we sampled a set of 'control regions in the genome. We randomly sampled 131 regions of the genome with the same length as the putative hotspots requiring that each region had <= 5% missing nucleotides in the reference, at least one SNP and did not overlap with existing hotspots or other control regions. We then examined differences in the GC content, proportion of bases overlapping coding regions and genetic diversity between our hotspots and the control regions using logistic regressions in R (R Core Team 2016). The logistic model we used was: Region type (i.e. hotspot or control) ~ GC content + genetic diversity + coding sequence overlap.

Results and Discussion

Heterozygosity and clonal genotypic structure

After variant calling and filtering, we obtained 79,166,349 invariant and 417,884 variant sites among the 36 *S. polyrhiza* samples, which resulted in an average observed heterozygosity of 0.000636 per site (Table S5). Grouping the samples into 12 distinct genets (i.e., 12 genotypes that are not clonal relatives of one another) resulted in a similar average observed heterozygosity of 0.000694 (Table S4). However, one genet (made up of samples RU448 and RU99) had approximately three times higher heterozygosity compared to the average of the other 11 genotypes (Table S4). This genotype was also highly divergent from the others (below) and over-contributed to the singleton category in the allele frequency spectrum. For these reasons, this genotype was excluded from most analyses below (unless stated otherwise). For the remaining 11 genets, we had 71,949,140 invariant sites and 142,106 variant sites and the average observed heterozygosity was 0.000588. We calculated $F_{IS} = 1 - H_{obs} / H_{exp}$ among the 11 genets at each site and found the average F_{IS} to be 0.044, where H_{obs} and H_{exp} are the observed and expected heterozygosity at a site, respectively (Figure 1).

Genetic distance

We constructed neighbour joining trees based on their pairwise genotypic and allelic distance among the 36 *S. polyrhiza* samples (Figure 2). Both trees revealed consistent clusters of genetically similar samples and the one pair of samples (RU99 and RU448) that was highly divergent. The 36 samples clustered into 12 groups based on the number of homozygous and heterozygous differences between pairs of samples (Table S4, Fig. S1). Individuals found in the same genotype group generally showed strong geographic associations; for example, all samples from Oklahoma belonged to the same genotype group (samples labelled CC and RC). Our most extensively sampled population from a large pond in Toronto (GP) had five samples from a single genotype group (GP8-1, GP10-3, GP6-5, GP4-2, and GP2-3), while a sixth sample (GP4-4) formed a separate genotype group with a sample from a nearby pond. On the other hand, there is no significant correlation between genetic and geographic distance among the 12 genets (genotypic distance: Pearson's r=0.092, p=0. 16; allelic distance: r=0.041, p=0.54) Three samples (RD24, BC RR2_1 and RU195) were the sole representative of their genet while all other genotypes had at least two samples (Table S4). We confirmed the groupings into distinct genets using k-means clustering. The model with 13 genets had the lowest BIC score (BIC= 305.7751), but that with 12 groups was the next best fit (BIC= 307.1549). Clustering with

k = 13 assigned the two Nova Scotia samples (HFA10, HFB11) to separate groups. However, based on the low pairwise genotypic and allelic distance between these two samples, we decided to represent the two samples as one genet in downstream analyses.

The unusual genet consisting of RU99 and RU448 was divergent in genotypic and allelic distances suggesting that was very different from the other genets. To investigate this further we examined sites where a focal genet possesses an alternate allele (i.e. heterozygote or homozygous for the alternate allele) but all other genotypes are homozygous for the reference allele. We found that the {RU99, RU448} genet possess 223652 sites where it is the only genet with the alternate allele, while the other 11 genotypes possess 7542 of these sites, on average (Table S4). Due to the strong differentiation of the {RU99, RU448} genet from the others and its abnormally high levels of heterozygosity (above), we removed it from downstream analyses.

Genetic diversity

Among the 11 *S. polyrhiza* genets, genetic diversity across all sites was very low (π = 0.000542) compared to other plants (Chen *et al* 2017). We estimated π_s to be 0.000463 at 4-fold degenerate (synonymous) sites and π_n to be 0.000229 at 0-fold (non-synonymous sites) resulting in π_n/π_s = 0.495 (Table 1), suggesting a relatively low level of selective constraint in *Spirodela* on amino acid mutations. This ratio is considerably higher than other plant species estimates to date (Chen *et al* 2017). Diversity at 2- or 3-fold degenerate sites was 0.000351 and as expected falls between diversity at 0-fold and 4-fold sites. Diversity at intergenic sites was 0.000732, which was higher than π_n. Values were similar for Watterson's estimate of diversity (θ_w) and π resulting in Tajima's *D* values close to zero for both synonymous and intergenic sites (Table 1). Tajima's *D* is slightly more negative at nonsynonymous sites, consistent with weak purifying selection.

The relatively high value of π_n/π_s could reflect a reduced genome-wide efficacy of natural selection ($N_e s$) due to low rates of sexual reproduction causing low N_e but may also reflect weaker selection (i.e., low *s*) on many genes, perhaps due to a simplified morphology and life cycle since divergence from monocot ancestors. Under the latter hypothesis, the elevated π_n/π_s may only occur in genes with little to no expression. Genes with lower expression had higher diversity at both synonymous and non-synonymous sites (Figure 3, Table S6). However, the reduction in diversity with increasing expression levels was proportionally faster for π_n than for π_s, which resulted in π_n/π_s values decreasing with increasing expression (e.g. π_n/π_s for low expression and high expression genes was 0.557 and 0.346, respectively). These results are consistent with purifying selection being stronger at genes with higher

expression levels, consistent with patterns observed in other species (Carneiro *et al.*, 2012); (Paape *et al.*, 2013; Williamson *et al.*, 2014) . Very high levels of π_n/π_s in genes with low or no expression may indicate that some fraction of these genes are not functionally important. However, even genes in the more highly expressed categories show relatively high π_n/π_s compared to other plants (Chen *et al.*, 2017), suggesting a genome-wide signal of low selection efficacy in *S. polyrhiza*.

