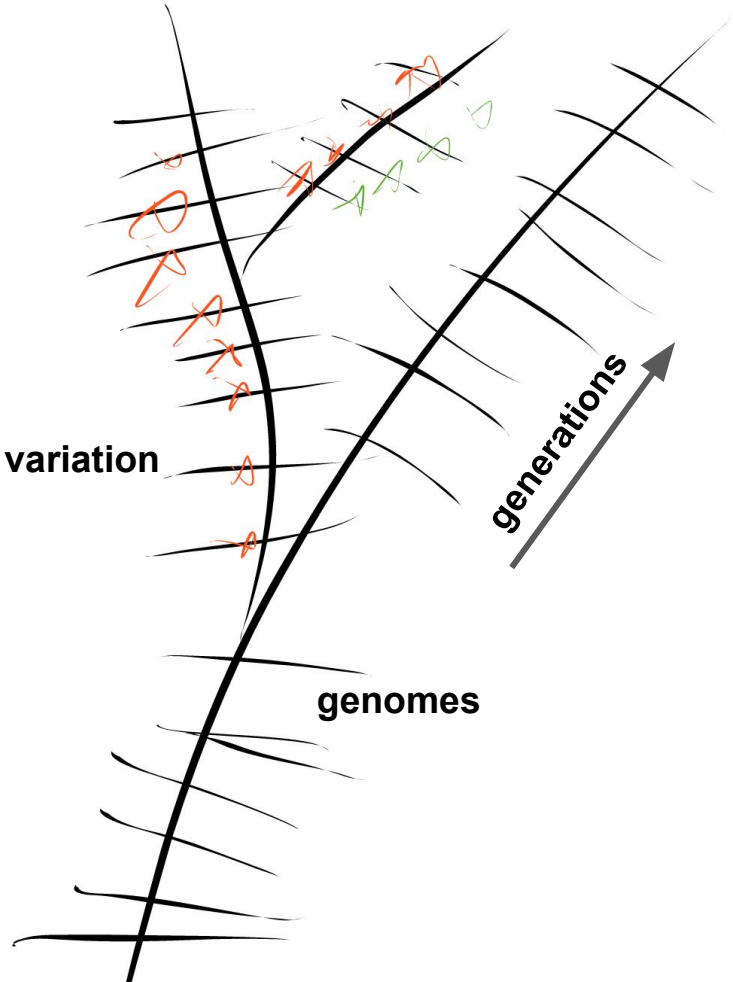
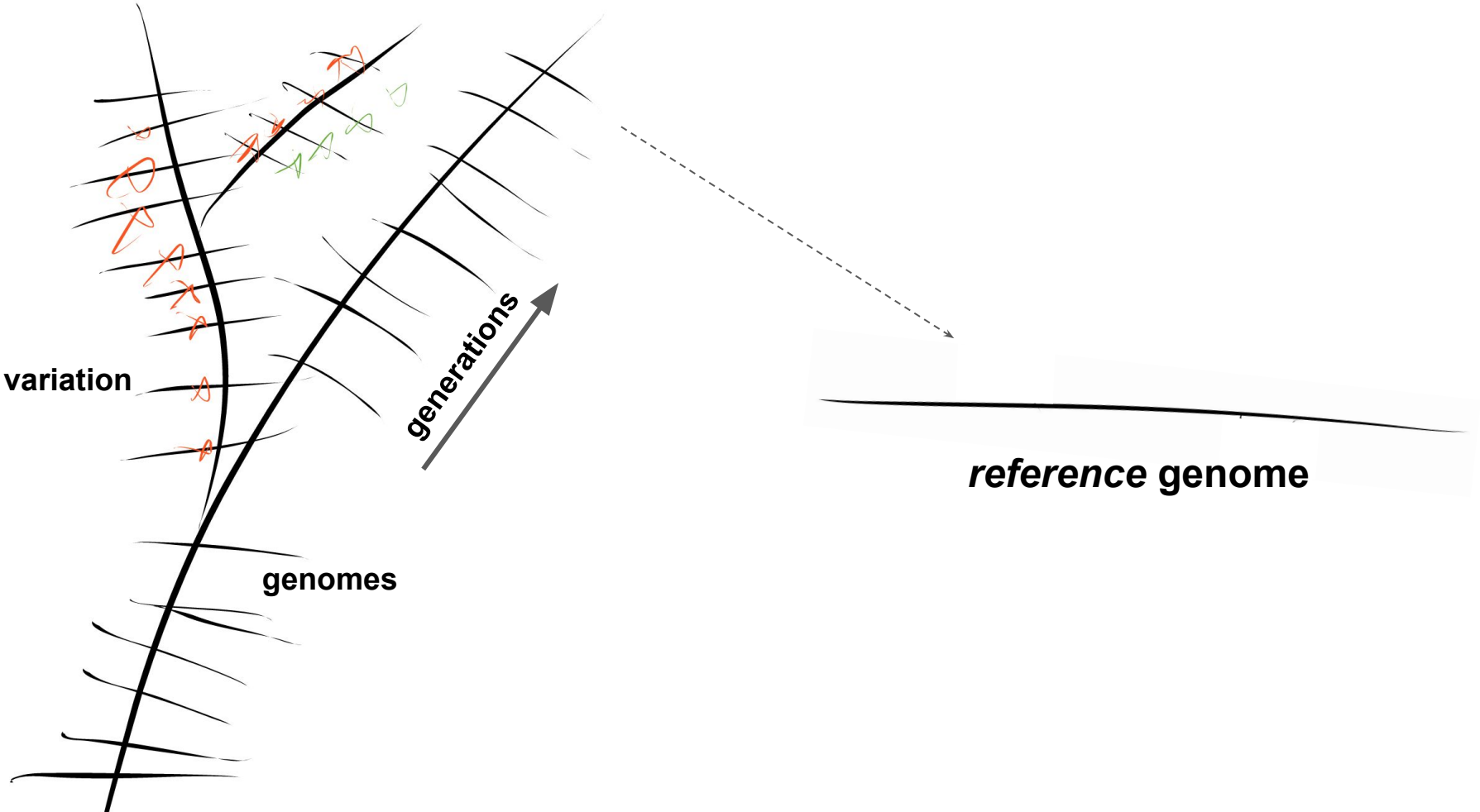
Building, understanding, and using pangenomes

Erik Garrison

University of Tennessee (UTHSC), Memphis

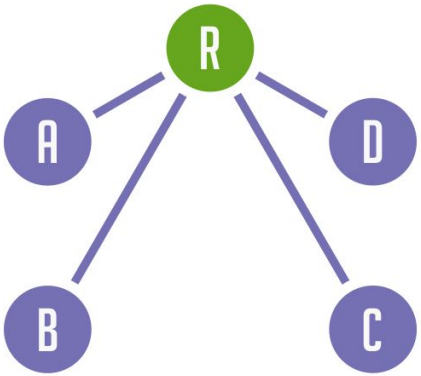
@Workshop on Genomics, Český Krumlov
January 13, 2024


variation
genomes
generations

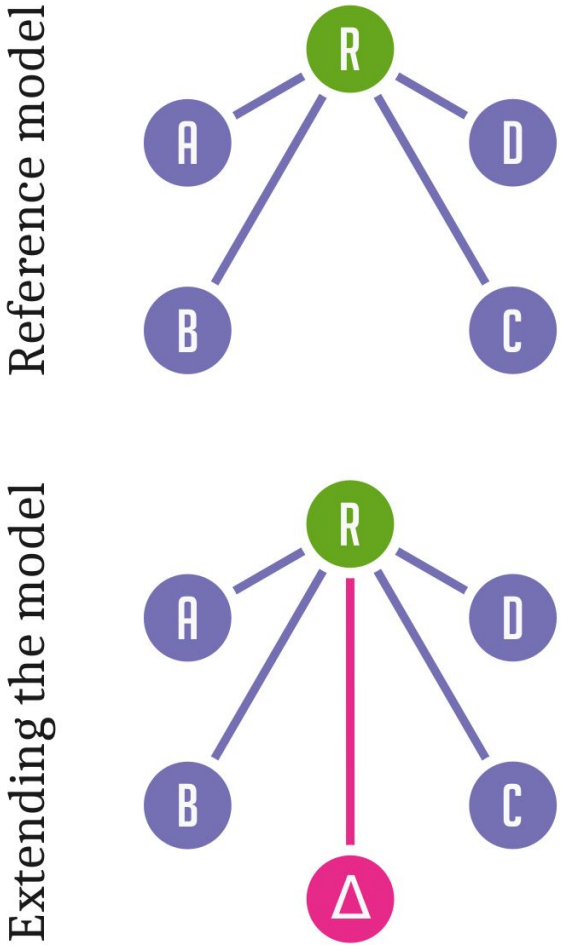

variation
generations
genomes
reference genome

Genomic

Reference model



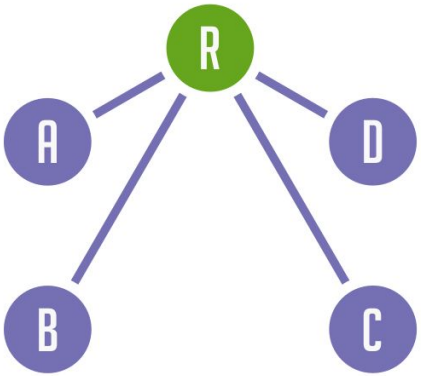
Genomic



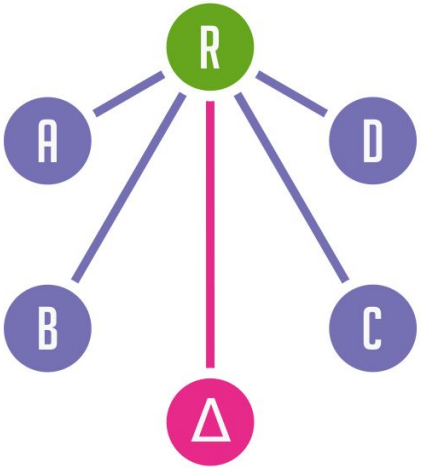
Δ: new genome; R: reference genome.
Figure from Eizenga et al., 2020.

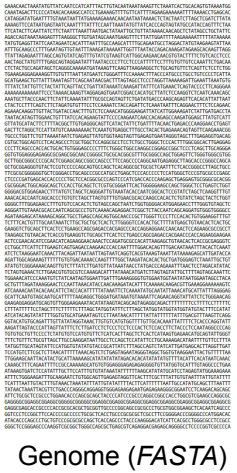
Genomic

Reference model

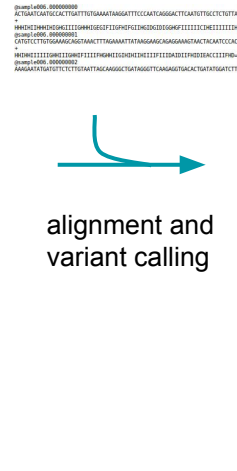


Extending the model





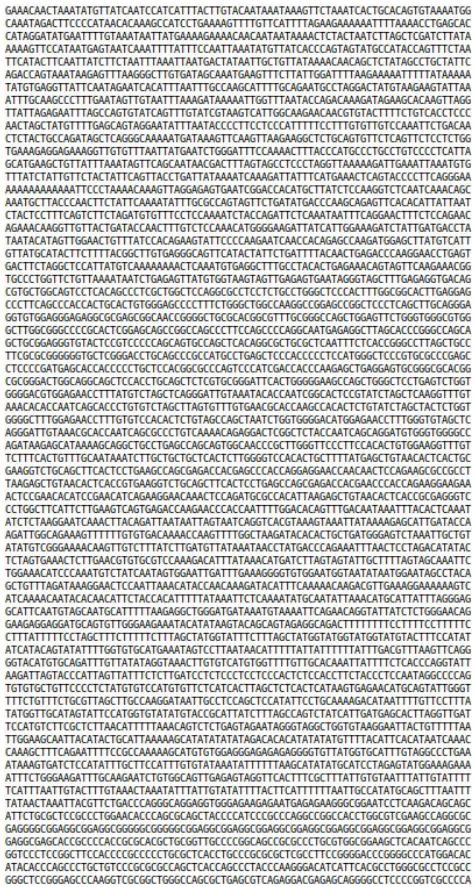
Genome (*FASTA*)

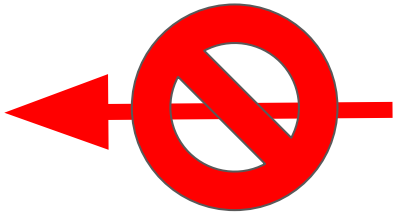


Variation (*VCF*)

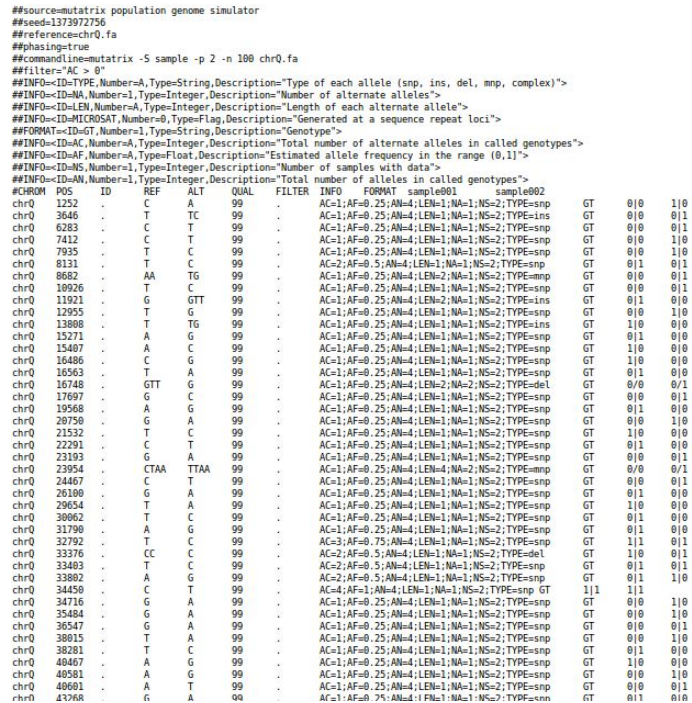
Reads (*FASTQ*)

alignment and variant calling





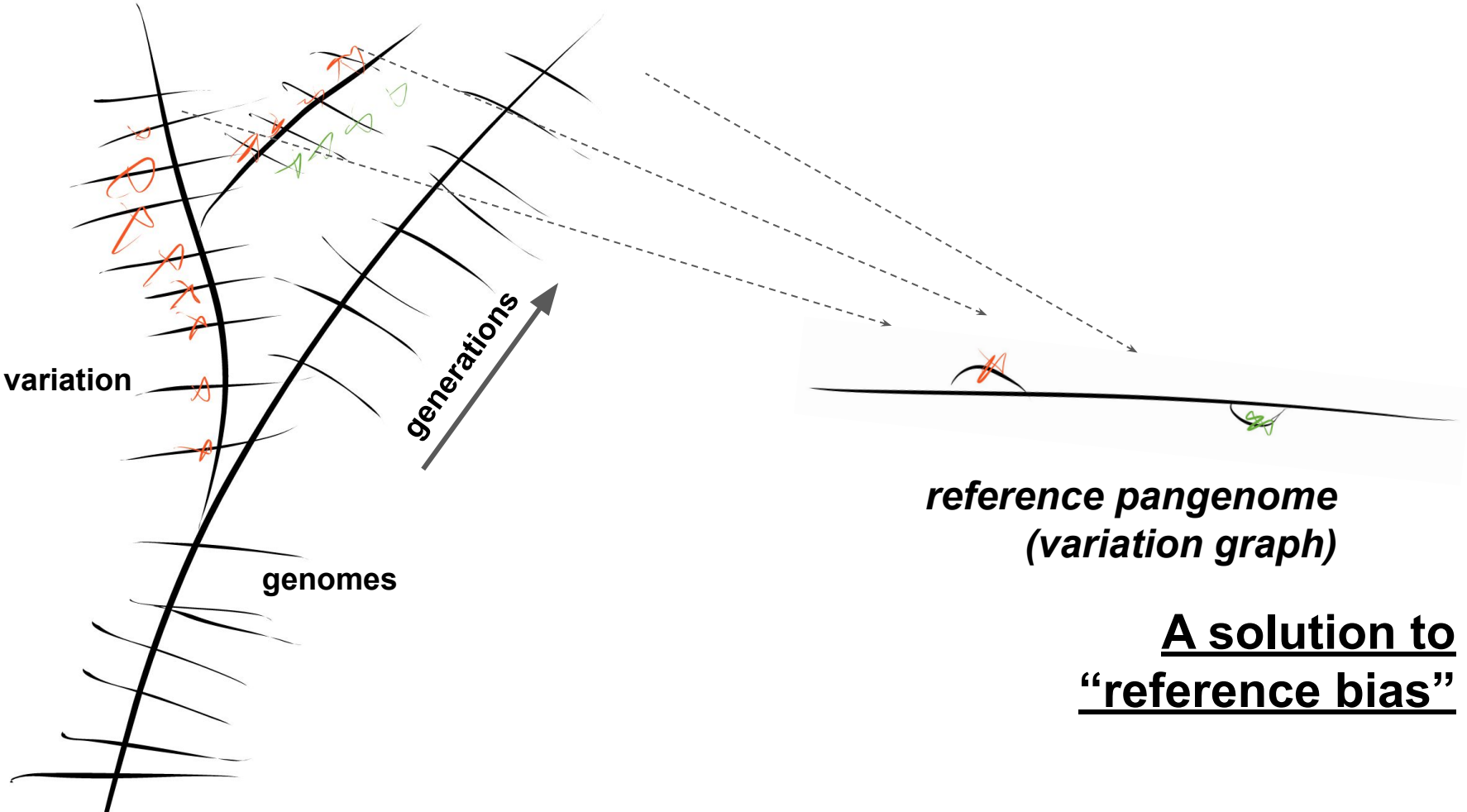
We cannot update a linear reference sequence



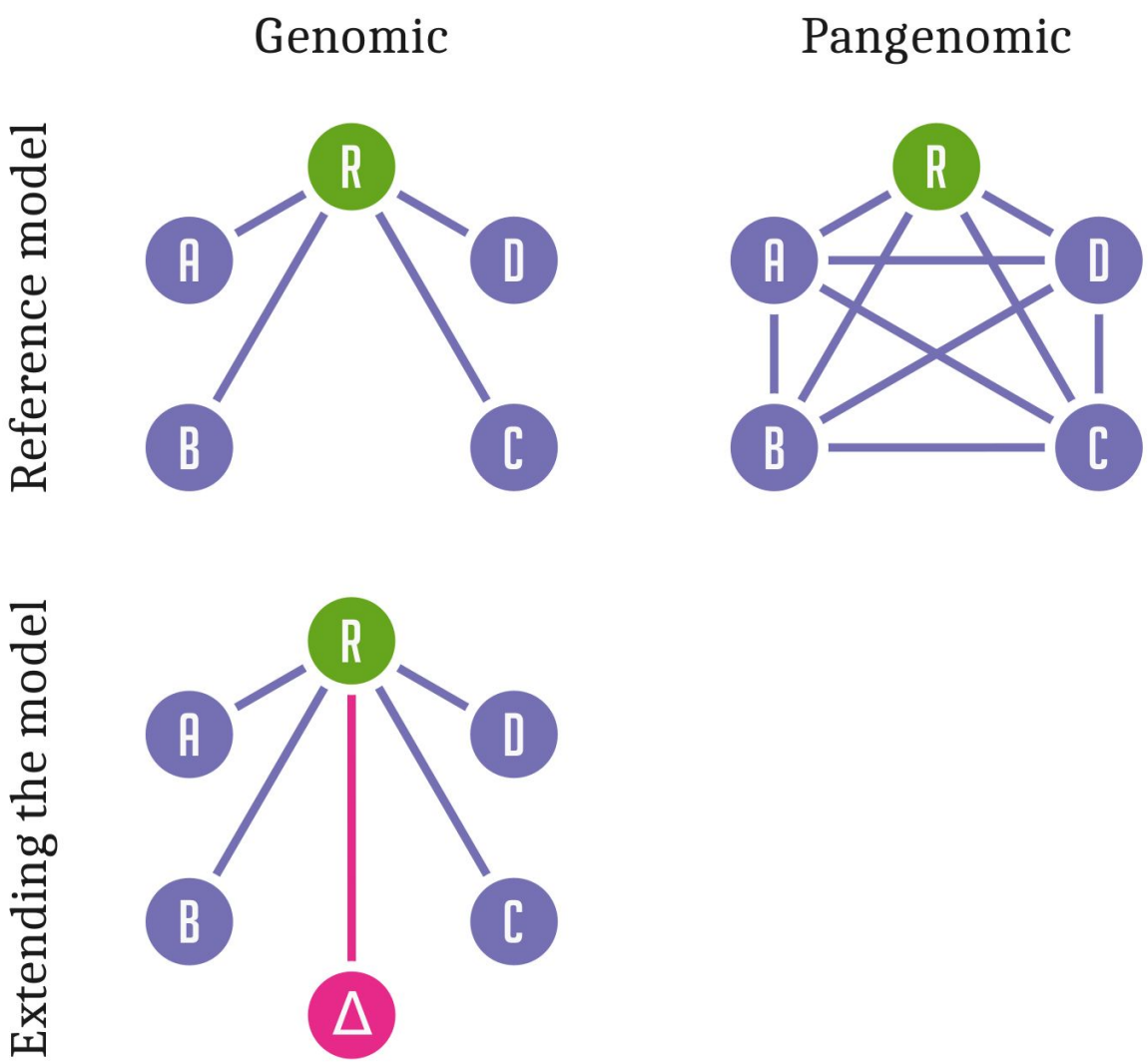
```
##source=mutatrix population genome simulator
##seed=1373972756
##reference=chrQ.fa
##phasing=true
##commandline=mutatrix -S sample -p 2 -n 100 chrQ.fa
##filter="AC > 0"
##INFO=<ID=TYPE,Number=A,Type=String,Description="Type of each allele (snp, ins, del, mnp, complex)">
##INFO=<ID=NA,Number=1,Type=Integer,Description="Number of alternate alleles">
##INFO=<ID=LEN,Number=A,Type=Integer,Description="Length of each alternate allele">
##INFO=<ID=MICROSAT,Number=0,Type=Flag,Description="Generated at a sequence repeat loci">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Total number of alternate alleles in called genotypes">
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated allele frequency in the range (0,1]">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
#CHROM  POS    ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  sample001       sample002
chrQ    1252   .       C       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    3646   .       T       TC      99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=ins      GT      0|0     0|1
chrQ    6283   .       C       T       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     0|1
chrQ    7412   .       C       T       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    7935   .       T       C       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    8131   .       T       C       99      .       AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=snp       GT      0|1     0|1
chrQ    8682   .       AA      TG      99      .       AC=1;AF=0.25;AN=4;LEN=2;NA=1;NS=2;TYPE=mnp      GT      0|0     0|1
chrQ    10926  .       T       C       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     0|1
chrQ    11921  .       G       GTT     99      .       AC=1;AF=0.25;AN=4;LEN=2;NA=1;NS=2;TYPE=ins      GT      0|1     0|0
chrQ    12955  .       T       G       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    13808  .       T       TG      99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=ins      GT      1|0     0|0
chrQ    15271  .       A       G       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    15407  .       A       C       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      1|0     0|0
chrQ    16486  .       C       G       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      1|0     0|0
chrQ    16563  .       T       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    16748  .       GTT     G       99      .       AC=1;AF=0.25;AN=4;LEN=2;NA=2;NS=2;TYPE=del      GT      0/0     0/1
chrQ    17697  .       G       C       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     0|1
chrQ    19568  .       A       G       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    20750  .       G       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    21532  .       T       C       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      1|0     0|0
chrQ    22291  .       C       T       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    23193  .       G       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     0|1
chrQ    23954  .       CTAA    TTAA    99      .       AC=1;AF=0.25;AN=4;LEN=4;NA=2;NS=2;TYPE=mnp      GT      0/0     0/1
chrQ    24467  .       C       T       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     0|1
chrQ    26100  .       G       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    29654  .       T       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      1|0     0|0
chrQ    30062  .       T       C       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    31790  .       A       G       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    32792  .       T       C       99      .       AC=3;AF=0.75;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      1|1     0|1
chrQ    33376  .       CC      C       99      .       AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=del       GT      1|0     0|1
chrQ    33403  .       T       C       99      .       AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=snp       GT      0|1     0|1
chrQ    33802  .       A       G       99      .       AC=2;AF=0.5;AN=4;LEN=1;NA=1;NS=2;TYPE=snp       GT      0|1     1|0
chrQ    34450  .       C       T       99      .       AC=4;AF=1;AN=4;LEN=1;NA=1;NS=2;TYPE=snp GT      1|1     1|1
chrQ    34716  .       G       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    35484  .       G       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    36547  .       G       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     0|1
chrQ    38015  .       T       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    38281  .       T       C       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
chrQ    40467  .       A       G       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      1|0     0|0
chrQ    40581  .       A       G       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     1|0
chrQ    40601  .       A       T       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|0     0|1
chrQ    43268  .       G       A       99      .       AC=1;AF=0.25;AN=4;LEN=1;NA=1;NS=2;TYPE=snp      GT      0|1     0|0
```

