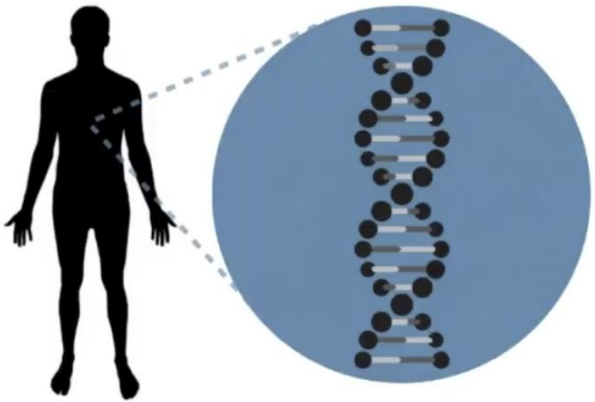
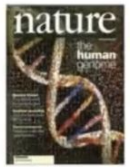
T2T genome assembly overview

Meleshko Dmitry,
June 16th, 2025

Human Reference Genome: Foundational for Genomic Research



The Human Reference Genome:

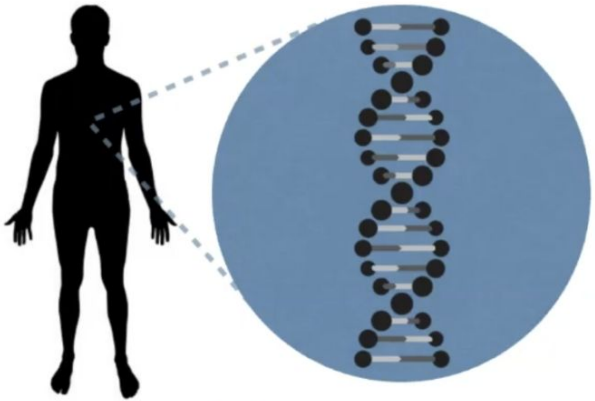


HGP, June 26, 2000
"Initial sequencing and analysis of the human genome"

Celera, June 26, 2000
"The sequence of the human genome"

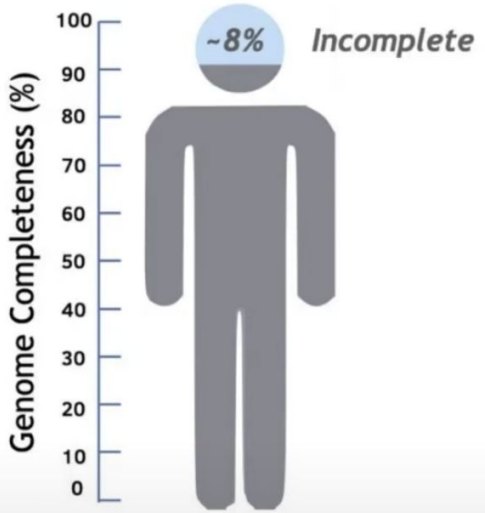
February 11, 2021
Celebrating a "***Genome revolution***"

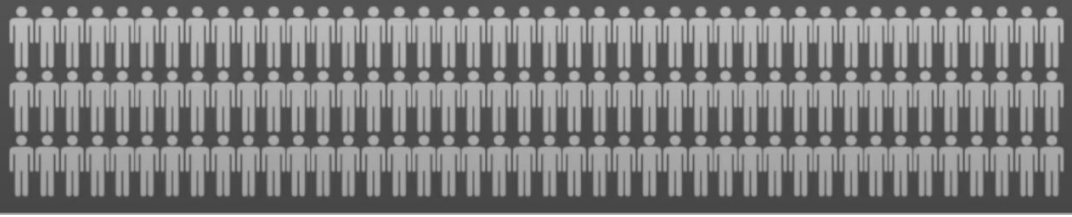
Human Reference Genome: Foundational for Genomic Research



The Human Reference Genome:

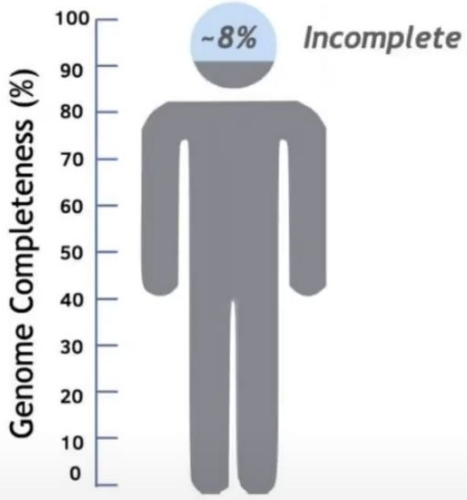
Need to generate and analyze *complete* human genomes



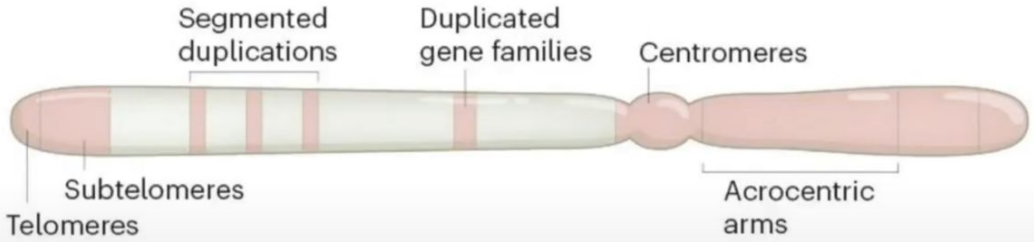


Technological Challenge:
Production of *Complete*
Reference Genomes

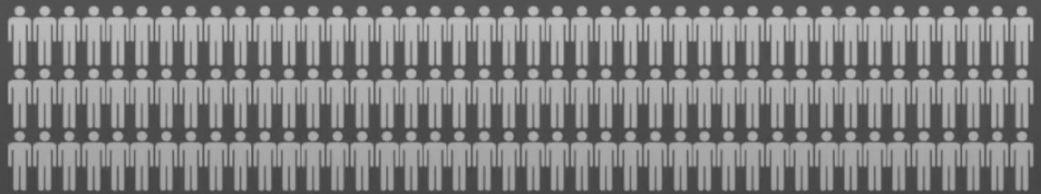
Need to generate and analyze *complete* human genomes



Human Genome Project (April 2003):
Focused Exclusively on Finishing 99% *Euchromatic* Regions.
Highly-Repetitive Heterochromatin Regions were not Included

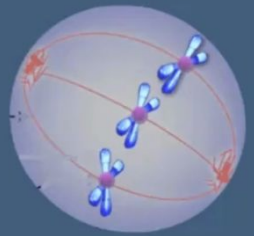


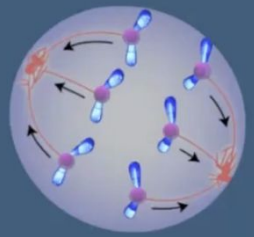
Miga, Nature 2020



Technological Challenge:
Production of *Complete* Reference Genomes

CENTROMERE FUNCTION

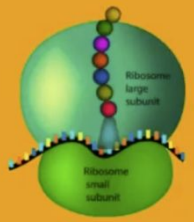




Contribute to Chromosome Cohesion

Regulate Centromere Function

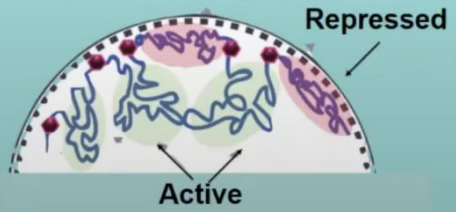
RIBOSOMES: TRANSLATION



Bringing together amino acids to form particular proteins

rDNA Arrays

GENOME SPATIAL ORGANIZATION



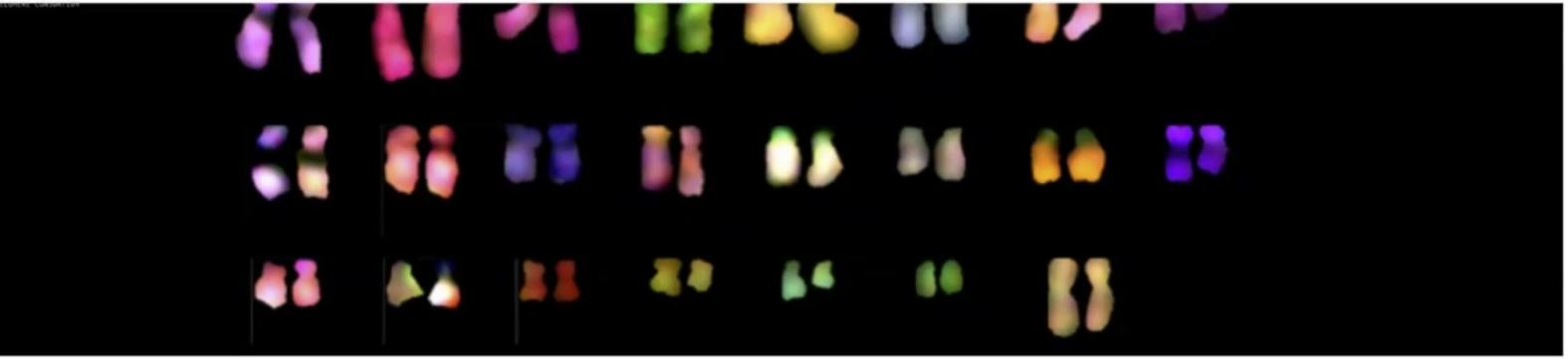
Typically Spatially distinct from active TAD compartments

GENOME INSTABILITY AND GENE FAMILIES

Large, Unresolved Intra and Inter Segmental Duplications

et's finish the human genome

T2T Working Group Home · Technology · Data · CHM13 Cell Line · Remaining Challenges · Who We Are · Join Us

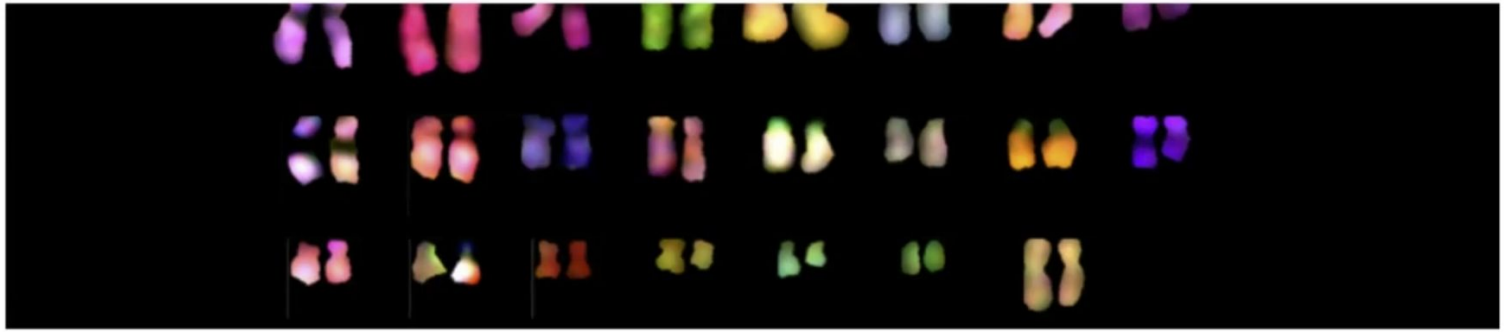


The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

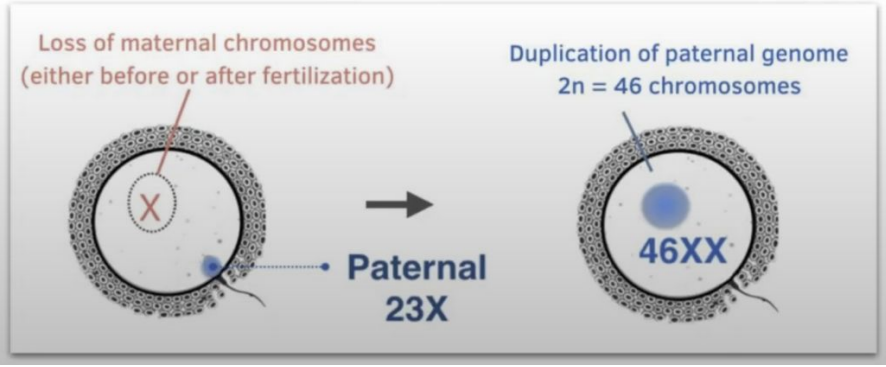
HM13 homozygous 46,XX cell line from Urvashi Surti, Pitt; SKY karyotype from Jennifer Gerton, Stowers

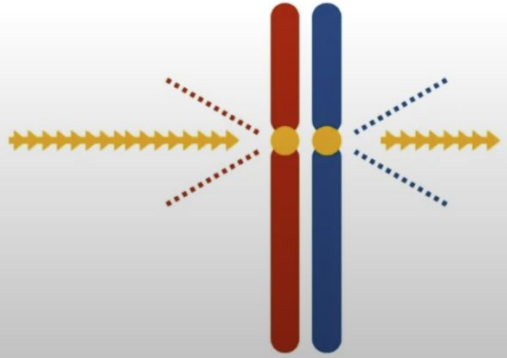
Effectively Haploid Genome

T2T Working Group Home · Technology · Data · CHM13 Cell Line · Remaining Challenges · Who We Are · Join Us Urvashi Surti

