Telomere-to-telomere assembly of a complete human X chromosome

Sergey Koren

Staff Scientist, Genome Informatics Section, NHGRI

National Human Genome Research Institute

@sergekoren

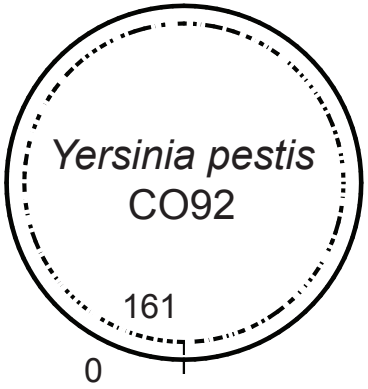
The **Forefront** of **Genomics**®

40 years of genome assembly

- **Rodger Staden (1979)**
  - “If the 5′ end of the sequence from one gel reading is the same as the 3′ end of the sequence from another the data is said to overlap. **If the overlap is of sufficient length to distinguish it from being a repeat** in the sequence the two sequences must be contiguous. The data from the two gel readings can then be joined to form one longer continuous sequence.”

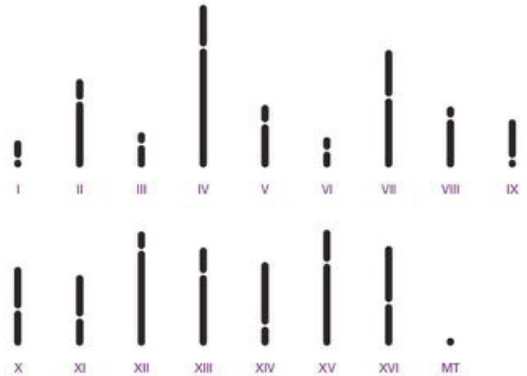
Are we there yet?

- **First complete *de novo* assemblies**
  - 2012: Bacteria (10^6 bp)



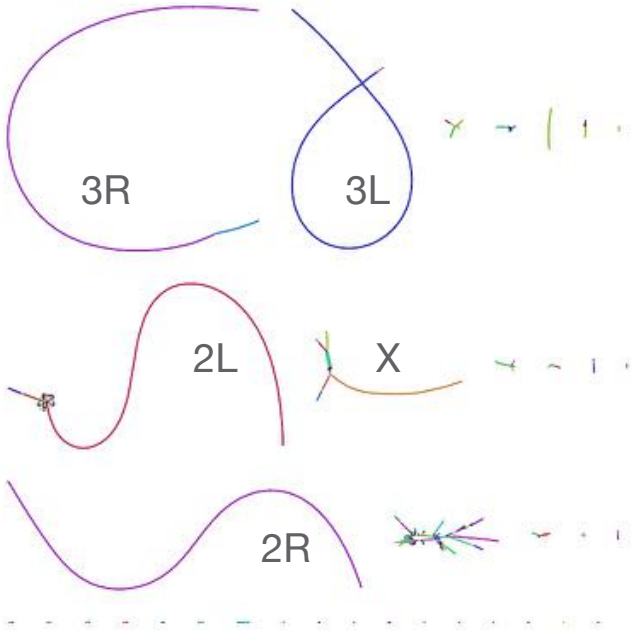
Are we there yet?

- **First complete *de novo* assemblies**
  - 2012: Bacteria (10^6 bp)
  - 2014: Yeast (10^7 bp)



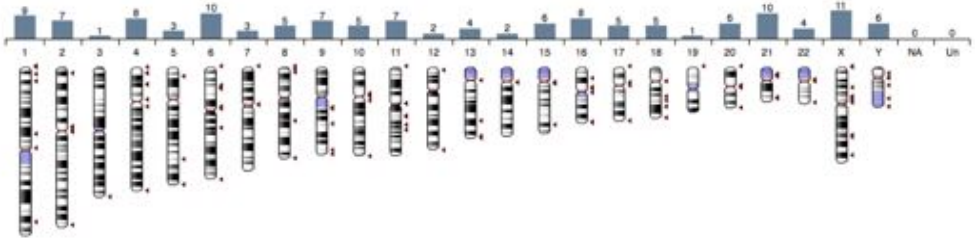
Are we there yet?

- **First complete *de novo* assemblies**
  - 2012: Bacteria (10^6 bp)
  - 2014: Yeast (10^7 bp)
  - 2014: Drosophila (10^8 bp)



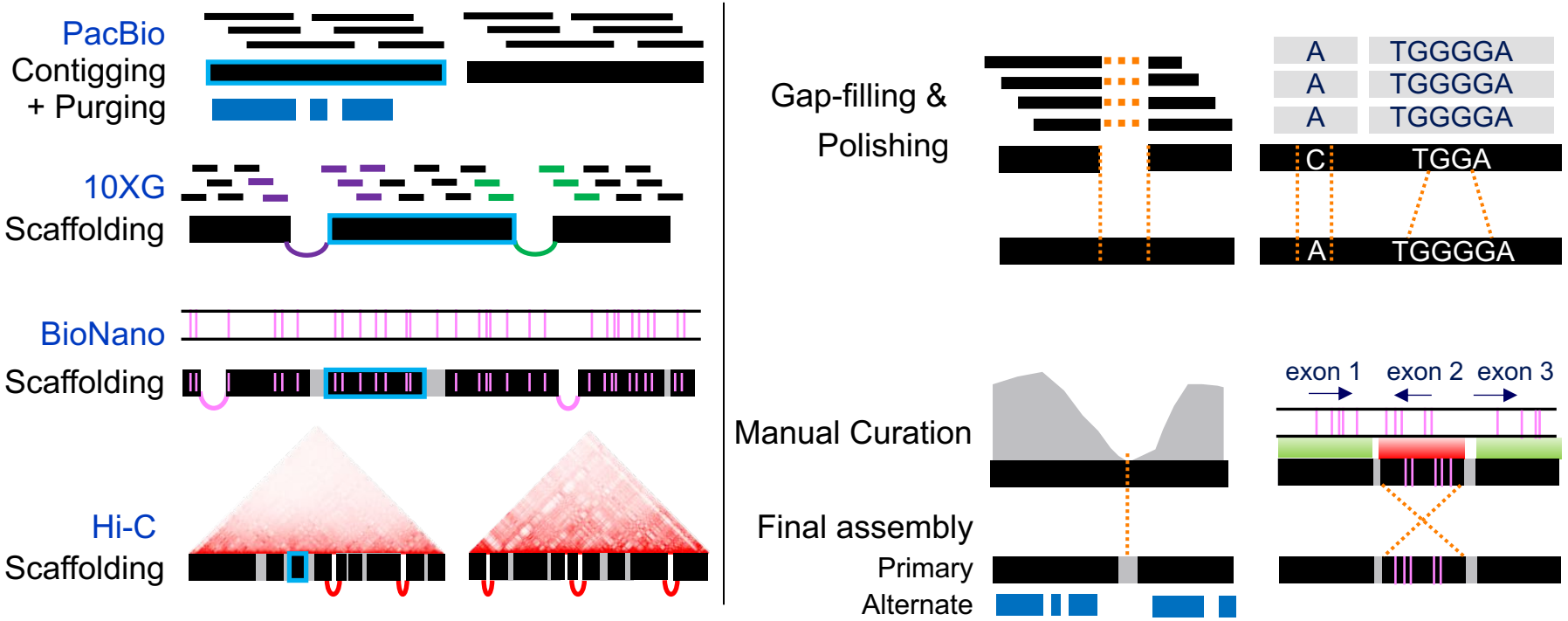
Are we there yet?

- **First complete *de novo* assemblies**
  - 2012: Bacteria (10^6 bp)
  - 2014: Yeast (10^7 bp)
  - 2014: Drosophila (10^8 bp)
  - ????: Human (10^9 bp)



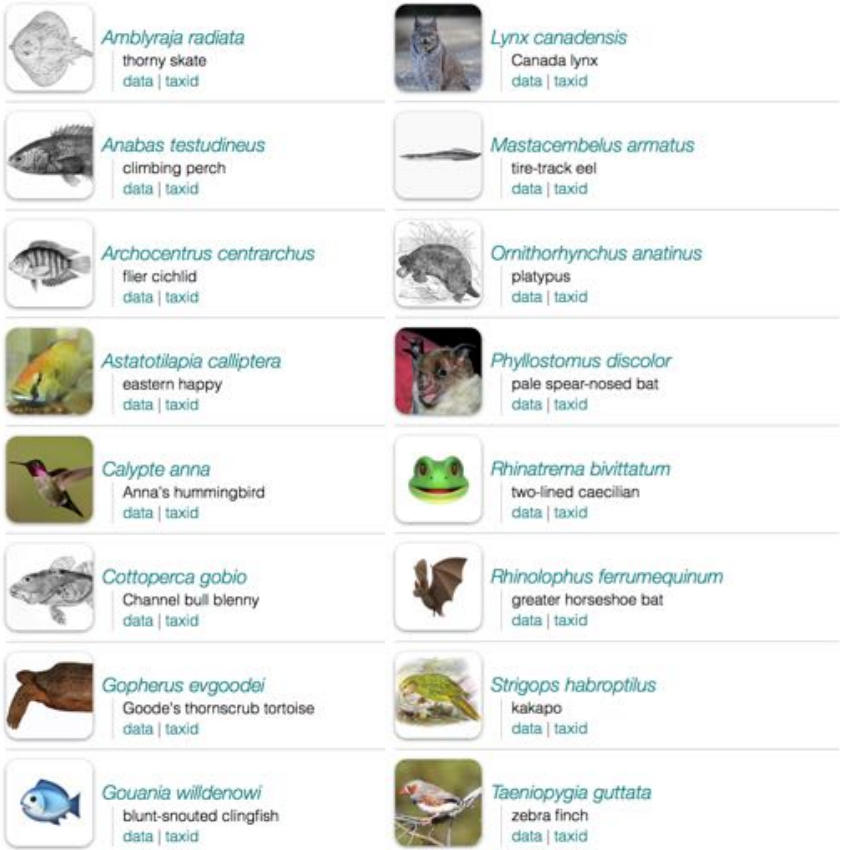
NIH NHGRI

The Vertebrate Genomes Project Pipeline



VGP GenomeArk: 1st data release

Jennifer Vashon of Maine Department of Inland Fisheries and Wildlife, left, and UMass lynx team coordinator, Tanya Lama, with an adult male lynx from northern Maine whose DNA was used to create first-ever whole genome for the species. The lynx has since been released to the wild. (MassWildlife photo / Bill Byrne)

