TESE DE DOUTORADO

Genômica comparativa de leveduras probióticas: *Saccharomyces cerevisiae* UFMG A-905 e cepas de *Saccharomyces boulardii.*

Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Departamento de Bioquímica e Imunologia

Programa de Pós-graduação em Bioinformática

Thiago Mafra Batista

Belo Horizonte, fevereiro de 2015.

Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Programa de Pós-graduação em Bioinformática

Genômica comparativa de leveduras probióticas: *Saccharomyces cerevisiae* UFMG A-905 e cepas de *Saccharomyces boulardii*.

Thiago Mafra Batista

Orientadora: Prof. Dra. Glória Regina Franco

Tese apresentada ao programa de Pós-graduação em Bioinformática do Departamento de Bioquímica e Imunologia da Universidade Federal de Minas Gerais, como parte integrante dos requisitos para obtenção do título de Doutor em Bioinformática.

Belo Horizonte, fevereiro de 2015.

À minha família,

com todo meu amor.

À minha orientadora, Chefa e amiga, Professora Dra. **Glória Regina Franco**, que sempre acreditou e confiou em mim. Obrigado por todas as palavras de carinho e duras quando necessário, todo ensinamento e valores científicos nestes cinco anos de ótimo convívio.

À comunidade virtual de Bioinformatas espalhada pelo mundo em www.biostars.org, www.seqanswers.com e em grupos de discussão, sempre prontamente compartilhando experiências e conhecimentos em **Bioinformática**. Sem eles este trabalho não seria o mesmo.

Aos Professores Dr. **Ieso de Miranda Castro** e Dr. **Jacques Robert Nicoli**, sempre prestativos.

Ao Dr. **Guilherme Oliveira** e a toda equipe do **CEBIO** – Centro de Excelência em Bioinformática – **FIOCRUZ**, sempre à disposição quando precisei.

Ao Professor Dr. **João Trindade Marques**, sempre atencioso.

À Dra. Marcela Drummond e **MyleusBiotecnologia.**

Ao Dr. **Adhemar Zerlotini** Neto, pelos primeiros ensinamentos de Bioinformática pacientemente a mim oferecidos.

Ao **LGB** – Laboratório de Genética Bioquímica, Professor Dr. Carlos Renato Machado (atleticano acima de qualquer coisa!), Professora Dra. Andréa Mara Macedo, Neuza Antunes Rodrigues, e demais colegas pelo crescimento científico e convívio durante todos estes anos.

Aos **Gloriosos** e agregados, em especial à Priscila Grynberg, Dani Durso, Léo Carnevalli, Michele, Ítalo, Raony e André, por toda contribuição e excelente convívio todos estes anos.

Ao Dr. **Rondon Mendonça de Pessoa Neto**, meu amigo e colega de doutorado desde os primeiros dias. Estivemos juntos nas boas e péssimas horas.

Aos grandes amigos do **DA Bio**, pela amizade, pelo papo científico, futebolístico, político e furado, em especial Fafinha, Gabriel Perfeito, Tintim, Heron, Slot, Túlio, Kajuru e demais membros da Cervajonka.

Ao amigo Dr. **Francisco Pereira Lobo**, pelos aconselhamentos científicos sempre que solicitado, e ao Laboratório Multiusuário de Bioinformática da Embrapa/Campinas, pelo suporte computacional a mim disponibilizado durante todo o doutorado.

Ao programa de Pós-graduação em Bioinformática da **UFPR**, por todo apoio computacional e científico neste trabalho, em especial à professora Dra. **Maria Berenice Reynaud Steffens**.

Ao Professor Dr. **Vasco Ariston de Carvalho Azevedo**, Coordenador da Pós-graduação em Bioinformática, sempre competente e solidário!

Aos meus queridos e **amadospais**, que nunca mediram esforços pela minha felicidade e do meu irmão, obrigado por tudo.

À minha **linda** e **amada** esposa Dra. **Juliana de Oliveira Santos**, sempre presente, acompanhando de perto todos os momentos e conquistando juntos, dia após dia, nossa felicidade. Eu te amo!

Ao CNPq, à Capes e à Fapemig.

*A luta pela vitória nem sempreé vantajosa aos fortes nem aos espertos.Mais cedo ou mais tardequem cativa a vitória é aquele que crê plenamente:**“EU CONSEGUIREI!”***

Filosofia do Sucesso - Napoleon Hill

Probióticos são microrganismos vivos presentes em alimentos e suplementos que, quando ingeridos em quantidades suficientes, podem conferir benefícios à saúde do hospedeiro. Isolada por Henri Boulard em 1920 durante um surto de cólera na Indochina (atual Vietnã), a levedura *Saccharomyces cerevisiae* var. *boulardii* é uma levedura não patogênica, termotolerante, e o único microrganismo eucarioto comercializado como probiótico no mundo todo para o tratamento de distúrbios gastrointestinais em humanos. A levedura *Saccharomyces cerevisiae* UFMG A-905 foi isolada a partir de uma coleção de *S. cerevisiae* que foram testadas *in vitro* simulando condições gastrointestinais e *in vivo*, pela capacidade de colonizar o trato gastrointestinal de ratos sem causar patologia. Posteriormente foi demonstrado o seu efeito protetor em animais gnotobióticos desafiados com *Salmonella typhimurium*, *Escherichia coli* e *Clostridium difficile*, caracterizando pela primeira vez, um potencial produto probiótico de origem brasileira. Neste estudo, apresentamos e comparamos a sequência do genoma da cepa *S. cerevisiae* UFMG A-905 e de três cepas de *S. boulardii*. Ambas as cepas possuem genoma com tamanho característico de *S. cerevisiae*, entre 11.4Mb e 11.6Mb, com média de 5.350 genes codificadores de proteínas preditos. A quantidade de termos de ontologia, domínios proteicos e contagem de *Enzime Code* foi semelhante entre as cepas probióticas, embora apenas a cepa Sb_ATCC MYA-796 tenha apresentado evidência de enriquecimento da categoria funcional *"Protein Binding"*, quando comparada à cepa não-probiótica S288c. As relações filogenéticas inferidas a partir do alinhamento de 415 proteínas ortólogas e construção de uma mega-árvore pelo método *Neighbor-Joining* foi capaz de evidenciar três grandes clados, formados por cepas de importância industrial, cepas laboratoriais e cepas fermentadoras alcoólicas, onde estão agrupadas, em um ramo monofilético, as cepas probióticas. As análises de SNVs sugerem uma conservação em nível de variação de nucleotídeos nos genomas probióticos, apresentando média de 51.943 variantes

em cada genoma, com média de uma variante a cada 230 bases, e 70% de impactos causados por estas variantes em comum nos quatro genomas. Foram encontrados 20 genes exclusivos às cepas probióticas e ausentes na cepa não-probiótica S288c, a maioria codificando para proteínas com função desconhecida. Foram encontrados 803 genes de S288c ausentes nas cepas probióticas, sendo 706 em comum às quatro cepas, codificando em sua maioria para proteínas relacionados à atividade de transposição e retroelementos. A ausência de genes codificadores de proteínas relacionadas ao influxo de íons e prótons, bem como proteínas de choque térmico e chaperonas, sugere um envolvimento indireto destes na resistência celular ao estresse ácido.

Probiotics are living microorganisms present in food and supplements that when ingested in sufficient amounts can confer health benefits. The yeast *Saccharomyces cerevisiae* var. *boulardii* was isolated by Henri Boulard in 1920 during a cholera outbreak in Indochina (current Vietnam). *S. boulardii* is non-pathogenic, thermotolerand and the only eukaryotic microorganism commercialized worldwide as a probiotic for the treatment of human gastrointestinal disorders. The yeast *Saccharomyces cerevisiae* UFMG A-905 was isolated from a collection of *S. cerevisiae* that was tested *in vitro* using simulated gastrointestinal conditions. It has also been tested *in vivo* for its ability to colonize mice gastrointestinal tract without causing any pathology. Recently, its prottective effect was demonstrated over gnotobiotic animals challenged with *Salmonella typhimurium*, *Escherichia coli* and *Clostridium difficile*, thus characterizing the first potential probiotic product of a Brazilian origin. In this study, we present the genome sequence of UFMG A-905 and three *S. boulardii* strains. All strains have characteristic *S. cerevisiae* genome sizes between 11.4Mb and 11.6Mb and on average 5,350 predicted protein-coding genes. The number of gene ontology terms, protein domains and Enzyme Code counts were similar between the probiotic strains, although only the Sb_ATCC MYA-796 strain has substantial evidence of enrichment of the functional category “Protein Binding”, when compared to the non-probiotic S288c strain. The phylogenetic relationships inferred from the alignment of 415 orthologous proteins and construction of a mega-tree by the Neighbor-Joiningmethod revealed the clustering of three major clades, composed by strains of industrial importance, laboratory strains and alcoholic fermentation strains, which were grouped in a monophyletic branch of probiotic strains. Analyses of SNVs suggest a conservation of variants in probiotic genomes, with an average of 51,943 variants per genome (one variant every 230 bases) and 70% of the predicted impacts of these variants are shared among all four genomes. Twenty unique genes were

found to be present in probiotic strains and absent in the non-probiotic strain S288c, most of these coding for proteins with unknown function. We found 803 S288c genes that were absent in the probiotic strains, being 706 common to all four strains, coding mostly for protein related to transposition activity and retrotransposons. The absence of genes related to the influx of protons and ions, as well as heat shock proteins and chaperones, suggests an indirect involvement in the cell resistance to acid stress.

BAM	Versão binária do arquivo SAM
CCD	*Charge-coupled device*
CDS	*Coding DNA Sequence*
Contig	Fragmento de sequencia de DNA gerado a partir da montagem das *reads*
dtRNA	Diversidade de RNA transportador
gaps	Fragmento de DNA não sequenciado
GFF	Arquivo de texto contendo informações referentes à anotação de sequencias.
GO	*Gene Ontology*
Kb	Kilo bases
k-mer	Fragmento de sequencia de tamanho k
KOG	*Eu**K**ariotic cluster of **O**rthologous **G**enes*
Mate pair	Tipo de biblioteca de sequenciamento onde são sequenciadas as extremidades um fragmento de DNA ≥ 1kb na orientação *reverse-forward* (RF)
Mb	Mega bases
Montagem*de novo*	Montagem sem a utilização de um genoma de referência.
mRNA	RNA mensageiro
N50	Métrica estatística que representa 50% do tamanho total do genoma contido em N bases.
NGS	*Next generation sequencing*
nr	Banco de dados de proteínas não-redundantes do Genbank
ORF	*Open Reading Frame*
Paired-end	Tipo de biblioteca de sequenciamento onde são sequenciadas as extremidades de um fragmento de DNA ≤ 1kb na orientação *forward-reverse* (FR)
pb	Pares de bases
PCR	*Polymerase Chain Reaction*
Phred	Valor de qualidade do sequenciamento, medido em escala logaritma, que representa a chance de erro de cada base sequenciada

Reads	Leituras individuais de fragmentos de DNA geradas pelo sequenciador de
RNA-Seq	Sequenciamento de RNA
SAM	*Sequence Alignment Map* - Arquivo de texto contendo informações de alinhamento de *reads*
Scaffold	Fragmento de sequencia de DNA gerado a partir da sobreposição de *contigs* ou a partir do mapeamento de *reads* pareadas ao longo dos *contigs*.
Script	Conjunto de instruções em códigos, escritas em linguagem computacional
SNP	*Single Nucleotide Polymorphism*
SNV	*Single Nucleotide Variation*
tRNA	RNA transportador
ttRNA	Total de RNA transportador

1. INTRODUÇÃO

1.1. Probióticos

Probióticos são microrganismos vivos que estão presentes em alimentos ou em suplementos alimentares e que conferem benefícios à saúde do hospedeiro quando ingeridos em quantidades suficientes [1]. Sua utilização em processos alimentícios, como na fermentação de leite, fabricação de pães e produção de bebidas alcoólicas e não alcoólicas é de longa data, mas somente a partir do último século foi que os efeitos benéficos causados pelos produtos lácteos fermentados foram atribuídos aos microrganismos probióticos [2].

Para ser inserido no grupo dos probióticos um microrganismo precisa atender a critérios rigorosos de sobrevivência no organismo do hospedeiro, como elevada temperatura corporal (37 ºC), alterações de pH ao passar pelo trato gastrointestinal, exposição ao suco gástrico, enzimas, tensão de oxigênio, sais biliares e pancreáticos. Além disso, existe a competição com a microbiota nativa, exigindo uma maior adaptação do microrganismo. Somando a esses fatores, é importante que o probiótico seja tolerante também aos produtos secundários provenientes de outros microrganismos e do próprio hospedeiro [3]. Todos esses limitantes podem levar à diminuição da sobrevivência das células do microrganismo e estes, por sua vez, devem permanecer estáveis, viáveis em níveis satisfatórios e ainda resistirem às condições de processamento do produto para comercialização como um probiótico [4].

Bactérias e leveduras têm sido amplamente exploradas no tratamento e prevenção de distúrbios gastrointestinais, sendo as bactérias do gênero *Lactobacillus e Bifidobacterium* as mais utilizadas. Embora a maioria dos probióticos comercializados atualmente sejam de origem bacteriana, a utilização de leveduras como probiótico é altamente vantajosa uma vez que são organismos imunes à ação de antibacterianos, não permanecem no hospedeiro por

mais de cinco dias [5], não colonizam o tubo digestivo [6] e não alteram os níveis populacionais da microbiota normal [7].

Em 1923, na Indochina (atual Vietnã), um microbiologista francês, Henri Boulard, estava à procura de uma linhagem de levedura que fosse capaz de suportar altas temperaturas para produzir um bom vinho. Durante esta época houve uma epidemia de cólera em uma das vilas e ele observou que a população preparava um chá da casca de uma fruta local (a lichia) para aliviar e até mesmo suprimir os sintomas da diarreia. Posteriormente, verificou-se que a fruta, na verdade, estava recoberta por uma levedura e a eficácia contra a diarreia se devia à presença desta levedura, que foi isolada, identificada e chamada de *Saccharomyces boulardii* (FLORASTOR, 2006). A levedura *S. boulardii*é uma cepa não patogênica, termotolerante (cresce na temperatura de 37 ºC) e atualmente de uso muito difundido na medicina humana [8–11]. A partir de 1960 iniciou-se a comercialização da levedura liofilizada, pelo "Laboratoire Biocodex" (Paris, França). Assim, seu uso como medicação para combate às diarreias foi difundido em toda Europa. Atualmente, a levedura é amplamente comercializada na Europa, Américas do Sul e do Norte, Ásia e África [8]. Os direitos de comercialização para a América do Sul foram adquiridos pela MERCK S.A. Indústrias Farmacêuticas [12].

S. boulardii é utilizada contra vários tipos de distúrbios gastrointestinais, como diarreia associada ao uso de antibióticos [13], tratamento da diarreia causada pelo *Clostridium difficile* [14]tanto nos casos de prevenção [15] quanto nos casos de recorrência da doença [16], prevenção e tratamento da diarreia do viajante [17], na manutenção do tratamento da doença de Crohn [18], na prevenção de diarreia em pacientes recebendo alimentação por sonda [19]. Além disso, existem ensaios clínicos mostrando o seu efeito benéfico sobre a microbiota de prematuros [20] e na diminuição da diarreia em pacientes com amebíase aguda [21].

Em animais, a administração de *S. cerevisiae* como probiótico fornece proteção contra lesões intestinais causadas por vários patógenos diarreicos. A vantagem de se trabalhar com levedura é que ela pode ser liofilizada, é rapidamente eliminada após interrupção da terapia e não é afetada pelo uso de antibacterianos [5]. Esta última propriedade é importante, pois algumas terapias associam a administração de probióticos com antibacterianos durante infecções gastrointestinais como, por exemplo, no caso de pacientes infectados por *Helicobacter pylori*, cuja terapia é uma combinação de drogas [22].

*S. cerevisiae*tem sido utilizada há mais de uma década como probiótico em suínos, devido ao seu efeito sobre o crescimento e desempenho reprodutivo e para a redução da morbidade e da mortalidade, especialmente em animais jovens. A levedura *S. cerevisiae* Sc47 é uma linhagem disponível comercialmente, originalmente destinada para a indústria alimentar e tem sido usada por cerca de 20 anos como probióticos de suínos e na pecuária, reduzindo a frequência de diarreia [23]. Recentemente, estudos mostraram que *S. boulardii* possui eficácia também no tratamento de diarreia associada a antimicrobianos em equinos [24].

Alguns dos mecanismos que podem ajudar a entender como as leveduras são capazes de proteger os hospedeiros contra agentes patogênicos incluem o estímulo ao sistema imune, a degradação de toxinas bacterianas por enzima proteolítica da levedura ou a inibição da sua aderência às células epiteliais gastrointestinais pela liberação de uma protease que degrada os receptores de toxina, como do *C. difficile,* e formação de um conglomerado levedura-bactéria pela adesão da bactéria à parede das células da levedura. Existem inúmeras linhagens de *S. cerevisiae,*mas apenas algumas foram estudadas e demonstraram propriedades probióticas [23].

Em um trabalho prévio, várias linhagens de *S. cerevisiae* isoladas de diferentes ambientes do Brasil (associados a insetos, frutas tropicais, queijo e produção de cachaça)

foram pré selecionadas em testes *in vitro*, (simulando as condições gastrointestinais) e posteriormente testadas *in vivo*,pela capacidade de colonizar o trato gastrointestinal sem causar patogenia, e pelo seu efeito protetor em animais gnotobióticos desafiados com *Salmonella typhimurium* e *C. difficile* [12]. Os resultados do estudo sugeriram que *S. cerevisiae* UFMG A-905 teria um potencial probiótico, sendo capaz de reduzir a ação de algumas bactérias patogênicas, assim como de reduzir níveis de translocação de *S. typhimurium* e estimulando o sistema imunológico do hospedeiro [25]. Em 2010, pesquisadores observaram um decréscimo da translocação de *E. coli* 10536 nos grupos tratados com *S. cerevisiae* UFMG A-905 em comparação ao grupo controle, esta diminuição foi significativa tanto para o tratamento com a levedura viável, quanto para a levedura inviável [26]. Vários gêneros de leveduras têm sido testados para uso como agente bioterapêutico, apesar de alguns outros gêneros apresentarem uma provável atividade probiótica, como por exemplo, *Kluyveromyces* [27], apenas o gênero *Saccharomyces* demonstrou, em experimentos *in vivo*, possuir propriedade que a enquadram na categoria de probiótico [25, 28].

Diversos trabalhos vêm sendo desenvolvidos buscando elucidar os mecanismos pela qual a levedura *S. boulardii* age no hospedeiro. Em 2009, Sant'Ana e colaboradores estudaram o efeito protetor de íons contra a morte celular induzida pelo estresse ácido em leveduras do gênero *Saccharomyces* e mostraram que na presença de íons a viabilidade celular é aumentada, sobretudo de *S. boulardii* [29]. Entretanto, pouco se conhece a nível genômico e proteico dos mecanismos envolvidos no efeito probiótico conferido pela levedura.

O advento do sequenciamento de última geração (NGS) em 2005 criou possibilidades sem precedentes para a caracterização de genomas, permitindo avanços significativos na compreensão da sua organização. Atualmente, as tecnologias NGS podem ser empregadas no sequenciamento de grandes genomas [30], na compreensão de diferenças genômicas

individuais dentro da mesma espécie [31], na caracterização do espectro de interação de proteínas ao DNA [32] e permite criar perfis de modificações epigenéticas no genoma [33].

Diversos organismos vem sendo resequenciados utilizando diferentes tecnologias de sequenciamento. Cada equipamento apresenta uma particularidade e diferem-se na construção da biblioteca, na maneira de obtenção da sequência, no tamanho e quantidade de *reads* geradas [34]. Embora as cepas de *S. cerevisiae* apresentem uma forte conservação genômica, algumas diferenças fisiológicas relevantes são atribuídas apenas à aquisição ou deleção de poucos genes, por exemplo, a ausência do gene *MALx3* em *S. cerevisiae* S288c leva a um fenótipo maltose-negativo, enquanto a aquisição do gene *ENA6* torna a cepa CEN.PK113-7D mais resistente a íons lítio [35]. A enorme quantidade de dados gerados nestas novas tecnologias nos permitem explorar e inferir, com uma maior confiabilidade estatística, acerca das diferenças genômicas entre organismos.

1.2. Genômica comparativa

A partir da publicação do *draft* do genoma humano em 2001 [36, 37], inúmeras perguntas foram surgindo e com elas, mecanismos para buscar estas respostas. Uma das abordagens mais poderosas para desvendar os segredos do genoma é a genômica comparativa, que permite analisar em larga escala diferentes genomas e assim, compreender os mecanismos e forças evolutivas atuantes sobre eles. Publicado em 2002, a análise comparativa do genoma do camundongo *Mus musculus* com o genoma humano, refinou a contagem de genes comuns entre mamíferos em aproximadamente 30 mil genes e permitiu estimar em 5% o percentual do genoma que é conservado entre mamíferos placentários, e portanto, funcional[38].

Uma das essências da genômica comparativa é de que a sequência de DNA que permanece conservada entre espécies é susceptível à manutenção da similaridade devido a pressões evolutivas, o que implica em uma função biológica. Entretanto, o inverso não é necessariamente verdade, pois uma sequência de DNA pode ter uma função biológica sem haver conservação com outras espécies. Isso é especialmente verdadeiro para alterações recentes de linhagens em que o tempo evolutivo ainda não proporcionou uma assinatura de conservação à sequência de DNA. Conservação da função biológica não implica necessariamente em identidade de sequencias [39].

