The Future of High-Quality Genomes, Transcriptomes & Epigenomes

Jonas Korlach

December 3, 2015
HIGH-QUALITY GENOMES, TRANSCRIPTOMES & EPIGENOMES

Genome

Transcriptome

Epigenome
HIGH-QUALITY GENOMES, TRANSCRIPTOMES & EPIGENOMES

Genome

Transcriptome

Epigenome
## HUMAN GENOME DE NOVO ASSEMBLIES

<table>
<thead>
<tr>
<th>Year</th>
<th>Technology</th>
<th>Assembler</th>
<th>Sample</th>
<th>Contig N50 (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>ABI 3730</td>
<td>Celera</td>
<td>HuRef</td>
<td>0.11</td>
</tr>
<tr>
<td>2009</td>
<td>Illumina GA</td>
<td>SOAP de novo</td>
<td>BGI YH</td>
<td>0.007</td>
</tr>
<tr>
<td>2010</td>
<td>454 GS Flx Titanium</td>
<td>Newbler</td>
<td>KB1</td>
<td>0.006</td>
</tr>
<tr>
<td>2010</td>
<td>Illumina GA</td>
<td>ALLPATHS-LG</td>
<td>NA12878</td>
<td>0.024</td>
</tr>
<tr>
<td>2013</td>
<td>454 GS, HiSeq, MiSeq</td>
<td>Newbler</td>
<td>RP11_0.7</td>
<td>0.13</td>
</tr>
<tr>
<td>2014</td>
<td>HiSeq, BAC clones</td>
<td>Reference-guided</td>
<td>CHM1</td>
<td>0.14</td>
</tr>
<tr>
<td>2014</td>
<td>PacBio RS II</td>
<td>FALCON</td>
<td>CHM1</td>
<td>4.38</td>
</tr>
<tr>
<td>2015</td>
<td>PacBio RS II</td>
<td>FALCON</td>
<td>CHM13</td>
<td>26.9 Mb</td>
</tr>
<tr>
<td>2015</td>
<td>PacBio RS II</td>
<td>FALCON</td>
<td>AK1</td>
<td>7.28</td>
</tr>
<tr>
<td>2015</td>
<td>PacBio RS II</td>
<td>FALCON</td>
<td>HuRef</td>
<td>10.38</td>
</tr>
<tr>
<td>2015</td>
<td>PacBio RS II</td>
<td>FALCON</td>
<td>PC-9*</td>
<td>3.58</td>
</tr>
<tr>
<td>2015</td>
<td>PacBio RS II</td>
<td>FALCON</td>
<td>SK-BR-3*</td>
<td>2.56</td>
</tr>
</tbody>
</table>

* cancer cell lines

**Data sources:**
- HuRef (Venter) (http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.0050254)
- BGI YH (http://genome.cshlp.org/content/20/2/265.abstract Table II)
- KB1 (http://www.nature.com/nature/journal/v463/n7283/full/nature08795.html)
- NA12878 (http://www.pnas.org/content/early/2010/12/20/1017351108.abstract Table3)
- CHM1 Illumina (http://www.ncbi.nlm.nih.gov/assembly/GCF_000306695.2)
DIPLOID ASSEMBLY GRAPH IN 2013

An unitig graph from Ler-0 + Col-0 data

The graph “diameter” ~ 12 M bp
Mean edge size=17.4 k bp

“Bubbles” caused by SV between homologous copies
Branching point caused by repeats
THE FALCON UNZIP PROCESS

Falcon Unzip

Agumented with haplotype information of each reads

Updated primary contig + “associate haplotigs”

Falcon

Associate contig 1 (Alternative allele)

Primary contig

Associate contig 2 (Alternative allele)
SOLVING THE DIPLOID ASSEMBLY PROBLEM

- **Falcon** (a polyploid-aware assembler): generating the contigs through the bubbles
- **Falcon Unzip**: identifying smaller variants and using them to separate the haplotypes
SOLVING THE DIPLOID ASSEMBLY PROBLEM

- Bubbles = big variants between the haplotypes

- Collapsed Path = smaller variants between the haplotypes

- Falcon (a polyploid-aware assembler): generating the contigs through the bubbles

- Falcon Unzip: identifying smaller variants and using them to separate the haplotypes
PUT EVERYTHING TOGETHER

4 major haplotype phased blocks

Un-phased region

~ 4.80 Mbp

Add missing haplotype specific nodes & edges

Remove edges that connect different haplotypes

The final graph comprises a primary contig (blue), a major haplotig (red) and other smaller haplotigs.
USING IN SILICO F1 FOR EVALUATING PHASING ACCURACY

- Two inbred lines, CVI-0 and Col-0, were sequenced separately about 1.5 years ago with P5C3 chemistry
- Characterize the variations between the two strains with the per-strain haploid assemblies:
  - High SV density: big SV every 80 kb
  - High SNP density: SNP every 100 to 300 bp
- In silico diploid dataset: mixture of the two datasets to emulate a diploid genome at about 80x coverage.

