DANIELA ARRUDA COSTA

GENÔMICA COMPARATIVA DE LINHAGENS DE *Saccharomyces E Kluyveromyces* DE INTERESSE BIOTECNOLÓGICO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, para obtenção do título de *Doctor Scientiae.*

VIÇOSA
MINAS GERAIS - BRASIL
2015

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
C837g
2015

Costa, Daniela Arruda, 1983-
Genômica comparativa de linhagens de *Saccharomyces* e *Kluyveromyces* de interesse biotecnológico / Daniela Arruda Costa. – Viçosa, MG, 2015.
xv,133f. : il. (algumas color.) ; 29 cm.

Inclui anexos.
Orientador: Luciano Gomes Fietto.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.93-103.

1. Leveduras. 2. Genomas. 3. *Saccharomyces cerevisiae*. 4. *Kluyveromyces. 5.* Biotecnologia. 6. Etanol. I. Universidade Federal de Viçosa. Departamento de Microbiologia. Programa de Pós-graduação em Microbiologia Agrícola. II. Título.

CDD 22. ed. 579.562

DANIELA ARRUDA COSTA

GENÔMICA COMPARATIVA DE LINHAGENS DE *Saccharomyces E Kluyveromyces* DE INTERESSE BIOTECNOLÓGICO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Microbiologia Agrícola, para obtenção do título de *Doctor Scientiae.*

APROVADA: 31 de julho de 2015.

_____	_____
Thiago Mafra Batista	Cláudio Lísias Mafra de Siqueira
_____	_____
Otávio José Bernardes Brustolini	Cynthia Cânedo Silva

_____
Luciano Gomes Fietto
(Orientador)

Aos meus queridos pais Simeão e Elza

Ao amado Pedro

DEDICO

AGRADECIMENTOS

Agradeço a Deus por me ajudar a perseverar nessa longa caminhada, repleta de caminhos tortuosos, abrindo portas e endireitando o caminho.

Ao orientador e amigo Dr. Luciano Gomes Fietto por mais uma vez aceitar o desafio de me conduzir na pesquisa acadêmica e contribuir de maneira ímpar para minha contínua formação pessoal e profissional: Bacharel, Mestre e agora Doutora em Microbiologia Agrícola. Agradeço os ensinamentos recebidos, os conselhos singelamente prestados e as oportunidades concedidas.

Ao coorientador Dr. Jeronimo Conceição Ruiz por me proporcionar um aprendizado excelente em uma área fantástica da biologia: a bioinformática.

À Universidade Federal de Viçosa e a Fiocruz-MG, professores e funcionários, por oferecerem um ambiente propício e incentivador ao desenvolvimento da ciência e pesquisa, pautado na qualidade do ensino.

Às agências de fomento à pesquisa: CAPES pela bolsa concedida, FAPEMIG e CNPq pelo apoio financeiro concedido aos projetos.

Agradeço aos Laboratórios do qual possuo um orgulho imenso de fazer parte: Laboratório de Biotecnologia Molecular (UFV) e Grupo Informática de Biossistemas (Fiocruz-MG). Levarei vocês no coração e na memória para sempre.

Em especial, agradeço aos meus pais, irmãos, e, ao Pedro, por me incentivarem e acreditarem no meu potencial, inclusive naqueles momentos difíceis em que eu estava descrente. Muito obrigada!

À minha família e amigos de BH, Viçosa, Ouro Preto e Badia: pelo incentivo, carinho e pelos maravilhosos momentos de descontração. O apoio de todos foi fundamental para o desenvolvimento e conclusão dessa tese.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E TERMOS EM INGLÊS

ADH	*Alcohol Dehydrogenase*. Ácool desidrogenase.
beads	Pequenas esferas magnéticas de tamanho de micrômetros usadas nas construções de bibliotecas de sequenciamento.
both	A predição de genes pelo programa Augustus pode ocorrer em ambas as fitas.
bp	*A base pair* (bp) é o par de bases nitrogenadas que conecta fitas complementares de DNA.
CAD	*Cinnamyl-Alcohol Dehydrogenase.* Cinamil ácool desidrogenase.
CBP	*Consolidated Bioprocessing*. Bioprocesso consolidado.
CCD	*Charge-Coupled Device*. Dispositivo de carga acoplada.
CDD	*Conserved Domain Database.* Banco de dados de domínios conservados.
CDS	*Coding DNA Sequence.* Sequência de DNA codificante.
cluster	Grupo.
contigs	Fragmento de sequência de DNA gerado a partir da montagem das *reads.*
ddNTP	*Dideoxynucleotide Triphosphates*. Didesoxirribonucleotídeos trifosfatados.
default	Parâmetro padrão para determinado programa.
DNA	*Deoxyribonucleic Acid.* Ácido desoxirribonucleico.
dNTP	*Deoxynucleotide Triphosphates*. Desoxirribonucleotídeos trifosfatados.
DP	*Dynamic Programming* ou DP. Programação dinâmica.
draft	É o rascunho do genoma.
F	F é a abreviatura de *False*. Parâmetro que considera a informação como falsa.
feature	Termo geral e genérico utilizado para descrever qualquer região genômica com alguma função anotada.
FTP	*File Transfer Protocol*. É uma forma versátil e rápida de transferir arquivos. Protocolo de Transferência de Arquivos.
gbk	*GenBank format*. Formato de sequências do banco GenBank.

GRAS	*Generally Recognized As Safe. Status* conferido a um organismo como geralmente reconhecido como seguro em se tratando de manipulação e segurança alimentar.
high throughput	Alto rendimento.
HMM	*Hidden Markov Model* ou HMM. Modelo Oculto de Markov.
HSP	*High-scoring Segment Pairs* ou HSP. É a significância estatística do alinhamento.
input	*Input* é o arquivo de entrada.
ITS	*Internal Transcriber Spacer.* Espaçador do transcrito interno que separa as regiões 18S e 28S do DNA ribossomal.
kb	A *kilobase pair* (kbp) é a unidade de medida de ácidos nucleicos, igual a 1.000 pares de bases. Kbp ou kb=1.000bp
kDa	É uma unidade de medida de massa atômica, dada em quilodáltons. KDa=1.000Da
Mb	*Megabase pair* (Mbp ou Mb) é a unidade de tamanho dos ácidos nucleicos, igual a um milhão de pares de bases. Mb=1.000.000bp
MDR	*Medium chain Dehydrogenases/Reductases* ou MDR. Redutases-desidrogenases de cadeia média.
NAD^+	*Nicotinamide Adenine Dinucleotide* ou NAD^+. Nicotinamida adenina dinucleotídeo.
$NADP^+$	*Nicotinamide Adenine Dinucleotide Phosphate* ou $NADP^+$. Nicotinamida adenina dinucleotídeo fosfato.
NGS	*Next Generation Sequencing*. Sequenciamento de nova geração.
NR	*Non Redundant* or NR. Banco de dados de proteínas não redundante.
ORF	*Open Reading Frame* ou ORF. Janela aberta de leitura.
output	*Output* é o arquivo de saída.
partial	É a predição de sequências parciais, ao invés, de ocorrer somente a predição de sequências completas.

path	É o caminho utilizado em Linux para chegar ao arquivo ou programa desejado.
PCR	*Polimerase Chain Reaction* ou PCR. Reação em cadeia da polimerase.
PE	*Paired Ends* ou PE. Terminações de leituras em pares.
query	É a sequência de busca.
reads	Leituras individuais de fragmentos de DNA geradas pelo sequenciador.
read length	É o comprimento das *reads* em pb.
scaffold	É a união de *contigs*.
SCP	*Single Cell Protein* ou SCP. Proteína unicelular.
SE	*Single Ends* ou SE. Terminações de leituras únicas.
short read assembly	Montagem através de fragmentos de pequeno tamanho.
SNP	*Single Nucleotide Polymorphim* ou SNP. Polimorfismo único de nucleotídeos.
SOLiD	*Sequencing by Oligonuclotide Ligation and Detection*.
SSF	*Simultaneous Saccharification and Fermentation* ou SSF. Sacarificação e fermentação simultâneas.
subject	É o banco utilizado na comparação com a sequência de busca.
T	T é a abreviatura de *True*. Parâmetro que considera a informação como verdadeira.
TAB	Formato tabular de separação de dados.
UTR	*Untraslated Region* ou UTR. Região não-traduzida ou região não-codante.
WGD	*Whole Genome Duplication* ou WGD. Duplicação do genoma inteiro.

RESUMO

COSTA, Daniela Arruda, D.Sc., Universidade Federal de Viçosa, julho de 2015. **Genômica comparativa de leveduras de interesse biotecnológico.** Orientador: Luciano Gomes Fietto. Coorientadores: Jeronimo Conceição Ruiz e Wendel Batista da Silveira.

O advento das tecnologias de sequenciamento de nova geração possibilitou avanços significativos no campo do conhecimento dos genomas dos organismos. Com a redução dos custos de sequenciamento e aumento na rapidez do processamento das amostras, o número de genomas sequenciados e disponíveis em bancos de dados cresceu exponencialmente. Paralelamente, ocorreu um aumento na propagação de erros associados à anotação das sequências, que gera um impacto negativo em pesquisas posteriores baseadas nestes dados. A revisão da anotação e a comparação entre as informações disponíveis em bancos de dados públicos com experimentos de validação oferece uma alternativa para a correção desses erros. Neste contexto, este trabalho realizou a re-anotação e padronização de 14 genomas de leveduras de interese biotecnológico e disponíveis em bancos de dados públicos. Para facilitar a análise, os genomas foram separados em dois grupos: o grupo 1 (11 linhagens de *Saccharomyces cerevisiae*) e o grupo 2 (3 espécies de *Kluyveromyces*). Após a re-anotação, os proteomas preditos foram submetidos ao algoritmo de agrupamento OrthoMCL, para identificação de grupos de ortólogos e parálogos. As sequências ortólogas desses organismos, juntamente com comparações por similaridade com o banco de dados NR e com o banco de domínios conservados CDD, possibilitou a criação de bancos de dados locais com uma grande riqueza de informações, como *link* de acesso para os resultados do BLAST . Análise dos dados dos genomas permitiu a comparação da família de proteínas álcool desidrogenase (ADH) dentro dos grupos. Na linhagem *S. cerevisiae* JAY 291 não foi possível identificar o gene ADH7. Adicionalmente, o grupo 2 não possui sequências similares ao ADH5.

ABSTRACT

COSTA, Daniela Arruda, D.Sc., Universidade Federal de Viçosa, July, 2015. **Comparative genomics between yeasts of biotechnological interest.** Adviser: Luciano Gomes Fietto. Co-advisers: Jeronimo Conceição Ruiz and Wendel Batista da Silveira.

The advent of new generation sequencing technologies has made capable substancial advances at genomes organism knowledge field. Lower sequencing costs along with faster sample processing, has grown a number of sequenced genomes and available databases exponentially. At the same time, associated annotation sequences errors became more common, causing a negative impact for upcoming researchs based on those databases. Review of annotation and comparison between available public databases with validation experiments offer an alternative solution for error repairs. In this context, this research performed a re-annotation and standarization of 14 yeast genomes of biotechnological interest and available on public database domain. In order to facilitate analysis, genomes were separated in two groups: group 1 (11 *Saccharomyces cerevisiae* strains) and group 2 (3 *Kluyveromyces* species). After their re-annotation, predicted proteomes were submitted to the *cluster*ing algorithm OrthoMCL for identification of orthologues and paralogues groups. The orthologues sequences of those organisms along with similarity comparisons with NR database and conserved domain database CDD, enabled the creation of local databases with an abundance of information, such as acess link to BLAST results. Genome data analysis made possible the comparison of alcohol dehydrogenase family (ADH) within the groups. The *S. cerevisiae* JAY 291 strain was not able to identify gene ADH7. Additionallly, group 2 does not have similarity sequences to ADH5.

1. INTRODUÇÃO

O advento das tecnologias de sequenciamento de nova geração, tecnologias NGS, possibilitou avanços significativos no campo do conhecimento dos genomas dos organismos. Com a redução dos custos de sequenciamento e aumento na rapidez do processamento das amostras, o número de genomas sequenciados e disponíveis em bancos de dados cresceu exponencialmente.

Paralelamente, a enorme quantidade de dados gerados e a falta de recursos humanos capacitados para trabalharem com estas informações representam desafios significantivos para a bioinformática. Dessa forma, armazenar racionalmente os dados, gerenciar soluções computacionais e a análise exploratória dos dados viabilizam a extração de informações valiosas a respeito da sequência genômica do organismo.

É sabido que os dados provêm de diferentes plataformas de sequenciamento, que por sua vez, utilizam diferentes metodologias experimentais, gerando sequências com tamanhos desiguais e em formatos distintos. Ademais, cada plataforma NGS possui vieses associados à metodologia, tal como, erros de nomeamento de bases. Para lidar com esses desafios, houve um grande desenvolvimento de programas e ferramentas para análise de dados, desde a avaliação da qualidade das leituras até a anotação e montagem dos genomas.

Os projetos de sequenciamento e anotação de genomas estão sendo realizados por diversos grupos de pesquisa, que adicionam um grau considerável de informação secundária (anotação estrutural e funcional, similaridades, referências) às sequências proteicas. Todavia, em grande parte dos casos, não ocorre uma padronização quanto ao tipo, quantidade e qualidade da informação fornecida. Além disso, após disponibilizar publicamente os dados, raramente se faz uma revisão sobre as informações acrescentadas.

Neste contexto, a propagação de erros de anotação de sequências torna-se progressiva e gera um impacto negativo em pesquisas posteriores. A revisão da anotação e a comparação entre as informações disponíveis

em bancos de dados públicos com experimentos de bancada oferece uma alternativa para a correção desses erros.

Buscando extrair informações relacionadas aos genomas de leveduras com interesse biotecnológico foi realizada a genômica comparativa entre linhagens disponíveis em bancos de dados públicos. O objetivo principal da genômica comparativa foi a execução de comparações par a par entre todas as sequências proteicas preditas por meio do algoritmo de agrupamento de sequências OrthoMCL (Li et al., 2003).

2. REVISÃO DE LITERATURA

2.1. Sequenciamento de genoma

Tecnologias de sequenciamento de DNA tornaram-se técnicas essenciais para diversas áreas da ciência; trazendo benefícios para as áreas de conhecimento que abrangem arqueologia, antropologia, genética, microbiologia, biotecnologia, biologia molecular e ciências forenses, entre outras (França et al., 2002).

Em 1977, Frederick Sanger e Alan Coulson publicaram dois trabalhos no qual reportavam metodologias eficazes para a determinação da sequência de DNA de organismos, abrindo as portas para uma completa revolução na biologia com o desvendamento da sequência completa de genes e genomas (Sanger et al., 1977; Sanger & Coulson, 1975). Em 1980, Sanger recebeu o segundo Prêmio Nobel de Química por ter realizado o sequenciamento do genoma do bacteriófago ϕx 174 de 5375 nucleotídeos através da técnica de sequenciamento por terminação da cadeia ou método didesoxi. O método de Sanger consiste na incorporação de desoxinucleotídeos e de didesoxinucleotídeos a uma cadeia de DNA em crescimento, tendo como molde o DNA de interesse. Uma vez que os ddNTPs são adicionados, a extensão da cadeia é interrompida pois esses didesoxinucleotídeos não apresentam um grupo hidroxila 3'(OH) necessário para a ligação do próximo dNTP. Como esses ddNTPs são marcados, podem ser detectados e a sequência dos nucleotídeos identificada.

Naquela época, o sequenciamento de Sanger era uma técnica laboriosa e utilizava materiais radioativos, o que estimulou a busca por melhorias na técnica. Depois de anos de pesquisa, a *Applied Biosystems* lançou a primeira máquina de sequenciamento automático (chamada AB370) em 1987, adotando a eletroforese por capilaridade para tornar o sequenciamento mais rápido e acurado. A AB370 podia detectar 96 bases nucleotídicas por vez, 500 kb por dia, e, cada leitura gerada possuía até 600 bases de comprimento. O modelo AB370xl pode gerar 2,88 Mb por dia

e alcançar 900 bases nucleotídicas de comprimento desde 1995 (Liu et al., 2012).

O sequenciamento por Sanger continua sendo a tecnologia que gera dados com maior qualidade, sendo ainda muito utilizado para detectar pequenas mutações nos genomas, os SNPs. Nos estudos de taxonomia e identificação de espécies, essa metodologia é extensamente utilizado através do sequenciamento da região ITS do DNA ribossomal (Chen et al., 2001).

Um estímulo ao desenvolvimento de novas tecnologias de sequenciamento de DNA foi o audacioso projeto de sequenciamento do genoma humano, iniciado em 1990. Este projeto reuniu cerca de 5.000 cientistas de 17 países diferentes com o objetivo de determinar a sequência inteira da eucromatina humana dentro de 15 anos. O avanço das metodologias de sequenciamento, resultado da união de pesquisas do consórcio público com o setor privado Celera Genomics liderado por Dr. Craig J. Venter permitiu que o projeto fosse concluído em 2003 com o sequenciamento de 99% das bases (Chial H., 2008). O projeto genoma humano estimulou o desenvolvimento de poderosas máquinas de sequenciamento que aumentavam a velocidade e a acurácia, enquanto simultaneamente reduziam o custo e a mão de obra (Liu et al., 2012).

As tecnologias de sequenciamento de nova geração, conhecidas como plataformas NGS, revolucionaram a biologia moderna ao permitir o sequenciamento de genomas completos de forma mais rápida, por meio do sequenciamento em paralelo, e a custos mais baixos. Antes do advento das tecnologias NGS, no ano de 2005 com o pirossequenciador 454 da Roche, o sequenciamento de genomas completos exigia a laboriosa etapa de clonagem de fragmentos de DNA e preparação de bibliotecas genômicas, o que tornava o processo dispendioso e demorado (Glenn, 2011; Thudi et al., 2012).

As plataformas NGS utilizam diferentes metodologias de sequenciamento, mas em comum, elas não precisam passar pela etapa de clonagem dos fragmentos de interesse. Adicionalmente, as tecnologias

NGS são capazes de gerar milhões de sequências em uma única corrida. Existem diversas plataformas NGS, dentre elas Roche 454, Illumina, AB SOLiD, Ion Torrent, Ion Proton, Pacbio, Nanopore que estão sendo utilizadas no sequenciamento 'de novo', re-sequenciamento de genoma e análises de genomas completos e de transcriptomas (Glenn, 2011; Liu et al., 2012; Thudi et al., 2012; Mohamed & Syed, 2013).

A plataforma 454 lançada em 2005 pela Roche foi o primeiro sistema de sequenciamento de nova geração a ser comercializada. Ao invés de utilizar ddNTPs para terminar a amplificação da cadeia, a tecnologia de pirosequenciamento depende da detecção de um pirofostato liberado durante a incorporação de nucleotídeos. Nesta metodologia, a biblioteca de DNA ligada a adaptadores específicos para o 454 é desnaturada em fita simples e capturada por microesferas de amplificação seguidas pelo PCR em emulsão (Liu et al., 2012). A leitura da sequência nesse sistema é realizada a partir de uma combinação de reações enzimáticas que se inicia com a liberação de um pirofosfato, oriundo da adição de um desoxinucleotídeo (dATP, dGTP, dCTP ou dTTP) à cadeia. Em seguida, esse pirofosfato é convertido para ATP, pela ATP sulfurilase, sendo este utilizado pela luciferase para oxidar a luciferina, produzindo um sinal de luz capturado por uma câmera CCD acoplada ao sistema (Carvalho & Silva, 2010).

A plataforma Illumina foi a segunda a alcançar o mercado sendo atualmente a tecnologia mais utilizada. O Illumina usa uma abordagem de sequenciamento por síntese no qual os quatro nucleotídeos marcados com fluorescência são adicionados simultaneamente aos canais na superfície de lâmina que contém moldes de nucleotídeos imobilizados, ou seja, sequências de adaptadores ligadas à superfície, juntamente com a DNA polimerase. Dessa forma, a amplificação é feita por meio de pontes que formam grupos. Esta plataforma é reconhecida como a mais adaptável e fácil de usar, com a qualidade superior dos dados e o comprimento adequado das leituras se tornou o método escolhido para muitos projetos de sequenciamento de genoma. O Illumina MiSeq pode gerar um *output* de

15G e leituras 2 X 300 pb (Glenn, 2011; Zhang et al., 2011; Quail et al., 2012; Thudi et al., 2012).

A plataforma NGS IonTorrent PGM, que utiliza a tecnologia de sequenciamento por semicondutor de íons, é capaz de aliar um baixo custo por base sequenciada com a versatilidade dos chips de sequenciamento, com *outputs* de 100Mb, 1Gb e 2Gb de dados gerados por corrida. Atualmente os protocolos de sequenciamento para IonTorrent PGM são capazes de gerar leituras de 200 e 400 pb (Liu et al., 2012; Seo et al., 2015).

Comparado à segunda geração de sequenciadores (454, Illumina, entre outros), o PacBio RS (o primeiro sequenciador apresentado pela PacBio) tem diversas vantagens. Em primeiro lugar, a preparação da amostra é muito rápida, dura entre 4 a 6 horas em vez de dias. Em segundo lugar, as corridas são rápidas, acabando geralmente em um dia. Em terceiro lugar, o comprimento médio de leitura é de 1300 pb, comprimento maior que aquele gerado pela tecnologia de segunda geração. Embora a taxa de transferência dos PacBioRS seja inferior aos sequenciadores de segunda geração, esta tecnologia é muito útil para os laboratórios clínicos, especialmente para pesquisa de microbiologia (Liu et al., 2012).

As plataformas descritas de sequenciamento de nova geração destacam-se entre outras tecnologias NGS por exibirem melhor desempenho e outras vantagens como tamanho das leituras, acurácia, aplicações, requerimento humano e estrutura de informática, dentre outras (Liu et al., 2012).

O acesso às tecnologias NGS está produzindo um aumento exponencial na quantidade de genomas completos sequenciados, no entanto, o processo de análise das sequências não está acompanhando o volume de dados gerados pelas plataformas NGS. Neste contexto, torna-se indispensável o uso de abordagens computacionais visando à integração dos dados gerados pelo sequenciamento em larga escala de DNA com a informação biológica existente e depositada em bancos de dados. A integração dessas duas áreas de conhecimento, biologia computacional e biologia experimental, permite uma melhor compreensão

do organismo biológico e fornece modelos matemáticos que podem direcionar os experimentos de bancada.

2.2. Montagem genômica

Após o sequenciamento do genoma, milhões de leituras são geradas e necessitam ser montadas, de modo a expressar a sequência original contida no genoma do organismo. Essa montagem é realizada através de programas que identificam as sobreposições entre as leituras obtidas, gerando sequências consenso maiores e ordenando-as em *contigs* a fim de obter um rascunho do genoma. Adicionalmente, plataformas NGS diferenciadas possuem erros e viés associados às suas técnicas que dificultam o processo de montagem.

Um viés associado às tecnologias NGS é o tamanho pequeno das leituras geradas, quando comparado ao sequenciamento de Sanger. A metodologia de Sanger gera leituras de até 900 pb, enquanto as plataformas NGS geram no máximo leituras de 700 pb, obtidas pelo 454 da Roche, enquanto que o Illumina gera leituras de até 300 pb. No entanto, apesar do pequeno tamanho das sequências obtidas em NGS, o custo por base sequenciada em NGS é muito menor quando comparado ao método de Sanger, sendo este um dos diferencias das metodologias NGS. Outro importante viés associado às plataformas NGS é o erro na montagem devido à existência de regiões repetitivas no genoma. Estas regiões podem ser mascaradas pelos algoritmos, de modo a produzir uma montagem errônea do genoma. Outros vieses associados às plataformas consistem em: erros de nomeamento de bases, erros de sequenciamento, alto conteúdo CG e mascaramento de adaptadores, dentre outros (Baker, 2012; Henson et al., 2012).

Uma estratégia utilizada para facilitar a montagem do genoma consiste em aumentar a cobertura do sequenciamento. Essa cobertura é realizada para dar uma maior confiança à existência daquela base na sequência do genoma, através do sequenciamento de uma região por repetidas vezes.

Existem diversos programas para montagem do genoma de organismos, cada um com suas vantagens e limitações. Os montadores SSAKE, Edena e Euler-sr necessitam de alta profundidade de cobertura, aproximadamente 50X, enquanto que os montadores Velvet, ABySS e SOAPdenovo precisam de 30X. Dentre os programas de montagem supracitados, tem-se que o SOAP denovo é a ferramenta mais rápida; o ABySS precisa de menos espaço na memória do computador; e o Velvet é um excelente montador, porém possui uma baixa velocidade de processamento (Warren et al., 2007; Zerbino & Birney, 2008; Simpson et al., 2009; Luo et al., 2012).

Com o intuito de melhorar a montagem do genoma várias estratégias computacionais e experimentais podem ser empregadas incluindo: a) verificação da qualidade de sequência nas regiões de junção de *contigs* e *supercontig*s; b) extensão do tamanho trecho de sequência utilizado como critério de corte na sobreposição de *contigs* e *supercontig*s; c) mapeamento das leituras obtidas nos *contigs* e *supercontig*s gerados; d) utilização de sequências geradas por várias bibliotecas de sequenciamento (bibliotecas do tipo *pair-end* e *mate-pair* com tamanhos de inserto diferentes); e) identificação de recombinações genômicas e análise dos dados que não entraram na montagem; e f) utilização de diferentes algoritmos empregando diferentes estratégias de montagem.

Um genoma adequadamente montado permite análises consistentes de genes e regiões regulatórias, sintenia, duplicações e relações evolutivas entre as espécies (Salzberg & Yorke, 2005).

2.3. Anotação de Genomas

Os avanços obtidos com as tecnologias NGS têm gerado uma quantidade muito grande de sequências biológicas, as quais por sua vez necessitam de elevada demanda computacional e profissional especializado para serem corretamente interpretadas, atribuindo significado biológico às informações obtidas pelo sequenciamento do genoma está no escopo da anotação genômica.

O processo de anotação é o processo de interpretação de dados brutos gerados pelo sequenciamento em informação biológica útil. Anotações descrevem o genoma, transformando sequências genômicas brutas em informações biológicas, integrando análises computacionais, experimentos biológicos complementares e conhecimentos biológicos (Lewis et al., 2000). A anotação de genomas é um processo que acontece em multi-passos, podendo seguir ordenadamente três categorias: anotação em nível de nucleotídeos, proteínas e de processos (Stein, 2001). Essas três categorias de anotação podem ser entendidas como onde, o quê e como (Figura 1).

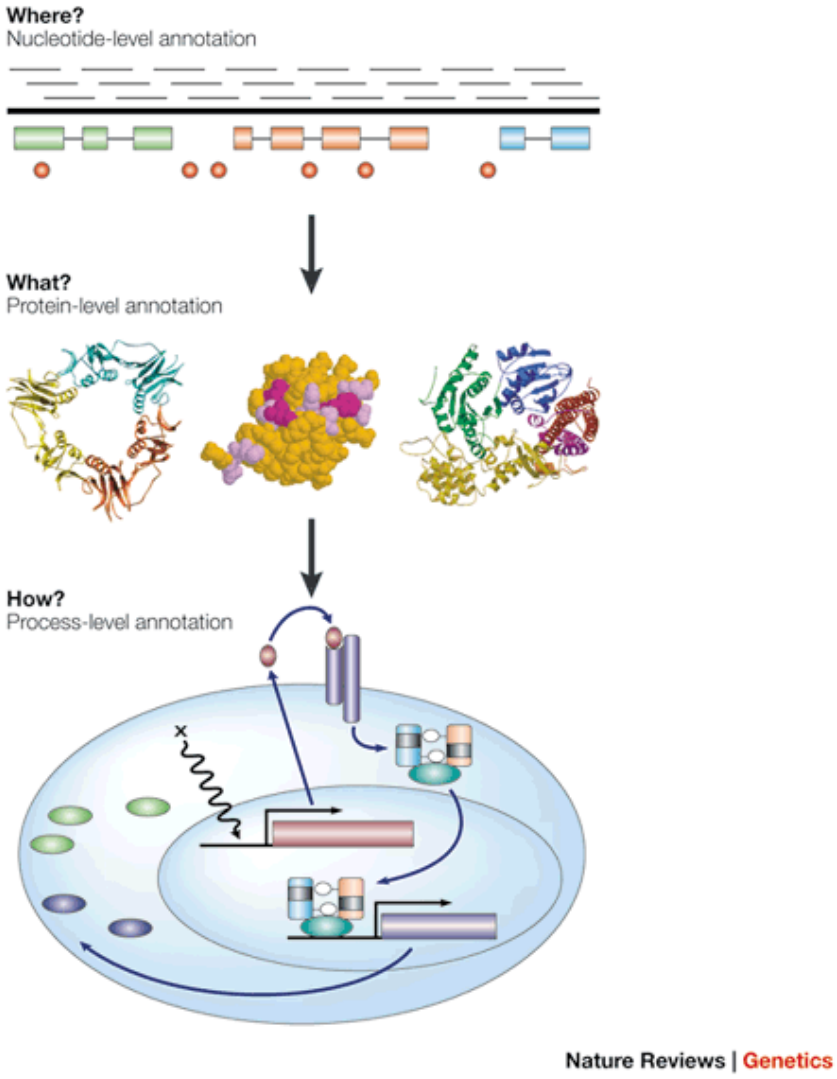


Figura 1: As três camadas de anotação de genoma: onde, o quê e como? (Stein, *et al*. 2001).

Quando o sequenciamento é finalizado, a empresa que o realizou ou o profissional responsável pela geração de dados disponibiliza ao bioinformata uma grande quantidade de pares de bases de DNA em arquivo de formato fastq, porém sem qualquer sentido biológico atribuído. O anotador inicia o laborioso trabalho por meio do mapeamento do genoma, buscando genes conhecidos e marcas genéticas previamente identificadas por validação biológica. Dessa forma, o anotador tenta localizar os tRNAs, rRNAs, RNAs não traduzidos, elementos repetitivos e evidências de duplicação gênica, de forma a mapear fisicamente o DNA para converter a informação bruta em marcos de referência biológica. Destaca-se que ao longo desse processo o mais importante é identificar a localização dessas marcas genéticas (Stein, 2001). A anotação em nível de nucleotídeos pode ser chamada de anotação estrutural, por localizar e delimitar o início e o fim da região anotada.

Finalizada a etapa de anotação estrutural, os anotadores anseiam por saber o que são estas marcas genéticas. Esse estágio de anotação utiliza de buscas em bancos de dados de sequências de nucleotídeos e proteínas dos organismos para nomear as ORFs preditas, atribuindo-lhes possíveis funções (Stein, 2001). A anotação em nível de nucleotídeo e proteína pode ser denominada de anotação funcional, uma vez que ela associa uma função biológica à região anotada.

Dos conjuntos de genes conhecidos, somente uma pequena parte são proteínas bem caracterizadas, o que dificulta o trabalho do anotador. As proteínas são classificadas em grupos ou famílias de acordo com as similaridades entre as sequências das espécies. Entretanto, existe um problema intrínseco a este processo: a evolução. Na natureza, mutações randômicas e recombinação levam a diversidades genéticas. Durante a evolução de uma família de proteína, um gene ancestral pode ser duplicado uma ou mais vezes, com suas cópias divergindo e formando genes ditos parálogos (Stein, 2001).

A importância da duplicação de genes em fornecer material genético para a evolução biológica é reconhecida desde os anos de 1930. Dados

genômicos demonstraram a abundância de genes duplicados nos organismos pesquisados. Lynch e Conery estimaram que genes duplicados surgissem (e se fixaram nas populações) a uma taxa aproximada de 1 gene a cada 100 milhões de anos em eucariotos como *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* e *Saccharomyces cerevisiae* (Zhang, 2003; Lynch & Conery, 2000). Genes duplicados são muitas vezes referidos como parálogos, os quais formam as famílias de genes. Posteriormente, muitos genes duplicados tornam-se pseudogenes, ou seja, sequências de DNA derivadas de genes funcionais e que por sofrerem determinadas mutações tornam-se não funcionais. Algumas metodologias estão ajudando a revelar os mecanismos por trás da evolução dos genes: caracterização das famílias de genes individuais, análise de sequências genômicas e modelagem genética de populações (Zhang, 2003).

Um caso bem relatado na literatura sobre duplicação gênica é o da levedura *S. cerevisiae*. Um ancestral desta levedura passou pelo processo WGD de duplicação de todo o genoma, no qual o número de seus cromossomos dobrou de 8 para 16 cromossomos. Espécies que não passaram pelo evento WGD possuem entre 6 e 8 cromossomos, como aquelas pertencentes ao gênero *Kluyveromyces* sp.. No entanto, houve uma perda maciça de 90% dos genes duplicados através de pequenas deleções, restando cerca de 550 pares de genes duplicados no genoma de *S. cerevisiae* (Kellis et al., 2004; Zhu et al., 2013).

A anotação funcional fornece uma informação rica e útil para estudos comparativos entre diferentes espécies através de proteínas. Por exemplo, uma proteína bem caracterizada de levedura envolvida na inicialização da replicação do DNA, pode ser predita numa sequência de genoma humano com a mesma função. No entanto, as sequências ortólogas podem ter funções diferentes, o que poderia confundir o bioinformata (Stein, 2001). De forma a evitar que informações equivocadas sejam associadas às sequências preditas, o anotador realiza o processo de curadoria manual das sequências. Neste processo, o anotador utiliza informações confiáveis provenientes de bancos de dados curados, dentre eles o Swiss-Prot, que

no ano de 2013 possuía 455.294 sequências de proteínas curadas (O' Donovan et al., 2002).

Relacionar a qual processo as proteínas preditas e anotadas está associado é o maior desafio da anotação do genoma. Os cientistas desejam saber como os genes e proteínas relacionam-se com diversos processos, como a manutenção do ciclo celular, morte celular, embriogênse e metabolismo, dentre outros. Tal como a anotação em nível de proteínas, assim como a anotação em nível de processos biológicos pode ser chamada de anotação funcional, atribuindo significado biológico a região estrutural identificada (Stein, 2001). Até o ano de 2000, não havia uma padronização de normas para realizar a anotação por diferentes grupos de pesquisa, dificultando o intercâmbio das informações obtidas. Assim criou-se o GO, *Gene Ontology*: um vocabulário padrão para descrever a função de genes eucariotos, dividindo-se em: função molecular, processo biológico e componente celular (Ashburner et al., 2000).

Este processo categorizado de anotação, estrutural e funcional, mostra-se uma abordagem mais ampla e integrativa para analisar toda a informação proveniente do sequenciamento do genoma do organismo desejado, sendo uma área do conhecimento muito útil para a compreensão do papel do organismo, sua função biológica e processo evolutivo.

2.4. Erros na Anotação de Genomas

Em 2001, os pesquisadores Devos e Valencia, discutiram os erros intrínsecos na anotação do genoma calculando a estimativa da magnitude de qualquer erro possível de anotação. Segundo esses autores, a predição da função das proteínas para genomas completos estava criando expectativas de um rápido progresso na biologia molecular, sendo um assunto ainda controverso devido a grande quantidade de erros existentes nas sequências anotadas. Neste trabalho, foram demonstrados erros de anotação em 8% dos três genomas analisados, os quais dependendo da definição de função utilizada poderiam chegar a 37% de anotação equivocada (Devos & Valencia, 2001). Outro importante trabalho corrobora

os erros na anotação demonstrando 8% de erros de anotação no genoma de *Mycoplasma genitalium* (Brenner, 1999). Jones e colaboradores observaram que a taxa de erros de anotação em bancos de dados curados também pode ser elevada, no GOSeqLite database, versão de Março de 2006, correspondia entre 28 a 30% dos dados supracitados (Jones et al., 2007).

Embora a análise de genomas completos seja esperada como uma influência positiva no desenvolvimento da biologia e da biomedicina, é importante estar consciente dos possíveis erros preditos pelas anotações que podem ter sido introduzidos por predições funcionais dos padrões durante uma análise preliminar (Devos & Valencia, 2001).

A anotação de genomas é um processo multinível e erros podem aparecer em diferentes etapas; seja o sequenciamento, resultado do procedimento de nomeamento de bases; seja a etapa de atribuição de função biológica aos genes (Poptsova & Gogarten, 2010).

2.5. O porquê de re-anotar genomas

Uma vez que o projeto genoma é finalizado e depositado em bancos de dados de domínio público, é comum que ocorra um segundo olhar por diversos grupos de pesquisa sobre a anotação original. Este processo de anotação sobre um genoma previamente anotado é definido como re-anotação, tendo como motivação: a descoberta de novas funções de genes e proteínas; a avaliação de novos métodos de predição gênica; a avaliação da reprodutibilidade da anotação; e, a diminuição dos erros de anotação (Ouzounis & Karp, 2002).

Este processo é caracterizado por diversos elementos. Como os anotadores não possuem acesso aos dados originais de sequenciamento, podem ser levados a cometer erros que poderiam ter sido detectados com a observação dos dados originais. A re-anotação é um processo intenso, demanda a avaliação do genoma inteiro, envolvendo um grande número de operações manuais, incluindo a correção de erros na anotação original. Além disso, a falta de um padrão ouro para julgar as anotações representa

além de um desafio, um sério problema, porque não haveria garantias de que aquela seria a melhor anotação para o genoma (Ouzounis & Karp, 2002).

Dado um padrão ouro para a anotação dos dados, duas medidas de acurácia precisam estar bem definidas: cobertura e precisão. A cobertura é definida como a taxa de verdadeiros positivos sobre a soma de verdadeiros positivos mais falsos negativos, então se não existe falsos negativos a cobertura é 100%. A precisão é definida como a proporção de verdadeiros positivos sobre a soma de verdadeiros positivos, mais falsos positivos - assim, se não há falsos positivos, a precisão é 100%. Uma medida combinada destes dois valores é exatidão, a qual é definida como a taxa verdadeira de casos (positiva mais negativa) sobre o número total de casos (em que "casos" são, por exemplo, o número de genes ou proteínas). Embora estas medidas nem sempre fossem utilizadas explicitamente nos projetos de anotação do genoma, elas são geralmente utilizadas de forma implícita nos argumentos sobre a forma de predição de acurácia (Ouzounis & Karp, 2002).

