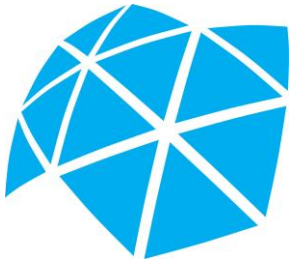

Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Biociências de Botucatu
Programa de Pós-Graduação em Biotecnologia

Busca e análise de lncRNA (*long non-coding* RNAs) importantes para a tolerância ao etanol em *Saccharomyces cerevisiae*

Lucas Farinazzo Marques

Botucatu – SP
2019

Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Biociências de Botucatu
Programa de Pós-Graduação em Biotecnologia

Busca e análise de lncRNA (*long non-coding* RNAs) importantes para a tolerância ao etanol em *Saccharomyces cerevisiae*

Mestrando: Lucas Farinazzo Marques
Orientador: Dr. Guilherme Targino Valente

Dissertação apresentada ao Programa de Pós-Graduação em Biotecnologia do Instituto de Biociências de Botucatu da Universidade Estadual Paulista "Júlio de Mesquita Filho", para obtenção do título de mestre.

Botucatu - SP
2019

M357b

Marques, Lucas Farinazzo
Busca e análise de lncRNA (long non-coding RNAs) importantes para a tolerância ao etanol em Saccharomyces cerevisiae / Lucas Farinazzo Marques. -- Botucatu, 2019
104 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Biociências, Botucatu
Orientador: Guilherme Targino Valente

1. long non-coding RNAs. 2. Saccharomyces cerevisiae. 3. Análise de transcriptoma. 4. Bioinformática. 5. Tolerância ao etanol. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências, Botucatu. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

“Somos feitos de milhares de partes com milhares de funções, todas funcionando em harmonia para nos manter vivos. Caso uma parte da nossa maquinaria imperfeita falhe, a vida falha. Isso nos faz perceber quão frágeis e falhos somos.”
- Ingun Black-Briar, Elder Scrolls V: Skyrim.

Dedico este trabalho a minha família que agora está longe e a minha companheira e aos meus amigos que sempre estão perto.

Agradecimentos

Agradeço aos meus pais, *Fortunato Massaud Marques* e *Regina Célia Farinazzo Marques*, por me proverem do melhor possível para me desenvolver como um cidadão consciente do meu papel na sociedade. Pensar em vocês me leva diretamente ao meu lar, onde sou amado e querido. Espero de verdade conseguir passar pelo menos um pouco de tudo aquilo que vocês me passaram de experiência de vida e sabedoria para dar início ao desenvolvimento de uma sociedade melhor. Graças a vocês não tenho medo da vida, da dor, do trabalho e do que vier pela frente. Amo muito vocês!

Agradeço ao meu irmão de sangue, *Daniel Farinazzo Marques* que, mesmo com suas limitações, ainda me ama incondicionalmente e que, graças a esse amor, eu sempre vou tentar ser uma pessoa melhor que ele espera que eu seja!

Agradeço a minha madrinha *Rosana Farinazzo*, minha falecida tia-avó *Therezinha Farinazzo* e minha falecida avó *Célia Moreira Farinazzo* que praticamente foram mães para mim, me criando quando meus pais estavam ausentes a trabalho. Sem vocês, eu não seria nem 10% do que sou hoje. Espero de coração que tanto minha tia-avó quanto minha avó estejam em um bom local agora papeando do jeito que sempre gostavam de fazer em vida. Amo vocês!

Agradeço à *Cristiane Barp* que conheci quase por acaso durante meu período de intercâmbio no Ciências sem Fronteiras nos EUA. Quem diria que essa guria se tornaria o amor da minha vida, com seu jeitinho e sotaque do sul, o encanto no seu sorriso e a alegria no seu olhar! Uma simples carona do aeroporto de Chicago até o hotel pago pela AbbVie me levou a conhecer minha companheira e confidente, quem diria! Sou muito grato a você, meu amor, por tudo que já passamos juntos e vamos ainda passar. Te amo muito!

Agradeço aos meus avós de coração, Ronaldo Marcelo Martins e Lizi Maria Martins. Vocês entraram na minha vida como vizinhos, porém, com o passar do tempo, depois de vários pedidos de ajuda para olhar o computador, para dar olhadas na televisão, de risadas de histórias passadas e de muitas jantas para me ensinar a cozinhar, vocês se tornaram muito mais do que isso. Agradeço muito pelos conselhos e pelo tempo juntos!

Agradeço aos meus irmãos de coração, *Carlos Eduardo de Souza, Yuri Lima, Rodolfo Rocha, Ryan Bishop, Silas Muniz, Rafael Araújo, Lucas Godinho, Bruce Oliveira, Lucas Pereira, Leonardo Rander, Michael Durfee, Bianca Felizardo, Isaque Fontinele, Luis César Bastos, Fagner Rezende e Isadora Travnik.* Tenho um imenso carinho por cada um de vocês e gratidão por tudo que já passamos juntos. É uma excelente sensação poder reencontrar com vocês depois de muito tempo e simplesmente não parecer que se passaram cinco minutos que ficamos longes. Vocês já presenciaram meus altos e baixos, me apoiaram em todos os momentos que precisei e, graças a isso, me sinto um ser humano melhor e mais preparado para a vida. Independente do que, vocês sempre serão pessoas especiais para mim e assim espero ser para vocês também!

Agradeço aos meus primos que foram criados junto comigo, *Tércio Sammuel Farinazzo*, *Leonardo Farinazzo, Pedro Farinazzo*, *Luiz Felipe Cremonezi*, *Fernanda Farinazzo*, *Igor Farinazzo* (que agora sou padrinho!) e *Isabella Farinazzo*. Passamos incontáveis tardes de finais de semana assistindo vários filmes (indo dos de comédia até os de terror!), jogando centenas de jogos (indo do Atari até o Playstation 4!), encenando para câmeras imaginárias em programas de TV. Graças a vocês, eu consegui dar asas a minha imaginação!

Agradeço ao meu orientador *Guilherme Targino Valente,* por me desafiar diariamente, inclusive antes mesmo do meu mestrado começar. Graças a tudo isso, eu cheguei na UNESP com uma mentalidade que agora vejo como inocente e, após meu período de mestrado, me sinto muito mais preparado do que quando entrei. Todos esses desafios foram colocados visando sempre o meu melhor. A principal frase que me marcou em todo nosso tempo junto foi: “Eu esperava mais de você” quando entreguei minha primeira versão da qualificação, foi um choque gigantesco para mim, porém acho que foi o maior desafio que já me foi colocado na minha vida. Graças a você, professor Guilherme, eu sempre tive que dar 120% de mim e fico extremamente grato por isso, sou agora uma pessoa e profissional melhor do que era antes. Sentirei falta dos seus sábios conselhos nas nossas reuniões quase diárias com minha graduação, porém os levarei para o resto da vida para replicá-los às próximas gerações. Não tenho palavras de quanto sou grato por tudo isso!

Agradeço aos meus mais recentes irmãos do laboratório que trabalhei durante o mestrado, SBGL (*Systems Biology and Genomic Laboratory)*, *Luiz Henrique Cardoso, Ivan Wolf, Lucas Lazari, Camila Cristina de Oliveira, Eric Kazuo, Amanda Piveta Schnepper* e *Guilherme Luz,* Muito obrigado pelos vários *brainstorms* que tivemos juntos, pelas horas de trabalho em conjunto, pelos papos enquanto esperávamos programas rodarem, pelas rodadas de cerveja, por todas as dicas de como sobreviver em Botucatu e por toda ajuda que vocês me proveram. Espero encontrar com vocês em futuras oportunidades para podermos trocar experiências e jogar papo fora!

Agradeço a todos os professores que tive o prazer ou de ter aula ou de ter algum tipo de contato em Botucatu, em especial os professores *Rafael Plana Simões* e *José Luiz Rybarczyk Filho* e as professoras *Rejane Maria Tommasini Grotto* e *Carla dos Santos Riccardi*. Aprendi muito com vocês, tanto como profissional quanto como pessoa. Obrigado por ajudarem a expandir meus horizontes!

Agradeço a todos os meus professores da Universidade Federal de Juiz de Fora, em especial a todos ex-orientadores que tive durante minha graduação em Ciências Biológicas: *Antônia Ribeiro, José Paulo Rodrigues Furtado de Mendonça, Claudia Avellar Freitas, Marcelo dos Santos Oliveira e Priscila Vanessa Zabala Capriles Goliatt.* Obrigado pelos ensinamentos, desafios propostos, pelas dicas, sugestões e críticas. Vocês foram e continuam sendo uma grande inspiração!

Agradeço à *Universidade Estadual Paulista Júlio de Mesquita Filho* (UNESP-Botucatu) por toda a infraestrutura oferecida, além do suporte acadêmico fornecido para minha formação como acadêmico e profissional. Também agradeço ao *Programa de Pós-Graduação em Biotecnologia* oferecido no *Instituto de Biociências* pelo suporte e apoio a minha formação.

Agradeço a *Fundação de Amparo à Pesquisa do Estado de São Paulo* (**Processo número 2017/14764-3**) pelo auxílio financeiro fornecido durante esses dois anos de trabalho do mestrado em Botucatu-SP.

Resumo

A levedura *Saccharomyces cerevisiae* é o microrganismo mais utilizado para a produção de etanol devido a sua alta capacidade fermentativa e resistência aos estresses oriundos desse processo. Entretanto, a própria concentração de etanol é um dos fatores mais limitantes no processo de produção desse combustível. Os aspectos da genômica funcional relacionada à tolerância ao etanol são ainda pouco esclarecidos, e nem mesmo se sabe se os lncRNAs tem papel nesse processo. Poucos lncRNAs foram identificados em *S. cerevisiae*, e nem mesmo se conhece as redes lncRNAs-proteínas nessa espécie e nem se podem codificar micropeptídeos. Nesse contexto, este trabalho visa identificar lncRNAs em linhagens de *S. cerevisiae* com diferentes níveis de tolerância ao etanol. Para isso, foi realizado a montagem dos lncRNAs, predição de ligações lncRNA-proteínas, buscas de micropeptídeos, análises de conservação genômica, estrutural e funcional dos lncRNAs, avaliação da influência do lncRNAs em regular as expressões de seus vizinhos e comparação dos resultados entre linhagens mais e menos tolerantes ao etanol. As análises de enriquecimento ontológico apontam para uma relação próxima entre os lncRNAs e a tolerância ao etanol e uma conservação funcional, embora os dados não reportem nenhuma conservação nem genômica nem estrutural. Além disso, variados tipos de prováveis regulações foram sugeridos, sendo a regulação em *trans* majoritariamente inversa entre os lncRNAs e seus genes-alvo, diferentemente da maioria das regulações em *cis*. Esses dados, quando comparados com a literatura para diversas espécies, acaba confirmando a conservação funcional dos lncRNAs e o papel dessas moléculas não-codificantes como reguladores. Por fim, sugerimos que os lncRNAs estão agindo na superação do estresse causado pela alta concentração de etanol.

Palavras-chave: RNAs longos não-codificantes, *Saccharomyces cerevisiae*, Bioinformática

Abstract

The yeast *Saccharomyces cerevisiae* is the most used microorganism for ethanol production due to its high fermentative capacity and resistance to different stressors along this process. However, the ethanol concentration is one of the most limiting factors of fuel production. The functional genomics aspects related to the ethanol tolerance are still unclear, and it is not clear if the lncRNAs really have a role in this process. Few lncRNAs were identified in *S. cerevisiae*, lncRNA-protein networks of this species are still unknown and also if they can code micropeptides. In this context, this thesis aims to identify lncRNAs and evaluate their roles in *S. cerevisiae* ethanol tolerance. Then, it was performed the assembling of lncRNAs, predictions of lncRNA-protein interactions, searches for potential micropeptides coding-lncRNAs, analysis of genomic, structural and functional conservation of lncRNAs, evaluation of the lncRNAs influence in regulating the expressions of their neighbors, and comparison between strains that are more and less tolerant to the ethanol. Moreover, many putative regulatory pathways were here suggested, being that most *trans* regulations act on an inversely manner between the expression of the lncRNAs and their target-genes, unlike observed in most of *cis* regulations. The current literature confirms the lncRNAs functional conservation here observed, and the role of these non-coding molecules as regulators. Finally, here we suggest that lncRNAs are acting to overcome the stress caused by the highest ethanol concentration.

Keywords: long non-coding RNAs, *Saccharomyces cerevisiae*, Bioinformatics

Lista de Ilustrações

Lista de Tabelas

Sumário

1. Introdução

1.1. Bioetanol como biocombustível

Atualmente, os combustíveis fósseis são as principais fontes de energia mundial (RABINOVITCH-DEERE et al., 2013), porém o grande crescimento das demandas energéticas acompanha o crescimento da população. O fato de 80% da energia consumida no mundo ser de fontes não-renováveis, o preço e a crescente demanda por esses produtos afeta diretamente a economia global (CHAKRABORTY et al., 2012). Por conta desse interesse mercadológico, além da crescente preocupação sobre mudança global e segurança energética (KARUPPIAH et al., 2008), muitos países têm focado no aprimoramento e no desenvolvimento de biocombustíveis, sendo o bioetanol o mais comercializado (CHAKRABORTY et al., 2012).

O termo biocombustível se refere principalmente a combustíveis líquidos ou até gasosos (REIJNDERS, 2006), sendo uma forma sustentável de produção energética produzidas a partir de biomassa de cana-de-açúcar, milho, materiais lignocelulósicos, e até de óleos de plantas (DEMIRBAS, 2017). O bioetanol é uma das alternativas mais promissoras dentre os biocombustíveis em relação aos combustíveis fósseis, podendo ser produzido por várias fontes renováveis ricas em carboidratos (ZABED et al., 2017). O Brasil e os Estados Unidos são os maiores produtores de bioetanol do mundo, usando a cana-de-açúcar e milho como fontes primárias, respectivamente (DIAS et al., 2012).

O processo mais comum para a produção desse tipo de combustível é a tecnologia de primeira geração, a qual baseia-se principalmente na fermentação dos açúcares simples extraídos diretamente da cana-de-açúcar, beterraba, grãos e sorgo (CHAKRABORTY et al., 2012). Porém, durante o processo de produção de etanol, a própria concentração desse composto é o principal estressor para as células (MUSSATTO et al., 2010).

Dentre os principais organismos fermentadores utilizados no processo de fermentação do etanol, destaca-se o fungo *Saccharomyces cerevisiae*, por ser uma espécie

mais interessante quando comparado com bactérias e outros fungos em várias características fisiológicas para o contexto industrial. Esse fungo tolera uma grande variação no pH, apresenta uma maior capacidade fermentativa e uma maior tolerância ao etanol e outros inibidores, além de serem praticamente inofensivas à saúde humana (LIN et al., 2012; MUSSATTO et al., 2010; PRASERTWASU et al., 2014). Essa levedura consegue, inclusive, produzir etanol em temperaturas de 40ºC-45ºC (BALAKUMAR; ARASARATNAM, 2012), meio ácido de até pH 4 (NARENDRANATH; POWER, 2005) e concentração de 13% de etanol no meio (GHAREIB; YOUSSEF; KHALIL, 1988). As linhagens de *S. cerevisiae* consideradas industriais, apresentam também uma maior performance na fermentação de etanol, decorrente principalmente da sua alta performance de crescimento celular, maior tolerância ao etanol, menor necessidade de nitrogênio e uma menor formação de subprodutos (CHIN, 2012).

Em decorrência da grande relevância desse fungo, tanto para a indústria quanto para análises laboratoriais, foi criado um banco comunitário para todos os dados genômicos relativos à *S. cerevisiae,* o SGD (*Saccharomyces Genome Database* (CHERRY et al., 2012)), Esse banco de dados fornece dados curados manualmente das mais diversas linhagens do fungo, além de demonstrar em várias rotas metabólicas as diferentes expressões e comportamentos dos genes e outras moléculas expressas por *S. cerevisiae*, inclusive apresentado dados de vários estudos importantes para o entendimento do funcionamento desse importante organismo, como os de tolerância à estressores externos (ex. etanol).

A tolerância à um estressor externo pode ser resumida como o potencial de sobrevivência das células durante uma exposição crônica a essa substância (STANLEY et al., 2010a). No geral, apesar de *S. cerevisiae* tolerar grandes concentrações de etanol quando comparado com outros organismos, o estresse gerado é também um grande desafio para a célula (YU et al., 2012), pois aumenta o processo de morte celular

acarretando na redução da produção como um todo. Assim, a obtenção de linhagens mais tolerantes ao etanol é de extrema importância para a indústria (STANLEY et al., 2010b), uma vez que tal feito poderia gerar economia significativa anualmente. Normalmente, a produção de etanol em escala industrial gera entre 10-14% (v/v) de etanol e o rendimento teórico da conversão de etanol deve ser cerca de 91,5% (BAI; ANDERSON; MOO-YOUNG, 2008)(BAI; ANDERSON; MOO-YOUNG, 2008). Consequentemente, vários estudos têm como foco a tolerância ao etanol baseada na hipótese de que as linhagens mais tolerantes ao etanol teriam melhores rendimentos e produtividade (FIEDUREK; SKOWRONEK; GROMADA, 2011; SHI; WANG; WANG, 2009; THAMMASITTIRONG et al., 2012).

1.2. Aspectos moleculares da tolerância ao etanol e criação de linhagens de *S. cerevisiae* mais tolerantes

O etanol age sobre: 1- as macroestruturas das células, causando modificações nos lipídeos e proteínas das membranas citoplasmáticas e parede celular (biossíntese e organização) (MA; LIU, 2010; STANLEY et al., 2010b; YU et al., 2012); além de aumentar a fluidez das membranas causando acidificação do citoplasma (MA; LIU, 2010); 2- a expressão dos genes e o *folding* proteico (DING et al., 2009; HALLSWORTH et al., 2003; INGRAM, 1990; MA; LIU, 2010); 3- as vias metabólicas responsáveis pela formação do etanol e em muitos outros processos, tais como a síntese de nucleotídeos, energia, redox-homeostase, oxidação do NADH, oxidação das pentose-fosfatos, produção do acetil-CoA e ácido acético, entre outros processos (DING et al., 2009; MA; LIU, 2010; STANLEY et al., 2010a).

Tem sido reportado (JIA; ZHANG; LI, 2010) que a engenharia genética tem propiciado a criação de linhagens de *S. cerevisiae* mais tolerantes ao etanol baseada na geração de mutações dirigidas ou no *screening* de fenótipos a partir de uma biblioteca de mutantes. No entanto, para a produção eficiente desses microrganismos, é necessário

compreender os mecanismos que influenciam a sensibilidade das células a esse composto e como elas se adaptam molecularmente frente ao estresse causado por ele (ZHANG et al., 2009). O resultado do funcionamento deste mecanismo foi abordado baseando-se em 5.400 curvas de crescimento de 36 linhagens de *S. cerevisiae* (industriais e laboratoriais) (KANG et al., 2019). Os autores concluíram que as linhagens produzidas para uso industrial possuem maior heterozigosidade, resultando em fusões genômicas durante os processos fermentativos afetando diretamente a expressão de fatores de transcrição. Quanto ao fenótipo, as fusões resultam em um metabolismo de etanol relativamente mais alto e com menor atividade do ciclo do ácido cítrico (CAC), levando ao acúmulo dos intermediários do CAC e, consequentemente, uma tolerância maior etanol quando comparado com linhagens laboratoriais. No entanto, detalhes moleculares ainda precisam de mais esclarecimentos, pois, como exemplo, muitos genes já determinados como importantes para a tolerância ao etanol podem estar *up*-regulados ou até mesmo não apresentar modificações na sua expressão durante o estresse por etanol (MA; LIU, 2010). Por outro lado, genes que estão *down*-regulados na presença desse composto nem sempre são requeridos na tolerância; nesses casos, tais genes podem acabar sendo candidatos para a construção de linhagens mais tolerantes ao etanol (por exemplo, diferentes mutações da linhagem BY4743) (KASAVI et al., 2014). Fica evidente que, na tentativa de se obter somente uma característica ignorando todos os fatores biológicos do organismo, o fenótipo desejado raramente é alcançado, isso porque é necessário uma compreensão maior dos mecanismos intrínsecos a essa espécie (JIA; ZHANG; LI, 2010).

Para entender melhor o funcionamento e como funcionam as características dos organismos, é importante a análise e compreensão do transcriptoma do organismo-alvo, uma vez que esse tipo de investigação é interessante para verificar o estado celular e entender os processos biológicos pelos quais a célula está passando (LI; LIM; LING, 2019). Essa verificação pode ser feita de várias formas e, no caso de *S. cerevisiae*, os estudos

que abordam a avaliação dos transcriptomas quanto à tolerância ao etanol, utilizaram majoritariamente técnicas de *microarrays*, havendo ainda poucos relatos do uso de RNA-Seq diretamente nesse foco (ex. Goud; Ulaganathan (2019) e Shekhawat et al. (2019)). Contudo, o RNA-Seq é uma ferramenta mais adequada para estudar transcriptomas por ser mais sensível para genes expressos em níveis muito altos ou muito baixos, além de fornecer um detalhamento muito maior para características transcricionais, tais como novas regiões transcritas (OSHLACK; ROBINSON; YOUNG, 2010; PETRYSZAK et al., 2014).

Em geral, as análises dos transcriptomas de *S. cerevisiae* têm revelado que genes relativos a homeostase iônica, relacionados a *heat shock*, síntese de trealose, de defesa antioxidante, de resposta à estressores, utilização de energia, mecanismos de transporte, metabolismo de lipídeos, equilíbrio redox, entre outros processos metabólicos estão relacionados à tolerância ao etanol (ALEXANDRE et al., 2001; CHANDLER et al., 2004; STANLEY et al., 2010b). Kasavi et al. (2014) conduziu uma análise de expressão gênica em duas linhagens de *S. cerevisiae* para avaliar a tolerância ao etanol, montando redes de interação entre proteínas (PPIs), permitindo descobrir 17 novos genes candidatos relacionados ao fenótipo em questão. Wohlbach et al. (2014) utilizaram RNA-Seq para explorar a tolerância ao etanol em linhagens naturais de *S. cerevisiae* tratadas com 5% de etanol e alta temperatura por 30 minutos. As análises revelaram que muitos genes ativados por estresse foram induzidos pelo tratamento. Além disso, diversos desses genes eram linhagem-específicos e os genes *up*-regulados, em geral, eram fatores de transcrição. No entanto, esse trabalho teve como foco a genômica estrutural em detrimento da genômica funcional. Além disso, há o problema de terem usado dois estressores simultaneamente, tornando muito difícil certificar-se quais das respostas eram de fato relativas ao etanol.

Além de todas essas abordagens usadas para entender as nuances das análises de transcriptoma em *S. cerevisiae* nas mais diversas condições, recentemente foi descoberto que os *long non-coding RNAs* (lncRNAs) perfazem grande parte de todos os RNAs das

células e que também são essenciais na regulação de vários processos nessa levedura, muitas vezes estando relacionados a resposta a estressores externos (NIEDERER; HASS; ZAPPULLA, 2017). Porém, é evidente a necessidade de um maior aprofundamento desse conhecimento para detalhar melhor a arquitetura e funções dos lncRNAs em *S. cerevisiae*.

1.3. Uma visão geral dos lncRNAs

Os lncRNAs são uma grande e diversa classe de RNA-reguladores, medindo pelo menos 200 nucleotídeos e perfazendo uma boa parte de um transcriptoma (CHEKANOVA, 2015). Em humanos e em camundongos, por exemplo, há 19.361 (~30,12%) e 12.168 (~23,70%) lncRNAs, respectivamente, segundo o GENCODE V29 (FRANKISH et al., 2019). E, para classificar todos esses lncRNAs, uma das classificações mais utilizadas seguem as diretrizes do GENCODE v7 (DERRIEN et al., 2012), classificando essas moléculas como *lincRNAs*, *intronic*, *sense* e *antisense*. Para armazenar todos esses dados, diversos bancos de lncRNAs vêm sendo criados, tais como o GREENC (contém mais de 120.000 lncRNAs de 37 espécies de plantas e seis de algas) (PAYTUVÍ GALLART et al., 2016) e o NONCODE (contém mais de 500.000 genes de lncRNA das diversas espécies, tais como fungos e alguns primatas) (ZHAO et al., 2016).

Em geral, os lncRNAs são responsáveis por inúmeras funções tais como atuar como reguladores epigenéticos, reguladores da expressão gênica de mRNAs e capturadores de microRNAs (ANDERSON et al., 2015). Os lncRNAs também estão diretamente envolvidos em diversos outros mecanismos moleculares que incluem, muitas vezes, interação com uma ou mais proteínas (ZHU et al., 2013). Alguns lncRNAs podem funcionar como “iscas moleculares”, podendo ligar-se a fatores de transcrição específicos para prevenir sua associação com o DNA (GEISLER; COLLER, 2013), além de poderem se ligar também a mRNAs antissentido (*antisense*), regulando-os de forma pós-transcricional (CARRIERI et al., 2012) e também servindo como arcabouço para a montagem de complexos

macromoleculares (TRIPATHI et al., 2010). Contudo, ainda há poucos estudos referentes aos lncRNAs e sua atuação em diversos mecanismos celulares (WILUSZ; SUNWOO; SPECTOR, 2009). Mais detalhes dos tipos de regulação dos lncRNA estão no item 1.4.

A literatura acerca dos lncRNAs em *S. cerevisiae* é ainda muito escassa, sendo que apenas 18 lncRNAs já foram devidamente descritos, anotados e curados para essa espécie (TILL; MACH; MACH-AIGNER, 2018). No entanto, pelo menos 75% do genoma de *S. cerevisiae* é transcrito, sendo os lncRNAs a maior parte dos 25% não transcritos nessa espécie (YAMASHITA; SHICHINO; YAMAMOTO, 2016). Nessa levedura, os lncRNAs identificados estão envolvidos em mudanças metabólicas, iniciação da diferenciação sexual e outros processos ainda pouco esclarecidos (YAMASHITA; SHICHINO; YAMAMOTO, 2016).

Apesar do nome e de serem classicamente colocados apenas na parte não-transcrita do transcriptoma, alguns lncRNAs podem codificar pequenos peptídeos bioativos. Como exemplo, um micropeptídeo descrito por Anderson et al. (2015) regula a performance muscular em camundongos. Esse peptídeo contém 46 aminoácidos e é potencialmente codificado por uma ORF (*open reading frame*) de 138 nts. Outro micropeptídeo oriundo de um lncRNA foi descrito por Nelson et al. (2016), o qual reporta também que tal molécula aumenta a performance muscular em camundongos. Em contraste aos peptídeos bioativos clássicos, os quais são clivados de grandes proteínas precursoras e têm como alvo os caminhos secretores pela sequência N-terminal de sinalização, os micropeptídeos são codificados a partir de pequenas ORFs, não necessitam da mesma sequência sinalizadora e são liberados no citoplasma logo após a tradução (CRAPPÉ; VAN CRIEKINGE; MENSCHAERT, 2014). O recente aumento desses tipos de estudos adiciona uma nova camada de complexidade no entendimento da genômica funcional, indicando o potencial em estudar os lncRNAs também como possíveis agentes codificantes de micropeptídeos eventualmente funcionais, sendo que atualmente já foram encontrados 47 sORFs

traduzidas em lncRNAs de leveduras (CHOI; KIM; NAM, 2018). Porém, dos 18 lncRNAs já descritos para *S. cerevisiae*, nenhum deles é codificante.

Como exemplo do papel dos lncRNAs já descritos em *S. cerevisiae*, o lncRNA codificado na região promotora do gene IME1 (o qual codifica um fator de transcrição responsável pela esporulação e pelo processo de meiose) possui a sua expressão reprimida pelo fator de transcrição RME1. Quando o RME1 se liga ao promotor do IME1, induz a expressão do lncRNA nesse promotor, iniciando então modificações epigenéticas (inserção de histonas SET2 metiltransferase e histonas SET3 desacetilase) no promotor do IME1 reprimindo sua expressão (YAMASHITA; SHICHINO; YAMAMOTO, 2016), levando ao controle do processo reprodutivo quando a célula está sob algum tipo de levando ao controle do processo reprodutivo quando a célula está sob algum tipo de estresse. Esse tipo de controle sob determinadas condições ambientais também é observado na expressão do lncRNA gal10 em *S. cerevisiae*. A expressão desse lncRNA induz a metilação de lisinas da histona H3 levando à desacetilação da região *downstream* a ele. Porém, flanqueando a região que sofre esse processo, estão os loci dos genes GAL10 e GAL1, os quais sofrem uma redução ou repressão de suas expressões por conta desse sistema. Além disso, foi demonstrado que o *decapping* do lncRNA GAL10 mediado pelas enzimas DCP-2, RAT1 e XRN1, pode levar à expressão dos genes GAL10 e GAL1 uma vez que a desacetilação não ocorrerá nos seus promotores (YAMASHITA; SHICHINO; YAMAMOTO, 2016). Outros exemplos de regulação dos lncRNAs mediante a um estressor são propostos nesta dissertação, como a relação entre o lncRNA transcr_19942 da linhagem laboratorial BMA64-1A e o gene YGR250C (responsável pela restauração de crescimento que foi interrompido por algum estressor), onde a interação entre os dois é alterada conforme a célula passa pelo estresse etanólico; ou também a alteração no tipo de interação entre o lncRNA transcr_21244 da linhagem laboratorial S288c e o gene YBR187W (atua no

transporte de cálcio e magnésio em organelas), que segue o mesmo princípio de alteração conforme o estressor ao qual a célula está exposta.

Os lncRNAs também participam do processo de adesão célula-célula em *S. cerevisiae* sob condições de formação de filamentos. Um gene chamado FLO11, o qual participa dessa adesão, é expresso de forma ativa somente quando o lncRNA ICR1 está reprimido. Isso ocorre quando o fator de transcrição FLO8 (o qual liga-se ao promotor do FLO11) induz a expressão *upstream* do lncRNA PWR1, o qual suprime a expressão do lncRNA ICR1 (YAMASHITA; SHICHINO; YAMAMOTO, 2016).Outros lncRNAs participam da regulação do ciclo celular (regulando a expressão de quinases dependentes de ciclinas), biossíntese de serinas, a regulação de outros genes responsáveis pela meiose (ex. ZIP2 e IME4), entre outros processos biológicos. Em suma, lncRNAs são reguladores que respondem muito bem a estímulos externos (YAMASHITA; SHICHINO; YAMAMOTO, 2016), os quais devem influenciar na tolerância ao etanol que será investigada neste trabalho.

Todos esses processos moleculares podem ser explicados em grande parte pela interação entre lncRNAs-proteínas, as quais são majoritariamente interações regulatórias. Para entender melhor essa relação, alguns preditores de interação entre essas moléculas podem ser utilizados. Porém, como citado por Xiao et al. (2017), os atuais preditores só utilizam os aspectos inerentes dos lncRNAs e das proteínas, ignorando a informação implícita nas topologias das redes biológicas associadas aos lncRNAs. Isso torna necessário a criação e utilização de preditores diversos considerando estas limitações, além da natureza codificante dos lncRNAs que também continua pouco conhecida principalmente por conta da sua grande heterogeneidade, tornando ainda mais difícil o entendimento das relações dos micropeptídeos codificados pelos lncRNAs.