9.49 Mb haplotype fused assembly graph
ARABIDOPSIS THALIANA ASSEMBLY COMPARISON

Falcon Unzip: 85% genome resolved to haplotigs with haplotig N50=973 kb

Expected 1n genome size = 135 Mb
LARGE INSERT LIBRARY PREPARATION PROTOCOLS

Unsupported Protocol

Please note: the shared protocols described herein may not have been validated by Pacific Biosciences and are provided as-is and without any warranty. Use of these protocols is offered to those customers who understand and accept the associated terms and conditions and wish to take advantage of their potential to help prepare samples for analysis using the PacBio® System. If any of these protocols are to be used in a production environment, it is the responsibility of the end user to perform the required validation.

Preparation > 30 kb SMRTbell™ Libraries Using the Megaruptor® Shearing and BluePippin™ Size-Selection System

Before You Begin

This document provides recommendations for preparing >30 kb size-selected SMRTbell libraries from 5 μg of starting sheared genomic DNA (gDNA).

Only high-quality, high molecular weight gDNA may be used for producing >30 kb libraries. To ensure success, gDNA size and integrity must be verified by pulsed field gel electrophoresis before beginning library preparation. Additionally, conditions for shearing gDNA to a size that can support producing >30 kb libraries must also be determined and verified empirically for each sample.

Overall yields of >30 kb libraries are typically 5-10%. For large genome projects, we recommend starting this procedure with >20 μg of high quality gDNA sample.

LARGE INSERT LIBRARY PREPARATION PROTOCOLS

Preparation of >30 kb SMRTbell™ Libraries Using Needle Shearing and BluePippin™ Size-Selection System

Before You Begin

This document provides recommendations for preparing >30 kb size-selected SMRTbell libraries from 5 µg of starting sheared genomic DNA (gDNA).

Only high-quality, high molecular weight gDNA may be used for producing >30 kb libraries. To ensure success, gDNA size and integrity must be verified by pulsed field gel electrophoresis before beginning library preparation. Additionally, conditions for shearing gDNA to a size that can support producing >30 kb libraries must also be determined and verified empirically for each sample.

Overall yields of >30 kb libraries are typically 5-10%. For large genome projects, we recommend starting this procedure with >20 µg of high quality gDNA sample.

Unsupported Protocol

Please note: the shared protocols described herein may not have been validated by Pacific Biosciences and are provided as-is and without any warranty. Use of these protocols is offered to those customers who understand and accept the associated terms and conditions and wish to take advantage of their potential to help prepare samples for analysis using the PacBio® System. If any of these protocols are to be used in a production environment, it is the responsibility of the end user to perform the required validation.

TOOLS FOR HIGH MOLECULAR WEIGHT DNA

Guidelines for Using the Sage Science™ Pippin Pulse

Overview
Pulsed-field gel electrophoresis is a strategy for resolving large DNA fragments (a typical direct-current agarose gel, fragments are separated based on size). Pulsed-field gels work by shifting direction of the electric field in a gel. DNA will change direction faster than larger molecules, making them switch, or 'pulsing', the electric field.

A single-pulse duration typically separates a relative known to be necessary to use a range of pulse times; this is a forward and/or reverse intervals from a lower limit.

The Pippin Pulse is programmed by entering a forward followed by adding a time increment to the programs. Adding additional time increments to either or both of the steps will continuously cycle for the duration different lengths of time reproducibly.


Guidelines for Using the BIO-RAD® CHEF Mapper® XA Pulsed Field Electrophoresis

Overview
Conventional electrophoresis uses single electrical fields to cause DNA fragments to migrate through a medium. Effective separation of DNA fragments up to ~20 kb. However, when imaged, larger fragments co-migrate and appear as large bands at the top of the gel. Pulsed field gel electrophoresis (PFGE) overcomes this problem by alternating the electrical field between spatially distinct pairs of electrodes. This technique results in the separation of DNA fragments up to 10 Mb by their reorientation and movement at different speeds through the pores of an agarose gel.

