Beyond the reach of homology: successive computational filters find yeast pheromone genes

Sriram Srikanth^a, Rachelle Gaudet and Andrew W. Murray

Dept. of Molecular and Cellular Biology, Harvard University

^aCurrent address: Dept. of Biology, Massachusetts Institute of Technology

Abstract

The mating of fungi depends on pheromones that mediate communication between two mating types. Most species use short peptides as pheromones, which are either unmodified (e.g., α -factor in *Saccharomyces cerevisiae*) or C-terminally farnesylated (e.g., **a**-factor in *S. cerevisiae*). Peptide pheromones have been found by genetics or biochemistry in small number of fungi, but their short sequences and modest conservation make it impossible to detect homologous sequences in most species. To overcome this problem, we used a four-step computational pipeline to identify candidate **a**-factor genes in sequenced genomes of the Saccharomycotina, the fungal clade that contains most of the yeasts: we require that candidate genes have a C-terminal prenylation motif, are fewer than 100 amino acids long, contain a proteolytic processing motif upstream of the potential mature pheromone sequence, and that closely related species contain highly conserved homologs of the potential mature pheromone sequence. Additional manual curation exploits the observation that many species carry more than one **a**-factor gene, encoding identical or nearly identical pheromones. From 332 fungal genomes, we identified strong candidate pheromone genes in **238** genomes, covering **13** clades that are separated from each other by at least 100 million years, the time required for evolution to remove detectable sequence homology. For one small clade, the *Yarrowia*, we demonstrated that our algorithm found the **a**-factor genes: deleting all four related genes in the **a**-mating type of *Yarrowia lipolytica* prevents mating.

Introduction

Fungi are an ancient lineage of eukaryotes whose extant members shared a common ancestor more than one billion years ago [1]. Within a species, the mating depends on signaling by diffusible pheromones between haploid cells of two different mating types (Figure 1A) [2]. One important group is the *Saccharomycotina*, which emerged roughly 400 million years ago and contains most of the yeasts: unicellular fungi that lack fruiting bodies. Many yeast species play key roles as model systems, in the food industry, and as crop and human pathogens. Finding the pheromones of economically important yeast will help to understand and control their mating as well as shed light on the evolution of these molecules and the receptors that recognize them.

The best studied mating system is that of the baker's yeast *Saccharomyces cerevisiae*. Genetic screens and biochemistry have identified pheromones, the proteins that produce and export them, pheromone receptors, and their associated signaling pathways. The two pheromones in *S. cerevisiae* are the 13-amino acid peptide α -factor produced by α -cells and **a**-factor, the 12-amino acid C-terminally farnesylated peptide produced by **a**-cells (Figure 1B). Multiple copies of α -factor are present in each of two mating factor genes (*MF α 1* and *MF α 2*) and are released by proteolysis in the endoplasmic reticulum allowing the pheromone to be exported in secretory vesicles [3, 4]. In contrast, **a**-factor is encoded in two loci (*MFA1* and *MFA2*) as a precursor peptide that undergoes maturation in the cytosol to produce a single pheromone molecule (Figure 1C) [5, 6]. **a**-factor maturation involves C-terminal farnesylation and carboxymethylation, followed by N-terminal proteolysis. These steps occur in the cytoplasm, after which **a**-factor is pumped across the plasma membrane by Ste6, a member of the ABC family of ATP-dependent transporters. In *S. cerevisiae*, maturation is essential for bioactivity; unmodified pheromone is trapped in the cytosol because it is not recognized and exported by Ste6 [5, 7].

The Saccharomycotina fungi lie within the Ascomycota clade and the distinction between an unmodified α -factor and a farnesylated **a**-factor is maintained in the Ascomycota. In the Basidiomycota, the sister clade to the Ascomycota, all mating pheromones are farnesylated, suggesting that the ancestral mating pheromones were farnesylated, poorly soluble molecules [8]. The peptide sequences of both **a**-like and α -like pheromones diverge across fungal species with cognate receptors and transporters co-evolving to maintain activity [9-11].

The first fungal pheromones were identified as lipidated peptides by biochemical isolation from the extracellular medium of basidiomycetes *Rhodospiridium toruloides* [12, 13] and *Tremella mesenterica* [14, 15]. Since then, pheromones have been biochemically isolated and characterized in laboratory yeasts, filamentous fungi and crop pathogens by isolation from extracellular medium [2, 8, 16, 17] but this approach depends on discovering the environmental conditions that stimulate mating and obtaining sufficient material to determine the pheromone's peptide sequence.

An alternative route to pheromone identification is to look for homologs of known pheromones in sequenced fungal genomes. Because pheromones are small peptides and are only modestly conserved this approach only works over a limited phylogenetic distance. In principle, this problem could be solved by exploiting conservation in the order of genes along a chromosome (synteny): looking for orthologs of the longer genes on either side of the *S. cerevisiae* **a**-factor and then searching, more sensitively, for a pheromone ortholog that lies in the stretch of DNA between these two genes. Unfortunately, synteny is less conserved than pheromone sequence. A search for homologs of one of the *S. cerevisiae* genes that encode **a**-factor, *MFA2*, finds hits in two clades that diverged roughly 100 million years ago from the clade that enjoyed a whole genome duplication and includes *S. cerevisiae*: in the nearer clade (which contains *Zygosaccharomyces rouxii* and *Torulospora delbrueckii*) the **a**-factor homologs retain their syntenic location, but in the

further one (which contains *Kluyveromyces lactis*, *Lachancea waltii* and *Eremothecium gossypii*) synteny has been lost [18].

We built a computational pipeline to identify farnesylated fungal pheromones in sequenced genomes. Our algorithm successively discards sequences that lack the C-terminal signal sequence for farnesylation, lack a sequence that controls proteolytic processing, and lack strongly conserved homologs in closely related species. Manual curation privileges sequences that occur more than once in the same genome and identifies all the known pheromones in the Saccharomycotina. Overall, we find known or candidate pheromone genes in 238 of 332 mined genomes. To test the ability to identify novel pheromones, we verified that deletion of all copies of the candidate farnesylated pheromone of *Yarrowia lipolytica* (whose last common ancestor with *S. cerevisiae* existed more than 300 million years ago) prevented mating of the **a**-cells of this species.

Results

Algorithmic sieve to find candidate pheromone open reading frames (ORFs)

Given the limitations of syntenic methods and the constraints on homology search for small unstructured peptides, we designed a new approach to identify farnesylated peptide pheromones from sequenced genomes using successive identification of conserved features of previously identified pheromones across the fungal lineage as a series of sieves to identify candidate pheromone sequences.

Biogenesis of **a**-factor in *S. cerevisiae* involves a number of essential post-translation modifications by dedicated enzymes [5] (Figure 1C). We assume that these conserved enzymes play a role in all fungal pheromone maturation, and therefore our algorithm identifies candidate

pheromone genes by searching for the sequence motifs that these enzymes recognize. The post-translational maturation can be broadly separated into two parts – C-terminal farnesylation and carboxy-methylation, followed by N-terminal proteolytic cleavage (Figure 1C).

In eukaryotes, prenylation is achieved by a conserved enzyme complex recognizing a conserved sequence motif [8]. The enzyme complex (Ram1/Ram2) prenylates by S-thiol addition of a farnesyl group (a C15 isoprenoid) to the cysteine in the C-terminal sequence CAAX, with C representing the modified cysteine, A being small aliphatic residues and X being a small amino acid.

In addition to the S-farnesyl modification, the pheromones need to be carboxymethylated for bioactivity [19, 20]. The C-terminal AAX residues of *S. cerevisiae* **a**-factor are recognized by enzymes that proteolytically remove the -AAX residues (Ste24 and Rce1), and then carboxymethylate the now C-terminal cysteine (Ste14) [5]. To begin, we consolidated the CAAX motifs of known pheromones in the Ascomycota into a dictionary to identify potential pheromone candidate ORFs assuming strong conservation of the motif as a very stringent cutoff (Supp Figure 1). To avoid biases based on the small number of known pheromones, we relaxed the CAAX detection dictionary based on studies of the possible combinatorial signatures of farnesylation, proteolysis and carboxymethylation in *S. cerevisiae* [21-23] (Supp Figure 1).

We then selected the list of ORFs ending in CAAX for candidates that have an in-frame methionine within 100 residues upstream of the stop codon to find ORF candidates that are within the expected size range of pheromone genes in Ascomycota (Figure 2A). In short, our pipeline considers genome assemblies contig-by-contig and assumes that the pheromones can be positioned anywhere. Since our algorithm relies on tracing the ORF starting from the Stop codon, it fails to locate many pheromone genes that contain introns, like those in *S. pombe*. We believe this is a

reasonable simplification of the algorithm given the lower frequency of introns in Hemiascomycota yeasts (2-14% of all genes), relative to other eukaryotes [24].

We ran this pipeline on a set of 332 published Saccharomycotina genomes [25] (Figure 2B and Table 1) and found 500-2000 candidate ORFs per genome, which is far larger than can be experimentally validated even for a single species. Subsequent to C-terminal farnesylation and carboxymethylation, *Sca*-factor precursors undergo N-terminal processing in two steps [5, 19]. The second proteolysis is the final step in maturation and is mediated by one of two Zn-dependent metalloendopeptidases (Axl1 and Ste23; Figure 1C) [26]. Random mutagenesis in *S. cerevisiae* α -factor revealed that the two residues immediately before and after the cleavage site are important for *ScAxl1* cleavage [19]. Comparing known Ascomycota pheromones, we noticed that the cleavage site is typified by N-X with X representing aromatic or uncharged residues and the peptidase cleaving C-terminal of the asparagine (Figure 1C and Supp. Figure 1A). Therefore, we required pheromone candidates to have an asparagine in frame with the conserved cysteine such that the mature pheromone is between 5-20 residues from residue X following the cleavage site N to the farnesylated C. Adding this sieve for an internal cleavage site between 5-20 residues from CAAX reduces the number of unique pheromone candidates to 100-700 per genome (Figure 2C). As expected, the distribution of pheromone candidates per genome scales with genome size similar to the number of annotated protein models (Figure 2C). Candidates are still too numerous for experimental validation, but further knowledge of pheromone biology can be used to identify the best candidates.

Yeast pheromones are encoded in multiple gene copies and are conserved among closely related species

Yeast pheromones of closely related species are quite similar, although the number of copies in each genome varies (e.g. 1 each in *Candida albicans* and *C. dubliensis* [27], 2 each in *Komagataella phaffi* and *K. pastoris* [28], and 1 to 3 in different *Saccharomycetaceae* [18]). The amino acid sequence of the ORFs is most strongly conserved within the mature pheromone peptide sequence which must interact both with the **a**-factor exporter, Ste6, and the **a**-factor receptor, Ste2. The *Saccharomycetaceae* is the largest yeast lineage across which a pheromone candidate is known to be conserved and we used its time of divergence as an evolutionary horizon within which species are likely to show strong pheromone homology (Figure 3A). We identified all monophyletic clades on the yeast phylogram with ancestral nodes within this time horizon: 300 species give rise to 23 phylogroups, each containing at least 2 species, and 32 species are singleton orphans with no related species within this horizon (Figure 3A, B). Among these 23 phylogroups are recognized clades of yeasts, including *Debaryomycetaceae*/*Metschnikowiaceae*, *Pichiaceae*, *Saccharomycodaceae*, *Saccharomycetaceae*, *Phaffomycetaceae* and *Yarrowia*, which contain well-known species (Figure 3B). By pooling the candidates of species within these groups, we expect that the best candidate pheromones will share two properties: they are strongly conserved and conservation is strongest in the mature region, between the predicted N- and C-terminal processing signals.

Farnesylated (**a**-factor-like) pheromones are often encoded in multiple copies in a yeast genome, presumably to secrete sufficient pheromone from non-motile, haploid cells. The copy number of the pheromone genes seems to vary rapidly even among closely related species [18, 29]. To identify the most likely pheromones, we looked for candidates that have multiple copies across genomes of at least some of the species in the clade (at least 2 homologous copies in the same genome, or homologous copies in 2 species of the same phylogroup). Since only the mature region

of the pheromone is bioactive, this filter identifies gene copies based on the conservation of the amino-acid sequence between the predicted protease site (Asn) and the stop codon. Specifically, candidates that are greater than 85% identical in the mature region are considered copies of the same predicted pheromone in the genome. Our algorithm only considers ORFs that maintain the reading frame between the start and stop codons in a single exon and will not identify gene copies that contain introns or frameshifts.

Of the 78 206 candidate ORFs that have both farnesylation and proteolysis sites (N...CAAX candidates) across 300 species in the 23 phylogenetic clades, 11 065 candidates have at least one other homologous candidate in the phylogroup. There are between 10-300 candidate genes per species, with false positive ORFs that are conserved among closely related species present in a number of clades (Figure 3C, middle). Manual curation of the multicopy candidates by similarity of amino acid composition to known pheromones and presence across multiple species narrows down the candidates to a total of 705 loci across 218 species with an average of 3.2 ORFs that encode potential pheromones per genome, and a range of 1-19 candidates per genome (Figure 3). These 705 candidates mostly account for a “best” pheromone candidate often encoded in multiple copies per genome.

The well characterized farnesylated pheromones of *Saccharomyces* species [6, 18], *C. albicans* and *C. dubliensis* [27], and *K. pastoris* and *K. phaffi* [28] clades are identified among candidates with multiple copies by our algorithm (Supp. Figure 1 and Supp. Table 1). The *Saccharomyces* pheromone candidates are a large group with many homologous copies among the genomes of the parent clade of *Saccharomycetaceae*. The pheromones of *C. albicans* and *C. dubliensis* are homologous to each other and only encoded in one copy in each genome. The *Komagatella* pheromones are identified by the presence of the same pheromone in the sequenced sister species

K. pseudopastoris and *K. populi*. Some known pheromone genes in *Saccharomycetaceae* identified by synteny are not >85% identical and were not identified in our filter. To account for slightly more divergent pheromones, we filtered for candidates that are >70% identical (17 125 of 78 206 candidates), expanding the list to 792 candidates in 238 species (Figure 3C, right and Supp. Table 1). Even with this expanded list, some species in the large clades of *Debaryomycetaceae/Metschnikowiaceae* (21 of 91 species), *Pichiaceae* (11 of 52 species) and *Phaffomycetaceae* (1 of 30 species) are missing candidates that are homologous to the best candidates in the most closely related species (Supp Figure 3). Unfortunately, there are many explanations for this, including lineage-specific selection for faster pheromone evolution, incomplete or poorly assembled genomes, or the presence of introns in pheromone genes.