1.4. Os processos de regulação gênica e os mecanismos conhecidos de regulação exercidos pelos lncRNAs

A regulação gênica inclui uma grande variedade de mecanismos que são utilizados pelas células para controlar a produção de produtos gênicos, que podem ser proteínas após o processo de tradução ou RNAs após o processo de transcrição. Por ser um processo com várias etapas, ele pode ser alterado de diversas formas levando a diferentes resultados para o organismo. As alterações podem ser simples modificações no DNA mediados por metilação ou fosforilação (BELL et al., 2011) até mecanismos mais complexos como mudanças nos *enhancers* durante o processo de recrutamento da RNA polimerase II e a transcrição de RNAs não-codificantes (KIM et al., 2010). Esses processos perpassam diferentes tipos de regulação em vias importantes tanto para humanos (como é o caso de sítios de CpG que são metilados levando ao silenciamento de um gene e, consequentemente, a regulação de transcrição em células cancerígenas (BIRD, 2002; SAXONOV; BERG; BRUTLAG, 2006)) quanto para a própria *S. cerevisiae* (ex. reprogramação do perfil transcricional da RNA polimerase II para aumento da tolerância ao etanol e da sua produtividade (QIU; JIANG, 2017)).

Além das alterações pré-transcricionais descritas, também existem casos em que a regulação ocorre entre o processo transcricional e de tradução (TZFIRA, 2008). Esses processos vão do desbaste do final da 5’ até a criação do RNA mensageiro poliadenilado, sendo todos mediados por várias proteínas que levarão, finalmente, a exportação desse RNA ao citoplasma (SINGH et al., 2015). Porém, inúmeras interferências podem acontecer durante esse processo. Como exemplo, a expressão das HSPs (*heat-shock proteins*) são inibidas sob estresse etanólico e isso é feito durante as etapas de processamento de mRNA por conta do aumento na exportação de muitas poli-As (IZAWA; INOUE, 2009). Outro exemplo é o lncRNA LAST (LncRNA-*Assisted Stabilization of Transcripts*) que age como

estabilizador de mRNA ao cooperar com a CNBP (CCHC-*type zinc finger nucleic acid binding protein*) para promover a estabilidade da cicilina D1 de mRNA (CAO et al., 2017).

Todos os processos de regulação gênica possuem em comum a presença de algum tipo de RNA não-codificante (ncRNAs), tais como os RNAs de transferência (tRNAs), os RNAs ribossomais (rRNAs), os microRNAs (miRNAs) e os RNAs longos não-codificantes (lncRNAs) (TZFIRA, 2008). Por conta da sua grande gama de formas de atuação nos processos transcricionais da célula, os lncRNAs podem ser enquadrados em quatro modos de atuação na regulação da expressão gênica: 1- como sinais (**Figura 1-I**), na qual a expressão do lncRNA pode refletir as ações em conjunto de fatores de transcrição; 2- como iscas (**Figura 1-II**), nos quais os lncRNAs podem titular fatores de transcrição e outras proteínas da cromatina; 3- como guias (**Figura 1-III**), nos quais os lncRNAs podem recrutar enzimas que alteram a cromatina para mirar em genes, seja em *cis* (na região próxima ao lncRNA) ou em *trans* (em genes-alvo distantes); 4- como *scaffolds* (**Figura 1-IV**), nos quais os lncRNAs podem carrear várias proteínas para formar complexos ribonucleoproteicos, podendo estabilizar estruturas nucleares ou servindo como complexos de sinalização (LI et al., 2019; WANG; CHANG, 2011).

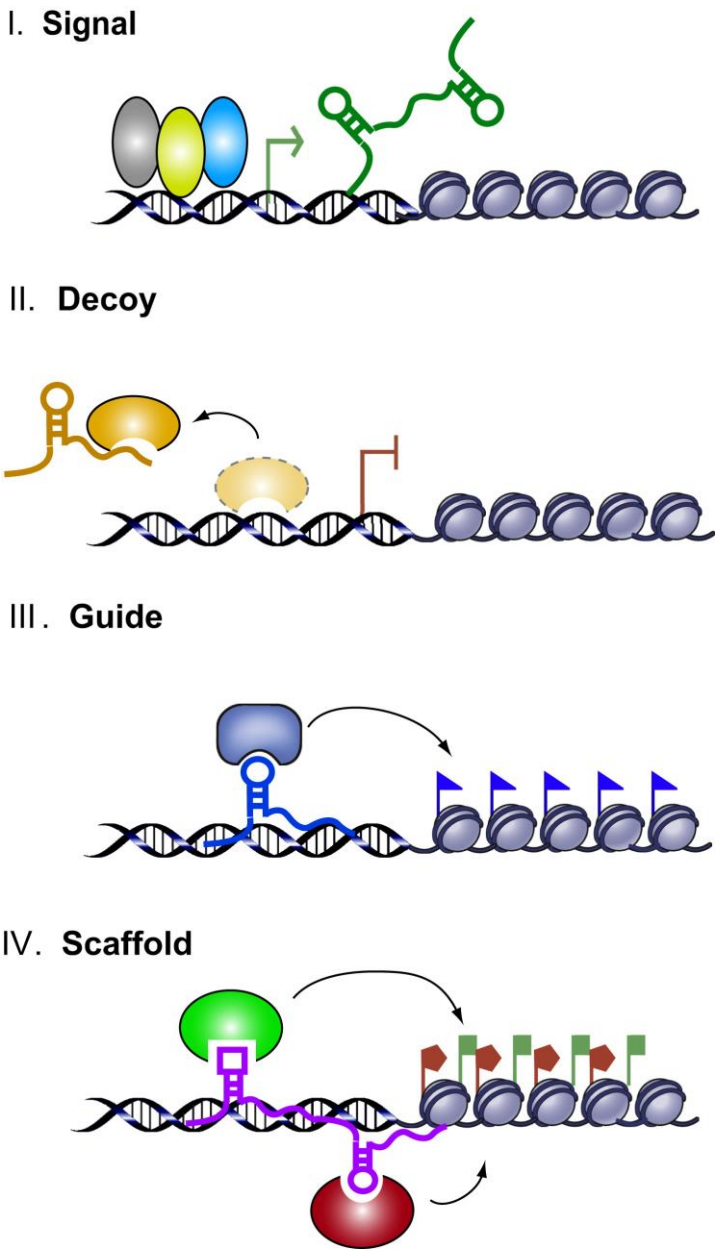

Figura 1 - Diagrama ilustrativo das quatro formas de regulação da expressão gênica. Figura retirada de (WANG; CHANG, 2011)

A atuação de um lncRNA na regulação pré-transcricional pode ser como sinal, guia ou *scaffold* no nível da cromatina (BROCKDORFF, 2013; MERCER; MATTICK, 2013), sendo que já foi reportado que essas moléculas podem participar no processo de metilação, como descrito por Wang et al. (2015), os quais identificaram a atuação do lncRNA *Dum* regulando a expressão do gene DPPa2 afetando a metilação do DNA. Um dos primeiros lncRNAs identificados, o HOTAIR (*Hox transcript antisense intergenic RNA*), também tem uma atuação pré-transcricional podendo coordenar a modificação da histona (TSAI et al., 2010).

Durante o processo transcricional a RNA polimerase II (RNA Pol II) se liga a região promotora do gene com o apoio de fatores de transcrição (TFs) e a transcrição termina quando a polimerase chega ao terminador (TZFIRA, 2008). Os lncRNAs também podem regular a expressão do gene se ligando-se diretamente aos TFs, ou à RNA Pol II ou até mesmo na ligação da polimerase com o promotor (LI et al., 2019). Por exemplo, o lncHIFCAR (*long noncoding HIF-1A co-activating RNA*) forma um complexo com a HIF-1A ao se ligar diretamente à ela, facilitando o recrutamento da HIF-1A e do cofator p300 para os promotores-alvo (SHIH et al., 2017). Além disso, existe também a indicação de que o promotor de um lncRNA pode competir pelo *enhancer* do promotor de um gene codificador de proteína. Cho et al. (2018) identificaram que o promotor do lncRNA PVT1 (que é independente do gene PVT1) compete com o promotor *Myc*, inibindo a expressão do gene *Myc* e conferindo uma função supressora de crescimento de tumores em humanos.

Logo após o processo transcricional, ainda existem vários níveis de regulação exercidas pelos lncRNAs. Como mostrado na **Figura 2**, esses diversos processos podem acontecer dentro do núcleo ou até no citoplasma celular. Essas formas de regulação podem ser: 1- por *splicing* alternativo (**Figura 2-1**), sendo que os lncRNAs competem com os pre-mRNAs na ligação com as proteínas regulatórias do *splicing*; 2- protegendo o mRNA do *decay* natural (**Figura 2-2**); 3- acelerando a degradação do mRNA (**Figura 2-3**); 4- reprimindo a tradução (**Figura 2-4**) ao fazer o recrutamento de repressores do processo traducional celular; 5- ativando a tradução (**Figura 2-5**) ao recrutar ribossomos para elevar a eficiência e velocidade do processo de tradução de mRNAs; 6- servindo de isca para microRNAs (**Figura 2-6**) fazendo uma associação funcional com esses pequenos RNAs competindo diretamente com essas pequenas moléculas (LI et al., 2019; YOON; ABDELMOHSEN; GOROSPE, 2013).

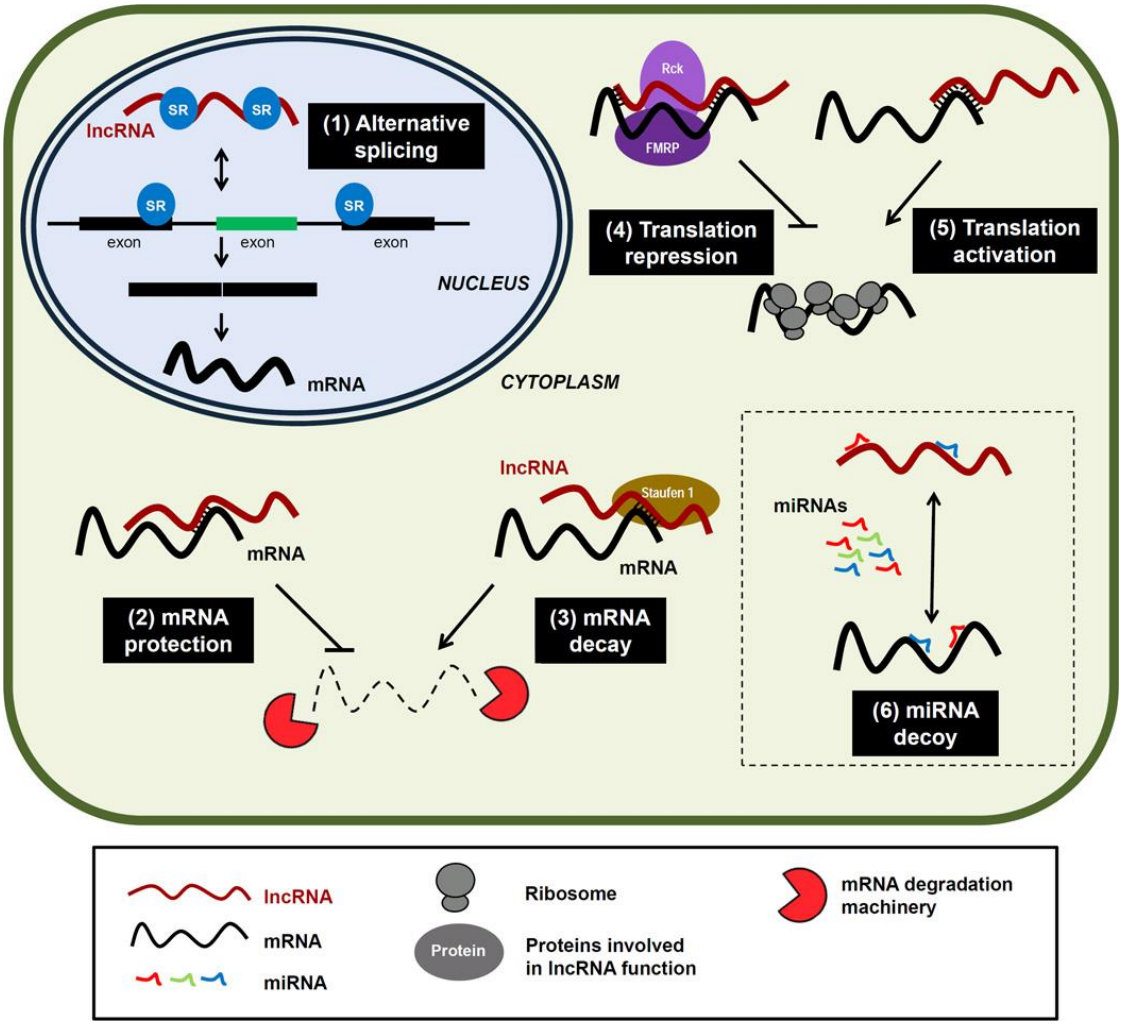

Figura 2 - Tipos de regulação pós-transcricional exercida pelos lncRNAs. As caixas pretas mostram os processos mais importantes que essas moléculas não-codificantes estão envolvidas. Figura retirada de (YOON; ABDELMOHSEN; GOROSPE, 2013).

É extremamente relevante também análises da região próxima à essas moléculas, já que muitos lncRNAs atuam principalmente como reguladores locais (GUIL; ESTELLER, 2012; ØROM et al., 2010) e suas funções podem explicar a observação dos níveis de expressão dos lncRNAs sendo muitas vezes diretamente correlacionado com a expressão de genes próximos (EBISUYA et al., 2008). Como exemplo, um estudo (ENGREITZ et al., 2016) que avaliou 12 loci genômicos que produzem lncRNAs em camundongos e reportou que 5 desses loci influenciam a expressão em *cis* do gene vizinho. Contudo, concluíram que existe a troca de informação entre sequências vizinhas, além dessa regulação ser um fenômeno prevalente envolvendo múltiplos mecanismos e sinais regulatórios *cis*.

Os processos de regulação gênica em *cis* possuem grande variação de formas de funcionamento, variando de regulações pré-transcricionais por meio da atuação na região promotora de genes próximos (inclusive funcionando de forma parecida à um *enhancer)*, ligação à RNA Polimerase II e até mesmo atuando nos mecanismos de *splicing* (ENGREITZ et al., 2016; WANG et al., 2011). Um estudo (JOUNG et al., 2017) conseguiu identificar 11 loci de lncRNAs que, após recrutar um ativador, fazia a mediação da resistência aos inibidores do gene BRAF em células humanas de melanoma, sendo que ainda conseguiram desenvolver um método em escala genômica de CRISPR-Cas9 que mira em mais de 10.000 sítios de início de transcrição de lncRNAs para identificar loci não-codificantes em sua região de vizinhança.

Contudo, identificar potenciais lncRNAs em *S. cerevisiae* que influenciam diretamente na expressão gênica da sua região vizinha, seja pela ótica da tolerância ao etanol, seja em condição ótima, é um tema muito relevante e ainda inexplorado. Para entender esse tipo de associação, foi usada a análise por *time-course* (BENDJILALI et al., 2017) nessa dissertação, a qual, com o passar do tempo, as relações dos lncRNAs com os genes classicamente conhecidos por terem relação com tolerância a diversos estressores ficam mais claras, já que podem revelar tendências de aumento ou diminuição de expressão. Existindo uma relação de níveis de expressão entre lncRNA-gene vizinho, esse processo pode ser replicado para outros organismos para identificação de inúmeros casos de regulação.

Essa dissertação irá colocar à luz potenciais lncRNAs em *S. cerevisiae* identificados nos transcriptoma de seis diferentes linhagens com diferentes níveis de tolerância ao etanol. No capítulo 1, será demonstrada a metodologia empregada para encontrar esses lncRNAs, a análise tanto da interação desses lncRNAs com proteínas-alvo sob uma ótica de entendimento da tolerância ao etanol e de enriquecimento ontológico dessas moléculas não-codificantes. A hipótese é que os lncRNAs são importantes no fenótipo de tolerância

ao etanol. No capítulo 2, as relações dos lncRNAs com as suas vizinhanças genômicas estão reportadas e analisadas sob a hipótese de que os lncRNAs influenciam significativamente seus genes vizinhos alterando sua expressão para regular diversos processos celulares. Ressalta-se que no capítulo 2, não é explorado a tolerância ao etanol uma vez que esse estressor é compreendido nesse capítulo como um agente causador de mudanças sistêmicas que possibilitam avaliar o impacto dos lncRNAs na regulação dos seus vizinhos sob a ótica de comparação entre uma condição ótima e uma condição extrema.

2. Objetivos

2.1. Objetivos geral

Buscar e avaliar o papel dos lncRNAs em *Saccharomyces cerevisiae* durante o estresse por etanol.

Avaliar as interações proteínas-lncRNAs e vizinhos-lncRNAs oriundas de dados de transcriptoma.

Analisar os dados de *time-course* durante o estresse ao etanol e avaliar o papel dos lncRNAs na regulação da expressão genica de seus vizinhos.

2.2. Objetivos específicos

1. Reconstruir os lncRNAs a partir dos dados de transcriptoma oriundos de experimentos de estresse máximo ao etanol com linhagens tolerantes e pouco tolerantes à esse composto;
2. Predizer a interação entre lncRNAs-proteínas para cada uma das linhagens escolhidas para o projeto;
3. Predizer o potencial desses lncRNAs em codificar micropeptídeos;
4. Avaliar e comparar as ontologias das proteínas-alvo dos lncRNAs entre as diferentes linhagens, buscando compreender se há um padrão para as linhagens mais tolerantes ao etanol;
5. Avaliar a expressão diferencial dos lncRNAs comparando tratamento *vs* controle.
6. Avaliar as vizinhanças genômicas dos lncRNAs e identificar o impacto dos lncRNAs na expressão desses genes.

Capítulo 1 – A montagem e avaliação dos *long non-coding RNAs*: *uma* análise da importância dessas moléculas no fenótipo de tolerância ao etanol.

Este capítulo reporta o procedimento realizado para a montagem dos lncRNAs identificados nesse trabalho com base em dados de RNA-Seq, além de buscar semelhanças entre os prováveis lncRNA aqui descritos e o potencial dos mesmos como codificadores de micropeptídeos. Além disso, são feitas análises das prováveis interações lncRNA-proteínas entre as linhagens com diferentes níveis de tolerância ao etanol, das ontologias dessas proteínas-alvo e os valores de expressão diferencial dos lncRNAs. Por último, análises de conservação funcional e estrutural são apresentadas.

3. Materiais e métodos

3.1. Descrição das linhagens, experimentos de tolerância e atribuição aos grupos HT e LT

O experimento realizado pelo grupo SBGL (*Systems Biology and Genomics Lab*) para separação das linhagens de *S. cerevisiae* em mais ou menos tolerantes ao etanol, contou com 6 diferentes linhagens que foram submetidas à determinação da máxima tolerância ao etanol em cultura celular ao longo de 1h de tratamento com etanol, seguindo o protocolo de Lewis et al. (2010); nos controles foram adicionados solução fisiológica. Para as linhagens BMA64-1A e S288c, as culturas celulares foram também tratadas ao longo de 2 e 4 horas em um experimento de *time-course*. As tolerâncias variaram entre 20-30% (v/v) de etanol e as linhagens foram separadas em duas classes utilizando aprendizado de máquina não supervisionado (WOLF, 2019): 1- HT, que compõem as linhagens mais tolerantes; 2- LT, que compõem as linhagens menos tolerantes (**Figura 3, Tabela 1**)(ALMEIDA, 2017)

Figura 3 - Desenho experimental da identificação e separação das 6 linhagens aqui utilizadas. A: Determinação da tolerância ao etanol para 13 linhagens e sua classificação de acordo com suas tolerâncias em concentrações crescentes de etanol; B: Cultivo após 1h de tratamento; C: Matriz com todas as medidas; D: Resultado do aprendizado não supervisionado; Círculos Vermelhos e azuis: as condições de tratamento e controle, respectivamente; Caixas em laranja: maior tolerância para cada linhagem; Círculos cinza: linhagens LT; Círculos Rosa: linhagens HT. Figura retirada de (ALMEIDA, 2017).

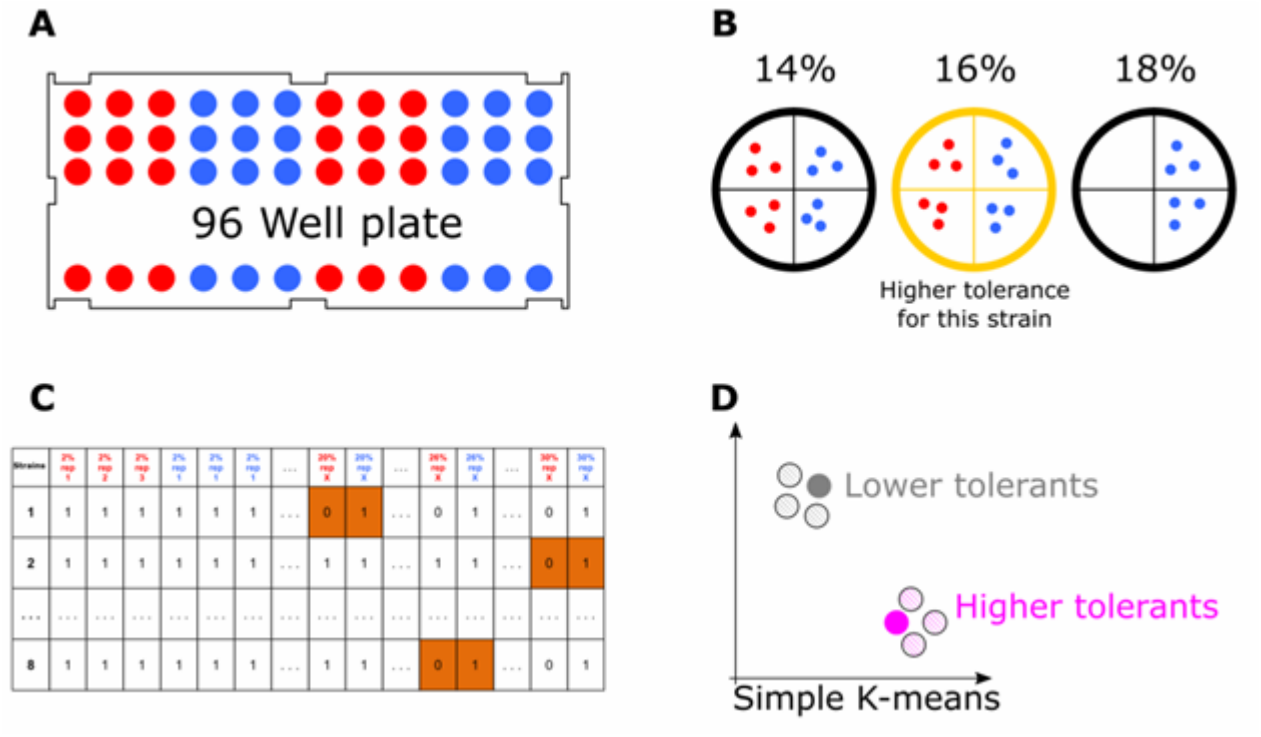

Tabela 1 - Descrição das linhagens utilizadas nesse estudo, resultados dos experimentos de tolerância ao etanol e aprendizado de máquina não supervisionado (ALMEIDA, 2017)

Linhagem	Maior nível de tol. EtOH	Grupo	Proteomas públicos	Empresa	Ploidia (locus MAT)
BMA64-1A	30%	HT	Não disponível	Euroscarf/20000A	a
BY4741	22%	LT	SGD/BY4741	Euroscarf/Y00000	a
BY4742	26%	HT	SGD/BY4742	Euroscarf/Y10000	alpha
SEY6210	20%	LT	SGD/SEY6210	NBRP/BY3553	alpha
X2180-1A	24%	HT	SGD/X2180-1A	NBRP/BY21559	a
S288c	20%	LT	SGD/S288c	NBRP/BY20118	alpha

Os RNAs e proteínas foram extraídos dos controles e tratamentos das linhagens listadas na **Tabela 1**. Os RNAs foram sequenciados usando Illumina HiSeq 2000 e após filtragens obteve-se no mínimo 30M de *reads* por biblioteca. As proteínas foram submetidas a um *shotgun-proteomics* (LC-MS/MS) (ALMEIDA, 2017; WOLF, 2019) utilizando o protocolo Almeida et al. (2019)

Com isso, foram gerados 36 transcriptomas e 36 proteomas, além do *time-course* para as linhagens BMA64-1A e S288c; todas as amostras de transcriptomas foram utilizadas para a identificação dos lncRNAs (ALMEIDA, 2017; WOLF, 2019).

3.2. Filtragem e identificação dos lncRNAs e micropeptídeos

Os *reads* das 36 bibliotecas de RNA-Seq foram filtrados por qualidade utilizando o programa Trimmomatic (BOLGER; LOHSE; USADEL, 2014) (parâmetros: Q-score ≥30, *read trimming* e exclusão de adaptadores) (WOLF, 2019). Posteriormente, seguiu-se o *pipeline* desenvolvido nesse trabalho, descrito na **Figura 4**, para a identificação dos lncRNAs.

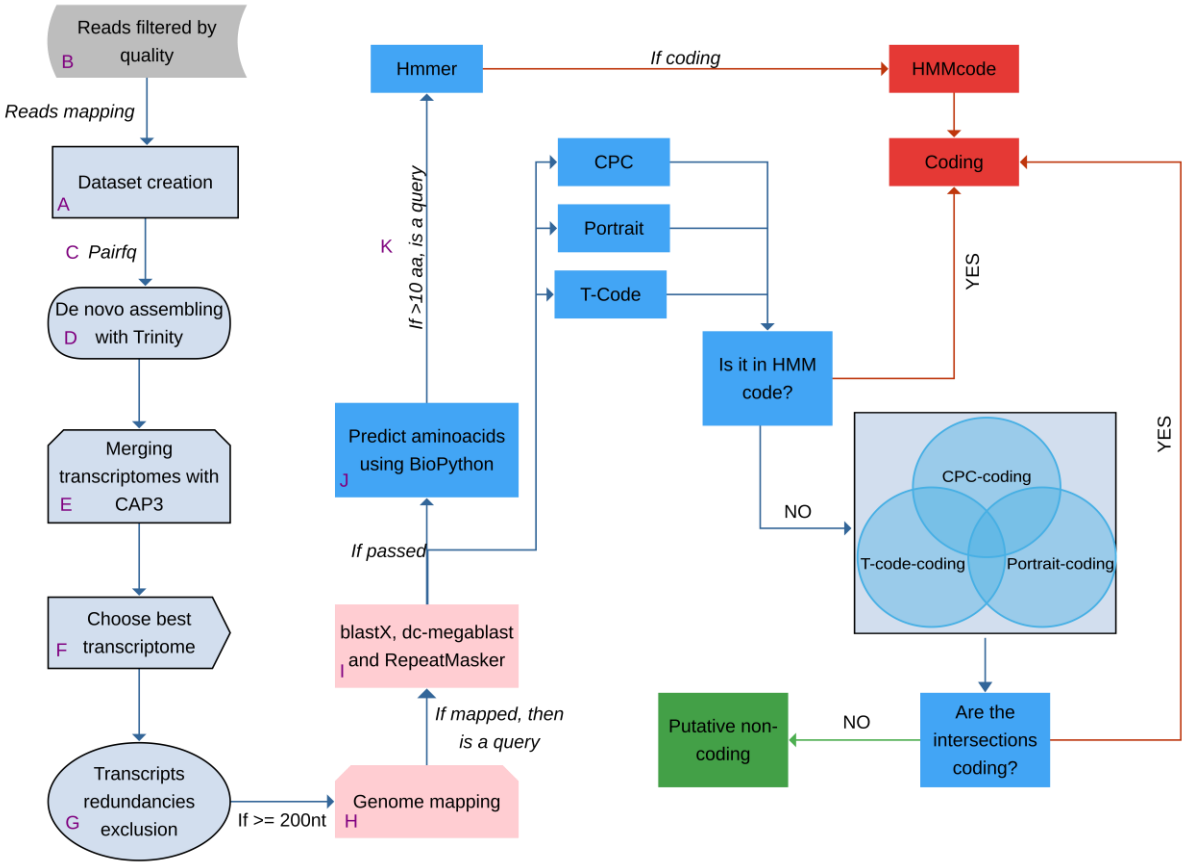

Figura 4 - Pipeline para identificação dos lncRNAs.

O *pipeline* inicia-se com a criação de um banco de dados com milhões de sequências incluindo: 1- CDS (sequências codificantes) e proteomas de eucariotos e bactérias; 2- genomas; 3- precursores de microRNA; 4- todas as famílias de ncRNAs no banco de dados do Rfam (KALVARI et al., 2018), sendo os lncRNAs as únicas exceções (número de acesso: Rfam 01884); 5- elementos móveis (**Figura 4B**): os dados descritos

estão abertos para acesso na tabela online disponível em: https://1drv.ms/x/s!Aju93ah3HMghguVbJaTvJ3P_VwnKyQ?e=jyOpI6, estando separados por “Espécie/linhagem, links dos genoma utilizados, CDS utilizados, proteoma utilizados, precursores de miRNAs e elementos móveis”. Depois, os *reads* filtrados pelo RNA-Seq advindos do Trimmomatic de cada linhagem foram alinhadas com as sequências de nucleotídeos do banco de dados criado usando o HISAT2 (KIM; LANGMEAD; SALZBERG, 2015) (parâmetros padrão) (**Figura 4A**). Os *reads* não alinhados foram separados e o *script* Pairfq (STATON; CHEF, 2016) (**Figura 4C**) foi utilizado para excluir *reads* que perderam um membro do par. Ao final, selecionamos apenas os pares de *reads* não-alinhados para a montagem *de novo* em virtude de serem *reads* sem similaridades com sequências codificantes, ncRNAs, elementos móveis e genomas contaminantes ou mitocondriais.

Para a montagem *de novo* (**Figura 4D**) dos *reads* filtrados, foi utilizada a estratégia de “*Single Assembler Multiple Parameters*” (HE et al. (2015). Para isso, os montadores Velvet/Oases (SCHULZ et al., 2012), Trinity (HAAS et al., 2013), IDBA-Tran (PENG et al., 2012) e rnaSPAdes (BANKEVICH et al., 2012) foram usados para determinar qual era o melhor algoritmo para esse conjunto de dados. Para o Velvet/Oases, rnaSPAdes e IDBA-Tran os *kmers* variaram de 19 a 81, porém para o Velvet/Oases e rnaSPAdes, foi adicionado também um *cutoff* de cobertura automática e opção sem *scaffolds*. Já para o Trinity, foi utilizado *kmers* de 19 a 31. Quando possível, todos os programas foram limitados a reportarem somente transcritos ≥200nts.

Após as montagens, todos os transcritos oriundos de um dado montador foram unificados em apenas um arquivo, as sequências ≥10% de nucleotídeos não-identificadas (“Ns”) foram removidas e uma remontagem foi feita usando o CAP3 (tamanho máximo de *gap* = 5, tamanho da porcentagem máxima de sobreposição = 20, ponto de parada do *clipping* = 1, porcentagem da identidade do ponto de corte = 80 e valor máximo da orientação reversa = 1) (HUANG, 1999) (**Figura 4E**). O procedimento de remontagem foi

realizado nas montagens oriundas de cada montador, independentemente, gerando ao final 1 transcriptoma por montador.

Após a execução de todos os passos anteriores, os *singlets* e *contigs* do CAP3 relativos a cada montador foram unificados e os *reads* usados nas montagens foram mapeados sobre tais montagens usando o Bowtie2 (parâmetros padrão) (LANGMEAD; SALZBERG, 2012). Um *score* comparando o número de sequências do *input*, número de sequências do *output*, tamanho médio dos transcritos e porcentagem de *reads* alinhados para cada remontagem do CAP3 foi calculado, permitindo elencar o Trinity como o melhor *assembler* tanto por apresentar um ganho final comparativamente alto, quanto por ser um algoritmo mais eficiente se tratando de uso de poder computacional (LIU et al., 2013) (os indicadores utilizados para alavancar como o Trinity como melhor montador estão disponíveis em https://1drv.ms/x/s!Aju93ah3HMghguVdxz29piiZUEC3bw?e=2bjlVy). Posteriormente, as redundâncias do melhor transcriptoma (oriundos do programa Trinity) foram removidas elencando apenas um representante por conjunto redundante. Para isso, foi usado o programa CD-Hit (identidade ≥ 98, porcentagem de alinhamento ≥ 99% e cobertura ≥ 99%) (FU et al., 2012) (**Figura 4G**). Todas as etapas supracitadas foram conduzidas usando os *reads* da linhagem S288c e, posteriormente, aplicou-se os melhores algoritmos e parâmetros para todas as outras linhagens.

