FEMS Yeast Research, 19, 2019, foz058

doi: 10.1093/femsyr/foz058
Advance Access Publication Date: 9 August 2019
Research Article

RESEARCH ARTICLE

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

Alexander P. Douglass, Kevin P. Byrne and Kenneth H. Wolfe*,†

UCD Conway Institute, School of Medicine, University College Dublin, Belfield, Dublin 4, Ireland

*Corresponding author: UCD Conway Institute, School of Medicine, University College Dublin, Belfield, Dublin 4, Ireland. Tel: +35317166712; E-mail: kenneth.wolfe@ucd.ie

One sentence summary: Using MGOB, a comparative genomics browser for methylotrophic yeasts, we show that centromeres have remained constant in location despite changing their structures.

Editor: Terrance Cooper
†Kenneth H. Wolfe, http://orcid.org/0000-0003-4992-4979

ABSTRACT

The yeast family Pichiaceae, also known as the 'methylotrophs clade', is a relatively little studied group of yeasts despite its economic and clinical relevance. To explore the genome evolution and synteny relationships within this family, we developed the Methylotroph Gene Order Browser (MGOB, http://mgob.ucd.ie) similar to our previous gene order browsers for other yeast families. The dataset contains genome sequences from nine Pichiaceae species, including our recent reference sequence of *Pichia kudriavzevii*. As an example, we demonstrate the conservation of synteny around the *MOX1* locus among species both containing and lacking the *MOX1* gene for methanol assimilation. We found ancient clusters of genes that are conserved as adjacent between Pichiaceae and Saccharomycetaceae. Surprisingly, we found evidence that the locations of some centromeres have been conserved among Pichiaceae species, and between Pichiaceae and Saccharomycetaceae, even though the centromeres fall into different structural categories—point centromeres, inverted repeats and retrotransposon cluster centromeres.

Keywords: comparative genomics; bioinformatics; centromeres

INTRODUCTION

The Pichiaceae is a very significant but somewhat an understudied family of budding yeasts. At least 30 species in this family are referred to as methylotrophs because they can grow on methanol as a sole carbon source, which is a trait that is not seen in any other yeasts (Riley *et al.* 2016). As a result, this family is commonly referred to as the 'methylotrophs clade'. Their ability to consume methanol is conferred by the *MOX1* gene for methanol oxidase (also known as alcohol oxidase—*AOX1* or *AOD1*). The promoter of *MOX1* is strongly induced by methanol, and this easily inducible genetic system has been exploited in biotechnology for the mass production of recombinant proteins in methylotrophic yeasts such as *Ogataea polymorpha* and *Komagataella phaffii* (Mattanovich *et al.* 2012). Other species in the family do not assimilate methanol. One of these is *Pichia kudriavzevii*, which is used in some traditional food fermentations and has a growing role in biotechnology due to its high resistance to multiple stresses. We recently generated a high-quality reference genome sequence for *Pi. kudriavzevii* and showed that this species is identical to *Candida krusei*, which is an opportunistic pathogen with a high intrinsic resistance to the antifungal drug fluconazole (Douglass *et al.* 2018).

Although research into species such as *O. polymorpha* and *Pi. kudriavzevii* has been carried out for several decades, it is

Received: 24 April 2019; **Accepted:** 8 August 2019

© FEMS 2019. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Table 1. Genomes and species included in the MGOB database.

Species	Strain	Number of genes	Genome size (Mbp)	Chromosomes	Scaffolds[a]	Reference
Pichia kudriavzevii	CBS 573	5140	10.81	5	5	(Douglass *et al.* 2018)
Pichia membranifaciens	NRRL Y-2026	5542	11.58	ND[d]	10	(Riley *et al.* 2016)
Brettanomyces bruxellensis	UMY321	5428	12.97	4–9[d]	8	(Fournier *et al.* 2017)
Ogataea polymorpha[b]	NCYC 495	5501	8.97	7	7	(Riley *et al.* 2016)
Ogataea parapolymorpha[b]	DL-1	5325	8.87	7	7	(Ravin *et al.* 2013)
Kuraishia capsulata	CBS 1993	5989	11.37	7	7	(Morales *et al.* 2013)
Komagataella pastoris[c]	NRRL Y-1603	5029	9.42	4	4	(Love *et al.* 2016)
Komagataella phaffii[c]	CBS 7435	5223	9.38	4	4	(Sturmberger *et al.* 2016)
Pachysolen tannophilus	NRRL Y-2460	5346	12.25	7–8[d]	9	(Riley *et al.* 2016)
Saccharomyces cerevisiae	S288C	5600	12.16	16	16	(Engel *et al.* 2014)
YGOB Ancestor	N/A	4754	N/A	8	8	(Gordon, Byrne and Wolfe 2009)

[a]Number of scaffolds larger than 5 kb, excluding mitochondrial DNA.
[b]*Ogataea polymorpha* and *O. parapolymorpha* are two separate but closely related species, which were both previously known as *H. polymorpha* (Kurtzman 2011a).
[c]*Komagataella phaffii* and *Ko. pastoris* are two separate but closely related species, which were both previously known as *Pi. pastoris* (Kurtzman 2009).
[d]Range of chromosome numbers estimated by pulsed field gel electrophoresis for multiple strains of *B. bruxellensis* (Hellborg and Piskur 2009) and for the type strain of *Pa. tannophilus* (Maleszka and Skrzypek 1990). ND, not determined.

only recently that molecular biology researchers have begun to appreciate that these species form a third clade (family Pichiaceae) of budding yeasts that is very separate from the two better-known clades that contain *Saccharomyces cerevisiae* and *C. albicans*. Consequently, it is often more informative to compare Pichiaceae species to each other than to *S. cerevisiae* or *C. albicans*.

It has recently been discovered that a small clade of species historically classified within the Pichiaceae uses a novel genetic code in which CUG is translated as alanine (CUG-Ala clade), whereas most Pichiaceae species use the standard genetic code (CUG-Leu2 clade) (Mühlhausen *et al.* 2016; Riley *et al.* 2016; Krassowski *et al.* 2018). The divergence between the CUG-Ala and CUG-Leu2 clades forms a deep evolutionary split, and it has been proposed that the CUG-Ala clade should be recognised as a family separate from, but sister to, the Pichiaceae (Shen *et al.* 2018).

Here we present MGOB (Methylotroph Gene Order Browser), a comparative genomics browser that enables gene orthology and synteny comparisons to be made among the genomes of Pichiaceae species. MGOB is based on an underlying software platform that we previously developed for the browsers YGOB (Yeast Gene Order Browser, which covers family Saccharomycetaceae, including *S. cerevisiae*) and CGOB (*Candida* Gene Order Browser, which covers families Debaryomycetacae and Metschnikowiaceae, including *C. albicans*), as well as OGOB for oomycete species (Byrne and Wolfe 2005; Maguire *et al.* 2013; McGowan, Byrne and Fitzpatrick 2019). MGOB incorporates data from nine Pichiaceae species for which well-annotated and highly contiguous genome sequences are available, including one (*Pachysolen tannophilus*) from the CUG-Ala clade and eight from the CUG-Leu2 clade (Krassowski *et al.* 2018).

MGOB can be used online interactively to compare the syntenic context around any gene in multiple Pichiaceae species. In this study, we also use the database underlying MGOB to explore the extents of synteny conservation within the three major clades of budding yeasts represented by the MGOB, YGOB and CGOB databases, and to investigate the evolution of centromere locations.

