Beyond one T2T human genome: What's next?

Arang Rhie

Staff Scientist
Genome Informatics Section, NHGRI, NIH

June 2, 2023

National Human Genome Research Institute

@ArangRhie

The Forefront of Genomics®

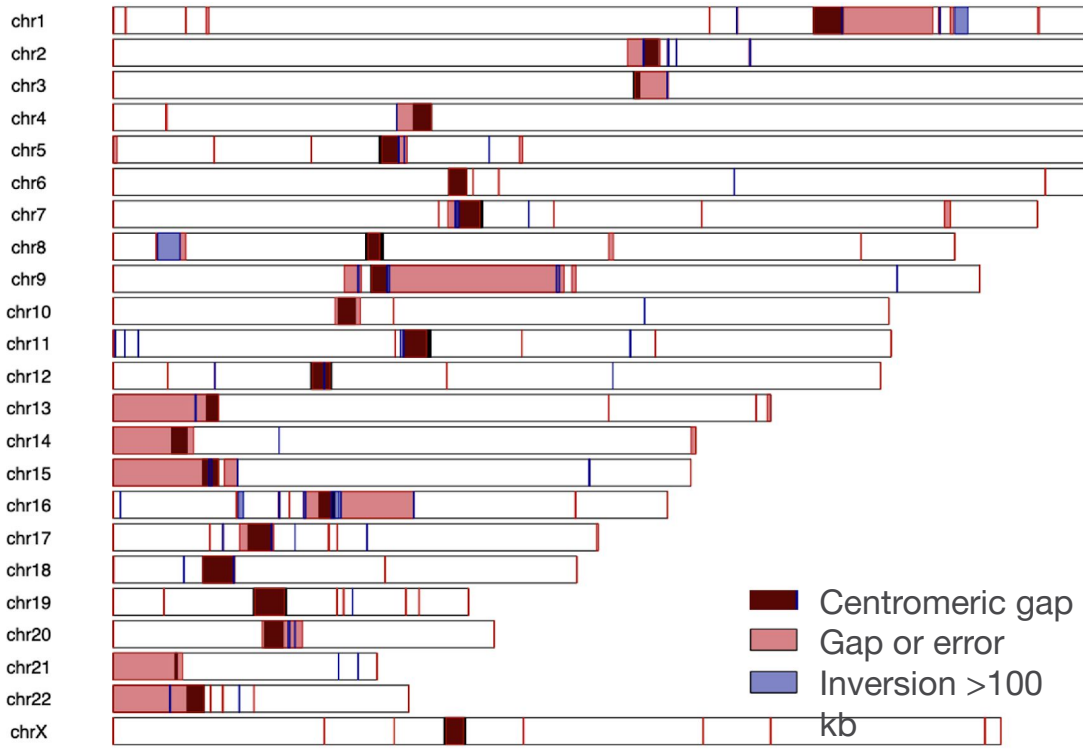
The human genome is finished!

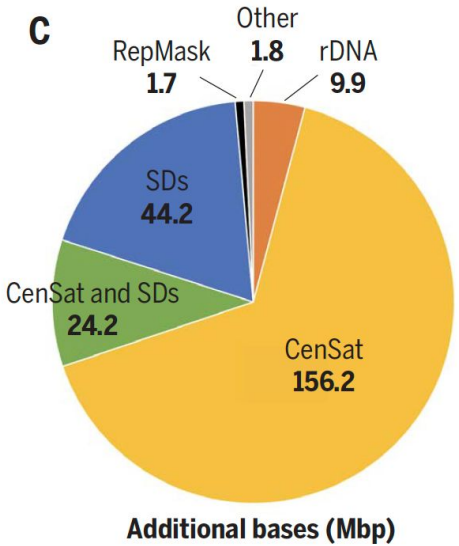
- The human genome is finally finished from T to T!
- Solved with combination of PacBio HiFi + ONT ultra-long
- Technology development: first 92% took 10 years, last 8% took 20 years
- We've been fooling ourselves for 20 years!

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

NIH NHGRI

The last 8% of the human genome





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

The human genome is finished?

- The human genome is finally finished from T to T! **except the Y!!**
- Solved with combination of PacBio HiFi + ONT ultra-long
- Technology development: first 92% took 10 years, last 8% took 20 years
- We've been fooling ourselves for 20 years!
  +1

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

NIH NHGRI

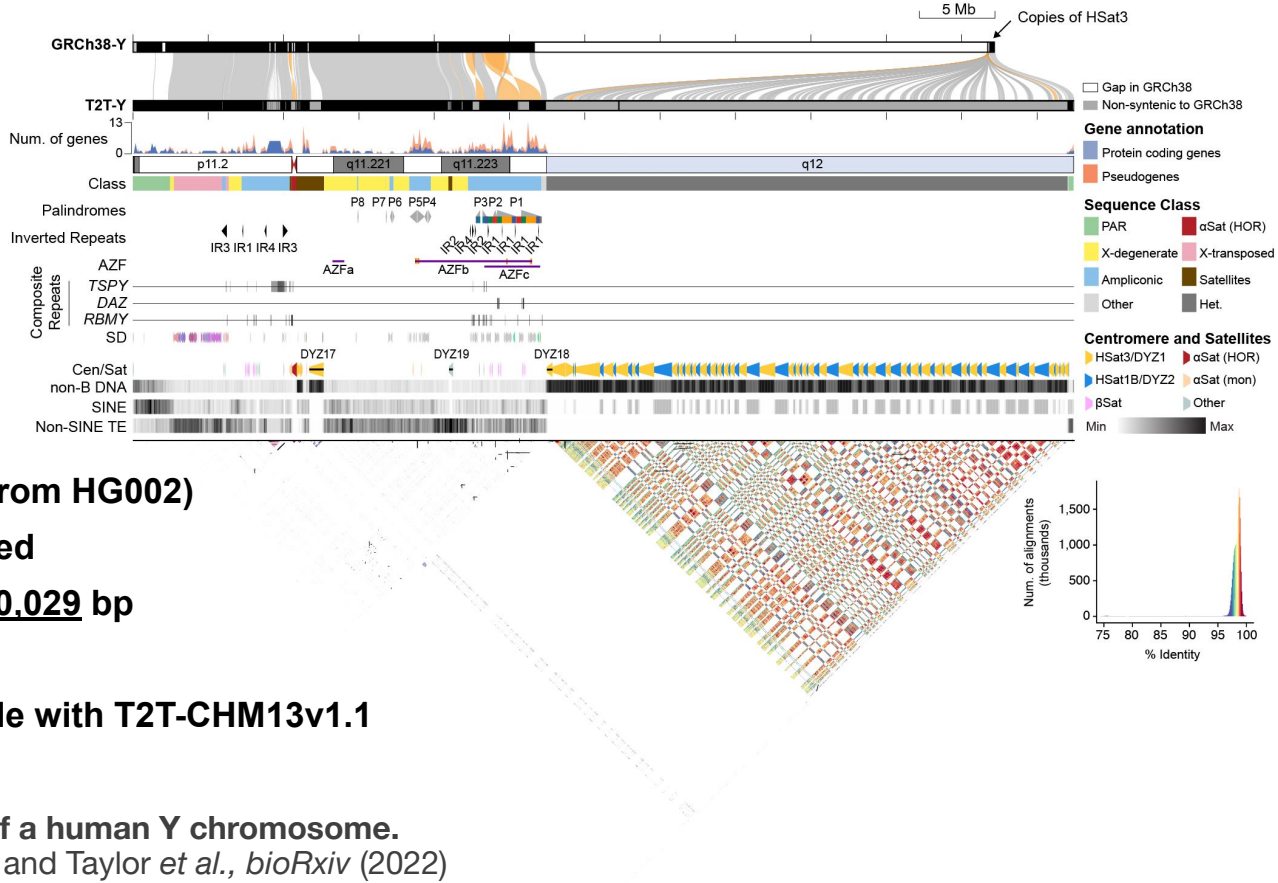
The Y chromosome is finished

CHM13 46,XX cell line from U. Surti, Pitt.
SKY karyotype from T. Potapova, Stowers

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

NIH NHGRI

The Y chromosome



T2T-CHM13v2.0

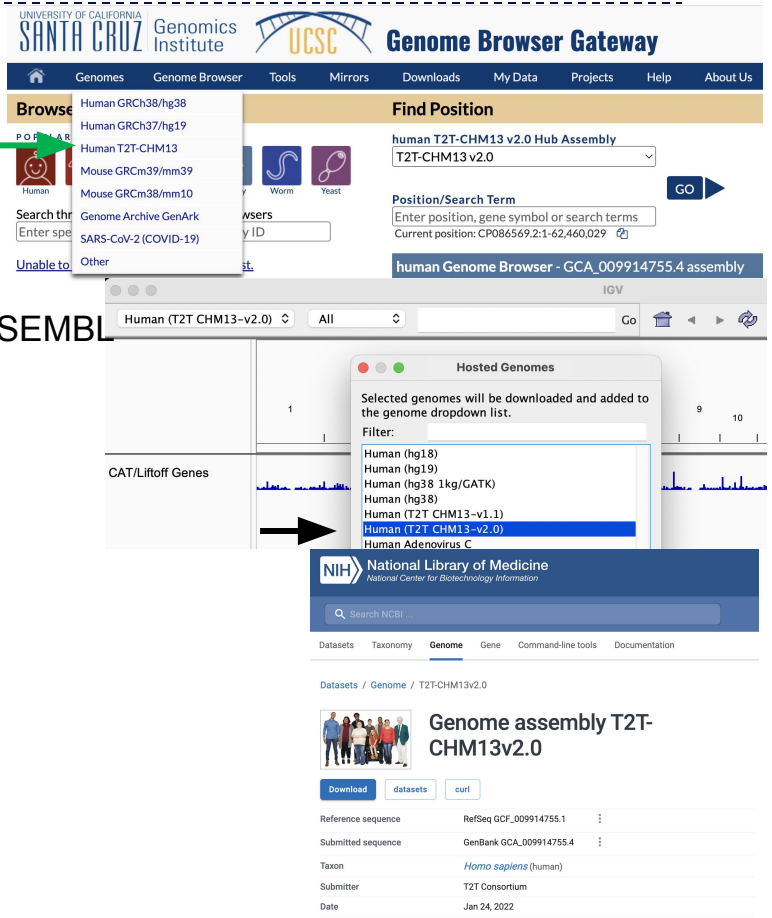
- **24 chromosomes (Y from HG002)**
- **0 unlocalized, unplaced**
- **3,054,832,041 + 62,460,029 bp**
- **No gaps**
- **Y is Q73.8, comparable with T2T-CHM13v1.1**

