

Discovering non-reference sequences to assemble a pan-genome from 910 African-ancestry individuals

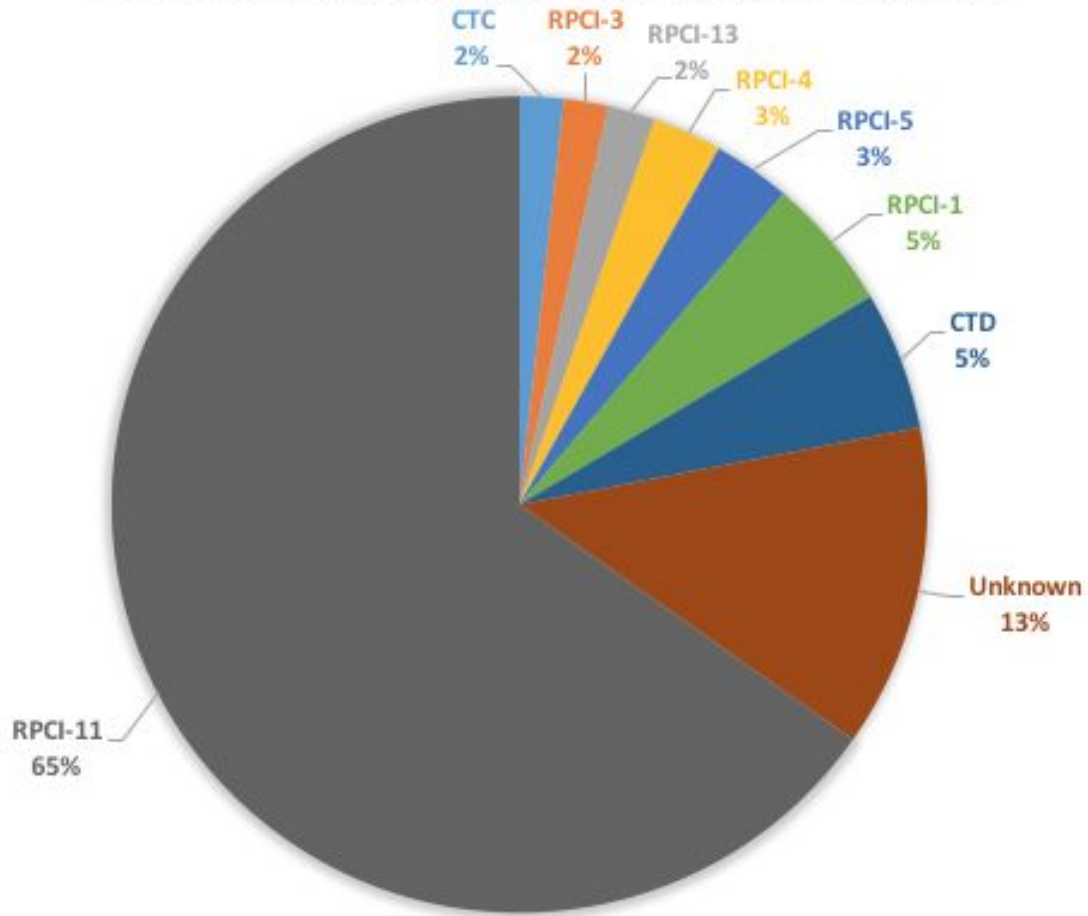
Rachel M Sherman

Johns Hopkins University, Salzberg Lab

Personal Genomics and International Cooperation

October 24, 2019

Human reference genome makeup



Source of BAC clones comprising the reference genome

The majority of the human reference is from one individual

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

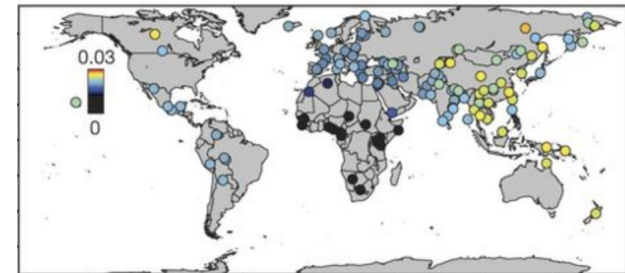
The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data