TARGETED SEQUENCING WITH PACBIO

- **Amplicon Sequencing**
  - 250 bp to >10 kb
  - Resolve complex genomic regions
  - HLA
  - Combination with droplet-based multiplex PCR (Raindance)

- **Target Capture Techniques**
  - 6kb fragments with Roche NimbleGen SeqCap EZ enrichment
TARGETED SEQUENCING

Please note: the shared protocols described herein may not have been validated by Pacific Biosciences and are provided as-is and without any warranty. Use of these protocols is offered to those customers who understand and accept the associated terms and conditions and wish to take advantage of their potential to help prepare samples for analysis using the PacBio RS II system. If any of these protocols are to be used in a production environment, it is the responsibility of the end user to perform the required validation.

Amplicon Template Preparation and Sequencing

Before You Begin

This procedure requires the following PacBio® products:

- SMRTbell™ Template Prep Kit
- DNA/Polymerase Binding Kit
- MagBead Kit for amplicons ≥ 1 kb
- DNA Sequencing Reagent
- DNA Internal Control Complex
- SMRT® Cells
- AMPure® PB beads

The PacBio System can be used to generate highly accurate sequences from amplicons ranging in size from several hundred bases to 10 kb or larger. Unlike sheared genomic DNA (gDNA), which is comprised of DNA fragments spanning a range of lengths, PCR products from one reaction are typically the same or similar lengths. This document describes methods for preparing PCR-amplified DNA for sequencing on the PacBio System.

Coming soon:

- HLA Getting Started Guide (12/11)
- Updated NimbleGen Target Capture / Barcoding (12/18)


NIMBLEGEN is a trademarks of Roche
FULL-LENGTH 16S SEQUENCING

Full-Length 16S Amplification, Library Preparation and Sequencing

With mock community samples, amplification with this protocol generates a discrete product with a yield adequate for library preparation. Sequencing results indicate good representation of community members for these samples, with low rates of chimeras.

Before You Begin
This procedure requires a KAPA HiFi HotStart PCR Kit from Kapa Biosystems and the following PCR primers:

27F: AGGTTYGATYMTGGCTCAG
1492R: RGYTACCTTACGACTT

Use only high-quality primers; damaged bases at the ends of the amplicons cannot be repaired by DNA Damage Repair enzymes. HPLC purification of PCR primers is recommended.

Additionally, this procedure requires the following PacBio® products:

- SMRTbell™ Template Prep Kit
- AMPure® PB beads
- DNA/Polymerase Binding Kit
- MagBead Kit and MagBeads for amplicons ≥1 kb
- DNA Sequencing Reagent
- DNA Internal Control Complex
- SMRT® Cells

Targeted release date 12/15-1/16
**UPDATED LOW INPUT PROTOCOLS**

**Procedures & Checklists**
- **Very Low (10 ng) Input 2 kb Template Preparation and Sequencing with Carrier DNA**

**Before You Begin**
To perform this procedure, you must have the PacBio®:
- Template Prep Kit
- AMPure® PB Beads
- DNA/Polymerase Binding Kit (P6 v2 or later)
- MagBead Kit
- DNA Sequencing Kit
- SMRT® Cells

Additional required materials:
- Carrier plasmid DNA, 10 μg (pBR322 or pUC, NEB or Thermo Scientific)
- Exo III, Exo VII, and Template prep buffer from the DNA Template Prep Kit (below).

This procedure can be used to prepare 1 kb to 3 kb libraries from 5 ng ofCarrier DNA, or from 1 kb to 3 kb amplicons. Note that when preparing libraries, 1 μg DNA, or from 1 kb to 3 kb amplicons. Note that when preparing libraries, you must use MagBead loading for sequencing.

**Shared Protocol 10 kb to 20 kb Template Preparation and Sequencing (with Low-Input DNA)**

**Before You Begin**
To perform this procedure, you must have the PacBio®:
- Template Prep Kit
- AMPure® P6 Beads
- DNA/Polymerase Binding Kit (P6 v2 or later)
- MagBead Kit
- DNA Sequencing Kit
- SMRT® Cells

This procedure can be used to prepare 10-20 kb libraries from 100 ng to 200 ng of sheared and concentrated DNA, minimally 100 ng into shearing. Note: for input amounts between 200 ng and 1 μg, the standard 10 kb library prep protocol may be used.