To further explore variation across genes in π_n/π_s, we categorized *S. polyrhiza* genes within each expression level category based on their evolutionary constraint, based on their *blastx* divergence from *Sorghum bicolor*, *Zea mays*, *Oryza sativa*. We observed that within each expression category, genes that are more conserved had lower π_n/π_s values (Figure 3, Table S7-S10). Nevertheless, π_n/π_s values are high in all expression/constraint categories. Only in the most highly expressed and highly constrained genes does π_n/π_s approach values typically observed in outcrossing plants (Chen *et al.*, 2017).

There are two patterns in synonymous diversity that are somewhat unexpected. Diversity at synonymous sites is low relative to intergenic sites, and it is also elevated in genes that are weakly expressed compared to those in high expression categories. There are two possible explanations for these patterns; synonymous sites may themselves be under purifying selection or they may be subject to the effects of background selection (or other forms of linked selection) from neighboring selected sites. If background selection is acting, we would predict that synonymous diversity should be reduced in regions with a higher density of functional sites. To test this, we examined the relationship between diversity across 50 kb windows for different site types using a linear model that includes coding site (CDS) density, GC content, and the CpG/GpC ratio, which has been used as an indicator of the level of DNA methylation (i.e. higher CpG/GpC implies less methylation (Hellsten *et al.*, 2013). As shown in Table 2, CDS density negatively affects diversity at all site types. The effect on synonymous sites is marginally nonsignificant but in the same direction as for other site types. The lack of significance may arise from the large number of windows lacking any synonymous SNPs, likely violating the assumptions of the model. Re-analyzing the data using only windows containing at least one SNP of each site type, reveals that the effect of CDS density is highly significant on all site types, including synonymous sites (Table S11). These results are consistent with background selection reducing diversity in regions with more sites under constraint. The CpG/GpC ratio has a significantly negative effect on diversity, most notably for intergenic sites, consistent with the idea that regions with low methylation (i.e., high CpG/GpC) have lower mutation rates.

Linkage disequilibrium and recombination heterogeneity

Levels of linkage disequilibrium (LD) among the genets are affected by the amount of recombination that has occurred in the past. Among the 11 *S. polyrhiza* genets, we observed that sites within 1 to 20 bp of each other had an average r^2 of 0.57 that decays to approximately 0.23 after a distance of 20 kb (Figs. 4a, b), after which there is a slower LD decay that continues to 100 kb. Between-scaffold LD is slightly but significantly lower (0.13) than LD at 100 kb (0.15), implying small residual LD at very large distances. This pattern of LD decay is comparable to that seen in the highly self-fertilizing species *Arabidopsis thaliana* (10-50kb, depending on sampling; (Nordborg *et al.*, 2005; Kim *et al.*, 2007) and *Medicago truncatula* (Branca *et al.*, 2011), whereas outcrossing populations often show much more rapid LD decay over several hundred base pairs (Foxe *et al.*, 2009; Mackay *et al.*, 2012).

On average across our 50 kb windows, our estimate of the population recombination rate ρ from LDhat is 0.00051. The ratio of ρ/θ is thus approximately 1, implying an effectively comparable rate of recombination and mutation. This ratio is considerably higher than that estimated in highly selfing species (Nordborg *et al.*, 2005; Branca *et al.*, 2011), while it is on the same order as outcrossing species (Wright *et al.*, 2003; Langley *et al.*, 2012). Since *Spirodela* is predominantly asexual, this high ratio of ρ/θ is somewhat surprising. There are several possibilities that might explain this discrepancy. First, the very low level of neutral diversity may reflect a very low mutation rate, potentially due to the low numbers of cell divisions per generation compared to other vascular plants. Recent estimates of mutation rate in this species are consistent with this possibility (Xu *et al.*, 2018; Sandler et al. in prep). Second, in partially asexual species, mitotic gene conversion and mitotic crossing over play a major role in reducing linkage disequilibrium in facultatively sexual organisms, but would not do so in selfers (Hartfield *et al.*, 2018). Thus, the combination of high ρ/θ and very low diversity is not inconsistent with predictions of a highly clonal organisms experiencing mitotic recombination and/or a low rate of per-base mutation.

Using LDhat, we identified 131 putative recombination hotspots in *S. polyrhiza* that possessed a recombination rate ten times higher than the average of the scaffold they belong to (Table S12). Compared to some species (e.g. mammals), these hotspots are weak in strength and very few in number. One possible explanation for their apparent rarity is that low levels of polymorphism are causing low power to detect hotspots. Indeed, hotspot presence was significantly correlated with levels of nucleotide diversity (π: Pearson's r = 0.28, p<0.0001, θ_W: r = 0.31, p<0.0001;). To test whether low SNP density limits our detection power, we randomly masked SNPs from our data such that each 2 kb had a maximum of 5 SNPs and then re-ran

LDhat and resampled appropriate control regions. In this SNP-masked dataset, we found only 50 putative hotspots which recovered ~36% of the putative hotspots in the full dataset; two of the 50 putative hotspots were not found in the full dataset. This suggests that there is bias in the unmasked dataset for detecting hotspots in SNP-dense regions.

