Patterns of population structure and complex haplotype sharing among field isolates of the green alga Chlamydomonas reinhardtii

Citation for published version:
Craig, R, Böndel, KB, Arakawa, K, Nakada, T, Ito, T, Bell, G, Colegrave, N, Keightley, P & Ness, RW 2019, 'Patterns of population structure and complex haplotype sharing among field isolates of the green alga Chlamydomonas reinhardtii', *Molecular Ecology*, vol. 28, no. 17, pp. 3977-3993. https://doi.org/10.1111/mec.15193

Digital Object Identifier (DOI):
10.1111/mec.15193

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Molecular Ecology

Patterns of population structure and complex haplotype sharing among field isolates of the green alga *Chlamydomonas reinhardtii*

Rory J. Craig[1,2], Katharina B. Böndel[1,3], Kazuharu Arakawa[4,5], Takashi Nakada[4,5,6], Takuro Ito[4,5], Graham Bell[7], Nick Colegrave[1], Peter D. Keightley[1] & Rob W. Ness[2]

1 Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3FL, Edinburgh, United Kingdom

2 Department of Biology, University of Toronto Mississauga, Mississauga, Ontario, L5L 1C6, Canada

3 Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, 70599 Stuttgart, Germany

4 Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, 997-0052, Japan

5 Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa, 252-0882, Japan

6 Faculty of Environment and Information Sciences, Yokohama National University, Yokohama, Kanagawa, 240-8501, Japan

7 Department of Biology, McGill University, Montreal, Quebec H3A 1B1, Canada

Correspondence: rory.craig@ed.ac.uk, rob.ness@utoronto.ca

Keywords: Population structure – Genetic differentiation – Microbial eukaryotes – *Chlamydomonas reinhardtii* – Admixture – Identity by descent

Abstract

The nature of population structure in microbial eukaryotes has long been debated. Competing models have argued that microbial species are either ubiquitous, with high dispersal and low rates of speciation, or that for many species gene flow between populations is limited, resulting in evolutionary histories similar to those of macroorganisms. However, population genomics approaches have seldom been applied to this question. Here, we analyse whole-genome re-sequencing data for all 36 confirmed field isolates of the green alga *Chlamydomonas reinhardtii*. At a continental scale, we report evidence for putative allopatric divergence, between both North American and Japanese isolates, and two highly differentiated lineages within N. America. Conversely, at a local scale within the most densely sampled lineage, we find little evidence for either spatial or temporal structure. Taken together with evidence for ongoing admixture between the two N. American lineages, this lack of structure supports a role for substantial dispersal in *C. reinhardtii* and implies that between-lineage differentiation may be maintained by reproductive isolation and/or local adaptation. Our results therefore support a role for allopatric divergence in microbial eukaryotes, while also indicating that species may be ubiquitous at local scales. Despite the high genetic diversity observed within the most well-sampled lineage, we find that pairs of isolates share on average ~9% of their genomes in long haplotypes, even when isolates were sampled decades apart and from different locations. This proportion is several orders of magnitude higher than the Wright-Fisher expectation, raising many further questions concerning the evolutionary genetics of *C. reinhardtii* and microbial eukaryotes generally.

Introduction

‘Everything is everywhere: but the environment selects’ (Baas Becking, 1934) has been a long-standing tenet of microbiology (O'Malley, 2008). Under this paradigm, dispersal is considered to be effectively unlimited, and the biogeography and evolutionary histories of microbial species should therefore be determined by ecology, rather than geography. For microbial eukaryotes (i.e. protists and other unicellular/colonial eukaryotes), this has been extended to the *ubiquity model* (Fenchel & Finlay, 2004; Finlay, 2002; Finlay & Fenchel, 1999), which predicts both cosmopolitan distributions and low rates of speciation, due to the extremely large population sizes and high dispersal of species. This view has been countered by the *moderate endemicity model* (Foissner, 1999, 2006, 2008)*,* which posits that dispersal is limited for many species, and as such the taxonomic diversity, biogeography, and evolution of microbial eukaryotes is generally expected to be more similar to that of macroorganisms. Exploring the validity of these opposing models is thus crucial for determining microbial eukaryotic biodiversity, for understanding the rate and mode of speciation in understudied lineages, and for providing insights into the ecology and evolutionary histories of individual species of interest.

Empirical tests of the two competing models have, however, largely been based on morphology, and their interpretation has been highly dependent on the species concept employed (Caron, 2009). DNA sequence-based studies of microbial eukaryotes are therefore of great importance, primarily to broadly delineate species (due to the prevalence of cryptic speciation (Lahr, Laughinghouse, Oliverio, Gao, & Katz, 2014)), but more specifically to characterise the nature of population structure within species. Genetic structure can arise as a result of barriers to gene flow formed by limited dispersal (allopatry or isolation by distance), reduced establishment of migrants (‘isolation by adaptation’), or more complex patterns caused by founder events (‘isolation by colonisation’) (Orsini, Vanoverbeke, Swillen, Mergeay, & De Meester, 2013). Exploring the extent of population structure and its causes can be used to test between the *ubiquity* and *moderate endemicity* models, as the former predicts a lack of divergence in allopatry or isolation by distance, and little evidence for recent speciation events, in contrast to what is observed in many plants and animals. Evidence for genetically structured populations has recently been reported across a variety of taxa and habitats, including examples from ciliates (Zufall, Dimond, & Doerder, 2013), amoebae (Douglas, Kronforst, Queller, & Strassmann, 2011; Heger, Mitchell, & Leander, 2013), diatoms (Casteleyn et al., 2010; Sjöqvist, Godhe, Jonsson, Sundqvist, & Kremp, 2015; Vanormelingen et al., 2015; Whittaker & Rynearson, 2017), dinoflagellates (Lowe, Martin, Montagnes, & Watts, 2012; Rengefors, Logares, & Laybourn-Parry, 2012), raphidophytes (Lebret, Tesson, Kritzberg, Tomas, & Rengefors, 2015), and fungi (Carriconde et al., 2008; Ellison et al., 2011). While many of these studies showed clear evidence for geographical structure (supporting the *moderate endemicity model*), the majority were limited in resolution due to the small number of marker loci used. Microbial eukaryotes remain severely understudied relative to their abundance and phylogenetic diversity (Pawlowski et al., 2012), and currently very few population genomics datasets exist for free-living species (Johri et al., 2017). Such datasets are required to fully capture patterns of genetic diversity within and between populations, to reveal complex patterns of migration and gene flow, and to identify loci putatively contributing to local adaptation and speciation.

Here, we analyse whole-genome re-sequencing data for all currently available *Chlamydomonas reinhardtii* field isolates. *C. reinhardtii* is a soil-dwelling unicellular green alga that is used extensively as a model organism for plant physiology, molecular and cell biology (Blaby et al., 2014; Harris, 2001, 2008), experimental evolution (Bell, 1997; Colegrave, 2002; Collins & Bell, 2004), and biofuel research (Scranton, Ostrand, Fields, & Mayfield, 2015). Despite its importance as a model system, very little is known about the ecology and evolutionary history of the species (Sasso, Stibor, Mittag, & Grossman, 2018). For many years *C. reinhardtii* had only been isolated from eastern North America, suggesting that the species may be endemic (Pröschold, Harris, & Coleman, 2005). However, isolates that are interfertile with N. American laboratory strains have since been discovered in Japan, implying a more cosmopolitan distribution (Nakada, Shinkawa, Ito, & Tomita, 2010; Nakada, Tsuchida, Arakawa, Ito, & Tomita, 2014). Two previous studies have reported evidence for population structure in field isolates of *C. reinhardtii* (Flowers et al., 2015; Jang & Ehrenreich, 2012), but sampling was limited to N. America, and between the studies a total of only 12 isolates were analysed, limiting the inferences that could be drawn. Furthermore, although there are excellent genomic resources available for *C. reinhardtii* (Blaby et al., 2014; Merchant et al., 2007), the low number of sequenced isolates has hindered the study of the population genetics of the species. *C. reinhardtii* has several attributes that make it a particularly interesting model for population genetics. Synonymous genetic diversity (~3%) and the estimated effective population size (~10^8) are amongst the highest reported in eukaryotes (Flowers et al., 2015), and its haploid state makes it highly amenable to studying recombination and evolutionary phenomena that would otherwise require haplotype phasing.

In this study we explore patterns of population structure inferred from 36 *C. reinhardtii* isolates sampled at three scales, (i) local, both between and within sites and time points in Quebec, (ii) within continent, between N. American isolates, and (iii) between continent, specifically between N. American and Japanese isolates. Overall, we report evidence for allopatric divergence, both between N. American and Japanese isolates, and putatively between two highly differentiated lineages in N. America, supporting the *moderate endemicity model* for the species. We find evidence for substantial admixture between the N. American lineages, providing some of the first insights into the ecology and dispersal capability of *C. reinhardtii*. Furthermore, within Quebec we find little signature of strong geographic or temporal structure. Finally, we report the extensive sharing of unexpectedly long genomic tracts likely to have been inherited identical by descent between pairs of isolates at local scales, and discuss several potential causes of this surprising result.

Materials and methods

Sampling and whole-genome re-sequencing

Sampling and whole-genome re-sequencing of the field isolates available from the Chlamydomonas Resource Centre (https://www.chlamycollection.org) has mostly been described previously. Briefly, sequencing data for 11 isolates sampled at eight locations between 1945 and 1994 were produced by Flowers et al. (2015), with the exception of CC-2932 (Jang & Ehrenreich, 2012). We obtained and sequenced the isolate CC-3268, since it was not included in previous studies. A total of 31 isolates (CC-3059 – CC-3089 in the collection), sampled in 1993/94 from two sets of fields ~80 km apart in Quebec (Farnham and MacDonald College), were first screened by Sanger sequencing of introns VI and VII of the *YPT4* gene, which are species-specific markers in volvocine algae (Liss, Kirk, Beyser, & Fabry, 1997). Eighteen isolates were confirmed as authentic *C. reinhardtii*, sequencing of which was described by Ness, Kraemer, Colegrave, and Keightley (2016). A further eight previously undescribed isolates (referred to as GB# in this study) were sampled from Farnham in 2016, using the protocol of Sack et al. (1994).

Data produced by Gallaher, Fitz-Gibbon, Glaesener, Pellegrini, and Merchant (2015) for the laboratory strains CC-1009 and CC-1010, which are descendants of the original isolation of *C. reinhardtii* in Massachusetts 1945, were also included. As all laboratory strains are hypothesised to have been derived from a single zygospore, the genomes of these strains consist of two parental haplotypes, although across all strains ~75% of the genome appears to have originated from one parent (Gallaher et al., 2015). CC-1009 and CC-1010 have inherited opposite parental haplotypes, and so together maximise the genetic variation present amongst the laboratory strains. Both strains were included in the analyses of population structure and admixture, where they can be analysed as genetically distinct at ~25% of genomic sites. For analyses where the independence of isolates was required (i.e. the calculation of population genetics statistics and the identification of identity by descent tracts), CC-1009 was excluded.

For the 2016 Farnham isolates and CC-3268, DNA was extracted by phenol-chloroform extraction following Ness, Morgan, Colegrave, and Keightley (2012). Whole-genome re-sequencing was performed on the Illumina HiSeq 2000 platform (100 bp paired-end reads) for the Farnham isolates, and on the Illumina Hiseq 4000 platform (150 bp paired-end) for CC-3268, both at BGI Hong Kong. The modified PCR conditions of Aird et al. (2011) were used during library preparation to accommodate the high GC-content of *C. reinhardtii* (mean nuclear GC = 64.1%). The Japanese isolates NIES-2463 and NIES-2464 were sequenced using the Illumina MiSeq platform (300 bp paired-end), full details of which will be presented elsewhere (Arakawa et al., manuscript in preparation).