Pheromone candidates are weakly associated with conserved DNA sequence motifs

Fungal mating depends on the transcription of mating genes controlled by master regulatory transcription factors [16, 30]. In *S. cerevisiae*, haploid cells are competent to mate because a transcription factor (TF), Ste12, binds the pheromone-responsive element (PRE motif) upstream of mating genes to induce transcription. Although we were initially attracted to the idea of using such motifs to help identify candidate pheromone genes, the specific transcription factor involved in regulating mating gene expression can change across groups of related species [31]. Specifically, Sorrells and colleagues identified the existence of Ste12 TF motifs upstream of mating genes across a large part of the budding yeasts (*Saccharomycotina*), however these motifs only control MATa-specific gene expression in the *Saccharomyces* lineage.

We considered 8 pheromone-activated genes and 7 MATa-specific genes essential to mating in *S. cerevisiae* expressed in haploid yeast cells that are either induced by pheromone stimulation or specifically induced in a-like haploid cells and found that most of these genes have homologs in the 332 sequenced yeast species (Figure 4A). We then used a computational search (MEME) [32] on upstream regions (inferred promoter region) of these homologs to identify conserved TF binding sites involved in mating (Figure 4A). Very few significant motifs were identified by restricting the search to these 15 mating promoters across the full set of 332 sequenced species (Supp Figure 4).

Similar to pheromone homolog detection, we improved motif discovery by pooling mating genes across species within an evolutionary time horizon where the predicted regulatory motif is conserved upstream of the a-specific mating genes (Figure 3A) [31]. We assume that mating TF binding sites are conserved within species that are separated by a time horizon smaller than that corresponding to the divergence of *S. cerevisiae* and *V. polyspora* (Figure 3A) [31]. This yields 37 clades with at least 2 species, accounting for 290 genomes (Figure 3B). MEME identified 52, 44, and 37 significant regulatory motifs (e-value < 0.01) when all mating genes, only pheromone-activated genes, or only a-specific genes were pooled, respectively (Supp Figure 4). These three pooling strategies only find potential motifs in 23, 21, and 16 of the 37 phylogenetic groups where regulatory motifs are expected to be conserved, highlighting the difficulty in finding regulatory motifs using a limited set of co-regulated promoters. Twenty-five of the 37 phylogenetic groups (266 of the 290 genomes) have a predicted conserved mating TF-binding motif identified from at least one of the three sets of mating genes. Several significant motifs are variations of the binding site for the mating master regulator Ste12 of the Saccharomycetaceae clade of yeast

(‘TGAAACA’). In the analyses below, we considered all identified significant MEME motifs to ensure accounting of yeast clades with divergent **a**-specific gene regulation.

We counted the presence of these identified mating TF motifs upstream of pheromone candidate genes in the presumed promoter region using the FIMO algorithm [33] and counted significant hits (p -value < 0.01). Of the 75 504 unique candidate loci across 290 species, 57 573 (76%) candidates across 266 species have at least one identified motif in their promoters. Of the 792 manually curated pheromone candidates in 238 species, 684 (86%) candidates in 216 species have at least one copy of any of the identified mating motifs in their promoter. The remaining 108 pheromone candidates are identified to have homologous pheromone candidates in related strains but do not have any identifiable regulatory motif upstream.

In summary, regulatory motifs may be useful to validate candidates identified by other approaches. But motifs are found upstream of nearly as many of the pheromone candidates prior to curation (76%) as in the manually curated set (86%) and are thus not sufficient to separate the true pheromones from the filtered [N...CAAX] candidates. Although the presence of mating-related TF-binding motifs in candidate promoters cannot be used to predict true pheromones, there is a statistical enrichment of mating-related motifs (from any or all subsets of mating genes), upstream of manually curated candidates (Figure 4B).

Identifying the *Y. lipolytica* **a-factor**

The *Yarrowia* clade of budding yeasts is used for the heterologous expression of proteins using hydrocarbons as a metabolic carbon source [34, 35], and the production of secondary metabolites [36]. In basic research, *Y. lipolytica* is a valuable model system to study fatty acid metabolism in

peroxisomes [37] and mitochondrial complex I in an obligate aerobe [38, 39]. Other than the 5 *Yarrowia* genomes included in the 332 genomes, we also analyzed a sixth genome available from NCBI for pheromone candidates (*Y. sp.* 30695; Figure 5A). The best candidate pheromone is present in multiple copies in each of the genomes (Figure 5A and Table 2). Comparing the flanking regions of copies in all the genomes confirms that all identified loci are unique within a genome, with pairs in the closest species pairs (*Y. divulgata* and *Y. deformans*; *Y. keelungensis* and *Y. sp.* 30695) being syntenic (Supp Figure 5C). The presumed mature peptide sequence of the pheromone is completely conserved within the clade with the exception of a single conservative mutation (F to Y; Figure 5A). Although there are several DNA polymorphisms in the mature peptide's sequence, they are synonymous mutations (Figure 5A). In contrast, there is both non-synonymous variation and indels in the N-terminal region of the pheromone precursor in the *Yarrowia* lineage (Figure 5A). This is consistent with previous experiments screening for mutants in *S. cerevisiae* which showed that the sequence identity of the proteolyzed N-terminal sequence does not affect the production of mature **a**-factor [19].

We showed that the *Y. lipolytica* MATA mating-type is functionally homologous to MAT α of *S. cerevisiae*: *Ylste2Δ* (YALI0F03905g, ortholog of the α -factor receptor) and *Ylste6Δ* (YALI0E05973g, ortholog of Ste6) deletions are mating-deficient in MATA, and a *Ylste3Δ* (YALI0F11913g, ortholog of the **a**-factor receptor) deletion is mating-deficient in MATB (Supp. Figure 5A). We deleted the best pheromone gene candidates identified by the algorithm (*YMF*A1, *YMF*A2, *YMF*A3) in MATA as single-, double- and triple-deletions (Figure 5B). We saw a decrease in the mating efficiency in all the deletion strains, suggesting that the candidates play a role in mating. However, we were surprised to find that the triple mutant did not eliminate mating. Comparing the copy number of the pheromone candidate in different *Yarrowia* species revealed

that *Y. lipolytica* is an outlier with the fewest pheromone copies (3) compared to other *Yarrowia* species (7, 9, 9, 10 and 14) (Figure 5A and Table 2). Because the triple mutant still retained mating activity, we hypothesized that there is at least one additional, intron-containing copy of the pheromone in the *Y. lipolytica* genome.

We searched translated genome fragments (using tBLASTN) for loci that might code for another copy of the best candidate protein sequence and noticed a hit on Chromosome F. This locus encodes a copy of the pheromone synonymous to another locus (*YIMFA3*) but with an intron that disrupts the ORF, masking it from our algorithm (Supp. Figure 5B). The intron matches the precise expectations of a *Y. lipolytica* intron in length, 5' splice site (5' ss), branch point (BP), and 3' splice site (3' ss) implying this is a true extra copy of the pheromone [24]. The DNA sequence of the intron-containing gene is a close homolog of one of the intron-less genes, suggesting that an intron jumped into the gene after its duplication from a genomic copy (Supp Figure 5B).

A single gene copy of the pheromone can support mating in other yeasts that mate under nutrient starvation at low efficiency like *S. pombe* [40]. We therefore made the full quadruple-deletion (*YIMFA1*, *YIMFA2*, *YIMFA3*, *YIMFA4*) and confirmed that this manipulation abrogates mating as effectively as deleting the pheromone transporter or receptor (Figure 5B). Thus, our experiments in the model yeast *Y. lipolytica* confirm that our pheromone detection algorithm can identify the true farnesylated pheromone of a yeast clade based on the conservation of maturation motifs and multiple homologous copies in closely related genomes. This algorithm will enable annotation of genomes to include the short pheromone genes across the full yeast lineage.

Discussion

We designed an algorithm that inputs raw genome sequence and outputs candidates of farnesylated pheromones in fungi. This is of particular importance for fungi that are of medical (*C. albicans* and *C. neoformans*) [41] and agricultural (*U. maydis* (corn smut)[42], *P. striiformis* (wheat stripe rust) [43] and *M. oryzae* (rice blast) [44]) relevance, because pheromone signaling and mating have been correlated with the development of alternate cell morphologies that play a role in virulence. We experimentally validated the candidate pheromone of the *Yarrowia* clade by genetic deletions and mating assays in *Y. lipolytica*. Although the regulatory pathways of mating in fungi are broadly conserved [31], we found that the conservation of pheromone sequence in closely related species is a stronger predictor of pheromones than conserved regulatory sequences in the promoters of these genes. Our method is currently limited by the inability to identify candidates that contain introns, but can be expanded to include models of introns in fungi [24].

In yeasts that have been used as laboratory models (*S. cerevisiae* and *S. pombe*) or for the industrial production of proteins and secondary metabolites (*K. pastoris* and *Y. lipolytica*), pheromones can be used as tools to control the cell biology and transcriptional program as needed. In Ascomycota (specifically in yeast-like fungi), α -factor-like pheromones remain the more frequently studied pheromone signaling cascade because the genes that encode them are easily identified by the presence of multiple repeats encoded in a single prepro-peptide and the pheromone molecules are easy to produce and biochemically well-behaved. However, it is important to identify farnesylated pheromones because they represent the ancestral form of pheromones, with the Basidiomycota, such as *U. maydis* and *T. mesenterica*, only possessing farnesylated pheromones [2, 17, 45]. Our method relies not on standard homology search, but rather is an algorithmic filter that narrows down the list of all possible ORFs to a smaller list of candidates that can be tested by experimental validation. Specifically, our current version identifies most copies of a predicted pheromone gene

but may not identify all copies (due to the presence of unaccounted introns, for example). This tool provides a platform to add other filters that might be relevant to specific clades of fungi, based on known pheromones of closely related fungi.

Finally, the divergence of pheromones across the yeast lineage gives us a window into the evolution of short, unstructured peptides. The peptide sequence of pheromones are presumably constrained by needing to co-evolve with their pheromone transporters [10] and receptors [9, 11]. In addition to sequence divergence over long evolutionary timescales, pheromones also show copy number variation between the genomes of sister species indicating selection for gene duplications that increase the level of pheromone secretion and thus promote efficient mating [46]. Modern methods indicate that short peptide-coding genes are common in genomes across the tree of life, but their identification and functional characterization remain difficult [47, 48]. The pheromones of the fungal lineage offer a class of short coding genes with clear physiological roles to study sequence and copy number evolution across a lineage of a billion years.

Methods

Strains and plasmids

Y. lipolytica pair of mating strains (ML16507 and ML16510) were a gracious gift from Joshua Truehart (DSM ltd) and are derivatives of the sequenced CLIB122 strain [49]. Genomic transformation was done using standard protocols [50, 51]. Gene deletions were done by replacing the ORF with an auxotrophic marker for *YILEU2* or *YIURA3*. To re-use the URA3 marker, deletions were made with a *YIURA3* fragment flanked by repeat regions (obtained from plasmid

PMB5082, a gift from Joshua Truehart) that can excise the URA3 cassette upon selection on 5FOA. Geneious Prime 2021.2.2 was used to align and compare candidate sequences.

Table 3 <i>Y. lipolytica</i> strains used in this study.		
<i>Y. lipolytica</i> strain label	Genotype	Source
ML16507	<i>MATA; ura3-302; leu2-270; lys1-11; Ku70Δ</i>	Gift from Joshua Truehart (DSM ltd.)
ML16510	<i>MATB; ura2-6861; leu2-3; ade1; Ku70Δ</i>	Gift from Joshua Truehart (DSM ltd.)
yaliSS001	ML16507; <i>Ylste2Δ:LEU2Yl</i>	
yaliSS005	ML16507; <i>Ylste6Δ:LEU2Yl</i>	This study
yaliSS007	ML16507; <i>Ylmfa2Δ:URA3Yl</i>	This study
yaliSS009	ML16507; <i>Ylmfa1Δ:URA3Yl</i>	This study
yaliSS011	ML16507; <i>Ylmfa3Δ:URA3Yl</i>	This study
yaliSS017, yaliSS018	ML16507; <i>Ylmfa1Δ:LEU2Yl, Ylmfa2Δ:URA3Yl</i>	This study
yaliSS019, yaliSS020	ML16507; <i>Ylmfa2Δ:URA3Yl, Ylmfa3Δ:LEU2Yl</i>	This study
yaliSS021, yaliSS022	ML16507; <i>Ylmfa1Δ:LEU2Yl, Ylmfa3Δ:URA3Yl</i>	This study
yaliSS023, yaliSS024	ML16507; <i>Ylmfa2Δ:LEU2Yl, Ylmfa3Δ:URA3Yl</i>	This study
yaliSS027, yaliSS028	ML16507; <i>Ylmfa1Δ</i>	This study
yaliSS029, yaliSS030	ML16507; <i>Ylmfa1Δ, Ylmfa3Δ:LEU2Yl</i>	This study
yaliSS031, yaliSS032	ML16507; <i>Ylmfa1Δ, Ylmfa2Δ:URA3Yl, Ylmfa3Δ:LEU2Yl</i>	This study
yaliSS035, yaliSS036	ML16507; <i>Ylmfa1Δ, Ylmfa4Δ</i>	This study
yaliSS037	ML16507; <i>Ylmfa1Δ, Ylmfa2Δ:LEU2Yl, Ylmfa4Δ</i>	This study
yaliSS039, yaliSS040	ML16507; <i>Ylmfa1Δ, Ylmfa3Δ:LEU2Yl, Ylmfa4Δ</i>	This study
yaliSS041	ML16507; <i>Ylmfa1Δ, Ylmfa2Δ:LEU2Yl, Ylmfa3Δ:URA3Yl, Ylmfa4Δ</i>	This study
yaliSS045, yaliSS046	ML16507; <i>Ylmfa4Δ</i>	This study
yaliSS047	ML16507; <i>Ylmfa1Δ, Ylmfa2Δ:LEU2Yl, Ylmfa3Δ, Ylmfa4Δ</i>	This study

NOTE: In *Y. lipolytica*, MATA and MATB mating-types are equivalent to MATa and MATα of *S. cerevisiae* respectively.

