Assembly of Repeat Content Using Next Generation Sequencing Data

Kurt LaButti[1*], Alan Kuo[1], Igor Grigoriev[1], Alex Copeland[1]

[1] LBNL - Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA

**To whom correspondence should be addressed*: Email: klabutti@lbl.gov

March 21, 2014

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Assembly of Repeat Content Using Next Generation Sequencing Data

K. LaButti, A. Kuo, I. Grigoriev, A. Copeland

Department of Energy Joint Genome Institute, 2800 Mitchel Drive, Walnut Creek, California 94598, USA

ABSTRACT

Repetitive organisms pose a challenge for short read assembly, and typically only unique regions and repeat regions shorter than the read length, can be accurately assembled. Recently, we have been investigating the use of Pacific Biosciences reads for *de novo* fungal assembly. We will present an assessment of the quality and degree of repeat reconstruction possible in a fungal genome using long read technology. We will also compare differences in assembly of repeat content using short read and long read technology.

INTRODUCTION

Repeats are a ubiquitous feature of the genomes of many organisms. They vary in abundance, size, and identity in each genome, and they are a major source of mis-assemblies in *de novo* genome assembly. The nuclear ribosomal DNA (rDNA) gene complex in eukaryotes is an important example of tandem repeated sequence. The genes encoding the ribosomal RNAs are organized into arrays, containing repetitive transcriptional units (Figure 1) on one or a few chromosomes [1]. The length and copy number of the repeat unit varies among eukaryotes, with fungi normally measuring between 7-12 kb and containing over 100 copies per genome [2]. rDNA repeat units are highly uniform with extremely low variation within a species, but differ more between species, suggesting that concerted evolution maintains the identity of the rDNA repeat arrays [2,3,4].

To assess whether long reads can resolve repetitive sequence, we searched for the rDNA in the assemblies of 21 fungal species, from 7 genera. The selected genomes range greatly in size (28-68 Mbp) and repeat content, (3-47%) (Table 1). All were sequenced and assembled by the JGI using either Sanger, 454, Illumina, or PacBio data, with the exception of *Aspergillus fumigatus* (Sanger, TIGR), *Aspergillus nidulans* (Sanger, Broad), and *Aspergillus oryzae* (Sanger, AIST); and *Laccaria bicolor* S238N-H70, sequenced by the JGI using both Illumina and PacBio reads.

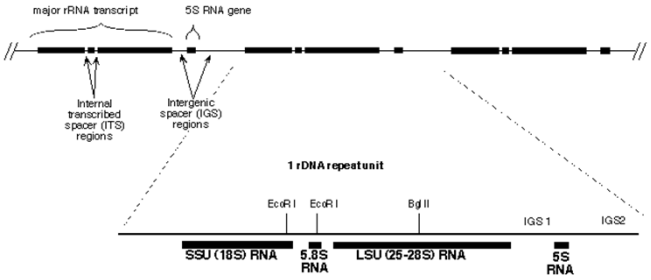

Figure 1. Physical map of the fungal rDNA operon [13]

MATERIALS & METHODS

Assemblies were evaluated as to whether components of the rDNA operon were accurately and completely reconstructed by comparing the assembled rDNA to previously characterized rDNA operons in the SILVA database [9] using BLAST+ [10] with the following parameters '-task megablast -perc_identity 85 -evalue 1e-3'. The presence or absence of 18S rDNA was quantified using alignments of at least 1650bp and 85% identity. A small sample of *Laccaria* S238N-H70 scaffolds containing ribosomal content were manually inspected. Copy number of the rDNA operon was estimated by alignment of a random subsample of the Illumina and PacBio read data with an in-house short read aligner, bbmap [11], and PacBio BLASR [12] respectively, to an assembled scaffold that contained a full copy of the operon. Genome-wide repeat detection was done with the JGI Annotation Pipeline.