Projetos de re-anotação tem sido realizados por diferentes grupos de pesquisa e um dado interessante demonstrando por Ouzounis & Karp (2002) refere-se a média de genes adicionais com funções novas preditas após o processo de re-anotação de genomas fica em torno de 7%. Esta informação acrescentada aos dados públicos é de grande valia aos pesquisadores, e, muitas vezes, estava negligenciada em dados mal analisados depositados em bancos de dados de domínio público.

2.6. Metodologias de Anotação

Embora as metodologias de anotação se diferenciem nos detalhes, elas compartilham um conjunto de propriedades (Yandell & Ence, 2012). O processo de anotação funcional e estrutural da sequência delimita o início e o fim da região anotada e atribuindo-lhe uma função biológica.

De maneira geral, para realizar a predição gênica, duas abordagens são utilizadas: técnicas *ab initio* e baseadas em similaridade. A informação

destes métodos é então combinada por meio de uma variedade de algoritmos, incluindo DP, programação dinâmica ou HMM, modelo oculto de Markov, para predição gênica em sequências genômicas (Do & Choi, 2006).

2.6.1. Anotação *Ab Initio*

Quando os algoritmos preditores se tornaram ferramentas disponíveis na década de 1990, como por exemplo, o *software* GeneScan (Burge & Karlin, 1997), eles revolucionaram a análise de genomas por fornecerem uma metodologia rápida e fácil de identificar genes em sequências de DNA montadas (Yandell & Ence, 2012).

A predição *ab initio* é assim referenciada por utilizar modelos matemáticos para identificar genes e determinar a localização de suas estruturas. A grande vantagem de preditores *ab initio* é que, em princípio, eles não necessitam de evidências externas para a identificação de um gene ou determinação da estrutura de um éxon ou íntron. Neste sentido, por lidar estritamente com a sequência de DNA, a predição *ab initio* pode ser referenciada como um método intrínseco de predição (Do & Choi, 2006).

Dentre as limitações para essa metodologia, destaca-se a capacidade de predição de UTRs, regiões não traduzidas, ou sítios de *splicing* alternativo. O treinamento do algoritmo é outro ponto limitante, os preditores utilizam características genômicas específicas do organismo, como frequências de uso de códon, distribuição de tamanhos de éxons e íntrons, para distinguir regiões gênicas e intergênicas e determinar as estruturas de éxons e íntrons. A maiora dos algoritmos possuem genomas de referência disponíveis que já foram treinados, e, a menos que o genoma que o pesquisador queira trabalhar seja próximo aos genomas disponibilizados pelos preditores, o pesquisador deverá fazer o treinamento do algoritmo (Yandell & Ence, 2012).

Augustus é um algoritmo para predição gênica *ab initio* em sequências eucaróticas baseado no modelo HMM. Pode-se fazer o

download de uma versão do programa no site http://bioinf.uni-greifswald.de/augustus/, para ser utilizado no servidor local ou, se o usuário preferir, pode utilizar as facilidades do programa na Internet. Augustus é baseado no modelo HMM, que define as distribuições de probabilidade para as diferentes seções de sequências genômicas. Íntrons, éxons, regiões intergênicas, entre outros, correspondem a estados no modelo e cada estado é pensado para criar regiões de DNA com determinadas probabilidades pré-definidas, de acordo com parâmetros estatísticos definidos pelo modelo HMM. Desta forma, é possível localizar genes preditos na sequência desejada. Este algoritmo possui um conjunto de anotações de espécies que já foram treinadas para predição gênica, dentre elas, encontram-se as leveduras: *S. cerevisiae* S288c e *Kluyveromyces lactis* (Stanke et al., 2006; Stanke & Morgenstern, 2005; Stanke et al., 2004).

2.6.2. Anotação por similaridade de sequências

Métodos de anotação por similaridade de sequências, ou extrínsecos, utilizam as informações derivadas de similaridade por meio de procedimentos de busca, considerando possíveis proteínas derivadas de uma lista de ORFs como consulta (Robison et al., 1994; Do & Choi, 2006).

Para predição gênica por similaridade de sequências são utilizados algoritmos de busca por similaridade de sequência para realizar comparações entre o genoma, como *query*, e diferentes bancos de dados de domínio público, como *subject* (Altschul et al., 1990).

2.6.3. Anotação Consenso Final

As combinações dos métodos de anotação *ab initio* e baseados em similaridade de sequências são determinantes na eficiência da anotação de genomas em projetos de sequenciamento (Fleischmann et al., 1995).

As abordagens de anotação descritas nos tópicos anteriores podem ser realizadas concomitantemente nos genomas de interesse, de forma a buscar um ganho de anotação proveniente de ambas as metodologias.

Quando o organismo em questão, possuir um conjunto de proteínas anotadas, estas podem ser incorporadas ao processo de re-anotação para obter um genoma com um nível de anotação e curadoria mais elevado.

2.7. Seleção Racional de Espécies

Os bancos de dados de domínio público possuem uma grande quantidade de genomas de organismos depositados em diferentes níveis de anotação e por diferentes grupos de pesquisa, dentre os quais citamos o GOLD, *Genomes OnLine Database* (http://www.genomesonline.org/), e o NCBI, *National Center for Biotechnology Information* (ftp://ftp.ncbi.nlm.nih.gov/genomes/) (Pagani et al., 2011). Existem bancos de dados não curados e curados, sendo que estes últimos passaram por um sistema de conferência manual da anotação por bioinformatas e especialistas nos organismos. Desta forma, os bancos de dados curados possuem informações muito mais confiáveis que os bancos de dados não curados, cabendo aos pesquisadores que consultam os últimos o poder de decisão ao verificar a estringência da informação que lhe é fornecida.

Tal como para outros organismos, existem bancos de dados de sequências biológicas de domínio público específicos para leveduras. O mais conhecido é o SGD, *Saccharomyces Genome Database*, que fornece informações biológicas para a levedura integradas a ferramentas de busca e análise, as quais permitem analisar os dados visando descobrir relações funcionais entre sequências e produtos gênicos em fungos e organismos eucariotos (Informação disponível em http://www.yeastgenome.org/, visualizada dia 10/06/2014) (Cherry et al., 1998). Outro importante banco de dados que está sendo atualizado constamente é o *JGI Fungal Project*, o qual é um banco de dados de genomas de fungos que tem como proposta sequenciar e anotar 1000 genomas de fungos. (http://genome.jgi.doe.gov/programs/fungi/1000fungalgenomes.jsf).

As leveduras possuem um papel crítico no entendimento da função molecular e evolução em eucariotos. Elas são pequenas, possuem genoma compacto, grande importância em vários processos fermentativos e são fáceis de serem manipuladas no laboratório, o que levou a análise e

sequenciamento de vários genomas desses micro-organismos (Brown et al., 2013; Souciet et al., 2009; Argueso et al., 2009; Dujon et al., 2006; Dujon et al., 2005; Goffeau et al., 1996).

A família Saccharomycetaceae é uma grande subdivisão dos hemiascomicetos, representando a maioria das sequências de genomas disponíveis. Esta subdivisão contém uma grande variedade de espécies, que embora compartilhem muitas propriedades fisiológicas e genômicas, representa um amplo espectro filogenético, sendo classificados em 14 clados distintos explorados de forma desigual (Souciet et al., 2009; Kurtzman, 2003; Kurtzman & Robnett, 2003). A levedura *S. cerevisiae* destaca-se como organismo modelo mais estudado, em detrimento de outras espécies dentro desta família, como a levedura *K. marxianus*.

Inserido neste contexto, o objetivo da seleção racional de espécies visa a distinção de um conjunto de genomas de organismos de interesse provenientes de um banco de dados de domínio público para estudar de forma mais aprofundada estes organismos e suas relações.

2.8. Genômica Comparativa

Segundo definição, a genômica comparativa lida com os processos de evolução através do alinhamento e análise de genes e genomas de organismos vivos ou extintos relacionados por diferentes graus de divergência evolutiva a partir de um ancestral comum (visualizado em 21/01/2015 e disponível em http://www.nature.com/*subject*s/comparative-genomics). As comparações são feitas geralmente em pares com referência a um terceiro grupo externo, ou através de observação de pares de sequências por meio de árvores filogenéticas.

O início dos estudos em genômica comparativa foi um marco decisivo para os estudos de evolução molecular, tendo dois tipos de sequências de genomas eram mais comumente comparados. Primeiro, as sequências de espécies proximamente relacionadas, no qual os genomas inteiros podiam ser alinhados em nível de ácidos nucleicos (Waterston et al., 2002). E aquelas sequências de genomas de espécies distantemente

relacionadas, que mostravam interessantes variações no estilo de vida e na fisiologia, mas que eram próximas o suficiente de forma que a maioria dos genes eram ortólogos (Aparicio et al., 2002). Os estudos de genômica comparativa iniciaram-se com a *S. cerevisiae*, visto que esta levedura foi o primeiro organismo eucarioto a ser sequenciado (Zarin & Moses, 2014).

A genômica comparativa de leveduras iniciou-se quando foram publicados o segundo e terceiro genomas de *S. cerevisiae*, as linhagens RM-11a (isolado de vinícola) e YJM789 (patógeno oportunista) respectivamente. Os pesquisadores começaram a investigar o significado funcional da variação genética na escala genômica. Estudos demonstraram que quase 60.000 SNPs e 6.000 inserções/deleções entre YJM789 e S288c, com heterogeneidade de densidade de polimorfismos, ao longo dos cromossomos e dentro de genes específicos (Engel & Cherry, 2013; Wei et al., 2007).

Sabe-se que as linhagens selvagens e laboratoriais de *S. cerevisiae* apresentam variações fisiológicas significativas. Isso pode ser devido ao fato de que os organismos mais adaptados às perturbações ambientais foram selecionados, transmitindo essas possíveis alterações no genoma para as próximas linhagens, o que permite que esse organismo tenha mais sucesso reprodutivo naquele ambiente. Esse raciocínio pode ser extrapolado para o entendimento das diferenças genômicas entre linhagens de *S. cerevisiae* industriais e laboratoriais. Leveduras industriais são constantemente submetidas a diferentes tipos de estresse e possuem determinados padrões genômicos que permitem a sua sobrevivência. Esses padrões podem ser duplicações gênicas, alterações no número de cópias de cromossomos e rearranjos intra ou inter cromossomais mediados por elementos transponíveis. Neste contexto, a análise comparativa de genomas de leveduras surge como uma poderosa ferramenta na busca por padrões genômicos que possam estar envolvidos no caráter industrial ou laboratorial destes organismos.

2.8.1. OrthoMCL

Para comparar relações evolutivas e funcionais entre genes de diferentes espécies é útil comparar ortólogos. Ortólogos são genes cujas sequências mais se assemelham a de um ancestral comum, mas que divergiram por especiação. Genes ortólogos mantém a mesma função em duas espécies que divergiram. Um exemplo de ortologia é a subunidade catalítica da DNA polimerase: POL2 em *S. cerevisiae* e POLE em humanos (Gibney et al., 2013; D'Urso & Nurse, 1997).

O conhecimento de todos os ortólogos de um gene é útil de muitas maneiras. Em primeiro lugar, o conhecimento de genes ortólogos pode ser utilizado para inferir funções a genes de diferentes espécies. Por exemplo, é sabida a função da POL2 e caso exista algum gene ortólogo que se agrupe no mesmo grupo da POL2, pode-se inferir que este gene tenha função na DNA polimerase. Em segundo lugar, análises de ortólogos podem ilustrar o contexto no qual o gene é importante. Se um gene ortólogo é conservado em todos os organismos, ele deve possuir um papel crucial na biologia básica da célula, como por exemplo, um gene conservado em fungos pode ser importante para um processo específico de fungos, como a formação de esporos. Em terceiro lugar, o conhecimento das espécies no qual um gene é conservado tem aplicações médicas. Caso um gene seja específico para fungos, a proteína codificada pode ser um candidato atrativo para drogas anti-fúngicas (Gibney et al., 2013).

Uma ferramenta que pode ser usada na genômica comparativa é o algoritmo OrthoMCL, que permite a identificação de grupos ortólogos entre os genomas. O OrthoMCL é uma ferramenta capaz de construir agrupamentos de ortólogos e parálogos entre dois ou mais organismos. Essa ferramenta é utilizada para encontrar regiões que preservam o conteúdo gênico em vários genomas, a partir de comparações feitas "todos contra todos" pelo blastp. O OrthoMCL utiliza o blastp para comparações de sequências e o algoritmo de agrupamento de Markov (Markov *cluster* – MCL) para a formação de grupos, o qual é baseado em probabilidade e teoria dos grafos (Li et al., 2003).

2.8.2. Artemis Comparison Tool (ACT)

Outra ferramenta muito utilizada na genômica comparativa é o *Artemis Comparison Tool* (ACT) (Carver et al., 2005). O ACT é um programa escrito em linguagem Java que permite visualizar pares de sequências, incluindo genomas completos. A visualização da comparação é feita através de um arquivo de comparação (resultado de uma busca por similaridade de sequências entre os organismos em estudo) e dos genomas anotados e de interesse, possibilitando a identificação de regiões sintênicas, inversões e rearranjos.

O ACT possui muitas vantagens: comparação de mais de dois genomas simultaneamente; leitura de diferentes formatos de sequências (EMBL ou GENBANK); flexibilidade do *zoom* para visualizar comparação do genoma completo; flexibilidade do *zoom* para observar finas escalas de comparações nas sequências de DNA ou proteínas; possibilidade de adição ou edição de anotações dos genomas comparados (Carver et al., 2005; Edwards & Holt, 2013).

2.8.3. Família Proteica de Interesse

Uma das abordagens da genômica comparativa deste trabalho foi o estudo de uma família proteica de interesse biotecnológico, especialmente no processo de produção de etanol.

2.8.3.1. ADHs: a importância das álcoois desidrogenases

Álcoois desidrogenases (ADHs) constituem uma grande família de enzimas responsáveis pela reversão da oxidação de álcoois a aldeídos com a concomitante redução de NAD^+ ou $NADP^+$. Estas enzimas foram identificadas não somente em leveduras, como em diversos outros eucariotos e em procariotos. Os ADHs de *S. cerevisiae* têm sido estudados há mais de meio século, e a disponibilização de genomas completos possibilitou uma maior compreensão sobre estas enzimas (De Smidt et al., 2008).

Fisiologicamente, a reação enzimática da família ADH possui um papel crítico no metabolismo de açúcar em *S. cerevisiae* e em outros organismos relacionados. Quase todo o carboidrato é utilizado fermentativamente, independente da disponibilidade de oxigênio, e uma isoenzima específica de ADH serve para regenerar o NAD^+ glicolítico, assim restaurando o balanço redox, através da redução de acetaldeído a etanol. Em condições aeróbicas, após o esgotamento de açúcar fermentável, ocorre a respiração do etanol acumulado novamente através de isoenzimas específicas de ADH. Dessa forma, em *S. cerevisiae* a ligação entre o metabolismo fermentativo ou respirativo (oxidativo) se dá através de uma reação por ADH, permitindo a utilização ótima da fonte de carbono (De Smidt et al., 2008).

Os ADHs (E.C. 1.1.1.1) são oxido-redutases que catalisam a oxidação reversível de álcoois a aldeídos ou cetonas, com a concomitante redução de NAD^+ ou $NADP^+$. Eles podem ser divididos em três superfamílias distintas: MDR, redutases-desidrogenases de cadeia média; redutases-desidrogenases de cadeia curta; e ADHS ativados por ferro (De Smidt et al., 2008; Kallberg et al., 2002). Genes ADHs clássicos incluem: *ADH1*, *ADH2*, *ADH3*, *ADH4*, *ADH5*, *ADH6* e *ADH7* (De Smidt et al., 2008).

Células eucarióticas possuem uma organização em membranas e compartimentos, os quais são especializados para várias funções biológicas. O conhecimento da localização celular destas enzimas é de extrema importância para o entendimento de suas funções e interações. Sabe-se que Adh1p e Adh2p estão localizadas no citosol; Adh3p e Adh4p encontram-se na matriz mitocontrial; Adh5p localiza-se no citoplasma; e os Adh6p e Adh7p ainda não possuem sua localização celular definida (De Smidt et al., 2008; Huh et al., 2003; Drewke et al., 1988; van Loon &Young, 1986).

ADH1 de levedura foi uma das primeiras enzimas a ser cristalizada (Negelein et al., 1937). Em leveduras, o gene *ADH1* constitutivo catalisa a redução de acetaldeído a etanol durante a fermentação da glicose. Esta enzima possui estrutura tetramérica composta por quatro subunidades

idênticas com 347 resíduos de aminoácidos cada (Savarimuthu et al., 2014).

ADH2 possui um alto grau de similaridade com ADH1, 90% em nível nucleotídico e 95% em nível de aminoácidos. As sequências de aminoácidos de ADH1 e ADH2 têm somente 22 resíduos de diferença de um total de 347 aminoácidos, e, não existem diferenças nos grupos diretamente envolvidos na catálise (Ganzhorn et al., 1987).

ADH3 possui identidade, em nível de sequência nucleotídica, com *ADH1* e *ADH2*, de 73% e 74% respectivamente (Young & Pilgrim, 1985). A enzima ADH3 possui estrutura tetramérica e a similaridade dos seus aminoácidos comparada a ADH1 e ADH2 foi identificada como 79% e 80%. Todo o sítio ativo, resíduos identificados à ligação de cofatores e ligação não catalítica ao Zinco estão conservados em *ADH1*, *ADH2* e *ADH3* (Jornvall, 1977). A estrutura e sequência amino terminal de ADH3 contribuem para a localização mitocondrial, importação e processamento (De Simdt et al., 2008; Mooney et al., 1990).

ADH4 é o gene mais distal no braço esquerdo do cromossomo VII e está situado próximo ao telômero, o que se torna interessante, pois genes localizados nessa extremidade codificam enzimas envolvidas na fermentação e glicólise (Walton et al., 1996; Mortimer & Schild, 1985). Análises de sequenciamento revelaram que ADH4 não é homólogo a outros ADHs de leveduras, estando distantemente relacionado às sequências de ADHs caracterizadas em outros eucariotos (Paquin & Williamson, 1986). ADH4 é uma proteína dimérica que ocorre em baixas concentrações em linhagens laboratoriais. Tal como as outras ADHs, ADH4 é ativada por íons de zinco (Drewke & Ciriacy, 1988).

ADH5 foi identificado pela primeira vez por meio do sequenciamento do cromossomo II de *S. cerevisiae*, e compartilha identidade de 76%, 77% e 70% com *ADH1*, *ADH2* e *ADH3*, respectivamente (Ladriere et al., 2000; Feldmann et al., 1994).

ADH6 foi o primeiro produto gênico caracterizado em *S. cerevisiae* como ADH de cadeia-média NADPH-dependente (Larroy et al., 2002a; González et al., 2000). A proteína ADH6 possui estrutura heterodimérica com duas subunidades de 40 kDa (Valencia et al., 2004). *ADH6* exibe conservação da assinatura de zinco, bem como sequências de aminoácidos e domínios de ligação de coenzima característicos enzimas MDR contendo zinco (Larroy et al., 2002a; González et al., 2000).

ADH7 possui 64% de identidade com ADH6, produto gênico com conformação homodímera do qual a atividade redutase é cerca de 5 vezes a atividade desidrogenase (Larroy et al., 2002b). ADH7 é uma família de enzimas estruturalmente relacionadas a ADHs cinamil (Larroy et al., 2002a,b).

3. OBJETIVO

O presente projeto objetivou realizar uma comparação entre genomas de leveduras de interesse biotecnológico por meio da re-anotação dos genomas e estudo de família proteica.

3.1. Objetivos Específicos

- Selecionar racionalmente organismos a serem estudados;
- Re-anotar dos genomas;
- Comparar os grupos de genes ortólogos e parálogos dos organismos selecionados;
- Criar um banco de dados local;
- Comparar família de proteínas que possa ser alvo de melhoramento genético.

4. MATERIAL E MÉTODOS

4.1. Seleção Racional de Espécies

Em Novembro/2013 foi realizada uma busca no banco de dados de domínio público do NCBI por genomas pertencentes à família Saccharomycetaceae. A palavra-chave utilizada no site do NCBI foi Saccharomycetaceae e a busca foi reduzida a Genome. Foi realizado o *download* dos arquivos contendo as sequências no formato de extensão GBK, escolhendo a opção gbk full, para o servidor local. Os dados foram transferidos para o servidor local utilizando o protocolo FTP do NCBI.

O *script* do Velvet assembly_stats.pl foi necessário para gerar dados qualitativos a respeito da montagem dos genomas. Segue a linha de comando do programa:

```
$ assembly_stats.pl [arquivo_entrada.mutifasta] >[arquivo_saída]
```

Onde:

- arquivo_entrada.mutifasta é o arquivo de união de toda a sequência do genoma, seja em *contig*s ou cromossomos.

Posteriormente, procedeu-se à escolha dos organismos a serem estudados de acordo com os critérios estabelecidos, quais sejam: interesse biotecnológico, aplicação industrial ou laboratorial, e sequência completa do genoma.

4.2. Tratamento dos Dados e Criação de Banco de Dados Local

Os arquivos contendo as sequências dos genomas foram formatados utilizando a ferramenta formatdb, um *script* do algoritmo BLAST que permite formatar as sequências de proteínas ou nucleotídeos para criar um banco de dados local. No entanto, o *script* formatdb somente aceita dados de entrada com formato do tipo FASTA ou ASN.1. Por conseguinte, foi necessário fazer a conversão do formato GBK para o formato FASTA, através do programa de conversão de formatos *reads*eq (Gilbert, 2003), como indicado na linha de comando abaixo:

```
$ java -jar readseq.jar -inform gbk -f fasta -o [nome_arquivo_saída.fasta] [arquivo_entrada.fasta]
```

Onde:

- -inform é o nome do formato de entrada dos dados, no caso foi gbk;
- -f é o nome do formato de saída dos dados, no caso foi fasta;
- -o é o redirecionamento da saída, deve-se colocar o nome do arquivo de saída a ser escrito e o arquivo de entrada.

Em seguida, procedeu-se à execução do formatdb através da seguinte linha de comando:

```
$ formatdb -i [arquivo_entrada] -p F -n [nome_arquivo_sem_extensão]
```

Onde:

- -i é o arquivo de entrada, no caso a sequência do genoma em formato FASTA;
- -p é o tipo de arquivo, o usuário deve selecionar F se a sequência for de nucleotídeos ou T se for de proteínas;
- -n é o nome base para os arquivos a serem criados, neste caso não é adicionado o formato de extensão do arquivo.

Adicionalmente, os genomas foram formatados através de *scripts* em Perl, para corrigir e padronizar informações.

Durante o desenvolvimento do projeto foi necessário escrever vários *script*s em linguagem de programação PERL, os quais foram essenciais para a manipulação e análise dos dados.

4.3. *Pipeline* para anotação e re-anotação funcional e estrutural

Os genomas selecionados e formatados foram submetidos ao processo de anotação ou re-anotação, com o *pipeline* de anotação seguindo os seguintes passos: predição *ab initio*, predição por similaridade de sequências e anotação consenso final.

4.3.1. Predição *ab initio*

A predição gênica *ab initio* foi realizada localmente no servidor através da utilização do programa Augustus, sendo o algoritmo treinado para os organismos *S. cerevisiae* e *K. lactis*. Segue a linha de comando executada para o Augustus:

```
$ augustus --strand=both --genemodel=partial --species=[espécie] [arquivo_entrada.fasta] > [arquivo_saída.gff]
```

Onde:

- --strand= é a predição de genes em ambas as fitas, somente na fita direta ou na fita reversa. Foi escolhido *both*;
- --genemodel= é o tipo de predição de genes permitida. Foi escolhida a *partial*, que permite a predição de genes incompletos nos limites da sequência. Essa opção é o *default* do Augustus;
- --species= é o identificador das espécies, no projeto foram utilizados os identificadores saccharomyces_cerevisiae_S288C e kluyveromyces_lactis, para as espécies *S. cerevisiae* e *K. lactis*, respectivamente.

Necessariamente, o Augustus utiliza como dados de entrada um arquivo em formato FASTA, retornando como saída um arquivo em formato GFF. O arquivo de saída de formato GFF foi submetido ao *script* coord-CDS_from_augustus_to_art.pl. Este *script* foi desenvolvido pelo Grupo Informática de Biossistemas FIOCRUZ/Minas com o objetivo de obter um arquivo com as coordenadas das CDS obtidas pela predição do Augustus, com o formato de entrada tabular adequado ao programa de visualização de dados Artemis.

4.3.2. Predição por similaridade de sequências

A predição por similaridade de sequências foi realizada através do algoritmo BLAST contra o banco de dados curado do Swiss-Prot. Esse

banco de dados de proteínas foi formatado usando o formatdb (conforme linha de comando descrita anteriormente), escolhendo a opção -p T para tipo de arquivo.

Anteriormente à execução do BLAST, os genomas foram divididos em fragmentos de 40.000 pb sem sobreposição de sequências utilizando o *script* desenvolvido em linguagem PERL split-bigfasta.pl. Como mostra a linha de comando:

```
$ split-bigfasta.pl [arquivo_entrada.fasta] [tamanhos_fragmentos] [sobreposição] > [arquivo_saída.frags]
```

Onde:

- O usário deve fornecer o arquivo de entrada em formato fasta, o tamanho dos fragmentos e a sobreposição, de forma a escrever um arquivo de saída com o arquivo fragmentado. Como boas práticas de programação o nome da extensão se refere ao arquivo FRAGS.

Após obter o arquivo fragmentado com extensão FRAGS, executou-se o formatdb, escolhendo a opção -p F, uma vez que este arquivo era de nucleotídeos. O passo seguinte foi realizar o BLAST do genoma formatado contra o Swiss-Prot formatado, de acordo com a seguinte linha de comando:

```
$ blastall -p tblastn -d [caminho_do_arquivo_formatado] -i [caminho_do_banco_de_dados_formatado] -o [nome_arquivo_saída] -v 5 -b 5 -e 0.000001 -a 30
```

Onde:

- -p é o tipo de programa de BLAST a ser executado. Foi escolhido o tblastn, que realiza comparações entre um *subject* de nucleotídeos traduzidos contra uma *query* de proteínas;
- -d é o *path* do *subject*. Para esta análise, o banco de dados foi o genoma previamente fragmentado e formatado;

- -i é o arquivo de entrada, a *query*. Foi utilizado o banco de dados do Swiss-Prot formatado como *query*;

- -o é o arquivo de saída a ser gerado. O nome desse arquivo indicava os parâmetros utilizados para rodar o programa;

- -v é o número de linhas no qual aparece a descrição dos *hit*s que será escrito no arquivo de saída. Para esta análise foi determinado 5 linhas;

- -b é o número de linhas de alinhamentos que irão ser escritos no arquivo de saída. Para esta análise foi determinado 5 linhas;

- -e é o *evalue*, valor de *score*, ou seja, probabilidade do *hit* acontecer ao acaso. Este valor determina o quão confiável é a similaridade da sequência. Foi escolhido o valor 10^{-6}, que é um valor muito estringente;

- -a é o número de processadores utilizados no processo. Para tornar a análise mais rápida foram usados 30 processadores.

Esses passos resultaram em arquivos contendo as sequências gênicas preditas no genoma contra o banco de dados Swiss-Prot. No entanto, estes arquivos precisavam ter suas coordenadas corrigidas e estar no formato TAB para serem lidos pelo Artemis.

Procedeu-se à utilização de três *script*s escritos em conjunto pelo Grupo Informática de Biossistemas para proceder à análise dos dados. O primeiro *script*, blast_parser_4_extract_coord_splittedDB.pl, teve como propósito o processamento de resultados BLAST obtidos da comparação do banco de dados Swiss-Prot contra um genoma divido em fragmentos de 40.000 pb sem sobreposição. O segundo *script*, correct_coord_from_blast_parser.pl, objetivou o processamento do resultado gerado pelo blast_parser_4_extract_coord_splittedDB.pl. O *script* corrigiu as coordenadas geradas pelo BLAST, levando em consideração que o genoma havia sido fragmentado, de forma a gerar as

coordenadas reais dos *hit*s obtidos. O outro objetivo desse *script* foi remover a redundância dos resultados gerados pelos BLAST, gerando um arquivo de formato tabular. De posse deste arquivo procedeu-se à anotação dos produtos gênicos encontrados por meio do terceiro *script*: anota-genome.pl. Por fim, este *script* anotava as coordenadas das CDS, previamente corrigidas, obtidas pelo BLAST no arquivo de formato tabular e atribuía as informações seguintes:

- Product: é o produto gênico, possível proteína predita pela análise. Neste campo além do produto gênico, consta a informação referente ao organismo no qual foi encontrada a sequência;

- Similarity é a similaridade entre a sequência predita e a encontrada no banco de dados. A similaridade é apresentada através do número identificador da proteína, valor de evalue, cobertura, similaridade, identidade e número do HSP, que é a signficância estatística do alinhamento;

- Blast_macth é o domínio eletrônico correspondente ao *hit* da proteína encontrada;

- Colour é o número correspondente à cor em que aparece o termo anotado. Neste projeto foram escolhidas *core*s diferentes para as CDS anotadas por predição *ab initio*, por similiaridade e consenso final, de forma a facilitar a visualização dos dados por meio do Artemis;

- Controlled_curation é o controle que o curador manual pode fazer das informações preditas. O curador anota nesse campo a existência ou não de informações conflitantes.

- Curation é o grupo responsável pela curadoria. Como o projeto foi uma colaboração entre a Universidade Federal de Viçosa e o Centro de Pesquisas René Rachou, este campo retorna a seguinte informação: “by Fiocruz- MG and UFV”;

- Note é o campo no qual o anotador pode acrescentar alguma informação;

- Systematic_id é o nome dado a cada termo anotado. Padronizou-se pelo nome do organismo, seguido da linhagem e posteriormente um número de 6 dígitos, que era único e específico para cada termo anotado.

4.3.3. Anotação consenso final

A anotação consenso final foi realizada através da integração dos arquivos obtidos pela predição *ab initio*, predição por similaridade e arquivos de anotação original dos genomas. Estes arquivos foram unidos em um arquivo único através do Artemis para ser processado.

A anotação final dos produtos gênicos se deu através do algoritmo BLAST contra o banco de dados do NR. Com o objetivo de não viciar a predição gênica foram retiradas as sequências relativas ao organismo estudado do banco de dados NR. Os bancos de dados NR sem as sequências dos organismos estudados, *S. cerevisiae* e *K. lactis*, foram criados a partir de listas contendo os identificadores únicos relativos a cada proteína anotada e depositada do organismo. Para isso foi feito o *download* dos arquivos de proteínas de formato FAA ou FA, fasta aminoácidos, a partir do FTP NCBI.

Dois *script*s foram escritos pelo Grupo Informática de Biossistemas para processar estes dados: remove-redundancia_genome-annotation.pl e anota-genome.pl. O primeiro *script* era responsável por remover a redundância de coordenadas que existia no arquivo e realizar um novo BLAST utilizando o banco de dados NR sem o organismo de referência. O tipo de programa BLAST definido foi o blastx, que utiliza como *query* o genoma fragmentado e formatado nas 6 fases de leitura traduzidas, e, como *subject* o banco de dados de proteínas NR sem o organismo de referência.

Posteriormente, estes arquivos de BLAST foram processados para produzir arquivos em formato tabular, contendo a anotação consenso final por meio do *script* anota-genome.pl. A visualização e comparação dos dados foi realizada por meio do Artemis (Art) e Artemis Comparison Tool (ACT).

4.4. Genômica Comparativa - OrthoMCL

A genômica comparativa para predição de ortologia foi realizada através do algoritmo OrthoMCL, sendo os organismos estudados divididos em dois grupos. O primeiro grupo continha as dez linhagens de *S. cerevisiae* e a linhagem de *S. boulardii*, e o segundo continha as espécies *Kluyveromyces* sp. de interesse: *K. lactis*, *K. marxianus* e *L. thermotolerans*.

Para processar os dados no OrthoMCL foi necessário criar arquivos separados dos proteomas preditos de todos os organismos em formato FA, fasta aminoácidos. Para evitar erros de processamento, ajustes adicionais foram feitos nestes arquivos. Estes ajustes corrigiam o cabeçalho das proteínas, deixando somente o identificador único de cada uma delas e retirando toda a parte da descrição da proteína. Para este tipo de correção foi escrito um programa de uma linha que removia toda informação que não era necessária, como é demonstrado na seguinte linha de comando:

```
$ for i in `ls *.fa`; do perl -pi -e 's/ .*//g' $i ; done
```

Primeiramente, foi necessário ajustar os paths dos programas usados pelo OrthoMCL, assim como o path o diretório dos dados de entrada, no arquivo orthomcl_module.pm. A execução do ORTHOMCLV1.4 ocorreu localmente através da execução da linha de comando:

```
$ orthomcl.pl --mode 1 --fa_files [arquivo_entrada.fa, arquivo_entrada.fa]
```

Onde:

- --mode é o modo de processamento do OrthoMCL. Neste projeto foi utilizado o 1 porque as análises foram feitas a partir

de arquivos em formato FASTA, iniciando pela execução do BLAST até o agrupamento pelo MCL;

- --fa_files são os arquivos em formato FA a serem agrupados. Um arquivo deve ser separado por vírgula do arquivo seguinte.

A análise das informações geradas se deu através de programas de parseamento. O arquivo all_orthomcl.out foi submetido ao processamento através do *script* orthomcl_table.pl, de forma a escrever um arquivo para cada c*luster* gerado pelo OrthoMCL. O comando foi executado da seguinte forma:

```
$ orthomcl_table.pl -i 1-taxa.txt  -o clusters-1-taxa.txt -f all.fa
```

Onde:

- -i é o *input*, no caso é o arquivo de saída do OrthoMCL;
- -o é o *output* que será gerado pelo *script*. Este arquivo contém os *hiperlinks* para os *clusters* gerados;
- -f é o arquivo de todas as sequências de entrada unido em um único arquivo em formato fasta, all.fa.

Após este passo, foi realizada uma checagem quanto à existência de identificador da espécie que aparece entre parênteses. Esta identificação foi removida utilizando o seguinte *script*:

```
$ for i in `ls *.txt`; do perl -pi -e s'/^.* //g' $i ; done
```

Para obter as sequências do OrthoMCL foi utilizado um programa de extração executado em *shell script* e escrito em *bash*. O *script* extract.pl extrai as sequências das proteínas através dos identificadores gerados no arquivo de saída, como mostra o comando:

```
$ for i in `ls *.txt` ; do ../../bin/extract.pl -i ../all.fa -o $i.fasta -l $i ; done &
```

Onde:

- -i é o input, arquivo que contém todas as sequências dos proteomas preditos all.fa;
- -o é o output, arquivos contendo as sequências das proteínas buscadas;
- -l é a lista contendo os identificadores das proteínas.

De posse dos arquivos dos grupos de proteínas em formato fasta, procedeu-se à execução de BLAST contra o banco de dados NR e contra o banco de dados CDD, como mostram as linhas de comando:

```
$ for i in `ls *.fasta`; do blastall -p blastp -d /storage/database/nr/nr -i $i -o $i--vs--nr.blastp_out-e6-v5-b5 -e 0.000001 -v 5 -b 5 -a 30; done &
```

```
$ for i in `ls *.fasta`; do rpsblast -i $i -d /storage/database/cdd/Cdd -o $i--vs--cdd.rpsblast_out-e6-v5-b5 -p T -e 0.000001 -v 5 -b 5 -a 30; done &
```

4.5. Banco de Dados Local

As informações obtidas através da re-anotação dos genomas e dos agrupamentos pelo OrthoMCL foram integradas e organizadas na forma de banco de dados local. Para tanto, foram criadas planilhas hiperlincadas para o EXCEL no sistema operacional do Windows utilizando os *script*s: separa-blast-v2.pl, make-hiperlinked_table.pl e split-fasta-hiperlinked_table.pl.

4.6. Comparação Filogenética

Foi realizado o *download* das sequências de proteínas dos ADHs de *S. cerevisiae*, ADH1 a ADH7, a partir do banco de dados Saccharomyces Genome Database (http://www.yeastgenome.org/). Estas sequências curadas pelo SGD foram utilizadas para realizar buscas por similaridade de sequências através do algoritmo BLAST nos genomas selecionados, por meio da execução da seguinte linha de comando:

```
$ blastall -p tblastn -d [caminho_do_banco_formatado] -i [caminho_do_arquivo] -o [nome_arquivo_saída] -v 40 -b 40 -e 0.000001 -a 30 &
```

Onde:

- -p é o tipo de programa selecionado de BLAST, foi escolhido o tblastn por se tratar de uma pesquisa em bancos de dados de nucleotídeos traduzidos a partir de sequências de proteínas;
- -d é o banco de dados, arquivos formatados referentes ao genoma do organismo;
- -i é o input, arquivo de união das proteínas ADHs curadas do SGD em formato fasta aminoácidos.

Após a execução do BLAST foram obtidos os arquivos contendo as sequências de ADHs com redundância de informações. Procedeu-se à remoção da redundância e anotação da *feature*, por meio dos *script*s remove-redundancia_genome-annotation.pl e anota-genome.pl, propiciando a visualização dos termos anotados pelo Artemis. A linha de comando acima foi executada duas vezes para cada organismo, modificando os valores de *e-value*: 10^{-4} e 10^{-6}.

As sequências anotadas de acordo com o SGD foram comparadas às sequências anotadas de acordo com o *pipeline* de anotação e com as sequências originalmente anotadas no genoma. Foram adicionadas notas às sequências curadas manualmente através dessa comparação. As sequências que não haviam sido preditas anteriormente foram adicionadas ao genoma final anotado.

O identificador das sequências curadas foi pesquisado nos agrupamentos obtidos através do algoritmo OrthoMCL de forma a retornar os *clusters* onde se agruparam tais sequências.

