Genomic and mitochondrial assembly of xylose-fermenting yeasts from Debaryomycetaceae family

Andrei Stecca Steindorff[1], Marcelo Soares Souza[2], João Ricardo Moreira de Almeida[3], Eduardo Fernandes Formighieri[4]

Abstract

Cellulosic biomass, especially sugarcane, is an abundant and underused substrate for biofuel production. The poor efficiency of many microbes to metabolize the xylose abundant in hemicellulose fraction of biomass creates challenges for microbial biofuel production derived from pentose sugars. In this study, we present the genome draft assembly of three yeasts from *Debaryomycetaceae* family, as well as the assembly and annotation of mitochondrial genomes from these isolates. Low coverage assemblies were used to assembly mtDNA. tRNA carring all aminoacids and rRNA (SSU, LSU) were found. Coding genes for COX, COB, NADH dehydrogenase complex, ATP synthase were found as well, excepting ATP8, absent in all three mtDNA. Different strategies were used to assembly total genomes. As result, assemblies with coverages ranging from 80X – 200X and mtDNA reads removed, showed better assembly metrics.

Introduction

The production of cellulosic biofuels presents an economic and environmental challenge and opportunity. From lignocellulosic materials, which include agricultural residues such as sugar-cane bagasse, xylose is the second-most abundant sugar, after glucose. Xylose consumption and consequent fermentation to alcohol on yeasts depends on assimilation enzymes, including xylose reductase, xylitol dehydrogenase and xylulokinase, as well as correct balance of NADH/NAD+ under anaerobic conditions. However, natural xylose fermentation remains slow

[1] Biólogo, doutor em Biologia Molecular, Universidade de Brasília, andreistecca@gmail.com
[2] Informata, Universidade Católica de Salvador, marcelo@libertais.org
[3] Biólogo, doutor em Microbiologia Aplicada, pesquisador da Embrapa Agroenergia, joao.almeida@embrapa.br
[4] Engenheiro-agrônomo, doutor em Biologia Funcional e Molecular, pesquisador da Embrapa Agroenergia, eduardo.formighieri@embrapa.br

and inefficient in the majority of yeasts tested so far. Therefore, improving xylose utilization in industrially relevant yeasts is essential for producing economically viable biofuels from cellulosic materials (WOHLBACH et al., 2011). In this aspect, yeasts from *Debaryomycetaceae* family are promising candidates for xylose-fermenting experiments (LOPES et al., 2016). Here we present the genome draft assembly of three yeasts from *Debaryomycetaceae* family, as well as the assembly and annotation of mitochondrial genomes from these isolates. All work was developed by Bioinformatics Research Group at the Bioinformatics and Bioenergy Laboratory – LBB. All software used are free to use and run on Linux OS.

Material and methods

Yeasts isolation, sequencing and quality control

Strains A1, A5 and A9 were obtained from wood decaying samples in Brasília, Distrito Federal, Brazil (unpublished data). For each strain, genomic DNA (gDNA) was isolated and sent for sequencing through two strategies: short inserts (Illumina Miseq paired-end 2x250bp) and long inserts (Illumina Hiseq 2000 paired-end long jump distance 2x125 bp with 3Kb insert size).

FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to evaluate the libraries quality before and after trimming. For quality trimming and sequence filtering, the software NGS QC Toolkit (version 2.3.3, www.nipgr.res.in/ngsqctoolkit.html) was employed to remove sequencing adapters residues and low quality reads.

The wet lab work was carried out by Microbial Biotechnology and Yeasts Research Group, and the analysis by Bioinformatics Research Group (LBB/Embrapa Agroenergia).

Genomes assemblies and mitochondrial genomes annotation

Mitochondrial genomes (mtDNA) were assembled using the premise that everything with high sequencing coverage will assemble first at low coverage assemblies. For this purpose, 1x and 2x coverage (based on genome size) assemblies were performed using AllPaths-LG (GNERRE et al., 2011). All assemblies (mtDNA and genomes) present same given coverage of short and

long inserts (2X assembly means 1X of short + 1X of long inserts, 40X is 20X short + 20X long, and so on).

The "clean" datasets for assemblies were obtained with mapping, after quality control, of fastq files onto the previously yeast assembled mitochondrial genome using bwa software (http://bio-bwa.sourceforge.net/), and unmapped reads were considered "mtDNA clean data", rescued using samtools function 'view -f 12 -F 256'.

Trimmed fastq files were used for *de novo* assembly using AllPaths-LG varying coverages from 40x to 240x. Metrics used on Figure 1 were calculated using a custom per script developed in house. The mitochondrial genomes were annotated using MITOS pipeline (BERNT et al., 2013). This pipeline predicts coding genes, tRNA and rRNA, and makes functional categorization of each structure.

