JOHNS HOPKINS
UNIVERSITY

JOHNS HOPKINS UNIVERSITY
CENTER FOR COMPUTATIONAL BIOLOGY
CCB

Assembly of a pan-genome from deep sequencing of 910 individuals of African descent

Rachel M Sherman

Johns Hopkins University

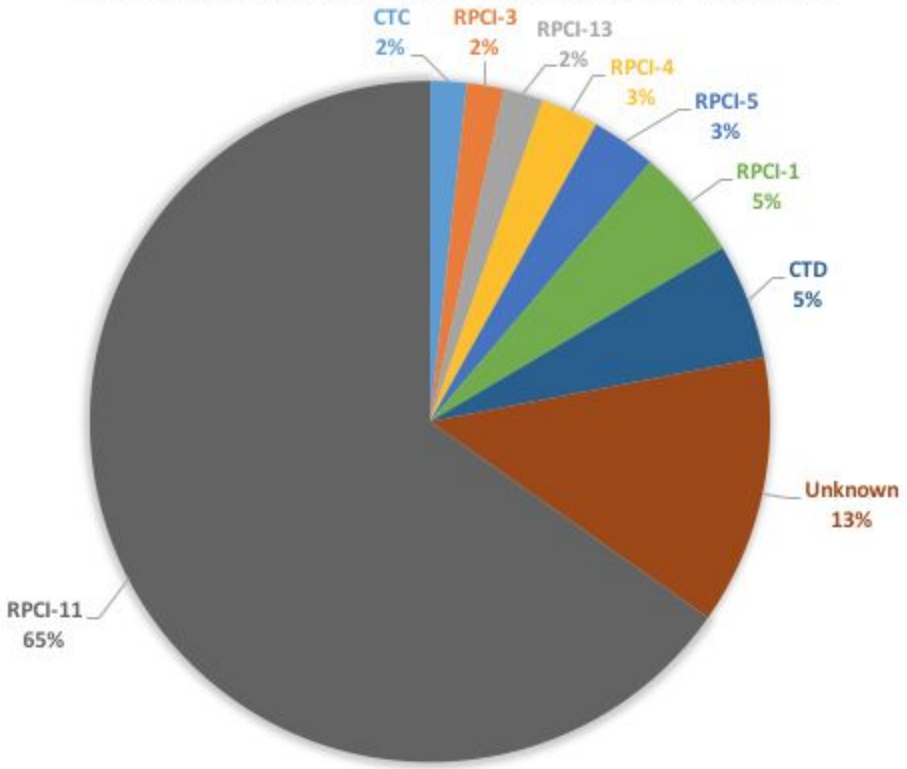
Salzberg Lab

RECOMB 2019

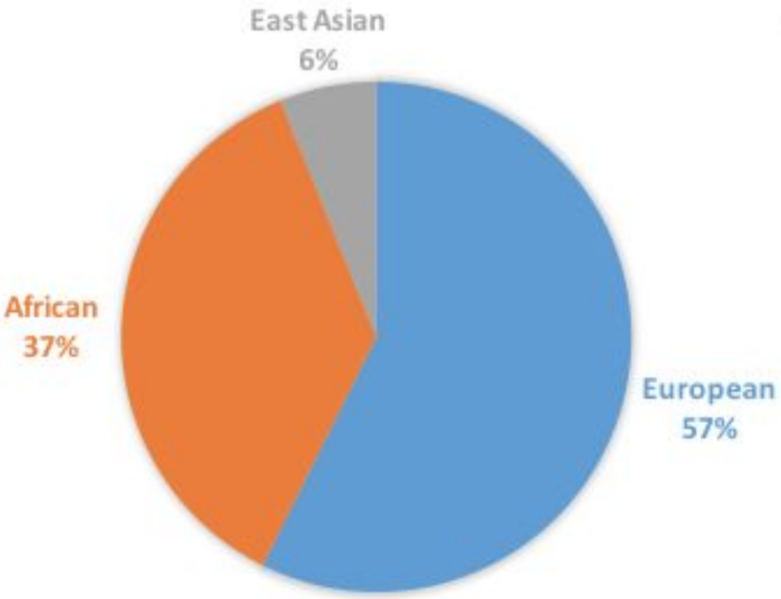
@rshermanjhu

Human reference genome makeup

The majority of the reference is from one individual



Source of BAC clones comprising the reference genome

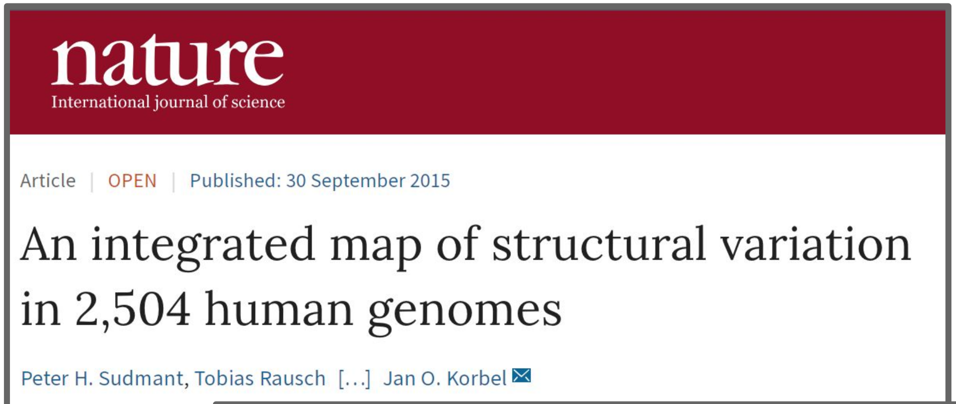


Rough ancestry inference of known reference BACs

Green *et al* (2010). Science.

Slide from Rachel Sherman, RECOMB 2019

Capturing human genetic diversity



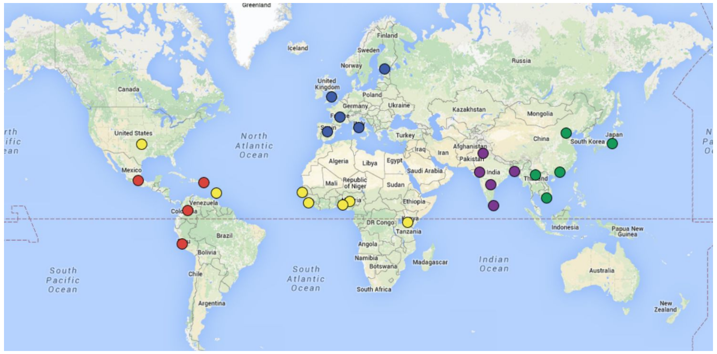
nature
International journal of science

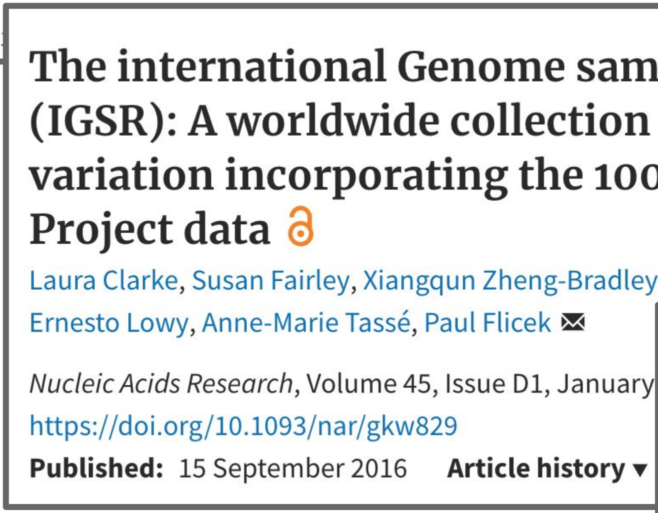
Article | OPEN | Published: 30 September 2015

An integrated map of structural variation in 2,504 human genomes

Peter H. Sudmant, Tobias Rausch [...] Jan O. Korbel ✉

Nature **526**, 75–81 (0





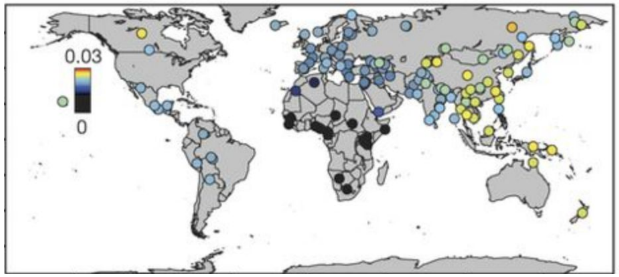
The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data

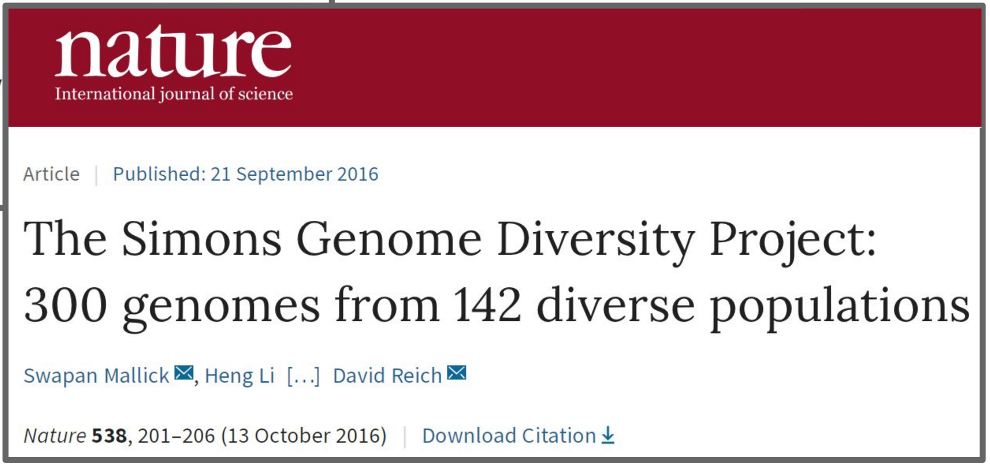
Laura Clarke, Susan Fairley, Xiangqun Zheng-Bradley, Ian Streeter, Emily Perry, Ernesto Lowy, Anne-Marie Tassé, Paul Flicek ✉

Nucleic Acids Research, Volume 45, Issue D1, January

https://doi.org/10.1093/nar/gkw829

Published: 15 September 2016 **Article history ▾**





nature
International journal of science

Article | Published: 21 September 2016

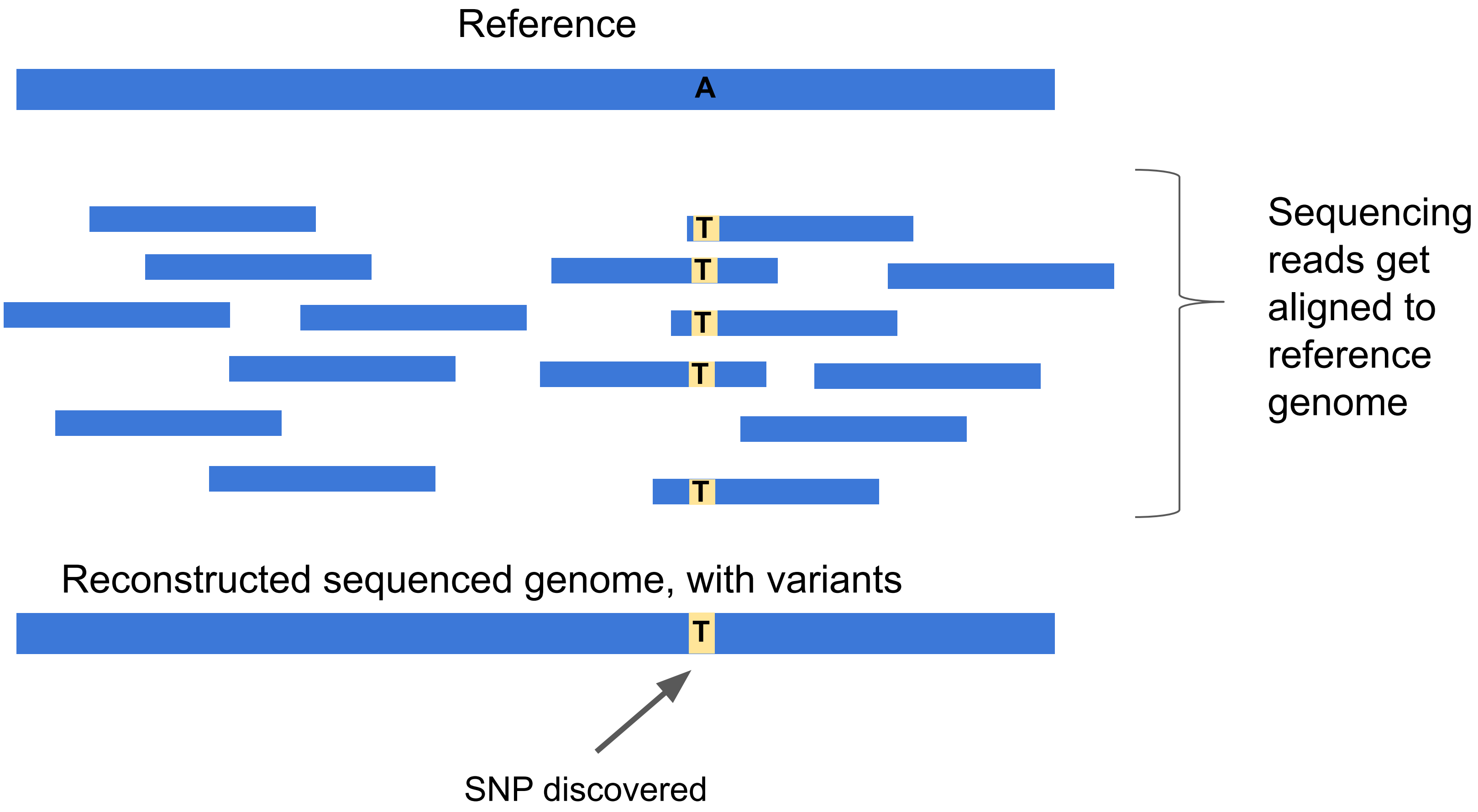
The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