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

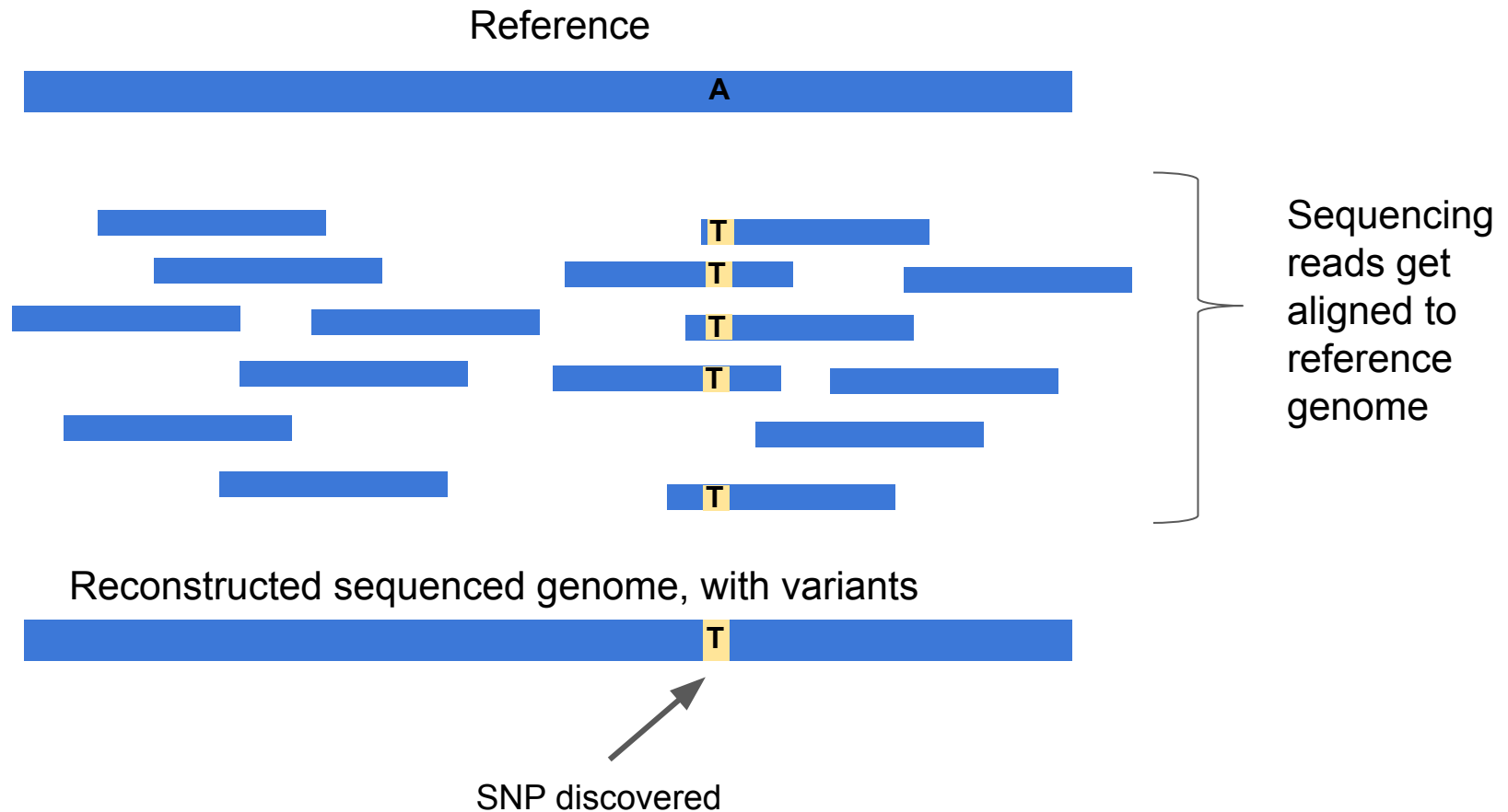
Article | Published: 21 September 2016

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

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

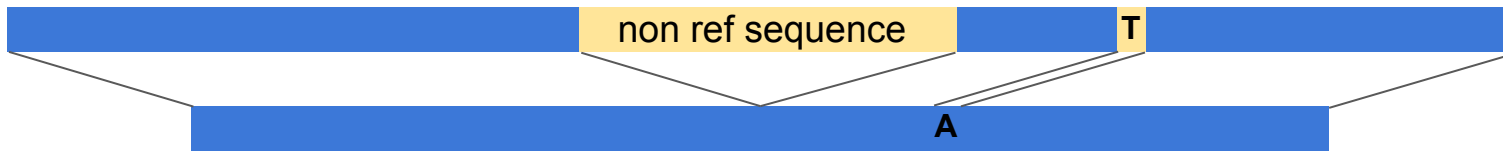
Nature 538, 201–206 (13 October 2016) | Download Citation ↓

Variant discovery via alignment

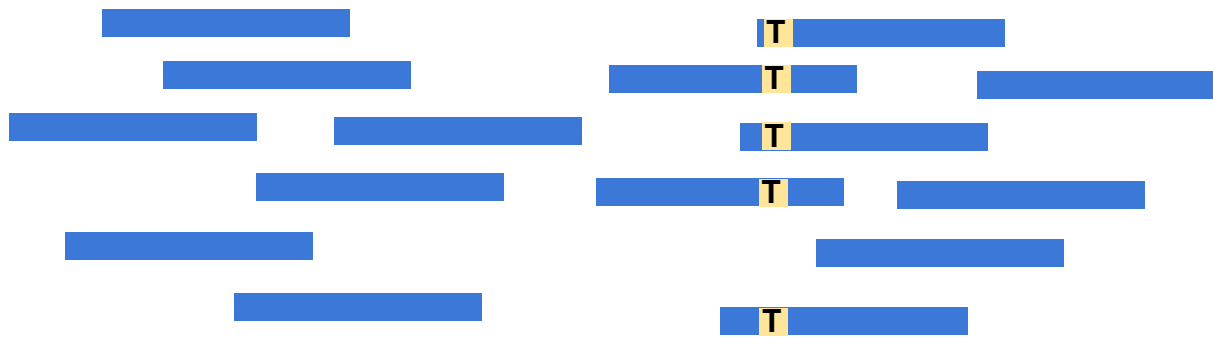


Sequences missed by alignment

True sequenced genome (unknown)

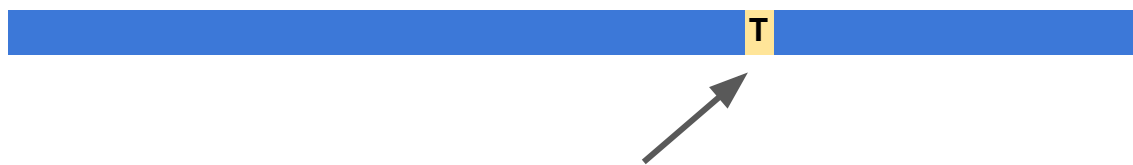


Reference



Sequencing reads get aligned to reference genome

Reconstructed sequenced genome, with variants



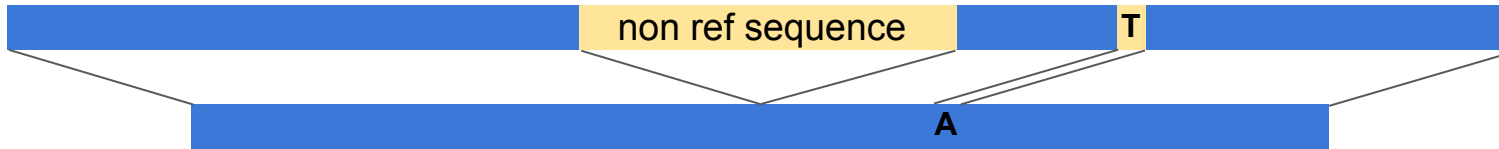
SNP discovered

Don't line up to ref, typically ignored

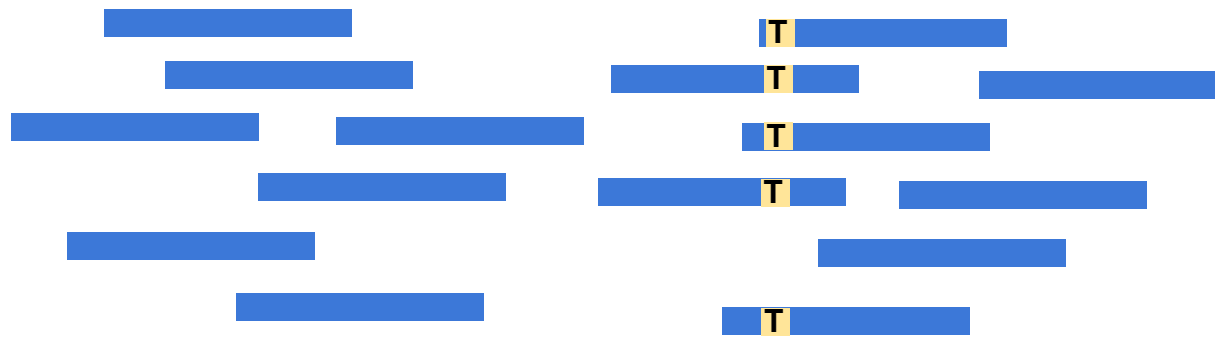


Sequences missed by alignment

True sequenced genome (unknown)



Reference



Sequencing reads get aligned to reference genome

Reconstructed sequenced genome, with variants



SNP discovered

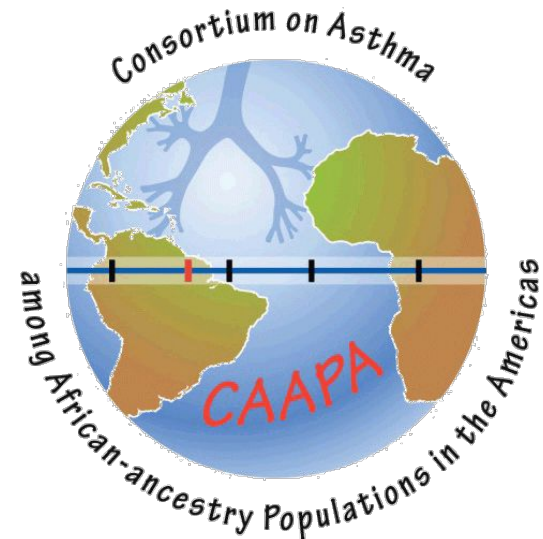


Don't line up to ref, typically ignored



African-ancestry population WGS data

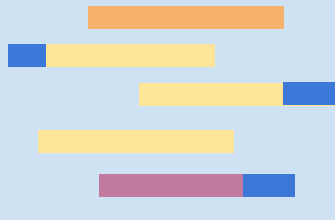
Cohort	Number of Samples
African American (Atlanta)	50
African American (Baltimore-DC)	50
African American (Chicago)	50
African American (Detroit)	50
African American (Jackson, MS)	50
African American (Nashville)	48
African American (NYC)	48
African American (San Francisco)	50
African American (Winston-Salem)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