Fungos produzem uma variedade de enzimas ativadas por carboidratos (CAZymes - *Carbohydrate-activee nzymes*) que desempenham um papel importante na interação patógeno-hospedeiro, promovendo a síntese, quebra, ou modificações na parede celular de plantas, facilitando a infecção ou o ganho nutricional. As análises comparativas de 103 genomas de fungos revelaram uma relação entre tamanho e diversidade destas enzimas com a estratégia nutricional e especificidade dela planta hospedeira [40].

O genoma de cepas selvagens e laboratoriais de *S. cerevisiae* apresentam variedades genômicas significativas[41]. Como efeito, as perturbações ambientais frequentemente selecionam cepas que apresentam duplicações gênicas, alterações no número de cópias de cromossomos, ou rearranjos intra ou intercromossomais mediados por transposons[42–44].

Neste contexto, a análise comparativa de genomas surge como uma poderosa ferramenta na busca por padrões genômicos que possam estar envolvidos no caráter probiótico que possuem as cepas *S. boulardii* e *S. cerevisiae* UFMG A-905.

1.3. Novas tecnologias de sequenciamentos de DNA/RNA

O rápido progresso das tecnologias de sequenciamento de DNA a partir da publicação do genoma humano foi impulsionado pela necessidade da redução de custos e aumento no rendimento e acurácia na aquisição das sequências genômicas. O primeiro sequenciador de nova geração foi o Genome Sequencer 454 lançado pela empresa 454 *Life Science Corporation* em 2005, seguido do lançamento do *Genome Analyzer* pela empresa Solexa em 2006 e pelo lançamento do SOLiD (***S**equencing by **O**ligo **Li**gation **D**etection*) pela empresa Agencourt. Estas empresas foram adquiridas pela Roche, Illumina e Applied Biosystems, respectivamente, que comercializam os sequenciadores mais utilizados pela comunidade científica.

O Roche GS 454 utiliza o método de Pirosequenciamento [45], que é dependente da detecção do pirofosfato liberado durante a incorporação do nucleotídeo. Nesta metodologia, o DNA é fragmento por sonicação, ligado a adaptadores específicos do 454, desnaturado em fita única, capturado por microesferas magnéticas, seguido de amplificação por PCR em emulsão. A reação de sequenciamento acontece em uma placa PicoTiter^{TM} contendo moléculas de dNTPs (dATP, dCTP, dGTP e dTTP) que complementarão a fita molde de DNA liberando um pirofostato, que será convertido em ATP pela enzima ATP-sulforilase, sendo esta utilizada como fonte de energia pela enzima luciferase para oxidar o substrato luciferina, produzindo oxiluciferina e um sinal de luz que é detectado por uma câmera CCD (*charge-coupled device*) de alta sensibilidade acoplada ao sequenciador. A detecção do sinal de luz é proporcional à quantidade de bases incorporadas. As bases não incorporadas são degradas pela enzima apirase, e então novos dNTPs são adicionados ao sistema de reação e a reação de pirosequenciamento é repetida [34, 46]. Em seu sistema mais atual, o 454 GS FLX Titanium XL+ gera *reads* com tamanho médio de 700 bases, com acurácia de 99,979% e *output* de 700

Mega bases de dados em uma corrida de sequenciamento com duração de 23 horas (http://454.com/products/gs-flx-system/index.asp). Com este método só é possível construir bibliotecas do tipo fragmento ou *single-reads*.

O SOLiD System™ da Applied Biosystem adota uma tecnologia diferente das demais tecnologias, utilizando uma enzima DNA ligase em vez de uma enzima DNA polimerase para a incorporação de oligonucleotídeos marcados com quatro diferentes fluoróforos para a detecção da sequência alvo. As bibliotecas de DNA são ligadas à microesferas magnéticas e amplificadas por PCR em emulsão[34]. A reação de sequenciamento acontece em uma lâmina (*flowcell*) onde as microesferas contendo uma única molécula de DNA são fixadas e entram em contato com os oligonucleotídeos, que são formados por duas bases, seguidas de três inosinas e três bases degeneradas, sendo que a última base degenerada é marcada com um fluoróforo contendo uma cor (azul, vermelho, verde ou amarelo) correspondente às duas primeiras bases aneladas. A especificidade da ligação é garantida pela ligação das duas primeiras bases dos oligos, enquanto as três inosinas seguintes se anelam de maneira inespecífica e as bases degeneradas são clivadas, liberando o sinal do fluoróforo para detecção. Ciclos posteriores são realizados utilizando *probes* de diferentes tamanhos, garantindo que haja a ligação de dois oligos por base do DNA molde[46]. Em sua versão mais atual, o SOLiD 5500xl gera pequenos fragmentos de sequências de até 75 bases, com acurácia de 99,999%, com um *output* de até 320 Giga bases de dados (http://www.lifetechnologies.com/br/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing.html). É possível construir bibliotecas de fragmento (*single-read*) e bibliotecas pareadas do tipo *paired-end* e *mate-pair* para sequenciamento com SOLiD. Estas duas últimas facilitam a identificação dos pequenos fragmentos de DNA no momento da montagem do genoma, uma vez que são sequenciados

dois fragmentos de uma molécula de DNA com tamanho de espaçamento conhecido entre eles. Posteriormente, a Applied Biosystem foi adquirida pela companhia Life Technologies.

O Genome Analyser, posteriormente lançado como Illumina HiSeq utiliza a tecnologia baseada no sequenciamento por síntese. Nesta metodologia, a biblioteca de DNA é ligada a adaptadores, desnaturada em fita simples e fixada a uma lâmina contendo adaptadores complementares aos adaptadores iniciais, seguida de uma amplificação por *Bridge-PCR*[47] (PCR em fase sólida, em uma tradução livre) para a formação de *clusters* contendo fragmentos de DNA clonal. A reação de sequenciamento acontece na presença da enzima DNA polimerase e de nucleotídeos marcados com diferentes fluoróforos que são detectáveis por uma câmera CCD no momento da síntese da molécula molde[46]. A Illumina possui vários modelos de sequenciadores que diferem principalmente na aplicação e quantidade de dados gerados. A versão HiSeq 2500 gera *reads* de tamanho máximo 2x150 bases e um *output* de até 1000 Giga bases com 6 dias de corrida. O MiSeq, versão compacta do sequenciador, gera *reads* de tamanho máximo 2x300 bases e um *output* de até 15 Giga bases com 55 horas de corrida (http://www.illumina.com/systems/sequencing.html).

A 3ª geração de sequenciadores de DNA se baseia na detecção direta do nucleotídeo sequenciado e não da detecção de um sinal. Destacam-se alguns sequenciadores, como o Ion Torrent™ e Ion Proton™ da empresa *Life Technologies* (http://www.lifetechnologies.com) que utilizam sequenciamento em um chip semicondutor que detecta a variação de pH durante a incorporação da base. O sequenciador PacBio™ da empresa *Pacific Bioscience* (http://www.pacificbiosciences.com) utiliza nucleotídeos marcados com fluoróforos e uma enzima DNA Polimerase acoplada ao fundo de uma microplaca que detecta o nucleotídeo no momento da incorporação da base. O MinIon™ da empresa *Oxford Nanopore Technologies* (https://ww.nanoporetech.com), é um sequenciador portátil com conexão USB que identifica as diferentes bases da molécula de DNA através de uma proteína sensora acoplada a um

nanoporo, e a transmissão dos dados é feita em tempo real para o computador. Nesta terceira geração os sequenciadores possuem características e aplicações não alcançadas na segunda geração, como *reads* muito grandes, detecção de modificações de bases e redução de custos[48].

Neste trabalho, utilizamos os sequenciadores de segunda geração SOLiD v4, Illumina MiSeq e HiSeq 2500, e analisamos dados provindos de sequenciamento realizado com o GS 454 FLX+.

2. OBJETIVOS.

2.1. Objetivo geral.

Comparar os genomas de cepas de leveduras probióticas e não probióticas em busca de genes ou elementos genômicos que possam estar associados a atividade probiótica.

2.2. Objetivos específicos

- Sequenciar e montar os genomas das cepas UFMG A-907 e Sb_17 utilizando tecnologias de nova geração (NGS);
- Sequenciar o mRNA de UFMG A-905 e Sb_17, montar os transcritos e utilizá-los na anotação do genoma destas cepas;
- Investigar a presença de SNVs nos genomas das cepas probióticas;
- Investigar genes presentes em cepas probióticas e ausentes em cepas não probióticas;
- Investigar genes presentes em cepas não probióticas e ausentes em cepas probióticas;
- Investigar genes relacionados à resistência ao estresse ácido;
- Investigar a relação filogenética entre as cepas probióticas e cepas não probióticas de *S. cerevisiae*;
- Analisar grupos funcionais enriquecidos/depletados nas cepas probióticas em comparação a cepas não probióticas.

3. MATERIAL E MÉTODOS.

3.1. Cultivo das células e extração de DNA/RNA

Células de *S. cerevisiae* var *boulardii* 17 e *S. cerevisiae* UFMG A-905 foram cultivadas em meio YPD suplementado com 2% de glicose até a fase estacionária. O DNA genômico foi extraído utilizando o kit Genomic-Tip 100/G (Qiagen, Germany), seguindo recomendações do fabricante e estocado a 4º C. O RNA total foi extraído utilizando o método do fenol ácido quente [49] e estocado a -80º C. As amostras de RNA foram purificadas utilizando Rneasy® MiniElute™ Cleanup Kit (Qiagen, Germany) seguindo as recomendações do fabricante. O RNA total foi quantificado usando Nanodrop ND-100 UV/Vis (NanoDrop Technologies, USA) e a qualidade do RNA foi comprovada por eletroforese em gel desnaturante. O RNA total foi enviado para sequenciamento na empresa BGI (*Beijing Genome Institute* – China) em tubos de RNAstable® (Biomátrica Company) para garantir a integridade do material biológico.

3.2. Construção de bibliotecas e sequenciamento de DNA/RNA

Utilizamos duas plataformas distintas de sequenciamento de nova geração (NGS), SOLiD e Illumina. Para sequenciamento na plataforma SOLiD foram construídas bibliotecas *mate-pair* com espaçamento de 1-2kb entre adaptadores e foram geradas *reads* pareadas de 50 bp. Foram utilizados dois sequenciadores Illumina, MiSeq e HiSeq, e para ambos foram construídas bibliotecas *paired-end* com distanciamento médio de 300 bp, e foram geradas *reads* pareadas de 151 bp e 101 bp, respectivamente.

3.3. Montagem dos genomas

As *reads* foram inicialmente analisadas quanto à qualidade do sequenciamento, utilizando o software FASTQC [50]. Leituras com valor de qualidade abaixo de *phred* 20 (para a plataforma SOLiD) e abaixo de *phred* 30 (para a plataforma Illumina) foram descartadas do conjunto de dados. As montagens *de novo* foram realizadas utilizando o algoritmo *De-Brujin Graph*.

Os dados obtidos pela plataforma SOLiD foram montados utilizando o Pipeline SOLiD^{TM}*de novo* accessory tools 2.0 (disponibilizado à época publicamente pela Applied Biosystem). O melhor valor de *k-mer* foi definido utilizando o *scriptVelvetOptimiser.pl* e os *contigs* foram gerados pelo software Velvet v1.2.07 [51].

Os dados obtidos pela plataforma Illumina foram montados utilizando o software SOAPdenovo v2.04 [52] com parâmetros padrão. Quando conveniente, foi realizada uma montagem híbrida combinando os *contigs* gerados pelas duas plataformas utilizando o software Zorro [53]. Através do software SSPACE [54] foram obtidos os *scaffolds* que posteriormente foram ordenados pelo programa CONTIGuator [55] utilizando como referência o genoma da cepa *S. cerevisiae* S288c. Os *gaps* remanescentes foram fechados pelo software GapCloser [52]. As métricas estatísticas de qualidade da montagem foram obtidas através de *scripts* em Perl, avaliando a soma dos *contigs*, qualidade de *contigs*, tamanho do maior *contig*, valor de N50, quantidade de *gaps* e completude do genoma, avaliada pela presença de genes ortólogos do KOG (Eu**K**ariotic clusters of **O**rthologous **G**roups) pelo software CEGMA (Core Eukariotic Genes Mapping Approach) v2.4 [56]. Os *scaffolds* mapeados no genoma de S288c foram visualizados graficamente e as imagens construídas com o software BRIG v 0.95 [57].

3.4. Anotação estrutural dos genomas

A predição gênica consiste em identificar no genoma as regiões biologicamente funcionais, que incluem genes codificadores de proteínas, bem como genes de RNA regulatórios. A busca por regiões codificadoras de proteínas foi realizada utilizando o software Maker2 [58] combinando a predição *ab initio* de dois preditores externos (Augustus e SNAP) e utilizando informação extrínseca de CDS (do inglês *Coding DNA Sequence*) e peptídeos presentes no genoma de *S. cerevisiae* S288c. O preditor Augustus utilizou o modelo intrínseco de S288c (disponíveis em arquivos *.fasta*), enquanto o preditor SNAP foi treinado utilizando o modelo gênico predito pelo CEGMA presente no arquivo GFF. Foram utilizados os seguintes comandos para treinar o SNAP:

```
$ cegma2zff output.cegma.gff genoma.fasta
$ fathom genome.ann genome.dna -categorize 1000
$ fathom -export 1000 -plus uni.ann uni.dna
$ forge export.ann export.dna
$ hmm-assembler.pl $genoma . > genome.cegmasnap.hmm
```

Para anotação das cepas A-905 e Sb_17, as *reads* de RNA-Seq foram alinhadas contra os *scaffolds* com o software Tophat2 [59]. As junções identificadas foram convertidas em um arquivo GFF pelo *script tophat2gff3* e utilizadas como informação intrínseca na anotação dos genomas. Para a identificação de tRNAs foi utilizado o software tRNAscan-SE [60], e para a identificação de regiões repetitivas foi utilizado o software RepeatMasker v4.0.5 com os seguintes parâmetros:

-species saccharomyces (busca por repetições presentes no gênero “saccharomyces”)

-s (força o algoritmo a ser mais estringente)

-excln (ignora as regiões que contenham ≥ 20 Ns)

3.5. Anotação funcional dos genomas

As possíveis ORFs (janelas de leitura aberta, do inglês ***O**pen **R**eading **F**rame*) identificadas na anotação estrutural foram mapeadas na busca de similaridade proteica contra o banco de dados de proteínas não redundantes (nr) do Genbank, utilizando Blastp [61] com valor de *e-value* 1e-10^{-3} e anotadas utilizando os termos de ontologia do Gene Ontology com o software Blast2GO v2.5.0 [62] e os termos de ontologia refinados com GOSlim. As assinaturas de domínios proteicos foram mapeadas utilizando Interproscan 5.0 incorporado ao software CLC Genomics Workbench 7.5 (www.clcbio.com) com o *plugin* Blast2GO Pro.

3.6. Relações filogenéticas

As relações filogenéticas entre as cepas probióticas e demais cepas não probióticas de *Saccharomyces* foi investigada alinhando proteínas ortólogas presentes nos genomas. Em setembro de 2014, haviam sido depositados 48 genomas no Ganbank ((http://www.ncbi.nlm.nih.gov/genome/genomes/15). Analisamos todos os 48 genomas quanto à completude (com o software CEGMA) e utilizamos apenas os genomas que apresentaram no mínimo 98% de completude. Desta forma, não utilizamos 13 genomas com completude inferior a 98% e 3 genomas redundantes. Conduzimos as análises utilizando 32 genomas de cepas de *S. cerevisiae*, os 4 genomas das cepas probióticas e como grupo externo, a levedura *Saccharomyces pastorianus.* O software CEGMA utiliza um core de 458 proteínas do KOG, e gera um arquivo de saída contendo todas as proteínas identificadas no genoma. Foram utilizadas apenas as proteínas presentes em todos os 37 genomas abaixo citados, totalizando 415 proteínas que foram concatenadas gerando um arquivo contendo uma única sequencia proteica para cada organismo. Utilizamos o software ClustalOmega v1.2.1 [63]

para o alinhamento par-a-par das sequências, construímos a árvore utilizando o método de matriz de distância, seguindo o modelo *Neighbour-Joining* sem ignorar os sítios com *gaps* e com suporte estatístico de 1000 replicatas de *bootstrap* utilizando o software Geneious (www.geneious.com). Abaixo, a relação das 37 leveduras utilizadas nesta análise com seus respectivos números de acesso ao genoma: *Saccharomyces pastorianus* (GI: 224836371), *S. cerevisiae* FostersB (GI: 323306170), *S. cerevisiae* FostersO (GI: 323310288), *S. cerevisiae* VL3 (GI: 323356373), *S. cerevisiae* LalvinQA23 (GI: 323349939), *S. cerevisiae* VIN13 (GI: 323338914), *S. cerevisiae* Kyokay nº 7 (GI: 347729985), *S. cerevisiae* ZTW1 (GI: 410375333), *S. cerevisiae* YJM789 (GI: 151946710), *S. cerevisiae* JAY291 (GI: 256274458), *S. cerevisiae* EC1118 (GCA_000218975.1), *S. cerevisiae* RM11-1a (GI: 61385875), *S. cerevisiae* CEN.PK113-7D (GI: 39881196), *S. cerevisiae* AWRI796 (GI: 323334825), *S. cerevisiae* T7 (GI: 330378716), *S. cerevisiae* YJSH1 (GI: 393396067), *S. cerevisiae* IR-2 (GI: 565476325), *S. cerevisiae* M3707 (GI: 478841355), *S. cerevisiae* M3836 (GI: 478841409), *S. cerevisiae* M3837 (GI: 478841692), *S. cerevisiae* M3838 (GI: 478841688), *S. cerevisiae* M3839 (GI: 478841952), *S. cerevisiae* N85 (GI: 646223699), *S. cerevisiae* NAM34-4C (GI: 565471015), *S. cerevisiae* NY1308 (GI: 516425807), *S. cerevisiae* P283 (GI: 570304287), *S. cerevisiae* P301 (GI: 570306094), *S. cerevisiae* R008 (GI: 570305200), *S. cerevisiae* R103 (GI: 570305647), *S. cerevisiae* Sigma1278b (GI: 295413815), *S. cerevisiae* BY4741 (download diretamente do SGD), *S. cerevisiae* W303 (GI: 402234185), *S. cerevisiae* S288c (GCA_000146045.2), *S. cerevisiae* UFMG A-905 (GI: 685248134) *S. boulardii* 17 (GI: 685248026), *S. boulardii* ATCC (GI: 699025856), *S. boulardii* EDRL (GI: 528490851).

3.7. Identificação de variantes (SNVs) nos genomas probióticos

A presença de variações de nucleotídeos nos genomas das cepas probióticas em relação à cepa de levedura não probiótica foi investigada mapeando as *reads* de Illumina utilizadas nas montagens, bem como as *reads* de 454 utilizadas na montagem do genoma de *S. boulardii* EDRL e depositadas no banco de dados SRA (*Sequence Reads Archive* – NCBI) sob o número de acesso SRR1105784, contra o genoma de *S. cerevisiae* S288c, versão EF3.64 (Ensembl), utilizando o software BWA [64]. O arquivo SAM, que contém todas as informações do mapeamento, foi processado utilizando o pacote Samtools [65] e as variantes (SNPs, indels e variações estruturais) identificadas utilizando o pacote GATK v3.2.2 [66] seguindo o *workflow* "*bestpractice*". SNPs (*Single Nucleotide Polymorphism*) são variações de apenas uma base na fita de DNA e indels são variações do tipo inserções ou deleções de bases. Para identificar as variantes e principalmente, excluir as falso-positivas, a primeira etapa do GATK consiste em identificar as regiões onde existe alguma variante, utilizando o parâmetro '-T *RealignerTargetCreator'*. Uma vez identificadas estas regiões de variantes, o algoritmo realiza um alinhamento local através do parâmetro '-T *IndelRealigner'*, corrigindo possíveis erros devido à presença dos indels. A última etapa é realizada pelo *script HaplotypeCaller*, que faz o *calling* dos SNPs e Indels simultaneamente através de uma remontagem-local nas regiões ativas. Uma vez identificadas, as variantes foram anotadas funcionalmente, utilizando o software SnpEff v 3.06 [67].

3.8. Presença e ausência de genes exclusivos nas cepas probióticas e na cepa não probiótica S288c.

A presença e ausência de genes nas quatro cepas probióticas em comparação com a cepa S288c foi investigada a partir do alinhamento das ORFs entre as cepas, utilizando o algoritmo Blast [61] com formato de saída *default* e *evalue* $1e10^{-3}$. Os dados foram manipulados utilizando programas nativos Unix como awk, sort, uniq e grep. Uma vez identificados, os genes foram submetidos à busca por ontologia utilizando a ferramenta *GoTermFinder.pl* do *Saccharomyces Genome Database* disponível em http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl.

3.9. Presença de genes relacionados à resistência ao estresse ácido.

Investigamos a presença de genes relacionados à resistência da célula ao estresse ácido ao qual são submetidas as cepas probióticas logo em contato com o suco gástrico humano. Foram mapeados 81 genes da cepa S288c contra os genomas das cepas probióticas. Na tabela 1 estão relacionados os genes analisados.

	Nome do gene	Tamanho	Descrição
1	YBL105C_PKC1	3456 pb	Serina/treonina proteína quinase; essencial para a remodelação da parede celular durante o crescimento
2	YBR001C_NTH2	2343 pb	Trealase neutra putativa, necessária para termotolerância; pode mediar resistência a outros estresses celulares.
3	YBR066C_NRG2	663 pb	Repressor de transcrição; medeia a repressão de glicose e regula negativamente o crescimento filamentoso; ativado em pulsos aleatórios de localização nuclear em resposta a glicose baixa
4	YBR072W_HSP26	645 pb	Proteína de choque térmico, com atividade chaperona; possui atividade de ligação ao mRNA
5	YBR126C_TPS1	1488 pb	Subunidade sintase do complexo trealose-6-P sintase/fosfatase; Sua expressão é induzida pela resposta ao estresse e reprimido pela via de Ras-cAMP; os níveis de TPS1 aumentam em resposta ao estresse de replicação de DNA e em resposta a exposição ao ácido bórico.

