Title

Evolutionary genomics of divergence and adaptation within the model fungi Neurospora crassa and Neurospora tetrasperma

Permalink

Author

Ellison, Christopher

Publication Date

2011

Evolutionary genomics of divergence and adaptation within the model fungi *Neurospora crassa* and *Neurospora tetrasperma*

By

Christopher Eugene Ellison

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 Rachel Brem
Professor N. Louise Glass

Spring 2011

Evolutionary genomics of divergence and adaptation within the model fungi *Neurospora crassa* and *Neurospora tetrasperma*

Abstract

Evolutionary genomics of divergence and adaptation within the model fungi *Neurospora crassa* and *Neurospora tetrasperma*

By

Christopher Eugene Ellison

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor John W. Taylor, Chair

The work presented in this dissertation involves the interpretation of patterns of DNA and gene expression variation to infer ways in which evolutionary events and processes have shaped the life histories of species of filamentous fungi within the genus *Neurospora*.

The first chapter involves a comparison between genome sequences from both mating types of the self-fertile fungus *Neurospora tetrasperma*. Self-fertility in this species is associated with a large, sex-linked region of suppressed recombination. The structural rearrangements involved in the suppression of recombination are identified here for the first time and a model for the evolution of this non-recombining chromosomal region is developed. Additionally, evidence is presented that suggests the *mat a*-linked region of suppressed recombination is accumulating deleterious alleles at a faster rate than the *mat A*-linked region and thus may be in the early stages of degeneration.

Discovering the genetic basis behind adaptive phenotypes has long been considered the holy grail of evolutionary genetics, yet most instances where adaptive alleles have been identified involved targeting candidate genes based on their having a function related to an obvious phenotype such as pigmentation. This forward-ecology approach is difficult for most fungi because they lack obvious phenotypes. In the second chapter of this dissertation, a reverse-ecology approach is utilized to identify candidate genes involved in local adaptation to cold temperature in two recently diverged populations of *Neurospora crassa*. High-resolution genome scans between populations were performed to identify genomic islands of extreme divergence. Two such islands were identified and found to contain genes whose functions, pattern of nucleotide polymorphism, and null phenotype are consistent with local adaptation.

The third chapter extends the DNA-based analyses presented in the second chapter to explore natural variation in gene expression within and between these same *N. crassa* populations. Intrapopulation variation in gene expression is harnessed to identify regulatory modules and to provide a tool for inferring functional information for unannotated genes. Divergence in regulatory networks between populations is assessed, providing insight into potential functional differences between these populations.

<u>**Table of Contents**</u>

Acknowledgements

First and foremost, I would like to thank my advisor John Taylor. None of this would have been possible without the valuable opportunities, insight, feedback and support that he provided. I am very lucky to have had, in addition to John, two other fantastic committee members: Rachel Brem and Louise Glass. Their expertise in regulatory variation and genetics, their support, ideas and suggestions all played a crucial part in the experiments, analyses, and writing presented here. Many other individuals have contributed in countless ways to this work, but the two that top the list are Jason Stajich and Tom Sharpton. I thank them for introducing me to the world of computational biology, the huge amount of time they spent teaching me about genomic analysis, UNIX, and Perl, and their invaluable help in troubleshooting and maintaining our seemingly cursed lab server. I have also benefited from the insight and advice of numerous great mycologists including Tom Bruns, David Jacobson, Namboori Raju, and Donald Natvig. The analyses presented in the first chapter were possible thanks to the skills of several people from the Joint Genome Institute including Alla Lapidus, Brian Foster, Andrea Aerts, Robert Riley, Erika Lindquist, and Igor Grigoriev, and their work sequencing, assembling, and annotating the *Neurospora tetrasperma* genomes. Throughout these past five years, the Glass Lab is the first place I have gone for questions about reagents, protocols, strains, and all other things *Neurospora* related, and I especially want to thank Betsy Hutchinson, Charles Hall, Juliet Welch, and David Kowbel for their help in that regard. Members of the Brem Lab have also been a great resource: OhKyu Yoon provided much appreciated advice regarding Illumina library prep and sequencing protocols, Tiffany Hsu prepared the Illumina libraries for the deletion strains used in the third chapter, and, as a rotation student, Angela Kaczmarczyk performed the first investigation into gene expression differences between *Neurospora* populations, paving the way for the analyses presented here. Illumina sequencing was performed at the QB3 Vincent J. Coates Genomics Sequencing Laboratory under the guidance of Leath Tonkin. The *Neurospora* strains used in the experiments described here were provided by the Fungal Genetics Stock Center. Support for this research was provided in part by the Chang-Lin Tien Graduate Fellowship. Tim Szaro is the engine that keeps the Bruns and Taylor labs running smoothly and the same is true for the PMB support staff vis-à-vis the entire department, especially Nikki Groen and the Graduate Student Affairs Officers Rocio Sanchez, Dana Jantz, and Kyle Dukart.

Last but certainly not least, I want to thank all those whose friendship provided the good times that served as a crucial distraction from the daily grind of research, especially Devin O'Connor, Hayley Oggel, Jeremy Wilbur, and Elaine Shapland, the PMB cohort of 2006, especially Adrien Burch and Solomon Stonebloom, and the entire Bruns-Taylor complex of students and Post-docs, past and present, whom I had the pleasure of overlapping with and celebrating many an occasion with. In addition, I am forever indebted to my wonderful parents, sister, and extended family for everything they have done for me and continue to do, and most importantly, to my incredible fiancée Kate, whose drive, accomplishments, enthusiasm, and all-around joyful outlook on life serve as a daily inspiration.

Introduction

The work presented in this dissertation involves the interpretation of patterns of DNA and gene expression variation to infer ways in which evolutionary events and processes have shaped the life histories of species of filamentous fungi within the genus *Neurospora*. The analyses described within these chapters parallel the development and take advantage of the sequence-based methods of evolutionary inference described below.

Charles Darwin published *On the Origin of Species* in 1859. An important piece missing from Darwin's work, however, was the mechanism behind his theory of "descent with modification". In an unfortunate twist of fate, Gregor Mendel published the results of his famous experiments with pea plants only six years after Darwin published the *Origin* but the importance of Mendel's work in establishing the existence of genetic traits, such as flower color and seed shape, that are inherited as discrete units, remained virtually unknown for another thirty-four years (Hartl and Orel 1992).

The rediscovery of Mendel's work in the early 1900's eventually led to the re-framing of Darwin's views in light of Mendel's evidence of heritability of discrete traits and with respect to the contemporary views that genes were heritable and that mutations in genes provide the raw variation that natural selection acts upon. This reconciliation is known as the Modern Synthesis and, because it allowed for the use of mathematical theory to predict the expected frequencies of mutations within a population and the contribution of mutations to an organism's fitness, led to the birth of the fields of population genetics and quantitative genetics in the 1930's (Charlesworth and Charlesworth 2009).

A string of landmark studies over the next several decades implicated DNA as the mechanism of inheritance (Avery et al. 1944), solved the structure of the molecule and the problem of how it might replicate (Watson and Crick 1953), and cracked the genetic code (Matthaei et al. 1962). Several years later, in 1968, Motoo Kimura published his theory of neutral evolution, claiming that most mutations are neutral and their frequency within a population is stochastic and controlled by genetic drift (Kimura 1968). This idea was contrary to the prevailing belief at the time that most, if not all, mutations in a population are under selection and represented a major paradigm shift that continues to be influential today (Nei et al. 2010).

During the early 1970's, the theoretical basis for how mutations behave in a population and they ways in which they interact to affect an organism's fitness were well established. What was missing, however, was the ability to directly interrogate the DNA itself in order to observe mutations first-hand. This breakthrough occurred in 1977 with the publication of the first complete genome sequence: the 5,375 bp genome of bacteriophage ΦX174 (Sanger et al. 1977a), and the advent of Sanger sequencing (Sanger et al. 1977b). The ability to determine the sequence of basepairs within a molecule of DNA revolutionized the entire field of biology and resulted in the development of sequence-based methods of evolutionary inference.

The first of these methods was originally designed as a way to measure the rate of evolution of a gene while accounting for the fact that synonymous sites, or those sites within a codon that are able to change without changing the resulting amino acid, will evolve at a faster rate than those sites that, if mutated, code for a different amino acid (nonsynonymous sites). These two rates are respectively referred to as *dS* and *dN*, or alternatively, *Ks* and *Ka*. The first study to infer evidence of positive selection by employing a formal statistical test in the comparison of these rates was performed in 1988 (Hughes and Nei 1988). The authors compared *dS* and *dN* between alleles of class I MHC genes within humans as well as mice and found that *dN* was significantly greater than *dS* in the region of the gene containing the antigen recognition site, consistent with overdominant selection (Hughes and Nei 1988). This test was appropriate and useful given the data at the time because it only required a pairwise comparison of DNA sequence from a single gene.

Several years after this test was conceptualized, McDonald and Kreitman developed a similar test, now known as the MK test, which compared the number of fixed differences in synonymous and nonsynonymous sites between species to the number of those that were variable within populations of the same species (McDonald and Kreitman 1991). In this case, a gene exhibiting an excess of nonsynonymous to synonymous fixed differences, relative to the amount of nonsynonymous to synonymous polymorphisms, was inferred to be under positive selection (McDonald and Kreitman 1991). In addition to the sequence data from a single gene from two different species, this test also required gene sequences from multiple individuals from the same population.

Advances in technology and approaches for assembling Sanger reads into chromosome-sized scaffolds over the next decade culminated in the sequencing of the human genome in 2001 (Lander et al. 2001; Venter et al. 2001). Shortly thereafter, the development of array-based genotyping methods, and more recently, next generation sequencing technology, has allowed for the examination of genome-wide variation across populations of humans and other organisms (Luikart et al. 2003; Shendure and Ji 2008). This large increase in the size of DNA sequence datasets allowed for a suite of entirely new approaches to evolutionary inference from DNA sequence (reviewed in (Nielsen 2005)). These approaches were largely based on examining variation in allele frequencies across the genome, the patterns of which can lead to inferences regarding the demographic history of a population as well as the genes that may have been recent targets of natural selection. Many of the methods used to identify the signature of positive selection in these datasets focus on the signature of a selective sweep. In this phenomenon, positive selection drives a beneficial mutation to fixation while dragging along a linked subset of neighboring variants with it, resulting in a localized genomic region of reduced nucleotide variability (Smith and Haigh 1974).

Concomitant with the development of these population genomic analyses, new methods for interpreting the patterns of gene expression variation within and between populations were also created. These methods include the ability to map the genetic architecture underlying variation in transcript levels across the genome (Brem et al. 2002) and to assess the influence of adaptive evolution with respect to differences in gene expression levels between lineages (Denver et al. 2005; Rifkin et al. 2005; Whitehead and Crawford 2006).

In the first chapter of this dissertation, analysis of codon usage and rates of evolution in the form of *dN* and *dS* are used to compare genome sequences from opposite mating-types of the self-fertile fungus *N. tetrasperma* and its close outbreeding relative, *N. crassa*. These comparisons allow insight into the evolutionary events that occurred during *N. tetrasperma*'s transition to self-fertility and into the evolutionary consequences of suppressed recombination. In the second chapter, population genomic analyses are employed to survey genome-wide DNA polymorphism and divergence within and between two recently diverged populations of *N. crassa*. These comparisons identify two genomic regions showing extremely large divergence between as well as reduced nucleotide variation within populations, the latter being suggestive of a selective sweep. The functions and null phenotype of two of the genes within these regions are consistent with local adaptation to low temperature. The third chapter extends the DNA based analyses between these *N. crassa* populations to gene expression variation. Genes that are differentially expressed between populations and novel regulons within populations are identified. Together, these chapters serve to increase the understanding of the evolution of this group of fungi, but also use *Neurospora* as a convenient model to elucidate general principles of genome evolution and adaptation that are relevant to all of life.

References

Avery OT, Macleod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. Journal of Experimental Medicine 79(2):137-58.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296(5568):752-5.

Charlesworth B, Charlesworth D. 2009. Darwin and genetics. Genetics 183(3):757-66.

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nature Genetics 37(5):544-8.

Hartl DL, Orel V. 1992. What did Gregor Mendel think he discovered? Genetics 131(2):245-53.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335(6186):167-70.

Kimura M. 1968. Evolutionary rate at the molecular level. Nature 217(5129):624-624.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nature Reviews Genetics 4(12):981-94.

Matthaei JH, Jones OW, Martin RG, Nirenberg MW. 1962. Characteristics and composition of RNA coding units. Proceedings of the National Academy of Sciences of the United States of America 48:666-77.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351(6328):652-4.

Nei M, Suzuki Y, Nozawa M. 2010. The neutral theory of molecular evolution in the genomic era. Annual Review of Genomics and Human Genetics 11:265-89.

Nielsen R. 2005. Molecular signatures of natural selection. Annual Review of Genetics 39:197-218.

Rifkin SA, Houle D, Kim J, White KP. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature 438(7065):220-3.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265(5596):687-95.

Sanger F, Nicklen S, Coulson AR. 1977b. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74(12):5463-7.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nature Biotechnology 26(10):1135-45.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23(1):23-35.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. Science 291(5507):1304-51.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171(4356):737-8.

Whitehead A, Crawford DL. 2006. Neutral and adaptive variation in gene expression. Proceedings of the National Academy of Sciences of the United States of America 103(14):5425-5430.

Chapter 1

Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus *Neurospora tetrasperma*

Abstract

A large region of suppressed recombination surrounds the sex-determining locus of the self-fertile fungus *Neurospora tetrasperma*. This region encompasses nearly one-fifth of the *N. tetrasperma* genome and the suppression of recombination is necessary for the maintenance of the self-fertile condition. The similarity of this sex-linked non-recombining segment to plant and animal sex chromosomes and its recent origin (<5 MYA), combined with a long history of genetic and cytological research, makes *N. tetrasperma* an ideal model for studying the evolutionary consequences of suppressed recombination. Here we compare genome sequences from two *N. tetrasperma* strains of opposite mating-type to determine if structural rearrangements are associated with the non-recombining region and to examine the effect of suppressed recombination for the evolution of the genes within it. We find a series of three inversions encompassing the majority of the region of suppressed recombination and provide evidence for two different types of rearrangement mechanisms: the recently proposed mechanism of inversion via staggered single strand breaks as well as ectopic recombination between transposable elements. In addition, we show that the *N. tetrasperma mat a* mating-type region appears to be accumulating deleterious substitutions at a faster rate than the other mating-type (*mat A*) and thus may be in the early stages of degeneration.

Introduction

The elimination of recombination can have a dramatic effect on the evolutionary trajectory of a genomic region. Without recombination, selection acts on linked genetic complexes rather than independent genetic elements. Theory predicts the accumulation of deleterious mutations and selfish genetic elements in the absence of recombinational purging and a reduced ability to fix adaptive mutations (Charlesworth and Charlesworth 2000; Charlesworth et al. 2005).

The genetic consequences of suppressed recombination have been best studied in the sex chromosomes of outcrossing eukaryotes, e.g., plants, insects and mammals, because the initial step in the formation of sex chromosomes is posited to be a cessation of recombination across a genomic region that includes the sex-determining locus (Charlesworth et al. 2005). The suppression of recombination across such a region will be selected for if it creates linkage between the sex-determining locus and other genes that are sexually antagonistic in that their functions are beneficial to only one of the sexes. The non-recombining region can be formed from the spread of recombinational suppressors or structural changes to the chromosome that prevent synapsis. In either case, studies in mammals, birds, and plants have shown evidence that present-day non-recombining regions are composed of multiple discrete blockage events that occurred at different time-points in the evolutionary history of the taxon in question (termed "evolutionary strata" by Lahn & Page (Charlesworth et al. 2005; Lahn and Page 1999).

Because large, non-recombining, sex-linked regions appear to have evolved independently across a diverse array of taxonomic groups, comparing the evolution of such regions across disparate taxa will make it possible to understand the evolutionary events associated with their formation as well as the basic genomic consequences of suppressed recombination.

In several species of fungi, the properties of the chromosomal region surrounding the mating-type locus have been studied extensively because of their similarities to the sex chromosomes of other organisms (Fraser and Heitman 2004; Fraser and Heitman 2005). The mating-type locus of *Cryptococcus neoformans* occurs within a ~100 kb region where recombination is suppressed due to multiple chromosomal rearrangements (Lengeler et al. 2002). The evolutionary history of this region includes the accumulation of transposable elements as well as gene conversion, gene loss and pseudogenization (Fraser et al. 2004; Metin et al. 2010). The mating-type region of *Ustilago hordei* resides within a 500 kb region that is characterized by suppressed recombination and chromosomal rearrangements (Lee et al. 1999) and the mating-type chromosomes of the fungus *Microbotryum violaceum* are heteromorphic and contain a region of suppressed recombination that has been roughly estimated to be 1,000 kb in size (Votintseva and Filatov 2009). DNA polymorphism within *M. violaceum* populations was assessed for evidence of genetic degeneration within the region of suppressed recombination, however, nucleotide variation within the non-recombining region did not stand out from the rest of the genome due to overall low levels of polymorphism and high linkage-disequilibrium (Votintseva and Filatov 2010), consistent with previous results showing that *M. violaceum* is predominantly selfing (Giraud et al. 2005).

The sex-determining locus of *Neurospora tetrasperma* (termed *mat*) is surrounded by a region of suppressed recombination that is approximately seven-fold larger than that of *M. violaceum* and seventy-fold larger than that of *C. neoformans* (Menkis et al. 2008). From the work presented here, we now know that it includes ca. 2000 genes and spans a distance of ca. 7.8

Mb, which represents 80% of the mating-type chromosome and approximately one-fifth of the *N. tetrasperma* genome. The formation of the region of suppressed recombination was part of a series of evolutionary events that allowed this species to become self-fertile and it is relatively young compared to most vertebrate sex chromosomes (less than ~4.5 MYA compared to ~160 MYA for the XY system in marsupial and placental mammals (Veyrunes et al. 2008) and the ZW system in snakes (O'Meally et al. 2010) and >120 MYA for the ZW system in birds (Mank and Ellegren 2007)). Due to the large size and recent origin of the non-recombining region, *N. tetrasperma* has emerged as an important model for the study of early sex chromosome evolution (Gallegos et al. 2000; Jacobson 2005; Menkis et al. 2008; Menkis et al. 2010; Merino et al. 1996) alongside other organisms with relatively young sex chromosomes such as medaka (Kondo et al. 2006), papaya (Liu et al. 2004), and several species of *Drosophila* (Charlesworth et al. 2005), (reviewed in (Fraser and Heitman 2005)).

Sex in *N. tetrasperma* is controlled by two mating type idiomorphs, *mat a* and *mat A*. The majority of isolates collected from nature are heterokaryotic: haploid nuclei of opposite mating type can be found in a single fungal individual, allowing the individual to be self-fertile (Raju 1992). Self-fertility is maintained between generations by the packaging of two nuclei of opposite mating type into a single sexual spore, however, if recombination occurs in the region between the mating-type locus and the centromere during meiosis, two nuclei of the same, rather than opposite, mating-type will be packaged into the spore (Gallegos et al. 2000; Merino et al. 1996; Raju and Perkins 1994). In this case, the resulting progeny will not be self-fertile. The region of suppressed recombination is therefore thought to have evolved to ensure the correct packaging of nuclei of opposite mating-type into the sexual spore, thus maintaining the heterokaryotic, self-fertile condition (Gallegos et al. 2000; Merino et al. 1996). Nevertheless, the heterokaryon does occasionally break down via the production of vegetative or sexual spores containing nuclei with only one of the two mating types (Raju 1992). The haploid, homokaryotic individual that grows from such a spore can mate with another homokaryotic individual of opposite mating type to restore the heterokaryotic condition (Raju 1992). Thus, the reproductive strategy of *N. tetrasperma* is likely to include repeated rounds of selfing with an occasional outcrossing event (Powell et al. 2001). Repeated selfing within a heterokaryon is also supported by previous work showing that allelic differences between *mat a* and *mat A* nuclei were confined to the non-recombining region of the mating chromosome (Merino et al. 1996).

A considerable body of work has built upon the initial observations of Howe and Haysman (Howe and Haysman 1966) that recombination was reduced on the *N. tetrasperma* mating-type chromosome. Raju (Raju 1992) and Raju and Perkins (Raju and Perkins 1994) determined the steps during ascus development that are required for correct packaging of *mat a* and *mat A* nuclei into a single ascospore. Merino *et al* (Merino et al. 1996) and Gallegos *et al* (Gallegos et al. 2000) confirmed that recombination is suppressed across the majority of the mating type chromosome and Gallegos *et al* visualized the non-recombining region as an anomalous unpaired region visible during pachytene. Jacobson (Jacobson 2005) reciprocally introgressed mating type chromosomes between *N. tetrasperma* and *N. crassa*. His results suggested that the *N. tetrasperma mat a* chromosome may be collinear with the *N. crassa mat A* chromosome while the *N. tetrasperma mat A* chromosome may be structurally rearranged. However, he also found evidence for the existence of genetic, rather than structural, modifiers of recombination and was unable to determine the relative contributions of these two phenomena with respect to the suppression of recombination in this region. More recently, Menkis *et al* (Menkis et al. 2008) sequenced 35 genes spanning the mating type chromosome from each of the two *N. tetrasperma mat a* and *mat A* strains used in this study. Based on sequence divergence between these genes, they predicted the existence of two evolutionary strata: one large pericentric region encompassing most of the chromosome, and another much smaller region distal to the *mat* locus. In two other studies, Menkis *et al* (Menkis et al. 2009) showed that *N. tetrasperma* is actually a species complex composed of nine genetically isolated lineages and found evidence of differences in the size of the non-recombining region among these lineages (Menkis et al. 2010). Together, these results suggest that there may be important differences in the region of suppressed recombination around the mating-type locus among the different *N. tetrasperma* lineages (Menkis et al. 2009; Menkis et al. 2008; Menkis et al. 2010). Whittle *et al* (Whittle and Johannesson 2011; Whittle et al. 2011b) have investigated patterns of codon usage and nonsynonymous substitution within the region of suppressed recombination in another *N. tetrasperma* lineage (strain P4492; lineage #1), one different from the strain that was the main subject of (Menkis et al. 2008), (Jacobson 2005), and the work presented here (strain P581; lineage #6). They found evidence for relaxed purifying selection within this region in the form of substitutions from preferred to nonpreferred codons and, in a branch-specific analysis using PAML, a higher dN/dS ratio along the *N. tetrasperma* branch for the genes in the non-recombining region compared to their *N. crassa* and *N. discreta* orthologs. However, these analyses assume that the location and number of non-recombining strata in their lineage of interest (lineage #1; see (Menkis et al. 2009)) is the same as those reported for lineage #6 (Menkis et al. 2008), despite the previous evidence suggesting that these regions may have evolved independently (Menkis et al. 2009; Menkis et al. 2008; Menkis et al. 2010). While several other recent studies have investigated the molecular evolution of the strain we study here (Nygren et al. 2011; Whittle et al. 2011c), reviewed in (Whittle et al. 2011a), none of these have examined the evolution and structure of the region of suppressed recombination in detail.

More than forty years of genetic and cytological analyses involving *N. tetrasperma* combined with the open question as to the degree in which changes in chromosome structure are involved in the suppression of recombination, made this organism an ideal candidate for a genome sequencing project. This sequencing project was undertaken by the Joint Genome Institute and involved the sequencing and assembly of two genomes: the haploid *mat A* and *mat a* strains derived from a single heterokaryotic *N. tetrasperma* isolate. These are the same strains studied in (Menkis et al. 2008) and (Jacobson 2005) and here we use their genome assemblies to investigate the mechanisms of recombination suppression as well as the evolutionary consequences for the genes residing within the non-recombining region. We find that the majority of the region of suppressed recombination is covered by several large chromosomal rearrangements. Additionally, we show that the young evolutionary stratum identified by Menkis *et al* (Menkis et al. 2008) is located in a region where the two mating chromosomes are collinear and identify an additional stratum, created by an inversion, on the opposite end of the chromosome. We propose a model for the sequence of events and mechanisms of rearrangement that produced the current orientation and show that the region of suppressed recombination within the *mat a* strain appears to be in the early stages of degeneration.

Results

Chromosomal Rearrangements

To determine if chromosomal rearrangements could be responsible for the suppression of recombination, we created a whole genome alignment between the *mat a* and *mat A* strains of *N. tetrasperma* as well as between each *N. tetrasperma* strain and one strain of the outgroup species *N. crassa*. We visualized large-scale synteny between these pairs using dotplots (Fig. 1). Synteny is strongly conserved across all genome pairs, with the exception of those involving the *N. tetrasperma mat A* mating type chromosome. Knowing that the *N. tetrasperma mat a* mating type chromosome is collinear with that of *N. crassa*, we concluded that a series of chromosomal rearrangements had occurred on the *N. tetrasperma mat A* chromosome.

Focusing on the mating type chromosome and the comparison between the two *N. tetrasperma* strains, we found that there have been two large inversions (approximately 5.3 and 1.2 Mb), a smaller inversion (68 kb), and an apparent translocation of a ~143 kb segment from one end of the chromosome to the other (Table 1; Fig. 2). Consistent with previous cytological data, the mating-type chromosome ends are collinear and the rearrangements encompass contiguous regions within the central portion of the chromosome, however, none of the rearrangement events include the *mat* locus itself. We propose that the evolution of the *mat A* chromosome can be explained by two overlapping inversions which, together, resulted in the movement of the 143 kb segment from one end of the chromosome to the other (Fig. 2). We therefore will hereafter refer to this segment as the "relocated" region. The alternative explanation for this movement, that a 143 kb genomic segment was excised from one end of the *mat* chromosome and reinserted into the other, appears less likely for reasons explained below.

Mechanisms of rearrangement

Ectopic recombination between transposable elements is generally believed to be a common mechanism for the generation of chromosomal inversions (Casals and Navarro 2007) and its role in creating chromosomal rearrangements has been experimentally verified in yeast (Argueso et al. 2008). However, a study by Ranz *et al* (Ranz et al. 2007) found no evidence for ectopic recombination in the majority of fixed inversions between species within the *Drosophila melanogaster* group. Instead, the authors found short duplications of non-repetitive sequence at the breakpoints of most of the inversions they identified and proposed a novel mechanism for the generation of inversions involving staggered single-strand breaks followed by nonhomologous end joining.

We compared the two breakpoints associated with each inversion shown in our model to determine if there was evidence for either ectopic recombination via transposable elements or the staggered break model of (Ranz et al. 2007). Interestingly, we find evidence for both types of rearrangement mechanisms, adding credence to the novel mechanism proposed by Ranz *et al* while also supporting the notion that ectopic recombination is a common mechanism underlying chromosomal rearrangements.

At the breakpoints of the 1.2 Mb inversion, we find in *mat A*, a short duplication of a 50 basepair segment that is unique in the *N. tetrasperma mat a* genome. The orientation of this duplication is consistent with those found in *Drosophila* (Ranz et al. 2007) and with our model of the rearrangement events. In *N. tetrasperma*, after the 1.2 Mb inversion, the duplicated segments would have been in an inverted orientation as in *Drosophila* (Ranz et al. 2007). The subsequent 5.3 Mb inversion, which contained the left duplication, would have then returned the leftmost duplication to the same orientation as the right duplication (Fig. 2).

The 5.3 Mb inversion is flanked by inverted *Mariner* transposable elements. The amino acid sequence of the transposase ORFs of the two elements are ~28% identical and ~59% similar to that of the *Pogo* family Mariner-3_AN from *Aspergillus nidulans* (Kapitonov 2003). Both transposase ORFs have multiple stop codons suggesting that they are no longer active.
These sequence features also support the inclusion of the relocated region within both inversions, providing additional support for our model of how this region moved from one end of the chromosome to the other (Fig. 2). We found no such sequence features supporting the alternative model, in which the inversions occurred independently from the relocation.

The small 68 kb inversion could also have resulted from ectopic recombination. Although the sequences located at the breakpoints of this inversion have no homology to any known transposable elements, both flanks contain several microsatellite regions and regions of low sequence complexity, creating several short blocks of micro-homology that could facilitate ectopic recombination.

Relative ages of rearrangement events

We used the whole genome alignment between the two *N. tetrasperma* strains to locate 192,225 nucleotide differences between the strains, more than 99% (190,728) of which are located within the boundaries of the non-recombining region on the mating-type chromosome. These two strains are homokaryons that were derived from a single heterokaryotic strain. The lack of nucleotide differences between these strains across most of the genome is most likely due to repeated inbreeding of the self-fertile heterokaryotic strain (Merino et al. 1996).

If nucleotide differences have accumulated on the mating-type chromosome because of recombination suppression, the number of neutral differences should be proportional to the amount of time since the formation of the non-recombining region. If the rearrangement events that we observed occurred at different evolutionary timepoints, they should have different levels of neutral nucleotide divergence with the oldest event being the most divergent.
As a measure of neutral divergence between the two *N. tetrasperma* strains, we calculated the number of synonymous substitutions per site (Ks) for each gene within each rearrangement event (Fig. 3). The large 5.3 Mb inversion was the rearrangement event that allowed *N. tetrasperma* to become self-fertile by suppressing recombination along the majority of the chromosomal region between the *mat* locus and the centromere. We have therefore compared the distribution of Ks values for genes within the large inversion to those distributions for each of the other rearrangement events.

We found no significant difference between Ks values for the 5.3 Mb inversion and the 1.2 Mb inversion, nor is there a significant difference between Ks values for the 5.3 Mb inversion and the *mat* proximal block (the 218 kb chromosomal region between the *mat* locus and the leftmost rearrangement event; see Fig. 3), suggesting that the suppression of recombination in the 5.3 Mb and 1.2 Mb inversions occurred at the same evolutionary time, and that these inversions also suppressed recombination in the *mat* proximal block (Bonferroni corrected MWU tests; $P = 0.6$ and $P = 1$, respectively; Fig. 3).

We found two chromosomal regions whose Ks values were much smaller compared to those from the 5.3 Mb inversion: the *mat* distal block (the 850 kb chromosomal region between the recombining left chromosome arm and the *mat* locus; see Fig. 3) and the 68 kb inversion (Bonferroni corrected MWU tests; $P < 0.001$ in both cases; Fig. 3). The suppression of recombination in these regions most likely occurred after the rearrangement event that created the large inversion.

