What’s new in the human (pan)genome

Adam M. Phillippy

CSHL Advanced Sequencing Technologies & Bioinformatics Analysis Course
November 8, 2023

@aphillippy.bsky.social 

National Human Genome Research Institute

The Forefront of Genomics®

A 20-year anniversary

Finishing the euchromatic sequence of the human genome

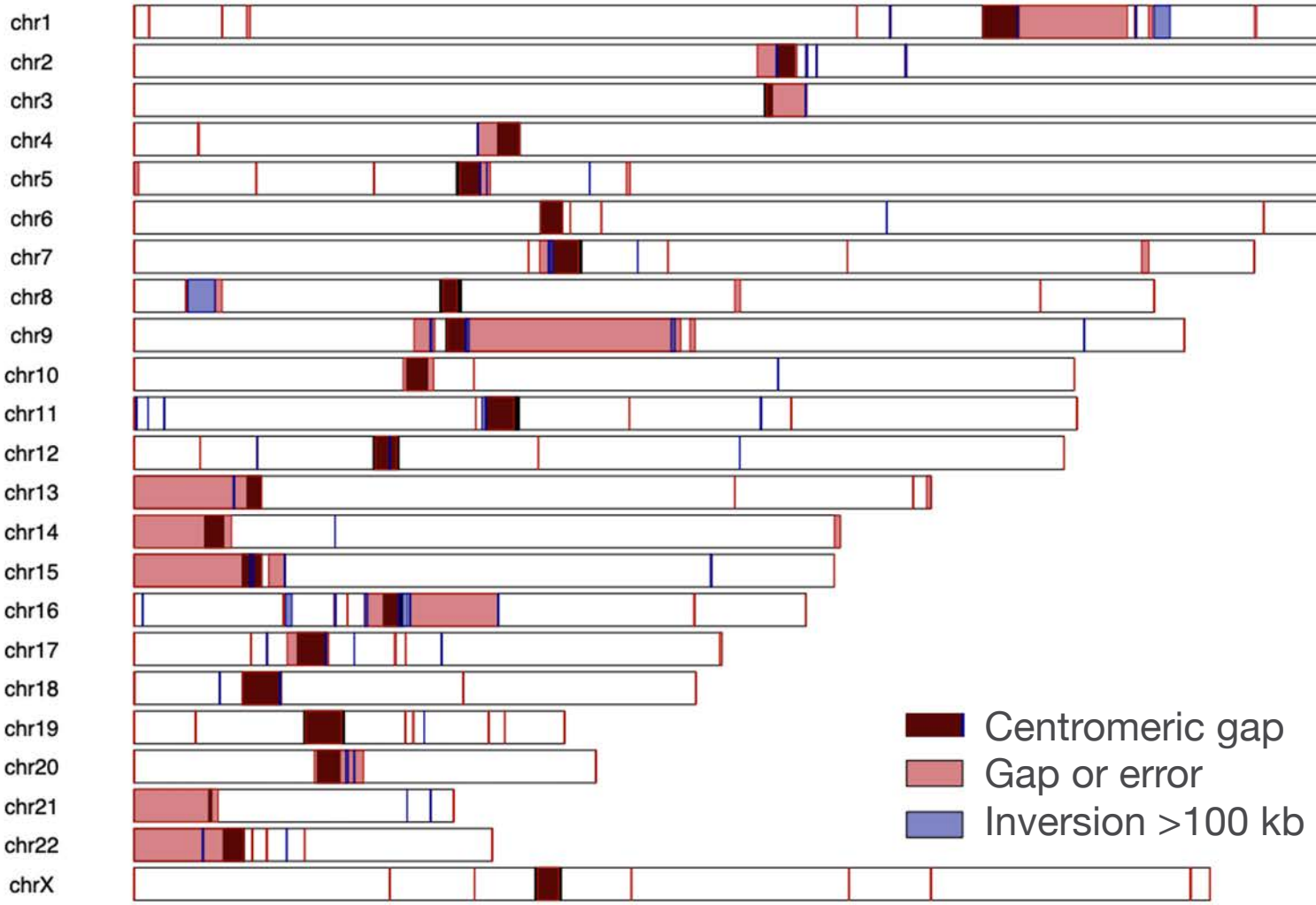
International Human Genome Sequencing Consortium*

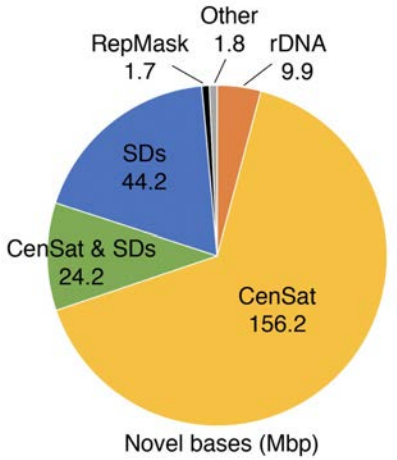
**A list of authors and their affiliations appears in the* Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

NIH NHGRI

What's new? Segdups and Satellites





What's there? Repeats!

- Centromeres
- Telomeres
- Segmental duplications
- Tandem gene arrays
- Acrocentric p-arms
- >200 Mbp of new sequence
- ~2,000 new genes predicted

NIH NHGRI

The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

Human reference then vs. now

- **2003 (hg35)**
- 2.85 Gbp
- 341 gaps
- 1 error per 100,000 (Q50)
- $5,000,000,000

- **2023 (chm13)**
- 3.12 Gbp
- 0 gaps
- 1 error per 10,000,000 (Q70)
- $10,000

- Mosaic, under-representative of segmental duplications, structural errors

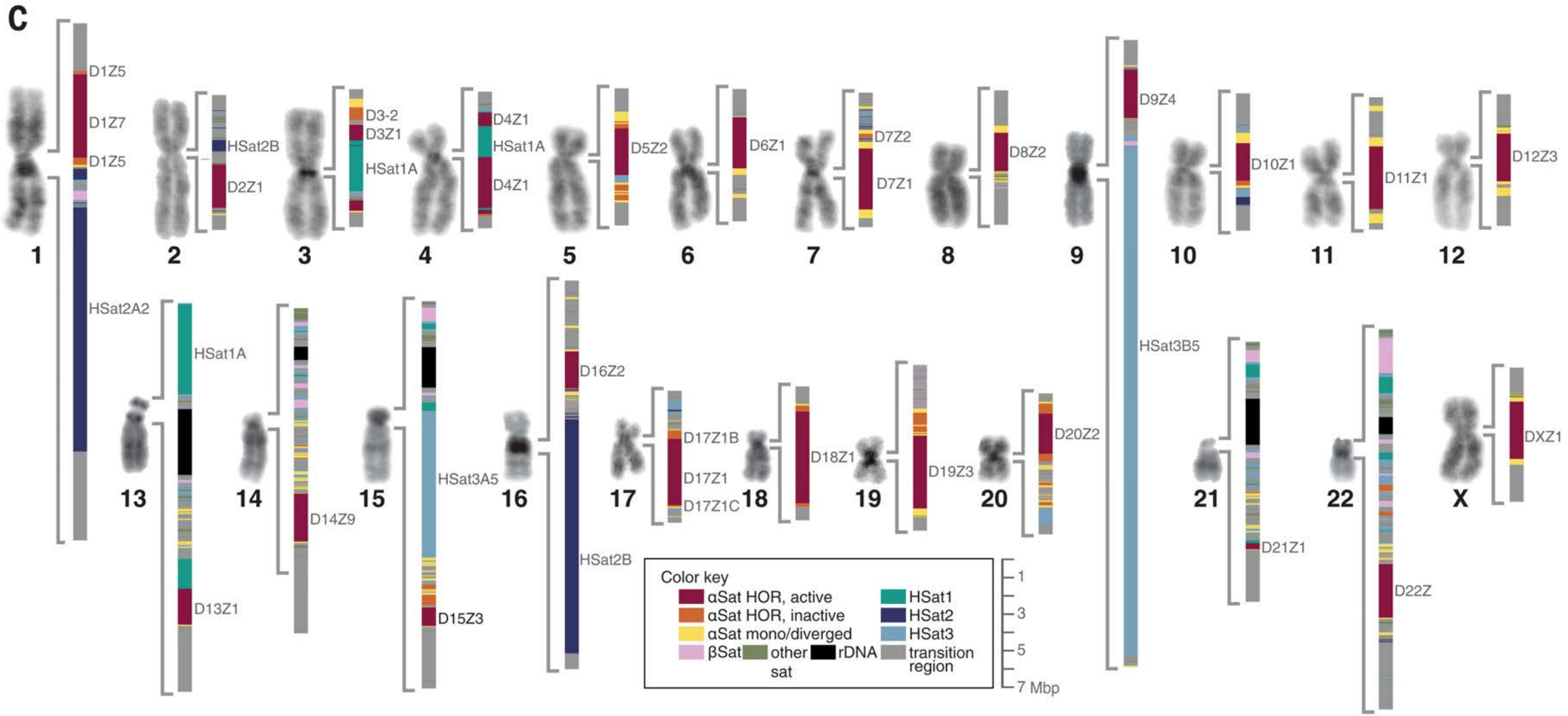
- 10-fold more representative of copy number and 3-fold more for inversion orientation

*than hg38

NIH NHGRI

The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

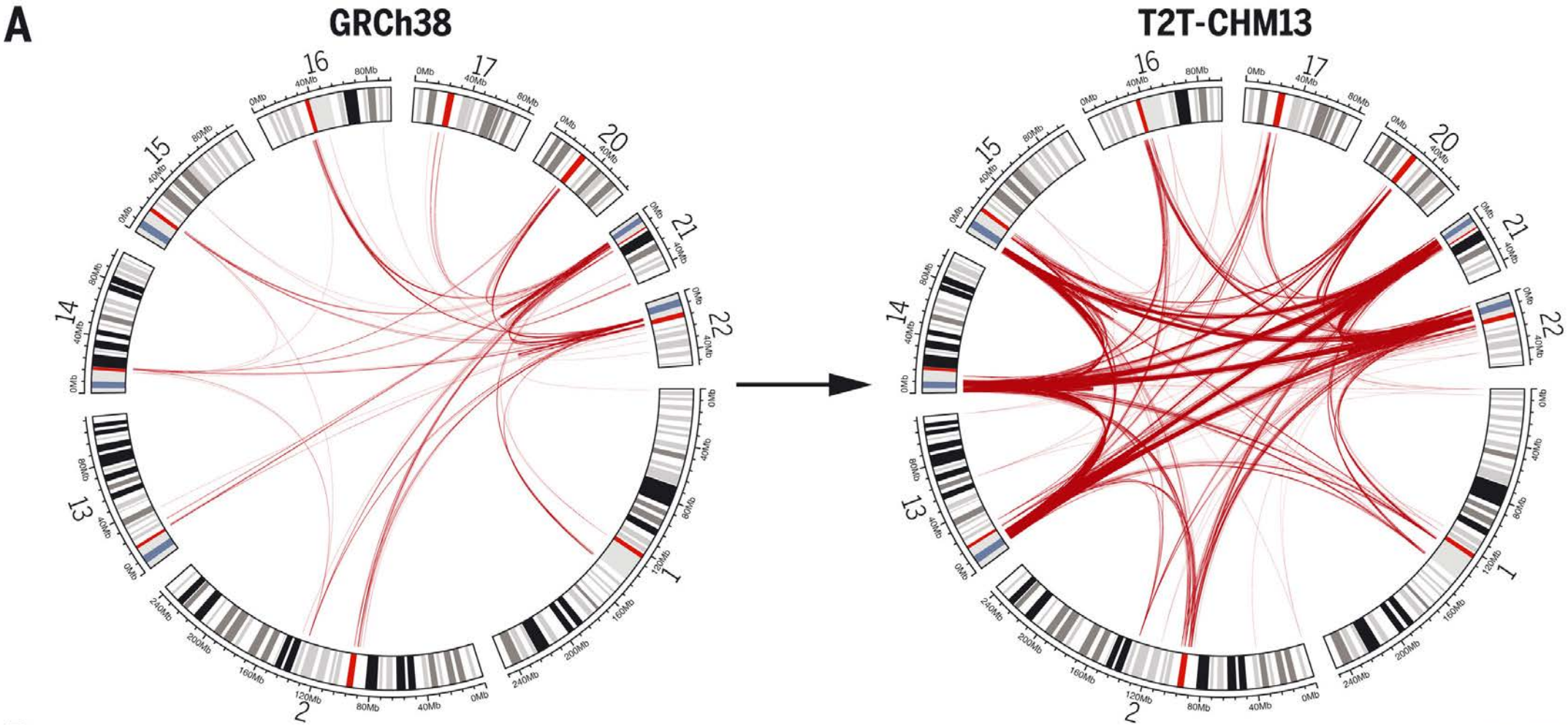
New satellites in CHM13



Complete genomic and epigenetic maps of human centromeres.
Altemose, et al. *Science* (2023)