Read mapping and variant calling

Read mapping and variant calling were performed as described by Ness et al. (2016). Briefly, reads were mapped to version 5.3 of the *C. reinhardtii* reference genome (Merchant et al., 2007) using the Burrows-Wheeler Aligner (BWA) v0.7.5a-r405 (Li & Durbin, 2009), using BWA-MEM with default settings. The plastid (NCBI accession NC_005353) and mitochondrial (NCBI accession NC_001638) genomes were appended to the reference, as was the minus mating type (*mt-*) locus (NCBI accession GU814015), since the reference genome isolate is *mt+*. Genotypes were called using the GATK v3.5 (DePristo et al., 2011) tool HaplotypeCaller, and the resulting per isolate Genomic Variant Call Files (gVCF) were combined to a species-wide Variant Call File (VCF) using GenotpyeGVCFs with the following non-default settings: sample_ploidy=1, includeNonVariantSites=true, heterozygosity=0.02, indel_heterozygosity=0.002.

Only invariant and biallelic sites were considered for analyses. Filters were applied independently on the genotype calls of each isolate, as opposed to per site. Retained genotypes required a minimum of three mapped reads, with the total depth not exceeding the average depth for the isolate in question plus four times the square root of the average depth (to remove regions with copy number variation (Li (2014)). Genotypes flanking 5 bp either side of an INDEL were filtered, to avoid false positives due to misaligned reads. Single nucleotide polymorphisms (SNPs) with a genotype quality (GQ) <20, or with <90% of the informative reads supporting the called genotype, were filtered. All sites from the ~600kb *mt+* (between the *NIC7* and *THI10* genes (De Hoff et al., 2013)) and *mt-* loci were filtered. For the population structure analyses no missing genotype data were allowed, resulting in the analysis of 1.44 million SNPs. For analyses comparing the different identified *C. reinhardtii* lineages (see *Results*), to maximise the number of callable sites a minimum of 50% of isolates within each lineage were required to have genotypes that passed filtering (with the exception of the Japanese isolates, where both were required), resulting in the analysis of 58.0% of sites genome-wide (61.77 Mb) and 74.4% of 4-fold degenerate sites (6.18 Mb).

Genomic site class annotations

Genomic coordinates for coding sequence (CDS) were downloaded for the *C. reinhardtii* genome annotation v5.3 from Phytozome (https://phytozome.jgi.doe.gov/pz/). Within CDS, 0-fold (0D) and 4-fold degenerate sites (4D) were defined relative to the reference genome. All "N" bases in the reference genome (~4 Mb) were removed. Any codons that overlapped more than one reading frame, or that contained more than one SNP, were filtered due to the difficulty in determining the degeneracy of sites in such cases.

Population structure analyses

To characterise patterns of species-wide populations structure, we used the haplotype-based method fineSTRUCTURE (Lawson, Hellenthal, Myers, & Falush, 2012). This approach utilises all variant sites, first using the Chromopainter algorithm to "paint" the chromosomes of every individual (the recipients) as a combination of haplotypes from all other individuals (the donors), so that the sites within each recipient haplotype coalesce most recently with the donor. This information can be plotted as a highly informative coancestry matrix (a heatmap summarising the number of haplotypes shared between all donor-recipient pairs), and is also used to probabilistically assign individuals to populations. fineSTRUCTURE v2.1.3 was run in "linked" mode, using the flag "-ploidy 1", and otherwise default parameters. Genetic distances between each SNP were calculated assuming a uniform recombination rate, based on the genome-wide estimate of 1.2 x 10^{-5} cM/bp obtained by Liu et al. (2018) from whole-genome re-sequencing of the progeny of crosses between the field isolates CC-2935 and CC-2936. Population structure was interpreted solely based on the coancestry matrix, as fineSTRUCTURE did not cluster isolates effectively into populations. This is likely due to extensive linkage disequilibrium (LD) and the low number of isolates, resulting in nearly all of the isolates exhibiting a unique relationship to each other in terms of genetic ancestry. As a secondary method, we also ran STRUCTURE (Falush, Stephens, & Pritchard, 2003; Pritchard, Stephens, & Donnelly, 2000), details of which are presented in the supplementary text.

As a complementary approach to visualise multilocus patterns of genetic similarity between isolates, a principal component analysis (PCA) was performed on 4D SNPs subsampled every 20 kb, based on the average decay of LD in *C. reinhardtii* (Flowers et al., 2015), using the R packages SNPRelate v1.8.0 and gdsfmt v1.10.1 (Zheng et al., 2012). A neighbour joining tree was produced using MEGA v7.0.26 (Kumar, Stecher, & Tamura, 2016) from all 4D sites, using the Tamura-Nei substitution model, and 1000 bootstrap replicates. To test for the presence of isolation by distance within the two identified N. American lineages (NA1 and NA2), a Mantel test (n=999 permutations) was performed independently for each lineage on a pairwise matrix of 4D genetic distance (calculated using MEGA, Tamura-Nei model) and geographic distance, using vegan v2.4-5 (Oksanen et al., 2017).

Mitochondrial and plastid haplotype networks

To explore patterns of population structure using the *C. reinhardtii* organelle genomes, sites that passed filtering were extracted for the mitochondrial genome (7.39 kb) and plastid CDS (18.25 kb).

PopART (Leigh & Bryant, 2015) was used to produce haplotype networks for each organelle using the TCS algorithm (Clement, Snell, & Walker, 2002). As the plastid genome is known to recombine in *C. reinhardtii* (Dürrenberger, Thompson, Herrin, & Rochaix, 1996; Ness et al., 2016), a haplotype based approach is suboptimal. However, given the short length (~204 kb) and low genetic diversity of the plastid genome (Ness et al., 2016), there was insufficient power to perform similar population structure/admixture analyses to those performed on the nuclear genome. There is no evidence that the mitochondrial genome recombines in *C. reinhardtii* (Hasan, Duggal, & Ness, 2019).

Admixture profiling and identification of putatively introgressed genomic regions

Following the signatures of admixture observed from the population structure analyses, we applied an *ad hoc* approach to identify and visualise putatively introgressed genomic regions derived from admixture between NA1 and NA2 individuals. Marker SNPs were assigned to each lineage by identifying sites where the within-lineage consensus allele (defined as an allele with ≥ 60% frequency) differed between the two lineages. This resulted in a total of 758,420 marker SNPs, or on average ~135 SNPs per 20 kb. For each isolate, the proportions of marker SNPs matching the NA1 or NA2 consensus were then calculated in 20 kb sliding windows (with 4 kb increments). Intervals of at least five overlapping windows exhibiting a majority of marker SNPs for the alternate lineage to which the isolate belonged were then merged to form putatively introgressed genomic intervals. To visualise the admixture analysis, for each isolate in discrete 20 kb windows the proportions of SNPs with NA1 and NA2 identities were plotted as a heat map along each chromosome.

Identification of genomic tracts inherited identical by descent

To quantify relatedness between isolates, we identified genomic tracts that are likely to have been inherited without recombination from a common ancestor (i.e. identical by descent) using the haploid-specific hidden Markov model hmmIBD (Schaffner, Taylor, Wong, Wirth, & Neafsey, 2018). This approach infers identical by descent tracts shared between pairs of individuals as genomic regions that are identical by state (allowing for genotyping error), based on SNP allele frequencies, the distance between SNPs in bases, and a genome-wide recombination rate. Additionally, the program estimates the expected proportion of the genome inherited identical by descent between pairs ($\hat{\pi}_{IBD}$) based on the average per-SNP probability of identity by descent, independent of the designation of tracts (Taylor et al., 2017). hmmIBD was run independently for each N. American lineage (NA1/NA2), assuming a recombination rate of 1.2 x 10^{-5} cM/bp (Liu et al., 2018) and otherwise default parameters. As we observed that the majority of identified tracts were within the range of the decay of LD in *C. reinhardtii* (~20 kb), tract length filters of >100 kb (~1.2 cM) and >500 kb (~6.0 cM) were applied. Identical by descent tracts have recently been defined using similar length cut-offs to explore population-level tract sharing (Wakeley & Wilton, 2016). Following Carmi et al. (2013), the cohort-averaged sharing was calculated for each isolate as the mean proportion of the genome shared identical by descent between the isolate in question and all other isolates in the sample.

Calculation of population genetics statistics within and between lineages

Genetic diversity was calculated as the average number of pairwise differences per site (π, Nei and Li (1979)) for each of the lineages (NA1/NA2/JPN), and for each sampling site and time point containing two or more isolates. As a measure of differentiation, F_{st} was calculated between each lineage using the approach of Hudson, Slatkin, and Maddison (1992), where within-population π was calculated as an unweighted mean of π for the two lineages in the comparison. As a measure of genetic distance between-lineages*,* we calculated the number of pairwise differences between two random sequences drawn from each lineage (d_{xy} Nei and Li (1979)).The proportions of fixed, shared and private polymorphisms were calculated for each between lineage comparison. All calculations were performed using custom Perl scripts.

Results

Whole-genome re-sequencing of Chlamydomonas reinhardtii field isolates

The species-wide sample consisted of 42 isolates, sampled from 11 sites/time points (fig. 1, detailed sampling and sequencing information table S1). Three isolate pairs and one isolate trio, all of which were sampled in Quebec, were found to be clonal (supplementary text, table S2). Although each isolate was derived from an independent soil sample, all identified clone mates were sampled at the same site and time, which has been observed previously in the case of the clonal pair CC-1952 and CC-2290 (Jang & Ehrenreich, 2012). Additionally, CC-3078 was found to be identical to the laboratory strain CC-1010, which was used in mating trials at the time of sampling (Sack et al., 1994) and therefore likely replaced the original isolate at that time. An additional 12 isolates, sampled in Quebec 1993/94, were found not to be *C. reinhardtii* (supplementary text, table S3). After retaining only one isolate for each clonal pair/trio, the final species-wide dataset comprised 36 isolates and 5.88 million SNPs, with $\pi_{\text{genome-wide}} = 0.0210$, $\pi_{4D} = 0.0288$, and $\pi_{0D} = 0.00657$. To our knowledge, this dataset encompasses all genetically-unique field isolates of *C. reinhardtii* (supplementary text).

Patterns of continental population structure

The species-wide analyses of population structure indicated that genetic variation in *C. reinhardtii* is geographically partitioned both between N. America and Japan, and within N. America. The neighbour joining tree (fig. 2a) and PCA (fig. 2b) were consistent with all isolates clustering as three distinct lineages, (i) a north eastern N. American lineage (NA1, 27 isolates) comprising the Massachusetts isolates and all Quebec isolates except CC-3079, (ii) an approximately Midwest/Mid-Atlantic/South USA lineage (NA2, eight isolates) comprising all isolates from Pennsylvania, North Carolina, Minnesota and Florida, as well as CC-3079, and (iii) a Japanese lineage (JPN) comprising both isolates from Kagoshima Prefecture, Japan. The N. American lineages were broadly consistent with the two groups described by Jang and Ehrenreich (2012), and our designation of these as NA1 and NA2 follows their previous labelling as group 1 and 2. The geographic distinction between NA1 and NA2 was most clearly shown by the genetic similarity of the Massachusetts and Quebec isolates (sampled ~320-350 km north), relative to the larger genetic distances observed between the Massachusetts isolates and CC-2344 (isolated only ~380 km south west, site PA2 in figure 1). The grouping of a single Quebec isolate, CC-3079, with NA2, was the only anomaly between these geographic groups, potentially indicating a recent migration event (see below).