Results and discussion

Regarding the mitochondrial assembly, as expected, the first scaffold from 1x or 2x assembly was mtDNA and the second ribosomal region. In order to verify the presence of residual pieces from mitochondrial genome in other scaffolds of clean assembly, nine complete mtDNA from yeasts were used in blastn comparison and no match was found. Table 1 shows the size of each mtDNA and coverage from each library. It reveals the importance of clean sequencing files before assembly due to the high coverage for such small region, compared to total genome.

Table 1. Mitochondrial genome assembly features. Percent values indicate the representation of mtDNA at sequenced data for each yeast.

	Size	%GC	Miseq	Coverage	LJD 3kb	Coverage
A1	23.3 Kb	29.89	1.80%	5,174X	1.20%	2,472X
A5	32.6 Kb	25.19	6.54%	14,392X	11.40%	12,895X
A9	32.6 Kb	23.41	3.40%	7,704X	3.80%	4,740X

Structural and functional annotation was performed using MITOS pipeline (BERNT et al., 2013). The Figure 1 presents linearized-mapping genome of the

three mtDNAs, all presenting low GC content, ranging from 23.41% to 29.89% (Table 1). They are composed of 39, 48 and 46 genes (A1, A5 and A9, respectively. Results in this order from now on), representing 63, 53 and 56% of total mtDNA and being transcribed by both strands. Among the genes predicted, there are 14, 19 and 20 protein coding genes, including all three subunits of cytochrome c oxidase (COX1, COX2 and COX3), the apocytochrome b (COB), two subunits of ATP synthase (ATP6 and ATP9) and seven subunits of NADH dehydrogenase complex 1 (NAD1-6 and NAD4L). LSU and SSU genes code the large and the small rRNA genes, whereas the other represent 24 tRNA genes able to carry all amino acids in all three yeasts. An interesting fact is the absence of gene ATP8 in all three mtDNA. We compared by sequence similarity (standalone tblastn) and did not find any match in mtDNA or nuclear genome scaffolds.