NIH NHGRI

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

T2T-CHM13 (hs1) as an alternate reference

- **github.com/marbl/CHM13**
- **github.com/marbl/CHM13-issues**
- **Analysis set**
  - Y PAR masked, Cambridge ref. MT
- **Gene Annotation**
  - Curated RefSeq/Liftoff, RefSeq, GENCODE CAT/Liftoff, ENSEMBL
- **1:1 Alignments and chains for liftOver**
  - hg19, hg38
- **Datasets**
  - RepeatMasker / CenSat annotation
  - Short-read and long-read alignments
  - Expression: RNA-Seq and Iso-Seq
  - Epigenetics: ENCODE, ONT/HiFi methylation calls
  - 1KGP, SGDP variant calls + bam
  - SGDP copy number estimates
  - ClinVar, dbSNP, GWAS, gnomAD lifted over from GRCh38



CHM13v2.0 = CHM13v1.1 + HG002Yv2.7

NIH NHGRI

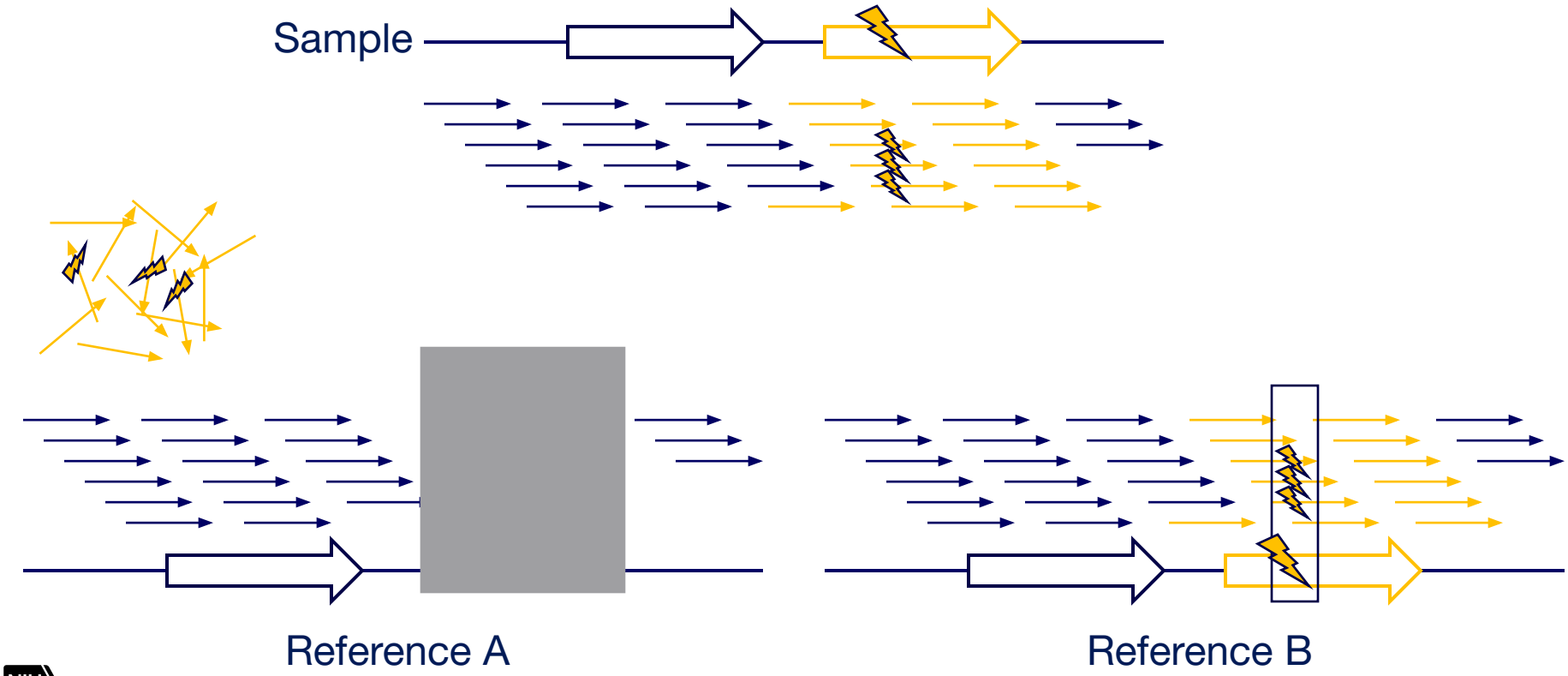
From T2T to Pan-Genome Reference

From T2T to HPR

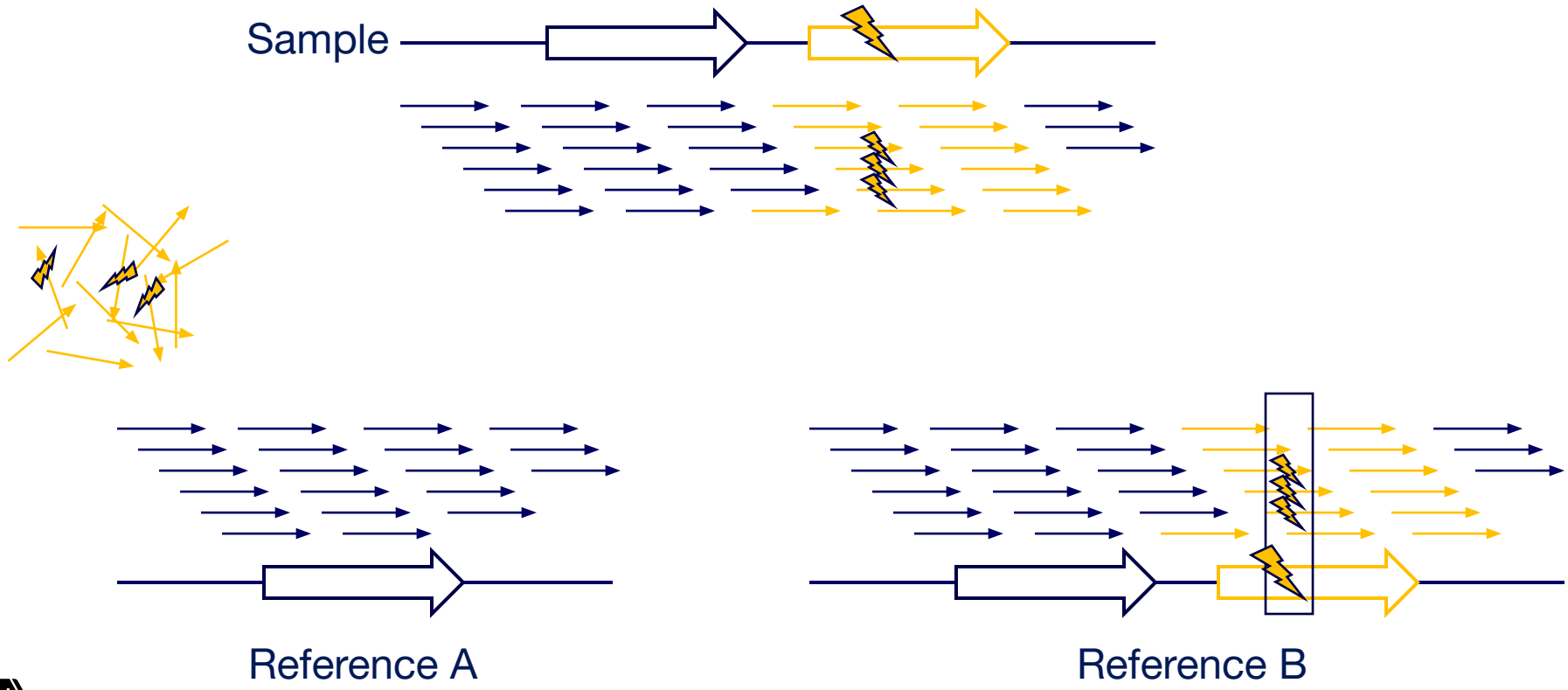
- **Telomere-to-Telomere Consortium**
- Enable complete, gapless assembly of human genomes
- “The complete sequence of a human genome” (2022)
- “The complete sequence of a human Y chromosome” (2022)

- **Human Pangenome Reference Consortium**
- Build a reference collection of 700+ T2T human haplotypes
- “A draft human pangenome reference” (2023)