Surprisingly and counter to our model of the order of rearrangement events, we found that the Ks values for genes within the relocated region were much larger than those for genes within the 5.3 Mb inversion (Bonferroni corrected MWU test; $P < 0.001$; Fig. 3). One interpretation of this result is that recombination was suppressed in this region much earlier than that of the 5.3 Mb inversion. The median Ks values between the *N. tetrasperma mat A* and *mat a* strains for the genes within the relocated region are approximately 2.7 times greater than those for the next most divergent regions. If the substitution rate was the same between these chromosomal regions, the suppression of recombination within the relocated region would have had to occur at an evolutionary timepoint that was more than 2-fold earlier than the other rearrangement events. However, the divergence between *N. tetrasperma mat A* and its close relative *N. crassa* is only 1.3 fold greater than the divergence between the two *N. tetrasperma* strains (median Ks between *N. tetrasperma* and *N. crassa* for genes outside of the non-recombining region: 0.093; median Ks between the *N. tetrasperma* strains for genes within the 5.3 and 1.2 Mb inversions: 0.070). This sequence of events would place the timepoint of the relocation well before the divergence of the species. Given that only one of the species now exhibits the relocation, this scenario is very unlikely.

Investigating the large sequence divergence of the relocated region

The alternative interpretation for the large sequence divergence of the relocated region is that it has been experiencing an elevated substitution rate relative to the rest of the non-recombining region. However, because we used only synonymous codon positions to calculate divergence, the elevated rate of substitution would have to pertain to nucleotide substitutions that did not change the amino acid, rather than an elevated rate of protein evolution. One possibility is that the elevated synonymous substitution rate is due to relaxed selection for codon usage. Another is that some of the annotated genes within this region are actually pseudogenes and thus, because they are completely unconstrained by purifying selection, are accumulating nucleotide substitutions at a faster rate than synonymous positions. To address the possibility that the divergence measures for this region are biased by pseudogenes, we repeated the analysis on a subset of genes from the region for which we had evidence of expression from the *mat A* expressed sequence tags (ESTs). The increased divergence was also present in the subset of genes showing evidence of expression, suggesting that the potential inclusion of pseudogenes in our analysis was not responsible for the increased divergence we observed.

Obviously, such a filter is not infallible because a minority of pseudogenes have been shown to be transcribed based on profiling of mRNA/ESTs from a variety of organisms (between 2% and 15% of pseudogenes studied in rice (Zou et al. 2009), *Arabidopsis* (Zou et al. 2009), humans (Zheng et al. 2007), and yeast (Lafontaine and Dujon 2010)). Assuming *Neurospora* also follows this general pattern, such a filter, though incomplete, would nevertheless eliminate the majority of pseudogenes from this region.

To test the hypothesis that relaxed selection for codon usage is occurring in the relocated region, we calculated the codon usage in the 100 most highly expressed genes in *N. crassa*. We used these values to calculate the Codon Adaptation Index (CAI) for each one-to-one ortholog between *N. crassa* and the two *N. tetrasperma* strains. CAI values range from 0-1 with higher values indicating more codon usage bias (Sharp and Li 1987). To control for any bias resulting from using *N. crassa* highly expressed genes to identify favored codons, we repeated this analysis with codon usage information gleaned from the 100 most highly expressed genes in *N. tetrasperma mat A* (as determined by EST coverage) and obtained similar results (Fig. S1). We compared the CAI values for genes from each *N. tetrasperma* strain to their ortholog in *N. crassa*. We found that the genes within the chromosomal region that was relocated in *N. tetrasperma mat A* have a median CAI that is lower than that of their *N. crassa* orthologs (one-sided permutation test: *N. tetrasperma mat A P*=0.0016)(Fig. 4). This situation is not seen for the other rearrangement events; to the contrary, the genes within the 68 kb inversion appear to be evolving higher codon usage bias in both *N. tetrasperma* strains (one-sided permutation test: *P*=0.026 [*mat A*], *P*=0.003 [*mat a*]; Fig. 4).

Given that most amino acid changing mutations are likely to be deleterious, it is reasonable to expect that relaxed selection across a genomic region would result in an increase in the number of nonsynonymous substitutions per site (Ka). To minimize inflation of Ka due to the inclusion of pseudogenes or misannotated genes in the analysis, we only used genes that had evidence of expression in the form of ESTs.

We found that the genes located within the relocated region of the *N. tetrasperma mat A* strain as well as their orthologs in *N. tetrasperma mat a*, both have a median Ka that is greater than that of the non-recombining region as a whole (one-sided permutation test: *N. tetrasperma mat A P*= 0.005; *N. tetrasperma mat a P*= 0.013; Fig. 5). Together, the elevated Ka and the reduced CAI of the genes in the relocated region suggest that it has experienced reduced purifying selection in the *N. tetrasperma* lineage since its divergence from *N. crassa*.

Evidence for asymmetrical degeneration within the non-recombining region

The origin of the *N. tetrasperma* region of suppressed recombination is approximately 37x younger than that of the XY sex chromosomes of marsupial and placental mammals (~4.5 MYA versus ~166 MYA) (Veyrunes et al. 2008). Unlike mammals, it is possible for a functionally diploid (heterokaryotic) *N. tetrasperma* individual to become haploid (homokaryotic). These haploid individuals grow and reproduce via mitotic spores in the laboratory and there is evidence of outcrossing between them in the wild (Menkis et al. 2009; Powell et al. 2001). These observations suggest that there should be selection to maintain the function of both copies of the sex-linked genes and one would therefore expect the non-recombining regions to be maintained intact. However, the two strains that we examine here are almost identical across their genomes except for the non-recombining region, implying a long history of inbreeding, and previous studies have observed high instances of sexual dysfunction when *N. tetrasperma* strains are outcrossed in the laboratory (Jacobson 1995; Saenz et al. 2001). These results suggest that the degree of outcrossing in nature may be limited, in which case selection against degeneration within the non-recombining regions of both mating-types may be reduced.

One signal of degeneration that has been found previously in *N. tetrasperma* (Whittle and Johannesson 2011; Whittle et al. 2011b), as well as in regions of reduced recombination in many other organisms (Bachtrog 2003; Betancourt and Presgraves 2002; Betancourt et al. 2009; Haddrill et al. 2007; Kliman and Hey 1993; Liu et al. 2004; Marais et al. 2008; Nicolas et al. 2005; Peichel et al. 2004; Presgraves 2005; Zhou et al. 2008), is the accumulation of deleterious alleles. The elimination of recombination across a genomic region will, in essence, lower the effective population size of that region, making fixation more likely for slightly deleterious mutations and less likely for slightly beneficial ones. Given that most amino-acid changing substitutions are deleterious, suppression of recombination should result in an increased proportion of nonsynonymous substitutions compared to synonymous substitutions (Ka/Ks) across the non-recombining region (Charlesworth and Charlesworth 2000).

The previous results from a different lineage within the *N. tetrasperma* species complex found a higher Ka/Ks ratio for a set of *N. tetrasperma* genes located within the non-recombining region but not for a set of genes from outside of this region, consistent with the genes within the non-recombining region being under relaxed purifying selection (Whittle and Johannesson 2011; Whittle et al. 2011b). To determine if there is evidence supporting this prediction for the genes within the non-recombining region of this *N. tetrasperma* lineage, we calculated Ka and Ks for all ortholog pairs between each *N. tetrasperma* strain and *N. crassa* and compared the Ka/Ks ratios for the genes within the non-recombining region to those for the genes outside it.

Consistent with the prediction of (Charlesworth and Charlesworth 2000) and the previous results from another *N. tetrasperma* lineage (Whittle and Johannesson 2011; Whittle et al. 2011b), we found that the genes within the non-recombining region of the *mat a* strain have a significantly higher median Ka/Ks ratio than the genes outside of the non-recombining region, but this pattern did not hold for the genes from the *mat A* strain (MWU test: *P*=0.001 and *P*=0.6, respectively; Fig. 6). One explanation for this is that the *mat a* region may be accumulating deleterious alleles at a faster rate than the *mat A* region. To further explore this possibility, we compared the accumulation of deleterious mutations within the non-recombining region in the form of pseudogenes, codon usage, and nonsynonymous substitutions between the two *N. tetrasperma* mating types. For the pseudogene analysis, we identified candidate pseudogenes from the set of *N. crassa* predicted proteins for which we were unable to find an ortholog in the *N. tetrasperma* set of predicted peptides. Confidently identifying pseudogenes can be difficult because misannotations in the *N. crassa* genome that incorrectly identify start sites or incorrectly predict ORFs that are not actually transcribed, can make the homologous region in *N. tetrasperma* appear to be a pseudogene. For these reasons, we required a candidate pseudogene in *N. tetrasperma* to have either a frameshift or nonsense mutation, full-length homology to functional genes in both *N. crassa* and *N. discreta*, and evidence of expression in *N. crassa*. Applying these conservative criteria, we found a total of ten candidate pseudogenes: two with nonsense and/or frameshift mutations in both *N. tetrasperma* genomes, five with such mutations only in the *N. tetrasperma mat a* genome, and three appearing only in the *N. tetrasperma mat A* genome (Table S1). While it is notable that we observed more pseudogenes on the *mat a* chromosome compared to *mat A*, the sample size (ten pseudogenes in total) is not large enough to confidently conclude that this represents evidence of degeneration.

As an additional approach to assess evidence of degeneration in one of the two *N. tetrasperma* sex-linked regions, we used *N. crassa* as an outgroup to assign the nucleotide substitutions that we identified within the non-recombining region to one of the two *N. tetrasperma* lineages. After normalizing by the total number of nucleotide substitutions, we compared the frequencies of each nonsynonymous substitution (at the codon level) between the two *N. tetrasperma* strains and found that nonsynonymous substitutions have occurred at higher frequencies on the *N. tetrasperma mat a* lineage (one-sided, paired MWU test: *P*= 2.749e-15; Fig. 7A), suggesting that the *N. tetrasperma mat a* strain may be accumulating deleterious substitutions at a higher rate than the *mat A* strain.

To provide additional evidence as to whether the *N. tetrasperma mat a* chromosome is in the early stage of degeneration, we used the codon usage table mentioned previously to identify synonymous changes involving the substitution of a more preferred codon to a less preferred codon and vice versa. After normalizing by the total number of synonymous substitutions within each lineage, we found a tendency for substitutions in *N. tetrasperma mat a* that involve a change to an unpreferred codon to have occurred at higher frequencies, although this difference is not significant at an α of 0.05 (one-sided, paired MWU test: *P*=0.072; Fig. 7B). We also found that substitutions involving a change to a preferred codon have occurred at lower frequencies compared to *N. tetrasperma mat A* (one-sided, paired MWU test: *P*=0.039; Fig. 7B). These results suggest that the *N. tetrasperma mat a* non-recombining region may be in the early stages of degeneration and are consistent with observations that, within a heterokaryotic *N. tetrasperma* individual, *mat A* nuclei outnumber *mat a* nuclei during growth and early sexual development (Johannesson, *pers com*).

Suppression of recombination in other systems has often been accompanied by the accumulation of transposable elements. We identified *de novo* repetitive elements as well as those with homology to known fungal elements in both *N. tetrasperma* strains and used a permutation test to determine if the region within the non-recombining region is enriched for transposons relative to the rest of the genome. Interestingly, the *mat A* strain has significantly more transposons in the genomic region where recombination is suppressed compared to the rest of the genome (*P*=0.0004), but the *mat a* strain does not (*P*=0.30)(Fig. 8).

Discussion

In this study we have used two high quality genome assemblies, representing the nuclei of opposite mating type derived from a single heterokaryotic strain, to discover a series of three inversions within the *N. tetrasperma* region of suppressed recombination. The location of these rearrangements and the collinearity of the chromosome ends are consistent with the cytology, genetic map data, and sequence divergence data from previous studies that have investigated the region of suppressed recombination. The identification of these rearrangements answers the question raised by Jacobson (Jacobson 2005) as to the relative influence of structural versus genetic modifiers in maintaining this non-recombining region: while we show that structural rearrangements encompass most of the non-recombining region, the *mat a* and *mat A* chromosomes are collinear in the chromosomal region surrounding the *mat* locus (Fig. 3). This region includes the more recent evolutionary stratum identified in (Menkis et al. 2008) (the ~850 kb *mat* distal block) as well as the *mat* proximal block (the ~218 kb chromosomal segment between the mating type locus and the first chromosomal rearrangement). Genetic evidence assures that this segment is within the non-recombining region, and the presence of sequence divergence within this region (Fig. 3) implies that it has long remained so, but it is unclear why recombination is suppressed. However, the *mat a* and *mat A* loci do not share sequence homology and in addition, we found that within each strain, the mat locus is flanked by regions that have been subject to Repeat Induced Point mutations (RIP (Galagan and Selker 2004)). It is possible that the RIPed regions, together with the idiomorphic mat loci, create enough sequence divergence to disrupt synapsis in this region and thus cause the suppression of recombination to extend past the distal portion of the *mat* locus.

We were able to confirm the finding in (Menkis et al. 2008) that the *mat* distal block has accumulated significantly less sequence divergence than the rest of the non-recombining region (Fig. 3). Although there is no structural difference between the two chromosomes in this region, it is possible that genetic factors are suppressing recombination (as hypothesized by Jacobson (Jacobson 2005) and similar to the *rec* genes of *N. crassa* (Catcheside 1975)), or there may have been fluctuations in the exact location of the boundary of the non-recombining region over evolutionary time.

By combining our analysis of the relative timing of the rearrangement events with the analysis of sequence features associated with the boundaries of these events, we have formulated a cohesive model for the structural evolution of the *N. tetrasperma* mating-type chromosome (Fig. 2). Based on sequence divergence, the last rearrangement event (a 68 kb inversion) occurred more recently than the first two and therefore represents a second evolutionary stratum, similar to what has been found in the sex chromosomes of other taxa (Charlesworth et al. 2005).

An additional similarity between the *N. tetrasperma* non-recombining region and the sex chromosomes from other systems is that it may be under reduced purifying selection. The non-recombining region in the *N. tetrasperma mat A* strain has accumulated an excess of transposons compared to the rest of genome and there is evidence that the *N. tetrasperma mat a* strain may be in the early stages of degeneration (Figures 6-8).

It is unclear why we observe an accumulation of transposons only on the *mat A* chromosome. One possible explanation is that we identified fewer transposons in the *mat a* strain because its genome assembly has a higher proportion of sequence bases in gaps compared to the *mat A* strain (Table S3). Alternatively, we speculate that the additional transposable elements in the *N. tetrasperma mat A* strain could be the result of "genome shock" (McClintock 1984): a single burst of activity specific to the *mat A* nucleus due to the release of transposon suppression during the reorganization of the mating-type chromosome. Of course, it is equally plausible that the relative timing of the chromosome reorganization and the transpositions was reversed, such that increased transposon activity enabled the 5.3 Mb inversion that led to self-fertility in *N. tetrasperma*, thereby linking the additional transposon copies to the new chromosome orientation.

Our finding that the *mat a* chromosome is accumulating deleterious alleles at a faster rate than the *mat A* chromosome is also somewhat unexpected and stands in contrast to previous theory (Bull 1978). *N. tetrasperma* individuals can grow as haploid homokaryons which are capable of outcrossing and acting as a maternal or paternal parent. Recessive deleterious alleles would not be sheltered under these conditions and, if *N. tetrasperma* exists often as a homokaryon in nature, there should be selection to maintain both copies of the sex-linked genes. However, there is evidence that outcrossing may be limited in nature (Jacobson 1995), in which case degeneration would be reasonable because most individuals would not leave their heterokaryotic, sheltered state. Additionally, asymmetric evolution between sex-determining chromosomes, at least with respect to differences in chromosome size, has been observed in haploid dioecious systems such as the liverwort *Marchantia polymorpha* (Yamato et al. 2007) and the fungus *M. violaceum* (Hood 2002). In neither of these systems, unfortunately, have the chromosomes of both mating-types been sequenced, making it impossible to determine if the size difference is due to asymmetric gain or loss (*i.e.* degeneration) of genetic elements.

While Whittle *et al* (Whittle et al. 2011b) report an asymmetry between the *mat a* and *mat A* chromosomes in terms of the frequency of substitutions to preferred codons in lineage #1 of the *N. tetrasperma* species complex, a different lineage than studied here, they do not report whether the deviation is statistically significant and a reanalysis of their results shows that it is not (Fisher's exact test: P=0.285; Table S6). Additionally, there does not appear to be a similar asymmetry with respect to nonsynonymous substitutions in a separate study using the same lineage (Whittle and Johannesson 2011). That we are able to observe such an asymmetry in both nonsynonymous substitutions and codon usage between the sex-linked regions in this *N.*

tetrasperma lineage may be due to the larger sample of genes we use here (~1300 genes from within the non-recombining region compared to 168 in (Whittle and Johannesson 2011) and 228 in (Whittle et al. 2011b)) or to differences between the lineage studied in (Whittle and Johannesson 2011; Whittle et al. 2011b) (lineage #1) and that studied here (lineage #6).

Future work comparing the structural rearrangements identified here to the other *N. tetrasperma* lineages would address the hypothesis of independent origins of this region of suppressed recombination and lead to further insight into its evolutionary history. The examination of sequence data from *N. tetrasperma* populations within these lineages would be useful for assessing the effect of reduced recombination on nucleotide diversity. Additionally, allele-specific gene expression experiments would be an ideal approach to determine whether there is evidence that the genes within the *mat A* non-recombining region are evolving increased expression to compensate for the increase in deleterious substitutions that are occurring within the *mat a* non-recombining region.

Materials and Methods

Genome sequencing and assembly

Both *N. tetrasperma* strains (FGSC 2508 *mat A* and FGSC 2509 *mat a*) were sequenced using a hybrid approach on Roche 454 pyrosequencing and Sanger platforms (Tables S3 and S4) and assembled with Newbler. The *N. tetrasperma mat A* assembly was post-processed to close gaps in-silico using JGI gapResolution software for the entire genome and targeted finishing for the mating-type chromosome. Statistics of both assemblies are summarized in Table S4.

Genome Annotation

Both *N. tetrasperma* assemblies were annotated using the JGI annotation pipeline with results deposited to the integrated fungal resource MycoCosm (http://jgi.doe.gov/fungi) for further analysis. Genome assembly scaffolds were masked using RepeatMasker (Smit et al. 1996-2010) and tRNAs were predicted using tRNAscan-SE (Lowe and Eddy 1997). Several gene predictors were used on the repeat-masked assembly: (i) ab initio FGENESH (Salamov and Solovyev 2000) and GeneMark (Isono et al. 1994), (ii) homology-based FGENESH+ and Genewise (Birney and Durbin 2000) seeded by BLASTx alignments against GenBank's database of non-redundant proteins, and (iii) direct mapping of EST-derived full-length genes to genome assembly. Genewise models were extended where possible using scaffold data to find start and stop codons. EST BLAT alignments (Kent 2002) were used to extend, verify, and complete the predicted gene models. From the resulting set of models, a non-redundant representative set of best models was selected (Table S5).

All predicted gene models were functionally annotated using SignalP (Nielsen et al. 1997), TMHMM (Melen et al. 2003), InterProScan (Zdobnov and Apweiler 2001), BLASTp (Altschul et al. 1990) against nr, and hardware-accelerated double-affine Smith-Waterman alignments (deCypherSW; http://www.timelogic.com/decypher_sw.html) against SwissProt (Boeckmann et al. 2003), KEGG (Kanehisa et al. 2008), and KOG (Koonin et al. 2004). KEGG hits were used to assign EC numbers (Gasteiger et al. 2003), and Interpro and SwissProt hits were used to map GO terms (Ashburner et al. 2000). Multigene families were predicted with the Markov clustering algorithm (MCL)(Enright et al. 2002) to cluster the proteins, using BLASTp alignment scores between proteins as a similarity metric.

ESTs

N. tetrasperma FGSC 2508 was grown in Vogel's liquid media (Vogel 1956) and total RNA was extracted by bead-beating in TRIzol (Invitrogen Life Science Technologies) with zirconia/silica beads (0.2 g, 0.5-mm diameter; Biospec Products). ESTs were sequenced using 454 pyrosequencing. The sequencing library protocol and sequence processing procedure are described in (Swarbreck et al. 2011).

Genome synteny

Whole genome synteny was assessed between the two *N. tetrasperma* strains as well as between each *N. tetrasperma* strain and the outgroup *Neurospora crassa*. Dotplots were created using the program MUMMER (Kurtz et al. 2004) to visualize synteny while more precise identification of rearrangement breakpoints was achieved using a combination of whole-genome orthology map construction with Mercator and whole-genome alignment with MAVID (Dewey 2007).

Identification of orthologs

Orthologs between the two *N. tetrasperma* strains as well as those between each *N. tetrasperma* strain and *N. crassa* were initially identified based on best-reciprocal-BLAST hits and further verified using synteny information from Mercator. This procedure resulted in the identification of a total of 7,693 single-copy orthologs between the three species.

Calculation of nonsynonymous and synonymous substitutions per site (Ka and Ks)

Ka/Ks ratios were calculated for each pair of orthologs between the two *N. tetrasperma* strains as well as between each *N. tetrasperma* strain and the outgroup *Neurospora crassa* using the modified Yang-Nielsen method in the program KaKs_Calculator (Zhang et al. 2006). It has previously been shown that the way in which the total number of synonymous sites are counted can bias the calculation of Ks (Bierne and Eyre-Walker 2003). For this reason, we also calculated the Ka and Ks values for all pairwise orthologs as in (Bierne and Eyre-Walker 2003), using only four-fold degenerate sites for Ks and the physical site definition for both Ka and Ks and found similar results (Figure S2). For consistency, all Ks results reported here are from the modified Yang-Nielsen method in the program KaKs_Calculator.

Identification of pseudogenes

We used the tfasty program within the FASTA sequence comparison package (Pearson et al. 1997) to perform translated searches (allowing for frameshifts and premature stop codons) against the *N. tetrasperma* genome sequence. As queries, we used the subset of *N. crassa* proteins from within the genomic region of suppressed recombination for which we were unable to find *N. tetrasperma* orthologs. We used a custom Perl script to parse the results to identify matches that showed frameshifts and/or nonsense mutations. We additionally extracted the genomic sequence for candidate pseudogenes and used the program Exonerate (Slater and Birney 2005) to create genomic DNA/protein sequence alignments to confirm the presence of nonsense mutations or frameshifts.

Codon Adaptation Index (CAI)

We used the *N. tetrasperma mat A* ESTs and *N. crassa* Illumina RNA-Seq data (Ellison et al. 2011) to identify the top 100 most highly expressed genes in each species. We used these genes to compute codon usage tables for each species using the *cusp* application in the EMBOSS package (Rice et al. 2000). We then used the EMBOSS application *cai* to calculate the codon adaptation index for each *N. tetrasperma* and *N. crassa* gene. All CAI analyses were performed twice, once with the *N. tetrasperma* codon usage table and once with the *N. crassa* codon usage table. The results were equivalent in all cases and the reported p-values are those from the *N. crassa* codon usage table. We defined substitutions to preferred codons as a change to a synonymous codon whose frequency of usage in the top 100 most highly expressed genes was at least 10-fold greater than that of the ancestral codon. Similarly, we defined substitutions to unpreferred codons as a change to a synonymous codon whose frequency of usage in the top 100 most highly expressed genes was at least 10-fold less than that of the ancestral codon.

Repetitive elements

Repetitive elements were identified *de novo* using the program RepeatModeler (Smit and Hubley 2008-2010). The results of RepeatModeler were added to a fungal-specific repeat library downloaded from RepBase (Jurka 2000) and repetitive elements were identified based on sequence homology using this library and the program RepeatMasker (Smit et al. 1996-2010). Centromeric regions in *Neurospora* can be easily identified because they are composed almost entirely of transposable element remnants. The number of repetitive elements per kilobase for the chromosomal segment within the *N. tetrasperma* non-recombining region and the homologous region in *N. crassa* were calculated after excluding centromeric regions so that these measures would not be confounded by differences between assemblies in the number of gaps in these repeat-dense regions.

Permutation tests

All permutation tests were performed using custom Perl scripts.

CAI and Ka:

This test was performed on both *N. tetrasperma* strains separately. Twenty-six genes (the number of genes within the relocated region) were drawn randomly from the set of all genes within the non-recombining region. Each gene's CAI was subtracted from that of its ortholog in *N. crassa* and the median of the differences was calculated for the random sample and compared to that of the real data (*i.e.* the genes within the relocated region). This procedure was repeated 10,000 times and the p-value was calculated as the proportion of random samples (out of the 10,000 permutations) that had a value greater than that of the real data.

Ka permutation tests were performed similarly except the median Ka of randomly sampled *N. tetrasperma mat A* and *mat a* ortholog pairs was compared to that of the genes within the relocated region.

Transposon enrichment:

We compared each *N. tetrasperma* strain separately to *N. crassa*. We took the total number of transposons within each genome and shuffled their locations. We then counted the number of shuffled transposons that landed within the boundaries of the non-recombining region in *N. tetrasperma* and within the homologous region in *N. crassa*. We subtracted the *N. crassa* sum from the *N. tetrasperma* sum and compared this difference to the true difference. The p-value was calculated as the proportion of permutations (out of 10,000 total) where the permuted difference was larger than the true difference.

Figures and Tables

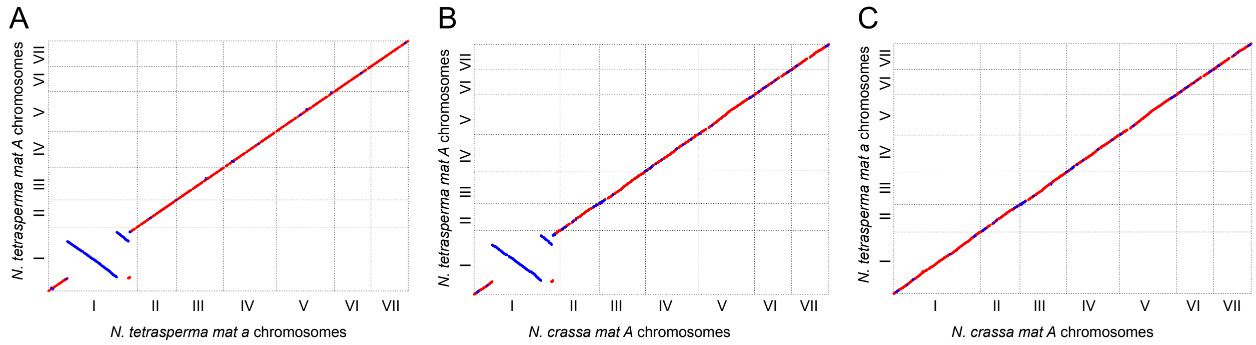

Figure 1. Whole genome synteny between *N. tetrasperma* strains and *N. crassa*.
(A) *N. tetrasperma mat A* compared to *N. tetrasperma mat a,* (B) *N. tetrasperma mat A* compared to *N. crassa*, (C) *N. tetrasperma mat a* compared to *N. crassa*. The mating type chromosome (chromosome I) is rearranged in the *N. tetrasperma mat A* strain compared to both *N. tetrasperma mat a* and *N. crassa*. Alignments occurring between positive strands are colored in red while those occurring between opposite strands are colored in blue and indicate inversions.

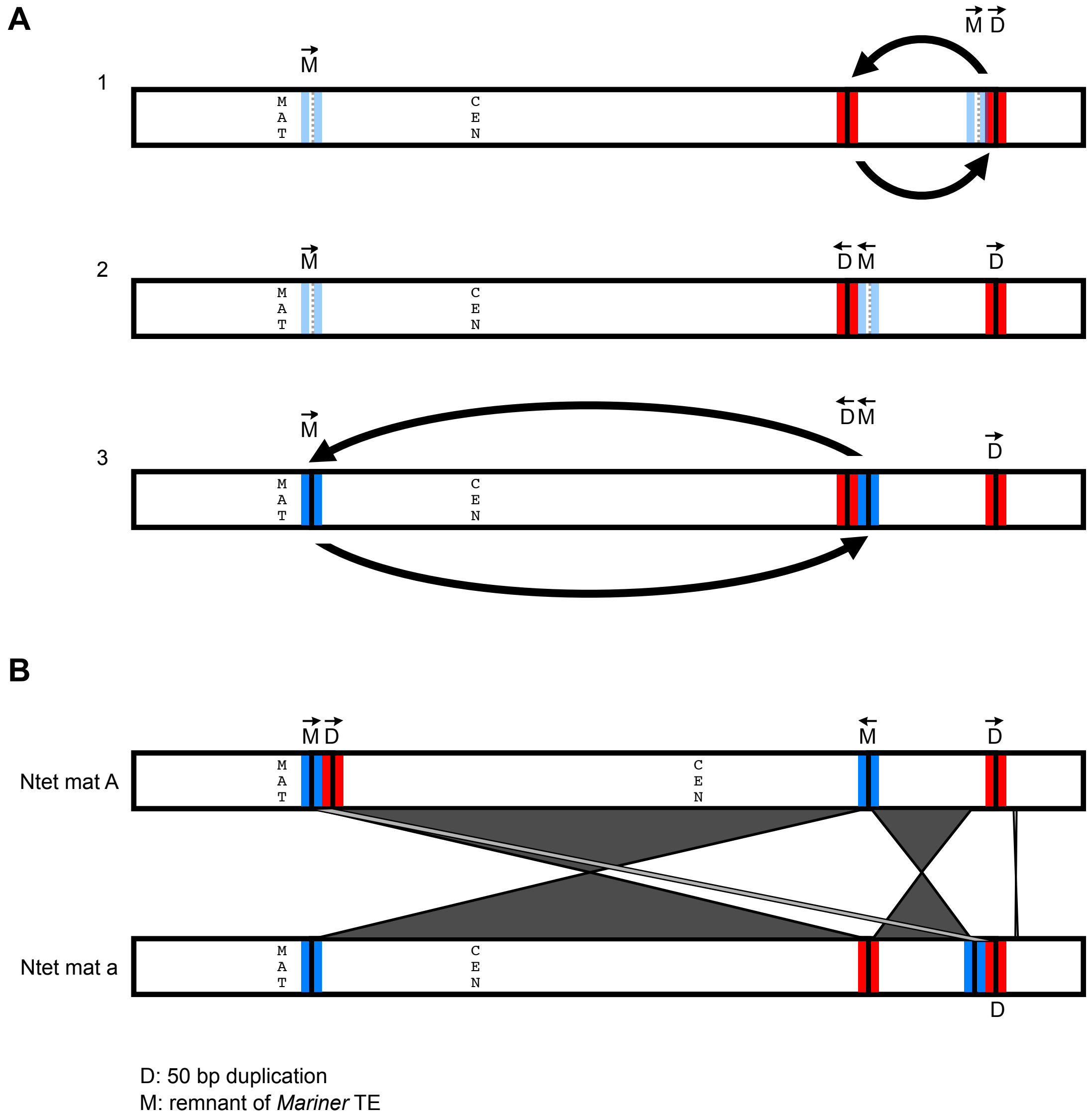

Figure 2. Model of the evolution of the *N. tetrasperma mat A* mating-type chromosome. The order of rearrangement events is shown in (A) and begins with the ancestral *mat A* chromosome (1) which was collinear with *mat a* and the mating-type chromosome of *N. crassa*. The 1.2 Mb inversion occurred first and produced the orientation in (2). This event was followed relatively quickly by the 5.3 Mb inversion (3). The 68 kb inversion, shown as the line at the far right of (B), occurred much later to produce the current arrangement of the *mat A* chromosome (B). The 1.2 Mb inversion (breakpoints shown in red) is flanked by unique 50 bp duplications (D) that would have been in an inverted orientation before the occurrence of the large inversion, consistent with rearrangement via staggered single-strand breaks. The 5.3 Mb inversion (breakpoints shown in blue) is flanked by *Mariner* transposable elements (M), consistent with rearrangement via ectopic recombination. *Mariner* remnants were not present in either of the homologous regions in the *mat a* chromosome. The overlapping nature of these two inversions explains the relocated genomic region. The 68 kb inversion is flanked by microsatellite containing, low-complexity sequence and may have occurred due to ecoptic recombination between blocks of micro-homology. MAT denotes the location of the mating-type locus while CEN shows the location of the centromere.