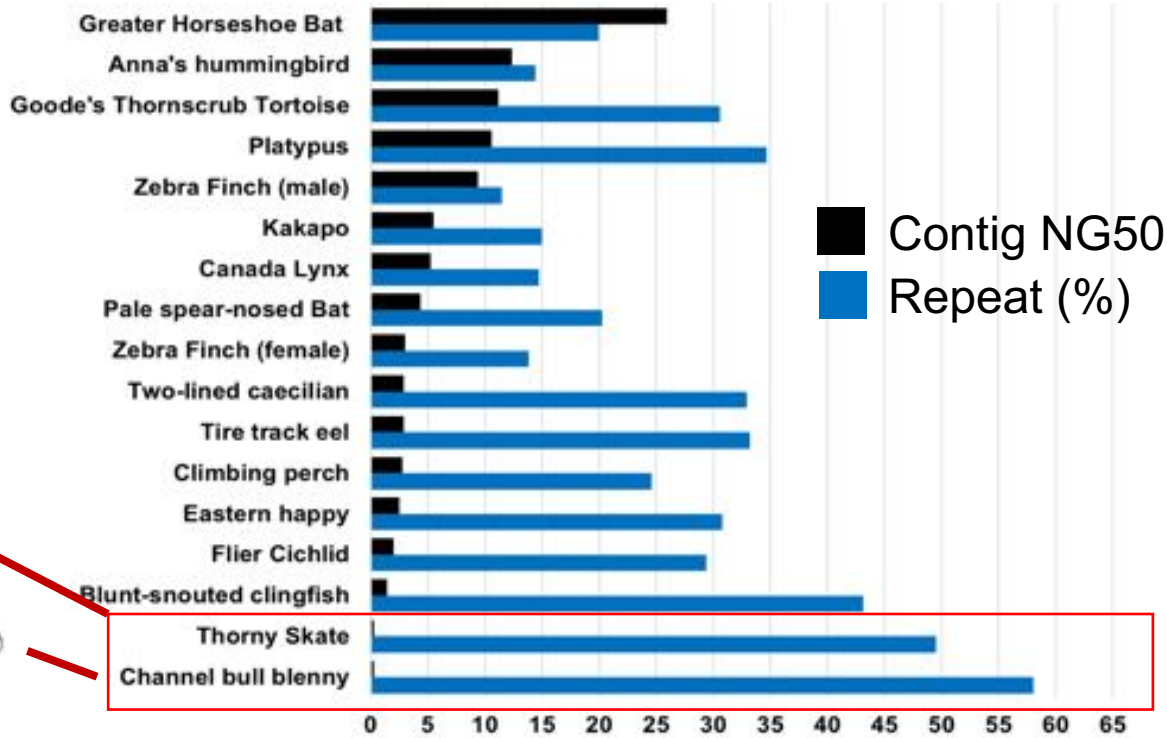


NIH NHGRI

https://VertebrateGenomesProject.org

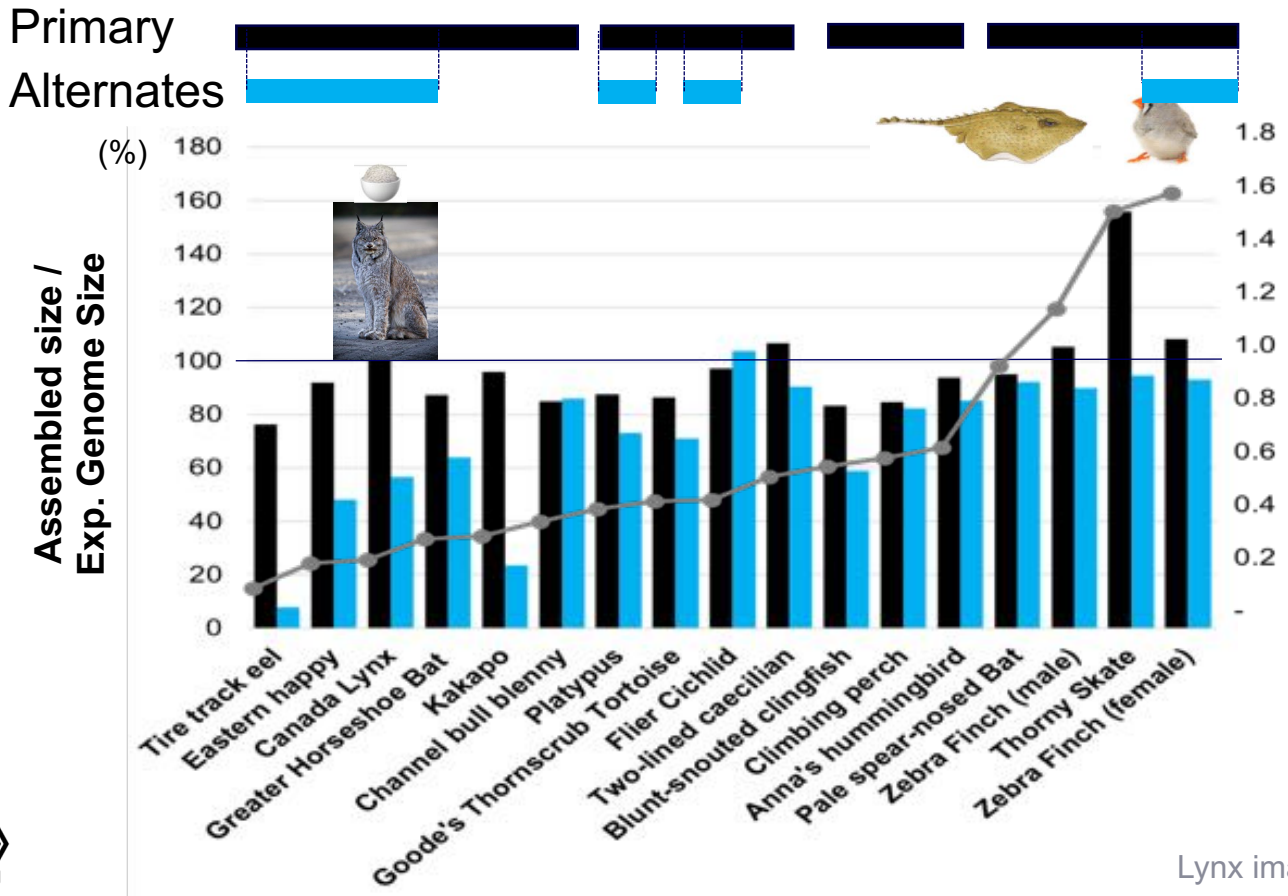
Not all genomes are equal

Contigs



NIH NHGRI

Heterozygosity causes allelic duplication



1.6% Image credit: wikimedia

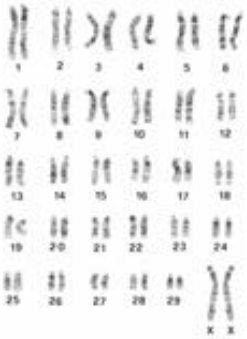
Falcon-unzip	Primary	Alt
Size (Gbp)	1.95	0.73
NG50 (Mbp)	2.6	0.02
BUSCO	94.2%	40.6%
Duplication	20.8%	3.4%