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

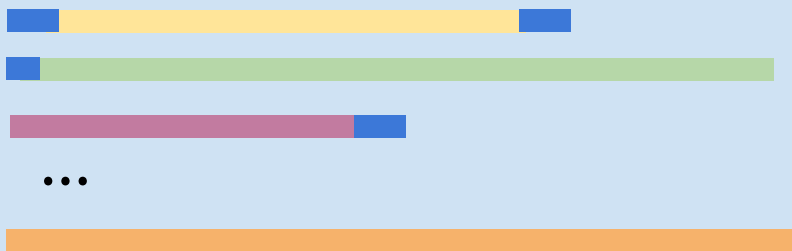
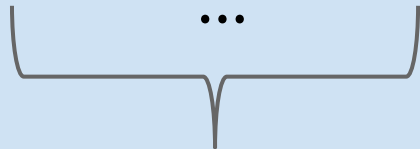
Analyzing the unaligned reads

Don't line up to ref,
typically ignored



✘ 910
people

Assembled with
MaSuRCA; removed
contaminants



≡ > 3.6 Gb sequence



methods



Sequencing
reads get
aligned to
reference
genome

Don't line up to ref,
typically ignored



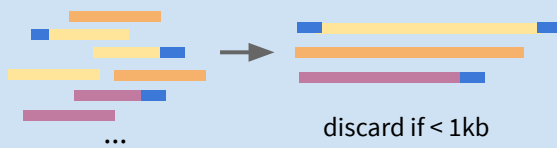
Analyzing the unaligned reads

Align reads to reference



bowtie2

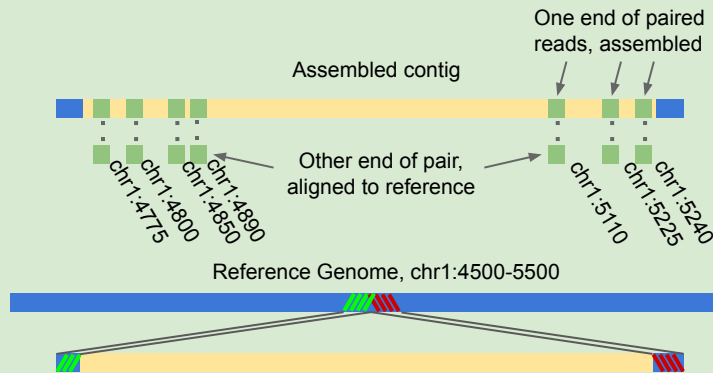
Assemble unaligned reads; filter contaminants



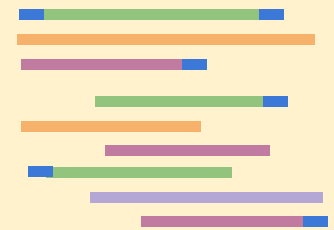
MaSuRCA, Centrifuge+BLAST

Attempt to place in GRCh38

(localize using mate pairs aligned to reference, then local alignment of ends)



Align contigs all-to-all to remove redundancy



nucmer; iteratively

Sequences missed by alignment

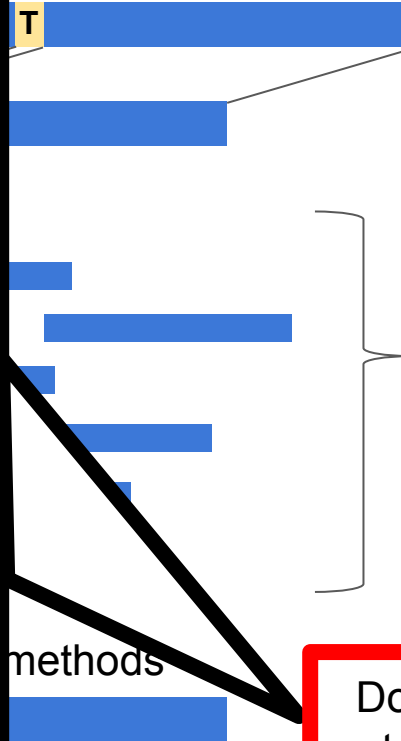
296.5 Mb *non-reference*
insertion sequences