Genome (*FASTA*)

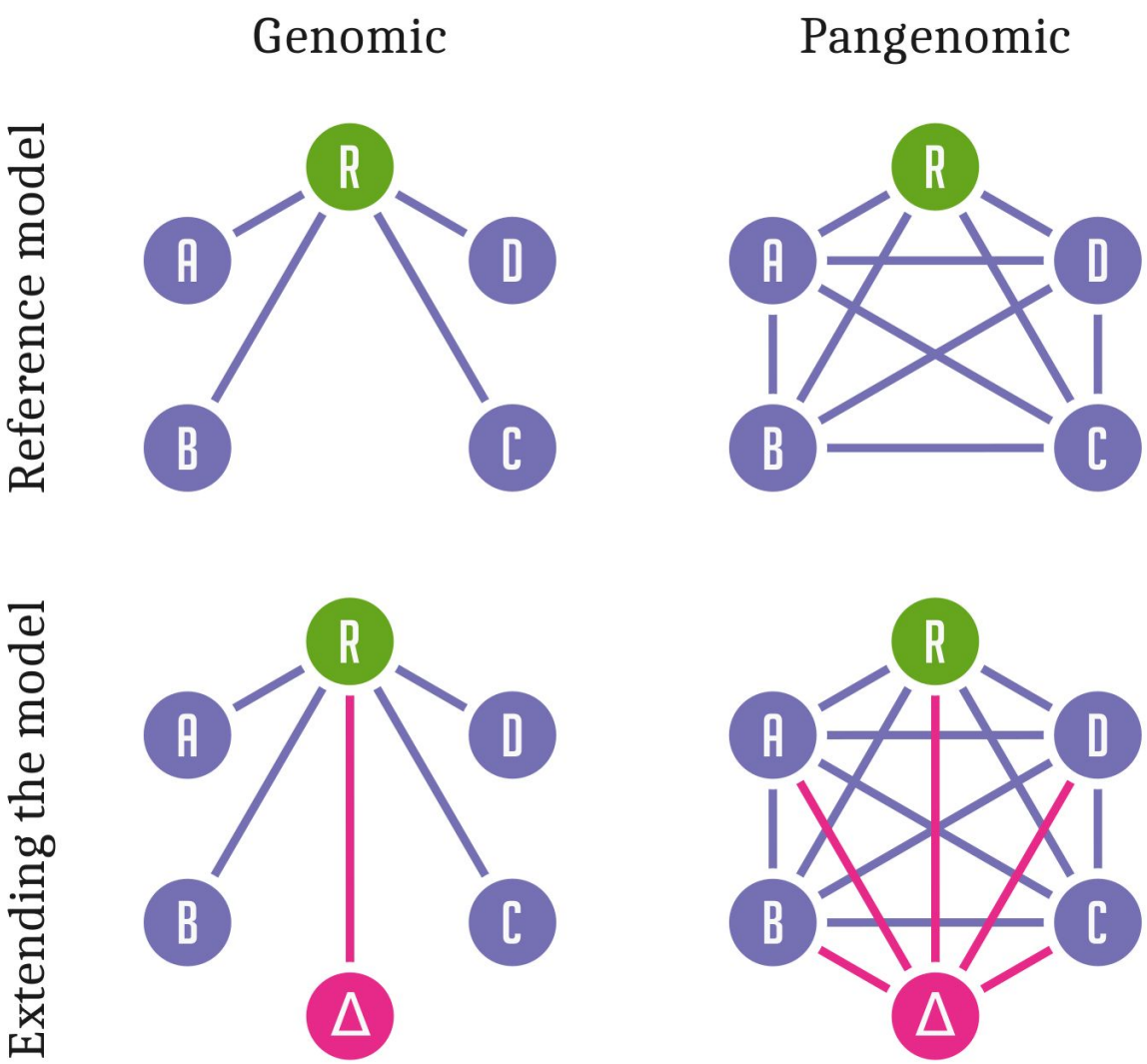
Variation (*VCF*)



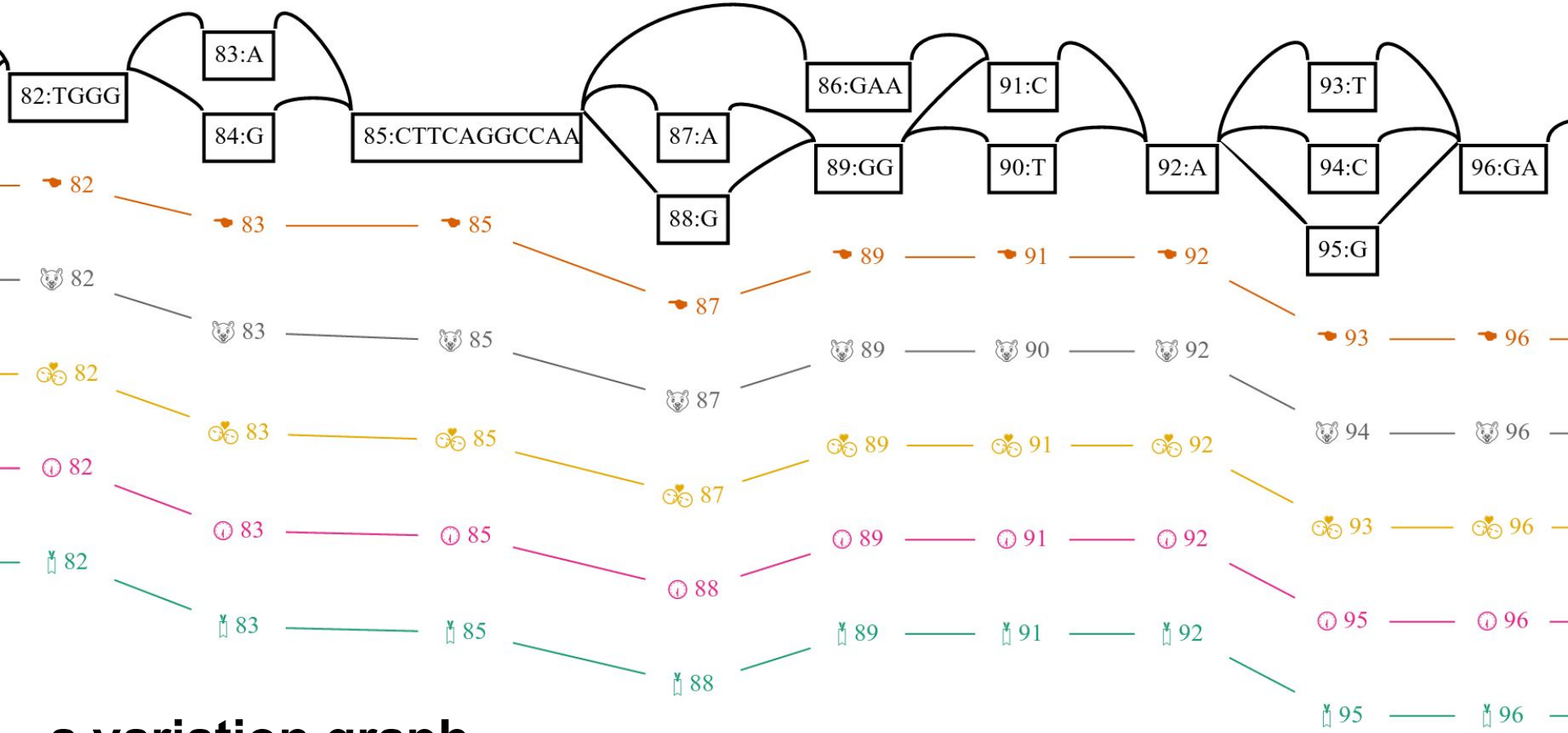
A solution to "reference bias"



Δ: new genome;
R: reference genome.
Figure from
Eizenga et al., 2020.

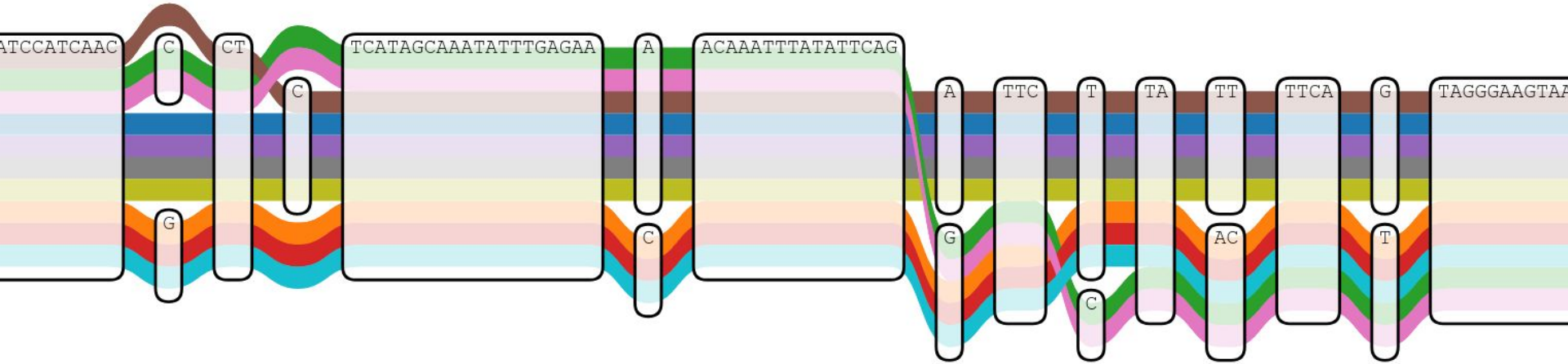


Δ: new genome;
R: reference genome.
Figure from
Eizenga et al., 2020.

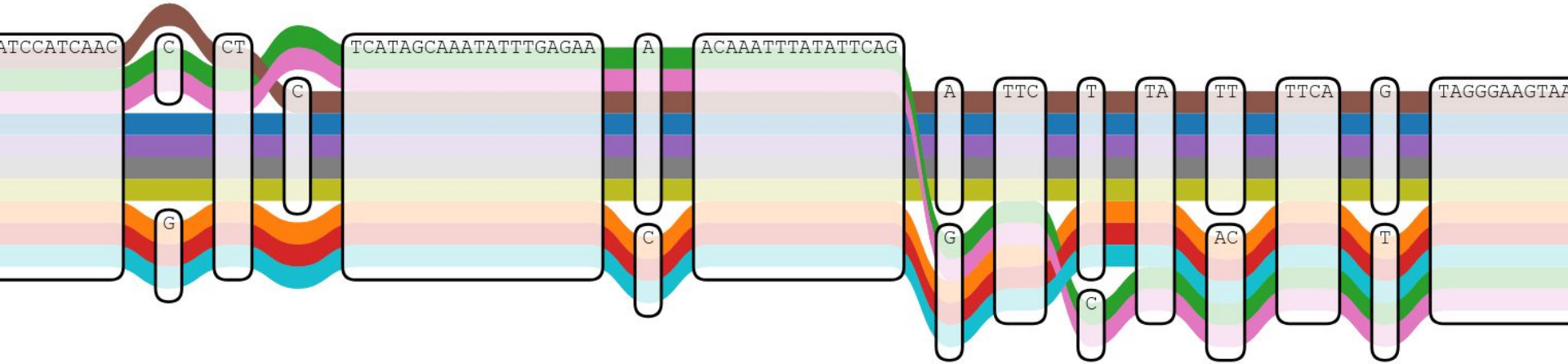


a variation graph

Variation graphs answer a key problem in bioinformatics:

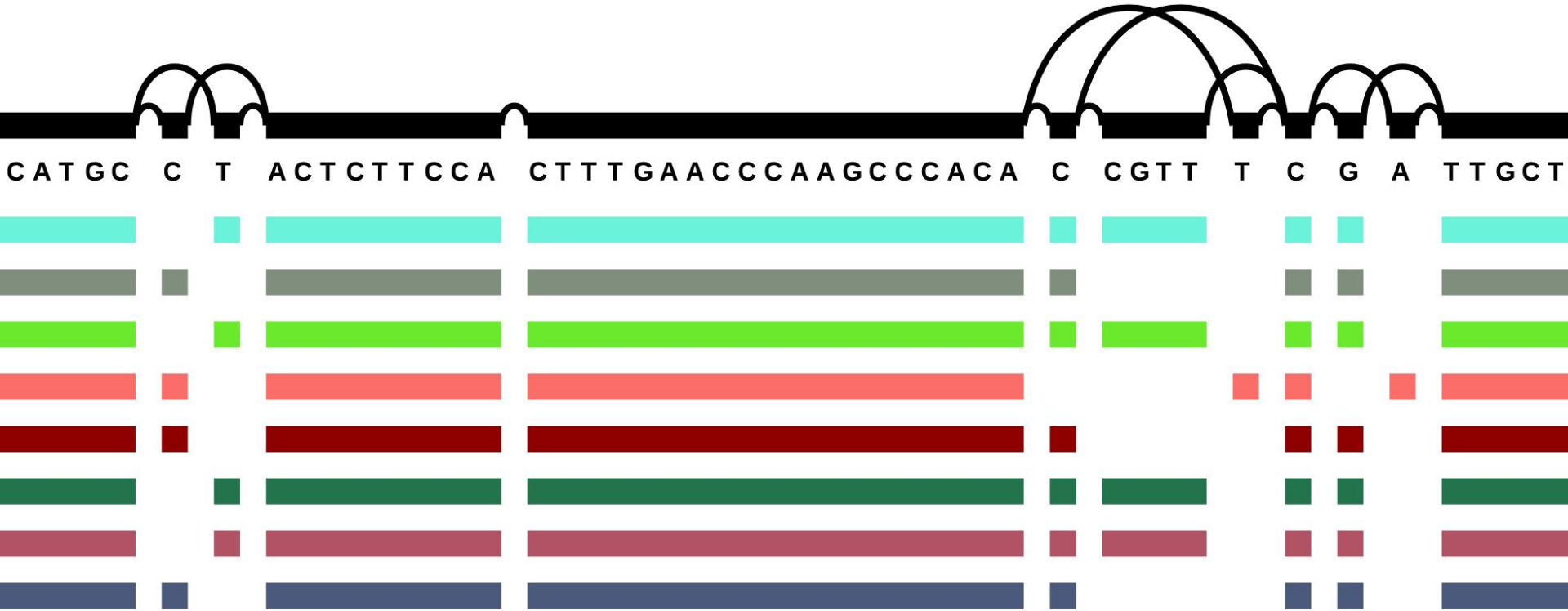


Variation graphs answer a key problem in bioinformatics:

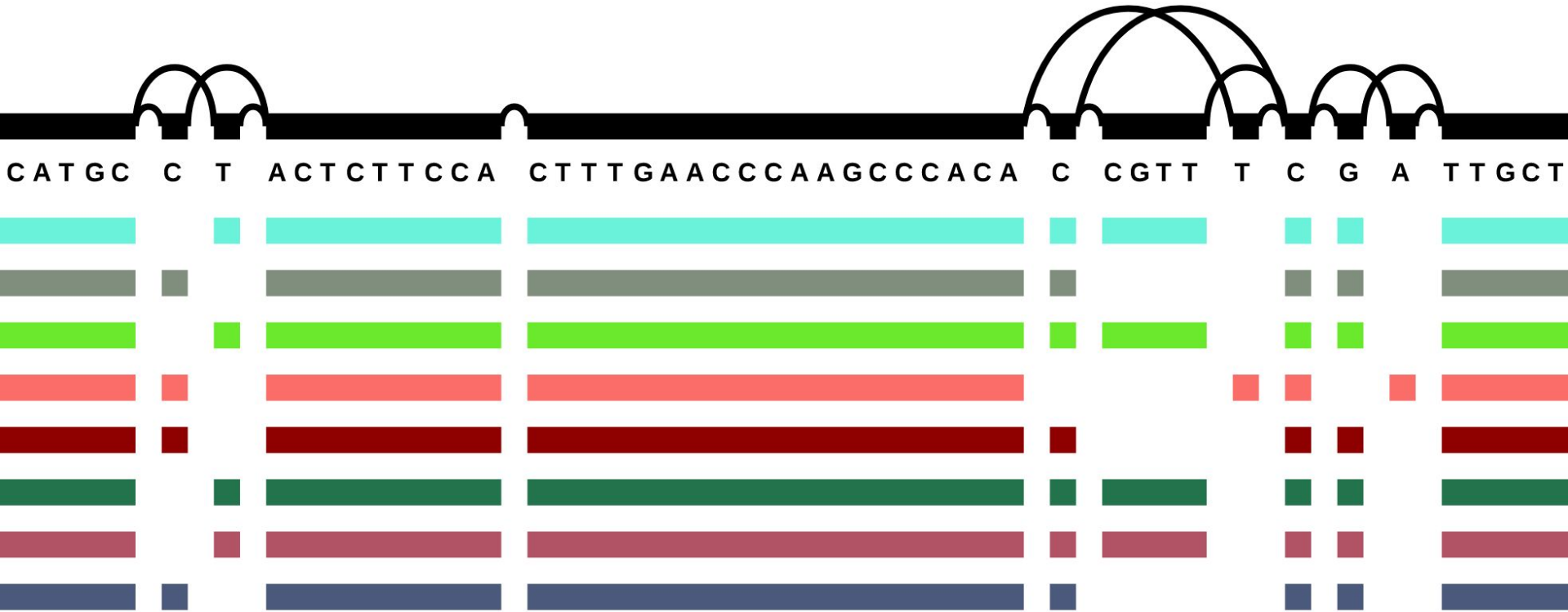


How to represent *both* sequences and *any* kind of variation between them.

variation graphs are ***pangenome*** *models*



variation graphs are *pangenome* models



… which give us a simple way to project many genomes into vector spaces.

New ideas often have a long history

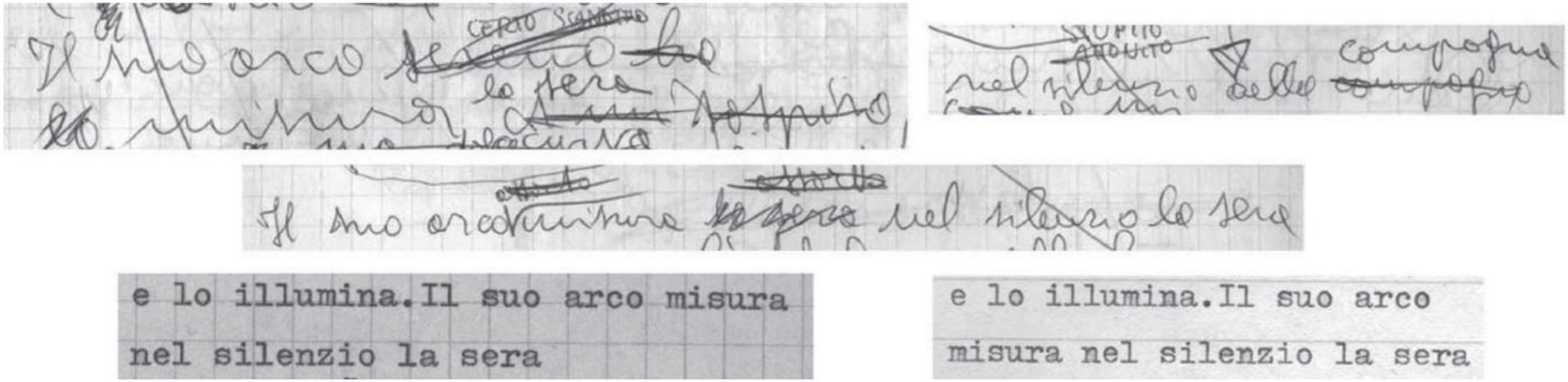
This all seems cool and “new” but ideas are rarely that.

