The *complete* sequence of a human genome

Adam M. Phillippy

NCI BTEP
September 16, 2021

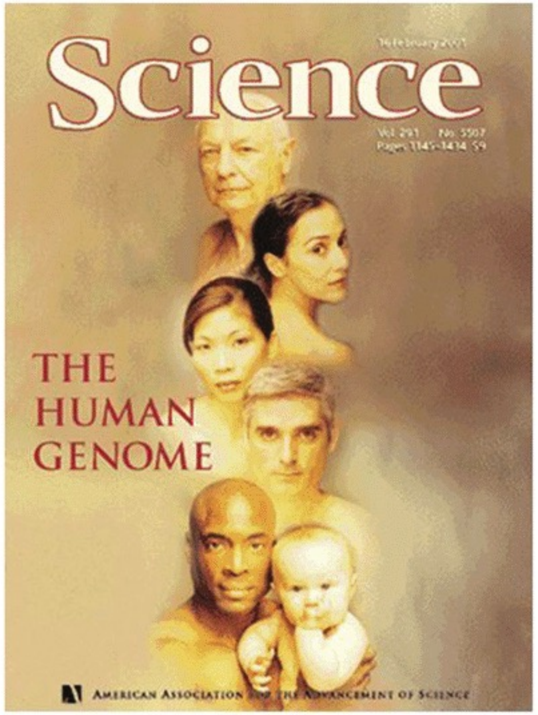
@aphillippy

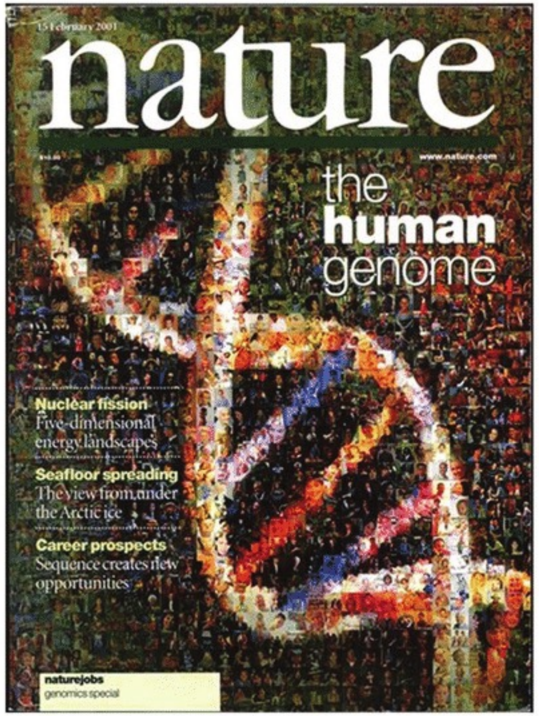
DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Human Genome Research Institute

The **Forefront** of **Genomics**®

I heard it was finished 20 years ago?



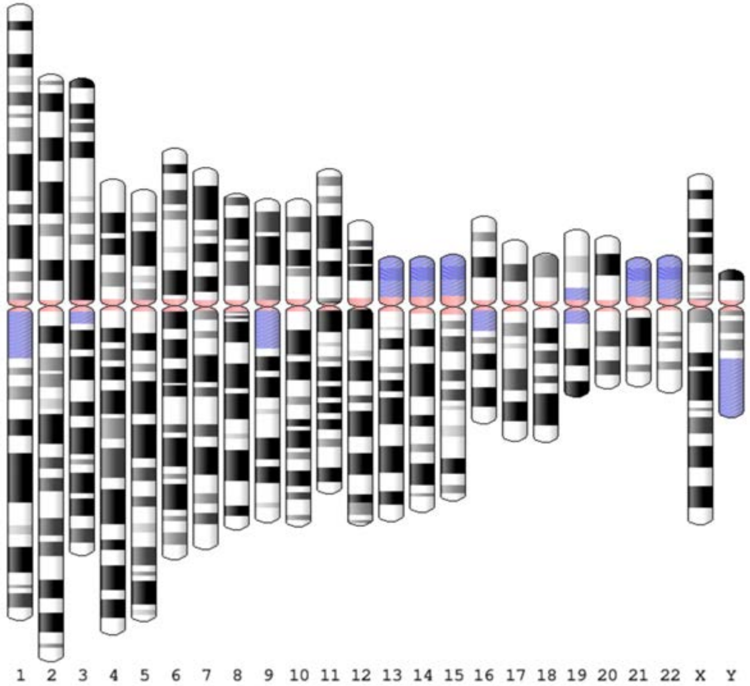

nature
the
human
genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

June 26, 2000

No!

- **And what’s missing is underappreciated**
  - “In the April 2003 version, there are less than 400 gaps and 99 percent of the genome is finished” (genome.gov)

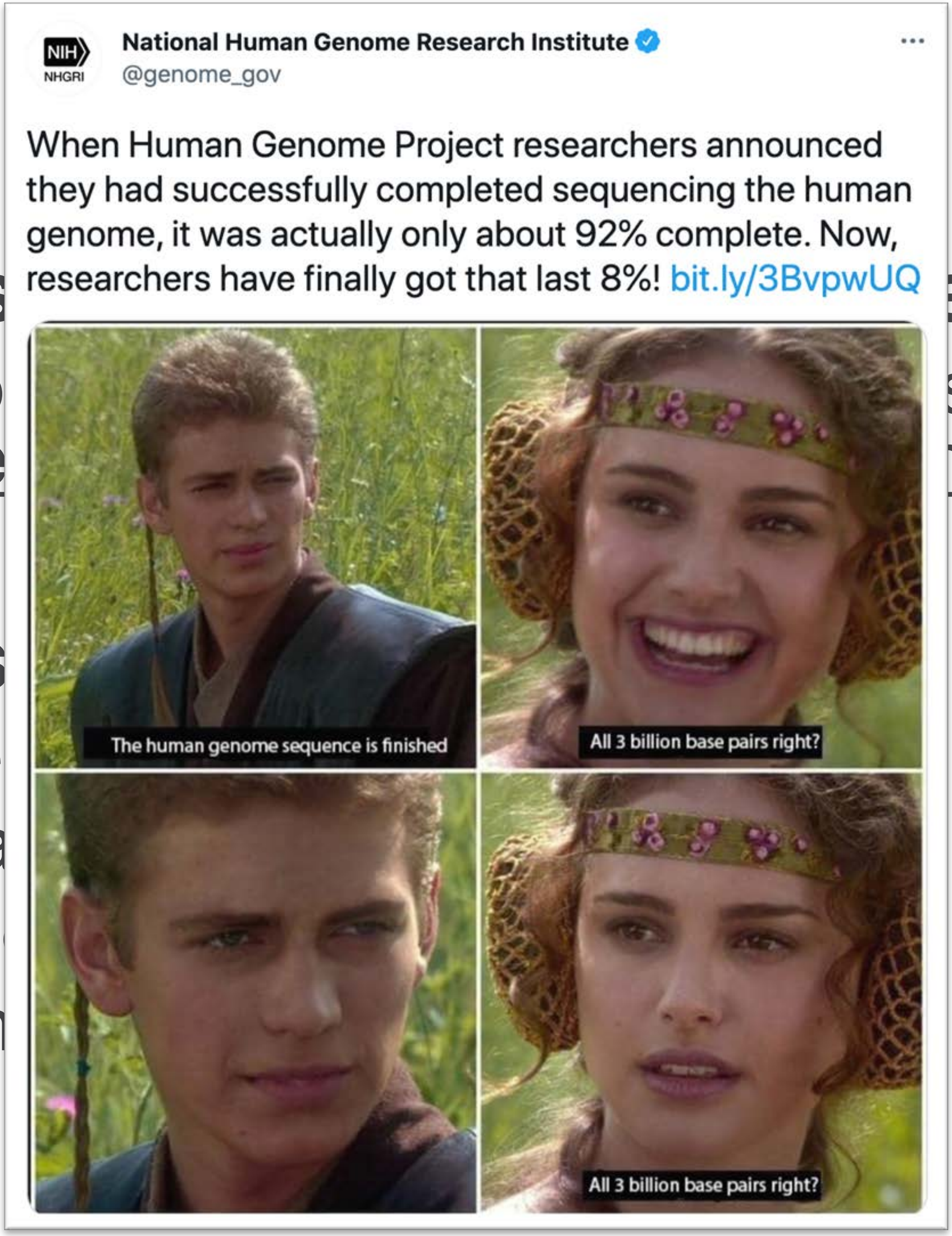
- **8% is missing or incorrect**
  - Centromeres and telomeres
  - Segmentally duplicated genes
  - Tandem gene arrays (e.g. rDNAs)
  - And an unknown number of errors…



NIH
NHGRI

No!

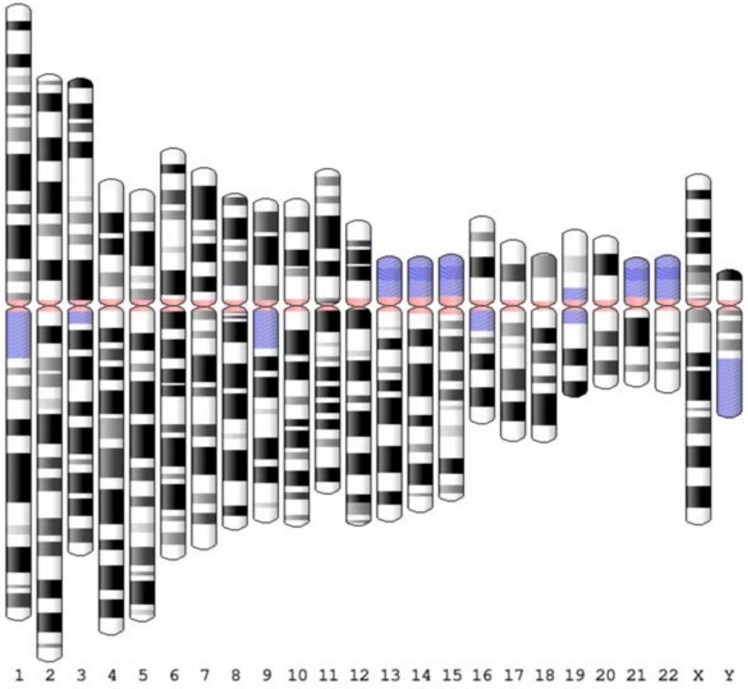
- **And what’s** **eciated**
  - “In the Ap ss than 400 gaps and 99 pe ned” (genome.gov)
- **8% is miss**
  - Centrome
  - Segmenta
  - Tandem g
  - And an un



National Human Genome Research Institute
@genome_gov

When Human Genome Project researchers announced they had successfully completed sequencing the human genome, it was actually only about 92% complete. Now, researchers have finally got that last 8%! bit.ly/3BvpwUQ

NIH NHGRI
The human genome sequence is finished
All 3 billion base pairs right?
All 3 billion base pairs right?



