BMC Research Notes

DATA NOTE

Open Access

Transcriptomics data of 11 species of yeast identically grown in rich media and oxidative stress conditions

William R. Blevins[1,2*], Lucas B. Carey[2,3] and M. Mar Albà[1,4]

Abstract

Objective: The objective of this experiment was to identify transcripts in baker's yeast (*Saccharomyces cerevisiae*) that could have originated from previously non-coding genomic regions, or de novo. We generated this data to be able to compare the transcriptomes of different species of Ascomycota.

Data description: We generated high-depth RNA sequencing data for 11 species of yeast: *Saccharomyces cerevisiae, Saccharomyces paradoxus, Saccharomyces mikatae, Saccharomyces kudriavzevii, Saccharomyces bayanus, Naumovia castelii, Kluyveromyces lactis, Lachancea waltii, Lachancea thermotolerans, Lachancea kluyveri*, and *Schizosaccharomyces pombe*. Using RNA-Seq from yeast grown in rich and oxidative conditions we created genome-guided de novo assemblies of the transcriptomes for each species. We included synthetic spike-in transcripts in each sample to determine the lower limit of detection of the sequencing platform as well as the reliability of our de novo transcriptome assembly pipeline. We subsequently compared the de novo transcripts assemblies to the reference gene annotations and generated assemblies that comprised both annotated and novel transcripts.

Keywords: RNA-seq, Yeast, Transcriptomics, De novo transcript assembly, De novo gene, Gene annotation

Objective

Due to pervasive transcription and pervasive translation in these yeast, new transcripts and ORFs can quickly appear in non-genic sequences and become exposed to selection. This process, known as de novo gene birth, can lead to the appearance of new genes with entirely novel functions. Our objective was to identify and characterize putative de novo genes in baker's yeast to further understand the phenomenon of de novo gene birth. To correctly classify putative de novo genes via the taxonomic conservation of these unique sequences, we need comparable data for a set of closely related species. Due to the similarity of molecular pathways to more complex eukaryotes coupled with their ease of growth in the lab, budding yeasts have proved to be a popular group of organisms for experiments ranging from experimental evolution to genetic engineering. We selected these 11 species based on their sparse taxonomic distribution, their amenability to growth in a custom rich media, the availability of genome assemblies, and their inclusion in previous studies of de novo genes in yeast. We have used novel transcripts assembled from our RNA-Seq data, taken together with the reference annotations, to generate a more complete transcriptome for each of the eleven species surveyed. We have estimated the time that each *S. cerevisiae* transcript originated in the yeast phylogeny using homology searches and genomic synteny [1]. As organisms modify their expression and translation of genes in response to stress, we sequenced the transcriptomes of all 11 species of yeast in both rich media and oxidative stress conditions to capture potential transcriptome variability.

The availability of complete gene annotations is key for genome-wide studies. The transcript assemblies provided contain hundreds of transcripts that were not present in the available annotations, and thus provide a more

*Correspondence: will.blevins@upf.edu
[1] Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Research Institute and Universitat Pompeu Fabra, Barcelona, Spain
Full list of author information is available at the end of the article

