The Role of Protein-coding and Regulatory Evolution in Speciation of Wild Yeast

Danting Jiang

Follow this and additional works at: https://scholarworks.wm.edu/honorstheses

Part of the Bioinformatics Commons, Developmental Biology Commons, Evolution Commons, Genomics Commons, Molecular Genetics Commons, and the Population Biology Commons

Recommended Citation


https://scholarworks.wm.edu/honorstheses/1205

This Honors Thesis is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Undergraduate Honors Theses by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
The Role of Protein-coding and Regulatory Evolution in Speciation of Wild Yeast

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Biology from The College of William and Mary

by

Danting Jiang

Accepted for Honors

Dr. Helen Murphy, Director

Dr. Mark Forsyth

Dr. Ross Iaci

Dr. Joshua Puzey

Williamsburg, VA
May 4, 2018
# Table of Contents

Acknowledgements ........................................................................................................ 3
Abstract ............................................................................................................................ 4
Introduction ...................................................................................................................... 5
  The Speciation of Microbial Eukaryotes ..................................................................... 5
  *Saccharomyces* Yeasts as Models for Evolutionary Genetic Studies ....................... 6
  Sympatric Speciation in *S. paradoxus* ..................................................................... 7
  The Migration of *S. paradoxus* .................................................................................. 8
  The “Evo-Devo” Debate ............................................................................................. 10
Research Aims ................................................................................................................ 12
Materials and Methods .................................................................................................. 13
  Strains and Sequencing of Genomes ........................................................................ 13
  Sequence Alignment .................................................................................................... 13
  Data Curation: Extracting Sequences ........................................................................ 14
  Data Analysis: Investigation into the Protein-coding Regions .................................. 15
  Investigation into the *cis*-regulatory Regions ......................................................... 16
  Investigation into the Mitochondrial Genes ............................................................... 17
  Data Visualization ....................................................................................................... 17
Results ............................................................................................................................. 17
  Section 1 (*SpB* vs. *SpA* Comparison) ................................................................ 17
  Section 2 (*A EU* vs *A MI* Comparison) ................................................................ 19
Discussion ....................................................................................................................... 23
  Patterns of Divergence ............................................................................................... 23
  Mito-nuclear Coevolution in *Saccharomyces* Yeasts ............................................. 24
  The *cis*-Regulatory Hypothesis ................................................................................ 24
  Limitations .................................................................................................................... 24
  Future Directions ........................................................................................................ 24
References ....................................................................................................................... 26
Appendix ......................................................................................................................... 32
Acknowledgements

First, I would like to thank my mom for supporting my education all the way through.

I would also like to thank Dr. Murphy for being a fantastic advisor. I still remember taking my freshman seminar with you three years ago. Thank you for introducing me to evolutionary biology and getting me involved in this project. I can’t appreciate more for the time and effort you’ve spent on me.

I would like to thank Dr. Forsyth for being a great friend and mentor. Thank you for being extremely supportive on everything I’ve come through in college. Thank you for making me feel like home as I stumble through the challenges of living and studying in a foreign country. I know I would not achieve what I have done without you on my team.

I would like to thank Dr. Iaci and Dr. Puzey for being on my defense committee. I would like to thank everyone in Murphy’s lab for all your support as I finish up my thesis, especially Rachel Rambadt for your help on this project.

Last but not least, I would like to thank all my friends for the best four years in my life, at William and Mary.
Abstract

The evolutionary processes leading to the generation of new species has been studied extensively in plants and animals; however, due to the challenges of studying microbes, microbial speciation has received less attention. This project aims to thoroughly characterize a case of eukaryotic microbial speciation on the genetic level, specifically, the effects of migration and divergence on the wild yeast, *Saccharomyces paradoxus*. Previous studies have shown there are two isolated populations— one in North America, and one in Europe— and a third migrant population that originally came from Eurasia but is currently inhabiting North America. The migrant population has been genetically diverging since its arrival and now avoids mating with the North American population, suggesting an on-going speciation process.

This research used publically available genomic data, as well as data collected in our lab, to quantify genetic differences between the three populations. I investigated all of the protein-coding genes in the wild yeast genome to determine the effects of migration and adaptation to a new environment. The results showed that the European and migrant populations are undergoing the very beginning of speciation. One nuclear gene, *PET111*, which encodes a mitochondrial regulatory protein, appears to have been under significant positive selection, indicating the possibility of mito-nuclear coevolution.

The importance of adaptive mutations in protein-coding vs. regulatory regions of the genome has been a hotly-debated topic in evolutionary developmental biology (“evo devo”). To address this controversial question, for each gene in the genome, I quantified neutral genetic divergence within the gene and compared it to the nucleotide diversity of the adjacent cis-regulatory regions. Confirming the “evo-devo” tenet, the results showed that more changes are accumulating in the cis-regulatory regions than in the protein-coding regions under neutral, and that the regulatory variation may be under selection in the diverging migrant population.
Introduction

The Speciation of Microbial Eukaryotes

Since the publication of Charles Darwin’s groundbreaking work in 1859, On the Origin of Species, evolutionary biology has become the cornerstone of modern biology. For decades, scientists have been studying the mechanisms generating differences among organisms, which ultimately lead to the formation of new species. Not only do studies on the mechanism of speciation shed light on the evolution of new species, but they ultimately address the generation of the biodiversity found on the planet. The most well-studied examples of speciation are in geographically isolated (i.e., allopatric) populations of plants and animals, as the phenotypic differences are easily observed and quantified in nature (Coyne 2010, Coyne and Orr, 2004).

Despite the ubiquitous presence of microbes on earth, their speciation, particularly that of microbial eukaryotes, has not yet been investigated systematically (Kuehne, et al., 2007). Similar to animals and plants, and unlike bacteria and archaea, microbial eukaryotes contain membrane-bound nuclei and can reproduce sexually (Heitman, et al., 2015). Theses similarities suggest the possibility that speciation models developed for multicellular organisms may be appropriate for microbial eukaryotes (Murphy and Zeyl, 2015). However, it is extremely challenging, if not impossible, to observe microbes in their natural environment (e.g., Boynton, et al., 2016). As such, it is difficult to identify the phenotypic and genetic variation relevant to fitness in microbes. Furthermore, many eukaryotic microbes have complex lifecycles with various levels of sexual reproduction (Xu, 2004). For these reasons, questions regarding the evolutionary processes generating new species in microbial eukaryotes still remain open.