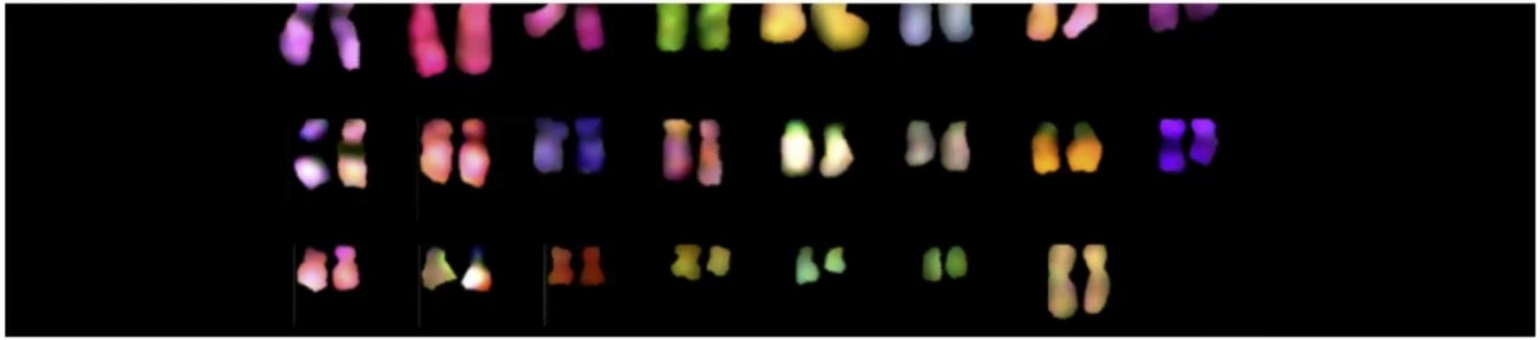


Cell Line (CHM13): Complete Hydatidiform Mole

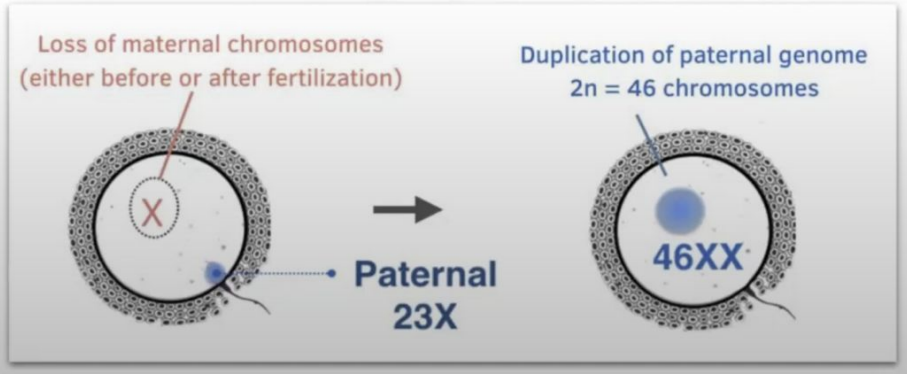


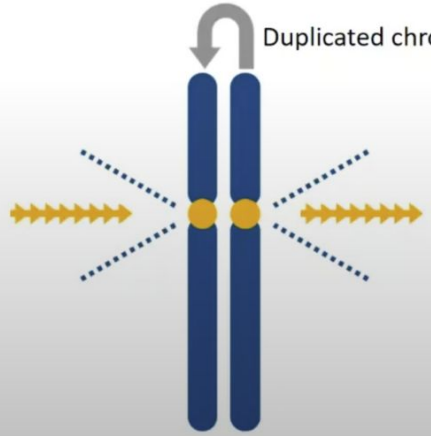


Effectively Haploid Genome



Cell Line (CHM13): Complete Hydatidiform Mole





CHM13 open data release

github.com/nanopore-wgs-consortium/chm13

- **120x** Nanopore
- **70x** PacBio CLR
- **30x** PacBio HiFi
- **50x** 10x Genomics
- **100x** PCR-free Illumina
- **35x** Arima Hi-C
- BioNano optical map
- PacBio Iso-Seq

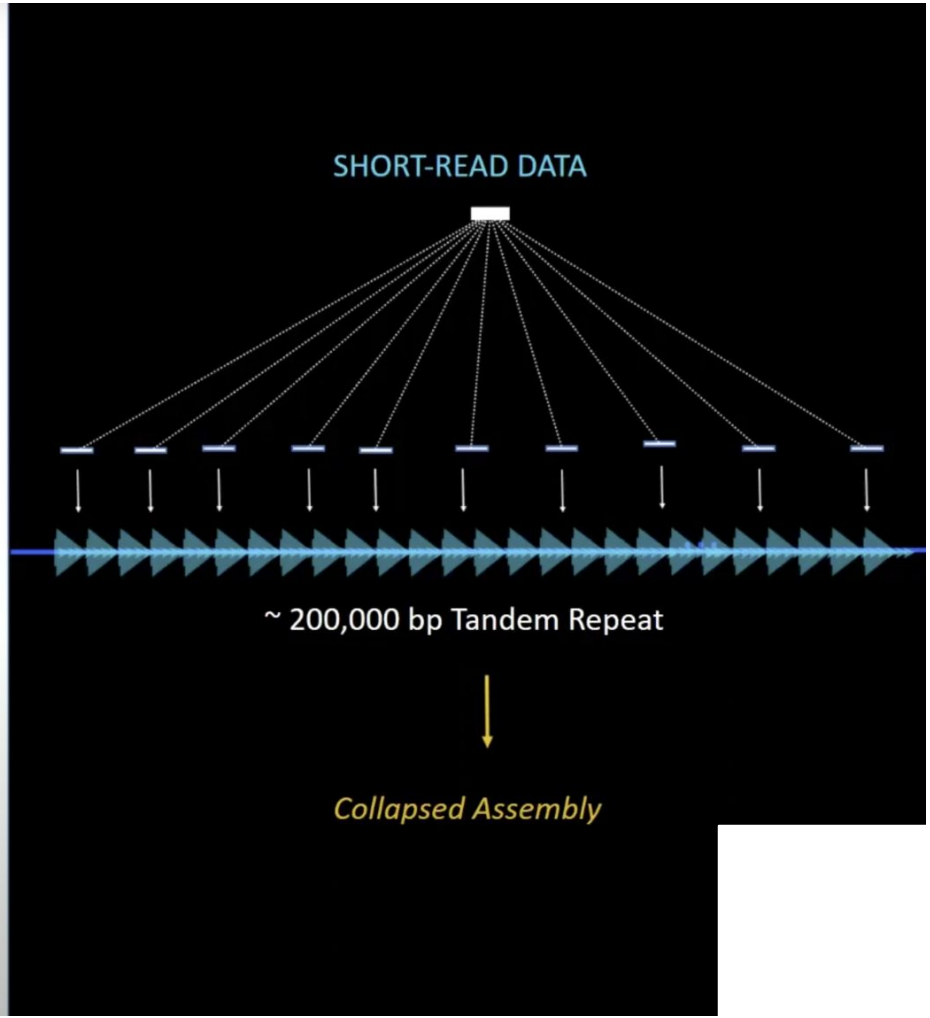
DNAnexus

amazon web services

UW, NHGRI, WashU, UC Santa Cruz, UC Davis, Nottingham, Arima Genomics

Repeat Assembly

- Short read data represent many exact copies within repeated regions equally.
- Are of insufficient length to span unique markers to inform linear ordering
- Often result in incorrect assemblies or are screened out.



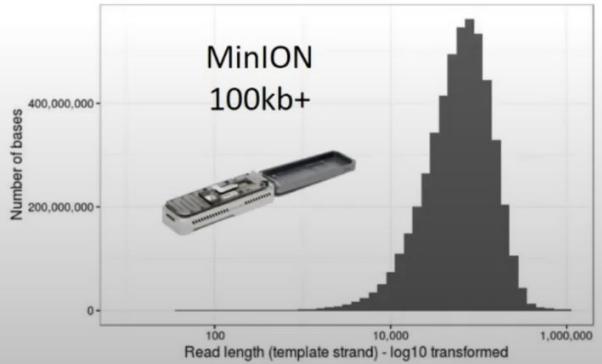
nature biotechnology

Article | OPEN | Published: 29 January 2018

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman & Matthew Loose - Show fewer authors

Nature Biotechnology **36**, 338–345 (2018) | Download Citation



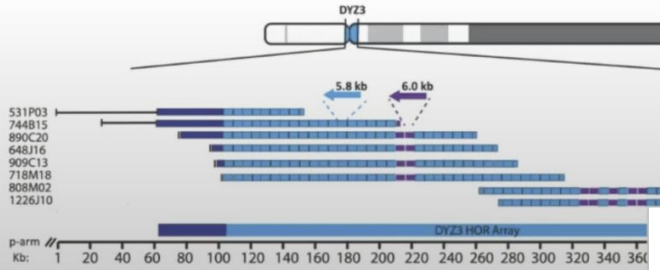
nature biotechnology

Brief Communication | OPEN | Published: 19 March 2018

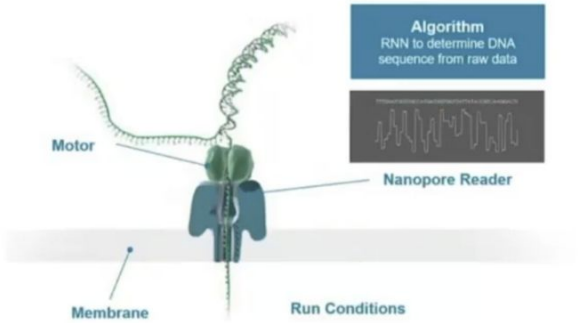
Linear assembly of a human centromere on the Y chromosome

Miten Jain, Hugh E Olsen, Daniel J Turner, David Stoddart, Kira V Bulazel, Benedict Paten, David Haussler, Huntington F Willard, Mark Akeson & Karen H Miga

Nature Biotechnology **36**, 321–323 (2018) | Download Citation



Repeat Assembly

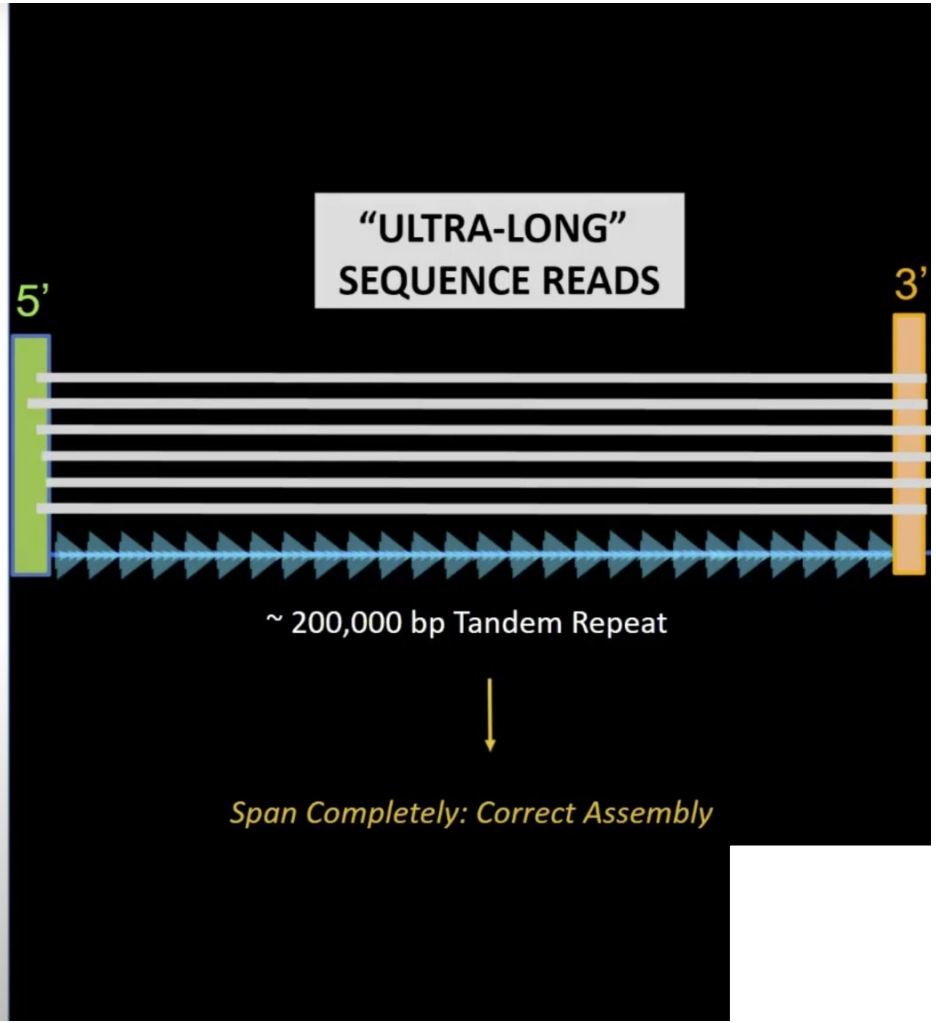


Nanopore Ultra-long (100+ kb)

Read lengths >1 Mbp are now possible

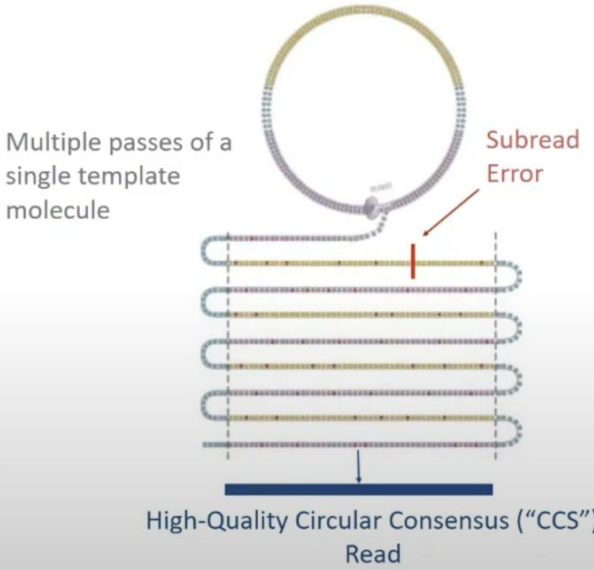
~96-98.3% single read accuracy (Recent bonito v0.35 with 'Q20' experimental models)

Nanopore sequencing and assembly of a human genome with ultra-long reads.
Jain et al. *Nature Biotechnology* (2018)

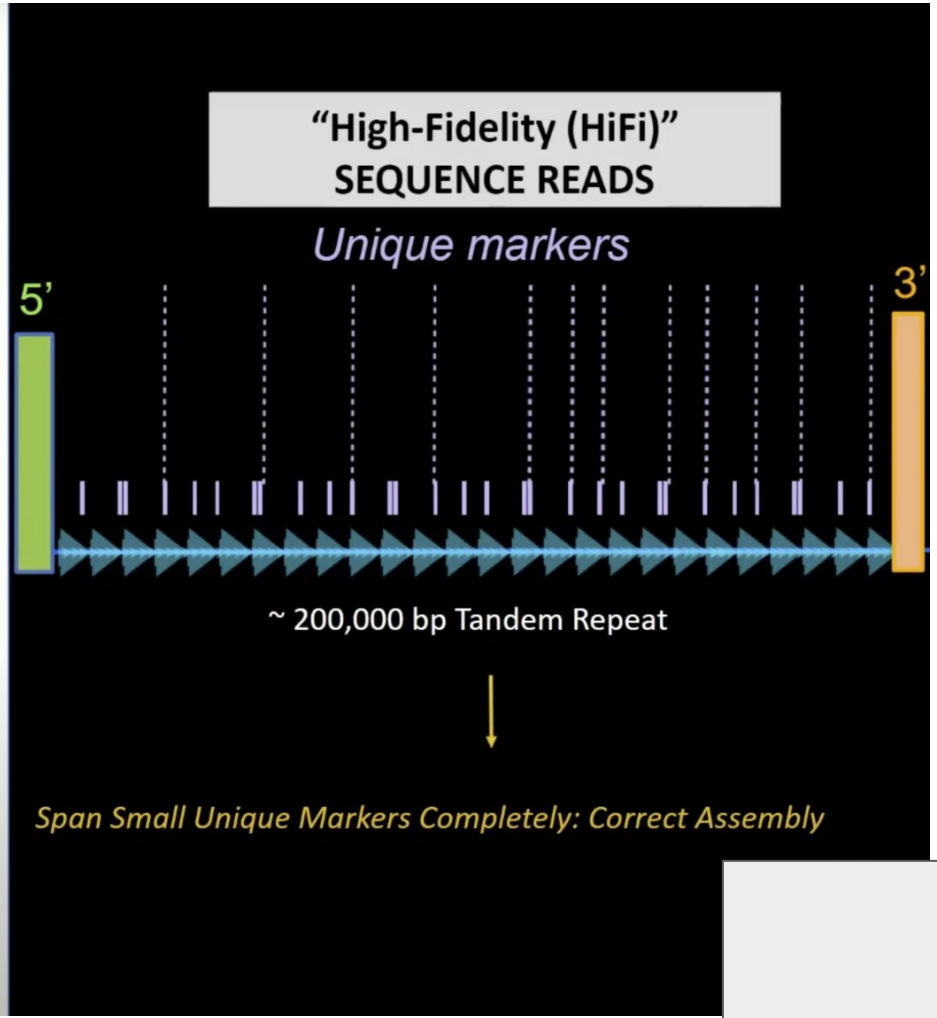


Repeat Assembly

- **PacBio High Fidelity (HiFi) Data**
  - 20 kb reads
  - **99.9% (Q30)** read quality

