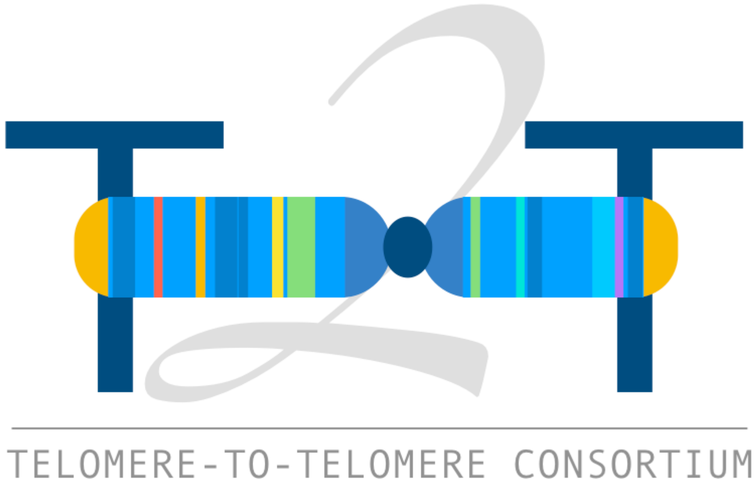
The T2T Genome and applications to variant calling

Dylan Taylor
Johns Hopkins University
November 18, 2022



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

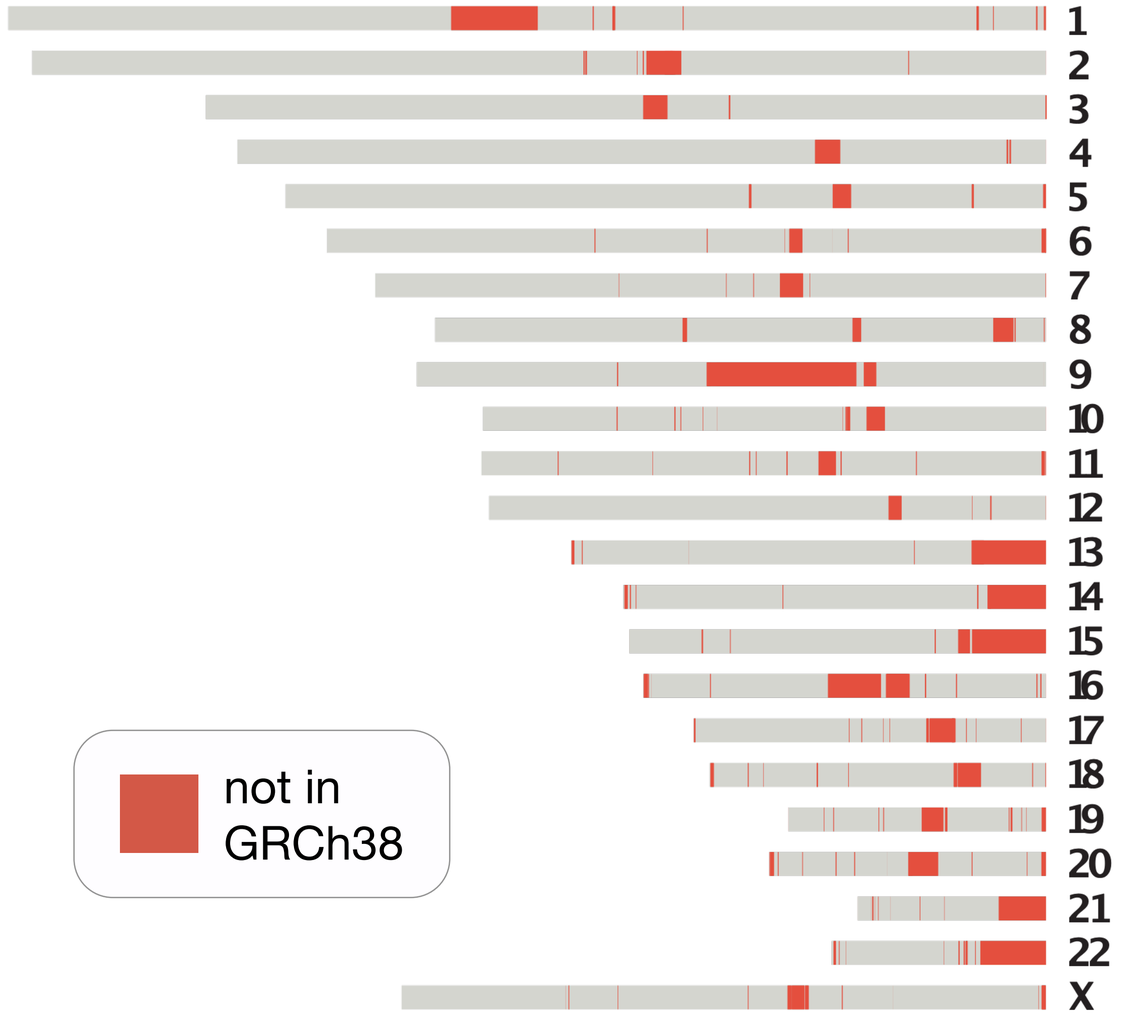
- What is a T2T Genome?
- Benchmarking on a diverse dataset
- Large-scale cloud analysis of short-read data
- Improvements to alignment/variant calling
- A T2T Y chromosome
- Future directions

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

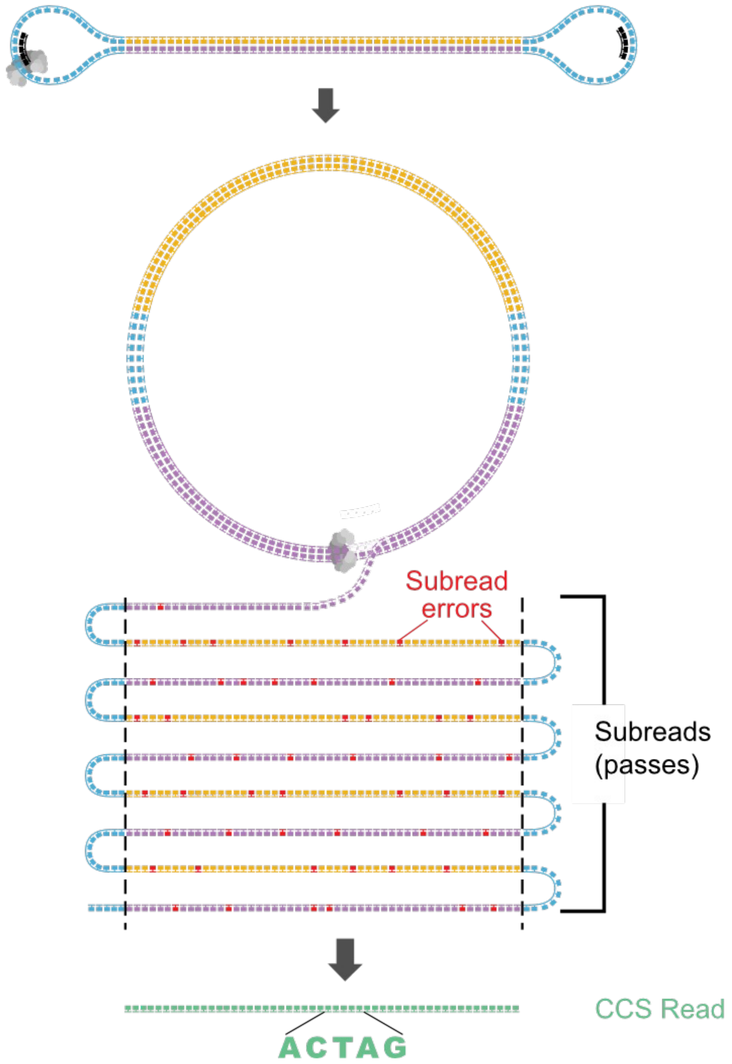
The “previous” human reference genome is incomplete

- 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



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

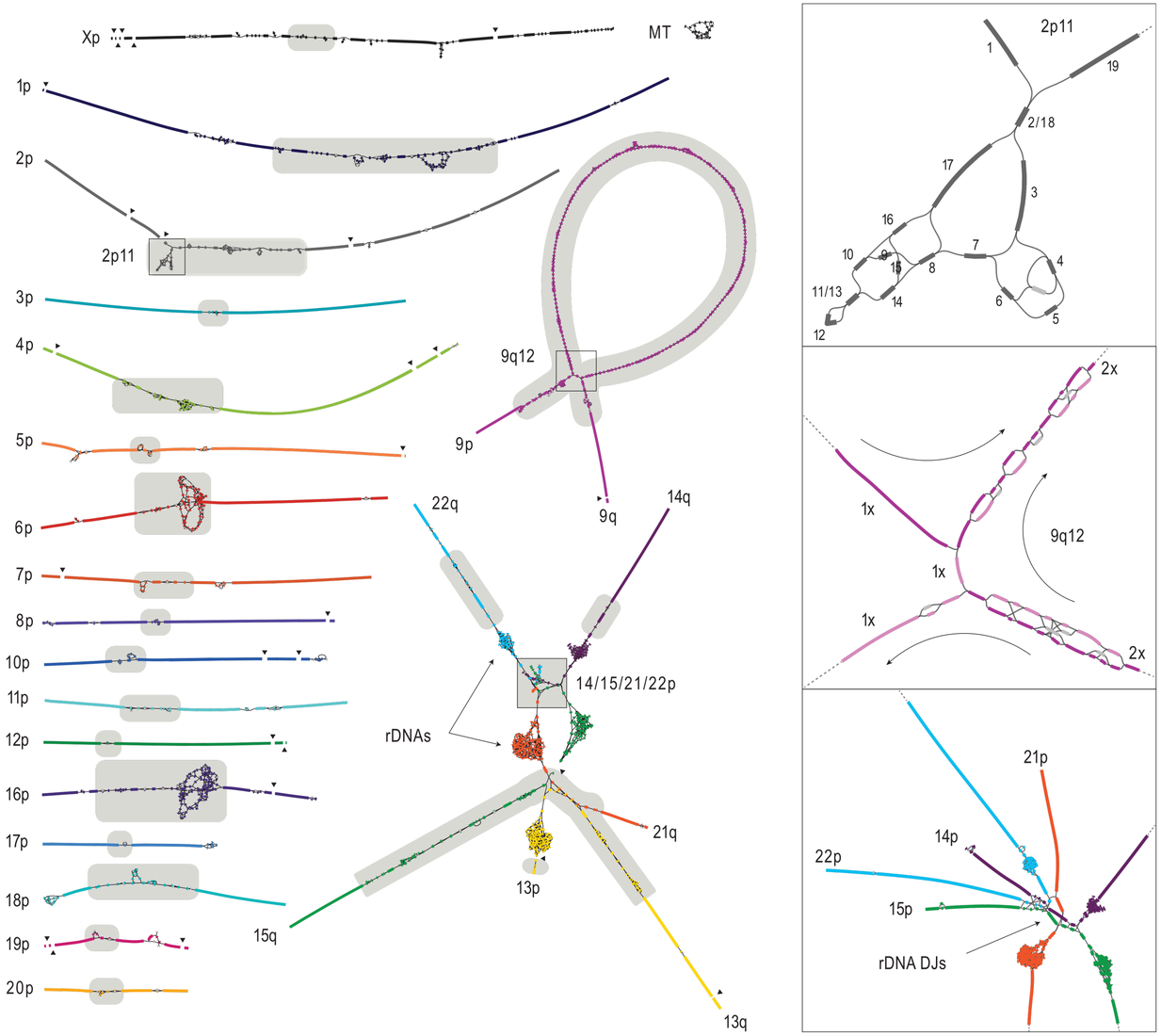
Assembling a complete Telomere-to-telomere genome



- Begin with highly accurate PacBio HiFi reads

Wenger *et al*., 2019, *Nat*. *Biotechnol.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.

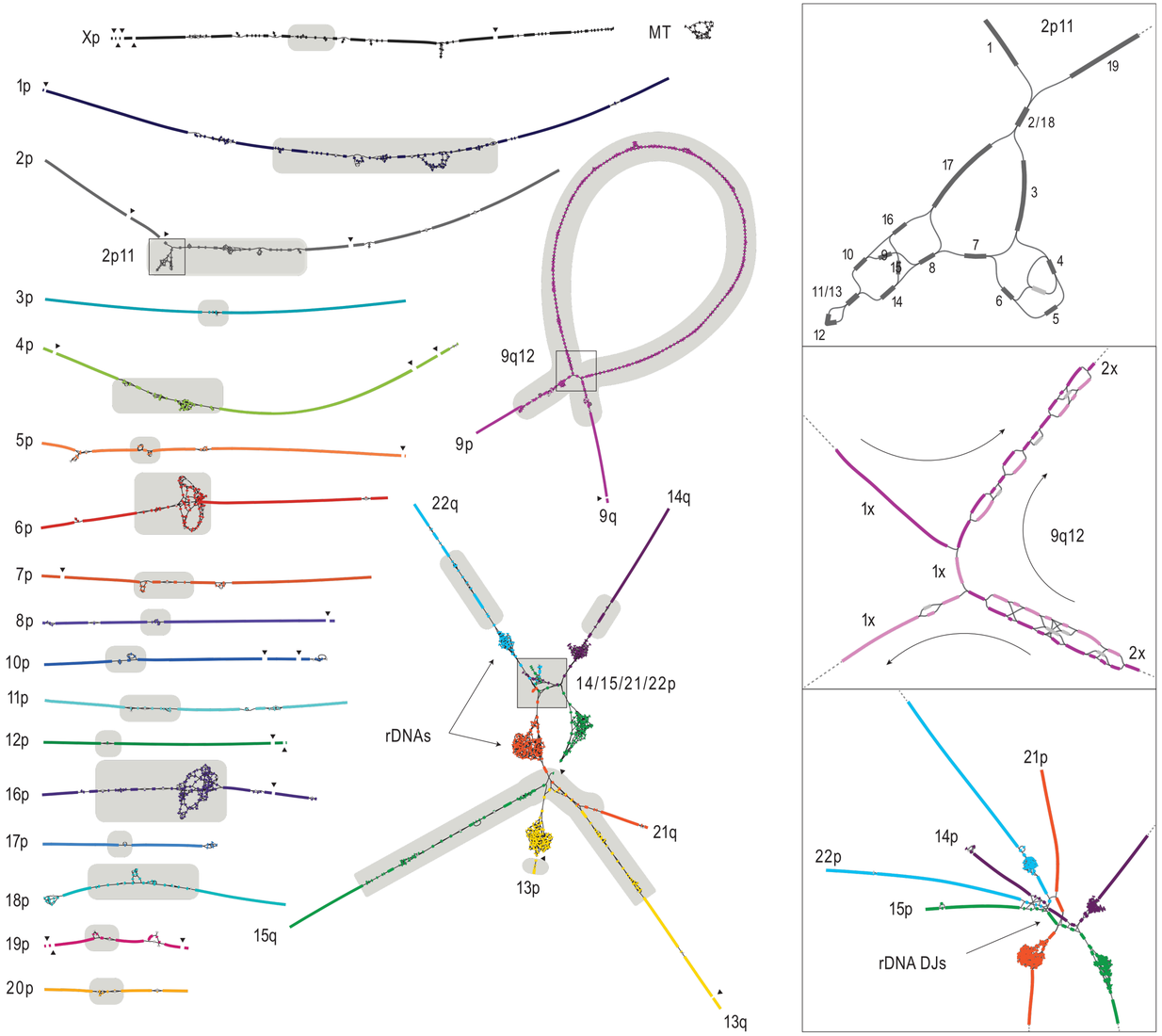
Assembling a complete Telomere-to-telomere genome



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

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

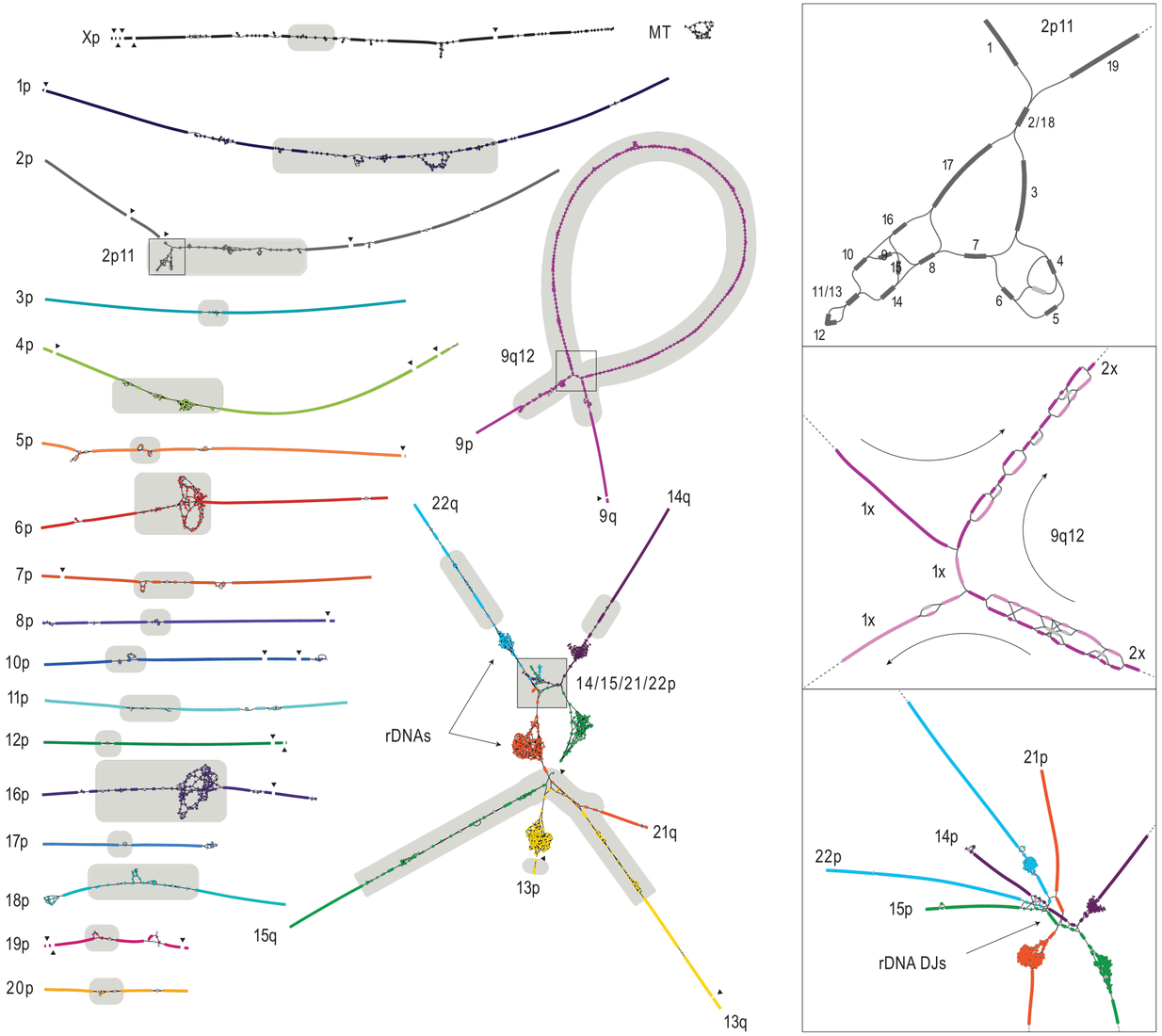
Assembling a complete Telomere-to-telomere genome



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

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

Assembling a complete Telomere-to-telomere genome

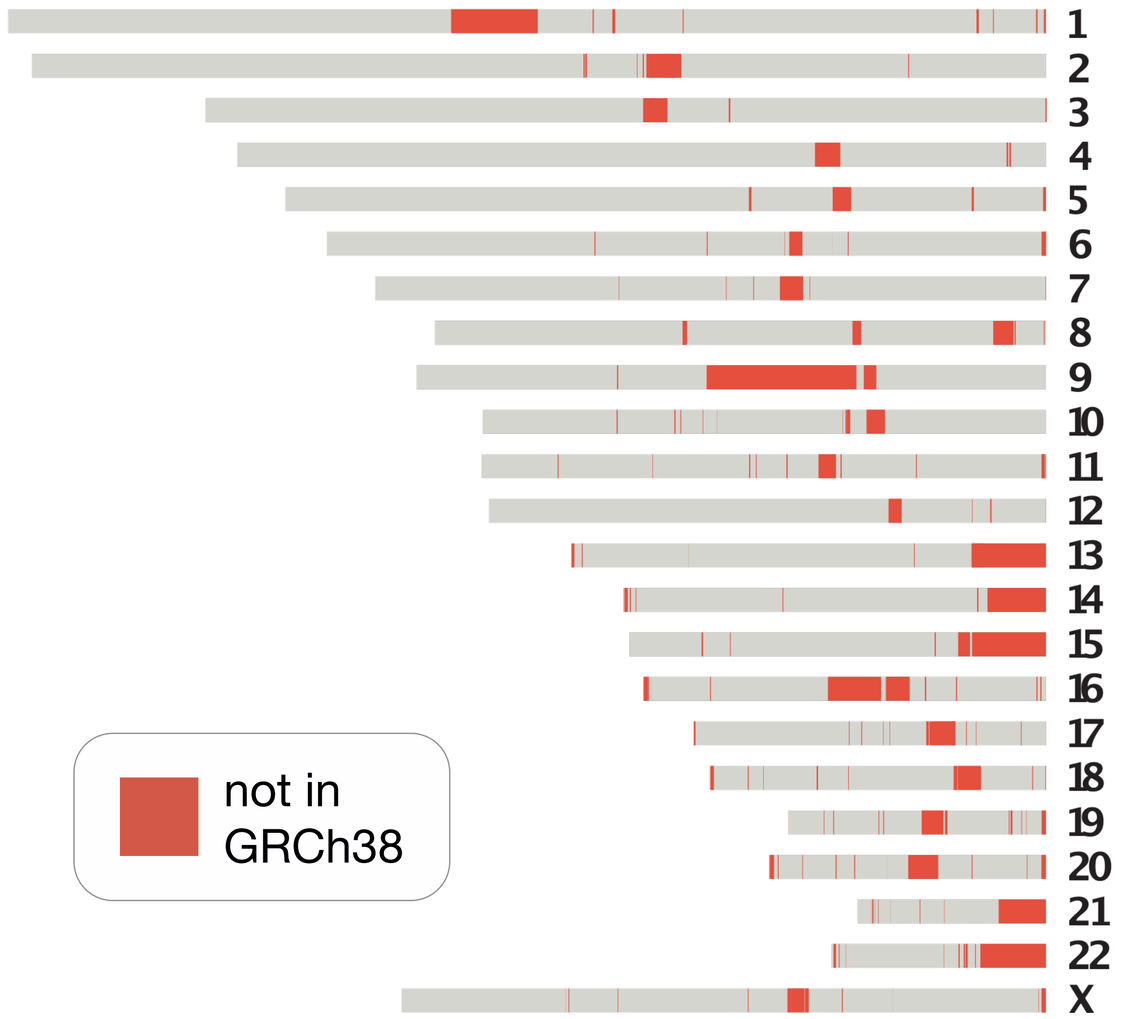


- Begin with highly accurate PacBio HiFi reads
- Using these, build high-resolution assembly string graph
- Most chromosome assemblies are nearly linear
- Ultralong ONT reads + read coverage used to resolve loops in the string graph

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

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

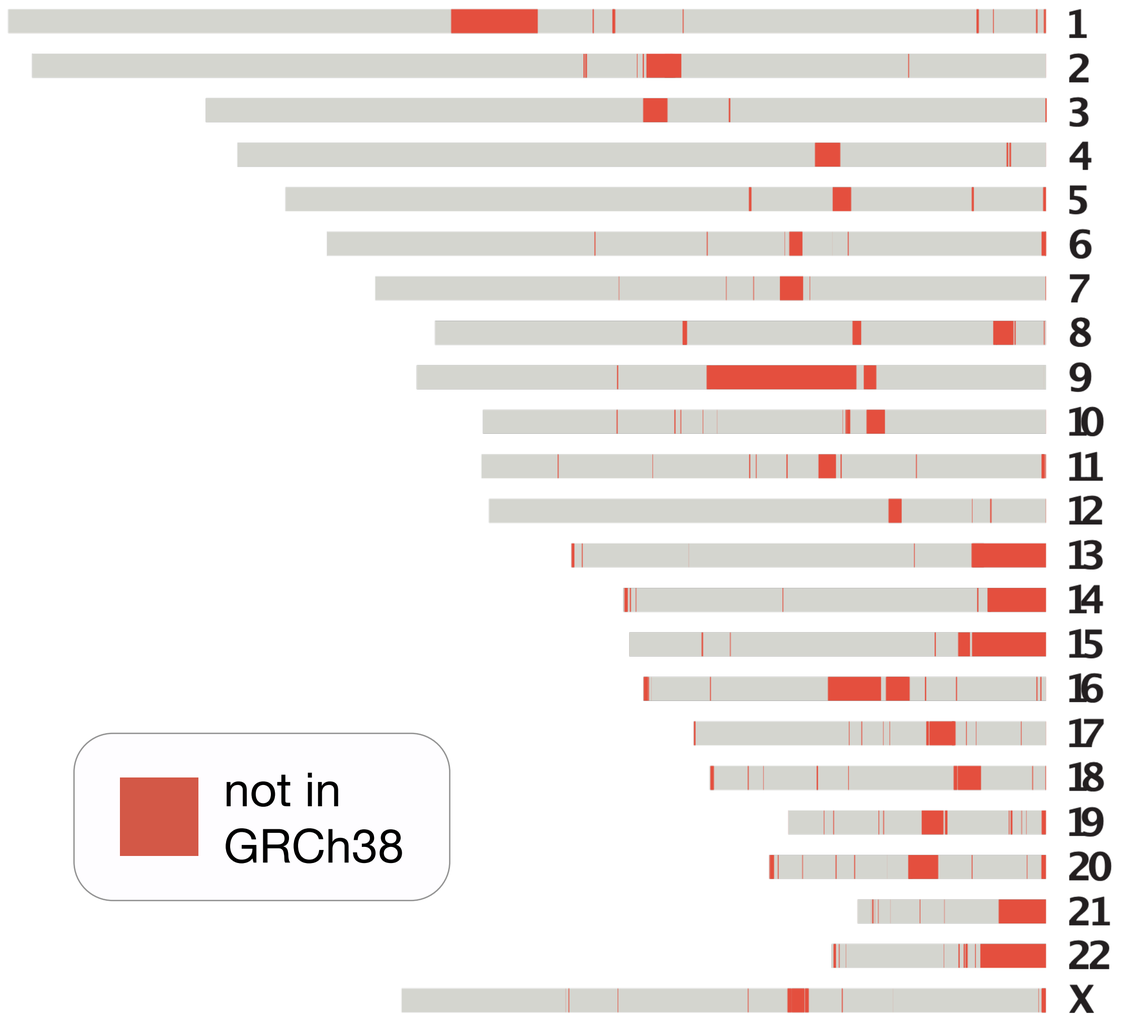


Nurk*, Koren*, Rhie*, Rautiainen*, *et al*., 2022, *Science*. The complete sequence of a human genome.