<table>
<thead>
<tr>
<th>Insert Size Target</th>
<th>Insert Size Range</th>
<th>Sheared and Concentrated DNA Amount</th>
<th>Ligation</th>
<th>DNA Damage Repair</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 kb to 20 kb</td>
<td>8 kb to 22 kb</td>
<td>100 to 200 ng</td>
<td>Blunt</td>
<td>Required</td>
</tr>
</tbody>
</table>

REPEAT EXPANSION DISEASES

Figure 1 | Location of expandable repeats responsible for human diseases. The sequence and location within a generic gene of expandable repeats that cause human diseases are shown, and the associated diseases are listed. BPES, blepharophimosis, ptosis and epicanthus inversus; CCD, cleidocranial dysplasia; CCHS, congenital central hypoventilation syndrome; DM, myotonic dystrophy; DRPLA, dentatorubral–pallidoluysian atrophy; EPM1, progressive myoclonic epilepsy 1; FRAXA, fragile X syndrome; FRAXE, fragile X mental retardation associated with FRAXE site; FRDA, Friedreich's ataxia; FXTAS, fragile X tremor and ataxia syndrome; HD, Huntington's disease; HDL2, Huntington's-disease-like 2; HFG, hand–foot–genital syndrome; HPE5, holoprosencephaly 5; ISSX, X-linked infantile spasms syndrome; MRGH, mental retardation with isolated growth hormone deficiency; OPMD, oculopharyngeal muscular dystrophy; SBMA, spinal and bulbar muscular atrophy; SCA, spinocerebellar ataxia; SPD, synpolydactyly.

REPEAT EXPANSION DISEASES

Epigenetic effects:

Mark A. Pook (2012). DNA Methylation and Trinucleotide Repeat Expansion Diseases, DNA Methylation - From Genomics to Technology, Dr. Tatiana Tatarinova (Ed.), ISBN: 978-953-51-0320-2,
AMPLIFICATION-FREE TARGETED ENRICHMENT

- Using Cas9 for targeting
- Sequences native DNA
- Compatible with multiple targets
- 1-3 μg of input

Example:
- Targeting HTT, FMR1, ALS & SCA10 in the same reaction:

<table>
<thead>
<tr>
<th>Reads on target</th>
<th>Enrichment factor</th>
<th>Reads on target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicate 1</td>
<td>38,600x</td>
<td>5.2%</td>
</tr>
<tr>
<td>Replicate 2</td>
<td>38,700x</td>
<td>5.4%</td>
</tr>
</tbody>
</table>
CRISPR/CAS9 SYSTEM

Some *in vivo* applications:
- Gene silencing
- Homology-directed repair
- Transient gene silencing or transcriptional repression
- Transient activation of endogenous genes
- Transgenic animals and embryonic stem cells
TARGET ENRICHMENT VIA CAS9 DIGESTION

Standard SMRTbell Template Library

Region of interest

Cas9 Digestion

Cas9 Target Site

Ligation of Capture Adapter

Polymerase Binding and Magbead Loading

Magbead Binds Only SMRTbell with Capture Adapter
CAG REPEATS IN HTT
CAG REPEAT COUNTS

Count

0 20 40 60 80 100 120 140 160 180

CAG Repeat Number

17

Control
CAG REPEAT COUNTS IN HT PATIENTS

- Widening repeat number distribution at the mutated allele is biological
- Obtained roughly equal number of sequenced molecules for normal and mutated alleles

Samples obtained from Vanessa Wheeler (Harvard Medical School)
FMR1 PRE-MUTATION SAMPLE

gDNA sample from Paul Hagerman (UC Davis)
>700 CGG REPEATS SEQUENCED FROM THE FMR1 GENE
FMR1 PRE-MUTATION SAMPLE

Normal allele (29 CGG repeats)

Mutated allele (~42 CGG repeats)

The mutant allele in this sample has a ~110 bp duplicated region preceding the CGG repeats
METHYLATION DETECTION OF FMR1 SAMPLE

- CGG repeat region appears to be heavily methylated (5mC)
AMPLIFICATION-FREE TARGETED ENRICHMENT

- Using Cas9 for targeting
- Sequences native DNA
- Compatible with multiple targets
- 1-3 μg of input

Example:
- Targeting HTT, FMR1, ALS & SCA10 in the same reaction:

<table>
<thead>
<tr>
<th></th>
<th>Enrichment factor</th>
<th>Reads on target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicate 1</td>
<td>38,600x</td>
<td>5.2%</td>
</tr>
<tr>
<td>Replicate 2</td>
<td>38,700x</td>
<td>5.4%</td>
</tr>
</tbody>
</table>
HIGH-QUALITY GENOMES, TRANSCRIPTOMES & EPIGENOMES
TARGETED ISO-SEQ USING ROCHE NIMBLEGEN SEQCAP ENRICHMENT

Full-length cDNA Target Sequence Capture Using SeqCap® EZ Libraries

Before You Begin

This document describes the process for enrichment of sample libraries using SeqCap EZ Libraries from Roche NimbleGen for subsequent sequencing on the PacBio® System.