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Finishing the human genome

- **Why does it matter?**

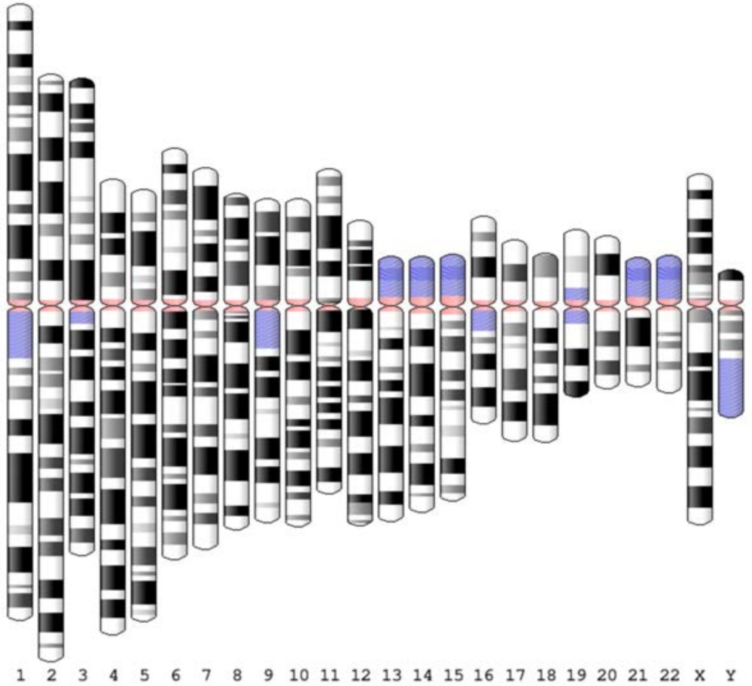
  Variation in these regions is unexplored

  Functional studies need sequence

  Reference gaps lead to artifacts

  We don’t know what we don’t know…

- **Why has it taken so long?**
  - Technological limitations
  - Genomic repeats



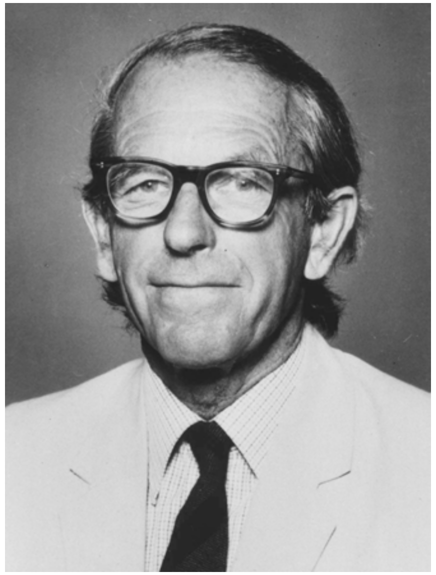
NIH
NHGRI

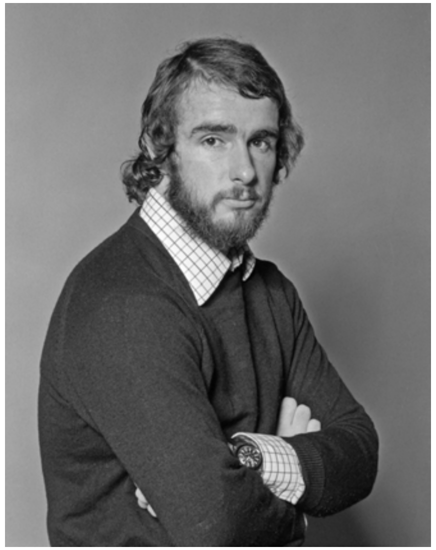
40 years of sequencing & assembly

"With modern fast sequencing techniques[1,2] and suitable computer programs it is now possible to sequence whole genomes without the need of restriction maps."

"If the overlap is of **sufficient length to distinguish it from being a repeat** in the sequence the two sequences must be contiguous."

— Rodger Staden, 1979





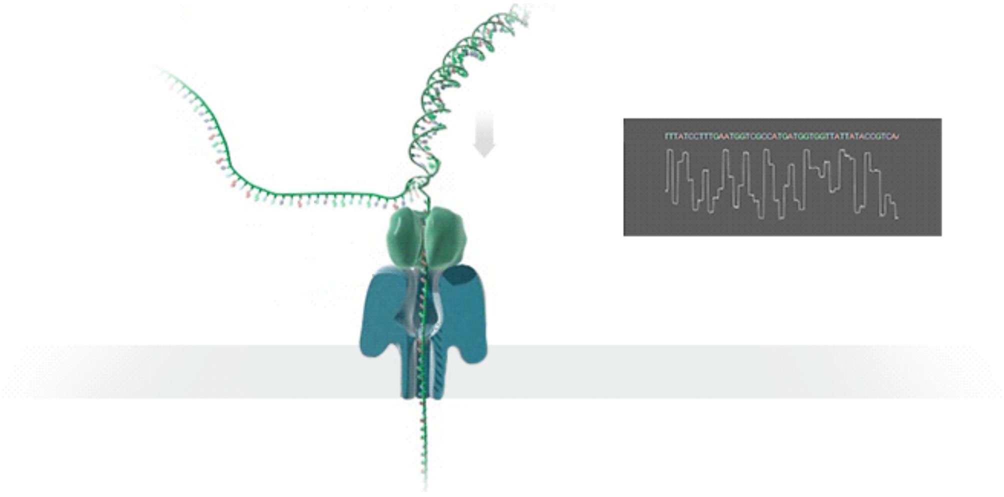
A strategy of DNA sequencing employing computer programs. Staden. *Nucleic Acids Research* (1979)

[1] Sanger and Coulson (1975), [2] Maxam and Gilbert (1977)

A new era of sequencing

Nanopore ultra-long sequencing

- **Nanopore UL**
  - >100 kb reads, up to 1 Mb
  - 95% (Q13) read quality
  - 99.9% (Q30+) assembly quality
- **Pros**
  - Outstanding length
  - Reads *span* repeats
- **Cons**
  - Lower throughput and quality



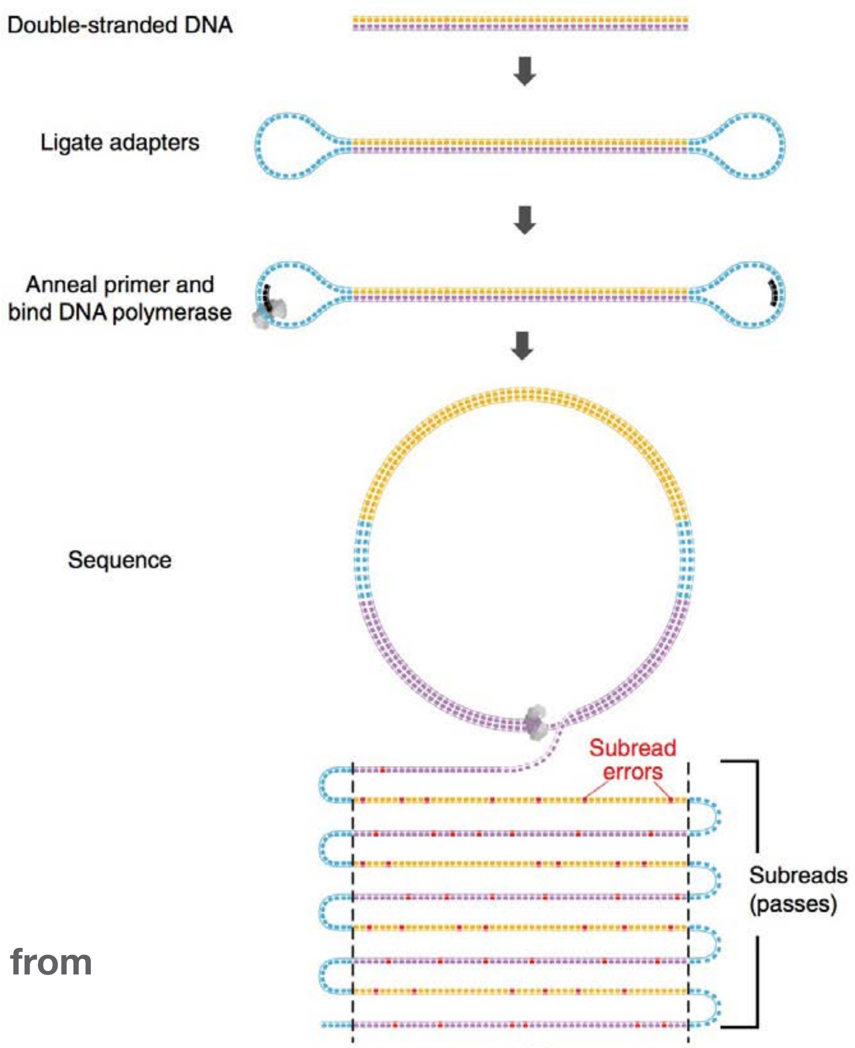
Nanopore sequencing and assembly of a human genome with ultra-long reads. Jain et al. *Nature Biotechnology* (2018)

Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Shafin et al. *Nature Biotechnology* (2020)

NIH NHGRI

Circular consensus sequencing

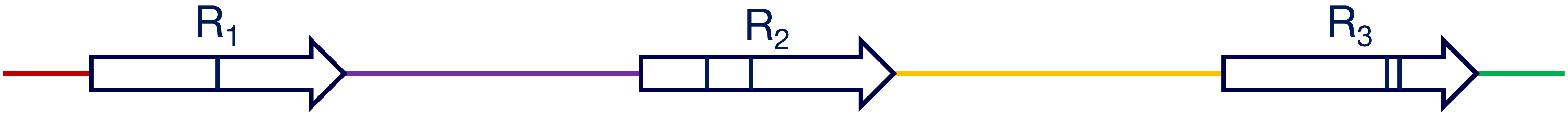
- **PacBio HiFi**
  - 20 kb reads
  - 99.9% (Q30) read quality
  - 99.9999% (Q60+) assembly quality
- **Pros**
  - Outstanding accuracy
  - Reads *distinguish* repeats
- **Cons**
  - Limited length and coverage



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

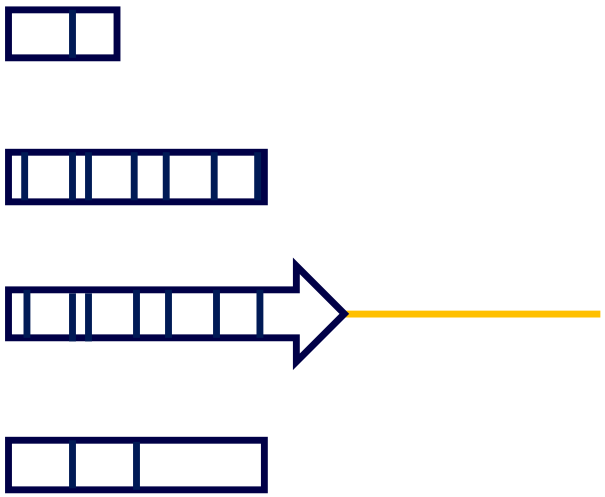
HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Nurk et al. *Genome Research* (2020)

“Sufficient length” depends on accuracy



- Where do the reads originate?