NIH NHGRI

New segdups in CHM13



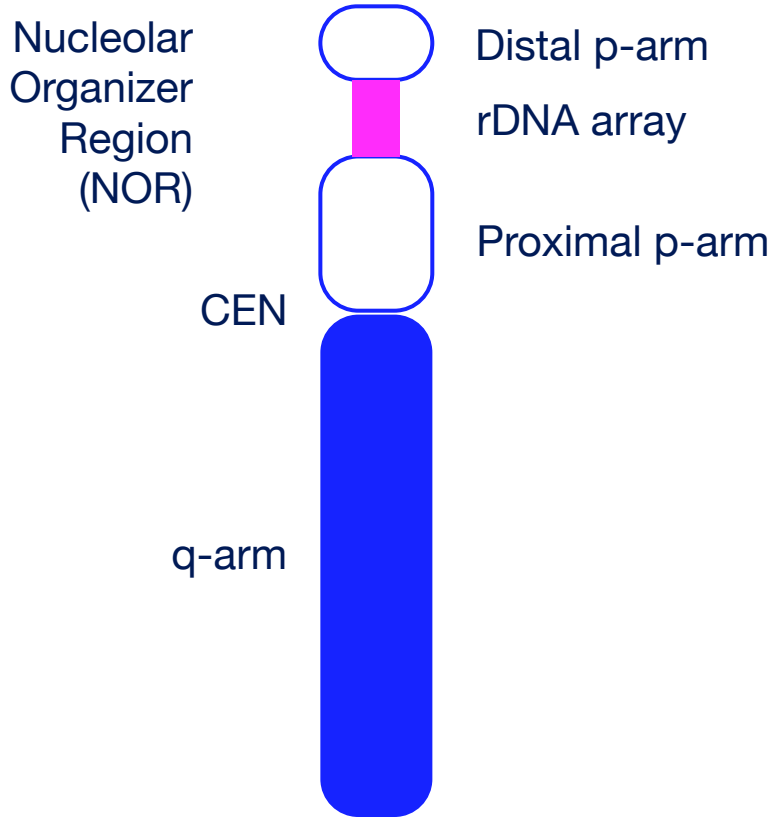
Segmental duplications and their variation in a complete human genome.
Vollger, et al. *Science* (2023)