Since geographical isolation is known to be a crucial factor in causing the divergence between plant and animal populations during allopatric speciation, the biogeography of microbes could also be important in understanding microbial speciation. Some have hypothesized that the distribution of eukaryotic microbial species is rarely constrained by geographical barriers due to their abundance in nature (Finlay, 2002), while others have done empirical research to show that biogeography in fact can play a role in allopatric speciation of sexually reproducing microbial eukaryotes (Kuehne, et al., 2007, Taylor, et al., 2006, Martiny, et al., 2006). Indeed, many questions regarding the speciation of eukaryotic microbes still remain.
Saccharomyces Yeasts as Models for Evolutionary Genetic Studies

Among all existing microbial eukaryotes, *Saccharomyces* yeasts are probably the most thoroughly studied group, with *Saccharomyces cerevisiae*, commonly known as the brewer’s yeast, serving as a major biomedical model for decades. Twenty years ago, *Saccharomyces cerevisiae* became the first eukaryotic species to have its genome entirely sequenced (Goffeau, et al., 1996, Dujon and Louis, 2017). Since then, whole genome sequencing projects have generated a wealth of data about the natural genetic variation of *S. cerevisiae* and related *Saccharomyces* species (Liti, et al., 2009; Schacherer, et al., 2009; Strope, et al., 2015; Yue, et al., 2017, Peter, et al., 2018). The availability of this data has greatly advanced our understanding of the evolution of eukaryotic genomes (Cherry, et al., 1998; Engel and Cherry, 2013; Yue, et al., 2017; Peters, et al., 2018).

The wild yeast *S. paradoxus* is a well-studied relative, and the closest known sister species, of *S. cerevisiae* (Johnson, et al., 2002). It is commonly found in woodlands on different continents around the world, with its genetic diversity largely related to biogeography (Dujon and Louis, 2017; Fig 1). Unlike *S. cerevisiae*, which has long been associated with human activity due to its use in beer brewing and bread making, there is no evidence that *S. paradoxus* has been domesticated by humans (Liti, et al., 2009). Therefore, *S. paradoxus* has been a popular model organism for evolutionary genetic studies on natural populations and could be an important model to study speciation in a sexual microbe.

Figure 1: A woodland near Lake Michigan where *S. paradoxus* was isolated. (Murphy and Zeyl, 2015)
Saccharomyces yeasts include sexual reproduction in their life cycle. Saccharomyces species have two mating types: \textit{a} and \textit{α}. Yeast mating proceeds through the production of mating-type specific pheromones, which are recognized by the opposite mating type, cell fusion and ultimately, nuclear fusion (Herskowitz, 1988; Fig 2). This process involves proteins called sexual adhesions that allow the cells to adhere to one another (Terrance and Lipke, 1987). They are encoded by \textit{AGA1}, \textit{AGA2}, \textit{FIG2}, and \textit{SAG1} (Dranginis, et al., 2007, Fig 3). Previous studies have shown that these genes have evolved significantly faster than other cell surface genes between \textit{S. cerevisiae} and its closest relative \textit{S. paradoxus}, suggesting that the sexual adhesin genes may play a role in mate recognition and potentially the speciation of these yeasts (Xie, et al., 2011, Fig 4).

\textbf{Figure 2:} Lifecycle of the budding yeast. (Herskowitz, 1988)

\textbf{Figure 3:} Schematic of the structure of a sexual adhesin. (Verstrepen and Klis, 2006)

\textbf{Figure 4:} Rates of non-synonymous and synonymous change in cell surface genes between \textit{S. cerevisiae} and \textit{S. paradoxus}. (Xie et al., 2011)

\textbf{Sympatric Speciation in \textit{S. paradoxus}}

Recent studies on the naturally-existing populations of \textit{S. paradoxus} in North America have shed light on a curious case of sympatric speciation. A case of on-going homoploid hybrid
speciation, or the formation of new hybrid species, was reported in indigenous populations (Leducq, et al., 2016). The nascent species was shown to have evolved from the hybridization of two genetically distinct and differentially distributed North American lineages; the hybrid has been reproductively isolated from its parent lineages due to chromosomal rearrangements and differences in ecological niche. This type of speciation event is similar to those found in polyploid plants (Rieseberg and Willis, 2007). Interestingly, another study focused on the divergence between the two parent lineages. To identify the pattern of divergence, researchers investigated the protein-coding regions in the genome and concluded that negative selection plays a greater role than positive selection in driving the divergence between the two incipience species of wild yeast (Eberlein, et al., 2017, Fig 5).

**The Migration of *S. paradoxus***

Challenging Baas Becking’s famous hypothesis that suggested “everything is everywhere, the environment selects,” (Bass-Becking, 1934), which was reasserted to describe the biogeography of microbes (Finlay, 2002), numerous studies on wild yeast have found evidence that these naturally existing eukaryotic microorganisms are diverging at different geographical locations over the globe (Kuehne, et al., 2007, Taylor, et al., 2006, Martiny, et al., 2006, Dujon and Louis, 2017).

One study on *S. paradoxus* inhabiting woodlands in North America and Eurasia (specifically, Europe and western Asia) showed that there are two clearly distinct genetic groups of the same species that have been evolving independently on the two continents for tens of millions of generations (Kuehne, et al., 2007). Some of the current North American resident strains show close genetic similarity to those that are found in Eurasia, but display significant genetic isolation against other North American residents that are found in the same environments. This provides strong evidence that there was a transcontinental migration event, which allowed
the wild yeast strains from Eurasia to colonize North America. The two distinct genetic groups now co-habit the same woodlands (Fig 6). For simplicity, the populations of Eurasian origin will be referred to as genetic group A (“SpA”), and the native north American residents will be referred to as genetic group B (“SpB”).

Although members of genetic group A and B, which evolved separately on two different continents, are now experiencing secondary contact upon migration, genetic admixture is prevented due to the decrease in offspring fitness of the hybrids. Experimental evidence shows that they are reproductively incompatible now—the hybrids between the genetic groups have low spore viability (Kuehne, et al., 2007; Murphy and Zeyl, 2015).