Table 4. “Best” candidate pheromones identified in <i>Y. lipolytica</i> for experimental validation.		
Label	Expt. Label (for lab material)	Chromosomal position
<i>YIMFA1</i>	YlPher01	Yali0F_6 [forward][2134586-2134685-2134697]
<i>YIMFA2</i>	YlPher02	Yali0F_6 [forward][2139263-2139362-2139374]
<i>YIMFA3</i>	YlPher03	Yali0E_5 [reverse][1012402-1012303-1012291]
<i>YIMFA4</i>	YlPher05	Yali0F_5 [forward][1323342-1323493-1323505]

NOTE: Candidates located by [contigID][strand][serial# of candidate sharing Stop][“Start-Cys-Stop” position in contig].

***Y. lipolytica* semi-quantitative mating assay**

All strains listed in Table 3 were constructed from our wildtype MATA strain (ML16507) using a chemical transformation protocol reported in the literature [51]. We modified a quantitative mating

protocol used in [52] to test the mating efficiency of strains with deletions of pheromone candidates against our MATB partner (ML16510). Briefly, exponential cultures in yeast extract peptone dextrose medium (YPD) of the mating gene deleted MATA strain and partner MATB strain were harvested, and 2.5×10^6 cells of each partner were mixed in 150 μ L sterile water + 0.02% (w/v) bovine serum albumin. Mating mixtures were transferred onto filters using a filter assembly (with the cells spreading to about 5mm radius), and the filters (with cells) were moved onto YM mating media plates (3 g/L yeast extract, 5 g/L Bacto-peptone, 5 g/L malt extract and 20 g/L Bacto-agar) [52]. These plates were incubated at 28°C in the dark for three days (70-74 h). The density of the inoculum cell suspensions was measured using a Coulter counter for normalization. After 3 days filters with the mating mixtures were moved into 3 mL YP + 2% (v/v) glycerol + 0.05% (w/v) dextrose and incubated on a roller drum at 30°C for 3 h to recover cells from filters. The cultures were transferred to microfuge tubes and sonicated to disrupt clumps, before using a Coulter counter to check the cell density to normalize for plating. The pelleted cells were resuspended in water + 0.02% (w/v) bovine serum albumin and plated on diploid selective media (CSM-Lys-Ade). The cell suspensions were counted using the Coulter counter to normalize the plating density. The mating efficiency was calculated as the number of diploid cells for a sample normalized to the number of diploid cells from the control mating (ML16507 + ML16510) performed on the same day (code on Github at <https://github.com/sriramsrikant/pheromoneFinder>). The experiment was repeated with biological replicates and plating replicates to account for intrinsic noise of mating mixtures.

Algorithmic filter to identify pheromone candidates

As described in the main text, we brute-force identified all stop codons (TAA, TAG, TGA) in the six frames of translation of a genome that are immediately preceded by the farnesylation motif,

CAAX (Supp Figure 1B). We further only considered loci that have a start codon (ATG) within 100 codons upstream of the stop codon in the same frame of translation. Presence of multiple Start (Met) codons within the region of up to 300 bases upstream of a CAAX-stop motif are accounted as multiple pheromone candidates that share a common CAAX-stop terminus and at least one internal Asn/N cleavage site (Figure 2A; 125 711 ORFs encoded in 87 326 unique CAAX-stop loci). We used the chromosomes or contigs as a natural split given that their length in fungi allows for reasonable single-core performance of the sieve for stop codons in all frames of translation. To ensure our pipeline is greedy in finding candidates we excluded concerns of chromosomal position because some fungi have poorly assembled genomes. Due to the incomplete assembly of chromosomes in many yeasts, pheromone candidates that bridge contigs are necessarily lost and cannot be reasonably identified. We distributed the calculation by contig across processors on the Odyssey cluster (Research computing, Harvard FAS).

All scripts to identify candidate loci and subsequent search for homologous pheromone candidates and mating-regulation motifs are custom-written in python. Pheromone candidate homologs are determined by measuring the pairwise sequence identity of the candidate ORFs from the potential protease site (N) to the stop codon (N(X_M)CAAX) across species within the phylogroup where pheromones are conserved (Figure 3A). About a seventh of candidates (11 065 of 78 206 [N...CAAX] candidates in conserved-pheromone phylogroups that have at least 2 species – 300 of 332 species) have at least one other candidate that is >85% sequence identical and are considered to have a homologous copy. We further expanded the set of considered homologous pheromone candidates to about a quarter (17 125 of 78 206) by considering all [N...CAAX] candidates in conserved-pheromone phylogroups that have at least one copy >70% sequence identical. Manual curation was done by aligning these candidates grouped by phylogroup and looking for large

blocks of near-identical sequences distributed across species and often present in more than one copy within a genome. Motifs involved in mating regulation were identified by using MEME [32] to look for motifs upstream of known mating genes (FIG1, PRM1, KAR4, KAR5, SST2, FUS3, STE12, FUS1 as pheromone-activated genes and STE2, STE6, ASG7, AXL1, RAM2, BAR1, AGA2 as **a**-specific genes in *S. cerevisiae*) [31] in all 332 yeast genomes. We identified at most 3 significant mating-related motifs upstream of all mating genes, pheromone-activated genes, or **a**-specific genes by grouping species into phylogroups where mating regulation is expected to be conserved (Figure 3A) [31]. FIMO was then used to identify significant hits of motifs from each phylogroup upstream of the candidates of species within that phylogroup.

Analysis and plotting of statistics of filters are written in python using Jupyter notebooks. All scripts and Jupyter notebooks are available on GitHub (<https://github.com/sriramsrikant/pheromoneFinder>).

Acknowledgements

We thank the members of the Murray lab and Gaudet lab for helpful discussions. S.S. was a Howard Hughes Medical Institute International Student Research fellow. This work was funded in part by NIGMS grant R01GM120996 to R.G. and NIH grant R01GM43987 and the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (#1764269 [NSF] and #594596 [Simons]) to A.W.M.

Figures

Figure 1. *S. cerevisiae* pheromones are produced by cleaving and modifying precursor peptides. (A) Haploid *S. cerevisiae* have two mating types, **a** and α . Their mating with each other is initiated by the secretion of diffusible peptide pheromones that are recognized by a G protein-coupled receptors (GPCR): **a**-cells (magenta) secrete the lipidated peptide pheromone **a**-factor, which is recognized by the **a**-factor receptor expressed on α -cells, while α -cells (cyan) secrete the peptide pheromone α -factor, which is recognized by the α -factor receptor expressed by **a**-cells. (B) Mating pheromones (underlined) are encoded within small precursor peptides by the *MFA1* and *MFA2* genes (for **a**-factor) and *MFa1* and *MFa2* genes (for α -factor). These peptides require several maturation steps before their secretion as biologically active molecules. (C) The modifications of initial products of the *MFA1* and *MFA2* genes that produce **a**-factor. Broadly, there are two stages of maturation, C-terminal modifications (S-thiol farnesylation, -AAX proteolysis (white bars), and carboxymethylation), followed by two steps of N-terminal proteolysis (two grey bars). The mature pheromone (black bar) is then exported from the cytosol through a dedicated ABC transporter, Ste6.

Supplementary Figure 1. Yeast peptide pheromones share common maturation motifs. (A) Lipidated peptide pheromones (*Sca*-factor like) of yeast species (Saccharomycotina) verified by biochemistry or genetics. Yeasts have different numbers of genes in their genomes encoding identical or near-identical mature pheromones sequences. (B) The farnesylation [CAAX] motif is defined by cysteine that is S-farnesylated, followed by two aliphatic residues (AA) and a final variable residue (X). (Top) Known yeast pheromones provide a dictionary of possible AAX amino acid residues to identify unannotated pheromone CAAX motifs in the genome. (Bottom) To ensure that the filter is not biased by known pheromones we expanded the dictionaries used in our filter to include alternate residues that maintain farnesylation in *S. cerevisiae* [21-23]. Our final filter thus searches for all possible [CAAX-stop] motifs with A₁, A₂, X defined by the bottom row.

Figure 2. Fungal pheromone candidates can be identified by a progressive filter for small open-reading frames that are farnesylated and cleaved by a protease associated with mating. (A) Our algorithm begins by looking for all possible short open-reading frames with C-terminal farnesylation (CAAX-stop) in the 332 available fungal genomes, resulting in 284 073 candidates. We filter again for an in-frame proteolytic-motif asparagine (N) important for the final step of maturation to produce bio-active pheromone. This resulted in 125 711 candidates, encoded in 87 326 unique farnesylated loci. (B) Phylogenetic tree of all 332 sequenced yeasts across Saccharomycotina, covering clades like Debaryomycetaceae/Metschnikowiaceae, Pichiaceae, Saccharomycodaceae, Saccharomycetaceae, Phaffomycetaceae and Yarrowia that contain species important for both basic biology and industrial production. The tree is based on these yeast genome sequences with estimated relative divergence times. Selected clades are labeled. (C) On the left, a scatter plot shows strong correlation the number of annotated protein-coding genes with genome size, which ranges between 8 and 27 Mbp for the 332 yeast genomes. On the right, a scatter plot

shows that the number of unique pheromone candidate loci is linearly correlated with genome size, ranging from 100 to 650 candidates, as expected of a greedy search for all possible pheromone candidates.

Supplementary Figure 2. Successive filters reduce the number of candidate pheromone loci.

(A) (Left) Matrix heatmap showing the pairwise distances amongst 332 sequenced yeast genomes evaluated using a reltime time-calibrated species divergence tree [25]. The scale bar indicates pairwise distance between extant species in millions of years. (Right) Histogram of phylogenetic distances between pairs of species from the heatmap confirms a hierarchical connection between clades, with yeasts within the same clade more related to each other than species from different clades. (B) Scatter plots of the number of [CAAX] candidates (left) and [N...CAAX] (right) candidates plotted against the genome size, where each circle represents a genome, and the diameter of the circles represents the number of annotated protein-coding genes in the genome. Successive filters in our algorithm (the presence of a C-terminal [CAAX] and an upstream, in-frame proteolytic site [N...CAAX]) reduce the number of candidate pheromone loci.

Figure 3. Fungal pheromones are conserved within closely related species and are often encoded in multiple copies in a genome. (A) Phylogenetic tree describing the evolutionary relationship of 332 yeasts with two horizons indicated. The solid black line indicates the maximum evolutionary distance at which mature pheromone sequences show detectable sequence homology, corresponding to the divergence time of *S. cerevisiae* and *Kluyveromyces fragilis*. The dashed black line indicates the maximum evolutionary distance at which mating genes share conserved regulatory motifs, corresponding to the divergence time of *S. cerevisiae* and *Vanderwaltozyma polyspora*. (B) Based on the pheromone homology time horizon (top), we separated 332 yeast genomes into 23 phylogroups of at least 2 species and 32 singletons. The most populated phylogroups correspond to the listed well-known clades where closely related species have been densely sequenced. These clades are represented in the tree by the corresponding colors. Phylogroups expected to share mating regulation (bottom) are smaller with the larger clades divided between multiple phylogroups as indicated; 290 genomes are distributed across 37 mating regulation phylogroups of at least 2 species with 42 singletons. (C) Number of pheromone candidates per species plotted for each of the 23 conserved-pheromone phylogroups (different colored circle swarms), where each circle corresponds to the number of candidates in a species, and each copy of a group of closely homologous sequences within a genome is counted separately. Selecting for candidates that have at least two homologous copies within the clade reduces the number of viable candidates per genome to 10-300 (center compared to left). Manual curation of candidates similar to known pheromones identifies the most likely pheromone(s) in each genome (right from center) for experimental validation. There are 1-19 curated candidates encoded in each species for experimental testing. Some phylogroups contain too few species and no candidates rose above the rest through manual curation.

Supplementary Figure 3. Manual curation from candidates that have homologous copies within a phylogroup identifies good pheromone candidates in most species. (A) Histograms of pairwise sequence identity between all [N...CAAX] candidates (calculated between potential proteolysis site to stop) identified from genomes within a phylogroup where pheromones are expected to be conserved. Because these species should have homologous pheromones, their pheromones are likely found among candidate pairs who are more than 85% identical (red shaded region). Sample histograms from phylogroup 14 (Saccharomycetaceae) and phylogroup 22 (Yarrowia) are shown. (B) List of species where manual curation of best pheromone candidates did not find any pheromone candidates.

Figure 4. Best pheromone candidates are weakly associated with motifs that regulate other mating genes. (A) We considered 15 genes essential to mating in *S. cerevisiae* that are either induced by pheromone stimulation or specifically induced in **a**-like haploid cells (see labels at bottom). The figure represents the sequences of 332 yeast genomes (rows) and the detection of homologs of each of these 15 specific mating genes (columns), with a black bar showing the failure to detect a homolog of the mating gene in a particular genome. (B) Although true candidates are not clearly identified by the presence of mating regulatory motifs in their promoters, there is a noticeable enrichment of motifs identified using all mating genes, pheromone-activated genes, **a**-specific genes, or the sum of all identified motifs of every class (cumulative) upstream of manually curated candidates (orange) compared to all candidates that have a homologous copy in genomes of the conserved-pheromone phylogroup (blue). Each box plot represents the distribution of motif counts upstream of every candidate, with the midline representing the median, the box and whiskers representing quartiles, and the dots representing outliers. The enrichment of MEME identified motifs upstream of curated candidates are found to be significant using a two-sample Kolmogorov-Smirnov test.

