Uma abordagem bioinformática para identificar QTLs (quantitative trait loci) em cepas industriais de *Saccharomyces cerevisiae.*

Palavras-Chave: BIOINFORMÁTICA, LEVEDURA, ETANOL, ESTRESSE OXIDATIVO

Eduardo Menoti Silva[1], Juliana Galhardo[1], Juliana José[1], Fellipe Mello[1], Gonçalo Amarantes[1], Marcelo Carazzolle[1]

[1] Instituto de Biologia, Unicamp.

1. Introdução

O Brasil se destaca no cenário mundial pelos altos níveis de produção de etanol, atividade fomentada desde 1975 pelo Programa Nacional do Álcool (Proálcool) instaurado em resposta às crises do petróleo no século XX. Ocupando hoje o posto de segundo maior produtor mundial de biocombustíveis, com uma produção média de 30 bilhões de litros de etanol por ano (RFA), a maior fonte do etanol brasileiro é a cana-de-açúcar, que já provou ser mais econômica, energética e ambientalmente sustentável do que outras alternativas como milho e beterraba (GOLDEMBERG, 2008). Vastas reservas de sacarose e outros açúcares encontrados na cana-de-açúcar servem de insumo para a fermentação alcoólica de leveduras *Saccharomyces cerevisiae* para sintetizar o que é conhecido como etanol de primeira geração. No entanto, o material lignocelulósico residual do bagaço da cana também pode ser reaproveitado para gerar o chamado etanol de segunda geração (DOS SANTOS, 2016).

Apesar do volume de etanol produzido por ano estar em uma crescente nas últimas duas décadas (RFA, 2021), essa produção está em descompasso com a demanda cada vez maior por biocombustíveis, devido principalmente às políticas governamentais destinadas a substituir os combustíveis fósseis. A eficiência da produção e a viabilidade econômica da geração de etanol depende em grande parte da performance fermentativa das leveduras, microrganismos que realizam a fermentação alcóolica nas usinas energéticas. Utilizar práticas biotecnológicas para criar cepas de leveduras mais produtivas e robustas contra os agentes estressores das usinas (e.g. altas temperaturas, estresse oxidativo, baixo pH) é, portanto, uma etapa importante para atender à demanda crescente de biocombustíveis.

O mapeamento de regiões gênicas associadas a traços quantitativos (QTLs) é uma estratégia comum para elucidar as bases genéticas dos fenótipos de interesse, possibilitando o design fino de novas linhagens robustas para determinados tipos específicos de estresse, ou fazer o aprimoramento genético das linhagens industriais já em uso. Esse mapeamento, no entanto, pode ser pouco elucidativo, com centenas de possíveis genes candidatos gerados como resultado. Em vista disso, o presente trabalho se propõe a desenvolver um pipeline de bioinformática que utiliza os mais de 1,000 genomas públicos de levedura sequenciados e fenotipados para identificar as mutações raras na população e correlacionar com as mutações identificadas nas análises de QTL. O objetivo é, partindo desses dados genômicos e fenotípicos, utilizar o ferramental da bioinformática para facilitar a identificação de genes com maior probabilidade de estarem relacionados ao fenótipo em estudo. Como estudo de caso será utilizado os dados já sequenciados do estudo de QTL sobre a resistência ao estresse oxidativo da levedura industrial PE-2, que está em curso no nosso laboratório de Genômica e bioenergia da Unicamp.