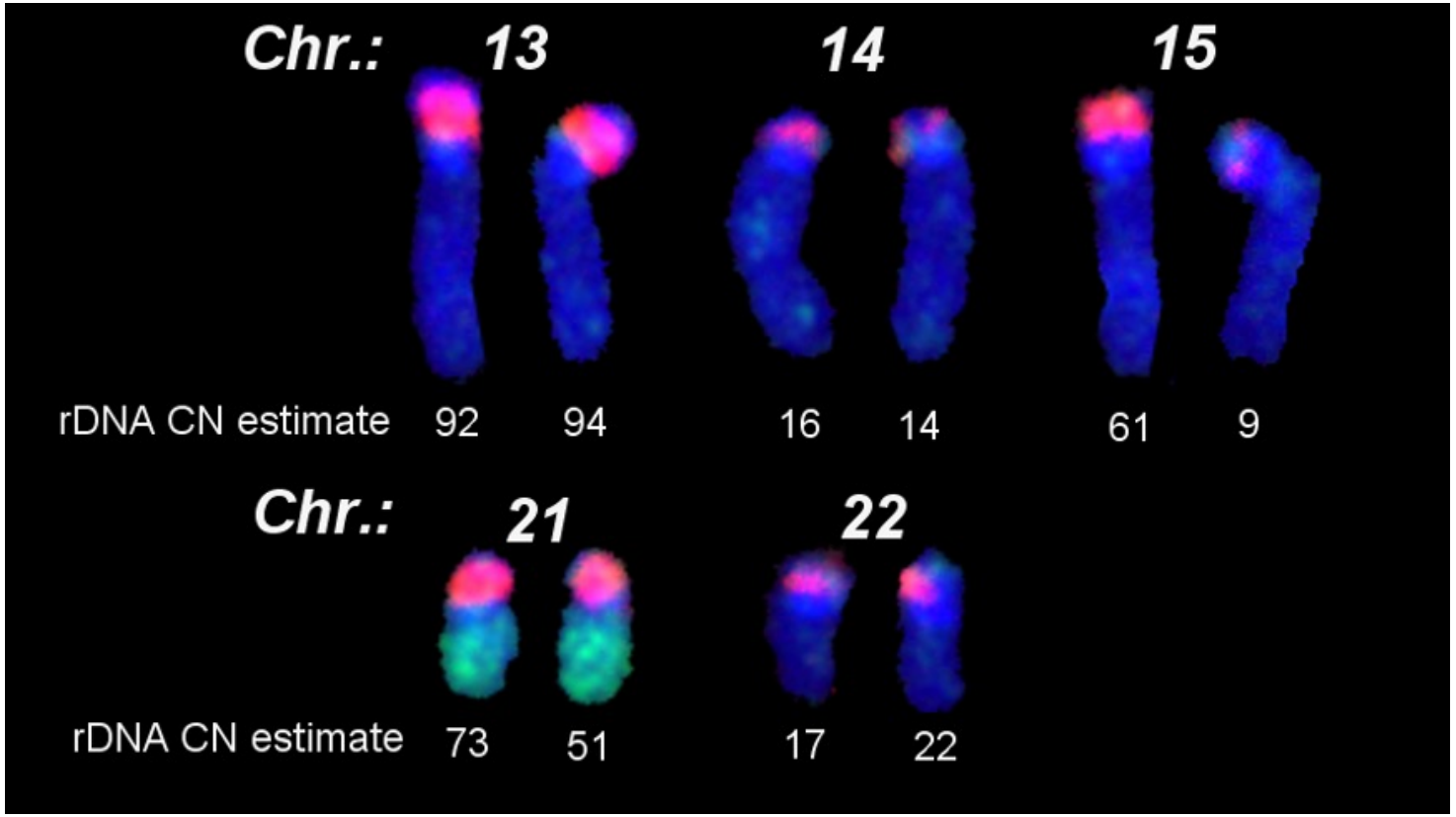
NIH NHGRI

The acrocentrics: Segdups and satellites everywhere

NIH
NHGRI

The human acrocentrics

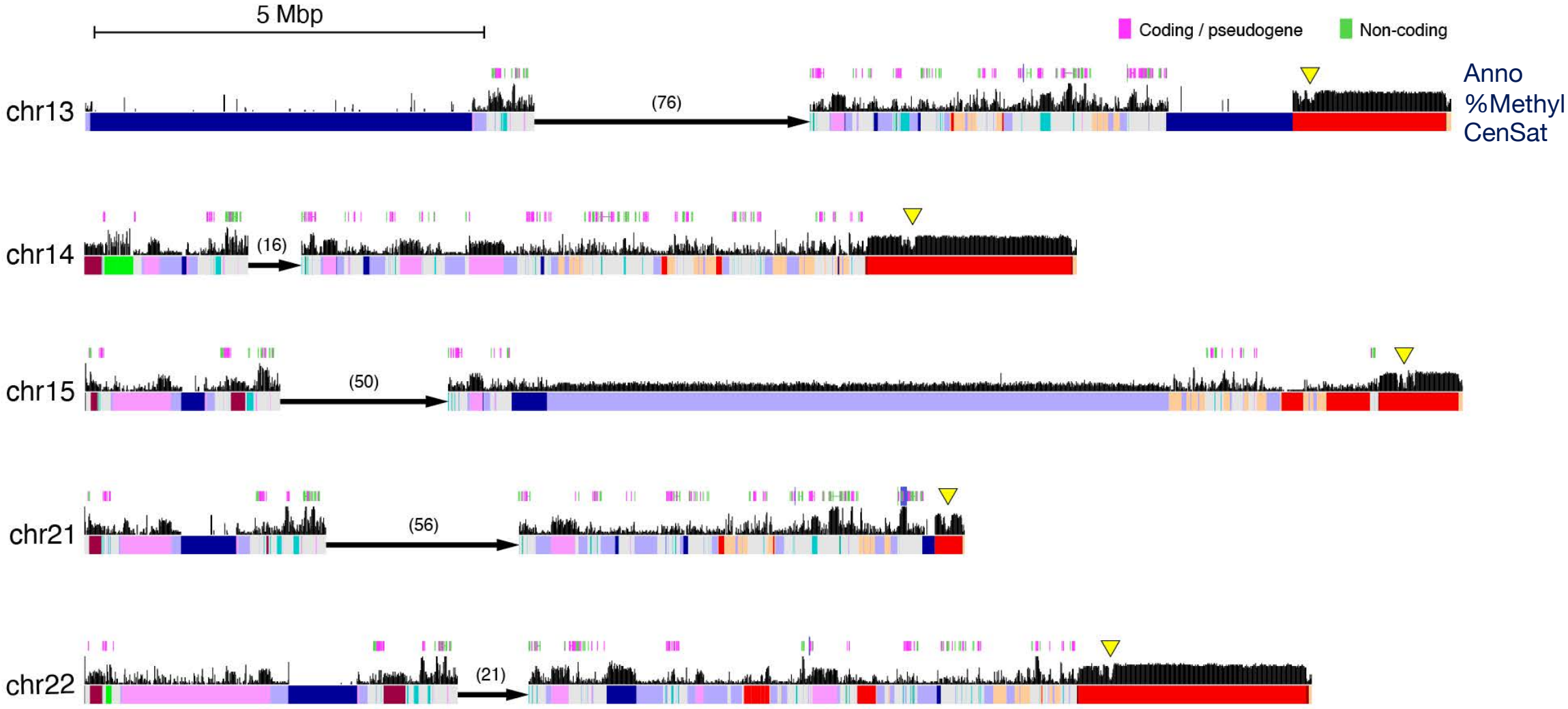




NIH NHGRI

CHM13 SKY by Tamara Potapova, Stowers Institute

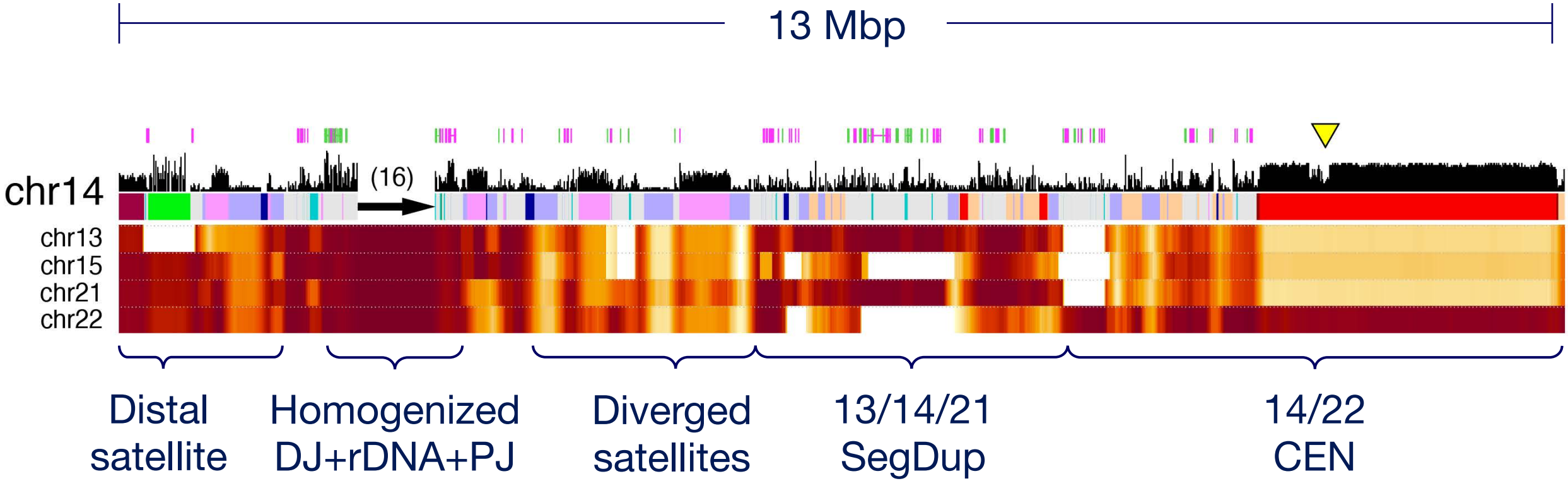
Short arms of the acrocentrics



The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

NIH NHGRI

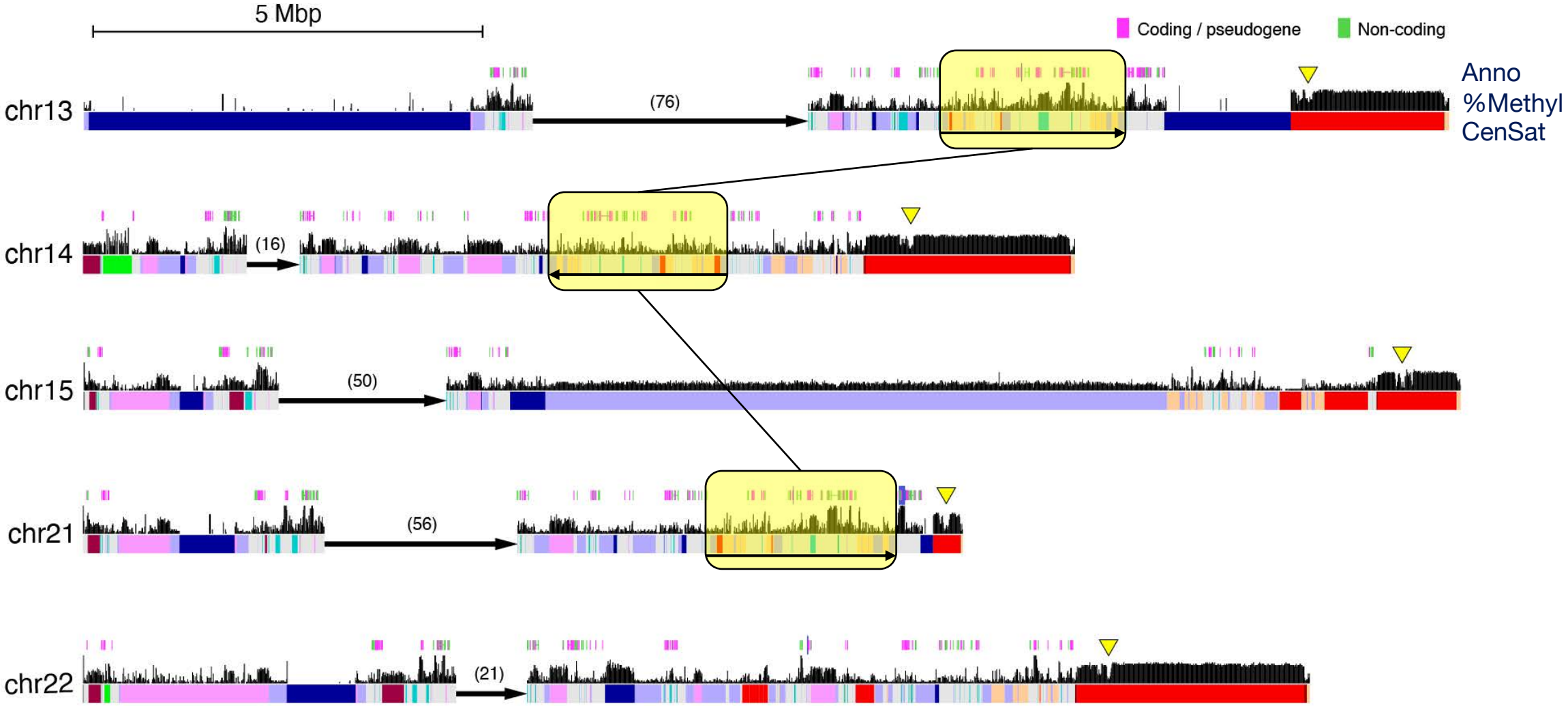
High sequence similarity between acros



The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

HSat1 BetaSat Alpha HOR
HSat3 Alpha MON Segdup

Inverted segdup on 13/14/21

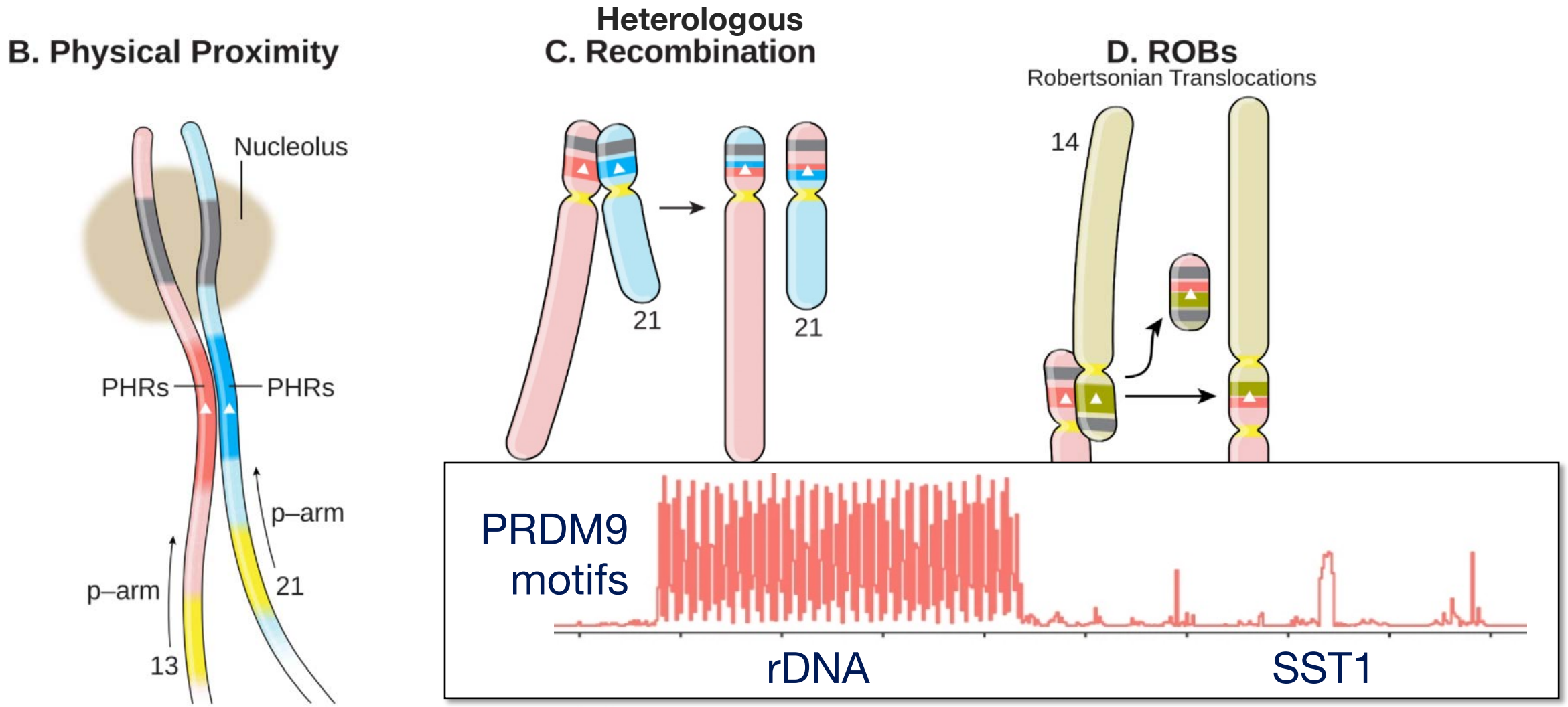


The complete sequence of a human genome.
Nurk, Koren, Rhie, Rautiainen, et al. *Science* (2022)

NIH NHGRI

HSat1 BetaSat Alpha HOR
HSat3 Alpha MON Segdup

Pseudo-homologous regions

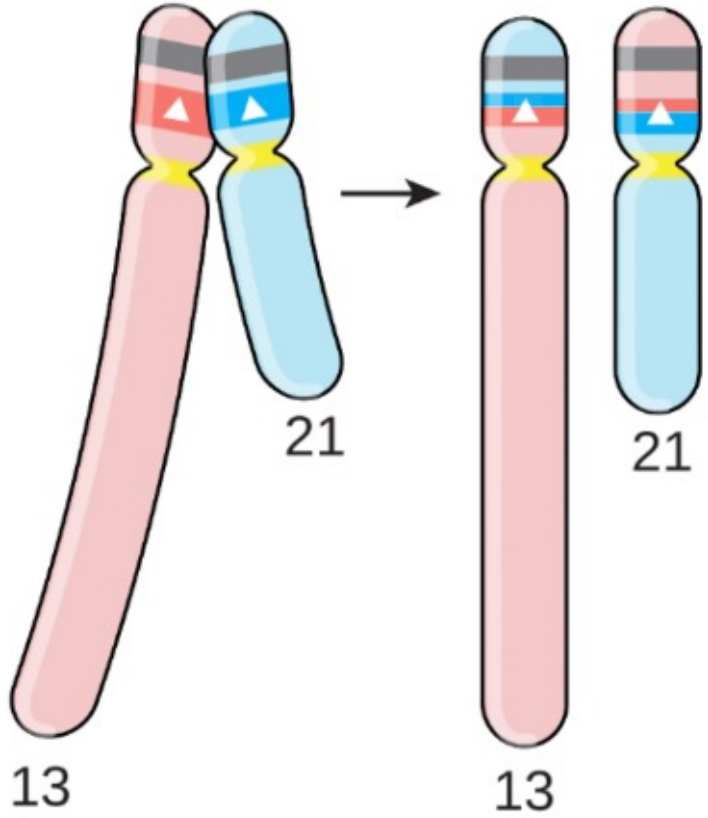


Recombination between heterologous human acrocentric chromosomes.
Guarracino, et al. *Nature* (2023)

NIH NHGRI

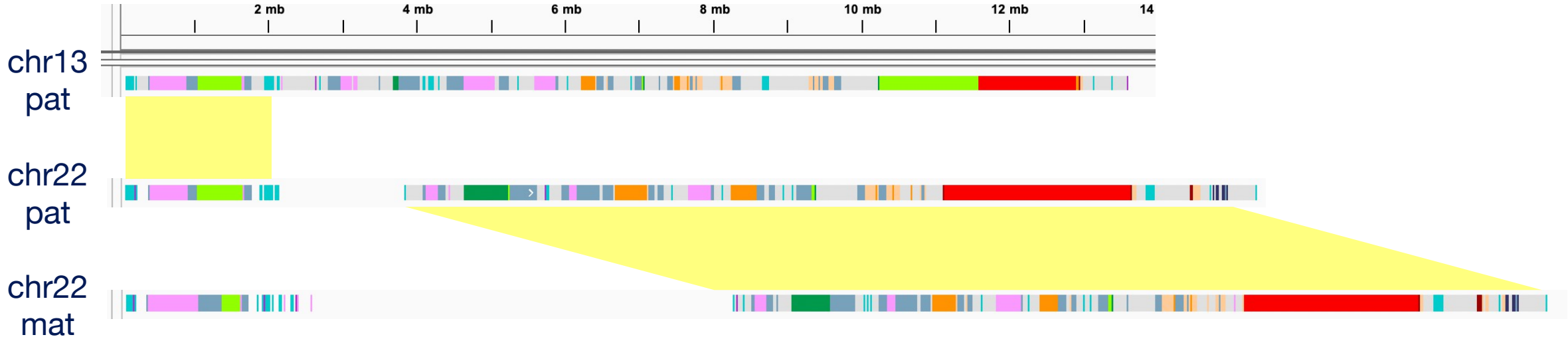
Why are rDNAs on the acros?