The coancestry matrix produced by fineSTRUCTURE corroborated the above results, with all isolates sharing many more haplotypes in within-lineage recipient-donor pairs, than in between-lineage pairs (fig. 2c). However, the patterns of haplotype sharing in both between- and within-lineage comparisons were not homogenous. There was evident sub-structure within NA2, with the North Carolina isolates clearly more closely related to each other than to the remaining NA2 isolates. Similar patterns of close relatedness were also evident within NA1 for several Quebec pairs. The between-lineage heterogeneity was indicative of admixture between NA1 and NA2 isolates. Specifically, a subset of NA1 isolates, marked by the dashed blue square in figure 2c, were the recipients of a greater number of NA2 haplotypes than the remaining NA1 isolates. The NA2 isolates CC-2344 and CC-3079 were the most frequent donors to NA1 isolates, which is notable given that they were sampled in the closest geographic proximity to Massachusetts/Quebec. The STRUCTURE analysis was congruent with admixture, with the majority of NA1 isolates (and in particular the subset outlined above) and CC-2344/CC-3079 appearing as admixed between the ancestral populations corresponding to NA1 and NA2 (fig. 2c/S1, supplementary text). Additionally, admixture potentially explained the variation on the first principal component of the PCA (fig. 2b), where NA1 axis coordinates were strongly correlated with the estimated proportion of introgressed genome from NA2 (see below) (R = 0.920, $p < 0.01$). A role for admixture was also supported by mitochondrial (fig. S2a) and plastid (fig. S2b) haplotype networks, although the patterns of population structure observed from the organelles were generally far less clear (supplementary text).

Finally, there was evidence for isolation by distance between NA2 isolates (Mantel's r^2 = 0.52, p = 0.01), but no significant pattern between NA1 isolates (fig. 3). A pattern of isolation by distance is consistent with the larger geographic range of the NA2 lineage, and the population sub-structure indicated by the fineSTRUCTURE analysis. Given the sparsity of sampling for this group, little can currently be concluded about the extent to which these isolates can be treated as a single evolutionary lineage.

Admixture profiling and identification of putatively introgressed genomic regions

To further explore the possibility of ongoing admixture between NA1 and NA2, local ancestry was profiled for each isolate. The proportions of marker SNPs matching either the NA1 or NA2 consensus alleles for each isolate in 20 kb windows were plotted as a heat map along each chromosome (chromosome 3 fig. 4a, all chromosomes fig. S3). For all NA1 isolates, large haplotype blocks indicative of recent introgression from NA2 were observed, and the total proportion of introgressed genome per NA1 isolate ranged from 5.4% to 21.9% (mean 12.7%, fig. 4b). The NA1 isolates designated as highly admixed from the fineSTRUCTURE analysis were found to have significantly more introgressed sequence than the remaining NA1 isolates (means 17.3% and 9.0%, respectively; Wilcoxon rank sum test, $W = 180$, $p = <0.01$), and in practice this categorical division separated the isolates into two groups with less than or greater than 15 Mb of introgressed sequence (~14% of the genome). The mean proportion of introgressed genome for NA2 isolates was lower at 7.7%, with only CC-3079 (17.6%) and CC-2344 (14.9%) exhibiting similarly substantial signatures of admixture. However, this does not necessarily imply that introgression from NA2 to NA1 is more prevalent than in the opposite direction, given that the current sampling of NA2 isolates is so limited, and that highly-admixed NA2 populations in close proximity to Massachusetts/Quebec may exist.

A mosaic pattern was observed across the genome of CC-3079, where on many chromosomes megabase-scale NA1 haplotypes were interspersed on an NA2 genomic background (e.g. chromosomes 3, 4, 6, 7, and 9) (fig. S4). However, far shorter transitions between NA1- and NA2-like sequences were also observed, conceivably due to older admixture events. Given that CC-3079 was the only NA2 isolate sampled in Quebec, it is surprising that only 17.6% of the genome was identified as introgressed. Indeed, some chromosomes (e.g. 1, 8, 10 and 16) had no NA1 haplotypes of a size indicative of very recent admixture. Such a pattern of introgression is consistent with at least one admixture event a small number of sexual generations in the past, although assuming all chromosomes undergo at least one crossover per meiosis, the presence of entirely NA2-like chromosomes suggests further mating with NA2 individuals since the putative admixture event(s). From the fineSTRUCTURE analysis, CC-3079 was most closely related to the Minnesota and Pennsylvania isolates, potentially indicating a northern source population from which a migration event could have occurred.

Identity by descent sharing and patterns of local population structure in the Quebec sample

To further explore patterns of relatedness within our sample, we used hmmIBD (Schaffner et al., 2018) to identify identical by descent tracts shared between pairs of isolates. The proportion of the genome shared identical by descent between each isolate pair (i.e. the total sharing) was then estimated using three metrics (i) $\hat{\pi}_{IBD}$, the total sharing estimated directly by hmmIBD from the average per-SNP probability of identity by descent, (ii) total sharing for tracts >100 kb, and (iii) total sharing for tracts >500 kb. The estimates differed substantially between metrics, since the absence of shorter tracts in the >100 kb and >500 kb datasets resulted in lower total sharing relative to $\hat{\pi}_{IBD}$ (table 1, fig. 5a for NA1 only). However, all three metrics were significantly and highly correlated (R = 0.848 – 0.968), and the interpretation of results was consistent across metrics, so the following results are given for tracts >100 kb.

As indicated by the fineSTRUCTURE analysis, there was substantial variation in relatedness between pairs within both NA1 and NA2. Across all NA1 pairs, the distribution of total sharing for tracts >100 kb was approximately normal, although a long tail of the distribution indicated the presence of pairs with a higher genomic fraction of shared tracts (fig. 5a). Total sharing was greater than zero for all

325 NA1 pairs (range 0.3% – 52.0%), and was 9.1% on average, an unexpectedly high figure given the very large effective population size of *C. reinhardtii* (see *Discussion*). The variation between isolate pairs may partly be explained by variation in admixture, since introgression is expected to reduce total sharing (Carmi et al., 2013). As expected under this scenario, the cohort-averaged sharing (a per isolate identity by descent summary statistic) for NA1 isolates was significantly negatively correlated with the inferred proportion of introgressed genome from NA2 (R = -0.675, p < 0.01). There was no signature that identical by descent tracts were highly concentrated in particular genomic regions, as ~99% of the genome was included in at least one pairwise tract, and the distribution of the average sharing across all NA1 pairs in 100 kb chromosomal windows was approximately normal (fig. 5b).

Given the prevalence of identity by descent tracts in NA1, it is unclear to what extent total sharing can be used as a proxy for relatedness. Nonetheless, following the assumption that the total sharing is at least partially indicative of the relatedness between a pair of isolates, this relationship can be used to explore local population structure within NA1, and specifically within Quebec. If genetic diversity is spatially or temporally structured at local scales in *C. reinhardtii*, it is expected that total sharing would be higher for within-site isolate pairs (Farnham and MacDonald College, ~80 km apart) relative to between-site pairs, and for within-time point pairs at the same site (Farnham 1993 and 2016) relative to between-time point pairs. There was, however, no support for either of these relationships, with no difference in total sharing for within-site pairs relative to between-site pairs (Wilcoxon rank sum test, W = 2228, p = 0.23), and no difference for within-time point pairs relative to between-time point pairs (Wilcoxon rank sum test, W = 5859, p = 0.40). Moreover, there was also no difference in total sharing for pairs within Quebec and Massachusetts, relative to pairs between Quebec and Massachusetts (Wilcoxon test rank sum test, W = 7054, p = 0.50), where the isolates were sampled ~320-350 km and ~50-70 years apart. Therefore, taken together with the lack of isolation by distance, there appears to be no strong signal of population structure within the current sampling of NA1.

Conversely, there were differences between the samples, with the average total sharing within MacDonald College 1994 (20.8%) and Farnham 2016 (17.3%) more than twice that of Farnham 1993 (7.1%). Samples with greater average total sharing exhibited lower putatively neutral genetic diversity (π_{4D}), resulting in the observation that diversity was marginally higher within a single sample (Farnham 1993 $\pi_{4D} = 0.0242$) than within the entire sampled lineage (NA1 $\pi_{4D} = 0.0236$, table 1). The lower average total sharing within Farnham 1993 may be explained by an increased rate of admixture within this sample, as the average proportion of introgressed genome was higher (14.8%) relative to MacDonald College 1994 (7.1%) and Farnham 2016 (12.8%) (fig. 4b). The Farnham 1993 isolate pairs make up the majority of the within-sample pairs in the above within vs between sample statistical comparisons, so the reduction in total sharing for this sample may explain the reported lack of significance. Regardless of this, the average total sharing between Farnham and MacDonald College (8.3%), and between Farnham 1993 and 2016 (8.0%), remain far greater than would be expected if there was strong spatial or temporal structure within Quebec.

In contrast to NA1, there was very little signature of close relatedness between NA2 isolates from different locations. Total sharing for between-location NA2 pairs was only 0.2% on average (table 1), corroborating the presence of population sub-structure in the lineage. However, within the North Carolina sample (the only site with more than one NA2 isolate), the average total sharing was 23.2%. Taken together with the results for NA1, the independent finding of very high total sharing between North Carolina isolate pairs suggests that *C. reinhardtii* haploid individuals may generally share a substantial proportion of their genomes identical by descent at local scales.

Genetic diversity within lineage, and genetic differentiation and divergence between lineages

Genetic diversity varied substantially between lineages (fig. 6a), with π_{4D} estimates of 0.0236, 0.0306, and 0.00123 for NA1, NA2, and JPN, respectively. Based on these estimates of putatively neutral diversity and a SNP mutation rate of 9.63 x 10^{-10} per site per generation estimated by re-sequencing of *C. reinhardtii* mutation accumulation lines by Ness, Morgan, Vasanthakrishnan, Colegrave, and

Keightley (2015), the estimated effective population sizes (N_e) for each lineage were 4.91 x 10^7 (NA1), 6.35 x 10^7 (NA2), and 2.56 x 10^6 (JPN) (following $\pi = 2N_e\mu$.) Thus, at least for the N. American lineages, these estimates are consistent with *C. reinhardtii* genetic diversity being amongst the highest reported in eukaryotes (Leffler et al., 2012). It is difficult to conclude to what extent the higher diversity of NA2 relative to NA1 reflects the sampling history of the species, since the NA2 isolates have been sampled over a far larger area with generally only one isolate per site (with the exception of the three NC isolates). Indeed, considering single sampling locations, π_{4D} estimated for only the three North Carolina isolates was 0.0190, lower than that calculated for the Farnham 1993 isolates (0.0242), and marginally lower than that for the three MacDonald College NA1 isolates (0.0193), which have a comparable incidence of identity by descent sharing as the North Carolina isolates (table 1).

Strikingly, genetic diversity for JPN was an order of magnitude lower than that for the N. American lineages, with the estimated π_{4D} of 0.00123 approximately 19 and 25 times lower than the estimated values for NA1 and NA2, respectively. Although based only on two isolates, this did not appear to be an artefact caused by high relatedness. Firstly, the isolates are of opposite mating types, and so are certainly not clonal. Secondly, genetic diversity appeared to be uniformly lower across the genome relative to N. American isolates, with no obvious long invariant tracts as observed for pairs of NA1 isolates (fig. 6b). Indeed, even for the extreme of highly related isolate pairs (e.g. GB119 and GB141, sharing ~50% of their genomes), and for the laboratory strains CC-1009 and CC-1010 (sharing ~75% of their genomes), pairwise genetic distances greatly exceeded that observed between the two JPN isolates (as shown by the branch lengths of the neighbour joining tree, fig. 2a).

