UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Sex, Alternative Lifestyles, and a Graphic Study of a Model

Permalink
https://escholarship.org/uc/item/2n26249r

Author
Hann-Soden, Christopher T

Publication Date
2018

Supplemental Material
https://escholarship.org/uc/item/2n26249r#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Sex, Alternative Lifestyles, and a Graphic Study of a Model

by

Christopher T. Hann-Soden

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John W. Taylor, Chair
Professor N. Louise Glass
Professor Michael Nachman

Fall 2018

Sex, Alternative Lifestyles, and a Graphic Study of a Model

Copyright 2018
by
Christopher T. Hann-Soden

Abstract

Sex, Alternative Lifestyles, and a Graphic Study of a Model

by

Christopher T. Hann-Soden

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor John W. Taylor, Chair

The genus of Ascomycete molds, *Neurospora*, has served as a valuable model for the cellular and molecular biology of multicellular fungi, and more recently for the evolution of fungi. In this thesis I present three studies of *Neurospora* which each expand the knowledge and utility of this model.

First, I expand collections of rare *Neurospora* species by 113 strains and 149 whole genome sequences, including the first conidial homothallic (self-fertile) *Neurospora* as well as genome sequences for endolichenic *Neurospora* from two species, both aconidial. These data provide insight into *Neurospora*'s mysterious lifestyle, and show that at least some *Neurospora* species live part of their lives as endosymbionts. These collections also expand the utility of *Neurospora* as a model for breeding system evolution by showing that every combination of reproductive system seen in *Ascomycota* is also seen in *Neurospora*.

Second, I use this collection of genomes, along with 43 previously published genomes, to study the population biology of *Neurospora* species with diverse reproductive ecologies. While homothallic *Neurospora* had previously been assumed to be dominantly or exclusively selfing, I show that at least one aconidial homothallic lineage is recombining at least as much as heterothallic (self-sterile) species. I further detail the population structure and biogeography of diverse groups of *Neurospora*, showing how differing ecologies and mating paradigms work at different scales to shape populations.

Third, I address the question: How does the rate of genomic rearrangements vary across the genome? Genomic rearrangements are an important source of adaptive variation, but studying them has been challenging due to a lack of suitable (i.e. whole genome) data and the computational complexity of the problem. Leveraging this collection of *Neurospora* genomes, I present a method to estimate the rearrangement rate across the genome. I phrase the problem in graph theoretic terms and find it to be an instance of the well known clique cover problem. I use this method to find evidence that in *N. crassa*, like other Eukaryotes, rearrangements are more common in the subtelomeric regions of the chromosomes, which facilitates the evolution of novel genes.

To My Teachers

I never listened to your advice.

Contents

List of Figures

List of Tables

Acknowledgments

This work was funded by NSF grant DEB-1257528, and CHS received support from the Philomathia Graduate Student Fellowship in the Environmental Sciences. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. Computation was performed on Berkeley Research Computing and Lawrence Berkeley National Lab's HPC cluster accessed through the Computational Genomics Research Laboratory.

Chapter 1

New Reproductive and Ecological Diversity in the Model Genus, *Neurospora*

Christopher Hann-Soden, Pierre Gladieux, Liliam Montoya, and John W. Taylor

Abstract

The model of multicellar fungi, *Neurospora*, is also becoming a model of fungal evolution. Yet evolutionary studies have been largely restricted to heterothallic (self-sterile) conidiating *Neurospora*, largely because the conidia they produce on burned vegetation are easy to cultivate from nature. All heretofore reported homothallic (self-fertile) *Neurospora* species have been aconidial and have not been associated with burned plants. In an effort to extend the *Neurospora* model to include homothallic fungi we sampled populations of *Neurospora* from soil across North America, as well as searched GenBank to find anonymous *Neurospora* records from new substrates. We have expanded collections of rare *Neurospora* species by 113 strains and 149 whole genome sequences, including: the first conidial homothallic *Neurospora*; genome sequences for endolichenic *Neurospora* from two species, both aconidial; and collections suitable for studying the population biology of aconidial *Neurospora* of all breeding systems.

1.1 Introduction

The ascomycete genus, *Neurospora*, has served as a valuable model for cellular and molecular biology (Davis, 2000; Davis and Perkins, 2002), and more recently genomics and evolution (Gladieux *et al.*, prep). Still, understanding of *Neurospora*'s biology is hampered by a lack of knowledge about basic aspects of its ecology; we know shockingly little about where

and how *Neurospora* lives in nature.

The Diverse Lifestyles of *Neurospora*

Studies of wild *Neurospora* have largely reflected the convenience of sampling, starting with *Neurospora*'s fortuitous infestation of French army bakeries (Ramsbottom and Stephens, 1935). These strains, and many other species of *Neurospora*, were easily spotted by their prodigious tufts of bright pink, yellow, or orange conidia. These conidial species have most often being collected from disturbed habitats, including rotting grain and fruit, corn cobs, oncom (a Sundanese fermented staple food), charred grain, sugarcane bagasse and filter mud, spent hops, steamed or kiln dried lumber, and burnt trees and shrubs (Ramsbottom and Stephens, 1935; Turner, 1987; Shaw, 1993; Perkins and Turner, 1988; Turner *et al.*, 2001; Jacobson *et al.*, 2004). However, tropical species of conidial *Neurospora* have also been found on living sugarcane and other grasses (Turner *et al.*, 2001), and have been known to be collected by honeybees (Turner, 1987; Shaw, 1993). Still, being common upon dead plant matter, *Neurospora* clearly has a saprobic niche, which is reflected by the abundance of enzymes for degrading plant biomass in *N. crassa*'s genome (Tian *et al.*, 2009).

Many of the substrates from which conidial *Neurospora* have been recovered have been burned or heat treated. In addition to the obvious examples, hops are boiled in the brewing process, bagasse is produced along with sugarcane press mud, which is heated, and compost piles can reach 75°C owing to heat from microbial activity. The frequent recovery of these species from burned or heated vegetation suggests a fire-adapted or thermophillic life cycle, which is further supported by the fact that *Neurospora*'s sexual spores are germinated by heat and combustion byproducts like furans (Jacobson *et al.*, 2004).

In contrast to the brightly colored conidial species, aconidial species have most often been isolated from dung or soil, but were not necessarilly recognized as members of the same genus until molecular evidence showed them to form a monophyletic clade with the condidiating species (Garcia *et al.*, 2004). Without distinctively colored conidia, aconidial but homothallic (self-fertile) *Neurospora* species are recognizible by *Neurospora*'s distinctive beaked perithecia which discharge either the eponymous ornate, spindle-shaped ascospores or the pitted, round ascospores characteristic of species formerly named *Gelasinospora*. Less research has been done on the aconidial species, but their infrequency on heat-treated plants and abundance on dung and soil suggests an alternative lifestyle.

The most cryptic of all *Neurospora* species are the aconidial and heterothallic (self-sterile) species that produce no identifiable morphological features in pure culture, appearing as a plain white mat of mycelium. Their true affinities emerge only when they are mated to produce characteristic perithecia and ascospores. These species have only been isolated once before when individuals of different mating type, newly-isolated from soil, serendipitously found each other on the plate and successfully reproduced (Glass *et al.*, 1990).

	Heterothallic	Pseudohomothallic	Homothallic
Conidial	F, O, P, S	F, S	S*
Aconidial	E*, S	E*, S	D, P, R, S

Table 1.1: Sampling sources of *Neurospora* by reproductive ecology. D - Dung, E - Endolichenic, F - burned/heated plants, O - Other, P - living Plant (stem or leaves), R - living plant (Root), S - Soil. *New in this work

Neurospora as an Endophyte

In temperate climates burned trees have proven a reliable source for the collection of natural populations of *Neurospora*, and the sudden appearance of *Neurospora* after forest fires has led to speculation about a missing reservoir of ascospores or living tissue (Jacobson *et al.*, 2004; Powell *et al.*, 2003). One hypothesis is that *Neurospora* species are endophytes that reproduce after their host plant is killed by the heat of fires. Subsequently, recovery of a *Neurospora* strain from *Acer ginnala* (Qi *et al.*, 2012) gave credence to this hypothesis, although we note that *Neurospora* ITS sequences had been previously discovered without comment in the roots of grasses (Khidir *et al.*, 2010). Since then, one aspect of the endophyte hypothesis was demonstrated by Kuo *et al.* (2014) who successfully inoculated the model tree, *Pinus sylvestris*, with a lab strain of *N. crassa*.

Surveying *Neurospora* From Novel Substrates

We are interested in *Neurospora* as a model system for evolution, and seek to expand its utility in this field by obtaining natural populations of diverse but rarely sampled *Neurospora*, such as the aconidial heterothallic species. In doing so we also sought to survey new corners of the natural environment for *Neurospora* and shed light on the secret lifestyle of this model organism. We hypothesized that the soil sampling method of Glass *et al.* (1990), which was previously used to discover the aconidial heterothallic strains, could provide a less biased, or at least alternative, source of *Neurospora* without the reliance on forest fires.

Specifically, we note that the most common species sampled in temperate North America, *N. discreta*, had not been reported from soil sampling and wondered if its ascospores reside in the soil. We also wanted to test the feasibility of sampling soil for natural populations of *Neurospora* for population genetic and evolutionary studies. We lastly hypothesized that if *Neurospora* is an endophyte or has other unknown habitats, it may exist anonymously within environmental culture collections.

To these ends, we sampled *Neurospora* from soil around North America, as well as retrieved previously collected *Neurospora* cultures identified by bioinformatic search. We identified these strains via whole genome sequencing, and present these new populations consisting of rare aconidial heterothallic and aconidial pseudohomothallic *Neurospora*, the first described endolichenic *Neurospora*, and the first reported conidial homothallic *Neurospora*.

Taxon	Morphology	Total Isolates	Delta Junction, Alaska	Eagle Summit, Alaska	Nome, Alaska	Pt. Reyes, California	Truckee, California	Littleton, Colorado	Hartwick Pines SP, Michigan	Hwy 32, Michigan	Alpena, Michigan
N. discreta sp. nov.	conidial, hom.	7	7								
N. cerealis	acon., hom.	13						13			
Cailleux IV	acon., hom.	18				6		11	1		
N. tetraspora	acon., pseudo.	11/24	1/1	6*/0	2*/0			1/3	0/7	1/12	0/1
N. sp. (Alaska)	acon., het.	10	7	2*					1		
N. sp. (Midwest)	acon., het.	8						2	4	2	
N. sp. (Mountain)	acon., het.	32				1	9	22			
Total		123	16	8	2	7	9	52	13	15	1

Table 1.2: Number of strains sequenced of different taxa from different sampling sites. For the pseudohomothallic *N. tetraspora* the number of haploid strains is given, followed by a slash, then the number of dikaryotic strains. *Endolichenic strains

1.2 Results and Discussion

Strain Collection and Identification

We collected soil samples from four broad geographic regions in North America, then germinated ascospores and isolated *Neurospora* from the soil following the protocol of (Glass *et al.*, 1990). Briefly, we heat-germinated mixed cultures of fungi from the soil then washed ascospores, shot by *Neurospora* perithecia, from the lids of the plates for a second round of germination and isolation. Once in pure culture, we were able to discern heterothallic strains from homothallic or pseudohomothallic strains because they do not produce perithecia or ascospores without a mate. In some cases we further discerned homothallic from pseudohomothallic strains by squashing mature perithecia and counting the number of spores within asci; Homothallic species produce 8 spores (4 from meiosis doubled by a round of mitosis) while pseudohomothallic species usually produce half as many due to their dikaryotic packaging. From this collection we sequenced the genomes of 113 *Neurospora* strains.

We merged this collection with a collection of *N. discreta* from burned trees in temperate North America, including 36 previously published genomes (Gladieux *et al.*, 2015), and 19 newly sequenced genomes. We also used the NCBI BLAST web server (Madden, 2002) to search the Nucleotide Collection database for the *N. crassa* ITS sequence. We filtered hits for unidentified environmental sequences, then searched the associated literature for culture-based studies. From this search, we were able to locate 10 endolichenic strains collected by U'Ren *et al.* (2012), and obtained them from the authors. We sequenced these strains, along with 7 strains of various species from culture collections, bringing the toal number of whole genome sequences to 185.

We have deposited all sequences and assemblies with the National Center for Biotechnology Information (NCBI) in the Sequence Read Archive and Genome databases, respectively, associated with the BioProject PRJNA487060. Additionally, we depositied preserved cultures of sequenced strains with the Fungal Genetics Stock Center (FGSC) at Kansas State University.

In order to identify these strains by phylogenetic analysis, we combined these data with 7 previously published genomes of various species (Ellison *et al.*, 2011a; Gioti *et al.*, 2013; Galagan *et al.*, 2003; Nowrousian *et al.*, 2010). We present details of genome sequences, including FGSC strain numbers, in Table A.2.

We assembled genomes *de novo* for all sequenced strains and assigned taxonomy via whole genome phylogenetic analysis. New strains fell into 12 clusters with 100% branch support by both bootstrap and Shimodiara-Hasegawa-like approximate Likelihood Ratio Test (SH-aLRT) (Figs. 1.1, 1.2, 1.3, and 1.4). Only one cluster contained both novel and named strains (*N. cerealis*), so we additionally utilized a multi-locus phylogeny (presented in Gladieux *et al.* (prep)) using sites that had previously been used for a comprehensive *Neurospora* phylogeny (Nygren *et al.*, 2011). From these data, as well as morphology, were were able to identify a cluster of aconidial pseudohomothallics as *N. tetraspora* and a cluster of conidial homothallics as a novel lineage within the *N. discreta* species group.

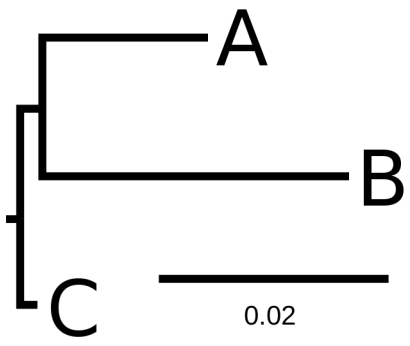

Figure 1.1: Relationship and scale of phylogenies shown on subsequent pages. Distances are in substitutions per base.

The multi-locus phylogeny placed *N. retispora*, *N. santi-flori*, and *N. novoguineensis* within a large clade of aconidial homothallic strains, and *N. dictyophora* and *N. saitoi* as close neighbors with 49% bootstrap support. As we cannot definitively assign taxonomy to any strain within this clade, we refer to the entire clade as the Cailleux IV clade, since it corresponds to Group IV as described by Cailleux (Cailleux, 1971; Garcia *et al.*, 2004).

A New Perspective of Neurospora Diversity in North America

While many *Neurospora* strains have been isolated from soil, it is unclear if the fire-adapted species like *N. discreta* reside in this habitat (Jacobson *et al.*, 2004; Turner *et al.*, 2001). Indeed, we sampled soil at a site in California where Jacobson *et al.* (2004) had collected 35 isolates of *N. discreta* after a fire in 2001, but recovered only aconidial heterothallic strains of a novel species. While *N. discreta* is clearly sexual its fruiting bodies have rarely been observed in nature. Where and when *N. discreta* has sex, or where it resides between forest fires, remains mysterious, but if it leaves a spore bank in the soil it is not as extensive as those left by aconidial species.

Where previous collections have been dominated by conidial heterothallic species such as *N. discreta* and *N. sitophila* in temperate regions, and *N. crassa*, *N. intermedia*, and the pseudohomothallic *N. tetrasperma* in tropical and subtropical regions (Jacobson *et al.*, 2004; Turner *et al.*, 2001), we found that soil is dominated by aconidial species with diverse breeding systems: heterothallic, pseudohomothallic, and homothallic alike. We found only 7 conidial strains within the soil, and these proved to be a type of semi-homothallic *N. discreta*.

A Novel Form of Homothallism

Of the 7 *N. discreta* strains recovered from soil, 4 produced asci with 8 ascospores like other homothallic species, but 3 appeared heterothallic. The homothallic-like strains produced lighter brown perithecia than other *N. discreta* species. Additionally, all 7 strains

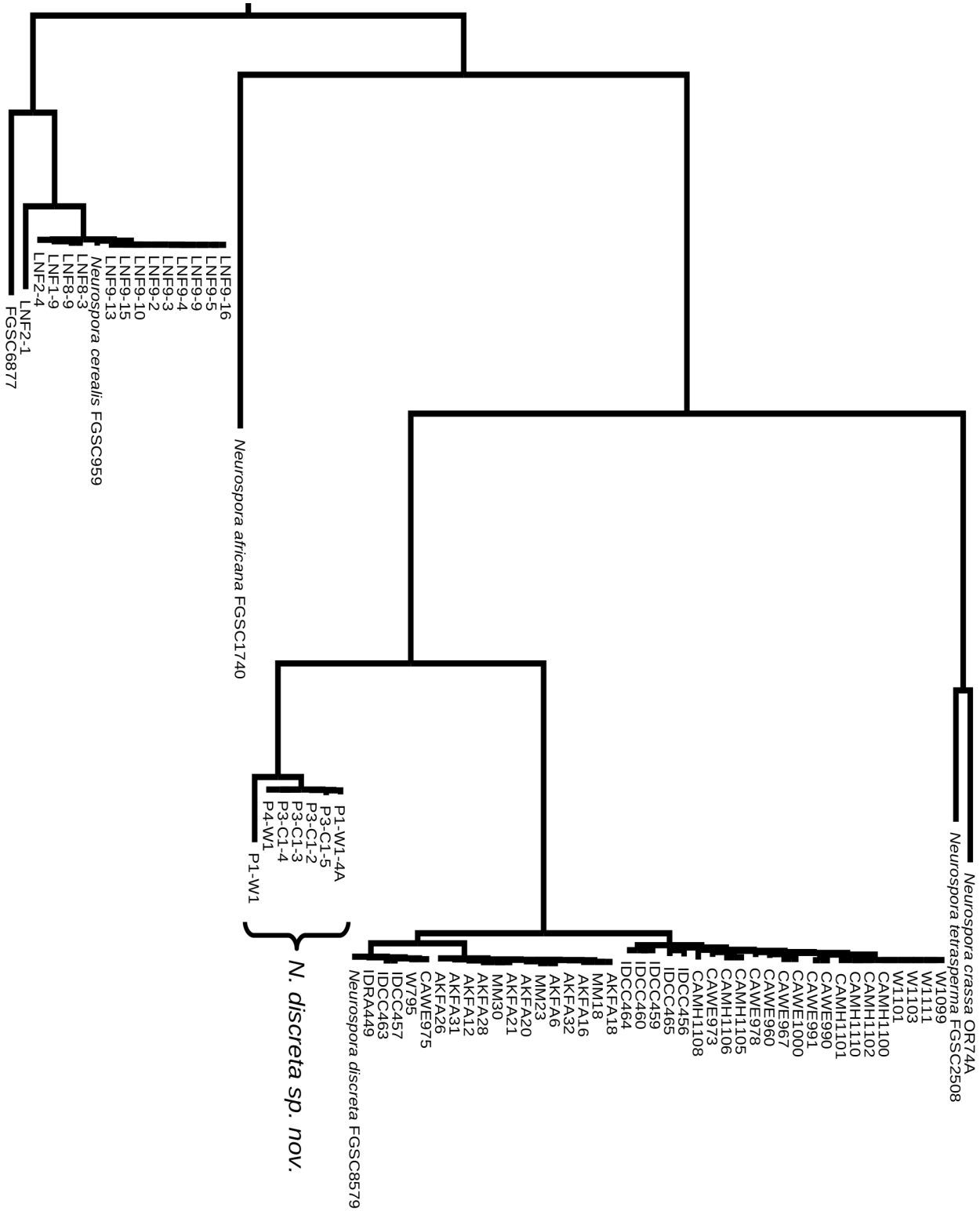

Figure 1.2: Phylogeny of *Neurospora* genomes within the "A" clade (Fig. 1.1).

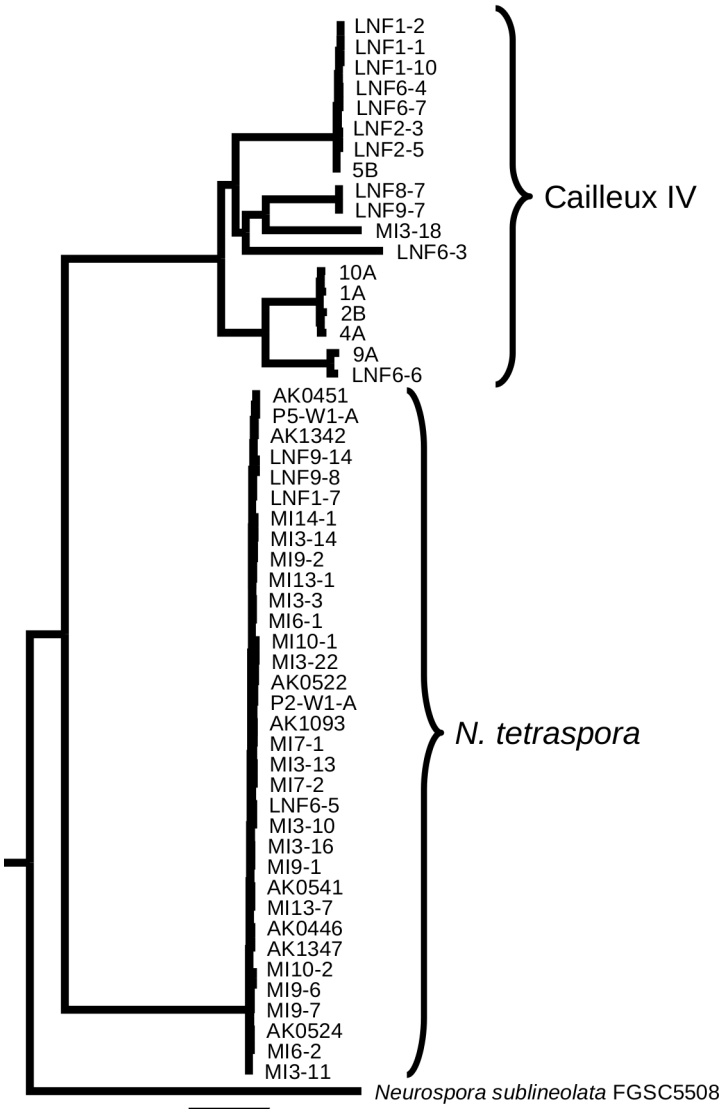

Figure 1.3: Phylogeny of *Neurospora* genomes within the "B" clade (Fig. 1.1).

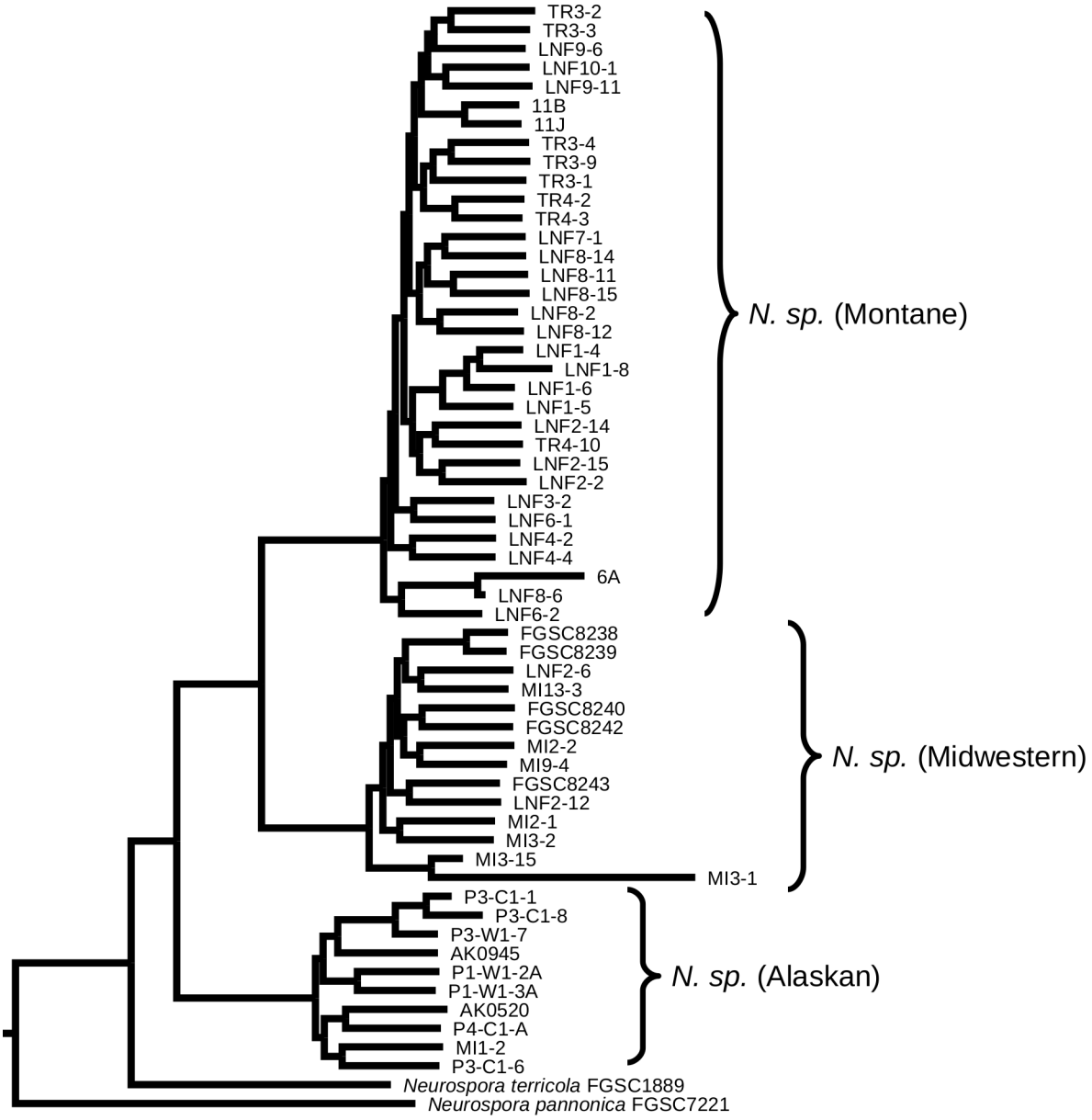

Figure 1.4: Phylogeny of *Neurospora* genomes within the "C" clade (Fig. 1.1).

appear to conidiate poorly and don't produce the large tufts of bright orange-yellow conidia as other *N. discreta* strains when grown under the same conditions. However, these coloration and conidiation differences fall within the range observed for other *Neurospora* species, such as *N. crassa* (Glass, 2018).