Para selecionar apenas os transcritos reais e excluir possíveis quimeras ou erros de montagens, o conjunto de transcritos não-redundantes de cada linhagem foram mapeados sobre seus próprios genomas usando o GMAP (cobertura ≥ 90%, identidade ≥99%, -k15, -B5, p3, sem *close indels* e *no-chimeras*) (WU; WATANABE, 2005) (**Figura 4H**).

Posteriormente, verificou-se novamente o potencial dos transcritos mapeados serem codificantes, ncRNAs não-lncRNAs, repetitivos ou sequências de genoma de vírus,

bactéria ou mitocôndrias, mapeando tais transcritos nos mesmos bancos de dados utilizado para filtrar os *reads*. Nesse ponto, os transcritos foram: 1- alinhados contra os proteomas usando Blastx (E-value = 0,00001 e 1 alinhamento *per query*); 2- alinhados contra o banco de ncRNAs usando o dcmegablast (E-value = 10E-5 e *word-size* = 11); 3- os ncRNAs e as repetições simples também foram buscadas usando o RepeatMasker (*crossmatch*, similaridade = 10, -s, -gccalc, pular inserção bacteriana e *word-length* = 4000) (SMIT; HUBLEY; GREEN, 2013), sobre o mesmo conjunto de dados do dcmegablast; 4- os transcritos foram alinhados contra genoma de bactérias, vírus e mitocôndrias usando o GMAP (cobertura ≥ 50%, identidade ≥ 50%, -k15, -B5, p3, sem *close indels* e *no-chimeras*) (**Figuras 4I e 5**). Os resultados dessas buscas foram filtrados tal como descrito na **Figura 6**.

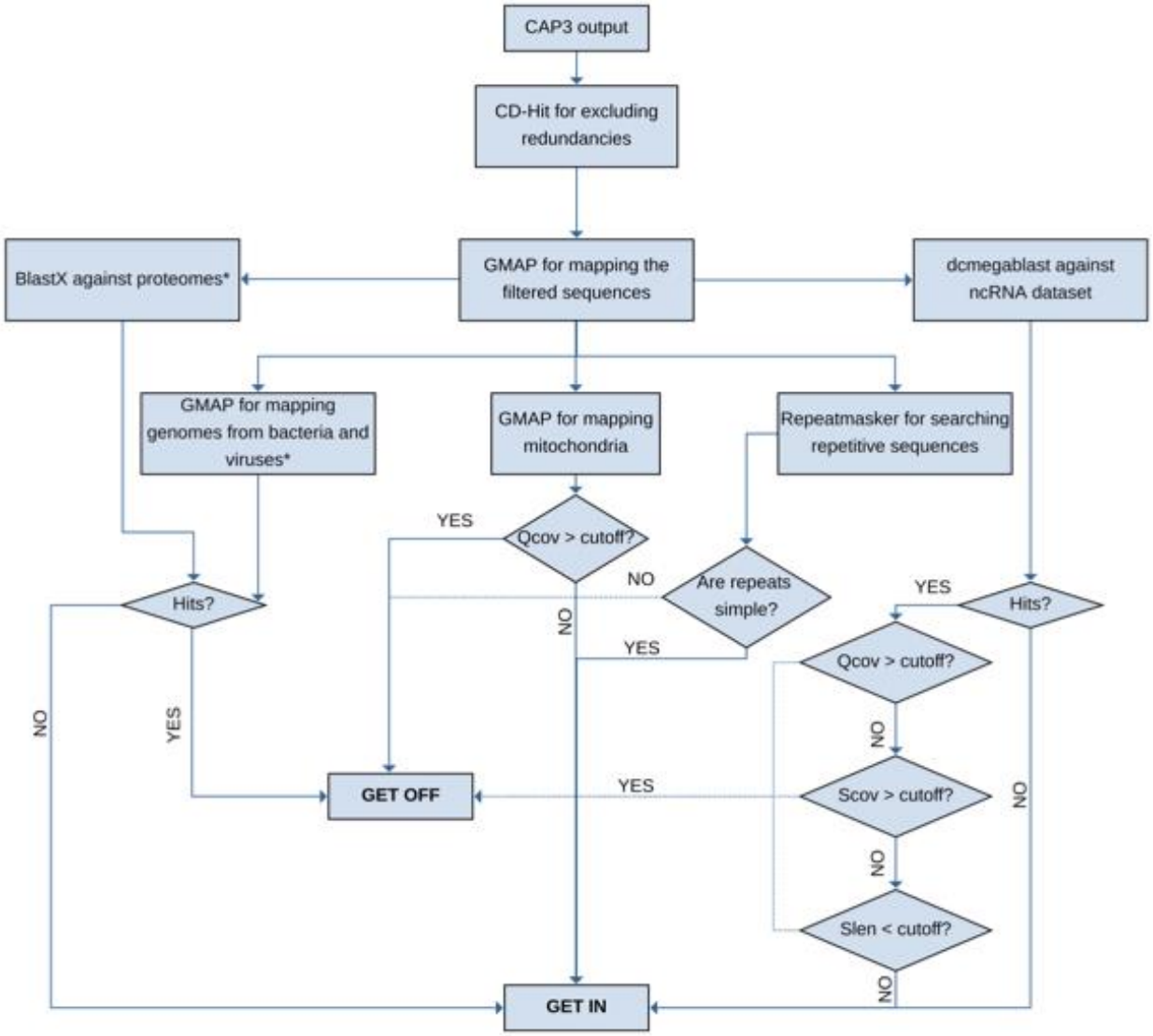

Figura 5 - Representação dos passos do algoritmo do CAP3 até a parte final do mapeamento dos transcritos. *, distribuição dos resultados com seus cut-offs estão detalhados apresentados em https://1drv.ms/x/s!Aju93ah3HMghgfcVcl3hL-sLHAYZ1w?e=G2WbEX.

Posteriormente, os programas Hmmer (JOHNSON; EDDY; PORTUGALY, 2010), Tcode (http://www.bioinformatics.nl/cgi-bin/emboss/help/tcode), Portrait (ARRIAL; TOGAWA; BRIGIDO, 2009) e CPC (KONG et al., 2007) foram usados para calcular o potencial de codificação das sequências selecionadas pelo passo anterior (**Figura 4K**). Para isso, os transcritos filtrados no passo **Figura 4K** foram traduzidos em proteínas (≥10 aa) usando o Getorf (RICE; LONGDEN; BLEASBY, 2000) (**Figura 4J**). Esses conjuntos de proteínas/transcritos foram submetidos ao Hmmer (usando a versão 31.0 do Pfam do http://pfam.xfam.org como bando de dados), Tcode, Portrait e CPC (todos nos parâmetros

padrão). Os transcritos que obtiveram *hits* no Hmmer, de acordo com os *cutoffs* reportado na **Figura 6** (https://1drv.ms/x/s!Aju93ah3HMghgfcVcl3hL-sLHAYZ1w?e=Omc7Dm), foram diretamente considerados como potencialmente codificantes. Para as outras sequências, se ao menos dois programas considerassem uma dada sequência sem potencial de codificação, de acordo com os *cutoffs* da tabela disponível em (https://1drv.ms/x/s!Aju93ah3HMghgfcVcl3hL-sLHAYZ1w?e=Omc7Dm), ela foi considerada como provável lncRNAs.

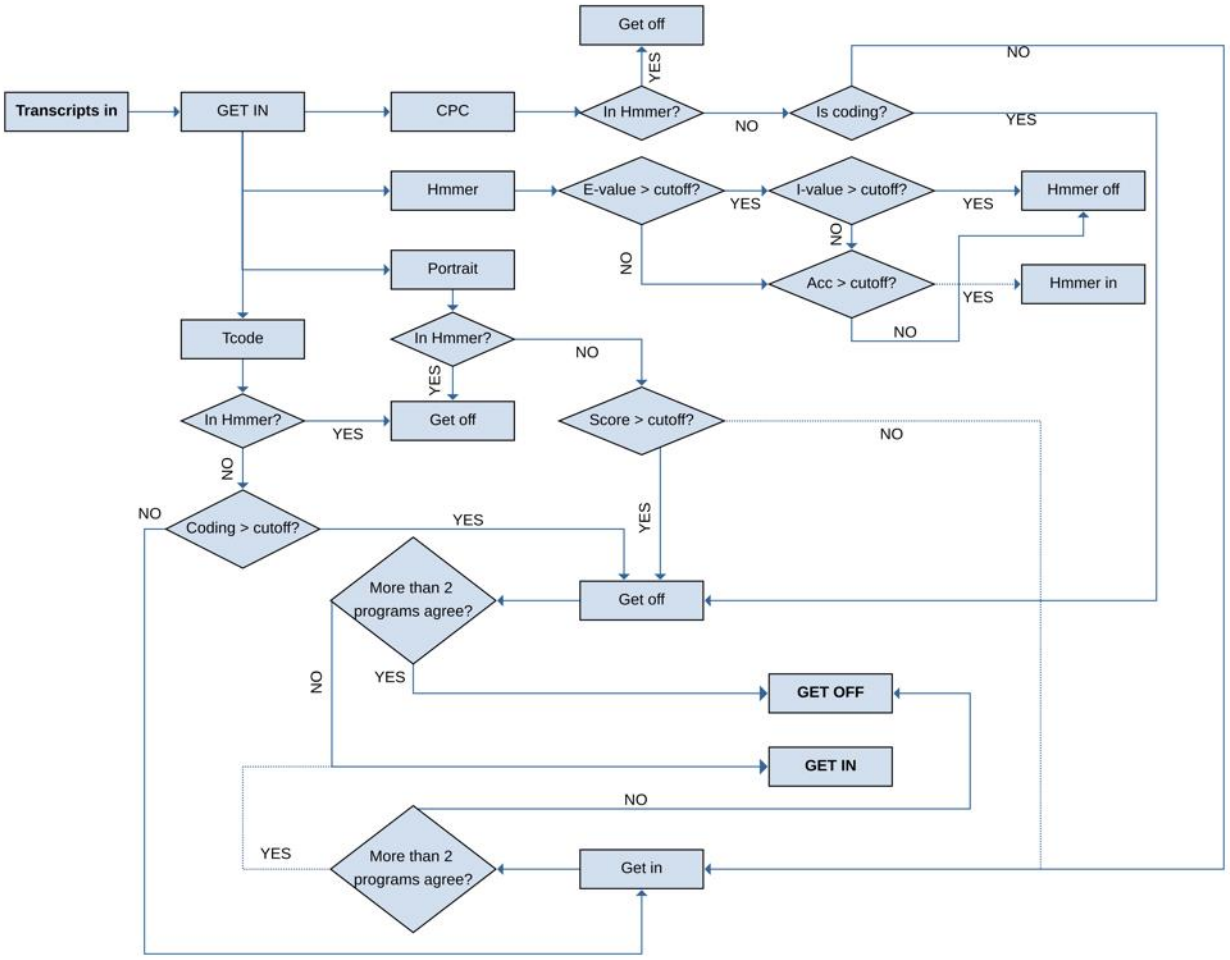

Figura 6 - Algoritmo do mapeamento geral dos transcritos e separação das prováveis moléculas não-codificantes. Transcritos “Hmmer off” foram considerados dentro da lista de prováveis RNAs longos não-codificantes. Transcritos “Hmmer in” foram comparados com o output dos outros programas utilizados. “Get in” - sequências consideradas lncRNAs. “Get off”, sequências codificantes.

3.3. Classificação dos prováveis lncRNAs e validação dos micropeptídeos

Os prováveis lncRNAs separados foram classificados de acordo com seus arranjos nos genomas seguindo as diretrizes do GENCODE v7 (DERRIEN et al., 2012) sendo então classificados como *lincRNAs, intronic, sense* e *antisense*; os lncRNAs que não se encaixam nessa classificação foram descritos como "*others*" (**Figura 7**). Foram assumidos então como: 1- *lincRNAs* os transcritos de *loci* intergênico (entre dois genes codificadores de proteínas); 2- *intronic* os transcritos localizados dentro da região intrônica dos genes codificadores de proteínas e sem intersecção com qualquer éxon; 3- *sense* caso um gene localiza-se dentro de um lncRNA sem nenhum éxon sobreposto; 4- *antisense* os lncRNAs localizados na fita oposta de um gene codificador de proteína e com uma intersecção com quaisquer éxons/íntrons; 5- *others* os casos que não se enquadram em nenhuma das classificações acima, como por exemplo um lncRNA o qual um dos seus éxons apresenta uma sobreposição com qualquer éxon de um gene codificador de proteína na mesma fita.

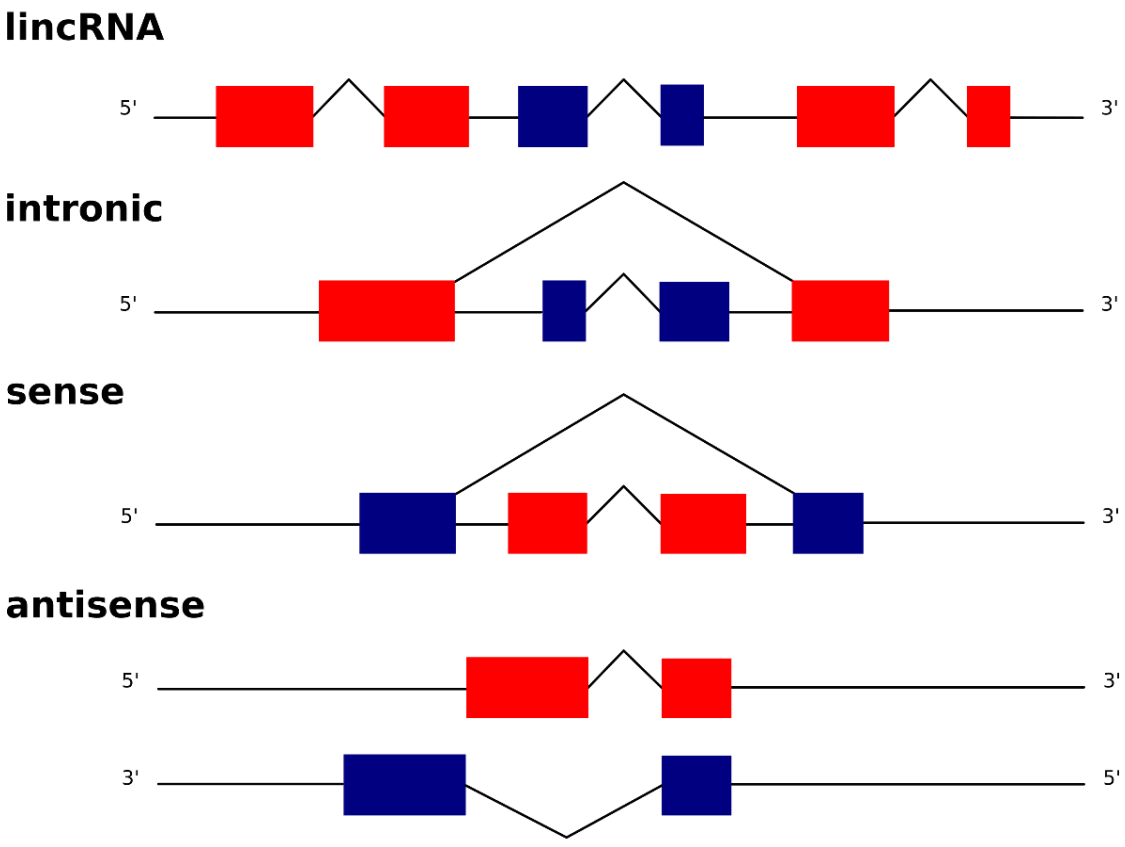

Figura 7 - Descrição das classificações dos lncRNAs. Os blocos vermelhos representam os éxons de genes codificadores de proteínas e os blocos azuis os éxons dos lncRNAs. As linhas pretas na horizontal são referentes à região genômica e as linhas pretas conectando os blocos representam os íntrons.

Além disso, os lncRNAs identificados com potencial de codificar micropeptídeos entre 10 a 99 aminoácidos foram classificados como prováveis lncRNAs codificantes de micropeptídeos. Ressalta-se que não houve nenhum caso de lncRNA que potencialmente codificasse uma proteína com mais de 100 aminoácidos, o qual poderia ser classificada como um suposto gene ainda não anotado.

Para certificar o real potencial dos lncRNAs codificarem micropeptídeos, foi feita a análise com base nos dados de proteômica descritos em Almeida, 2017 e, para isso, foi utilizado o programa Trans-Proteomic Pipeline (TPP) (DEUTSCH et al., 2010), usando o pipeline SpectaST (parâmetros no *default*). A análise de peptídeos foi feita pelo Peptide Prophet dentro do TPP (probabilidade ≤0,05, peptídeos mínimos considerados 7) e depois foram filtrados todos os resultados com probabilidade ≥0,6 e número de peptídeos ≥2.

O programa CD-Hit (FU et al., 2012) foi utilizado para fazer grupamentos de lncRNAs entre todas as linhagens com o intuito de verificar possíveis conservações em termos de sequências. Para isso, os lncRNAs identificados na linhagem S288c (por ser a linhagem com o maior número de lncRNAs identificados) foram utilizados como *queries* contra as outras linhagens utilizando parâmetros de identidade ≥80-99%. Assumiu-se como homólogos os lncRNAs de uma dada linhagem que agrupam com algum lncRNA da linhagem S288c.

3.4. Predição das interações lncRNAs-proteínas, análise de enriquecimento ontológico das proteínas-alvo e análise de expressão diferencial

As expressões diferenciais foram calculadas utilizando o programa DESEQ2 (LOVE; HUBER; ANDERS, 2014) para todos os genes e lncRNAs aqui encontrados e

anotados. Foi definido como transcritos diferencialmente expressos aqueles com *false discovery rate* (FDR) <0,01 quando comparado tratamento *vs. c*ontrole. Para as linhagens BMA64-1A e S288c, foram feitas também as contagens das expressões do *time course* (2 e 4 horas).

As predições de interação lncRNAs-proteínas para cada linhagem foram realizadas utilizando o programa lncPro (LU et al., 2013) e as interações com probabilidade ≥95% foram selecionadas. Posteriormente, mapeou-se manualmente as interações selecionadas dos lncRNAs diferencialmente expressos sobre os mapas do banco de dados do KEGG (KANEHISA et al., 2017). Estes mapas foram primeiramente montados de forma automatizada utilizando o pacote Pathview (LUO et al., 2009) e, para isso, Wolf (2019) utilizou duas abordagens: (I) somente genes diferencialmente expressos de forma exclusiva em HT e LT plotados separadamente e (II) todos os *fold-changes,* independente da significância estatística dos genes, foram plotados para todas as linhagens, individualmente, para obter-se uma visão geral sobre o estado das vias. Além disso, metabólitos diferencialmente abundantes e totais também foram plotados nos mapas.

Para identificar as prováveis funções dos lncRNAs aqui montados, foi usada uma estratégia de *guilt-by-association*. Primeiramente, redes de interação entre proteínas (PPIs) de *S. cerevisiae* foram baixadas do BioGrid (CHATR-ARYAMONTRI et al., 2015) e Mint (LICATA et al., 2012) e, após terem sido filtradas e unificadas, uma rede para cada linhagem foi criada (WOLF, 2019); ao final, essas redes PPIs com as redes lncRNA-proteínas já citadas foram unificadas. Então, esse arquivo unificado foi utilizado no programa SAFE (BARYSHNIKOVA, 2016) (limite de 0,5% da distância; limite de 75% usando o método de similaridade de *landscape* de Jaccard e tamanho mínimo de *landscape* = 10; *landscapes* multi-regionais removidos; *background* com todos nós da rede), para cada linhagem, independentemente, a fim de identificar as ontologias das comunidades as quais cada

lncRNA pertencem. Posteriormente, a plataforma REVIGO (*allowed similarity=0.7*) (SUPEK et al., 2011) foi utilizada para sumarizar as ontologias dessas comunidades.

3.5. Busca por ortologias dos lncRNAs nas diferentes linhagens e avaliação das semelhanças estruturais

Para checar possíveis conservação genômicas dos lncRNAs entre as diferentes linhagens aqui estudadas utilizou-se uma abordagem de análise das sintenias das vizinhanças genômicas. Para isso, uma busca dos genes codificadores de proteína com ao menos 1 nucleotídeo de sobreposição aos lncRNAs foi realizada manualmente visualizando os dados no programa Artemis (CARVER et al., 2012). Além disso, a seleção dos vizinhos localizado até 1.500 nts *up-stream* ou *down*-stream foi realizada utilizando scripts *in-house* e o pacote closestBED da suíte de aplicativos BEDTools (ignorando sequências que se sobrepõe, separando distâncias de até 1500 nts na mesma fita e mostrando somente o primeiro resultado) (QUINLAN; HALL, 2010); essas anotações de vizinhos foram realizados somente para as linhagens S288c e BMA64-1A. Por fim, dois vizinhos codificantes de cada lncRNA da linhagem S288c foram buscados em todas as outras linhagens a fim de checar se nelas esses marcadores também possuem algum lncRNA como vizinho.

Com o objetivo de avaliar possíveis conservações funcionais dos lncRNAs nas diferentes linhagens aqui estudadas, a abordagem de avaliação das estruturas secundárias foi utilizada. Para isso, foram selecionados quatro representantes de uma linhagem HT (BMA64-1A) e quatro representantes de uma linhagem LT (S288c) que interagem com a mesma proteína (YLR106C) para fazer uma comparação intra-linhagem (**Tabela 2**); esse gene (YLR106C) foi escolhido devido a sua importância na viabilidade celular (BASSLER et al., 2010; MATSUO et al., 2014) e pela grande quantidade de lncRNAs interagindo com ele (um total 87 lncRNAs, considerando todas as linhagens). Também serão analisados os genes YER008C e YPL242C e as ligações desses com os lncRNAs por conta do alto

número de ligações que eles recebem dos lncRNAs (57 e 49 lncRNAs respectivamente considerando todas as linhagens) e suas importâncias biológicas. As comparações inter-linhagem foram feitas escolhendo apenas um representante para cada linhagem. Então foi utilizada a ferramenta CROSSalign (DELLI PONTI et al., 2017, 2018) (parâmetros *default*) para comparar os perfis das estruturas secundárias das prováveis moléculas de lncRNAs selecionadas. Esse programa calcula um *Structural Distance Score (SDS)* entre os perfis gerados, sendo que quando SDS ~0, maior é a similaridade em quesito de estrutura secundária. Ambos eixos mostram o *CROSS Global Score (CGS)* para ambas estruturas comparadas, sendo que o *CGS* >0 indica um nucleotídeo duplicado e *CGS* <0 indica um nucleotídeo não-duplicado. Então, *CGS* ~1 indica uma maior quantidade de nucleotídeos duplicados portanto, perfis mais similares, ao passo que valores ~-1 indicam uma maior quantidade de nucleotídeos não-duplicados, ou seja, perfis mais distintos. O algoritmo calcula também um *p-value* sendo que *p*<0,05 representa significância estatística dos *scores*. No geral, *SDS* >0,10 são considerados diferentes em termos de estrutura secundária quando *p-value <=0,05.*

Tabela 2 - lncRNAs com potencial de interagir com o gene YLR106C separados por linhagem. Os lncRNAs em itálico são aqueles utilizados para a comparação inter-linhagens.

Linhagem	**Nome do lncRNA**
BMA64-1A	Transcr_6240, *Transcr_19942*, Transcr_20548, Transcr_22010
BY4742	*Transcr_10027*
BY4741	*Transcr_3737*
S288c	*Transcr_2076*, Transcr_12508, Transr_13050, Transcr_21244
SEY6210	*Transcr_1139*
X2180-1A	*Transcr_6988*

4. Resultados

4.1. Identificação dos lncRNAs e análises de expressão gênica

Foram isolados um total de 377 prováveis lncRNAs com potencial capacidade de codificarem micropeptídeos e 638 prováveis lncRNA completamente não codificantes, considerando as 6 linhagens estudadas (**Tabela 3**); foram também considerados completamente não codificantes também aqueles lncRNAs cujas *Open Reading Frames* (ORFs) não começavam com o aminoácido metionina. Quanto ao tamanho dos lncRNAs isolados para todas as linhagens, observa-se uma distribuição que segue uma lei-de-potência, contudo, a maioria deles possuem tamanho na faixa entre 200-400 nts, independente da linhagem analisada (**Figura 8**).

Tabela 3 - Prováveis lncRNAs identificados por linhagem. *, dados contrapostos com as informações de proteômica via espectrometria de massas (ALMEIDA, 2017).

Linhagem	N. total de lncRNAs identificados	N. de lncRNAs potencialmente codificantes de micropeptídeos*	N. de lncRNAs completamente não codificantes
BMA64-1A	227	78	149
BY4742	147	69	78
BY4741	129	52	77
S288c	259	91	168
SEY6210	120	49	71
X2180-1A	87	34	53

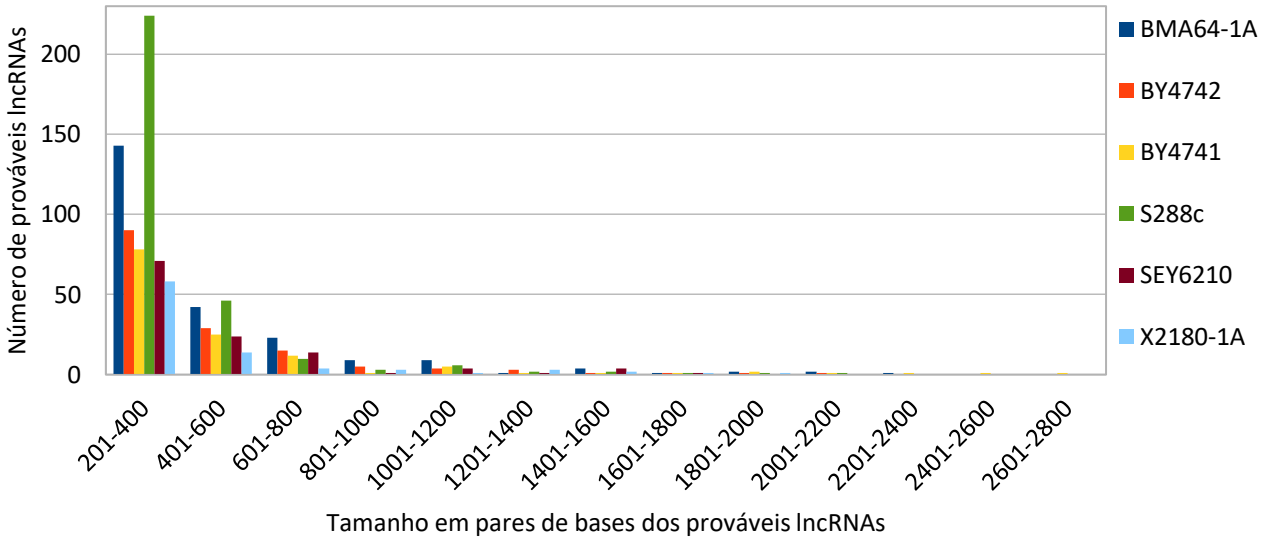

Figura 8 - Distribuição do tamanho dos lncRNAs anotados.

As possíveis ORFs dos lncRNA com potencial de codificar micropeptídeos (**Tabela 3**) também foram anotadas no GFF de cada linhagem. A checagem do potencial de codificação de micropeptídeos com base nos dados de proteômica oriundos de espectrometria de massas (ALMEIDA, 2017) revelou não haver nenhum *hit* com as amostras obtidas e, portanto, com os dados disponíveis, provavelmente esses lncRNAs não codificam micropeptídeos.

A classificação dos biotipos de todos lncRNAs recuperados (**Tabela 4**) revela que grande parte deles é classificada como '*others*', por terem pelo menos uma pequena interseção com o éxon de algum outro gene na mesma fita ou por estarem no final de um *contig* ou de um cromossomo.

Tabela 4 - Classificação dos biotipos separados por linhagem.

Biotipo	BMA64-1A	BY4742	BY4741	S288c	SEY6210	X2180-1A
Sense	0%	0%	0%	1%	0%	0%
Antisense	14%	4%	6%	20%	8%	4%
Intronic	0%	0%	0%	0%	0%	0%
lincRNA	1%	0%	1%	1%	0%	0%
Other	85%	96%	93%	78%	92%	96%

Um total de 273 lncRNAs foram diferencialmente expressos comparando-se tratamento *vs*. controle, mostrando uma tendência dessas moléculas serem *up-reguladas* num tratamento com estresse máximo de etanol (**Tabela 5**).

Tabela 5 - Número de lncRNAs diferencialmente expressos para cada linhagem.

	BMA64-1A	BY4742	BY4741	S288c	SEY6210	X2180-1A
Down	11	8	14	28	15	10
Up	20	27	52	48	22	18

A comparação da expressão diferencial desses lncRNAs com genes codificadores de proteína de cada linhagem (**Figura 9**) demonstra que os lncRNAs geralmente são menos expressos do que os genes codificantes das mesmas linhagens. Essa menor expressão pode ser observada também na **Figura 10**, onde todos os *Fold-Changes (FC)* com seus respectivos *p-values* foram esboçados, mostrando uma maior concentração na faixa entre -1,5 e 1,5 de FC para os lncRNAs de todas as linhagens juntas.

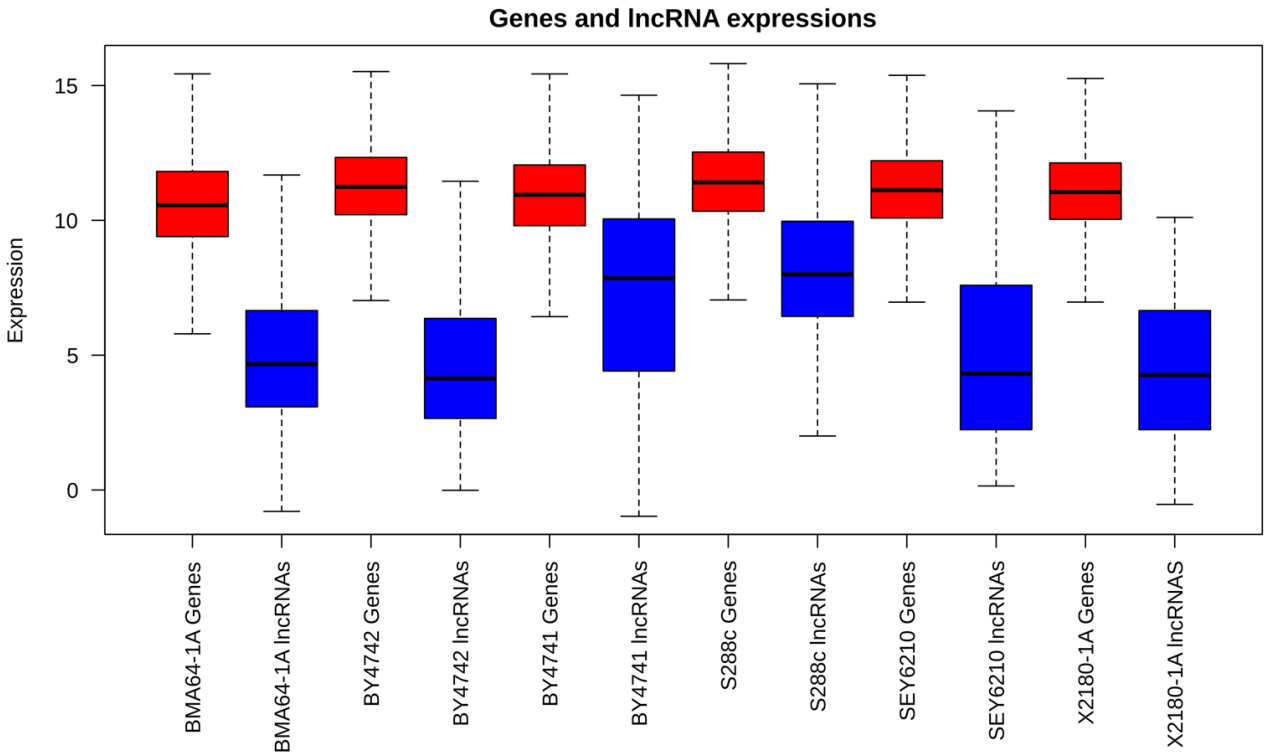

Figura 9 - Expressão dos *long-noncoding RNAs* comparados com a expressão dos genes codificantes.

