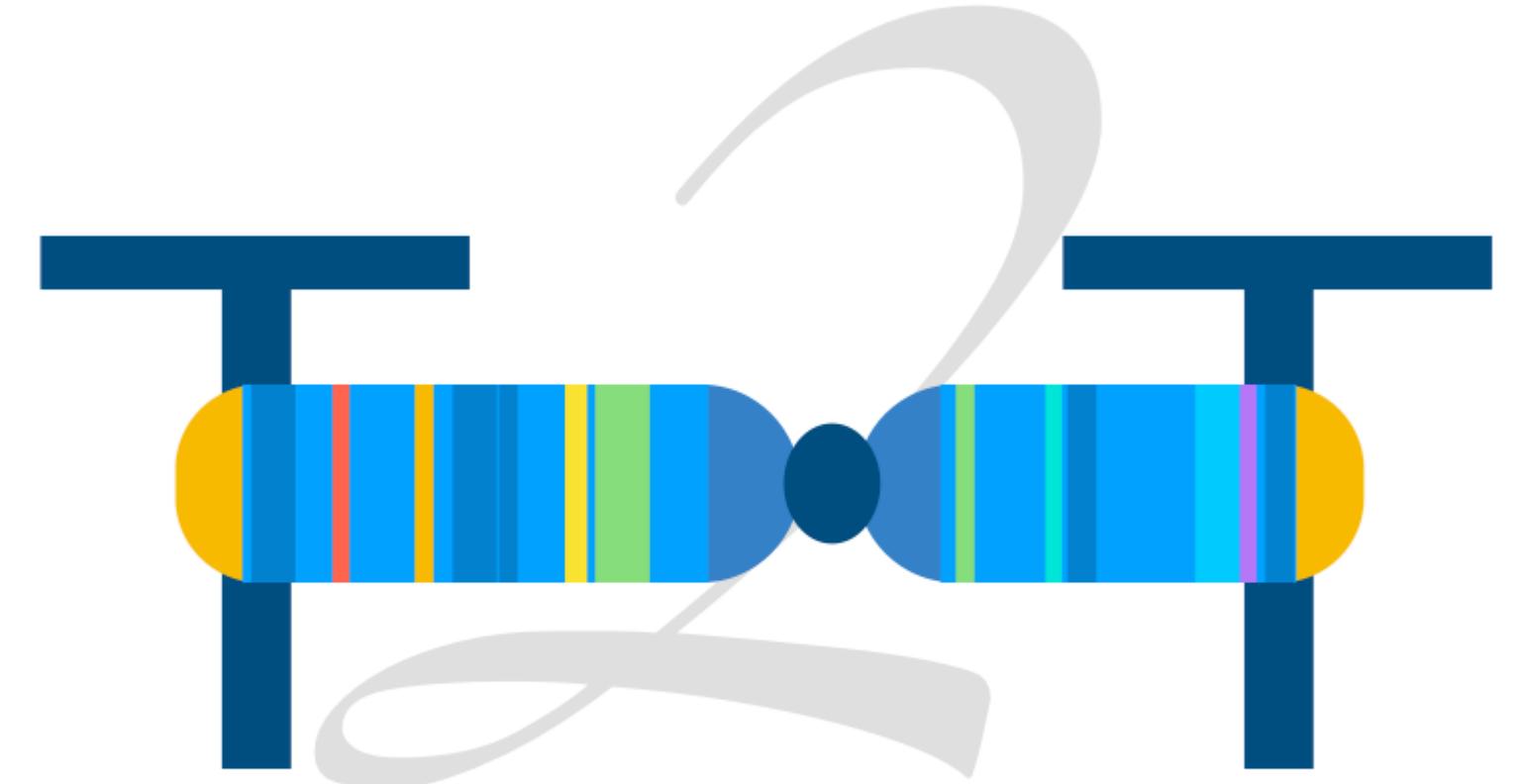


# The T2T Genome and applications to variant calling

Dylan Taylor  
Johns Hopkins University  
November 18, 2022



---

TELOMERE-TO-TELOMERE CONSORTIUM

# Lesson Overview

## Lecture Portion

- What is a T2T Genome?
- Benchmarking on a diverse dataset
- Large-scale cloud analysis of short-read data
- Improvements to alignment/variant calling
- A T2T Y chromosome
- Future directions

## Live-coding Portion

- Explore improvements to short-read alignment
  - Improvements in medically-relevant genes
- Run variant calling on short-read alignments
- Explore improvements to variant calling

# Lesson Overview

## Lecture Portion

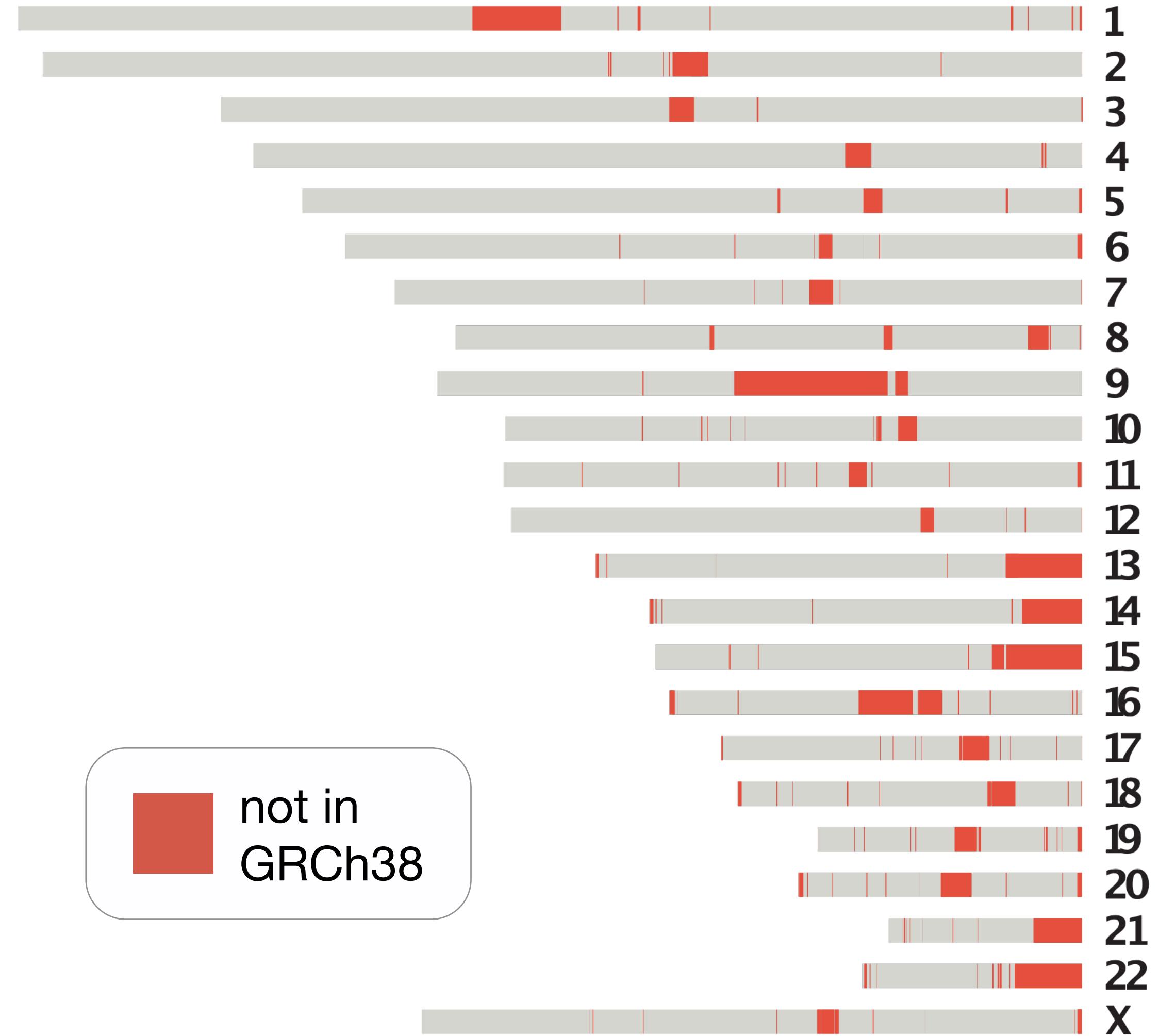
- What is a T2T Genome?
- Benchmarking on a diverse dataset
- Large-scale cloud analysis of short-read data
- Improvements to alignment/variant calling
- A T2T Y chromosome
- Future directions

## Live-coding Portion

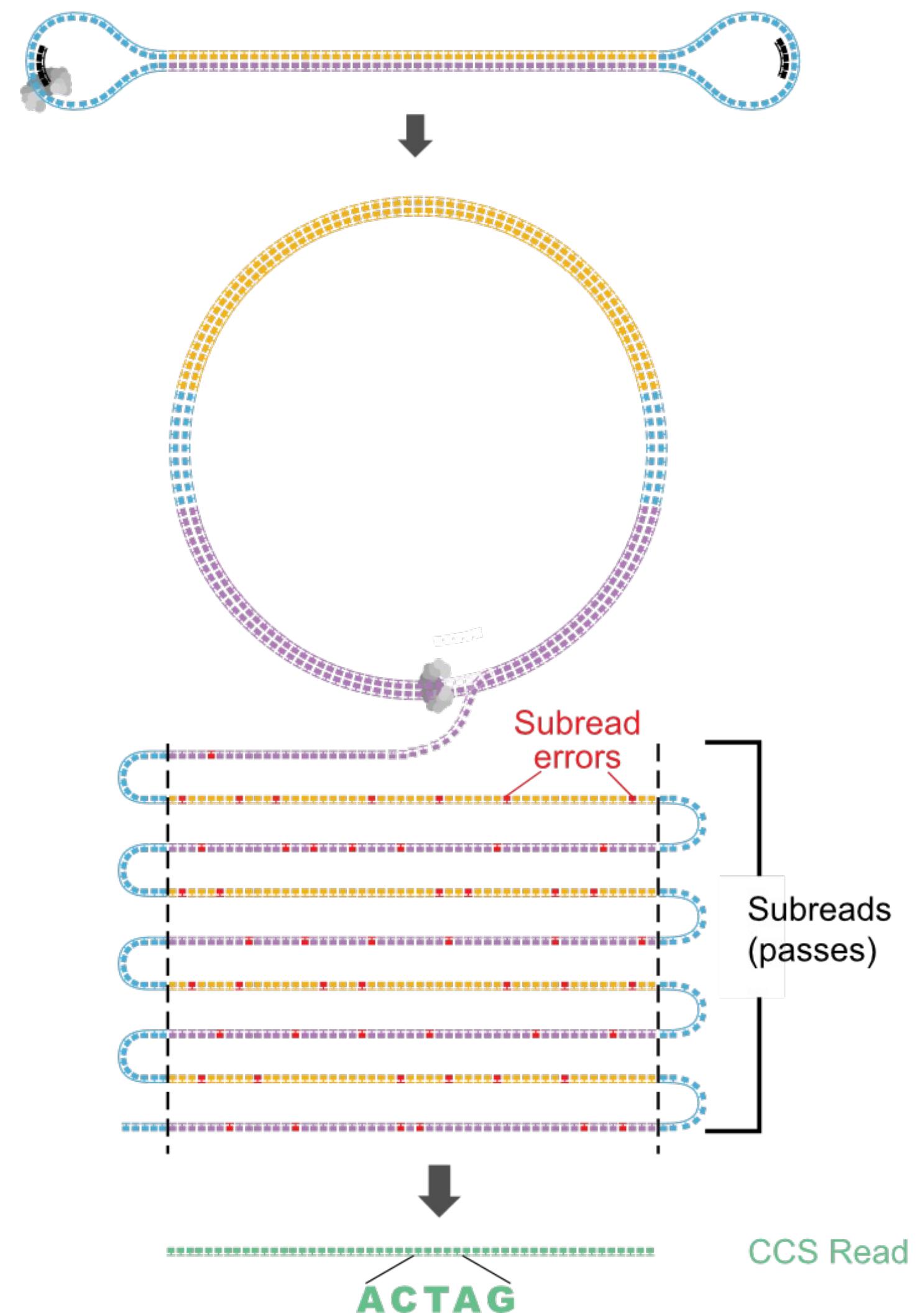
- Explore improvements to short-read alignment
  - Improvements in medically-relevant genes
  - Run variant calling on short-read alignments
- Explore improvements to variant calling

# The “previous” human reference genome is incomplete

- The human reference genome was first drafted in 2000, using short-read sequencing
- **GRCh38** has missing, incorrect, and/or artificial sequences
  - Centromeric regions
  - Segmental duplications
  - rDNA arrays

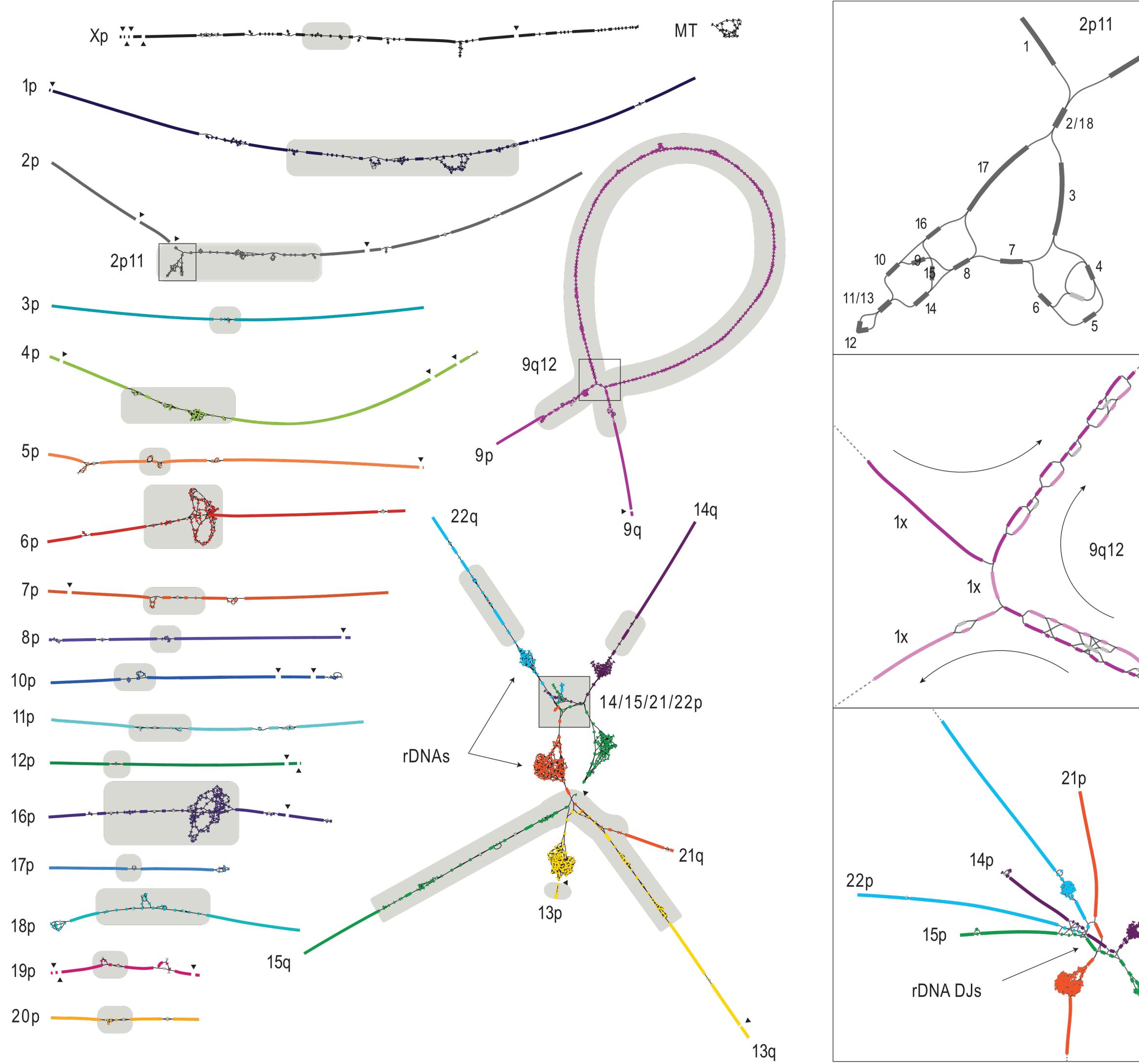


# Assembling a complete Telomere-to-telomere genome



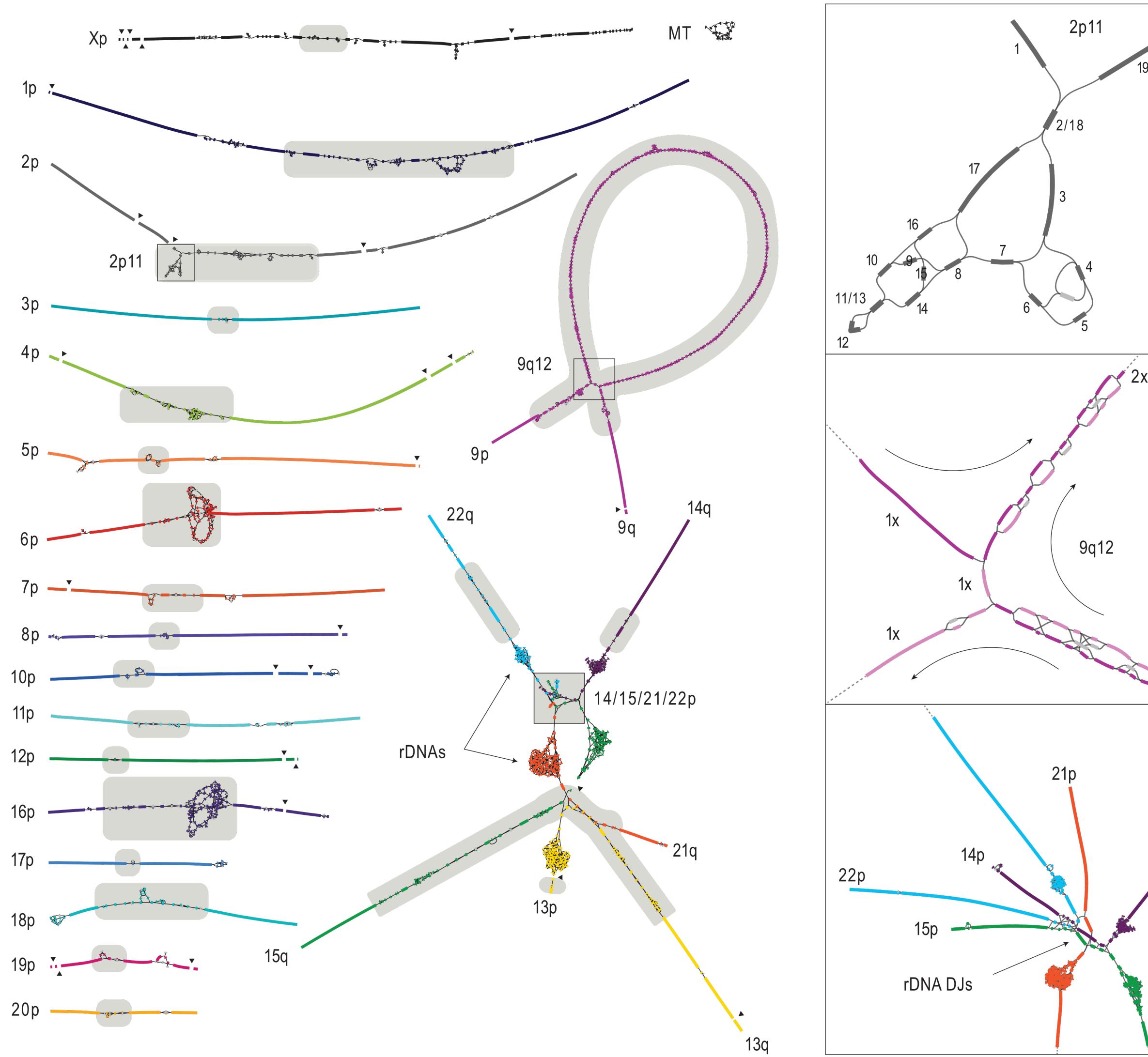
- Begin with highly accurate PacBio HiFi reads

# Assembling a complete Telomere-to-telomere genome



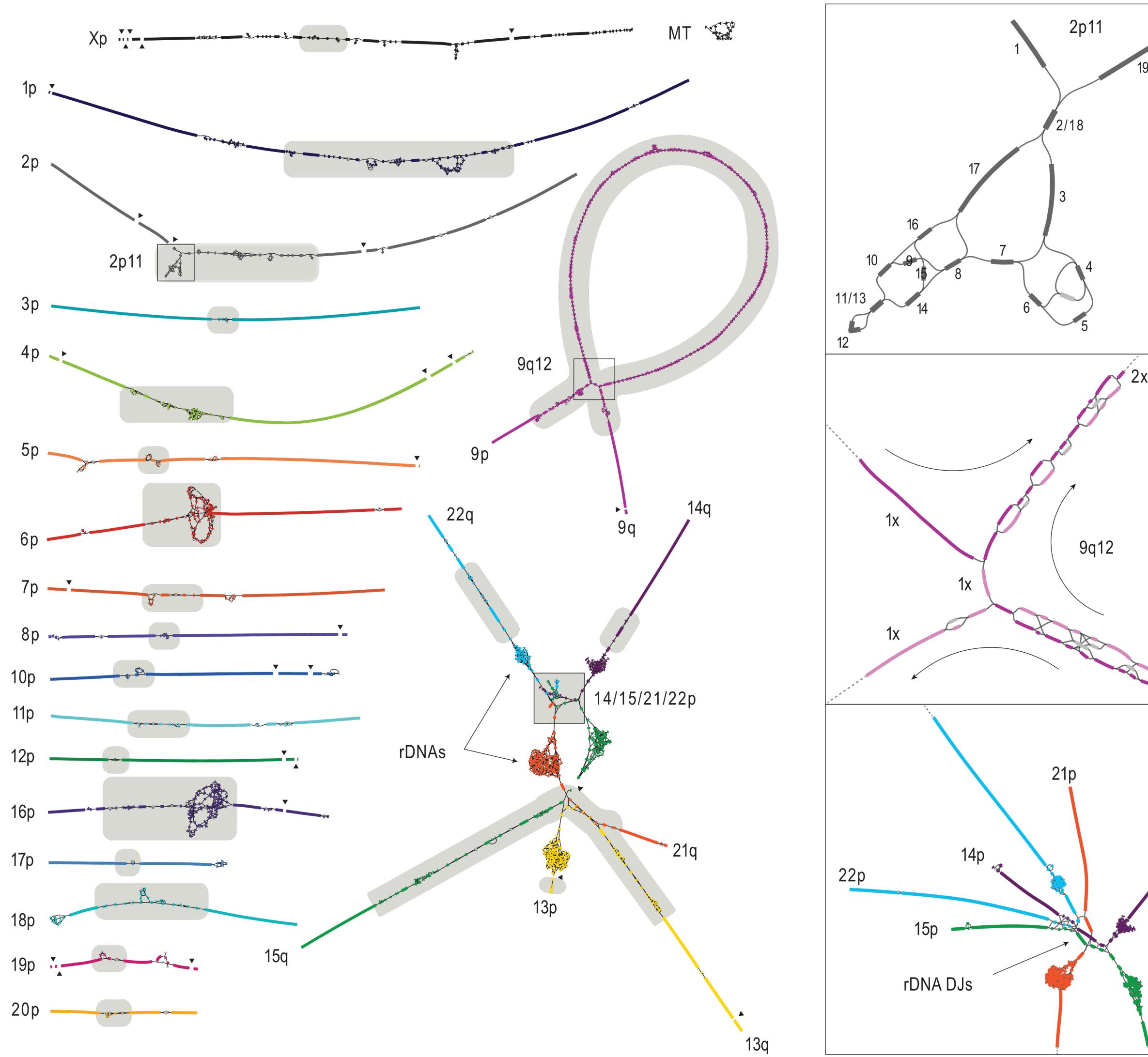
- Begin with highly accurate PacBio HiFi reads
- Using these, build high-resolution assembly string graph

# Assembling a complete Telomere-to-telomere genome



- Begin with highly accurate PacBio HiFi reads
- Using these, build high-resolution assembly string graph
- Most chromosome assemblies are nearly linear

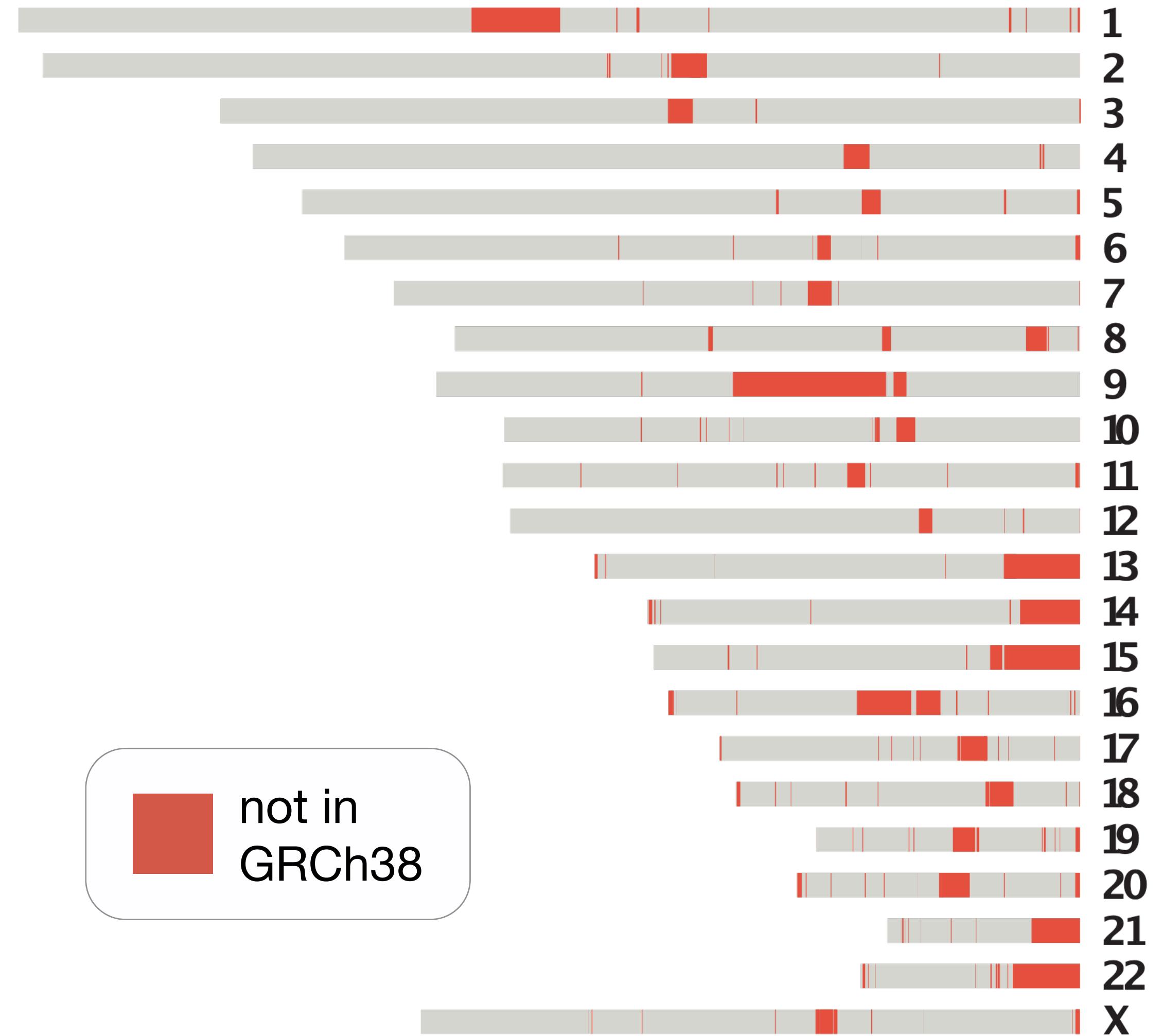
# Assembling a complete Telomere-to-telomere genome



- Begin with highly accurate PacBio HiFi reads
- Using these, build high-resolution assembly string graph
- Most chromosome assemblies are nearly linear
- Ultralong ONT reads + read coverage used to resolve loops in the string graph

# T2T-CHM13 is the *complete* sequence of a human genome

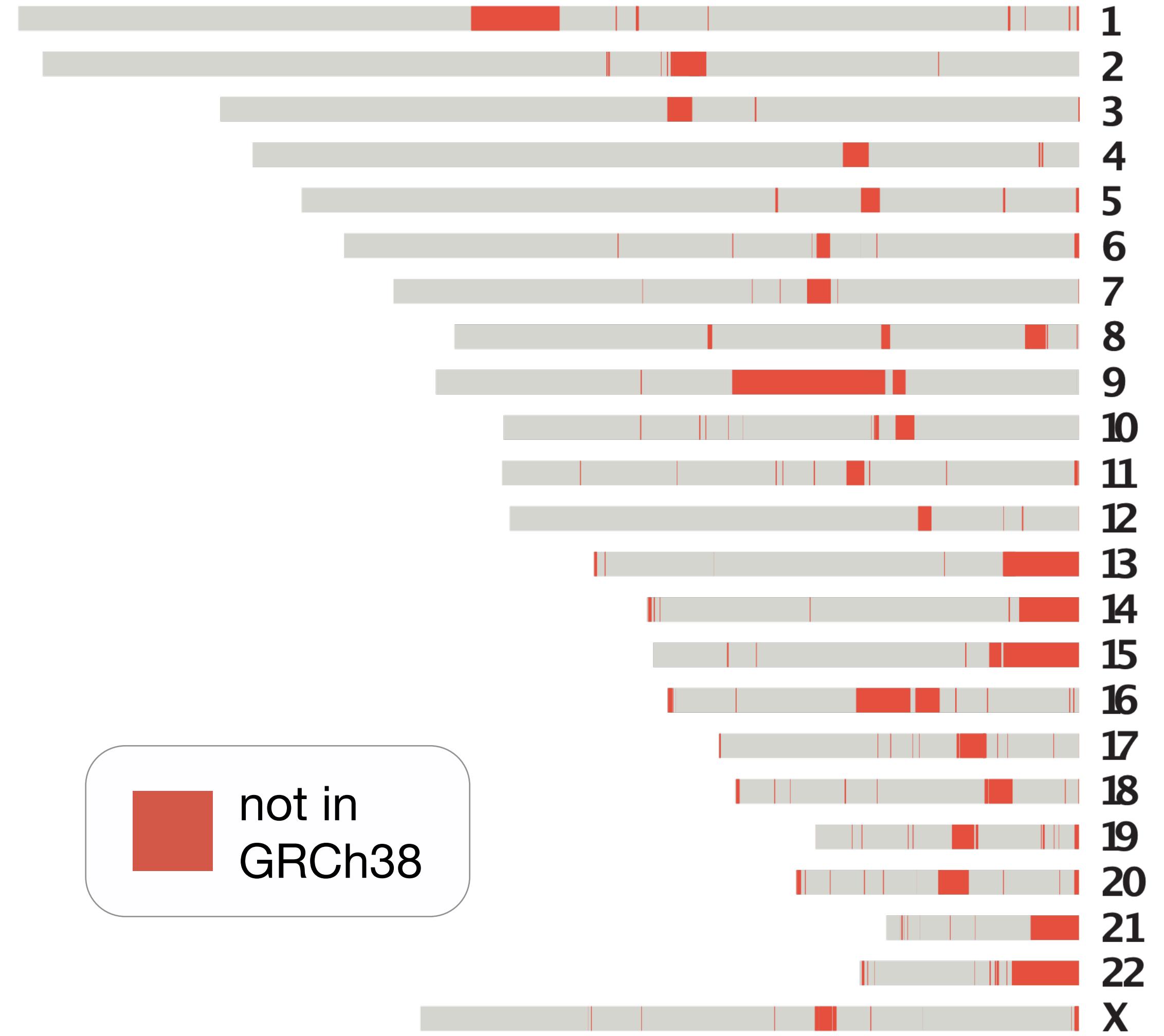
- CHM13v1.1 genome size is 3.055 Gbp, with zero Ns
- Every chromosome is telomere-to-telomere, quality estimated >Q70
- ~200 Mbp (3–6%) of new sequence vs. GRCh38, fixes thousands of errors
- 140 new putative protein-coding genes



# T2T-CHM13 is the *complete* sequence of a human genome

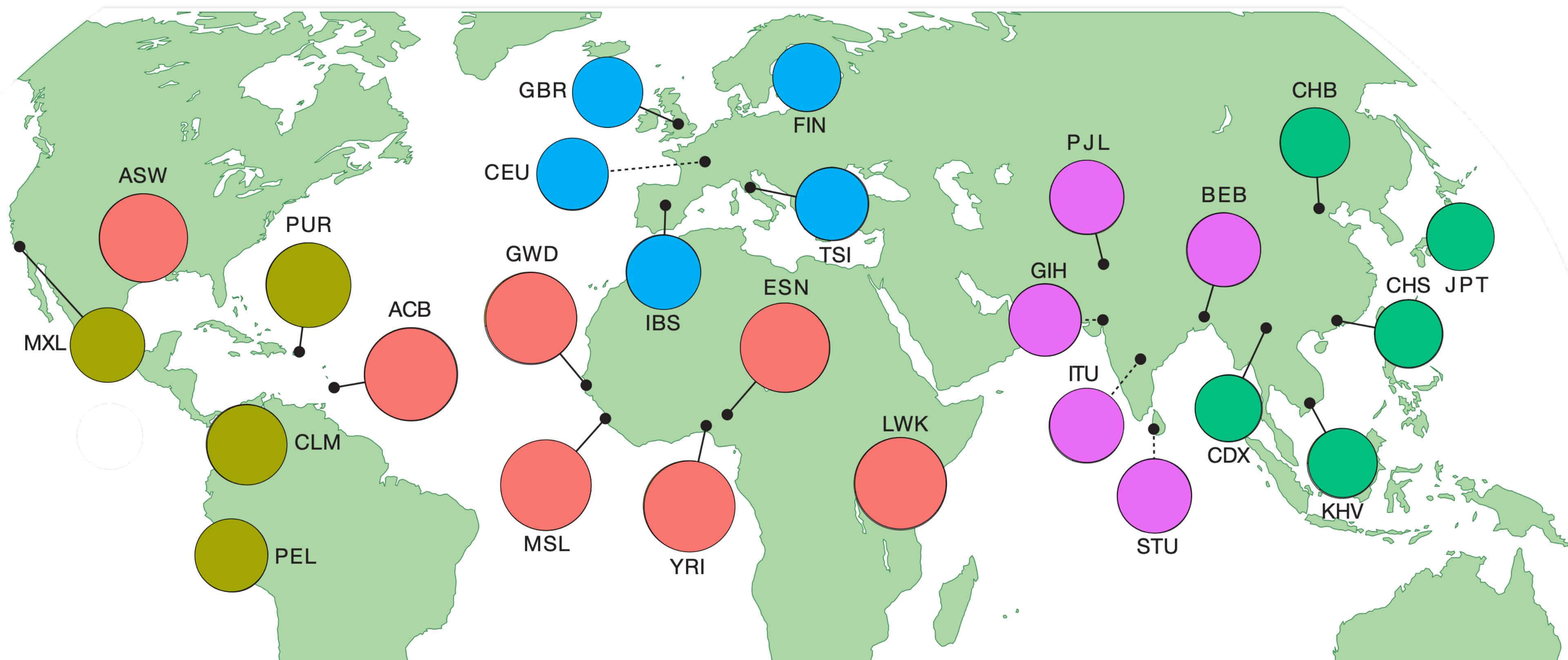
- CHM13v1.1 genome size is 3.055 Gbp, with zero Ns
- Every chromosome is telomere-to-telomere, quality estimated >Q70
- ~200 Mbp (3–6%) of new sequence vs. GRCh38, fixes thousands of errors
- 140 new putative protein-coding genes

**Most accurate assembly ever produced**



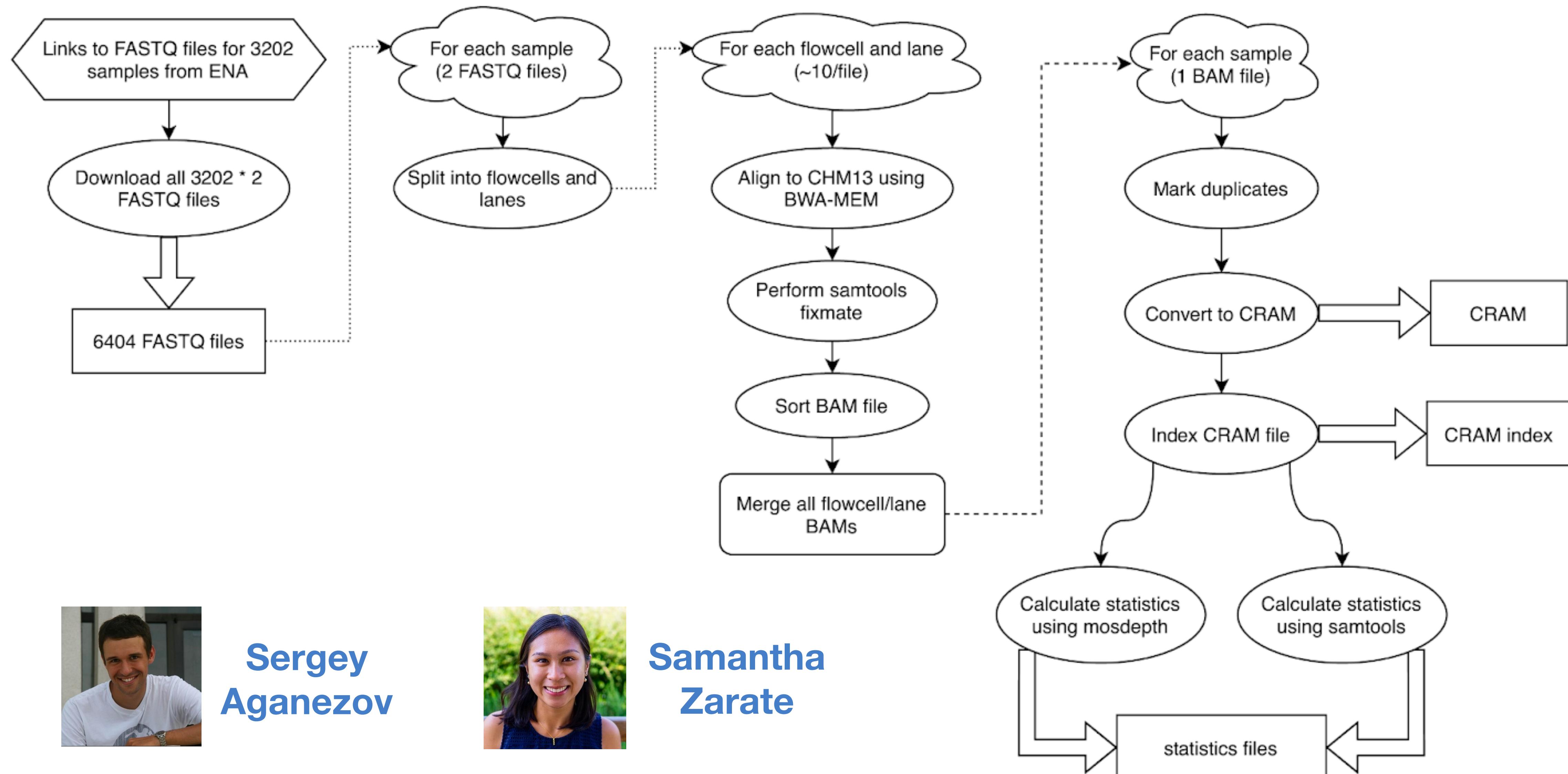
# Analyzing diverse, short-read data with T2T-CHM13