To look more broadly at recombination rate heterogeneity, we estimated average ρ in 50 kb windows across the genome, and tested for correlations with several genomic features. Previous work in plants suggest that recombination rates are elevated upstream of coding regions, and are associated with demethylated promoter regions. For these reasons, we examined the effects of coding sequence density and the CpG/GpC ratio. ρ is significantly positively correlated with coding sequence density (Pearson's r=0.20, p<0.0001), GC content (r=0.13, p<0.0001), CpG/GpC (r=0.16, p<0.0001) and π (r = 0.08, p = 0.005). Since these variables are likely to be non-independent of each other, we also ran the linear model *(ρ ~ GC content+CpG/GpC+coding sequence density+ π)* (Table 3). From this analysis, both coding sequence density and CpG/GpC positively affect ρ, while GC content now has a negative effect. The effects of coding density and CpG/GpC are consistent with recombination being biased towards open chromatin, as seen in other plants (Hellsten *et al.*, 2013; Rodgers-Melnick *et al.*, 2016).

Conclusion

The observation of low diversity could be explained by either low mutation rate or a low N_e. Direct estimates of the mutation rate indicate that the mutation rate is very low (i.e., one to two orders of magnitude lower than *Arabidopsis*; Xu *et al.*, 2018, Sandler et al *in prep*). Based on this low estimate, Xu *et al* (2018) inferred that N_e is quite large (~10^6); our own estimate of the mutation rate (Sandler et al. in prep) and diversity is broadly consistent with a large N_e.

At the outset, we had expected to find N_e to be low because this plant produces primarily by cloning so the effects of linked selection could be large (Agrawal and Hartfield 2016). However, background selection will not reduce N_e as much as it would if the mutation rate was higher. Nonetheless, the effects of linked selection are important in this system as N_e is likely much smaller than N. While no direct estimate of the census population size N exists, we expect it is massive. There are ~6 million ponds across Canada and the USA [flyways.us] so it is not unreasonable to speculate that census population size exceeds 10^9.

Our observation that ρ/θ is close to 1 implies that the effective recombination is close to the mutation rate. From this, we can infer the effective recombination rate in this species is several orders of magnitude lower than it is in outcrossing species such as *Drosophila* which

also have ρ/θ values close to 1 but also have much higher mutation rates than in *Spirodela*. Presumably, this low effective recombination rate occurs because of the low rate of sex. If the recombination rate per meiosis is 10^{-8} (comparable to other species), but the effective recombination rate is 10^{-10} (i.e., equal to the mutation rate, see Xu et al. 2018), then we infer the rate of sex is $10^{-10}/10^{-8} = 10^{-2}$. This simplistic calculation ignores the potential importance of mitotic gene conversion (which contributes to the inferred estimate of ρ), so the true rate of sex may be substantially lower. Thus, low mutation rate and low rates of sexual reproduction are likely contributing to our patterns of diversity and linkage disequilibrium.

Though we infer the rate of sex is low at the individual level, we see evidence of the effects of recombination at the species level. First, linkage disequilibrium declines with distance. The decline in LD is not dissimilar to that observed in selfing plants such as *Arabidopsis thaliana* (Nordborg *et al.*, 2005; Kim *et al.*, 2007) and *Medicago truncatula* (Branca *et al.*, 2011) that outcross at a low rate. Second, π_s is lower in gene dense regions, a pattern expected if recombination localizes background selection effects to tightly linked regions of the genome.

The most perplexing observation is the high π_n/π_s relative to other species, given that the effective population size is estimated to be large (Xu et al. 2018). The simplest explanation is that selection (s) tends to be weak in *S. polyrhiza*. This could be because of relaxed selection on many genes due to the diminutive form and lifestyle relative to other angiosperms. Alternatively, selection may not be realized much of the time because local populations often consist of a single clone so there is not competitive selection among genotypes. Finally, the high π_n/π_s might be explained by non-equilibrium conditions. It has been pointed out that neutral diversity takes longer to build up to its equilibrium levels than selected diversity, which can result in a transiently elevated π_n/π_s (Simons *et al.*, 2014; Brandvain & Wright, 2016). Our estimates of Tajima's D are negative, which could reflect recovery from bottlenecks. Disentangling the reasons for the high π_n/π_s remains a challenge for future work.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (AFA and SIW). We thank Jade Lavallee, Victor Mollov, and Niroshini Epitawalage for help with plant growth and extractions and Adrian Platts for bioinformatics assistance.

References

Agrawal AF, Hartfield M. **2016**. Coalescence with background and balancing Selection in systems with bi- and uniparental reproduction: contrasting partial asexuality and selfing. *Genetics* **202**: 313–326.

Ament-Velásquez SL, Figuet E, Ballenghien M, Zattara EE, Norenburg JL, Fernández-Álvarez FA, Bierne J, Bierne N, Galtier N. **2016**. Population genomics of sexual and asexual lineages in fissiparous ribbon worms (Lineus, Nemertea): hybridization, polyploidy and the Meselson effect. *Molecular Ecology* **25**: 3356–3369.

Arunkumar R, Ness RW, Wright SI, Barrett SCH. **2015**. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics* **199**: 817–829.

Balloux F, Lehmann L, de Meeûs T. **2003**. The population genetics of clonal and partially clonal diploids. *Genetics* **164**: 1635–1644.

