Resolving the 'Dark Matter' in Genomes

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

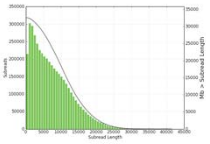
Introduction

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:

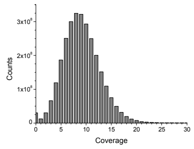
- **Long sequence reads:**
    - CHM1TERT:
        - Human cell line from complete hydatidiform mole
        - Equivalent of a haploid human genome, lack of allelic variation
        - Many associated datasets available for validation
    - Sequencing stats:
        - Total number of reads: 3,679,463
        - Total number of bases: 32,559,803,198
        - **Half of bases in reads: >10,985 bp**
        - 5% of sequenced DNA inserts: >18,060 bp
        - **Longest sequenced DNA insert: 41,460 bp**
        - PacBio® RS II sequencing time: 10 days



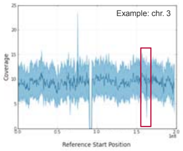
In collaboration with M. Chaisson, M. Malig, E. Eichler (HHMI, U of Washington)
http://blog.pacificbiosciences.com/2013/10/data-release-long-read-shotgun.html

- **Lack of sequence context bias, better mapping:**

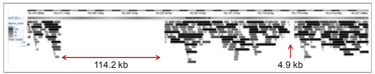
Coverage histogram:



Remapping against human reference:



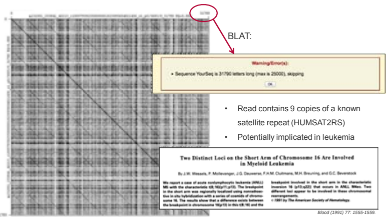
Direct detection of structural variation (deletions in the example shown here):



Confirmed by read-pair (CH17BES) and read depth (Illumina® Seq CHM1) data

- **Resolve repetitive regions:**

Single 31.8 kb read mapping to edge of centromere on chromosome 16:


BLAT:
Warning/Error(s):
• Sequence YourSeq is 31790 letters long (max is 25000), skipping
OK

- Read contains 9 copies of a known satellite repeat (HUMSAT2RS)
- Potentially implicated in leukemia

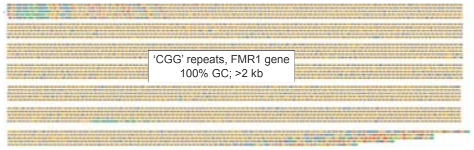
Two Distinct Loci on the Short Arm of Chromosome 16 Are Involved in Myeloid Leukemia

By J.W. Wessels, P. Mollevanger, J.G. Dauwerse, F.H.M. Cluitmans, M.H. Breuning, and G.C. Beverstock

Blood (1991) 77: 1555-1559.

Trinucleotide Repeat Expansions

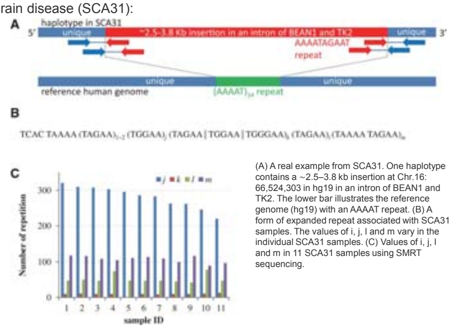
SMRT sequencing of previously unsequenceable, fragile X syndrome 'CGG' repeat full mutation allele:



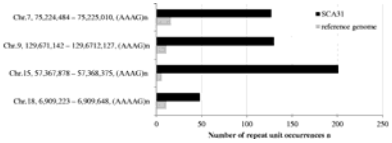
From: Loomis et al. (2013) Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Research* 23: 121-128.

Short Tandem Repeats

Locating and sequencing expanded short tandem repeats associated with a brain disease (SCA31):



(A) A real example from SCA31. One haplotype contains a ~2.5–3.8 kb insertion at Chr.16: 66,524,303 in hg19 in an intron of BEAN1 and TK2. The lower bar illustrates the reference genome (hg19) with an AAAAT repeat. (B) A form of expanded repeat associated with SCA31 samples. The values of i, j, l and m vary in the individual SCA31 samples. (C) Values of i, j, l and m in 11 SCA31 samples using SMRT sequencing.

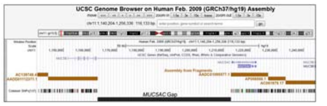


Sizes of the common STRs, $(AAAG)_n$ and $(AAAAG)_n$, at four genomic positions in the SCA31 sample and reference genome. Note that individual STR occurrences are significantly expanded in the SCA31 sample.

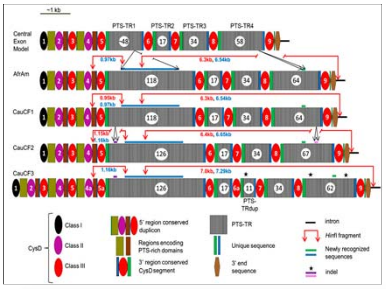
From: Doi *et al.* (2013) Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics* doi: 10.1093/bioinformatics/btt647

Complex Repeats

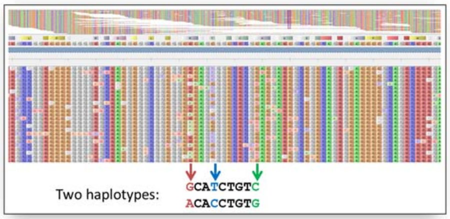
Resolving a previous gap in the human genome reference, updated in GRCh38:


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
MUC5AC Gap

MUC5AC genomic region: Current status and research design to cover the *MUC5AC* genomic gap. *(A)* Annotation tracks for the *MUC5AC* region excerpted from the UCSC genome browser (http://genome.ucsc.edu) (GRCh37/hg19) with notes added to emphasize the gap. The current gap in the *MUC5AC* gene is situated between a set of exons (blue vertical bars along arrowed line) that are 5′ (of note, the 5′ exons are incorrectly annotated to *MUC5B)* and a set of 3′ exons (that are correctly annotated to *MUC5AC*). The entire region is in general disarray.



