

Comparative genomic analysis reveals metabolic mechanisms for *Kluyveromyces marxianus*' fast growth during evolution

Wenjuan Mo

State Key Laboratory of Genetic Engineering

Haiyan Ren

State Key Laboratory of Genetic Engineering

Xianmei Yang

State Key Laboratory of Genetic Engineering

Tongyu Luo

State Key Laboratory of Genetic Engineering

Wanlin Lu

State Key Laboratory of Genetic Engineering

Jungang Zhou

State Key Laboratory of Genetic Engineering

Yao Yu

State Key Laboratory of Genetic Engineering

Ji Qi

State Key Laboratory of Genetic Engineering

Hong Lv (✉ honglv@fudan.edu.cn)

State Key Laboratory of Genetic Engineering

Research

Keywords: *Kluyveromyces marxianus*, Fast growth, Comparative genomic analysis, RNA-seq analysis, ATP production, Respiratory chain

Posted Date: December 16th, 2019

DOI: <https://doi.org/10.21203/rs.2.18764/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Using yeast fermentation to produce bioethanol, is an economic and renewable way to tackle the rapid increase in fuel consumption. Faster cell growth rate guarantees the superior result of fermentation course. The “non-conventional” yeast *Kluyveromyces marxianus* is the known fastest-growing eukaryote on the earth. Although its wide application in industry, the molecular mechanisms for its fast growth have seldom been discovered.

Results

We first carried out a comparative genome content analysis for *K. marxianus* evolution in Saccharomycetaceae and identified the gain/lost genes as well as highly copied genes during *K. marxianus* speciation. Then RNA-seq analyses for *K. marxianus* and *S. cerevisiae* at different time points along cultivation were performed, to infer the function of *K. marxianus*-specific genes and to find out the difference in homologous gene expression patterns between the two species. RNA-seq results were further validated with RT-qPCR analysis. Genome content analysis shows the highly intense events of genes' gain/loss happened at the occurrence of Saccharomycetaceae and *Kluyveromyces*, and *K. marxianus* has particularly high copy numbers of genes participating in iron transport and biotin biosynthesis. The RNA-seq analysis reveals 40% of *K. marxianus*-specific genes were up-regulated and may participate in glucose transport and mitochondrial function. Furthermore, compared to *S. cerevisiae*, *K. marxianus* has developed two features in homologous gene expressions to ensure its fast growth: (1) enhanced fundamental expressions of TCA cycle and respiratory chain genes at the beginning of cultivation; (2) tightly co-up-regulated expressions of respiratory chain, F₀F₁ ATPase, and glucose transporter genes at its fastest growth phase. Those co-expressions are mainly ascribed to the higher number of significant motifs in the upstream sequences of involved genes in *K. marxianus* than in *S. cerevisiae*, indicating the importance of transcriptional network remodelling during evolution.

Conclusions

This study gives insights into the possible mechanisms of *K. marxianus*' fast growth trait, via efficiently supplying sufficient energy for cell growth, meanwhile reinforcing glucose transport to guarantee the competition for environmental resources. Our findings present a theoretical support for *K. marxianus*' prospective application in industry, and give clues for further rational construction of fast-growing yeast strains.

Background

Within the last decade, *K. marxianus* as one of 'non-conventional' yeasts, has the traits of food safety, utilization of multiple carbon sources, thermotolerance, and fast growth [1,2]. It has been successfully applied in ethanol fermentation [3,4] and heterologous protein production [5]. The faster growth rate

guarantees the superior result of fermentation course [6]. *K. marxianus* is the fastest-growing eukaryote known so far [7]. The fastest growth rate for *K. marxianus* is 0.80 h^{-1} [7], for *K. lactis* is 0.50 h^{-1} [1], for *Saccharomyces cerevisiae* is 0.37 h^{-1} [1], and for *Pichia pastoris* is 0.18 h^{-1} [8]. Though the famous recognition of *K. marxianus*' fast growth, relatively little is known about the underlying mechanisms that carry out this phenotype.

Based on the literatures to our knowledge, through the pH-auxostat cultivation passage of *K. marxianus*, it elevates its own growth rate by enlarging cellular surface area with activated membrane processes [7]. However, how *K. marxianus* ensures its fast growth compared to other yeast species, especially in the aspects of molecular genetics and intracellular metabolic flow regulation, has not been reported so far.

Currently, five *K. marxianus* strains have been completely sequenced, including KCTC 17555 [9], DMB1 [10], CCT 7735 [11], DMKU 3-1042 [12], and NBRC1777 [13]. According to those genome data, *K. marxianus* has only a few specific genes with totally unknown function, which are not found in other species. The majority of *K. marxianus* genes have homologies in other species (e.g. *S. cerevisiae*) with reported functions. What are the possible functions of *K. marxianus*-specific genes, and how do those specific genes and homologous genes contribute to *K. marxianus*' fast growth trait? These questions should be further explored.

Transcriptome studies for *K. marxianus* have focused on its physiological properties, such as xylose catabolism [12,14], high temperature resistance [12], and ethanol tolerance [3]. These current studies mainly analyzed *K. marxianus*' gene expression changes in different environments. However, the investigation in parallel for homologous gene expression pattern comparison between *K. marxianus* and other yeast species has been little reported. There is a challenge to carry out this, since the homologous gene matches between different species are often many-to-many. If use one-to-one match, many important information would be missed. Thus we use the concept of 'gene family' to approach this problem, which will be described in the following part detailedly.

In this study, we carried out a complete genome sequencing and annotation of the *K. marxianus* FIM1 strain used in our laboratory. Then comparative genomic analysis for *K. marxianus* evolution in Saccharomycetaceae was performed, to find out the gained/lost genes during *K. marxianus* speciation and its particularly highly copied genes. Furthermore, *S. cerevisiae* was chosen as the control strain and RNA-seq analyses of *K. marxianus* and *S. cerevisiae* at different time points along cultivation were performed. Then we compared homologous gene expression patterns as well as metabolic pathways between those two species, to scope out *K. marxianus*'s unique regulation patterns. Combining the analysis of genome content evolution and homologous gene expression regulation, we hope the mystery of *K. marxianus*' fast growth could be revealed.

Results

Genome evolution analysis of *K. marxianus* explores the genetic basis of its rapid growth

We performed a genome sequencing of the *K. marxianus* FIM1 strain used in this study, and 8 nuclear chromosome sequences were obtained with total size of 10.9 Mbp. Then genome was annotated via “Yeast Genome Annotation Pipeline” [15] for gene identification and via non-redundant (nr) database for functional annotation, and finally the GFF3 file was obtained (NCBI Genbank accession numbers CP015054 to CP015061).

To observe the gained/lost genes during *K. marxianus* speciation in Saccharomycetales, a polygenetic tree was first reconstructed (Fig. 1), which contains 12 species in Saccharomycetaceae, 1 in Phaffomycetaceae, and 1 in Dipodascaceae as outgroup. Herein the concept of ‘gene family’ was introduced (see Methods for more detailed derivation), that each gene family is a set of homologous genes among those 15 yeast species. There were 636 gene families containing a single copy of gene in each species. We concatenated those single-copy genes in each species into a super gene, and reconstructed a phylogenetic tree (Fig. 1) using the maximum likelihood method. The derived phylogenetic tree (Fig. 1) was consistent with the reported tree which contains *K. marxianus* [16]. Then the number of gained or lost gene families at each node (Fig. 1) was calculated. The high-intense events of genes gain/lost when Saccharomycetaceae occurrence (node 3) and Kluyveromyces occurrence (node 6) indicate the formation of evolutionary family and genus may be substantially attributed to a great change in gene contents.

For *K. marxianus*, the genes gained at each node during its speciation were analyzed in terms of functional category (Fig. 2). As shown in Fig.2, at nodes 2 and 3, introduced genes were mainly involved in transcription, cell cycle regulation, mitochondrial morphology, carbon metabolism, etc, suggesting the ancestor of *K. marxianus* endowed with common physiological characteristics of Saccharomycetaceae. At nodes 4 ~ 7, it was majorly introduced with genes in DNA synthesis, chromosome segregation, ATP production, glucose transport, and lactose metabolism, implying that along with the emergence of Kluyveromyces genus, the cell cycle apparatus and ATP synthesis had experienced further refinement which may contribute to growth acceleration, meanwhile the resource catabolism may be adjusted to Kluyveromyces-characterized environments, such as milk. At the last node, genes involved in vacuole function and pre-mRNA processing were introduced. However, there are still a number of *K. marxianus*-specific genes with totally unknown function, their contribution to growth process may be partly reflected by RNA-seq analysis.

We further compared the copy number of genes in different yeast species (Fig. 3A). It was found that compared to other species, genes involved in flocculation, iron transport, and biotin biosynthesis have particularly high copies in *K. marxianus*, i.e. 11, 13, and 3, respectively, which may contribute to its fast growth trait.