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

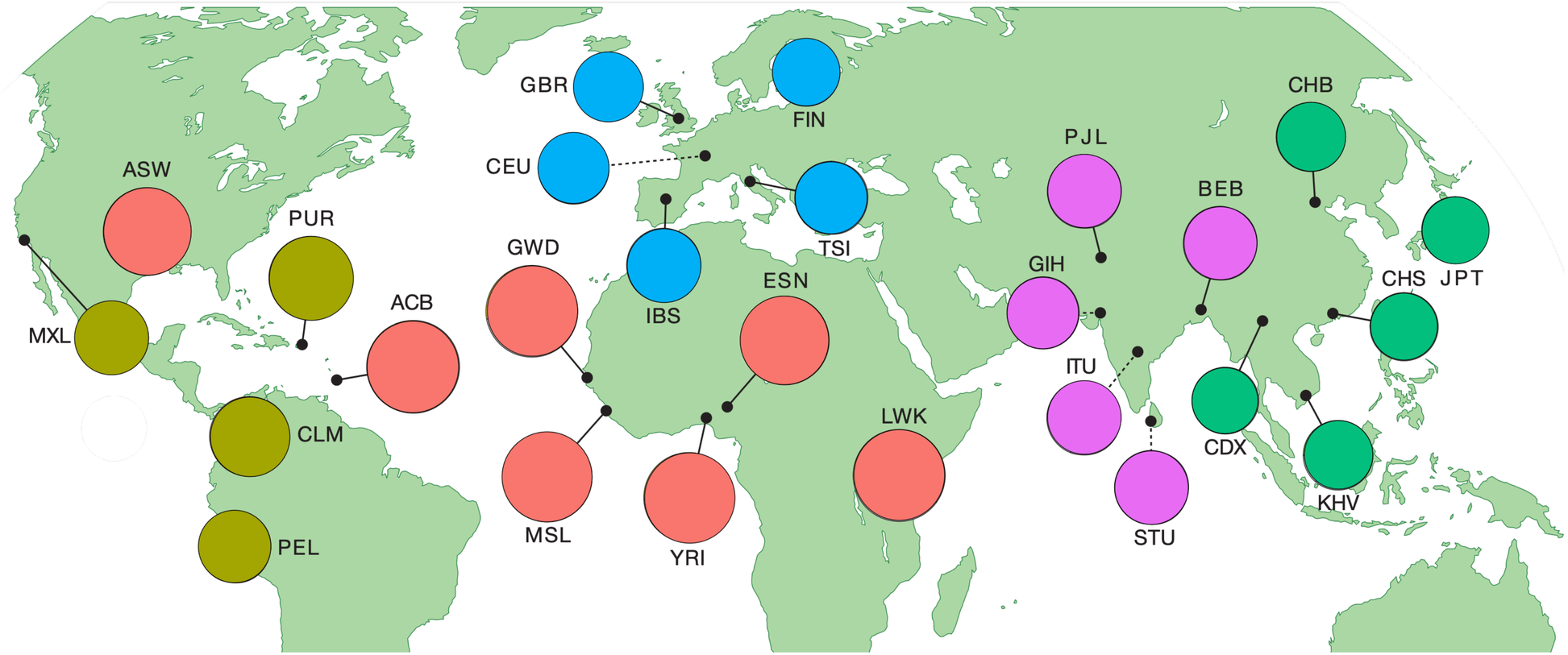
Most accurate assembly ever produced



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

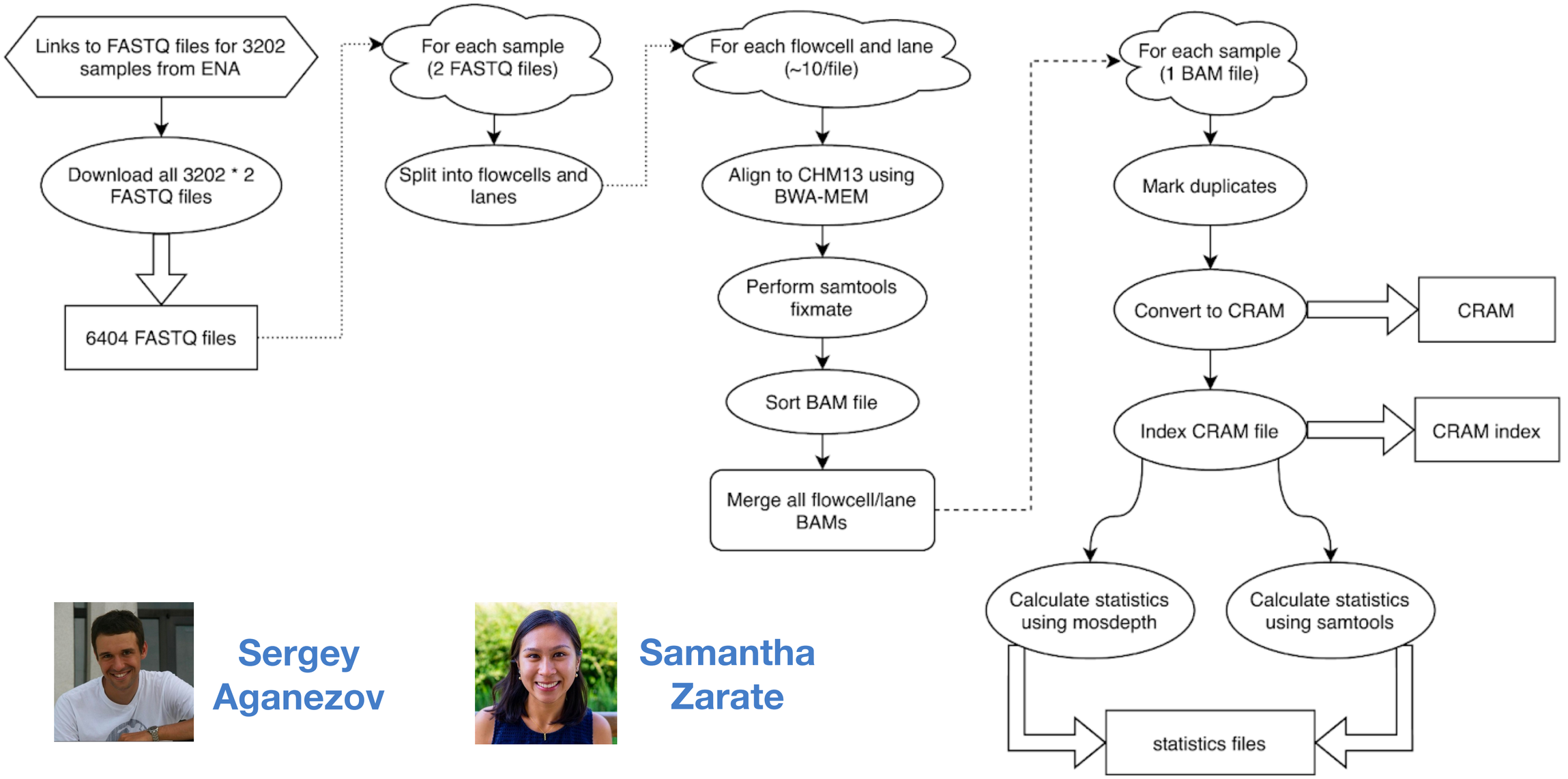
Analyzing diverse, short-read data with T2T-CHM13

- 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.

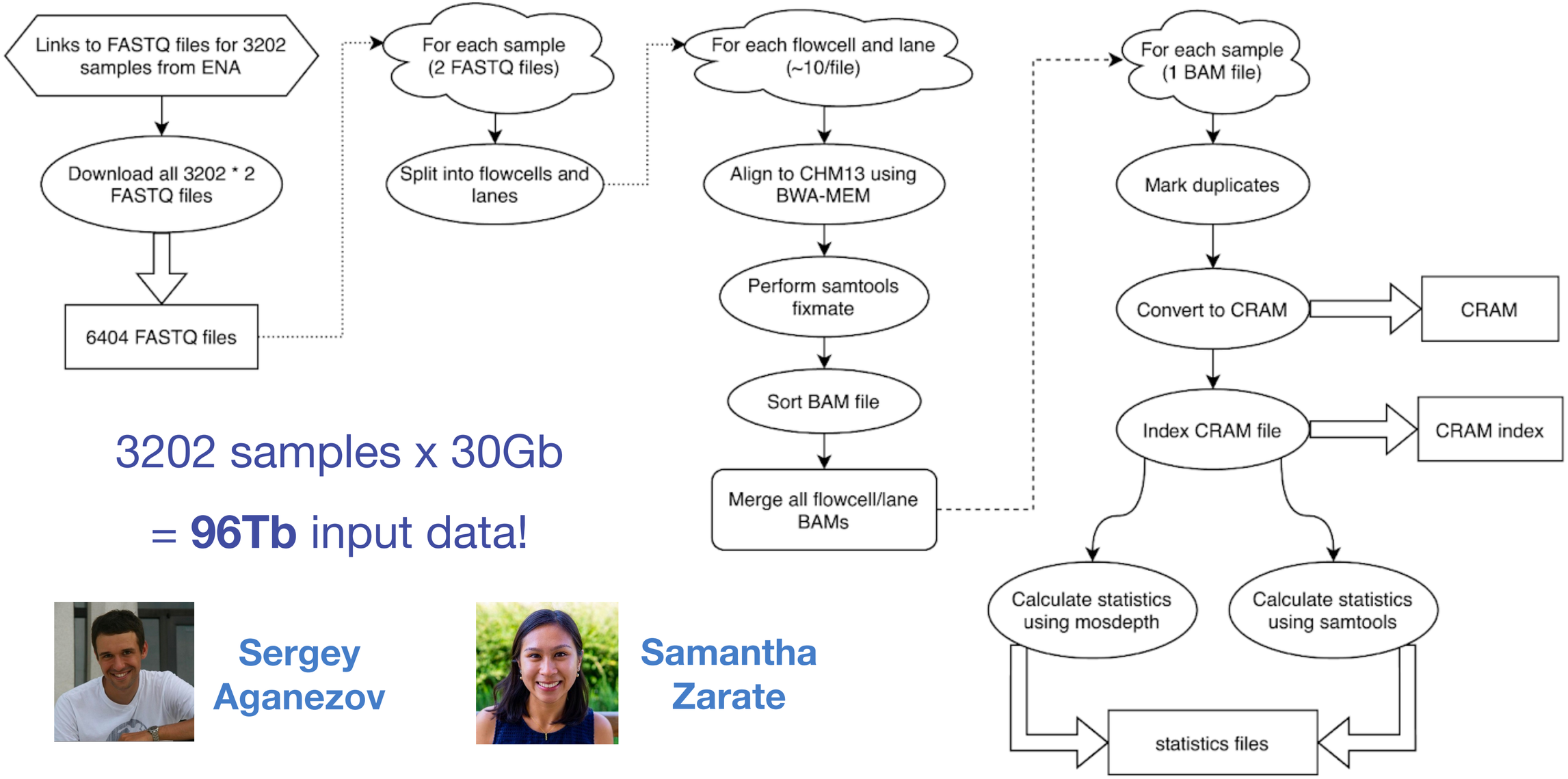
Population-scale short-read alignment and variant calling



Sergey Aganezov

Samantha Zarate

Population-scale short-read alignment and variant calling



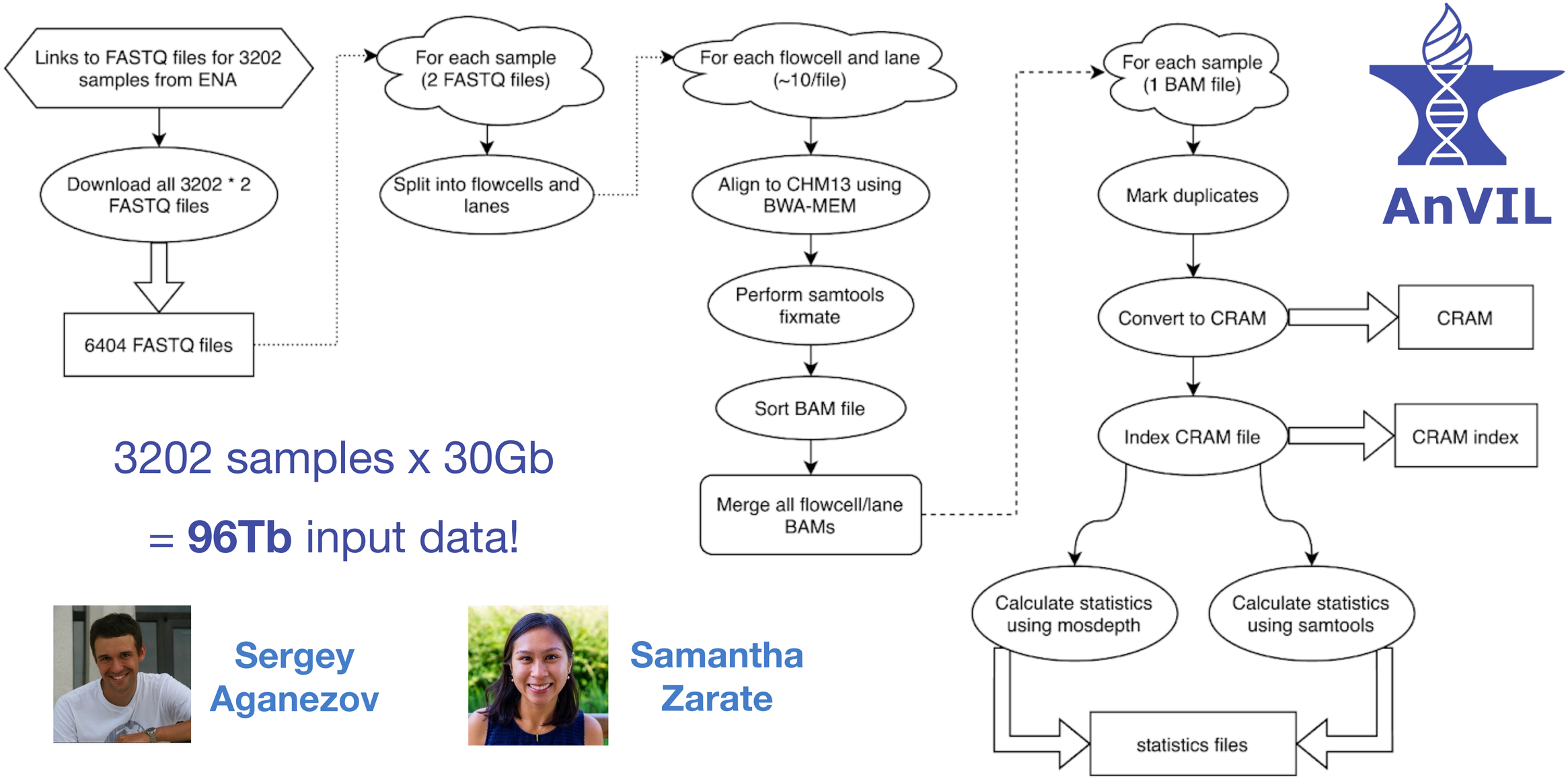
3202 samples x 30Gb

= **96Tb** input data!

Sergey Aganezov

Samantha Zarate

Population-scale short-read alignment and variant calling



AnVIL

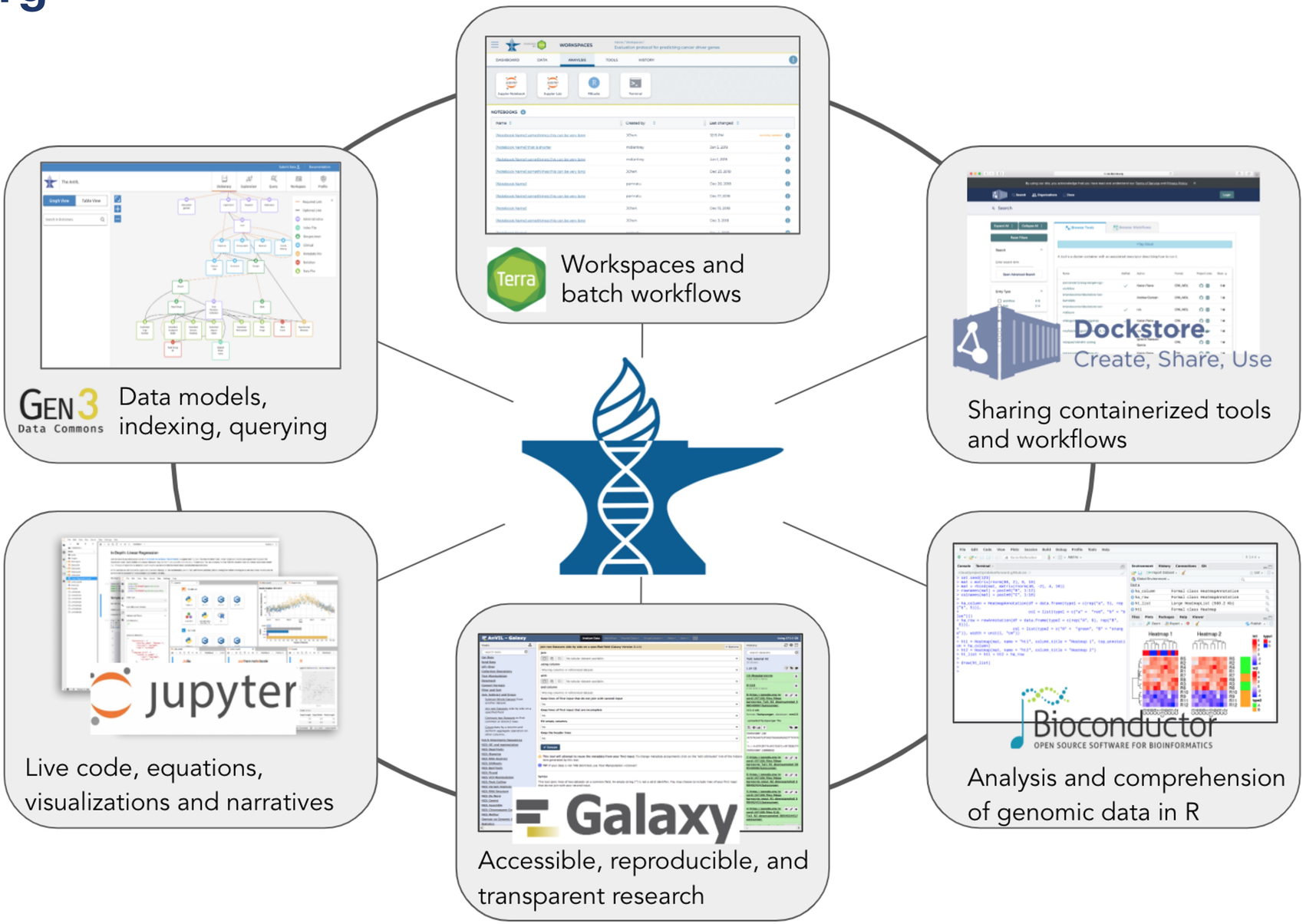
3202 samples x 30Gb

= **96Tb** input data!

Sergey Aganezov

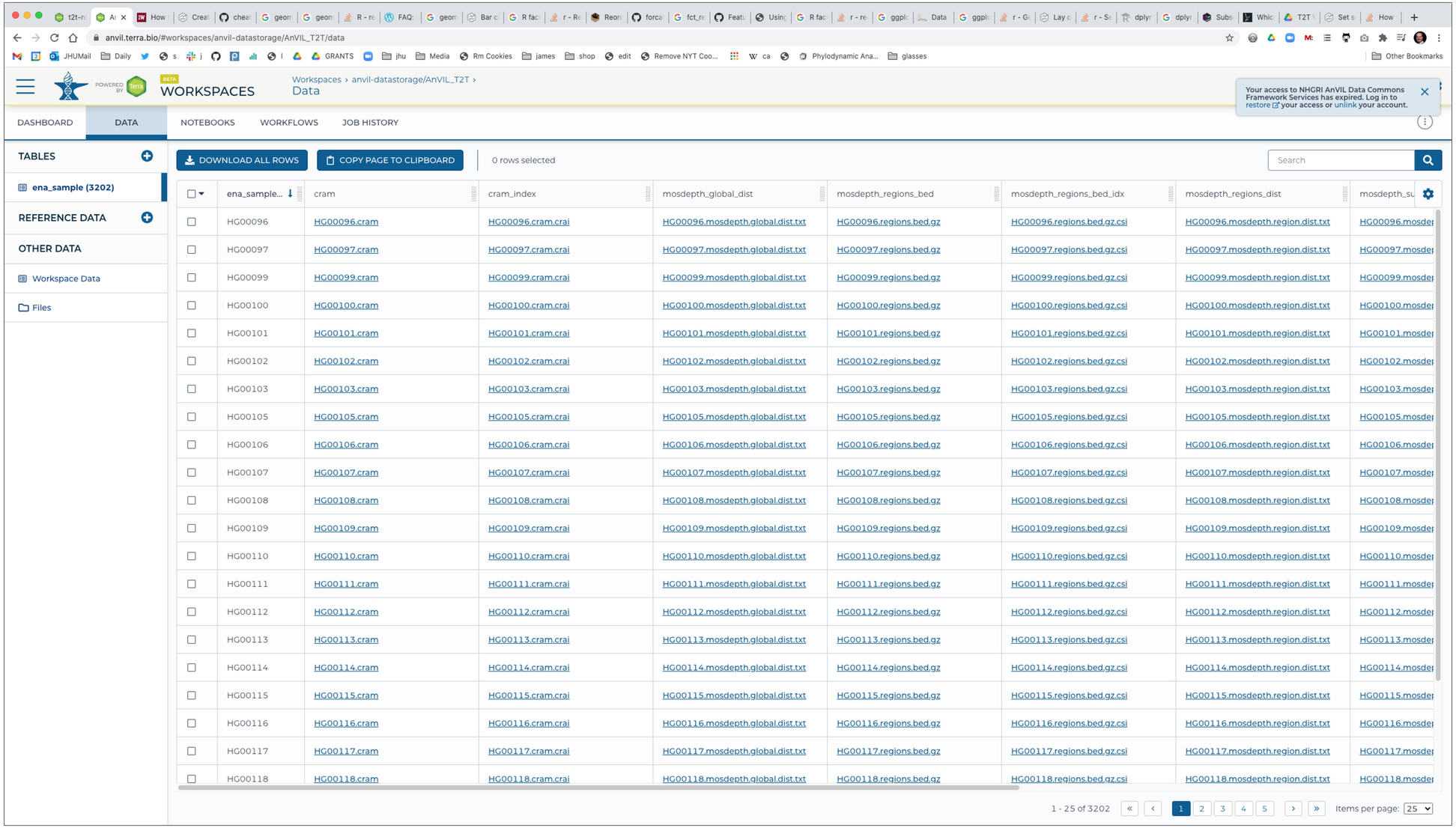
Samantha Zarate

anvilproject.org

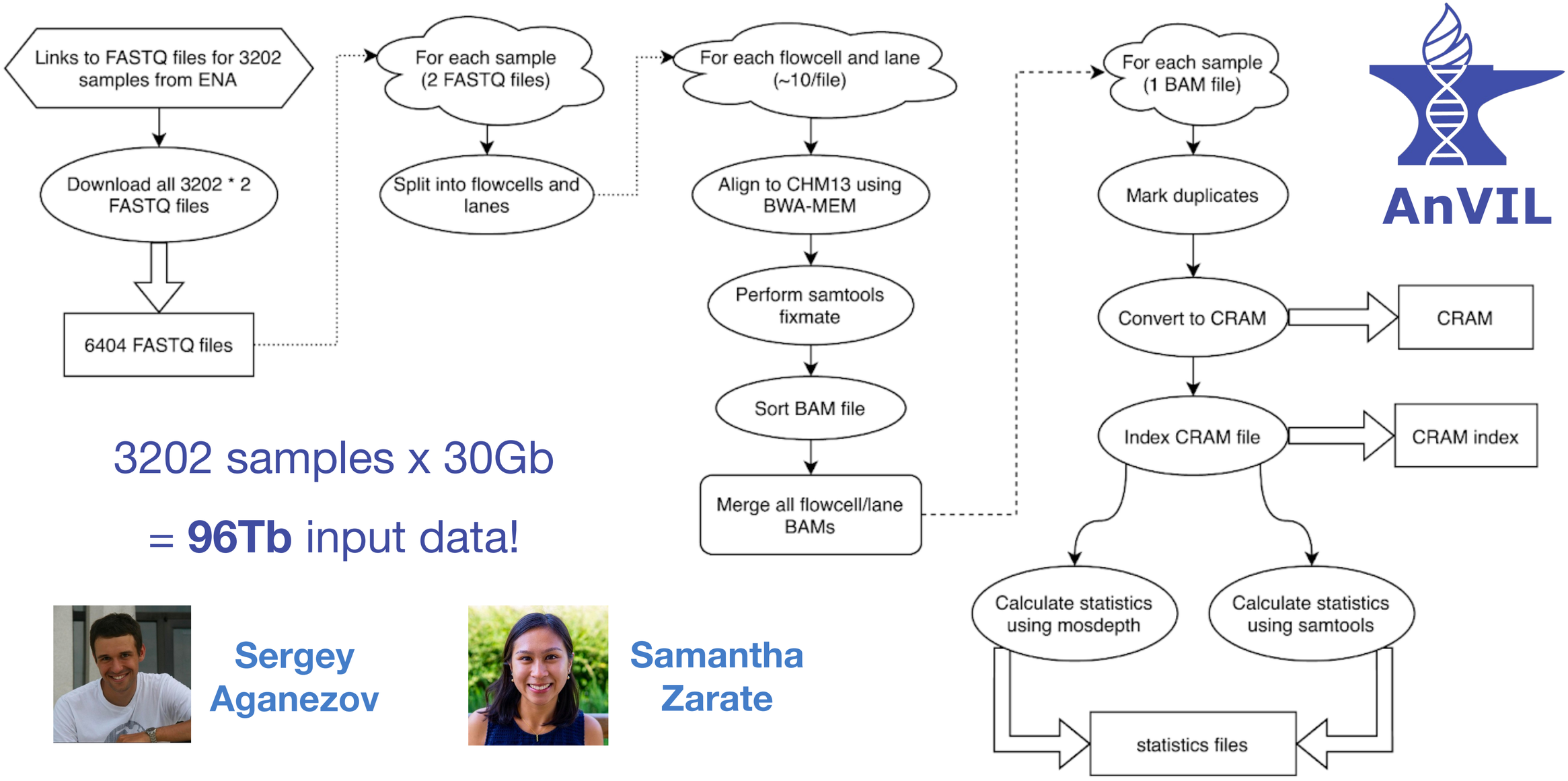


Schatz *et al.*, 2022, *Cell Genomics*. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space.

