Genetic basis of priority effects: insights from nectar yeast

Manpreet K. Dhami1, Thomas Hartwig2 and Tadashi Fukami1

1Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA
2Department of Plant Biology, Carnegie Institution for Science, 260 Panama Street, Stanford, CA 94305, USA

Priority effects, in which the order of species arrival dictates community assembly, can have a major influence on species diversity, but the genetic basis of priority effects remains unknown. Here, we suggest that nitrogen scavenging genes previously considered responsible for starvation avoidance may drive priority effects by causing rapid resource depletion. Using single-molecule sequencing, we de novo assembled the genome of the nectar-colonizing yeast, Metschnikowia reukaufii, across eight scaffolds and complete mitochondrion, with gap-free coverage over gene spaces. We found a high rate of tandem gene duplication in this genome, enriched for nitrogen metabolism and transport. Both high-capacity amino acid importers, GAP1 and PUT4, present as tandem gene arrays, were highly expressed in synthetic nectar and regulated by the availability and quality of amino acids. In experiments with competitive nectar yeast, Candida rancensis, amino acid addition alleviated suppression of C. rancensis by early arrival of M. reukaufii, corroborating that amino acid scavenging may contribute to priority effects. Because niche pre-emption via rapid resource depletion may underlie priority effects in a broad range of microbial, plant and animal communities, nutrient scavenging genes like the ones we considered here may be broadly relevant to understanding priority effects.

1. Introduction

Many processes affect species diversity in ecological communities, but one process that is receiving renewed interest is priority effects, where the order of species arrival determines community assembly [1–3]. In communities of plants (e.g. [4]), animals (e.g. [5]), fungi (e.g. [6]), and microbes (e.g. [7]), species that arrive at newly disturbed or formed habitats often inhibit colonization by late-arriving species via resource pre-emption and/or modification [8]. Growing evidence indicates that these inhibitory priority effects limit species diversity of the local community via competitive exclusion [8]. At the same time, species diversity at the regional scale is enhanced by priority effects causing local communities to diverge in species composition as a result of variable arrival history [9]. To our knowledge, however, no study has explicitly identified genes affecting species traits responsible for resource pre-emption or modification, leaving the genetic mechanisms behind priority effects unknown.

One system where priority effects have been systematically studied is the microbial communities that develop in floral nectar [10–12]. Nectar is a ubiquitous resource not just for pollinating animals, but also for the microbes that are introduced via pollinators [13]. A small number of yeast and bacterial species, which comprise the nectar microbiome, can tolerate the osmotic pressure caused by high sugar concentrations and the nutritional scarcity caused by low amino acid concentrations [13,14]. Recent studies show that these species engage in resource competition, resulting in strong and pervasive priority effects [10,11]. Even complete exclusion by early arriving species is frequently observed, with direct consequences for microbial species diversity within and across flowers and their mediation of plant–pollinator mutualism [10,11,15]. Because population growth in nectar appears limited by the availability of amino acids [10,12], which the
microbes deplete rapidly [8], pre-emption of amino acids by early colonizers is a plausible explanation for priority effects in this system. Moreover, the strength of priority effects can be predicted by the sensitivity of each species’ growth rate to amino acid concentration as well as the amount of overlap in amino acid utilization between species [11], further implicating amino acids as the limiting resource that drives priority effects in this system. However, little is known about the genes involved in resource pre-emption or modification, leaving the molecular mechanisms behind priority effects elusive.

In this study, we focus on Metschnikowia reukaufii (Saccharomycetales: Metschnikowiaceae), a ubiquitous nectar yeast [16–18] that exhibits particularly strong priority effects in competition with other yeasts and bacteria [11,12]. We first present a high-quality draft genome of M. reukaufii and compare its characteristics to other yeast genomes. Minimal genetic data are currently available for nectar yeasts (but see [19]), but other yeasts, including Saccharomyces cerevisiae, Clavispora lusitaniae, and Debaryomyces Hansenii, have long been used as model organisms for molecular genetics. Although not closely related to nectar yeasts, these well-studied genomes provide a basis that can inform the analysis of nectar yeasts. Additionally, we compare the M. reukaufii genome to those of several other species that are not as well characterized but are more closely related to M. reukaufii (see also the electronic supplementary material). These comparative analyses facilitated identification of candidate genes that may determine how M. reukaufii interacts with other inhabitants of the nectar microbiome. To provide supportive evidence for these genes, we also report the results of gene expression assays and interspecific competition experiments. Together, our results suggest that extensive gene duplication that enables efficient amino acid scavenging may underlie the strong priority effects exhibited by M. reukaufii.