To perform this procedure, you must have reviewed the User Bulletin- Guidelines for preparing cDNA Libraries for Isoform Sequencing (Iso-Seq™ Analysis). Below are the four available procedures for specific project requirements.

- Procedure & Checklist - Isoform Sequencing (Iso-Seq™ Analysis) using the Clontech® SMARTer® PCR cDNA Synthesis Kit and No Size Selection
- Procedure & Checklist - Isoform Sequencing (Iso-Seq™ Analysis) using the Clontech® SMARTer® PCR cDNA Synthesis Kit and Manual Agarose-gel Size Selection
- Procedure & Checklist - Isoform Sequencing (Iso-Seq™ Analysis) using the Clontech® SMARTer® PCR cDNA Synthesis Kit and the BluePippin™ Size Selection System
- Procedure & Checklist - Isoform Sequencing (Iso-Seq™ Analysis) using the Clontech® SMARTer® PCR cDNA Synthesis Kit and SageELF™ Size Selection System

NIMBLEGEN and SEQCAP are trademarks of Roche
TARGETED ISO-SEQ USING ROCHE NIMBLEGEN SEQCAP ENRICHMENT

- Effective targeting of genes of interest:

- Greater number of novel isoforms:

NIMBLEGEN and SEQCAP are trademarks of Roche
Illumina is a trademark of Illumina
TARGETED ISO-SEQ USING ROCHE NIMBLEGEN SEQCAP ENRICHMENT

Identification of novel mRNA isoforms
BARCODING SAMPLES FOR ISOFORM SEQUENCING (ISO-SEQ ANALYSIS)

- More efficient library preparation
- 16 bp barcodes
- Allows pooling of up to 4-8 samples
Develop a strategy to optimize transcript discovery

**Approach: multiplex tissue & size selection**

Tyson Clark

Six Maize Tissues

- Pollen
- v7 immature tassel
- 20 DAP embryo
- 20 DAP endosperm
- v8 immature ear
- 14 DAG root

Six Sized Libraries

- <1 kb
- 1-2 kb
- 2-3 kb
- 3-5 kb
- 4-6 kb
- 5-10 kb

Presented by Doreen Ware (USDA, CSHL) at East Coast UGM, June 2015
Comparison Between Iso-seq and Maize v3 FGS annotation

17% Same isoforms with existing V3 gene models
57% Novel isoforms from known genes
3% Novel isoforms from novel genes
4% Isoforms located in V3 intron
4% V3 isoforms located in Iso-seq intron
2% Isoforms exonic overlap with V3 loci on opposite strand
0.5% Exonic overlap with V3 but don’t share any intron chain
6% Partially match (contained) with V3 isoforms
3% Show up in WGS but not in FGS

Iso-seq has 111,151 transcripts, corresponding to 26946 genome loci (~70% gene loci)

Presented by Doreen Ware (USDA, CSHL) at East Coast UGM, June 2015
HIGH-QUALITY GENOMES, TRANSCRIPTOMES & EPIGENOMES
BACTERIAL BASE MODIFICATION DETECTION USING PACBIO