Swapan Mallick ✉, Heng Li [...] David Reich ✉

Nature **538**, 201–206 (13 October 2016) | Download Citation ⤓

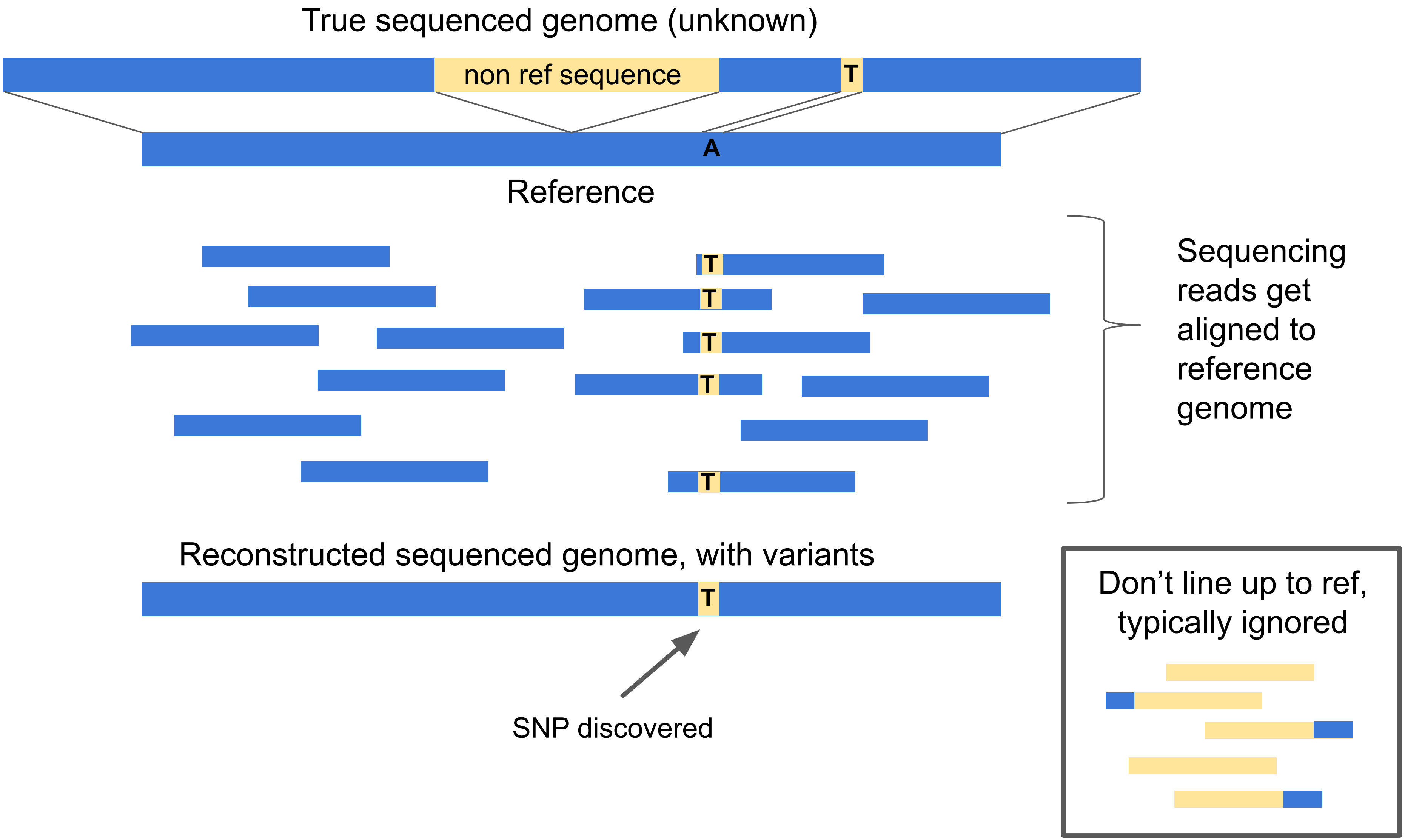
Slide from Rachel Sherman, RECOMB 2019

Variant discovery via alignment


Reference
A
T
T
T
T
T
Sequencing reads get aligned to reference genome
Reconstructed sequenced genome, with variants
T
SNP discovered

Slide from Rachel Sherman, RECOMB 2019

Sequences missed by alignment



True sequenced genome (unknown)

Reference

Sequencing reads get aligned to reference genome

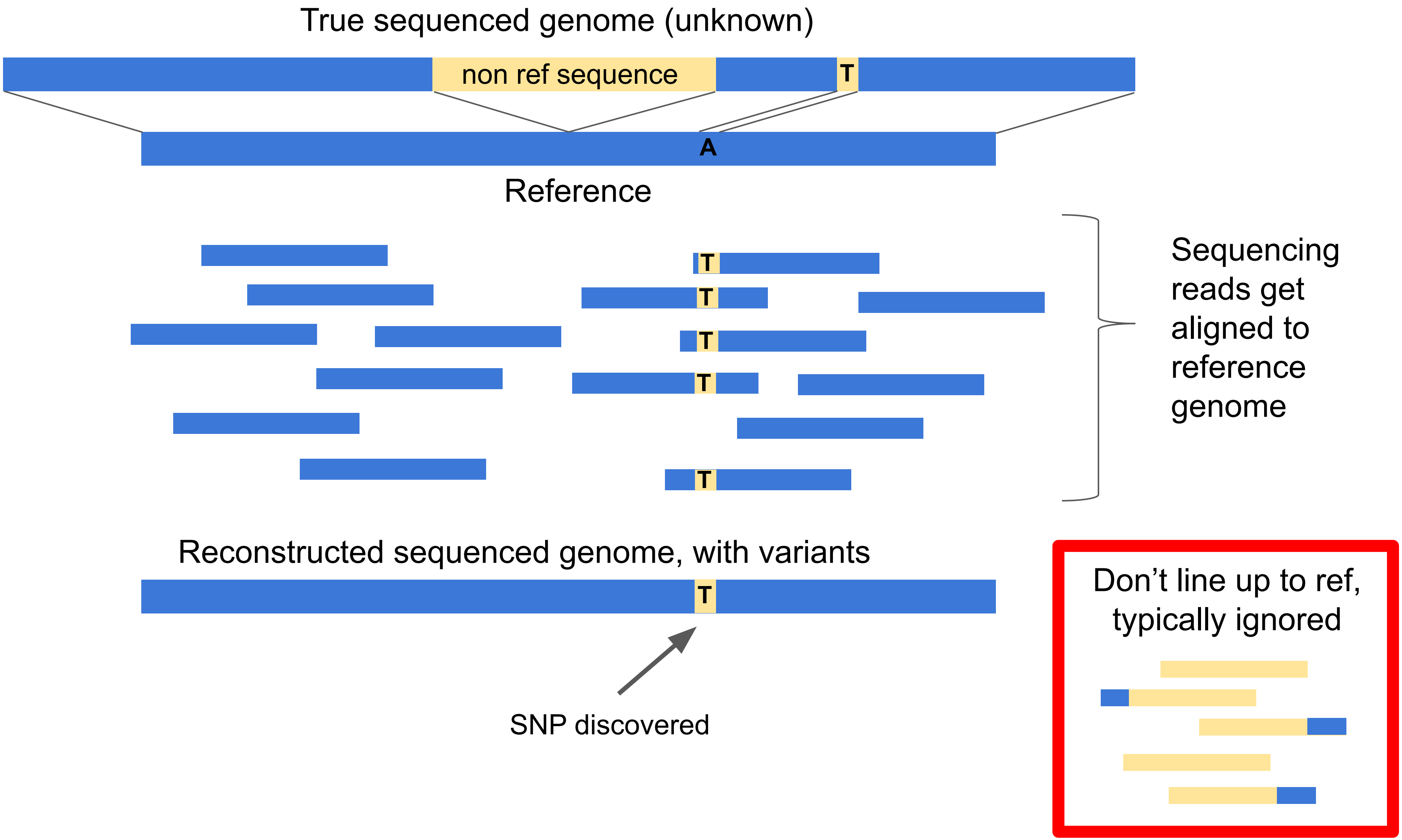
Reconstructed sequenced genome, with variants

SNP discovered

Don’t line up to ref, typically ignored

Slide from Rachel Sherman, RECOMB 2019

Sequences missed by alignment



True sequenced genome (unknown)

Reference

Sequencing reads get aligned to reference genome

Reconstructed sequenced genome, with variants

SNP discovered

Don’t line up to ref, typically ignored

African-ancestry population WGS data

Cohort	Number of Samples
African American (Atlanta)	50
African American (Baltimore-DC)	50
African American (Chicago)	50
African American (Detroit)	50
African American (Jackson, MS)	50
African American (Nashville)	48
African American (NYC)	48
African American (San Francisco)	50
African American (Winston-Salem)	50
Barbados	49
Brazil	47
Colombia	50
Dominican Republic	47
Gabon	34
Honduras	50
Jamaica	50
Palenque	34
Nigeria	50
Puerto Rico	53

Consortium on Asthma among African-ancestry Populations in the Americas

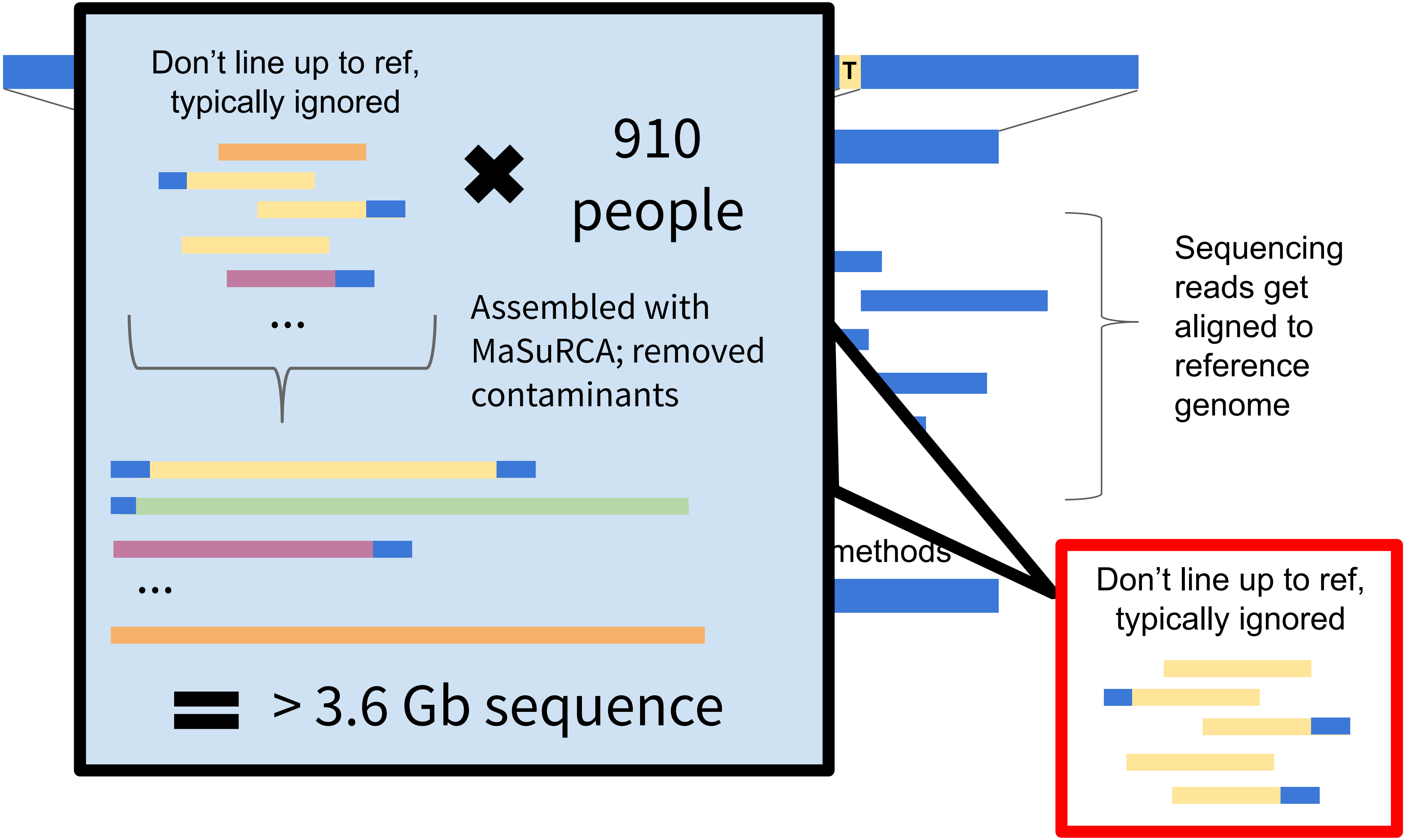
CAAPA

Data was collected from 19 distinct cohorts across the Americas, the Caribbean, and Africa resulting in 910 analyzed samples.

Sherman *et al* (2019). Nature Genetics.