Lynx image credit: Jacob W. Frank

NIH NHGRI

Resolving haplotypes

The *genomes* assembly problem



Esperanza

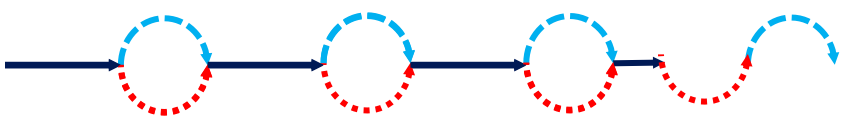
Duke, highland sire

Molly, yak dam

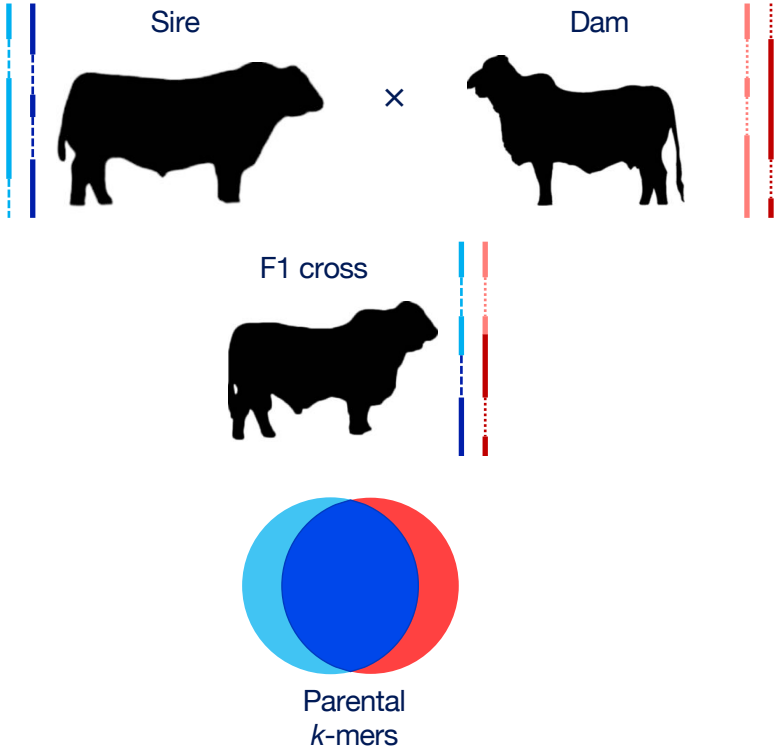
~1% heterozygosity

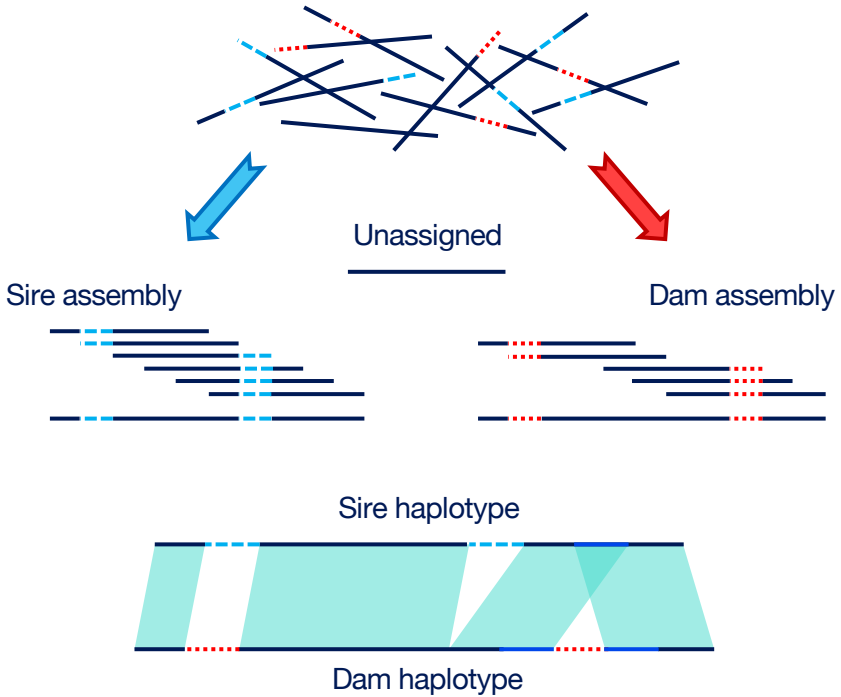
NIH NHGRI

State of the art: pseudo-haplotype



Trio binning with TrioCanu



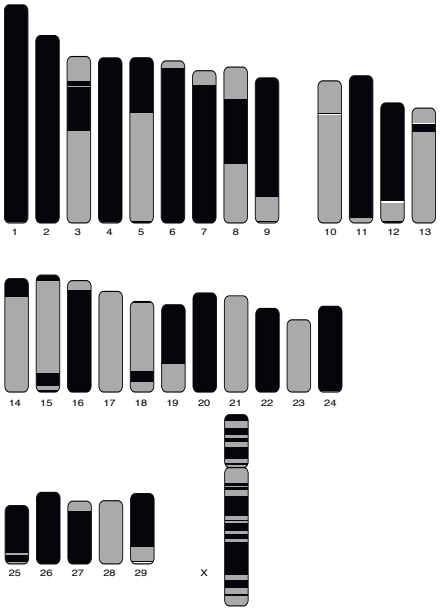


NIH NHGRI

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

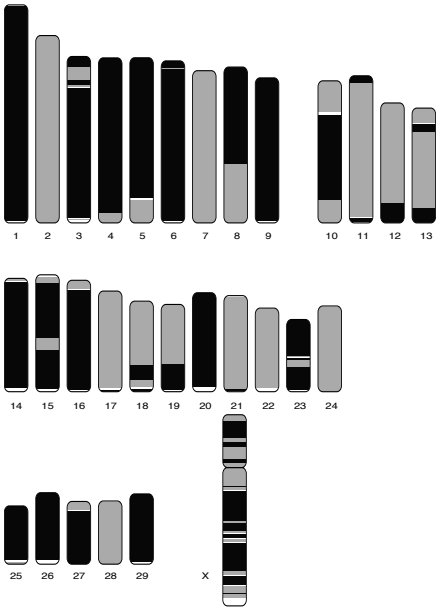
Esperanza: The nearly perfect diploid

Sire (Highland)



Esperanza

Dam (yak)



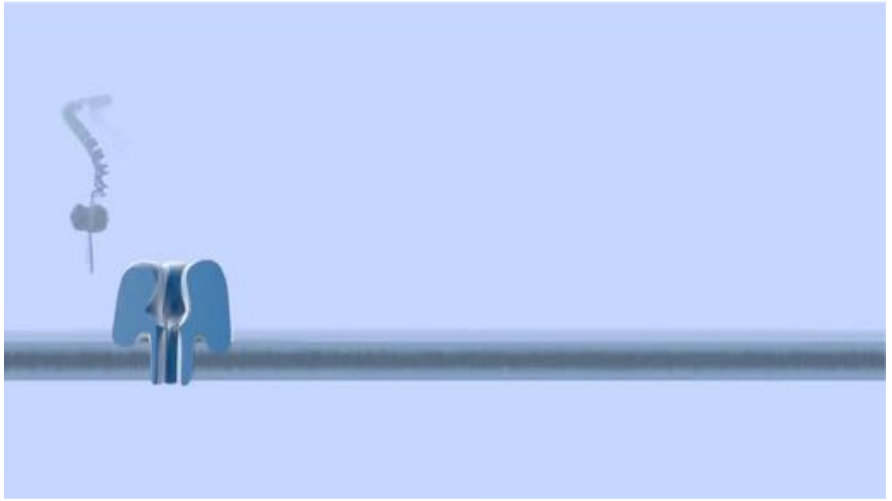
125x PacBio coverage (~60x per haplotype), **TrioCanu haplotig NG50 70 Mbp, BUSCOs 94%**

NIH NHGRI

Resolving repeats

Real-time data from a nanopore

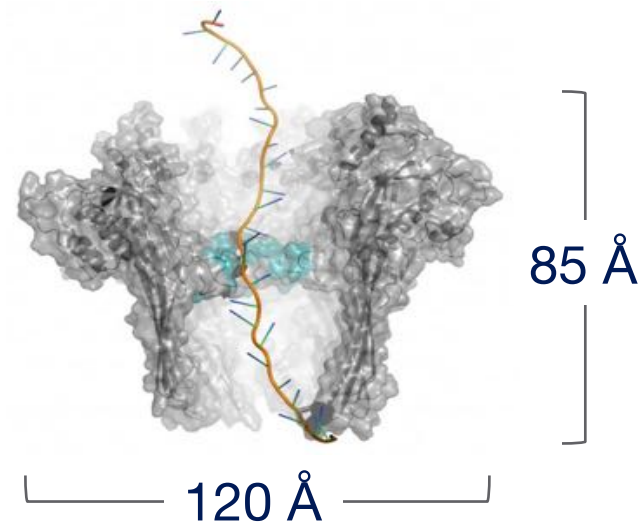
$1000 (free) instrument
$100 / bacterial genome
85–90% read accuracy

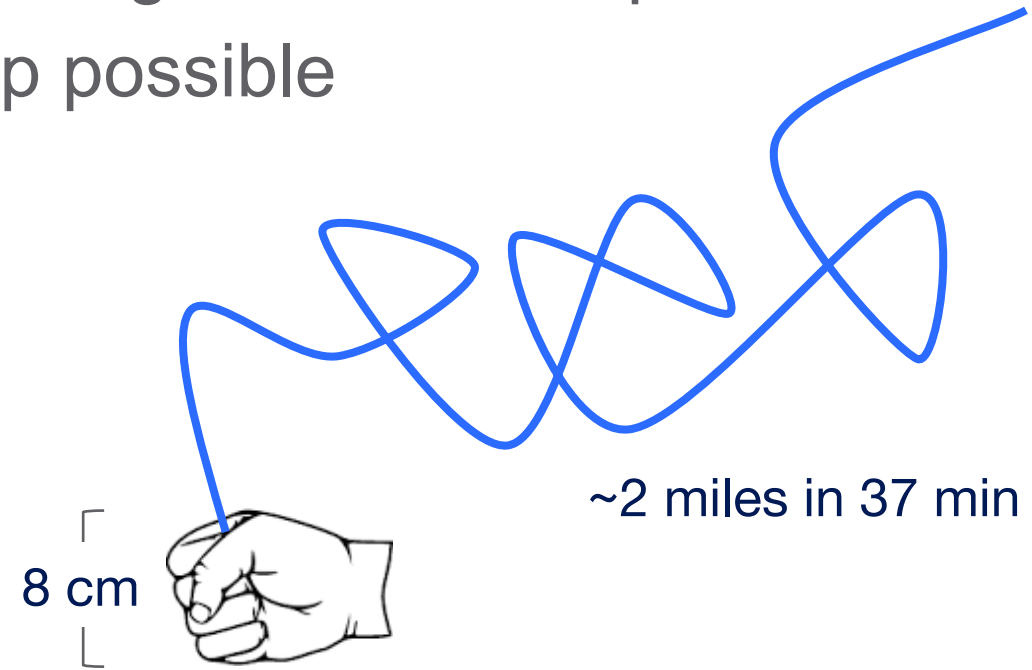


Oxford nanopore technologies

Ultra-long read nanopore sequencing

- ONT R9 pore: *E. coli* CsgG membrane protein
- Read lengths >1 Mbp possible

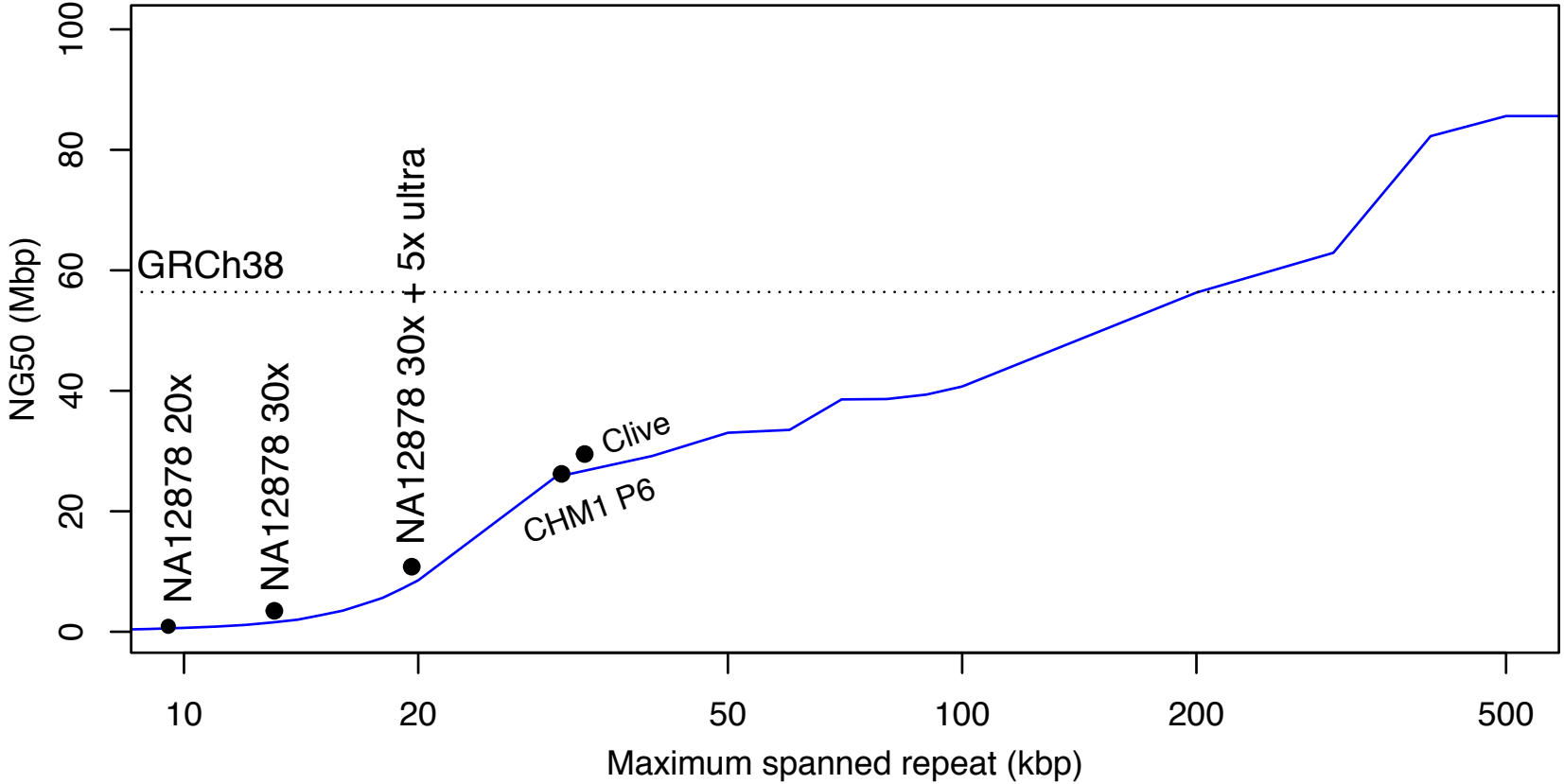




*Assuming 3.4 Å per bp, 1 Mbp = 3,400,000 Å (0.34 mm) = 40,000x height of the pore