MATERIALS AND METHODS

Sources of genome sequence data are listed in Table 1. Fully contiguous chromosome sequences, including annotated centromeres, were available for *Pi. kudriavzevii*, *O. polymorpha*, *Ku. capsulata* and *Ko. phaffii*. *Brettanomyces bruxellensis* (also known as *Dekkera bruxellensis*) has full-length annotated chromosomes, but its centromeres have not been identified (Fournier *et al.* 2017). The *Ko. pastoris* genome is very similar to the *Ko. phaffii* genome (90% DNA sequence identity), with only two reciprocal translocations between them (Love *et al.* 2016), and the locations of centromeric inverted repeats (IRs) are conserved between them. Similarly, *O. parapolymorpha* is very similar to *O. polymorpha*, with no translocations between them (Hanson, Byrne and Wolfe 2014), and the locations of centromeric retrotransposon clusters are conserved. The *Pi. membranifaciens* genome sequence consists of relatively short scaffolds (Riley *et al.* 2016), but we included this species because it is the type species of the genus *Pichia* (Kurtzman 2011c).

The pillars of homologous genes in MGOB were constructed by using BLAST and syntenoBLAST (Maguire *et al.* 2013) to identify syntenic orthologs in each species (or ohnologs in *S. cerevisiae*). Pillars were then checked and edited by manual curation. Specifically, every time a genome was to be added to MGOB, reciprocal best hits in BLASTP searches (with a conservative cutoff), against the most closely related previously loaded genome, established an initial layer of homology with which to load the genome into MGOB. syntenoBLAST was then used to interpret weaker BLAST scores in combination with synteny information, systematically searching for putative orthologs by looking for singleton pillars that could be merged into another pillar on the basis of a BLASTP hit to at least one gene in the pillar, provided that the assignment was also supported by the syntenic context. After constructing the initial set of MGOB pillars in this manner, we then used computer scripts to search for situations where pairs of nearby partially filled pillars could potentially be merged on the basis that (i) there is a sequence similarity between the two pillars and (ii) no species occurs in both of the pillars. Each candidate pair of pillars of this type was examined manually and merged if considered to be orthologous.

Conservation of centromere adjacency was investigated by calculating the distance (in number of genes) from the centromere for every gene in the dataset, for species with known centromere locations. Average distances to centromeres were calculated for every possible pair of genes from different species in the same MGOB pillar. For each pair of species, the distances were then sorted to find the gene pairs with the lowest average

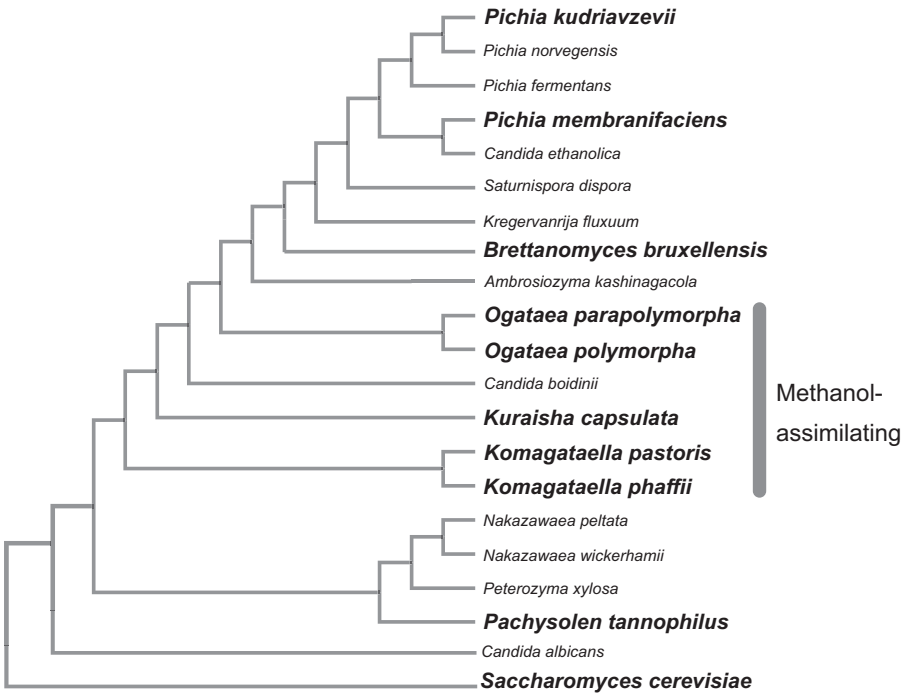

Figure 1. Phylogenetic tree of the Pichiaceae family, based on Kurtzman and Robnett (2010) and Douglass *et al.* (2018). Species included in the MGOB dataset are shown in bold. Species capable of assimilating methanol are highlighted.

distance to the centromere in the two species. In order to test the statistical significance of these centromeric adjacencies, we developed a simulated dataset of centromere distances, based on randomised gene pairs. Pillar content was shuffled so that the orthology relationships among genes were randomised, without changing the location of each gene on its chromosome, and without changing the locations of centromeres.

RESULTS AND DISCUSSION

The MGOB dataset and interface

MGOB version 1.0 includes data for nine Pichiaceae species (Table 1). The phylogenetic relationship among them, and their relationship to other species in the family, is shown in Fig. 1. The dataset used in MGOB includes every Pichiaceae species for which a high-contiguity and reasonably well annotated genome sequence was available in mid-2018. We excluded some species for which the only available sequence was fragmented into a large number of contigs or scaffolds. As non-Pichiaceae reference genomes, the dataset also includes *S. cerevisiae* (genome version R64-1-1 from SGD), and the gene order inferred for the 'Ancestor' of the post-Whole Genome Duplication (WGD) clade in YGOB (Gordon, Byrne and Wolfe 2009).

Similar to the existing YGOB framework (Byrne and Wolfe 2005, 2006), MGOB consists of (i) a curated database of homology assignments, (ii) a software engine for assessing synteny across genomes, supporting (iii) a web interface that allows users to visualise the syntenic context of any gene. MGOB works by storing sets of homologous genes in 'pillars', and representing genes visually along horizontal 'tracks', which represent segments of a chromosome, presenting an output screen, which is a matrix with pillars as columns and tracks as rows (Fig. 2). Each horizontal track shows genes from a chromosomal region in one species, but *S. cerevisiae* has two tracks due to the WGD. The web interface to MGOB is publicly available at http://mgob.ucd.ie. Details of the interface are explained in Fig. 2.

MOX1 and nitrate cluster loci

This example screenshot (Fig. 2) is centred on the *MOX1* gene of *O. polymorpha*, which is required for the assimilation of methanol as a carbon source (Ito *et al.* 2007; Yurimoto, Oku and Sakai 2011). This trait is only seen in Pichiaceae, but it is not universally present in all species in the family (Ravin *et al.* 2013). *MOX1* orthologs are present in *Ogataea*, *Komagataella* and *Kuraishia*, but absent in *Brettanomyces*, *Pichia* and *Pachysolen*. A comparison to the phylogenetic tree in Fig. 1 suggests that methanol assimilation was gained after the divergence of *Pachysolen*, and was later lost in the common ancestor of *Pichia* and *Brettanomyces*. Interestingly, the *Pichia* and *Brettanomyces* genomes show conserved synteny with *Ogataea* in a block of three to four genes that spans the *MOX1* locus, even though *MOX1* itself has been deleted, as first noted by Ravin *et al.* (2013) for *Brettanomyces*. There is no conservation of synteny between the locations of *MOX1* or its flanking genes in the genomes of *Ogataea, Komagataella* and *Kuraishia*. Some methylotrophs contain two separate, unlinked, genes for isozymes of methanol oxidase (Ito *et al.* 2007), but none of the genomes in MGOB are from species of this type.