6	YBR140C_IRA1	9279 pb	Proteína de ativação da proteína GTPasica; regulador inibitório da via RAS-cAMP
7	YBR160W_CDC28	897 pb	Ciclina dependente de quinase (CDK), subunidade catalítica; regulador mestre de mitose e meiose; aumenta a abundância em estresse de replicação de DNA;
8	YBR182C_SMP1	1359 pb	Fator de transcrição MADS-box envolvido na resposta ao estresse osmótico;
9	YBR260C_RGD1	2001 pb	Proteína (RhoGAP) para Rho3p e Rho4p-activação da GTPase; possivelmente envolvidos no controle da organização do citoesqueleto de actina
10	YBR295W_PCA1	3651 pb	Transporte de cádmio do tipo P ATPase; pode também ter um papel na homeostase de cobre e ferro;
11	YBR296C_PHO89	1725 pb	Co-transportador de Na+/Pi na membrana plasmática; ativo na fase inicial de crescimento
12	YCR021C_HSP30	999 pb	Proteína de choque térmico, regulador negativo da H (+) - ATPase Pma1p; proteína de resposta ao estresse;
13	YDL138W_RGT2	2292 pb	Regulador do transporte de glucose;
14	YDL185W_VMA1	3216 pb	ATPase vacuolar de membrana.
15	YDL194W_SNF3	2655 pb	Proteína de membrana sensor de glicose, regula o transporte; também detecta frutose e manose;
16	YDR001C_NTH1	2256 pb	Trealase, degrada trealose; necessário para termotolerância e pode mediar resistência a outros estresses celulares; pode ser fosforilada por Cdc28p; inibida por Dcs1p;
17	YDR028C_REG1	3045 pb	Subunidade reguladora da proteína fosfatase Glc7p tipo 1; envolvida na regulação negativa de genes reprimidos por glicose;
18	YDR038C_ENA5	3276 pb	Bomba de sódio ATPase do tipo P; envolvida no influxo de Na+ e Li+ para permitir tolerância ao sal;
19	YDR039C_ENA2	3276 pb	Bomba de sódio ATPase do tipo P; envolvida no influxo de Na+ e Li+ para permitir tolerância ao sal;
20	YDR040C_ENA1	3276 pb	Bomba de sódio ATPase do tipo P; envolvida no influxo de Na+ e Li+ para permitir tolerância ao sal;
21	YDR043C_NRG1	696 pb	Repressor de transcrição; recruta o complexo Cyc8p-Tup1p aos promotores; medeia a repressão de glicose e regula negativamente uma variedade de processos como resposta ao pH alcalino;
22	YDR074W_TPS2	2691 pb	Envolvida na síntese da trealose, armazenamento de carboidratos; expressão é induzida por condições de estresse e reprimido pela via de Ras-cAMP;
23	YDR171W_HSP42	1128 pb	Proteína de choque térmico, com atividade chaperona; envolvida na reorganização do citoesqueleto após choque térmico;
24	YDR173C_ARG82	1068 pb	Multiquinase inositol polifosfato (IPMK); tem também atividade difosfoinositol polifosfato sintase;
25	YDR216W_ADR1	3972 pb	Fator de transcricao do tipo Zinc finger responsivo a fontes de carbono. Requerido para a transcrição do gene ADH2 reprimido por glicose, de genes de proteínas peroxissomais e de genes requeridos para a utilização de etanol, glicerol e ácidos graxos
26	YDR258C_HSP78	2436 pb	Coopera com Ssc1p na termotolerância mitocondrial após choque térmico; capaz de prevenir a agregação de proteínas deformadas, bem como ressolubilizar agregados de proteínas;
27	YDR477W_SNF1	1902 pb	Serina/treonina proteína quinase ativada por AMP; necessário para a transcrição de genes reprimidos por glicose, tolerância térmica, esporulação e biogênese de peroxissomos; regula o crescimento filamentoso em resposta à fome;

28	YDR533C_HSP31	714 pb	Possível chaperona e cisteína protease; necessária para a reprogramação transcricional e para a sobrevivência em fase estacionária; semelhante a Hsp31p, Hsp32p, e Sno4p.
29	YEL011W_GLC3	2115 pb	Enzima ramificadora de glicogênio. Envolvida no acumulo de glicogênio.
30	YER129W_SAK1	3429 pb	Proteína do complexo SNF1serina/treonina quinase; desempenha papel no crescimento de pseudo-hifas
31	YFL014W_HSP12	330 pb	Envolvida na manutenção da organização durante condições de estresse; induzida por choque térmico, estresse oxidativo e osmótico, fase estacionária, depleção de glicose, oleato e álcool; regulada pelas vias HOG e Ras-Pka;
32	YFR014C_CMK1	1341 pb	Proteína quinase dependente de calmodulina; pode desempenhar um papel na resposta ao estresse;
33	YFR015C_GSY1	2127 pb	Glicogênio sintase; expressão induzida pela limitação de glicose, privação de nitrogênio, estresse ambiental e pela entrada em fase estacionária;
34	YGL006W_PMC1	3522 pb	ATPase vacuolar transportadora de Ca^{2+}; evita a inibição do crescimento por ativação da calcineurina na presença de concentrações elevadas de cálcio;
35	YGL008C_PMA1	2757 pb	H+-ATPase de membrana do tipo P2; bombeia prótons para fora da célula; principal regulador do pH citoplasmático e do potencial de membrana; Hsp30p desempenha um papel na regulação Pma1p
36	YGL035C_MIG1	1515 pb	Fator de transcrição envolvido na repressão da glicose; regula o crescimento filamentoso junto a Mig2p em resposta à depleção de glicose;
37	YGL071W_AFT1	2073 pb	Fator de transcrição envolvido na homeostase e utilização de ferro;
38	YGL179C_TOS3	1683 pb	Proteína quinase relacionada e funcionalmente redundante com Elm1p e Sak1p para a fosforilação e ativação de Snf1p;
39	YGL209W_MIG2	1149 pb	Proteína do tipo dedo de zinco repressora da transcrição; coopera com Mig1p na repressão do gene induzida por glicose;
40	YGL248W_PDE1	1110 pb	Controle de glicose e sinalização intracelular de AMPc induzida por acidificação;
41	YGR217W_CCH1	6120 pb	Canal de cálcio dependente de voltagem; envolvida no influxo de cálcio em resposta a pressões ambientais.
42	YHL027W_RIM101	1878 pb	Proteína dedo de zinco repressora da transcrição; envolvida na montagem da parede celular
43	YHR030C_SLT2	1455 pb	Serina/treonina MAP quinase envolvida na regulação da manutenção da integridade da parede celular, progressão do ciclo celular, e retenção de mRNA em choque térmico;
44	YIL050W_PCL7	858 pb	Forma um complexo quinase funcional com Pho85p envolvida no metabolismo do glicogênio.
45	YJL141C_YAK1	2424 pb	Serina/treonina proteína quinase; componente de um sistema de detecção de glicose que inibe o crescimento em resposta à baixa disponibilidade de nutriente;
46	YJL159W_HSP150	1242 pb	Proteína de choque térmico; segregada e ligada de forma covalente à parede da célula por meio de pontes de beta-1,3-glucano e dissulfureto; necessária para a estabilidade da parede celular;
47	YJL164C_TPK1	1194 pb	Subunidade catalítica da proteína quinase dependente de cAMP; promove o crescimento vegetativo em resposta a nutrientes através da via de sinalização de Ras-AMPc
48	YJR090C_GRR1	3456 pb	Atua como uma proteína ubiquitina-ligase direcionando a degradação dos substratos;

49	YKL048C_ELM1	1923 pb	Serina/treonina proteína quinase que regula a morfogênese celular; necessária para a regulação de outras quinases, tais como Kin4p;
50	YKL062W_MSN4	1893 pb	Ativador de transcrição em resposta ao estresse; ativado em resposta a diversas condições de estresse;
51	YKL166C_TPK3	1197 pb	Subunidade catalítica da proteína quinase dependente de cAMP; promove o crescimento vegetativo em resposta a nutrientes através da via de sinalização de Ras-AMPc
52	YKL190W_CNB1	604 pb	Subunidade reguladora da calcineurina, uma proteína fosfatase Ca++/calmodulina que regula Crz1p em condições de estresse;
53	YKR058W_GLG1	1851 pb	Glicogênio glicosiltransferase;
54	YLL026W_HSP104	2727 pb	Proteína de choque térmico que coopera com Hsp40 e Hsp70 para redobrar e reativar proteínas agregadas previamente desnaturadas; responsivo ao estresse;
55	YLR044C_PDC1	1692 pb	Maior das três isoenzimas piruvato descarboxilase; enzima chave na fermentação alcoólica; envolvida no catabolismo de aminoácidos;
56	YLR113W_HOG1	1308 pb	Proteína quinase envolvida na osmoregulação; controla a realocação global de RNAPII em condições de choque térmico;
57	YLR138W_NHA1	2958 pb	Na+/H+ Antiporter
58	YLR258W_GSY2	2118 pb	Glicogênio sintase; expressão induzida pela limitação de glicose, privação de nitrogênio, estresse ambiental e entrada em fase estacionária;
59	YLR259C_HSP60	1719 pb	Chaperonina mitocondrial tetradecamerica; previne a agregação e medeia o redobramento de proteínas após o choque térmico;
60	YLR310C_CDC25	4770 pb	Regula indiretamente adenilato ciclase através da ativação de Ras1p e Ras2p, estimulando a troca de GDP por GTP; envolvida no ciclo celular.
61	YLR332W_MID2	1131 pb	Proteína de membrana que atua como um sensor para a sinalização da integridade da parede celular;
62	YLR342W_FKS1	5631 pb	Envolvida na síntese e manutenção da parede celular;
63	YMR037C_MSN2	2115 pb	Ativador de transcrição em resposta a estresse; liga-se ao DNA em elementos de resposta ao estresse em genes responsivos;
64	YNL027W_CRZ1	2037 pb	Fator de transcrição que ativa genes de resposta ao estresse;
65	YNL098C_RAS2	969 pb	Proteína de ligação GTP; regula em resposta a privação de nitrogênio, esporulação e crescimento filamentoso;
66	YNL291C_MID1	1647 pb	Proteína de membrana; permite o influxo de Ca^{2+} estimulado por feromônio;
67	YOL016C_CMK2	1344 pb	Proteína quinase calmodulina-dependente; pode desempenhar um papel na resposta ao estresse;
68	YOL081W_IRA2	9240 pb	Proteína de ativação de GTPAse; regulador inibitório da via RAS-cAMP
69	YOR002W_ALG6	1635 pb	Alfa 1,3 glicosiltransferase; envolvida na transferência de oligossacarídeos durante a glicosilação de proteínas;
70	YOR008C_SLG1	1137 pb	Sensor de tradução ativado por estresse; envolvido na manutenção da integridade da parede celular e na organização do citoesqueleto;
71	YOR020C_HSP10	321 pb	Inibe a atividade de ATPase de Hsp60p, uma chaperonina mitocondrial; envolvida no dobramento de proteínas e de triagem nas mitocôndrias;
72	YOR087W_YVC1	2028 pb	Canal vacuolar de cátions; medeia a liberação de Ca^{2+} a partir do vacúolo em resposta a choque hiperosmótico;
73	YOR101W_RAS1	930 pb	Proteína de ligação GTP; regula em resposta a privação de nitrogênio, esporulação e crescimento filamentoso;

74	YOR178C_GAC1	2382 pb	Subunidade reguladora da transcrição de genes HSF-regulados em condições de estresse;
75	YOR391C_HSP33	714 pb	Possível chaperona e cisteína protease; necessária para a reprogramação transcricional e para a sobrevivência em fase estacionária; semelhante a Hsp31p, Hsp32p, e Sno4p.
76	YPL203W_TPK2	1143 pb	Proteína quinase dependente de cAMP; promove o crescimento vegetativo em resposta a nutrientes através da via de sinalização de Ras-AMPc
77	YPL240C_HSP82	2130 pb	Necessária para a sinalização de feromônio e regulação negativa de Hsf1p;
78	YPL280W_HSP32	714 pb	Possível chaperona e cisteína protease; necessária para a reprogramação transcricional e para a sobrevivência em fase estacionária; semelhante a Hsp31p, Hsp32p, e Sno4p.
79	YPR026W_ATH1	3636 pb	Trealase ácida necessária para a utilização de trealose extracelular; envolvida na degradação da trealose intracelular durante a retomada do crescimento após o estresse salino;
80	YPR160W_GPH1	2709 pb	Glicogênio fosforilase necessária para a mobilização de glicogênio; regulada por fosforilação mediada por AMPc;
81	YPR184W_GDB1	4611 pb	Possui atividades glicanotransferase e alfa-1,6-amiloglucosidase; necessária para a degradação do glicogênio; promove o aumento da abundância de proteínas em resposta ao estresse de replicação de DNA.

*Tabela 1**:** Relação dos genes envolvidos à resistência ao estresse ácido. Estão descritos o nome sistemático do gene separado por um underline (_) do nome da proteína, seguido do tamanho e descrição do mesmo.*

4. RESULTADOS.

4.1. Montagem *de novo, scaffolding* e ordenação dos *contigs*

O sequenciamento do genoma da cepa UFMG A-905 realizado com o Illumina MiSeq gerou 16.322.453 de *reads* de até 151 pb com valor de qualidade acima de *phred* 30 e cobertura estimada em 203 vezes. Para a montagem *de novo*, o melhor valor de *k-mer* utilizado pelo SOAPdenovo foi 81, gerando 2026 *contigs* que foram então estendidos em 838 *scaffolds*, que foram então ordenados por mapeamento no genoma de referência de S288c, resultando em 16 *superscaffolds* (mapeados na referência) e 194 *scaffolds* (não mapeados na referência) compreendendo 11.430,366 pb, tendo o maior *scaffold* 1.434.165 pb, tamanho médio de 54.430pb, valor de N_{50} igual a 882 kb, 38,2% de c.nteúdo GC, 324 *gaps* e 98,79% de completude. Para depósito do genoma no *Genbank*, os *superscaffolds* foram quebrados nos Ns, gerando 534 *contigs*, que foram depositados com o arquivo *.agp*, que contém as coordenadas dos *contigs* nos *scaffolds*, sob o número de acesso GCA_000733235.3.

O sequenciamento do genoma da cepa Sb_17 foi realizado em duas plataformas NGS, SOLiD 4 e Illumina MiSeq. Um total de 81,4 milhões de *reads* de 50 pb foram obtidas com o SOLiD 4. Utilizando o *script SOLiD_preprocess_filter_v2.pl* foram excluídas do conjunto de dados aproximadamente 60% das leituras por critérios de qualidade, resultando em 35,8 milhões de leituras com qualidade acima de *phred* 20 e com a cobertura estimada em 148 vezes. A montagem *de novo* realizada com o Velvet resultou em 968 *contigs*. O sequenciamento com o MiSeq gerou 17,3 milhões de *reads* de até 151 pb com valor de qualidade acima de *phred* 30 e com a cobertura estimada em 215 vezes. A montagem *de novo* realizada com o SOAPdenovo resultou em 2064 *contigs*. Uma montagem híbrida combinando os *contigs-velvet* e os *contigs-soap* foi realizada com o software Zorro.pl resultando em 719

contigs, que foram estendidos em 106 *scaffolds* e então ordenados mapeando contra o genoma de referência de S288c, resultando em 16 *superscaffolds* (mapeados na referência) e 83 *contigs* (não mapeados na referência) compreendendo 11.677.572 pb com o maior *scaffold* 1.496.914 pb, valor de N_{50} igual a 880 kb, 38,2% de conteúdo GC, 395 *gaps* e 100% de completude. Para depósito do genoma no *Genbank*, os *superscaffolds* foram quebrados nos Ns, gerando 442 *contigs*, que foram depositados com o arquivo *.agp,* que contém as coordenadas dos *contigs* nos *scaffolds*, sob o número de acesso GCA_000734875.3.

O sequenciamento da cepa Sb_ATCC MYA-796 foi conduzido pelo Dr. Bruno Douradinha (Center of Vaccine Research – University of Pittsburgh, USA) e as *reads* nos foram cedidas para montagem do genoma. Foi utilizado o Illumina HiSeq, que gerou 53,4 milhões de leituras pareadas de 101 pb. Após filtro de qualidade, foram retiradas do conjunto de dados aproximadamente 9,6% das leituras por critério de qualidade, resultando em 48,3 milhões de leituras com valor de qualidade acima de *phred* 30. A montagem *de novo* resultou em 1042 *contigs*, que foram estendidos em 331 *scaffolds* e ordenados mapeando contra o genoma de S288c, resultando em 16 *superscaffolds* (mapeados na referência) e 93 *contigs* (não mapeados na referência) compreendendo 11.420.333 pb com o maior *scaffold* 1.455.881 pb, valor de N_{50} igual a 882 kb, 38.1% de conteúdo GC, 90 *gaps* e 100% de completude. Para depósito do genoma no *Genbank*, os *superscaffolds* foram quebrados nos Ns, gerando 193 *contigs*, que foram depositados com o arquivo *.agp,* que contém as coordenadas dos *contigs* nos *scaffolds*, sob o número de acesso GCA_000769245.1.

Em nossas análises utilizamos o genoma da cepa Sb_EDRL [68] sequenciado com o GS 454 FLX+ e publicado em 2013 por um grupo de pesquisadores indianos com a cobertura estimada em 50 vezes. Os *contigs* foram obtidos do *Genbank* (número de acesso ATCS00000000.1) e então ordenados contra o genoma de S288c, resultando em 16 *superscaffolds* (mapeados na referência) e 93 *contigs* (não mapeados na referência),

compreendendo 11.482.966 pb, com o maior *scaffold* 1.456.401 pb, valor de N_{50} igual a 895 kb, 38,3% de conteúdo GC, 2 *gaps* e 100% de completude.

As métricas estatísticas da montagem e ordenação dos *scaffolds* das cepas probióticas estão sumarizados na tabela 2. Os *scaffolds* de maior tamanho em todas as quatro cepas correspondem ao cromossomo IV de *S. cerevisiae*. A quantidade de *contigs* não mapeados no genoma de referência foi pequena em todas as quatro cepas, e a soma destes *contigs* varia entre 125kb e 247kb.

	UFMG A-905	**Sb_17**	**Sb_ATCC**	**Sb_EDRL***
Tecnologia NGS	MiSeq	Solid e MiSeq	HiSeq	454
Cobertura	~203x	~366x	~403x	~50x
Contigs	534	442	193	194
Superscaffolds	16	16	16	16
Contigs não mapeados	194	83	91	91
Soma dos *contigs* não mapeados	247.651 pb	232.531 pb	144.594 pb	124.923 pb
Soma total	11.430.366 pb	11.677.572 pb	11.406.924 pb	11.482.966 pb
Maior *contig*	1.434.165 pb	1.496.914 pb	1.455.881 pb	1.456.401 pb
Menor *contig*	505 pb	524 pb	491 pb	509 pb
Média de tamanho dos *contigs*	54.430 pb	117.955 pb	107.612 pb	107.317 pb
N50	882.501 pb	880.009 pb	882.821 pb	895.096 pb
Contigs N50	6	6	6	6
GC%	38,2	38,2	38,1	38,3
Gaps (N)	324	395	88	2
Completude	98,79%	100%	100%	100%

*Tabela 2: Métricas estatísticas das montagens dos genomas. *O genoma da cepa Sb_EDRL não foi montado neste trabalho, apenas analisado junto aos demais.*

A figura 1 ilustra os *superscaffolds* de UFMG A-905, Sb_17, Sb_ATCC e Sb_EDRL mapeados nos cromossomos I a VI de S288c. O percentual de similaridade entre as sequências é representado de acordo com a transparência da cor referente à cepa alinhada.