AnVIL: Data Table



Population-scale short-read alignment and variant calling



AnVIL

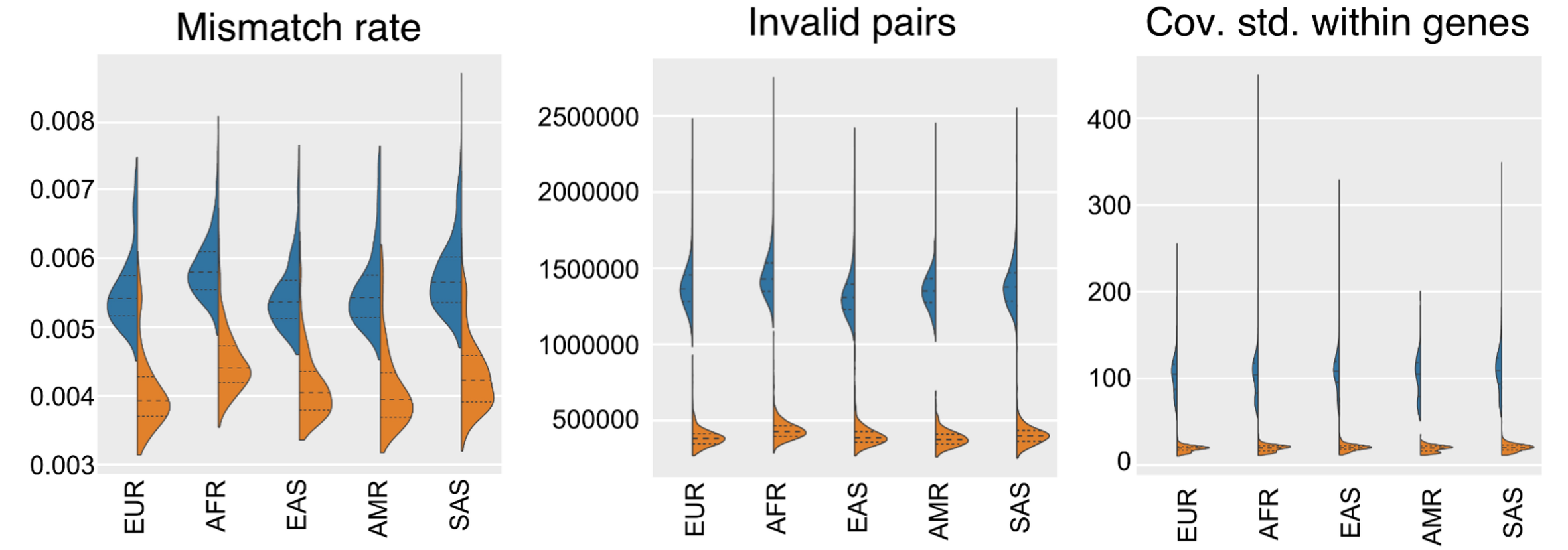
3202 samples x 30Gb

= **96Tb** input data!

Sergey Aganezov

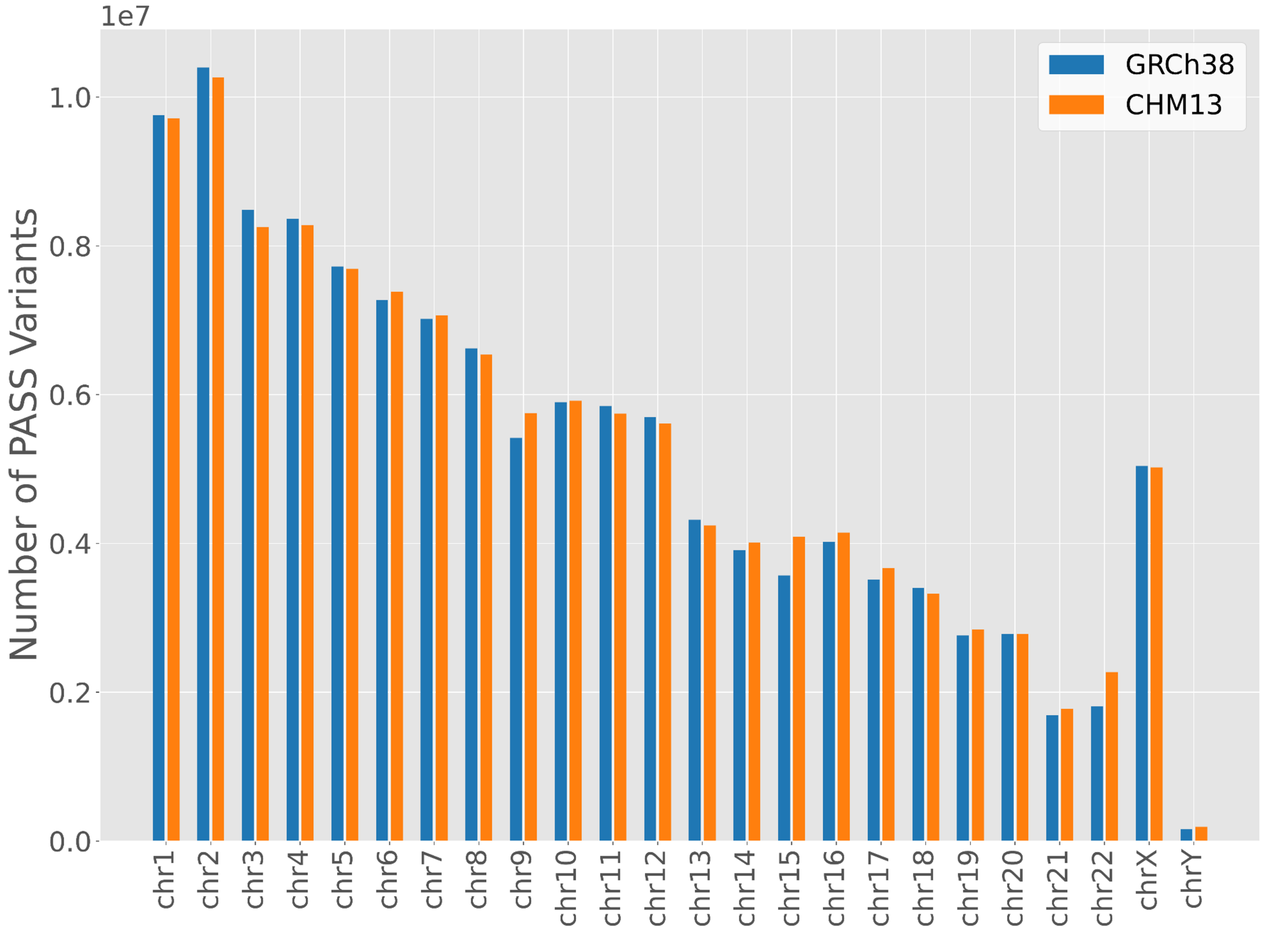
Samantha Zarate

T2T-CHM13 improves short-read alignment



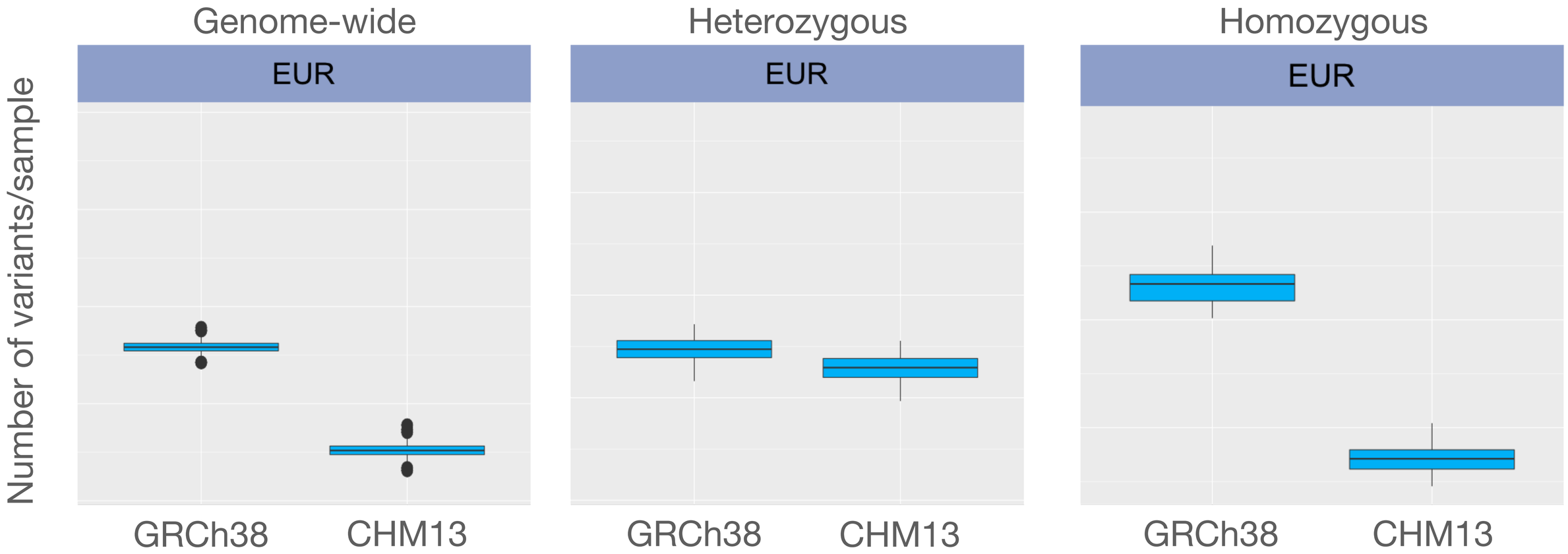
Sergey Aganezov

T2T-CHM13 allows discovery of more variants



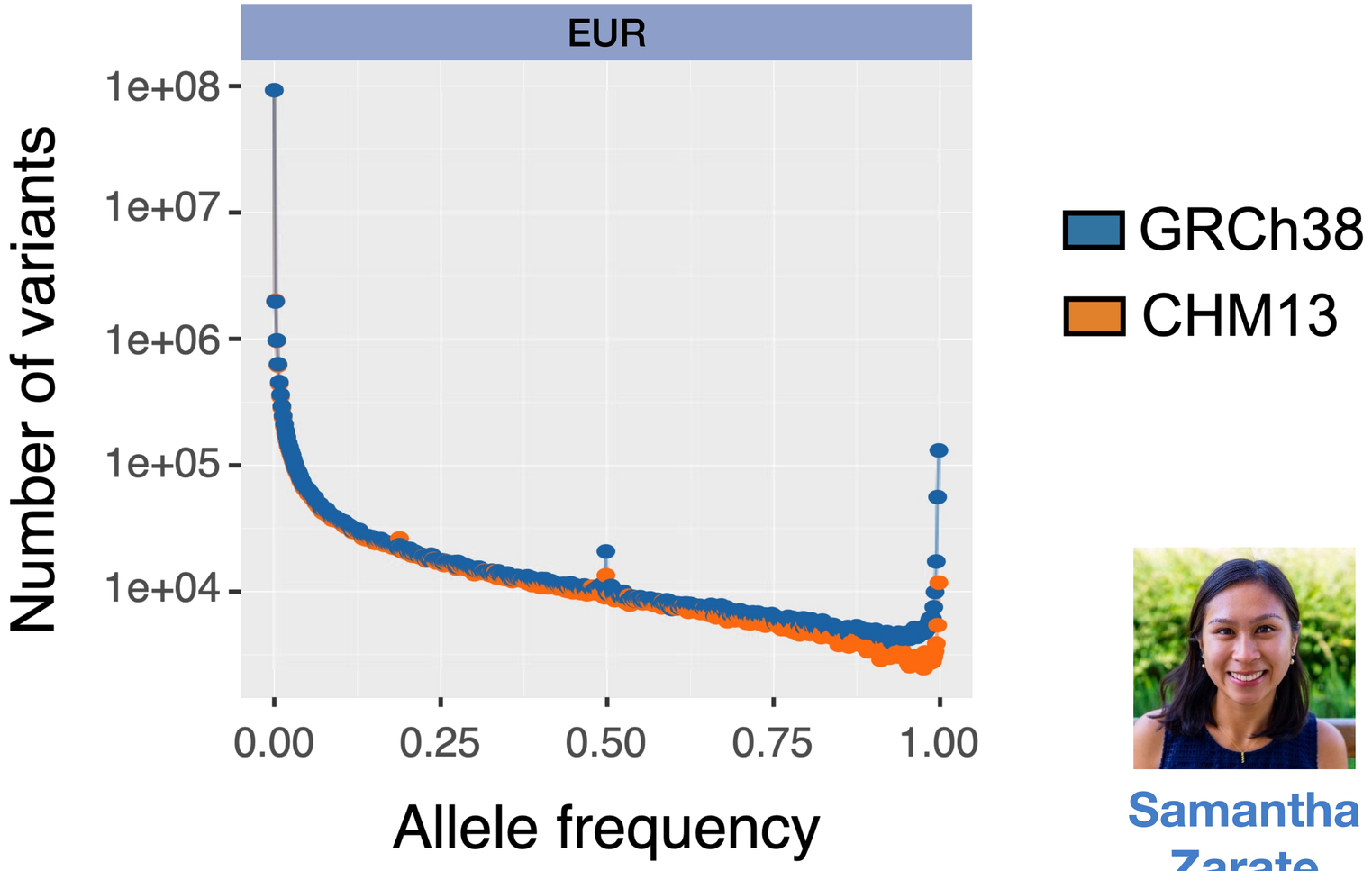
	GRCh38	CHM13
# of PASS variants	125,484,020	126,591,489

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



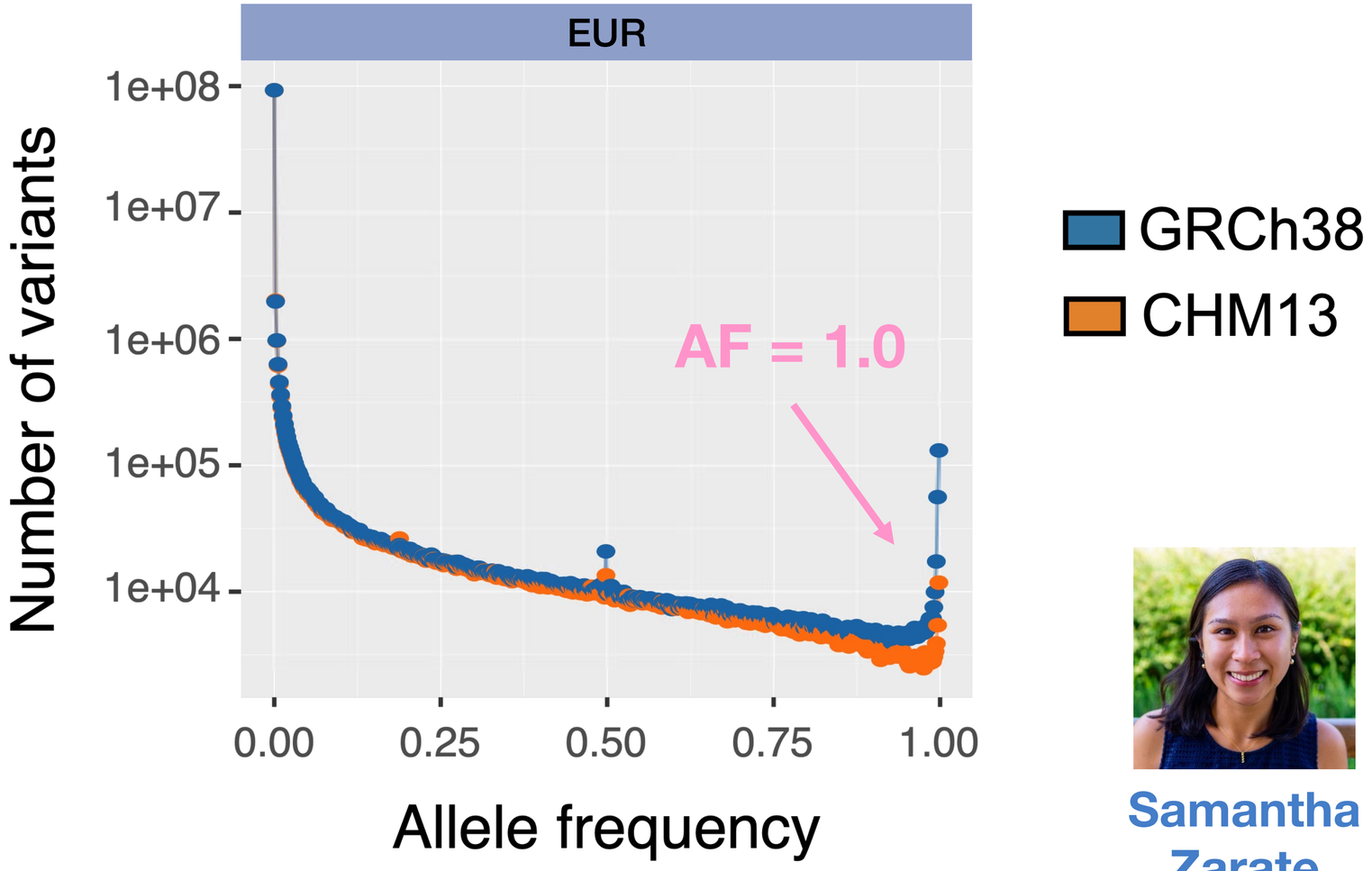
Samantha Zarate

T2T-CHM13 improves short-read variant calling

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

GRCh38 ...ATGGCAATATG...

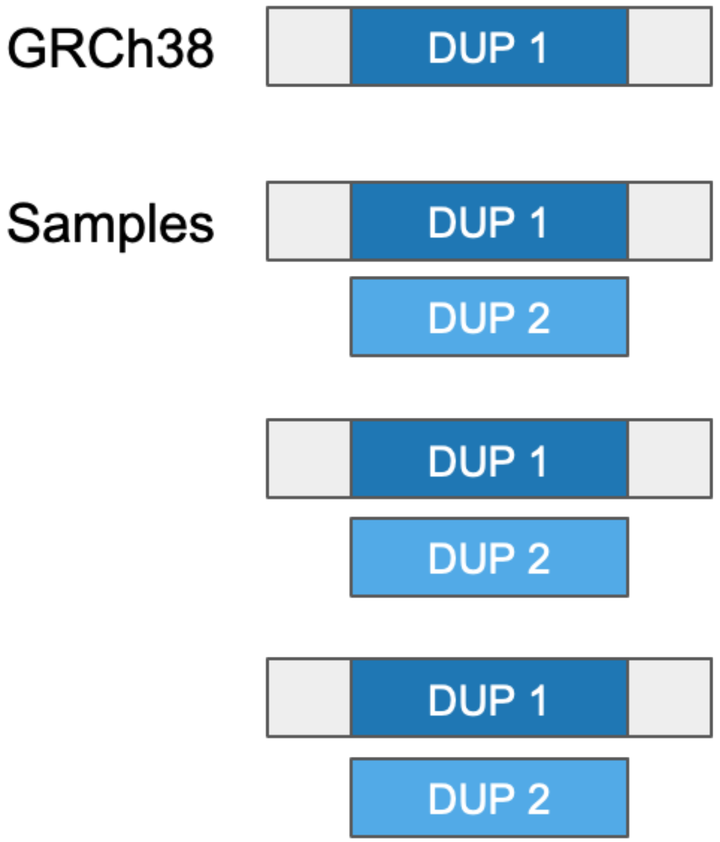
Samples
...ATGGTAATATG...
...ATGGTAATATG...
...ATGGTAATATG...
...ATGGTAATATG...
...ATGGTAATATG...
...ATGGTAATATG...