T2T addresses the completeness bias



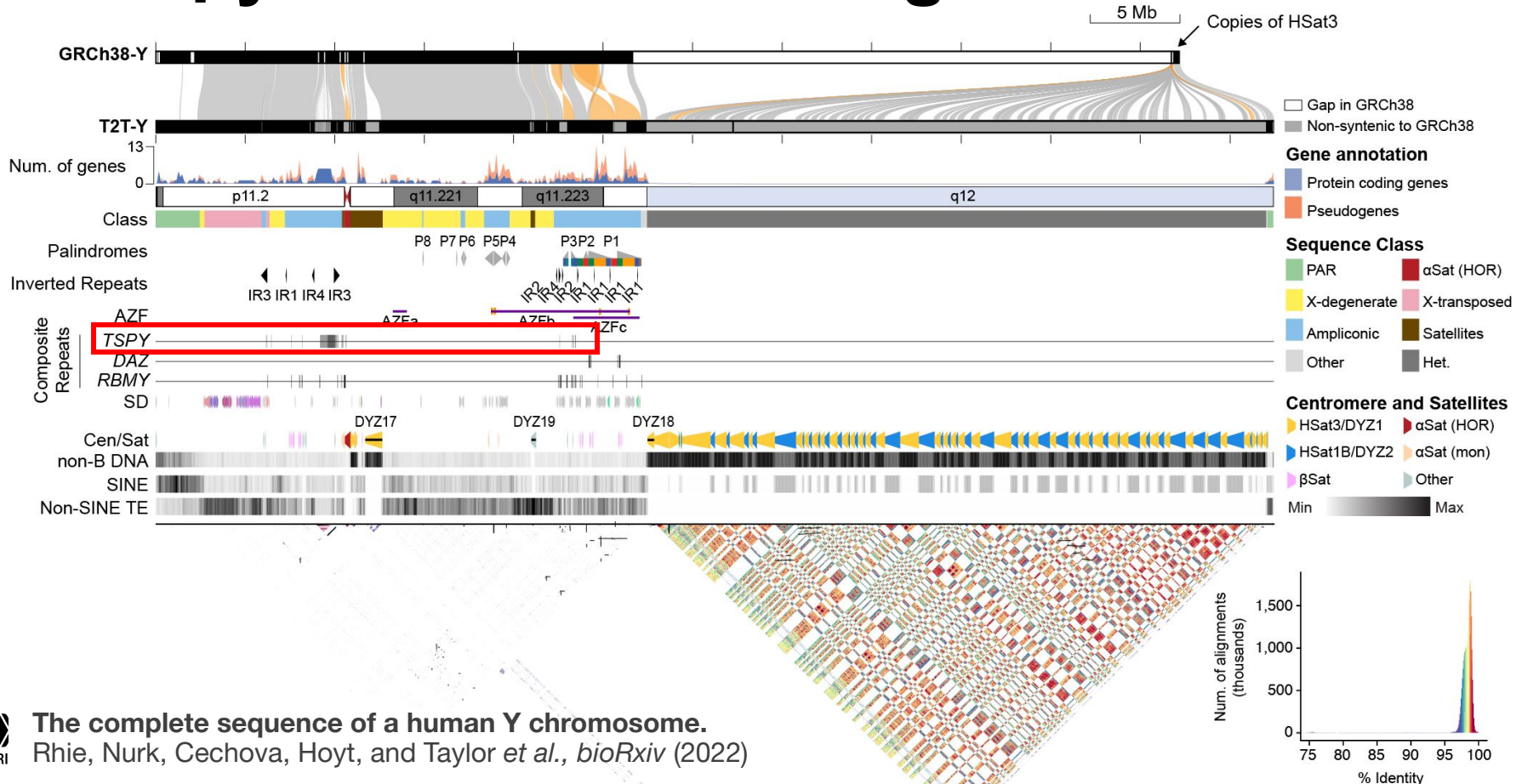
HPRC addresses the representation bias



NIH
NHGRI

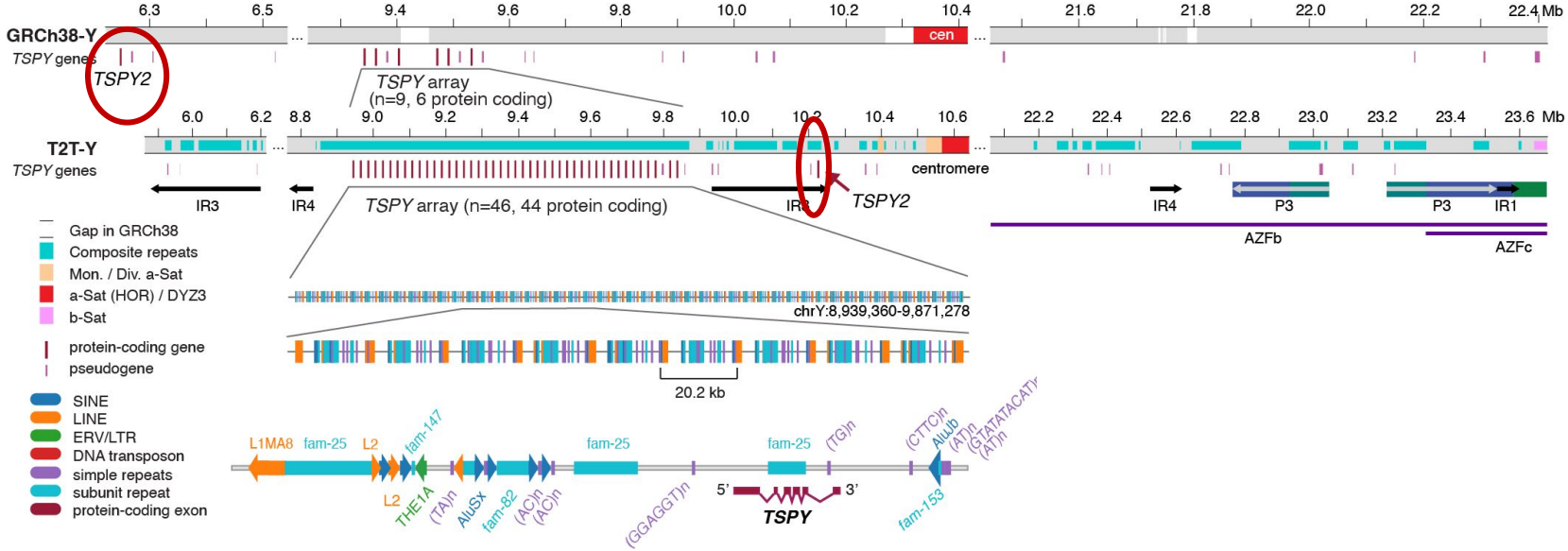
Lessons learnt from 1 T2T

1. Copy-number variable genes



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

TSPY gene annotation

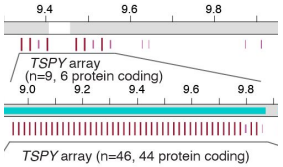


- HG002 has 45 *TSPY* protein-coding copies and many more pseudogenes
- 44 found in the *TSPY* composite array (8.9-9.8 Mb), ~50kb gap in GRCh38
- *TSPY2* at ~10 Mb (J and all other Y haplogroups), not at ~6.2 Mb (R and Q)

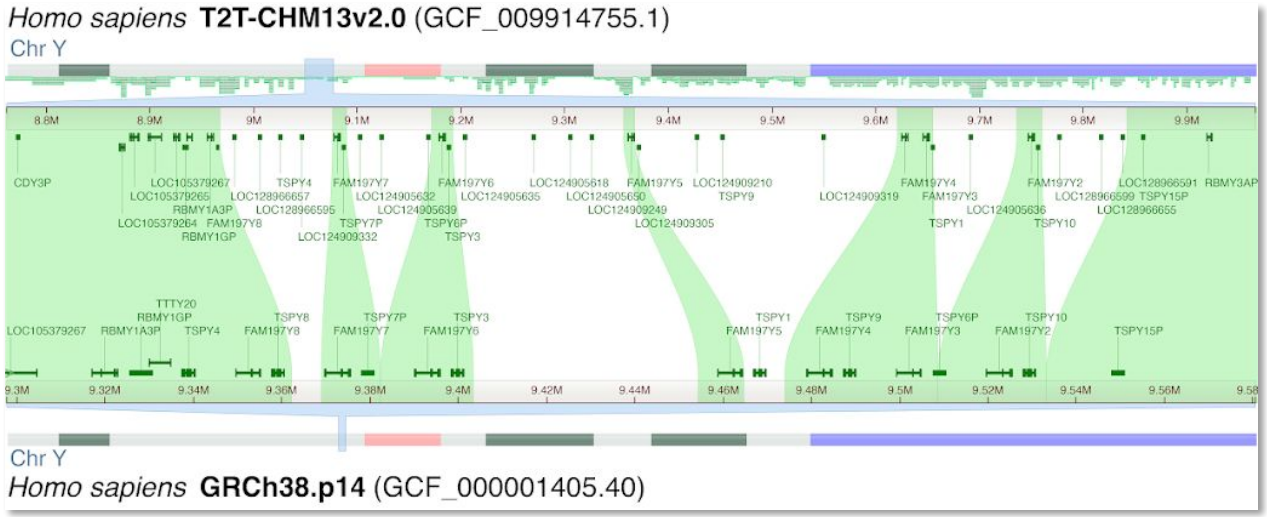
NIH NHGRI

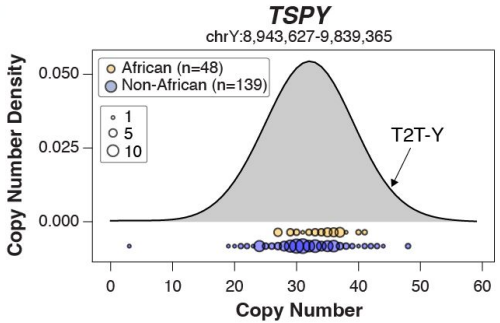
TSPY: Testis specific protein Y-linked; spermatogenesis

How do you want to name each *TSPY* copy?



- RefSeq gene annotation
  - All additional *TSPY* copies are named "*LOCXXXXXXXXX*"
  - *TSPY* is copy number variable among different individuals