2. Results and discussion

(a) Sequencing and assembly

Genomic DNA from M. reukaufii strain MRI was sequenced using high coverage (82×) single-molecule PacBio sequencing, supplemented by Illumina short-read sequencing (35×; electronic supplementary material, figure S1). PacBio sequencing yielded a total of 1.6 billion filtered bases and an average read length of more than 12 kb (electronic supplementary material, figure S2). The initial draft diploid assembly had a total size of 19.2 Mb (8×; figure 1A), an N50 of 1.2 Mb, and high-consensus accuracy of 99.994% (electronic supplementary material, figure S2). The initial genome assembly contained multiple incompletenesses (figure 1A); electronic supplementary material, figure S3), but was fragmented into 128 contigs. PacBio long read sequencing alleviates the assembly of complex and repetitive regions, but high levels of heterozygosity commonly found in wild yeasts [20] pose a challenge [21]. Indeed, not only was 80% of the M. reukaufii genome contained in the 17 largest contigs, but smaller 120 contigs also self-aligned within the remaining eight. To demarcate the haplotypes within this diploid genome, rectify base-level errors, and correct putative misjoins, we used a custom pipeline optimized for highly polymorphic genomes and included organelle assembly (electronic supplementary material, figure S1). Integration of a haplo-aware assembly improved the overall quality of the M. reukaufii reference genome. Resolved into eight scaffolds, the haploid reference genome had a total length of 15,522,805 bp and a tripled N50 of 3.4 Mb (figure 1C; electronic supplementary material, figure S3). At 15.5 Mb, similar to M. bicuspidata, we found evidence of genome expansion in this species, when compared with C. lusitaniae (see the electronic supplementary material, figure S4a). Telomeres are complex repeats and difficult to assemble, but in our case, the majority of the scaffolds contained a telomeric core motif (5’-AAATAATCATCA TCCCT-3’) at one or both ends, illustrating high contiguity.

To estimate the completeness of our assembly, we used the core set of highly conserved eukaryotic genes (CEG) [21]. With 96.4% of the CEG (239/248) present, the M. reukaufii genome ranks similar to other reference yeast genomes, with only some well-curated genomes (e.g. S. cerevisiae, 97.6%) showing slightly higher coverage (figure 1B).

(b) Annotation and alternative codon usage

Maker-predicted 6106 gene models of which 83% were supported by homology evidence from ESTs and/or proteins [22]. The average gene length was 1771 bp. Compared with S. cerevisiae S288c (12.3 Mb, 5404 genes) and C. lusitaniae (12.1 Mb, 5926 genes), M. reukaufii genome has a typical number of predicted protein-coding genes. Gene ontology mapping obtained hits for 5266 genes (86%), with 4689 genes (77%) assigned to a functional category and the remaining 577 genes uncharacterized.

Yeasts that belong to the CUG-Ser clade encode a unique seryl-tRNA<sub>CAC</sub> through which the CUG codon is encoded as serine (Ser) and rarely as the standard leucine (Leu) [23]. Because sister Metschnikowiaceae species, C. lusitaniae and M. bicuspidata, belong to this clade [23], we investigated whether M. reukaufii may encode CUG as Ser instead of Leu. Purine 33 (G<sup>53</sup>) in the C. albicans Ser-tRNA<sub>CAC</sub> anticodon loop, which replaces a conserved pyrimidine (U<sup>53</sup>) found in all other tRNAs, is the main element that lowers the rate of leucylation [24]. Structural analysis of M. reukaufii tRNAs revealed that the predicted CUG-tRNAs carry G instead of U at position 33 as well as the Ser discriminator base G<sup>27</sup>, as observed in C. albicans (electronic supplementary material, figure S4b). However, a third Ser-identity element, G3C3 run in the variable loop, is missing similar to M. bicuspidata. This supports the recently proposed stepwise accumulation of tRNA<sub>CAC</sub>Ser features in the evolution of alternative CUG translation [23]. Alignment of conserved CUG-sites further supports the alternative codon usage in M. reukaufii. Using the Bagheera pipeline, we predicted CUG-usage for 113 sites across 30 conserved proteins [25]. Seventy-two per cent of these sites suggested alternative codon usage (electronic supplementary material, table S3). However, without direct measures such as sequencing of tryptic peptides, the actual rate of leucylation in M. reukaufii remains to be determined.