- 1000 Genomes Project (1KGP): 3202 samples from 5 continental groups
- 30x sequencing by the New York Genome Center

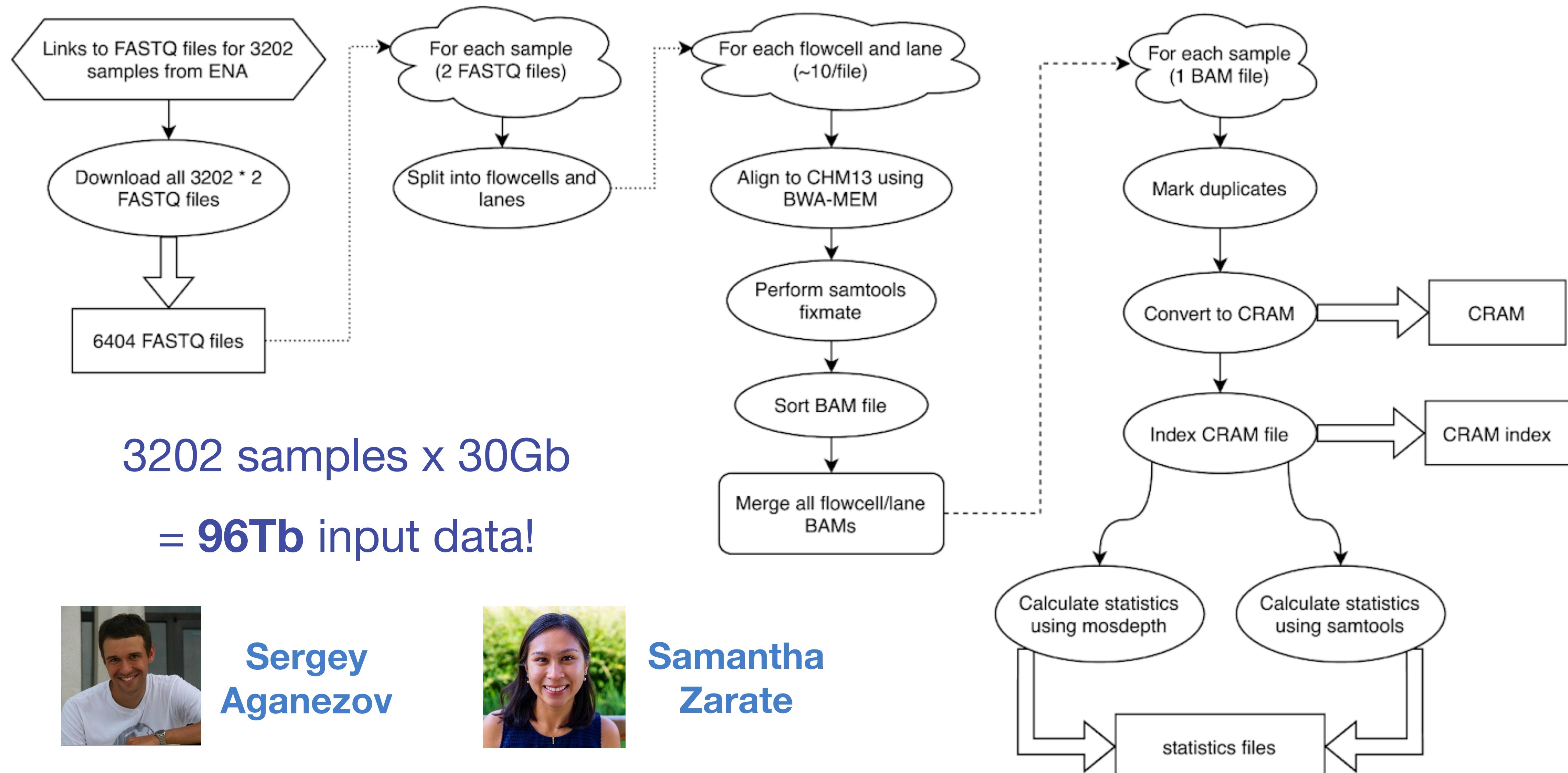


**Byrska-Bishop et al., 2022, Cell.**  
**High coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios.**

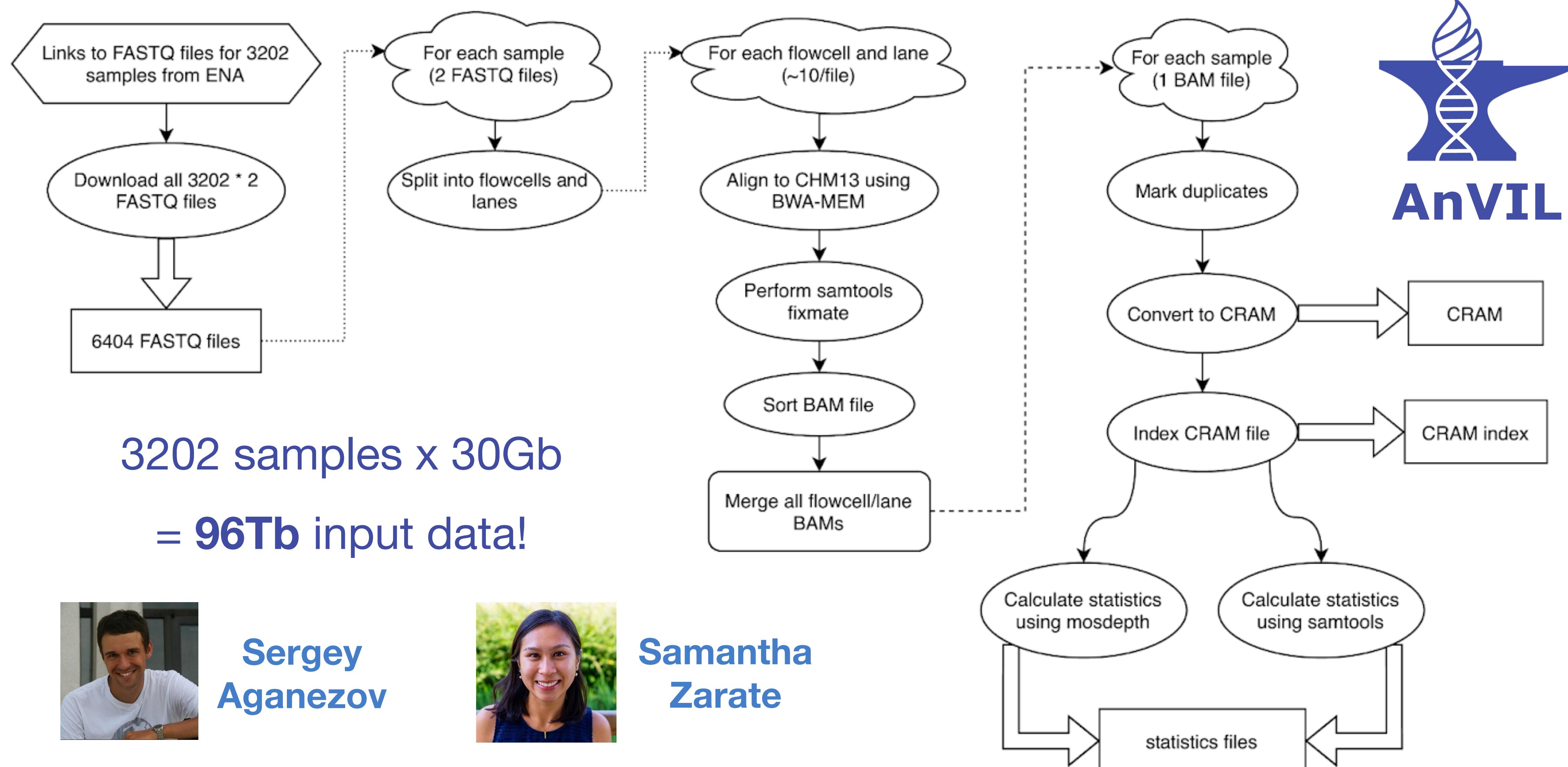
# Population-scale short-read alignment and variant calling

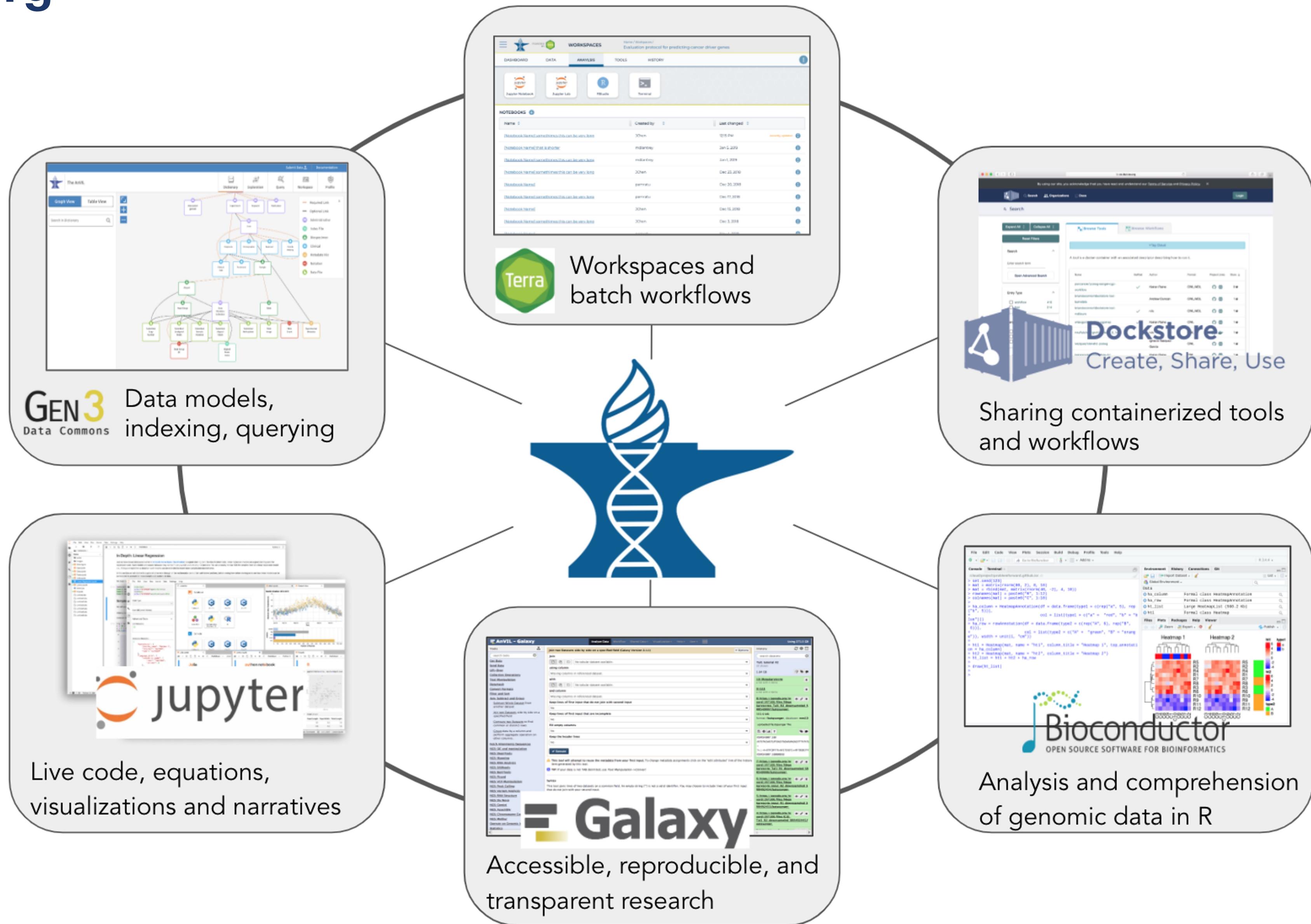


# Population-scale short-read alignment and variant calling



# Population-scale short-read alignment and variant calling





# AnVIL: Data Table

Screenshot of the AnVIL Data Table interface on a web browser.

The browser address bar shows: `anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T/data`.

The page title is "Workspaces > anvil-datastorage/AnVIL\_T2T > Data".

The navigation menu includes: DASHBOARD, DATA (selected), NOTEBOOKS, WORKFLOWS, and JOB HISTORY.

A message in the top right corner states: "Your access to NHGRI AnVIL Data Commons Framework Services has expired. Log in to restore your access or unlink your account." with a close button.

The main content area is titled "TABLES" and shows a table with 3202 rows. The columns are: ena\_sample..., cram, cram\_index, mosdepth\_global\_dist, mosdepth\_regions\_bed, mosdepth\_regions\_bed\_idx, mosdepth\_regions\_dist, mosdepth\_regions\_gz, and a gear icon.

Table Headers:

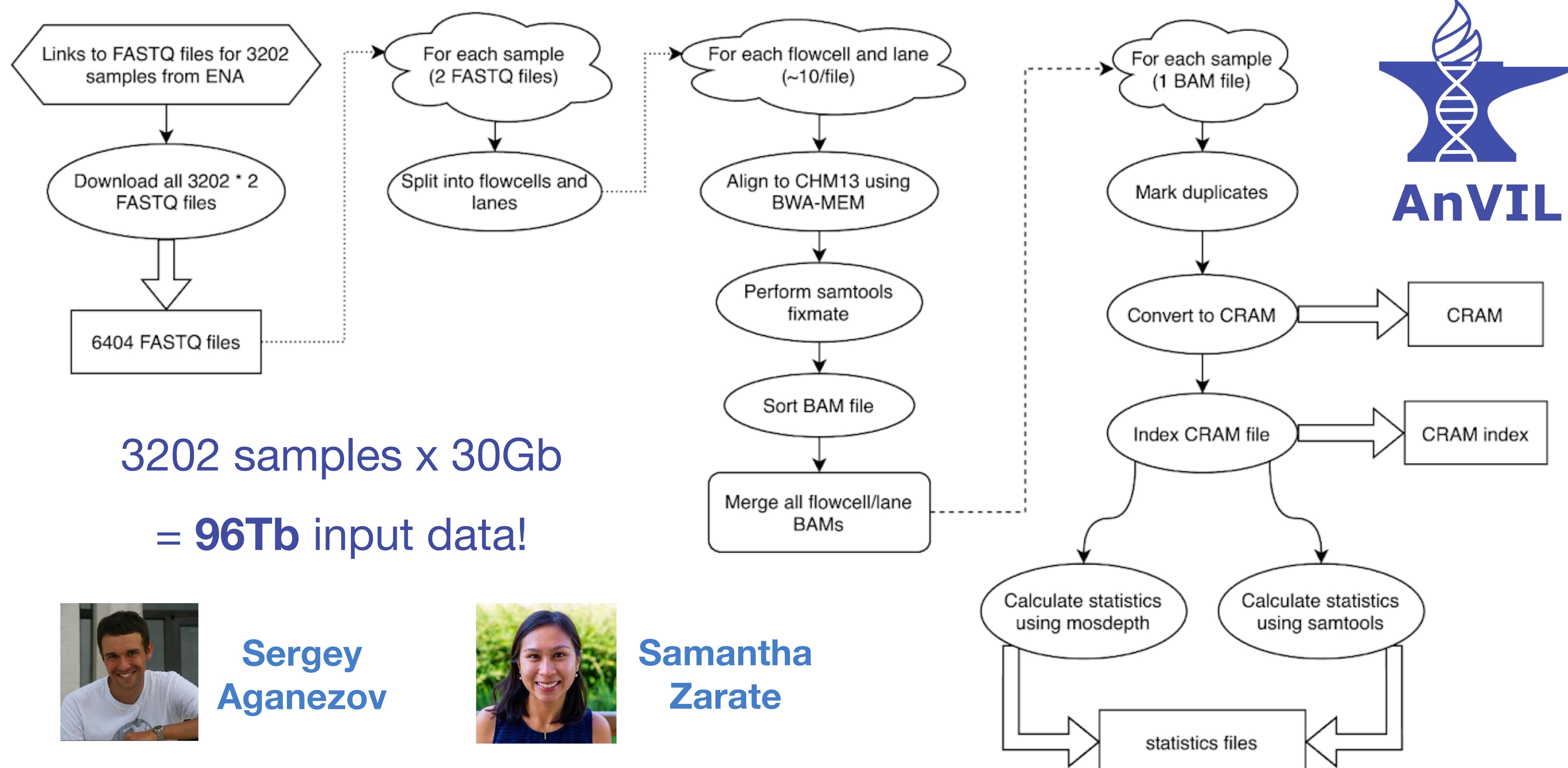
| ena_sample... | cram | cram_index | mosdepth_global_dist | mosdepth_regions_bed | mosdepth_regions_bed_idx | mosdepth_regions_dist | mosdepth_regions_gz |  |
|---------------|------|------------|----------------------|----------------------|--------------------------|-----------------------|---------------------|--|
|---------------|------|------------|----------------------|----------------------|--------------------------|-----------------------|---------------------|--|

Table Rows (Sample):

| HG00096 | [HG00096.cram](#) | [HG00096.cram.crai](#) | [HG00096.mosdepth.global.dist.txt](#) | [HG00096.regions.bed.gz](#) | [HG00096.regions.bed.gz.csi](#) | [HG00096.mosdepth.region.dist.txt](#) | [HG00096.mosdepth.region.gz](#) |  |
| HG00097 | [HG00097.cram](#) | [HG00097.cram.crai](#) | [HG00097.mosdepth.global.dist.txt](#) | [HG00097.regions.bed.gz](#) | [HG00097.regions.bed.gz.csi](#) | [HG00097.mosdepth.region.dist.txt](#) | [HG00097.mosdepth.region.gz](#) |  |
| HG00099 | [HG00099.cram](#) | [HG00099.cram.crai](#) | [HG00099.mosdepth.global.dist.txt](#) | [HG00099.regions.bed.gz](#) | [HG00099.regions.bed.gz.csi](#) | [HG00099.mosdepth.region.dist.txt](#) | [HG00099.mosdepth.region.gz](#) |  |
| HG00100 | [HG00100.cram](#) | [HG00100.cram.crai](#) | [HG00100.mosdepth.global.dist.txt](#) | [HG00100.regions.bed.gz](#) | [HG00100.regions.bed.gz.csi](#) | [HG00100.mosdepth.region.dist.txt](#) | [HG00100.mosdepth.region.gz](#) |  |
| HG00101 | [HG00101.cram](#) | [HG00101.cram.crai](#) | [HG00101.mosdepth.global.dist.txt](#) | [HG00101.regions.bed.gz](#) | [HG00101.regions.bed.gz.csi](#) | [HG00101.mosdepth.region.dist.txt](#) | [HG00101.mosdepth.region.gz](#) |  |
| HG00102 | [HG00102.cram](#) | [HG00102.cram.crai](#) | [HG00102.mosdepth.global.dist.txt](#) | [HG00102.regions.bed.gz](#) | [HG00102.regions.bed.gz.csi](#) | [HG00102.mosdepth.region.dist.txt](#) | [HG00102.mosdepth.region.gz](#) |  |
| HG00103 | [HG00103.cram](#) | [HG00103.cram.crai](#) | [HG00103.mosdepth.global.dist.txt](#) | [HG00103.regions.bed.gz](#) | [HG00103.regions.bed.gz.csi](#) | [HG00103.mosdepth.region.dist.txt](#) | [HG00103.mosdepth.region.gz](#) |  |
| HG00105 | [HG00105.cram](#) | [HG00105.cram.crai](#) | [HG00105.mosdepth.global.dist.txt](#) | [HG00105.regions.bed.gz](#) | [HG00105.regions.bed.gz.csi](#) | [HG00105.mosdepth.region.dist.txt](#) | [HG00105.mosdepth.region.gz](#) |  |
| HG00106 | [HG00106.cram](#) | [HG00106.cram.crai](#) | [HG00106.mosdepth.global.dist.txt](#) | [HG00106.regions.bed.gz](#) | [HG00106.regions.bed.gz.csi](#) | [HG00106.mosdepth.region.dist.txt](#) | [HG00106.mosdepth.region.gz](#) |  |
| HG00107 | [HG00107.cram](#) | [HG00107.cram.crai](#) | [HG00107.mosdepth.global.dist.txt](#) | [HG00107.regions.bed.gz](#) | [HG00107.regions.bed.gz.csi](#) | [HG00107.mosdepth.region.dist.txt](#) | [HG00107.mosdepth.region.gz](#) |  |
| HG00108 | [HG00108.cram](#) | [HG00108.cram.crai](#) | [HG00108.mosdepth.global.dist.txt](#) | [HG00108.regions.bed.gz](#) | [HG00108.regions.bed.gz.csi](#) | [HG00108.mosdepth.region.dist.txt](#) | [HG00108.mosdepth.region.gz](#) |  |
| HG00109 | [HG00109.cram](#) | [HG00109.cram.crai](#) | [HG00109.mosdepth.global.dist.txt](#) | [HG00109.regions.bed.gz](#) | [HG00109.regions.bed.gz.csi](#) | [HG00109.mosdepth.region.dist.txt](#) | [HG00109.mosdepth.region.gz](#) |  |
| HG00110 | [HG00110.cram](#) | [HG00110.cram.crai](#) | [HG00110.mosdepth.global.dist.txt](#) | [HG00110.regions.bed.gz](#) | [HG00110.regions.bed.gz.csi](#) | [HG00110.mosdepth.region.dist.txt](#) | [HG00110.mosdepth.region.gz](#) |  |
| HG00111 | [HG00111.cram](#) | [HG00111.cram.crai](#) | [HG00111.mosdepth.global.dist.txt](#) | [HG00111.regions.bed.gz](#) | [HG00111.regions.bed.gz.csi](#) | [HG00111.mosdepth.region.dist.txt](#) | [HG00111.mosdepth.region.gz](#) |  |
| HG00112 | [HG00112.cram](#) | [HG00112.cram.crai](#) | [HG00112.mosdepth.global.dist.txt](#) | [HG00112.regions.bed.gz](#) | [HG00112.regions.bed.gz.csi](#) | [HG00112.mosdepth.region.dist.txt](#) | [HG00112.mosdepth.region.gz](#) |  |
| HG00113 | [HG00113.cram](#) | [HG00113.cram.crai](#) | [HG00113.mosdepth.global.dist.txt](#) | [HG00113.regions.bed.gz](#) | [HG00113.regions.bed.gz.csi](#) | [HG00113.mosdepth.region.dist.txt](#) | [HG00113.mosdepth.region.gz](#) |  |
| HG00114 | [HG00114.cram](#) | [HG00114.cram.crai](#) | [HG00114.mosdepth.global.dist.txt](#) | [HG00114.regions.bed.gz](#) | [HG00114.regions.bed.gz.csi](#) | [HG00114.mosdepth.region.dist.txt](#) | [HG00114.mosdepth.region.gz](#) |  |
| HG00115 | [HG00115.cram](#) | [HG00115.cram.crai](#) | [HG00115.mosdepth.global.dist.txt](#) | [HG00115.regions.bed.gz](#) | [HG00115.regions.bed.gz.csi](#) | [HG00115.mosdepth.region.dist.txt](#) | [HG00115.mosdepth.region.gz](#) |  |
| HG00116 | [HG00116.cram](#) | [HG00116.cram.crai](#) | [HG00116.mosdepth.global.dist.txt](#) | [HG00116.regions.bed.gz](#) | [HG00116.regions.bed.gz.csi](#) | [HG00116.mosdepth.region.dist.txt](#) | [HG00116.mosdepth.region.gz](#) |  |
| HG00117 | [HG00117.cram](#) | [HG00117.cram.crai](#) | [HG00117.mosdepth.global.dist.txt](#) | [HG00117.regions.bed.gz](#) | [HG00117.regions.bed.gz.csi](#) | [HG00117.mosdepth.region.dist.txt](#) | [HG00117.mosdepth.region.gz](#) |  |
| HG00118 | [HG00118.cram](#) | [HG00118.cram.crai](#) | [HG00118.mosdepth.global.dist.txt](#) | [HG00118.regions.bed.gz](#) | [HG00118.regions.bed.gz.csi](#) | [HG00118.mosdepth.region.dist.txt](#) | [HG00118.mosdepth.region.gz](#) |  |

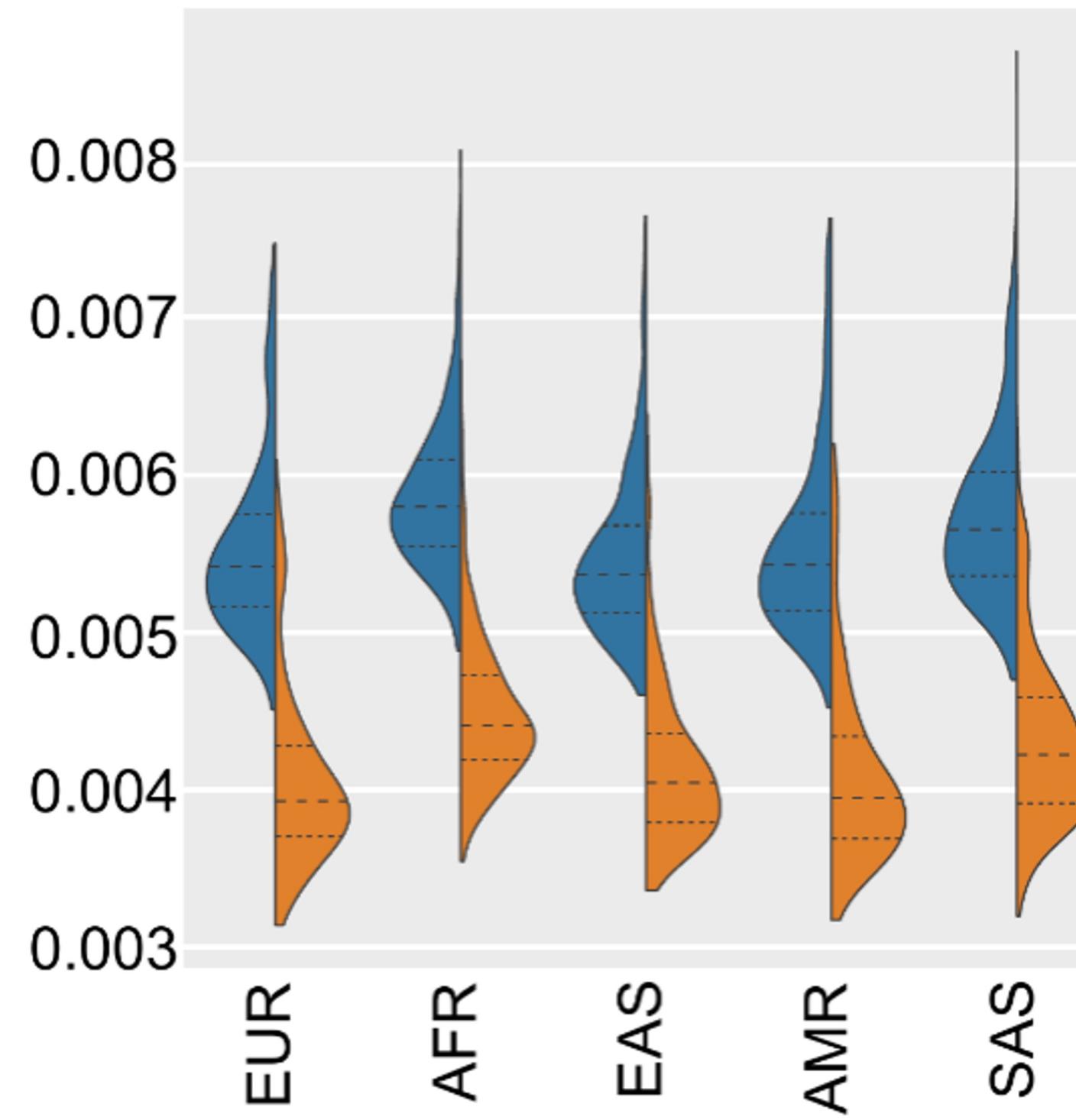
Pagination: 1 - 25 of 3202 | [«](#) [«](#) [1](#) [2](#) [3](#) [4](#) [5](#) [»](#) [»](#) Items per page: [25](#)

# Population-scale short-read alignment and variant calling

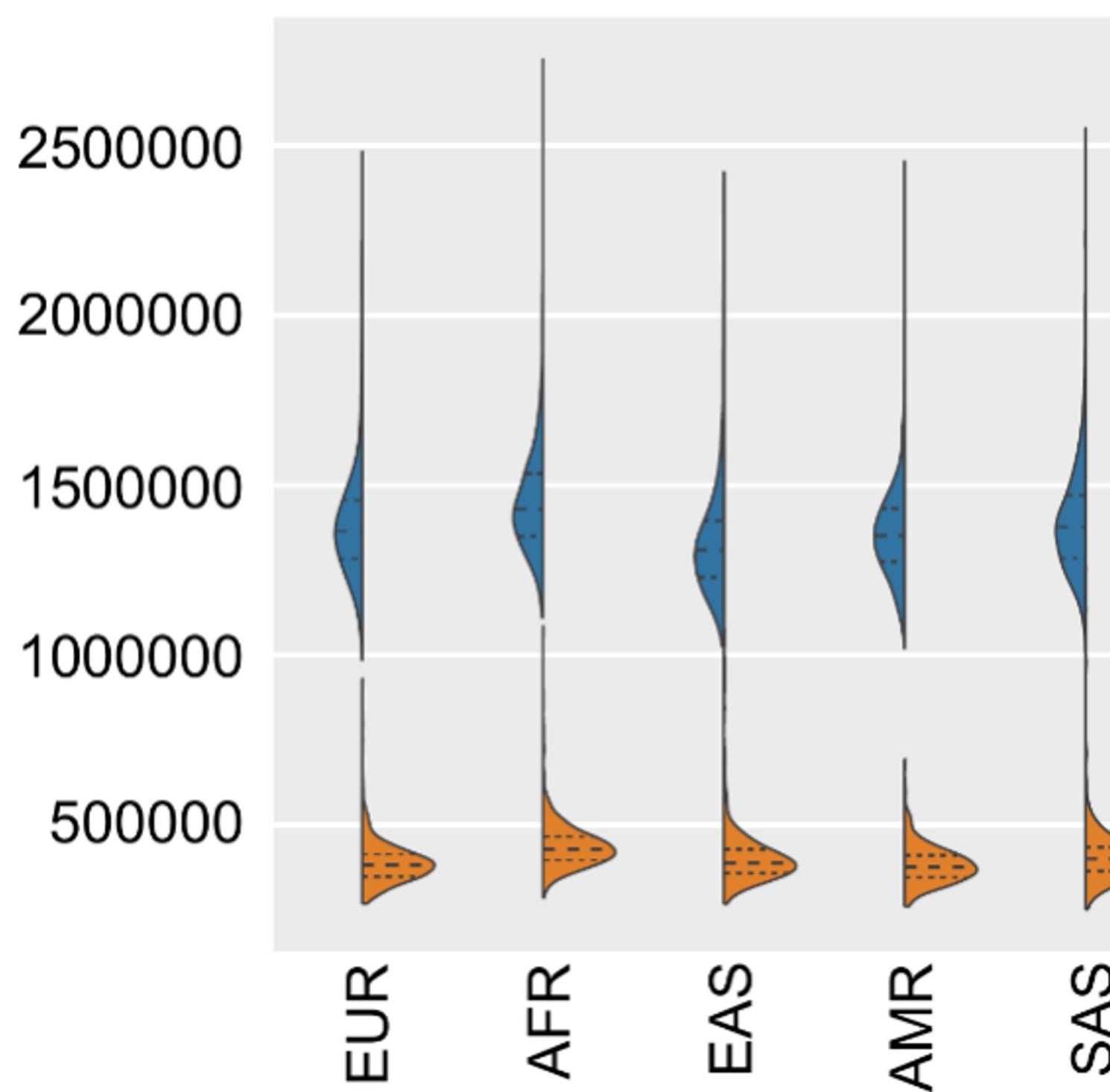


# T2T-CHM13 improves short-read alignment

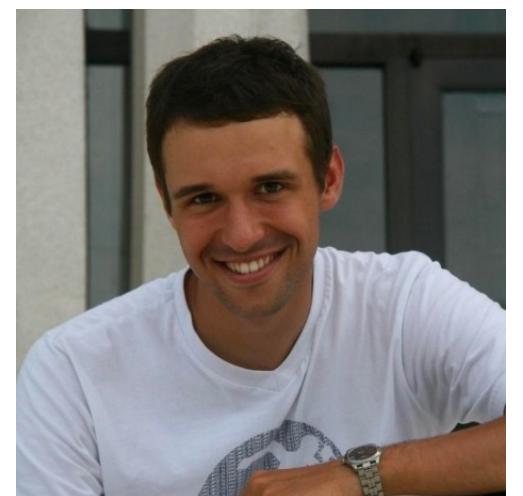
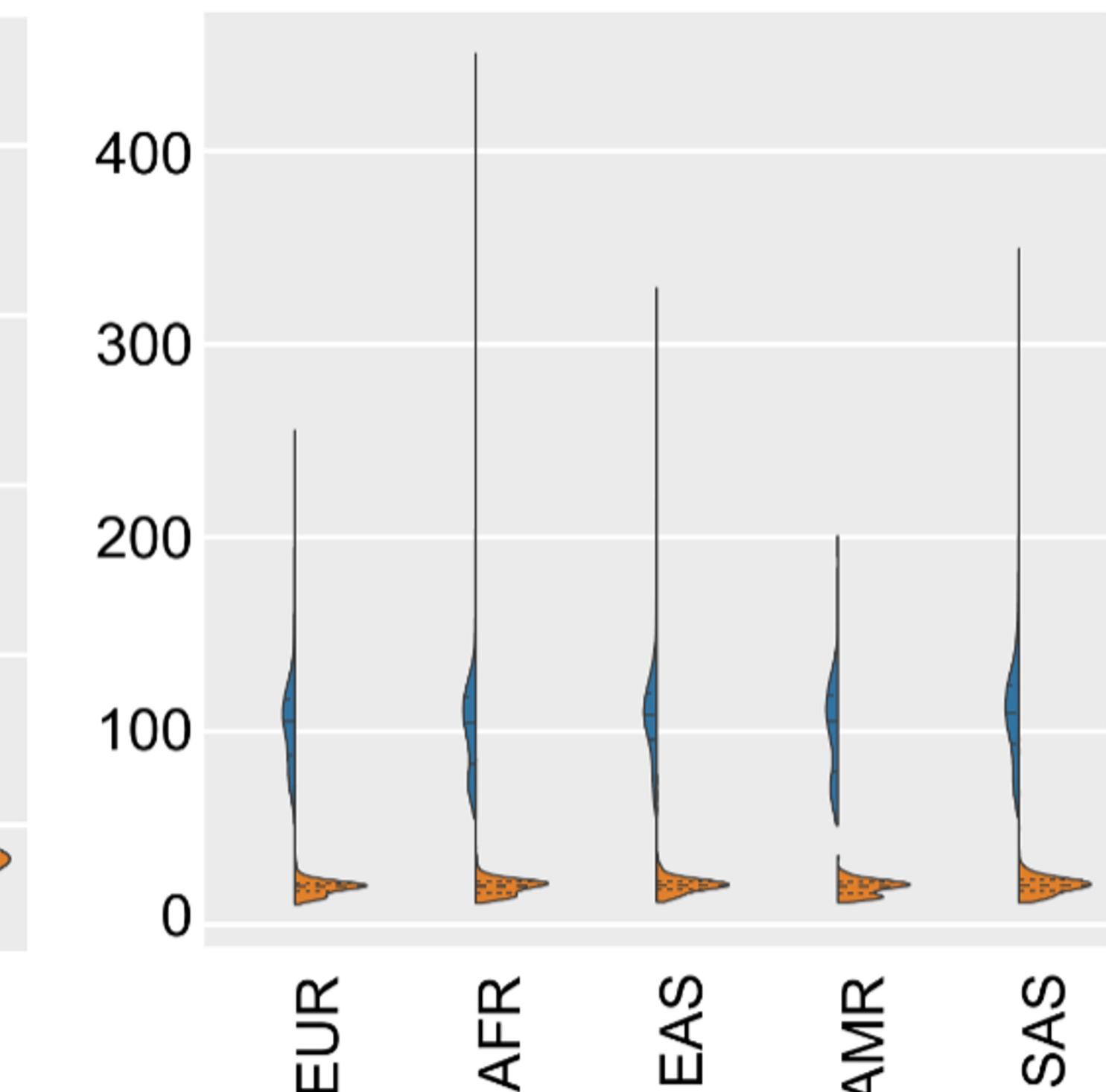
Mismatch rate