Table 1 Overview of data files

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
S. cerevisiae rich medium RNA-seq	SRR8690267_1.fastq; SRR8690267_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690267
S. paradoxus rich medium RNA-seq	SRR8690268_1.fastq; SRR8690268_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690268
S. mikatae rich medium RNA-seq	SRR8690269_1.fastq; SRR8690269_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690269
S. kudriavzevii rich medium RNA-seq	SRR8690270_1.fastq; SRR8690270_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690270
S. bayanus rich medium RNA-seq	SRR8690271_1.fastq; SRR8690271_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690271
N. castelii rich medium RNA-seq	SRR8690272_1.fastq; SRR8690272_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690272
K. lactis rich medium RNA-seq	SRR8690273_1.fastq; SRR8690273_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690273
L. waltii rich medium RNA-seq	SRR8690274_1.fastq; SRR8690274_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690274
L. waltii replicate 2 rich medium RNA-seq	SRR8690275_1.fastq; SRR8690275_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690275
L. thermotolerans rich medium RNA-seq	SRR8690276_1.fastq; SRR8690276_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690276
L. kluyveri rich medium RNA-seq	SRR8690261_1.fastq; SRR8690261_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690261
Schizo. pombe rich medium RNA-seq	SRR8690262_1.fastq; SRR8690262_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690262
S. cerevisiae oxidative stress RNA-seq	SRR8690267_1.fastq; SRR8690267_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690267
S. paradoxus oxidative stress RNA-seq	SRR8690268_1.fastq; SRR8690268_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690268
S. mikatae oxidative stress RNA-seq	SRR8690257_1.fastq; SRR8690257_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690257
S. kudriavzevii oxidative stress RNA-seq	SRR8690258_1.fastq; SRR8690258_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690258
S. bayanus oxidative stress RNA-seq	SRR8690259_1.fastq; SRR8690259_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690259
N. castelii oxidative stress RNA-seq	SRR8690260_1.fastq; SRR8690260_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690260
K. lactis oxidative stress RNA-seq	SRR8690265_1.fastq; SRR8690265_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690265
L. waltii oxidative stress RNA-seq	SRR8690266_1.fastq; SRR8690266_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690266
L. waltii replicate 2 oxidative stress RNA-seq	SRR8690278_1.fastq; SRR8690278_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690278
L. thermotolerans oxidative stress RNA-seq	SRR8690277_1.fastq; SRR8690277_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690277
L. kluyveri oxidative stress RNA-seq	SRR8690280_1.fastq; SRR8690280_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690280
Schizo. pombe oxidative stress RNA-seq	SRR8690279_1.fastq; SRR8690279_2.fastq	.fastq	https://identifiers.org/ncbi/insdc.sra:SRR8690279
S. cerevisiae transcriptome assembly	s_cerevisiae_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
S. paradoxus transcriptome assembly	s_paradoxus_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
S. mikatae transcriptome assembly	s_mikatae_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2

Table 1 (continued)

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
S. kudriavzevii transcriptome assembly	s_kudriavzevii_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
S. bayanus rich medium RNA-seq	s_bayanus_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
N. castelii transcriptome assembly	n_castellii_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
K. lactis transcriptome assembly	k_lactis_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
L. waltii transcriptome assembly	l_waltii_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
L. waltii replicate 2 transcriptome assembly	l_waltii_rep2_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
L. thermotolerans transcriptome assembly	l_thermotolerans_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
L. kluyveri transcriptome assembly	l_kluyveri_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
Schizo. pombe transcriptome assembly	schizo_pombe_annotations_with_novel.gff	.gff	https://doi.org/10.6084/m9.figshare.7851521.v2
Sources of yeast strains and reference genomes	Yeast_strain_source.pdf	.pdf	https://doi.org/10.6084/m9.figshare.7851521.v2

The raw RNA-Seq data (.fastq files) are available for download on the SRA [7], and the transcriptome assembly annotations (.gff files), consisting of novel transcripts combined with the reference annotations for each species, are available on Figshare [8] as well as a table with the information about the source of each strain and reference genome

complete view of the gene content of each organism than previous annotations. These transcriptomes can be used as a basis to discover new encoded proteins, to study the evolution of yeast gene families and to investigate the changes in gene expression across different Saccharomycotina species. The addition of the ERCC Spike-into all samples also allows for the benchmarking of different de novo transcriptome assembly protocols.

Data description

We grew 11 species of yeast in two conditions:

1. *Rich medium* The yeast were grown in 20 mL of a custom rich medium [2], which was shown to accommodate various species of yeast, in 50 mL Erlenmeyer flasks at 30 °C. Cells were harvested in log growth phase at an OD_{600} of approximately 0.25.
2. *Oxidative stress* The same isogenic populations of yeast were grown in parallel, identical to the first condition. However, 30 min prior to harvesting the cells, hydrogen peroxide was added to a final concentration of 1.5 mM; we used a time period of 30 min to maximize the cellular response to stress [3], and a concentration of 1.5 mM H_2O_2 as we observed the yeast to grow approximately twice as slowly at this concentration.

After extraction, purification, and polyA selection of the RNA, synthetic spike-in transcripts from the ERCC RNA Spike-in kit [4] were added to each sample in order to assess the performance and limitations of our pipeline. After library preparation, the libraries were pooled into two batches (normal/stress) and sequenced in one lane on the Illumina HiSeq 2500 (paired-end, stranded, 50 bp long). This generated > 20 million high-quality strand-specific read pairs per sample (Table 1).

After taking some quality control measures with our raw RNA-Seq data, we mapped the reads to their respective genomes (Table 1) and assembled de novo transcriptomes using the program Trinity version 2.1.0 [5]. We created a non-redundant set of features from the reference annotations combined with our de novo assembled transcripts; de novo assembled transcripts which correspond to annotated features according to Cuffmerge version 2.2.0 [6] were discarded, while those that did not were considered to be novel; we identified an average of 700 novel transcripts per species [1] (Table 1). The majority of these novel transcripts were found to be expressed in both conditions, but dozens of transcripts were only expressed in one condition or the other. Using the ERCC RNA Spike-in [4], we calculated that the lower limit of detection for annotated features in our pipeline was 2 TPM, and the lower limit of expression necessary to reliably assemble novel transcripts was 15 TPM; over

half of the unannotated transcripts that we assembled were expressed above this conservative threshold of 15 TPM in at least one of the two conditions.