RNA-seq analysis revealed that 40% of *K. marxianus*-specific genes were up-regulated and may participate in glucose transport and mitochondrial related function

To analyze the difference of homologous genes’ expression pattern between different species in the same cultivation condition, we chose the type strain *S. cerevisiae* as reference and cultivated *K.*

marxianus and *S. cerevisiae* in YPD medium at 30°C. Then RNA-seq analysis was performed at different time points (1h, 4h, 6h, 12h, 24h, 48h, and 72h) during their cultivation, respectively, based on their growth curves (Additional file 1: Figure S1). Additional file 2 provides expression FPKM values of all genes. The gfold algorithm was conducted to measure gene family's differential expression (see Methods).

For those gene families in Saccharomycetale, there are 55 gene families only containing *K. marxianus* genes, which are called KMS (*K. marxianus*-specific) genes herein. As shown by the RNA-seq analysis (Fig. 3B), using 1h as control, 28 KMS gene families had significant expression changes during cultivation, i.e. with 2 fold or more changes at some time point after 1h. Furthermore, 23 gene families were up-regulated (Fig. 3B) and occupied approximately 40% of the total KMS genes. To explore their possible functions, Pearson correlation coefficient (denoted as r) was calculated between KMS gene and gene with known function, based on the hypothesis that strongly co-expressed genes are likely to function in the same or closely related biological pathways [17,18]. Genes with the highest r value (at least $r > 0.8$) were used to infer the possible function of KMS genes. The predicted function of KMS genes and the corresponding r value were listed in Fig. 2B. It was found that the up-regulated KMS genes may mainly participate in glycogen metabolism, glucose transport, and mitochondrial related function, while the down-regulated KMS genes may be involved in ergosterol, leucine, and purine biosynthesis, which give a clue for the fast growth from the perspective of KMS genes.

Mitochondrial function related genes and highly copied genes were particularly up-regulated in *K. marxianus* compared to in *S. cerevisiae*

To analyze the difference of homologous gene expressions between *K. marxianus* and *S. cerevisiae*, gene family was preferably used, which contains a total set of homologous genes and thus is suitable for the many-to-many correspondence of homologous genes between two species. The expression value of a gene family in a species was the sum of the contained genes' expressions. There were 3653 gene families contain both *K. marxianus* and *S. cerevisiae* genes, in which 2535 gene families were identified as differentially expressed in *K. marxianus* or *S. cerevisiae*, using respective 1h as control. Those gene families were then clustered into 12 groups according to their expression patterns (Fig. 4, denoted as C1, C2, ..., C12), using k-means method calculated on Euclidean distance. The significantly enriched GO terms (p value < 0.05) for each cluster were listed in Table 1 (more details are provided in Additional file 3).

As illustrated in Fig. 4, at the early stage (i.e. time points 4h and 6h), the numbers of differentially expressed genes in *K. marxianus* are much higher than those in *S. cerevisiae*. Furthermore, *K. marxianus* and *S. cerevisiae* showed dramatic expression change since 12h and 24h, respectively, which are in line with the time point of OD₆₀₀'s quick increment in the growth curves (Additional file 1: Figure S1).

For the commonality of gene expression pattern between *K. marxianus* and *S. cerevisiae*, most genes involved in ribosome biogenesis and protein translation were both evidently down-regulated during cultivation (C1 ~ C3 in Fig. 4), and genes participating in trehalose biosynthesis, TCA cycle, respiratory chain, and autophagy were both heavily up-regulated (C9 ~ C11 in Fig. 4). For the dissimilarity of gene

expression between the two species, genes for the intracellular signal transduction (e.g. MAPK cascade) were unchanged in *K. marxianus* and up-regulated in *S. cerevisiae* (C12 in Fig. 4). Genes involved in DNA replication and cell cycle were down-regulated in *K. marxianus* and nearly unchanged in *S. cerevisiae* (C4 in Fig. 4). Notably, genes participating in biotin biosynthesis, iron transport, flocculation, respiratory chain, and mitochondrion assembly were up-regulated in *K. marxianus* but unchanged in *S. cerevisiae* (C7 in Fig. 4), which are consistent with the particularly high copy numbers in *K. marxianus*. Overall, those findings suggest ribosome biogenesis, signal transduction, and cell cycle may not be key to the fast growth phenotype in *K. marxianus*, instead, mitochondrial function related genes and highly copied genes may strongly contribute to its fast growth.

Fundamental expression of genes involved in TCA cycle, respiratory chain, and ATP synthesis exhibit higher levels in *K. marxianus* compared to in *S. cerevisiae*

Regarding expression at 1h as the fundamental level for gene transcription, we compared this between *K. marxianus* and *S. cerevisiae*. Using *S. cerevisiae* 1h as control, there were 357 gene families with significantly higher expressions and 437 gene families with lower expressions in *K. marxianus* 1h. The enriched GO term for those genes were listed in Table 2 (see details in Additional file 4). It is clear that compared to *S. cerevisiae*, *K. marxianus* has enhanced fundamental expression in TCA cycle, respiratory chain, and ATP synthesis, while genes involved in cellular response to drug and amino acid transport had relatively lower expressions. We further validated the fundamental expression comparison of respiratory chain genes and ATP synthesis genes by RT-qPCR analysis, whose results in Fig. 5 were perfectly consistent with the findings derived by RNA-seq. Therefore, at the fundamental expression level, *K. marxianus* has remarkably higher gene expressions in mitochondrial function, including respiratory chain and ATP synthesis, which may greatly contribute to its fast growth.

***K. marxianus* and *S. cerevisiae* have significant difference in metabolic pathways during their fastest growth phase**

Due to the cultivation experiments (Additional file 1: Figure S1), the fastest growth phase for *K. marxianus* was 6h ~ 12h and for *S. cerevisiae* was 12h ~ 24h herein. This was also confirmed by the RNA-seq analysis (Fig. 4), i.e. there was a genome-wide dramatic expression change between 6h and 12h for *K. marxianus* and between 12h and 24h for *S. cerevisiae*. We further analyzed homologous gene expression change during the fastest growth phase.

For *K. marxianus*, treating 6h as control, gfold values at 12h were calculated, while for *S. cerevisiae*, treating 12h as control, gfold values at 24h were calculated. Then gene families were ordered according to *K. marxianus*' and *S. cerevisiae*'s gfold values, respectively, as shown in Fig. 6A and Fig. 6B. The majority of enriched GO for the up-regulated and down-regulated genes were also provided in Fig. 6 (details in Additional file 5). Gene expression changes during *K. marxianus*' fastest growth phase were mainly consistent with those in *S. cerevisiae*, but there are also some inconsistency (Fig. 6). The enriched GO terms for the consistent and inconsistent parts were listed in Table 3 (details in Additional file 6). In the consistency, TCA cycle and respiratory chain were both up-regulated in *K. marxianus* and *S.*

cerevisiae, while cell cycle related genes had little expression change in both species. In the inconsistency, glucose import, iron homeostasis, and ATP biosynthesis were up-regulated in *K. marxianus* but unchanged in *S. cerevisiae*. While genes for protein folding and response to heat were unchanged in *K. marxianus* but up-regulated in *S. cerevisiae*. Those difference in metabolic pathways during their fastest growth phase, may contain the key for unlocking *K. marxianus*' fast growth mystery.

***K. marxianus* preferentially co-up-regulated respiratory chain, ATP synthesis, and glucose import during the fastest growth phase**

We further compared the changes in central metabolic pathways between *K. marxianus* and *S. cerevisiae* during their fastest growth phase. As shown in Fig. 7, the central metabolism was divided into material metabolism (in the left part, including glucose transport, glycolysis, and TCA cycle) and energy metabolism (in the right part, including respiratory chain, ATP synthesis, iron transport, and heme biosynthesis). For the material metabolism, glucose import was particularly enhanced in *K. marxianus* but unchanged in *S. cerevisiae*, while glycolysis and TCA cycle were consistent between *K. marxianus* and *S. cerevisiae*. For the energy metabolism, iron transport and heme biosynthesis, which take part in respiratory chain, were specially up-regulated in *K. marxianus* but unchanged in *S. cerevisiae*. For the respiratory chain itself (not including the downstream ATP synthesis), it was enhanced both in *K. marxianus* and *S. cerevisiae*, with a stronger up-regulation in *K. marxianus*. For the ATP synthesis via F_0F_1 ATPase, it was strengthened in *K. marxianus* but slightly down-regulated in *S. cerevisiae*, suggesting in *K. marxianus* the respiratory chain and ATP biogenesis are particularly tightly coupled. Notably, as aforementioned, during this phase, glucose transport was also evidently up-regulated in *K. marxianus* but not in *S. cerevisiae*. The implication for the co-up-regulated glucose import and respiratory chain in *K. marxianus* will be explored in great details in the discussion part.