The initial genome assembly contained multiple incomplete mitochondrial (mt) genomes mis-assembled as forward and reverse repeats, probably due to the large differences in sequence coverage (40× of nuclear DNA) and GC content. Subsequent integration of a bait-mapping approach optimized for organelles assembled the 29,534 bp mitochondrial (figure 1D). Of the 14 core mitochondrial-encoded genes found in most fungi [26], the M. reukaufii mitochondrial harbours 13, including COX1, NAD2, NAD3, NAD4, COX2, NAD6, NAD1,COX3, ATP6, COB, NAD4 L, NAD5, ATP9 (figure 1D; electronic supplementary material, figure S5), with ATP8 absent. These genes, in addition to 23 tRNAs and four
Figure 1. Genome of Metschnikowia reukaufii. (a) Flower of host plant sticky monkeyflower (M. aurantiacus), zoomed inset, scanning electron micrograph of M. reukaufii cells (blue) attached to cells and trichome of sticky monkeyflower floral tube that contains nectar (gold), Scale bar, 5 μm, false-coloured (b) CEGMA completion of M. reukaufii (Mreu, this study), compared to published yeast genomes, from nectar specialists (yellow: S. bombicola (Sbom), H. valbyensis (Hval), Metschnikowiaceae species (green: M. bicuspidata (Mbic), C. lusitaniae (Clus)) and closely related CUG-Ser clade members (blue: C. tenuis (Cten), D. hansenii (Dhan), H. burtonii (Hbur), C. tanzawaensis (Ctan), Sp. passalidarum (Spal) and S. cerevisiae S288c reference genome (Scer). (c) Scaffolds representing nuclear genome: predicted genes (black lines) and predicted repeats (red lines). (d) Circular mitochondrial genome: predicted genes, tRNAs and rRNAs, sequencing coverage, GC content and sequence similarity to C. lusitaniae, D. hansenii and S. cerevisiae mitochondrial genomes.
rRNAs (missing tRNAy and four rRNAs), comprise the mitochondrion (GC content = 33%).

(c) Tandem gene duplication
Genomic analyses of model organisms have shown that over a third of all protein-coding genes belong to multigene families that arise from duplications, e.g. whole-genome duplication, segmental duplication, and tandem gene duplication [27]. Most sequenced hemiascomyceye yeast genomes harbour 1.5–2% of tandem gene arrays (TGAs), with some notable exceptions. Debaromyces Hansenii, a marine yeast that tolerates extreme salt stress, has 4.4% of all genes arranged in TGAs [28]. Metchnikowia reukaufii showed a relatively high number of 363 TGAs, representing 5.9% of all annotated genes. Of these, 227 are novel TGAs compared to C. lusitaniae (electronic supplementary material, table S1). The largest fraction of these TGAs comprised genes involved in nitrogen metabolism, that of either cellular- (48 genes, 21%) or organo-nitrogen (18 genes, 8%, figure 2a). Another enriched fraction represented genes involved in cellular transport (27 genes, 12%, figure 2a). These transporters include homologues of ATP binding cassette and oligopeptide transporters, as well as amino acid permeases (electronic supplementary material, table S1).

(d) Duplication of nitrogen transport and metabolism genes
Two of the tandem duplicated genes, MreuGAP1-1 and MreuGAP1-2A, are of particular interest given that their closest homologue in S. cerevisiae is the general amino acid permease1 (GAP1). In fungi, amino acid uptake is mediated by yeast amino acid transporters (YATs, amino acid-polymamine-organocation superfamily), of which 18 have been functionally characterized in S. cerevisiae [29]. YAT members share a common topology with 12 trans-membrane domains and cytosolic N and C terminals. Most YATs, such as the histidine permease1 (HIP1), have low capacity and are most active when amino acids are abundant in the growth medium [29]. They opportunistically transport specific amino acids for protein synthesis. By contrast, GAP1, a high-capacity transporter of all naturally occurring amino acids and analogues, is active when amino acids are limiting [29]. GAP1, therefore, functions as a scavenger of amino acids for nitrogen when supply is low. In fact, experimental evolution in S. cerevisiae has shown that GAP1 is a recurring locus for adaptation to nitrogen-limited environments [30,31] and additional GAP1 copies confer an average fitness advantage of 24–44% in nitrogen-limited media [31].