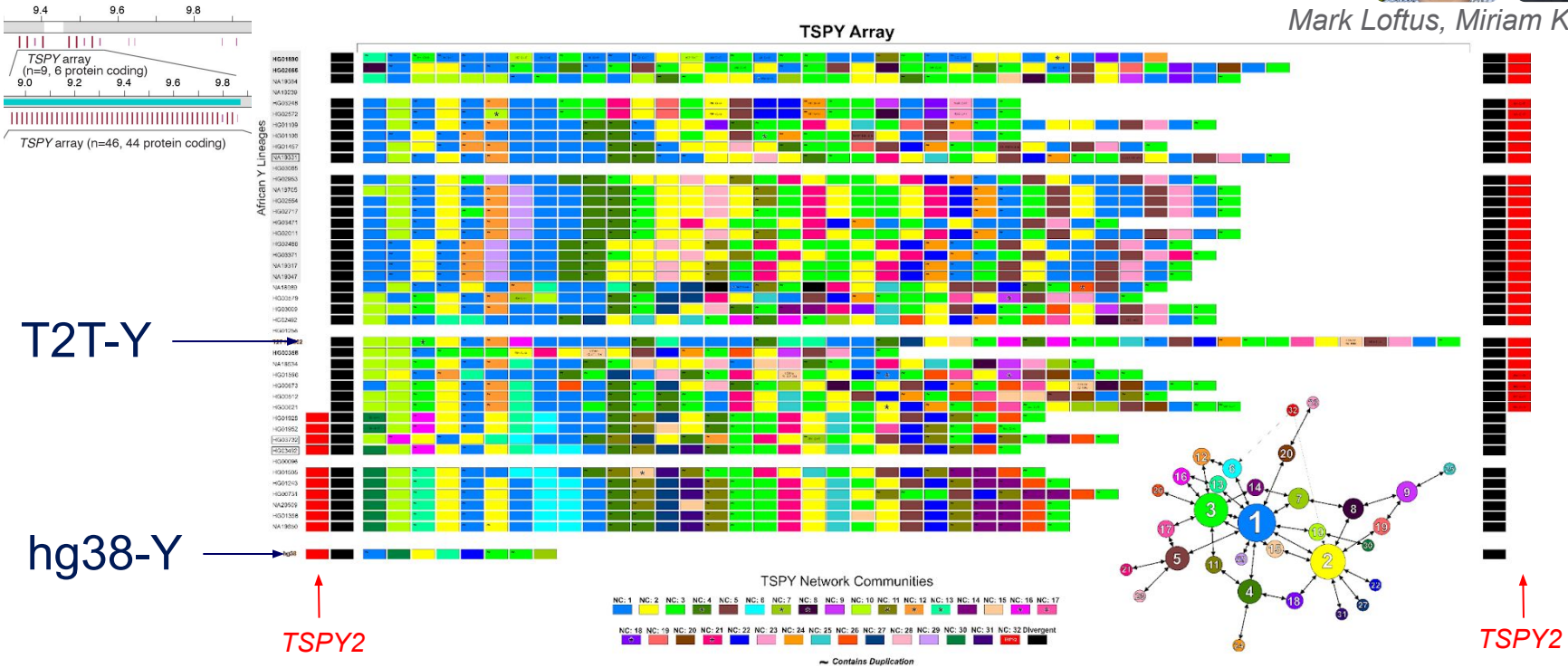


NIH NHGRI

Visualization: NCBI Comparative Genome Viewer

Not all *TSPY*s are the same, we need better representation

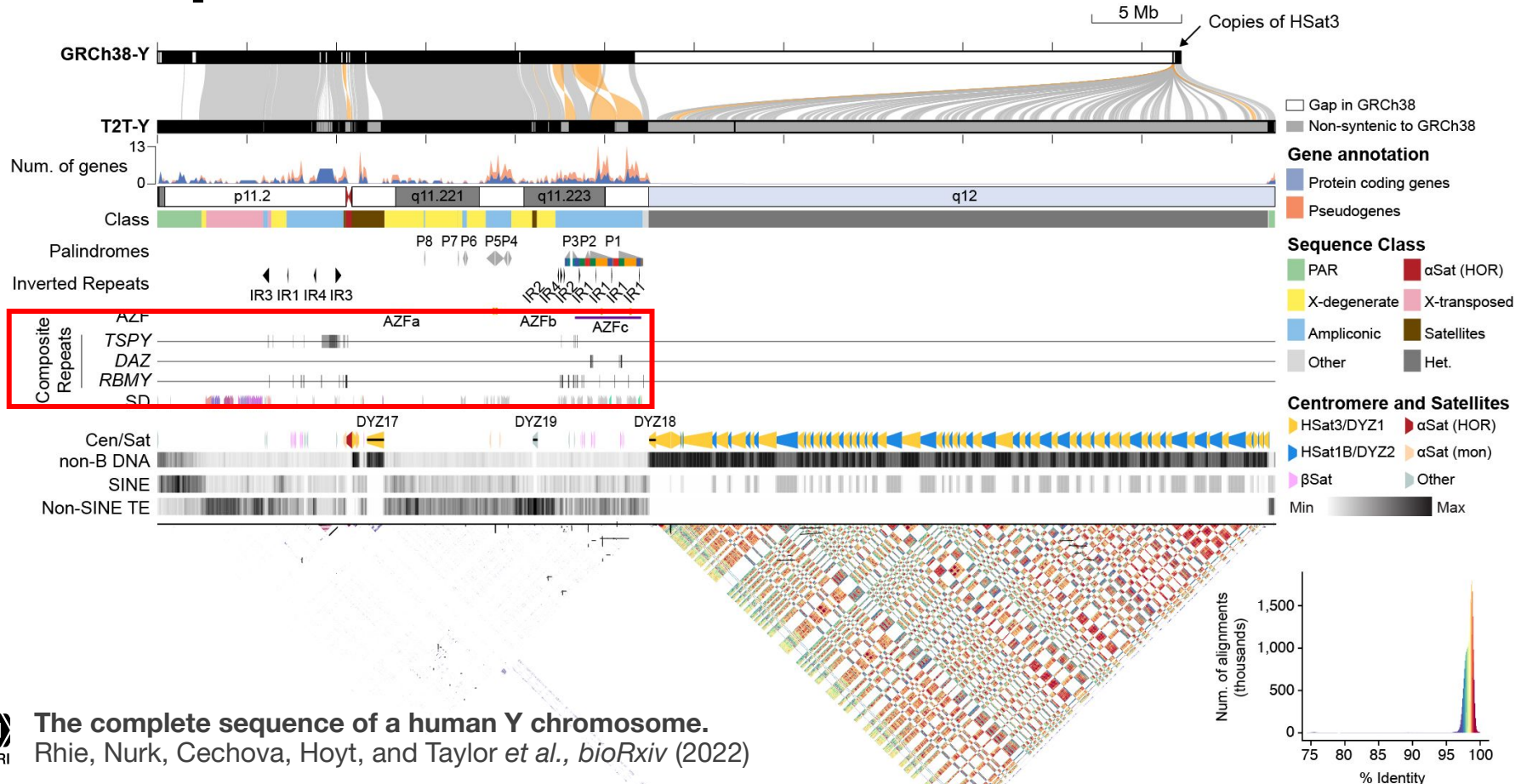
Mark Loftus, Miriam Konkel



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

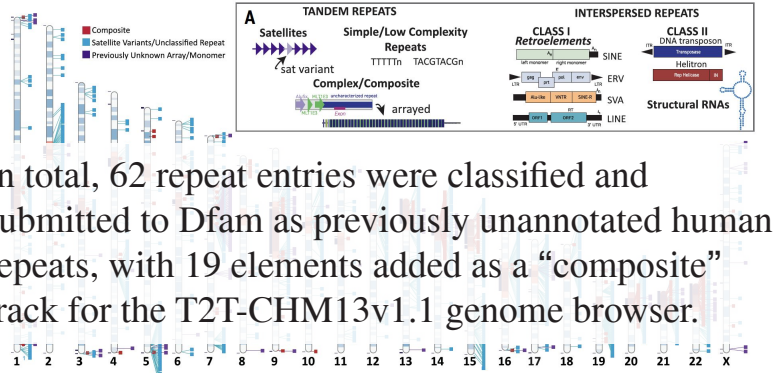
NIH NHGRI

2. Repeat annotation



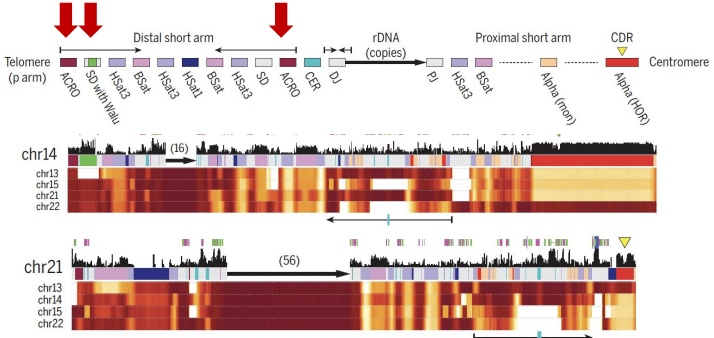
The complete sequence of a human Y chromosome.
Rhie, Nurk, Cechova, Hoyt, and Taylor *et al., bioRxiv* (2022)

Annotating the unknown

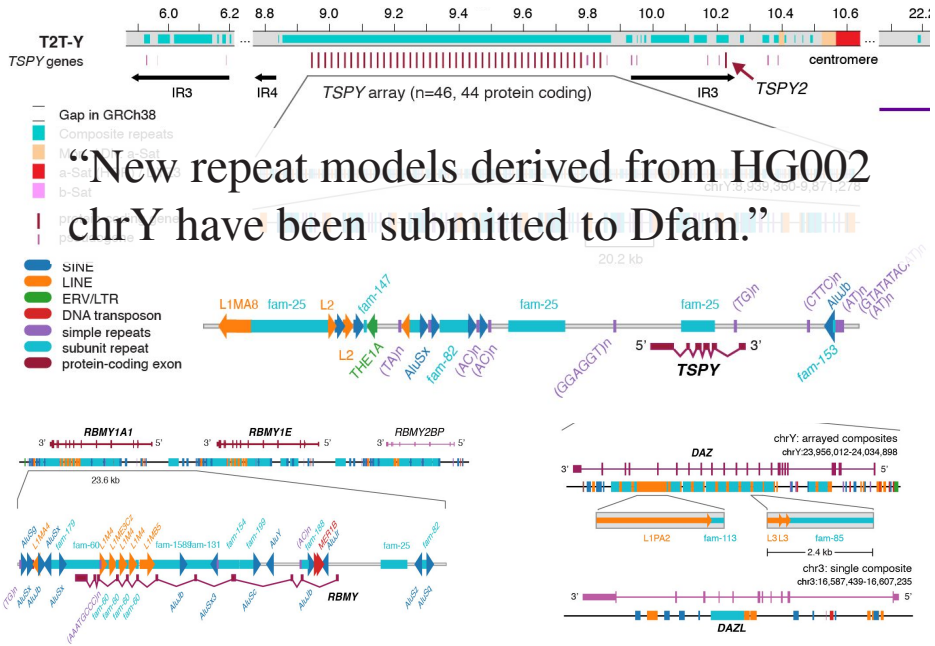


In total, 62 repeat entries were classified and submitted to Dfam as previously unannotated human repeats, with 19 elements added as a "composite" track for the T2T-CHM13v1.1 genome browser.