Barraclough TG, Fontaneto D, Ricci C, Herniou EA. **2007**. Evidence for inefficient selection against deleterious mutations in cytochrome oxidase I of asexual bdelloid rotifers. *Molecular Biology and Evolution* **24**: 1952–1962.

Barrett SCH. **2002**. Evolution of sex: the evolution of plant sexual diversity. *Nature Reviews Genetics* **3**: 274.

Barrett SCH. **2015**. Influences of clonality on plant sexual reproduction. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 8859–8866.

Bog M, Baumbach H, Schween U, Hellwig F, Landolt E, Appenroth K-J. **2010**. Genetic structure of the genus Lemna L. (Lemnaceae) as revealed by amplified fragment length polymorphism. *Planta* **232**: 609–619.

Branca A, Paape TD, Zhou P, Briskine R, Farmer AD, Mudge J, Bharti AK, Woodward JE, May GD, Gentzbittel L, *et al.* 2011. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proceedings of the National Academy of Sciences of the United States of America* **108**: E864–70.

Brandvain Y, Slotte T, Hazzouri KM, Wright SI, Coop G. **2013**. Genomic identification of founding haplotypes reveals the history of the selfing species *Capsella rubella*. *PLoS Genetics* **9**: e1003754.

Brandvain Y, Wright SI. **2016**. The limits of natural selection in a nonequilibrium world. *Trends in Genetics* **32**: 201–210.

Burgarella C, Gayral P, Ballenghien M, Bernard A, David P, Jarne P, Correa A, Hurtrez-Boussès S, Escobar J, Galtier N, *et al.* 2015. Molecular evolution of freshwater snails with contrasting mating systems. *Molecular Biology and Evolution* **32**: 2403–2416.

Butlin R. **2002**. Evolution of sex: The costs and benefits of sex: new insights from old asexual lineages. *Nature Reviews Genetics* **3**: 311–317.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. **2012**. Evidence for widespread positive and purifying selection across the European rabbit (*Oryctolagus cuniculus*) genome. *Molecular Biology and Evolution* **29**: 1837–1849.

Charlesworth B, Charlesworth D. **1997**. Rapid fixation of deleterious alleles can be caused by Muller's ratchet. *Genetical Research* **70**: 63–73.

Charlesworth D, Morgan MT, Charlesworth B. **1993**. Mutation accumulation in finite outbreeding and inbreeding populations. *Genetics Research* **61**: 39–56.

Charlesworth B, Nordborg M, Charlesworth D. **1997**. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research* **70**: 155–174.

Chen J, Glémin S, Lascoux M. **2017**. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Molecular Biology and Evolution* **34**: 1417–1428.

Crawford DJ, Landolt E. **1993**. Allozyme studies in *Spirodela* (Lemnaceae): variation among conspecific clones and divergence among the species. *Systematic Botany* **18**: 389.

Cutter AD. **2006**. Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. *Genetics* **172**: 171–184.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, *et al.* 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491–498.

Escobar JS, Cenci A, Bolognini J, Haudry A, Laurent S, David J, Glémin S. **2010**. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). *Evolution* **64**: 2855–2872.

Fisher RA. **1941**. Average excess and average effect of a gene substitution. *Annals of Eugenics* **11**: 53–63.

Foxe JP, Slotte T, Stahl EA, Neuffer B, Hurka H, Wright SI. **2009**. Recent speciation associated with the evolution of selfing in *Capsella*. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 5241–5245.

Glémin S, Bazin E, Charlesworth D. **2006**. Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proceedings of the Royal Society Series B* **273**: 3011–3019.

Glémin S, Galtier N. **2012**. Genome evolution in outcrossing versus selfing versus asexual species. *Methods in Molecular Biology* **855**: 311–335.

Glémin S, Muyle A. **2014**. Mating systems and selection efficacy: a test using chloroplastic sequence data in angiosperms. *Journal of Evolutionary Biology* **27**: 1386–1399.

Glémin S, Ronfort J. **2013**. Adaptation and maladaptation in selfing and outcrossing species: new mutations versus standing variation. *Evolution* **67**: 225–240.

Golding GB, Strobeck C. **1980**. Linkage disequilibrium in a finite population that is partially selfing. *Genetics* **94**: 777–789.

559 **Goodwillie C, Kalisz S, Eckert CG**. **2005**. The evolutionary enigma of mixed mating systems in
560 plants: occurrence, theoretical explanations, and empirical evidence. *Annual Review of Ecology,*
561 *Evolution, and Systematics* **36**: 47–79.

562 **Hartfield M**. **2016**. Evolutionary genetic consequences of facultative sex and outcrossing.
563 *Journal of Evolutionary Biology* **29**: 5–22.

564 **Hartfield M, Wright SI, Agrawal AF**. **2016**. Coalescent times and patterns of genetic diversity
565 in species with facultative sex: effects of gene conversion, population structure, and
566 heterogeneity. *Genetics* **202**: 297–312.

567 **Hartfield M, Wright SI, Agrawal AF**. **2018**. Coalescence and linkage disequilibrium in
568 facultatively sexual diploids. *Genetics* **210**: 683–701.

569 **Heller R, Maynard Smith J**. **1978**. Does Muller's ratchet work with selfing? *Genetics Research*
570 **32**: 289–293.

571 **Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, Schmutz J, Willis JH,**
572 **Rokhsar DS**. **2013**. Fine-scale variation in meiotic recombination in *Mimulus* inferred from
573 population shotgun sequencing. *Proceedings of the National Academy of Sciences* **110**: 19478–
574 19482.