MreuGAP1-1 and 2A are predicted to encode 585 amino acids, located on the minus strand of scaffold 3 and 1, respectively (figure 2b). They share 64% and 62% protein identity with ScerGAP1, and 80% and 74% identity to C. lusitaniae’s orthologue (CLUG_00762), respectively. In contrast to ScerGAP1’s single-copy arrangement, M. reukaufii has two GAP1 TGAs. The first contains MreuGAP1-1 and 175 bp downstream a second transporter more closely related to ScerHIP1, with only 56% protein identity between them. Clavispora lusitaniae shares the same anti-parallel TGA, suggesting that this duplication originated before the divergence of these Metchnikoviales species (figure 2b). However, the second GAP1 TGA with three additional GAP1 homologues (MreuGAP1-2A, 2B and 2C) is unique to M. reukaufii (figure 2b). Phylogenetic analysis supports the expansion of GAP1 in M. reukaufii (figure 2c). The fact that we identified numerous long PacBio reads contiguous over the GAP1 TGAs is evidence for the physical linkage of these genes, given the random nature of chimeric artefacts.

All four MreuGAP1 homologues share conserved LysP (COG0833) and AA permease domains (pfam00324) (electronic supplementary material, figure S9). The consensus amphipathic region (CAR domain) of ScerGAP1 forms the amino acid translocation channel with the critical residues N959, S961, S967 and R986 [31]. MreuGAP1-1, 2A to C, but not MreuHIP1, had all of these residues conserved (electronic supplementary material, figure S9). Similarly, the C-terminus, critical for transport and amino acid sensing in ScerGAP1, was highly conserved in MreuGAP1-1 and 2A to C (electronic supplementary material, figure S9). The conservation of all residues critical for amino acid transport in MreuGAP1-1 and 2A to C supports the preservation of function.

In order to ascertain the function of the MreuGAP1 genes, we quantified their expression under conditions that mimic floral nectar. Nitrogen resources are commonly low in sticky monkeyflower nectar, from which the M. reukaufii strain was isolated, with amino acids such as proline and glutamine ranging between 30 and 50 μM [10]. Quantitative RT-PCR showed that all MreuGAP1 homologues were expressed under synthetic nectar environments with low amino acid availability (figure 3a,b).

Excessive uptake of amino acids can be detrimental to yeast growth [32]. To prevent toxic intracellular accumulation, ScerGAP1 is under nitrogen catabolite repression (NCR), which controls its abundance transcriptionally and post-translationally [32], a mechanism not shared by low-capacity permeases such as HIP1. Indeed, the promoter regions of MreuGAP1 homologues harboured UASNCR GLN3/GAT1 [33] activator binding site (GATAAG) upstream of their transcriptional start. Consistent with this regulation, the expression of all four MreuGAP1 homologues was strongly induced (20- to 350-fold) under synthetic nectar, compared with conditions with abundant nitrogen sources (YM) (figure 3c). Because the overall nutrient composition of YM differs from synthetic nectar, we evaluated the regulation of MreuGAP1 homologues in nectar with varying levels of high- and low-quality nitrogen sources, with MreuHIP1 used as a reference. As predicted, a good nitrogen source, glutamine, suppressed the expression of all MreuGAP1 homologues, with GAP1-2A inhibited by 97.5% at 10-fold (400 μM) of the typical standing concentration found in sticky monkeyflower nectar (approx. 40 μM) [10]. By contrast, the mRNA abundance of MreuHIP1 increased by 259% under the same conditions (figure 3b). At high concentrations, both proline (poor nitrogen source) and urea (non-amino acid nitrogen source) repressed MreuGAP1-1 and 2A-C transcription, but the relative repression of each homologue was lower compared with glutamine. With proline as sole nitrogen source, even at 200 times (8 mM) the steady-state nectar concentration, GAP1-2A expression was reduced by only 30.5% (figure 3b). MreuHIP1 expression, on the other hand, increased twofold (figure 3b).