Pangenomes and variation graphs have a long† history.

(†for genomics)

St. Agatha pipetting a biosample into a nanopore sequencer c. 1420


pipette
biosample
nanopore
sequencer



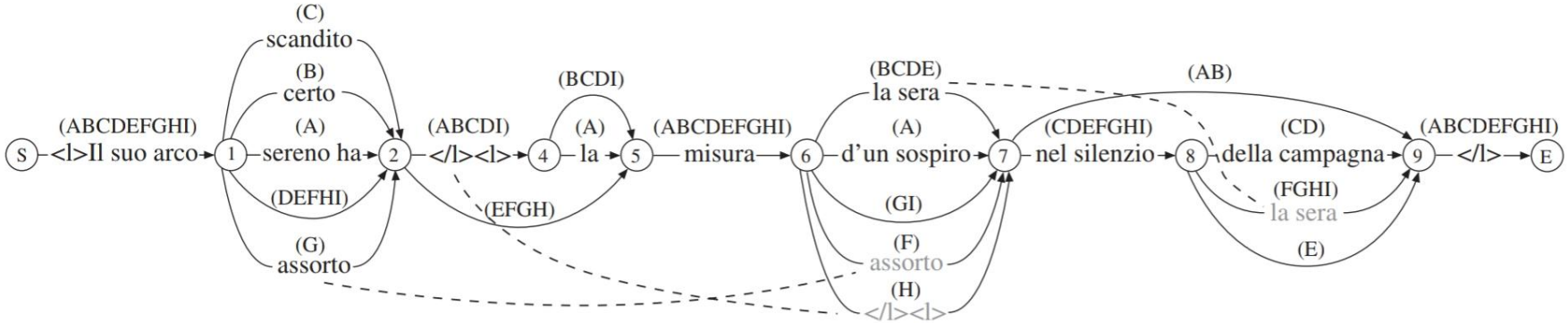
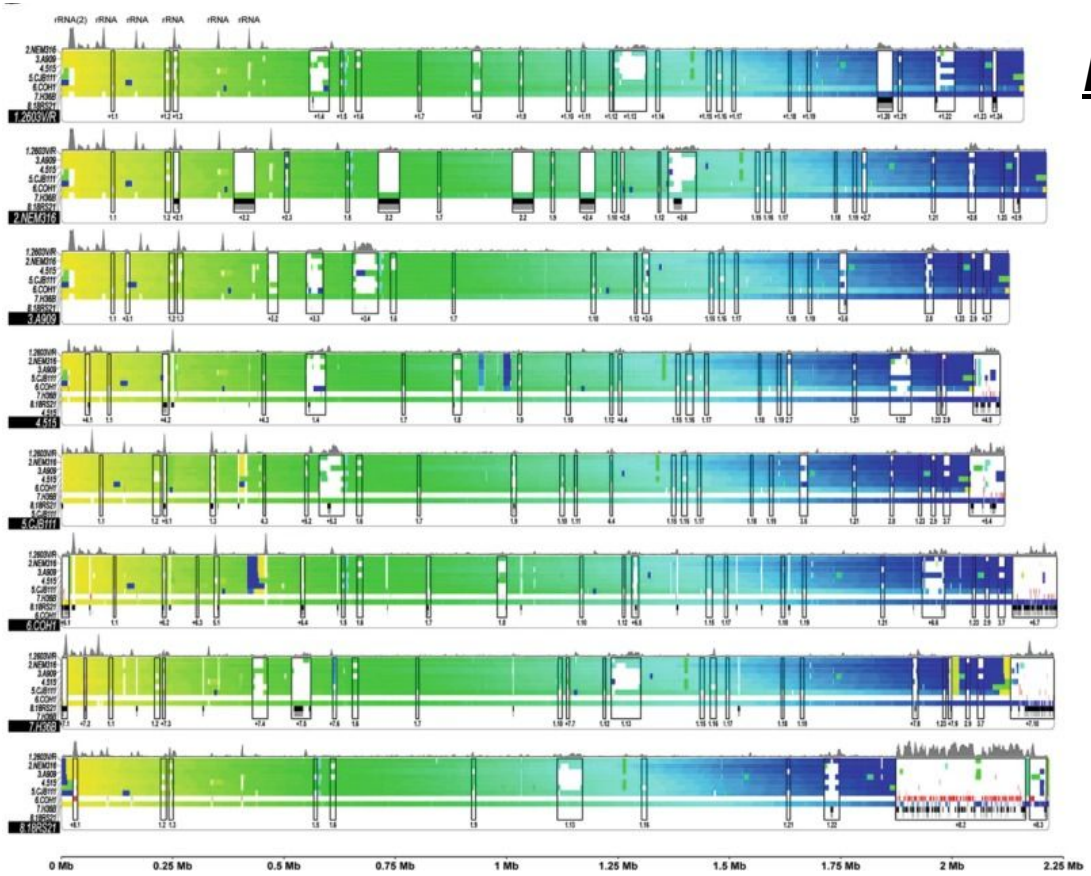


Fig. 5. A variant graph.

variation graphs: not a new idea!

nine versions Valerio Magrelli's poem "Campagna Romana" (1981)



Group B Streptococcus assemblies from 2002

Pangenome: not a new concept

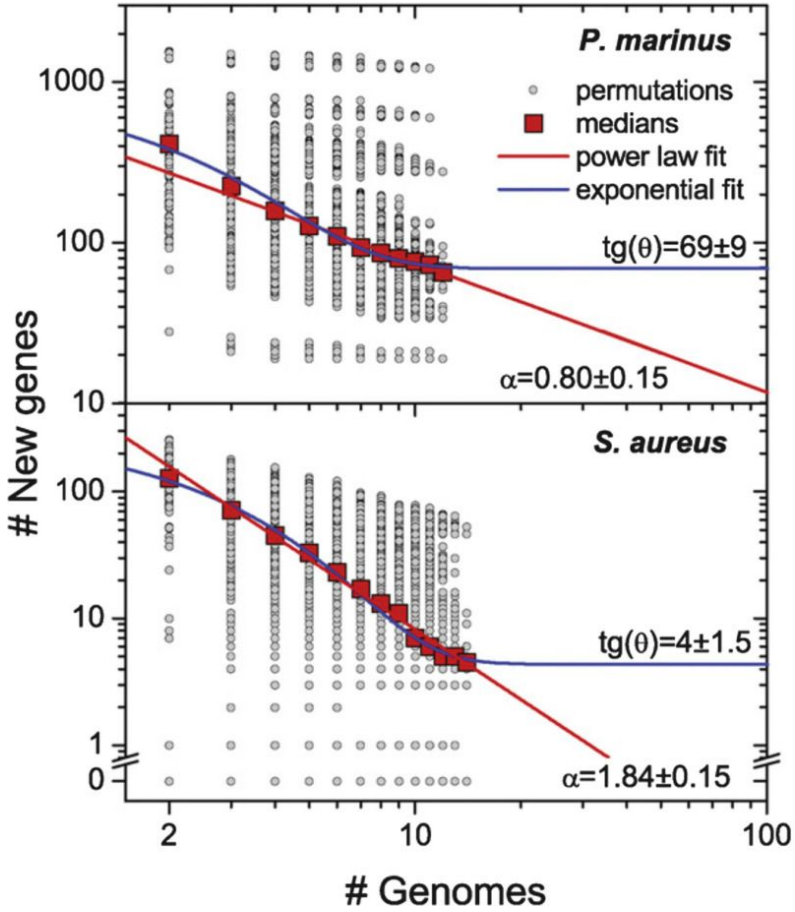
First collections of multiple genomes from the same species demonstrated substantial differences.

This was unexpected and required new theory to understand.

A single reference is not enough to explain genomic diversity. Even many genomes may not be enough.

Some genes are shared among all individuals: these are “core”, while others are not—we call them “accessory”.

https://doi.org/10.1007/978-3-030-38281-0



The exponent α determines whether the pangenome is open (α ≤ 1) or closed (α > 1). The top panel shows data for an open pangenome species, P. marinus; the bottom panel for a closed pangenome species, S. aureus

Lessons from language modeling: Heaps' law

A pangenome is:

Closed: our observations of new genes with new genomes diminish.

Open: we continue to see new genes as we add more genomes.

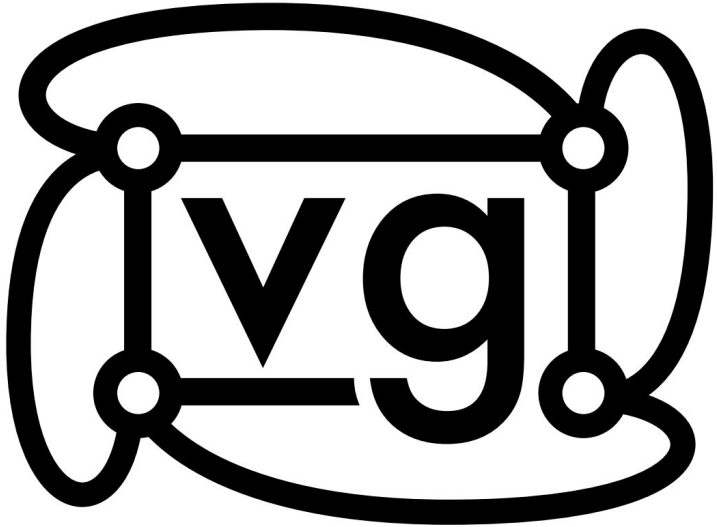
https://doi.org/10.1007/978-3-030-38281-0

Pangenome research timeline

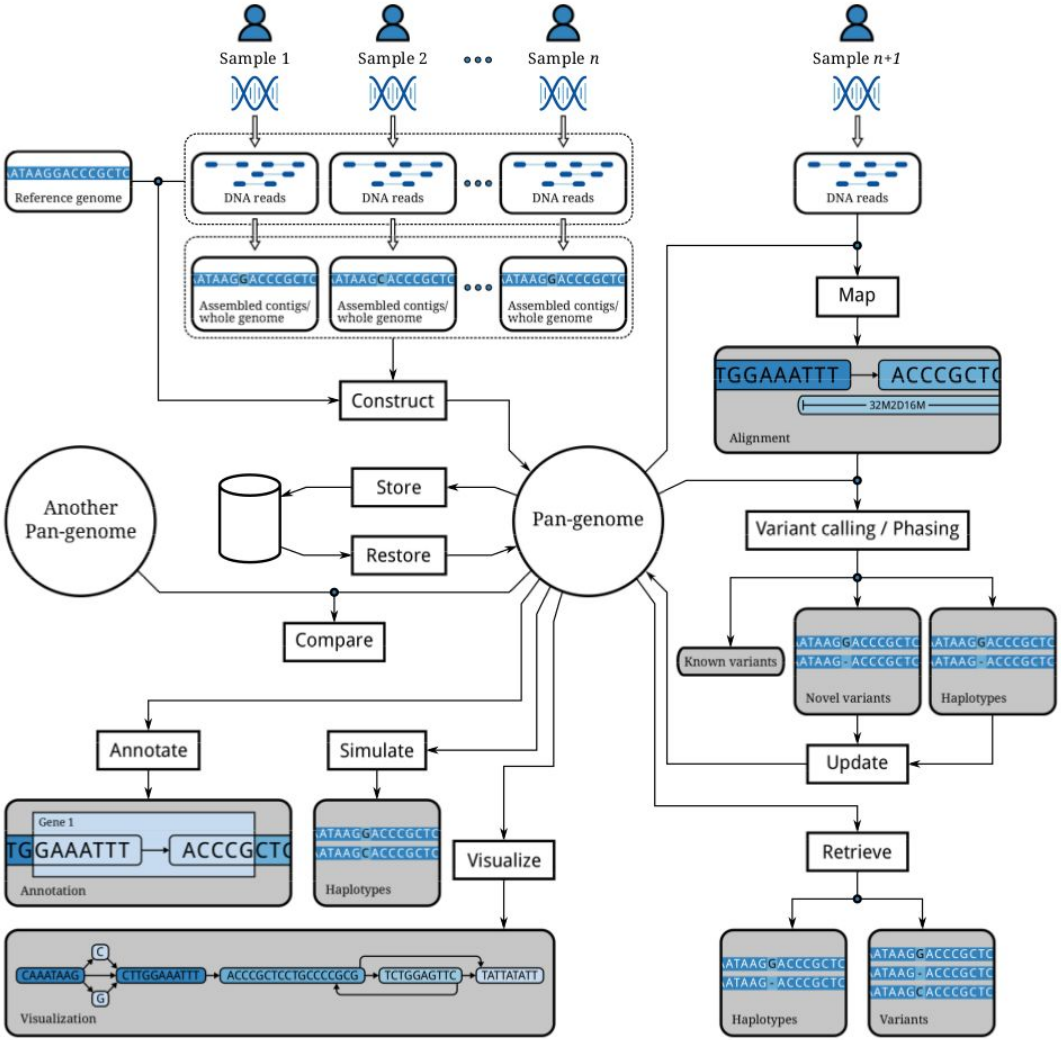
2000-2010s: counting genes

~2015: let’s take it to the sequence level (genome graphs)

2020s: complete assemblies (T2T pangenomes)



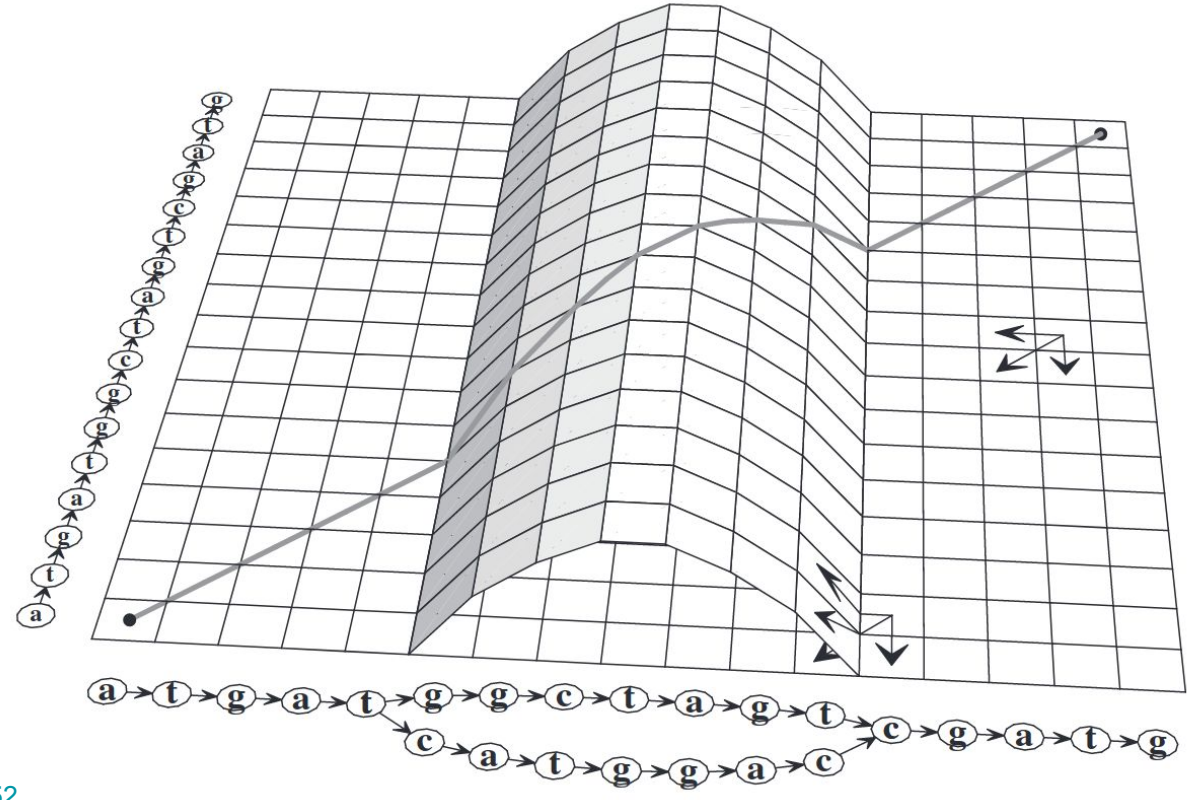
github.com/vgteam/vg



"Computational pan-genomics: status, promises and challenges" https://doi.org/10.1093%2Fbib%2Fbbw089

Wait! You can align sequences to graphs?

yup… we can generalize most standard bioinformatic algorithms to graphs, as in Partial Order Alignment →



https://doi.org/10.1093/bioinformatics/18.3.452

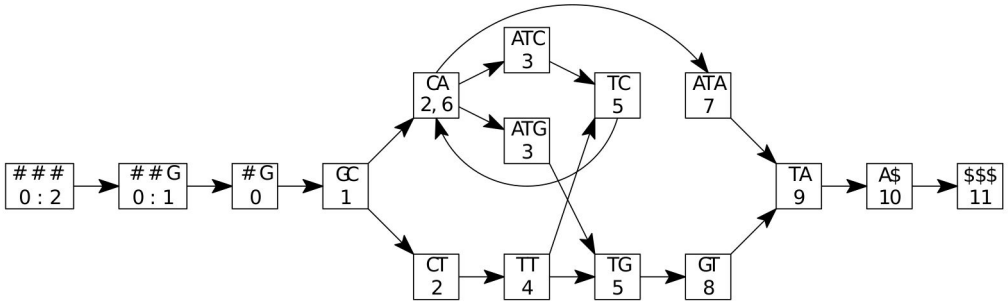
And the FM-index?

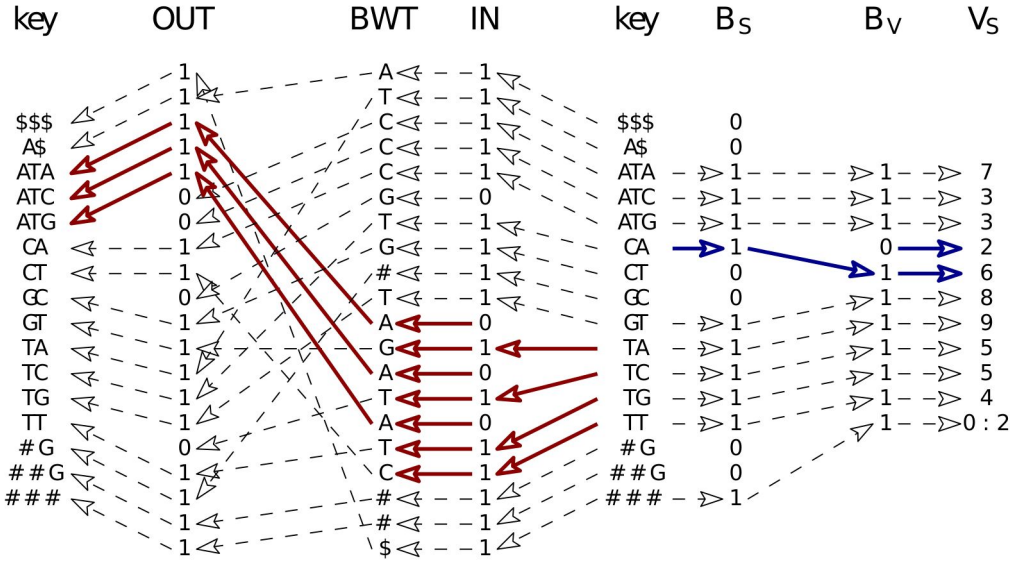
Jouni Sirén generalized the FM-index to work on a transformation of the variation graph (technically a de Bruijn graph with k=256).

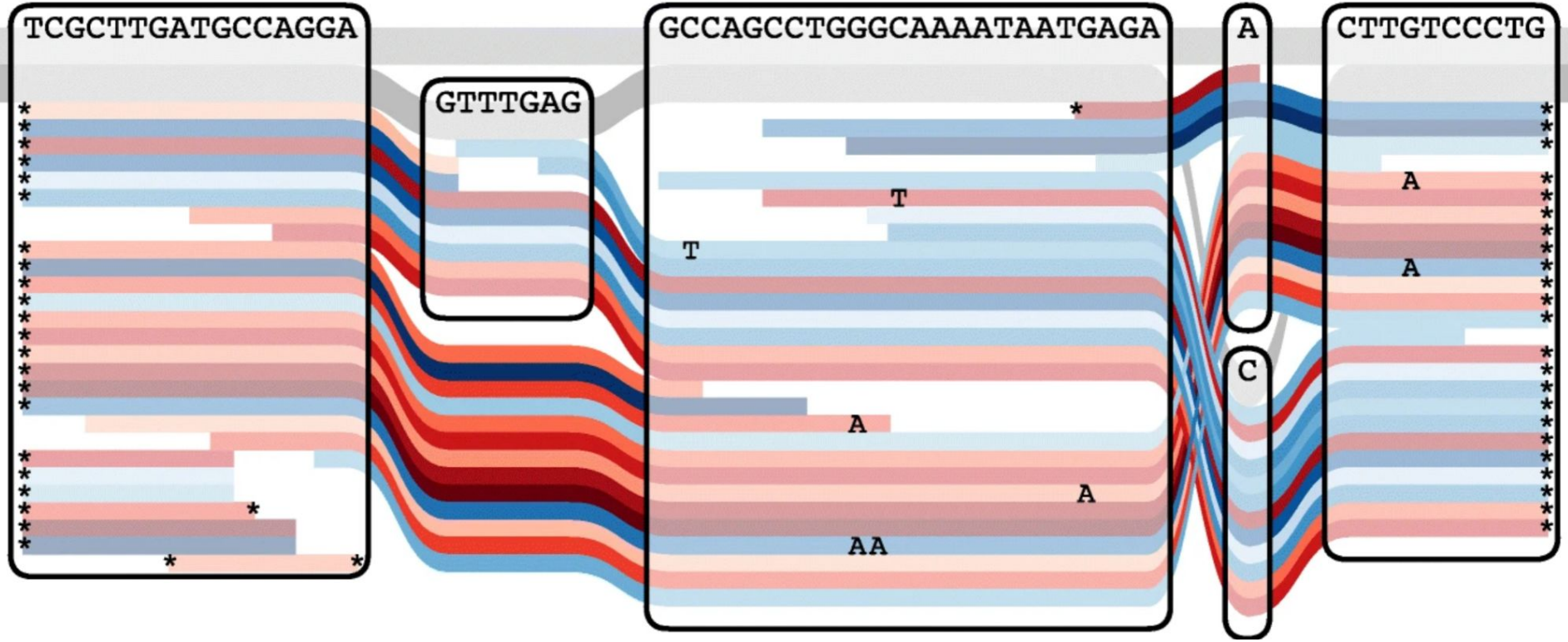
GCSA2 →

We use it to find MEMs just as in bwa mem.