Supplementary Figure 4. Motifs of mating regulators are found by MEME in mating genes within phylogroups. Histograms of e-values of MEME motifs identified in promoters of all mating genes (left), pheromone-activated genes (middle) or **a**-specific genes (right). (Top row) Motifs identified when analysis is done in individual species. This analysis does not identify many significant MEME motifs (e-value < 0.01). (Middle row) When mating promoters are pooled across species in conserved-mating-regulation phylogroups (290 species across 37 phylogroups), several significant motifs are identified. (Bottom row) Histograms of e-values of significant MEME motifs identified from analyzing all 290 genomes from 37 mating regulation conserved phylogroups collectively for all mating genes (52 motifs), pheromone-stimulated genes (44 motifs) and **a**-specific genes (37 motifs) in 266 genomes.

Figure 5. All species in Yarrowia clade of yeasts have a homologous farnesylated pheromone that is encoded in multiple copies per genome. (A) Phylogenetic tree of Yarrowia species (and outgroup in grey) from the set 332 of sequenced yeast of genomes [25], along with *Y. sp 30695*

which is a sister species of *Y. keelungensis*. The genome of *Y. sp. 30695* was analyzed independently to identify pheromone candidates. The translated ORFs of manually curated candidates from each species are aligned and ordered according to the phylogenetic relationship between the species. The red shaded region represents the candidate mature pheromone sequence, showing no non-synonymous variation across *Yarrowia*, except for a conservative change of phenylalanine (F) to tyrosine (Y) in *Y. lipolytica*. At the top are alignment logos of nucleotide and amino-acid sequences, which illustrate that non-synonymous variation is enriched in the proteolyzed N-terminal fragment, compared to the mature pheromone and conserved motifs. Synonymous variation is still present in the mature pheromone sequence (see wobble bases of nucleotide logo). **(B)** The mating efficiency of MATA haploid derivatives of *Y. lipolytica* with combinations of the four pheromone loci deleted was evaluated using a semi-quantitative mating protocol. Measurement for each genotype is represented as a group of at least 3 replicate measurements performed as both biological and technical replicates. Single- double- and triple-mutants of *YIMFA* genes show reduced mating, but only the quadruple deletion of all pheromone loci is deficient in mating to a comparable degree as the receptor (*Ylste2Δ*) and pheromone exporter (*Ylste6Δ*) deleted strains.

Supplementary Figure 5. *Y. lipolytica* mating-type MATA is functionally homologous to *S. cerevisiae* MATa and MATB is functionally homologous to MATa. (A) Plating of *Y. lipolytica* mating mixtures from semi-quantitative mating experiments on medium on which only diploids can grow. Deletions of homologs of MATa specific mating genes (α -factor receptor, Ste2 / YALI0F03905g and the α -factor exporter, Ste6 / YALI0E05973g) are deleterious to mating in the *Y. lipolytica* MATA haploid background. Correspondingly, deletion of a homolog of a MATa specific mating gene (α -factor receptor, Ste3 / YALI0F11913g) shows a mating defect in the *Y. lipolytica* MATB haploid. **(B)** *Y. lipolytica* genomic locus *YIMFA4* encodes the candidate pheromone and contains a predicted intron. (Top) Genomic locus in *Y. lipolytica* that contains a predicted intron identified by motifs (5' ss, BP, 3' ss) conserved across *Y. lipolytica* introns [24]. The corresponding cDNA and its translation (single-letter amino acid code) is aligned to the locus. (Bottom) The cDNA from intron-containing pheromone is aligned to the three intron-less genomic loci identified from the *Y. lipolytica* genome, showing a number of synonymous variants (darker letters) with only a single non-synonymous variant (gray and red boxes) in the N-terminal region (consistent with the pattern of variation across all *Yarrowia*; Figure 5A). **(C)** Aligning the 1 kbp upstream and downstream flanking regions of pheromones across the *Yarrowia* lineage (Figure 5A) is a measure of synteny of pheromone genes across *Yarrowia* genomes. Each row and column correspond to pheromone candidates in *Yarrowia* species with between 3 to 14 copies grouped by genome in the same order as in Figure 5A. Pairwise % identity from the alignment of the flanking regions of each locus displayed as a heat-map marking group of pheromones contained in each species. The low off-diagonal values within candidates of a species are consistent with no candidate within a species being double-counted. The high off-diagonal values all represent syntenic candidates in most-closely related pairs of species (*Y. divulgata* and *Y. deformans*; *Y. keelungensis* and *Y. sp. 30695*).

Table 1. Results of computational filter on 332 yeast genomes.

Species	#scaffolds	Genome Size (bp)	# of protein-coding genes	# of [CAAX] candidate s	# of [N...CAAX] candidates	# of unique [N...CAAX] loci	phylogenetic group	regulation phylogenetic group
yHAB160_kazachstania_kunashirensis_160519	353	10985687	5288	584	323	241	phyloGroup_14	phyloGroup_27
lachancea_fantastica	7	11336659	5002	769	313	247	phyloGroup_14	phyloGroup_30
yHMPu5000034615_saturnispor a_zaruensis_160519	391	9901215	5023	562	267	189	phyloGroup_10	phyloGroup_20
lipomyces_starkeyi	117	21267709	8889	1964	785	531	phyloGroup_53	phyloGroup_77
ashbya_aceri	8	8894523	4322	868	250	178	phyloGroup_14	phyloGroup_32
yHMPu5000034655_botryozym a_nematodophila_160519	28196	18790024	11043	455	239	186		
saccharomyces_paradoxus	16	11744610	5038	714	330	226	phyloGroup_14	phyloGroup_27
yHMPu5000035279_tortispora_starmeri_160928	1255	9585486	4978	524	203	157		
yHMPu5000034610_saturnispor a_hagleri_160519	336	10264027	5018	630	304	220	phyloGroup_10	phyloGroup_20
yHMPu5000034612_saturnispor a_saitoi_160519	312	9875405	4910	410	237	164	phyloGroup_10	phyloGroup_20
yHMPu5000034665_kodamaea_laetipori_160519	145	15043375	6050	913	431	299	phyloGroup_1	phyloGroup_2
yHMPu5000035011_candida_p yralidae_160519	412	12476905	5373	771	340	232	phyloGroup_1	phyloGroup_2
yHMPu5000035044_barnettozyma_californica_160519	276	11899360	5810	926	408	282	phyloGroup_17	phyloGroup_37
yHMPu5000034635_nadsonia_f ulvescens_160519	14721	17772825	7096	586	279	195	phyloGroup_20	phyloGroup_42
yHMPu5000035297_priceomyces_castillae_160928	597	10856769	5158	656	302	208	phyloGroup_1	phyloGroup_2
candida_sojae	511	11747307	5083	600	324	232	phyloGroup_1	phyloGroup_2

ambrosiozyma_monospora	118	24978438	7752	2056	974	646	phyloGroup_10	phyloGroup_20
yHMPu5000034886_ogataea_tr chalophila_160519	522	14167188	5807	802	415	278	phyloGroup_10	phyloGroup_20
yHMPu5000034654_aciculocon idium_aculeatum_160519	1065	18270442	7084	946	435	301	phyloGroup_1	phyloGroup_2
yHMPu5000034673_blastobotry s_nivea_160519	7527	15589204	9671	1097	458	312	phyloGroup_38	
yHMPu5000034637_ogataea_p opuliabae_160519	366	11802454	5489	843	367	257	phyloGroup_10	phyloGroup_20
yHMPu5000034669_blastobotry s_raffinofermentans_160519	61	12458252	6242	1000	324	220	phyloGroup_36	phyloGroup_59
kazachstania_naganishii	13	10845821	4925	837	366	240	phyloGroup_14	phyloGroup_27
yHMPu5000034667_blastobotry s_serpentis_160519	847	13890435	6716	923	466	310		
yHMPu5000041818_magnusio myces_tetrasperma_160928	1083	25980888	8711	1651	728	495		
metschnikowia_ipomoeae	549	18900759	6107	1547	626	427	phyloGroup_1	phyloGroup_3
yHMPu5000035022_candida_e mberorum_160519	308	13533920	5471	993	520	347	phyloGroup_1	phyloGroup_2
yHMPu5000035048_barnettozy ma_salicaria_160519	304	11010013	5586	666	311	230	phyloGroup_17	phyloGroup_37
yHMPu5000035268_wickerham omyces_hampshirensis_160928	1101	12058197	6241	576	296	224	phyloGroup_17	phyloGroup_38
yHMPu5000034881_torulaspora _pretoriensis_160519	157	11585715	5670	751	303	218	phyloGroup_14	phyloGroup_28
yHMPu5000035689_candida_m ycetangii_160613	146	12161848	5720	1099	438	291	phyloGroup_17	phyloGroup_38
yHMPu5000035243_zygoascus _meyerae_160928	242	12250757	5629	885	371	268		
yHMPu5000041862_candida_g olubevii_160928	882	14937814	5700	1065	448	310	phyloGroup_1	phyloGroup_3
candida_succiphila	22	12149879	5254	907	408	285	phyloGroup_10	phyloGroup_19
ascoidea_rubescens	63	17503998	6318	715	336	246	phyloGroup_12	

clavispora_lusitaniae	9	12114892	5149	1111	426	302	phyloGroup_ 1	phyloGroup_ _3
yHMPu5000034623_pichia_norvegensis_160519	856	10269774	5054	484	217	169	phyloGroup_ 10	phyloGroup_ _20
lachancea_kluyveri	8	11345814	5189	714	329	233	phyloGroup_ 14	
yHMPu5000035629_yueomyces_sinensis_160519	1143	13063945	5545	493	245	197	phyloGroup_ 14	phyloGroup_ _26
zygosaccharomyces_bailii	27	10268813	4606	730	343	232	phyloGroup_ 14	phyloGroup_ _28
yHMPu5000034613_saturnispora_serradocipensis_160519	645	10317717	5005	655	298	218	phyloGroup_ 10	phyloGroup_ _20
yHMPu5000034865_zygosaccharomyces_kombuchaensis_160519	656	10284548	4527	747	329	227	phyloGroup_ 14	phyloGroup_ _28
yHMPu5000034754_lipomyces_arxii_160519	1357	19329961	8351	1675	678	451	phyloGroup_ 53	phyloGroup_ _77
yHMPu5000034957_hanseniaspora_osmophila_160519	1899	12156859	5050	984	475	302	phyloGroup_ 13	phyloGroup_ _24
candida_apicola	40	9769876	4833	765	302	196	phyloGroup_ 30	phyloGroup_ _52
yHMPu5000034998_cephaloascus_albidus_160519	2177	16590907	7315	845	468	339	phyloGroup_ 2	phyloGroup_ _9
yHAB158_kazachstania_viticola_160519	1062	12238665	5828	582	302	209	phyloGroup_ 14	phyloGroup_ _27
candida_orthopsilosis	8	12659401	5212	943	489	345	phyloGroup_ 1	phyloGroup_ _2
yHMPu5000041824_debaryomyces_subglobosus_160928	291	11562645	5552	613	348	245	phyloGroup_ 1	phyloGroup_ _1
yHMPu5000034624_pichia_nakasei_160519	347	10430795	4358	560	291	199	phyloGroup_ 10	phyloGroup_ _20
naumovozyma_dairenensis	11	13527580	5651	657	338	244	phyloGroup_ 14	phyloGroup_ _27
lachancea_quebecensis	51	10225810	4672	738	305	213	phyloGroup_ 14	phyloGroup_ _30
hanseniaspora_valbyensis	647	11464179	5010	303	189	119	phyloGroup_ 13	phyloGroup_ _25
saccharomyces_cerevisiae	17	12157105	5917	692	330	248	phyloGroup_ 14	phyloGroup_ _27

candida_glabrata	13	12318245	5143	878	388	271	phyloGroup_14	phyloGroup_27
yHMPu5000035045_barnettozyma_hawaiiensis_160613	194	10617442	5371	816	342	226	phyloGroup_17	phyloGroup_37
metschnikowia_mauinuiana	270	17273017	6060	1348	580	378	phyloGroup_1	phyloGroup_3
spathaspora_passalidarum	8	13182099	5666	883	422	300	phyloGroup_1	phyloGroup_2
zygosaccharomyces_rouxii	7	9764635	4623	561	290	192	phyloGroup_14	phyloGroup_28
lachancea_waltii	713	10912112	5216	705	271	194	phyloGroup_14	phyloGroup_30
metschnikowia_dekortum	159	16312347	5408	1547	594	364	phyloGroup_1	phyloGroup_3
candida_auris	99	12498763	5207	887	370	258	phyloGroup_1	phyloGroup_3
scheffersomyces_stipitidis	9	15441079	5995	981	423	306	phyloGroup_1	phyloGroup_2
candida_dubliniensis	8	14618422	5949	746	403	281	phyloGroup_1	phyloGroup_2
scheffersomyces_lignosus	19	16592489	6075	1063	462	357	phyloGroup_1	phyloGroup_2
yHAB136_kazachstania_bromelicearum_160519	1136	11408292	5308	592	332	255	phyloGroup_14	phyloGroup_27
candida_versatilis	19	9313034	5476	704	289	195		
yHMPu5000034897_ogataea_glucozyma_160519	533	13866424	5858	886	352	243	phyloGroup_10	phyloGroup_20
yHMPu5000035041_ambrosiomyces_pseudovanderkluftii_160519	439	14456618	5927	1166	639	408	phyloGroup_10	phyloGroup_20
spathaspora_arborariae	41	12870582	5749	535	300	223	phyloGroup_1	phyloGroup_2
yHMPu5000034901_ogataea_methylovora_160519	181	12846970	6089	583	305	226	phyloGroup_10	phyloGroup_19
kazachstania_africana	12	11130140	5122	561	280	206	phyloGroup_14	phyloGroup_27
yHMPu5000026274_komagataella_populi_160519	171	9213105	4885	516	221	171	phyloGroup_6	phyloGroup_14