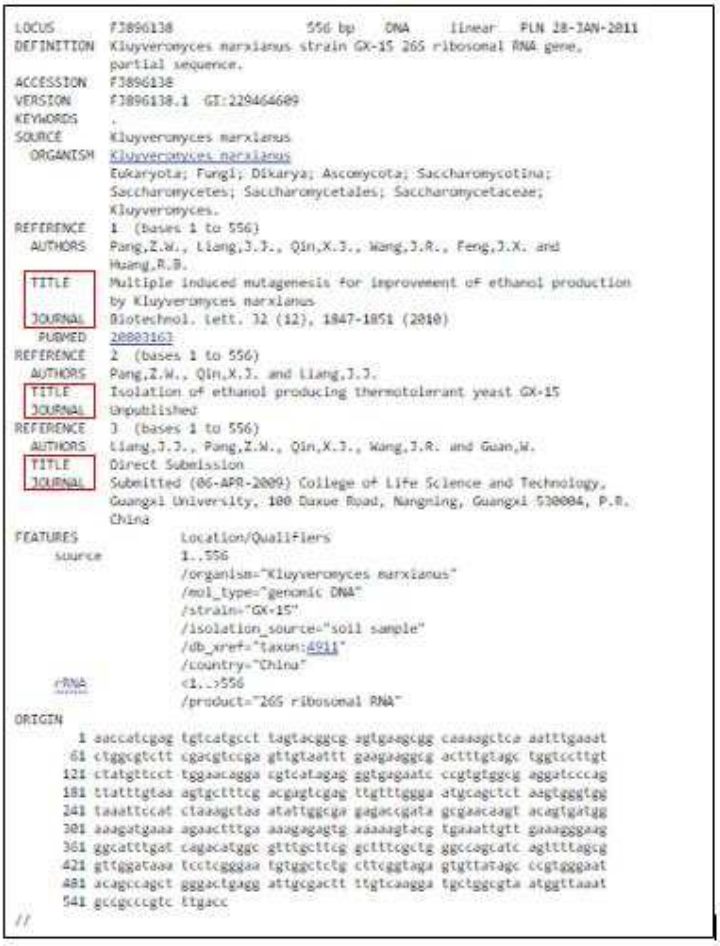
Posteriormente, essas sequências foram submetidas a alinhamentos através da plataforma CLC Workbench. Os alinhamentos

múltiplos foram feitos para cada grupo de ADH anotado de forma a comparar estas sequências em diferentes organismos. Para tanto foi utilizado o *Muscle* e o *Clustalw*. Em seguida, procedeu-se à criação de árvores filogenéticas pelo alinhamento obtido, utilizando o método de *Neibhor Joining*.

5. RESULTADOS E DISCUSSÃO

5.1. Seleção Racional de Espécies

O banco de dados de domínio público GenBank hospedado no NCBI foi escolhido para realizar as buscar das sequências dos genomas de leveduras Saccharomycetaceae a serem analisadas. O NBCI disponibiliza os dados em diferentes formatos para *downloads*. Com o intuito de obter uma maior quantidade de informações relacionadas aos dados, optamos por sempre fazer o *download* das sequências no formato mais completo, que é o GBK full (Figura 2).



```
LOCUS       FJ896138               556 bp    DNA     linear   PLN 28-JAN-2011
DEFINITION  Kluyveromyces marxianus strain GX-15 26S ribosomal RNA gene,
            partial sequence.
ACCESSION   FJ896138
VERSION     FJ896138.1  GI:229464609
KEYWORDS    .
SOURCE      Kluyveromyces marxianus
  ORGANISM  Kluyveromyces marxianus
            Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
            Saccharomycetes; Saccharomycetales; Saccharomycetaceae;
            Kluyveromyces.
REFERENCE   1  (bases 1 to 556)
  AUTHORS   Pang,Z.W., Liang,J.J., Qin,X.J., Wang,J.R., Feng,J.X. and
            Huang,R.B.
  TITLE     Multiple induced mutagenesis for improvement of ethanol production
            by Kluyveromyces marxianus
  JOURNAL   Biotechnol. Lett. 32 (12), 1847-1851 (2010)
   PUBMED   20803163
REFERENCE   2  (bases 1 to 556)
  AUTHORS   Pang,Z.W., Qin,X.J. and Liang,J.J.
  TITLE     Isolation of ethanol producing thermotolerant yeast GX-15
  JOURNAL   Unpublished
REFERENCE   3  (bases 1 to 556)
  AUTHORS   Liang,J.J., Pang,Z.W., Qin,X.J., Wang,J.R. and Guan,W.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-APR-2009) College of Life Science and Technology,
            Guangxi University, 100 Daxue Road, Nanning, Guangxi 530004, P.R.
            China
FEATURES             Location/Qualifiers
     source          1..556
                     /organism="Kluyveromyces marxianus"
                     /mol_type="genomic DNA"
                     /strain="GX-15"
                     /isolation_source="soil sample"
                     /db_xref="taxon:4911"
                     /country="China"
     rRNA            <1..>556
                     /product="26S ribosomal RNA"
ORIGIN
        1 aaccatcgag tgtcatgcct tagtacggcg agtgaagcgg caaaagctca aatttgaaat
       61 ctggcgtctt cgacgtccga gttgtaattt gaagaaggcg actttgtagc tggtccttgt
      121 ctatgttcct tggaacagga cgtcatagag ggtgagaatc ccgtgtggcg aggatcccag
      181 ttatttgtaa agtgctttcg acgagtcgag ttgtttggga atgcagctct aagtgggtgg
      241 taaattccat ctaaagctaa atattggcga gagaccgata gcgaacaagt acagtgatgg
      301 aaagatgaaa agaactttga aaagagagtg aaaaagtacg tgaaattgtt gaaagggaag
      361 ggcatttgat cagacatggc gtttgcttcg gctttcgctg ggccagatc agttttagcg
      421 gttggataaa tcctcgggaa tgtggctctg cttcggtaga gtgttatagc ccgtgggaat
      481 acagccagct gggactgagg attgcgactt ttgtcaagga tgctggcgta atggttaaat
      541 gccgcccgtc ttgacc
//
```

Figura 2: Exemplo de sequência em formato GBK full, em destaque pode-se observar o campo “TITLE JOURNAL”.

O formato GBK full permitiu que fosse realizada, sempre que necessária, a interconversão de formatos sem a perda de informações originais. Foram transferidas para o servidor local todas as sequências de

genomas completos pertencentes à família Saccharomycetaceae disponibilizados no NBCI em Novembro de 2013.

Para a construção da Tabela 1 e posterior seleção dos organismos, o formato GBK forneceu uma riqueza de informações que facilitou a pesquisa sobre o interesse biotecnológico e propriedades do organismo. As informações sobre o interesse biotecnológico foram buscadas no campo "TITLE JOURNAL" que contém o título do artigo de referência para os dados relacionados (Figura 2). As informações referentes ao *Status* da montagem do genoma foram extraídas do próprio site do NCBI em "Genome Assembly and Annotation Report".

A Tabela 1 representa os 71 genomas de leveduras Saccharomycetaceae disponíveis na data mencionada e copiados para o servidor local em formato GBK full. Destes genomas, 38 são de espécies diferentes de *S. cerevisiae* e 33 são linhagens pertencentes à espécie *S. cerevisiae*. A linhagem *S. boulardii* EDRL foi incluída no grupo de leveduras *S. cerevisiae*. Fato que pode ser justificado pelo recente estudo de hibridização genômica de genomas completos, que reclassificou *S. boulardii* como pertencente à espécie *S. cerevisiae* (Edwards-Ingram et al., 2007; Edwards-Ingram et al., 2004).

A análise do levantamento de genomas da família Saccharomycetaceae revelou dados interessantes, destacando-se o número de leveduras pertencentes a uma única espécie em detrimento ao total de espécies sequenciadas desta família. O advento das novas tecnologias NGS juntamente com o interesse biotecnológico pela *S. cerevisiae* refletem claramente a distribuição discrepante de dados genômicos na família Saccharomycetaceae, sendo quase 47% dos genomas pertencentes a uma única espécie: *S. cerevisiae* (Tabela 1). A predominância de genomas sequenciados pertencentes à espécie *S. cerevisiae* pode ser justificada pelo grande interesse biotecnológico neste organismo, além da ampla utilização em diversos processos industriais, como fabricação de pães, vinhos, cerveja e produção de etanol. Os 33 genomas de *S. cerevisiae* distribuem-se pelas diversas atividades:

produção de etanol (9 genomas), vinho (9), cerveja (2), *sake* (2), panificação (2), interesse clínico (1), efeito probiótico (1), laboratoriais (5), isolado do exsudato de uma árvore (1) e do ambiente de um cânion (1).

Tabela 1: Tabela comparativa de genomas pertencentes à família Saccharomycetaceae: organismos, descrição de propriedades e interesse biotecnológico, e, *status* da montagem do genoma (acesso em Novembro/2013).

(continua)

Genomas	Interesse Biotecnológico/ Propriedades	*Status* da montagem
Ashbya aceri	Produção de riboflavina. Patógeno de plantas.	Cromossomos sem gaps
Ashbya gossypii ATCC 10895	Fungo filamentoso. Agente patogênico para planta. Utilizado na indústria para produção de vitamina B12.	Cromossomos sem gaps
Candida glabrata CBS 138	Levedura ascomiceto patogênica. É o segundo agente causador mais frequente de candidíase em seres humanos.	Cromossomos com gaps
Candida glabrata CCTCC M202019	Levedura ascomiceto patogênica. É o segundo agente causador mais frequente de candidíase em seres humanos.	*Scaffolds*
Dekkera bruxellensis AWRI 1499	Causador da deterioração do vinho.	*Contig*s
Dekkera bruxellensis CBS 2499	Causador da deterioração do vinho.	*Scaffolds*
Eremothecium cymbalariae DBVPG 7215	Presença de micélio aéreo.	Cromossomos com gaps
Kazachstania africana CBS 2517	Levedura ascomiceta.	Cromossomos com gaps
Kazachstania naganishii CBS 8797	Espécie pós WGD.	Cromossomos com gaps
Kluyveromyces aestuarii ATCC 18862	Levedura ascomiceta encontrada em habitats marinhos.	*Contig*s
Kluyveromyces lactis strain NRRL Y-1140 *	Levedura ascomiceta usada para estudos genéticos e aplicações industriais, como produção de beta-galactosidase.	Cromossomos sem gaps
Kluyveromyces marxianus var marxianus KCTC 17555 *	Levedura termotolerante, polimórfica. Produção de etanol, enzimas, biorrremediação.	*Scaffolds*
Kluyveromyces wickerhamii UCD 54-210	Levedura ascomiceta associada com o intestino de moscas *Drosophila* sp..	*Contig*s

Tabela 1: Tabela comparativa de genomas pertencentes à família Saccharomycetaceae: organismos, descrição de propriedades e interesse biotecnológico, e, *status* da montagem do genoma (acesso em Novembro/2013).

(continuação)

Genomas	Interesse Biotecnológico/ Propriedades	*Status* da montagem
Komagataella pastoris CBS 7435	Produção de proteínas recombinantes, como alvos para drogas e de uso terapêuticos. Levedura modelo para proliferação de peroxissomos e assimilação de metanol.	Cromossomos com gaps
Komagataella pastoris DSMZ 70382	Produção de proteínas recombinantes, como alvos para drogas e de uso terapêuticos. Levedura modelo para proliferação de peroxissomos e assimilação de metanol.	*Contig*s
Komagataella pastoris GS 115	Produção de proteínas recombinantes, como alvos para drogas e de uso terapêuticos. Levedura modelo para proliferação de peroxissomos e assimilação de metanol.	Cromossomos com gaps
Lachancea kluyveri NRRL Y-12651	Levedura sequenciada para estudos de genômica comparativa.	Cromossomos
Lachancea thermotolerans CBS 6340 *	Levedura diplóide associada com frutas, *Drosophila* sp. e insetos.	Cromossomos com gaps
Lachancea waltii NCYC 2644	Relaciona-se a *S. cerevisiae*, porém divergiu antes da duplicação do genoma inteiro (WGD).	*Contigs*
Naumovozyma castellii CBS 4309	Membro do grupo de leveduras de brotamento *Saccharomyces* sp..	Cromossomos com gaps
Naumovozyma castellii NRRLY 12630	Membro do grupo de leveduras de brotamento *Saccharomyces* sp..	*Contigs*

Tabela 1: Tabela comparativa de genomas pertencentes à família Saccharomycetaceae: organismos, descrição de propriedades e interesse biotecnológico, e, *status* da montagem do genoma (acesso em Novembro/2013).

(continuação)

Genomas	Interesse Biotecnológico/ Propriedades	*Status* da montagem
Naumovozyma dairenensis CBS 421	Levedura isolada de frutas secas.	Cromossomos com gaps
Pachysolen tannophilus NRRL Y-2460	Levedura capaz de fermentar D-xilose a álcool.	*Contigs*
Saccharomyces arboricola H-6	Levedura isolada na China que é próxima a *S. cerevisiae.*	Cromossomos
Saccharomyces bayanus 623-6C	Usada na vinificação e sequenciada para estudos de genômica comparativa.	*Contigs*
Saccharomyces bayanus MCYC 623	Usada na vinificação e sequenciada para estudos de genômica comparativa.	*Contigs*
Saccharomyces boulardii EDRL *	Propriedades probióticas.	*Contigs*
Saccharomyces cerevisiae AWRI 1631	Derivado haplóide de uma linhagem comercial de vinho sul africano N96.	*Contigs*
Saccharomyces cerevisiae AWRI 796	Isolado de vinho vermelho sul-africano.	Cromossomos
Saccharomyces cerevisiae CBS 7960 *	Isolado da produção de etanol a partir do caldo de cana de açúcar brasileira.	*Scaffolds*
Saccharomyces cerevisiae CEN PK 133 7D	Linhagem laboratorial, que pode ser importante modelo biotecnológico.	Cromossomos
Saccharomyces cerevisiae CLIB 215	Isolado de uma panificadora da Nova Zelândia.	*Scaffolds*
Saccharomyces cerevisiae CLIB 324	Isolado de uma panificadora do Vietnã.	*Scaffolds*
Saccharomyces cerevisiae EC 1118	Linhagem industrial utilizada na produção de vinho.	*Scaffolds*
Saccharomyces cerevisiae EC 9-8	Derivado haplóide de um isolado do canion de Israel.	*Scaffolds*
Saccharomyces cerevisiae FL 100	Linhagem laboratorial.	*Scaffolds*
Saccharomyces cerevisiae Fosters B	Linhagem de cerveja comercial *ale*.	Cromossomos

Tabela 1: Tabela comparativa de genomas pertencentes à família Saccharomycetaceae: organismos, descrição de propriedades e interesse biotecnológico, e, *status* da montagem do genoma (acesso em Novembro/2013).

(continuação)

Genomas	Interesse Biotecnológico/ Propriedades	*Status* da montagem
Saccharomyces cerevisiae Fosters O	Linhagem de cerveja comercial *ale.*	*Contig*s
Saccharomcyes cerevisiae JAY 291 *	Linhagem haplóide derivada da levedura *S. cerevisiae* PE-2 da indústria brasileira de bioetanol.	*Contigs*
Saccharomyces cerevisiae Lalvin QA23	Linhagem do vinho branco português Vinho Verde.	*Scaffolds*
Saccharomyces cerevisiae M3707 *	Isolado diplóide de um destilador para uso no bioprocesso consolidado de bioetanol.	*Scaffolds*
Saccharomyces cerevisiae M3836 *	Haplóide do isolado M3707 para uso no bioprocesso consolidado de bioetanol.	*Scaffolds*
Saccharomyces cerevisiae M3837 *	Haplóide do isolado M3707 para uso no bioprocesso consolidado de bioetanol.	*Scaffolds*
Saccharomyces cerevisiae M3838 *	Haplóide do isolado M3707 para uso no bioprocesso consolidado de bioetanol.	*Scaffolds*
Saccharomyces cerevisiae M3839 *	Haplóide do isolado M3707 para uso no bioprocesso consolidado de bioetanol.	*Scaffolds*
Saccharomyces cerevisiae NY 1308 *	Linhagem industrial utilizada para fermentação de etanol na China.	*Contigs*
Saccharomyces cerevisiae PW 5	Isolado do vinho da palma Nigerian Rapha.	*Scaffolds*
Saccharomyces cerevisiae RM 11-1ª	Isolado haplóide derivado de um vinhedo da Califórnia.	*Scaffolds*
Saccharomyces cerevisiae S288c *	Genoma de referência da *S. cerevisiae* (Goffeau et al., 1996)	Cromossomos
Saccharomyces cerevisiae Sigma 1278b	Linhagem laboratorial.	Cromossomos
Saccharomyces cerevisiae T7	Isolado do exsudato da árvore do carvalho missouri.	*Scaffolds*
Saccharomyces cerevisiae UC-5	Levedura japonesa do *sake*.	*Scaffolds*
Saccharomyces cerevisiae Kyokai 7	Levedura japonesa do *sake*.	*Scaffolds*

Tabela 1: Tabela comparativa de genomas pertencentes à família Saccharomycetaceae: organismos, descrição de propriedades e interesse biotecnológico, e, *status* da montagem do genoma (acesso em Novembro/2013).

(continuação)

Genomas	Interesse Biotecnológico/ Propriedades	*Status* da montagem
Saccharomyces cerevisiae Vin 13	Linhagem de vinho branco sul africano.	Cromossomos
Saccharomyces cerevisiae VL3	Linhagem de vinho branco francês.	Cromossomos
Saccharomyces cerevisiae W303	Linhagem laboratorial.	Cromossomos
Saccharomyces cerevisiae ZTW1 *	Isolado da mistura de milho para produção de bioetanol chinês.	*Contigs*
Saccharomyces cerevisiae X *Saccharomyces kudriavzevii*	Libera grandes quantidades de tiol na produção de vinho, despertando o interesse de produtores de vinho.	*Contigs*
Saccharomyces cerevisiae YJM 789	Derivado haplóide de um isolado do pulmão de um paciente com AIDS.	*Contigs*
Saccharomyces cerevisiae YJM269	Isolado do vinho da uva *Austrian Blauer Portugieser.*	*Scaffolds*
Saccharomyces kudriavzevii IFO 1802	Levedura de brotamento, próxima a *S. cerevisiae*, separada aproximadamente por 10-12 milhões de anos de evolução.	*Scaffolds*
Saccharomyces kudriavzevii ZP591	Levedura de brotamento, próxima *a S. cerevisiae*, separada aproximadamente por 10-12 milhões de anos de evolução.	*Scaffolds*
Saccharomyces mikatae IFO 1815	Sequenciada para estudos de genômica comparativa.	*Contigs*
Saccharomyces paradoxus NRRL Y17217	Levedura utilizada na produção de vinhos e em estudos de genômica comparativa.	*Contigs*
Saccharomyces pastorianus CCY48-91	Levedura de cerveja *pilsen*.	*Contigs*
Saccharomyces pastorianus weihenstephan 34-70	Levedura de cerveja *pilsen*.	*Contigs*
Tetrapisispora blattae CBS 6284	Levedura isolada do intestino de uma barata.	Cromossomos com gaps

Tabela 1: Tabela comparativa de genomas pertencentes à família Saccharomycetaceae: organismos, descrição de propriedades e interesse biotecnológico, e, *status* da montagem do genoma (acesso em Novembro/2013).

(conclusão)

Genomas	Interesse Biotecnológico/ Propriedades	*Status* da montagem
Tetrapisispora phaffii CBS 4417	Levedura "assassina" usada para controlar leveduras contaminantes em processos industriais e na produção de vinho. Usada para controlar fungos patogênicos de plantas e para desenvolver anti-fúngicos para homens e animais.	Cromossomos
Torulaspora delbrueckii CBS 1146	Levedura usada para a produção de vinho e cerveja de trigo. Tolerante a vários fatores de stress como: congelamento, sal e desequilíbrio osmótico.	Cromossomos com gaps
Vanderwaltozyma polyspora DSM 70294	Levedura que representa a linhagem pós-WGD mais divergente.	*Scaffolds*
Zygosaccharomyces rouxii CBS 732	Levedura halotolerante e osmotolerante que provoca a deterioração de alimentos.	Cromossomos com gaps

Fonte: Banco de Dados NCBI (acesso em Novembro/2013, no site ftp://ftp.ncbi.nlm.nih.gov/genomes/)

Notas: * Leveduras selecionadas racionalmente.

Os dados referentes aos outros 38 genomas disponíveis da família Saccharomycetaceae reafirmam o interesse biotecnológico desta família, incluindo espécies além da *S. cerevisiae* (Tabela 1). Destes organismos 21 têm potencial biotecnológico diversificado: produção de etanol, vinho e cerveja; isolados clínicos; alvos para drogas e uso terapêutico, entre outros. Em conjunto, estes achados corroboram o fato de que os clados existentes dentro desta família são explorados de forma desigual (Souciet et al., 2009; Kurtzman 2003; Kurtzman & Robnett, 2003).

De acordo com os critérios estabelecidos selecionou-se 14 genomas para o trabalho (Tabela 2). As 14 linhagens, as quais estão destacadas na Tabela 1 com um asterisco, foram divididas em dois grupos de estudo (Tabela 2). O primeiro grupo continha as 11 leveduras *S. cerevisiae* de interesse para a produção de etanol, englobando leveduras industriais e laboratoriais. O segundo grupo continha as leveduras pertencentes ao complexo *Kluyveromyces* sp., que possuem propriedades biotecnológicas a serem exploradas por grupos de pesquisa de etanol, como por exemplo, a termotolerância da linhagem *K. marxianus* demonstrada em artigos publicados pelo grupo (Costa et al., 2014; De Souza et al., 2012).

Adicionalmente, na Tabela 2 estão os identificadores do projeto de sequenciamento, a versão e data da montagem disponibilizada, o banco de dados utilizado e as referências de artigos relacionados às linhagens. Estas informações são de extrema importância para reprodução dos experimentos, visto que os grupos que depositaram os genomas podem depositar versões atualizadas dos dados com informações adicionais nos bancos de dados. Como observado, foi feito um novo *download* de alguns genomas no site do JGI, uma vez que o NCBI não disponibilizava a informação de anotação referente a estes organismos.

Tabela 2: Espécies selecionadas racionalmente para a genômica comparativa: identificador do projeto, versão e data da montagem, banco de dados e referência.

Linhagens de Leveduras	BioProject	GenBank Assembly ID	Disponível (mês/ano) Fonte	Referência
S. cerevisiae S288c *[1]	PRJNA43747	GCA_000146055.2	07-08/2013 NCBI	Goffeau et al., 1996
S. cerevisiae CBS 7960 *[1]	PRJNA60391	GCA_000192455.1	03/2011 NCBI	Borneman et al., 2013
S. cerevisiae JAY 291 *[1]	PRJNA32809	GCA_000182315.2	05/2012 NCBI	Argueso et al., 2009
S. cerevisiae M3707 *[1]	PRJNA174688	GCA_000365045.1	04/2013 JGI	Brown et al., 2013
S. cerevisiae M3836 *[1]	PRJNA174689	GCA_000365065.1	04/2013 JGI	Brown et al., 2013
S. cerevisiae M3837 *[1]	PRJNA174690	GCA_000365085.1	04/2013 JGI	Brown et al., 2013
S. cerevisiae M3838 *[1]	PRJNA174691	GCA_000365105.1	04/2013 JGI	Brown et al., 2013
S. cerevisiae M3839 *[1]	PRJNA174692	GCA_000365125.1	04/2013 JGI	Brown et al., 2013
S. cerevisiae NY1308 *[1]	PRJNA202086	GCA_000416405.1	06/2013 NCBI	Zheng et al., 2013
S. cerevisiae ZTW1 *[1]	PRJNA174065	GCA_000308935.1	10/2012 NCBI	Zheng et al., 2013
S. boulardii EDRL *[1]	PRJNA207020	GCA_000442675.1	08/2013 NCBI	Kathri et al., 2013
K. lactis NRRL Y-1140 *[2]	PRJNA13835	GCA_000002515.1	07/2004 NCBI	Zivanovic et al., 2005; Dujon et al., 2004
K. marxianus KCTC 17555 *[2]	PRJNA89605	GCA_000299195.1	09/2012 NCBI	Jeong et al., 2012
L. thermotolerans CBS 6340 *[2]	PRJNA39575	GCF_000142805.1	06/2009 NCBI	Genolevures Consortium, 2009

Fonte: Bancos de Dados NCBI e JGI (acesso em Novembro e Dezembro/2013).

Notas: *[1] Leveduras pertencentes ao grupo 1: espécies *S. cerevisiae*.

*[2] Leveduras pertencentes ao grupo 2: espécies *Kluyveromyces* sp.

A análise qualitativa da montagem dos genomas permitiu uma visão geral das diferenças dos genomas selecionados (Tabela 3). De uma maneira geral, o conteúdo CG variou entre 38% e 40%, com exceção de *L. thermotolerans*, que possui um conteúdo de 47,30%. Isso pode ser explicado, pelo fato de que um segmento cromossômico rico em CG apresenta uma sintenia conservada em *L. thermotolerans*, é desprovido de elementos transponíveis, e replica mais tarde que outros cromossomos durante a fase S (Genolevures Consortium, 2009; Payen et al., 2009).

Sabe-se que o par CG está ligado por três pontes de hidrogênio enquanto que o par AT é ligado por duas ligações de hidrogênio, espera-se que os organismos de crescimento em temperaturas mais elevadas possuam uma proporção maior de pares GC do que o AT. Neste sentido, a porcentagem do conteúdo GC pode revelar informações sobre a estabilidade e habitat desse organismo. No entanto, os dados existentes na literatura sobre essa proposição, variação de CG e temperatura, ainda são conflitantes (Hao & Wu, 2010; Hickey & Singer, 2004; Xia et al., 2002). Neste sentido, não é possível afirmar que o alto conteúdo de CG de *L. thermotolerans* esteja ligado à sua termotolerância.

Com relação aos *contig*s foi realizada uma contagem para caracterizar: o número dos *contig*s na montagem; a soma dos pares de bases (pb) dos *contigs*, que é o tamanho do genoma; o tamanho médio dos *contigs* em pb; a mediana referente ao tamanho dos *contigs* em pb, que é a medida mais robusta para separar o conjunto dos valores dos tamanhos dos *contigs* ao meio; e os valores N50 dos *contigs*, que fornece uma estatística da mediana, de tal modo que 50% de todo o conjunto está contido em *contigs* iguais ou maiores do que o valor N50. Para *S. cerevisiae* CBS 7960 foram obtidos os menores valores de média e mediana, o que é condizente com o alto número de *contigs* deste genoma 2.367, refletindo em um genoma muito fragmentado e que necessita de um novo sequenciamento para refletir a sequência real do genoma (Tabela 3).

Os genomas dos organismos selecionados apresentam um tamanho médio entre 10 e 12 Mb, foram sequenciados por diferentes plataformas e

montados em cromossomos sem gaps ou somente disponibilizados em *contigs* sem orientação e ordenação. No útlimo caso, os *contigs* apresentam-se como sequências consenso, numeradas do maior tamanho de *contigs* para o menor tamanho de *contigs*, o que não referencia sua ordem no genoma nem orientação direta ou reversa complementar. A orientação dos *contigs* é necessária para indicar se o *contigs* deve ser montado na forma direta, ou utilizando o reverso complementar da sequência. A ordenação dos *contigs* é necessária para identificar em qual ordem os *contigs* devem ser representados dentro dos cromossomos. Dessa forma, a ordenação e orientação dos *contigs* são muito importantes para se obtiver um genoma bem montado e que retrate o DNA do organismo em estudo. Pela Tabela 3 pode-se observar diferentes casos de ordenação e orientação de *contigs*, por exemplo, *K. lactis*, que possui *contigs* ordenados e orientados em cromossomos, e *S. boulardii*, que possui *contigs* sem ordenação e orientação.

Em conjunto, os dados das Tabelas 1, 2 e 3 permitiram a seleção racional das espécies com o objetivo de distinguir um conjunto de genomas de organismos de interesse, provenientes de um banco de dados de domínio público, para estudar de forma mais aprofundada estes organismos e suas relações. A fim de complementar as informações obtidas foi feita uma revisão literária das pesquisas e informações existentes sobre os organismos selecionados.

Tabela 3: Dados qualitativos da montagem dos genomas.

Linhagens de Leveduras	%CG	N° de *Contigs*	Soma de tamanhos dos *Contigs* (pb)	Média de tamanhos dos *Contigs* (pb)	Mediana de tamanhos dos *Contigs* (pb)	Valor de N50	Tecnologia de Sequenciamento
S. cerevisiae S288c *[1]	38,16	17	12.157.105	715.123	745.751	924.431	Diversos métodos em colaboração de laboratórios
S. cerevisiae CBS 7960 *[1]	38	2367	12.215.386	5.160	1.152	18.761	454; ABI 3730
S. cerevisiae JAY 291 *[1]	38,10	453	11.538.054	25.470	10.133	64.336	454; Solexa
S. cerevisiae M3707 *[1]	38,20	97	11.459.946	118.143	71.100	291.573	Illumina
S. cerevisiae M3836 *[1]	38,20	119	11.498.676	96.627	40.633	259.428	Illumina
S. cerevisiae M3837 *[1]	38,20	113	11.482.540	101.615	65.740	231.369	Illumina
S. cerevisiae M3838 *[1]	38,20	84	11.521.811	137.164	68.836	355.957	Illumina
S. cerevisiae M3839 *[1]	38,20	120	11.490.006	95.750	54.654	231.486	Illumina
S. cerevisiae NY1308 *[1]	38,20	35	11.514.196	328.977	213.154	546.517	454; Illumina
S. cerevisiae ZTW1 *[1]	38,30	33	11.414.768	345.902	311.976	556.921	454; Sanger
S. boulardii EDRL *[1]	38,30	194	11.482.966	59.190	2.324	247.679	454
K. lactis NRRL Y-1140 *[2]	38,71	6	10.689.156	1.781.526	1.753.957	1.753.957	WGS Sanger
K. marxianus KCTC 17555 *[2]	40,10	119	10.845.407	91.137	865	1.189.284	Illumina GAIIx
L. thermotolerans CBS 6340 *[2]	47,30	8	10.392.862	1.299.107	1.513.537	1.513.537	WGS Sanger

Fonte: Banco de Dados NCBI (acesso em Novembro e Dezembro/2013).

Notas: *[1] Leveduras pertencentes ao grupo 1: espécies *S. cerevisiae*.

*[2] Leveduras pertencentes ao grupo 2: espécies *Kluyveromyces* sp.

5.1.1. Grupo 1: Linhagens de *Saccharomyces cerevisiae*

A levedura *S. cerevisiae* foi o primeiro organismo eucarioto a ser sequenciado, sendo o principal modelo experimental utilizado para genômica funcional (Ooi et al., 2006; Goffeau et al.,1996). Diversas linhagens deste organismo foram inteiramente sequenciadas, constituindo em um extenso volume de informações a ser explorado de forma mais aprofundada, para permitir um melhor entendimento do organismo e suas relações filogenéticas.

Associada com atividades humanas por milhares de anos em atividades como panificação, fabricação de cerveja e produção de vinho, a levedura *S. cerevisiae* passou por pressões seletivas para características desejadas específicas. Como resultado, as linhagens deste organismo são altamente especializadas, com leveduras que produzem vinho não são boas produtoras de cerveja e vice-versa (Borneman et al., 2011), apresentando considerável diversidade de linhagens conforme descrito na Tabela 1.

Linhagens selecionadas de *S. cerevisiae* são amplamente utilizadas pelo setor sucro-energético brasileiro, por possuírem uma combinação de alta eficiência fermentativa com prolongada persistência na safra. Algumas destas linhagens tornaram-se disponíveis comercialmente no final da década de 1990, e, atualmente, mais da metade das destilarias brasileiras usam uma, ou mais comumente uma mistura de duas ou mais, destas linhagens selecionadas para iniciar o processo fermentativo, produzindo bilhões de galões de etanol por ano. As linhagens mais amplamente utilizadas pelas usinas brasileiras são PE-2, CAT-1 e BG-1 por apresentarem uma notável capacidade de competir com linhagens nativas, sobrevivência e dominância durante o processo fermentativo industrial. Na safra de 2007/2008, as linhagens PE-2 e CAT-1 foram usadas em cerca de 150 destilarias, representando cerca de 60% do etanol combustível produzido no Brasil (Basso et al., 2008). Uma das linhagens selecionadas neste projeto foi justamente o derivado haplóide da PE-2, que pode ser amplamente utilizado nas indústrias brasileiras, possui características

industriais desejadas ao micro-organismo fermentador, portanto, diferindo-se de leveduras laboratoriais.

5.1.1.1. Saccharomyces cerevisiae S288c

A linhagem laboratorial *S. cerevisiae* S288c foi o primeiro organismo eucarioto a ter o genoma completamente sequenciado através de um esforço conjunto mundial (Goffeau et al., 1996). S288c possui uma genealogia complexa, mas é derivada primariamente (~88% do seu genoma) da linhagem EM93, a qual foi isolada de um figo apodrecido na Califórnia Central, Estados Unidos, em 1938 (Mortimer & Johnston, 1986). Os 12% restantes do seu genoma vêm de cinco diferentes progenitores: dois isolados naturais (EM126 isolado em 1939 de um figo apodrecido na Califórnia Central e NRRL YB-210 isolada de bananas apodrecidas da Costa Rica em 1942) e três isolados de linhagens de panificação industriais (Yeast Foam, FLD e LK) (Engel & Cherry, 2013; Mortimer & Johnston, 1986).

Este organismo possui 16 cromossomos e cerca de 6000 genes, sendo extensamente estudado por grupos de pesquisa por ser um organismo modelo eucarioto, de fácil manipulação e possuir s*tatus* GRAS de segurança para ser utilizado na indústria alimentícia e para manipulação laboratorial.

Pelas características citadas previamente este organismo é geralmente utilizado como organismo laboratorial de referência para estudos comparativos com outras linhagens e espécies, assim como foi utilizado no presente trabalho. Os dados apresentados pela Tabela 3 mostram que este organismo possui 38,16% de conteúdo CG, 17 *contigs* que se referem aos 16 cromossomos e ao DNA mitocondrial.

5.1.1.2. Saccharomyces cerevisiae CBS 7960

A linhagem industrial *S. cerevisiae* CBS 7960 é um isolado da produção de etanol a partir do caldo de cana-de-açúcar brasileira (Borneman et al., 2013).

Os dados do sequenciamento indicam que a informação deste genoma está truncada, uma vez que se distribui em 2.367 *contigs* (Tabela 3). Dessa forma, a informação sequenciada não está organizada em cromossomos e para facilitar o trabalho de comparação de genomas, unimos todos os *contigs* em um arquivo único que permitia a comparação com outros genomas através do visualizador ACT. Este procedimento de união de *contigs* ou cromossomos foi padronizado para todos os organismos selecionados. Por um lado, a união dos dados permitia a visualização de blocos sintênicos, mas por outro lado a união sem orientação e ordenação de *contigs* levou à obtenção de muitas CDS truncadas, que atrapalhavam a análise, por superestimar o número de CDS desse organismo em relação aos demais (ver item 4.2).

5.1.1.3. Saccharomyces cerevisiae JAY 291 (PE-2)

A linhagem industrial *S. cerevisiae* JAY 291, comercializada como PE-2 pela indústria brasileira Fermentec, é um isolado da usina sucro-energética brasileira a partir da produção do etanol de primeira geração, aquele proveniente do caldo da cana-de-açúcar (Argueso et al., 2009). A linhagem JAY 291 é amplamente utilizada pelo setor industrial devido a sua elevada produtividade e robustez, consideravalmente superiores ao padrão de laboratório S288c. Além disso, estudos posteriores demonstraram a capacidade superior de fermentação de JAY 291 no processo de produção de etanol de segunda geração, em comparação a leveduras não industriais (Costa et al., 2014).

Os dados de sequenciamento e montagem desse organismo demonstram que o genoma está fragmentado em um alto número de *contigs*, 453, e que 50% dos *contigs* possuem um tamanho maior ou igual a 64.336 pb (Tabela 3). Em comparação com o genoma da linhagem S288c, seria como se cada cromossomo estivesse fragmentado em 10 segmentos sem orientação e ordenação.

5.1.1.4. Saccharomyces cerevisiae M3707 e seus quatro derivados haplóides M3836, M3837, M3838 e M3839

A linhagem diplóide *S. cerevisiae* M3707 e seus quatro derivados haplóides M3836, M3837, M3838 e M3839 foram isolados de destiladores comerciais (Brown et al., 2013). Estas linhagens possuem o potencial para serem utilizadas no CBP, bioprocesso consolidado para produção de bioetanol. Em 2013 foi anunciado o sequenciamento do genoma destas linhagens abrindo novos caminhos para o desenvolvimento do CBP (Brown et al, 2013).

Os dados apresentados na Tabela 3 mostram que pelo fato destes genomas terem sido sequenciados na mesma plataforma (Ilumina), estão submetidos aos mesmos erros de sequenciamento e geraram dados facilmente comparáveis entre si. O número de *contigs* variou de 84 a 120, e os valores de N50 ficaram entre 231.486 e 355.957, para as linhagens M3838 e M3839, respectivamente.

5.1.1.5. Saccharomyces cerevisiae NY1308

A linhagem industrial *S. cerevisiae* NY1308 é utilizada para fermentação de etanol na China, tendo sido sequenciada por pesquisadores do *Institute of Microbiology* da *Zhejiang University* e submetida ao banco de dados do NCBI em 2013 (Zheng, 2013).

Os dados gerados pelo sequenciamento em duas plataformas, 454 e Ilumina, resultaram em 35 *contigs* e um valor de N50 maior (546.517) do que o sequenciamento realizado na mesma época somente pelo Ilumina, para as linhagens: M3707, M3836, M3837, M3838 e M3839 (Tabela 3). À vista disso, a utilização de mais de uma tecnologia de sequenciamento traz benefícios para a qualidade do genoma montado.

5.1.1.6. Saccharomyces cerevisiae ZTW1

A linhagem industrial *S. cerevisiae* ZTW1 é um isolado da mistura de milho para a produção de bioetanol chinês. Esta linhagem é altamente tolerante a condições de estresse, acúmulo de etanol no meio, produção

de bioprodutos e possui elevadas taxas de crescimento, sendo estas características desejáveis em leveduras industriais (Zhang et al., 2015; Zheng et al., 2014; Zheng et al., 2013).

Os dados obtidos pelo sequenciamento nas plataformas, Sanger e 454, revelam um genoma dividido em 33 *contigs*, com 50% dos *contigs* possuindo um tamanho maior que 556.921 pb (Tabela 3). A união de duas plataformas de sequenciamento, Sanger e 454, favoreceu a montagem dos *contigs* e por consequência forneceu dados mais acurados para a anotação. O grupo responsável pelo sequenciamento de ZTW1 foi o mesmo grupo que sequenciou NY 1308 (Zheng et al., 2013).