Most other homothallic *Neurospora* species contain genes from both mating type idiomorphs, *mata* and *matA*, often with the full complement of *mata-1*, *matA-1*, *matA-2*, and *matA-3* (Gioti *et al.*, 2012). We sought to confirm if the homothallic-like strains, the heterothallic-like strains, both, or niether shared this common adaptation for homothallism by searching for the homologues of *N. crassa* (OR74A v. 12)'s versions of these genes with BLAST. Surprisingly, 5 strains contain only the *mata-1* gene in the mating type locus, as is typical for *mata* heterothallic strains. One phenotypically heterothallic strain is of an unknown type due to poor sequence quality. The genome sequence of the last strain, which is phenotypically homothallic, appears to be of multiple strains or a polyploid, as the genome assembly was approximately 3 times the size of the other strains (126 Mbp) and contains both one homologue of the *matA* mating type idiomorph (consisting of *matA-1*, *matA-2*, and *matA-3*) and two homologues of *mata-1*, all on separate scaffolds. This genome sequence additionally contains 2 and 3 homologues of NCU01961 and NCU01956, respectively, which typically flank the mating type locus Gioti *et al.* (2012).

Being isolated from an unorthodox substrate (i.e. soil instead of burned trees), having an unusual and inconsistent phenotype of mixed homothallism and apparent heterothallism, and having a peculiar architecture in their mating type locus, these strains are an oddity within *N. discreta* and further investigation is needed to characterize their breeding system and ecology.

Endolichenic Neurospora

Of the endolichenic strains isolated from Alaska, eight clustered with aconidial and pseudohomothallic *N. tetraspora*, and two with a new clade of aconidial and heterothallic *Neurospora* (Table 1.2). As a pseudohomothallic species, *N. tetraspora* is capable of having nuclei of both mating types present in the same strain. Based upon a heterothallic-like morphology and the presence of only a single mating type idiomorph in the genome, we judged 3/16 soil-isolated *N. tetraspora* strains to be haploid, but judged all eight of the endolichenic strains to be haploid (Table 1.2). While the number of haploid strains isolated from soil likely reflects our biased sampling in choosing between small haploid spores and large dikaryotic ones under the dissecting microscope, the preponderance of haploid strains isolated from lichen is in stark contrast to the conidial pseudohomothallic *N. tetrasperma*, where only 10% of ascospores from heterokaryotic cultures are spontaneously haploid (Raju, 1992).

The discovery of endolichenic as well as soil-derived strains shows that *N. tetraspora* and the aconidial heterothallic species can live part of their lifecycle as endosymbionts, likely in a haploid form, and have sexual spores that reside in the soil. Further support for the endosymbiont hypothesis comes from ITS sequences indicating that *Neurospora* is present in

grasses and other plants of more southern latitudes (Herrera *et al.*, 2010; Loro *et al.*, 2012; Khidir *et al.*, 2010).

Many aconidial *Neurospora* have been isolated from dung (Garcia *et al.*, 2004) and other fungi have been reported to be both endophytic and associated with dung (Herrera *et al.*, 2010; Mrqueza *et al.*, 2012). Whether *Neurospora* consumes dead plant matter after plants burn, are grazed upon, shed their leaves, or die from other causes, it makes sense for this opportunistic saprobe to be lying in wait. We hypothesize that different *Neurospora* species have evolved to withstand different disturbances and feast upon their recently deceased hosts.

Reproductive Diversity in Neurospora

With the discovery of conidial homothallic *Neurospora*, species of this genus with all combinations of conidiation and heterothalism, pseudohomothallism, or homothallism are now known, with homothallism and pseudohomothallism likely having been evolved mutliple times (Gioti *et al.*, 2012). In our study of the population genetics of these strains, we show that while homothallic species are self-fertile, the Cailleux IV clade appears to be outcrossing as much as heterothallic species (Chapter 2). The different reproductive modes of *Neurospora* may not have evolved in response to intrinsic pressures, but rather extrinsic ecological pressures.

Rather than homothallism evolving due to the automatic or other advantage of selfing, it may be an adaptation for universal mating compatibility. Universal compatibility may be advantageous when access to mates is limited, such as after being dispersed to a new habitat, while selfing could also allow persistence in new environments by producing hardy spores. On the other hand, selection to maintain heterothallism could be relaxed when the risk of selfing is low due to an abundance of access to mates.

Different potential niches may alter dispersal and mixing of *Neurospora* populations, shifting the balance of forces affecting breeding systems. Grazing animals could transport *Neurospora* great distances or mix together many individuals within their guts. In Alaska, where lichen is an important part of the ecosystem (Cornelissen *et al.*, 2004), Caribou or other herbivores that feed on tundra lichen may make up an integral part of these species' life cycles. Fire-adapted *Neurospora* may require conidia to colonize an area rapidly after a fire, which may put them at greater risk of encountering individuals of the same genet in nature. Species that reside in trees and consume deciduous leaves every season may be regularly dispersed within falling leaves. Further understanding of *Neurospora*'s habitat and lifecycle will help us to understand the reasons behind *Neurospora*'s diverse reproductive systems.

Conclusions

1. The full spectrum of Ascomycete reproductive systems are represented in *Neurospora*. We report the first conidial homothallic *Neurospora* strains.

2. *Neurospora* has diverse unknown ecologies. We confirm that *Neurospora* exists as an endosymbiont in nature, adding to a growing body of evidence that *Neurospora* is a genus of saprobic endosymbionts.

3. Soil sampling has proven to be an efficient and repeatable means to collect natural populations of diverse *Neurospora* species. We have sampled a collection of aconidial *Neurospora* suitable for population level analysis.

4. The two disjoint pools of *Neurospora*, one from burned plants or compost piles, and the other from ascospores in the soil, suggests different habitats or lifecycles for these groups.

1.3 Materials and Methods

Culture, DNA Extraction, and Genome Assembly

We grew cultures and extracted DNA following the protocols of Gladieux *et al.* (2015). The UC Berkeley Functional Genomics Laboratory prepared libraries, and the Vincent J. Coates Genomic Sequencing Laboratory sequenced the strains on Illumina HiSeq machines. We assembled genomes *de novo* for all sequenced strains using the A5 assembly pipeline (Tritt *et al.*, 2012; Coil *et al.*, 2015). Details about libraries, sequencing, and assembly statistics are given in Table A.2.

Phylogenetic Analysis

We first aligned all strains to *N. crassa* and discarded genomes that had less than 10 Mbp of alignment as unlikely to be *Neurospora*. From the remaining set of genomes we built a phylogenetic tree using the whole genome phylogenetic analysis pipeline described in Chapter 3, which makes use of MUMmer (Kurtz, 1999; Delcher *et al.*, 1999, 2002; Kurtz *et al.*, 2004), MAFFT (Katoh *et al.*, 2002; Katoh and Standley, 2013), a rewrite of TIGER (Cummins and McInerney (2011), Chapter 3) that employs the partitioning method of Rota *et al.* (2017), and IQTree (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017; Thi Hoang *et al.*, 2018; Chernomor *et al.*, 2016).

Our pipeline for the whole-genome phylogenetic analysis and our rewrite of TIGER can be found at `https://github.com/channsoden/hannsoden-bioinformatics/`.

Chapter 2

Lack of Linkage and Efficient Selection Evince Outcrossing in Self-Fertile *Neurospora*

Christopher Hann-Soden, Pierre Gladieux, Liliam Montoya, and John W. Taylor

Abstract

Within the model mold genus, *Neurospora*, at least nine transitions from self-sterility to self-fertility have occured. Current theory posits that these transitions are driven by the intrinsic reproductive advantages of selfing, but that long-term costs make these lineages evolutionary dead-ends. However, the relationship between the ability to self and actual patterns of inbreeding has not been established within *Neurospora*. We compare populations of self-sterile and self-fertile *Neurospora*, as well as populations that produce mitotic spores and those that do not, to understand how differences in reproductive systems alter demography. Using a dataset of nearly 200 genomes that represent all unstudied combinations of mitotic and meiotic reproductive systems, we find evidence that a diverse group of self-fertile *Neurospora* have a similar pattern of outbreeding to self-sterile species. However, we also find evidence that self-fertile species are highly inbred at limited scales, which is associated with decreased efficacy of selection. Taken together, these results show how different mating paradigms at different scales work together to shape populations.

2.1 Introduction

The causes and consequences of sexual reproduction have perplexed biologists at least since the time of Darwin, who stated "We do not even in the least know the final cause of sexuality" (1862). Since then science has produced a great deal of theory and data on

the subject, and today sex remains a perennial topic. Much of the theory of the evolution of sex, in needing to explain the near universal preponderance of sex in Eukaryotes, posits advantages that outweight sex's numerous disadvantages. The existence of asexual species counterpoints this theory and the study of asexual species refines the theory by allowing us to characterize the parameters under which sex is lost as well as the consequences of doing so. Ascomycete fungi have diverse sexual and asexual systems, and the full spectrum of Ascomycete breeding systems is represented in the genus, *Neurospora* (Chapter 1). We compare the population structure and patterns of selection across *Neurospora* species with disparate breeding systems in order to better understand the evolutionary forces acting upon breeding systems.

The Dead-End Hypothesis of Sexual Evolution

Of many biological phenomena, sex is notable in that it provides no direct fitness advantage to the individual; neither in terms of offspring produced, energetics, defenses, nor alteration of the habitat. Rather, sex incurs some severe costs ranging from expenditure of metabolic resources to increased risk of disease or predation. The effect of sexual outbreeding, relative to other modes of reproduction, is to create recombined genotypes in the next generation. Sex must clearly play a role not in organismal biology, but in the evolution of a species. Sex is thus a meta-adaptation which has evolved to direct evolution itself.

Even within an evolutionary context, however, sex poses some serious disadvantages. While sex increases the diversity of a population, sex can also break apart co-adapted alleles (Otto, 2009). Furthermore, in anisogamous species males incur a two-fold transmission disadvantage (Smith, 1971; W.M. Lewis, 1987). Still, the vast majority of eukaryotes undergo sex, even if rarely, despite its tremendous costs (Otto, 2009). The increased diversity produced by sex is thus clearly advantageous, an inference supported by experiments showing that sexual populations can adapt faster than asexual populations (Goddard *et al.*, 2005).

The dead-end hypothesis seeks to explain the paradox of sex: Lineages that forgo sex (such as through self-crossing) are evolutionary dead-ends in that either by accumulation of deleterious alleles or a susceptibility to disturbances resulting from a lack of diversity, asexual or inbred lineages are at increased risk of extinction (Igic and Busch, 2013). The dead-end hypothesis admits that clonality may bear short-term advantages, but requires that any strictly clonal species be short lived. How short lived clonal species must be, though, is an open question (Birky, 2010). Current research aims to weigh the balance of forces that lead to sexual evolution and to explain the considerable variation in degree of outbreeding. Indeed, while sex is clearly indispensable, vertebrate animals are exceptional in their limited clonal reproduction compared to the rest of *Eukarya* (Avise, 2015).

Neurospora: A Model for Sexual Evolution

An ideal natural experiment to study the effects of sex on wild populations would require numerous closely related species that vary only in their independently evolved abilities to

reproduce as clones, as recombined progeny, or by both means. Fungi provide an excellent system for the study of sexual evolution because their diverse reproductive modes approximate such systems. Additionally, fungi are abundant, often culturable, potentially immortal in lab, and often have small haploid genomes that facilitate genomics.

The genus *Neurospora* has emerged as a model system for multicellular fungi (Gladieux *et al.*, prep). *Neurospora* have low repetitive content in their genomes due to multiple genomic defense mechanisms (Galagan and Selker, 2004; Shiu *et al.*, 2001), and deep knowledge about the molecular model, *N. crassa*, provides a strong reference point for the mechanisms that underly evolution (Davis, 2000; Davis and Perkins, 2002). The cosmopolitan nature and ease of sampling *Neurospora* has enabled population genomic studies by providing large collections of wild isolates (Perkins and Turner, 1988; Turner *et al.*, 2001). To date, comparative genomics of *Neurospora* populations has been used to find evidence of natural selection, identify adaptive genes, and discern demographic history (Ellison *et al.*, 2011b; Gladieux *et al.*, 2015), while a comparison of *Neurospora* genomes has shown evidence for unidirectional transitions to selfing associated with genomic degeneration (Gioti *et al.*, 2012, 2013).

In terms of reproductive systems, conidial *Neurospora* species produce mitotic asexual spores while aconidial species do not, and various species display one of the three breeding systems; heterothallism, homothallism, or pseudohomothallism. Heterothallic species are haploid and have two mating types that each must mate with the opposite, thus rendering intra-haploid selfing impossible (although intra-diploid selfing, as in selfing plants, is still possible). Homothallic species are self-fertile and self readily in culture. While they undergo meiosis, this haploid selfing is quite nearly clonal. Homothallic species typically contain genes from both mating types in a single haploid strain. Gioti *et al.* (2012) showed that each of several major homothallic lineages had unique organizations of mating type genes, while the mating type locus is conserved in heterothallic lineages, leading them to conclude that heterothallism is ancestral in *Neurospora* and homothallism has evolved multiple times. Lastly, in pseudohomothallic species dikaryotic strains contain haploid nuclei of both mating types. In this system nuclei of opposite mating type are typically packaged into single ascospores after meiosis, although single nuclei are sometimes packaged into haploid spores to create strains that behave as heterothallics.

Until now, studies have focused on conidial heterothallic species (especially *N. crassa* and *N. discreta*) and the conidial pseudohomothallic species, *N. tetrasperma*. Conidial homothallic species have been unknown, but we recently discovered the first species of this type (Chapter 1). Gioti *et al.* (2013) advanced the understanding of aconidial *Neurospora* when they sequenced the genomes of four aconidial homothallic species and compared them to the three other sequenced *Neurospora* to report that the homothallic species had: a higher rate of nonsynonymous substitutions to synonymous substitutions; reduced codon usage bias in highly expressed genes; and reduced transmission of transposons. Based on their findings, Gioti *et al.* (2013) concluded that the homothallic species appeared to be undergoing genomic degradation and relaxed selection due to a lack of outbreeding, in concordance with the dead-end hypothesis. This interpretation, however, relies on the untested assumption that homothallic *Neurospora* are in fact clonal, or at least much more so than other species

- a limitation which Gioti *et al.* recognize. In contrast, aconidial heterothallic and pseudohomothallic species of *Neurospora*, although known, are poorly studied.

Ancient Asexual Scandals

Many organisms have been supposed to be clonal until new evidence has shown otherwise. Perhaps the most famous "asexual scandal," the Bdelloid rotifers, were recently shown not to be "scandalous" at all as they exchange genes within and between species through either a sexual or parasexual system (Gladyshev *et al.*, 2008; Signorovitch *et al.*, 2015; Debortoli *et al.*, 2016). Within the fungi, the Glomeromycota form essential symbioses with around 80% of vascular plants (Smith and Read, 2008), and due to their unique biology were hypothesized to be asexual (Jany and Pawlowska, 2010). However, genetic and population genetic evidence has now shown the presence of recombined haplotypes of Glomeromycetes (Pawlowska and Taylor, 2004; Croll and Sanders, 2009; den Bakker *et al.*, 2010). More recently, genomic evidence was used to identify putative mating type genes as well as both haploid and dikaryotic strains, as is typical in other sexual fungi (Ropars *et al.*, 2016). With the molecular revolution, swathes of fungal diversity once regarded as either asexual (anamorphs) or sexual (teleomorphs) have had to be recognized as coupled halves of the same organisms' life cycles (Taylor, 2011). Even within *Neurospora*, the pseudohomothallic *N. tetrasperma* was hypothesized to be inbred until maintenance of intraspecific polymorphisms in the rapidly evolving *het* loci showed it to be outbreeding (Powell *et al.*, 2001). Rather than an example of clonality through selfing, *N. tetrasperma* has since become a model for the evolution of sex chromosomes because the different genetic contents at the mating type loci create a region of suppressed recombination similar to early sex chromosomes (Menkis *et al.*, 2008). Moreover, introgression from heterothallic species appears to be important for the maintenance of the mating type loci in this species (Sun *et al.*, 2012).

These debunked scandals demonstrate the necessity to delineate breeding systems (the biological capability of an organism, such as heterothallism) from mating systems (the mating patterns in nature, such as inbreeding or outbreeding) (Billiard *et al.*, 2012). As no population of homothallic *Neurospora* has been studied there is a pressing need for a close examination of their mating systems before strong claims can be made regarding their breeding system.

A Comparison of Breeding Systems in Neurospora

To better understand the consequences of different breeding systems in *Neurospora* and fungi generally, we sought to compare the demography and patterns of evolution between *Neurospora* with diverse breeding systems. We examine populations of all six combinations of breeding systems and conidiation except for conidial pseudohomothallism, which has been examined by others (Ellison *et al.*, 2011a; Corcoran *et al.*, 2016; Sun *et al.*, 2017). In light of Gioti *et al.* (2013)'s evidence that homothallic *Neurospora* are evolutionary dead-ends, we first sought to elucidate the relationship between *Neurospora*'s breeding systems and its

mating systems, if any. That is, we ask the question, "Are homothallic *Neurospora* inbred or outbred?" We further sought to measure the age of these lineages in terms of nucleotide diversity. If homothallic species are, as suspected, inbred, then we should not expect to find diverse long-lived lineages according to the dead-end hypothesis. Lastly, we sought generally to see if there are evolutionary or biogeographic consequences, such as reduced efficacy of selection or increased distribution, for transitioning from heterothallism to homothallism.

We show that all demonstrably long-lived species are outbred, including both heterothallic and pseudohomothallic species and, unexpectedly, a group of homothallic *Neurospora*. We find that although different *Neurospora* species have similar reproductive modes, they appear to have widely diverse demographic and geographic patterns which may reflect their particular ecologies. Furthermore, we find that the mixed mating system of selfing and outcrossing in homothallic *Neurospora* species leads to the seemingly paradoxical observation of more evidence for frequent recombination between distant strains than between close relatives. Overall, the different evolutionary paradigms at different scales suggests that different evolutionary forces prevail at these scales. Furthermore, this model genus demonstrates the flexibility of fungi to maintain gene flow through counter-intuitive and as of yet unknown means.

2.2 Results

Strains and Populations

We analyzed a collection of genome sequences from 192 strains of *Neurospora* that we present in Chapter 1. This dataset includes 113 strains which we recently isolated from soils in North America, 55 strains of *N. discreta* isolated from burned pine trees in North America, 10 strains isolated from within lichens in Alaska, and 14 strains of various species collected around the world. The approximate sampling locations of strains from North America used in population analysis are shown in Fig. 2.1. Details of strains and genome sequences are given in Table A.2.

We had previously clustered the strains into groups via whole genome phylogenetic analysis (Chapter 1). We identified 13 clusters of more than one genome with 100% branch support by both bootstrap and Shimodiara-Hasegawa-like approximate Likelihood Ratio Test (SH-aLRT). From each cluster we selected the genome with the best assembly, as judged by the highest N50 score, to serve as a reference. We reconstructed the whole genome phylogeny using only the reference genomes as well as the previously published *Neurospora* and *Sordaria macrospora* genomes (Gioti *et al.*, 2013; Galagan *et al.*, 2003; Nowrousian *et al.*, 2010). The resulting phylogeny, made from 11 Mbp of aligned sequence, has 100% support for all branches by both bootstrap and SH-aLRT, and is shown along with the number of genomes from each cluster in Fig. 2.3.

Three clades contained several closely related clusters: first, the previously published *N. discreta* PS4B complex; second, a clade containing three clusters of entirely unnamed

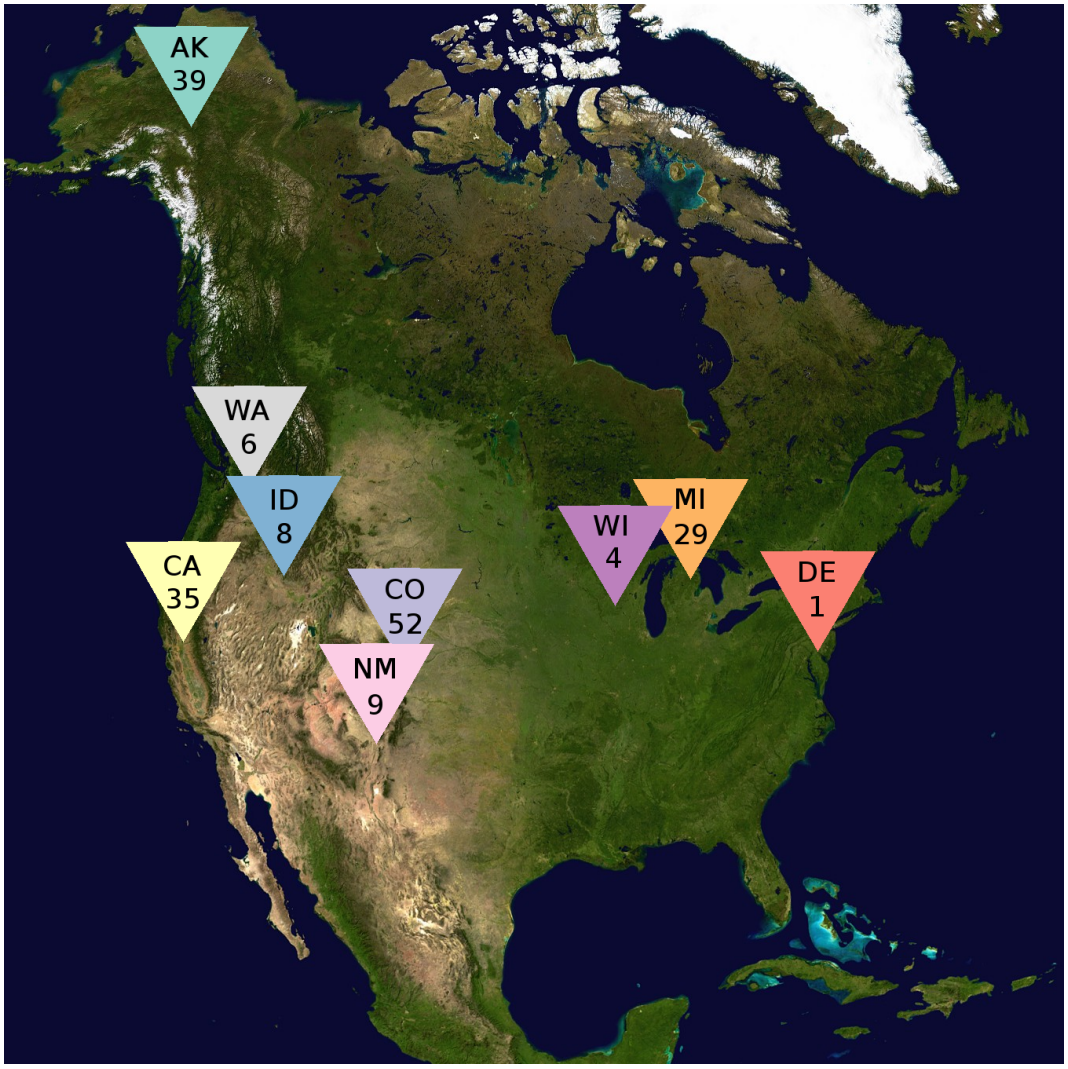

Figure 2.1: Map of sampling regions by state. Each colored maker shows a state where samples were collected and the number of strains collected from the state.

aconidial, heterothallic strains; and third, a homothallic clade consisting of two small clusters and several doubletons and singletons. Not knowing the degree to which these clusters have diverged, we analyzed these clades in three manners (Fig. 2.2): collectively, treating each cluster as a distinct population within a species; collectively, treating all strains as members of a single population; and individually, treating each cluster as a distinct species.

Where we discovered strains with identical genomes within clusters we retained just one of the strains as representative of the genet. Whether or not we employed this type of clone correction, we obtained the same qualitative results and, unless otherwise stated, base our analyses on unique genets.

Where clusters contain strains that have been taxonomically identified we use the existing taxonomic name or population label. Within the Cailleux IV clade of aconidial homothlalics (Cailleux, 1971; Garcia *et al.*, 2004), we label clades by representative strains (e.g. LNF6-4, 9A, and 4A). We label the three aconidial heterothallic species by the descriptive geographic adjectives, Midwestern, Montane, and Alaskan.

Within each group we aligned the reads to the reference (or best reference for larger clades) to identify variants. We used ADMIXTURE (Alexander *et al.*, 2009) to learn and assign population ancestry to the strains within each group. For each inferred population we calculated population statistics and linkage disequilibrium (LD) using custom scripts available at `https://github.com/channsoden/neurospora_popgen`.

Population Structure of *Neurospora* Species

Conidial Heterothallic *Neurospora discreta*

In our reanalysis of the *N. discreta* PS4B complex, we again recognize the California-Washington (CAWA), New Mexico-Washington (NMWA), and Alaska-Europe (AKEU) populations. ADMIXTURE and phylogenetic analysis reveals these three distinct populations, two of which have overlapping ranges in Washington (Figs. 2.3 and 2.4). The presence of a

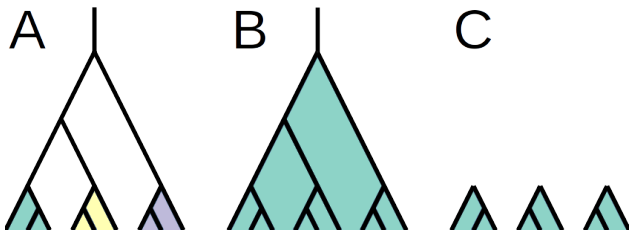

Figure 2.2: Conceptual diagram of different scales of analysis. A) We analyzed clusters as seperate populations for between group measurements such as Dxy and Fst. B) We used the null model of a single population to check for recombination between clusters that is not observed within clusters. C) We analyzed clusters individually as seperate species for the most accurate measurements of within-group statistics such as π and to check for fine-grained structure.