To gain insight into the underlying mechanism for the co-expression of respiratory chain and ATP synthesis in *K. marxianus* but not in *S. cerevisiae*, we analyzed the upstream of the involved genes. The upstream 1kb of 27 genes in respiratory chain and 14 genes in ATP synthesis were analyzed via MEME software online [19] for motif discovery. There were six significant motifs (E-value < 0.05) located in those upstream regions in *K. marxianus* (Fig. 8A), but only one significant motif in *S. cerevisiae* (Fig. 8B). The sequences and location of such motifs were provided in Fig. 8. Similarly, the upstream 1kb of 27 genes in respiratory chain and 9 genes in glucose import were also analyzed by MEME. Three significant motifs were detected in *K. marxianus* (Fig. 8C) and one motif identified in *S. cerevisiae* (Fig. 8D). The above findings provide molecular support for the particular co-expression of glucose transport, respiratory chain, and ATP synthesis in *K. marxianus*, indicating the importance of regulatory region evolution during *K. marxianus* speciation.

Discussion

In this study, upon the genome content analysis of *K. marxianus* evolution in Saccharomycetaceae and the transcriptome analysis between *K. marxianus* and *S. cerevisiae*, we found the mechanisms for *K.*

marxianus' fast growth may be closely related to ATP production. This finding coincidentally meets the widely recognized concept that ATP is the energy in living cells and the key to drive energy-consuming processes such as growth [20,21]. In the following, we will discuss the traditional dilemma for the trade-off in ATP production and the new metabolic strategy for how *K. marxianus* overcomes this, as discovered in this work.

Cell's growth rate is tightly connected with ATP production [21]. Generally, higher growth rate relies on higher ATP production yield as well as production rate [21]. However, heterotrophic organisms usually have to face the trade-off between ATP production rate and yield [22]. In details, during cells' degrading substrates with higher free energy into products with lower free energy, the free energy difference between substrate and product can be partly conserved into ATP production and partly dissipated to drive the degradation reaction. If substrate (such as glucose) is catabolized through the 'glycolysis - respiratory chain' route (Fig. 9A, blue arrow), 32 ATPs are generated. In this case, free energy difference is nearly totally preserved into ATPs and the reaction is in thermodynamic equilibrium, resulting in the very slow rate of substrate degradation (including glucose import), which hinders the competition with other species for environmental resources. If cells adopt the route of 'glycolysis - fermentation' for substrate degradation (Fig. 9A, red arrow), only 2 ATPs are produced, the other part of free energy difference is used to promote the degradation reaction, triggering the fast glucose import, which is a remarkable advantage in resource competition. However, the very few ATP yield can hardly supply sufficient energy for cell growth.

We found *K. marxianus* may have developed two new strategies during evolution for ATP production to guarantee its fastest growth. (1) *K. marxianus* may co-up-regulate respiratory chain and glucose transport, which does not exist in *S. cerevisiae* (Fig. 7, Fig. 9B). This was supported both by the specific genes and homologous genes. For KMS genes, due to their predicted function based on Pearson correlation coefficient, genes possibly involved in respiratory chain and glucose transport were up-regulated in the fastest growth phase (Fig. 3B). For homologous genes, genes participating in respiratory chain and ATP synthesis were highly up-regulated in *K. marxianus* than in *S. cerevisiae* at the fundamental level (Table 2, Fig. 5), while genes involved in respiratory chain (including iron transport and heme biosynthesis) as well as in glucose import were co-up-regulated in *K. marxianus* but not in *S. cerevisiae* during their fastest growth phase (Fig. 7). These findings suggest *K. marxianus* may gain a new strategy of coupling glucose import with respiratory chain, to ensure during the high yield of ATP production (by respiratory chain), available medium resources are also strongly absorbed (by glucose import). This particular co-up-regulation in *K. marxianus*, was further supported by the larger motif numbers in the upstream of respiratory chain and glucose transporter genes in *K. marxianus* (Fig. 8C) than in *S. cerevisiae* (Fig. 8D), implying *K. marxianus* may acquire the new metabolic strategy via gene regulation region evolution. This is consistent with previous reports that the cis-regulatory systems are dynamically evolved in Ascomycete fungi [23].

K. marxianus may tightly co-up-regulate respiratory chain and ATP synthesis related F_0F_1 ATPase during the fastest growth phase (Fig. 7). In contrast, in *S. cerevisiae*'s fastest growth phase, respiratory chain

was up-regulated but F_0F_1 ATPase was unchanged (Fig. 7). Notably, the up-regulation of genes involved in response to heat and protein folding in *S. cerevisiae* but unchanged in *K. marxianus* (Table 3), is consistent with the imperfect coupling of respiratory chain and ATP synthesis in *S. cerevisiae*. It has been reported that when respiratory chain and ATP synthesis are uncoupled, a large part of the free energy difference derived from electron transport chain are transformed to heat [24], and this leads to the up-regulation of cellular response to heat and molecular chaperones for maintaining protein folding [25]. The remarkably higher number of motifs in the upstream of respiratory chain and F_0F_1 ATPase genes in *K. marxianus* (Fig. 8A) than in *S. cerevisiae* (Fig. 8B), suggesting *K. marxianus* may have particularly optimized its regulation regions to ensure highly efficient production of ATP.

Finally, in this work we attempted to make function prediction for *K. marxianus*-specific (KMS) genes. Based on the recognition that genes with strong co-expression may be in the same or closely related pathway [17,18], we used the function known gene with the highest Pearson correlation coefficient to infer the function of KMS genes. The predicted function of KMS genes were generally accordant with the findings of homologous genes (Fig. 9B). However, we found in some cases, several genes possessing very close correlation coefficients with a KMS gene, may have distinct functions. Therefore, there is still a long way to unravel the functions of KMS genes by experiment validation and to re-evaluate their contribution to the fast-growth phenotype.

Conclusions

In this study, based on the comparative genome study for gene content and expression pattern, we found *K. marxianus*' fast growth may be attributed to the preferentially enhanced fundamental expressions of TCA cycle and respiratory chain genes, as well as the co-up-regulation of glucose transporter, respiratory chain and ATP synthesis genes during the fastest growth phase. Those conclusions underscore the importance of genome-wide rewiring of transcriptional network during evolution. To sum up, our findings provide a theoretical support for *K. marxianus*' wide industrial application, and also propose a practicable means to explore species' complex phenotype formation by combining genome evolution and homologous gene expression analysis.

Methods

Yeast strains and culture conditions

K. marxianus FIM1 strain used in this work was deposited at China General Microbiological Culture Collection Center (CGMCC) with a reference number of 10621. *S. cerevisiae* S288c strain was used in this study. FIM1 and S288c were transferred from YPD plates to 10 mL tube containing 3 mL YPD medium (2% glucose, 2% peptone, 1% yeast extract). Strains were inoculated and grown at 30°C and 220 rpm for overnight, then transferred to 150 mL flasks with 50 mL YPD medium to start at an OD_{600} of 0.1. After that the flasks were shaken at 30°C and 220 rpm.

Genome sequencing, assembly, and annotation

The nuclear DNA was extracted from *K. marxianus* FIM1 using standard protocols. DNA libraries harboring 300-bp, 3-kb, and 8-kb inserts were subsequently constructed. Paired-end sequencing of these DNA libraries was then performed on an IlluminaHiSeq 2000 and PacBio RS sequencer by using standard protocols. Low quality short reads were trimmed or ignored by using Trimmomatic. The SOAPdenovo [26] was then applied to the short reads passing the filtering procedure to generate scaffolds. Protein-coding genes were predicted from the assembled genome by using the “Yeast Genome Annotation Pipeline” [15], and then blasted against NCBI non-redundant (nr) protein database with E-value cutoff as 10^{-5} for functional annotation. tRNAs were predicted by tRNAscan-SE [27] and rRNAs were predicted by Barrnap (<https://github.com/tseemann/barrnap>). *K. marxianus* FIM1’s complete genome sequence and annotation GFF3 file have been deposited in NCBI Genome with GenBank accession No. CP015054 to CP015061.

Identification of gene families and phylogenetic analysis

To study the gain/loss of genes during *K. marxianus* evolution, we downloaded 14 genome sequences in Saccharomycetales (Fig. 1). Gene families of all protein sequences in the total genomes were identified as follows. First, an all-against-all BLAST was applied on these protein sequences with E-values $< 10^{-5}$. Second, global protein similarities were calculated using InParanoid [28] and those matched with both sufficient gene coverage ($> 75\%$) and alignment identity ($> 50\%$) were left for further analyses. Third, orthologue clusters, i.e. the ‘gene families’, were identified by comparing protein alignments using OrthoMCL [29]. Multiple sequence alignment of proteins in each gene family was obtained using MUSCLE [30] with default parameters and was further trimmed with trimAL [31]. The gene families containing single-copy genes within each genome were concatenated into a super gene for reconstruction of phylogenetic tree, using the maximum likelihood method with 100 bootstrap replicates. The Dollo Parsimony analysis was performed on all the gene families of the 15 yeasts by using PHYLIP package to analyze the gain and loss of gene families at each node along phylogenetic tree.