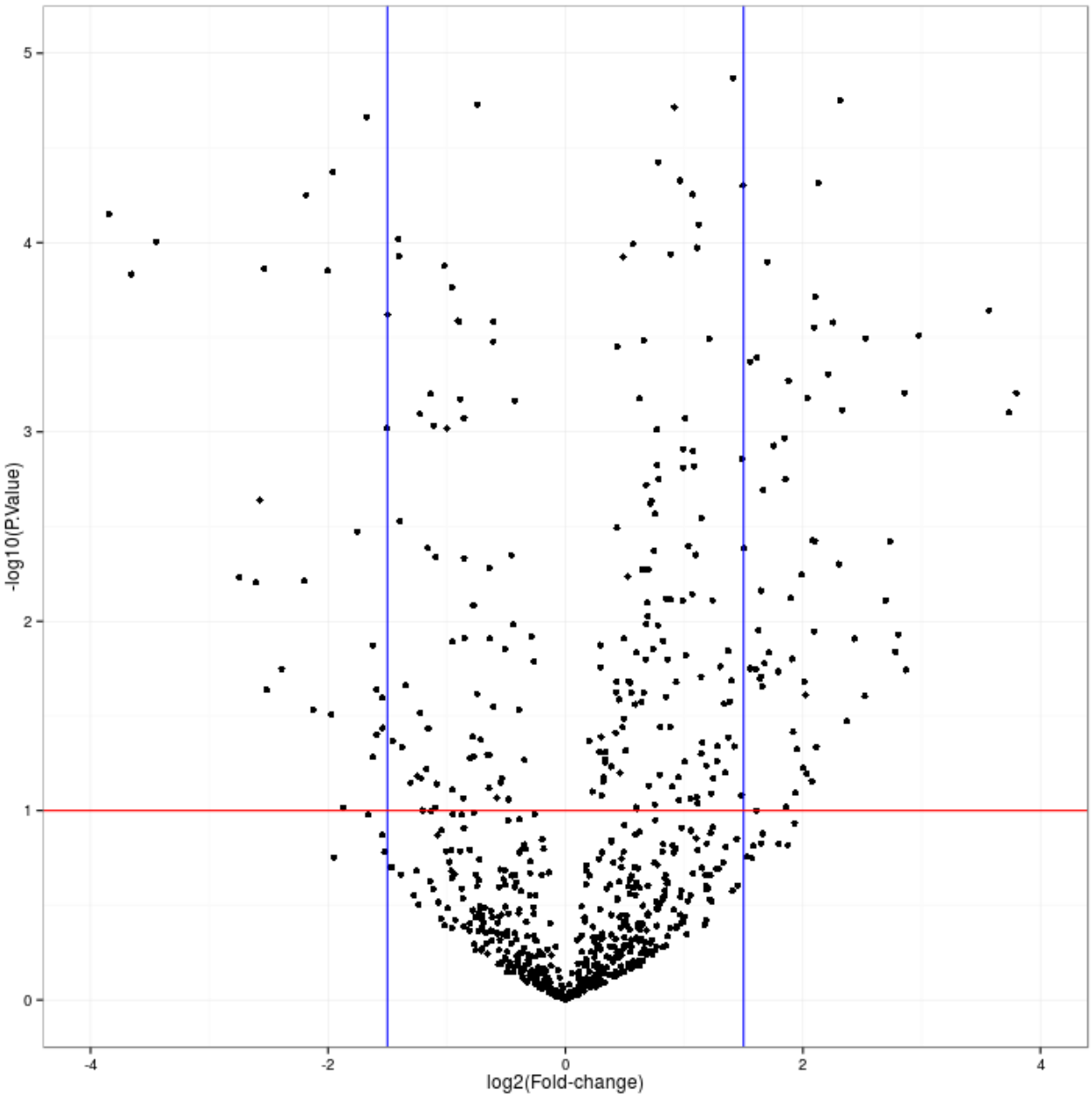

Figura 10 – Dispersão dos *FC* com seus respectivos *P-values* dos lncRNAs de todas as linhagens.

4.2. Análises para inferências funcionais e evolutivas dos lncRNAs

4.2.1. Assinando possíveis funções aos lncRNAs por guilt-by-association

Após a identificação das possíveis interações de cada lncRNA geradas pelo programa lncPro (*cutoff* ≥0,95 de probabilidade, estabelecido em virtude das probabilidades possuírem uma distribuição normal) (**Tabela 6**), as proteínas-alvo dos lncRNAs exclusivas das linhagens HT e exclusivas das linhagens LT foram separadas (**Figura 11**).

Tabela 6 - Número de possíveis interações lncRNA-proteína com probabilidade ≥95%.

Linhagem	N. de lncRNAs que possuem ao menos uma interação	N. de proteínas-alvo	N. de possíveis interações do lncRNA com as proteínas-alvo
BMA64-1A	36	159	244
BY4742	25	200	386
BY4741	17	230	286
S288c	44	394	726
SEY6210	20	174	233
X2180-1A	17	86	153

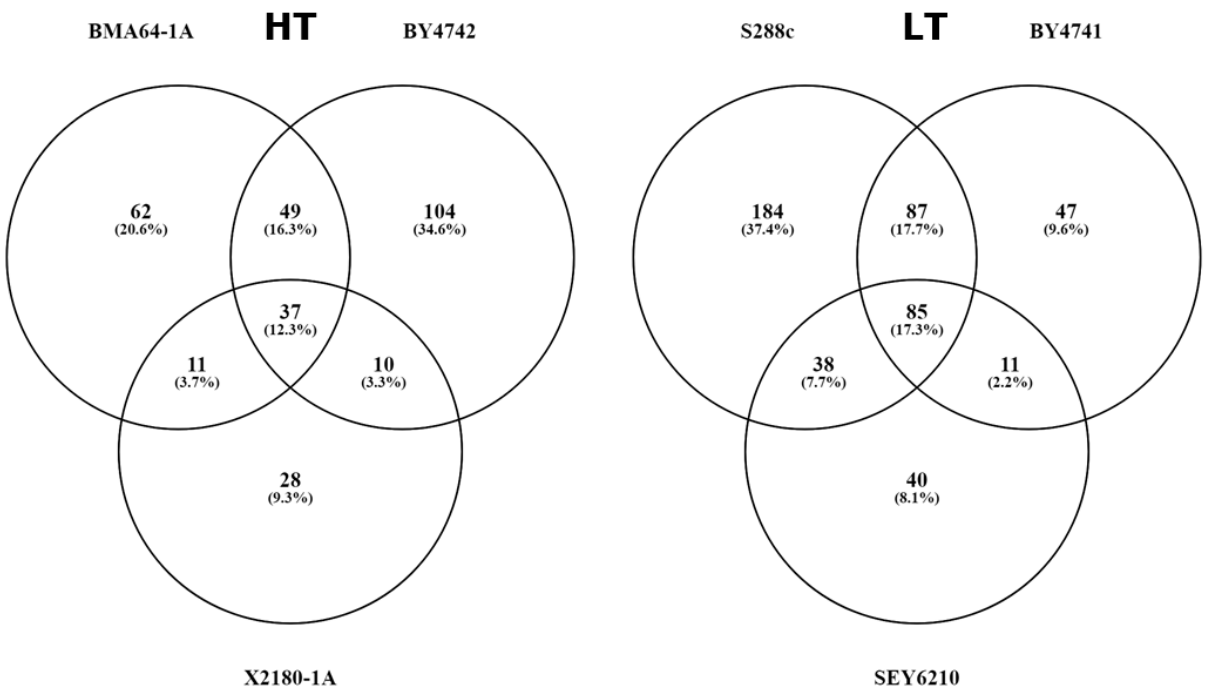

Figura 11 - Diagrama de Venn reportando as quantidades de proteínas-alvo dos lncRNAs exclusivas das linhagens HT e LT.

A análise das ontologias das proteínas-alvo exclusivas de HT (37 proteínas) e exclusivas de LT (85 proteínas) gerados pela análise via SAFE e sumarizados pelo REVIGO, reportam que para as HTs existem termos com relações metabólicas (*ubiquitin-dependent protein catabolism, metabolism)*, estruturais (*actin filamento reorganization, cell*

division) (**Figura 12**). Já em LTs, há termo relacionado à transporte (*vacuolar transport*) e grande concentração de termos relacionados à metabolismo (*lipid metabolismo, response to ionizing radiation, protein sumoylation, cellular metabolism*) (**Figura 13**).

Foi possível mapear nos mapas do KEGG (com base nas redes lncRNAs-proteínas) um total de 46 lncRNAs em um total de 153 mapas por linhagem (**Tabela 7**), sendo que existem possíveis lncRNAs que se repetem em vários mapas. Um exemplo disso é o transcr_18666 da linhagem LT S288c, o qual está presente 72 vezes nos mapeamentos. Um outro exemplo é o transcr_3746 da linhagem HT X2180-1A, o qual está presente 37 vezes nos mapas.

Tabela 7 - Síntese do mapeamento dos lncRNAs nas vias do KEGG Pathways. Os mapas estão disponíveis em https://1drv.ms/u/s!Aju93ah3HMghgckhY39BfKtVkt2Jag?e=sfOp5H. Nas figuras do link aqui citado, a faixa de cores de azul até roxo indica o grau de expressão diferencial dos genes, a faixa de cores verde até amarelo indicam a abundância dos metabólitos plotados, os quadrados laranjas mostram o nome do lncRNA mapeado sendo que o sinal '+' indica que ele está *up*-regulado e o sinal '-' indica que ele está *down*-regulado.

Linhagem	**N. de lncRNAs únicos dentro dos mapas**	**N. de mapas os quais os lncRNAs foram alocados**
BMA64-1A	6	15
BY4742	8	30
BY4741	7	25
S288c	10	37
SEY6210	8	20
X2180-1A	8	26

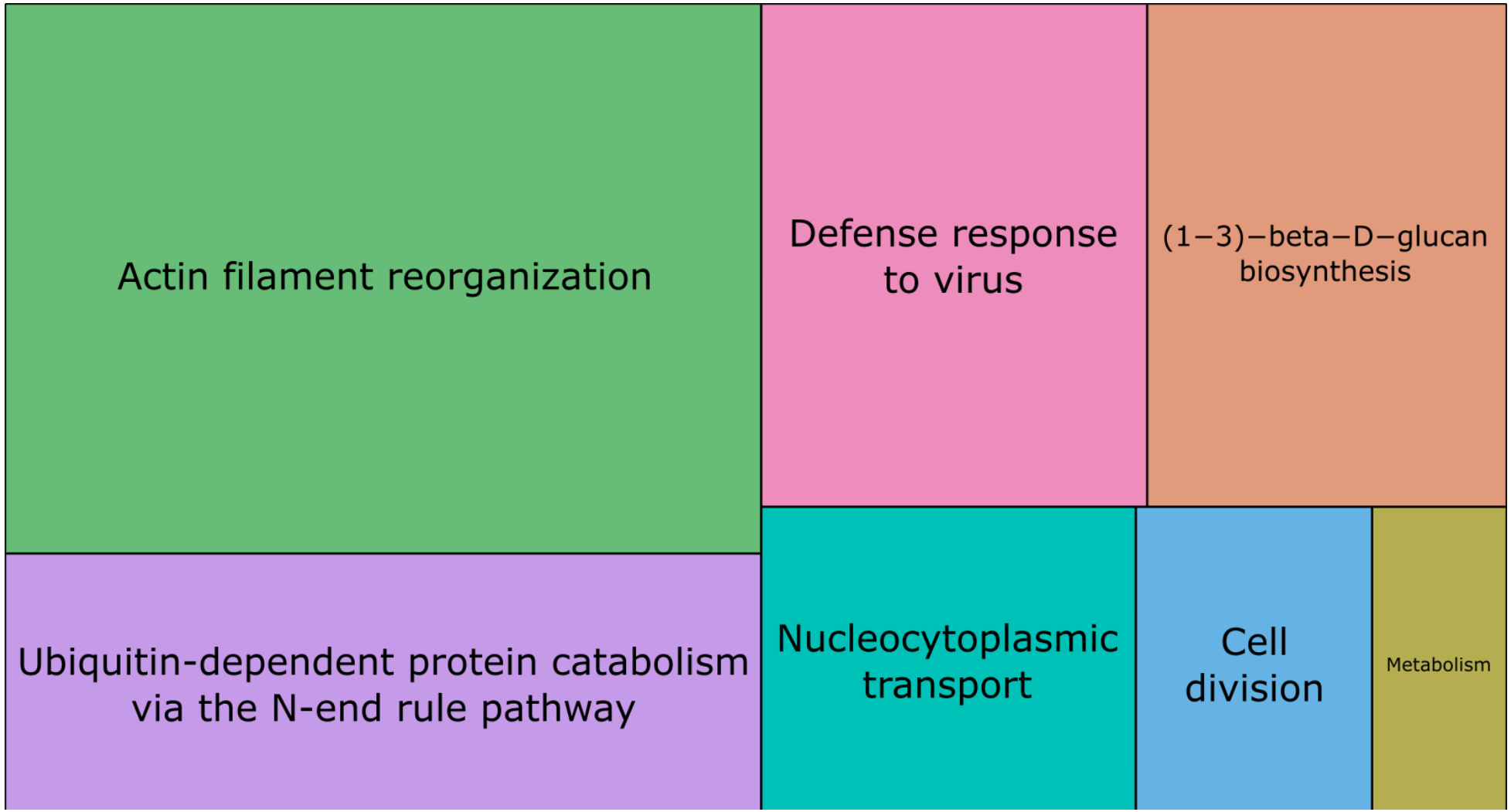

Figura 12 - Processos biológicos gerados pelo REVIGO com base nos Gene Ontology (GOs) das proteínas-alvo para HTs.

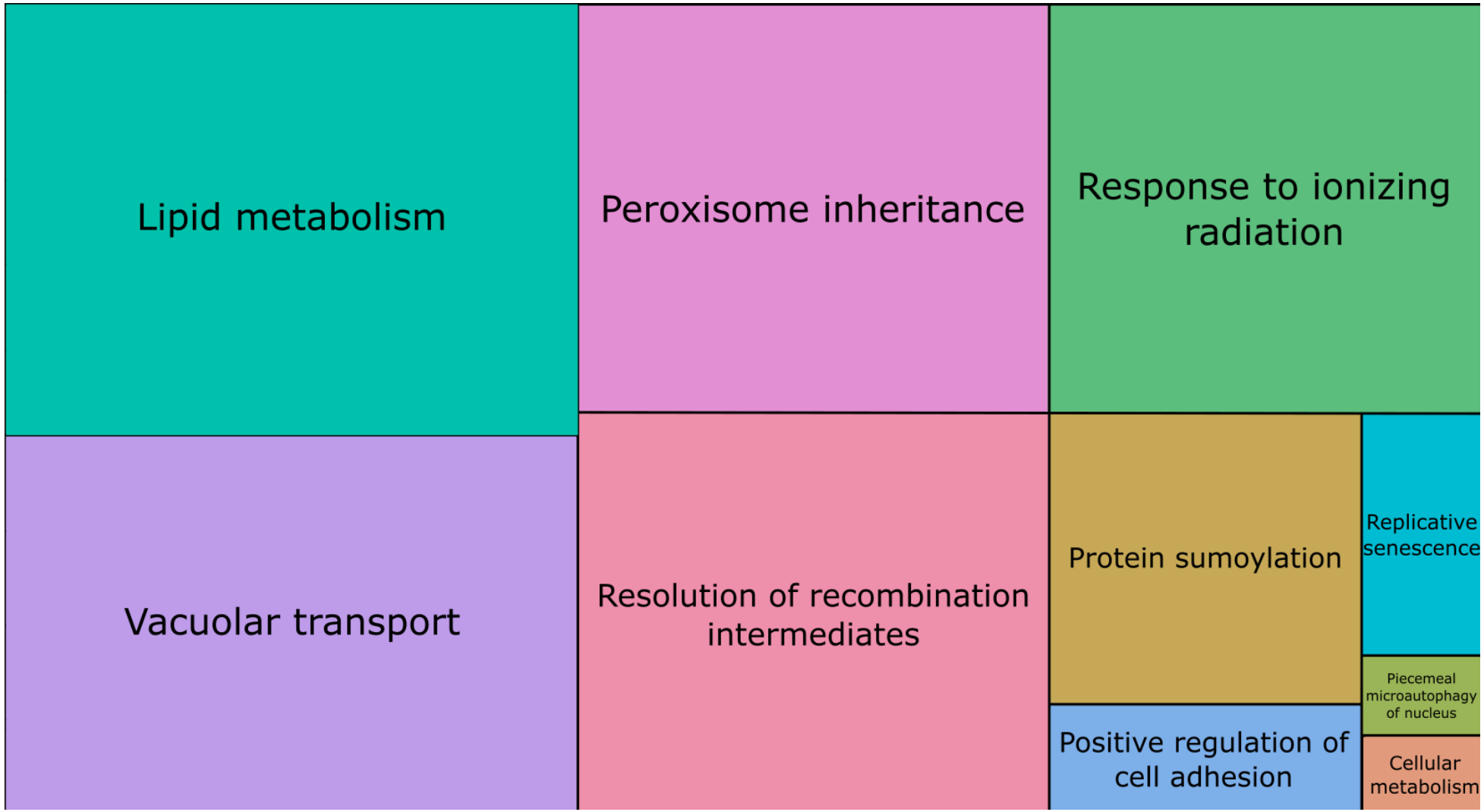

Figura 13 - Processos biológicos gerados pelo REVIGO com base nos GOs das proteínas-alvo para LTs.

4.2.2. Avaliação da conservação (colocar outro termo, como conservação das sequências) dos lncRNAs e busca de regiões sintênicas conservadas

Quando a conservação dos lncRNAs entre as linhagens é analisada tomando como base as sequências identificadas na linhagem S288c, é possível observar que os lncRNAs têm baixa conservação em termos de sequência, sendo que apenas um lncRNA da S288c está presente em todas as linhagens (transcr_22584), considerando uma identidade mínima de 80%. Ao restringir a identidade mínima para 99%, o número de lncRNAs similares aos observados na S288c nas outras linhagens diminui para apenas 7 transcritos (**Figura 13**).

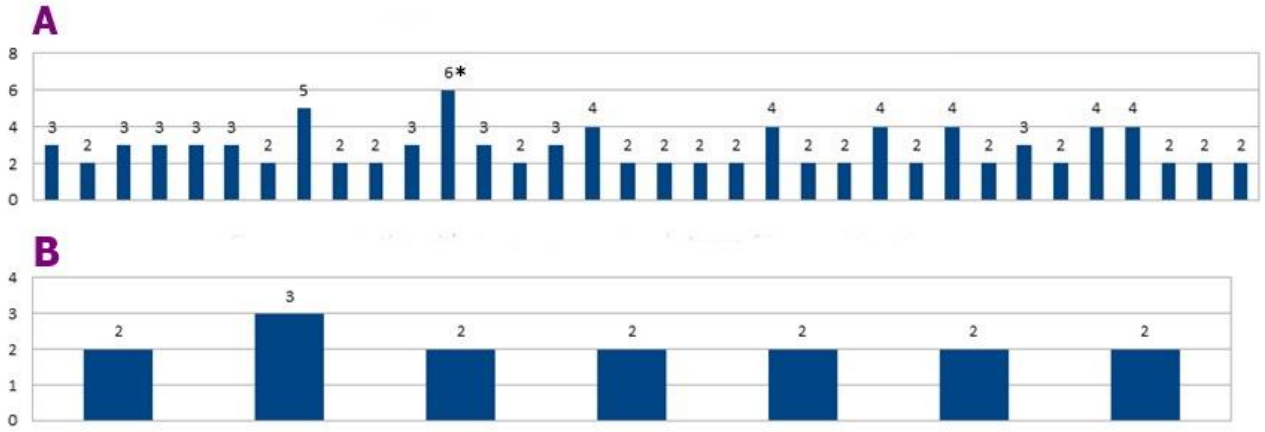

Figura 14 - Análise de agrupamento para identificar o número de sequências similares entre as linhagens quando comparados com a linhagem S288c. Cada barra representa um lncRNA e o número acima delas representa em quantas linhagens está presente tal lncRNA. A, agrupamento baseado em sequências com 80% de identidade entre si. B, agrupamento baseado em sequências com 99% de identidade entre si. *, lncRNA presente em todas linhagens.

A falta de conservação entre as linhagens também é observada ao comparar as estruturas secundárias dos lncRNAs que se ligam ao gene YLR106C. Como mostrado na **Figura 15** e **16**, quando compara-se um lncRNA da S288c (uma linhagem LT) e da BMA64-1A (uma linhagem HT) com os lncRNAs de todas as outras linhagens, é possível notar que os perfis não são similares um vez que os *scores* variam em entre 0,103-0,214 e os *p-values* variam entre 0,48-1; contudo, essas estruturas são diferentes, corroborando a falta

de conservação da estrutura conservadas já reportada para outros organismos (NITSCHE; STADLER, 2017). A falta de conservação estrutural também é observada nas comparações intra-linhagens tanto nas linhagens S288c como na BMA64-1A (**Figura 17** e **18**). Na S288c, os *scores de perfil* variam entre 0,102-0,200 e contendo *p-values* entre 0,45-1, sendo então estruturas não-similares.

A falta de semelhanças estruturais pode ser observada em outros genes importantes para a célula. Usando a mesma metodologia descrita para o gene YLR106C, foram comparados os lncRNAs que interagem com o gene YER008C, gene este ligado diretamente à processos de exocitose (HAARER et al., 1996; TERBUSH et al., 1996). Os resultados reportam também uma ausência de conservação estrutural, obtendo-se *scores* variando de 0,099-0,176 e *p-values* de 0,26-1. Algo similar pode ser observado também nos lncRNAs interatores do gene YPL242C, o qual codifica uma proteína essencial para o padrão de agregação da levedura (EPP; CHANT, 1997); *scores* de perfil de 0,101-0,114 e *p-value* de 0,41.

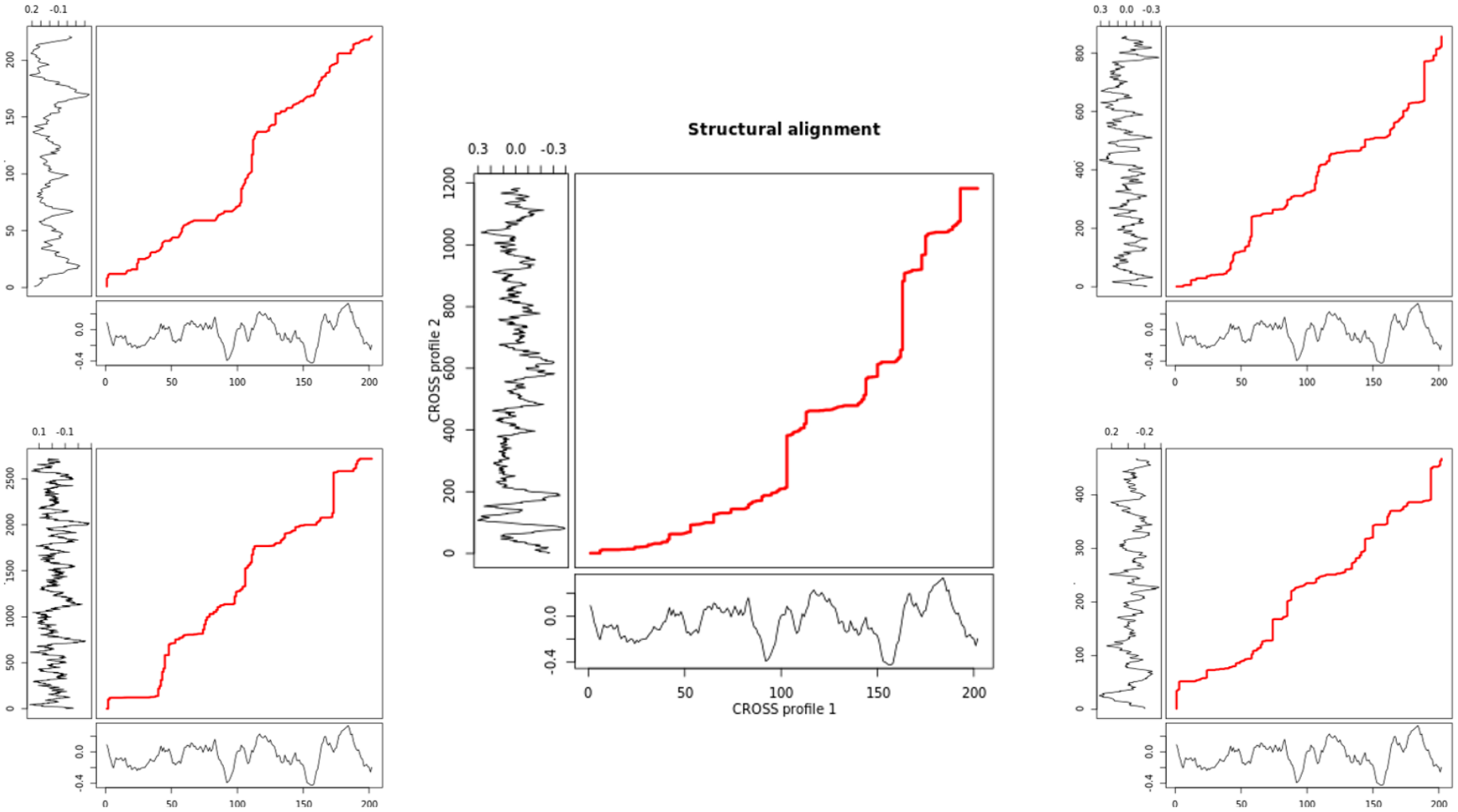

Figura 15 – No centro está o alinhamento estrutural entre os lncRNAs S288c *vs.* BMA64-1A. À esquerda está o alinhamento S288c *vs.* LTs. À direita está o alimento S288c *vs.* HTs. Nos dois eixos de cada gráfico estão os perfis estruturais obtidos com pelo *CROSS Score Global* para os dois lncRNAs, que significa: *score* > 0 é um nucleotídeo duplicado; *score* < 0 é um nucleotídeo não-duplicado.

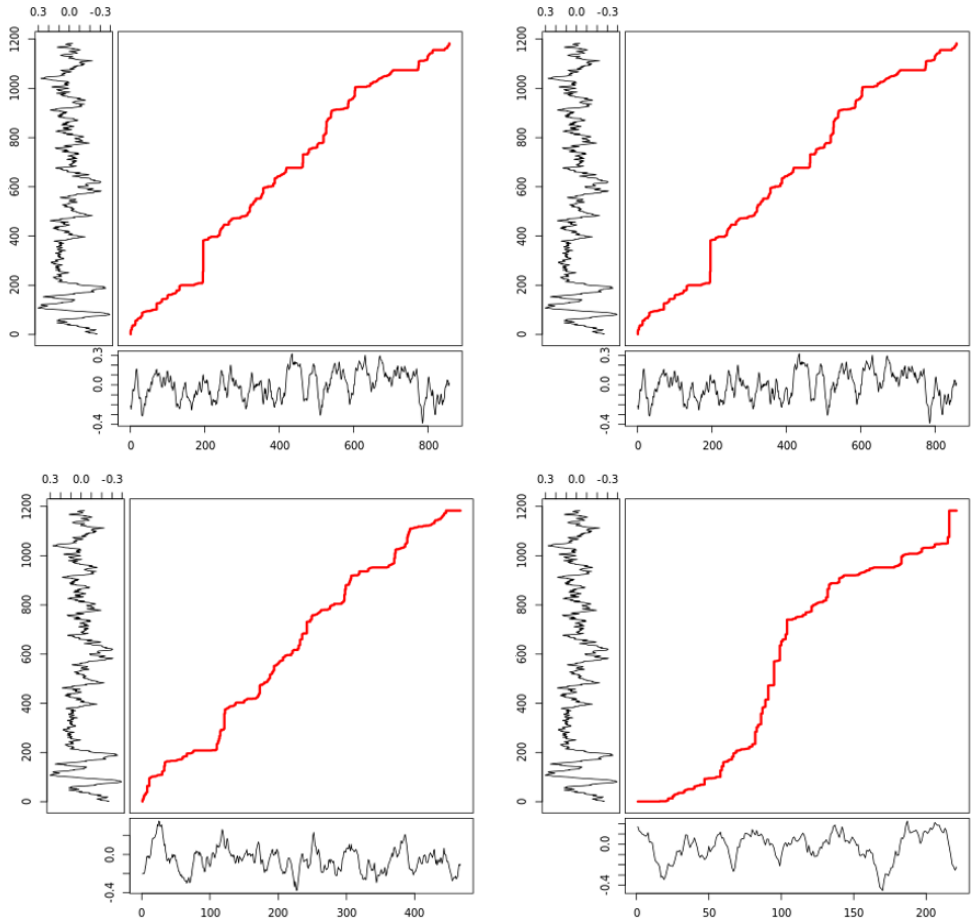

Figura 16 - À esquerda, o alinhamento BMA64-1A vs. LTs. À direita, o alimento BMA64-1A vs. HTs. Nos dois eixos de cada gráfico estão os perfis estruturais obtidos com pelo *CROSS Score Global C* para os dois lncRNAs, que significa *score* > 0 é um nucleotídeo duplicado e *score* < 0 é um nucleotídeo não-duplicado.