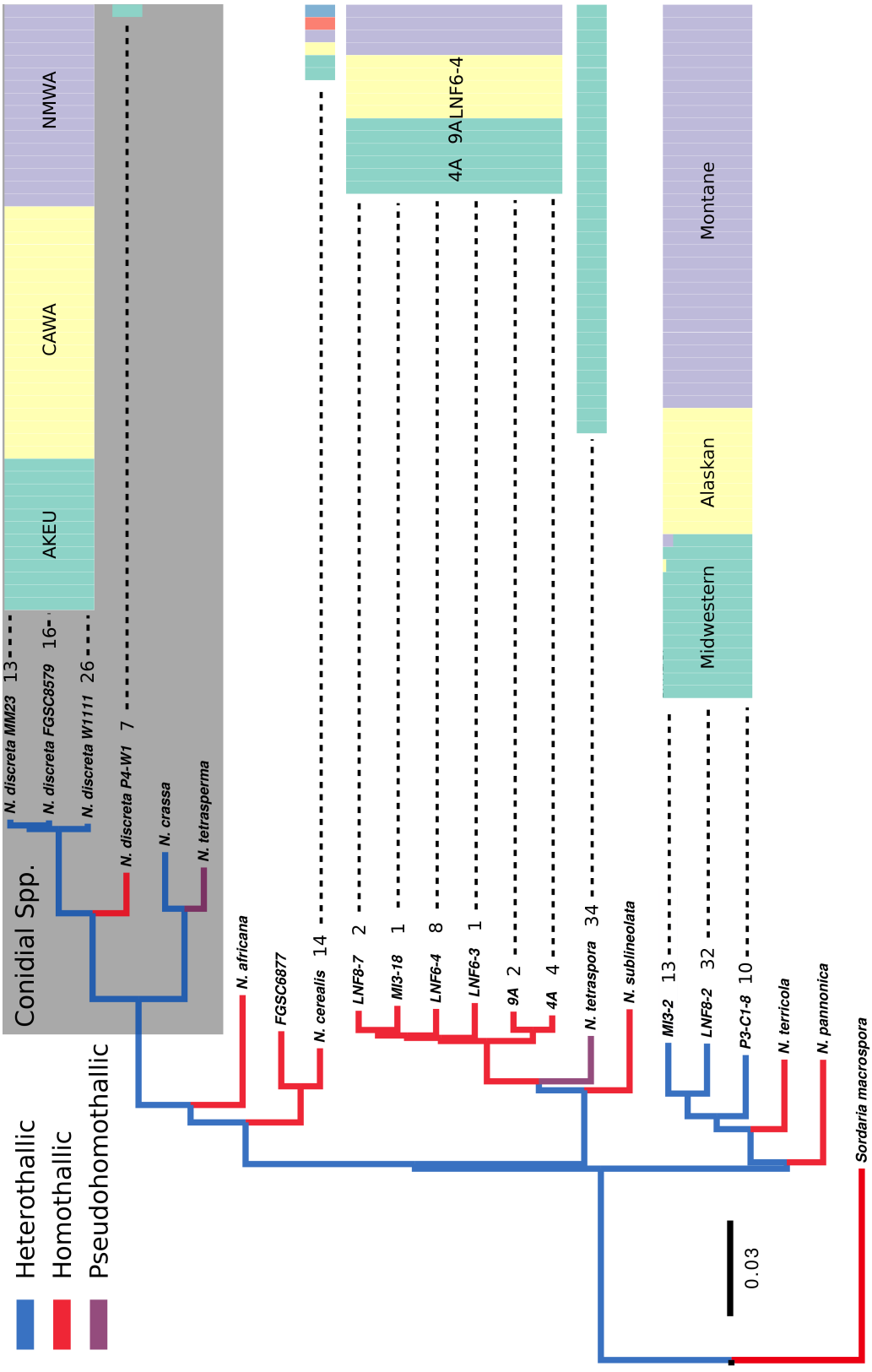

Figure 2.3: Caption on next page.

Figure 2.3: Whole genome phylogeny of representative *Neurospora* species from each major clade and *S. macrospora*.The representative strains for alternatively labeled clades are MM23 (AKEU), FGSC8579 (NMWA), W1111 (CAWA), MI3-2 (Midwestern), LNF8-2 (Montane), and P3-C1-8 (Alaskan). The number of genome sequences within each clade is connected by a dashed line to the results of ADMIXTURE analysis after removal of clones. In the ADMIXTURE plots each bar represents an individual genome, colored by the proportion of the genome that belongs to each inferred population. The colors are arbitrary, and identical colors between plots purely coincidental.

clone out of 13 strains in the AKEU population and 6 clones out of 26 strains in the CAWA population indicates some level of asexual propagation in this conidial species (Fig. 2.5A). In spite of the presence of clones, the patterns of LD were consistent with histories of recombination, decaying to 50% of their maximum values (LD_{50}) at 785 bp, 988 bp, and 18,573 bp for CAWA, NMWA, and AKEU, respectively (Fig. 2.5B). These values indicate a history of recombination for the CAWA and NMWA populations that is similar to the other sequenced conidial heterothallic, *N. crassa* (700-850 bp) and somewhat more recombined than the heterothallic yeast, *Saccharomyces cerevisiae* (3,000 bp), while there has been somwhat less recombination in the AKEU population than the homothallic yeast, *S. paradoxus* (9,000 bp) (Ellison *et al.*, 2011b).

Conidial Homothallic *Neurospora discreta*

Seven conidiating strains including P4-W1 fall into the *N. discreta* species complex based on a multi-locus phylogeny, and are genetically, morphologically, and perhaps ecologically distinct (Chapter 1). These strains were isolated from ascospores in the soil, while all other *N. discreta* strains were isolated from conidia on burned trees. Upon the same media some strains appear heterothallic, while others produce 8 spored perithecia as homothallic species do. Lastly, 5/7 strains of both heterothallic appearance and homothallic appearance have only the *mata* mating type idiomorph, while only P1-W1 has both the *mata* and *matA* idiomorphs, and the last strain is of unknown type.

Surprisingly, based upon alignment of all strains' reads to P4-W1 we find no evidence that any of these 7 strains are not identical clones of each other, as there are no more SNPs discernible beyond sequencing error. However, in addition to both mating type idiomorphs, the assembly of P1-W1 has a large amount of extra content compared to the assembly of P4-W1, including two copies of the *mata-1* gene and being around three times as large. Despite this, a higher percentage of reads from P1-W1 aligned to the assembly of P4-W1 than P4-W1's reads did to its own assembly, and these reads evidenced only 368 SNPs between the two strains, while aligning P4-W1's reads to its own assembly evidenced almost as many (336) SNPs that must be erroneous.

While we cannot account for the phenotypic variation between these strains or the bizarre genome of P1-W1, this species apparently possesses two methods of clonal reproduction, by

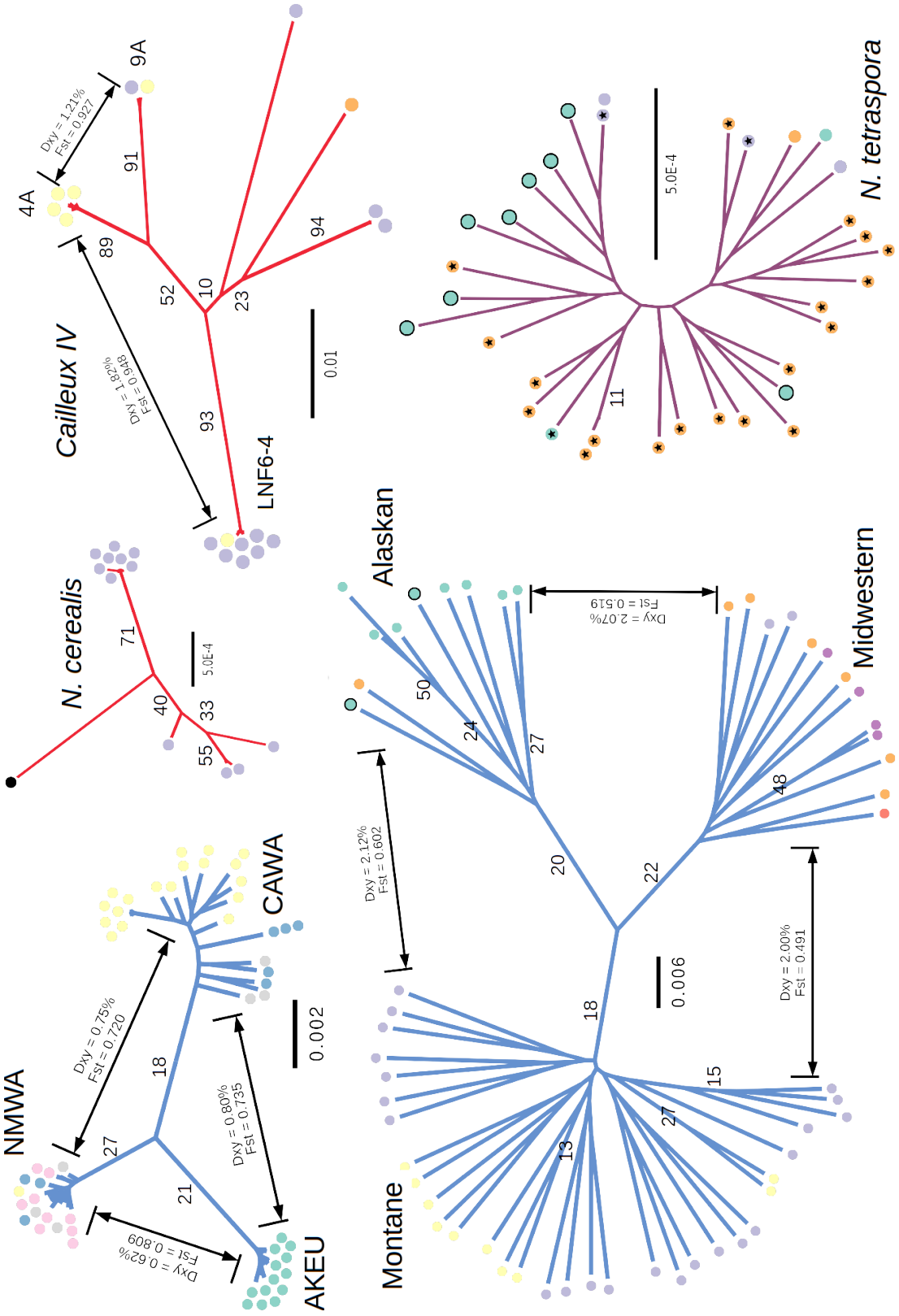

Figure 2.4: Caption on next page.

Figure 2.4: Whole genome phylogenies for each of the major clades of *Neurospora* studied here. Branches are color-coded as in Fig. 2.3. Circles at the tips of branches represent individual strains colored by their geographic origin as in (Fig. 2.1), except the black circle which represents the type strain of *N. cerealis* of unknown origin. Black edges around circles indicate strains that were isolated from within lichens and stars indicate dikaryotic strains. The percent of orthologues that support each internal branch is given for internal branches when over 10%. Scale bars for each phylogeny are in substitutions per base.

conidia and by self-fertilized ascospores, and it is therefore plausible that all seven strains should be recent clones of each other. These strains were isolated from meiotic spores in three separate soil samples, demonstrating that this clone is abundantly selfing at the collection site.

Aconidial Heterothallic Species

We identified three aconidial heterothallic species that demonstrate no capacity for asexual reproduction. Each species forms an unstructured star-like phylogeny with more divergence between groups ($D_A = 1.09\%, 1.16\%, 1.33\%$) than within each group ($\pi = 0.78\%$, 0.79%, and 1.03% analyzed together; Fig. 2.5C analyzed separately), and the smallest branch distance between the three (Midwestern to Montane, 0.0367 substitutions per site) is greater than that between clearly established species such as the conidiating heterothallic *N. crassa* and the conidiating pseudohomothallic *N. tetrasperma* (0.0352 substitutions per site).

LD decays extremely rapidly, reaching 50% of the maximum after only 33 bp, 36 bp, and 87 bp for the Montane, Midwestern, and Alaskan populations, respectively (Fig. 2.5B). This steep decay of LD in these unstructured populations indicates a high frequency of recombination relative to processes that create linkage.

These three species form outbred populations that, like the populations of *N. discreta*, have overlapping but distinct geographic ranges (Fig. 2.4). The Montane population appears to be a sub-alpine species, being isolated from elevations of 1,600m and 1,800m in the mountains of Colorado and California, respectively. The Alaskan population was all isolated from the subarctic of Alaska, except for a single strain from the old growth pine forest of Hartwick Pines State Park, Michigan. Apart from the one Alaskan isolate, however, Michigan is dominated by individuals of the Midwestern population, which reaches across the North East and Midwest at least from the site in Colorado to Wisconsin and Delaware.

Aconidial Pseudohomothallic *Neurospora tetraspora*

Like the heterothallic species, the aconidial pseudohomothallic *N. tetraspora* appears highly outbred. Cross-validation by ADMIXTURE strongly supports a single population, and LD decay is similar to that in *N. discreta*, decaying to 50% in only 20 bp, but not reaching equilibrium until around 170 Kbp (Fig. 2.5B). Despite producing self-fertile dikaryotic

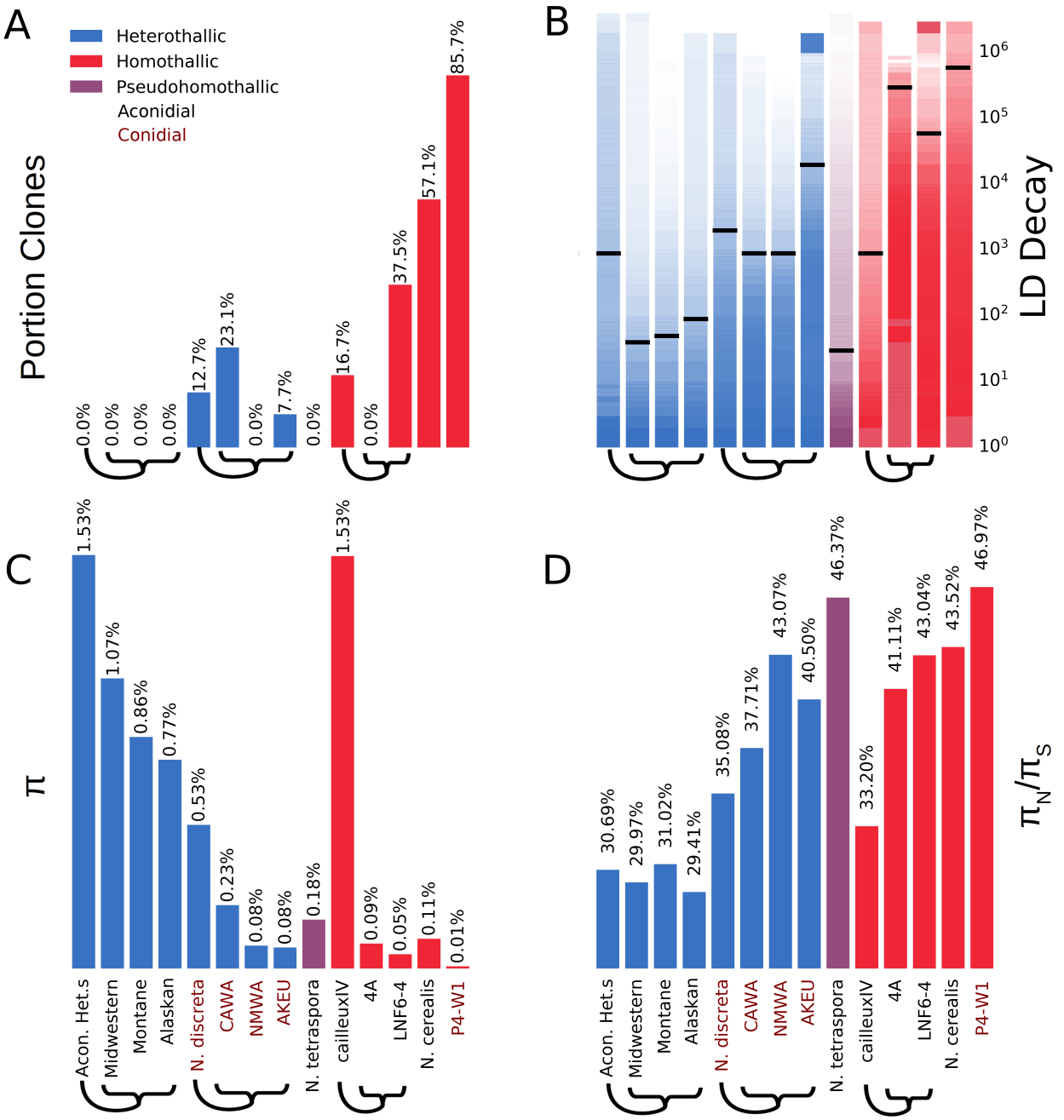

Figure 2.5: Outbred lineages are more diverse and experience greater purifying selection than inbred lineages. Within each population bars show (A) the percent of each population that we discarded after clone correction, (B) the distance at which LD decays to varying extents, (C) the pairwise nucleotide diversity (π), and (D) the ratio of nonsynonymous to synonymous diversity (π_N/π_S). In (B), fully opaque colors represent the maximum LD observed of any distance class, transparency (white) represents no LD, and black bars show LD_{50} values. The labels of conidial groups are shown in brown, while aconidial groups are in black. The three groups, "Acon. Het.s," *N. discreta*, and "Cailleux IV" contain the subgroups "Midwestern," "Montane," and "Alaskan;" AKEU, NMWA, and CAWA; and LNF6-4 and 4A; respectively.

spores, we found no clones among the 34 strains. Unlike the heterothallic species, however, *N. tetraspora* does not show the same degree of geographic structure. The single population spans at least from Michigan to Colorado and all the way North to Alaska.

Aconidial Homothallic Species

The population of *N. cerealis* collected from Colorado appears similar to, but much more clonal than, the populations of heterothallic *N. discreta*. All eight strains isolated from a single soil sample are clones of each other ($mean = 385$ variants between strains, $sd = 62$), leaving only 6 unique genets (Fig. 2.5A), of which two more strains from another soil sample are extremely similar but not identical (2,426 variants). The type specimen (FGSC 959) is the most diverged from the other strains, but unfortunately we could find no record of its origin. Sequence diversity in *N. cerealis* was similar to *N. discreta* (Fig. 2.5C), but LD decayed over a much greater distance than heterothallic *N. discreta*, not reaching 50% until the farthest observed distances at more than 760 Kbp (Fig. 2.5B).

Cailleux IV contained two groups and two doubletons that each have similar pairwise diversity to the other *Neurospora* species except the very diverse aconidial heterothallic species (Fig. 2.5C). The group containing LNF6-4 included three redundant clones from the same soil samples and five genets (Fig. 2.5A), two of which were isolated from the same soil sample. In the two groups large enough to measure, LD decayed over a distance up to three orders of magnitude greater than *N. discreta* ($LD_{50} = 60$ Kbp in LNF6-4, and $LD_{50} = 372$ Kbp in 4A, Fig. 2.5B).

At this finest scale, all homothallic species meet the expectations of dominantly clonal species. However, as Cailleux IV is a monophyletic clade of homothallic *Neurospora*, the different lineages within Cailleux IV could represent independent derivations of homothallism or could have arisen from a single event. To determine if a single or multiple transitions to homothallism occured within Cailleux IV, we compared the architecture of the genes in the mating type locus in the manner of Gioti *et al.* (2012). While we detected some structural variation within the mating type locus, the position and orientation of mating type genes was consistent among all strains, indicating a single origin of homothallism (Fig. 2.6). We therefore consider two possibilities for the evolution of the Cailleux IV clade, both of which challenge different assumptions of the evolution of selfing species.

First, each group within Cailleux IV may be separate clonal lineages from the common homothallic ancestor. Given the divergence between these lineages (0.0256-.0370 substitutions per base to the root) they must then be members of a relatively long-lived group of homothallics. Unless the generation time is substantially lower or the mutation rate substantially higher within these lineages, Cailleux IV would be approximately as old as the *N. discreta* species complex or the primates (Rogers and Gibbs, 2014). The existence of such a lineage is unlikely under the dead-end hypothesis, although groups at least as diverse have been argued to represent the tail end of the distribution of asexual lineage ages (Birky, 2010).

Second, Cailleux IV may be a single species and each group within Cailleux IV mostly-clonal propagations of recombining genets. This would imply that homothallism, at least in Cailleux IV, is not an adaptation for clonality at all, but a co-option of the meiotic process for propagation.

To test the second hypothesis, we measured linkage between variants identified within the entire clade by analyzing one representative from each of the six major lineages as a single population. In contrast to the slow linkage decay within groups, linkage between groups decayed rapidly ($LD_{50} = 814$ bp, Fig. 2.5B). We performed a similar analysis on the three populations of heterothallic *N. discreta* and the three aconidial heterothallic species, but observed patterns of LD between groups that were similar to or slightly elevated above the patterns within groups, as expected for diverged populations. The only explanation for the distance decay of LD we can offer is recombination among the lineages, and since we used only a single representative from each of the lineages this pattern of LD decay cannot be explained by within-group recombination.

Similar to the reduced linkage between groups, while around 90% of phylogenies from individual orthologous segments agreed on the clustering of groups, the internal branches of the tree are only supported by 10%-52% of orthologous segments (Fig. 2.4). We furthermore counted the number of sites with a tree-consistent pattern of alleles (i.e. those site where one allele is shared by a monophyletic group). The majority of sites indicated either convergent evolution or recombination, as only 128,932 out of 284,946 sites (45%) were tree consistent. Despite the fact that most sites do not support the tree, the tree is preferred over alternative configurations because its internal branches are supported by 18,489-63,267 sites, while the two most common tree-inconsistent patterns occurred only 19,082 and 15,307 times. Taken together, these data indicate conflicting signals between different regions of the genome, a pattern that is incompatible with clonality alone.

Clonality Leads to Relaxed Selection on Substitutions but not Rearrangements

Gioti *et al.* (2013) found an elevated ratio of nonsynonymous to synonymous mutations (D_N/D_S) in homothallic lineages compared to non-homothallic lineages. Assuming that nonsynonymous mutations are typically deleterious and synonymous mutations neutral, this elevation indicates a relaxation of selection due to the inability of non-recombining lineages to remove deleterious alleles and supports the dead-end hypothesis for *Neurospora*. We sought to see if this conclusion extends to the variation that exists within populations of homothallic lineages and when using a direct measurement of the relative rate of recombination (LD decay), rather than breeding system as a proxy for recombination.

To this end, we determined the ratio of pairwise site diversity at nonsynonymous sites to synonymous sites (π_N/π_S) within populations and then used the ratios for comparison among populations. We found that those populations with the shortest decay of linkage (i.e. the most recombined) often had the most pairwise diversity and the lowest π_N/π_S (Fig.

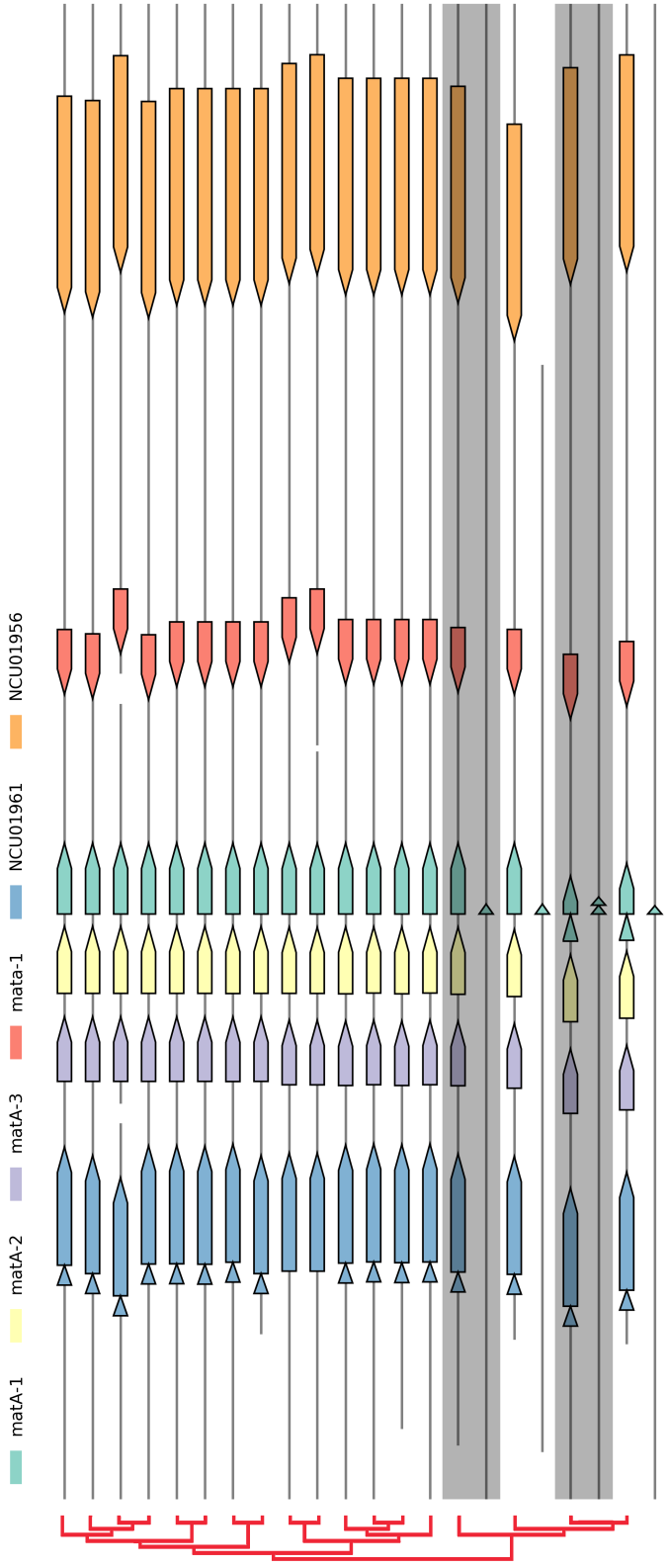

Figure 2.6: Caption on next page.

Figure 2.6: Genetic architecture of the mating type locus in all strains of Cailleux IV. Grey lines represent scaffolds of the genome assemblies, while colored arrows represent genes or gene segments identified by BLAST using the *N. crassa* orthologues. When gene duplication across multiple scaffolds is evident we show these multiple scaffolds of the genome in alternating light and dark shaded bands.

2.5). From this result, we infer that the increase in the number of nonsynonymous mutations seen in poorly mixed lineages is likely due to a lack of purifying selection rather than an increased mutational rate. However, this tendency, which is consistent with the theoretical predictions surrounding the dead-end hypothesis, did not correlate with breeding system in that the homothallic lineage Cailleux IV appeared among the most recombined.

Notably, the ratio of nonsynonymous to synonymous diversity was quite high within the highly inbred lineages of Cailleux IV, but much lower for the diversity between these groups. This discrepancy indicates that selection is not acting strongly at the spatiotemporal scale of these fine lineages, but is still effective at the scale between them.