Invalid pairs



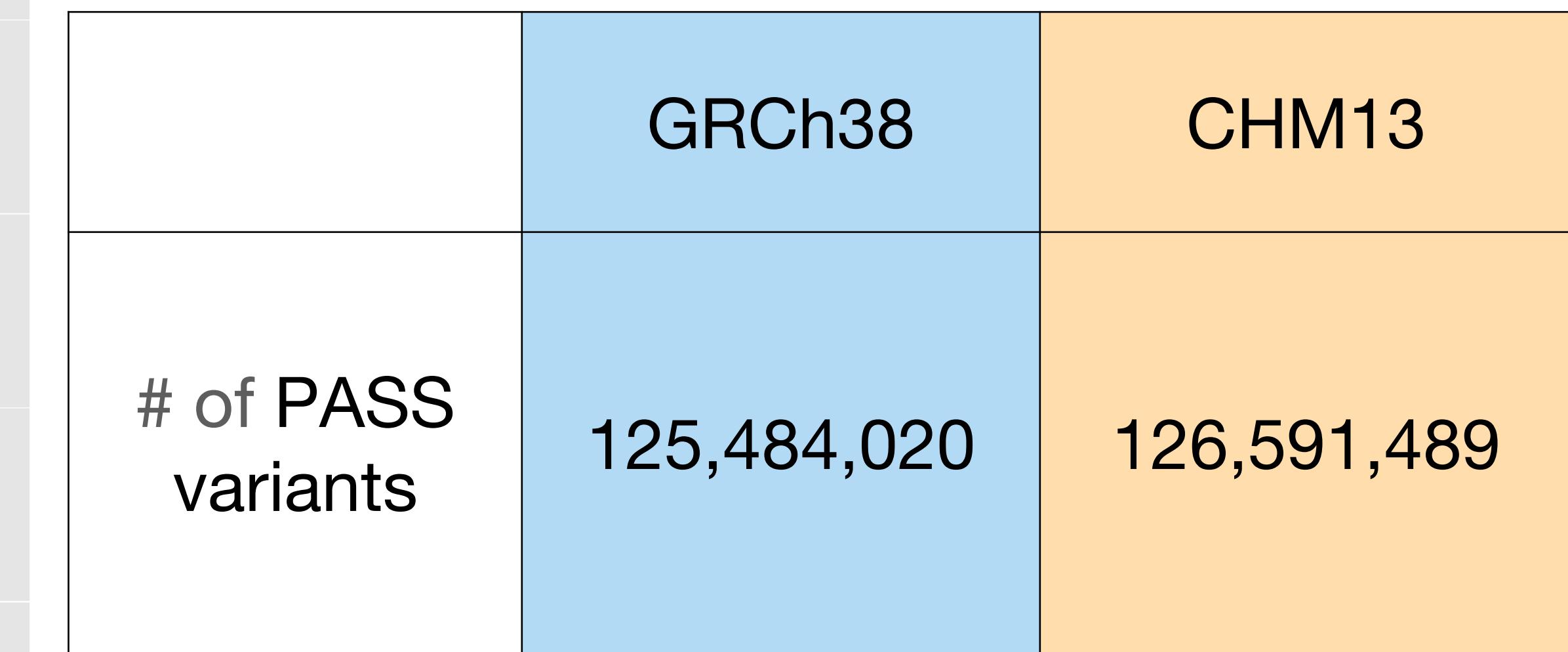
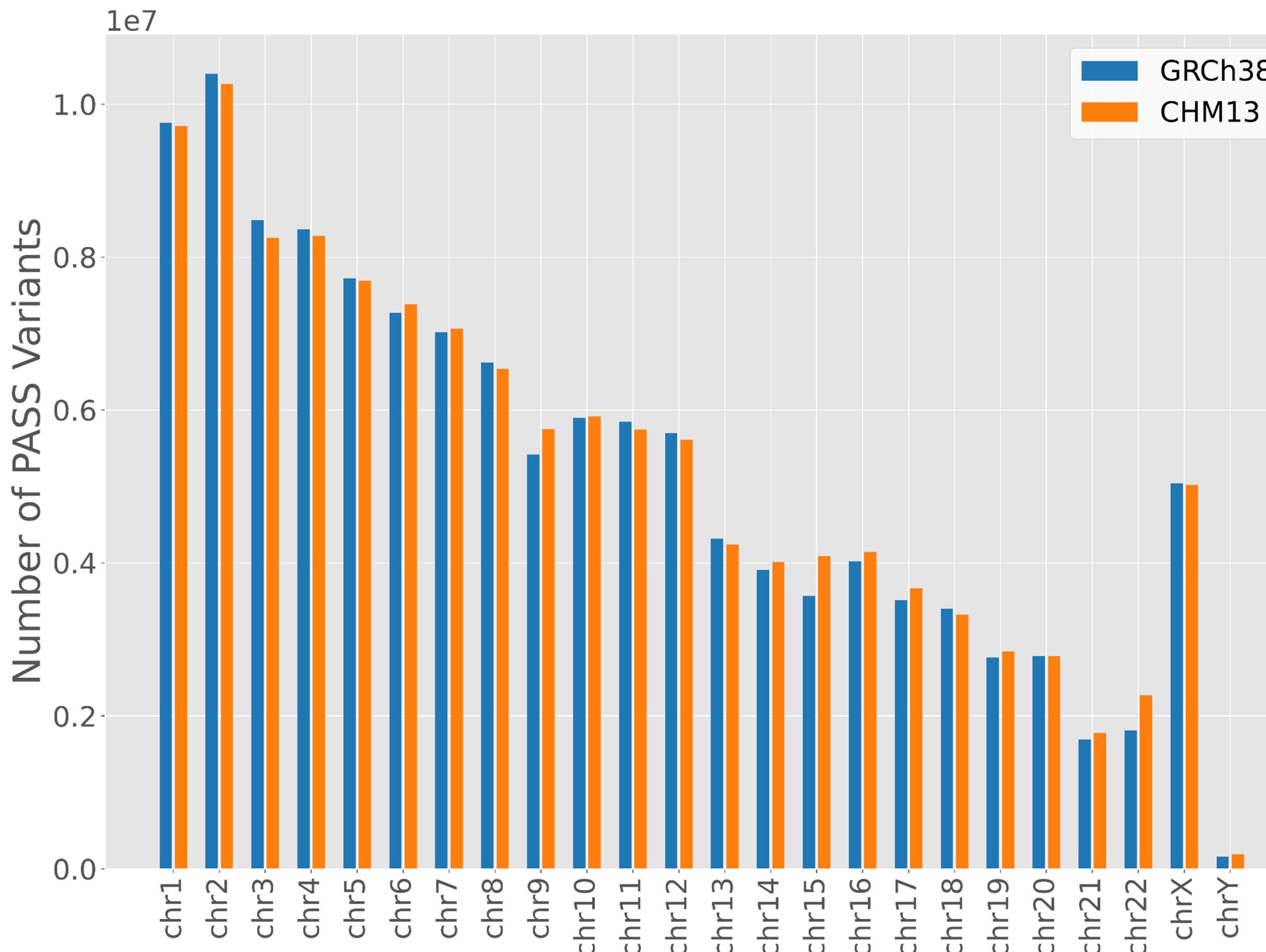
Cov. std. within genes



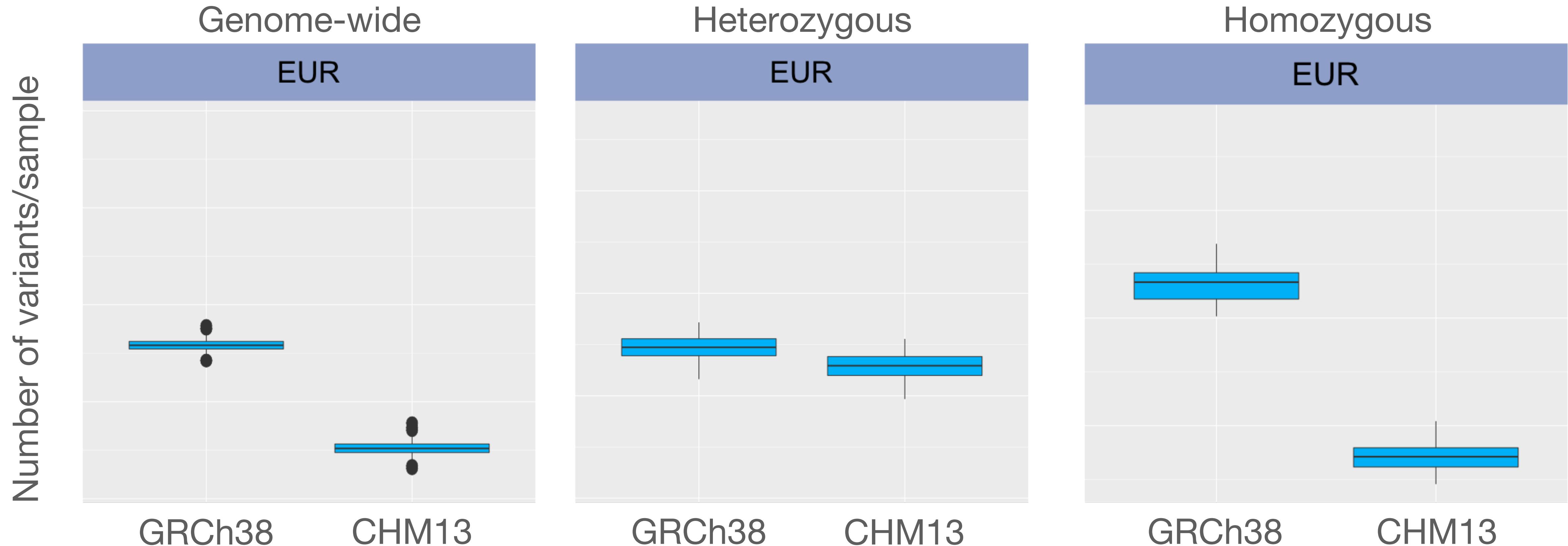
Sergey  
Aganezov

GRCh38      CHM13

# T2T-CHM13 allows discovery of more variants

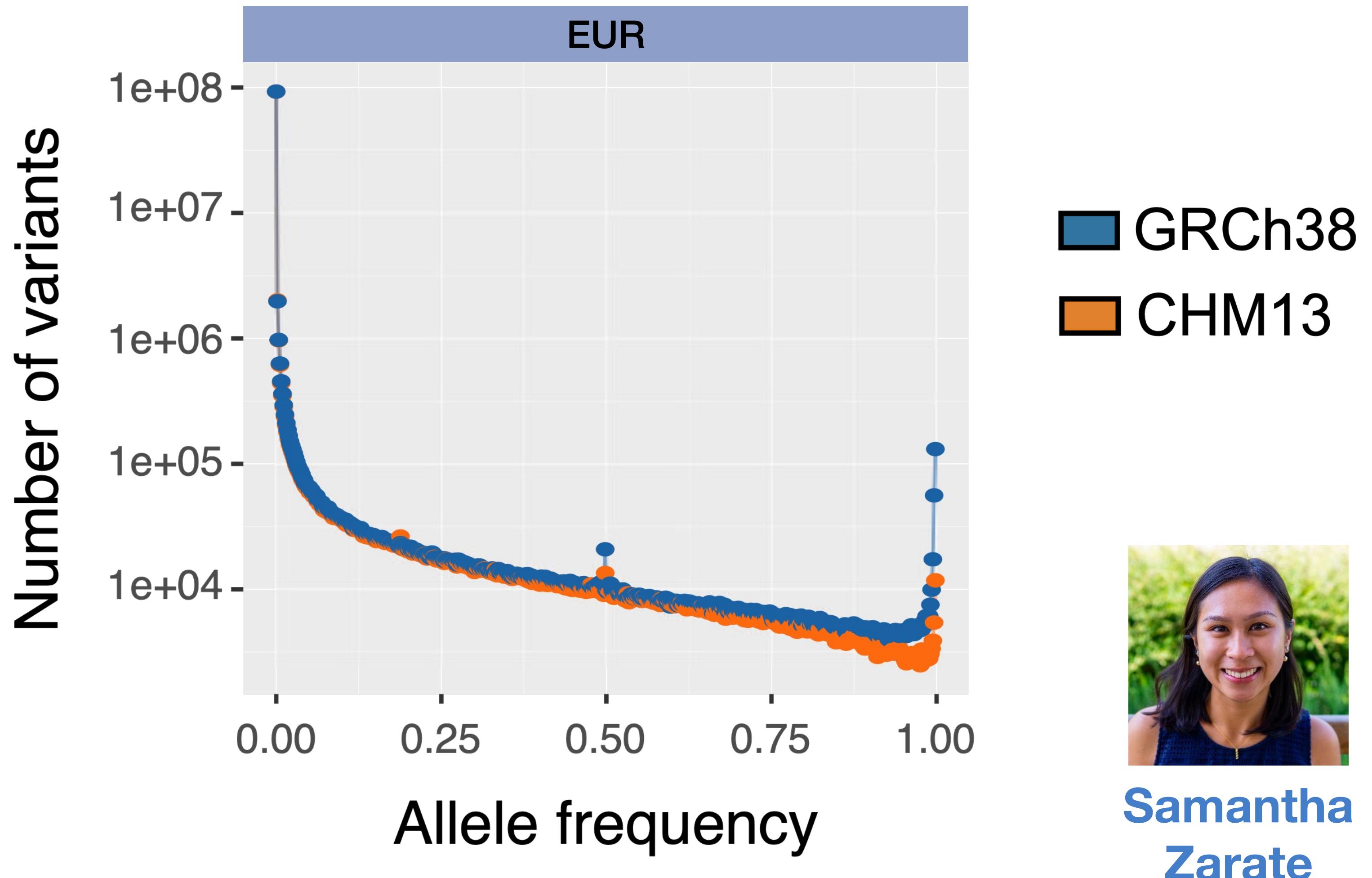


# T2T-CHM13 reduces per-sample variant counts



# T2T-CHM13 improves short-read variant calling

- Excess of fixed variants when using GRCh38 is due to reference errors



# T2T-CHM13 improves short-read variant calling

- Excess of fixed variants when using GRCh38 is due to reference errors

GRCh38    ...ATGG**CA**ATATG...

Samples    ...ATGG**TA**ATATG...

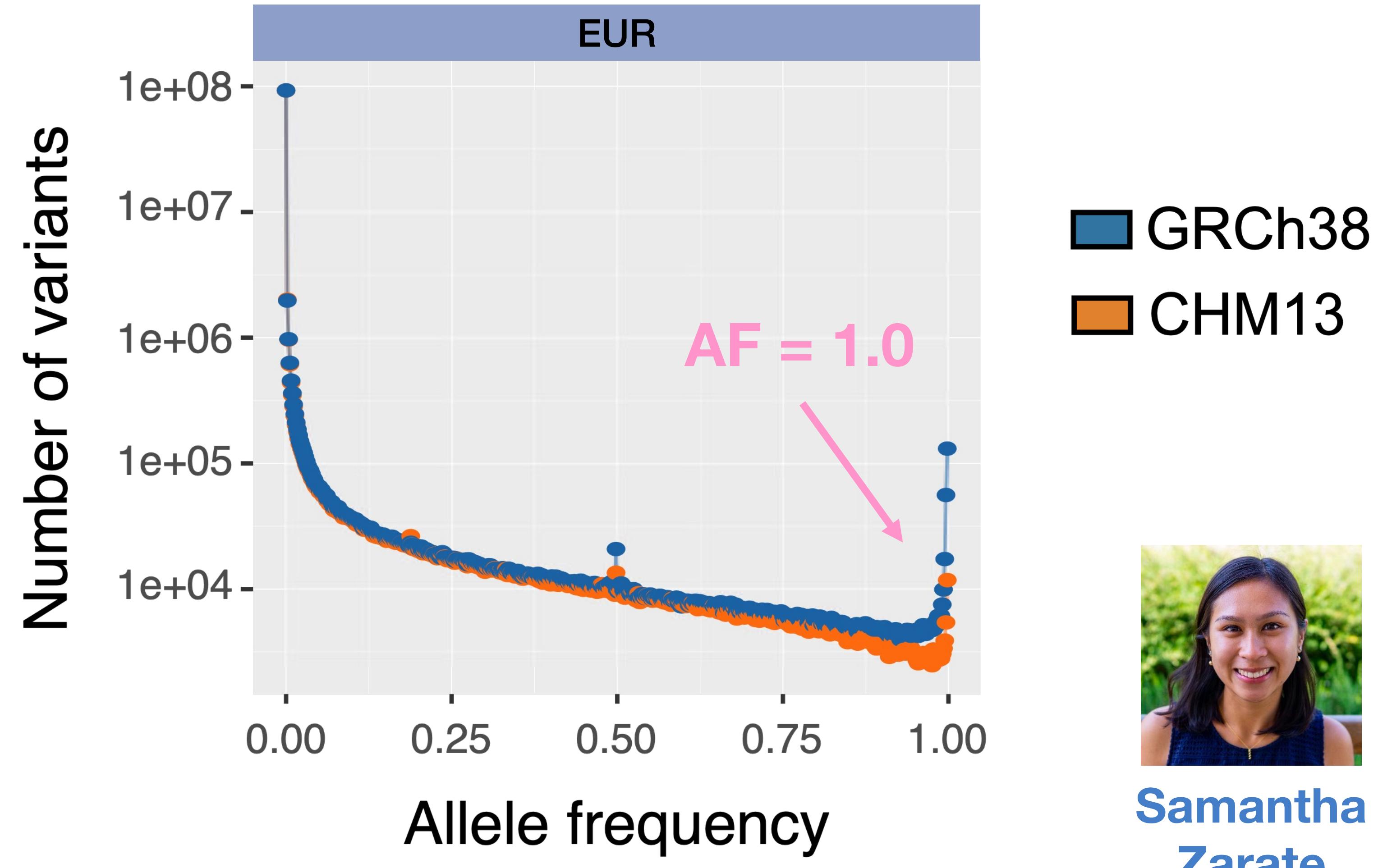
              ...ATGG**TA**ATATG...

              ...ATGG**TA**ATATG...

              ...ATGG**TA**ATATG...

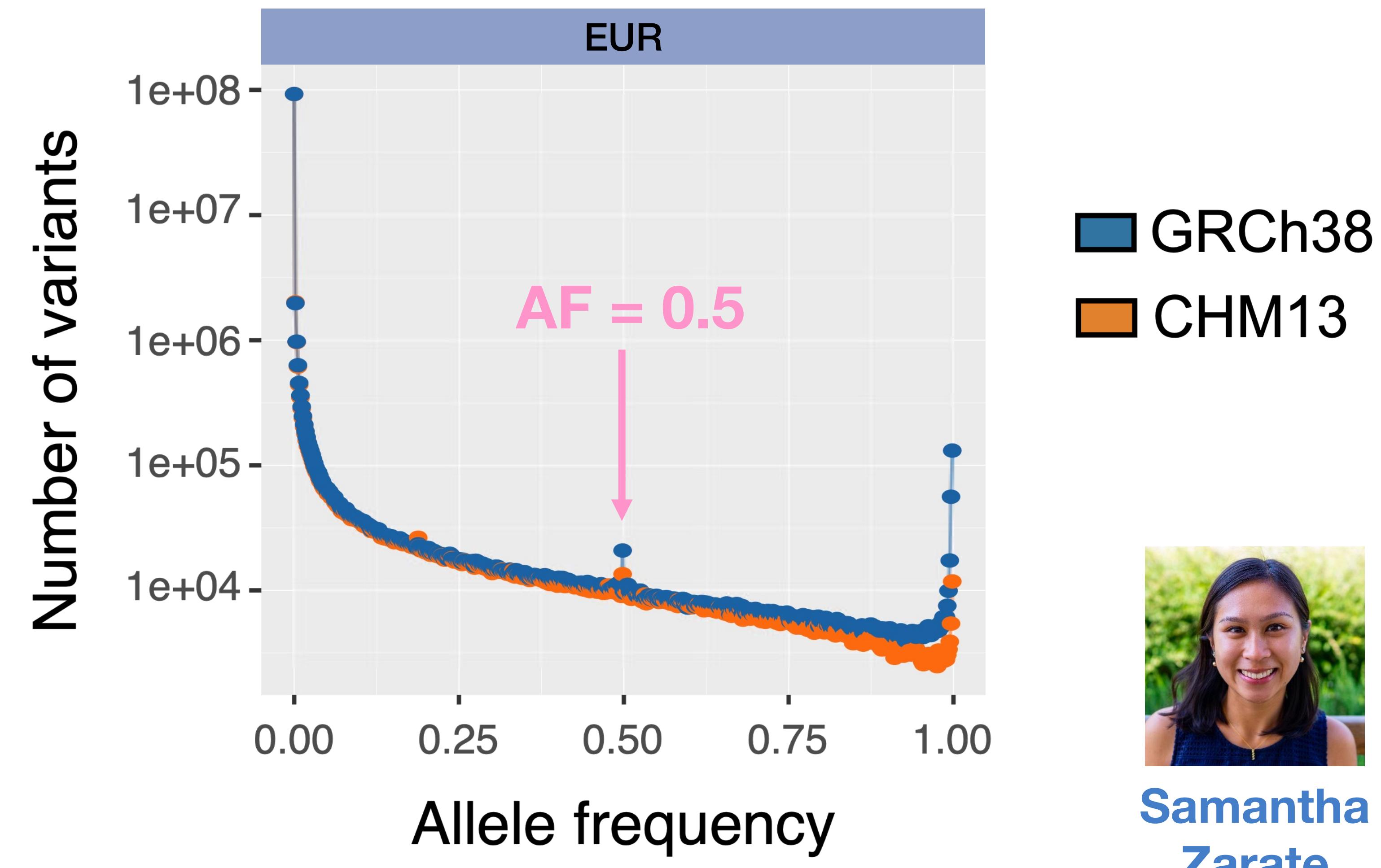
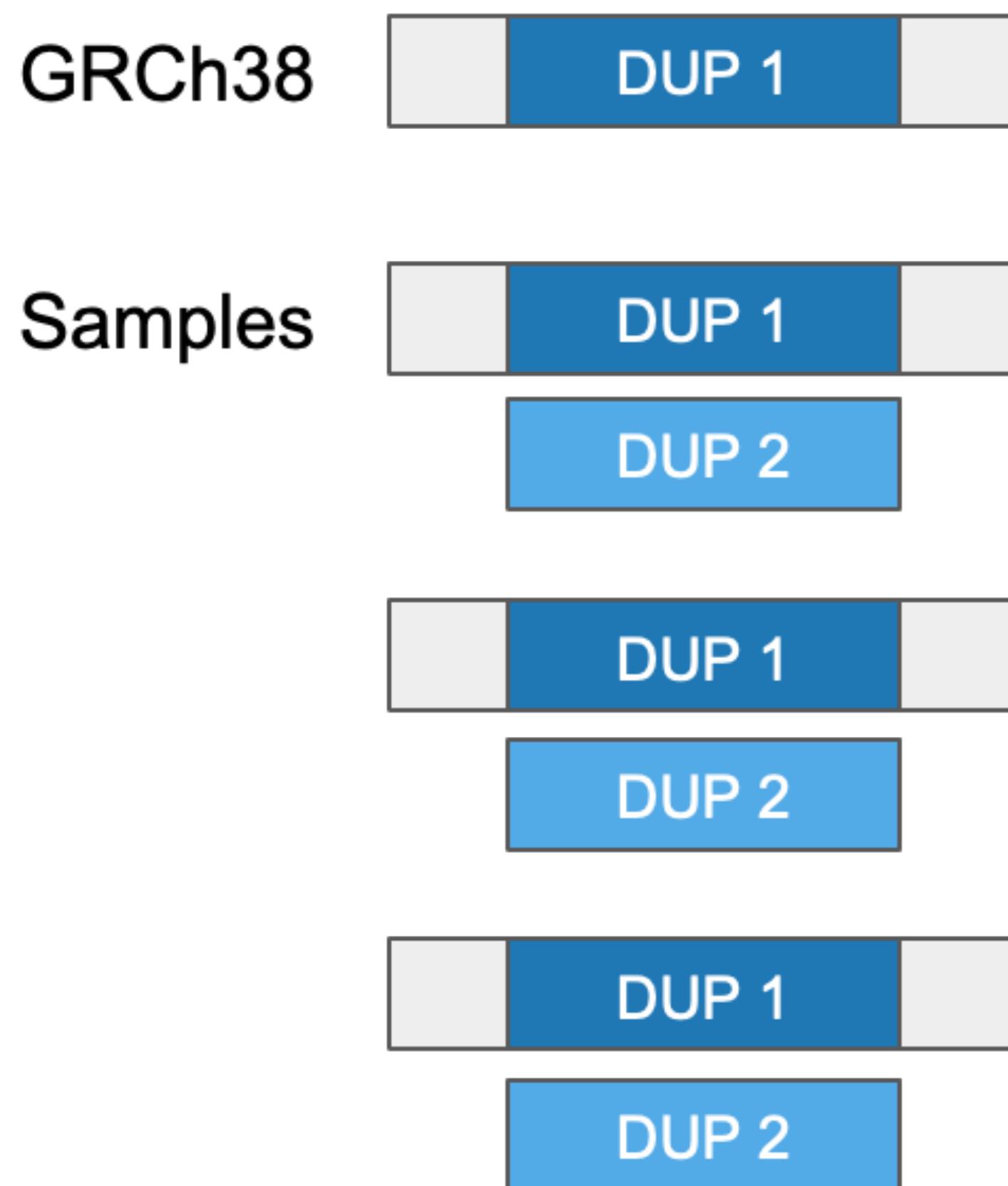
              ...ATGG**TA**ATATG...

              ...ATGG**TA**ATATG...



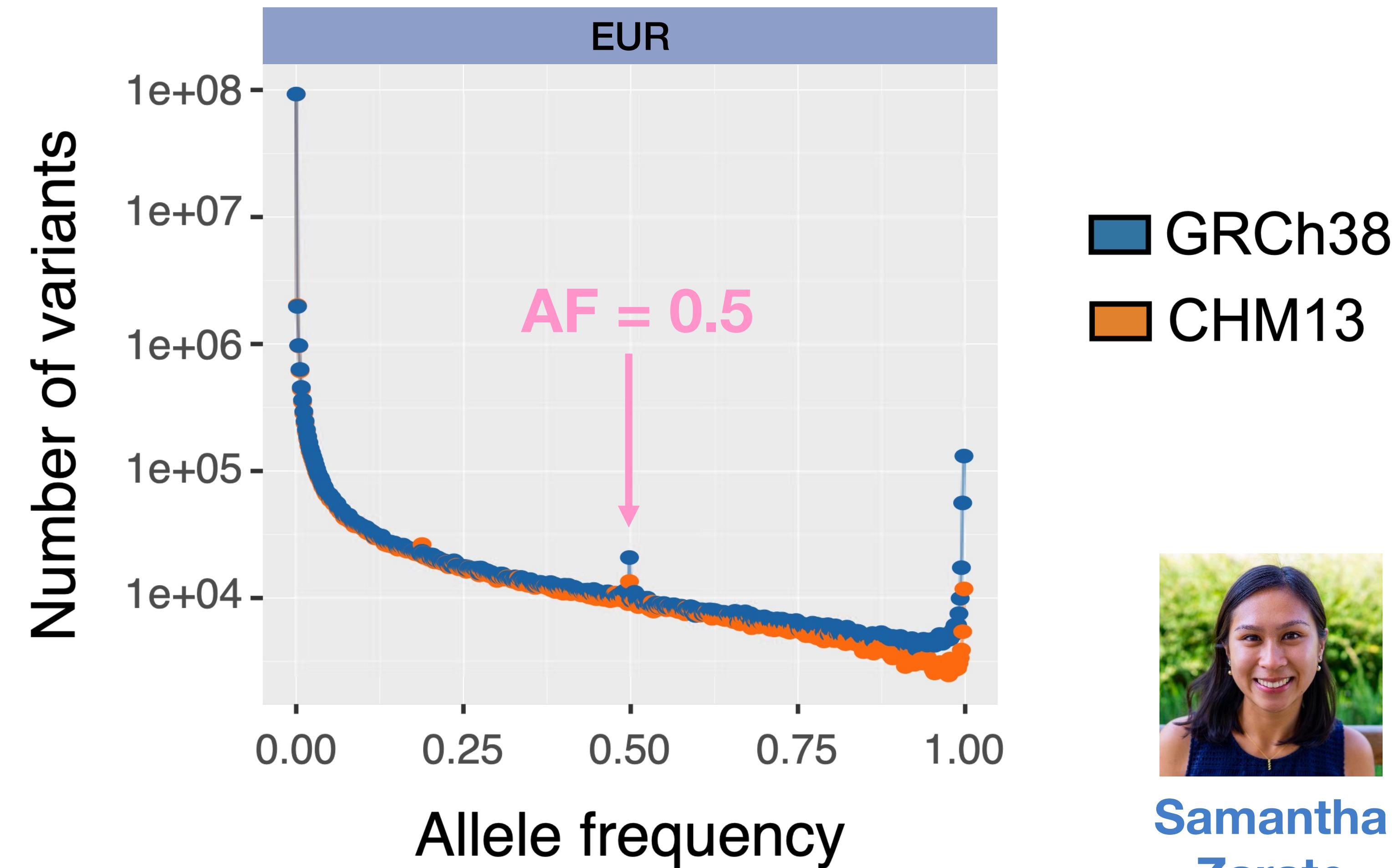
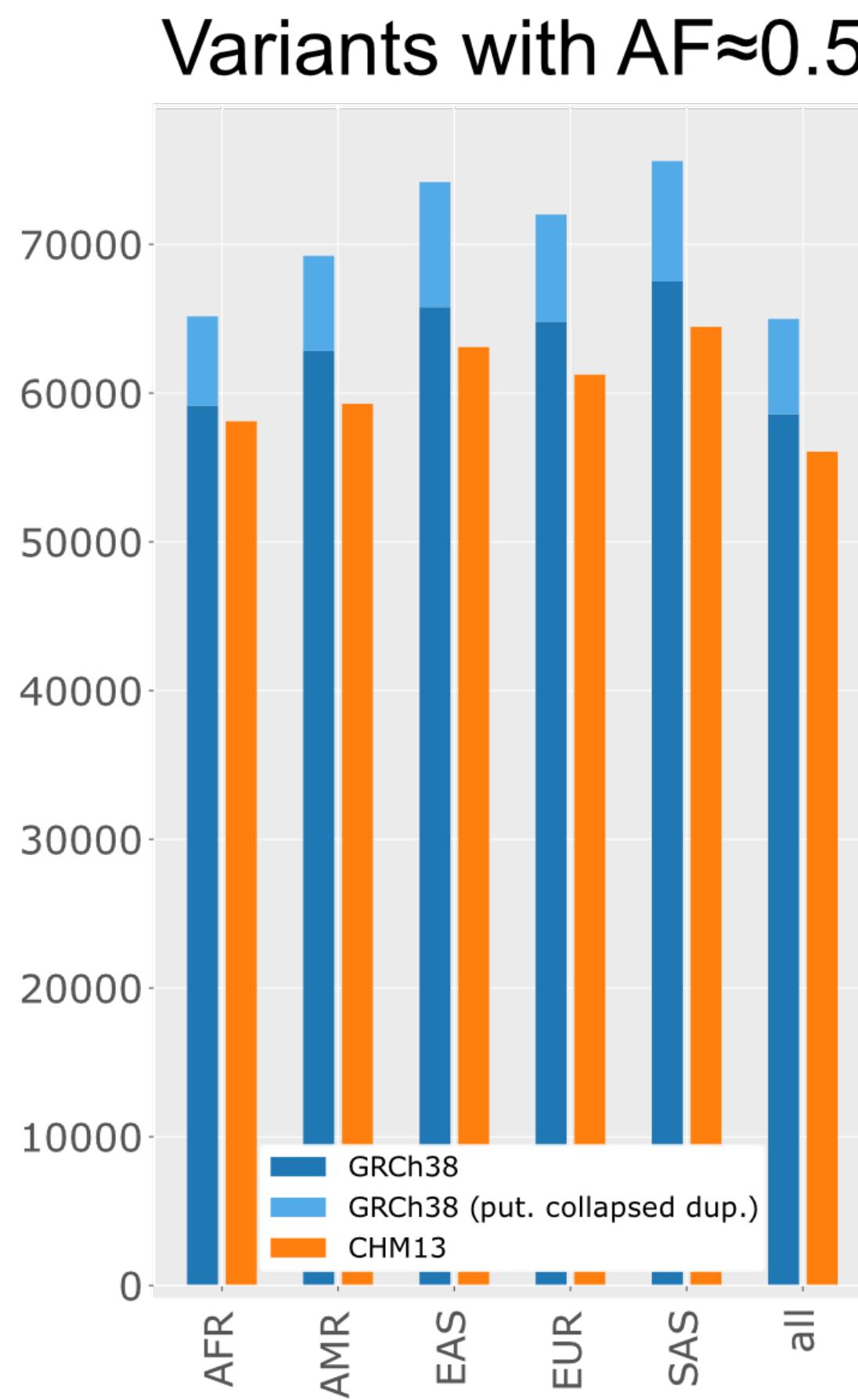
# T2T-CHM13 improves short-read variant calling

- Fewer false heterozygous variants from GRCh38 collapsed duplications



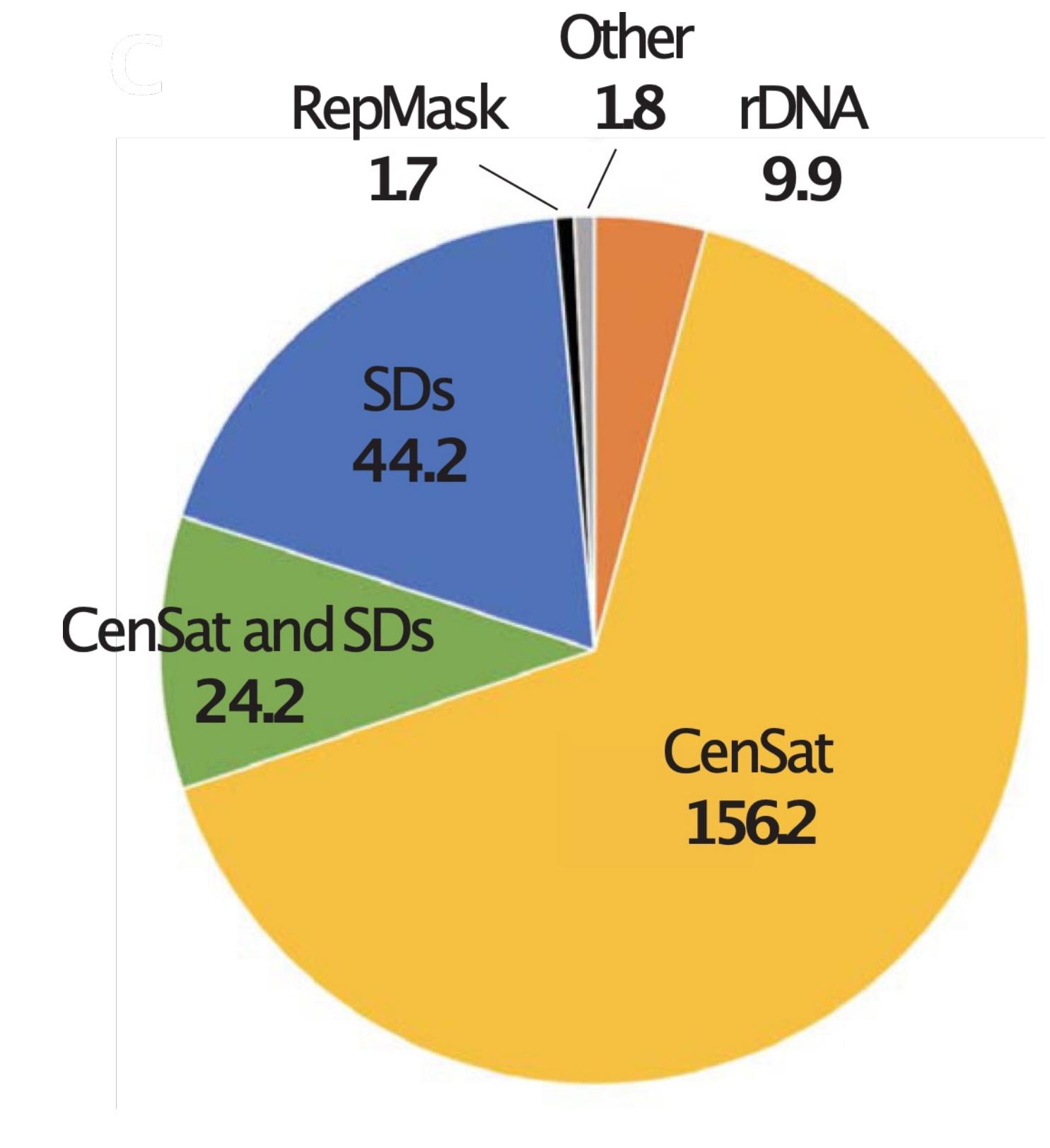
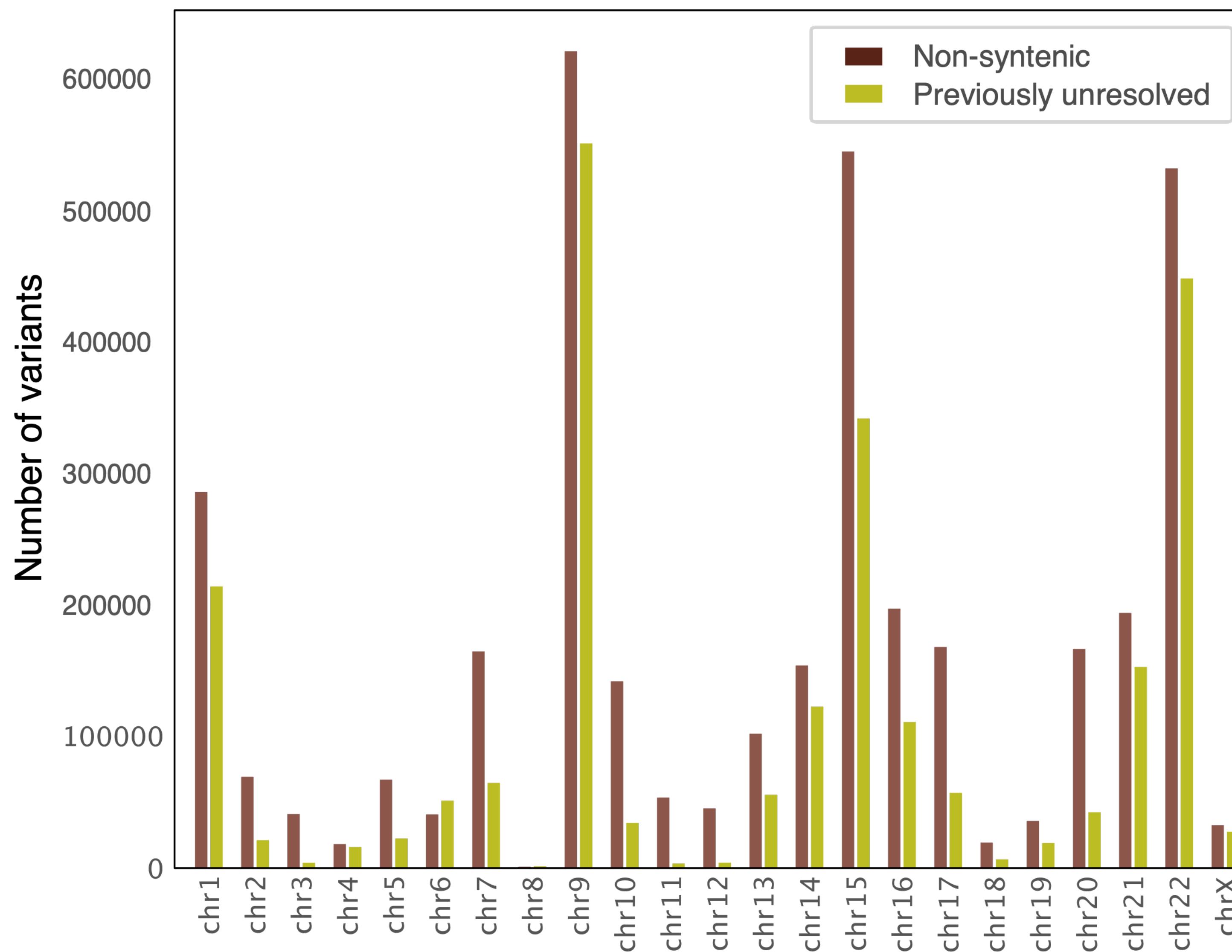
# T2T-CHM13 improves short-read variant calling

