PROSPECTAR A FUNCIONALIDADE DE RNA LONGOS NÃO CODANTES (LONG NON-CODING RNA) EM FERMENTAÇÕES INDUSTRIAIS DE ETANOL 1G E 2G ATRAVÉS DE REDES DE CO-EXPRESSÃO

Palavras-Chave: Redes de co-expressão, RNA longo não-codante, Bioinformática

Autores(as):

Giovana Leme (autora) [Unicamp]

Prof(ª). Dr(ª). Gonçalo Amarante Guimarães Pereira (orientador) [Unicamp]

Prof(ª). Dr(ª). Lucas Miguel de Carvalho (co-orientador) [Unicamp]

Prof(ª). Dr(ª). Marcelo Falsarella Carazzolle (co-autor) [Unicamp]

Doutorando Jovanderson Jackson Barbosa da Silva (co-autor) [Unicamp]

INTRODUÇÃO:

O bioetanol é uma fonte de energia promissora por possuir baixo impacto econômico e contribuir para a diminuição nas emissões dos gases de efeito estufa (GEEs), sendo assim uma opção mais limpa e renovável (Dos Santos et al., 2016). O Brasil produz um biocombustível denominado "Etanol de primeira geração (etanol 1G)", uma fonte energética limpa e renovável, advinda da fermentação do suco e melaço de cana-de-açúcar através da levedura *Saccharomyces cerevisiae* (Carvalho-Neto et al., 2015). O etanol de segunda geração (etanol 2G) é produzido a partir do aproveitamento da biomassa lignocelulósica da cana-de-açúcar (palha e bagaço), gerada a partir da produção de açúcar e etanol. A biomassa lignocelulósica, por sua vez, é composta por celulose (um homopolímero linear de glicose como regiões cristalinas e amorfas), hemicelulose (um heteropolímero amorfo e ramificado de hexoses e pentoses) e lignina (Cunha et al., 2019). *S. cerevisiae*, principal organismo utilizado na produção industrial, não é capaz de assimilar celulose e hemicelulose diretamente pois a cepa selvagem de *S. cerevisiae* possui a capacidade de fermentar a glicose naturalmente, mas não possui um metabolismo próprio para o consumo de xilose. Desse modo, ao converter efetivamente a xilose em etanol, o custo de produção de biocombustíveis 2G deve diminuir significativamente (Peng et al., 2012), e, para tentar suprir tais gargalos, a engenharia genética é utilizada juntamente com as ferramentas de prospecções da Bioinformática. A biologia de sistemas tenta entender o metabolismo celular através da integração de vários dados ômicos e estratégias e utiliza, além de outras estratégias, o uso dos dados experimentais de

interatoma e das redes de interação entre moléculas de diversos dados ômicos. Assim, um grande esforço é dedicado à caracterização dos dados obtidos, principalmente pela identificação de elementos genômicos funcionais como RNAs mensageiros (mRNAs) e RNAs longos não codificantes (lncRNAs). Os lncRNAs são definidos como transcritos >200 nucleotídeos e que não são traduzidos em proteínas e só recentemente seu papel como reguladores da expressão gênica e sua ligação com doenças genéticas foi revelado (Camargo et al., 2020). Nesse contexto, nosso laboratório tem construído cepas de leveduras geneticamente modificadas para o consumo da xilose e desenvolvido diversos estudos de dados de transcriptoma em fermentações de etanol 1G e 2G com foco em genes codificadores de proteínas. Sendo assim, o objetivo do trabalho foi prospectar a funcionalidade de RNA longos não codantes (long non-coding RNA) em fermentações industriais de etanol 1G e 2G através de redes de co-expressão construídas a partir dos dados de transcriptomas.

METODOLOGIA:

Para realizar este projeto, utilizamos dados públicos de fermentação de etanol 2G que estão depositados no SRA (Sequence Read Archive) sob o número de acesso [SRA057038], oriundos do projeto do artigo de Carvalho, *et. al.*, 2021 (Carvalho *et al,.* 2021). As amostras são provenientes da levedura da espécie *S. cerevisiae* geneticamente modificada para o consumo de xilose (cepa PE-2) e os pontos de fermentação foram coletadas em dornas industriais de fermentação 2G, em diferentes pontos do tempo de fermentação, na usina da GranBio no estado de Alagoas. Os dados de fermentação de etanol 1G que serão utilizados podem ser acessados pelo número de acesso SRA057038. Nesta fermentação de etanol 1G os dados são oriundos de uma fermentação industrial 1G típica (Carvalho-Netto et al, 2015), usando a levedura selvagem PE-2 e os dados foram coletados na destilaria Nova América (Maracaí-SP, Brasil). A análise de RNA-Seq Inicialmente analisamos a qualidade dos reads dos transcriptomas utilizando o FastQC (Leggett, 2013) e removemos os adaptadores com o Trimmomatic (Bolger, 2014). Após a análise de qualidade, os reads foram mapeados contra o genoma de referência de *S. cerevisiae* usando o software Hisat2 (Kim et al,. 2015) A reconstrução dos transcritos a partir do alinhamento foi realizada com o software StringTie (Pertea, et al. 2015). Para quantificar a expressão dos genes, os reads foram alinhados contra o transcriptoma de *S. cerevisiae* com o kallisto (Bray et al., 2016), o que gera uma matriz com o TPM (Transcripts Per Million) de cada transcrito. Para identificar os RNAs longos não-codificantes, utilizamos o RNASamba (Camargo et al., 2020), que é um software de predição baseado em inteligência artificial. Para seu treinamento, os dados para geração do modelo de treinamento de sequências codantes e não codantes foram retirados do Ensembl (https://www.ensembl.org/index.html). Em seguida, aplicamos um filtro para deixar RNAs com comprimento acima de 200 pb e com TPM acima 1. A análise de expressão diferencial dos RNA

não-codificantes e dos genes diferencialmente expressos em fermentação 2G foram realizadas através do Sleuth (Pimentel et al., 2017), considerando um FDR <= 0.05 e |b| > 1.5 (“b” é o valor da aproximação do fold-change pelo Sleuth). Para a fermentação 1G, devido à falta de réplicas biológicas, utilizamos o pacote DEGseq para identificar os RNA não-codificantes e dos genes diferencialmente expressos (p-value <= 0.05 and |fold-change| >1.5) (Wang et al, 2010). Para entender o papel dos RNAs longos não-codificantes vindos de diferentes condições de fermentação (1G e 2G) sobre os genes codantes, realizamos uma análise de redes de co-expressão utilizando o WGCNA. Após a quantificação dos genes, a partir da matriz de expressão gênica completa utilizamos um agrupamento hierárquico de amostras para remover valores discrepantes. Em seguida, o parâmetro soft threshold power (β) foi determinado para garantir uma rede livre de escala. A função potência foi utilizada para transformar a matriz de correlação de Pearson em uma matriz de adjacência, que foi então transformada em uma matriz de sobreposição topológica (TOM). Usando um algoritmo de corte dinâmico, um agrupamento hierárquico foi realizado para agrupar genes semelhantes em um mesmo módulo. Reconstruímos as redes de co-expressão de cada módulo e verificaremos a presença de RNA não codantes em cada uma delas com corte de correlação de 0.60. As redes e subsequentes análises serão realizadas no Cytoscape (Shannon, P. 2003). Com as redes estabelecidas, as funcionalidades dos RNA não codificantes em fermentações de etanol 1G e 2G serão inferidas indiretamente pela anotação funcional dos genes codantes pertencentes à rede. Por todos os genes (codantes e não codantes) estarem co-expressos, iremos supor que ambos atuam com o mesmo papel no metabolismo. As vias KEGG enriquecidas e seus módulos serão identificados através das funções enrichKEGG() e enrichMKEGG() do pacote ClusterProfiler (Wu et al,. 2021) usando R 4.0. O enriquecimento dos processos GO será realizado pelo ShinyGO (Ge, et al. 2020)

RESULTADOS E DISCUSSÃO:

Os dados da fermentação de etanol 1G provém de uma fermentação típica na destilaria Nova América (Maracaí-SP, Brasil) utilizando a cepa selvagem de levedura e os pontos de coleta (em horas) utilizados foram 01, 04, 07, 10, 12 e 15. Por outro lado, os dados da fermentação de etanol 2G foram obtidos na usina Granbio no estado de Alagoas através de uma cepa de levedura modificada para o consumo de xilose e os pontos de coleta (em horas) foram 7, 12, 16 e 24. Em primeira instância, o software FastQC foi utilizado para analisar a qualidade dos reads do transcriptoma. Nenhum deles apresentou adaptadores (deixando inviável o uso do removedor de adaptadores) e todos apresentaram um alto valor phred (> 30). Na fermentação 1G, o tamanho médio de cada read foi 36 pb, sendo em média 2.3017.257 reads por amostra. Já na fermentação 2G, o número de reads em cada amostra variou

de 3.786.300 a 10.799.903 pb e o tamanho médio de cada read variou de 35 a 101 pb. Nenhuma amostra apresenta sequências com baixa qualidade.