Further investigation into the three populations of *S. paradoxus* involved in this migration event, (the European (A_EU), migrants (A_Mi) from Eurasia, and the North American (B_NA)), showed that the migrants have evolved mating preferences against the North American residents (Murphy and Zeyl, 2015). This finding is consistent with the process of reinforcement, or selection to avoid mismating, and provides further evidence for on-going allopatric speciation of natural microbial eukaryotic population. The migrants are likely currently under stronger selection to avoid mating mistakes as they are in secondary contact with the native North American residents. While adapting to the new environment, the migrant population is expected to further diverge from its parent European population.

Unlike the previously described example of sympatric speciation, this case of speciation is similar to the more "classic" cases of speciation: allopatric divergence, accumulation of
genetic differences, and reinforcement upon secondary contact. This suggests that these sexual eukaryotic microbes may undergo speciation in a manner similar to multicellular organisms.

The molecular mechanism through which the mate choice behavior evolved in this incipient species of *S. paradoxus* is still undetermined. Furthermore, the genomic effects of the migration event, and subsequent adaptation to a new environment, have also not been investigated. The *populations involved in this transoceanic migration event of* *S. paradoxus* *can serve as an important model for studying the speciation of sexual microbial eukaryotes and may give some insights into the evolution of mating behavior in unicellular organisms.*

**The “Evo-Devo” Debate**

Mutations are the ultimate source of genetic variation upon which evolutionary forces act. Under selection, adaptive mutations may accumulate and lead to divergence among populations, and ultimately, contribute to the formation of new species. In the past decade, there has been a debate about the type of adaptive mutations that are more likely to accumulate in the genome and lead to differences between species: those that change the actual amino acid composition of a protein (mutations in the protein-coding regions of a gene), or those that change the location, time, and amount of expression of different genes (mutations in the regulatory regions of a gene) (Stapley, et al., 2010; Stern and Orgogozo, 2008; Hoekstra and Coyne, 2007; Seehausen, et al., 2014). The first type of mutation, those in protein-coding regions, refer to changes within an open reading frame of a gene that can be transcribed and translated into protein. The second type of mutation, those in regulatory regions, are part of the non-coding regions that regulate the transcription of nearby genes, but are themselves not transcribed or translated into protein. The most commonly studied type of regulatory region is the *cis*-regulatory region, which includes the promoter located immediately upstream of the transcription start site and the terminator sequence directly following the coding region (Fig 7).

![Figure 7: Schematic of the types of mutations that can cause adaptive phenotypic differences among organisms. (Carroll, 2008)](image_url)
The gene expression of a cell is controlled by its gene regulatory network. As transcription factor proteins bind to *cis*-regulatory regions within the genome, expression of different genes is often tightly controlled (Carroll, 2008). In evolutionary developmental biology (“evo-devo”), a long established hypothesis predicts that adaptive mutations causing morphological variation should be more common in the *cis*-regulatory regions (Carroll, 2008; Stern and Orgogozo, 2008). Based on this hypothesis, *cis*-regulatory evolution, rather than protein-coding, would contribute more to the overall evolutionary process leading to speciation (Seehausen, et al., 2014). This hypothesis has been tested in numerous taxa, including *S. cerevisiae*, and studies have found evidence that supports this “evo-devo” theory: adaptive mutations are more likely to exist in *cis*-regulatory than protein-coding regions of the genome, implying that *cis*-regulatory evolution might contribute more to the divergence between species (Borneman, et al., 2007; Schaefke, et al., 2013; Emerson, et al., 2010; Tirosh, et al., 2009). Similar patterns supporting the *cis*-regulatory hypothesis have also been found in *Drosophila* species and humans (Wittkopp, et al., 2004; Andolfatto, 2005; Haygood, et al., 2010).

In contrast to the *cis*-regulatory hypothesis, some researchers have suggested that the contribution of structural variation (i.e., protein-coding mutations) to adaptation and speciation is of paramount importance (Hoekstra and Coyne, 2007). Interestingly, based on research done in rodents, some researchers argued that although more adaptive changes tend to appear in regulatory regions, the fitness effects of amino acid substitution may surpass that of regulatory mutations, suggesting the predominant role of protein-coding evolution (Halligen, et al., 2013). In addition, other studies have shown that in primates, since most of the changes in regulatory regions are strongly-deleterious, the fraction of adaptive regulatory mutations under positive selection is, in fact, very low (Gaffney, et al., 2008; Keightley, et al., 2005; Necsulea and Kaessmann, 2014).

As the “evo-devo” debate goes on, there is still a need to explore the different roles of protein-coding and regulatory evolution in speciation. As we enter the genomic era, with the research opportunities provided by next generation sequencing technology, we are now able to collect more empirical data to resolve this controversy, especially by using well-studied model organisms such as the *Saccharomyces* yeasts (Stapley, et al., 2010; Necsulea and Kaessmann, 2014; Seehausen, et al., 2014).
Research Aims

The three populations of *S. paradoxus* that are involved in the transcontinental migration event provide an opportunity to study the process of speciation, divergence, and adaptation on two different timescales. The North American and European populations have been diverging for approximately 38,000 years on different continents (assuming 2,920 generations per year), show hybrid incompatibility, and thus represent the end of the speciation process (Kuehne, et al., 2007). In contrast, the European and its migrant population are in the beginning of diverging and have been separated from one another for only 1,200 years (assuming 2,920 generations per year) as the migrant adapts to a new environment (Murphy and Zeyl, 2015). Taken together, these populations represent a powerful system to study the speciation of sexual microbial eukaryotes in nature. Using a comparative population genomic approach, we can study the two populations that are deeply diverged and the two populations that recently started to diverge.

In light of the established hypothesis on the evolution of macroorganisms, my research into the three wild yeast populations addressed the following questions on the speciation of sexual microbial eukaryotes:

**North American (SpB) vs. European/Migrant (SpA) Comparison**

The European and North American populations, which have undergone divergence due to long-term geographical isolation, are now ending the speciation process; there appears to be reinforcement and evolution of mating preference behavior in the migrants, which are now in secondary contact with the North American population.

**Aim 1a:** What are the patterns of divergence between *SpA* and *SpB*? Can we detect selection in the genome? Are the majority of genes under purifying or positive selection?

**Aim 1b:** Are the sexual adhesin genes, which are involved in yeast mating and known to evolve at a high rate, different in *A_{MI}* than *A_{EU}*? Are these genes undergoing positive selection as a final touch on the speciation process between *SpA* and *SpB*?