1. Illumina (short + accurate):
2. CLR (midsize + noisy):
3. Nanopore (long + noisy):
4. HiFi (midsize + accurate):

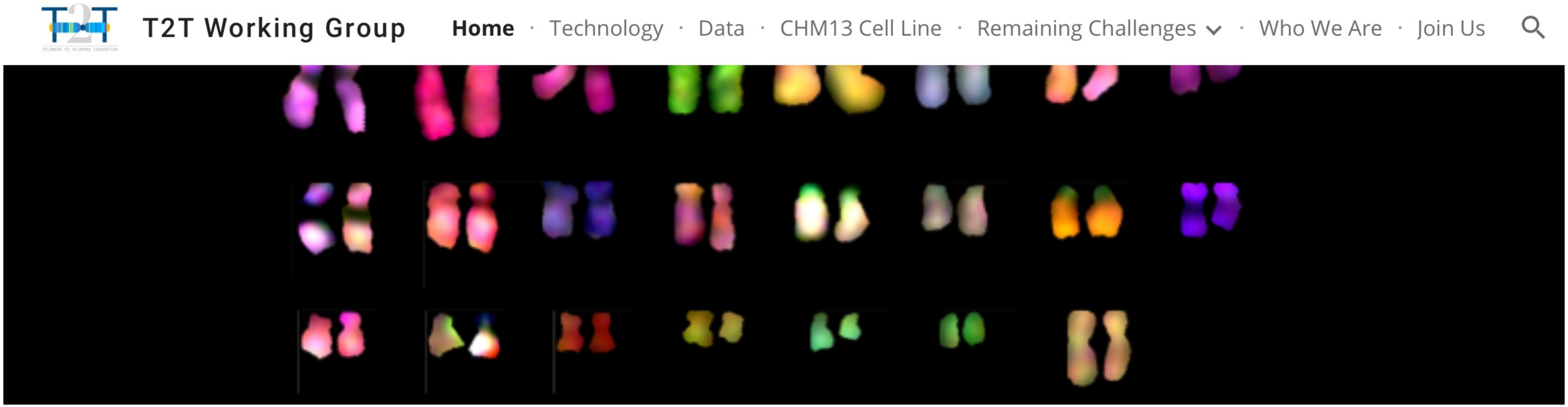


Depth of coverage matters for continuity too, but we’ll assume equivalent coverage for now…

Finishing the human genome

Let’s finish a human genome (2018)

Karen Miga, UCSC

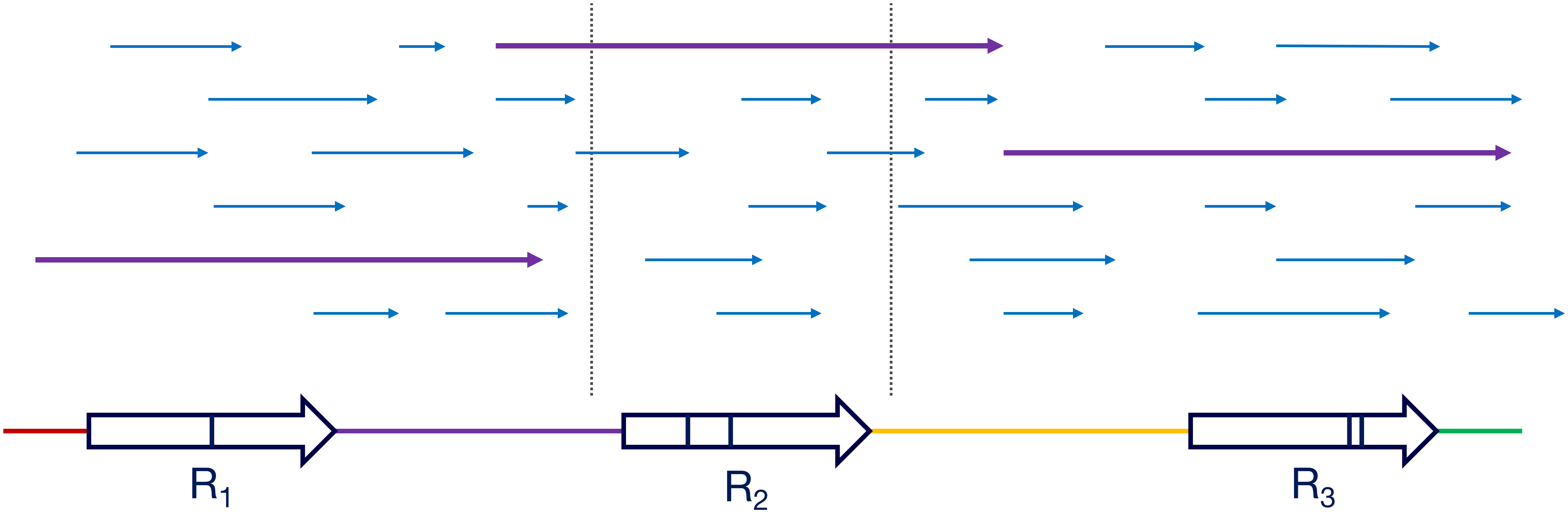


The Telomere-to-Telomere (T2T) consortium is an open, community-based effort to generate the first complete assembly of a human genome.

NIH NHGRI

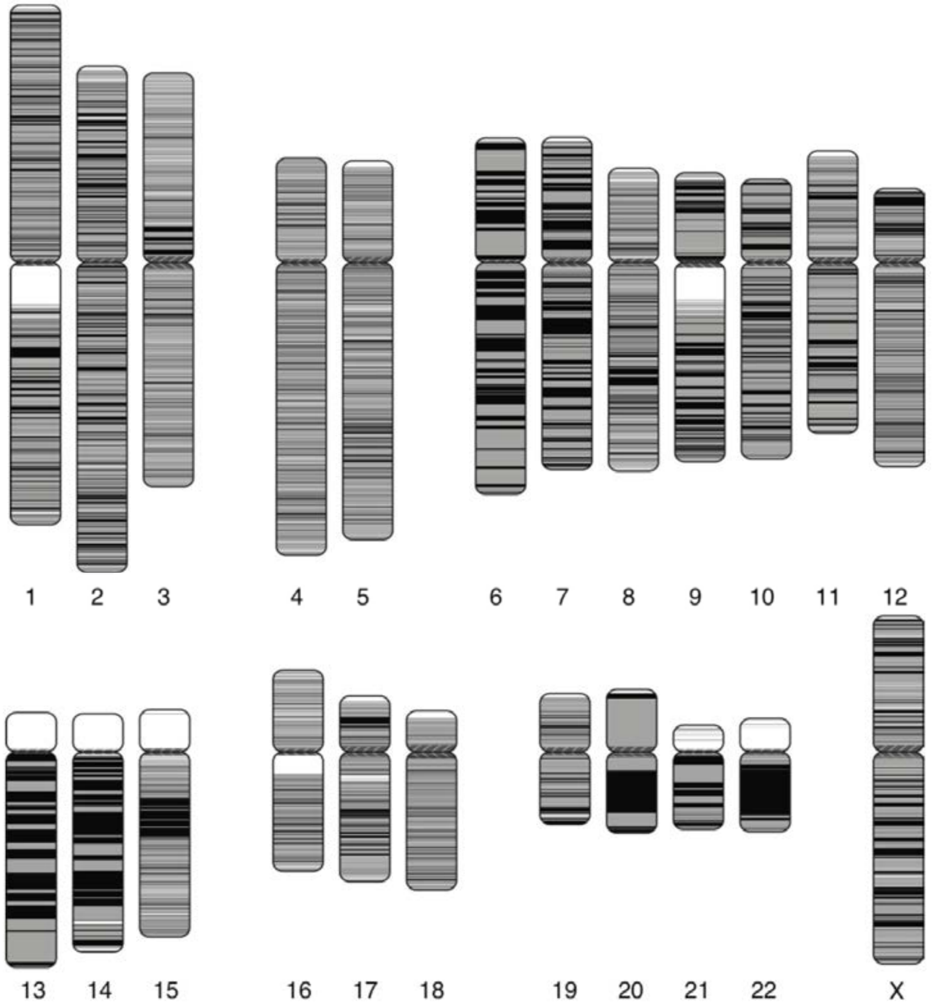
CHM13 homozygous 46,XX cell line from Urvashi Surti, Pitt; SKY karyotype from Jennifer Gerton, Stowers