- Fewer false heterozygous variants from GRCh38 collapsed duplications



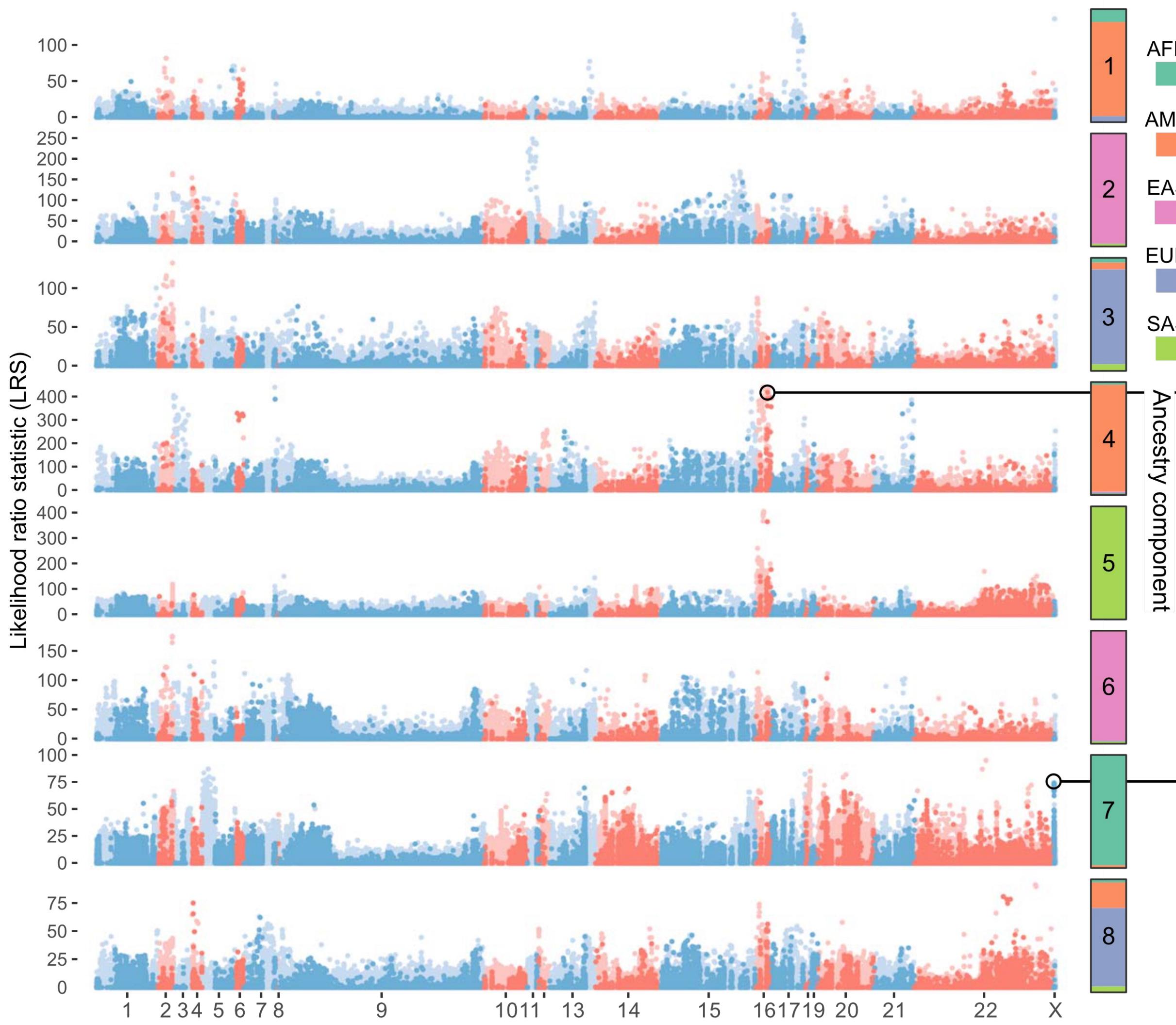
Samantha  
Zarate

# “Novel” variants revealed by T2T-CHM13

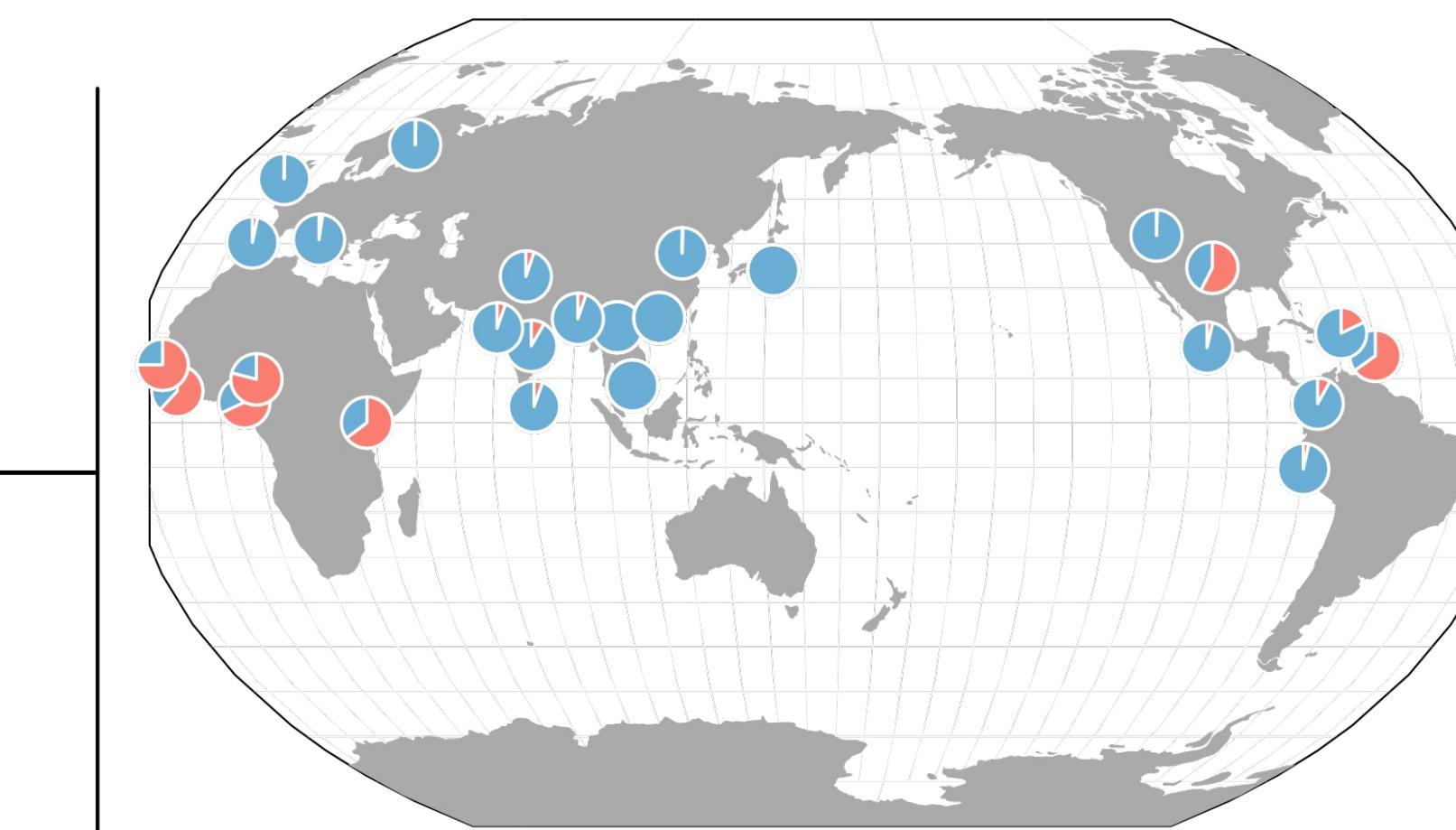


**Samantha  
Zarate**

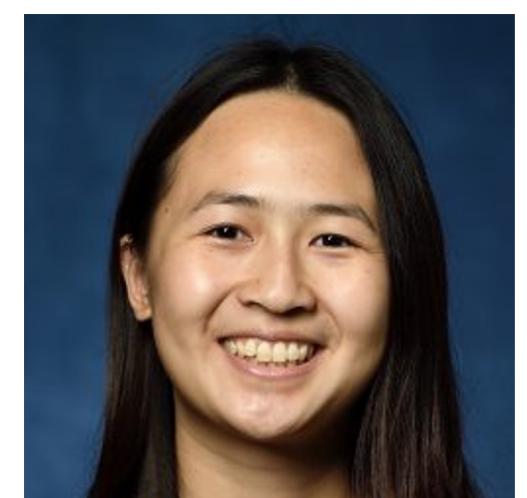
# Extreme AF differentiation of novel variants



chr16:37828623 A / T



chrX:36684515 A / AT



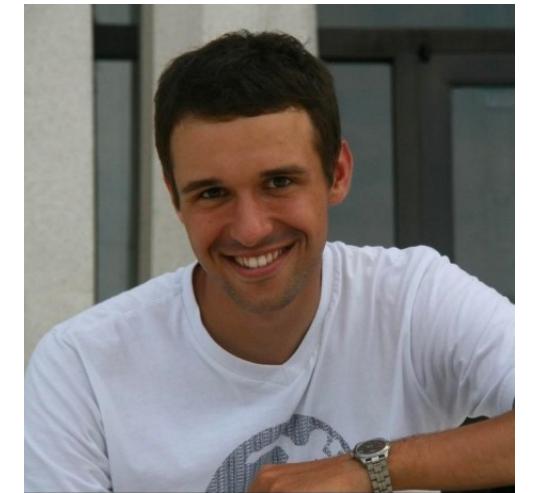
Stephanie  
Yan

# Long-read alignment and variant calling with T2T-CHM13

- 17 diverse samples from the Human Pangenome Reference Consortium and Genome in a Bottle
- PacBio HiFi data + 14 samples with ONT data

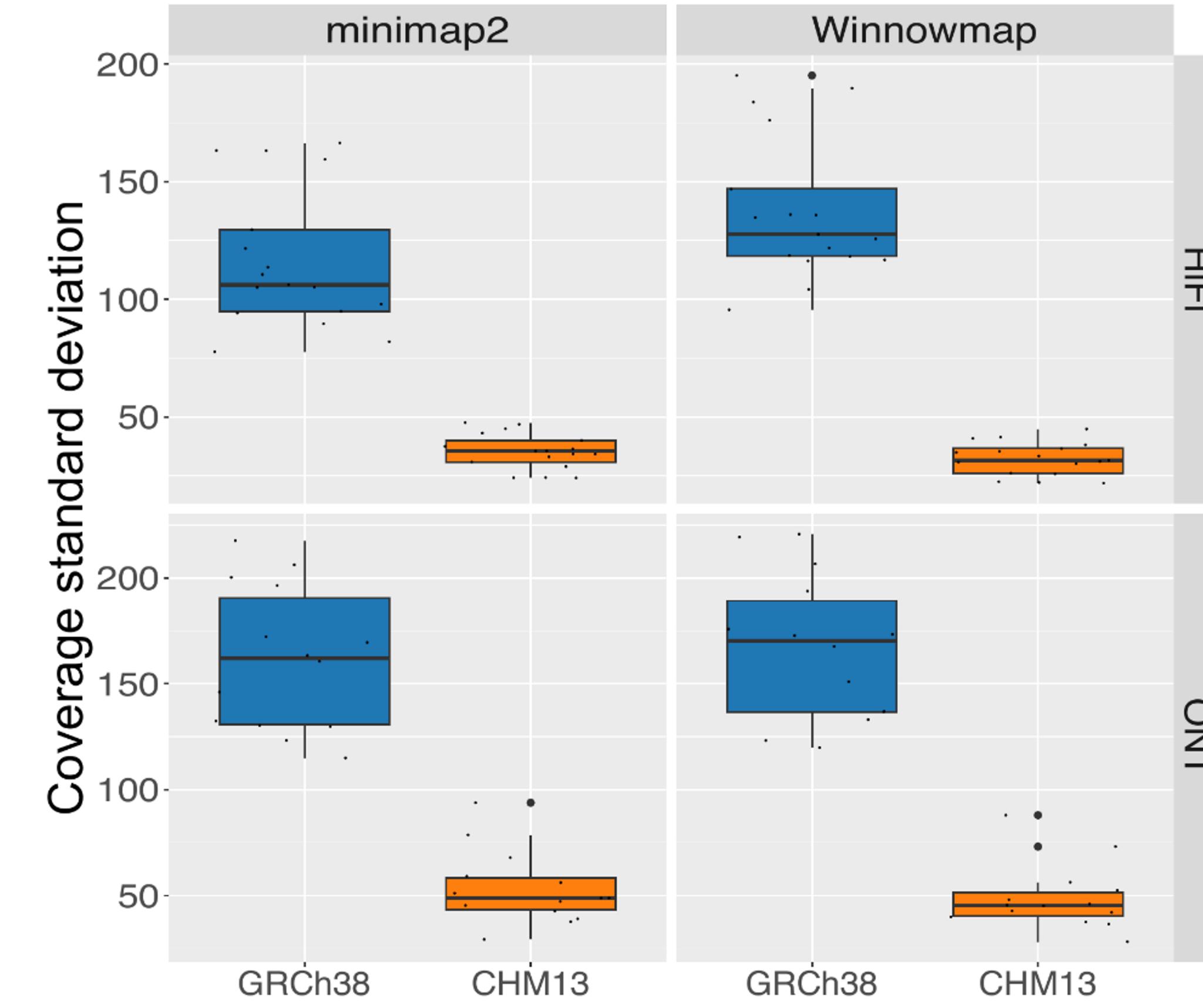
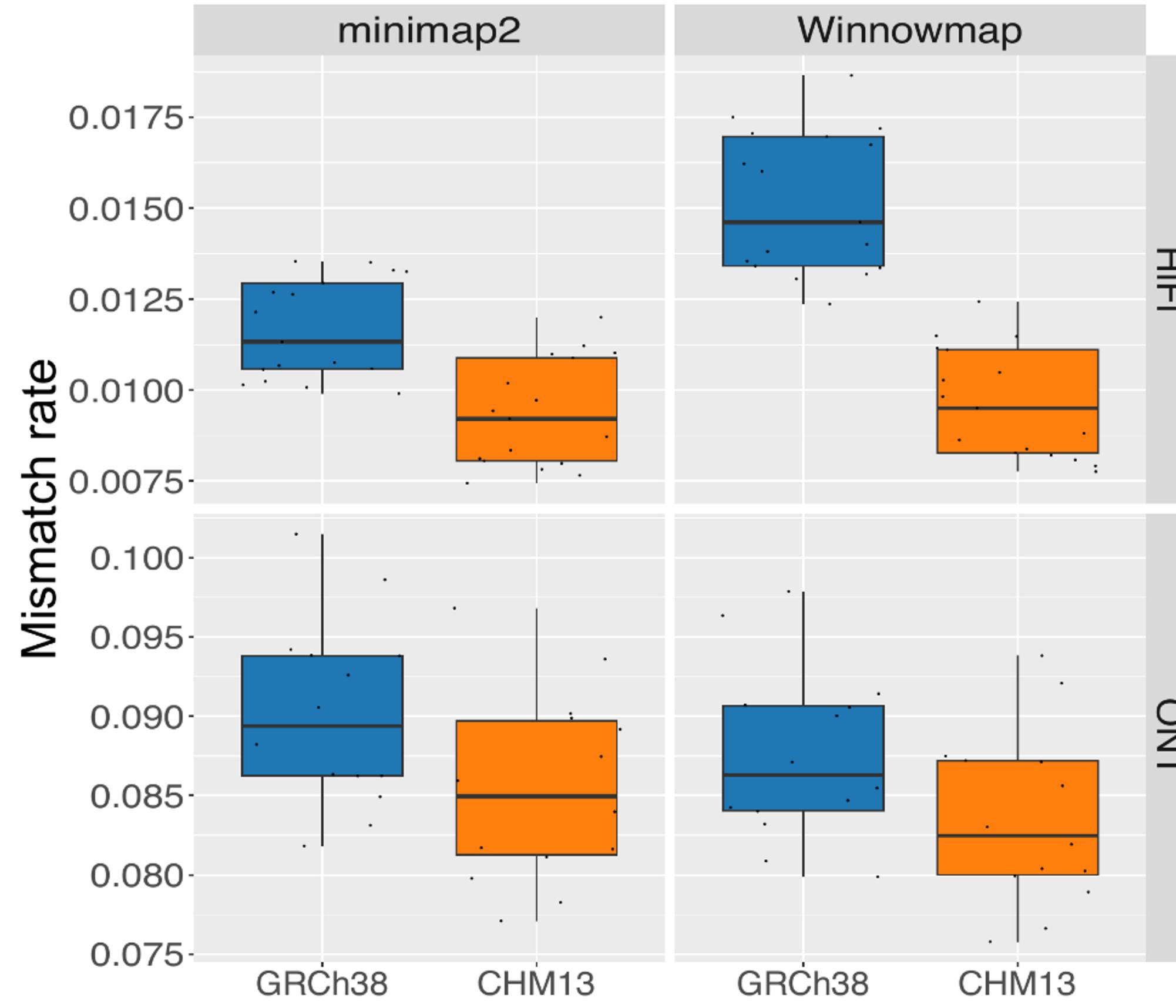


Melanie  
Kirsche

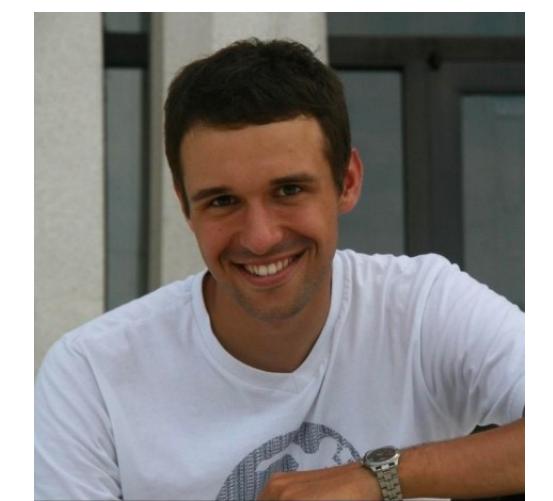


Sergey  
Aganezov

# T2T-CHM13 improves alignment for long reads

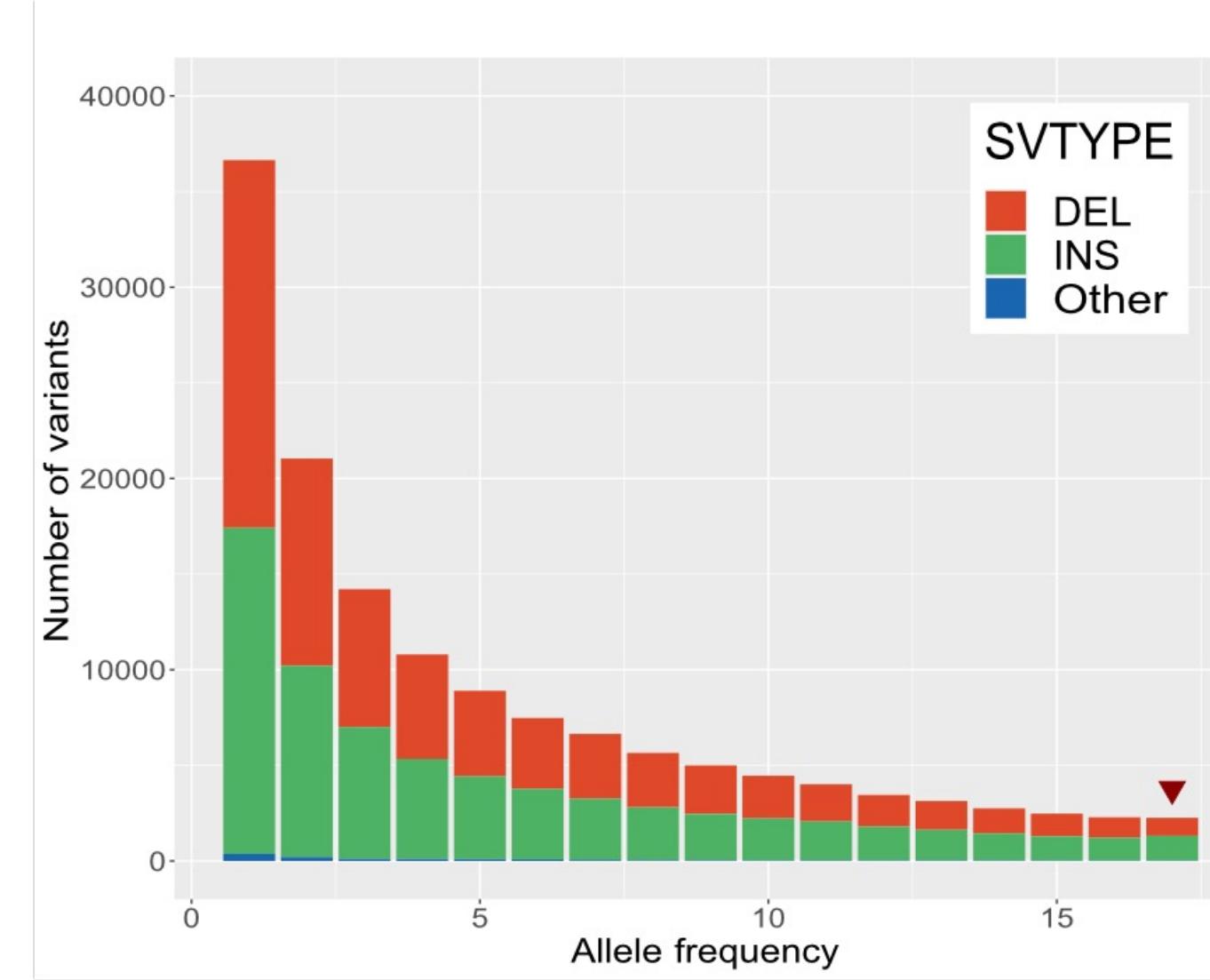
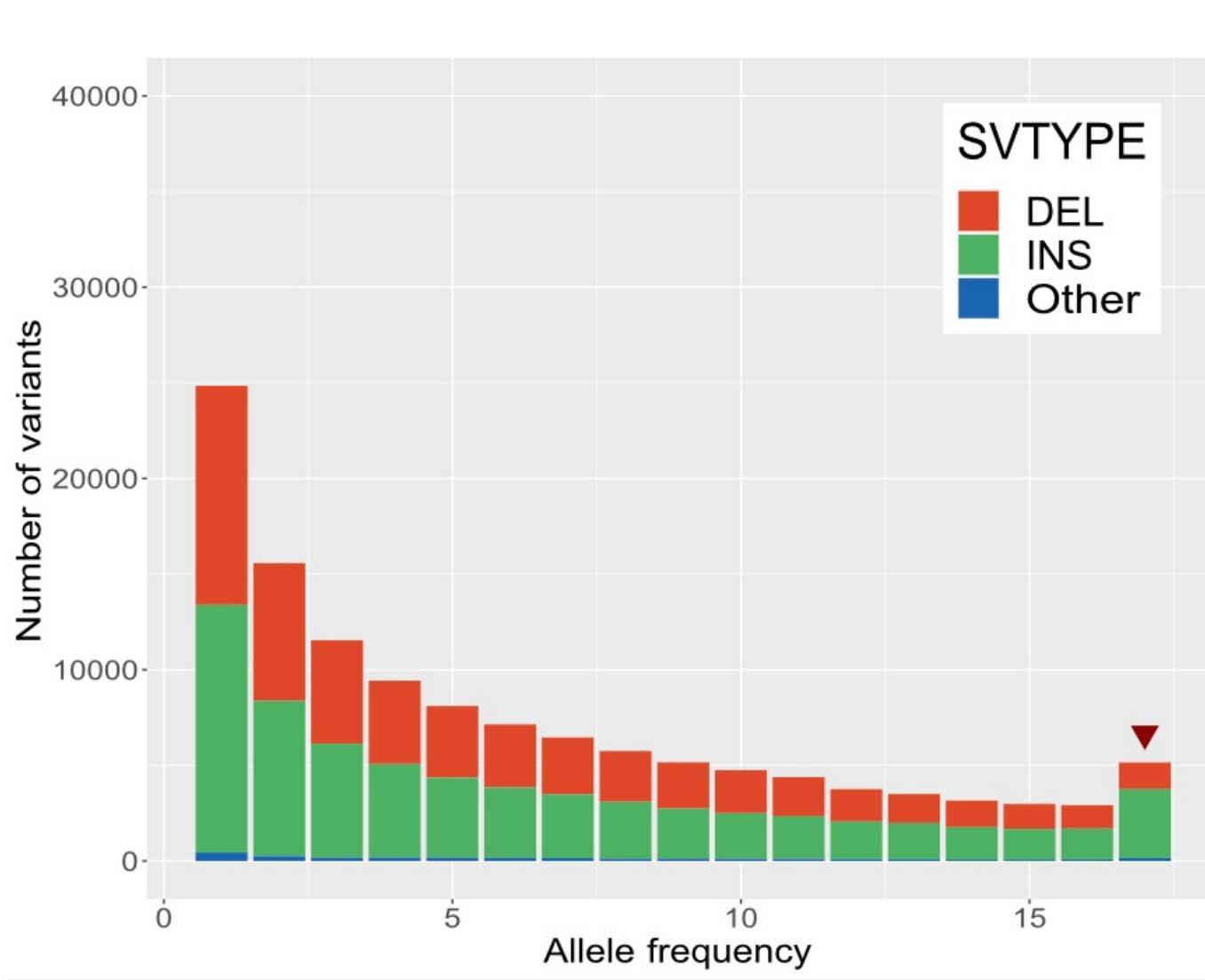


Melanie  
Kirsche



Sergey  
Aganezov

# T2T-CHM13 improves SV-calling with long reads

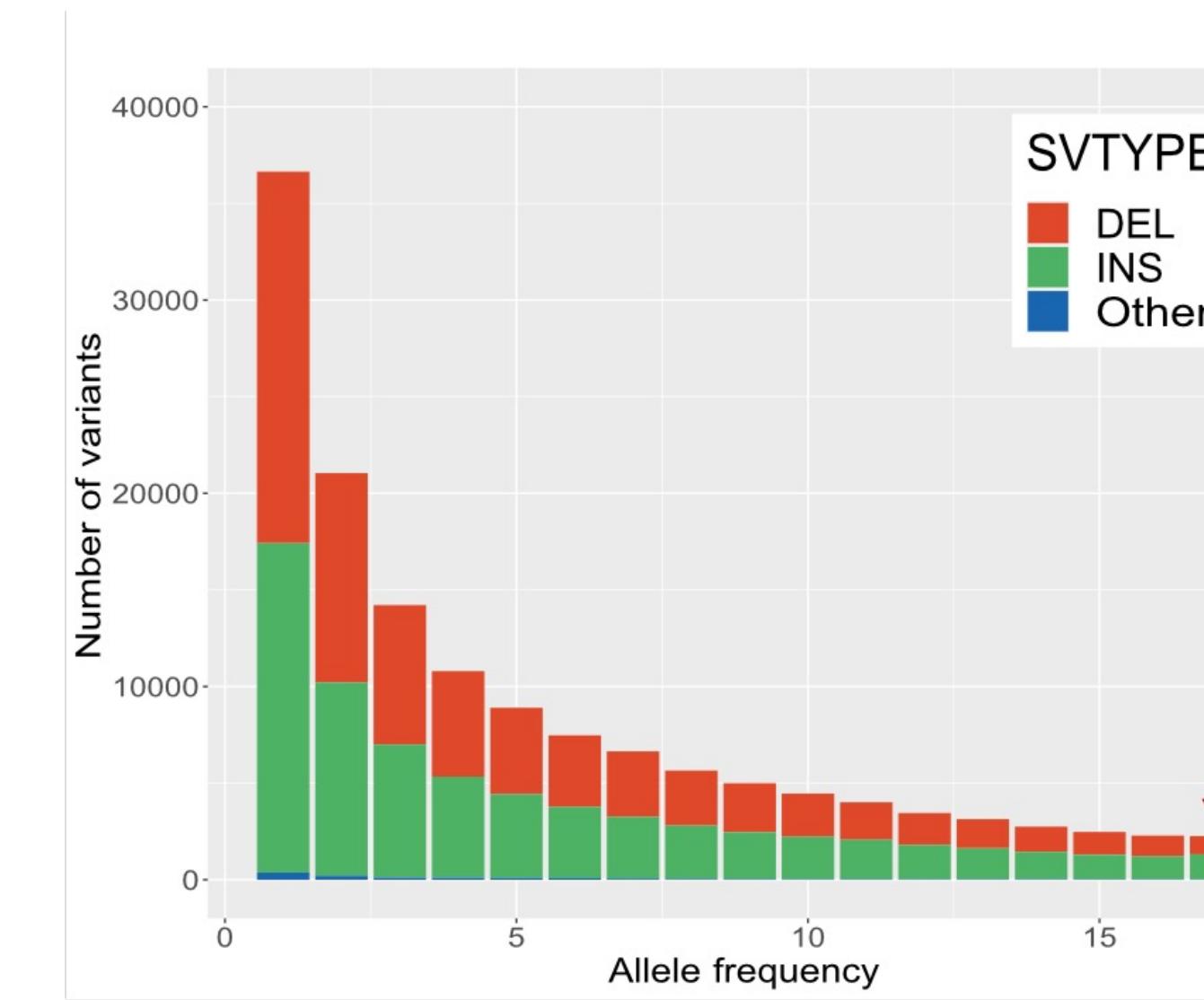
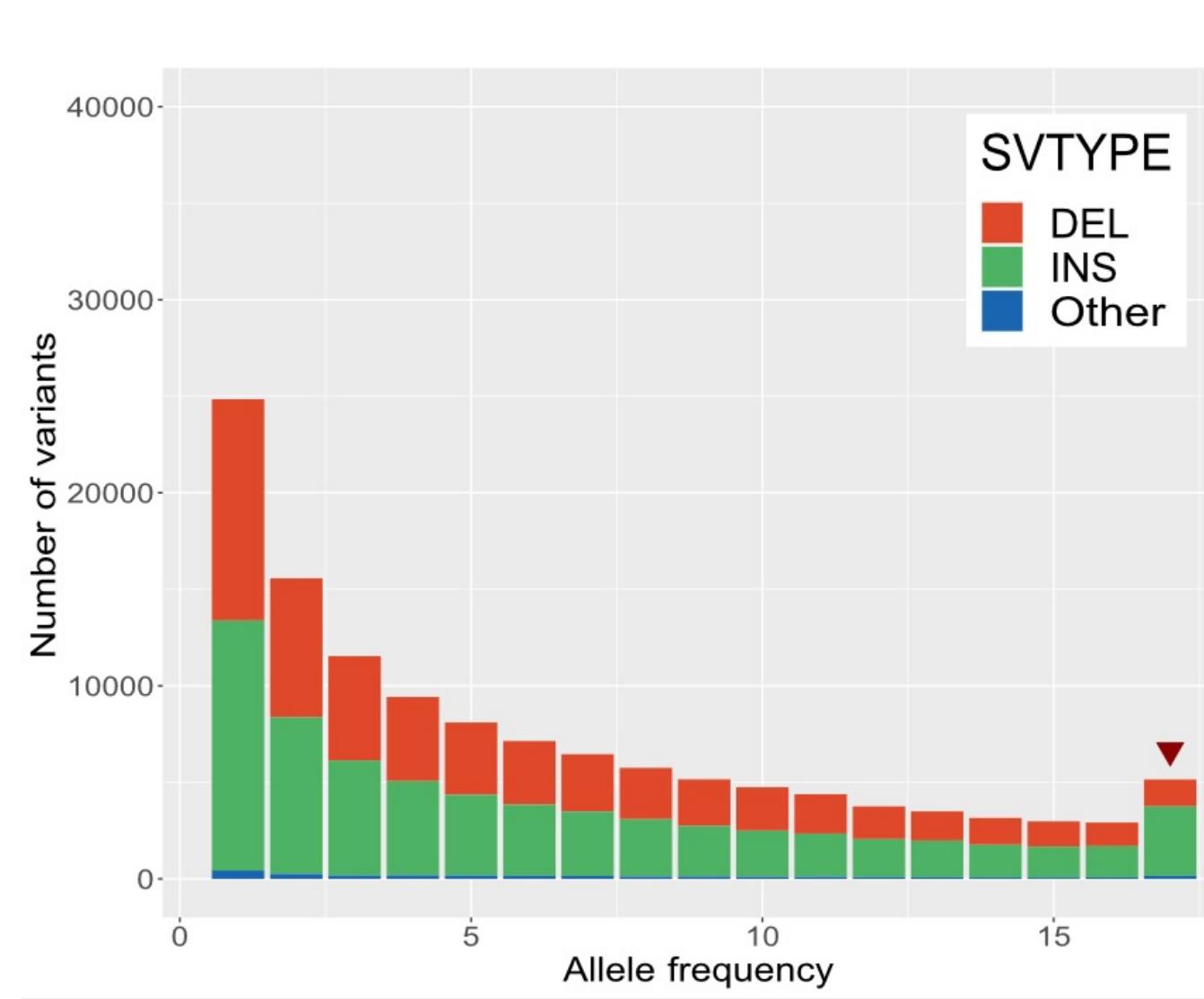


Melanie  
Kirsche

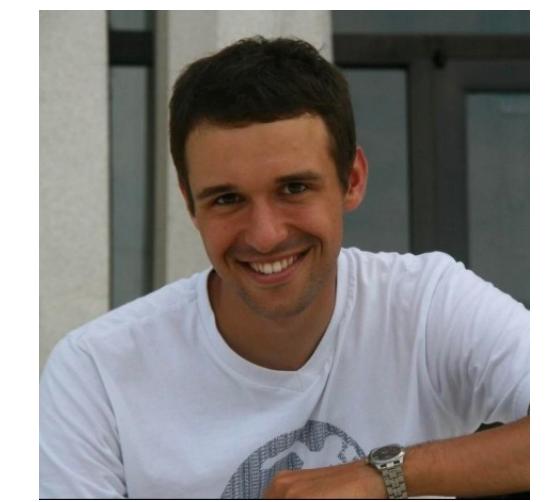
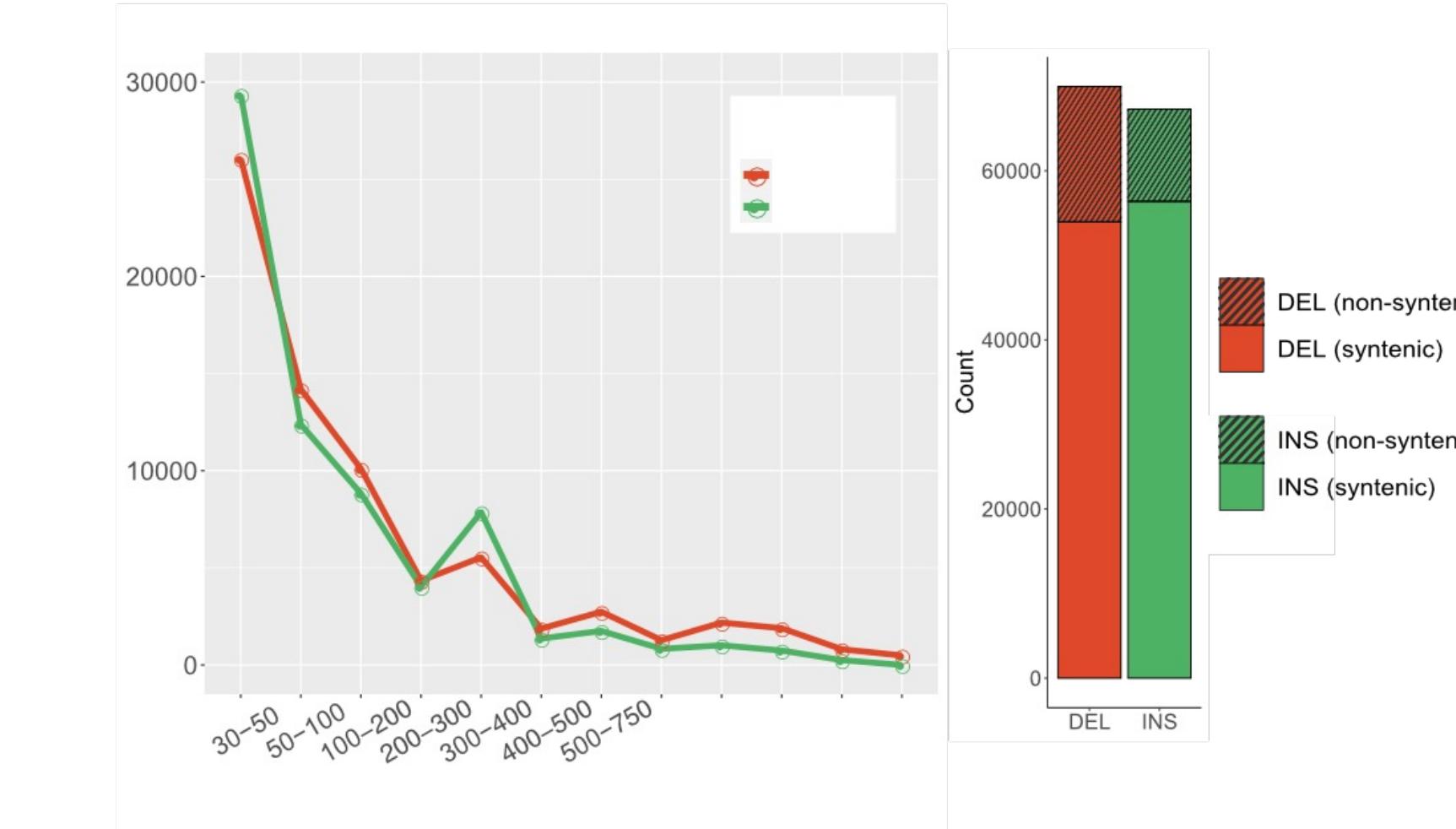
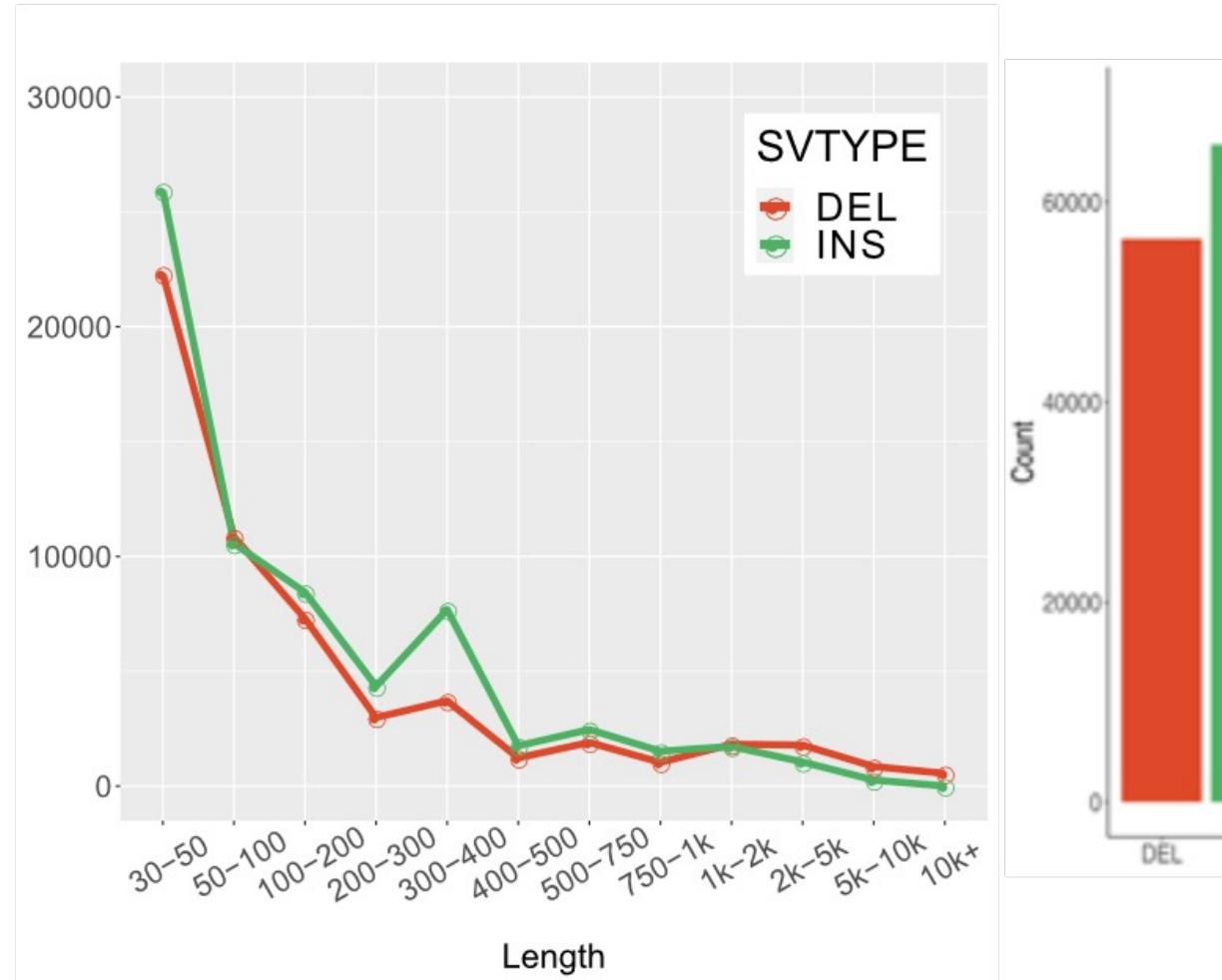