5.1.1.7. Saccharomyces boulardii EDRL

A linhagem industrial *S. boulardii* EDRL é um probiótico mundialmente utilizado para alivar os efeitos de doenças gastro-intestinais severas e controle de diarréias devido ao uso de antibióticos (Kathri et al., 2013). Estudos de hibridização demonstraram que esta levedura pertence à espécie *S. cerevisiae* (Edwards-Ingram et al., 2007; Edwards-Ingram et al., 2004). Devido às suas propriedades biotecnológicas, amplo uso industrial e pouco conhecimento do seu genoma, recentemente sequenciado (Kathri et al., 2013), optou-se por incluí-la no presente estudo. Desta forma, a levedura *S. boulardii* foi incluída no grupo de leveduras *S. cerevisiae*.

O sequenciamento da linhagem EDRL pela plataforma 454 foi o primeiro referente à espécie probiótica *S. boulardii* publicado. Seu genoma distribui-se por 194 *contigs* e o valor N50 foi 247.679 (Tabela 3).

5.1.2. Grupo 2: Linhagens pertencentes ao gênero *Kluyveromyces* sp.

Dentre as leveduras, o gênero *Kluyveromyces* é de grande interesse do ponto de vista industrial, devido a suas características fisiológicas e o *Status* GRAS de segurança. O gênero *Kluyveromyces* pertence à família Saccharomycetaceae, mesma família da levedura *S. cerevisiae*. Estas

leveduras são conhecidas como sendo protoplóides, pelo fato de não terem passado pela duplicação de todo genoma através do WGD e possuem 6, 7 ou 8 cromossomos (Souciet et al., 2009) A espécie *K. marxianus*, por exemplo, pode ser utilizada para a produção de enzimas, biomassa, SCP proteína unicelular e expressão de proteínas heterólogas (Lane & Morrisey, 2010).

5.1.2.1. Kluyveromyces marxianus KCTC 1755

A levedura *K. marxianus* é uma levedura termotolerante que tem sido pesquisada devido ao seu potencial biotecnológico. Geralmente *K. marxianus* é encontrada em queijos e derivados lácteos, sendo uma espécie altamente polimórfica. O número de cromossmo varia de 6 a 12, sendo 8 cromossomos o mais comum. A linhagem KCTC 1755 tem tamanho de 10,9 Mb e foi isolada de pozol, que é uma massa de milho fermentada mexicana, e tem sido usada para a produção de enzimas industriais como β-galactosidases e inulinases (Jeong et al., 2012; Martins et al., 2002; Rouwenhorst 1988).

Embora a levedura *K. lactis* tenha sido reconhecida como a levedura modelo dentro do gênero *Kluyveromyces*, *K. marxianus* possui inúmeras vantagens sobre *K. lactis* de forma poder se tornar levedura modelo, com potenciais aplicações biotecnológicas. *K. marxianus* pode crescer em uma maior variedade de substratos e a altas temperaturas, exibem uma alta taxa de crescimento específico e produzem menos etanol na presença de açúcar excessivo. Esse promissor potencial biotecnológico juntamente com a busca por um organismo fermentador para biomassas lignocelulósicas motivou o sequenciamento da linhagem *K. marxianus* KCTC 1755. Este foi o primeiro genoma de *K. marxianus* disponibilizado para domínio público e foi publicado por Jeong e colaboradores em 2012 (Jeong et al., 2012).

De acordo com a Tabela 3, este genoma foi sequenciado na plataforma Ilumina, distribui-se por 119 *contigs* e tem 50% dos *contigs* com tamanho acima de 189.284 pb.

5.1.2.2. Kluyveromyces lactis

A levedura *K. lactis* foi sequenciada pelo grupo Genolevures Consortium em 2004, possuindo 6 cromossomos (nomeados de A-F) além do DNA mitocondrial. É a levedura modelo para estudos do metabolismo de lactose e de metabolismo Crabtree negativa, ou seja, possui baixos rendimentos fermentativos em ambientes aeróbios, sendo considerada uma levedura respiro-fermentativa (Dujon et al., 2004; Gonzáles-Siso et al., 1996).

Os dados do sequenciamento deram suporte à escolha deste organismo como referência no Grupo 2, seu genoma divide-se por 7 *contigs*, os cromossomos A-F e o DNA mitocondrial, com o valor de N50 1.753.957, tendo sido sequenciado através da metodologia WGS, portanto pelo método de Sanger, que produz fragmentos maiores que facilitaram a montagem desse genoma em cromossomos (Tabela 3).

5.1.2.3. Lacchancea thermotolerans

A levedura *Kluyveromyces thermotolerans* foi originalmente isolado por Filipov (como *Saccharomyces veronae* em 1932) a partir da fermentação geléia de ameixa. Linhagens de *K. thermotolerans* são normalmente associados com frutas, moscas *Drosophila* e, possivelmente, outros insetos que se alimentam de plantas. Primeiramente, esta espécie foi classificada como pertencente ao gênero *Kluyveromyces* devido ao fato dos esporos serem liberados somente após a formação dos ascos. Apesar da sua nomenclatura, a capacidade de crescer em altas temperaturas é variável, no entanto, uma linhagem originária da decomposição de pêra nas ilhas Cayman cresceu vigorosamente a 37° C (Informação visualizada dia 26/04/2015 em http://genolevures.org/klth.html).

O genoma diplóide foi sequenciado em 2000 e apresenta-se distribuído em oito cromossmos (nomeados de A-H) em um tamanho de 10,4 Mb, excluindo o DNA repetitivo (Informação visualizada dia 26/04/2015 em http://genolevures.org/klth.html).O grande interesse neste organismo deve-se ao fato dessa levedura ser utilizada na produção do vinho

Marlborough, a partir da uva Sauvignon blanc (Gobbi et al., 2013; Anfang et al., 2008). A característica de termotolerância encontrada em muitas linhagens desse organismo poderia ser muito útil no processo de produção de etanol celulósico SSF, uma vez que as enzimas celulolíticas utlizadas nesse processo requerem altas temperaturas para agirem com maior rendimento.

Segundo nosso trabalho, este genoma está dividido em 10 *contigs*, apesar de ser apresentado em oito cromossomos (Tabela 3). Isso se deve ao fato de terem sido inseridos dois *gaps* entre *contigs* ordenados e orientados, para se fechar o cromossomo. Dessa forma, o valor de N50 é alto, 1.513.537pb.

5.2. Anotação e Re-anotação de Genomas

Após selecionar os genomas e unir os dados de sequenciamento em arquivos únicos, os organismos foram submetidos à anotação e re-anotação de acordo com o *pipeline* do Grupo Informática de Biossistemas: predição *ab initio*, predição por similaridade de sequências e anotação consenso final (Tabela 4).

Os organismos *S. cerevisiae* CBS 7960, *S. cerevisiae* NY1308, *S. boulardii* EDRL e *K. marxianus* KCTC 17555 não possuíam anotação disponível para *download* nos bancos de dados de domínio público (Tabela 4). Com isso, toda e qualquer informação que obtivemos para estes genomas foi considerada como ganho de anotação. O restante dos organismos já possuía algum grau de anotação, sendo incorporada às informações obtidas através do *pipeline* de re-anotação. Dessa forma, o ganho de anotação calculado para estes genomas foi a diferença entre a anotação obtida pelo *pipeline* e a anotação original dos dados. Com a exceção de *K. lactis* NRRL Y-1140, todos os genomas possuíram um ganho de anotação. É importante destacar que os genomas de *S. cerevisiae* M3707, 3836, 3837, 3838 e 3839 possuíam dados somente de anotação estrutural pelo *software fgenes,* o banco de dados não reportou a anotação funcional dos termos anotados.

Tomando como genoma de referência a levedura *S. cerevisiae* S288c para o grupo 1, observou-se um aumento no número de informação codificante em relação às 5.906 proteínas deste organismo (Tabela 4). Isto ocorre inclusive com o próprio genoma de *S. cerevisiae* S288c, que após a re-anotação teve um possível ganho de 322 CDS. Entretanto, esse ganho de anotação deve ser olhado com cautela, uma vez que o genoma de *S. cerevisiae* S288c é altamente conhecido e curado manualmente pelo SGD. Essas 322 CDS a mais podem ser explicadas pelo fato de que todos os cromossomos desse organismo, incluindo o DNA mitocondrial, foram unidos em um arquivo único, o que pode ter levado à formação errônea de CDS. Adicionalmente, o *script* remove-redundancia_genome-annotation.pl anota CDS multi-éxons como CDS diferentes, em um arquivo separado. Para corrigir este problema, seria necessário realizar a curadoria manual de todos os genomas, que se tornou inviável de ser executada dentro do plano do trabalho. Há ainda que se considerar que algumas dessas CDS possam ser realmente CDS novas que não foram anotadas, fato este que deve ser investigado mediante o laborioso processo de curadoria manual.

A linhagem *S. cerevisiae* CBS 7960 retornou o maior número de CDS de todas as leveduras do grupo 1. Foram anotadas 6.796 CDS, valor que indica a presença de 890 CDS a mais que a anotação original do genoma de referência *S. cerevisiae* S288c. Isso pode ser explicado pelo alto número de *contigs* desse genoma, que por estar muito segmentado resultou em na identificação de muitas CDS truncadas. Uma sugestão para verificar a precisão do protocolo de anotação seria o re-sequenciamento dessa levedura, buscando fechar os *contigs*, re-ordená-los e orientá-los.

Para o grupo 2, o genoma de referência foi a *K. lactis* NRRL Y-1140 por ser o organismo mais bem conhecido deste grupo. Coincidentemente, esse foi o único caso que não houve ganho de anotação. Após o *pipeline* de re-anotação houve uma perda de 34 CDS anotadas, que pode ser explicado pelo valor estringente de *e-value* utilizado no *pipeline*.

De uma maneira geral, era esperado que obtivéssemos um ganho de anotação, visto que a literatura tem reportado um valor de 7% de genes adicionados após a re-anotação dos genomas. Para afirmar que o aumento

no número de CDS anotadas refletiu em ganho real de anotação faz-se necessário realizar a curadoria manual em pesquisas posteriores.

Adicionalmente, é importante destacar que a anotação e re-anotação dos genomas através de um *pipeline* único permitiu uma melhor comparação entre as sequências desses organismos. A padronização dos formatos das sequências, nomenclatura, termos de anotação e grau de curadoria foi estabelecida com critérios rigorosos. Com isso, mesmo que continuem existindo erros de anotação ficou mais fácil detectá-los e levá-los em consideração durante as análises comparativas.

Tabela 4: *Status* final da anotação de proteínas incluindo os genomas mitocondriais.

Linhagens de Leveduras	Dados Originais	Predição *Ab initio* (Augustus)	Predição por similaridade (Swiss-Prot db)	Anotação Consenso (nr db)	Ganho de Anotação	Porcentagem de Anotação
S. cerevisiae S288c *[1]	5.906	5.466	6.585	6.228	322	6,47%
S. cerevisiae CBS 7960 *[1]	-	5.763	7.053	6.796	6.796	100%
S. cerevisiae JAY 291 *[1]	5.197	5.389	6.395	6.096	899	17,2%
S. cerevisiae M3707 *[1]	5.974	5.364	6.291	6.128	154	2,57%
S. cerevisiae M3836 *[1]	5.984	5.367	6.309	6.134	150	2,50%
S. cerevisiae M3837 *[1]	5.969	5.369	6.296	6.128	159	2,66%
S. cerevisiae M3838 *[1]	5.991	5.382	6.323	6.139	148	2,47%
S. cerevisiae M3839 *[1]	5.989	5.357	6.309	6.139	150	1,02%
S. cerevisiae NY1308 *[1]	-	5.374	6.310	5.996	5.996	100%
S. cerevisiae ZTW1 *[1]	5.197	5.349	6.278	6.000	803	1,15%
S. boulardii EDRL *[1]	-	5.348	6.355	6.034	6.034	100%
K. lactis NRRL Y-1140 *[2]	5.076	4.902	5.154	5.042	- 34	-0,67%
K. marxianus KCTC 17555 *[2]	-	4.838	5.214	5.033	5.033	100%
L. thermotolerans CBS 6340 *[2]	5.092	4.459	5.332	5.146	54	1,01%

Fonte: Banco de Dados GenBank NCBI (acesso em Novembro e Dezembro/2013).

Notas: *[1] Leveduras pertencentes ao grupo 1: espécies *S. cerevisiae*.

*[2] Leveduras pertencentes ao grupo 2: espécies *Kluyveromyces* sp..

As Tabelas 5 e 6 representam um achado interessante do trabalho no que tange à propagação de informações através das buscas por similaridades de sequência. A Tabela 5 representa a distribuição de termos relacionados ao desconhecimento da função da CDS anotada nos dados originais, enquanto que a Tabela 6 representa essa mesma distribuição nos dados re-anotados. Essa busca feita pelos termos não era sensível às diferenças de maiúsculo e minúsculo nos termos procurados: *unknown*, *hypothetical*, *putative* e *undefined*.

De uma maneira geral, após a re-anotação houve um aumento significativo do termo *putative*. Isso ocorre devido a um viés intrínseco à metodologia de anotação utilizada, a qual coloca sempre no início da descrição da CDS o termo *putative*. Esta nomenclatura pode ser explicada, baseada no fato de que para confirmar a função predita da CDS seria necessário cruzar dados genômicos e experimentais, por meio do processo de curadoria manual.

O aumento geral desses termos que não definem as CDS leva a uma hipótese errônea de que o *pipeline* utilizado não trouxe ganho de informação. Isso aconteceu porque a busca por similaridade de sequências retornou as 10 primeiras linhas da descrição da *query* desejada, propagando informações imprecisas do banco de dados. Mais uma vez, para retornar somente a informação desejada pode-se fazer a curadoria manual das CDS, ou realizar a busca somente contra um banco de dados curados, recuperando-se somente o primeiro *hit* da sequência buscada.

Tabela 5: Distribuição de genes originalmente anotados sem função descrita, de acordo com os termos mais frequentes para descrever esta condição.

Linhagens de Leveduras	unknown	hypothetical	putative	undefined	Soma dos termos em inglês
S. cerevisiae S288c *[1]	15	1654	428	0	2.097
S. cerevisiae CBS 7960 *[1]	-	-	-	-	-
S. cerevisiae JAY 291 *[1]	930	5	986	0	1.921
S. cerevisiae M3707 *[1]	0	0	0	0	0
S. cerevisiae M3836 *[1]	0	0	0	0	0
S. cerevisiae M3837 *[1]	0	0	0	0	0
S. cerevisiae M3838 *[1]	0	0	0	0	0
S. cerevisiae M3839 *[1]	0	0	0	0	0
S. cerevisiae NY1308 *[1]	-	-	-	-	-
S. cerevisiae ZTW1 *[1]	963	3	969	0	1.935
S. boulardii EDRL *[1]	-	-	-	-	-
K. lactis NRRL Y-1140 *[2]	195	737	184	0	1.116
K. marxianus KCTC 17555 *[2]	-	-	-	-	-
L. thermotolerans CBS 6340 *[2]	201	974	180	0	1.355

Fonte: Banco de Dados NCBI (acesso em Novembro e Dezembro/2013).

Notas: *[1] Leveduras pertencentes ao grupo 1: espécies *S. cerevisiae*.

*[2] Leveduras pertencentes ao grupo 2: espécies *Kluyveromyces* sp..

Tabela 6: Distribuição de genes originalmente re-anotados sem função descrita, de acordo com os termos mais frequentes para descrever esta condição.

Linhagens de Leveduras	unknown	hypothetical	putative	undefined	Soma dos termos em inglês
S. cerevisiae S288c *[1]	86	607	6.511	0	7.204
S. cerevisiae CBS 7960 *[1]	72	714	7.072	0	7.858
S. cerevisiae JAY 291 *[1]	39	576	6.317	0	6.932
S. cerevisiae M3707 *[1]	59	493	6.397	0	6.949
S. cerevisiae M3836 *[1]	60	489	6.415	0	6.964
S. cerevisiae M3837 *[1]	62	495	6.398	0	6.955
S. cerevisiae M3838 *[1]	58	490	6.424	0	6.972
S. cerevisiae M3839 *[1]	63	481	6.416	0	6.960
S. cerevisiae NY1308 *[1]	50	443	6.246	0	6.739
S. cerevisiae ZTW1 *[1]	42	451	6.233	0	6.726
S. boulardii EDRL *[1]	47	853	6.273	0	7.173
K. lactis NRRL Y-1140 *[2]	13	4.230	5.073	0	9.316
K. marxianus KCTC 17555 *[2]	10	4.175	5.068	0	9.253
L. thermotolerans CBS 6340 *[2]	16	5.340	5.193	0	10.549

Notas: *[1] Leveduras pertencentes ao grupo 1: espécies *S. cerevisiae*.

*[2] Leveduras pertencentes ao grupo 2: espécies *Kluyveromyces* sp.

5.3. Genômica Comparativa

A genômica comparativa foi realizada com os dados do proteoma predito dos organismos por meio do algoritmo de *cluster*ização OrthoMCL. Os parâmetros escolhidos estavam de acordo com o recomendado pelos autores (Li *et al.,* 2003).

Foram realizados dois experimentos de agrupamento de proteínas: um para cada grupo previamente definido. O conjunto de todas as proteínas preditas do grupo 1, o qual representa 11 linhagens de *S. cerevisiae*, possui um total de 67.818 proteínas. Enquanto que o total de proteínas do grupo 2, o qual remete às 3 espécies *Kluyveromyces* sp., foi 15.221 proteínas (Figura 3).

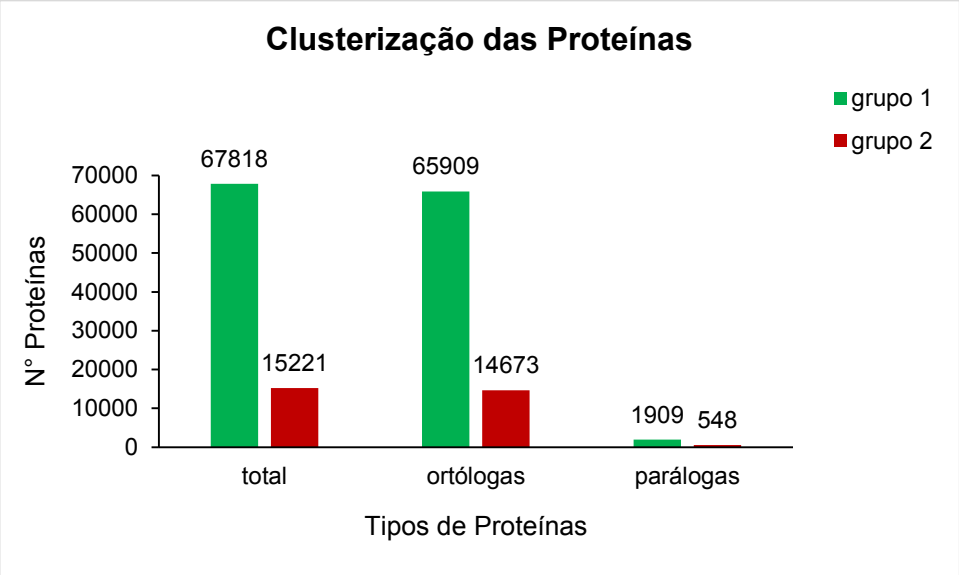


Figura 3: Distribuição das proteínas ortólogas e possíveis parálogas para os grupos 1 e 2.

Para o grupo 1, obteve-se 65.909 proteínas agrupadas em *clusters* de genes ortólogos, ou seja, 97,18% do total de proteínas submetidas ao agrupamento pelo OrthoMCL. As proteínas possíveis parálogas representaram uma fração bem menor: 2,81% o que corresponde a 1.909 proteínas do grupo 1 (Figura 3).

Para o grupo 2, obteve-se 14.673 proteínas agrupadas em *clusters* de genes ortólogos, ou seja, 96,40% do total de proteínas submetidas ao

agrupamento pelo OrthoMCL. As proteínas possíveis parálogas representaram uma fração bem menor: 3,60% o que corresponde a 548 proteínas do grupo 2 (Figura 3).

O agrupamento das proteínas ortólogas em *clusters* contendo somente uma espécie até *clusters* que continham todas as espécies permite inferir o genoma *core* dos grupos além das possíveis proteínas espécie-específicas. O genoma *core* é composto pelo conjunto de proteínas que está presente em todos os organismos do grupo.

Para o grupo 1, foram agrupadas 11 linhagens de *S. cerevisiae*, portanto o genoma *core* é composto por 60.322 proteínas que estão presentes em todas as linhagens, o que significa 88,96% do total do proteoma predito para o grupo 1 (Tabela 7). O grupo 2, que possui as 3 espécies de *Kluyveromyces* sp., o genoma *core* é composto pelas 13.798 proteínas que estão presentes em todas as três espécies, o que significa 96,40% do total do proteoma predito para o grupo 2 (Tabela 7).

Tabela 7: Distribuição dos grupos de ortólogos para os grupos 1 e 2.

N° taxas (espécies)	N° de *clusters* Grupo 1	N° Proteínas Grupo 1	N° de *clusters* Grupo 2	N° Proteínas Grupo 2
1 taxa	21	54	53	152
2 taxa	159	328	338	723
3 taxa	83	287	4.494	13.798
4 taxa	88	375	-	-
5 taxa	105	536	-	-
6 taxa	42	267	-	-
7 taxa	33	238	-	-
8 taxa	41	357	-	-
9 taxa	58	585	-	-
10 taxa	237	2.544	-	-
11 taxa	5.233	60.332	-	-

Como o grupo 1 contém linhagens pertencentes à mesma espécie, era esperado que o genoma *core* compartilhado entre elas fosse maior que o genoma *core* do grupo 2. No entanto, isso não aconteceu como descrito

anteriormente. Isso pode ser devido à diferença discrepante do número de espécies presentes nos grupos 1 e 2.

A construção de grupos de ortólogos para organismos eucariotos incluem complicações como: duplicação de genes e redundância de funcionalidade, proteínas com estruturas de multi-domínios e a existência de sequências genômicas incompletas depositadas nos bancos de dados (Doolittle el at., 1995; Heinkoff et al., 1997). Dessa forma, o algoritmo OrthoMCL ao conseguir distinguir parálogos recentes (ex: duplicação gênica ocorrendo depois de uma especiação, como as múltiplas β–tubulinas encontradas no genoma humano) dos parálogos antigos (duplicação antes da especiação) proporciona a distinção desses grupos de acordo com suas diferentes funções. O algoritmo permite distinguir a redudância funcional da divergência através da inclusão dos parálogos recentes em grupos de ortólogos que possuem um melhor *hit* dentro da própria espécie que com outras espécies (Li et al., 2003). Neste sentido, a análise dos grupos separadamente permitiu identificar os parálogos recentes das linhagens dos grupos 1 e 2.

5.3.1. Análise dos *clusters* do Grupo 1: Linhagens de *S. cerevisiae*

As 65.909 sequências de proteínas ortólogas do grupo 1 resultaram em uma matriz (65909 X 65909) para realização do BLAST *all-against-all*, no qual todas as sequências foram comparadas contra todas através de um blastp. Os resultados do BLAST permitiram identificar sequências *inparalogs*, conhecidos como parálogos recentes, para todas as linhagens (Tabela 8). As matrizes utilizadas para a obtenção dos parálogos recentes encontram-se na seção de Anexos (Anexos 1 e 2).

Tabela 8: Identificação de parálogos recentes para os organismos do grupo 1.

Organismos	Sequências com o melhor *hit* dentro dessa espécie	Pares de sequências identificadas como Melhor *Hit*/Melhor Recíproco
S. cerevisiae S288c	656	1.663
S. cerevisiae CBS 7960	544	332
S. cerevisiae JAY 291	387	232
S. cerevisiae M3707	419	300
S. cerevisiae M3836	410	297
S. cerevisiae M3837	420	298
S. cerevisiae M3838	417	319
S. cerevisiae M3839	410	306
S. cerevisiae NY1308	471	301
S. cerevisiae ZTW1	421	290
S. boulardii EDRL	425	279

A Tabela 8 representa o número de sequências identificadas como parálogos recentes dentro de cada espécie, ganhando destaque a linhagem *S. cerevisiae* S288c para o maior número de sequências com o melhor *hit* com esta levedura, ou seja, 656 sequências. As linhagens *S. cerevisiae* CBS 7960 e *S. cerevisiae* NY 1308 ficaram com o segundo (544) e terceiros (471) maiores números de parálogos recentes, respectivamente. A linhagem *S. cerevisiae* JAY 291 retornou o menor valor de parálogos recentes, 387 sequências. As outras linhagens apresentaram valores similares de parálogos recentes.

Quanto aos pares de sequências identificadas como melhor *hit*/melhor recíproco dentro das linhagens, novamente, a levedura *S. cerevisiae* S288c destacou-se com 1.633 *hit*s. Em segundo lugar, ficou a linhagem *S. cerevisiae* CBS 7960 e em terceiro a linhagem *S. cerevisiae* M3838. A linhagem *S. cerevisiae* JAY 291 obteve o menor número de sequências com o melhor *hit* (Tabela 8). Em conjunto, estes dados mostram que os parálogos recentes estão presentes de maneira considerável nos genomas destes organismos e que a duplicação do genoma de *S. cerevisiae* pode ser estudada também através da exploração dos parálogos recentes.

5.3.2. Análise dos *clusters* do Grupo 2: Linhagens pertencentes ao gênero *Kluyveromyces* sp.

As 14.673 sequências de proteínas ortólogas do grupo 2 resultaram em uma matriz (14673 X 14673) para realização do BLAST *all-against-all*, no qual todas as sequências foram comparadas contra todas através de um blastp. Os resultados do BLAST permitiram identificar sequências *inparalogs*, conhecidos como parálogos recentes, para todas as espécies (Tabela 9). As matrizes utilizadas para a obtenção dos parálogos recentes encontram-se na seção de Anexos (Anexos 3 e 4).

Tabela 9: Identificação de parálogos recentes para os organismos do grupo 2.

Organismos	Sequências com o melhor *hit* dentro dessa espécie	Pares de sequências identificadas como Melhor *Hit*/Melhor Recíproco
K. lactis NRRL Y-1140	235	165
K. marxianus KCTC 17555	239	156
L. thermotolerans CBS 6340	413	288

A Tabela 9 representa o número de sequências identificadas como parálogos recentes dentro de cada espécie, sendo que *L. thermotolerans* CBS 6340 possui o maior número de sequências com o melhor *hit* com esta levedura, ou seja, 413 sequências recentes. As outras duas espécies possuem uma quantidade similar de sequências dentro de cada espécie.

Quanto aos pares de sequências identificadas como melhor *hit*/melhor recíproco dentro das linhagens, novamente, a levedura *L. thermotolerans* CBS 6340 destacou-se com 288 *hit*s. As outras duas espécies, *K. lactis* NRRL Y-1140 e *K. marxianus* KCTC 17555, obtiveram valores similares de sequências com o melhor *hit*.

Em relação ao grupo 1, o grupo 2 apresentou valores significavamente menores para os parálogos recentes, aqueles que duplicaram após a especiação. Isso pode ser explicado pelo fato de que as espécies do grupo 2 não passaram pelo fenômeno de duplicação do

genoma inteiro WGD, sendo conhecidas como leveduras protoplóides (Souciet et al., 2009).

5.4. Criação de Banco de Dados Local

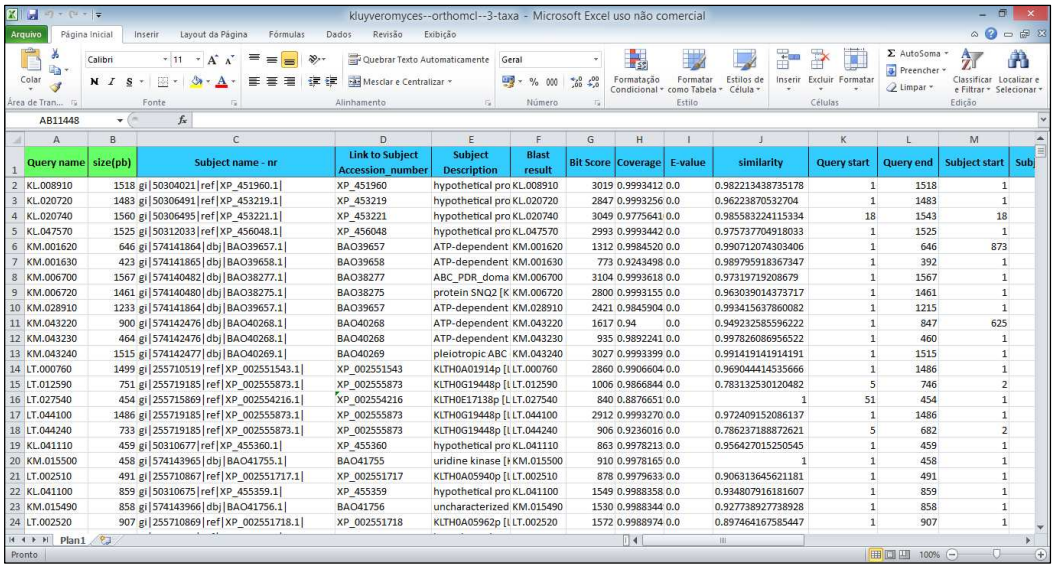
As informações obtidas através da re-anotação dos genomas e dos agrupamentos pelo OrthoMCL foram integradas e organizadas na forma de banco de dados local. O banco de dados local foi produzido para os dois grupos de proteomas preditos, sendo que para cada grupo foi feita a separação das sequências de acordo com o número de espécies agrupadas.

Para o grupo 1 (linhagens de *S. cerevisiae*) foram construídas 11 planilhas hiperlincadas contendo as sequências dos agrupamentos de 1 taxa até 11 taxa. Para o grupo 2 (espécies de *Kluyveromyces*) foram construídas 3 planilhas hiperlincadas contendo as sequências dos agrupamentos de 1 taxa até 3 taxa. Na Figura 4 podemos observar o layout das planilhas criadas. Para ambos os grupos, as células das planilhas retornavam as seguintes informações:

- *Query name:* identificador da sequência da proteína e link de acesso à sequência. Ex: SC_CBS7960.063270;
- *Size (aa):* quantidade de aminoácidos da proteína;
- *Subject name - nr/ Subject name - cdd:* identificador da sequência através de busca por similaridade nos bancos CDD e NR. Ex: gi|207342317|gb|EDZ70110.1|;
- *Link to Subject Accession_number:* link para acesso a sequência identificada no NCBI;
- *Subject Description:* descrição da proteína de acordo com o banco utilizado CDD ou NR. Ex: YMR109Wp-like protein [Saccharomyces cerevisiae AWRI1631]Myo5p [Saccharomyces cerevisiae EC1118]Myo5p [Saccharomyces cerevisiae P301];

- *Blast result:* link para acesso local ao BLAST realizado para a sequência.
- *Bit Score:* é uma pontuação normalizada expressa em bits que permite estimar a magnitude do espaço de busca, apresentando a significância da similaridade;
- *Coverage:* é a cobertura dada em porcentagem da sequência busca obtida em alinhamentos;
- *E-value:* é o parâmetro que define a chance ou a probabilidade do *hit* ocorrer ao acaso;
- *Similarity:* é o valor dado em porcentagem da pontuação positiva na matriz de substituição;
- *Query start:* é o aminoácido de início da sequência buscada;
- *Query end:* é o aminoácido final da sequência buscada;
- *Subject start:* é o aminoácido de início da sequência correspondente no banco de dados;
- *Subject end:* é o aminoácido final da sequência correspondente no banco de dados.

A criação desse banco de dados abre infinitas possibilidades de pesquisa. Este banco de dados está disponibilizado para o Laboratório de Biotecnologia Molecular/UFV e para o Grupo Informática de Biossistemas/Fiocruz. No entanto, colaborações futuras podem ser estabelecidas para um melhor aproveitamento do banco de dados e divulgação dos dados obtidos. Outra possibilidade que está sendo considerada é a divulgação *online* dos dados através da criação de um banco de dados.



kluyveromyces--orthomcl--3-taxa - Microsoft Excel uso não comercial

	Query name	size(pb)	Subject name - nr	Link to Subject Accession_number	Subject Description	Blast result	Bit Score	Coverage	E-value	similarity	Query start	Query end	Subject start	Subj
2	KL.008910	1518	gi|50304021|ref|XP_451960.1|	XP_451960	hypothetical pro	KL.008910	3019	0.9993412	0.0	0.982213438735178	1	1518	1	
3	KL.020720	1483	gi|50306491|ref|XP_453219.1|	XP_453219	hypothetical pro	KL.020720	2847	0.9993256	0.0	0.96223870532704	1	1483	1	
4	KL.020740	1560	gi|50306495|ref|XP_453221.1|	XP_453221	hypothetical pro	KL.020740	3049	0.9775641	0.0	0.985583224115334	18	1543	18	
5	KL.047570	1525	gi|50312033|ref|XP_456048.1|	XP_456048	hypothetical pro	KL.047570	2993	0.9993442	0.0	0.975737704918033	1	1525	1	
6	KM.001620	646	gi|574141864|dbj|BAO39657.1|	BAO39657	ATP-dependent	KM.001620	1312	0.9984520	0.0	0.990712074303406	1	646	873	
7	KM.001630	423	gi|574141865|dbj|BAO39658.1|	BAO39658	ATP-dependent	KM.001630	773	0.9243498	0.0	0.989795918367347	1	392	1	
8	KM.006700	1567	gi|574140482|dbj|BAO38277.1|	BAO38277	ABC_PDR_doma	KM.006700	3104	0.9993618	0.0	0.97319719208679	1	1567	1	
9	KM.006720	1461	gi|574140480|dbj|BAO38275.1|	BAO38275	protein SNQ2 [K	KM.006720	2800	0.9993155	0.0	0.963039014373717	1	1461	1	
10	KM.028910	1233	gi|574141864|dbj|BAO39657.1|	BAO39657	ATP-dependent	KM.028910	2421	0.9845904	0.0	0.993415637860082	1	1215	1	
11	KM.043220	900	gi|574142476|dbj|BAO40268.1|	BAO40268	ATP-dependent	KM.043220	1617	0.94	0.0	0.949232585596222	1	847	625	
12	KM.043230	464	gi|574142476|dbj|BAO40268.1|	BAO40268	ATP-dependent	KM.043230	935	0.9892241	0.0	0.997826086956522	1	460	1	
13	KM.043240	1515	gi|574142477|dbj|BAO40269.1|	BAO40269	pleiotropic ABC	KM.043240	3027	0.9993399	0.0	0.991419141914191	1	1515	1	
14	LT.000760	1499	gi|255710519|ref|XP_002551543.1|	XP_002551543	KLTH0A01914p [L	LT.000760	2860	0.9906604	0.0	0.969044414535666	1	1486	1	
15	LT.012590	751	gi|255719185|ref|XP_002555873.1|	XP_002555873	KLTH0G19448p [L	LT.012590	1006	0.9866844	0.0	0.783132530120482	5	746	2	
16	LT.027540	454	gi|255715869|ref|XP_002554216.1|	XP_002554216	KLTH0E17138p [L	LT.027540	840	0.8876651	0.0	1	51	454	1	
17	LT.044100	1486	gi|255719185|ref|XP_002555873.1|	XP_002555873	KLTH0G19448p [L	LT.044100	2912	0.9993270	0.0	0.972409152086137	1	1486	1	
18	LT.044240	733	gi|255719185|ref|XP_002555873.1|	XP_002555873	KLTH0G19448p [L	LT.044240	906	0.9236016	0.0	0.786237188872621	5	682	2	
19	KL.041110	459	gi|50310677|ref|XP_455360.1|	XP_455360	hypothetical pro	KL.041110	863	0.9978213	0.0	0.956427015250545	1	459	1	
20	KM.015500	458	gi|574143965|dbj|BAO41755.1|	BAO41755	uridine kinase [K	KM.015500	910	0.9978165	0.0	1	1	458	1	
21	LT.002510	491	gi|255710867|ref|XP_002551717.1|	XP_002551717	KLTH0A05940p [L	LT.002510	878	0.9979633	0.0	0.906313645621181	1	491	1	
22	KL.041100	859	gi|50310675|ref|XP_455359.1|	XP_455359	hypothetical pro	KL.041100	1549	0.9988358	0.0	0.934807916181607	1	859	1	
23	KM.015490	858	gi|574143966|dbj|BAO41756.1|	BAO41756	uncharacterized	KM.015490	1530	0.9988344	0.0	0.927738927738928	1	858	1	
24	LT.002520	907	gi|255710869|ref|XP_002551718.1|	XP_002551718	KLTH0A05962p [L	LT.002520	1572	0.9988974	0.0	0.897464167585447	1	907	1	

Figura 4: Representação de parte da planilha hiperlincada do grupo 2 de acordo com o proteoma predito presente nos três taxas: *K. lactis* NRRL Y-1140, *K. marxianus* KCTC 17555 e *L. thermotolerans* CBS 6340.

5.5. Comparação Filogenética: Proteínas Álcoois Desidrogenases

A criação do banco de dados local facilitou o desenvolvimento de um trabalho de comparação filogenética entre sequências de álcoois desidrogenases dos organismos investigados. Dessa forma, foram realizadas comparações das sequências das 7 proteínas álcoois desidrogenases de *S. cerevisiae* disponíveis no banco de dados SGD para os dois grupos separadamente.

É importante destacar que a busca por ADHs nos genomas foi realizada com dois diferentes valores de *e-value*, para que se pudessem obter dados mais estringentes, no caso do valor de *e-value* ser 10^{-6}, e menos estringentes, 10^{-4}. Essa estratégia foi utilizada para que fossem recuperadas todas as sequências de ADHs dos genomas. No entanto, não houve diferença nos resultados obtidos para ambos os grupos.

A análise dos grupos foi baseada nas sequências que foram preditas pelo *pipeline* de anotação, confirmadas pela similaridade contra ADHs do banco de dados curado SGD, e, agrupadas pelo OrthoMCL. A busca de ADHs por similaridade através do BLAST, utilizando como *query* as sequências de ADHs do SGD e como *subject* os genomas das espécis,

retornou sequências de ADHs que não haviam sido agrupadas. Portanto, estas sequências ficaram fora da análise dos alinhamentos múltiplos.