This seeds alignment to the graph.



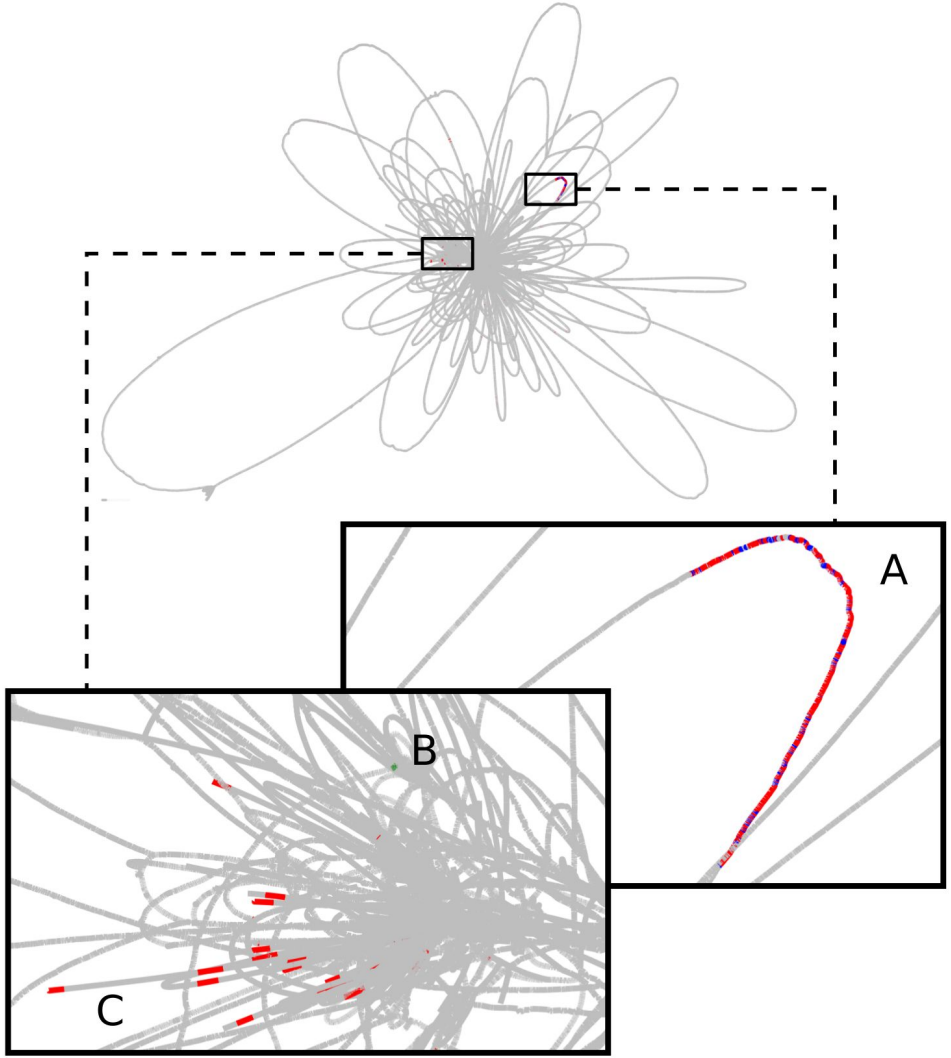




reads aligned to a variation graph

and long reads too!

a pacbio read vs. a yeast graph:



nature biotechnology

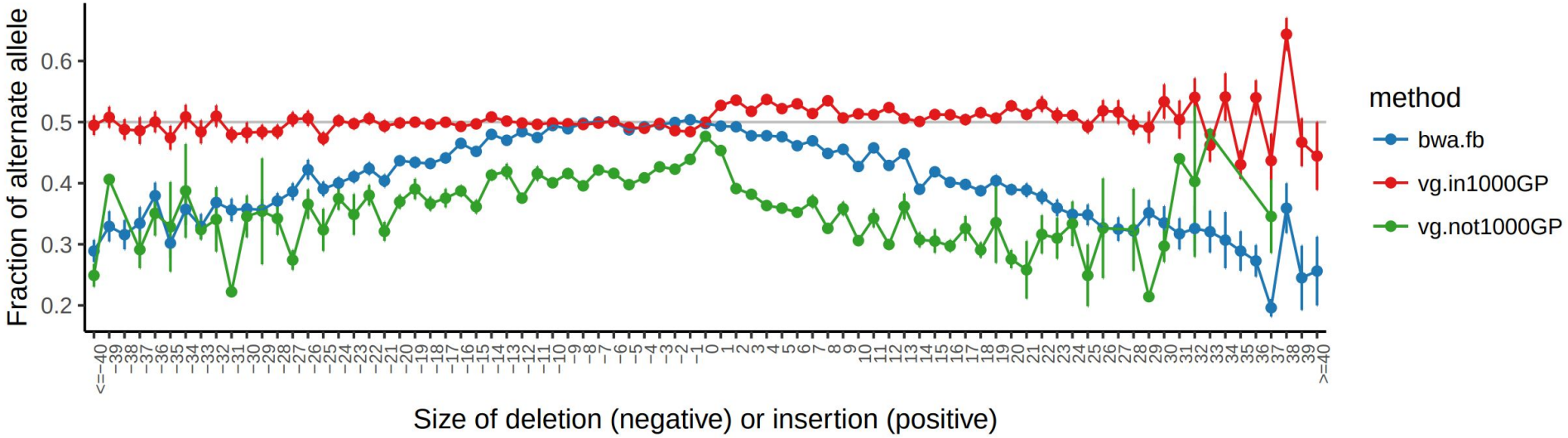
Letter | Published: 20 August 2018

Variation graph toolkit improves read mapping by representing genetic variation in the reference

Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten & Richard Durbin

Nature Biotechnology **36**, 875–879 (2018) | Download Citation

vg resolves reference bias at known indels in *HG002*



50x 2x150bp Illumina sequencing of HG002

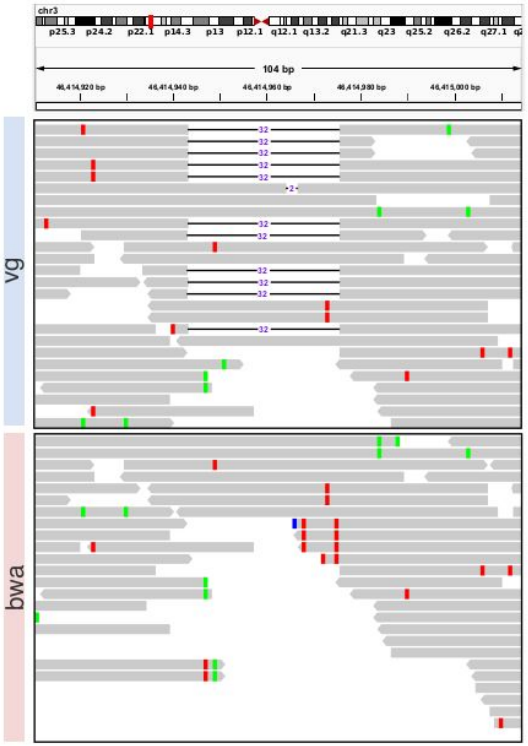
Genome Biology

Research | Open Access | Published: 17 September 2020

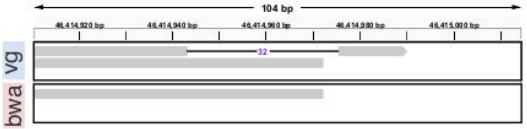
Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph

Rui Martiniano, Erik Garrison, Eppie R. Jones, Andrea Manica & Richard Durbin

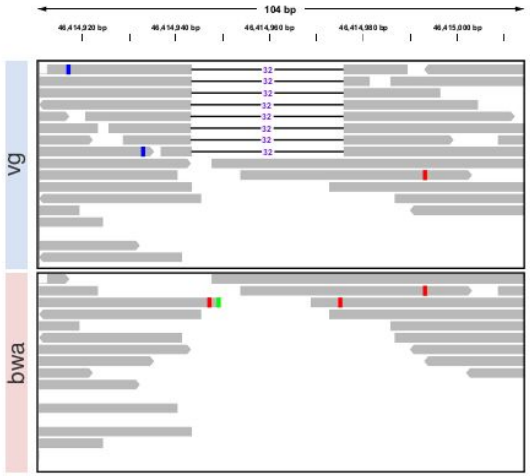
Yamnaya (Early Bronze Age Kazakhstan)



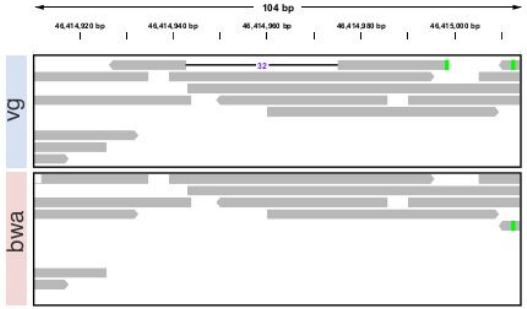
6DT3 (Roman Britain)



12880A (Iron Age Britain)



15577A (Anglo-Saxon Britain)



Using variation graphs to observe CCR5-delta in ancient samples

Rui Martiniano

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

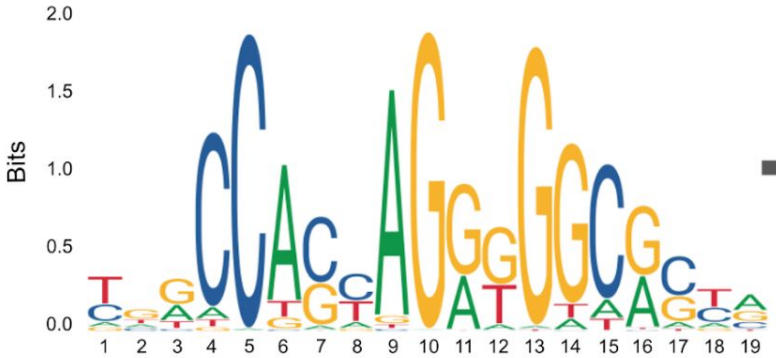
RESEARCH ARTICLE

GRAFIMO: Variant and haplotype aware motif scanning on pangenome graphs

Manuel Tognon, Vincenzo Bonnici, Erik Garrison, Rosalba Giugno, Luca Pinello

A

CTCF Motif (MA139.1)

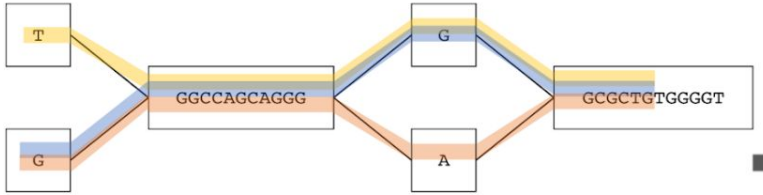


Processed motif PWM

S1	G	G	G	C	C	A	G	C	A	G	G	G	G	G	C	G	C	T	G
S2	T	G	G	C	C	A	G	C	A	G	G	G	G	G	C	G	C	T	G
S3	G	G	G	C	C	A	G	C	A	G	G	G	A	G	C	G	C	T	G
A	754	812	858	716	545	943	686	773	955	504	873	713	**581**	716	771	883	750	781	890
C	893	832	736	982	993	620	945	928	605	51	491	613	51	577	975	620	938	902	852
G	**775**	924	931	662	51	761	905	734	699	993	952	941	**992**	979	536	942	898	772	886
T	**901**	822	792	677	424	758	573	870	616	459	555	875	545	737	732	611	680	888	721

B

Pangenome variation graph (VG)



Reference motif candidate: GGGCCAGCAGGGGGCGCTG

Haplotype motif candidate: TGGCCAGCAGGGGGCGCTG

Haplotype motif candidate: GGGCCAGCAGGGAGCGCTG

Retrieved motif occurrences and haplotype frequencies

Sequence	Log-odds score	*P*-value	*q*-value	Reference	Haplotype frequency
GGGCCAGCAGGGGGCGCTG	28.22	$7.51e^{-12}$	$3.86e^{-6}$	non ref.	32
TGGCCAGCAGGGGGCGCTG	26.16	$3.12e^{-10}$	$7.72e^{-6}$	ref.	5063
GGGCCAGCAGGGAGCGCTG	19.43	$1.71e^{-7}$	$1.73e^{-4}$	non ref.	1

Genome Biology

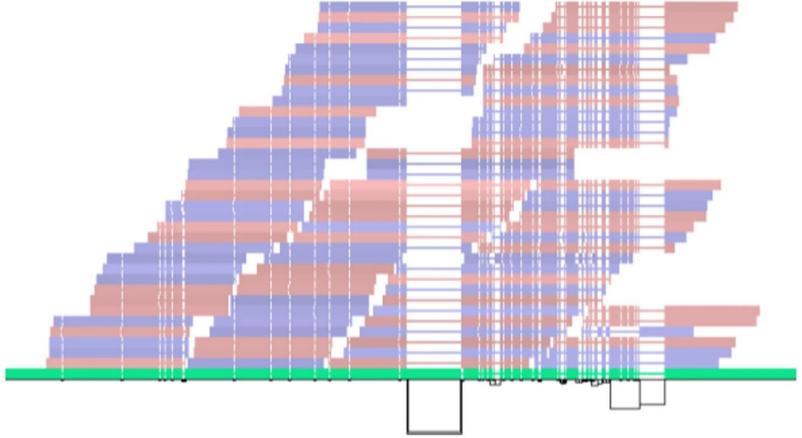
Method | Open Access | Published: 12 February 2020

Genotyping structural variants in pangenome graphs using the vg toolkit

Glenn Hickey, David Heller, Jean Monlong, Jonas A. Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T. Dawson, Erik Garrison, Adam M. Novak & Benedict Paten

Exonic deletion in the HGSVC dataset correctly genotyped by vg

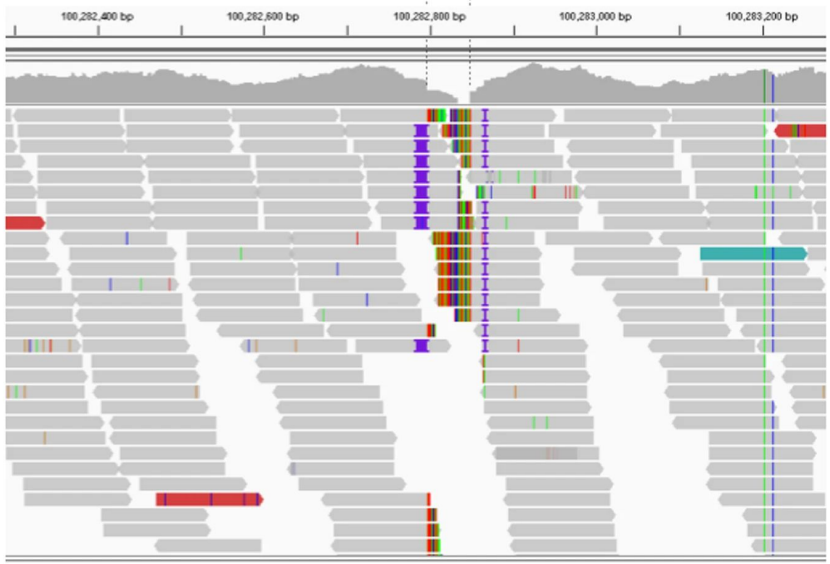
a)



reads vs. graph

deletion

b)



reads vs. GRCh38

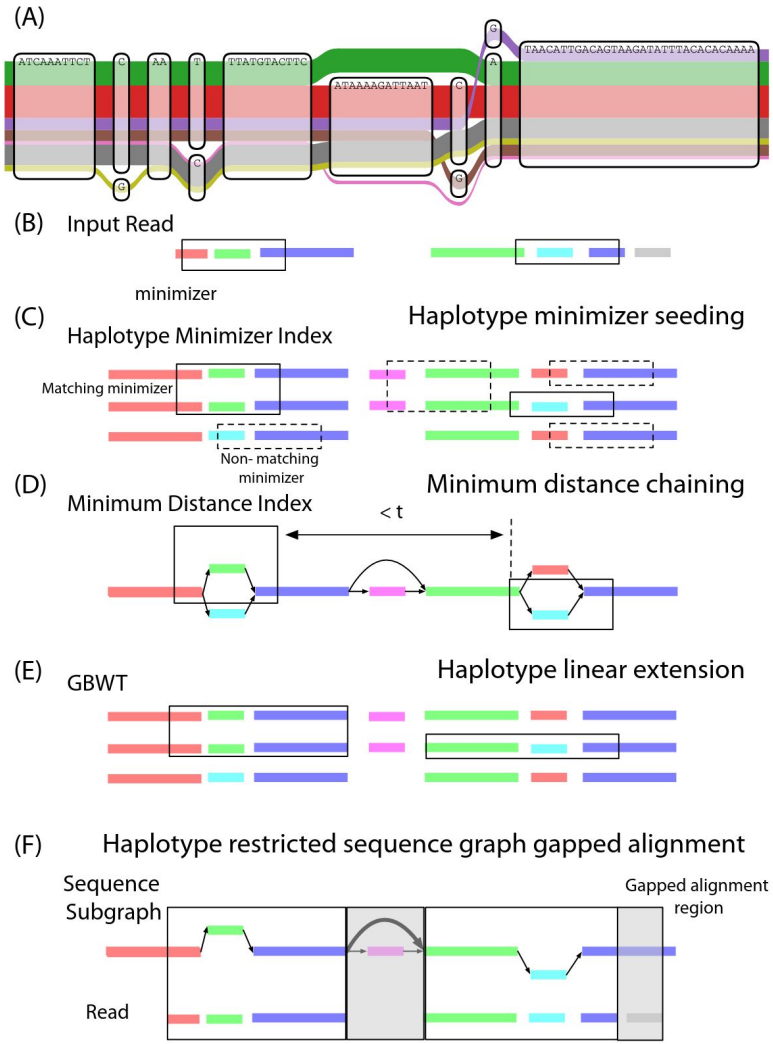
RESEARCH ARTICLE | GENOMICS

Pangenomics enables genotyping of known structural variants in 5202 diverse genomes

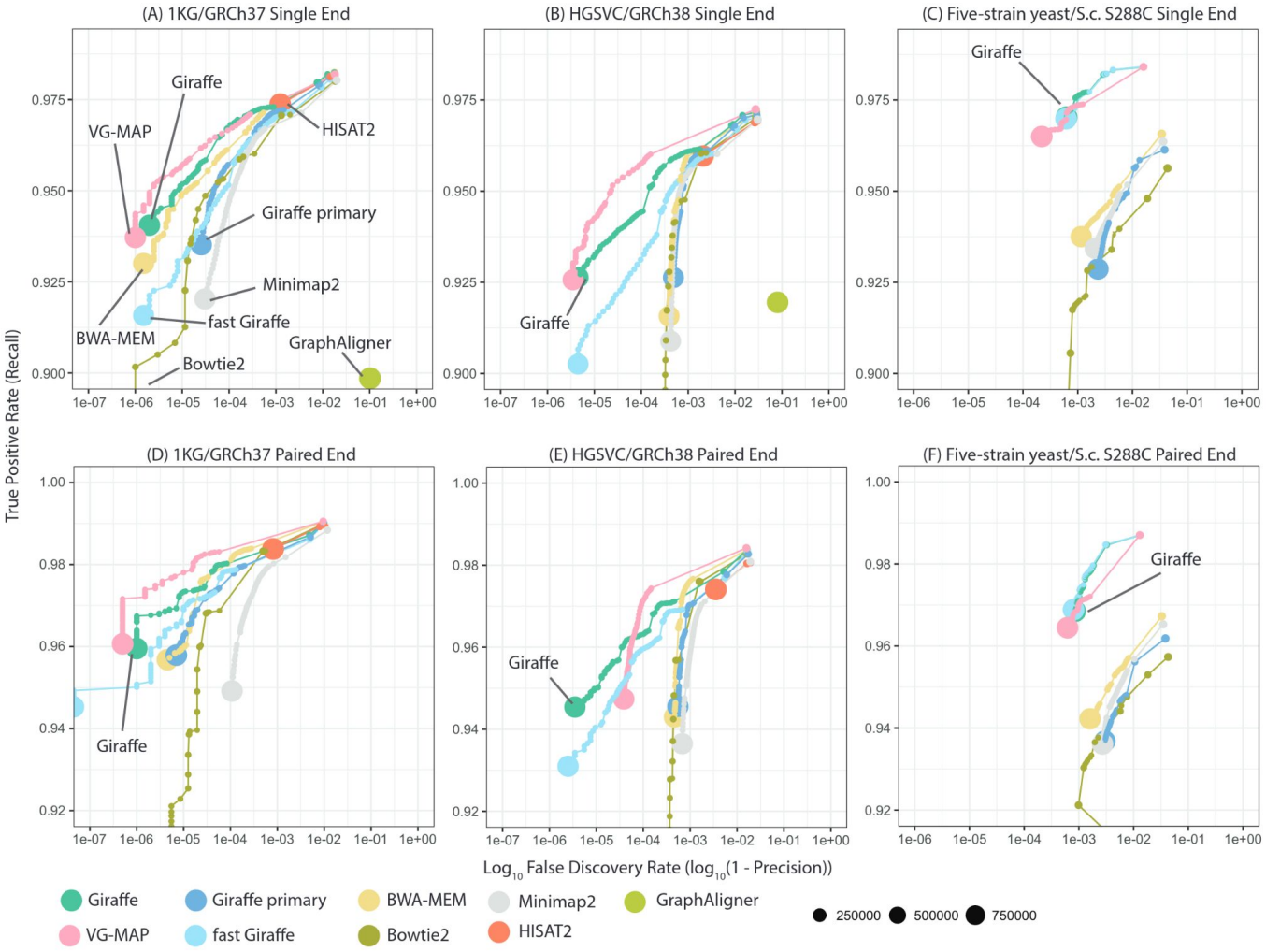
JOUNI SIRÉN, JEAN MONLONG, XIAN CHANG, ADAM M. NOVAK, JORDAN M. EIZENGA, CHARLES MARKELLO, JONAS A. SIBBESEN, GLENN HICKEY, PI-CHUAN CHANG, ANDREW CARROLL, NAMRATA GUPTA, STACEY GABRIEL, THOMAS W. BLACKWELL, AAKROSH RATAN, KENT D. TAYLOR, STEPHEN S. RICH, JEROME I. ROTTER, DAVID HAUSSLER, ERIK GARRISON, AND BENEDICT PATEN fewer Authors Info & Affiliations

