Montagem do genoma de *Spathaspora arborariae*, uma levedura fermentadora de xilose, para a produção de biocombustíveis

Edmar Melo dos Santos[1]
Francisco Pereira Lobo[2]

A levedura *Spathaspora arborariae* foi isolada de madeira em decomposição coletada nos ecossistemas de mata atlântica e cerrado brasileiros, e tem recebido bastante atenção, em função da sua capacidade de utilizar xilose para produzir quantidades consideráveis de etanol (até $Y_{e/s}$~0.50 g etanol g-1 xilose) (CADETE et al., 2009). A biomassa celulósica é uma fonte de biocombustíveis subutilizada, e leveduras do gênero Spathaspora possuem potencial biotecnológico para prover genes, enzimas e o arcabouço genômico para engenharia de linhagens visando à produção eficiente de etanol a partir de biomassa renovável (WOHLBACH et al., 2011). O presente trabalho visou montar o genoma de *S. arborariae* a partir de dados de sequenciamento genômico.

O sequenciamento do genoma foi realizado utilizando a estratégia de *whole-genome shotgun* com a plataforma 454. A montagem final foi realizada utilizando o montador Newbler com os parâmetros-padrão (MARGULIES et al., 2005). A técnica de eletroforese de campo pulsátil, *pulsed field gel eletrophoresis* (PFGE), foi utilizada para se estimar o número de cromossomos e o tamanho médio do genoma de *S. arborariae*. A predição gênica *ab initio* foi realizada utilizando o programa GeneMark-hmm com os parâmetros padrão, e treinado com os arquivos hmm de *Saccharomyces cerevisisae* (BORODOVSKY; LOMSADZE, 2011). A anotação gênica automática foi realizada utilizando o programa BLAST+

[1] Ciências Biológicas, UNICAMP, edmarms@cnptia.embrapa.br
[2] Embrapa Informática Agropecuária, francisco.lobo@embrapa.br

(CAMACHO et al., 2009) para realizar buscas no banco de dados nr, com pontos de corte de: 1) e-value de 10-5; 2) identidade mínima de 50% e 3) porcentagem mínima de alinhamento da query de 80%. A predição de genes de RNA ribossomal (rRNA) foi feita utilizando o programa RNAmmer com os parâmetros padrão (LAGESEN et al., 2007), e a predição de RNA transportador (tRNA) foi realizada utilizando o programa tRNAscan-SE, escolhendo-se os parâmetros para maximizar a sensibilidade (LOWE; EDDY, 1997). Para validar a montagem utilizamos as únicas sequências disponíveis de *S. arborariae* no NCBI, as quais correspondiam a diferentes porções do gene de rRNA dessa espécie. Essas sequências de rRNA foram então alinhadas ao gene de rRNA predito a partir do genoma de *S. arborariae*, de modo a detectar possíveis erros de montagem quando comparadas à sequência conhecida de rRNA dessa espécie (CADETE et al., 2009).

O dado bruto resultante do sequenciamento compreendia um total de 915.700 *reads* contendo 657.682 *paired-ends*, totalizando 291.670.584 nucleotídeos sequenciados. O tamanho estimado do genoma de acordo com o PFGE foi de 12 Mb, de modo que a cobertura média de sequenciamento foi de 23X. A montagem final continha 439 *contigs* e 41 *scaffolds*, e um tamanho final de 12.708.019 pb após excluírmos os 162.563 nucleotídeos não-determinados (Ns ou Xs, aproximadamente 1% da montagem).

O genoma montado possui um N50 de ~679 kb (6 *scaffolds*) e um N90 de ~ 202 kb (18 *scaffolds*), e um conteúdo GC de 31,7%, compatível com outros genomas proximamente relacionados, filogeneticamente (WOHLBACH et al., 2011). Nós detectamos 6595 genes de tamanho superior a 100 nucleotídeos, dos quais 5569 possuíam sequências similares no banco de dados nr. Localizamos os genes de rRNA no *scaffold* 9, e 187 genes de tRNA espalhados ao longo do genoma. As sequências de rRNA disponíveis no banco de dados NCBI alinharam-se com 100% de identidade ao gene de rRNA predito em nosso estudo, demonstrando que a nossa montagem é coerente com as únicas sequências de *S. arborariae* disponíveis a partir de outras fontes. As próximas etapas do projeto compreenderão a busca por possíveis genes envolvidos no metabolismo de xilose.

Agradecimentos

À Embrapa, por fornecer a bolsa do estagiário e a infraestrutura computacional para a realização deste trabalho.

Referências

BORODOVSKY, M.; LOMSADZE, A. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. In: BAXEVANIS, A. D. et al. **Current Protocols in Bioinformatics**. New York: J. Willey, 2011. Cap. 4:Unit 4.6.1-10. Doi: 10.1002/0471250953.bi0406s35.

CADETE, R. M.; SANTOS, R. O.; MELO, M. A.; MOURO, A.; GONCALVES, D .L.; STAMBUK, B. U.; GOMES, F. C.; LACHANCE, M. A.; ROSA, C. A. *Spathaspora arborariae* sp. nov., a d-xylose-fermenting yeast species isolated from rotting wood in Brazil. **FEMS Yeast Research**, v. 9, n. 8, p. 1338-1342, Dec. 2009.

CAMACHO, C.; COULOURIS, G.; AVAGYAN, V. M. A. N.; PAPADOPOULOS, J.; BEALER, K.; MADDEN, T. L. BLAST+: architecture and applications. **BMC Bioinformatics**, v. 10, p. 421, Dec. 2009. Doi: 10.1186/1471-2105-10-421.

LAGESEN, K.; HALLIN, P.; RODLAND, E. A.; STAERFELDT, H. H.; ROGNES, T.; USSERY, D. W. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. **Nucleic Acids Research**, v. 35, n. 9, p. 3100-3108, Apr. 2009.

LOWE, TM.; EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Research**, v. 25, p. 955-964, 1997.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, n. 437, p. Sept. 376-380, 2005. Doi:10.1038/nature03959

WOHLBACH, D. J.; KUO, A.; SATO, T. K.; POTTS, K. M.; SALAMOV, A. A.; LABUTTI, K. M.; SUN, H.; CLUM, A.; PANGILINAN, J. L.; LINDQUIST, E. A.; LUCAS, S.; LAPIDUS, A.; JIN, M.; GUNAWA, C.; BALAN, V.; DALE, B. E.; JEFFRIES, T. W.; ZINKELL, R.; BARRY, K. W.; GRIGORIEV, I. V.; GASCH, A. P. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. **PNAS**, v. 108, p. 13212-13217, 2011. Proceedings of the National Academy of Sciences of the United States of America.