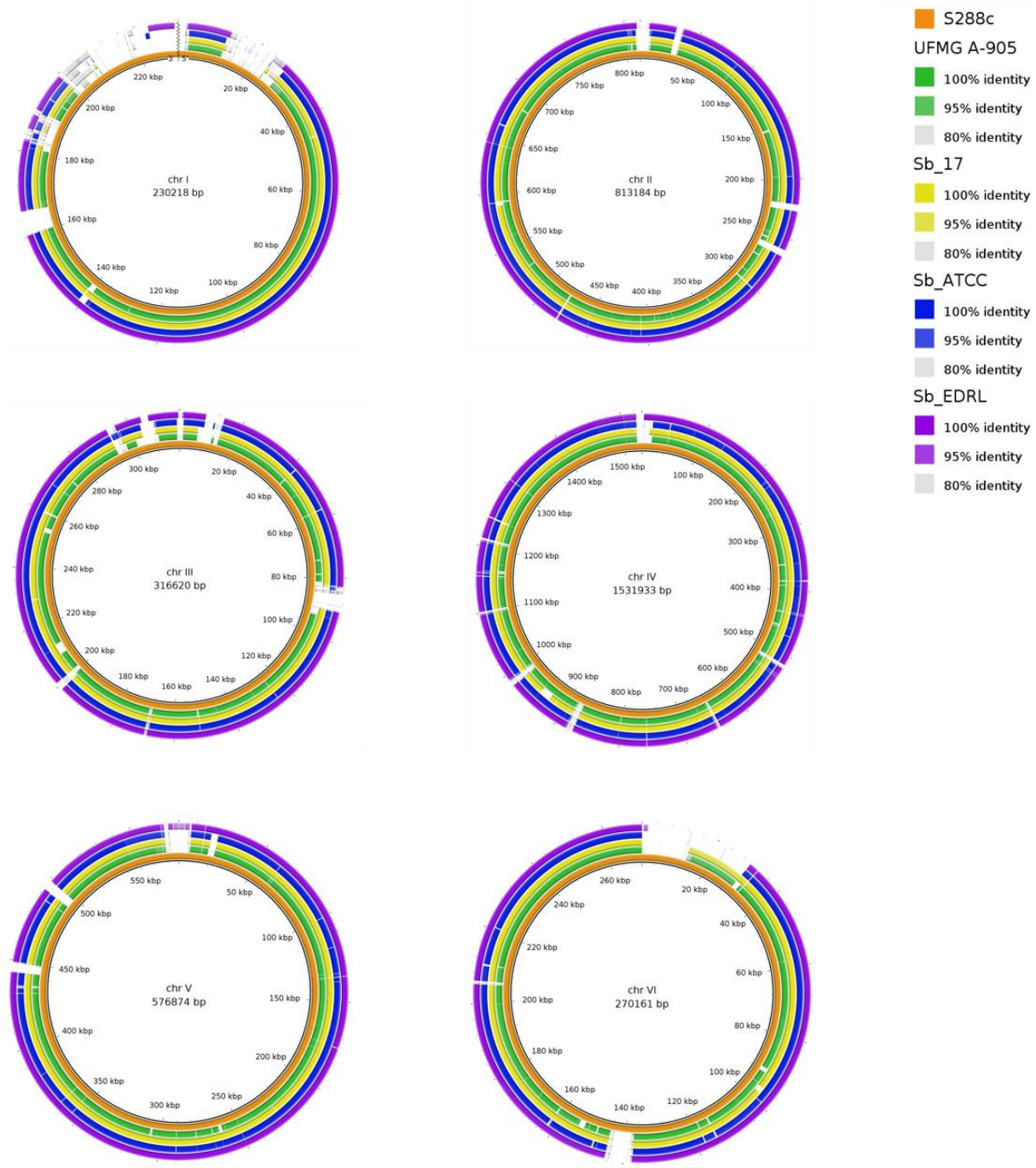

Figura 1: Brig plot dos superscaffolds das cepas probióticas mapeados nos cromossomos I a VI de S288c, representada em laranja. Em verde está representada a cepa UFMG A-905, em amarelo a cepa Sb_17, em azul a cepa Sb_ATCC e em roxo a cepa Sb_EDRL.

A figura 2 ilustra os *superscaffolds* de UFMG A-905, Sb_17, Sb_ATCC e Sb_EDRL mapeados nos cromossomos VII a XII de S288c. O percentual de similaridade entre as sequências é representado de acordo com a transparência da cor referente à cepa alinhada.

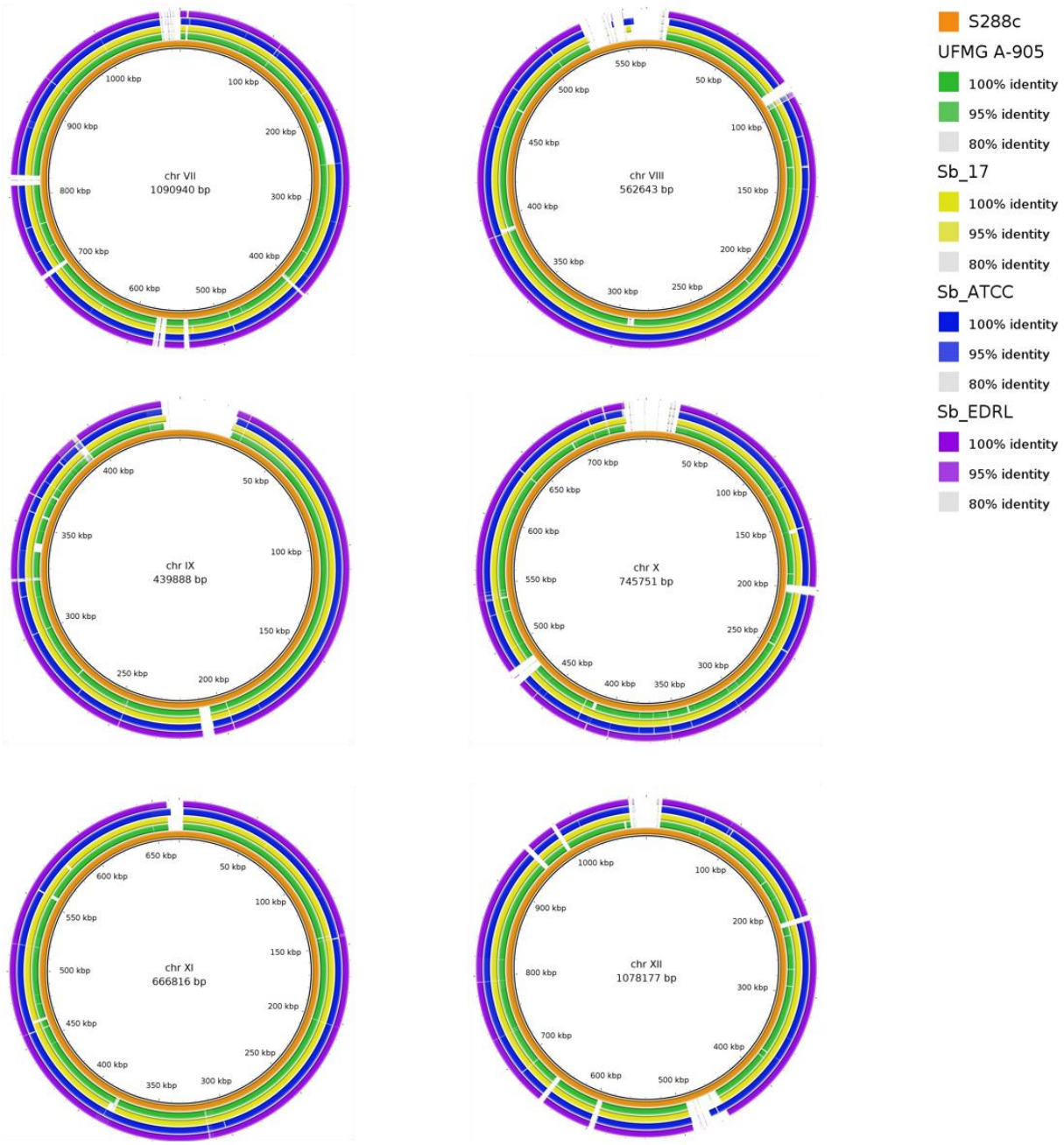

Figura 2: Brig plot dos superscaffolds das cepas probióticas mapeados nos cromossomos VII a XII de S288c, representada em laranja. Em verde está representada a cepa UFMG A-905, em amarelo a cepa Sb_17, em azul a cepa Sb_ATCC e em roxo a cepa Sb_EDRL.

A figura 3 ilustra os *superscaffolds* de UFMG A-905, Sb_17, Sb_ATCC e Sb_EDRL mapeados nos cromossomos XIII a XVI de S288c. O percentual de similaridade entre as sequências é representado de acordo com a transparência da cor referente à cepa alinhada.

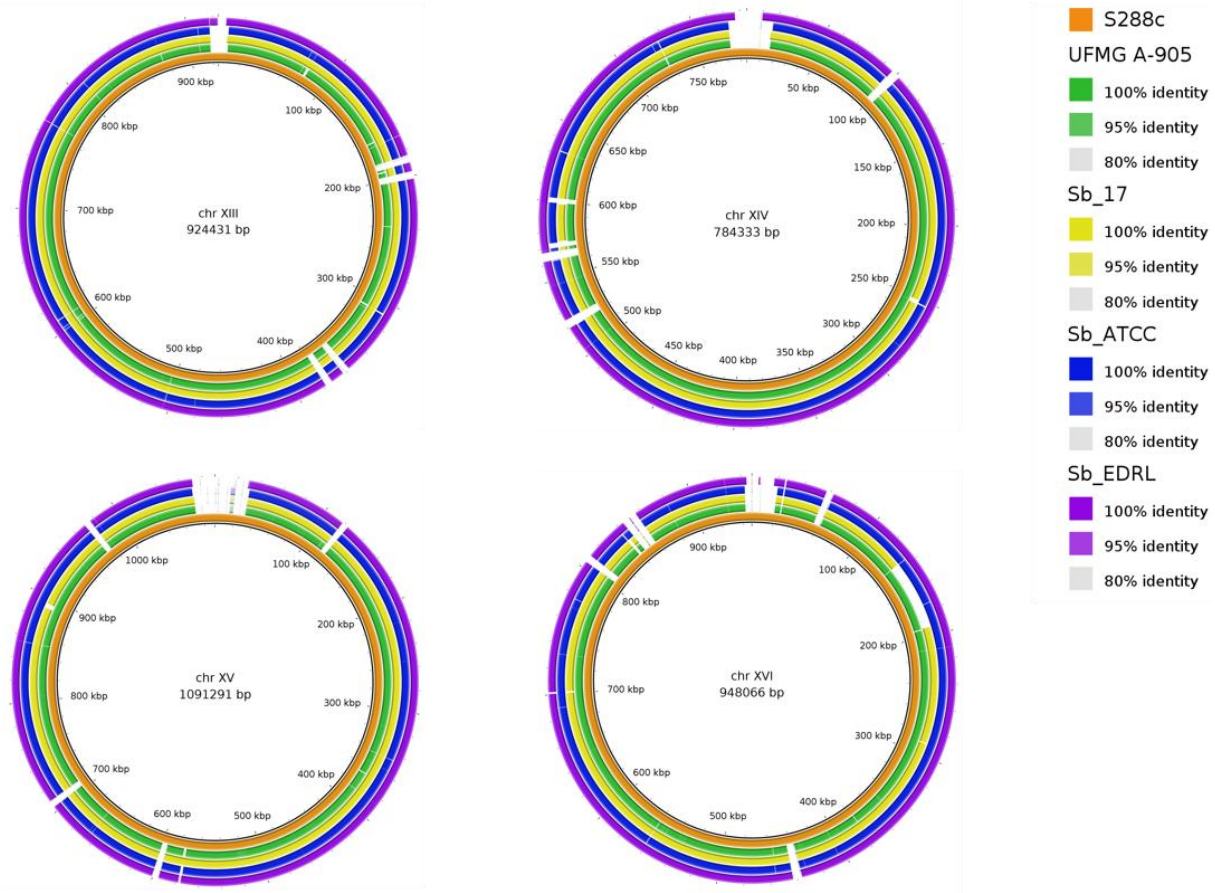

Figura 3: Brig plot dos superscaffolds das cepas probióticas mapeados nos cromossomos XIII a XVI de S288c, representada em laranja. Em verde está representada a cepa UFMG A-905, em amarelo a cepa Sb_17, em azul a cepa Sb_ATCC e em roxo a cepa Sb_EDRL.

4.2. Anotação estrutural dos genomas

A predição gênica identificou possíveis ORFs (*Open Reading Frames*) nos genomas, que foram mapeadas contra o banco de dados de proteínas não redundantes – nr, do NCBI (*National Center of Biothecnology Information*). Aproximadamente 98,4% de todas as ORFs preditas em todas as quatro cepas revelaram similaridade significante (*evalue cutoff* $\leq 10^{-3}$) com proteínas do nr. A tabela 3 sumariza os resultados obtidos de predição gênica e tRNAs presentes nos genomas das cepas probióticas. A diversidade de tRNAs está relacionada ao anticodon que o tRNA carrega.

	UFMG A-905	Sb_17	Sb_ATCC	Sb_EDRL
ORFs preditas	5335	5441	5305	5326
ORFs mapeadas x nr	5252	5357	5221	5233
% ORFs mapeadas	98,44%	98,45%	98,41%	98,40%
ttRNAs	311	287	273	276
dtRNAs	41	42	43	41

Tabela 3: ORFs e tRNAs preditos nos genomas das cepas probióticas. ttRNAs = total de tRNAs; dtRNAs = diversidade de tRNAs.

O resultado da busca por elementos repetitivos nos genomas probióticos, bem como no genoma da cepa não-probiótica S288c estão sumarizados na tabela 4. As cepas probióticas apresentaram uma redução em aproximadamente 75% de bases em regiões repetitivas do tipo LTR, embora a quantidade de elementos identificados tenha sido próxima da cepa referência, variando entre 5% na cepa Sb_17 e 16% na cepa Sb_ATCC. Elementos do tipo Ty1/copia são os mais abundantes em leveduras e por conseguinte, foram os que apresentaram maior redução na quantidade de bases identificadas nas cepas probióticas, embora a contagem do número de elementos do tipo Ty3/GIPSY/DIRS1 apresentou discrepância (maior contagem) quando comparada à cepa referência. Isso acontece devido à comparação das sequências de interesse com os elementos repetitivos do banco de dados interno do RepeatMasker (neste caso, elementos repetitivos presentes no gênero *Saccharomyces*), as inserções e deleções que fragmentaram essas regiões foram contadas como um elemento. As repetições simples, são compostas por regiões de di, tri, tetra, penta e hexa repetições maiores do que 20pb, e apresentavam contagem e frequência semelhantes entre as cepas.

Pequenos pseudogenes de RNAs também foram mapeados pelo RepeatMasker, com contagens semelhantes nas cepas UFMG A-905 e Sb_17, aproximadamente 10% da contagem encontrada na cepa S288c, ao passo que as cepas Sb_ATCC e Sb_EDRL apresentaram aproximadamente 50% e 70%, respectivamente, da quantidade de bases de pequenos RNAs encontrada na cepa referência.

Cepas analisadas	Tipos de elementos	Quantidade de elementos	Tamanho dos elementos	Percentual no genoma
S288c (referência)	LTRs	514	406714 pb	3,35 %
	Ty1/Copia	452	381547 pb	3,14 %
	Gypsy/DIRS1	62	25167 pb	0,21 %
	Não classificado	19	50995 pb	0,42 %
	Pequenos RNAs	6	12034 pb	0,10 %
	Repetições simples	2976	127844 pb	1,05 %
UFMG A-905	LTRs	456	104078 pb	0,91 %
	Ty1/Copia	386	88506 pb	0,77 %
	Gypsy/DIRS1	70	15572 pb	0,14 %
	Não classificado	6	4371 pb	0,04 %
	Pequenos RNAs	3	1329 pb	0,01 %
	Repetições simples	2912	121006 pb	1,06 %
Sb_17	LTRs	488	112859 pb	0,97 %
	Ty1/Copia	411	95312 pb	0,82 %
	Gypsy/DIRS1	77	17547 pb	0,15 %
	Não classificado	2	4992 pb	0,04 %
	Pequenos RNAs	4	1167 pb	0,01 %
	Repetições simples	2900	121205 pb	1,04 %
Sb_ATCC	LTRs	433	102336 pb	0,90 %
	Ty1/Copia	366	86707 pb	0,76 %
	Gypsy/DIRS1	67	15629 pb	0,14 %
	Não classificado	3	3575 pb	0,03 %
	Pequenos RNAs	6	5959 pb	0,05 %
	Repetições simples	2823	131486 pb	1,15 %
Sb_EDRL	LTRs	468	116602 pb	1,02 %
	Ty1/Copia	393	98312 pb	0,86 %
	Gypsy/DIRS1	75	18290 pb	0,16 %
	Não classificado	6	7101 pb	0,06 %
	Pequenos RNAs	14	8015 pb	0,07 %
	Repetições simples	2735	116595 pb	1,02 %

Tabela 4: Resultado do mapeamento de elementos repetitivos identificados nas cinco cepas analisadas.

4.3. Anotação funcional dos genomas

A anotação funcional com Blast2GO foi capaz de identificar categorias funcionais de ontologia em 69% das ORFs preditas (aproximadamente 3700 ORFs) nos quatro genomas probióticos. A busca por domínios proteicos realizada pelo Interproscan identificou assinaturas proteicas em 93% das ORFs preditas (aproximadamente 5000 ORFs) nos quatro genomas probióticos (figura 4).

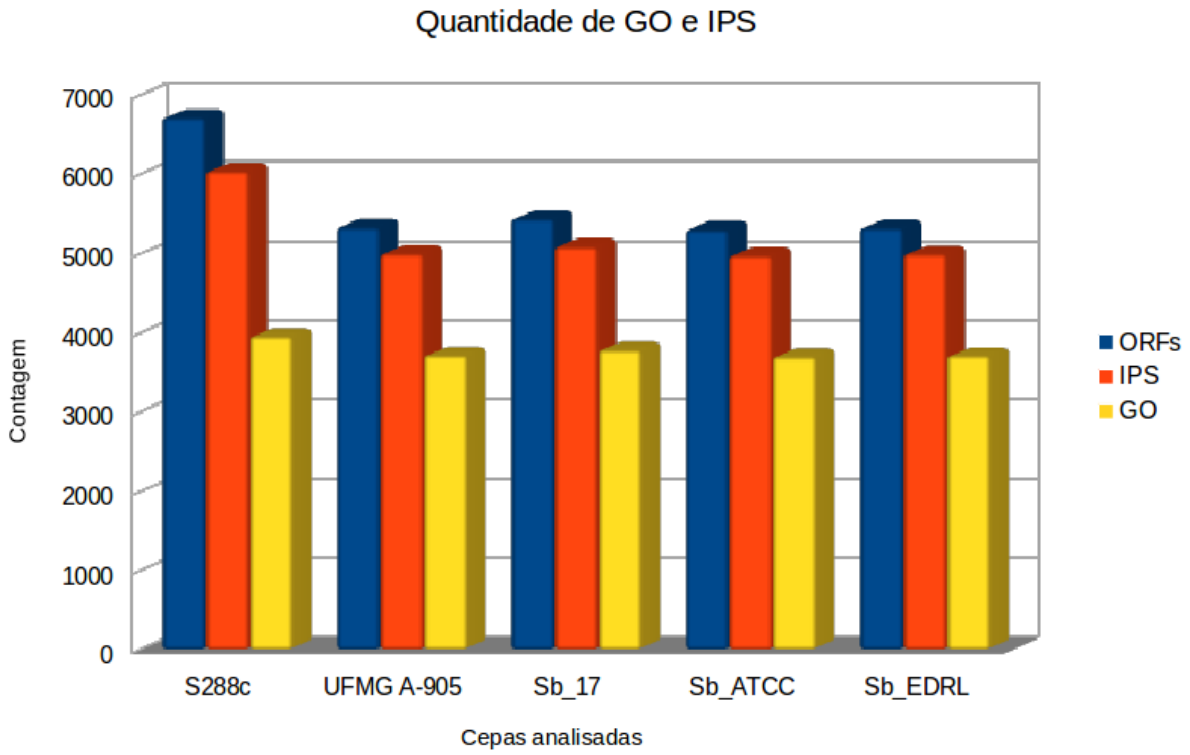

Figura 4: Quantidade de GO e IPS mapeados nos genomas. Em azul está representado o total de ORFs em cada genoma, em laranja estão representadas as assinaturas proteicas identificadas no Interproscan e em amarelo a quantidade de GOs identificados.

A distribuição de *Enzime Code* (EC) das seis principais classes (Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases e Ligases) foi semelhante nos quatro genomas probióticos e quando comparado com o genoma de S288c (figura 5).

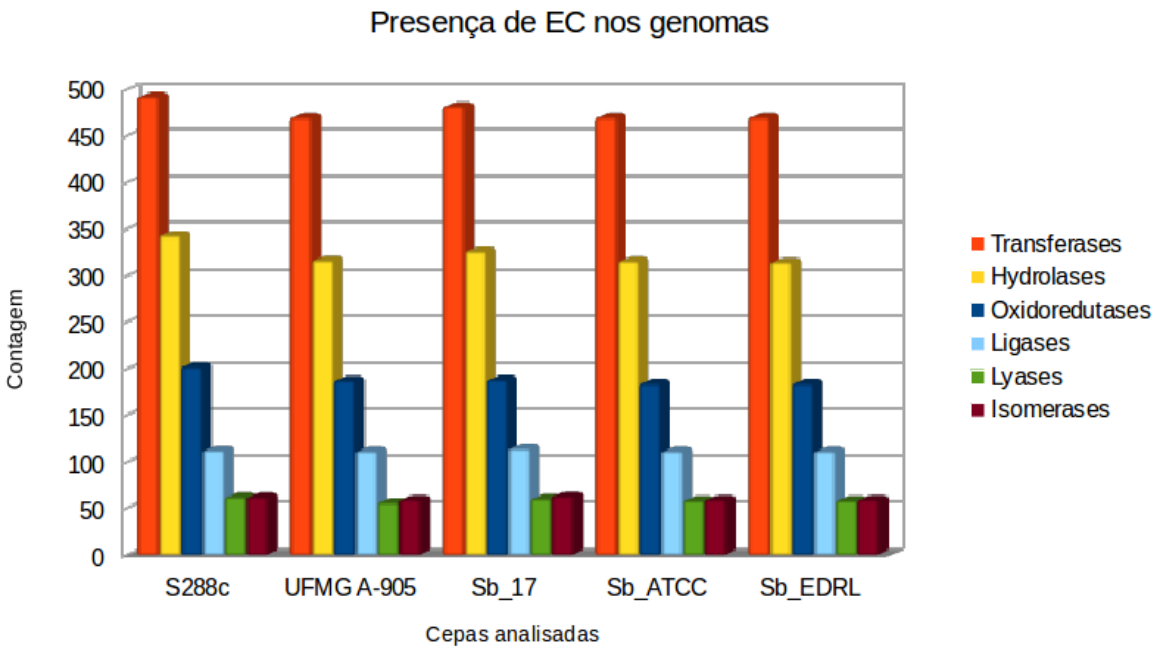

Figura 5: Contagem de EC (Enzime Code) presente nos genomas analisados.

