Single Molecule, Real-Time Sequencing for Base Modification Detection in Eukaryotic Organisms: *Coprinopsis cinerea*

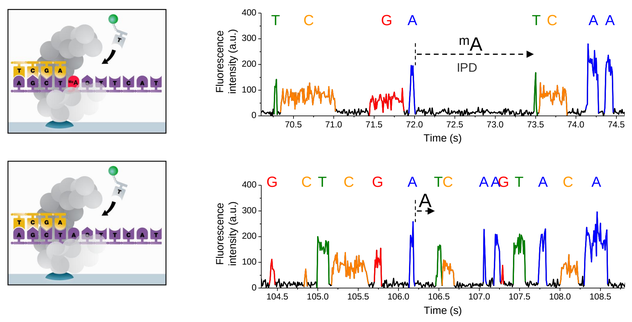
Khai Luong[1], Tyson A. Clark[1], Matthew Boitano[1], Yi Song[1], Stephen W. Turner[1], Jonas Korlach[1]
Lukas Chavez[2], Patricia J Pukkila[3], Yun Huang[2], Virginia K. Hench[3], Willaim Pastor[2], Lakshminarayan M. Iyer[5], Suneet Agarwal[4], L. Aravind Iyer[5], Anjana Rao[2]
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025
[2] La Jolla Insitute for Allergy & Immunology, 9420 Athena Circle, La Jolla, CA 92037
[3] University of North Carolina at Chapel Hill, Department of Biology, CB#3280, Chapel Hill, NC 27599-3280
[4] Harvard Stem Cell Institute, Holyoke Center, Suite 727W, 1350 Massaachusetts Ave., Cambridge, MA 02138
[5] NCBI-CBB, 8600 Rockville Pike, MSC 6075, Bethesda, MD 20894-6075

Introduction

Single Molecule Real-Time (SMRT®) DNA sequencing provides a wealth of kinetic information beyond the extraction of the primary DNA sequence, and this kinetic information can provide for the direct detection of modified bases present in genomic DNA. This method has been demonstrated for base modification detection in prokaryotes at base and strand resolutions. In eukaryotes, the common base modifications known to exist are the cytosine variants including methyl, hydroxymethyl, formyl and carboxyl forms. Each of these modifications exhibits different signatures in SMRT kinetic data, allowing for unprecedented possibilities to differentiate between them in direct sequencing data. We present early results of directly sequencing different base modifications in eukaryotic genomic DNA using this method.

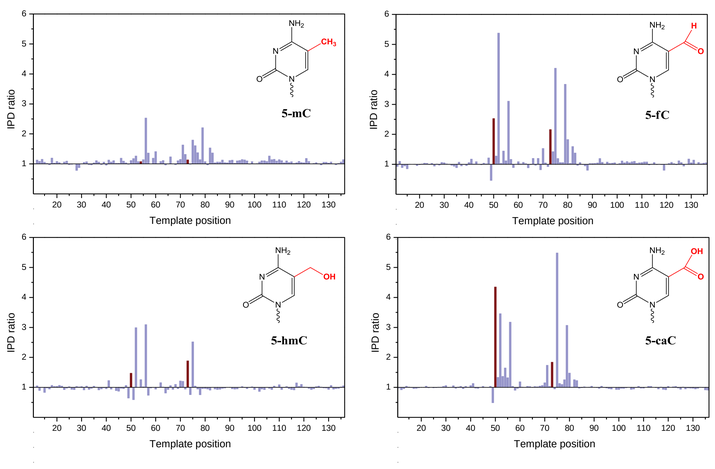
Methods

Base Modification Detection by SMRT® Sequencing



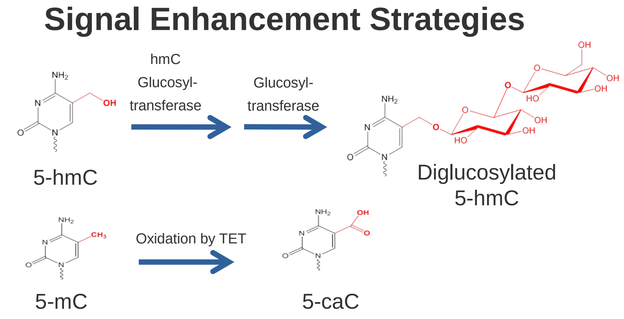
- Inter-pulse duration (IPD) is the time between the previous and current base incorporations
- IPD for a base complement to a modified base is, on average, longer than to a canonical base
- At every position, compare the observed IPDs to the expected IPD distribution

SMRT® Sequencing of the Four Forms of Cytosine



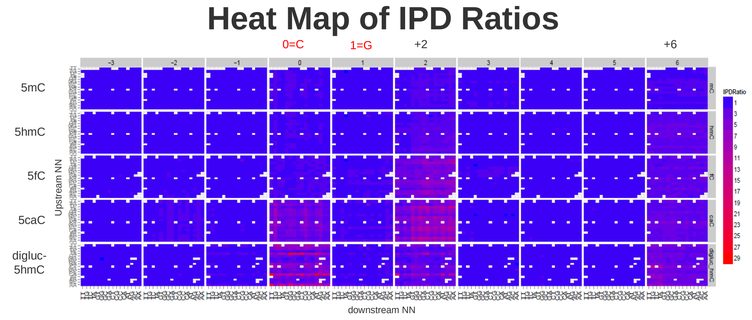
- IPD ratio kinetograms show reproducible footprint signatures for different modifications
- Salient Feature #1: Variants of C have three strongest peaks at position 0, +2, and +6 in the 5’ direction