2. Metodologia

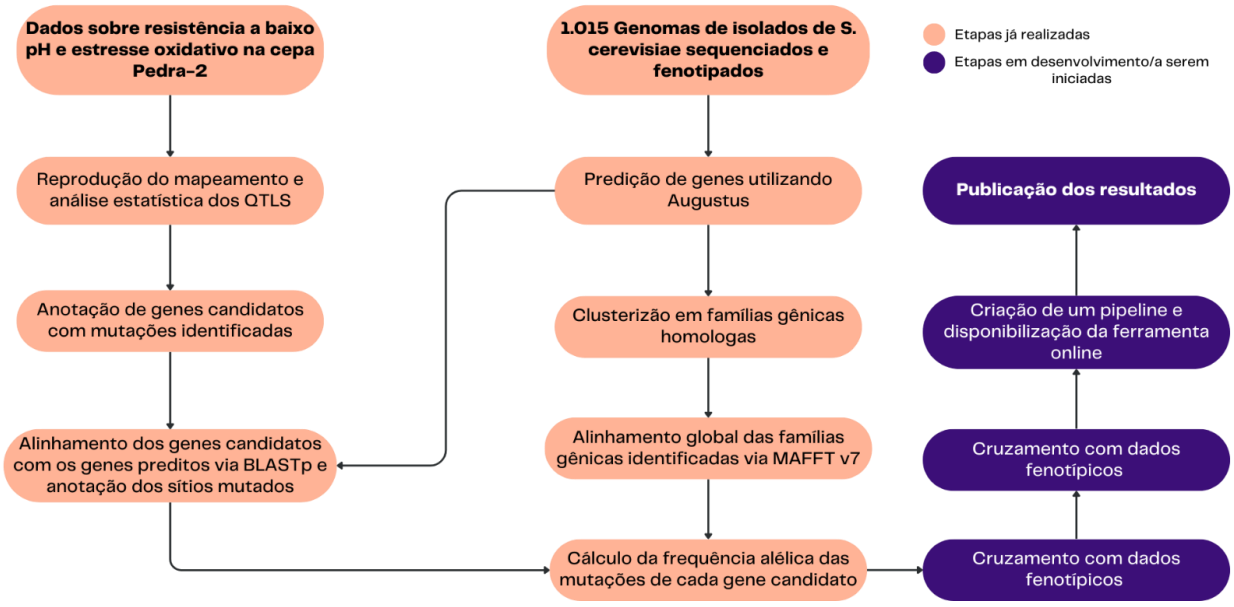

Figura 1. Fluxo de trabalho seguido durante a execução do projeto

2.1 Reprodução do mapeamento e análise estatística de QTLs.

A análise de QTLs foi realizada a partir de milhões de reads pareados (2x100 pb) sequenciados, com o alinhamento dessas sequências no genoma de referência da levedura. O alinhamento foi feito usando o Bowtie2 (LANGMEAD; SALZBERG, 2012), com arquivos de saída no formato SAM, convertidos em arquivos BAM através do SAMTools (LI et al., 2009). Duplicatas de PCR em potencial foram removidas com o Picard "MarkDuplicates" (“Picard Tools – By Broad Institute”, 2019), mantendo apenas o par com a maior qualidade de mapeamento. A identificação das mutações (SNP, inserções e deleções) foi feita usando o GATK (VAN; O’CONNOR, 2020). A análise estatística dos QTLs foi realizada conforme Magwene et al (MAGWENE; WILLIS; KELLY, 2011a), usando o QTLseqR (MANSFELD; GRUMET, 2018), e envolveu o cálculo de uma estatística G modificada para cada SNP. A suavização foi realizada com um Nadaraya-Watson ou kernel de suavização tricube, ponderando os SNPs vizinhos pela distância relativa do SNP focal.

2.2 Conjunto de dados

Foram utilizados dados genômicos de 1.015 isolados da levedura *Saccharomyces cerevisiae*, todos sequenciados e fenotipados, e disponíveis em bancos de dados públicos. Dentre eles, 1.011 genomas foram coletados pelo trabalho de Peter *et al.* (2018). Para estes genomas, foi feito primeiramente a predição gênica com a ferramenta BUSCO, que compreende um dataset de genes ortólogos conservados em ao menos 90% dos indivíduos de um grupo filogenético. Os quatro genomas remanescentes correspondem a isolados de leveduras industriais, descobertas em usinas de etanol brasileiras e com predição gênica disponível publicamente.

2.3 Clusterização de famílias gênicas homólogas.

Este trabalho compõe parte da tese da doutoranda Juliana Pimentel Galhardo. A clusterização das famílias gênicas foi realizada com base na similaridade da sequência de cada gene em conjunto com informações filogenéticas, utilizando o algoritmo Orthofinder v2.2.6 (EMMS; KELLY, 2019). Este é baseado em uma comparação blastp de ‘todos contra todos’, com um corte de e-value permissivo, e clusterização pelo algoritmo MCL (Markov Cluster Algorithm) para posterior reconstrução filogenética de cada ortogrupo.

Durante o desenvolvimento do projeto, foi identificada a ocorrência de erros nos frames de leitura dos genes preditos, causando stop códons prematuros em determinadas sequências. Para driblar este problema evitando a repetição de toda a predição gênica, foi utilizado o software TransDecoder ("TransDecoder", 2023) que busca por open reading frames (ORFs) nas regiões de codificação preditas a fim de corrigir a imprecisão nos frames de leitura. Assim, a análise de clusterização em famílias gênicas está em via de ser refeita com as sequências com frames corrigidos utilizando o algoritmo JustOrthologs (MILLER; PICKETT; RIDGE, 2019).