Sergey  
Aganezov

# T2T-CHM13 improves SV-calling with long reads



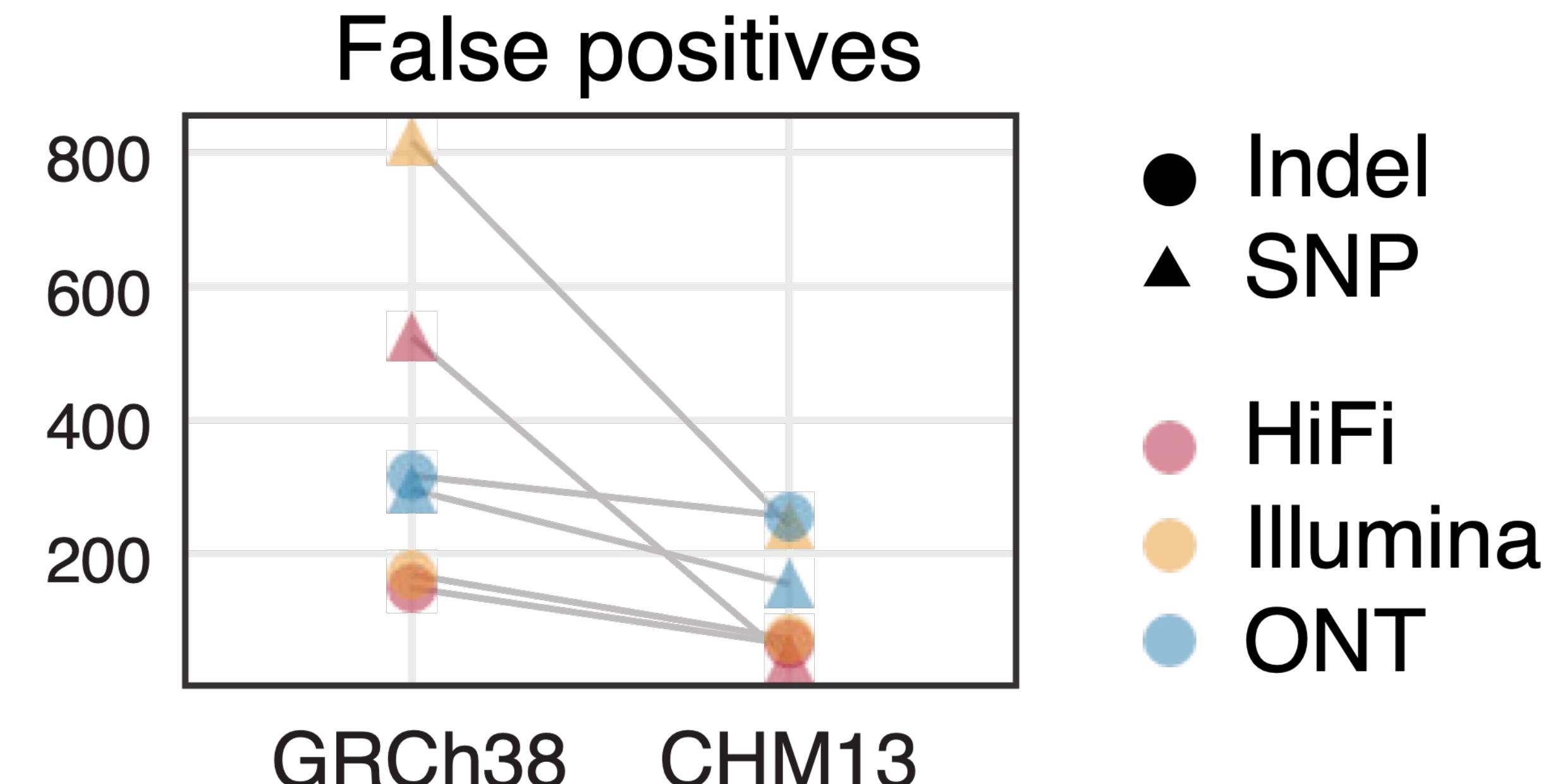
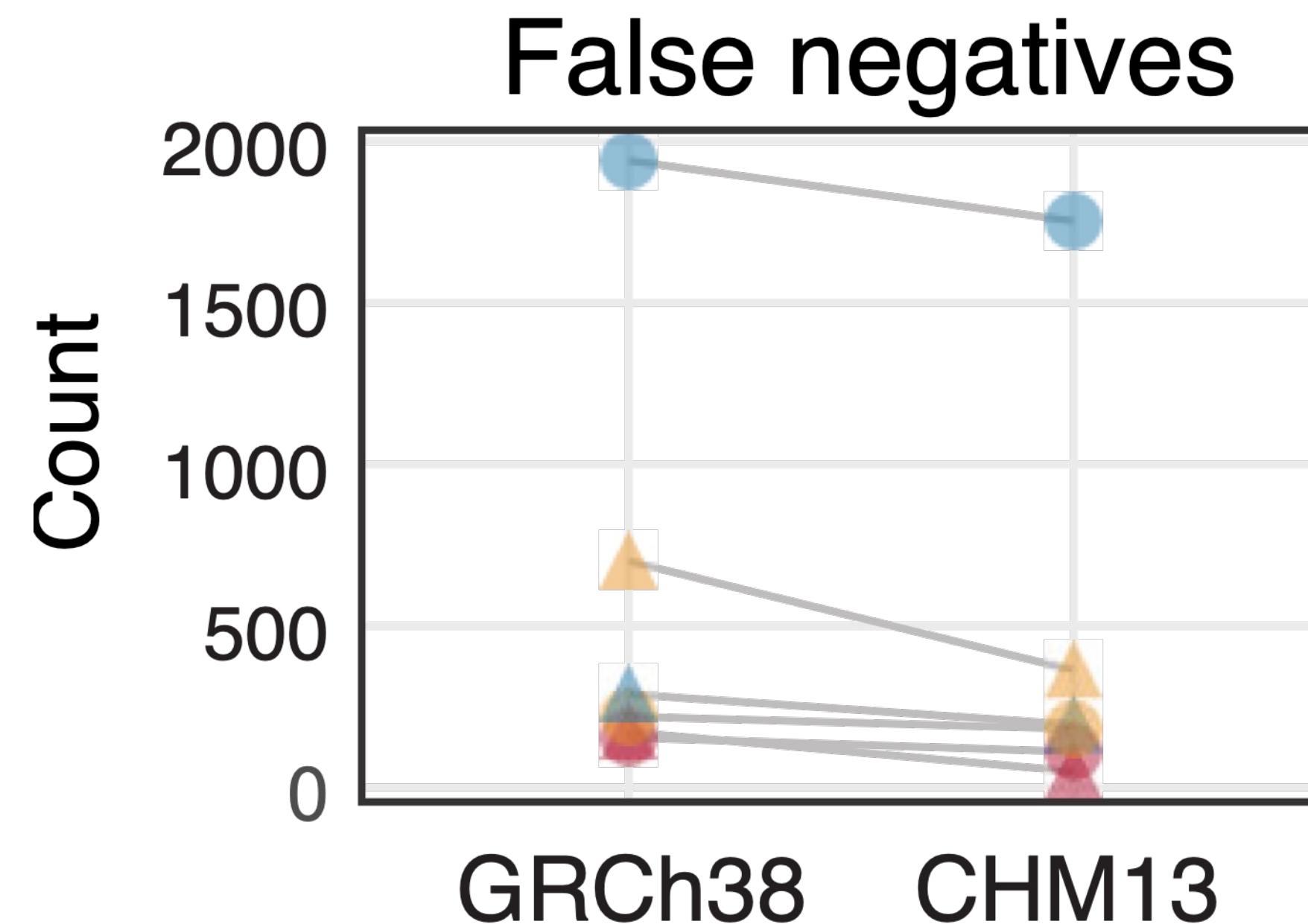
Melanie  
Kirsche



Sergey  
Aganezov

# T2T-CHM13 improves clinical genomics variant calling

- 273 challenging, medically relevant genes
- Benchmarked with sequencing data from HG002



Daniela  
Soto



Megan  
Dennis



Justin  
Zook

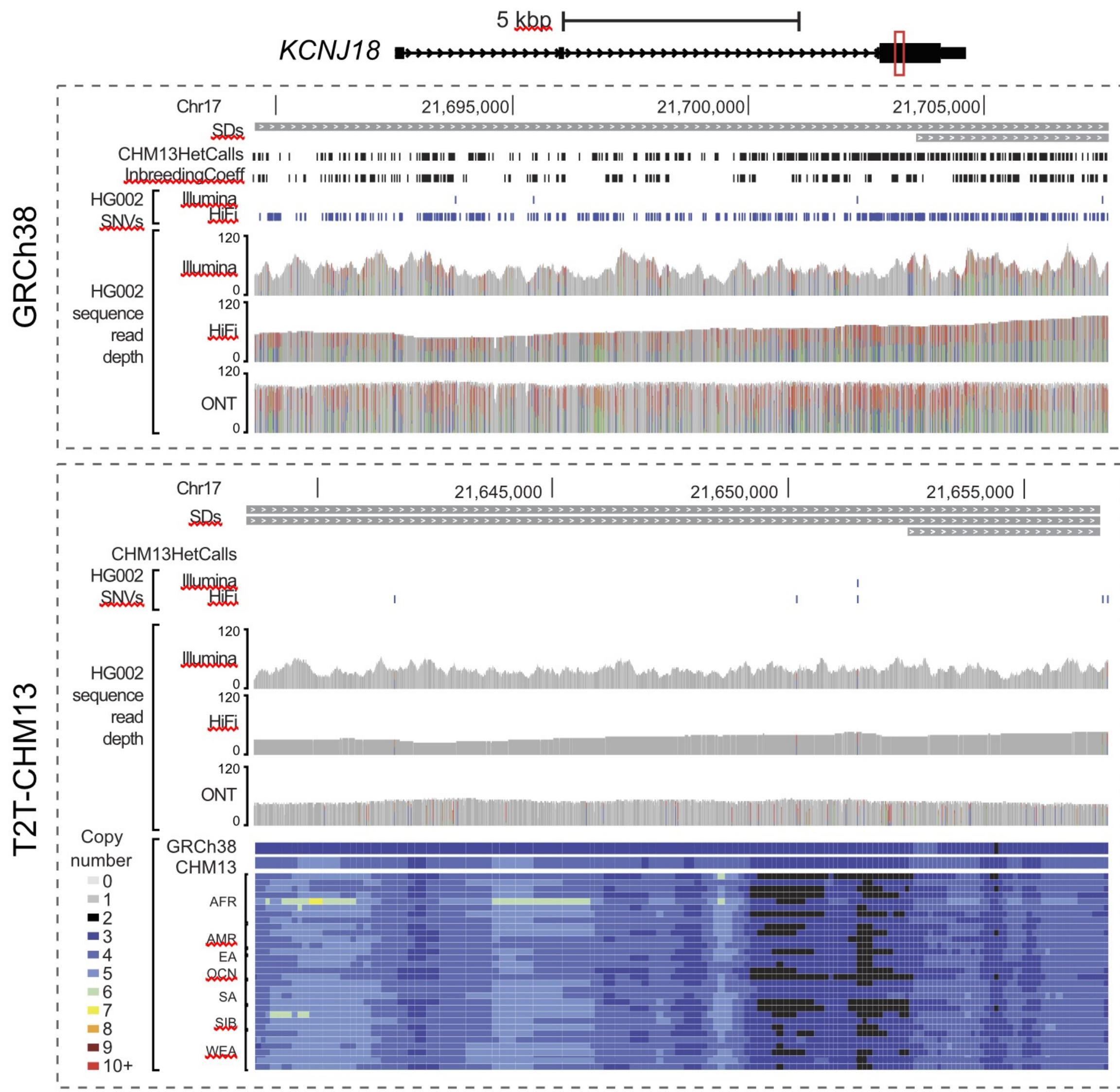


Fritz  
Sedlazeck



Danny  
Miller

# T2T-CHM13 improves clinical genomics variant calling



Daniela  
Soto



Megan  
Dennis



Justin  
Zook



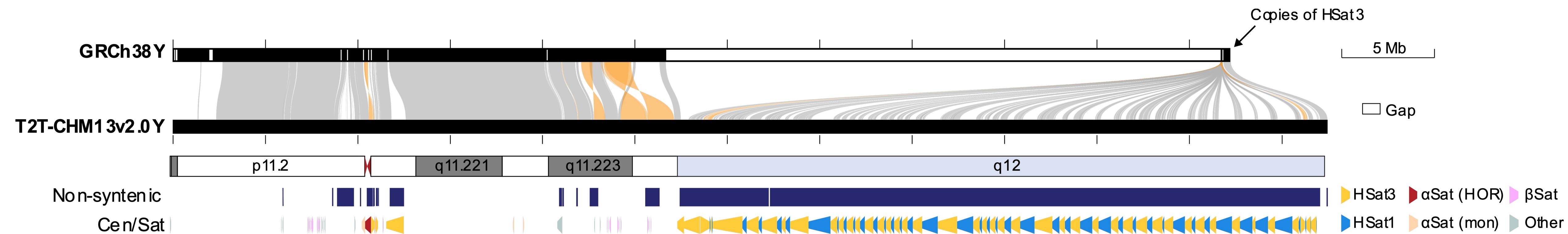
Fritz  
Sedlazeck



Danny  
Miller

# A telomere-to-telomere Y chromosome

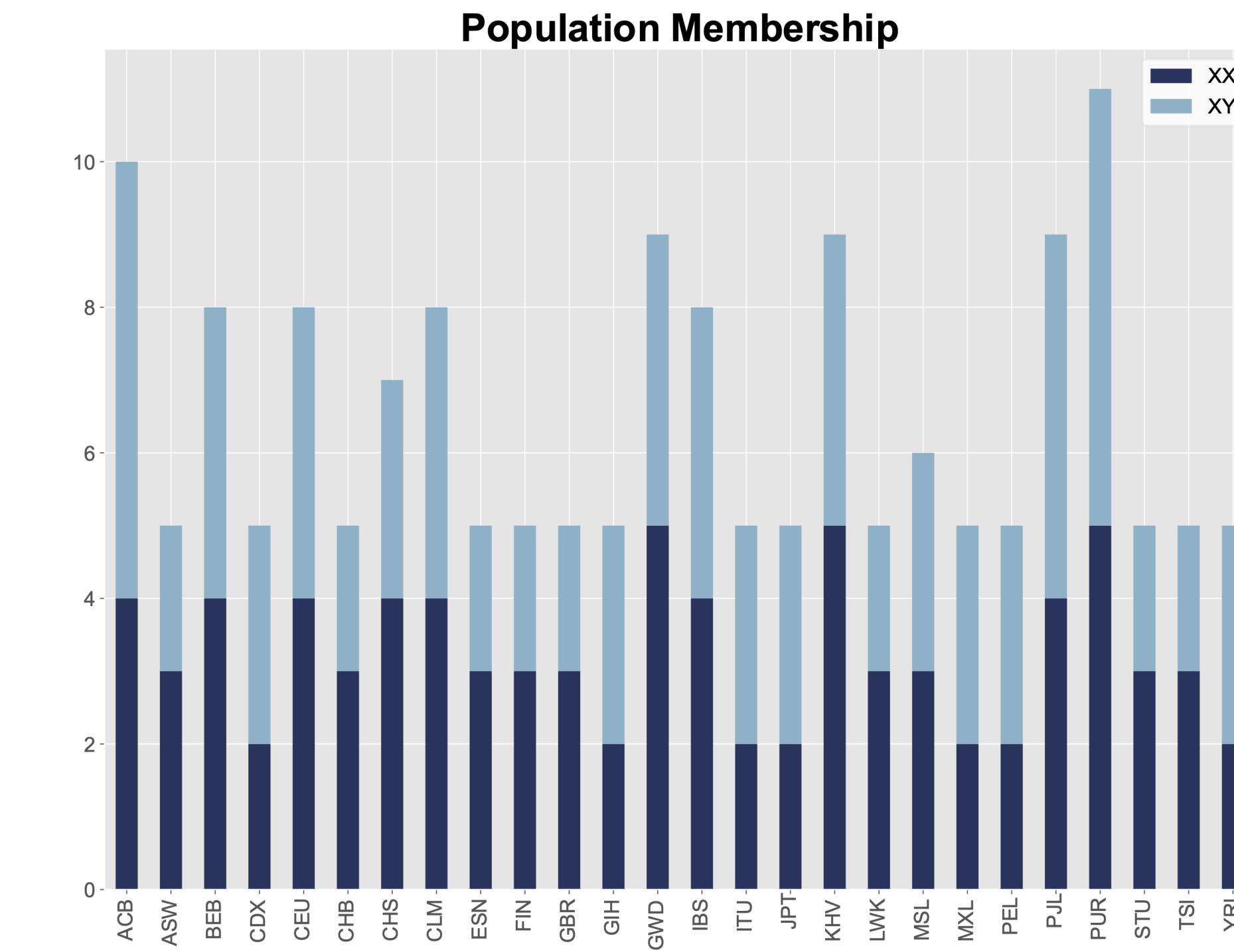
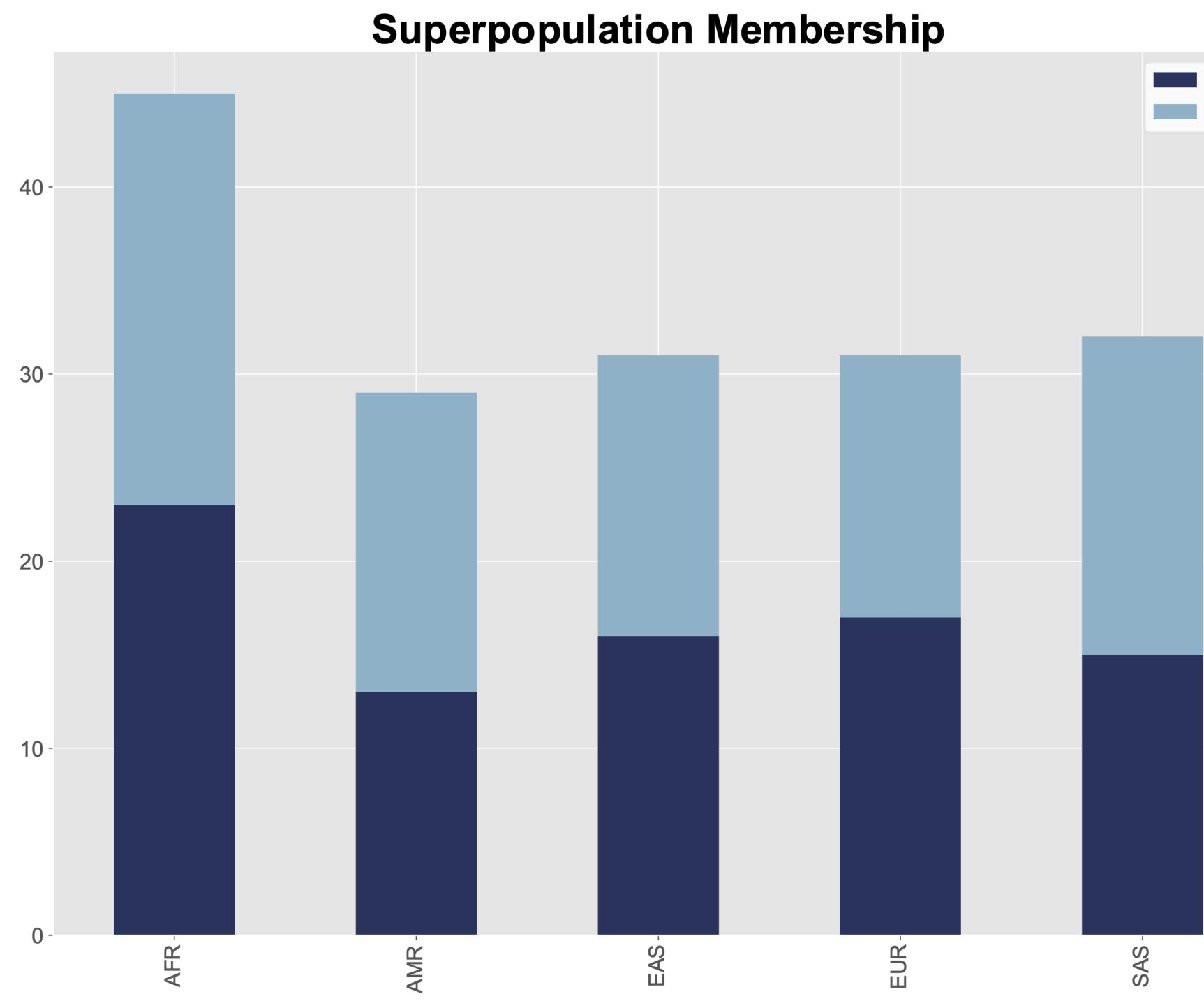
- CHM13 is an XX complete hydatidiform mole
- T2T-CHM13 v1 used the **GRCh38** Y chromosome
- More than 50% of the GRCh38 Y chromosome assembly is missing
- Used long-read sequencing to generate a complete Y chromosome assembly from HG002 cell line, completing the **T2T-CHM13 v2.0** reference



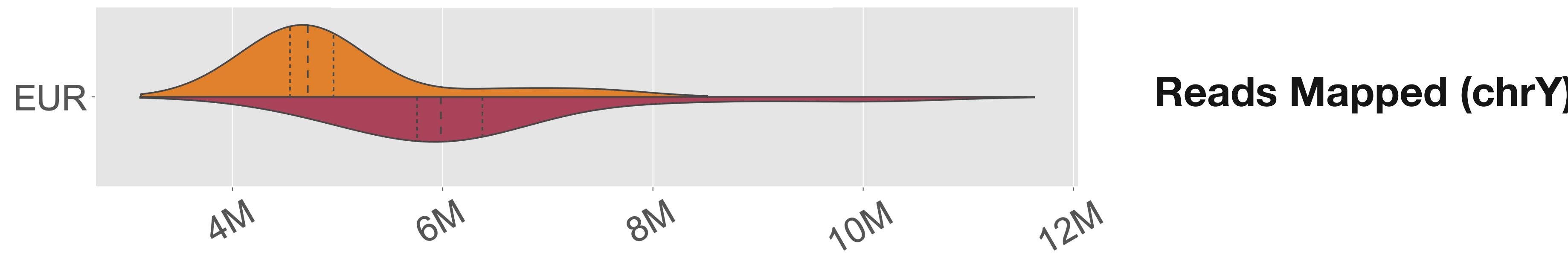
Arang  
Rhee

# Analyzing diverse, short-read data with T2T-CHM13 v2.0

- 1000 Genomes Project (1KGP): 168 samples (84 XX, 84 XY)
- 30x sequencing by the New York Genome Center



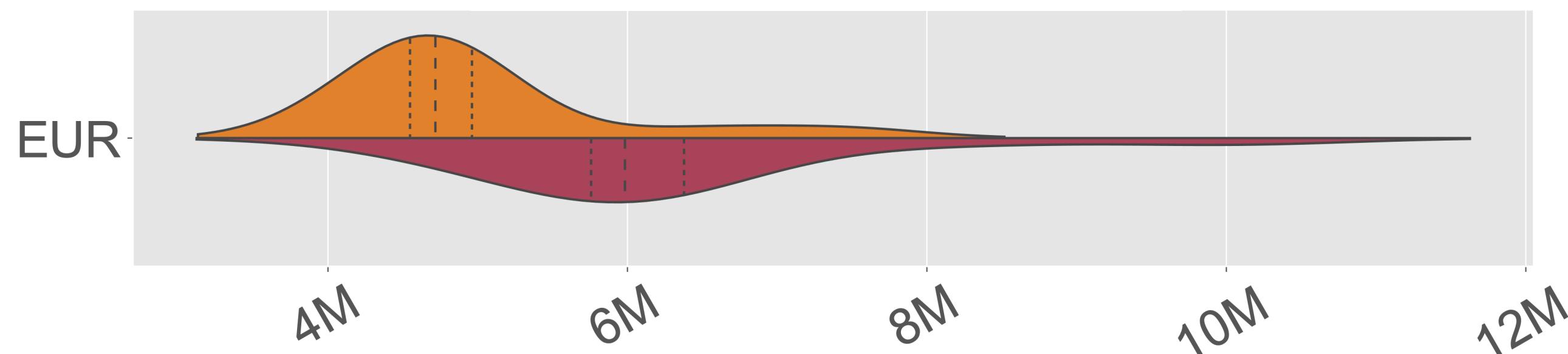
# HG002Y improves short-read alignment



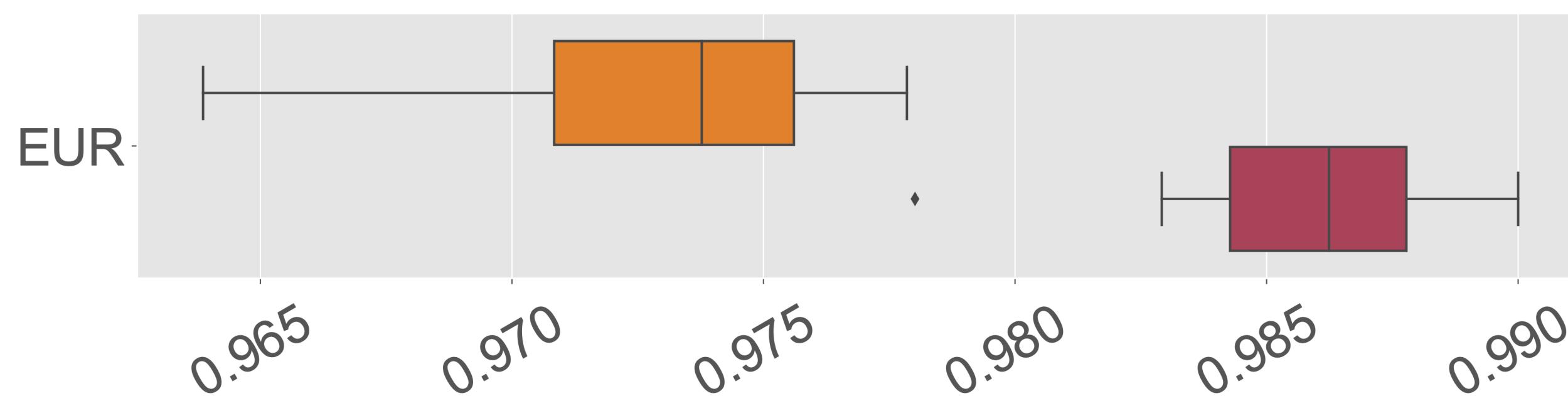
CHM13v1 (GRCh38Y)

CHM13v2 (HG002Y)

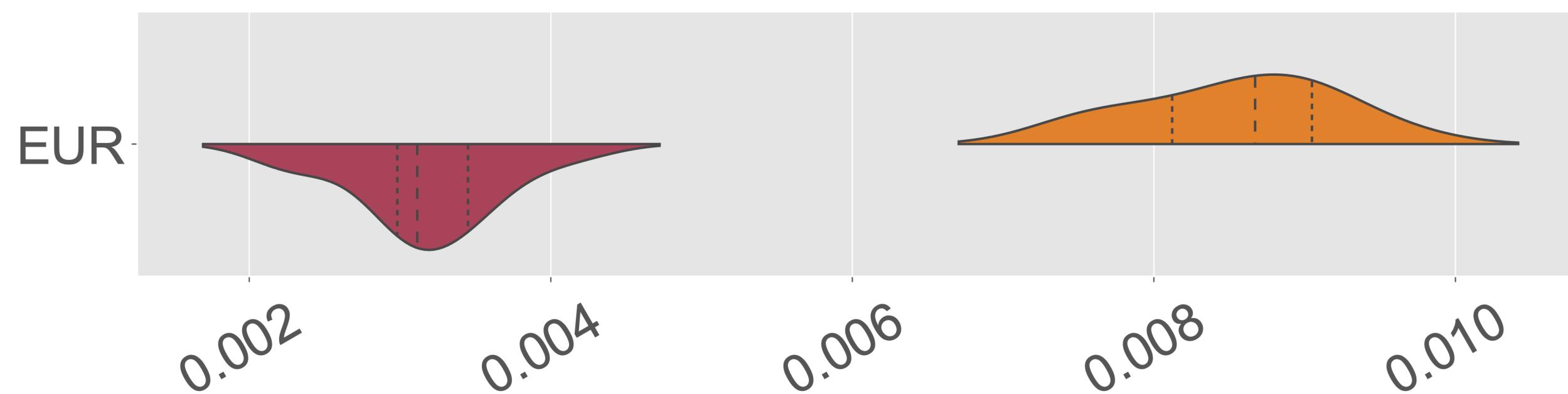
# HG002Y improves short-read alignment



**Reads Mapped (chrY)**



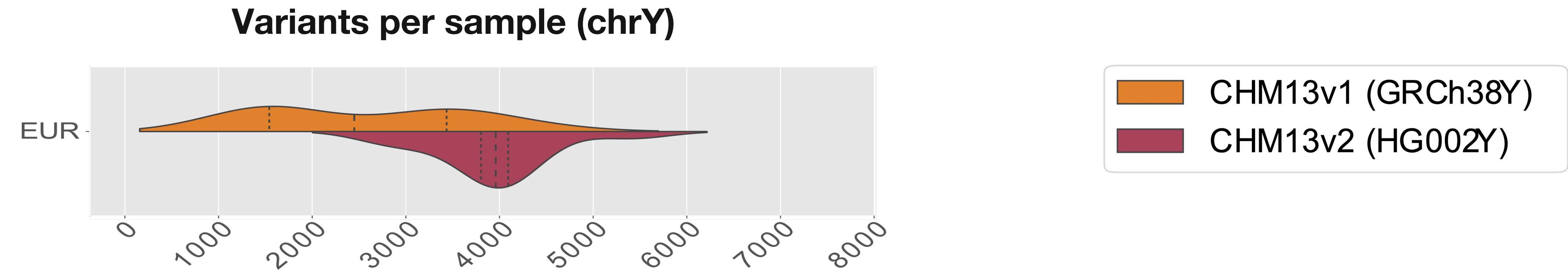
**Proportion of reads  
properly paired (chrY)**



**Mismatch rate (chrY)**

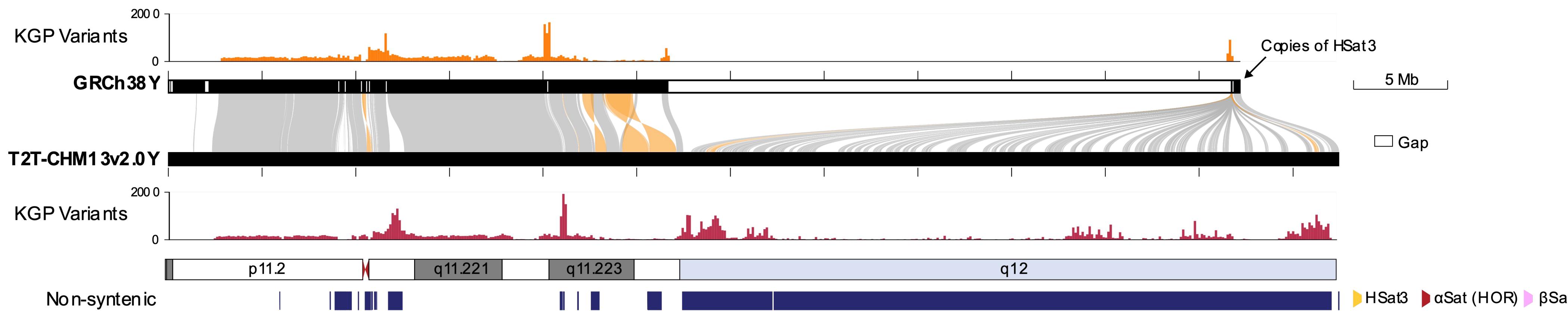
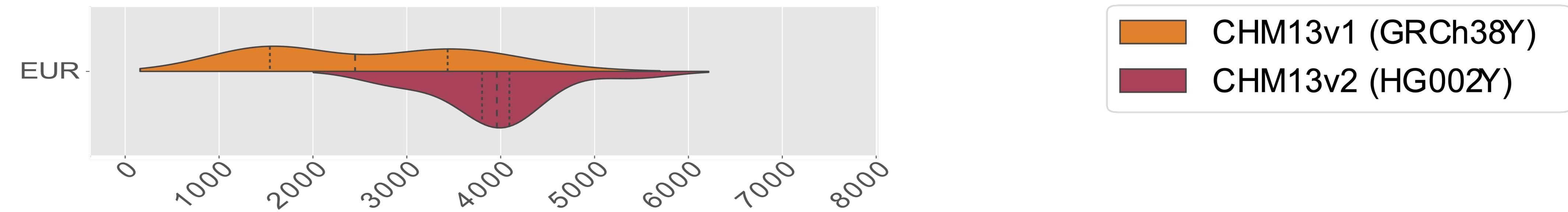
CHM13v1 (GRCh38Y)  
CHM13v2 (HG002Y)