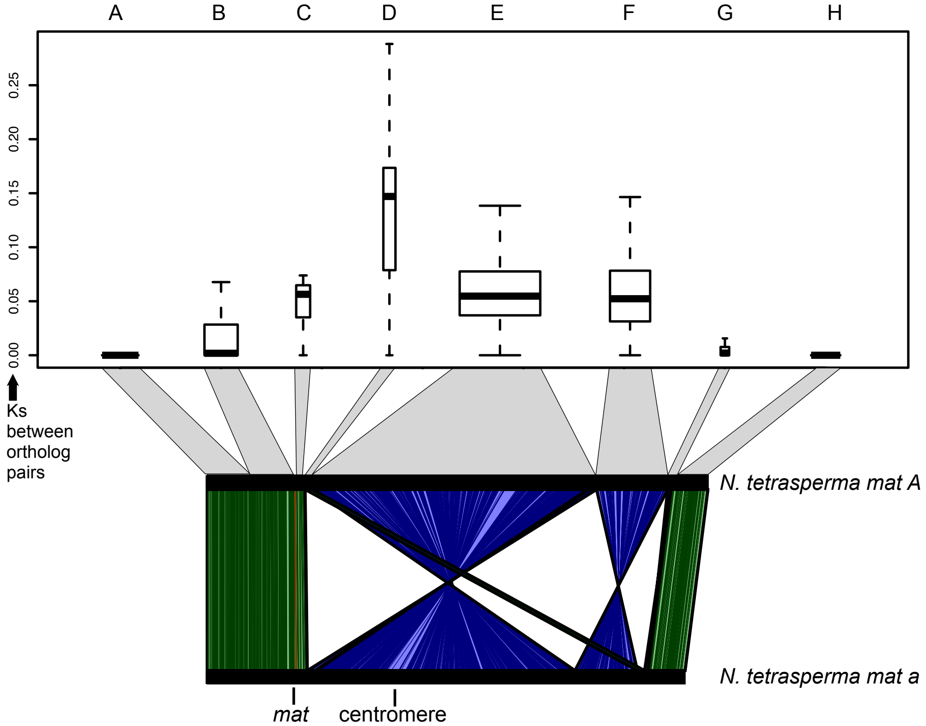

Figure 3. Relative ages of rearrangement events.

The bottom panel shows the two *N. tetrasperma* mating-type chromosomes with lines connecting pairs of orthologous genes. The top panel shows the distribution of synonymous substitutions per site (Ks) between ortholog pairs for chromosomal regions that have been rearranged and for two other chromosomal segments that are collinear between the two *N. tetrasperma* strains but have sequence divergence: an 850 kb segment distal to the *mat* locus and a 218 kb segment proximal to the *mat* locus. The amount of synonymous sequence divergence between orthologs within a given rearrangement event will be proportional to the amount of time since the event occurred. Using the 5.3 Mb inversion as a point of reference, the genes within the relocated region have a distribution of Ks values that is significantly larger, while the genes within the *mat* distal unknown block and small 68 kb inversion both have Ks distributions that are significantly smaller (Bonferroni-corrected MWU test: *P*<0.001 in all cases). The letters A-H denote the chromosomal regions that are referred to in the main text: A: recombining left arm, B: *mat* distal block, C: *mat* proximal block, D: 143 kb relocated region, E: 5.3 Mb inversion, F: 1.2 Mb inversion, G: 68 kb inversion, H: recombining right arm.

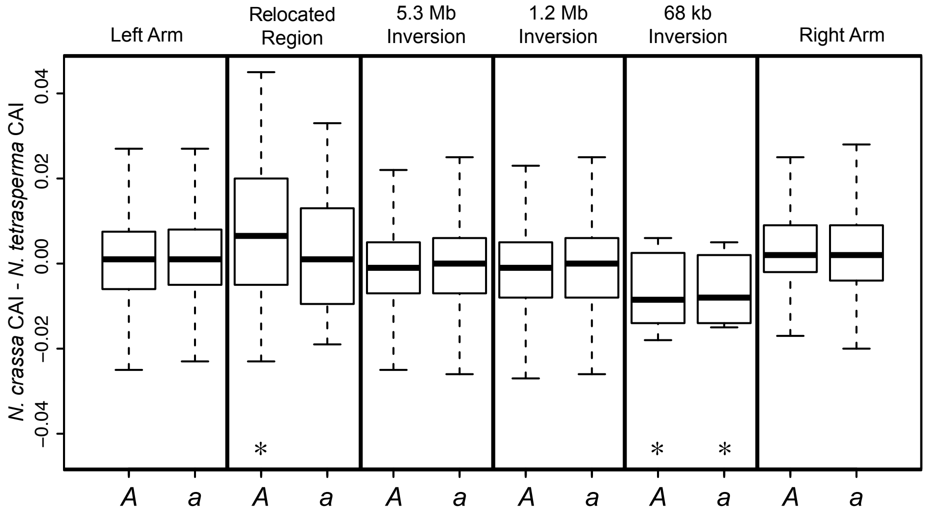

Figure 4. Relaxed selection for codon usage in the *N. tetrasperma mat A* genes within the relocated region.

The CAI is a measure of codon usage bias and the difference between the *N. crassa* CAI and those from each *N. tetrasperma* strain was calculated for every set of three-way orthologs. The genes within the *N. tetrasperma mat A* relocated region appear to be evolving reduced codon usage bias (one-sided permutation test: *P*=0.0016) while the genes within the small inversion appear to be evolving increased codon usage bias (one-sided permutation test: *P*=0.026 [*mat A*], *P*=0.003 [*mat a*]). Outliers are not shown.

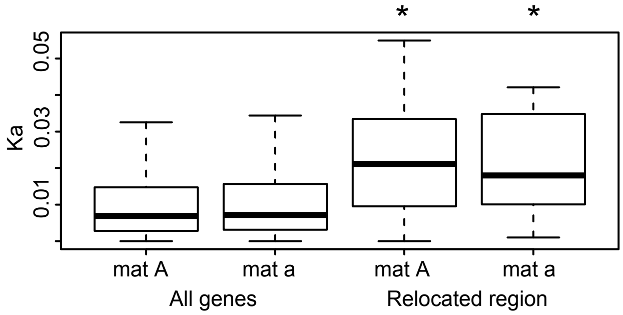

Figure 5. Increased nonsynonymous divergence in the *N. tetrasperma mat A* and *mat a* genes within the relocated region.

The number of nonsynonymous substitutions per site (Ka) between *N. tetrasperma mat A* and *mat a* orthologs within the relocated region are significantly larger than those values for the entire non-recombining region (one-sided permutation test: *P*= 0.005 [*mat A*], *P*= 0.013 [*mat a*]). Although Ka is larger for these genes, no gene has a Ka/Ks ratio significantly greater than one implying that the increased Ka is due to the accumulation of slightly deleterious amino acid substitutions rather than adaptive evolution. Outliers are not shown.

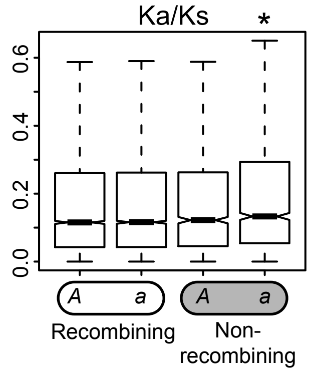

Figure 6. Reduced efficiency of selection in the *mat a* non-recombining region
The ratio of nonsynonymous substitutions per site (Ka) to synonymous substitutions per site (Ks) was calculated for every pair of orthologs between each *N. tetrasperma* strain and *N. crassa*. The distribution of Ka/Ks ratios was compared for genes inside of and outside of the non-recombining region. Consistent with the evidence in Fig. 7, genes within the region of suppressed recombination in the *mat a* strain, but not the *mat A* strain, have significantly larger Ka/Ks ratios than those outside of the non-recombining region (MWU test: *P*=0.001 and *P*=0.6, respectively). Outliers are not shown.

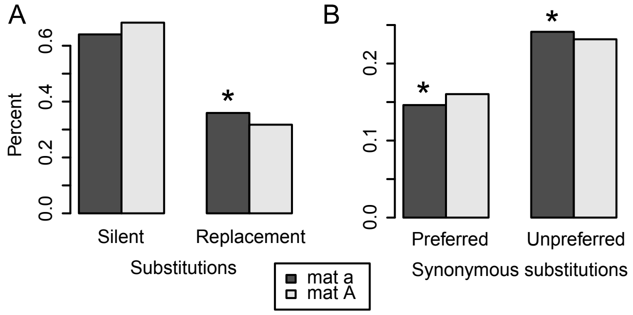

Figure 7. Evidence that the *N. tetrasperma mat a* mating-type chromosome may be in the early stages of degeneration.
Nucleotide substitutions occurring within the non-recombining region were assigned to either the *N. tetrasperma mat A or mat a* lineage depending upon which allele was present in *N. crassa*. Examination of this set of polarized substitutions showed that nonsynonymous substitutions have occurred at higher frequencies in the *N. tetrasperma mat a* lineage compared to *mat A* (one-sided, paired MWU test: *P*= 2.749e-15). In addition, examination of the set of polarized synonymous substitutions showed that, in the *N. tetrasperma mat a* lineage, preferred substitutions have occurred at significantly lower frequencies and there is a trend toward significance with respect to unpreferred substitutions having occurred at higher frequencies, compared to the *mat A* strain (one-sided, paired MWU test: preferred: *P*=0.039; unpreferred: *P*=0.072). Preferred substitutions were defined as a change to a synonymous codon whose frequency of usage in the top 100 most highly expressed genes in *N. crassa* is at least 10-fold larger than that of the ancestral codon while unpreferred substitutions were defined as the opposite (the usage frequency of the derived codon is at least 10-fold smaller than that of the ancestral codon).

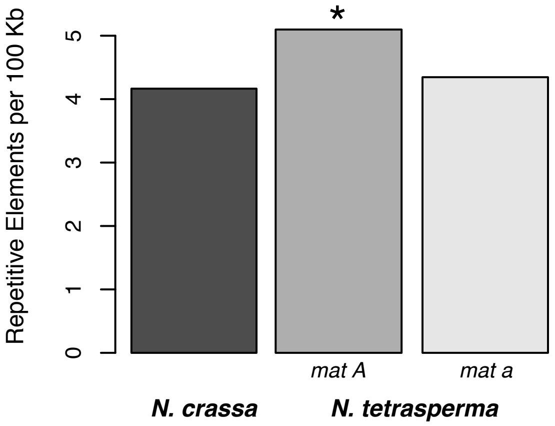

Figure 8. The *N. tetrasperma mat A* region of suppressed recombination is enriched for repetitive elements.
Repetitive elements were identified *de novo* and based on homology to known elements. Significance was assessed using a permutation test (*P*=0.0004). Using the same test, the homologous region from *N. tetrasperma mat a* did not have significantly more repetitive elements than *N. crassa* (*P*=0.30).

Chromosomal region	Size (kb)	Single-copy orthologs
Recombining left arm	930	198
mat distal block	850	166
mat proximal block	218	29
Relocated region	143	22
5.3 Mb inversion	5300	972
1.2 Mb inversion	1200	245
68 kb inversion	68	11
Recombining right arm	728	128

Table 1. Summary of the *N. tetrasperma* mating chromosome regions
The approximate size of each chromosomal region and number of single-copy orthologs are listed for the eight segments of the *N. tetrasperma* mating chromosome that are discussed in this study.

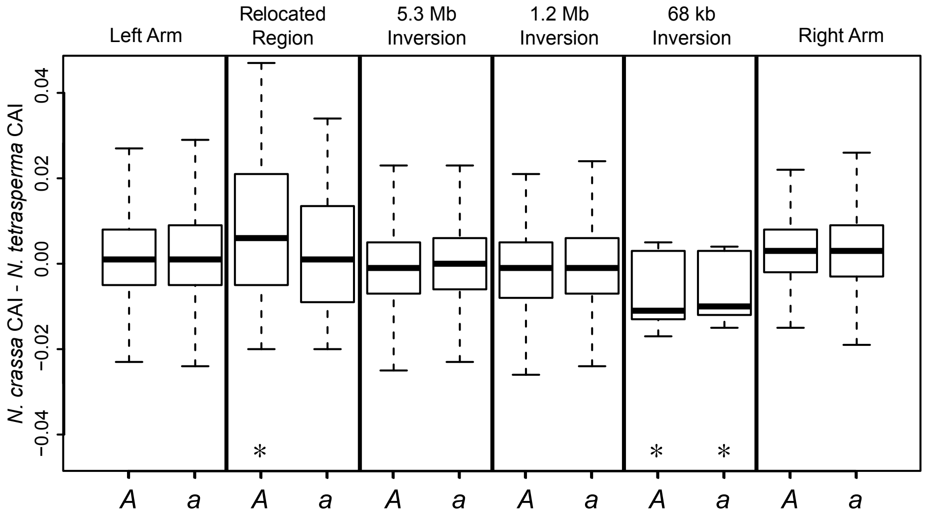

Figure S1. Re-analysis of Codon Adaptation Index (CAI) differences between *N. crassa* and *N. tetrasperma*.
The analysis shown in Figure 4 was based on codon usage in *N. crassa*. To verify that the results shown in Figure 4 do not depend on the species used to obtain codon usage information, we repeated the analysis with codon usage information from *N. tetrasperma mat A* and obtained similar results. The genes within the *N. tetrasperma mat A* relocated region appear to be evolving reduced codon usage bias (one-sided permutation test: *P*=0.0015) while the genes within the small inversion appear to be evolving increased codon usage bias (one-sided permutation test: *P*=0.0006 [*mat A*], *P*=0.0003 [*mat a*]). Outliers are not shown.

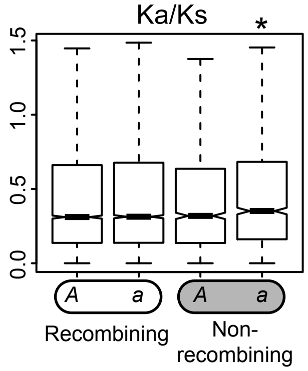

Figure S2. Re-analysis of the ratio of nonsynonymous (Ka) and synonymous (Ks) substitutions per site between recombining and non-recombining regions of the genome.
The results presented in Figure 6 were reanalyzed using only four-fold degenerate sites. Genes within the region of suppressed recombination in the *mat a* strain, but not the *mat A* strain, have significantly larger Ka/Ks ratios than those outside of the non-recombining region (MWU test: *P*=0.04 and *P*=0.4, respectively).

Gene ID	Function	Strain	Mutation
NCU03134	class I alpha-mannosidase 1A	*mat A*	FS
NCU01896	Unknown	*mat A*	FS
NCU02817	Unknown	*mat A*	PS
NCU00765	amino acid permease 2	*mat a*	PS
NCU02882	Unknown	*mat a*	PS
NCU03288	Unknown	*mat a*	PS
NCU10125	Unknown	*mat a*	PS
NCU08351	Unknown	*mat a*	PS
NCU11480	Major Facilitator Superfamily	Both	Mat A: PS; mat a: FS
NCU03275	Unknown	Both	Mat A: PS; mat a: PS and FS

Table S1. Pseudogenes present within the *N. tetrasperma* non-recombining region. Pseudogenes were identified as genes in one or both of the *N. tetrasperma* strains where mutations or insertions/deletions resulted in premature stop codons (PS) and/or frameshifts (FS) relative to their orthologs in *N. discreta* and *N. crassa*.

Nuclear Genome Assembly	Neurospora tetrasperma FGSC 2508 mat A v2.0	Neurospora tetrasperma FGSC 2509 mat a v1.0
Sequencing platform	Hybrid 454/Sanger	Hybrid 454/Sanger
Scaffold count	81	307
All Contig count	551	861
Scaffold sequence bases total	39.1 Mb	39.1 Mb
Scaffolded (Large) Contig sequence bases total	38.5 Mb	38.1 Mb
Estimated % sequence bases in gaps	1.7%	2.5%
Scaffold N50 / L50	3 / 5.7 Mb	3 / 5.7 Mb
Contig N50 / L50	89 / 134.9 Kb	117 / 99.8 Kb
Number of scaffolds > 50.0 Kb	7	7
% in scaffolds > 50.0 Kb	99.2%	98.2%

Table S2. Assembly statistics

Library	Library Type	Raw Reads	Raw Bases	Trimmed Bases	Assem Reads	Assem Bases	Coverage	Insert	Std Dev
FHCI	SANG	232,238	242,846,510	180,826,637	228,004	176,589,447	4.58x	2822	705
FHCN	SANG	209,117	181,677,551	156,903,722	205,454	154,099,062	3.99x	8558	2139
FHCO	SANG	52,883	44,594,073	34,316,772	49,195	32,285,178	0.84x	37937	9484
GBNH	454	701,361	269,672,184	268,587,339	679,800	261,469,578	6.78x		
GCSS	454PE	1,329,018	329,435,293	287,196,298	1,229,007	277,086,984	7.18x	17894	4473
GSHS	454PE	1,718,929	358,804,691	317,600,455	1,565,911	305,662,912	7.92x	3550	887
Total		4,243,546	1,427,030,302	1,245,431,223	3,957,371	1,207,193,161	31.29x		

Table S3. Sequencing libraries for *Neurospora tetrasperma* FGSC 2508 *mat A*

Library	Library Type	Raw Reads	Raw Bases	Trimmed Bases	Assem Reads	Assem Bases	Coverage	Insert	Std Dev
FHFB	SANG	39,341	29,329,148	23,656,065	36,529	21,613,054	0.57x	39775	9943
GCXS	454	2,224,123	824,396,125	821,323,297	2,083,410	770,085,673	20.18x		
GGUB	454PE	306,099	90,121,826	67,446,884	260,379	59,046,315	1.55x	20556	5139
GGUI	454	1,849,195	787,934,733	785,260,954	1,761,107	748,048,169	19.60x		
GHTG	454PE	968,920	257,042,191	230,156,414	874,202	215,706,069	5.65x	13704	3426
Total		5,387,678	1,988,824,023	1,927,843,614	5,015,627	1,814,499,280	47.55x		

Table S4. Sequencing libraries for *Neurospora tetrasperma* FGSC 2509 *mat a*

Nuclear Genome Annotation	*Neurospora tetrasperma* FGSC 2508 mat A v2.0	*Neurospora tetrasperma* FGSC 2509 mat a v2.0
# gene models	10,380	11,192
Gene density	265	286
Avg. gene length	1836	1800
Avg. protein length	468	443
Avg. exon frequency	2.72 exons/gene	2.7 exons/gene
Avg. exon length	579	548
Avg. intron length	154	152
% complete gene models (with start and stop codons)	94%	92%
% genes with homology support	91%	87%
% genes with Pfam domains	50%	46%

Table S5. Gene model statistics

	mat a	***mat A***
NPR to PR	134	128
PR to NPR	232	263

Table S6. Assessment of results from Whittle *et al* 2011b.

The numbers reported in Whittle *et al* 2011b were compared to determine if there is evidence for significant asymmetry in synonymous codon substitutions between the *N. tetrasperma mat a* and *mat A* non-recombining regions. There is not a statistically significant difference between these two regions in the proportion of non-preferred (NPR) to preferred (PR) substitutions relative to preferred to non-preferred (Fisher's exact test: *P*=0.285).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3):403-10.

Argueso JL, Westmoreland J, Mieczkowski PA, Gawel M, Petes TD, Resnick MA. 2008. Double-strand breaks associated with repetitive DNA can reshape the genome. Proceedings of the National Academy of Sciences of the United States of America 105(33):11845-50.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25(1):25-29.

Bachtrog D. 2003. Protein evolution and codon usage bias on the neo-sex chromosomes of *Drosophila miranda*. Genetics 165(3):1221-32.

Betancourt AJ, Presgraves DC. 2002. Linkage limits the power of natural selection in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America 99(21):13616-20.

Betancourt AJ, Welch JJ, Charlesworth B. 2009. Reduced effectiveness of selection caused by a lack of recombination. Current Biology 19(8):655-60.

Bierne N, Eyre-Walker A. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics 165(3):1587-97.

Birney E, Durbin R. 2000. Using GeneWise in the *Drosophila* annotation experiment. Genome Research 10(4):547-548.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research 31(1):365-370.

Bull JJ. 1978. Sex chromosomes in haploid dioecy: unique contrast to Muller's theory for diploid dioecy. American Naturalist 112(983):245-250.

Casals F, Navarro A. 2007. Inversions: The chicken or the egg? Heredity 99(5):479-480.

Catcheside DG. 1975. Occurrence in wild strains of *Neurospora crassa* of genes controlling genetic recombination. Australian Journal of Biological Sciences 28(2):213-25.

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 355(1403):1563-1572.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity 95(2):118-128.

Dewey CN. 2007. Aligning multiple whole genomes with Mercator and MAVID. Methods in Molecular Biology 395:221-36.

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proceedings of the National Academy of Sciences of the United States of America 108(7):2831-6.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-84.

Fraser JA, Diezmann S, Subaran RL, Allen A, Lengeler KB, Dietrich FS, Heitman J. 2004. Convergent evolution of chromosomal sex-determining regions in the animal and fungal kingdoms. PLoS Biology 2(12):2243-2255.

Fraser JA, Heitman J. 2004. Evolution of fungal sex chromosomes. Molecular Microbiology 51(2):299-306.

Fraser JA, Heitman J. 2005. Chromosomal sex-determining regions in animals, plants and fungi. Current Opinion in Genetics & Development 15(6):645-651.

Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends in Genetics 20(9):417-423.

Gallegos A, Jacobson DJ, Raju NB, Skupski MP, Natvig DO. 2000. Suppressed recombination and a pairing anomaly on the mating-type chromosome of *Neurospora tetrasperma*. Genetics 154(2):623-633.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research 31(13):3784-3788.

Giraud T, Jonot O, Shykoff JA. 2005. Selfing propensity under choice conditions in a parasitic fungus, *Microbotryum violaceum*, and parameters influencing infection success in artificial inoculations. International Journal of Plant Sciences 166(4):649-657.

Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biology 8(2):R18.

Hood ME. 2002. Dimorphic mating-type chromosomes in the fungus *Microbotryum violaceum*. Genetics 160(2):457-461.

Howe HB, Haysman P. 1966. Linkage group establishment in *Neurospora tetrasperma* by interspecific hybridization with *N. crassa*. Genetics 54(1P1):293-&.

Isono K, McIninch JD, Borodovsky M. 1994. Characteristic features of the nucleotide sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer program GeneMark. DNA Res 1(6):263-9.

Jacobson DJ. 1995. Sexual dysfunction associated with outcrossing in *Neurospora tetrasperma*, a pseudohomothallic Ascomycete. Mycologia 87(5):604-617.

Jacobson DJ. 2005. Blocked recombination along the mating-type chromosomes of *Neurospora tetrasperma* involves both structural heterozygosity and autosomal genes. Genetics 171(2):839-843.

Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends in Genetics 16(9):418-20.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, et al. 2008. KEGG for linking genomes to life and the environment. Nucleic Acids Research 36(Database issue):D480-4.

Kapitonov VVaJ, J. . 2003. Mariner-3_AN, a family of DNA transposons in the Aspergillus nidulans genome. Repbase Reports 3(11).

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Research 12(4):656-64.

Kliman RM, Hey J. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Molecular Biology and Evolution 10(6):1239-58.

Kondo M, Hornung U, Nanda I, Imai S, Sasaki T, Shimizu A, Asakawa S, Hori H, Schmid M, Shimizu N, et al. 2006. Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. Genome Research 16(7):815-26.

Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, et al. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5(2):R7.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biology 5(2):R12.

Lafontaine I, Dujon B. 2010. Origin and fate of pseudogenes in Hemiascomycetes: a comparative analysis. BMC Genomics 11:260.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. Science 286(5441):964-967.

Lee N, Bakkeren G, Wong K, Sherwood JE, Kronstad JW. 1999. The mating-type and pathogenicity locus of the fungus *Ustilago hordei* spans a 500-kb region. Proceedings of the National Academy of Sciences of the United States of America 96(26):15026-15031.

Lengeler KB, Fox DS, Fraser JA, Allen A, Forrester K, Dietrich FS, Heitman J. 2002. Mating-type locus of *Cryptococcus neoformans*: A step in the evolution of sex chromosomes. Eukaryotic Cell 1(5):704-718.

Liu Z, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu Q, Pearl HM, Kim MS, Charlton JW, Stiles JI, et al. 2004. A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature 427(6972):348-52.

Lowe TM, Eddy SR. 1997. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 25(5):955-964.

Mank JE, Ellegren H. 2007. Parallel divergence and degradation of the avian W sex chromosome. Trends in Ecology & Evolution 22(8):389-391.

Marais GA, Nicolas M, Bergero R, Chambrier P, Kejnovsky E, Moneger F, Hobza R, Widmer A, Charlesworth D. 2008. Evidence for degeneration of the Y chromosome in the dioecious plant *Silene latifolia*. Current Biology 18(7):545-9.

McClintock B. 1984. The significance of responses of the genome to challenge. Science 226(4676):792-801.

Melen K, Krogh A, von Heijne G. 2003. Reliability measures for membrane protein topology prediction algorithms. Journal of Molecular Biology 327(3):735-744.

Menkis A, Bastiaans E, Jacobson DJ, Johannesson H. 2009. Phylogenetic and biological species diversity within the *Neurospora tetrasperma* complex. Journal of Evolutionary Biology 22(9):1923-1936.

Menkis A, Jacobson DJ, Gustafsson T, 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.

Menkis A, Whittle CA, Johannesson H. 2010. Gene genealogies indicates abundant gene conversions and independent evolutionary histories of the mating-type chromosomes in the evolutionary history of *Neurospora tetrasperma*. BMC Evolutionary Biology 10:234.

Merino ST, Nelson MA, Jacobson DJ, Natvig DO. 1996. Pseudohomothallism and evolution of the mating-type chromosome in *Neurospora tetrasperma*. Genetics 143(2):789-799.

Metin B, Findley K, Heitman J. 2010. The mating type locus (MAT) and sexual reproduction of *Cryptococcus heveanensis*: Insights into the evolution of sex and sex-determining chromosomal regions in fungi. PLoS Genetics 6(5):Article No.: e1000961.

Nicolas M, Marais G, Hykelova V, Janousek B, Laporte V, Vyskot B, Mouchiroud D, Negrutiu I, Charlesworth D, Moneger F. 2005. A gradual process of recombination restriction in the evolutionary history of the sex chromosomes in dioecious plants. PLoS Biology 3(1):e4.

Nielsen H, Engelbrecht J, Brunak S, Von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering 10(1):1-6.

Nygren K, Strandberg R, Wallberg A, Nabholz B, Gustafsson T, Garcia D, Cano J, Guarro J, Johannesson H. 2011. A comprehensive phylogeny of *Neurospora* reveals a link between reproductive mode and molecular evolution in fungi. Molecular Phylogenetics and Evolution advance online publication, 23 March 2011(DOI 10.1016/j.ympev.2011.03.023).

O'Meally D, Patel HR, Stiglec R, Sarre SD, Georges A, Graves JAM, Ezaz T. 2010. Non-homologous sex chromosomes of birds and snakes share repetitive sequences. Chromosome Research 18(7):787-800.

Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. Genomics 46(1):24-36.

Peichel CL, Ross JA, Matson CK, Dickson M, Grimwood J, Schmutz J, Myers RM, Mori S, Schluter D, Kingsley DM. 2004. The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. Current Biology 14(16):1416-24.

Powell AJ, Jacobson DJ, Natvig DO. 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.

Presgraves DC. 2005. Recombination enhances protein adaptation in *Drosophila melanogaster*. Current Biology 15(18):1651-6.

Raju NB. 1992. Functional heterothallism resulting from homokaryotic conidia and ascospores in *Neurospora tetrasperma*. Mycological Research 96:103-116.