Limitations

A limitation of this dataset is that there are not multiple replicates for each species/condition, except for *L. waltii,* which has two replicates in each condition. We also would like to acknowledge that the concentration of hydrogen peroxide we used to induce an oxidative stress response (1.5 mM) was higher than the concentration used in other studies of oxidative stress response in yeast (0.1–1 mM).

Abbreviations

RNA-Seq: RNA sequencing; TPM: transcripts per million, a normalized measure of mRNA abundance; ERCC: External RNA Control Consortium; mM: millimolar, a measure of concentration; H_2O_2: hydrogen peroxide.

Authors' contributions

MMA and LBC designed and funded the experiments. WRB carried out the experiments and wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements

We thank Dr. Ksenia Pugach and the Verstreppen lab for cultures of several species of yeast, and the Sequencing Facilities at the Center for Regulatory Genomics (CRG) for the library preparation and sequencing. We also thank Lorena Espinar, Bernat Blasco-Moreno, and Leire de Campos-Mata for their help in preparing the samples.

Competing interests

The authors declare that they have no competing interests.

Availability of data materials

The data described in this Data Note can be freely and openly accessed on the Sequence Read Archive (SRA) with Project ID SRP187756 http://identifiers.org/ncbi/insdc.sra:SRP187756 [7] and on Figshare at https://doi.org/10.6084/m9.figshare.7851521.v2 [8] (Table 1). We have performed numerous additional analyses using this data which are available from the corresponding author on reasonable request; more information can be found in the text and supplementary material of our preprint "Frequent birth of de novo genes in the compact yeast genome" [1].

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

The work was funded by the following grants: (1) BFU2015-65235-P Ministerio de Economía e Innovación (Spanish Government)-FEDER (EU). (2) BFU2015-68351-P Ministerio de Economía e Innovación (Spanish Government)-FEDER (EU). (3) MDM-2014-0370 "Maria de Maeztu" Programme for Units of Excellence in R&D (Spanish Government). (4) 2017SGR1054 Agència de Gestió d'Ajuts Universitaris i de Recerca Generalitat de Catalunya. (5) 2017SGR01020 Agència de Gestió d'Ajuts Universitaris i de Recerca Generalitat de Catalunya. (6) Predoctoral fellowship (FI, Generalitat de Catalunya) to WRB. Grants 1, 3 and 4 were used to cover lab expenses and to obtain the RNA samples from yeast cultures. Grants 2 and 5 were used for RNA sequencing. Grant 6 covered the salary of WRB. These funding bodies had no role in the design of the study, collection of the data, analysis of the results, or writing of the manuscript.

Publisher's Note

Author details

[1] Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Research Institute and Universitat Pompeu Fabra, Barcelona, Spain. [2] Single Cell Behavior Group, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain. [3] Center for Quantitative Biology and Peking-Tsinghua Joint Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China. [4] Catalan Institution for Research and Advanced Studies, Barcelona, Spain.

Received: 19 March 2019 Accepted: 25 April 2019
Published online: 03 May 2019

References

1. Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Canas JL, Espinar L, et al. Frequent birth of de novo genes in the compact yeast genome. bioRxiv. 2019. https://doi.org/10.1101/575837.
2. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. The role of nucleosome positioning in the evolution of gene regulation. PLoS Biol. 2010. https://doi.org/10.1371/journal.pbio.1000414.
3. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell. 2000;11(12):4241–57. https://doi.org/10.1091/mbc.11.12.4241.
4. Rna E, Consortium C. Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics. 2005;6:150. https://doi.org/10.1186/1471-2164-6-150.
5. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644–52. https://doi.org/10.1038/nbt.1883.
6. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78. https://doi.org/10.1038/nprot.2012.016.
7. Transcriptomic comparison of 11 species of yeast in rich media and oxidative stress conditions. Sequence Read Archive. 2019. http://identifiers.org/ncbi/insdc.sra:SRP187756. Accessed 16 Mar 2019.
8. Blevins, W. Combination of novel transcripts from de novo assembly and genes from reference annotations for 11 species. figshare. 2019. https://doi.org/10.6084/m9.figshare.7851521.v2.