Slide from Rachel Sherman, RECOMB 2019

Sequences missed by alignment



Don’t line up to ref, typically ignored

✖ 910 people

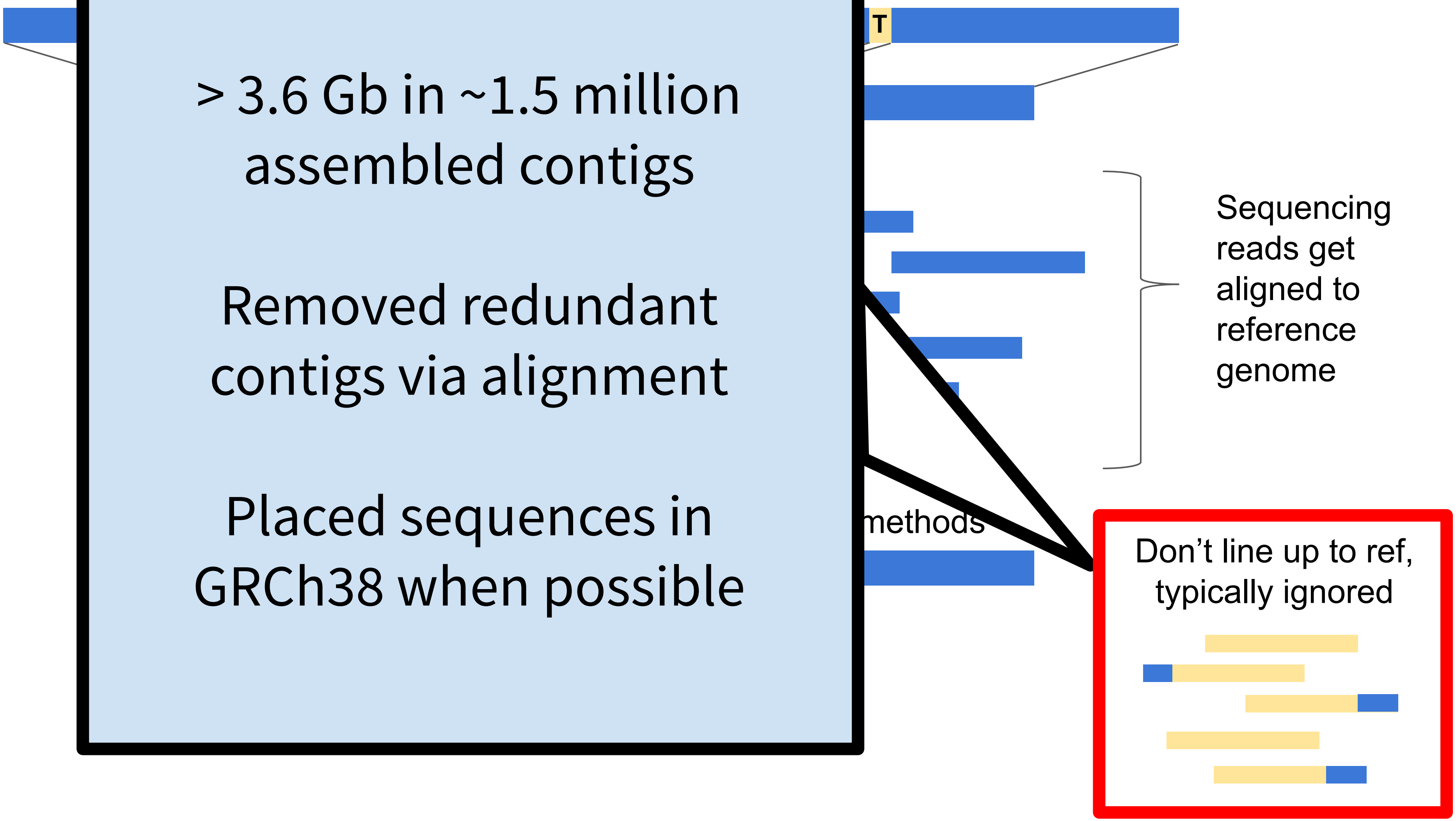
Assembled with MaSuRCA; removed contaminants

= > 3.6 Gb sequence

Sequencing reads get aligned to reference genome

Don’t line up to ref, typically ignored

Sequences missed by alignment



> 3.6 Gb in ~1.5 million assembled contigs

Removed redundant contigs via alignment

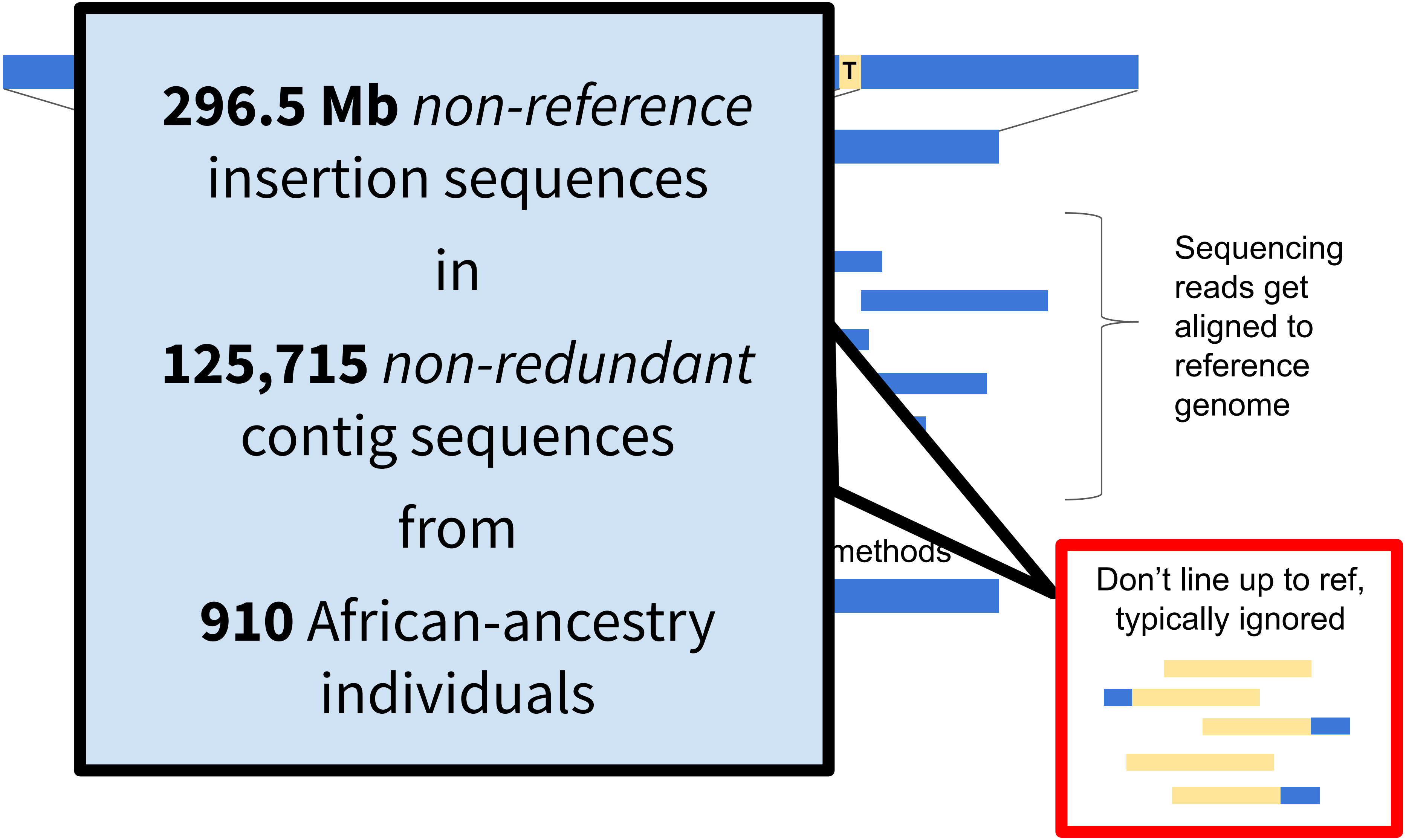
Placed sequences in GRCh38 when possible

Sequencing reads get aligned to reference genome

Don't line up to ref, typically ignored

Slide from Rachel Sherman, RECOMB 2019

Sequences missed by alignment



296.5 Mb *non-reference* insertion sequences

in

125,715 *non-redundant* contig sequences

from

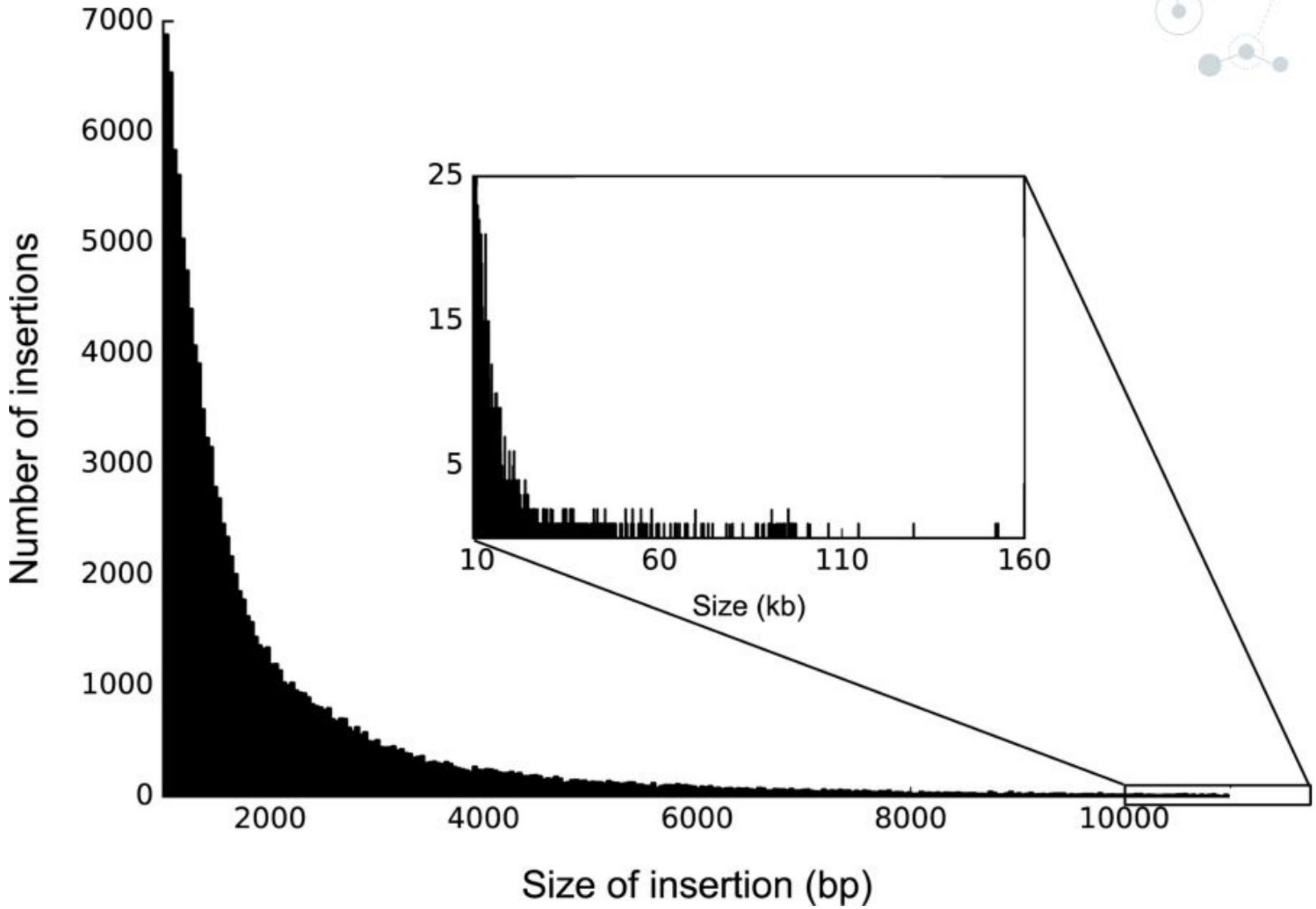
910 African-ancestry individuals

Sequencing reads get aligned to reference genome

Don’t line up to ref, typically ignored

Slide from Rachel Sherman, RECOMB 2019

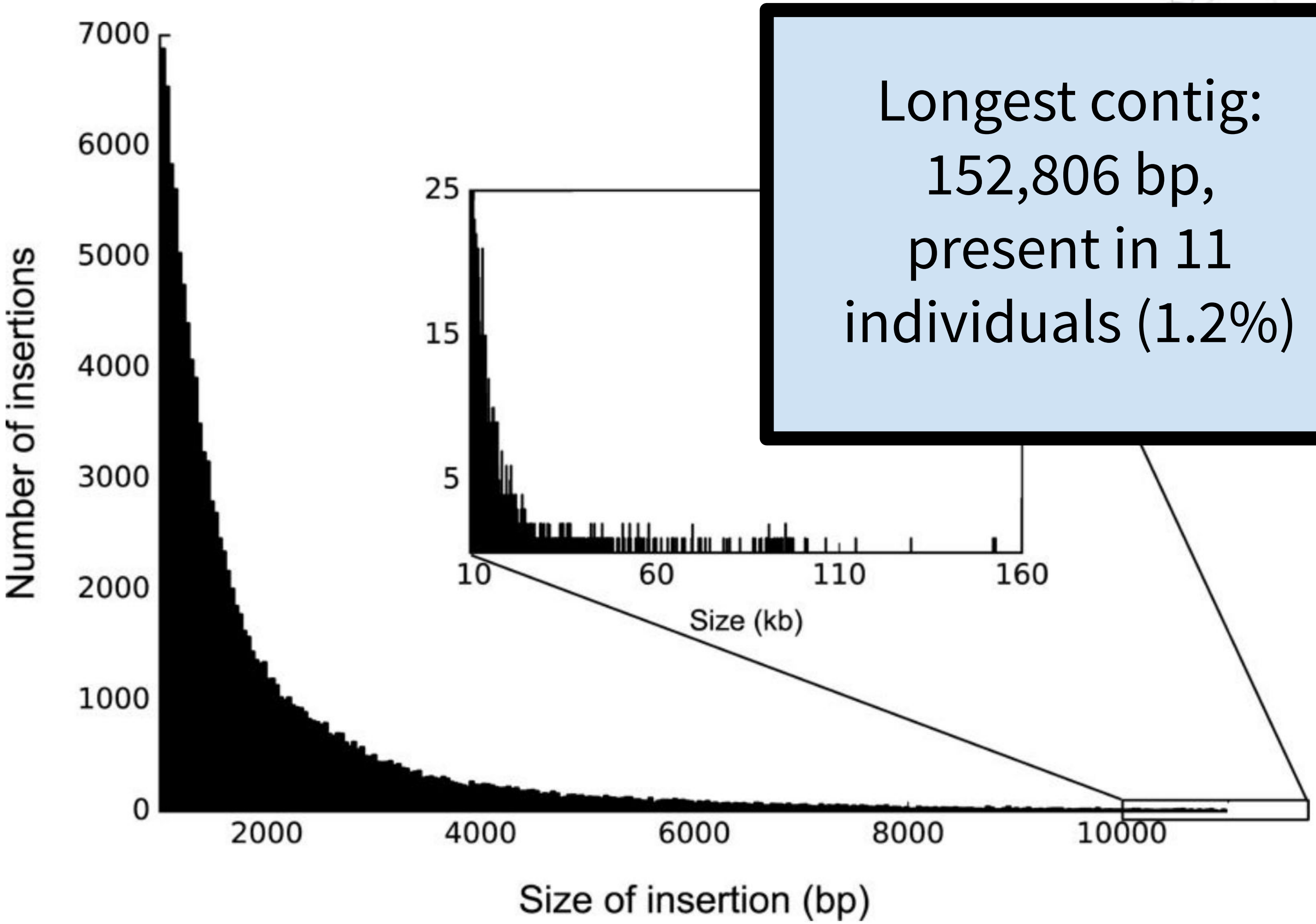
Pan-genome contig size distribution



Sherman *et al* (2019).

