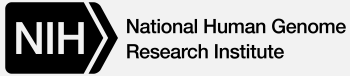
The era of long reads

Sergey Koren

Staff Scientist, Genome Informatics Section, NHGRI



@sergekoren

The Forefront of Genomics®

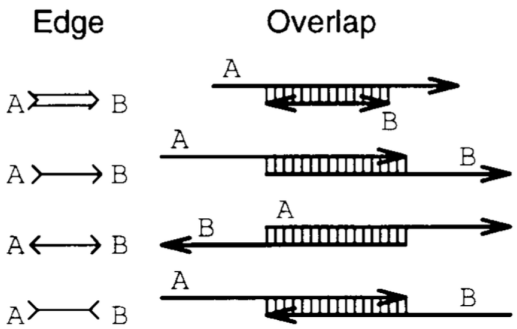
Assembly review

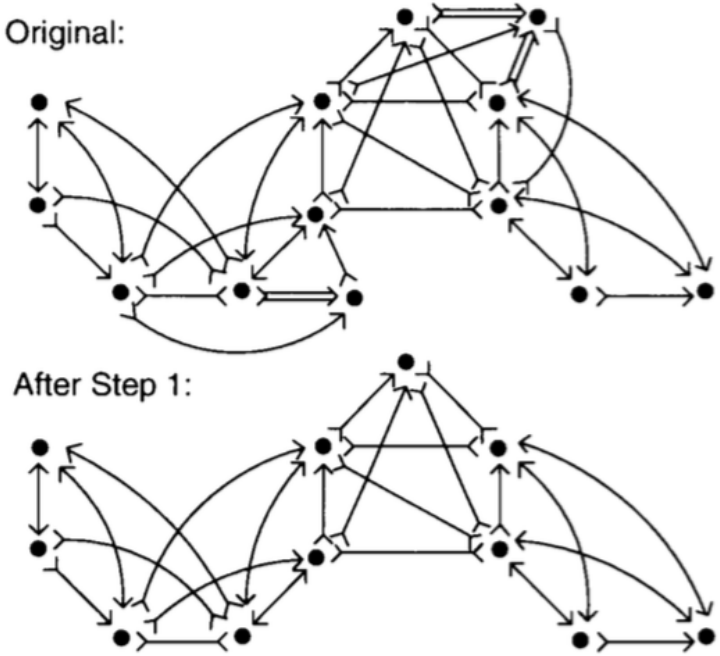
Genome Assembly

- Assembling a puzzle with a billion pieces

Assembly the Celera way

- Step 0:
  - Find overlaps



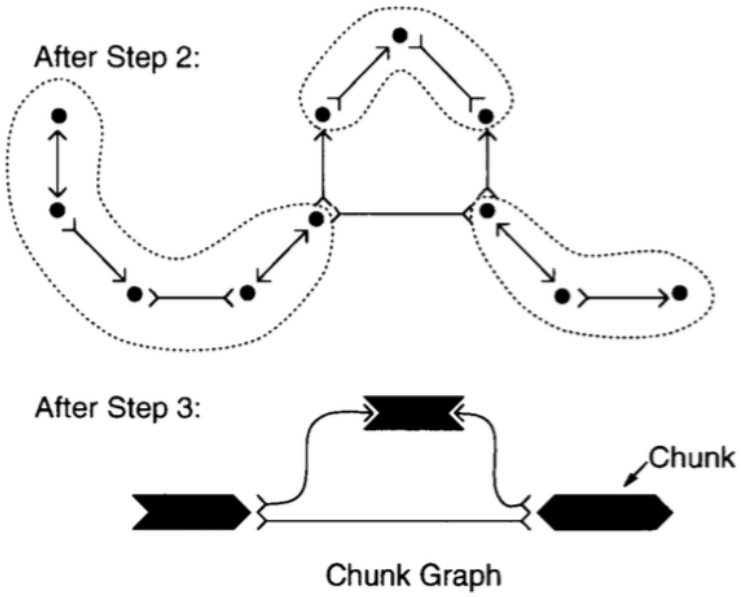


- Step 1:
  - Remove contained

Toward Simplifying and Accurately Formulating Fragment Assembly.
Myers. *Journal of Computational Biology* (1995)

Assembly the Celera way

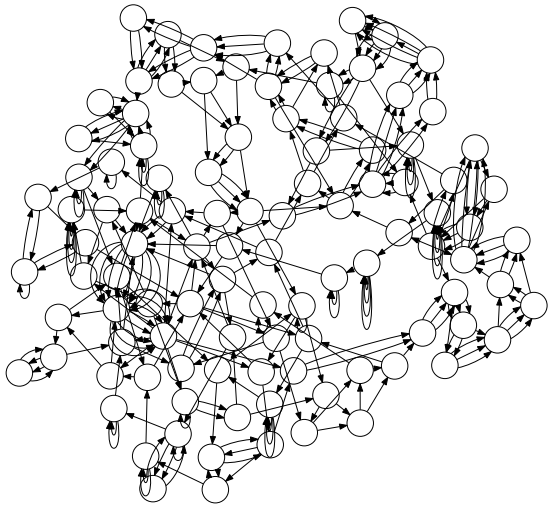
- Step 2:
  - Transitive reduction
- Step 3:
  - Collapse unique
- Output
  - “Unitigs”



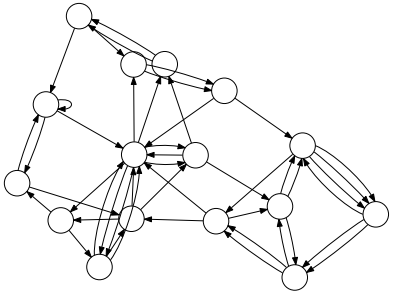
Toward simplifying and accurately formulating fragment assembly.
Myers. *Journal of Computational Biology* (1995)

Read length matters (*E. coli*)

k = 50



k = 1,000

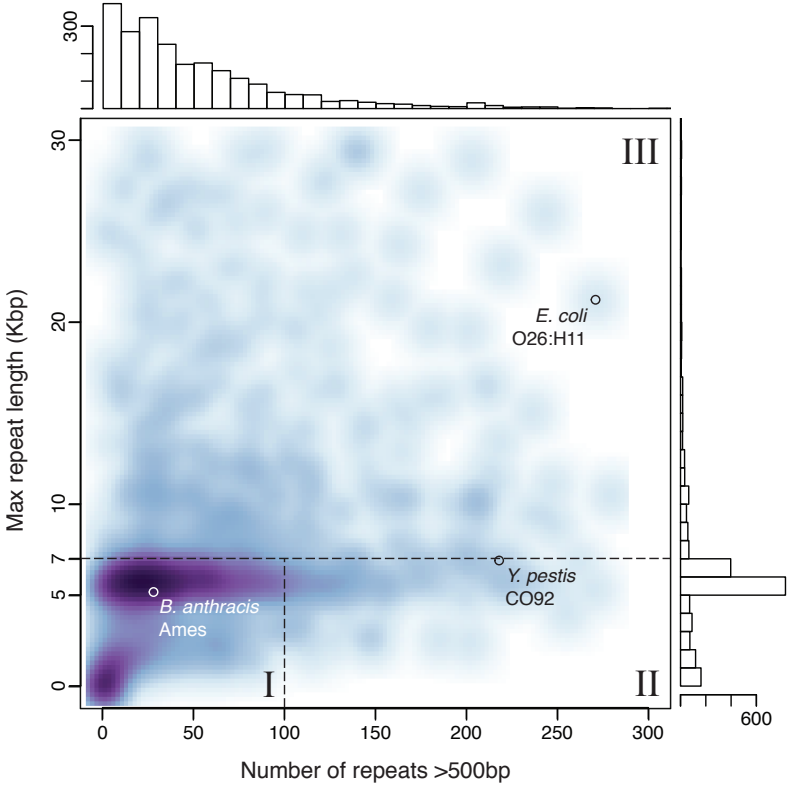


k = 7,000

* No errors, perfect coverage, uniform read length

- **One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly.** Koren and Phillippy. *Current Opinion Microbiology* (2015)

How long are microbial repeats?