In S. cerevisiae, a second amino acid scavenger active under nitrogen limitation is the high-capacity proline utilization 4 permease (PUT4). PUT4 and GAP1 are the main importers of the imino acid proline [34,35]. Although less preferred than glutamine, proline is one of the most abundant nitrogen sources in many natural environments of yeasts, including sticky monkeyflower nectar [10,36]. The genome of M. reukaufii features three parallel duplicated PUT4 homologues on the minus strand of scaffold 6 (figure 2b). They
are predicted to encode 567 (MreuPUT4A), 565 (MreuPUT4B) and 566 (MreuPUT4C) amino acid proteins, approximately 3800 bp and 1450 bp apart, respectively (figure 2b). Multiple PacBio long reads contiguous over the three PUT4 homologues support the physical linkage of these genes. The qRT-PCR showed that all three MreuPUT4 genes were expressed under synthetic nectar conditions (electronic supplementary material, figure S8). Well characterized in S. cerevisiae and Aspergillus nidulans, the amino acid residues critical for transport and substrate-specificity of PUT4 have been experimentally verified [37]. MreuPUT4A to C not only share all conserved residues essential for transport, but also those unique to PUT4 members of the YAT clade (electronic supplementary material, figure S9). Oligopeptide transporters were also tandem duplicated in the M. reukaufii genome, e.g. five copies of an OPT2 homologue (electronic supplementary material, figure S9).
Figure 3. Amino acid scavenging genes for priority effects caused by nitrogen pre-emption. (a) Expression of four GAP1 homologues of *M. reukaufii* cells growing for 4 h in either synthetic nectar (20% sucrose) with 0.04 mM proline or glutamine as nitrogen source or YM (approx. 400 mM mixed amino acids) compared to HIP1. All GAP1 homologues are expressed in nectar, but repressed in YM media. (b) NCR of GAP1 genes in *M. reukaufii* cells growing for 4 h in 20% sucrose supplemented with 0.04, 0.4, 4 or 8 mM of the poor (proline), good (glutamine) and non-amino acid (urea) nitrogen source. Expression of GAP1 homologues, but not HIP1, is strongly repressed by glutamine, and to a much lower extent by proline and urea. Data show mean ± s.d. relative to 0.04 mM concentrations; n = 3 biological replica each with two technical replica. (c) (left panel) Growth of the 2-day late-arriving *C. rancensis* was significantly repressed in the presence of early arriving *M. reukaufii* when either no nutrients were restocked ($P_{adj} = 5 \times 10^{-5}$) or only sucrose was restocked ($P_{adj} = 15 \times 10^{-6}$) (10 mg ml$^{-1}$ per day after *C. rancensis* arrival). This effect was reversed when 40 mM amino acids were supplied daily post *C. rancensis* arrival ($P_{adj} = 0.646$). No observed significant difference between their population densities when grown alone ($P_{adj} = 0.221$) (right panel). Data show mean ± s.e.
supplementary material, table S1). Similar to GAP1 and PUT4, OPT2 has been shown critical for growth under nitrogen starvation [38].

More generally, we found that genes involved in nitrogen metabolism were highly enriched (21% and 8%, figure 2a) among TGAs identified in the M. reukaufii genome. One example is MreuAOC1A and B, which are homologous to the Cu²⁺-containing amine oxidase of Ogataea polymorpha (OpolAOC1) [39]. MreuAOC1A and B encode 609 and 620 amino acid proteins located 1729 bp apart on the minus strand of scaffold 1 (figure 2d) and share 78% and 77% protein identity to C. lusitaniae’s homologue (CLUG_00771), respectively. OpolAOC1 has been studied thoroughly and its overexpression is linked to increased cell growth [39]. The complete list of M. reukaufii TGAs contains many more examples of genes with predicted functions in amino acid biosynthesis or catabolism (electronic supplementary material, table S1), indicating that these families may have undergone adaptive expansion.

(e) Nitrogen scavenging as a mechanism of priority effects

Based on our findings, we hypothesize that rapid depletion of amino acids promoted by gene duplications is a key mechanism of the priority effects observed in M. reukaufii. If M. reukaufii colonizes nectar prior to other yeasts, it should severely limit amino acid availability for late-arriving species.