Gene copy number analysis

To analyze gene copy numbers among different species in Saccharomycetales, we first found out the replicated genes in *K. marxianus* according to the same functional annotation within a gene family, then counted the copy number of those genes in other species based on their replication in the gene family.

Sample preparation for RNA-seq

For comparing the RNA-seq data of *K. marxianus* and *S. cerevisiae* along their growth course, strains were previously inoculated in 50 mL YPD medium overnight under agitation at 220rpm at 30°C, then grown into a new flask containing 50 mL fresh YPD medium at initial OD₆₀₀ of 0.1. After that, 150 L samples were collected at 1h, 4h, 6h, 12h, 24h, 48h, and 72h. Samples for RNA-seq analyses were both in biological two replicates. Total RNA was extracted using the ZR Fungal/Bacterial RNA MiniPrep™

(Zymo Research, CA). The samples were then sent to the Genergy Biotechnology company (Shanghai, China) for quality and quantity evaluation and sequencing.

RNA-seq analysis and differential expression identification

We obtained 15.1 million pair-end reads on average for each RNA sample. After initial QC, short 150 bp reads were mapped to the reference genomes of *K. marxianus* FIM1 and *S. cerevisiae* S288c (Saccharomyces Genome Database) using HISAT2.1 [32]. Finally the FPKM values for all genes were obtained (Additional file 2). Due to the fact that there usually exists a many-to-many correspondence for the homologous genes between two species, and a gene family contains a total set of homologous genes, we used gene family to compare homologous genes' expression difference between *K. marxianus* and *S. cerevisiae*. For each gene family in a species, its expression is the sum of FPKM values from the contained genes. Then gene family's differential expression was calculated using the GFOLD algorithm [33], which is roughly equivalent to the raw fold change \log_2 ratio value, but takes into account the uncertainty of gene expression measurement by RNA-seq, thus is more reliable than the raw fold change [33]. Two fold or more changes, i.e. $|\text{gfold}| \geq 1$, were defined as significantly differentially expressed herein. Gene family with $\text{gfold} \geq 1$ is defined as up-regulated and with $\text{gfold} \leq -1$ is referred as down-regulated. When analyzing gene expression change within a species along cultivation time, expression at 1h was as control, then gene family's gfold value at a subsequent time point was calculated. When comparing fundamental expressions between *K. marxianus* and *S. cerevisiae*, *S. cerevisiae* 1h was chosen as control, the gfold value for *K. marxianus* 1h vs *S. cerevisiae* 1h was then calculated.

RT-qPCR analysis

Genes' expression levels were determined by real-time reverse transcription PCR (RT-qPCR). To compare genes' fundamental expressions between *K. marxianus* and *S. cerevisiae*, strains were diluted into YPD medium at an initial concentration of 0.1 OD₆₀₀, then cells were collected at 1 h. The total RNA was isolated using ZR Fungal/Bacterial RNA MiniPrep kit (R2014, Zymoresearch) and cDNA was obtained by reverse transcription using PrimeScript RT (RR037A, Takara). Analysis of cDNA was conducted on a LightCycler 480 (Roche Applied Science, Germany) with the SYBR Premix Ex Taq II (RR820A, Takara). The expressions of individual genes were normalized against the level of 18S rRNA. Primers used for RT-qPCR were listed in Additional file 1: Table S1.

Function prediction for *K. marxianus*-specific genes based on correlation coefficient

Inspired by the idea of using gene co-expression network to predict the function of hypothetical genes in *Aspergillus niger* [17], in this study we employed Pearson correlation coefficient (denoted as r) to infer the function of *K. marxianus*-specific (KMS) genes, which was calculated between a KMS gene and a function known gene along the time points during cultivation. If $r > 0.8$, the function known gene can be considered as candidate for function prediction, finally the gene with the highest r value was chosen for function inference.

Abbreviations

KM: *Kluyveromyces marxianus*; SC: *Saccharomyces cerevisiae*; KMS: *K. marxianus*-specific; GO: Gene Ontology; TCA cycle: tricarboxylic acid cycle; ROS: reactive oxygen species; RT-qPCR: real-time reverse transcription PCR

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors consent for publication.

Availability of data and material

We provide all the necessary data for the publication of this article. All additional data are present in the article and the additional material documents.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (No. 31970068 and No. 31770094 to HL, No. 31770244 to JQ, and No. 31771384 to YY).

Authors' contributions

WJM, YY, JGZ, and JQ performed the genome sequencing and assembly. WJM, TYL, and JQ carried out genome annotation and upload to NCBI. XMY, HYR and TYL carried out the strain cultivation and RNA-seq sample preparation. WLL and JQ reconstructed phylogenetic tree. WJM and JQ analyzed gain/lost genes and high-copied genes. WJM, JQ, and HL analyzed the RNA-seq data. HYR performed RT-qPCR analysis. WJM analyzed the motifs of the upstream sequences. WJM and HL prepared the manuscript. HL organized this research project. All authors have read and approved the final version of the manuscript.

Acknowledgements

The authors acknowledge financial support from the National Natural Science Foundation of China (No. 31970068 and No. 31770094 to HL, No. 31770244 to JQ, and No. 31771384 to YY).

References

1. Lane MM, Morrissey JP. *Kluyveromyces marxianus*: a yeast emerging from its sister's shadow. *Fungal Biol Rev*. 2010;24(1-2):17-26.
2. Marcisauskas S, Ji B, Nielsen J. Reconstruction and analysis of a *Kluyveromyces marxianus* genome-scale metabolic model. *BMC Bioinformatics*. 2019;20(1):551.
3. Mo W, Wang M, Zhan R, Yu Y, He Y, Lu H. *Kluyveromyces marxianus* developing ethanol tolerance during adaptive evolution with significant improvements of multiple pathways. *Biotechnol Biofuels*. 2019;12(63).
4. Zoppellari F, Bardi L. Production of bioethanol from effluents of the dairy industry by *Kluyveromyces marxianus*. *New Biotechnol*. 2013;30(6SI):607-13.
5. Gombert AK, Madeira JV, Cerdanv M, Gonzalez-Siso M. *Kluyveromyces marxianus* as a host for heterologous protein synthesis. *Appl Microbiol Biot*. 2016;100(14):6193-208.
6. Gao J, Yuan W, Li Y, Xiang R, Hou S, Zhong S, Bai F. Transcriptional analysis of *Kluyveromyces marxianus* for ethanol production from inulin using consolidated bioprocessing technology. *Biotechnol Biofuels*. 2015;8(115).
7. Groeneveld P, Stouthamer AH, Westerhoff HV. Super life - how and why 'cell selection' leads to the fastest-growing eukaryote. *FEBS J*. 2009;276(1):254-70.
8. Gao M, Zheng Z, Wu J, Dong S, Li Z, Jin H, Zhan X, Lin C. Improvement of specific growth rate of *Pichia pastoris* for effective porcine interferon-alpha production with an on-line model-based glycerol feeding strategy. *Appl Microbiol Biot*. 2012;93(4):1437-45.
9. Jeong H, Lee D, Kim SH, Kim H, Lee K, Song JY, Kim BK, Sung BH, Park JC, Sohn JH, Koo HM, Kim JF. Genome sequence of the thermotolerant yeast *Kluyveromyces marxianus* var. *marxianus* KCTC 17555. *Eukaryotic Cell*. 2012;11(12):1584-85.
10. Suzuki T, Hoshino T, Matsushika A. Draft genome sequence of *Kluyveromyces marxianus* strain DMB1, isolated from sugarcane bagasse hydrolysate. *Genome Announc*. 2014;2(4):e714-33.
11. Silveira WB, Diniz RHS, Cerdan ME, Gonzalez-Siso MI, Souza RDA, Vidigal PMP, Brustolini OJB, de Almeida Prata ERB, Medeiros AC, Paiva LC, Nascimento M, Ferreira EG, dos Santos VC, Bragança CRS, Fernandes TAR, Colombo LT, Passos FML. Genomic sequence of the yeast *Kluyveromyces marxianus* CCT 7735 (UFV-3), a highly lactose-fermenting yeast isolated from the Brazilian dairy industry. *Genome Announc*. 2014;2(6):e1114-36.
12. Lertwattanasakul N, Kosaka T, Hosoyama A, Suzuki Y, Rodrussamee N, Matsutani M, Murata M, Fujimoto N, Suprayogi, Tsuchikane K, Limtong S, Fujita N, Yamada M. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnol Biofuels*. 2015;8(47).
13. Inokuma K, Ishii J, Hara KY, Mochizuki M, Hasunuma T, Kondo A. Complete genome sequence of *Kluyveromyces marxianus* NBRC1777, a nonconventional thermotolerant yeast. *Genome Announc*. 2015;3(2):e00389-15.
14. Schabort DTWP, Letebele PK, Steyn L, Kilian SG, du Preez JC. Differential RNA-seq, multi-Network analysis and metabolic regulation analysis of *Kluyveromyces marxianus* reveals a compartmentalised response to xylose. *PLoS One*. 2016;11(6):e0156242.
15. Proux-Wera E, Armisen D, Byrne KP, Wolfe KH. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics*. 2012;13:237.
16. Wolfe KH, Armisen D, Proux-Wera E, OhEigeartaigh SS, Azam H, Gordon JL, Byrne KP. Clade- and species-specific features of genome evolution in the Saccharomycetaceae. *FEMS Yeast Res*. 2015;15:fov0355.
17. Schaepe P, Kwon MJ, Baumann B, Gutschmann B, Jung S, Lenz S, Nitsche B, Paege N, Schuetze T, Cairns TC, Meyer V. Updating genome annotation for the microbial cell factory *Aspergillus niger* using gene co-expression networks. *Nucleic Acids Res*. 2019;47(2):559-69.
- 18.

Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T. A gene ontology inferred from molecular networks. *Nature Biotechnol.* 2013;31(1):38. 19. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren JR, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37S:W202-08. 20. Hara KY, Kondo A. ATP regulation in bioproduction. *Microb Cell Fact.* 2015;14:198. 21. Chen Y, Nielsen J. Energy metabolism controls phenotypes by protein efficiency and allocation. *PNAS.* 2019;116(35):17592-97. 22. Pfeiffer T, Schuster S, Bonhoeffer S. Cooperation and competition in the evolution of ATP-producing pathways. *Science.* 2001;292(5516):504-07. 23. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol.* 2004;2(12):e398. 24. Nakamura T, Matsuoka I. Calorimetric studies of heat of respiration of mitochondria. *J Biochem.* 1978;84(1):39-46. 25. GAGE DJ, NEIDHARDT FC. Adaptation of *Escherichia coli* to the uncoupler of oxidative phosphorylation 2, 4-dinitrophenol. *J Bacteriol.* 1993;175(21):7105-08. 26. Luo R, Liu B, Xie Y, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience.* 2012;1:18. 27. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997;25(5):955-64. 28. O'Brien KP, Remm M, Sonnhammer E. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 2005;33(SI):D476-80. 29. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003;13(9):2178-89. 30. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004;5(113):1-19. 31. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25(15):1972-73. 32. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 2015;12(4):121-357. 33. Feng J, Meyer CA, Wang Q, Liu JS, Liu XS, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics.* 2012;28(21):2782-88.

Additional Files

Additional file 1: Supplementary information. A file containing 1 supplementary figures and 1 supplementary table. **Fig S1.** Growth curves of *K. marxianus* and *S. cerevisiae* in YPD medium at 30°C. **Table S1.** Primers for RT-qPCR analysis.

Additional file 2: Transcriptome data. A file containing the expression data of all genes in *K. marxianus* and in *S. cerevisiae* based on RNA-seq analysis, respectively.

Additional file 3: Detailed information of enriched GO terms. A file containing the enriched GO terms (pvalue < 0.05) for 12 clusters (name C1, C2, ... , C12) of differentially expressed genes in *K. marxianus* or *S. cerevisiae* using respective 1h as control.

Additional file 4: Detailed information of enriched GO terms. A file containing the enriched GO terms for higher expressed and lower expressed genes in *K. marxianus* 1h compared to *S. cerevisiae* 1h.

Additional file 5: Detailed information of enriched GO terms. A file containing enriched GO terms for the up-regulated and down-regulated genes in *K. marxianus* and *S. cerevisiae* during their fastest growth phase, respectively.

Additional file 6: Detailed information of enriched GO terms. A file containing enriched GO terms for the differentially expressed genes with consistent and inconsistent patterns between *K. marxianus* and *S. cerevisiae* during their respective fastest growth phase.

Tables

Table 1. Enriched GO terms for differentially expressed gene family clusters with species-respective 1h as control.

Cluster	Number of gene families	Enriched GO
1	128	Ribosome biogenesis, tRNA processing, valine biosynthesis, pre-replicative complex assembly involved in DNA replication
2	132	Ribosome biogenesis, tRNA processing, translation, glycolysis, cell wall formation, amino acid catabolism
3	209	Translation, protein folding, purine biosynthesis, fatty acid biosynthesis, steroid biosynthesis, cytokinesis
4	344	Protein glycosylation, DNA replication, nucleotide biosynthesis, cell cycle, lipid metabolism, vesicle transport
5	208	RNA processing, iron metabolism, ribosome biogenesis, protein glycosylation, kinetochore assembly, mitosis
6	302	DNA replication, amino acid degradation, glycogen biogenesis, glycolysis, respiratory chain, DNA repair, vacuole transport
7	271	Biotin biosynthesis, amino acid transport, respiratory chain, flocculation, iron transport, mitochondrion assembly
8	128	Glucose transport, ATP biogenesis, TCA cycle, endocytosis, autophagy, sporulation, proton transport
9	120	Flocculation, fructose metabolism, trehalose biosynthesis, autophagy, mitophagy, TCA cycle, respiratory chain, cell aging
		Xylose catabolism, glycogen biosynthesis, TCA cycle, response to oxidative stress, autophagy, apoptosis

11	97	Acetate transport, ammonium transport, acetate biosynthesis, NADPH regeneration, fatty acid beta-oxidation
12	480	intracellular signal transduction, MAPK cascade, cell wall organization, osmosensory signalling

Table 2. Enriched GO for differentially expressed genes in *K. marxianus* 1h compared to *S. cerevisiae* 1h.

Type	Number of gene families	Enriched GO
Higher expressed	357	TCA cycle, flocculation, aerobic respiration, ATP synthesis, mitochondrial transport, mitochondrial translation
Lower-expressed	437	cellular response to drug, amino acid transport, lysine biosynthesis, ribosome biogenesis, fatty acid catabolism

Table 3. Comparison of homologous gene expression pattern during the fastest growth phase

Trend	Type	Number	Enriched GO
consistent	KM ↑ vs. SC ↑	229	Flocculation, TCA cycle, autophagy, mitochondrial electron transport, xylose catabolism, response to ROS
	KM ↓ vs. SC ↓	330	Translation, ribosome biogenesis, tRNA processing, sterol biosynthesis, fatty acid biosynthesis
	KM - vs. SC -	1072	Phosphorylation, amino acid transport, cell cycle, biotin biosynthesis, DNA repair, osmosensory signaling
inconsistent	KM ↑ vs. SC -	368	Glucose import, iron ion homeostasis, ATP biosynthesis, respiratory chain complex IV assembly
	KM ↑ vs. SC ↓	24	Glycine catabolism, IMP biosynthesis, mitochondrial proton-transporting ATP synthase complex assembly
inconsistent	KM ↓ vs. SC -	321	Protein glycosylation, steroid biosynthesis, DNA replication, ER to Golgi vesicle-mediated transport
	KM ↓ vs. SC ↑	44	Lactate catabolism, lipid storage, phospholipid biosynthesis, response to pheromone
inconsistent	KM - vs. SC	344	ammonium transport, protein folding, pentose-phosphate shunt, response to heat

↑

KM - 329
vs. SC
↓

ribosome biogenesis, rRNA processing, translation, cell wall organization, NADH oxidation

In this table, 'Number' represents the counted number of gene families, '↑', '↓', and '-' denote up-regulation, down-regulation, and no change during the fastest growth phase, respectively.