**European (A_{EU}) vs. Migrant (A_{MI}) Comparison**

For the migrant strains colonizing North America, the split from their parent European population is evolutionarily much more recent than the separation between the native North American and European residents. I was interested in 1) characterizing the beginning of the
process of divergence and adaptation to a new environment and 2) acquiring evidence to the “evo-devo” debate on the location of adaptive mutations occurring in the genome.

**Aim 2:** Are there any genes that may indicate the adaptation of AMI to their new living environment? Can we identify any specific class of genes under positive selection upon migration?

**Aim 3:** Where in the genome are new mutations accumulating under selection: regulatory regions or protein-coding regions? Can we find any genomic evidence in support of or against the cis-regulatory hypothesis?

**Materials and Methods**

**Strains and Sequencing of Genomes**

This research focused on strains from the three populations: 4 European resident strains which I refer to as AEU; 14 North American migrant strains, which I refer to as AMI; and 8 North American resident strains, which I refer to as BNA. Whole genome sequence data were obtained for each strain from our own sequencing effort, or downloaded from NCBI (Accession: PRJNA277692) (Leducq, et al., 2016). Reference genomes CBS432 and YPS138, for the A and B genetic groups respectively, were obtained from the Yeast Population Reference Panel (YRPR; https://yjx1217.github.io/Yeast_PacBio_2016/data/) (Yue, et al., 2017). See Appendix Table 1 for strain names, genetic group, sampling location, and sequencing coverage information.

**Whole Genome Sequencing**

For the data generated for this study, yeast cultures were grown overnight in 10 ml of YPD (1% yeast extract, 2% peptone, 2% dextrose). Genomic DNA was extracted from ~400-800 mg of cells using anion-exchange gravity flow columns (Genomic-Tip, Qiagen) and sent to the Duke Genome Sequencing & Analysis Facility. Libraries were prepared according to Illumina’s protocol and used for 50bp single-read sequencing on an Illumina HiSeq2000 sequencer. All strains were multiplexed on a single lane.

**Sequence Alignment**
Raw sequencing reads of all strains from genetic group A (A_{EU} & A_{MI}) were aligned to the CBS432 reference genome (Yue, et al., 2017) and reads from genetic group B (B_{NA}) were aligned to YPS138 as the reference genome (Yue, et al., 2017) using BWA (Li, et al., 2009); SNPs were called using Freebayes (Garrison and Marth, 2012). Using a custom Python-based pipeline, SNPs were filtered for an overall quality of 40 or greater, and a frequency of 0.3 or greater, and used to generate a genome that represented the strain.

**Data Curation: Extracting Sequences**

**Protein-coding Regions**

The reference open reading frames for all genes of the two genetic groups were obtained from the coding sequence data of CBS432 (for genetic group A) and YPS138 (for genetic group B) published in Yue, et al., 2017. Excluding the non-reference genes (those that do not exist in Saccharomyces Genome Database (SGD; https://www.yeastgenome.org/)) and ambiguously annotated genes (due to computational constraints), a final list of 5352 protein-coding genes in genetic group A was generated for downstream analysis; 5312 of these genes were shared by both genetic group A & B. Using the genome annotations for the two reference strains (CBS432 & YPS138), genes with introns were separated from those without introns.

Custom pipelines were written in Python to extract the coding sequence for each gene (with no introns) from all 26 genomes by calling BLASTN (Altschul, et al., 1990). For each gene that contained introns, using annotated coordinates (Yue, et al., 2017), sequences of the whole gene were extracted and the exons were concatenated to obtain the complete coding sequence.

- *SpB vs. SpA* Comparison (26 genomes)

Open reading frames in genetic group A and B turned out to be highly variable for a large number of genes. To maintain the integrity of the downstream analyses, only genes whose length was within 6 nucleotides (2 amino acids) of the reference genome were kept for analysis. This resulted in a total of 4761 genes to be analyzed for the B_{NA} vs A_{EU}/A_{MI} comparison (out of the initial 5312). The coding regions were aligned using MAFFT alignment program (Katoh, et al., 2002), and the aligned sequences were inspected for non-sense mutations. Rare sequences with indels causing frameshifts were removed from the analysis.

- *A_{EU} vs. A_{MI}* Comparison (18 genomes)
Since strains in genetic group A were more similar, insertions and deletions were less common between \( A_{EU} \) & \( A_{MI} \). Coding regions of all 5352 genes shared within genetic group A were aligned using MAFFT alignment program, and the aligned sequences were inspected for non-sense mutations. Rare sequences with indels causing frameshifts were removed from the analysis.

**Sexual Adhesin Genes**

To study the evolution of sexual adhesin genes (i.e., \textit{AGA1}, \textit{AGA2}, \textit{FIG2}, and \textit{SAG1}) between \textit{SpA} \& \textit{SpB}, we focused on the subset of 73 cell surface genes listed in Xie, et al., 2011. The coding sequences of these genes (annotated by their common names) were extracted from all 26 genomes (\( A_{EU} \), \( A_{MI} \) & \( B_{NA} \)) by “blasting” reference sequences downloaded from YRPR and SGD. Extracted sequences were aligned using MAFFT alignment program and curated manually for non-sense mutations. Rare sequences with indels causing frameshifts were removed from the analysis.

**Cis-regulatory Regions**

Based on multiple studies done on \textit{S. cerevisiae} and \textit{S. paradoxus}, most functional transcription factor binding sites in the promoter next to each gene (proximal regulatory elements) can be found in the region up to 200 base pairs (bp) upstream of the transcription starting site (Schaeffke, et al., 2015; Lin, et al., 2010). In contrast to the promoter, another important \textit{cis}-regulatory element located at the 3’ region after each gene, known as the terminator, has been overlooked in yeast (Curran, et al., 2013). Thus, to study the \textit{cis}-regulatory elements in the yeast genome, the following regions were investigated: (1) 200 bp upstream from the transcription starting site and (2) 200 bp downstream from the stop codon for every gene. Using the gene coordinates generated by BLASTN in the coding sequence extraction (see above), custom Python programs were used to extract the 200bp upstream and downstream sequences from all genomes of genetic group A (\( A_{EU} \) & \( A_{MI} \)). These sequences were then aligned using MAFFT. Sequences with rare mutations causing frameshift (i.e. insertions and deletions) in the coding regions were removed.