Another experimentally characterised region for which MGOB revealed new information is the nitrate assimilation cluster. This cluster was first described in *Ogataea* and contains *YNT1* (transporter for the uptake of nitrate), *YNR1* (nitrate reductase), *YNI1* (nitrite reductase), *YNA1* and *YNA2* (transcription factors) (Perez *et al.* 1997; Ávila *et al.* 2002; Silvestrini *et al.* 2015). *Ogataea polymorpha* contains two highly similar clusters on two different chromosomes, whereas *O. parapolymorpha* has only one. It has previously been suggested that the nitrate cluster was acquired

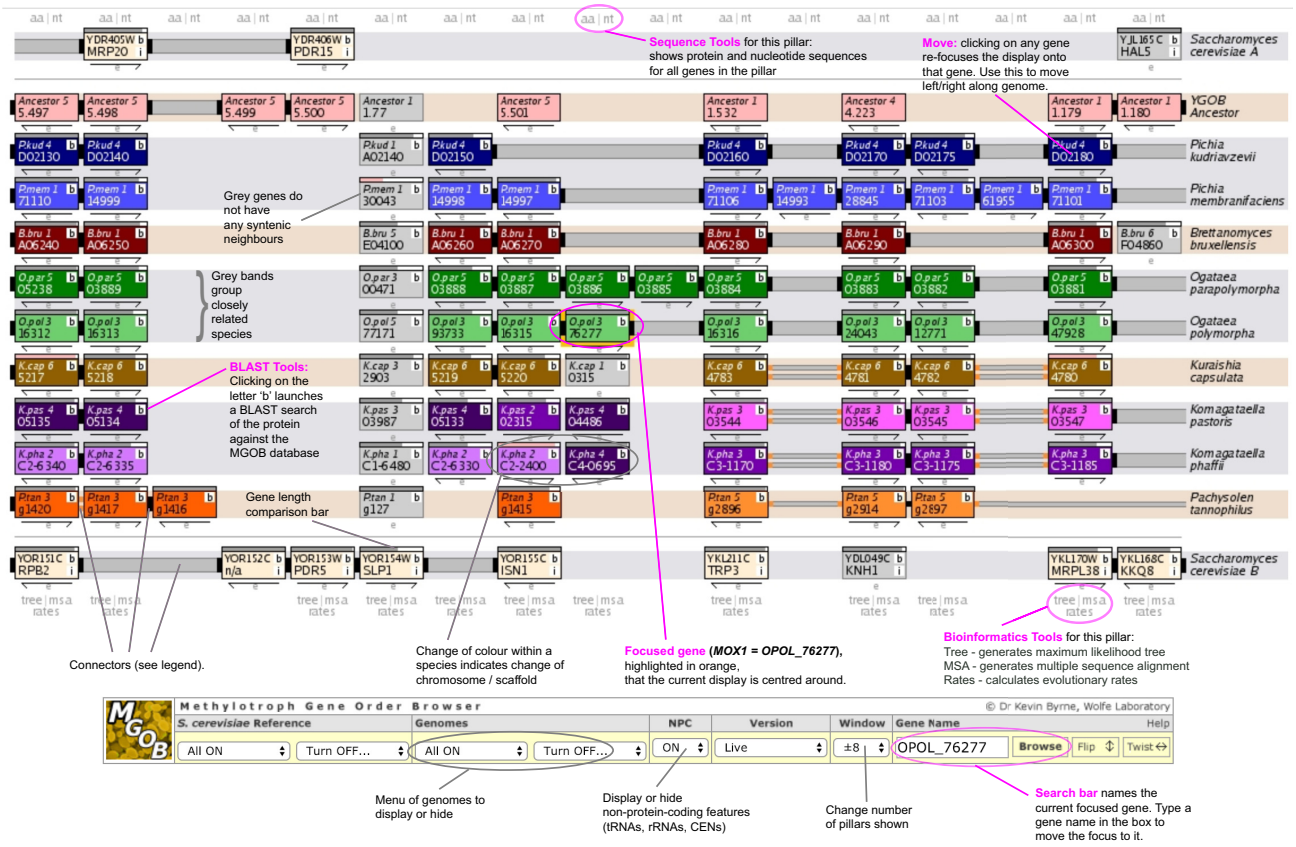

Figure 2. Annotated screenshot of the MGOB web interface. The most important features are labelled in magenta. Each box represents a gene. This screenshot is focused on the *MOX1* (OPOL_76277) gene of *O. polymorpha*, in the centre of the screen and surrounded by an orange outline. Vertical columns (pillars) show orthologous genes in each species, where present. Horizontal rows (tracks) show sections of chromosome from each species, around the pillar containing the focused gene. The connectors between genes in the same track are drawn in different styles to indicate different levels of adjacency: immediately neighboring genes (thick black connectors and thick grey lines; clicking on these shows the intergenic DNA sequence), genes <5 positions apart (two thin grey lines), genes 5–20 positions apart (one thin grey line), endpoints of inversions (orange marks on connectors, e.g. between the *TRP3* and *KNH1* orthologs in four species).

by a horizontal transfer from the Pezizomycotina after the common ancestor of *Brettanomyces, Kuraishia* and *Ogataea* diverged from *Komagataella* (Morales *et al.* 2013). However, using MGOB we find that three of these genes (*YNT1*, *YNA2* and *YNR1*) are present and clustered in *Pa. tannophilus* (BLASTP E-values in the range 1e−158 to 0.0; Fig. S1, Supporting Information). Additionally, we also found orthologs of the transcription factor *YNA2* in both *Komagataella* species, which do not assimilate nitrate (Kurtzman 2011b). We suggest that the nitrate cluster was acquired by the ancestor of the entire Pichiaceae family, followed by losses of some genes in *Pachysolen* and *Komagataella*.

Comparison of sequence divergence and synteny divergence in the MGOB, YGOB and CGOB datasets

Our three Gene Order Browsers (GOBs) contain data from three large families of budding yeasts: Pichiaceae (MGOB), Saccharomycetaceae (YGOB) and Debaryomycetaceae/Metschnikowiaceae (CGOB). We investigated how these datasets compare in terms of their levels of sequence diversity and synteny conservation. To measure sequence diversity within each GOB, we calculated the level of protein sequence identity for all orthologs between all pairs of species within the GOB, using ClustalW alignments (Larkin *et al.* 2007). We then plotted the distribution of sequence identity levels, similar to the approach taken by Dujon *et al.* (2004). For YGOB, we used only data from non-WGD genera (*Kluyveromyces*, *Lachancea*, *Eremothecium*, *Zygosaccharomyces*, *Torulaspora*), to allow us to compare its synteny conservation relative to MGOB and CGOB without the complication of post-WGD gene deletions.

Figure 3 shows the distribution of protein sequence identity levels between all pairs of genomes within each of the three GOBs. In YGOB, the majority of pairwise distributions centre around 50–55% amino acid sequence identity, with the curves for a very few highly similar species pairs (eg. *Z. rouxii* vs. *Z. bailii*) lying to the right. A similar pattern is observed in the CGOB species. The MGOB species, in contrast, show lower levels of sequence conservation with amino sequence identity for most species pairs centred on 40%. These plots show that, in general, interspecies orthologs in MGOB are more divergent from each other than in the other databases. Within MGOB, the two *Ogataea* species are the most similar pair, followed by the two *Komagataella* species. The two *Pichia* species (*Pi. kudriavzevii* and *Pi. membranifaciens*) form a more divergent pair with a peak at 70% identity, which is approximately the same as *Z. rouxii* vs. *Z. bailii*, or *C. albicans* vs. *C. tropicalis* (Fig. 3).