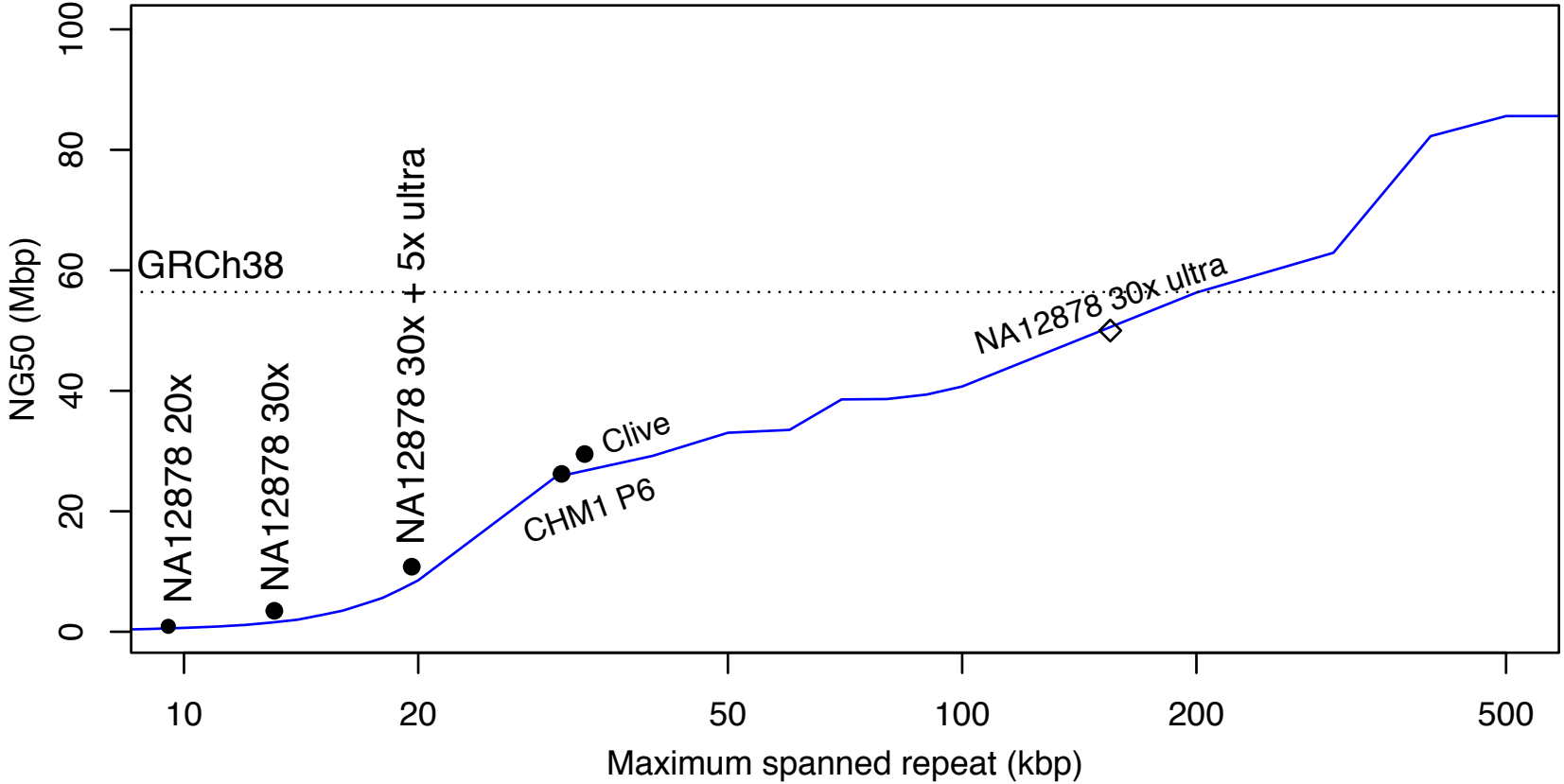
NIH NHGRI

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

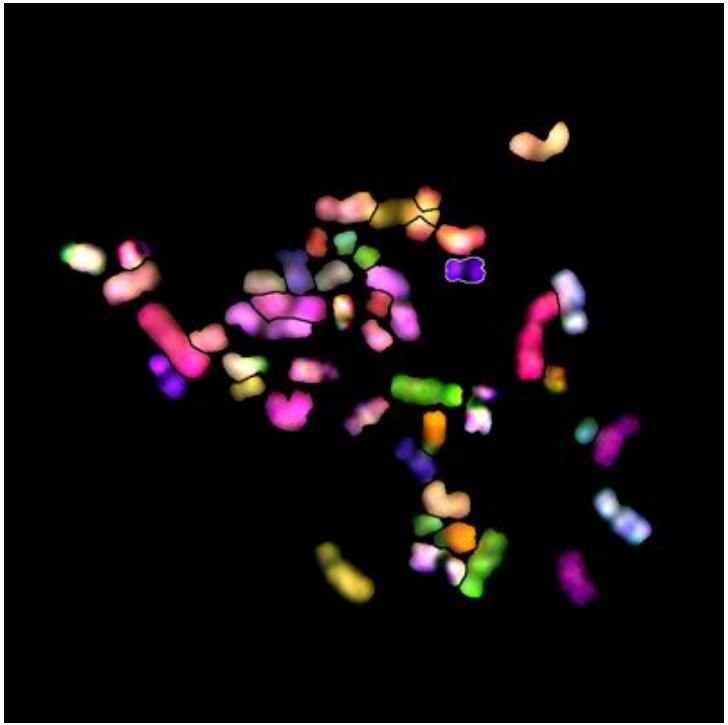


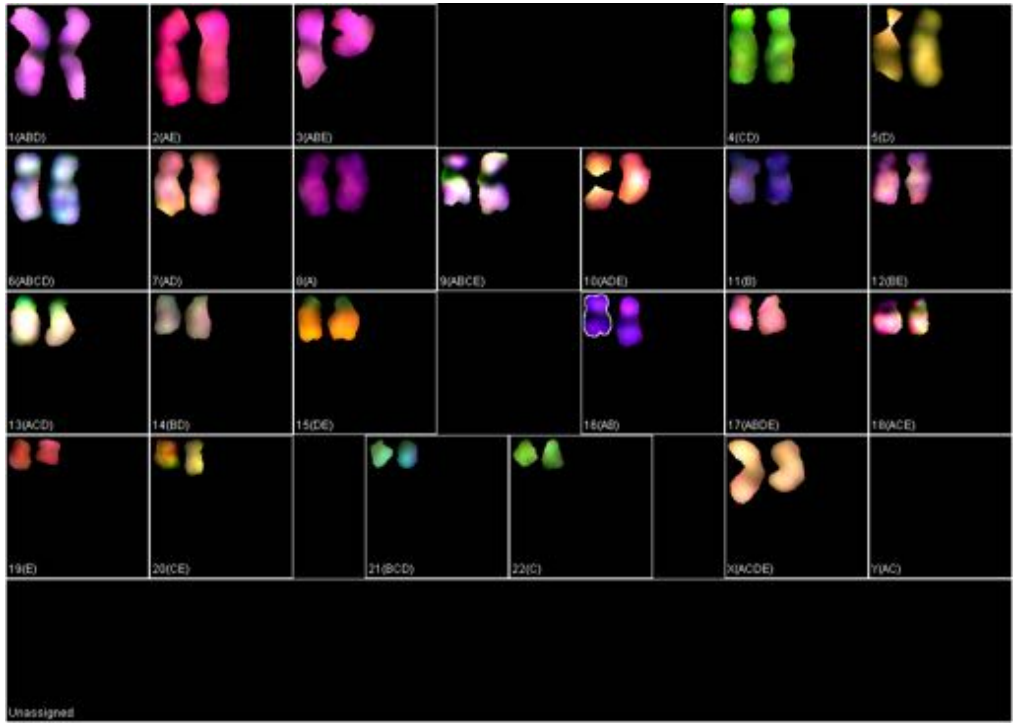
Nanopore sequencing and assembly of a human genome with ultra-long reads.
Jain, Koren, Miga, Quick, Rand, Sasani, Tyson, et al. *Nature Biotech* (2018)

The One Genome Project

- **The human reference is incomplete**
  - 368 unresolved issues, 102 gaps
  - Segmental duplications, rDNAs
  - Centromeres, telomeres, heterochromatin
- **These gaps contain important information**
  - Missing reference sequence leads to mapping artifacts
  - Variation in these gaps is unexplored (e.g. rDNAs)
  - We don’t know what we don’t know…
- **Long-read sequencing can close these gaps**