From telomere to telomere: The transcriptional and epigenetic state of human repeat elements.
Hoyt et al. *Science* (2022)



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

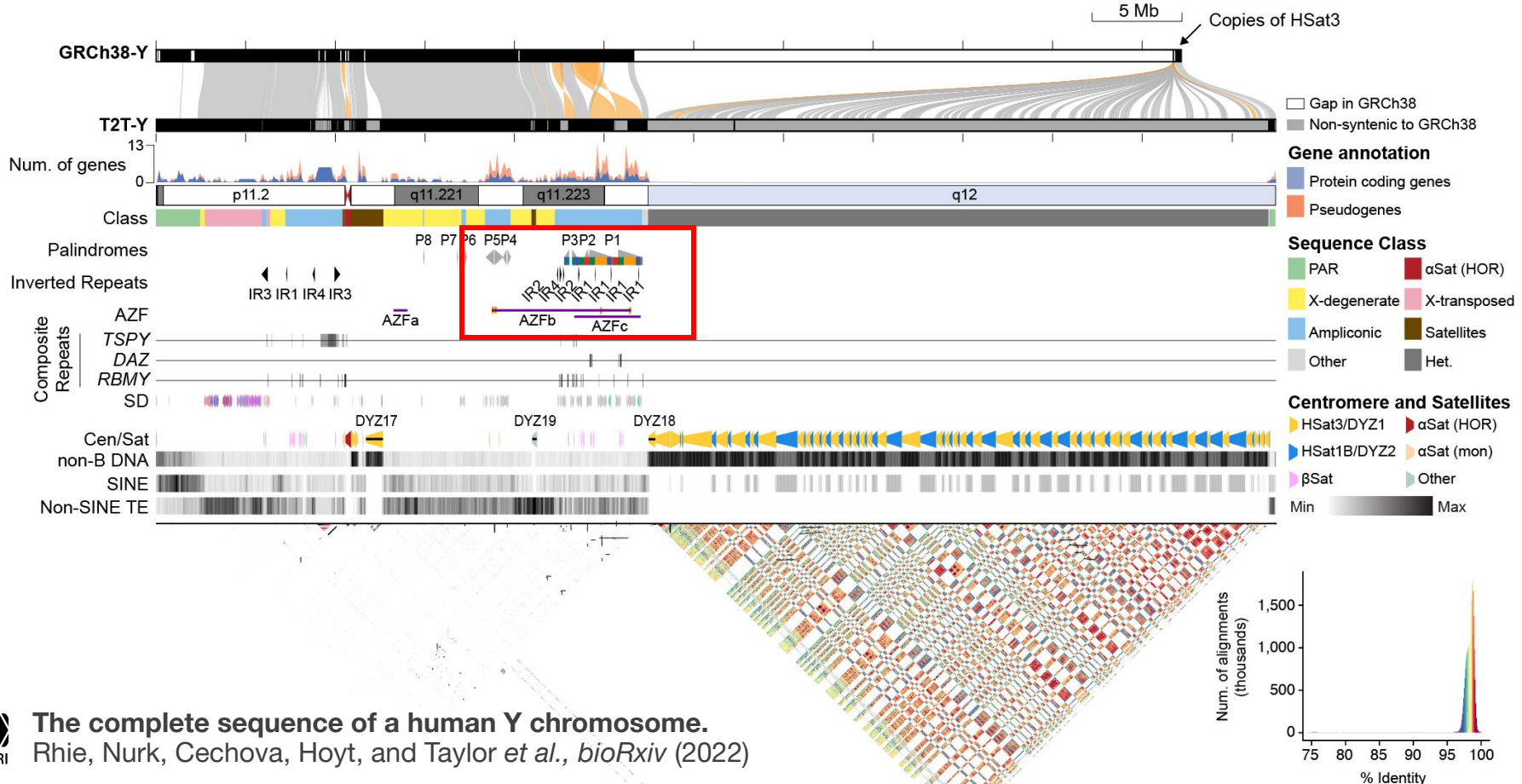


"New repeat models derived from HG002 chrY have been submitted to Dfam."

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

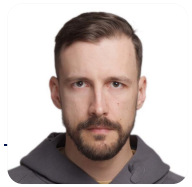
NIH NHGRI

3. Palindromes!

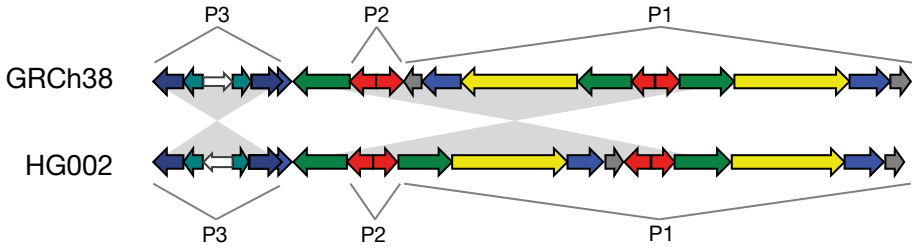


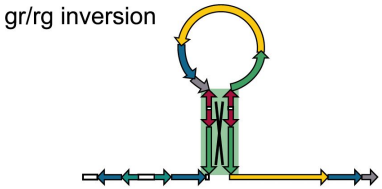
The complete sequence of a human Y chromosome.
Rhie, Nurk, Cechova, Hoyt, and Taylor *et al., bioRxiv* (2022)

Inversions in palindromes



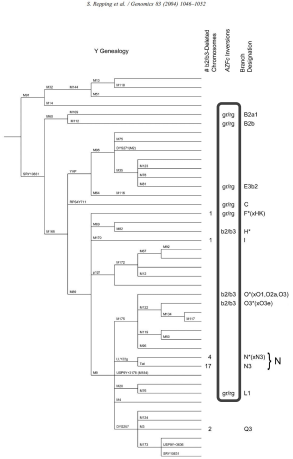
Most inversions are neutral, but things go wrong when deleted



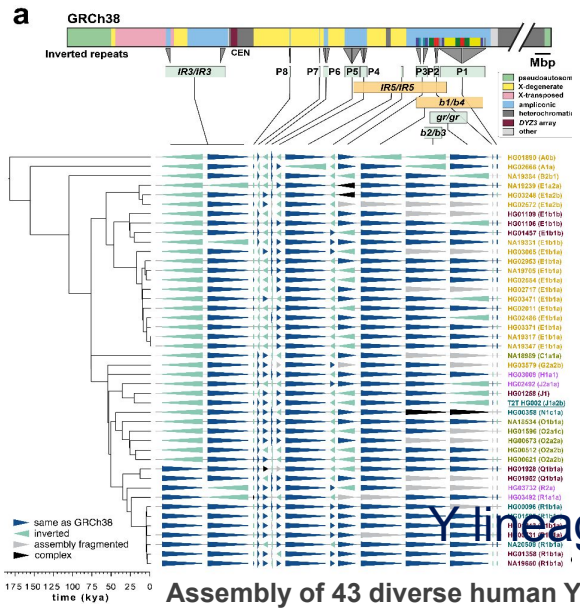


In 6 / 44 genealogical branches

Repping et al. Genomics (2004)



- P1-P2 gr/rg inversion
- P3 inversion



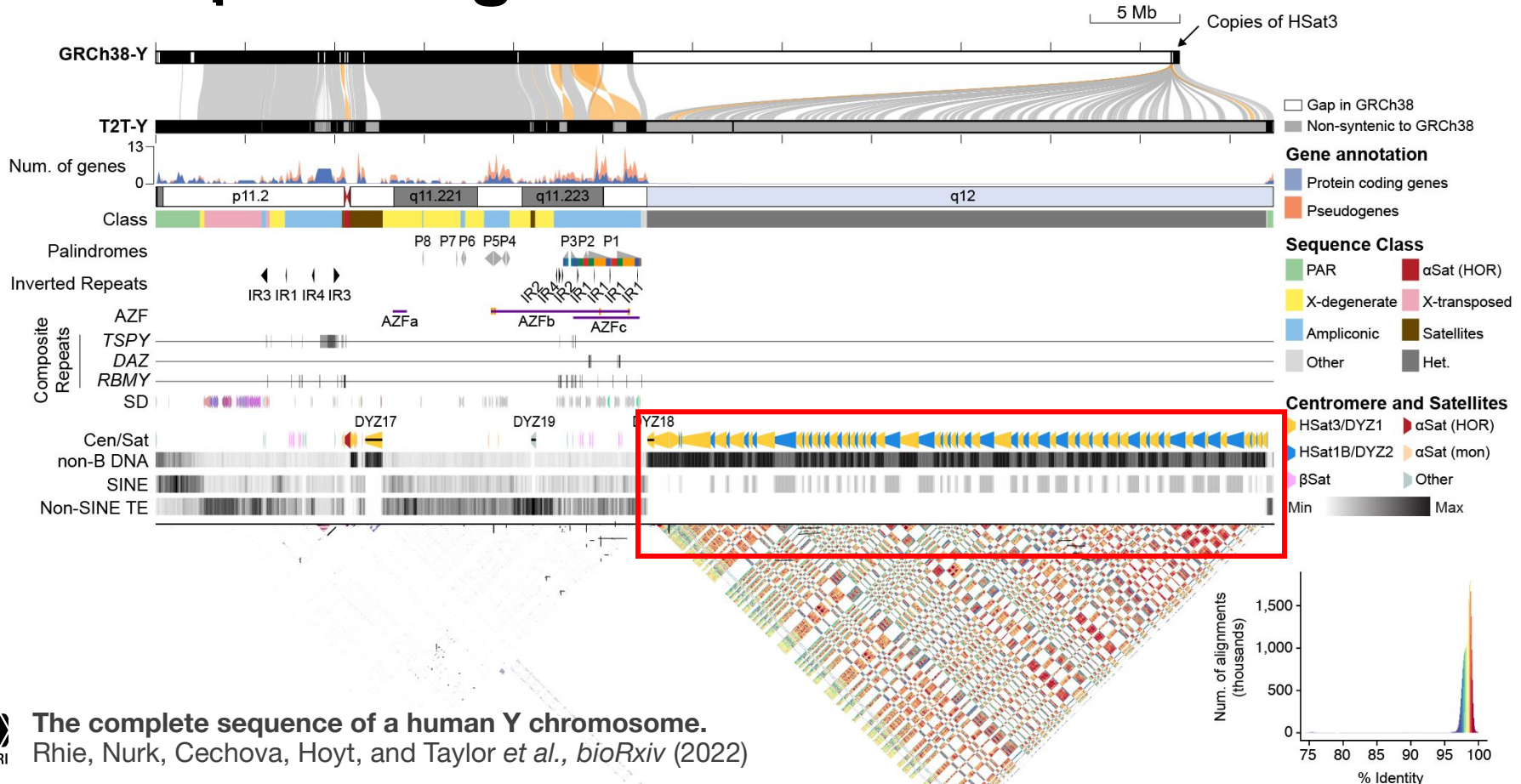
Y lineage specific?