7 kbp golden threshold

- **Reducing assembly complexity of microbial genomes with single-molecule sequencing.**
  Koren et al. *Genome Biology* (2013)

A new era of sequencing

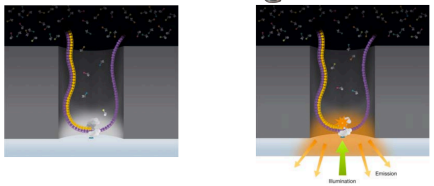
NIH
NHGRI

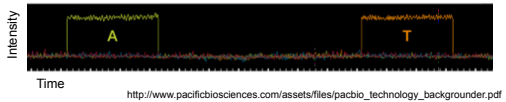
PacBio Sequel II

- Single Molecule sequencer (one DNA strand)
  - Ligate adapters to make a bell 
  - Load molecules onto zero mode waveguides
  - Real-time polymerase sequencing
  - Video analysis

- Capable of sequencing long molecules
  - 10-60 kbp

- High error (85-90% accuracy) but random
  - Can read shorter reads multiple times
  - Converges to near-perfect consensus

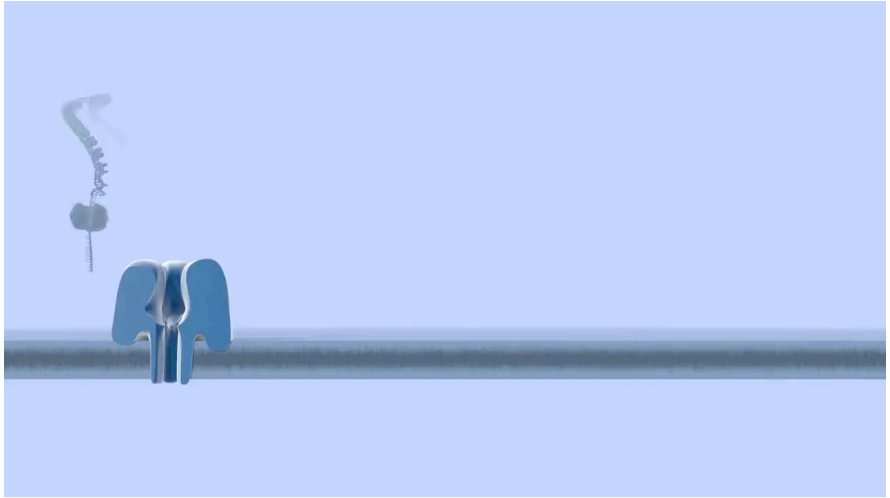




http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

Oxford Nanopore MinION

$1000 (free) instrument
$100 / bacterial genome
85–95% read accuracy



Oxford nanopore technologies

Long read assembly in practice

NIH
NHGRI

Real data is messy

- Every technology has its own quirks
- Tools developed for one don’t work on others
- Best tool may not be the theoretically optimal but best engineered

Example: PacBio Sequel II

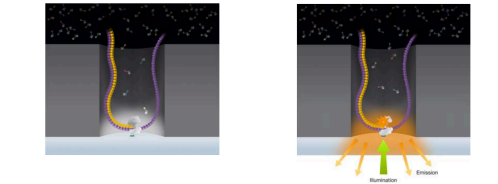
- Single Molecule sequencer (one DNA strand)
  - Ligate adapters to make a bell 
  - Load molecules onto zero mode waveguides
  - Real-time polymerase sequencing
  - Video analysis

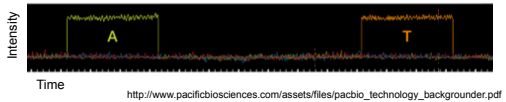
What can go wrong

- More than 1 read loaded into a well
  - Chimeric sequence when basecaller mixes them

- Read goes around adapter
  - Same sequence (forward then complement strand)

- Secondary DNA structure slows down/confuses polymerase





http://www.pacificbiosciences.com/assets/files/pacbio_technology_backgrounder.pdf

Example: Oxford Nanopore MinION

- Single Molecule sequencer (one DNA strand)
    - Ligation or transposase to add adapter
    - Load molecules onto flowcell guides
    - DNA denatured in real time and passed through pore
    - Signal analysis to identify bases

What can go wrong

- Two reads pass through same pore quickly
    - Chimeric sequence when not detected
    - Can be same as PacBio chimera (fwd then comp)

- Continuous current mistaken for empty pore
    - Single read split into multiple parts

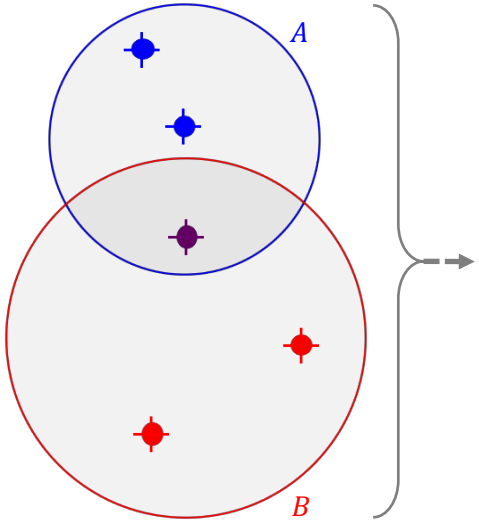
- DNA structure re-folding on the other side of the pore
    - Can make one strand higher error than the other

In summary

- Long-read data is noisy
  - Base errors
  - Chimeric reads
  - *Solution:* read clustering, correction, and trimming
- Overlaps are long, and graph is big
  - All-pairs alignment is slow
  - Full graph is a giant tangle (due to repeats)
  - *Solution:* MinHash “best” overlap graph
- *D. melanogaster* results
  - Celera Assembler v8: **630,000** CPU hours, 15 Mbp NG50
  - Canu v1: **500** CPU hours, 21 Mbp NG50

Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.
Berlin et al. *Nature Biotechnology* (2015)

Fast overlapping with MinHash



$S(A \cup B) = \{42, 64, 66, \mathbf{82}, 87\}$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

On the resemblance and containment of documents. Broder (1997)

tf-idf weighted MinHash

The Stolen White Elephant by Mark Twain

The Tell-Tale Heart by Edgar Allan Poe

Shooting an Elephant by George Orwell

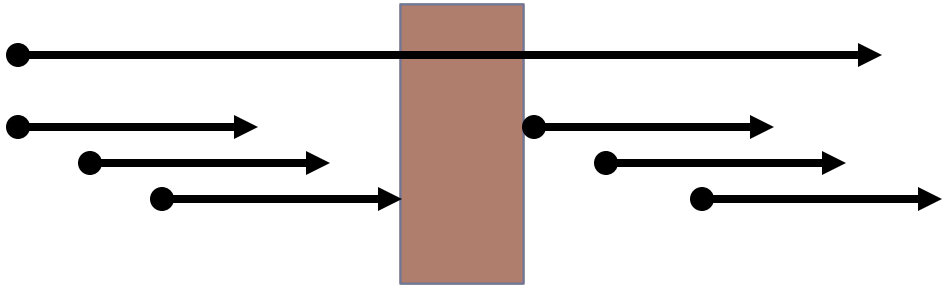
A few extra details

- Throw hashes in hash table for all-pairs speedup
  - Only look at reads sharing some minimum number
- Jaccard based on k-mers, want a base error rate
  - Estimate from k-mers in the first round of overlapping
  - Compute exactly in the second round for contigging
- *tf-idf* weighted MinHash
  - Common repeats more likely to get larger hash value
  - Distinctive words more likely to get smaller hash value
  - Lower memory and runtime *without* k-mer filtering
- Keep position for each hash
  - Can be used to approximate the overlap bounds
  - (See German tank problem)