The One Genome: CHM13hTERT





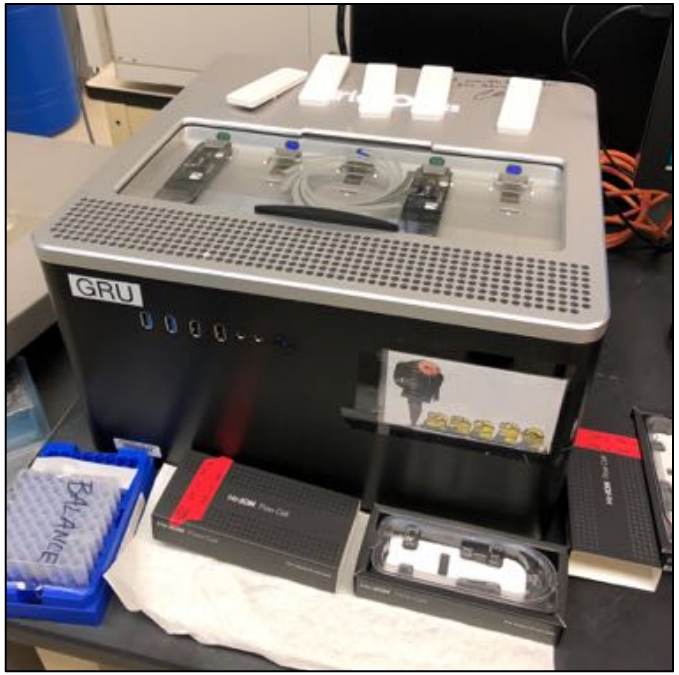
N=46; XX

Cell line from Urvashi Surti, Pitt; SKY karyotype from Jennifer Gerton, Stowers

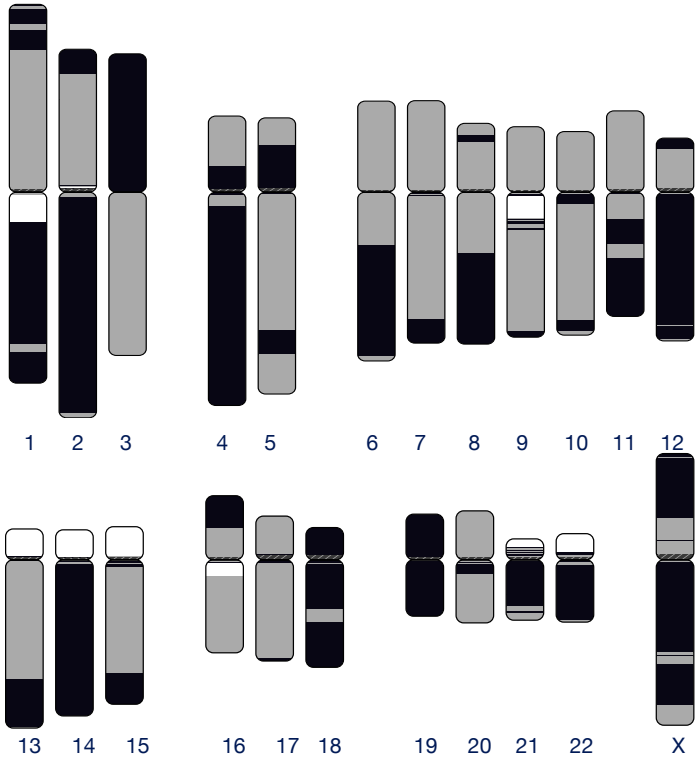
NIH NHGRI

CHM13 sequencing at NISC

- From May 1/18 – Jan 8/19
  - 94 MinION/GridION flow cells
  - 11.1M reads
  - 155 Gb (1.6 Gb / flow cell) (50x)
  - 99 Gb in reads >50 kb (32x)
  - 78 Gb in reads >70 kb (25x)
  - Max mapped read length 1.04 Mb
- Assembled with Canu



The human genome, 2017



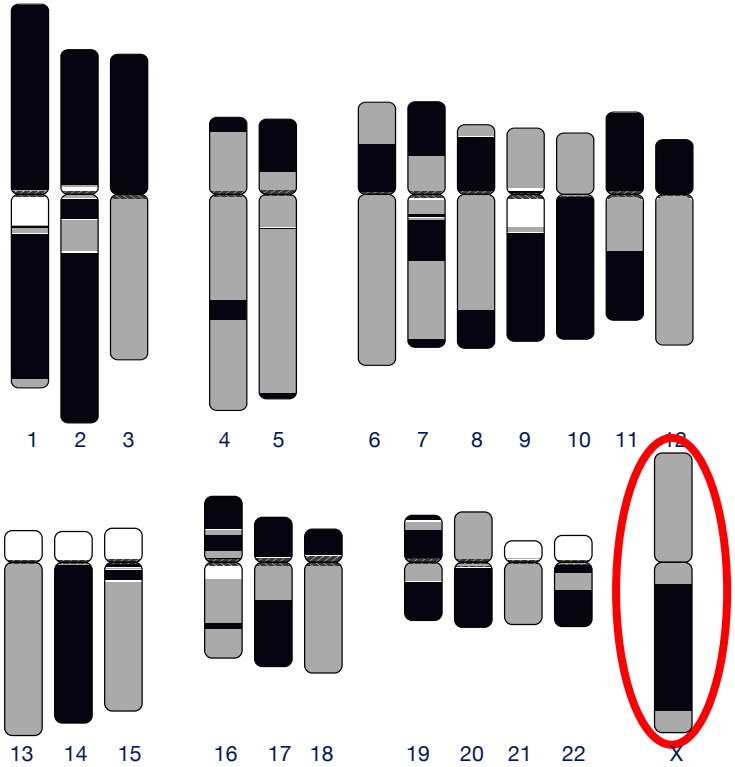
GRCh38

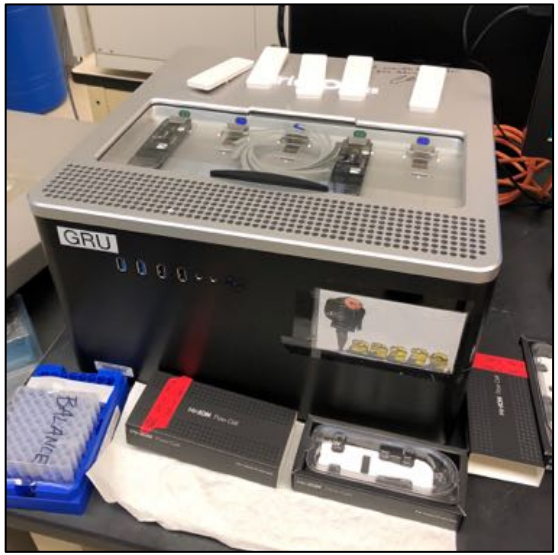
The Genome Reference Consortium consists of:

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

GRCh38 NG50 contig 56.4 Mbp

The human genome, 2018

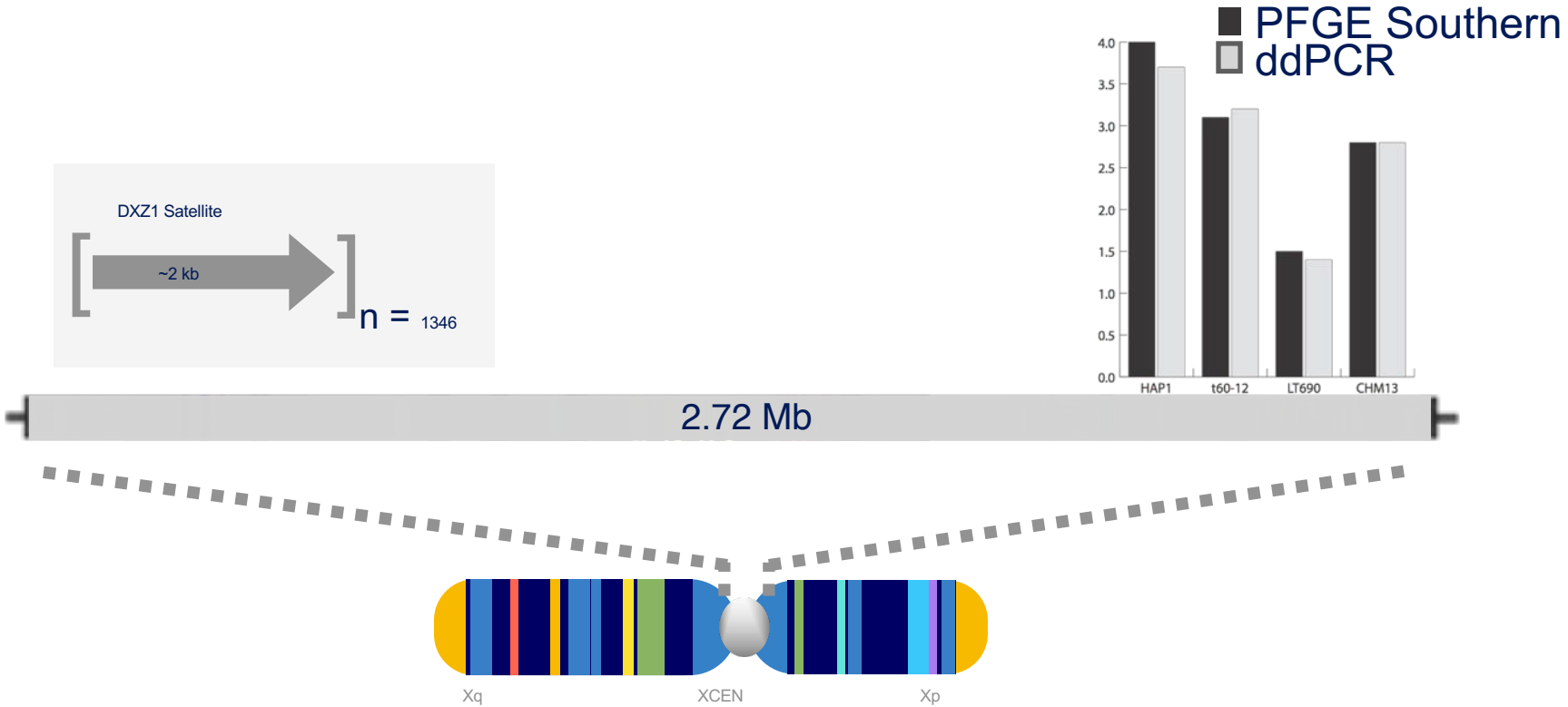




CHM13 NG50 contig 74.8 Mbp (70x PacBio + 35x UL ONT)