in

125,715 *non-redundant*
contig sequences

from

910 African-ancestry
individuals

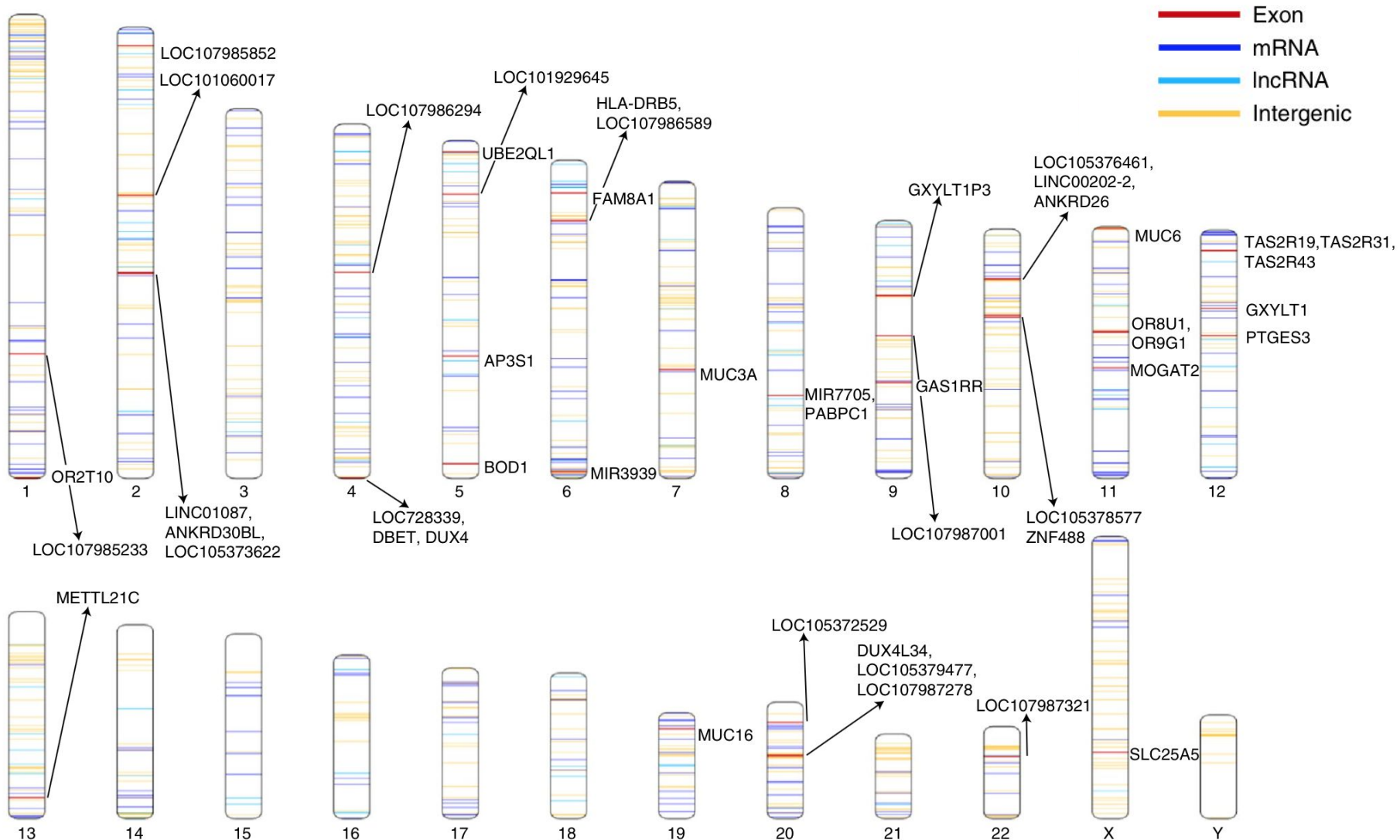


Sequencing
reads get
aligned to
reference
genome

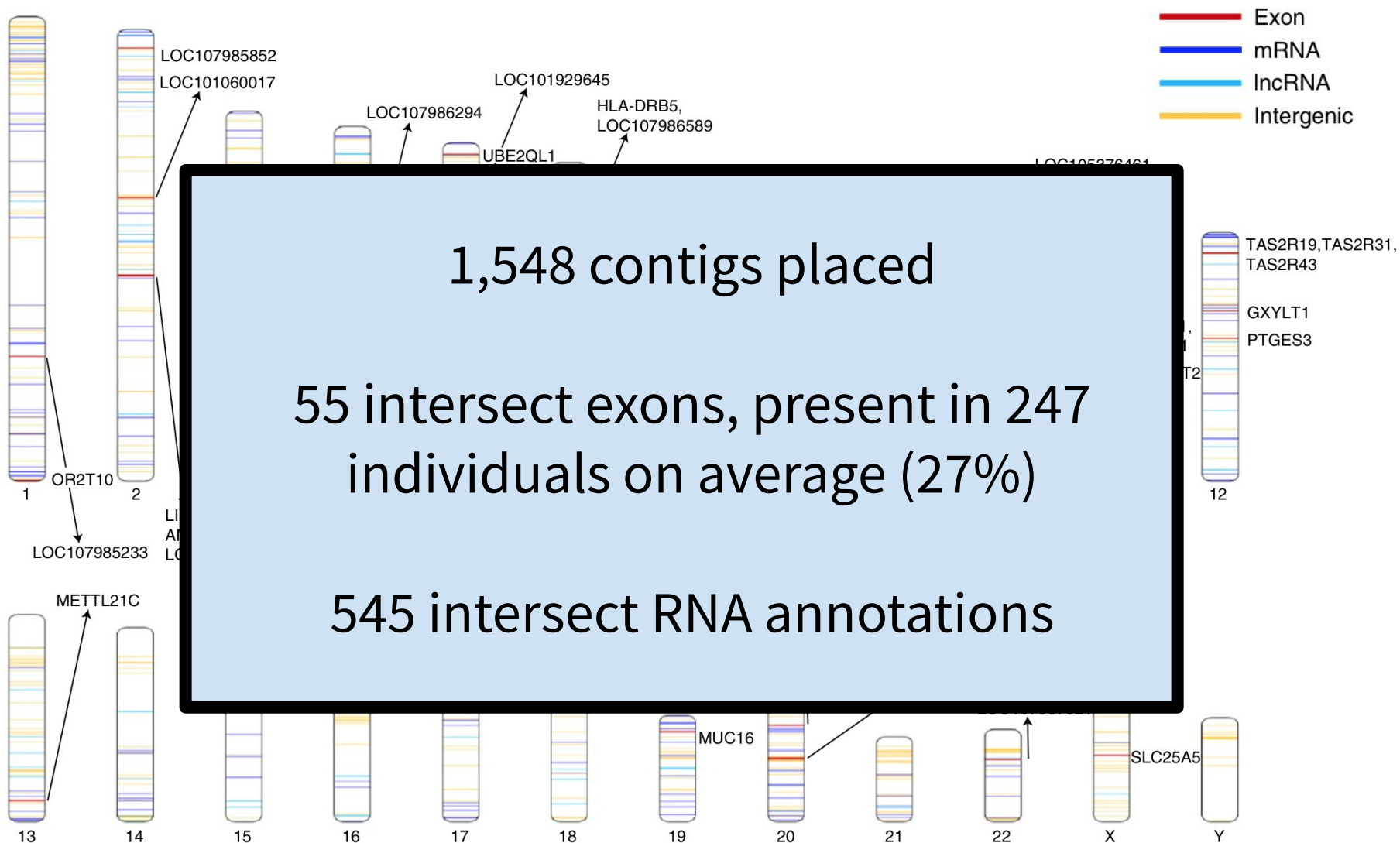
Don't line up to ref,
typically ignored





Pan-genome insertion locations



Pan-genome insertion locations









Pan-genome insertion locations



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

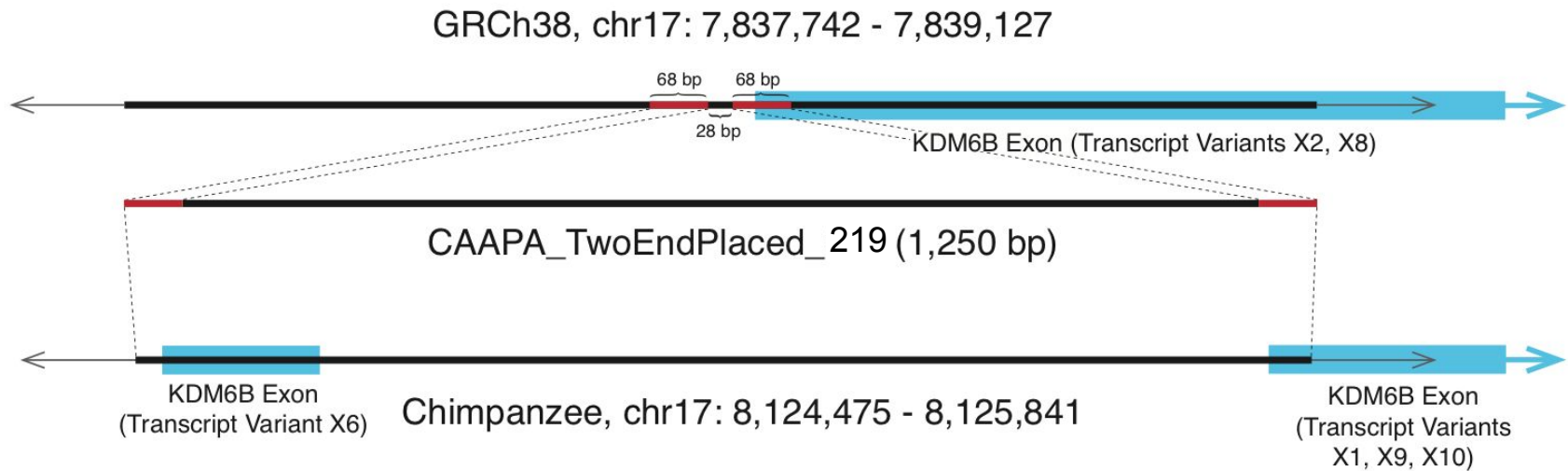
Rachel M. Sherman ^{1,2*}, Juliet Forman^{1,3}, Valentin Antonescu¹, Daniela Puiu¹, Michelle Daya⁴, Nicholas Rafaels⁴, Meher Preethi Boorgula⁴, Sameer Chavan⁴, Candelaria Vergara ⁵, Victor E. Ortega⁶, Albert M. Levin⁷, Celeste Eng⁸, Maria Yazdanbakhsh ⁹, James G. Wilson¹⁰, Javier Marrugo¹¹, Leslie A. Lange⁴, L. Keoki Williams¹², Harold Watson¹³, Lorraine B. Ware¹⁴, Christopher O. Olopade¹⁵, Olufunmilayo Olopade¹⁶, Ricardo R. Oliveira¹⁷, Carole Ober¹⁸, Dan L. Nicolae¹⁶, Deborah A. Meyers¹⁹, Alvaro Mayorga²⁰, Jennifer Knight-Madden²¹, Tina Hartert¹⁴, Nadia N. Hansel⁵, Marilyn G. Foreman²², Jean G. Ford²³, Mezbah U. Faruque²⁴, Georgia M. Dunston²⁵, Luis Caraballo¹¹, Esteban G. Burchard²⁶, Eugene R. Bleecker¹⁹, Maria I. Araujo²⁷, Edwin F. Herrera-Paz ²⁸, Monica Campbell⁴, Cassandra Foster⁵, Margaret A. Taub²⁹, Terri H. Beaty ³⁰, Ingo Ruczinski³¹, Rasika A. Mathias^{5,30}, Kathleen C. Barnes⁴ and Steven L. Salzberg ^{1,2,29,31*}