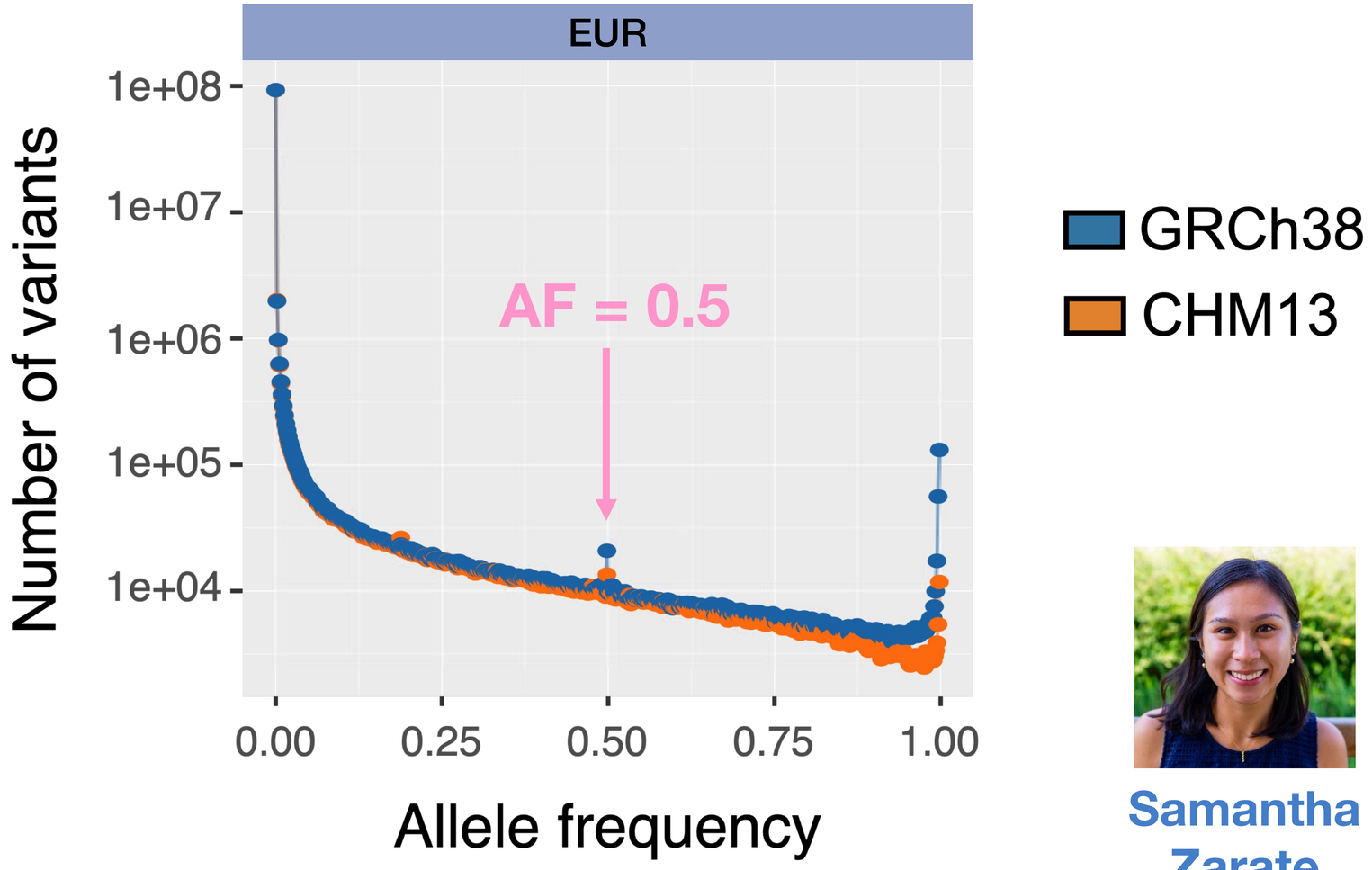


Samantha Zarate

T2T-CHM13 improves short-read variant calling

- Fewer false heterozygous variants from GRCh38 collapsed duplications

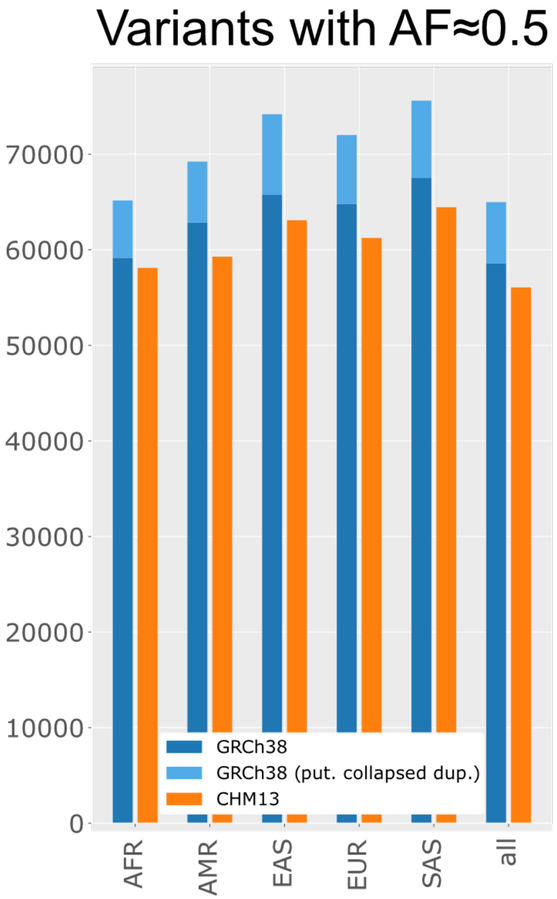


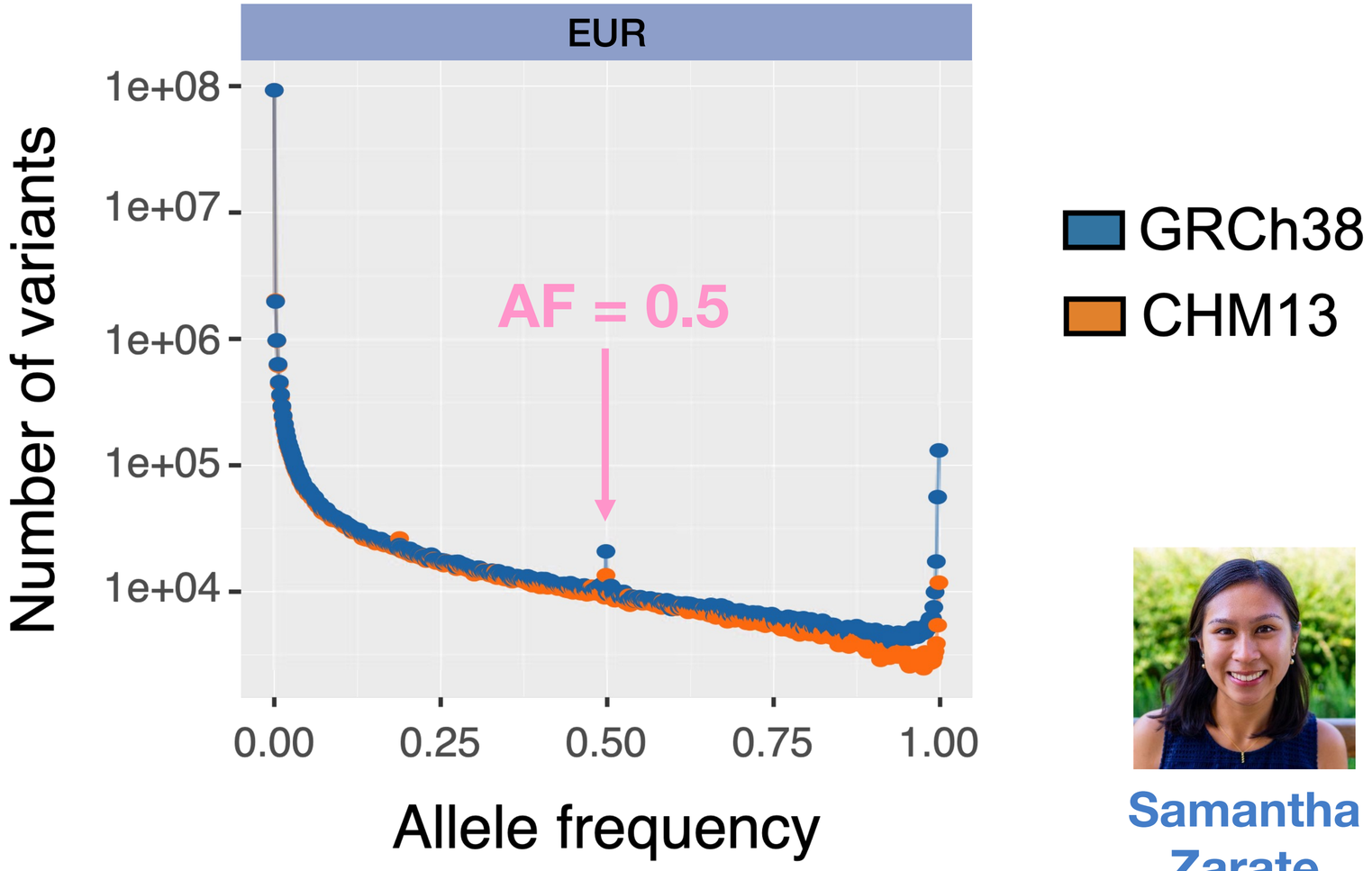


Samantha Zarate

T2T-CHM13 improves short-read variant calling

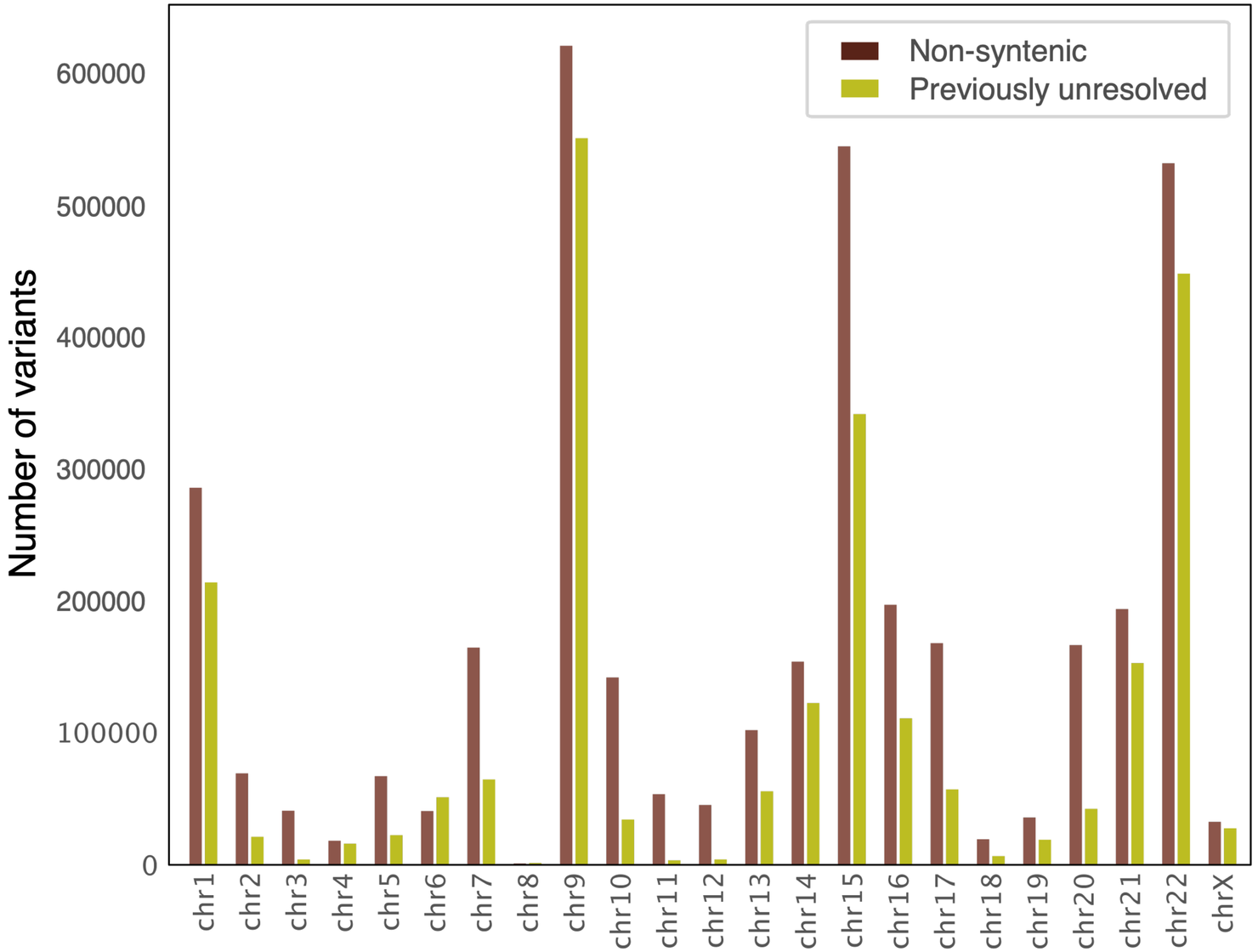
- Fewer false heterozygous variants from GRCh38 collapsed duplications

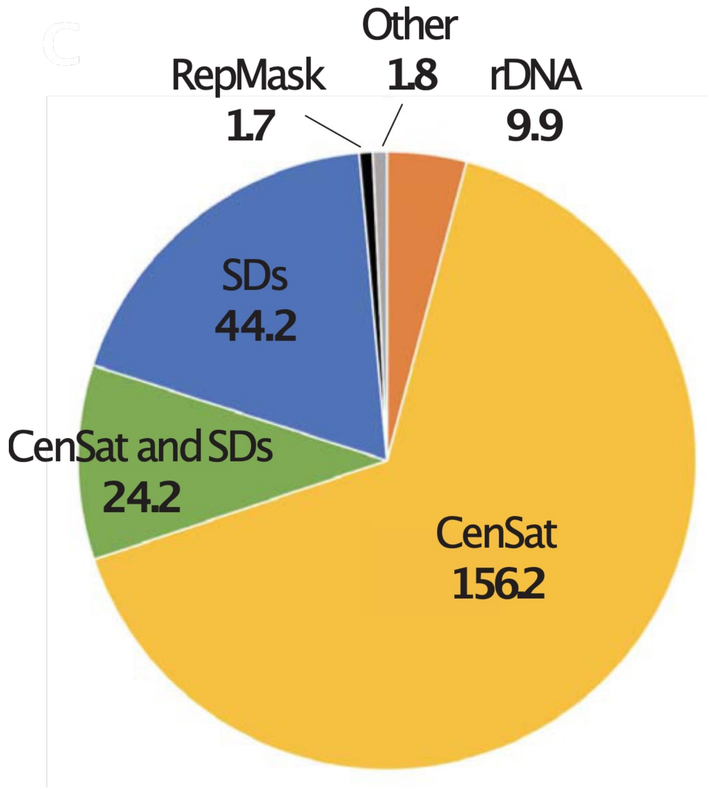




Samantha Zarate

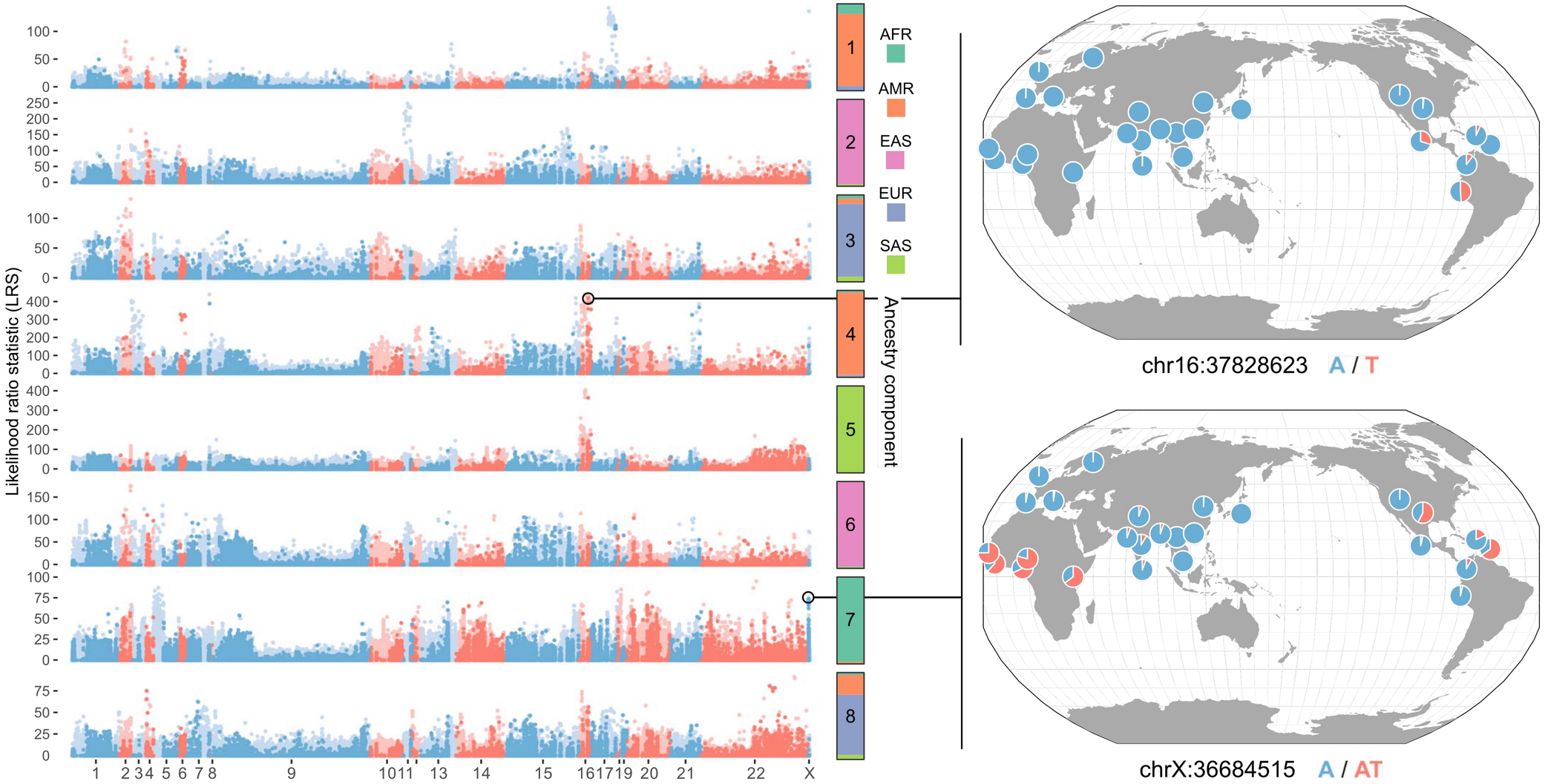
“Novel” variants revealed by T2T-CHM13





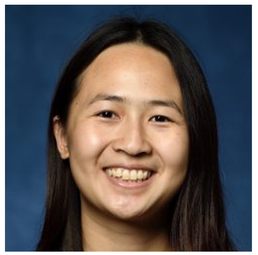
Samantha Zarate

Extreme AF differentiation of novel variants



chr16:37828623 A / T

chrX:36684515 A / AT



Stephanie Yan

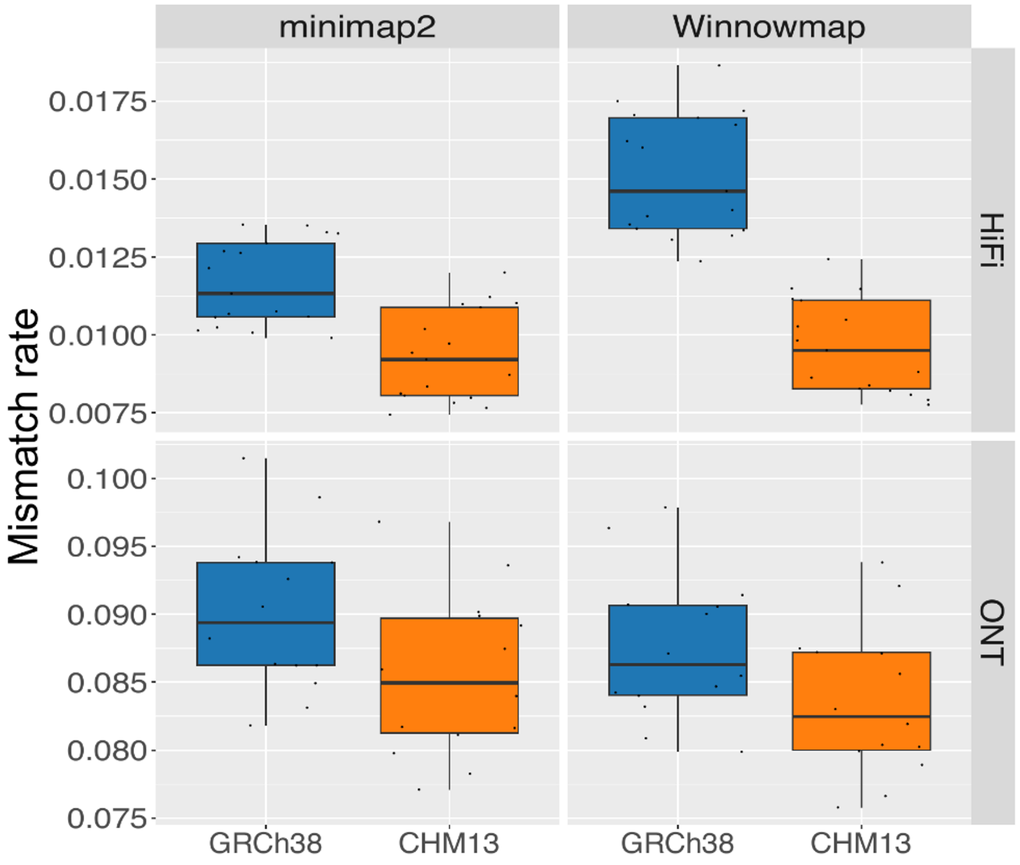
Long-read alignment and variant calling with T2T-CHM13

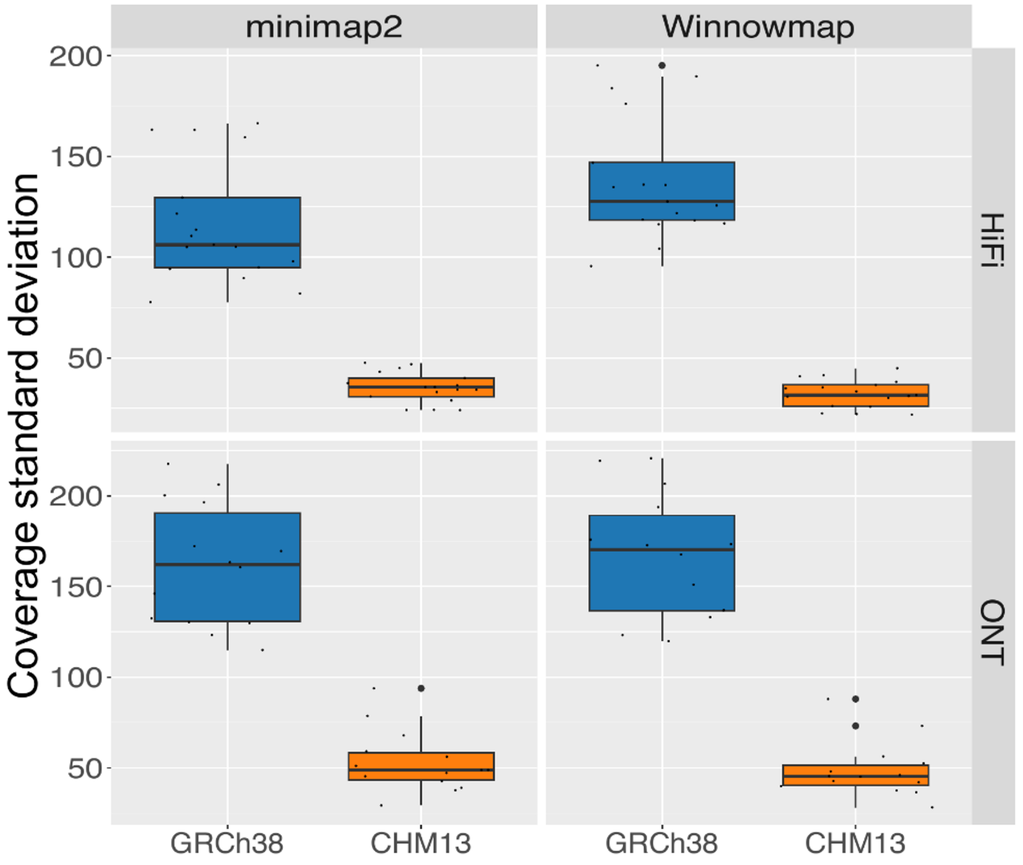
- 17 diverse samples from the Human Pangenome Reference Consortium and Genome in a Bottle
- PacBio HiFi data + 14 samples with ONT data

Melanie Kirsche

Sergey Aganezov

T2T-CHM13 improves alignment for long reads

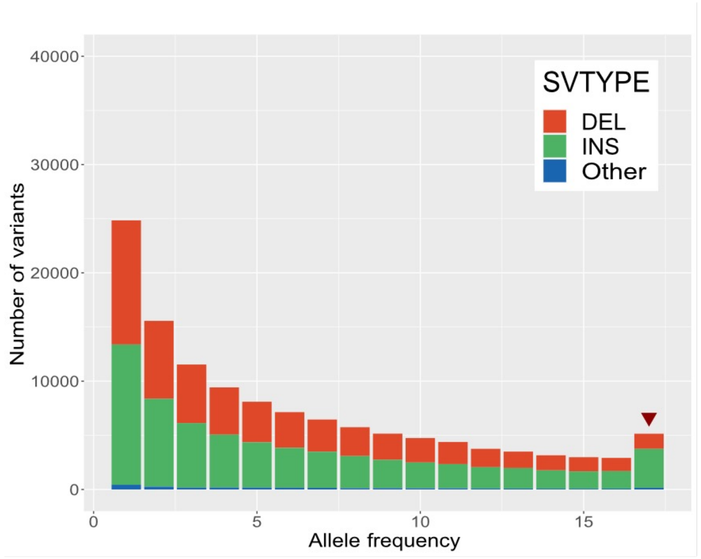


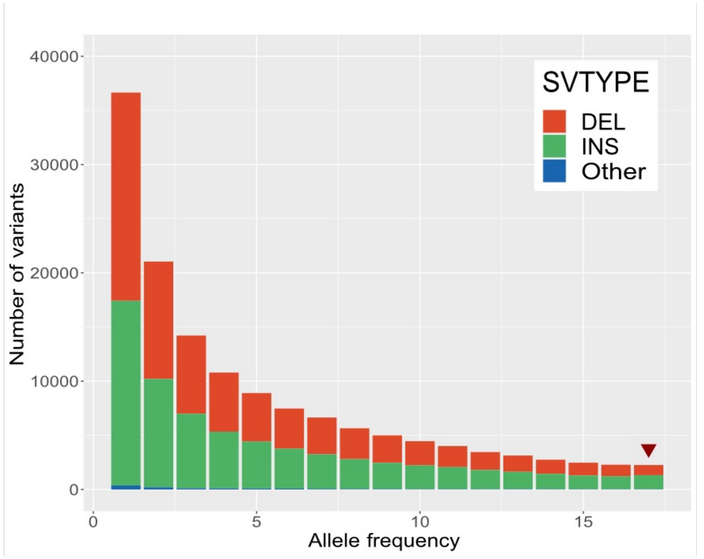


Melanie Kirsche

Sergey Aganezov

T2T-CHM13 improves SV-calling with long reads

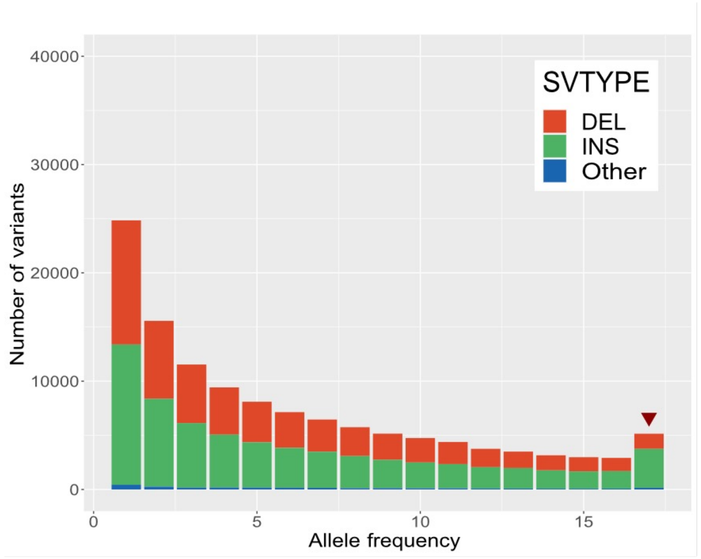


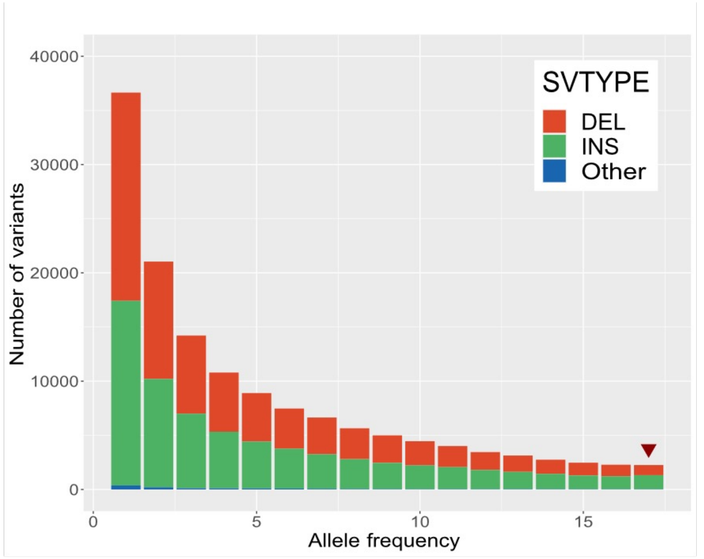


Melanie Kirsche

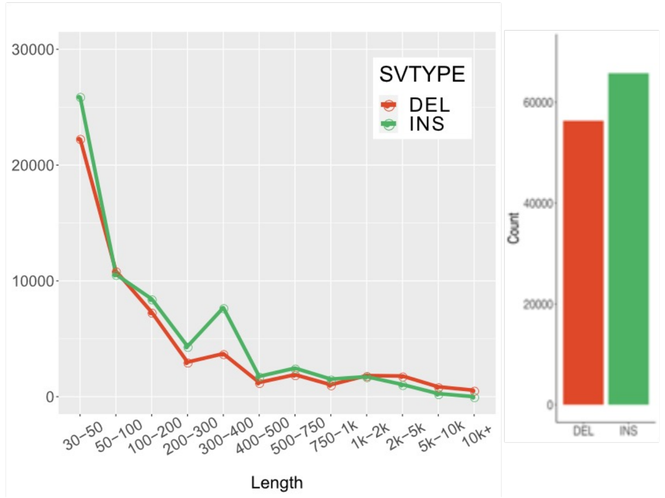
Sergey Aganezov

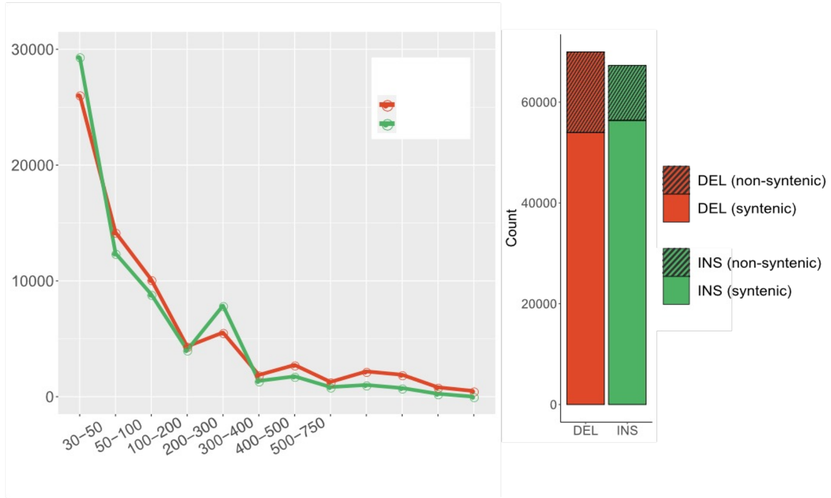
T2T-CHM13 improves SV-calling with long reads