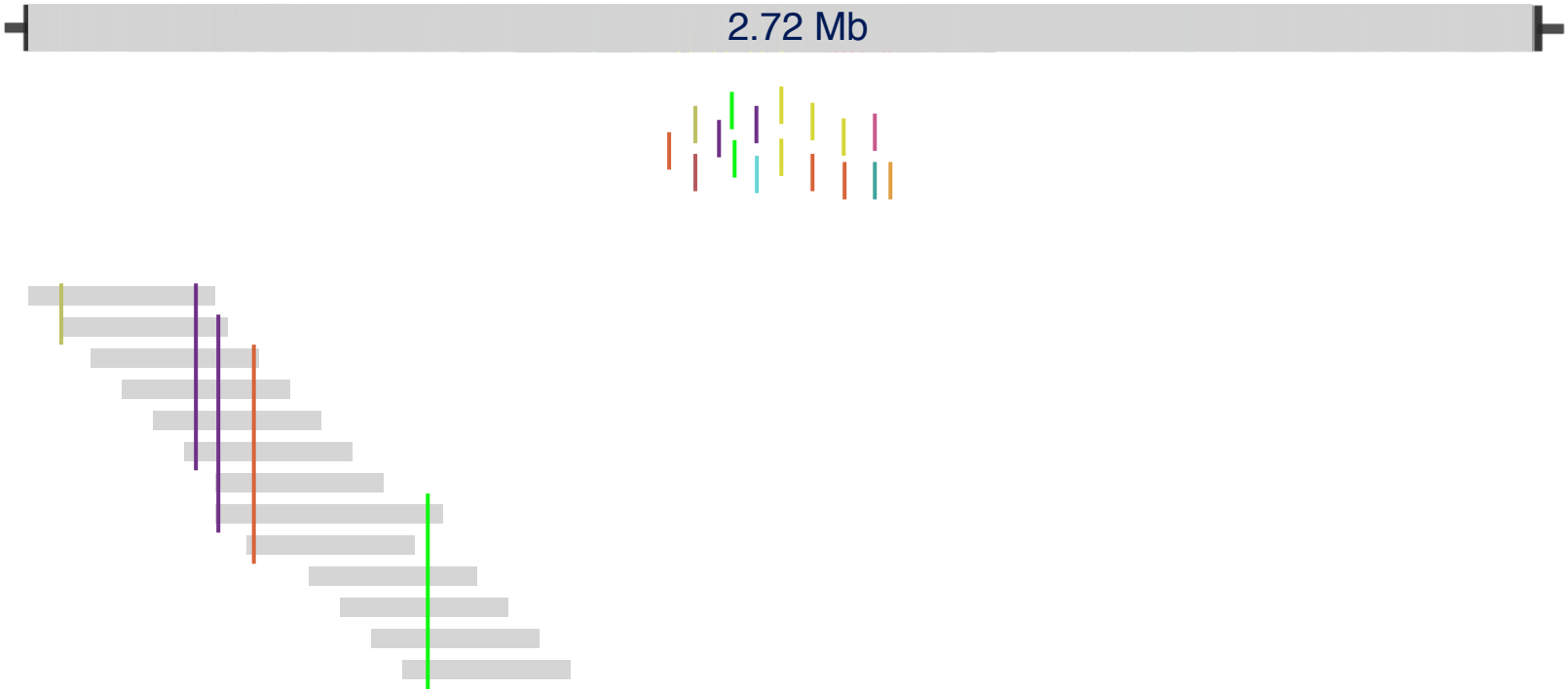
NIH NHGRI

ChrX centromere detail

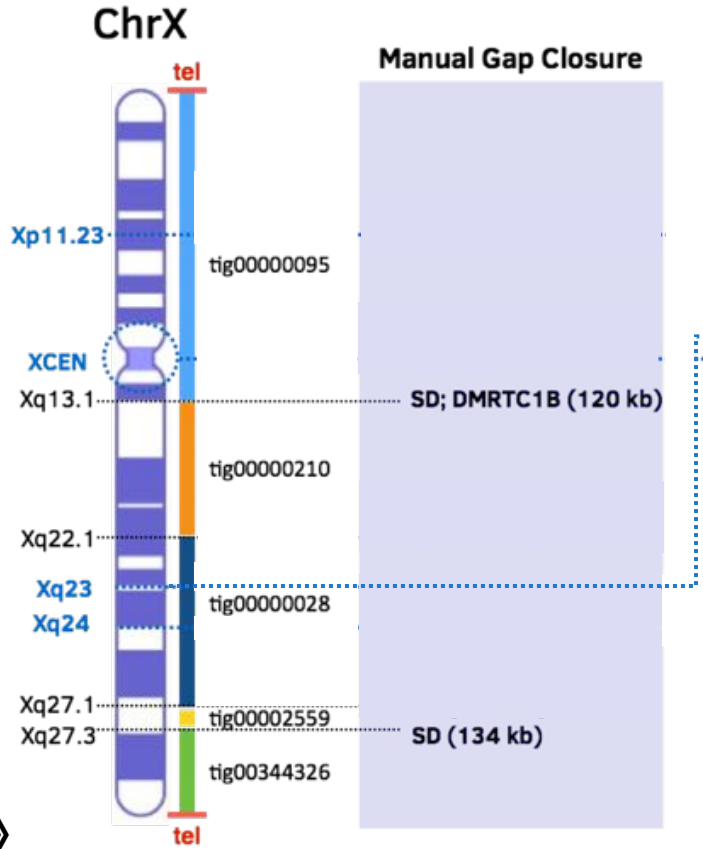


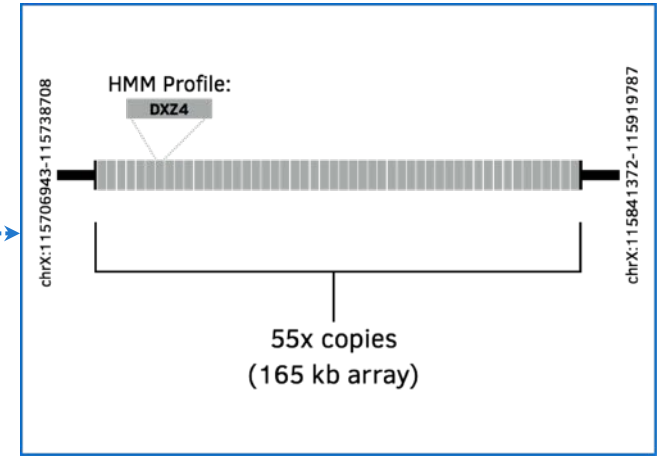
NIH
NHGRI

Repeat resolution using variants



First complete X chromosome

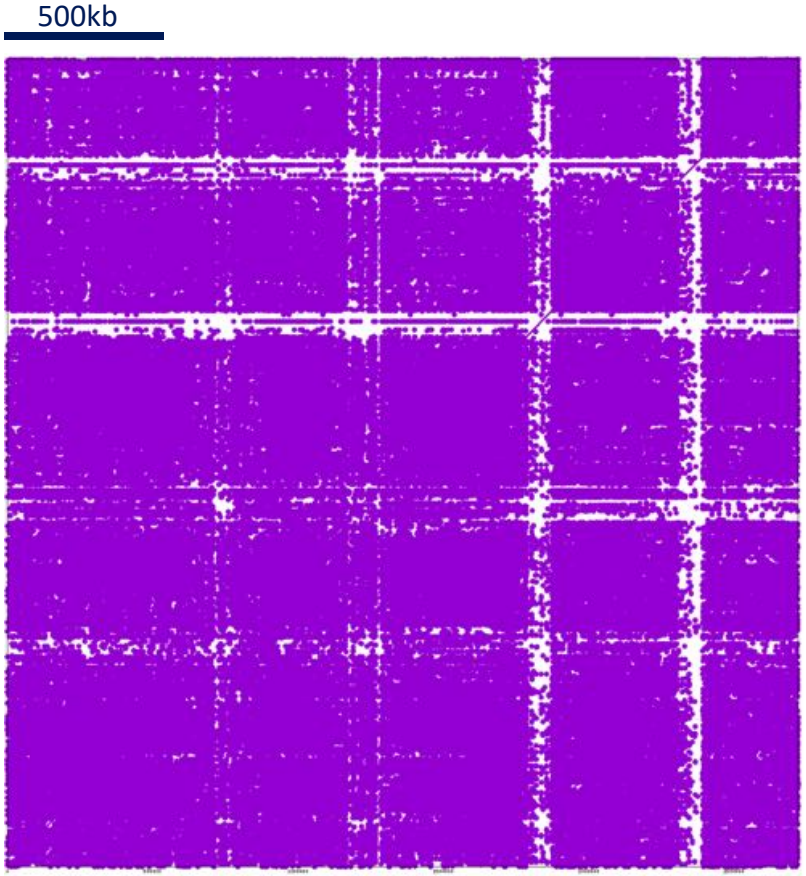




DXZ4 arrays contained between 12 and 100 3.0 kb repeat units with an average array containing 57 (Tremblay *et al* 2011)

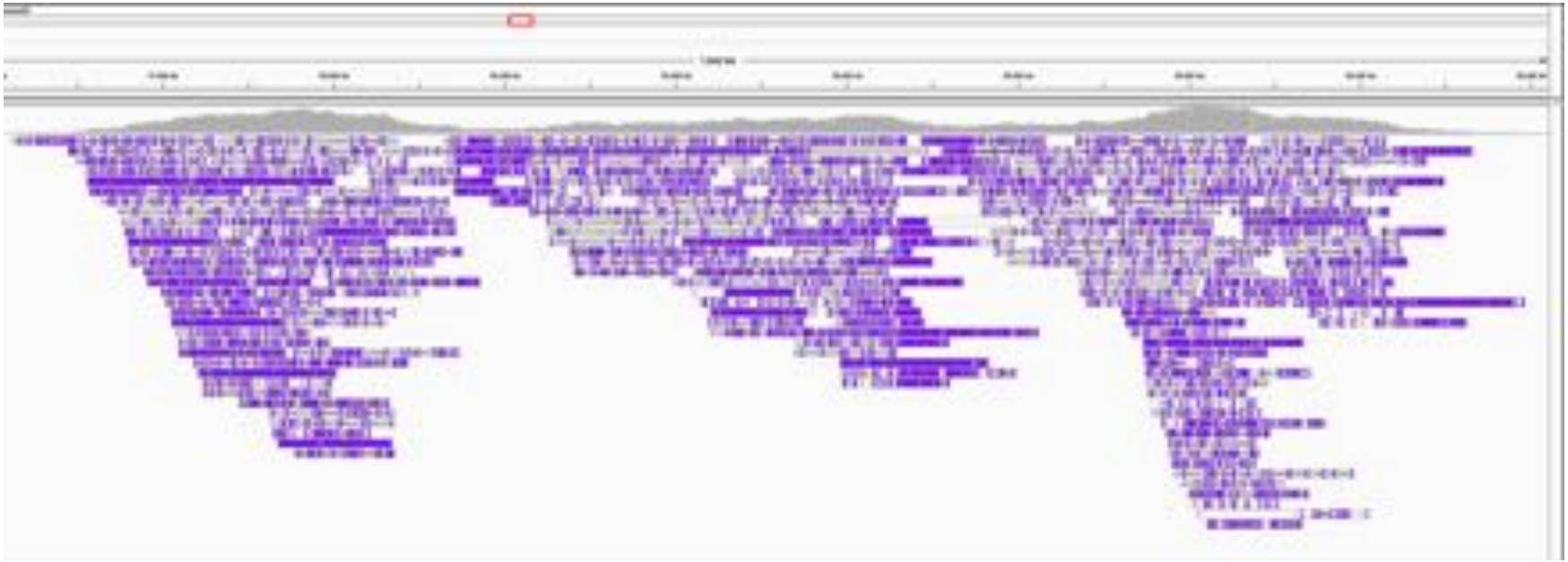
NIH NHGRI

How to validate 1300 copy 2kb repeat?

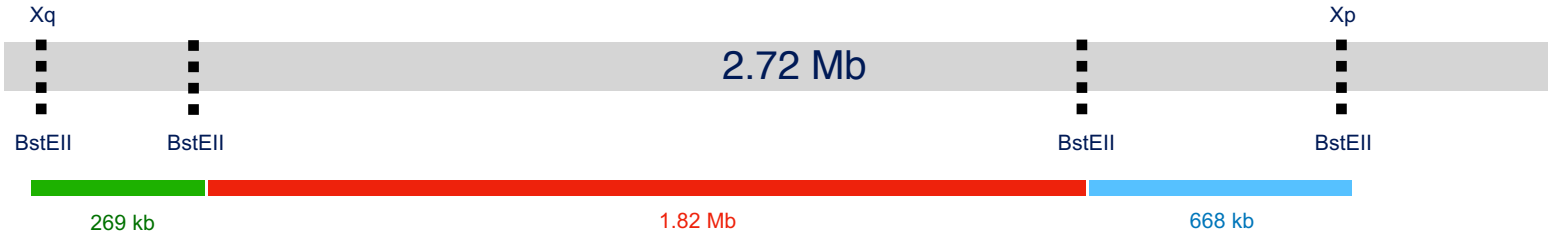


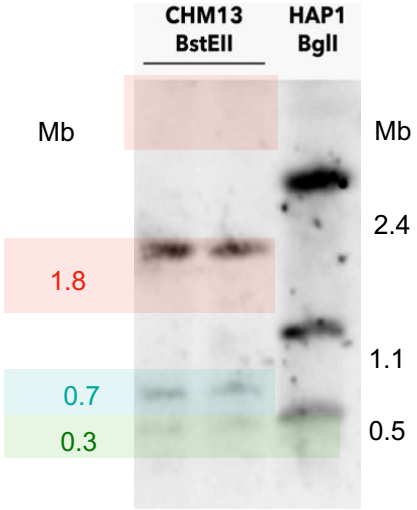
>=500bp, >=97% idy

Independent markers confirm structure



Lab validation supports structure





NIH
NHGRI

Almost there!