**Data Analysis: Investigation into the Protein-coding Regions**
**dN/dS Analysis**

To detect a signal for positive selection, the rates of non-synonymous (dN) and synonymous (dS) change was calculated for each pair-wise comparison for all coding sequences using the yn00 program of Phylogenetic Analysis by Maximum Likelihood (PAML) software (Yang, 2007). Estimates of dN and dS were based on counting methods proposed by Yang and Nielsen (2000). Pair-wise comparisons were categorized by the populations to which the two strains belong (i.e. B⁷NA vs B⁷NA, A⁷MI vs A⁷MI, A⁷EU vs A⁷EU, B⁷NA vs A⁷EU, B⁷NA vs A⁷MI, or A⁷EU vs A⁷MI).

**MK test**

To identify specific genes under significant positive selection, the McDonald–Kreitman (McDonald and Kreitman, 1991) test was performed on all genes using the MKT function of PopGenome (Pfeifer, et al., 2014) package in R. Populations were defined as A⁷EU, A⁷MI and B⁷NA. A Fisher’s exact test was used to determine the statistical significance of the evidence for selection. A neutrality index, which indicates the direction of selection, was also calculated using PopGenome. It is calculated using the following equation proposed in Rand and Kann, 1996:

\[
\text{Neutrality index } = \frac{\text{no. of polymorphic nonsynonymous sites}}{\text{no. of fixed nonsynonymous sites}} \times \frac{\text{no. of polymorphic synonymous sites}}{\text{no. of fixed synonymous sites}}.
\]

**Investigation into the cis-regulatory Regions**

**Neutral Model**

To determine the level of potentially adaptive mutations in cis-regulatory regions, we compared regulatory variation to a neutral model. The rate of neutral evolution at a given locus was estimated by the synonymous nucleotide diversity (\(\pi_s\)) of that gene’s coding region. Nucleotide diversity of the corresponding regulatory region (\(\pi\)) was also calculated for 200 base pairs upstream from the transcription start site and 200 base pairs downstream from the stop codon for each gene. Under neutral evolution, the diversity in the regulatory regions is expected to be equal to the diversity of the synonymous sites in the coding regions. I compared the \(\pi_s\) and \(\pi\) (upstream and downstream) for each gene and looked for any deviation from the neutral evolution model that would indicate evidence for selection force. This approach had the advantage of controlling for variation in mutation rate across the genome.
Nucleotide diversity ($\pi$) and Synonymous nucleotide diversity ($\pi_s$)

Nucleotide diversity measures the level of polymorphism within a population for a given sequence. Synonymous nucleotide diversity is the nucleotide diversity measured at only synonymous substitution sites. The two values for each locus were calculated by calling the diversity.stats function in PopGenome and then dividing by the number of nucleotides (n.site) of that sequence (Pfeifer, et al., 2014). This function implemented the nucleotide diversity estimation by Nei’s method (1979). $\pi$ and $\pi_s$ were calculated for 1) AEU and AMI as separate populations and 2) AEU and AMI as one population.

Investigation into the Mitochondrial Genes

The protein-coding genes in the wild yeast mitochondrial genomes were also investigated. The mitochondrial genomes of the 26 strains from genetic group A and B were sequenced and aligned using the same data sources and computational methods as described for the nuclear genomes (Appendix Table 1). To obtain the level of synonymous and nonsynonymous substitutions in mitochondrial protein-coding genes, the same procedure to extract their open reading frames and conduct dN/dS analysis on the coding sequences was followed.

Data Visualization

All data and results were plotted in R using gglot2 for visual analysis (Wickham, 2009).

Results

Section 1 (SpB vs. SpA Comparison)

Aim 1a: Patterns of Divergence

The dN/dS analysis and MK test for all protein-coding regions in the wild yeast genome yielded consistent results regarding the patterns of divergence between the two genetic groups. For both AEU vs. BNA and AMI vs. BNA comparisons, the majority of protein-coding genes appear to be under purifying selection, whereas only a few of them are under positive selection (Fig 8a-b; Appendix Table 2&3). These results agree with the divergence patterns observed in another study of incipient wild yeasts, which investigated the two distinct lineages of native North American residents, SpB and SpC (Eberlein, et al., 2017, Fig 5). The dN/dS analysis from the
current study also shows that more genes have a dN/dS ratio greater than 1 within each genetic group than between genetic groups (Fig 8c), suggesting that the changes in these protein-coding regions are likely deleterious in the long-term and might be purified from the populations as the process of divergence continues.

Figure 8: Genetic divergence between genetic group A and B. (A-B) MK test of all protein-coding genes. Points above the dashed line are significantly different than neutral; dark shaded box indicates significant positive selection. (C) dN/dS analysis of all protein-coding genes. Points above the line indicate positive selection; pink: B_{NA} vs B_{NA}; emerald: A_{MI} vs A_{MI}; red: A_{EU} vs A_{EU}; green: B_{NA} vs A_{EU}; blue: B_{NA} vs A_{MI}; brown: A_{EU} vs A_{MI}.

**Aim 1b: Sexual Adhesin Genes**

The dN/dS analysis of the 73 cell surface genes showed no evidence that the sexual adhesion genes were under positive selection between genetic group A and B, although they are
clearly accumulating more nonsynonymous substitutions than other classes of cell surface genes (Fig 9a), as was shown for comparisons between *S. cerevisiae* and *S. paradoxus* (Fig 4)(Xie, et al., 2011). Furthermore, the dN/dS ratios of the sexual adhesin genes between *A EU vs A MI* do not indicate strong evidence for positive selection since the migration event (Fig 9b). Investigation of the alignment of the exact amino acid sequences encoded by these genes shows that there are no fixed differences in the migrants in any of the four sexual adhesin genes (Appendix Fig A2). Therefore, mutations in the sexual adhesin genes are unlikely to explain the evolution of mating *SpA vs SpB* preference in *A MI*.

**Figure 9:** Genetic divergence of 73 cell surface genes, including 4 sexual adhesin genes. (A-B) Points above the line indicate positive selection; red: sexual adhesin genes; black: other cell surface genes.