The NA1 and NA2 lineages were highly differentiated, both genome-wide ($F_{st} = 0.25$) and at putatively neutral 4D sites ($F_{st} = 0.24$) (table 2). Only 30.6% of the 7.19 million SNPs segregating in the N. America sample were shared between the lineages, with 37.3% private to NA1, and 31.8% private to NA2. Results were similar for 4D SNPs, with a slightly higher percentage shared between the lineages (33.0%). Despite the majority of SNPs being private to either lineage, only 0.3%

(genome-wide) and 0.2% (4D) of SNPs were fixed, consistent both with admixture and the expected weak force of genetic drift due to the high effective population size of the species. The average number of pairwise differences between the lineages (d_{xy}) was estimated as 0.0274 (genome-wide) and 0.0364 (4D), and thus two sequences drawn randomly between NA1 and NA2 contained 54.2% more differences than two NA1 sequences, and 19.0% more differences than two NA2 sequences (for 4D sites, based on comparison to within-lineage π_{4D}). After masking introgressed regions for both lineages, the overall percentage of shared SNPs decreased to 19.8% and 22.6%, F_{st} increased to 0.34 and 0.32, and d_{xy} increased to 0.0281 and 0.0374 (all for genome-wide and 4D sites, respectively). Surprisingly, the JPN lineage was no more genetically distant from NA1 (4D d_{xy} = 0.0343) and NA2 (4D d_{xy} = 0.0376), than NA1 and NA2 were from each other.

Discussion

In this study we have used genome-wide data to explore patterns of population structure across field isolates of *C. reinhardtii*. Taking advantage of the haploid state of the isolates, we have applied haplotype-based analyses to characterise structure at both continental and local scales, and to infer patterns of admixture between the two identified N. American lineages. In what follows, we contextualise these findings within the ongoing debate concerning the nature of biogeography and speciation in microbial eukaryotes, and discuss further insights concerning the evolutionary history and ecology of *C. reinhardtii*. Finally, we discuss the surprising prevalence of identity by descent sharing between isolates sampled at local scales.

The North American biogeography of Chlamydomonas reinhardtii

Based on current sampling, the evidence for three geographically distinct lineages of *C. reinhardtii* strongly contradicts the predictions of the *ubiquity model*, under which little geographic population structure is expected. Interestingly, there are notable similarities between the observed biogeography of *C. reinhardtii*, and the best studied microbial eukaryote in this context, *Saccharomyces paradoxus.* This wild yeast has been shown to form a species complex, comprising highly differentiated lineages on different continents, suggesting allopatric divergence and speciation (Koufopanou, Hughes, Bell, & Burt, 2006; Kuehne, Murphy, Francis, & Sniegowski, 2007; Liti et al., 2009). Within N. America, two allopatric lineages of *S. paradoxus* have been described, which exhibit signatures of local adaptation and reproductive isolation characteristic of incipient species (Charron, Leducq, & Landry, 2014; Leducq et al., 2014; Leducq et al., 2016). Similar to *C. reinhardtii*, one lineage has a more restricted range in the north east, while the other is widely distributed to the south and west, with a sympatric zone occurring along Lake Ontario and the St. Lawrence River (Charron et al., 2014). This biogeography is consistent with allopatric divergence in the Atlantic and Mississippian glacial refugia during the last glacial maximum (~110,000 – 12,000 year ago), which has been documented in numerous plants and animals (Charron et al., 2014). Thus, although as a morphological entity *S. paradoxus* fulfils the '*everything is everywhere'* maxim, it in fact consists of several cryptic species that have undergone allopatric speciation events, including a putative event in glacial refugia contemporaneous with several plants and animals.

Whether glacial refugia can explain the biogeography of the two N. American *C. reinhardtii* lineages will largely be contingent on further sampling, especially in what would be expected to be the north eastern limits of the NA2 range (i.e. south west of New England and the St. Lawrence River). However, the observed biogeography is consistent with such a scenario, under which NA1 would have persisted in the Atlantic regufium (located east of the Appalachians), before re-colonising Massachusetts and Quebec. This could also explain the sub-structure observed for NA2, which may have a markedly different evolutionary history to NA1, with the possibility of multiple refugia (e.g. Mississippian, Virginia/Carolinas Atlantic coast, and further south) connected by varying amounts of gene flow at different times. Furthermore, the two lineages cannot easily be explained by climate or other environmental factors, since NA2 includes both one of the most northerly (CC-1952, Minnesota) and the most southerly (CC-2343, Florida) isolates, and the Massachusetts and

Pennsylvania sites presumably share similar environments. However, we have not explicitly tested any environmental variables in this study, and this will form an important aspect of future research.

That essentially all NA1 isolates exhibit signatures of admixture with NA2 individuals supports a role for substantial dispersal in *C. reinhardtii*. Given that the length of the observed introgressed haplotypes are considerably longer than the physical distance over which LD decays in the species, admixture is likely to have occurred in the relatively recent past. Furthermore, that a single highly-admixed NA2 isolate (CC-3079) was present within our small Quebec sample suggests that both migration and gene flow are ongoing. Under such a scenario, that the two lineages remain so highly differentiated in the face of migration and gene flow potentially indicates the presence of reproductive isolation and/or local adaptation. However, there is currently no evidence for either reproductive isolation or local adaptation in *C. reinhardtii*, and isolates of all three identified lineages cross successfully in the laboratory (Nakada et al., 2014; Pröschold et al., 2005). Nonetheless, there are substantial phenotypic differences between isolates (Flowers et al., 2015), and it should be possible to re-visit such variation in the context of the two N. American lineages, and to further test for reproductive isolation in the laboratory (e.g. via fitness assays of 'hybrid' progeny).

The mosaic genome of CC-3079 also provides further insights into the ecology of *C. reinhardtii*. The observed pattern cannot simply be explained by an NA2 migrant arriving in Quebec and subsequently mating with only NA1 individuals, as several chromosomes show no signature of recent introgression, implying that mating between other NA2 individuals occurred after the inferred admixture event(s). This could be explained if CC-3079 were itself a migrant from an unsampled location in which both NA1 and NA2 individuals occur in sympatry and hybridise. Alternatively, an NA2 ancestor of CC-3079 may have migrated to Quebec, implying the presence of other NA2 individuals at the site. Almost nothing is known about the dispersal capability and mechanisms in *C. reinhardtii*, although there is abundant evidence for the passive dispersal of dormant propagules (such as the *C. reinhardtii* zygospore) of various species (De Meester, Gómez, Okamura, & Schwenk, 2002). As such propagules are resistant to environmental stresses, they can be transported over long distances via biotic (e.g. birds and insects), abiotic (e.g. wind and water), or anthropomorphic vectors. Additionally, as *C. reinhardtii* zygospores adhere to each other (Harris, 2008), a single migration event may have the potential to introduce many migrant individuals of both mating types, which could explain the implied presence of other NA2 individuals at the sampling site.

The Japanese isolates and the wider biogeography of Chlamydomonas reinhardtii

Although the evolutionary history of the Japanese isolates is essentially unresolved based on current sampling, their inclusion in this study at least indicates that *C. reinhardtii* on different continents may be expected to form substantially divergent lineages. However, under a model of allopatric divergence between N. American and Japanese *C. reinhardtii,* it is surprising that the JPN lineage is no more genetically distinct from either NA1 or NA2, than NA1 and NA2 are from each other. One speculative explanation is that the Japanese isolates were derived from a third unsampled N. American lineage that underwent divergence from NA1 and NA2 simultaneously (e.g. in Pacific or Beringian refugia), before migration to Japan. Water birds are thought to be a major mechanism of algal dispersal (Kristiansen, 1996), and western N. America, and in particular Alaska, is linked to Japan by the flyways of several migratory bird species. Alternatively, gradual dispersal across the Bering land bridge could also give rise to a similar pattern, leading to the prediction that any East Asian and Alaskan *C. reinhardtii* may be genetically similar. The strikingly low genetic diversity of the two Japanese isolates relative to the N. American lineages is also surprising. If the lineage was established from a larger population by migration (which could in principal occur from a single zygospore), then such a founder effect would be expected to reduce diversity via a severe bottleneck (De Meester et al., 2002). Supporting this hypothesis, any population present in Kagoshima must be geologically young, as a result of the formation of the Aira Caldera ~30,000 years ago, and the Akahoya eruption ~7,000 years ago (Machida & Arai, 2003).

As a result of the historic difficulty in isolating *C. reinhardtii* (Pröschold et al., 2005), it is likely that the current sampling primarily reflects the distribution of researchers. Intercontinental distributions of more conspicuous Volvocalean algae have been documented (e.g. Kawasaki, Nakada, and Tomita (2015)), and given the geographic distance between eastern N. America and Japan, it would not be surprising if *C. reinhardtii* is shown to have a considerably wider distribution in the future. However, far more extensive sampling across multiple regions and habitats, alongside improvements in sampling methodology, will be required to address this.

Patterns of population structure and genetic diversity at a local scale

In facultatively sexual organisms, under certain conditions clonal erosion can generate population structure and reduce genetic diversity at local scales (Vanoverbeke & De Meester, 2010). Prior to this study, almost nothing was known about the local structure of genetic diversity in *C. reinhardtii,* and it was unknown whether a single site would be dominated by clonal lineages. Although our sample contained a small number of clonal pairs/trios, the majority of isolates sampled at single sites were genetically distinct, and diversity at single sites and time points was of the same magnitude as the total lineage diversity. Although the extent of identity by descent sharing appeared to vary between sites and time points in Quebec, we found no evidence for strong population structure at this scale. The lack of structure observed in space further supports the considerable dispersal potential of *C. reinhardtii*. The lack of structure observed in time could potentially be explained by long-term zygospore dormancy, which would result in isolates sampled many years apart being separated by far fewer sexual generations than would otherwise be expected. Such a phenomenon is known in other chlorophyte algae, where dormant zygospores are capable of forming propagule banks (Fryxell, 1983), and it is known that *C. reinhardtii* zygospores are resistant to both long-term freezing and desiccation (Harris, 2008). Propagule banks have also been hypothesised to contribute to high levels of genetic diversity, as populations can be re-seeded with haplotypes present at previous time points (Rengefors, Kremp, Reusch, & Wood, 2017; Shoemaker & Lennon, 2018), and therefore long-term zygospore dormancy could be a contributing factor to the high diversity estimated for *C. reinhardtii*.

As detailed previously, *C. reinhardtii* population genetics analyses have been hindered by the absence of a suitable set of isolates, and the lack of understanding as to what constitutes a 'population' in the species. The high genetic diversity found at single sites in this study now presents the opportunity to use samples from single sites (e.g. Farnham 1993) for future analyses. Furthermore, given the lack of structure between sites/time points, the entire Quebec sample could conceivably be analysed together. Although the extent of identity by descent sharing between these isolates requires further explanation (see below), the delineation of a group of isolates suitable for population genetics analyses has the potential to greatly enhance the use of *C. reinhardtii* in evolutionary biology research.

Broader perspectives on microbial biogeography and speciation

Taken together with the evolutionary history of *S. paradoxus*, our interpretation of *C. reinhardtii* continental population structure supports a role for allopatric differentiation (and potentially speciation) in microbial eukaryotes. This permits the rejection of the *ubiquity model* in these cases, supporting the more similar rates of speciation between microbial eukaryotes and macroorganisms predicted by the *moderate endemicity model*, and implying that microbial species may be far more speciose than existing taxonomic descriptions suggest. It is worth noting, however, that the *moderate endemicity model* does not predict frequent allopatric speciation (instead favouring various forms of non-allopatric speciation) (Foissner, 2008), and in this sense the model may need to be revised. De Meester et al. (2002) detailed the role of glacial refugia in speciation events for various zooplankton, and it may be that similar allopatric events are also commonplace in microbial eukaryotes. However, it is unclear to what extent the results for two terrestrial species can be extrapolated, and the exploration of similar patterns across a far larger range of species is obviously required to fully address this question.