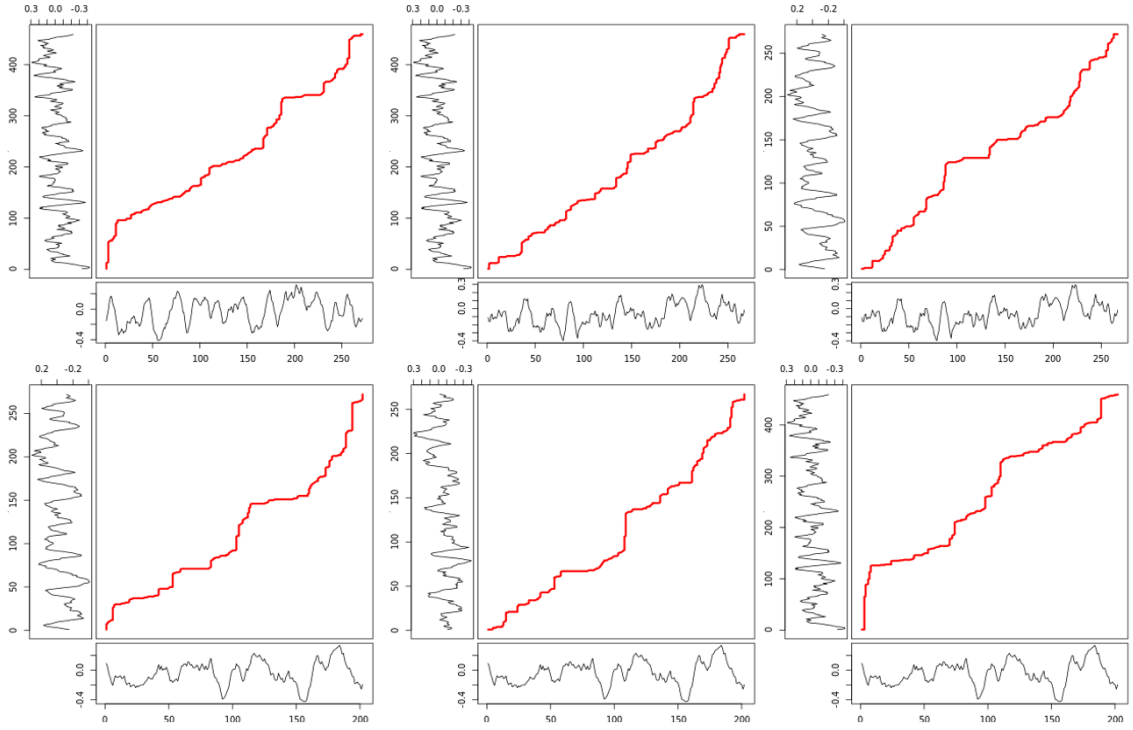

Figura 17 - Os perfis das quatro estruturas secundárias dos lncRNAs da S288c comparadas entre si. Nos dois eixos de cada gráfico estão os perfis estruturais obtidos com pelo *CROSS Score Global* para os dois lncRNAs, que significa *score* > 0 é um nucleotídeo duplicado e *score* < 0 é um nucleotídeo não-duplicado.

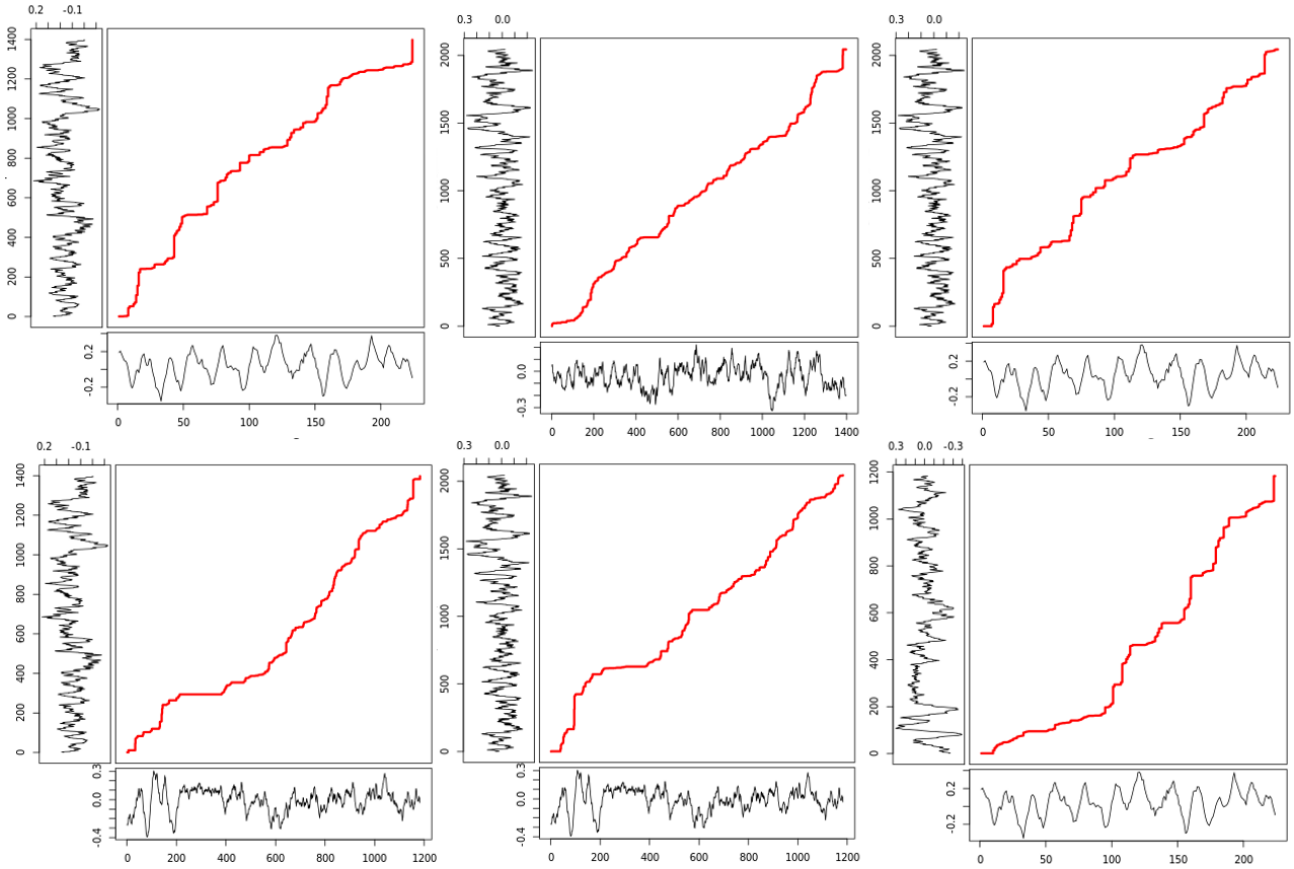

Figura 18 - Os perfis das quatro estruturas secundárias dos lncRNAs da BMA64-1A comparadas entre si.

Por fim, os lncRNAs não apresentam sintenias conservadas uma vez que, tomando como base sempre dois genes codificantes marcadores da linhagem S288c que possuem um lncRNA vizinho ou um lncRNA em sobreposição com um gene, não se observou presença de lncRNAs nas proximidades, 1500 nts *up* e *down*, desses marcadores em outras linhagens. Fazendo as análises de sintenia, foram identificados 110 genes com pelo menos 1 nucleotídeo de sobreposição com os prováveis lncRNAs em todas as linhagens, além de um total de 483 genes vizinhos aos lncRNAs; todos com a mesma orientação dos lncRNAs. Os genes com sobreposição se concentram nas linhagens BMA64-1A e S288c, e os genes vizinhos são mais distribuídos entre as diferentes linhagens (**Tabela 8**).

Tabela 8 - Número de genes com pelo menos um nucleotídeo de sobreposição aos lncRNAs identificados e dos genes com distâncias de até 1.500 nt up-stream ou down-stream ao loci não-codificante.

Linhagem	N. genes com sobreposição	N. genes vizinhos
BMA64-1A	17	105
BY4742	7	98
BY4741	7	69
S288c	68	96
SEY6210	6	58
X2180-1A	5	57

5. Discussão

5.1. Os lncRNAs são de diferentes biótipos e não apresentam uma conservação estrutural e nem ortológica

Como mostrado na **Tabela 4**, existe uma grande discrepância no número de prováveis lncRNAs com a classificação '*others'* que foram isolados e, então, será dada uma maior ênfase nos membros desse grupo. Os lncRNAs descritos com sobreposição de pelo menos um nucleotídeo com um gene ou no final de um *contig* ou cromossomo (classificado como "others") representam ~90% de todos os lncRNAs identificados. Quando se trata de outros organismos como vertebrados, utilizar a classificação proposta pelo GENCODE é apropriada. Porém, para *S. cerevisiae* essa classificação não é apropriada, primeiro em virtude de possuírem menos regiões intrônicas quando comparadas aos vertebrados, enviesando as classificações, e segundo por haver muitos casos de lncRNAs nas regiões terminais de cromossomos e *contigs*. O genoma humano, por exemplo, possui ~25% de sequências intrônicas (SAKHARKAR; CHOW; KANGUEANE, 2004) enquanto que o genoma de *S. cerevisiae* apresenta somente ~5% de íntrons (AST, 2004; CHERRY et al., 2012). Finalmente, por conta dessa diferença e dos dados observados, conclui-se que a classificação utilizada com base no GENCODEv7 é incapaz de atender a demanda de análise dos lncRNAs de *S. cerevisiae,* uma vez que em humanos e outros vertebrados

existe um grande número de sobreposições de lncRNAs e genes codificadores de proteínas completas, ou seja, um gene está totalmente inserido em um lncRNA ou vice-versa (NING et al., 2017). Portanto, existe uma necessidade de padronizar as classificações ou elaborar um outro método classificatório, algo já relatado na literatura também por outros autores como sendo um problema (MA; BAJIC; ZHANG, 2013) em virtude da grande gama de diferentes classificações existentes para esse tipo de RNA. Apesar dos problemas classificatórios, os lncRNAs sobrepostos à genes codificadores de proteínas em humanos têm papel principalmente na gênese de tumores e podem também servir como marcador evolutivo, como mostrado por Ning et al. (2017); por conta disso, no presente trabalho decidimos manter esses lncRNAs nas análises subsequentes.

Apesar de não seguir a mesma tendência de outros organismos, os lncRNAs identificados na *S. cerevisiae* seguem um padrão bem similar a outros organismos quando se trata da conservação dessas moléculas não-codificantes. Recentemente foi sugerido (NOVIELLO et al., 2018) que, ao contrário dos genes codificantes de proteínas, os lncRNAs tendem a preservar sua maquinaria regulatória ao invés de preservar sua sequência, estrutura secundária e ortologia, quando compara-se diferentes organismos. Isso pode ser observado nos dados aqui coletados já que existe uma tendência dos lncRNAs atuarem em vias semelhantes nas linhagens com fenótipos similares, mesmo havendo baixa conservação estrutural e ausência de sintenia conservada. Essa conservação funcional pode ser explicada por padrões evolutivos dos lncRNAs (NITSCHE; STADLER, 2017), já que, mesmo que essas moléculas não sejam um grupo homogêneo, elas conseguem atuar em conjunto numa grande quantidade de vias biológicas. Um exemplo disso está demonstrado nas análises ontológicas das **Figuras 12 e 13**, as quais reportam uma predominância de alvos dos lncRNAs atuando diretamente em controles metabólicos das células quando submetidas ao etanol, além de mecanismos estruturais e de transporte que podem ajudar a célula a tolerar diversos estressores. Essas constatações reforçam a

hipótese de que os lncRNAs, mesmo não conservando suas estruturas entre diferentes linhagens, acabam conservando sua função biológica ao longo da evolução. Um exemplo disso são as linhagens BY4741 (linhagem LT) e BY4742 (linhagem HT), as quais foram desenvolvidas com base na linhagem BY4743 descendente direta da linhagem S288c (linhagem LT) (BRACHMANN et al., 1998), demonstrando que pode ter acontecido uma possível conservação biológica de características funcionais compartilhadas dessas três linhagens e talvez das funcionalidades dos seus lncRNAs.

Além da falta de conservação entre os lncRNAs, também foi possível observar que, ao contrário do reportado para outras espécies de leveduras (CHOI; KIM; NAM, 2018), não foi identificado aqui nenhum lncRNA codificando micropeptídeos com base nos dados de proteômica obtidos em estudos anteriores (ALMEIDA, 2016). Mesmo com esses resultados, ainda podem estar presentes lncRNAs codificadores de micropeptídeos, pois como discutido em Almeida et al. (2019), os dados de proteômica obtidos foram coletados utilizando um cromatógrafo de baixo fluxo, e, mesmo havendo uma baixa recuperação, o algoritmo desenvolvido para identificação dos lncRNAs conseguiu identificar 377 lncRNAs potencialmente codificantes de micropeptídeos (**Tabela 3**). Portanto, uma análise usando um HPLC deve geraria uma maior quantidade de picos massa/carga e, consequentemente, uma maior probabilidade de assinar peptídeos referente à esses lncRNAs isolados.

5.2. Os lncRNAs atuam em vias diferentes de acordo com o fenótipo

A identificação dos papeis dos lncRNAs foi abordada com base na estratégia de *guilt-by-association* seguida de análise de enriquecimento das proteínas-alvo dos lncRNAs. De forma geral, conclui-se que alguns lncRNAs devem possuir uma função relativamente conservada dentro dos genótipos embora esses papéis sejam diferentes quando se compara diferentes fenótipos.

A identificação dos papéis dos lncRNAs foi abordada com base na estratégia de *guilt-by-association* seguida de análise de enriquecimento das proteínas-alvo dos lncRNAs. De forma geral, conclui-se que alguns lncRNAs devem possuir uma função relativamente conservada dentro dos fenótipos. De fato, essas análises permitiram observar que em HTs, os lncRNAs devem atuar tanto em processos de catabolismo de proteínas como em processos de divisão celular. Nesse último caso, aqui sugere-se que os lncRNAs podem afetar negativamente a proliferação celular pois, como observado por Almeida (2017) e Wolf (2019), as linhagens HT possuem uma menor viabilidade e crescimento celular do que as LTs após o tratamento. Como descrito por Huang et al. (2017), o aumento do catabolismo de proteínas acaba por aumentar também o consumo de glicose e, consequentemente, a produção de etanol metabólico. Dessa forma, aqui sugere-se que os lncRNAs nas HTs podem estar contribuindo para uma maior produção de etanol (hipótese a ser futuramente testada). Além disso, alguns alvos estão relacionados à resposta ao etanol. Como já descrito nos trabalhos anteriores (ALMEIDA, 2017; WOLF, 2019) as HTs possuem uma maior tolerância ao estresse etanólico.

O metabolismo de lipídios é um traço muito característico das proteínas-alvo (DE OLIVA NETO et al., 2013; FRAENKEL, 1982; PFEIFFER; MORLEY, 2014) os lncRNAs das LTs, demonstrando um provável papel dos lncRNAs nesse processo ou atuando na sua região de vizinhança ou atuando nas suas proteínas-alvo. Ainda, termos relativos a transporte vacuolar, senescência replicativa e sumolização proteica, ligados ao fato de que as linhagens LTs morrem menos e crescem mais que as HTs durante a condição de tratamento (ALMEIDA, 2017; WOLF, 2019), conduz à hipótese de que os lncRNAs podem estar contribuindo no controle desses processos.

Os papéis dos lncRNAs ainda podem ser analisados tomando como base sua expressão média geral e correlaciona-las com os dados ontológicos, já que, como mostrado na **Figura 9**, duas linhagens LT possuem, no geral, um maior número de lncRNAs com

expressão maior do que todas as outras linhagens, podendo ter relação com o papel dessas moléculas no catabolismo do etanol metabólico, levando a um maior nível de expressão frente ao estressor que estão enfrentando. Além disso, é importante também notar que, em geral, os lncRNAs possuem expressões menores do que os genes por terem caráter regulatório (LI et al., 2019), algo que mesmo sob estresse não há modificações bruscas nos níveis de expressão.

5.3. Mapeamento dos lncRNAs

O mapeamento dos lncRNAs em mapas do KEGG revela que, em geral, os lncRNAs possuem uma expressão diferencial discordante das suas proteínas-alvo. Em outras palavras, na maioria das vezes quando o lncRNA está *up-regulado* sua proteína-alvo está *down-regulada* e vice-versa. Com base nesses dados, propõe-se que esses lncRNAs podem inversamente estarem regulando suas proteínas-alvo. Essa relação já foi descrita para humanos (LI et al., 2018b) e até para o fungo *Schizosaccharomyces pombe* (GARG et al., 2018); a proximidade evolutiva entre *S. cerevisiae* e *S. pombe* reforça o comportamento inverso de gene-lncRNA pois, em *S. pombe,* a inativação da transcrição do lncRNA *prt 2* (seja por deleção ou mutação do seu promotor) leva a maior expressão dos genes-alvo *pho84* e *pho1*, levando a conclusão do papel repressivo por interferência transcricional deste lncRNA.

O mapa da **Figura 19** (uma linhagem HT), assim como em *S. pombe* (GARG et al., 2018), é um exemplo da relação inversa de um lncRNA com sua proteína-alvo. Esse exemplo suplementa a informação de enriquecimento observado em LTs relativo ao catabolismo do etanol metabólico, uma vez que as HTs crescem menos que as LTs após estresse severo de etanol (ALMEIDA, 2017). Contudo, supõe-se que esta regulação pode estar sendo mediada pelo lncRNA nessa via, algo que já foi reportado para humanos (LI et al., 2017). No caso citado, o lncRNA lncHR1 de humanos regula negativamente a

expressão da SREBP-1c (uma proteína regulatória de esteróis), ou seja, quando o lncRNA está muito expresso, ele acaba levando a um menor expressão desse gene e, consequentemente, a uma menor expressão das enzimas relacionadas tais como a FAZ e a Acetil-CoA carboxilase (sendo esta última, a mesma proteína que o lncRNA identificado neste trabalho está também regulando negativamente). Isso é possível devido ao fato dos lncRNAs poderem controlar alostericamente processos transcricionais (LONG et al., 2017) formando complexos com outras proteínas que acabam interagindo com fatores de transcrição, modificando o tipo de interação da polimerase com o promotor; isso é um dos tipos de interações mais importantes dos lncRNAs. Porém, aqui não reportamos nenhum caso de interação física predita de um lncRNA com um fator de transcrição.

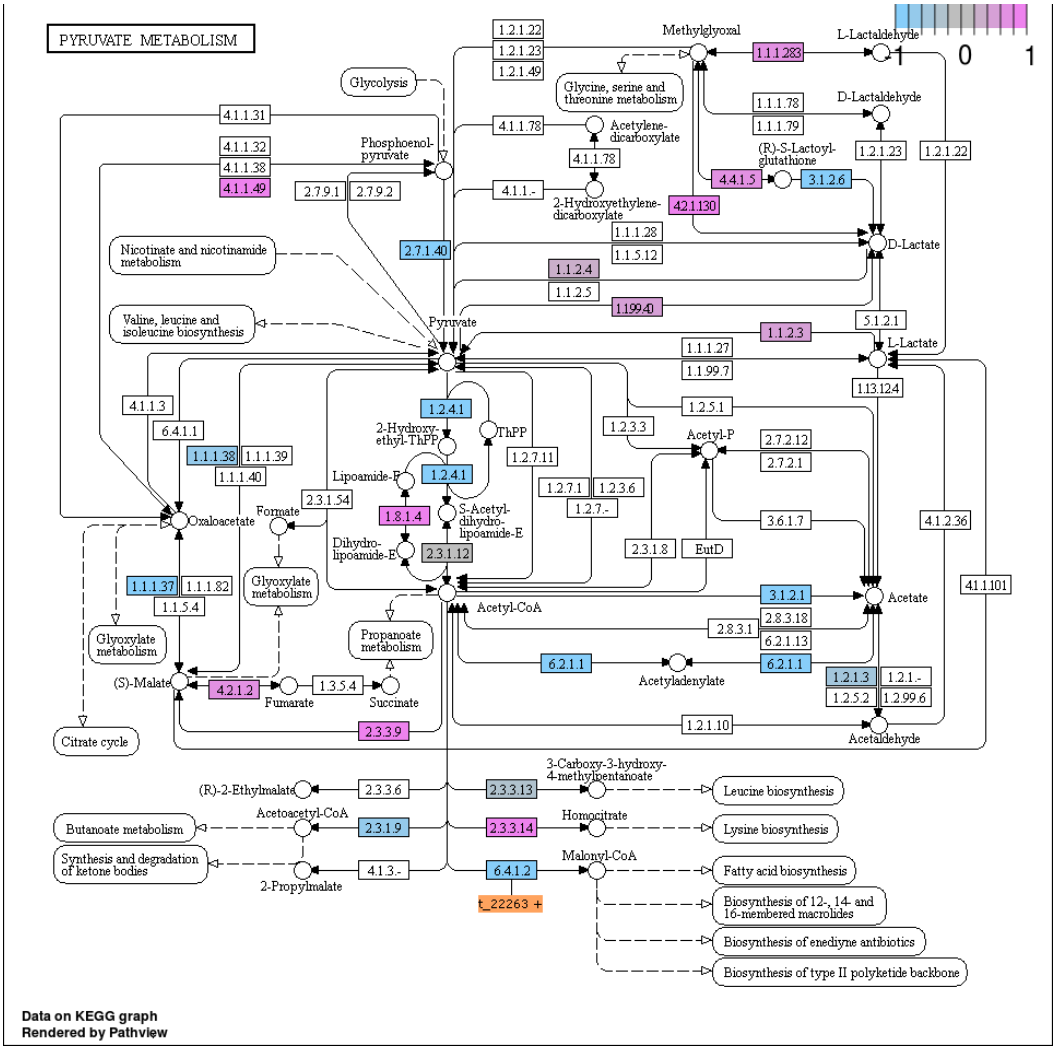

Figura 19 - Mapa mostrando o metabolismo do piruvato na linhagem HT BMA64-1A.

Outro exemplo da relação de regulação inversa dos lncRNA pode ser observado no mapa da **Figura 20**. Esse caso ainda serve como provável explicação do motivo das LTs terem uma maior viabilidade celular do que as HTs (ALMEIDA, 2017; WOLF, 2019). Nesse caso, provavelmente o lncRNA está reprimindo a tradução da enzima-alvo já que, frente ao estímulo do estressor, existe uma maior abundância do principal metabólito dessa via. Essa abundância é fundamental para célula já que a maior disponibilidade de L-Lisina está diretamente relacionada com a diminuição dos efeitos de toxicidade do etanol no organismo (WANG et al., 2014; WARD et al., 1972). A formação do metabólito *L-2-aminoadipate 6-semialdehyde* (derivada da lisina, chamada de alisina) foi descrita como regulada epigeneticamente pelo lncRNA GATA6-AS em camundongos (NEUMANN et al., 2018). Essa interação ocorre entre o RNA e o gene LOXL2 (responsável pela deaminação da lisina em alisina), o qual é negativamente influenciado pelo lncRNA e, assim, uma maior expressão do lncRNA leva a uma diminuição de alisina. Essa relação tem ligação direta com a importância da lisina nas análises aqui realizadas, uma vez que supõe-se, que aumentando a expressão do lncRNA e diminuindo a do gene LYS2 (possui função homóloga ao LOXL2) na amostra S288c, pode não haver conversão de resíduos de lisina para alisina, levando a uma maior disponibilidade desse metabólito para a célula.

O papel dos lncRNAs como “esponjas” de proteínas é um outro tipo de regulação relacionadas que pode ser empregado para explicar as interações físicas inversas entre lncRNAs e seus alvos (CHARLEY; WILUSZ, 2014; KIM et al., 2016). Essa interação é possível pois, uma célula sob estresse pode sinalizar para os lncRNAs a necessidade de inibir a produção de determinada proteína (PORTO; DAULATABAD; JANGA, 2019), consequentemente fazendo um tipo de tamponamento do efeito de tal proteína. Duss et al. (2014) mostraram em *Pseudomonas fluorescens*, que o RNA não-codificante RsmZ atua num mecanismo de tamponamento proteico no qual ele consegue sequestrar e armazenar a proteína RsmE protegendo o ncRNA da degradação por RNase.

De forma contrastante às regulações inversas descritas acima, também existem casos em que o lncRNA acompanha a expressão diferencial da sua proteína-alvo, como mostrado na **Figura 21**. Por se tratar de uma linhagem LT, nesse caso, supõe-se que esse lncRNA *up*-regulado pode estar ligando e estabilizando a ATP-Sintase (de forma pós-transcricional) ao mascarar seu sítio de ubiquitinação, induzindo o acúmulo da proteína-alvo (ZHANG et al., 2015); isso provavelmente decorre da regulação positiva da proteína-alvo feita pelo lncRNA, consequentemente aumentando a atividade dessa proteína. A importância dos lncRNAs nesta etapa da produção energética foi demonstrada na própria *S. cerevisiae* recentemente (DU MEE et al., 2018), a qual descreveram a importância na regulação feita pelo lncRNA CUT60 no gene ATP16. Foi observado que, quando ocorre a inibição deste RNA, ocorre tanto um mal funcionamento de um dos principais genes responsáveis pela ATP-Sintase (ATP16) quanto também uma perda do genoma mitocondrial, levando a uma menor taxa de crescimento celular. O lncRNA identificado neste trabalho também tem como alvo alguns dos genes responsáveis pelo funcionamento da ATP-Sintase: da mesma forma que o lncRNA CUT60 é importante para a expressão do gene ATP16, o lncRNA transcr_18666 possui expressão *up*-regulada juntamente como seu alvo, levando a hipótese que esse RNA pode estar aumentando a produção energética como um todo visto que essa enzima é responsável pelo transporte de H+ dentro da mitocôndria.

No caso analisado da levedura sob estresse etanólico, os mecanismos aqui discutidos são interessantes à célula pois ela acaba usando todos os tipos de “defesa” possíveis para combater a toxicidade etanólica, aumentando a expressão dos seus lncRNAs para tamponar proteínas que não sejam de interesse celular naquele momento de combate à molécula do etanol, ou ligando-se à determinadas proteínas de interesse para a célula estabilizando-as, ou até modificando a interação da polimerase com o promotor de determinados genes.

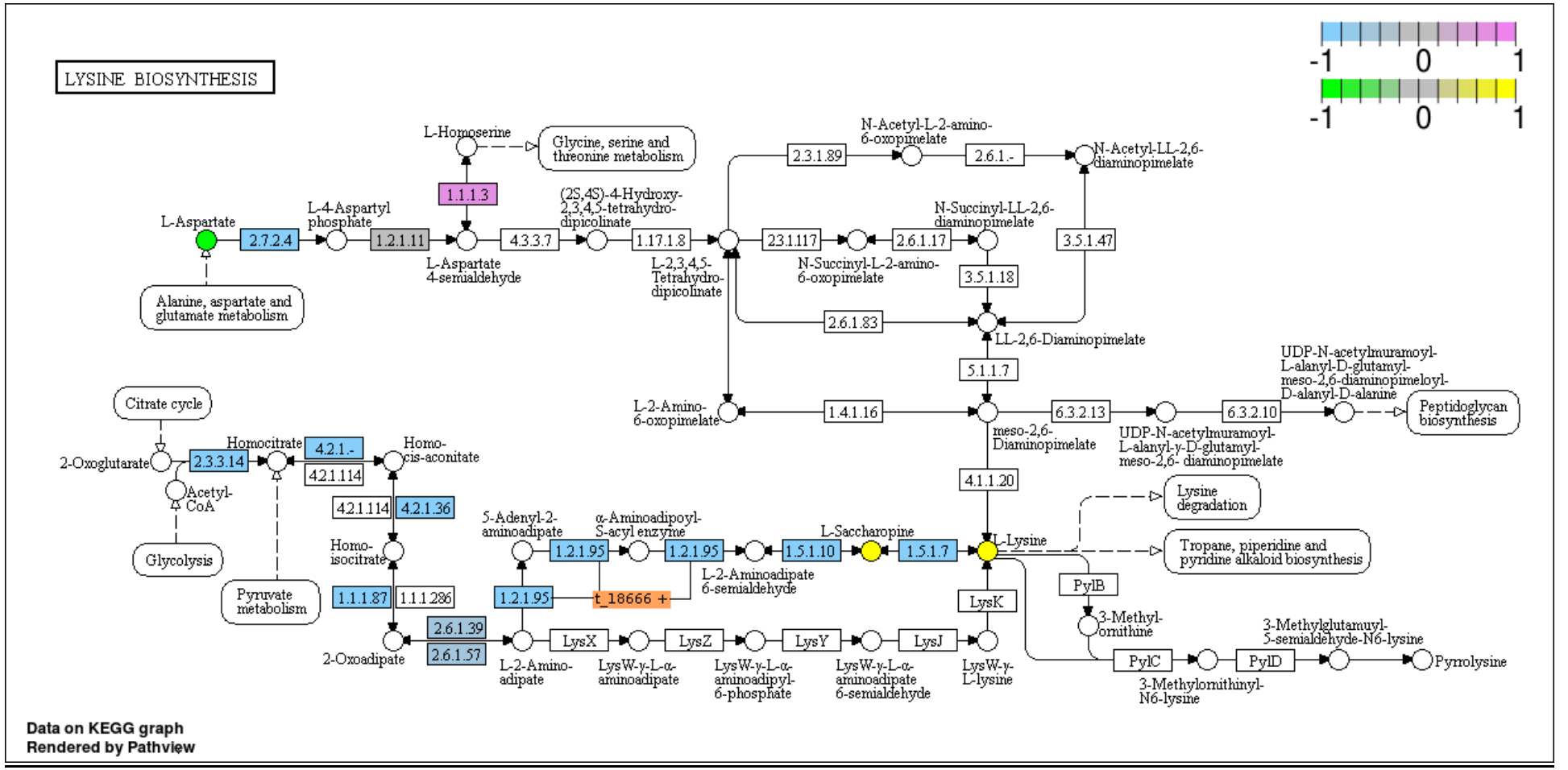

Figura 20 - Mapa mostrando a biossíntese de lisina na linhagem LT S288c.

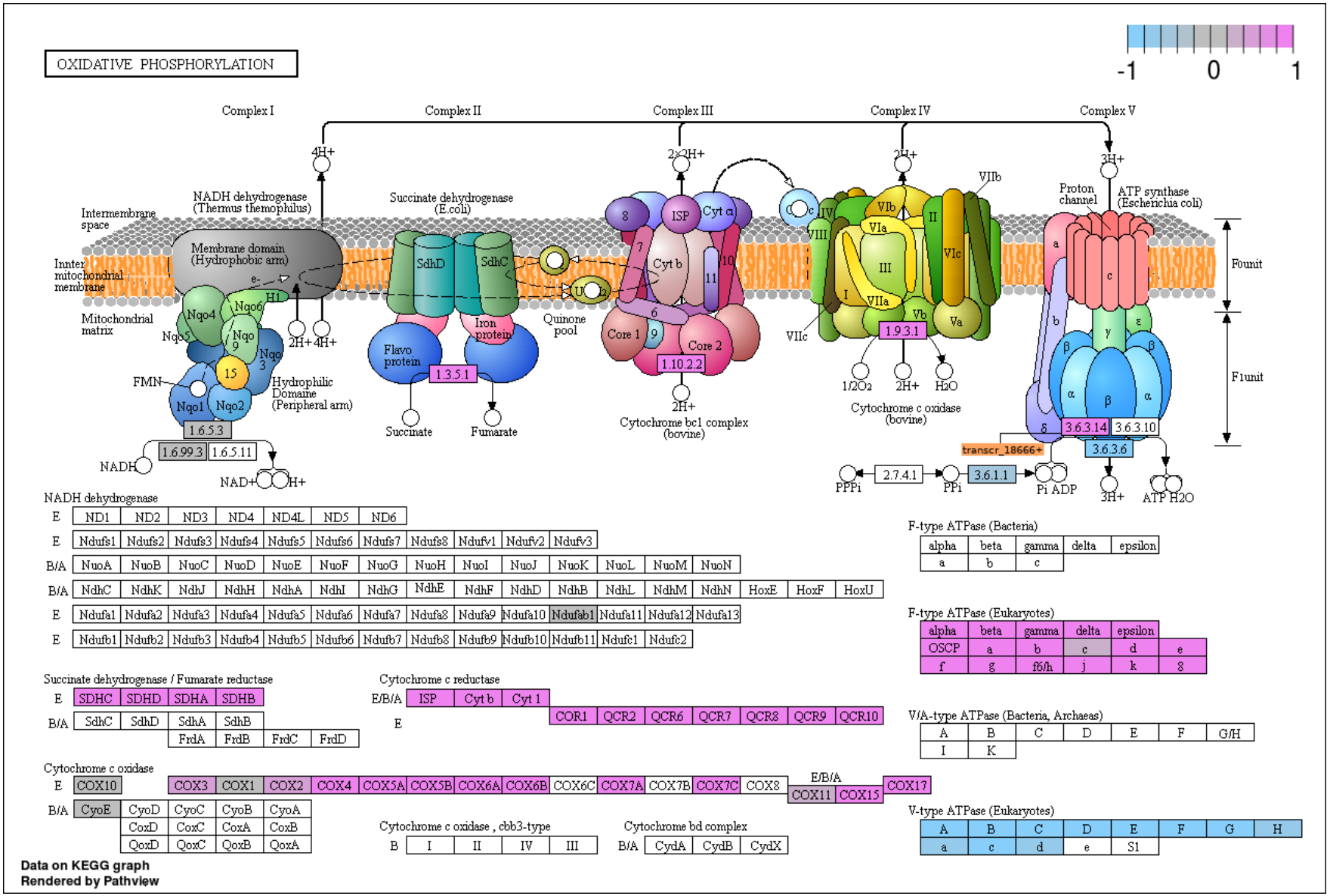

Figura 21 - Mapa mostrando a fosforilação oxidativa na linhagem LT S288c.