SCIENCE • 17 Dec 2021 • Vol 374, Issue 6574 • DOI: 10.1126/science.abg8871

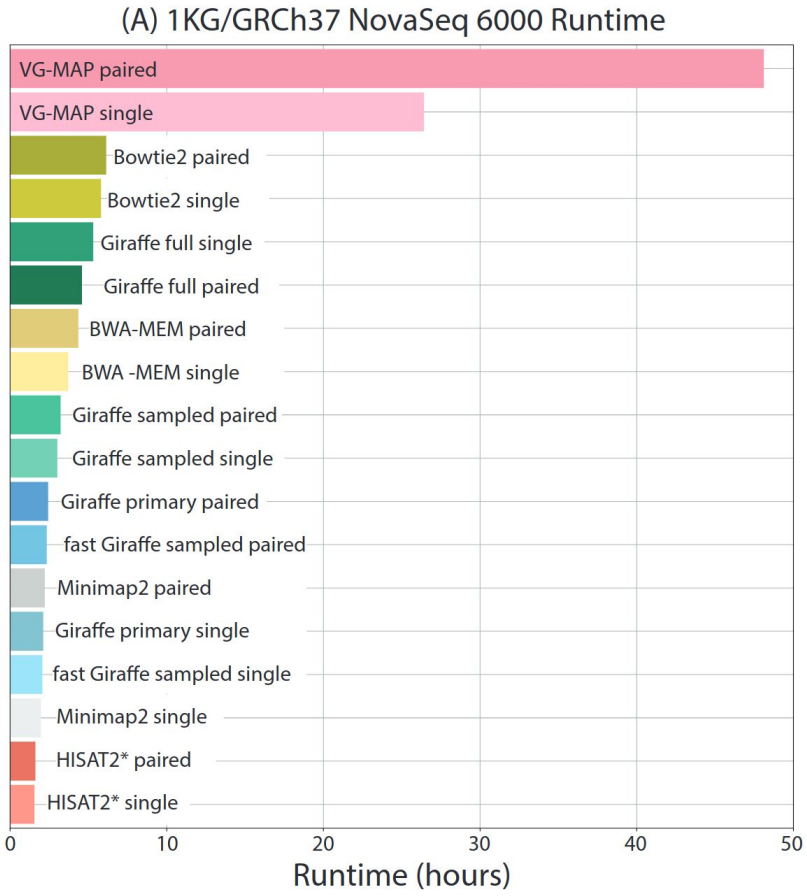
vg giraffe: approach

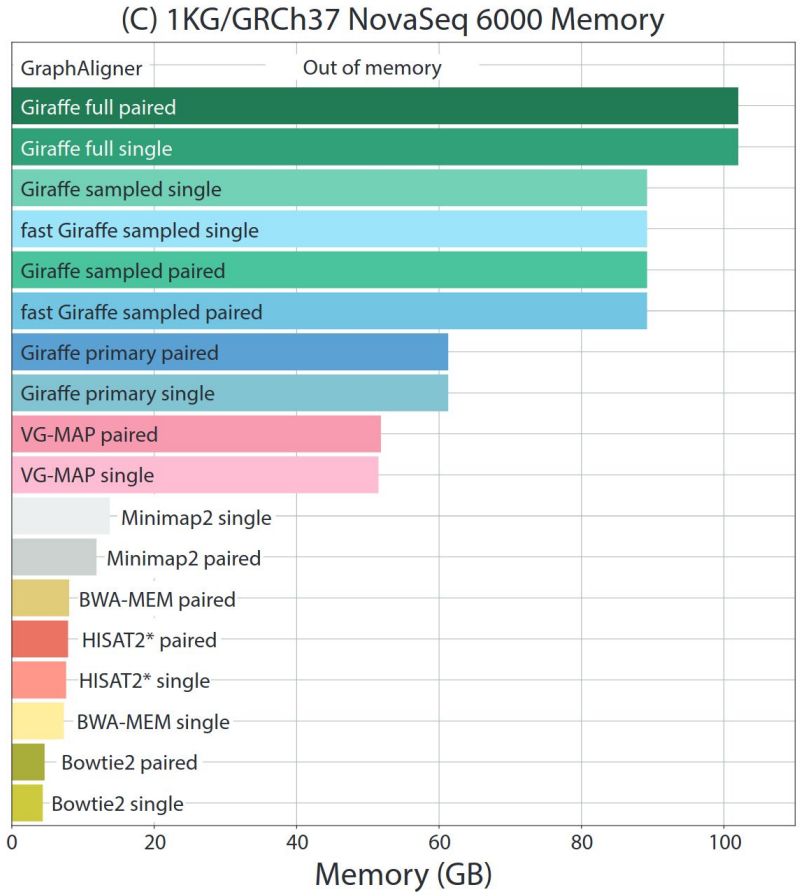


vg giraffe is accurate enough

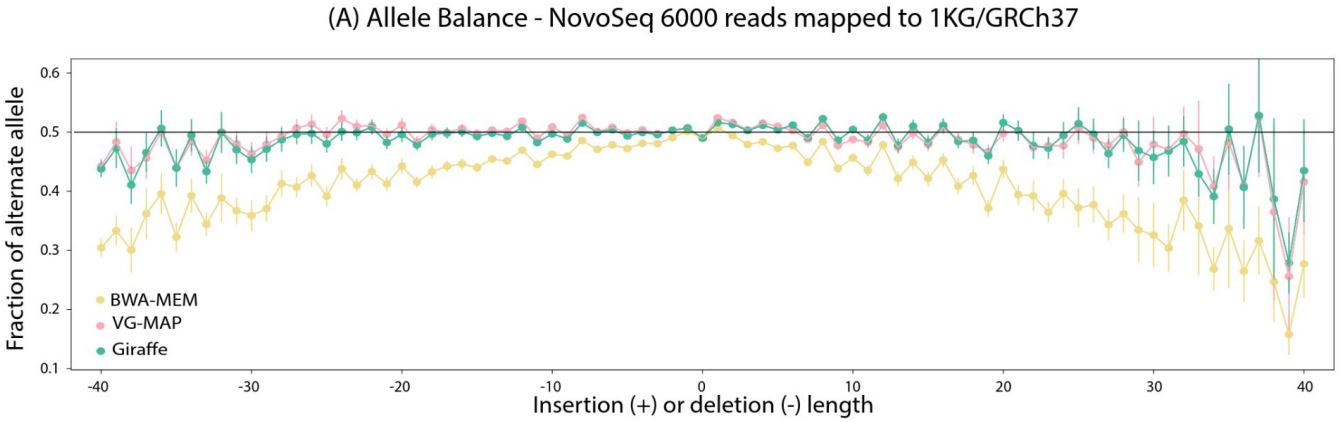


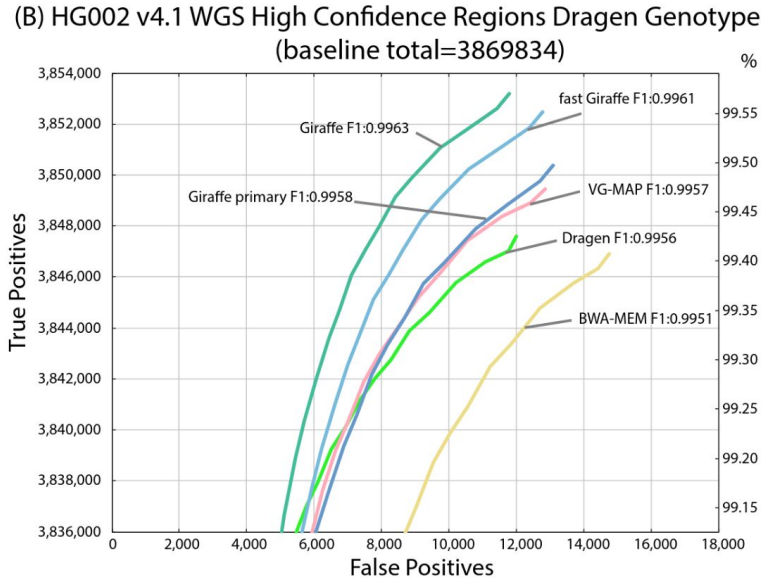
vg giraffe is *very fast*

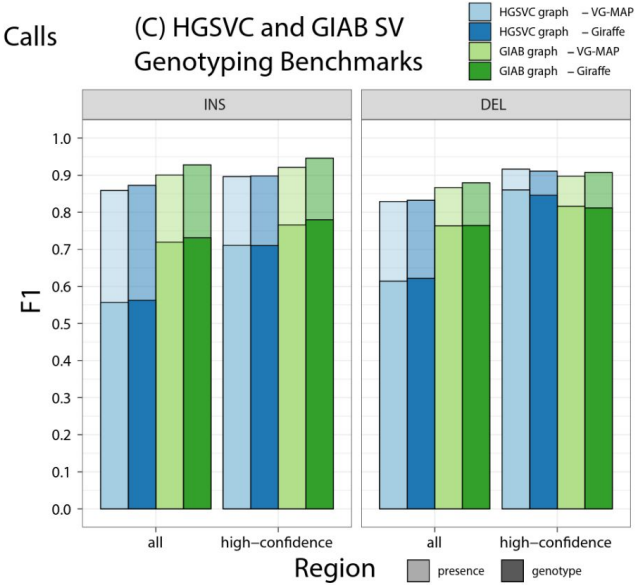




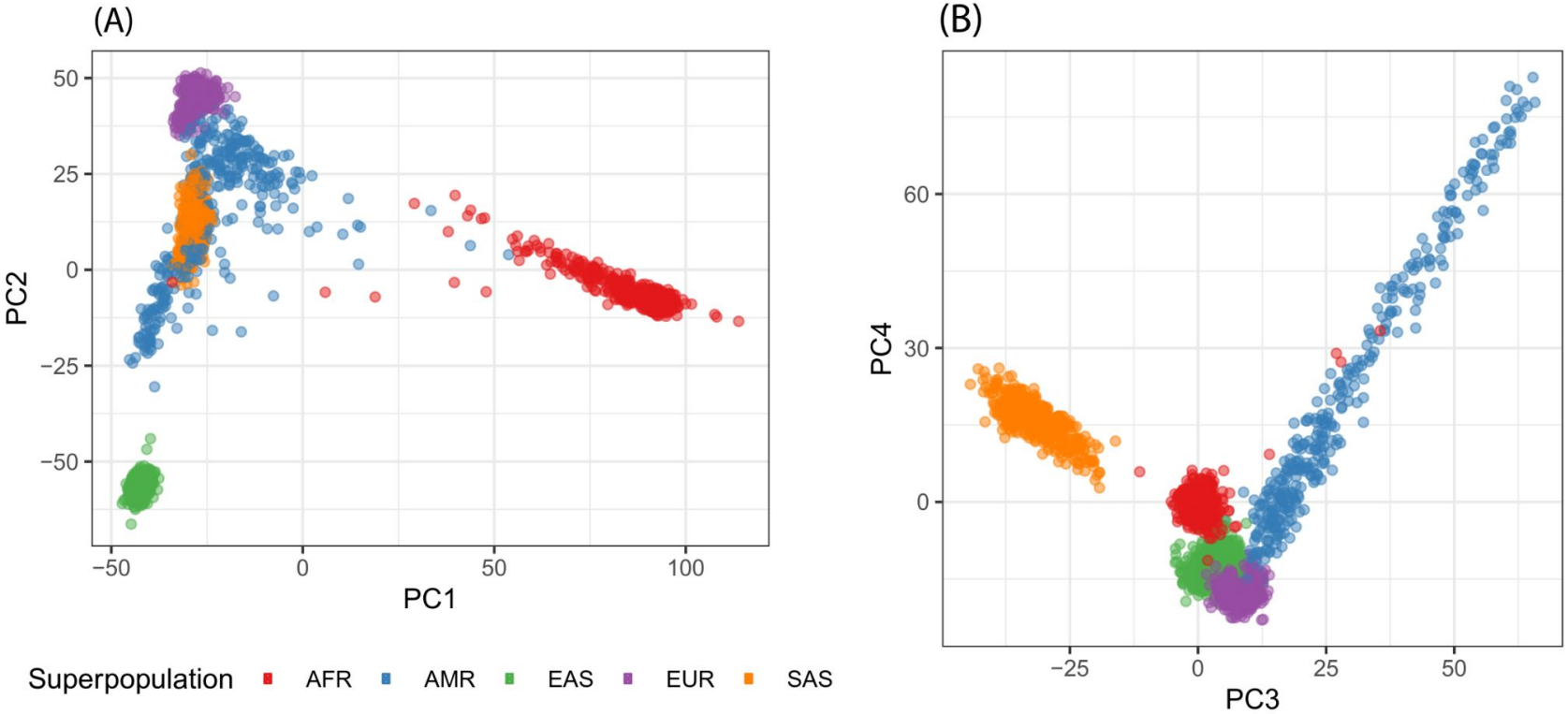
vg giraffe improves variant calling







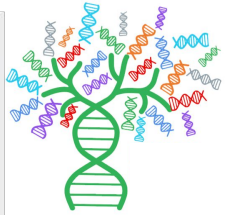
vg giraffe lets us scale: PCA from SVs in 5k genomes



The human pangenome project

Human Pangenome Reference Consortium

- Improve representation of **global genomic diversity** (>350 diverse diploid references)
- **Prioritizing quality**: we aim to release a complete (T2T) and comprehensive map of genome variation
- **Develop a new, non-linear reference data structure** and foster an innovative ecosystem of pangenomic tools
- Outreach, Education and Implementation
- **First draft is available**!



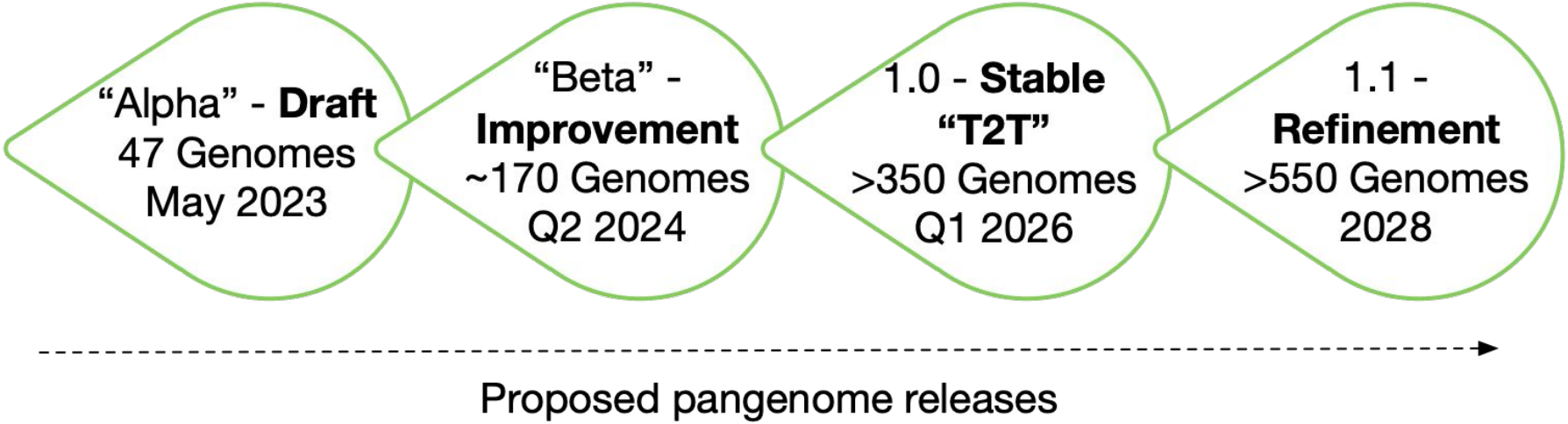
The Human Pangenome

Composed of three As:

- **Assemblies**
  - Haplotype resolved, soon T2T, but also 37, 38, T2T-CHM13.
- **Alignment**
  - Provides canonical homology information
- **Annotations**
  - Genes, etc. Should be consistent with alignment


Population sampling and representation
Technology Production
Phased/Finished T2T Assemblies
Pangenome and new workflows/tooling
International Pangenome Project
Raw reads
Global Alliance for Genomics & Health
Embedded Ethics and Policy:
Inter-disciplinary ethics working group/oversight committee

Human Pangenome Timeline



Building a draft human pangenome

Article

A draft human pangenome reference

https://doi.org/10.1038/s41586-023-05896-x

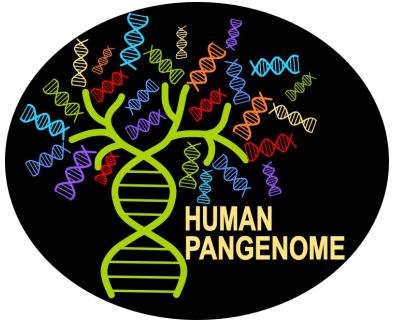
Received: 9 July 2022

Accepted: 28 February 2023

Published online: 10 May 2023

Open access

Check for updates

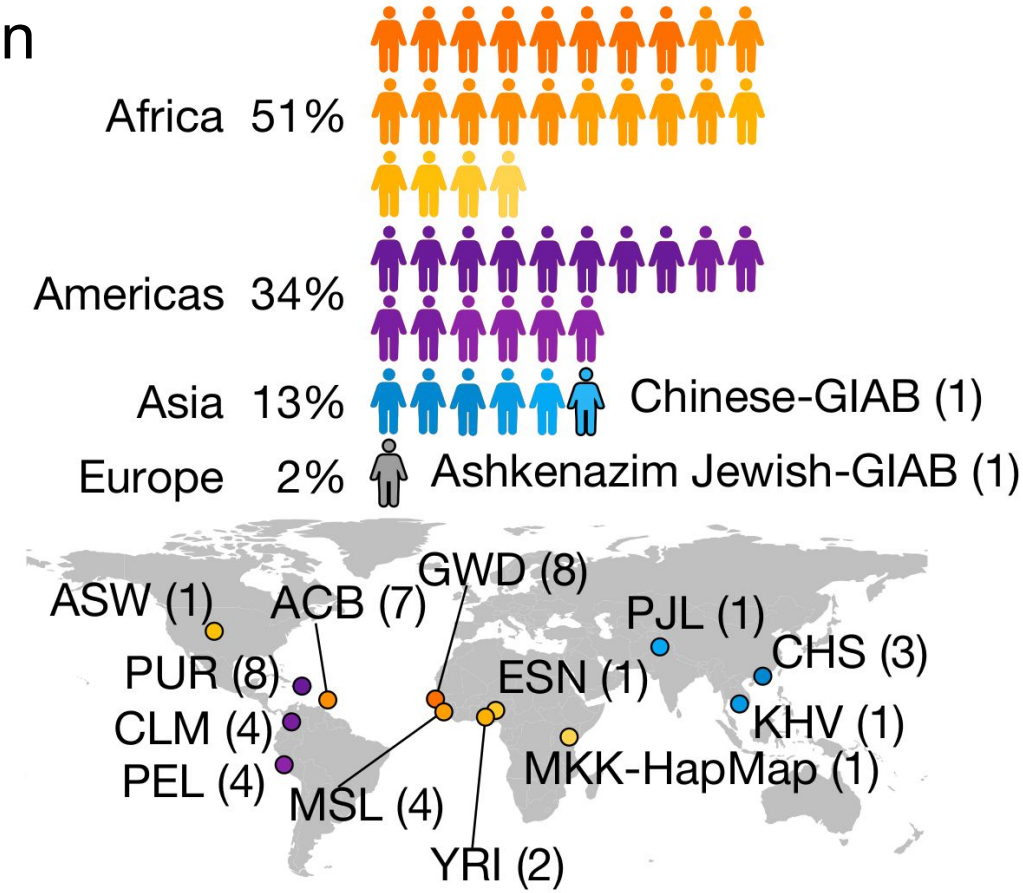


Wen-Wei Liao[1,2,3,60], Mobin Asri[4,60], Jana Ebler[5,6,60], Daniel Doerr[5,6], Marina Haukness[4], Glenn Hickey[4], Shuangjia Lu[1,2], Julian K. Lucas[4], Jean Monlong[4], Haley J. Abel[7], Silvia Buonaiuto[8], Xian H. Chang[4], Haoyu Cheng[9,10], Justin Chu[9], Vincenza Colonna[8,11], Jordan M. Eizenga[4], Xiaowen Feng[9,10], Christian Fischer[11], Robert S. Fulton[12,13], Shilpa Garg[14], Cristian Groza[15], Andrea Guarracino[11,16], William T. Harvey[17], Simon Heumos[18,19], Kerstin Howe[20], Miten Jain[21], Tsung-Yu Lu[22], Charles Markello[4], Fergal J. Martin[23], Matthew W. Mitchell[24], Katherine M. Munson[17], Moses Njagi Mwaniki[25], Adam M. Novak[4], Hugh E. Olsen[4], Trevor Pesout[4], David Porubsky[17], Pjotr Prins[11], Jonas A. Sibbesen[26], Jouni Sirén[4], Chad Tomlinson[12], Flavia Villani[11], Mitchell R. Vollger[17,27], Lucinda L. Antonacci-Fulton[12], Gunjan Baid[28], Carl A. Baker[17], Anastasiya Belyaeva[28], Konstantinos Billis[23], Andrew Carroll[28], Pi-Chuan Chang[28], Sarah Cody[12], Daniel E. Cook[28], Robert M. Cook-Deegan[29], Omar E. Cornejo[30], Mark Diekhans[4], Peter Ebert[5,6,31], Susan Fairley[23], Olivier Fedrigo[32], Adam L. Felsenfeld[33], Giulio Formenti[32], Adam Frankish[23], Yan Gao[34], Nanibaa' A. Garrison[35,36,37], Carlos Garcia Giron[23], Richard E. Green[38,39], Leanne Haggerty[23], Kendra Hoekzema[17], Thibaut Hourlier[23], Hanlee P. Ji[40], Eimear E. Kenny[41], Barbara A. Koenig[42], Alexey Kolesnikov[28], Jan O. Korbel[23,43], Jennifer Kordosky[17], Sergey Koren[44], HoJoon Lee[40], Alexandra P. Lewis[17], Hugo Magalhães[5,6], Santiago Marco-Sola[45,46], Pierre Marijon[5,6], Ann McCartney[44], Jennifer McDaniel[47], Jacquelyn Mountcastle[32], Maria Nattestad[28], Sergey Nurk[44], Nathan D. Olson[47], Alice B. Popejoy[48], Daniela Puiu[49], Mikko Rautiainen[44], Allison A. Regier[12], Arang Rhie[44], Samuel Sacco[30], Ashley D. Sanders[50], Valerie A. Schneider[51], Baergen I. Schultz[33], Kishwar Shafin[28], Michael W. Smith[33], Heidi J. Sofia[33], Ahmad N. Abou Tayoun[52,53], Françoise Thibaud-Nissen[51], Francesca Floriana Tricomi[23], Justin Wagner[47], Brian Walenz[44], Jonathan M. D. Wood[20], Aleksey V. Zimin[49,54], Guillaume Bourque[55,56,57], Mark J. P. Chaisson[22], Paul Flicek[23], Adam M. Phillippy[44], Justin M. Zook[47], Evan E. Eichler[17,58], David Haussler[4,58], Ting Wang[12,13], Erich D. Jarvis[32,58,59], Karen H. Miga[4], Erik Garrison[11]✉, Tobias Marschall[5,6]✉, Ira M. Hall[1,2]✉, Heng Li[9,10]✉ & Benedict Paten[4]✉

