Integrative Biology of a Fungus: Using PacBio® SMRT® Sequencing to Interrogate the Genome, Epigenome, and Transcriptome of *Neurospora crassa*

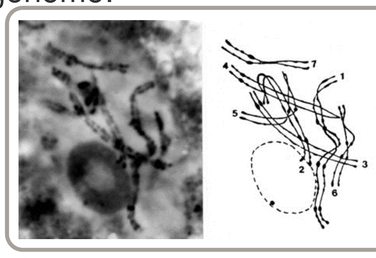
Jane Yeadon[1]*, Kristi Kim[2]*, Elizabeth Tseng, Susana Wang[2], Joan Wilson[2], David Catchside[1], Jane Landolin[2]
*Joint first authors, [1]Flinders University, Adelaide, Australia
[2]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Abstract

PacBio SMRT® Sequencing has the unique ability to directly detect base modifications in addition to the nucleotide sequence of DNA. Because eukaryotes use base modifications to regulate gene expression, the absence or presence of epigenetic events relative to the location of genes is critical to elucidate the function of the modification. Therefore an integrated approach that combines multiple omic-scale assays is necessary to study complex organisms. Here, we present an integrated analysis of three sequencing experiments: 1) DNA sequencing, 2) base-modification detection, and 3) Iso-seq™ analysis, in *Neurospora crassa*, a filamentous fungus that has been used to make many landmark discoveries in biochemistry and genetics. We show that *de novo* assembly of a new strain yields complete assemblies of entire chromosomes, and additionally contains entire centromeric sequences. Base-modification analyses reveal candidate sites of increased interpulse duration (IPD) ratio, that may signify regions of 5mC, 5hmC, or 6mA base modifications. Iso-seq method provides full-length transcript evidence for comprehensive gene annotation, as well as context to the base-modifications in the newly assembled genome. Projects that integrate multiple genome-wide assays could become common practice for identifying genomic elements and understanding their function in new strains and organisms.

Genome: *de novo* Assembly

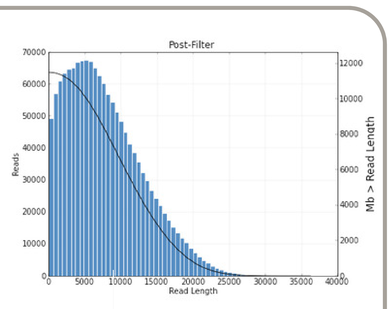
The *Neurospora crassa* genome:
- Seven chromosomes
- Approximately 40 Mb



We sequenced a new strain (T1) of *N. crassa,* which is an A mating type strain derived by DG Catcheside from a cross between the Em a 5297 and Em A 5256 strains he obtained from Stirling Emerson in 1955.

Sequencing statistics

Bases: 9 Gb (225 X)
Bases per cell: 500 Mb
Mean read length: 7.8 kb
N50 read length: 10.4 kb
Chemistry: C2

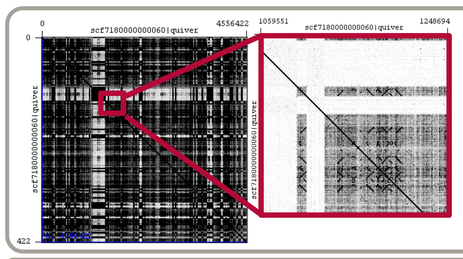


Assembly statistics

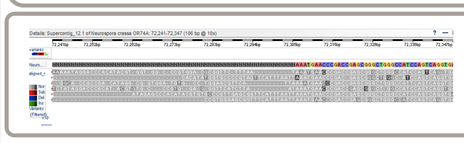
Pre-assembly seed cutoff length: 15.9 kb
Total assembled bases: 41.7 Mb
Total contigs: 20
N50 contig length: 5.3 Mb *(limited by chrom. size)*
Algorithm: HGAP.1 *(Chin et al.)*
Software: SMRT Analysis v.2.0.1 *(free, open-source)*

We assembled four entire chromosomes into a single contig and the remaining three chromosomes were assembled with only 1-2 gaps remaining. By comparison, each "supercontig" annotated in the reference genome has between 38 – 89 gaps.