Slide from Rachel Sherman, RECOMB 2019

Pan-genome contig size distribution

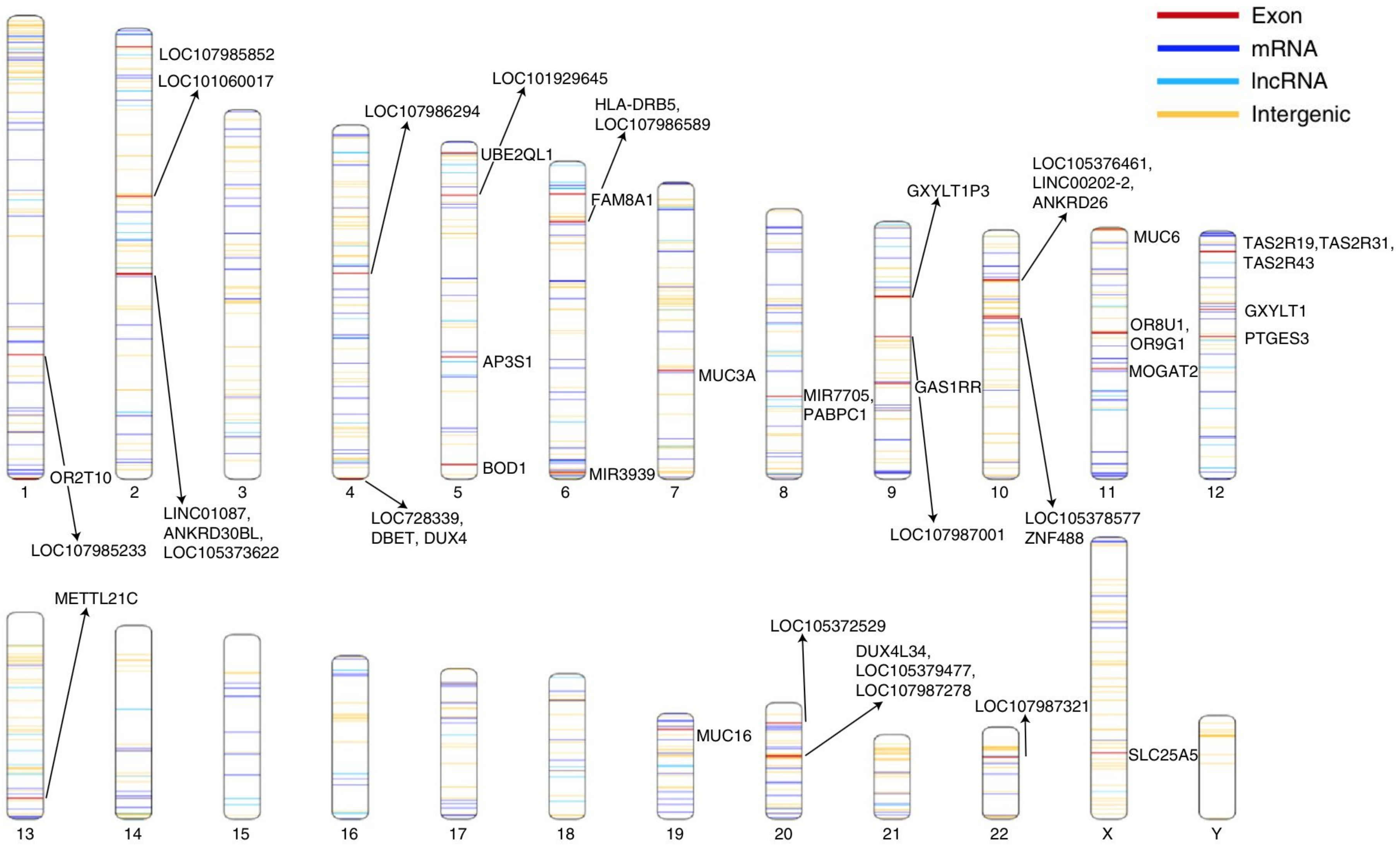


Longest contig: 152,806 bp, present in 11 individuals (1.2%)

Sherman *et al* (2019).

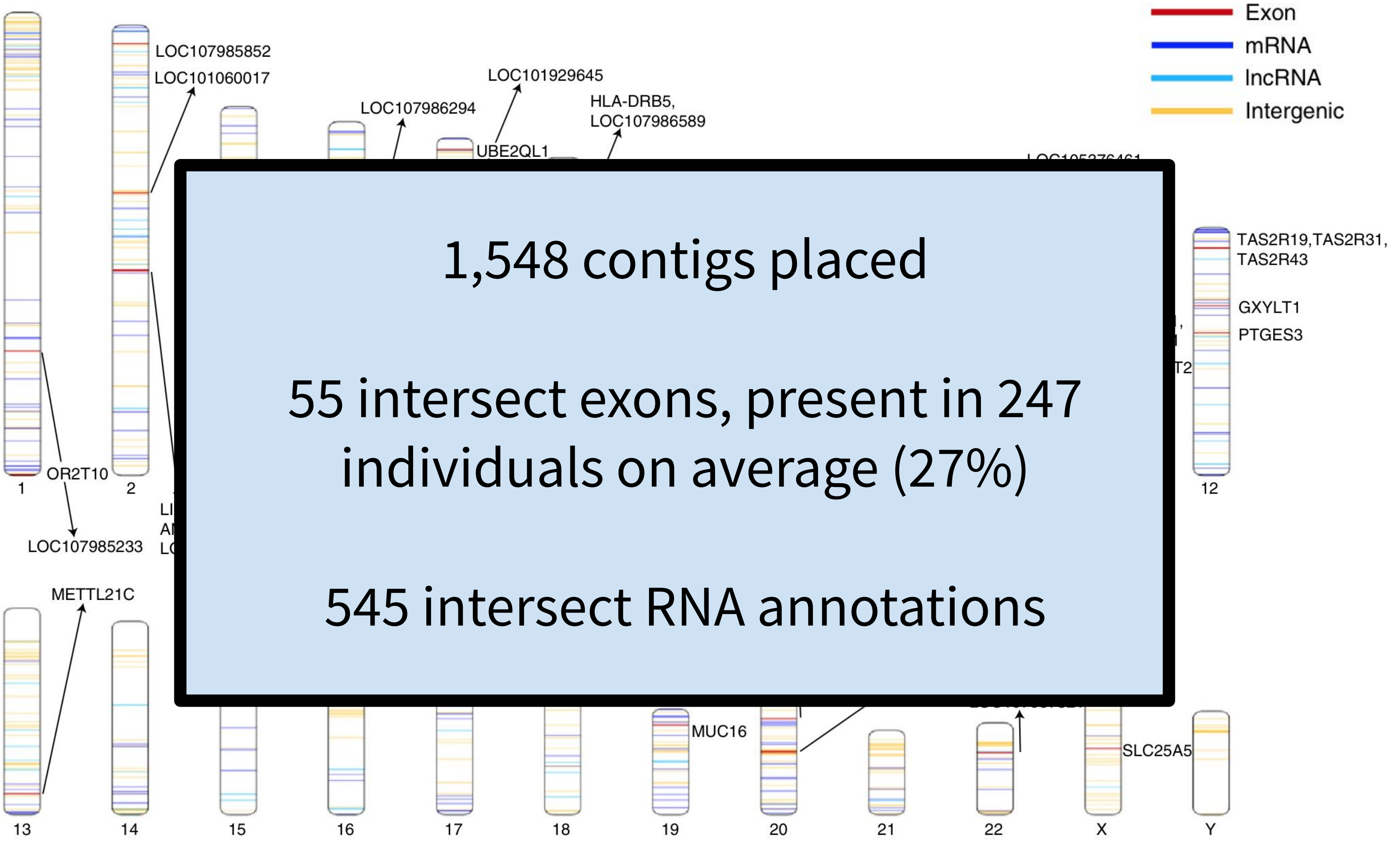
Slide from Rachel Sherman, RECOMB 2019

Pan-genome insertion locations



Sherman *et al* (2019).

Pan-genome insertion locations



1,548 contigs placed

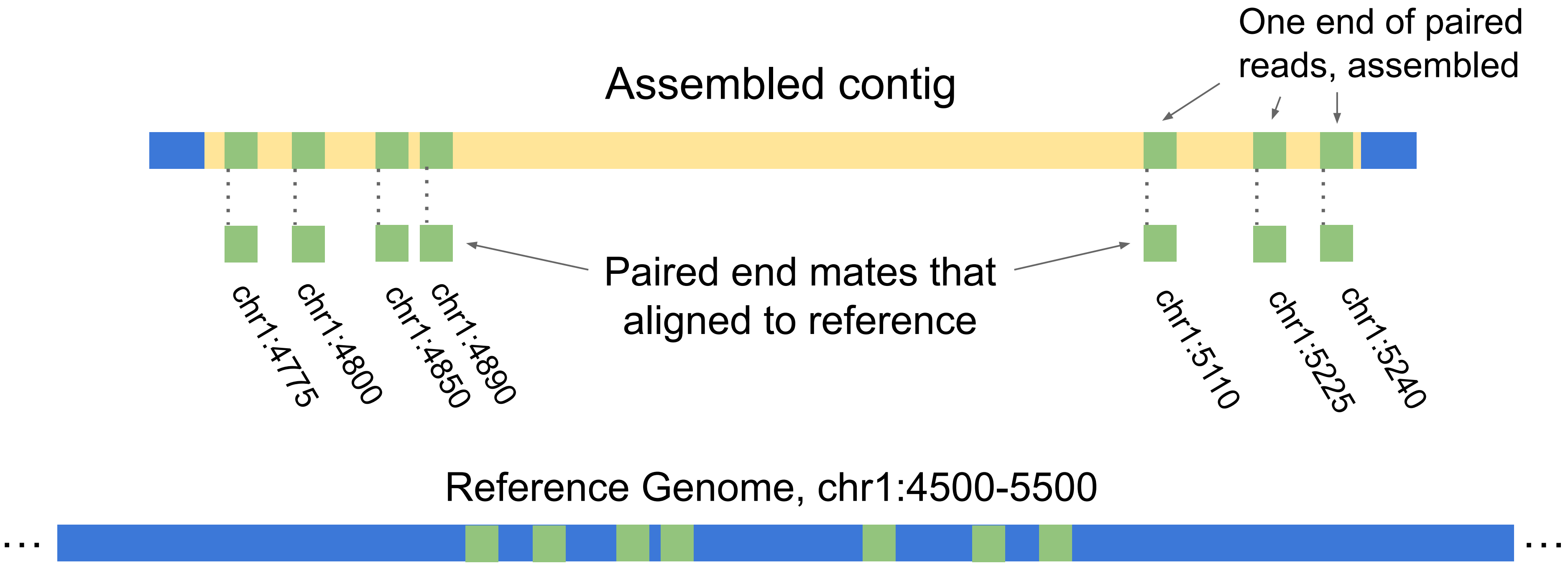
55 intersect exons, present in 247 individuals on average (27%)

545 intersect RNA annotations

Sherman *et al* (2019).

Slide from Rachel Sherman, RECOMB 2019

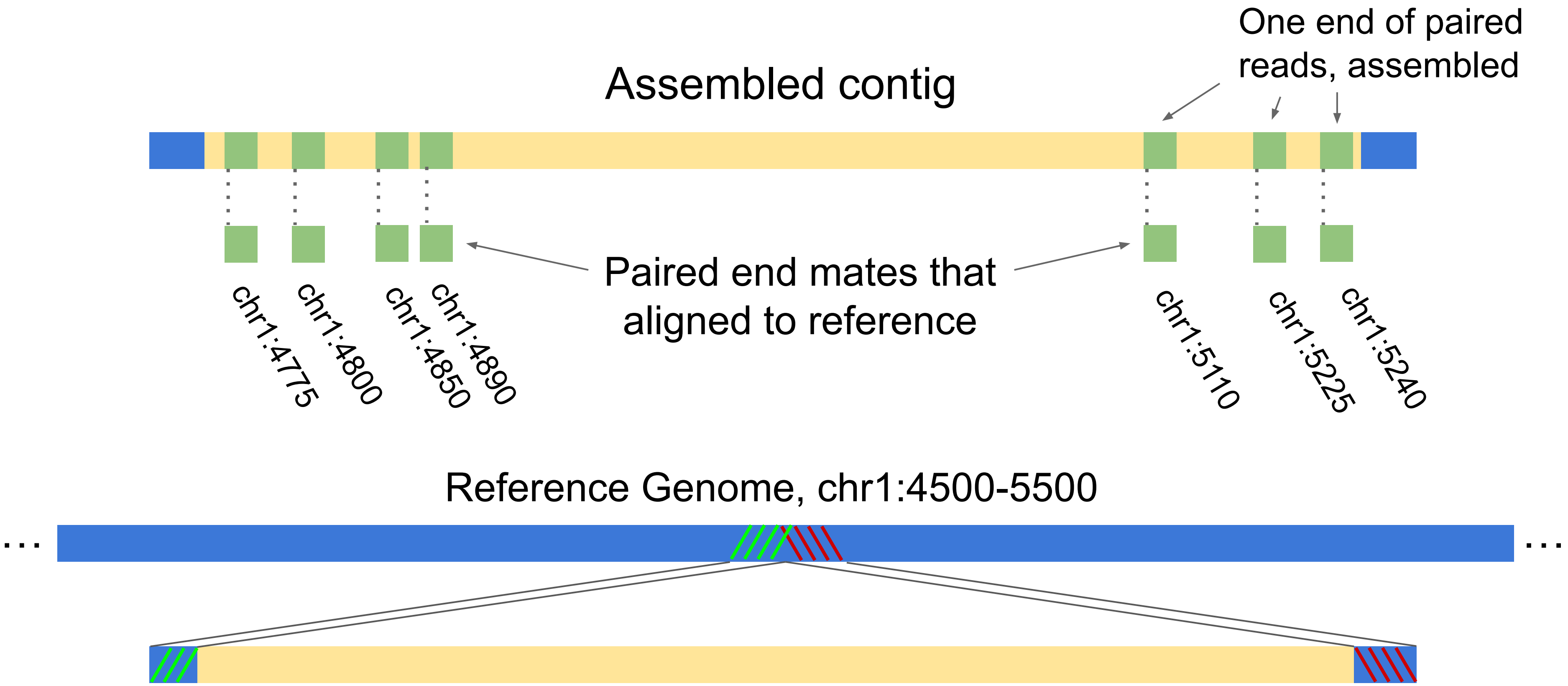
Placing assembled contigs


One end of paired reads, assembled
Assembled contig
Paired end mates that aligned to reference
chr1:4775
chr1:4800
chr1:4850
chr1:4890
chr1:5110
chr1:5225
chr1:5240
Reference Genome, chr1:4500-5500
...
...