Despite having a linkage pattern similar to *N. discreta*, *N. tetraspora* had the highest π_N/π_S ratio. Indeed, *N. tetraspora*'s π_N/π_S was nearly as high as that measured for sequencing errors between the 7 clones of P4-W1, which should be entirely neutral (Fig. 2.5D).

Another estimator of selection, Tajima's D, compares the number of variant sites to the average pairwise diversity within a population. Extreme values of Tajima's D can arise due to several factors, and we observed no obvious relationship between Tajima's D and the pattern of LD (Fig. 2.7). Tajima's D did indicate, however, that most *Neurospora* species have an overabundance of rare alleles except *N. discreta*. Gladieux *et al.* (2015) hypothesized that the *N. discreta* populations in North America were expanding from refugia following glacial retreat, and the general overabundance of rare alleles in all populations except CAWA may indicate that similar expansions have been common in *Neurospora*.

We furthermore estimated the rate of rearrangements (as break points), inversions, and the average size of inversions along branches in the *Neurospora* phylogeny. We found no difference in any of the rate of rearrangement breaks, rate of inversions, or the average size of inversions between 21 heterothallic or pseudohomothallic branches and the 7 homothallic branches (Kruskal-Wallis $p > 0.1$, Fig. 2.8).

2.3 Discussion

Lack of Clonality and The Dead-End Hypothesis in Cailleux IV

Given the likely single origin of homothallism in Cailleux IV, we posit two scenarios of homothallic evolution. First, that the species are clonal as previously supposed, and second that Cailleux IV is a single recombining species.

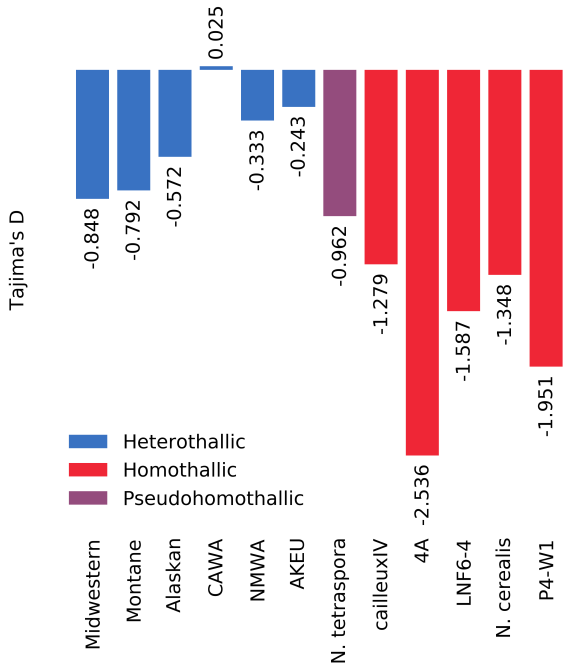

Figure 2.7: Tajima's D measurements for different *Neurospora* populations. Negative values indicate an overabundance of rare alleles relative to neutrality, and positive values indicate overly-balanced allele frequencies.

Under the hypothesis of clonality, multiple clonal lineages must have survived for a long time without an accumulation of deleterious alleles, making Cailleux IV an "asexual scandal." Under the dead-end hypothesis such lineages should be rare and accepting this hypothesis requires either 1) accepting the clade as a lucky find, 2) rejecting the dead-end hypothesis, or 3) supposing some unknown mechanism to purge deleterious alleles and escape Mueller's ratchet. Still, in order to explain the distance decay of linkage between lineages or the preponderance of variants discordant with a single evolutionary tree we must invoke not only an ancient asexual species, but an ancient asexual individual akin to the debunked Glomeromycete heterokaryon hypothesis.

Unlike vertebrate animals, we have no evidence that *Neurospora* segregates a germ line from a somatic line, and a colony may thus be viewed as a population of mitotically cloned nuclei. Ascospores contain only a single haploid nucleus, though, and so meiosis is an extreme bottleneck on this population. If a single colony were to persist for a great length of time without undergoing meiosis it could accumulate many diverged nuclei, then in a single round of selfing could create recombined progeny that resemble a sexual population with linkage that decays with distance and discordant patterns of shared variation. In order to explain the pattern observed in Cailleux IV, however, this single colony would have had to have undergone a feat of evolutionary gymnastics: First, it must have lived longer than the oldest

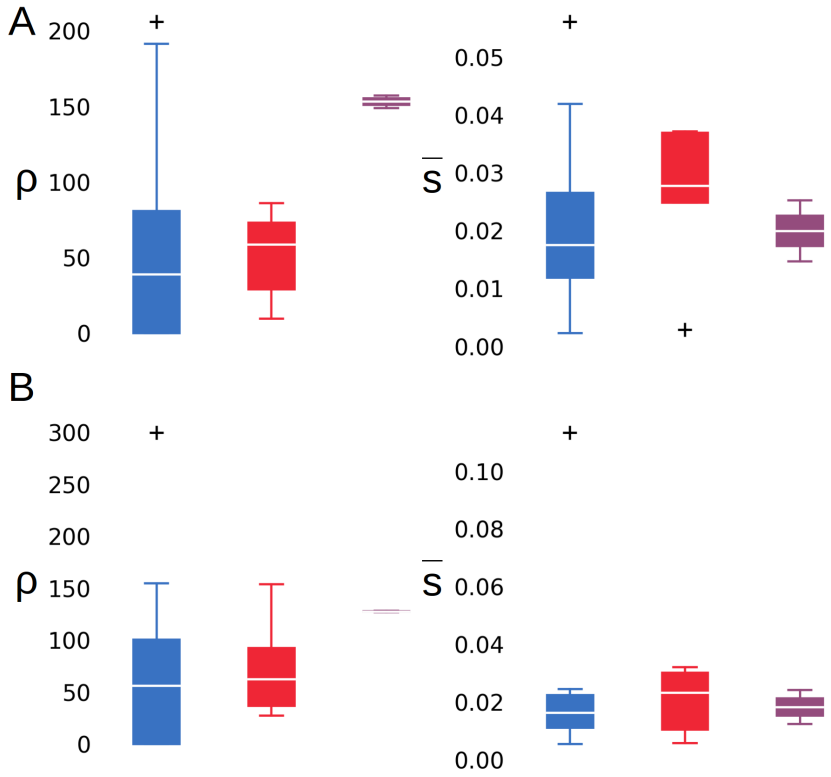

Figure 2.8: Estimated rates (ρ) and mean sizes ($\bar{s}$) of rearrangements along *Neurospora* lineages. No significant difference is seen between heterothallic (blue), homothallic (red), or pseudohomothallic (purple) lineages when analyzed using either all orthologous segments (A) or only orthologous segments internal on the scaffolds (B).

known fungal individual (estimated at 8,650 years) (Ferguson *et al.*, 2003) and potentially for millions of years; second, it must only recently have undergone a round of selfing; and third, its progeny must have independently spread across North America.

Under the hypothesis of recombination we need not challenge existing theory surrounding the dead-end hypothesis, as the patterns of LD decay and shared variation are consistent with a sexual species. The patterns of LD and diversity within *N. cerealis* and P4-W1, on the other hand, are compatible with short-lived, dominantly clonal, degenerative lifestyles. Still, the possibility exists that these groups are simply members of a larger, diverse, recombining population. In their comprehensive multilocus phylogeny of *Neurospora*, Nygren *et al.* (2011) showed that most homothallic taxa exist within diversified clades similar to Cailleux IV. We therefore think the latter scenario is quite plausible, as *N. cerealis*, for example, may share a common origin and be recombining with close neighbors *N. reticulata* and *N. minuta*.

The similar rate of inversions between the heterothallic and homothallic lineages indicates continued selection on genomic rearrangements in contrast to the relaxed selection on protein coding genes within the inbred homothallic lineages. Selection against inversions and other

rearrangements is thought to occur during meiosis, as crossing-over within heterokaryotypic crosses can yield inviable chromosomes (Kirkpatrick, 2010). Haploid selfing, as homothallic species can do, is not subject to this pressure because the homologous chromosomes are identical. That inversions are not accruing any faster within homothallic lineages than heterothallic lineages is consistent with their outcrossing frequently enough in nature to select against these and other rearrangements.

While Fst is extremely high between groups (>0.9, Fig. 2.4) in Cailleux IV, which typically indicates divergence due to a large proportion of fixed differences, a high Fst is also expected when comparing multiple samples of the same genet to multiple samples of a different genet. To illustrate this point, imagine tissue samples from many different organs were sequenced from two humans selected randomly from a population. Both individuals are obviously members the same recombining population, but the relationship between their different tissues is clonal in the same way that different meiotic progeny from haploid selfing are clonal. Since they were randomly selected from the population both their genotypes at the moment of conception represent selections from the gene pool. Yet, when comparing only these two individuals the differences between them appear "fixed." As they develop they incur different sets of mitotic mutations, and those tissues with a more recent developmental origin will share more of these mutations. Fst between these two individuals represents the comparison of variation between their tissues to variation between them. The number of mutations within each person being much smaller than the number of allelic differences between them, Fst will thus be very large.

Based upon their comparative genomic analysis, Gioti *et al.* (2013) found additional evidence supporting the dead-end hypothesis for homothallic lineages, including less efficient codon usage within highly expressed genes in the homothallic species, *N. africana*, and decreased transmission of transposable elements within homothallic genomes, as predicted for a dominantly selfing lifestyle. Gioti *et al.* did not examine any member of Cailleux IV, and the evident sexuality of Cailleux IV does not refute their findings nor disprove the dead-end hypothesis, and is in fact consistent with its predictions. Rather, accepting the hypothesis of clonality in Cailleux IV would demand explanation as to why selection is not relaxed (as indicated by the low π_N/π_S) in this homothallic clade, as well as challenge the dead-end hypothesis' tenant of rare ancient asexuals. Instead, we believe the evidence shows that at least some homothallic species are likely to remain outbred. Homothallism should therefore not be taken as an assumption of clonality, but must be considered as a co-option of meiosis for production of hardy propagules.

Siblings In Dataset

There is considerable variability among the percent of orthologous sequences that support different branches in the trees, with some distal branches having unusually high support values of around 25% or 50% in an otherwise interbred population (Fig. 2.4). Such a pattern is expected when sampling is not random and close relatives are present within the sample. Because only a few cross-over events occur per meiosis, haploid siblings should have identical

tracks over around half their genomes. Given another generation of outcrossing, haploid cousins should share around 25% variation. Given that we sampled soil from relatively small geographic areas within sites (often under a kilometer) and isolated multiple strains per soil sample, the presence of close relatives is plausible.

A notable exception is the 52% concordance between groups 4A and 9A in Cailleux IV, which includes samples collected in both California and Colorado. This long branch, representing approximately half of the divergence between these strains and the root of the clade, and the 52% concordance of orthologues is consistent with them being sibling genets. Similarly, these two lineages share the most alleles of any set in the group (63,267 sites compared to 28,687 in the next most similar pair). While a pair of cross-continent siblings may seem implausible, both LNF6-4 and 9A demonstrate that even a continent-spanning genet is possible.

Relaxed Selection in *N. tetraspora* and an Overabundance of Rare Alleles in *Neurospora*

N. tetraspora's overabundance of rare alleles compared to *N. discreta* (Fig. 2.7) might indicate population expansion, a hypothesis which is corroborated by the species' wide geographic range and lack of population structure. An abundance of rare alleles can also emerge due to purifying selection, but this explanation is contradicted by the extremely high π_N/π_S ratio in *N. tetraspora*. In contrast, purifying selection might explain the overabundance of rare alleles in the outbred species. Tajima's D is also known to be negatively biased with increasing sample size (Subramanian, 2016), so the even more negative values within the inbred homothallic lineages likely still underestimate the abundance of rare alleles.

Conclusions

With the addition of the populations we examined here, populations of *Neurospora* genomes from the full range of breeding systems in Ascomycota, with and without clonal conidia, have now been studied. With multiple independent lineages examined for some breeding systems, *Neurospora* is now poised as a model for all forms of sexual evolution.

We observe in the homothallic lineages studied here a link between clonal population structure and a lack of selection that is consistent with the body of theory surrounding the dead-end hypothesis. We find that the homothallic group, Cailleux IV, is long-lived and experiencing efficient selection, owing to this homothallic lineage outcrossing as much as or in some cases more than heterothallic *Neurospora* species in addition to producing clonal offspring.

Due to its ability and frequency of selfing, Cailleux IV has a population structure that is "inverted" from that of obligately sexual organisms. Under the paradigm of sexual reproduction, the branches of the evolutionary tree bifurcate as species (and eventually populations) diverge from each other, terminating in star-shaped phylogenies for panmictic populations.

Within Cailleux IV, however, different lineages produced through recombination join in a star-like phylogeny, which then branch out clonally at the tips of the major lineages. Without the context of the other lineages each lineage would easily be misinterpreted as a clonal species.

Biologists, and mycologists especially, should be cautious about interpreting demographic data before the full ecology or distribution of an organism is known. Here, we take an in-depth look at a demographic pattern that has previously deceived research on pathogenic fungi such as *Crytpococcus gattii* and *Aspergillus fumigatus* (Taylor *et al.*, 2015). In the context of epidemic pathogens uneven sampling is particularly likely, as sampling will be biased towards clinical or agricultural isolates that have been spread by human activity. For these reasons, any claim of clonality, particularly for organisms of medical or economic importance, should be qualified by the domain of sampling.

2.4 Materials and Methods

Annotation

We annotated genomes using Augustus (Stanke, 2003; Stanke *et al.*, 2006), with *N. crassa* as the reference species and intron hints obtained by mapping proteins predicted in *N. crassa* onto each focal genome using exonerate (Slater and Birney, 2005).

Phylogenetic Analysis

We built phylogenetic trees using the whole genome phylogenetic analysis pipeline described in Chapter 3, which makes use of MUMmer (Kurtz, 1999; Delcher *et al.*, 1999, 2002; Kurtz *et al.*, 2004), MAFFT (Katoh *et al.*, 2002; Katoh and Standley, 2013), a rewrite of TIGER (Cummins and McInerney (2011), Chapter 3) that employs the partitioning method of Rota *et al.* (2017), and IQTree (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017; Thi Hoang *et al.*, 2018; Chernomor *et al.*, 2016). Additionally, we built trees for each orthologous segment larger than 100 sites separately in the same manner, but without rate partitioning, and applied them as support values on the concatenated tree using IQTree's `--sup` option. We built the species tree using only representatives from each cluster and the trees of closely related clusters in Fig. 2.4 using the same pipeline.

Our pipeline for whole-genome phylogenetic analysis and our rewrite of TIGER can be found at `https://github.com/channsoden/hannsoden-bioinformatics`

Variant Calling

For each group, we aligned reads to the reference genome using Bowtie 2 (Langmead *et al.*, 2009) with the `--sensitive` preset and `--no-discordant` option, then removed duplicate reads with Picard Tools MarkDuplicates (The Broad Institute, 2009). When multiple

sequencing runs or experiments were available for a single strain we merged the aligned reads using Picard Tools. We called variants for each strain individually using GATK HaplotypeCaller, then combined the haplotype calls with GATK CombineGVCFs and called final genotypes using GATK GenotypeGVCFs (McKenna *et al.*, 2010), setting a maximum of 4 alternate allele types.

Determination of Clones

As a control for the error in variant calling, we aligned the reference reads to the reference assembly and called SNPs along with the related strains. We considered two strains to be clones of each other when they were not significantly more different from each other than the reference was to itself.

For this determination, we assumed that sequencing and calling errors are normally distributed, and empirically measured the mean and standard deviation of calling errors from all references aligned to themselves. We then initialized an EM algorithm on the set of all pairwise differences to categorize them into either clones or distinct, utilizing a 1% significance threshold for inclusion into the clonal set of pairs and stopping iteration when the mean number of errors was within 1 of the previous step.

This method failed to identify three strains of P4-W1 as clones of the reference due to their higher sequencing error. However, these strains still varied by less then any other non-clone in our data set, and had only 83 total shared variants out of around 10,000 called sites. This pattern would be consistent with a very recent clonal ancestor, except that a pattern of LD was detected within 200 bp. Given the almost complete lack of shared variation, this indicates that the uncommon alleles are spatially clustered within around 200 bp. As our sequencing reads were 150 bp long, this pattern can only be explained by the presence of low-quality reads within the dataset, in accordance with the lower read quality of these strains.

Population Inference and Statistics

We inferred population ancestry using ADMIXTURE (Alexander *et al.*, 2009), selecting the minimum value of K within 5% of the lowest cross-validation (CV) error. We report the Fst values calculated by ADMIXTURE.

We calculated the number of pairwise differences between members of a population (π) and between members of two populations (Dxy), adjusted by call quality, from GATK's VCF output. To calculate π_N/π_S, we examined each codon as predicted by the reference's annotation. For each biallelic variant within a codon, we calculated the probability of all possible different contexts from the allele frequencies of any alleles at the other two sites. We then summed the probabilities that the two alleles are synonymous and multiplied this by the site quality to achieve the joint probability that the call is correct and that the allele is synonymous. We similarly summed the probabilities that the two alleles are nonsynonymous

to determine the converse joint probability. We used these adjusted quality scores to calculate π_S and π_N, respectively, ignoring all sites with more than 2 alleles.

From the value of π and the number of called variant sites we calculated Tajima's D in the usual manner (Tajima, 1989).

Measurement of Linkage Disequilibrium

Within each group we calculated LD between biallelic sites for which at least two strains shared the minor allele. In order to account for variability in confidence of genotype calls, we calculated a composite quality score which is not probabilistically justified, but can be easily calculated and applied *post hoc* to LD calculations which are not quality aware. For each site, we calculated the composite variant quality as the product of the mean quality score for all samples times the site quality score.

We calculated the coefficient of correlation of LD (r^2) between pairs of sites using VCFtools `hap-r2` for haploid populations and `geno-r2` for *N. tetraspora* (Danecek *et al.*, 2011). To reduce the number of comparisons to be made we calculated LD only between each site and 100 of its nth neighbors, where $n \in N = \{1, ..., n_i, ..., 30000\}$ and

$$n_i = \begin{cases} i, & i \leq 15 \\ \|10^{(1.2+(i-16)(\log 30000 - 1.2) \div 84))}\|, & i > 15 \end{cases}$$

This sampling scheme yields measurements for pairs over a large semi-continuous range, since the distance between nth pairs varies. We binned pairs by distance in a step-wise log manner (i.e. 1, 2, 3, ..., 10, 20, ..., 10^6). We calculated the mean LD and mean distance within each bin, weighted by the product of the composite quality scores of the two sites.

We assume that LD decays logarithmically toward an equilibrium value determined in part by the sample size and distribution of allele frequencies. We estimated the equilibrium LD from the data by assuming that the most distant 10% of pairs will be unlinked and thus at equilibrium. We calculated LD_{50} as the mean distance between the first bin with a mean r^2 under 50% of the maximum and the last bin with a mean r^2 over 50% of the maximum.

Estimation of Inversion Rates & Sizes

We used a distance-based framework to estimate the rate and average size of inversions along the branches of the *Neurospora* phylogeny. First, we aligned representative genomes of each species to each other in a pairwise manner using Murasaki (Popendorf *et al.*, 2010) and OSFinder (Hachiya *et al.*, 2009) in the manner described in Chapter 3. From these orthologous segments, we use GRIMM (Tesler, 2002b,a) to calculate the multichromosomal rearrangement distance between each pair of genomes (b). We additionally counted the number of inversions (i) between each pair as the minimum number of inverted segments when scaffolds can be rotated arbitrarily. We measured the total length of inverted sequence (v) and the alignment length (a) as the sum of both matches and mismatches in the respective

segments. As the adjacencies of the ends of scaffolds in draft genomes are unknown, we measured these values using all segments and using only internal segments with known adjacencies.

Since inverted sequences are highly susceptible to reversion, we correct the distance, v, in the manner of Tajima and Nei (1984). The instantaneous rate of accumulation of inverted sequence is balanced by the negative feedback of reversion, following the differential equation:

$$\frac{dv}{dt} = \rho\bar{s}(a - v) - \rho\bar{s}v \tag{2.1}$$

Where ρ is the rate of inversion events per site and $\bar{s}$ is the mean size of those inversions. The solution to this equation yields the familiar distance correction for an equilibrium frequency of 0.5:

$$\rho\bar{s}t = -\frac{1}{2}\ln\left(1 - \frac{2v}{a}\right) \tag{2.2}$$

We use the Phylip program FITCH (Felsenstein, 2009) to estimate branch lengths on the *Neurospora* phylogeny from the distance matrices of rearrangements, inversions, and length of inverted sequence. We relate these different evolutionary distances to the substitution distance, μt, to calculate rearrangements per substitution, $b/\mu ta$, and inversions per substitution, $\rho/\mu = i/\mu ta$. Finally, we calculate the average size of inversions as the quotient of the two two inversion distances, $\bar{s} = \frac{i}{a}/\rho t$.

Chapter 3

Estimation of Rearrangement Break Rates Across the Genome

Christopher Hann-Soden, Ian Holmes, and John W. Taylor

Abstract

Genomic rearrangements provide an important source of novel functions by recombining genes and motifs throughout and between genomes. However, understanding how rearrangement functions to shape genomes is hard because reconstructing rearrangements is a combinatoric problem which often has many solutions. In lieu of reconstructing the history of rearrangements, we answer the question of where rearrangements are occurring in the genome by remaining agnostic to the types of rearrangement and solving the simpler problem of estimating the rate at which double-strand breaks occur at every site in a genome. We phrase this problem in graph theoretic terms and find that it is a special case of the minimum cover problem for an interval graph. We employ and modify existing algorithms for efficiently solving this problem. We implement this method as a Python program, named BRAG, and use it to estimate the break rates in the genome of the model Ascomycete mold, *Neurospora crassa*. We find evidence that rearrangements are more common in the subtelomeric regions of the chromosomes, which facilitates the evolution of novel genes.

3.1 Introduction

As the molecular revolution allowed scientists to understand the evolutionary import of genetic domains by comparing genes that varied in sequence, the genomic revolution allows us to dissect the importance of genome architecture by comparing genomes that vary in structure.

The Importance of Genome Structure

The chromosomes of both eukaryotes and prokaryotes are known to be highly structured, and that structure is important for gene regulation, the coevolution of genes, DNA replication, and cell division. Moreover, the position of genes within the genome reflects their evolutionary importance. In prokaryotes, genes are biased toward being coded in the leading strand, with conserved essential genes showing higher strand bias (Rocha and Danchin, 2003a,b). Similarly, an emerging trend in eukaryotic genomics, including our model organism of choice, *Neurospora crasa*, has been the concentration of essential genes to the center of chromosomes and the use of subtelomeric regions as sources of genetic innovation (Batada and Hurst, 2007; Kasuga and Glass, 2008; Brown *et al.*, 2010). A similar trend in the genomics of fungal pathogens is the two-speed genome model, whereby rapidly evolving genes associated with pathogenesis and adaptation to new hosts are found in regions rich with transposons and repetitive elements (Dong *et al.*, 2015).

Given the evident importance of genome structure, it is unsurprising that genomes seem to have evolved mechanisms that stabilize their structure and regulate architectural evolution. In organisms from humans to plasmodium, genes have been shown to be shuffled between the subtelomeric regions, and subtelomeric regions contain unusually high levels of gene duplications and novel genes (Linardopoulou *et al.*, 2005; Cerón-Romero *et al.*, 2018; Kasuga and Glass, 2008). The presence of transposons in the "high-speed" regions of fungal pathogens have been shown to increase the rate of rearrangements both passively by supplying regions of homology and directly by providing a mechanism for rearrangement (Faino *et al.*, 2016). Rearrangements between subtelomeres are also thought to be aided by blocks of repetitive domains, and in humans there is little evidence for direct action of transposons (Linardopoulou *et al.*, 2005).

Within these regions of rapid evolution rearrangement serves as an engine for genetic novelty. In other circumstances, though, rearrangement can act to preserve the genome. Tight linkage of cooperative alleles can ensure co-segregation and co-evolution. Inversions, and perhaps other rearrangements, surrounding cooperative clusters can reduce the risk of separation through recombination as well as prevent repair processes, leading to the development of "supergenes" (Thompson and Jiggins, 2014). Rearrangements are hypothesized to lead to the evolution of specialized functions (Tigano and Friesen, 2016; Brown *et al.*, 2004; Hoffmann and Rieseberg, 2008) and sex chromosomes (Wang *et al.*, 2012; Lahn and Page, 1999; Bachtrog, 2013; Fraser and Heitman, 2004) because of their role in linking advantageous haplotypes. Similarly, recombination is suppressed on the mating-type chromosomes of the fungi *Neurospora tetrasperma* and *Micobotryum lychnidis-dioicae*, and this suppression is linked to an accumulation of rearrangements on these chromosomes (Menkis *et al.*, 2008; Ellison *et al.*, 2011a; Badouin *et al.*, 2015). In *N. tetrasperma*, both structural and non-structural suppression have been demonstrated (Jacobson, 2005), although the non-structural mechanisms predate the inversions on this chromosome (Sun *et al.*, 2017).

Under the opposing pressures of maintaining order and generating novelty, the genome structure itself has evolved to promote potentially adaptive rearrangements and mitigate

the harm of maladapative rearrangements. Similarly, genome defense mechanisms against selfish genetic elements have arisen. For example, in ascomycete fungi Repeat Induced Point Mutation (RIP) acts to silence transposons (Galagan and Selker, 2004). The existence of RIP and structured regions of rapid evolution add to a growing body of evidence indicating that genomes have evolved mechanisms to pick themselves up by their own bootstraps.

Despite the essential functions and evolutionary importance of genome architecture, our understanding of genome architecture has been hampered by the difficulty of assaying the structure of a genome and of inferring the evolutionary history of related genomes. Chromosome mapping techniques such as G-banding, linkage mapping, restriction mapping, happy mapping, and FISH are skill and labor intensive, and only reveal genome organization on the macro- and meso- scales. Genome sequencing can reveal the complete ordering of a genome, but until recently the cost and low quality of sequencing has precluded large scale comparative studies. Now, thanks to the ever decreasing cost and increasing quality of genome sequencing, comprehensive studies of rearrangements of all sizes are possible.

Phylogenetic Inference of Rearrangements

A dizzying array of new high-throughput assays are enabling the study of genome architecture, but analysis of these piles of data has become the primary challenge. Measuring the pace of rearrangement within different chromosomal domains would ideally involve reconstructing the history of those domains and counting the occurrences of different rearrangements. However, the difficulty of inferring rearrangements makes this task impractical for the increasingly large datasets now available.