Raju NB, Perkins DD. 1994. Diverse Programs of Ascus Development in Pseudohomothallic Species of *Neurospora*, *Gelasinospora*, and *Podospora*. Developmental Genetics 15(1):104-118.

Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. 2007. Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biology 5(6):1366-1381.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics 16(6):276-7.

Saenz GS, Stam JG, Jacobson DJ, Natvig DO. 2001. Heteroallelism at the het-c locus contributes to sexual dysfunction in outcrossed strains of *Neurospora tetrasperma*. Fungal Genetics and Biology 34(2):123-129.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in *Drosophila* genomic DNA. Genome Research 10(4):516-522.

Sharp PM, Li WH. 1987. The Codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15(3):1281-1295.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.

Smit A, Hubley R. 2008-2010. RepeatModeler Open-1.0. <http://www.repeatmasker.org>.

Smit A, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

Swarbreck SM, Lindquist EA, Ackerly DD, Andersen GL. 2011. Analysis of leaf and root transcriptomes of soil-grown *Avena barbata* plants. Plant and Cell Physiology 52(2):317-32.

Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, Alsop AE, Gruzner F, Deakin JE, Whittington CM, Schatzkamer K, et al. 2008. Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Research 18(6):965-973.

Vogel HJ. 1956. A convenient growth medium for *Neurospora* (Medium N). Microbial Genetics Bulletin 13:42–43.

Votintseva AA, Filatov DA. 2009. Evolutionary strata in a small mating-type-specific region of the smut fungus *Microbotryum violaceum*. Genetics 182(4):1391-1396.

Votintseva AA, Filatov DA. 2010. DNA polymorphism in recombining and non-recombing mating-type-specific loci of the smut fungus *Microbotryum*. Heredity advance online publication, 17 Nov 2010 DOI 10.1038/hdy.2010.140.

Whittle CA, Johannesson H. 2011. Evidence of the accumulation of allele-specific non-synonymous substitutions in the young region of recombination suppression within the mating-type chromosomes of Neurospora tetrasperma. Heredity advance online publication, 9 March 2011 (DOI 10.1038/hdy.2011.11).

Whittle CA, Nygren K, Johannesson H. 2011a. Consequences of reproductive mode on genome evolution in fungi. Fungal Genetics and Biology advance online publication, 27 February 2011(DOI 10.1016/j.fgb.2011.02.005).

Whittle CA, Sun Y, Johannesson H. 2011b. Degeneration in codon usage within the region of suppressed recombination in the mating type chromosomes of *Neurospora tetrasperma*. Eukaryotic Cell 10(4):594-603.

Whittle CA, Sun Y, Johannesson H. 2011c. Evolution of synonymous codon usage in *Neurospora tetrasperma* and *Neurospora discreta*. Genome Biology and Evolution advance online publication, 14 March 2011(DOI 10.1093/gbe/evr018).

Yamato KT, Ishizaki K, Fujisawa M, Okada S, Nakayama S, Fujishita M, Bando H, Yodoya K, Hayashi K, Bando T, et al. 2007. Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. Proceedings of the National Academy of Sciences of the United States of America 104(15):6472-6477.

Zdobnov EM, Apweiler R. 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9):847-8.

Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics & Bioinformatics 4(4):259-63.

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, et al. 2007. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17(6):839-51.

Zhou Q, Wang J, Huang L, Nie W, Liu Y, Zhao X, Yang F, Wang W. 2008. Neo-sex chromosomes in the black muntjac recapitulate incipient evolution of mammalian sex chromosomes. Genome Biol 9(6):R98.

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. 2009. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. Plant Physiology 151(1):3-15.

Chapter 2

Population genomics and local adaptation in wild isolates of a model microbial eukaryote

Christopher E. Ellison[1], Charles Hall[1], David Kowbel[1], Juliet Welch[1], Rachel B. Brem[2], N. Louise Glass[1], John W. Taylor[1*]

Departments of [1]Plant and Microbial Biology and [2]Molecular and Cell Biology, University of California, Berkeley, CA 94720-3102, USA.

Abstract

Elucidating the connection between genotype, phenotype and adaptation in wild populations is fundamental to the study of evolutionary biology, yet it remains an elusive goal, particularly for microscopic taxa, which comprise the majority of life. Even for microbes that can be reliably found in the wild, defining the boundaries of their populations and discovering ecologically relevant phenotypes has proved extremely difficult. Here, we have circumvented these issues in the microbial eukaryote *Neurospora crassa* by utilizing a "reverse-ecology" population genomic approach that is free of *a priori* assumptions about candidate adaptive alleles. We performed Illumina whole-transcriptome sequencing of 48 individuals to identify single nucleotide polymorphisms. From these data, we discovered two cryptic and recently diverged populations, one in the tropical Caribbean basin and the other endemic to subtropical Louisiana. We conducted high-resolution scans for chromosomal regions of extreme divergence between these populations and found two such genomic "islands." Through growth-rate assays, we found that the subtropical Louisiana population has a higher fitness at low temperature (10°C) and that several of the genes within these distinct regions have functions related to the response to cold temperature. These results suggest the divergence islands may be the result of local adaptation to the 9°C difference in average yearly minimum temperature between these two populations. Remarkably, another of the genes identified using this unbiased, whole-genome approach is the well-known circadian oscillator *frequency*, suggesting that the 2.4° – 10.6° difference in latitude between the populations may be another important environmental parameter.

This chapter was previously published in the *Proceedings of the National Academy of Sciences*:

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proceedings of the National Academy of Sciences of the United States of America 108(7):2831-6.

Author contributions: R.B.B., N.L.G., and J.W.T. designed research; C.E.E., C.H., D.K., and J.W. performed research; C.E.E., C.H., D.K., R.B.B., and N.L.G. contributed new reagents/ analytic tools; C.E.E. and C.H. analyzed data; and C.E.E., R.B.B., N.L.G., and J.W.T. wrote the paper.

Introduction

Discovering the genetic basis behind adaptive phenotypes has long been considered the holy grail of evolutionary genetics. While there are now several studies that have succeeded in identifying genes responsible for such phenotypes, the majority of them utilize a "forward-ecology" approach where candidate genes are identified based on their having a function related to conspicuous traits such as pigmentation (Jiggins and McMillan 1997; Nachman et al. 2003; Pool and Aquadro 2007; Storz et al. 2009). A paucity of obvious phenotypic traits has been a major impediment for studying adaptation in microbes because these organisms are, by nature, inconspicuous. However, next-generation sequencing technology has made it possible for individual labs to acquire whole-genome sequence information across populations. This innovation has enabled an unbiased "reverse-ecology" approach where genes with functions related to ecologically relevant traits can be identified by examining patterns of genetic diversity within and between populations along environmental gradients and identifying candidate genes as those that fall within genomic regions showing the signature of recent positive selection and/or divergent adaptation between populations (Li et al. 2008).

The feasibility of such an approach has been illustrated by several recent studies in macrobes including plants (Turner et al. 2010), insects (Turner et al. 2005; Turner et al. 2008), mice (Harr 2006), and fish (Hohenlohe et al. 2010). However, even in these cases, populations had been identified *a priori* based on: candidate phenotypes associated with tolerance for serpentine soil (Turner et al. 2010), assortative mating in nature (Turner et al. 2005; Turner et al. 2008), the extremes of latitudinal clines (Turner et al. 2008), or morphology and geographic isolation (Harr 2006; Hohenlohe et al. 2010). By contrast, here we use comparative population genomics to simultaneously recognize populations *de novo* and identify candidate adaptive phenotypes.

We chose the filamentous, fungal genus *Neurospora* for this study because it is an ideal system for studying the evolutionary genomics of wild populations. Species within the genus are haploid and have relatively small genomes (40Mb) (Galagan et al. 2003). Thousands of wild strains have been collected from around the world and are available from the Fungal Genetic Stock Center, several phylogenies have been published that together provide broad taxon sampling across the genus (Dettman et al. 2003a; 2006; Menkis et al. 2009), and there is a nearly complete gene deletion collection for *N. crassa* (Colot et al. 2006).

Although *Neurospora* is a microbe, in terms of evolution, it is very similar to more developmentally complex animals. The genus is broadly distributed but also shows patterns of geographic endemism and both intrinsic and extrinsic barriers to reproduction are acting to maintain species boundaries (Dettman et al. 2003a; Dettman et al. 2003b). Additionally, Dettman *et al* (Dettman et al. 2008) have shown that reproductive isolation arises between strains of *Neurospora* evolved in the lab under different selective regimes, suggesting that local adaptation may be an important contributor to divergence between *Neurospora* populations in nature.

Unlike yeast, there is evidence that most species of *Neurospora* (including *N. crassa*) are highly outbred (Powell et al. 2003). Finally, species of *Neurospora* are haploid, free-living heterotrophs with two sexes (*mat a* and *mat A*). Each sex can produce mitotic spores that act as gametes, which may be widely disseminated. This strategy is similar to that of coral and should result in long-distance gene flow between geographically separated populations.

Here, we have discovered two previously unknown and recently diverged populations of *Neurospora crassa* (Ascomycota) by resequencing transcriptomes from 48 individuals collected from the Caribbean basin. These two populations are exposed to different local environments (subtropical versus tropical) and exhibit "islands" of divergence in genomic regions containing genes whose functions, patterns of nucleotide polymorphism, and null phenotype are consistent with local adaptation.

Results and Discussion

Population genomics. We genotyped 48 isolates of *Neurospora crassa* from the Caribbean basin, South America, and Africa (Table S1) by identifying ~135K single nucleotide polymorphisms (SNPs) from Illumina mRNA sequence tags. We estimated a SNP false positive rate of 1/18000 based on sequencing mRNA from the reference strain (Galagan et al. 2003). Using Bayesian clustering of allele frequencies (Corander et al. 2008)(Fig. S1) and phylogenetic inference using Bayesian methods (Huelsenbeck and Ronquist 2001)(Fig. 1), we found strong support for two cryptic populations in the dataset: one endemic to Louisiana and the other including isolates from Florida, Haiti, and the Yucatan (referred to as the Caribbean population). This genetic structure is also supported by our relatively high F_{ST} estimate of 0.19. These populations were not found by previous phylogenetic studies (Dettman et al. 2003a) and, in laboratory crosses, between-population reproductive compatibility is indistinguishable from that within populations (Dettman et al. 2003b).

We used a diffusion-based approach implemented in the software package ∂a∂i (Gutenkunst et al. 2009) to infer demographic parameters for these two populations under an isolation with asymmetric migration model (Fig 2; Table 1). To assess the goodness-of-fit of this model and to obtain uncertainty estimates for the demographic parameters, we used the *Neurospora* parameters to simulate 100 datasets in *ms* (Hudson 2002). The optimized log-likelihood and the sum-of-squares of the residuals for the real data fall within the boundaries of those values from the simulated data, implying that our model is neither grossly inappropriate for our data, nor is it an example of extreme overfitting (Fig. 2B). However, as shown in the heat-maps (Fig. 2A), there is an excess of high-frequency derived alleles in our data compared to what is predicted by the model. To see if this pattern was an artifact resulting from the misidentification of ancestral alleles, we fit our model to two additional datasets: In the first we included two outgroups (*N. tetrasperma* FGSC #2508 and *N. discreta* FGSC #8579)(http://genome.jgi-psf.org/Neudi1/Neudi1.home.html ; http://genome.jgi-psf.org/Neute_matA2/Neute_matA2.home.html) and restricted our dataset to include only those SNPs where both outgroups shared the same allele. In the second, we applied the misidentification correction that is part of the ∂a∂i package. We still observed an excess of high frequency derived alleles in both of these cases suggesting this pattern is not an artifact (Fig. S2). A similar pattern has been found in wild populations of *Arabidopsis* and *Oryza* (Caicedo et al. 2007; Morton et al. 2009) and in (Caicedo et al. 2007), the authors found that this pattern could be explained by either a complex demographic history including population bottlenecks and high migration rates or pervasive genome-wide positive selection in the form of selective sweeps, both of which are plausible scenarios for our *Neurospora* populations.

We infer a relatively high population migration rate from Louisiana into the Caribbean (0.77 effective migrants per generation) and about one-sixth that rate in the other direction (Table 1). We also inferred a relatively recent divergence time (~0.4 MYA) between the two populations, in agreement with the small proportion of fixed differences (9.4%; Table 2). Although we cannot eliminate the possibility that these populations diverged in complete allopatry, this scenario seems unlikely given the high dispersal potential of fungi (Taylor et al. 2006), and the fact that the Louisiana population is closer to the Caribbean population (~1000 km) than some of the Caribbean localities are to each other (~1000-1600 km). However, the strong support for population structure from multiple methods and the migration estimates from ∂a∂i both suggest that current migration is not sufficient to overcome genetic drift.

Comparisons to *Saccharomyces*. This dataset also provides an important benchmark for comparison to the recent *Saccharomyces* population genomic study (Liti et al. 2009)(Table 3). Our population genetic summary statistics indicate that, as predicted previously, *Neurospora* is much more outbred than *Saccharomyces* (Powell et al. 2003). Linkage disequilibrium decays to half its maximum value at a physical distance of ~0.78 kb compared to ~3 kb in *S. cerevisiae* and ~9 kb in *S. paradoxus*. Additionally, nucleotide diversity within each *Neurospora* population is more than 2-fold greater than that found in the UK population of *S. paradoxus* and the Wine/European cluster of *S. cerevisiae*. We have also used the same approach as in Liti *et al* (Liti et al. 2009) to estimate the number of deleterious nonsynonymous polymorphisms segregating in the *Neurospora* populations. Consistent with *Neurospora* being more outbred, we estimate that approximately 34% of non-synonymous polymorphisms are deleterious in the two *Neurospora* populations, which is about half the amount estimated in *Saccharomyces* (Liti et al. 2009).

Finally, ~50% of the SNPs we identified are still segregating within each of the two *Neurospora* populations whereas the majority of polymorphisms identified by Liti *et al* are fixed within each non-mosaic *S. cerevisiae* lineage. Given that their effective population sizes are around an order of magnitude smaller than the estimates for *S. cerevisiae* (Skelly et al. 2009) and *S. paradoxus* (Tsai et al. 2008), the amount of ancestral variation that remains within these two *Neurospora* populations implies that they are much more recently diverged than any found in the *Saccharomyces* study. Such recent divergence makes these populations an ideal system for the study of incipient speciation and adaptation.

Genomic islands of divergence between populations of *Neurospora*. To identify candidate genomic islands of divergence, we conducted sliding window estimates of three different population genetic parameters: F_{ST} (Wright 1950), Tajima's *D* (Tajima 1989), and Dxy (Nei 1987). For each parameter, we identified empirical outliers in the 0.5% quantile. F_{ST} measures relative divergence and is the most commonly used metric in studies of heterogeneous genomic divergence (Beaumont 2005). We additionally use Dxy, a measure of absolute divergence, following the recommendation of Noor and Bennet (Noor and Bennett 2009). To our knowledge, Tajima's *D* has not been previously used for this purpose but, in principle, scans of the combined populations should produce large positive values for regions showing low within-population polymorphism and high between-population divergence, making it similar to a relative divergence measure.

We were surprised to discover little overlap between the significant regions identified by the three different metrics. Out of a total of 37 regions, only two were identified by all three metrics (Fig. S3). Interestingly, removing from analysis the sites that fell within these regions still resulted in a phylogeny with strong support for the two populations (Fig. S8). As such, these two major loci are not the sole drivers for the population structure that we observe. This finding is consistent with the results of Bayesian clustering of allele frequencies (Fig. S1) where the two populations were delineated based only on differences in allele frequencies and lends credence to our model in which, despite the presence of gene flow between populations, genetic drift and/or natural selection has resulted in genome-wide differences in allele frequencies between populations.

Apart from these two candidate islands of divergence, we observed little overlap between the top-scoring regions across the three different metrics of population divergence applied to the data. A total of 35 regions were called significant in some but not all analyses. To investigate these discrepancies, we examined these regions in more detail.

Genomic loci that did not achieve consensus across our tests for divergence fit into three major classes: A) block-like haplotypes that do not perfectly sort by population, B) regions where relative divergence is high but absolute divergence does not stand out from the genomic background and C) regions where absolute divergence is high but relative divergence does not stand out from the genomic background (Fig. S4). Patterns (A) and (B) were predicted by Noor and Bennett (Noor and Bennett 2009) and may result from an inversion or other barrier to recombination that was segregating in the ancestral population. In pattern (A), both haplotypes are still segregating in each population. In pattern (B), relative divergence is high because alternate haplotypes have become fixed in each population, but absolute divergence does not stand out from the genomic background because the region itself contains few polymorphic sites. Finally, the pattern described in (C) was only seen in Dxy outliers. Absolute divergence is high because there are a large number of polymorphisms in the region but relative divergence does not stand out from the genomic background because most of the polymorphisms are still segregating in both populations. We conclude that analyses of inter-population divergence using only a single measure (*e.g.* F_{ST}) are susceptible to the identification of false-positives. This result is troubling because it suggests that any such study, along with those using datasets of much lower resolution, will be unable to distinguish these misleading divergence outliers from true islands of divergence, potentially drastically overestimating the prevalence of this phenomenon (Noor and Bennett 2009).

Genes inside divergence islands have functions and patterns of variation consistent with local adaptation. The difference in latitude between the Louisiana and Caribbean populations suggests that they may have experienced differences in selective forces related to environmental parameters such as day length and average yearly minimum temperature (5.0°C for Welsh, Louisiana and 13.8°C for Homestead, Florida (PRISM 2003)). We sought to investigate whether our candidate genomic islands of divergence between these two populations could harbor genetic factors that are locally adapted.

The first divergence island is on chromosome three and contains a pattern of nucleotide variation consistent with independent selective sweeps within each population: an excess of variants segregating at low frequency and reduced π (average number of intrapopulation pairwise differences) within both populations, relative to the flanking regions (Fig. 3 and Fig. S5). We find both the Caribbean and Louisiana haplotypes present among the outgroup strains (Fig. 3), but strains from the same locality always have the same haplotype. These facts are consistent with either a history of gene migration among populations or the presence of both haplotypes in the ancestral population, followed by the sweeping of a single haplotype to fixation within populations.

This region contains the genes *plc-1* (phospholipase C), an *MRH4*-like mitochondrial DEAD box RNA helicase and the unnamed gene NCU06247 (inferred to encode an outer mitochondrial membrane protein (Schmitt et al. 2006)). Coincidentally, Gavric *et al* (Gavric et al. 2007) observed this same pattern of divergence in *N. crassa plc-1*, but, lacking the context of the two different populations and the genome-wide sampling presented here, could not explain it. We also found another mitochondrial DEAD box RNA helicase (homolog of the yeast gene *MSS116*) as one of twelve genes in the Louisiana population that show the signature of positive selection by the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991) (Table S2). We did not expect to find the *MRH4*-like RNA helicase in this case because the MK test is confounded by the reduced within-population polymorphism in the genomic islands of divergence. RNA helicases are key factors in the microbial cold response (Hunger et al. 2006), making it tempting to speculate that they are important to Louisiana *N. crassa* which experience minimum temperatures almost 9°C lower than their Caribbean relatives.

The second divergence island is on chromosome seven and was identified by the highest observed values of all three divergence measures. It shows an unusually large number of variable sites, the majority of which are fixed between populations (Fig. 3). At first glance, this pattern appeared to be consistent with the action of repeated selective sweeps within each population. A selective sweep within each population should produce an excess of alleles segregating at low frequency and a reduction in the number of polymorphic sites (Andolfatto 2001). This prediction holds true for the Louisiana population, in which Tajima's *D* for the region is negative and π decreases relative to the flanking regions (Fig. S6). This pattern, however, is not seen in the Caribbean Basin population (Fig. S6). Additionally, all non-Louisiana strains have the same haplotype and the boundaries of the distinct region in the Louisiana population vary among individuals (Fig. 3). Together, these observations point to the introgression of a genomic region as a single "migrant tract" (Pool and Nielsen 2009) into Louisiana from a more genetically diverged population or species that we did not sample. Under this model, the introgressed haplotype would have rapidly spread through the Louisiana population, explaining why nucleotide polymorphism within this region is reduced in this population but not the Caribbean, while the non-uniformity of the region's boundaries could be due to recombination that occurred after the introgression (Pool and Nielsen 2009).

The sweep to fixation of this region within the Louisiana population implies that it contains a gene that may confer a local selective advantage over the ancestral haplotype. Among the five genes in this region is the circadian oscillator gene *frequency* (*frq*), the subject of a significant body of work using *Neurospora crassa* as a model for understanding the circadian clock (e.g. (Aronson et al. 1994; McClung et al. 1989; Merrow et al. 1999)). Also present are an *NSL1*-like kinetochore MIND complex subunit, a *SEC14*-like phosphatidylinositol/phosphatidylcholine transfer protein, a *PAC10*-like prefoldin alpha subunit, and a gene of unknown function (NCU02261). As with the helicases, it is tempting to speculate that *frq* is involved in adaptation, in this case related to differences in local photoperiod associated with the 2.4 to 10.6 degree difference in latitude between the Louisiana population and various Caribbean population localities.

Characterizing the candidate adaptive phenotypes. The distributions of these two populations in conjunction with the RNA helicase and major circadian oscillator that we find within these genomic islands of divergence suggest two major environmental factors that may be promoting local adaptation: temperature and day length. Here we have chosen to focus on the response to low temperature. We chose to focus only on low temperature, rather than both low and high temperature for several reasons. The global distribution of *Neurospora crassa* is mainly tropical, implying that the extension of its range into more temperate Louisiana is a derived condition

(Turner et al. 2001). In addition, there is a 9°C difference in the mean annual minimum temperature between Welsh, Louisiana and Homestead, Florida, but only a 0.7°C difference in the mean annual maximum temperature (PRISM 2003). Thus, although winter in Louisiana is noticeably cooler than winter in the tropics, the summers are equally warm.

We predicted that individuals from the Louisiana population would exhibit higher fitness in cold temperature relative to individuals from the Caribbean. To test this prediction, we measured the growth rate of ten randomly chosen individuals from the Louisiana population and ten from the Caribbean population at 10°C and 25°C. For each individual, we calculated its growth rate at 10°C as a percentage of its growth rate at 25°C and found, as predicted, the reduction in growth rate at 10°C for strains from the Louisiana population is significantly less than that for strains from the Caribbean population, consistent with Louisiana strains exhibiting higher fitness at lower temperatures (P=0.031; one-sided MWU test; Fig. 4A).

To begin to address the potential role in cold adaptation of the candidate genomic islands of divergence identified by our sequence analysis, we used strains from the *N. crassa* deletion collection (Colot et al. 2006) to determine whether genes in these islands were involved in low-temperature growth.

Preliminary growth experiments on null mutants of each locus at 10°C suggested a cold temperature growth defect in deletions of the *MRH4*-like RNA helicase, the *PAC10*-like prefoldin subunit, and the unannotated gene NCU06247 (Fig. S7). To control for unlinked lesions introduced during generation of the deletion strains, and to verify reproducibility, we crossed each marked deletion strain to an unmarked tester strain and compared the growth rate of progeny with and without the deletion marker cassette. This experiment confirmed significant growth defects resulting from deletion of either of the two annotated genes but not NCU06247 (Fig. 4B-D). The importance of the *MRH4*-like RNA helicase for growth at cold temperature in *N. crassa* is consistent with work on RNA helicases in many other systems (Hunger et al. 2006; Kim et al. 2008; Schade et al. 2004), though the relatively modest effect of deleting this locus may be a consequence of functional redundancy among the 18 known helicases in the *N. crassa* genome, as has been suggested in *Arabidopsis* (Kim et al. 2008).

Taken together, our results indicate that the *MRH4*-like RNA helicase and the *PAC10*-like prefoldin subunit are critical for wild-type growth in cold temperatures in *N. crassa*, lending credence to the model that these genomic islands of divergence are the result of adaptation to low temperature. It should be noted, however, that large-scale fluctuations in climate have taken place since the divergence of these populations ~0.4 MYA (Petit et al. 1999). Although there have been four major glacial events during this time period, at ~0.4 MYA the planet was in the middle of an interglacial period with an ice volume and surface temperature that is remarkably similar to current levels (Petit et al. 1999). In addition, although temperatures were cooler in absolute terms during the glaciations, paleontological studies based on pollen and plant microfossils indicate that the relative difference in temperature between the Florida peninsula and that of Louisiana was still present during the last glacial maximum (Jackson et al. 2000). There is no evidence that this most recent glacial period was much more severe than those that preceded it (Petit et al. 1999), indicating that the Florida/Louisiana temperature difference likely was maintained since the divergence of these two populations.

Future work will be needed to establish the relationship between sequence variants at these loci, cold tolerance, and other environmental parameters such as day length that are relevant to the Caribbean and Louisiana populations. It is especially interesting to consider the possibility that the genes in these distinct genomic regions may be interacting in both the response to cold and the circadian rhythm given that the circadian clock of *N. crassa* exhibits temperature compensation and can be entrained by temperature in addition to light (Liu and Bell-Pedersen 2006).

Conclusion

Here we have illustrated the utility of combining a "reverse ecology" genome-scan approach with functional characterization of the resulting candidate genes to identify ecologically relevant phenotypes in organisms that are difficult to study in nature. The major benefit of this approach, compared to a purely candidate gene approach, is that it provides a relatively unbiased look across the whole genome, allowing for identification of genes whose role in adaptation may not have been expected *a priori*. As it becomes easier to obtain large amounts of DNA sequence data, this type of approach is becoming increasingly common and will help facilitate the study of ecologically important non-model systems.

Although this approach has been demonstrated in other systems (Harr 2006; Hohenlohe et al. 2010; Turner et al. 2010; Turner et al. 2005; Turner et al. 2008), this is the first time it has been used with a microbe, which is where it may prove to be the most useful. It can be difficult to apply this type of approach to populations of non-model organisms because it generally needs to be combined with a nearly complete reference genome assembly or an unfinished assembly paired with a genetic map (Storz and Wheat 2010). However, compared to macrobes, most microbes have smaller genomes with a lower repeat density and low-cost, high-quality *de novo* genome assemblies from short reads have been achieved for both fungi (Nowrousian et al. 2010) and bacteria (Reinhardt et al. 2009). These features of microbes suggest that it is feasible to produce a reference genome assembly from a single individual while additionally resequencing many other individuals at low coverage to obtain polymorphism data which can be used for the genome scan. Furthermore, microbes are generally more amenable to genetic transformation, which may aid in the functional characterization of the candidate genes identified in the genome scan.

Materials and Methods

Identification of single nucleotide polymorphisms (SNPs). Messenger RNA-Seq reads that did not map uniquely were discarded. Read alignments from each strain were pooled and SNPs were identified using a Bayesian approach implemented in the program GigaBayes (Hillier et al. 2008). To be included in the final set of high-quality SNPs, a candidate site was required to be biallelic and needed to meet or exceed the following criteria: coverage of five reads per allele, individual base qualities of 10, aggregate base qualities of 40, and Bayesian genotype probability of 0.90. To further reduce the number of potential false positives, singletons were discarded. Sites with missing data (*i.e.* the allele of one or more individuals was unidentifiable because it did not meet the above criteria) were excluded from analysis. Using these criteria, we found 5640 genes that had at least one SNP out of the approximately 9800 genes in the genome. These 5640 genes had an average of 14.4 SNPs per gene.

Analysis of population demographics. Demographic parameters were estimated from the Louisiana and Caribbean joint allele frequency spectrum using a diffusion-based approach implemented in the program ∂a∂i (Gutenkunst et al. 2009). To control for the potential misidentification of ancestral states, we fit the model to two additional datasets: one where we used two outgroups (*N. tetrasperma* FGSC #2508 (http://genome.jgi-psf.org/Neute_matA2/Neute_matA2.home.html) and *N. discreta* FGSC #8579 (http://genome.jgi-psf.org/Neudi1/Neudi1.home.html)) and one where we applied a correction that is part of the ∂a∂i package. The results were nearly identical (Fig. S2) and we report the parameters estimated from the uncorrected spectrum

Growth rate assays. All strains used in the growth rate assays were *a* mating type. The location of the hyphal front was recorded at regular intervals until it reached the other end of the tube. Each strain was grown in triplicate in constant darkness inside 25°C and 10°C incubators. Crosses involving null mutants were made on synthetic crossing medium (Westergaard and Mitchell 1947) to the *fluffy* mating type tester strain. The *fluffy* strain contains a mutation at a single locus which makes it aconidate and highly fertile (Lindegren 1933). All progeny used in growth rate assays were screened to ensure that they produced macroconidia and thus did not have the *fluffy* mutation.

See Supporting Information below for detailed materials and methods.

ACKNOWLEDGEMENTS. We thank M. Slatkin and D. Matute for helpful discussion and reading an earlier version of this manuscript. We also thank R. Gutenkunst for help with ∂a∂i and the Fungal Genetic Stock Center for providing the strains used in this study. This work was supported by NIH grant NIGMS R01RGM081597 to JWT, NLG and RB and the Chang-Lin Tien Graduate Fellowship to CEE.

Figures and Tables

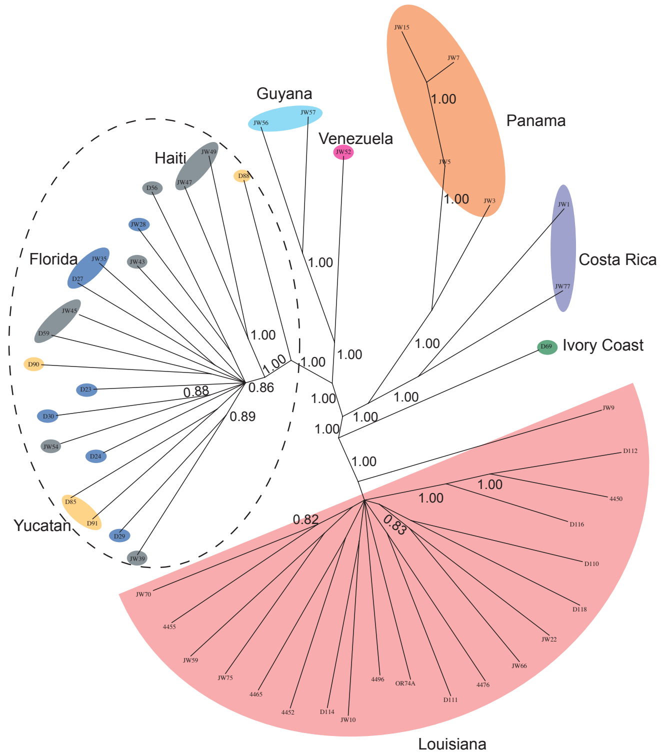

Figure 1. Unrooted Bayesian phylogeny showing cryptic population structure. The tree was inferred from 135035 polymorphic sites. The support values are clade posterior probabilities and values less than 0.75 are not shown. Each taxon is colored according to collection locality. The pink semicircle and dashed ellipse correspond to the Louisiana and Caribbean (Florida, Haiti, Yucatan) populations, respectively. The remaining clades likely represent additional populations that we are unable to resolve given their small sample size.