To begin to test this hypothesis, we conducted competition experiments against one of the closest known relatives of M. reukaufii, Candida rancensis (Metschnikowia-cae) [40], commonly found in monkeyflower nectar [41]. C. rancensis is one of the strongest competitors of M. reukaufii, likely due to resource-use overlap [10], but it is severely disadvantaged when M. reukaufii colonizes the nectar first [10]. Using established methods to quantify priority effects [10,11] and to simulate realistic nectar microbiome interactions, we inoculated M. reukaufii into synthetic nectar 2 days before C. rancensis. Consistent with previous studies [10–12], M. reukaufii suppressed the growth of C. rancensis when it was given a head start (figure 3c; \( P_{adj} = 5 \times 10^{-7} \)). This effect remained even when 10 mg ml⁻¹ of sucrose was replenished in the nectar medium every 24 h (figure 3c; \( P_{adj} = 1.5 \times 10^{-7} \)). By contrast, when 40 mM of amino acids were supplied every 24 h, the negative effect of early M. reukaufii arrival was rescinded and the two species coexisted at similar population densities (figure 3c; \( P_{adj} = 0.646 \)). These results support the hypothesis of amino acid limitation as a cause of priority effects.

To test the hypothesis unequivocally, competition experiments using GAP1 loss of function mutants would be particularly powerful. In addition, genomic and competition analyses with other species of nectar-inhabiting yeasts should be informative because the strength of priority effects varies among these species [10–12]. Given that several nectar-inhabiting yeasts that exert priority effects with M. reukaufii are similar to M. reukaufii both phylogenetically [10] and ecologically [11], GAP1 duplication may well be a mechanism shared by many of the nectar yeast species in the Metschnikowia-cae family. One hypothesis that we believe is worth testing in future research is whether the GAP1 copy number varies among species and correlates with the strength of priority effects.

3. Conclusion

Our newly assembled M. reukaufii genome allowed us to identify candidate genes underlying the priority effects observed in nectar. In particular, we have highlighted the hypothesis that extensive genome expansion, especially in high-capacity amino acid transporter genes such as GAP1 and PUT4, allows M. reukaufii to exert strong priority effects against other nectar microbes. Further investigation will be required to unequivocally test this hypothesis. Nonetheless, the findings presented here lay a molecular foundation on which to build a better understanding of how species assemble into ecological communities. Niche pre-emption via rapid resource depletion may underlie priority effects in a broad range of microbial, plant and animal communities [8,42]. For this reason, nutrient scavenging genes like the ones considered in the yeast here may be broadly relevant to understanding priority effects.

4. Material and methods

(a) Strain

Metschnikowia reukaufii, strain MR1, was isolated from floral nectar of the sticky monkeyflower (Mimulus aurantiacus) at the Jasper Ridge Biological Preserve, Stanford, California, USA (37°24’29”N and 122°13’39”W) as previously described [10]. Overnight cultures (at 25°C) from single colonies grown in yeast media (YM) were used for DNA extraction process.

(b) DNA extraction, sequencing and genome assembly

PacBio: gDNA was extracted from 7 \( \times \) 10⁶ cells using the Genomic-tips 100/G platform (Qiagen), optimized for yeast cells. The gDNA was examined for bacterial contamination via PCR amplification of the 16S rRNA gene using universal primers (electronic supplementary material, table S2), but no amplicons were returned. SMRTbell libraries were obtained using the ’Procedure and Checklist – 20 kb Template Preparation using BluePippin™ Size Selection’ protocol at the UC Davis Sequencing Core. The p5 sequencing polymerase and Magbeads (Pacific Biosciences, Menlo Park, CA, USA) were used to bind the SMRTbell templates annealed to sequencing primers. Two SMRT Cells were run on the PacBio RS II system using P6C4 chemistry, with an on-plate concentration of 150 pM and a 240-min data collection mode.

Illumina: gDNA was prepared from 10⁶ cells using the DNeasy Blood & Tissue Kit (Qiagen). Similar to PacBio gDNA assessment, 16S rRNA PCR was used to verify the absence of bacterial contamination. Genome library was prepared using the Illumina Sequencing Library Preparation protocol as described previously [43]. The library was run on a single lane of an Illumina HiSeq 2500 sequencer with 101 bp paired-end mode.