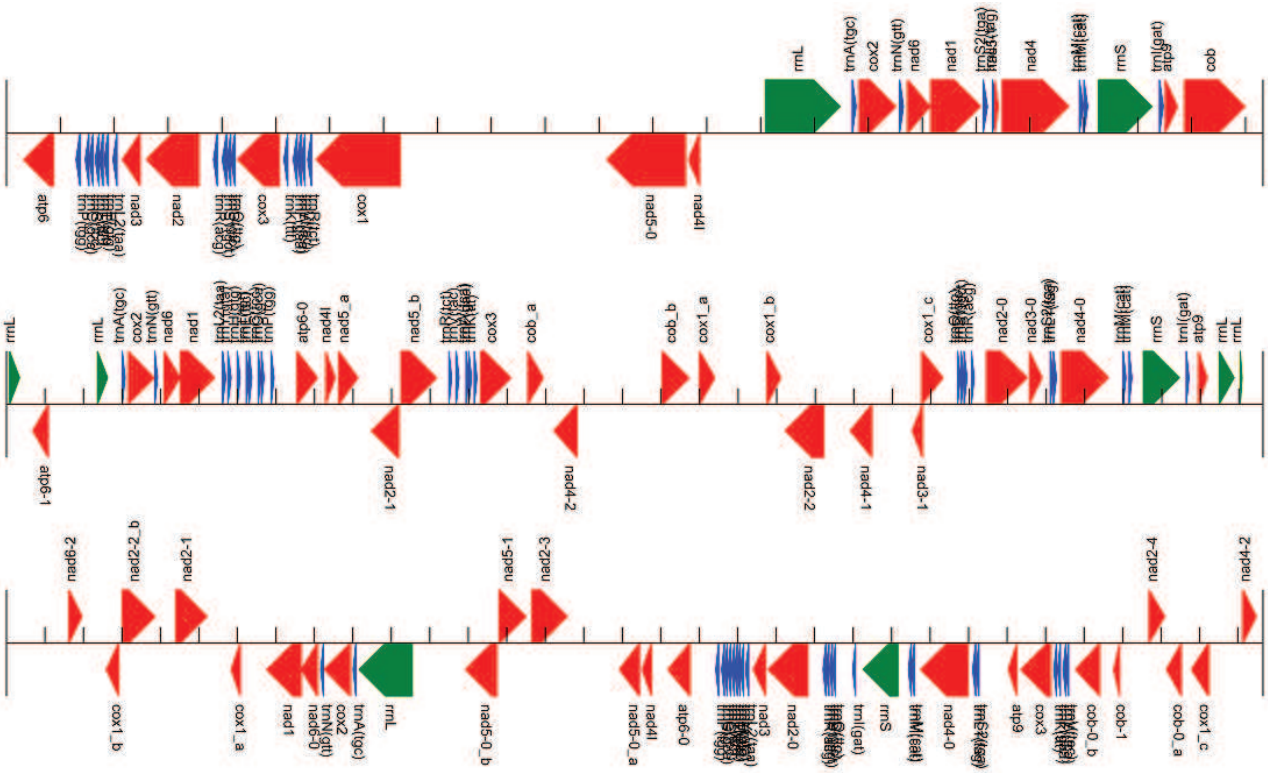

Figure 1. Structural and functional annotation of mitochondrial genomes. Coding genes in red: subunits of cytochrome c oxidase (COX1, 2, 3), the apocytochrome b (COB), ATP synthase (ATP6 and9) and NADH dehydrogenase complex (NAD1-6 and NAD4L); rRNA in green: LSU and SSU genes; tRNA in blue: trnA(tgc), trnC(gca), trnD(gtc), trnE(ttc), trnF(gaa), trnG(tcc), trnH(gtg), trnI(gat), trnK(ttt), trnL1(tag), trnM(cat), trnN(gtt), trnP(tgg), trnQ(ttg). trnR(acg), trnS(gct), trnT(tgt), trnV(tac), trnW(tca), trnY(gta).