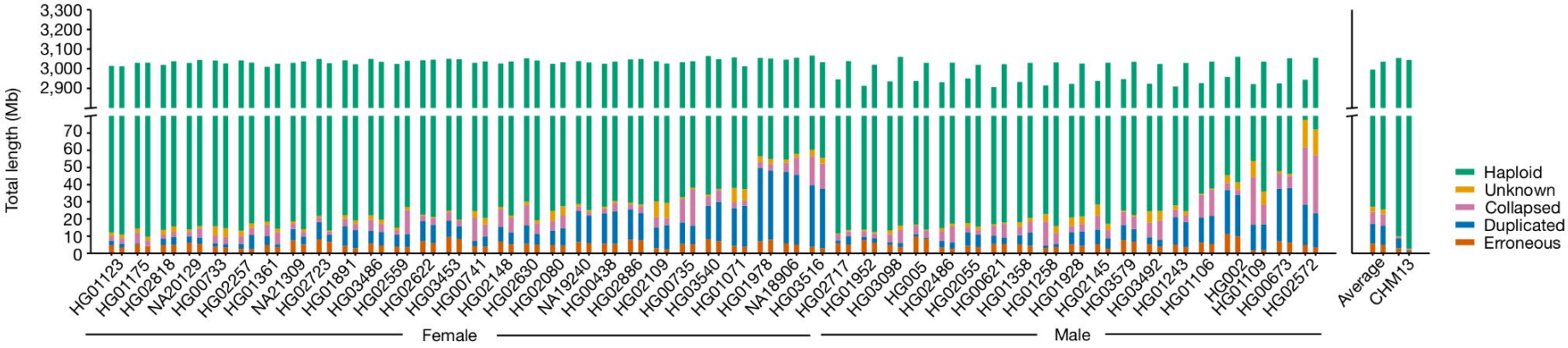
Draft pangenome composition

Sample selection was constrained by:

- trio status in Coriell biobank (-Europeans)
- low cell line passage count (--Europeans)
- genetic diversity (+++Africans)
- drift (+Asians, ++Americas)



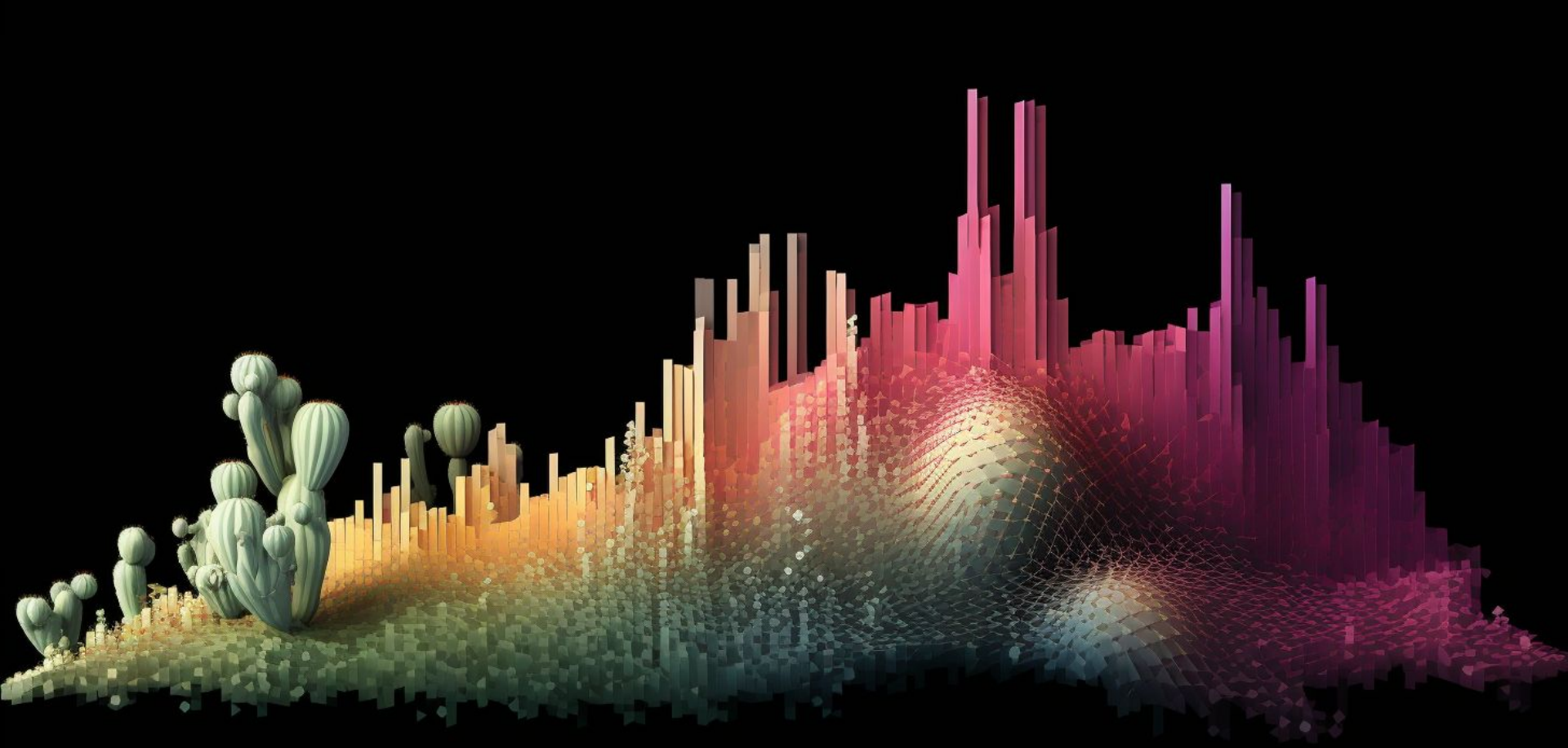
Amazing assemblies approach reference quality



Haplotype-resolved assemblies from trio-hifiasm.

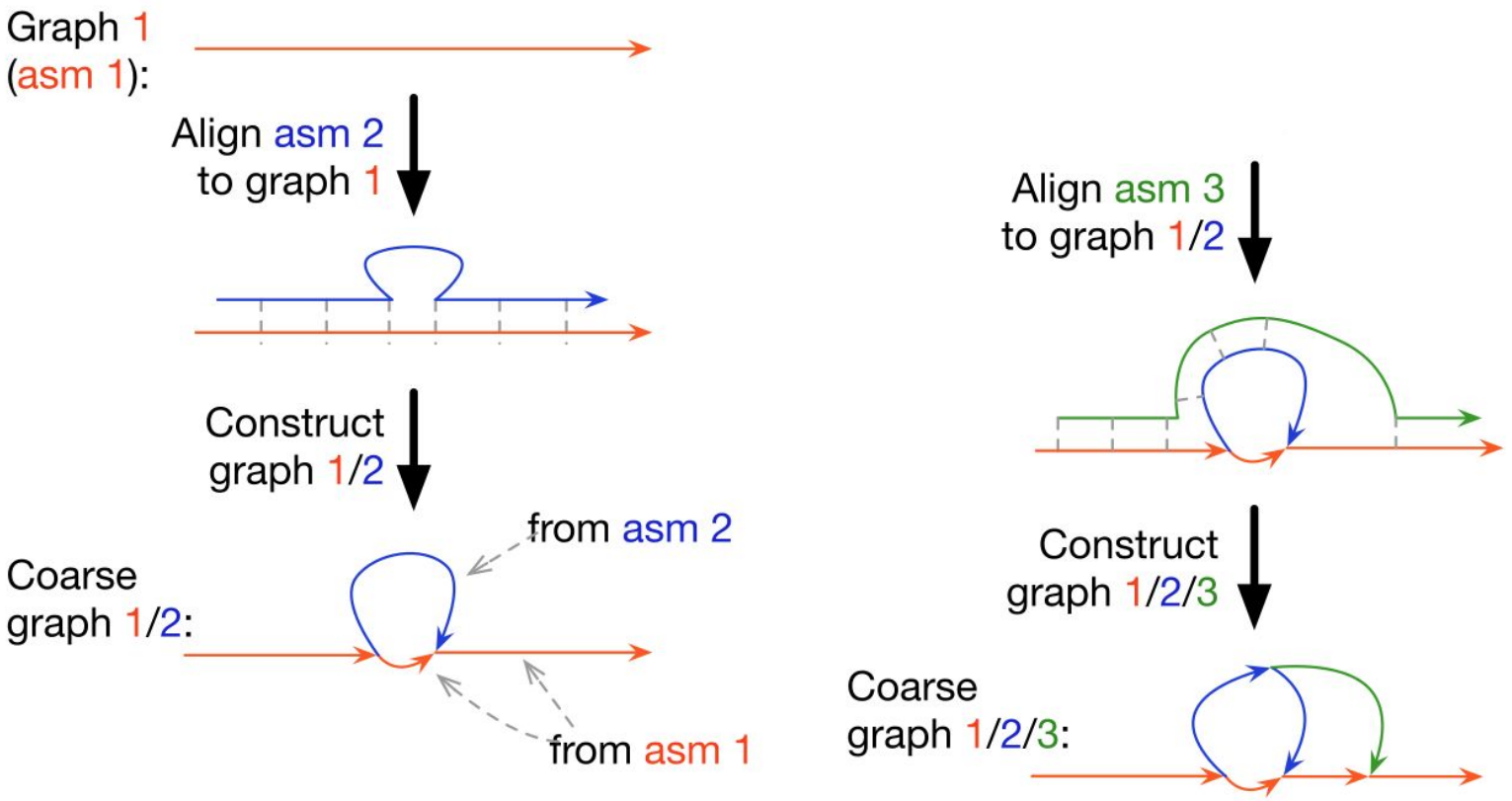
They are really good, according to realignment of reads to the assemblies and model of assembly completeness—nearly as good as T2T-CHM13!

Mobin Asri



Then we made 5 pangenome (reference) graphs…

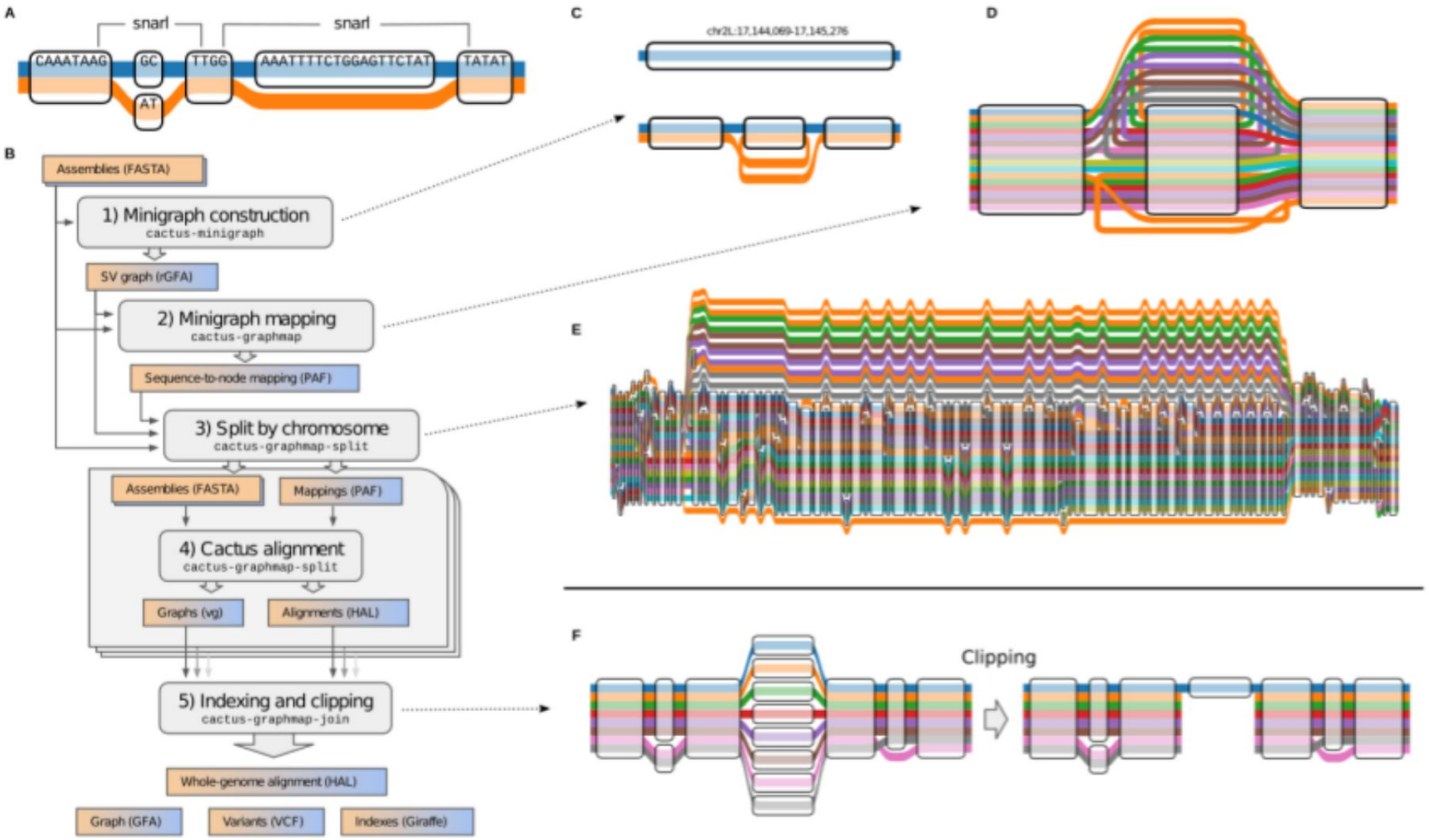
Minigraph



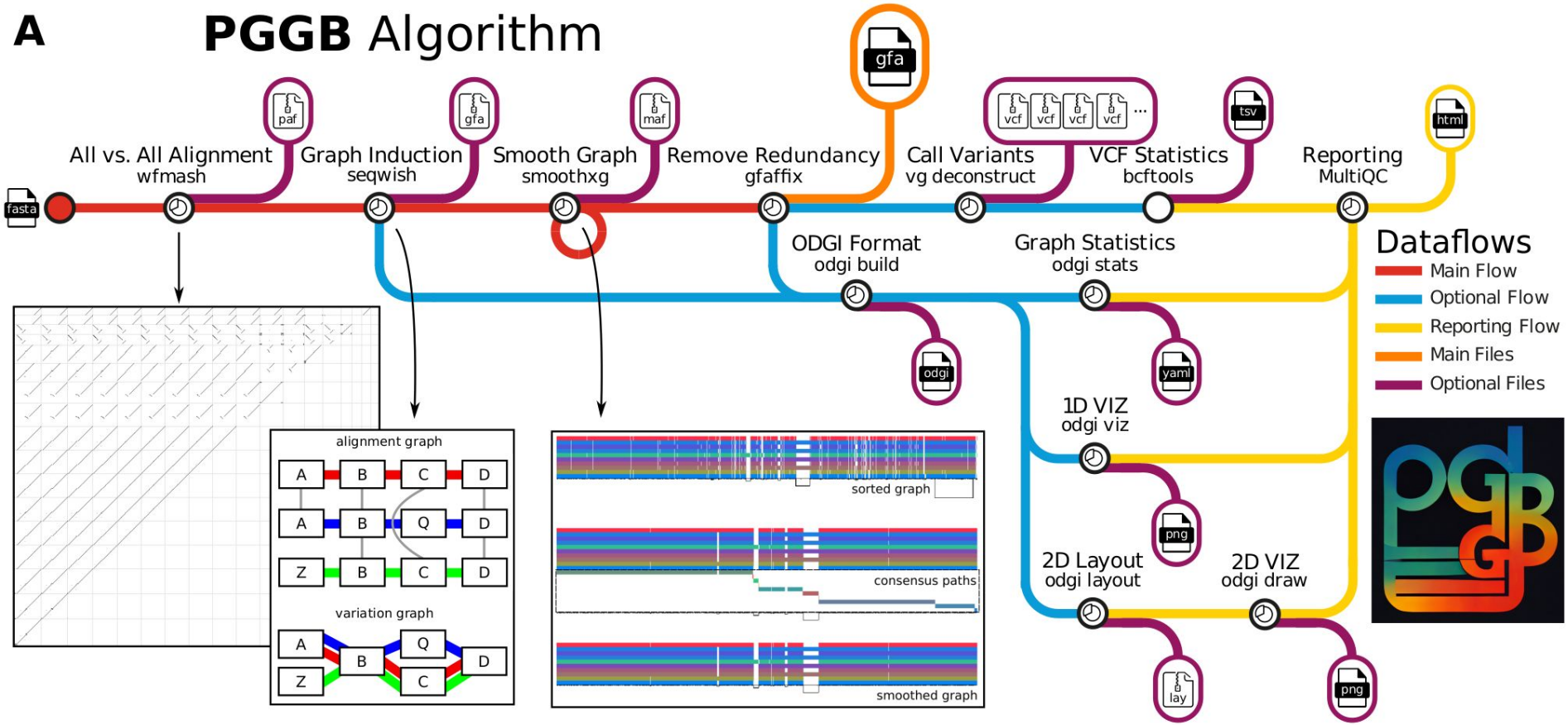
https://doi.org/10.1186/s13059-020-02168-z

Minigraph-Cactus

https://doi.org/10.1038/s41587-023-01793-w



https://doi.org/10.1101/2023.04.05.535718



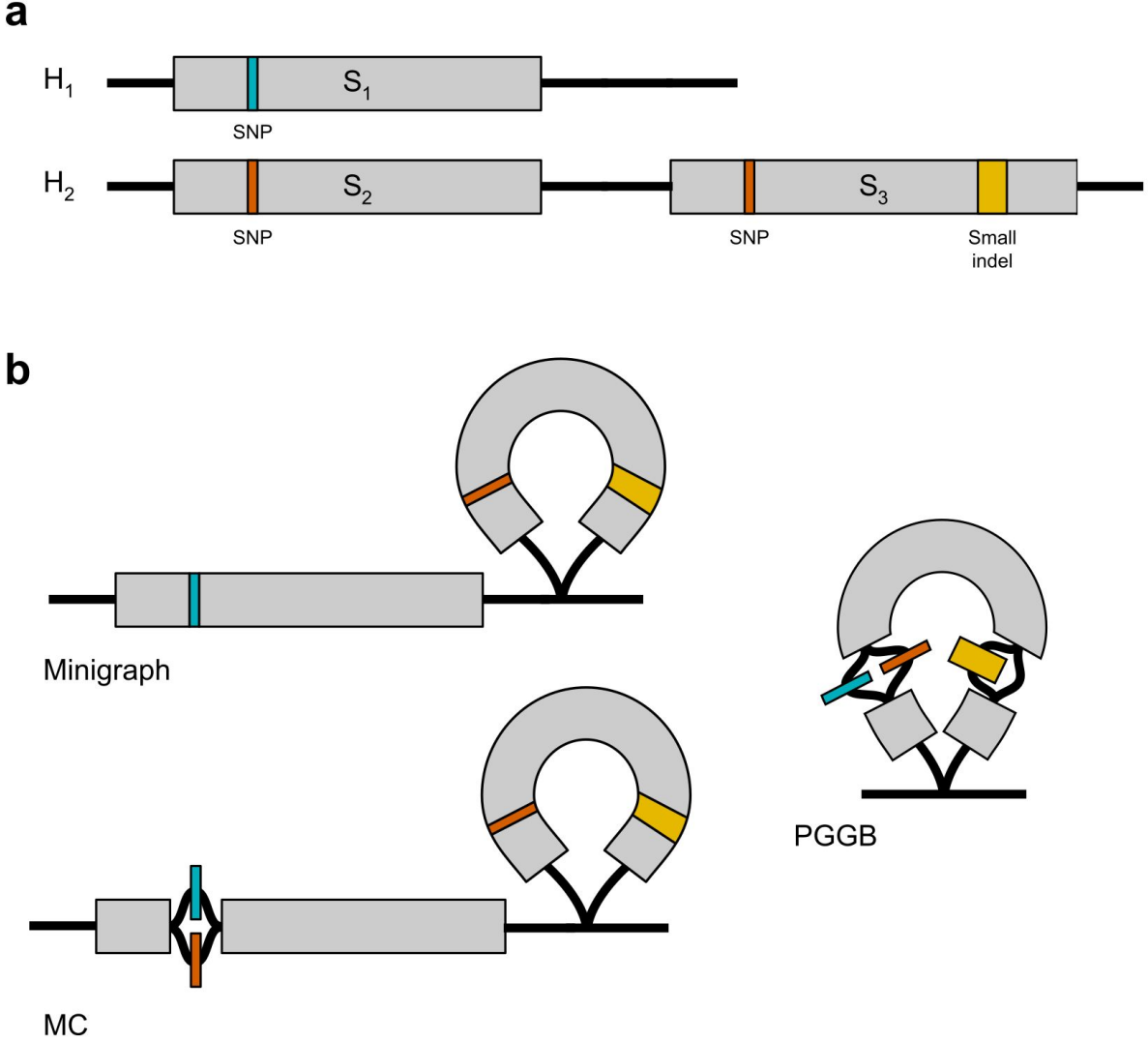
Key conceptual differences between HPRC pangenome construction methods

minigraph: just SVs, no complex stuff, one reference.