Phylogenetic inference is dependent upon the inference of homology between the sequences under study, that is, the alignment. Sequence alignments represent sequences as rows in a matrix where columns represent a homology relation between characters in the sequence. Unfortunately, while pairwise alignment is relatively efficient, the problem of multiple sequence alignment (MSA) is known to be NP-hard (Elias, 2006) and remains a computational challenge. Even matrix sequence alignments, though, are insufficient for aligning genomes because they cannot account for rearrangements between the sequences.

One approach to multiple genome alignment (MGA) is to find and align collinear regions between the genomes, then represent each genome as an ordering on an improper subset of the collinear regions. Inferring rearrangements from the ordering then requires combinatorial analysis. Graph based MGAs improve this paradigm by representing collinear segments as nodes and each genome as a path between the nodes. Edges in such a graph thus represent adjacencies between sequences, or non-homologous phosphodiester bonds. MGA graphs have the elegant property that rearrangements are equivalent to swapping edges in the graph, and reconstructing possible histories of rearrangements is thus equivalent to converting this multigraph into a graph where all genomes follow the same path (Alekseyev and Pevzner, 2009). Reconstructing genome histories in this way is known as the Multiple Genome Rearrangement Problem (MGRP), which is known to be NP-complete in at least some cases (Caprara, 1999), and the complexity of MGRP balloons as more genomes are added since

each genome adds both more observed rearrangements that fragment the collinear regions and another ordering of those regions.

Furthermore, this sort of MGA suffers from the philosophical problem of splitting MGA into two steps with different rules: first the NP-hard MSA that assumes no rearrangement, and second the NP-complete analysis of rearrangements that assumes the truth of the MSAs. This paradigm conflicts with the biological truth of evolution, which operates under a single set of rules. While we classify mutations (e.g. as single nucleotide polymorphisms, deletions, or transpositions) based upon our observation of the products of evolution (i.e. sequences), these classifications have an ambiguous relationship to the mechanisms that caused them. Cactus graphs, a recursive alignment graph where each node is itself an alignment graph, allow for iterative refinement of the inferences of homology in the context of the larger genome structure (Paten *et al.*, 2011). Cactus graphs thus provide a unified model of homology that accounts for both changes in the nitrogenous bases of the DNA and the phosphodiester bonds that link them.

Estimating Rearrangement Rates

Even given a perfect alignment methodology, there are often multiple equally good solutions to the MGR problem, leading to significant ambiguity in the sequence of events (Sankoff and Blanchette, 1998).

Distance based phylogenetic methods can provide an imperfect solution by counting the number of double-strand breaks between two genomes. Breaks occur when the phosphodiester backbone of both strands of a chromosome are broken and not reformed. When chromosome number is maintained, breaks only become permanent when two double-strands are broken and the ends of the DNA fragments are swapped. Thus, the break rate is less than or equal to twice the rearrangement rate (Sankoff and Blanchette, 1998). Every missing adjacency in a comparison of genome orderings or every branch point in an MGA graph represents a **breakpoint**, and breaks are thus easily quantified.

Distance based methods, though, become less reliable at larger phylogenetic distances. While the risk of reversion of rearrangements is lower than with substitution mutations, alignments between distantly related genomes are more difficult to achieve and have more gaps, which can lead to an underestimate of break distances. Furthermore, information about the local density of breaks is lost when reduced to a simple count.

Rather than attempt to infer and quantify the number of rearrangement events along a phylogeny, we re-frame the problem as measuring the degree of conservation, or conversely fragility, of the set of phosphodiester bonds within a genome. Rearrangements by their nature break one set of bonds and form a new set of bonds, and so rearrangements represent a decay process on the set of phosphodiester bonds. The fragility of bonds in this context is underlain by the chemical instability of the bonds, but perhaps more importantly reflects selection that acts to carry those changes into the next generation or eliminate them from the population.

Within this framework, we present a method to estimate the Break Rates Across a Genome (BRAG). BRAG uses pairwise alignments between the genome of interest, termed the reference, and a set of related genomes, thus obviating the need for a difficult multiple genome alignment, or even all-against-all pairwise alignment. BRAG employs a novel, interval graph based approach for which efficient algorithms have been previously described. Additionally, BRAG provides a detailed survey of break rates across the genome by computing the likelihood landscape of the break rate at every site in the genome.

3.2 New Approaches

The breakpoints in pairwise alignments evidence at least one double-strand break within a region of the reference genome at some point in evolutionary time between the reference and a query. Different queries, having different evolutionary relationships to the reference genome, paint different patterns of breaks on to the reference (Fig. 3.6). These breakpoints are a relationship only between the reference and particular query, so we abbreviate them qbreaks to distinguish them from actual evolutionary events.

Using draft assemblies the state of the bonds on the ends of scaffolds is unknown. We term the qbreaks formed by the ends of scaffolds "false" qbreaks since they may represent break events or simply incomplete assembly, while we term qbreaks for which we have positive evidence for "true."

Under a paradigm of parsimony, we seek to explain the observed qbreaks with the fewest evolutionary events. Multiple qbreaks can be explained by the same evolutionary event when they overlap each other on the reference. Finding the fewest events that explains all qbreaks is a case of the **minimum clique cover problem**. The pattern of qbreaks can be represented as an interval graph by making each qbreak a node then connecting qbreaks that overlap on the reference by an edge (Fig. 3.7A). A **clique** is a set of fully connected nodes, and thus represents a set of qbreaks that could be explained by a single event, and a clique cover is a set of cliques that explains all qbreaks.

The evolutionary relationship of the genomes (i.e. the tree) imposes and additional restriction on this problem, however. Events must be placed on branches in the tree, and thus not every clique represents a valid evolutionary event (Fig. 3.7B). We term a clique that is consistent with the evolutionary tree a **tbreak**.

We adapt existing algorithms for clique covers of interval graphs to enumerate all solutions to the minimum clique cover by tbreaks (MCCT) problem. Notably, the magnitude of this problem can be substantially reduced by splitting the graph into disconnected subgraphs, each of which can be solved independently. Assuming that breaks represent a decay process on the bonds in the reference, we then calculate the joint likelihood for a given break rate over all maximum parsimony solutions for all sites in the genome.

We implemented this approach in Python. BRAG depends upon the Python packages NumPy (Oliphant, 2006), Pandas (McKinney, 2010), StatsModels (Seabold and Perktold,

2010), SciPy (Jones *et al.*, 01), BioPython (Cock *et al.*, 2009), ETE (Huerta-Cepas *et al.*, 2016), and Matplotlib (Hunter, 2007).

BRAG, and all code used in the analysis presented in this paper, are available at `https://github.com/channsoden/BRAG`. All code is licensed under the 2-clause BSD license. Genomes and alignments used in this study are available at `https://osf.io/ak54t/`.

Our pipeline for the whole-genome phylogenetic analysis and our rewrite of TIGER can be found at `https://github.com/channsoden/hannsoden-bioinformatics`.

Glossary

affiliations - x^* - A qbreak is affiliated with a tbreak if the tbreak contains the qbreak. The set of all tbreaks that contain a qbreak are the qbreak's affiliations.

breakpoint - A phosphodiester bond, or set of contiguous bonds, that is present in the reference genome but not the query.

clique - A set of vertices in an undirected graph such that each is adjacent to every other.

clique problem - The problem of enumerating all maximal cliques in an undirected graph.

cover - A set of cliques of a graph whose union includes every vertex in the graph

history - h - A solution to the tbreak cover problem, which represents a maximum parsimony evolutionary history of the reference genome.

interval graph - A graph of a multiset of interval periods, such that each vertex represents an interval over space or time, and intervals that overlap one another are represented by an edge drawn between their corresponding vertices.

maximal - A set of elements of a set or graph so that no other element may be added while maintaining a property. With regards to cliques, a maximal clique is a clique that cannot be made larger by adding another vertex and still remain a clique.

maximum - A set of elements of a set or graph that is the largest of its kind. With regards to cliques, a maximum clique is the largest clique in a graph.

minimal - A set of elements of a set or graph that no other element may be removed from while maintaining a property. With regards to covers, a minimal cover is a set of cliques that no clique can be removed from and still cover every element of the graph.

$\subset$-minimal - A qbreak (or vertex of another interval graph) whose affiliations contains no other qbreak's affiliations as a proper subset.

minimum - A set of elements of a set or graph that is the smallest of its kind. With regards to covers, a minimum cover is the smallest cover of a graph.

minimum clique cover problem - The problem of finding the smallest cover by cliques of a graph.

multiset - $\{a, b, b\}$ - An unordered collection of elements which may be, but are not necessarily, unique.

powerset - $\mathcal{P}(X)$ - The set of all subsets of a set.

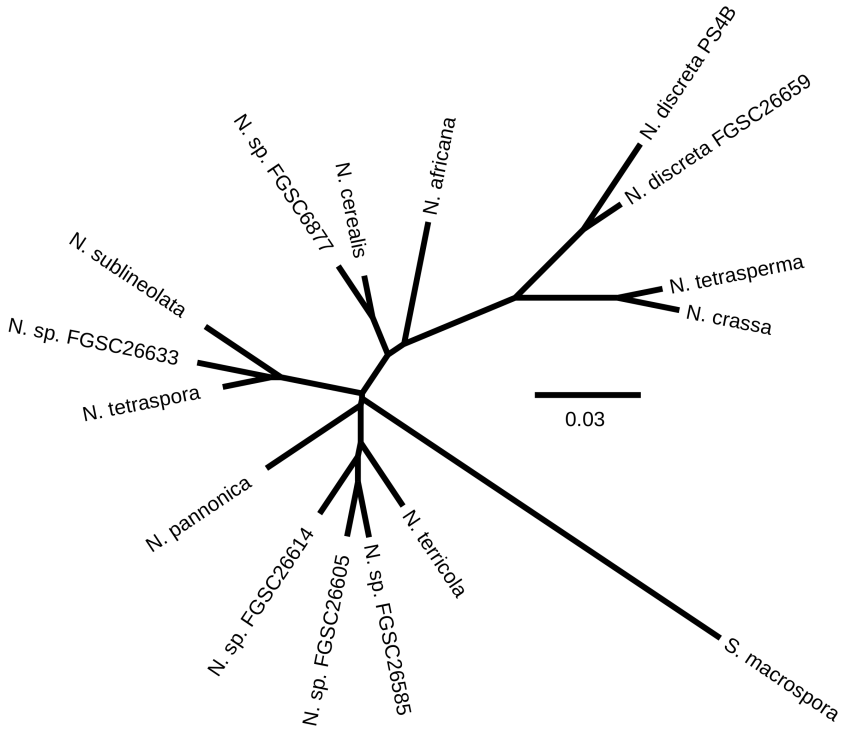

Figure 3.1: The whole genome phylogeny of all *Neurospora* species and an outgroup (*Sordaria macrospora*) for which at least one genome is available. All branches have 100% bootstrap and SH-aLRT support.

qbreak - A region of the reference genome between two orthologous segments that are adjacent in the reference but not in the query.

set - $\{a, b, c\}$ - An unordered collection of unique elements.

tbreak - τ - A clique of qbreaks that is also tree consistent.

tree consistent - A tbreak is tree consistent if it contains all and only qbreaks from queries that lie on one side of one branch in the phylogenetic tree.

3.3 Results

We employed BRAG to examine the pattern of rearrangements in the genome of the model filamentous ascomycete mold, *Neurospora crassa*. *Neurospora*'s relatively small haploid genomes with low repetitive content, as well as ease of sampling diverse species, make it an attractive model for genomics and the study of evolution (Gladieux *et al.*, prep; Palma-Guerrero *et al.*, 2013; Heller *et al.*, 2016; Zhao *et al.*, 2015; Stajich *et al.*, 2009). Here, we examine the finished genome of *N. crassa* by comparing it to draft genomes of 14 *Neurospora* species and an outgroup, *Sordaria macrospora* (Fig. 3.1). The extensive knowledge of *N.*

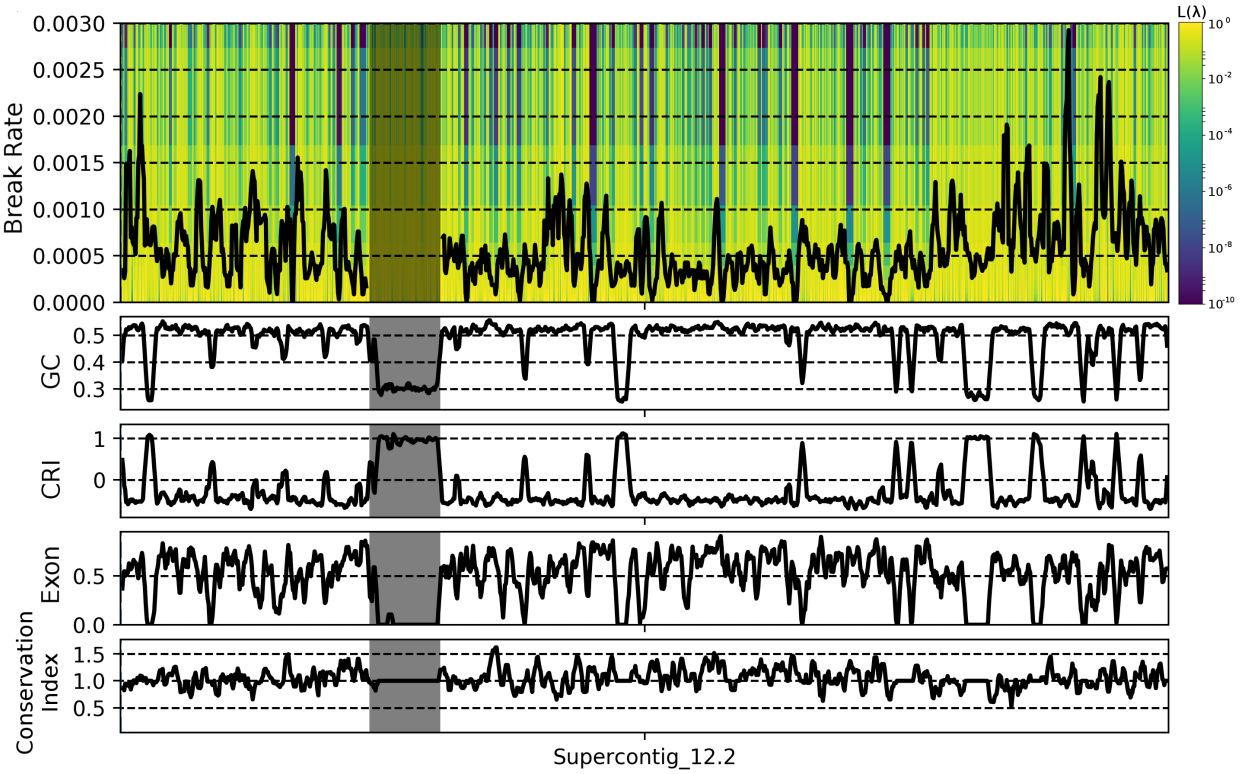

Figure 3.2: Likelihood landscape of break rates along chromosome 2 of *N. crassa*. Yellow indicates high likelihood, while purple indicates low likelihood. The black line shows the maximum likelihood estimate of the break rate for sliding windows with a step of 4,000bp and width of 20,000bp. The break rate was not estimated within the centromere, which is shaded on all tracks. Sliding window calculations of the G/C content, Composite RIP Index (a measure of repetitive content), portion of exonic sequence, and a conservation index are shown below the break rate. High values indicate greater portion of conserved sequence, while low values indicate greater portion of recently evolved sequence.

crassa biology also allows us to correlate the break rate to other features of the genome with the aim of dissecting which factors affect rearrangement dynamics.

BRAG took only 6.5 min to run on a single thread of an Intel Xeon E5-2670 v2 CPU, and used 3.5 GB peak memory. However, solving the minimum clique cover by tbreaks (MCCT) problem took only 3 min 33 s and 494 MB peak memory, with the remaining resources used to draw figures.

We identified 15,450 true (19,544 true or false) qbreaks and 12,356 true (16,176 true or false) tbreaks. The tbreaks formed 10,801 true (14,409 true or false) disconnected subgraphs. Only 23 true (25 true or false) subgraphs had multiple maximally parsimonious solutions, resulting in 60 true (77 true or false) solutions to the MCCT problem. The resulting estimates of the break rate for chromosome 2 is shown in Fig. 3.2.

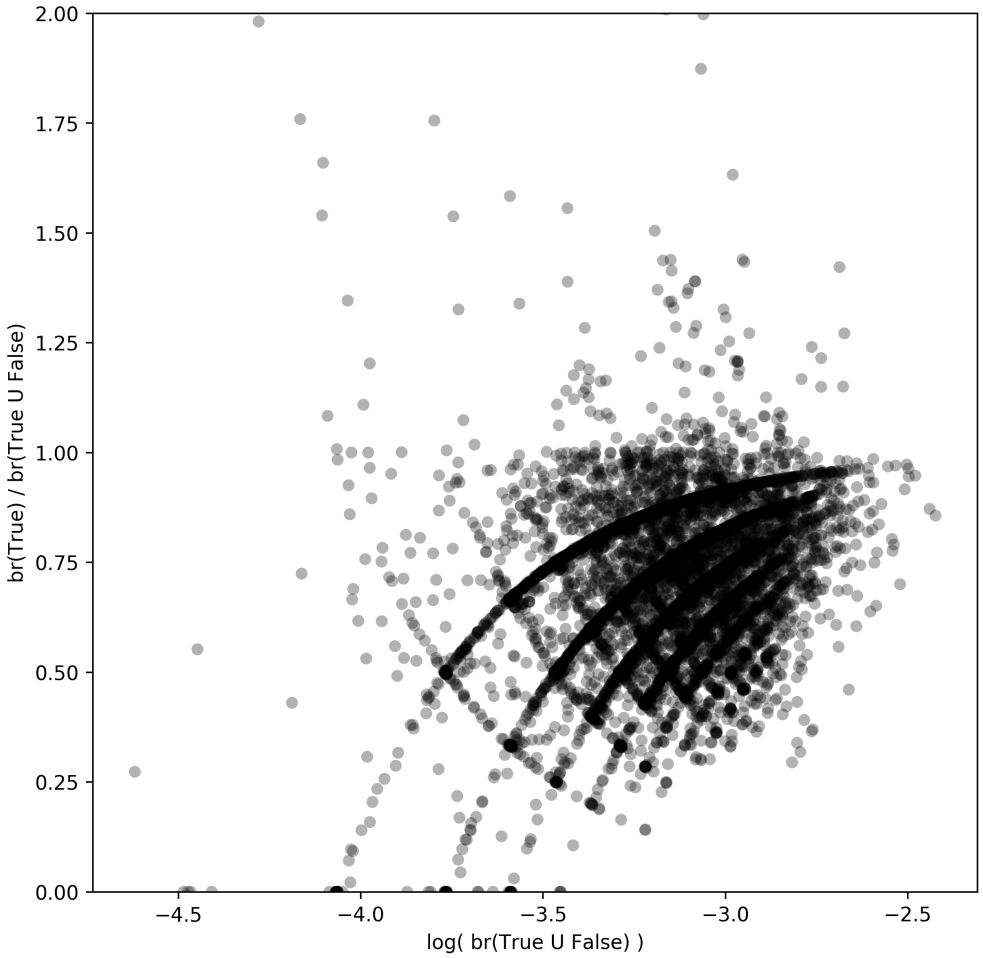

Figure 3.3: Using false qbreaks leads to greater overestimation of the break rate in regions with a low break rate. Each point represents the maximum likelihood estimate of the break rate across a 20,000bp window. Values greater than 1 represent windows where including false qbreaks led to a lower estimation of the break rate, while values lower than 1 represent windows where including false qbreaks led to a greater estimate of the break rate. The 12.4% of windows where the estimate did not depend upon false breaks are not shown.

The maximum likelihood estimate of the break rate when only true qbreaks were used was identical to the estimate when true and false qbreaks were used for 12.4% of sliding windows. 89.8% of non-identical windows showed an elevated rate when using both true and false qbreaks, with a mean increase of 42.9%. The over-estimation of the break rate when using false qbreaks was stronger in regions with a lower break rate ($p < 10^{-126}$, Fig. 3.3), indicating the higher ratio of poorly assembled regions to real qbreaks in these regions.

Factors Affecting the Break Rate

We predicted that breaks would be more common in regions with high repetitive content due to the presence of transposable elements, and less common in gene-dense regions. In *Neurospora*, RIP silences transposons by selectively mutating CpA dinucleotides to TpA dinucleotides in non-unique sequences (Cambareri *et al.*, 1989). As such, regions that have been repetitive in *Neurospora*'s past bear a genetic signature, summarized by the Composite RIP Index (CRI) (Lewis *et al.*, 2009). Positive CRI values indicate RIP action, and thus repetitive sequence, while zero or negative CRI values indicate unique sequence.

We performed multiple linear regression analysis to identify factors correlated with the break rate. While a model incorporating G/C content, CRI, the portion of protein coding sequence, the portion of exonic sequence, and an index of the phylogenetic conservation of genes within a region (described below) found all factors to be significant ($p < 0.001$), the combined effect only explained 5.4% of the variation. Due to the action of RIP, all factors are significantly collinear because RIP enriches the genome for GC content while nullifying genes. We therefore performed independent linear regression between the break rate and CRI, exon density for nonrepetitive regions with CRI < 0, and conservation index. While significant, the effects of both exon density and CRI were vanishingly slight ($R^2 = 0.009$ and 0.0005, respectively). Only the level of phylogenetic conservation of the genes within a region had any appreciable effect on the break rate ($R^2 = 0.029$ in the independent linear model).

We looked for evidence of rapidly evolving subtelomeric regions by performing linear regression between the distance to the telomere and the break rate for each of the 14 chromosomal arms. After Benjamini-Hochberg correction for a false-discovery rate of 0.05, we found a significant negative relationship between distance from the telomere and the break rate for all but one of the 14 chromosome arms (Fig. 3.4A). A 0.5-1 Mbp region of elevated breakage is apparent on the most distal tips of many of the chromosome arms (e.g. the left arms of chromosomes 1, 6, and 7, and the right arms of all chromosomes).

Overall, we find evidence that GC content, repetitive content, and gene density have weak effects, while the level of phylogenetic conservation of genes in a region and the position along the chromosome arm have larger effects. However, collinearity between factors obstructs inference of independent effects and may be leading to spurious observations in the minor factors. Repetitive content was also negatively correlated with distance to the telomeres in 9/14 chromosome arms, which may explain the slight correlation between CRI, GC content, and the break rate. Similarly, our conservation index was positively correlated with distance

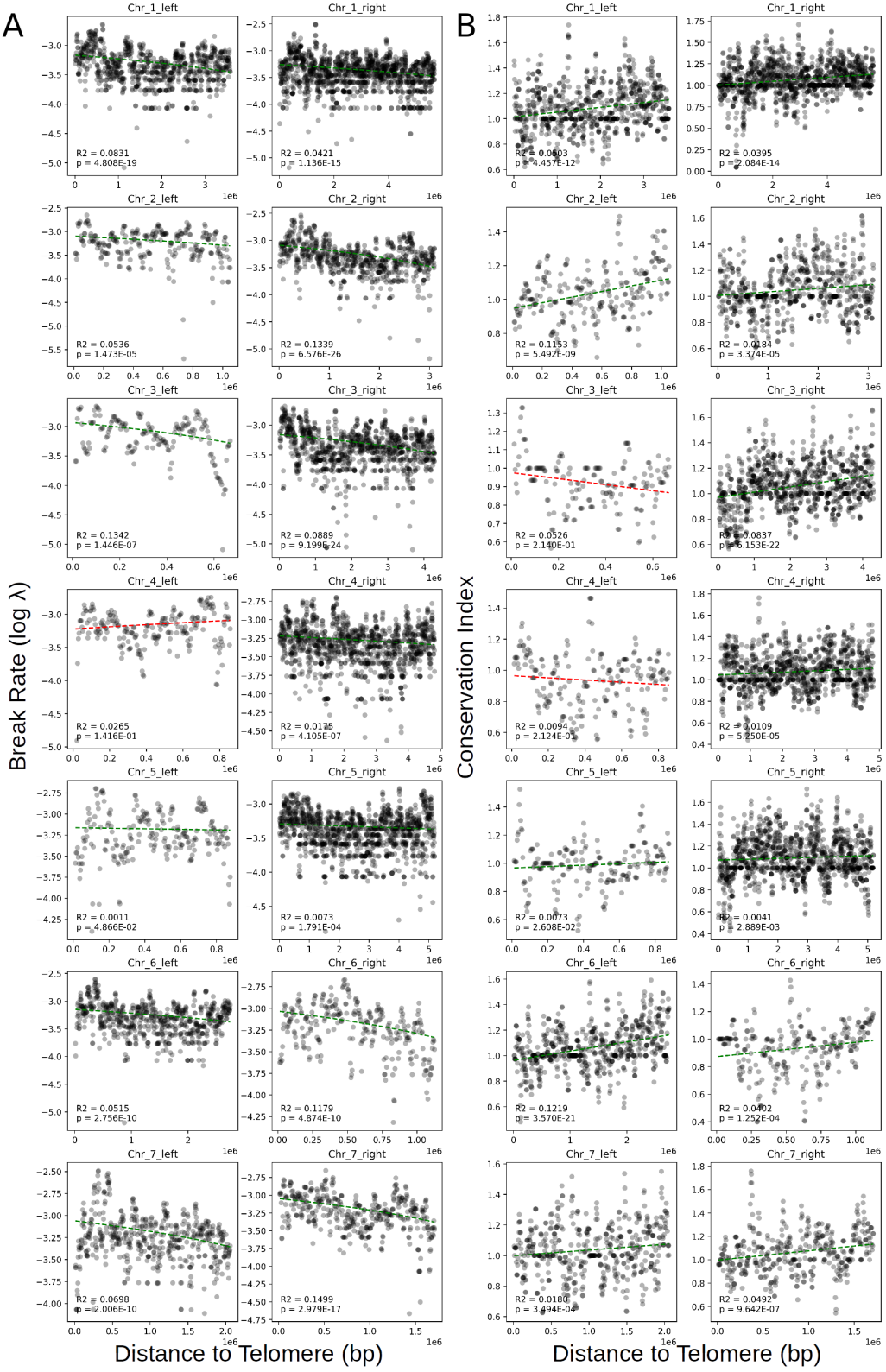

Figure 3.4: Caption on next page.