Através dos identificadores das sequências de ADHs confirmadas pelos resultados do BLAST, foram recuperados os *clusters* ao qual faziam parte, e então foram realizados os alinhamentos múltiplos.

5.5.1. Comparação filogenética de ADHs do Grupo 1

Para o grupo 1, as sequências de ADHs recuperadas dos *clusters* foram alinhadas de acordo com o tipo de ADH do *cluster*, portanto foram construídos 7 alinhamentos, um para cada tipo de ADH (Figuras 5 a 11). Além disso, ao final foi feito um alinhamento global de todas as sequências de ADHs em uma única análise (Figura 12).

Para estas análises foram recuperadas as sequências de proteínas de 8 *clusters* que haviam sido agrupados de acordo com a ortologia existente entre eles. Os *clusters* foram: Orthomcl571, Orthomcl902, Orthomcl3080, Orthomcl3504, Orthomcl5422, Orthomcl5451, Orthomcl5520 e Orthomcl5585.

O *cluster* Orthomcl571 era composto por 11 sequências de proteínas. Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD1), que é membro da família de redutases/desidrogenases de cadeia média, cujo identificador é cd05283. Estas proteínas estão envolvidas na redução de cinamildeídos a cinamil álcoois no último passo no metabolismo de “monolignal” na parede das células. O banco de dados NR identificou estas proteínas com similaridade com álcool desidrogenase VI (ADH6) e similaridade acima de 95%.

O *cluster* Orthomcl902 era composto por 11 sequências de proteínas. Uma observação curiosa a respeito desse grupo é o fato que de acordo com a similaridade pesquisada no CDD, o grupo possui um domínio LPO, oxidoredutase lactaldeído: propanodiol, que é um domínio típico de enterobactérias. Essa incoerência pode ser devido ao fato de possuir uma desidrogenase ativado por ferro III, que em enterobactérias, catalisa a interconversão de L-lactaldeído e L-1,2-propanodiol. Por outro lado, o banco de dados NR identificou as proteínas do *cluster* Orthomcl902 com

similaridade com álcool desidrogenase IV (ADH4) e similaridade acima de 93%. No entanto, essa predição inesperada pelo CDD, pode ser explicada, de acordo com o banco de dados SGD, pelo fato de que a ADH4 é uma enzima dimérica que demonstra ter atividade zinco-dependente, apesar da semelhança com desidrogenase ativado por ferro.

Os *clusters* Orthomcl3080 e Orthomcl3504 eram composto por 11 sequências de proteínas. Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD3), que é membro da família de redutases/desidrogenases de cadeia média, cujo identificador é cd08297. Tal como CAD1, as proteínas CAD3 estão envolvidas na redução de cinamildeídos a cinamil álcoois no último passo no metabolismo de "monolignal" na parede das células. O banco de dados NR identificou as proteínas do Orthomcl3080 com similaridade com álcool desidrogenase III (ADH3) e similaridade acima de 92%. As proteínas do Orthomcl3504 apresentaram com similaridade com álcool desidrogenase V (ADH5) e similaridade acima de 99%.

Os *clusters* Orthomcl5422 e Orthomcl5451 eram compostos por 10 sequências de proteínas. Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD3), cujo identificador no CDD é cd08297. Os dois *clusters* não apresentaram sequência relativa à *S. cerevisiae* JAY 291. Para Orthomcl5422 foram identificadas as proteínas com similaridade com álcool desidrogenase I (ADH1) e similaridade acima de 91% com o banco de dados NR. Para Orthomcl 5451, o banco de dados NR identificou estas proteínas com similaridade com álcool desidrogenase II (ADH2) e similaridade acima de 90%.

O *cluster* Orthomcl5520 era composto por 9 sequências de proteínas. Neste *cluster* não foram agrupadas sequências relativas às linhagens *S. cerevisae* JAY 291 e *S. cerevisae* NY 1308. Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD1), cujo identificador é cd05283. O banco de dados NR identificou estas proteínas com similaridade com álcool desidrogenase VII (ADH7) e similaridade acima de 90%.

O *cluster* Orthomcl5585 era composto por 7 sequências de proteínas. Neste *cluster* não foram agrupadas sequências relativas às

linhagens *S. cerevisae* CBS 7960, *S. cerevisae* JAY 291, *S. cerevisae* ZTW1 e *S. boulardii* EDRL. Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD1), cujo identificador é cd05283. O banco de dados NR identificou estas proteínas com similaridade com álcool desidrogenase VII (ADH7) e similaridade acima de 68%.

Adicionalmente, existiam 3 sequências de ADHs do grupo 1 que não foram agrupadas em *clusters* e que foram preditas pela comparação com o banco de dados SGD, são elas: SC_JAY.032860, SC_JAY.040400 e SC_JAY.043970. Estas sequências haviam sido preditas como ADH1, ADH2 e ADH6, e possuíam um tamanho de 180, 155 e 363 resíduos de aminoácidos, respectivamente. Possivelmente, as sequências de ADH1 e ADH2 de *S. cerevisae* JAY 291 não foram agrupadas nos *clusters* pela falta do domínio CAD que confere a função desidrogenase a essas proteínas. Estas sequências estão truncadas, faltando a parte C-terminal, já que ADH1 e ADH2 de *S. cerevisae* do SGD contém 348 resíduos de aminoácidos.

Quanto ao ADH6, *S. cerevisae* JAY 291 apresentou duas sequências de proteínas, uma agrupada no *cluster* Orthomcl571 e a outra sequência predita, SC_JAY.043970, mas não agrupada em *clusters*. Para linhagem *S. cerevisiae* JAY 291 também não foi encontrada a sequência do ADH7. A checagem dos dados originais revelou que as sequências ADH1, ADH2 e ADH7 não haviam sido preditas como tais, necessitando de uma análise mais aprofundada no genoma.

As 3 sequências de *S. cerevisae* JAY 291 que não foram agrupadas pelo OrthoMCL foram adicionadas ao alinhamento múltiplo para confirmar nossa hipótese sobre a exclusão das sequências deste agrupamento.

5.5.1.1. ADH1

Para analisar a relação filogenética entre os ADH1 do grupo 1 foram executados dois alinhamentos múltiplos separadamente. O primeiro alinhamento contava somente as 10 sequências presentes no *cluster* Orthomcl5422 (Anexo 5). Ao segundo alinhamento foi adicionada a

sequência predita SC_JAY.032860, às sequências do *cluster* Orthomcl5422, dessa forma contendo todas as sequências de ADH1 (Anexo 6).

Observa-se que nos alinhamentos obtidos é clara a diferença de tamanho entre a sequência SC_JAY.032860, que apresentava somente 180 resíduos de aminoácidos, e as outras sequências que foram agrupadas no *cluster* Orthomcl5422, que possuíam 348 resíduos de aminoácidos, tal como o ADH1 do banco de dados curado SGD (Anexos 5 e 6).

As matrizes resultantes desses dois alinhamentos reiteram a informação já evidenciada nos alinhamentos, através de números (Figuras 5 e 6). Além disso, as matrizes possibilitam uma maior facilidade para comparar as identidades existentes entre as sequências e a diferença nos resíduos de aminoácidos entre elas.

		1	2	3	4	5	6	7	8	9	10
SC_CBS7960.051600	1		100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,71	98,85
SC_M3707.010430	2	0		100,00	100,00	100,00	100,00	100,00	100,00	99,71	98,85
SC_M3836.000720	3	0	0		100,00	100,00	100,00	100,00	100,00	99,71	98,85
SC_M3837.010500	4	0	0	0		100,00	100,00	100,00	100,00	99,71	98,85
SC_M3838.012460	5	0	0	0	0		100,00	100,00	100,00	99,71	98,85
SC_M3839.018130	6	0	0	0	0	0		100,00	100,00	99,71	98,85
SC_ZTW1.050710	7	0	0	0	0	0	0		100,00	99,71	98,85
Sb.007250	8	0	0	0	0	0	0	0		99,71	98,85
SC_NY1308.022930	9	1	1	1	1	1	1	1	1		99,14
SC_S288c.052710	10	4	4	4	4	4	4	4	4	3	

Figura 5: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas do *cluster* Orthomcl5422 de ADH1 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

		1	2	3	4	5	6	7	8	9	10	11
SC_CBS7960.051600	1		100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,71	98,85	51,72
SC_M3707.010430	2	0		100,00	100,00	100,00	100,00	100,00	100,00	99,71	98,85	51,72
SC_M3836.000720	3	0	0		100,00	100,00	100,00	100,00	100,00	99,71	98,85	51,72
SC_M3837.010500	4	0	0	0		100,00	100,00	100,00	100,00	99,71	98,85	51,72
SC_M3838.012460	5	0	0	0	0		100,00	100,00	100,00	99,71	98,85	51,72
SC_M3839.018130	6	0	0	0	0	0		100,00	100,00	99,71	98,85	51,72
SC_ZTW1.050710	7	0	0	0	0	0	0		100,00	99,71	98,85	51,72
Sb.007250	8	0	0	0	0	0	0	0		99,71	98,85	51,72
SC_NY1308.022930	9	1	1	1	1	1	1	1	1		99,14	51,72
SC_S288c.052710	10	4	4	4	4	4	4	4	4	3		51,72
SC_JAY291.032860	11	168	168	168	168	168	168	168	168	168	168	

Figura 6: Matriz resultante do alinhamento múltiplo entre todas as sequências de proteínas ADH1 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

Na Figura 5 podemos observar que as linhagens *S. cerevisiae* S288c e *S. cerevisiae* NY 1308 possuem resíduos de aminoácidos diferentes das demais linhagens. A linhagem *S. cerevisiae* S288c possui 4 resíduos de aminoácidos diferentes, quando comparada às linhagens de *S. cerevisiae* CBS 7960, M3707, M3836, M3837, M3838, M3839, ZTW1 e *S. boulardii* EDRL, e 3 resíduos diferentes quando comparada à linhagem NY 1308.

O resíduo de aminoácido 59 da linhagem *S. cerevisiae* S288c é a substituição de um triptofano (T) por uma valina (V). Esta substituição tem caráter conservativo por serem dois aminoácidos polares. Os resíduos de aminoácidos nas posições 129 e 159 da linhagem *S. cerevisiae* S288c são substiuições de um ácido glutâmico (E) por uma glutamina (Q), sendo uma substituição do tipo não conservativa, um aminoácido polar ácido por um aminoácido polar neutro. A substituição conservativa de valina (V) por isoleucina (I) ocorre tanto em *S. cerevisiae* S288c quanto em NY 1308, e, pelo fato da natureza química dos aminoácidos serem similares, diminui a chance de uma alteração grave na estrutura da proteína (Figura 5). Estas substituições estão refletidas na variação da porcentagem de identidade entre as sequências, entre 98,85% a 100% de identidade. Considerando que a linhagem *S. cerevisiae* S288c foi extensamente estudada e sequenciada, acreditamos que estes 4 resíduos diferentes possam estar relacionados à característica laboratorial da linhagem. Quanto aos 3 resíduos diferentes em *S. cerevisiae* NY 1308, não podemos inferir tal

informação. Visto que este genoma foi sequenciado somente uma vez e pode conter erros no nomeamento das bases.

A proteína referente à SC_JAY291.032860 não possui parte de sua sequência. Está faltando a sequência inicial dessa proteína, 168 resíduos de aminoácidos iniciais (Anexo 6), o que levou a uma baixa porcentagem de identidade, 51,72%, em comparação às demais linhagens (Figura 6). A parte faltante dessa proteína poderia estar em outro *contigs*, contudo devido aos estringentes parâmetros utilizados ela não foi recuperada. Uma alternativa a esta suposição levaria ao fato dessa sequência estar em uma extremidade de um *contigs* e ter sido descartada por causa de baixos valores de qualidade. Com isso, sugere-se um re-sequenciamento dessa proteína, para obter a sequência correta do ADH1 de *S. cerevisiae* JAY 291, uma vez que não foi possível afimar que SC_JAY291.032860 seria o ADH1. De modo suplementar, esta análise evidencia a necessidade de uma correta montagem, ordenação e orientação de *contigs*, para que não ocorram interferências na análise final dos dados genômicos.

5.5.1.2. ADH2

Para analisar a relação filogenética entre os ADH2 do grupo 1 foram executados dois alinhamentos múltiplos separadamente. O primeiro alinhamento contava somente as 10 sequências presentes no *cluster* Orthomcl5451 (Anexo 7). Ao segundo alinhamento foi adicionada a sequência predita SC_JAY.040400, às sequências do *cluster* Orthomcl5451, dessa forma contendo todas as sequências de ADH2 (Anexo 8).

Observa-se que nos alinhamentos obtidos é clara a diferença de tamanho entre a sequência SC_JAY.040400, que apresentava a mais 160 resíduos de aminoácidos, e as outras sequências que foram agrupadas no *cluster* Orthomcl5451, que possuíam 348 resíduos de aminoácidos, tal como o gene *ADH2* do banco de dados curado SGD (Anexos 7 e 8).

Na Figura 7 podemos observar que somente a linhagem diplóide *S. cerevisiae* M3707 e seus quatro derivados haplóides M3836, M3837, M3838 e M3839, provenientes de destiladores comerciais, possuíam

sequências idênticas. As outras linhagens possuíam diferença entre os resíduos de aminoácidos variando entre 2 a 12 resíduos. A linhagem ZTW1 seria mais similar às linhagens de destiladores comerciais, sequenciada por Brown e colaboradores em 2013 (Brown et al., 2013). Em seguida, a linhagem *S. cerevisiae* S288c seria mais similar, contendo uma variação de 5 resíduos de aminoácidos e identidade de 98,56%. As linhagens com maior quantidade de resíduos diferentes foram *S. cerevisiae* NY 1308 e *S. boulardii* EDRL.

A Figura 8 evidencia os 160 resíduos a mais que estão presentes na sequência SC_JAY.040400, o que torna baixa a sua identidade em comparação com as outras linhagens. Podemos perceber que houve uma união equivocada de 160 resíduos de aminoácidos no início de SC_JAY.040400, os quais não permitiram que esta sequência fosse corretamente agrupada com o *cluster* Orthomcl5451.

		1	2	3	4	5	6	7	8	9	10
SC_CBS7960.035500	1		100,00	100,00	100,00	100,00	100,00	98,56	99,14	97,41	97,41
SC_M3707.011390	2	0		100,00	100,00	100,00	100,00	98,56	99,14	97,41	97,41
SC_M3836.009940	3	0	0		100,00	100,00	100,00	98,56	99,14	97,41	97,41
SC_M3837.011400	4	0	0	0		100,00	100,00	98,56	99,14	97,41	97,41
SC_M3838.026780	5	0	0	0	0		100,00	98,56	99,14	97,41	97,41
SC_M3839.037070	6	0	0	0	0	0		98,56	99,14	97,41	97,41
SC_S288c.047490	7	5	5	5	5	5	5		99,43	98,28	96,55
SC_ZTW1.045720	8	3	3	3	3	3	3	2		98,28	97,13
Sb.022250	9	9	9	9	9	9	9	6	6		97,70
SC_NY1308.017400	10	9	9	9	9	9	9	12	10	8	

Figura 7: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas do *cluster* Orthomcl5451 de ADH2 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

		1	2	3	4	5	6	7	8	9	10
SC_JAY291.040400	1		68,50	68,50	68,50	68,50	68,50	67,52	67,91	66,73	66,73
SC_M3707.011390	2	160		100,00	100,00	100,00	100,00	98,56	99,14	97,41	97,41
SC_M3836.009940	3	160	0		100,00	100,00	100,00	98,56	99,14	97,41	97,41
SC_M3837.011400	4	160	0	0		100,00	100,00	98,56	99,14	97,41	97,41
SC_M3838.026780	5	160	0	0	0		100,00	98,56	99,14	97,41	97,41
SC_M3839.037070	6	160	0	0	0	0		98,56	99,14	97,41	97,41
SC_S288c.047490	7	165	5	5	5	5	5		99,43	98,28	96,55
SC_ZTW1.045720	8	163	3	3	3	3	3	2		98,28	97,13
Sb.022250	9	169	9	9	9	9	9	6	6		97,70
SC_NY1308.017400	10	169	9	9	9	9	9	12	10	8	

Figura 8: Matriz resultante do alinhamento múltiplo entre todas as sequências de proteínas ADH2 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

5.5.1.3. ADH3

Para analisar a relação filogenética entre os ADH3 do grupo 1 foi realizado um único alinhamento múltiplo, contendo todas as 11 sequências de ADH3 agrupadas no *cluster* Orthomcl3080 (Anexo 9).

Para ADH3, alinhamento múltiplo mostrou que todas as 11 linhagens do Grupo 1 possuíam esta proteína com 375 resíduos de aminoácidos (Anexo 9). A linhagem *S. cerevisiae* NY 1308 sofreu uma substituição do aminoácido prolina, conservado nas outras 10 linhagens, por arginina no resíduo 54. Essa substituição necessitaria ser confirmada por um re-sequenciamento, uma vez que foi uma substituição do tipo não conservativa, uma vez que a prolina é um aminoácido apolar e a argingina é uma aminoácido polar carregado positivamente.

Na Figura 9, pode-se verificar que todas as sequências possuíram uma identidade de 100%, com exceção para *S. cerevisiae* NY 1308, que apresentou um percentual de identidade de 99,73% entre essa sequência e as demais linhagens, justamente por possuir 1 resíduo de aminoácido diferente. Nossos resultados estão de acordo com o banco de dados SGD, no qual ADH3 de *S. cerevisae* possui 375 aminoácidos, o que vai de acordo com os resultados obtidos para este alinhamento.

		1	2	3	4	5	6	7	8	9	10	11
SC_CBS7960.029330	1		100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,73
SC_JAY291.009990	2	0		100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,73
SC_M3707.013760	3	0	0		100,00	100,00	100,00	100,00	100,00	100,00	100,00	99,73
SC_M3836.007600	4	0	0	0		100,00	100,00	100,00	100,00	100,00	100,00	99,73
SC_M3837.013750	5	0	0	0	0		100,00	100,00	100,00	100,00	100,00	99,73
SC_M3838.029130	6	0	0	0	0	0		100,00	100,00	100,00	100,00	99,73
SC_M3839.034710	7	0	0	0	0	0	0		100,00	100,00	100,00	99,73
SC_S288c.045180	8	0	0	0	0	0	0	0		100,00	100,00	99,73
SC_ZTW1.043430	9	0	0	0	0	0	0	0	0		100,00	99,73
Sb.035530	10	0	0	0	0	0	0	0	0	0		99,73
SC_NY1308.015070	11	1	1	1	1	1	1	1	1	1	1	

Figura 9: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas do *cluster* Orthomcl3080 de ADH3 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

5.5.1.4. ADH4

Para analisar a relação filogenética entre os ADH4 do grupo 1 foi realizado um único alinhamento múltiplo, contendo todas as 11 sequências de ADH4 agrupadas no *cluster* Orthomcl902 (Anexo 10).

O alinhamento múltiplo mostrou que somente 2 linhagens do Grupo 1, *S. cerevisiae* S288c e *S. cerevisiae* ZTW1, possuíam esta proteína com 382 resíduos de aminoácidos, assim como a ADH4 curada do banco SGD (Anexo 10). As outras 9 linhagens possuíam uma sequência de 465 resíduos de aminoácidos, ou seja, um adicional de 83 resíduos de aminoácidos erroneamente unidos no início da sequência.

A linhagem diplóide *S. cerevisiae* M3707 e seus quatro derivados haplóides M3836, M3837, M3838 e M3839, provenientes de destiladores comerciais, possuíam sequências idênticas com identidade de 100% (Figura 10). A diferença destas 5 linhagens para *S. cerevisiae* NY 1308 e demais linhagens foi a substituição do resíduo de aminoácido polar básico de arginina (R) por aminoácido polar neutro: a serina(S). As outras linhagens possuíam diferença entre os resíduos de aminoácidos variando entre 2 a 5 resíduos. Desconsiderando os 83 resíduos erroneamente

incorporados, a linhagem *S. boulardii* EDRL foi a linhagem que mais se diferenciou das outras.

		1	2	3	4	5	6	7	8	9	10	11
SC_M3707.029200	1		100,00	100,00	100,00	100,00	99,57	99,57	99,78	98,92	81,94	81,72
SC_M3836.023520	2	0		100,00	100,00	100,00	99,57	99,57	99,78	98,92	81,94	81,72
SC_M3837.029130	3	0	0		100,00	100,00	99,57	99,57	99,78	98,92	81,94	81,72
SC_M3838.022250	4	0	0	0		100,00	99,57	99,57	99,78	98,92	81,94	81,72
SC_M3839.013290	5	0	0	0	0		99,57	99,57	99,78	98,92	81,94	81,72
SC_CBS7960.059160	6	2	2	2	2	2		100,00	99,78	99,35	81,94	81,72
SC_JAY291.021250	7	2	2	2	2	2	0		99,78	99,35	81,94	81,72
SC_NY1308.049480	8	1	1	1	1	1	1	1		99,14	82,15	81,94
Sb.037910	9	5	5	5	5	5	3	3	4		81,29	81,08
SC_ZTW1.018900	10	84	84	84	84	84	84	84	83	87		99,74
SC_S288c.022660	11	85	85	85	85	85	85	85	84	88	1	

Figura 10: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas do *cluster* Orthomcl902 de ADH4 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

5.5.1.5. ADH5

Para analisar a relação filogenética entre os ADH5 do grupo 1 foi realizado um único alinhamento múltiplo, contendo todas as 11 sequências de ADH5 agrupadas no *cluster* Orthomcl3504 (Anexo 11).

A análise do alinhamento múltiplo mostrou que 10 linhagens do Grupo 1 possuíam esta proteína com 351 resíduos de aminoácidos, com exceção de *S. cerevisiae* ZTW1 que possuía 343 resíduos, faltando os 8 resíduos de aminoácidos iniciais (Anexo 11). Adicionalmente, houve uma substituição conservativa de um aminoácido polar neutro por outro, no caso específico foi glutamina (Q) por treonina (T) em *S. cerevisiae* ZTW (Anexo 11 e Figura 11).

		1	2	3	4	5	6	7	8	9	10	11
SC_CBS7960.023690	1		0	0	0	0	0	0	0	0	0	8
SC_JAY291.059280	2	0		0	0	0	0	0	0	0	0	8
SC_M3707.023340	3	0	0		0	0	0	0	0	0	0	8
SC_M3836.031690	4	0	0	0		0	0	0	0	0	0	8
SC_M3837.023290	5	0	0	0	0		0	0	0	0	0	8
SC_M3838.019370	6	0	0	0	0	0		0	0	0	0	8
SC_M3839.026270	7	0	0	0	0	0	0		0	0	0	8
SC_NY1308.034700	8	0	0	0	0	0	0	0		0	0	8
SC_S288c.007160	9	0	0	0	0	0	0	0	0		0	8
Sb.017730	10	0	0	0	0	0	0	0	0	0		8
SC_ZTW1.003720	11	9	9	9	9	9	9	9	9	9	9	

Figura 11: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas do *cluster* Orthomcl3504 de ADH5 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

Esses dados demonstram o quanto o gene *ADH5* é conservado em todas as linhagens, portanto *ADH5* não está ligado ao caráter industrial ou laboratorial das linhagens. Sabe-se que o gene *ADH5* possui um parálogo, *ADH1,* que emergiu do evento de duplicação genômica WGD (Byrne & Wolfe, 2005). Neste sentido, é importante destacar que o gene parálogo de *ADH5*, o *ADH1*, possui justamente na linhagem laboratorial 4 resíduos de aminoácidos diferentes das demais linhagens. Portanto, estas substituições podem ser assinaturas genéticas de leveduras industriais, devendo se melhor investigado.

5.5.1.6. ADH6

Para analisar a relação filogenética entre os ADH6 do grupo 1 foi realizado um único alinhamento múltiplo, contendo todas as 11 sequências de ADH6 agrupadas no *cluster* Orthomcl571 (Anexo 12).

A análise do alinhamento múltiplo mostrou que 10 linhagens do Grupo 1 possuíam esta proteína com 360 resíduos de aminoácidos, com exceção de *S. cerevisiae* CBS 7960 que possuía 372 resíduos, diferentemente dos 360 resíduos de ADH6 encontrados na proteína do banco SGD (Anexo 12).

Para facilitar a análise, vamos primeiramente focar nas 10 linhagens que possuem o mesmo tamanho de sequência: 360 resíduos de aminoácidos (Anexo 12). Estas linhagens possuem o ADH6 extremamente conservado com somente uma alteração conservativa em 5 exemplares (Figura 12). A linhagem diplóide *S. cerevisiae* M3707 e seus quatro derivados haplóides M3836, M3837, M3838 e M3839, provenientes de destiladores comerciais, possuíam uma única substituição no resíduo 277 de isoleucina (I) por valina (V). A confirmação dessa substituição, através do sequenciamento em outra plataforma NGS ou Sanger, não irá alterar o alto grau de conservação existente entre estas linhagens, que já possuem identidade de 99,72%. Isso indica que a proteína ADH6, a qual está envolvida em processos de resposta a stress (Tkach et al., 2012; Larroy et al., 2003), não caracteriza uma assinatura do caráter industrial ou laboratorial das linhagens.

A linhagem *S. cerevisiae* CBS 7960 até o resíduo 336 possui alta identidade com as outras sequências. Até esse ponto, a única substituição que ocorreu em CBS 7960 foi conservativa: modificação de uma isoleucina (I) por uma valina (V). No entanto, a parte terminal da sequência é completamente diferente dos outros ADH6, os últimos 36 resíduos provavelmente foram unidos de forma errônea a esta sequência. Como este genoma está muito fragmentado em diversos *contigs* e os parâmetros utilizados para corte foram estringentes não conseguimos localizar os resíduos finais dessa sequência.

		1	2	3	4	5	6	7	8	9	10	11
SC_JAY291.024590	1		100,00	100,00	100,00	100,00	99,72	99,72	99,72	99,72	99,72	90,59
SC_NY1308.017580	2	0		100,00	100,00	100,00	99,72	99,72	99,72	99,72	99,72	90,59
SC_S288c.047660	3	0	0		100,00	100,00	99,72	99,72	99,72	99,72	99,72	90,59
SC_ZTW1.045880	4	0	0	0		100,00	99,72	99,72	99,72	99,72	99,72	90,59
Sb.022420	5	0	0	0	0		99,72	99,72	99,72	99,72	99,72	90,59
SC_M3707.011230	6	1	1	1	1	1		100,00	100,00	100,00	100,00	90,32
SC_M3836.010110	7	1	1	1	1	1	0		100,00	100,00	100,00	90,32
SC_M3837.011240	8	1	1	1	1	1	0	0		100,00	100,00	90,32
SC_M3838.026620	9	1	1	1	1	1	0	0	0		100,00	90,32
SC_M3839.037240	10	1	1	1	1	1	0	0	0	0		90,32
SC_CBS7960.064420	11	35	35	35	35	35	36	36	36	36	36	

Figura 12: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas do *cluster* Orthomcl571 de ADH6 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior

esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

5.5.1.7. ADH7

Para analisar a relação filogenética entre os ADH7 do grupo 1 foi construído um alinhamento múltiplo com as sequências dos *clusters* Orthomcl5520 e Orthomcl5585. Decidimos unir os *clusters* para um único alinhamento pelo fato das sequências possuírem domínios conservados de ADH7. Dessa forma, foram unidas 9 sequências presentes no *cluster* Orthomcl5520 com as 7 sequências presentes no *cluster* Orthomcl5585 (Anexo 13).

De acordo com o banco de dados SGD, o gene *ADH7* possui 361 resíduos de aminoácidos, tal como as 9 sequências do *cluster* Orthomcl5520. Pela Figura 13 podemos entender o motivo da separação dos ADH7 em dois *clusters*. Primeiramente, existe uma discrepância no tamanho das sequências entre os dois *clusters*, o *cluster* Orthomcl5520 possui sequências com 361 resíduos de aminoácidos, enquanto que o *cluster* Orthomcl5585 possui sequências com diversos tamanhos, 66, 85 e 129 resíduos de aminoácidos. Isso indica que as 7 sequências do *cluster* Orthomcl5585 (*S. cerevisiae* M3707, M3836, M3837, M3838, M3839, NY 1308 e S288c) não sejam funcionais, podendo representar pseudogenes que permaneceram após o evento WGD.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
SC_M3707.020710	1		100,00	100,00	100,00	100,00	100,00	99,45	99,45	99,45	14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_M3836.054390	2	0		100,00	100,00	100,00	100,00	99,45	99,45	99,45	14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_M3837.020660	3	0	0		100,00	100,00	100,00	99,45	99,45	99,45	14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_M3838.021970	4	0	0	0		100,00	100,00	99,45	99,45	99,45	14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_M3839.023610	5	0	0	0	0		100,00	99,45	99,45	99,45	14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_ZTW1.007040	6	0	0	0	0	0		99,45	99,45	99,45	14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_CBS7960.053910	7	2	2	2	2	2	2		100,00	100,00	14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_S288c.010530	8	2	2	2	2	2	2	0		100,00	14,68	14,68	14,68	14,68	14,68	16,07	12,19
Sb.058660	9	2	2	2	2	2	2	0	0		14,68	14,68	14,68	14,68	14,68	16,07	12,19
SC_M3707.058570	10	308	308	308	308	308	308	308	308	308		100,00	100,00	100,00	100,00	65,89	77,65
SC_M3836.055880	11	308	308	308	308	308	308	308	308	308	0		100,00	100,00	100,00	65,89	77,65
SC_M3837.059350	12	308	308	308	308	308	308	308	308	308	0	0		100,00	100,00	65,89	77,65
SC_M3838.059090	13	308	308	308	308	308	308	308	308	308	0	0	0		100,00	65,89	77,65
SC_M3839.058550	14	308	308	308	308	308	308	308	308	308	0	0	0	0		65,89	77,65
SC_NY1308.048230	15	303	303	303	303	303	303	303	303	303	44	44	44	44	44		51,16
SC_S288c.021190	16	317	317	317	317	317	317	317	317	317	19	19	19	19	19	63	

Figura 13: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas dos *clusters* Orthomcl5520 e Orthomcl5585 de ADH7 do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes

encontrados entre as sequências.

Outro motivo para haver a separação dos ADH7 em *cluster* pode ser o alto grau de identidade entre as 9 sequências do *cluster* Orthomcl5520 (Figura 13). Só existem dois resíduos de aminoácidos diferentes, nas linhagens *S. cerevisiae* CBS 7960, *S. cerevisiae* S288c e *S. boulardii* EDRL, em relação à igualdade proteica das outras 6 sequências, *S. cerevisiae* M3707, M3836, M3837, M3838, M3839 e ZTW1. A primeira substituição conservativa ocorreu no resíduo 243, sendo um aminoácido polar neutro treonina (T) nas 5 linhagens por um resíduo de aminoácido polar neutro asparagina (N). A segunda modificação foi de caráter não conservativo, no qual um aminoácido polar básico, lisina (K) nas 5 linhagens, por um aminoácido polar ácido, glutamato (E).

A Figura 14 representa uma análise filogenética global de todas as sequências dos ADHs para as 11 linhagens do grupo 1. Dessa forma, podemos observar melhor a separação das sequências de acordo com o tipo do ADH. A sequência SC_JAY291.032860 havia sido alinhada juntamente com ADH1, entretanto a Figura 21 mostra que esta sequência está mais relacionada ao ADH2. Com isso, não foi encontrada a sequência ADH1 da linhagem *S. cerevisiae* JAY 291. Outra sequência que não foi predita para esta linhagem foi o ADH7. Dessa forma, um PCR poderia confirmar a existência destas sequências no genoma.

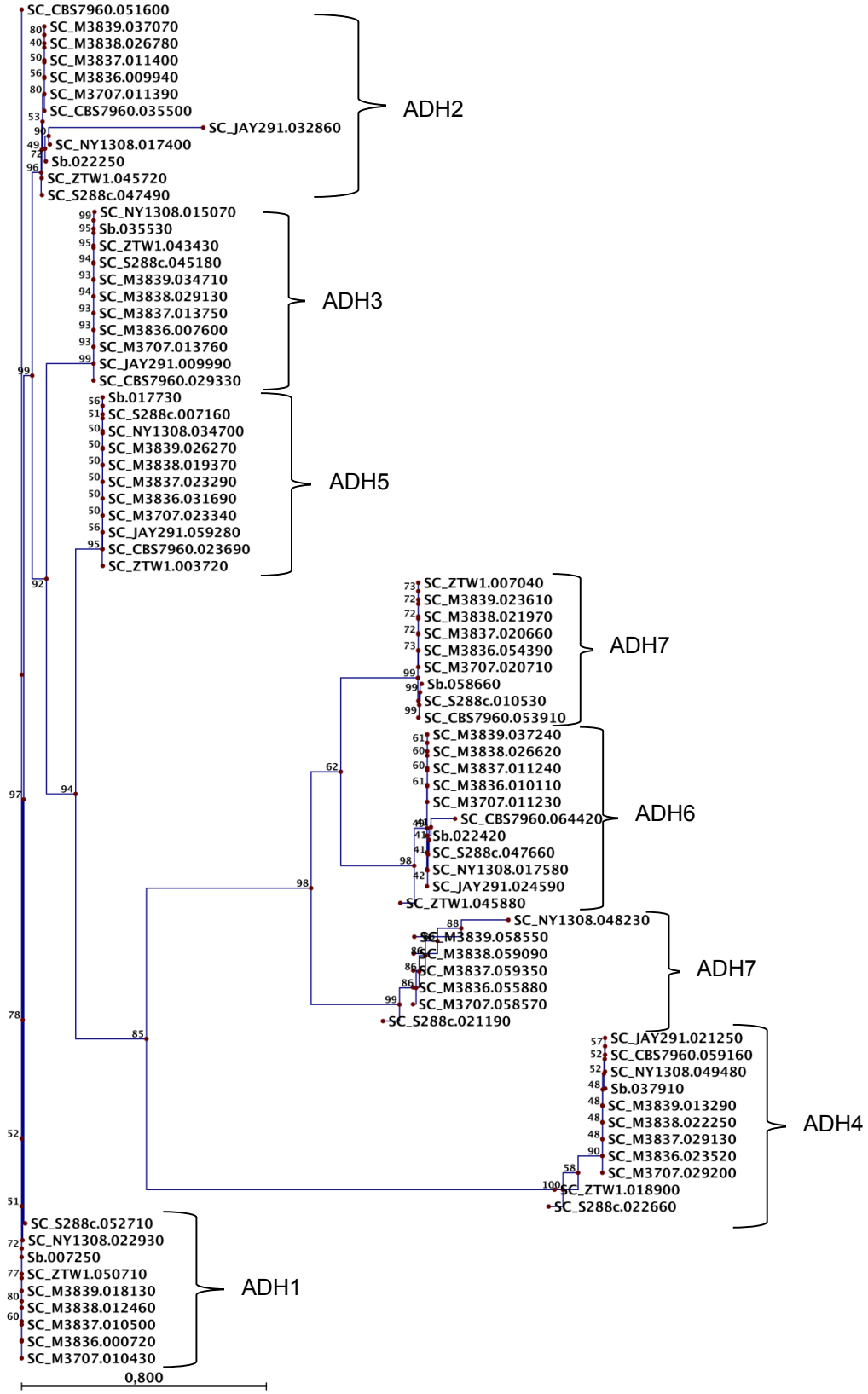


Figura 14: Árvore filogenética de sequências de proteínas ADHs relativa ao grupo 1 pela ferramenta CLC Workbench.

5.5.2. Comparação filogenética de ADHs do Grupo 2

Para analisar a relação filogenética entre os ADHs do grupo 2 foi realizado um único alinhamento múltiplo, contendo todas as 19 sequências de ADHs (Anexos 14 e 15). Para isso, foi feita a união das 17 sequências agrupadas nos 5 *clusters* (Orthomcl59, Orthomcl126, Orthomcl416, Orthomcl716 e Orthomcl4838) e das duas sequências adicionais preditas e não agrupadas (LT.022570, ADH2, e LT.023600, ADH7).

O *cluster* Orthomcl59 era composto por 5 sequências de proteínas, sendo duas de *K. lactis*, duas de *K. marxianus* e uma de *L. thermotolerans*. Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD3), que é membro da família de redutases/desidrogenases de cadeia média, cujo identificador é cd08297. Tal como CAD1, as proteínas CAD3 estão envolvidas na redução de cinamildeídos a cinamil álcoois no último passo no metabolismo de "monolignal" na parede das células. O banco de dados NR identificou estas proteínas com similaridade com ADH1 e ADH2. As sequências KL.046170 e KM.042970 apresentaram similaridade acima de 87% com ADH2. As sequências KL.047320, KM.041820 e LT.042560 apresentaram similaridade com ADH1 acima de 91%.

O *cluster* Orthomcl126 era composto por 4 sequências de proteínas, sendo uma sequência de *K. Lactis* (KL.019350), uma de *K. marxianus* (KM.032680) e duas de *L. thermotolerans* (LT.007790 e LT.027450). Assim como ocorreu com o grupo de ortólogos de ADH4, este grupo apresentou similaridade no CDD com um domínio LPO, oxidoredutase lactaldeído: propanodiol, que é um domínio típico de enterobactérias. Por outro lado, o banco de dados NR identificou as sequências com similaridade ADH4 acima de 81% com ADH4, com exceção da sequência KL.019350 que foi identificada como ADH somente.

O *cluster* Orthomcl716 era composto por 3 sequências de proteínas, uma de cada organismo. Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD3), que é membro da família de redutases/desidrogenases de cadeia média, cujo identificador é cd08297. O banco de dados NR identificou que duas sequências,

KL.044190 e KM.034210, possuíam similaridade com ADH4 acima de 91%. A outra sequência (LT.014940) possuía similaridade de 92% com ADH3.

O *cluster* Orthomcl4838 era composto por 2 sequências de proteínas, uma de *K. lactis* (KL.008680) e a outra de *K. marxianus* (KM.001900). Em comum, estas sequências possuíam um domínio das cinamil álcool desidrogenases (CAD3) e identificador cd08297. Ambas as sequências possuíam similaridade acima de 92% com ADH3 de acordo com o banco de dados NR.