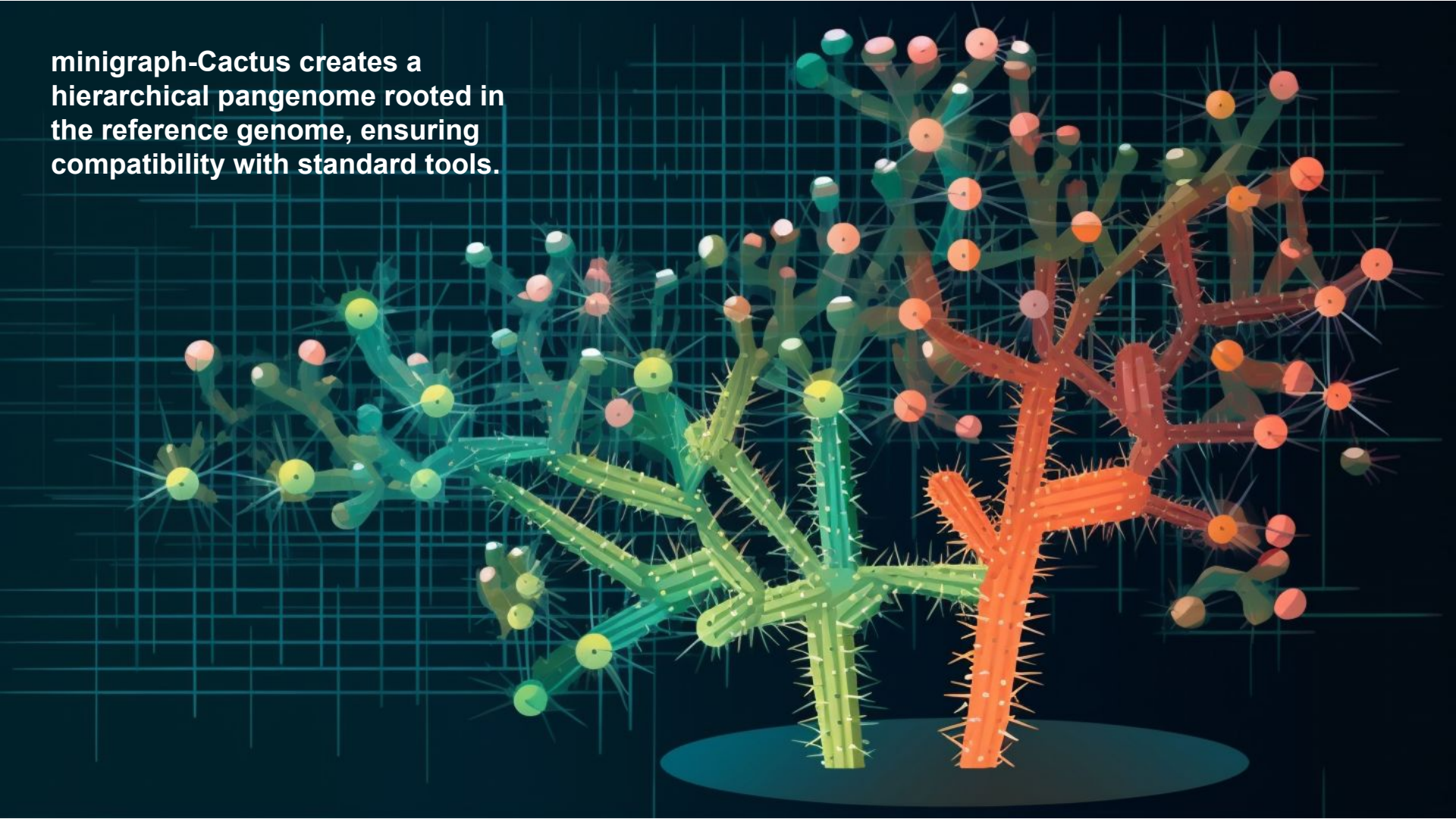
minigraph-cactus: add SNPs, clean up the breakpoints, useful for alignment, one reference.

pggb: everything-vs-everything, hard to align to, useful for studying evolution and pangenome structure at all scales, all genomes are references.

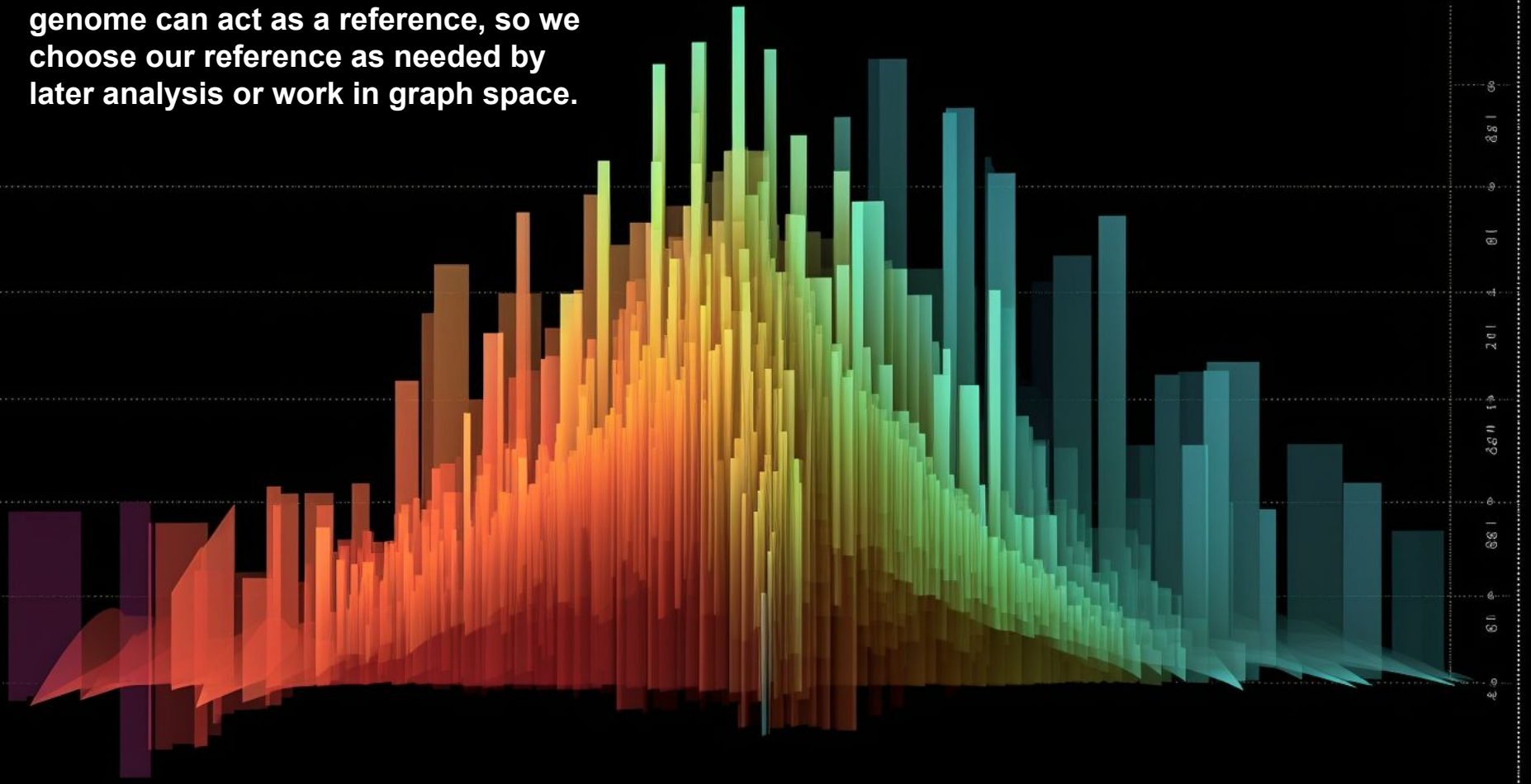
“Collapse” in high-copy repeats →



minigraph-Cactus creates a hierarchical pangenome rooted in the reference genome, ensuring compatibility with standard tools.

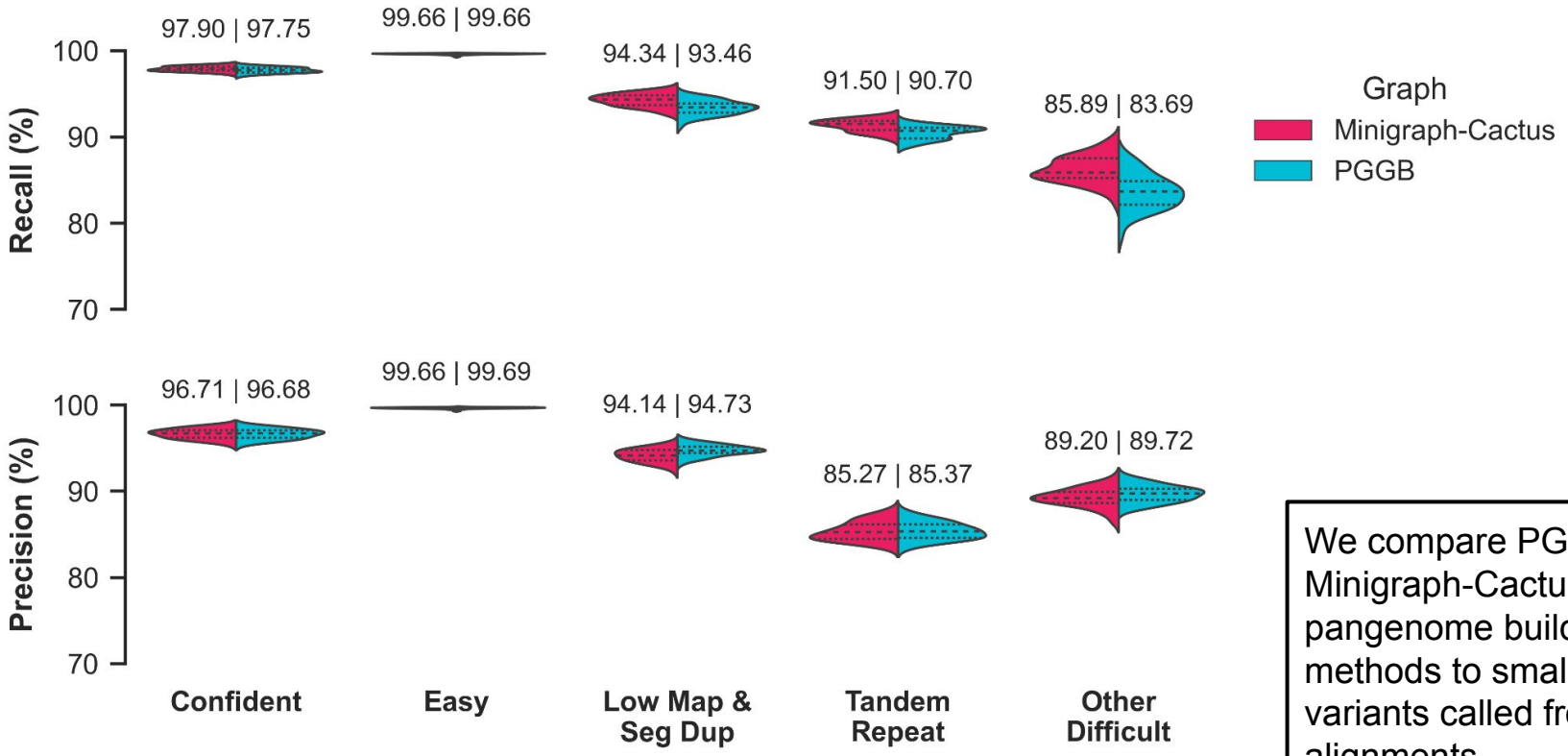


PGGB creates graphs in which each genome can act as a reference, so we choose our reference as needed by later analysis or work in graph space.



Multiple graph building methods show consistent quality

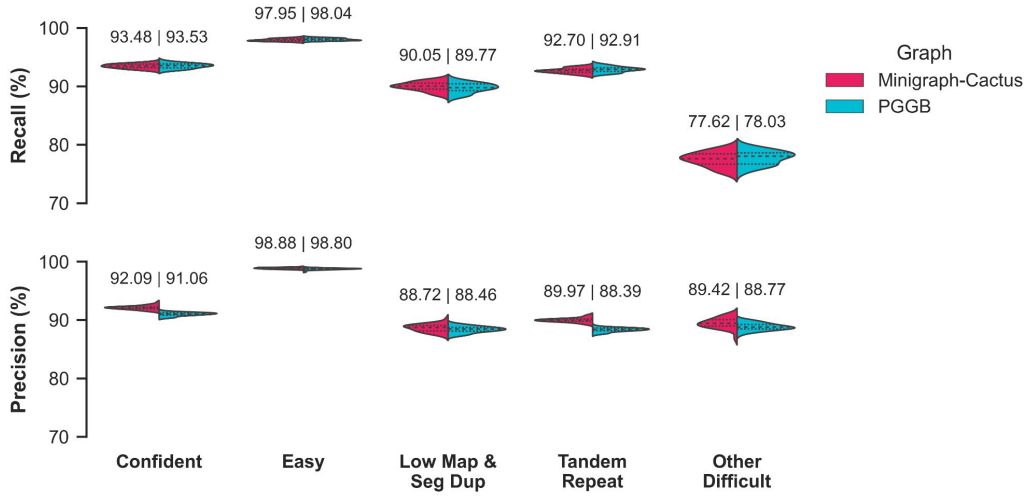
* variants extracted from graphs with vg deconstruct

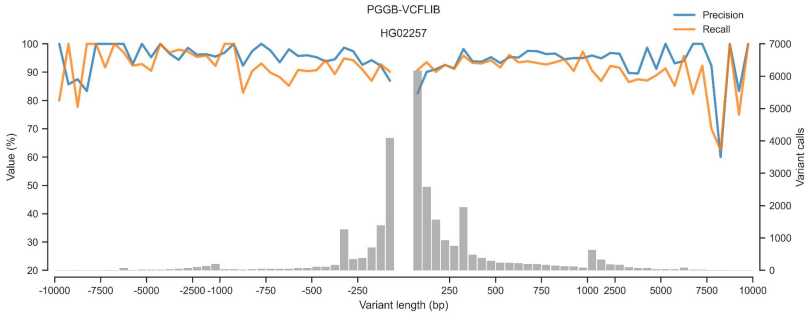


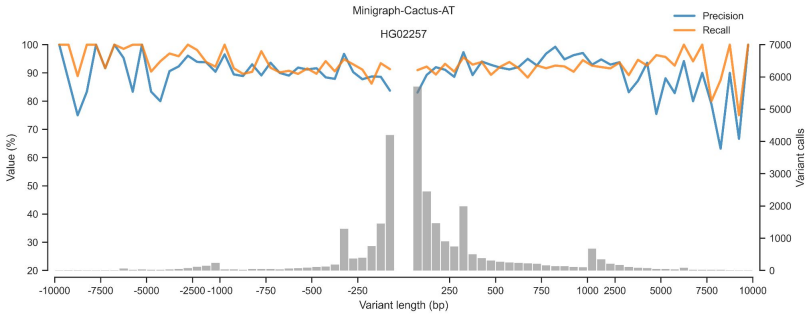
We compare PGGB and Minigraph-Cactus pangenome building methods to small variants called from HiFi alignments.

Wen-Wei Liao

The graphs accurately characterize structural variants





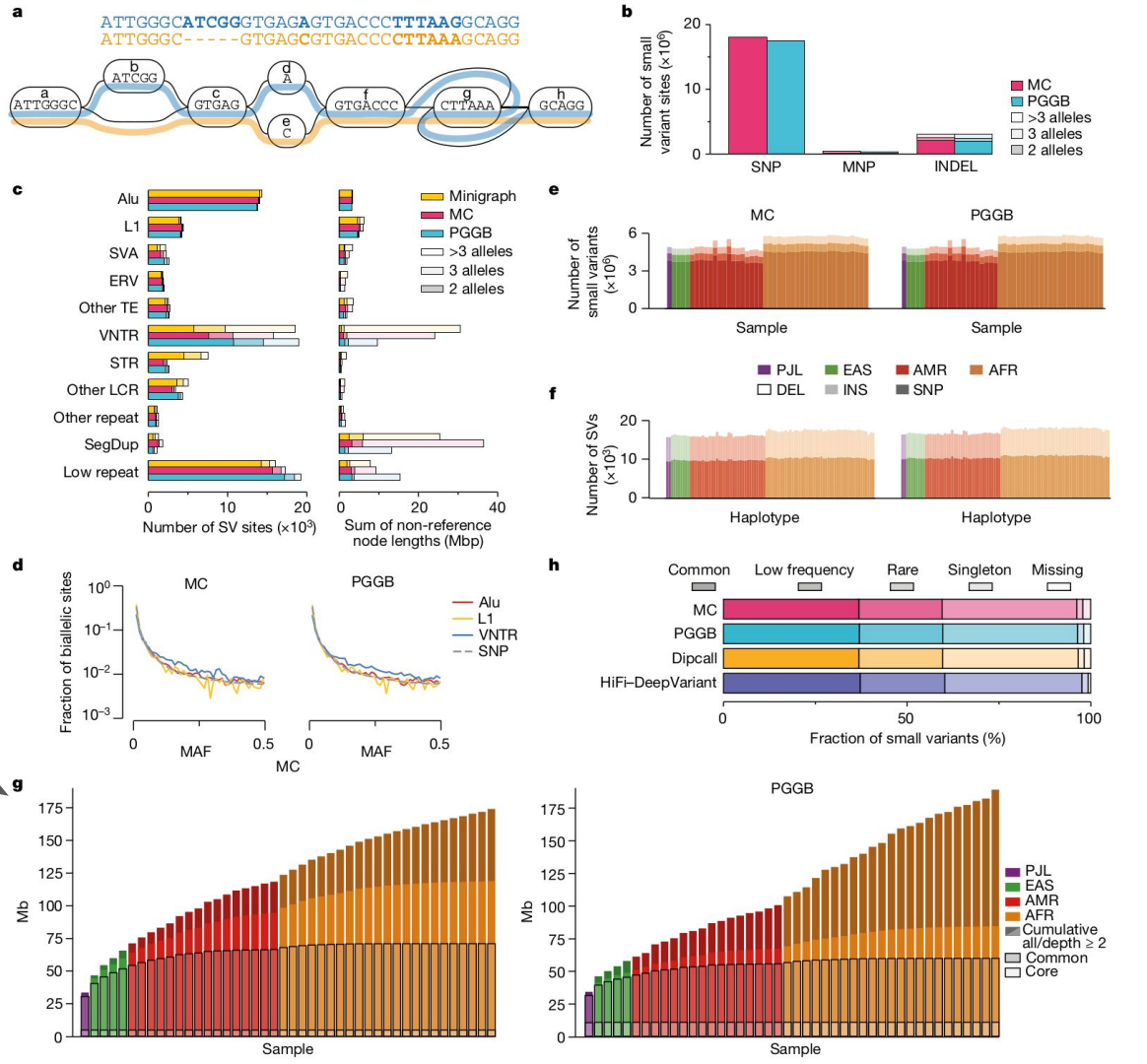


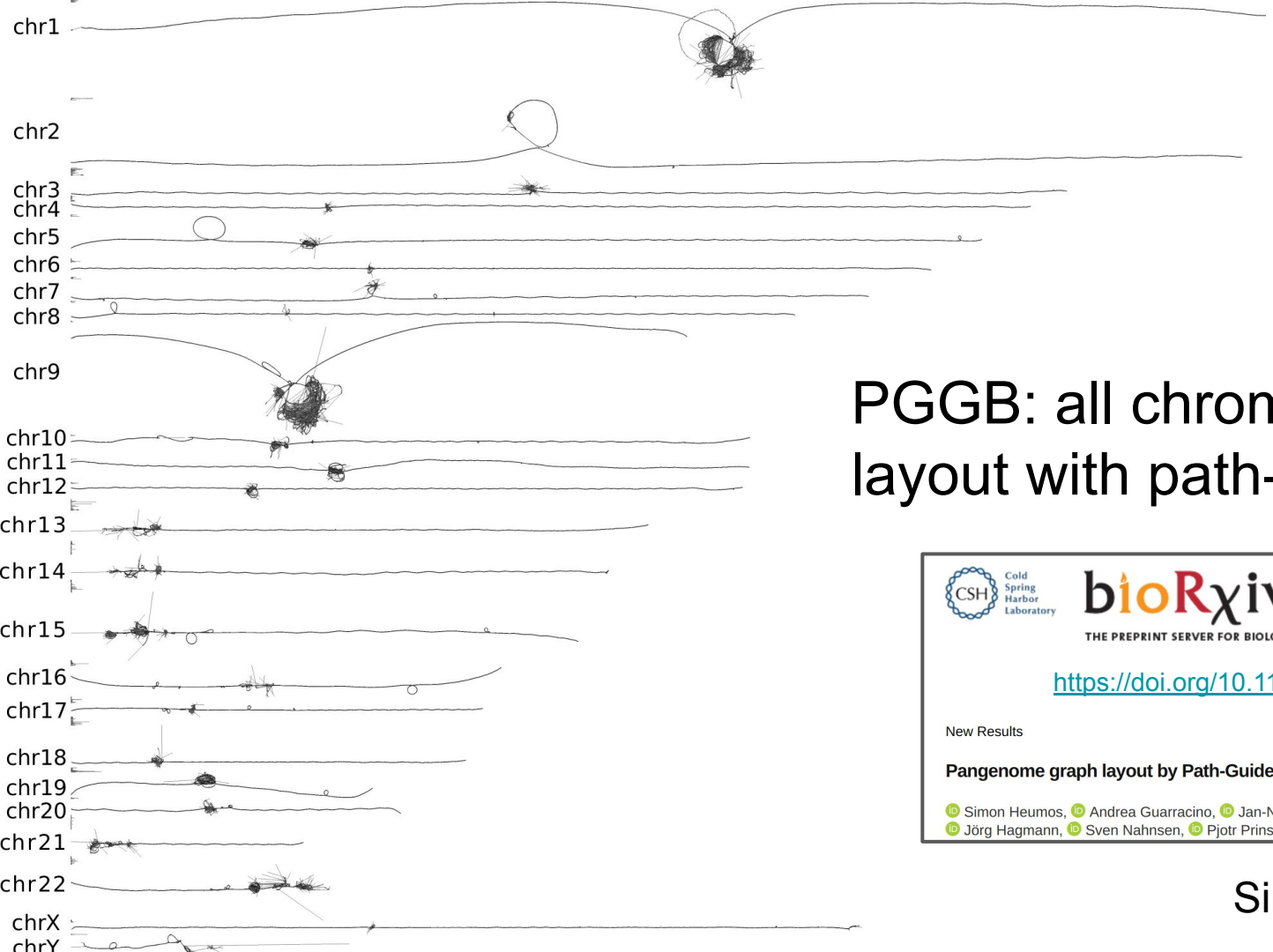
Comparing to consensus calls made by many reference-based SV callers.