575 **Hicks LE**. **1932**. Flower production in the Lemnaceae. *The Ohio Journal of Science* **32**: 115-
576 132.

577 **Hudson RR, Kaplan NL**. **1995**. The coalescent process and background selection.
578 *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* **349**:
579 19–23.

580 **Jarne P, Auld JR**. **2006**. Animals mix it up too: the distribution of self-fertilization among
581 hermaphroditic animals. *Evolution* **60**: 1816–1824.

582 **Jombart T, Ahmed I**. **2011**. adegenet 1.3-1: new tools for the analysis of genome-wide SNP
583 data. *Bioinformatics* **27**: 3070–3071.

584 **Jordan WC, Courtney MW, Neigel JE**. **1996**. Low levels of intraspecific genetic variation at a
585 rapidly evolving chloroplast DNA locus in North American duckweeds (Lemnaceae). *American*
586 *Journal of Botany* **83**: 430–439.

587 **Kamran-Disfani A, Agrawal AF**. **2014**. Selfing, adaptation and background selection in finite
588 populations. *Journal of Evolutionary Biology* **27**: 1360–1371.

589 **Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D,**
590 **Nordborg M**. **2007**. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature*
591 *Genetics* **39**: 1151–1155.

592 **Kimura M**. **1962**. On the probability of fixation of mutant genes in a population. *Genetics* **47**:
593 713–719.

594 **Laenen B, Tedder A, Nowak MD, Toräng P, Wunder J, Wötzel S, Steige K, Kourmpetis Y,**
595 **Odong T, Drouzas AD, *et al.* 2018**. Demography and mating system shape the genome-wide
596 impact of purifying selection in *Arabis alpina*. *Proceedings of the National Academy of Sciences*
597 **115**: 816-821.

Landolt E. **1986**. *Biosystematic investigations in the family of duckweeds (Lemnaceae), volume 2. The family of Lemnaceae—a monographic study, volume 1*. Zurich: Veroffentlichungen des Geobotanischen Institutes der ETH, Stiftung Rubel, in Zurich (71 Heft).

Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, Langley SA, Suarez C, Corbett-Detig RB, Kolaczkowski B, *et al.* 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* **192**: 533–598.

Les DH, Landold CDJ, and EGJ. **2002**. Phylogeny and systematics of Lemnaceae, the duckweed family. *Systematic Botany* **27**: 221–240.

Lloyd DG. **1979**. Some reproductive factors affecting the selection of self-fertilization in plants. *The American Naturalist* **113**: 67–79.

Lutz KA, Wang W, Zdepski A, Michael TP. **2011**. Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnology* **11**: 54.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, *et al.* 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* **482**: 173–178.

Mark Welch D, Meselson M. **2000**. Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288**: 1211–1215.

Maynard-Smith J. **1978**. *The evolution of sex*. Cambridge University Press Cambridge.

McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. **2004**. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**: 581–584.

Muller HJ. **1932**. Some genetic aspects of sex. *The American Naturalist* **66**: 118–138.

Ness RW, Siol M, Barrett SCH. **2012**. Genomic consequences of transitions from cross- to self-fertilization on the efficacy of selection in three independently derived selfing plants. *BMC Genomics* **13**: 611.

Nordborg M. **2000**. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**: 923–929.

Nordborg M, Donnelly P. **1997**. The coalescent process with selfing. *Genetics* **146**: 1185–1195.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, *et al.* 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **3**: e196.

Orr HA. **2000**. The rate of adaptation in asexuals. *Genetics* **155**: 961–968.

Otto SP. **2009**. The evolutionary enigma of sex. *The American Naturalist* **174**: S1–S14.

Paape T, Bataillon T, Zhou P, J Y Kono T, Briskine R, Young ND, Tiffin P. **2013**. Selection, genome-wide fitness effects and evolutionary rates in the model legume *Medicago truncatula*. *Molecular Ecology* **22**: 3525–3538.

635 **Paland S, Lynch M**. **2006**. Transitions to asexuality result in excess amino acid substitutions.
636 *Science* **311**: 990–992.

637 **Paradis E**. **2011**. *Analysis of Phylogenetics and Evolution with R*. Springer Science & Business
638 Media.

639 **Paradis E, Claude J, Strimmer K**. **2004**. APE: Analyses of Phylogenetics and Evolution in R
640 language. *Bioinformatics* **20**: 289–290.

641 **Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D**. **2011**. Reduced efficacy of natural
642 selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. *Genome Biology*
643 *and Evolution* **3**: 868–880.

644 **R Core Team, N**. **2016**. R: A language and environment for statistical computing [Computer
645 software manual]. Vienna, Austria.

646 **Rice WR**. **2002**. Experimental tests of the adaptive significance of sexual recombination. *Nature*
647 *Reviews Genetics* **3**: 241–251.

648 **Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES**. **2016**. Open chromatin reveals the
649 functional maize genome. *Proceedings of the National Academy of Sciences of the United*
650 *States of America* **113**: E3177–84.

651 **Rogers AR, Huff C**. **2009**. Linkage disequilibrium between loci with unknown phase. *Genetics*
652 **182**: 839–844.

653 **Schoen DJ, Brown AH**. **1991**. Intraspecific variation in population gene diversity and effective
654 population size correlates with the mating system in plants. *Proceedings of the National*
655 *Academy of Sciences of the United States of America* **88**: 4494–4497.

656 **Simons YB, Turchin MC, Pritchard JK, Sella G**. **2014**. The deleterious mutation load is
657 insensitive to recent population history. *Nature Genetics* **46**: 220–224.