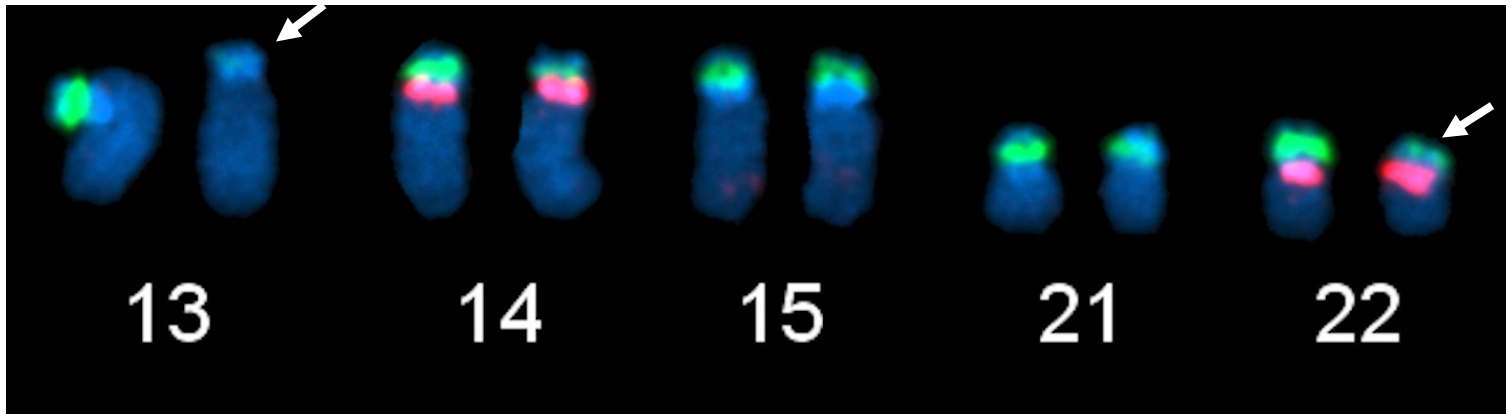
- **rDNAs seem to appear either:**
  - In one array on a metacentric chr
  - In multiple arrays on short arms
- **Why?**
  - Permissive of crossovers
  - Maintains rDNA concerted evolution
- **Conclusion**
  - Human NORs not likely chr-specific



NIH
NHGRI

HG002 distal copy between 13/22





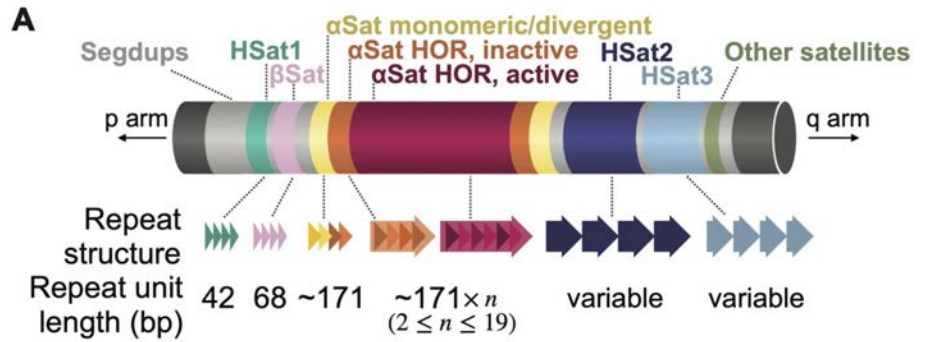
HG002 SKY by Tamara Potapova, Stowers Institute

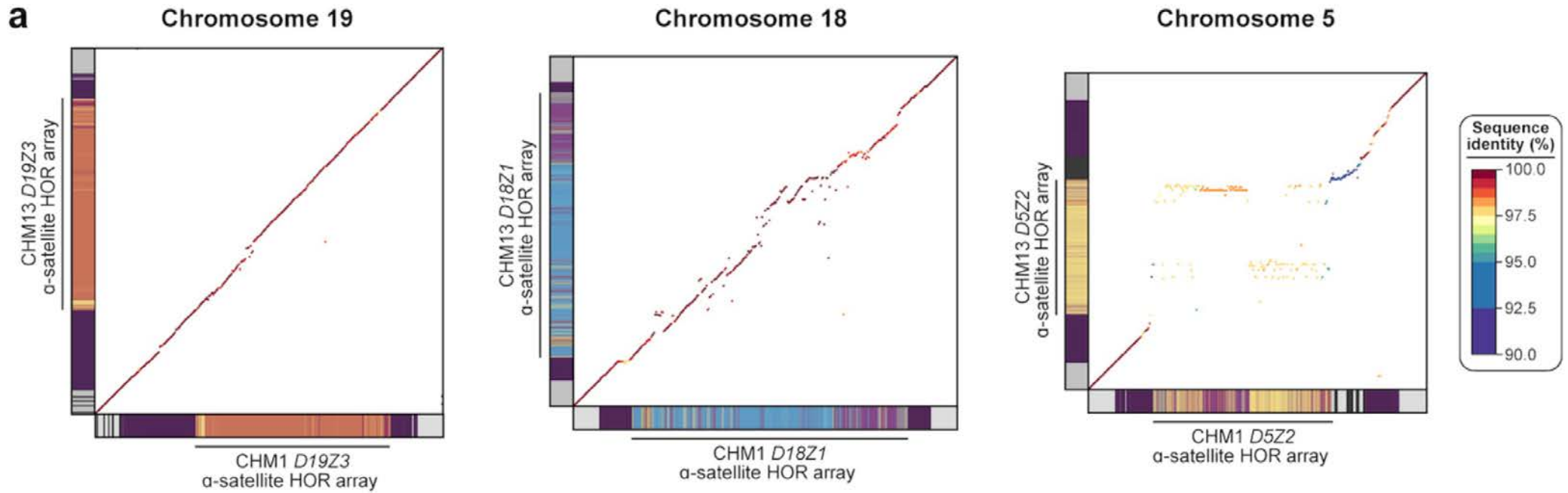
NIH NHGRI

Satellite visualization

NIH
NHGRI

Satellites are often unalignable

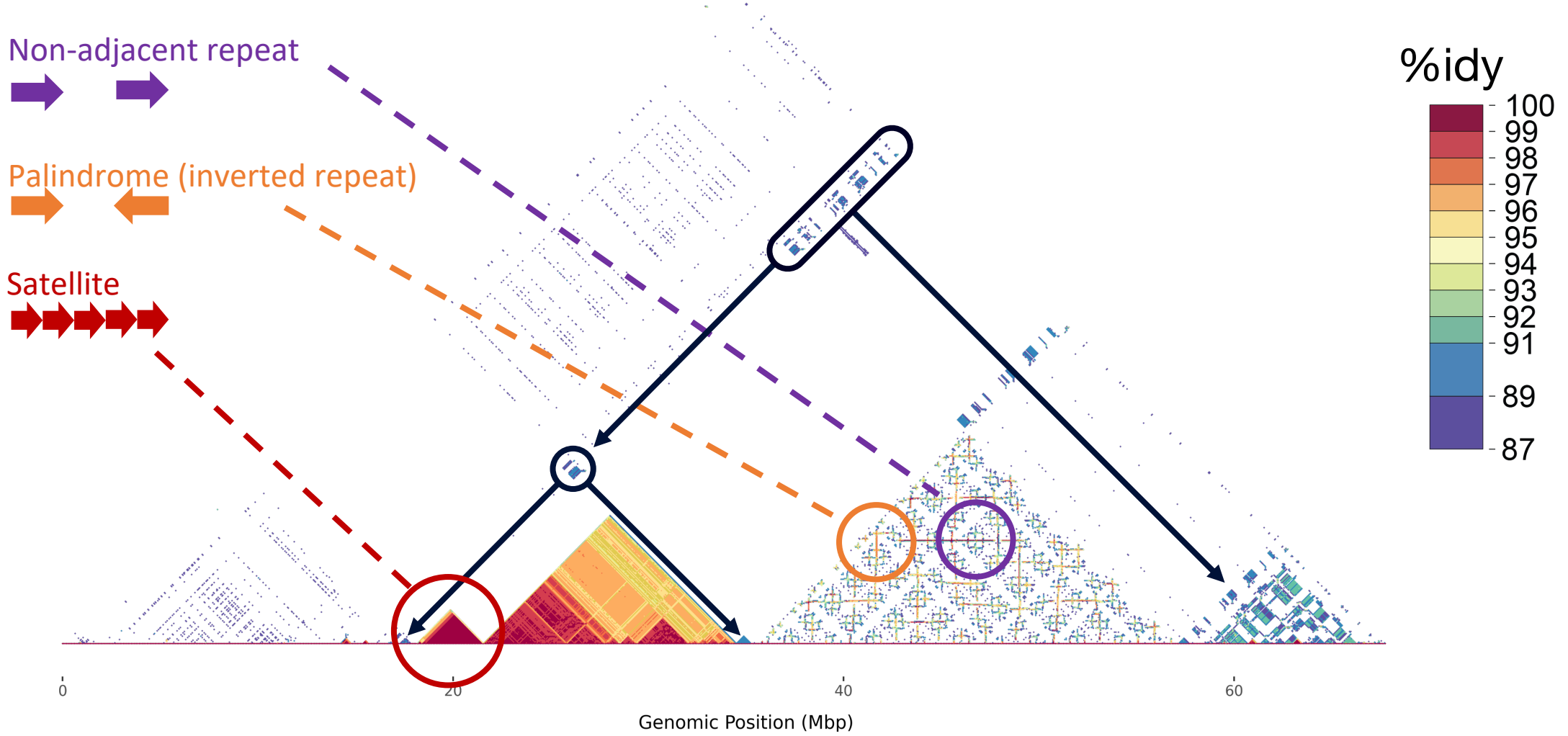




The variation and evolution of complete human centromeres.
Logsdon, et al. *bioRxiv* (2023)