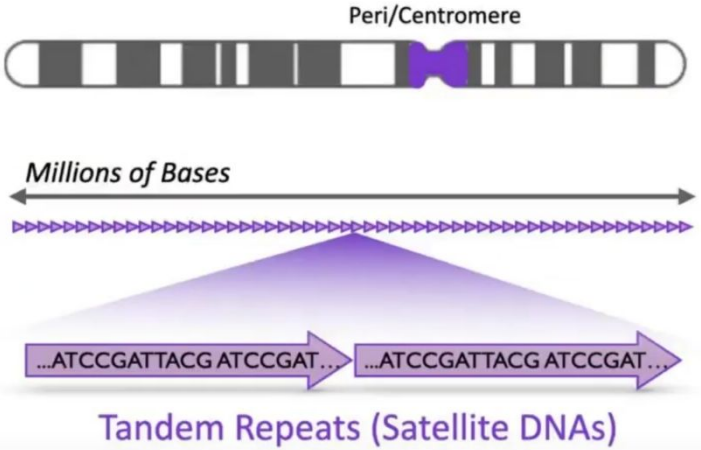


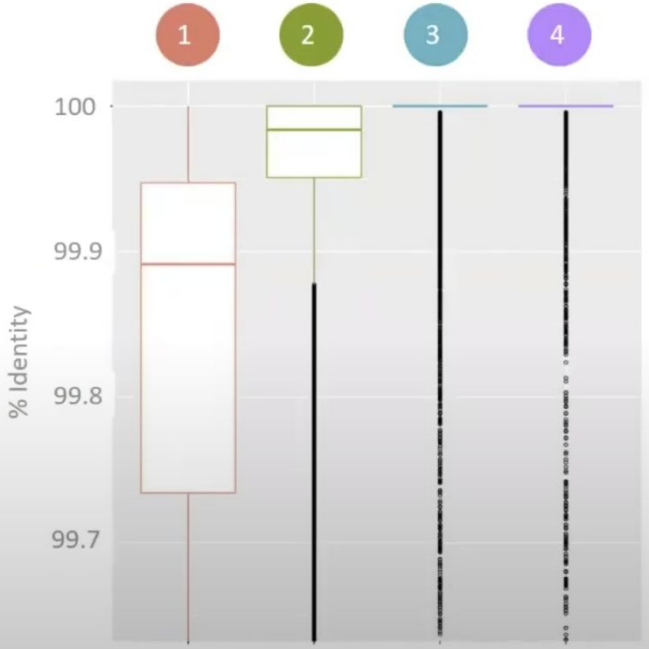
Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Wenger et al. *Nature Biotechnology* (2019)



Accurate repeat assemblies from HiFi reads (HiCanu)

Sergey Nurk

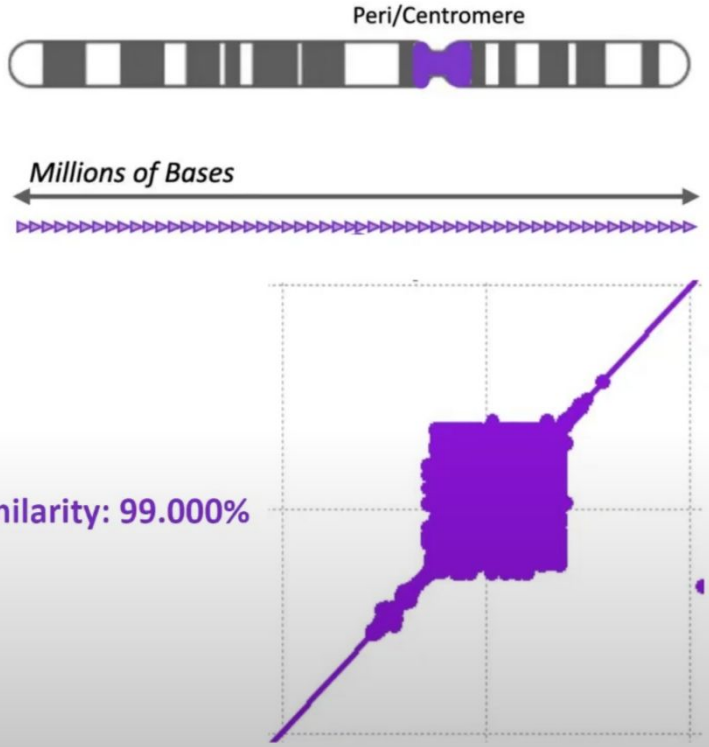




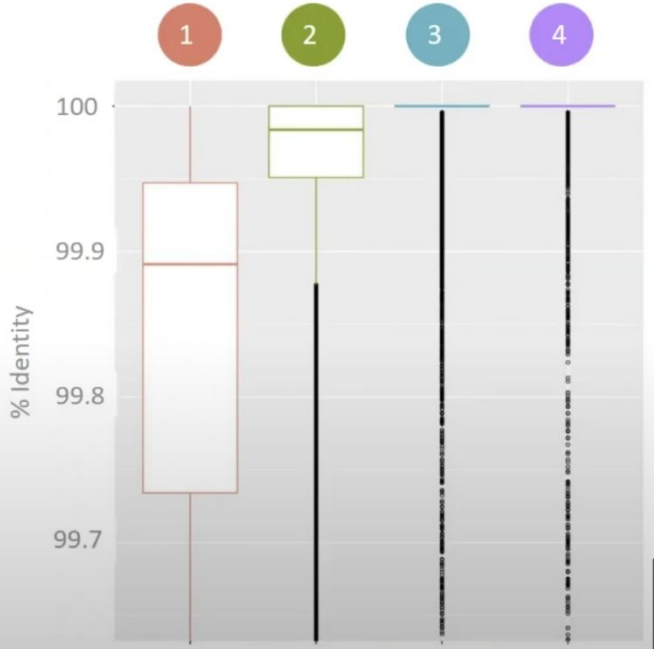
Nurk *et al.* HiCanu: accurate assembly of segment
satellites, and allelic variants from high-fidelity long

Accurate repeat assemblies from HiFi reads (HiCanu)

Sergey Nurk



Self-similarity: 99.000%

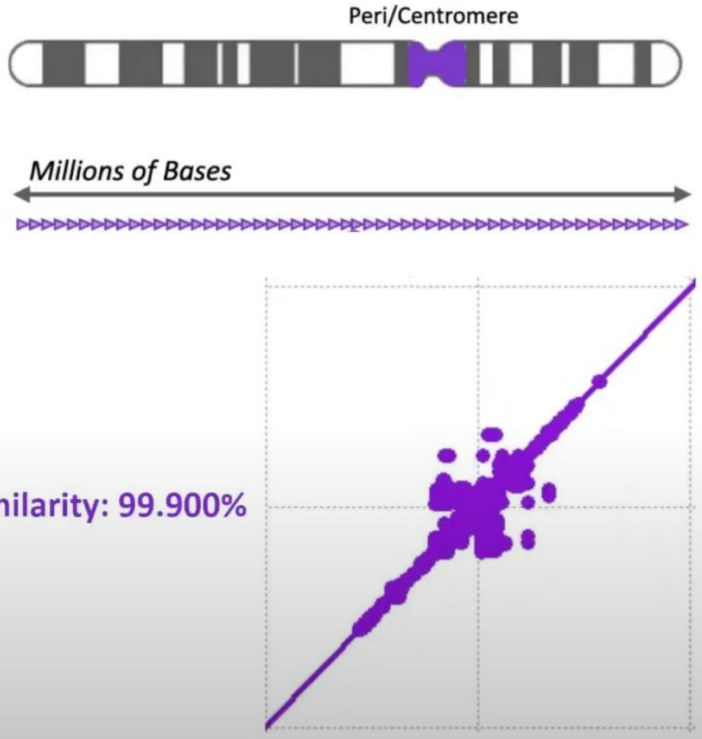


Slide Credit: Sergey Koren

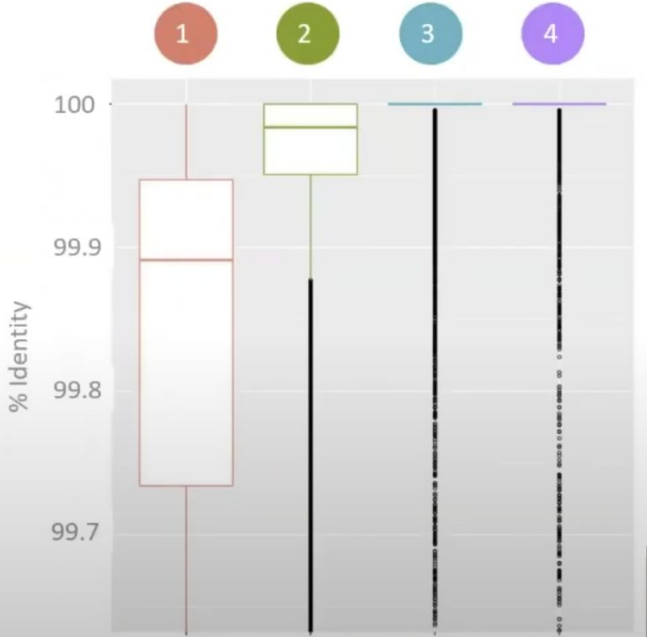
Nurk *et al.* HiCanu: accurate assembly of segment
satellites, and allelic variants from high-fidelity long

Accurate repeat assemblies from HiFi reads (HiCanu)

Sergey Nurk



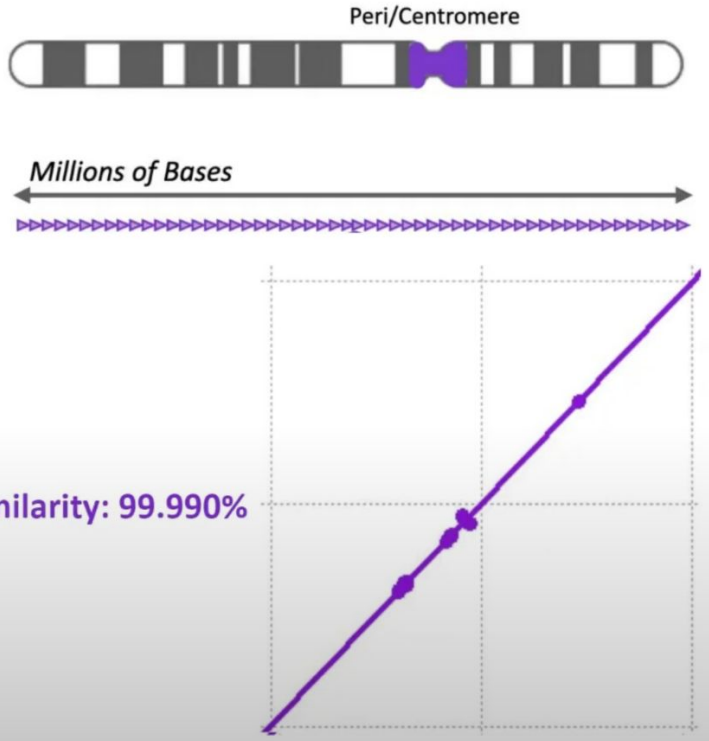
Slide Credit: Sergey Koren



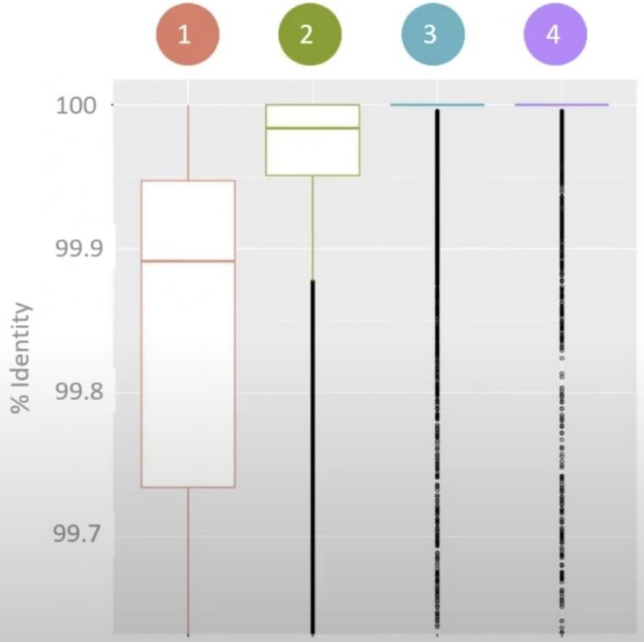
Nurk *et al.* HiCanu: accurate assembly of segment
satellites, and allelic variants from high-fidelity long

Accurate repeat assemblies from HiFi reads (HiCanu)

Sergey Nurk



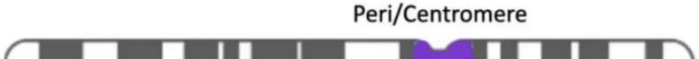
Slide Credit: Sergey Koren

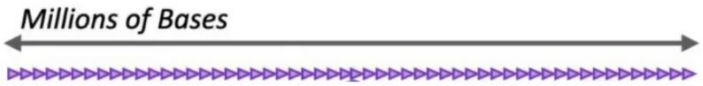


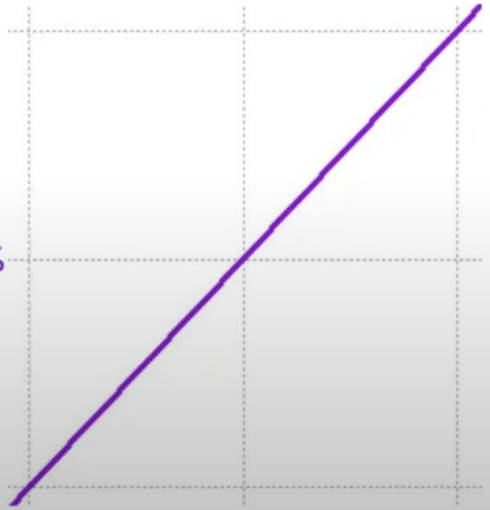
Nurk *et al.* HiCanu: accurate assembly of segment
satellites, and allelic variants from high-fidelity long