5.4. Estudo da potencial interação entre os lncRNAs e genes codificadores de proteínas sobrepostos a eles

Vários genes sobrepostos aos lncRNAs foram identificados, sendo que vários deles são relacionados a funções estruturais, funções transcricionais e funções metabólicas. O gene YGR087C (piruvato descarboxilase) é um dos principais exemplos de sobreposição a um lncRNAs na linhagem BMA64-1A (a linhagem mais tolerante ao etanol). Além deste gene estar diretamente relacionado ao mapa mostrado na **Figura 19**, reforçando o potencial regulatório dos lncRNAs tanto em *cis* quanto em *trans*, o lncRNA GCASPC em humanos (MA et al., 2016) também está diretamente relacionado à piruvato descaboxilase, atuando como isca molecular e *down*-regulando os níveis dessa enzima. Porém, em humanos é uma regulação *trans*, não havendo até o momento nenhum relato de regulação *cis* para esse gene.

Em termos evolutivos, as configurações de sobreposição no genoma são principalmente afetadas pela própria origem de genes evolutivamente sobrepondo-se. Consequentemente, levando ao aparecimento de casos de sobreposição dos lncRNAs sendo que há relatos dessas sobreposições ajustando os níveis de expressão dos pares genes/lncRNA (NING et al., 2017). Porém, como não foram identificados casos gerais da regulação da região de sobreposição gênica, o foco da análise da região *cis* dos lncRNAs será os vizinhos dessas moléculas não-codificantes.

Capítulo 2 – O papel dos lncRNAs como reguladores da expressão genica de seus vizinhos

Neste capítulo sumariza-se as abordagens realizadas para a análise das possíveis formas que os lncRNAs identificados nessa dissertação podem regular a expressão dos seus genes vizinhos e como podem alterar essa regulação com o passar do tempo, tanto em uma situação de estresse quanto em condições ótimas de crescimento. Contudo, aqui não será explorado aspectos relativos à tolerância ao etanol, sendo então reportado esse estressor como um agente causador de mudanças sistêmicas que possibilitam avaliar o impacto dos lncRNAs como reguladores de seus vizinhos.

3. Materiais e métodos

3.1. Separação dos pares genes vizinhos-lncRNA com seus respectivos coeficientes angulares e identificação da região promotora dos genes-vizinhos

Todos os prováveis lncRNAs tiveram seus dados de expressão de *time-course* comparados com seus vizinhos (identificados tal como descrito no Capítulo 1) e com genes aleatoriamente selecionados no genoma. Primeiramente, para cada lncRNA, foram capturados aleatoriamente no genoma 10 genes codificantes não-vizinhos ao lncRNA sob avaliação. Posteriormente, foi calculada a correlação de Pearson dos dados de expressão entre cada lncRNA com o seu vizinho e entre o lncRNA e os outros 10 genes não-vizinhos aleatoriamente escolhidos; obtendo ao final 11 valores de correlação de expressão para cada lncRNAs. Esses valores foram estatisticamente comparados utilizando o teste T de Student e assumiu-se que um lncRNA impacta a expressão de seu vizinho caso a correlação da expressão entre ele e seu vizinho tenha diferença estatística (*p-value* ≤ 0,05) das correlações com os genes aleatoriamente escolhidos. Essa análise foi realizada tanto para os dados controle quanto tratamento e estão disponíveis em https://1drv.ms/x/s!Aju93ah3HMghgfVIBAZorouriBXW9Q?e=IuVQ2f

Com base nas correlações significativas (*p-value* ≤ 0.05) entre os lncRNAs e seus genes vizinhos, os valores de expressão no *time-course* desses pares foram isolados e, com base nesses valores, foi possível calcular *k-scores*; esses *scores* são os coeficientes angulares das retas os quais neste trabalho serviram para modelar como os lncRNAs influenciam as expressões de seus vizinhos ao longo do tempo. Para isso, foram definidos k1 e k2 como sendo, respectivamente, a comparação das expressões de determinado gene-lncRNA em 2 horas *vs.* 1 hora e entre 4 horas *vs.* 2 horas (**Equações 1 e 2**). Os tipos de perfis regulatórios que os lncRNAs podem exercer sobre os seus vizinhos foram definidos como o comportamento da expressão deles em cada momento do *time-course*. O cálculo dos coeficientes angulares permite verificar a tendência de *up*-expressão (caso k>0)

e a tendência de *down*-expressão (caso k<0) tanto para os genes quanto para os lncRNAs. Em suma, como mostrado na **Tabela 9**, é possível traçar quatro comportamentos regulatórios.

$$k1 = \frac{Expressão\ 2hrs - Expressão\ 1hr}{2} \qquad k2 = \frac{Expressão\ 4hrs - Expressão\ 2hr}{2}$$

Equações 1 e 2 – Equações para cálculo dos coeficientes angulares.

Tabela 9 - Perfis regulatórios baseados nos coeficientes angulares de cada membro do par gene-lncRNA. Relação "Similar" representa uma situação a qual tanto o lncRNA quanto gene possuem o mesmo perfil de expressão. Relação "Diferente" representa uma situação a qual o lncRNA possui um perfil de expressão diferente do seu vizinho.

Casos	lncRNA *up*	lncRNA *down*
Gene *up*	Relação Similar	Relação Diferente
Gene *down*	Relação Diferente	Relação Similar

Os perfis de expressão de um lncRNA e de seu vizinho ao longo do *time-course* foram definidos como *"Same"* (Relação Similar) ou "*Not Same*" (Relação Diferente) com base nos valores de k1 e k2 (exemplo na **Figura 22**). Para isso, se o k1 de um gene não apresentasse perfil similar ao k1 ou ao k2 do seu lncRNA-par, o par foi definido como *Not Same* ("Relação Diferente"). Caso o k1 do gene tenha perfil similar ao k1 ou ao k2 do seu lncRNA-par, o perfil foi classificado como "Relação Similar" (*Same*) (**Tabela 10**). Além disso, esses dados foram analisados separadamente levando em consideração: correlação entre os vizinhos (positiva ou negativa), situação (controle ou tratamento) e linhagem (BMA64-1A ou S288c). Esses dados estão disponíveis em https://1drv.ms/x/s!Aju93ah3HMghgfVKMaw8PWkh351TTg?e=0Ny13D (BMA64-1A) e https://1drv.ms/x/s!Aju93ah3HMghgfVL7wyhcmndrYTmww?e=jfHsBo (S288c).

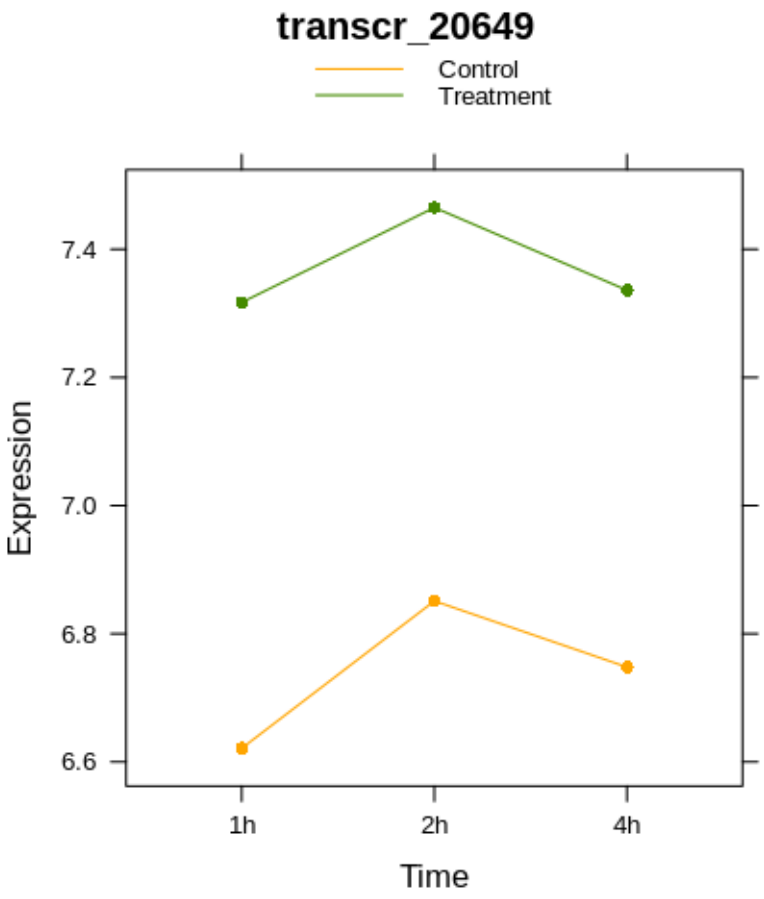

Figura 22 – Perfis de expressão de um lncRNA da S288c (transcr_20649). O k1 é positivo (valor de 0,24), ou seja, a expressão está subindo de 1 a 2 horas enquanto o k2 é negativo (valor de -0,42), ou seja, a expressão está diminuindo de 2 a 4 horas. Perfil semelhante tanto na situação de tratamento como na situação controle.

Tabela 10 - Sumarização das possíveis relações entre lncRNA/genes.

Casos	**K1 lncRNA positivo**	**K1 lncRNA negativo**	**K2 lncRNA positivo**	**K2 lncRNA negativo**
K1 gene positivo	*Same*	*Not-Same*	-	-
K1 gene negativo	*Not-Same*	*Same*	-	-
K2 gene positivo	*Same*	*Not-Same*	*Same*	*Not-Same*
K2 gene negativo	*Not-Same*	*Same*	*Not-Same*	*Same*

A identificação da região promotora dos genes aqui analisados foi feita com base no banco de dados Yeastract (TEIXEIRA et al., 2018) a fim de avaliar tanto o posicionamento do lncRNAs em relação ao próprio gene (se está na região 5’ ou 3’ dele) quanto em relação ao promotor desse gene (se está na região 5’ ou 3’ ou se está sobreposto à ele).

4. Resultados

4.1. Identificação dos prováveis tipos de regulação que os lncRNAs exercem sobre seus vizinhos

Os dados de significância das correlações entre os lncRNAs com seus vizinhos reais comparados com 10 genes aleatoriamente escolhidos demonstram que há uma tendência de aumento no número de correlações significativas entre os lncRNAs e os seus vizinhos quando as linhagens observadas (BMA64-1A e S288c) estão sob condição de tratamento (**Tabela 11**). Além disso, há uma tendência dos vizinhos acompanhar a expressão dos lncRNAs e uma tendência de aumento das correlações negativas sob condição de estresse etanólico. Ressalta-se ainda que todos os lncRNAs identificados estão sempre na mesma orientação dos seus vizinhos.

Tabela 11 - Correlações significantes entre os lncRNAs com seus vizinhos verdadeiros quando comparados com o valor de expressão de 10 genes aleatórios.

Condições	Significante	Não-Significante
BMA64-1A Controle	267 (62%)	161 (38%)
BMA64-1A Tratamento	315 (74%)	113 (26%)
S288c Controle	287 (58%)	210 (42%)
S288c Tratamento	354 (71%)	143 (29%)

Tabela 12 - Correlações positivas e negativas significativas entre os lncRNAs e seus vizinhos verdadeiros.

Condições	Negativo	Positivo
BMA64-1A Controle	106 (40%)	161 (60%)
BMA64-1A Tratamento	137 (44%)	178 (46%)
S288c Controle	94 (33%)	193 (67%)
S288c Tratamento	146 (41%)	208 (59%)

Quando observa-se as mudanças pontuais das correlações significativas (tanto negativas quanto positivas em ambas situações (controle e tratamento)) nas linhagens BMA64-1A e S288c (**Figura 23**), é possível encontrar 3 padrões: 1- pares que somente possuem correlação significativa em uma única condição (ou controle ou tratamento),

perfazendo a maioria dos casos (52% para S288c e 46,3% para BMA64-1A) com predominância das correlações positivas (**Figuras 23-1, 23-4 e 24**); 2- o segundo maior percentual de correlações significativas observadas refere-se a um estado de manutenção das correlações após o tratamento (**Figuras 23-2 e 24**); 3- uma menor percentagem das correlações refere-se a uma troca de estado após o tratamento (a correlação passa de positiva para negativa ou vice-versa) (**Figuras 23-3 e 24**).

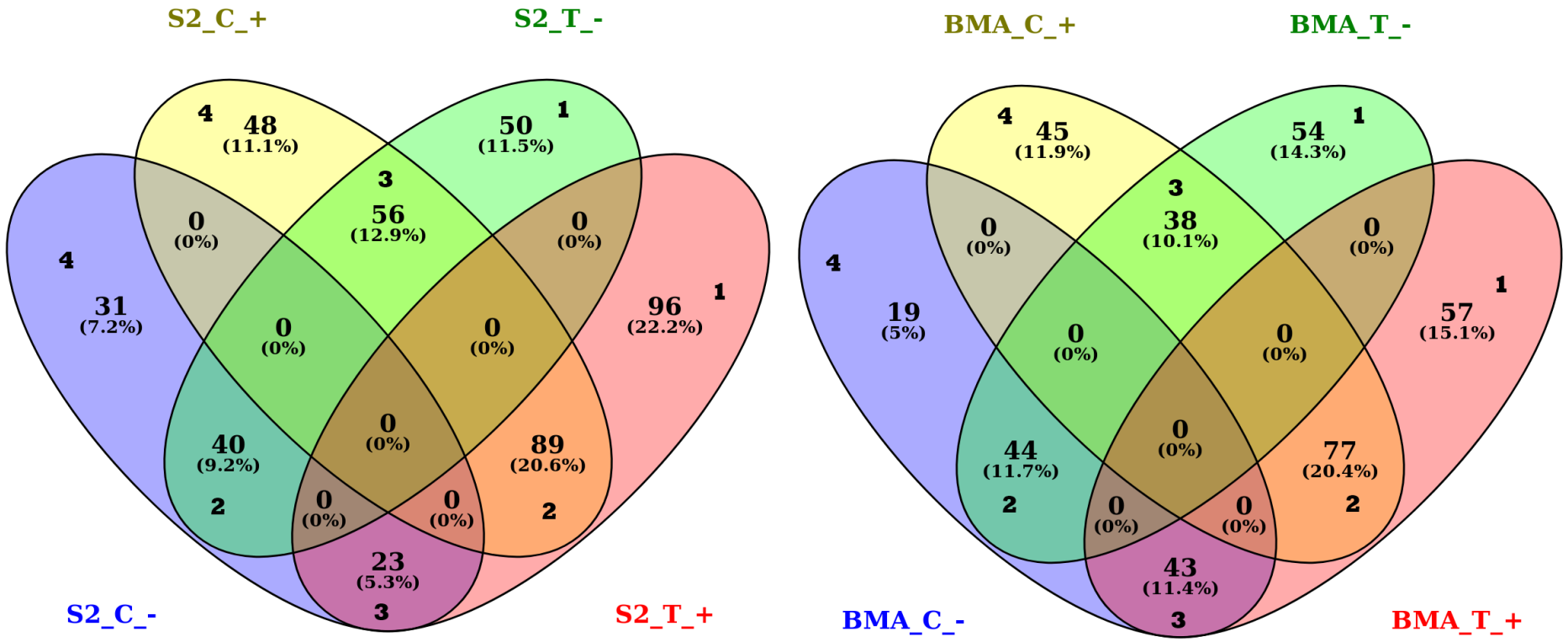

Figura 23 - Diagramas de Venn reportando o comportamento dos lncRNAs quando passam da situação de controle para tratamento. A esquerda observa-se os dados da linhagem S288c e a direita os dados da linhagem BMA64-1A.

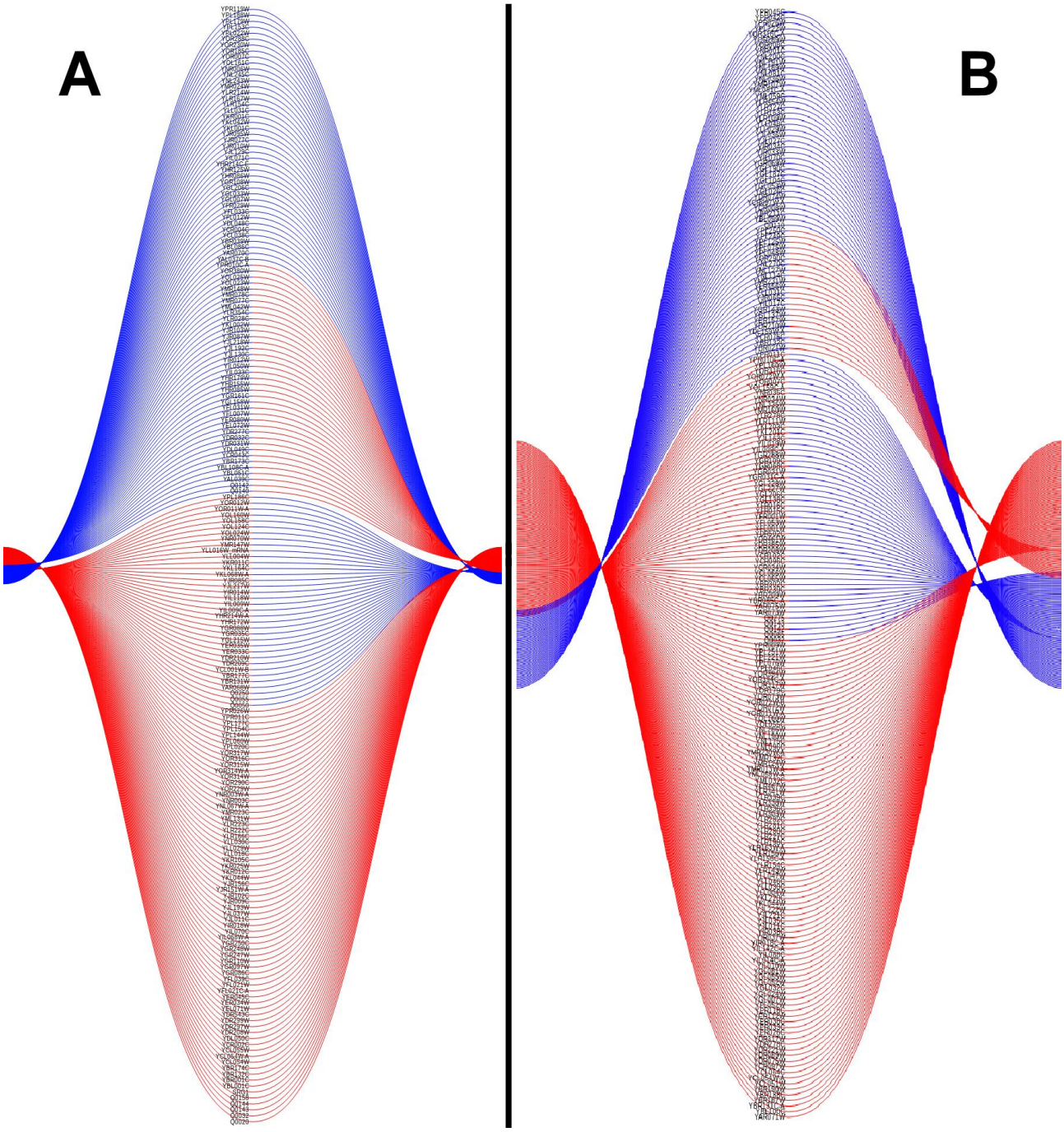

Figura 24 - Comportamento dos genes vizinhos aos lncRNAs quando passam da situação de controle para tratamento. A, BMA64-1A; B, S288c. A esquerda de cada diagrama está a situação controle e a direita a tratamento, sendo que a cor vermelha representa correlações positivas e a cor azul correlações negativas.

É possível identificar que as relações concordantes, ou seja, aquelas em que o gene está *up*-expresso e o lncRNA também está *up*-expresso, por exemplo, são mais abundantes nas duas linhagens avaliadas e nas duas situações analisadas (**Tabela 13**). Além disso, reporta-se uma diferença significativa nos níveis de expressão e nas mudanças das relações dos lncRNAs/genes-vizinhos com o passar do tempo (**Figuras 26, 27**). Contudo, é possível constatar que mesmo a célula não estando em uma condição ótima, parece não

haver uma clara mudança no perfil regulatório das moléculas aqui descritas, apesar de haver alguns casos específicos em que esse perfil pode ser alterado quando se avança de 1h para 2h e posteriormente para 4h.

Tabela 13 - Dados dos perfis regulatórios observados comparando os tempos por meio dos coeficientes angulares na relação lncRNA-gene vizinho. *, o número bruto e percentagens estão apresentados.

Situação	***BMA64-1A Controle**	***BMA64-1A Tratamento**	***S288c Controle**	***S288c Tratamento**
Relação	*Same/Not-Same*	*Same/Not-Same*	*Same/Not-Same*	*Same/Not-Same*
K1	160 = 59,92% / 107 = 40,08%	168 = 53,33% / 147 = 46,67%	189 = 65,85% / 98 = 34,15%	198 = 55,93% / 156 = 44,07%
K2	155 = 58,05% / 113 = 41,95%	158 = 50,16% / 157 = 49,84%	171 = 59,59% / 116 = 40,41%	216 = 61,02% / 138 = 38,98%

A presença desses dois tipos de relações antagônicas entre si (tanto concordantes, ou seja, relação na qual que o gene e o lncRNA são ambos *up* ou ambos *down*-expressos, quanto discordantes, ou seja, relação na qual o gene é *up* e o lncRNA é *down*-expresso e vice-versa) pode ser elucidada pontualmente pela própria posição genômica dos genes e dos lncRNAs. Por exemplo, como mostrado na **Figuras 25**, a relação entre as expressões gênicas do lncRNA e do seu gene vizinho podem decorrer das suas posições genômicas, ou seja, a posição do lncRNA estando à 3’ ou à 5’ do gene pode influenciar no tipo de relação lncRNA-gene. A **Figura 25-A** mostra uma Relação Discordante a qual o lncRNA está na região 3’ do gene e a **Figura 25-B** uma Relação Concordante a qual o lncRNA está na região 5’ do gene (alocando-se na região promotora).

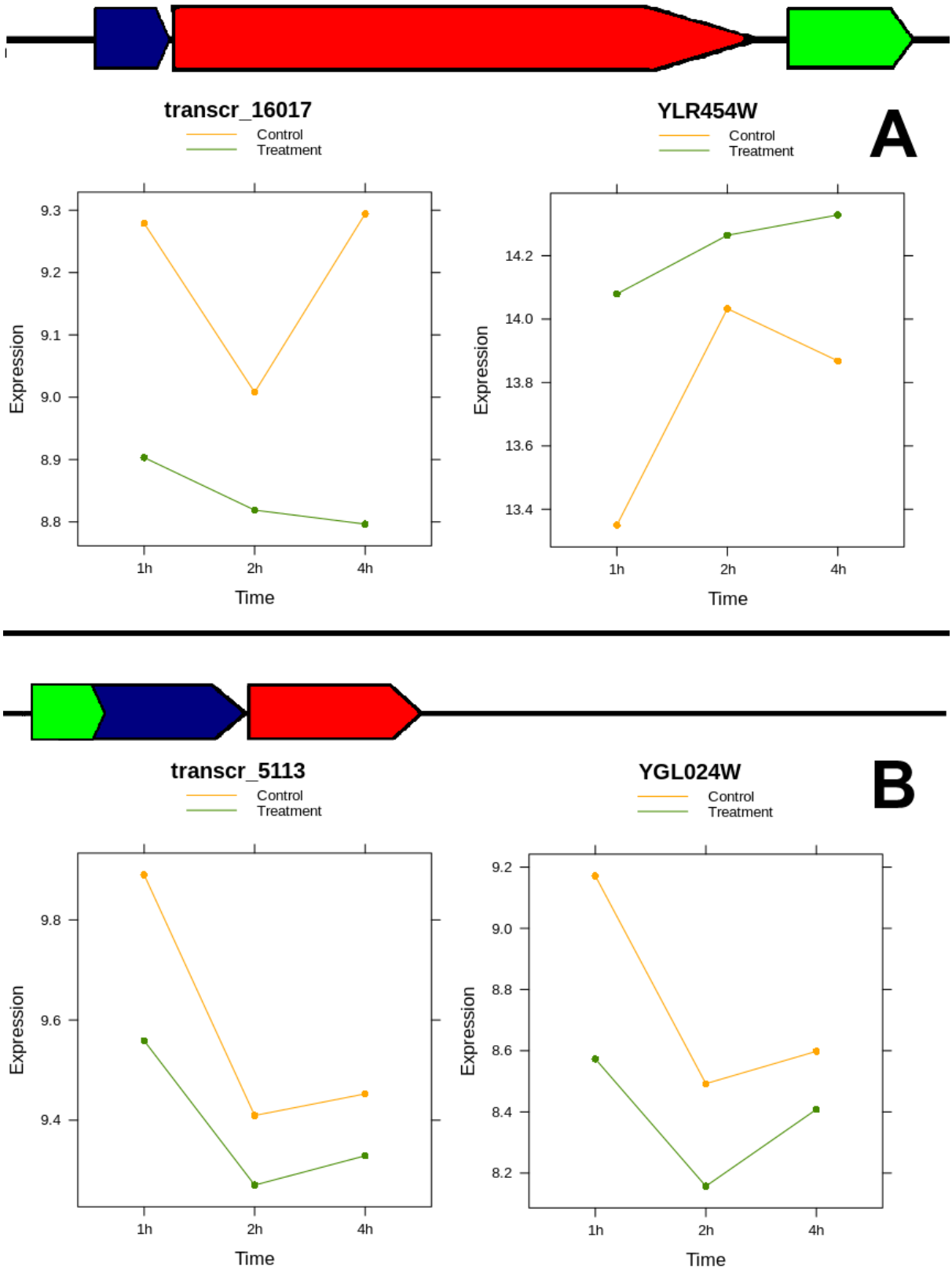

Figura 25 - Na parte superior de ambas imagens, demonstração ilustrativa do posicionamento da região promotora (em azul) do gene (em vermelho) YLR454W (**A**), YGL024W (**B**) e dos lncRNAs vizinhos (em verde), sendo o transcr_16017 (**A**) e o transcr_5113 (**B**). Os gráficos mostram a expressão gênica nos três pontos da análise do *time course.*

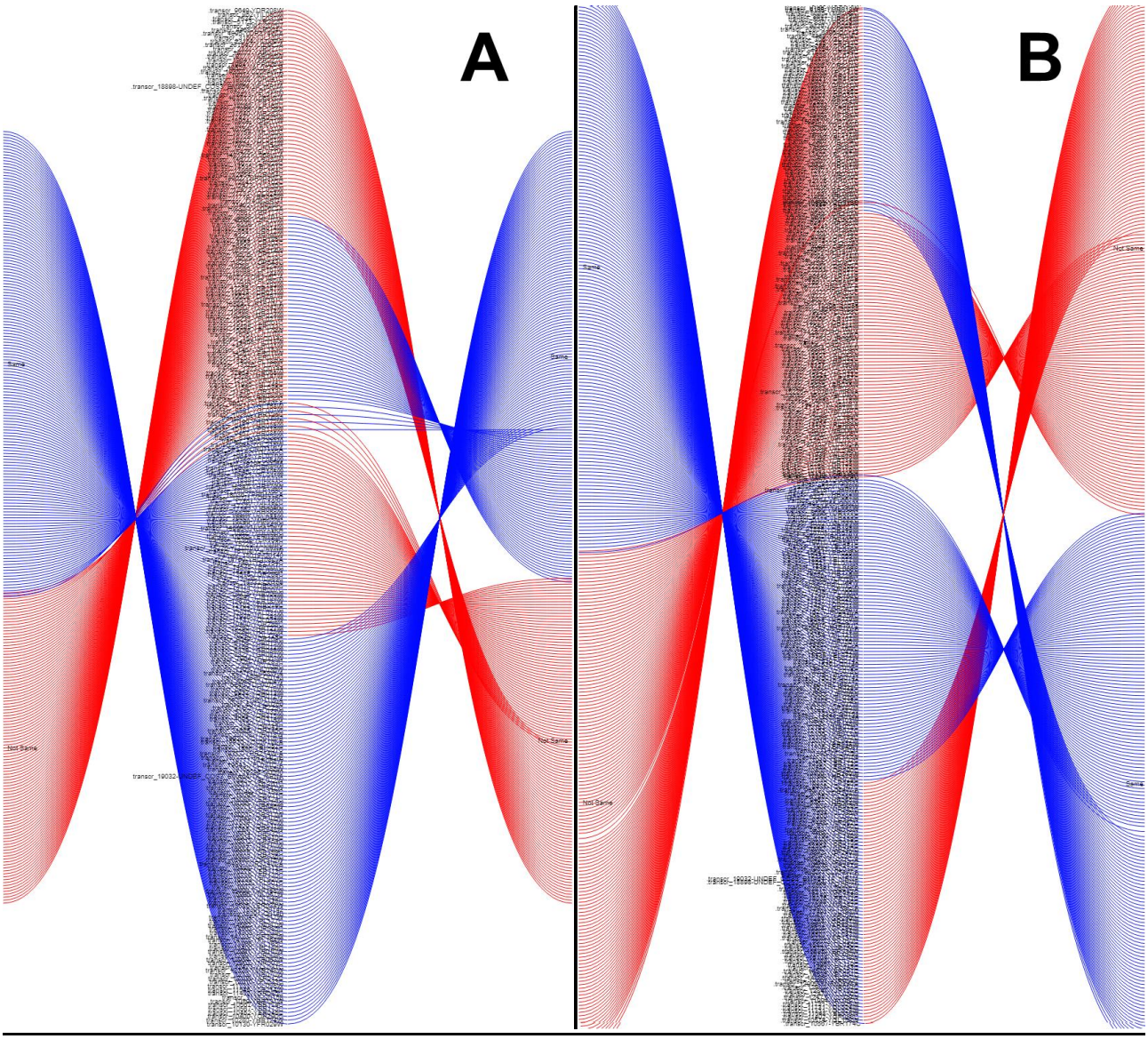

Figura 26 - Diagramas aluviais mostrando o comportamento dos coeficientes angulares entre os tempos analisados para a linhagem BMA64-1A. A = Situação controle; B = Situação tratamento. A esquerda de cada diagrama estão os dados de k1 e a direita os dados de k2. A cor vermelha representa a relação *Not Same* e a azul a relação *Same*.