NIH NHGRI

StainedGlass & ModDotPlot



NIH NHGRI

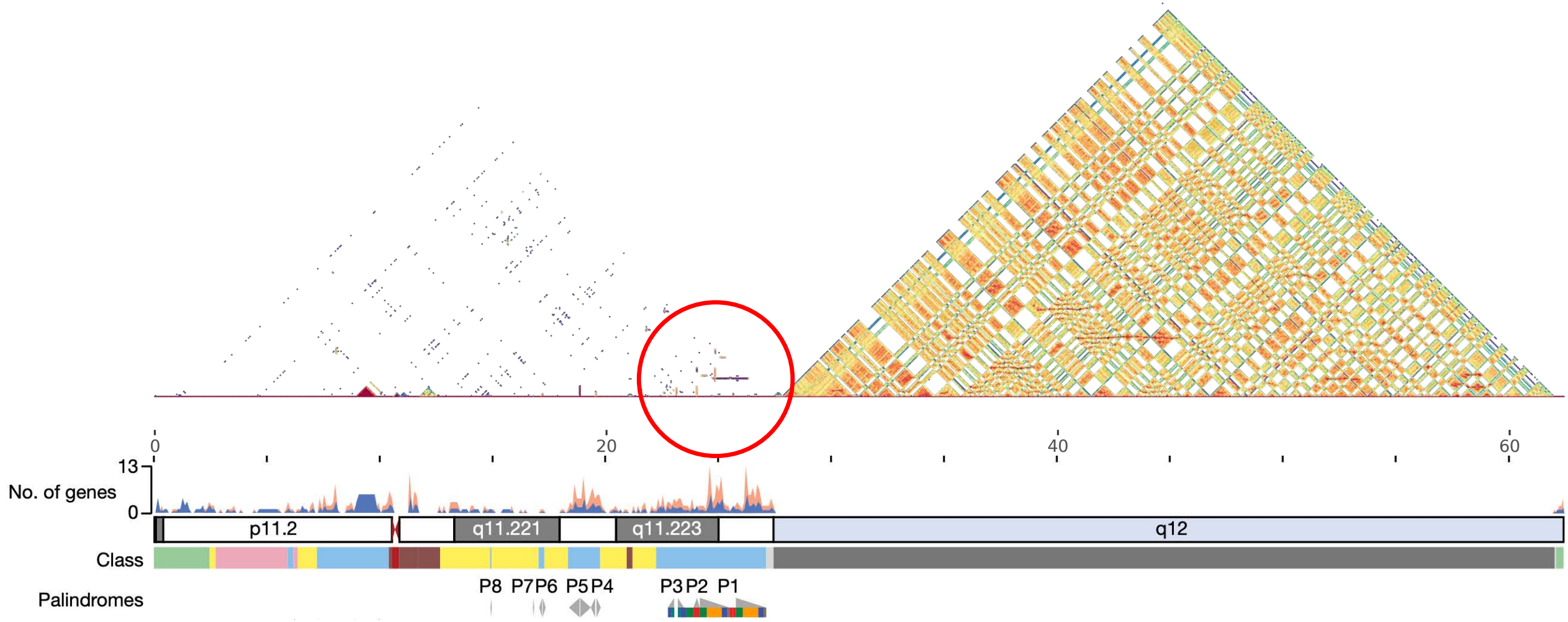
StainedGlass: Interactive visualization of massive tandem repeat structures with identity heatmaps. Vollger, et al. *Bioinformatics* (2022)

https://github.com/mrvollger/StainedGlass
https://github.com/marbl/ModDotPlot

The Y chromosome: A wannabe acrocentric

NIH
NHGRI

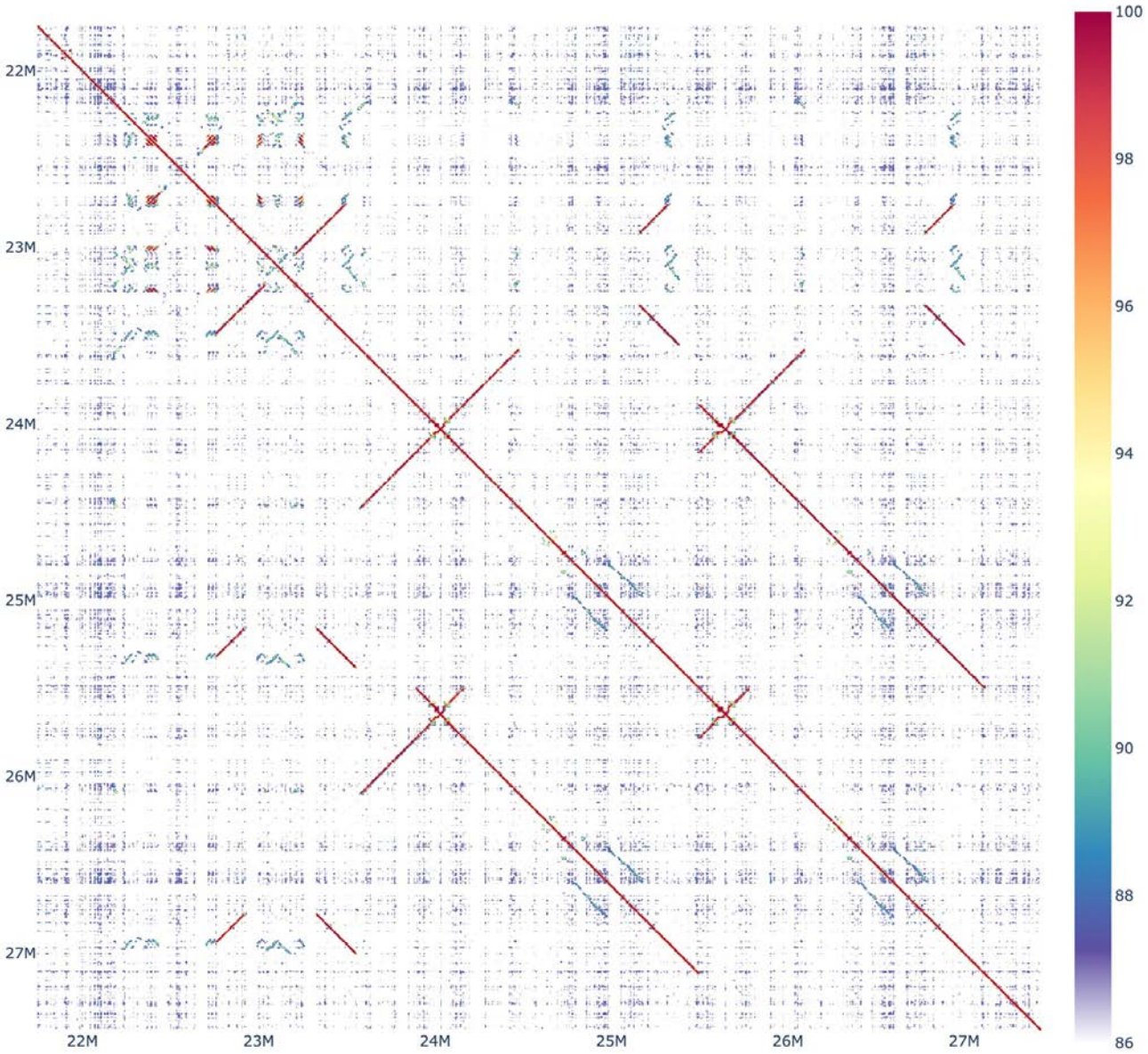
Human HG002 T2T-Y



The complete sequence of a human Y chromosome.
Rhie, Nurk, et al. *Nature* (2023)