13 14 15 16 17 18 19 20 21 22 X Y

19, TAS2R31, 43
1
3

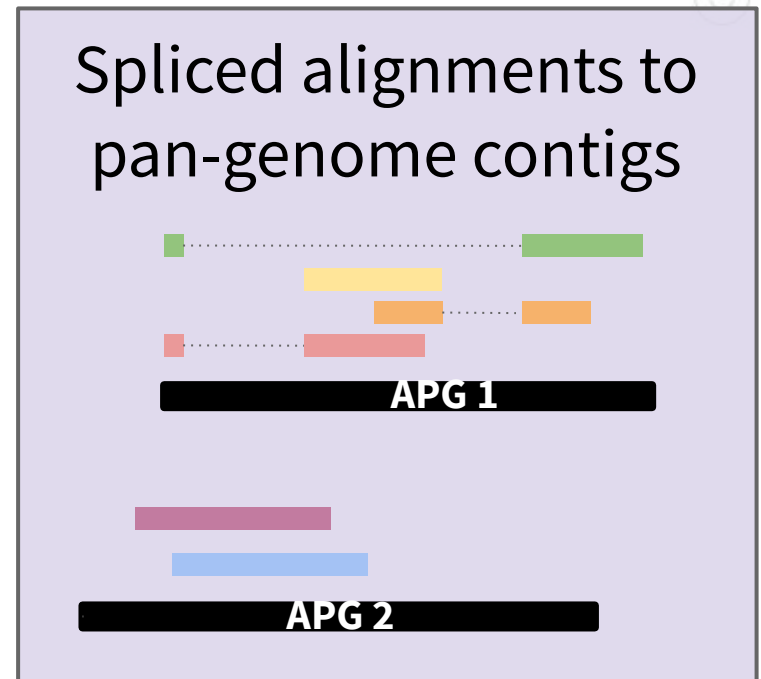
Are any of these sequences transcribed?

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



GTEEx reads align to pan-genome contigs

Preliminary analysis with 93 GTEEx RNA-seq samples
31 tissues, 3 samples each



GTEEx reads align to pan-genome contigs

Preliminary results on 93 samples

Alignments to 17,878 of 125,715 APG contigs (14%)

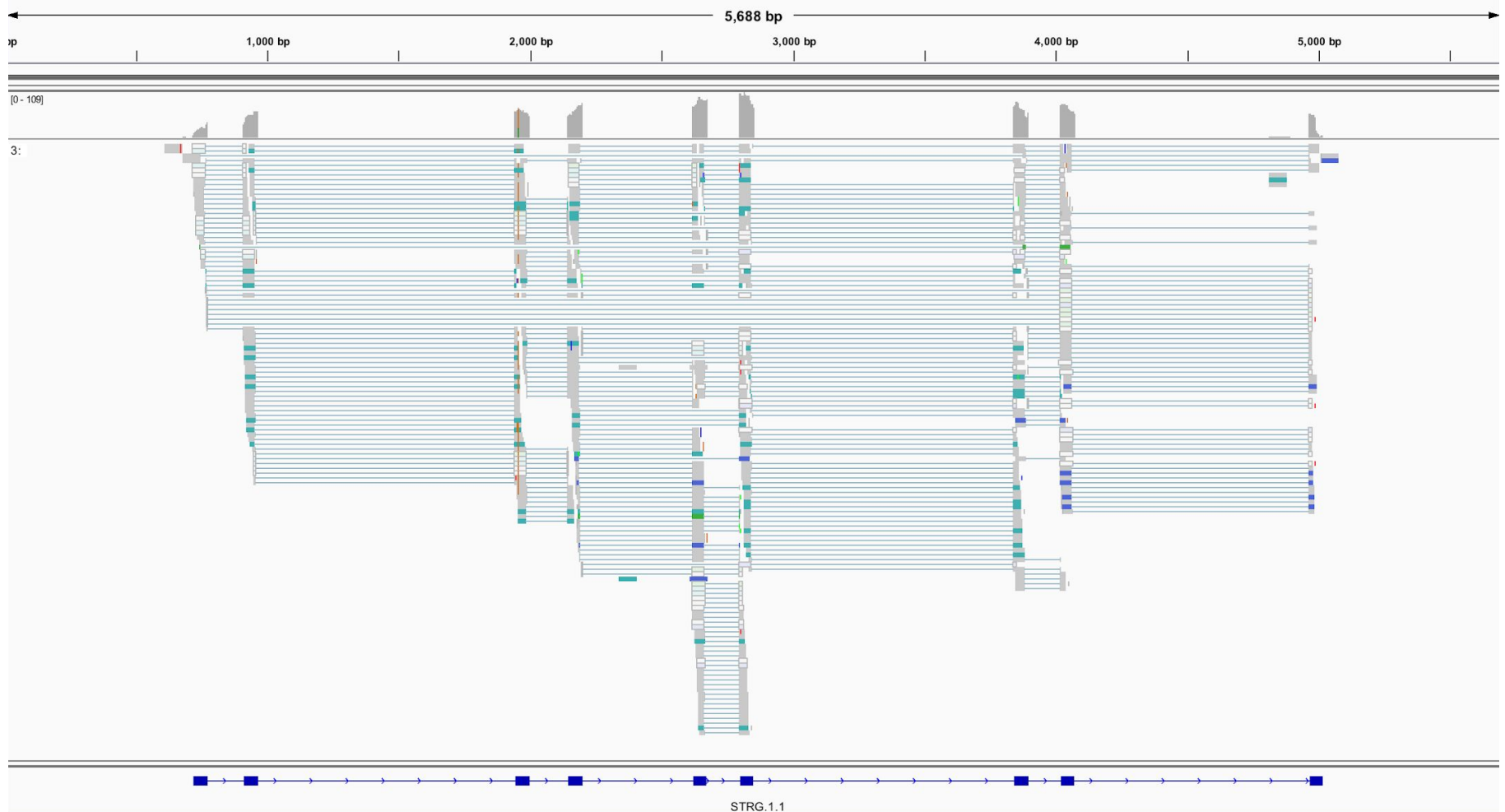
Canonically spliced alignments to 268 contigs (0.2%)

Ideal candidates are:

- Non-repetitive (unmasked)
- Many reads align with splicing
- Stringtie can assemble into transcript(s)