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

NIH NHGRI

Full of ampliconic genes *RBMY, DAZ, BPY, CDY*, …

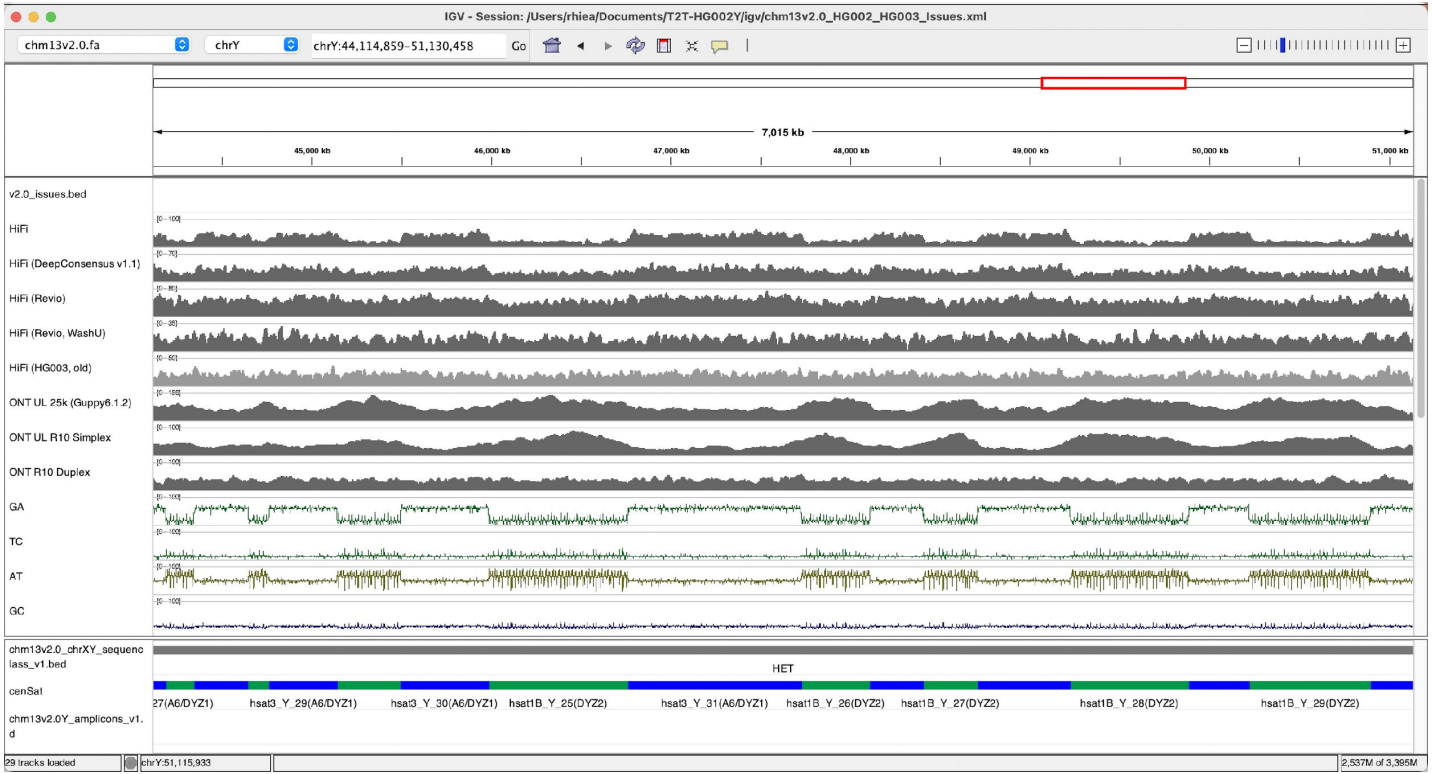
4. Sequencing biases



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

Assemblies will never be the same

SMRTbell Express Template Prep Kit
2.0
3.0
1.0
UL Protocol
R10 Duplex



Revio data: PacBio, Katy Munson, Bob Fulton. ONT R10 data: Mitten Jain, HPRC

Duplex, good or bad?

Minknow software sees what looks like a blocked pore and kicks the molecule out

ONT is actively working to fix issue

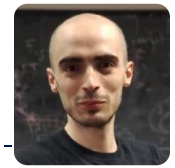
Remaining assembly challenges

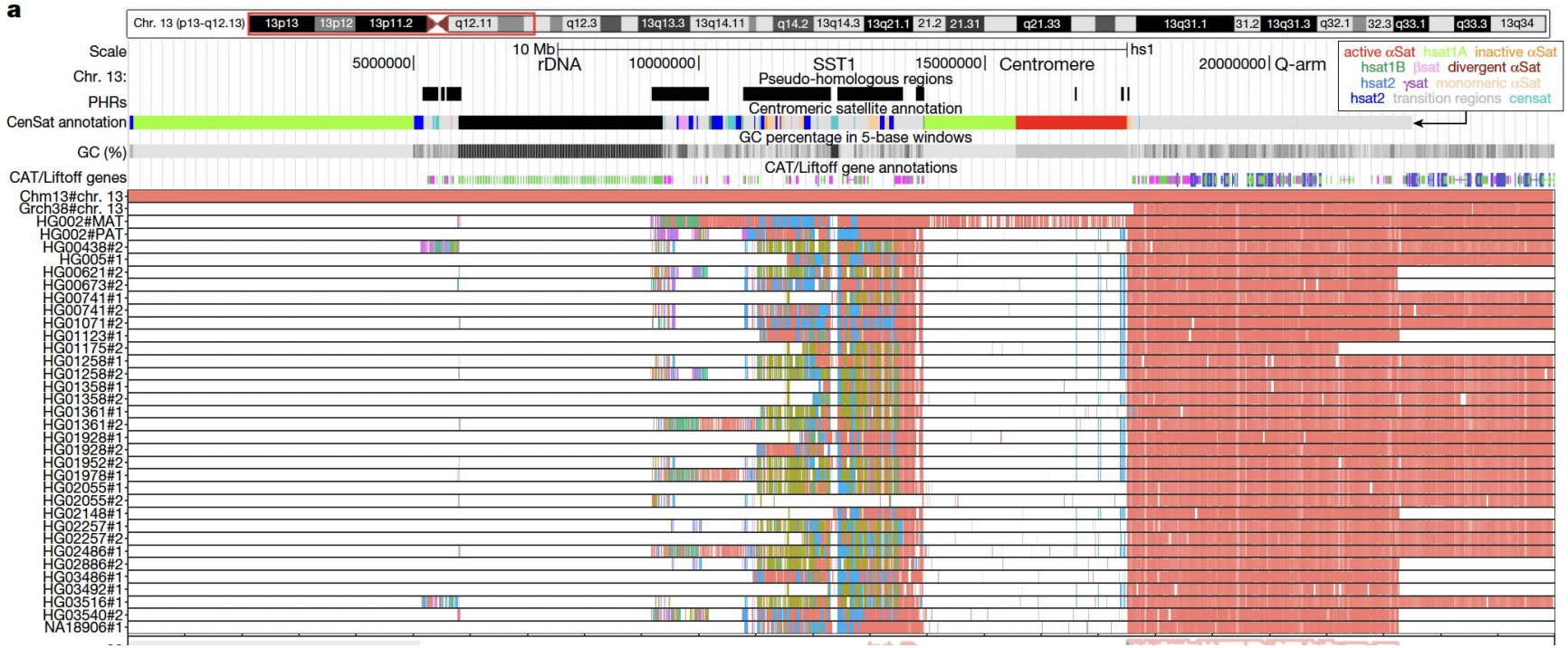
- **Pairing rDNA distal junctions to the proximal**
  Major bottleneck in any genome assembly
  5 acrocentric chromosomes x 2 arms in human
  Additional data can help: Hi-C, Pore-C, Strand-Seq
- **Polishing diploid genomes**
  Benchmark HG002 diploid genome (HG002-Q100)

Opportunities in the T2T pan-genome era

NIH
NHGRI

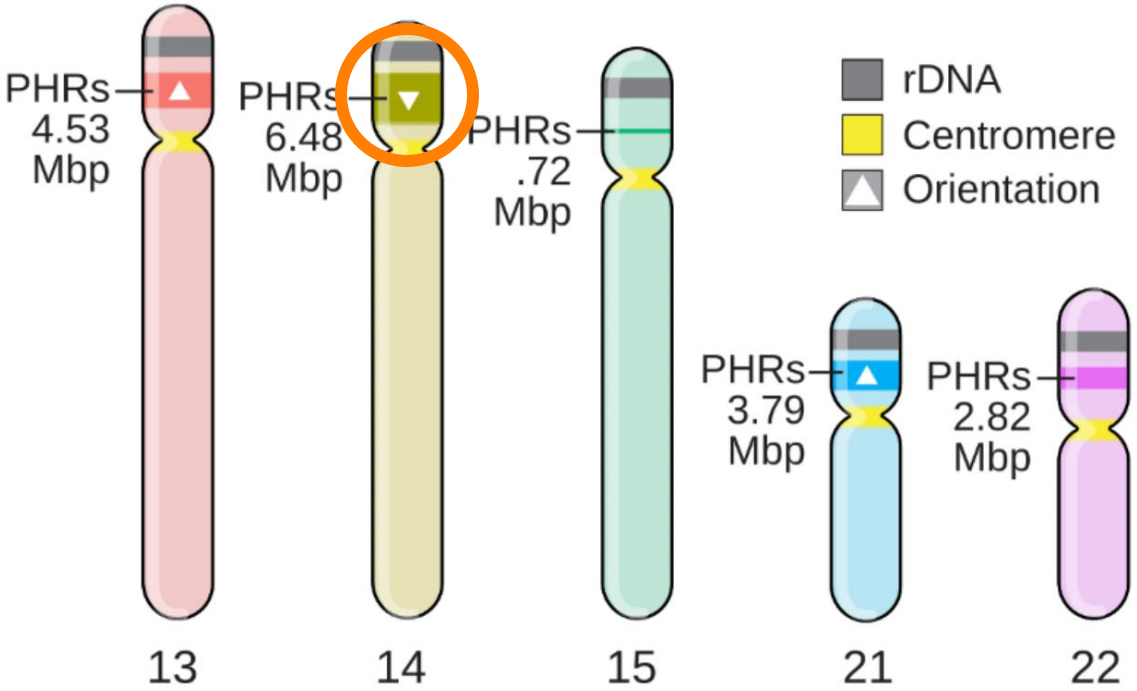
Pseudo-homologous region in acrocentrics