Figures

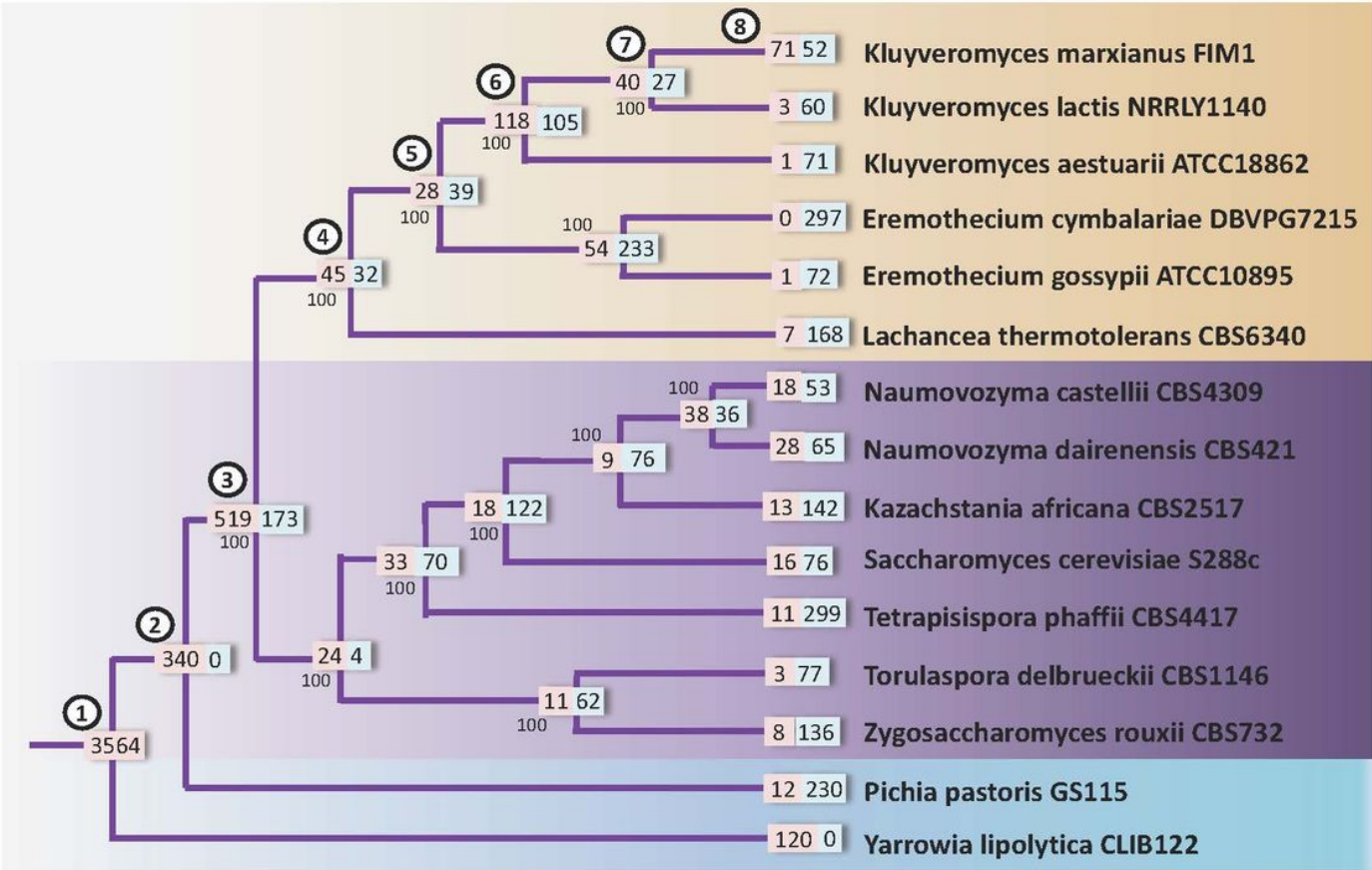


Figure 1

Phylogenetic tree of *K. marxianus* evolution in Saccharomycetales. The tree was generated based on the concatenated sequence of 636 single-copy genes without long-branch score heterogeneity. The bootstrap value 100 at each node point indicates the correctness of inferred topology. The blue part contains Y.

lipolytica in Dipodascaceae and *P. pastoris* in Phaffomycetaceae as outgroup. The purple and the brown parts represent the two major clades in Saccharomycetaceae, respectively, i.e., the one close to *Saccharomyces* and the one close to *Kluyveromyces*. At each node, the number in circle denotes the serial number in phylogenetic tree, and the numbers of gained and lost gene families were provided in pink and blue boxes, respectively.

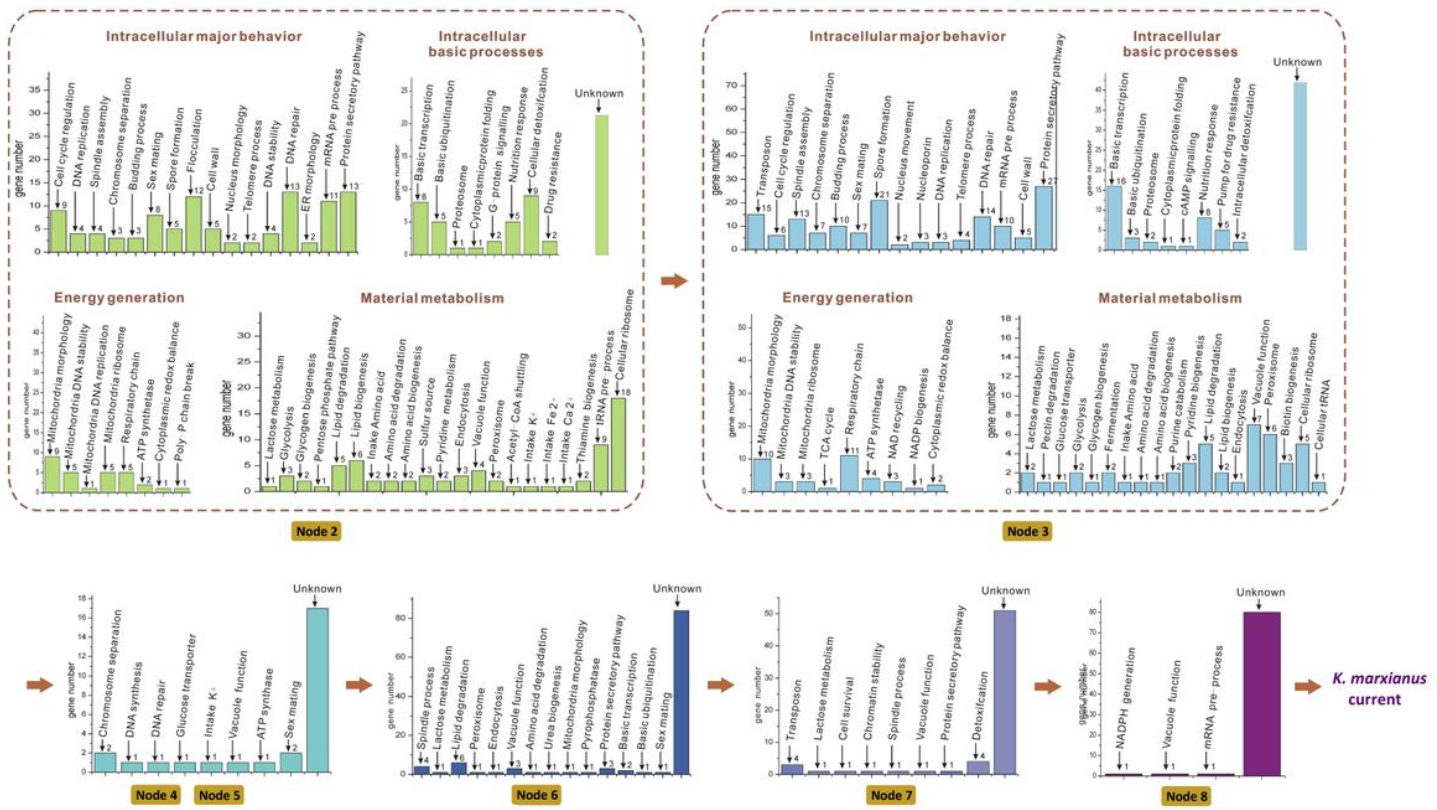


Figure 2

Functional categories of genes gained at each node along *K. marxianus* speciation. The number upon each functional category column denotes the counted number of *K. marxianus* gained genes involved in the function process.

inferred by the known function of the highest correlated gene, were provided on the left, while the corresponding correlation coefficients were listed on the right.