(i) Genome assembly

The combined read maps of the two SMRT cells were filtered and trimmed to remove low-quality sequences. We ranked different PacBio de novo assemblies generated by SMRTAnalysis (v. 1.2, patch 5), Celera (v. 8.3 RC2) and DipSPAdes (v. 3.6.2), based on their N50 value, number of contigs, average contig size and mapping quality scores against approximately 35 × coverage MR1 Illumina reads. The draft assembly generated by SMRTAnalysis outperformed all other assemblies and was therefore chosen for the assembly pipeline (electronic supplementary material, figure S2). The SMRTAnalysis whitelisting protocol was used to exclude possible bacteria-contaminated reads from the filtered sub-reads pool. The initial MR1_a1 draft consisted of 128 contigs,
tools and improve gene model quality. The consensus mitochondrion calls were generated using SNAPhmm, AUGUSTUS and Exonerate. Predicted complete open reading frames were retained. Gene models were invoked by MAKER for gene prediction. Only genes with exons longer than 300 bases were used. In addition, AUGUSTUS (v. 3.0.2) and GeneMarkHMM (v. 4.21) were used to subsequently train SNAPhmm (release 2013-11-29).

CEGMA (v. 2.5) was used in conjunction with geneID (v. 1.4), the RAST (v. 3.2) and the Prochloro [22] genome annotation pipelines. Masker (v. 4.0.6) [22] using the latest RepBase repeat libraries were also used [23,44,46]. Within the MAKER pipeline, RepeatMasker and RepeatFinder were employed for repetitive element detection. The integrative pipeline presented here makes use of three successive tiers, diploid draft, haplotype reconstruction and independent organelle assembly. It allows de novo assembly efforts, especially for organisms with high rates of heterozygosity, to near-complete genomes with minimal need for manual curation, ideal for downstream applications such as automated gene annotation/ontology, phylogenetics and resequencing/variant discovery.

(ii) Mitochondrion assembly
The haploid assembly contained a partial mitochondrial sequence but showed assembly errors with direct and inverted repeats assembled into three separate contigs of 29, 48 and 128 kb length. To assemble the MR1 mitochondrion, we made use of a custom pipeline (electronic supplementary material, figure S1). The partial mitochondrial sequence from the haploid assembly contained the conserved COX1 sequence, which was used for a bait-mapping approach with the Illumina reads. The generated sequence was used with MITOBim (v. 1.8) and MIRA4 for further mapping for 31 iterations to construct the mitochondrial sequence. Additionally, we used the mitochondrial version of C. hirsutae as reference for bait-mapping with MIRA4 to construct the draft MR1 mitochondrial sequence, followed by 21 iterations of MITOBim/MIRA4 to improve the final MR1 mitochondrial sequence. The consensus sequence was further error-corrected with Pilon.

The integrative pipeline presented here makes use of three successive tiers, diploid draft, haplotype reconstruction and independent organelle assembly. It allows de novo assembly efforts, especially for organisms with high rates of heterozygosity, to near-complete genomes with minimal need for manual curation, ideal for downstream applications such as automated gene annotation/ontology, phylogenetics and resequencing/variant discovery.

(c) Genome annotation, gene ontology and genetic features
We used the ab initio gene prediction pipeline MAKER (v. 2.3.18) to annotate the genome [44]. EST evidence from M. fructicola and M. bicuspidata was used to train MAKER [23,45]. Protein evidence from M. bicuspidata, D. hansenii, Candida tenuis and C. hirsutae were also used [23,44,46]. Within the MAKER pipeline, RepeatMasker (v. 4.0.6) [22] using the latest RepeatBase repeat libraries was employed for soft masking. To refine initial gene models, CEGMA (v. 2.5) was used in conjunction with geneID (v. 1.4), genewise (v. 2.2.3-rnc7) and HMMER (v. 3.1b2) to identify highly conserved eukaryotic 'core' genes. The identified core genes were used to subsequently train SNAPHmm (release 2013-11-29).

In addition, AUGUSTUS (v. 3.0.2) and GeneMarkHMM (v. 4.21) were invoked by MAKER for gene prediction. Only genes with predicted complete open reading frames were retained. Gene calls were generated using SNAPHmm, AUGUSTUS and Exonerate (v. 2.2.0), using the evidence sets detailed above. Six consecutive MAKER runs were used to further train prediction tools and improve gene model quality. The consensus mitochondrial sequence was annotated using the MITOS web portal. We selected a range of species in a closely related sister clade (CUG-Ser) for which published genomes were available, other ascomycetous yeast species, and the manually created reference genome of S. cerevisiae in order to evaluate the assembly completeness of M. reukaufii genome, using CEGMA genes as reference.