Strategy: sequence the heck out of it



>120x Nanopore ultra-long coverage

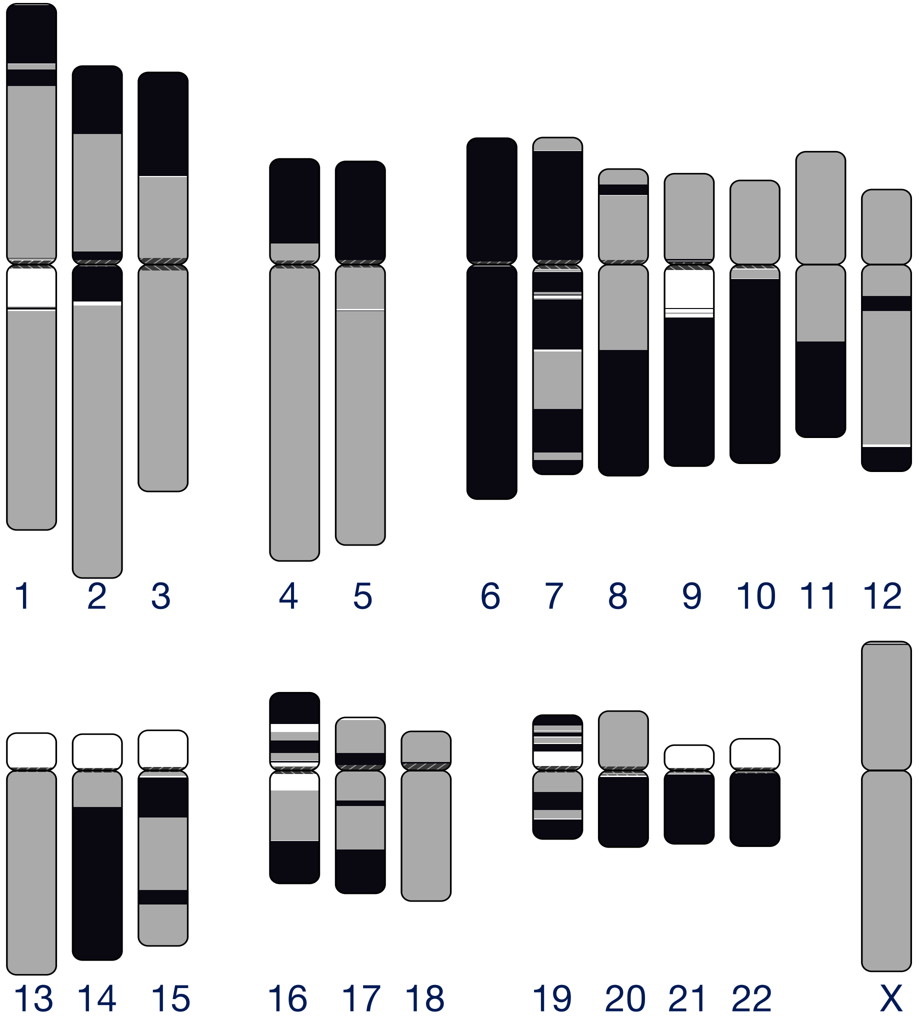
HGP/Sanger assembly (2001)



ref28 / hg10 : N50 0.5 Mbp

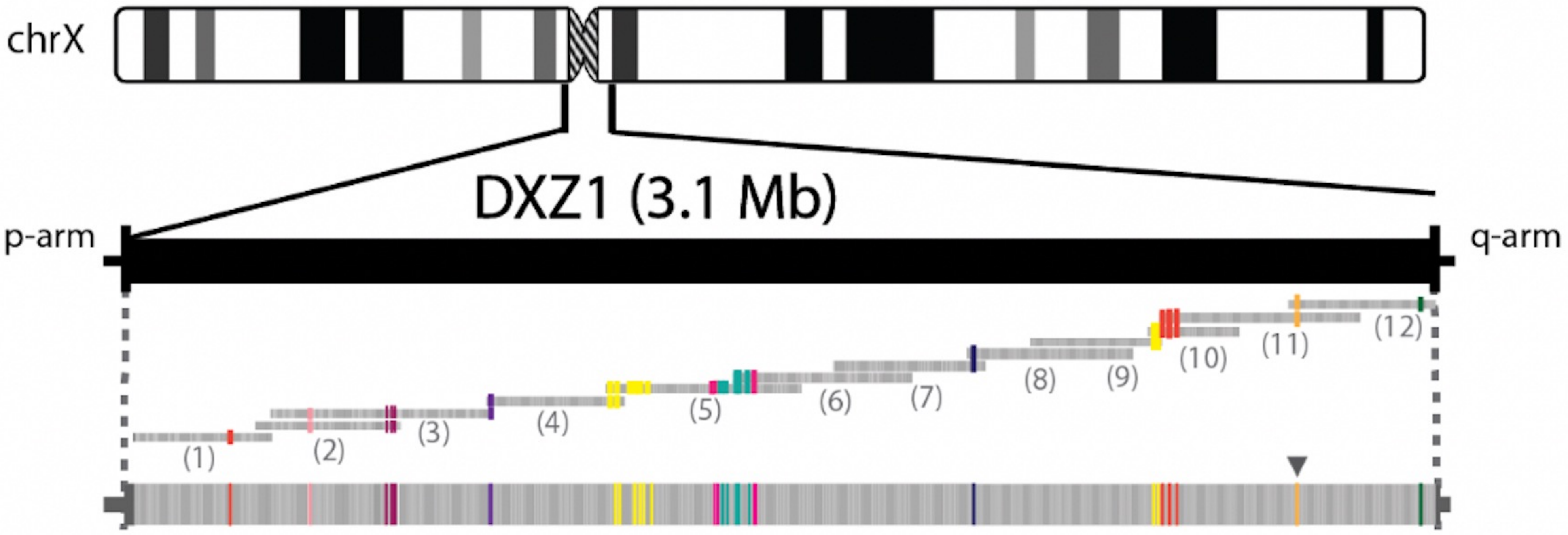
T2T/ONT assembly (2019)

Sergey Koren & Shelise Brooks, NHGRI



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

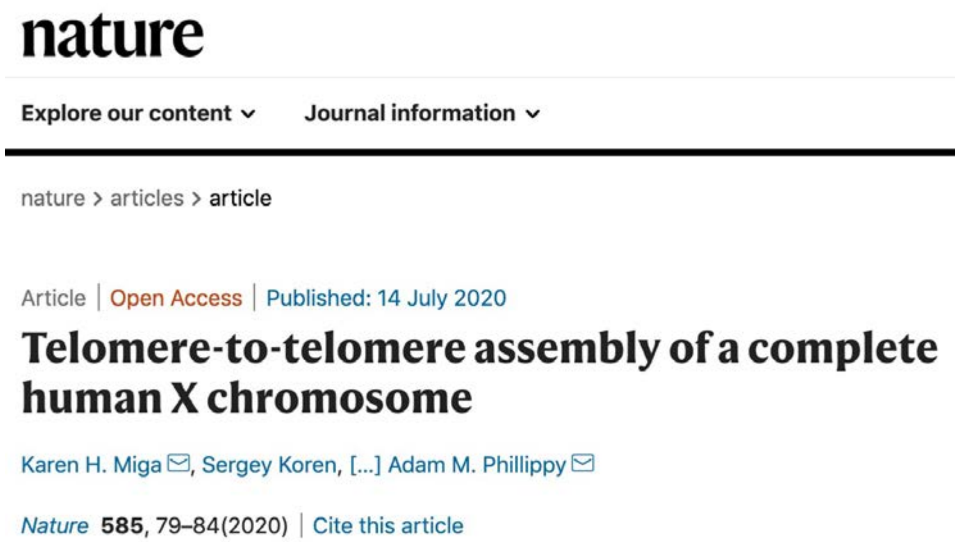
Nanopore backbone (2019)



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

Complete chromosomes X and 8!