Non-reference exons present in insertions

5.7 kb CAAPA contig placed within MUC19 on chr12



Non-reference exons present in insertions

5.7 kb CAAPA contig placed within MUC19 on chr12

American Journal of Respiratory Cell and Molecular Biology

Home > All AJRCMB Issues > Vol. 45, No. 2 | Aug 01, 2011

Cloning and Characterization of Human *MUC19* Gene

Lingxiang Zhu ¹, Pakkei Lee ², Dongfang Yu ³, Shasha Tao ^{1,4}, and Yin Chen ¹

+ Author Affiliations

<https://doi.org/10.1165/rcmb.2010-0312OC> PubMed: [21075863](https://pubmed.ncbi.nlm.nih.gov/21075863/)

Received: July 27, 2010 Accepted: September 27, 2010

Abstract

Full Text

References

Supplements

Cited by

PDF

Abstract

The most recently discovered gel-forming mucin, *MUC19*, is expressed in both salivary glands and tracheal submucosal glands. We previously cloned the 3'-end partial sequence (AY236870), and here report the complete sequencing of the entire *MUC19* cDNA. One highly variable region (HVR) was discovered in the 5' end of *MUC19*. A total of 20 different splicing variants were detected in HVR, and 18 variants are able to translate into proteins along with the rest of the *MUC19* sequence. The longest variant of *MUC19* consists of 182 exons, with a transcript of approximately 25 kb. A central exon of approximately

We need more than just a variant catalog

nature
International journal of science

Letter | [Open Access](#) | Published: 26 July 2017

Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference

Lasse Maretty, Jacob Malte Jensen [...] Mikkel Heide Schierup 

Genome Biology

[Home](#) [About](#) [Articles](#) [Submission Guidelines](#)

Method | [Open Access](#) | Published: 31 July 2019

HUPAN: a pan-genome analysis pipeline for human genomes

[Zhongqu Duan](#), [Yuyang Qiao](#), [Jinyuan Lu](#), [Huimin Lu](#), [Wenmin Zhang](#), [Fazhe Yan](#), [Chen Sun](#), [Zhiqiang Hu](#), [Zhen Zhang](#), [Guichao Li](#), [Hongzhan Chen](#), [Zhen Xiang](#), [Zhenggang Zhu](#), [Hongyu Zhao](#), [Yingyan Yu](#)  & [Chaochun Wei](#) 

nature
COMMUNICATIONS 

Article | [Open Access](#) | Published: 24 November 2016

An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes

Yun Sung Cho, Hyunho Kim, Hak-Min Kim, Sungwoong Jho, JeHoon Jun, Yong Joo Lee, Kyun Shik Chae, Chang Geun Kim, Sangsoo Kim, Anders Eriksson, Jeremy S. Edwards, Semin Lee, Byung Chul Kim, Andrea Manica, Tae-Kwang Oh, George M. Church  & Jong Bhak 



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

HOME |

Search

New Results

Genotyping structural variants in pangenome graphs using the vg toolkit

Glenn Hickey, David Heller, Jean Monlong, Jonas Andreas Sibbesen, Jouni Siren, Jordan Eizenga, Eric Dawson, Erik Garrison, Adam Novak, Benedict Paten

doi: <https://doi.org/10.1101/654566>

Acknowledgments

Steven Salzberg

Daniela Puiu

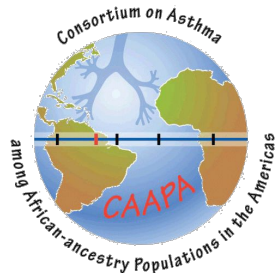
Valentin Antonescu

Juliet Forman

William Cho



JOHNS HOPKINS
UNIVERSITY



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 Antonescu¹, Daniela Puiu¹, Michelle Daya⁴, Nicholas Rafaels⁴, Meher Preethi Boorgula⁴, Sameer Chavan⁴, Candelaria Vergara⁵, Victor E. Ortega⁶, Albert M. Levin⁷, Celeste Eng⁸, Maria Yazdanbakhsh⁹, James G. Wilson¹⁰, Javier Marrugo¹¹, Leslie A. Lange⁴, L. Keoki Williams¹², Harold Watson¹³, Lorraine B. Ware¹⁴, Christopher O. Olopade¹⁵, Olufunmilayo Olopade¹⁶, Ricardo R. Oliveira¹⁷, Carole Ober¹⁸, Dan L. Nicolae¹⁶, Deborah A. Meyers¹⁹, Alvaro Mayorga²⁰, Jennifer Knight-Madden²¹, Tina Hartert¹⁴, Nadia N. Hansel⁵, Marilyn G. Foreman²², Jean G. Ford²³, Mezbah U. Faruque²⁴, Georgia M. Dunston²⁵, Luis Caraballo¹¹, Esteban G. Burchard²⁶, Eugene R. Bleecker¹⁹, Maria I. Araujo²⁷, Edwin F. Herrera-Paz²⁸, Monica Campbell⁴, Cassandra Foster⁵, Margaret A. Taub²⁹, Terri H. Beaty³⁰, Ingo Ruczinski³¹, Rasika A. Mathias^{5,30}, Kathleen C. Barnes⁴ and Steven L. Salzberg^{1,2,29,31*}

A decorative network diagram consisting of interconnected nodes and lines, rendered in light gray and blue tones, framing the central text. The nodes vary in size and some are highlighted with a blue outline or filled with a dark blue color. The lines connecting them are thin and light gray.

Questions?