Blast2GO (v. 2.7.2) [47] was used with BLASTP searches against the NCBI fungi dataset, filtered using Blast2GO annotation algorithm, and GO, and enzyme code were annotated using the GO database. Interproscan results were imported into Blast2GO and merged with GO annotations. Annotation statistics were produced using Eval (v. 2.2.8) and Geneious (v. 8.0.2). Repeat identification and annotation was generated using the REPET pipeline [48].

Whole genomes of M. bicuspidata NRRL YB-4993 [23] and C. hirsutae ATCC 42720 [44] were compared against the genome of M. reukaufii MR1-a14. Pairwise alignments were performed for coding sequences of predicted gene models using adaptive seeds [49]. The SynFind synteny search pipeline identified syntenic blocks by chaining the hits from large-scale alignment tool (LAST) with a distance cut-off of 20 genes apart, and with at least four gene pairs per syntenic block [50]. The syntenic blocks were screened further using QUOTAALIGN to retain one-to-one blocks and remove weak blocks [50]. The outputs were visually inspected to confirm the structural similarity of the M. reukaufii genome to other genomes. Resulting syntenic gene blocks were used to identify orthologies between the genome pairs. Duplicated genes (TGAs) were mapped with Blast2GO to classify their biological function.

Manually selected tandem duplicated genes such as GAPI and PUT4 were analysed further. Pairwise alignments of amino acid sequences from previously discussed species with those of M. reukaufii genes of interest were performed in Geneious using the Blosum62 cost matrix with free end gaps, followed by manual curation. Maximum parsimony and likelihood trees were calculated using PAUP* (heuristic search with TBR branch swapping and 100 bootstraps) [51] and RaxML 7.2.8 (GAMMA BLOSUM62 protein model, 100 bootstraps) [52].

(d) RNA isolation and qRT-PCR
Total RNA was isolated from M. reukaufii as described [53], with the following modifications: post supernatant removal, cell pellet was resuspended in 100 μl water, immediately frozen in liquid N and stored at −80 °C. Frozen cells were re-suspended in 1.5 ml 50:50 (v/v) RNA buffer and phenol/chlorophorm/isomyl alcohol (PCI, 25:24:1 (v/v/v)), mixed and incubated at 65°C for 1 h, mixing every 15 min. After centrifugation, supernatant was washed twice with PCI and incubated with one volume 8 M lithium chloride for 2.5 h (−20°C) for precipitation, followed by two washes in cold 80% ethanol. Final pellets were re-suspended in 100 μl TE buffer. For qRT-PCR, RNA was pre-treated with TURBO DNA-free DNase I (ThermoFisher), and 2.5 μg RNA/sample was synthesized to cDNA by reverse transcriptase (Maxima First Strand Kit, ThermoFisher). The TAF10 gene (ScerTAF10, Mreu_scf5_1.979) was used as internal control. qRT-PCR was performed with SensiMix SYBR & Fluorescein Kit (Bioline) in a LightCycler 480 Instrument II (Roche). Developed qRT-PCRs were tested for efficiency using a dilution series of target cDNA and their amplification efficiencies ranged between 90–105% (R2 > 0.95) and single melting peak observed for each amplicon (electronic supplementary material, figure S10 and table S2). All qRT-PCR reactions included three biological replicates and two technical replicates each.

(e) Competition experiments
Sterile synthetic nectar was prepared as previously described [11], with 4 mM amino acids. On day 0, 200 M. reukaufii cells were introduced into sterile microcosms containing 9 μl of nectar. To induce priority effects, 200 cells of Candida rancensis were introduced to half of the microcosms after 48 h. Thereafter, nutrient supplements of either 40 mM amino acids, 10 mg ml−1 sucrose or water were provided every 24 h. Monocultures of M. reukaufii and C. rancensis in conditions described above were maintained as positive controls, as well as negative controls (no yeast). The experiment was independently replicated twice, consisting of six biological replicates/treatment. At the end of day 4, 0.5 μl nectar from each treatment microcosm was plated onto YM agar and
References

30. Gresham D, Usaite R, Germann SM, Lisby M, Botsstein D, Regenberg B. 2010 Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the