Signal Enhancement Strategies



- Chemical or enzymatic treatment of the DNA sample can be used to increase signal intensity

Heat Map of IPD Ratios



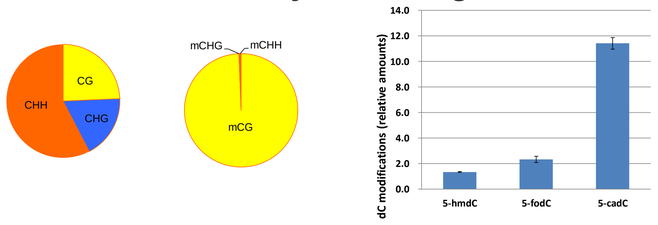
- Using template sequence context NN*CGNN, the heat map shows sequence-dependent variability in the kinetic signature
- Salient Feature #2: IPD ratio intensity increases from 5-mC to higher oxidative states the of methyl group in nearly all sequence contexts

Model Organism

- *Coprinus cinereus* (*C. cinerea okayama* 7#130)**:** a multicellular basidiomycete fungus with a typical mushroom form that undergoes a complete sexual cycle.

- Small genome ~36 Mb:
  13 chromosomes

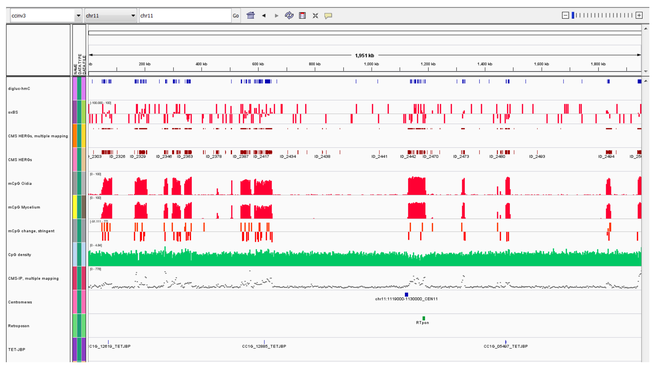
Variants of cytosine in genome



Results

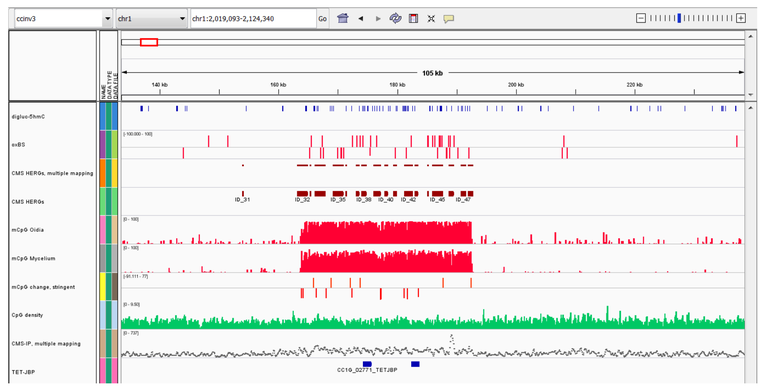
- Used SMRT sequencing to sequence complete *C. cinerea* genome to ~100x coverage using both 800 bp and 8 kb libraries
  - Native genome
  - Diglucosylation (enhance 5-hmC)

Correlation with other Methods



- Global comparison shows that SMRT detection of diglucosylation agrees with other methods, in particular in CpG-rich regions.

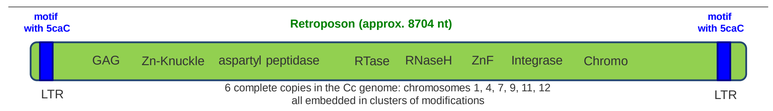
Correlation with Gene Annotations



- Modification rich regions overlap with TET/JBP transposons
- Discovered unique motif of ~200 bp that is methylated across several chromosomes and overlaps with novel retroposons

5’-…CACAGGTTACTGCGGAG*C*GCAGCAGAGATAAATTAGAGAA…-3

- Comparison of base modification analysis of native vs. diglucosylated samples reveals prominence of caC in LTR of novel retroposons (6 complete copies in chr. 1,4,7,9,11, and 12)



- Pattern of oxidized 5-mC enrichment is predictive for centromeres

Conclusions

- SMRT Sequencing can detect the four known variants of C found in eukaryote genomes, 5-mC, 5-hmC, 5-fC, and 5-caC
- Enhancement strategy combined with native genome sequencing can increase differentiation between some variants of C
- There are 40 copies of TET/JBP transposons in *C. cinerea*, contained within large regions of oxidized 5-mC in CpG regions
- Centromeres are also strongly marked with oxidized methylcytosines
- A novel set of retroposons bears caC in a specific motif in its long terminal repeat