saccharomyces_arboricola	35	11619520	5013	736	315	233	phyloGroup_14	phyloGroup_27
yHMPu5000034643_candida_sc hataavii_160519	234	13749826	5660	1201	464	321	phyloGroup_1	phyloGroup_8
yHMPu5000034890_ogataea_ra menticola_160519	802	13049329	5890	465	201	164	phyloGroup_10	phyloGroup_19
candida_boidinii_JCM9604	32	19377211	6029	896	439	319	phyloGroup_10	
yHMPu5000035691_candida_sa lmanticensis_160519	216	12318838	5902	793	302	219		
yHMPu5000026145_ambrosioz yma_vanderkliftii_160519	4911	15665411	5988	929	489	348	phyloGroup_10	phyloGroup_20
yHAB164_kazachstania_aerobia _160519	533	12085512	5221	744	410	289	phyloGroup_14	phyloGroup_27
yHMPu5000041822_dipodascus _geniculatus_160928	303	15199412	6924	833	354	243	phyloGroup_24	phyloGroup_46
lodderomyces_elongisporus	28	15547272	5924	1117	508	355	phyloGroup_1	phyloGroup_2
nakaseomyces_nivariensis	29	11687903	5206	641	305	217	phyloGroup_14	phyloGroup_27
yHMPu5000035302_candida_fr agi_160928	453	12105387	5376	676	335	229	phyloGroup_1	phyloGroup_8
yHMPu5000034893_ogataea_p hilodendra_160519	197	8882715	4715	570	235	155	phyloGroup_10	phyloGroup_20
metschnikowia_hamakuensis	2097	19049992	6342	1485	613	405	phyloGroup_1	phyloGroup_3
yHMPu5000034627_pichia_hee dii_160519	1086	10776783	5208	413	232	164	phyloGroup_10	phyloGroup_20
yHAB143_kazachstania_siamen sis_160519	967	11904546	5180	532	323	236	phyloGroup_14	phyloGroup_27
yHMPu5000034675_blastobotry s_mokoenaii_160519	637	13798552	6191	1116	412	276		
hyphopichia_burtonii	27	12403110	5189	442	232	164	phyloGroup_1	phyloGroup_4
yHMPu5000041863_candida_g orgasii_160928	175	11363463	5577	485	252	191	phyloGroup_1	phyloGroup_5
yHMPu5000035677_kodamaea _ohmeri_160519	200	12373457	5362	930	383	261	phyloGroup_1	phyloGroup_2

yHMPu5000035696_hanseniasp ora_singularis_160519	1225	9236024	4985	270	139	106	phyloGroup_ 13	phyloGroup _25
metschnikowia_matae	2664	21760885	7279	1646	687	468	phyloGroup_ 1	phyloGroup _3
metschnikowia_hawaiiensis	1714	18614590	6220	1451	638	405	phyloGroup_ 1	phyloGroup _3
yHMPu5000035645_yarrowia_ divulgata_160519	488	21444462	6691	1925	602	426	phyloGroup_ 22	phyloGroup _44
nakaseomyces_delphensis	30	10969104	5105	606	262	199	phyloGroup_ 14	phyloGroup _27
yHMPu5000041840_candida_w ancherniae_160928	339	10375075	4825	709	219	155	phyloGroup_ 1	phyloGroup _3
yHMPu5000035643_yarrowia_ bubula_160519	811	20726424	6797	1680	578	396	phyloGroup_ 22	phyloGroup _44
lachancea_lanzarotensis	24	11092131	4871	830	330	237	phyloGroup_ 14	phyloGroup _30
saprochaete_clavata	339	17715251	6749	828	398	297		
tetrapisipora_phaaffii	17	12115070	5295	658	341	258	phyloGroup_ 14	phyloGroup _26
yHMPu5000035032_candida_g atunensis_160519	660	10928885	5262	717	347	244	phyloGroup_ 1	phyloGroup _2
brettanomyces_anomalous	30	12880761	5423	858	437	312	phyloGroup_ 10	phyloGroup _20
yHMPu5000034892_ogataea_pi lisensis_160519	1431	14405693	5782	860	395	303	phyloGroup_ 10	phyloGroup _20
yHMPu5000034625_pichia_kud riavzevii_160519	1082	11673292	5217	812	367	276	phyloGroup_ 10	phyloGroup _20
yHMPu5000041833_candida_ta mmaniensis_160928	358	10827769	5523	434	243	186	phyloGroup_ 1	phyloGroup _5
yHMPu5000034742_lipomyces _suomiensis_160519	376	14093042	6211	1200	492	320		
yHMPu5000034876_tetrapisip ora_iriomotensis_160519	576	12058124	5775	573	333	243	phyloGroup_ 14	phyloGroup _26
yHMPu5000035252_yamadazy ma_nakazawae_160928	1033	11163843	5331	620	325	235	phyloGroup_ 1	phyloGroup _5
metschnikowia_similis	1440	17474979	5654	1549	596	387	phyloGroup_ 1	phyloGroup _3
yHMPu5000034710_kluyverom yces_dobzhanskii_160519	235	10845893	4882	691	349	238	phyloGroup_ 14	phyloGroup _31

naumovozya_castellii	10	11219539	5264	675	329	241	phyloGroup_14	phyloGroup_27
candida_tropicalis	24	14630139	5975	902	513	341	phyloGroup_1	phyloGroup_2
yHMPu5000034973_danielozy ma_ontarioensis_160519	284	10726519	5149	781	323	198	phyloGroup_1	phyloGroup_4
yHMPu5000035707_candida_h eveicola_160519	85	13075526	5216	888	309	222	phyloGroup_1	phyloGroup_3
kluveromyces_marxianus	36	11165408	5110	766	373	264	phyloGroup_14	phyloGroup_31
metschnikowia_hibisci	86	11317242	5119	822	363	241	phyloGroup_1	phyloGroup_3
kuraishia_capsulata	7	11371246	5553	846	358	229	phyloGroup_9	phyloGroup_17
hanseniaspora_vinae	305	11383302	4661	906	441	298	phyloGroup_13	phyloGroup_24
ogataea_methanolica	32	15086366	5855	834	408	294	phyloGroup_10	phyloGroup_20
yHMPu5000034933_kuraishia_ molischiana_160519	305	10381940	5537	775	274	167	phyloGroup_9	phyloGroup_17
nakaseomyces_castellii	36	10286478	4689	754	359	258	phyloGroup_14	phyloGroup_27
yarrowia_keelungensis	41	21824367	6382	1605	528	361	phyloGroup_22	phyloGroup_44
yHMPu5000034660_diddensiell a_caesifluorescens_160519	355	12022407	6246	836	326	228		
yHMPu5000035654_tortispora_ ganteri_160519	1044	9626561	5266	624	254	173	phyloGroup_48	phyloGroup_72
lachancea_fermentati	8	10264457	4879	625	309	221	phyloGroup_14	phyloGroup_30
cyberlindnera_fabianii_JCM360 1	12	12303847	5752	1253	478	304	phyloGroup_17	phyloGroup_38
yHMPu5000041829_debaryomy ces_fabryi_160928	480	11965231	5618	570	320	235	phyloGroup_1	phyloGroup_1
yHMPu5000041678_debaryomy ces_prosopidis_160928	323	12928139	6201	712	378	268	phyloGroup_1	phyloGroup_1
yHMPu5000035301_pichia_terr icola_160928	390	12280907	5195	687	355	255	phyloGroup_10	phyloGroup_20
geotrichum_candidum	28	24836057	7139	1618	652	460		

yHMPu5000035690_candida_v artiovaarae_160613	237	11254085	5617	1293	554	318	phyloGroup_ 17	phyloGroup _38
eremothecium_cymbalariae	8	9669424	4471	643	312	222	phyloGroup_ 14	phyloGroup _32
yHMPu5000034978_cyberlindn era_mrakii_160519	628	12792944	6157	1010	446	301	phyloGroup_ 17	phyloGroup _38
yHAB133_kazachstania_unispo ra_160519	1935	12832539	5539	588	321	242	phyloGroup_ 14	phyloGroup _27
yHMPu5000035699_cyberlindn era_maclurae_160613	2336	11046364	5725	1165	409	261	phyloGroup_ 17	phyloGroup _38
yHMPu5000035674_kregervanr ija_delftensis_160613	2069	12710168	6042	452	254	168	phyloGroup_ 10	phyloGroup _20
arxula_adeninivorans	4	11804616	5939	965	339	218	phyloGroup_ 36	phyloGroup _59
yHMPu5000034758_lipomyces _japonicus_160519	690	13913255	6513	1027	484	322		
yHMPu5000035684_kloeckera_ hatyaiensis_160519	1209	10010803	4589	683	311	229	phyloGroup_ 13	phyloGroup _25
yHMPu5000035670_phaffomyc es_opuntiae_160613	986	13277281	5913	516	269	185	phyloGroup_ 17	phyloGroup _37
yHMPu5000034885_ogataea_zs oltii_160519	409	11719984	5466	730	319	203	phyloGroup_ 10	phyloGroup _20
metschnikowia_shivogae	93	10722219	4902	982	385	235	phyloGroup_ 1	phyloGroup _3
candida_homilientoma	8	12176763	5435	843	347	220	phyloGroup_ 1	phyloGroup _4
lachancea_mirantina	8	10117267	4852	647	260	196	phyloGroup_ 14	phyloGroup _30
yHMPu5000035040_ambrosioz yma_malecae_160519	180	12751984	5316	1032	408	281	phyloGroup_ 10	phyloGroup _20
yHMPu5000035031_candida_kr uisii_160519	232	12779385	5754	865	405	265	phyloGroup_ 1	phyloGroup _2
yHMPu5000034761_lipomyces _lipofer_160519	996	19856226	7989	1839	769	490		
yHMPu5000035658_starmera_a methionina_160613	1669	11902503	5836	535	287	217	phyloGroup_ 16	phyloGroup _35
yHMPu5000034681_blastobotry s_americana_160519	159	12598369	5788	809	373	261	phyloGroup_ 36	phyloGroup _58

yHMPu5000035318_hyphopichi a_heimii_160928	5312	13374215	5882	731	356	247	phyloGroup_ 1	phyloGroup _4
candida_tenuis	25	10747050	4945	699	295	194	phyloGroup_ 1	phyloGroup _5
metschnikowia_cerradonensis	2263	21032656	6801	1642	654	444	phyloGroup_ 1	phyloGroup _3
yHMPu5000034646_wickerham iella_cacticola_160519	1891	12938510	5846	1047	411	293		
metschnikowia_kamakouana	235	15649791	5586	1349	512	345	phyloGroup_ 1	phyloGroup _3
yHMPu5000041693_debaryomy ces_nepalensis_160928	210	11468613	5575	685	339	234	phyloGroup_ 1	phyloGroup _1
yHMPu5000034862_zygotorula spora_florentina_160519	485	11090184	5014	712	316	228	phyloGroup_ 14	phyloGroup _28
tetrapisispora_blatiae	10	14048593	5377	603	347	261	phyloGroup_ 14	phyloGroup _26
yHMPu5000035688_cyberlindn era_xylosilytica_160519	464	13952255	6510	1488	460	281	phyloGroup_ 17	phyloGroup _38
yHMPu5000034903_ogataea_na ganishii_160519	256	12020319	5340	905	353	251	phyloGroup_ 10	phyloGroup _19
yHMPu5000034674_blastobotry s_muscicola_160519	1987	14881377	6761	1136	419	284		
yHMPu5000035046_barnettozy ma_populi_160613	726	11533356	5594	988	387	259	phyloGroup_ 17	phyloGroup _37
yHMPu5000034979_cyberlindn era_misumaiensis_160519	361	10543864	5588	1135	430	265	phyloGroup_ 17	phyloGroup _38
yHMPu5000034661_dipodascus _albidus_160519	725	16263702	7267	809	391	274	phyloGroup_ 24	phyloGroup _46
yHMPu5000035641_yamadazy ma_scolyti_160613	301	11319754	5207	586	274	191	phyloGroup_ 1	phyloGroup _5
yHMPu5000035671_phaffomyc es_antillensis_160613	1461	12301330	5931	484	270	191	phyloGroup_ 17	phyloGroup _37
yHMPu5000041855_candida_as calaphidarum_160928	1113	10739842	4959	859	288	181	phyloGroup_ 1	
yHMPu5000034671_blastobotry s_peoriensis_160519	888	14179073	6289	995	338	241	phyloGroup_ 36	phyloGroup _58
candida_tanzawaensis	16	13136204	5321	866	385	264	phyloGroup_ 1	phyloGroup _2
alloascoidea_hylecoeti	136	24815695	9631	1011	574	445		