Genetic variants and organization of the MUC5AC central exon. Sequence schematics representing the MUC5AC central exon from the African American subject and three white subjects with cystic fibrosis produced by *de novo* assembly were compared, as a group and individually, with the central exon model.



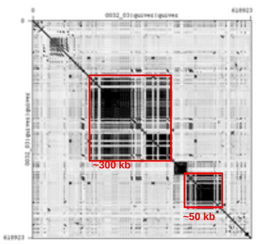
From: Guo et al. (2014) Genome Reference and Sequence Variation in the Large Repetitive Central Exon of Human MUC5AC. *Amer. J. Respiratory Cell & Mol. Biol.* 50:223-32

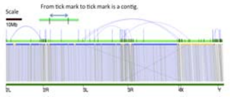
Y Chromosome

De Novo Drosophila assembly:

	Release 5 Reference genome	*De novo* assembly
chr2L	6 pieces	6 pieces
chr2R	27 pieces	2 pieces
chr3L	22 pieces	1 piece
chr3R	15 pieces	3 pieces
chr4	2 pieces	3 pieces
chrX	3 pieces	42 pieces
Effort	many libraries shotgun Sanger + BAC + Opgen + manual finishing 10+ years $millions	1 week – collect DNA 1 week – sample prep 6 days – sequencing 3 weeks – assembly ~$15,000

618 kb Y-chromosome contig:



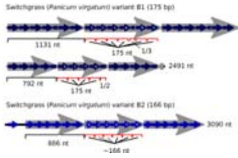


Only ~1% of chromosome Y represented in reference (Release 5). Anticipating >50% can be assembled with this data, including complex regions shown here.

http://blog.pacificbiosciences.com/2014/01/data-release-preliminary-de-novo.html

Centromeres

Novel patterns of higher order repeat structures in switchgrass centromeres:

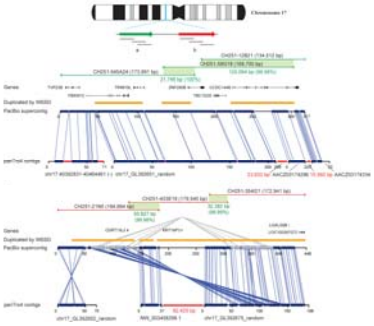

Switchgrass (Panicum virgatum) variant B1 (175 bp)
7032 nt
1131 nt
175 nt
1/3
2491 nt
792 nt
175 nt
1/2
Switchgrass (Panicum virgatum) variant B2 (166 bp)
3090 nt
886 nt
~166 nt

Pacific Biosciences sequencing shows homogeneity of repeat arrays and detects long higher order repeat structures. Although no repeat variant mixing was detected in the PacBio reads, several higher order repeat (HOR) structures were found in longer PacBio reads. These HOR structures consisted of a mixture of complete and truncated repeats. Two switchgrass variant B1 centromere reads with higher order structure and one switchgrass variant B2 centromere repeat are shown.

From: Melters *et al.* (2013) Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, 14:R10

Segmental Duplications

Upgrading a chimpanzee genomic region:

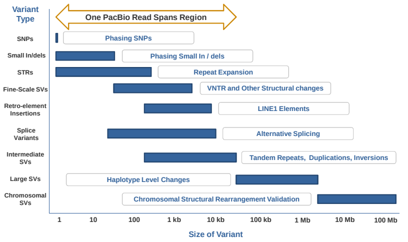


Sequence and assembly of six large-insert clones (CH251) from two segmental duplication blocks (red and green) are aligned to their corresponding sequences from the 17p11.2 Smith-Magenis region of the chimpanzee reference assembly (panTro4). 31% (241/766 kbp) of the chimpanzee sequence is missing within the working draft assembly. Gaps in the panTro4 contigs are indicated in red. Gene annotations are shown based on a custom liftover from RefSeq annotations of GRCh37 in the corresponding regions of 17p11.2. The missing sequence corresponds to high-identity segmental duplications (orange bars represent segmental duplications predicted by whole-genome shotgun sequence detection or WSSD).

From: Huddleston *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Research*, doi:10.1101/gr.168450.113

Summary

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


Variant Type
One PacBio Read Spans Region
SNPs
Phasing SNPs
Small In/dels
Phasing Small In / dels
STRs
Repeat Expansion
Fine-Scale SVs
VNTR and Other Structural changes
Retro-element Insertions
LINE1 Elements
Splice Variants
Alternative Splicing
Intermediate SVs
Tandem Repeats, Duplications, Inversions
Large SVs
Haplotype Level Changes
Chromosomal SVs
Chromosomal Structural Rearrangement Validation
1 kb
10 kb
100 kb
1 Mb
10 Mb
100 Mb
Size of Variant