# HG002Y improves variant calling in diverse samples



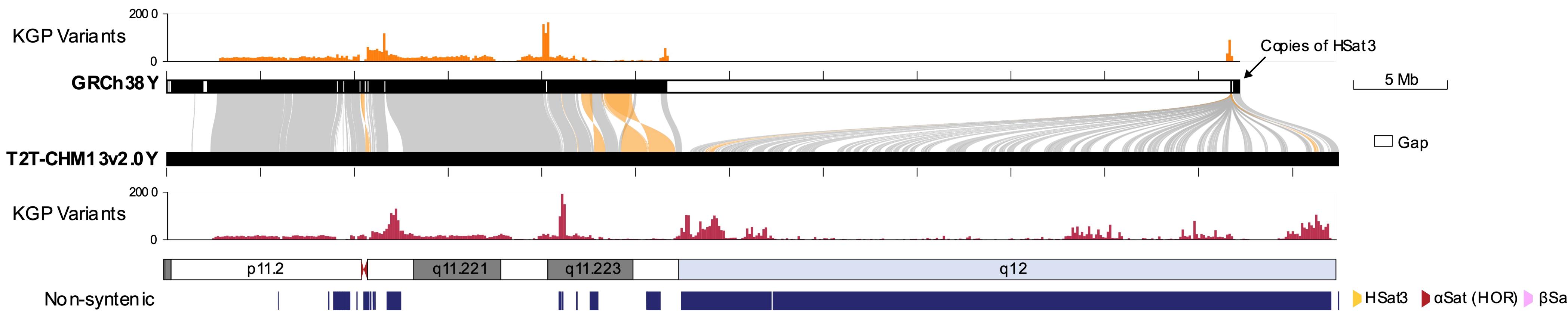
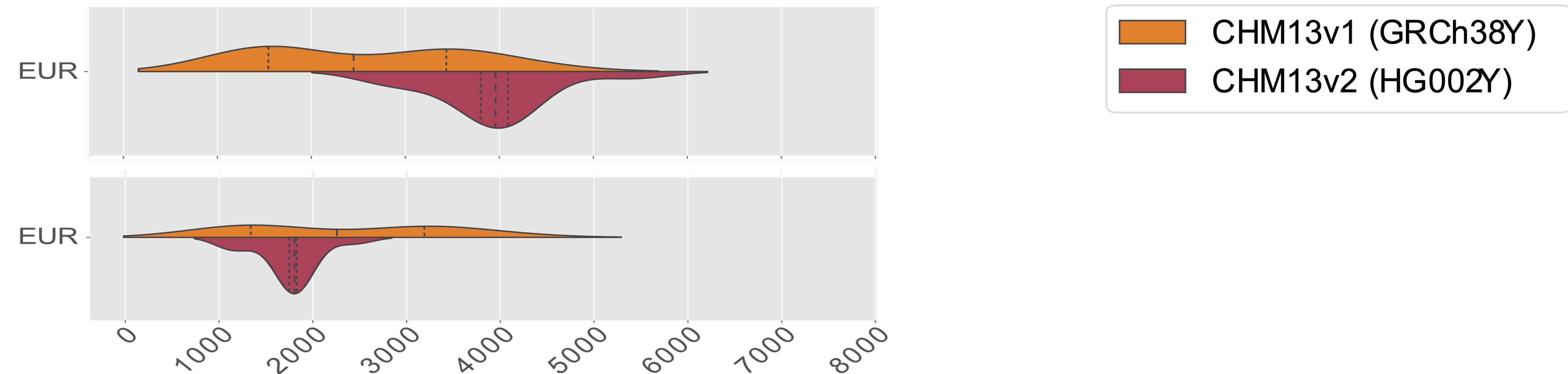
# HG002Y improves variant calling in diverse samples

Variants per sample (chrY)



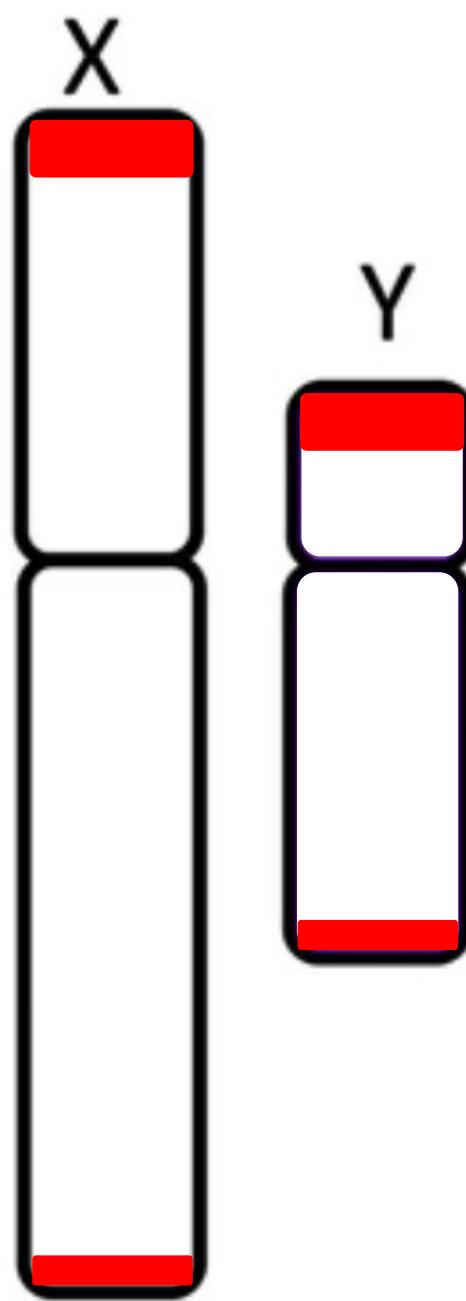
# HG002Y improves variant calling in diverse samples

Variants per sample (chrY syntenic)

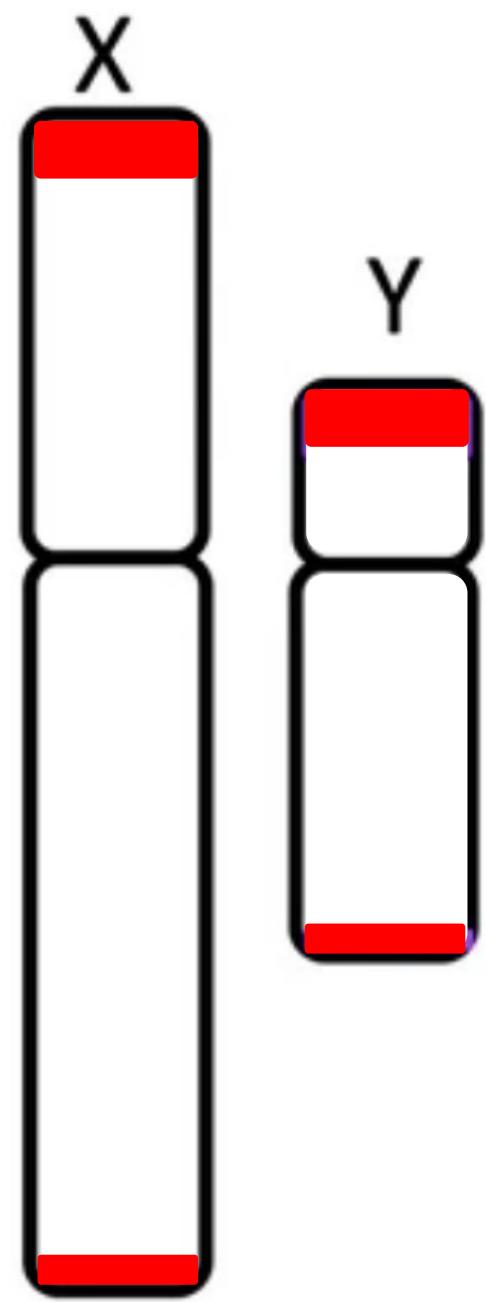


# Pseudo-autosomal regions on the sex chromosomes

XX Samples



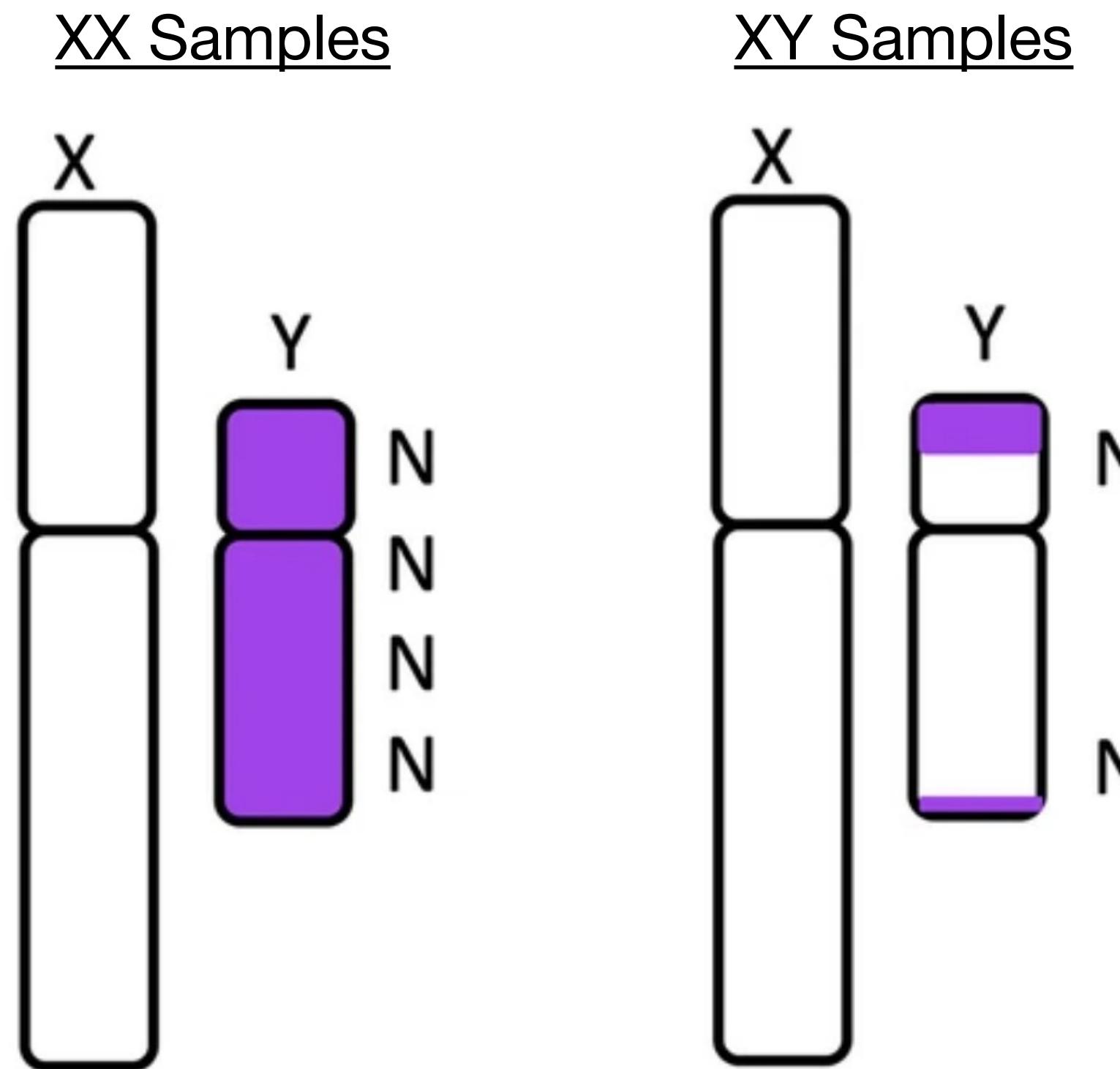
XY Samples



Adapted from Olney, K. C., et al. (2020) *BMC*

# Karyotype-specific alignment

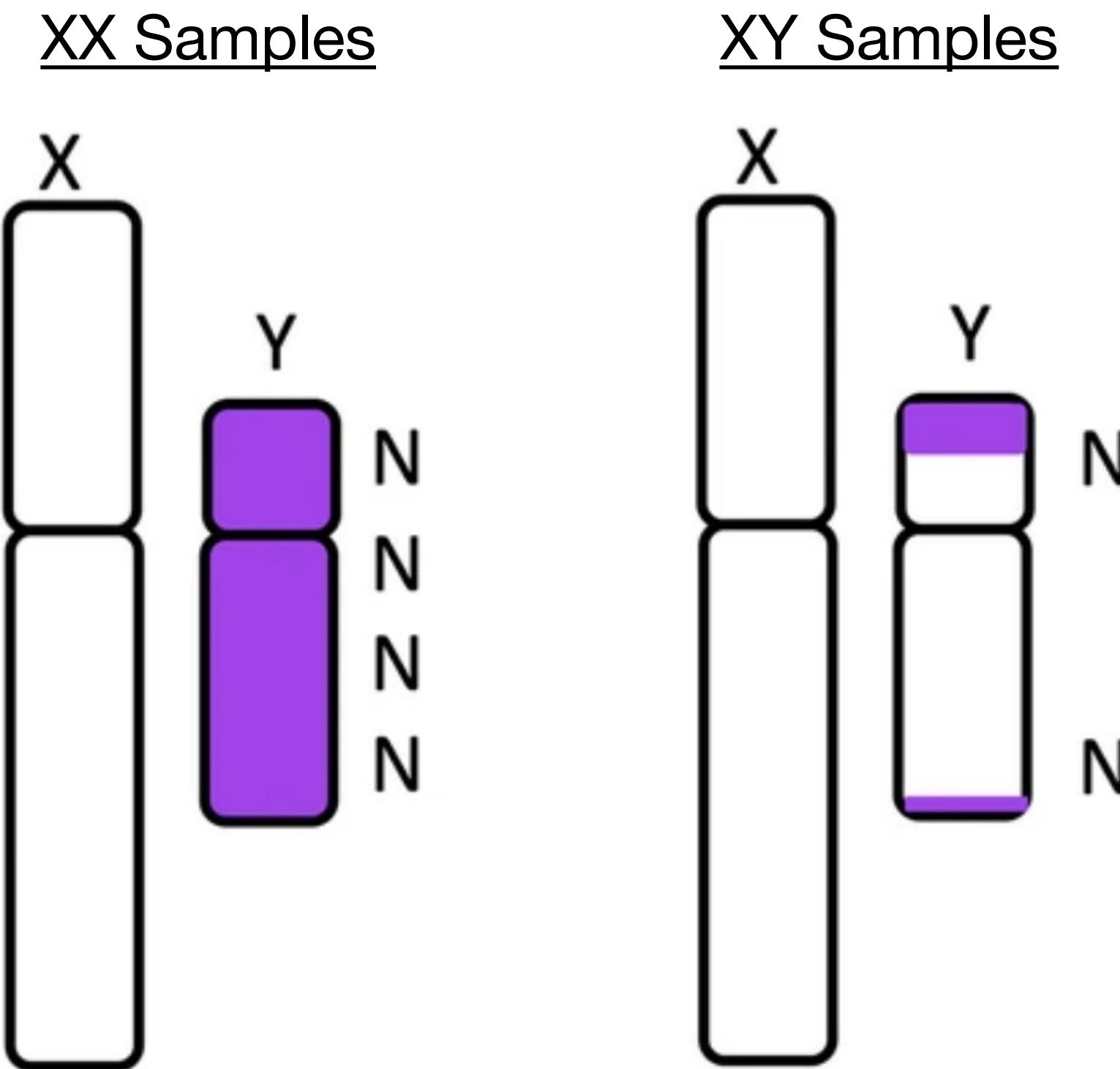
Used **XYalign** for alignment



Adapted from Olney, K. C., et al. (2020) *BMC*

# Karyotype-specific alignment

Used **XYalign** for alignment



Adapted from Olney, K. C., et al. (2020) *BMC*

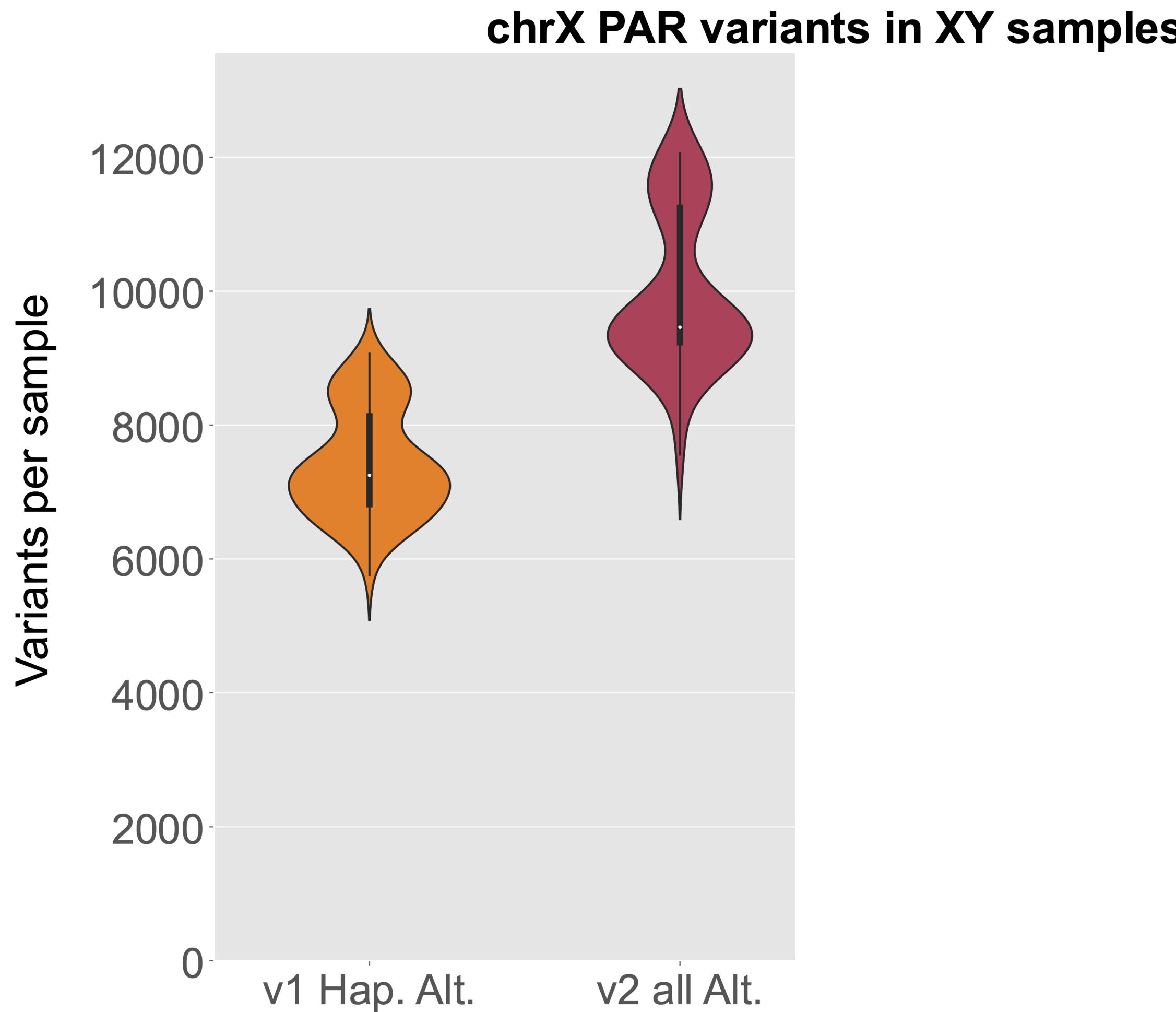
## Variant Calling Pipeline

- XX samples
  - chrX: diploid
  - chrY: hard masked
- XY samples:
  - chrX PAR: diploid
  - chrX non-PAR: haploid
  - chrY non-PAR: haploid
  - chrY PAR: hard masked

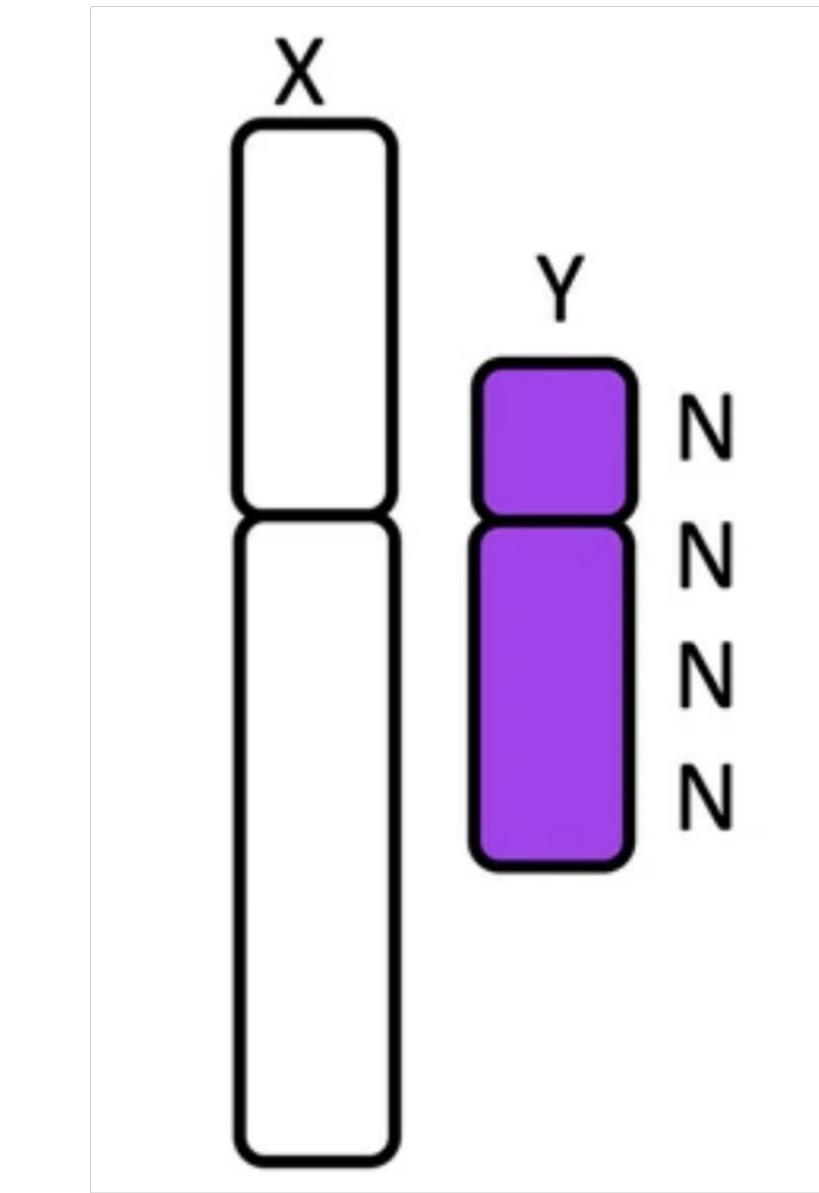


**Samantha  
Zarate**

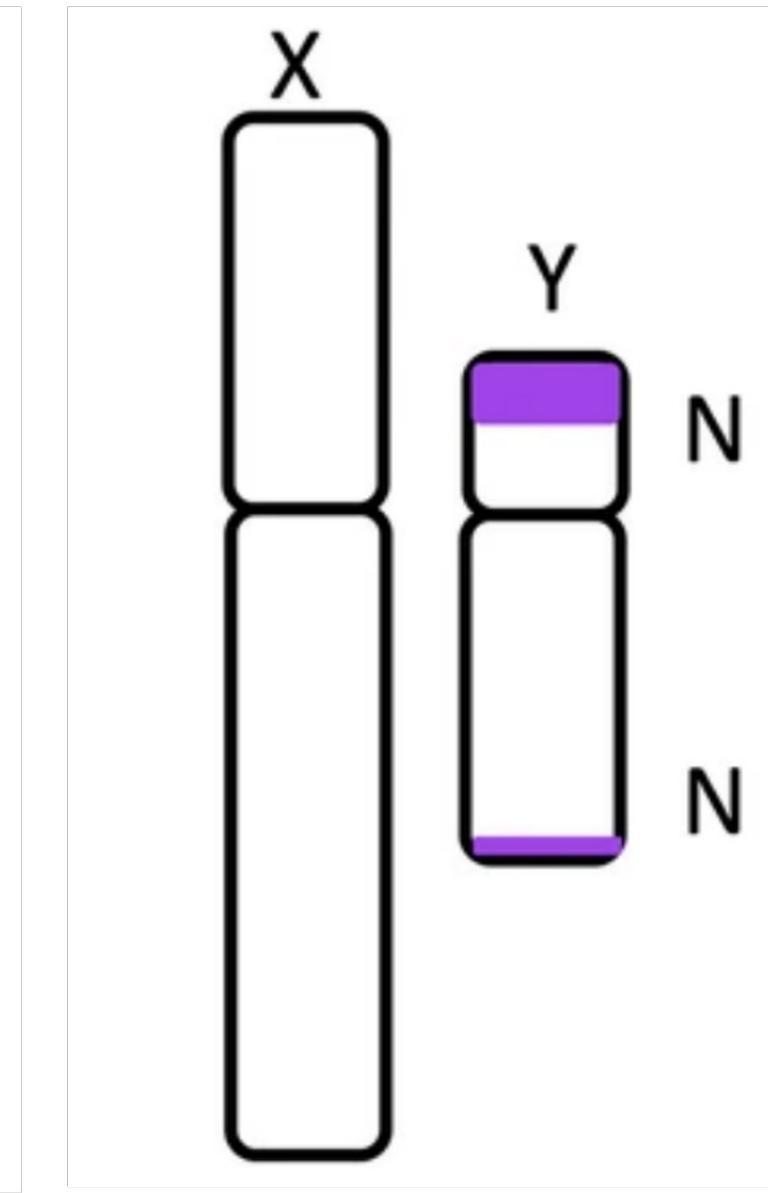
# Karyotype-specific alignment elucidates PAR genotypes



XX Samples



XY Samples

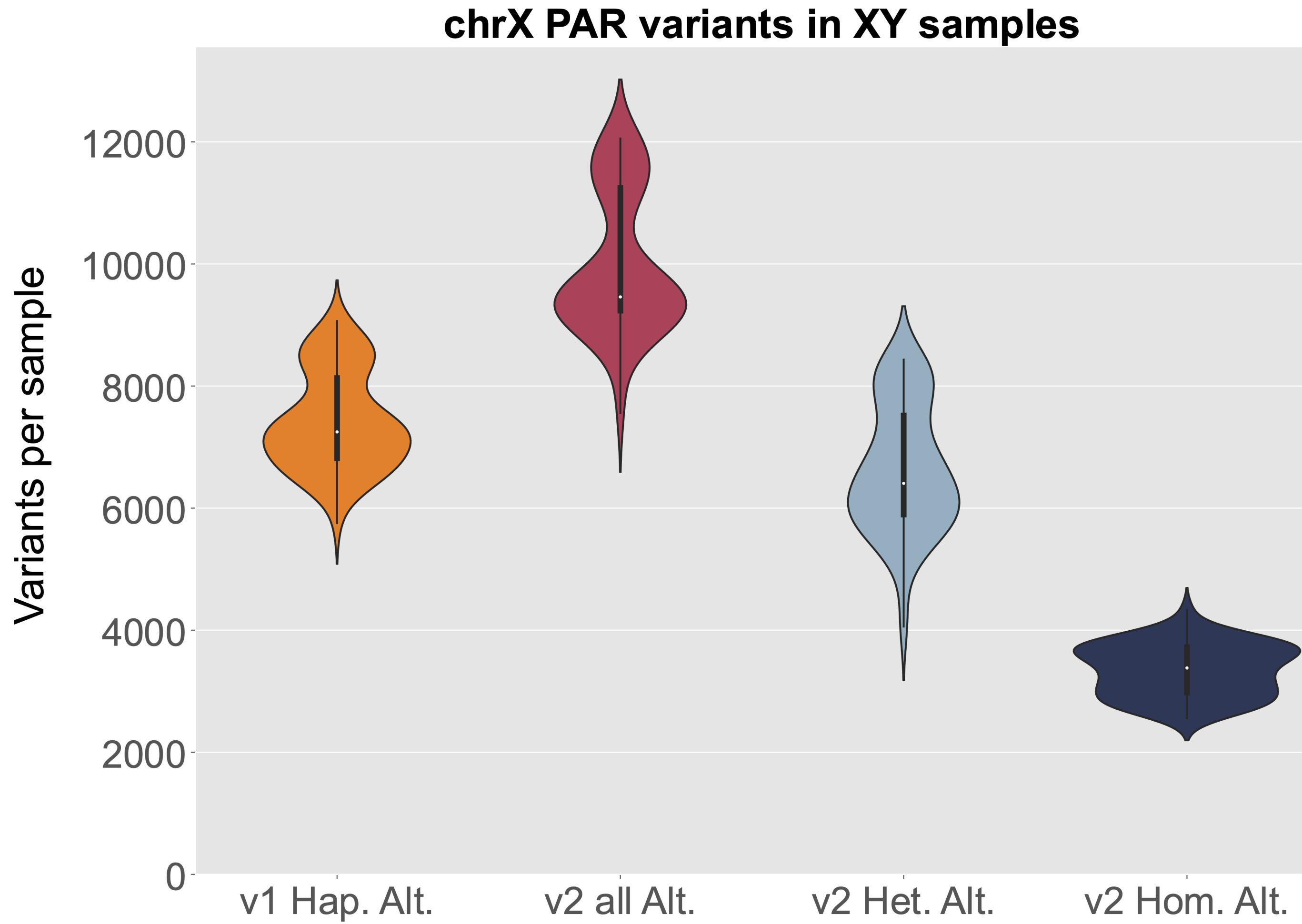


Adapted from Olney, K. C., et al. (2020) *BMC*

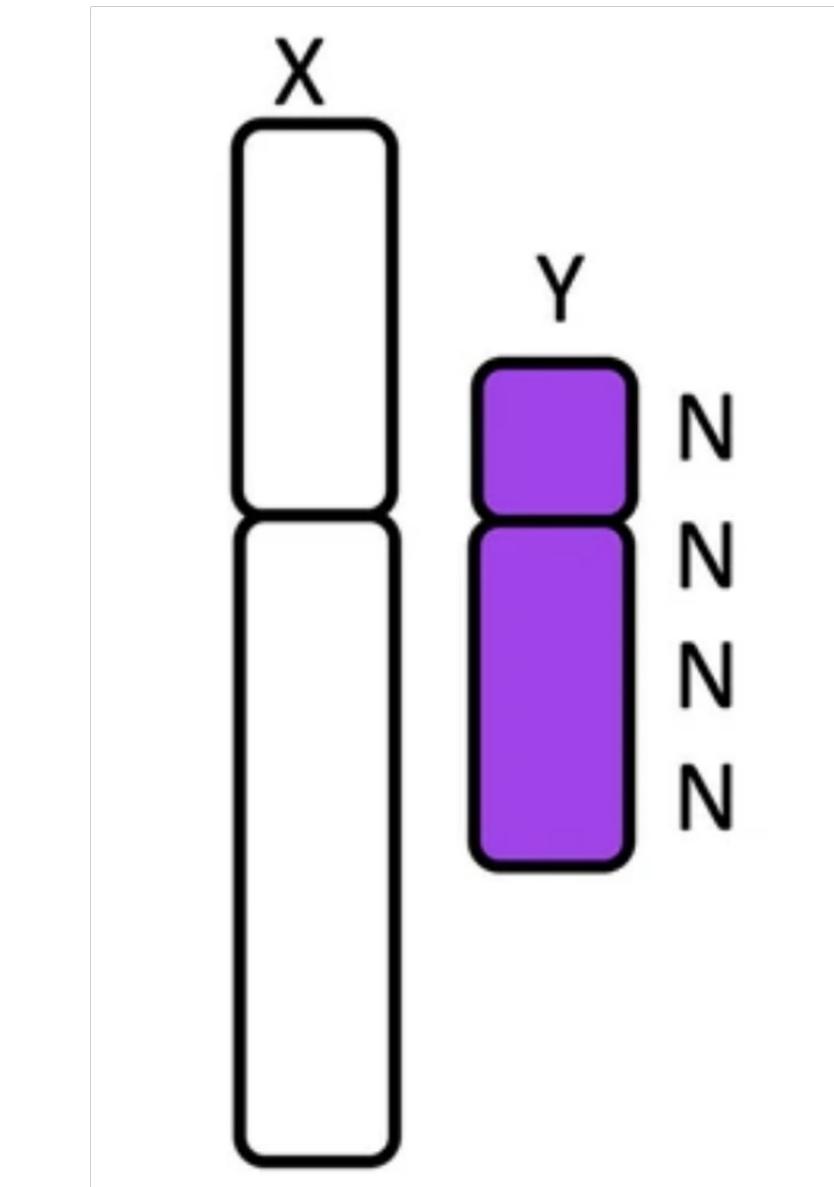


**Samantha  
Zarate**

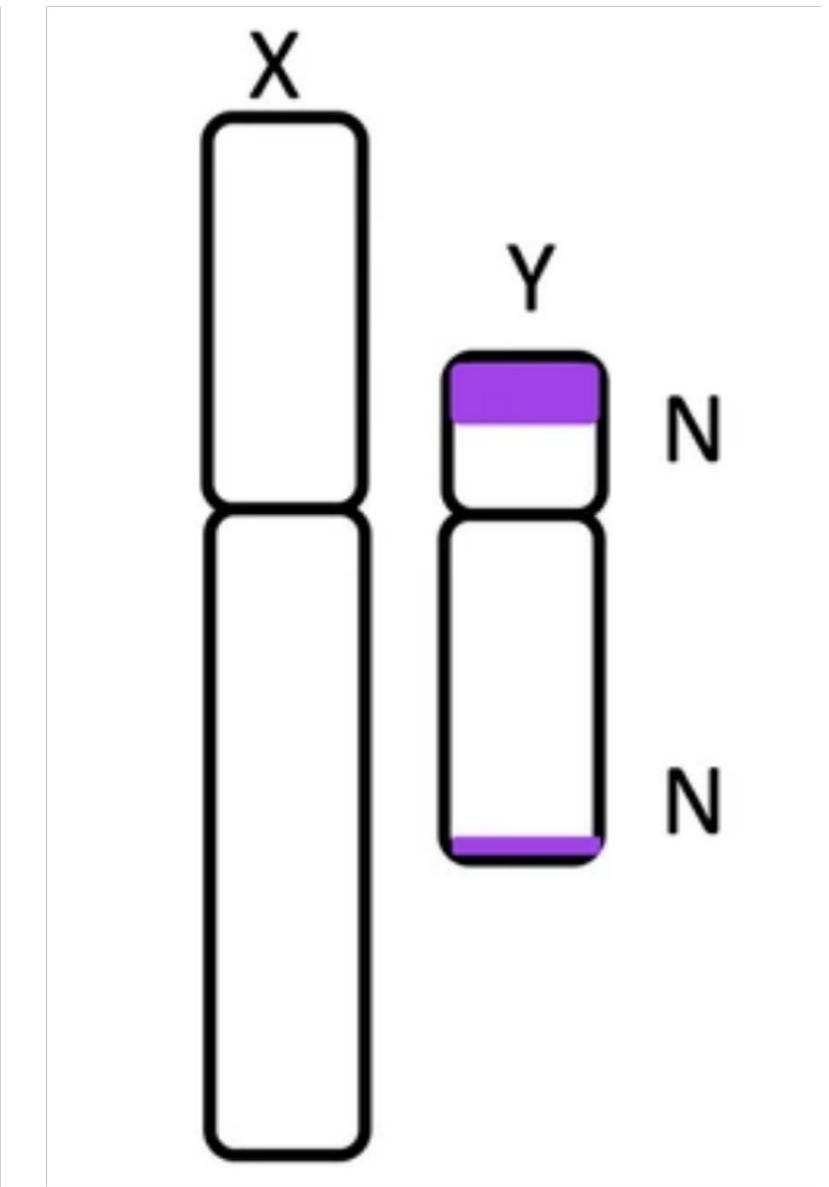
# Karyotype-specific alignment elucidates PAR genotypes



XX Samples



XY Samples



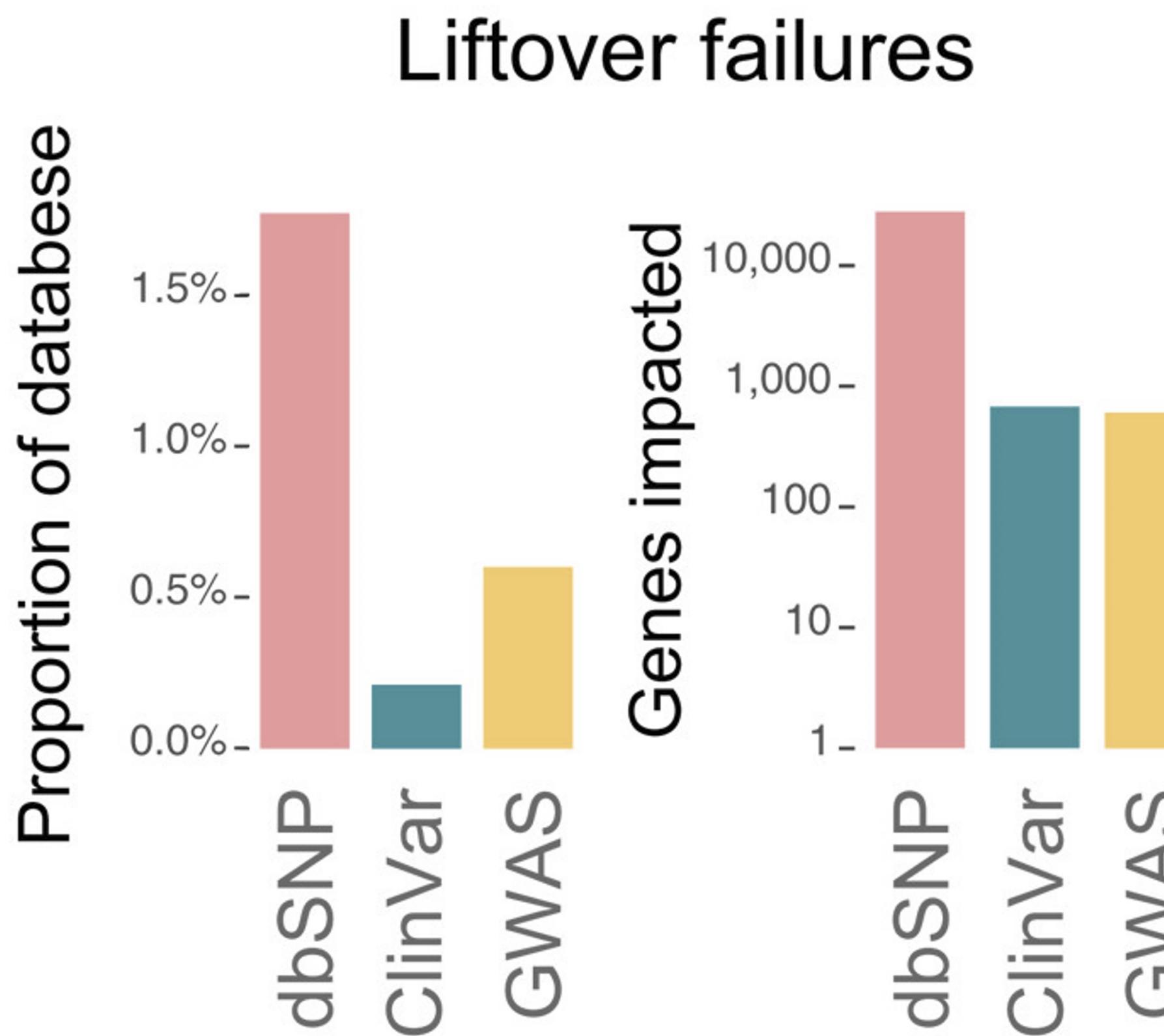
Adapted from Olney, K. C., et al. (2020) *BMC*