**Section 2 (A EU vs A MI Comparison)**

**Aim 2: Genes Under Positive Selection**

Both the dN/dS analysis and MK test of all protein-coding genes within genetic group A show that the migrant and European populations are undergoing the beginning of speciation process (Fig 10ab). As expected for a recent migration event, the level of overall divergence between the populations is very low: despite the existence of nonsynonymous substitution in a number of genes, few of these mutations are fixed between the two populations. Excitingly, one protein-coding gene (out of a total of 5352), *PET 111*, was identified as under significant positive selection based on the MK test result (Fig 10c). The exact alignment of this gene
indicates the presence four fixed nonsynonymous substitutions between $A_{EU}$ and $A_{MI}$ (Appendix Fig A3), confirming the significant result from the MK test. In addition, the dN/dS ratios of $PET111$ between all $A_{EU}$ and $A_{MI}$ pairwise comparisons are above 1, which also appear as evidence for positive selection within this gene (Fig 10d). I also investigated the level of divergence of $PET111$ between $SpA$ and $SpB$. Both the dN/dS analysis and MK test show that this gene is not under positive selection between the two genetic groups, thus indicating that the changes accumulated between $A_{EU}$ and $A_{MI}$ are more likely the effect of migration (Appendix Table 4).

Figure 10. Genetic divergence within genetic group A. (A) MK test for all-protein coding genes. Points above the dashed line are significantly different than neutral; dark shaded box indicates significant positive selection. (B) dN/dS analysis of all protein-coding genes. Points above the line indicates positive selection; green: $A_{EU}$ vs $A_{MI}$; red: $A_{EU}$ vs $A_{EU}$; blue: $A_{MI}$ vs $A_{MI}$; (C) MK test and dN/dS analysis both show evidence that $PET111$ is under positive selection. Red: $PET111$; black/gray: other genes.
PET111, located in the yeast nuclear genome, is known to encode for a mitochondrial regulatory protein that controls the expression of COX2 (cytochrome c oxidase subunit II), a mitochondrial protein essential for respiration (Mulero and Fox, 1992; Poutre and Fox, 1986). Investigation into the eight protein-coding genes in the mitochondria shows that nonsynonymous changes in COX2 are occurring both between and within AEU and AMI, and thus not much can be concluded from this dN/dS analysis (Fig 11).

**Aim 3: Cis-regulatory Evolution**

Based on our neutral model, there is clear evidence that within genetic group A, more new mutations are accumulating in the cis-regulatory than the synonymous sites in the protein-coding regions in the genome. Similar patterns are found in both upstream and downstream 200 base pair analyses (Fig 11a-b). Interestingly, it appears that overall the \( \pi/\sigma_s \) values tend to be higher in AMI than AEU (Fig 11c-f). This may suggest that the cis-regulatory elements in the migrant population are currently under selection as a result of the migration event followed by the adaptation to new environment, although the level of difference in \( \pi/\sigma_s \) between the two populations has to be measured by appropriate statistical testing to ensure its significance.

Viewing AEU and AMI as one population, the results also allow us to identify one locus with fixed changes in its adjacent cis-regulatory elements (both upstream and downstream) between the two populations. We need to further investigate the reasoning behind the existence of this outlier gene.
Figure 11. Mutations in cis-regulatory vs. protein-coding regions within genetic group A.

(A-F) Points above the line have more changes in cis-regulatory regions (upstream or downstream 200 bp) than synonymous mutations within the gene.
Discussion

This study thoroughly investigated the genomic consequences of a transcontinental migration event that involves three populations of *S. paradoxus* from two diverging genetic groups. We characterized the patterns of divergence between two deeply-diverged populations and two recently-diverged populations of wild yeast, which has allowed us insights into the speciation of naturally existing sexual microbes. One mitochondrial regulatory protein, *PET11*, was identified as the only one under significant positive selection, likely to be the result of migration, and suggests the possibility of mito-nuclear coevolution. In addition, we have found evidence supporting the *cis*-regulatory hypothesis; that is, more new mutations are accumulating in the *cis*-regulatory elements upon migration than would be expected under neutral evolution. Finally, despite observing an elevated rate of non-synonymous change in the sexual adhesin genes, the molecular mechanism of the evolved mate preference remains elusive. Thus, questions regarding the evolution of mating choice in the migrant population still remain.

Patterns of Divergence

The overall pattern we observed was the level of genetic divergence between A*EU* and A*MI* is much lower than that between *SpA* and *SpB*. This makes sense because the two genetic groups have been evolving independently for a much longer time than the relatively recent migration event. This result is consistent with the pattern obtained from the nine intron loci investigated in Kuehne, et al., 2007. Interestingly, we observed more evidence for purifying selection *between* the genetic groups than *within* each genetic group, implying that most mutations in the protein-coding regions are deleterious in the long-term, a trend that has been shown in models of macroorganisms (Eberlein, et al., 2017). Nonetheless, the pattern that emerged in the deeper-divergence was beginning to form in the less-divergent populations. This suggests that in the long-term, the migrant and Eurasian populations will show a pattern of divergence similar to that exhibited by the North American resident and Eurasian populations. We compiled two lists of genes that are under significant positive selection between genetic groups A and B, and populations A*MI* and A*EU*, respectively (Appendix Table 2-3). Further inspection into the two lists may provide more insights on the effects of migration, adaptation, and speciation of wild yeast.
Mito-nuclear Coevolution in Saccharomyces Yeasts

Numerous lines of evidence suggest that PET111, a mitochondrial regulatory protein, has been under positive selection since the migration of A_EU. This raises the possibility of mito-nuclear coevolution in SpA of S. paradoxus. There have been multiple reports on the interaction between mitochondrial and nuclear genomes in Saccharomyces yeasts, suggesting the presence of mito-nuclear coevolution in these sexual microbes (Wolters, et al., 2015; Paliwal, et al., 2014; Leducq, et al., 2017). In addition, variation in carbon sources is known to affect respiration and fermentation of Saccharomyces yeasts (Duenas-Sanchez, et al., 2012; Gancedo, 2008). Studies have also shown that divergence in substrate use due to spatial distribution in S. paradoxus can play a role in speciation (Leducq, et al., 2016; Samani, et al., 2015; Deken, 1966). These findings inspired us to further investigate whether mito-nuclear incompatibility within the two population of SpA could be a possible result of the migration event and adaptation to new environment, likely leading to the incipient speciation of A_EU. With the help of Rachel Rambadt, we are currently investigating the compatibility of the mitochondrial and PET111 alleles in A_EU and A_MI strains to test mito-nuclear incompatibility within SpA (Fox, et al., 1991).

