Population Genomics of the *Ogataea polymorpha* Species Complex

Sara J. Hanson[1], Paul Fuchs[1], Georgie Nahass[1], and Kenneth H. Wolfe[2]

[1]Department of Molecular Biology, Colorado College, Colorado Springs, CO 80903

[2] UCD Conway Institute, School of Medicine, University College Dublin, Dublin 4, Ireland

Abstract

The methylotrophic yeast *Ogataea polymorpha* has long been a useful system for recombinant protein production, as well as a model system for methanol metabolism, peroxisome biogenesis, thermotolerance, and nitrate assimilation. More recently, it has become an important model for the evolution of mating-type switching. *O. polymorpha* performs mating-type switching by inverting a 19-kilobase region of the genome to move the mating-type (*MAT*) genes between expressed and transcriptionally repressed regions. This mechanism of mating-type switching has evolved independently multiple times in the *Ogataea* clade. Here, we present a population genomics analysis of 50 strains within the *Ogataea polymorpha* species complex from the USDA-NRRL and Phaff yeast culture collections, including representatives from the species *O. polymorpha, O. parapolymorpha, O. haglerorum*, and *O. angusta*. In addition to examining the population structure and genetic variation within and between these species, we also examine the structure and evolution of the *MAT* region across strains to better understand how flip/flop mating-type switching has impacted the genome.

The *Ogataea polymorpha* Species Complex is a Group of Methylotrophic Yeasts

There are four species within the *Ogataea polymorpha* species complex: *Ogataea polymorpha* (Opol), *Ogataea parapolymorpha* (Opar), *Ogataea angusta* (Oang), and *Ogataea haglerorum*(Ohag).

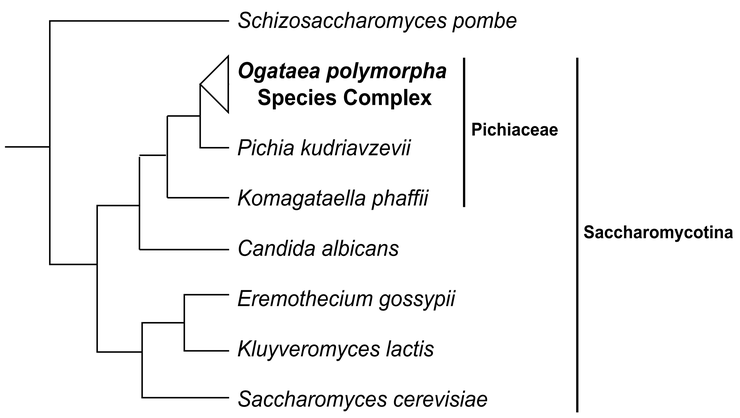

Figure 1. Relationship of *Ogataea polymorpha* species complex to other yeasts in Saccharomycotina. Phylogeny drawn is based on phylogenomic study by Shen et al. 2018. Branch lengths do not reflect actual divergence rates.

Short Read Sequencing and Genome Assembly of 46 Isolates of the *O. polymorpha* Species Complex

Illumina 2 x 150-bp reads were assembled *de novo* for strains from the USDA NRRL, CBS-KNAW, and Phaff collections for ~100X genome coverage.

Table 1. Summary of Genome Assembly Statistics for Strains Used in Study.

Species	# strains sequenced	Mean genome size (Mb)	Mean %GC	Mean. N50 (kb)	Mean SNPs/kb[1]	Mean indels/kb[1]
Ogataea polymorpha	11	8.94	47.71	611.32	3.81	0.17
Ogataea parapolymorpha[2]	3	8.90	47.74	479.63	4.26	0.16
Ogataea angusta	10	8.89	49.45	715.62	5.49	0.20
Ogataea haglerorum	22	8.86	49.36	545.04	2.27	0.11

[1] Variants determined vs. reference genome sequence (Opol = NCYC495; Opar = DL-1; Oang = 61-244; Ohag = 81-453-3)
[2] 2 of 3 sequenced strains appear to be identical to DL-1

Assembly and Conservation of the Mating-Type (*MAT*) Region in the *O. polymorpha* Species Complex