Accurate repeat assemblies from HiFi reads (HiCanu)

Sergey Nurk







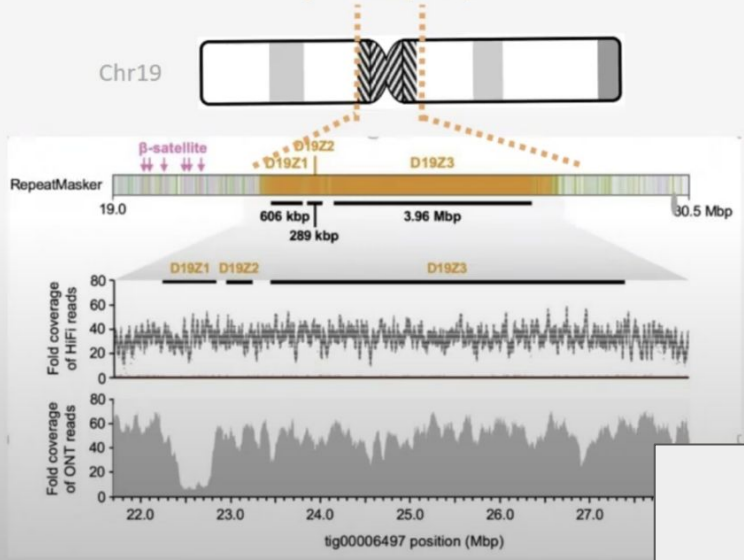
Self-similarity: 99.999%

Slide Credit: Sergey Koren

High Fidelity (Sanger-quality) Reads

Entire centromeric regions are being assembled

(~5 Mb repeats)



Nurk et al

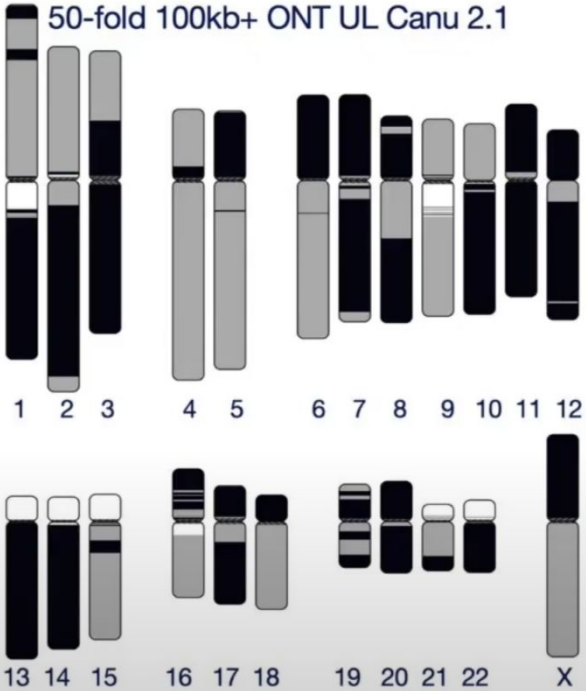
Genome assembly, 2001

NIH NHGRI

ref28 / hg10 : N50 0.5 Mbp

Genome assembly, 2019

Canu

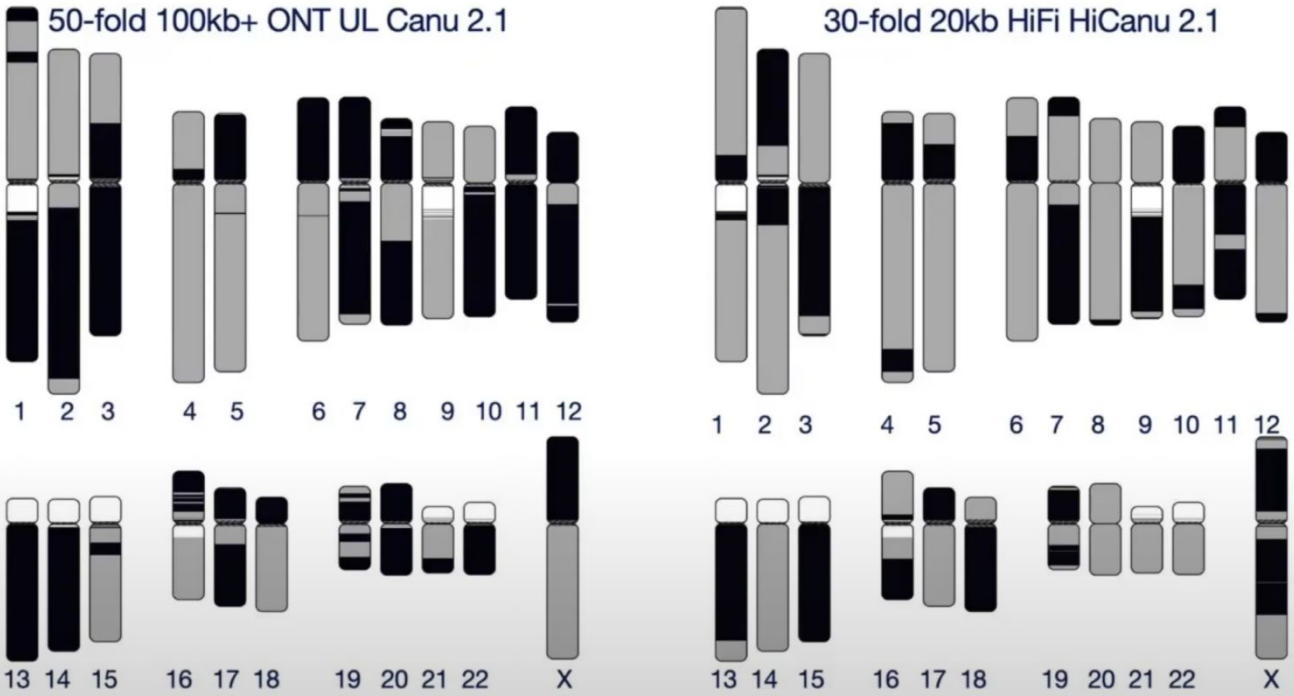


NIH NHGRI

Telomere-to-telomere assembly of a complete human X chromosome. Miga et al. *Nature* (2020)

Telomere-to-telomere assembly of a complete human X chromosome. Miga et al. *Nature* (2020)

HiFi and ONT UL are complementary



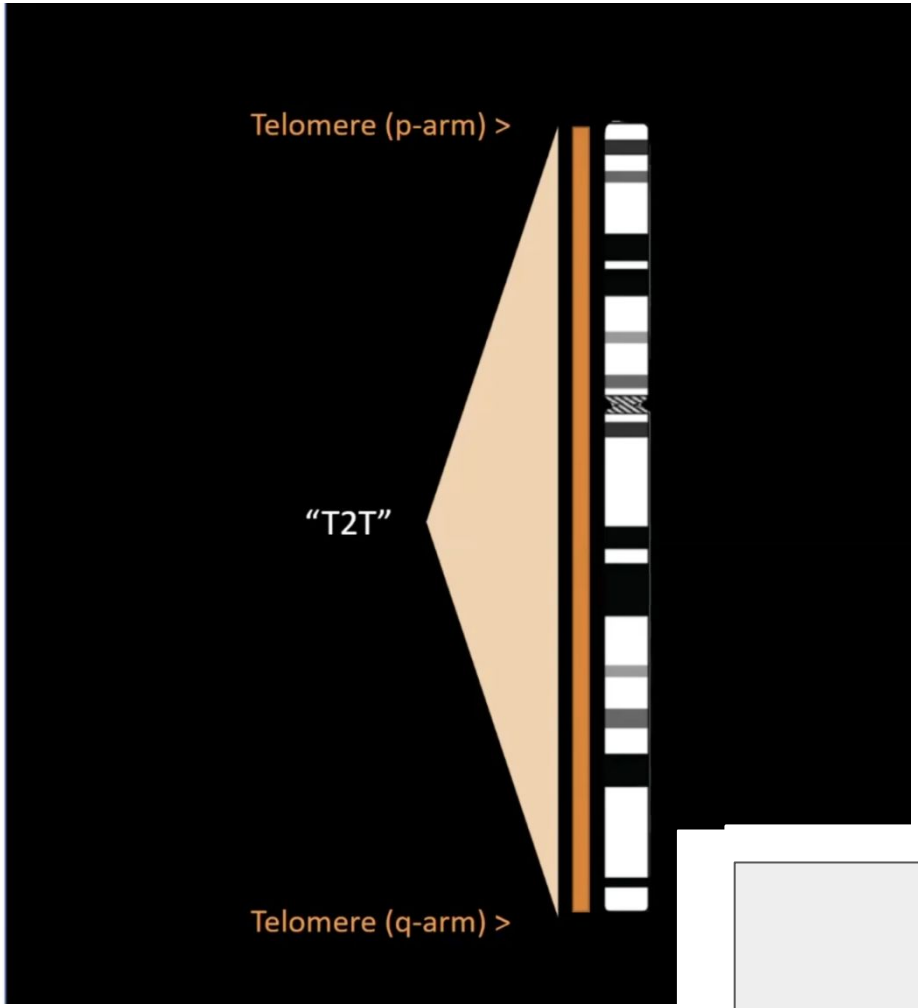
NIH NHGRI

Telomere-to-telomere assembly of a complete human X chromosome. Miga et al. *Nature* (2020)

Telomere-to-telomere chromosome assemblies

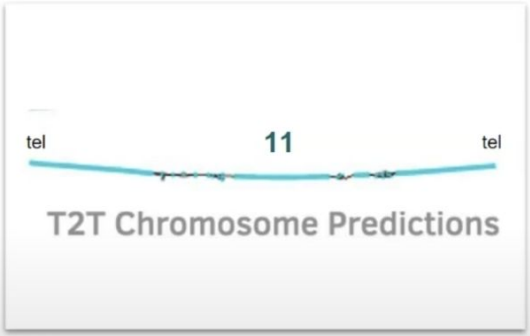
- HiFi: Construct string graph from long perfect overlaps
- Nanopore for “hard” tangles

Sergey Nurk

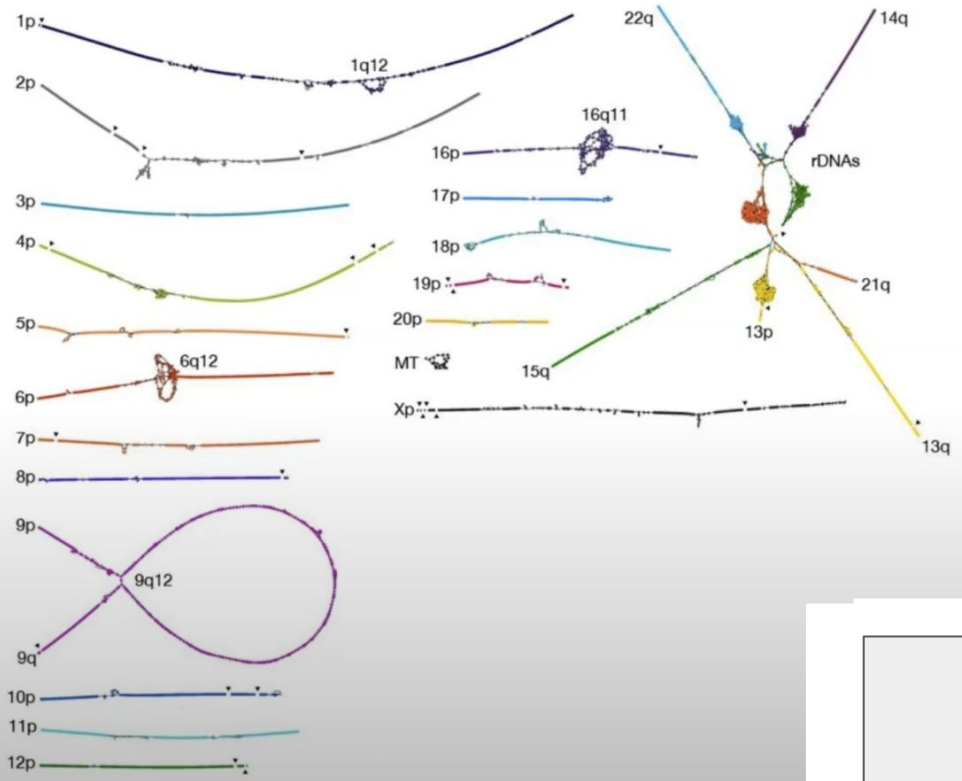


Custom assembly pipeline for T2T Reconstruction

Sergey Nurk Sergey Koren Adam Phillippy

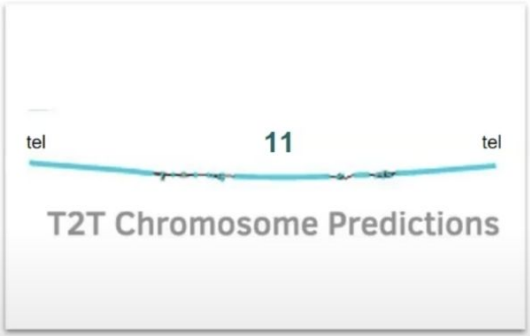


chr11

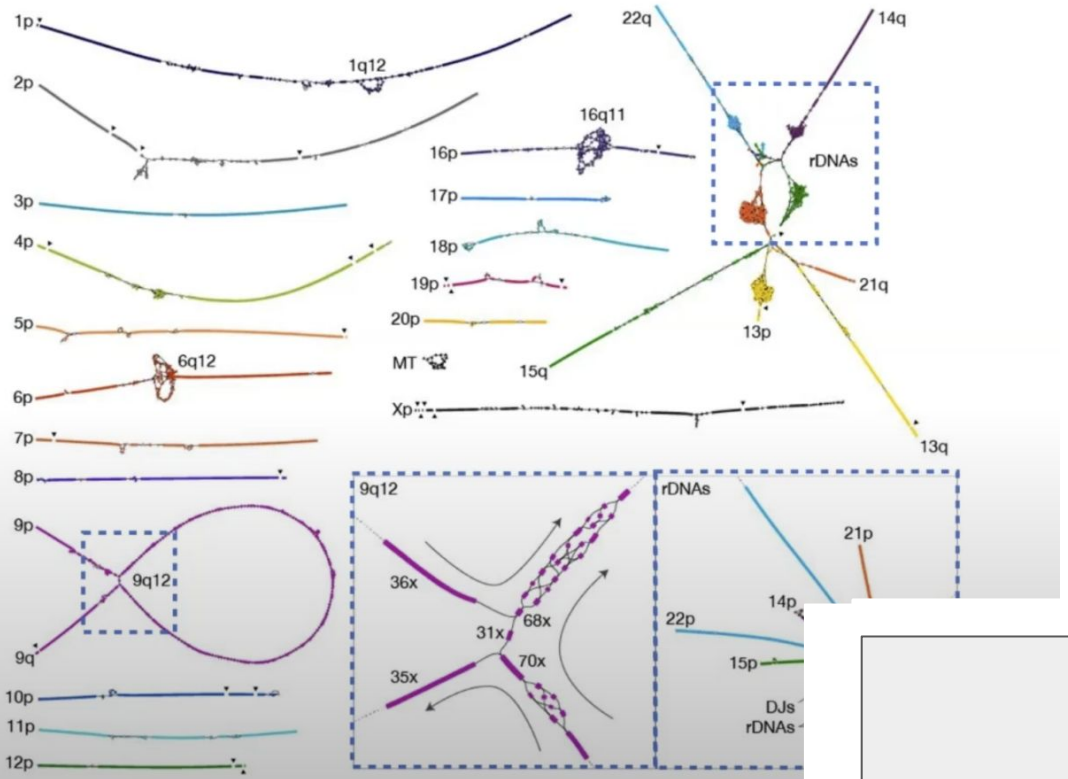


Custom assembly pipeline for T2T Reconstruction

Sergey Nurk *Sergey Koren* *Adam Phillippy*



chr11



The complete sequence of a human genome*

- **GRCh38.p13** (no alts)
  - 24 chromosomes
  - 42 unlocalized
  - 127 unplaced
  - 2,948,627,755 bp
  - 161.4 Mbp of Ns
  - Uncertain quality