Melanie Kirsche

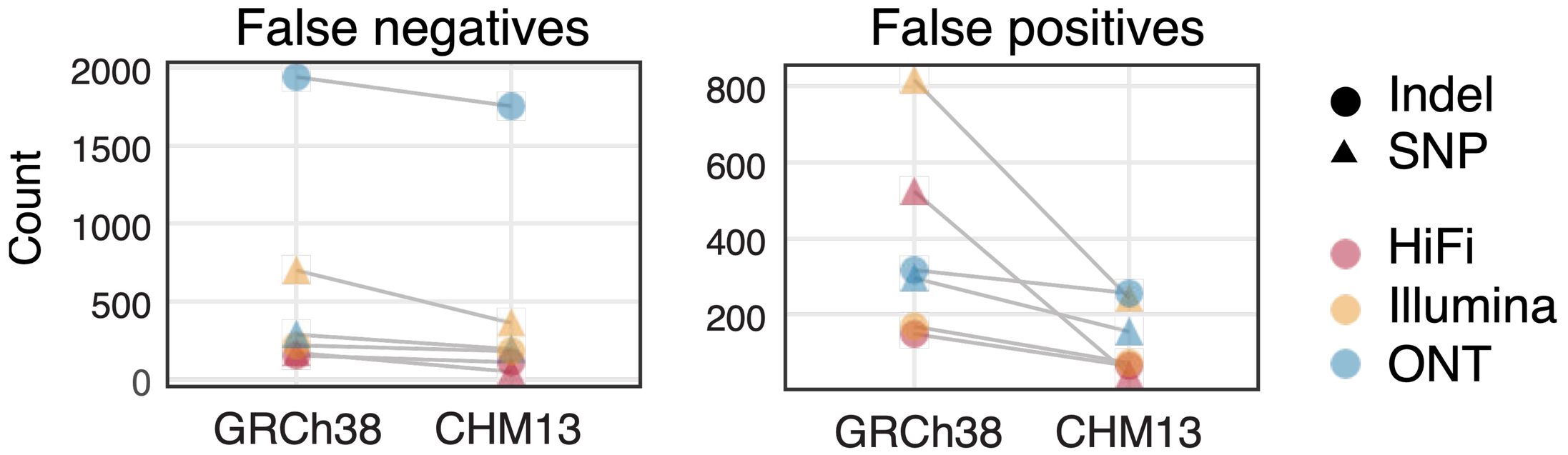


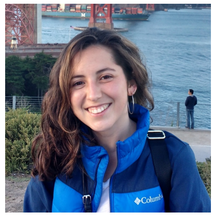


Sergey Aganezov

T2T-CHM13 improves clinical genomics variant calling

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





Daniela Soto

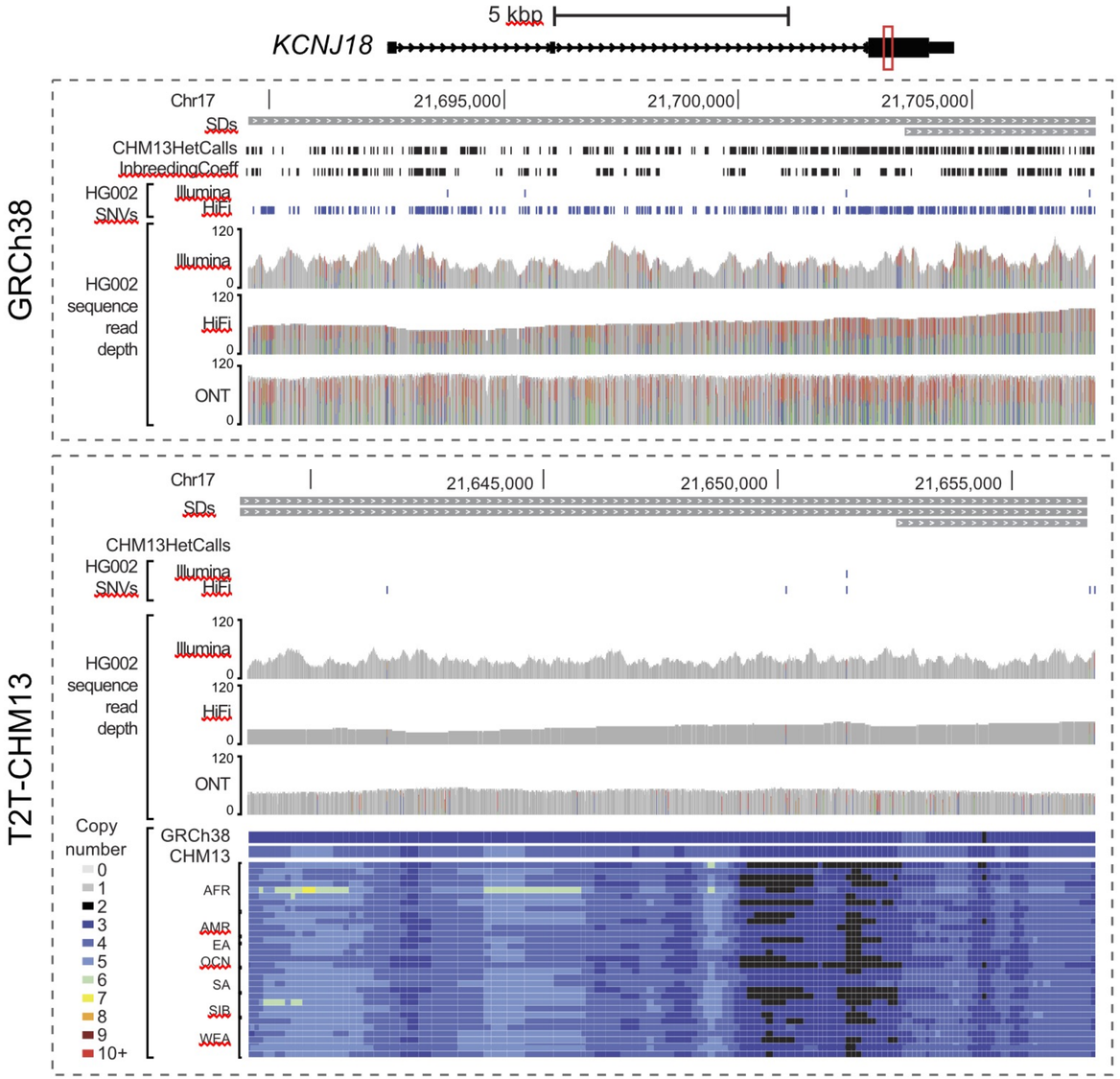
Megan Dennis

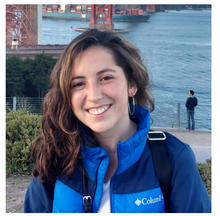
Justin Zook

Fritz Sedlazeck

Danny Miller

T2T-CHM13 improves clinical genomics variant calling





Daniela Soto

Megan Dennis

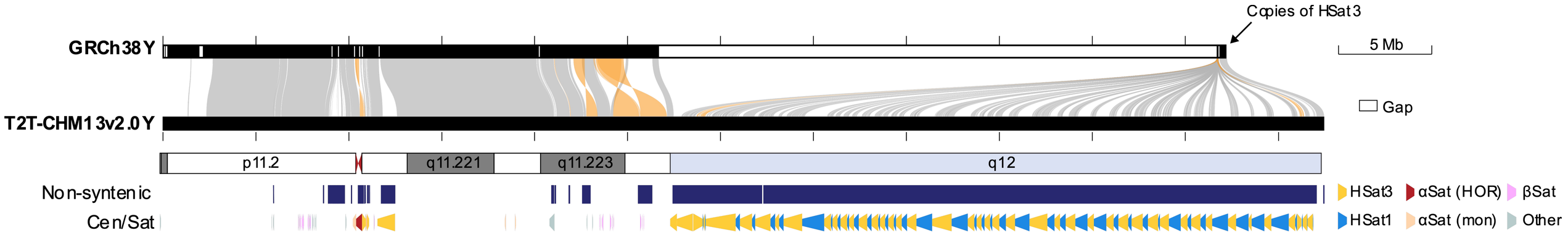
Justin Zook

Fritz Sedlazeck

Danny Miller

A telomere-to-telomere Y chromosome

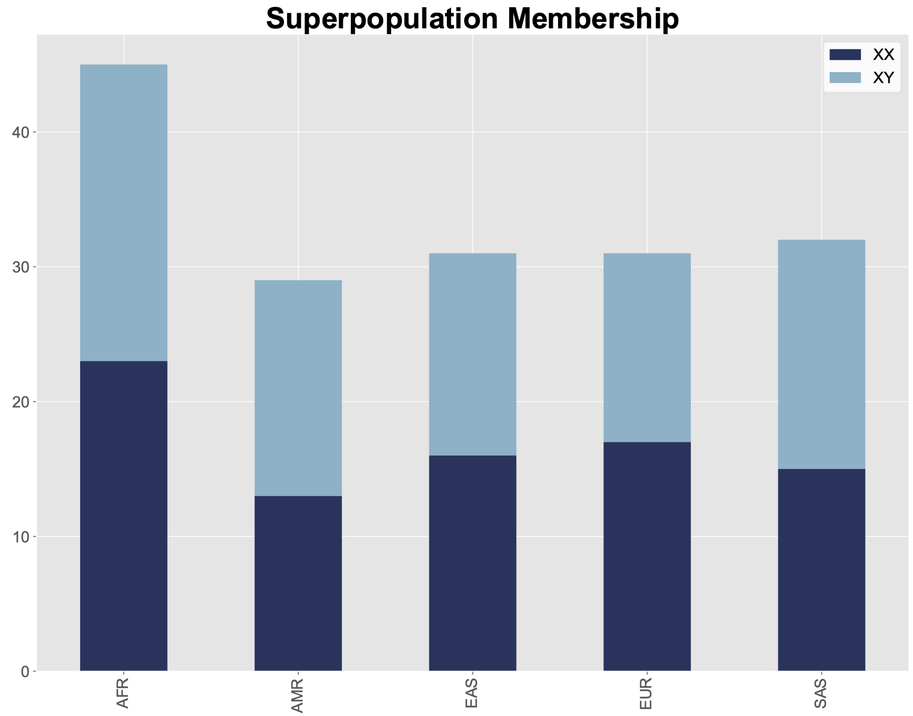
- CHM13 is an XX complete hydatidiform mole
- T2T-CHM13 v1 used the **GRCh38** Y chromosome
- More than 50% of the GRCh38 Y chromosome assembly is missing
- Used long-read sequencing to generate a complete Y chromosome assembly from HG002 cell line, completing the **T2T-CHM13 v2.0** reference

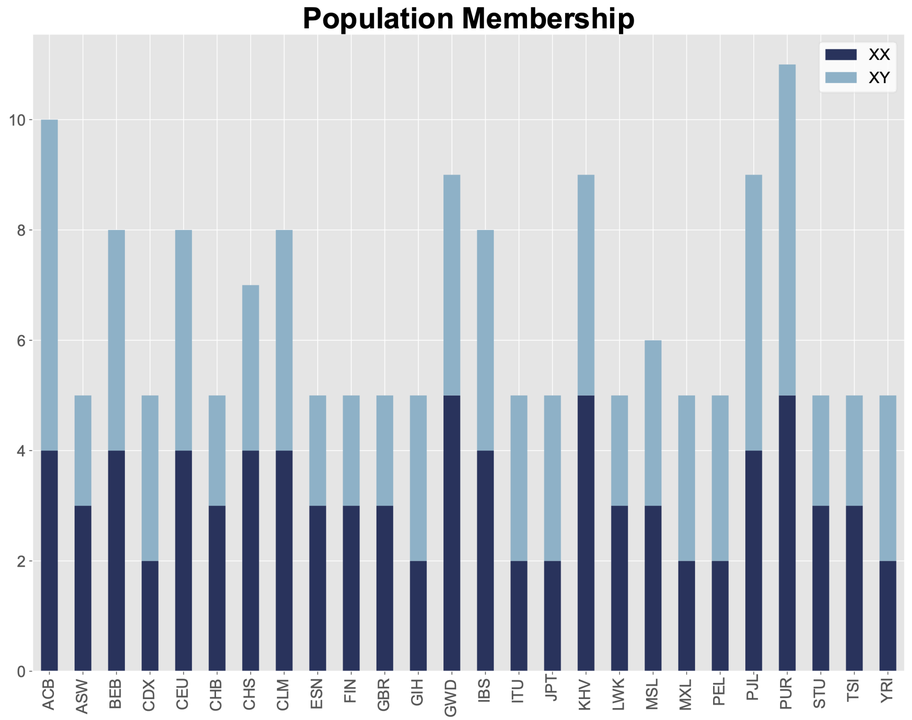


Arang Rhie

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

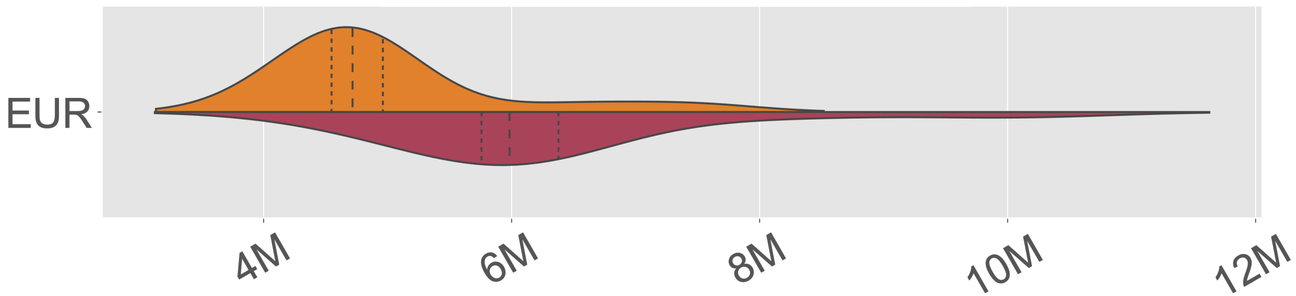
- 1000 Genomes Project (1KGP): 168 samples (84 XX, 84 XY)
- 30x sequencing by the New York Genome Center





Byrska-Bishop *et al*., 2021, *bioRxiv*. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios.

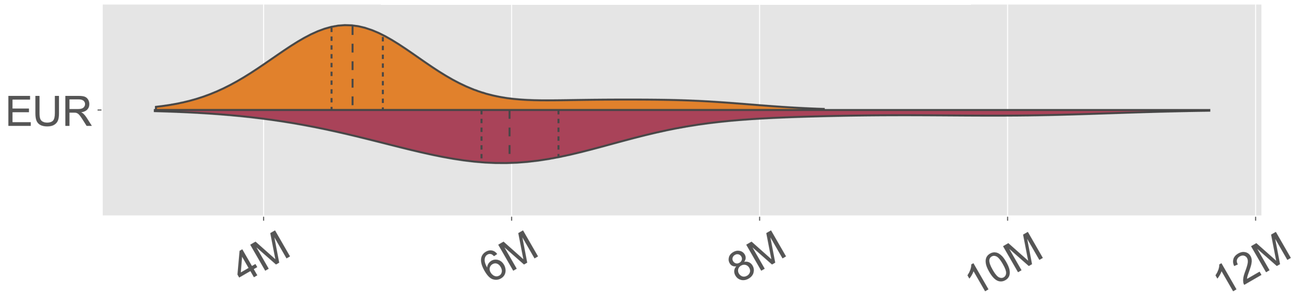
HG002Y improves short-read alignment



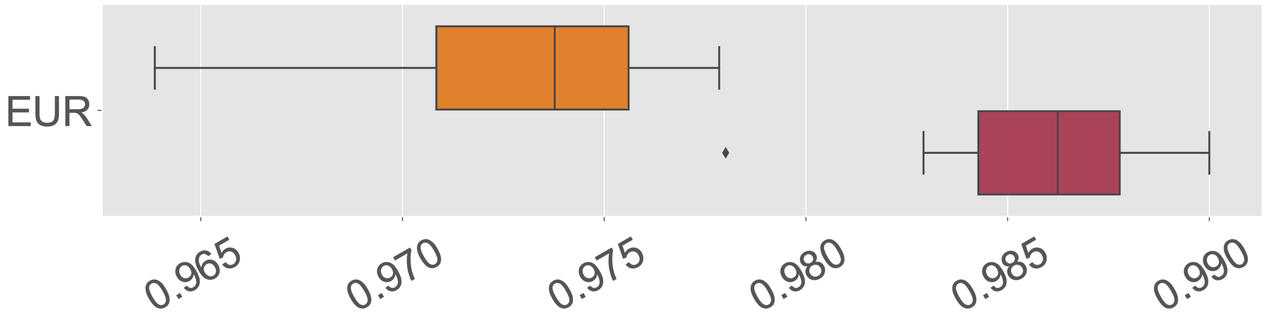
Reads Mapped (chrY)

CHM13v1 (GRCh38Y)
CHM13v2 (HG002Y)

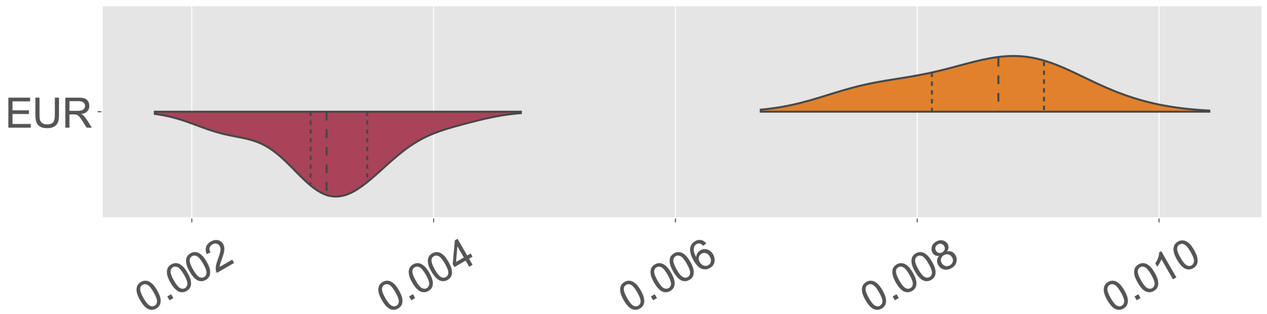
HG002Y improves short-read alignment



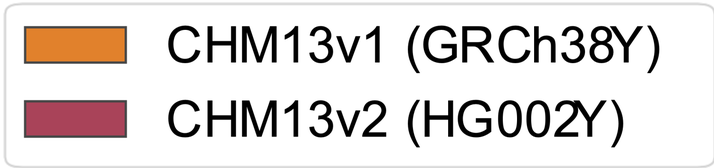
Reads Mapped (chrY)



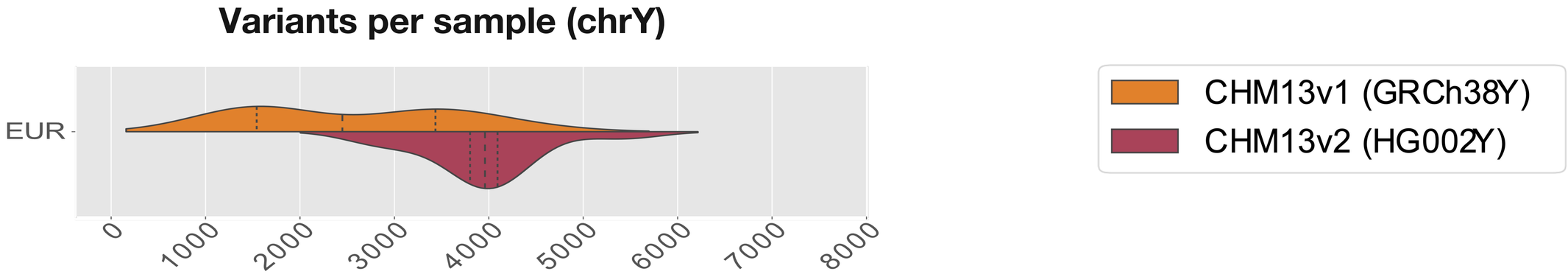
Proportion of reads properly paired (chrY)



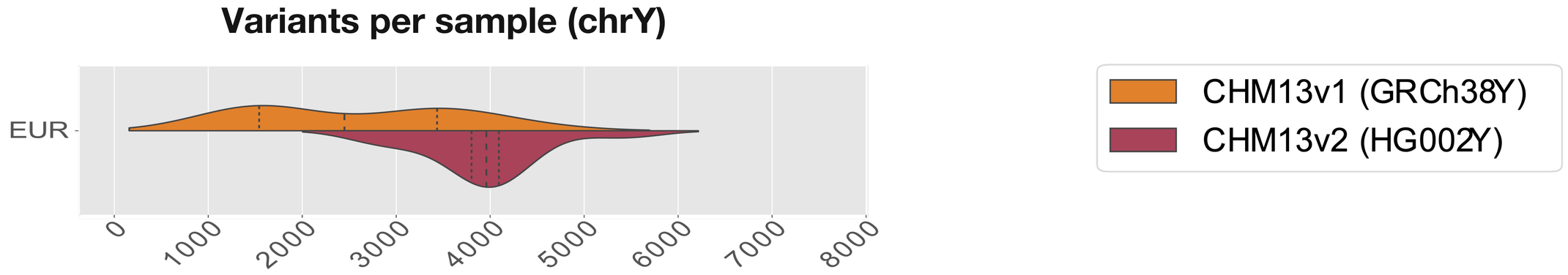
Mismatch rate (chrY)