O *cluster* Orthomcl416 era composto por 3 sequências de proteínas, uma de cada organismo. Este *cluster* apresentou um domínio conservado de cinamil álcool desidrogenases (CAD1), que é membro da família de redutases/desidrogenases de cadeia média, cujo identificador é cd05283. O banco de dados NR identificou estas proteínas (KL.036310, KM.002200 e LT.005410) com similaridade com ADH6 acima de 96% de similaridade.

Dessa forma, foram encontrados *clusters* e sequências relativos aos grupos: ADH1 (3 sequências de Orthomcl59), ADH2 (LT.022570 e 2 sequências de Orthomcl59), ADH3 (Orthomcl4838 e LT.014940), ADH4 (Orthomcl126 e duas sequências do Orthomcl716), ADH6 (Orthomcl416) e ADH7 (LT.023600). Dessa forma, não foram encontradas sequências relacionadas ao ADH5, um parálogo ao gene *ADH1*, envolvida na produção de etanol (Byrne & Wolfe, 2005; Smith et al., 2004). Diferentemente, do ADH5 do grupo 1 que apresenta esta proteína convervada em todas as linhagens. Outro achado interessante do grupo 2 é a ausência do ADH7, envolvido na tolerância a aldeído e na produção de álcool fúsel (Larroy et al., 2002), nas linhagens de *K. lactis* e *K. marxianus*. Posteriormente, deverá ser realizado um PCR para confirmar a ausência desses ADHs nas linhagens.

A Figura 15 representa a filogenia entre as sequências de ADHs do grupo 2. Diferentemente do grupo 1, a separação das sequências no grupo 2 não permitiu a exata classificação dos ADHs. Mais uma vez, a bioinformática direcionou os experimentos de biologia molecular a serem realizados, e estes se mostram essenciais para confirmação dos resultados.

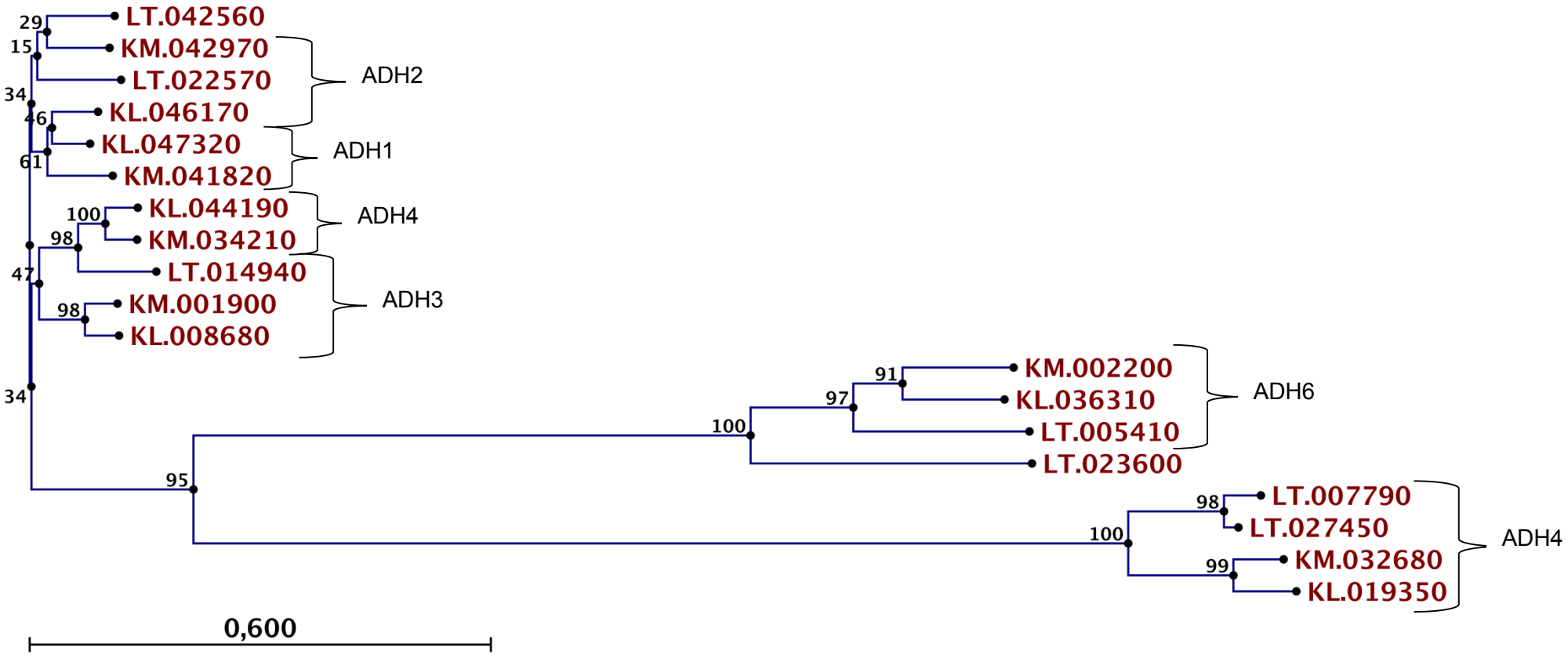


Figura 15: Árvore filogenética de sequências de proteínas ADHs relativa ao grupo 2 pela ferramenta CLC Workbench.

6. CONCLUSÕES

Nossas principais conclusões a partir da análise dos dados são:

- A seleção racional de espécies foi fundamental para o desenvolvimento do estudo de genômica comparativa, permitindo buscar assinaturas genômicas específicas dos organismos;
- Foi realizada a re-anotação de 14 genomas de leveduras de interesse biotecnológico;
- A re-anotação dos organismos foi essencial para a padronização das análises;
- A criação dos bancos de dados dos proteomas preditos mostrou-se uma ferramenta útil para análises e extremamente amigável;
- Não foi encontrada a sequência do gene ADH7 para a linhagem *S. cerevisiae* JAY 291, uma linhagem industrial extensamente utilizada no setor sucro-alcoeeiro brasileiro;
- Nos organismos pertencentes ao grupo 2, *Kluyveromyces* sp, não foram identificadas sequências relativas aos genes ADH5;
- Nas espécies do grupo 2, *K. lactis* e *K. marxianus*, não foram identificadas sequências do gene *ADH7*;
- Os experimentos de bioinformática poderão auxiliar o direcionamento dos experimentos de bancada, podendo validar por PCR os resultados obtidos através da bioinformática

7. REFERÊNCIAS BIBLIOGRÁFICAS

Altschul, S. F. et al. Basic local alignment search tool. **Journal of molecular biology,** v. 215, n. 3, p. 403-410, 1990. ISSN 0022-2836.

Anfang, N.; Brajkovich, M.; Goddard, M. R. Co-fermentation with *Pichia kluyveri* increases varietal thiol concentrations in Sauvignon Blanc. **Australian Journal of Grape and Wine Research,** v. 15, n. 1, p. 1-8, 2009. ISSN 1755-0238.

Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. **Science,** v. 297, n. 5585, p. 1301-1310, 2002. ISSN 0036-8075.

Argueso, J. L. et al. Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. **Genome research,** v. 19, n. 12, p. 2258-2270, 2009. ISSN 1088-9051.

Baker, M. De novo genome assembly: what every biologist should know. **Nature methods,** v. 9, n. 4, p. 333-337, 2012. ISSN 1548-7091.

Basso, L. C. et al. Yeast selection for fuel ethanol production in Brazil. **FEMS yeast research,** v. 8, n. 7, p. 1155-1163, 2008. ISSN 1567-1364.

Borneman, A. R. et al. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. **PLoS genetics,** v. 7, n. 2, p. e1001287, 2011. ISSN 1553-7404.

Borneman, A. R.; Pretorius, I. S.; Chambers, P. J. Comparative genomics: a revolutionary tool for wine yeast strain development. **Current opinion in biotechnology,** v. 24, n. 2, p. 192-199, 2013. ISSN 0958-1669.

Brenner, S. E. Errors in genome annotation. **Trends in Genetics,** v. 15, n. 4, p. 132-133, 1999. ISSN 0168-9525.

Brown, S. D. et al. Genome sequences of industrially relevant *Saccharomyces cerevisiae* strain M3707, isolated from a sample of distillers yeast and four haploid derivatives. **Genome announcements,** v. 1, n. 3, p.

e00323-13, 2013. ISSN 2169-8287.

Burge, C.; Karlin, S. Prediction of complete gene structures in human genomic DNA. **Journal of molecular biology,** v. 268, n. 1, p. 78-94, 1997. ISSN 0022-2836.

Carvalho, M. C. d. C. G.; Silva, D. C. G. d. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural,** v. 40, n. 3, p. 735-744, 2010. ISSN 0103-8478.

Carver, T. J. et al. ACT: the Artemis comparison tool. **Bioinformatics,** v. 21, n. 16, p. 3422-3423, 2005. ISSN 1367-4803.

Chen, Yi-Ching et al. Polymorphic internal transcribed spacer region 1 DNA sequences identify medically important yeasts. **Journal of clinical microbiology**, v. 39, n. 11, p. 4042-4051, 2001.

Cherry, J. M. et al. SGD: *Saccharomyces genome database*. **Nucleic acids research,** v. 26, n. 1, p. 73-79, 1998. ISSN 0305-1048.

Chial, H. DNA sequencing technologies key to the Human Genome Project. **Nature Education,** v. 1, n. 1, p. 219, 2008.

Chinwalla, A. T. et al. Initial sequencing and comparative analysis of the mouse genome. **Nature,** v. 420, n. 6915, p. 520-562, 2002. ISSN 0028-0836.

Costa, D. A. et al. Physiological characterization of thermotolerant yeast for cellulosic ethanol production. **Applied microbiology and biotechnology,** v. 98, n. 8, p. 3829-3840, 2014. ISSN 0175-7598.

De Smidt, O.; Du Preez, J. C.; Albertyn, J. The alcohol dehydrogenases of *Saccharomyces cerevisiae*: a comprehensive review. **FEMS yeast research,** v. 8, n. 7, p. 967-978, 2008. ISSN 1567-1364.

de Souza, C. J. A. et al. The influence of presaccharification, fermentation temperature and yeast strain on ethanol production from sugarcane bagasse. **Bioresource technology**, 2012. ISSN 0960-8524.

Devos, D.; Valencia, A. Intrinsic errors in genome annotation. **Trends in Genetics,** v. 17, n. 8, p. 429-431, 2001. ISSN 0168-9525.

Do, J. H.; Choi, D. Computational approaches to gene prediction. **Journal of microbiology ,** v. 44, n. 2, p. 137, 2006. ISSN 1225-8873.

Doolittle, R. F. The multiplicity of domains in proteins. **Annual review of biochemistry,** v. 64, n. 1, p. 287-314, 1995. ISSN 0066-4154.

Drewke, C.; Ciriacy, M. Overexpression, purification and properties of alcohol dehydrogenase IV from *Saccharomyces cerevisiae*. **Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression,** v. 950, n. 1, p. 54-60, 1988. ISSN 0167-4781.

Drewke, C.; Thielen, J.; Ciriacy, M. Ethanol formation in adh0 mutants reveals the existence of a novel acetaldehyde-reducing activity in *Saccharomyces cerevisiae*. **Journal of bacteriology,** v. 172, n. 7, p. 3909-3917, 1990. ISSN 0021-9193.

Dujon, B. Hemiascomycetous yeasts at the forefront of comparative genomics. **Current opinion in genetics & development,** v. 15, n. 6, p. 614-620, 2005. ISSN 0959-437X.

Dujon, B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. **Trends in Genetics,** v. 22, n. 7, p. 375-387, 2006. ISSN 0168-9525.

Dujon, B. et al. Genome evolution in yeasts. **Nature,** v. 430, n. 6995, p. 35-44, 2004. ISSN 0028-0836.

D'Urso, G.; Nurse, P. *Schizosaccharomyces pombe* cdc20+ encodes DNA polymerase ε and is required for chromosomal replication but not for the S phase checkpoint. **Proceedings of the National Academy of Sciences,** v. 94, n. 23, p. 12491-12496, 1997. ISSN 0027-8424.

Edwards, D. J.; Holt, K. E. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. **Microb Inform Exp,** v. 3, n. 1, p. 2, 2013.

Edwards-Ingram, L. et al. Genotypic and physiological characterization of Saccharomyces boulardii, the probiotic strain of *Saccharomyces cerevisiae*. **Applied and environmental microbiology,** v. 73, n. 8, p. 2458-2467, 2007. ISSN 0099-2240.

Edwards-Ingram, L. C. et al. Comparative genomic hybridization provides new insights into the molecular taxonomy of the *Saccharomyces sensu stricto* complex. **Genome research,** v. 14, n. 6, p. 1043-1051, 2004. ISSN 1088-9051.

Engel, S. R.; Cherry, J. M. The new modern era of yeast genomics: Community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the *Saccharomyces Genome Database.* **Database,** v. 2013, p. bat012, 2013. ISSN 1758-0463.

Feldmann, H. et al. Complete DNA sequence of yeast chromosome II. **The EMBO Journal,** v. 13, n. 24, p. 5795, 1994.

Fleischmann, R. D. et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. **Science,** v. 269, n. 5223, p. 496-512, 1995. ISSN 0036-8075.

França, L. T.; Carrilho, E.; Kist, T. B. A review of DNA sequencing techniques. **Quarterly reviews of biophysics,** v. 35, n. 02, p. 169-200, 2002. ISSN 1469-8994.

Ganzhorn, A. et al. Kinetic characterization of yeast alcohol dehydrogenases. Amino acid residue 294 and substrate specificity. **Journal of Biological Chemistry,** v. 262, n. 8, p. 3754-3761, 1987. ISSN 0021-9258.

Gibney, Patrick A. et al. Phylogenetic Portrait of the *Saccharomyces cerevisiae* Functional Genome. **G3: Genes| Genomes| Genetics**, v. 3, n. 8, p. 1335-1340, 2013.

Gilbert, D. Sequence File Format Conversion with Command-Line *Reads*eq. **Current Protocols in Bioinformatics**, p. A. 1E. 1-A. 1E. 4, 2003. ISSN 0471250953.

Glenn, T. C. Field guide to next-generation DNA sequencers. **Molecular ecology resources,** v. 11, n. 5, p. 759-769, 2011. ISSN 1755-0998.

Gobbi, M. et al. *Lachancea thermotolerans* and *Saccharomyces cerevisiae* in simultaneous and sequential co-fermentation: a strategy to enhance acidity and improve the overall quality of wine. **Food microbiology,** v. 33, n. 2, p. 271-281, 2013. ISSN 0740-0020.

Goffeau, A. et al. Life with 6000 genes. **Science,** v. 274, n. 5287, p. 546-567, 1996. ISSN 0036-8075.

González, E. et al. Characterization of a (2R, 3R)-2, 3-Butanediol Dehydrogenase as the *Saccharomyces cerevisiae* YAL060W Gene Product disruption and induction of the gene. **Journal of Biological Chemistry,** v. 275, n. 46, p. 35876-35885, 2000. ISSN 0021-9258.

Henikoff, S. et al. Gene families: the taxonomy of protein paralogs and chimeras. **Science,** v. 278, n. 5338, p. 609-614, 1997. ISSN 0036-8075.

Henson, J.; Tischler, G.; Ning, Z. Next-generation sequencing and large genome assemblies. **Pharmacogenomics,** v. 13, n. 8, p. 901-915, 2012. ISSN 1462-2416.

Hickey, D. A.; Singer, G. Genomic and proteomic adaptations to growth at high temperature. **Genome biology,** v. 5, p. /2004/5/10/117-/2004/5/10/117, 2004. ISSN 1465-6906.

Huh, W.-K. et al. Global analysis of protein localization in budding yeast. **Nature,** v. 425, n. 6959, p. 686-691, 2003. ISSN 0028-0836.

Jeong, H. et al. Genome sequence of the thermotolerant yeast *Kluyveromyces marxianus var. marxianus* KCTC 17555. **Eukaryotic cell,** v. 11, n. 12, p. 1584-1585, 2012. ISSN 1535-9778.

Jones, C. E.; Brown, A. L.; Baumann, U. Estimating the annotation error rate of curated GO database sequence annotations. **BMC bioinformatics,** v. 8, n. 1, p. 170, 2007. ISSN 1471-2105.

Jörnvall, H. The primary structure of yeast alcohol dehydrogenase.

European Journal of Biochemistry, v. 72, n. 3, p. 425-442, 1977. ISSN 1432-1033.

Kallberg, Y. et al. Short-chain dehydrogenases/reductases (SDRs). **European Journal of Biochemistry,** v. 269, n. 18, p. 4409-4417, 2002. ISSN 1432-1033.

Kellis, M.; Birren, B. W.; Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. **Nature,** v. 428, n. 6983, p. 617-624, 2004. ISSN 0028-0836.

Khatri, I. et al. Gleaning evolutionary insights from the genome sequence of a probiotic yeast *Saccharomyces boulardii*. **Gut Pathog,** v. 5, n. 1, p. 30, 2013.

Kurtzman, C. P. Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma* and *Zygotorulaspora*. **FEMS yeast research,** v. 4, n. 3, p. 233-245, 2003. ISSN 1567-1364.

Kurtzman, C. P.; Robnett, C. J. Phylogenetic relationships among yeasts of the '*Saccharomyces complex*' determined from multigene sequence analyses. **FEMS yeast research,** v. 3, n. 4, p. 417-432, 2003. ISSN 1567-1364.

Ladrière, J.-M. et al. *Kluyveromyces marxianus* exhibits an ancestral Saccharomyces cerevisiae genome organization downstream of ADH2. **Gene,** v. 255, n. 1, p. 83-91, 2000. ISSN 0378-1119.

Lane, M. M.; Morrissey, J. P. *Kluyveromyces marxianus*: a yeast emerging from its sister's shadow. **Fungal Biology Reviews,** v. 24, n. 1, p. 17-26, 2010. ISSN 1749-4613.

Larroy, C. et al. Properties and functional significance of *Saccharomyces cerevisiae* ADHVI. **Chemico-biological interactions,** v. 143, p. 229-238, 2003. ISSN 0009-2797.

Larroy, C. et al. Characterization of the *Saccharomyces cerevisiae* YMR318C (ADH6) gene product as a broad specificity NADPH-dependent alcohol dehydrogenase: relevance in aldehyde reduction. **Biochem. J,** v. 361, p. 163-172, 2002a.

Larroy, C.; Pares, X.; Biosca, J. A. Characterization of a *Saccharomyces cerevisiae* NADP (H)-dependent alcohol dehydrogenase (ADHVII), a member of the cinnamyl alcohol dehydrogenase family. **European Journal of Biochemistry,** v. 269, n. 22, p. 5738-5745, 2002b. ISSN 1432-1033.

Liu, L. et al. Comparison of next-generation sequencing systems. **BioMed Research International,** v. 2012, 2012. ISSN 1110-7243.

Lopes, M. R. et al. Production and Characterization of β-Glucanase Secreted by the Yeast *Kluyveromyces marxianus*. **Applied biochemistry and biotechnology**, v. 172, n. 5, p. 2412-2424, 2014. ISSN 0273-2289.

Luo, Ruibang et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. **Gigascience**, v. 1, n. 1, p. 18, 2012.

Lynch, M.; Conery, J. S. The evolutionary fate and consequences of duplicate genes. **Science,** v. 290, n. 5494, p. 1151-1155, 2000. ISSN 0036-8075.

Martins, D. B. G. et al. The β-galactosidase activity in *Kluyveromyces marxianus* CBS6556 decreases by high concentrations of galactose. **Current microbiology,** v. 44, n. 5, p. 379-382, 2002. ISSN 0343-8651.

Mohamed, S.; Syed, B. A. Commercial prospects for genomic sequencing technologies. **Nature reviews Drug discovery,** v. 12, n. 5, p. 341-342, 2013. ISSN 1474-1776.

Mooney, D.; Pilgrim, D.; Young, E. Mutant alcohol dehydrogenase (ADH III) presequences that affect both in vitro mitochondrial import and in vitro processing by the matrix protease. **Molecular and cellular biology,** v. 10, n. 6, p. 2801-2808, 1990. ISSN 0270-7306.

Mortimer, R. K.; Johnston, J. R. Genealogy of principal strains of the yeast

genetic stock center. **Genetics,** v. 113, n. 1, p. 35-43, 1986. ISSN 0016-6731.

Negelein, E.; Wulff, H. Kristallisation des proteins der acetaldehydreduktase. **Biochem. Z,** v. 289, p. 436-437, 1937.

O'Donovan, C. et al. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. **Briefings in bioinformatics,** v. 3, n. 3, p. 275-284, 2002. ISSN 1467-5463.

Pagani, I. et al. The Genomes OnLine Database (GOLD) v. 4: *Status* of genomic and metagenomic projects and their associated metadata. **Nucleic acids research,** v. 40, n. D1, p. D571-D579, 2012. ISSN 0305-1048.

Paquin, C. E.; Williamson, V. M. Ty insertions at two loci account for most of the spontaneous antimycin A resistance mutations during growth at 15 degrees C of *Saccharomyces cerevisiae* strains lacking ADH1. **Molecular and cellular biology,** v. 6, n. 1, p. 70-79, 1986. ISSN 0270-7306.

Payen, C. et al. Unusual composition of a yeast chromosome arm is associated with its delayed replication. **Genome research,** v. 19, n. 10, p. 1710-1721, 2009. ISSN 1088-9051.

Quail, M. A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. **BMC genomics,** v. 13, n. 1, p. 341, 2012. ISSN 1471-2164.

Raj, S. B.; Ramaswamy, S.; Plapp, B. V. Yeast alcohol dehydrogenase structure and catalysis. **Biochemistry,** v. 53, n. 36, p. 5791-5803, 2014. ISSN 0006-2960.

Robison, K.; Gilbert, W.; Church, G. M. Large scale bacterial gene discovery by similarity search. **Nature genetics,** v. 7, n. 2, p. 205-214, 1994.

Rouwenhorst, R. J. et al. Production, distribution, and kinetic properties of inulinase in continuous cultures of *Kluyveromyces marxianus* CBS 6556. **Applied and environmental microbiology,** v. 54, n. 5, p. 1131-1137, 1988. ISSN 0099-2240.

Sanger, F.; Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of molecular biology,** v. 94, n. 3, p. 441-448, 1975. ISSN 0022-2836.

Sanger, F.; Nicklen, S.; Coulson, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences,** v. 74, n. 12, p. 5463-5467, 1977. ISSN 0027-8424.

Seo, Seung Bum et al. Underlying Data for Sequencing the Mitochondrial Genome with the Massively Parallel Sequencing Platform Ion Torrent™ PGM™. **BMC genomics**, v. 16, n. 1, p. 1-10, 2015.

Simpson, Jared T. et al. ABySS: a parallel assembler for short read sequence data**. Genome research**, v. 19, n. 6, p. 1117-1123, 2009.

Siso, M. G. The biotechnological utilization of cheese whey: a review. **Bioresource Technology,** v. 57, n. 1, p. 1-11, 1996. ISSN 0960-8524.

Souciet, J.-L. et al. Comparative genomics of protoploid Saccharomycetaceae. **Genome research,** v. 19, n. 10, p. 1696-1709, 2009. ISSN 1088-9051.

Stanke, M. et al. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. **BMC bioinformatics,** v. 7, n. 1, p. 62, 2006. ISSN 1471-2105.

Stein, L. Genome annotation: from sequence to biology. **Nature reviews genetics,** v. 2, n. 7, p. 493-503, 2001. ISSN 1471-0056.

Steven, L.; Salzberg, J. Beware of mis—assembled genomes. **Bioinformatics,** v. 21, n. 4, p. 320-4, 2005.

Sánchez, Ó. J.; Cardona, C. A. Trends in biotechnological production of fuel ethanol from different feedstocks. **Bioresource technology,** v. 99, n. 13, p. 5270-5295, 2008. ISSN 0960-8524.

Thudi, M. et al. Current state-of-art of sequencing technologies for plant genomics research. **Briefings in Functional Genomics,** v. 11, n. 1, p. 3-11, 2012. ISSN 2041-2649.

Tkach, J. M. et al. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. **Nature cell biology,** v. 14, n. 9, p. 966-976, 2012. ISSN 1465-7392.

Valencia, E. et al. Apo and Holo structures of an NADP (H)-dependent cinnamyl alcohol dehydrogenase from *Saccharomyces cerevisiae*. **Journal of molecular biology,** v. 341, n. 4, p. 1049-1062, 2004. ISSN 0022-2836.

Van Loon, A.; Young, E. T. Intracellular sorting of alcohol dehydrogenase isoenzymes in yeast: a cytosolic location reflects absence of an amino-terminal targeting sequence for the mitochondrion. **The EMBO journal,** v. 5, n. 1, p. 161, 1986.

Warren, René L. et al. Assembling millions of short DNA sequences using SSAKE. **Bioinformatics**, v. 23, n. 4, p. 500-501, 2007.

Wei, W. et al. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. **Proceedings of the National Academy of Sciences,** v. 104, n. 31, p. 12825-12830, 2007. ISSN 0027-8424.

Xia, X. et al. Genomic changes in nucleotide and dinucleotide frequencies in Pasteurella multocida cultured under high temperature. **Genetics,** v. 161, n. 4, p. 1385-1394, 2002. ISSN 0016-6731.

Yandell, M.; Ence, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics,** v. 13, n. 5, p. 329-342, 2012. ISSN 1471-0056.

Young, E. T.; Pilgrim, D. Isolation and DNA sequence of ADH3, a nuclear gene encoding the mitochondrial isozyme of alcohol dehydrogenase in *Saccharomyces cerevisiae*. **Molecular and cellular biology,** v. 5, n. 11, p. 3024-3034, 1985. ISSN 0270-7306.

ZERBINO, Daniel R.; BIRNEY, Ewan. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome research**, v. 18, n. 5, p. 821-829, 2008.

Zhang, J. Evolution by gene duplication: an update. **Trends in ecology & evolution,** v. 18, n. 6, p. 292-298, 2003. ISSN 0169-5347.

Zhang, J. et al. The impact of next-generation sequencing on genomics. **Journal of genetics and genomics,** v. 38, n. 3, p. 95-109, 2011. ISSN 1673-8527.

Zhang, K. et al. Genomic reconstruction to improve bioethanol and ergosterol production of industrial yeast *Saccharomyces cerevisiae*. **Journal of industrial microbiology & biotechnology,** v. 42, n. 2, p. 207-218, 2015. ISSN 1367-5435.

Zhao, X.; Bai, F. Yeast flocculation: New story in fuel ethanol production. **Biotechnology advances,** v. 27, n. 6, p. 849-856, 2009. ISSN 0734-9750.

Zheng, D. et al. Construction of novel *Saccharomyces cerevisiae* strains for bioethanol active dry yeast (ADY) production. 2013. ISSN 1932-6203.

Zheng, D.-Q. et al. Genomic structural variations contribute to trait improvement during whole-genome shuffling of yeast. **Applied microbiology and biotechnology,** v. 98, n. 7, p. 3059-3070, 2014. ISSN 0175-7598.

Zheng, H.; Wu, H. Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species. **BMC bioinformatics,** v. 11, n. Suppl 11, p. S7, 2010. ISSN 1471-2105.

Zhu, Y.; Lin, Z.; Nakhleh, L. Evolution after whole-genome duplication: a network perspective. **G3: Genes| Genomes| Genetics,** v. 3, n. 11, p. 2049-2057, 2013. ISSN 2160-1836.

Zivanovic, Y. et al. Complete nucleotide sequence of the mitochondrial DNA from *Kluyveromyces lactis*. **FEMS yeast research,** v. 5, n. 4-5, p. 315-322, 2005. ISSN 1567-1364.

8. ANEXOS

Anexo 1: Matriz de identificação de pares de ortólogos entre as linhagens do grupo 1. O lado superior direito representa os pares de ortólogos por meio de sequências que possuem o melhor *hit* entre si. O lado inferior esquerdo representa os pares de ortólgos que foram identificados como Melhor *Hit*/Melhor Recíproco.

Organismos	CBS 7960	JAY 291	M3707	M3836	M3837	M3838	M3839	NY1308	S288c	ZTW1	EDRL
CBS 7960	-	12.604	12.512	12.505	12.504	12.524	12.539	12.468	12.736	12.442	12.597
JAY 291	6.128	-	11.924	11.924	11.919	11.932	11.943	11.883	12.124	11.891	11.947
M3707	6.185	6.196	-	12.162	12.151	12.176	12.147	11.886	12.013	11.914	11.910
M3836	6.183	6.188	6.538	-	12.159	12.170	12.150	11.888	12.012	11.921	11.907
M3837	6.186	6.200	6.557	6.526	-	12.170	12.148	11.881	12.022	11.907	11.914
M3838	6.219	6.206	6.568	6.592	6.576	-	12.169	11.897	12.031	11.920	11.924
M3839	6.217	6.210	6.552	6.550	6.556	6.581	-	11.916	12.054	11.933	11.943
NY1308	6.154	6.187	6.316	6.309	6.306	6.341	6.334	-	11.981	11.843	11.856
S288c	6.763	6.331	6.546	6.606	6.618	6.725	6.621	6.423	-	11.972	12.072
ZTW1	6.167	6.195	6.372	6.348	6.354	6.390	6.371	6.313	6.452	-	11.858
EDRL	6.239	6.217	6.309	6.284	6.297	6.327	6.318	6.283	6.514	6.304	-

Anexo 2: No lado superior direito tem-se a matriz de identificação de pares de co-ortólogos entre as linhagens do grupo 1. No lado inferior esquerdo tem-se a matriz de identificação de peso médio (average weight) entre as linhagens do mesmo grupo. Em negrito, encontram-se os pesos médios de cada linhagem comparada a si própria.

Organismos	CBS 7960	JAY 291	M3707	M3836	M3837	M3838	M3839	NY1308	S288c	ZTW1	EDRL
CBS 7960	**244,75**	142	163	162	167	173	158	156	249	153	158
JAY 291	217,09	**295,69**	108	107	111	126	106	102	254	127	109
M3707	219,47	228,87	**296,62**	92	104	129	97	110	170	154	120
M3836	219,49	228,80	225,73	**297,76**	101	109	85	103	165	139	120
M3837	219,69	228,54	225,32	225,68	**290,87**	121	105	106	214	153	130
M3838	219,71	228,91	225,67	225,91	225,80	**299,85**	114	109	226	158	136
M3839	219,43	228,39	225,64	225,70	225,41	225,80	**295,05**	97	184	142	117
NY1308	217,96	227,81	228,97	228,77	228,92	229,18	228,50	**281,70**	159	124	101
S288c	222,85	227,20	232,83	233,31	233,54	234,67	233,09	230,27	**299,71**	184	230
ZTW1	218,00	227,33	227,52	227,67	227,55	228,10	227,56	227,14	230,29	**296,16**	135
EDRL	217,31	227,41	228,58	228,51	228,31	228,76	228,11	228,11	229,14	227,66	**292,17**

Anexo 3: Matriz de identificação de pares de ortólogos entre as linhagens do grupo 2. O lado superior direito representa os pares de ortólogos por meio de sequências que possuem o melhor *hit* entre si. O lado inferior esquerdo representa os pares de ortólgos que foram identificados como Melhor *Hit*/Melhor Recíproco.

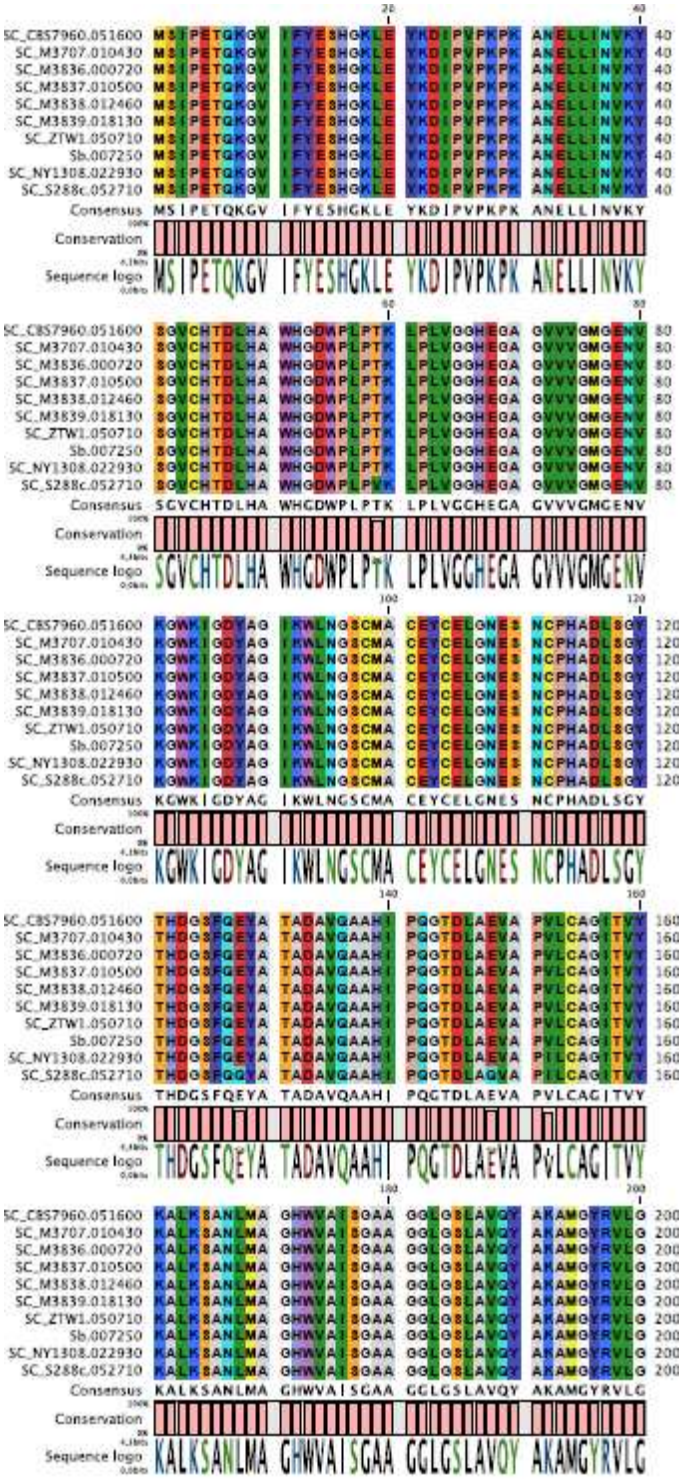
Organismos	NRRL Y-1140	KCTC 17555	CBS 6340
NRRL Y-1140	-	9.910	9.799
KCTC 17555	4.971	-	9.791
CBS 6340	4.747	4.742	-

Anexo 4: No lado superior direito tem-se a matriz de identificação de pares de co-ortólogos entre as linhagens do grupo 2. No lado inferior esquerdo tem-se a matriz de identificação de peso médio (average weight) entre as linhagens do mesmo grupo. Em negrito, encontram-se os pesos médios de cada linhagem comparada a si própria.

Organismos	NRRL Y-1140	KCTC 17555	CBS 6340
NRRL Y-1140	**204,08**	61	112
KCTC 17555	195,19	**275,32**	121
CBS 6340	137,88	138,94	**187,56**

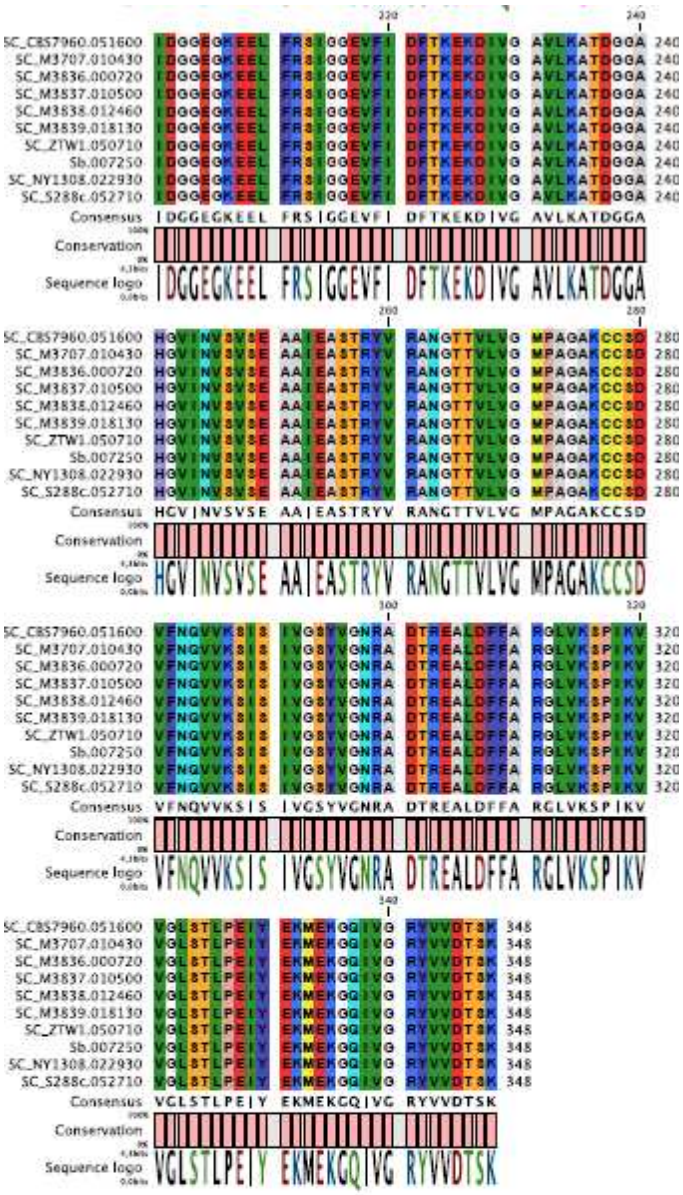
Anexo 5: Alinhamento múltiplo, entre as sequências de proteínas ADH1 do grupo 1 agrupadas no *cluster* Orthomcl5422, evidenciado pelo programa CLC Workbench.