The cis-Regulatory Hypothesis

We have found evidence that the regulatory regions are accumulating variation faster than other neutral regions. This could imply regulatory variation is an important part of adapting to a new environment. This result is consistent with work done in a variety of taxa from yeast to mammals, supporting the cis-regulatory hypothesis. It also suggests that the evolution of sexual microbial eukaryotes may not be that different from the evolution of macroorganisms.

Limitations

The dN/dS analysis considers the signature of selection in the whole genes. There may be important individual codons that are under selection that would be lost using this approach. While this work was attempting to understand large patterns, a sliding-window FST analysis may uncover significant sites (Seehausen, et al., 2014).

Future Directions
In preparation for submitting a manuscript on these results, I will conduct more computational analyses to further complete our investigation into the three populations of *S. paradoxus* that are involved in the transcontinental migration event. First, appropriate statistical tests will be done for our neutral model in order to measure the statistical significance of differences between the $\pi/\pi_s$ ratios of $A_{EU}$ and $A_{MI}$, which should yield more compelling evidence regarding the level of *cis*-regulatory changes that occurred after migration. Moreover, we still need to investigate the outlier genes that appeared in our *cis*-regulatory analysis and make sure there is no substantial error in our genome alignments, which may have generated this result. Second, to complement our population genomics analysis, I will also conduct a genome-wide $F_{ST}$ sliding window analysis to obtain more evidence of divergence between the two populations after migration. This will allow me to determine whether any nucleotides/regions of the genome (both coding and regulatory) are fixed for differences between the populations. I expect *PET111* to be a significant hit in this genome-wide $F_{ST}$-analysis. Finally, I will construct a phylogenetic tree using whole-genome sequences to show more evidence for the divergence within *SpA* as a result of migration and adaptation to new environment.
References


Xu, J. (2004). The prevalence and evolution of sex in microorganisms. *Genome, 47*(5), 775-780. doi:10.1139/g04-037


Yue, J., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., . . . Liti, G. (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics, 49*(6), 913-924. doi:10.1038/ng.3847
# Appendix

<table>
<thead>
<tr>
<th>Original Strain</th>
<th>Genetic Group</th>
<th>Location</th>
<th>Isolation</th>
<th>Coverage</th>
<th>Strain Ref</th>
<th>Sequence Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBS432</td>
<td>A (Eur)</td>
<td>Moscow, Russia</td>
<td>Oak tree exudate</td>
<td>ref genome</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>YPS3 (N7)</td>
<td>A (Eur)</td>
<td>Leningrad, Russia</td>
<td>Oak tree exudate</td>
<td>70X</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS5 (N9)</td>
<td>A (Eur)</td>
<td>Tashkent, Uzbekistan</td>
<td>Oak tree exudate</td>
<td>73X</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS12 (N34)</td>
<td>A (Eur)</td>
<td>Voronezh, Russia</td>
<td>Oak tree exudate</td>
<td>73X *</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS642</td>
<td>A (Mig)</td>
<td>Buck Hill Falls, PA</td>
<td>Oak tree exudate</td>
<td>80X</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS644</td>
<td>A (Mig)</td>
<td>Buck Hill Falls, PA</td>
<td>Oak tree exudate</td>
<td>79X</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS695</td>
<td>A (Mig)</td>
<td>Tuscarora Forest, PA</td>
<td>Oak tree exudate</td>
<td>70X</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS708</td>
<td>A (Mig)</td>
<td>Tuscarora Forest, PA</td>
<td>Oak tree exudate</td>
<td>74X *</td>
<td>c</td>
<td>This study</td>
</tr>
<tr>
<td>YPS714</td>
<td>A (Mig)</td>
<td>Tuscarora Forest, PA</td>
<td>Oak tree exudate</td>
<td>70X *</td>
<td>c</td>
<td>This study</td>
</tr>
<tr>
<td>YPS744</td>
<td>A (Mig)</td>
<td>Tuscarora Forest, PA</td>
<td>Oak tree exudate</td>
<td>79X</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>LL2012_001</td>
<td>A (Mig)</td>
<td>Hull, QC</td>
<td>Bark- Oak tree</td>
<td>70X</td>
<td>d</td>
<td>g</td>
</tr>
<tr>
<td>yHKS172</td>
<td>A (Mig)</td>
<td>Dunes State Park, IN</td>
<td>Bark- <em>Populus deltoides</em></td>
<td>23X *</td>
<td>e</td>
<td>g</td>
</tr>
<tr>
<td>yHKS320</td>
<td>A (Mig)</td>
<td>Sheboygan,WI</td>
<td>Bark- <em>Acer saccharum</em></td>
<td>13X</td>
<td>e</td>
<td>g</td>
</tr>
<tr>
<td>yHKS341</td>
<td>A (Mig)</td>
<td>Fayette, MI</td>
<td>Soil- <em>Acer</em></td>
<td>8X</td>
<td>e</td>
<td>g</td>
</tr>
<tr>
<td>yHKS402</td>
<td>A (Mig)</td>
<td>Fayette, MI</td>
<td>Soil- <em>Acer</em></td>
<td>17X *</td>
<td>e</td>
<td>g</td>
</tr>
<tr>
<td>yHKS421</td>
<td>A (Mig)</td>
<td>Fayette, MI</td>
<td>Soil- <em>Acer</em></td>
<td>9X</td>
<td>e</td>
<td>g</td>
</tr>
<tr>
<td>yHRM71</td>
<td>A (Mig)</td>
<td>Madison, WI</td>
<td>Soil-Unknown</td>
<td>22X</td>
<td>e</td>
<td>g</td>
</tr>
<tr>
<td>yHRM73</td>
<td>A (Mig)</td>
<td>Madison, WI</td>
<td>Soil-Unknown</td>
<td>14X</td>
<td>e</td>
<td>g</td>
</tr>
<tr>
<td>YPS138</td>
<td>B(NA)</td>
<td>Media, PA</td>
<td>Soil- Oak tree</td>
<td>ref genome</td>
<td>b</td>
<td>f</td>
</tr>
<tr>
<td>YPS484</td>
<td>B (NA)</td>
<td>Grand Bend, On</td>
<td>Oak tree exudate</td>
<td>64X</td>
<td>b</td>
<td>g</td>
</tr>
<tr>
<td>YPS616</td>
<td>B (NA)</td>
<td>New Brunswick, NJ</td>
<td>Oak tree exudate</td>
<td>55X</td>
<td>b</td>
<td>g</td>
</tr>
<tr>
<td>YPS618</td>
<td>B (NA)</td>
<td>New Brunswick, NJ</td>
<td>Oak tree exudate</td>
<td>55X</td>
<td>b</td>
<td>g</td>
</tr>
<tr>
<td>YPS631</td>
<td>B (NA)</td>
<td>New Brunswick, NJ</td>
<td>Oak tree exudate</td>
<td>55X</td>
<td>b</td>
<td>g</td>
</tr>
<tr>
<td>YPS646</td>
<td>B (NA)</td>
<td>Buck Hill Falls, PA</td>
<td>Oak tree exudate</td>
<td>69X *</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS664</td>
<td>B (NA)</td>
<td>Buck Hill Falls, PA</td>
<td>Oak tree exudate</td>
<td>72X</td>
<td>b</td>
<td>This study</td>
</tr>
<tr>
<td>YPS668</td>
<td>B (NA)</td>
<td>Buck Hill Falls, PA</td>
<td>Oak tree exudate</td>
<td>76X</td>
<td>b</td>
<td>This study</td>
</tr>
</tbody>
</table>