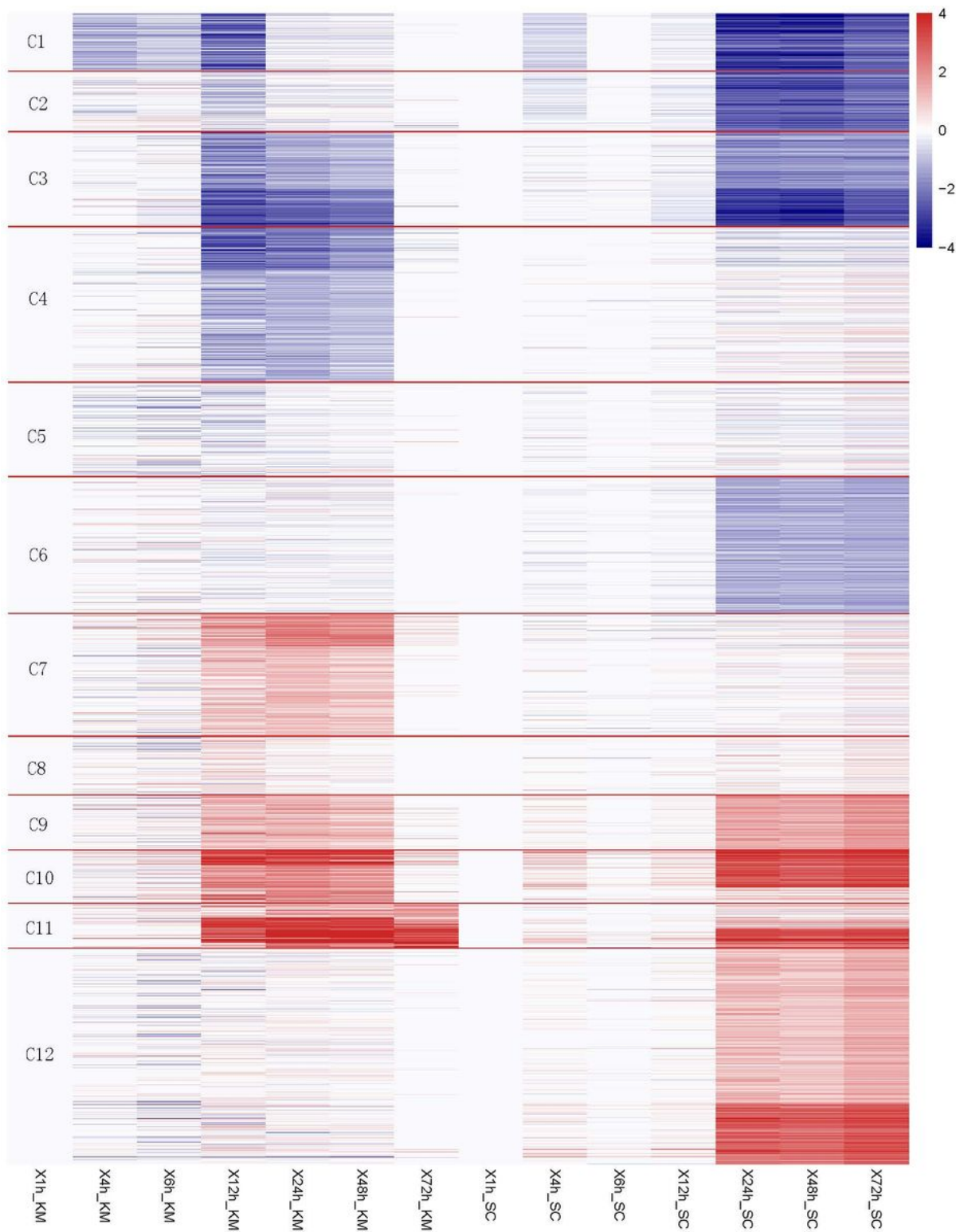


Figure 4

Clustering expression patterns of homologous gene family in *K. marxianus* and *S. cerevisiae*. Gene families with 2 fold or more changes at least at one time point after 1h in *K. marxianus* or *S. cerevisiae* were presented in this figure. The horizontal axis carries out the time points along cultivation for *K.*

marxianus (on the left, abbreviated as 'KM') and *S. cerevisiae* (on the right, abbreviated as 'SC'). Each line represents the differential expression (i.e. gfold value) of a gene family containing homologous genes in *K. marxianus* and *S. cerevisiae*, using respective 1h as control. Using k-means based on Euclidean distance, gene families were clustered into 12 clusters, denoted as C1, C2, ..., C12 in the first column. The clusters C4, C6, C7, and C12 are apparently different between *K. marxianus* and *S. cerevisiae*.

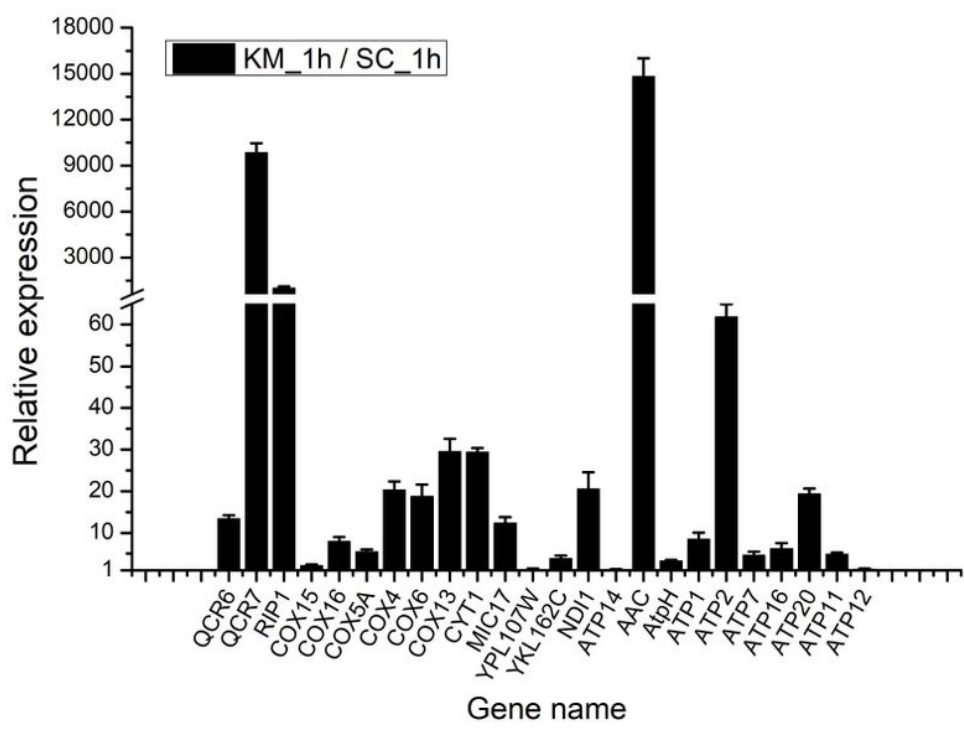


Figure 5

RT-qPCR analysis of genes' fundamental expression in *K. marxianus* compared to *S. cerevisiae*. 'KM_1h' and 'SC_1h' denote gene expression at 1h in *K. marxianus* and *S. cerevisiae*, respectively. Using SC_1h as control, the ratio of KM_1h vs SC_1h was provided. Note that the starting point for the vertical axis is 1, which directly displays *K. marxianus*' higher expression than *S. cerevisiae*. Some genes for respiratory chain (e.g. YNL134C) were nearly undetectable in *S. cerevisiae* at 1h, thus have no value in the figure.

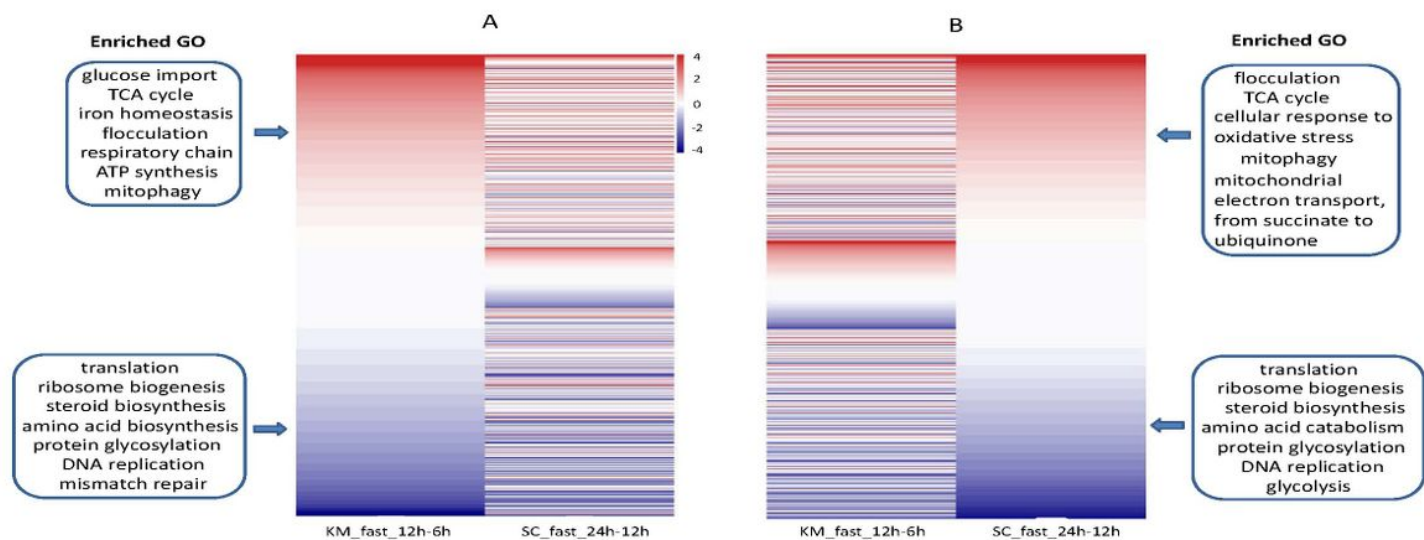


Figure 6

Comparison of homologous gene expression change during fastest growth phase. (A) Gene families ordered by *K. marxianus*' gfold values. (B) Gene families ordered by *S. cerevisiae*'s gfold values. In each subfigure, the left and right heatmaps are for *K. marxianus*' and *S. cerevisiae*'s gfold values, respectively, which were calculated due to 12h vs. 6h for *K. marxianus* and 24h vs. 12h for *S. cerevisiae*, representing genes' expression change during the fastest growth phase. Enriched GO terms for the up-regulated and down-regulated genes in *K. marxianus* were provided in Fig. A and for *S. cerevisiae* were provided in Fig. B.