Doutores

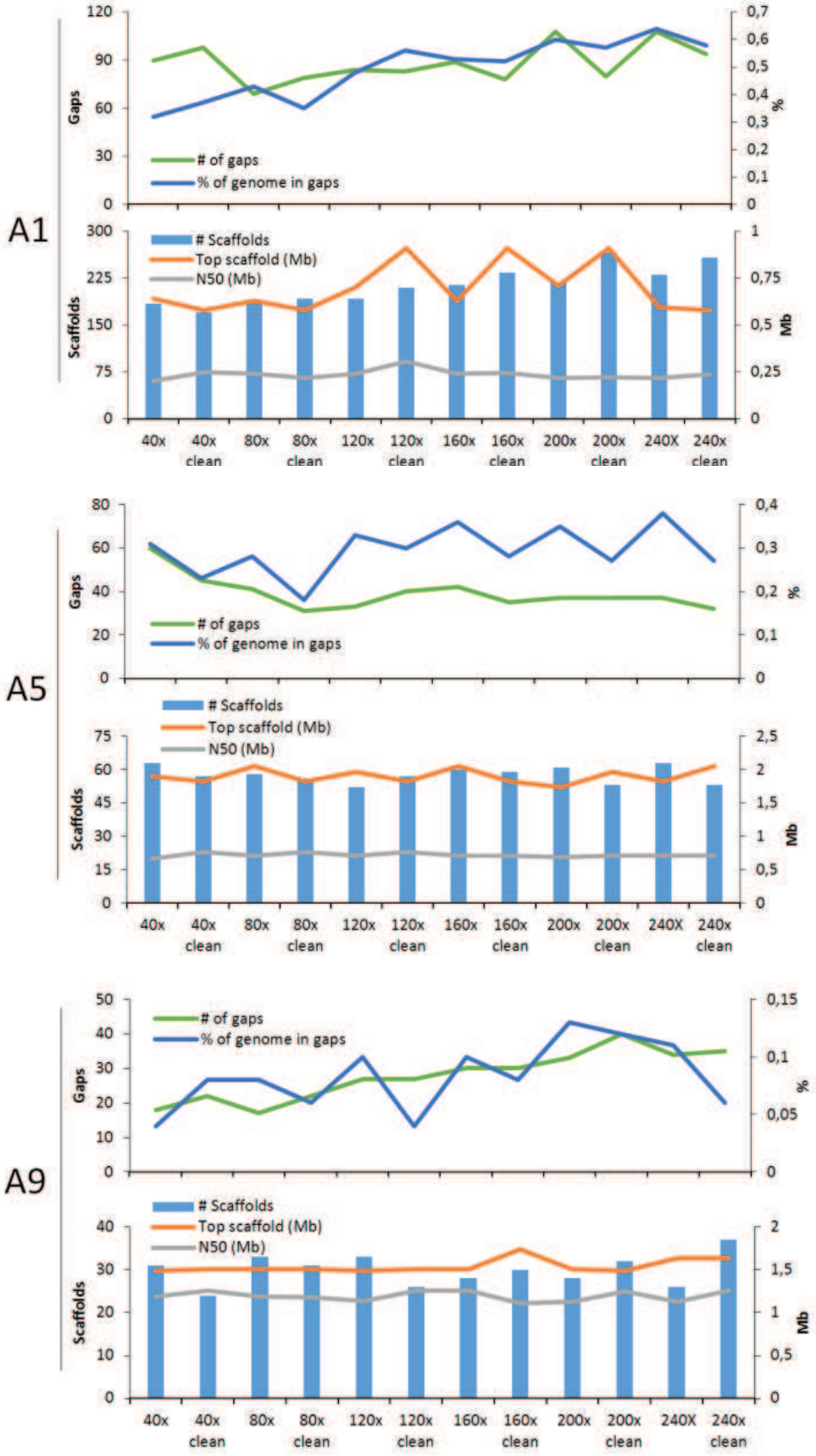

Figure 2. Assembly profile of yeasts genome using different coverages. All assemblies were performed using AllPaths-LG software. "Clean" in this context means that clean assemblies does not contain mtDNA sequences.

Concerning nuclear genome, aiming to reconstruct the genome as accurately as possible, we performed different rounds of genome assembly, varying total genome coverage from 40X to 240X. Figure 2 shows relevant metrics regarding each assembly: Number of gaps, percentage of genome in gaps, number of scaffolds, size of major scaffold and N50. Genome size variation was minimal between assemblies (<0,1%), therefore, was not included in the figure.

The number and size of genome gaps included into the assembly grow up in the same pace of coverage, with the exception of A5 yeast, that reduces and stabilizes after 80X (Figure 2). Other metrics did not change when we increase coverage. The most relevant information of the Figure 2 is that assemblies without organelle reads (clean assemblies) were improved in all metrics for all yeasts analyzed. Based on Figure 1, the sweet spot for better assembly metrics was between 80x and 200x for all yeasts. The size of genomes was comparable with yeasts from *Debaryomycetaceae* family: A1 – 14.5 Mb; A5 – 14.8; A9 – 10.3 Mb. The next step will be the structural and functional annotation of these three yeasts genomes and analysis regarding the phylogenetic relationship between mitochondrial genes and ribosomal regions of yeasts.

Conclusions

In this study, we assembled and annotated the mitochondrial genome from three different xylose-fermenting yeasts. tRNA carring all aminoacids and rRNA (SSU, LSU) were found. Coding genes (COX, COB, NADH dehydrogenase complex, ATP synthase) were found as well, excepting ATP8, absent in all three mtDNA.

After that, nuclear genome was assembled with and without (clean) mitochondrial sequences. The clean approach improved significantly the final assembly in basically all coverages used. This result showed the importance of clean sequences from “not nuclear” regions before assemblies in order to remove coverage bias and therefore, improve genome assemblies. The next step of this work will be the structural and functional annotation of these nuclear genomes.

Financial support

This study was supported by the Brazilian Agricultural Research Corporation (Embrapa) through a grant provided by 'Banco Nacional de Desenvolvimento Econômico e Social' (BNDES).

References

BERNT, M.; DONATH, A.; JÜHLING, F.; EXTERNBRINK, F.; FLORENTZ, C.; FRITZSCH, G.; PÜTZ, J.; MIDDENDORF, M.; STADLER, P. F. MITOS: improved de novo metazoan mitochondrial genome annotation. **Molecular Phylogenetics and Evolution,** San Diego, v. 69, n. 2, p. 313-319, 2013.

GNERRE, S.; MACCALLUM, I.; PRZYBYLSKI, D.; RIBEIRO, F. J.; BURTON, J. N.; WALKER, B. J.; SHARPE, T.; HALL, G.; SHEA, T. P.; SYKES, S.; BERLIN, A. M.; AIRD, D.; COSTELLO, M.; DAZA, R.; WILLIAMS, L.; NICOL, R.; GNIRKE, A.; NUSBAUM, C.; LANDER, E. S.; JAFFE, D. B. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. **Proceedings of the National Academy of Sciences of the United States of America, Washington, DC,** v. 108, n. 4, p. 1513-1518, 2011.

LOPES, M. R.; MORAIS, C. G.; KOMINEK, J.; CADETE, R. M.; SOARES, M. A.; UETANABARO, A. P.; FONSECA, C.; LACHANCE, M. A.; HITTINGER, C. T.; ROSA, C. A. Genomic analysis and D-xylose fermentation of three novel *Spathaspora* species: *Spathaspora girioi* sp. nov., *Spathaspora hagerdaliae* f. a., sp. nov. and *Spathaspora gorwiae* f. a., sp. nov. **Fems Yeast Research,** Oxford, v. 16, n. 4, 2016.

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.; GUNAWAN, C.; BALAN, V.; DALE, B. E.; JEFFRIES, T. W.; ZINKEL, R.; BARRY, K. W.; GRIGORIEV, I. V.; GASCH, A. P. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. **Proceedings of the National Academy of Sciences of the United States of America, Washington, DC,** v. 108, n. 32, p. 13212-13217, 2011.