- **Large repeats remain challenging**
  - Satellite arrays, SegDups, rDNAs, …
- **Haploid assembly is solved by long reads**
  - Complete human reference in 1–2 years
- **Diploid assembly is solved by trios/Hi-C/ss**
  - Complete *haplotypes* will become the new norm
- **Enabling reference genomes for all species**
  - Genome10k and the Vertebrate Genomes Project

The next 20 years

NIH
NHGRI

Bioinformatics is currently artesinal

Scaling up, training volunteers

Marcela Uliano

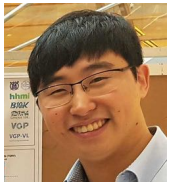
Leibniz Institute for Zoo and Wildlife Research
IN THE FORSCHUNGSVERBUND BERLIN E.V.

BeGenDiv
Berlin Center for Genomics in Biodiversity Research

Giulio Formenti

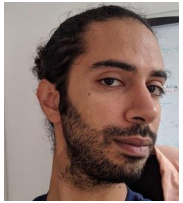
Simona Secomandi

Univ. of Milan



Chul Lee

Seoul Nat. Univ.



Maxmilian Driller

Calvinna Caswara

Majid Vafadar

Freie Univ. Berlin

DNAnexus

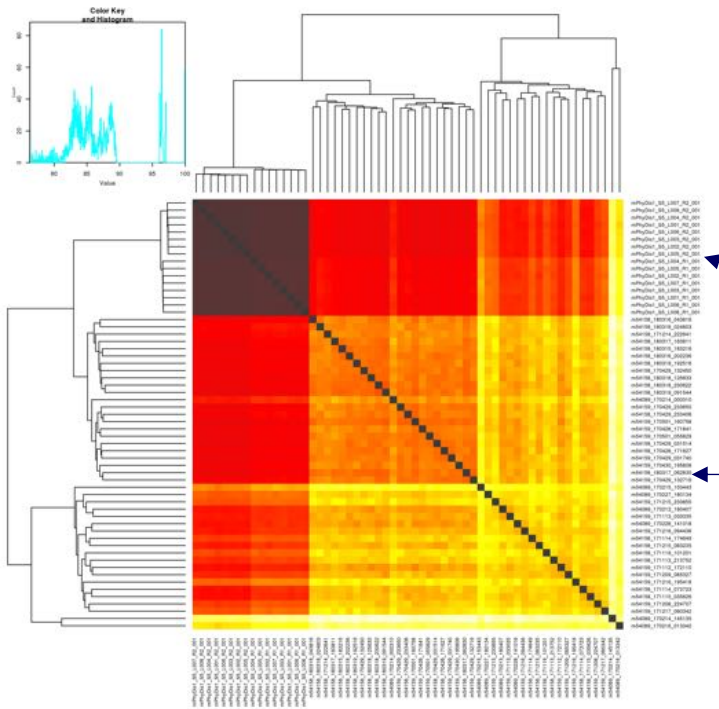
JOHNS HOPKINS UNIVERSITY

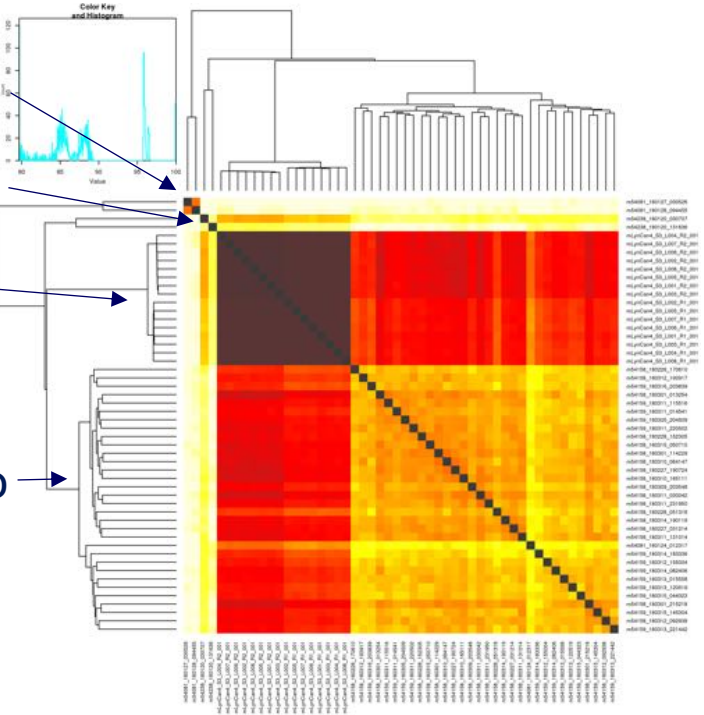
Chai Fungtammasan

Nicholas Hill

NIH NHGRI

Automating sequence QC





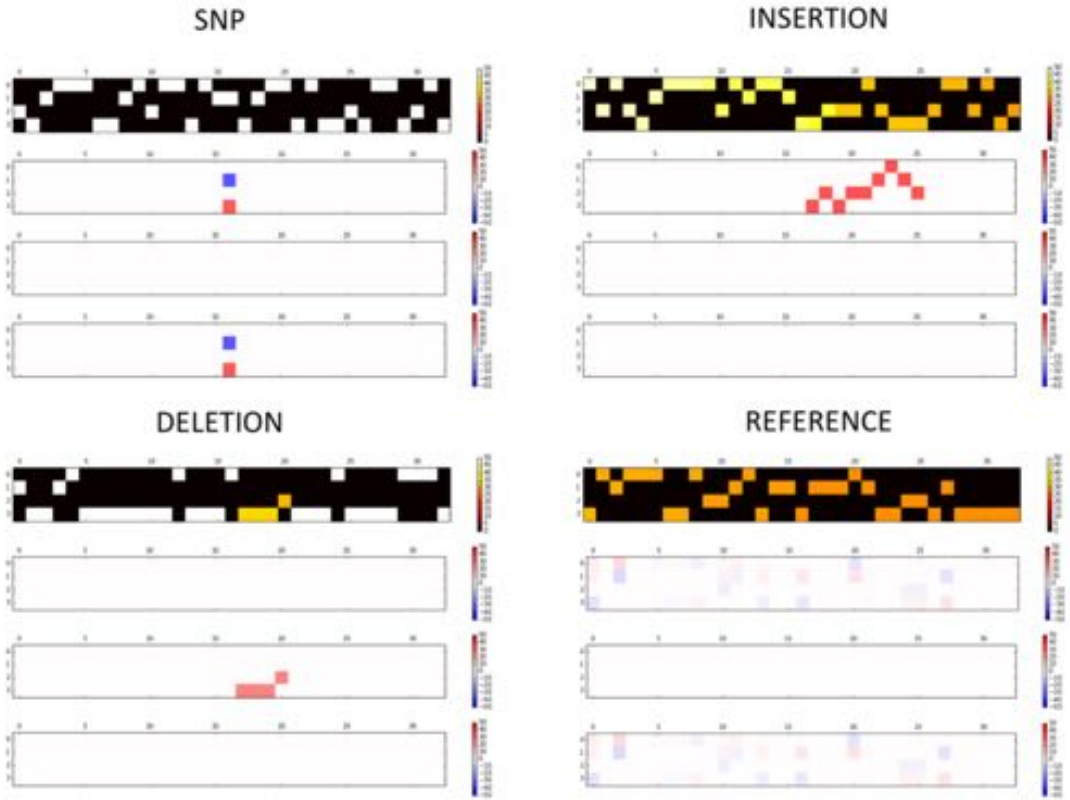
Example of no contamination (mPhyDis1)

Rice SMRT cells in Canadian Lynx

NIH NHGRI

Mash: fast genome and metagenome distance estimation using MinHash.
Ondov et al. 2016

Automating assembly QC?



Luo et al. Clairvoyante: a multi-task convolutional deep neural network for variant calling in Single Molecule Sequencing. 2018

First trainee results

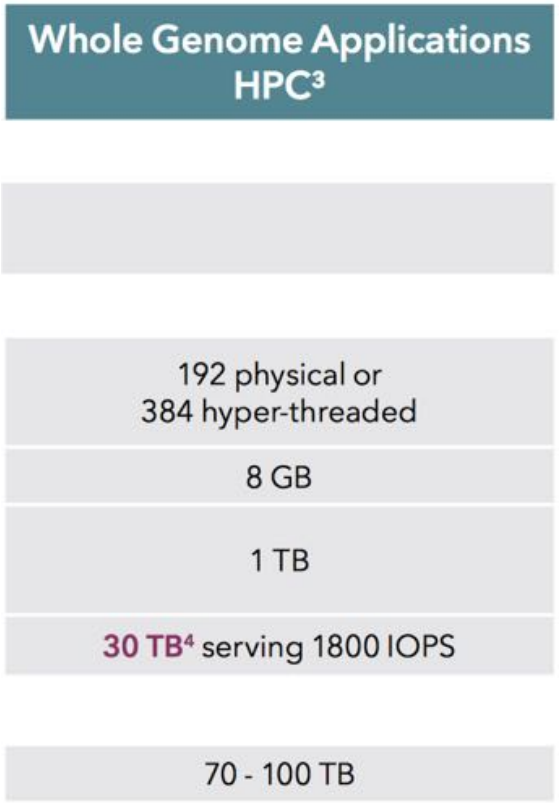
Carmine Bee-eater

Total bp	1,182.18 Mbp
Number of contigs	723
NG50	10.27 Mbp
Max contig size	41.28 Mbp

Common brushtail possum

Total bp	3,450.80 Mbp
Number of contigs	3,124
NG50	4.60 Mbp
Max contig size	35.90 Mbp