Após a análise de qualidade, realizamos o mapeamento dos reads de etanol 1G e 2G contra o genoma de *S. cerevisiae*, usando o HISAT2. Desse modo, foi possível obter a porcentagem dos reads que foram mapeados apenas uma vez no genoma, bem como aqueles mapeados múltiplas vezes ou aqueles que não foram mapeados. Para a fermentação 1G, observamos que a maior porcentagem dos reads em cada amostra encontra-se mapeado apenas uma vez no genoma da levedura. Uma pequena porcentagem se encontra mapeada múltiplas vezes e uma porcentagem ínfima dos reads não se encontra mapeada no genoma. Na fermentação 2G, tem-se de modo semelhante à fermentação 1G que a maior parte do número de reads encontra-se mapeado no genoma apenas uma vez. A partir das matrizes de expressão obtidas através do kallisto, extraímos a expressão dos RNAs não-codificantes com TPM > 1 para cada tempo de fermentação. Observa-se que em etanol 1G, o número de RNAs não-codificantes com expressão com TPM > 1 é maior na primeira hora comparada às demais horas. Cerca de 596 RNAs não codificantes são expressos na primeira hora enquando nas demais o número máximo em atividade é 401. Na fermentação 2G, não é observada tal diferença. O número médio de RNA longo não codificante que está em atividade varia entre 505 e 523 durante todo o período de coleta. Com base nesses resultados, podemos concluir que muitos RNAs não-codantes são expressos durante a fermentação de etanol 1G e 2G, algo que não está descrito na literatura ainda.

Durante a fermentação 1G e 2g, encontram-se respectivamente 39 e 154 e non-codings sendo expressos, respectivamente, sendo que 19 estão presentes de modo simultâneo nas duas condições. Na análise das redes de coexpressão na fermentação 1G, foram observados 105 módulos, cujo mais abundante apresentou 244 transcritos. Já na análise da fermentação 2G, foram encontrados 57 módulos e o mais abundante apresentou 788 transcritos. Na fermentação 2G, dentro da condição “7H” da fermentação, destaca-se o módulo “lightgreen”, “palevioletred”, “light yellow”, que apresentam Já na condição “12H”, destacam-se os módulos “red”, “brown” e principalmente “orange”. Na condição “16H”, destacam-se os módulos “salmon”, “blum” e “magenta” e na condição “24H”, “maroon”, “saddle brown” e “green”.

Nas próximas etapas do projeto, cada módulo será investigado mais afundo, será investigado se há ou não a presença de non-codings. Será feita a inferência da função de cada non-coding identificado.

BIBLIOGRAFIA

Bolger, A. M.; Lohse, M.; & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170, 2014.

Bray, N.; Pimentel, H.; Melsted, P. et al. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525–527, 2016.

Camargo A. P.; Sourkov V.; Pereira G. A. G.; Carazzolle M. F. RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. NAR Genom Bioinform, 2(1), 2020.

Carvalho L. M. et al.Understanding the differences in 2G ethanol fermentative scales through omics data integration. FEMS Yeast Res. 21(4):foab030, 2021.

Carvalho-Netto, V. O. et al. Saccharomyces cerevisiae transcriptional reprograming due to bacterial contamination during industrial scale bioethanol production. Microbial cell factories, v. 14, n. 1, p. 13, 2015.

Cunha, T. J. et al. Xylose fermentation efficiency of industrial Saccharomyces cerevisiae yeast with separate or combined xylose reductase/xylitol dehydrogenase and xylose isomerase pathways. Biotechnology for biofuels, v. 12, n. 1, p. 20, 2019.

Dos Santos, V. L. et al. Second-generation ethanol: the need is becoming a reality. Industrial Biotechnology, v. 12, n. 1, p. 40-57, 2016.

ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics, 36(8), 2628-2629, 2020

Kim, D.; Langmead, B., & Salzberg, S. L. (2015).

HISAT: a fast spliced aligner with low memory requirements. Nature methods, 12(4), 357-360.

Leggett, R. M.; Ramirez-Gonzalez, R. H.; Clavijo, B. J.; Waite, D. & Davey, R. P. Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. Frontiers in genetics, 4, 288, 2013.

Pimentel, H., Bray, N., Puente, S. et al. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods 14, 687–690, 2017.

Peng, B. et al. Improvement of xylose fermentation in respiratory-deficient xylose-fermenting Saccharomyces cerevisiae. Metabolic Engineering, v. 14, n. 1, p. 9-18, 2012.

Pertea, M.; Pertea, G. M.; Antonescu, C. M.; Chang, T. C.; Mendell, J. T. & Salzberg, S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology, 33(3), 290-295, 2015.

Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504, 2003.

Wang L., et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics, 26:136–8, 2010.

World Economic Situation and Prospects 2020. Disponível em: (Acesso: 25 Abril 2022).

Wu, T., et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation, 2(3), 100141, 2021.