Sequence divergence in genes and rearrangements of gene order along chromosomes both accumulate over evolutionary time so it is expected that these two quantities will be correlated (Dujon *et al.* 2004; Rolland and Dujon 2011; Vakirlis *et al.* 2016). To examine this correlation in our data, we calculated the number of shared adjacencies between every pair of species in each GOB.

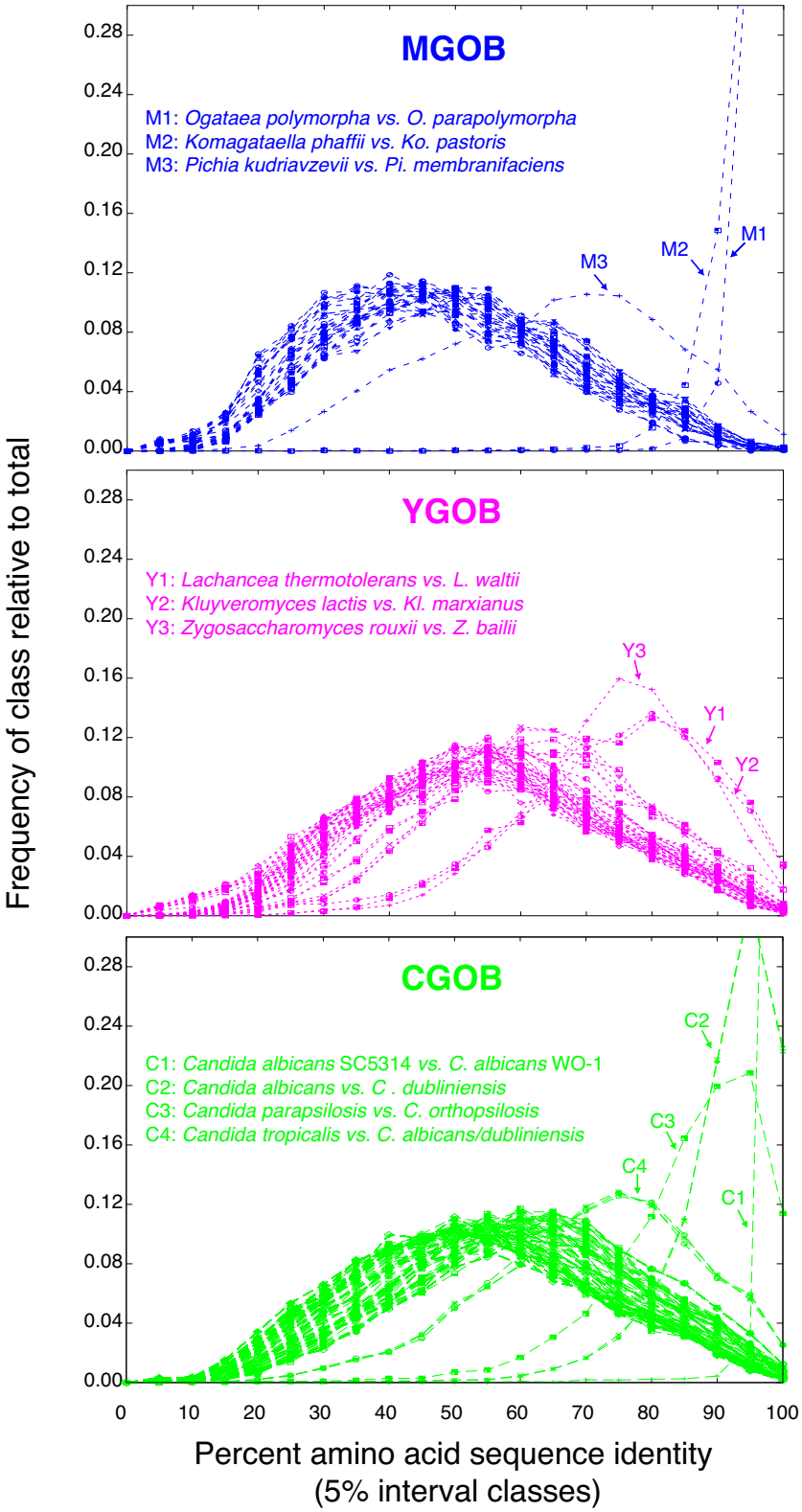

Figure 3. Distribution of sequence identity in orthologous proteins, for all pairs of genomes in each GOB. Each curve compares all orthologous proteins from one pair of genomes. The X-axis is % protein sequence identity (in 5% bins) and the Y-axis is the fraction of proteins with that level of sequence identity. Labels M1–M3, Y1–Y3 and C1–C4 identify the curves from the three to four closest genome pairs in each GOB. Complete lists of all the species included in each plot are given in Table S1 (Supporting Information).

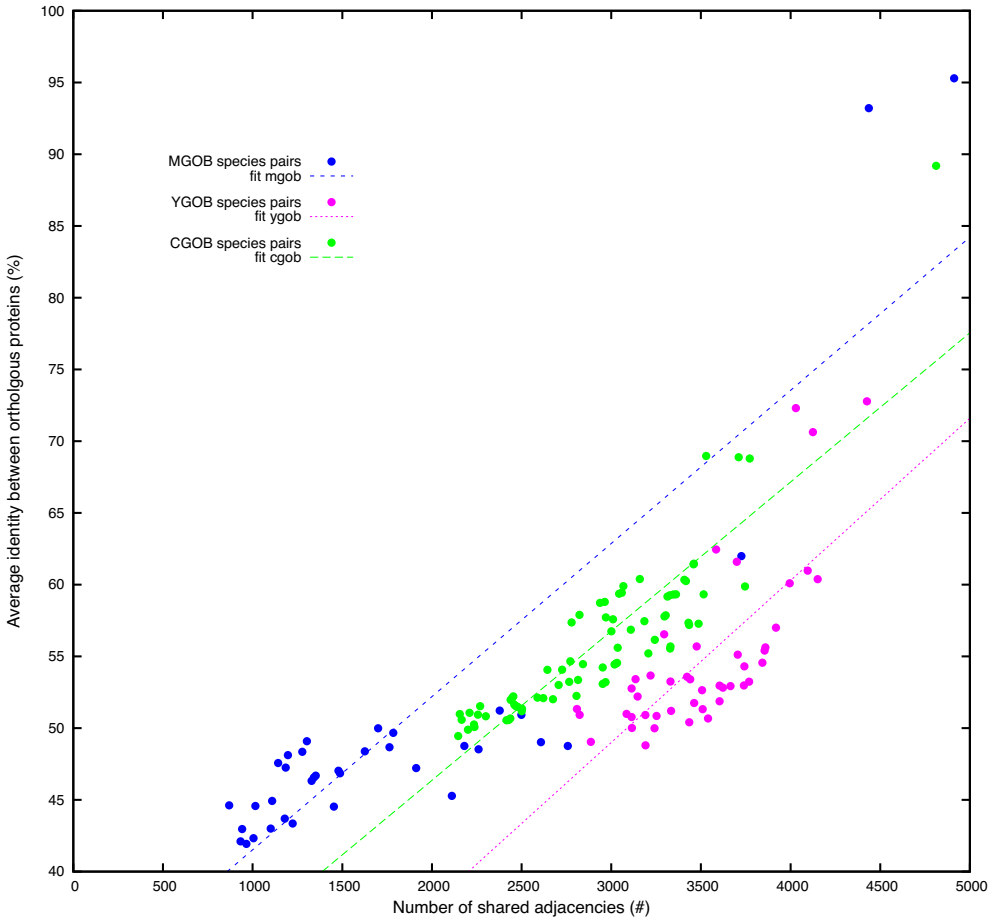

Figure 4. Correlation between sequence divergence and gene order divergence. The relationship between average protein sequence identity (Y-axis) and the total number of shared gene adjacencies (X-axis) is plotted. Each point is a pair of species in MGOB (blue), YGOB (pink) or CGOB (green).