**Samantha  
Zarate**

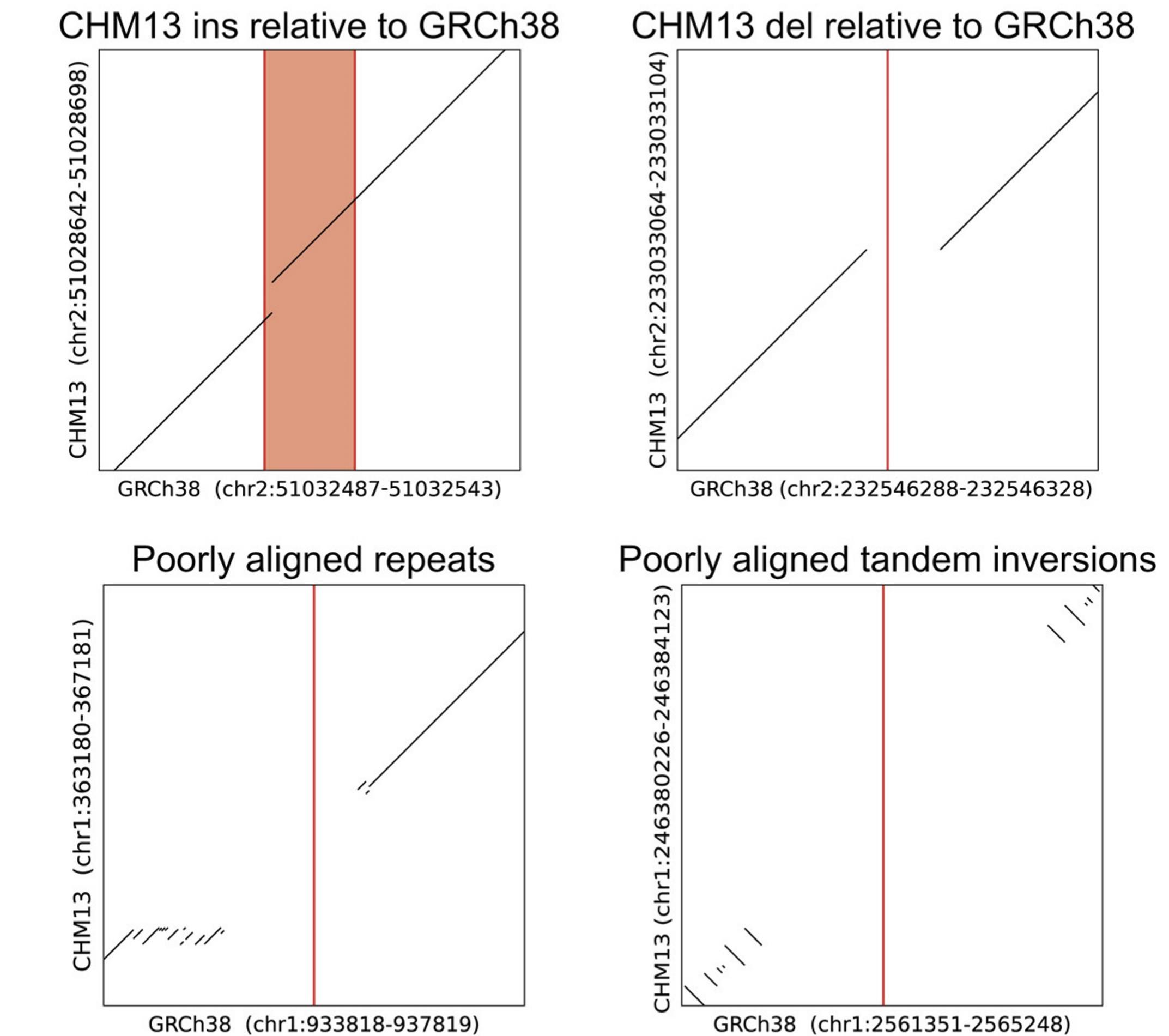
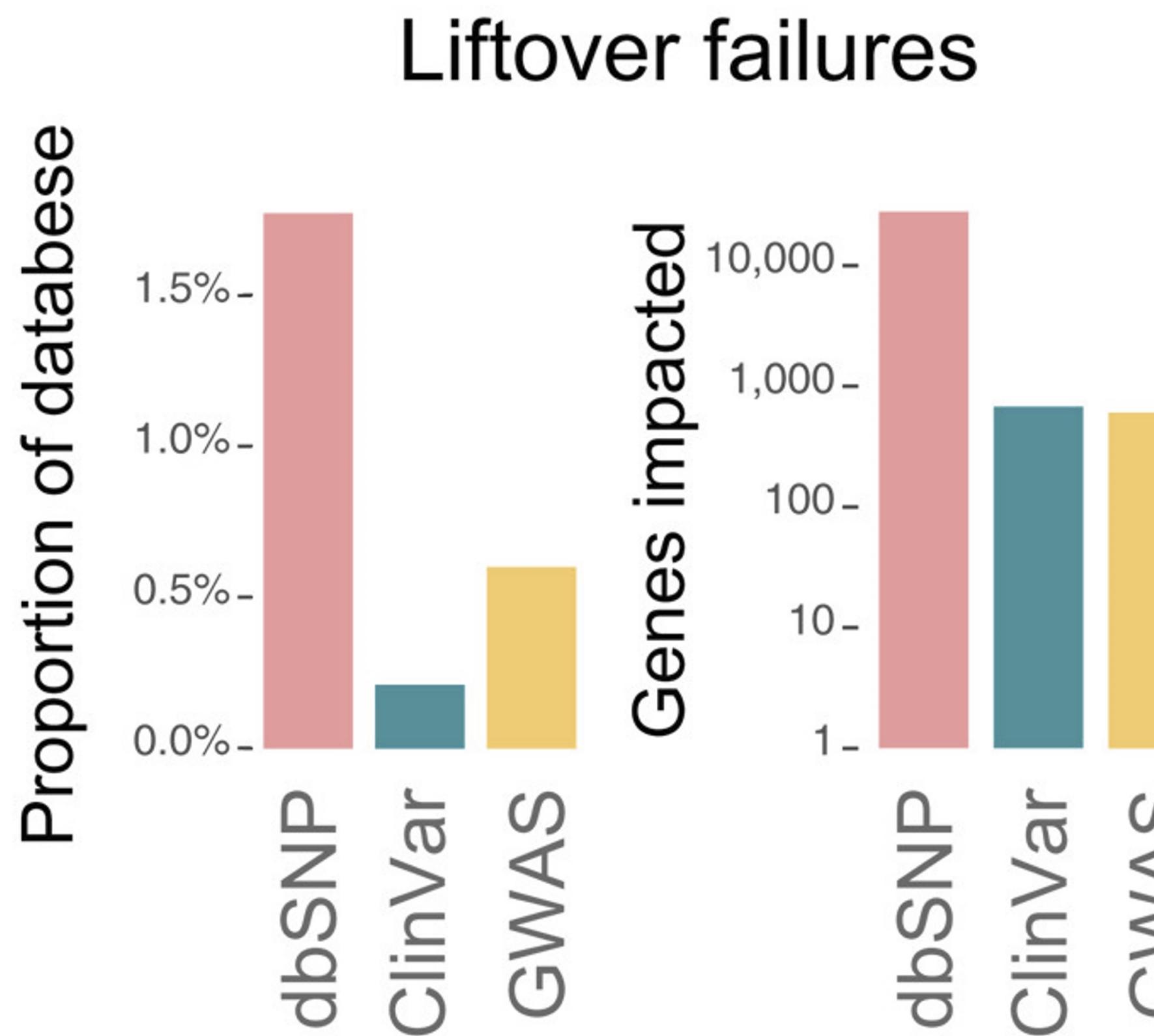
# Liftover of genetic variation databases to T2T-CHM13

- Lifted ClinVar, NCBI dbSNP, and NHGRI-EBI GWAS Catalog to CHM13



# Liftover of genetic variation databases to T2T-CHM13

- Lifted ClinVar, NCBI dbSNP, and NHGRI-EBI GWAS Catalog to CHM13v1.1



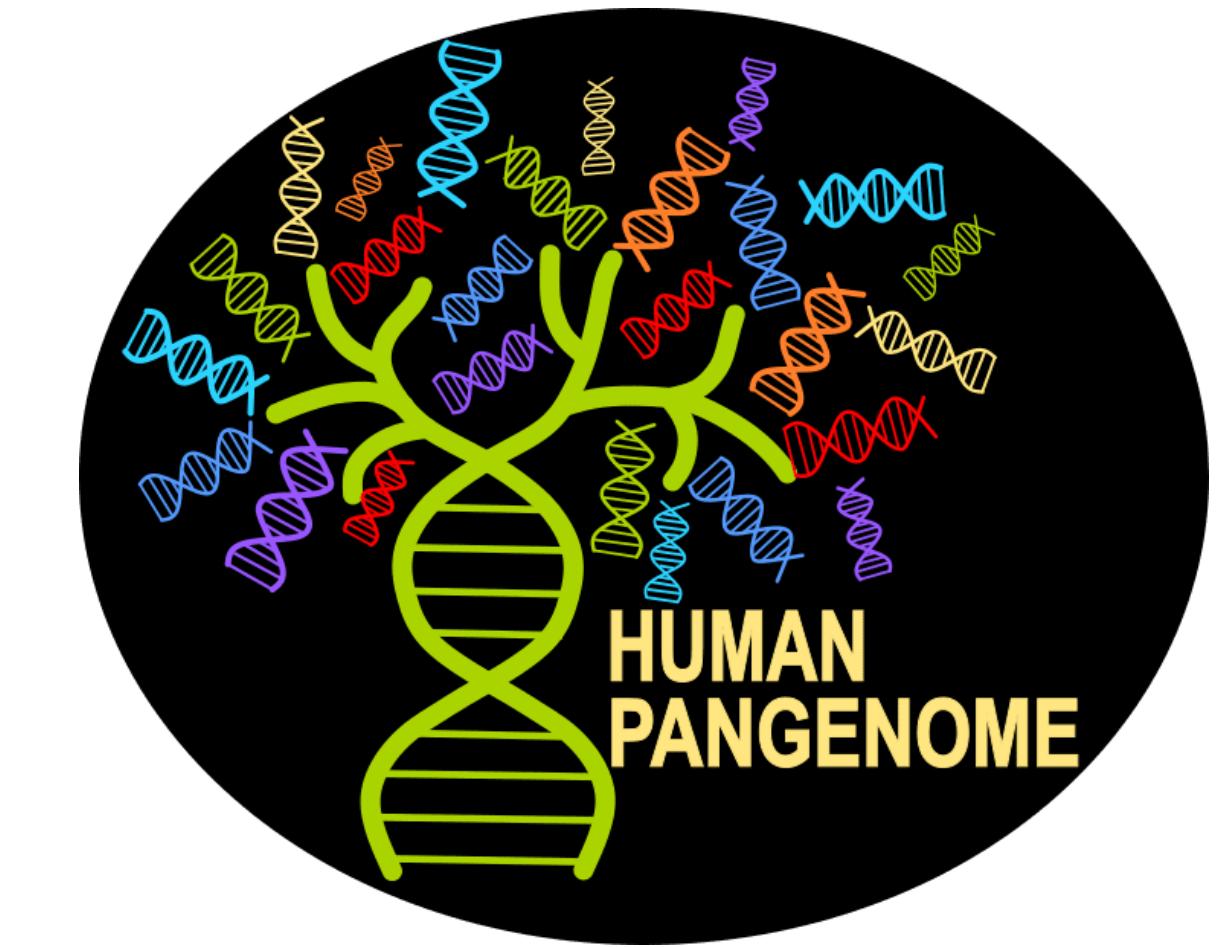
# Liftover of genetic variation databases to HG002Y

- Lifted ClinVar, NCBI dbSNP, and NHGRI-EBI GWAS Catalog from GRCh38 chrY to HG002 chrY

| Database               | GRCh38    | Number of Variants      |                                  |
|------------------------|-----------|-------------------------|----------------------------------|
|                        |           | Lifted to T2T-CHM13v2.0 | Failed liftover to T2T-CHM13v2.0 |
| dbSNP155<br>(chrY)     | 2,480,588 | 2,355,634 (95.0%)       | 124,954 (5.0%)                   |
| Clinvar (chrY)         | 48        | 48 (100%)               | 0                                |
| GWAS Catalog<br>(chrY) | 26        | 26 (100%)               | 0                                |

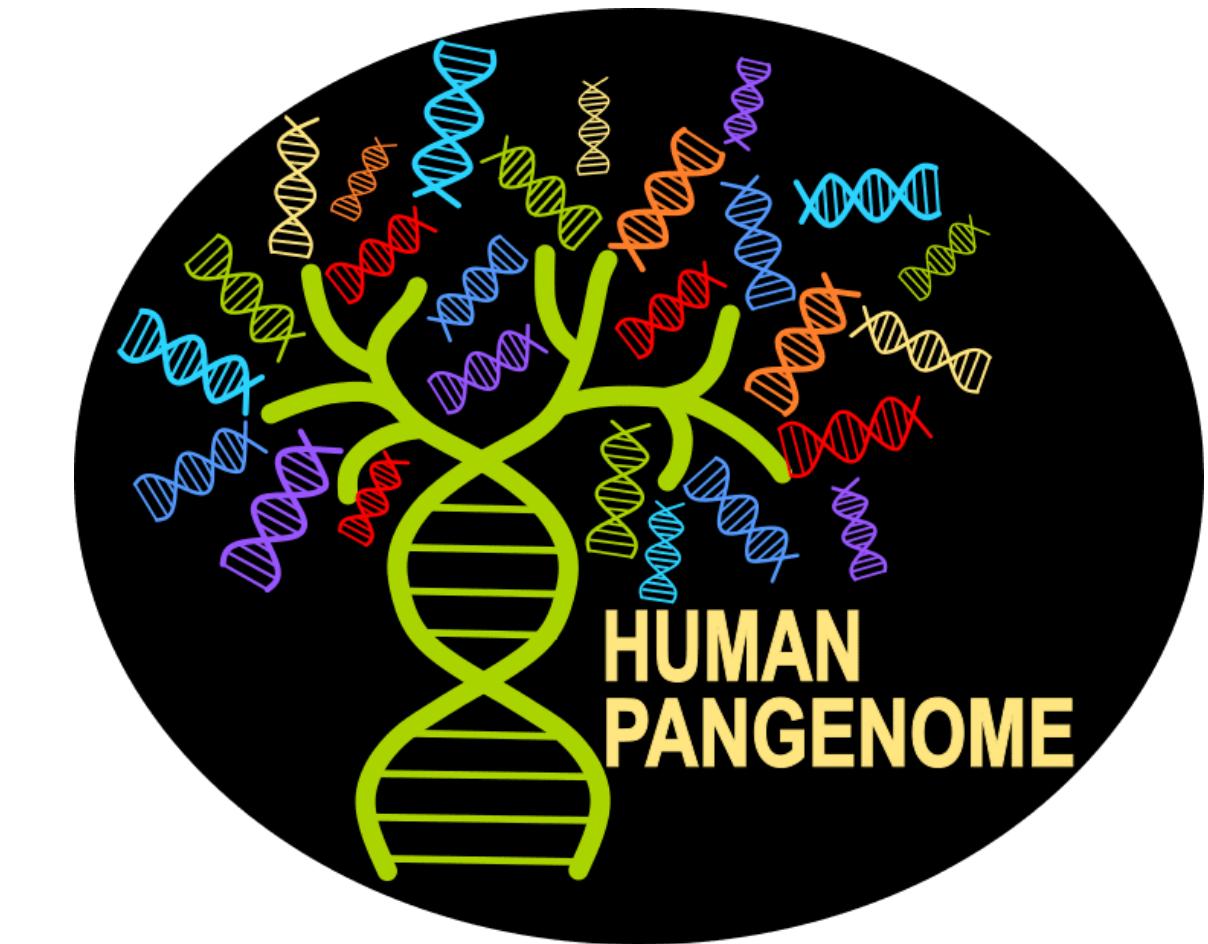
# The future of Telomere-to-telomere assemblies

- Collaboration with the Human Pangenome Reference Consortium (HPRC) – more Telomere-to-Telomere human genomes!
- Generate a high-quality reference capturing the scope of human genetic diversity



# The future of Telomere-to-telomere assemblies

- Collaboration with the Human Pangenome Reference Consortium (HPRC) – more Telomere-to-Telomere human genomes!
- Generate a high-quality reference capturing the scope of human genetic diversity
- Produce more complete genomes for organisms across the tree of life

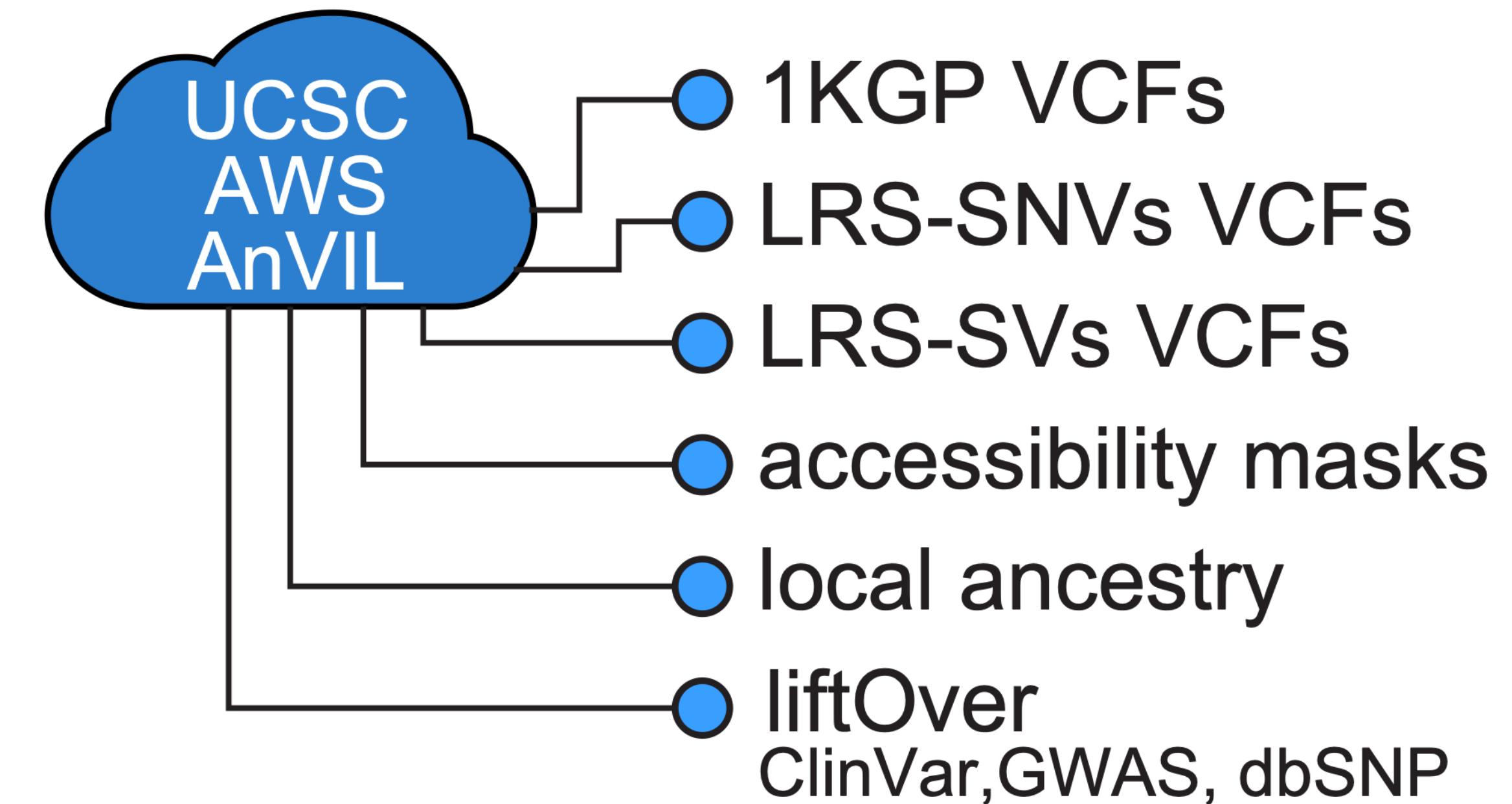


## The genetic and epigenetic landscape of the *Arabidopsis* centromeres

Matthew Naish†, Michael Alonge†, Piotr Włodzimierz†, Andrew J. Tock, Bradley W. Abramson, Anna Schmücker, Terezie Mandáková, Bhagyshree Jamge, Christophe Lambing, Pallas Kuo, Natasha Yelina, Nolan Hartwick, Kelly Colt, Lisa M. Smith, Jurriaan Ton, Tetsuji Kakutani, Robert A. Martienssen, Korbinian Schneeberger, Martin A. Lysak, Frédéric Berger, Alexandros Bousios, Todd P. Michael, Michael C. Schatz\*, Ian R. Henderson\*

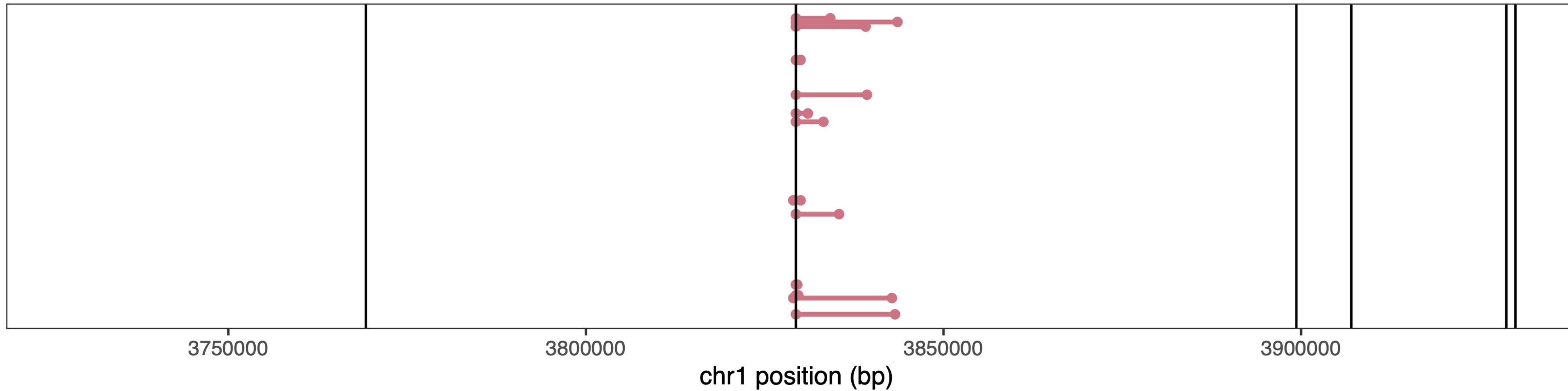
# Conclusion

- T2T-CHM13 improves alignment and variant calling analysis across all populations
- Newly resolved regions of the genome reveal 230 Mbp of sequence for analysis for the first time
- A complete telomere-to-telomere assembly of the Y chromosome from HG002
- T2T has generated resources for using T2T-CHM13v2.0 as a reference genome

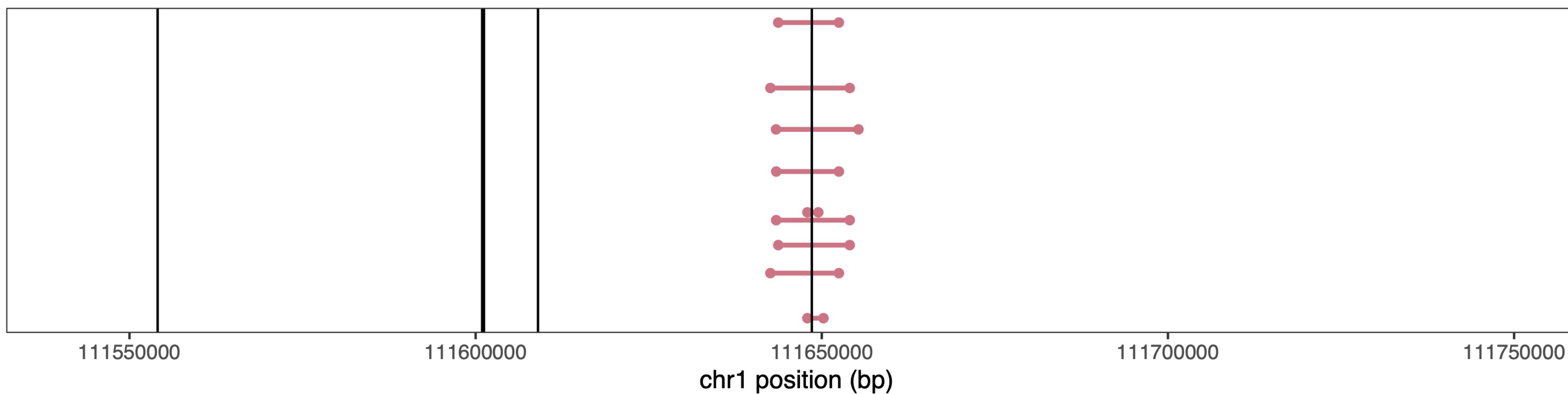


**extra slides**

# GRCh38 contains artificial haplotype boundaries

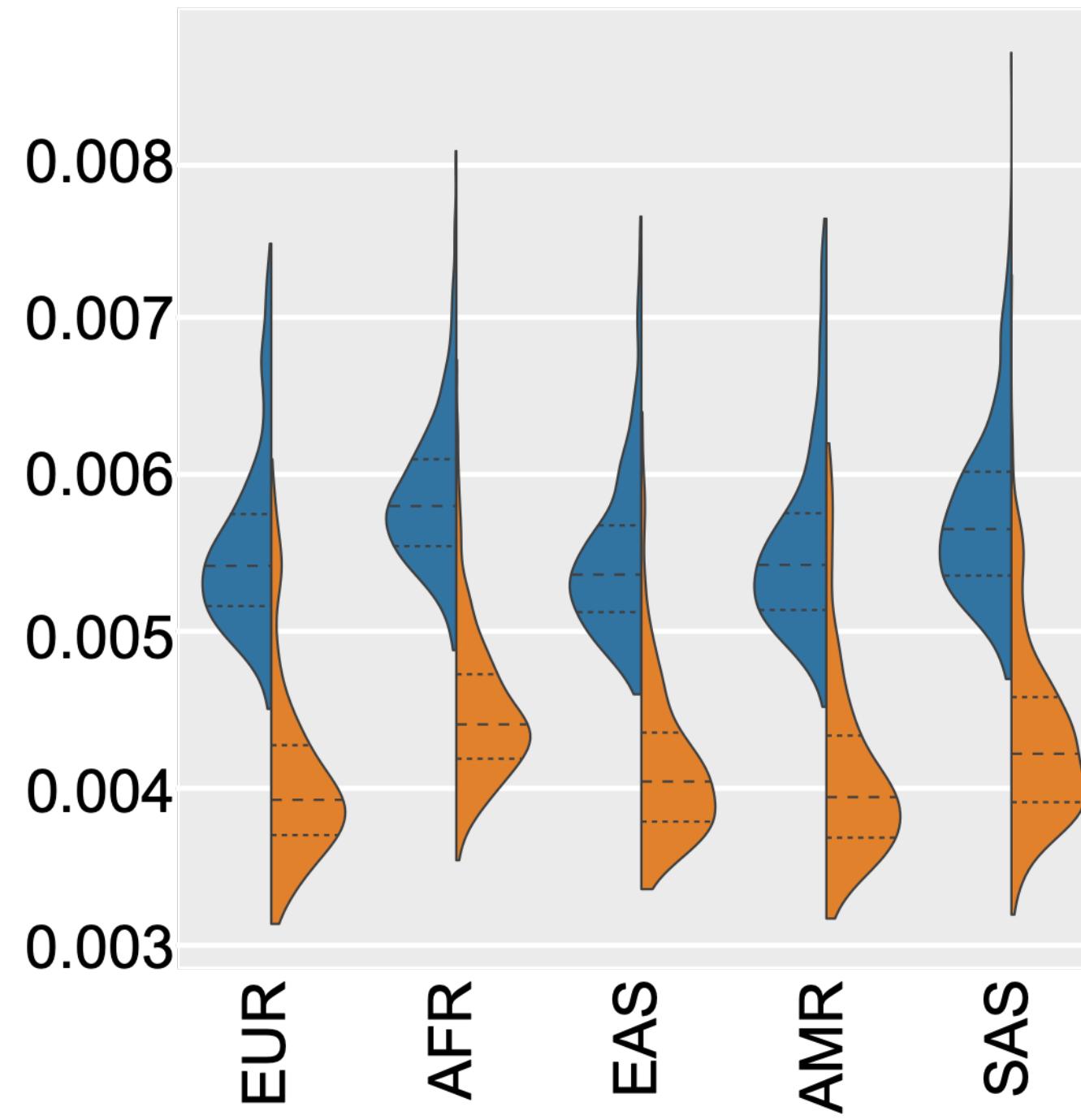


# Rajiv McCoy

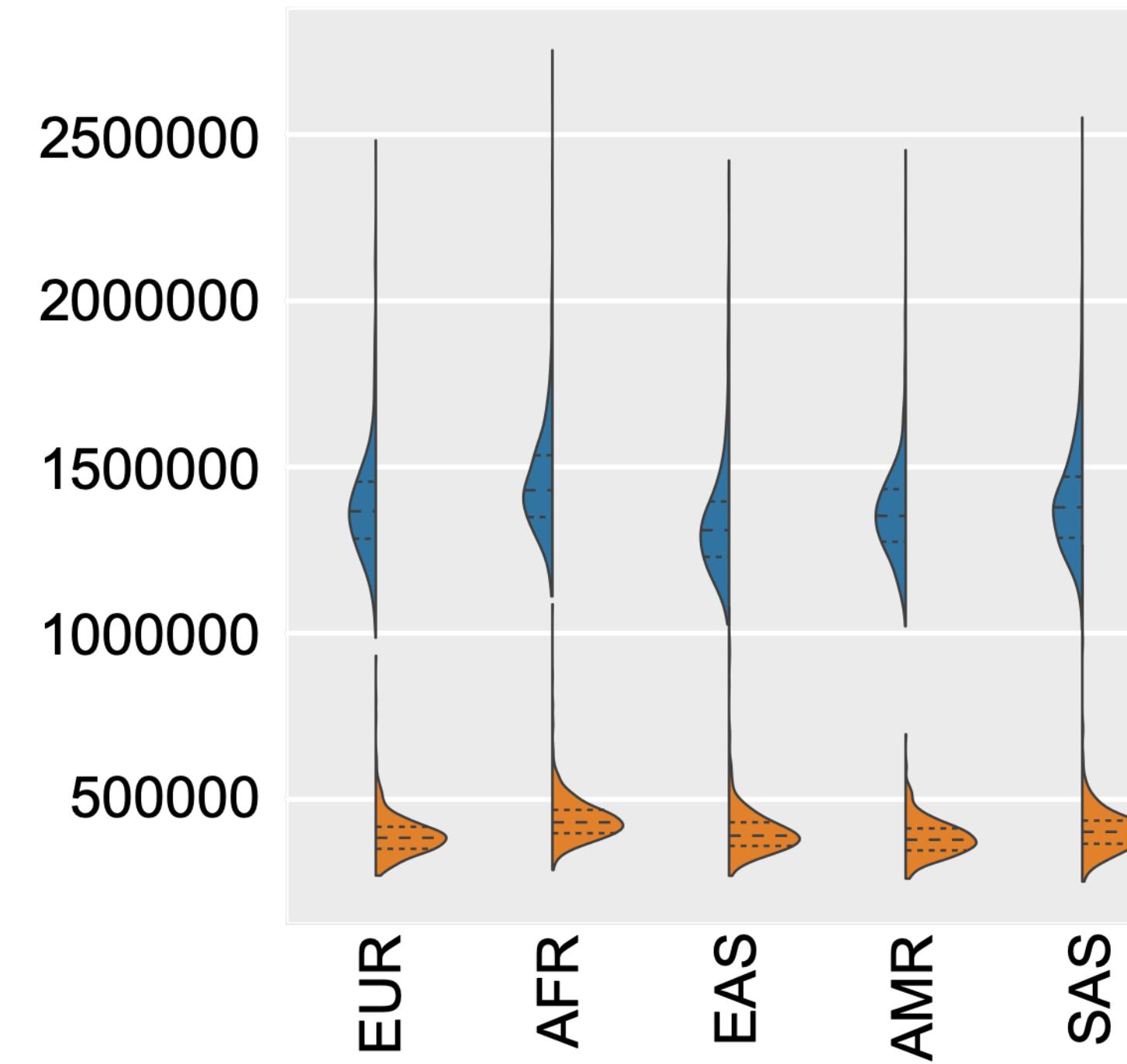


# T2T-CHM13 improves short-read alignment

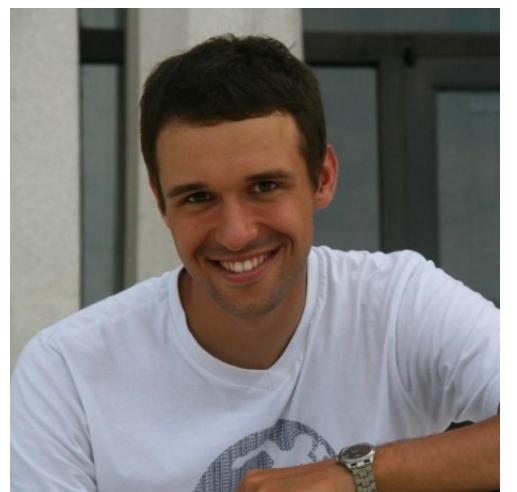
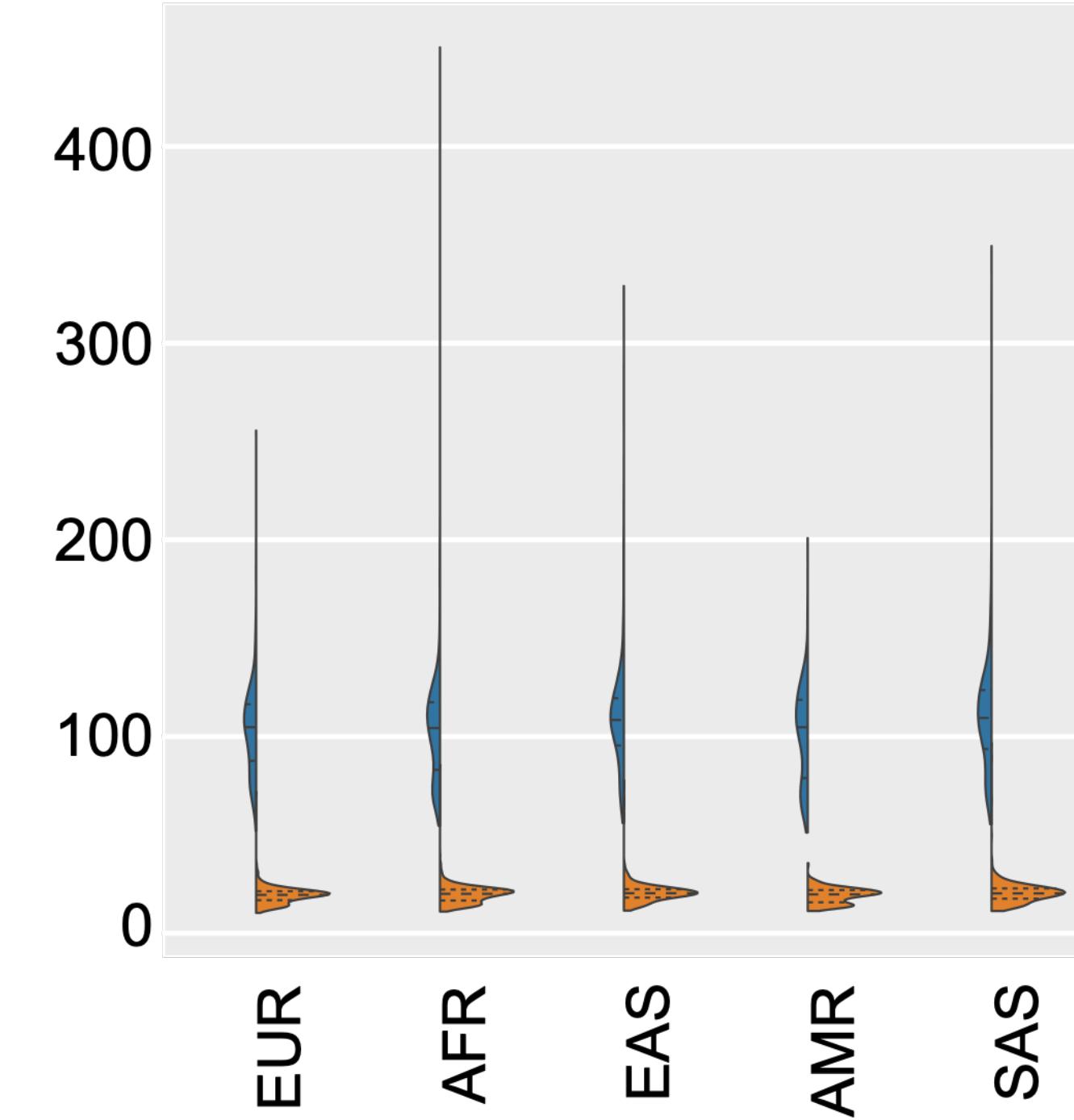
Mismatch rate