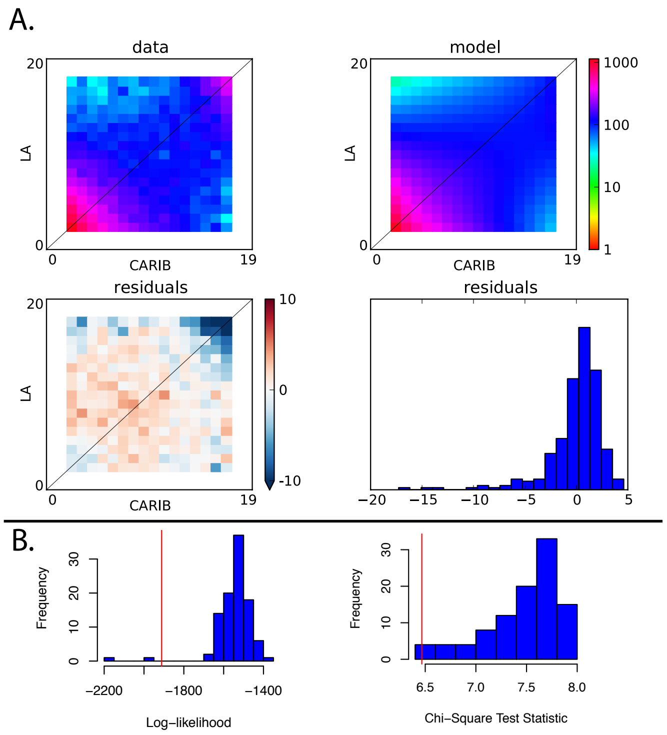

Figure 2. Comparison of the *N. crassa* joint allele frequency spectrum to that expected under an isolation with migration demographic model.

To determine whether our model of isolation with asymmetrical migration is appropriate for our data, we used two metrics to compare the joint allele frequency spectrum for the Louisiana and Caribbean populations of *Neurospora* to that expected under our model: (A) visual examination of the allele frequency spectra and calculation of the residuals between model and data, and (B) log-likelihood and Pearson's χ^2 goodness-of-fit tests.

(A) Heat-maps showing the joint allele frequency spectrum for the Louisiana and Caribbean populations compared to that expected under a simple isolation with migration model. The color of a cell at position [X,Y] in the matrix corresponds to the number of derived alleles (relative to *N. tetrasperma*; see Methods) that are at frequency X in the Caribbean population and frequency Y in the Louisiana population. The residuals represent the difference in the number of alleles predicted by the model compared to that found in the data for each bin in the spectrum (red: model predicts too many; blue: model predicts too few). The cluster of blue bins in the upper right corner of the heat-map of residuals implies that the model predicts too few high-frequency derived alleles, or in other words, there is an excess of high-frequency derived alleles in the data. Similar patterns have been found in *Arabidopsis* and rice and may be due to a history of repeated bottlenecks and high migration and/or genome-wide selective sweeps (Caicedo et al. 2007; Morton et al. 2009) (B) Goodness-of-fit tests using the likelihood and Pearson's X^2 statistics and based on the results of 100 coalescent simulations in *ms* (Hudson 2002) under the demographic parameters inferred for the two *Neurospora* populations. Better fits have likelihood and X^2 values closer to zero. The values from fitting the real data are shown by the red line. These values fall within those from the simulations indicating that the model is neither grossly inappropriate for the data, nor is it an example of extreme overfitting.

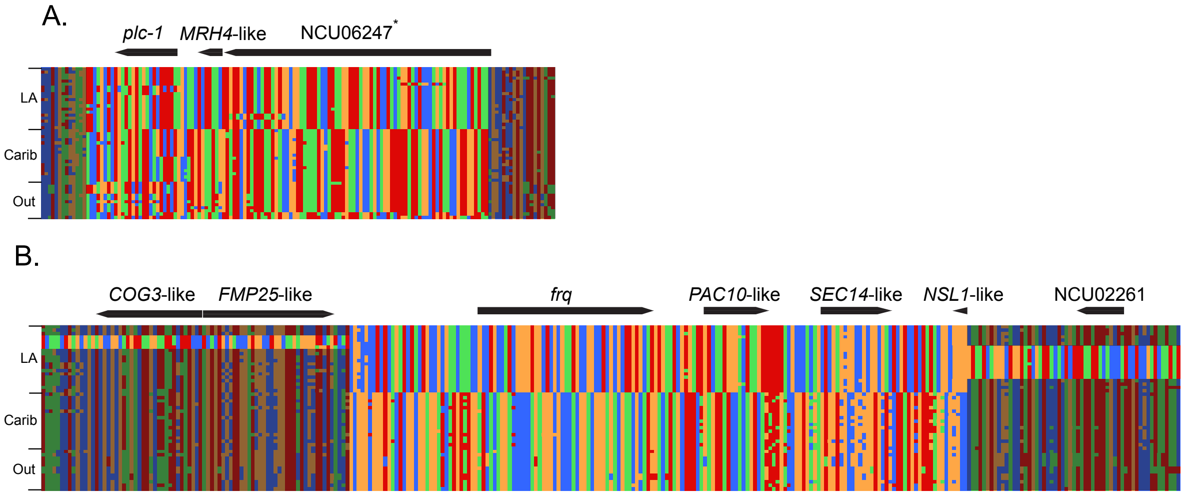

Figure 3. Genomic islands of divergence genotype matrices
Each column is a polymorphic site and each row contains the genotype for a particular strain. The flanking regions surrounding the divergence outliers are shaded and are shown to accentuate the distinct patterns of nucleotide polymorphism within the divergence outlier regions. Strains are grouped by population of origin. LA=Louisiana, Carib=Caribbean (Florida, Haiti, and the Yucatan), and Out=outgroups from Central America, South America, and Africa. The matrix in (A) is the 10 kilobase divergence island on chromosome three. The matrix in (B) is the 27 kilobase divergence island on chromosome seven. *The function of NCU06247 is unknown but the protein localizes to the outer mitochondrial membrane (Schmitt et al. 2006).

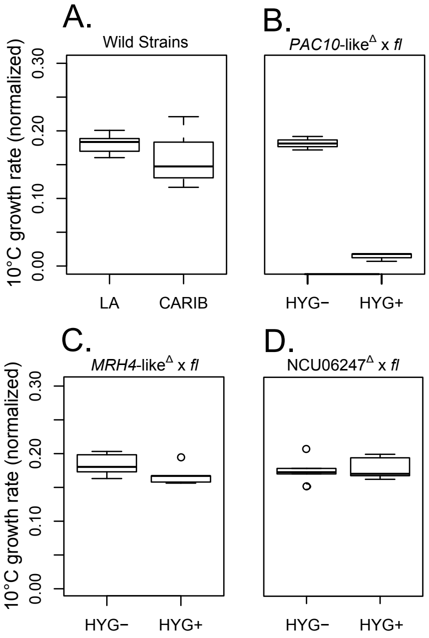

Figure 4. Adaptation to cold temperature and functional characterization of genes within divergence islands

A. For ten strains from each population, growth rate at 10°C was calculated as a percentage of that at 25°C. Strains from Louisiana grow significantly faster than those from the Caribbean at 10°C (*P*=0.031; one-sided MWU test).

B.-D. Three null mutant strains were crossed to the *fluffy* mating-type tester strain and the growth rates of progeny with the hygromycin deletion cassette were compared to progeny with the wild-type allele (hygromycin sensitive). The growth rate at 10°C was calculated as a percentage of that at 25°C. The progeny with the *PAC10*-like null allele and the *MRH4*-like null allele grew significantly slower at 10°C while those with the NCU06247 null allele did not (one-sided MWU test; *P*=0.05, *P*=0.048, and *P*=0.5, respectively; sample size = 3 wild-type and 3 mutant progeny for (B), 5 wild-type and 5 mutant progeny for (C) and (D)).

Populations	F_{ST}	$\theta\pi$	LD decay (kb)	Deleterious SNPs (%)
Caribbean	NA	0.0023	0.85	34.0
Louisiana	NA	0.0024	0.70	33.1
Between *Neurospora* populations	0.191	0.0029	NA	NA
S. paradoxus (UK)	NA	0.0010	~9.0	NA
S. cerevisiae (Wine/European)	NA	0.0011	~3.0	61.0

Table 1. Population demographic parameters and uncertainty estimates
This table shows the maximum-likelihood parameter estimates for an isolation with asymmetrical migration model fitted to the joint allele frequency spectrum for the Caribbean and Louisiana *N. crassa* populations. Uncertainty estimates were obtained by calculating the mean of the point estimate and the corresponding standard deviation for each model parameter from 100 datasets simulated in *ms* (Hudson 2002) using the *Neurospora* maximum-likelihood parameters.

Parameter	Point Estimate	Mean	S.D.
Ancestral N_e	344049	343700	9822.7
LA N_e after split	1151961	1211000	488347.4
CARIB N_e after split	364841	387000	209200.4
Divergence time	494547	514500	165472.5
Effective migration rate: CARIB into LA	0.1297	0.1639	0.1376
Effective migration rate: LA into CARIB	0.7651	0.9032	0.4705
N_e: Effective population size			

Table 2. Population genetic summary statistics: comparison to *Saccharomyces*.
Summary statistics for the Caribbean and Louisiana populations of *N. crassa* compared to those for the Wine/European population of *S. cerevisiae* and the UK population of *S. paradoxus* calculated by Liti *et al* (Liti et al. 2009). $\theta\pi$ is the average number of pairwise differences between individuals, LD decay is the average physical distance over which the coefficient of linkage disequilibrium (r^2) decays to half its maximum value, and deleterious SNPs are an estimate of the number of amino-acid changing polymorphisms that are deleterious, calculated as in (Liti et al. 2009). NA indicates that either the statistic is not applicable or the value was not calculated by Liti *et al* (Liti et al. 2009).

Category	Percent of Total SNPs
Fixed in both populations	9.4%
Polymorphic in both populations	48.7%
Fixed in LA, polymorphic in CARIB	19.0%
Fixed in CARIB, polymorphic in LA	22.9%

Table 3. Summary of Single Nucleotide Polymorphisms (SNPs).
We identified 135,035 SNPs from the pooled sequence information obtained from all isolates used in this study. This table shows the percentage of the total SNPs that fall into a given category. LA=Louisiana population, CARIB=Florida-Haiti-Yucatan population.

Supporting Information

Detailed Materials and Methods

Strain collection. Isolates of *Neurospora crassa* were provided by the Fungal Genetic Stock Center (FGSC) and were chosen with the goal of deeply sampling the diversity available from the Caribbean basin and broadly including strains from South and Central America as well as Africa (Table S1).

Library preparation and sequencing. *Neurospora* strains were initially germinated on Vogel's minimal medium (Vogel 1956). After growth was observed, a hyphal plug was transferred to a petri dish containing Bird medium (Metzenberg 2004) overlaid by cellophane and left to grow under constant light for 24 hours before harvesting. Total RNA was extracted from the mycelia by bead-beating in TRIzol (Invitrogen Life Science Technologies) with zirconia/silica beads (0.2 g, 0.5-mm diameter; Biospec Products). Messenger RNA was purified using oligo (dT) beads (Invitrogen Life Science Technologies) following the manufacturers protocol. Illumina sequencing libraries were prepared from messenger RNA using the Illumina sample prep kit. Sequencing produced between 5.7-16.7 million, 36 base pair reads per strain.

Identification of single nucleotide polymorphisms (SNPs). Messenger RNA-Seq reads from each strain were mapped to version 9 of the reference strain (*N. crassa* FGSC 2489) (Galagan et al. 2003) genome assembly. Mapping was performed using the program MOSAIK (version 1)(Hillier et al. 2008) allowing up to 4 mismatches per 36 bp read. Reads that did not map uniquely were discarded. Read alignments from each strain were pooled and SNPs were identified using a Bayesian approach implemented in the program GigaBayes (Hillier et al. 2008). To be included in the final set of high-quality SNPs, a candidate site was required to be biallelic and needed to meet or exceed the following criteria: coverage of five reads per allele, individual base qualities of 10, aggregate base qualities of 40, and Bayesian genotype probability of 0.90. To further reduce the number of potential false positives, singletons were discarded. Sites with missing data (i.e. the allele of one or more individuals was unidentifiable because it did not meet the above criteria) were excluded from analysis. Using these criteria, we found 5640 genes that had at least one SNP out of the approximately 9800 genes in the genome. These 5640 genes had an average of 14.4 SNPs per gene. After a preliminary phylogenetic analysis, the dataset was clone-corrected by discarding strains with nearly identical genotypes (i.e. those on very short branches). The false-positive rate was estimated by running the SNP-calling pipeline on five lanes of RNA-Seq reads from biological replicates of the reference strain (*N. crassa* FGSC 2489) (Galagan et al. 2003).

Identification of population structure. The presence of population structure was inferred based on the results from Bayesian clustering of allele frequencies and phylogenetic inference. Clustering of allele frequencies was performed using the program Bayesian Analysis of Population Structure (BAPS)(Corander et al. 2008) on a set of 3590 unlinked (separated by at least 5 kilobases) sites. The phylogeny was estimated using the same set of unlinked sites as well as the entire dataset using the Bayesian method implemented in Mr. Bayes (Huelsenbeck and Ronquist 2001). We also inferred a phylogeny for the same dataset except excluding sites within the two genomic islands of divergence using the combined rapid bootstrap and maximum-likelihood search algorithm in RAxML on the CIPRES webserver (Stamatakis et al. 2008).

Analysis of population demographics. Demographic parameters were estimated from the Louisiana and Caribbean joint allele frequency spectrum using a diffusion-based approach implemented in the program ∂a∂i (Gutenkunst et al. 2009). Singletons were masked because their identification is prone to false positives. The joint frequency spectrum data were polarized using the *N. tetrasperma* genome sequence (FGSC 2508) and were fit to the split_mig demographic model (customized to allow for asymmetric migration rates). To control for the potential misidentification of ancestral states, we fit the model to two additional datasets: one where we used two outgroups (*N. tetrasperma* FGSC #2508 (http://genome.jgi-psf.org/Neute_matA2/Neute_matA2.home.html) and *N. discreta* FGSC #8579 (http://genome.jgi-psf.org/Neudi1/Neudi1.home.html)) and one where we applied a correction that is part of the ∂a∂i package. The results were nearly identical (Fig. S2) and we report the parameters estimated from the uncorrected spectrum. The maximum likelihood estimates of the demographic parameters of this model were inferred using the Nelder-Mead optimization method in ∂a∂i and the resulting estimates were used to simulate 100 datasets in *ms* (Hudson 2002). As in (Gutenkunst et al. 2009) and (Murray et al. 2010) these simulations were used to check the goodness of fit of the model to the data and to generate uncertainty estimates for the demographic parameters. Divergence time in generations was converted to real time using a mutation rate estimate of 7.82 x 10^{-9} obtained from the average sequence divergence (7.27%) and divergence time (4.65 MYA) between *N. crassa* and *N. tetrasperma* (Menkis et al. 2008).

Identifying signals of positive selection. We performed sliding window calculations of Tajima's D and θπ within each population to identify signals of recent selective sweeps. To identify genes showing an excess of amino acid substitutions, we conducted the McDonald-Kreitman (MK) test using *N. tetrasperma* as an outgroup and the perl script used in (Holloway et al. 2007). We filtered the results to exclude genes that showed limited variability as explained in (Fay et al. 2001; 2002) and corrected for multiple hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg 1995).

Genome scans for divergence. We calculated each divergence measure for 10 kb overlapping windows across the genome. We used the method described in (Hudson et al. 1992) to calculate FST and the Bio::PopGen module in BioPerl (Stajich et al. 2002; Stajich and Hahn 2005) to calculate θπ and Tajima's D. We calculated Dxy as the average number of pairwise differences between populations using a custom perl script. We set a cutoff for outliers for each parameter by examining the distribution of values from each window and finding the value that is larger than 99.5% of the data (the 0.5% quantile).

Temperature data

The mean annual minimum and maximum temperatures of Homestead, Florida and Welsh, Louisiana (calculated for the time period 1971-2000) were accessed via the PRISM Climate Group website (PRISM 2003) (version 2 of the 800m PRISM data set).

Growth rate assays. All strains used in the growth rate assays were *a* mating type. Each strain was germinated on a Vogel's minimal medium (VMM) (Vogel 1956) petri dish and, after growing overnight, a hyphal plug was transferred to a race tube (Ryan et al. 1943) containing VMM. The location of the hyphal front was recorded at regular intervals until it reached the other end of the tube. Each strain was grown in triplicate in constant darkness inside 25°C and 10°C incubators. Crosses involving null mutants were made on synthetic crossing medium (Westergaard and Mitchell 1947) to the fluffy mating type tester strain. The fluffy strain contains a mutation at a single locus which makes it aconidate and highly fertile (Lindegren 1933). All progeny used in growth rate assays were screened to ensure that they produced macroconidia and thus did not have the fluffy mutation.

<u>Supplemental Figures</u>

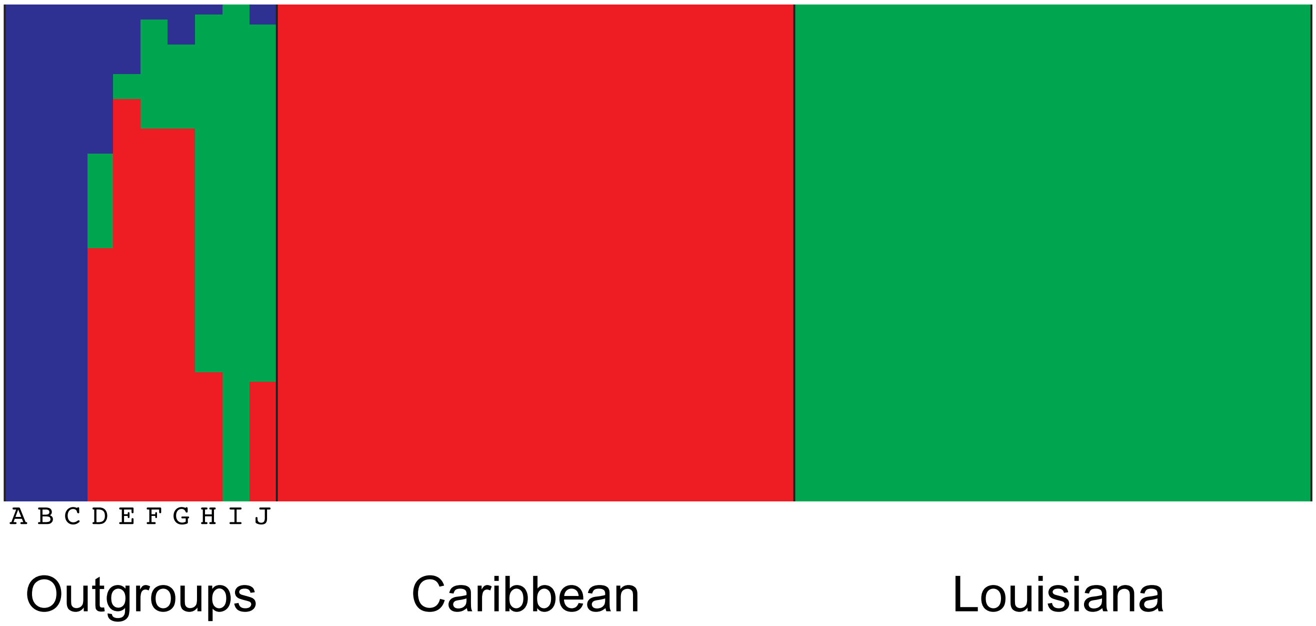

Figure S1. Bayesian Analysis of Population Structure (BAPS) Admixture Clustering Using 3590 unlinked sites, we clustered isolates by allele frequencies for K (number of partitions/populations) equal to 1 through 10. K=3 had the highest log-likelihood and is shown here. Each population is represented by a different color and each individual corresponds to a vertical bar. Bars are split into different colors if there is evidence of admixture. Outgroups: A-D: Panama, E: Costa Rica, F: Ivory Coast, G: Costa Rica, H: Guyana, I: Venezuela, J: Guyana.

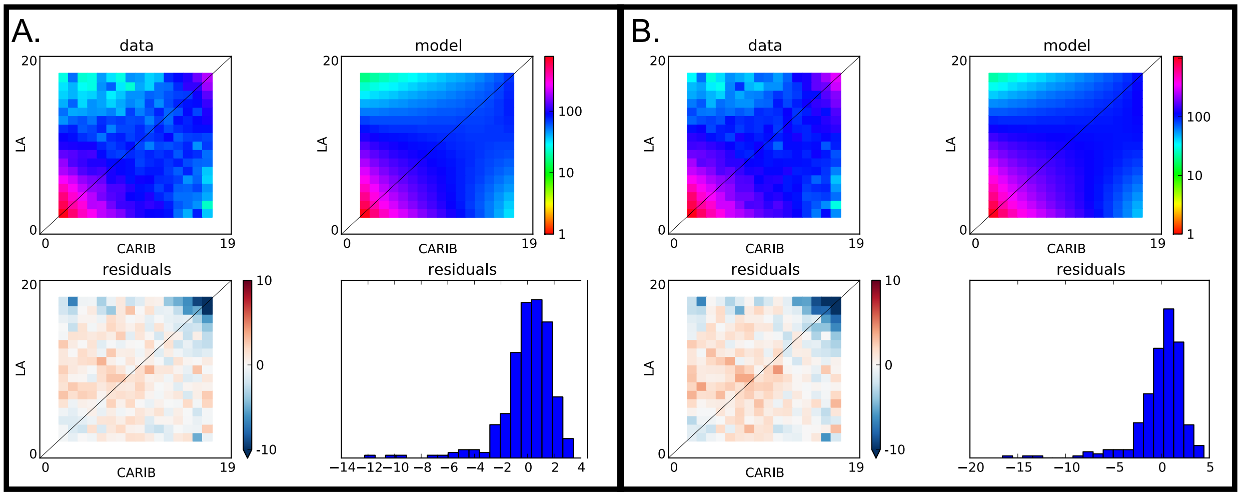

Figure S2. Assessing the possibility of misidentification of ancestral states in ∂a∂i. To determine if the excess of high frequency derived alleles observed in our data could be due to misidentification of ancestral alleles, we fit the isolation with migration model to two additional datasets. In (A) we included an additional outgroup (*N. discreta* FGSC #8579) and restricted our set of SNPs to include only those where the two outgroups had the same allele. In (B) we used a correction for ancestral misidentification that is part of the ∂a∂i package. The excess of high frequency derived alleles is still present in both cases suggesting this pattern is not an artifact.

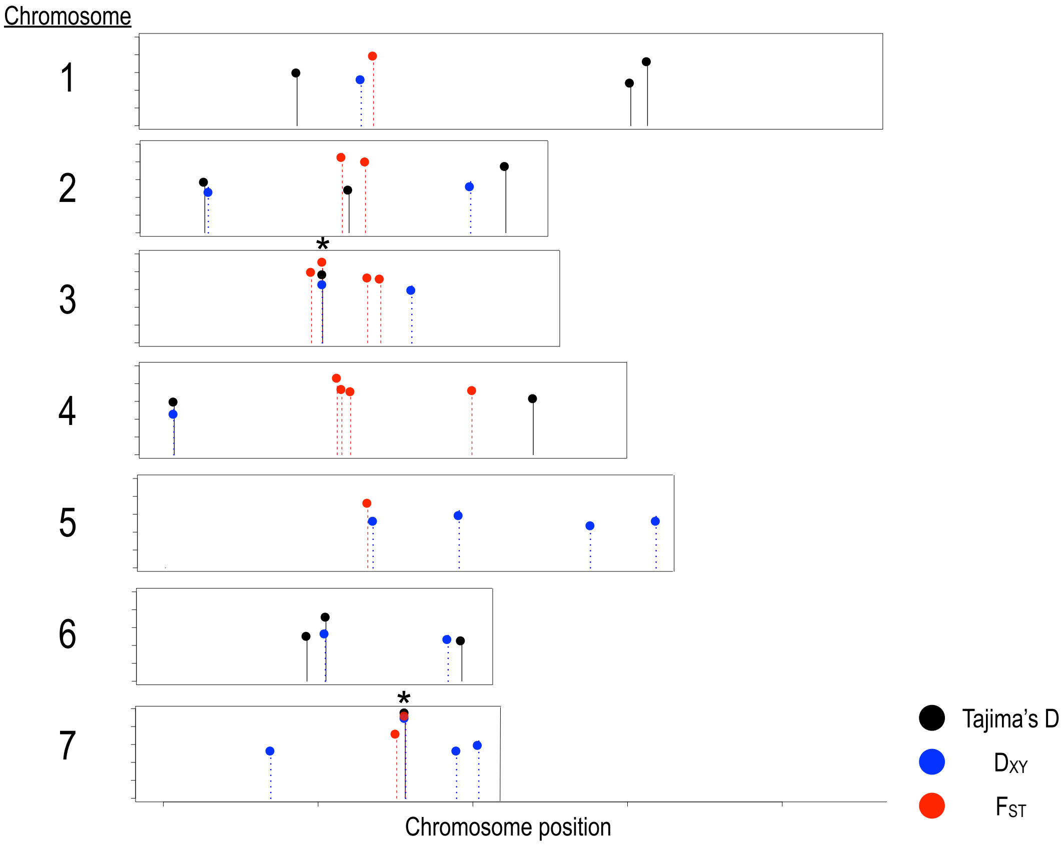

Figure S3. Genomic regions showing significant interpopulation divergence.
Three population genetic parameters were used to quantify interpopulation divergence: FST (shown as dashed red lines), Tajima's D (solid black lines), and Dxy (dotted blue lines). Values of each measure have been scaled by their maximum and only significant outliers are shown (0.5% quantile). Asterisks denote regions that are supported by all three measures.

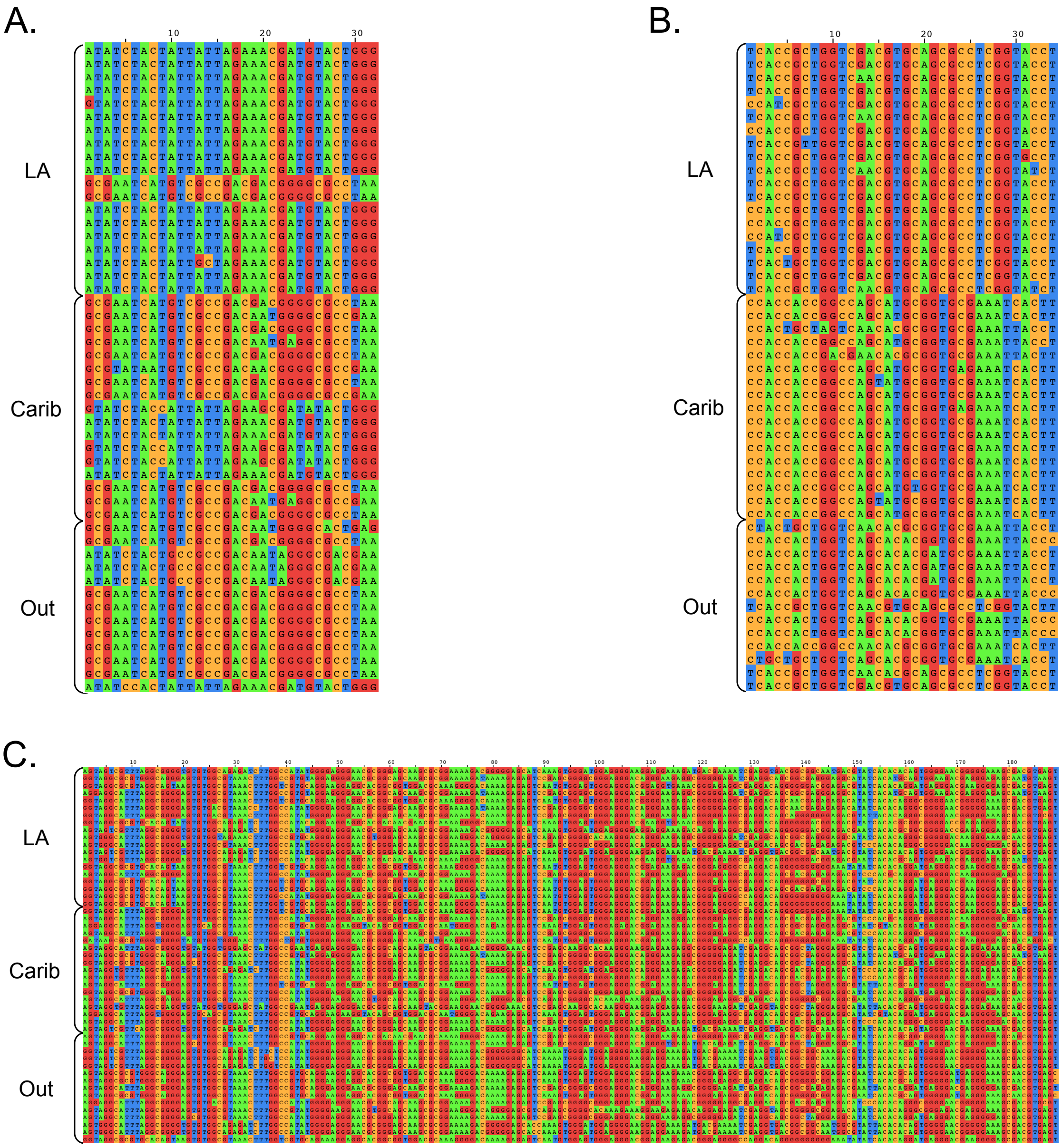

Figure S4. Examples of the three types of divergence outliers that were not truly distinct between populations.