Table 2. Percentages of shared adjacencies between each species in the MGOB dataset, including *S. cerevisiae* and the YGOB Ancestor (Anc).

	Anc	*P.kud.*	*P.mem.*	*B.bru.*	*O.par.*	*O.pol.*	*K.cap.*	*K.pas.*	*K.pha.*	*P.tan.*
S.cer.	50%	9%	8%	10%	9%	10%	10%	10%	10%	13%
Anc		15%	14%	18%	16%	17%	19%	17%	18%	23%
P.kud.	14%		**73%**	41%	51%	54%	28%	21%	24%	18%
P.mem.	12%	**67%**		35%	43%	45%	24%	18%	20%	16%
B.bru.	16%	39%	35%		40%	42%	22%	17%	19%	18%
O.par.	14%	49%	45%	41%		**92%**	32%	25%	28%	21%
O.pol.	15%	50%	45%	41%	**89%**		32%	24%	27%	22%
K.cap.	15%	24%	22%	20%	28%	30%		27%	29%	22%
K.pas.	16%	22%	20%	19%	27%	27%	32%		**88%**	24%
K.pha.	16%	23%	21%	19%	28%	28%	34%	**85%**		24%
P.tan.	20%	18%	16%	18%	21%	22%	24%	23%	24%	

Numbers represent the percentage of genes in one species (rows) that have shared adjacencies in the other species (columns). Bold cells show intragenus comparisons. The matrix is not perfectly symmetrical due to the variation in the number of genes among species. Full species names are given in Table 1.

This quantity is the number of orthologs in the two species that are immediate chromosomal neighbours (in both genomes) of another pair of orthologs.

All three GOB datasets showed clear correlations between levels of protein sequence identity and levels of synteny conservation, in pairs of species (Fig. 4). The three datasets have almost identical slopes for linear regression fits (0.010–0.011), although their extrapolated Y-axis intercepts are different (YGOB: 15% sequence identity; CGOB: 26%; MGOB: 31%). Most of the MGOB data do not overlap the range of data in the other GOBs, as they are lower on both axes. Interestingly, for any given level of sequence identity, the number of shared adjacencies is lower in MGOB than in CGOB, which, in turn, is lower than YGOB (Fig. 4). This result shows that the Pichiaceae species are more diverged than the other families, as measured by both sequence divergence and loss of synteny. These differences highlight the importance of using separate GOBs for different yeast families that cover different branches of the phylogenetic tree. We did

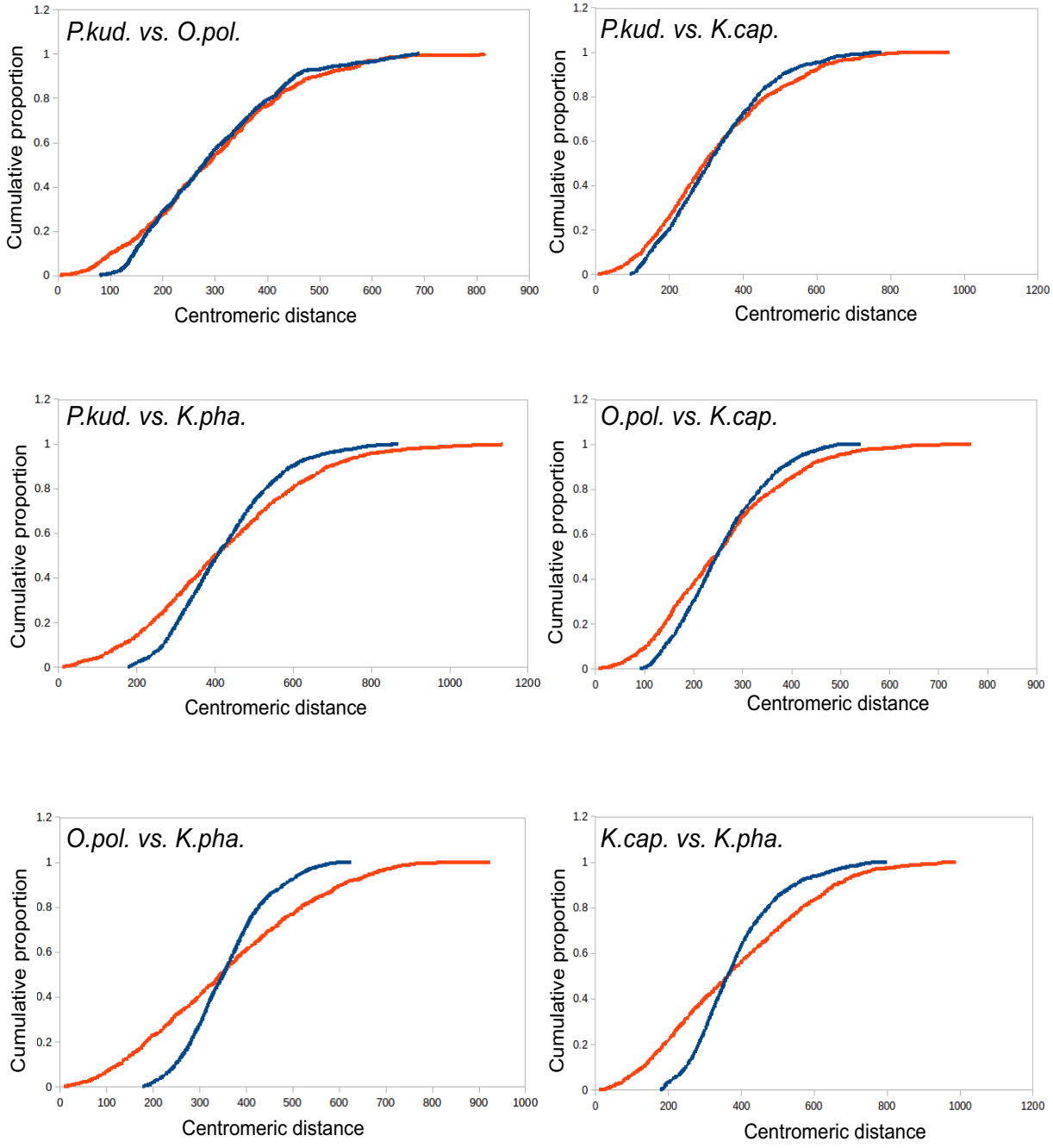

Figure 5. Interspecies conservation of centromere linkage. Each plot shows one species pair. Red curves (real data) show the cumulative distribution of the average distance to the centromere in the two species, for genes in a pillar (i.e. orthologs), after sorting the pillars in increasing order of distance. Blue curves (randomisations) show the means of cumulative distributions from 1000 simulations in which pillar content was shuffled (see MATERIALS AND METHODS). Distances were measured as numbers of genes.

not attempt to incorporate the MGOB, CGOB and YGOB data into a single browser because the level of conserved synteny between the different yeast families is too low.