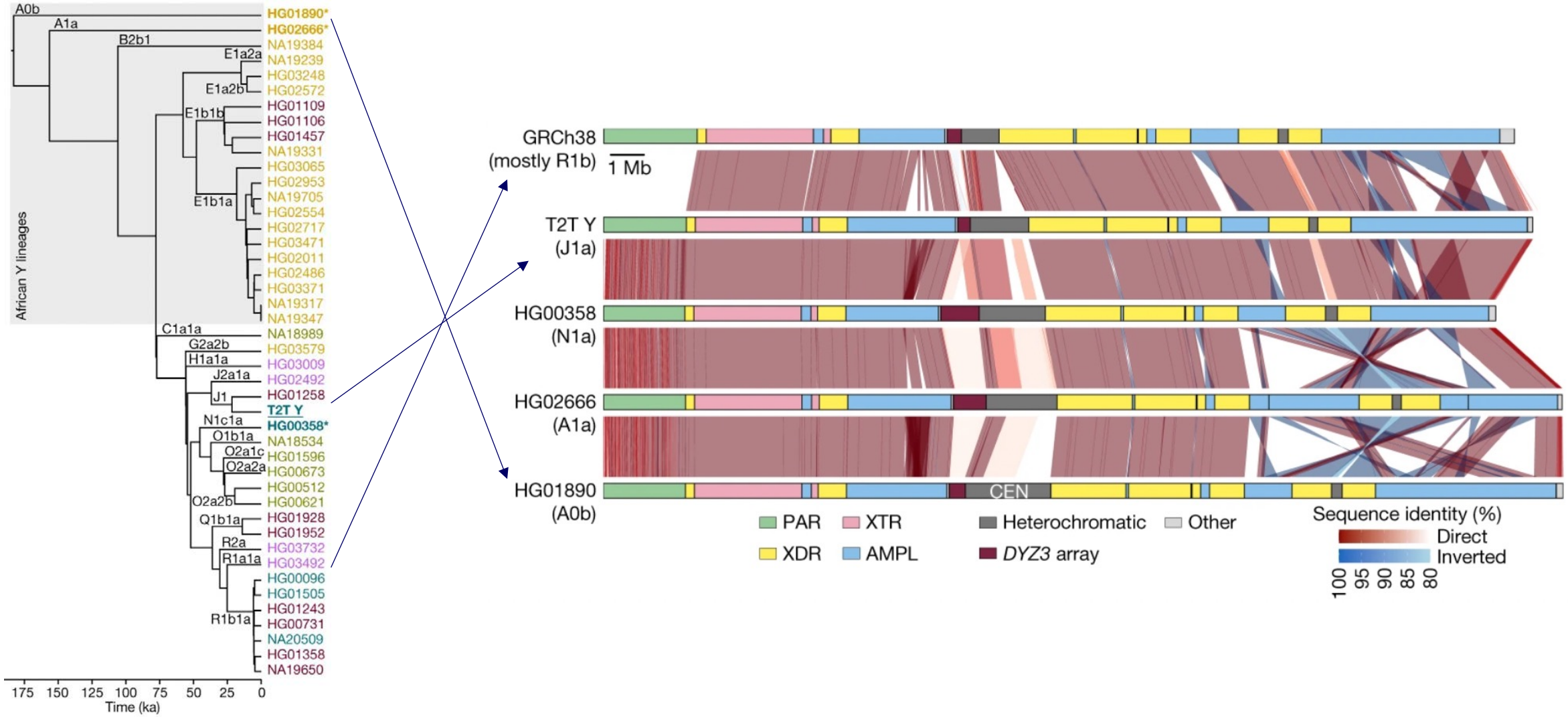
NIH NHGRI

T2T-Y palindromic region



NIH
NHGRI

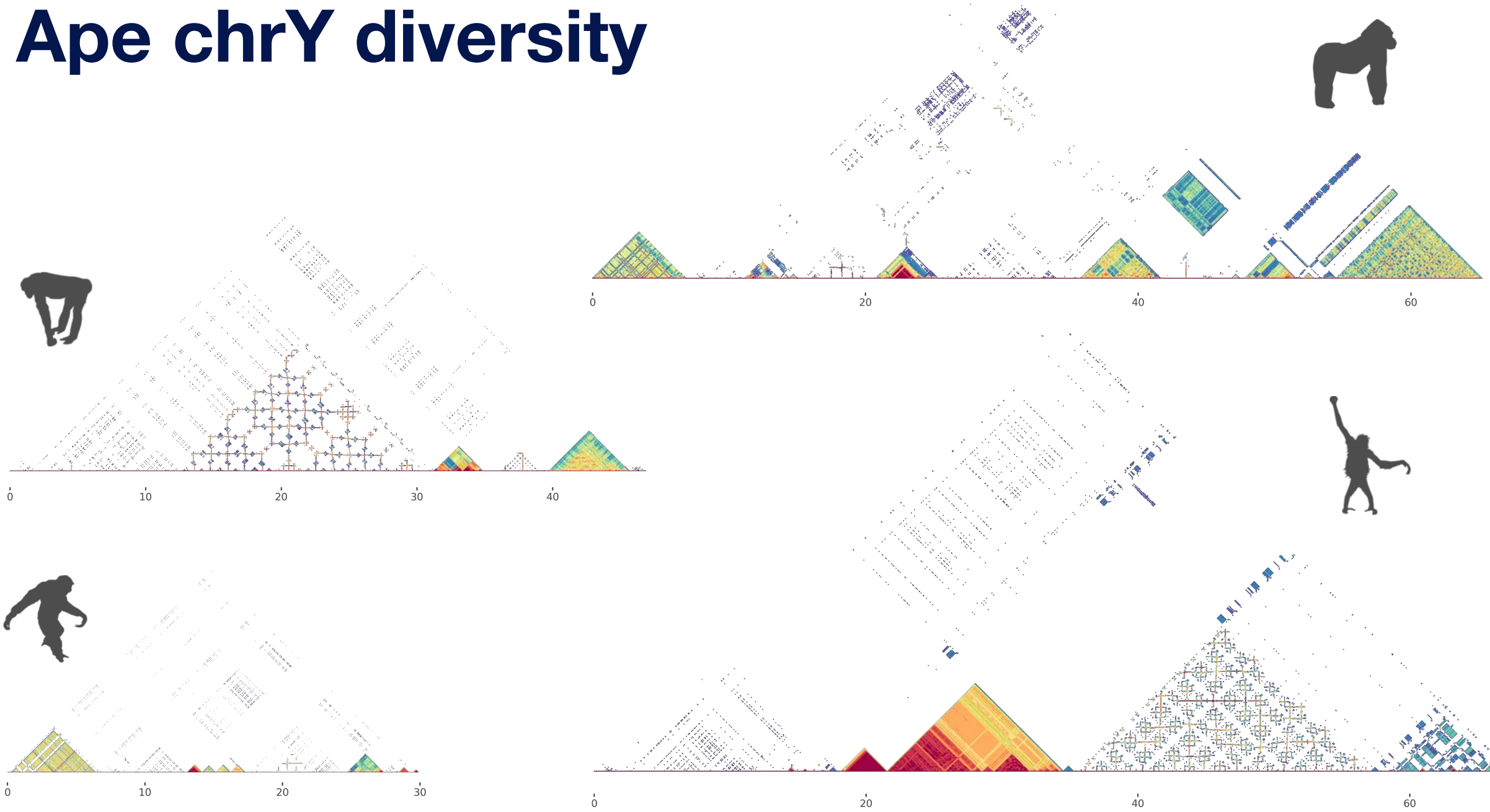
Human chrY variation



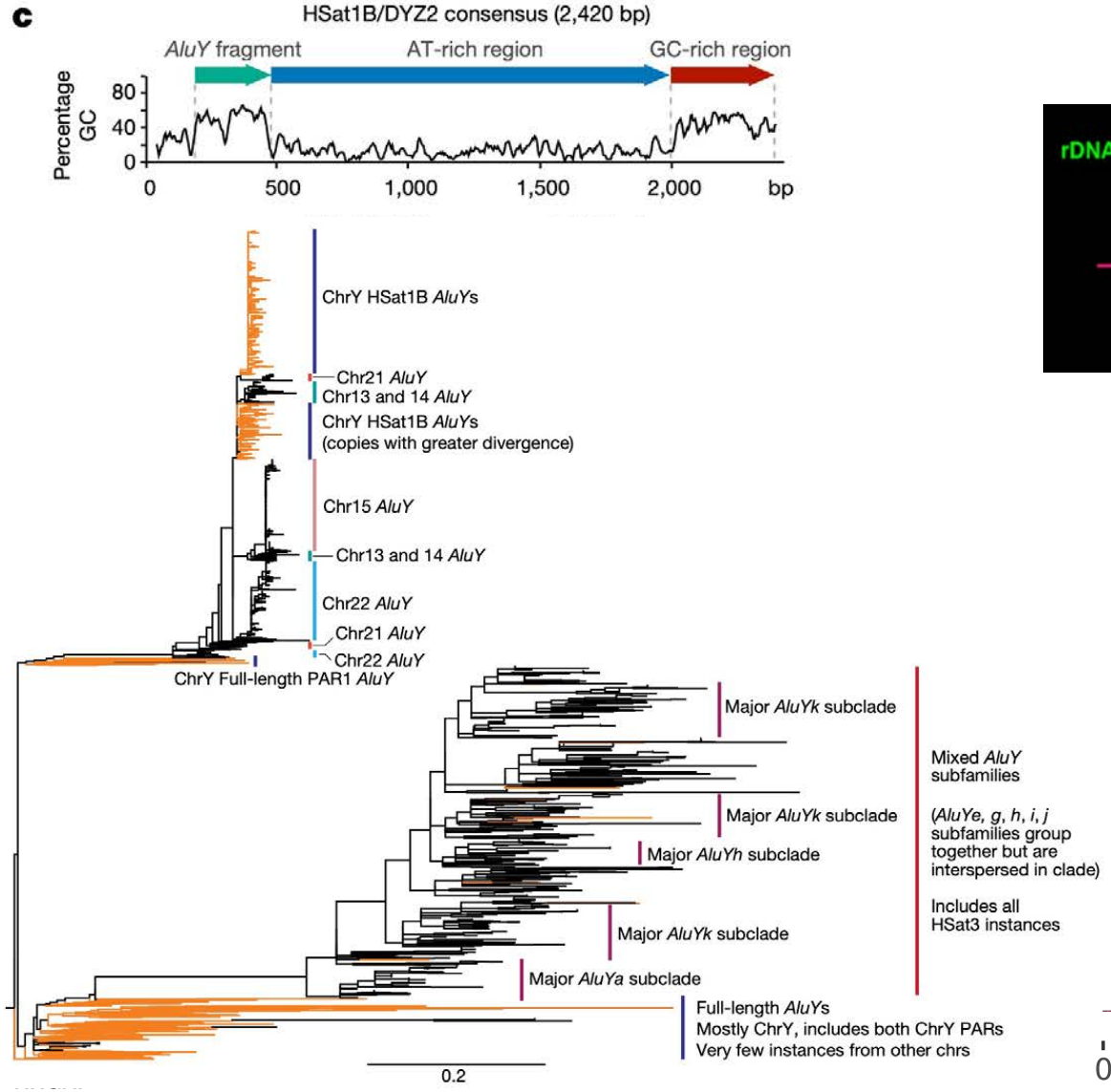
Assembly of 43 human Y chromosomes reveals extensive complexity and variation.
Hallast, et al. *Nature* (2023)

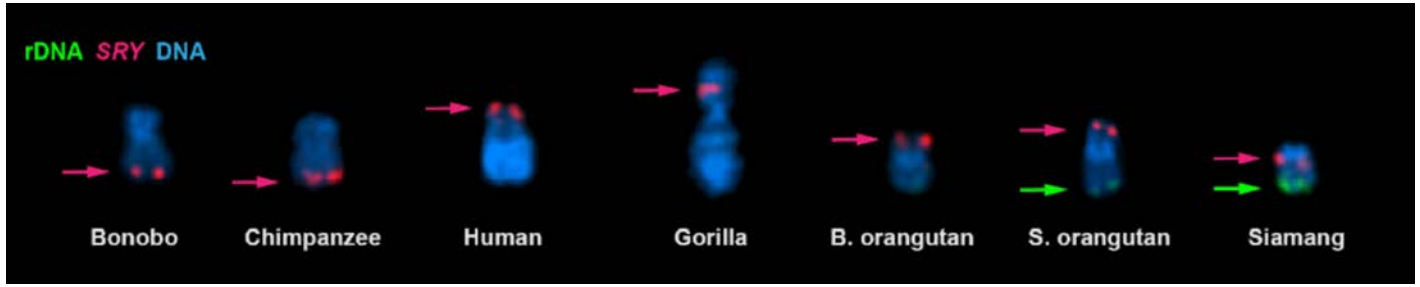
NIH NHGRI

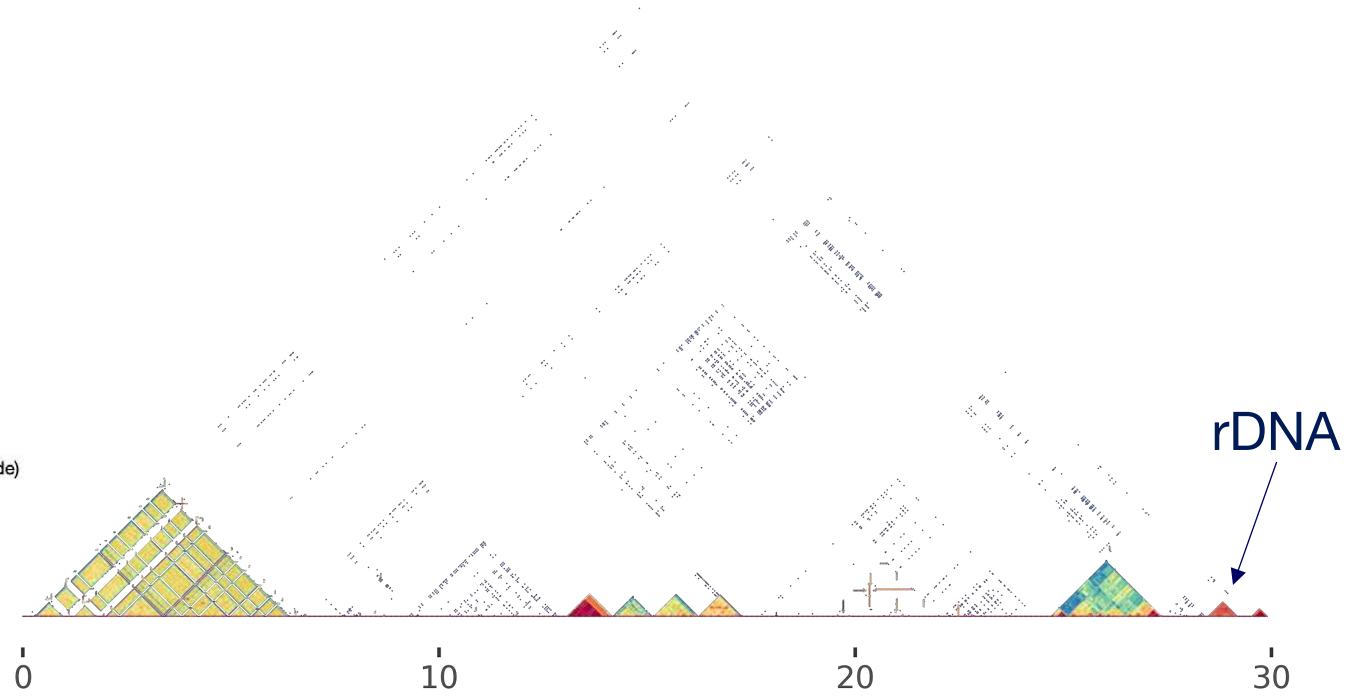
Ape chrY diversity



Crosstalk between acros and chrY

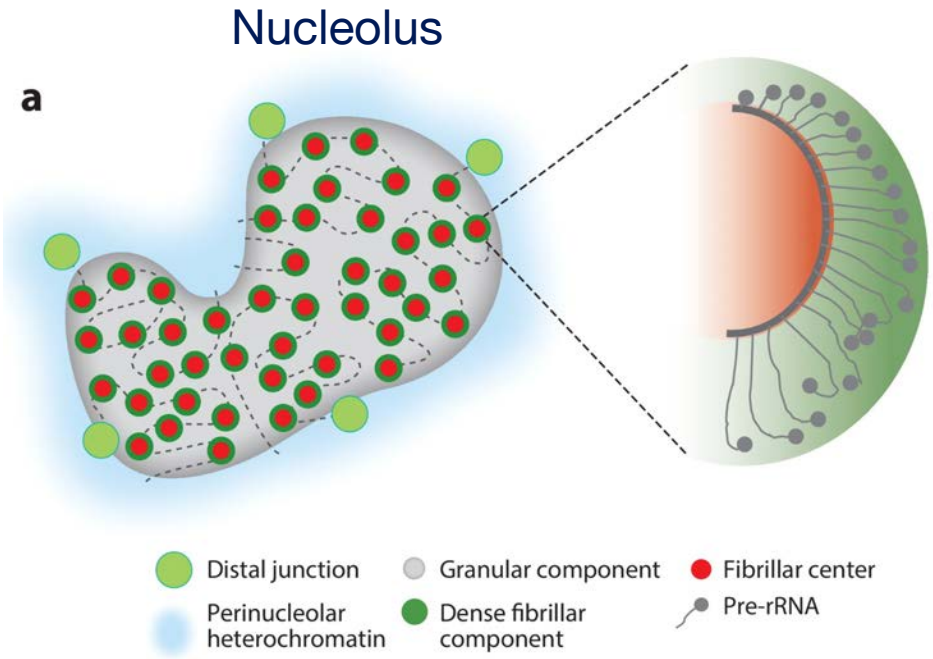






Why does Y behave like an acro?

