Resolving the 'Dark Matter' in Genomes

Jonas Kollbach
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

Genomes have many regions that are difficult to resolve with traditional sequencing techniques:
- Extreme Sequence Contexts
  - High GC
  - Low GC
  - Low complexity (di-, trinucleotide, ...)
- Structural Variation
  - Inversions, insertions, deletions
- Simple & Complex Repeats
  - Microsatellites, VNTRs, centromeres, telomeres
- Highly Polymorphic Regions
  - HLA, KIR
- Mobile Elements
  - LINE, Alu, ...
- Palindromes
- Full-length Transcripts

Single Molecule, Real-Time (SMRT™) Sequencing has excellent performance characteristics to resolve these regions:
- Long sequence reads:
  - SMRTbell
    - High-quality outputs from complete palindromic reads
    - Isolated from nominal human genome
    - Many associated datasets available for comparison
- Sequencing stats:
  - Single 31.8 kb read mapping to edge of centromere on chromosome 16:
  - PacBio® RS II sequencing time: 10 days
  - Total number of bases: 32,559,803,198
  - Many associated datasets available for validation
  - Equivalent of a haploid human genome, lack of allelic variation
  - Human cell line from complete hydatidiform mole

- Longest sequenced DNA insert: 41,460 bp
- Half of bases in reads: >10,985 bp

Trinucleotide Repeat Expansions

SMRT sequencing of previously unsequenceable, fragile X syndrome 'CGG' repeat full mutation allele:

Y Chromosome

De Novo Drosophila assembly:

Centromeres

Novel patterns of higher order repeat structures in switchgrass centromeres:

Segmental Duplications

Upgrading a chimpanzee genomic region:

Complex Repeats

Resolving a previous gap in the human genome reference, updated in GRCh38:

Summary

PacBio reads span over four orders of magnitude of genomic length scales, facilitating discovery and validation of many types of structural variation:

Genetic variants and organization of the MHC/GAC central locus, which is one of the six largest cancer genomic regions, are being resolved by MHC/GAC (HLA-DQA1, -DQB1, -DRB1, etc); the 64,000 base-pairs of the six large-insert clones produced as part of the Pacific Biosciences de novo assembly were sequenced individually with a standard model.

Pacific Biosciences, PacBio, SMRT, SMRTbell, iso-Seq, 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.