Synteny conservation within Pichiaceae

Whereas Fig. 4 gives an overview of genome divergence in the three yeast families represented by the three GOBs, we also examined synteny conservation within the Pichiaceae in more detail. Table 2 shows levels of shared adjacency, expressed as a percentage of the number of genes in the genome, for each pair of species in Pichiaceae. The highest levels of adjacency conservation are within the *Ogataea*, *Komagataella* and *Pichia* species pairs, in that order, matching the order of sequence conservation in these genera (Fig. 3). Between genera, synteny conservation is highest between *Pichia* and *Ogataea* (43–54%), and lowest between *Pichia* and *Pachysolen* (16–18%) (Table 2). The low level of synteny conservation between *Pachysolen* and other genera is consistent with its phylogenetic position in the CUG-Ala clade and as an outgroup to the other MGOB species (Fig. 1).

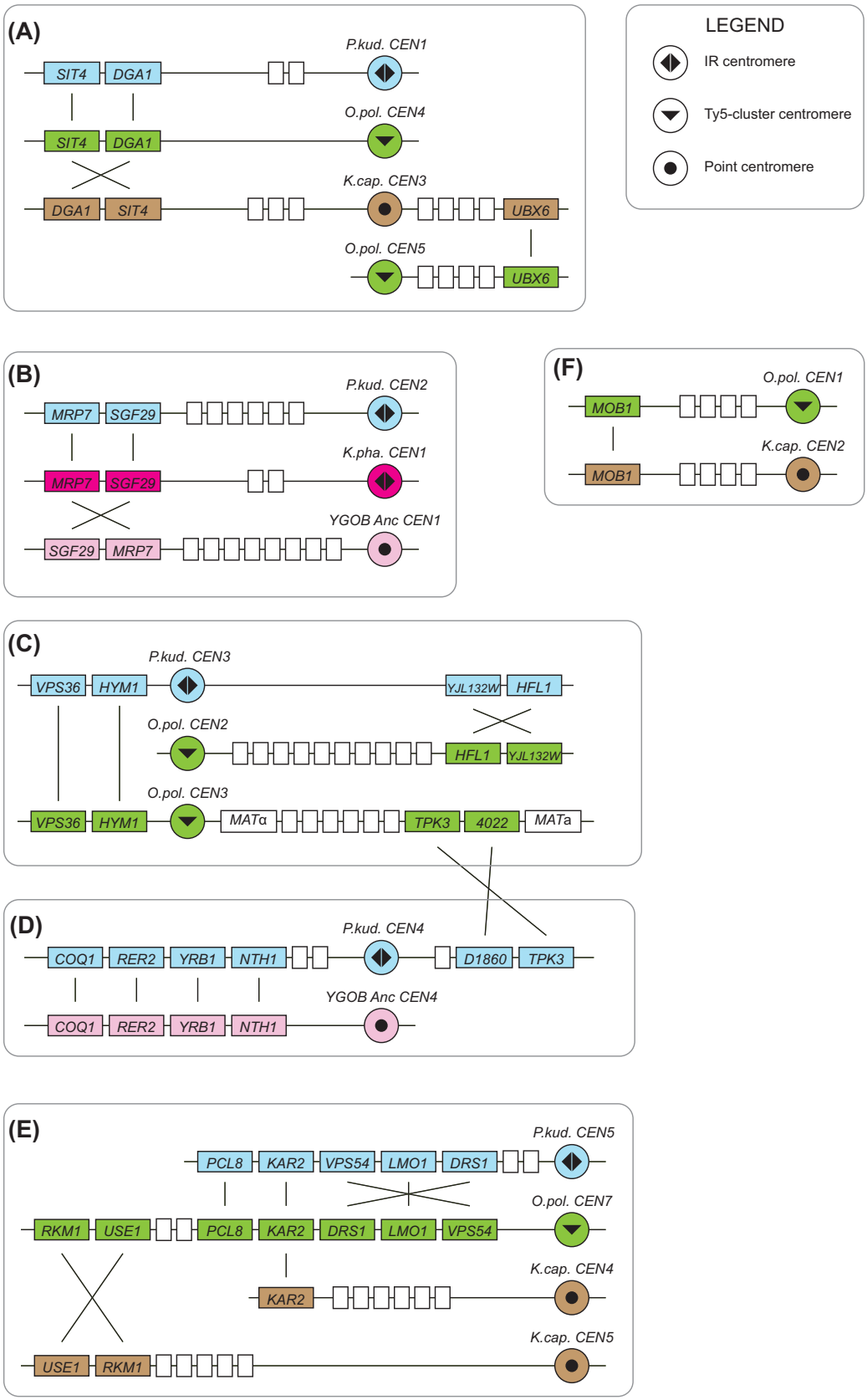

Figure 6. Centromeres are conserved in location but not in structure. **A–E**, relationships involving the five centromeres of *Pi. kudriavzevii*. **F**, relationship between *O. polymorpha CEN1* and *Ku. capsulata CEN2*. Circles represent centromeres and rectangles represent protein-coding genes. Small white rectangles represent genes without orthologs in these regions in other species. Connecting lines indicate the conservation of gene order, including some small inversions.

Interestingly, *B. bruxellensis* shows accelerated levels of both sequence divergence and genome rearrangement, relative to *Pichia*. *Ogataea* is an outgroup to *Brettanomyces* + *Pichia* (Fig. 1), but the *Brettanomyces/Pichia* pair shows fewer shared adjacencies and more sequence divergence than the *Ogataea/Pichia* pair (Table 2; Fig. 3).

We found that levels of synteny conservation between genera are lower in Pichiaceae than in Saccharomycetaceae. For non-WGD Saccharomycetaceae species, the proportion of adjacencies shared between genera ranges from 54% (*Kluyveromyces marxianus* vs. *Eremothecium cymbalariae*) to 82% (*Zygosaccharomyces rouxii* vs. *Torulaspora delbrueckii*). In contrast, in Pichiaceae, it ranges from only 16 to 54% (Table 2). Considering that the YGOB dataset includes genome sequences from all known genera of family Saccharomycetaceae, whereas the MGOB dataset is relatively incomplete, this result suggests that Pichiaceae encompasses a deeper evolutionary divergence than Saccharomycetaceae. Consistent with this, Shen *et al.* (2018) estimated that the deepest divergence within Pichiaceae (including the CUG-Ala clade) is 204 Myr, whereas within Saccharomycetaceae it is 114 Myr.

Ancient synteny

The YGOB 'Ancestor' gene order is the order of genes along chromosomes that was inferred to have existed in the ancestral Saccharomycetaceae species that underwent WGD (Gordon, Byrne and Wolfe 2009). Even though evidence now indicates that the WGD was the result of hybridisation between two distinct species, the two parents of the hybrid appear to have had almost no differences in their gene orders (Gordon, Byrne and Wolfe 2009; Marcet-Houben and Gabaldón 2015), so the YGOB Ancestral gene order is still a useful concept. The proportion of shared adjacencies between the YGOB Ancestor and the MGOB species ranges from 12 to 23% (Table 2). It is interesting that *Pachysolen*'s level of synteny conservation to the Ancestor (20–23%) is the highest among Pichiaceae (Table 2) and approximately the same as between *Pachysolen* and other Pichiaceae species (16–24%), even though the Ancestor represents a different yeast family, Saccharomycetaceae. These results suggest that the *Pachysolen* genome is less rearranged, relative to the common ancestor of the two families, than other Pichiaceae genomes.