The extent of identity by descent sharing between Chlamydomonas reinhardtii isolates

The original motivation for identifying identical by descent tracts was to quantify between-pair relatedness and explore patterns of local population structure. However, the most surprising result of these analyses was the finding that on average a pair of NA1 isolates share 9.1% of their genomes in tracts >100 kb, and that an even higher proportion was independently observed between the three isolates sampled in North Carolina. Even more unexpectedly, isolates from Massachusetts and Quebec (sampled ~50-70 years apart) share 8.6% of their genomes identical by descent on average. This highlights a striking dichotomy: how can essentially the entire sampled population appear to share recent ancestry, yet genetic diversity be maintained at a high level? Although much of our understanding of identity by descent in populations has been built upon pedigrees (Thompson, 2013), population-level theory has recently been developed for tracts defined based on arbitrary genetic length cut-offs (Carmi et al., 2013; Carmi, Wilton, Wakeley, & Pe'er, 2014; Palamara, Lencz, Darvasi, & Pe'er, 2012). Using equation 4 of Carmi et al. (2013), and based on the estimated N_e for NA1 and a minimum tract length of 100 kb (~1.2 cM), the average proportion of the genome shared identical by descent between a pair of individuals in a Wright-Fisher population is expected to be ~0.00017%, four order of magnitude lower than observed.

Although we currently lack an explanation for this discrepancy, there are a number of possibilities that can currently be considered. Firstly, *C. reinhardtii* evidently does not meet the assumptions of a Wright-Fisher population, and therefore a stochastic process may be responsible. Clonal reproduction is expected to result in a high variance in reproductive success (Tellier & Lemaire, 2014), and zygospore dormancy would result in overlapping generations, although further theoretical work will be needed to address the effects of such processes on identity by descent. Secondly, it is conceivable that many long shared genomic tracts could arise in a population as a result of pervasive positive selection combined with long-range effects of selection on linked sites. Frequent adaptive evolution and the resulting effects of hitchhiking on linked sites has recently been evoked to explain the low observed diversity in the ubiquitous phytoplankton species *Emiliania huxleyi* (Filatov, 2019).

Although *C. reinhardtii* obviously differs from this case with respect to genetic diversity, if pervasive positive selection acted mostly on standing variation in the species, it is possible that soft selective sweeps could result in multiple haplotypes rising to high frequency, while maintaining high genetic diversity. Thirdly, if there is a high diversity of structural variants segregating in *C. reinhardtii* populations there may be recombination suppression between certain haplotypes. Physical recombination has only been studied between a very small number of *C. reinhardtii* isolates (Kathir et al., 2003; Liu et al., 2018), and additional experimental work will be required to further explore recombination in the species. In a broader sense, empirical studies of other species with similar life cycles will also be crucial to determining the generality of this result.

Conclusions

C. reinhardtii is divided into three geographically distinct lineages based on current sampling, supporting the *moderate endemicity model* of microbial eukaryote biogeography. *C. reinhardtii* is likely to have substantial dispersal capability, implying that reproductive isolation and/or local adaptation may be maintaining genetic differentiation between the two N. American lineages in the face of ongoing migration and gene flow. High dispersal may also prevent the evolution of population structure at local geographic scales. Within two independent populations an extremely high incidence of identity by descent sharing was observed, raising several interesting questions regarding the evolutionary genetics of *C. reinhardtii*.

Acknowledgements

We thank Ahmed Hasan for useful discussions and comments on the manuscript, and Yuki Takai for technical assistance in sequencing. We thank three anonymous reviewers for their insightful comments on an earlier version of this manuscript. This project was funded by a grant (BB/H006109/1) from the UK Biotechnology and Biological Sciences Research Council (BBSRC), and has received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 694212). This work was also supported in part by research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan. Rory Craig is supported by a BBSRC EASTBIO Doctoral Training Partnership grant, and Rob Ness is supported by a Natural Sciences and Engineering Research Council Discovery grant and Canada Foundation for Innovation JELF.

Data accessibility

Sequencing reads for isolates sequenced in this study have been deposited under the accession numbers PRJEB33012 (ENA, N. American isolates) and PRJNA547760 (SRA, Japanese isolates). Code used to perform analyses is available at: https://github.com/rorycraig337/Chlamydomonas_reinhardtii_population_structure

Author contributions

R.J.C., N.C., P.D.K. & R.W.N. conceived the study. R.J.C., K.B.B. & R.W.N. performed analyses. K.A, T.N., T.I. & G.B. performed sampling. R.J.C., K.A. & R.W.N. performed sequencing. R.J.C. wrote the manuscript together with P.D.K & R.W.N. All authors read and commented on the final version of the manuscript.

Figures and tables

Table 1. Average proportions of the genome shared identical by descent between isolate pairs.

Population/Comparison	$\hat{\pi}_{IBD}$	**Average total sharing >100 kb tracts (%)**	**Average total sharing >500 kb tracts (%)**	π_{4D}	**Number of isolate pairs**
NA1	23.6	9.11	2.64	0.0236	325
Massachusetts	36.2	16.9	3.50	0.0188	1
Quebec	23.4	9.18	2.78	0.0237	276
Farnham 1993	22.2	7.13	1.18	0.0242	91
MacDonald College 1994	35.2	20.8	10.2	0.0193	3
Farnham 2016	29.3	17.3	9.04	0.0218	21
Massachusetts – Quebec	24.3	8.55	1.76	/	48
Farnham 1993 - MacDonald College 1994	23.2	8.31	2.52	/	42
Farnham 1993 - Farnham 2016	21.8	7.99	2.00	/	98
NA2	9.41	2.77	0.959	0.0306	28
North Carolina	32.6	23.2	8.95	0.0190	3
NA2 between-locations	0.0595	0.217	0.00	/	12

Proportions of the genome shared identical by descent (i.e. total sharing) are shown for the total predicted by hmmIBD ($\hat{\pi}_{IBD}$), for tracts >100 kb, and for tracts > 500 kb. The number of isolate pairs refers to the total number of pairwise comparisons contributing to the average total sharing. For each lineage, average total sharing is shown for the subsets of isolates discussed in the main text (e.g. North Carolina for NA2), and comparisons between subsets are labelled as the two subsets separated by a hyphen (e.g. Farnham 1993 – Farnham 2016).

Table 2. Differentiation and divergence between the three lineages (NA1 26 isolates, NA2 eight isolates, JPN two isolates).

		NA1 - NA2	**NA1- JPN**	**NA2 - JPN**	**NA1 - NA2 (introgression masked)**
SNPs	genome-wide	7,188,929	4,496,586	4,167,903	6,379,381
	4D	881,984	598,261	562,782	798,407
shared (%)	genome-wide	30.6	0.279	0.222	19.8
	4D	33.0	0.315	0.261	22.6
private A (%)	genome-wide	37.3	88.9	84.4	36.1
	4D	36.5	90.1	85.7	35.6
private B (%)	genome-wide	31.8	1.22	1.32	42.0
	4D	30.4	1.00	1.12	40.1
fixed (%)	genome-wide	0.301	9.67	14.0	2.21
	4D	0.194	8.60	12.9	1.64
F_{st}	genome-wide	0.25	0.64	0.59	0.34
	4D	0.24	0.63	0.58	0.32
d_{xy}	genome-wide	0.0274	0.0256	0.0283	0.0281
	4D	0.0364	0.0343	0.0376	0.0374

For private SNPs, A is the first lineage in the comparison, and B the second. Introgression masked refers to the NA1 – NA2 comparison after removing genomic regions identified as introgressed for each individual.

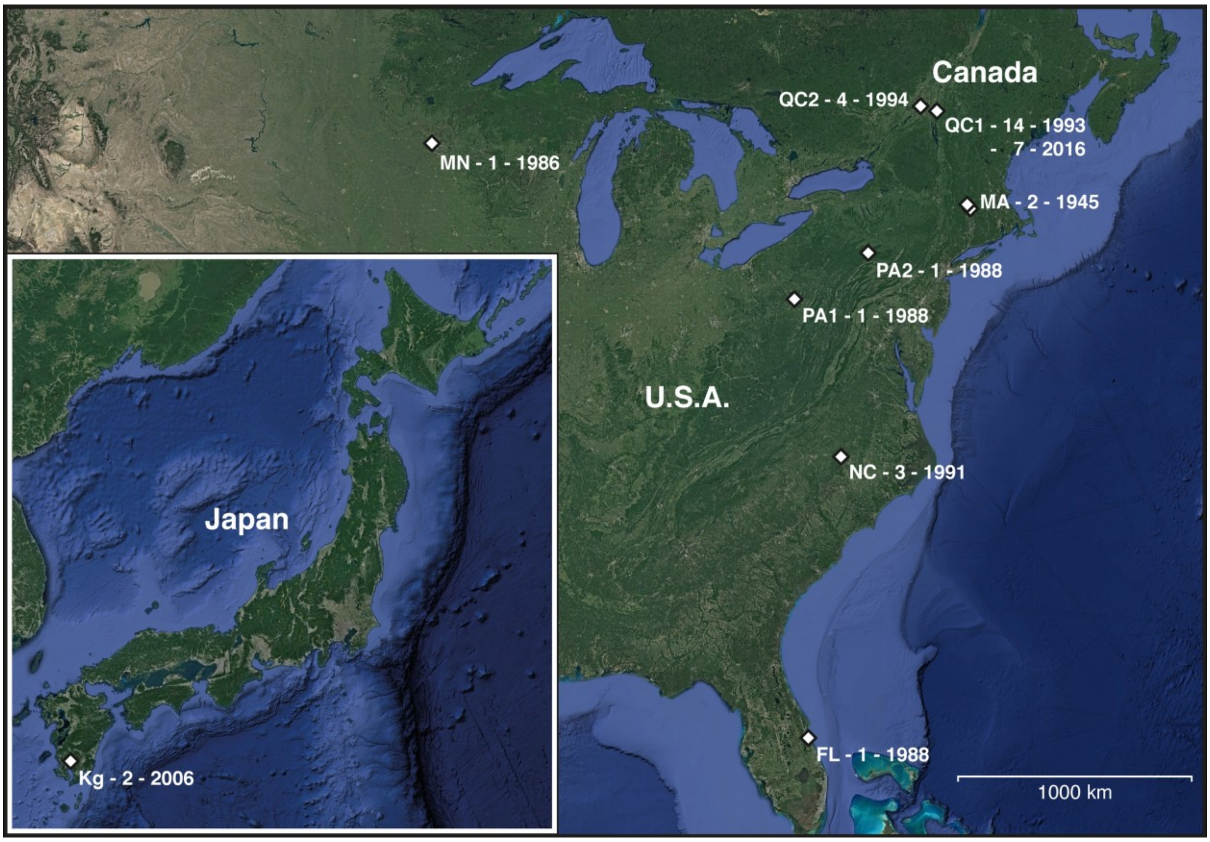

Figure 1. Sampling locations and years for all field isolates included in analyses. Format is 'site – number of isolates – year', where the number of isolates refers to genetically unique (i.e. non-clonal) samples. Location abbreviations are as follows: QC – Quebec, MA – Massachusetts, PA – Pennsylvania, NC – North Carolina, MN – Minnesota, FL – Florida, Kg – Kagoshima Prefecture. Quebec refers to two separate sites, Farnham (QC1, 21 total isolates) and MacDonald College (QC2, four isolates). The Massachusetts isolates are also from two sites ~13 km apart, and one site/isolation is represented by two laboratory strains in the species-wide dataset (see main text).