A distribuição de GO por níveis de categorias em Processos Biológicos (PB), Componente Celular (CC) e Função Molecular (MF) foi semelhante entre as cepas probióticas e a cepa não probiótica (S288c), com exceção do GO 0005515 *(Protein Binding)* que se mostrou enriquecido na cepa Sb_ATCC quando realizado o teste *FischerExact* com múltiplos testes de correção de FDR (*cutoff* ≤ 0.05). Foram encontrados 396 genes relacionados à categoria *Protein Binding* na cepa Sb_ATCC enquanto a cepa referência e demais cepas probióticas apresentaram aproximadamente 96 genes relacionados à categoria acima.

GO term	Nome	Tipo	FDR	Single test p-Value	# Sb_ATCC	#S288c	Acima/abaixo
0005515	Protein binding	F	$1.2e^{-29}$	$1.6e^{-32}$	326	96	Acima

4.4. Relações filogenéticas

A história evolutiva recuperada a partir do alinhamento das 415 proteínas ortólogas concatenadas está representada na figura 6. Como raiz foi utilizado o genoma da levedura *Saccharomyces pastorianus* que foi agrupada separadamente das demais. Distinguimos três grandes clados, destacados por diferentes cores e que refletem a importância biotecnológica das cepas. No clado vermelho, com suporte estatístico de 100%, estão agrupadas as leveduras de importância industrial, produtoras de bioetanol Sce_IR2, Sce_NAM34-4c, Sce_NY1308, Sce_ZTW1, Sce_M3837, Sce_M3707, Sce_M3836, Sce_M3838, Sce_YJSH1, Sce_M3839, a cepa Sce_T7 isolada de um Carvalho (árvore) no estado de Missouri nos EUA, e duas cepas produtoras de vinho e saque, respectivamente, Sce_N85 e Sce_Kyokay7. A levedura Sce_YJM789, uma cepa oportunista patógena humana isolada de um paciente imunodeprimido em 1989, foi agrupada em um ramo, com suporte estatístico de 70%, entre o clado de importância industrial e o clado das cepas laboratoriais (destacado em azul), contendo as cepas Sce_Sigma1278b, Sce_CEN.PK113-7D, Sce_W303, Sce_BY4741 e S288c, com 100% de suporte estatístico. Logo abaixo foi formado um ramo, com 63% de suporte estatístico, com a cepa produtora de bioetanol Sce_JAY291, seguido do clado das cepas fermentadoras alcoólicas (destacado em verde) produtoras de vinho Sce_R008, Sce_R103, Sce_P301, Sce_P283, Sce_EC1118, Sce_RM11-1a, Sce_VL3, Sce_VIN13, Sce_AWRI796 e LalvinQa23, e as cepas produtoras de cerveja Ale Sce_FostersB e Sce_FostersO. Dentro deste clado, destaca-se o ramo em verde-limão, com alto suporte estatístico, onde estão agrupadas as cepas probióticas Sb_17, UFMGA-905, Sb_EDRL e Sb_ATCC.

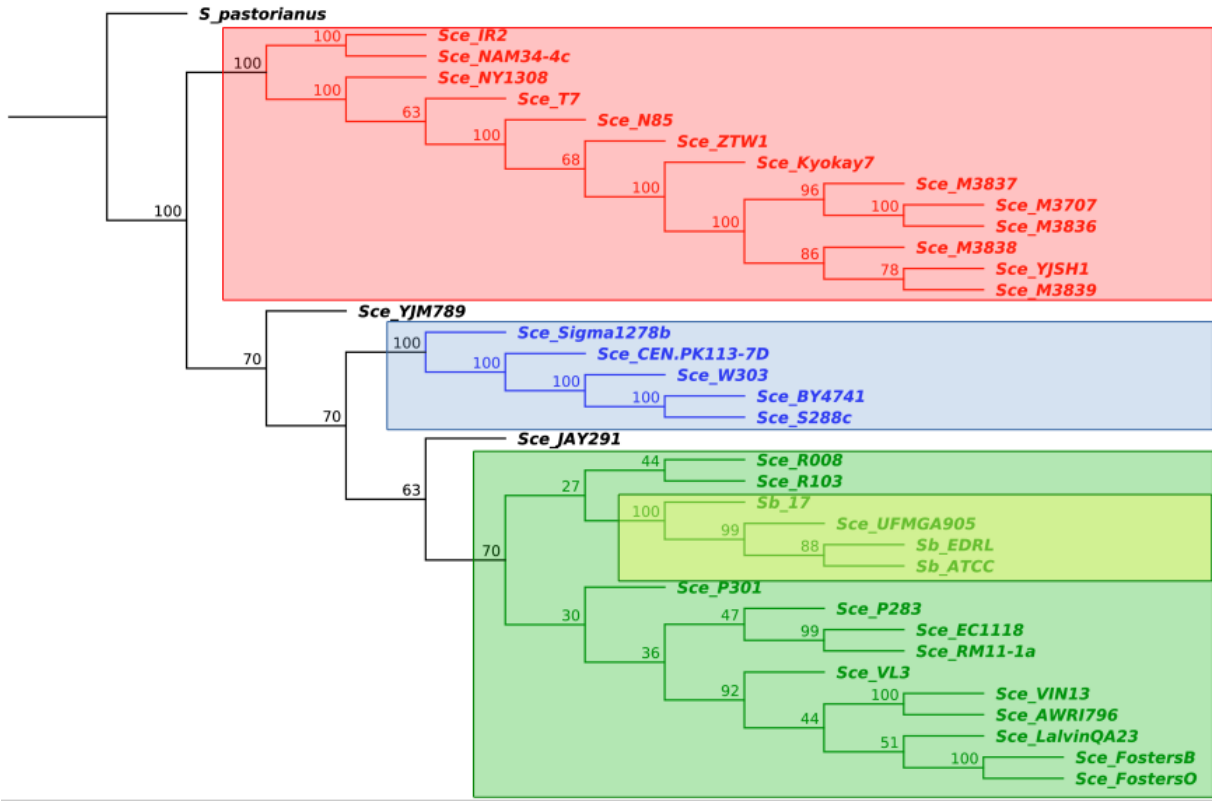

Figura 6: Árvore filogenética construída a partir do alinhamento de 415 proteínas ortólogas. A história evolutiva foi inferida usando o método Neighbour Joining com suporte estatístico de 1000 replicatas de bootstrap. O percentual de árvores em que as cepas se agruparam é mostrado nos nós e a análise foi conduzida utilizando o software Geneious. Em vermelho esta destacado o clado das cepas de importância industrial, em azul está destacado o clado das cepas laboratoriais e em verde está destacado o clado das cepas fermentadoras, com destaque para o ramo em verde-limão onde estão agrupadas as cepas probióticas.

4.5. Análise de variantes (SNVs)

A análise de variantes no genótipo das quatro cepas probióticas foi realizada a partir do alinhamento das *reads* contra o genoma de referência de *S. cerevisiae.* As cepas UFMG905, Sb_17 e Sb_ATCC foram sequenciadas com Illumina gerando *reads* curtas (101pb – 151pb) com alta cobertura (>300x), enquanto a cepa Sb_EDRL foi sequenciada utilizando 454 FLX+, gerando *reads* longas (~800pb) com cobertura estimada em 50x. Nas quatro cepas probióticas, foi encontrada uma variante a cada 245, 237, 230 e 222 bases, respectivamente para UFMG A-905, Sb_17, Sb_ATCC e Sb_EDRL, com média de 51,943

variantes em cada genoma. Na tabela 5 estão descritas as contagens de variantes encontradas nas quatro cepas. A quantidade de variantes homozigotas e heterozigotas foi semelhante entre as cepas, embora Sb_EDRL (sequenciada com 454 FLX+) tenha apresentado uma contagem elevada (em negrito) de inserções em apenas um alelo em relação às demais cepas, bem como uma elevação na contagem total de deleções (entre 25 e 32%).

	UFMG A-905			**Sb_17**			**Sb_ATCC**			**Sb_EDRL**		
	Hom	**Het**	**Total**	**Hom**	**Het**	**Total**	**Hom**	**Het**	**Total**	**Hom**	**Het**	**Total**
SNP	42.394	3.231	45.625	42.667	3.408	46.075	43.424	3.623	47.047	44.098	3.885	47.983
Inserção	1.911	388	2.299	1.928	390	2.318	1.917	663	2.580	1.402	**1.457**	2.859
Deleção	2.059	336	2.395	2.086	332	2.418	2.086	561	2.647	2.704	822	**3.526**
Total	46.394	3.955	50.319	46.681	4.130	50.811	47.427	4.847	52.274	48.204	6.164	54.368

Tabela 5: Quantidade de variantes por tipo nos quatro genomas probióticos.

O SnpEff reporta os possíveis impactos que uma variante causará ao fenótipo produzido e os classifica em quatro grupos, sendo HIGH (provavelmente causa uma proteína truncada, ou a perda de função, ou alguma troca que acarretará no decaimento do transcrito, exemplo, ganho de *stop* codon e mudança de *frameshift*), MODERATE (variantes sem interrupções que podem alterar a eficiência da proteína, exemplo, variante *missense* e deleção *inframe*), LOW (assume que a maior parte das variantes sejam inofensivas ou que seja improvável alguma alteração na proteína, exemplo, variantes sinônimas) e MODIFIER (variantes não codificadoras, ou em regiões de genes não codificadores de proteínas, ou quando não há evidência de impacto, por exemplo, variantes em regiões *downstream* ao gene). Quando localizadas em regiões codificadoras dos genes as variantes podem causar algum efeito funcional na proteína a ser traduzida. Esse efeito pode ser chamado *Missense* quando a alteração do nucleotídeo afetará o códon e consequentemente o aminoácido traduzido. É chamado *Nonsense*, quando a alteração do nucleotídeo acarretará em um

stopcodon prematuro. É chamado de mutação silenciosa, quando a alteração no nucleotídeo, embora altere o codon, não alterará o aminoácido traduzido. As variantes encontradas e anotadas com SnpEff nos quatro genomas probióticos estão representadas na figura 7. A quantidade de impactos nos quatro níveis nas quatro cepas foi semelhante, embora a cepa Sb_EDRL tenha apresentado contagem pouco acima das demais cepas.

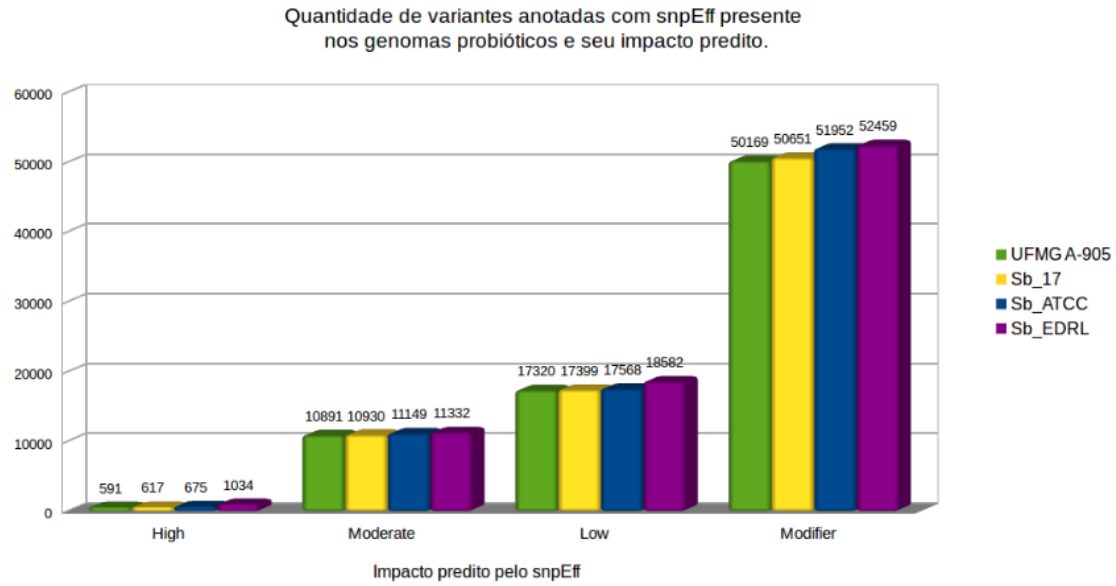

Figura 7: Quantidade de variantes encontradas nos quatro níveis de impacto nas quatro cepas probióticas, preditos pelo SnpEff.

Embora a contagem total dos impactos tenha sido semelhante entre as quatro cepas, nem todas as variantes foram encontradas em todos os genomas; houve uma diferença na quantidade de impactos presentes em comum em todas as quatro cepas, ou seja, mapeados no mesmo cromossomo e na mesma posição em relação à referência. A figura 8 representa em um gráfico de pizza, sendo as fatias correspondentes aos tipos de impactados de variantes encontrados em comum nas cepas probióticas, e que tenham apresentado valor de qualidade do mapeamento igual ou superior a *phred* 30.

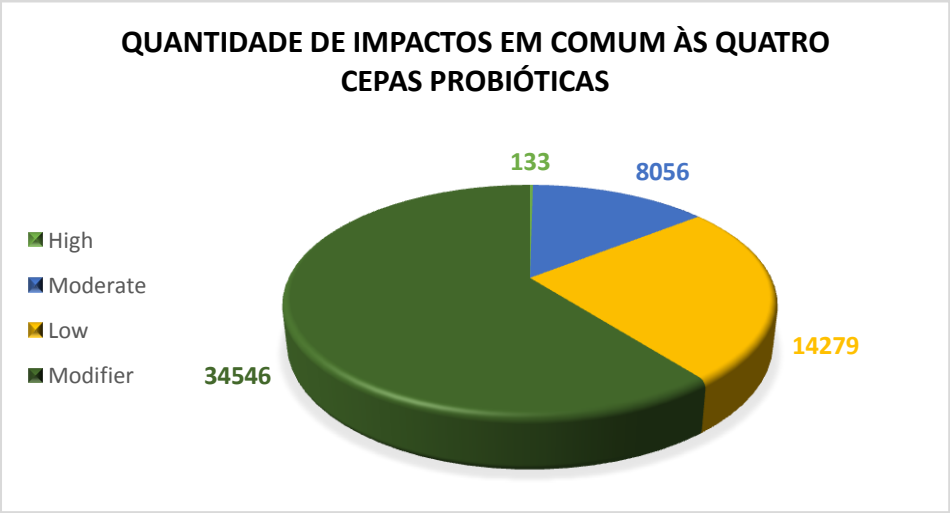

Figura 8: Quantidade de impactos de variantes anotados com SnpEff encontrados em comum aos quatro genomas probióticos.

Dos 133 impactos com anotação do tipo High, 11 resultam em *frameshift*, 1 impacto em um sitio aceptor de splicing, 16 causam a perda de *start* códon, 76 resultam em ganho de *stop* códon e 29 resultam na perda do *stop* códon. Dos impactos do tipo High, 111 genes são afetados em uma posição em comum nas quatro cepas probióticas, 8 genes são afetados duas vezes na mesma posição nas quatro cepas probióticas (HPC2, IMD4, RPL16A, YAL067W-A, YER097W, YHL037C, YHR028W-A e YOL079W) e 2 genes sofreram três impactos do tipo High em comum nas cepas probióticas (NSP1 e OSW2). A relação completa dos genes e impactos do tipo High estão descritos no Anexo.

As cepas sequenciadas com alta cobertura apresentaram valores semelhantes de variantes causadoras de mutação *missense*, *nonsense* e silenciosa, enquanto a cepa Sb_EDRL apresentou contagem um pouco acima das demais para mutações *missense* e silenciosa. Quando a alteração de uma única base acontece de uma base purina para outra purina (A/G), ou de uma base pirimidina para outra pirimidina (C/T), essa mutação é conhecida como transição. Quando essa alteração acontece de uma base purina para uma pirimidina (e vice-versa), ela é conhecida como transversão. A quantidade de transições e transversões, bem como a relação transições/transversões encontradas foi semelhante entre as quatro cepas. Na

tabela 6 estão descritas as contagens de mutações por classe funcional que afetam as regiões gênicas, bem como a contagem de transições e transversões e sua relação.

	UFMG A-905	**Sb_17**	**Sb_ATCC**	**Sb_EDRL**
Missense	10.586	10.622	10.815	11.121
Nonsense	109	110	117	116
Silenciosa	17.320	17.398	17.567	18.579
Transições	33.529	33.826	34.442	34.968
Transversões	12.096	12.249	12.605	13.015
Relação Trans/Transv	2,7719	2,7615	2,7324	2,6867

Tabela 6: Quantidade de efeitos por classe funcional e quantidade de transições e transversões presentes nos genomas probióticos.

A densidade de variantes ao longo dos cromossomos das 4 cepas probióticas foi semelhante entre as quatro cepas, embora a cepa Sb_EDRL tenha apresentado valores acima das demais em algumas regiões. As figuras 9 e 10 representam a densidade de variações encontradas ao longo de 1kb ou 10kb nos 16 cromossomos das cepas probióticas.

Figura 9: Densidade de SNVs ao longo dos cromossomos I a VIII das cepas probióticas. Em verde está representado a cepa UFMG A-905, em amarelo está representada a cepa Sb_17, em azul está representada a cepa Sb_ATCC e em rosa está representada a cepa Sb_EDRL.

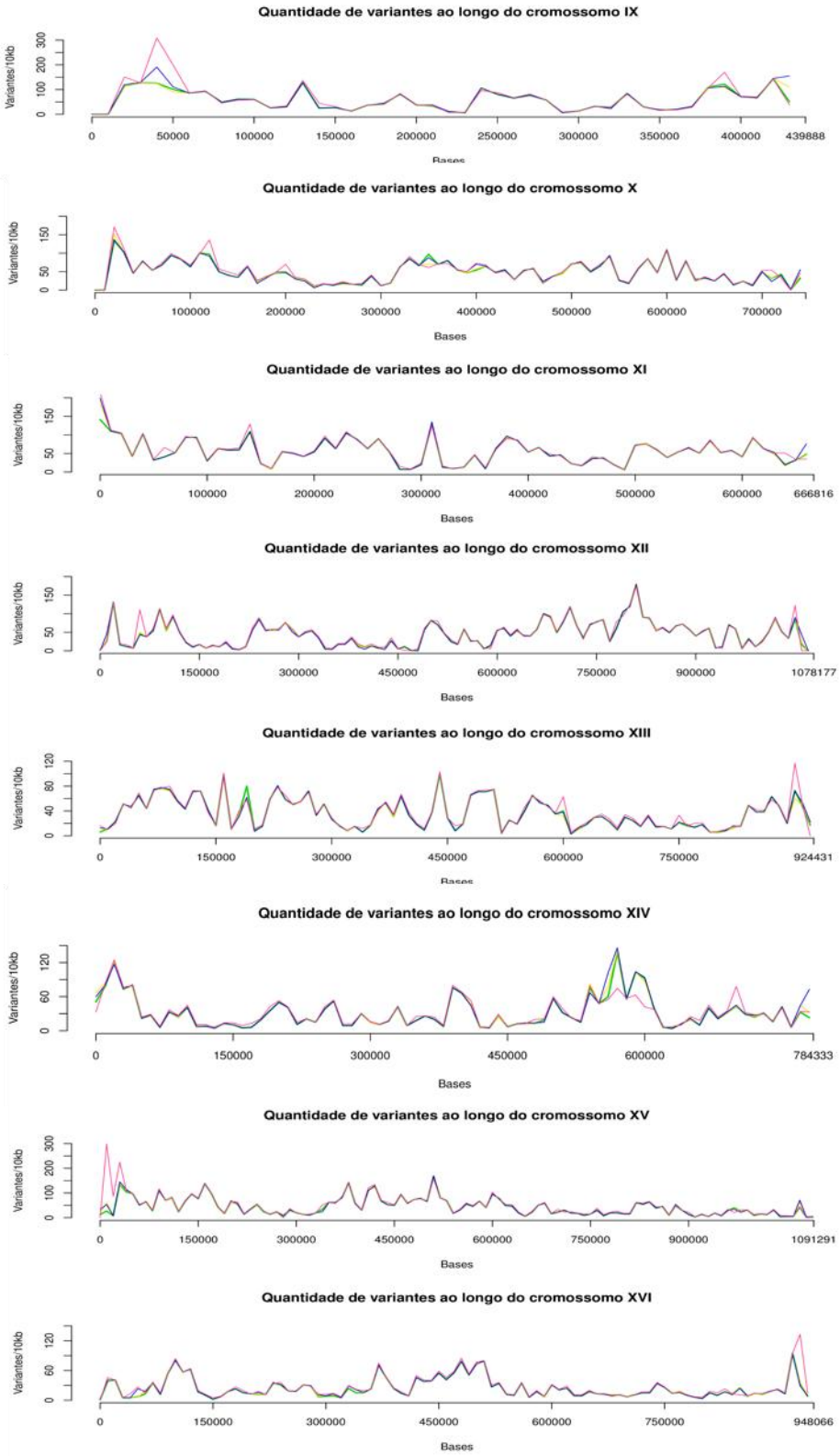

Figura 10: Densidade de SNVs ao longo dos cromossomos IX a XVI das cepas probióticas. Em verde está representado a cepa UFMG A-905, em amarelo está representada a cepa Sb_17, em azul está representada a cepa Sb_ATCC e em rosa está representada a cepa Sb_EDRL.