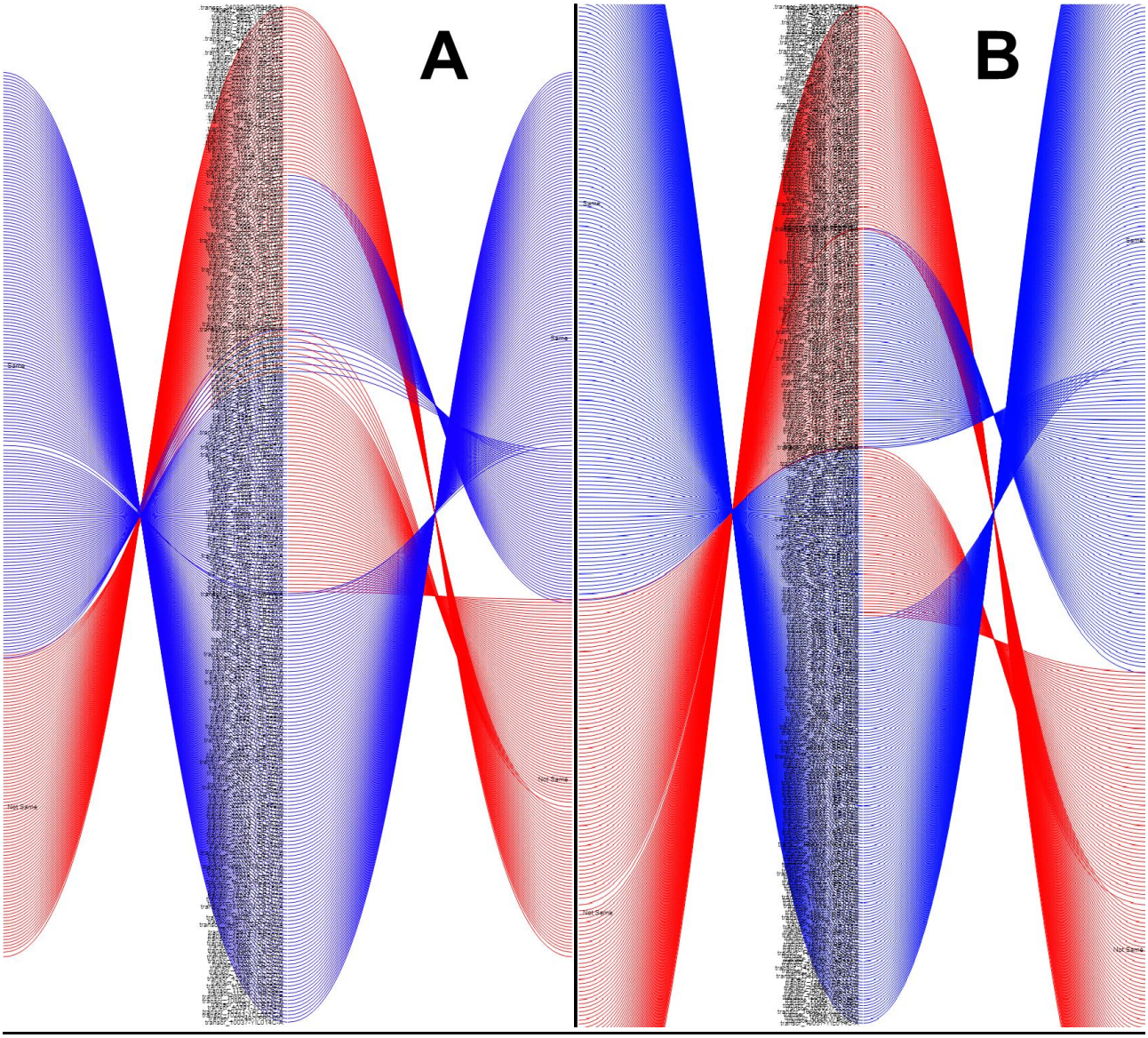

Figura 27 - Diagramas aluviais mostrando o comportamento dos coeficientes angulares entre os tempos analisados para a linhagem S288c. A = Situação controle; B = Situação tratamento. A esquerda de cada diagrama estão os dados de k1 e a direita os dados de k2. A cor vermelha representa a relação *Not Same* e a azul a relação *Same*.

5. Discussão

5.1. Estudo da potencial interação entre os lncRNAs e genes codificadores de proteínas em sua região de vizinhança

Ao observarmos a diferença de significância da **Tabela 11**, a tendência de aumentar as correlações significativas após o tratamento leva a elaboração da hipótese de que, sob estresse, os lncRNA atuam muito acentuadamente e de forma pós-transcricional

para regular os genes responsáveis pela resposta do estresse ao etanol. A característica pós-transcricional fica ainda mais clara quando se observa que os lncRNA e seus vizinhos possuem sempre a mesma orientação (**Figura 28**)**.** Tal fato é esperado devido às diversas formas existentes de regulação dos lncRNAs, tais como as descritas no Capítulo 1.

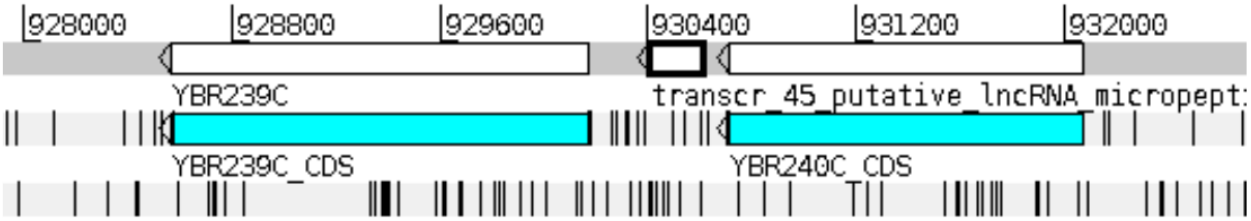

Figura 28 - Região genômica de uma linhagem LT (S288c). Os genes (nesse caso, YBR239C e YBR240C) sempre possuem a mesma orientação dos lncRNA vizinho (nesse caso, transcr_45).

Concomitante a essas constatações, os lncRNAs não tiveram suas relações com seus vizinhos muito alteradas após o tratamento com etanol (**Figura 23 e 24, Tabela 11**), indicando uma manutenção de comportamento dos pares frente ao estresse. Esse tipo de influência na vizinhança já foi observada em vários organismos, tal como o caso do lncRNA LAIR de arroz (WANG et al., 2018). A super-expressão desse lncRNA *up*-regula a expressão de vários genes da família LRK ao se ligar à histona e modificar as proteínas OsMOF e OsWDR5 nas células da planta. Em *Drosophila,* foi identificado correlações positivas entre os níveis dos lncRNAs e seus genes vizinhos (SCHOR et al., 2018). Em humanos o lncRNA sloyfley pode influenciar na isquemia cerebral por conta de regular o seu gene vizinho Peg3 (LI et al., 2018a); sua super-expressão resulta em uma maior expressão do seu vizinho. Com essas informações, sugere-se que quando a célula passa por um estresse, ocorre o aumento do número de lncRNAs que passam a ser significativamente correlacionados com seus vizinhos (como mostrado na **Tabela 11**), ou seja, ocorre o aumento no número de correlações significantes em ambas linhagens analisadas quando elas passam da situação controle para a situação tratamento. Esse aumento na significância no tratamento potencialmente ocorre por conta das sinalizações

celulares feitas decorrente da situação de estresse. Nesse caso, quando a célula precisa combater um estressor, ela ativa mecanismos de resposta (AUESUKAREE, 2017) sendo que os próprios lncRNAs podem atuar de forma bem mais discreta em *cis* do que em *trans*, direta ou indiretamente nos diversos mecanismos celulares envolvidos. Apesar da diferença no nível de pares significantes ser baixo entre controle e tratamento (12% na BMA64-1A e 13% na S288c), as análises dos termos ontológicos relacionados à esses lncRNAs que fazem parte de pares que se tornaram significantes sob estresse, reporta que o metabolismo de lipídeos e resposta a feromônios (ligado ao estresse) está enriquecido na linhagem S288c, ao passo que na BMA64-1A os termos relativos ao metabolismo do propionato e resposta ao furfural estão enriquecidos. Todos esses termos estão diretamente ligados a respostas ao estresse quando analisadas as correlações dos lncRNAs com seus vizinhos, respondendo à uma maior produção de energia intracelular, tal como ocorre no metabolismo de lipídeos, ou atuando numa melhor sinalização intracelular (como a resposta ao furfural) (**Figura 29** e **30**).

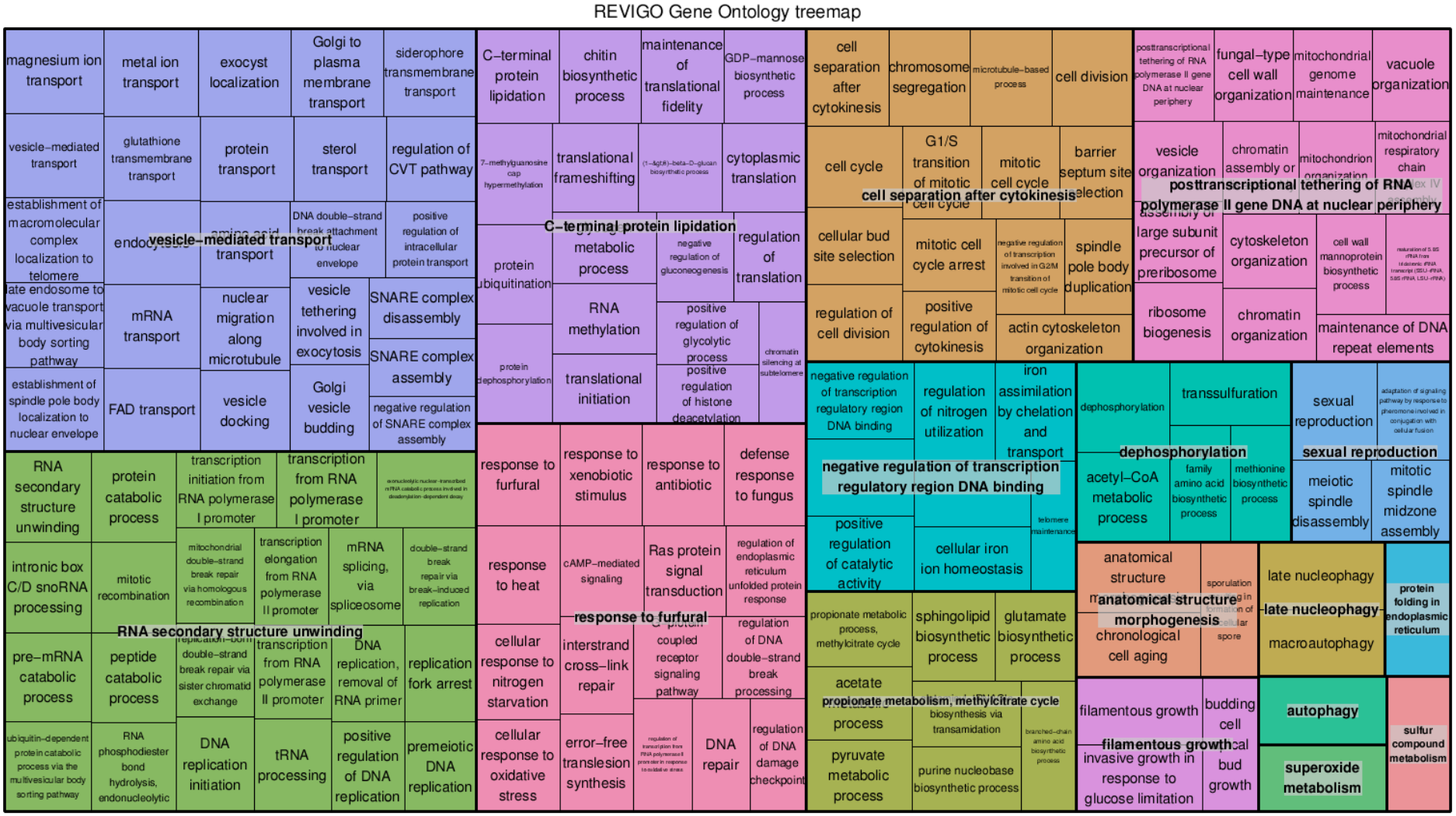

Figura 29 - Processos biológicos gerados pelo REVIGO com base nos GOs dos lncRNA que passaram a ser significantes sob condição de tratamento em BMA64-1A.

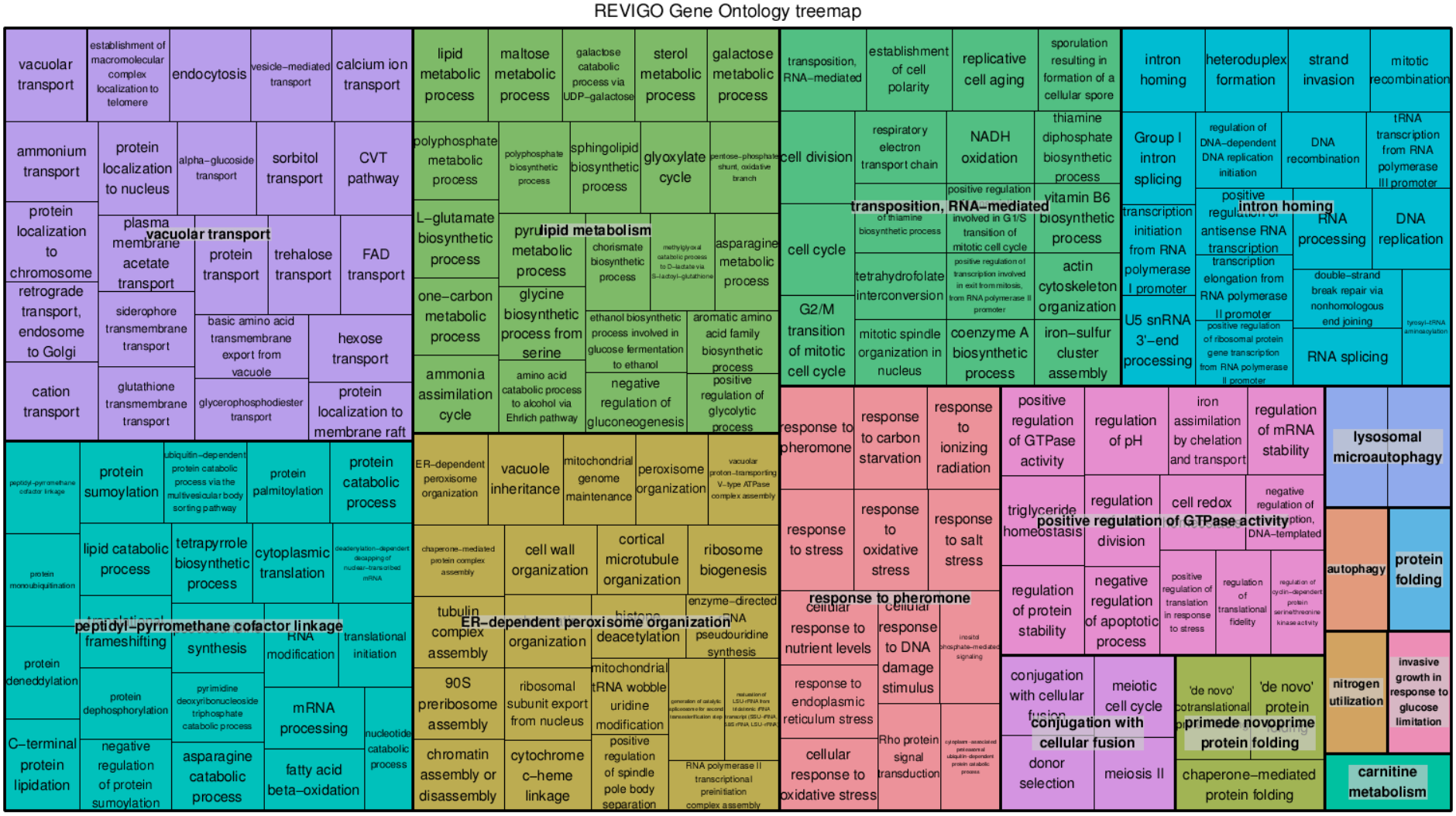

Figura 30 - Processos biológicos gerados pelo REVIGO com base nos GOs dos lncRNA que passaram a ser significantes sob condição de tratamento em S288c.

Além disso, quando comparamos os resultados da análise de interação lncRNA-proteína e interação lncRNA-genes vizinhos, é clara a diferença nas tendências de regulação. Na primeira análise, há uma maior tendência dos lncRNAs terem expressões inversas às suas proteínas-alvo, algo que pode ser explicado pelo caráter de regulação à distância dessas moléculas não-codificantes. Porém, na segunda análise, acontece um efeito diferente, pois, como descrito acima, existe uma maior tendência dos lncRNAs terem expressões similares em *cis* na mesma orientação do que quando comparado com regulações *trans*, tal como já descrito para outros organismos (WANG et al., 2019; YAN et al., 2017). A contraditória característica de haver dois tipos de regulação no mesmo organismo já foi descrita na literatura para mamíferos, na qual o lncRNA *ANRIL* (KONG; HSIEH; ALONSO, 2018; SMOLLE et al., 2017) atua tanto em *cis* (na sua região de vizinhança) quanto em *trans* (distante do seu locus gênico) por meio de modificação de complexos da cromatina. No entanto, aqui apresenta-se o primeiro relato para *S. cerevisiae*.

5.2. Os lncRNAs como potenciais reguladores da expressão gênica de sua vizinhança em S. cerevisiae

Como já aqui reportado, os lncRNAs exercem variados tipos de regulação sobre suas regiões vizinhas. Nesta dissertação, observou-se uma maior tendência dos lncRNAs apresentarem expressões similares dos seus vizinhos, tanto na S288c quanto na BMA64-1A, entre 1-2h, intervalo reportado pelo coeficiente angular k1. Porém, ao longo do tempo (entre 2h e 4h), observa-se que, de forma geral, as relações tendem a se equalizar (**Tabela 13**), principalmente sob estresse. Isso permite propor que os lncRNA podem estar alterando a forma com que regulam a região de vizinhança ao longo do tempo, passando a reprimir ou ativar a transcrição do gene vizinho, por exemplo. Como já demonstrado em outros trabalhos (LI et al., 2018a; TRAN et al., 2016), há diferentes formas dos lncRNAs alterarem

sua regulação ao longo do tempo e de acordo com a necessidade metabólica e estrutural da célula, independente do organismo, indo da indução de uma maior migração de poli-ribossomos para a base dos dendritos por meio de ligações locais à receptores para atender demandas proteicas nas sinapses até o aumento da tradução de determinada proteína em eventos de especiação dos megacariócitos na medula óssea.

Tanto a análise do comportamento geral da relação lncRNA-gene vizinho quanto do comportamento pontual entre os tempos analisados, acabam concordando que a maior parte dos perfis são similares, ou seja, o lncRNA possui uma expressão no mesmo sentido à do seu gene vizinho (lncRNA *up* e gene *up*, por exemplo). Porém, no último ponto coletado (4h), a diferença entre a quantidade de expressões tende a se equalizar. Essa diferença também pode estar relacionada com as regulações pré-transcricionais: considerando os lncRNAs desse trabalho como "guias" que levam à alteração da cromatina dos genes regulados, tal como já observado (WANG; CHANG, 2011), a regulação poderia acontecer eficientemente na região de vizinhança preferencialmente quando o lncRNA está *downstream* ao promotor do gene vizinho (BÖHMDORFER; WIERZBICKI, 2015; YU et al., 2018). Isso foi observado no fungo *Schizossacharomyces pombe* o qual, sob condições normais, o lncRNA não atua diretamente na regulação do seu vizinho. Porém, quando *S. pombe* fica em um meio com escassez de glicose (situação de estresse), seus lncRNAs exercem um papel crítico na modulação da cromatina e numa subsequente ativação do gene. Os autores ainda discutem que provavelmente existem dois tipos de complexos de RNA Polimerase II envolvidos nessa regulação, um que transcreve o lncRNA e outro que se associa a TATA box, explicando a função local de regulação da indução da cromatina (SENMATSU et al., 2019); contudo, sugerimos que uma regulação semelhante à da *S. pombe* pode explicar a diferença nos dados descritos aqui quanto às relações controle *vs* tratamento, principalmente por conta da proximidade evolutiva dessa espécie com *S. cerevisiae*. Alternativamente, sugerimos que com o desenrolar do processo transcricional,

as expressões diferentes tendem a diminuir, uma vez que a característica de equalização das expressões também pode estar relacionada ao tempo que leva para a maquinaria celular fazer o processo de transcrição, já que em fungos geralmente o tempo médio desse processo é de 2-30 mRNAs/hora (PELECHANO; CHÁVEZ; PÉREZ-ORTÍN, 2010).

Além disso, como mostrado na **Tabela 13**, as diferenças regulatórias dos lncRNAs sob diferentes condições ficam evidentes quando se compara os tempos em uma forma mais global, ou seja, sem uma perspectiva de especificidades de cada tempo. Então, observa-se a maior parte das relações concordantes (tanto o lncRNA quanto o seu gene-vizinho estão ou *up* ou *down*-expressos) sendo mantidas tanto na situação controle quanto tratamento. No entanto, com o passar do tempo existe um aumento das relações discordantes, levantando a hipótese de que esses lncRNAs possivelmente também estejam inversamente regulando os seus genes-vizinho. Essa característica inversa já foi observada em outros organismos, como em humanos (GARDING et al., 2013), no qual dois lncRNAs (*DLEU1* e *DLEU2*) causam modificações nas histonas nas suas vizinhanças, levando a uma transcrição mais ativa e uma demetilação significante do DNA nos sítios de início dos lncRNAs. Além disso, observou-se que a super-expressão desses lncRNAs leva à um mecanismo supressor do tumor, sub-expressando os genes vizinhos daquela região cromossômica. Um exemplo dessa característica dos lncRNAs aqui analisados, pode ser observado na linhagem BMA64-1A: existe uma “Relação Concordante” entre o lncRNA transcr_19942 e o gene YGR250C na condição de controle a qual torna-se uma “Relação Discordante” na condição de tratamento. Curiosamente, esse gene é responsável pela restauração do crescimento interrompido por algum tipo de estresse (KOJIMA et al., 2016), e nesse caso, sugere-se que o lncRNA provavelmente alterou sua relação com o gene ao longo do tempo objetivando permitir a tradução do gene YGR250C, de forma que a proteína codificada por esse gene só é mais expressa quando o lncRNA é menos expresso. Existem também um exemplo para a linhagem S288c: o lncRNA transcr_21244 e o gene YBR187W

possuem uma “Relação Concordante” na condição de controle e uma “Relação Discordante” sob tratamento. Esse gene é um transportador de cálcio e magnésio no Golgi, tornando a célula haplo-insuficiente no caso de menor expressão desse gene (THINES et al., 2018). Porém, ao contrário do observado na linhagem BMA64-1A, aqui provavelmente a regulação se deu pela modificação das histonas, já que houve apenas uma pequena mudança no nível de expressão do gene quando comparado aos níveis de expressão no controle *vs.* tratamento, uma vez que a regulação mediado por modificações na cromatina ser mais lenta (ex. na **Figura 31** de Kornienko et al. (2013)).

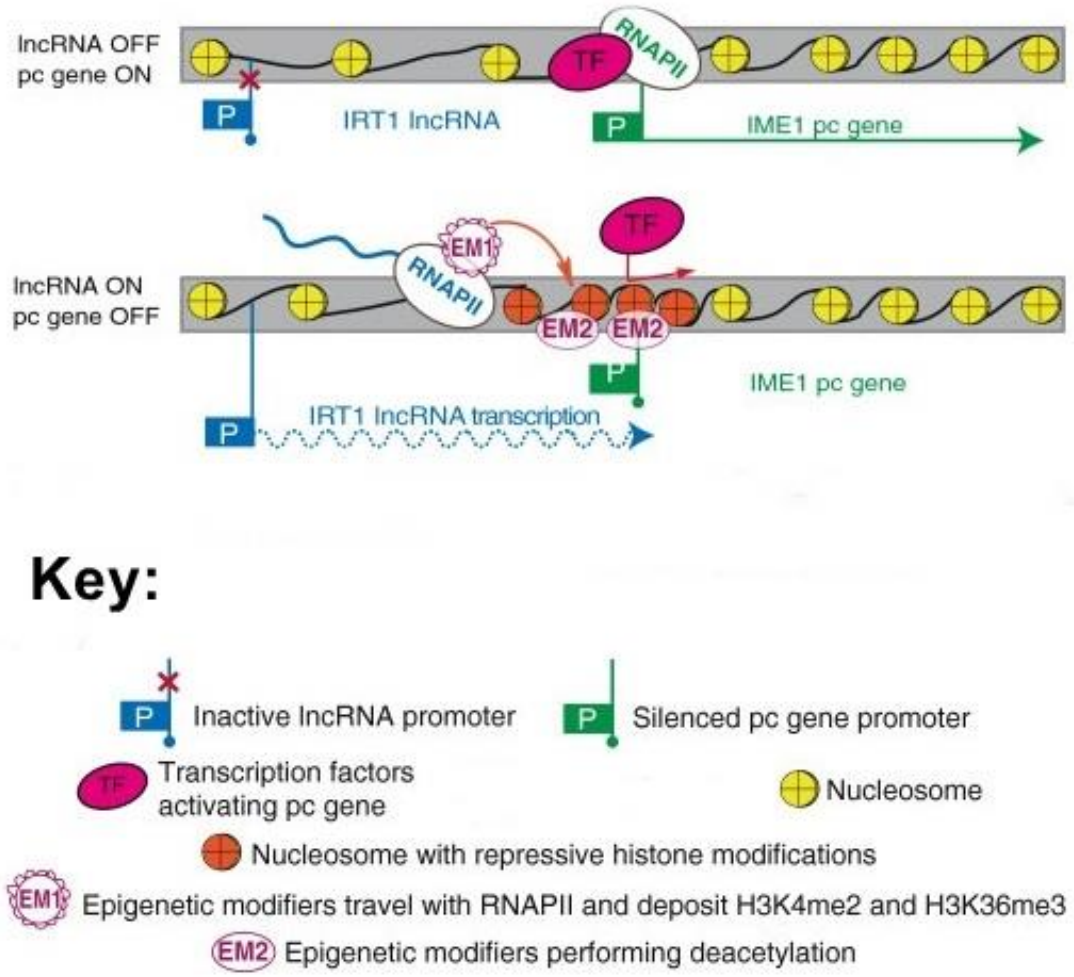

Figura 31 – Interferência da transcrição mediada pelo silenciamento das mudanças da cromatina em *S. cerevisiae*. Acima, ausência do lncRNA *IRT1* que permite a expressão gênica de *IME1*. Abaixo, a RNAPII fazendo a transcrição do lncRNA que carreia os EMs que depositam modificações repressoras de histona no promotor do *IME1*, levando ao bloqueio da ligação de TFs e consequentemente ao silenciamento do gene. Figura adaptada de l et al. (2013).

A importância da regulação gênica pelos lncRNAs fica também evidente na região genômica próxima à essas moléculas. Por exemplo, a **Figura 25-B** mostra um lncRNA

sendo parte do promotor de um gene e, já que os lncRNAs podem atuar como *enhancers* diretamente na região promotora de um determinado gene (KIM; HEMBERG; GRAY, 2015; LAI et al., 2013) mediando o *looping* da cromatina levando à uma ativação coordenada de determinado gene (KADAUKE; BLOBEL, 2009), é provável que o lncRNA transcr_5113 pode estar atuando dessa forma sob seu vizinho YGL024W. De fato, os gráficos dos níveis de expressão também demonstram que ambas as moléculas podem ser reflexos dessa regulação pré-transcricional, já que apresentam níveis semelhantes de expressão ao longo do tempo. Já a **Figura 25-A** mostra uma situação a qual um lncRNA está logo após o final do gene, o qual sugerimos que ele realize uma regulação pós-transcricional. Nesse exemplo, o lncRNA transcr_16017 sob condição de tratamento, tem seu nível de expressão reduzindo ao passo que seu gene vizinho tem sua expressão proporcionalmente aumentada. Aqui, sugerimos que esse lncRNA pode estar regulando o vizinho via um *splicing* alternativo que acaba afetando a transcrição. Isso foi observado para o lncRNA *MALAT1*, o qual interage com fatores de *splicing* e influência na distribuição deles e outros fatores de *splicing* em domínios nucleares. A diminuição no nível de *MALAT1* muda significativamente o *splicing* alternativo do conjunto de pré-mRNAs afetados, aumentando sua expressão (CIARLO et al., 2013; TRIPATHI et al., 2010).

6. Conclusão

Esta é o primeiro trabalho a analisar os lncRNAs frente à tolerância ao etanol em *S. cerevisiae*, além de ter propiciado o maior catálogo disponível até então dessas sequências nessa espécie.

Em uma visão geral, trabalhamos sob a hipótese de que os lncRNAs têm um importante papel no controle da expressão dos seus respectivos vizinhos e que, quando analisados com base na rede de interações lncRNAs-proteínas aqui descritas, existem informações que se sobrepõem e se complementam em grande parte, apesar de possivelmente haver diferenças entre o controle transcricional feito pelos lncRNAs das suas proteínas-alvo (em *trans*) e dos seus genes vizinhos (em *cis*).

Como demonstrado nos exemplos acima, as regulações da região de vizinhança do lncRNA (regulação *cis*) são várias, indo desde regulações pré-transcricionais (atuando em conjunto à cromatina e modificando histonas), às pós-transcricionais (atuando diretamente no processo traducional e sendo fator chave no processo de *splicing* alternativo). Isso acaba levando à uma ligação entre as características regulatórias dos lncRNAs em *cis* e em *trans,* já que elas parecem se relacionar. Quando em *trans*, os lncRNAs aqui identificados parecem atuar de forma mais intensa para combater algum estressor, além de terem uma atuação em uma maior quantidade de vias biológicas (por exemplo, o transcr_18666 da linhagem S288c, o qual está presente 72 vezes nos mais variados mapas metabólicos), ao passo que em *cis,* os lncRNAs parecem não ter uma diferença tão intensa entre as expressões sob condições ótimas *vs.* estresse. Porém, por conta do seu caráter de regulador local (ou seja, acabam tendo uma atuação em uma menor quantidade de locais), esses lncRNAs acabam tendo uma expressão em geral menos variável, levando à manutenção dos processos celulares independente dos fatores externos.

Adicionalmente, confirmamos uma pobre conservação estrutural entre os lncRNA, ao passo que há uma forte conservação evolutiva no que tange aos seus papéis sistêmicos, mostrando a tendência entre diferentes organismos do papel evolutivo desses RNAs.

Conclui-se então, com base nas redes de interação lncRNA-proteínas e suas possíveis forma de interação em *trans*, nas análises de enriquecimento com termos ontológicos diretamente relacionados à tolerância ao estressor aqui discutido e nas análises das relações dos genes vizinhos com suas formas de interação locais em *cis*, que há fortes evidências de que os *long non-coding* RNAs realmente possuem papel ativo no controle da tolerância ao etanol e até no controle da tolerância à outros estressores nas linhagens de *S. cerevisiae* analisadas, decorrente do seu leque de formas de regulação propostas neste trabalho.

7. Referências

ALEXANDRE, H. et al. Global gene expression during short-term ethanol stress in Saccharomyces cerevisiae. **FEBS letters**, v. 498, n. 1, p. 98–103, 1 jun. 2001.

ALMEIDA, L. F. DE. Análise das linhagens de Saccharomyces cerevisiae expostas ao estresse por etanol. 2017.

ALMEIDA, L. F. DE et al. Development and comparative analysis of yeast protein extraction protocols for mass spectrometry. **Analytical Biochemistry**, v. 567, p. 90–95, fev. 2019.

ANDERSON, D. M. et al. A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. **Cell**, v. 160, n. 4, p. 595–606, 12 fev. 2015.

ARRIAL, R. T.; TOGAWA, R. C.; BRIGIDO, M. DE M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus Paracoccidioides brasiliensis. **BMC Bioinformatics**, v. 10, n. 1, p. 239, 4 dez. 2009.