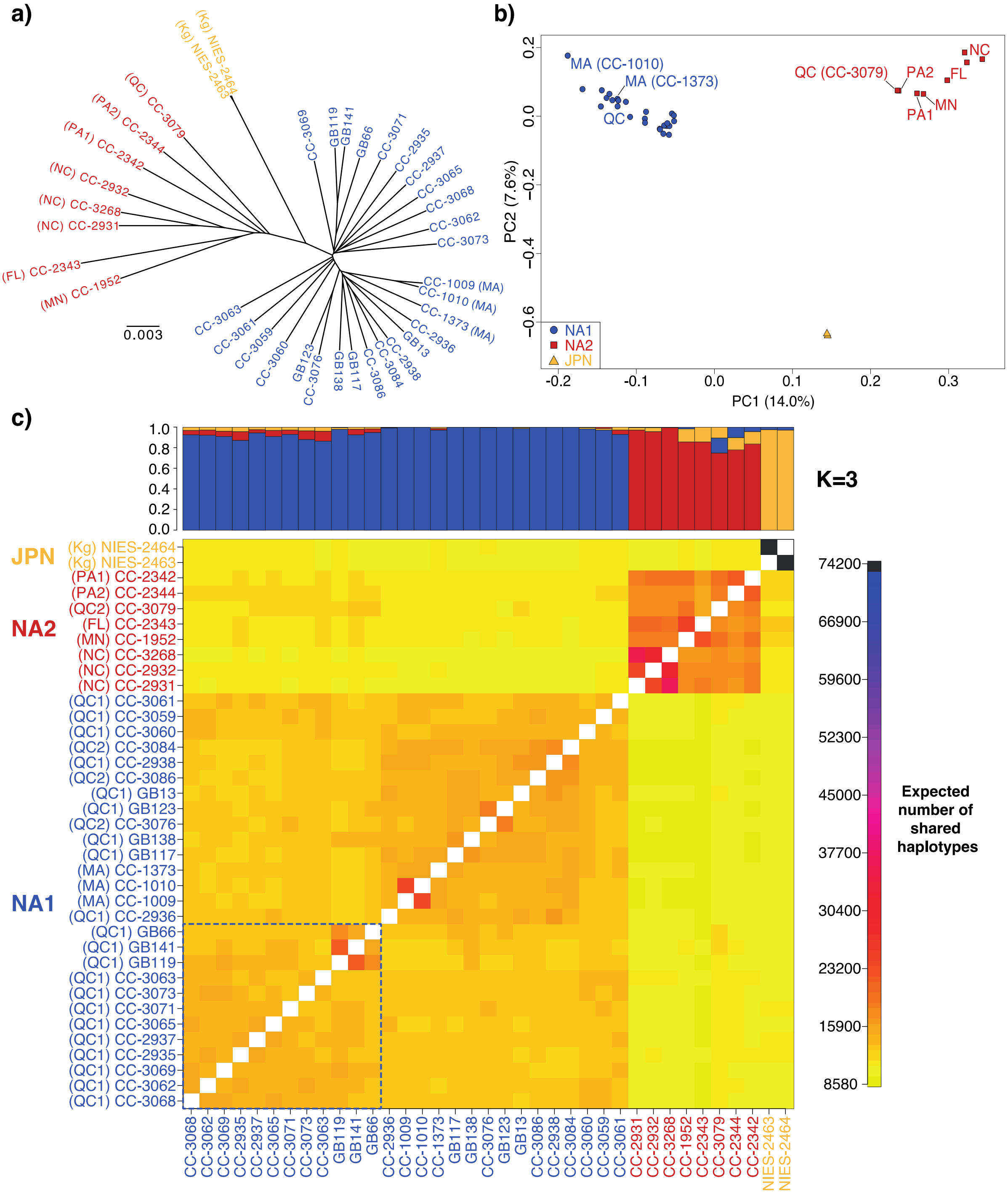

Figure 2. Results of the population structure analyses. a) Neighbour joining tree of all 4D sites, with NA1 isolates coloured blue, NA2 isolates red, and JPN isolates yellow. All nodes had >70% bootstrap support, with the exception of the node connecting CC-3069 with GB119/GB141/GB66. b) The first and second axes of the PCA. c) fineSTRUCTURE coancestry matrix, in which the colour of the cells represents the expected number of shared haplotypes between donor (columns) and recipient (rows) isolate pairs. The blue dashed square marks a subset of highly admixed NA1 isolates. Sampling locations for each isolate are provided on the y-axis (see figure 1 for abbreviations). A STRUCTURE plot for three populations is shown above the matrix (see figure S1 for additional population numbers).

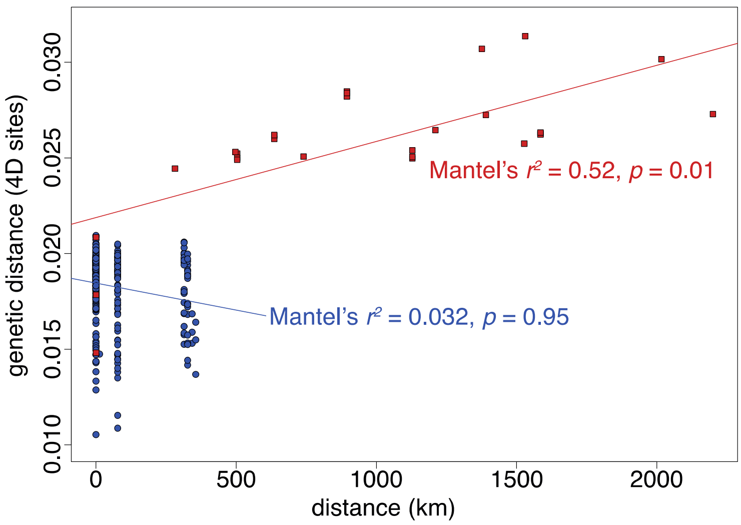

Figure 3. Mantel tests performed on matrices of genetic distance and geographical distance within NA1 (blue) and NA2 (red).

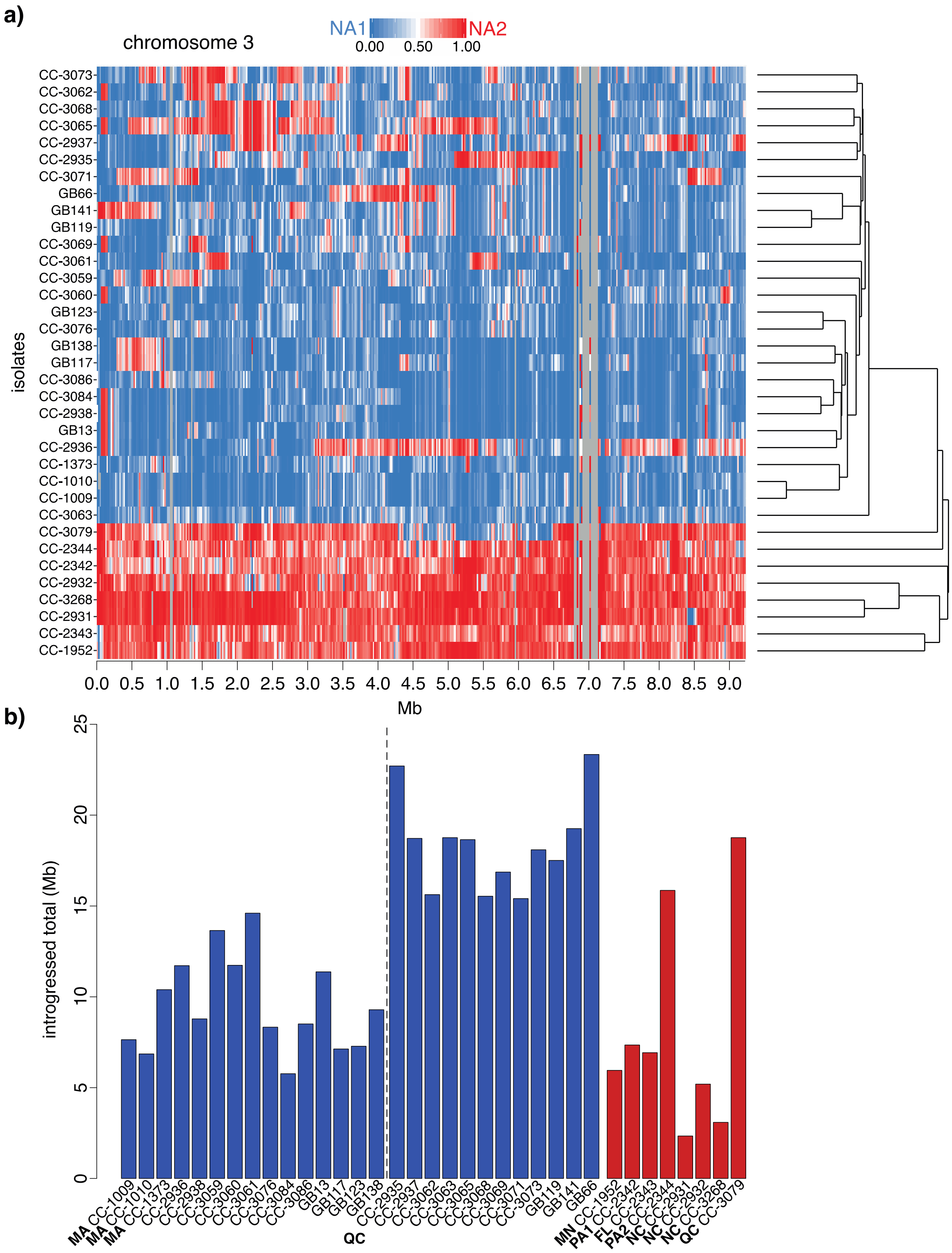

Figure 4. Admixture profiling. a) For each isolate, the proportion of NA1 and NA2 marker SNPs in 20 kb windows along chromosome 3 plotted as a heat map, with 0 (dark blue) representing 100% NA1 SNPs, and 1 (dark red) representing 100% NA2 SNPs. Windows containing no sites/SNPs are shown in grey. Chromosome 3 was randomly selected, see figure S3 for all chromosomes. b) Per isolate total of introgressed sequence, with NA1 isolates shown in blue (with bars representing the total length of introgressed sequence from NA2), and NA2 isolates shown in red. The NA1 isolates to the right of the dashed line are those that were designated as highly admixed from the fineSTRUCTURE analysis.

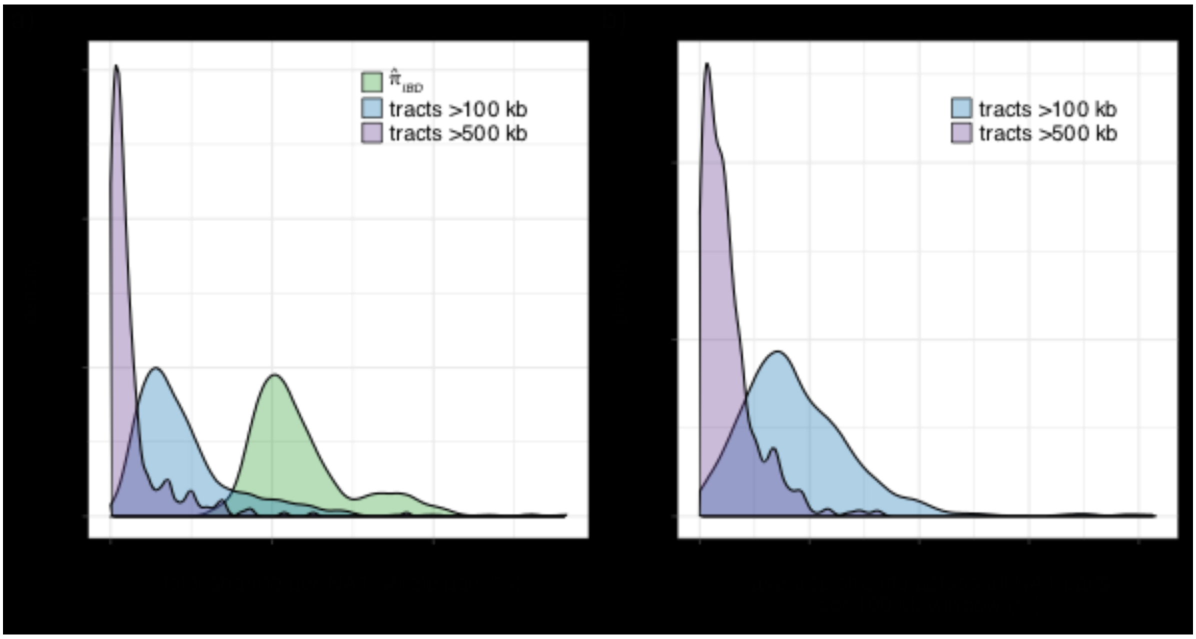

Figure 5. NA1 identity by descent analyses a) Density plot of the estimates of total sharing across all 325 isolate pair comparisons for NA1, shown for the three definitions of identity by descent. b) Density plot of the mean sharing across all 325 NA1 pairs per 100 kb chromosomal window, shown for tracts >100 kb and >500 kb.