* And it's written in Java

Overlap-based correction and trimming

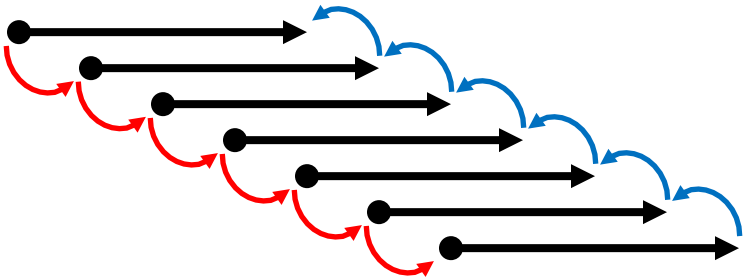
- Every (long) read corrected by its overlaps
  - Consensus called for covered bases
  - Missing coverage suggests low-quality or chimeras
  - Read correction acc: >99% PacBio, <98% Nanopore



- Data cleaning is key to assembly
  - Necessary, not glamorous

Best overlap graph

- After transitive reduction, only best are left
  - With enough coverage, nearly a global alignment
  - Find the “best” 5’ and 3’ overlap for each read
  - Build a graph from these edges

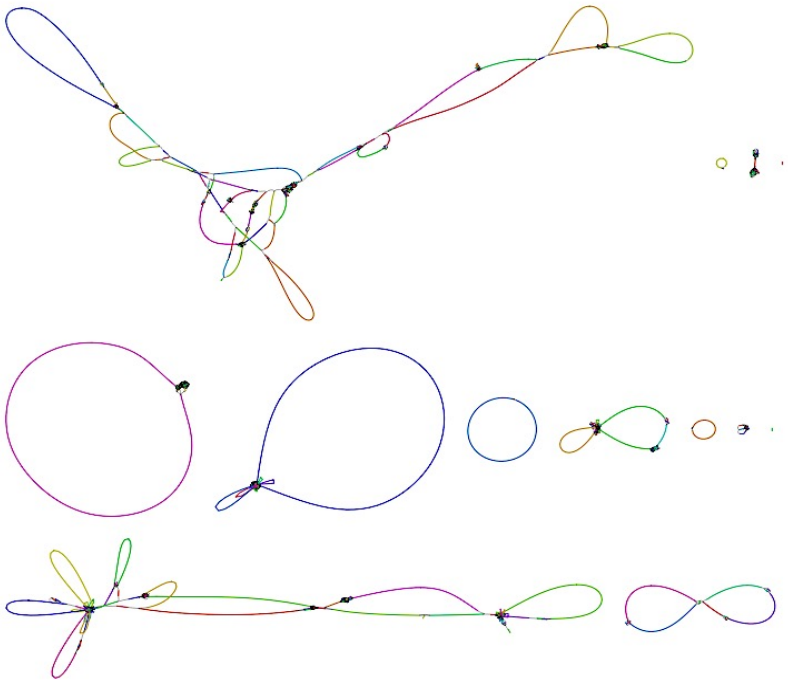


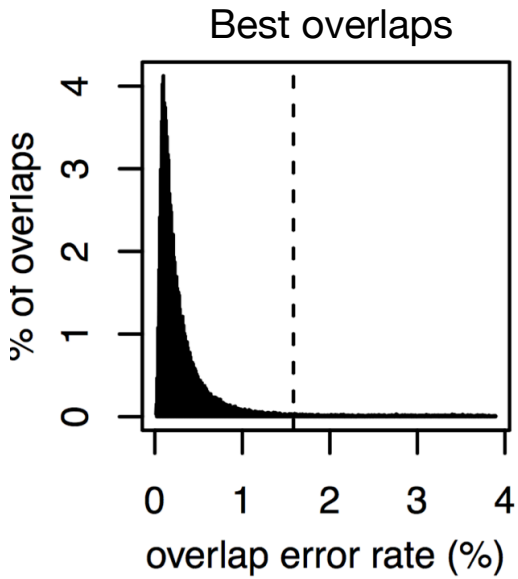
- Greedy approach, can be mislead by repeats
  - Works great if given only “true” overlaps

Check your work

- Overlap filtering + greedy = pretty good
  - Automatically split divergent repeats and alleles
- Can still make mistakes, so…
  - Annotate repeats within contigs using overlaps
  - Check repeats for spanning reads
  - Check local error rate across each contig
  - Break on suspicion of misjoin
- Complete the graph with non-best overlaps

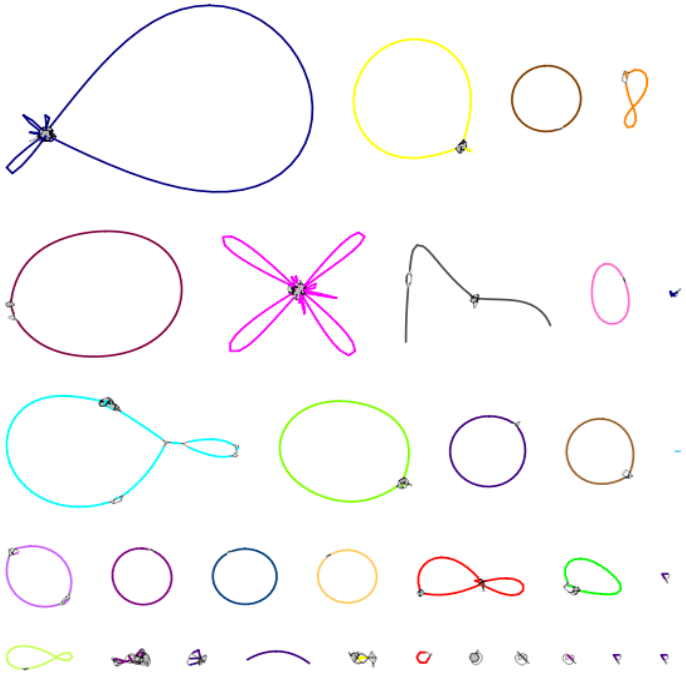
Repeat and haplotype separation

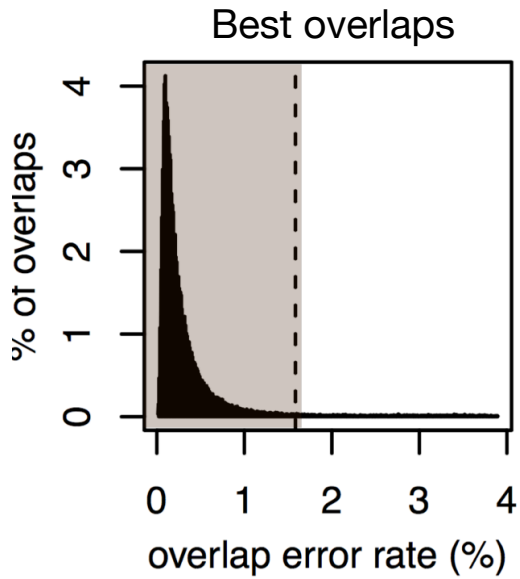




- Don’t know the read error rate a priori

Repeat and haplotype separation





- Differentiate true from false overlaps

Can long reads solve assembly?