4.6. Genes presentes nas cepas probióticas e ausentes na cepa S288c

A busca por genes presentes nas cepas probióticas e ausentes na cepa referência não probiótica (S288c) retornou 20 candidatos, sendo que 4 destes genes foram encontrados duplicados (destacados com * na tabela 7) em, pelo menos, uma das quatro cepas. Apenas dois dos vinte genes exclusivos não foram encontrados em uma das cepas probióticas, yil169c-like protein (de aproximadamente 450 pb) não encontrada na cepa Sb_EDRL e um gene de 216 pb que codifica uma proteína hipotética não foi encontrado na cepa Sb_ATCC. Dentre os vinte genes, foi encontrado um gene exclusivo da cepa *S. cerevisiae* EC1118 (de 237 pb), um gene da cepa *S. cerevisiae* CENPK113-7D (de 225 pb) e um gene da cepa *S. cerevisiae* JAY291 (de 1263 pb), ambos codificando para proteínas hipotéticas. Os genes com as respectivas anotações e tamanhos estão descriminados na tabela 7.

Descrição da anotação via GO		**Tamanho do gene**			
		UFMG A-905	**Sb_17**	**Sb_ATCC**	**Sb_EDRL**
1	ykl215c-like protein	3864 pb	3864 pb	3864 pb	3864 pb
2	Proteína hipotética	183 pb	183 pb	*183 pb	183 pb
3	EC1118_1J19_0562p	237 pb	237 pb	237 pb	237 pb
4	yil169c-like protein	450 pb	498 pb	432 pb	-
5	N-acetiltransferase mpr1 envolvida na tolerância ao estresse oxidativo via metabolismo de prolina	690 pb	690 pb	690 pb	690 pb
6	yfr012w-like protein	666 pb	447 pb	666 pb	666 pb
7	Proteína hipotética	*231 pb	231 pb	231 pb	*231 pb
8	Proteína hipotética	*189 pb	189 pb	*189 pb	*189 pb
9	Proteína hipotética CENPK1137D_4762	225 pb	225 pb	225 pb	225 pb
10	Proteína hipotética	195 pb	195 pb	195 pb	195 pb
11	Proteína hipotética	192 pb	*192 pb	*192 pb	*192 pb

12	Proteína hipotética	216 pb	216 pb	-	216 pb
13	Proteína hipotética C1Q_05653 (Sce_JAY291)	1263 pb	1263 pb	1263 pb	1263 pb
14	Floculação	1182 pb	1182 pb	1182 pb	1182 pb
15	Transportador membranar de ácido nicotínico	1512 pb	1512 pb	1512 pb	1512 pb
16	Fator de transcrição C6	1962 pb	1962 pb	1962 pb	1962 pb
17	Proteína hipotética	246 pb	246 pb	246 pb	246 pb
18	Proteína hipotética	219 pb	*219 pb	219 pb	219 pb
19	Proteína hipotética	198 pb	198 pb	198 pb	198 pb
20	Proteína hipotética	204 pb	204 pb	204 pb	204 pb

Tabela 7: Genes presentes nas cepas probióticas e ausentes na cepa S288c. Os genes duplicados nas respectivas cepas estão indicados com asterisco (*).

4.7. Genes ausentes nas cepas probióticas e presentes na cepa S288c.

Os 6717 genes anotados da cepa S288c foram mapeados contra as CDS preditas e anotadas das cepas probióticas. Foram identificados 803 genes ausentes em, pelo menos, uma das cepas probióticas e 706 ausentes nas quatro cepas, como visualizado na figura 11, criada com a ferramenta Venny (http://bioinfogp.cnb.csic.es/tools/venny/).

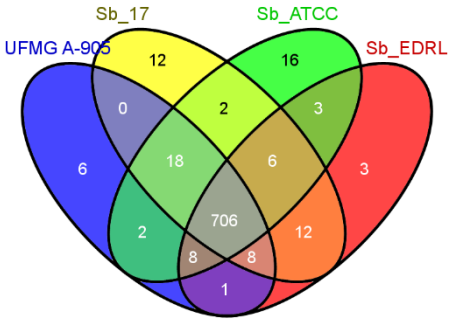

Figura 11: Diagrama de Venn ilustrando a quantidade de genes da cepa S288c ausentes nas cepas probióticas.

Para o mapeamento funcional, foram retirados deste grupo de 706 genes ausentes nas cepas probióticas, 24 genes codificados pelo genoma mitocondrial, pois nosso interesse a priori, é apenas em genes nucleares. Analisamos então, 682 genes nucleares e identificamos os Processos Biológicos, Funções Moleculares e Componentes Celular aos quais estes genes estão relacionados, como descrito na tabela 8. O gene YIL080W foi excluído dos resultados pois não apresenta categoria de GO descrita, e se trata de um pseudogene TyB Gag-Pol deletado.

	Termo GO	#Prob.	%	#Background	%	P-value	FDR (%)	Expect. Falso positivo
PB	Transposição mediada por RNA	94/681	13,8	110/7167	1,5	$1,81e^{-78}$	0.00	0.00
	Transposição	94/681	13,8	113/7167	1,6	$3,39e^{-76}$	0.00	0.00
FM	Atividade DNA Polimerase dirigida por RNA	46/681	6,8	49/7167	0,7	$4,00e^{-42}$	0.00	0.00
	Atividade DNA Polimerase dirigia por DNA	46/681	6,8	62/7167	0,9	$1,81e^{-32}$	0.00	0.00
	Atividade DNA Polimerase	46/681	6,8	65/7167	0,9	$6,21e^{-31}$	0.00	0.00
	Atividade Ribonuclease	47/681	6,9	84/7167	1,3	$5,64e^{-22}$	0.00	0.00
	Nucleotidil transferase	49/681	7,2	94/7167	1,3	$1,41e^{-17}$	0.00	0.00
	Atividade Peptidase	47/681	6,9	140/7167	2,0	$2,28e^{-13}$	0.00	0.00
	Atividade Nuclease	47/681	6,9	142/7167	2,0	$4,26e^{-13}$	0.00	0.00
	Atividade Hidrolase	49/681	7,2	289/7167	4,0	0,00441	0.01	0.04
CC	Retrotransposon nucleocapsideo	94/681	13,8	94/7167	1,3	$2.71e^{-97}$	0.00	0.00

Tabela 8: Termos de ontologia (GO) dos genes de S288c ausentes nas quatro cepas probióticas. Estão descritos os Processos Biológicos (PB), Função Molecular (FM) e Componente Celular (CC) encontrados com diferença significativa em comparação com os termos GO dos genomas das leveduras depositados no Saccharomyces Genome Database.

Destacaram-se como enriquecidos os processos relacionados à atividade de transposição e retro transposição com 94 dos 681 genes ausentes nas cepas probióticas em cada um destes processos, bem como genes relacionados à atividade transcricional mediados

por DNA e RNA e enzimas com atividade peptidase, nuclease, hidrolase e nucleotidiltransferase.

4.8. Genes relacionados à resistência ao estresse ácido.

Genes relacionados à resistência ao estresse ácido pelo qual as leveduras probióticas são submetidas em contato com o suco gástrico humano foram investigados nas quatro cepas probióticas. Foram analisados genes de choque térmico, fatores de transcrição, proteínas quinases, ATPases, proteínas de membrana, chaperonas, entre outras (a relação completa está descrita na tabela 1). Consideramos o tamanho da *query*, tamanho da *subject* e tamanho do alinhamento para inferir sobre a presença ou ausência dos genes nas cepas probióticas. Do total de 81 genes analisados, 54 foram mapeados nas quatro cepas probióticas, 8 apresentaram evidência de alinhamento em ao menos três cepas probióticas e 19 não apresentaram evidência de alinhamento ou há divergência entre o tamanho da *query*, da *subject* e do alinhamento. Estes 19 genes estão descritos na tabela 9. No anexo a relação de todos os genes e o resultado do blast.

	Nome do gene	Classe funcional
1	YDL185W_VMA1	Bomba ATPase
2	YDR039C_ENA2	Bomba ATPase
3	YDR040C_ENA1	Bomba ATPase
4	YKL190W_CNB1	Caucineurina
5	YDR258C_HSP78	Choque Térmico
6	YJL159W_HSP150	Choque Térmico
7	YPL240C_HSP82	Choque térmico
8	YJR090C_GRR1	Envolvimento em ubiquitinação
9	YBR066C_NRG2	Envolvimento na transcrição
10	YNL027W_CRZ1	Envolvimento na transcrição
11	YLR310C_CDC25	Envolvimento no ciclo celular
12	YOR178C_GAC1	Fosfatase
13	YNL098C_RAS2	GTPase
14	YOR101W_RAS1	GTPase
15	YBL105C_PKC1	Proteína quinase

16	YJL164C_TPK1	Proteína quinase
17	YKL166C_TPK3	Proteína quinase
18	YPL203W_TPK2	Proteína quinase
19	YLR332W_MID2	Sensor de sinalização

Tabela 9: Genes relacionados à resistência ao estresse ácido sem evidência de alinhamento nos genomas das cepas probióticas.

5. DISCUSSÃO.

Probióticos vêm sendo alvo de crescente interesse científico nos últimos anos. Em uma busca realizada no Pubmed (http://www.ncbi.nlm.nih.gov/pmc/) em novembro de 2014, o termo "probiotic" retornou 6.322 artigos, sendo 778 relacionados ao termo "*Saccharomyces*" e destes, 477 artigos foram relacionados ao termo "*Saccharomyces boulardii*". Embora o conhecimento sobre os efeitos probióticos causados por essa levedura tenha aumentado nos últimos anos, pouco se sabe sobre o genoma de *S. boulardii*. A levedura *S. cerevisiae* foi o primeiro microrganismo eucarioto a ter o genoma sequenciado em 1996, abrindo uma gama de possibilidades para pesquisas genéticas, permitindo a clonagem de seus genes, caracterização e identificação, além da compreensão dos mecanismos celulares e fisiológicos deste micro-organismo [69]. Com o advento das novas tecnologias de sequenciamento aliada ao processamento de dados através de análises computacionais avançadas, podemos explorar fronteiras ainda desconhecidas no genoma destes microrganismos, sequenciando com alta cobertura e em tempo muito menor, a um custo cada vez mais reduzido. O primeiro genoma de *S. boulardii* depositado em banco de dados públicos foi a cepa EDRL [68] de origem indiana, em 2013, cujo o foco do trabalho foi a análise de um grupo de proteases e fosfatases presentes no referido genoma e em cepas de *S. cerevisiae*.

Nosso trabalho propôs, além de sequenciar utilizando avançadas plataformas de sequenciamento de última geração, montar, anotar o genoma de duas cepas probióticas, uma isolada no Brasil (UFMG A-905) e outra comercializada no país (Sb_17), analisar de forma ampla incluindo o genoma das cepas Sb_ATCC MYA-796 (mantida nos EUA) e Sb_EDRL (comercializada na indiana), buscando encontrar assinaturas genômicas em comum ao grupo de leveduras probióticas em comparação a cepa não probiótica S288c.

A primeira corrida de sequenciamento das leveduras UFMG A-905 e Sb_17 foi realizada utilizando o sequenciador SOLiD 4, com bibliotecas *mate-pair*. Devido a problemas de contaminação bacteriana, os dados de UFMG A-905 foram descartados. A montagem *de novo* da cepa Sb_17 não foi satisfatória, sendo necessária uma nova corrida de sequenciamento que foi realizada utilizando o sequenciador Illumina MiSeq, com bibliotecas *paired-end*. Foram gerados dados de alta qualidade, que foram montados *de novo*. Devido a falta de softwares específicos capazes de realizar uma montagem *denovo* combinando as *reads* de SOLiD e MiSeq da cepa Sb_17, uma montagem híbrida combinando os *contigs* obtidos com cada uma das montagens foi a alternativa viável para aproveitamento dos dois dados de sequenciamento desta cepa. A montagem *de novo* da cepa UFMG A-905, sequenciada apenas em uma plataforma (MiSeq), gerou métricas estatísticas satisfatórias. As *reads* geradas com o sequenciador MiSeq foram utilizadas para gerar os *scaffolds*, que foram posteriormente alinhados utilizando o genoma da cepa S288c como referência, gerando 16 *superscaffolds* que representam os cromossomos da levedura.

A cepa Sb_ATCC foi sequenciada utilizando Illumina HiSeq em um trabalho prévio [70] de um grupo do Centro de Vacinas da Universidade de Pittsburgh, que gentilmente nos cedeu os dados brutos (*reads*) da referida cepa para incluirmos em nosso trabalho. Os dados foram então submetidos ao mesmo *pipeline* de montagem que utilizamos para as cepas UFMG A-905 e Sb_17: montagem de novo, *scaffolding* utilizando a informação de pareamento das *reads*, e alinhamento no genoma de referência (S288c).

O genoma das três cepas, UFMG A-905, Sb_17 e Sb_ATCC foram disponibilizados publicamente no Genbank como montagem a nível cromossomal.

Os *contigs* da cepa Sb_EDRL, sequenciada com GS 454 FLX+ e disponibilizados publicamente no Genbank, foram obtidos e submetidos à última etapa do pipeline de montagem (alinhamento contra o genoma de referência), a fim de obtermos os 16 *superscaffolds*.

A sintenia encontrada ao alinharmos os 16 cromossomos das quatro cepas probióticas aos cromossomos da cepa referência S288c reflete uma conservação a nível genômico presente nas cepas probióticas, com ausência de partes do genoma em regiões semelhantes, embora a origem dos dados montados de cada cepa tenha sido diferente.

O *pipeline* de anotação foi semelhante para os quatro genomas, diferenciando pelo uso de dados de transcritos montados a partir de dados de RNA-Seq das cepas UFMG A-905 e Sb_17. A quantidade de ORFs preditas semelhante nas quatro cepas e a alta quantidade de ORFs mapeadas contra o nr (98,4%) sugerem uma montagem consistente, uma excelente eficiência do pipeline de anotação e a não necessidade de utilização de dados de RNA-Seq na anotação de genomas de leveduras, uma vez que os preditores *ab initio* do pipeline de anotação foram semelhantemente efetivos na anotação dos quatro genomas.

A quantidade de genes de tRNA encontrados na cepa UFMG A-905 foi maior do que nas três cepas de *S. boulardii* e do que na cepa não-probiótica S288c (que possui 295 genes de tRNA), embora a diversidade de genes de tRNA não tenha sido refletida no total de genes. Satapathy e colaboradores, em um trabalho envolvendo a análise comparativa de 441 estirpes bacterianas, mostraram um aumento da diversidade de tRNA em bactérias termófilas (temperatura de crescimento acima de 50º C) quando comparadas a bactérias não-termófilas, ao passo que bactérias psicrófilas (temperatura de crescimento abaixo de 25º C) apresentam aumento no total de tRNA em comparação a bactérias não-psicrófilas, sugerindo que a adaptação à temperatura de crescimento afeta essas características genômicas em bactérias[71]. Não existem na literatura trabalhos descrevendo a diversidade e quantidade de tRNAs em diferentes leveduras correlacionando com a temperatura de crescimento ou qualquer outra característica fisiológica/metabólica. Nossos dados sugerem que o aumento no ttRNA encontrado na cepa UFMG A-905 pode estar correlacionado à necessidade de

adaptação do microrganismo às condições extremas encontradas em seu ambiente de origem (mosto da produção de Cachaça).

Elementos transponíveis (TEs) são seguimentos de DNA repetitivo que podem ser inseridos em novas localizações cromossômicas e que muitas vezes podem fazer cópias duplicadas de si mesmo durante o processo[72]. Compõe uma das principais fontes de variabilidade genética em procariotos e eucariotos. Em bactérias, os TEs são responsáveis pela transferência de genes causadores de resistência a antibióticos[73]. O elemento transponível do tipo P é encontrado exclusivamente em *Drosophilamelanogaster*, e amplamente utilizado para estudos de mutagênese e na criação de moscas geneticamente modificadas para estudos genéticos[74]. Em leveduras, a única classe de TEs encontrada são os elementos do tipo LTR retrotransposons. O resultado da busca por TEs nos genomas analisados, sugere uma perda conservativa de LTRs nos genomas probióticos quando comparados ao genoma da cepa não-probiótica S288c. A redução de aproximadamente 75% de bases em regiões de LTRs foi comparável nas cepas probióticas, embora a quantidade de números de elementos tenha sido próxima à da cepa S288c; isso devido a fragmentação destes elementos repetitivos por inserções ou deleções que são contados como um elemento pelo algoritmo do RepeatMasker. Em 2002, Dunham e colaboradores, trabalhando com análises de CGH (*Comparative Genomic Hybridization*) envolvendo oito cepas de leveduras em experimentos sob condições de limitação de glicose, mostraram que a maioria dos rearranjos cromossomais estão relacionados a pontos de quebram em sequências de TEs[44]. Embora o estudo não ofereça evidências diretas de que os rearranjos promovam algum *fitness,* o fato de múltiplas cepas aprestarem o mesmo ponto de quebra, apesar de suas origens independentes, reforça a sugestão de que os pontos de rearranjos funcionem como base para adaptação. Três das oito cepas tiveram pontos de quebra em regiões de Ty próximas ao gene CIT1, que codifica a enzima citrato sintase, uma importante enzima do ciclo do ácido cítrico. Os autores

especulam que o rearranjo pode ativar os elementos Ty levando a desrepressão do CIT1 na presença de glicose. Como é um regulador chave do ciclo do ácido cítrico, a ativação de CIT1 pode promover a desrepressão de outros genes do clico e resultar no fenótipo adaptativo[44]. Os autores sugerem ainda que sequências de elementos Ty possam se tornar fixas em rearranjos, por oferecem vantagens fenotípicas seletivas.

Bleykasten-Grosshans e colaboradores publicaram um trabalho em 2013 envolvendo a análise comparativa de 41 genomas de leveduras e mostraram a existência de consideráveis diferenças entre as cepas em relação aos elementos Ty. Embora não tenham encontrado um padrão na quantidade de conteúdo de LTRs, existe um viés para um maior teor de LTR em cepas laboratoriais quando comparados às demais cepas[75]. Uma vez que nossas análises identificaram quebras cromossomais (figuras 1 a 3) e perdas semelhantes no conteúdo de LTRs entre as cepas probióticas (tabela 4), sugerimos que essa conservação deletéria possa estar relacionada, de alguma maneira, ao *fitness* probiótico.

A anotação funcional via Blast2GO identificou categorias funcionais de ontologia e assinaturas de domínios proteicos equivalentes nas cepas probióticas, bem com a quantidade de Enzime Code entre as cepas probióticas e quando comparadas à cepa S288c, sugerindo uma similaridade global a nível proteico/funcional. A análise de enriquecimento utilizando teste *FischerExact* identificou uma categoria funcional, *Binding Protein* (GO 0005515) enriquecida apenas na cepa Sb_ATCC em comparação com a cepa S288c. Nemcová e colaboradores já mostraram que a habilidade de ligação à matriz extracelular é fundamental para habilitar bactérias a colonizarem o intestino e exercerem uma função probiótica[76].

Há muita controvérsia em relação à classificação taxonômica de *S. boulardii* dentro do gênero *Saccharomyces*. Por meio de diferentes metodologias de análises moleculares, há pesquisadores que classificam *S. boulardii* como uma espécie distinta de *S. cerevisiae* [77, 78], outros que a classificam como um grupo separado dentro de espécies de *S. cerevisiae*

[79], outros que a classificam como uma espécie geneticamente muito próxima ou quase idêntica à *S. cerevisiae*, embora apresentem diferenças metabolicamente e fisiologicamente [80]. A abordagem filogenética utilizada em nosso estudo, concatenando e alinhando 415 proteínas ortólogas presentes em 37 genomas de espécies de *Saccharomyces* (incluindo as cepas probióticas aqui descritas e o *outgroup S. pastorianus*) foi capaz de agrupar as cepas de acordo com características metabólicas/biotecnológicas em três grandes clados, das cepas de importância industrial/biotecnológica (figura 6, em vermelho), o clado das cepas laboratoriais (figura 6, em azul), e o clado das cepas fermentadoras alcoólicas (figura 6, em verde), onde estão agrupas em um ramo monofilético, as cepas probióticas (figura 6, em amarelo). A abordagem de múltiplas proteínas concatenadas produz um aumento da acurácia da inferência filogenética quando utilizado o método *Neighbour-Joining* [81]. Ao alinharmos grupos de proteínas ortólogas para construir uma mega-árvore, estamos garantindo uma relação simétrica entre os genomas, mesmo não descartando posições incongruentes, uma vez que elas fornecem informações adicionais para a resolução de ramos curtos [81]. Sendo assim, sugerimos que *S. boulardii* é uma cepa de *S. cerevisiae* e não uma subespécie de *Saccharomyces*.

A análise de variantes presentes nos genomas probióticos confirmou o carácter diploide destas cepas. Os genomas sequenciados com Solid/Illumina (UFMG A-905, Sb_17 e Sb_ATCC) tiveram contagem semelhante na quantidade de variantes, enquanto o genoma da cepa Sb_EDRL, sequenciada com 454 FLX+ apresentou contagem um pouco maior que as demais, sobretudo de inserções, devido à natureza da tecnologia empregada neste sequenciador, que apresenta alta taxa de erros em homopolímeros [34]. A semelhança na quantidade de impactos preditos entre as cepas, inclusive com a presença de 54% dos impactos preditos em comum às quatro cepas probióticas, sugere uma conservação genômica em nível de variantes nos genomas probióticos.