+97 Mbp

-150 Mbp

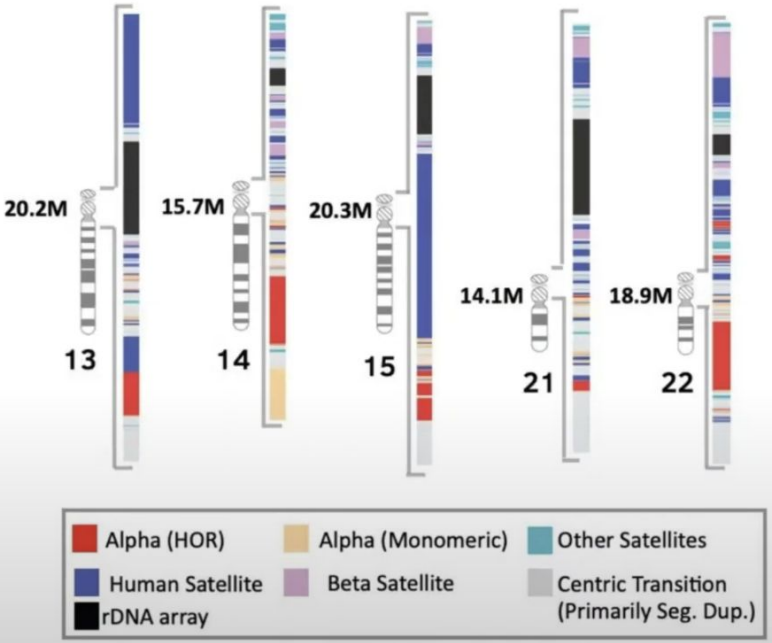
- **T2T CHM13** (no hets)
  - 23 chromosomes (no Y)
  - 0 unlocalized
  - 0 unplaced
  - 3,045,441,701 bp
  - 11.5 Mbp of “known” Ns
  - ~Q70, no known hom SVs

Estimated CHM13 genome size of **3.057 Gbp**
100-190 Mbp (3-6%) of new sequence vs. GRCh38

genomeinformatics.github.io

*5 N-gaps internal to the rDNA arrays remain, GRCh38 aligned with `nucmer -mum -l 500 -c 10000 -g 5000`

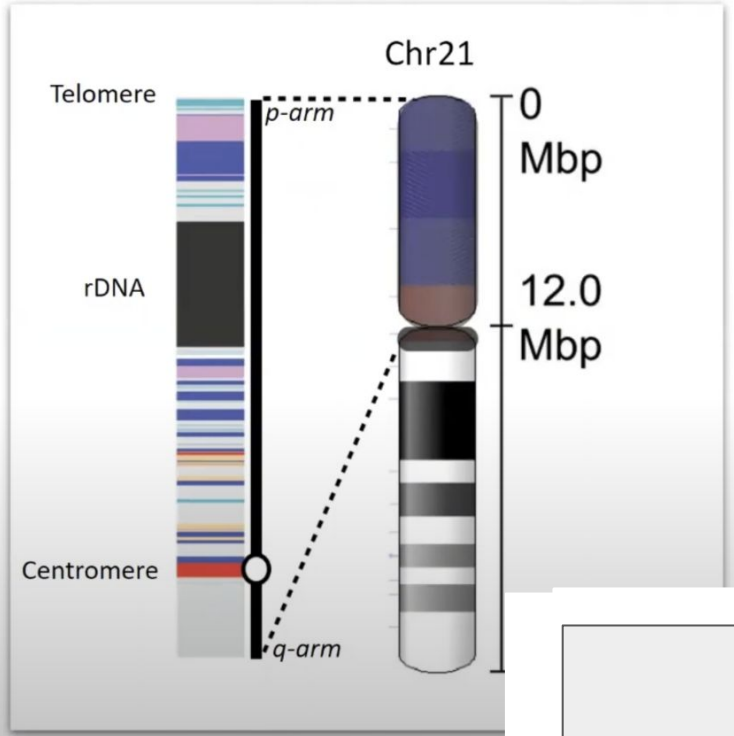
Acrocentric Short Arms



Acrocentric Short Arm References: ~90 Mbps

Satellite DNA, rDNAs, and Segmental Duplications

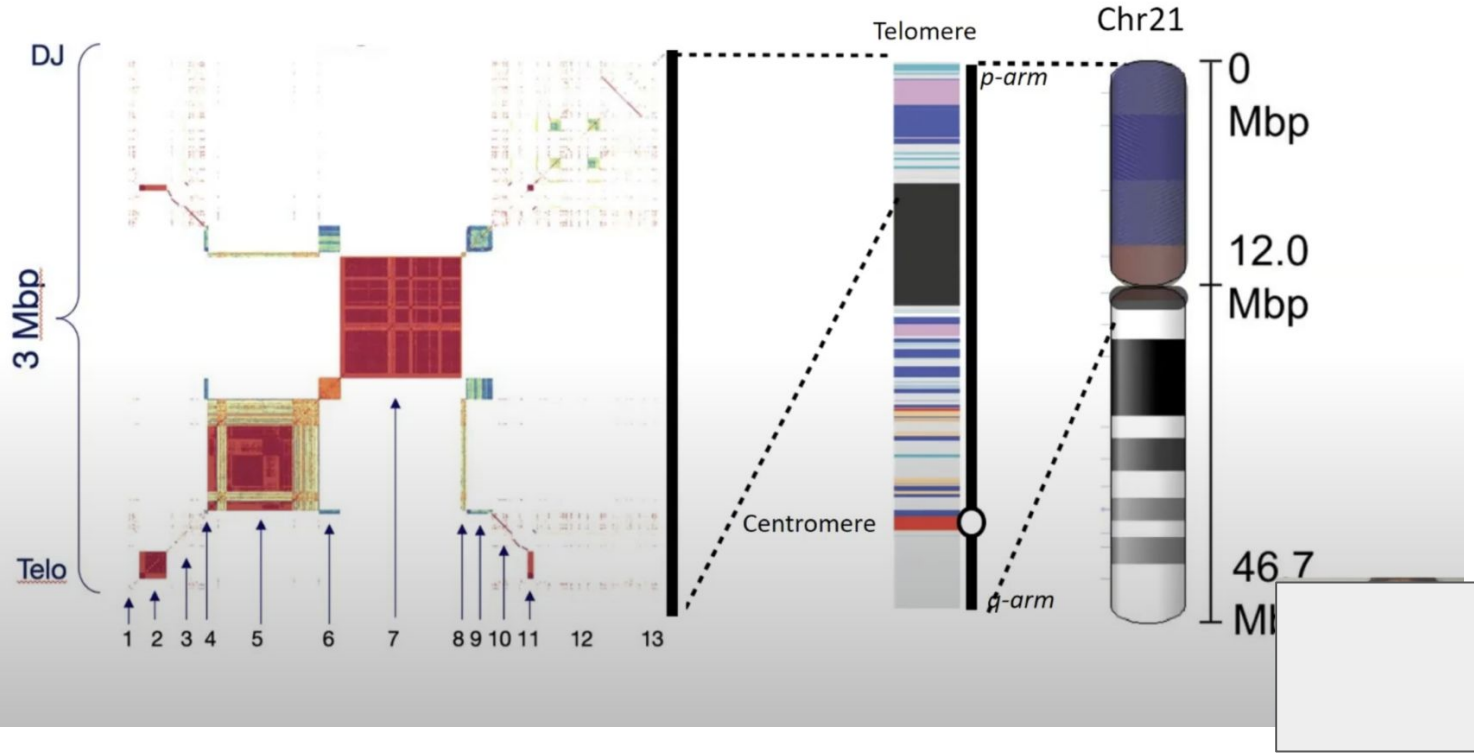
(*with hundreds of candidate gene annotations)



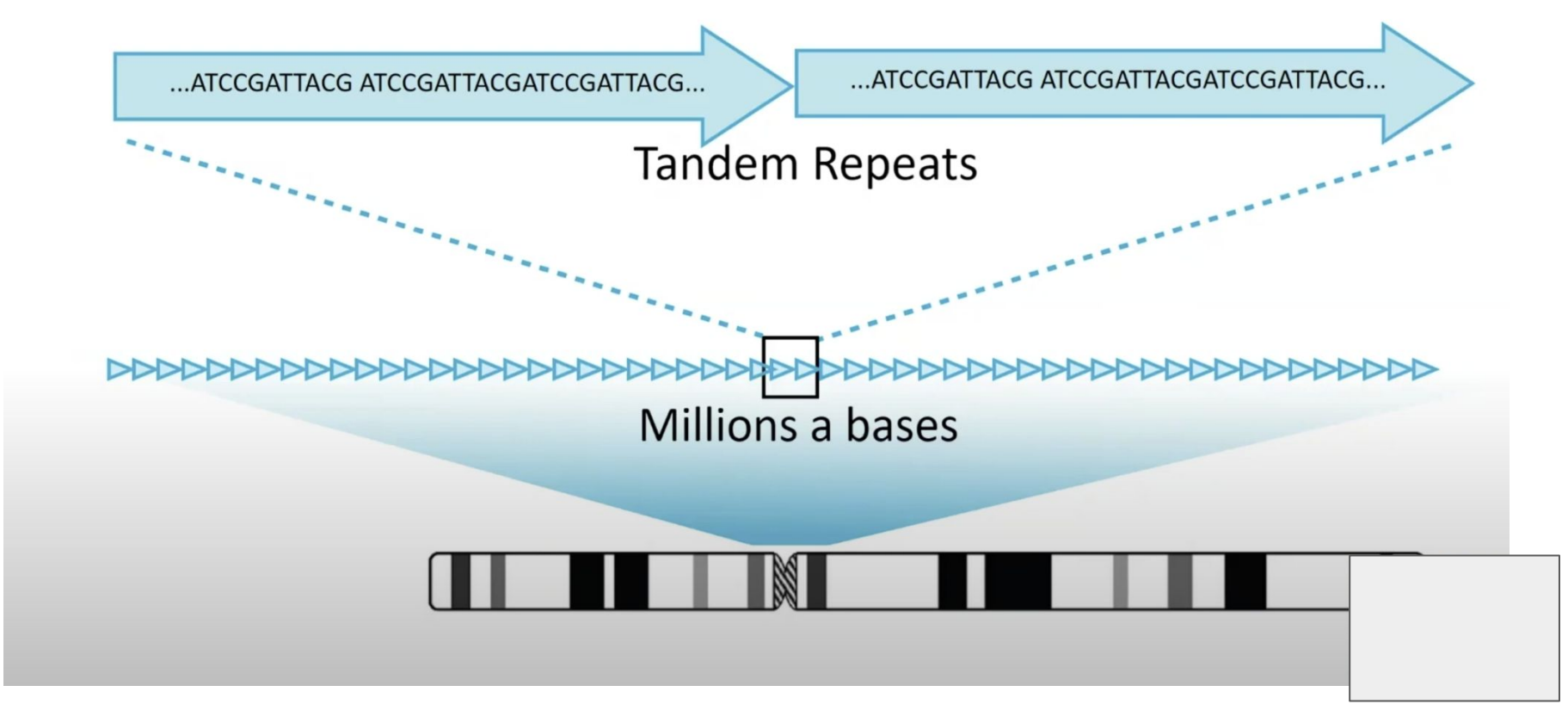
Acrocentric Short Arms

Adam Phillippy

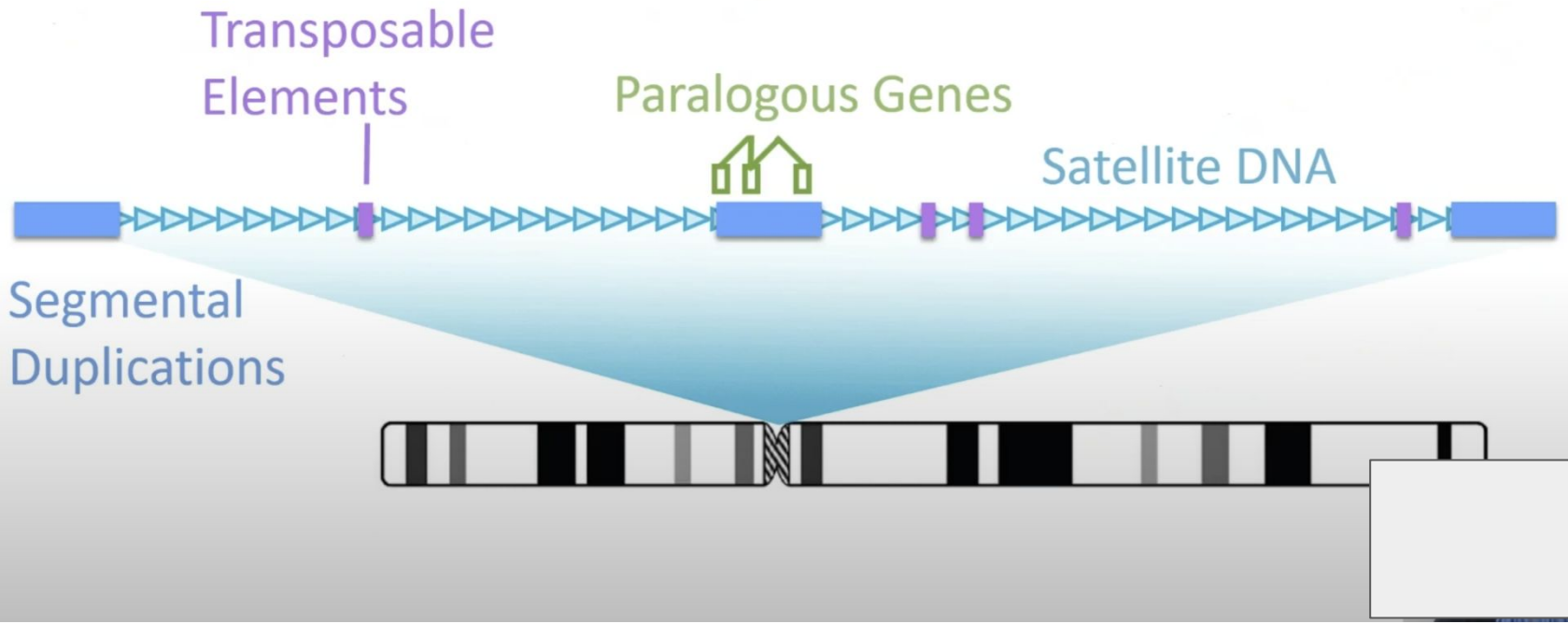
1. Telomere
2. ACRO tandem
3. Inverted segdup
4. HSat2,3
5. BSat
6. **HSat2,3**
7. **HSat1**
8. BSat
9. HSat2,3
10. Inverted segdup
11. ACRO tandem
12. rDNA DJ
13. rDNA array