Invalid pairs



Cov. std. within genes



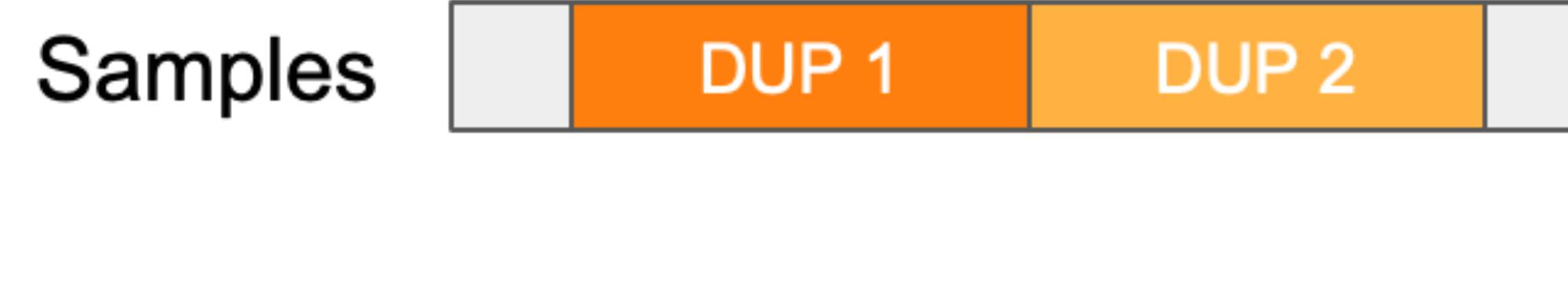
Sergey  
Aganezov

GRCh38

CHM13

# T2T-CHM13 improves short-read variant calling

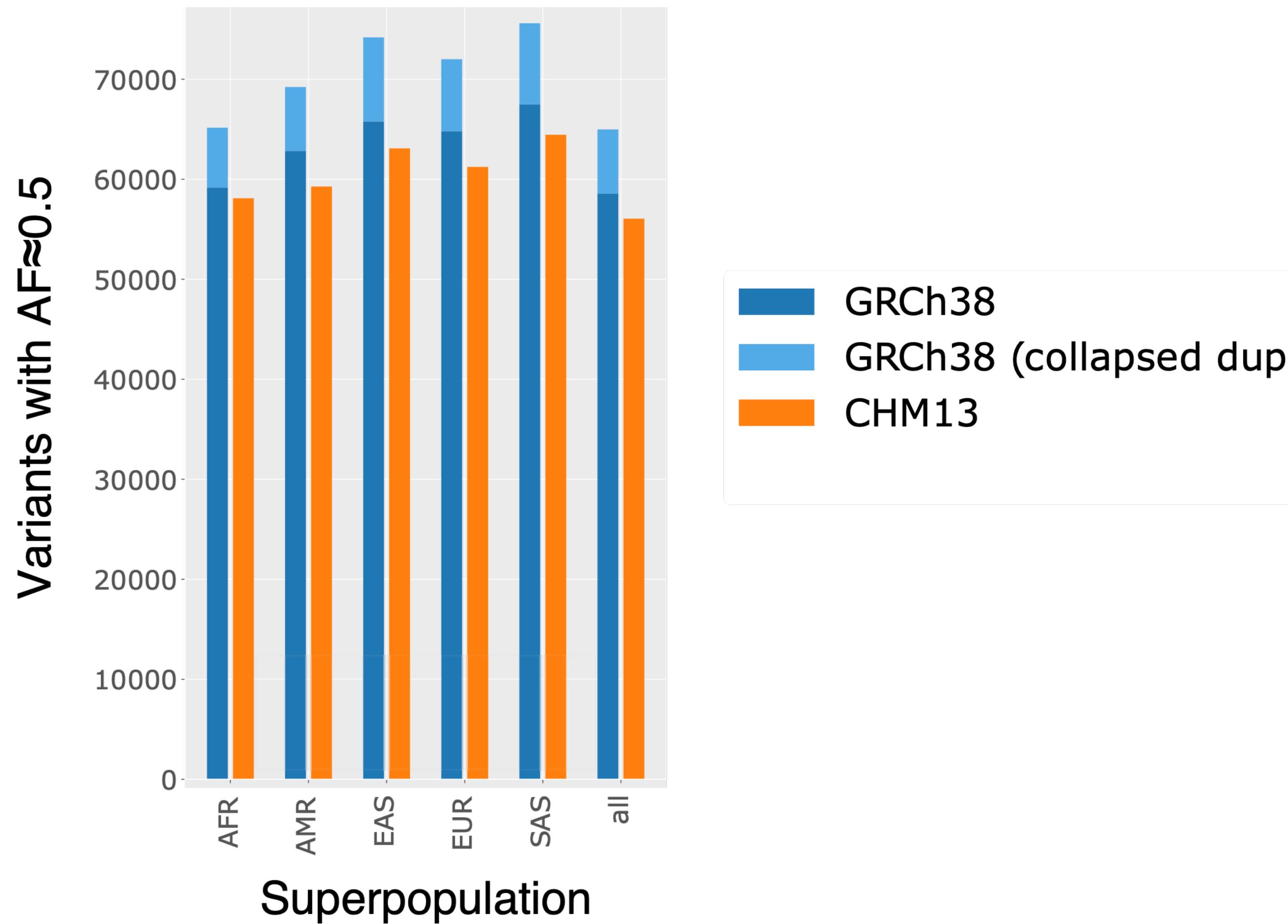
- Fewer false heterozygous variants from GRCh38 collapsed duplications



**Samantha  
Zarate**

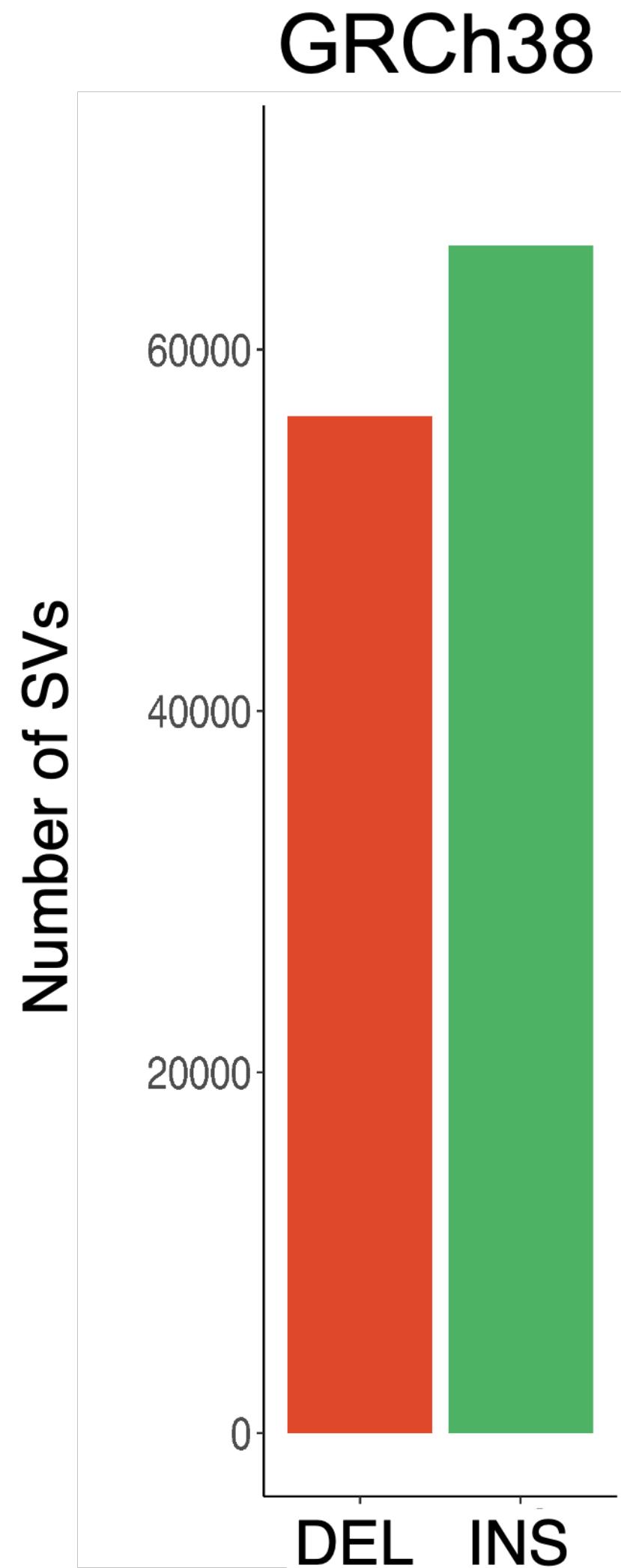
# T2T-CHM13 improves short-read variant calling

- Fewer spurious heterozygous variants from GRCh38 collapsed duplications



**Samantha  
Zarate**

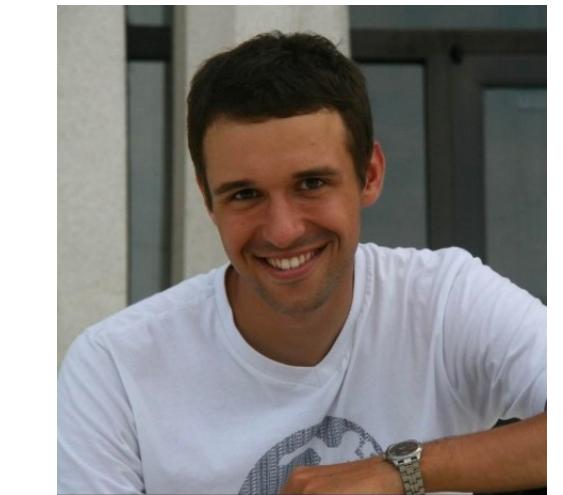
# T2T-CHM13 fixes the insertion-deletion misbalance in GRCh38



- Excess GRCh38 insertions represent missing or incomplete sequences

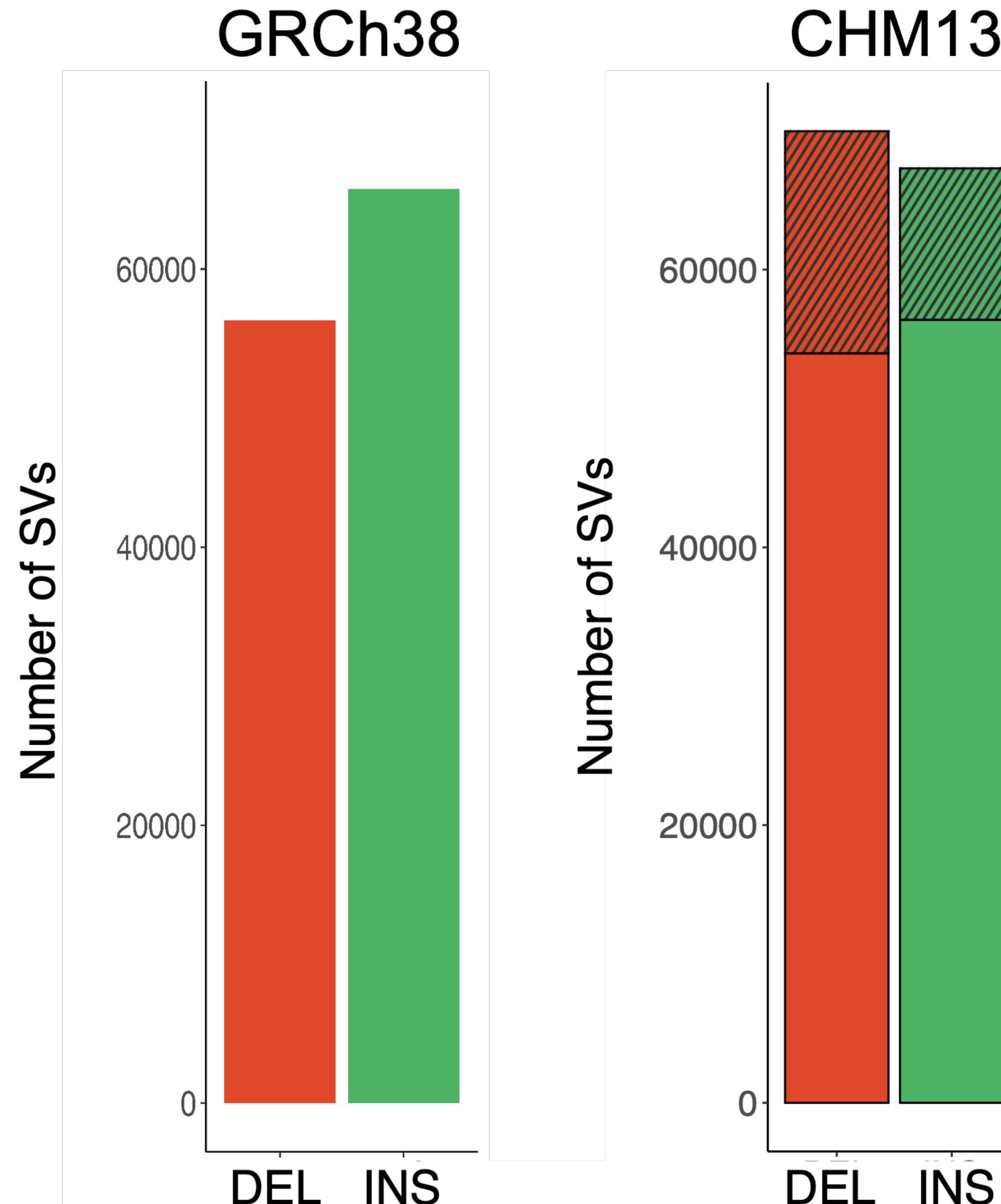


Melanie  
Kirsche

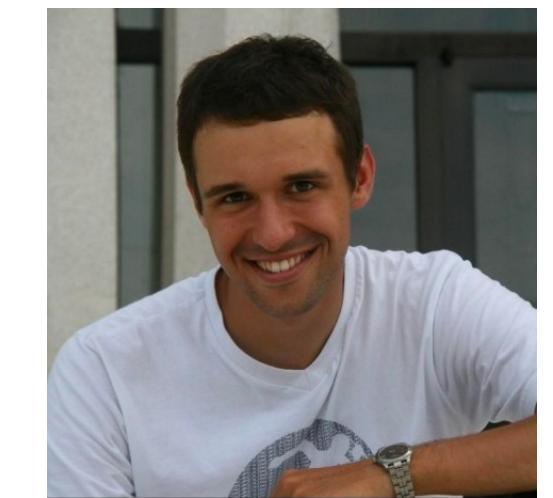


Sergey  
Aganezov

# T2T-CHM13 fixes the insertion-deletion misbalance in GRCh38

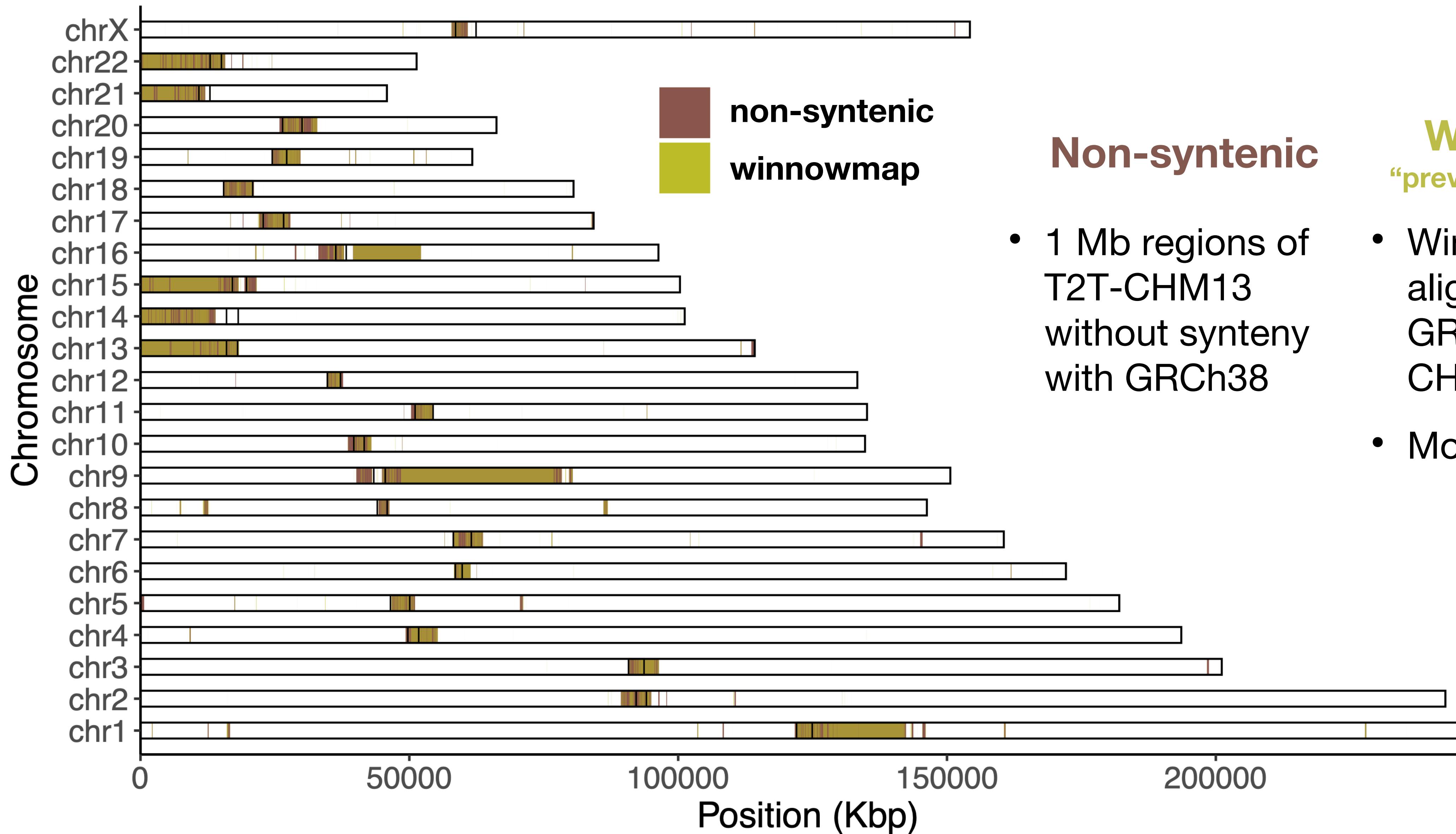


Melanie  
Kirsche



Sergey  
Aganezov

# Identifying new sequences revealed by T2T-CHM13



**Winnowmap**  
“previously unresolved”

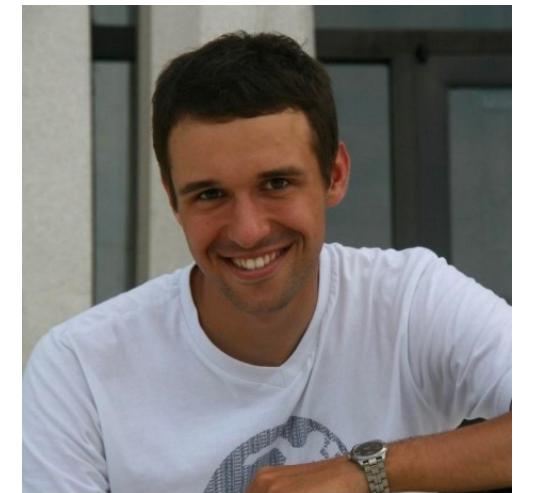
- Winnowmap alignments of GRCh38 to T2T-CHM13
- More strict

# Long-read alignment and variant calling with T2T-CHM13

- 17 diverse samples from the Human Pangenome Reference Consortium and Genome in a Bottle
- PacBio HiFi data + 14 samples with ONT data

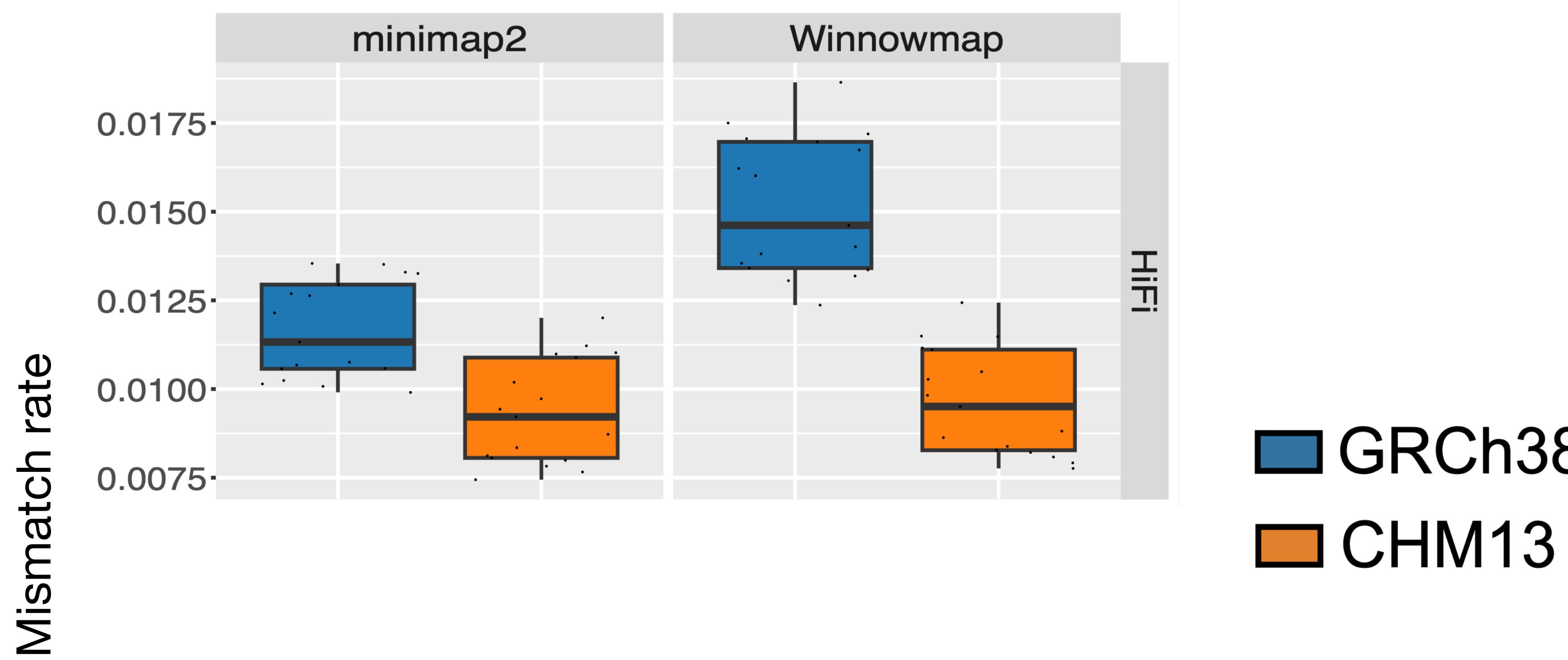


Melanie  
Kirsche

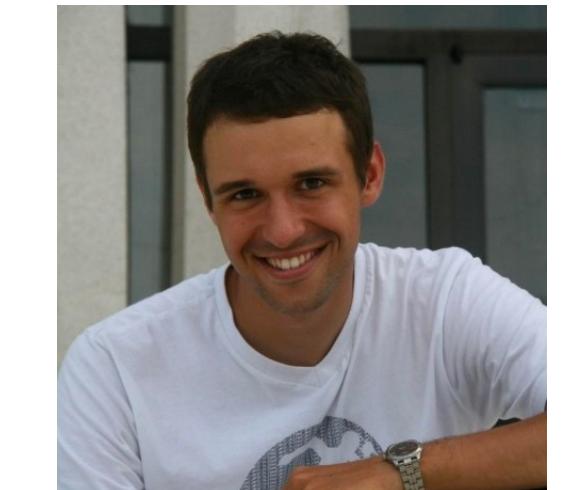


Sergey  
Aganezov

# T2T-CHM13 lowers alignment mismatch rates for long reads

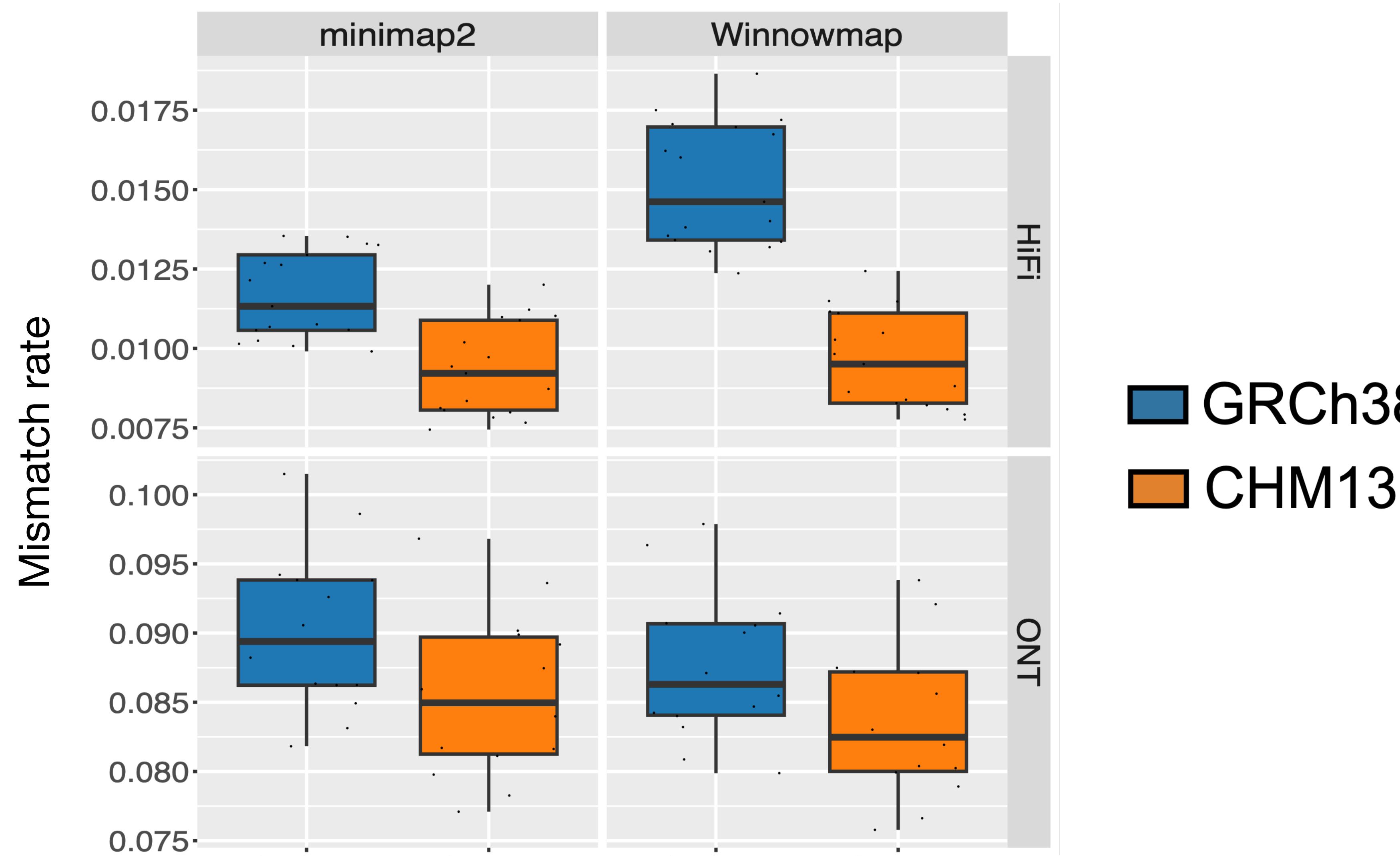


Melanie  
Kirsche

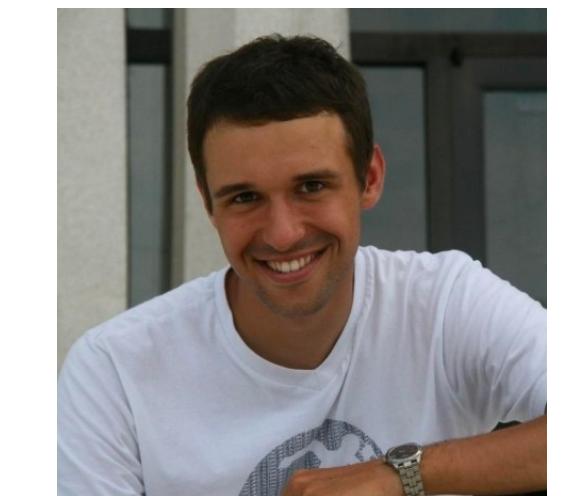


Sergey  
Aganezov

# T2T-CHM13 lowers alignment mismatch rates for long reads



Melanie  
Kirsche



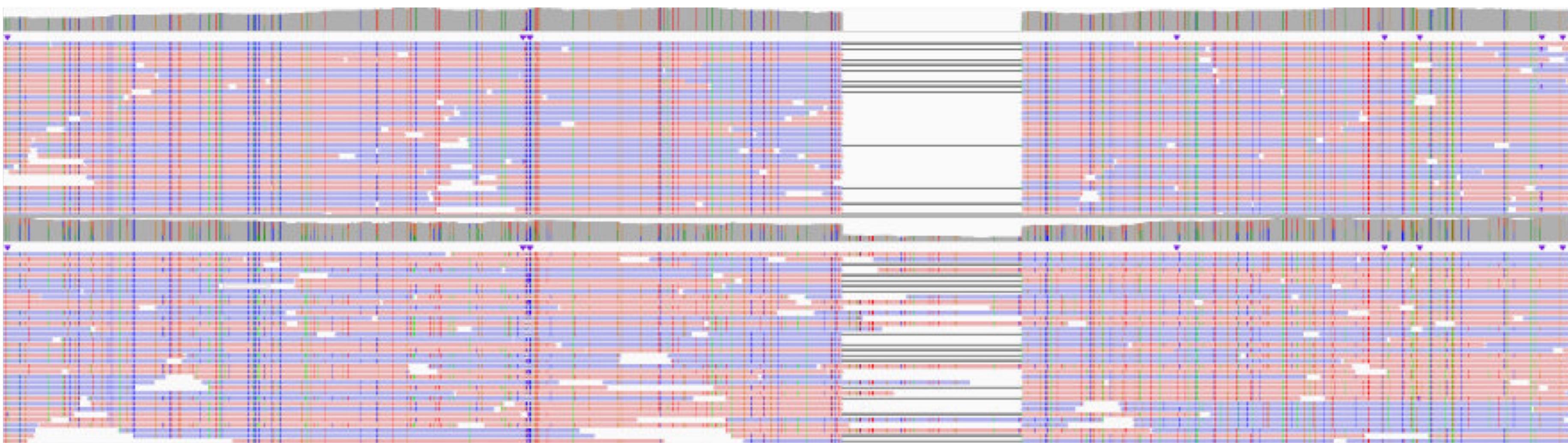
Sergey  
Aganezov

# T2T-CHM13 improves resolution of structural variants

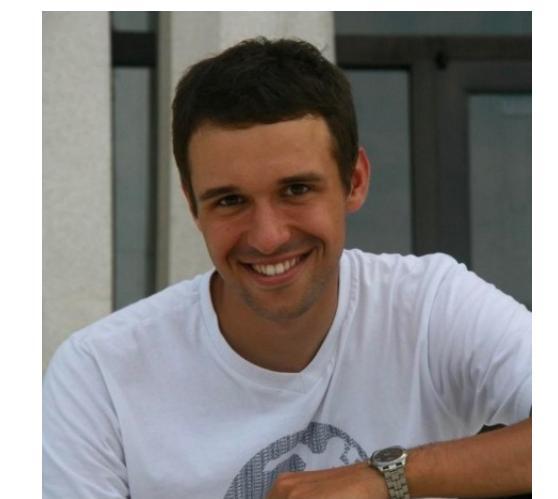
GRCh38



CHM13



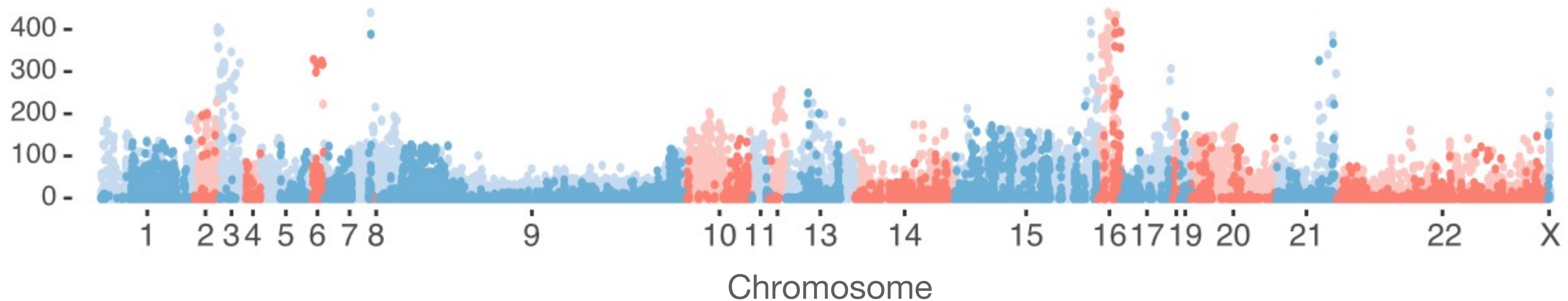
Melanie  
Kirsche



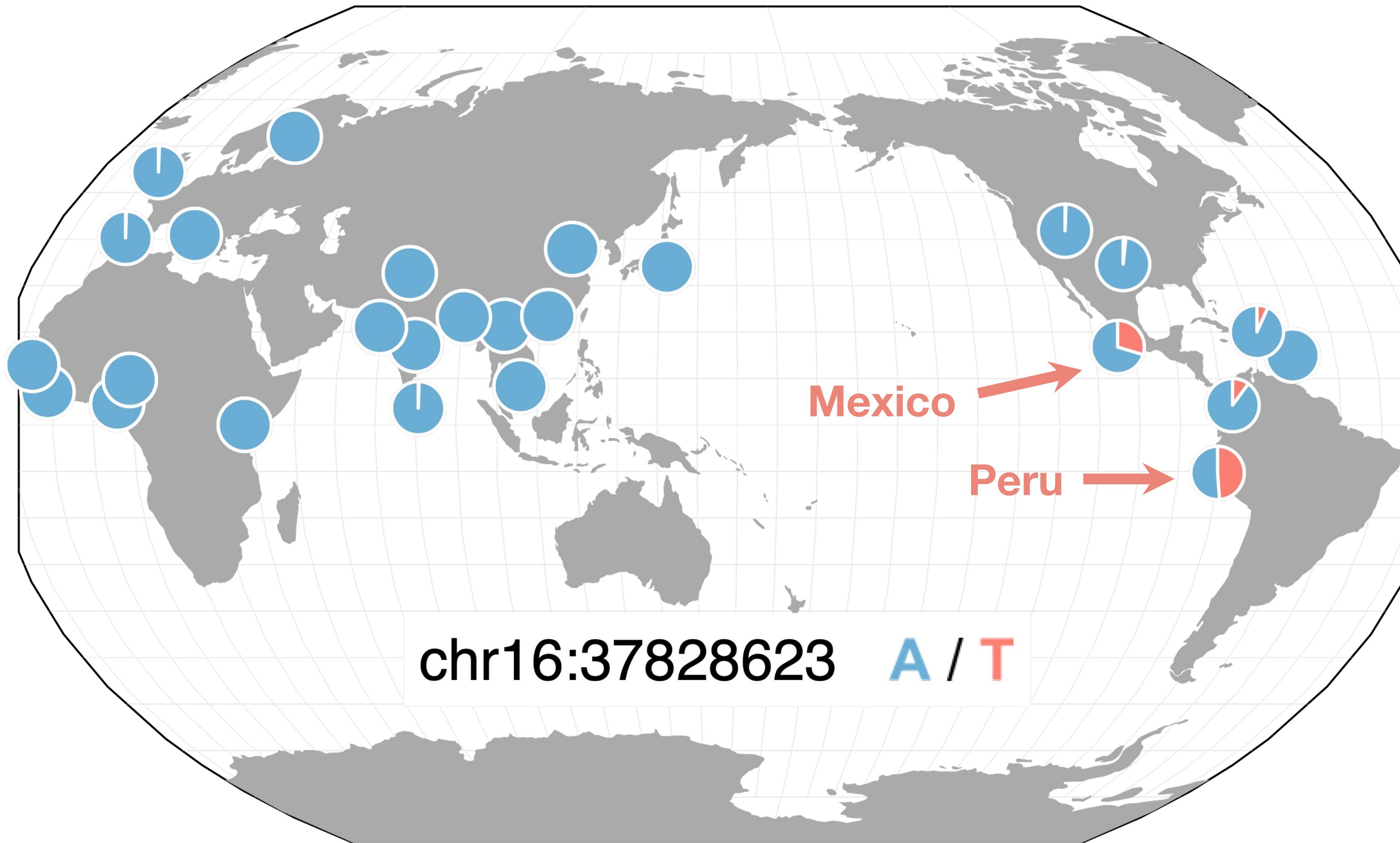
Sergey  
Aganezov

# Evolutionary signatures in novel regions of T2T-CHM13

- Search for novel variants with strong allele frequency differences between populations
- Variants with extreme AF differences are enriched for targets of selection



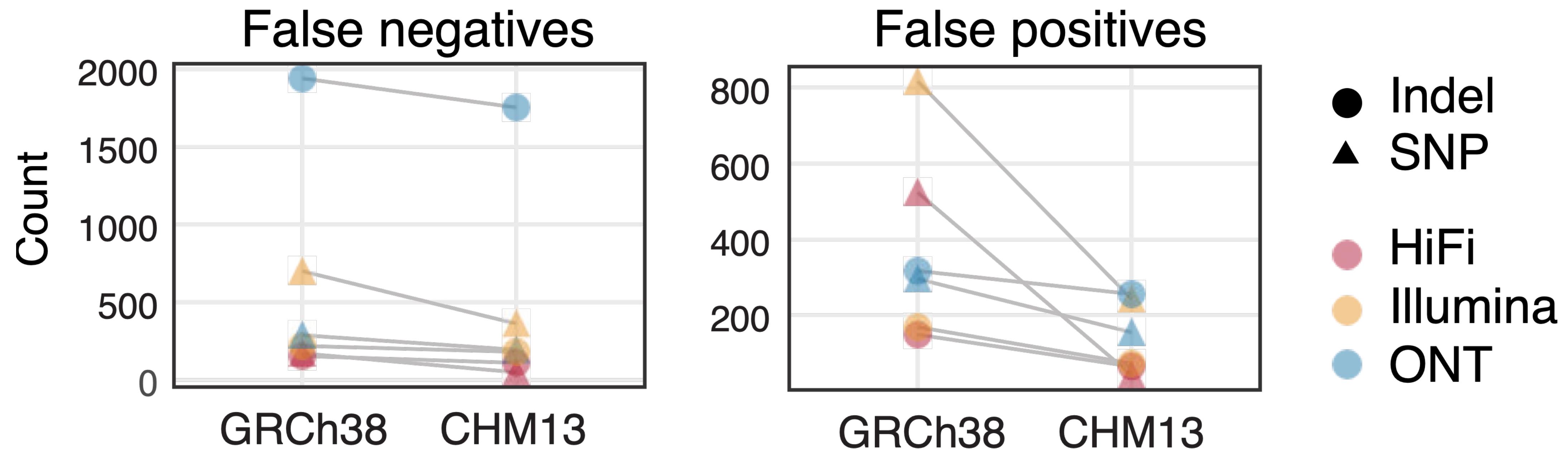
# Evolutionary signatures in novel regions of T2T-CHM13



Located in a  
centromeric  
alpha satellite  
on chr16

# T2T-CHM13 improves clinical genomics variant calling

- 273 challenging, medically relevant genes
- Benchmarked with sequencing data from HG002



Daniela  
Soto



Megan  
Dennis



Justin  
Zook



Fritz  
Sedlazeck



Danny  
Miller