- A comparative analysis of methylome profiles of Campylobacter jejuni sheep abortion isolate and gastroenteric strains using PacBio data
- A random six-phase switch regulates pneumococcal virulence via global epigenetic changes.
- Analysis of the Campylobacter jejuni Genome by SMRT DNA Sequencing Identifies Restriction Modification Motifs.
- Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence.
- Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles.
- Characterization of DNA methyltransferase specificities using single molecule, real-time DNA sequencing.
- Comparing the genomes of Helicobacter pylori clinical strain UM032 and Mice adapted derivatives.
- Complete genome sequence of BS49 and draft genome sequence of BS34A, Bacillus subtilis strains carrying Tn916.
- Complete genome sequence of the Clostridium difficile laboratory strain 630¿erm reveals differences from strain 630, including translocation of the mobile element CTn5.
- Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic.
- Efficient and accurate whole genome assembly and methylome profiling of E. coli.
- Epigenetics: Reading methylated genomes.
- Exploring the roles of DNA methylation in the metal-reducing bacterium Shewanella oneidensis MR-1.
- Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single molecule real-time sequencing.
- Genome-wide Methylation Patterns in Salmonella enterica Subsp. enterica Serovars.
- Genomic mapping of phosphorothioates reveals partial modification of short consensus sequences.
- Going beyond five bases in DNA sequencing.
- Methylome diversification through changes in DNA methyltransferase sequence specificity.
- ModM DNA methyltransferase methylome analysis reveals a potential role for Moraxella catarrhalis phasevarions in otitis media.
- Molecular Analysis of Asymptomatic Bacteriuria Escherichia coli Strain VR50 Reveals Adaptation to the Urinary Tract by Gene Acquisition.
- REBASE—a database for DNA restriction and modification: enzymes, genes and genomes.
- Sequence data for Clostridium autoethanogenum using three generations of sequencing technologies.
- Short communication: Single molecule, real-time sequencing technology revealed species and strain-specific methylation patterns of 2 Lactobacillus strains.
- Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes.
- The complex methylome of the human gastric pathogen Helicobacter pylori.
- The functions of DNA methylation by CcrM in Caulobacter crescentus: a global approach.
- The genomic landscape of the verrucomicrobial methanotroph Methylacidiphilum fumariolicum SolV.
- The methylome of the gut microbiome: disparate Dam methylation patterns in intestinal Bacteroides dorei.
- The methylomes of six bacteria.
- Type I restriction enzymes and their relatives.

Entering the era of bacterial epigenomics with single molecule real time DNA sequencing
Brigid M Davis, Michael C Chao and Matthew K Waldor

DNA modifications, such as methylation guide numerous critical biological processes, yet epigenetic information has not routinely been collected as part of DNA sequence analyses. Recently, the development of single molecule real time (SMRT) DNA sequencing has enabled detection of modified nucleotides (e.g. 6mA, 4mC, 5mC) in parallel with acquisition of primary sequence data, based on analysis of the kinetics of DNA synthesis reactions. In bacteria, genome-wide mapping of methylated and unmethylated loci is now feasible. This technological advance sets the stage for comprehensive, mechanistic assessment of the effects of bacterial DNA methyltransferases (MTases) — which are ubiquitous, extremely diverse, and largely uncharacterized — on gene expression, chromosome structure, chromosome replication, and other fundamental biological processes. SMRT sequencing also enables detection of damaged DNA and has the potential to uncover novel DNA modifications.

neuropsychiatric disorders [1]. Derivatives of methylated cytosine, including hydroxymethylcytosine and the recently discovered formylcytosine and 5-carboxyl cytosine [2,3], may likewise regulate gene expression, although potentially by a different mechanism(s) [4]. The importance of DNA methylation is also well-established in several model proteobacteria, including Escherichia coli and Caulobacter crescentus, in which methylation of adenine residues (by Dam and CcrM, respectively) is pivotal in the control of chromosome replication, DNA repair, and gene expression (reviewed in [5,6]). However, the enzymes that mediate DNA methylation in prokaryotes are far more diverse than in eukaryotes [7], and the nature, extent and consequences of DNA modification have not been extensively investigated for most of the bacterial kingdom.
The Pacific Biosciences SMRT sequencing technology not only gives the order of As, Gs, Cs and Ts in the genome, but it also can tell when individual bases are modified. The strongest data presently is for m6A and m4C, but sometimes m5C can be detected reliably and improvements in the technology are planned, which would be successful at detecting all 3 methylated bases. By looking at the sequences flanking the methylated base, consensus sequences (motifs) can be defined that are the recognition sequence for the methylase that produced that particular modification.

REBASE always welcomes submissions of methylation data obtained by Pacific Biosciences sequencing as this can be extraordinarily helpful in determining recognition sequences for restriction systems. The methylated motifs that can be generated by SMRT Analysis of your sequence data can be entered into REBASE and the genome sequence information is scanned for the presence of RM system genes. The matching between the RM genes and the motifs is carried out by REBASE staff and the results can be kept private or made public depending on your wishes.

Send us yours!

View existing PacBio-REBASE data...
Recent issue of *Cell*:

**N6-METHYLADENINE IN EUKARYOTES**

An Adenine Code for DNA: A Second Life for N6-Methyladenine

Holger Heyn* and Manel Esteller**

*Instituto Catalán de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Catalonia, Spain
**Department of Physiological Sciences II, School of Medicine, University of Barcelona, 08026 Barcelona, Catalonia, Spain

DNA N6-methyladenine (6mA) protects against restriction enzymes in bacteria. However, isolated reports have suggested additional activities and its presence in other organisms, such as unicellular eukaryotes. New data now find that 6mA may have a gene regulatory function in green alga, worm, and fly, suggesting m6A as a potential "epigenetic" mark.
Recent issue of *Cell*:

DNA Methylation on N⁶-Adenine in *C. elegans*

Graphical Abstract

Authors

Eric Lehman, T. Geo.

