

## **SUPPLEMENTARY MATERIAL**

### **The Methylotroph Gene Order Browser (MGOB) reveals conserved synteny and ancestral centromere locations in the yeast family Pichiaceae**

Alexander P. Douglass<sup>1</sup>, Kevin P. Byrne<sup>1</sup>, Kenneth H. Wolfe<sup>1\*</sup>

<sup>1</sup>UCD Conway Institute, School of Medicine, University College Dublin, Dublin 4, Ireland

\*Corresponding author. Email: [kenneth.wolfe@ucd.ie](mailto:kenneth.wolfe@ucd.ie)

**Figure S1.** Organisation of the nitrate assimilation cluster across the Pichiaceae. Genes in colors belong to the cluster. The two flanking genes in *Ogataea* (white) are genes specific to that genus that have no known association with nitrate. In *B. bruxellensis*, the transcription activators *YNA1* and *YNA2* are present on a different chromosome than the enzyme and transporter genes. The cluster in *O. polymorpha* has been duplicated and is found in two chromosomes: chr. 1 at bases 1511489-1523693, and chr. 7 at bases 18794-6589. The *O. parapolyomorpha* cluster is located on chr. 1 at bases 933071-944111.

**Figure S2.** Genes in ancient adjacency pairs evolve slowly. The curves show the cumulative distribution of nonsynonymous sequence divergence ( $K_A$ ) in orthologs between *P. kudriavzevii* and *O. polymorpha*, as a proxy for a gene's evolutionary rate. The blue curve is for 387 genes that are in ancient adjacencies, i.e. genes that are conserved as neighbors in all nine MGOB species as well as in the YGOB Ancestor. The red curve is for all other orthologs (4152 loci). The left-shifted nature of the anciently adjacent set shows that their values of  $K_A$  tend to be lower, and thus that these genes tend to be slower evolving. The difference is statistically significant ( $P = 8e-6$  by Kolmogorov-Smirnov test). Each point is a  $K_A$  value for an orthologous gene pair, sorted in increasing order of  $K_A$ .

**Table S1.** Genomes included in Figure 3, which shows an amino acid sequence identity curve for every possible pair of species within each GOB. Only non-WGD species were used from YGOB.

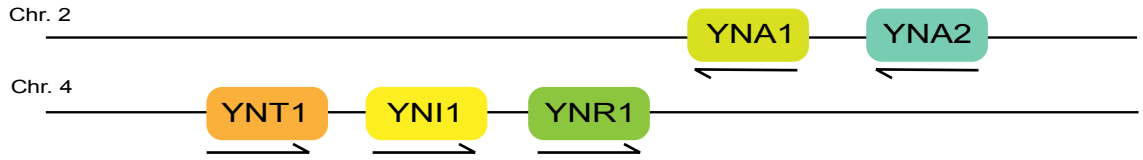
<b>MGOB species name</b>	<b>YGOB species name</b>	<b>CGOB species name</b>
<i>Pichia kudriavzevii</i>	<i>Zygosaccharomyces rouxii</i>	<i>Candida albicans</i> (strain SC5314)
<i>Pichia membranifaciens</i>	<i>Zygosaccharomyces bailii</i>	<i>Candida albicans</i> (strain WO1)
<i>Brettanomyces bruxellensis</i>	<i>Torulasporea delbrueckii</i>	<i>Candida dubliniensis</i>
<i>Ogataea polymorpha</i>	<i>Kluyveromyces lactis</i>	<i>Candida tropicalis</i>
<i>Ogataea parapolyomorpha</i>	<i>Kluyveromyces marxianus</i>	<i>Candida parapsilosis</i>
<i>Kuraishia capsulata</i>	<i>Eremothecium gossypii</i>	<i>Candida orthopsilosis</i>
<i>Komagataella pastoris</i>	<i>Eremothecium cymbalariae</i>	<i>Lodderomyces elongisporus</i>
<i>Komagataella phaffii</i>	<i>Lachancea kluyveri</i>	<i>Debaryomyces hansenii</i>
<i>Pachysolen tannophilus</i>	<i>Lachancea thermotolerans</i>	<i>Scheffersomyces stipitis</i>
	<i>Lachancea waltii</i>	<i>Candida tenuis</i>
		<i>Spathaspora passalidarum</i>
		<i>Meyerozyma guilliermondii</i>
		<i>Candida lusitanae</i>