Use mate pair information to localize contigs to a GRCh38 location

Slide from Rachel Sherman, RECOMB 2019

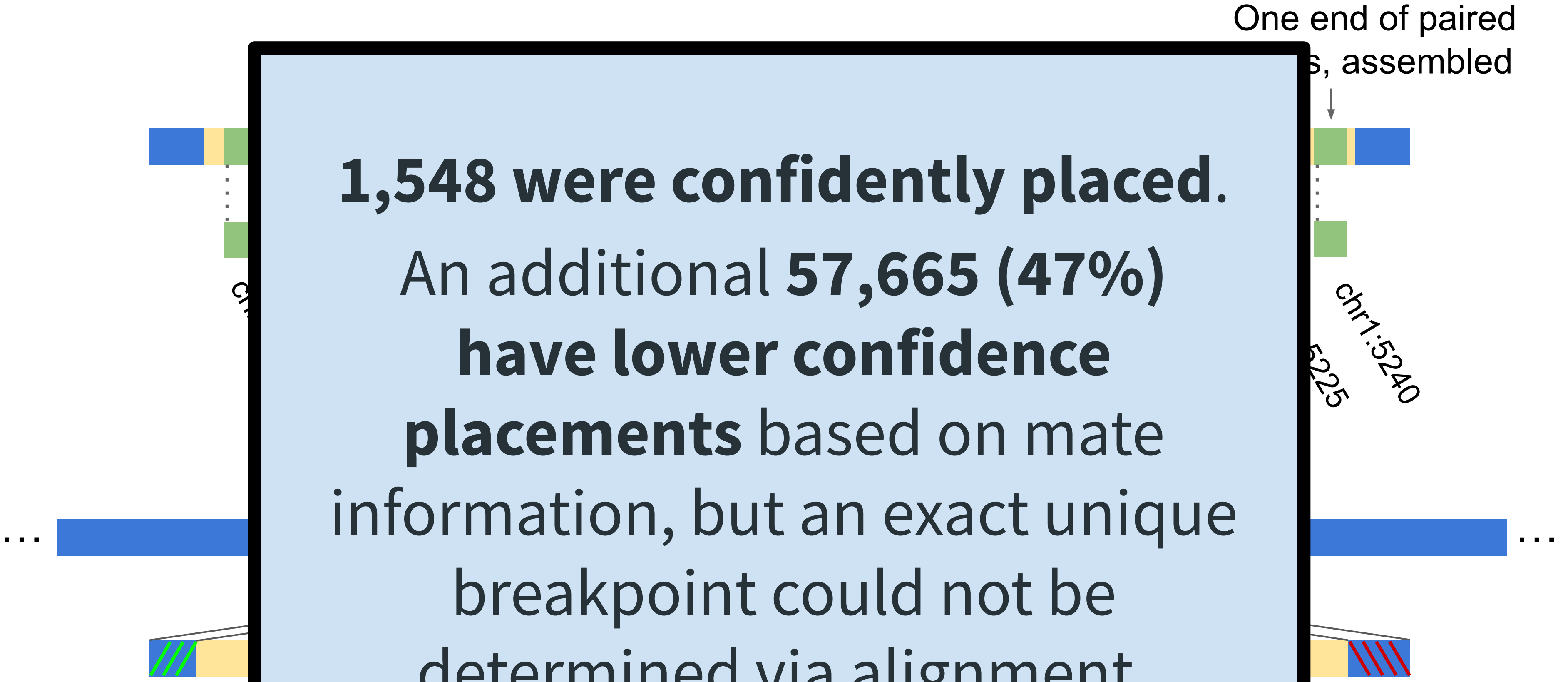
Placing assembled contigs


One end of paired reads, assembled
Assembled contig
Paired end mates that aligned to reference
chr1:4775
chr1:4800
chr1:4850
chr1:4890
chr1:5110
chr1:5225
chr1:5240
Reference Genome, chr1:4500-5500
...
...

Align contig to mate-indicated region and look for unique alignments of ends in correct orientation.

Slide from Rachel Sherman, RECOMB 2019

Placing assembled contigs



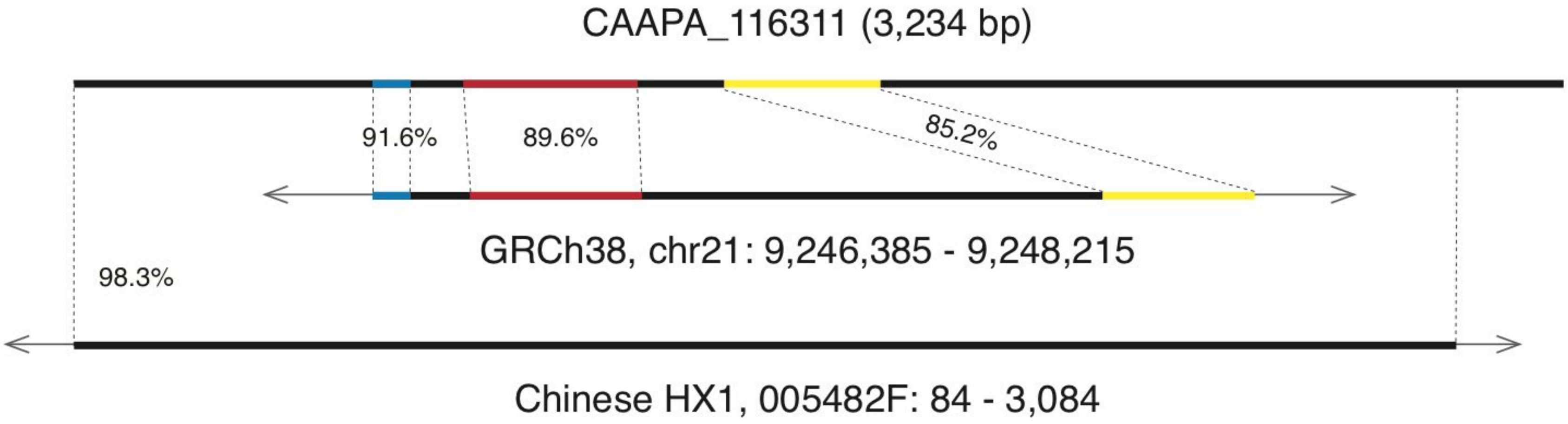
1,548 were confidently placed.

An additional **57,665 (47%) have lower confidence placements** based on mate information, but an exact unique breakpoint could not be determined via alignment.

Slide from Rachel Sherman, RECOMB 2019

Pan-genome contigs in Asian assemblies

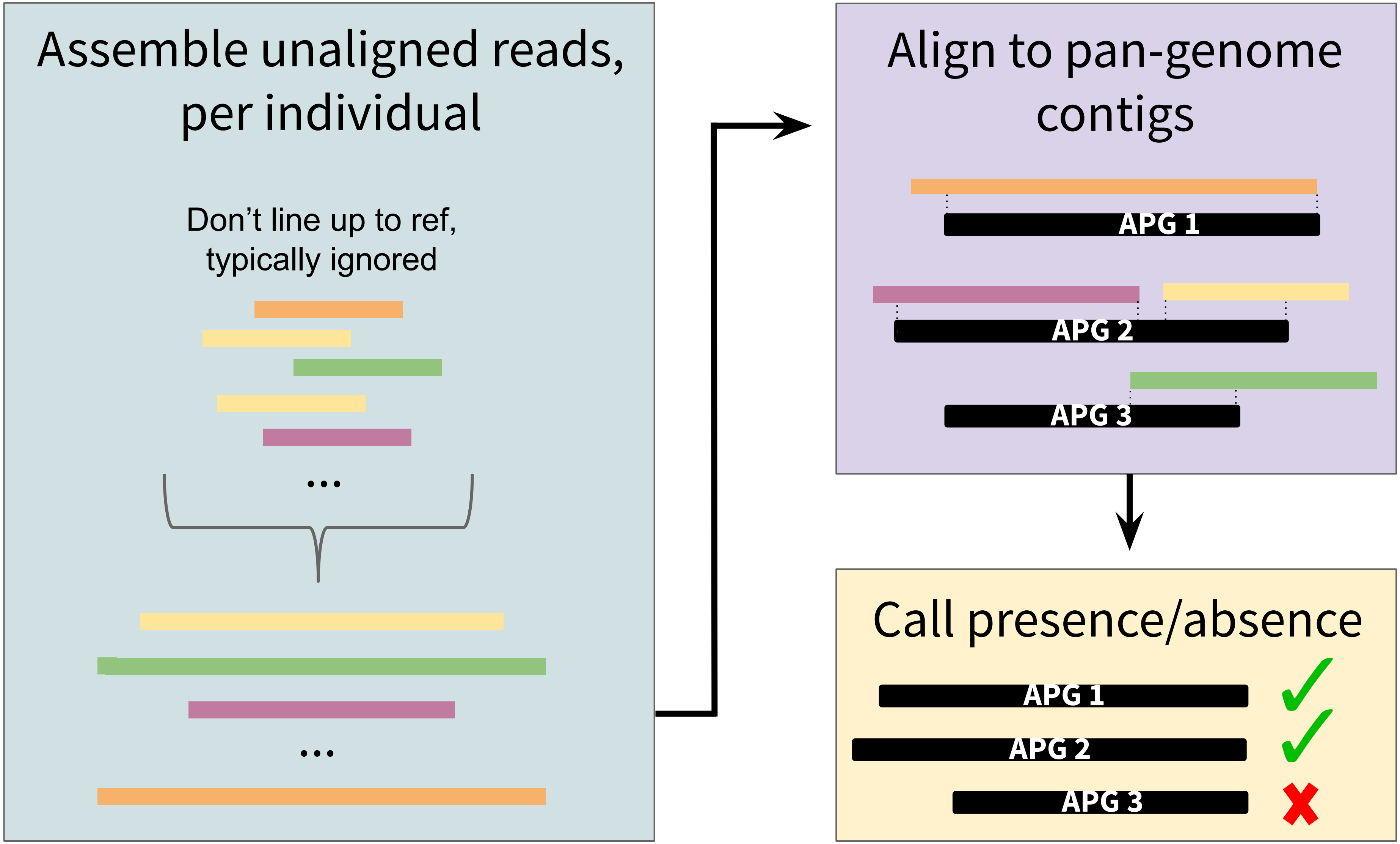
42,207 contigs totaling 120.7 Mb align to Chinese or Korean long read assemblies (HX1 or KOREF)


CAAPA_116311 (3,234 bp)
91.6%
89.6%
85.2%
GRCh38, chr21: 9,246,385 - 9,248,215
98.3%
Chinese HX1, 005482F: 84 - 3,084

Sherman *et al* (2019).