ORGANISM	PLATFORM	LEN (MB)	REPEAT (%)	18S partial (n_scaf)	18S full length (n_scaf)
Aspergillus fumigatus	Sanger	29	3-5	2	1
Aspergillus nidulans	Sanger	30	3-5	0	0
Aspergillus niger	Sanger	35	3-5	1	1
Aspergillus oryzae	Sanger	38	3-5	1	1
Aspergillus aculeatus	454	35	3-5	7	0
Aspergillus carbonarius	454	36	3-5	4	0
Aspergillus brasiliensis	Illumina	36	3-5	1	0
Aspergillus glaucus	Illumina	28	3-5	0	0
Aspergillus versicolor	Illumina	33	3-5	0	0
Aspergillus wentii	Illumina	31	3-5	1	0
Aspergillus campestris	PacBio	28	3-5	7	6
Aspergillus novofumigatus	PacBio	32	3-5	8	8
Aspergillus ochraceoroseus	PacBio	28	3-5	8	5
Aspergillus steynii	PacBio	38	3-5	21	17
Coccodinium bartshchii	Illumina	28	7	0	0
Guyanagaster necrorhiza	Illumina	54	18	0	0
Laccaria bicolor S238N-H70	**Illumina**	**57**	**47**	**1**	**0**
Laccaria bicolor S238N-H70	**Ill+PacBio**	**57**	**47**	**1**	**0**
Laccaria bicolor S238N-H70	**PacBio**	**68**	**47**	**98**	**68**
Morchella conica	Illumina	48	25	1	0
Morchella conica	Ill-norm	48	25	2	1
Phoma tracheiphila	Illumina	34	15	1	0
Phoma tracheiphila	Ill-norm	34	15	1	1
Thielavia antarctica	Illumina	40	45	0	0
Thielavia hyrcaniae	Illumina	31	14	1	0
Thielavia hyrcaniae	Ill-norm	31	14	2	0

Table 1. Selected fungal genome assemblies and their size and estimated repeat content, sorted by name and data type (color). The last two columns represent the number of scaffolds that contain partial (>=100bp,>=85% identity) and full length (>=1670bp,>=85% identity) hits to small subunit 18S.

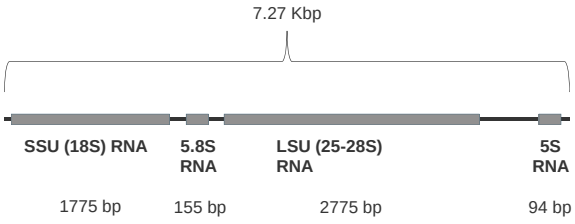

Figure 3. Average lengths of ribosomal DNA components and overall operon length for *Laccaria bicolor* S238N-H70.

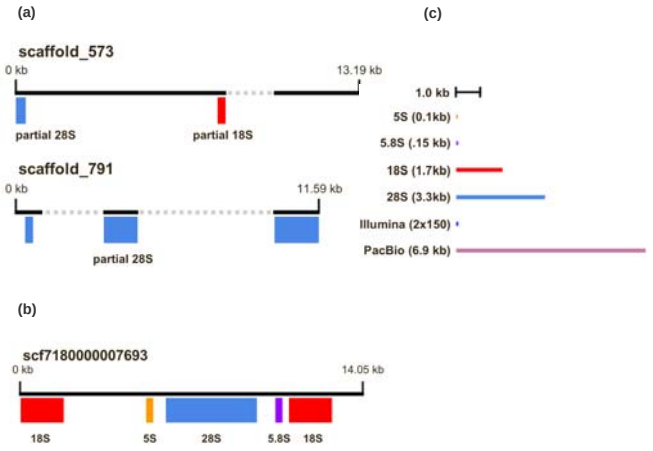

Figure 2. Physical map of *Laccaria* Illumina **(a)** and PacBio **(b)** assembly scaffolds with ribosomal content. Contiguous sequence is represented by a horizontal black line while dashed gray lines depict gaps. Ribosomal components are represented by colored boxes below the scaffold. **(c)** Legend depicting approximate relative sizes of ribosomal units and average read sizes for Illumina and PacBio.

RESULTS

- Long read assemblies clearly reconstruct rDNA operons better than short read assemblies
- Short read assemblies (Illumina) contain partial or no reconstructed 18S SSU (Table 1)
- Long read assemblies (Sanger, 454) contain at least some fully reconstructed 18S
- Longest read assemblies (PacBio) contain most full length copies of 18S
- Illumina *Laccaria* assembly contains 2 scaffolds with partial rDNA components (Figure 2a)
- PacBio *Laccaria* assembly contains 68 scaffolds with at least one full length 18S (Figure 2b)
  - Assembly contains full copy of rDNA major transcript, 7.27 Kbp in length (Figure 3)
  - Assembly is 10 Mbp larger than Illumina equivalent
  - Annotation suggests 10 Mbp more repetitive sequence
- Read alignment suggests copy number of 30-120 rDNA operons in *Laccaria*

CONCLUSIONS

Short read assemblies, though economical, may lack accurately reconstructed repeats including rDNA. Assemblies using long read data appear to more accurately represent the entire genome, as repeat sequences spanned partially or entirely by long reads can often be assembled as well as the unique portions of the genome. We also determined that the size of the rDNA operon in the fungus *Laccaria bicolor* S238N-H70 is approximately 7.27 Kbp and contains between 30 and 120 copies of the rDNA operon.

REFERENCES

See handout.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.