Glennis Logsdon, UW



nature

Explore our content ⌄ Journal information ⌄

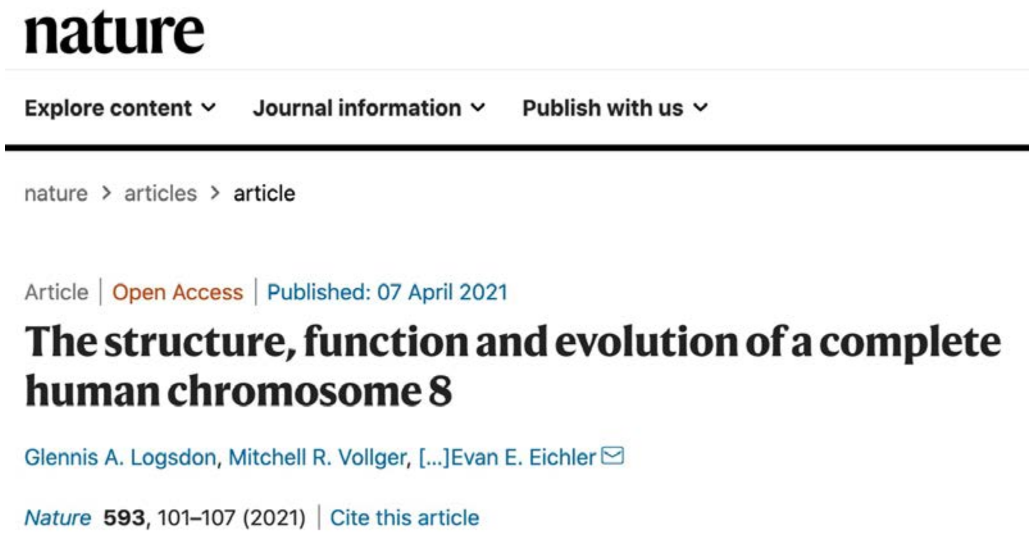
nature > articles > article

Article | Open Access | Published: 14 July 2020

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

Karen H. Miga, Sergey Koren, [...] Adam M. Phillippy

Nature **585**, 79–84(2020) | Cite this article



nature

Explore content ⌄ Journal information ⌄ Publish with us ⌄

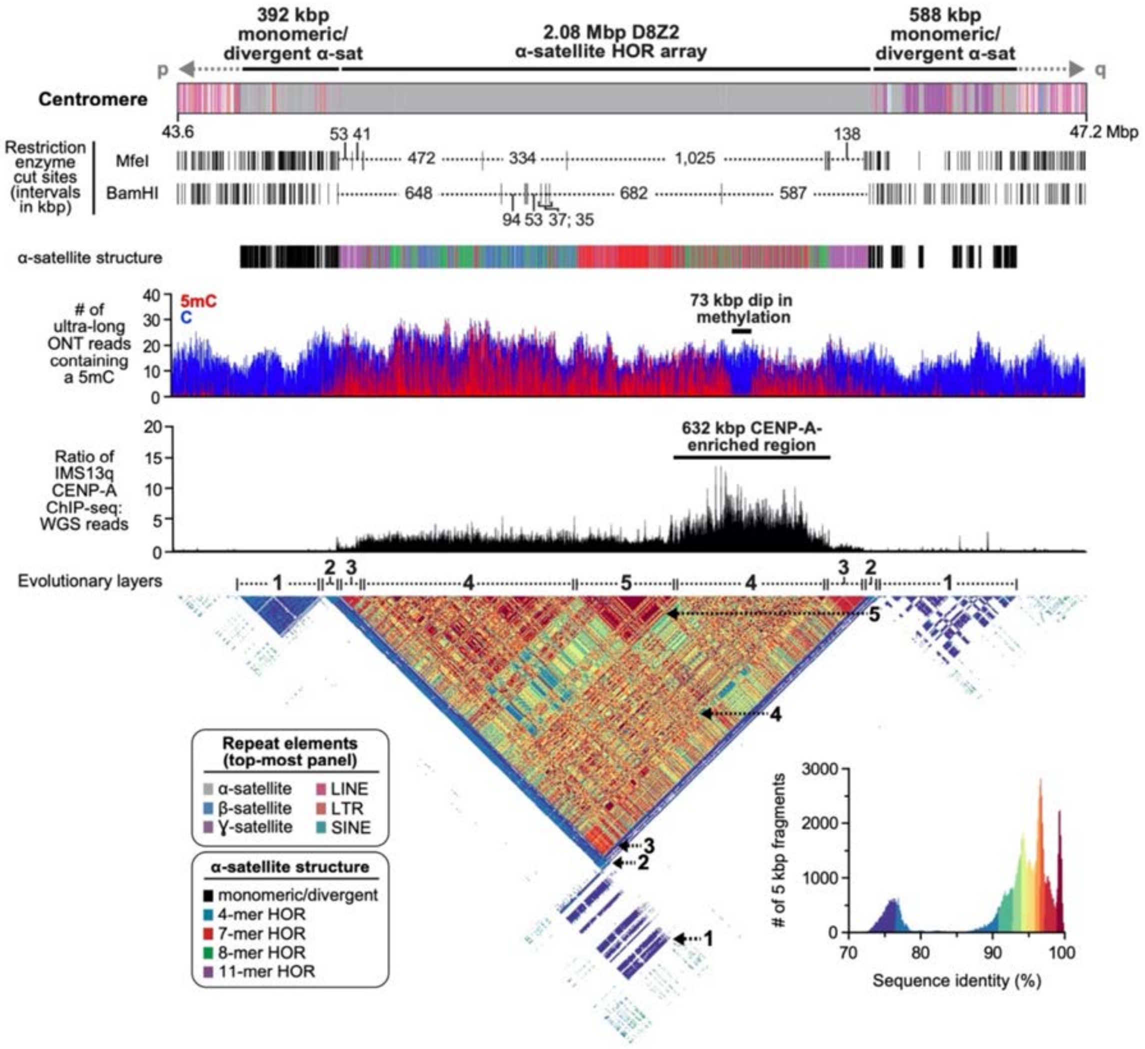
nature > articles > article

Article | Open Access | Published: 07 April 2021

The structure, function and evolution of a complete human chromosome 8

Glennis A. Logsdon, Mitchell R. Vollger, [...]Evan E. Eichler

Nature **593**, 101–107 (2021) | Cite this article



Can we speed this up?

A graph-first approach

Sergey Nurk, NHGRI

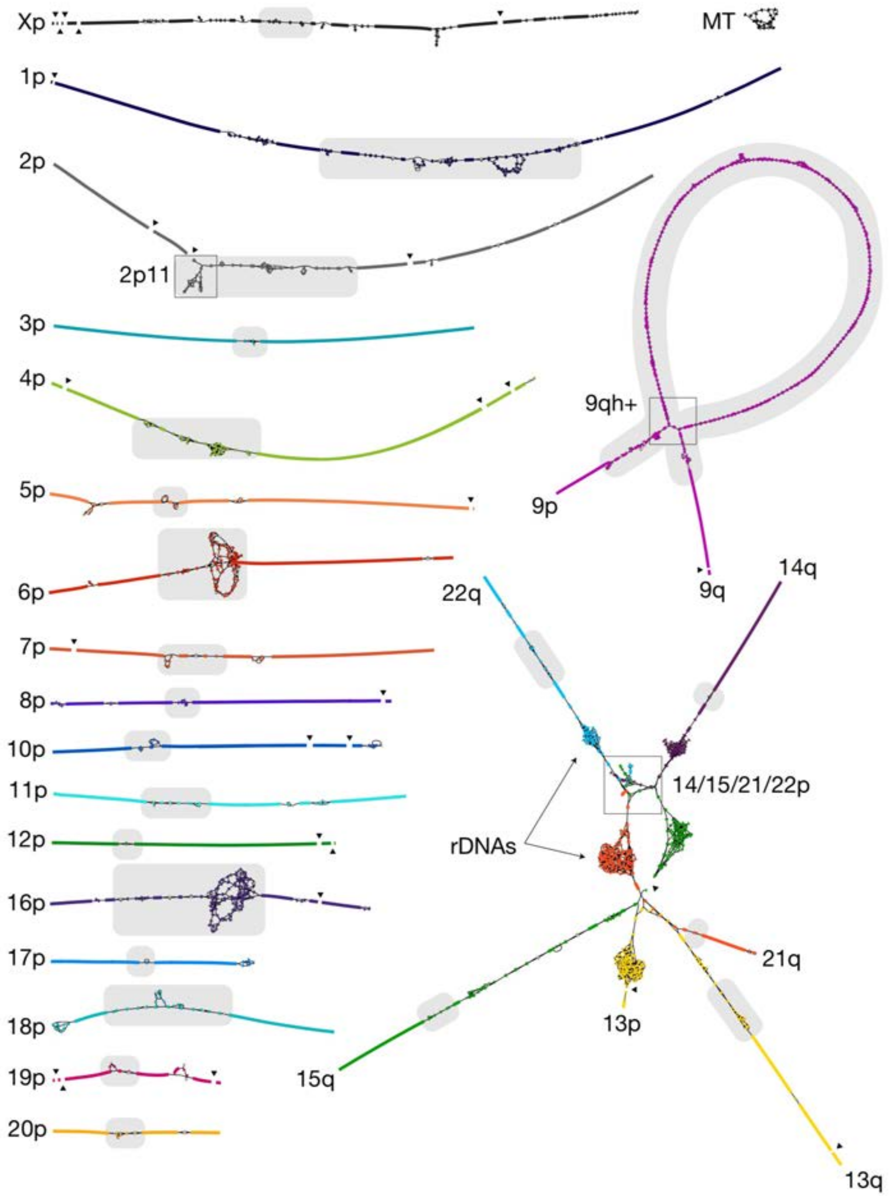
1. HiFi string graph
   - Homopolymer compression (CAAAAT → CAT)
   - Read cleaning and correction
   - String graph from long *perfect* overlaps
2. Hamiltonian walks for easy tangles
3. Nanopore walks for hard tangles
4. Use only HiFi for consensus (decompression)

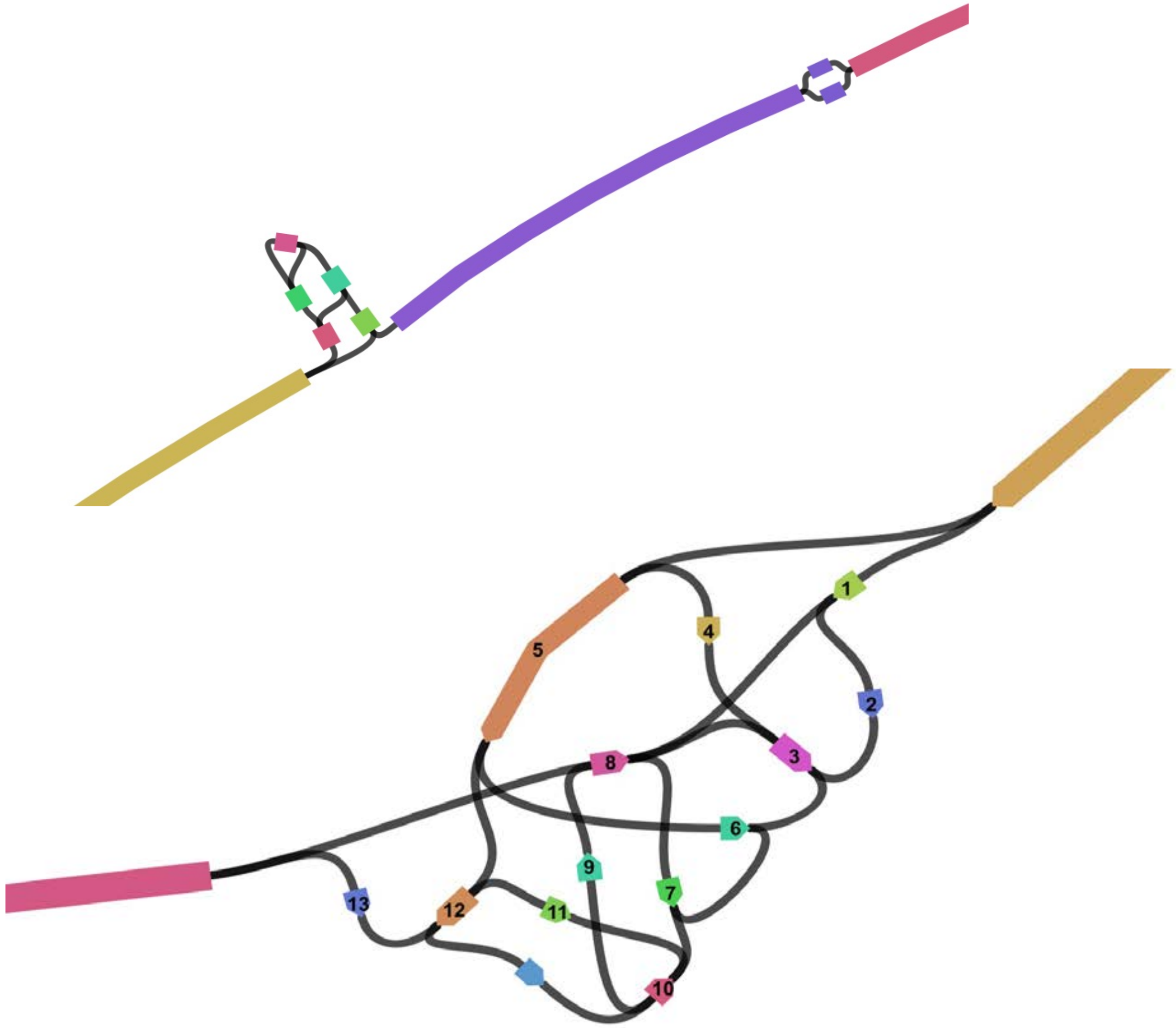
NIH NHGRI

HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Nurk et al. *Genome Research* (2020)

CHM13 HiFi assembly graph (2020)