Slide from Rachel Sherman, RECOMB 2019

Pan-genome contigs in other WGS cohorts



Assemble unaligned reads, per individual

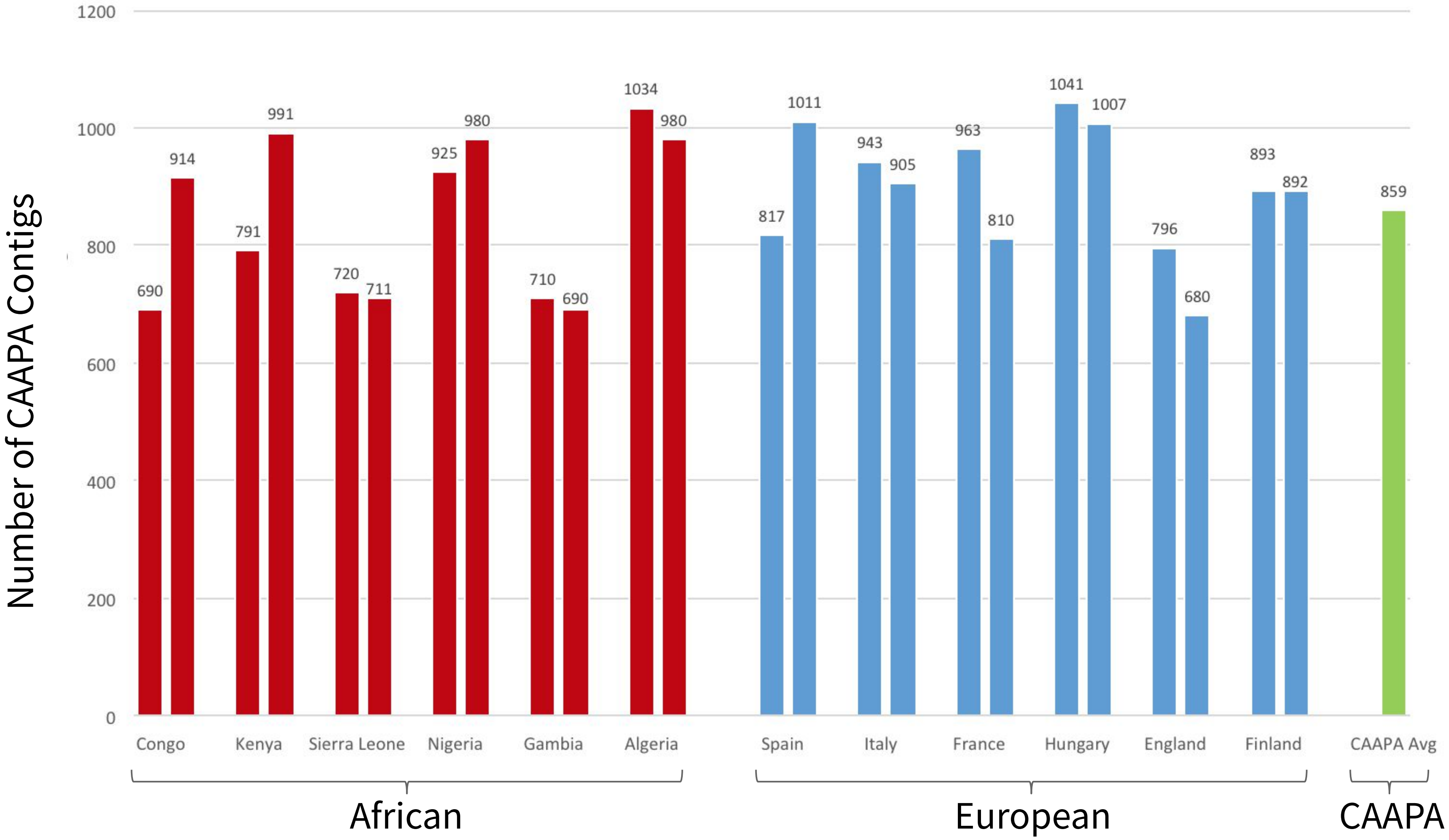
Don’t line up to ref, typically ignored

Align to pan-genome contigs

Call presence/absence

Slide from Rachel Sherman, RECOMB 2019

Pan-genome contigs in SGDP populations

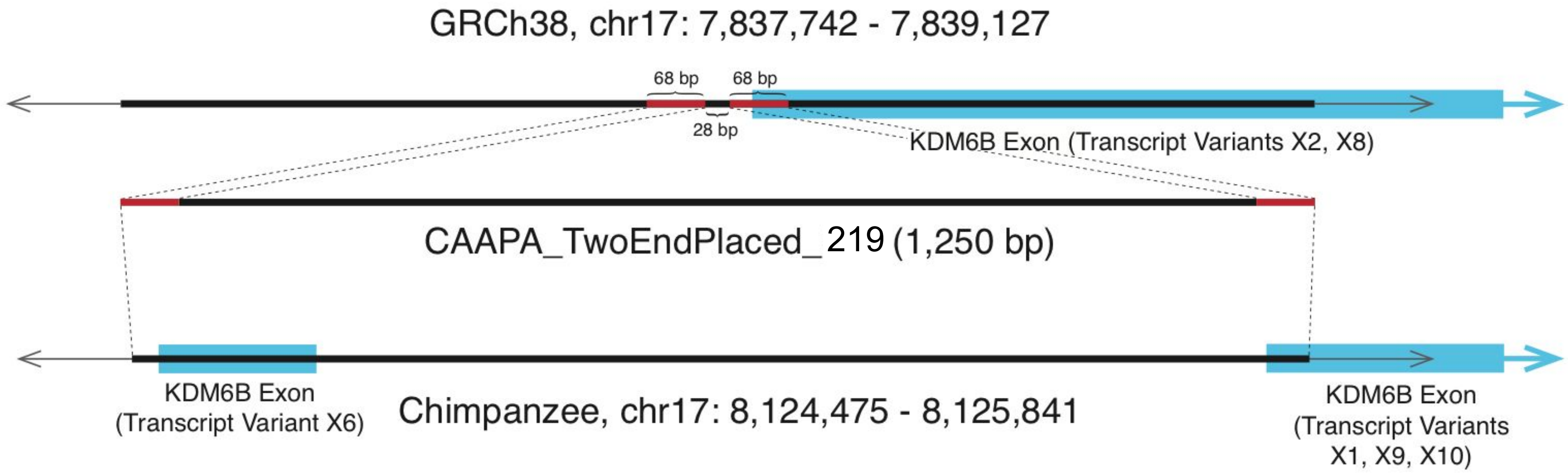


data from Sherman *et al* (2019), *Simons Genome Diversity Project samples from* Mallick *et al* (2016).

Slide from Rachel Sherman, RECOMB 2019

Are any of these sequences transcribed?

Insertion in at least 769 individuals (85%), intersects a known primate exon in KDM6B that isn't annotated in GRCh38:

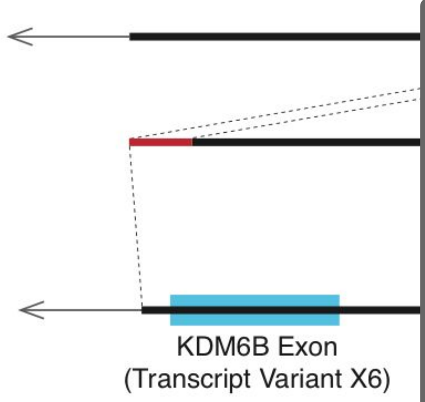

GRCh38, chr17: 7,837,742 - 7,839,127
68 bp
68 bp
28 bp
KDM6B Exon (Transcript Variants X2, X8)
CAAPA_TwoEndPlaced_219 (1,250 bp)
KDM6B Exon
(Transcript Variant X6)
Chimpanzee, chr17: 8,124,475 - 8,125,841
KDM6B Exon
(Transcript Variants
X1, X9, X10)

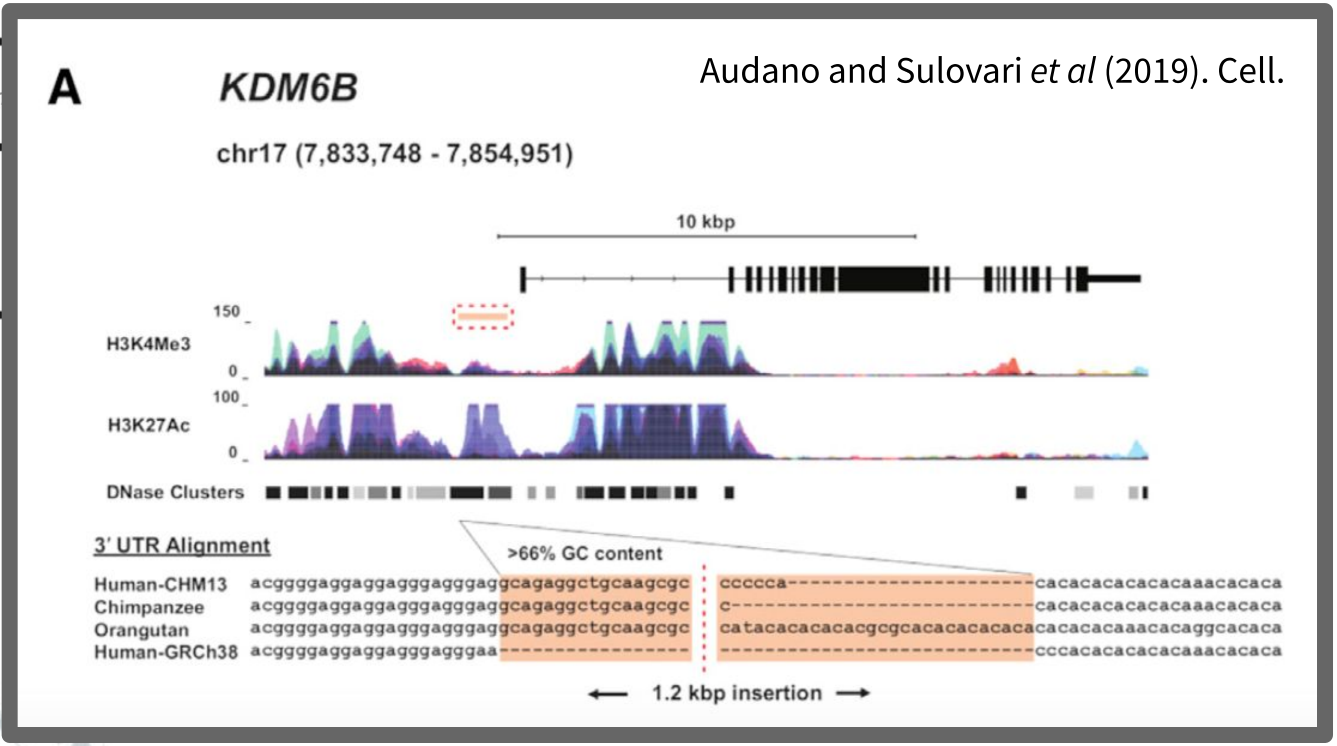
Slide from Rachel Sherman, RECOMB 2019

Are any of these sequences transcribed?

Insertion in at least 769 individuals (85%), intersects a known primate exon in KDM6B that isn't annotated in GRCh38:

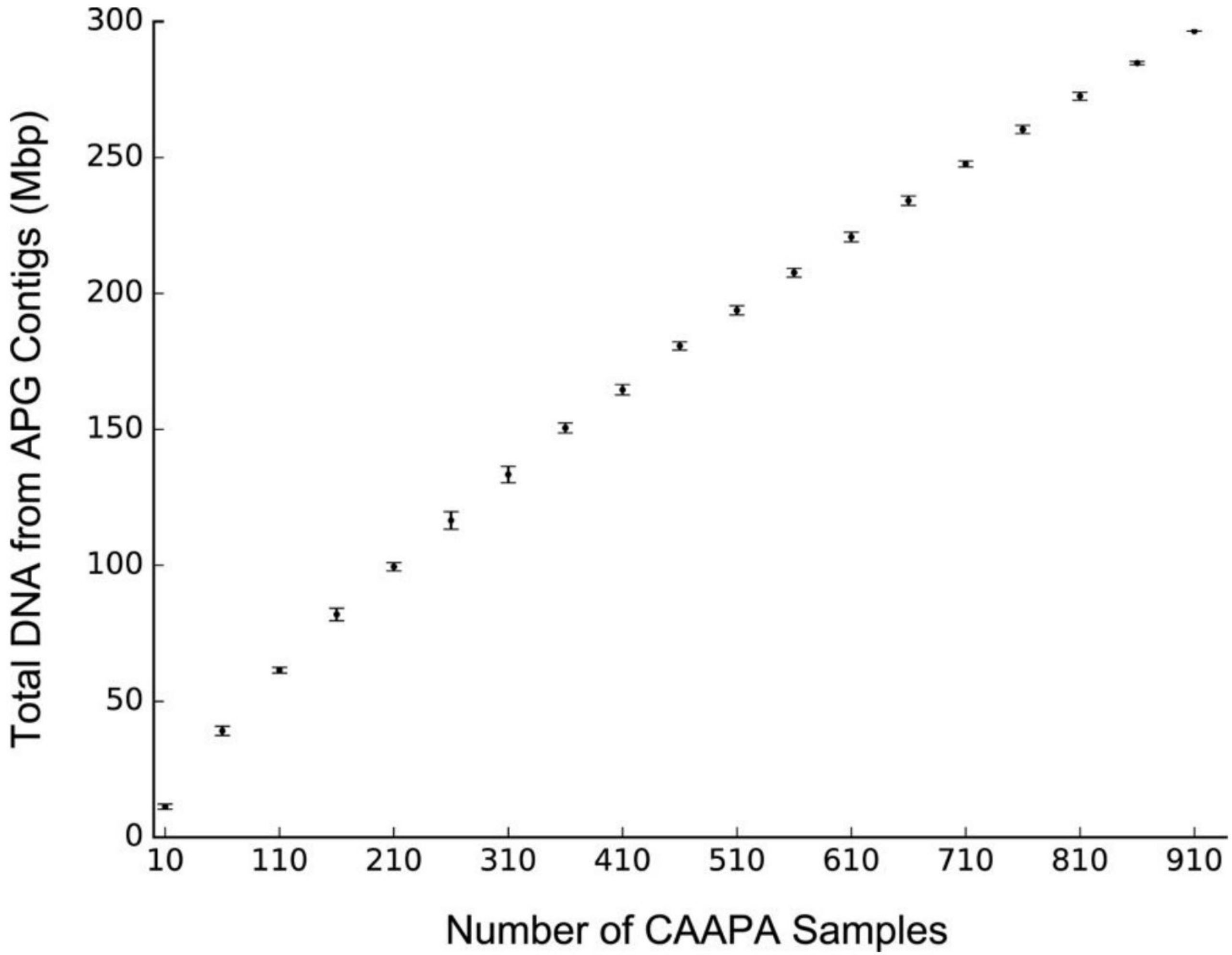
GRCh38, chr17: 7,837,742 - 7,839,127




Audano and Sulovari *et al* (2019). Cell.

Slide from Rachel Sherman, RECOMB 2019

The pan-genome is still open



Sherman *et al* (2019).