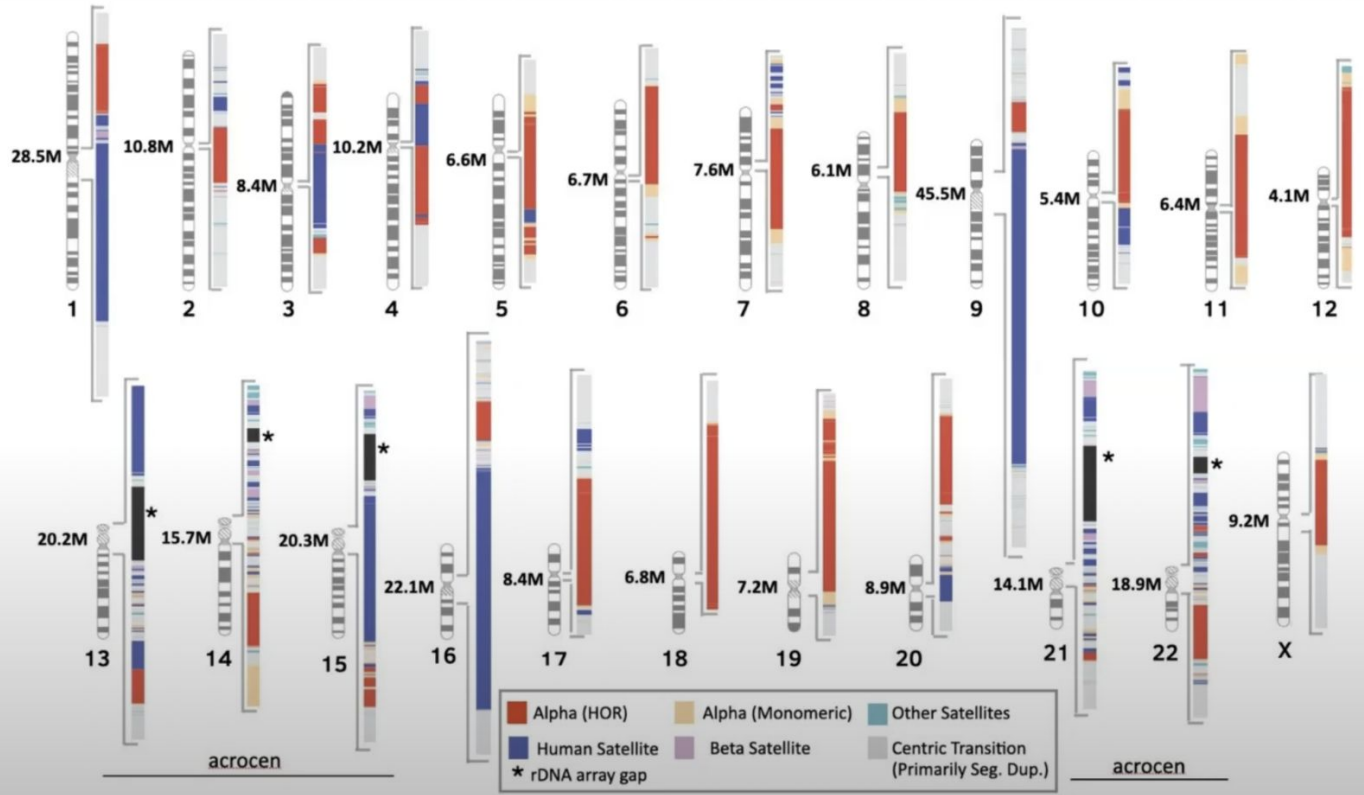
Largest gaps are in human centromeric regions



Centromeric regions: A Genomic Ecosystem Defined by Repeats

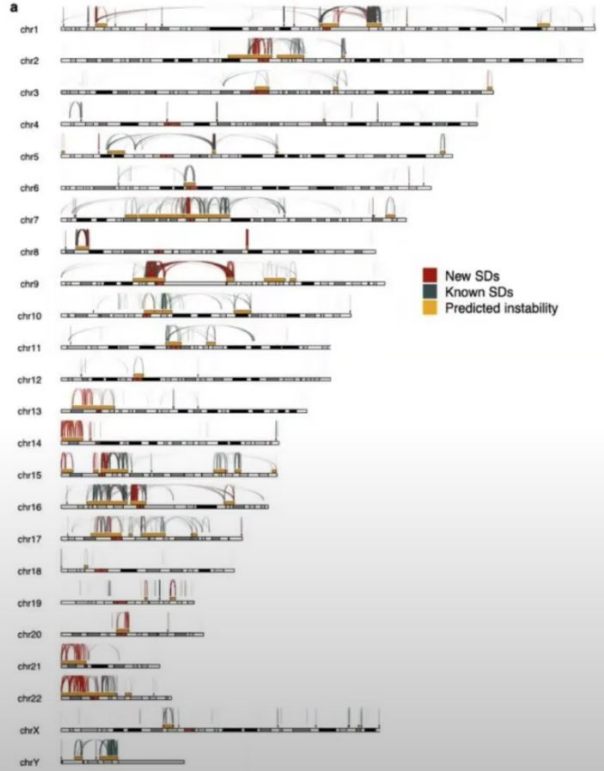


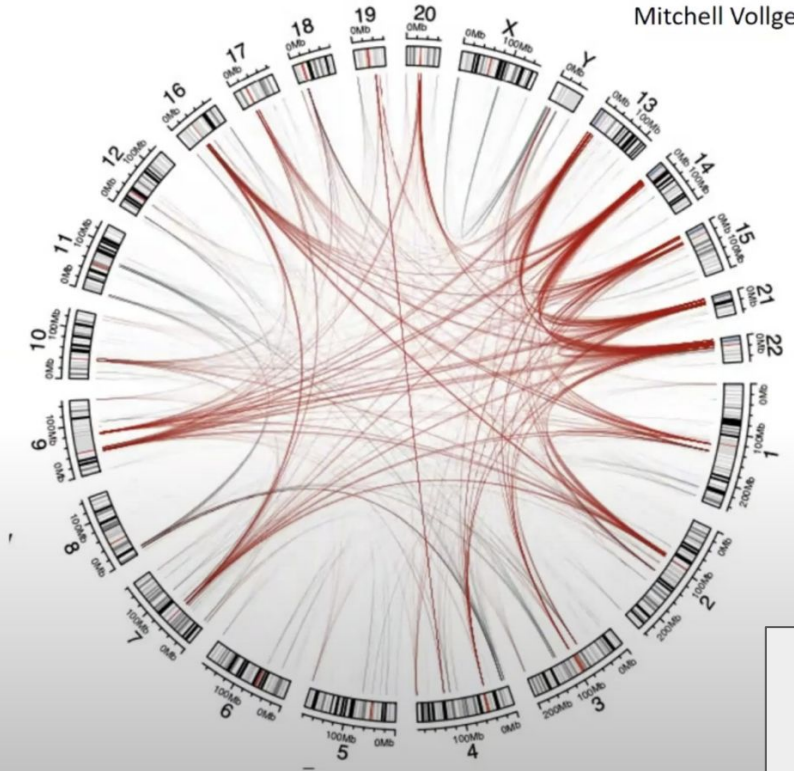
High Resolution Maps of Human Peri/Centromeric Regions



Comprehensive maps of segmental duplications

Mitchell Vollger Evan Eichler

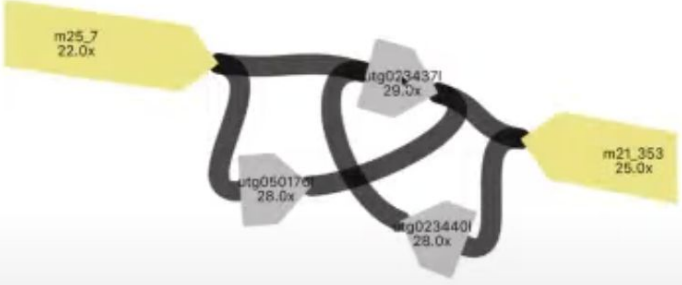




Verkko assembler

Simplifying the HiFi graph

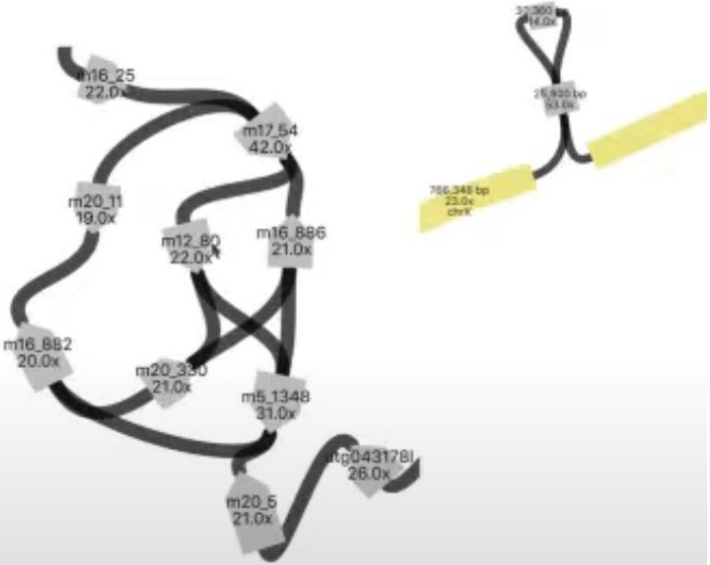
- Walking simple paths
- Aligning Nanopore reads

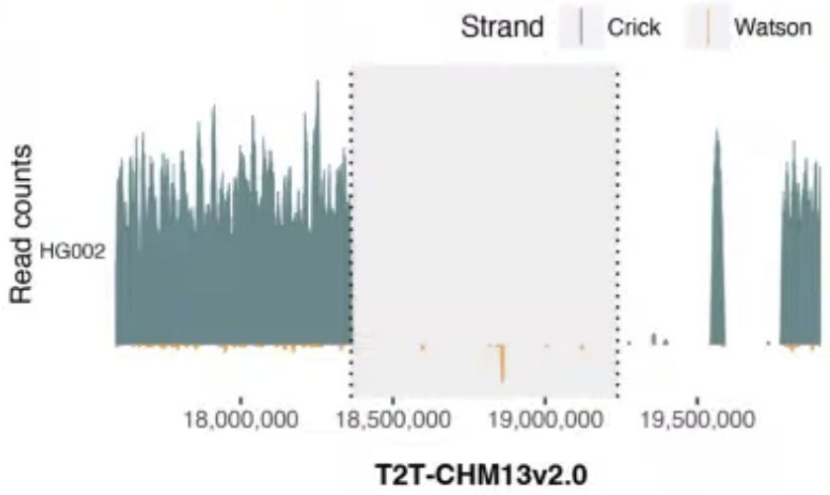


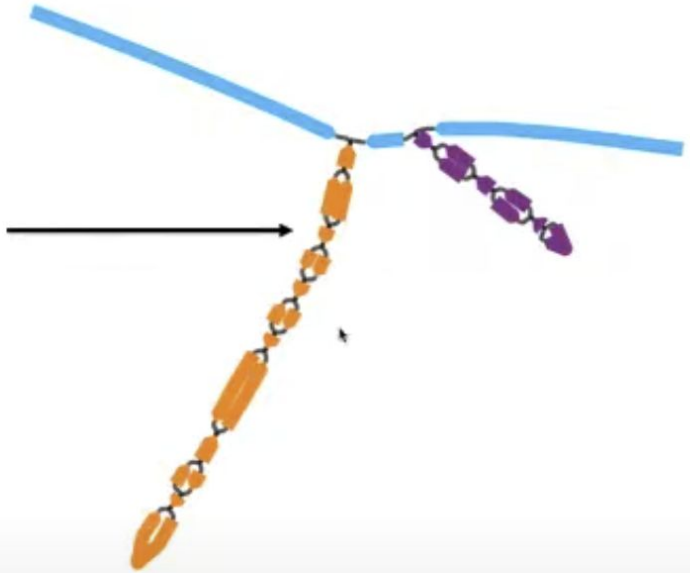
Simplifying the HiFi graph

- Walking simple paths

- Aligning Nanopore reads







P5 mis-assembly (~800 kb)

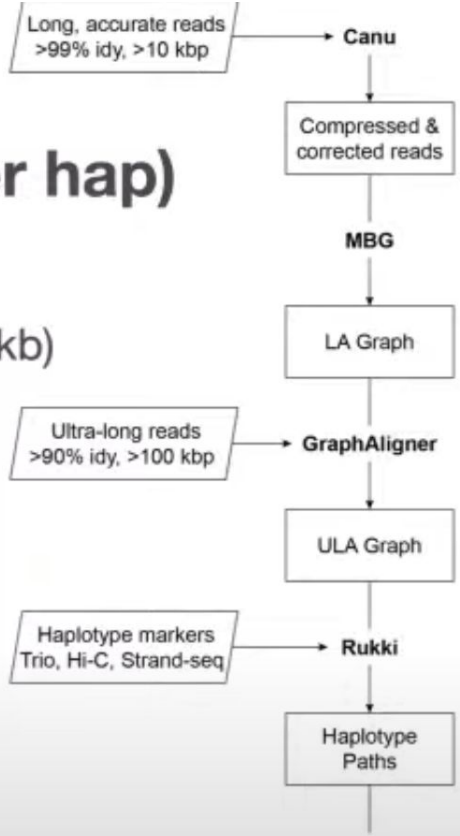
~20 kbp 100% identical

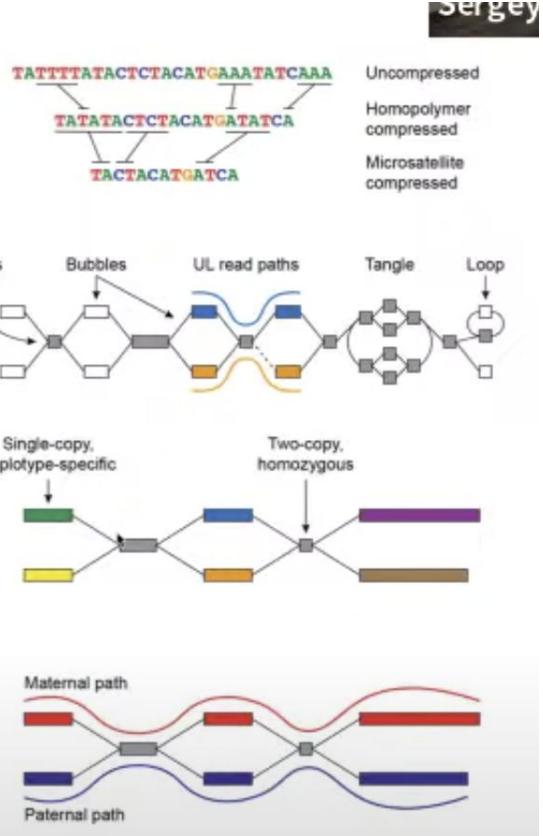
Verkko!