## Table 1: Strain and Sequence Data.
* indicates possible aneuploidy or CNV


Fig A1. Coverage information for whole-genome sequencing.
Fig A2. Sequence alignments of the four sexual adhesin genes.

YLL004W
YOR004W
YKR008W
YPL020C
YOR032C
YNR033W
YHL038C
YKL046C
YDL054C
YIL071C
YLR073C
YBR074W
YIL079C
YBL088C
YER092W
YDL099W
YPL099C
YPL116W
YJL118W
YML118W
YOR118W
YDR128W
YPR128C
YOR129C
YJL149W
YMR157C
YPL163C
YER167W
YDL177C
<table>
<thead>
<tr>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDL179W</td>
</tr>
<tr>
<td>YLR181C</td>
</tr>
<tr>
<td>YDL183C</td>
</tr>
<tr>
<td>YDR197W</td>
</tr>
<tr>
<td>YMR207C</td>
</tr>
<tr>
<td>YDR217C</td>
</tr>
<tr>
<td>YBR242W</td>
</tr>
<tr>
<td>YNL245C</td>
</tr>
<tr>
<td>YLR263W</td>
</tr>
<tr>
<td>YMR285C</td>
</tr>
<tr>
<td>YDR309C</td>
</tr>
<tr>
<td>YMR312W</td>
</tr>
<tr>
<td>YLR342W</td>
</tr>
<tr>
<td>YDR389W</td>
</tr>
<tr>
<td>YDR438W</td>
</tr>
<tr>
<td>YDR469W</td>
</tr>
</tbody>
</table>

Table 2. List of genes under positive selection between A\textsubscript{EU} & B\textsubscript{NA}

<table>
<thead>
<tr>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOR004W</td>
</tr>
<tr>
<td>YGL013C</td>
</tr>
<tr>
<td>YOL015W</td>
</tr>
<tr>
<td>YKL017C</td>
</tr>
<tr>
<td>YPL020C</td>
</tr>
<tr>
<td>YOL027C</td>
</tr>
<tr>
<td>YAL028W</td>
</tr>
<tr>
<td>YOR032C</td>
</tr>
<tr>
<td>YHL038C</td>
</tr>
<tr>
<td>YKL046C</td>
</tr>
<tr>
<td>YBL054W</td>
</tr>
<tr>
<td>YDL054C</td>
</tr>
<tr>
<td>YJL062W</td>
</tr>
<tr>
<td>Gene ID</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>YIL071C</td>
</tr>
<tr>
<td>YBR074W</td>
</tr>
<tr>
<td>YIL079C</td>
</tr>
<tr>
<td>YKR082W</td>
</tr>
<tr>
<td>YER092W</td>
</tr>
<tr>
<td>YPL099C</td>
</tr>
<tr>
<td>YPR106W</td>
</tr>
<tr>
<td>YML116W</td>
</tr>
<tr>
<td>YML118W</td>
</tr>
<tr>
<td>YPR128C</td>
</tr>
<tr>
<td>YOR129C</td>
</tr>
<tr>
<td>YNL139C</td>
</tr>
<tr>
<td>YJL149W</td>
</tr>
<tr>
<td>YMR157C</td>
</tr>
<tr>
<td>YPL163C</td>
</tr>
<tr>
<td>YOR172W</td>
</tr>
<tr>
<td>YDL179W</td>
</tr>
<tr>
<td>YDL183C</td>
</tr>
<tr>
<td>YDR197W</td>
</tr>
<tr>
<td>YMR207C</td>
</tr>
<tr>
<td>YDR217C</td>
</tr>
<tr>
<td>YLR226W</td>
</tr>
<tr>
<td>YNL245C</td>
</tr>
<tr>
<td>YMR253C</td>
</tr>
<tr>
<td>YMR255W</td>
</tr>
<tr>
<td>YLR263W</td>
</tr>
<tr>
<td>YMR312W</td>
</tr>
<tr>
<td>YLR342W</td>
</tr>
<tr>
<td>YDR469W</td>
</tr>
</tbody>
</table>

**Table 3.** List of genes under positive selection between $A_{MI}$ & $B_{NA}$. 
Fig A3. Sequence alignment of PET111.

<table>
<thead>
<tr>
<th></th>
<th>P1_ non syn</th>
<th>P2_ non syn</th>
<th>P1_ syn</th>
<th>P2_ syn</th>
<th>D_non syn</th>
<th>D_syn</th>
<th>neutrality .index</th>
<th>fisher.P.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1 vs. P2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.06666667</td>
</tr>
<tr>
<td>P1 vs. P3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>38</td>
<td>54</td>
<td>0.355263</td>
<td>0.30824048</td>
</tr>
<tr>
<td>P2 vs. P3</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>38</td>
<td>53</td>
<td>0.278947</td>
<td>0.12128554</td>
</tr>
</tbody>
</table>

Table 4. MK test results for PET111. (P1: AEU, P2: AMI, P3: BNA, D: fixed sites)