658 **Slotte T, Foxe JP, Hazzouri KM, Wright SI**. **2010**. Genome-wide evidence for efficient positive
659 and purifying selection in *Capsella grandiflora*, a plant species with a large effective population
660 size. *Molecular Biology and Evolution* **27**: 1813–1821.

661 **Tang J, Zhang F, Cui W, Ma J**. **2014**. Genetic structure of duckweed populations of *Spirodela*,
662 *Landoltia* and *Lemna* from Lake Tai, China. *Planta* **239**: 1299–1307.

663 **Tucker AE, Ackerman MS, Eads BD, Xu S, Lynch M**. **2013**. Population-genomic insights into
664 the evolutionary origin and fate of obligately asexual *Daphnia pulex*. *Proceedings of the National*
665 *Academy of Sciences of the United States of America* **110**: 15740–15745.

666 **Vrijenhoek RC**. **1998**. Animal clones and diversity. *Bioscience* **48**: 617–628.

667 **Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo MC, Lomsadze A,**
668 **Borodovsky M, Kerstetter RA, Shanklin J, *et al.* 2014**. The *Spirodela polyrhiza* genome
669 reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nature*
670 *Communications* **5**: 3311.

671 **Whitton J, Sears CJ, Baack EJ, Otto SP**. **2008**. The dynamic nature of apomixis in the
672 angiosperms. *International Journal of Plant Sciences* **169**: 169–182.

673 **Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI**.
674 **2014**. Evidence for widespread positive and negative selection in coding and conserved
675 noncoding regions of *Capsella grandiflora*. *PLoS Genetics* **10**: e1004622.

676 **Wright SI, Lauga B, Charlesworth D**. **2002**. Rates and patterns of molecular evolution in
677 inbred and outbred *Arabidopsis*. *Molecular Biology and Evolution* **19**: 1407–1420.

678 **Wright SI, Lauga B, Charlesworth D**. **2003**. Subdivision and haplotype structure in natural
679 populations of *Arabidopsis lyrata*. *Molecular Ecology* **12**: 1247–1263.

680 **Xue H, Xiao Y, Jin Y, Li X, Fang Y, Zhao H, Zhao Y, Guan J**. **2012**. Genetic diversity and
681 geographic differentiation analysis of duckweed using inter-simple sequence repeat markers.
682 *Molecular Biology Reports* **39**: 547–554.

683 **Xu Y, Ma S, Huang M, Peng M, Bog M, Sree KS, Appenroth K-J, Zhang J**. **2015**. Species
684 distribution, genetic diversity and barcoding in the duckweed family (Lemnaceae). *Hydrobiologia*
685 **743**: 75–87.

686 **Xu S, Stapley J, Gablenz S, Boyer J, Appenroth KJ, Sree SK, Gershenzon J, Widmer A,**
687 **Huber MSC**. **2018**. Low genetic variation is associated with low mutation rate in the giant
688 duckweed. *biorxiv* doi: https://doi.org/10.1101/381574
689
690
691

692 **Table 1. Genetic diversity and Tajima's D estimates**

Site type	π	θ_w	Tajima's D
Intergenic	0.00073 2	0.00074 7	-0.0826
Intron	0.00044 9	0.00046 0	-0.107
2-, 3-fold degen.	0.00035 1	0.00037 1	-0.229
Nonsynonymous	0.00022 9	0.00025 3	-0.402
Synonymous	0.00046 3	0.00047 9	-0.143
Nonsyn. / Syn.	0.495	0.510	

693 *Synonymous and nonsynonymous refer to sites that are 4-fold and 0fold degenerate, respectively

694
695
696

697 **Table 2**. Coefficients from linear model of diversity (π) of different site types (per 50 kb

698 window) as function of coding site (CDS) density, % GC, and the CpG/GpC ratio.

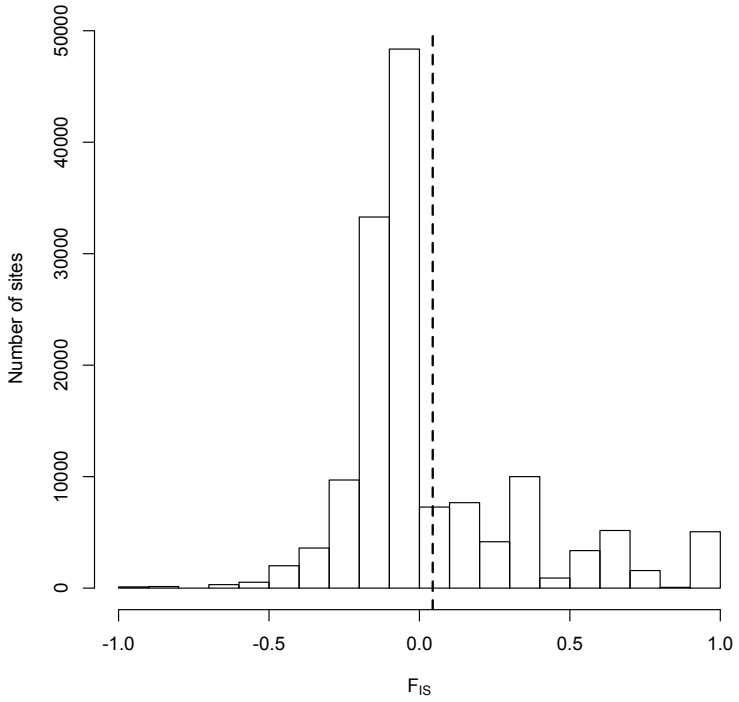
Site type	Variable	Estimate	t	P-value
Intergenic	CDS density	-0.00061	-2.17	0.030
	% GC,	-0.00038	-1.18	0.237
	CpG:GpC	-0.00063	-4.71	< 0.0001
Intronic	CDS density	-0.00096	-4.52	< 0.0001
	% GC,	0.00017	0.72	0.474
	CpG:GpC	-0.00021	-2.06	0.039
0 fold	CDS density	-0.00102	-5.40	< 0.0001
	% GC,	0.00044	2.02	0.043
	CpG:GpC	-0.00006	-0.64	0.523
4 fold	CDS density	-0.00056	-1.59	0.111
	% GC,	-0.00008	-0.19	0.848
	CpG:GpC	0.00007	0.44	0.661