Mikko Rautiainen, NHGRI

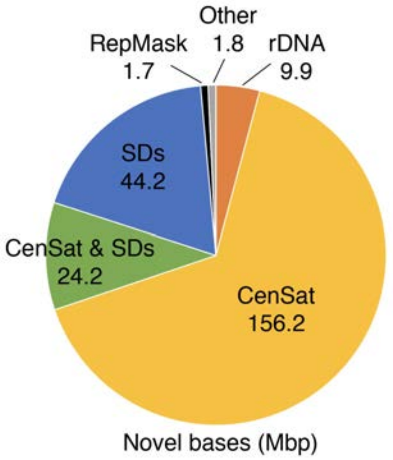


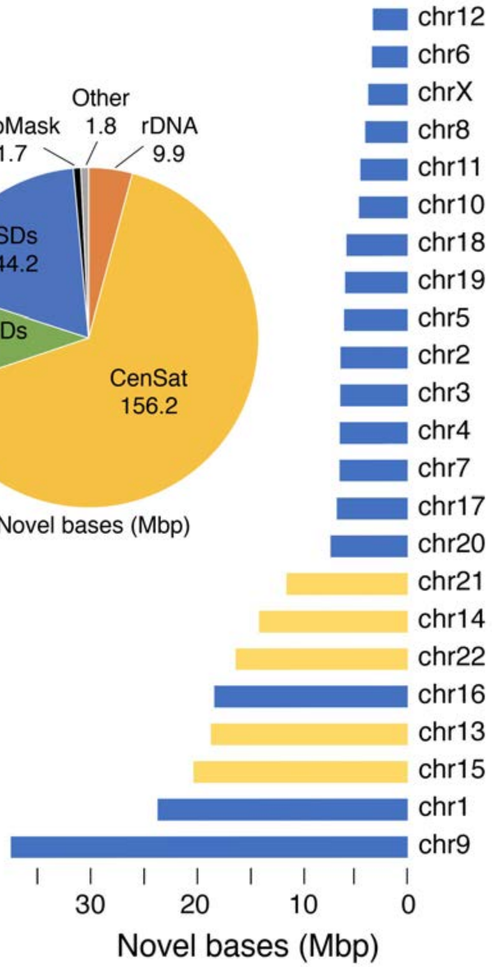


One year later…

The *complete* sequence of a human genome

- **GRCh38.p13 (no alts)**
  - 24 chromosomes
  - 42 unlocalized
  - 127 unplaced
  - 2,922,212,712 bp
  - 130.6 Mbp of gaps
  - Uncertain quality
- **CHM13v1.1 (no hets)**
  - 23 chromosomes (no Y)
  - 0 unlocalized
  - 0 unplaced
  - 3,054,832,041 bp
  - No gaps
  - ~Q70, no known SVs





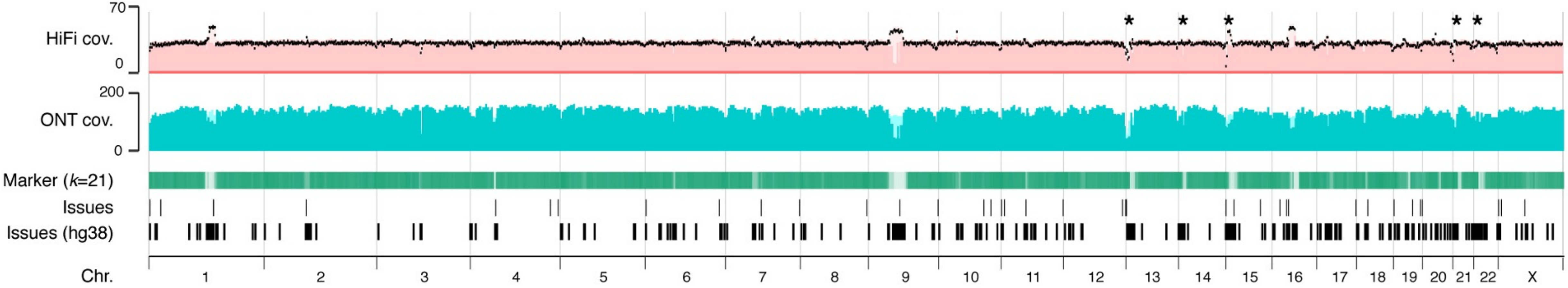
Estimated CHM13 genome size of **3.055 Gbp**
>200 Mbp of *new* sequence vs. GRCh38

2,226 new genes (115 predicted protein coding)

NIH
NHGRI

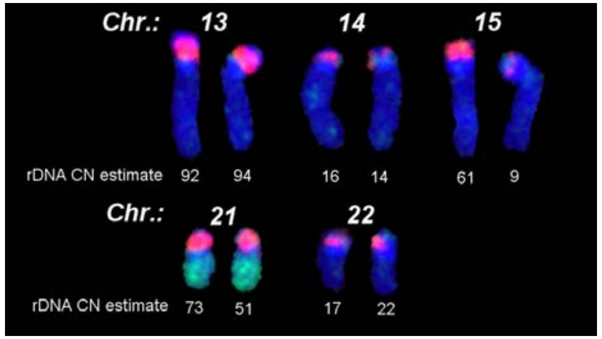
CHM13 assembly validation

Arang Rhie, NHGRI



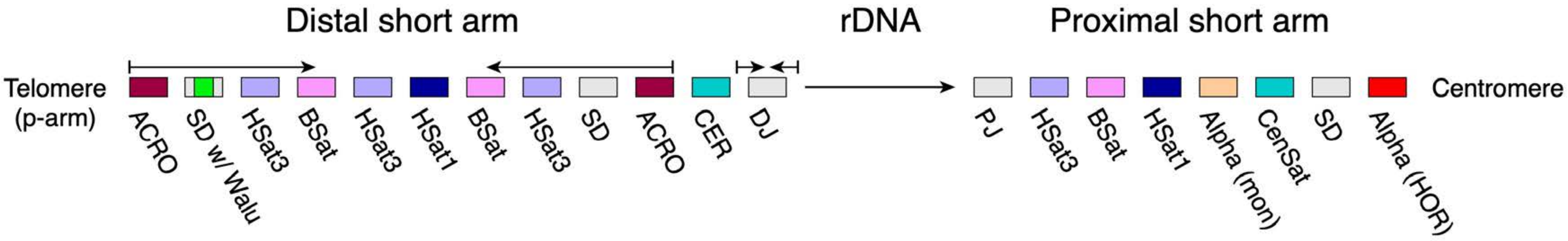
Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies.
Mc Cartney et al. *bioRxiv* (2021)

The acrocentrics revealed

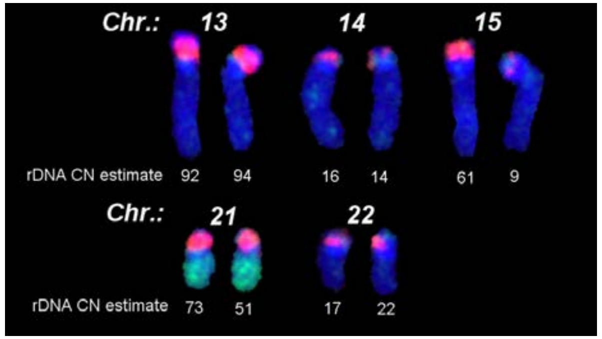

Chr.: 13 14 15
rDNA CN estimate 92 94 16 14 61 9
Chr.: 21 22
rDNA CN estimate 73 51 17 22

Tamara Potapova, Stowers

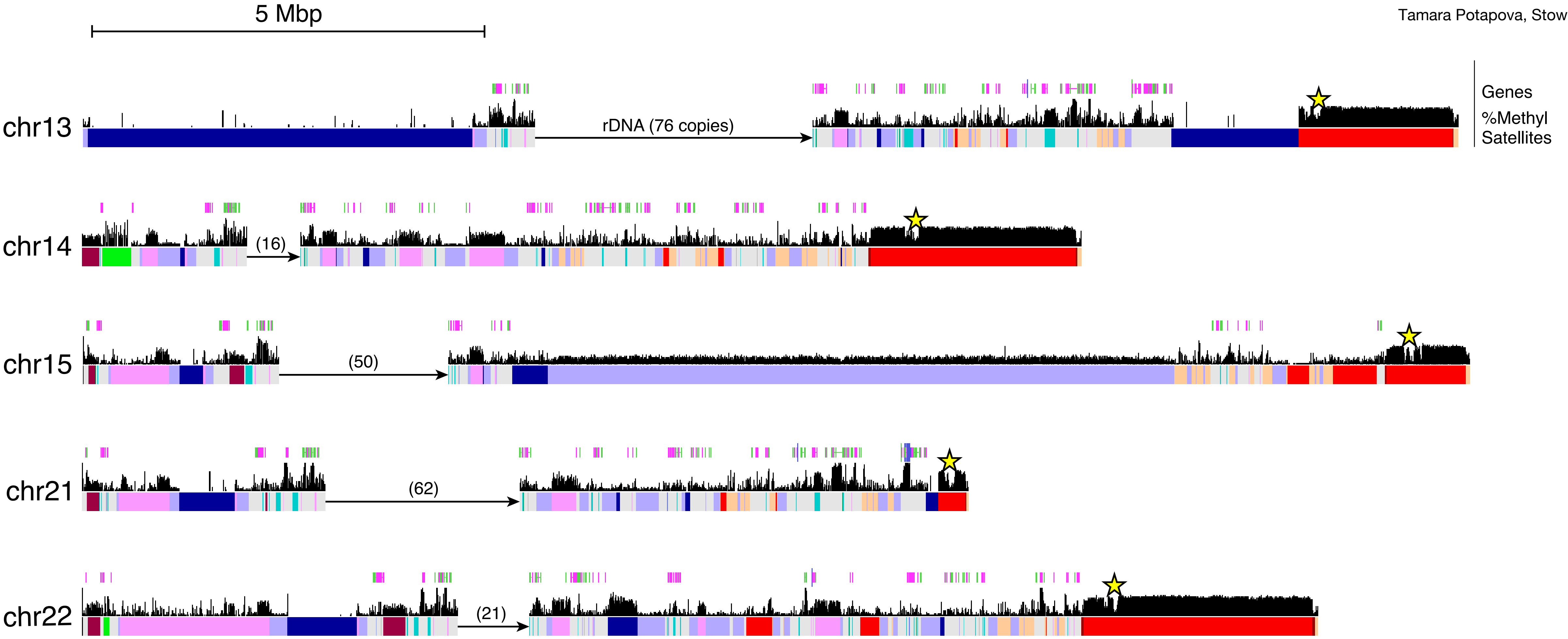
- 66.1 Mbp of new sequence
- Dynamic sources of segmental duplication
- Median inter-chromosomal identity 98.7%
- No unique 5 kbp windows at 80% identity
- 96% can be found elsewhere in the genome