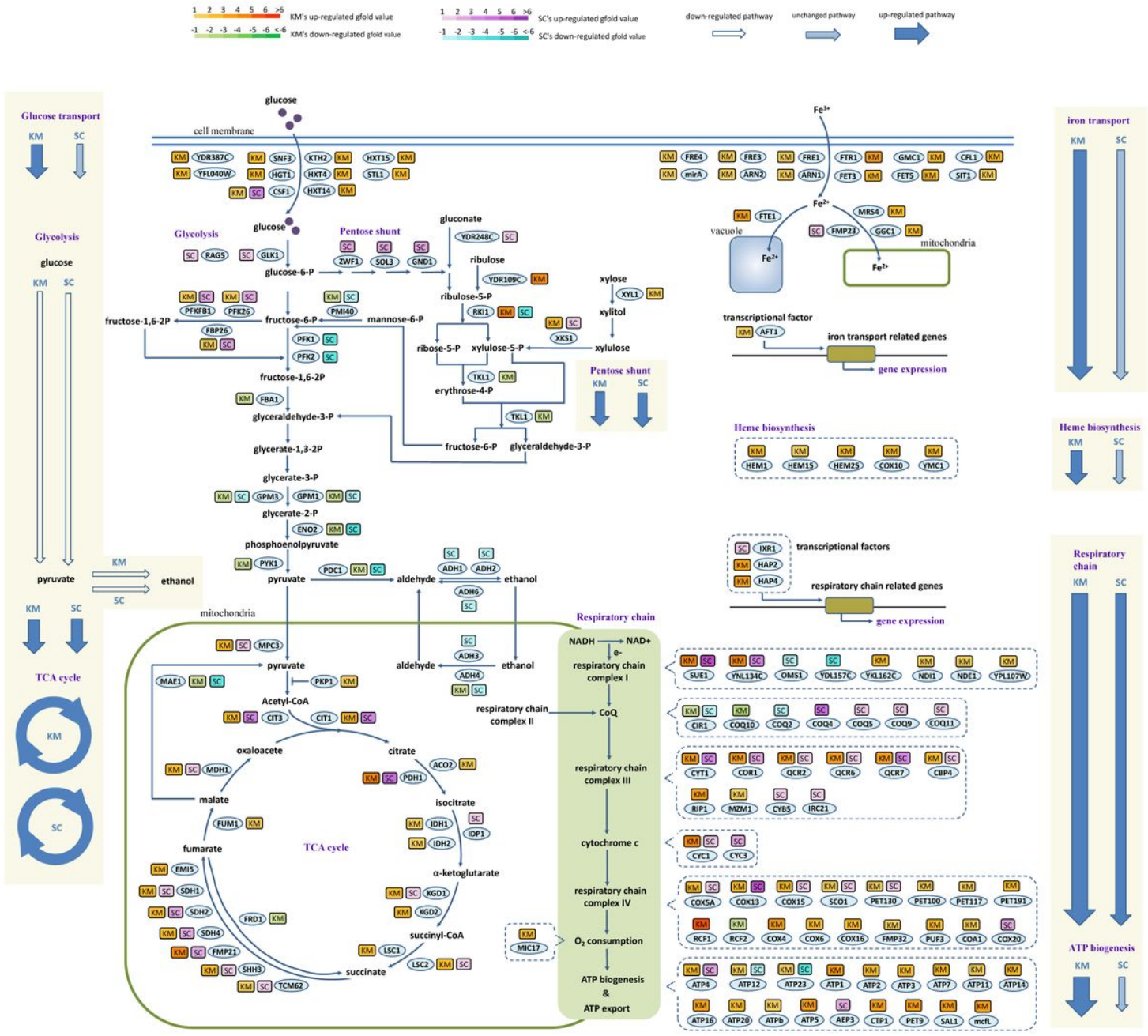


Figure 7

Comparison of central metabolic pathways during fastest growth phase between *K. marxianus* and *S. cerevisiae*. For each gene, the nearby colour box denotes its up-regulation or down-regulation. The red system and green system stand for up-regulation and down-regulation in *K. marxianus*, respectively. The purple system and blue system stand for up-regulation and down-regulation in *S. cerevisiae*, respectively. The correspondence of colour and gfold value is quantified by the colour bars in figure's upper part. For the changes in pathway, the hollow arrow, slant filled arrow, and blue solid arrow denote the generally down-regulated, unchanged, and up-regulated state of the pathway, respectively.

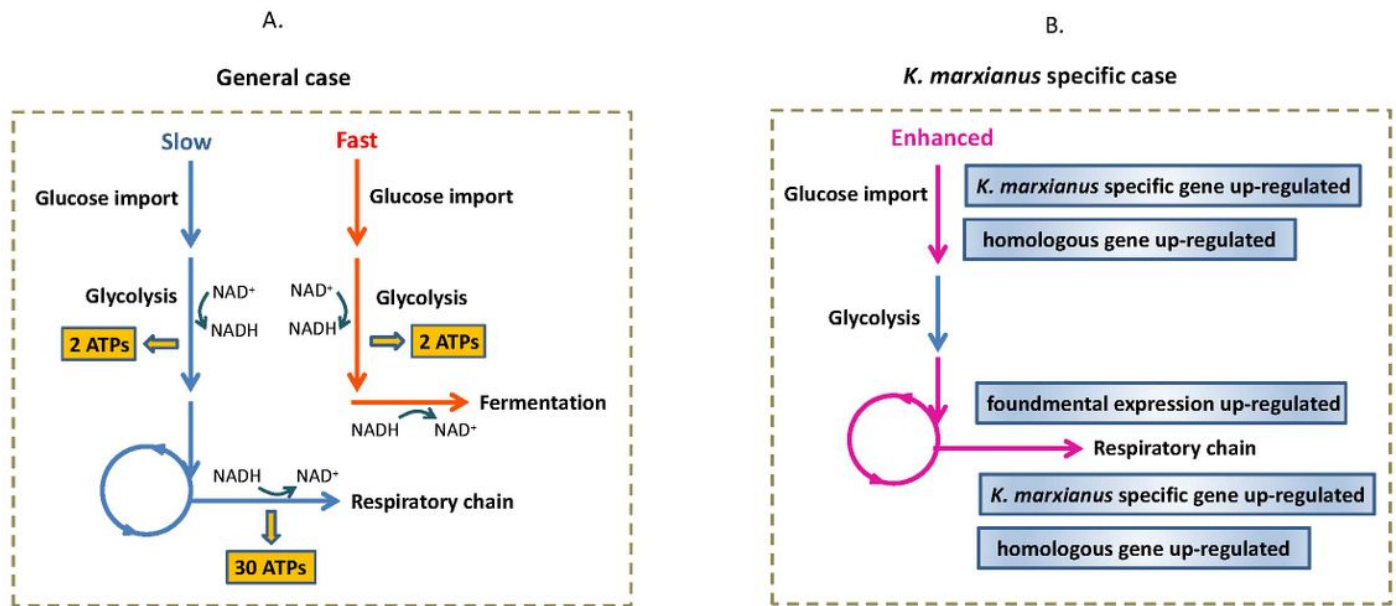


Figure 9

Diagram of ATP production routes in cells. (A) General case for the trade-off between ATP production yield and rate. (B) *K. marxianus* specific case for simultaneously high yield of ATP and high import of glucose. In Fig. A, a general trade-off between the yield and rate of ATP production exists, i.e. if substrate degradation is via the 'glycolysis - respiratory chain' way (denoted in blue), the yield of ATP is high but the rate of glucose import is very low, which leads to a disadvantage in competing for environmental resources; if substrate is degraded by the route of 'glycolysis - fermentation' (denoted in red), the rate of glucose import is high but ATP production yield is very low, which hardly supports sufficient energy demand for cell fast growth. In Fig. B, *K. marxianus* may have developed a new strategy to overcome the dilemma for ATP production and glucose import, by simultaneously up-regulating the genes involved in respiratory chain and glucose transport.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)
- [Additionalfile4.xlsx](#)
- [Additionalfile2.xlsx](#)
- [Additionalfile5.xlsx](#)
- [Additionalfile6.xlsx](#)
- [Additionalfile3.xlsx](#)