(continua)



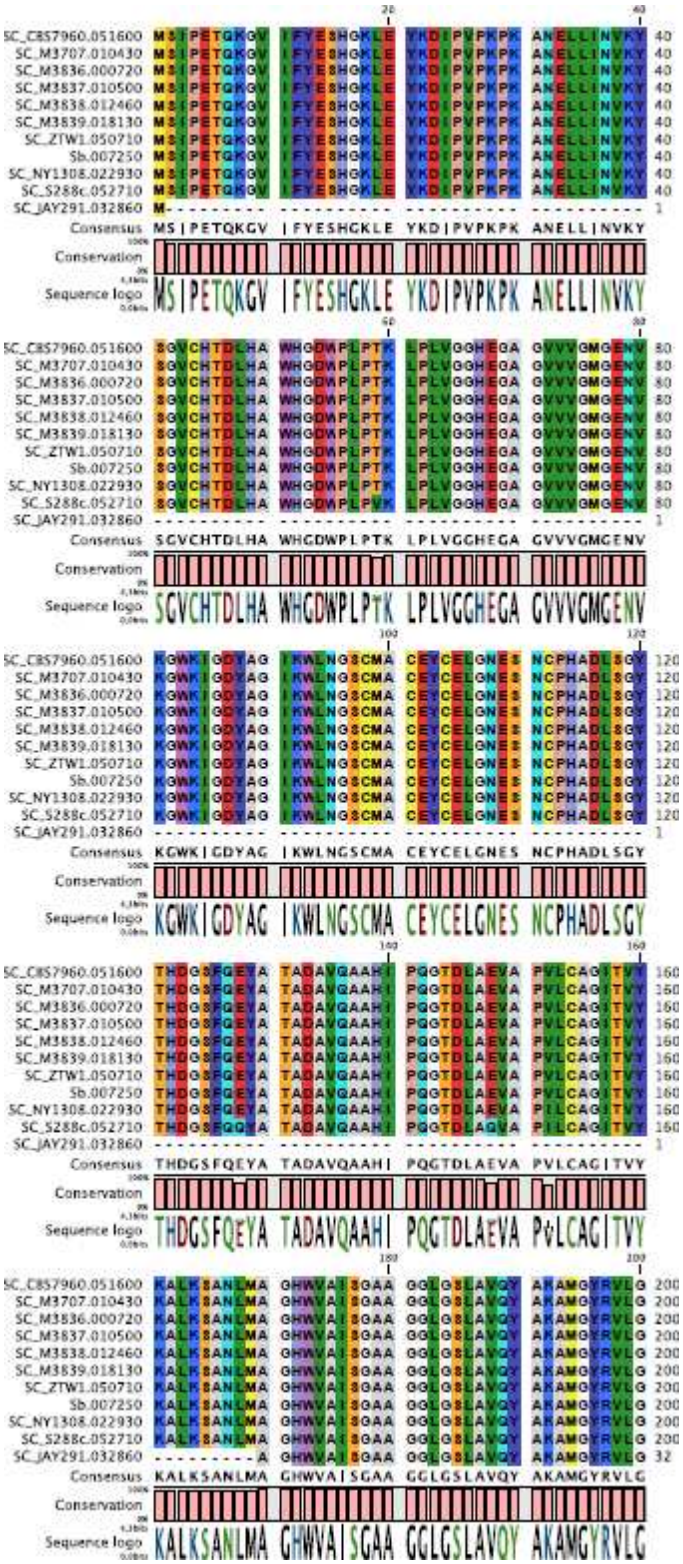
Anexo 5: Alinhamento múltiplo, entre as sequências de proteínas ADH1 do grupo 1 agrupadas no *cluster* Orthomcl5422, evidenciado pelo programa CLC Workbench.

(conclusão)



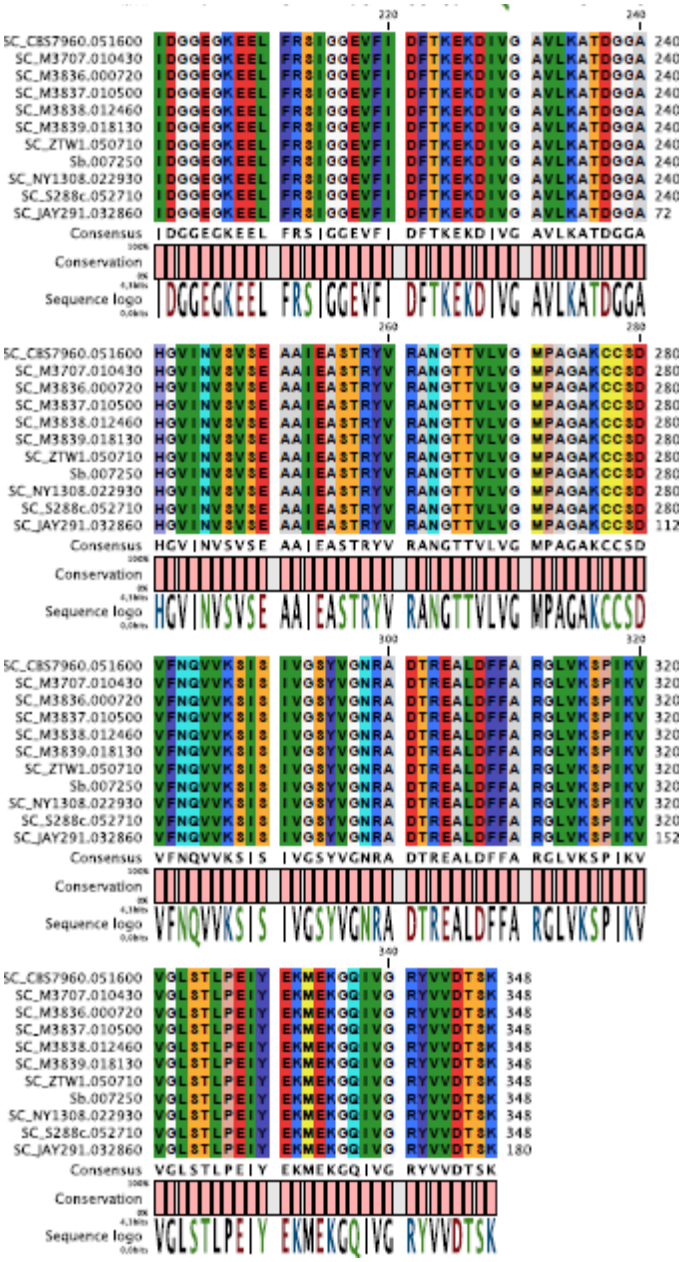
Anexo 6: Alinhamento múltiplo, entre todas as sequências de proteínas ADH1, as sequências do *cluster* Orthomcl5422 e a sequência predita SC_JAY.032860, evidenciado pelo programa CLC Workbench.

(continua)



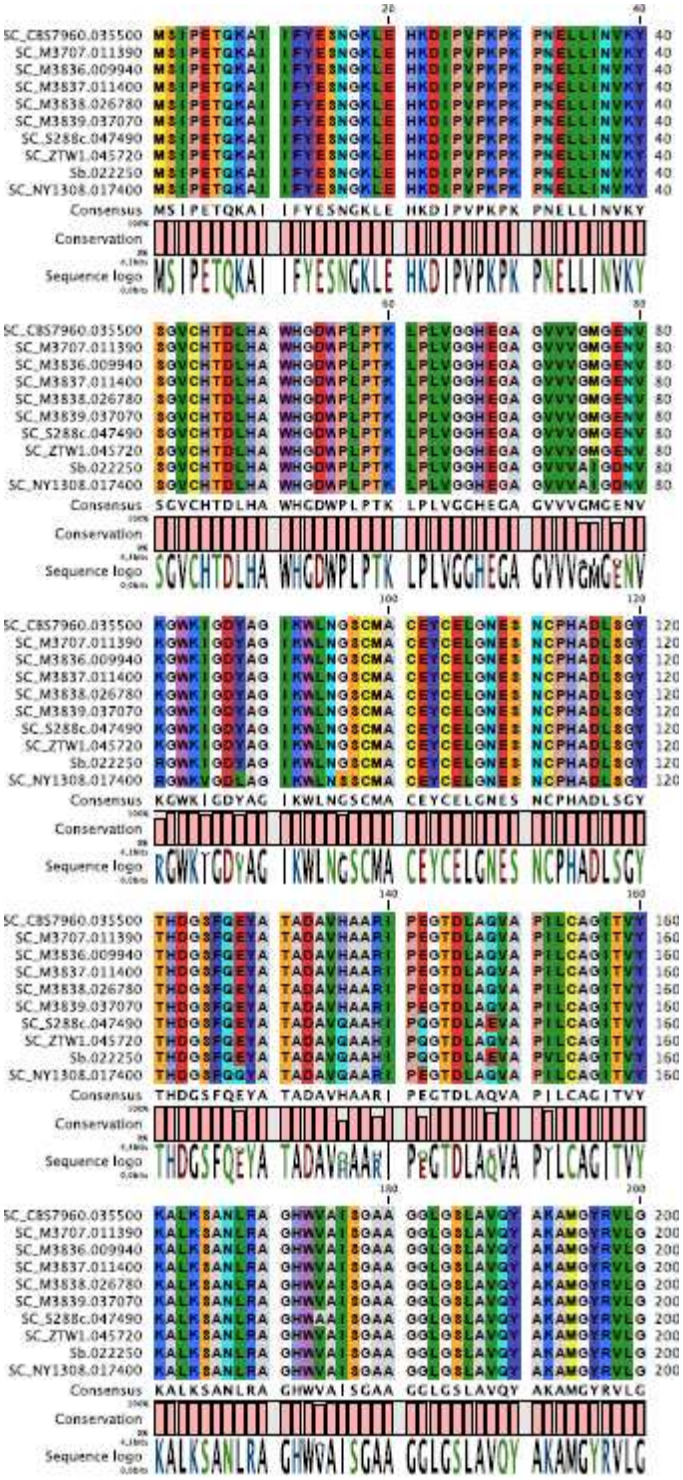
Anexo 6: Alinhamento múltiplo, entre todas as sequências de proteínas ADH1, as sequências do *cluster* Orthomcl5422 e a sequência predita SC_JAY.032860, evidenciado pelo programa CLC Workbench.

(conclusão)



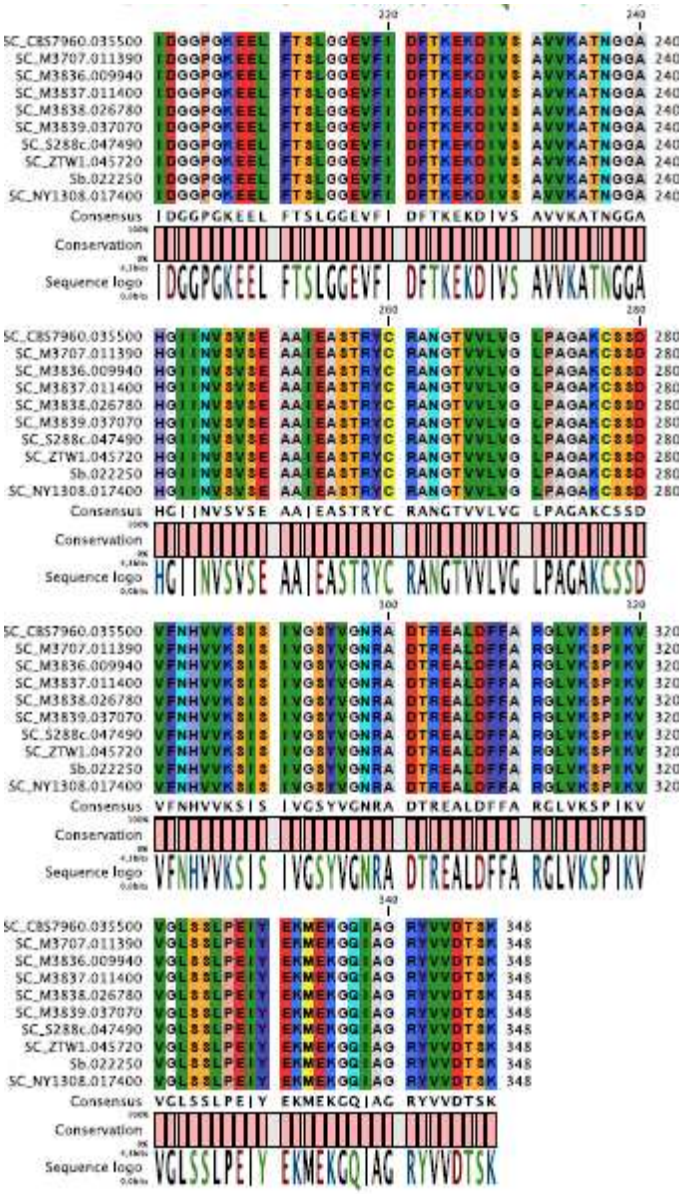
Anexo 7: Alinhamento múltiplo, entre as sequências de proteínas ADH2 do grupo 1 agrupadas no *cluster* Orthomcl5451, evidenciado pelo programa CLC Workbench.

(continua)



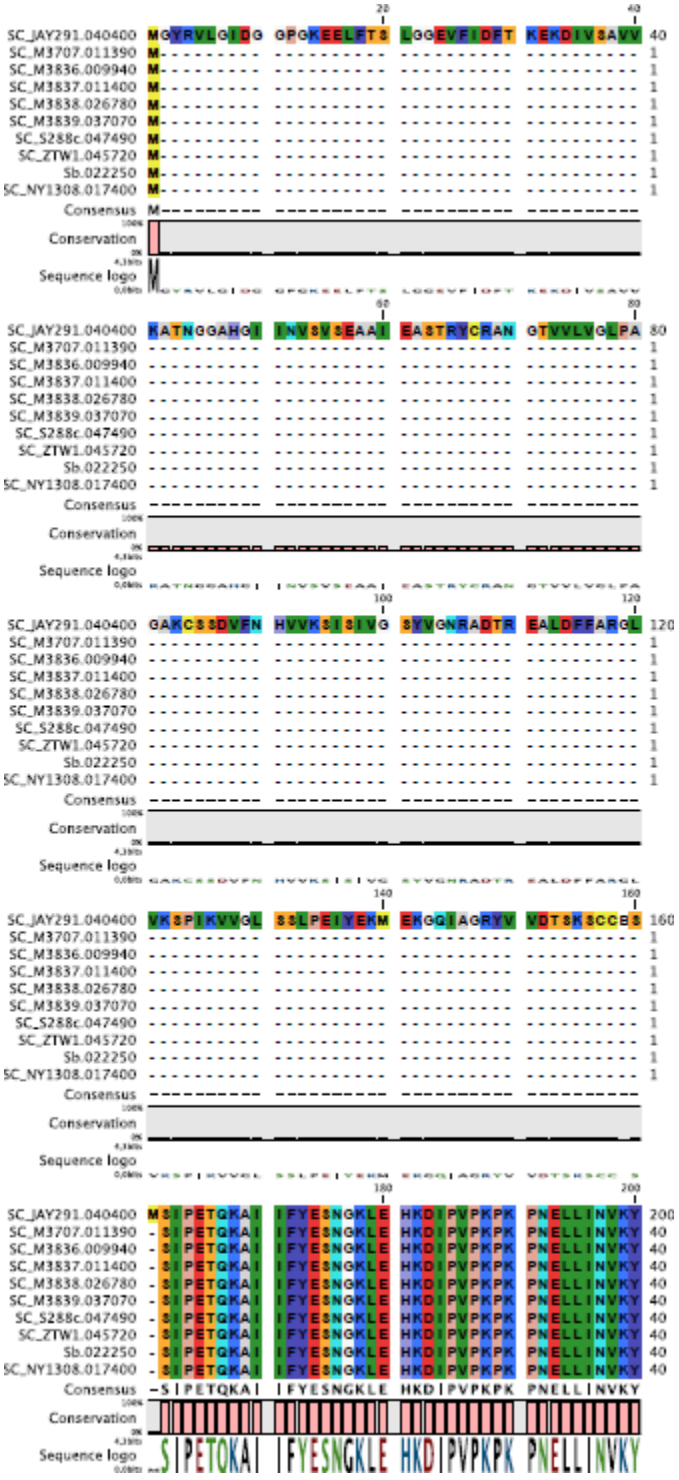
Anexo 7: Alinhamento múltiplo, entre as sequências de proteínas ADH2 do grupo 1 agrupadas no *cluster* Orthomcl5451, evidenciado pelo programa CLC Workbench.

(conclusão)



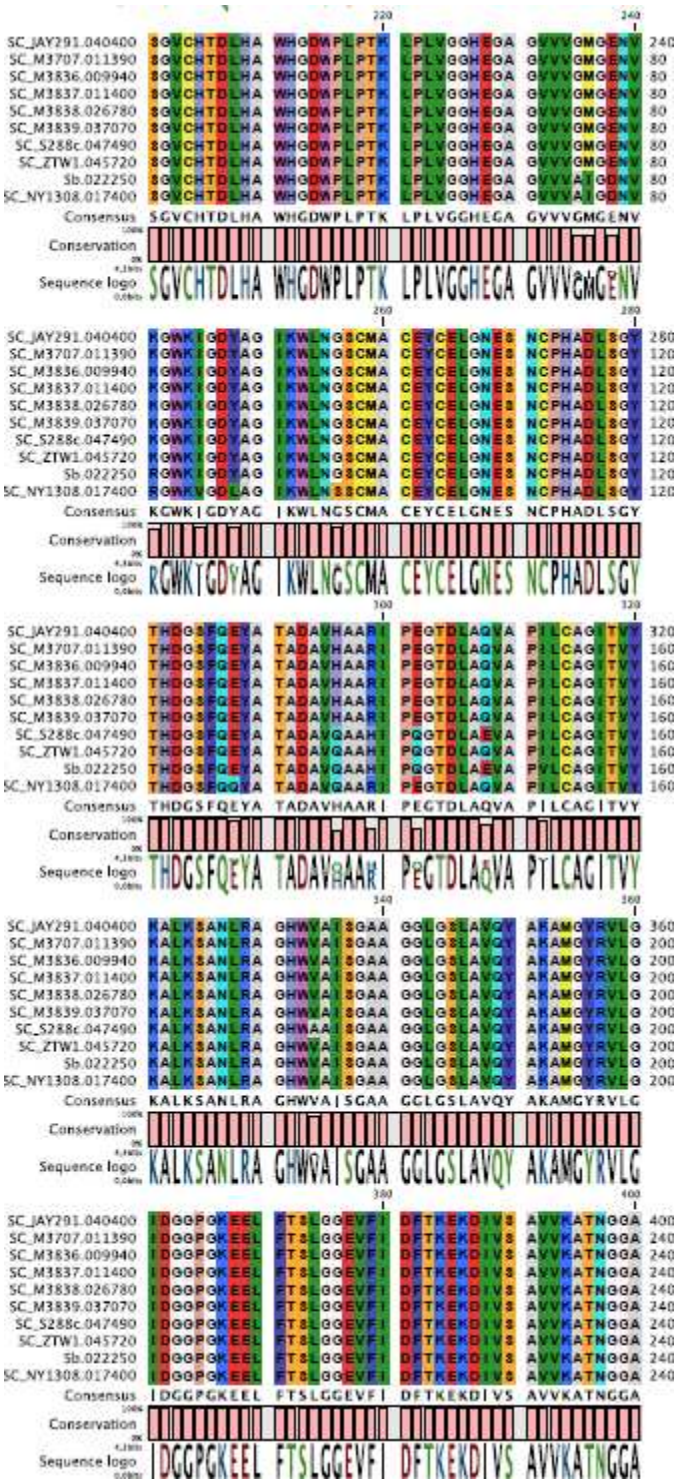
Anexo 8: Alinhamento múltiplo, entre todas as sequências de proteínas ADH2, as sequências do *cluster* Orthomcl5451 e a sequência predita SC_JAY.040400, evidenciado pelo programa CLC Workbench.

(continua)



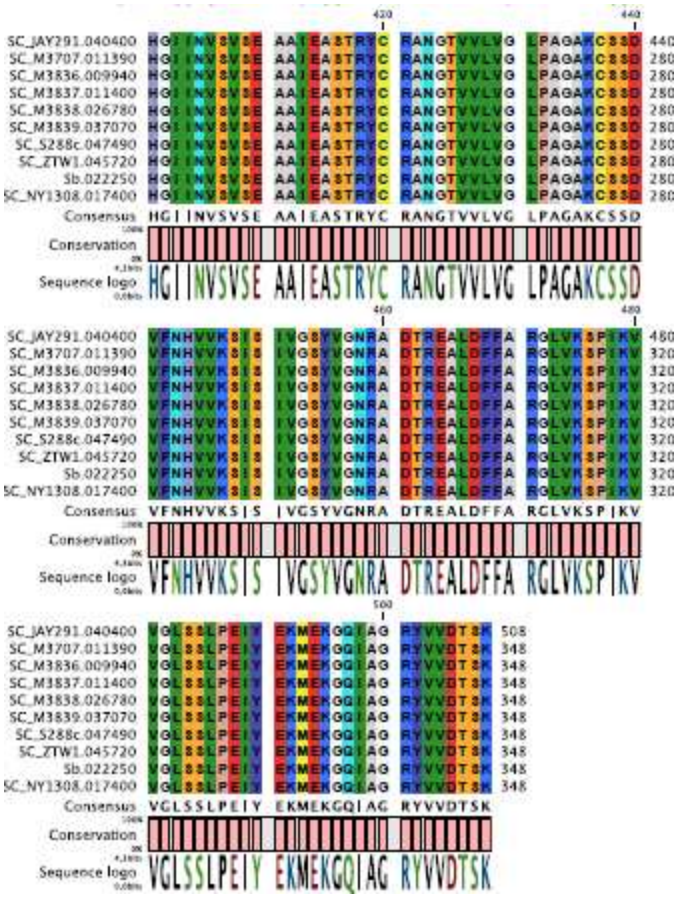
Anexo 8: Alinhamento múltiplo, entre todas as sequências de proteínas ADH2, as sequências do *cluster* Orthomcl5451 e a sequência predita SC_JAY.040400, evidenciado pelo programa CLC Workbench.

(continuação)



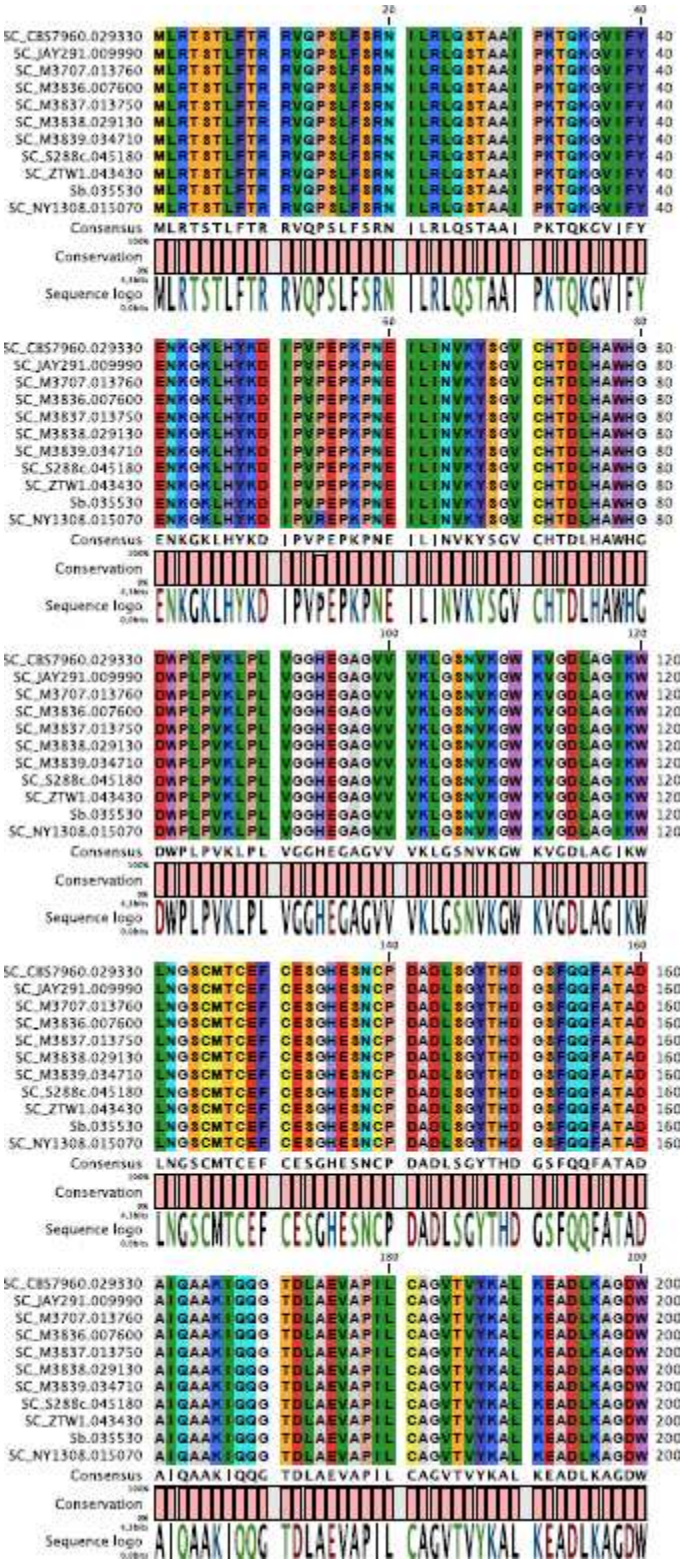
Anexo 8: Alinhamento múltiplo, entre todas as sequências de proteínas ADH2, as sequências do *cluster* Orthomcl5451 e a sequência predita SC_JAY.040400, evidenciado pelo programa CLC Workbench.

(conclusão)



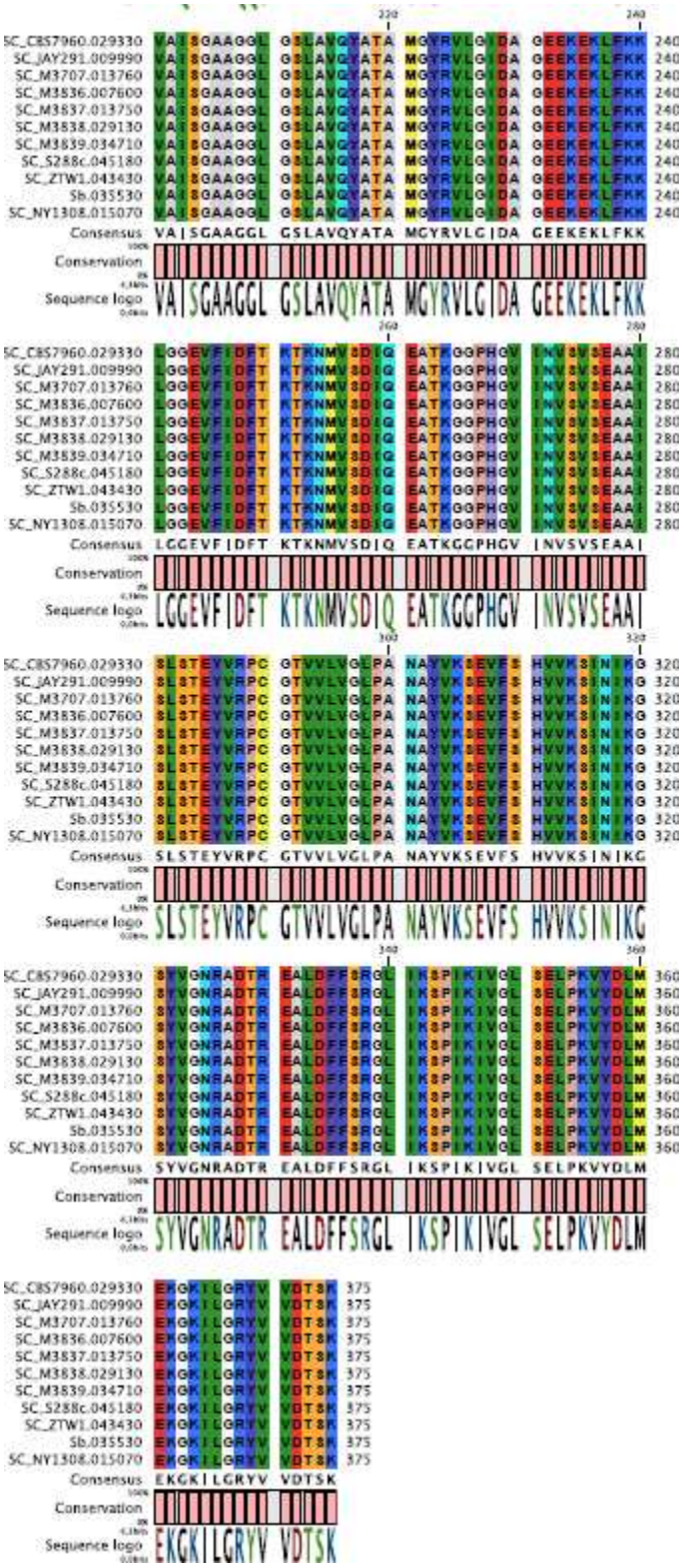
Anexo 9: Alinhamento múltiplo, entre as sequências de proteínas ADH3 do grupo 1 agrupadas no *cluster* Orthomcl3080, evidenciado pelo programa CLC Workbench.

(continua)



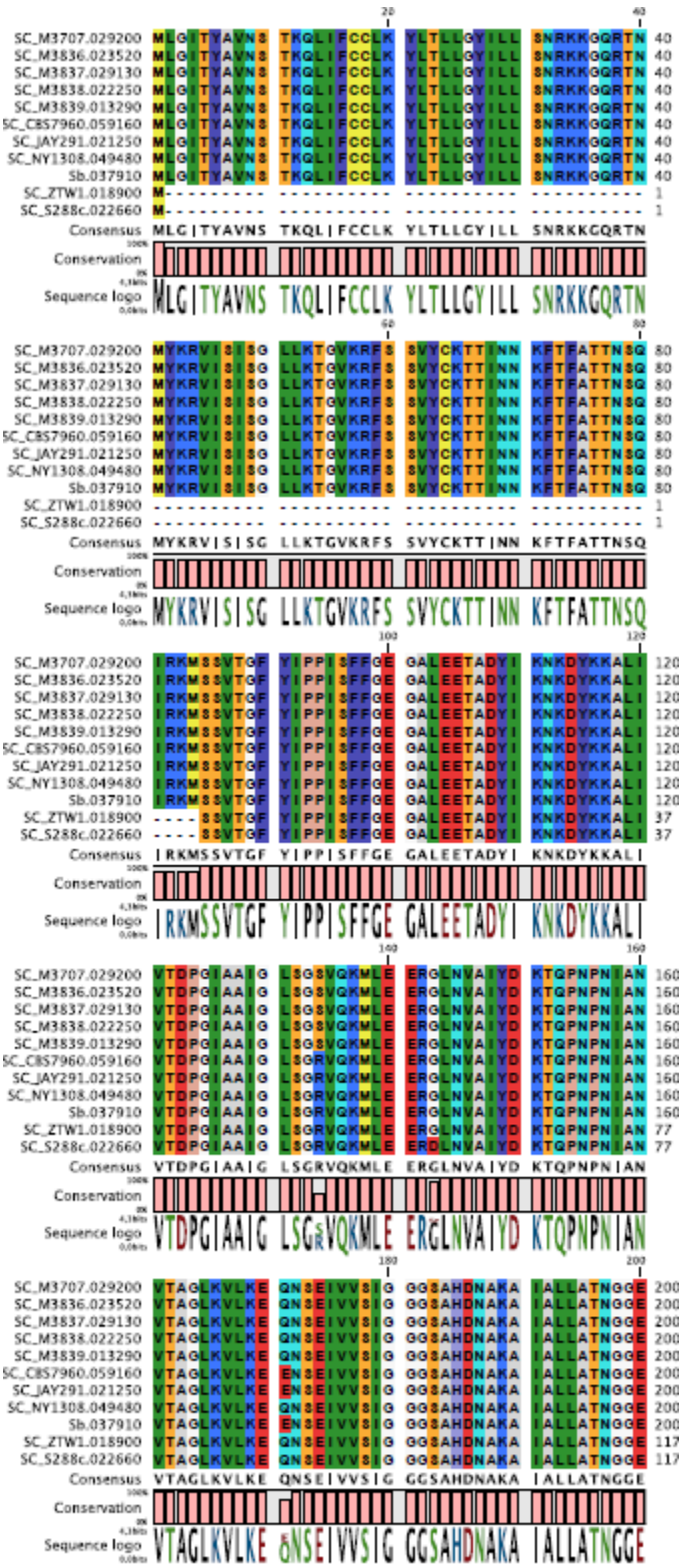
Anexo 9: Alinhamento múltiplo, entre as sequências de proteínas ADH3 do grupo 1 agrupadas no *cluster* Orthomcl3080, evidenciado pelo programa CLC Workbench.

(conclusão)



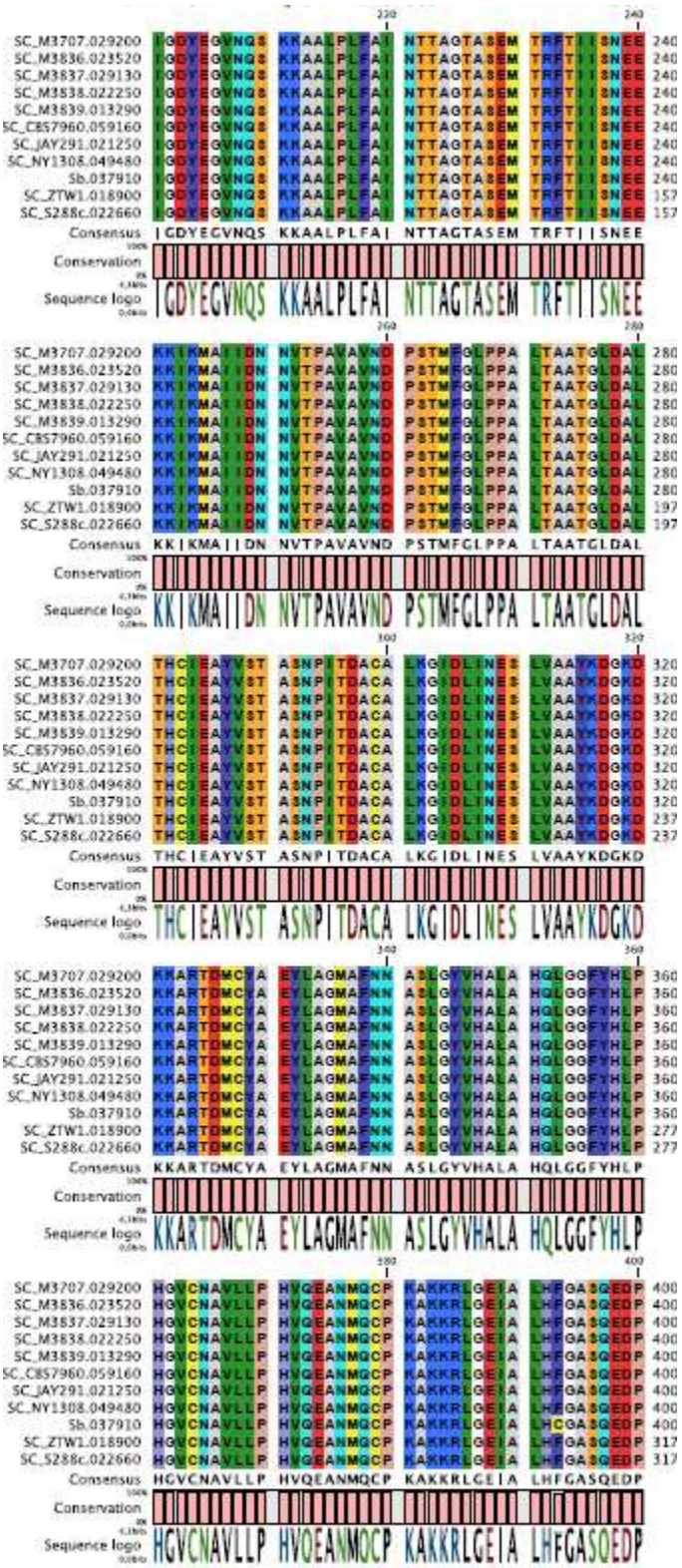
Anexo 10: Alinhamento múltiplo, entre as sequências de proteínas ADH4 do grupo 1 agrupadas no *cluster* Orthomcl902, evidenciado pelo programa CLC Workbench.

(continua)



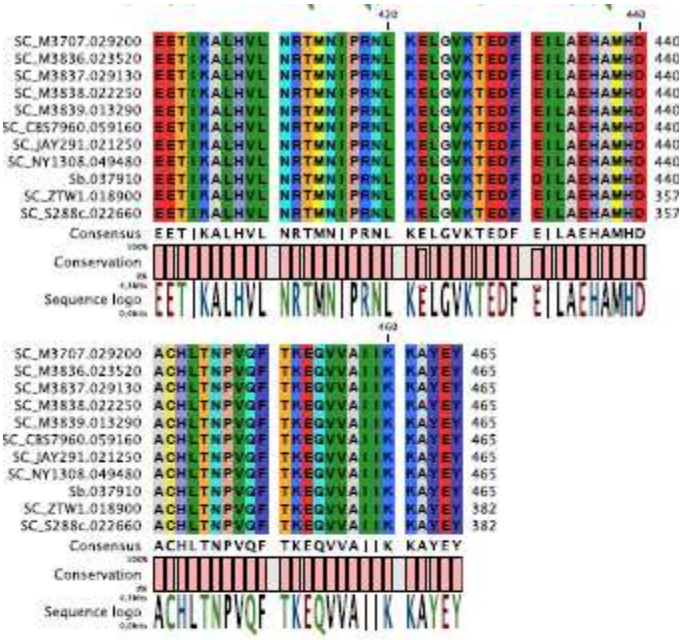
Anexo 10: Alinhamento múltiplo, entre as sequências de proteínas ADH4 do grupo 1 agrupadas no *cluster* Orthomcl902, evidenciado pelo programa CLC Workbench.