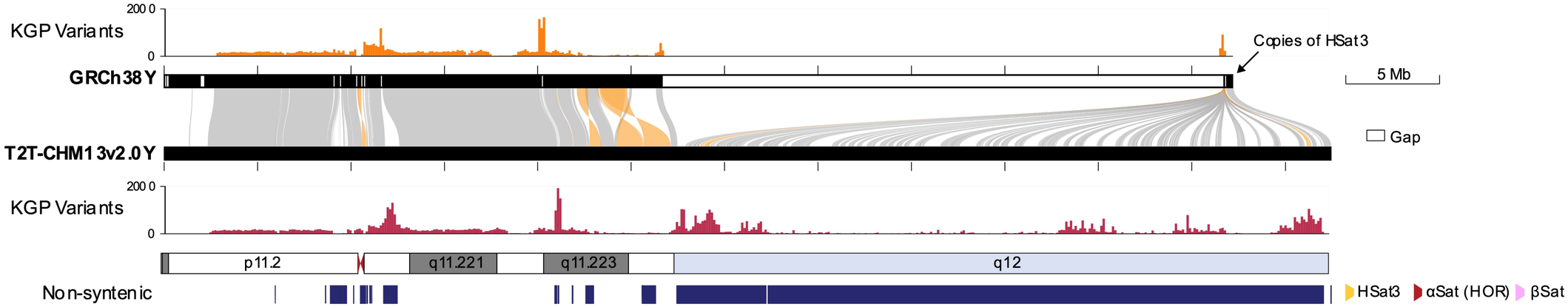


HG002Y improves variant calling in diverse samples

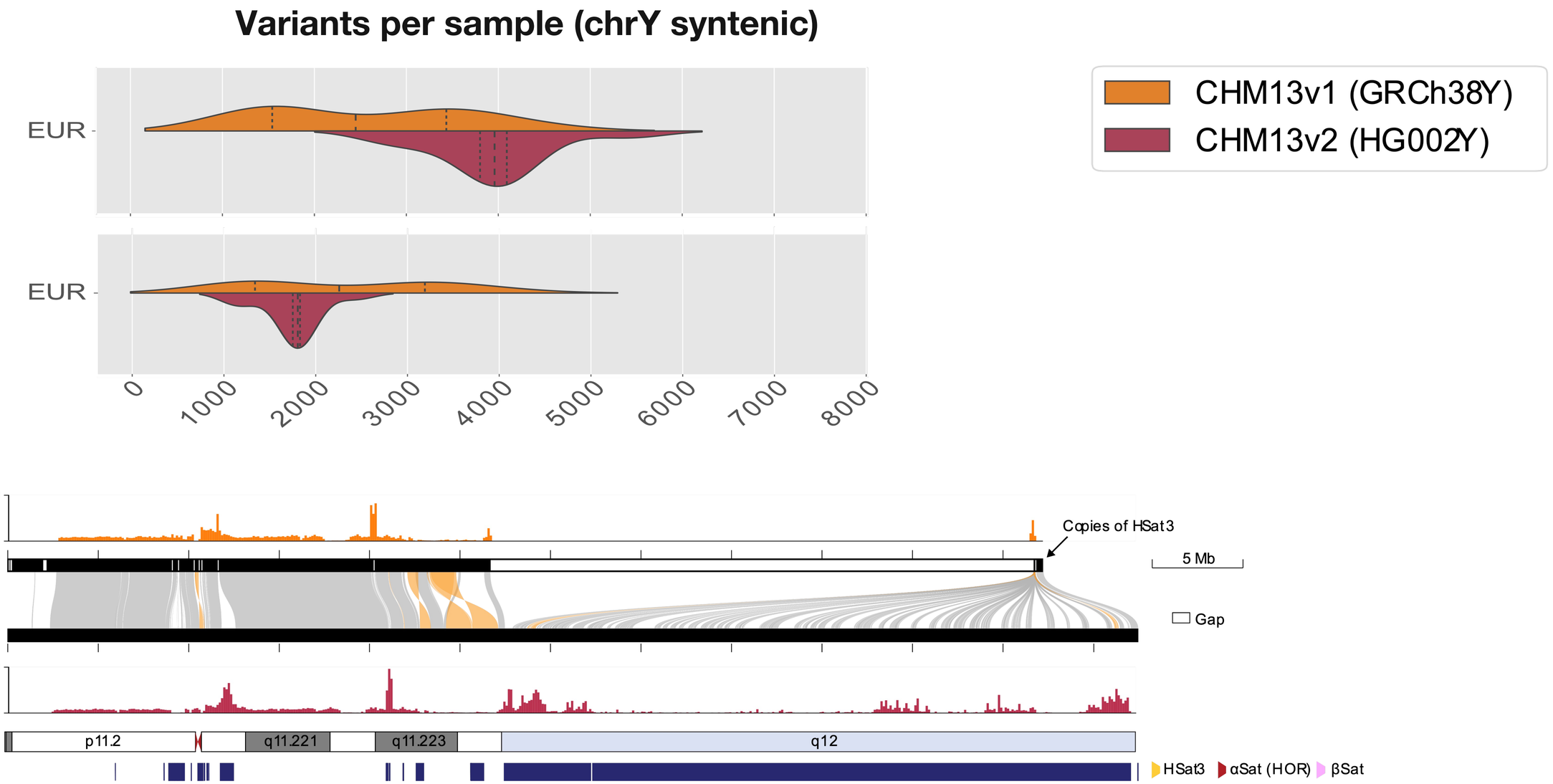


HG002Y improves variant calling in diverse samples



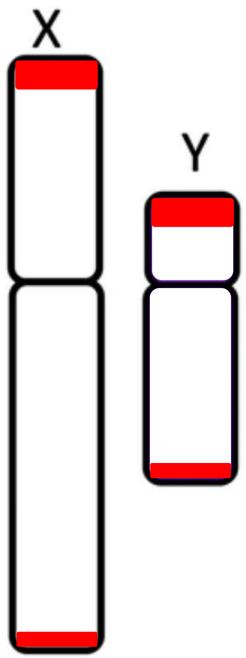


HG002Y improves variant calling in diverse samples

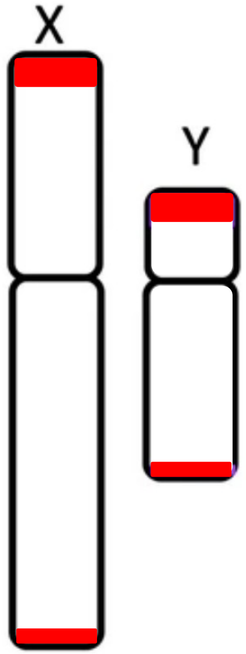


Pseudo-autosomal regions on the sex chromosomes

XX Samples



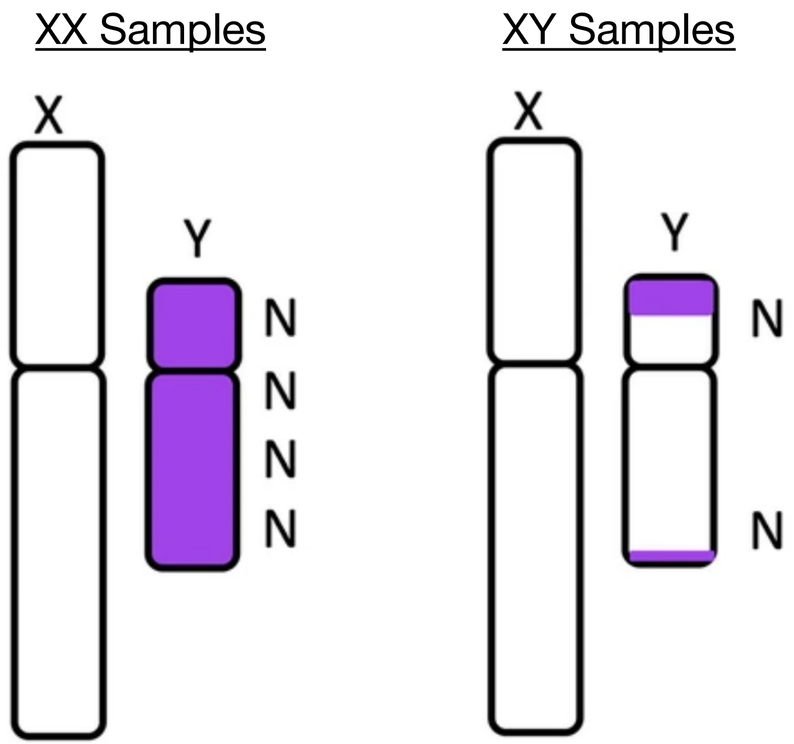
XY Samples



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

Karyotype-specific alignment

Used **XYalign** for alignment

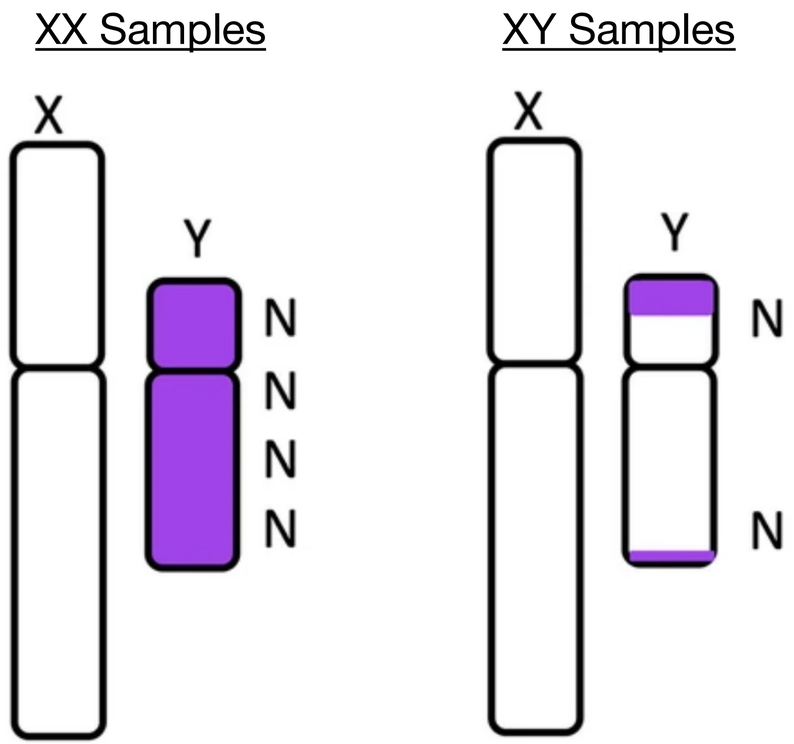


Adapted from Olney, K. C., et al. (2020) *BMC*

Webster, T. H., et al., 2019, *GigaScience*.
Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data.

Karyotype-specific alignment

Used **XYalign** for alignment



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

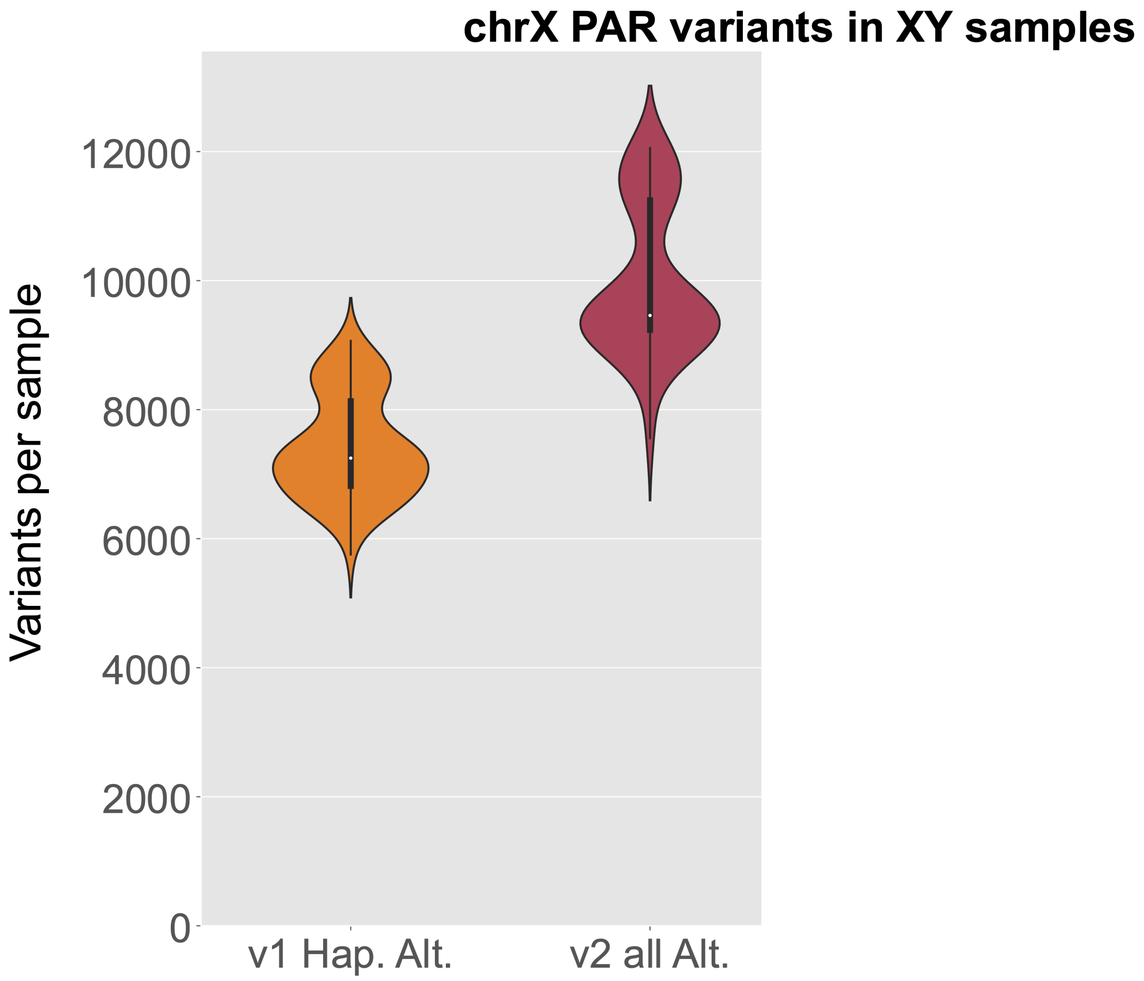
Variant Calling Pipeline

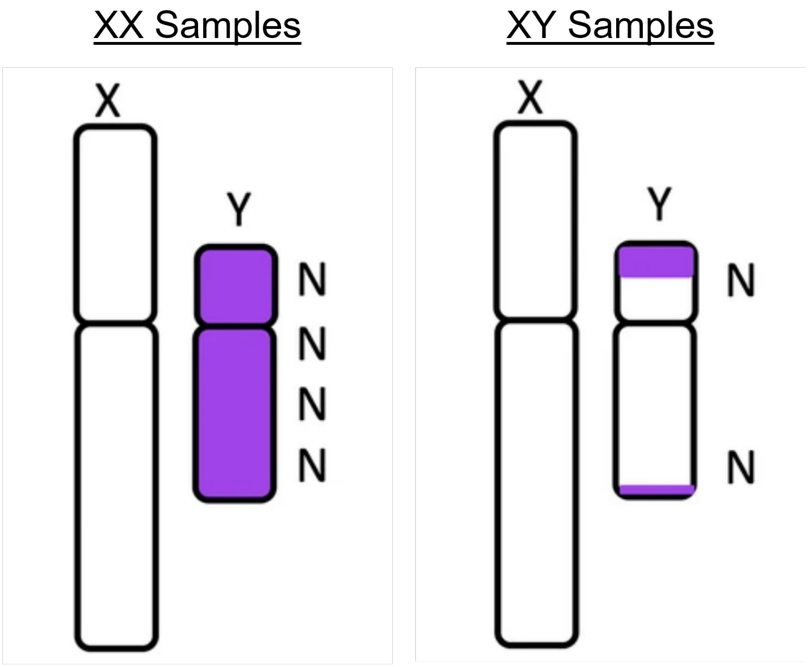
- XX samples
  - chrX: diploid
  - chrY: hard masked
- XY samples:
  - chrX PAR: diploid
  - chrX non-PAR: haploid
  - chrY non-PAR: haploid
  - chrY PAR: hard masked

Samantha Zarate

Webster, T. H., et al., 2019, *GigaScience*. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data.

Karyotype-specific alignment elucidates PAR genotypes



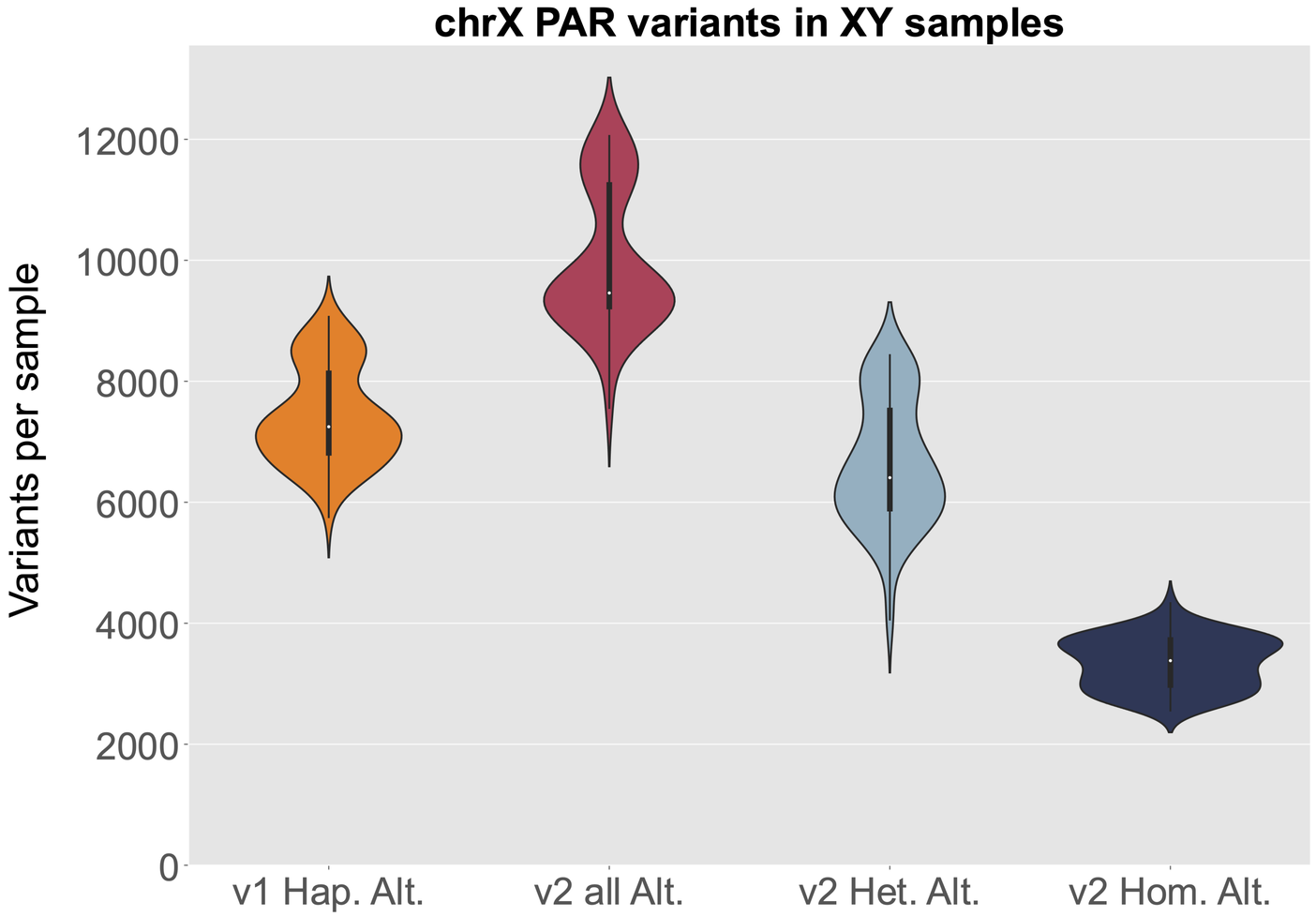


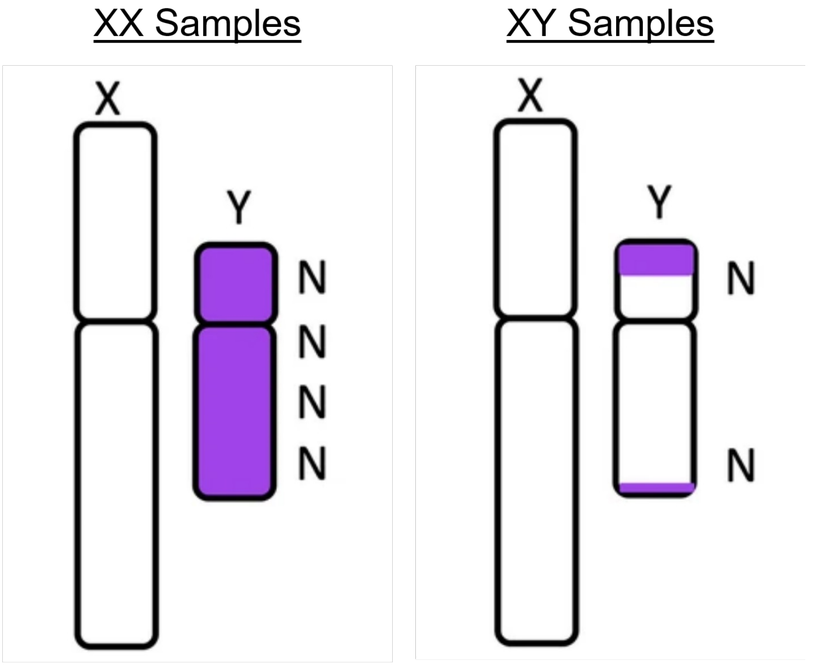
Adapted from Olney, K. C., et al. (2020) *BMC*

Samantha Zarate

Webster, T. H., et al., 2019, *GigaScience*.
Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data.

Karyotype-specific alignment elucidates PAR genotypes





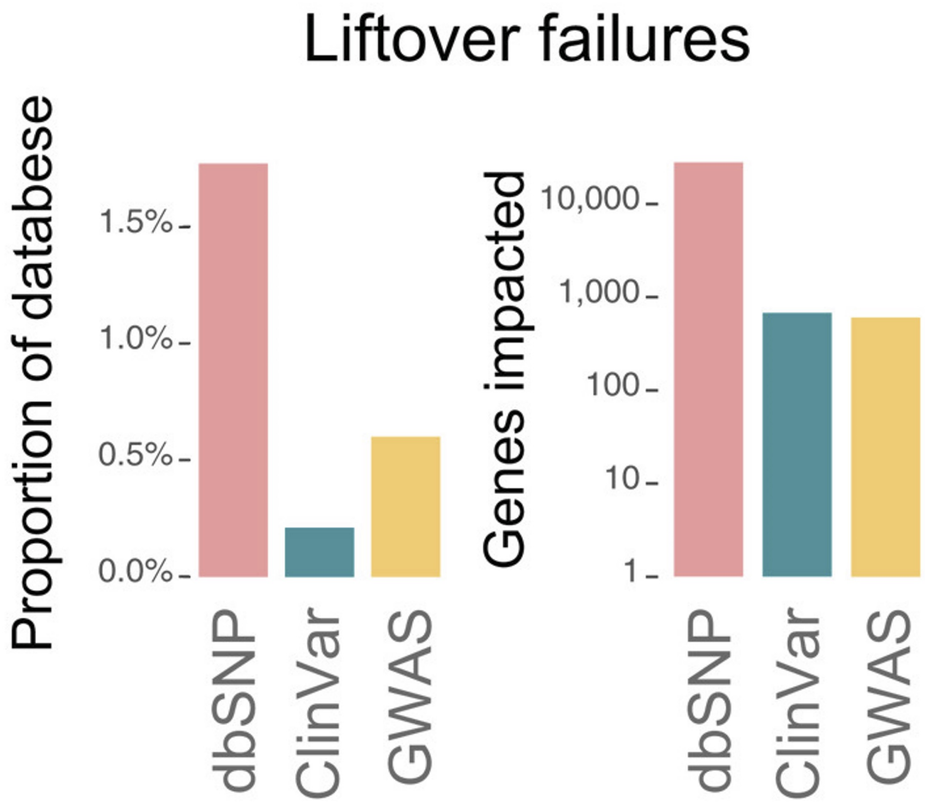
Adapted from Olney, K. C., et al. (2020) *BMC*

Samantha Zarate

Webster, T. H., et al., 2019, *GigaScience*.
Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data.