Cell

N⁶-Methyldeoxyadenosine (6mA) Transcription Start Sites

Graphical Abstract

Chlamydomonas Genome

Leading Edge Minireview

Methylation levels:
- 6mA in *Drosophila*
- 6mA in *C. elegans*

Proximity to neighboring 6mA sites:
- Proximity to neighboring 6mA sites

AGAA motif density:
- AGAA motif density

GAGG motif density:
- GAGG motif density

Oxidative stress density:
- Oxidative stress density

60-100% methylation density
- 60-100% methylation density

20-80% methylation density
- 20-80% methylation density

10-20% methylation density
- 10-20% methylation density

Transposon expression, suggesting a potential role of 6mA in regulating gene expression.
SMRT BISULFITE SEQUENCING

Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS)

Yao Yang¹, Robert Sebra²,², Benjamin S Pullman¹, Wangjiong Qiao¹, Inga Peter¹, Robert J Desnick¹, C Ronald Geyer³, John F DeCoteau¹ and Stuart A Scott¹

Abstract

Background: DNA methylation has essential roles in transcriptional regulation, imprinting and other cellular processes, and aberrant CpG methylation is directly involved in the development of many diseases and cancers. To address the need for a quantitative and highly multiplexed approach with long read lengths for targeted CpG methylation analysis, we developed single molecule real-time bisulfite sequencing (SMRT-BS).

Results: Optimized bisulfite conversion and PCR conditions enabled the amplification of CpG islands and amplicons to ~1.5 kb, and targeting overlapping 625–1491 bp amplicons to SMRT-BS enabled single molecule real-time bisulfite sequencing (SMRT-BS). Across all amplicon lengths, CpG methylation quantitation was correlated with reduced variability for immediately methylated regions. SMRT-BS was validated by orthogonal bisulfite sequencing and methylation levels were assessed by simultaneously targeting four distinct CpG island amplicons in hematological malignancy cell lines to SMRT-BS (average depth of 110X), which quantitative methylation levels across all interrogated CpG sites and cell lines.

Conclusions: SMRT-BS is a novel, accurate and cost-effective targeted CpG methylation analysis approach with high multiplexing potential and minimal internal PCR artifacts. Increased sensitivity for interrogating longer amplicons (~1.0 kb) and the previously reported bisulfite sequencing Technologies of unamplified genomic DNA should be considered when measuring methylation using SMRT-BS.

Keywords: DNA methylation, CpG islands, Bisulfite sequencing, Long-read sequencing, Third generation sequencing, Single-molecule real-time (SMRT) sequencing, Pacific Bioscience
Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS)

Yao Yang¹, Robert Sebra¹,², Benjamin S Pullman³, Wangqiong Qiao¹, Inga Peter¹, Reed A Young¹, Ronald C Geyer³, John F DeCoteau³ and Stuart A Scott³

Abstract

Background: DNA methylation has essential roles in transcriptional regulation, imprinting, and other cellular processes, and aberrant CpG methylation is directly involved in the pathogenesis of some disorders and many cancers. To address the need for a quantitative and highly multiplexing method with long read lengths for targeted CpG methylation analysis, we developed single-molecule real-time bisulfite sequencing (SMRT-BS).

Results: Optimized bisulfite conversion and PCR conditions enabled the amplification of up to ~1.5 kb and the generation of overlapping 625-1491 bp amplicons to SMRT-BS input across all amplicon lengths (r = 0.972) and low standard deviations (50.10) between independent triplicates. Higher variability in CpG methylation quantitation was correlated with reduced coverage of target regions. SMRT-BS was validated by orthogonal bisulfite-RT-PCR and second generation sequencing (r = 0.931; 174 CpG sites); however, longer amplicons had reduced but very acceptable correlation with both orthogonal methods (r = 0.89) compared to amplicons less than ~1.0 kb (r = 0.940 ± 0.095) and r = 0.948. The utility of SMRT-BS was assessed by simultaneously subjecting four distinct CpG island amplifications from two human hematoepoietic malignancy cell lines to SMRT-BS (average depth of 110x), which accurately measured methylation levels across all interrogated CpG sites and cell lines.