699

700

Table 3. Linear model examining variation in estimates of ρ for 50 kb windows as a function of window characters: CDS density, % GC, the CpG/GpC ratio and diversity (π). Statistical model: lm(rho ~ CDS density + % GC+ CpG/GpC ratio + π)

	Estimate	Standard error	*t*	*p*
Intercept	0.5236	0.1127	4.645	< 0.0001
CDS density	0.8997	0.1833	4.909	< 0.0001
% GC	-0.8643	0.3647	-2.37	0.0179
CpG:GpC	0.2687	0.0989	2.716	0.0067
π	-30.0521	20.2418	-1.485	0.1379



710

711 **Figure 1**. Distribution of F_{IS} for 11 *S. polyrhiza* genets.

712

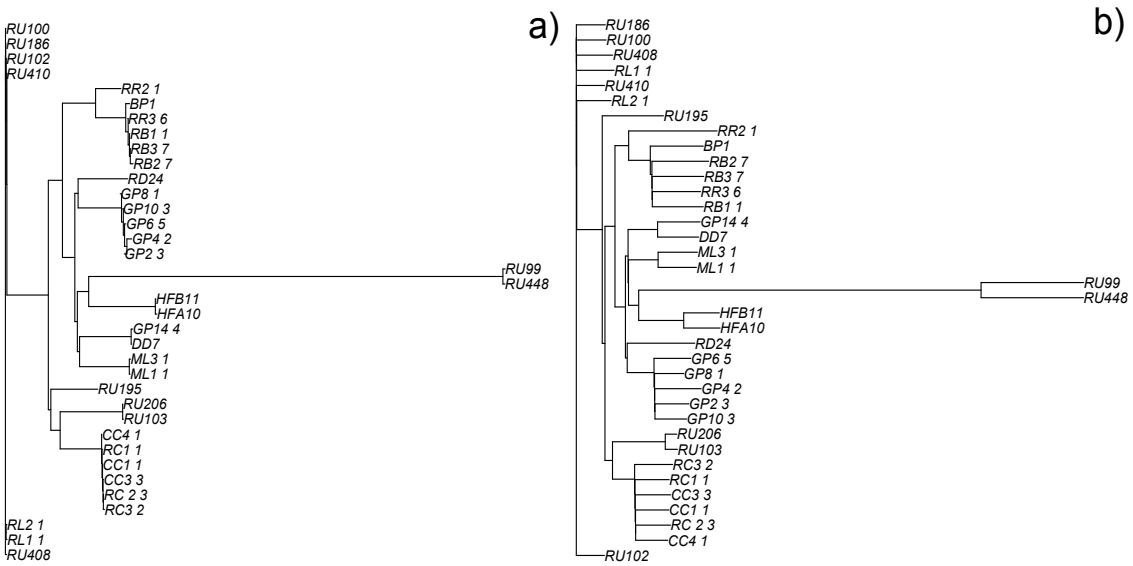


Figure 2. Neighbour joining trees for 36 *S. polyrhiza* samples constructed using pairwise (a) genotypic and (b) allelic distances.