Slide from Rachel Sherman, RECOMB 2019

Acknowledgments

Steven Salzberg

Daniela Puiu

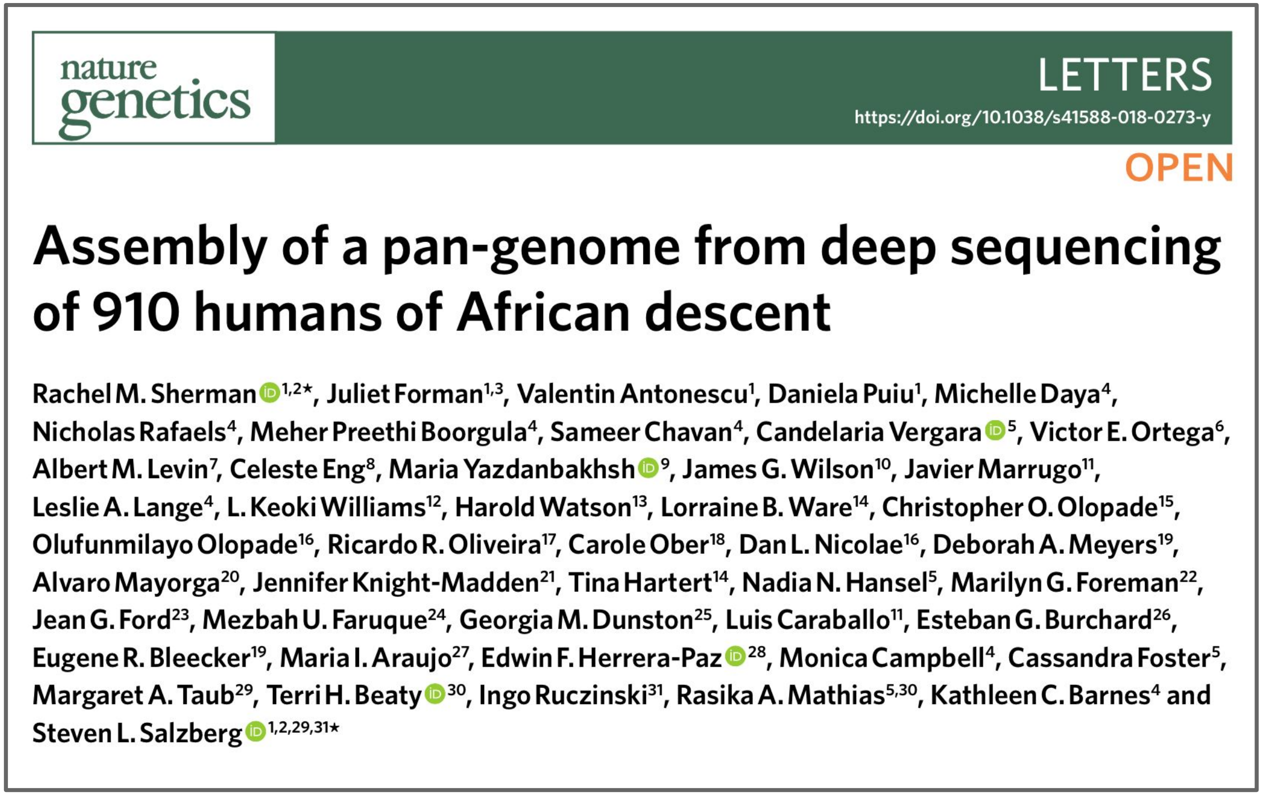
Valentin Antonescu

Juliet Forman

JOHNS HOPKINS
UNIVERSITY

Consortium on Asthma among African-ancestry Populations in the Americas
CAAPA

NIH National Heart, Lung, and Blood Institute



nature genetics

LETTERS

https://doi.org/10.1038/s41588-018-0273-y

OPEN

Assembly of a pan-genome from deep sequencing of 910 humans of African descent

Rachel M. Sherman[1,2*], Juliet Forman[1,3], Valentin Antonescu[1], Daniela Puiu[1], Michelle Daya[4], Nicholas Rafaels[4], Meher Preethi Boorgula[4], Sameer Chavan[4], Candelaria Vergara[5], Victor E. Ortega[6], Albert M. Levin[7], Celeste Eng[8], Maria Yazdanbakhsh[9], James G. Wilson[10], Javier Marrugo[11], Leslie A. Lange[4], L. Keoki Williams[12], Harold Watson[13], Lorraine B. Ware[14], Christopher O. Olopade[15], Olufunmilayo Olopade[16], Ricardo R. Oliveira[17], Carole Ober[18], Dan L. Nicolae[16], Deborah A. Meyers[19], Alvaro Mayorga[20], Jennifer Knight-Madden[21], Tina Hartert[14], Nadia N. Hansel[5], Marilyn G. Foreman[22], Jean G. Ford[23], Mezbah U. Faruque[24], Georgia M. Dunston[25], Luis Caraballo[11], Esteban G. Burchard[26], Eugene R. Bleecker[19], Maria I. Araujo[27], Edwin F. Herrera-Paz[28], Monica Campbell[4], Cassandra Foster[5], Margaret A. Taub[29], Terri H. Beaty[30], Ingo Ruczinski[31], Rasika A. Mathias[5,30], Kathleen C. Barnes[4] and Steven L. Salzberg[1,2,29,31*]

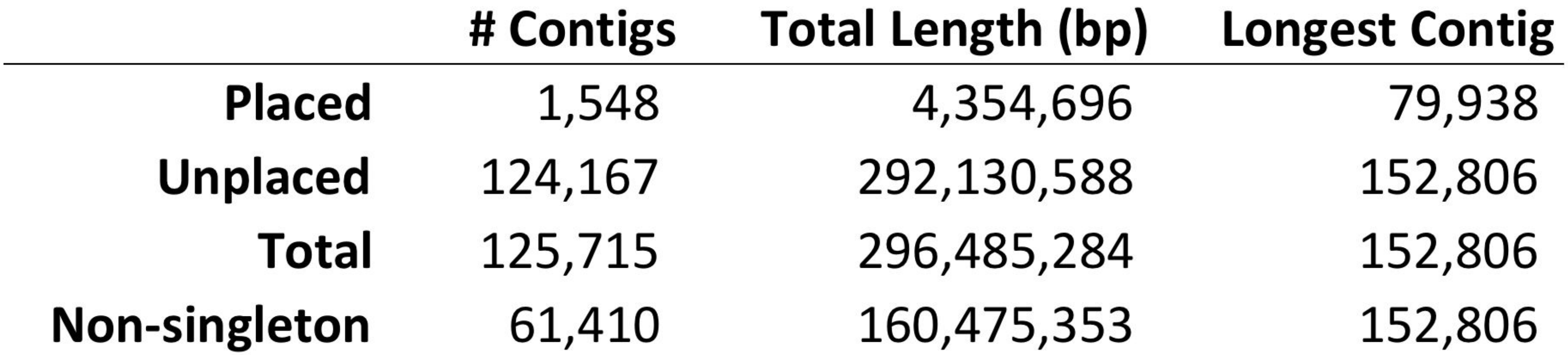
@rshermanjhu

Questions?

Additional Slides

Slide from Rachel Sherman, RECOMB 2019

Pan-genome stats



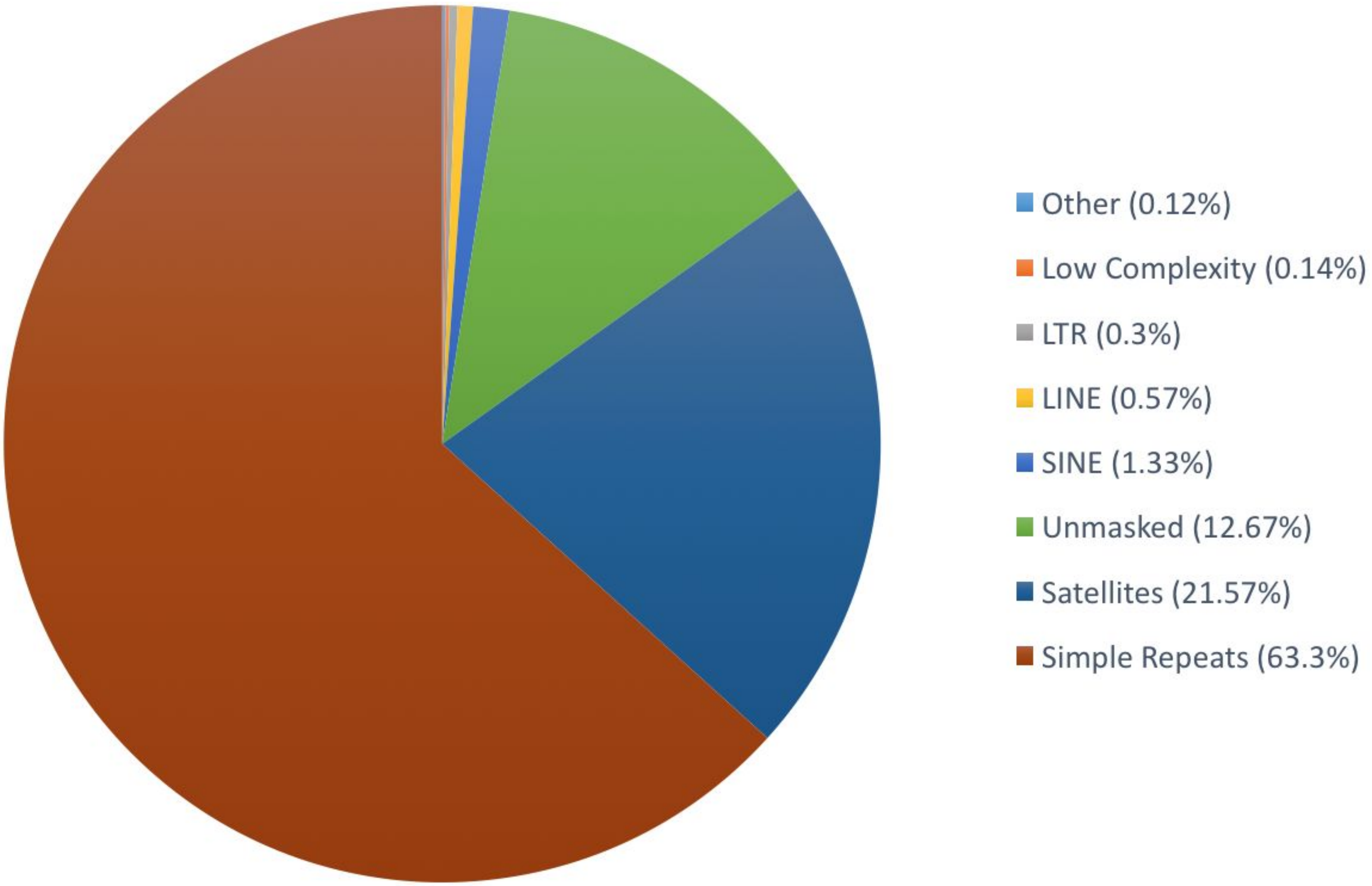
	# Contigs	Total Length (bp)	Longest Contig
Placed	1,548	4,354,696	79,938
Unplaced	124,167	292,130,588	152,806
Total	125,715	296,485,284	152,806
Non-singleton	61,410	160,475,353	152,806

- 51% of contigs are singletons
- 34% of contigs align to HX1 or KOREF
- 98% of contigs have some alignment to Chimpanzee or Rhesus Macaque, demonstrating these are not contaminants

Sherman *et al* (2019).

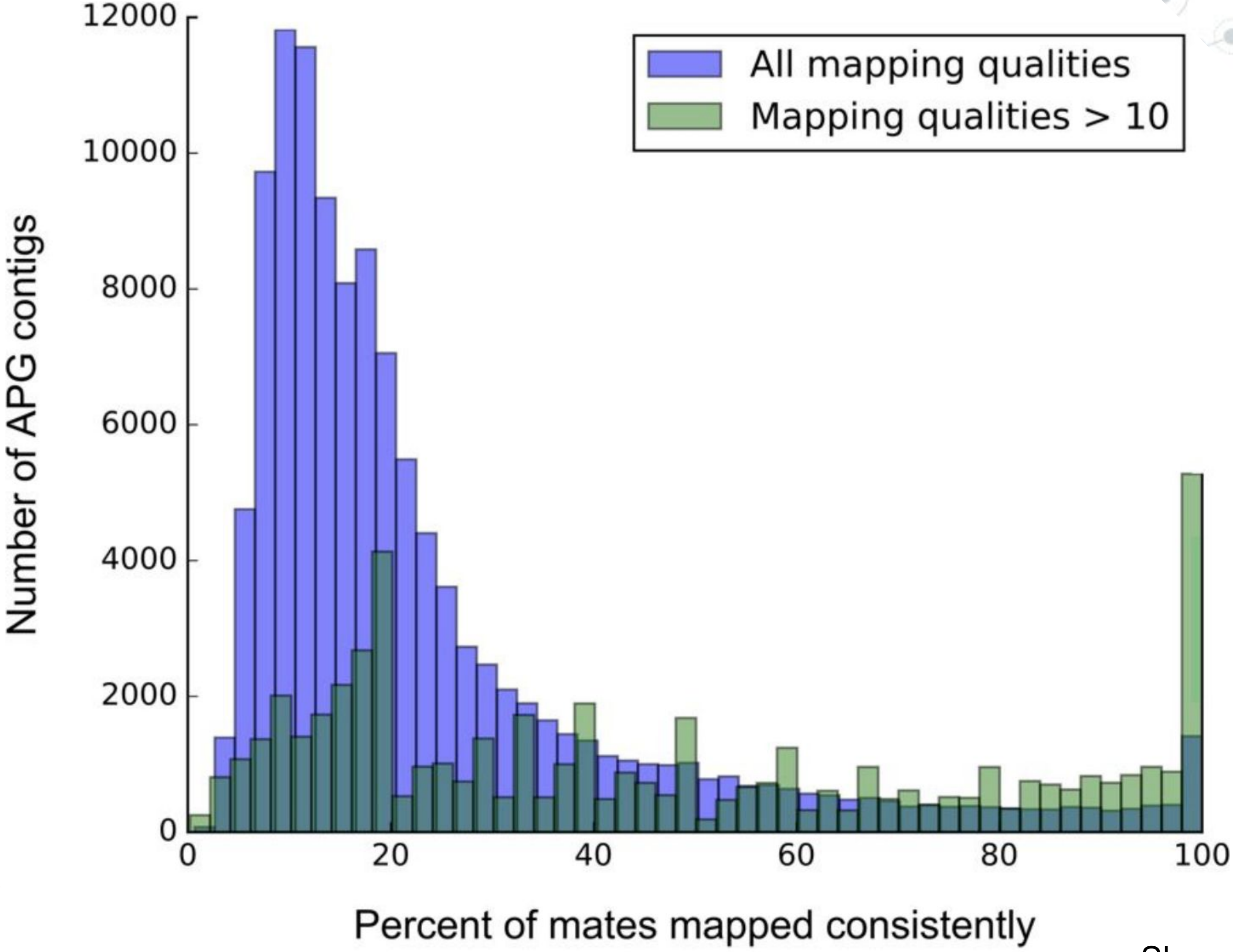
Slide from Rachel Sherman, RECOMB 2019

Repeat content in pan-genome contigs



Slide from Rachel Sherman, RECOMB 2019

Placement of contigs based on mapped mates



Sherman *et al* (2019).