A duplicação gênica é uma importante fonte de novos genes e exerce papel crucial influenciando na evolução molecular [82]. Essa duplicação pode acontecer em duas escalas: a nível genômico, duplicando grandes blocos do genoma e a nível gênico, com pequenas duplicações gênicas ou de pequenos segmentos do genoma [83]. A presença de um grupo de 20 genes (em sua maioria, hipotéticos sem função conhecida) ausentes na cepa não probiótica S288c e presentes nas quatro cepas probióticas sugere um papel importante no caráter probiótico ao qual pertencem estas cepas, sobretudo pela presença do gene N-acetyltransferase Mpr1, responsável pela proteção celular ao estresse oxidativo, via metabolismo de L-prolina e L-arginina [84]. Quatro destes genes, todos hipotéticos sem função conhecida, estão duplicados nos genomas probióticos. O gene transportador membranar de ácido nicotínico, também presente nos genomas probióticos e ausente em S288c, possui domínio MSF_1 que pertence à superfamília dos facilitadores de membrana, e atuam principalmente na absorção de açúcares, mas também no transporte de drogas, metabólitos, oligossacarídeos e aminoácidos [85], podendo estar envolvido no caráter probiótico destas cepas. Técnicas de biologia molecular podem ser aplicadas para inferir a função destes genes, como expressão da proteína recombinante seguida de ensaios *in vitro*, nocaute gênico seguido da avaliação do fenótipo e localização celular por hibridização *in situ*. Dos 803 genes de S288c ausentes em, pelo menos, uma cepa probiótica, 706 (88%) destes genes estão ausentes nas quatro cepas, com grande presença de genes envolvidos em processos de transposição.

A ausência de genes relacionados à resposta ao estresse ácido, como ATPase de membrana e bombas de sódio (VMA1, ENA1, ENA2), bem como proteínas quinase (PKC1 TPK3, TPK2) podem estar envolvidos indiretamente na maior resistência das cepas probióticas quando expostas às condições de estresse induzido por baixo pH, como em condições de contato com o suco gástrico humano. Sant'Ana e colaboradores [86] mostraram

que cepas de *S. cerevisiae* mutantes (*ena1-4Δ*) apresentam maior viabilidade celular quando expostas a pH 2 quando comparadas à cepa W303. Durso & Grynberg e colaboradores, analisaram o perfil de expressão gênica utilizando a técnica de *microarray* das cepas de UFMG A-905, Sb_17 e W303 expostas a condição de estresse ácido simulando as condições gástricas, e visualizaram um enriquecimento da categoria funcional "*response to salt stress (GO 0009651)*" na cepa não-probiótica (W303). O gene ENA1 é um dos onze genes relacionados a esta categoria funcional, e utilizando qPCR, não foi identificado qualquer sinal de expressão de ENA1 nas cepas probióticas (artigo em preparação). Os autores chamam ainda a atenção pelo inesperado fato da inibição da categoria funcional relacionada ao dobramento proteico, ao qual fazem parte genes de choque térmico como HSP82, considerando que, sob condições de estresse, uma célula utiliza chaperonas para prevenir o acúmulo de proteínas no citoplasma (artigo em preparação).

Leveduras *S. cerevisiae* fazem parte da alimentação humana há séculos, seja produzindo bebidas como o vinho ou alimentos como o pão, entretanto, a única levedura utilizada e reconhecida como probiótico é a cepa *S. boulardii*. Este trabalho foi o primeiro a analisar de forma ampla o genoma das cepas probióticas *S. boulardii* e *S. cerevisiae* UFMG A-905, mostrando semelhanças genômicas entre elas e em comparação com a cepa não-probiótica *S. cerevisiae* S288c. Ao disponibilizarmos estes dados oferecemos à comunidade científica a base para futuros estudos genômicos, e incentivamos a continuidade pela elucidação dos mecanismos que possam estar envolvidos no caráter probiótico.

6. CONCLUSÕES.

Nossas principais conclusões a partir das análises dos dados, são:

- ✓ *S. cerevisiae* UFMG A-905 possui similaridade genômica maior com *S. boulardii* do que com outras cepas não-probióticas de *Saccharomyces cerevisiae*, o que reforça ainda mais seu caráter probiótico;
- ✓ A presença de um maior número de genes de tRNAs na cepa UFMG A-905 pode estar relacionada à adaptação ao ambiente inóspito ao qual foi isolada;
- ✓ A sintenia e ausência de regiões de LTRs foi semelhante nos genomas probióticos, sugerindo uma conservação entre as cepas em comparação com a cepa não-probiótica S288c;
- ✓ Embora tenha sido evidenciado em apenas uma das quatro cepas probióticas, o enriquecimento da categoria funcional "*Binding Protein*" reforça o caráter probiótico destas cepas em comparação com a cepa S288c;
- ✓ Existem 20 genes presentes nas cepas probióticas e ausentes na cepa referência S288c, codificando em sua maioria, para proteínas hipotéticas;
- ✓ Os genes de S288c ausentes nas cepas probióticas são, em sua maioria, genes relacionados a atividade de transposição e retroelementos;
- ✓ Com base nas relações filogenéticas inferidas através de 415 genes ortólogos, *S. boulardii* é uma cepa de *S. cerevisiae*, e não uma subespécie do gênero *Saccharomyces*;
- ✓ Existe uma deleção de genes envolvidos no influxo de íons e prótons nas cepas probióticas, sugerindo um envolvimento indireto na resistência celular ao estresse ácido sofrido pelas leveduras probióticas;

7. BIBLIOGRAFIA.

1. Girardina M, Seidmana EG: **Indications for the Use of Probiotics in Gastrointestinal Diseases**. *Digestive diseases* 2011, **29**.
2. Tomasik PJ, Tomasik P: **Probiotics and Prebiotics**. *Cereal Chemistry* 2003, **80**:113–117.
3. Holzapfel WH, Haberer P, Snel J, Schillinger U, Huis in't Veld JH: **Overview of gut flora and probiotics.***International journal of food microbiology* 1998, **41**:85–101.
4. STANTON C, DESMOND C, FITZGERALD G, ROSS RP: **Probiotic health benefits: reality or myth?**. *Australian journal of dairy technology* , **58**:107–113.
5. H.Blehaut, J. Massot, G.W.Elmer RHL: **DISPOSITION KINETICS OF *SACCHAROMYCES BOULARDII* IN MAN AND RAT**. *Biopharmaceutics & Drug Disposition, 1989* 1989:353–364.
6. Fuller R: **Probiotics - The Scientific Basis**. *Chapman & Hall, Reading UK* 1992:1–8.
7. Klein S, Elmer G, McFarland L, Surawicz C, Levy R: **Recovery and Elimination of the Biotherapeutic Agent, *Saccharomyces boulardii,* in Healthy Human Volunteers**. *Pharmaceutical Research* 1993, **10**:1615–1619 LA – English.
8. McFarland, L.V., Bernasconi P: ***Saccharomyces boulardii*: A review of an innovative biotherapeutic agent.***Microb Ecol Health Dis* 1993, **6**:157–171.
9. Czerucka D, Dahan S, Mograbi B, Rossi B, Rampal P: ***Saccharomyces boulardii* Preserves the Barrier Function and Modulates the Signal Transduction Pathway Induced in Enteropathogenic *Escherichia coli*-Infected T84 Cells**. 2000, **68**:5998–6004.
10. Czerucka D, Rampal P: **Experimental effects of *Saccharomyces boulardii* on diarrheal pathogens**. *Microbes and Infection* 2002, **4**:733–739.
11. Kelesidis T, Pothoulakis C: **Efficacy and safety of the probiotic *Saccharomyces boulardii* for the prevention and therapy of gastrointestinal disorders.***Therapeutic advances in gastroenterology* 2012, **5**:111–25.
12. Martins FS, Nardi RMD, Arantes RME, Rosa CA, Neves MJ, Nicoli JR: **Screening of yeasts as probiotic based on capacities to colonize the gastrointestinal tract and to protect against enteropathogen challenge in mice.***The Journal of general and applied microbiology* 2005, **51**:83–92.
13. Pozzoni P, Riva A, Bellatorre AG, Amigoni M, Redaelli E, Ronchetti A, Stefani M, Tironi R, Molteni EE, Conte D, Casazza G, Colli A: ***Saccharomyces boulardii* for the**

Prevention of Antibiotic-Associated Diarrhea in Adult Hospitalized Patients: A Single-Center, Randomized, Double-Blind, Placebo-Controlled Trial. *Am J Gastroenterol* 2012, **107**:922–931.

14. Surawicz CM: **Probiotics, antibiotic-associated diarrhoea and *Clostridium difficile* diarrhoea in humans.***Best practice research Clinical gastroenterology* 2003, **17**:775–783.
15. Castagliuolo I, Riegler MF, Valenick L, Lamont T, Pothoulakis C: ***Saccharomyces boulardii* Protease Inhibits the Effects of *Clostridiumdifficile* Toxins A and B in Human Colonic Mucosa**. 1999.
16. Surawicz CM, McFarland L V, Greenberg RN, Rubin M, Fekety R, Mulligan ME, Garcia RJ, Brandmarker S, Bowen K, Borjal D, Elmer GW: **The search for a better treatment for recurrent *Clostridium difficile* disease: use of high-dose vancomycin combined with *Saccharomyces boulardii*.***Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2000, **31**:1012–7.
17. Williams NT: **Probiotics**. *American Journal of Health-System Pharmacy* 2010, **67** (6):449–458.
18. Garcia Vilela E, De Lourdes De Abreu Ferrari M, Oswaldo Da Gama Torres H, Guerra Pinto A, Carolina Carneiro Aguirre A, Paiva Martins F, Marcos Andrade Goulart E, Sales Da Cunha A: **Influence of *Saccharomyces boulardii* on the intestinal permeability of patients with Crohn's disease in remission**. *Scandinavian Journal of Gastroenterology* 2008, **43**:842–848.
19. Bleichner G, Bléhaut H, Mentec H, Moyse D: ***Saccharomyces boulardii* prevents diarrhea in critically ill tube-fed patients**. *Intensive Care Medicine* 1997, **23**:517–523 LA – English.
20. Costalos C, Skouteri V, Gounaris A, Sevastiadou S, Triandafilidou A, Ekonomidou C, Kontaxaki F, Petrochilou V: **Enteral feeding of premature infants with *Saccharomyces boulardii***. *Early human development* 2003:89–96.
21. Mansour-Ghanaei F, Dehbashi N, Yazdanparast K, Shafaghi A: **Efficacy of *Saccharomyces boulardii* with antibiotics in acute amoebiasis.***World journal of gastroenterology : WJG* 2003, **9**:1832–3.
22. Armuzzi A, Cremonini F, Ojetti V, Bartolozzi F, Canducci F, Candelli M, Santarelli L, Cammarota G, De Lorenzo A, Pola P, Gasbarrini G GA: **Effect of Lactobacillus GG Supplementation on Antibiotic-Associated Gastrointestinal Side Effects during *Helicobacter pylori* Eradication Therapy: A Pilot Study**. *Digestion* 2001, **63**.

23. Pérez-Sotelo LS, Talavera-Rojas M, Monroy-Salazar HG, Lagunas-Bernabé S, Cuarón-Ibargüengoytia JA, Jimenez RM V-CJ: **In vitro evaluation of the binding capacity of *Saccharomyces cerevisiae* Sc47 to adhere to the wall of *Salmonella* spp.***Rev Latinoam Microbiologia* 2005, **47**:70–75.

24. Boyle AG, Magdesian KG, Gallop R, Sigdel S, Durando MM: ***Saccharomyces boulardii* viability and efficacy in horses with antimicrobial-induced diarrhoea**. *Veterinary Record* 2012.

25. Martins FS, Rodrigues ACP, Tiago FCP, Penna FJ, Rosa CA, Arantes RME, Nardi RMD, Neves MJ, Nicoli JR: ***Saccharomyces cerevisiae* strain 905 reduces the translocation of *Salmonella enterica* serotype Typhimurium and stimulates the immune system in gnotobiotic and conventional mice.***Journal of Medical Microbiology* 2007, **56**(Pt 3):352–359.

26. Generoso S V, Viana M, Santos R, Martins FS, Machado JAN, Arantes RME, Nicoli JR, Correia MITD, Cardoso VN: ***Saccharomyces cerevisiae* strain UFMG 905 protects against bacterial translocation, preserves gut barrier integrity and stimulates the immune system in a murine intestinal obstruction model.***Archives of Microbiology* 2010, **192**:477–484.

27. Tiago FCP, Martins FS, Rosa C a., Nardi RMD, Cara DC, Nicoli JR: **Physiological characterization of non-Saccharomyces yeasts from agro-industrial and environmental origins with possible probiotic function**. *World Journal of Microbiology and Biotechnology* 2008, **25**:657–666.

28. Martins FS, Elian SD a, Vieira AT, Tiago FCP, Martins AKS, Silva FCP, Souza ELS, Sousa LP, Araújo HRC, Pimenta PF, Bonjardim C a, Arantes RME, Teixeira MM, Nicoli JR: **Oral treatment with *Saccharomyces cerevisiae* strain UFMG 905 modulates immune responses and interferes with signal pathways involved in the activation of inflammation in a murine model of typhoid fever.***International journal of medical microbiology : IJMM* 2011, **301**:359–64.

29. Dos Santos Sant'Ana G, Da Silva Paes L, Vieira Paiva AF, Fietto LG, Totola AH, Magalhães Trópia MJ, Lemos DS, Lucas C, Fietto JLR, Brandão RL, Castro I de M: **Protective effect of ions against cell death induced by acid stress in *Saccharomyces***. *FEMS Yeast Research* 2009, **9**:701–712.

30. Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth L V, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, Mitreva M, Cook L, Delehaunty KD, Fronick C, Schmidt H, Fulton L a, Fulton RS, Nelson JO, Magrini V, Pohl C, Graves T a,

Markovic C, Cree A, Dinh HH, Hume J, Kovar CL, Fowler GR, Lunter G, Meader S, Heger A, et al.: **Comparative and demographic analysis of orang-utan genomes.***Nature* 2011, **469**:529–33.

31. Africa W: **A map of human genome variation from population-scale sequencing.***Nature* 2010, **467**:1061–73.
32. Park PJ: **ChIP-seq: advantages and challenges of a maturing technology.***Nature reviews Genetics* 2009, **10**:669–80.
33. Li N, Ye M, Li Y, Yan Z, Butcher LM, Sun J, Han X, Chen Q, zhang X, Wang J: **Whole genome DNA methylation analysis based on high throughput sequencing technology**. *Methods* 2010, **52**:203–212.
34. Metzker ML: **Sequencing technologies - the next generation.***Nature reviews Genetics* 2010, **11**:31–46.
35. Nijkamp JF, van den Broek M, Datema E, de Kok S, Bosman L, Luttik M a, Daran-Lapujade P, Vongsangnak W, Nielsen J, Heijne WHM, Klaassen P, Paddon CJ, Platt D, Kötter P, van Ham RC, Reinders MJT, Pronk JT, de Ridder D, Daran J-M: **De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial biotechnology.***Microbial cell factories* 2012, **11**(March):36.
36. Lander ES: **Initial impact of the sequencing of the human genome.***Nature* 2011, **470**:187–197.
37. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans C a, Holt R a, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark a G, Nadeau J, McKusick V a, Zinder N, et al.: **The sequence of the human genome.***Science (New York, NY)* 2001, **291**(February):1304–1351.
38. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, et al.: **Initial sequencing and comparative analysis of the mouse genome.***Nature* 2002, **420**(December):520–562.
39. Alföldi J, Lindblad-Toh K: **Comparative genomics as a tool to understand evolution and disease**. *Genome Research* 2013, **23**:1063–1068.

40. Zhao Z, Liu H, Wang C, Xu J: **Correction: comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi**. *BMC genomics* 2014, **15**:6.

41. Microbiologie L De, De P, Viala P, Technologie C De, Bidenne C, Blondin B, Dequin S, Vezinhet F: **Analysis of the chromosomal DNA polymorphism of wine strains of *Saccharomyces cerevisiae***. *Current Genetics* 1992, **22**:1–7.

42. Adams J, Puskas-Rozsa S, Simlar J, Wilke CM: **Adaptation and major chromosomal changes in populations of *Saccharomyces cerevisiae***. *Current Genetics* 1992, **22**:13–19.

43. Pérez-Ortín JE, Querol A, Puig S, Barrio E: **Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains.***Genome research* 2002, **12**:1533–1539.

44. Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D: **Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*.***Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**:16144–16149.

45. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P: **Real-time DNA sequencing using detection of pyrophosphate release.***Analytical biochemistry* 1996, **242**:84–89.

46. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems**. *Journal of Biomedicine and Biotechnology* 2012, **2012**.

47. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G: **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies**. *Nucleic Acids Research* 2006, **34**.

48. Schadt EE, Turner S, Kasarskis A: **A window into third-generation sequencing**. *Human Molecular Genetics* 2010, **19**:227–240.

49. Green, M.R. Sambrook J: *Molecular Cloning: A Laboratory Manual*. 4th editio. New York: Cold Spring Harbor Laboratory Press; 2012:362.

50. Andrews S: **FastQC A Quality Control tool for High Throughput Sequence Data**. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* .

51. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.***Genome research* 2008, **18**:821–9.

52. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.***GigaScience* 2012, **1**:18.

53. Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OVC, Missawa SK, Galzerani F, Costa GGL, Vidal RO, Noronha MF, Dominska M, Andrietta MGS, Andrietta SR, Cunha AF, Gomes LH, Tavares FCA, Alcarde AR, Dietrich FS, McCusker JH, Petes TD, Pereira GAG: **Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production.***Genome research* 2009, **19**:2258–70.

54. Boetzer M, Henkel C V, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.***Bioinformatics (Oxford, England)* 2011, **27**:578–9.

55. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A: **CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes.***Source code for biology and medicine* 2011, **6**:11.

56. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.***Bioinformatics (Oxford, England)* 2007, **23**:1061–7.

57. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson S a: **BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons.***BMC genomics* 2011, **12**:402.

58. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.***BMC bioinformatics* 2011, **12**:491.

59. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.***Genome biology* 2013, **14**:R36.

60. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.***Nucleic acids research* 1997, **25**:955–64.

61. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.***BMC bioinformatics* 2009, **10**:421.

62. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.***Bioinformatics (Oxford, England)* 2005, **21**:3674–6.

63. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.***Molecular systems biology* 2011, **7**:539.

64. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.***Bioinformatics (Oxford, England)* 2009, **25**:1754–60.

65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.***Bioinformatics (Oxford, England)* 2009, **25**:2078–9.

66. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo M a: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.***Genome research* 2010, **20**:1297–303.

67. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.***Fly* 2012, **6**:80–92.

68. Khatri I, Akhtar A, Kaur K, Tomar R, Prasad GS, Ramya TNC, Subramanian S: **Gleaning evolutionary insights from the genome sequence of a probiotic yeast *Saccharomyces boulardii*.***Gut pathogens* 2013, **5**:30.

69. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 Genes**. *Science* 1996, **274**:546–567.

70. Douradinha B, Reis VCB, Rogers MB, Torres FAG, Evans JD, Marques ETA: **Novel insights in genetic transformation of the probiotic yeast *Saccharomyces boulardii*.***Bioengineered* 2014, **5**:21–9.

71. Satapathy SS, Dutta M, Ray SK: **Higher tRNA diversity in thermophilic bacteria: A possible adaptation to growth at high temperature**. *Microbiological Research* 2010, **165**:609–616.

72. Wessler SR: *Eukaryotic Transposable Elements: Teaching Old Genomes New Tricks*. 2006:138–165.

73. AJF, Griffiths, Miller JH, Suzuki DT et al: **Molecular nature of transposable elements in eukaryotes.** In *An Introduction to Genetic Analysis*. 7th editio. Edited by Freeman WH. New York; 2000.

74. Kidwell MG, Kidwell JF, Sved J a.: **Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination**. *Genetics* 1977, **86**:813–833.

75. Bleykasten-grosshans C, Friedrich A, Schacherer J: **Genome-wide analysis of intraspecific transposon diversity in yeast.***BMC genomics* 2013, **14**:399.

76. Styriak I, Nemcova R, Chang Y-H, Ljungh a.: **Binding of extracellular matrix molecules by probiotic bacteria**. *Letters in Applied Microbiology* 2003, **37**:329–333.

77. Cardinali G, Martini A: **Electrophoretic karyotypes of authentic strains of the sensu stricto group of the genus *Saccharomyces.****International journal of systematic bacteriology* 1994, **44**:791–7.

78. Hennequin C, Thierry A, Richard GF, Lecointre G, Gaillardin C, Dujon B, Nguyen H V: **Microsatellite Typing as a New Tool for Identification of *Saccharomyces cerevisiae* Strains Microsatellite Typing as a New Tool for Identification of *Saccharomyces cerevisiae* Strains**. 2001.

79. McCullough MJ, Clemons K V, McCusker JH, Stevens D a: **Species identification and virulence attributes of *Saccharomyces boulardii* (nom. inval.).***Journal of clinical microbiology* 1998, **36**:2613–7.

80. Fietto JLR, Fietto LG, Neves MJ, Nicoli JR, Castro IM: **Molecular and physiological comparisons between**. *Yeast* 2004, **621**:615–621.

81. Gadagkar SR, Rosenberg MS, Kumar S: **Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree.***Journal of experimental zoology Part B, Molecular and developmental evolution* 2005, **304**:64–74.

82. Wolfe KH, Li W-H: **Molecular evolution meets the genomics revolution.***Nature genetics* 2003, **33 Suppl**(march):255–65.

83. Guan Y, Dunham MJ, Troyanskaya OG: **Functional analysis of gene duplications in *Saccharomyces cerevisiae*.***Genetics* 2007, **175**:933–43.