yarrowia_deformans	44	20877144	6509	1819	627	419	phyloGroup_22	phyloGroup_44
cyberlindnera_jadinii	76	13018415	5860	1230	552	362	phyloGroup_17	phyloGroup_38
metschnikowia_drakensbergensis	564	11968805	5078	1022	385	257	phyloGroup_1	phyloGroup_3
yHMPu5000034932_kuraishia_ogatae_160519	135	14739522	5573	1489	572	380		
yHMPu5000034611_saturnisporina_mendoncae_160519	669	9100453	4657	371	202	140	phyloGroup_10	phyloGroup_20
yHMPu5000035633_candida_hispaniensis_160519	344	10692747	5389	667	327	235		
yHAB134_kazachstania_turicensis_160519	972	14233920	5937	745	390	288	phyloGroup_14	phyloGroup_27
yHMPu5000026142_citeromyces_matritensis_160519	3451	20321446	7868	1581	660	462	phyloGroup_7	phyloGroup_15
candida_arabinofermentans	62	13234078	5674	647	338	236	phyloGroup_10	phyloGroup_20
yHMPu5000034967_candida_freyeyschussii_160519	187	10982976	5527	724	297	213	phyloGroup_17	phyloGroup_38
candida_carpophila	10	10242926	4920	765	307	200	phyloGroup_1	phyloGroup_7
komagataella_pastoris	4	9216378	4831	563	257	184	phyloGroup_6	phyloGroup_14
torulaspora_delbrueckii	8	9220678	4506	564	248	167	phyloGroup_14	phyloGroup_28
saccharomyces_uvarum	16	11502913	5004	767	316	226	phyloGroup_14	phyloGroup_27
yHMPu5000034884_peterozyma_toletana_160519	422	13669850	6276	815	407	284	phyloGroup_5	phyloGroup_13
ogataea_parapolyomorpha	7	8874589	4518	705	271	173	phyloGroup_10	phyloGroup_20
yHMPu5000034629_pichia_exiguua_160519	1055	11263787	5064	475	285	202	phyloGroup_10	phyloGroup_20
yHMPu5000035261_candida_pionderosae_160928	2820	14862131	7327	851	427	309		
yHMPu5000026152_torulaspora_franciscana_160519	307	10548653	5046	652	261	198	phyloGroup_14	phyloGroup_28

saccharomyces_eubayanus	29	11683510	5043	746	324	230	phyloGroup_14	phyloGroup_27
yHAB154_kazachstania_transvalensis_160519	2362	12771112	4897	680	366	262	phyloGroup_14	phyloGroup_27
yHMPu5000035296_priceomyces_carsonii_160928	215	11729577	5670	567	260	197	phyloGroup_1	phyloGroup_2
yHMPu5000035282_trigonopsis_vinaria_160928	551	11377523	5420	719	293	211		
yHMPu5000035659_saturnisporia_dispora_160519	645	10120266	4881	588	289	206	phyloGroup_10	phyloGroup_20
yHMPu5000026137_ambrosiozyma_ambrosiae_160519	183	14868748	6191	670	344	274	phyloGroup_10	phyloGroup_20
yHMPu5000034607_saccharomycopsis_capsularis_160519	1346	17823225	6736	1144	436	303	phyloGroup_11	phyloGroup_21
lachancea_dasiensis	8	10701617	4889	748	293	219	phyloGroup_14	phyloGroup_30
nadsonia_fulvescens_var_elongata	20	13746944	5861	592	294	206	phyloGroup_20	phyloGroup_42
lachancea_thermotolerans	8	10392861	4808	710	270	190	phyloGroup_14	phyloGroup_30
yHMPu5000035652_torulasporamaleae_160613	264	9267644	4312	614	222	166	phyloGroup_14	phyloGroup_28
yHMPu5000035639_wickerhamomyces_canadensis_160613	365	12510583	5950	804	427	295	phyloGroup_17	phyloGroup_37
yHMPu5000035037_candida_montana_160519	153	12551108	5819	691	352	254	phyloGroup_17	phyloGroup_37
sugiyamaella_lignohabitans	5	15982127	6447	1025	418	294		
yHMPu5000034648_candida_restringae_160519	250	13987553	5371	1046	433	296	phyloGroup_1	phyloGroup_2
yHMPu5000034899_ogataea_kodamae_160519	408	8701465	4551	611	260	178	phyloGroup_10	phyloGroup_20
yHMPu5000035335_candida_blatatae_160928	785	12290805	5526	1013	402	255	phyloGroup_1	phyloGroup_3
yHMPu5000026124_ogataea_henricii_160519	497	12604709	5567	786	293	205	phyloGroup_10	phyloGroup_20
sporopachydermia_quercuum	15	16411463	5968	911	460	333	phyloGroup_18	
yHMPu5000034952_citeromyces_siamensis_160519	8808	25257042	12786	1775	790	550	phyloGroup_7	phyloGroup_15

yHMPu5000035675_kregervanr ija_fluxuum_160613	2365	13058076	6539	516	289	191	phyloGroup_ 10	phyloGroup _20
yarrowia_lipolytica	6	20502978	6433	1953	738	473	phyloGroup_ 22	phyloGroup _44
ogataea_polymorpha	7	8974850	4289	693	245	165	phyloGroup_ 10	phyloGroup _20
eremothecium_coryli	19	9094934	4439	631	302	223	phyloGroup_ 14	phyloGroup _32
yHAB155_kazachstania_spence rorum_160519	2023	13322408	6540	655	361	265	phyloGroup_ 14	phyloGroup _27
ascoidea_asiatica	71	20313720	7632	439	205	161	phyloGroup_ 12	
yHMPu5000034974_deakozym a_indianensis_160519	363	11777613	5535	959	292	221	phyloGroup_ 28	phyloGroup _50
eremothecium_sinecaudum	8	8948761	4314	512	246	186	phyloGroup_ 14	phyloGroup _32
yHMPu5000034709_kluyverom yces_aestuarii_160519	184	10059533	4833	571	289	207	phyloGroup_ 14	phyloGroup _31
pachysolen_tannophilus	198	12597116	5800	447	232	165		
tortispora_caseinolytica	6	9180161	4587	663	247	177	phyloGroup_ 48	phyloGroup _72
yHMPu5000035681_candida_g otoi_160519	513	13358213	5752	789	383	254	phyloGroup_ 1	phyloGroup _4
metschnikowia_proteae	731	12525497	5184	1041	416	259	phyloGroup_ 1	phyloGroup _3
yHMPu5000034990_candida_c orydali_160519	160	11657245	5412	587	374	270	phyloGroup_ 1	phyloGroup _2
metschnikowia_arizonensis	1012	16315752	5892	1316	522	364	phyloGroup_ 1	phyloGroup _3
spathaspora_girioi	2380	14809507	6733	678	374	263	phyloGroup_ 1	phyloGroup _2
yHMPu5000034875_tetrapisisp ora_fleetii_160519	1792	12364389	5563	598	341	265	phyloGroup_ 14	phyloGroup _26
yHAB132_kazachstania_martini ae_160519	592	11811926	5708	470	242	181	phyloGroup_ 14	phyloGroup _27
yHMPu5000035665_middelhov enomyces_tepae_160613	317	15078393	6443	888	446	317		
yHMPu5000035686_cyberlindn era_saturnus_160519	650	13429652	6564	1129	498	334	phyloGroup_ 17	phyloGroup _38

yHMPu5000035242_zygoascus_ofunaensis_160928	676	13241094	5725	915	350	256		
yHMPu5000034883_peterozyma_xylosa_160519	234	12957929	5982	751	371	249	phyloGroup_5	phyloGroup_13
yHMPu5000035673_candida_orba_160613	690	11580103	5742	432	236	175	phyloGroup_17	phyloGroup_37
yHMPu5000035286_candida_azyma_160928	450	12206970	5949	703	384	245	phyloGroup_17	phyloGroup_37
candida_parapsilosis	8	12998393	5280	1020	503	359	phyloGroup_1	phyloGroup_2
lachancea_meyersii	8	11261819	4955	775	303	222	phyloGroup_14	phyloGroup_30
hanseniaspora_uvarum	335	9503661	4256	382	188	143	phyloGroup_13	phyloGroup_25
metschnikowia_kipukae	158	11256265	5084	943	398	243	phyloGroup_1	phyloGroup_3
vanderwaltozyma_polyspora	281	14674591	5679	606	307	245	phyloGroup_14	phyloGroup_26
yHMPu5000041743_candida_hawaiiiana_160928	335	11909429	5098	1003	371	243	phyloGroup_1	phyloGroup_3
yHMPu5000035703_cyberlindnera_americana_160613	417	11631002	5848	839	342	226	phyloGroup_17	phyloGroup_38
yHMPu5000035672_phaffomyces_thermotolerans_160613	630	11427259	5625	389	213	161	phyloGroup_17	phyloGroup_37
yHMPu5000034597_candida_stellimalicola_160519	276	10858432	5226	698	260	186	phyloGroup_16	phyloGroup_35
metschnikowia_santacerciliae	704	20387451	6489	1621	632	430	phyloGroup_1	phyloGroup_3
yHMPu5000034887_ogataea_trichaloabstinens_160519	527	11332404	5641	703	312	228	phyloGroup_10	phyloGroup_20
starmerella_bombicola_JCM9596	16	9561028	4706	739	276	203	phyloGroup_30	phyloGroup_52
yHMPu5000034614_saturnisporasilvae_160519	300	9300146	4587	640	266	185	phyloGroup_10	phyloGroup_20
yHMPu5000034986_candida_oregonensis_160519	551	10987101	5328	943	340	220	phyloGroup_1	phyloGroup_3
spathaspora_gorwiae	2530	14889443	6678	730	370	279	phyloGroup_1	phyloGroup_2

yHMPu5000034757_lipomyces _doorenjongii_160519	568	12255567	5910	1109	486	327		
nakaseomyces_bacillisporus	37	11106681	4844	678	319	227	phyloGroup_ 14	phyloGroup _27
metschnikowia_bicuspidata	48	16055203	5241	1376	519	334	phyloGroup_ 1	phyloGroup _3
yHMPu5000034946_ambrosioz yma_oregonensis_160519	1058	15395101	6680	508	295	216	phyloGroup_ 10	phyloGroup _20
yHMPu5000034988_candida_fr uctus_160519	190	11442680	5017	868	326	204	phyloGroup_ 1	phyloGroup _3
candida_intermedia	12	13027022	5533	1056	435	302	phyloGroup_ 1	phyloGroup _3
yHMPu5000035667_kurtzmanie lla_cleridarum_160928	418	12128693	5478	566	321	218	phyloGroup_ 1	phyloGroup _8
yHAB153_kazachstania_rosinii _160519	455	12267339	5378	1139	522	344	phyloGroup_ 14	phyloGroup _27
yHMPu5000034604_sporopach ydermia_lactativora_160519	1572	16710198	6204	1013	424	294	phyloGroup_ 18	
yHMPu5000034904_ogataea_n onfermentans_160519	57	10823021	5798	762	297	215	phyloGroup_ 10	phyloGroup _20
yHMPu5000034918_nakazawae a_holstii_160519	699	13538229	6206	850	495	334	phyloGroup_ 5	phyloGroup _12
metschnikowia_aberdeeniae	191	10686963	4807	898	334	209	phyloGroup_ 1	phyloGroup _3
yHMPu5000034877_tetrapisip ora_namnaonensis_160519	1578	12741674	5748	585	317	250	phyloGroup_ 14	phyloGroup _26
millerozyma_acaciae	10	11065882	4893	675	298	220	phyloGroup_ 1	phyloGroup _1
yHMPu5000034866_zygosacch aromyces_bisporus_160519	418	10605989	4717	737	336	243	phyloGroup_ 14	phyloGroup _28
lachancea_cidri	8	10147735	4740	668	291	208	phyloGroup_ 14	phyloGroup _30
nakaseomyces_bracarensis	40	12413499	5432	637	301	215	phyloGroup_ 14	phyloGroup _27
yHMPu5000034631_martiniozy ma_abiesophila_161215	11816	16284914	12589	333	220	167	phyloGroup_ 10	phyloGroup _20
yHMPu5000034963_hanseniasp ora_clermontiae_160519	794	9003421	4450	590	251	179	phyloGroup_ 13	phyloGroup _25

yHMPu5000041713_debaryomyces_maramus_160928	341	11126639	5559	608	339	229	phyloGroup_1	phyloGroup_1
eremothecium_gossypii	7	9095748	4327	759	280	173	phyloGroup_14	phyloGroup_32
spathaspora_hagerdaliae	1873	16614112	6953	830	440	314	phyloGroup_1	phyloGroup_2
metschnikowia_borealis	839	20333103	6574	1691	683	469	phyloGroup_1	phyloGroup_3
candida_sorboxylosa	37	10714862	4456	877	447	287	phyloGroup_10	phyloGroup_20
yHMPu5000026256_zygotorula_spora_mrakii_161215	568	10393571	4772	721	311	223	phyloGroup_14	phyloGroup_28
yHMPu5000034891_ogataea_pini_160519	118	11571877	5407	749	341	233	phyloGroup_10	phyloGroup_20
yHAB148_kazachstania_intestinalis_160519	2126	10108516	4961	659	274	204	phyloGroup_14	phyloGroup_27
nakazawaea_peltata	11	11653826	5458	593	294	212	phyloGroup_5	phyloGroup_12
yHAB159_kazachstania_solicola_160519	1524	13235758	6019	885	410	263	phyloGroup_14	phyloGroup_27
yHMPu5000034670_blastobotrys_proliferans_160519	479	14685634	7445	1267	525	361	phyloGroup_38	
wickerhamomyces_anomalus	46	14145566	6207	734	398	289	phyloGroup_17	phyloGroup_36
meyerozyma_guilliermondii	9	10609954	5010	834	363	242	phyloGroup_1	phyloGroup_7
priceomyces_haplophilus	9	10517128	4983	657	283	188	phyloGroup_1	phyloGroup_2
yHMPu5000035271_wickerhamomyces_bovis_160928	410	11715122	5850	880	420	286	phyloGroup_17	phyloGroup_37
yHMPu5000034636_ogataea_nitratoaversa_160519	910	14344126	5880	776	420	312	phyloGroup_10	phyloGroup_20
kluyveromyces_lactis	6	10689156	4962	678	374	259	phyloGroup_14	phyloGroup_31
pichia_membranifaciens	11	11584685	4969	756	317	213	phyloGroup_10	phyloGroup_20
yHMPu5000035244_candida_incommunis_160928	570	12142731	5696	929	377	262	phyloGroup_28	phyloGroup_50