Each matrix covers a 10 kb genomic region. Matrix rows contain the genotype for an individual while columns contain the alleles found at each polymorphic site. (A) This pattern was found in outliers for all three divergence measures and shows a genomic region where two distinct haplotypes are segregating within each population at opposite frequencies. (B) This pattern was found in F_{ST} divergence outliers and shows a region where haplotypes are fixed between populations but there are few polymorphic sites within the 10 kb genomic region (absolute divergence is not unusual). (C) This pattern was found in Dxy divergence outliers and shows a genomic region that is unusually SNP-dense but relative divergence between populations is not unusual.

Regions shown: A: chr 1: 2541000-2551000; B: chr 4: 2221000-2231000; C: chr 1: 6281000-6291000.

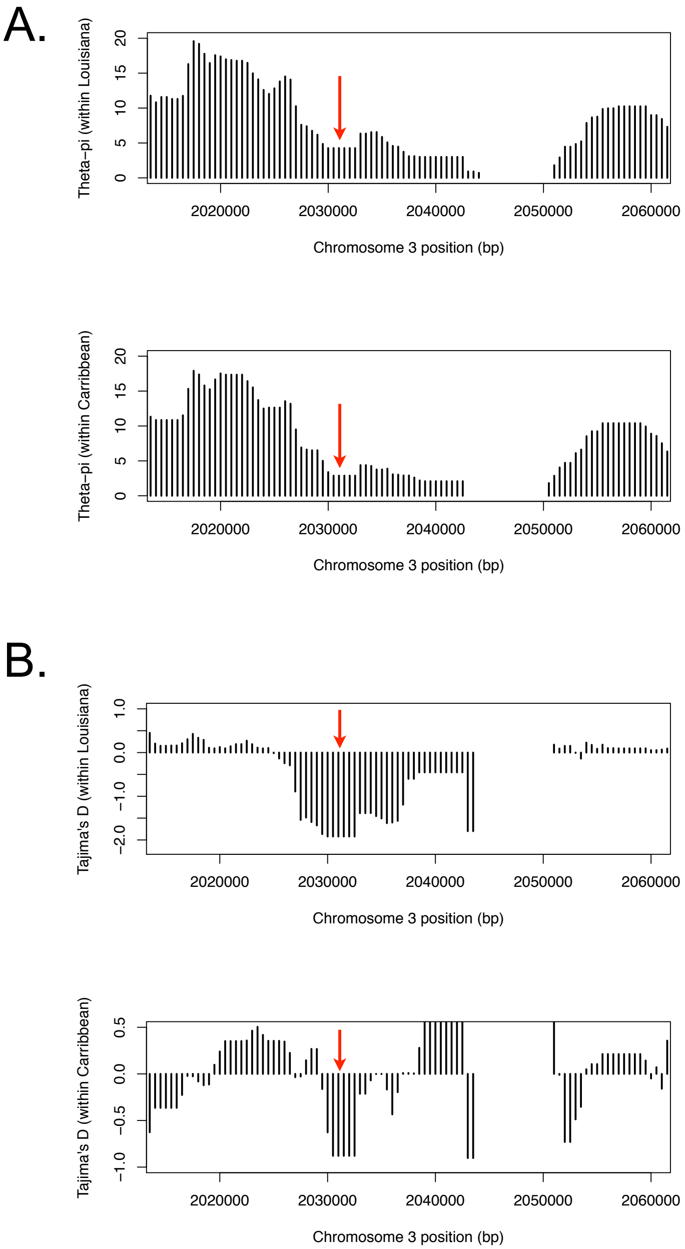

Figure S5. Within-population Tajima's *D* and θπ for the island-like region on chromosome three.
Each graph shows the results for a 10 kb sliding window calculation centered on the outlier region on chromosome three. Each calculation was performed within the Louisiana and Caribbean populations separately. The average number of pairwise differences (θπ) is shown in (A) while the within-population value of Tajima's *D* is shown in (B). Each vertical bar represents the value for a single 10 kb window and the red arrow points to the first window that encompasses the outlier region. Regions without vertical bars represent windows whose number of polymorphic sites falls below an arbitrary threshold (Tajima's *D*: 10, θπ: 5). Here, both the Louisiana population and the Caribbean population show an excess of low frequency alleles and a reduction in nucleotide diversity around the outlier region, consistent with independent selective sweeps within each population.

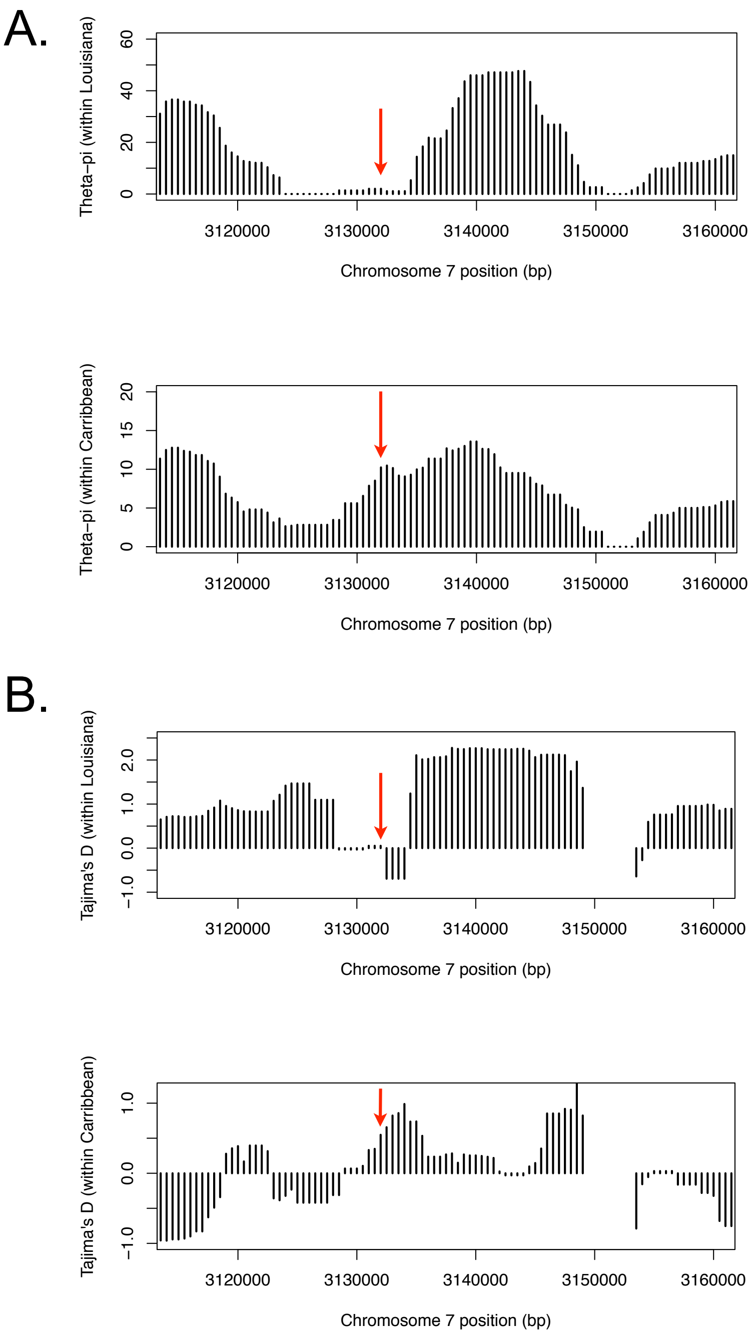

Figure S6. Within-population Tajima's *D* and $\theta\pi$ for the island-like region on chromosome seven.

Each graph shows the results for a 10 kb sliding window calculation centered on the chromosome seven outlier region. Each calculation was performed within the Louisiana and Caribbean populations separately. The average number of pairwise differences ($\theta\pi$) is shown in (A) while the within-population value of Tajima's *D* is shown in (B). Each vertical bar represents the value for a single 10 kb window and the red arrow points to the first window that encompasses the outlier region. Regions without vertical bars represent windows whose number of polymorphic sites falls below an arbitrary threshold (Tajima's *D*: 10, $\theta\pi$: 5). By comparing the graphs, it is apparent that the Louisiana population shows an excess of low frequency alleles and a reduction in nucleotide diversity around the outlier region, consistent with a recent selective sweep, while the Caribbean population does not.

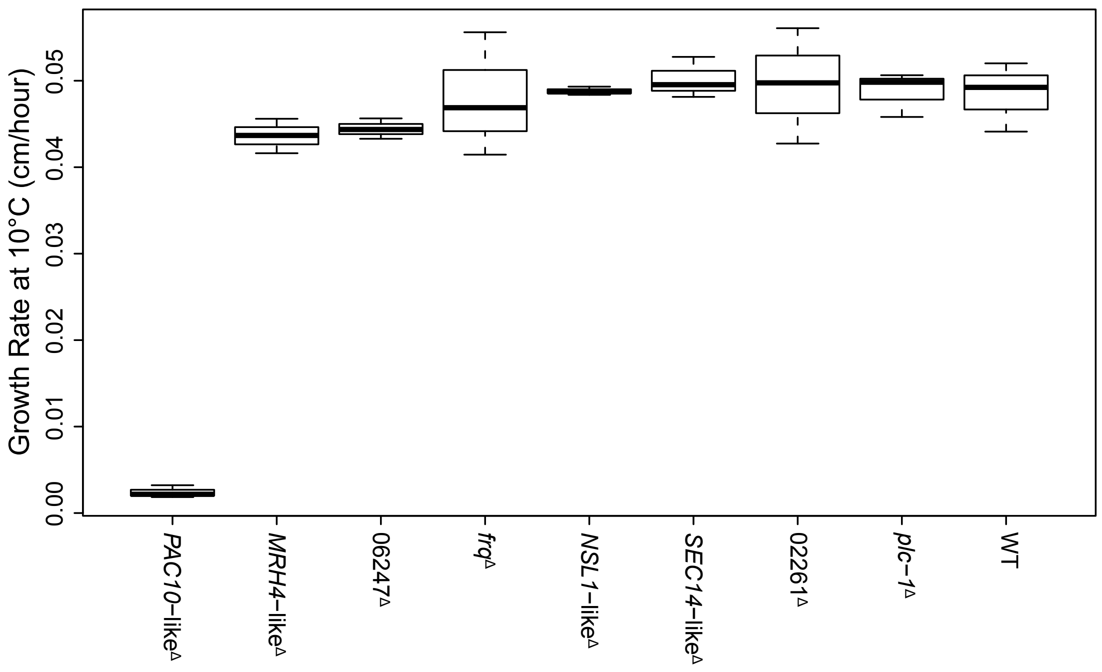

Figure S7. Growth rate at 10°C of null mutants for genes within divergence islands
Null mutants for each gene within the two divergence islands were grown in triplicate at 10°C along with the wild-type lab strain FGSC 2489.

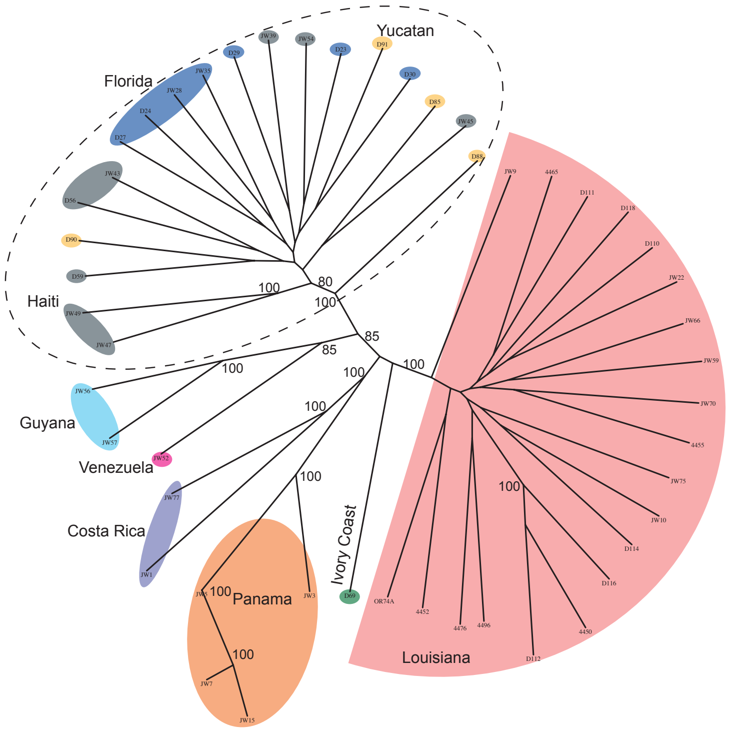

Figure S8. Phylogenetic support for population structure is maintained after excluding SNPs from within the genomic islands of divergence

This phylogeny was inferred with the same dataset as in Figure 1 except that the polymorphic sites residing within the two genomic islands of divergence were excluded. The tree was inferred using the RAxML combined rapid bootstrap and maximum-likelihood search algorithm. Bootstrap support values less than 0.80 are not shown. The colors shown here are equivalent to those in Figure 1.

Supplemental Tables

Strain Number	FGSC	Perkins	Mat	Strain provenance	Collection site	Substrate
D110	8870	4448	A	Dettman, J.	Franklin, Louisiana	Sugarcane
D111	8871	4449	a	Dettman, J.	Franklin, Louisiana	Sugarcane
D112	8872	4453	A	Dettman, J.	Franklin, Louisiana	Sugarcane
D114	8874	4464	A	Dettman, J.	Franklin, Louisiana	Sugarcane
D116	8876	4481	a	Dettman, J.	Franklin, Louisiana	Sugarcane
D118	8878	4491	a	Dettman, J.	Franklin, Louisiana	Sugarcane
D23	8783	1409	A	Dettman, J.	Homestead, Florida	grass
D24	8784	1410	A	Dettman, J.	Homestead, Florida	grass
D27	8787	1417	A	Dettman, J.	Homestead, Florida	grass
D29	8789	1465	A	Dettman, J.	Homestead, Florida	grass
D30	8790	1470	a	Dettman, J.	Homestead, Florida	grass
D56	8816	3424	A	Dettman, J.	Carrefour Dufort, Haiti	grass
D59	8819	3427	a	Dettman, J.	Carrefour Dufort, Haiti	Sugarcane
D69	8829	3684	a	Dettman, J.	Tiassale, Ivory Coast	grass
D85	8845	4130	a	Dettman, J.	Kabah, Yucatan, Mexico	soil isolation, unburnt
D88	8848	4150	a	Dettman, J.	Sayil, Yucatan, Mexico	soil isolation, unburnt
D90	8850	4154	A	Dettman, J.	Uxmal, Yucatan, Mexico	soil isolation, unburnt
D91	8851	4155	A	Dettman, J.	Uman, Yucatan, Mexico	soil isolation, unburnt
JW01	851		A	Welch.J	Costa Rica	unknown
JW03	1131		A	Welch, J.	Panama	unknown
JW05	1133		a	Welch, J.	Panama	unknown
JW07	1165		a	Welch, J.	Panama	unknown
JW09	2229		A	Welch, J.	Welsh.LA	burned grass
JW10	2229		A	Welch, J.	Welsh, LA	burned grass
JW15	1132		a	Welch,J	Panama	unknown
JW22	3223		A	Welch, J.	Elizabeth, LA	pine burn
JW28	3968		A	Welch,J	Okeechobee, FL	unknown
JW35	3975		a	Welch, J.	Florida	reeds, grass burn
JW39	4708		A	Welch, J.	Haiti	wood, grass burn
JW43	4712		a	Welch, J.	Haiti	sugarcane burn
JW45	4713		A	Welch, J.	Haiti	cane, grass burn
JW47	4715		a	Welch, J.	Haiti	cane, grass burn
JW49	4716		A	Welch, J.	Haiti	grass burn
JW52	4730		A	Welch, J.	Venezuela	grass burn
JW54	4824		A	Welch, J.	Haiti	bushes burn
JW56	5910		A	Welch, J.	Digitima Creek,Guiana	vegetation burn
JW57	5914		A	Welch, J.	Torani Canal,Guiana	palm wood burn
JW59	3200		a	Welch, J.	Coon.LA	burned stumps
JW66	3211		a	Welch, J.	Sugartown,LA	pine burn
JW70	3199		A	Welch, J.	Coon.LA	burned stumps
JW75	3943		a	Welch, J.	Houma,LA	sugarcane burn
JW77	6203		A	Welch, J.	Aguda Rd, Costa Rica	wood burn
		4450	a	Perkins, D.	Franklin, Louisiana	sugarcane burn
		4452	a	Perkins, D.	Franklin, Louisiana	sugarcane burn
		4455	a	Perkins, D.	Franklin, Louisiana	sugarcane burn
		4465	a	Perkins, D.	Franklin, Louisiana	sugarcane burn
		4476	a	Perkins, D.	Franklin, Louisiana	sugarcane burn
		4496	a	Perkins, D.	Franklin, Louisiana	sugarcane burn
NCU02261Δ	16978		a	FGSC	Marrero, Louisiana	unknown
NSL1-like NCU02262Δ	16925		a	FGSC	Marrero, Louisiana	unknown
SEC14-like NCU02263Δ	16212		a	FGSC	Marrero, Louisiana	unknown
PAC10-like NCU02264Δ	16061		a	FGSC	Marrero, Louisiana	unknown
frq NCU02265Δ	11554		a	FGSC	Marrero, Louisiana	unknown
plc-1 NCU06245Δ	11411		a	FGSC	Marrero, Louisiana	unknown
MRH4-like NCU06246Δ	12864		a	FGSC	Marrero, Louisiana	unknown
NCU06247Δ	12056		a	FGSC	Marrero, Louisiana	unknown
fluffy	4317		a	FGSC	Marrero, Louisiana	unknown
WT strain 74-OR8-1a	988		a	FGSC	Marrero, Louisiana	unknown
WT strain 74-OR23-1VA	2489		A	FGSC	Marrero, Louisiana	unknown

Table S1. Strains used in this study
All strains were provided by the Fungal Genetic Stock Center (FGSC).

Gene ID	Protein Function	LA *P*-value	Carib *P*-value
NCU01104	RNA helicase MSS116	*P*=0.027	NS
NCU02325	GMP synthase	*P*<0.005	*P*<0.005
NCU03539	ribonucleoside-diphosphate reductase large chain (un-24)	*P*<0.005	*P*=0.001
NCU04349	proteinase inhibitor PBI2, alpha-ketoacid dehydrogenase kinase	*P*<0.005	*P*<0.005
NCU04837	mitochondrial oxodicarboxylate carrier	*P*=0.005	*P*=0.014
NCU05006	fatty acid omega-hydroxylase (P450foxy)	*P*=0.003	*P*=0.034
NCU05425	oxoglutarate dehydrogenase precursor	*P*<0.005	*P*=0.002
NCU06871	glucan synthase	*P*=0.005	NS
NCU06877	phosphatidylinositol transfer protein	*P*=0.017	*P*=0.030
NCU02505	probable PYC2 Pyruvate carboxylase 2	*P*<0.005	*P*=0.012
NCU08336	succinate dehydrogenase (ubiquinone) flavoprotein precursor, mitochondrial	NS	*P*=0.016
NCU09209	galactose oxidase precursor	*P*=0.027	NS

Table S2. Genes under positive selection as identified by the McDonald-Kreitman (MK) Test.
The MK test identifies genes showing a significant excess of amino acid substitutions. The test was conducted using polymorphism information for each population separately. The *P*-values shown are the result of using the Benjamini-Hochberg correction for multiple hypothesis testing using a cutoff of *P*=0.05. LA=Louisiana, Carib=Caribbean, NS=Not Significant.

References

Andolfatto P. 2001. Adaptive hitchhiking effects on genome variability. Current Opinion in Genetics & Development 11(6):635-641.

Aronson BD, Johnson KA, Loros JJ, Dunlap JC. 1994. Negative feedback defining a circadian clock: autoregulation of the clock gene *frequency*. Science 263(5153):1578-84.

Beaumont MA. 2005. Adaptation and speciation: what can F-st tell us? Trends in Ecology & Evolution 20(8):435-440.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological 57(1):289-300.

Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, Polato NR, Olsen KM, Nielsen R, McCouch SR, et al. 2007. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genetics 3(9):1745-56.

Colot HV, Park G, Turner GE, Ringelberg C, Crew CM, Litvinkova L, Weiss RL, Borkovich KA, Dunlap JC. 2006. A high-throughput gene knockout procedure for *Neurospora* reveals functions for multiple transcription factors. Proceedings of the National Academy of Sciences of the United States of America 103(27):10352-10357.

Corander J, Marttinen P, Siren J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics 9:539.

Dettman JR, Anderson JB, Kohn LM. 2008. Divergent adaptation promotes reproductive isolation among experimental populations of the filamentous fungus *Neurospora*. BMC Evolutionary Biology 8:35.

Dettman JR, Jacobson DJ, Taylor JW. 2003a. A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote *Neurospora*. Evolution 57(12):2703-20.

Dettman JR, Jacobson DJ, Taylor JW. 2006. Multilocus sequence data reveal extensive phylogenetic species diversity within the *Neurospora discreta* complex. Mycologia 98(3):436-46.

Dettman JR, Jacobson DJ, Turner E, Pringle A, Taylor JW. 2003b. Reproductive isolation and phylogenetic divergence in *Neurospora*: comparing methods of species recognition in a model eukaryote. Evolution 57(12):2721-41.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics 158(3):1227-34.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature 415(6875):1024-6.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422(6934):859-868.

Gavric O, dos Santos DB, Griffiths A. 2007. Mutation and divergence of the phospholipase C gene in Neurospora crassa. Fungal Genetics and Biology 44(4):242-249.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5(10):e1000695.

Harr B. 2006. Genomic islands of differentiation between house mouse subspecies. Genome Research 16(6):730-7.

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang WC, et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. Nature Methods 5(2):183-188.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genetics 6(2):e1000862.

Holloway AK, Lawniczak MKN, Mezey JG, Begun DJ, Jones CD. 2007. Adaptive gene expression divergence inferred from population genomics. PLoS Genetics 3(10):2007-2013.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18(2):337-8.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132(2):583-9.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754-5.

Hunger K, Beckering CL, Wiegeshoff F, Graumann PL, Marahiel MA. 2006. Cold-induced putative DEAD box RNA helicases CshA and CshB are essential for cold adaptation and interact with cold shock protein B in *Bacillus subtilis*. Journal of Bacteriology 188(1):240-8.

Jackson ST, Webb RS, Anderson KH, Overpeck JT, Webb T, Williams JW, Hansen BCS. 2000. Vegetation and environment in Eastern North America during the Last Glacial Maximum. Quaternary Science Reviews 19(6):489-508.

Jiggins CD, McMillan WO. 1997. The genetic basis of an adaptive radiation: warning colour in two *Heliconius* species. Proceedings of the Royal Society of London Series B-Biological Sciences 264(1385):1167-1175.

Kim JS, Kim KA, Oh TR, Park CM, Kang H. 2008. Functional characterization of DEAD-Box RNA helicases in *Arabidopsis thaliana* under abiotic stress conditions. Plant and Cell Physiology 49(10):1563-1571.

Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse Ecology" and the power of population genomics. Evolution 62(12):2984-2994.

Lindegren CC. 1933. The genetics of *Neurospora* III: Pure bred stocks and crossing over in *N. Crassa* Bulletin of the Torrey Botanical Club 60(3):133-154.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458(7236):337-341.

Liu Y, Bell-Pedersen D. 2006. Circadian rhythms in *Neurospora crassa* and other filamentous fungi. Eukaryotic Cell 5(8):1184-1193.

McClung CR, Fox BA, Dunlap JC. 1989. The *Neurospora* clock gene *frequency* shares a sequence element with the *Drosophila* clock gene *period*. Nature 339(6225):558-62.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351(6328):652-4.

Menkis A, Bastiaans E, Jacobson DJ, Johannesson H. 2009. Phylogenetic and biological species diversity within the *Neurospora tetrasperma* complex. Journal of Evolutionary Biology 22(9):1923-1936.

Menkis A, Jacobson DJ, Gustafsson T, 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.

Merrow M, Brunner M, Roenneberg T. 1999. Assignment of circadian function for the *Neurospora* clock gene *frequency*. Nature 399(6736):584-6.

Metzenberg RL. 2004. Bird Medium: an alternative to Vogel Medium. Fungal Genetics Newsletter 51:19-20.

Morton BR, Dar V-u-N, Wright SI. 2009. Analysis of site frequency spectra from *Arabidopsis* with context-dependent corrections for ancestral misinference. Plant Physiology 149(2):616-624.

Murray C, Huerta-Sanchez E, Casey F, Bradley DG. 2010. Cattle demographic history modelled from autosomal sequence variation. Proceedings of the Royal Society of London Series B-Biological Sciences 365(1552):2531-9.

Nachman MW, Hoekstra HE, D'Agostino SL. 2003. The genetic basis of adaptive melanism in pocket mice. Proceedings of the National Academy of Sciences of the United States of America 100(9):5268-5273.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Noor MA, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity 103(6):439-44.

Nowrousian M, Stajich JE, Chu ML, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC, et al. 2010. De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genetics 6(4):e1000891.

Petit JR, Jouzel J, Raynaud D, Barkov NI, Barnola JM, Basile I, Bender M, Chappellaz J, Davis M, Delaygue G, et al. 1999. Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. Nature 399(6735):429-436.

Pool JE, Aquadro CF. 2007. The genetic basis of adaptive pigmentation variation in *Drosophila melanogaster*. Molecular Ecology 16(14):2844-2851.

Pool JE, Nielsen R. 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181(2):711-9.

Powell AJ, Jacobson DJ, Salter L, Natvig DO. 2003. Variation among natural isolates of *Neurospora* on small spatial scales. Mycologia 95(5):809-819.

PRISM. 2003. Oregon Climate Service, Oregon State University. http://www.prismclimate.org.

Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL. 2009. De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. Genome Research 19(2):294-305.

Ryan FJ, Beadle GW, Tatum EL. 1943. The tube method of measuring the growth rate of *Neurospora*. American Journal of Botany 30((10)):784-799.

Schade B, Jansen G, Whiteway M, Entian KD, Thomas DY. 2004. Cold adaptation in budding yeast. Molecular Biology of the Cell 15(12):5492-5502.

Schmitt S, Prokisch H, Schlunck T, Camp DG, Ahting U, Waizenegger T, Scharfe C, Meitinger T, Imhof A, Neupert W, et al. 2006. Proteome analysis of mitochondrial outer membrane from *Neurospora crassa*. Proteomics 6(1):72-80.

Skelly DA, Ronald J, Connelly CF, Akey JM. 2009. Population genomics of intron splicing in 38 *Saccharomyces cerevisiae* genome sequences. Genome Biology and Evolution:466-478.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Research 12(10):1611-8.

Stajich JE, Hahn MW. 2005. Disentangling the effects of demography and selection in human history. Molecular Biology and Evolution 22(1):63-73.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Systematic Biology 57(5):758-71.

Storz JF, Runck AM, Sabatino SJ, Kelly JK, Ferrand N, Moriyama H, Weber RE, Fago A. 2009. Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. Proceedings of the National Academy of Sciences of the United States of America 106(34):14450-14455.

Storz JF, Wheat CW. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. Evolution 64(9):2489-2509.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585-95.

Taylor JW, Turner E, Townsend JP, Dettman JR, Jacobson D. 2006. Eukaryotic microbes, species recognition and the geographic limits of species: examples from the kingdom Fungi. Philosophical Transactions of the Royal Society B-Biological Sciences 361(1475):1947-1963.

Tsai IJ, Bensasson D, Burt A, Koufopanou V. 2008. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. Proceedings of the National Academy of Sciences of the United States of America 105(12):4957-4962.

Turner BC, Perkins DD, Fairfield A. 2001. *Neurospora* from natural populations: a global study. Fungal Genetics and Biology 32(2):67-92.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nature Genetics 42(3):260-U42.

Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biology 3(9):e285.

Turner TL, Levine MT, Eckert ML, Begun DJ. 2008. Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. Genetics 179(1):455-473.

Vogel HJ. 1956. A convenient growth medium for *Neurospora* (Medium N). Microbial Genetics Bulletin 13:42–43.

Westergaard M, Mitchell HK. 1947. *Neurospora*. V. A synthetic medium favoring sexual reproduction. American Journal of Botany 34(10):573-577.

Wright S. 1950. Genetical Structure of Populations. Nature 166(4215):247-249.

Chapter 3

Inference of regulatory networks within and gene expression divergence between populations of the model fungus *Neurospora crassa*

Abstract

Examination of natural variation in gene expression levels within and between lineages is beginning to provide important insight into transcriptome evolution, much as the study of population genetics has increased our knowledge of genome evolution. However, the study of gene expression regulation is a much younger field. For example, although there are hundreds of species of fungi with sequenced genomes, few have standardized measures of transcription levels that are comparable between lineages, and fewer still have measures of variation in transcription between individuals from the same population. Here we used Illumina RNA-Seq to examine variation in gene expression patterns during active growth of the model filamentous fungus *Neurospora crassa*. We examined genome-wide expression profiles from a total of 67 wild isolates, 48 from Louisiana and 19 from the Caribbean, and found that ~17% of the genes were differentially expressed between these two populations which diverged approximately 0.5 MYA. We also identified a set of candidate genes whose change in expression patterns between these two populations may be due to directional selection and a suite of genes involved in nitrogen metabolism that appear to have been specifically up-regulated in the Louisiana population. We additionally harness the natural variation in gene expression levels within the Louisiana population to identify regulatory modules, including candidate targets of a previously uncharacterized homeobox transcription factor.