2.4 Alinhamento múltiplo de proteínas e cálculo da frequência alélica

Para identificação de mutações raras, foi feito o alinhamento múltiplo das proteínas dos ortogrupos. Esta etapa foi desenvolvida em ambiente Linux utilizando tanto a ferramenta BLAST para construção do banco de dados relativo às proteínas dos ortogrupos e realização do alinhamento em si, quanto as linguagens Bash e Python para o desenvolvimento dos scripts de parser dos alinhamentos e identificação de mutações. Foram utilizados os genes os genes SEA2 e SEA3 da cepa de leveduras FMY097 como teste para os scripts desenvolvidos.

Os alinhamentos foram feitos em linha de comando utilizando o software BLAST (Basic Local Alignment Search Tool). Para parsear os resultados e avaliar a qualidade dos alinhamentos foi utilizado como base o BLAST-QC (TORKIAN et al., 2020) script desenvolvido em Python. Foram feitas modificações no script de Torkian *et* al para que informações adicionais sobre query coverage, ocorrência e posição de mutações fossem acusadas nos arquivos gerados como output. Os parâmetros definidos para o alinhamento foram um corte de E-value de 1e-20, identidade mínima de 70% e query coverage de 40%.As proteínas componentes de cada ortogrupo clusterizado pelo OrthoFinder foram alinhadas globalmente utilizando MAFFT v7 (KATOH; STANDLEY, 2013) para posterior utilização no script de cálculo das frequências alélicas das mutações identificadas.

Com base nas mutações identificadas, será realizado o cálculo da frequência alélica por meio de um script Perl que, com base no alinhamento global das proteínas de cada ortogrupo, checa o número de ocorrências de cada mutação identificada nos 1.015 genomas. O output gerado pelo script incluirá as frequências alélicas das mutações, bem como as cepas em que elas ocorrem. Com os isolados *S. cerevisiae* já fenotipados, será possível buscar por correlações entre as mutações e os fenótipos de interesse em etapas posteriores.

3. Resultados e Discussão

O presente projeto parte da reprodução do mapeamento de QTLs em leveduras feita a partir dos dados obtidos na tese do doutor Alessandro Luis Venega Coradini, Cuma prequisa investigou QTLs relacionados à resistência a baxios pHs e ao estresse oxidativo no genoma da levedura industrial Pedra-2. Utilizando a metodologia de Bulk Segregant Analysis, foi feito o sequenciamento dos segregantes isolados ao final do protocolo de mapeamento de QTLs e os reads obtidos foram alinhados com a sequência do genoma de referência CEN.PIK113-7D a fim de identificar SNPS. No total, 43502 SNPs de alta confiança (3.5 SNPs/Kb) entre o pool bom e o pool ruim foram selecionados para a análise de QTL após uma filtragem de qualidade. O mapeamento seguiu o método G' para Bulk-Segregant Analysis proposto por Magwene *et al.* (MAGWENE; WILLIS; KELLY, 2011b), e os dados foram analisados em R através do pacote QTLseqr. Para condições de estresse oxidativo especificamente, uma comparação de genoma inteiro do perfil de SNPs entre dois bulks de segregantes permitiu a identificação de três principais regiões nos cromossomos II, IV e XV.

Tabela **1**: Sumário dos dados de sequenciamentos dos dois pools avaliados.

	Indivíduos	**Pares de Reads**	**Unique (%)**	**Multiple (%)**	**Cobertura**
Pool Bom	81	14.830.448	83.1	1.2	247x
Pool Ruim	77	13.530.082	78.6	1.2	225x

Para facilitar a interpretação de regiões contendo potenciais QTLs o gráfico foi plotado em função do p-valor, uma derivativa direta do G'. As regiões de pico acima do False Discovery Rate (FDR) de 0.01 ($p<0.01$) foram consideradas regiões candidatas (Figura 2).