The acrocentrics revealed

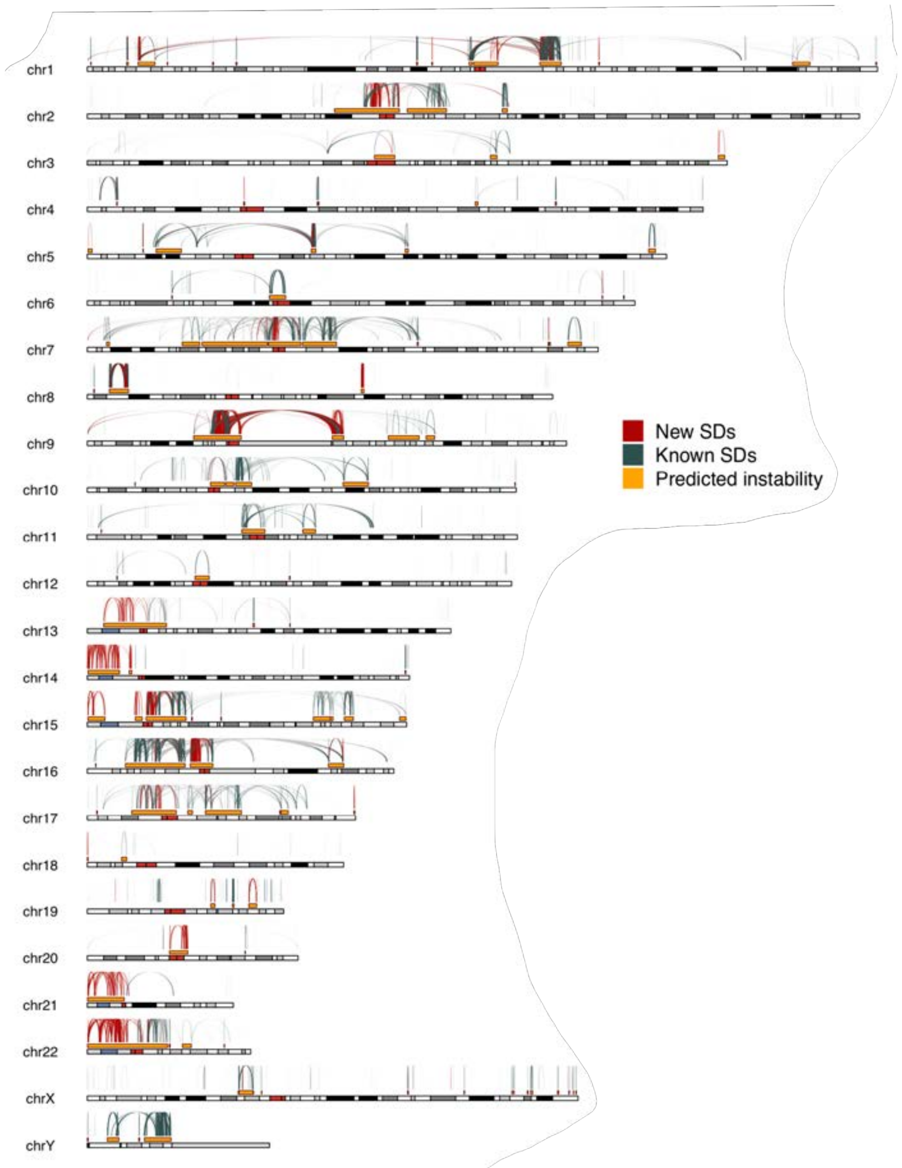

Chr.: 13 14 15
rDNA CN estimate 92 94 16 14 61 9
Chr.: 21 22
rDNA CN estimate 73 51 17 22