Introduction

Transcription level bridges the path from genotype to phenotype and variation at this juncture plays an important role in phenotypic evolution. With the advent of the microarray and, more recently, next generation sequencing technology, it has become possible to measure genome-wide transcription levels across many different samples. Early microarray experiments focused on measuring differences in gene expression between tissues (Schena et al. 1995) or across experimental conditions (Lashkari et al. 1997) using a single strain or accession. As the technology and algorithms for quantitating gene expression have been optimized, new approaches for utilizing these types of data have been developed, including the comparison of expression divergence between species and across individuals within populations (reviewed in (Ranz and Machado 2006) and (Fay and Wittkopp 2008)), mapping of the genetic architecture underlying variation in transcript levels (reviewed in (Rockman and Kruglyak 2006)), and the use of correlated expression levels between genes to identify regulatory networks and modules (Nayak et al. 2009; Yvert et al. 2003).

Yeast has played a major role in efforts to understand mechanisms influencing transcriptional regulation as well as the genetic architecture underlying natural variation in gene expression levels (e.g. (Brem et al. 2002; Gasch et al. 2000; Gerke et al. 2009; Harbison et al. 2004; Hughes et al. 2000; Lee et al. 2002; Nagalakshmi et al. 2008)). In fact, the Ascomycete fungi in general represent an ideal system for studying the evolution of regulatory networks (Gasch et al. 2004; Thompson and Regev 2009; Wohlbach et al. 2009) as there are over 500 species of Ascomycetes with genome sequencing projects either in progress or already completed (Kyrpides 1999) that together represent over 500 million years of evolution (Berbee and Taylor 2010). However, while microarray or expressed sequence tag (EST) based measurements of transcription exist for some of these species, few besides *S. cerevisiae* and other members of the hemiascomycete clade have measurements of transcription levels within and between populations.

A prime candidate for exploration of transcriptional variation between wild isolates is the filamentous ascomycete *Neurospora crassa*. *N. crassa* has served as a model organism for dissecting the genetic basis of circadian rhythm (Loros and Dunlap 2001), self/non-self recognition (Glass and Dementhon 2006), RNA silencing (Fulci and Macino 2007; Shiu and Metzenberg 2002), and biomass degradation (Tian et al. 2009). This fungus is also beginning to emerge as a model for understanding the evolutionary genetics and genomics of wild populations (Dettman et al. 2003a; Dettman et al. 2003b; Ellison et al. 2011; Jacobson et al. 2004; Turner et al. 2010; Whittle et al. 2011). Well-assembled and high-quality genome sequences are available for three species within the genus: *N. crassa*, *N. tetrasperma* and *N. discreta*, thousands of wild isolates have been collected from around the world and are available from the Fungal Genetic Stock Center, several well-characterized phylogenies exist for the genus (Dettman et al. 2003a; 2006; Menkis et al. 2009; Villalta et al. 2009), and there is a nearly complete gene deletion collection in *N. crassa* (Colot et al. 2006; Dunlap et al. 2007). More recently, a population genomic analysis by Ellison *et al* (Ellison et al. 2011) discovered two cryptic and recently diverged populations of *N. crassa* from the Caribbean basin, one endemic to Louisiana and the other containing isolates from Florida, Haiti, and the Yucatan. This study also identified two genomic regions that are extremely divergent between the two populations and that contain genes whose pattern of sequence variation and null phenotype are consistent with adaptation to low temperature.

Here we explore genome-wide variation in gene expression within and between these same populations. We have identified genes that are differentially expressed between the two populations, including a set of genes involved in nitrogen metabolism that appear to be specifically and strongly up-regulated in the Louisiana population as well as a set of candidate genes whose expression divergence may be due to positive selection. In addition, we inferred candidate regulatory modules from clusters of coexpressed genes within the Louisiana population and assessed the utility of these clusters for predicting transcription factor targets. We performed Illumina RNA-seq profiling on several strains containing deletions in genes encoding predicted transcription factors and, for each one, we compared the set of genes that were differentially expressed in the deletion strain to the set of genes that were coexpressed with the transcription factor in the Louisiana population. For one of these, the uncharacterized homeobox transcription factor NCU05257, we found a highly significant overlap between these two gene sets.

Results

Gene expression measurements

We measured transcription levels across a total of 67 *N. crassa* isolates, 48 from Louisiana and 19 from the Caribbean. These measures were taken from entire colonies after ~20 hours of active growth on Bird minimal medium (Metzenberg 2004) and we measured expression for 8,823 of the 9,733 predicted *N. crassa* genes.

Gene expression clusters

Clustering of coexpressed genes is a common method of analyzing gene expression data and has been shown to be useful for inferring gene networks (Nayak et al. 2009) and for predicting functions of uncharacterized genes (Eisen et al. 1998). Here we have clustered genes whose expression is correlated between Louisiana individuals. To assess the effect of using different correlation coefficient cutoffs in the clustering algorithm, we inferred clusters using five cutoffs ranging from 0.2 to 0.6 in 0.1 increments. To determine the size of clusters expected by chance at each of these cutoffs, we inferred clusters from datasets where expression measurements were permuted between individuals. The minimum size of clusters found in the real data but not in the permuted data varied from 45 genes at a cutoff of 0.2 to four genes at a cutoff of 0.6 (Table S1). After eliminating clusters with sizes below or equal to those found in the permutated data at the same correlation coefficient cutoff, the number of clusters remaining varied from 49 at a cutoff of 0.2 to 356 at a cutoff of 0.6 (Table 1). We chose the dataset created using the 0.4 correlation coefficient cutoff for further analysis because it represents a balance between the total number of clusters created and the total number of genes included in clusters (Table 1).

To investigate the utility of the clusters for inferring regulatory networks and functional information about uncharacterized genes in *Neurospora*, we assessed whether they contained genes of related function. *N. crassa* genes have been associated with Functional Catalogue (FunCat) terms by the Munich Information Center for Protein Sequences (MIPS) (Ruepp et al. 2004). We used these terms to determine if the clusters contain genes of related function by performing functional enrichment tests for each cluster of nine or more genes. We found that 84% (142 out of 170) of the clusters with at least two genes with FunCat terms were significantly enriched for one or more functional terms (hypergeometric *P*<=0.1). In all, approximately 40% of the genes that we were able to cluster (2244 out of 5526) are annotated as *hypothetical protein*. Because these clusters are enriched for particular functions, we conclude that they represent a valuable resource for the inference of functional information for these uncharacterized genes based on their clustering with genes of known function.

Expression profiling of transcription factor deletion strains

The vast majority of annotated transcription factors in *N. crassa* have no known function or target set. To investigate the utility of these clusters for making inferences about regulatory networks, we assessed whether they could be used to predict transcription factor target genes. We predicted that coexpression clusters containing a transcription factor should also be enriched for targets of that factor. To test this prediction, we obtained deletion strains for four transcription factors chosen at random from four different clusters. For each strain, we extracted RNA from an entire colony after ~20 hours of active growth on Bird medium (Metzenberg 2004), and used Illumina RNA-seq to compare gene expression levels between the deletion mutant and the wild-type strain. The four factors that we investigated are: NCU05257 [uncharacterized homeobox gene], NCU05767 [divergent paralog of the *Sordaria macrospora pro-1* gene (Masloff et al. 2002)], NCU07039 [*asd-4*: defective in ascus development (Feng et al. 2000)], and NCU07705 [uncharacterized, light-induced Zn(2)-Cys(6) transcription factor (Smith et al. 2010)]. Other than the ascus and ascospore defects previously observed for the *asd-4* mutant (Feng et al. 2000), no morphological defects have been reported for mutants in any of these genes and no obvious defects were observed in our experience with these cultures. For two of the four transcription factor deletion strains, we found overlap between genes that we identified as differentially expressed in the deletion strain and genes that clustered with the TF, however the overlap was statistically significant for only one of these, the homeobox containing transcription factor NCU05257 (Fisher's exact test: *P*=1.84e-09)(Table 2 and Table 3).

NCU05257 has previously been identified as a putative target of the bZIP transcription factor *cpc-1* (ortholog of yeast *GCN4*) (Tian et al. 2007) and, in the wild strains from Louisiana, its expression is highly correlated with that of two aminotransferases, a FAD-dependent oxidoreductase, MFS monocarboxylate transporter, and cystathionine beta-lyase, as well as three other genes (Table 3), all of which are differentially expressed when this factor is deleted. The presence of a cystathionine beta-lyase and two aminotransferases among the putative targets of this transcription factor, combined with the results implicating this factor as a target of *cpc-1* (Tian et al. 2007), suggest that it may also play a role in amino acid biosynthesis or the response to amino acid starvation. Although only one of the four transcription factors examined here showed a significant overlap between the genes that were differentially expressed in the deletion strain and the genes that were coexpressed with the factor in the Louisiana population, we conclude that the coexpression clusters represent a useful starting point for generating predictions about the targets of uncharacterized regulators of transcription.

Differential expression between populations

The Louisiana strains are genetically divergent from a neighboring population of Caribbean isolates (Ellison et al. 2011). To survey expression divergence between these populations on a genomic scale, we performed RNA-seq on an additional 29 Louisiana strains, resulting in a total of 48 isolates from Louisiana and 19 from the Caribbean. We calculated significance estimates for differential expression for each gene and identified a total of 1539 genes as being differentially expressed (DE) between the two populations (wilcoxon test, Benjamini-Hochberg corrected *P*<=0.05), including five of the seven genes located in the genomic regions of extreme divergence identified in (Ellison et al. 2011)(Table S2). Compared to the entire set of expressed genes, the set of DE genes is enriched for several functional categories, the most significant of which are *RNA processing* and *rRNA processing* (Benjamini-Hochberg corrected P<0.001 in both cases)(Table S3). Constraining our total gene set to those that we identified as being differentially expressed between populations, we identified 15 functional terms that were enriched in the set of genes with higher expression in the Louisiana population and 22 enriched in the set of genes with higher expression in the Caribbean population (Fig. 1).

We next aimed to identify candidate genes whose change in expression between the two populations may have been due to positive selection. We used the same approach that has been implemented in several previous studies (Blekhman et al. 2008; Gilad et al. 2006; Meiklejohn et al. 2003; Nuzhdin et al. 2004) where putative adaptive changes in gene expression are identified as those cases where variation in gene expression level is low within populations and divergence in expression level is high between populations. We additionally inferred the ancestral and derived states for the genes showing this pattern using the Panamanian population of *N. crassa* as an outgroup. The strongest imaginable divergence in expression between these populations would be a case where the within-population distributions of expression levels for a particular gene are completely non-overlapping between populations. We see no instances of such a case for the genes measured here, presumably because of the recent divergence of these populations and the relatively large amount of ancestral polymorphisms shared between them (~50%) (Ellison et al. 2011). We did, however, find 154 genes with non-overlapping inter-quartile ranges, 86 of which are also among the top 5% of genes with the largest ratio of interpopulation variation to between population divergence in expression level (Table S4).

From the subset of these 86 genes for which some aspect of function is known, there are several interesting groups that stand out, including three genes known to be involved in chromatin remodeling: an uncharacterized gene that contains a functional domain homologous to that of the yeast *SPT2* negative regulator of transcription, a homolog of the yeast *NHP6A/B* nucleosome remodeling HMG proteins, and a GNAT family acetyltransferase, and several cytoskeleton related genes including a microtubule-associated CRIPT family protein, a *STU1*-like component of the mitotic spindle, and septin (Table S4). In addition, three of the 86 genes are Zn(II)2Cys6 binuclear cluster transcription factors. Out of the 94 genes in the genome with the Zn(II)2Cys6 binuclear cluster domain, the presence of three within this set of 86 genes represents a trend toward significant enrichment (hypergeometric *P*=0.094).

Using an enrichment test and FunCat terms, we also found that these 86 genes are enriched for *catabolism of nitrogenous compounds* (hypergeometric *P*=0.043). However, although the enrichment is statistically significant, only two of the 86 genes have this functional term: the allantoicase encoding gene *alc-1* and malate synthase, homolog of the duplicated yeast genes *MLS1* and *DAL7* (Hartig et al. 1992; Wong and Wolfe 2005). *DAL7* is involved in allantoin degradation and is believed to play a role in preventing the buildup of the glyoxylate endproduct of the purine degradation pathway (Fernandez et al. 1993).

We also examined the entire set of DE genes for others that may have functions related to nitrogen metabolism. We assembled a list of *N. crassa* genes known to be involved in nitrogen metabolism based on manual inspection of primary literature as well as GO and FUNCAT terms. From the combined list of 116 genes, we identified 16 that are significantly differentially expressed between the two populations and, using the Panama strains for polarization, show evidence of having changed expression in the Louisiana population. Of these 16 genes, a total of 13 (including *alc-1* and malate synthase) have been up-regulated, significantly more than expected by chance (hypergeometric *P*=0.022) (Fig. 2). Interestingly, one of these 13 genes is the transcription factor *pco*-1, which has been shown to be required for growth on purines as the sole nitrogen source and binds to the allantoicase promoter (Liu and Marzluf 2004).

Together, the up-regulation in the Louisiana population of this suite of genes related to nitrogen metabolism, the 86 genes with low intrapopulation expression variation and high interpopulation divergence, and the finding that most of the *N. crassa* genes previously identified as candidate targets of positive selection from sequence-based analyses (Ellison et al. 2011) are also differentially expressed, indicate that positive selection has played a small but potentially important role in shaping the regulatory divergence between these two populations.

Discussion

The clusters that we identify here are generated from transcriptional profiles taken from entire colonies of *Neurospora* growing on minimal media. These clusters obviously do not represent the full regulatory complexity of this organism because we are missing expression information for regulons that are only induced under specific conditions, such as those involved in mating. Despite this limitation, we found 142 clusters that are significantly enriched for one or more FunCat terms and that, together, contain a total of 923 genes of unknown function. We provided additional evidence of the utility of these clusters by identifying candidate targets of a homeobox transcription factor for which the target set was previously unknown.

In addition to assessing within population variation, we have also identified genes that are significantly differentially expressed between populations. We find that ~17% of the genes with measurable expression are significantly differentially expressed between these two populations (1539 out of 8823 genes). This percentage is higher than that found in pairwise comparisons of gene expression at the start of metamorphosis between three *Drosophila* species (between ~5% and 11% of genes were found to be differentially expressed between species pairs) (Rifkin et al. 2003) as well as that found for metabolic genes between populations of whitefish (~12%) (Whitehead and Crawford 2006) but falls within the range found between human and chimp (between ~8% and 32%, depending upon the tissue studied) (Khaitovich et al. 2005). Compared to these previous studies, we have expression information for at least four-fold more individuals per population, so the larger number of differentially expressed genes that we identify relative to the previous studies in *Drosophila* and whitefish may simply be the result of having increased statistical power.

We also see a strong asymmetry between populations when we compare the functional terms of the differentially expressed genes with higher expression in one population to those with higher expression in the other (Fig. 1). While this asymmetry suggests that there has been divergence between these two populations in gene expression programs during active growth, these changes may be due to a small number of trans-acting variants that are fixed between the populations. Mapping of eQTL within the Louisiana population is currently underway and may shed light on the genetic architecture underlying the differences we see here.

We have also identified a set of candidate genes whose differences in expression between populations may be due to directional selection. Interestingly, we find a slight enrichment of Zn(II)2Cys6 binuclear cluster transcription factors within this set of genes. This result, albeit not as significant, is similar to that of (Gilad et al. 2006), who found an enrichment of transcription factors within a set of genes identified as having changed expression level along the human lineage due to directional selection. More striking, however, is the set of genes related to nitrogen metabolism whose expression appears to have been specifically up-regulated in the Louisiana population (Fig. 2). Among these genes, allantoicase shows the strongest pattern of differential expression and is also one of the 86 genes whose change in expression between the two populations may be due to directional selection (Fig. 2). This enzyme is involved in the degradation of purines and catalyzes the conversion of allantoate (also known as allantoic acid) to (S)-ureidoglycolate and urea (Reinert and Marzluf 1975). Interestingly, one of the endproducts of this pathway is glyoxylate and the other nitrogen related gene within the set of 86, malate synthase, converts glyoxylate to malate as part of the glyoxylate cycle (Flavell and Woodward 1971). While none of the other enzymes of the purine degradation pathway are differentially expressed between these two populations, the transcription factor *pco*-1 is up-regulated in the Louisana population (Fig. 2). This gene product is required for growth on purines as the sole nitrogen source and binds to the allantoicase promoter (Liu and Marzluf 2004). It may be that the increased expression of allantoicase in the Louisiana population is due to the increased expression of *pco-1*, however, the three other genes whose upstream regions *pco-1* has been shown to bind (xanthine dehydrogenase, uricase, and allantoinase) (Liu and Marzluf 2004) show no difference in expression level between the two populations.

A possible explanation for the up-regulation in Louisiana of the entire suite of genes involved in nitrogen metabolism is that there is a difference in nitrogen repression between these two populations. The strains in this study were grown on media containing ammonium as the nitrogen source and, although *N. crassa* can utilize a variety of compounds as nitrogen sources, ammonium and glutamine are preferred and many of the pathways for the use of alternative sources are repressed in the presence of these molecules (Wiame et al. 1985). However, the expression of the *pco-1* transcription factor has been shown to be independent of nitrogen repression (Liu and Marzluf 2004), suggesting that, at least for allantoicase and *pco-1*, differences in nitrogen repression between the two populations cannot explain the differences in transcription level.

N. crassa is able to grow on media containing allantoate as the sole nitrogen source (Reinert and Marzluf 1975) and it is interesting to note that some nitrogen-fixing plants are known to preferentially transport and store fixed nitrogen as allantoin and allantoate (Schubert 1986). However, little is known about whether *N. crasssa* exhibits any preference for plant growth substrate in nature or whether there are differences in growth substrates between these two populations. Additionally, and perhaps more importantly, it is also currently unknown whether there are differences between these two populations when wild strains are grown on media with allantoate as the sole nitrogen source. These experiments are currently underway.

In this study we examined genome-wide gene expression levels within and between two populations of the model fungus *Neurospora crassa* and harnessed within-population variation in gene expression levels to identify clusters of coexpressed genes. The differences in gene expression between these two populations suggests that, although they are recently diverged and lack any striking differences in morphological phenotype, there may be additional functional differences between them besides the recently discovered difference in growth rate at low temperature (Ellison et al. 2011). The identification of genes whose change in expression may be due to directional selection provides a set of candidate genes that merit further investigation in this regard. The examination of intra-population variation in gene expression has allowed us to predict a set of targets for a previously uncharacterized transcription factor and provides an important resource for imputing functional information for the ca. 50% of genes in the *Neurospora* genome that are of unknown function.

Methods

Sequencing

We used the Illumina RNA-seq dataset from (Ellison et al. 2011) and have added data from an additional 29 Louisiana individuals here, bringing the total to 48 Louisiana individuals and 19 Caribbean individuals. The approaches used for library preparation, sequencing, and identification of single nucleotide polymorphisms for these additional isolates were identical to those in (Ellison et al. 2011). The complete set of strains used is listed in Table S5.

Gene expression quantification

We mapped the Illumina reads to the *Neurospora crassa* OR74A version 10 reference assembly (Galagan et al. 2003) using Tophat (Trapnell et al. 2009). We required reads to map uniquely with a maximum of two mismatches and minimum and maximum intron lengths of 40 and 200 basepairs respectively for both mapping, coverage, and split-segment searches.

We supplied Tophat with the Broad Institute's February 2010 release of the *N. crassa* version 4 annotation and also allowed the program to search for novel splice junctions, however, for purposes of expression quantification, we collapsed all putative isoforms into a single expression measure per gene. We used a custom Perl script and the *N. crassa* annotation to calculate the number of raw reads overlapping each gene model and used the third quartile method from (Bullard et al. 2010) to normalize read counts between lanes. Genes were excluded where more than half of the individuals analyzed had an expression level of zero. For the clustering analysis, read counts were also normalized by gene length.

Functional enrichment tests

The Munich Information Center for Protein Sequences (MIPS) has developed a database of functional categories (FunCats) (Ruepp et al. 2004) that have been associated with *N. crassa* gene IDs. Similar to GO terms (Ashburner et al. 2000), these categories consist of five hierarchical levels of increasing specificity, each containing increasingly smaller subsets of genes from the levels above them. We used a custom Perl script and the R function *phyper* to perform enrichment tests using these categories. We used the Benjamini-Hochberg (Benjamini and Hochberg 1995) correction for multiple hypothesis testing within each level and a corrected P-value cutoff of 0.1.

Differential expression between populations

We calculated normalized expression measurements for the 19 Caribbean individuals reported in (Ellison et al. 2011) as described above and identified genes showing differential expression between the two populations using the wilcoxon test and Benjamini-Hochberg correction (Benjamini and Hochberg 1995) for multiple hypothesis testing. We also used three individuals from Panama that were identified as a separate population in the analysis using Bayesian clustering of allele frequencies conducted in (Ellison et al. 2011). The Panama strains were used to polarize the change in expression level for genes that were differentially expressed between the Louisiana and the Caribbean populations. The change in expression was inferred to have occurred in the population whose median expression level was furthest from that of the Panama strains.

The ratio of expression variation within populations to divergence between populations was calculated, in a manner analogous to F_{ST} (Hudson et al. 1992), by dividing the average of the absolute value of the $\log_{10}$ fold changes between all pairwise comparisons within populations to the average of all pairwise comparisons between populations and subtracting this ratio from one.

Expression clusters

We used the Statistics::RankCorrelation Perl module to calculate Spearman's rank correlation coefficients between all pairwise comparisons of the 8361 *N. crassa* genes using expression levels from the 48 Louisiana individuals. We took the absolute value of each correlation coefficient and created clusters of genes whose expression is correlated between individuals using the agglomerative hierarchical clustering method described in (Horan et al. 2008). This approach uses the hclust package in R (Team 2009) to perform complete linkage clustering followed by a custom hierarchical threshold clustering method developed by (Horan et al. 2008) to create discrete clusters from the hclust output. We employed this method using correlation coefficient thresholds ranging from 0.2 to 0.6. For each threshold, we assessed the false discovery rate by analyzing ten datasets where the expression levels for each gene were randomly permuted between individuals.

Transcription factor deletion strains

The four transcription factor deletion strains were obtained from the *N. crassa* deletion collection (Colot et al. 2006). Culture and RNA extraction techniques were identical to those described in (Ellison et al. 2011). Illumina RNA-seq reads were mapped as described above, genes with raw read counts less than five were excluded, and the R package DEseq (Anders and Huber 2010) was used to identify differentially expressed genes between each deletion strain and the wild-type reference strain OR74A. For DEseq analyses we normalized read counts between lanes using the *estimateSizeFactors* function that is part of the DEseq package. In order to assess differential expression without replicates, we made the assumption that the majority of genes are not truly differentially expressed between the wild-type and deletion strains and therefore estimated the expression variance across the pooled samples as described in (Anders and Huber 2010).

For each transcription factor deletion strain, we took the intersection of the list of differentially expressed genes from DEseq and the list of the genes that clustered with the transcription factor in the expression cluster analysis. We then determined if the overlap was larger than expected by chance using Fisher's exact test.

Figures and Tables

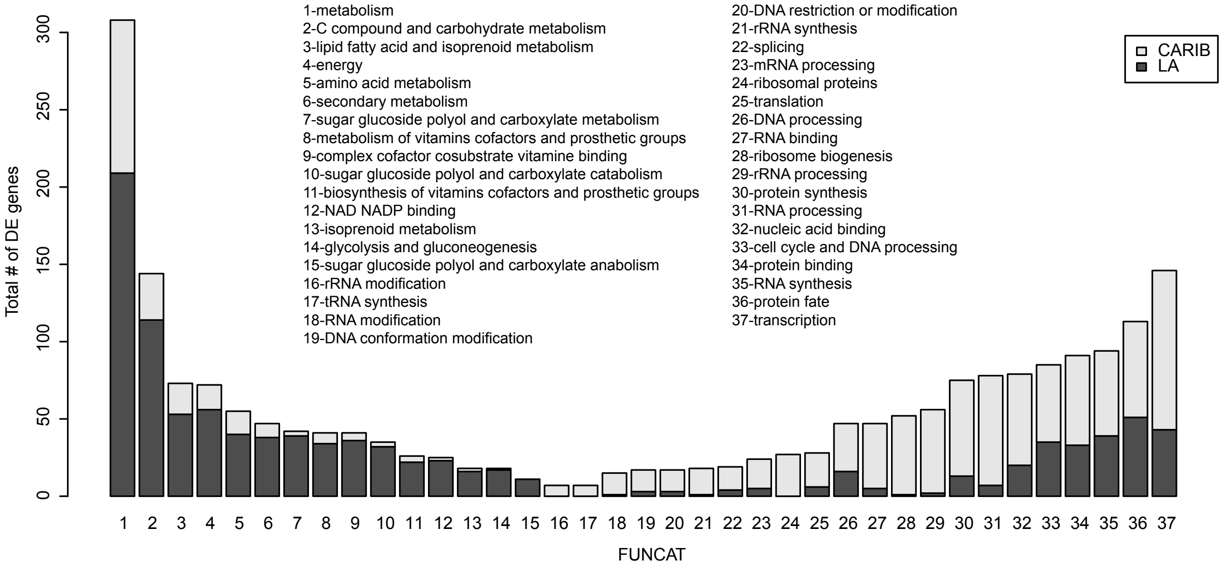

Figure 1. Functional asymmetry in gene expression between the Caribbean and Louisiana *N. crassa* populations.

FunCat enrichment tests were performed for genes with significantly higher expression in the Louisiana population compared to the Caribbean (and vice versa) out of the total pool of differentially expressed genes. Each term shown here was significantly enriched in one of these two sets: numbers 1-15 are the enriched terms for the set of more highly expressed Louisiana genes and numbers 16-37 are the enriched terms for the set of more highly expressed Caribbean genes. The height of each bar is equal to the total number of differentially expressed genes associated with that FunCat term while the shaded segments show the number of genes that are more highly expressed in one population versus the other.

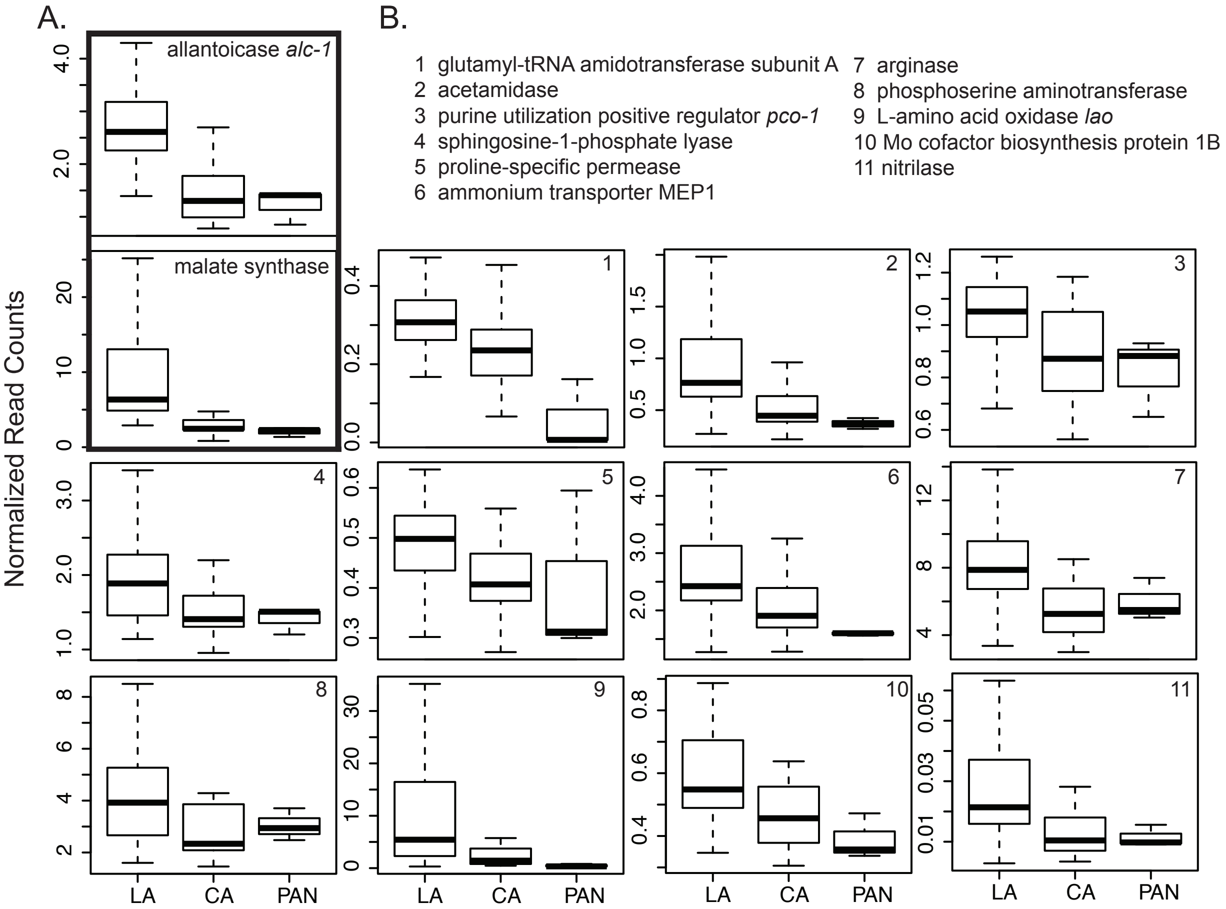

Figure 2. Genes involved in nitrogen metabolism have been specifically up-regulated in the *N. crassa* Louisiana population.

The *N. crassa* genes coding for allantoicase and malate synthase are among the set of 86 genes whose change in expression level between populations may be due to directional selection. Using the Panama (PAN) individuals to polarize the change in expression, these genes also show strong up-regulation in Louisiana (A). The expression levels of an additional 11 genes involved in nitrogen metabolism also appear to have been specifically up-regulated in the Louisiana population (B). Out of a total of 16 differentially expressed, nitrogen related genes whose expression has changed in the Louisiana population, 13 of these are up-regulated, significantly more than expected by chance (hypergeometric *P*=0.022).