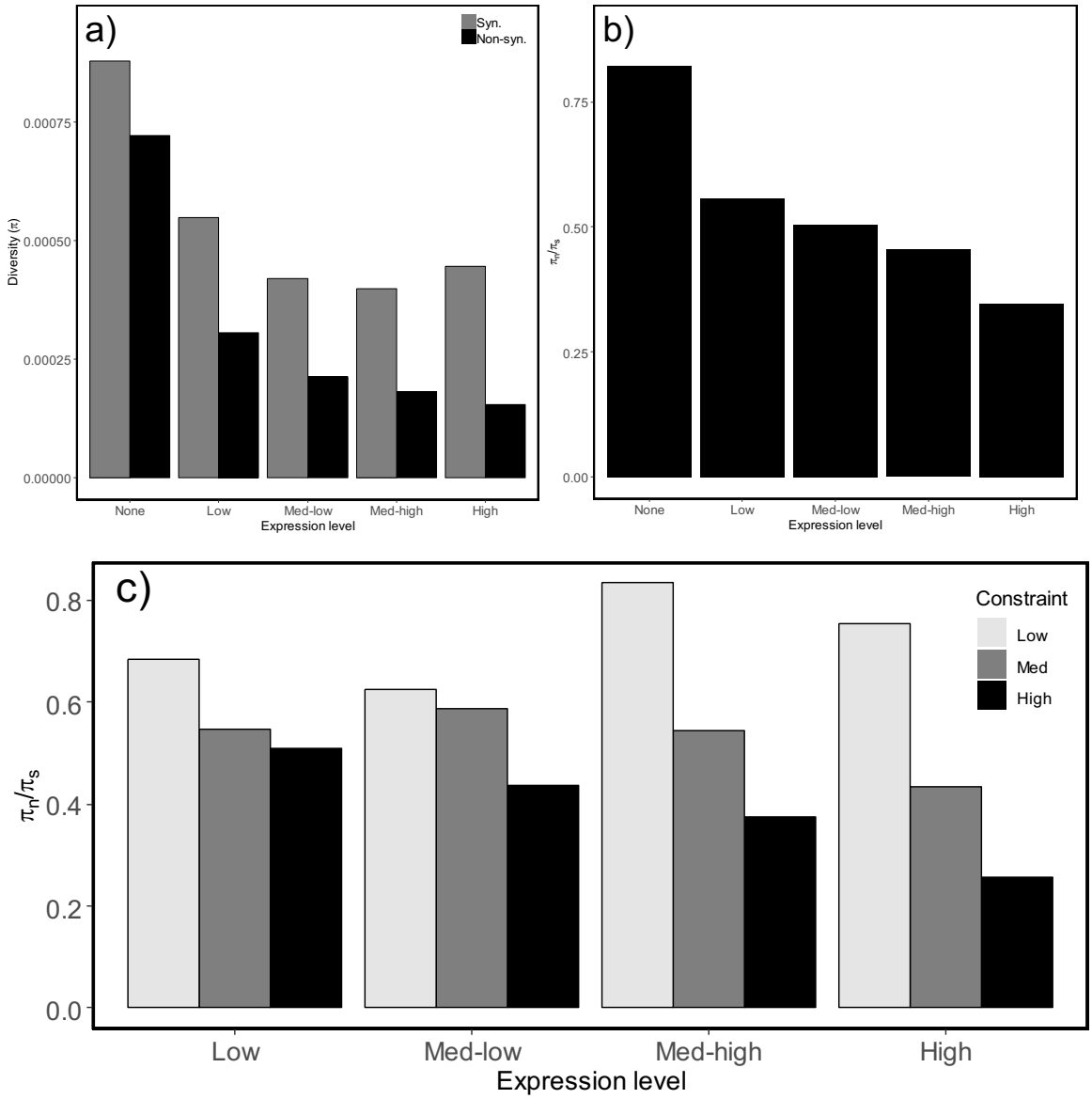


Figure 3. (a) π_s and π_n for *S .polyrhiza* at genes with varying expression levels. (b) π_n / π_s in *S. polyrhiza* genes that have different expression levels within *S. polyrhiza* tissue. (c) π_n / π_s in *S. polyrhiza* genes that have different expression levels within *S. polyrhiza* tissue. Within each expression level category, genes are separated into low, mid and high evolutionary constraint based on their divergence to homologous genes in *Sorghum bicolor*, *Zea mays*, *Oryza sativa*. *S. polyrrhiza* consists of 11 genets.

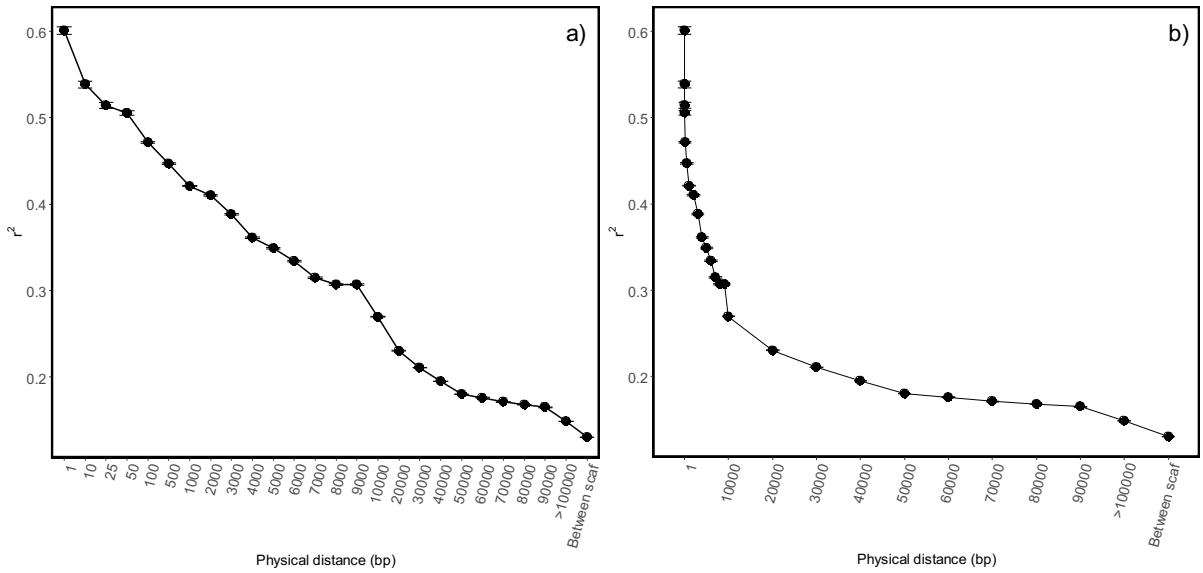


Figure 4. (a) Linkage disequilibrium among 11 genets of *S. polyrhiza,* measured as mean r^2, decaying with distance (bp) between pairs of sites. Each point contains r^2 values binned by physical distance. For example, bins contain pairs of loci that are 1-10, 10-25, 50-100, …, etc. bp apart. The last point indicates the average r^2 between scaffolds. (b) Same as (a) but with the distance shown on a linear scale.