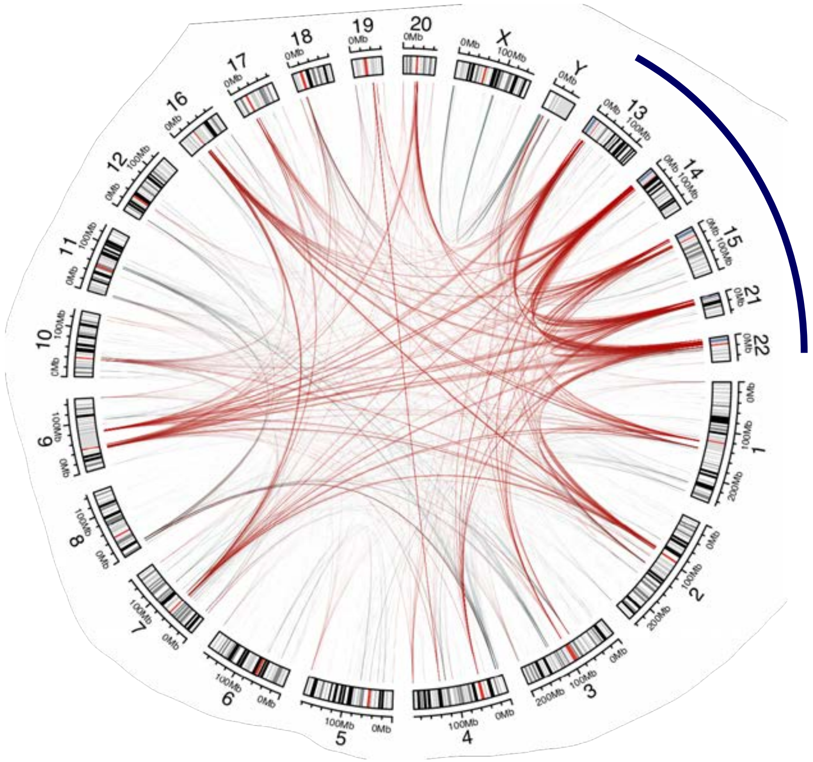
Tamara Potapova, Stowers

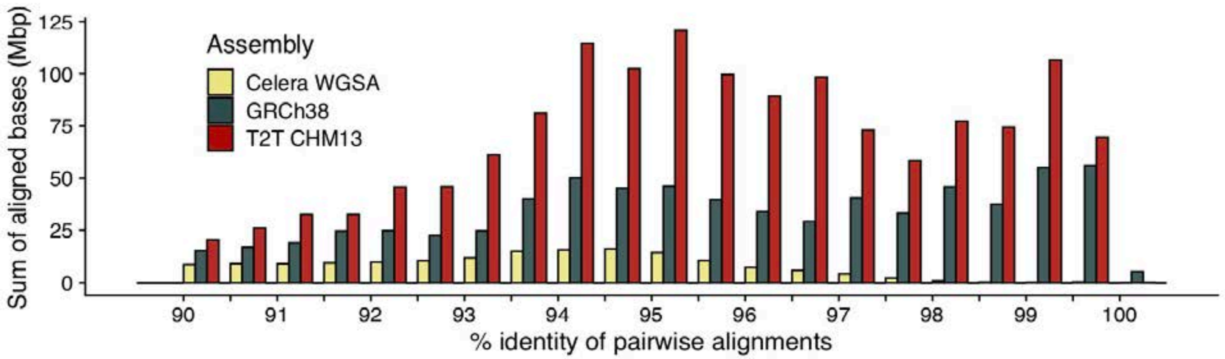


Many new segmental duplications

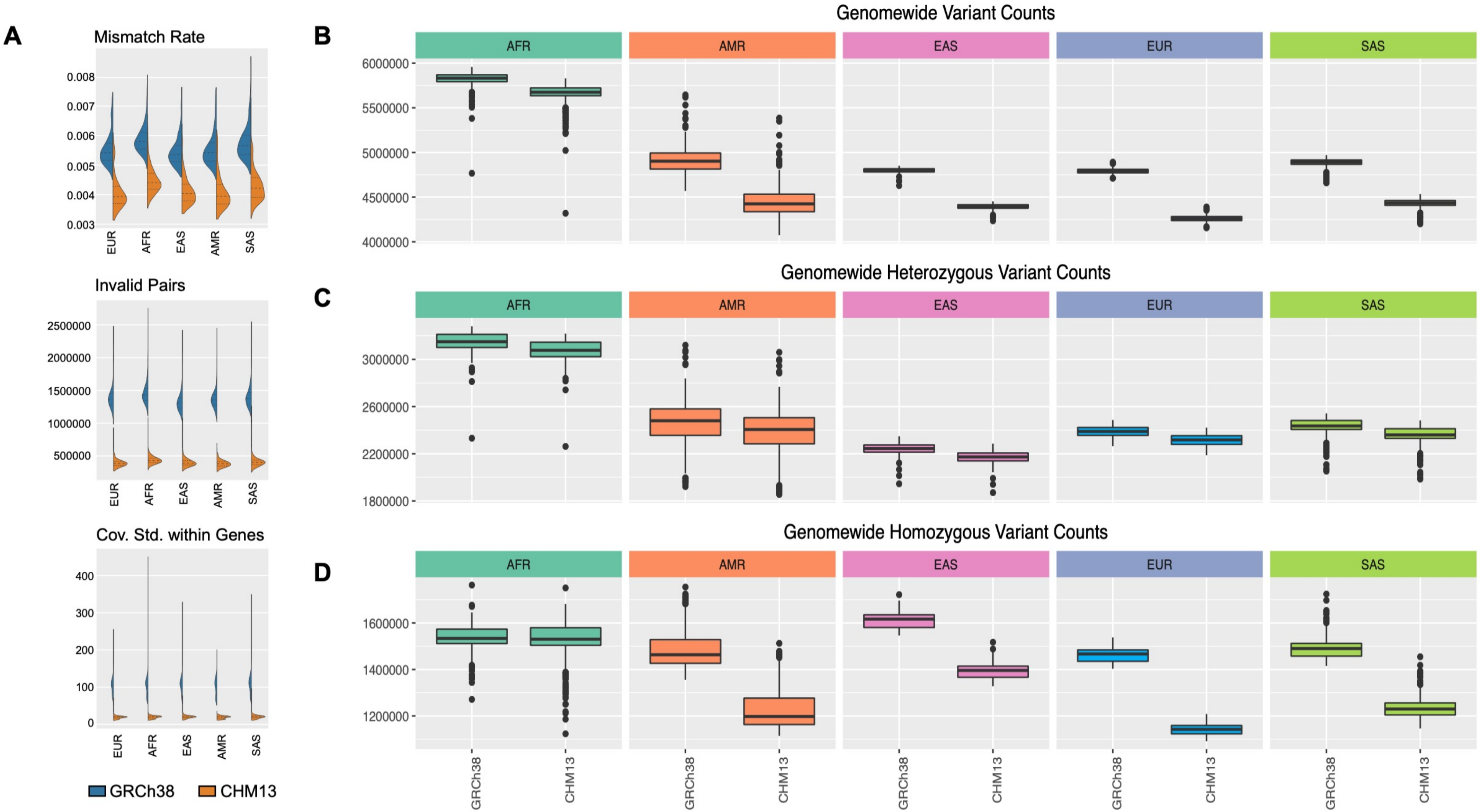
Mitchell Vollger, UW





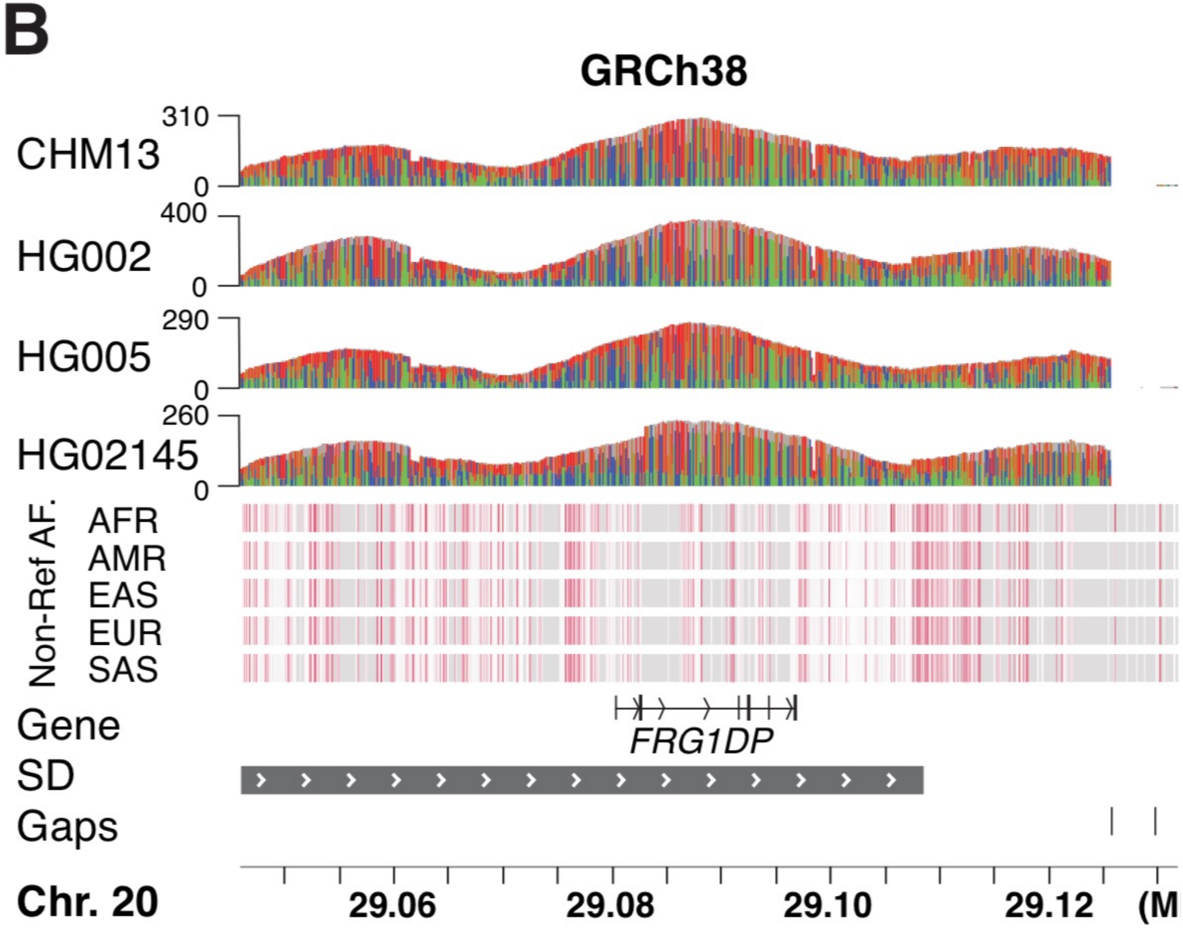


A more accurate reference sequence



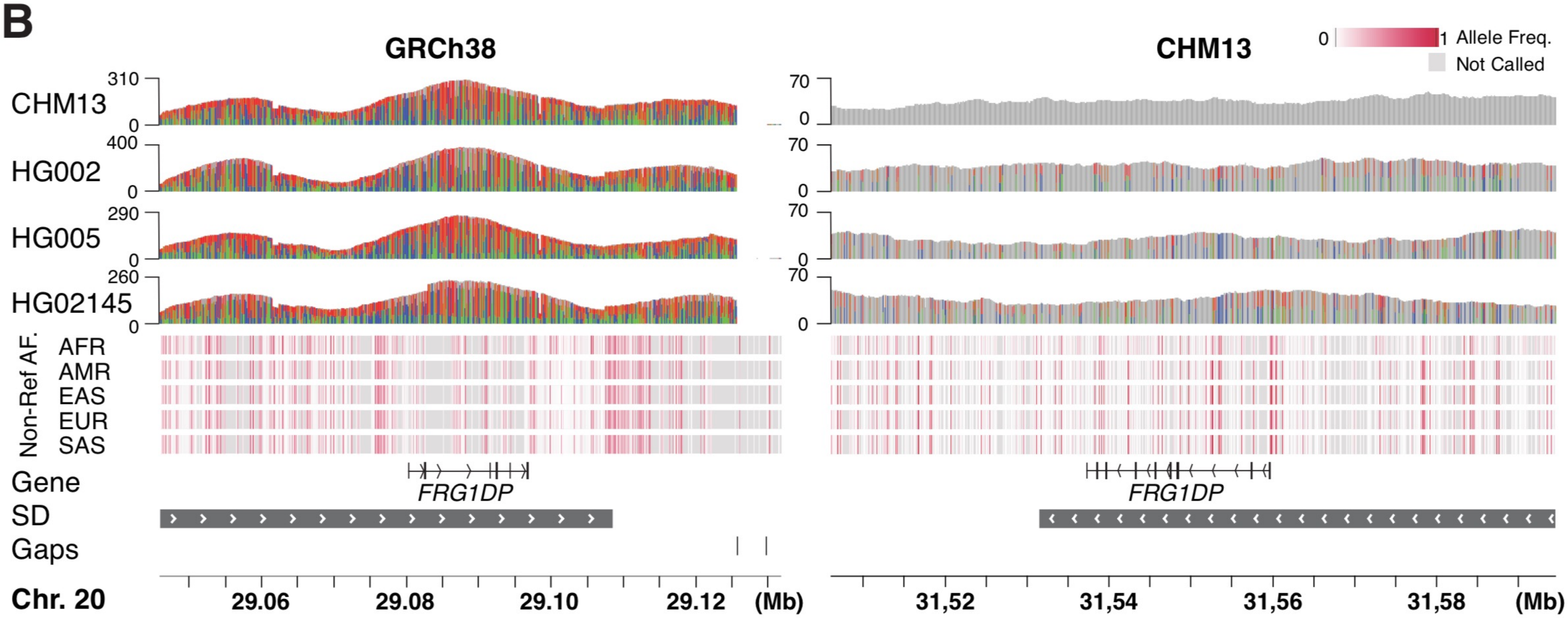
Aganezov *et al.*

Newly resolved paralogs fix old ones



Paralogs of FSHD Region Gene 1 (FRG1). 23 paralogs in CHM13, only 9 in GRCh38.

Newly resolved paralogs fix old ones



Paralogs of FSHD Region Gene 1 (FRG1). 23 paralogs in CHM13, only 9 in GRCh38.

Compared to GRCh38, CHM13…

- Is a *complete* genome
- Represents a natural haplotype
- Corrects systematic errors in GRCh38 (SVs, dups)
- Improves both long and short read mapping
- Eliminates >10k false variants per sample*
- Identifies >2M new variants in 1000G datasets
- Adds ~2,000 new genes (~100 protein coding)

Including 12-fold reduction in false positives for 269 medically relevant genes

T2T bioRxiv preprints

TELOMERE-TO-TELOMERE CONSORTIUM

The complete sequence of a human genome
Nurk, Koren, Rhie, Rautiainen, Eicher, Miga,, Phillippy, *et al.*

Complete genomic and epigenetic maps of human centromeres
Altemose, Alexandrov, Miga, *et al.*

Segmental duplications and their variation in a complete human genome
Vollger, Eichler, *et al.*

Epigenetic patterns in a complete human genome
Gershman, Miga, Timp, *et al.*

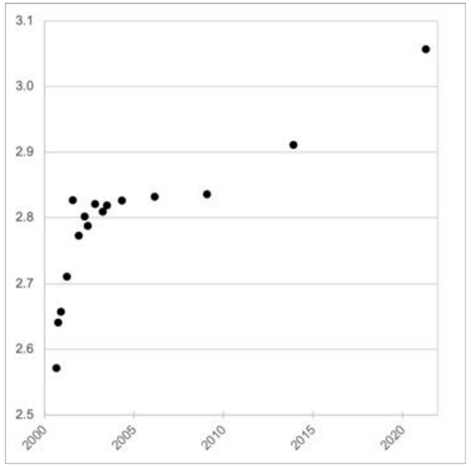
A complete reference genome improves analysis of human genetic variation
Aganezov, Yan, Soto, Kirsche, Zarate, McCoy, Dennis, Zook, Schatz, *et al.*

The transcriptional and epigenetic state of human repeat elements
Hoyt, O'Neill, *et al.*

Summary thoughts

TELOMERE-TO-TELOMERE CONSORTIUM

- **The human genome is *finally* complete**
  - The most bases ever added to the genome
  - CHM13 is a better reference for mapping
  - More variants within repeats than expected
  - New genes and structures uncovered



- **PacBio HiFi is a powerful new data type**
  - Accurate, yet continuous, assembly graphs
  - Nanopore and/or Hi-C for gaps, tangles, and phasing

NIH
NHGRI

What is *the* reference?

- **GRC if you must**
  - 20 years of accumulated resources
- **T2T for everything else**
  - Improved accuracy and reduced bias
  - Only option for 8% of the genome
- **Pangenome for the future**
  - Complete catalog of human genomic variation

What’s next for the T2T?

- **Y chromosome**
  - Coming (very) soon!
- **Human Pangenome Reference Consortium**
  - 250+ *diploid* HiFi genomes
  - Reference pangenome data structures
  - UCSC, UW, WashU, Rockefeller, NHGRI…
  - https://github.com/human-pangenomics/
- **ModT2T**
  - Zebrafish, fly, mouse, primates…

HGP started it, T2T finished it

T2T
TELOMERE-TO-TELOMERE CONSORTIUM

DNAnexus

amazon web services

PACBIO®

Oxford NANOPORE Technologies

ARIMA GENOMICS

NIH
NHGRI