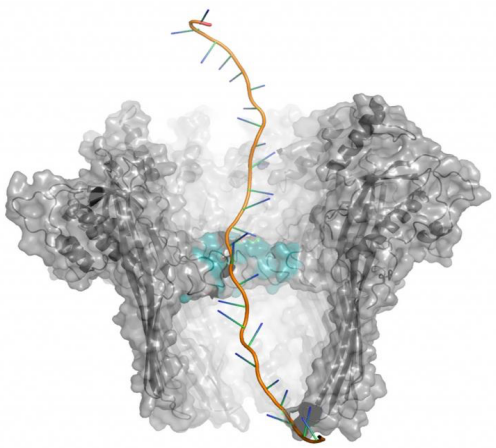
Yes

How long do reads need to be, for human?

- **How long are the repeats?**
  - 7 kbp LINEs
  - 1 Mbp+ rDNA arrays
  - 1 Mbp+ centromere arrays
  - 10 Mbp+ heterochromatin blocks
- **Coverage and accuracy matter too**
  - 1,000X of 100 bp reads at 100% accuracy? NO
  - 10X of 10,000,000 bp reads at 100% accuracy, YES
  - 100X of 100,000 bp reads at 90% accuracy, MAYBE?

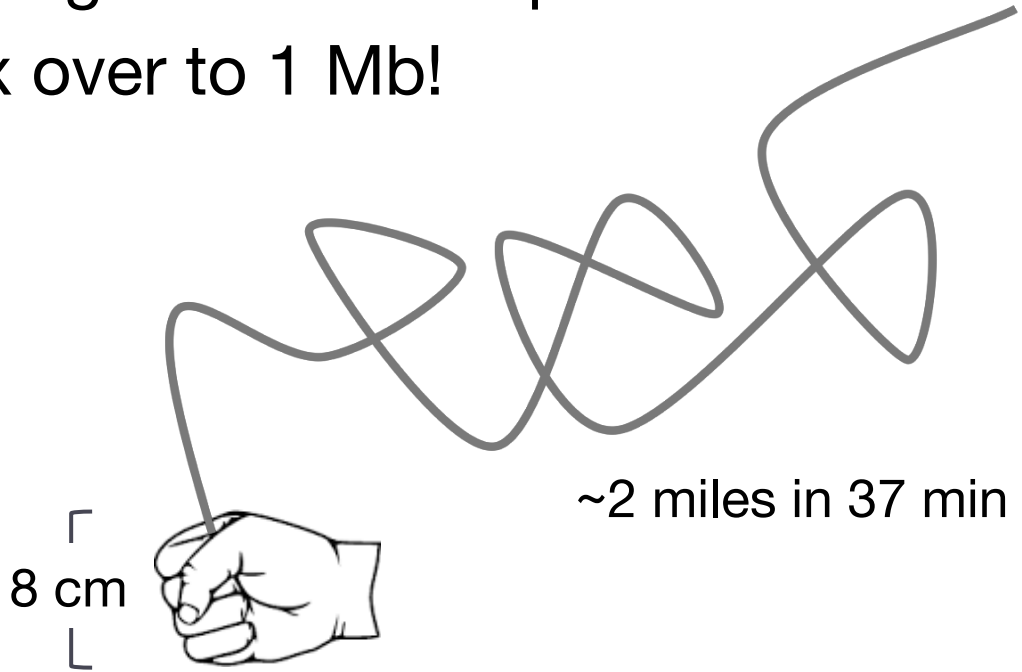
Ultra-long read sequencing

- ONT R9 pore: *E. coli* CsgG membrane protein
- 100 kb read N50, max over to 1 Mb!



85 Å

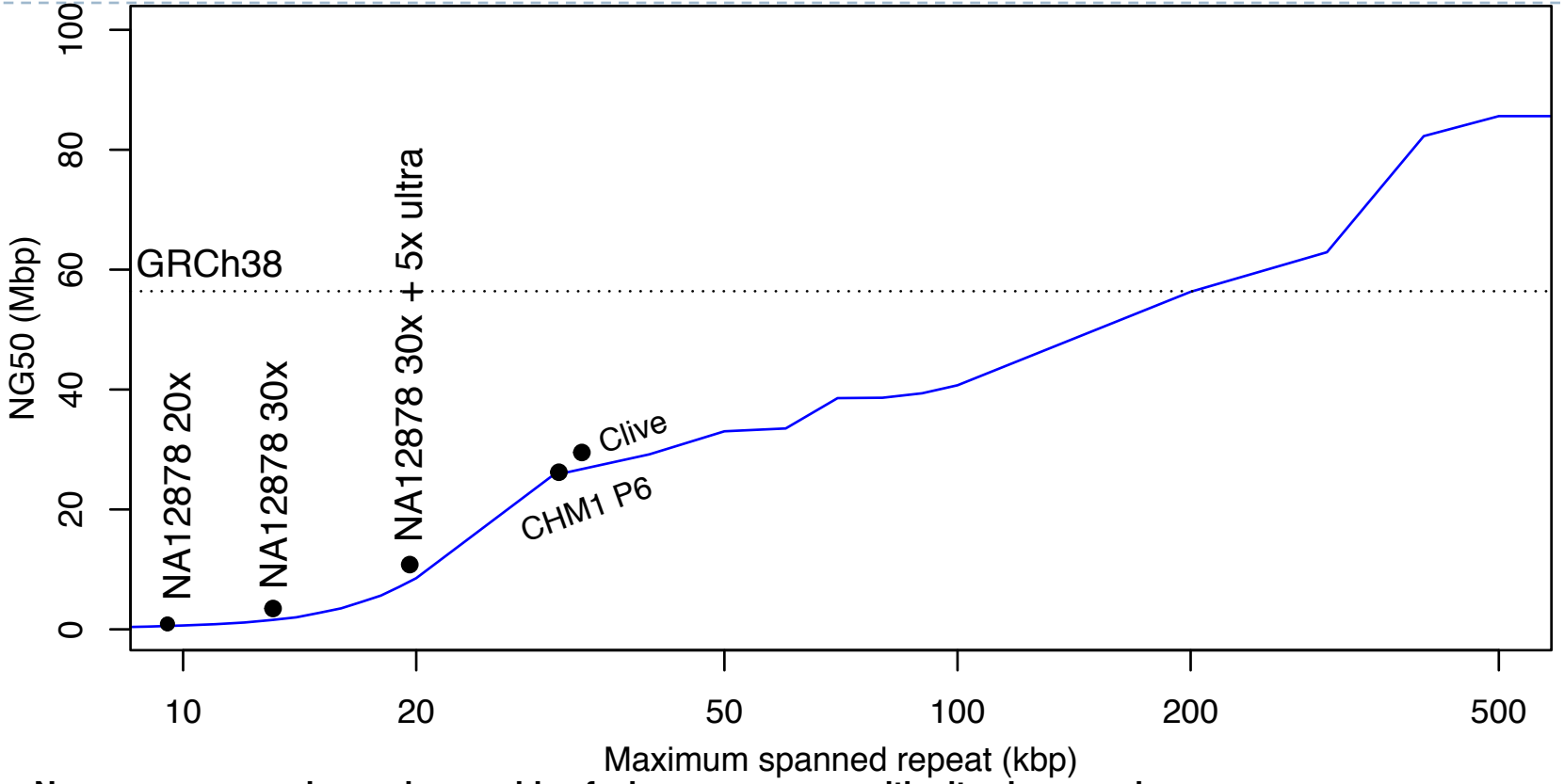
120 Å



*Assuming 3.4 Å per bp, 1 Mbp = 3,400,000 Å (0.34 mm) = 40,000x height of the pore