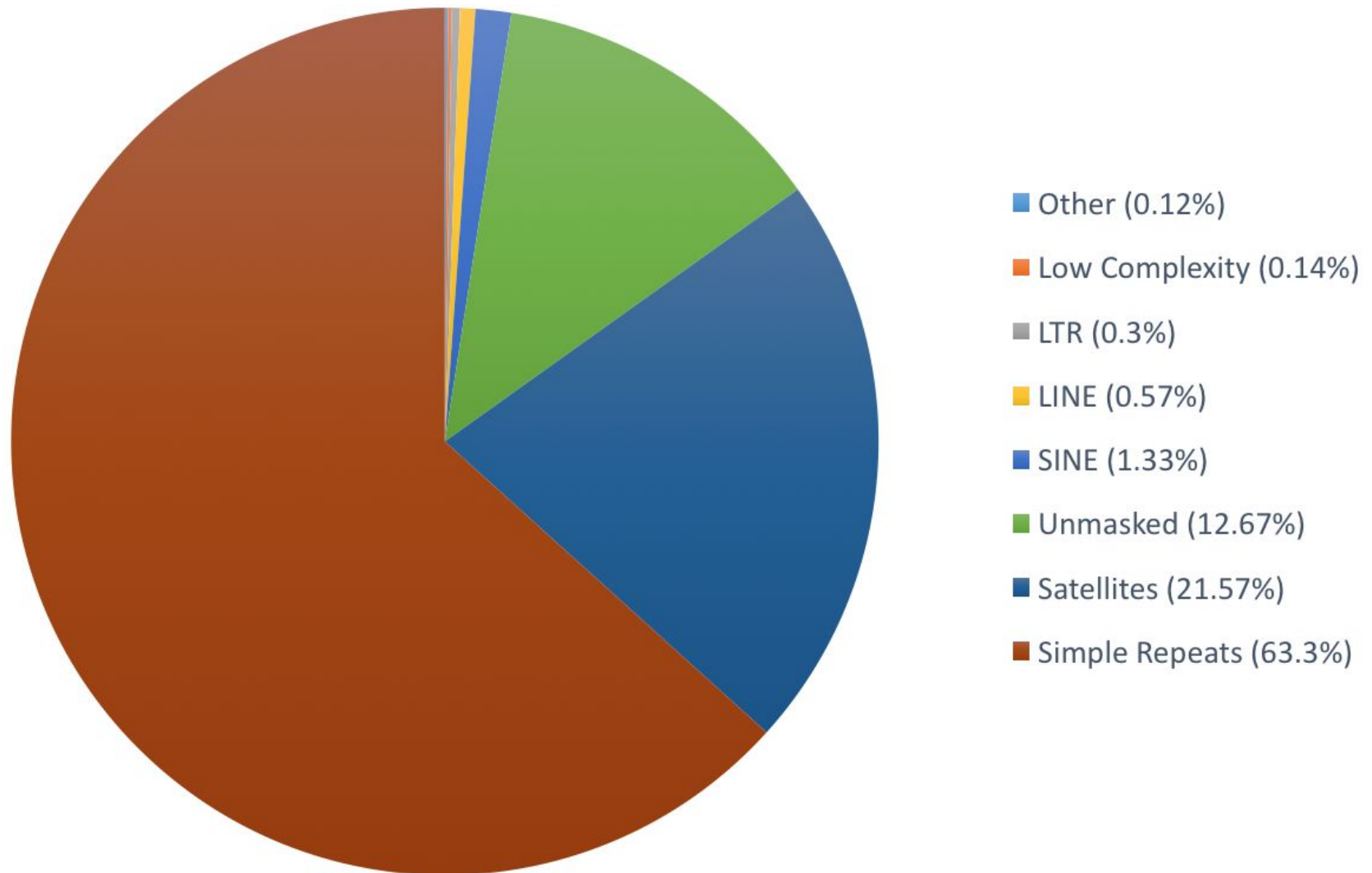
Additional Slides

Pan-genome stats

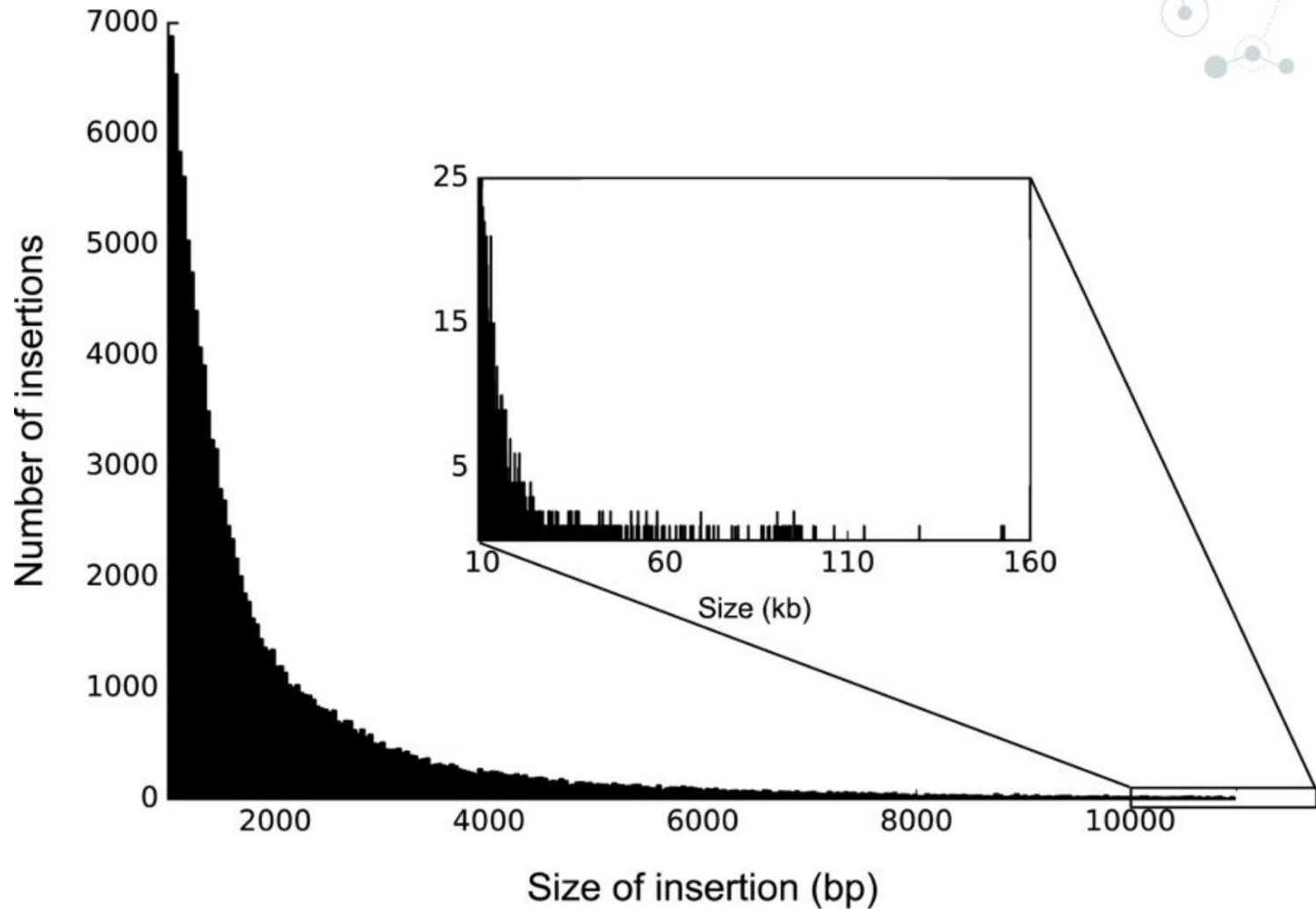
	# Contigs	Total Length (bp)	Longest Contig
Placed	1,548	4,354,696	79,938
Unplaced	124,167	292,130,588	152,806
Total	125,715	296,485,284	152,806
Non-singleton	61,410	160,475,353	152,806