- **Sequencing recipe (per hap)**
  - 20-25x high accuracy
    - (Pac Bio HiFi, Duplex, HERRO)
  - 15-20x ONT ultra-long (>100 kb)
  - 20x Illumina Trio or Hi-C
  - Available from conda

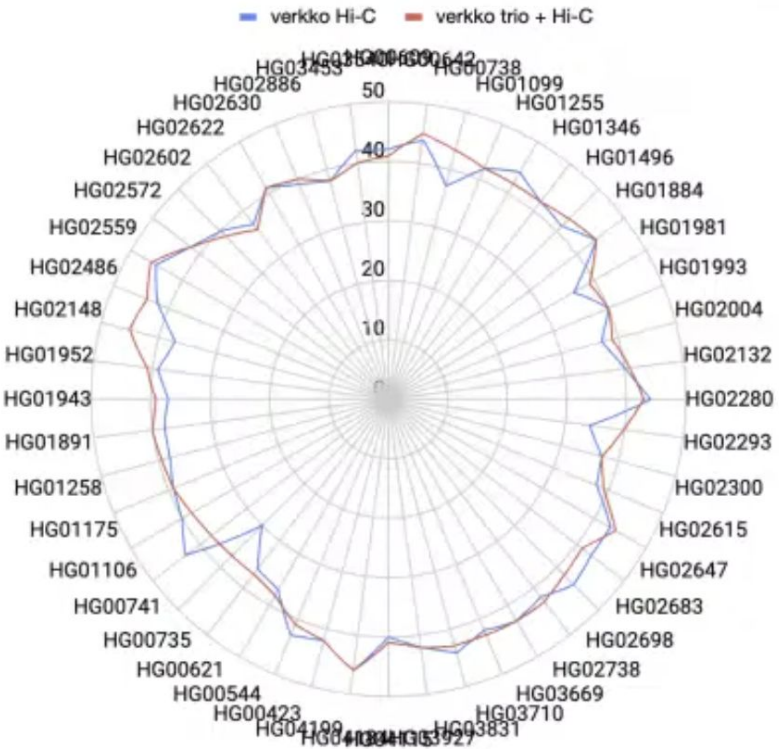
Verkko!

- **Sequencing recipe (per hap)**
  - 20-25x high accuracy
    - (Pac Bio HiFi, Duplex, HERRO)
  - 15-20x ONT ultra-long (>100 kb)
  - 20x Illumina Trio or Hi-C
  - Available from conda
- **Verkko pipeline**
  - Read correction
  - Sparse multiplex DBG
  - ONT graph simplification
  - Walk haplotypes



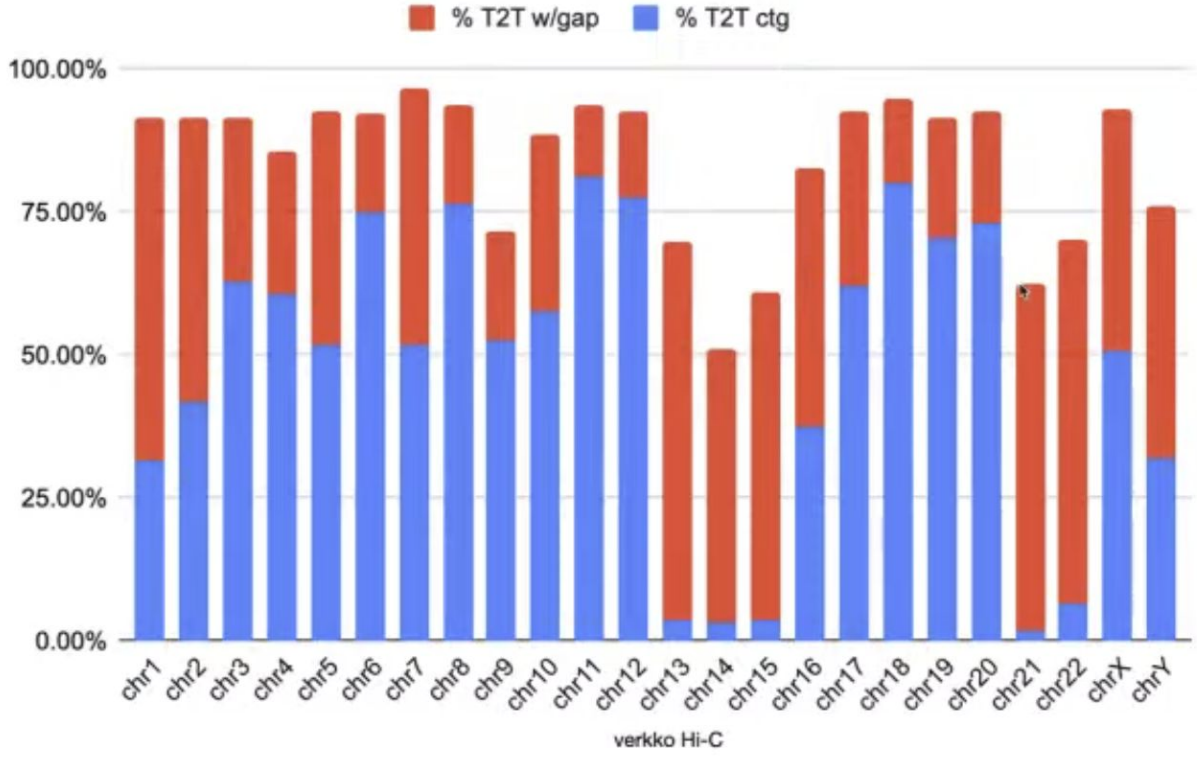


- Our assemblies are strong, 40/46 T2T scaffold average on 101 human samples
  - 62x HiFi, 34x >100 kbp ONT (158x total), 68x Hi-C



NIH
NHGRI

- Another view, by chromosome, 52% T2T contig, 91% T2T scaffold



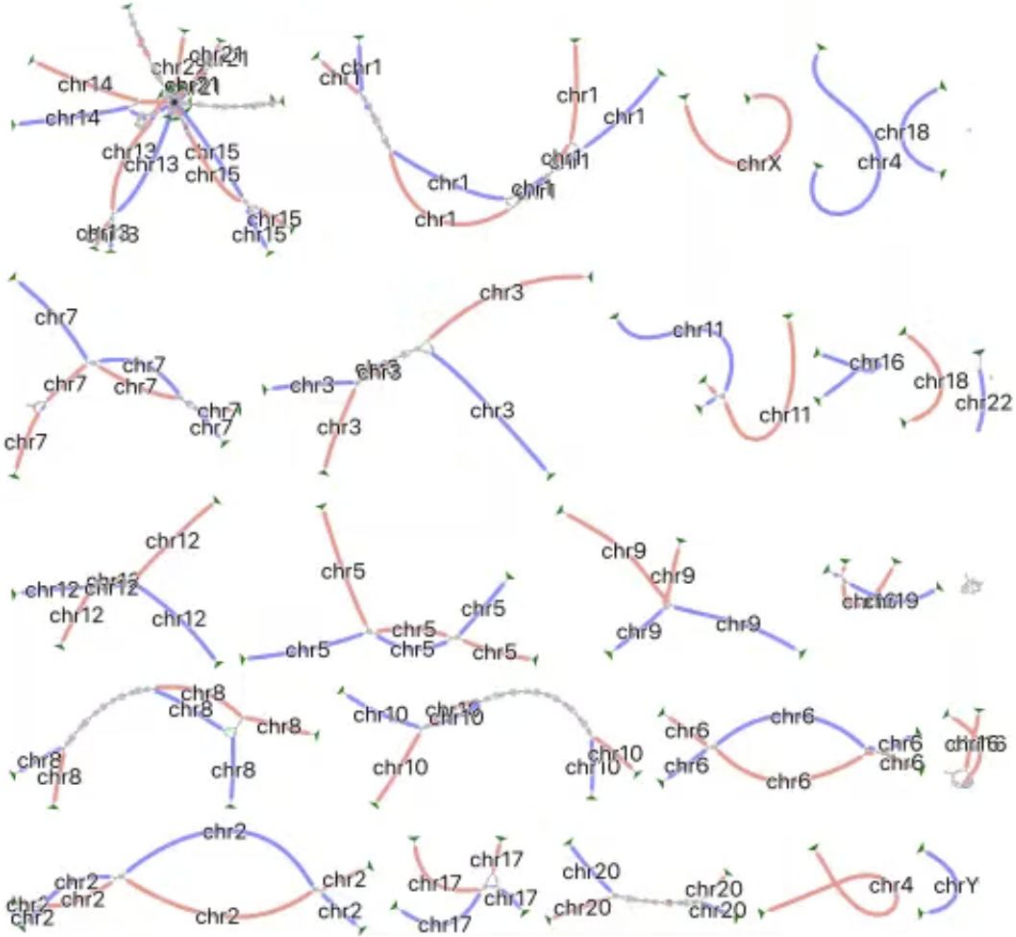
NIH NHGRI

How do we know it’s any good?

T2T QC

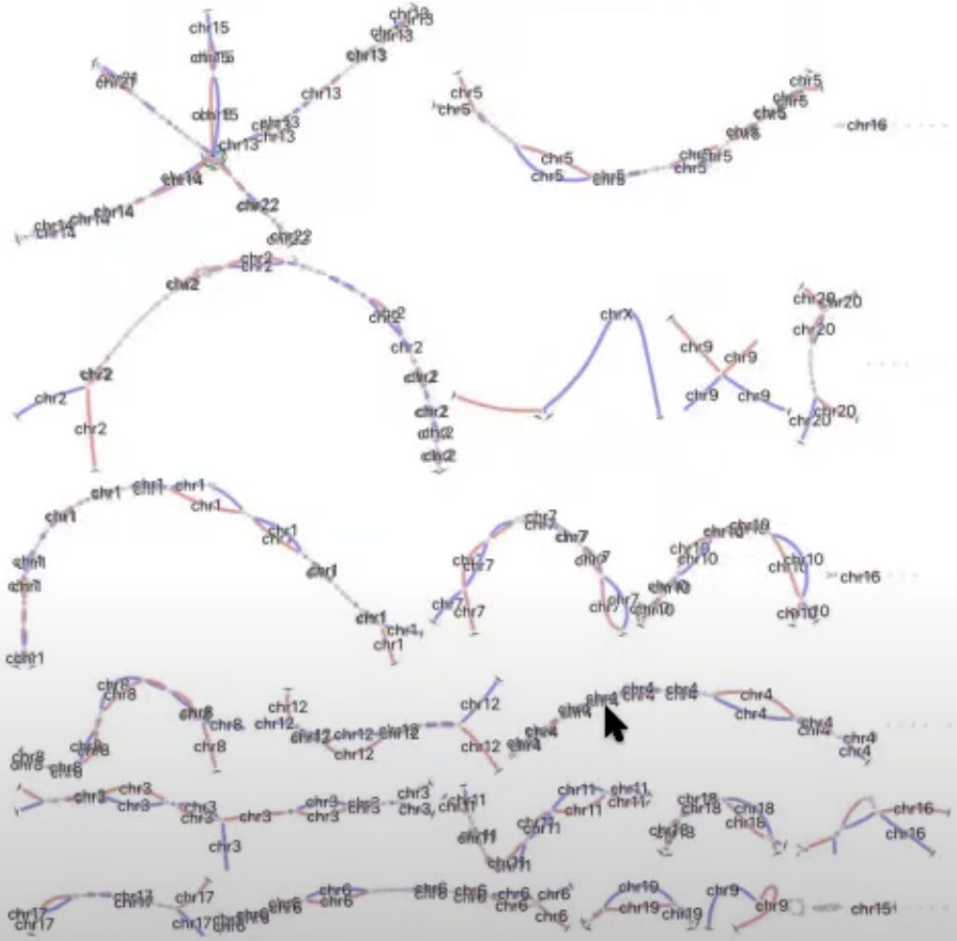
- T2T contigs and scaffolds
- QV
  - Merqury, yak
- Hamming & switch error rate
  - If trio data available
- Missing/duplicated core genes
  - Compleasm, busco
- non-T2T contiguity metrics
  - N50, L50
- Alignment-based evaluation
  - NucFreq, Flagger, VerityMap

Graph, the "Good, the meh and the ugly"



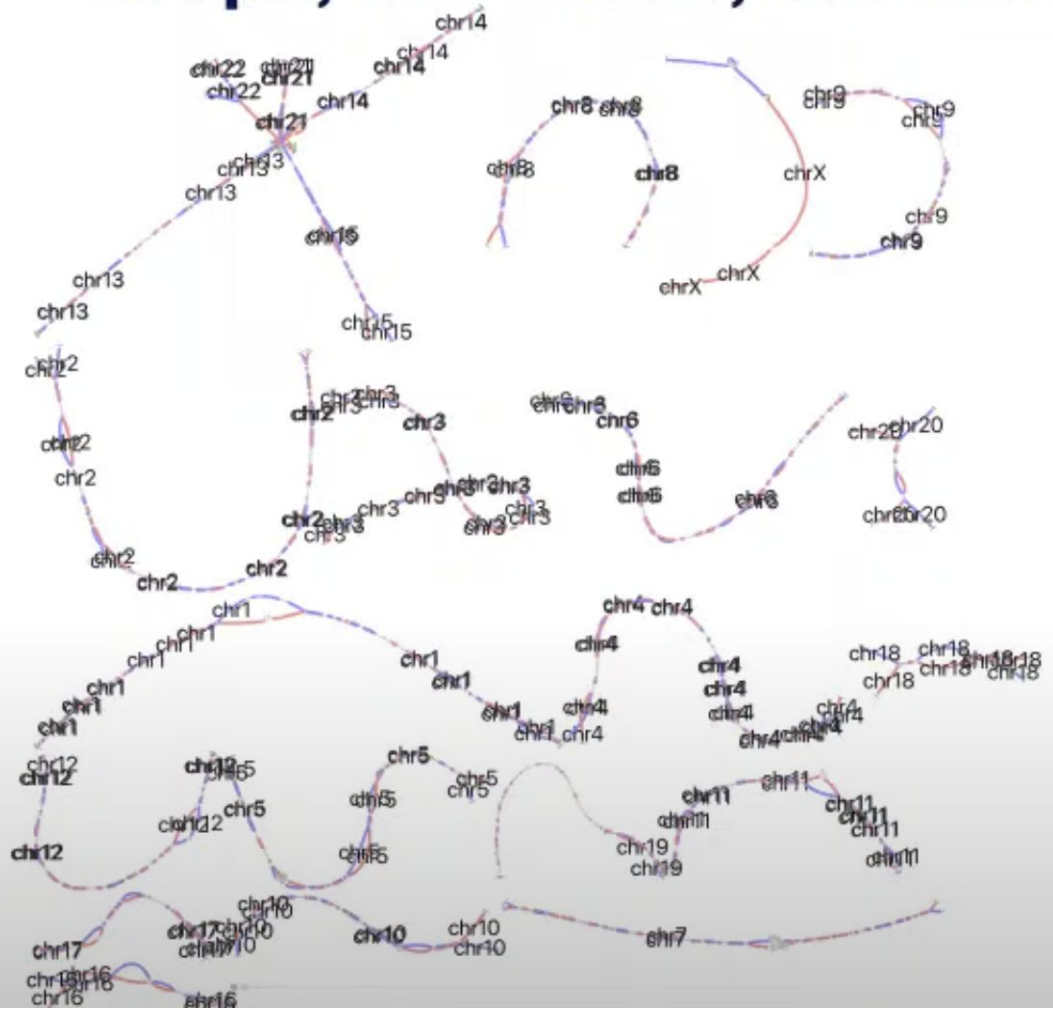
Q20 Herro-correct, ≅300 nodes
- 44 T2T scfs (36 ctgs)