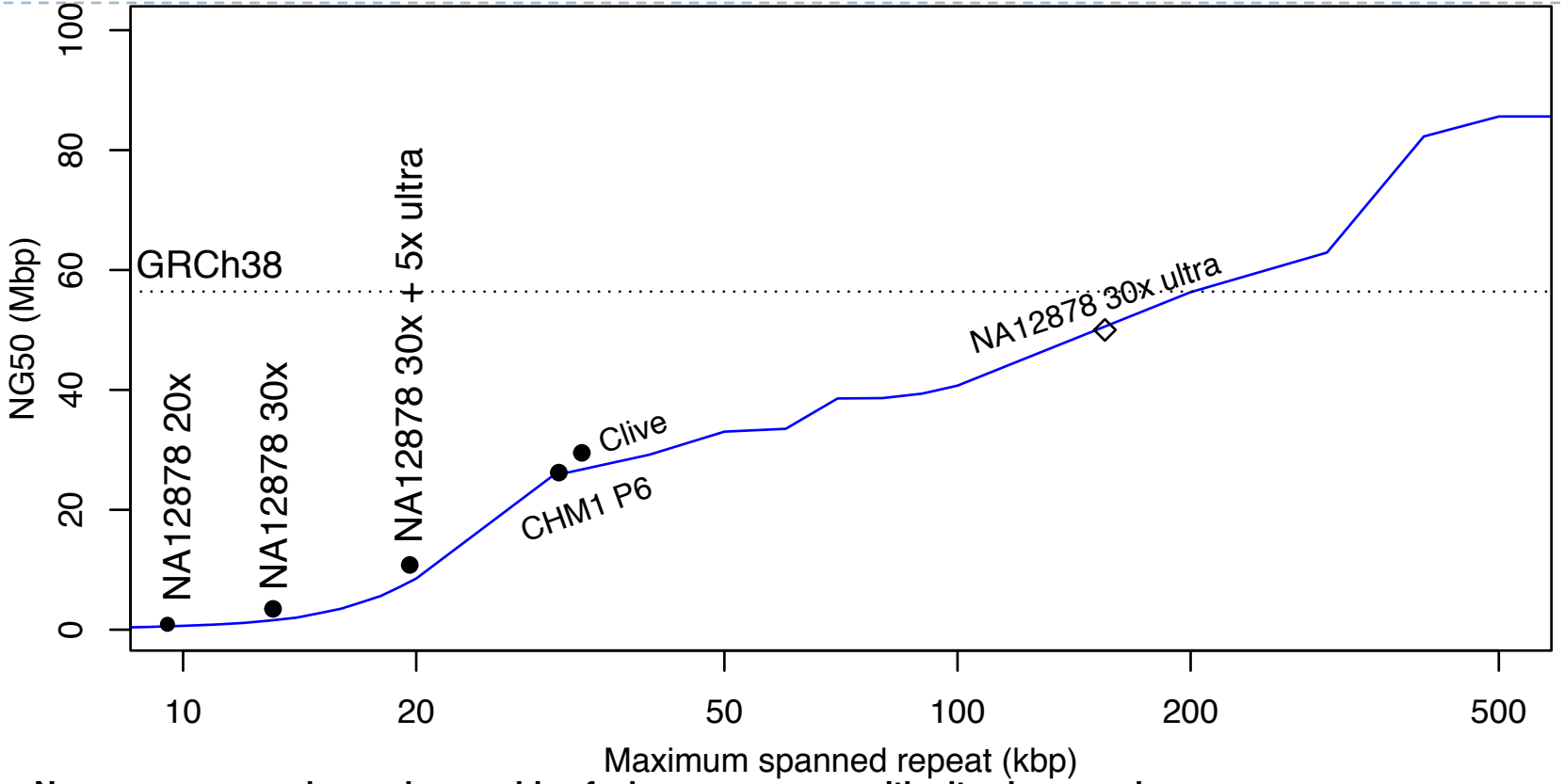
- http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/ (Josh Quick & Nick Loman, U. Birmingham)

Ultra-long read benefits



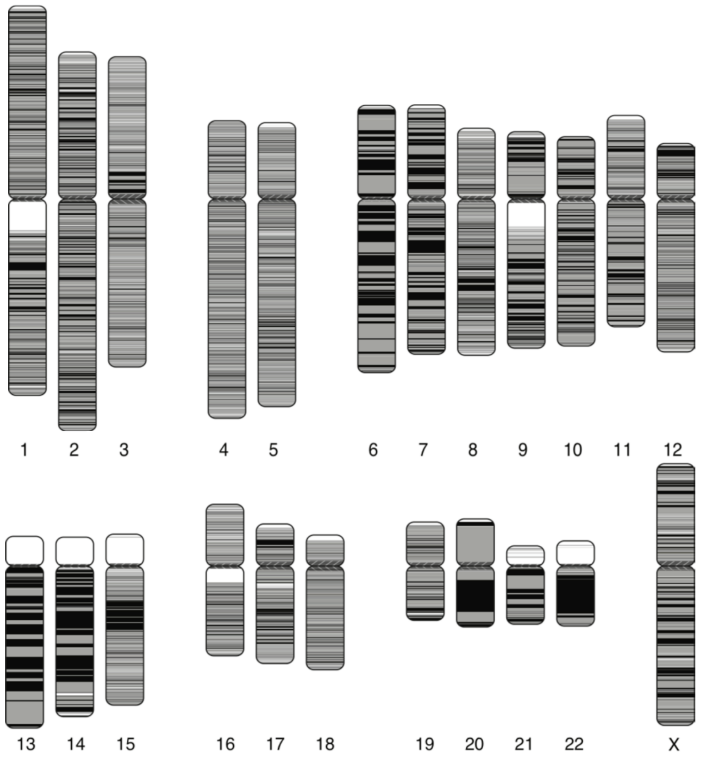
- **Nanopore sequencing and assembly of a human genome with ultra-long reads.**
  Jain, Koren, Miga, Quick, Rand, Sasani, Tyson, et al. *Nature Biotech* (2018)

Ultra-long read benefits



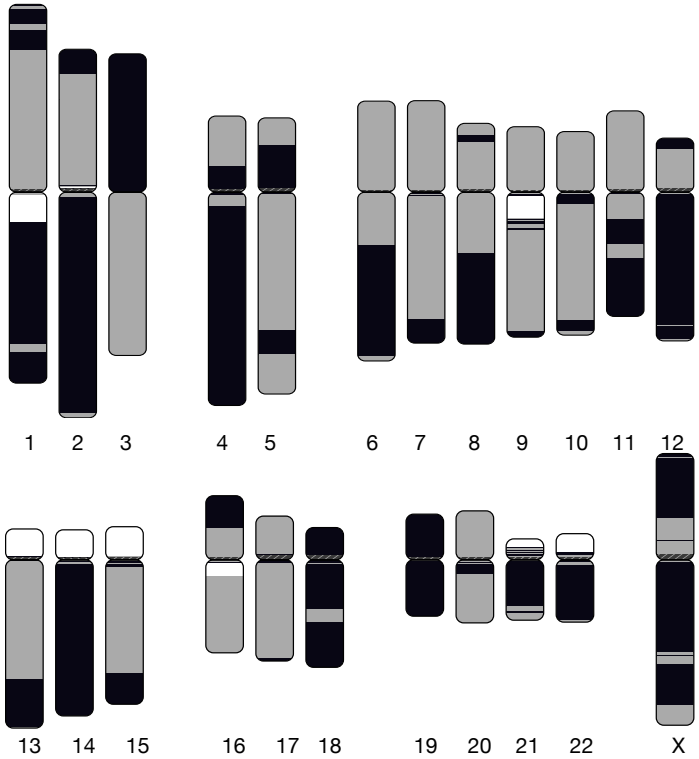
- **Nanopore sequencing and assembly of a human genome with ultra-long reads.**
  Jain, Koren, Miga, Quick, Rand, Sasani, Tyson, et al. *Nature Biotech* (2018)