The fact that about one-fifth of gene adjacencies are shared between the YGOB Ancestor and Pichiaceae species (Table 2) suggests the existence of ancient pairs of neighbouring genes that have been preserved as neighbours during hundreds of millions of years of evolution. We searched for gene pairs that are conserved as immediate chromosomal neighbours in all nine MGOB species and the YGOB Ancestor, and found 205 such pairs. These 205 pairs correspond to 4% of adjacencies in the YGOB Ancestor, and are located in 181 syntenic blocks. The longest block consists of six genes: *CPR6*, *RPO21*, *BPL1*, *CRD1*, *CCT4* and*CDC123* [loci Anc_7.313 to Anc_7.318 in the Ancestral genome nomenclature (Gordon, Byrne and Wolfe 2009)]. The second longest consists of four genes: *COQ1*, *RER2*, *YRB1* and *NTH1* (Anc_3.202 to Anc_3.205, which is close to a centromere). There is no obvious functional link between the genes in either of these clusters. There were also 19 triplets, and the remaining 160 were pairs. These anciently syntenic regions encompass 387 genes in total.

Genes that are part of these ancient adjacencies were found to be slower evolving than the rest of the genome (Fig. S2, Supporting Information). The average non-synonymous divergence (K_A) in these ancient adjacency genes is 0.46, compared to 0.53 in other genes, for comparisons between *Pi. kudriavzevii* and *O. polymorpha* orthologs. The rate difference is statistically significant ($P = 8e{-}6$ by the Kolmogorov–Smirnov test). These genes are also more likely to be essential in *S. cerevisiae* (29% essential vs. 17% for other genes; $P = 1.2e{-}5$ by Fisher's exact test). This result is consistent with the previous observation that regions containing essential genes are less likely to undergo rearrangements in multiple ascomycete families (Fischer *et al.* 2006). By Gene Ontology analysis (Table S2, Supporting Information), we found that the lists of genes involved in ancient adjacencies are enriched in ribosomal protein genes (both cytosolic and mitochondrial) and genes for subunits of RNA polymerase. The enrichment of ribosomal protein and RNA polymerase genes explains why anciently adjacent genes are more likely to be slow-evolving and essential. However, the ancient adjacencies generally do not involve pairs of ribosomal protein genes or RNA polymerase genes, but genes of these types adjacent to other genes.

Conserved centromere locations

Centromeres have been characterised in four methylotroph species in our dataset: *Pi. kudriavzevii*, *Ko. phaffii*, *O. polymorpha* and *Ku. capsulata*. The centromeres are annotated in each of these genomes and are shown as features in MGOB (they have names such as *Pkud_CEN1*; see also the online help pages). The four species show an enormous diversity of centromere structures, so we were curious to investigate whether there is any conservation of centromere locations. In *Ko. phaffii* and *Pi. kudriavzevii*, centromeres consist of simple IR structures. Each chromosome has two near-identical sequences in opposite orientations, separated by a unique central region (Coughlan *et al.* 2016; Douglass *et al.* 2018). Each *Ko. phaffii* centromere is ~6 kb long, whereas each *Pi. kudriavzevii* centromere is ~35 kb. In *O. polymorpha*, centromeric regions contain no large IRs, but instead contain clusters of a Ty5-like retrotransposon and its long terminal repeats in regions of ~10 kb that are devoid of other genes (Ravin *et al.* 2013; Hanson, Byrne and Wolfe 2014). This retrotransposon is found only near centromeres, but the exact position of the functional centromere within the retrotransposon cluster is not known. In *Ku. capsulata*, centromere locations were mapped by chromosome conformation capture (3C) experiments (Morales *et al.* 2013; Marie-Nelly *et al.* 2014). The *Ku. capsulata* centromeres do not contain IRs or retrotransposons, but five of the seven chromosomes contain a conserved sequence motif of ~200 bp (Morales *et al.* 2013), which we refer to as a point centromere, although it does not contain the CDE I-II-III elements characterised in *S. cerevisiae*. The YGOB Ancestor is also inferred to have had point centromeres, because all its descendants have point centromeres (Gordon, Byrne and Wolfe 2011).

We found that some genes had remained centromere-proximal over long evolutionary periods during methylotroph evolution. To search for such genes, we first calculated the distance of each gene from its centromere, for every gene in an MGOB pillar, for all species with known centromere locations. For each pair of species, we then sorted the distances to find the pillars with the lowest average distance in the two species. To test whether these putatively conserved centromere-proximal genes exist to a greater extent than expected by chance, we compared the observed data to 1000 randomisations in which pillar content was shuffled, for each species pair (see MATERIALS AND METHODS). The results (Fig. 5) show that, in all six possible pairs of species, there are more conserved centromere-proximal pillars than expected by chance. For short distances from the centromere, there is an excess of pillars in the observed

data (red curves), compared to the null expectation (blue curves). The excess extends out to a distance of at least 150 genes from the centromere for every species pair (Fig. 5). Every species pair showed a highly significant difference between distributions (Kolmogorov–Smirnov tests, $P < 1e{-}8$ in each pair). Thus, more genes are close to centromeres in multiple species than expected by chance. The most likely explanation for this pattern is that the linkage between the genes and the centromeres is an ancient feature of the genomes, inherited from their common ancestor, and relatively undisturbed by genomic rearrangements. In turn, this explanation implies that the centromeres of different species are orthologous, or at least that the centromere locations are orthologous. We are not suggesting that there is any connection between the function of the genes and the function of the centromere.

To identify putative ancient centromere locations, we identified all genes that are ≤10 genes away from a centromere in at least two species. There are 23 such genes (Fig. 6). They define synteny relationships between the centromeric regions of all five *Pi. kudriavzevii* chromosomes and centromeric regions in other methylotrophs (Fig. 6A–C and E) and/or the YGOB Ancestor (Fig. 6B and D). The largest conserved blocks are a five-gene block shared by *Pi. kudriavzevii CEN5* and *O. polymorpha CEN7*, and a four-gene block shared by *Pi. kudriavzevii CEN4* and the YGOB Ancestor (Fig. 6D and E). As well as involving all five *Pi. kudriavzevii* centromeres, the synteny relationships involve six of the seven *O. polymorpha* centromeres, four of the seven *Ku. capsulata* centromeres, one of the four *Ko. phaffii* centromeres and two of the eight YGOB Ancestral centromeres (Fig. 6). In every case, the synteny relationship between a pair of chromosomes does not span the centromere itself and is present on only one side, suggesting that centromeres may have been frequent sites of chromosomal breakage during evolution.

Although the regions of synteny near centromeres shown in Fig. 6 are short, the result in Fig. 5 indicates that a significant level of interspecies synteny conservation extends into much larger regions around the centromeres. Together, these results show that the approximate locations of some centromeres have been conserved among multiple distantly related Pichiaceae species and have therefore been reasonably stable for up to 200 Myr. Moreover, some centromere locations are conserved between Pichiaceae and Saccharomycetaceae (i.e. the YGOB Ancestor). The syntenic regions involve similarities of gene content between centromeres with different structures—IRs, Ty5-like clusters and point centromeres (Fig. 6). Therefore, during Pichiaceae evolution, centromeres seem to have been able to change their structures without changing their locations. Given the limited extent of synteny conservation, we cannot tell whether centromeres with new structures were formed at exactly the same sites as previous old centromeres, or just close to them. A similar situation occurs in the genus *Naumovozyma*, which transitioned from one type of point centromere to another without making major changes in centromere location (Kobayashi *et al.* 2015). What the ancestral structure of centromeres was in Pichiaceae and why so much upheaval of centromere structure occurred in budding yeasts remain unanswered questions.

CONCLUSIONS

MGOB provides a resource for exploring gene orthology and synteny relationships among Pichiaceae species. This yeast family shows significant differences from the model organism *S. cerevisiae* in many aspects of biology, including centromere structures, control of mating types and even the genetic code in some species. It is likely that further insights into the evolutionary history of yeasts will be gained using MGOB.