Figure 3.4: A) Break rate estimates (λ) are higher in the distal regions of chromosomes and lower toward the centromeres. B) The distal regions of chromosomes are enriched for species-specific genes, while the proximal regions are enriched for core conserved genes. Linear models are shown as green dashed lines when significantly negative (A) or significantly positive (B), and red dashed lines when not. Distances are in millions of base pairs.

to the telomeres in 12/14 chromosome arms, indicating linkage between novel genes and rearrangements in a spatially organized manner (Fig. 3.4B).

Fragile Regions Are a Reservoir of Novel Genes

If rapidly rearranging (i.e. fragile) regions of the genome serve as engines for genetic novelty, then we hypothesized that fragile regions will harbor more species-specific genes that could be involved in recent adaptation while stable regions of the genome will harbor more highly conserved genes. To test this hypothesis, we categorized each gene in the genome by the phylogenetic level at which the gene is lineage specific, as reported by Kasuga *et al.* (2009). Kasuga *et al.* reported 5 levels of lineage specificity: species (*N. crassa*), subphylum (Pezizomycotina), phylum (Ascomycota), subkingdom (Dikarya), and all of cellular life.

We calculated a conservation index for regions in the genome as 1 minus the exon density of genes unique to *N. crassa* plus the exon density of genes common to all life. The resulting scale ranges from 0, indicating a region composed entirely of species-specific exons, to 2, indicating a region composed entirely of highly conserved exons. Values near 1 indicate either little exonic sequence or an even mix of species-specific and highly conserved exons.

We then ranked all the genes in *N. crassa* by the break rate of the window they occur in and compared the distribution of ranks for genes from each level of phylogenetic specificity (Fig. 3.5A). The mean break rate ranks, from species-specific to common in all cellular life, were 2,162, 2,502, 2,764, 2,386, and 2,614. We rejected the null hypothesis that the distribution of break rates was identical for each level of phylogenetic specificity with a Kruskal-Wallis test ($p < 10^{-24}$), and species-specific genes had the fastest mean rank break rate.

This pattern is especially apparent between the most fragile and most conserved regions of the genome. The most fragile 2 Mbp (about 5%) of the genome contained 553 genes, 451 of which were phylogenetically classified, and the most stable 2 Mbp of the genome contained 488 genes, 414 of which were phylogenetically classified. The most fragile regions had roughly the same number of genes as the most stable regions for genes that were specific at the subphylum, phylum, and subkingdom levels, but the fragile regions had far more species-specific genes while the stable regions had far more genes that were common to all cellular life (Fig. 3.5B).

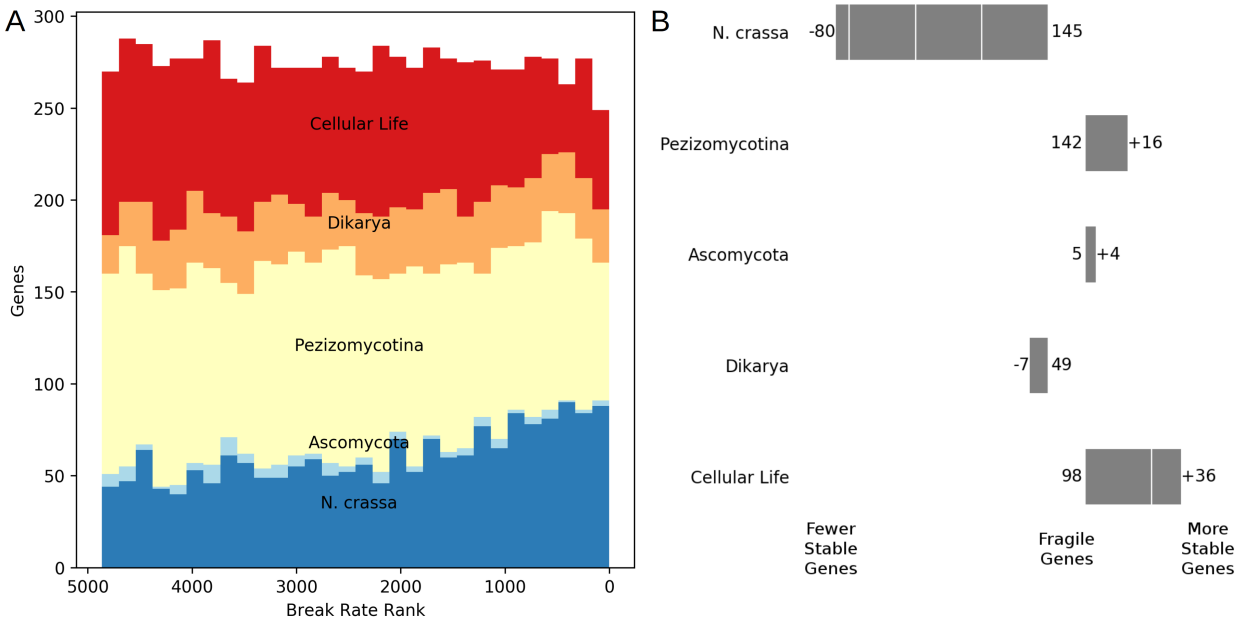

Figure 3.5: A) Stacked histograms of the number of genes within the five levels of phylogenetic specificity at different break rate percentiles show that the distribution of break rates around *N. crassa* specific genes is higher than in more highly conserved genes. B) The difference in number of genes from each category is most apparent between the most fragile (right side in A) and most conserved (left side in A) regions of the genome. For each level of phylogenetic specificity, the difference between the number of genes in the most stable 2 Mbp of the genome and in the most fragile 2 Mbp is shown. The number of genes of each category in fragile regions is given in the center.

The Two-Speed Genome of *Neurospora crassa*

Overall, we find that rearrangements abound across the genome of *N. crassa*, especially within the subtelomeric regions involved in recent adaptation. As has been described in humans (Linardopoulou *et al.*, 2005) and plasmodium (Cerón-Romero *et al.*, 2018), we observe subtelomeres that are enriched for rearrangements and recently evolved genes and are approximately 200 Kbp to 1 Mbp long.

Selection plays a clear role in filtering out rearrangements that occur within genes while permitting rearrangements in intergenic regions. In *Neurospora*, this selective filter is at least partly extended to rearrangements that contain entire genes, since a genomic defense mechanism known as Meiotic Silencing of Unpaired DNA (MSUD) silences all copies of genes for which at least one copy does not pair with a syntenous copy on the homologous chromosome during meiosis (Shiu *et al.*, 2001). While MSUD may be contributing to the correlation between the break rate and level of conservation, the subtelomeres still show a propensity for rearrangement beyond a mere selective filter. If the elevated break rate within the subtelomeres were due merely to an increased tolerance for gene disruption then

we would expect subtelomeres to be relatively gene sparse. To the contrary, we observe more genes in the most highly rearranged portions of the genome than in the most stable. It appears that selection not only filters out rearrangements that land within genes, but also favors a higher level structure of where rearrangement occurs.

Furthermore, the manner in which rearrangement creates novel genes for adaptation appears structured, since the placement of species-specific genes is extremely non-random ($\chi^2 = 5922$, df = 1, p = 0). Species-specific genes are far more likely to be surrounded by other species-specific genes than they are to neighbor a highly conserved gene. In comparing the neighborhoods of the 4,675 *N. crassa* specific genes and 6,967 genes common to cellular life, if the genome were randomly ordered we would expect 1,866 pairs of specific genes to be neighbors, 4,176 pairs of conserved genes to be neighbors, and 5,582 divergent pairs to be neighbors. In actuality, though, there are 3,858, 6,168, and 1,598 pairs of each category, respectively.

Still, the majority of variation in the break rate remains unexplained by the factors examined here, and this may reflect the heterogeneous effects of selection as well as biases in the patterns of incomplete data. However, the consistent partitioning of genes into rapidly evolving or conserved regions by organisms across Eukarya leads us to believe that there is significantly more to be learned about how organisms control the movement of genes throughout their genome.

Rearrangements can be an important source of adaptive variation (Brown *et al.*, 2010; Cerón-Romero *et al.*, 2018; Linardopoulou *et al.*, 2005; Miles *et al.*, 2016; Steenwyk and Rokas, 2018; Soucy *et al.*, 2015), and this source appears to be cultivated by eukaryotic genomes. In primates and likely other organisms, the structured repetition of subtelomeric regions facilitates rearrangements between rapidly evolving genes, while transposons aid in the transfer of adaptive genes in fungal pathogens. In *N. crassa*, an organism that keeps a meticulously clean house through mechanisms such as RIP and MSUD, the permissibility of rearrangement in the subtelomeres suggests that they are gardens of variation, rather than wild landscapes.

3.4 Discussion

Our implementation of BRAG has demonstrated that it is possible to take a comprehensive look at the rearrangement landscape of a genome with relatively little computational power. By remaining agnostic about the mechanisms of rearrangements BRAG should scale highly favorably with the addition of more or more distantly related genomes. However, BRAG's results are highly dependent upon correct inference of orthology and the underlying phylogenetic tree, and this scalability comes at the cost of potential idiosyncrasies in the results.

Scalability of BRAG

In their 2009 paper, Vandal *et al.* prove that the complexity of the algorithm to enumerate minimal covers is $O(2^{m-2})$, where $m = |\mathrm{T}|$, or the number of tbreaks in the graph. Solving disconnected subgraphs of T independently thus provides a significant advantage for this exponential time algorithm and allows for a roughly linear increase in computational time with increasing genome size. We further suspect that, for qbreak graphs, the rate at which m increases as new genomes are added will be slow because of the hierarchical nature of evolutionary trees.

Given the phylogeny relating the reference and queries, for any region of the reference genome the lowest nonzero maximum likelihood estimate of the break rate is equal to one divided by the product of the total tree length and the length of the region. This represents a kind of limit of detection for the break rate, since while the maximum likelihood estimate for an unbroken region is 0, from a biological perspective the break rate is more reasonably regarded as less than this limit. Much of the likelihood landscape across the genome is composed of regions without breaks, with a maximum likelihood of 0 and a negative slope determined by the width of the region. Improving the sensitivity of the break rate estimate is best achieved by adding genomes to the analysis that are distantly related to all other individuals, as deeper branches add more evolutionary time to the tree overall, increasing the quotient of the limit of detection (Fig. 3.8).

At the same time, adding deeply branching genomes is not likely to add much complexity to the tbreak graph. The tbreak graph is, tautologically, sparse in conserved regions, so additional tbreaks are unlikely to tie Gordian knots. In more rapidly breaking regions deeply branching genomes are unlikely to provide information, since they are more likely to fall behind an existing inferred tbreak and be effectively masked (Fig. 3.8). Adding deep-branching genomes thus provides more information about conserved regions of the genome by providing more opportunity to observe rare events, but does not provide much information about rapidly rearranging regions.

Within regions of the genome that have at least one qbreak, the accuracy of placing tbreaks on the tree is limited to the discrete internode distances (the branch lengths). Shorter internal branches allow for more precise placement of tbreaks, and thus more precise estimates of the break rate. The additional internally branching genomes are similarly unlikely to increase the complexity of the tbreak graph, but rather just to refine the placement of tbreaks.

Effects of Poorly Resolved Trees and Paralogy

While the inclusion of more data should generally yield better results, including ambiguously placed genomes in the analysis must be avoided. An incorrect tree topology will systematically duplicate tbreaks, leading to wild overestimates of the break rate. Such a situation can arise when there is recombination between individuals included in the analysis. Recombination results in different trees being true for different regions of the genome. Thus,

a systematic bias will be introduced for regions that do not match the consensus tree. Including three individuals from a recombining population with architectural variation could nearly double the estimate of the break rate for half the genome. Horizontal transfer between species in the analysis could similarly multiply the break rate within the transfered region. Recombination will also lead to poorly resolved branches, which is why BRAG is only suitable for high-confidence phylogenies, or when regions that don't match the consensus are masked (e.g. the centromeres in this analysis).

Similarly, paralogous segments that are misidentified as orthologous can lead to systematic overestimation of the break rate. Such a situation could occur following gene duplication and pseudogenization. If the orthologous copy is pseudogenized, then alignment may identify the working paralogue as orthologous, leading to three breaks observed where there should be one. Alignment methodologies that either allow for multiple homology relationships (e.g. cactus graphs) or that favor synteny over sequence identity should therefore be used. The former solution is particularly appealing. Although reading cactus alignments is not implemented by BRAG, we believe a method that traces paths on cactus graphs to identify tbreaks, combined with BRAG's tbreak placement on trees and Poisson rate modeling would be particularly promising.

Effects of Genome Assembly Quality

BRAG is sensitive to the quality of the genome assemblies used, although we think the substantial similarity in break patterns using only "true" or both "true" and "false" qbreaks indicates that the general trends are robust to this sensitivity. In simulations using a similar framework, Zheng and Sankoff (2016) found detection of rearrangement events to be surprisingly robust to assembly fragmentation. Still, while we use estimates using only "true" or both "true" and "false" qbreaks as lower and upper bounds, respectively, this characterization is not entirely accurate since the presence or absence of a qbreak can complete, disrupt, grow, or shrink a tbreak. Zheng and Sankoff describe a superior treatment of ambiguous breaks: converting them into greater temporal censoring by allowing tbreaks to be placed across a subset of valid branches.

Limitations and Biases

BRAG remains agnostic about the mechanism and nature of the underlying rearrangements by measuring the age of existing bonds, with the assumption that younger bonds replaced more fragile bonds in the past. In balanced rearrangements, where two bonds are broken and the ends rejoined in a new arrangement, this assumption makes intuitive sense. But in unbalanced rearrangements where the total number of bonds is increased or decreased the number of tbreaks observed does not necessarily match the number of double-strand breaks or joins that occurred. Insertions break one bond and rejoin the ends with a series of new bonds, but are observed as a single large tbreak. Conversely, deletions break two bonds and join the ends in a single bond, but are still observed as a single tbreak,

this time of size 1. The point estimate of the break rate at the site of the deletion will be much higher than the insertion due to the tbreak's smaller size, despite the fact that the events are reversals of each other. However, the estimates should be substantially similar after smoothing (e.g by sliding windows), and we believe that BRAG captures the essential nature of fragility as the evolutionary tolerance of a region for rearrangement.

BRAG favors choosing tbreaks that are higher up on the tree (i.e. closer to the reference) over tbreaks that are closer to the queries when both choices are equally parsimonious. This preference is due to only considering tbreaks that are maximal partitions of maximal cliques in our search for covers. Such a situation arises in Fig. 3.7. τ_1 could be equally well explained by a tbreak on the ancestral branch of A and B or by a tbreak on the branch leading to B. Our algorithm dismisses the latter tbreak and favors the former tbreak because is is maximal.

This limitation results in two biases: an overestimation of the break rate in such regions, and a clustering of inferred breakpoints, such that the overestimation is much greater in one region while the neighboring region is underestimated. The overestimation of the break rate is due to selecting the highest possible placement of the tbreaks, and thus inferring the shortest possible observed evolutionary time in that region. The clustering bias arises because the placement of the tbreaks is bounded by the overlapping region of its qbreaks, so the two tbreaks evidenced by the shared qbreak will tend to be closer together. For example, in Fig. 3.7, if $min(A_1) > min(B_1)$ (i.e. the left endpoint of A_1 is greater than the left endpoint of B_1) then the choice of placement for the tbreak would lie between $min(A_1)$ and $max(B_1)$ (the right endpoint of B_1), while if $min(A_1) \leq min(B_1)$ then the choice of placement for the tbreak would be the same as a tbreak on the branch leading to B (i.e. between $min(B_1)$ and $max(B_1)$).

There are biological reasons to favor a clustering bias. Indeed, we add to the evidence that rearrangements are more common in domains involved in recent adaptation. However, this bias would be better modeled explicitly. We acknowledge this limitation to our methodology, but hope that BRAG will provide better resolution on the placement of break sites, enabling future models to account for these and yet unsupposed effects. With additional work, it may be possible to explore histories utilizing these non-maximal tbreaks.

Assumptions of BRAG

In addition to exploring non-maximal tbreaks, it may be advantageous to explore non-minimum histories. The restriction to minimum histories is imposed by the assumption of maximum parsimony, which is itself a weak assumption. A better method would employ a purely maximum likelihood approach, integrating across both minimum and less likely histories. Such a set of solutions could be very large, and a sampling methodology like that described by Vandal *et al.* could be utilized to explore the space of histories in a manner similar to Bayesian tree exploration and produce a more complete probabilistic model of the evolutionary history.

Strain	Taxonomy	Scaffolds	Genome Size	Longest Scaffold	N50	Sequencing Technology	Read Length	Raw Cov	Assembler	Version	Reference
FGSC1740	N. africana	1965	37046517	286522	59428	Illumina GA II	55/76	53	A5-miseq 20140604	8012016	Gioti *et al.* (2013)
FGSC26649	N. cerealis	91	39288751	2583041	954150	Illumina HS4000	151	51	A5-miseq 20140604	6032016	Chapter 1
OR74A	N. crassa	20	41037538	9798893	6000761	Sanger	varied	20	Arachne	12[a]	Galagan *et al.* (2003)
FGSC8579	N. discreta PS4B	176	37302679	4845008	2306035	NA	NA	NA	NA	FungiDB-3.1	Joint Genome Institute
FGSC26659	N. discreta sp.	941	48283943	698905	112110	Illumina HS4000	151	55	A5-miseq 20140604	10132017	Chapter 1
FGSC7221	N. pannonica	2915	41672953	156060	33804	Illumina HS4000	151	41	A5-miseq 20140604	10252017	Chapter 1
FGSC6877	N. sp.	908	38131013	547143	152004	Illumina HS4000	151	66	A5-miseq 20140604	10132017	Chapter 1
FGSC26585	N. sp.	142	51833329	4326500	1770967	Illumina HS4000	151	73	A5-miseq 20140604	6032016	Chapter 1
FGSC26605	N. sp.	269	57662837	2488616	693409	Illumina HS4000	151	98	A5-miseq 20140604	6032016	Chapter 1
FGSC26614	N. sp.	330	53494617	2057027	517442	Illumina HS4000	151	41	A5-miseq 20140604	10132017	Chapter 1
FGSC26633	N. sp.	213	38307754	2116602	535266	Illumina HS4000	151	73	A5-miseq 20140604	6032016	Chapter 1
FGSC5508	N. sublineolata	5272	34946819	73969	14101	Illumina GA II	55/76	34	A5-miseq 20140604	8012016	Gioti *et al.* (2013)
FGSC1889	N. terricola	4655	42004691	228174	54719	Illumina HS4000	151	39	A5-miseq 20140604	10132017	Chapter 1
FGSC2508	N. tetrasperma	81	39146333	9915204	5681182	Roche 454/Sanger	varied	31	Newbler	FungiDB-3.1	Ellison *et al.* (2011b)
FGSC26668	N. tetraspora	127	43558099	4246662	1100031	Illumina HS4000	151	75	A5-miseq 20140604	6032016	Chapter 1
FGSC10222	Sordaria macrospora	4783	39955017	2538341	498327	Solexa/454	varied	85	Velvet 0.7.31	FungiDB-3.1	Nowrousian *et al.* (2010)

[a]Corrected for the assembly error detected by Galazka *et al.* (2016)

Table 3.1: Genome assemblies used in this work.

In addition to assuming a maximum parsimony history, BRAG assumes that the reference genome sequence and the organismal ecology remain unchanged throughout it's evolutionary history. In reality, the evolving sequence of the genome (as well as the structure itself) influence the rearrangement dynamics. Rearrangements may become common in a region along a lineage following null-functionalization or an alteration of the epigenetic structure. Our inference of the rearrangement rate in such a region does not reflect the evolutionary dynamics of the reference itself, but rather an integration of it's historical and historical-adjacent states that are found in the tree.

Similarly, demographic changes, and external or internal ecological shifts across the tree can change the rearrangement dynamics. Such a shift is seen in *Neurospora* where species have consistently but independently transitioned from self-sterility to self-fertility (Nygren *et al.*, 2011). This shift is associated with profound changes in evolutionary paradigm (Gioti *et al.*, 2013), however we found that it is not associated with changes in rearrangement rates (Chapter 2). Nevertheless, our results in *N. crassa* should be seen as a partial integration of the effect of self-fertility, induced by this lineage's proclivity for transitions to self-fertility.

The ability to assay rearrangement dynamics on a large scale, facilitated by advances in sequencing and analysis methodologies, brings new questions into the fields of molecular biology, evolution, and ecology, such as the effect of sexual transitions in *Neurospora*. In particular, we believe there is still much to learn about evolution within the fields of molecular and ecological phylogenomics.

3.5 Materials and Methods

Here, we describe the BRAG method in detail.

Genomes and Phylogeny

We utilized a dataset of 15 *Neurospora* genomes and one *Sordaria* genome to serve as the outgroup. Sequencing and assembly methodologies, assembly statistics, and references are described in Table 3.1. We selected the well studied and complete genome of *N. crassa* as the reference genome.

Since the BRAG methodology attempts to place break events on the branches of a phylogenetic tree to maximize parsimony, using the correct tree topology is essential to accuracy. Furthermore, the break rates estimated by BRAG are scaled by the branch lengths of the tree, so results are biased by inaccurate branch lengths. An accurate alignment and high quality tree are therefore essential to correct inference.

We identified orthologous sequences between the reference and all queries in a pairwise manner with the MUMmer package (Kurtz, 1999; Delcher *et al.*, 1999, 2002; Kurtz *et al.*, 2004), first finding maximal alignments with nucmer with a gap length of 2000bp, then filtering for the highest scoring 1-to-1 set of alignments with delta-filter. We filtered for regions of the reference that had aligned orthologous regions to all queries. We then performed multi-

ple sequence alignment between orthologues of all such regions using MAFFT (Katoh *et al.*, 2002; Katoh and Standley, 2013) using the –globalpair setting for global alignment, 1000 maximum refinement iterations, and the JTT 10 model (Jones *et al.*, 1992). We discarded regions on the ends of alignments for which any sequence had missing data before concatenating all alignments into a single 11,400,437 character alignment with 1,648,053 parsimony-informative sites. We partitioned the concatenated alignment into bins by their approximate rate of evolution using TIGER (Cummins and McInerney, 2011), which we re-wrote to fix errors, optimize for genome-scale alignments, and bin using the method described by Rota *et al.* (2017). Finally, we searched for the maximum-likelihood tree with IQTree using the -m MFP option for model selection, generating 1000 ultrafast bootstraps, and estimating branch support using 1000 replicates in the non-parametric Shimodaira-Hasegawa-like approximate Likelihood Ratio Test (SH-aLRT) (Nguyen *et al.*, 2015; Kalyaanamoorthy *et al.*, 2017; Thi Hoang *et al.*, 2018; Chernomor *et al.*, 2016).

Our whole genome phylogeny of *Neurospora* is fully supported by both metrics, and thus should be suitable for BRAG analysis.

Whole Genome Alignment

For BRAG analysis, we roughly aligned all the genomes to the *N. crassa* reference genome using Murasaki and OSfinder (Popendorf *et al.*, 2010; Hachiya *et al.*, 2009). Murasaki identifies roughly similar short sequences, or anchors, between genomes. We configured Murasaki to find 36-mer anchors with at least 28 matches. OSfinder identifies collinear chains of anchors and attempts to find the most likely set of orthologous segments from non-overlapping chains. We configured OSfinder to find orthologous segments at least 1,000 bp long.

Query Breaks

For each pairwise genome alignment we define breakpoints to a query (or **qbreaks**) as the regions on the reference genome between two orthologous segments (Fig. 3.6). Qbreaks have both left and right endpoints and a query, with the property that no two qbreaks of the same query are overlapping.

Orthologous segments are defined by sequence, or nucleotide base, identity and are closed intervals on the reference sequence. As such, breakpoints occur on the bonds between nucleotides and qbreaks are the open intervals between the orthologous segments. When two orthologous segments are immediately adjacent there is thus a qbreak between them encompassing a single bond.

Additionally, due to fragmentation of the query genomes, orthologous segments that are on the ends of the scaffolds could be adjacent on a chromosome or not. We define qbreaks where one or both orthologous segments are internal on a scaffold, and thus their adjacencies are known, as "true." Conversely, qbreaks where both orthologous segments are on the ends of a scaffold in the query as "false," equivalent to "can't tell" vertices described by Zheng and Sankoff (2016). As a conservative estimate, we calculate break rates utilizing only

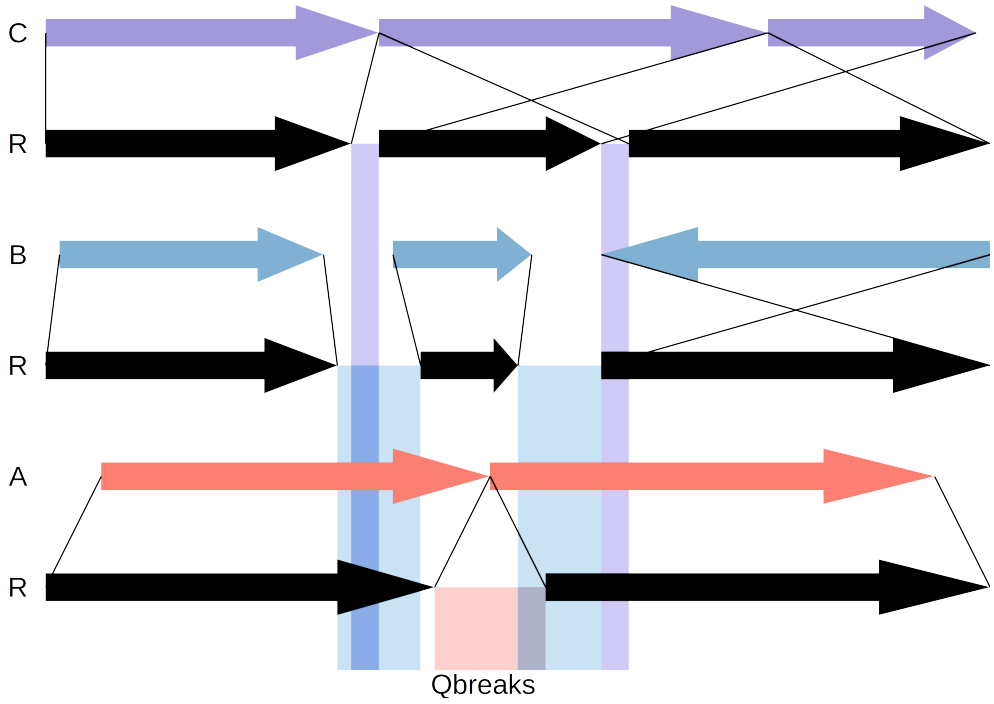

Figure 3.6: Three toy alignments between query genomes (A, B, and C) and the reference (R). Orthologous segments between the reference and each query are shown as colored block arrows, with thin black lines showing the orthology relationship. Qbreaks are represented as shaded rectangles emanating from the reference at regions between orthologous segments, and colored by their corresponding query alignment. No two qbreaks of the same color can overlap, but qbreaks of different colors (i.e. queries) can overlap.