Human genome, 2001



ref28 / hg10 : N50 0.5 Mbp

The human genome, 2017



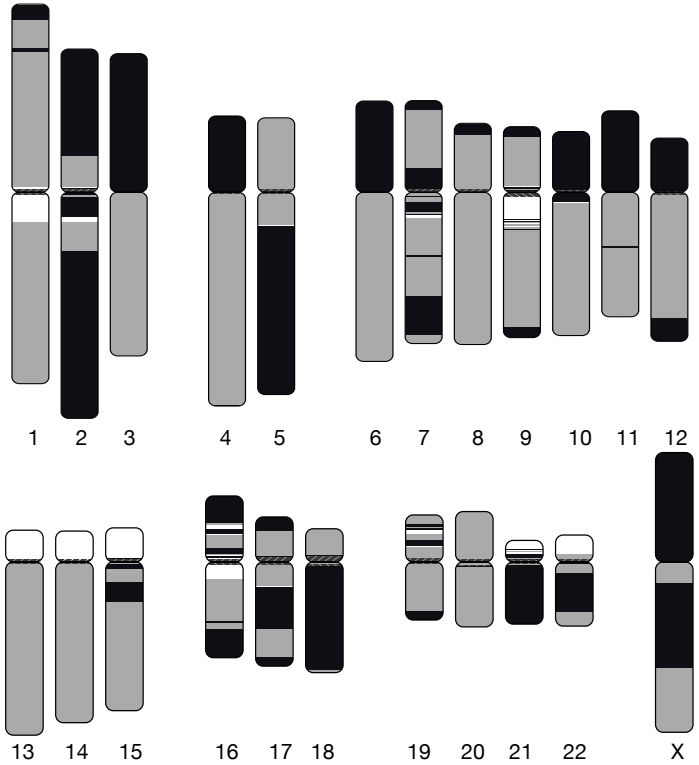
GRCh38 NG50 contig 56.4 Mbp

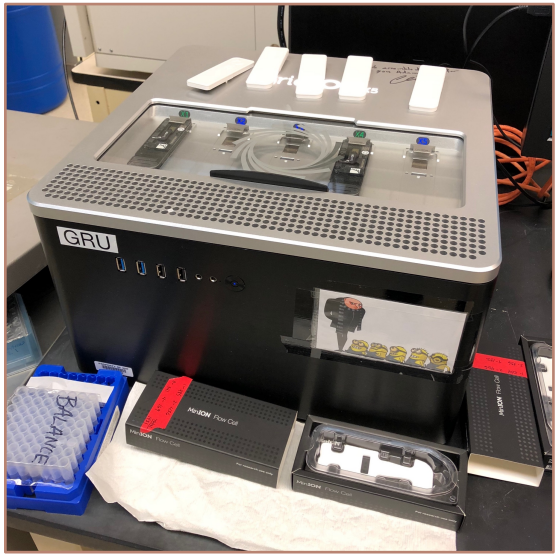
GRCh38

The Genome Reference Consortium consists of:

- Wellcome Sanger Institute
- The McDonnell Genome Institute at Washington University
- The European Bioinformatics Institute
- The National Center for Biotechnology Information

The human genome, 2018



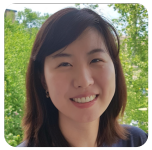


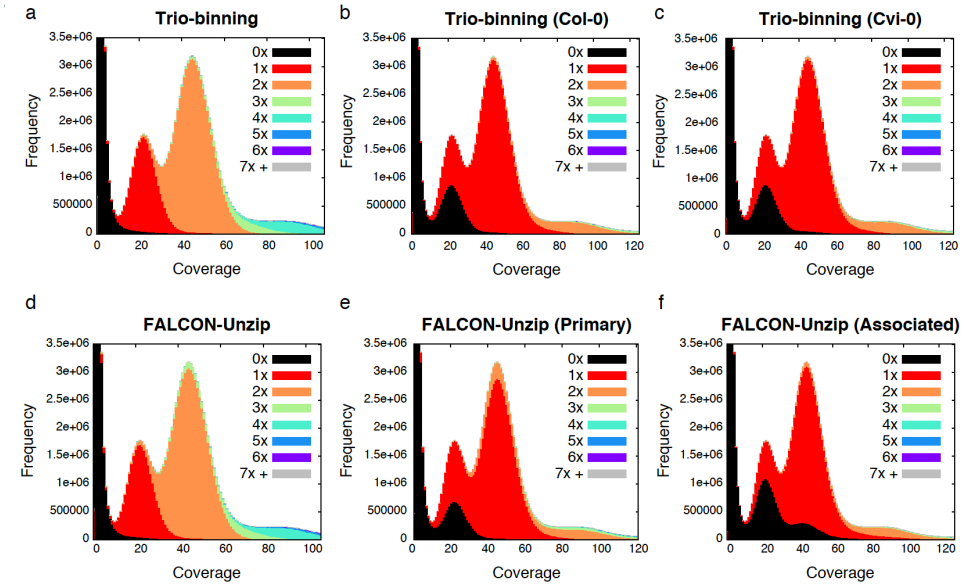
CHM13 NG50 contig 79.5 Mbp (50x UL ONT)

An assembly is a hypothesis

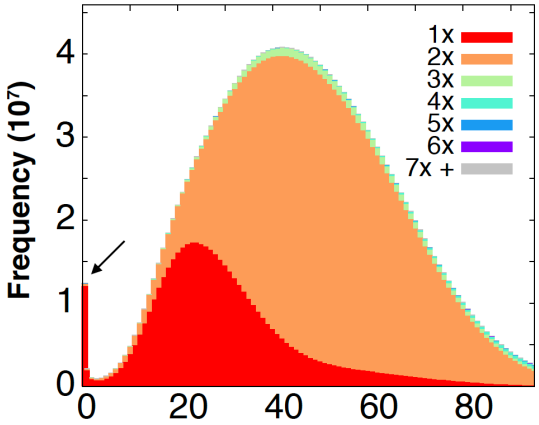
NIH
NHGRI

K-mers as a measure of completeness





- K-mers only in assembly (misassembled bps)
- Haplotype completeness
- Over-assembled (duplications)
- Repeat copies ~ exp. copies?

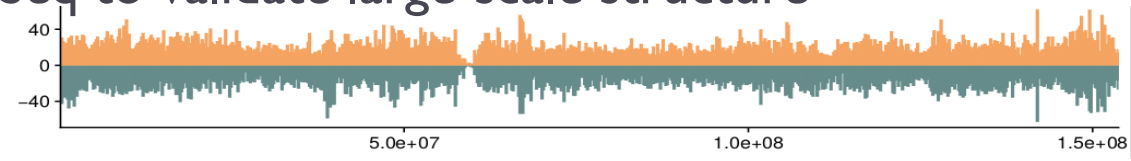


KAT Spectra-cn plots:
https://github.com/TGAC/KAT

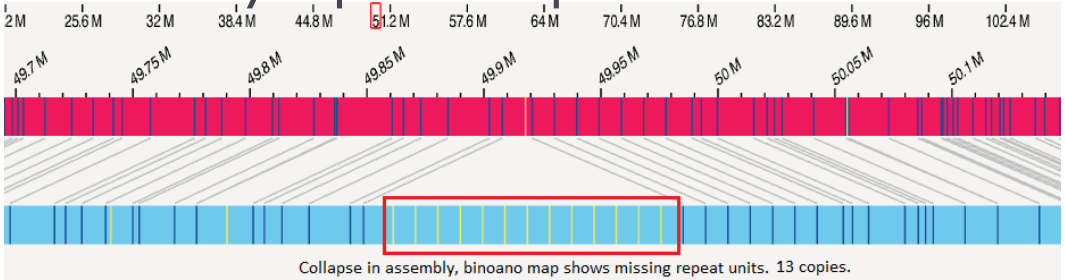
Mapleson *et al.*, Bioinformatics (2016)

Complementary technologies

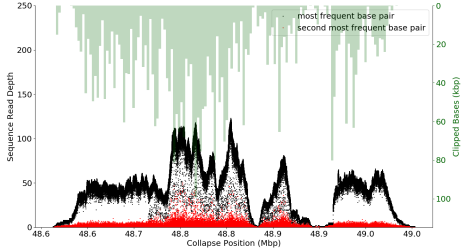
- StrandSeq to validate large-scale structure

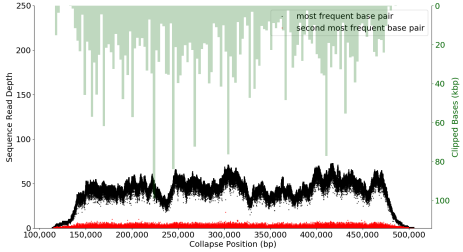


- BioNano to identify repeat collapse/errors



- Mapping to identify low-quality regions





Who said assembly wasn't cool?