yHMPu5000034760_lipomyces_kononenkoe_160519	754	17286122	7718	1455	647	419	phyloGroup_53	phyloGroup_77
yHMPu5000034632_candida_at_hensensis_160519	42	9665742	4611	776	314	204	phyloGroup_1	phyloGroup_7
metschnikowia_continentalis	802	21835504	6668	1705	741	504	phyloGroup_1	phyloGroup_3
yHMPu5000034939_komagatae_lla_pseudopastoris_160519	398	9184203	4604	547	258	191	phyloGroup_6	phyloGroup_14
yHMPu5000035640_yamadazy_ma_philogaea_160519	592	11284508	4854	750	369	243	phyloGroup_1	phyloGroup_5
wickerhamiella_domercqiae	4	8470812	4162	717	250	188		
metschnikowia_bowlesiae	2171	17565573	5826	1550	623	395	phyloGroup_1	phyloGroup_3
metschnikowia_lockheadii	2140	20094425	6718	1812	705	460	phyloGroup_1	phyloGroup_3
yHMPu5000034749_lipomyces_mesembrius_160519	547	17138379	7511	1564	630	441	phyloGroup_53	phyloGroup_77
yHMPu5000034748_lipomyces_oligophaga_160519	124	13644737	5873	1013	463	313		
lachancea_nothofagi	8	11313798	5086	827	357	247	phyloGroup_14	phyloGroup_30
yHMPu5000035033_candida_cr_etensis_160519	137	11230935	5169	778	326	225	phyloGroup_1	phyloGroup_2
yHMPu5000034950_citeromyces_hawaiiensis_160519	3163	19639324	7349	1678	644	460	phyloGroup_7	phyloGroup_15
yHAB147_kazachstania_taiensis_160519	483	13629483	5201	876	389	270	phyloGroup_14	phyloGroup_27
candida_albicans	9	14323546	6226	734	387	281	phyloGroup_1	phyloGroup_2
yHMPu5000034594_starmera_q_uercuum_160519	265	11746186	5736	771	372	262	phyloGroup_16	
yHMPu5000035274_wickerhamomyces_alni_160928	1389	12460614	6112	893	408	285	phyloGroup_17	phyloGroup_37
yHMPu5000026197_brettanomyces_custersianus_160519	219	10457826	5062	642	289	203	phyloGroup_10	phyloGroup_20
metschnikowia_matae_maris	2377	21578388	7336	1658	673	465	phyloGroup_1	phyloGroup_3
yHMPu5000035047_barnettozyma_pratensis_160519	469	12622213	5854	872	467	322	phyloGroup_17	phyloGroup_37

yHMPu5000035695_hanseniasp ora_pseudoguilliermondii_160519	917	9078015	4438	597	282	190	phyloGroup_ 13	phyloGroup_ _25
yHMPu5000034999_cephaloasc us_fragrans_160519	427	13760617	5865	735	376	288	phyloGroup_ 2	phyloGroup_ _9
ambrosiozyma_kashinagacola	23	12320962	5611	795	427	279	phyloGroup_ 10	phyloGroup_ _20
yHMPu5000035662_meyerozy ma_caribbica_160613	295	10509558	5027	768	295	200	phyloGroup_ 1	phyloGroup_ _7
saccharomycopsis_malanga	44	16719687	6187	1125	499	352	phyloGroup_ 11	phyloGroup_ _21
debaryomyces_hansenii	7	12152486	5806	661	346	262	phyloGroup_ 1	phyloGroup_ _1
candida_infanticola	22	8152740	4489	584	237	152		
yHMPu5000034902_ogataea_m inuta_160519	180	9133187	4457	651	261	182	phyloGroup_ 10	phyloGroup_ _20
yHMPu5000034947_ambrosioz yma_philentoma_160519	474	17710905	6975	852	408	306	phyloGroup_ 10	phyloGroup_ _20
saccharomyces_mikatae	16	11450123	4914	644	327	226	phyloGroup_ 14	phyloGroup_ _27
yHMPu5000035325_cyberlindn era_petersonii_160928	6324	12138127	5769	1060	386	254	phyloGroup_ 17	phyloGroup_ _38
yHMPu5000034606_priceomyc es_medius_160519	176	11083848	5280	580	296	206	phyloGroup_ 1	phyloGroup_ _2
yHMPu5000034605_spencerma rtinsiella_europaea_160519	363	14731014	6401	1091	391	269		
yHMPu5000035651_torulaspora _microellipsoides_160519	302	10979968	5298	716	356	258	phyloGroup_ 14	phyloGroup_ _28
dekkera_bruxellensis	84	13385232	5600	938	430	309	phyloGroup_ 10	phyloGroup_ _20
babjeviella_inositolovora	49	15221050	5752	1017	361	256		
yHAB166_kazachstania_yakush imaensis_160519	898	12964544	5511	582	298	223	phyloGroup_ 14	phyloGroup_ _27
wickerhamomyces_ciferrii	364	15901152	6645	606	359	258	phyloGroup_ 17	phyloGroup_ _36
yHMPu5000035650_trigonopsis _variabilis_160613	338	11545724	5469	733	284	203		
yHMPu5000034622_pichia_occ identalis_160519	880	12048800	5456	908	367	261	phyloGroup_ 10	phyloGroup_ _20

yHMPu5000035687_cyberlindn era_suaveolens_160613	886	12423684	6088	1023	449	299	phyloGroup_ 17	phyloGroup _38
yHMPu5000035679_candida_rh agii_160613	333	12462493	5587	780	370	256	phyloGroup_ 1	phyloGroup _4
yHMPu5000035018_candida_ca nberraensis_160519	260	14815857	5763	990	476	330	phyloGroup_ 1	phyloGroup _2
saccharomyces_kudriavzevii	16	11305202	4887	766	355	249	phyloGroup_ 14	phyloGroup _27
wickerhamia_fluorescens	20	13178500	5297	769	388	278	phyloGroup_ 1	phyloGroup _2

Table 2. Conserved pheromone candidates in *Yarrowia* clade

<i>Yarrowia lipolytica</i> CLIB122	yarrowia_lipolytica/[YALI0F][forward][1][2139263-2139362- 2139374][Translate:MSKAIPRAYGTDSYRVNSNVTIVGGYRTFQPSSCVIA*] yarrowia_lipolytica/[YALI0F][forward][1][2134586-2134685- 2134697][Translate:MSKAIPRAYGTDSYRVNSNVTIVGGYRTFQPSSCVIA*] yarrowia_lipolytica/[YALI0E][reverse][1][1012402-1012303- 1012291][Translate:MSKAIPRDYGTDSYRVNSNVTIVGGYRTFQPSSCVIA*] gi 49650673 emb CR382132.1 :1323506-133234[intron-containing][Translate: MSKAIPRDYGTDSYRVNSNVTIVGGYRTFQPSSCVIA*]
<i>Yarrowia deformans</i> J CM_1694_BCIW_NC BI	yarrowia_deformans/[gi 1002317476 dbj BCIW01000013.1][reverse][1][569522- 569426- 569414][Translate:MSVVKREYGNESYRVNSNVTIVGGYRTYQPSSCVIA*] yarrowia_deformans/[gi 1002317485 dbj BCIW01000004.1][forward][3][359951- 360209- 360221][Translate:MIMETTETPALRASIRYQKPPLQTHINISNLLQFPIKQPTPTT TTTTTTTTTIMSVVKREYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*] yarrowia_deformans/[gi 1002317472 dbj BCIW01000017.1][forward][2][66377- 66635- 66647][Translate:MTYNFIGSSYLEHYQVASNTSPSKYKYL PSTILELQQTTKTTT TPITTTTTTTMSVVKREYGNDSYRVNSNVTIVGGYRTYQPSSCVIA*] yarrowia_deformans/[gi 1002317485 dbj BCIW01000004.1][reverse][1][144986- 144890- 144878][Translate:MSVVKREYGNDSYRVNSNVTIVGGYRTYQPSSCVIA*] yarrowia_deformans/[gi 1002317485 dbj BCIW01000004.1][forward][1][511956- 512052- 512064][Translate:MSVVKREYGNESYRVNSNVTIVGGYRTYQPSSCVIA*] yarrowia_deformans/[gi 1002317479 dbj BCIW01000010.1][forward][1][912729- 912825- 912837][Translate:MSVVKREYGNESYRVNSNVTIVGGYRTYQPSSCVIA*] yarrowia_deformans/[gi 1002317481 dbj BCIW01000008.1][reverse][1][69294- 69198-69186][Translate:MSVVKREYGNDSYRVNSNVTIVGGYRTYQPSSCVIA*]
<i>Yarrowia divulgata</i> Y ADI0_NCBI	yHMPu5000035645_yarrowia_divulgata_160519/flattened_line_47/[reverse][1][116 579-116483- 116471][Translate:MSVVKREYGNESYRVNSNVTIVGGYRTYQPSSCVIA*] yHMPu5000035645_yarrowia_divulgata_160519/flattened_line_167/[forward][1][39 26-4022-4034][Translate:MSVVKREYGNESYRVNSNVTIVGGYRTYQPSSCVIA*]

	<p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_147 [reverse][1][29621-29525-29513][Translate:MSVVKREYGNESYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_346 [forward][1][7340-7436-7448][Translate:MSVVKREYGNESYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_141 [forward][1][50636-50732-50744][Translate:MSVVKREYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_19 [forward][1][220678-220774-220786][Translate:MSVVKREYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_79 [forward][1][12173-12269-12281][Translate:MSVVKREYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_368 [reverse][1][2146-2050-2038][Translate:MSVVKREYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_159 [forward][1][57250-57346-57358][Translate:MSVVKREYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035645_yarrowia_divulgata_160519/ flattened_line_195 [forward][1][32625-32721-32733][Translate:MSVVKREYGNQSYRVNSNVTIVGGYRTYQPSSCVIS*]</p>
<i>Yarrowia keelungensis</i> _JCM_14894_BCJD_N CBI	<p>yarrowia_keelungensis/ gi 1002317714 dbj BCJD01000001.1 [reverse][2][540603-540393-540381][Translate:MDSRQYNNFYSPQYKYHLNPPLFLHQTSTTTTTTTMGSAPIRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarrowia_keelungensis/ gi 1002317714 dbj BCJD01000001.1 [reverse][1][242794-242695-242683][Translate:MGSAPIRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarrowia_keelungensis/ gi 1002317709 dbj BCJD01000006.1 [reverse][1][423585-423486-423474][Translate:MGSAPIRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarrowia_keelungensis/ gi 1002317714 dbj BCJD01000001.1 [reverse][1][402036-401937-401925][Translate:MGSAPIRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarrowia_keelungensis/ gi 1002317714 dbj BCJD01000001.1 [forward][1][264198-264297-264309][Translate:MGSAPIRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarrowia_keelungensis/ gi 1002317712 dbj BCJD01000003.1 [forward][1][907356-907455-907467][Translate:MGEAIPRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarrowia_keelungensis/ gi 1002317711 dbj BCJD01000004.1 [forward][1][1838215-1838314-1838326][Translate:MGEAIPRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarrowia_keelungensis/ gi 1002317713 dbj BCJD01000002.1 [reverse][1][338601-338502-338490][Translate:MGSAPIRRYGTASYRVNSNVTIVGGYRTYQPSSCVIS*]</p> <p>yarrowia_keelungensis/ gi 1002317714 dbj BCJD01000001.1 [reverse][1][1692828-1692729-1692717][Translate:MGSAPIRRYGTASYRVNSNVTIVGGYRTYQPSSCVIA*]</p>
<i>Yarrowia sp.</i> _JCM_30 695_BCLX_NCBI	<p>yarsp2 BCLX01000001.1 [forward][1][2795617-2795716-2795728][Translate:MGSAPIRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarsp2 BCLX01000001.1 [forward][1][3085794-3085893-3085905][Translate:MGSAPIRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p>

	<p>yarsp2 BCLX01000007.1 [forward][1][1018344-1018443-1018455][Translate:MGSAIPRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarsp2 BCLX01000001.1 [forward][1][2934643-2934742-2934754][Translate:MGSAIPRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarsp2 BCLX01000001.1 [reverse][1][3073012-3072913-3072901][Translate:MGSAIPRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarsp2 BCLX01000005.1 [reverse][1][1822017-1821918-1821906][Translate:MGEAIPRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarsp2 BCLX01000002.1 [forward][1][3456539-3456638-3456650][Translate:MGEAIPRAYGTDSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarsp2 BCLX01000001.1 [forward][1][1641875-1641974-1641986][Translate:MGSAIPRRYGTASYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yarsp2 BCLX01000002.1 [reverse][1][771285-771186-771174][Translate:MGSAIPRRYGTASYRVNSNVTIVGGYRTYQPSSCVIS*]</p>
<i>Yarrowia_bubula_YAB U0_NCBI</i>	<p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_177 [forward][1][13360-13465-13477][Translate:MSGSKAIPRDYGNESYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_4 [forward][2][53746-53908-53920][Translate:MSTPVFIFHQTTTTTTTTTMSGSKAIPRDYGNESYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_73 [forward][1][126719-126824-126836][Translate:MSGSKAIPRDYGNESYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_135 [forward][1][11955-12060-12072][Translate:MSGSKAIPRDYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_57 [reverse][2][58408-58183-58171][Translate:MYP SLKCSFSAKSILHCFILQYINPSSPPQITTKHHTLSLMSGSKAIPRDYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_247 [forward][1][9255-9360-9372][Translate:MSGSKAIPRDYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_4 [reverse][1][407758-407653-407641][Translate:MSGSKAIPRDYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_169 [forward][1][18491-18596-18608][Translate:MSGSKAIPRDYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_29 [reverse][1][80548-80443-80431][Translate:MAGSKAIPRDYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_0 [reverse][1][542840-542735-542723][Translate:MSGSKAIPRNYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_95 [forward][1][139800-139905-139917][Translate:MSGSKAIPRNYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_227 [forward][1][37480-37585-37597][Translate:MSGSKAIPRNYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p> <p>yHMPu5000035643_yarrowia_bubula_160519/ flattened_line_129 [forward][1][79430-79535-79547][Translate:MSGSKAIPRNYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]</p>

yHMPu5000035643_yarrowia_bubula_160519//flattened_line_55 [forward][1][78869-78974-78986][Translate:MSGSRAIPRNYGNQSYRVNSNVTIVGGYRTYQPSSCVIA*]
--