"true" qbreaks, which represents a lower-bound on the true break rate. For an upper-bound estimate, we calculate the break rate using both "true" and "false" qbreaks.

Identification of Breakpoints

Given the set of qbreaks, we seek to identify double-strand breaks that have occurred along the branches of the evolutionary tree connecting the reference to the queries. Multiple qbreaks can be explained by the same double-strand break event when they overlap and when their queries share a branch leading to the reference. Under the paradigm of maximum parsimony, we favor these breaks that can explain the existence of the most qbreaks with the fewest number of events. Similarly, we do not consider the possibility of reversions, which would require the same sites to break again at the same time and the ends of the fragments to exchange back to their original orientation and adjacencies - an event we assume to have a vanishingly small probability.

To this end, we frame the problem in graph theoretic terms and find that it is a case of the **minimum clique cover problem**. Well known in mathematics as one of Karp's

original 21 NP-complete problems (Karp, 1972), the minimum clique cover problem is any problem that is equivalent to the following: Knowing which students at a school share a class, find the minimum number of classes that must be offered and which students are in which classes.

Tree Consistent Breaks

We define a tree consistent breakpoint (or **tbreak**) as a set of overlapping qbreaks whose queries form a **tree consistent** partition (Fig. 3.7). In the manner of Alekseyev and Pevzner (2009), a partition of queries is tree consistent if it or its complement form a monophyletic group. That is, a partition of queries is tree consistent if it is a bipartition present in the tree. A tbreak's placement in the tree corresponds to this bipartition (or branch). We use $\tau \in \mathrm{T}$ to denote a tbreak within the set of all tbreaks. Since tbreaks are bounded by their member qbreaks, the left endpoint of a tbreak ($min(\tau)$) is the maximum left endpoint of it's members. Analogously, a tbreak's right endpoint ($max(\tau)$) is the minimum right endpoint of it's members.

Finding the most parsimonious history of double-strand breaks for a given genome over the observed tree is equivalent to finding the smallest set of tbreaks that includes each qbreak at least once. We note that this solution may include a single qbreak in multiple tbreaks, corresponding to inferring multiple break events within the region of a qbreak. Such an inference is consistent with the data since qbreaks evidence at least one break within them. Multiple hits in the same region could be masked by a concomitant deletion associated with the region, or by sequence divergence following null functionalization at the site of the break. However, we note that this method does create a bias in the placement of double-strand breaks, that is, to make them more clustered.

Synteny as an Interval Graph

Let the set of qbreaks be vertices in a graph where undirected edges represent an overlap between two qbreaks. Such a graph is known as an **interval graph**. We find it interesting to note that the queries of the qbreaks are a coloring of the interval graph, and the chromatic number of the graph is therefore at most the number of queries.

A tbreak is a set of fully connected qbreaks in the interval graph (i.e. a **clique**) whose members form a tree consistent partition. Finding the smallest set of tbreaks that includes each qbreak at least once is equivalent to finding a minimum cover using tbreaks. The minimum clique cover problem is NP-hard for arbitrary graphs, but can be solved in polynomial time for interval graphs. We follow a similar method to that described by Vandal *et al.* (1997) to enumerate all minimum covers of the interval graph using tbreaks.

A clique is considered **maximal** if it is not a proper subset of any other clique (i.e. it cannot be grown by adding another qbreak). A clique cover merely needs to contain each qbreak at least once, so can be composed of overlapping cliques. A small non-maximal clique in a cover can therefore always be replaced by a maximal clique that is its superset.

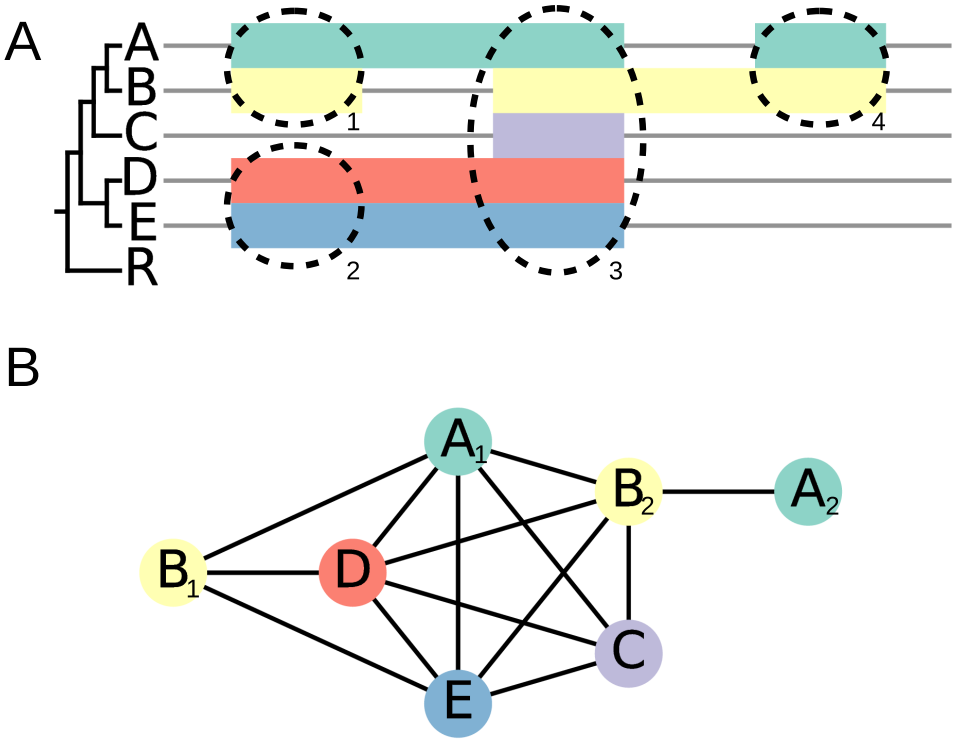

Figure 3.7: A) A toy alignment of five genomes (A, B, C, D, and E) to a reference genome (R). Each row (and color) corresponds to an alignment of a query to the reference. The horizontal position corresponds to the position in the reference genome. Grey lines represent orthologous segments between the reference and corresponding query, while colored bars represent qbreaks. The cladogram of the six genomes is shown on the left side, which is rooted on the branch leading to the reference. Dashed circles represent tbreaks $\{\tau_1, \tau_2, \tau_3, \tau_4\}$ B) The interval graph of the qbreaks in A. Qbreaks are labeled and colored according to their query, and where more than one qbreak occurs in an alignment to a query they are indexed by their positions. Black edges are drawn between qbreaks that overlap each other. Note that tbreaks τ_3 and τ_4 are both maximal cliques (as can be seen in B) and maximal tbreaks (as can be seen in the tree on A), while τ_1 is a maximal tbreak but not a maximal clique since $\{A_1, B_1, D, E\}$ forms a clique. τ_2 is neither a maximal clique nor a maximal tbreak, since it is a subset of τ_3, but is discovered (and subsequently discarded) by our algorithm.

A minimum clique cover composed of maximal cliques therefore exists. Minimum covers composed of non-maximal cliques can also exist, but the set of non-maximal cliques can be very large. In order to keep the problem tractable we consider only tbreaks that are maximal partitions of maximal cliques. This simplification leads to the breakpoint clustering bias previously discussed.

We refer to the multiple solutions to the tbreak cover problem as **histories**, as they represent multiple evolutionary histories that are equally parsimonious.

The unique feature of this problem relative to other minimum clique cover problems is the restriction of cliques to be tree consistent, rather than simply maximal. Once maximal cliques are found, they must be partitioned into their maximal tbreaks. Additionally, the method of Vandal *et al.* must be modified since maximal cliques have a strict ordering, while tbreaks are only partially ordered (Fig. 3.7). Fortunately, the low chromatic number of the qbreak interval graph ensures relatively low connectivity within the graph, leading to a single, obvious, best solution for large sections of the graph.

Estimation of the Break Rate

The resulting histories represent equally likely sets of inferred double-strand breaks that occurred within a known range of space (across the genome) and time (along the evolutionary tree). From these data, we seek to make a maximum likelihood estimate of the break rate across the genome. We compute such an estimate by calculating the joint likelihood across the four forms of uncertainty: 1) the underlying rate that produced the observed count of breaks, 2) the temporal placement of the break within each tbreak in the history, 3) the spatial placement of the break within each tbreak in the history, and 4) the selection of the history.

Due to the rarity of events the resulting landscape is highly rugged, with high rates estimated at bonds where breaks are observed and rates near zero estimated within unbroken regions. To make more even estimates of the break rate across a region, we apply a rectangular (sliding window) smoothing function to the counts of breaks across the genome and calculate likelihoods within the windows.

Poisson Model of Double-Strand Breaks

Starting with the uncertainty of the underlying rate, we model rearrangements as an independent Poisson process for each phosphodiester bond in the genome. Individual bonds break, or decay, at a constant rate, λ, and the break rates across the genome are denoted $\{\lambda_1, ..., \lambda_N\}$ where N is half the number of phosphodiester bonds in the genome. For a given region of the genome defined by the closed interval $[p, q]$, the break rate within that region is estimated as the number of breaks observed, c, divided by the amount of evolutionary time that region is observed, t.

$$\hat{\lambda}_{[p,q]} = \sum_{i=p}^{q} \lambda_i = \frac{c}{t} \tag{3.1}$$

The maximum likelihood estimate of the break rate, λ_i, for an individual bond within the region, $i \in [p, q]$, is equal to the total rate divided by the number of bonds.

$$\hat{\lambda}_i = \frac{\hat{\lambda}_{[p,q]}}{q - (p-1)} \tag{3.2}$$

The likelihood function of the break rate for a region can be determined from the likelihood function for the mean number of breaks within the region. From the rate for the region, the mean number of breaks within the region is simply $\mu_{[p,q]} = \lambda_{[p,q]} t$. The likelihood function for the mean number of events in a Poisson process is

$$L(\mu \,|\, c) = P(c \,|\, \mu) = \frac{e^{-\mu}\mu^c}{c!} \tag{3.3}$$

so it follows that:

$$L(\lambda_{[p,q]} \,|\, c) = \frac{e^{-\mu_{[p,q]}}\mu_{[p,q]}^c}{c!} \tag{3.4}$$

We use this model to compute the expectation and likelihood landscape of λ for sliding windows across the reference genome.

Observed Evolutionary Time

The time for which we observe a phosphodiester bond is equal to the combined length of branches of the phylogeny that can be inferred to share that bond. As a decay process, we cannot observe a bond after it is broken. However, we can still observe multiple breaks for a given bond, since speciation acts as a birth process for the bond (Fig. 3.8). We represent the observed evolutionary time for each bond in the genome as $\{t_1, ..., t_N\}$, where t_i is dependent on the pattern of tbreaks at that bond.

The position of a break along the branch of a tbreak cannot be known, a condition known as "interval-censoring." Assuming that new bonds after a break are exaclty as stable as the ancestral bonds or that breaks are rare events, the probability density function of the time to an event over an interval of time given that exactly one event occurred is approximately uniform. We therefore account for the temporal uncertainty in determining the evolutionary time observed for the region of the tbreak by simply using the midpoint of the tbreak's branch, which is the expectation for the observed time.

When multiple tbreaks overlap spatially, each break masks the evolutionary history on the tree beyond the break from the perspective of the reference. For the region of overlap, we therefore calculate the total length of the masked branches and subtract it from the total length of the tree to determine the observed evolutionary time.

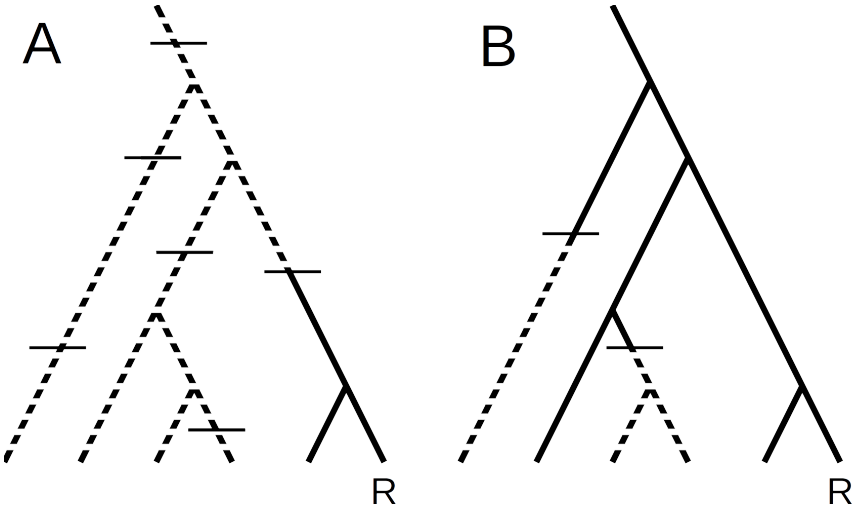

Figure 3.8: A hypothetical phylogeny showing break events at a rapidly breaking bond (A) and a relatively conserved bond (B). From the perspective of the reference state (labeled "R"), multiple hits beyond the most recent cannot be observed. The unobservable evolutionary time is masked (dashed lines) beyond these most recent events. The observed lifetime of the bond is equal to the sum of the lengths of the solid branches. At rapidly breaking bonds (A), a break in a recent ancestor of the reference is more likely to have occurred, and deeply branching queries provide no new information. At conserved bonds (B), however, deeply branching queries add significantly to the length of the tree observed and thus provide a greater chance for observing rare events.

Spatial Placement of Breaks within Tbreaks

Just as breaks cannot be precisely placed in evolutionary time, breaks are also interval-censored spatially. When two orthologous segments of a query are immediately adjacent to each other, the inferred qbreak between them consists of a single phosphodiester bond and the break can be precisely placed. Where there is a gap in the alignment, however, the placement of the break on any particular bond can not be known. Such qbreaks represent "hidden" breakpoints (Sankoff and Blanchette, 1998). This situation can arise in repetitive or other regions where our ability to align the genomes is poor, when the rearrangement is associated with an insertion or deletion, or where the sequence has degenerated around the site of the break following rearrangement. We expect degeneration of conserved elements at the break point of a rearrangement to be common since rearrangements often result in null-functionalization.

As in the larger genome, we expect the likelihood of a break to be heterogeneous across the bonds within a tbreak. In the absence of an adequate predictive model that incorporates both structural predilection and selective constraints, though, we use a uniform model for the spatial placement of breaks within a tbreak. The likelihood that a tbreak (τ) is an

inference of a break within the interval $[p, q]$ is simply the number of bonds that the interval overlaps with the tbreak divided by the number of bonds the tbreak covers. Or formally:

$$L(\ \rho(\tau, p, q)\ ) = \frac{max\{0,\quad min[p,\ max(\tau)] - (max[q,\ min(\tau)] - 1)\}}{max(\tau) - [min(\tau) - 1]} \tag{3.5}$$

Given multiple tbreaks that overlap the region $[p, q]$, denoted $\mathrm{T}_{[p,q]}$, the combinations of tbreaks that could be placed within the region is the powerset of $\mathrm{T}_{[p,q]}$. Let $\chi \in \mathcal{P}(\mathrm{T}_{[p,q]})$ denote a combination of tbreaks potentially in the region $[p, q]$ and n be the number of overlapping tbreaks, $|\mathrm{T}_{[p,q]}|$. The joint likelihood for each possible combination of tbreak placements is the product of the likelihoods that each tbreak in the combination, $\tau \in \chi$, is placed in the region, $L(\ \rho(\tau, p, q)\ )$, times the product of the likelihoods that each tbreak not in the combination, $\tau \in \mathrm{T}_{[p,q]} \setminus \chi$, is not placed in the region, $(1 - L(\ \rho(\tau, p, q)\ ))$, divided by the number of combinations, 2^n:

$$L(\ \chi \mid \mathrm{T}_{[p,q]}\ ) = \frac{1}{2^n} \cdot \prod^{\chi} L(\ \rho(\tau, p, q)\ ) \cdot \prod^{\mathrm{T}_{[p,q]} \setminus \chi} (1 - L(\ \rho(\tau, p, q)\ )) \tag{3.6}$$

Combining Histories

Given the set of k maximally parsimonious histories, $H = \{h_1, ..., h_k\}$, the marginal likelihood of a given rate at a site is the sum of the likelihoods over all histories. Under the paradigm of parsimony, though, we assume that all histories are equally likely, so:

$$L(\lambda_i|H) = \sum_{j=1}^{k} L(\lambda_i|h_j) \tag{3.7}$$

In computing the histories, though, it is easiest to compute a set of solutions for each disconnected (and therefore independent) subgraph of the qbreak graph. As described later, some tbreaks are obviously essential to all minimum cover solutions and their members can be effectively eliminated from the qbreak graph. The elimination of qbreaks from the graph can create disconnected subgraphs that may be solved independently, but whose qbreaks overlap spatially and which must be combined to achieve the full solution.

The resulting histories from the algorithm described below consist of a multiset of independent solutions, each of which is a multiset of minimum covers for a disconnected subgraph. The complete number of histories, being the product of the number of covers of each subgraph, has the potential to be prohibitively large. However, in estimating the break rate at a particular bond in the genome, those subgraphs that don't include a qbreak that overlaps the bond can be disregarded, reducing to one or a few the number of subgraphs that need to be considered at a time.

For a qbreak graph that is composed of κ disconnected subgraphs, let $\Gamma = \{\gamma_1, ..., \gamma_\kappa\}$ be the family of minimum cover solutions for each disconnected subgraph of the qbreak graph. Each subgraph solution, γ, is itself the family of the minimum tbreak covers of the

corresponding subgraph, and these covers are themselves sets of tbreaks. The set of histories is therefore equal to the unions of the κ-tuples of covers resulting from the κ-fold Cartesian product of all $\gamma \in \Gamma$. Let f be this function such that $f(\Gamma) = H$.

For each region defined by a unique pattern of tbreaks $[p, q]$, we remove all tbreaks that don't overlap the region to form $\Gamma_{[p,q]}$ which is the multiset of multisets of covers consisting only of tbreaks that overlap the region. We define the function, g such that $g(\Gamma, p, q) = \Gamma_{[p,q]}$. Clearly, $\Gamma_{[p,q]}$ should be equivalent to Γ within the region $[p, q]$, so that $L(\lambda_i|\Gamma) = L(\lambda_i|\Gamma_{[p,q]})$ where $i \in [p, q]$. If we define g as a recursive function that removes all non-overlapping tbreaks from a multiset and any multiset within that multiset, then $g(f(\Gamma), p, q) = f(g(\Gamma, p, q)) = H_{[p,q]}$. Since g will completely empty most covers within the subgraphs, most $\gamma \in \Gamma_{[p,q]}$ will be a multiset of one or more empty sets. Any such subgraph, $\gamma_\emptyset$, will be the multiplicative identity with regards to $L(\lambda_i|f(\Gamma_{[p,q]}))$, since $f(\{\gamma_\emptyset, \gamma_j\})$ will equal the multiset composed of γ_j unioned to itself $|\gamma_\emptyset|$ times. Remembering that all histories are assumed equally likely, this means $L(\lambda_i|\{\gamma_\emptyset, \gamma_j\}) = L(\lambda_i|\{\gamma_j\})$, and all such subgraphs can be discarded in calculating $H_{[p,q]}$.

Consolidation of Uncertainty

We note that c, the number of breaks counted in the region $[p, q]$, is equal to $|\chi|$, the number of tbreaks placed in the region. Additionally, each history, $h \in H_{[p,q]}$, is a possible set of tbreaks that could occur within the region $[p, q]$, that is, a possible instance of $\mathrm{T}_{[p,q]}$. Summing across possible histories, $h_j \in H_{[p,q]}$, and again across the possible placement combinations, $\chi_i \in \mathcal{P}(h_j)$, we combine equations 3.4, 3.6 and 3.7, above, to compute the likelihood of a given break rate, λ over the region $[p, q]$:

$$L(\lambda_{[p,q]} \mid H_{[p,q]}) = \sum_{j=1}^{k} \sum_{i=1}^{2^n} L(\mu_{[p,q]} \mid |\chi_i|) \cdot L(\chi_i|h_j) \tag{3.8}$$

While the calculation of joint likelihoods of all combinations within the region takes $O(k2^n)$ time, $k = |H_{[p,q]}|$ and $n = |h_j \in H_{[p,q]}|$ should both be be quite small except for large trees in regions of the genome that were ancestrally conserved but have diverged in multiple more recent lineages. This computational burden is multiplied by the sampling resolution: the number of rates for which the likelihood is computed, and the number of windows across the genome.

Algorithm to Enumerate Maximum Parsimony Histories

Here, we describe a method to solve the tbreak cover problem and enumerate all maximum parsimony histories. First, we recognize that all maximal tbreaks can be formed by partitioning the maximal cliques in the qbreak interval graph. After finding this set of plausible tbreaks, we reduce the size of the problem substantially in the manner of Vandal *et al.*, and finally recursively enumerate the minimum covers using these tbreaks.

Algorithm 1 Exhaustive enumeration of minimum covers of an interval graph of $\subset$-minimal qbreaks. The choice of which qbreak to assign to x in the implementation of this algorithm can be made arbitrarily, as all non-redundant qbreaks will eventually be visited by successively deeper levels of recursion. x^* is the set of maximal tbreaks x is affiliated with. The algorithm is initialized with a set of connected non-simplicial qbreaks, the simplicial qbreaks connected to the first set, and the essential tbreaks that cover the simplicial qbreaks.

```
minimum_covers(qbreaks, covered_qbreaks, accepted_tbreaks):
    if qbreaks = ∅:
        return {accepted_tbreaks}
    else:
        covers ← ∅
        x ← any qbreak within qbreaks
        for τ in x*:
            remaining_qbreaks ← qbreaks \ τ
            new_covered ← covered_qbreaks ∪ τ
            new_accepted ← accepted_tbreaks ∪ {τ}
            new_covers ← minimum_covers(remaining_qbreaks,
                                        new_covered,
                                        new_accepted)
            covers ← covers ∪ new_covers
        return all_minimum(covers)

all_minimum(covers):
    l ← min{|C| : C ∈ covers }
    return {C ∈ covers : |C| = l}
```

Finding Maximal Cliques

Via transitivity, any tbreak that is an improper subset of a non-maximal clique must be contained in a maximal clique. Therefore we need only consider the maximal cliques of the qbreak interval graph. Finding all maximal cliques in a graph is known as the **clique problem**, which is well known to be NP-complete for arbitrary graphs, but is equivalent to a sorting problem for interval graphs. We use an algorithm very similar to the one described by Gupta *et al.* (1982), which follows:

First, we sort the left and right endpoints of all qbreaks together, with right endpoints preceding left endpoints when their positions are equal. We proceed through the sorted list of endpoints, adding each qbreak to a set when we encounter its left endpoint, and removing it from the set when we encounter its right endpoint. At each left endpoint that immediately precedes a right endpoint we record the set as a maximal clique.

Partitioning Maximal Cliques into Tbreaks

We partition maximal cliques into tbreaks by finding their tree-consistent subsets. Remembering that cliques are defined as fully connected sets of graph elements, we note that every subset of a clique is itself a clique. Furthermore, since the tree is hierarchical, for any two non-identical tbreaks that contain the same qbreak one must be a proper subset of the other, and therefore is not maximal. There is therefore a single partitioning of a maximal clique into maximal tbreaks which is equivalent to the minimum partition. With respect to the graph of qbreaks as a whole, however, some of these tbreaks may be non-maximal. These non-maximal tbreaks arise when neighboring maximal cliques are partitioned such that one tbreak is a subset of a neighboring tbreak. For example, in Fig. 3.7, τ_2 is maximal within the clique $\{A_1, B_1, D, E\}$ but not within the clique $\{A_1, B_2, C, D, E\}$. Such non-maximal tbreaks can be discarded following the same logic as for discarding non-maximal cliques.

To partion each maximal clique into its maximal tbreaks, we first find the origin of the reference state on the rooted tree by climbing from the reference leaf until we reach the ancestral branch that subtends all taxa with the reference state (i.e. that are not represented in the clique). If the origin branch is not the root of the tree, we place a tbreak on the origin branch and remove qbreaks with queries on the other side of the break from the clique. While qbreaks remain in the clique, we select a qbreak arbitrarily and climb the tree to find the highest branch that subtends only queries represented in the clique. We place a tbreak on that branch and remove all qbreaks with queries that are under it.

This process of partitioning maximal cliques into tbreaks can yield multiple instances of the same tbreak (i.e. tbreaks whose members are identical) in adjacent cliques. We cull these redundant tbreaks to produce the set of maximal tbreaks.

Enumerating Histories

From the set of plausible tbreaks, we enumerate all minimum covers in the manner of Vandal *et al.*. Prior to enumeration, Vandal *et al.* make several observations that enable a significant reduction of the size of the problem, which we summarize here.

First, we consider only covers of the **$\subset$-minimal** qbreaks. We refer to the set of tbreaks that a qbreak belongs to as the **affiliations** of the qbreak, denoted x^*, where x is a qbreak. Qbreaks whose affiliations contains no other set of affiliations as a proper subset are termed "$\subset$-minimal." For example, in Fig. 3.7, A_1 and B_2 are not $\subset$-minimal because their affiliations ($\{\tau_1, \tau_3\}$ and $\{\tau_3, \tau_4\}$, respectively) are subset by the affiliations of C ($\{\tau_3\}$), which is $\subset$-minimal.

The affiliations of a qbreak that is not $\subset$-minimal contain exactly the affiliations of at least one $\subset$-minimal qbreak. Each qbreak need only be covered by a single tbreak, so any tbreak that covers the $\subset$-minimal qbreak also covers any non $\subset$-minimal qbreaks whose affiliations contain the affiliations of the $\subset$-minimal qbreak. Therefore, the minimum covers of the $\subset$-minimal qbreaks are exactly the minimum covers of all the qbreaks. For example,

in Fig. 3.7, any tbreak cover of C will also cover A_1 and B_2, so A_1 and B_2 can be ignored in finding minimum covers.