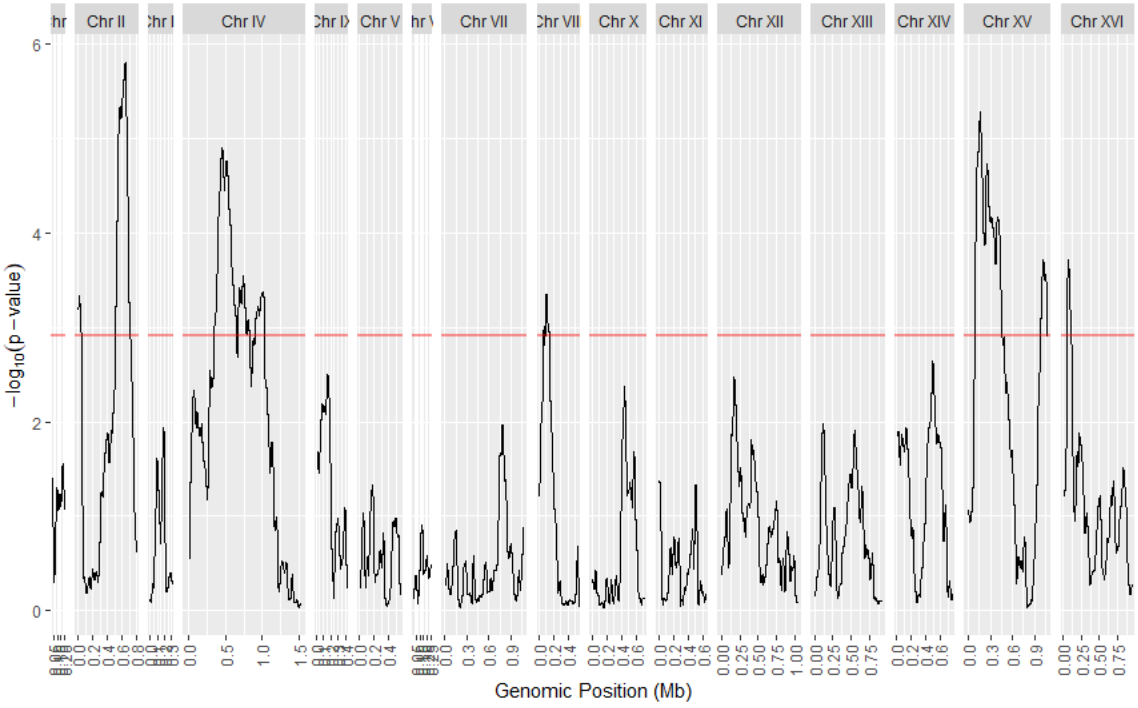

Figura 2. Mapeamento de loci relacionados à resistência ao estresse oxidativo pela análise do sequenciamento de genoma completo de pools de segregantes. O eixo X indica a posição dos cromossomos; o eixo Y indica o valor do -log10(p-valor) calculador para

É nítida a existência de três regiões principais localizadas nos cromossomos II, IV e XV exibindo p-valores acima do limiar de 0.01. Todavia, tais os picos são amplos, albergando dezenas de milhares de pares de bases. Tamanha abrangência dificulta a resolução de QTL a nível de gene por conta do alto número de genes candidatos localizados nestas regiões.

A clusterização em ortogrupos gerou os resultados representados na Tabela 2. No entanto, foram identificados erros nos frames de leitura de alguns dos genes preditos via Augustus, e está em execução uma nova clusterização utilizando o algoritmo JustOrthologs, uma alternativa mais recente ao OrthoFinder que explora a conservação da estrutura dos genes, utilizando o comprimento das regiões de sequência codificante e as porcentagens de dinucleotídeos para identificar ortólogos, reduzindo o tempo de execução em até 96% e obtendo resultados com precisão comparável. A expectativa é que, por meio dessa nova clusterização, seja possível obter informações mais precisas e úteis sobre as relações filogenéticas das leveduras.

Tabela 2. Sumário dos dados obtidos através do agrupamento dos ortólogos via OrthoFinder.

Ortogrupos	Genes ortólogos	Genes ortólogos não atribuídos	Ortólogos de cópia única
25683	5811132	15230	4

Nas etapas de alinhamento e cálculo da frequência alélica, foram utilizados como estudo de caso os genes SEA2 e SEA3 do isolado de levedura FMY097, para o qual foram identificadas mutações raras em outros trabalhos. Utilizando o script em Python desenvolvido *in house*, foi feito um alinhamento local dos dois genes contra um banco de dados contendo as proteínas preditas dos 1,015 genomas de leveduras. O script compara a sequência dos genes fornecidos com a sequência correspondente nos 1,015 genomas, identificando mutações e a frequência com que ocorrem. Foram identificadas para SEA2 4 mutações, e 5 mutações para SEA3.

	F195S	D756-	S830A	L1124S
Freq. SEA2	1.21%	66.44%	53.51%	1.59%

	P620S	D678E	F861Y	E908G	P918S
Freq. SEA3	97.20%	96.81%	2.02%	96.81%	4.15%

Tabelas 3 e 4. Mutações encontradas nos genes SEA2 e SEA3 e suas respectivas frequências. Nesta nomenclatura, a primeira letra diz respeito ao aminoácido presente na sequência de consulta, o número representa a posição da mutação em relação à sequência de referência e a última letra representa o aminoácido presente nesta mesma sequência. Quando presente, o hífen representa gaps no alinhamento.