**Table S2.** Over-represented GOslim terms in genes in ancient adjacencies.

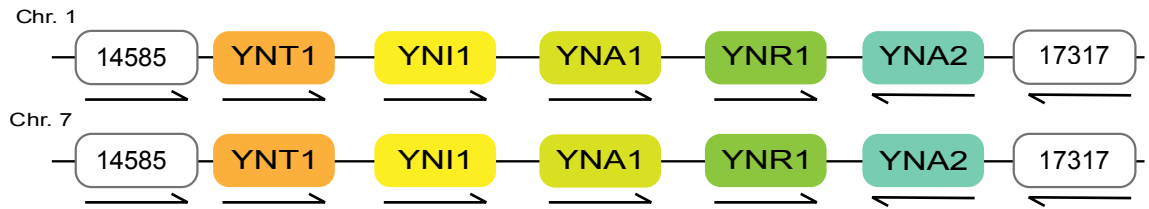
Anciently adjacent genes		Rest of the genome			
# of genes with GO term	%	# of genes with GO term	%	<i>P</i>	GOslim term
16	4.2%	48	1.0%	2.0e-5	Nucleotidyltransferase activity
38.5	10.1%	230	4.9%	2.9e-4	Ribosome
38.5	10.1%	241	5.2%	8.1e-4	Structural molecule activity
40	10.4%	264	5.7%	9.6e-4	Ribosome biogenesis

All GOslim terms that are statistically significantly over-represented ( $P < 0.001$ , Chi-squared test) are listed. No terms are significantly under-represented at this level. Ancient adjacencies are defined as pillars that are adjacent in the YGOB Ancestor and all 9 Pichiaceae genomes in MGOB. The list of genes with the term “nucleotidyltransferase activity” includes 10 subunits of RNA polymerase complexes (*RPB4*, *RPB5*, *RPB7*, *RPB9*, *RPC19*, *RPC25*, *RPC82*, *RPO21*, *RPO26*, *RPO41*). The lists of genes with the terms “ribosome”, “ribosome biogenesis”, and “structural molecule activity” overlap extensively and include 16 cytosolic ribosomal protein genes (*RPL11*, *RPL12*, *RPL13*, *RPL14*, *RPL17*, *RPL21*, *RPL28*, *RPL35*, *RPS0*, *RPS1*, *RPS2*, *RPS11*, *RPS15*, *RPS16*, *RPS23*, *RPS25*) and 13 mitochondrial ribosomal protein genes (*MRP7*, *MRP21*, *MRPL11*, *MRPL16*, *MRPL19*, *MRPL23*, *MRPL31*, *MRPL32*, *RSM10*, *RSM18*, *RSM23*, *RSM24*, *RSM26*). This analysis was conducted by connecting MGOB pillars to GOslim terms via the *S. cerevisiae* genes they contain; pillars without *S. cerevisiae* genes were ignored, leaving 383 (out of 387) pillars in ancient adjacencies, which were compared to the remaining 4676 pillars that contain *S. cerevisiae* genes and are not in ancient adjacencies.

*Brettanomyces bruxellensis*



*Ogataea polymorpha*



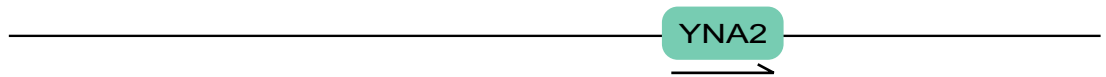
*Ogataea parapolyomorpha*



*Kuraishia capsulata*



*Komagataella phaffii*



*Pachysolen tannophilus*