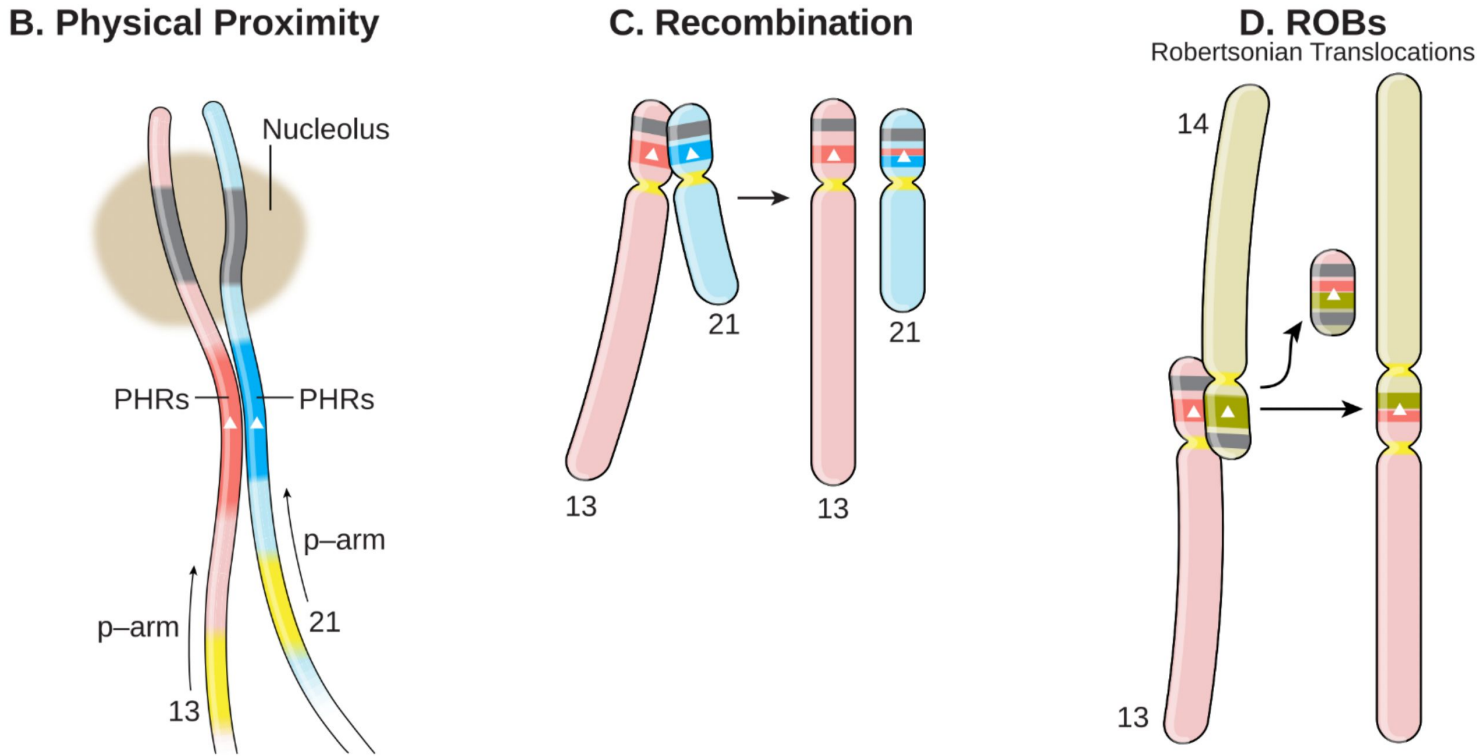
Recombination between heterologous human acrocentric chromosomes
Guarracino, … Gerton, Phillippy, Colonna, Garrison. *Nature* (2023)

PHRs on the acrocentric short arms



Recombination between heterologous human acrocentric chromosomes
Guarracino, … Gerton, Phillippy, Colonna, Garrison. *Nature* (2023)

Recombination of the acrocentrics



Recombination between heterologous human acrocentric chromosomes.
Guarracino, … Gerton, Phillippy, Colonna, Garrison. *Nature* (2023)

Summary

- **One human genome *finished T2T***
  - 8% is newly accessible with long reads, only 1% with short
- **Working toward a global *pangenome***
  - More sequence = more variants = more associations
  - Challenges ahead:
    multi-copy gene annotation, repeat annotation, palindromes, sequencing biases
- **The genome is *more dynamic* than appreciated**
  - Happening precisely in the hardest to assemble regions
- ***You are your own best reference***
  - Most complete and accurate way to measure somatic variants

Team T2T, HPRC, (…and many more)


The Forefront
of Genomics®

Please Vote ☺

PARTNERSHIP FOR PUBLIC SERVICE

SAMUEL J. HEYMAN SERVICE TO AMERICA MEDALS

People's Choice A

How it Works

1. Read the finalists' profiles to learn about their accomplishments,
2. Vote for your top picks using the poll on this page. You may select as many finalists as you like.
3. Vote again! You may submit your vote once every 24 hours, and you can vote for the same or different finalists each time.
4. Throughout the voting period, we will narrow the field and reset the poll for the remaining top contenders.

Important Dates

- **Monday, May 8**: Voting Begins
- **Monday, May 29:** Top 12 Announced
- **Monday, June 19:** Top 6 Announced
- **Friday, June 30:** Voting Closed

The Sammies People's Choice award gi
year's finalists.

The People's Choice selection is based solely on the
from the selection of the category award winners. A
and the category award for which they are a finalist.

Vote now!

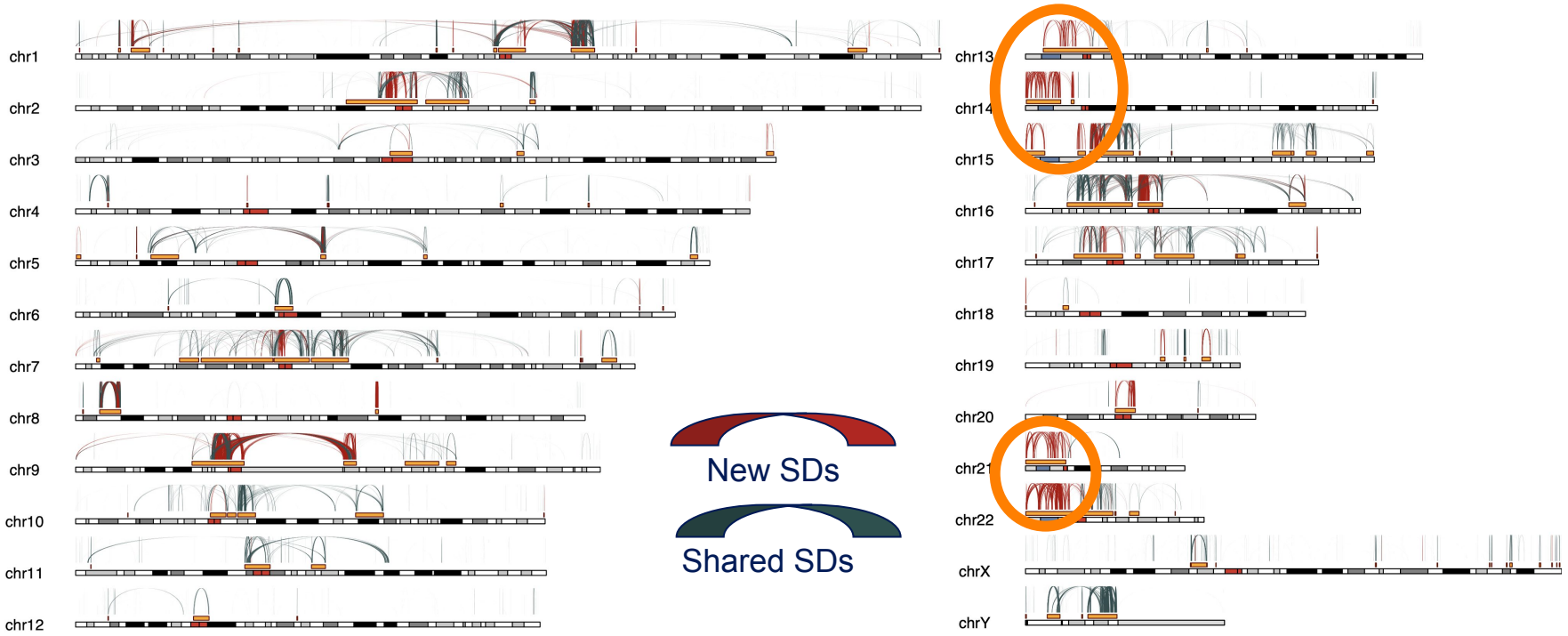
Click on name(s) below to vote for your favorite Sammies finalist(s)!

- [x] Adam Phillippy, Sergey Koren, Arang Rhie and the Telomere-to-Telomere Team (NIH)
- [ ] Anne Lord Bailey and Caitlin Rawlins (VA)
- [ ] Blane Workie, Robert Gorman, Jessica Ilich and the Aviation Consumer Protection Team (DOT)

Google "Sammies people's choice award"

NIH NHGRI

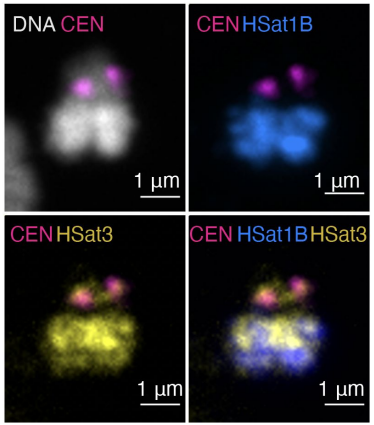
60% of T2T SDs are new or structurally different from GRCh38

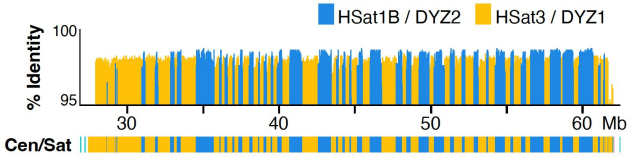


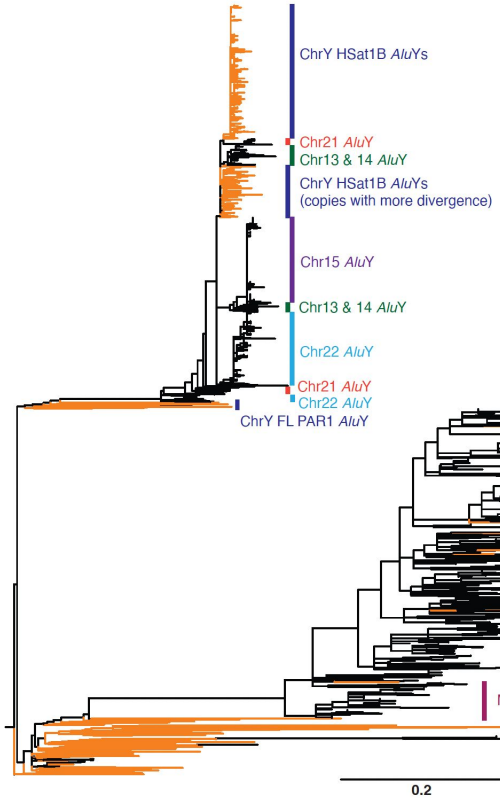
Segmental duplications and their variation in a complete human genome.
Vollger, et al. *Science* (2022)