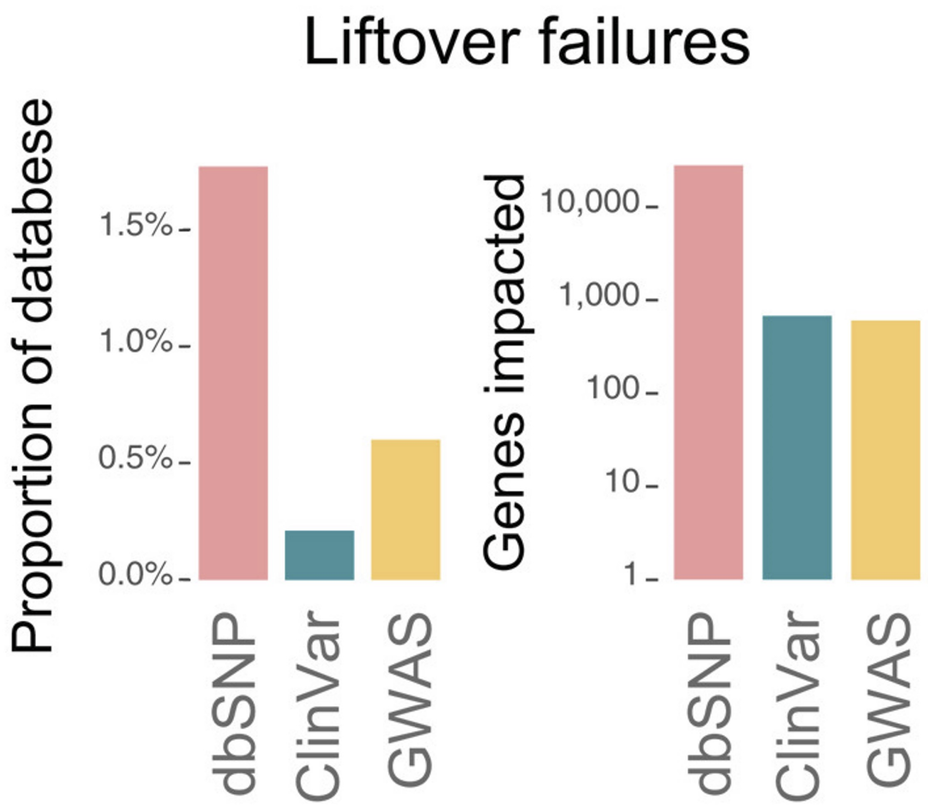
Liftover of genetic variation databases to T2T-CHM13

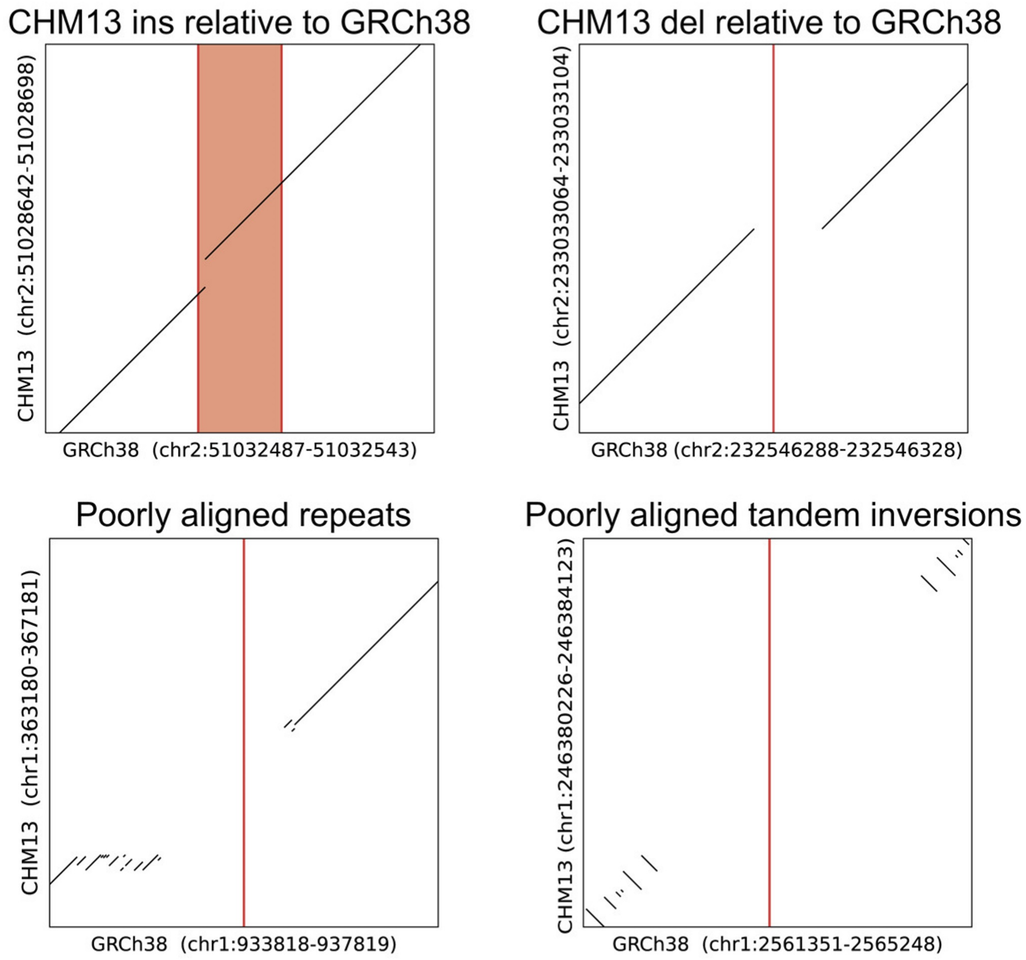
- Lifted ClinVar, NCBI dbSNP, and NHGRI-EBI GWAS Catalog to CHM13



Liftover of genetic variation databases to T2T-CHM13

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





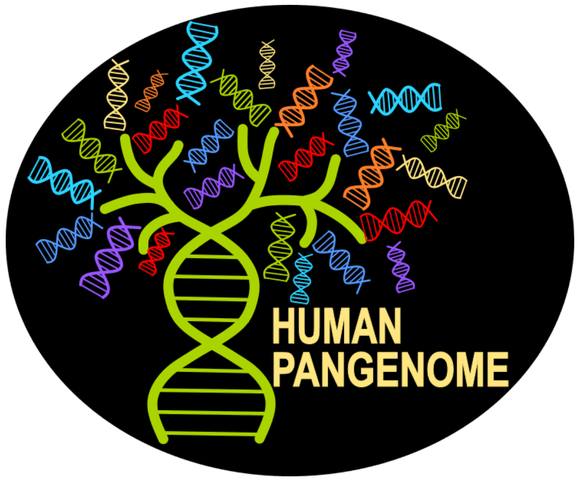
Liftover of genetic variation databases to HG002Y

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

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

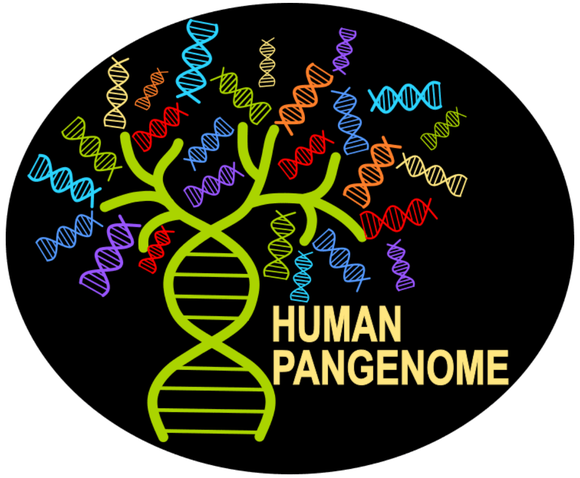
The future of Telomere-to-telomere assemblies

- 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

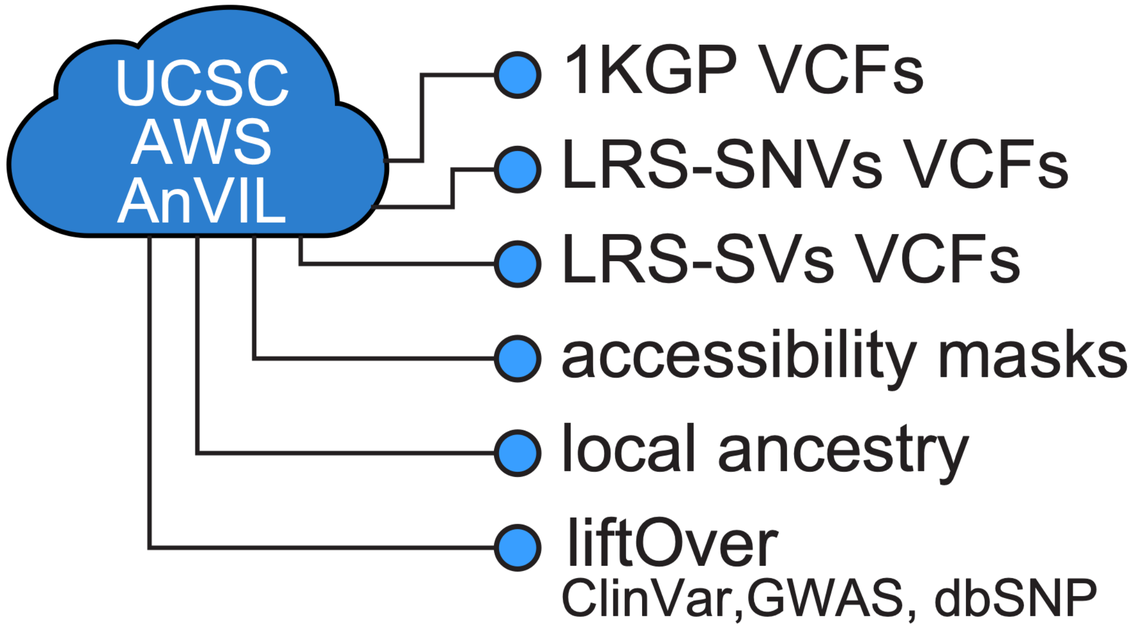


The genetic and epigenetic landscape of the *Arabidopsis* centromeres

Matthew Naish†, Michael Alonge†, Piotr Wlodzimierz†, 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*

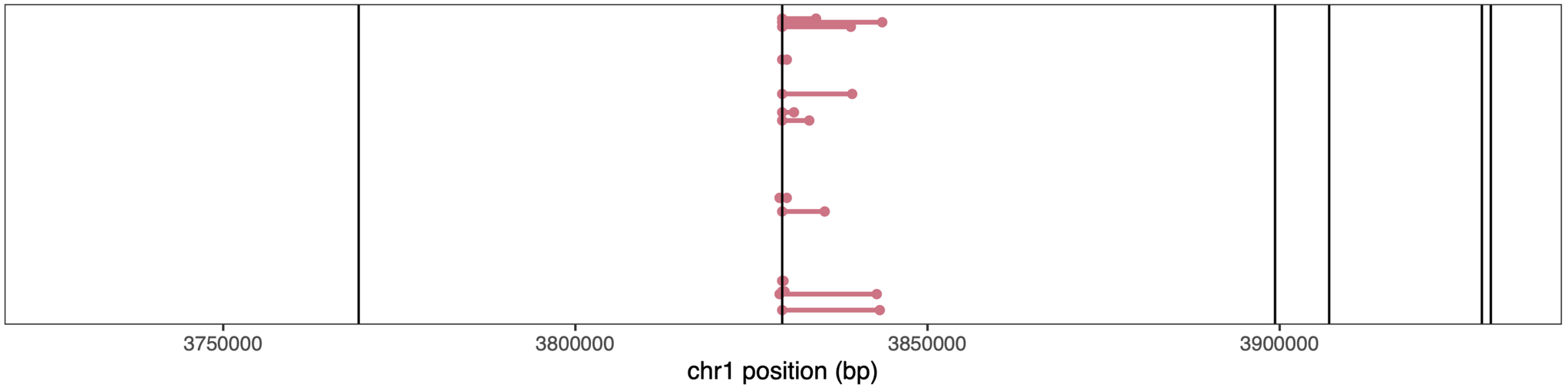
Conclusion

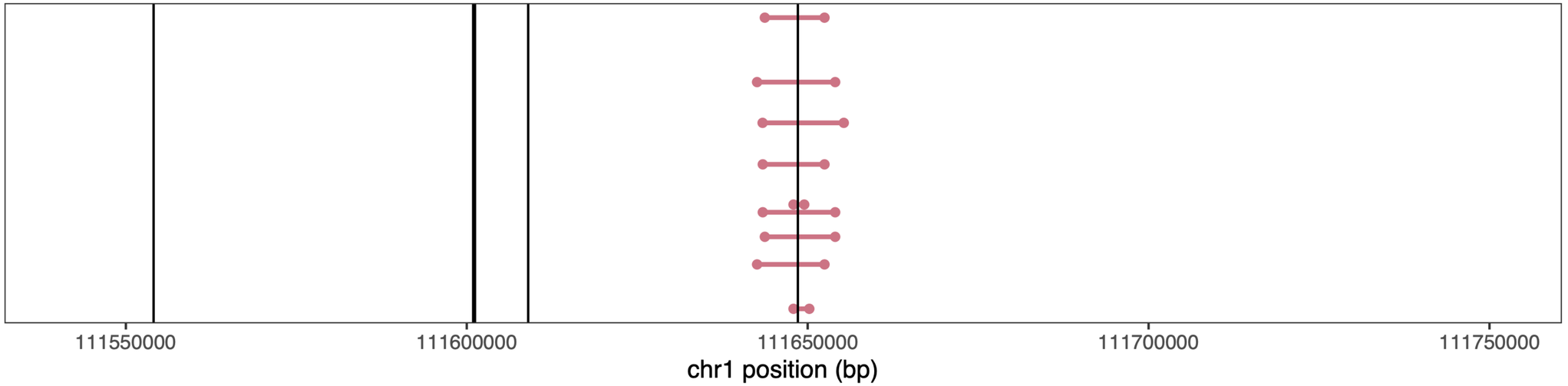
- T2T-CHM13 improves alignment and variant calling analysis across all populations
- Newly resolved regions of the genome reveal 230 Mbp of sequence for analysis for the first time
- A complete telomere-to-telomere assembly of the Y chromosome from HG002
- T2T has generated resources for using T2T-CHM13v2.0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



extra slides

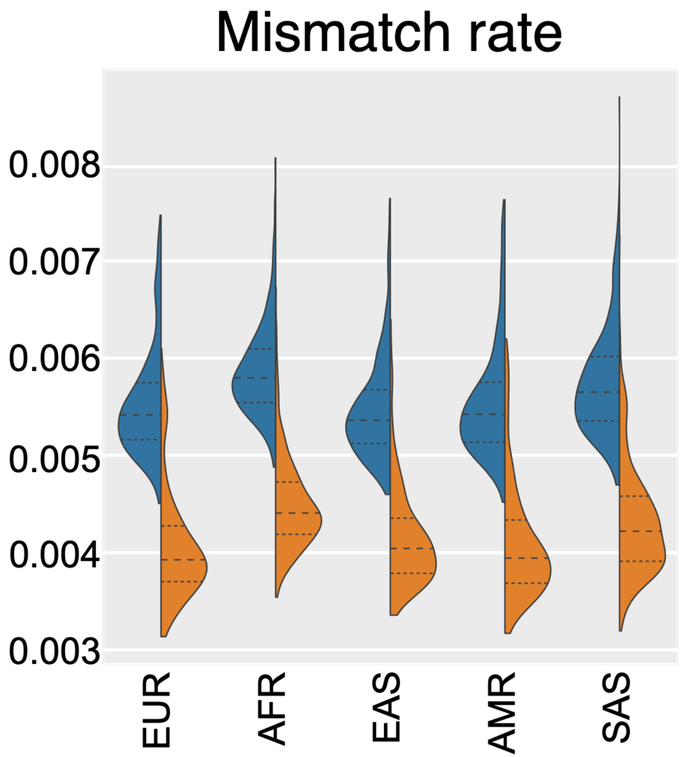
GRCh38 contains artificial haplotype boundaries

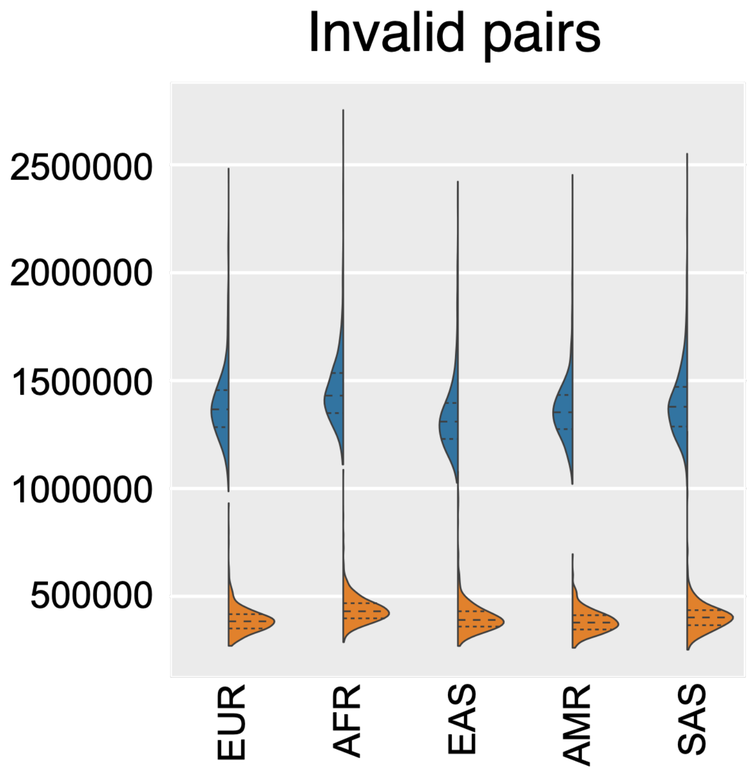


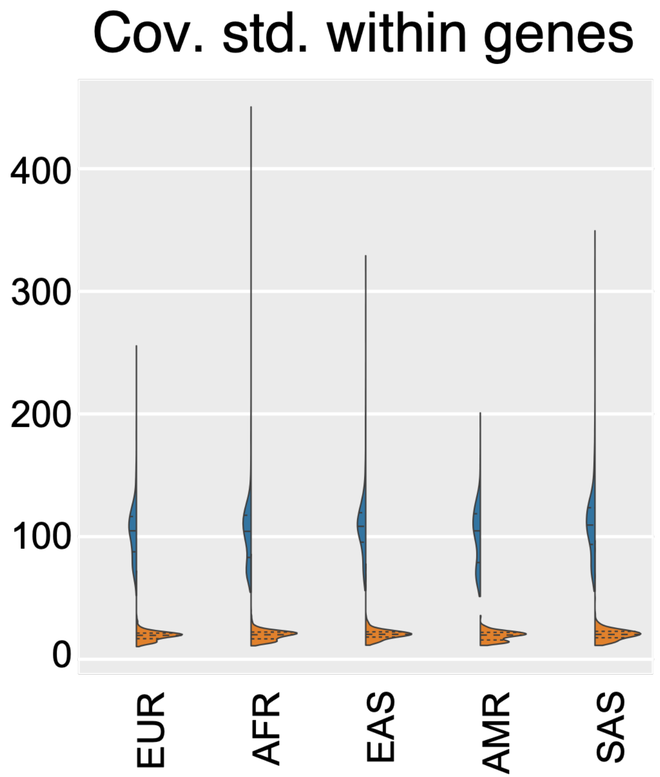


Rajiv McCoy

T2T-CHM13 improves short-read alignment





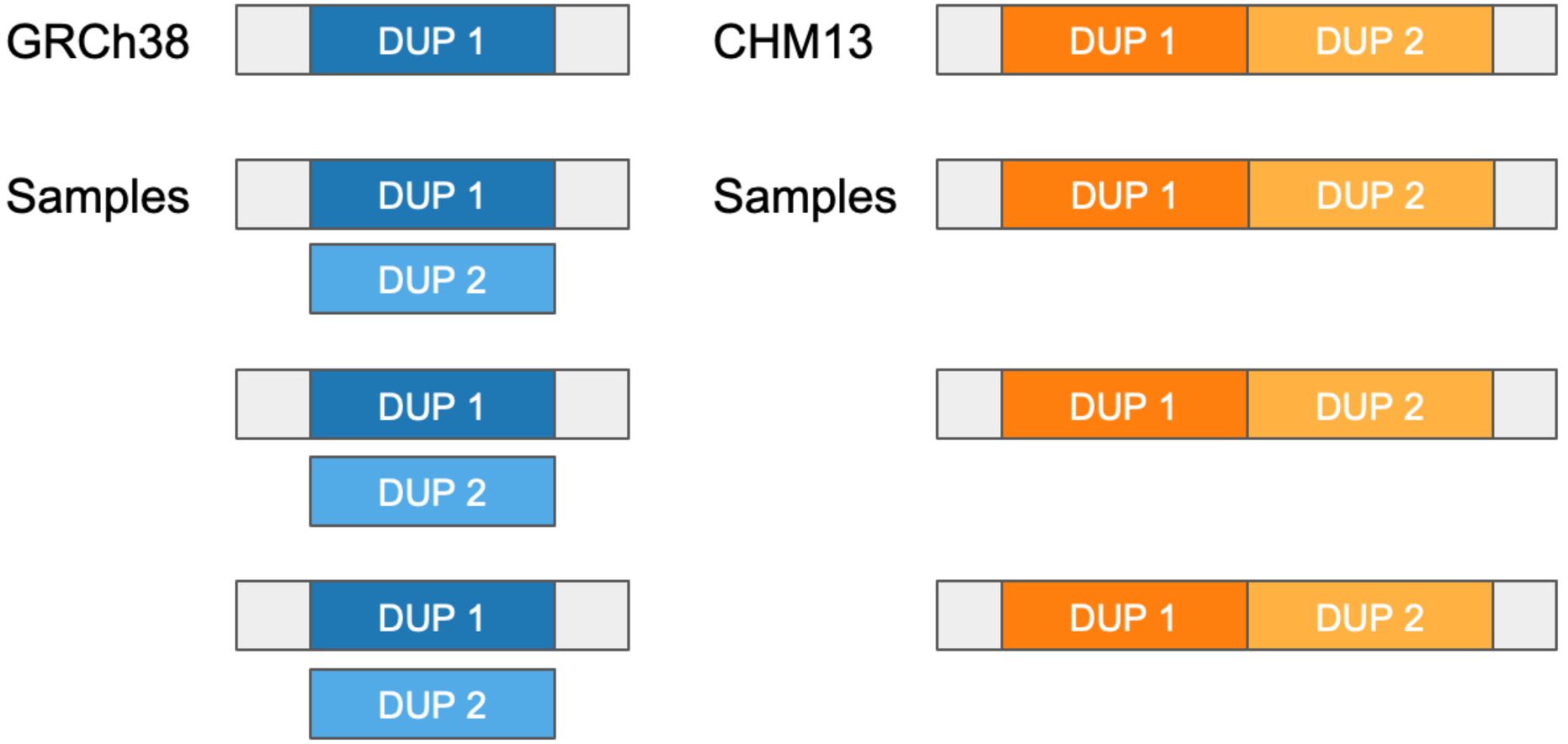


Sergey Aganezov

GRCh38 CHM13

T2T-CHM13 improves short-read variant calling

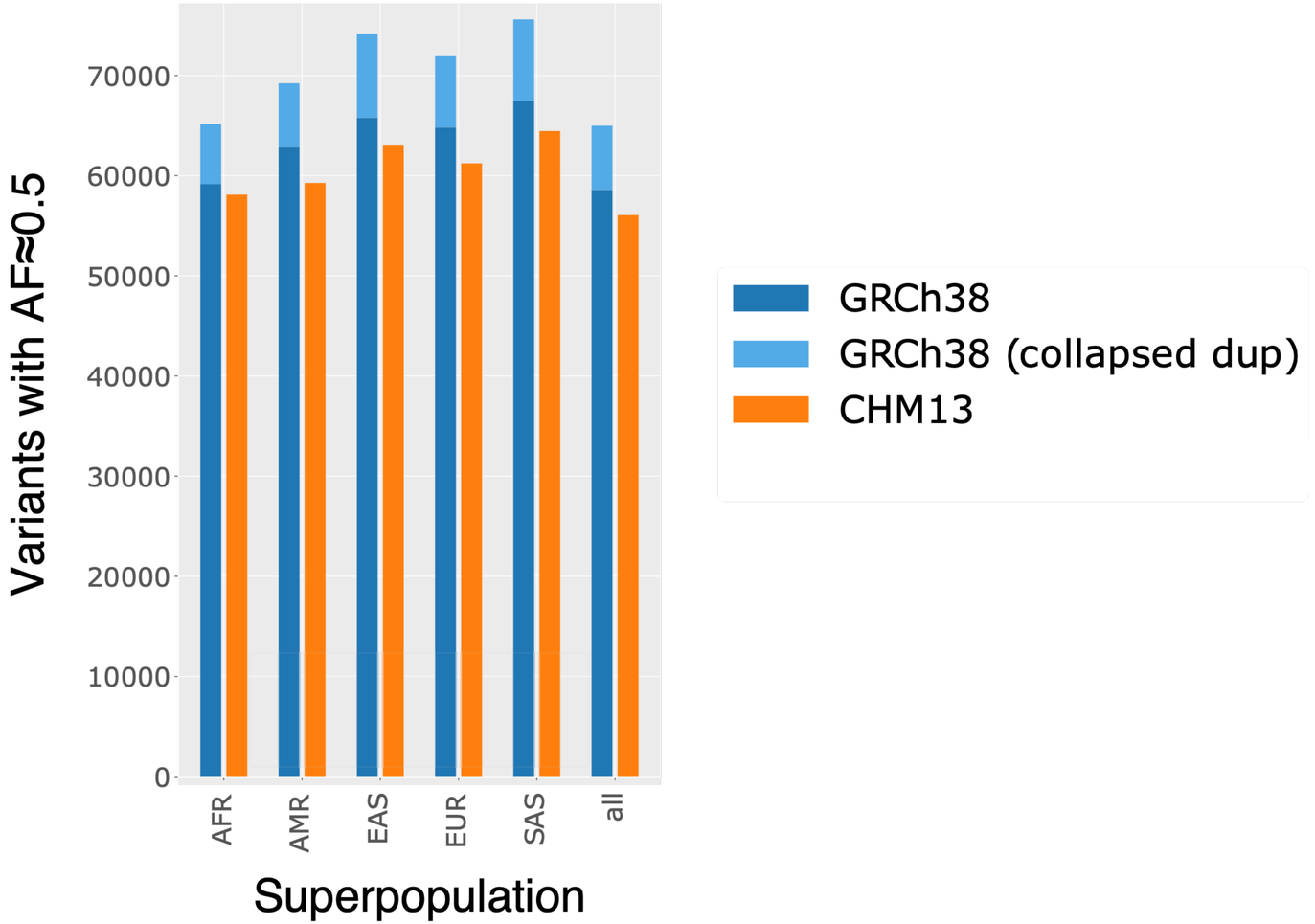
- Fewer false heterozygous variants from GRCh38 collapsed duplications



Samantha Zarate

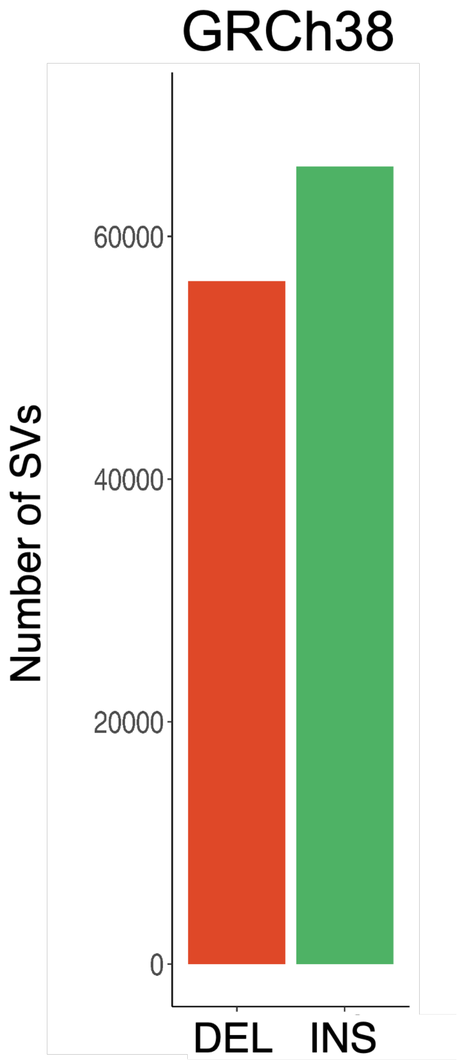
T2T-CHM13 improves short-read variant calling