(continuação)



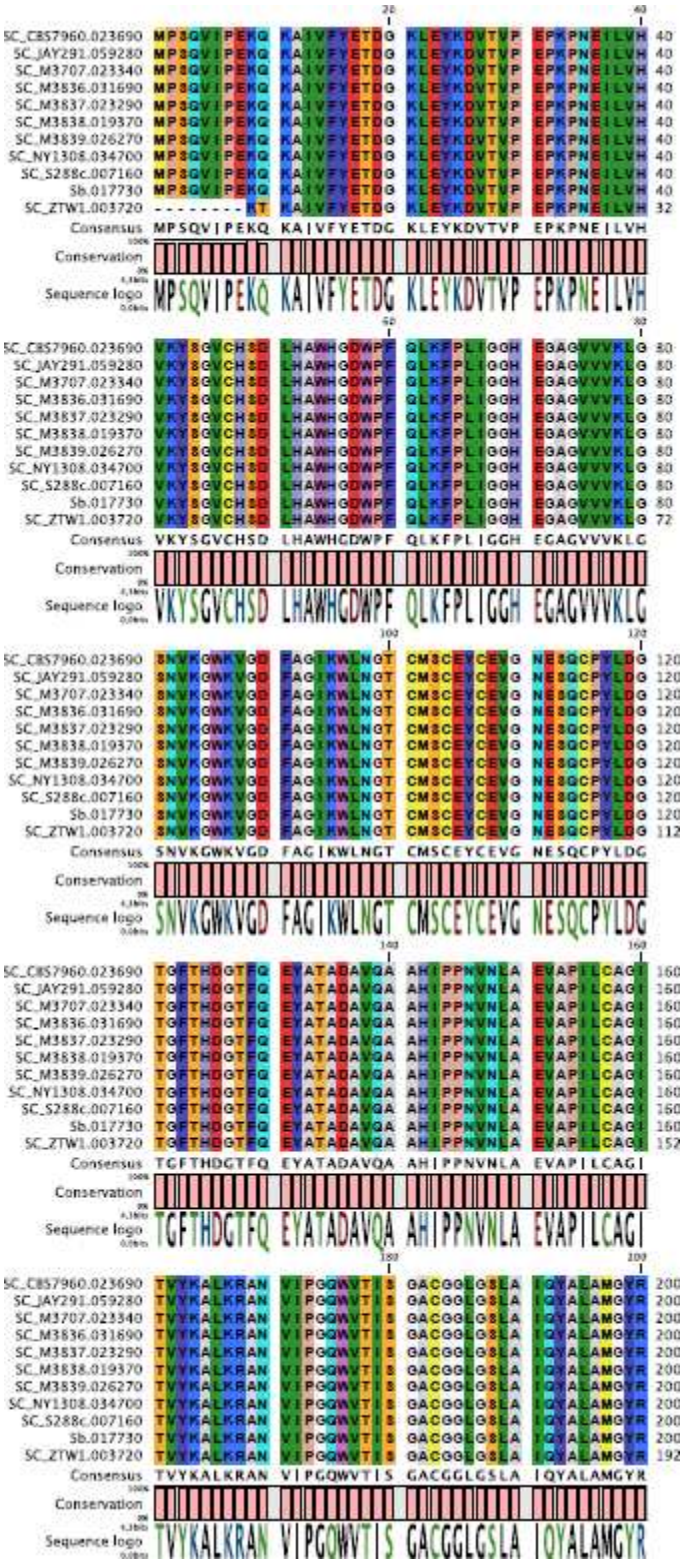
Anexo 10: Alinhamento múltiplo, entre as sequências de proteínas ADH4 do grupo 1 agrupadas no *cluster* Orthomcl902, evidenciado pelo programa CLC Workbench.

(conclusão)



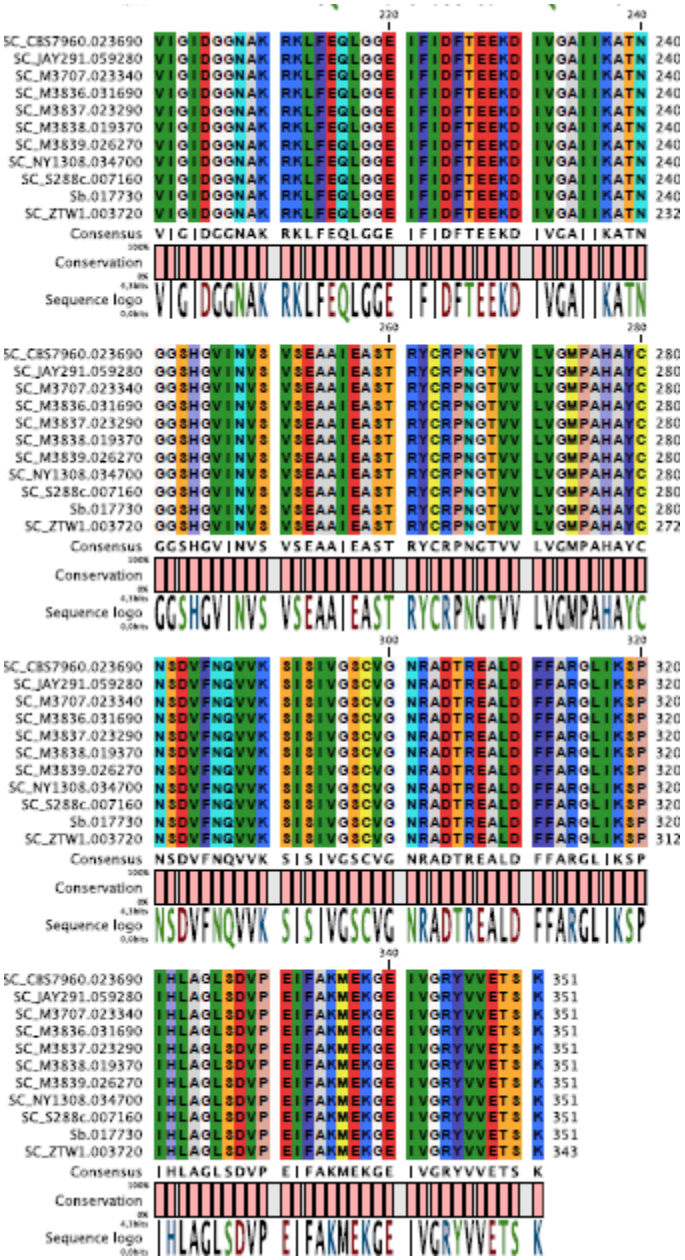
Anexo 11: Alinhamento múltiplo, entre as sequências de proteínas ADH5 do grupo 1 agrupadas no *cluster* Orthomcl3504, evidenciado pelo programa CLC Workbench.

(continua)



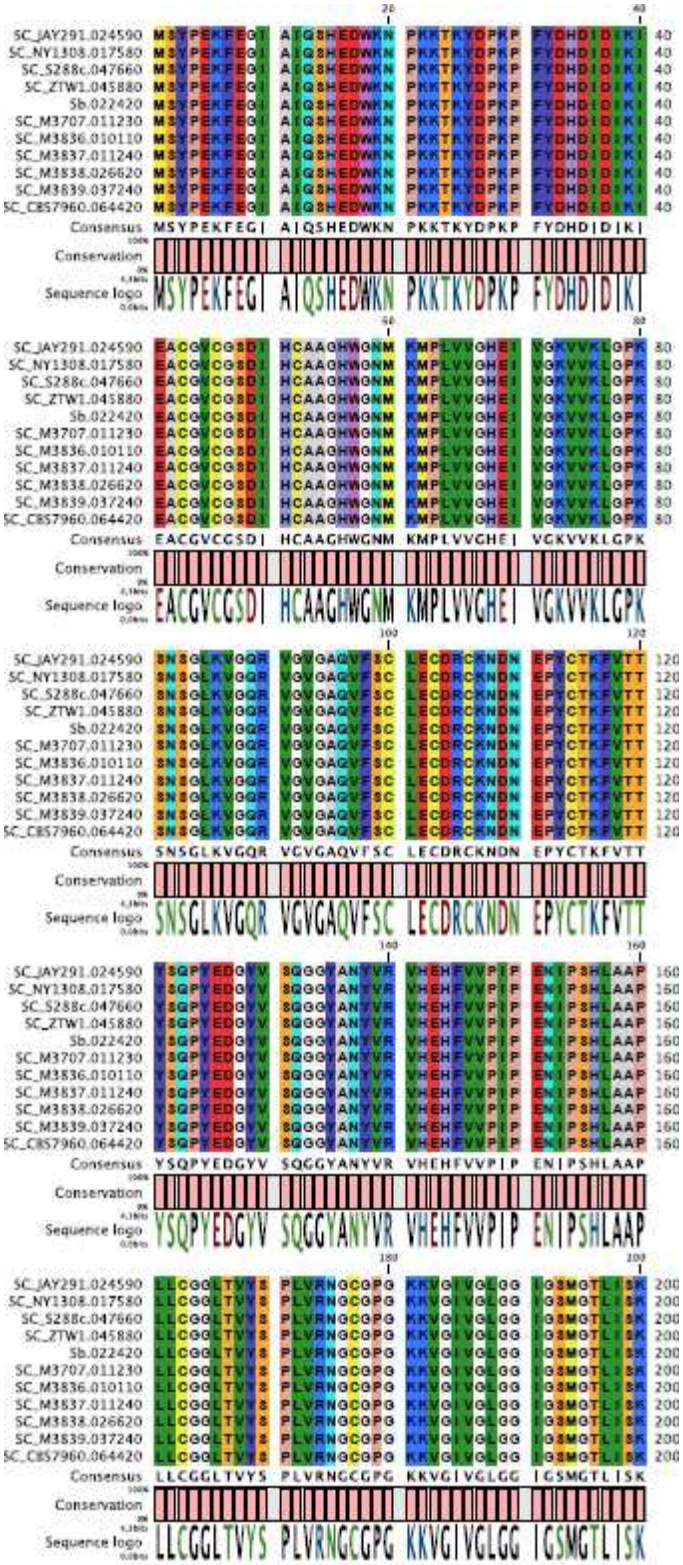
Anexo 11: Alinhamento múltiplo, entre as sequências de proteínas ADH5 do grupo 1 agrupadas no *cluster* Orthomcl3504, evidenciado pelo programa CLC Workbench.

(conclusão)



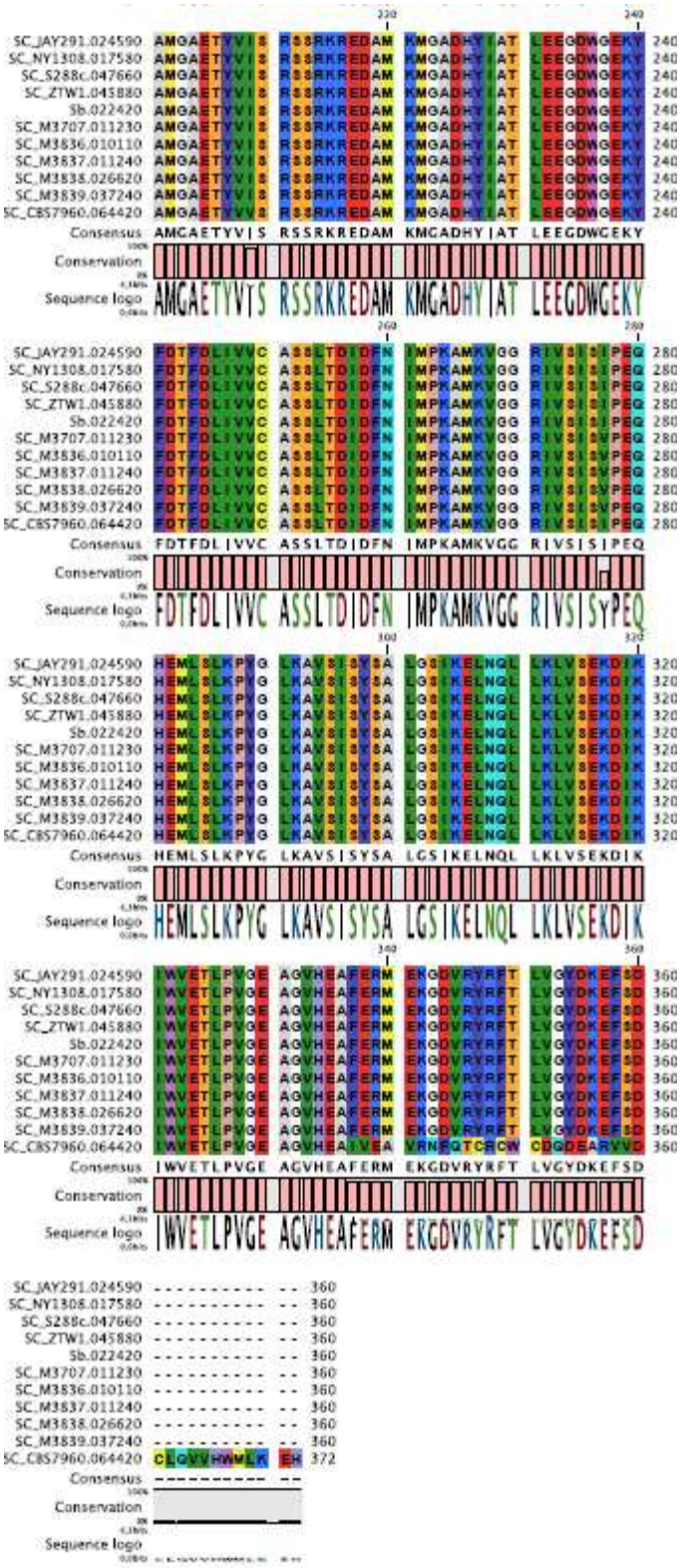
Anexo 12: Alinhamento múltiplo, entre as sequências de proteínas ADH6 do grupo 1 agrupadas no *cluster* Orthomcl571, evidenciado pelo programa CLC Workbench.

(continua)



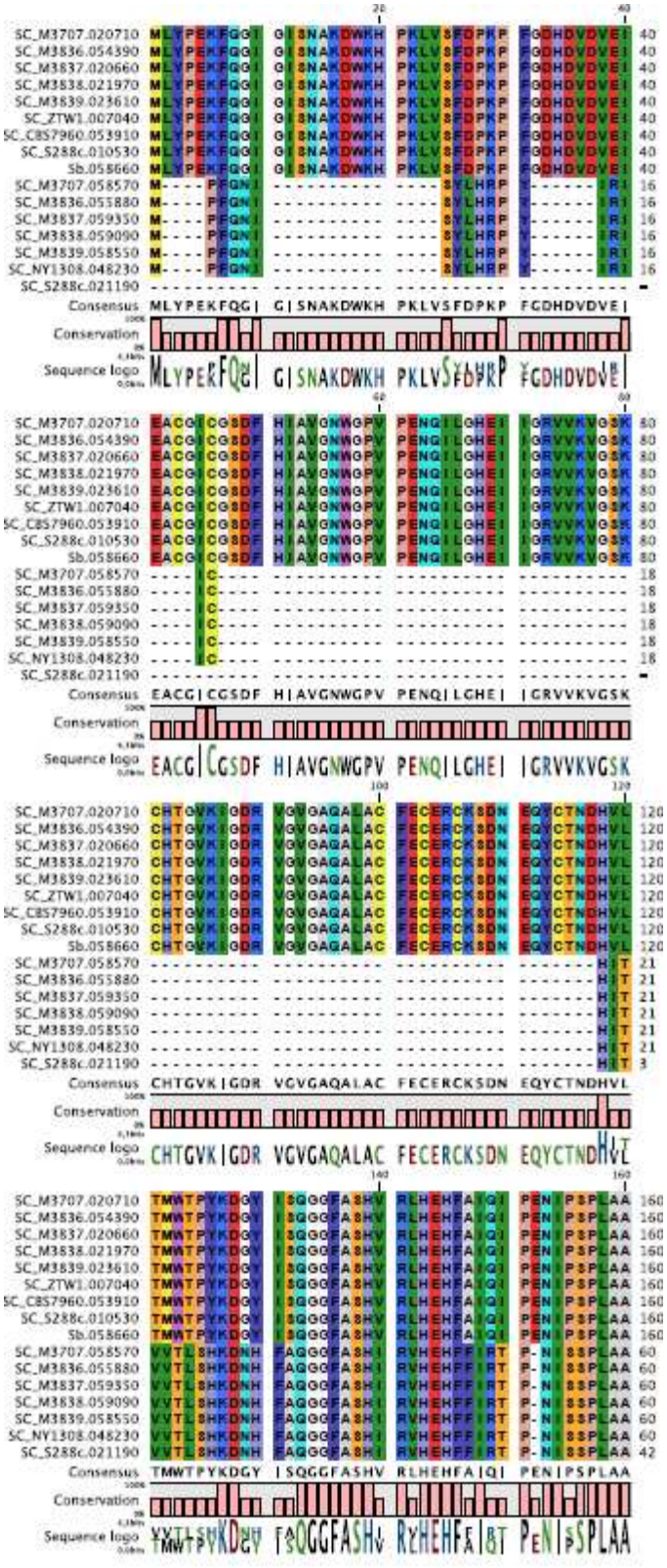
Anexo 12: Alinhamento múltiplo, entre as sequências de proteínas ADH6 do grupo 1 agrupadas no *cluster* Orthomcl571, evidenciado pelo programa CLC Workbench.

(conclusão)



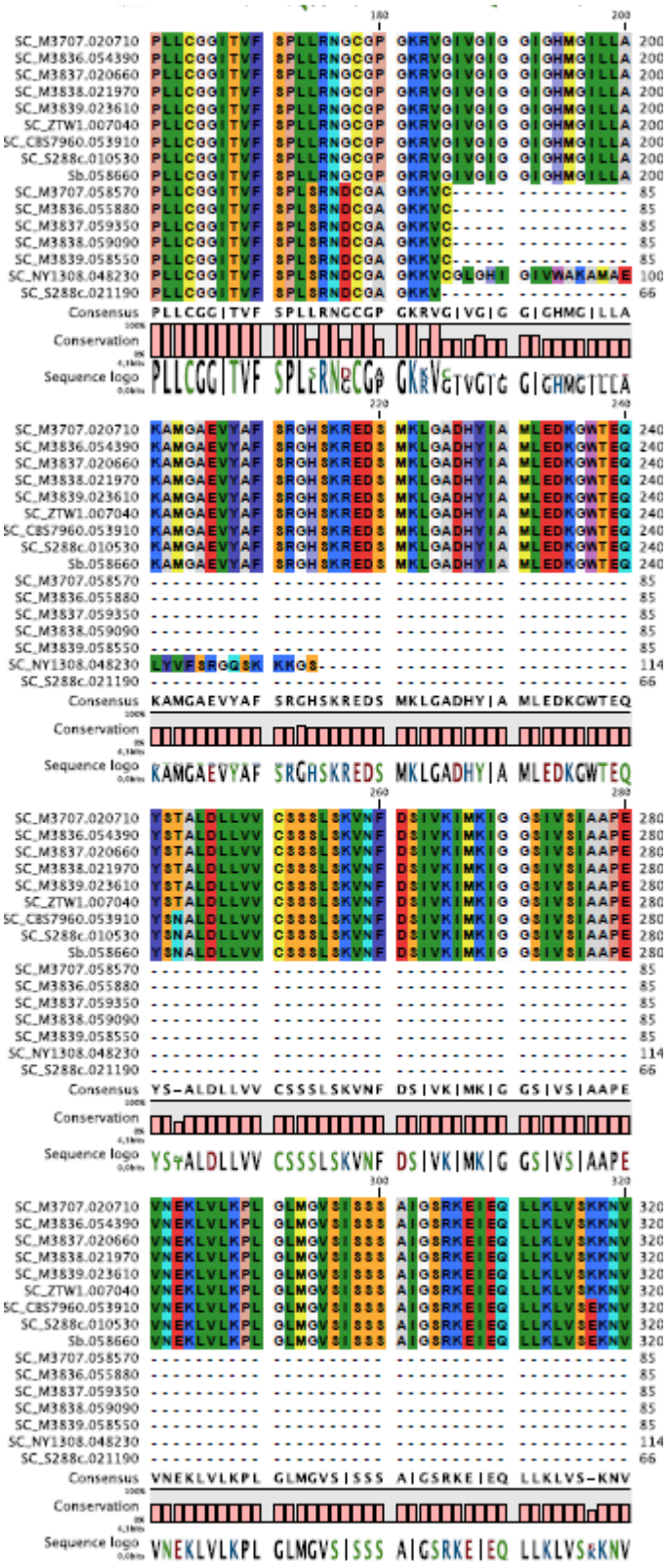
Anexo 13: Alinhamento múltiplo, entre as sequências de proteínas ADH7 do grupo 1 agrupadas nos *clusters* Orthomcl5520 e Orthomcl5585, evidenciado pelo programa CLC Workbench.

(continua)



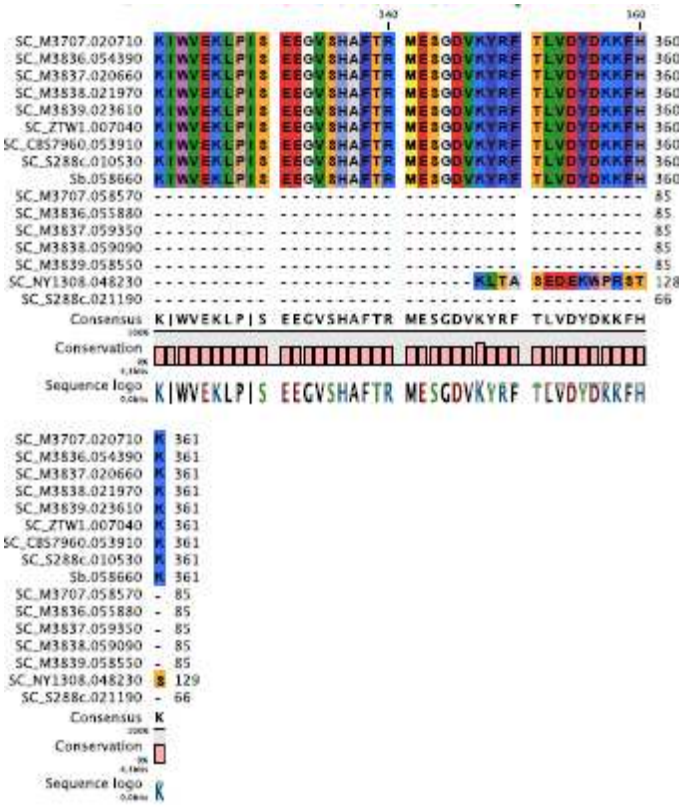
Anexo 13: Alinhamento múltiplo, entre as sequências de proteínas ADH7 do grupo 1 agrupadas nos *clusters* Orthomcl5520 e Orthomcl5585, evidenciado pelo programa CLC Workbench.

(continuação)



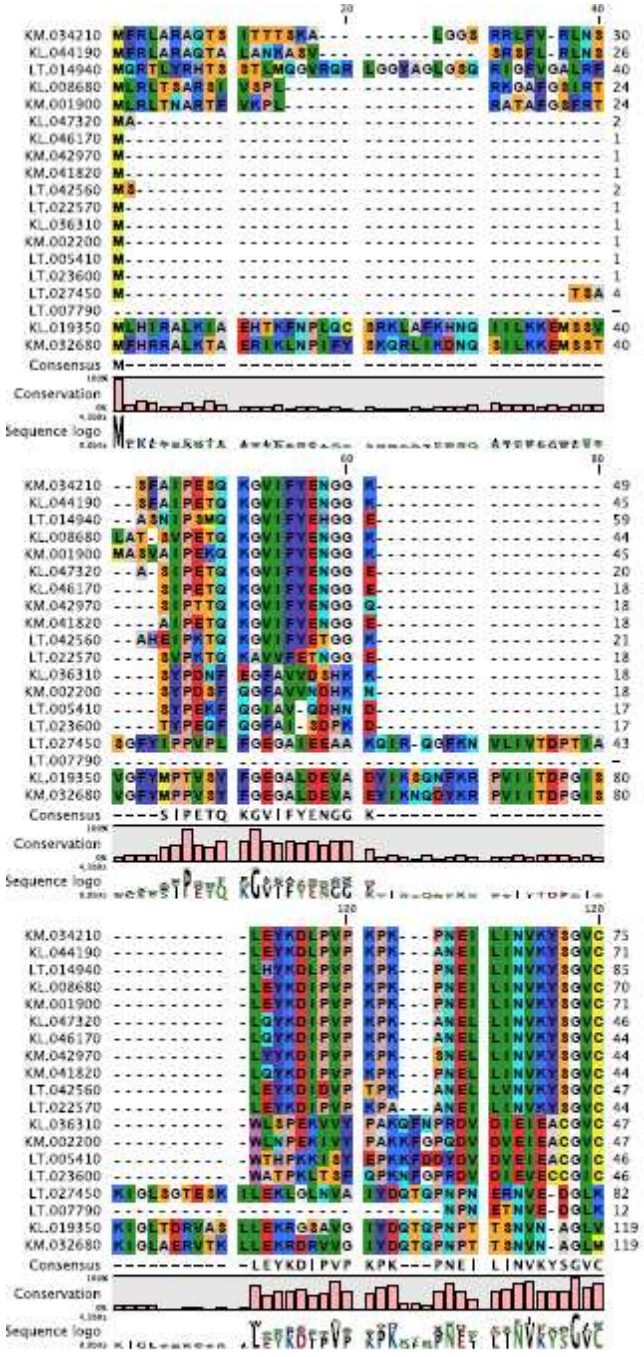
Anexo 13: Alinhamento múltiplo, entre as sequências de proteínas ADH7 do grupo 1 agrupadas nos *clusters* Orthomcl5520 e Orthomcl5585, evidenciado pelo programa CLC Workbench.

(conclusão)



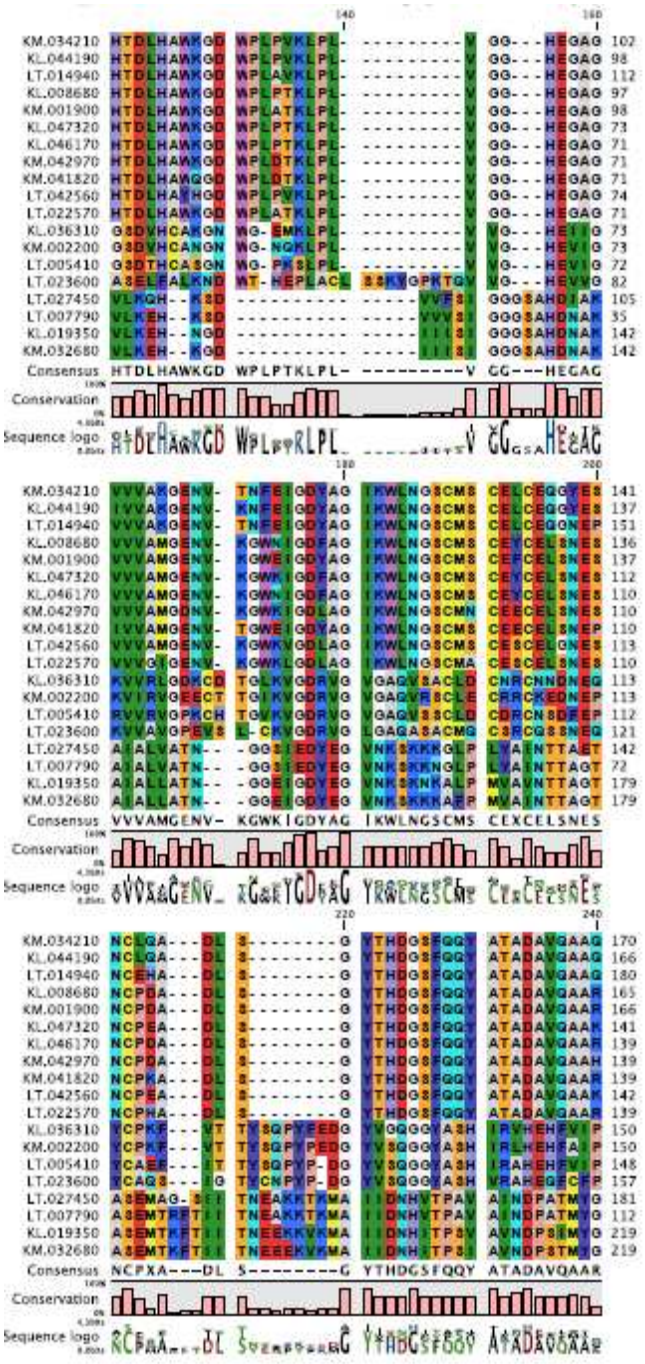
Anexo 14: Alinhamento múltiplo, entre as sequências de proteínas ADHs do grupo 2 agrupadas em *clusters* pelo OrthoMCL e preditas por similaridade, evidenciado pelo programa CLC Workbench.

(continua)



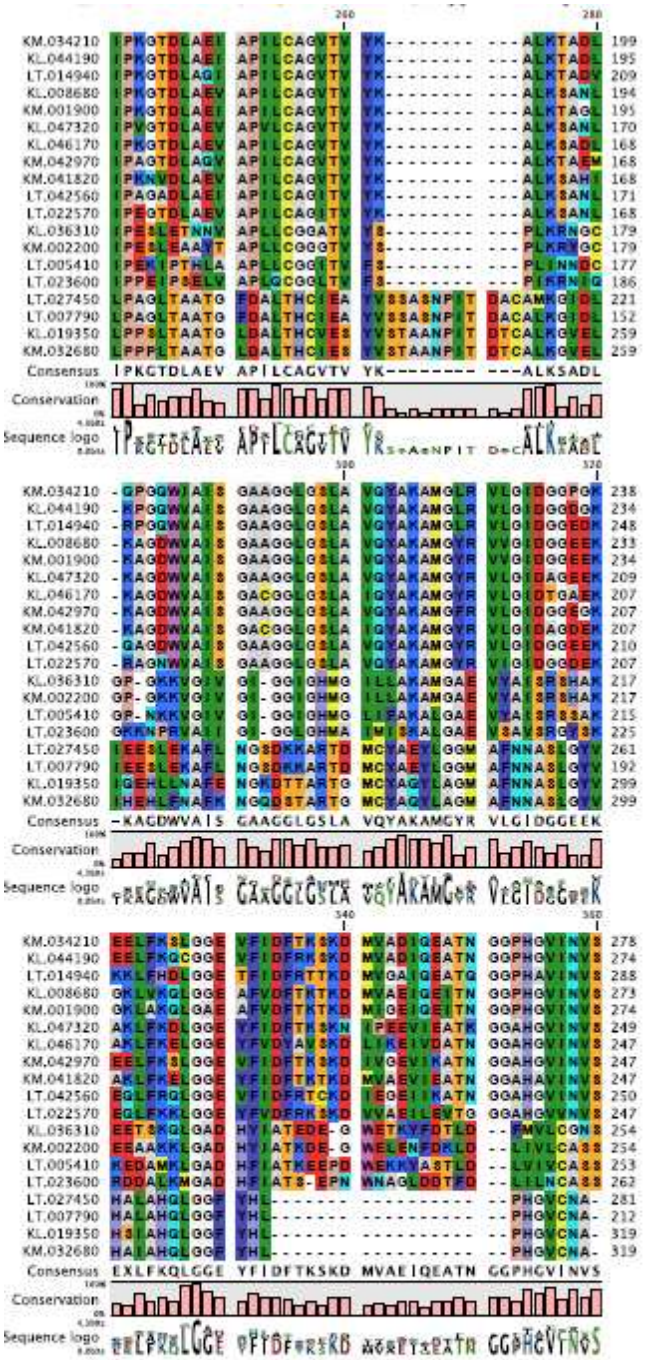
Anexo 14: Alinhamento múltiplo, entre as sequências de proteínas ADHs do grupo 2 agrupadas em *clusters* pelo OrthoMCL e preditas por similaridade, evidenciado pelo programa CLC Workbench.

(continuação)



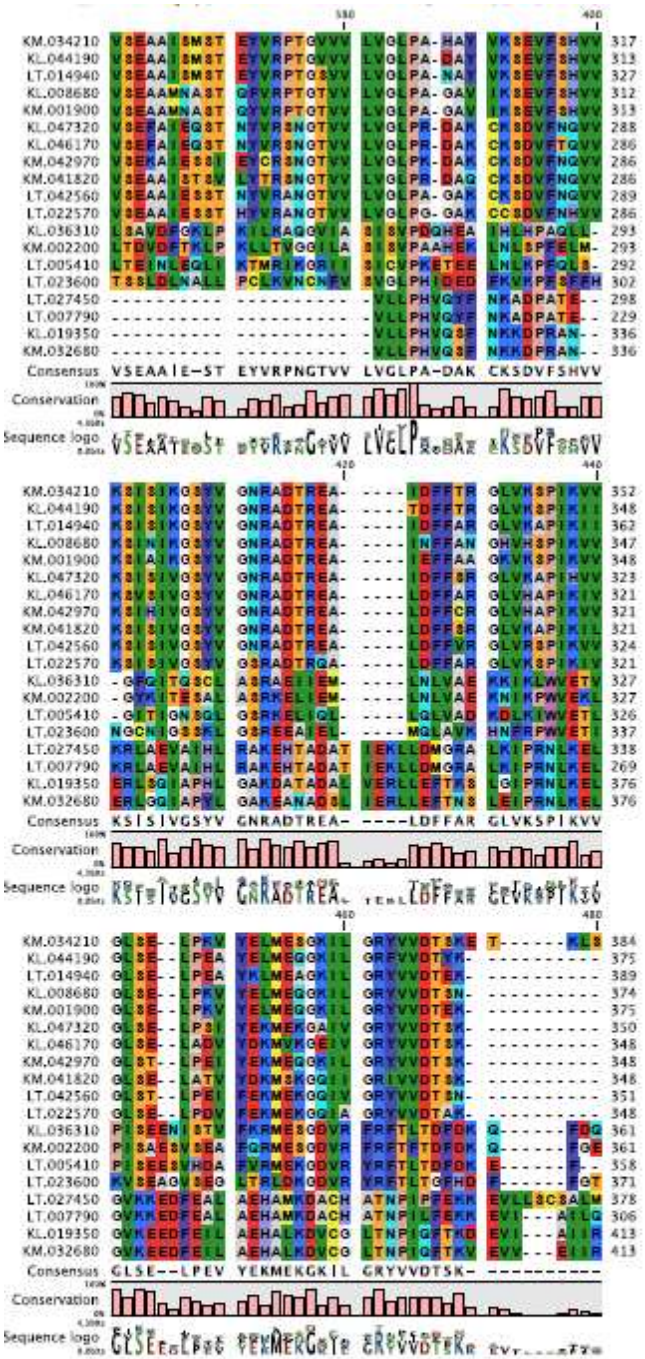
Anexo 14: Alinhamento múltiplo, entre as sequências de proteínas ADHs do grupo 2 agrupadas em *clusters* pelo OrthoMCL e preditas por similaridade, evidenciado pelo programa CLC Workbench.

(continuação)



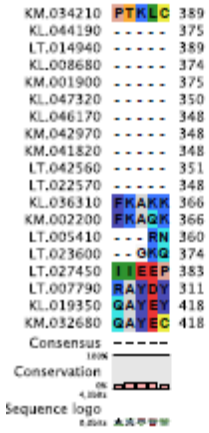
Anexo 14: Alinhamento múltiplo, entre as sequências de proteínas ADHs do grupo 2 agrupadas em *clusters* pelo OrthoMCL e preditas por similaridade, evidenciado pelo programa CLC Workbench.

(continuação)



Anexo 14: Alinhamento múltiplo, entre as sequências de proteínas ADHs do grupo 2 agrupadas em *clusters* pelo OrthoMCL e preditas por similaridade, evidenciado pelo programa CLC Workbench.

(conclusão)



Anexo 15: Matriz resultante do alinhamento múltiplo entre as sequências de proteínas dos *clusters* e proteínas pretidas para ADHs do grupo 1. O lado superior direito representa a porcentagem de identidade entre as sequências. O lado inferior esquerdo representa o número de aminoácidos diferentes encontrados entre as sequências.

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
KM.034210	1		88,69	79,70	75,77	77,30	71,21	69,67	71,72	70,18	70,44	68,64	22,96	22,22	21,29	21,26	13,95	12,56	13,47	13,68
KL.044190	2	44		80,46	76,98	78,57	73,33	72,53	74,13	73,07	73,07	71,73	22,19	21,95	22,03	22,79	13,64	12,44	12,84	13,47
LT.014940	3	81	76		72,38	74,68	69,41	68,12	69,15	69,67	68,38	67,10	21,20	21,69	21,52	21,09	13,03	12,05	13,05	13,05
KL.008680	4	95	87	108		91,47	77,54	77,27	75,67	74,33	74,40	75,13	23,75	22,50	21,57	22,85	13,26	12,00	12,42	12,63
KM.001900	5	89	81	99	32		75,73	75,20	74,93	74,13	73,60	73,60	23,94	22,94	22,03	22,55	14,57	13,38	13,26	13,68
KL.047320	6	112	100	119	84	91		89,14	86,57	85,71	84,90	82,00	24,20	23,94	24,32	24,28	14,06	13,72	12,21	12,42
KL.046170	7	118	103	124	85	93	38		83,91	86,78	81,20	81,90	24,60	24,06	24,46	24,15	14,55	14,29	12,42	12,63
KM.042970	8	110	97	120	91	94	47	56		82,76	83,76	80,17	24,87	24,60	23,91	23,62	13,86	13,53	11,79	12,00
KM.041820	9	116	101	118	96	97	50	46	60		78,35	79,02	24,87	24,87	23,91	24,41	13,86	13,28	12,84	12,84
LT.042560	10	115	101	123	96	99	53	66	57	76		82,05	24,40	25,20	24,80	22,92	14,51	13,68	12,21	12,21
LT.022570	11	122	106	128	93	99	63	63	69	73	63		25,40	24,87	25,27	22,57	13,86	13,78	12,00	12,63
KL.036310	12	312	312	327	305	305	285	282	281	281	285	279		75,96	62,40	44,21	9,63	8,79	8,70	8,92
KM.002200	13	315	313	325	310	309	286	284	282	281	282	281	88		65,40	45,53	8,94	8,54	8,70	8,28
LT.005410	14	318	308	321	309	308	280	278	280	280	279	275	138	127		49,73	8,92	9,05	8,26	8,26
LT.023600	15	326	315	333	314	316	290	289	291	288	296	295	212	207	189		8,13	7,43	7,32	7,53
LT.027450	16	401	399	414	399	393	379	376	379	379	377	379	394	397	398	407		75,52	60,10	60,57
LT.007790	17	376	373	387	374	369	346	342	345	346	347	344	363	364	362	374	94		53,83	53,59
KL.019350	18	411	414	413	416	412	417	416	419	414	417	418	430	430	433	443	168	193		86,36
KM.032680	19	410	411	413	415	410	416	415	418	414	417	415	429	432	433	442	166	194	57	