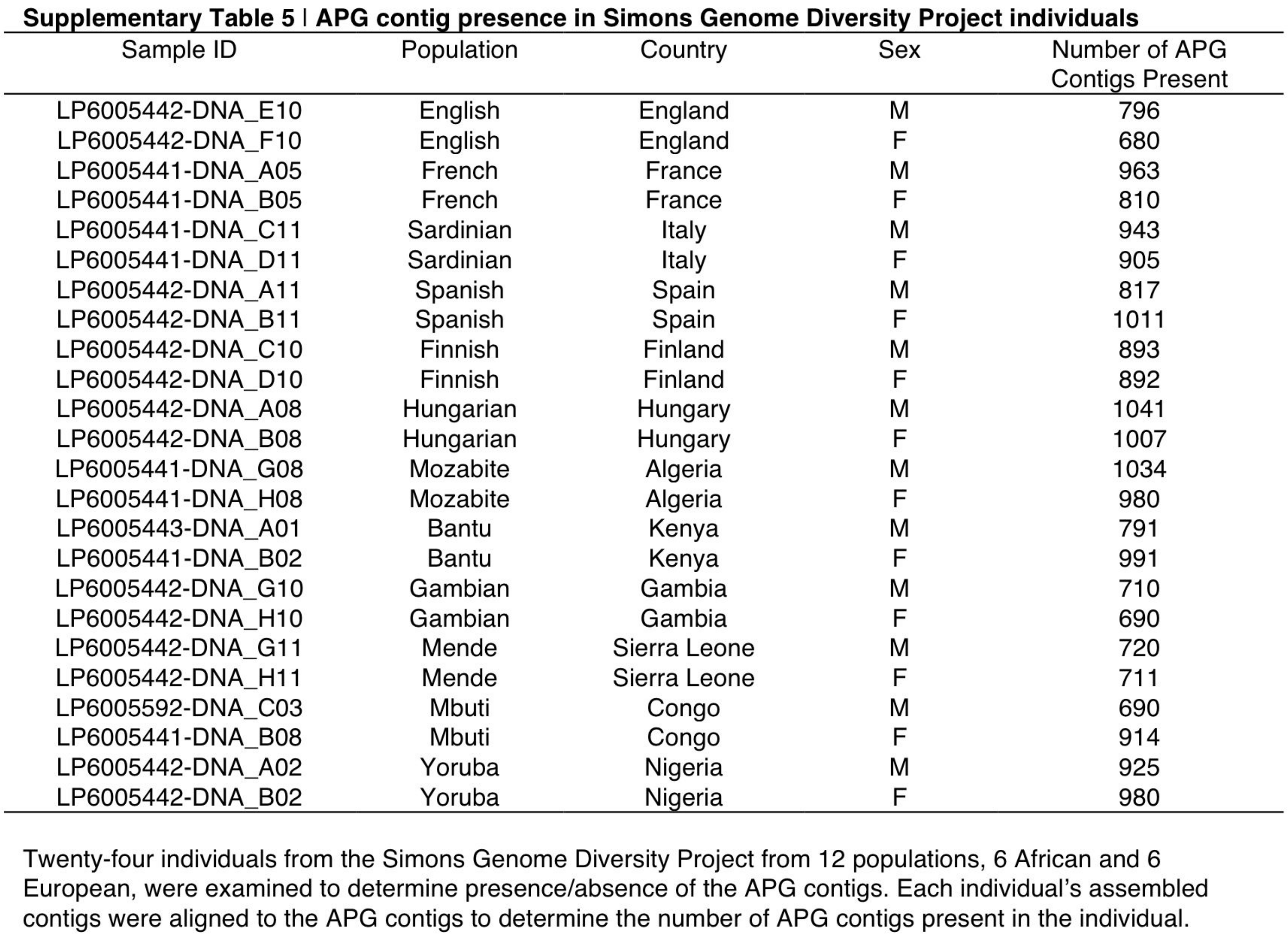
Slide from Rachel Sherman, RECOMB 2019

Pan-genome contig presence/absence

	Number of Contigs	Mean # Insertions Per Individual	Mean # Individuals per Insertion
Two Ends Placed	302	120 (39.7%)	363 (of 910)
One End Placed	1,246	212 (17.0%)	155 (of 910)
Unplaced	124,167	527 (0.4%)	4 (of 910)
Total	125,715	859 (0.7%)	6 (of 910)
Shared within CAAPA	33,599	758 (2.2%)	21 (of 910)

Sherman *et al* (2019).

Pan-genome contigs in SGDP individuals



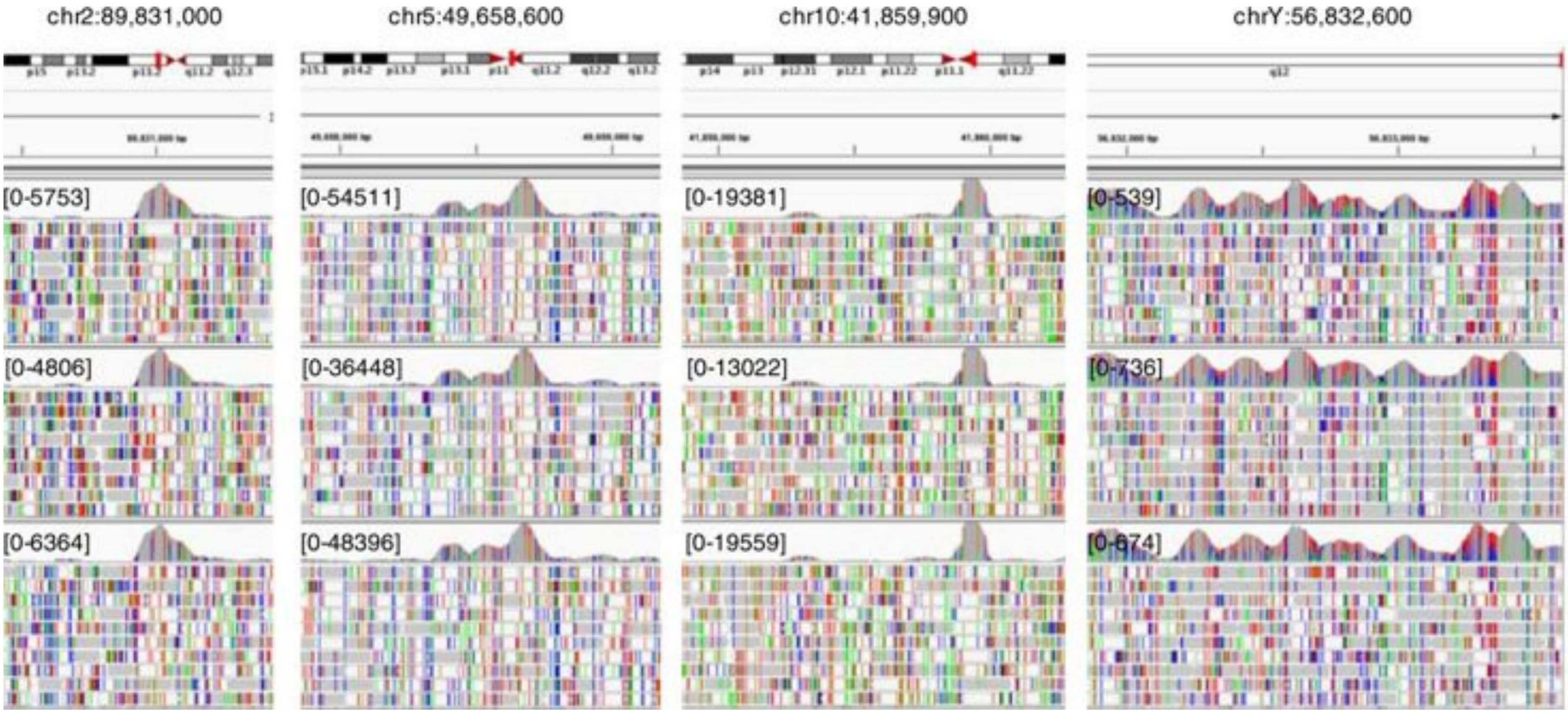
Supplementary Table 5 | APG contig presence in Simons Genome Diversity Project individuals

Sample ID	Population	Country	Sex	Number of APG Contigs Present
LP6005442-DNA_E10	English	England	M	796
LP6005442-DNA_F10	English	England	F	680
LP6005441-DNA_A05	French	France	M	963
LP6005441-DNA_B05	French	France	F	810
LP6005441-DNA_C11	Sardinian	Italy	M	943
LP6005441-DNA_D11	Sardinian	Italy	F	905
LP6005442-DNA_A11	Spanish	Spain	M	817
LP6005442-DNA_B11	Spanish	Spain	F	1011
LP6005442-DNA_C10	Finnish	Finland	M	893
LP6005442-DNA_D10	Finnish	Finland	F	892
LP6005442-DNA_A08	Hungarian	Hungary	M	1041
LP6005442-DNA_B08	Hungarian	Hungary	F	1007
LP6005441-DNA_G08	Mozabite	Algeria	M	1034
LP6005441-DNA_H08	Mozabite	Algeria	F	980
LP6005443-DNA_A01	Bantu	Kenya	M	791
LP6005441-DNA_B02	Bantu	Kenya	F	991
LP6005442-DNA_G10	Gambian	Gambia	M	710
LP6005442-DNA_H10	Gambian	Gambia	F	690
LP6005442-DNA_G11	Mende	Sierra Leone	M	720
LP6005442-DNA_H11	Mende	Sierra Leone	F	711
LP6005592-DNA_C03	Mbuti	Congo	M	690
LP6005441-DNA_B08	Mbuti	Congo	F	914
LP6005442-DNA_A02	Yoruba	Nigeria	M	925
LP6005442-DNA_B02	Yoruba	Nigeria	F	980

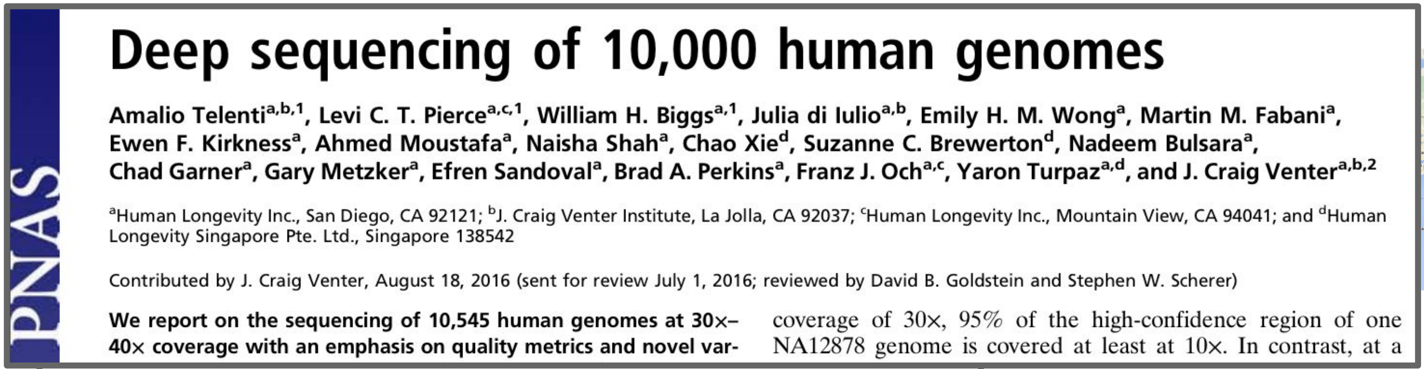
Twenty-four individuals from the Simons Genome Diversity Project from 12 populations, 6 African and 6 European, were examined to determine presence/absence of the APG contigs. Each individual's assembled contigs were aligned to the APG contigs to determine the number of APG contigs present in the individual.

Sherman *et al* (2019).

Underrepresented reference elements?



Capturing human genetic diversity



PNAS

Deep sequencing of 10,000 human genomes

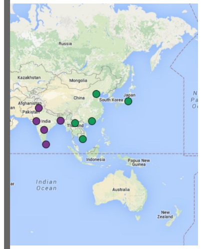
Amalio Telenti[a,b,1], Levi C. T. Pierce[a,c,1], William H. Biggs[a,1], Julia di Iulio[a,b], Emily H. M. Wong[a], Martin M. Fabani[a], Ewen F. Kirkness[a], Ahmed Moustafa[a], Naisha Shah[a], Chao Xie[d], Suzanne C. Brewerton[d], Nadeem Bulsara[a], Chad Garner[a], Gary Metzker[a], Efren Sandoval[a], Brad A. Perkins[a], Franz J. Och[a,c], Yaron Turpaz[a,d], and J. Craig Venter[a,b,2]

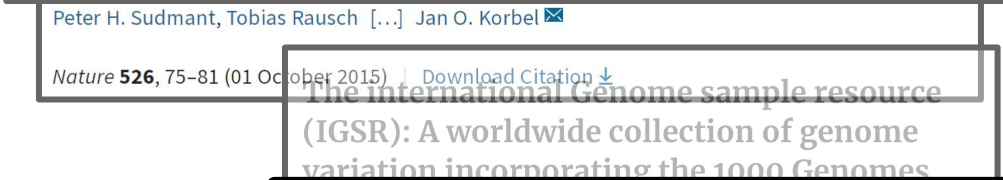
[a]Human Longevity Inc., San Diego, CA 92121; [b]J. Craig Venter Institute, La Jolla, CA 92037; [c]Human Longevity Inc., Mountain View, CA 94041; and [d]Human Longevity Singapore Pte. Ltd., Singapore 138542

Contributed by J. Craig Venter, August 18, 2016 (sent for review July 1, 2016; reviewed by David B. Goldstein and Stephen W. Scherer)

We report on the sequencing of 10,545 human genomes at 30×–40× coverage with an emphasis on quality metrics and novel var-

coverage of 30×, 95% of the high-confidence region of one NA12878 genome is covered at least at 10×. In contrast, at a



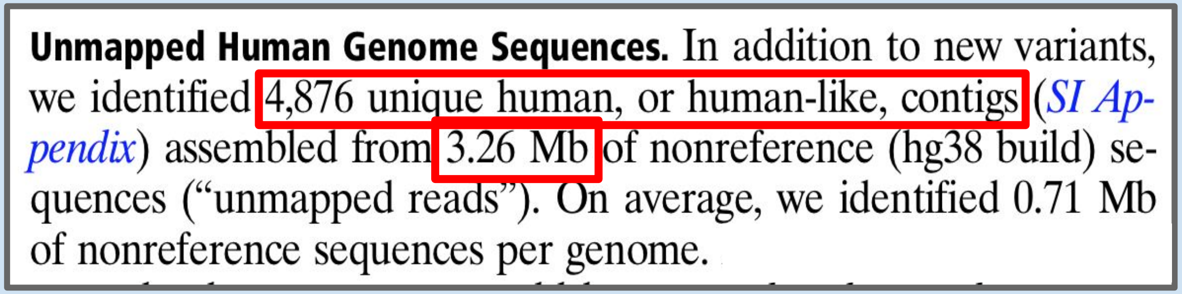


Peter H. Sudmant, Tobias Rausch […] Jan O. Korbel

Nature **526**, 75–81 (01 October 2015) | Download Citation

The International Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes

10k genomes, 3.26 Mb novel sequence



Unmapped Human Genome Sequences. In addition to new variants, we identified 4,876 unique human, or human-like, contigs (*SI Appendix*) assembled from 3.26 Mb of nonreference (hg38 build) sequences ("unmapped reads"). On average, we identified 0.71 Mb of nonreference sequences per genome.

Project: … populations

Nature 538, 201–206 (13 October 2016) | Download Citation