"All the News That's Fit to Print"

The New York Times

Washington Edition

Today, mostly cloudy, a few showers, a thunderstorm, high 79. Tonight, a passing shower late, low 57. Tomorrow, a morning shower, high 64. Weather map is on Page A18.

VOL. CLXV . . . No. 57,189 © 2016 The New York Times FRIDAY, APRIL 1, 2016 $2.50

MARK KUNITOMI

Each line is a piece of the Aedes aegypti mosquito genome that could not be fit together without data scientists hope to produce.

Tweets Inspire Rival Scientists To Come Together to Fight Zika

By AMY HARMON

With the Zika virus spreading largely unchecked in Latin America and the Caribbean by way of a now-notorious insect, some of the

or cattle, for desirable — or, at least, less undesirable — traits, such as a preference for biting animals other than humans.

Five Leading Female Players Allege U.S. Soccer Wage Bias

By ANDREW DAS

U.S. Soccer, the governing body for the sport in America, pays the members of the men's and women's national teams who

women's athletes, said they were being shortchanged on everything from bonuses to appearance fees to per diems.

Minimum Pay For New York Rises in Deal

$15 in City by End of '18 — Slower Elsewhere

By JESSE McKINLEY and VIVIAN YEE

ALBANY — Gov. Andrew M. Cuomo and state legislative leaders announced on Thursday that they had reached a budget agreement that would raise the minimum wage in New York City to $15 by the end of 2018, but initiate slower increases elsewhere, even in the city's wealthy suburbs.

For Mr. Cuomo, the wage agreement came with clear concessions, as some lawmakers outside the city won a softer phase-in period. Long Island and Westchester County will not reach a $15 wage for nearly six years; areas north of Westchester are assured only of reaching $12.50 by 2021.

In announcing the $15 wage, New York became the second state to embrace that threshold; California lawmakers passed a similar measure only hours earlier on Thursday. The movement began in earnest in New York City in late 2012, when fast-food workers began the so-called Fight for $15, which became a nationwide effort to increase wages and support unions.

The wage issue was one of the final obstacles to reaching a timely accord on the next state budget, estimated to be more than $150 billion, which mixes high-profile policy measures with nuts-and-bolts infrastructure spending and aid to schools.

With the minimum wage deal

OPPONENTS SEE WISCONSIN VOTE HOBBLING TRUMP

KEY TEST AS RACE SHIFTS

Momentum Challenged — State G.O.P. Lines Up in Opposition

By TRIP GABRIEL

JANESVILLE, Wis. — The danger signs are mounting for Donald J. Trump in Wisconsin: Right-wing radio hosts are flaying him, Gov. Scott Walker and other elected Republicans have endorsed Senator Ted Cruz, and a new poll showed Mr. Cruz with a 10-percentage-point lead in the state before Tuesday's primary.

The Stop Trump movement may never have another opportunity like the one here, where resistance to Mr. Trump was running high even before his campaign became consumed by a new round of controversies, from his mocking of Mr. Cruz's wife to the arrest of his campaign manager to his comments in favor of punishing women who get abortions.

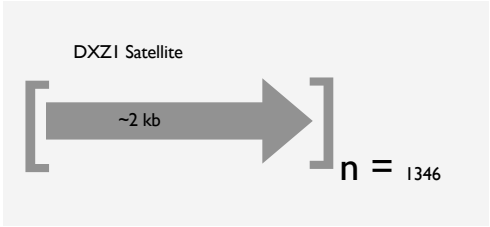
If Mr. Trump is dealt a setback in the Wisconsin primary, including a potential sweep by Mr. Cruz of all 42 delegates, it would be his most prominent reversal since his second-place finish in the Iowa caucuses in February. And it would show Mr. Trump's vulnerability before the race moves to New York and other North-