Correlation cutoff	Number of clusters	Minimum size of clusters	Genes in clusters
0.2	49	45	7038
0.3	66	25	5512
0.4	188	9	5526
0.5	275	6	4364
0.6	356	4	3199

Table 1. Summary of regulon clusters
Shown here are summaries of the clusters of genes whose expression levels were correlated between Louisiana individuals. The cluster sets were generated using correlation coefficient cutoffs ranging from 0.2 to 0.6. The cutoff of 0.4 results in the second-largest number of genes assigned to clusters, behind the 0.2 cutoff. However, a cutoff of 0.2 results in a much smaller total number of clusters. For this reason, we chose the set generated using the 0.4 cutoff for further analysis.

TF ID	Annotation	PFAM	# of DE genes	Cluster size	Overlap	FET P-value
NCU05257	homeobox and C2H2 transcription factor	Homeobox/ zf-C2H2 Zinc finger	43	59	8	1.84e-09
NCU05767	paralog of *Sordaria macrospora pro-1*	Fungal Zn(2)-Cys(6) binuclear cluster domain	53	27	1	0.17
NCU07039	*asd-4*	GATA zinc finger	55	19	0	--
NCU07705	C6 finger domain-containing protein	Fungal Zn(2)-Cys(6) binuclear cluster domain	86	35	0	--

Table 2. Comparison of the genes from transcription factor expression clusters to those that were differentially expressed in the transcription factor deletion strain.
Gene expression levels were compared between four transcription factor deletion strains and the isogenic wild-type strain to identify genes that were differentially expressed (DE) in the mutant. The significance of the overlap between the set of DE genes and the set of genes whose expression levels across the 51 wild Louisiana strains was correlated with that of the transcription factor was assessed using Fisher's exact test (FET).

ID	Annotation	DE P-value	Coexpression correlation
NCU00522	cystathionine beta-lyase	2.3104e-05	0.78
NCU00535	alanyl-tRNA synthetase	0.0054	0.84
NCU02019	FAD dependent oxidoreductase	1.5469e-08	0.82
NCU02543	aspartate aminotransferase	0.0027	0.92
NCU05045	MFS monocarboxylate transporter	1.8375e-09	0.90
NCU05256	hypothetical	0.0013	0.81
NCU07126	Acetyltransferase (GNAT) family	0.0503	0.83
NCU11365	aminotransferase	0.0533	0.82

Table 3. Overlap between expression cluster and differentially expressed genes for the homeobox transcription factor NCU05257.

The genes shown here were differentially expressed compared to wild-type in the NCU05257 deletion strain and their expression is also highly correlated with that of NCU05257 across the wild Louisiana strains. DE P-value is the Benjamini-Hochberg corrected P-value resulting from the test for differential expression between the wild-type and deletion mutant. The coexpression correlation column shows the Spearman's rank correlation coefficient between the expression levels of each gene and NCU05257 across the wild Louisiana strains.

Cluster Size	Real 0.2	Perm 0.2	Real 0.3	Perm 0.3	Real 0.4	Perm 0.4	Real 0.5	Perm 0.5	Real 0.6	Perm 0.6
1	0	0	0	0	22	4.3	672	1884.9	3139	8148.9
2	0	0	1	0.4	253	885.4	927	3085.7	0891	361
3	0	0	5	4.8	119	369.6	289	239.2	252	1.7
4	0	0	73	357.4	221	1237.6	171	24.9	113	0
5	0	0	38	119.4	110	139.9	87	0.5	60	0
6	0	0.1	42	146.6	75	44.3	61	0	55	0
7	0	2	34	176.1	51	9.2	36	0	22	0
8	4	28.4	40	404.4	28	1.5	25	0	18	0
9	4	17.7	30	83.3	17	0	17	0	17	0
10	3	18.8	28	39.4	19	0	28	0	10	0
11	0	21.9	19	16.2	12	0	14	0	7	0
12	2	32.8	18	9.7	15	0	13	0	4	0
13	7	37.7	5	2.4	6	0	9	0	8	0
14	5	52.2	6	0.8	13	0	3	0	4	0
BINS OF 5										
20	23	313.6	31	0.4	36	0	21	0	7	0
25	10	44.3	19	0	19	0	13	0	5	0
30	14	10.9	9	0	10	0	9	0	1	0
35	6	1.4	11	0	8	0	5	0	3	0
40	1	0.1	8	0	7	0	3	0	0	0
45	8	0	4	0	4	0	3	0	1	0
50	3	0	3	0	4	0	2	0	1	0
55	4	0	1	0	2	0	3	0	0	0
60	3	0	5	0	2	0	2	0	1	0
65	5	0	3	0	0	0	1	0	2	0
70	0	0	0	0	2	0	1	0	0	0
75	4	0	3	0	0	0	0	0	0	0
80	3	0	2	0	0	0	0	0	0	0
85	0	0	2	0	1	0	0	0	0	0
90	2	0	1	0	1	0	2	0	0	0
95	0	0	0	0	0	0	1	0	0	0
100	3	0	0	0	1	0	0	0	0	0
105-435	22	0	14	0	9	0	3	0	2	0

Table S1. Comparison of cluster sizes between regulon clusters and randomly permutated data.

Clusters of genes showing correlated expression between Louisiana individuals were created using correlation coefficient cutoffs ranging from 0.2 to 0.6. For each cutoff, 10 additional datasets were clustered where expression levels were randomly permuted between individuals for each gene. This table shows the size of the clusters for the actual dataset compared to the average size across the ten permutations, for each correlation coefficient cutoff.

ID	Name	Function	Corrected P-value
NCU06245	*plc-1*	Phospholipase	0.036
NCU06246	*MRH4*-like	mitochondrial DEAD box RNA helicase	0.030
NCU06247	--	outer mitochondrial membrane protein	0.009
NCU02263	*SEC14*-like	phospholipid transport	0.029
NCU02265	*frq*	major circadian oscillator	0.000005

Table S2. Differential expression of genes from the "divergence islands" of Ellison et al. 2011.

Five of the seven genes identified in (Ellison et al. 2011) as being extremely divergent between the *N. crassa* Louisiana and Caribbean populations are also differentially expressed.

FunCat	P-value
Transcription	0.033
Pentose phosphate pathway	0.095
RNA processing	0.00008
RNA modification	0.046
Ribosome biogenesis	0.095
Nucleic acid binding	0.057
Sugar glucoside polyol and carboxylate metabolism	0.036
rRNA synthesis	0.043
rRNA processing	0.000000004
rRNA modification	0.012
RNA binding	0.012
nucleolus	0.036

Table S3. Functional enrichment of genes found to be differentially expressed between populations
The set of differentially expressed genes that were associated with one or more FunCat terms was compared to the total set of expressed genes. The terms shown here were found to be significantly enriched (corrected P<= 0.1) in the set of differentially expressed genes.

ID	FUNCTION	PFAM DOMAINS	DIR	1-(V/D)
NCU00115	rRNA-processing_protein_FCF2	Fcf2_pre-rRNA_processing	LA_DOWN	0.40
NCU00150	phosphoribosylformimino-5-aminoimidazole_carboxamide_ribotide_isomerase	Histidine_biosynthesis_protein/ phosphoribosylformimino-5-aminoimidazole_carboxamide ribotide_isomerase	CA_DOWN	0.42
NCU00290	ABC_transporter	ABC-2_type_transporter	CA_UP	0.27
NCU00344	hypothetical_protein	Fungal_Zn(2)-Cys(6) binuclear_cluster_domain	LA_UP	0.36
NCU00351	cript_family_protein	Microtubule-associated_protein_CRIPT	LA_UP	0.35
NCU00776	tyrosinase	central_domain_of_tyrosinase	LA_DOWN	0.28
NCU00801	MFS_lactose_permease	MFS_transporter_sugar_porter_family	CA_DOWN	0.26
NCU00864	TIM-barrel_enzyme_family	2-nitropropane_dioxygenase/TIM-barrel_signal_transduction_protein	LA_UP	0.29
NCU00918	hypothetical_protein	NA	LA_UP	0.59
NCU00980	hypothetical_protein	SPT2_chromatin_protein	LA_DOWN	0.25
NCU01097	hypothetical_protein	Fungal_Zn(2)-Cys(6) binuclear_cluster_domain	CA_UP	0.28
NCU01374	CutC_family_protein	CutC_family	LA_UP	0.33
NCU01659	dDENN_domain-containing_protein	dDENN_domain/DENN_domain/ uDENN_domain	CA_DOWN	0.27
NCU01771	DNA_repair_protein_rhp57	KaiC/Rad51	LA_DOWN	0.28
NCU01816	allantoicase-I	allantoicase/Allantoicase_repeat	LA_UP	0.52
NCU01857	DEAD/DEAH_box RNA_helicase	DEAD/DEAH_box_helicase/ DSHCT_domain/ Helicase _Cterminal_domain	LA_UP	0.30
NCU01998	septin	Septin	LA_UP	0.26
NCU02060	zinc_metallopeptidase	WLM_domain/ Zn-finger_in_Ran_binding_protein	CA_UP	0.30
NCU02223	hypothetical_protein	Transcription_initiation_factor_TFIID_subunit_A	CA_DOWN	0.42
NCU02265	frequency	Frequency_clock_protein	CA_DOWN	0.41
NCU02529	hypothetical_protein	NA	LA_DOWN	0.37
NCU02597	hypothetical_protein	Amidohydrolase_family	CA_DOWN	0.36
NCU02598	hypothetical_protein	NA	LA_DOWN	0.26
NCU02713	conidial_separation-1	Zinc_finger_C2H2_type	CA_DOWN	0.27
NCU02732	hypothetical_protein	NA	LA_UP	0.44
NCU03056	hypothetical_protein	NA	LA_UP	0.71
NCU03057	hypothetical_protein	DUF3500	LA_UP	0.50
NCU03067	hypothetical_protein	NA	LA_UP	0.39
NCU03141	lysophospholipase	Lysophospholipase_catalytic_domain	LA_DOWN	0.31
NCU03242	serine/threonine-protein_kinase_cbk1	Protein_kinase_domain/ Protein_tyrosine_kinase	LA_DOWN	0.34
NCU03617	haloacid_dehalogenase	HAD-superfamily_hydrolase subfamily_IA_variant_2/ HAD-like_hydrolase/HAD_type_II	CA_DOWN	0.36
NCU03641	beta-glucosidase_2	Glycosyl_hydrolase_family_3_C_and_N_terminal_domains	LA_UP	0.28
NCU03859	MYG1_protein	UPF0160	LA_DOWN	0.31
NCU03883	vacuolar_transporter chaperone_1	Domain_of_unknown_function_DUF	LA_DOWN	0.32
NCU03948	amidohydrolase_3	Amidohydrolase_family	LA_UP	0.31
NCU03960	hypothetical_protein	NA	LA_DOWN	0.42
NCU03966	Mitochondrial Rho_GTPase_1	EF_hand_associated/ Miro-like_protein/Ras_family/ small_GTP-binding_protein_domain	CA_DOWN	0.40
NCU03979	biotin_synthase	Biotin_and_Thiamin_Synthesis associated_domain/biotin_synthase/ Radical_SAM_superfamily	LA_UP	0.27
NCU04070	hypothetical_protein	NA	CA_DOWN	0.39
NCU04071	carboxymuconate_cyclase	3-carboxy-cis,cis-muconate_lactonizing_enzyme	CA_DOWN	0.29
NCU04096	serine/threonine-protein_kinase_3	Protein_kinase_domain/ Protein_tyrosine_kinase	CA_UP	0.33
NCU04138	hypothetical_protein	NA	LA_DOWN	0.30
NCU04167	hypothetical_protein	NA	LA_UP	0.34
NCU04354	DEAD_box_family_helicase	DEAD/DEAH_box_helicase/ Helicase_Cterminal_domain/ Type_III_restriction_enzyme_res_subunit	CA_DOWN	0.27

NCU04518	hypothetical_protein	NA	CA_UP	0.46
NCU04641	FAD_dependent_oxidoreductase	FAD_dependent_oxidoreductase	LA_UP	0.36
NCU04776	hypothetical_protein	NA	CA_UP	0.27
NCU05164	short_chain_dehydrogenase/reductase_SDR	KR_domain/NAD_dependent_epimerase/dehydratase_family/ short_chain_dehydrogenase	LA_UP	0.31
NCU05209	hypothetical_protein	RTA1_like_protein	CA_DOWN	0.39
NCU05900	hypothetical_protein	NA	CA_UP	0.27
NCU06088	hypothetical_protein	DUF2962	LA_DOWN	0.29
NCU06212	trafficking_protein_particle complex_subunit_6B	Transport_protein_particle_(TRAPP)_component	CA_UP	0.26
NCU06223	hypothetical_protein	NA	CA_DOWN	0.41
NCU06284	VPS-associated_protein_35	VPS-associated_protein_35	CA_DOWN	0.30
NCU06387	hypothetical_protein	NA	LA_UP	0.41
NCU06390	hypothetical_protein	PAS_domain_S-box/PAS_fold	LA_DOWN	0.33
NCU06731	hypothetical_protein	NA	LA_UP	0.38
NCU07693	stu-1	CLASP_N_terminal	LA_DOWN	0.48
NCU07705	C6_finger_domain-containing_protein	Fungal_Zn(2)Cys(6) binuclear_cluster_domain	LA_UP	0.37
NCU07824	MDM10	DUF3722)	CA_UP	0.53
NCU07849	thiamine-4	pfkB_family_carbohydrate_kinase/ phosphomethylpyrimidine_kinase/ TENA/THI-4/PQQC_family	LA_UP	0.26
NCU07966	calcium-transporting_ATPase_3	ATPase_P-type HAD_superfamily_subfamily_IC/ Cation_transporter/ATPase_N-terminus/Cation_transporting_ATPase_C-terminus/E1-E2_ATPase/ HAD-like_hydrolase/ potassium/sodium_efflux_P-type_ATPase_fungal-type	LA_DOWN	0.42
NCU08032	tRNAsplicing_endonucleasesubunitsen34	tRNA_intron_endonuclease	LA_DOWN	0.33
NCU08033	hypothetical_protein	Tat_pathway_signal_sequence	LA_UP	0.31
NCU08045	hypothetical_protein	NA	LA_UP	0.33
NCU08384	xylose_reductase	Aldo/keto_reductase_family	LA_UP	0.37
NCU08436	alpha/beta_hydrolase	alpha/beta_hydrolase_fold/Ndr_family	LA_UP	0.29
NCU08986	hypothetical_protein	NA	LA_DOWN	0.32
NCU09336	hypothetical_protein	NA	LA_UP	0.28
NCU09427	G-protein_coupled_receptor	7_transmembrane_receptor/ G_protein_coupled_glucose_receptorregulating_Gpa2/ Slime_mold_cyclic_AMP_receptor	LA_UP	0.45
NCU09503	hypothetical_protein	BRCA1_C_Terminus_domain	LA_UP	0.26
NCU09806	hypothetical_protein	NA	LA_DOWN	0.35
NCU09995	hypothetical_protein	HMG_box	CA_UP	0.32
NCU10007	malate_synthase	Malate_synthase/malate_synthase_A	LA_UP	0.37
NCU10040	hypothetical_protein	NA	LA_UP	0.45
NCU10069	hypothetical_protein	NA	LA_UP	0.71
NCU11043	hypothetical_protein	NA	CA_UP	0.26
NCU11423	hypothetical_protein	NA	CA_UP	0.29
NCU11442	hypothetical_protein	NA	LA_DOWN	0.44
NCU11468	hypothetical_protein	NA	LA_DOWN	0.25
NCU11629	hypothetical_protein	NA	CA_UP	0.40
NCU11767	hypothetical_protein	Acetyltransferase_(GNAT)_family	LA_DOWN	0.29
NCU11788	hypothetical_protein	DUF2781	CA_DOWN	0.48
NCU11856	hypothetical_protein	NA	CA_DOWN	0.45
NCU11919	hypothetical_protein	NA	LA_UP	0.46
NCU12150	ywbE	DUF2196	LA_UP	0.57

Table S4. Candidate genes whose change in expression level between populations may be due to directional selection.

The gene id number, functional annotation, and PFAM domains are shown for each gene. The two columns on the far right show the polarized change in expression inferred using the Panama population as an outgroup and the relative divergence of expression levels between populations (calculated as the ratio of expression differences within populations [V] to those between populations [D] and subtracted from one).

Strain Number	FGSC	Perkins	Mat	Strain provenance	Collection site	Substrate
D110	8870	4448	A	Dettman, J.	Franklin, LA	sugarcane
D111	8871	4449	a	Dettman, J.	Franklin, LA	sugarcane
D112	8872	4453	A	Dettman, J.	Franklin, LA	sugarcane
D114	8874	4464	A	Dettman, J.	Franklin, LA	sugarcane
D116	8876	4481	a	Dettman, J.	Franklin, LA	sugarcane
D118	8878	4491	a	Dettman, J.	Franklin, LA	sugarcane
D23	8783	1409	A	Dettman, J.	Homestead, FL	grass
D24	8784	1410	A	Dettman, J.	Homestead, FL	grass
D27	8787	1417	A	Dettman, J.	Homestead, FL	grass
D29	8789	1465	A	Dettman, J.	Homestead, FL	grass
D30	8790	1470	a	Dettman, J.	Homestead, FL	grass
D56	8816	3424	A	Dettman, J.	Carrefour Dufort, Haiti	grass
D59	8819	3427	a	Dettman, J.	Carrefour Dufort, Haiti	Sugarcane
D69	8829	3684	a	Dettman, J.	Tiassale, Ivory Coast	grass
D85	8845	4130	a	Dettman, J.	Kabah, Yucatan, MX	soil, unburnt
D88	8848	4150	a	Dettman, J.	Sayil, Yucatan, MX	soil, unburnt
D90	8850	4154	A	Dettman, J.	Uxmal, Yucatan, MX	soil, unburnt
D91	8851	4155	A	Dettman, J.	Uman, Yucatan, MX	soil, unburnt
JW05	1133		a	Welch, J.	Panama	unknown
JW07	1165		a	Welch, J.	Panama	unknown
JW09	2229		A	Welch, J.	Welsh, LA	burned grass
JW10	2229		A	Welch, J.	Welsh, LA	burned grass
JW15	1132		a	Welch,J	Panama	unknown
JW22	3223		A	Welch, J.	Elizabeth, LA	pine burn
JW28	3968		A	Welch,J	Okeechobee, FL	unknown
JW35	3975		a	Welch, J.	Florida	reeds, burnt
JW39	4708		A	Welch, J.	Haiti	wood, burnt
JW43	4712		a	Welch, J.	Haiti	sugarcane burn
JW45	4713		A	Welch, J.	Haiti	sugarcane burn
JW47	4715		a	Welch, J.	Haiti	sugarcane burn
JW49	4716		A	Welch, J.	Haiti	grass burn
JW54	4824		A	Welch, J.	Haiti	bushes burn
JW59	3200		a	Welch, J.	Coon, LA	burned stumps
JW66	3211		a	Welch, J.	Sugartown, LA	pine burn
JW70	3199		A	Welch, J.	Coon, LA	burned stumps
JW75	3943		a	Welch, J.	Houma, LA	sugarcane burn
		4450	a	Perkins, D.	Franklin, LA	sugarcane burn
		4452	a	Perkins, D.	Franklin, LA	sugarcane burn
		4455	a	Perkins, D.	Franklin, LA	sugarcane burn
		4465	a	Perkins, D.	Franklin, LA	sugarcane burn
		4476	a	Perkins, D.	Franklin, LA	sugarcane burn
		4496	a	Perkins, D.	Franklin, LA	sugarcane burn
		4463	A	Perkins, D.	Franklin, LA	sugarcane burn
		4471	a	Perkins, D.	Franklin, LA	sugarcane burn
		4486	A	Perkins, D.	Franklin, LA	sugarcane burn
		4487	A	Perkins, D.	Franklin, LA	sugarcane burn
		4489	a	Perkins, D.	Franklin, LA	sugarcane burn
		4497	a	Perkins, D.	Franklin, LA	sugarcane burn
		4498	a	Perkins, D.	Franklin, LA	sugarcane burn
	847		A	Lein	Louisiana	sugarcane burn

D113	8873	4454	a	Dettman, J.	Franklin, LA	sugarcane
D119	8879	4500	a	Dettman, J.	Franklin, LA	sugarcane
JW20	3212		A	Welch,J	Ravenswood, LA	bonfire
JW76	3943		a	Welch,J	Houma, LA	sugarcane burn
JW159	2221		a	Welch,J	Houma, LA	sugarcane burn
JW160	2222		A	Welch,J	Iowa, LA	grass burn
JW162	2223		a	Welch,J	Iowa, LA	grass burn
JW164	2224		a	Welch,J	Marrero, LA	wood burn
JW167	2228		a	Welch,J	Roanoke, LA	grass burn
JW176		513	a	Welch,J	Welsh, LA	unknown
JW179		517	a	Welch,J	Roanoke. LA	unknown
JW184		531	a	Welch,J	Iowa, LA	grass burn
JW210		4493	a	Welch,J	Franklin, LA	sugarcane burn
JW216		511	a	Welch,J	Welsh, LA	unknown
JW222		880	a	Welch,J	Coon, LA	burned stumps
JW230		4509	a	Welch,J	Georgia Plantation,LA	unknown
JW233		492	a	Welch,J	Houma, LA	sugarcane burn
JW234		502	A	Welch,J	Houma, LA	sugarcane burn
JW238		515	A	Welch,J	Welsh, LA	unknown
JW242		521	A	Welch,J	Welsh, LA	burned grass
JW250		853	A	Welch,J	Sugartown.LA	unknown
OR74A	2489		A	FGSC	Marrero, LA	unknown

Table S5. Strains used in this study
All strains were provided by the Fungal Genetic Stock Center (FGSC).

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biology 11(10):R106.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25(1):25-29.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological 57(1):289-300.

Berbee ML, Taylor JW. 2010. Dating the molecular clock in fungi - how close are we? Fungal Biology Reviews 24(1-2):1-16.

Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genetics 4(11):e1000271.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296(5568):752-5.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11:94.

Colot HV, Park G, Turner GE, Ringelberg C, Crew CM, Litvinkova L, Weiss RL, Borkovich KA, Dunlap JC. 2006. A high-throughput gene knockout procedure for *Neurospora* reveals functions for multiple transcription factors. Proceedings of the National Academy of Sciences of the United States of America 103(27):10352-10357.

Dettman JR, Jacobson DJ, Taylor JW. 2003a. A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote *Neurospora*. Evolution 57(12):2703-20.

Dettman JR, Jacobson DJ, Taylor JW. 2006. Multilocus sequence data reveal extensive phylogenetic species diversity within the *Neurospora discreta* complex. Mycologia 98(3):436-46.

Dettman JR, Jacobson DJ, Turner E, Pringle A, Taylor JW. 2003b. Reproductive isolation and phylogenetic divergence in *Neurospora*: comparing methods of species recognition in a model eukaryote. Evolution 57(12):2721-41.

Dunlap JC, Borkovich KA, Henn MR, Turner GE, Sachs MS, Glass NL, McCluskey K, Plamann M, Galagan JE, Birren BW, et al. 2007. Enabling a community to dissect an organism: overview of the *Neurospora* functional genomics project. Advances in Genetics 57:49-96.

Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95(25):14863-8.

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proceedings of the National Academy of Sciences of the United States of America 108(7):2831-6.

Fay JC, Wittkopp PJ. 2008. Evaluating the role of natural selection in the evolution of gene regulation. Heredity 100(2):191-9.

Feng B, Haas H, Marzluf GA. 2000. ASD4, a new GATA factor of *Neurospora crassa*, displays sequence-specific DNA binding and functions in ascus and ascospore development. Biochemistry 39(36):11065-11073.

Fernandez E, Fernandez M, Rodicio R. 1993. Two structural genes are encoding malate synthase isoenzymes in *Saccharomyces cerevisiae*. FEBS Letters 320(3):271-5.

Flavell RB, Woodward DO. 1971. Metabolic role, regulation of synthesis, cellular localization, and genetic control of the glyoxylate cycle enzymes in *Neurospora crassa*. Journal of Bacteriology 105(1):200-10.

Fulci V, Macino G. 2007. Quelling: post-transcriptional gene silencing guided by small RNAs in *Neurospora crassa*. Current Opinion in Microbiology 10(2):199-203.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422(6934):859-868.

Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. 2004. Conservation and evolution of cis-regulatory systems in ascomycete fungi. PLoS Biology 2(12):e398.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. 2000. Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell 11(12):4241-57.

Gerke J, Lorenz K, Cohen B. 2009. Genetic interactions between transcription factors cause natural variation in yeast. Science 323(5913):498-501.

Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 440(7081):242-5.

Glass NL, Dementhon K. 2006. Non-self recognition and programmed cell death in filamentous fungi. Current Opinion in Microbiology 9(6):553-8.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431(7004):99-104.

Hartig A, Simon MM, Schuster T, Daugherty JR, Yoo HS, Cooper TG. 1992. Differentially regulated malate synthase genes participate in carbon and nitrogen metabolism of *S. cerevisiae*. Nucleic Acids Research 20(21):5677-86.

Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, Zhu JK, Cushman JC, Gollery M, Girke T. 2008. Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiology 147(1):41-57.

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132(2):583-9.

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al. 2000. Functional discovery via a compendium of expression profiles. Cell 102(1):109-26.

Jacobson DJ, Powell AJ, Dettman JR, Saenz GS, Barton MM, Hiltz MD, Dvorachek WH, Jr., Glass NL, Taylor JW, Natvig DO. 2004. *Neurospora* in temperate forests of western North America. Mycologia 96(1):66-74.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309(5742):1850-1854.

Kyrpides NC. 1999. Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. Bioinformatics 15(9):773-4.

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. 1997. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proceedings of the National Academy of Sciences of the United States of America 94(24):13057-62.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298(5594):799-804.

Liu TD, Marzluf GA. 2004. Characterization of *pco-1*, a newly identified gene which regulates purine catabolism in *Neurospora*. Current Genetics 46(4):213-227.

Loros JJ, Dunlap JC. 2001. Genetic and molecular analysis of circadian rhythms in *Neurospora*. Annual Review of Physiology 63:757-94.

Masloff S, Jacobsen S, Poggeler S, Kuck U. 2002. Functional analysis of the C6 zinc finger gene pro1 involved in fungal sexual development. Fungal Genetics and Biology 36(2):107-16.

Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America 100(17):9894-9.

Menkis A, Bastiaans E, Jacobson DJ, Johannesson H. 2009. Phylogenetic and biological species diversity within the *Neurospora tetrasperma* complex. Journal of Evolutionary Biology 22(9):1923-1936.

Metzenberg RL. 2004. Bird Medium: an alternative to Vogel Medium. Fungal Genetics Newsletter 51:19-20.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320(5881):1344-9.

Nayak RR, Kearns M, Spielman RS, Cheung VG. 2009. Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. Genome Research 19(11):1953-62.

Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM. 2004. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. Molecular Biology and Evolution 21(7):1308-17.

Ranz JM, Machado CA. 2006. Uncovering evolutionary patterns of gene expression using microarrays. Trends in Ecology & Evolution 21(1):29-37.

Reinert WR, Marzluf GA. 1975. Regulation of the purine catabolic enzymes in *Neurospora crassa*. Archives of Biochemistry and Biophysics 166(2):565-74.

Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nature Genetics 33(2):138-44.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. Nature Reviews Genetics 7(11):862-72.

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, et al. 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research 32(18):5539-45.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467-70.

Schubert KR. 1986. Products of biological nitrogen fixation in higher plants: Synthesis, Transport, and Metabolism. Annual review of plant physiology 37:539-574.

Shiu PK, Metzenberg RL. 2002. Meiotic silencing by unpaired DNA: properties, regulation and suppression. Genetics 161(4):1483-95.

Smith KM, Sancar G, Dekhang R, Sullivan CM, Li S, Tag AG, Sancar C, Bredeweg EL, Priest HD, McCormick RF, et al. 2010. Transcription factors in light and circadian clock signaling networks revealed by genomewide mapping of direct targets for neurospora white collar complex. Eukaryotic Cell 9(10):1549-56.

Team RDC. 2009. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Thompson DA, Regev A. 2009. Fungal regulatory evolution: cis and trans in the balance. FEBS Letters 583(24):3959-65.

Tian C, Beeson WT, Iavarone AT, Sun J, Marletta MA, Cate JH, Glass NL. 2009. Systems analysis of plant cell wall degradation by the model filamentous fungus *Neurospora crassa*. Proceedings of the National Academy of Sciences of the United States of America 106(52):22157-62.

Tian C, Kasuga T, Sachs MS, Glass NL. 2007. Transcriptional profiling of cross pathway control in *Neurospora crassa* and comparative analysis of the Gcn4 and CPC1 regulons. Eukaryotic Cell 6(6):1018-29.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105-1111.

Turner E, Jacobson DJ, Taylor JW. 2010. Reinforced postmating reproductive isolation barriers in Neurospora, an Ascomycete microfungus. Journal of Evolutionary Biology 23(8):1642-56.

Villalta CF, Jacobson DJ, Taylor JW. 2009. Three new phylogenetic and biological *Neurospora* species: *N. hispaniola*, *N. metzenbergii* and *N. perkinsii*. Mycologia 101(6):777-89.

Whitehead A, Crawford DL. 2006. Neutral and adaptive variation in gene expression. Proceedings of the National Academy of Sciences of the United States of America 103(14):5425-5430.

Whittle CA, Nygren K, Johannesson H. 2011. Consequences of reproductive mode on genome evolution in fungi. Fungal Genetics and Biology advance online publication, 27 February 2011(DOI 10.1016/j.fgb.2011.02.005).

Wiame JM, Grenson M, Arst HN, Jr. 1985. Nitrogen catabolite repression in yeasts and filamentous fungi. Advances in Microbial Physiology 26:1-88.

Wohlbach DJ, Thompson DA, Gasch AP, Regev A. 2009. From elements to modules: regulatory evolution in Ascomycota fungi. Current Opinion in Genetics & Development 19(6):571-8.

Wong S, Wolfe KH. 2005. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nature Genetics 37(7):777-82.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L. 2003. Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genetics 35(1):57-64.