- Fewer spurious heterozygous variants from GRCh38 collapsed duplications



Samantha Zarate

T2T-CHM13 fixes the insertion-deletion misbalance in GRCh38

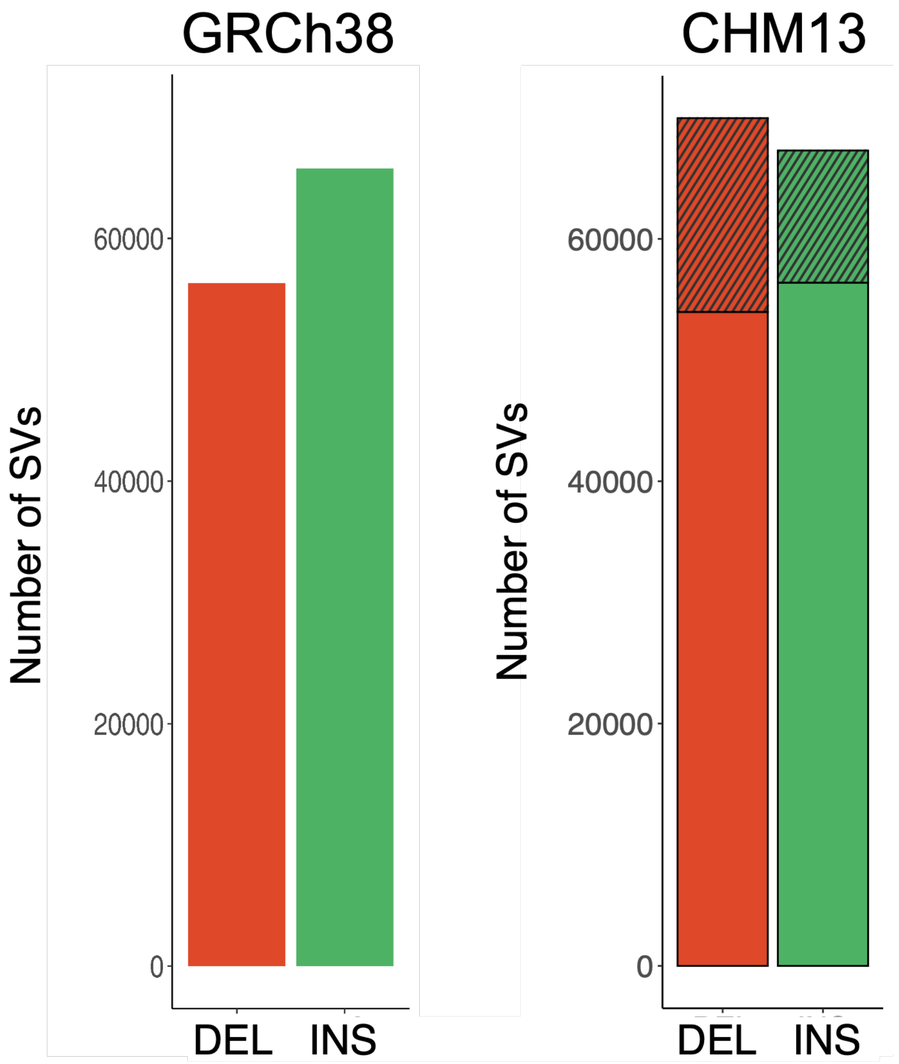


- Excess GRCh38 insertions represent missing or incomplete sequences

Melanie Kirsche

Sergey Aganezov

T2T-CHM13 fixes the insertion-deletion misbalance in GRCh38



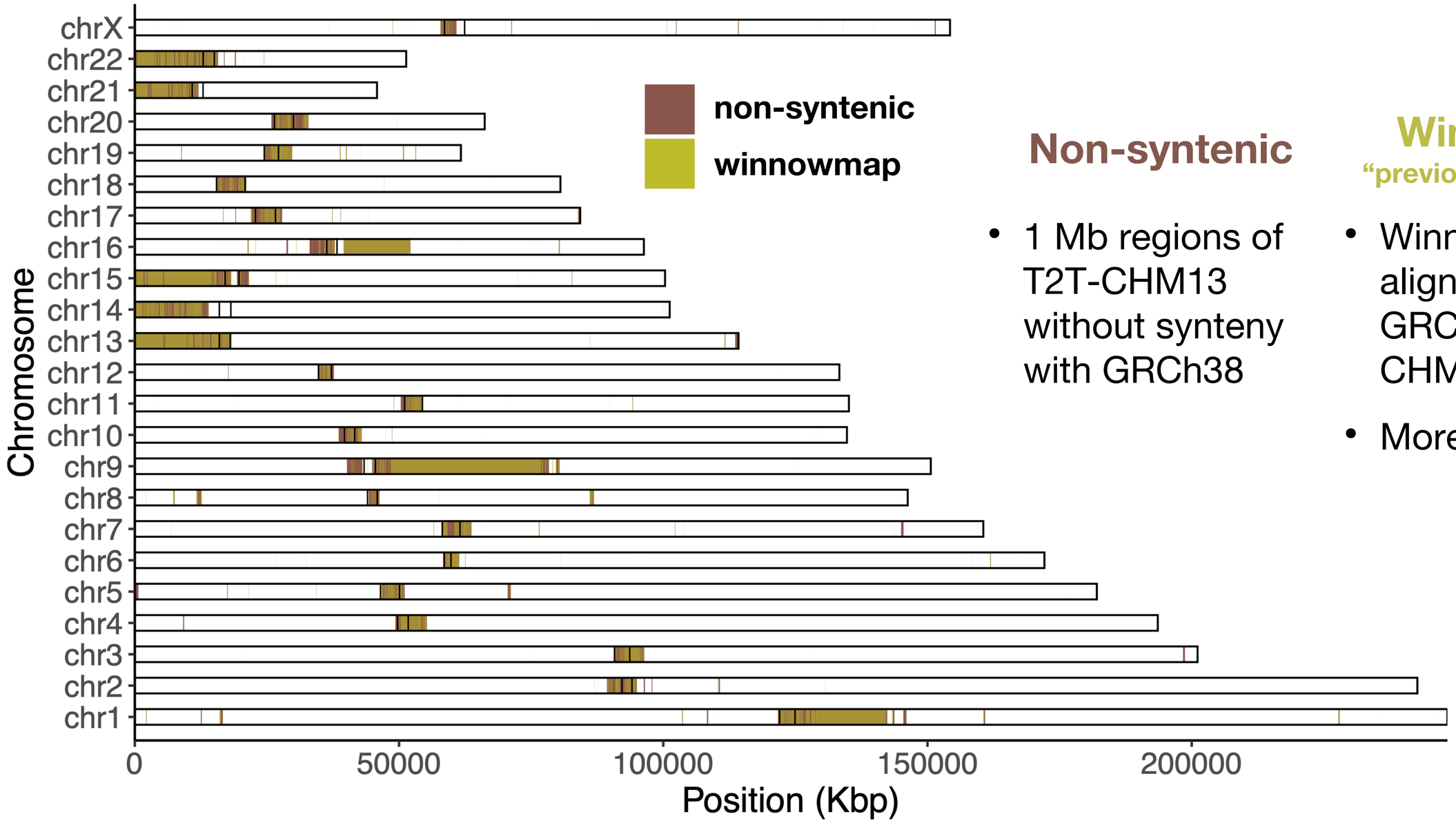
- Excess GRCh38 insertions represent missing or incomplete sequences

sequences only in CHM13

Melanie Kirsche

Sergey Aganezov

Identifying new sequences revealed by T2T-CHM13



Non-syntenic

- 1 Mb regions of T2T-CHM13 without synteny with GRCh38

Winnowmap

"previously unresolved"

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

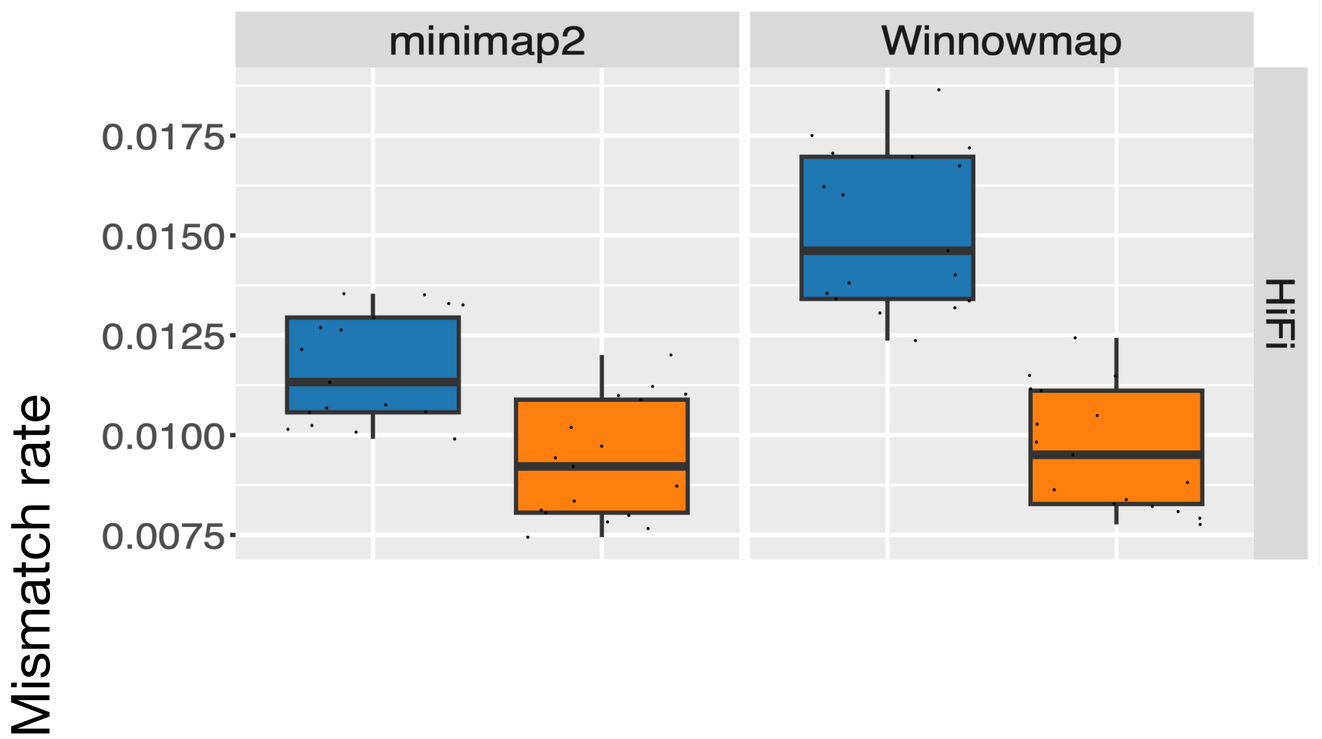
Long-read alignment and variant calling with T2T-CHM13

- 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

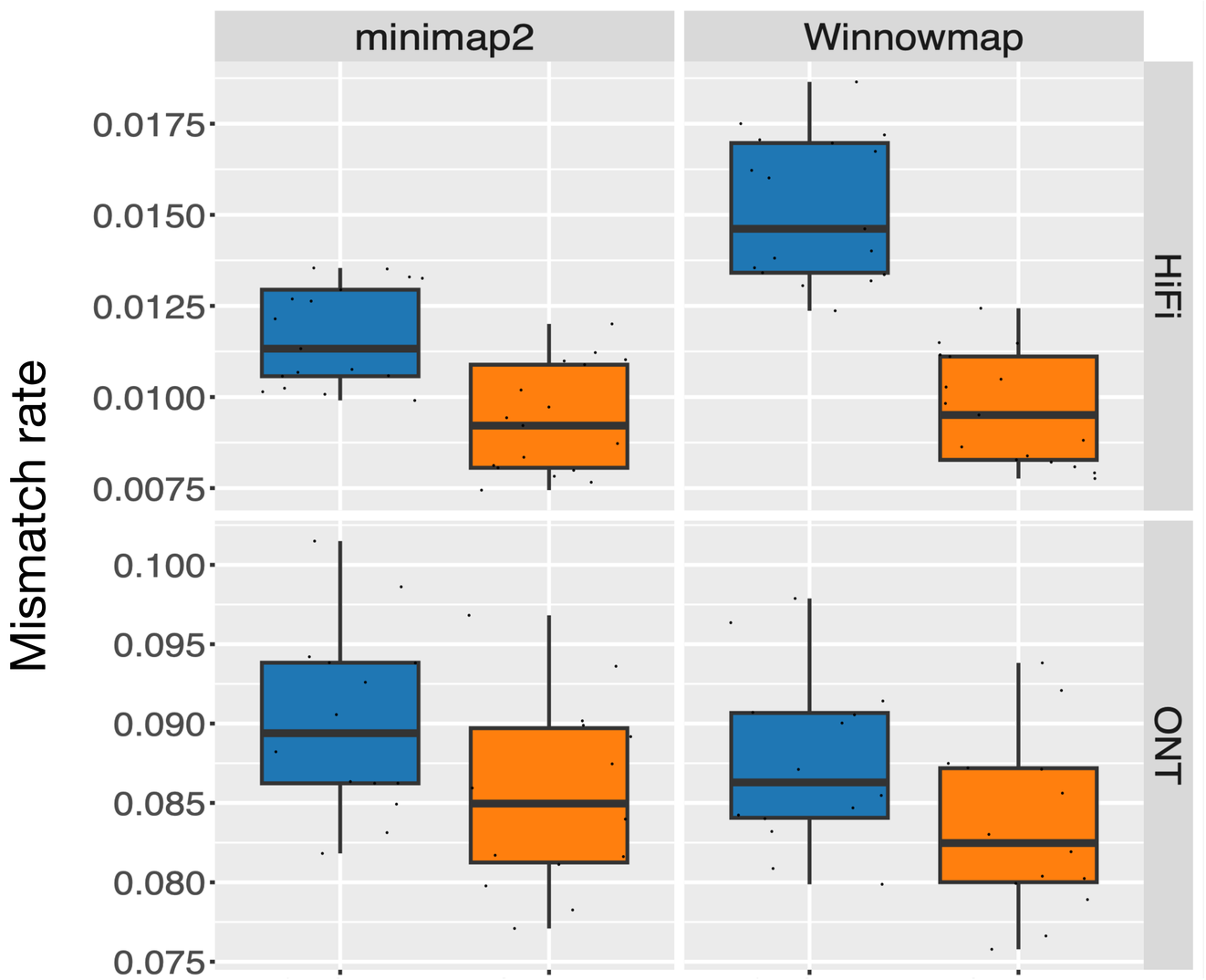
T2T-CHM13 lowers alignment mismatch rates for long reads



Melanie Kirsche

Sergey Aganezov

T2T-CHM13 lowers alignment mismatch rates for long reads



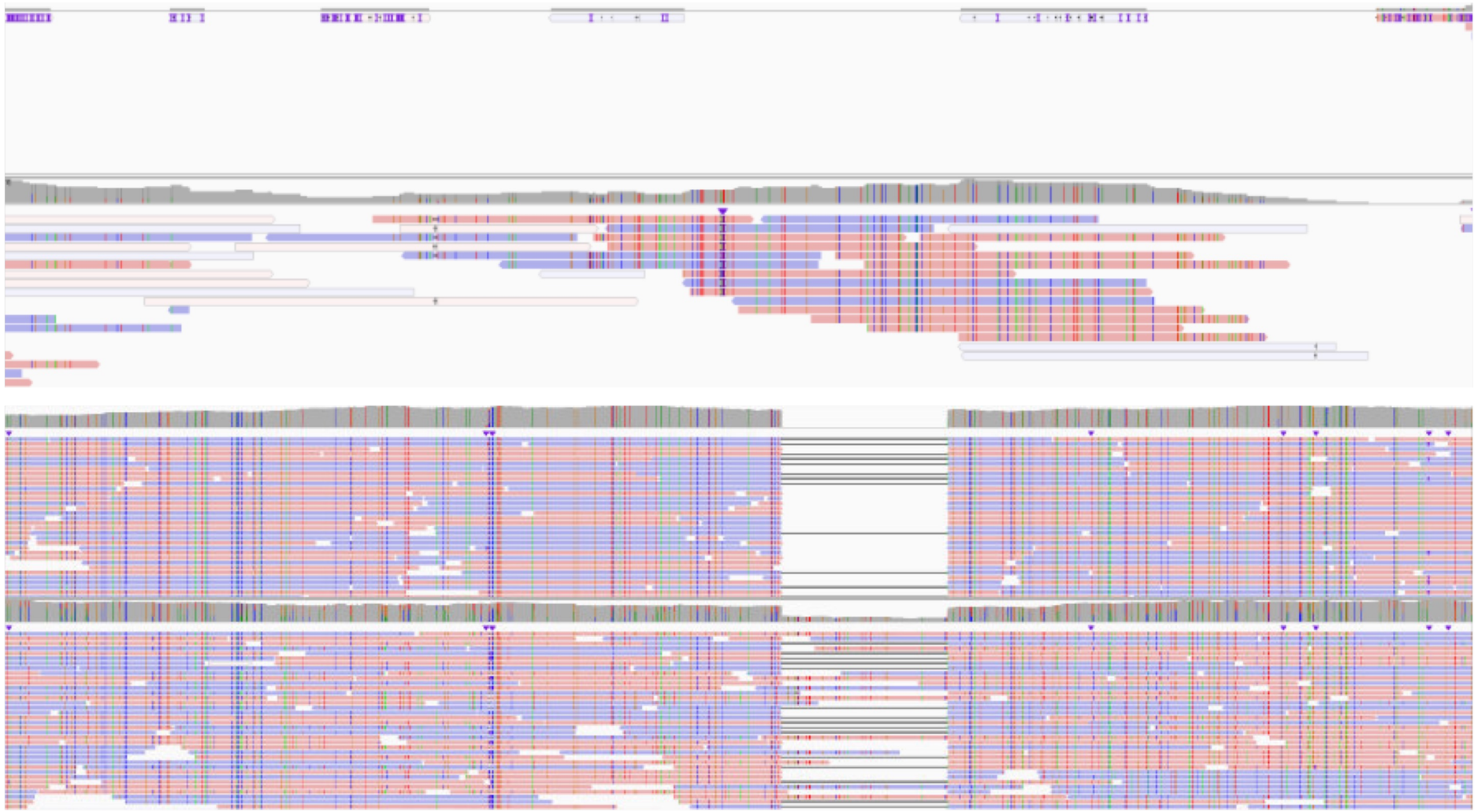
Melanie Kirsche

Sergey Aganezov

T2T-CHM13 improves resolution of structural variants

GRCh38

CHM13

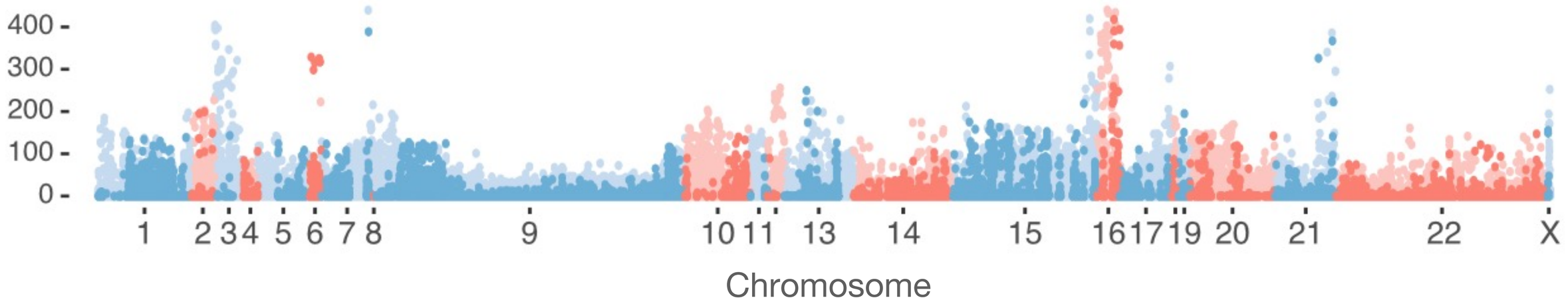


Melanie Kirsche

Sergey Aganezov

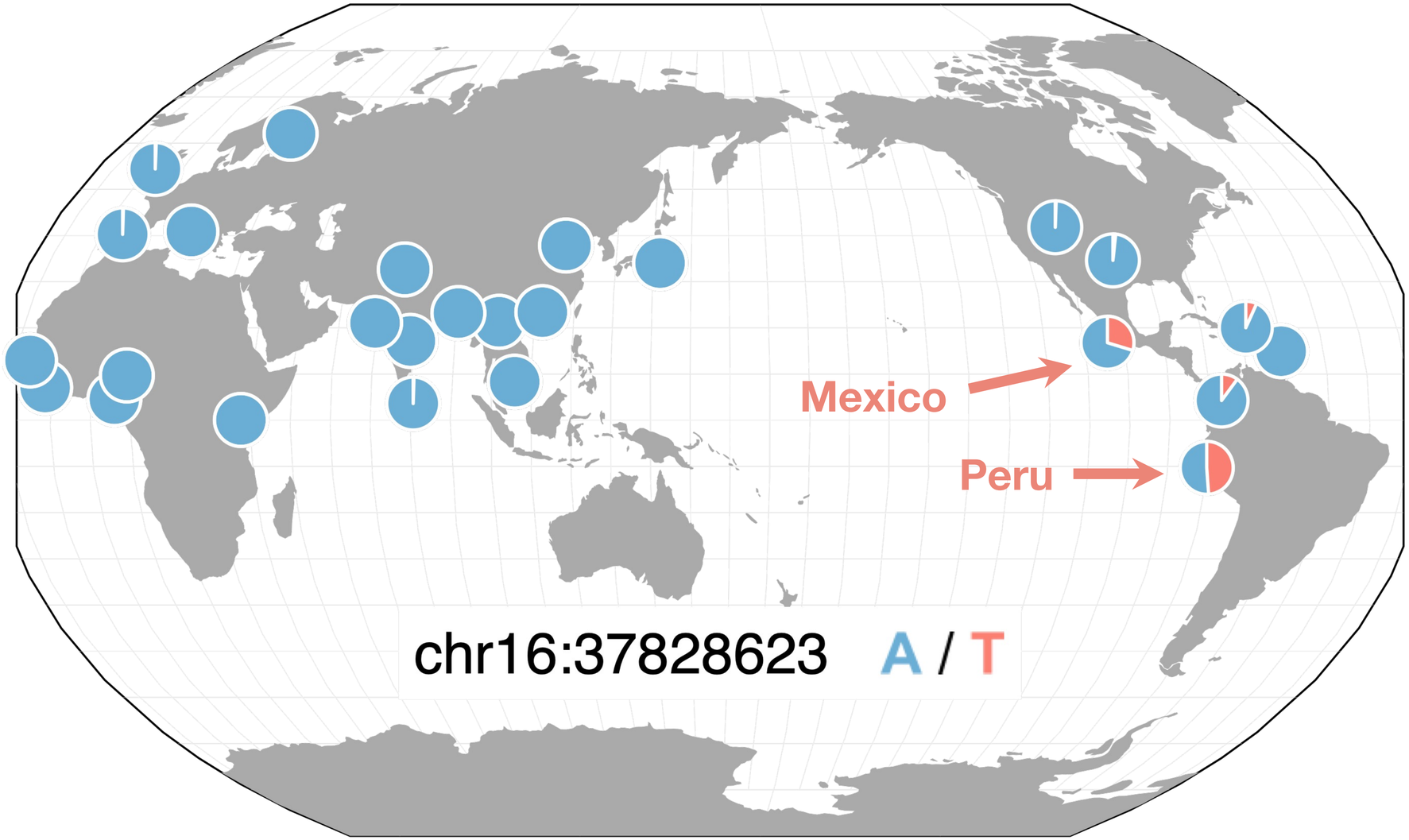
Evolutionary signatures in novel regions of T2T-CHM13

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



Cheng *et al*., 2022, *MBE*. Detecting selection in multiple populations by modeling ancestral admixture components.

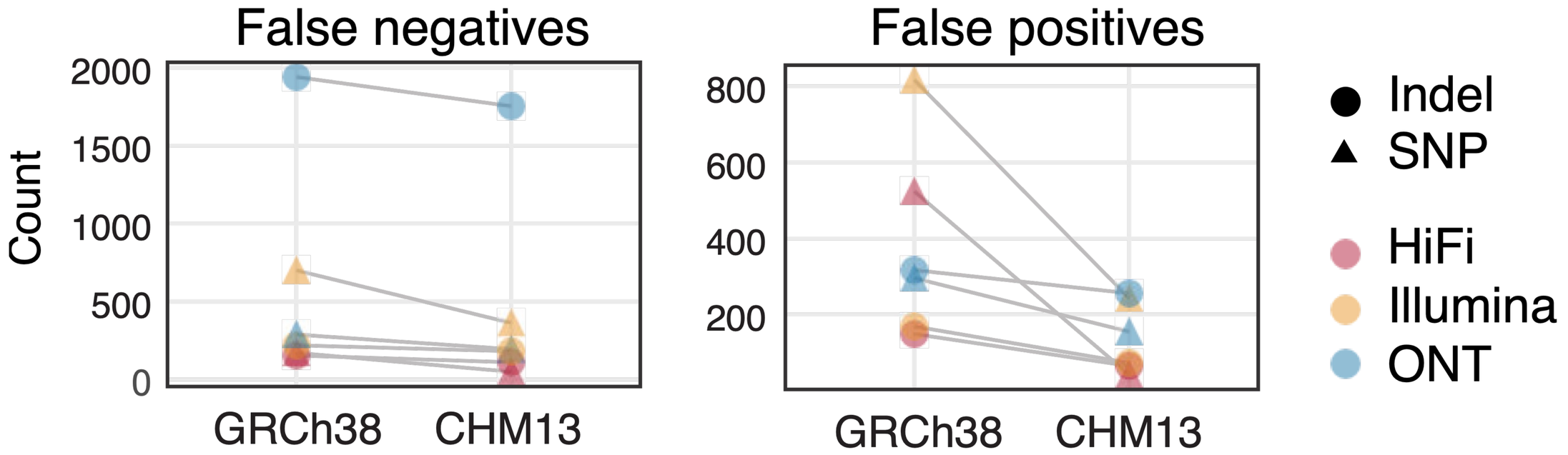
Evolutionary signatures in novel regions of T2T-CHM13

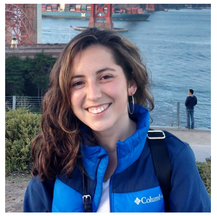


Located in a centromeric alpha satellite on chr16

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