Chromosome (reference scaffold)	Length	# gaps	Assembled Contigs	# gaps	Extra Sequence
Supercontig 12.1	9.7 Mb	89	Contig_54 (6.6 Mb) Contig_63 (3.4 Mb)	1	300 kb
Supercontig 12.2	**4.5 Mb**	**56**	**Contig 60 (4.5 Mb)**	**0**	**78 kb**
Supercontig 12.3	**5.3 Mb**	**45**	**Contig 59 (5.3 Mb)**	**0**	**24 kb**
Supercontig 12.4	6.0 Mb	47	Contig 57 (6.2 Mb) Contig 58 (12 kb)	1	181 kb
Supercontig 12.5	**6.4 Mb**	**42**	**Contig 56 (6.4 Mb)**	**0**	**(21 kb)**
Supercontig 12.6	**4.2 Mb**	**38**	**Contig 62 (4.3 Mb)**	**0**	**101 kb**
Supercontig 12.7	4.3 Mb	43	Contig 69 (2.6 Mb) Contig 70 (1.7 Mb) Contig 61 (20 kb)	2	102 kb



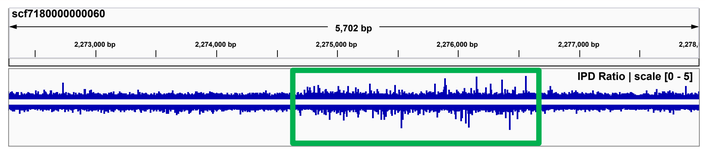
A self-self dot-plot of a 4.5 Mb contig (contig 60). The entire 276 kb centromere composed of low complexity sequence was successfully assembled *de novo*.



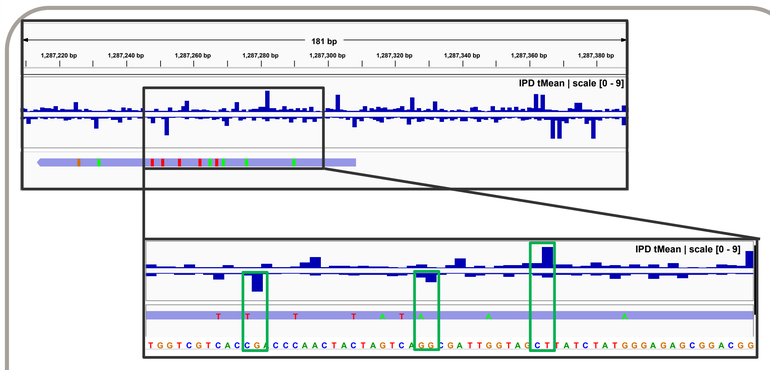
An example of a gap that has been spanned with PacBio sequence.

Epigenome: Base Modifications

PacBio SMRT sequencing has the unique ability to detect base modifications based on increased interpulse duration (IPD) between two base-incorporation events.



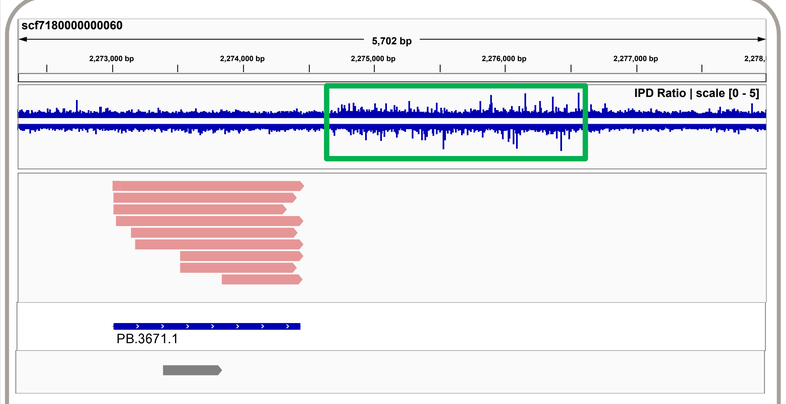
We can detect large regions of increased IPD. Above is an example of a 2 kb region on contig 60, where IPD is increased in both the forward and reverse directions.