Second, we can include all essential tbreaks *a priori*. Qbreaks that are affiliated with a single tbreak are termed "**simplicial**," and the tbreak that covers them is termed "essential." All simplicial qbreaks are necessarily $\subset$-minimal. We need simply append these essential tbreaks to covers of the non-simplicial qbreaks.

Third, we can find minimum covers of the disconnected subgraphs of the remaining non-simplicial, $\subset$-minimal qbreaks and combine them to achieve the full solutions as described previously.

We then exhaustively enumerate the minimum covers for each disconnected sub-graph using Algorithm 1, initialized with the set of qbreaks in the subgraph, the simplicial qbreaks adjacent to the graph, and the essential tbreaks, respectively. Vandal *et al.* first presented the substance of this algorithm as Algorithm 3.5, although with an error. We present the corrected form here, as well as modify the algorithm slightly due to the lack of a linear ordering on simplicial qbreaks (simplicial elements in an interval graph have a natural ordering) and to return only minimum, rather than all minimal, covers. For further details and proofs regarding the enumeration of minimal covers we refer the reader to Vandal *et al.*'s full paper.

References

Alekseyev, M. A. and Pevzner, P. A. (2009). Breakpoint graphs and ancestral genome reconstructions. *Genome Research*, **19**(5), 943–957.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Avise, J. C. (2015). Evolutionary perspectives on clonal reproduction in vertebrate animals. *PNAS*, **112**(29), 8867–8873.

Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of y-chromosome degeneration. *Nature Reviews Genetics*, **14**, 113–124.

Badouin, H., Hood, M. E., Gouzy, J., Aguileta, G., Siguenza, S., Perlin, M. H., Cuomo, V. O. P. A., Fairhead, C., Branca, A., and Giraud, T. (2015). Chaos of rearrangements in the mating-type chromosomes of the anther-smut fungus microbotryum lychnidis-dioicae. *Genetics*, **200**(4), 1275–1284.

Batada, N. N. and Hurst, L. D. (2007). Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nature Genetics*, **39**, 945–949.

Billiard, S., Lopez-Villavicencio, M., Hood, M. E., and Giraud, T. (2012). Sex, outcrossing and mating types: unsolved questions in fungi and beyond. *Journal of Evolutionary Biology*, **25**, 10201038.

Birky, Jr, C. W. (2010). Positively negative evidence for asexuality. *Journal of Heredity*, **101**, S42–S45.

Brown, C. A., Murray, A. W., and Verstrepen, K. J. (2010). Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology*, **20**(10), 895–903.

Brown, K. M., Burk, L. M., Henagan, L. M., and Noor, M. A. F. (2004). A test of the chromosomal rearrangement model of speciation in drosophila pseudoobscura. *Evolution*, **58**(8), 1856–1860.

Cailleux, R. (1971). Recherches sur la mycoflore coprophile centrafricainee. le genres sordaria, gelasinospora, bombardia. *Bulletin de la Societe Mycologique de France*, **87**, 461626.

Cambareri, E., Jensen, B., Schabtach, E., and Selker, E. (1989). Repeat-induced g-c to a-t mutations in neurospora. *Science*, **244**(4912), 1571–1575.

Caprara (1999). Formulations and hardness of multiple sorting by reversals. *RECOMB '99: Proceedings of the Third Annual International Conference on Computational Molecular Biology*, pages 84–93.

Cerón-Romero, M. A., Nwaka, E., Owoade, Z., and Katz, L. A. (2018). Phylochromomap, a tool for mapping phylogenomic history along chromosomes, reveals the dynamic nature of karyotype evolution in plasmodium falciparum. *Genome Biology and Evolution*, **10**(2), 553–561.

Chernomor, O., von Haeseler, A., and Quang Minh, B. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Syst Biol*, **65**, 997–1008.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Coil, D., Jospin, G., and Darling, A. (2015). A5-miseq: an updated pipeline to assemble microbial genomes from illumina miseq data. *Bioinformatics*, **31**(4), 587–589.

Corcoran, P., Anderson, J. L., Jacobson, D. J., Sun, Y., Ni, P., Lascoux, M., and Johannesson, H. (2016). Introgression maintains the genetic integrity of the sex-determining chromosome of the fungus neurospora tetrasperma. *Genome Research*.

Cornelissen, J. H. C., Callaghan, T. V., Alatalo, J. M., Michelsen, A., Graglia, E., Hartley, A. E., Hik, D. S., Hobbie, S. E., Press, M. C., Robinson, C. H., Henry, G. H. R., Shaver, G. R., Phoenix, G. K., Jones, D. G., Jonasson, S., III, F. S. C., Molau, U., Neill, C., Lee, J. A., Melillo, J. M., Sveinbjrnsson, B., and Aerts, R. (2004). Global change and arctic ecosystems: is lichen decline a function of increases in vascular plant biomass? *Journal of Ecology*, **89**(6), 984–994.

Croll, D. and Sanders, I. R. (2009). Recombination in glomus intraradices, a supposed ancient asexual arbuscular mycorrhizal fungus. *BMC Evolutionary Biology*, **9**, 13.

Cummins, C. A. and McInerney, J. O. (2011). A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology*, **60**(6), 833844.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R., Lunter, G., Marth, G., Sherry, S. T., McVean, G., Durbin, R., and Group, . G. P. A. (2011). The varian call format and vcftools. *Bioinformatics*, **27**(15), 2156–2158.

Darwin, C. (1862). On the two forms, or dimorphic condition, in the species of primula, and on their remarkable sexual relations. *Botanical Journal of the Linnean Society*, **6**(22), 77–96.

Davis, R. H. (2000). *Neurospora: contributions of a model organism.* Oxford University Press, New York, NY.

Davis, R. H. and Perkins, D. D. (2002). Neurospora: a model of model microbes. *Nature Reviews*, **3**, 397–403.

Debortoli, N., Li, X., Eyres, I., Fontaneto, D., Hespeels, B., Q.Tang, C., Flot, J.-F., and Doninck, K. V. (2016). Genetic exchange among bdelloid rotifers is more likely due to horizontal gene transfer than to meiotic sex. *Current Biology*, **26**(6), 723–732.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research*, **27**(11), 2369–2376.

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparision. *Nucleic Acids Research*, **30**(11), 2478–2483.

den Bakker, H. C., VanKuren, N. W., Morton, J. B., and Pawlowska, T. E. (2010). Clonality and recombination in the life history of an asexual arbuscular mycorrhizal fungus. *Molecular Biology and Evolution*, **27**(11), 2474–2486.

Dong, S., Raffaele, S., and Kamoun, S. (2015). The two-speed genomes of filamentous pathogens: waltz with plants. *Current Opinion in Genetics and Development*.

Elias, I. (2006). Settling the intractability of multiple alignment. *Journal of Computational Biology*.

Ellison, C. E., Stajich, J. E., Jacobson, D. J., Natvig, D. O., Lapidus, A., Foster, B., Aerts, A., Riley, R., Lindquist, E. A., Grigoriev, I. V., and Taylor, J. W. (2011a). Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus neurospora tetrasperma. *Genetics*, **189**(1), 55–69.

Ellison, C. E., Hall, C., Kowbel, D., Welch, J., Brem, R. B., Glass, N., and Taylor, J. W. (2011b). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *PNAS*, **108**(7), 2831–2836.

Faino, L., Seidl, M. F., Shi-Kunne, X., Pauper, M., van den Berg, G. C., Wittenberg, A. H., and Thomma, B. P. (2016). Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Research*.

Felsenstein, J. (2009). Phylip (phylogeny inference package) version 3.7a. *Distributed by the Author. Department of Genome Sciences, University of Washington, Seatle*.

Ferguson, B., Dreisbach, T., Parks, C., Filip, G., and Schmitt, C. (2003). Coarse-scale population structure of pathogenic armillaria species in a mixed-conifer forest in the blue mountains of northeast oregon. *Canadian Journal of Forest Research*, **33**, 612–623.

Fraser, J. A. and Heitman, J. (2004). Evolution of fungal sex chromosomes. *Molecular Microbiology*, **51**(2), 299–306.

Galagan, J. E. and Selker, E. U. (2004). Rip: the evolutionary cost of genome defense. *TRENDS in Genetics*, **20**(9), 417–423.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C. B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M. A., Werner-Washburne, M., Selitrennikoff, C. P., Kinsey, J. A., Braun, E. L., Zelter, A., Schulte, U., Kothe, G. O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R. L., Perkins, D. D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R. J., Osmani, S. A., DeSouza, C. P. C., Glass, L., Orbach, M. J., Berglund, J. A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D. O., Alex, L. A., Mannhaupt, G., Ebbole, D. J., Freitag, M., Paulsen, I., Sachs, M. S., Lander, E. S., Nusbaum, C., and Birren, B. (2003). The genome sequence of the filamentous fungus neurospora crassa. *Nature*, **422**, 859–868.

Galazka, J. M., Klocho, A. D., Uesaka, M., Honda, S., Selker, E. U., and Freitag, M. (2016). Neurospora chromosomes are organized by blocks of importin alpha-dependent heterochromatin that are largely independent of h3k9me3. *Genome Research*, **26**(8), 1069–1080.

Garcia, D., Stchigel, A. M., Cano, J., Guarro, J., and Hawksworth, D. L. (2004). A synopsis and re-circumscription of neurospora (syn. gelasinospora) based on ultrastructural and 28s rdna sequence data. *Mycological Research*, **108**(10), 11191142.

Gioti, A., Mushegian, A. A., Strandberg, R., Stajich, J. E., and Johannesson, H. (2012). Unidirectional evolutionary transitions in fungal mating systems and the role of transposable elements. *Molecular Biology and Evolution*, **29**(10), 3215–3226.

Gioti, A., Stajich, J. E., and Johannesson, H. (2013). Neurospora and the dead-end hypothesis: Genomic consequences of selfing in the model genus. *Evolution*, **67**(12), 3600–3616.

Gladieux, P., Wilson, B. A., Perraudeau, F., Montoya, L. A., Kowbel, D., HannSoden, C., Fischer, M., Sylvain, I., Jacobson, D. J., and Taylor, J. W. (2015). Genomic sequencing reveals historical, demographic and selective factors associated with the diversification of the fireassociated fungus neurospora discreta. *Molecular Ecology*, **24**(22), 5657–5675.

Gladieux, P., Bellis, F. D., Hann-Soden, C., Svedberg, J., Johannesson, H., and Taylor, J. W. (in prep). Neurospora from natural populations: Population genomics insights into the life history of a model microbial eukaryote.

Gladyshev, E. A., Meselson, M., and Arkhipova, I. R. (2008). Massive horizontal gene transfer in bdelloid rotifers. *Science*, **320**(5880), 1210–1213.

Glass, N. (2018). personal communication.

Glass, N., Metzenberg, R. L., and Raju, N. B. (1990). Homothallic sordariaceae from nature: The absence of strains containing only the a mating type sequence. *Experimental Mycology*, **14**, 274–289.

Goddard, M. R., Godfray, H. C. J., and Burt, A. (2005). Sex increases the efficacy of natural slection in experimental yeast populations. *Nature*, **434**(7033), 636.

Gupta, U., Lee, D., and Leung, J.-T. (1982). Efficient algorithms for interval graphs and circular-arc graphs.

Hachiya, T., Osana, Y., Popendorf, K., and Sakakibara, Y. (2009). Accurate identification of orthologous segments among multiple genomes. *Bioinformatics*, **25**, 853–860.

Heller, J., Zhao, J., Rosenfield, G., Kowbel, D. J., Gladieux, P., and Glass, N. (2016). Characterization of greenbeard genes involved in long-distance kind discrimination in a microbial eukaryote. *PLOS Biology*, **14**(4), e1002431.

Herrera, J., Poudel, R., and Khidir, H. H. (2010). Molecular characterization of coprophilous fungal communities reveals sequences related to root-associated fungal endophytes. *Microbial Ecology*, **61**(2), 239–244.

Hoffmann, A. A. and Rieseberg, L. H. (2008). Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution, and Systematics*, **39**, 21–42.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, **33**(6), 1635–1638.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, **9**, 90–95.

Igic, B. and Busch, J. W. (2013). Is self-fertilization an evolutionary dead end? *New Phytologist*, **198**(2), 1143–1150.

Jacobson, D. J. (2005). Blocked recombination along the mating-type chromosomes of neurospora tetrasperma involves both structural heterozygosity and autosomal genes. *Genetics*, **171**(2), 839–843.

Jacobson, D. J., Powell, A. J., Dettman, J. R., Saenz, G. S., Barton, M. M., Hiltz, M. D., William H. Dvorachek, J., Glass, N., Taylor, J. W., and Natvig, D. O. (2004). Neurospora in temperate forests of western north america. *Mycologia*, **96**(1), 66–74.

Jany, J. and Pawlowska, T. E. (2010). Multinucleate spores contribute to evolutionary longevity of asexual glomeromycota. *The American Naturalist*, **175**(4), 424–435.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, **8**, 275–282.

Jones, E., Oliphant, T., Peterson, P., *et al.* (2001–). SciPy: Open source scientific tools for Python. [Online; accessed Oct. 25 2017].

Kalyaanamoorthy, S., Quang Minh, B., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). Modelfinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, **14**, 587–589.

Karp, R. M. (1972). Reducability among combinatorial problems. *Complexity of Computer Computations*, pages 85–103.

Kasuga, T. and Glass, N. (2008). Dissecting colony development of neurospora crassa using mrna profiling and comparative genomics approaches. *Eukaryotic Cell*, page 15491564.

Kasuga, T., Mannhaupt, G., and Glass, N. (2009). Relationship between phylogenetic distribution and genomic features in neurospora crassa. *PLoS ONE*, **4**(4), e5286.

Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

Khidir, H., Eudy, D., Porras-Alfaro, A., Herrera, J., Natvig, D., and Sinsabaugh, R. (2010). A general suite of fungal endophytes dominate the roots of two dominant grasses in a semiarid grassland. *Journal of Arid Environments*, **74**, 35–42.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS Biology*, **8**(9), e1000501.

Kuo, H.-C., Hui, S., Choi, J., Asiegbu, F. O., Valkonen, J. P. T., and Lee, Y.-H. (2014). Secret lifestyles of neurospora crassa. *Scientific Reports*, **4**, 5135.

Kurtz, S. (1999). Reducing the space requirement of suffix trees. *Software-Practice and Experience*, **29**(13), 1149–1171.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, **5**, R12.

Lahn, B. T. and Page, D. C. (1999). Four evolutionary strata on the human x chromosome. *Science*, **286**(5441), 964–967.

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

Lewis, Z. A., Honda, S., Khlafallah, T. K., Jeffress, J. K., Freitag, M., Mohn, F., Schubeler, D., and Selker, E. U. (2009). Relics of repeat-induced point mutaiton in direct heterochromatin formation in neurospora crassa. *Genome Research*.

Linardopoulou, E. V., Williams, E. M., Fan, Y., Friedman, C., Young, J. M., and Trask, B. J. (2005). Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature*, **437**, 94–100.

Loro, M., Valero-Jimnez, C. A., Nozawa, S., and Mrquez, L. M. (2012). Diversity and composition of fungal endophytes in semiarid northwest venezuela. *Journal of Arid Environments*, **85**, 46–55.

Madden, T. (2002). *The BLAST Sequence Analysis Tool*, chapter 16. National Center for Biotechnology Information (US), Bethesda, MD.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, **20**, 1297–1303.

McKinney, W. (2010). Data structures for statistical computing in python. In *9th Python in Science Conference*.

Menkis, A., Jacobson, D. J., Gustafsson, T., and Johannesson, H. (2008). The mating-type chromosome in the filamentous ascomycete neurospora tetrasperma represents a model for early evolution of sex chromosomes. *PLOS Genetics*, **4**(3), e1000030.

Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., Gould, K., Mead, D., Drury, E., O'Brien, J., Ruano Rubio, V., MacInnis, B., Mwangi, J., Samarakoon, U., Ranford-Cartwright, L., Ferdig, M., Hayton, K., zhuan Su, X., Wellems, T., Rayner, J., McVean, G., and Kwiatkowski, D. (2016). Indels, structural variation, and recombination drive genomic diversity in plasmodium falciparum. *Genome Research*, **26**, 1288–1299.

Mrqueza, S. S., F.Bills, G., Herreroa, N., and igo Zabalgogeazcoa (2012). Non-systemic fungal endophytes of grasses. *Fungal Ecology*, **5**(3), 289–297.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Quang Minh, B. (2015). Iq-tree: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol Biol Evol*, **32**, 268–274.

Nowrousian, M., Teichert, I., Masloff, S., and Kck, U. (2010). Whole-genome sequencing of sordaria macrospora mutants identifies developmental genes. *G3*, **2**(2), 261–270.

Nygren, K., Strandberg, R., Wallberg, A., Nabholz, B., Gustafsson, T., Garca, D., Cano, J., Guarro, J., and Johannesson, H. (2011). A comprehensive phylogeny of neurospora reveals a link between reproductive mode and molecular evolution in fungi. *Molecular Phylogenetics and Evolution*, **59**, 649–663.

Oliphant, T. E. (2006). *A guide to NumPy*. Trelgol Publishing, USA.

Otto, S. P. (2009). The evolutionary enigma of sex. *The American Naturalist*, pages S1–S14.

Palma-Guerrero, J., Hall, C. R., Kowbel, D., Welch, J., Taylor, J. W., Brem, R. B., and Glass, N. (2013). Genome wide association identifies novel loci involved in fungal communication. *PLOS Genetics*, **9**(8), e1003669.

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., and Haussler, D. (2011). Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, **21**(9), 1512–1528.

Pawlowska, T. E. and Taylor, J. W. (2004). Organization and genetic variation in individuals of arbuscular mycorrhizal fungi. *Nature*, **427**(6976), 733.

Perkins, D. D. and Turner, B. C. (1988). Neurospora from natural populations: Toward the population biology of a haploid eukaryote. *Experimental Mycology*, **12**(2), 91–131.

Popendorf, K., Hachiya, T., Osana, Y., and Sakakibara, Y. (2010). Murasaki: A fast, parallelizable algorithm to find anchors from multiple genomes. *PLoS ONE*, **5**(9), e12651.

Powell, A. J., Jacobson, D. J., and Natvig, D. O. (2001). Allelic diversity at the het-c locus in neurospora tetrasperma confirms outcrossing in nature and reveals an evolutionary dilemma for pseudohomothallic ascomycetes. *Journal of Molecular Evolution*, **52**(1), 94–102.

Powell, A. J., Jacobson, D. J., Salter, L., and Natvig, D. O. (2003). Variation among natural isolates of neurospora on small spatial scales. *Mycologia*, **95**(5), 809–819.

Qi, F., Jing, T., and Zhan, Y. (2012). Characterization of endophytic fungi from acer ginnala maxim. in an artificial plantation: media effect and tissue-dependent variation. *PLoS ONE*, **7**, e46785.

Raju, N. B. (1992). Functional heterothallism resulting from homokaryotic conidia and ascospores in neurospora tetrasperma. *Mycological Research*, **96**(2), 103–116.

Ramsbottom, J. and Stephens, F. (1935). Neurospora in britain. *Transactions of the British Mycological Society*, **19**(3), 215–220.

Rocha, E. P. and Danchin, A. (2003a). Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nature Genetics*, **34**, 377–378.

Rocha, E. P. and Danchin, A. (2003b). Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Research*, **31**(22), 6570–6577.

Rogers, J. and Gibbs, R. A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature Reviews Genetics*, **15**, 347–359.

Ropars, J., Toro, K. S., Noel, J., Pelin, A., Charron, P., Farinelli, L., Marton, T., Krger, M., Fuchs, J., Brachmann, A., and Corradi, N. (2016). Evidence for the sexual origin of heterokaryosis in arbuscular mycorrhizal fungi. *Nature Microbiology*, **1**, 16033.

Rota, J., Malm, T., and Wahlberg, N. (2017). A simple method for data partitioning based on relative evolutionary rates. *PeerJ Preprints*, **5**, e3414v1.

Sankoff, D. and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, **5**(3), 555–570.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Shaw, D. E. (1993). Honeybees collecting neurospora spores from steamed pinus logs in queensland. *Mycologist*, **7**(4), 182–185.

Shiu, P. K. T., Raju, N. B., Zickler, D., and Metzenberg, R. L. (2001). Meiotic silencing by unpaired dna. *Cell*, **107**(7), 905–916.

Signorovitch, A., Hur, J., Gladyshev, E., and Meselson, M. (2015). Allele sharing and evidence for sexuality in a mitochondrial clade of bdelloid rotifers. *Genetics*, **200**, 581–590.

Slater, G. S. C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**(31).

Smith, J. M. (1971). What use is sex? *Journal of Theoretical Biology*, **30**(2), 319–335.

Smith, S. E. and Read, D. (2008). *Mycorrhizal Symbiosis*. Academic Press, New York, NY, 3 edition.

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, **16**, 472–482.

Stajich, J. E., Berbee, M. L., Blackwell, M., Hibbett, D. S., James, T. Y., Spatafora, J. W., , and Taylor, J. W. (2009). The fungi. *Current Biology*, **19**(18), R840–R845.

Stanke, M. (2003). *Gene Prediction with a Hidden Markov Model*. Ph.D. thesis, Universitt Gttingen.

Stanke, M., Schffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden markov model that uses hints from external sources. *BMC Bioinformatics*, **7**(62).

Steenwyk, J. L. and Rokas, A. (2018). Copy number variation in fungi and its implications for wine yeast genetic diversity and adaptation. *Frontiers in Microbiology*, **9**, 288.

Subramanian, S. (2016). The effects of sample size on population genomic analyses–implications for the tests of neutrality. *BMC Genomics*, **17**, 123.

Sun, Y., Corcoran, P., Menkis, A., Whittle, C. A., Andersson, S. G. E., and Johannesson, H. (2012). Large-scale introgression shapes the evolution of the mating-type chromosomes of the filamentous ascomycete neurospora tetrasperma. *PLOS Genetics*, **8**(7), e1002820.

Sun, Y., Svedberg, J., Hiltunen, M., Corcoran, P., and Johannesson, H. (2017). Large-scale suppression of recombination predates genomic rearrangements in neurospora tetrasperma. *Nature Communications*, **8**, 1140.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, **123**(3), 585–595.

Tajima, F. and Nei, M. (1984). Estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, **1**(3), 269–285.

Taylor, J. W. (2011). One fungus= one name: Dna and fungal nomenclature twenty years after pcr. *IMA fungus*, **2**(2), 113–120.

Taylor, J. W., Hann-Soden, C., Branco, S., Sylvain, I., and Ellison, C. E. (2015). Clonal reproduction in fungi. *PNAS*, **112**(29), 8901–8908.

Tesler, G. (2002a). Efficient algorithms for multichromosomal genome rearrangements. *Journal of Computer and System Sciences*, **65**(3), 587–609.

Tesler, G. (2002b). Grimm: genome rearrangements web server. *Bioinformatics*, **18**(3), 492–493.

The Broad Institute (2009). Picard. `https://broadinstitute.github.io/picard/`.

Thi Hoang, D., Chernomor, O., von Haeseler, A., Quang Minh, B., and Sy Vinh, L. (2018). Ufboot2: Improving the ultrafast bootstrap approximation. *Mol Biol Evol*, **35**(2), 518–522.

Thompson, M. and Jiggins, C. (2014). Supergenes and tehir role in evolution. *Heredity*, **113**, 1–8.

Tian, C., Beeson, W. T., Iavarone, A. T., Sun, J., Marletta, M. A., Cate, J. H. D., and Glass, N. (2009). Systems analysis of plant cell wall degradation by the model filamentous fungus neurospora crassa. *PNAS*, pages 1–6.

Tigano, A. and Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, **25**, 2144–2164.

Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE*, **7**(9), e42304.

Turner, B. C. (1987). Two ecotypes of neurospora intermedia. *Mycologia*, **79**(3), 425–432.

Turner, B. C., Perkins, D. D., and Fairfield, A. (2001). Neurospora from natural populations: A global study. *Fungal Genetics and Biology*, **32**, 67–92.

U'Ren, J. M., Lutzoni, F., Miadlikowska, J., Laetsch, A. D., and Arnold, A. E. (2012). Host and geographic structure of endophytic and endolichenic fungi at a continental scale. *American Journal of Botany*, **99**(5), 898–914.

Vandal, A. C., Conder, M. D., and Gentleman, R. (1997). Minimal covers of maximal antichains for interval orders.

Wang, J., Na, J.-K., Yu, Q., Gschwend, A. R., Han, J., Zeng, F., Aryal, R., VanBuren, R., Murray, J. E., Zhang, W., Navajas-Prez, R., Feltus, F., Lemke, C., Tong, E. J., Chen, C., Man Wai, C., Singh, R., Wang, M.-L., Jia Min, X., Alam, M., Charlesworth, D., Moore, P. H., Jiang, J., Paterson, A. H., , and Ming, R. (2012). Sequencing papaya x and y^h chromosomes reveals molecular basis of incipient sex chromosome evolution. *PNAS*, **109**(34), 13710–13715.

W.M. Lewis, J. (1987). *The cost of sex*, pages 33–58. Springer Basel AG, Basel.

Zhao, J., Gladieux, P., Hutchison, E., Bueche, J., Hall, C., Perraudeau, F., and Glass, N. (2015). Identification of allorecognition loci in neurospora crassa by genomics and evolutionary approaches. *Molecular Biology and Evolution*, **32**(9), 2417–2432.

Zheng, C. and Sankoff, D. (2016). Locating rearrangement events in a phylogeny based on highly fragmented assemblies. *BMC Genomics*, **17**(Suppl 1), 1–5.

Appendix A

Tables of Strains

A.1 Strain Names

New strains collected by the authors and presented here are identified with three part names. The first part of strain names is an acronym of the place name or geographic region from which they were collected; the second part of the name is a number identifying the soil sample from which the strain was isolated; and the third part of the name is a number identifying the particular isolation. For example, strain "LNF6-4" was the 4th attempted isolation from the 6th soil sample collected in the area of the Lower North Fork wildfire in Colorado. For strains without an area acronym, such as "4A", the number represents the soil sample, and the letter indicates which single spore was kept from the soil germination. Soil samples 1-10 were collected in Pt. Reyes, California, while sample 11 was collected from Truckee, California.

LNF	Lower North Fork
MI	Michigan
TR	Truckee River

Table A.1: Expansion of acronyms used in strain names.

A.2 Strains Used

Table A.2: This table is available as a CSV file from `https://osf.io/f3vuy/`.