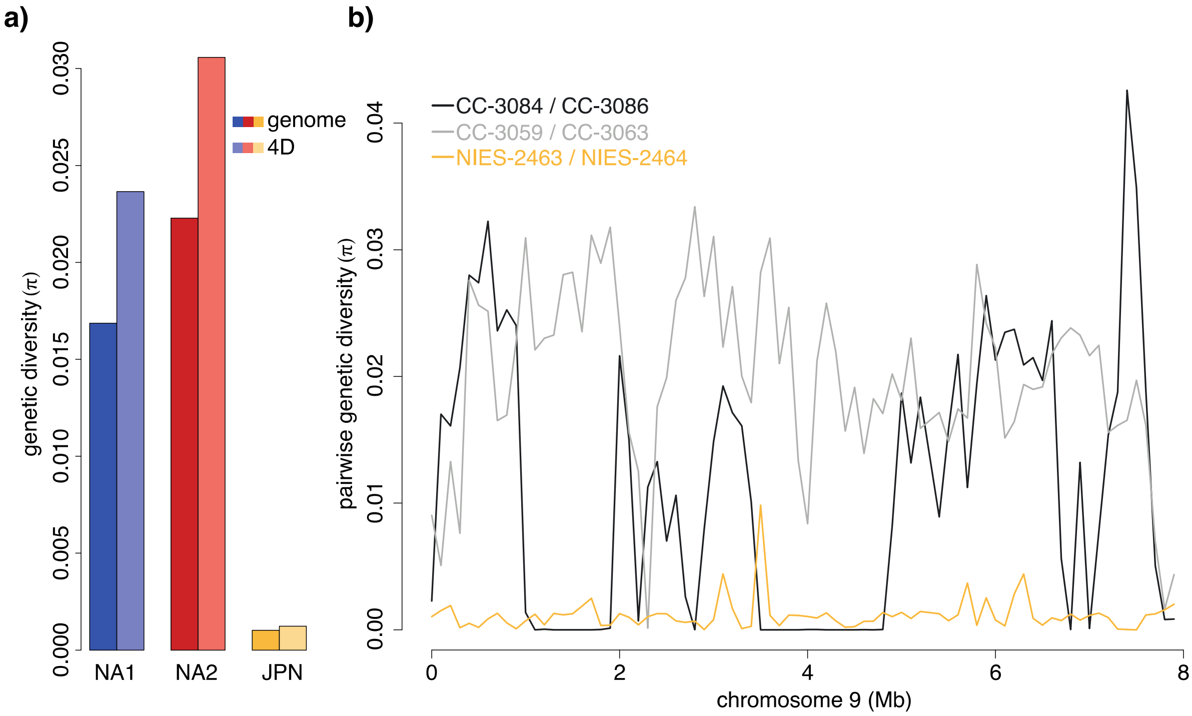

Figure 6. Summary of *C. reinhardtii* genetic diversity. A) Genome-wide and 4D within-lineage genetic diversity for NA1, NA2 and JPN. b) A comparison of pairwise genetic diversity estimated along chromosome 9 in 100 kb windows, for the JPN isolates, and for Quebec isolate pairs exhibiting a low (CC-3059 – CC-3063) and high (CC-3084 – CC-3086) incidence of identity by descent sharing.

References

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., . . . Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology, 12*(2), R18. doi:10.1186/gb-2011-12-2-r18

Baas Becking, L. G. M. (1934). *Geobiologie of Inleiding tot de Milieukunde*. The Hague: Van Stockum & Zoon.

Bell, G. A. C. (1997). Experimental evolution in *Chlamydomonas*. I. Short-term selection in uniform and diverse environments. *Heredity, 78*, 490-497. doi:10.1038/hdy.1997.77

Blaby, I. K., Blaby-Haas, C. E., Tourasse, N., Hom, E. F., Lopez, D., Aksoy, M., . . . Prochnik, S. (2014). The *Chlamydomonas* genome project: a decade on. *Trends in Plant Science, 19*(10), 672-680. doi:10.1016/j.tplants.2014.05.008

Carmi, S., Palamara, P. F., Vacic, V., Lencz, T., Darvasi, A., & Pe'er, I. (2013). The variance of identity-by-descent sharing in the Wright-Fisher model. *Genetics, 193*(3), 911-928. doi:10.1534/genetics.112.147215

Carmi, S., Wilton, P. R., Wakeley, J., & Pe'er, I. (2014). A renewal theory approach to IBD sharing. *Theoretical Population Biology, 97*, 35-48. doi:10.1016/j.tpb.2014.08.002

Caron, D. A. (2009). Past President's address: protistan biogeography: why all the fuss? *Journal of Eukaryotic Microbiology, 56*(2), 105-112. doi:10.1111/j.1550-7408.2008.00381.x

Carriconde, F., Gardes, M., Jargeat, P., Heilmann-Clausen, J., Mouhamadou, B., & Gryta, H. (2008). Population evidence of cryptic species and geographical structure in the cosmopolitan ectomycorrhizal fungus, *Tricholoma scalpturatum*. *Microbial Ecology, 56*(3), 513-524. doi:10.1007/s00248-008-9370-2

Casteleyn, G., Leliaert, F., Backeljau, T., Debeer, A. E., Kotaki, Y., Rhodes, L., . . . Vyverman, W. (2010). Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences of the United States of America, 107*(29), 12952-12957. doi:10.1073/pnas.1001380107

Charron, G., Leducq, J. B., & Landry, C. R. (2014). Chromosomal variation segregates within incipient species and correlates with reproductive isolation. *Molecular Ecology, 23*(17), 4362-4372. doi:10.1111/mec.12864

Clement, M., Snell, Q., & Walker, P. (2002). TCS: Estimating gene genealogies. *Proceedings of the 16th International Parallel and Distributed Processing Symposium, 2:184*.

Colegrave, N. (2002). Sex releases the speed limit on evolution. *Nature, 420*(6916), 664-666. doi:10.1038/nature01191

Collins, S., & Bell, G. (2004). Phenotypic consequences of 1,000 generations of selection at elevated CO2 in a green alga. *Nature, 431*(7008), 566-569. doi:10.1038/nature02945

De Hoff, P. L., Ferris, P., Olson, B. J. S. C., Miyagi, A., Geng, S., & Umen, J. G. (2013). Species and population level molecular profiling reveals cryptic recombination and emergent asymmetry in the dimorphic mating locus of *C. reinhardtii*. *PLoS Genetics, 9*(8). doi:10.1371/journal.pgen.1003724

De Meester, L., Gómez, A., Okamura, B., & Schwenk, K. (2002). The Monopolization Hypothesis and the dispersal-gene flow paradox in aquatic organisms. *Acta Oecologica, 23*(3), 121-135. doi:10.1016/S1146-609x(02)01145-1

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), 491-498. doi:10.1038/ng.806

Douglas, T. E., Kronforst, M. R., Queller, D. C., & Strassmann, J. E. (2011). Genetic diversity in the social amoeba *Dictyostelium discoideum*: population differentiation and cryptic species. *Molecular Phylogenetics and Evolution, 60*(3), 455-462. doi:10.1016/j.ympev.2011.05.007

Dürrenberger, F., Thompson, A. J., Herrin, D. L., & Rochaix, J. D. (1996). Double strand break-induced recombination in *Chlamydomonas reinhardtii* chloroplasts. *Nucleic Acids Research, 24*(17), 3323-3331. doi:10.1093/nar/24.17.3323

Ellison, C. E., Hall, C., Kowbel, D., Welch, J., Brem, R. B., Glass, N. L., & Taylor, J. W. (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences of the United States of America, 108*(7), 2831-2836. doi:10.1073/pnas.1014971108

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics, 164*(4), 1567-1587.

Fenchel, T., & Finlay, B. J. (2004). The ubiquity of small species: Patterns of local and global diversity. *Bioscience, 54*(8), 777-784. doi:10.1641/0006-3568(2004)054[0777:Tuossp]2.0.Co;2

Filatov, D. A. (2019). Extreme Lewontin's paradox in ubiquitous marine phytoplankton species. *Molecular Biology and Evolution, 36*(1), 4-14. doi:10.1093/molbev/msy195

Finlay, B. J. (2002). Global dispersal of free-living microbial eukaryote species. *Science, 296*(5570), 1061-1063. doi:10.1126/science.1070710

Finlay, B. J., & Fenchel, T. (1999). Divergent perspectives on protist species richness. *Protist, 150*(3), 229-233. doi:10.1016/S1434-4610(99)70025-8

Flowers, J. M., Hazzouri, K. M., Pham, G. M., Rosas, U., Bahmani, T., Khraiwesh, B., . . . Purugganan, M. D. (2015). Whole-genome resequencing reveals extensive natural variation in the model green alga *Chlamydomonas reinhardtii*. *Plant Cell, 27*(9), 2353-2369. doi:10.1105/tpc.15.00492

Foissner, W. (1999). Protist diversity: estimates of the near-imponderable. *Protist, 150*(4), 363-368. doi:10.1016/S1434-4610(99)70037-4

Foissner, W. (2006). Biogeography and dispersal of micro-organisms: A review emphasizing protists. *Acta Protozoologica, 45*(2), 111-136.

Foissner, W. (2008). Protist diversity and distribution: some basic considerations. *Biodiversity and Conservation, 17*(2), 235-242. doi:10.1007/s10531-007-9248-5

Fryxell, G. A. (1983). *Survival strategies of the algae*. New York, NY: Cambridge University Press.

Gallaher, S. D., Fitz-Gibbon, S. T., Glaesener, A. G., Pellegrini, M., & Merchant, S. S. (2015). *Chlamydomonas* genome resource for laboratory strains reveals a mosaic of sequence variation, identifies true strain histories, and enables strain-specific studies. *Plant Cell, 27*(9), 2335-2352. doi:10.1105/tpc.15.00508

Harris, E. H. (2001). *Chlamydomonas* as a Model Organism. *Annual Review of Plant Physiology and Plant Molecular Biology, 52*, 363-406. doi:10.1146/annurev.arplant.52.1.363

Harris, E. H. (2008). *The Chlamydomonas Sourcebook (Second Edition): Introduction to Chlamydomonas and Its Laboratory Use.* San Diego, CA: Academic Press.

Hasan, A. R., Duggal, J. K., & Ness, R. W. (2019). Consequences of recombination for the evolution of the mating type locus in *Chlamydomonas reinhardtii*. *New Phytologist*. doi:10.1111/nph.16003

Heger, T. J., Mitchell, E. A., & Leander, B. S. (2013). Holarctic phylogeography of the testate amoeba *Hyalosphenia papilio* (Amoebozoa: Arcellinida) reveals extensive genetic diversity explained more by environment than dispersal limitation. *Molecular Ecology, 22*(20), 5172-5184. doi:10.1111/mec.12449

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics, 132*(2), 583-589.