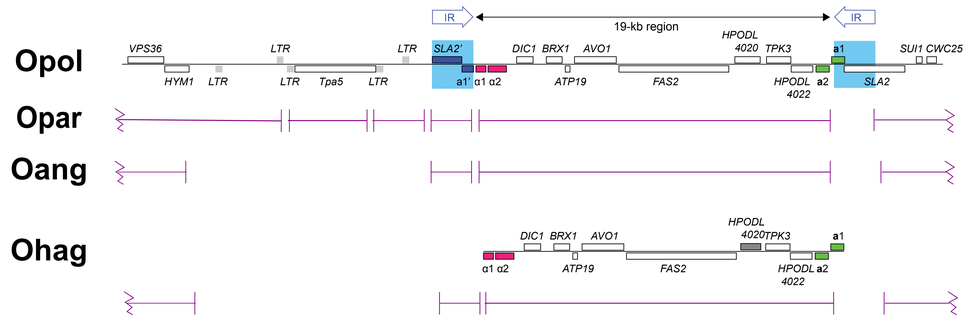

Figure 2. Structure and assembly of the *MAT* region across *O. polymorpha* species complex. The previously published (Hanson et al. 2014, Maekawa et al. 2014, Hanson et al. 2017) reference assembly of the *MAT* region in *O. polymorpha* NCYC495 is given. The *MAT* genes (highlighted in pink and green) are found at opposite ends of a 19-kb region that is flanked by 2-kb inverted repeat (IR) sequences (blue boxes) and adjacent to a Tpa5 and LTR-containing centromere. Coverage of the region by contigs indicated in purple for a representative strain of each of the other species is provided below. Assembly across the adjacent centromere is incomplete in all strains, and for each, the genome assembly resulted in a single contig for the IR sequence that had ~2X higher coverage than the rest of the genome, a separate small contig for the *MAT* region between the IR, and larger contigs for the left and right arms of the chromosome. In Ohag, one of the genes found in the *MAT* region (*HPODL_4020*) has undergone pseudogenization (highlighted in grey).

Genomewide Nucleotide Diversity in *Ogataea* Species

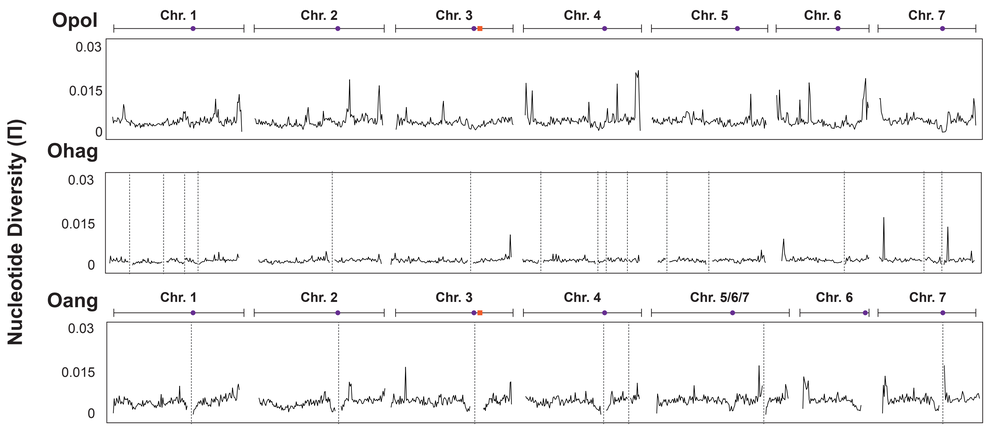

Figure 3. Nucleotide Diversity Across Genome in Opol, Ohag, and Oang. Nucleotide diversity (∏) was calculad in 10-kb windows across the genome of each species. The Ohag and Oang genomes were aligned to Opol NCYC495 chromosomes. Rearrangements have occurred between Oang chromosmes 5, 6, and 7 with respect to the Opol genome. For each chromosome, the centromer position is indicated as a purple circle, and the *MAT* region posiiton on chromosome 3 is indicated by an orange box.

Population Structure of *Ogataea* Species

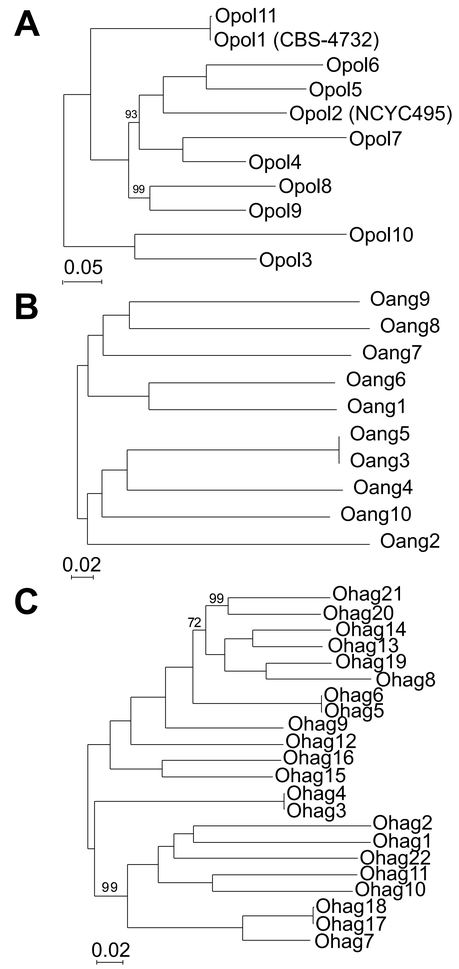

Figure 4. SNP phylogeny of *Ogataea* species. PhyML maximum likelihood analysis (GTR model, 100 bootstrap replicates) of (A) 123,424 SNPs in *O. polymorpha* (B) 149,029 SNPs in *O. angusta* and (C) 82,570 SNPs in *O. haglerorum*. All nodes have 100% support except those indicated.

Acknowledgements

Sequencing was performed by BGI Genomics. Technical suport was provided by Kevin Byrne. Funding for this work was provided through by a Colorado College Benezet start-up grant and a Natrual Sciences Division research grant.

Works Cited

Hanson et al. 2014 PNAS 111(45):E4851-8.
Hanson et al. 2017 PLoS Genetics 13(11): e1007092.
Maekawa and Kaneko 2014 PLoS Genetics 10(11):e1004796.
Shen et al. 2018 Cell 175(6): 1533-1545.