Conclusions: SMRT-BS is a novel, accurate and cost-effective targeted CpG methylation platform with high degree of multiplexing with minimal donor PCR artifacts. Increased sequencing depth will allow amplification of longer regions (~1.0 kb) and the previously reported bisulfite sequencing unmethylated DNA should be considered when measuring intermediately methylated CpG sites. SMRT-BS is capable of interrogating ~1.5 kb amplification across all CpG islands in the human genome.

Keywords: DNA methylation, CpG islands, Bisulfite sequencing, Long-read sequence sequencing, Single-molecule real-time (SMRT) sequencing, Pacific Bioscience

---

CGGBP1 mitigates cytosine methylation at repetitive DNA sequences

Praoon Agarwal¹, Paul Cellier², Markus Hsi-Yang Fritz², Vladimir Bener², Helena Jernberg Wildlund¹, Bengt Westermark¹ and Umashankar Singh³

Abstract

Background: CGGBP1 is a repetitive DNA-binding transcription regulator with target sites at CpG-rich sequences such as CGG repeats and Alu-SINES and L1-LINES. The role of CGGBP1 as a possible mediator of CpG methylation however remains unknown. CGGBP1-mediated repetitive DNA sequences cytosine methylation is a major mechanism of transcriptional repression. Concordantly, generich regions typically carry lower levels of CpG methylation than the repetitive elements. It is well known that at interspersed repeats Alu-SINES and L1-LINES high levels of CpG methylation constitute a transcriptional silencing and retrotransposon inactivating mechanism.

Results: Here, we have studied genome-wide CpG methylation with or without CGGBP1-depletion. By high throughput sequencing of bisulfite-treated genomic DNA we have identified CGGBP1 to be a negative regulator of CpG methylation at repetitive DNA sequences. In addition, we have studied CpG methylation alterations on Alu and L1 retrotransposons in CGGBP1-depleted cells using a novel bisulfite-treatment and high throughput sequencing approach.

Conclusions: The results clearly show that CGGBP1 is a possible bidirectional regulator of CpG methylation at Alus, and acts as a repressor of methylation at L1 retrotransposons.
New Results

**Landscape of CpG methylation of individual repetitive elements**

Yuta Suzuki, Jonas K Orbach, Stephen W. Turner, Tatsuya Tsukahara, Junko Taniguchi, Hideaki Yurino, Wei Qu, Jun Yoshimura, Yuji Takahashi, Jun Mitsui, Shoji Tsuji, Hirokazu Takeda, Shinichi Morishita

doi: http://dx.doi.org/10.1101/018531

Abstract

Determining the methylation state of regions with high copy numbers is challenging for second-generation sequencing, because the read length is insufficient to map uniquely, especially when repetitive regions are long and nearly identical to each other. Single-molecule real-time (SMRT) sequencing is a promising method for observing such regions, because it is not vulnerable to GC bias, it performs long read lengths, and its kinetic information is sensitive to DNA modifications. Here, we propose a novel algorithm that combines the kinetic information for neighboring CpG sites and increases the confidence in identifying the methylation states of those sites.
COMPREHENSIVE VIEW OF A CANCER GENOME & EPIGENOME

- PC-9 lung cancer cell line

- Apply PacBio for:
  - *De novo* long-read assembly of a cancer genome
  - Characterization of gene fusions
  - Genome-wide methylome characterization

- Sequenced both drug-sensitive as well as drug-resistant sample

Collaboration with M. Classon, V. Janakiraman, E. Stawiski, S. Durinck, S. Seshagiri (Genentech) & Y. Suzuki, S. Morishita (U of Tokyo)
EPIGENOME CHARACTERIZATION

- Methylation status of CpG islands (https://github.com/hacone/AgIn)
- Chr4: FGFR3 (fibroblast growth factor 3)

Collaboration with M. Classon, V. Janakiraman, E. Stawiski, S. Durinck, S. Seshagiri (Genentech) & Y. Suzuki, S. Morishita (U of Tokyo)
ALLELE-RESOLVED EPIGENOME CHARACTERIZATION

Work by Y. Suzuki, S. Morishita (U of Tokyo)
ALLELE-RESOLVED EPIGENOME CHARACTERIZATION

Work by Y. Suzuki, S. Morishita (U of Tokyo)
HIGH-QUALITY GENOMES, TRANSCRIPTOMES & EPIGENOMES
some pins represent multiple placements
http://www.pacbio.com/publications

*through November 2015*