SUPPLEMENTARY DATA

Supplementary data are available at FEMSYR online.

FUNDING

This work was supported by the Wellcome Trust (105341/Z/14/Z) and Science Foundation Ireland (13/IA/1910).

Conflict of interest. None declared.

REFERENCES

Ávila J, González C, Brito N *et al.* A second Zn(II)(2)Cys(6) transcriptional factor encoded by the *YNA2* gene is indispensable for the transcriptional activation of the genes involved in nitrate assimilation in the yeast *Hansenula polymorpha*. *Yeast* 2002;**19**:537–44.

Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res* 2005;**15**:1456–61.

Byrne KP, Wolfe KH. Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic Acids Res* 2006;**34**:D452–455.

Coughlan AY, Hanson SJ, Byrne KP *et al.* Centromeres of the yeast *Komagataella phaffii (Pichia pastoris)* have a simple inverted-repeat structure. *Genome Biol Evol* 2016;**8**:2482–92.

Douglass AP, Offei B, Braun-Galleani S *et al.* Population genomics shows no distinction between pathogenic *Candida krus e* i and environmental *Pichia kudriavzevii*: One species, four names. *PLoS Pathog* 2018;**14**:e1007138.

Dujon B, Sherman D, Fischer G *et al.* Genome evolution in yeasts. *Nature* 2004;**430**:35–44.

Engel SR, Dietrich FS, Fisk DG *et al.* The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)* 2014;**4**:389–98.

Fischer G, Rocha EP, Brunet F *et al.* Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet* 2006;**2**:e32.

Fournier T, Gounot JS, Freel K *et al.* High-quality *de novo* genome assembly of the *Dekkera bruxellensis* yeast using nanopore MinION sequencing. *G3 (Bethesda)* 2017;**7**:3243–50.

Gordon JL, Byrne KP, Wolfe KH. Additions, losses and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet* 2009;**5**:e1000485.

Gordon JL, Byrne KP, Wolfe KH. Mechanisms of chromosome number evolution in yeast. *PLoS Genet* 2011;**7**:e1002190.

Hanson SJ, Byrne KP, Wolfe KH. Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus *Saccharomyces cerevisiae* system. *Proc Natl Acad Sci USA* 2014;**111**:E4851–4858.

Hellborg L, Piskur J. Complex nature of the genome in a wine spoilage yeast, *Dekkera bruxellensis*. *Eukaryot Cell* 2009;**8**:1739–49.

Ito T, Fujimura S, Uchino M *et al.* Distribution, diversity and regulation of alcohol oxidase isozymes, and phylogenetic relationships of methylotrophic yeasts. *Yeast* 2007;**24**:523–32.

Kobayashi N, Suzuki Y, Schoenfeld LW *et al*. Discovery of an unconventional centromere in budding yeast redefines evolution of point centromeres. *Curr Biol* 2015;**25**:2026–33.

Krassowski T, Coughlan AY, Shen XX *et al*. Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. *Nat Commun* 2018;**9**:1887.

Kurtzman CP. Biotechnological strains of *Komagataella (Pichia) pastoris* are *Komagataella phaffii* as determined from multigene sequence analysis. *J Ind Microbiol Biotechnol* 2009;**36**:1435–8.

Kurtzman CP. *Ogataea* Y. Yamada, K. Maeda & Mikata (1994). In: Kurtzman CP , Fell JW, Boekhout T (eds). *The Yeasts, A Taxonomic Study*. Amsterdam: Elsevier Science, 2011a, 645–71.

Kurtzman CP. *Komagataella* Y. Yamada, Matsuda, Maeda & Mikata (1995). In: Kurtzman CP , Fell JW, Boekhout T (eds). *The Yeasts, A Taxonomic Study*. Amsterdam: ElsevierScience, 2011b, 491–5.

Kurtzman CP. *Pichia* E.C. Hansen (1904). In: Kurtzman CP , Fell JW , Boekhout T (eds) . *The Yeasts, A Taxonomic Study*. Amsterdam: Elsevier Science, 2011c, 685–707.

Kurtzman CP, Robnett CJ. Systematics of methanol assimilating yeasts and neighboring taxa from multigene sequence analysis and the proposal of *Peterozyma* gen. nov., a new member of the Saccharomycetales. *FEMS Yeast Res* 2010;**10**:353–61.

Larkin MA, Blackshields G, Brown NP *et al*. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007;**23**:2947–8.

Love KR, Shah KA, Whittaker CA *et al*. Comparative genomics and transcriptomics of *Pichia pastoris*. *BMC Genomics* 2016;**17**:550.

McGowan J, Byrne KP, Fitzpatrick DA. Comparative analysis of oomycete genome evolution using the Oomycete Gene Order Browser (OGOB). *Genome Biol Evol* 2019;**11**:189–206.

Maguire SL, OhEigeartaigh SS, Byrne KP *et al*. Comparative genome analysis and gene finding in *Candida* species using CGOB. *Mol Biol Evol* 2013;**30**:1281–91.

Maleszka R, Skrzypek M. Assignment of cloned genes to electrophoretically separated chromosomes of the yeast *Pachysolen tannophilus*. *FEMS Microbiol Lett* 1990;**57**:79–82.

Marcet-Houben M, Gabaldón T. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biol* 2015;**13**:e1002220.

Marie-Nelly H, Marbouty M, Cournac A *et al*. Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinformatics* 2014;**30**:2105–13.

Mattanovich D, Branduardi P, Dato L *et al*. Recombinant protein production in yeasts. *Methods Mol Biol* 2012;**824**:329–58.

Morales L, Noel B, Porcel B *et al*. Complete DNA sequence of *Kuraishia capsulata* illustrates novel genomic features among budding yeasts (Saccharomycotina). *Genome Biol Evol* 2013;**5**:2524–39.

Mühlhausen S, Findeisen P, Plessmann U *et al*. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. *Genome Res* 2016;**26**:945–55.

Perez MD, Gonzalez C, Avila J *et al*. The *YNT1* gene encoding the nitrate transporter in the yeast *Hansenula polymorpha* is clustered with genes *YNI1* and *YNR1* encoding nitrite reductase and nitrate reductase, and its disruption causes inability to grow in nitrate. *Biochem J* 1997;**321**:397–403.

Ravin NV, Eldarov MA, Kadnikov VV *et al*. Genome sequence and analysis of methylotrophic yeast *Hansenula polymorpha* DL1. *BMC Genomics* 2013;**14**:837.

Riley R, Haridas S, Wolfe KH *et al*. Comparative genomics of biotechnologically important yeasts. *Proc Natl Acad Sci USA* 2016;**113**:9882–7.

Rolland T, Dujon B. Yeasty clocks: dating genomic changes in yeasts. *C R Biol* 2011;**334**:620–8.

Shen XX, Opulente DA, Kominek J *et al*. Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* 2018;**175**:1533–45 e1520.

Silvestrini L, Rossi B, Gallmetzer A *et al*. Interaction of Yna1 and Yna2 is required for nuclear accumulation and transcriptional activation of the nitrate assimilation pathway in the yeast *Hansenula polymorpha*. *PLoS One* 2015;**10**:e0135416.

Sturmberger L, Chappell T, Geier M *et al*. Refined *Pichia pastoris* reference genome sequence. *J Biotechnol* 2016;**235**:121–31.

Vakirlis N, Sarilar V, Drillon G *et al*. Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. *Genome Res* 2016;**26**:918–32.

Yurimoto H, Oku M, Sakai Y. Yeast methylotrophy: metabolism, gene regulation and peroxisome homeostasis. *Int J Microbiol* 2011;**2011**:101298.