NOTE: Candidates are identified by [contigID][strand][serial# of candidate sharing Stop][“Start-Cys-Stop” position in contig][Translation of locus]

References

1. Lücking, R., Huhndorf, S., Pfister, D.H., Plata, E.R., and Lumbsch, H.T. (2017). Fungi evolved right on track. *Mycologia* 101, 810-822.
2. Jones, S.K., Jr., and Bennett, R.J. (2011). Fungal mating pheromones: choreographing the dating game. *Fungal Genet Biol* 48, 668-676.
3. Kurjan, J., and Herskowitz, I. (1982). Structure of a yeast pheromone gene (MF alpha): a putative alpha-factor precursor contains four tandem copies of mature alpha-factor. *Cell* 30, 933-943.
4. Singh, A., Chen, E.Y., Lugovoy, J.M., Chang, C.N., Hitzeman, R.A., and Seeburg, P.H. (1983). *Saccharomyces cerevisiae* contains two discrete genes coding for the alpha-factor pheromone. *Nucleic Acids Res* 11, 4049-4063.
5. Chen, P., Sapperstein, S.K., Choi, J.D., and Michaelis, S. (1997). Biogenesis of the *Saccharomyces cerevisiae* Mating Pheromone a-Factor. *The Journal of Cell Biology* 136, 251-269.
6. Michaelis, S., and Herskowitz, I. (1988). The a-factor pheromone of *Saccharomyces cerevisiae* is essential for mating. *Molecular and Cellular Biology* 8, 1309-1318.
7. Michaelis, S. (1993). STE6, the yeast a-factor transporter. *Seminars in Cell Biology* 4, 17-27.
8. Caldwell, G.A., Naider, F., and Becker, J.M. (1995). Fungal lipopeptide mating pheromones: a model system for the study of protein prenylation. *Microbiol Rev* 59, 406-422.
9. Seike, T., Nakamura, T., and Shimoda, C. (2015). Molecular coevolution of a sex pheromone and its receptor triggers reproductive isolation in *Schizosaccharomyces pombe*. *Proceedings of the National Academy of Sciences of the United States of America* 112, 4405-4410.
10. Srikant, S., Gaudet, R., and Murray, A.W. (2020). Selecting for Altered Substrate Specificity Reveals the Evolutionary Flexibility of ATP-Binding Cassette Transporters. *Curr Biol* 30, 1689-1702 e1686.
11. Martin, S.H., Wingfield, B.D., Wingfield, M.J., and Steenkamp, E.T. (2011). Causes and consequences of variability in peptide mating pheromones of ascomycete fungi. *Mol Biol Evol* 28, 1987-2003.
12. Kamiya, Y., Sakurai, A., Tamura, S., Takahashi, N., Abe, K., Tsuchiya, E., and Fukui, S. (1978). Isolation of RhodotorucineA, a Peptidyl Factor Inducing the Mating Tube Formation in *Rhodospiridium toruloides*. *Agricultural and Biological Chemistry* 42, 1239-1243.

13. Kamiya, Y., Sakurai, A., Tamura, S., Takahashi, N., Abe, K., Tsuchiya, E., Fukui, S., Kitada, C., and Fujino, M. (1978). Structure of rhodotorucine A, a novel lipopeptide, inducing mating tube formation in *Rhodospiridiumtoruloides*. *Biochemical and biophysical research communications* 83, 1077-1083.
14. Sakagami, Y., Isogai, A., Suzuki, A., Tamura, S., Kitada, C., and Fujino, M. (1979). Structure of Tremrogen A-10, a Peptidal Hormone Inducing Conjugation Tube Formation in *Tremella mesenterica*. *Agricultural and Biological Chemistry* 43, 2643-2645.
15. Sakagami, Y., Isogai, A., Suzuki, A., Tamura, S., Tsuchiya, E., and Fukui, S. (1978). Isolation of a Novel Sex Hormone, Tremrogen A-10, Controlling Conjugation Tube Formation in *Tremella mesenterica* Fries. *Agricultural and Biological Chemistry* 42, 1093-1094.
16. Bennett, R.J., and Turgeon, B.G. (2016). Fungal Sex: The Ascomycota. *Microbiol Spectr* 4.
17. Coelho, M.A., Bakkeren, G., Sun, S., Hood, M.E., and Giraud, T. (2017). Fungal Sex: The Basidiomycota. *Microbiol Spectr* 5.
18. OhEigeartaigh, S.S., Armisen, D., Byrne, K.P., and Wolfe, K.H. (2011). Systematic discovery of unannotated genes in 11 yeast species using a database of orthologous genomic segments. *Bmc Genomics* 12.
19. Hoyer, G., Kistler, A., Nouvet, F.J., George, C.M., Boyle, M.L., and Michaelis, S. (2006). *Saccharomyces cerevisiae* a-factor mutants reveal residues critical for processing, activity, and export. *Eukaryot Cell* 5, 1560-1570.
20. Marr, R.S., Blair, L.C., and Thorner, J. (1990). *Saccharomyces-Cerevisiae*-Ste14 Gene Is Required for CooH-Terminal Methylation of a-Factor Mating Pheromone. *Journal of Biological Chemistry* 265, 20057-20060.
21. Berger, B.M., Kim, J.H., Hildebrandt, E.R., Davis, I.C., Morgan, M.C., Hougland, J.L., and Schmidt, W.K. (2018). Protein Isoprenylation in Yeast Targets COOH-Terminal Sequences Not Adhering to the CaaX Consensus. *Genetics* 210, 1301-1316.
22. Stein, V., Kubala, M.H., Steen, J., Grimmond, S.M., and Alexandrov, K. (2015). Towards the systematic mapping and engineering of the protein prenylation machinery in *Saccharomyces cerevisiae*. *PLoS One* 10, e0120716.
23. Trueblood, C.E., Boyartchuk, V.L., Picologlou, E.A., Rozema, D., Poulter, C.D., and Rine, J. (2000). The CaaX Proteases, Afc1p and Rce1p, Have Overlapping but Distinct Substrate Specificities. *Molecular and Cellular Biology* 20, 4381-4392.
24. Neugeglise, C., Marck, C., and Gaillardin, C. (2011). The intronome of budding yeasts. *C R Biol* 334, 662-670.

25. Shen, X.X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* 175, 1533-1545 e1520.
26. Adames, N., Blundell, K., Ashby, M.N., and Boone, C. (1995). Role of Yeast Insulin-Degrading Enzyme Homologs in Pheromone Processing and Bud Site Selection. *Science* (New York, N.Y.) 270, 464-467.
27. Dignard, D., El-Naggar, A.L., Logue, M.E., Butler, G., and Whiteway, M. (2007). Identification and characterization of MFA1, the gene encoding *Candida albicans* a-factor pheromone. *Eukaryot Cell* 6, 487-494.
28. Heisteringer, L., Moser, J., Tatro, N.E., Valli, M., Gasser, B., and Mattanovich, D. (2018). Identification and characterization of the *Komagataella phaffii* mating pheromone genes. *FEMS Yeast Res* 18.
29. Seike, T., Shimoda, C., and Niki, H. (2019). Asymmetric diversification of mating pheromones in fission yeast. *PLoS Biol* 17, e3000101.
30. Wong Sak Hoi, J., and Dumas, B. (2010). Ste12 and Ste12-like proteins, fungal transcription factors regulating development and pathogenicity. *Eukaryot Cell* 9, 480-485.
31. Sorrells, T.R., Booth, L.N., Tuch, B.B., and Johnson, A.D. (2015). Intersecting transcription networks constrain gene regulatory evolution. *Nature* 523, 361-365.
32. Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2, 28-36.
33. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017-1018.
34. Nicaud, J.M. (2012). *Yarrowia lipolytica*. *Yeast* 29, 409-418.
35. Liu, H.H., Ji, X.J., and Huang, H. (2015). Biotechnological applications of *Yarrowia lipolytica*: Past, present and future. *Biotechnol Adv* 33, 1522-1546.
36. Markham, K.A., and Alper, H.S. (2018). Synthetic Biology Expands the Industrial Potential of *Yarrowia lipolytica*. *Trends Biotechnol* 36, 1085-1095.
37. Dulermo, R., Gamboa-Melendez, H., Ledesma-Amaro, R., Thevenieau, F., and Nicaud, J.M. (2015). Unraveling fatty acid transport and activation mechanisms in *Yarrowia lipolytica*. *Biochimica et biophysica acta* 1851, 1202-1217.
38. Kerscher, S., Dröse, S., Zwicker, K., Zickermann, V., and Brandt, U. (2002). *Yarrowia lipolytica*, a yeast genetic system to study mitochondrial complex I. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1555, 83-91.

39. Parey, K., Brandt, U., Xie, H., Mills, D.J., Siegmund, K., Vonck, J., Kuhlbrandt, W., and Zickermann, V. (2018). Cryo-EM structure of respiratory complex I at work. *eLife* 7.
40. Kjaerulff, S., Davey, J., and Nielsen, O. (1994). Analysis of the structural genes encoding M-factor in the fission yeast *Schizosaccharomyces pombe*: identification of a third gene, *mfm3*. *Mol Cell Biol* 14, 3895-3905.
41. Boral, H., Metin, B., Dogen, A., Seyedmousavi, S., and Ilkit, M. (2018). Overview of selected virulence attributes in *Aspergillus fumigatus*, *Candida albicans*, *Cryptococcus neoformans*, *Trichophyton rubrum*, and *Exophiala dermatitidis*. *Fungal Genet Biol* 111, 92-107.
42. Chacko, N., and Gold, S. (2012). Deletion of the *Ustilago maydis* ortholog of the *Aspergillus* sporulation regulator *medA* affects mating and virulence through pheromone response. *Fungal Genet Biol* 49, 426-432.
43. Zhu, X., Liu, W., Chu, X., Sun, Q., Tan, C., Yang, Q., Jiao, M., Guo, J., and Kang, Z. (2018). The transcription factor *PstSTE12* is required for virulence of *Puccinia striiformis* f. sp. *tritici*. *Mol Plant Pathol* 19, 961-974.
44. Li, Y., Que, Y., Liu, Y., Yue, X., Meng, X., Zhang, Z., and Wang, Z. (2015). The putative Ggamma subunit gene *MGG1* is required for conidiation, appressorium formation, mating and pathogenicity in *Magnaporthe oryzae*. *Curr Genet* 61, 641-651.
45. Raudaskoski, M., and Kothe, E. (2010). Basidiomycete mating type genes and pheromone signaling. *Eukaryot Cell* 9, 847-859.
46. Moore, T.I., Chou, C.S., Nie, Q., Jeon, N.L., and Yi, T.M. (2008). Robust spatial sensing of mating pheromone gradients by yeast cells. *PLoS One* 3, e3865.
47. Couso, J.P., and Patraquim, P. (2017). Classification and function of small open reading frames. *Nat Rev Mol Cell Biol* 18, 575-589.
48. Andrews, S.J., and Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nat Rev Genet* 15, 193-204.
49. Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., et al. (2004). Genome evolution in yeasts. *Nature* 430, 35-44.
50. Burke, D., Dawson, D., and Stearns, T. (2000). *Methods in Yeast genetics: A Cold Spring Harbor Laboratory course manual*, (Cold Spring Harbor Laboratory Press).
51. Verbeke, J., Beopoulos, A., and Nicaud, J.M. (2013). Efficient homologous recombination with short length flanking fragments in *Ku70* deficient *Yarrowia lipolytica* strains. *Biotechnol Lett* 35, 571-576.

52. Rosas-Quijano, R., Gaillardin, C., and Ruiz-Herrera, J. (2008). Functional analysis of the MATB mating-type idiomorph of the dimorphic fungus *Yarrowia lipolytica*. *Curr Microbiol* 57, 115-120.

Figure 1. *S. cerevisiae* pheromones are produced by cleaving and modifying precursor peptides.

Supp Figure 1.

A	Species	Sequence of mature pheromone	Copies in genome	
	<i>Saccharomyces cerevisiae</i>	YIIKGVFWD PAC(S-farnesyl) YIIKGLFWD PAC(S-farnesyl)	2	
	<i>Vanderwaltozyma polyspora</i>	YIVKGWFWD PEC(S-farnesyl)	3	
	<i>Kluyveromyces lactis</i>	WIIPGFVWP QC(S-farnesyl)	1	
	<i>Komagatella pastoris</i>	SNQVYYGGNRYLALYGC(S-farnesyl) SHQVYHNGNKYLALYGC(S-farnesyl)	2	
	<i>Candida albicans</i>	AVRSVSTGNCCSTC(S-farnesyl)	1	
	<i>Schizosaccharomyces pombe</i>	YTPKVPYMC(S-farnesyl)	3	
B	CAAX motif identified from	Position A ₁	Position A ₂	Position X
	Known Ascomycota pheromones	A I L M S T V	A I L M S T V	A C L M Q S V
	Mutational analysis in <i>S. cerevisiae</i>	A D G I L S T V	A G H I L M N Q S T V	A C G I L M Q S N V

Figure 2. Fungal pheromone candidates can be identified by a progressive filter for small open-reading frames that are farnesylated and cleaved by protease associated with mating.

Supp Figure 2.

Figure 3. Fungal pheromones are conserved within closely related species and are often encoded in multiple copies in a genome.

Supp Figure 3.

Figure 4. Best pheromone candidates are weakly associated with motifs that regulate other mating genes.

Supp Figure 4.