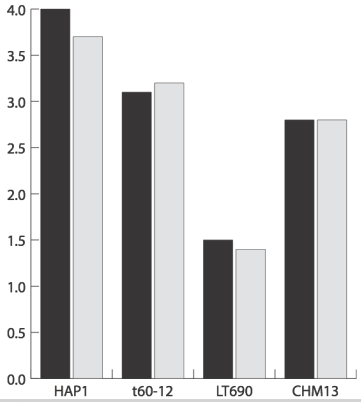
- April 1, 2016

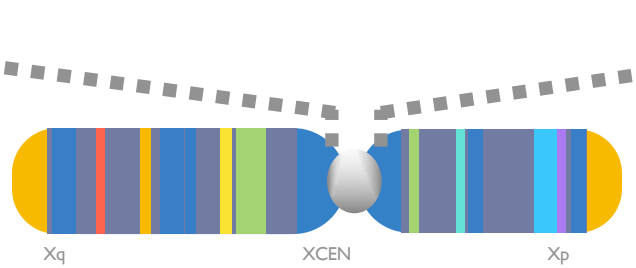
Assembly is not solved

NIH
NHGRI

X Centromere Detail

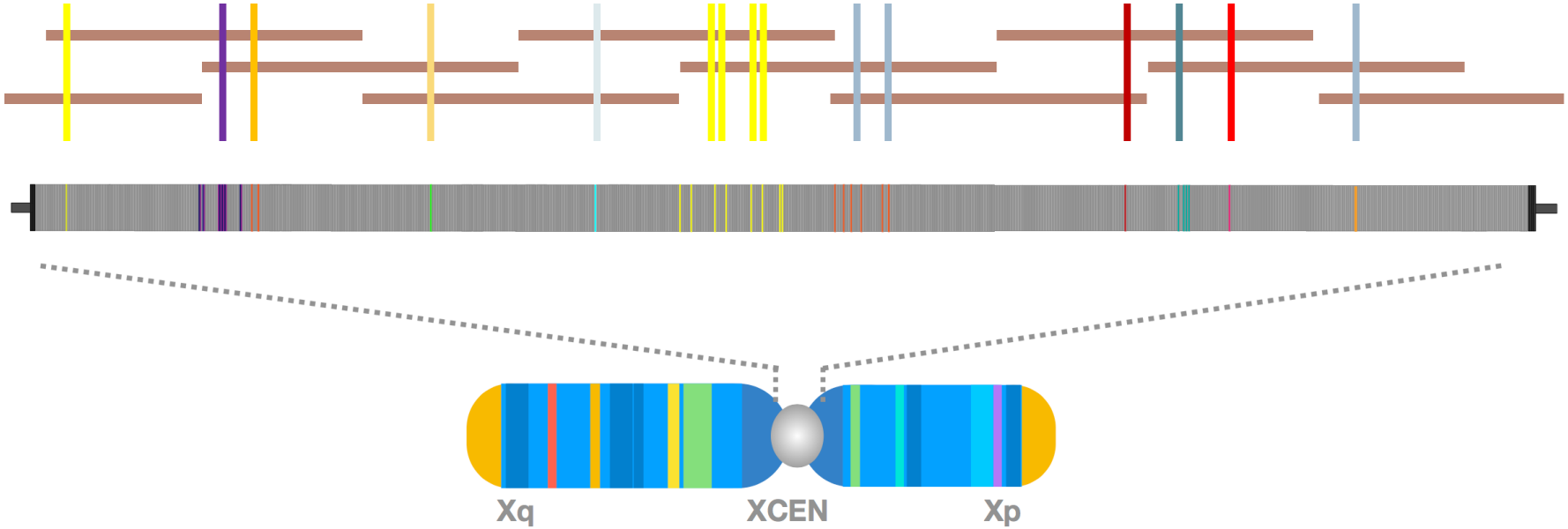






Stitching across the X centromere

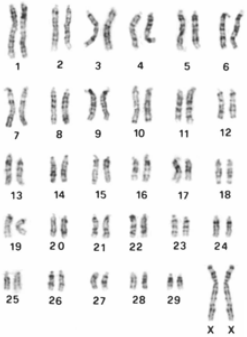
- Unique structural variants from PacBio
- Unique k-mers confirmed by Duplex-Seq



There isn’t a single “genome”

NIH
NHGRI

The *genomes* assembly problem



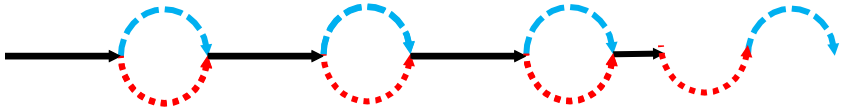
Esperanza

Duke, highland sire

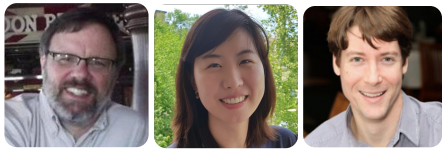
Molly, yak dam

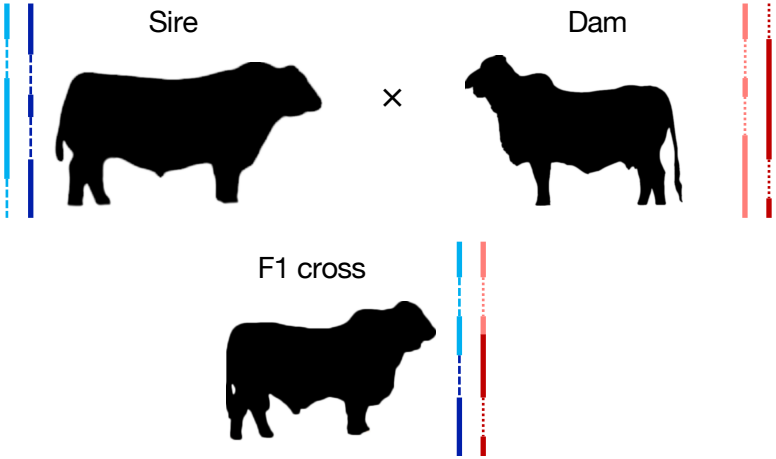
~1% heterozygosity

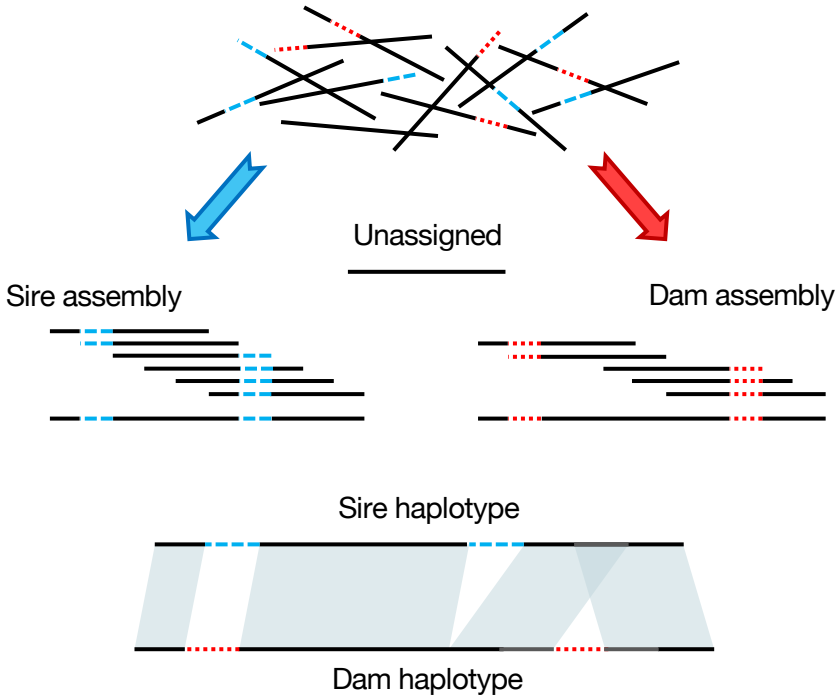
State of the art: pseudo-haplotype



Trio binning with TrioCanu



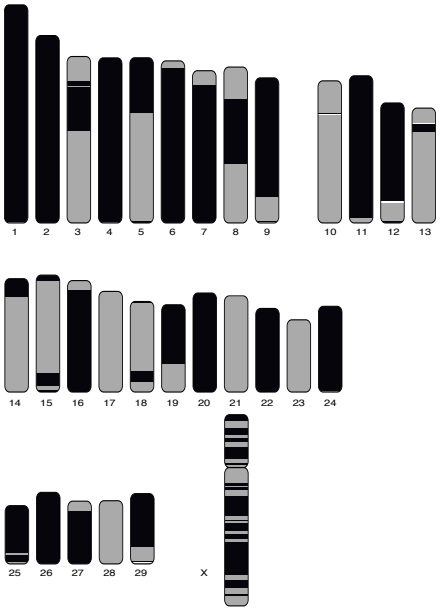




- **Complete assembly of parental haplotypes with trio binning.**
  Koren, Rhie et al. 2018

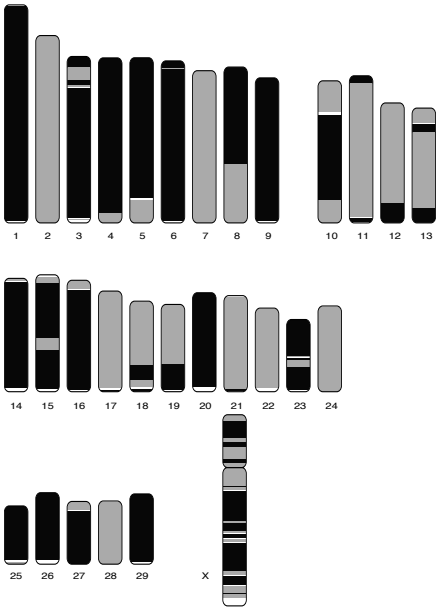
Esperanza: The nearly perfect diploid

Sire (Highland)



Esperanza

Dam (yak)



125x PacBio coverage (~60x per haplotype), **TrioCanu haplotig NG50 70 Mbp, BUSCOs 94%**

Acknowledgements

genomeinformatics.github.io

- Sergey Koren
- Brian Walenz
- Alexander Dilthey
- Arang Rhie
- Brian Ondov
- Chirag Jain
- Anna Sappington

AD CJ SK BO AP AR AS BW

canu.readthedocs.io

- Sergey Koren
- Brian Walenz
- Konstantin Berlin
- Jason Miller

- GM12878 collaborators
- T2T collaborators
- VGP collaborators
- Cattle Collaborators