- 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

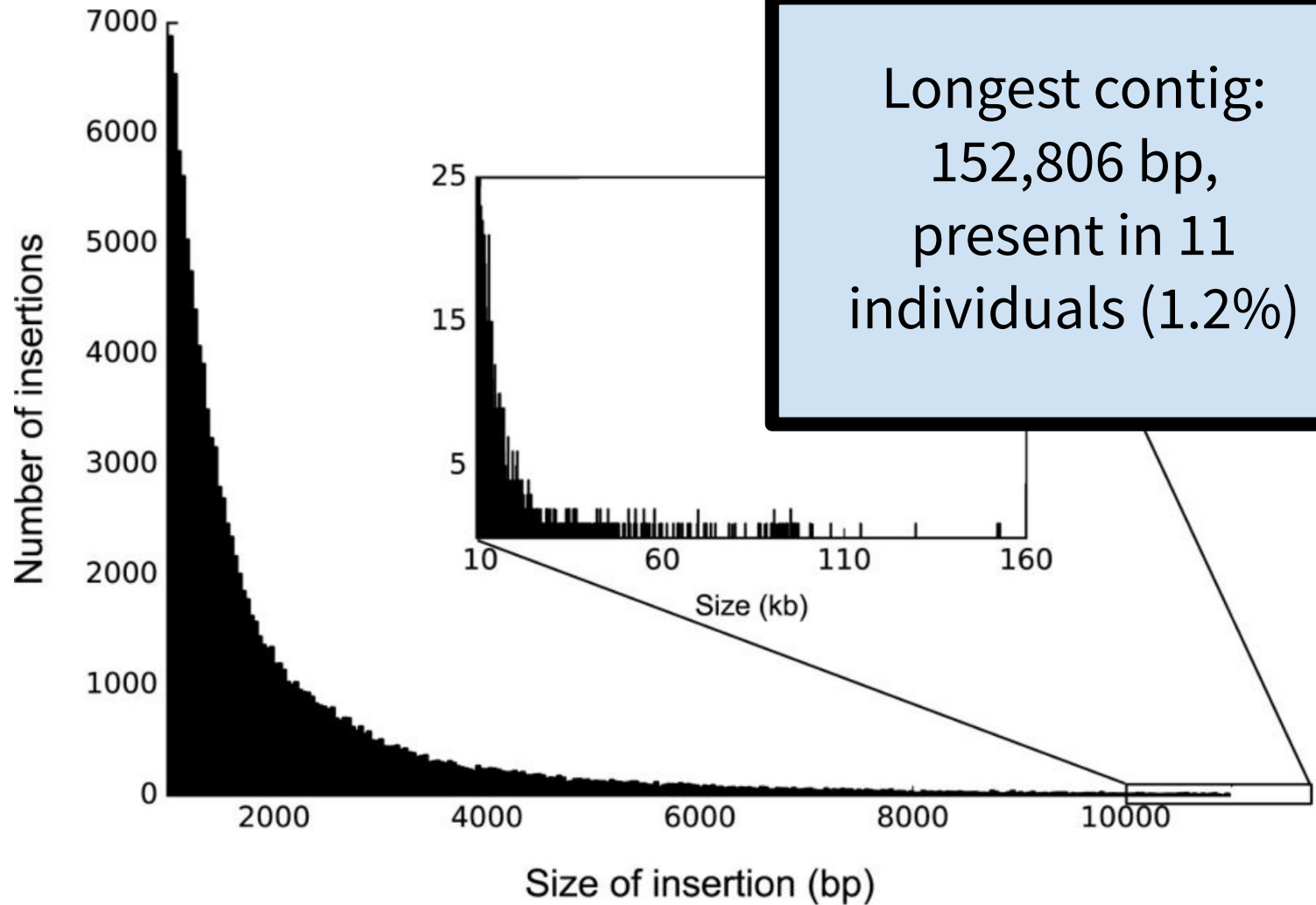
Repeat content in pan-genome contigs



Pan-genome contig size distribution



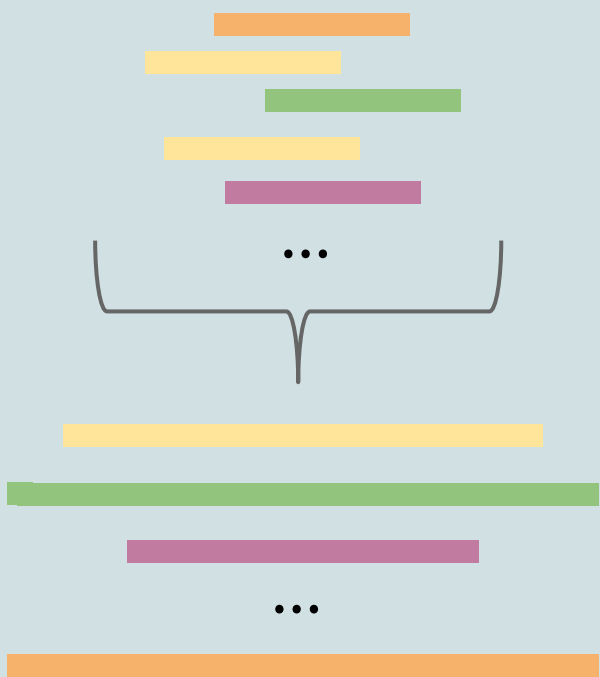
Pan-genome contig size distribution



Pan-genome contigs in other WGS cohorts

Assemble unaligned reads,
per individual

Don't line up to ref,
typically ignored



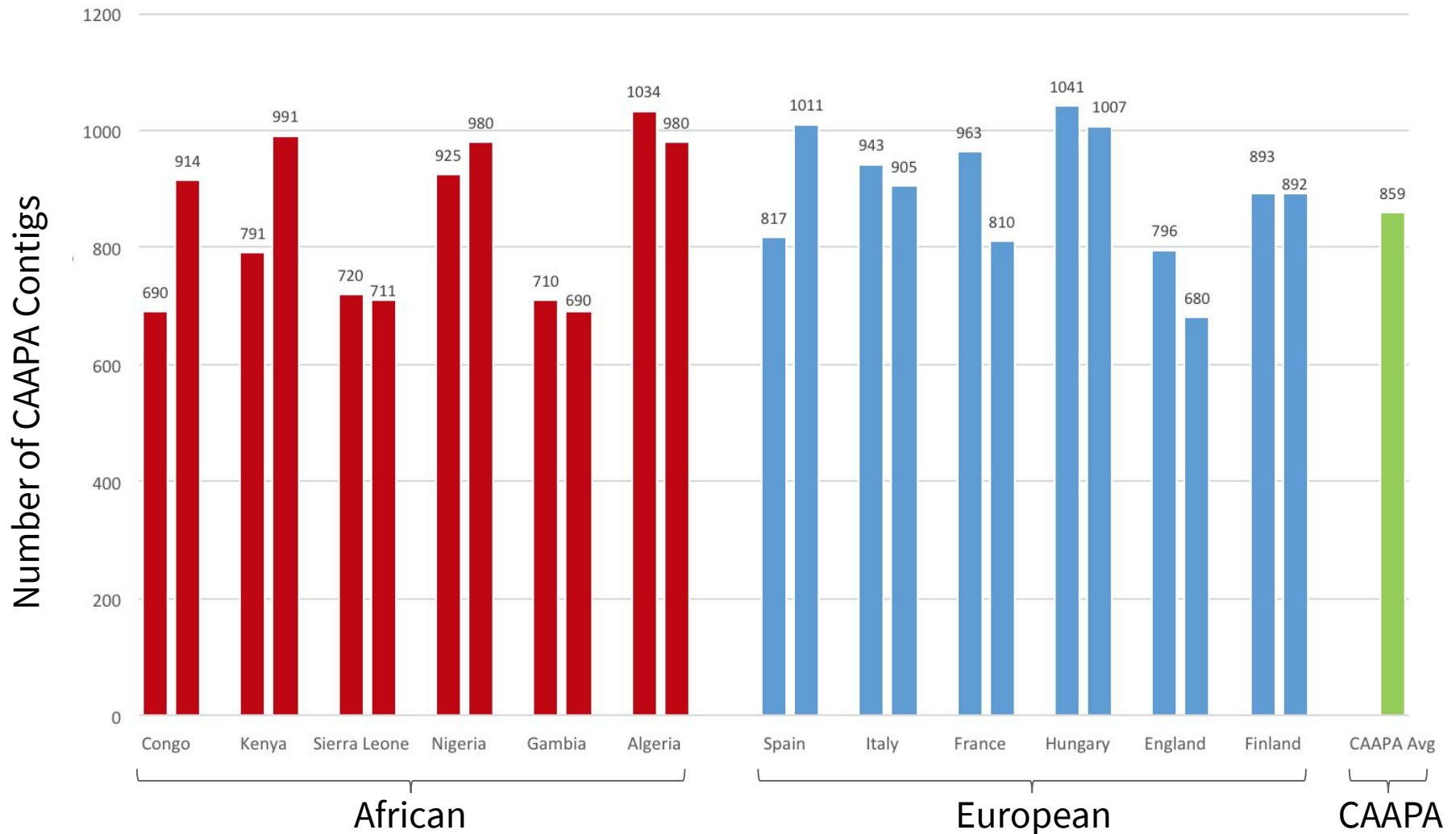
Align to pan-genome
contigs



Call presence/absence



Pan-genome contigs in SGRP populations



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

The pan-genome is still open