AST, G. How did alternative splicing evolve? **Nature reviews. Genetics**, v. 5, n. 10, p. 773–82, out. 2004.

AUESUKAREE, C. Molecular mechanisms of the yeast adaptive response and tolerance to stresses encountered during ethanol fermentation. **Journal of Bioscience and Bioengineering**, v. 124, n. 2, p. 133–142, ago. 2017.

BAI, F. W.; ANDERSON, W. A.; MOO-YOUNG, M. Ethanol fermentation technologies from sugar and starch feedstocks. **Biotechnology Advances**, v. 26, n. 1, p. 89–105, 1 jan. 2008.

BALAKUMAR, S.; ARASARATNAM, V. Osmo-, thermo- and ethanol- tolerances of Saccharomyces cerevisiae S1. **Brazilian Journal of Microbiology**, v. 43, n. 1, p. 157–166, mar. 2012.

BANKEVICH, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455–477, maio 2012.

BARYSHNIKOVA, A. Systematic Functional Annotation and Visualization of Biological Networks. **Cell Systems**, v. 2, n. 6, p. 412–421, jun. 2016.

BASSLER, J. et al. The AAA-ATPase Rea1 drives removal of biogenesis factors during multiple stages of 60S ribosome assembly. **Molecular cell**, v. 38, n. 5, p. 712–21, 11 jun. 2010.

BELL, J. T. et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. **Genome Biology**, v. 12, n. 1, p. R10, 2011.

BENDJILALI, N. et al. Time-Course Analysis of Gene Expression During the

Saccharomyces cerevisiae Hypoxic Response. **G3: Genes|Genomes|Genetics**, v. 7, n. 1, p. 221–231, jan. 2017.

BIRD, A. DNA methylation patterns and epigenetic memory. **Genes & Development**, v. 16, n. 1, p. 6–21, 1 jan. 2002.

BÖHMDORFER, G.; WIERZBICKI, A. T. Control of Chromatin Structure by Long Noncoding RNA. **Trends in Cell Biology**, v. 25, n. 10, p. 623–632, out. 2015.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics (Oxford, England)**, v. 30, n. 15, p. 2114–20, 1 ago. 2014.

BRACHMANN, C. B. et al. Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. **Yeast (Chichester, England)**, v. 14, n. 2, p. 115–32, 30 jan. 1998.

BROCKDORFF, N. Noncoding RNA and Polycomb recruitment. **RNA**, v. 19, n. 4, p. 429–442, 1 abr. 2013.

CAO, L. et al. LAST, a c-Myc-inducible long noncoding RNA, cooperates with CNBP to promote CCND1 mRNA stability in human cells. **eLife**, v. 6, 4 dez. 2017.

CARRIERI, C. et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. **Nature**, v. 491, n. 7424, p. 454–7, 15 nov. 2012.

CARVER, T. et al. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. **Bioinformatics**, v. 28, n. 4, p. 464–469, 2012.

CHAKRABORTY, S. et al. Biomass to biofuel: a review on production technology. **Asia-Pacific Journal of Chemical Engineering**, v. 7, p. S254–S262, ago. 2012.

CHANDLER, M. et al. **A genomic approach to defining the ethanol stress response in the yeast Saccharomyces cerevisiae**. [s.l: s.n.]. v. 54

CHARLEY, P. A.; WILUSZ, J. Sponging of cellular proteins by viral RNAs. **Current Opinion in Virology**, v. 9, p. 14–18, dez. 2014.

CHATR-ARYAMONTRI, A. et al. The BioGRID interaction database: 2015 update. **Nucleic acids research**, v. 43, n. Database issue, p. D470-8, jan. 2015.

CHEKANOVA, J. A. Long non-coding RNAs and their functions in plants. **Current opinion in plant biology**, v. 27, p. 207–16, out. 2015.

CHERRY, J. M. et al. Saccharomyces Genome Database: The genomics resource of budding yeast. **Nucleic Acids Research**, v. 40, n. D1, p. 700–705, 2012.

CHIN, Y.-W. Comparison of Ethanol Fermentation Properties between Laboratorial and Industrial Yeast Strains using Cassava Hydrolysate. **Korean Journal of Microbiology and Biotechnology**, v. 40, n. 3, p. 220–225, 28 set. 2012.

CHO, S. W. et al. Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. **Cell**, v. 173, n. 6, p. 1398–1412.e22, maio 2018.

CHOI, S.-W.; KIM, H.-W.; NAM, J.-W. The small peptide world in long noncoding RNAs. **Briefings in Bioinformatics**, 29 jun. 2018.

CIARLO, E. et al. An intronic ncRNA-dependent regulation of SORL1 expression affecting A formation is upregulated in post-mortem Alzheimer's disease brain samples. **Disease Models & Mechanisms**, v. 6, n. 2, p. 424–433, 1 mar. 2013.

CRAPPÉ, J.; VAN CRIEKINGE, W.; MENSCHAERT, G. Little things make big things happen: A summary of micropeptide encoding genes. **EuPA Open Proteomics**, v. 3, p. 128–137, 1 jun. 2014.

DE OLIVA NETO, P. et al. **The Brazilian technology of fuel ethanol fermentation-yeast inhibition factors and new perspectives to improve the technology**. [s.l: s.n.]. v. 1

DELLI PONTI, R. et al. A high-throughput approach to profile RNA structure. **Nucleic Acids Research**, v. 45, n. 5, p. e35–e35, 17 mar. 2017.

DELLI PONTI, R. et al. A Method for RNA Structure Prediction Shows Evidence for Structure in lncRNAs. **Frontiers in Molecular Biosciences**, v. 5, 3 dez. 2018.

DEMIRBAS, A. Tomorrow's biofuels: Goals and hopes. **Energy Sources, Part A: Recovery, Utilization, and Environmental Effects**, v. 39, n. 7, p. 673–679, 3 abr. 2017.

DERRIEN, T. et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. **Genome Research**, v. 22, n. 9, p. 1775–1789, 2012.

DEUTSCH, E. W. et al. A guided tour of the Trans-Proteomic Pipeline. **PROTEOMICS**, v. 10, n. 6, p. 1150–1159, 25 jan. 2010.

DIAS, M. O. S. et al. Integrated versus stand-alone second generation ethanol production from sugarcane bagasse and trash. **Bioresource Technology**, 2012.

DING, J. et al. Tolerance and stress response to ethanol in the yeast Saccharomyces cerevisiae. **Applied Microbiology and Biotechnology**, v. 85, n. 2, p. 253–263, 16 nov. 2009.

DU MEE, D. J. M. et al. Efficient termination of nuclear lncRNA transcription promotes mitochondrial genome maintenance. **eLife**, v. 7, 5 mar. 2018.

DUSS, O. et al. Structural basis of the non-coding RNA RsmZ acting as a protein sponge. **Nature**, v. 509, p. 588, 14 maio 2014.

EBISUYA, M. et al. Ripples from neighbouring transcription. **Nature Cell Biology**, v. 10, n. 9, p. 1106–1113, 10 set. 2008.

ENGREITZ, J. M. et al. Local regulation of gene expression by lncRNA promoters,

transcription and splicing. **Nature**, v. 539, n. 7629, p. 452–455, 26 nov. 2016.

EPP, J. A.; CHANT, J. An IQGAP-related protein controls actin-ring formation and cytokinesis in yeast. **Current biology : CB**, v. 7, n. 12, p. 921–9, 1 dez. 1997.

FIEDUREK, J.; SKOWRONEK, M.; GROMADA, A. Selection and adaptation of Saccharomyces cerevisae to increased ethanol tolerance and production. **Polish journal of microbiology**, v. 60, n. 1, p. 51–8, 2011.

FRAENKEL, D. Carbohydrate Metabolism. **Cold Spring Harbor Monograph Archive**, n. 11B, p. 39, 1982.

FRANKISH, A. et al. GENCODE reference annotation for the human and mouse genomes. **Nucleic Acids Research**, v. 47, n. D1, p. D766–D773, 8 jan. 2019.

FU, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics**, v. 28, n. 23, p. 3150–3152, dez. 2012.

GARDING, A. et al. Epigenetic Upregulation of lncRNAs at 13q14.3 in Leukemia Is Linked to the In Cis Downregulation of a Gene Cluster That Targets NF-kB. **PLoS Genetics**, v. 9, n. 4, p. e1003373, 4 abr. 2013.

GARG, A. et al. A long noncoding (lnc)RNA governs expression of the phosphate transporter Pho84 in fission yeast and has cascading effects on the flanking prt lncRNA and pho1 genes. **Journal of Biological Chemistry**, v. 293, n. 12, p. 4456–4467, 23 mar. 2018.

GEISLER, S.; COLLER, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. **Nature Reviews Molecular Cell Biology**, v. 14, n. 11, p. 699–712, 9 nov. 2013.

GHAREIB, M.; YOUSSEF, K. A.; KHALIL, A. A. Ethanol tolerance of Saccharomyces cerevisiae and its relationship to lipid content and composition. **Folia microbiologica**, v. 33, n. 6, p. 447–52, 1988.

GOUD, B. S.; ULAGANATHAN, K. RNA-seq analysis of transcriptomes for assessing stress tolerance of S. cerevisiae strain, NCIM3186. **bioRxiv**, p. 609370, 1 jan. 2019.

GUIL, S.; ESTELLER, M. Cis-acting noncoding RNAs: friends and foes. **Nature Structural & Molecular Biology**, v. 19, n. 11, p. 1068–1075, 6 nov. 2012.

HAARER, B. K. et al. SEC3 mutations are synthetically lethal with profilin mutations and cause defects in diploid-specific bud-site selection. **Genetics**, v. 144, n. 2, p. 495–510, out. 1996.

HAAS, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. **Nature Protocols**, v. 8, n. 8, p. 1494–1512, 11 ago. 2013.

HALLSWORTH, J. E. et al. Compatible solutes protect against chaotrope (ethanol)-induced, nonosmotic water stress. **Applied and environmental microbiology**, v. 69, n. 12, p. 7032–4, dez. 2003.

HE, B. et al. Optimal assembly strategies of transcriptome related to ploidies of eukaryotic organisms. **BMC Genomics**, v. 16, n. 1, p. 65, 2015.

HUANG, M. et al. Efficient protein production by yeast requires global tuning of metabolism. **Nature Communications**, v. 8, n. 1, p. 1131, 25 dez. 2017.

HUANG, X. CAP3: A DNA Sequence Assembly Program. **Genome Research**, v. 9, n. 9, p. 868–877, 1 set. 1999.

INGRAM, L. O. Ethanol tolerance in bacteria. **Critical reviews in biotechnology**, v. 9, n. 4, p. 305–19, 1990.

IZAWA, S.; INOUE, Y. Post-transcriptional regulation of gene expression in yeast under ethanol stress. **Biotechnology and Applied Biochemistry**, v. 53, n. 2, p. 93, 1 jun. 2009.

JIA, K.; ZHANG, Y.; LI, Y. Systematic engineering of microorganisms to improve alcohol tolerance. **Engineering in Life Sciences**, v. 10, n. 5, p. 422–429, out. 2010.

JOHNSON, L. S.; EDDY, S. R.; PORTUGALY, E. Hidden Markov model speed heuristic and iterative HMM search procedure. **BMC Bioinformatics**, v. 11, n. 1, p. 431, 18 dez. 2010.

JOUNG, J. et al. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. **Nature**, v. 548, p. 343, 11 ago. 2017.

KADAUKE, S.; BLOBEL, G. A. Chromatin loops in gene regulation. **Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms**, v. 1789, n. 1, p. 17–25, jan. 2009.

KALVARI, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. **Nucleic Acids Research**, v. 46, n. D1, p. D335–D342, 4 jan. 2018.

KANEHISA, M. et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. **Nucleic acids research**, v. 45, n. D1, p. D353–D361, 2017.

KANG, K. et al. Linking genetic, metabolic, and phenotypic diversity among Saccharomyces cerevisiae strains using multi-omics associations. **GigaScience**, v. 8, n. 4, abr. 2019.

KARUPPIAH, R. et al. Energy optimization for the design of corn-based ethanol plants. **AIChE Journal**, v. 54, n. 6, p. 1499–1525, jun. 2008.

KASAVI, C. et al. A system based network approach to ethanol tolerance in Saccharomyces cerevisiae. **BMC Systems Biology**, v. 8, n. 1, p. 90, 8 dez. 2014.

KIM, D.; LANGMEAD, B.; SALZBERG, S. L. HISAT: a fast spliced aligner with low memory requirements. **Nature Methods**, v. 12, p. 357, 9 mar. 2015.

KIM, J. et al. LncRNA OIP5-AS1/cyrano sponges RNA-binding protein HuR. **Nucleic Acids**

Research, v. 44, n. 5, p. 2378–2392, 18 mar. 2016.
KIM, T.-K. et al. Widespread transcription at neuronal activity-regulated enhancers. **Nature**, v. 465, n. 7295, p. 182–187, 14 maio 2010.
KIM, T.-K.; HEMBERG, M.; GRAY, J. M. Enhancer RNAs: A Class of Long Noncoding RNAs Synthesized at Enhancers: Figure 1. **Cold Spring Harbor Perspectives in Biology**, v. 7, n. 1, p. a018622, 5 jan. 2015.
KOJIMA, R. et al. Identification of multi-copy suppressors for endoplasmic reticulum-mitochondria tethering proteins in Saccharomyces cerevisiae. **FEBS letters**, v. 590, n. 18, p. 3061–70, 2016.
KONG, L. et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. **Nucleic acids research**, v. 35, n. Web Server issue, p. W345-9, jul. 2007.
KONG, Y.; HSIEH, C.-H.; ALONSO, L. C. ANRIL: A lncRNA at the CDKN2A/B Locus With Roles in Cancer and Metabolic Disease. **Frontiers in Endocrinology**, v. 9, 24 jul. 2018.
KORNIENKO, A. E. et al. Gene regulation by the act of long non-coding RNA transcription. **BMC Biology**, v. 11, n. 1, p. 59, 2013.
LAI, F. et al. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. **Nature**, v. 494, n. 7438, p. 497–501, 17 fev. 2013.
LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, 4 abr. 2012.
LEWIS, J. A. et al. Exploiting Natural Variation in Saccharomyces cerevisiae to Identify Genes for Increased Ethanol Resistance. **Genetics**, v. 186, n. 4, p. 1197–1205, dez. 2010.
LI, B. T.; LIM, J. X.; LING, M. H. T. Analyzing Transcriptome-Phenotype Correlations. In: **Encyclopedia of Bioinformatics and Computational Biology**. [s.l.] Elsevier, 2019. p. 819–824.
LI, D. et al. Identification of a novel human long non-coding RNA that regulates hepatic lipid metabolism by inhibiting SREBP-1c. **International Journal of Biological Sciences**, v. 13, n. 3, p. 349–357, 2017.
LI, H. et al. Expression profile of long non-coding RNAs in cardiomyocytes exposed to acute ischemic hypoxia. **Molecular Medicine Reports**, 13 nov. 2018a.
LI, Y. et al. Long noncoding RNA BDNF-AS inversely regulated BDNF and modulated high-glucose induced apoptosis in human retinal pigment epithelial cells. **Journal of Cellular Biochemistry**, v. 119, n. 1, p. 817–823, jan. 2018b.
LI, Z. et al. The Role of Long Noncoding RNAs in Gene Expression Regulation. In: **Gene Expression Profiling in Cancer**. [s.l.] IntechOpen, 2019.

LICATA, L. et al. MINT, the molecular interaction database: 2012 update. **Nucleic Acids Research**, v. 40, n. D1, p. D857–D861, jan. 2012.

LIN, Y. et al. Factors affecting ethanol fermentation using Saccharomyces cerevisiae BY4742. **Biomass and Bioenergy**, v. 47, p. 395–401, 1 dez. 2012.

LIU, S. et al. De Novo Transcriptome Assembly in Chili Pepper (Capsicum frutescens) to Identify Genes Involved in the Biosynthesis of Capsaicinoids. **PLoS ONE**, v. 8, n. 1, p. e48156, jan. 2013.

LONG, Y. et al. How do lncRNAs regulate transcription? **Science Advances**, v. 3, n. 9, p. eaao2110, 27 set. 2017.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology**, v. 15, n. 12, p. 550, 5 dez. 2014.

LU, Q. et al. Computational prediction of associations between long non-coding RNAs and proteins. **BMC Genomics**, v. 14, n. 1, p. 651, 2013.

LUO, W. et al. GAGE: generally applicable gene set enrichment for pathway analysis. **BMC Bioinformatics**, v. 10, n. 1, p. 161, 2009.

MA, L.; BAJIC, V. B.; ZHANG, Z. On the classification of long non-coding RNAs. **RNA Biology**, v. 10, n. 6, p. 924–933, 27 jun. 2013.

MA, M. et al. Long Noncoding RNA GCASPC , a Target of miR-17-3p, Negatively Regulates Pyruvate Carboxylase–Dependent Cell Proliferation in Gallbladder Cancer. **Cancer Research**, v. 76, n. 18, p. 5361–5371, 15 set. 2016.

MA, M.; LIU, Z. L. Mechanisms of ethanol tolerance in Saccharomyces cerevisiae. **Applied Microbiology and Biotechnology**, v. 87, n. 3, p. 829–845, 13 jul. 2010.

MATSUO, Y. et al. Coupled GTPase and remodelling ATPase activities form a checkpoint for ribosome export. **Nature**, v. 505, n. 7481, p. 112–116, 2 jan. 2014.

MERCER, T. R.; MATTICK, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. **Nature Structural & Molecular Biology**, v. 20, n. 3, p. 300–307, 5 mar. 2013.

MUSSATTO, S. I. et al. Technological trends, global market, and challenges of bio-ethanol production. **Biotechnology advances**, v. 28, n. 6, p. 817–830, 2010.

NARENDRANATH, N. V.; POWER, R. Relationship between pH and Medium Dissolved Solids in Terms of Growth and Metabolism of Lactobacilli and Saccharomyces cerevisiae during Ethanol Production. **Applied and Environmental Microbiology**, v. 71, n. 5, p. 2239–2243, 1 maio 2005.

NELSON, B. R. et al. A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. **Science**, v. 351, n. 6270, p. 271 LP-275, 15 jan. 2016.

NEUMANN, P. et al. The lncRNA GATA6-AS epigenetically regulates endothelial gene expression via interaction with LOXL2. **Nature Communications**, v. 9, n. 1, p. 237, 2018.

NIEDERER, R. O.; HASS, E. P.; ZAPPULLA, D. C. Long Noncoding RNAs in the Yeast S. cerevisiae. In: [s.l: s.n.]. p. 119–132.

NING, Q. et al. The Evolution and Expression Pattern of Human Overlapping lncRNA and Protein-coding Gene Pairs. **Scientific Reports**, v. 7, p. 42775, 27 mar. 2017.

NITSCHE, A.; STADLER, P. F. Evolutionary clues in lncRNAs. **Wiley Interdisciplinary Reviews: RNA**, v. 8, n. 1, p. e1376, jan. 2017.

NOVIELLO, T. M. R. et al. Detection of long non–coding RNA homology, a comparative study on alignment and alignment–free metrics. **BMC Bioinformatics**, v. 19, n. 1, p. 407, 6 dez. 2018.

ØROM, U. A. et al. Long Noncoding RNAs with Enhancer-like Function in Human Cells. **Cell**, v. 143, n. 1, p. 46–58, out. 2010.

OSHLACK, A.; ROBINSON, M. D.; YOUNG, M. D. From RNA-seq reads to differential expression results. **Genome biology**, v. 11, n. 12, p. 220, 2010.

PAYTUVÍ GALLART, A. et al. GREENC: a Wiki-based database of plant lncRNAs. **Nucleic Acids Research**, v. 44, n. D1, p. D1161–D1166, 4 jan. 2016.

PELECHANO, V.; CHÁVEZ, S.; PÉREZ-ORTÍN, J. E. A Complete Set of Nascent Transcription Rates for Yeast Genes. **PLoS ONE**, v. 5, n. 11, p. e15442, 16 nov. 2010.

PENG, Y. et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. **Bioinformatics**, v. 28, n. 11, p. 1420–1428, 1 jun. 2012.

PETRYSZAK, R. et al. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. **Nucleic acids research**, v. 42, n. Database issue, p. D926-32, jan. 2014.

PFEIFFER, T.; MORLEY, A. An evolutionary perspective on the Crabtree effect. **Frontiers in Molecular Biosciences**, v. 1, 21 out. 2014.

PORTO, F. W.; DAULATABAD, S. V.; JANGA, S. C. Long Non-Coding RNA Expression Levels Modulate Cell-Type-Specific Splicing Patterns by Altering Their Interaction Landscape with RNA-Binding Proteins. **Genes**, v. 10, n. 8, p. 593, 6 ago. 2019.

PRASERTWASU, S. et al. Efficient process for ethanol production from Thai Mission grass (Pennisetum polystachion). **Bioresource Technology**, v. 163, p. 152–159, 1 jul. 2014.

QIU, Z.; JIANG, R. Improving Saccharomyces cerevisiae ethanol production and tolerance via RNA polymerase II subunit Rpb7. **Biotechnology for Biofuels**, v. 10, n. 1, p. 125, 15 dez. 2017.

QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic

features. **Bioinformatics**, v. 26, n. 6, p. 841–842, 15 mar. 2010.
RABINOVITCH-DEERE, C. A. et al. Synthetic Biology and Metabolic Engineering Approaches To Produce Biofuels. **Chemical Reviews**, v. 113, n. 7, p. 4611–4632, 10 jul. 2013.
REIJNDERS, L. Conditions for the sustainability of biomass based fuel use. **Energy Policy**, v. 34, n. 7, p. 863–876, 1 maio 2006.
RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: The European Molecular Biology Open Software Suite. **Trends in Genetics**, v. 16, n. 6, p. 276–277, jun. 2000.
SAKHARKAR, M. K.; CHOW, V. T. K.; KANGUEANE, P. Distributions of exons and introns in the human genome. **In silico biology**, v. 4, n. 4, p. 387–93, 2004.
SAXONOV, S.; BERG, P.; BRUTLAG, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. **Proceedings of the National Academy of Sciences**, v. 103, n. 5, p. 1412–1417, 31 jan. 2006.
SCHOR, I. E. et al. Non-coding RNA Expression, Function, and Variation during Drosophila Embryogenesis. **Current Biology**, v. 28, n. 22, p. 3547–3561.e9, nov. 2018.
SCHULZ, M. H. et al. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. **Bioinformatics**, v. 28, n. 8, p. 1086–1092, abr. 2012.
SENMATSU, S. et al. lncRNA transcriptional initiation induces chromatin remodeling within a limited range in the fission yeast fbp1 promoter. **Scientific Reports**, v. 9, n. 1, p. 299, 2019.
SHEKHAWAT, K. et al. RNA-seq based transcriptional analysis of Saccharomyces cerevisiae and Lachancea thermotolerans in mixed-culture fermentations under anaerobic conditions. **BMC Genomics**, v. 20, n. 1, p. 145, 18 dez. 2019.
SHI, D.; WANG, C.; WANG, K. Genome shuffling to improve thermotolerance, ethanol tolerance and ethanol productivity of Saccharomyces cerevisiae. **Journal of Industrial Microbiology & Biotechnology**, v. 36, n. 1, p. 139–147, 10 jan. 2009.
SHIH, J.-W. et al. Long noncoding RNA LncHIFCAR/MIR31HG is a HIF-1α co-activator driving oral cancer progression. **Nature Communications**, v. 8, n. 1, p. 15874, 22 ago. 2017.
SINGH, G. et al. The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. **Annual Review of Biochemistry**, v. 84, n. 1, p. 325–354, 2 jun. 2015.
SMIT, A.; HUBLEY, R.; GREEN, P. RepeatMasker Open-4.0. 2013.
SMOLLE, M. et al. Current Insights into Long Non-Coding RNAs (LncRNAs) in Prostate Cancer. **International Journal of Molecular Sciences**, v. 18, n. 2, p. 473, 22 fev. 2017.
STANLEY, D. et al. The ethanol stress response and ethanol tolerance of Saccharomyces

cerevisiae. **Journal of applied microbiology**, v. 109, n. 1, p. 13–24, jul. 2010a.

STANLEY, D. et al. Transcriptional changes associated with ethanol tolerance in Saccharomyces cerevisiae. **Applied Microbiology and Biotechnology**, v. 88, n. 1, p. 231–239, 27 set. 2010b.

STATON, E.; CHEF, B. Pairfq: Pairfq version 0.16.1. **Zenodo**, 2016.

SUPEK, F. et al. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. **PLoS ONE**, v. 6, n. 7, p. e21800, 18 jul. 2011.

TEIXEIRA, M. C. et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. **Nucleic Acids Research**, v. 46, n. D1, p. D348–D353, 4 jan. 2018.

TERBUSH, D. R. et al. The Exocyst is a multiprotein complex required for exocytosis in Saccharomyces cerevisiae. **The EMBO journal**, v. 15, n. 23, p. 6483–94, 2 dez. 1996.

THAMMASITTIRONG, S. N.-R. et al. Ethanol Production Potential of Ethanol-Tolerant Saccharomyces and Non-Saccharomyces Yeasts. **Polish journal of microbiology**, v. 61, n. 3, p. 219–221, 28 set. 2012.

THINES, L. et al. The yeast protein Gdt1p transports Mn2+ ions and thereby regulates manganese homeostasis in the Golgi. **The Journal of biological chemistry**, v. 293, n. 21, p. 8048–8055, 2018.

TILL, P.; MACH, R. L.; MACH-AIGNER, A. R. A current view on long noncoding RNAs in yeast and filamentous fungi. **Applied Microbiology and Biotechnology**, v. 102, n. 17, p. 7319–7331, 4 set. 2018.

TRAN, N. et al. The AS-RBM15 lncRNA enhances RBM15 protein translation during megakaryocyte differentiation. **EMBO reports**, v. 17, n. 6, p. 887–900, 26 jun. 2016.

TRIPATHI, V. et al. The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. **Molecular Cell**, v. 39, n. 6, p. 925–938, set. 2010.

TSAI, M.-C. et al. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. **Science**, v. 329, n. 5992, p. 689–693, 6 ago. 2010.

TZFIRA, T. Molecular Biology of the Cell. Fifth Edition. By Bruce Alberts , Alexander Johnson , Julian Lewis , Martin Raff , Keith Roberts , and , Peter Walter ; with problems by John Wilson and Tim Hunt. Garland Science . New York: Taylor & Francis Group. $142.00 (. **The Quarterly Review of Biology**, v. 83, n. 3, p. 311–311, set. 2008.

WANG, H. et al. The relationship between lysine 4 on histone H3 methylation levels of alcohol tolerance genes and changes of ethanol tolerance in S accharomyces cerevisiae. **Microbial Biotechnology**, v. 7, n. 4, p. 307–314, jul. 2014.

WANG, K. C. et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. **Nature**, v. 472, n. 7341, p. 120–124, 20 abr. 2011.

WANG, K. C.; CHANG, H. Y. Molecular Mechanisms of Long Noncoding RNAs. **Molecular Cell**, v. 43, n. 6, p. 904–914, 2011.

WANG, L. et al. LncRNA Dum interacts with Dnmts to regulate Dppa2 expression during myogenic differentiation and muscle regeneration. **Cell Research**, v. 25, n. 3, p. 335–350, 17 mar. 2015.

WANG, P. et al. Identification and functional prediction of cold-related long non-coding RNA (lncRNA) in grapevine. **Scientific Reports**, v. 9, n. 1, p. 6638, 29 dez. 2019.

WANG, Y. et al. Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice. **Nature Communications**, v. 9, n. 1, p. 3516, 2018.

WARD, C. O. et al. Effect of lysine on toxicity and depressant effects of ethanol in rats. **Toxicology and Applied Pharmacology**, v. 22, n. 3, p. 422–426, jul. 1972.

WILUSZ, J. E.; SUNWOO, H.; SPECTOR, D. L. Long noncoding RNAs: functional surprises from the RNA world. **Genes & Development**, v. 23, n. 13, p. 1494–1504, 1 jul. 2009.

WOHLBACH, D. J. et al. Comparative genomics of Saccharomyces cerevisiae natural isolates for bioenergy production. **Genome biology and evolution**, v. 6, n. 9, p. 2557–66, set. 2014.

WOLF, I. **Identificação de assinaturas sistêmicas associadas à tolerância ao etanol em linhagens de Saccharomyces cerevisiae**. [s.l.] UNESP - Botucatu, 2019.

WU, T. D.; WATANABE, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. **Bioinformatics**, v. 21, n. 9, p. 1859–1875, 1 maio 2005.

XIAO, Y.; ZHANG, J.; DENG, L. Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. **Scientific Reports**, v. 7, n. 1, p. 3664, 16 dez. 2017.

YAMASHITA, A.; SHICHINO, Y.; YAMAMOTO, M. The long non-coding RNA world in yeasts. **Biochimica et Biophysica Acta - Gene Regulatory Mechanisms**, v. 1859, n. 1, p. 147–154, 2016.

YAN, P. et al. Cis- and trans-acting lncRNAs in pluripotency and reprogramming. **Current Opinion in Genetics & Development**, v. 46, p. 170–178, out. 2017.

YOON, J.-H.; ABDELMOHSEN, K.; GOROSPE, M. Posttranscriptional Gene Regulation by Long Noncoding RNA. **Journal of Molecular Biology**, v. 425, n. 19, p. 3723–3730, out. 2013.

YU, F. et al. LnChrom: a resource of experimentally validated lncRNA–chromatin

interactions in human and mouse. **Database**, v. 2018, 1 jan. 2018.

YU, K. O. et al. Increased ethanol production from glycerol by Saccharomyces cerevisiae strains with enhanced stress tolerance from the overexpression of SAGA complex components. **Enzyme and Microbial Technology**, v. 51, n. 4, p. 237–243, set. 2012.

ZABED, H. et al. Bioethanol production from renewable sources: Current perspectives and technological progress. **Renewable and Sustainable Energy Reviews**, v. 71, p. 475–501, 1 maio 2017.

ZHANG, A. et al. LncRNA HOTAIR Enhances the Androgen-Receptor-Mediated Transcriptional Program and Drives Castration-Resistant Prostate Cancer. **Cell Reports**, v. 13, n. 1, p. 209–221, out. 2015.

ZHANG, Y. et al. The importance of engineering physiological functionality into microbes. **Trends in Biotechnology**, v. 27, n. 12, p. 664–672, dez. 2009.

ZHAO, Y. et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. **Nucleic Acids Research**, v. 44, n. D1, p. D203–D208, 4 jan. 2016.

ZHU, J. et al. Function of lncRNAs and approaches to lncRNA-protein interactions. **Science China Life Sciences**, v. 56, n. 10, p. 876–885, 5 out. 2013.

8. Projetos realizados durante o mestrado

Introdução

Durante o período do mestrado foi possível desenvolver uma série de projetos em parceria que geraram manuscritos sem prejuízo para o desenvolvimento do projeto principal desta dissertação.

Artigos em preparação:

1- Tolerância ao etanol em linhagens com diferentes tolerâncias ao etanol sob uma perspectiva integrativa. Ivan R. Wolf, Lucas F. Marques, et al.

2- Montagem e identificação de lncRNAs diferencialmente expressos em tomateiros em resposta à infecção por dois diferentes tipos de fungos, em parceria com o pós-doutorando Eder Marques da Silva da USP-CENA.

3- Montagem de banco de dados de anticorpos anti-HCV em cachorros, em parceria com a professora Rejane Maria Tommasini Grotto da UNESP Botucatu.

4- Análises em larga-escala de genes expressos em mosquitos controle *Aedes aegypti* e com mutação para análise da dispersão da dengue, em parceria com o professor Jayme Augusto de Souza Neto e o pós-doutorando Bruno Tinoco Nunes da UNESP Botucatu.