Graph, the “Good, the meh and the ugly”



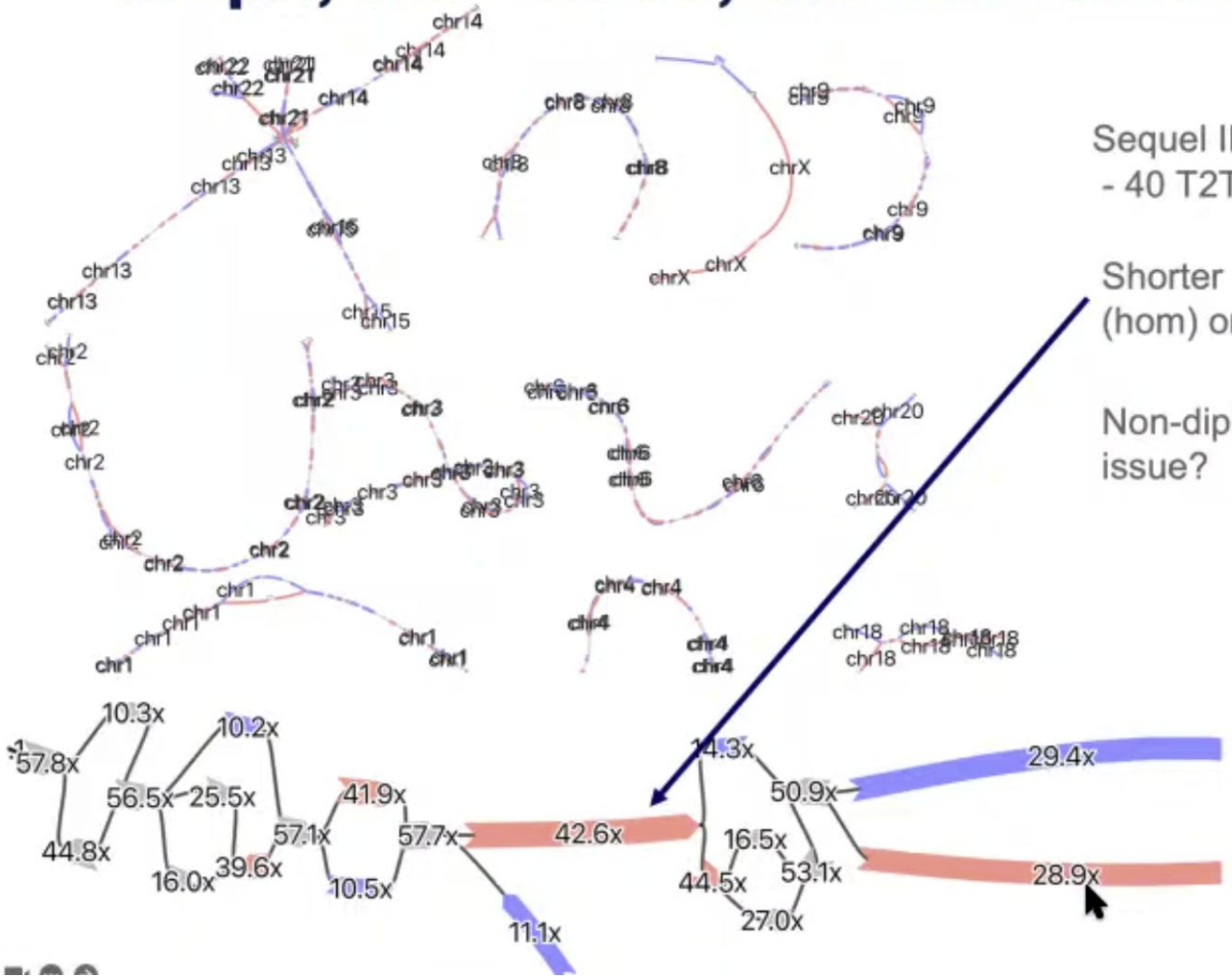
Revio/R10 YR4, ≅900 nodes
- 37 T2T scfs (25 ctgs)

Graph, the “Good, the meh and the ugly”



Sequel II, R9 ≅2000 nodes
- 40 T2T scfs (31 ctgs)

Graph, the “Good, the meh and the ugly”

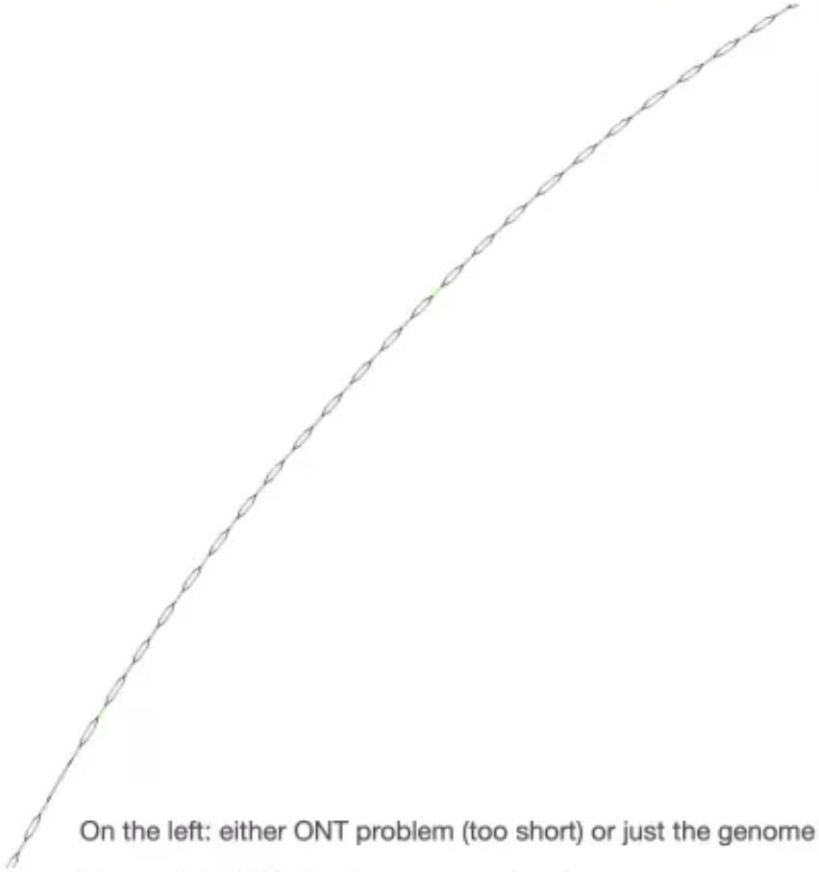


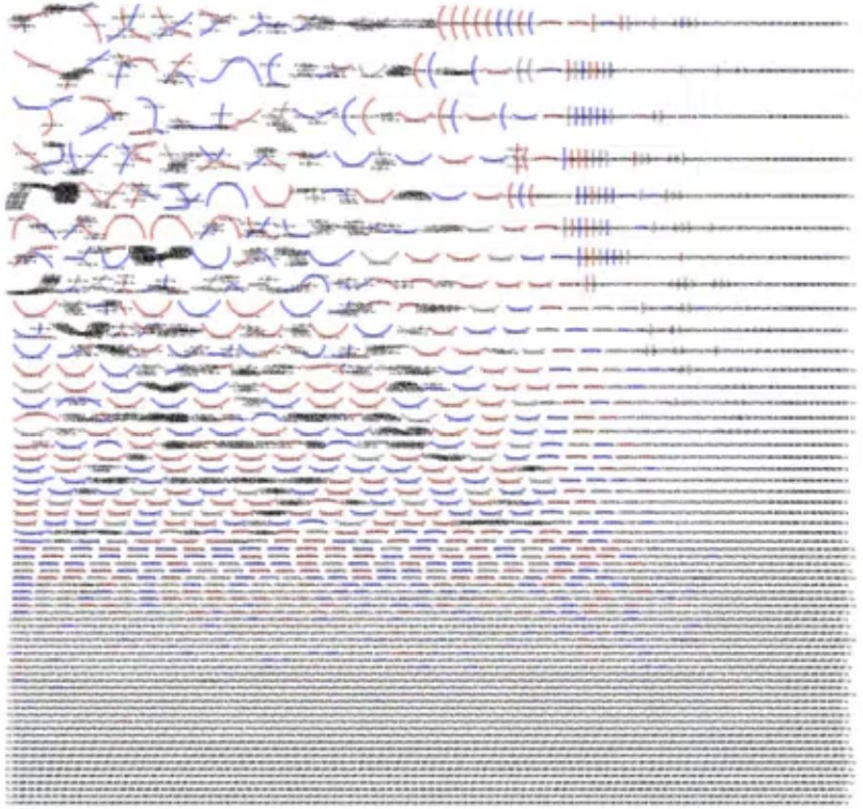
Sequel II, R9 ≅2000 nodes
- 40 T2T scfs (31 ctgs)

Shorter nodes, more gray (hom) or yellow (switch)

Non-diploid structure, cell line issue?

Too few T2T, fragmented assembly



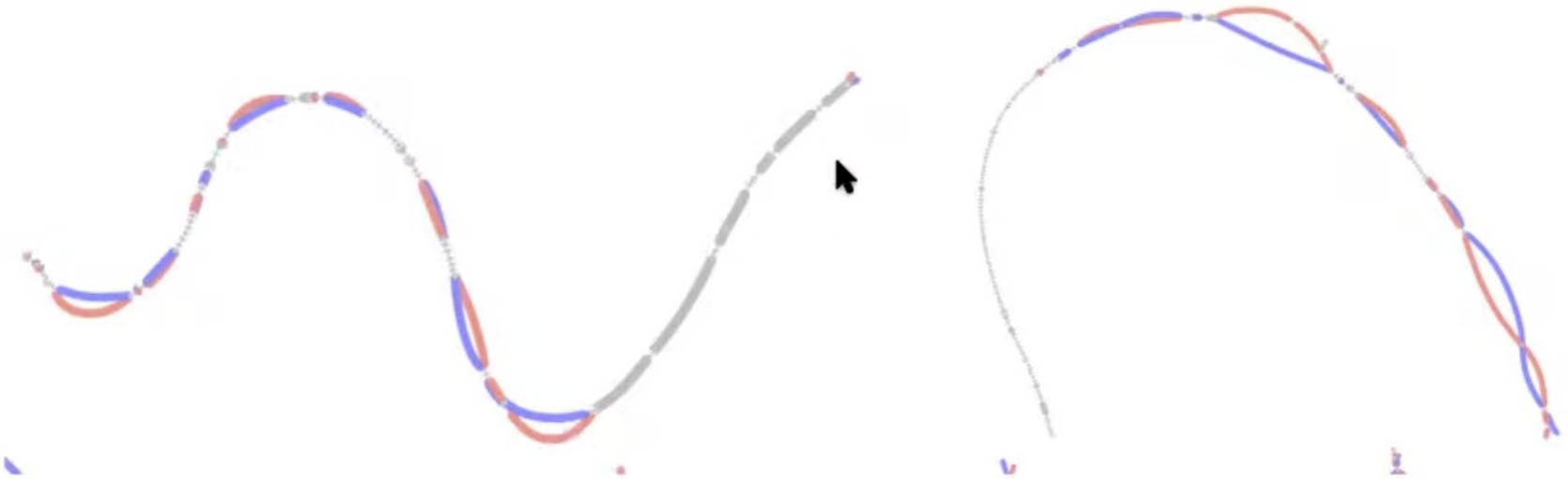


On the left: either ONT problem (too short) or just the genome structure

On the right: HiFi ultra-low input protocol problems

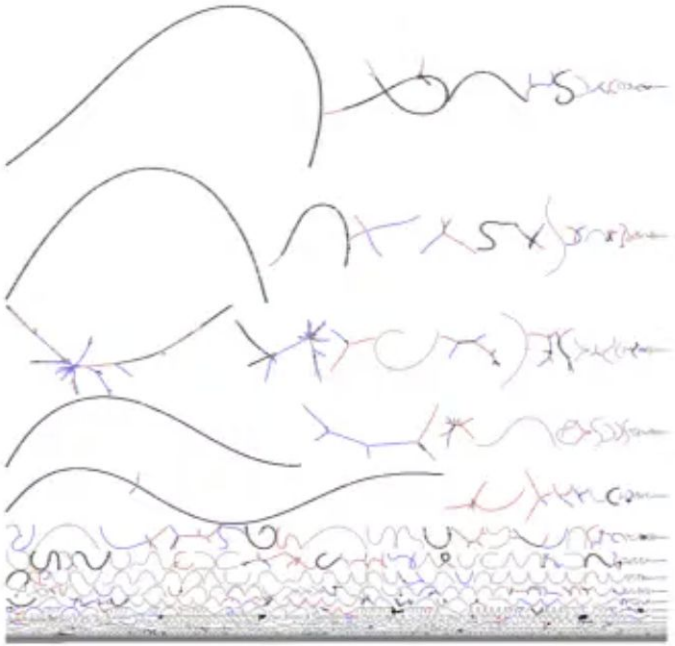
NIH NHGRI

Phasing issues: large homozygous regions



- Different chromosomes of same bonobo sample
- Left is phased correctly (long nodes), right - lots of unassigned (and so missing genes)

Heterozygosity level matters!



- Verkko can have problems with both very high and very low heterozygosity
- Sometimes this may even happen in the same sample!

Large tandem repeats

- Large (few Mb) tandem repeats is quite typical issue preventing verkko from T2T.
- Verkko/rukki heuristics stops because there are multiple large “blue” extensions for a large blue node here.
- Usually random walk will not add many errors here