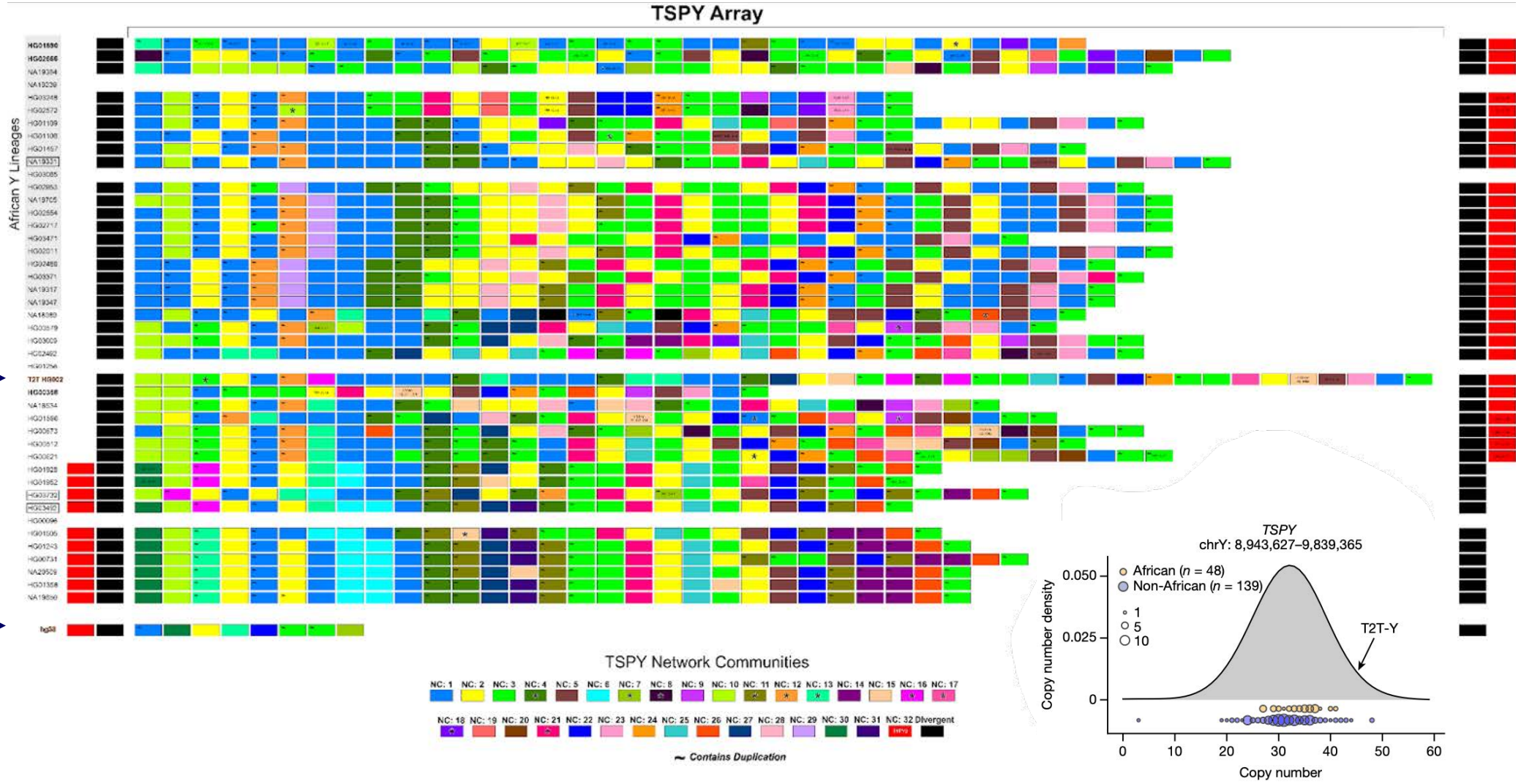
- Low gene density
- Lack of recombination
- Accumulation of satellites
- Association with nucleoli



Nucleolus figure by Brian McStay, University of Galway

Mapping fails to characterize such dynamic regions properly

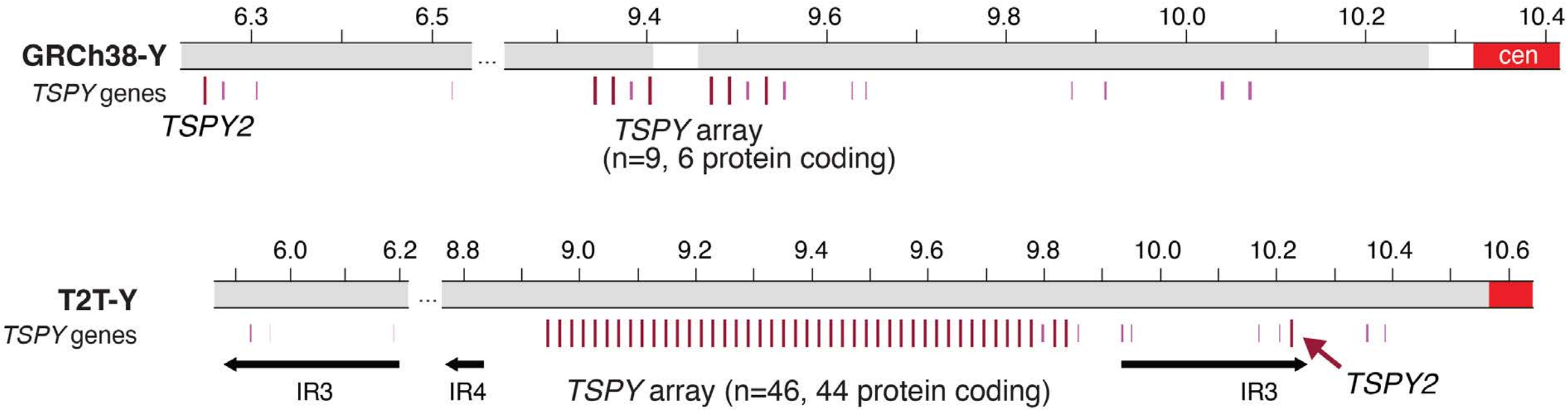
TSPY gene array



Assembly of 43 diverse human Y chromosomes reveals extensive complexity and variation.
Hallast, et al. *bioRxiv* (2022)

NIH NHGRI

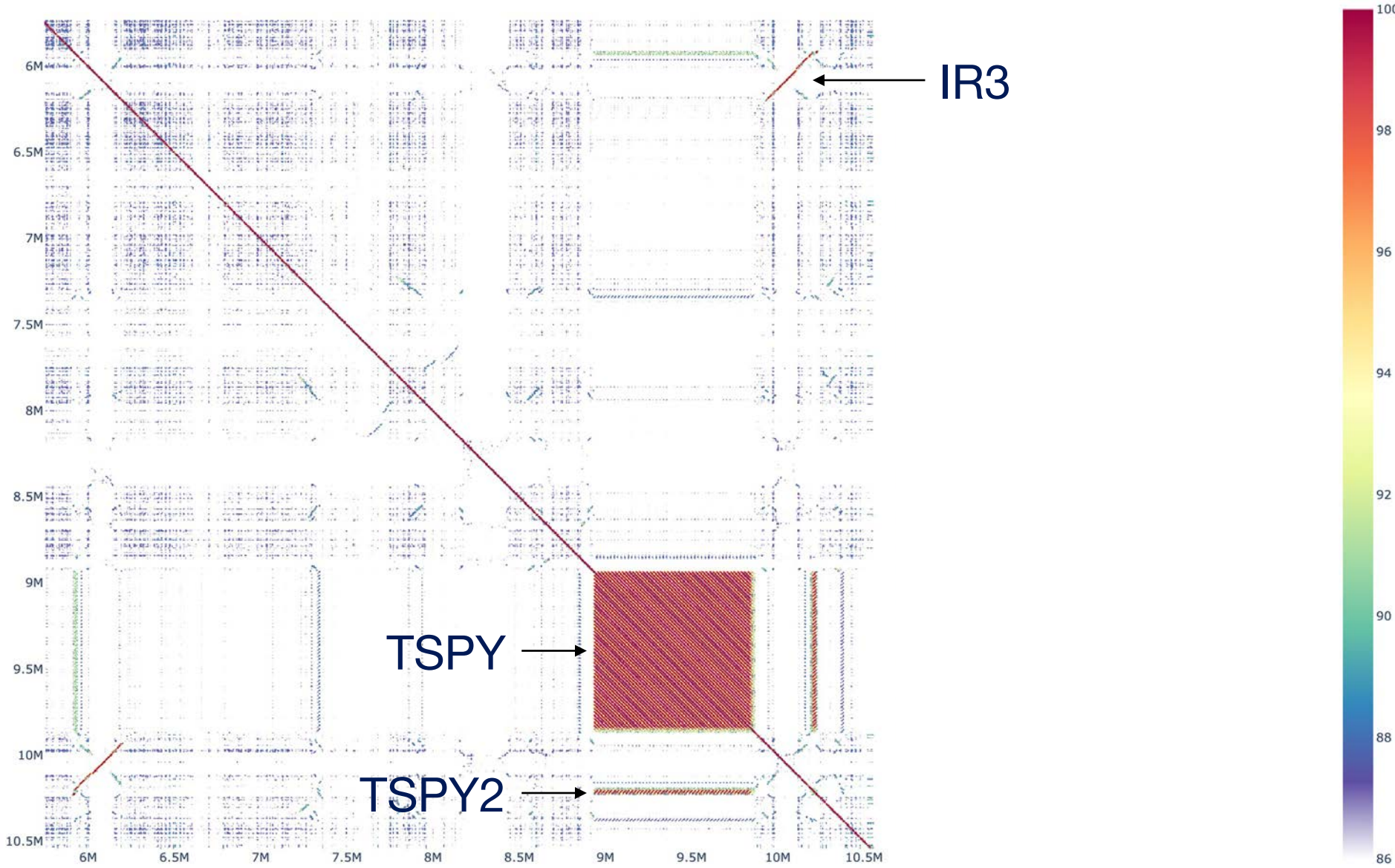
TSPY2 gene conversion



The complete sequence of a human Y chromosome.
Rhie, Nurk, et al. *bioRxiv* (2022)

NIH NHGRI

TSPY region



NIH
NHGRI

Mapped ONT reads clip at TSPY2

GRCh38 ref

HG002 reads

Not just a problem on chrY

Lost in the WASH. The functional human WASH complex I gene is on chromosome 20

Daniel Cerdán-Vélez, Michael L. Tress

doi: https://doi.org/10.1101/2023.06.14.544951

This article is a preprint and has not been certified by peer review [what does this mean?].