4. Conclusões

Neste projeto, está sendo desenvolvida uma pipeline bioinformática capaz de refinar os resultados das análises de QTL, ajudando a poupar tempo e custos na identificação de genes quantitativos relacionados a fenótipos de interesse industrial. Até o presente momento, foi possível validar o funcionamento dos scripts desenvolvidos para identificação e cálculo da frequência de mutações raras em estudos de caso nos genes SEA2 e SEA3. Ademais, foram feitas as análises necessárias de predição gênica, clusterização e alinhamento necessárias para fundamentar o funcionamento do script em desenvolvimento, além de estar em curso uma nova clusterização dos genes em ortogrupos. Esta etapa segue em paralelo às etapas subsequentes previstas no plano de trabalho, desenvolvidas ainda com os dados antigos e que serão corrigidas *a posteriori* com os dados corretos de clusterização uma vez que este passo estiver finalizado. Em etapas posteriores, irá ser feita a busca por mutações em todos os genes albergados pelas janelas do genoma definidas pela análise de QTLs (Figura 2), além da análise estatística buscando correlacionar às mutações raras encontradas e o fenótipo de resistência a estresse oxidativo, utilizado como estudo de caso. O pipeline final, no entanto, poderá ser utilizado para estudar qualquer fenótipo quantitativo de leveduras.

Bibliografia

DOS SANTOS, L. V.; DE BARROS GRASSI, M. C.; GALLARDO, J. C. M.; PIROLLA, R. A. S.; CALDERÓN, L. L.; DE CARVALHO-NETTO, O. V.; PARREIRAS, L. S.; CAMARGO, E. L. O.; DREZZA, A. L.; MISSAWA, S. K.; TEIXEIRA, G. S.; LUNARDI, I.; BRESSIANI, J.; PEREIRA, G. A. G. Second-Generation Ethanol: The Need is Becoming a Reality. **Industrial Biotechnology**, vol. 12, no. 1, p. 40–57, 2016. https://doi.org/10.1089/ind.2015.0017.

EMMS, D. M.; KELLY, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. **Genome Biology**, vol. 20, no. 1, p. 1–14, 2019. https://doi.org/10.1186/s13059-019-1832-y

GOLDEMBERG, J. The Brazilian biofuels industry. **Biotechnology for Biofuels**, vol. 1, p. 1–7, 2008. https://doi.org/10.1186/1754-6834-1-6.

KATOH, K.; STANDLEY, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. **Molecular Biology and Evolution**, vol. 30, no. 4, p. 772–780, 2013. https://doi.org/10.1093/molbev/mst010.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, vol. 9, no. 4, p. 357–359, 2012. https://doi.org/10.1038/nmeth.1923.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, vol. 25, no. 16, p. 2078–2079, 2009. https://doi.org/10.1093/bioinformatics/btp352.

MAGWENE, P. M.; WILLIS, J. H.; KELLY, J. K. The statistics of bulk segregant analysis using next generation sequencing. **PLoS Computational Biology**, vol. 7, no. 11, p. 1–9, 2011a.

MANSFELD, B. N.; GRUMET, R. QTLseqr: An R Package for Bulk Segregant Analysis with Next-Generation Sequencing. **The Plant Genome**, vol. 11, no. 2, p. 180006, 2018.

MILLER, J. B.; PICKETT, B. D.; RIDGE, P. G. JustOrthologs: A fast, accurate and user-friendly ortholog identification algorithm. **Bioinformatics**, v. 35, n. 4, p. 546–552, 2019.

PETER, J.; DE CHIARA, M.; FRIEDRICH, A.; YUE, J.-X.; PFLIEGER, D.; BERGSTRÖM, A.; SIGWALT, A.; BARRE, B.; FREEL, K.; LLORED, A.; CRUAUD, C.; LABADIE, K.; AURY, J.-M.; ISTACE, B.; LEBRIGAND, K.; BARBRY, P.; ENGELEN, S.; LEMAINQUE, A.; WINCKER, P.; … SCHACHERER, J. Genome evolution across 1,011 Saccharomyces cerevisiae isolates Species-wide genetic and phenotypic diversity. **Nature**, 2018. .

TORKIAN, B. et al. BLAST-QC: Automated analysis of BLAST results. **Environmental Microbiomes**, v. 15, n. 1, p. 1–8, 2020.

HAAS, B. J.; **TransDecoder**. Disponível em: <https://github.com/TransDecoder/TransDecoder>. Acesso em: 4 maio. 2023.