We can also detect strand-specific increased IPD at base-level resolution. The figure above contains examples of punctate signals of increased IPD in the zeta-eta 5S rRNA region known to be methylated in association with a genome defense mechanism in fungi called repeat induced point mutation (RIP). Repetitive DNA (mostly CpA's) become methylated, which induces mutations that ultimately limit accumulation of transposable elements.

Integration

A well assembled genome is an important backbone for annotating functional elements such as genes and epigenetic marks. The rich genomic context obtained by combining datasets and assays is greater than the sum of the individual parts.



The 2 kb region with increased IPD identified earlier can now be placed relative to a transcript (PB.3671.1) whose length was underestimated by the reference annotation. It is apparent now that the 2 kb region is immediately downstream of the transcript.

Transcriptome: Iso-seq™ Analysis

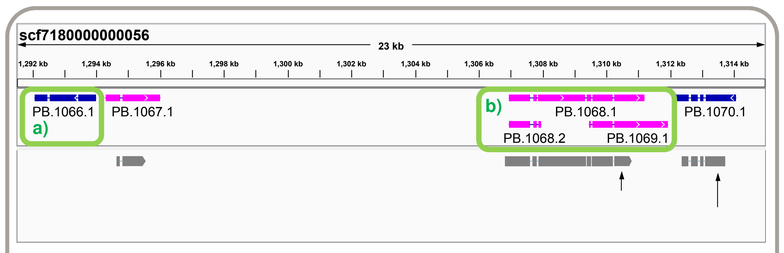
We isolated and sequenced polyA+ RNA from the same T1 strain. A cDNA library was prepared and separated into three size fractions (1-2 kb, 2-3 kb, 3-6 kb) according to the Iso-Seq protocol "Using Clontech cDNA Prep and BluePippin™ Size Selection". Transcripts were obtained by applying the bioinformatics pipeline for consensus clustering (https://github.com/PacificBiosciences/cDNA_primer).

Consensus clustering statistics

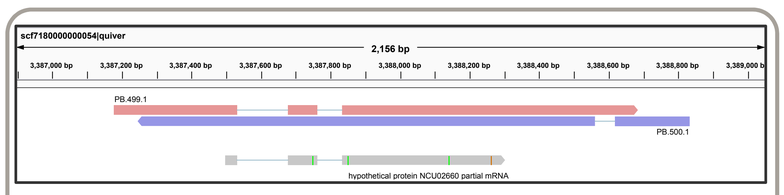
- Identified 31,890 full-length cDNAs
- Aligned 68,244,024 bases
  - 99.94% accuracy (0.057% error)
  - 11,932 (0.017%) insertions
  - 15,502 (0.022%) deletions
  - 11,482 (0.016%) mismatches

Transcript analysis statistics

- 7,626 unique transcripts
  - Found 4823 named/known transcripts
  - Validated 2829 hypothetical transcripts that had no prior mRNA evidence
  - Discovered 144 new transcripts



a) A new gene (PB.1066) is detected. It is composed of one isoform (PB.1066.1) with three exons transcribed on the negative strand and bidirectional to PB.1067.1. **b)** Two additional isoforms (PB.1.1068.2, and .3) are detected for gene NCU04664T0.



We also found many examples of antisense transcription. Here, a new transcript (PB.500.1) is antisense to NCJ02660, a hypothetical protein annotated as a "partial mRNA". The full transcript (PB.499.1) is now validated with Iso-Seq analysis.

References

- Chin CS, et al., Nonhybrid finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods, 2013 Jun 10; 10(6):563-9
- SMRT Analysis v 2.1.1 Software suite is free and open-source, and can be downloaded at http://pacbiodevnet.com.
- Iso-Seq™ sample preparation protocol can be downloaded at http://www.smrtcommunity.com/Share/Protocol?id=a1q70000000HqSvAAK&strRecordTypeName=Protocol

Pacific Biosciences, PacBio, SMRT, SMRTbell and the Pacific Biosciences logo are trademarks of Pacific Biosciences of California, Inc. All other trademarks are the property of their respective owners. © 2014 Pacific Biosciences of California, Inc. All rights reserved.