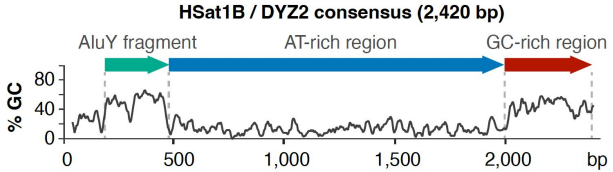
NIH NHGRI

Yqh and *Alu*Ys







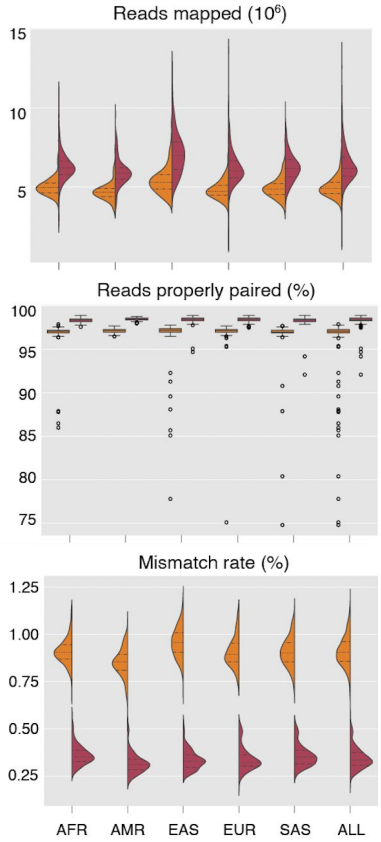


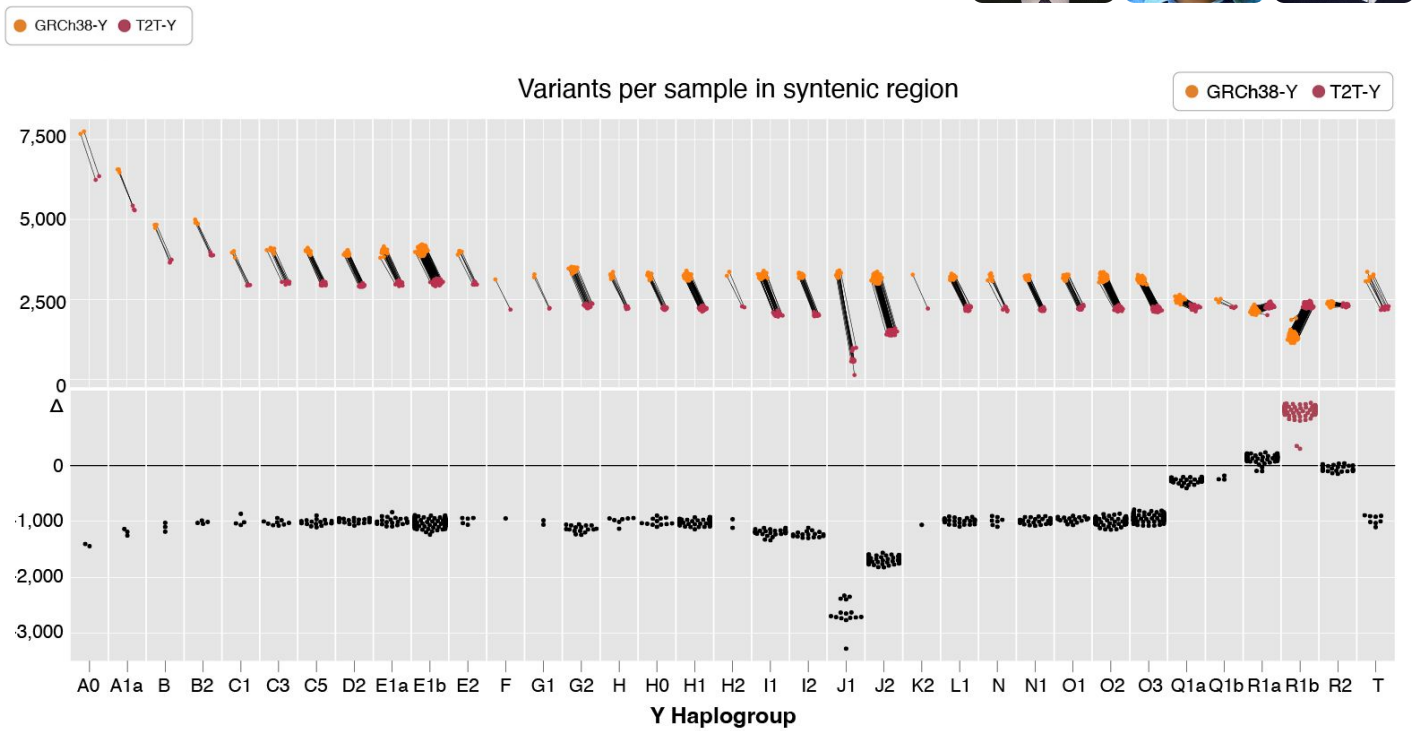
*Alu*Ys intermingled between HSat3s in DYZ1

Initial seeding events that created these arrays were independent of one another, yet were derived from *Alu*Y elements from PAR1

NIH NHGRI

T2T-Y improves variant calling

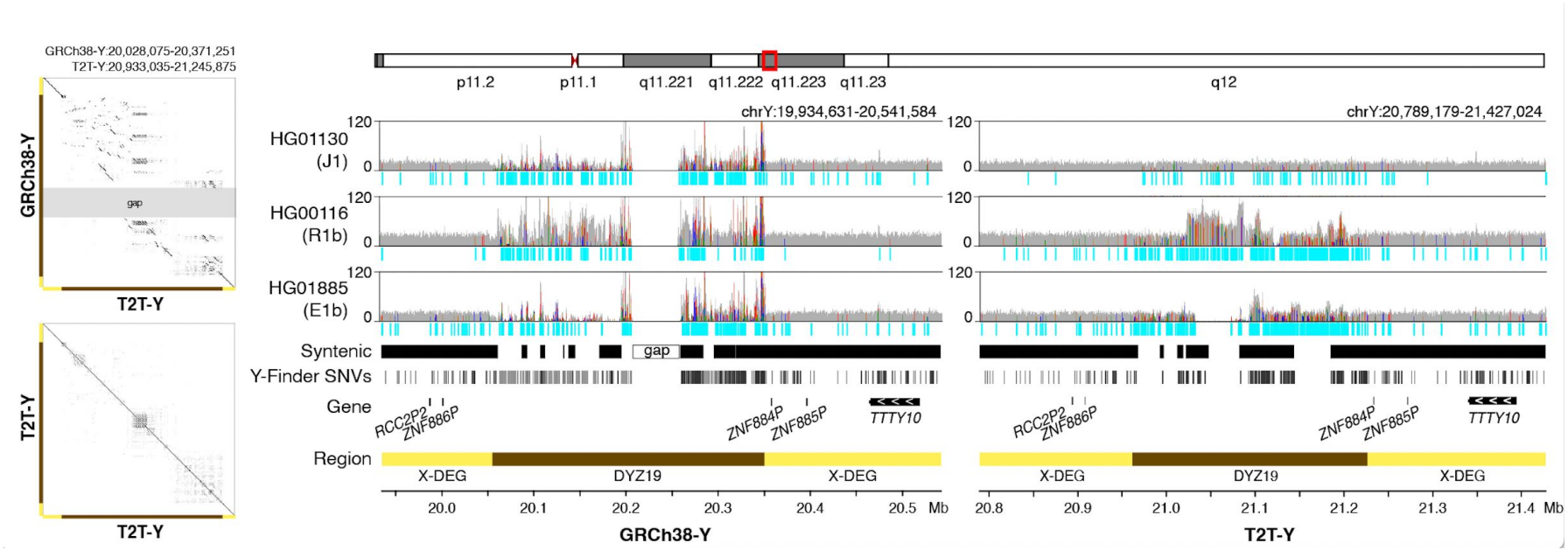




Unrelated 1,233 XY individuals from 1000 Genomes Project

NIH NHGRI

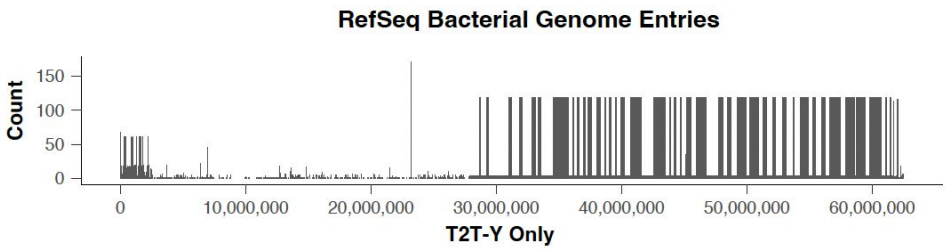
Example: Variant calling around DYZ19

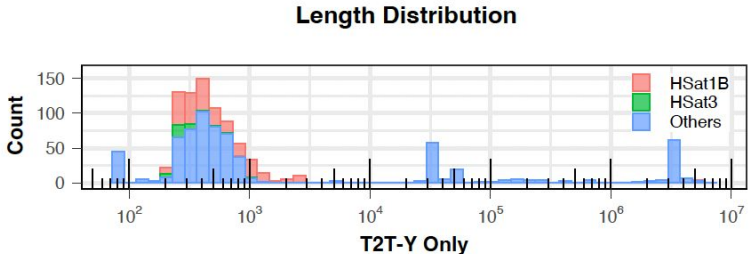


Y adds 578 kb of accessible sequence (+4.2%)

What about non-human?

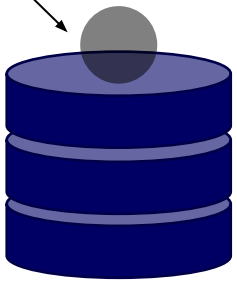
Human Y contamination in Bacterial Genomes





"What's in my microbial sample, collected from a human?"

Human Y mislabeled



RefSeq Bacterial DB

Remove sequences aligning to GRCh38 / hg19

Query against RefSeq Bacterial DB

Identify what's abundant

~~"Male samples have more abundant microbial community of … !"~~