84. Nasuno R, Hirano Y, Itoh T, Hakoshima T, Hibi T, Takagi H: **Structural and functional analysis of the yeast N-acetyltransferase Mpr1 involved in oxidative stress tolerance via proline metabolism.***Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**:11821–6.

85. Marger MD, Saier MH: **A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport.** *Trends in biochemical sciences* 1993, **18**:13–20.

86. Dos Santos Sant'Ana G, Da Silva Paes L, Vieira Paiva AF, Fietto LG, Totola AH, Magalhães Trópia MJ, Lemos DS, Lucas C, Fietto JLR, Brandão RL, Castro I de M: **Protective effect of ions against cell death induced by acid stress in *Saccharomyces***. *FEMS Yeast Research* 2009, **9**:701–712.

8. ANEXOS.

8.1. Relação das 415 proteínas do KOG utilizadas para o alinhamento e inferência filogenética.

KOG0002	KOG0291	KOG0524	KOG0862	KOG1159
KOG0003	KOG0292	KOG0530	KOG0871	KOG1180
KOG0018	KOG0302	KOG0534	KOG0876	KOG1185
KOG0025	KOG0313	KOG0544	KOG0878	KOG1211
KOG0047	KOG0318	KOG0556	KOG0880	KOG1235
KOG0062	KOG0327	KOG0559	KOG0888	KOG1241
KOG0073	KOG0328	KOG0563	KOG0894	KOG1255
KOG0077	KOG0329	KOG0567	KOG0922	KOG1268
KOG0084	KOG0330	KOG0602	KOG0927	KOG1272
KOG0092	KOG0344	KOG0622	KOG0933	KOG1291
KOG0100	KOG0346	KOG0625	KOG0935	KOG1299
KOG0102	KOG0357	KOG0631	KOG0937	KOG1300
KOG0103	KOG0358	KOG0650	KOG0938	KOG1301
KOG0122	KOG0359	KOG0659	KOG0948	KOG1342
KOG0142	KOG0362	KOG0675	KOG0952	KOG1349
KOG0173	KOG0363	KOG0679	KOG0959	KOG1350
KOG0174	KOG0364	KOG0683	KOG0964	KOG1351
KOG0175	KOG0365	KOG0687	KOG0969	KOG1355
KOG0177	KOG0366	KOG0688	KOG0985	KOG1358
KOG0179	KOG0367	KOG0727	KOG0989	KOG1367
KOG0180	KOG0371	KOG0728	KOG0991	KOG1370
KOG0181	KOG0372	KOG0729	KOG0996	KOG1373
KOG0182	KOG0373	KOG0734	KOG1036	KOG1374
KOG0183	KOG0394	KOG0741	KOG1047	KOG1390
KOG0184	KOG0400	KOG0756	KOG1058	KOG1393
KOG0185	KOG0402	KOG0758	KOG1062	KOG1394
KOG0188	KOG0418	KOG0767	KOG1068	KOG1415
KOG0190	KOG0419	KOG0780	KOG1077	KOG1430
KOG0209	KOG0420	KOG0784	KOG1078	KOG1433
KOG0211	KOG0424	KOG0785	KOG1088	KOG1439
KOG0225	KOG0434	KOG0787	KOG1098	KOG1448
KOG0233	KOG0441	KOG0788	KOG1099	KOG1458
KOG0258	KOG0450	KOG0815	KOG1112	KOG1463
KOG0261	KOG0460	KOG0817	KOG1123	KOG1466
KOG0264	KOG0462	KOG0820	KOG1131	KOG1468
KOG0271	KOG0466	KOG0829	KOG1137	KOG1487
KOG0276	KOG0468	KOG0852	KOG1145	KOG1491
KOG0279	KOG0481	KOG0853	KOG1148	KOG1494
KOG0285	KOG0495	KOG0857	KOG1149	KOG1498
KOG0289	KOG0523	KOG0861	KOG1158	KOG1506

KOG1523	KOG1762	KOG2472	KOG2825	KOG3237
KOG1526	KOG1769	KOG2481	KOG2833	KOG3239
KOG1531	KOG1770	KOG2509	KOG2854	KOG3273
KOG1532	KOG1772	KOG2519	KOG2855	KOG3275
KOG1533	KOG1774	KOG2529	KOG2874	KOG3284
KOG1534	KOG1775	KOG2531	KOG2877	KOG3297
KOG1535	KOG1779	KOG2535	KOG2908	KOG3311
KOG1539	KOG1781	KOG2537	KOG2909	KOG3313
KOG1540	KOG1782	KOG2555	KOG2916	KOG3318
KOG1549	KOG1784	KOG2572	KOG2930	KOG3320
KOG1556	KOG1795	KOG2574	KOG2948	KOG3330
KOG1562	KOG1800	KOG2575	KOG2952	KOG3343
KOG1566	KOG1816	KOG2606	KOG2957	KOG3349
KOG1567	KOG1872	KOG2613	KOG2967	KOG3361
KOG1568	KOG1889	KOG2617	KOG2971	KOG3372
KOG1596	KOG1915	KOG2623	KOG2981	KOG3387
KOG1597	KOG1936	KOG2636	KOG3013	KOG3404
KOG1626	KOG1942	KOG2638	KOG3022	KOG3405
KOG1636	KOG1979	KOG2653	KOG3031	KOG3406
KOG1637	KOG1980	KOG2680	KOG3048	KOG3411
KOG1641	KOG1986	KOG2700	KOG3049	KOG3418
KOG1643	KOG1992	KOG2703	KOG3052	KOG3428
KOG1644	KOG2004	KOG2707	KOG3064	KOG3430
KOG1654	KOG2014	KOG2711	KOG3079	KOG3432
KOG1662	KOG2017	KOG2719	KOG3090	KOG3442
KOG1664	KOG2035	KOG2726	KOG3106	KOG3448
KOG1668	KOG2036	KOG2728	KOG3147	KOG3449
KOG1688	KOG2044	KOG2732	KOG3149	KOG3453
KOG1691	KOG2047	KOG2738	KOG3157	KOG3457
KOG1692	KOG2067	KOG2749	KOG3163	KOG3459
KOG1712	KOG2104	KOG2754	KOG3164	KOG3463
KOG1722	KOG2189	KOG2757	KOG3167	KOG3475
KOG1723	KOG2270	KOG2767	KOG3174	KOG3479
KOG1727	KOG2276	KOG2770	KOG3180	KOG3480
KOG1728	KOG2292	KOG2772	KOG3185	KOG3482
KOG1733	KOG2303	KOG2775	KOG3188	KOG3498
KOG1742	KOG2311	KOG2781	KOG3189	KOG3499
KOG1746	KOG2321	KOG2783	KOG3204	KOG3503
KOG1750	KOG2387	KOG2784	KOG3205	KOG3506
KOG1754	KOG2415	KOG2785	KOG3218	KOG3954
KOG1755	KOG2446	KOG2792	KOG3222	KOG3974
KOG1758	KOG2451	KOG2803	KOG3229	KOG4392
KOG1760	KOG2467	KOG2807	KOG3232	KOG4655

8.2. Genes que sofreram impacto do tipo HIGH nos quatro genomas probióticos.

	Tipo de impacto	Efeito do impacto	Variante	Gene
1	FRAME_SHIFT		aat/	YIL012W
2	FRAME_SHIFT		aca/CGCTTaca	YER046W-A
3	FRAME_SHIFT		aca/	YER145C-A
4	FRAME_SHIFT		agaaca/	YGL193C
5	**FRAME_SHIFT**		**ata/atGa**	**YER097W**
6	FRAME_SHIFT		atg/	YPL025C
7	FRAME_SHIFT		cga/	YER186C
8	FRAME_SHIFT		gga/	MNS1
9	FRAME_SHIFT		ggc/	SPH1
10	FRAME_SHIFT		tac/	YFL032W
11	FRAME_SHIFT		tat/Gtat	YBR134W
12	SPLICE_SITE_ACCEPTOR			CNB1
13	START_LOST	MISSENSE	aTg/aAg	COA1
14	START_LOST	MISSENSE	aTg/aCg	YGL006W-A
15	START_LOST	MISSENSE	aTg/aCg	YJR107W
16	START_LOST	MISSENSE	aTg/aCg	YNR062C
17	START_LOST	MISSENSE	aTg/aGg	TAD1
18	START_LOST	MISSENSE	atG/atA	AKR2
19	START_LOST	MISSENSE	atG/atA	ATG29
20	START_LOST	MISSENSE	atG/atA	SSK22
21	START_LOST	MISSENSE	atG/atA	YDL094C
22	START_LOST	MISSENSE	atG/atA	YFR035C
23	**START_LOST**	**MISSENSE**	**atG/atA**	**YHL037C**
24	START_LOST	MISSENSE	atG/atA	YKR075W-A
25	START_LOST	MISSENSE	atG/atA	YPR002C-A
26	START_LOST	MISSENSE	atG/atC	GTT2
27	**START_LOST**	**MISSENSE**	**atG/atT**	**YAL067W-A**
28	**START_LOST**	**MISSENSE**	**Atg/Gtg**	**YHL037C**
29	**STOP_GAINED**	**NONSENSE**	**Aaa/Taa**	**YAL067W-A**
30	STOP_GAINED	NONSENSE	Aag/Tag	YGR051C
31	**STOP_GAINED**	**NONSENSE**	**Aga/Tga**	**OSW2**
32	STOP_GAINED	NONSENSE	Caa/Taa	ACT1
33	STOP_GAINED	NONSENSE	Caa/Taa	AMA1
34	STOP_GAINED	NONSENSE	Caa/Taa	CRT10
35	**STOP_GAINED**	**NONSENSE**	**Caa/Taa**	**OSW2**
36	STOP_GAINED	NONSENSE	Caa/Taa	RFA2
37	STOP_GAINED	NONSENSE	Caa/Taa	RPL37A
38	STOP_GAINED	NONSENSE	Caa/Taa	TSC11
39	STOP_GAINED	NONSENSE	Caa/Taa	YBR109W-A
40	STOP_GAINED	NONSENSE	Caa/Taa	YCR087W
41	STOP_GAINED	NONSENSE	Caa/Taa	YKL147C
42	STOP_GAINED	NONSENSE	Caa/Taa	YML116W-A
43	STOP_GAINED	NONSENSE	Caa/Taa	ZPS1
44	STOP_GAINED	NONSENSE	Cag/Tag	ECM33

45	STOP_GAINED	NONSENSE	Cag/Tag	EMI1
46	STOP_GAINED	NONSENSE	Cag/Tag	HXT8
47	STOP_GAINED	NONSENSE	Cag/Tag	KIN28
48	STOP_GAINED	NONSENSE	Cag/Tag	RPL13B
49	STOP_GAINED	NONSENSE	Cag/Tag	YLR317W
50	STOP_GAINED	NONSENSE	Cag/Tag	YML099W-A
51	STOP_GAINED	NONSENSE	Cag/Tag	YNL235C
52	STOP_GAINED	NONSENSE	Cga/Tga	ECM1
53	STOP_GAINED	NONSENSE	Cga/Tga	IWR1
54	**STOP_GAINED**	**NONSENSE**	**Cga/Tga**	**NSP1**
55	STOP_GAINED	NONSENSE	Cga/Tga	VMR1
56	STOP_GAINED	NONSENSE	Cga/Tga	YGR176W
57	**STOP_GAINED**	**NONSENSE**	**Cga/Tga**	**YHR028W-A**
58	STOP_GAINED	NONSENSE	Cga/Tga	YOR282W
59	STOP_GAINED	NONSENSE	Gaa/Taa	ABZ1
60	STOP_GAINED	NONSENSE	Gaa/Taa	PUP1
61	STOP_GAINED	NONSENSE	Gaa/Taa	YCL012C
62	STOP_GAINED	NONSENSE	Gaa/Taa	YFL015C
63	STOP_GAINED	NONSENSE	Gaa/Taa	YML057C-A
64	STOP_GAINED	NONSENSE	Gag/Tag	YGL258W-A
65	STOP_GAINED	NONSENSE	Gga/Tga	YLR444C
66	**STOP_GAINED**	**NONSENSE**	**taC/taA**	**NSP1**
67	STOP_GAINED	NONSENSE	taC/taA	TIR4
68	STOP_GAINED	NONSENSE	taC/taA	YHR071C-A
69	STOP_GAINED	NONSENSE	taC/taG	YBR090C
70	STOP_GAINED	NONSENSE	taC/taG	YLR400W
71	STOP_GAINED	NONSENSE	taC/taG	YOR024W
72	STOP_GAINED	NONSENSE	taT/taA	IRC14
73	**STOP_GAINED**	**NONSENSE**	**taT/taA**	**YOL079W**
74	STOP_GAINED	NONSENSE	taT/taG	YFL021C-A
75	STOP_GAINED	NONSENSE	tCa/tAa	YLL058W
76	STOP_GAINED	NONSENSE	tCa/tAa	YOR050C
77	**STOP_GAINED**	**NONSENSE**	**tCa/tGa**	**YER097W**
78	STOP_GAINED	NONSENSE	tCa/tGa	YJL182C
79	STOP_GAINED	NONSENSE	tCg/tAg	YEL067C
80	STOP_GAINED	NONSENSE	tgC/tgA	GAC1
81	STOP_GAINED	NONSENSE	tGg/tAg	AGP3
82	STOP_GAINED	NONSENSE	tGg/tAg	COX5B
83	STOP_GAINED	NONSENSE	tGg/tAg	DSF1
84	**STOP_GAINED**	**NONSENSE**	**tGg/tAg**	**NSP1**
85	STOP_GAINED	NONSENSE	tGg/tAg	PRM9
86	STOP_GAINED	NONSENSE	tGg/tAg	YBR113W
87	STOP_GAINED	NONSENSE	tGg/tAg	YIR020C
88	STOP_GAINED	NONSENSE	tGg/tAg	YMR193C-A
89	STOP_GAINED	NONSENSE	tGg/tAg	YNL285W
90	**STOP_GAINED**	**NONSENSE**	**tgG/tgA**	**IMD4**
91	STOP_GAINED	NONSENSE	tgG/tgA	MAL33
92	STOP_GAINED	NONSENSE	tgG/tgA	PGM2
93	STOP_GAINED	NONSENSE	tgG/tgA	TRM2
94	STOP_GAINED	NONSENSE	tgG/tgA	YAR028W
95	STOP_GAINED	NONSENSE	tgG/tgA	YDR010C

96	**STOP_GAINED**	**NONSENSE**	**tgG/tgA**	**YHR028W-A**
97	STOP_GAINED	NONSENSE	tgG/tgA	YLR428C
98	**STOP_GAINED**	**NONSENSE**	**tgG/tgA**	**YOL079W**
99	STOP_GAINED	NONSENSE	tgG/tgA	YOL134C
100	STOP_GAINED	NONSENSE	tgT/tgA	YMR316C-A
101	STOP_GAINED	NONSENSE	tTa/tAa	MET1
102	STOP_GAINED	NONSENSE	tTa/tAa	YDR215C
103	STOP_GAINED	NONSENSE	tTa/tGa	UBP15
104	STOP_GAINED	NONSENSE	tTa/tGa	YPL277C
105	**STOP_LOST**	**MISSENSE**	**Taa/Caa**	**IMD4**
106	**STOP_LOST**	**MISSENSE**	**Taa/Caa**	**OSW2**
107	STOP_LOST	MISSENSE	Taa/Caa	PAU7
108	STOP_LOST	MISSENSE	Taa/Caa	QDR2
109	**STOP_LOST**	**MISSENSE**	**Taa/Caa**	**RPL16A**
110	STOP_LOST	MISSENSE	Taa/Caa	YHR022C
111	STOP_LOST	MISSENSE	Taa/Caa	YNL324W
112	STOP_LOST	MISSENSE	Taa/Caa	YOR072W
113	STOP_LOST	MISSENSE	Taa/Gaa	CRS5
114	STOP_LOST	MISSENSE	Taa/Gaa	KDX1
115	STOP_LOST	MISSENSE	taA/taT	YAL031W-A
116	STOP_LOST	MISSENSE	taA/taT	YEL074W
117	STOP_LOST	MISSENSE	tAa/tTa	APA2
118	**STOP_LOST**	**MISSENSE**	**Tag/Cag**	**HPC2**
119	STOP_LOST	MISSENSE	taG/taC	BBP1
120	STOP_LOST	MISSENSE	taG/taT	NFT1
121	STOP_LOST	MISSENSE	tAg/tGg	ABM1
122	STOP_LOST	MISSENSE	tAg/tGg	FLO8
123	**STOP_LOST**	**MISSENSE**	**tAg/tGg**	**HPC2**
124	STOP_LOST	MISSENSE	tAg/tGg	NIT1
125	STOP_LOST	MISSENSE	tAg/tGg	YDR114C
126	STOP_LOST	MISSENSE	tAg/tGg	YJR079W
127	STOP_LOST	MISSENSE	tAg/tGg	YKL223W
128	STOP_LOST	MISSENSE	Tga/Cga	HAC1
129	**STOP_LOST**	**MISSENSE**	**Tga/Cga**	**RPL16A**
130	STOP_LOST	MISSENSE	Tga/Gga	CMC4
131	STOP_LOST	MISSENSE	tgA/tgG	PSP2
132	STOP_LOST	MISSENSE	tGa/tTa	MUD1
133	STOP_LOST	MISSENSE	tGa/tTa	YBR197C

8.3. Resultado do blast dos genes relacionados à resistência ao estresse ácido.

Resultado do Blast: em verde está representada a evidência de similaridade via alinhamento, em vermelho está representada a ausência de evidência de similaridade via alinhamento e em amarelo está representada uma similaridade parcial *query* x *subject* no alinhamento. Estão representados 54 genes em verde, 20 em amarelo e 7 em vermelho.

		UFMG A-905	Sb_17	Sb_ATCC	Sb_EDRL
1	YBL105C_PKC1				
2	YBR001C_NTH2				
3	YBR066C_NRG2				
4	YBR072W_HSP26				
5	YBR126C_TPS1				
6	YBR140C_IRA1				
7	YBR160W_CDC28				
8	YBR182C_SMP1				
9	YBR260C_RGD1				
10	YBR295W_PCA1				
11	YBR296C_PHO89				
12	YCR021C_HSP30				
13	YDL138W_RGT2				
14	YDL185W_VMA1				
15	YDL194W_SNF3				
16	YDR001C_NTH1				
17	YDR028C_REG1				
18	YDR038C_ENA5				
19	YDR039C_ENA2				
20	YDR040C_ENA1				
21	YDR043C_NRG1				
22	YDR074W_TPS2				
23	YDR171W_HSP42				
24	YDR173C_ARG82				
25	YDR216W_ADR1				
26	YDR258C_HSP78				
27	YDR477W_SNF1				
28	YDR533C_HSP31				
29	YEL011W_GLC3				
30	YER129W_SAK1				
31	YFL014W_HSP12				
32	YFR014C_CMK1				
33	YFR015C_GSY1				
34	YGL006W_PMC1				

35	YGL008C_PMA1				
36	YGL035C_MIG1				
37	YGL071W_AFT1				
38	YGL179C_TOS3				
39	YGL209W_MIG2				
40	YGL248W_PDE1				
41	YGR217W_CCH1				
42	YHL027W_RIM101				
43	YHR030C_SLT2				
44	YIL050W_PCL7				
45	YJL141C_YAK1				
46	YJL159W_HSP150				
47	YJL164C_TPK1				
48	YJR090C_GRR1				
49	YKL048C_ELM1				
50	YKL062W_MSN4				
51	YKL166C_TPK3				
52	YKL190W_CNB1				
53	YKR058W_GLG1				
54	YLL026W_HSP104				
55	YLR044C_PDC1				
56	YLR113W_HOG1				
57	YLR138W_NHA1				
58	YLR258W_GSY2				
59	YLR259C_HSP60				
60	YLR310C_CDC25				
61	YLR332W_MID2				
62	YLR342W_FKS1				
63	YMR037C_MSN2				
64	YNL027W_CRZ1				
65	YNL098C_RAS2				
66	YNL291C_MID1				
67	YOL016C_CMK2				
68	YOL081W_IRA2				
69	YOR002W_ALG6				
70	YOR008C_SLG1				
71	YOR020C_HSP10				
72	YOR087W_YVC1				
73	YOR101W_RAS1				
74	YOR178C_GAC1				
75	YOR391C_HSP33				
76	YPL203W_TPK2				
77	YPL240C_HSP82				
78	YPL280W_HSP32				
79	YPR026W_ATH1				
80	YPR160W_GPH1				
81	YPR184W_GDB1				

8.1. Artigos científicos publicados e em processo de submissão a revistas internacionais.

BATISTA, T.M.; MARQUES JR., E.T.A.; Franco, GR; DOURADINHA, B.**Draft Genome Sequence of the Probiotic Yeast *Saccharomyces cerevisiae* var. *boulardii* Strain ATCC MYA-796**. Genome Announcements, v. 2, p. e01345-14-e01345-14, 2014.

BATISTA, TM,Castro, IE,Araujo, FMG, Salim, ACM, Cardoso, DC, Martins, FS,Brandão, RL, Drummond, MG, Oliveira, GC, Rosa, CA, Nicoli, JRand Franco, Gr. **Genomic comparison of the two probiotic yeasts *Saccharomyces cerevisiae* UFMG A-905 and *Saccharomyces cerevisiae* var. *boulardii* with the non-probiotic *Saccharomyces cerevisiae* S288c.**Em preparação para submissão à revista Integrative Biology. Impact Factor 4.45.

Durso DF & Grynberg P, Dias L.L.C., **BATISTA, TM**, Nicoli, J.R., Brandão R.L., Castro I.M., Franco G.R.**Genome Expression Response of Probiotics Yeasts Under Simulated Gastric Enviroment**. Em preparação.