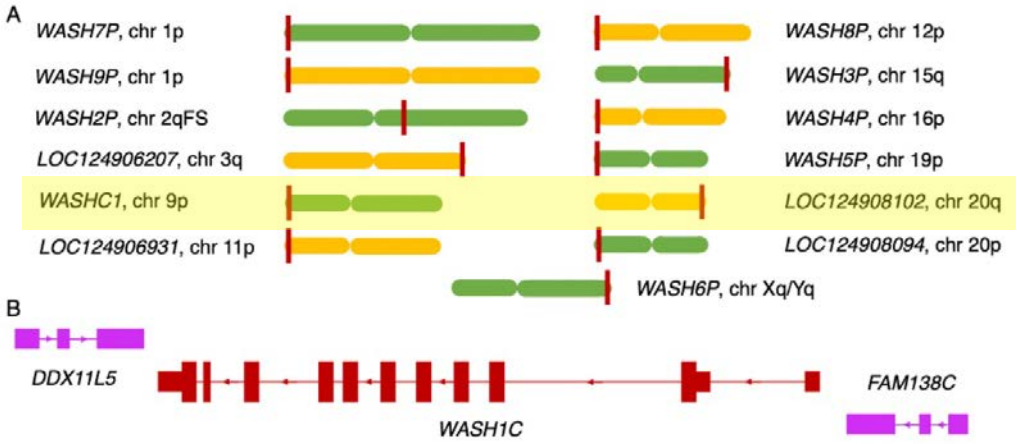
Abstract

The WASH1 gene produces a protein that forms part of the developmentally important WASH complex. The WASH complex activates the Arp2/3 complex to initiate branched actin networks at the surface of endosomes. As a curiosity, the human reference gene set includes nine WASH1 genes. How many of these are pseudogenes and how many are *bona fide* coding genes is not clear.

Eight of the nine WASH1 genes reside in rearrangement and duplication-prone subtelomeric regions. Many of these subtelomeric regions had gaps in the GRCh38 human genome assembly, but the recently published T2T-CHM13 assembly from the Telomere to Telomere (T2T) Consortium has filled in the gaps. As a result, the T2T Consortium has added four new WASH1 paralogues in previously unannotated subtelomeric regions.

Here we show that one of these four novel WASH1 genes, *LOC124908094*, is the gene most likely to produce the functional WASH1 protein. We also demonstrate that the other twelve WASH1 genes derived from a single *WASH8P* pseudogene on chromosome 12. These 12 genes include WASHC1, the gene currently annotated as the functional WASH1 gene.

We propose *LOC124908094* should be annotated as a coding gene and all functional information relating to the *WASHC1* gene on chromosome 9 should be transferred to *LOC124908094*. The remaining WASH1 genes, including *WASHC1*, should be annotated as pseudogenes. This work confirms that the T2T assembly has added at least one functionally relevant coding gene to the human reference set. It remains to be seen whether other important coding genes are missing from the GRCh38 reference assembly.



Are subtelomeres undergoing heterologous recombination like the short arms?

NIH NHGRI

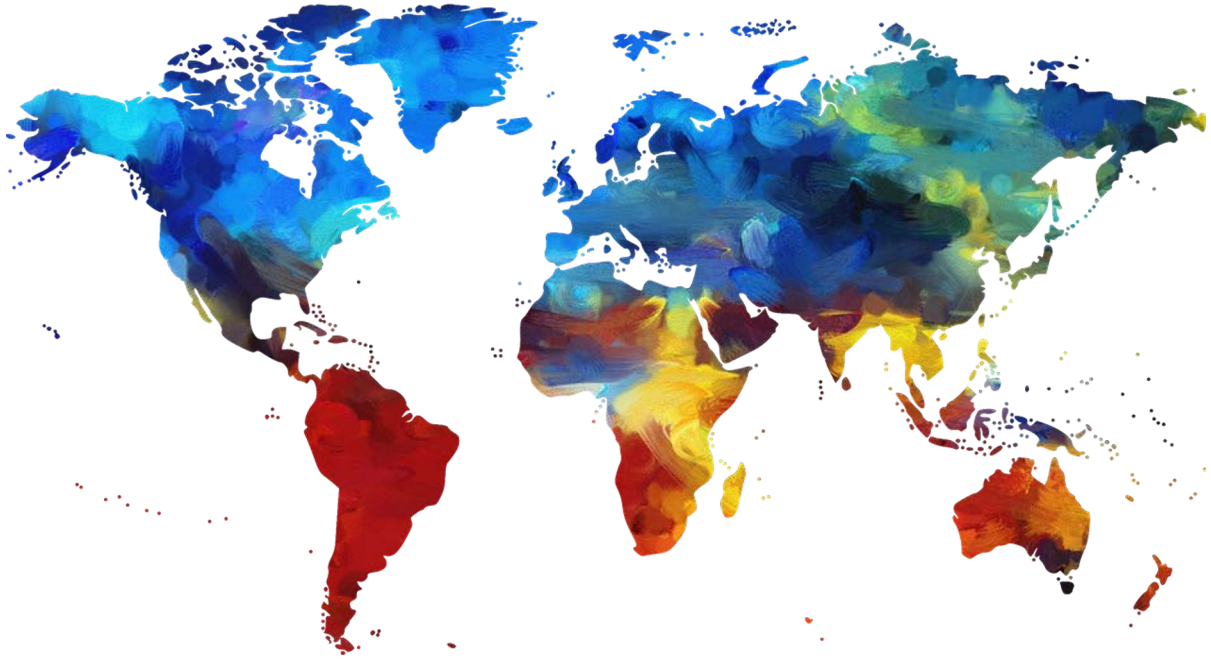
The issues with mapping

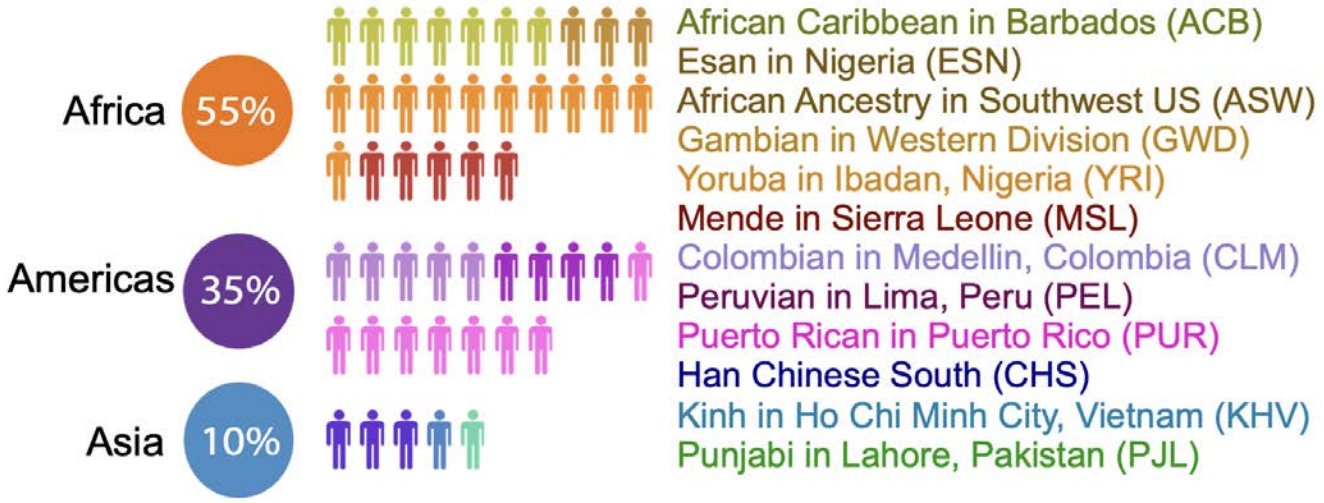
- **Structural variation is rampant**
  - Enriched around segdups and satellites
- **Satellites and gene arrays evolve differently**
  - Unequal crossing over, recombination
- **Segdups prone to mutation and gene conversion**
  - Important for human adaptation and disease
- **Genes are not always where you think**
  - Troublesome when mapping long reads
  - The location of the functional copy matters

What’s the solution?

NIH
NHGRI

A draft human pangenome reference





v1: HiFi assemblies (released)
v2: HiFi+ONT assemblies
v3: T2T assemblies

48 initial samples from 1KGP

- Immortalized 1000G cell lines
- Cover allelic and geographic diversity
- Availability of low passage cell lines
- Availability of trios/parental data
- Aiming for 350+ diploid genomes

NIH NHGRI

A draft human pangenome reference.
Liao, et al. *Nature* (2023)

The best reference is no reference

- **What is the diploid genome of this individual?**
  - Variant calls are just an incomplete genome assembly
  - De novo assembly is costly

- **Working towards “genome inference”**
  - Use all information available (e.g. reads + pangenome)
  - Infer the complete, diploid genome of the individual
  - Genome assembly with priors

We need better benchmarks

- **GIAB benchmarks have had a tremendous effect**
  - But, there is no incentive to do well on the hardest 10%

- ***Variant* benchmarks are fundamentally flawed**
  - NIST RMs are defined as a list of variants vs. hg38
  - You're only told ~90% of what's in the tube

- ***Genome* benchmarks are the solution**
  - The genome sequence itself is the truth
  - T2T-HG002v1.0 "Q100" assembly released last week

2000: Draft genomes
2010: Reference genomes
2020: T2T genomes
2030: Genomes

NIH
NHGRI

The future is complete genomes

1. Routine *assembly* of T2T genomes
2. Comprehensive pangenome reference database
3. Cheap and routine *inference* of T2T genomes
4. ML-based annotation of personal genomes
5. Accurate functional/somatic view over lifespan

Will be the norm in ~10 years (and cheap)

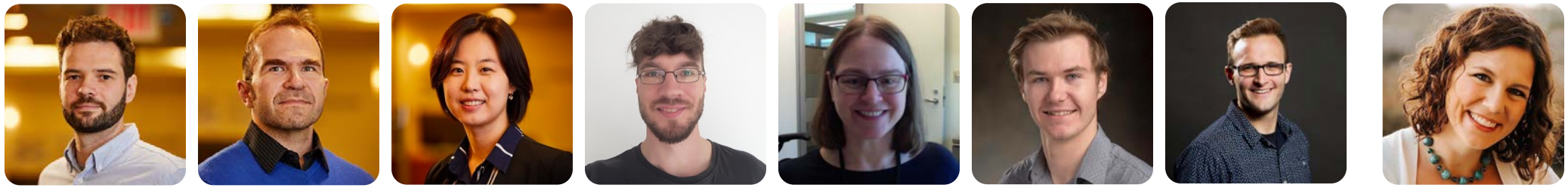
Resources

GitHub /MarBL

- Verkko assembler
  - https://github.com/marbl/verkko
- T2T-CHM13 reference
  - https://github.com/marbl/chm13
- HG002 diploid genome benchmark
  - https://github.com/marbl/hg002
- Human Pangenome Reference Consortium
  - https://humanpangenome.org
- ModT2T
  - https://www.genomeark.org/

NIH NHGRI

Team T2T, HPRC, (…and many more)



Contact me if interested in joining the T2T mailing list!

NIH NHGRI