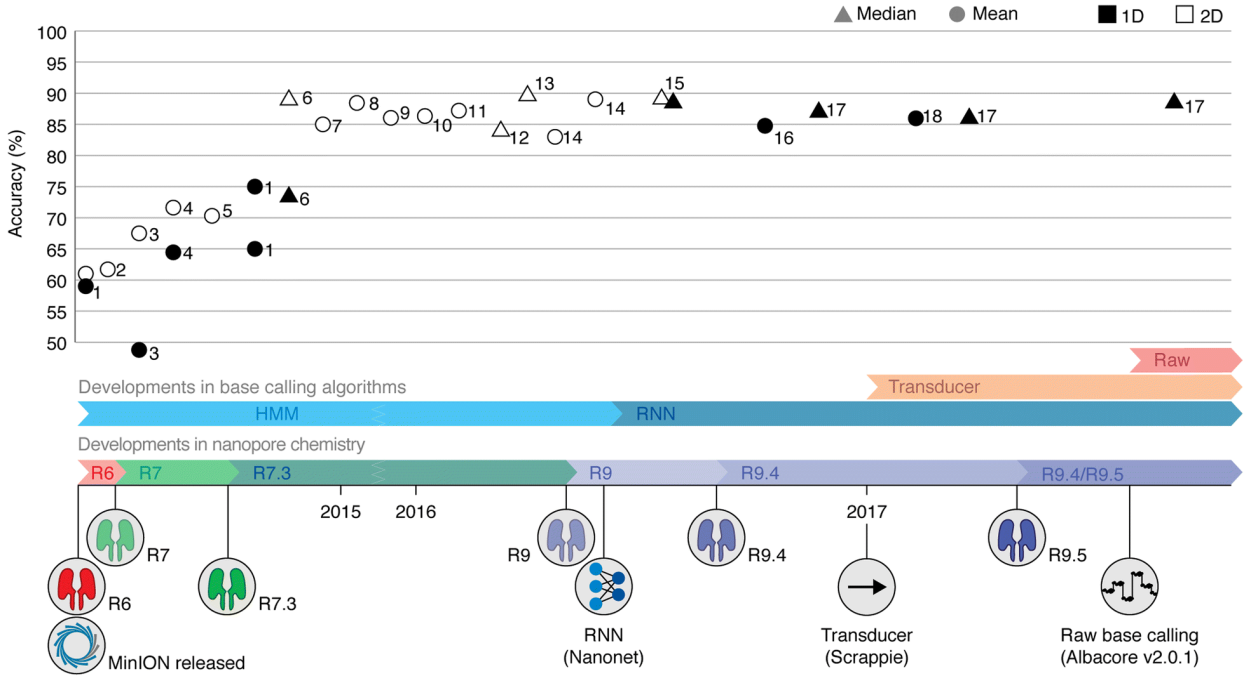
Evolving compute requirements

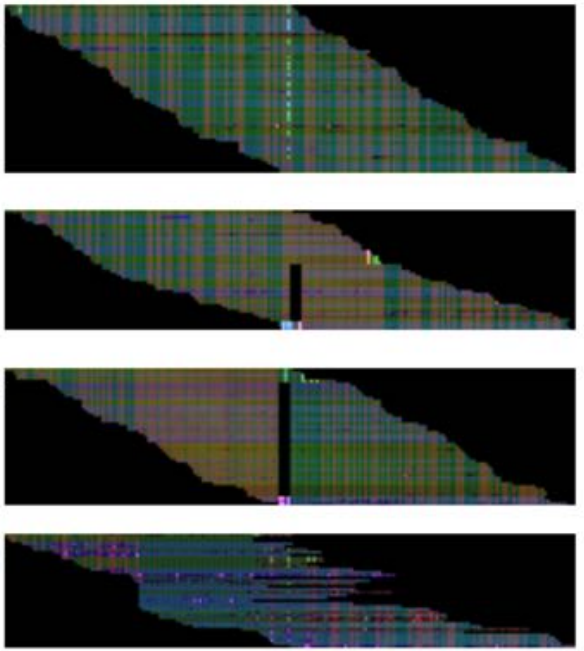


These are biowulf norm nodes!

https://www.pacb.com/wp-content/uploads/SMRT-Analysis-Brochure-Gain-a-deeper-understanding-of-your-sequencing-data.pdf

ML gaining popularity



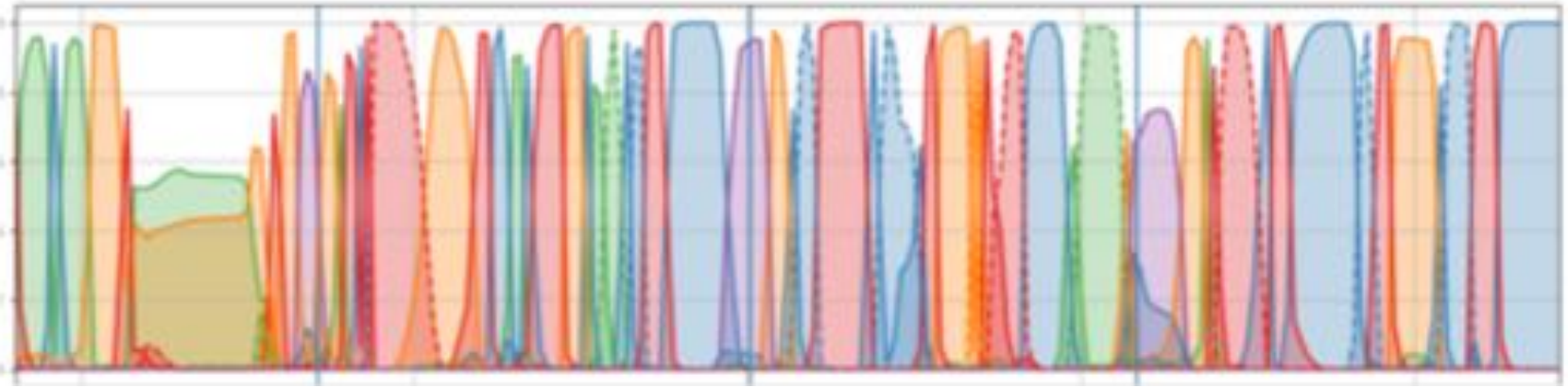


From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Rang et al. 2018

A universal SNP and small-indel variant caller using deep neural networks. Poplin et al. 2018

NIH
NHGRI

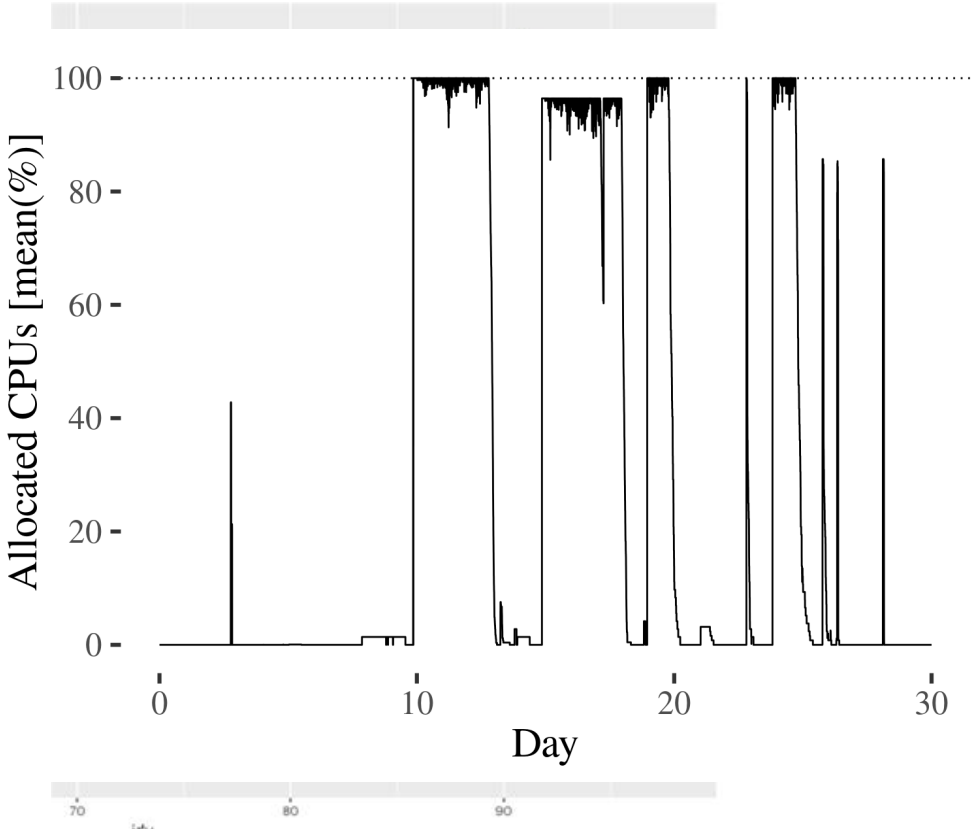
Yet another nanopore basecaller: flappie



- Runtime for 2.5k reads = 15.14h (8 cores)
  - 11m reads / 2.5k = 1/4000th of total
  - 15.14 * 8 * 4000 = 515,282 CPU h!
  - Reservation!

Tools changed too fast

- New GPU basecaller released before reservation
  - Same job took 0.03 h
  - 0.03 * 1 * 4000 = 130 CPU h
  - Finished in < 3 hrs
- Used reservation for other analysis but under-utilized
- But at least it is better



Democratized sequencing is here

- $750 40x human genome*
  - 15–90m sample prep
  - 1m–48h runtime
  - 0.1Tb per hour (6Tb in 48h)
  - 1.44GB/s raw data stream
- >100kb reads possible
- Pay for runs, not instrument
  - “Sequence until”
  - “Read until”
- Primary data to retain?
  - 10tb/genome @ 10k genomes
    - 100 petabytes
      - SRA = 5 petabytes
      - $0.5-2 M/yr on S3

https://www.nanoporetech.com

* Assuming 2,880 flowcell commitment and 120 Gbp/cell

Democratized sequencing is here

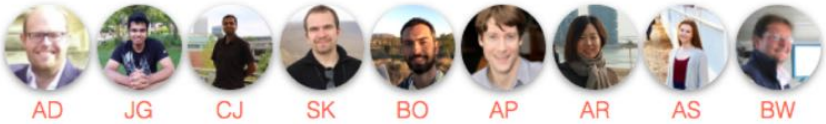
- How to democratize expertise/HPC?
- What is the role of HPC in a peer-to-peer world?
- Lots of small groups doing compute
  - No experience with “proper” HPC usage
  - Jobs that run fine on a laptop don’t translate to HPC

Acknowledgements

Karen Miga

genomeinformatics.github.io

- **Adam Phillippy**
- **Arang Rhie**
- **Brian Walenz**
- Alexander Dilthey



- 🐂🐄🐖🐐 Tim Smith
- 🐄 John Williams
- 🐄 Sarah Kingan
- 🐖 Brittney Keel
- 🐂🐄 Ben Rosen
- 🐂🐄 Derek Bickhart
- 🐐 Jessica Petersen
- 🐐 Mike Heaton
- 🐐 Ed Rice
- 👨‍🦲 Michael Hunkapiller
- 🐟 Kerstin Howe
- 🐟 Shane McCarthy
- 🦪 Erich Jarvis
- 🦪 Olivier Fedrigo
- 🚢 VGP assembly group

NIH NHGRI


The Forefront
of Genomics