Jang, H., & Ehrenreich, I. M. (2012). Genome-wide characterization of genetic variation in the unicellular, green alga *Chlamydomonas reinhardtii*. *PLoS One, 7*(7), e41307. doi:10.1371/journal.pone.0041307

Johri, P., Krenek, S., Marinov, G. K., Doak, T. G., Berendonk, T. U., & Lynch, M. (2017). Population genomics of *Paramecium* species. *Molecular Biology and Evolution, 34*(5), 1194-1216. doi:10.1093/molbev/msx074

Kathir, P., LaVoie, M., Brazelton, W. J., Haas, N. A., Lefebvre, P. A., & Silflow, C. D. (2003). Molecular map of the *Chlamydomonas reinhardtii* nuclear genome. *Eukaryot Cell, 2*(2), 362-379. doi:10.1128/ec.2.2.362-379.2003

Kawasaki, Y., Nakada, T., & Tomita, M. (2015). Taxonomic revision of oil-producing green algae, *Chlorococcum Oleofaciens* (Volvocales, Chlorophyceae), and its relatives. *Journal of Phycology, 51*(5), 1000-1016. doi:10.1111/jpy.12343

Koufopanou, V., Hughes, J., Bell, G., & Burt, A. (2006). The spatial scale of genetic differentiation in a model organism: the wild yeast *Saccharomyces paradoxus*. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1475), 1941-1946. doi:10.1098/rstb.2006.1922

Kristiansen, J. (1996). Dispersal of freshwater algae - a review. *Hydrobiologia, 336*(1-3), 151-157. doi:10.1007/Bf00010829

Kuehne, H. A., Murphy, H. A., Francis, C. A., & Sniegowski, P. D. (2007). Allopatric divergence, secondary contact and genetic isolation in wild yeast populations. *Current Biology, 17*(5), 407-411. doi:10.1016/j.cub.2006.12.047

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution, 33*(7), 1870-1874. doi:10.1093/molbev/msw054

Lahr, D. J., Laughinghouse, H. D. t., Oliverio, A. M., Gao, F., & Katz, L. A. (2014). How discordant morphological and molecular evolution among microorganisms can revise our notions of biodiversity on Earth. *Bioessays, 36*(10), 950-959. doi:10.1002/bies.201400056

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genetics, 8*(1), e1002453. doi:10.1371/journal.pgen.1002453

Lebret, K., Tesson, S. V. M., Kritzberg, E. S., Tomas, C., & Rengefors, K. (2015). Phylogeography of the freshwater raphidophyte *Gonyostomum Semen* confirms a recent expansion in Northern Europe by a single haplotype. *Journal of Phycology, 51*(4), 768-781. doi:10.1111/jpy.12317

Leducq, J. B., Charron, G., Samani, P., Dubé, A. K., Sylvester, K., James, B., . . . Landry, C. R. (2014). Local climatic adaptation in a widespread microorganism. *Proceedings of the*

Royal Society B: Biological Sciences, 281(1777), 20132472. doi:10.1098/rspb.2013.2472

Leducq, J. B., Nielly-Thibault, L., Charron, G., Eberlein, C., Verta, J. P., Samani, P., . . . Landry, C. R. (2016). Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nature Microbiology, 1*, 15003. doi:10.1038/nmicrobiol.2015.3

Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., . . . Przeworski, M. (2012). Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology, 10*(9), e1001388. doi:10.1371/journal.pbio.1001388

Leigh, J. W., & Bryant, D. (2015). PopART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution, 6*(9), 1110-1116. doi:10.1111/2041-210x.12410

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics, 30*(20), 2843-2851. doi:10.1093/bioinformatics/btu356

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Liss, M., Kirk, D. L., Beyser, K., & Fabry, S. (1997). Intron sequences provide a tool for high-resolution phylogenetic analysis of volvocine algae. *Current Genetics, 31*(3), 214-227.

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., . . . Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature, 458*(7236), 337-341. doi:10.1038/nature07743

Liu, H., Huang, J., Sun, X., Li, J., Hu, Y., Yu, L., . . . Yang, S. (2018). Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. *Nature Ecology & Evolution, 2*(1), 164-173. doi:10.1038/s41559-017-0372-7

Lowe, C. D., Martin, L. E., Montagnes, D. J. S., & Watts, P. C. (2012). A legacy of contrasting spatial genetic structure on either side of the Atlantic-Mediterranean transition zone in a marine protist. *Proceedings of the National Academy of Sciences of the United States of America, 109*(51), 20998-21003. doi:10.1073/pnas.1214398110

Machida, H., & Arai, F. (2003). *Atlas of tephra in and around Japan* (rev. ed.). Tokyo: University of Tokyo Press (in Japanese).

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., . . . Grossman, A. R. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science, 318*(5848), 245-250. doi:10.1126/science.1143609

Nakada, T., Shinkawa, H., Ito, T., & Tomita, M. (2010). Recharacterization of *Chlamydomonas reinhardtii* and its relatives with new isolates from Japan. *Journal of Plant Research, 123*(1), 67-78. doi:10.1007/s10265-009-0266-0

Nakada, T., Tsuchida, Y., Arakawa, K., Ito, T., & Tomita, M. (2014). Hybridization between Japanese and North American *Chlamydomonas reinhardtii* (Volvocales, Chlorophyceae). *Phycological Research, 62*(3), 232-236. doi:10.1111/pre.12061

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America, 76*(10), 5269-5273. doi:10.1073/pnas.76.10.5269

Ness, R. W., Kraemer, S. A., Colegrave, N., & Keightley, P. D. (2016). Direct estimate of the spontaneous mutation rate uncovers the effects of drift and recombination in the *Chlamydomonas reinhardtii* plastid genome. *Molecular Biology and Evolution, 33*(3), 800-808. doi:10.1093/molbev/msv272

Ness, R. W., Morgan, A. D., Colegrave, N., & Keightley, P. D. (2012). Estimate of the spontaneous mutation rate in *Chlamydomonas reinhardtii*. *Genetics, 192*(4), 1447-1454. doi:10.1534/genetics.112.145078

Ness, R. W., Morgan, A. D., Vasanthakrishnan, R. B., Colegrave, N., & Keightley, P. D. (2015). Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. *Genome Research, 25*(11), 1739-1749. doi:10.1101/gr.191494.115

O'Malley, M. A. (2008). 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Biological and Biomedical Sciences, 39*(3), 314-325. doi:10.1016/j.shpsc.2008.06.005

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., . . . Wagner, H. (2017). vegan: Community Ecology Package. R package version 2.4-5. doi:https://CRAN.R-project.org/package=vegan

Orsini, L., Vanoverbeke, J., Swillen, I., Mergeay, J., & De Meester, L. (2013). Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Molecular Ecology, 22*(24), 5983-5999. doi:10.1111/mec.12561

Palamara, P. F., Lencz, T., Darvasi, A., & Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics, 91*(6), 1150-1150. doi:10.1016/j.ajhg.2012.11.006

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., . . . de Vargas, C. (2012). CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology, 10*(11), e1001419. doi:10.1371/journal.pbio.1001419

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Pröschold, T., Harris, E. H., & Coleman, A. W. (2005). Portrait of a species: *Chlamydomonas reinhardtii*. *Genetics, 170*(4), 1601-1610. doi:10.1534/genetics.105.044503

Rengefors, K., Kremp, A., Reusch, T. B. H., & Wood, A. M. (2017). Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. *Journal of Plankton Research, 39*(2), 165-179. doi:10.1093/plankt/fbw098

Rengefors, K., Logares, R., & Laybourn-Parry, J. (2012). Polar lakes may act as ecological islands to aquatic protists. *Molecular Ecology, 21*(13), 3200-3209. doi:10.1111/j.1365-294X.2012.05596.x

Sack, L., Zeyl, C., Bell, G., Sharbel, T., Reboud, X., Bernhardt, T., & Koelewyn, H. (1994). Isolation of four new strains of *Chlamydomonas reinhardtii* (Chlorophyta) from soil samples. *Journal of Phycology, 30*(4), 770-773. doi:10.1111/j.0022-3646.1994.00770.x

Sasso, S., Stibor, H., Mittag, M., & Grossman, A. R. (2018). From molecular manipulation of domesticated *Chlamydomonas reinhardtii* to survival in nature. *Elife, 7*. doi:10.7554/eLife.39233

Schaffner, S. F., Taylor, A. R., Wong, W., Wirth, D. F., & Neafsey, D. E. (2018). hmmIBD: software to infer pairwise identity by descent between haploid genotypes. *Malaria Journal, 17*(1), 196. doi:10.1186/s12936-018-2349-7

Scranton, M. A., Ostrand, J. T., Fields, F. J., & Mayfield, S. P. (2015). *Chlamydomonas* as a model for biofuels and bio-products production. *The Plant Journal, 82*(3), 523-531. doi:10.1111/tpj.12780

Shoemaker, W. R., & Lennon, J. T. (2018). Evolution with a seed bank: The population genetic consequences of microbial dormancy. *Evolutionary Applications, 11*(1), 60-75. doi:10.1111/eva.12557

Sjöqvist, C., Godhe, A., Jonsson, P. R., Sundqvist, L., & Kremp, A. (2015). Local adaptation and oceanographic connectivity patterns explain genetic differentiation of a marine diatom across the North Sea-Baltic Sea salinity gradient. *Molecular Ecology, 24*(11), 2871-2885. doi:10.1111/mec.13208

Taylor, A. R., Schaffner, S. F., Cerqueira, G. C., Nkhoma, S. C., Anderson, T. J. C., Sriprawat, K., . . . Buckee, C. O. (2017). Quantifying connectivity between local *Plasmodium falciparum* malaria parasite populations using identity by descent. *PLoS Genetics, 13*(10), e1007065. doi:10.1371/journal.pgen.1007065

Tellier, A., & Lemaire, C. (2014). Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular Ecology, 23*(11), 2637-2652. doi:10.1111/mec.12755

Thompson, E. A. (2013). Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics, 194*(2), 301-326. doi:10.1534/genetics.112.148825

Vanormelingen, P., Evans, K. M., Mann, D. G., Lance, S., Debeer, A. E., D'Hondt, S., . . . Vyverman, W. (2015). Genotypic diversity and differentiation among populations of two benthic freshwater diatoms as revealed by microsatellites. *Molecular Ecology, 24*(17), 4433-4448. doi:10.1111/mec.13336

Vanoverbeke, J., & De Meester, L. (2010). Clonal erosion and genetic drift in cyclical parthenogens - the interplay between neutral and selective processes. *Journal of Evolutionary Biology, 23*(5), 997-1012. doi:10.1111/j.1420-9101.2010.01970.x

Wakeley, J., & Wilton, P. R. (2016). Coalescent and models of identity by descent. In R. M. Kliman (Ed.), *Encyclopedia of Evolutionary Biology* (Vol. 1, pp. 287-292). Oxford: Academic Press.

Whittaker, K. A., & Rynearson, T. A. (2017). Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. *Proceedings of the National Academy of Sciences of the United States of America, 114*(10), 2651-2656. doi:10.1073/pnas.1612346114

Zheng, X. W., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics, 28*(24), 3326-3328. doi:10.1093/bioinformatics/bts606

Zufall, R. A., Dimond, K. L., & Doerder, F. P. (2013). Restricted distribution and limited gene flow in the model ciliate *Tetrahymena thermophila*. *Molecular Ecology, 22*(4), 1081-1091. doi:10.1111/mec.12066