Wen-Wei Liao

MC and PGGB show very similar pictures of the pangenome

- T2T-CHM13 adds **~200MB of heterochromatin** to reference.
- Draft pangenome adds **~100MB of polymorphic euchromatin** (and a lot more heterochromatin),
- **0.6-4.4 Mb of additional genic sequences per haplotype** compared to GRCh38 (38 gene CNVs/haplotype).





PGGB: all chromosomes, layout with path-guided SGD

Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

https://doi.org/10.1101/2023.09.22.558964

New Results

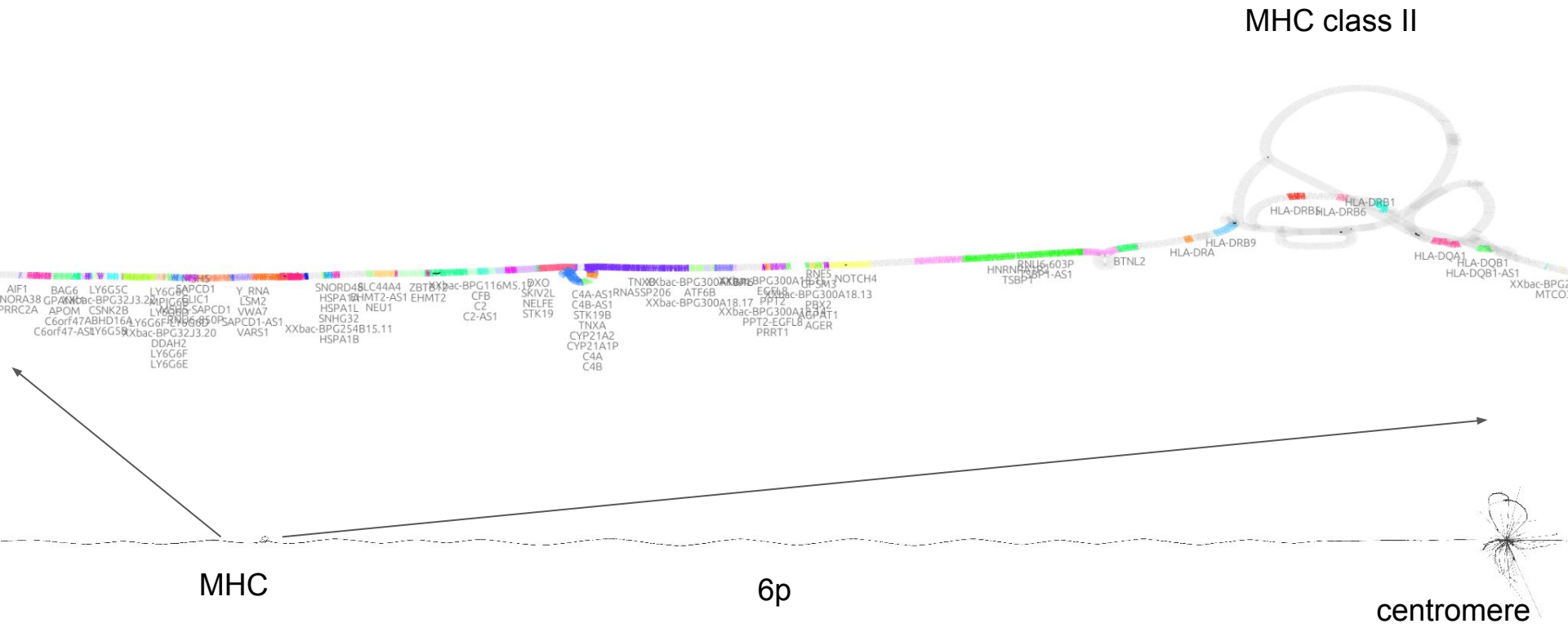
Follow this preprint

Pangenome graph layout by Path-Guided Stochastic Gradient Descent

Simon Heumos, Andrea Guarracino, Jan-Niklas M. Schmelzle, Jiajie Li, Zhiru Zhang, Jörg Hagmann, Sven Nahnsen, Pjotr Prins, Erik Garrison

Simon Heumos

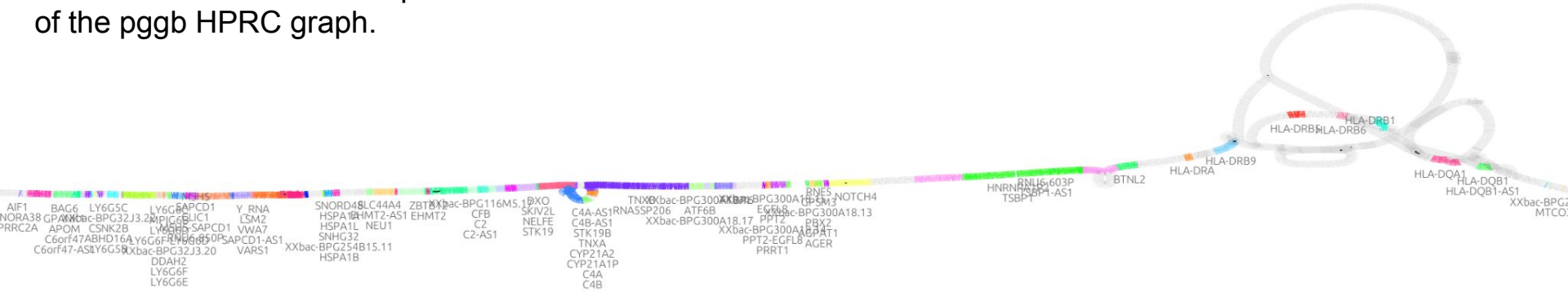
C4A/B in pggb graph



C4A/B in pggb graph

MHC class II

Genbank annotations on top of the pggb HPRC graph.



https://github.com/chfi/gfaestus

Christian Fischer (UTHSC)

C4A/B in pggb graph

Genbank annotations on top of the pggb HPRC graph.

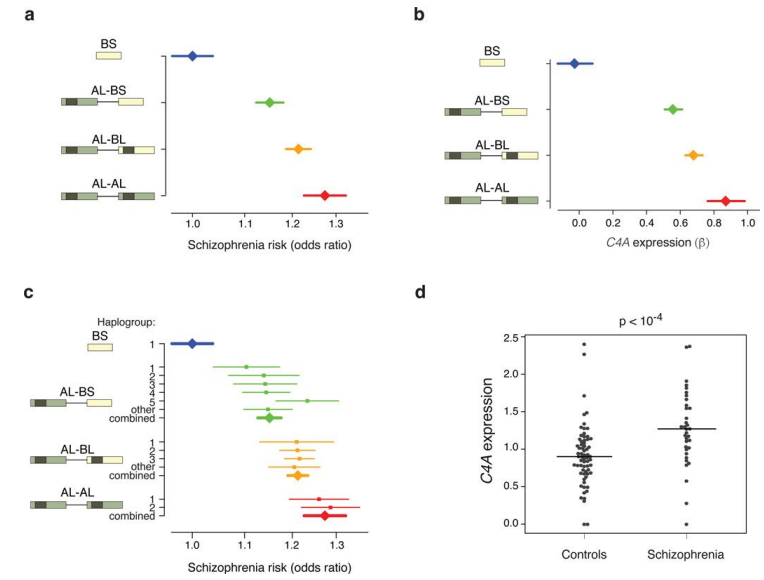
MHC class II

C4A/B

Christian Fischer (UTHSC)

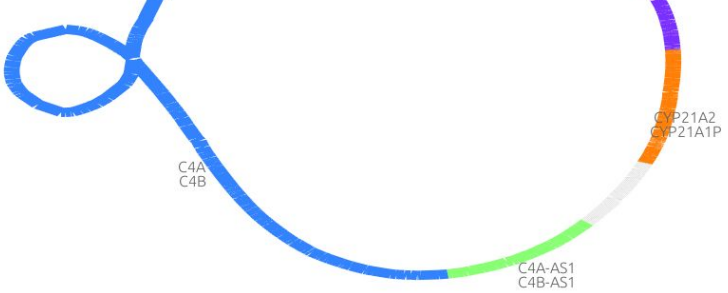
https://github.com/chfi/gfaestus

C4A/B in pggb graph



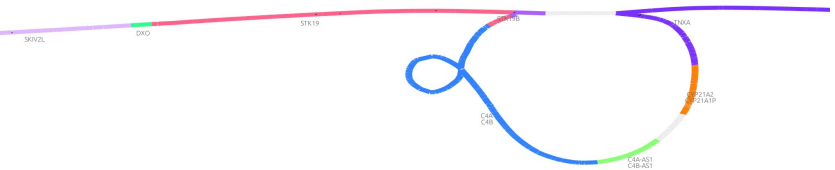
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4752392/

copy number is related to schizophrenia risk

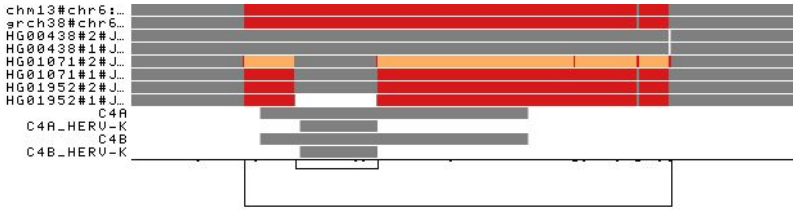


Christian Fischer (UTHSC)

We learn that genome evolution is often *nonlinear*

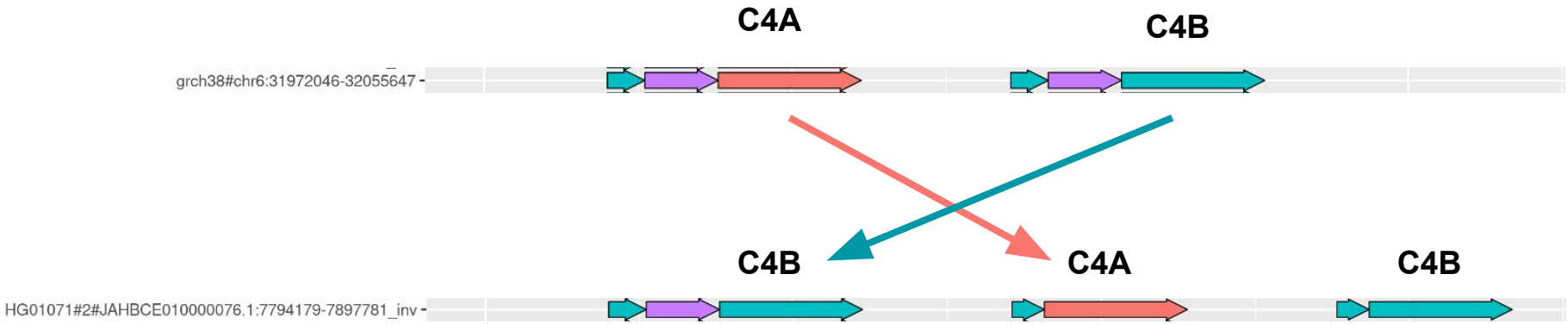


complement component 4 locus



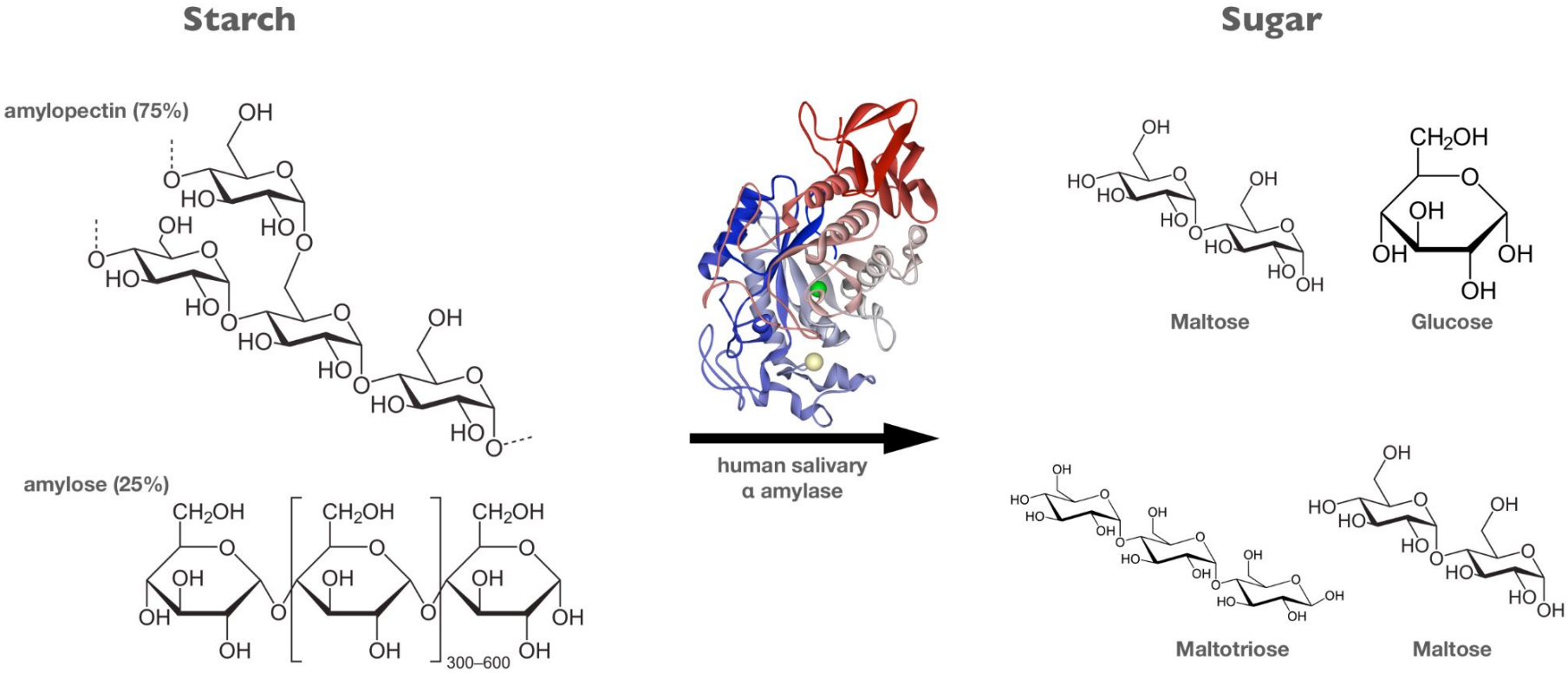
Large SVs predominantly occur at VNTRs which are simply loops in our pggb graphs.

We learn that genome evolution is often *nonlinear*

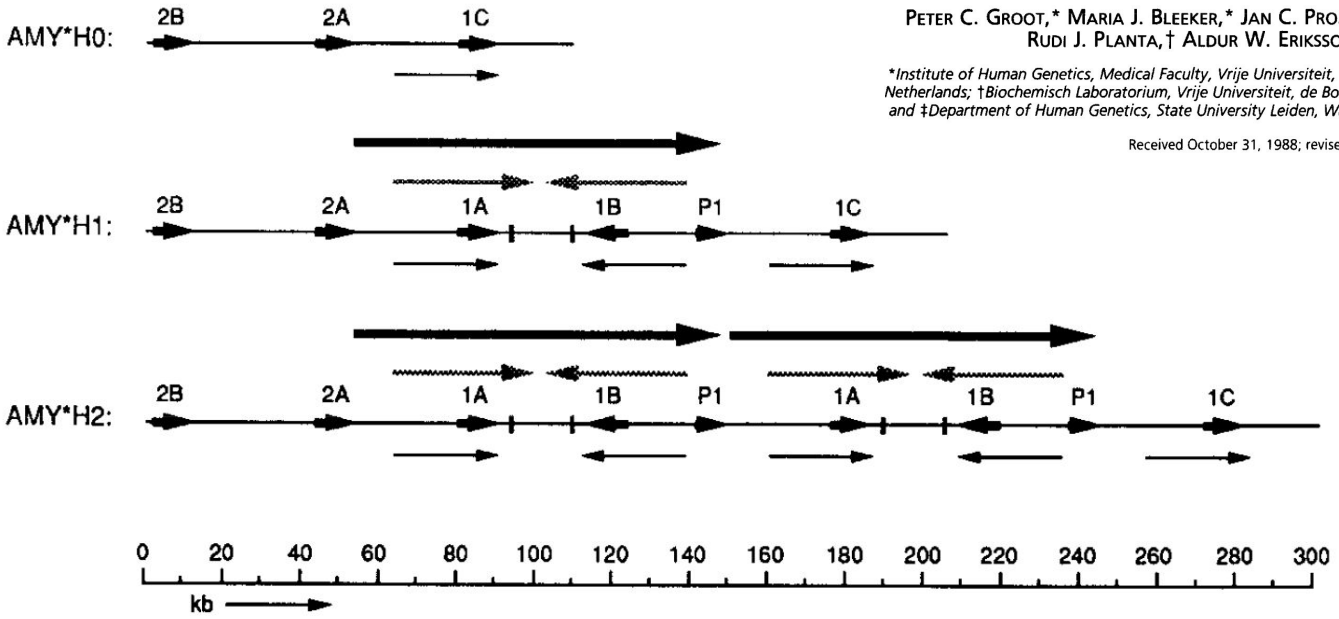


The human pangenome exposes selection at the amylase locus

Amylase digests starch into sugar



Amylase is a multi-copy gene family



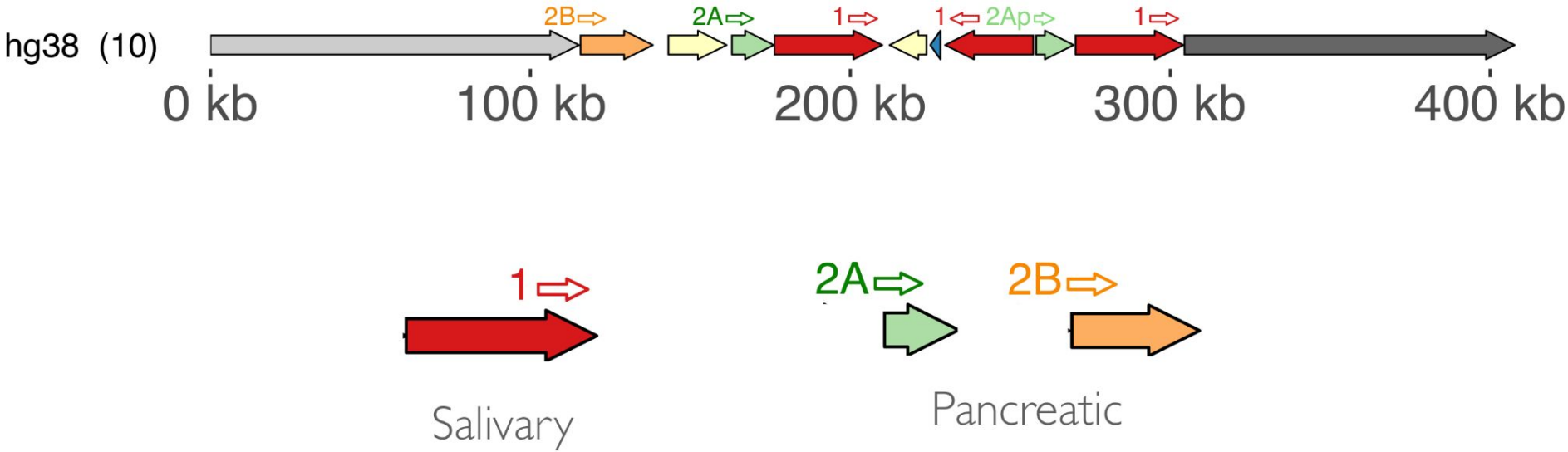
The Human α-Amylase Multigene Family Consists of Haplotypes with Variable Numbers of Genes

PETER C. GROOT,* MARIA J. BLEEKER,* JAN C. PRONK,* FRÉ ARWERT,* WILLEM H. MAGER,† RUDI J. PLANTA,† ALDUR W. ERIKSSON,* AND RUNE R. FRANTS‡

*Institute of Human Genetics, Medical Faculty, Vrije Universiteit, van der Boechorststraat 7, 1081 BT, Amsterdam, The Netherlands; †Biochemisch Laboratorium, Vrije Universiteit, de Boelelaan 1083, 1081 HV, Amsterdam, The Netherlands; and ‡Department of Human Genetics, State University Leiden, Wassenaarseweg 72, 2333 AL, Leiden, The Netherlands

Received October 31, 1988; revised February 10, 1989

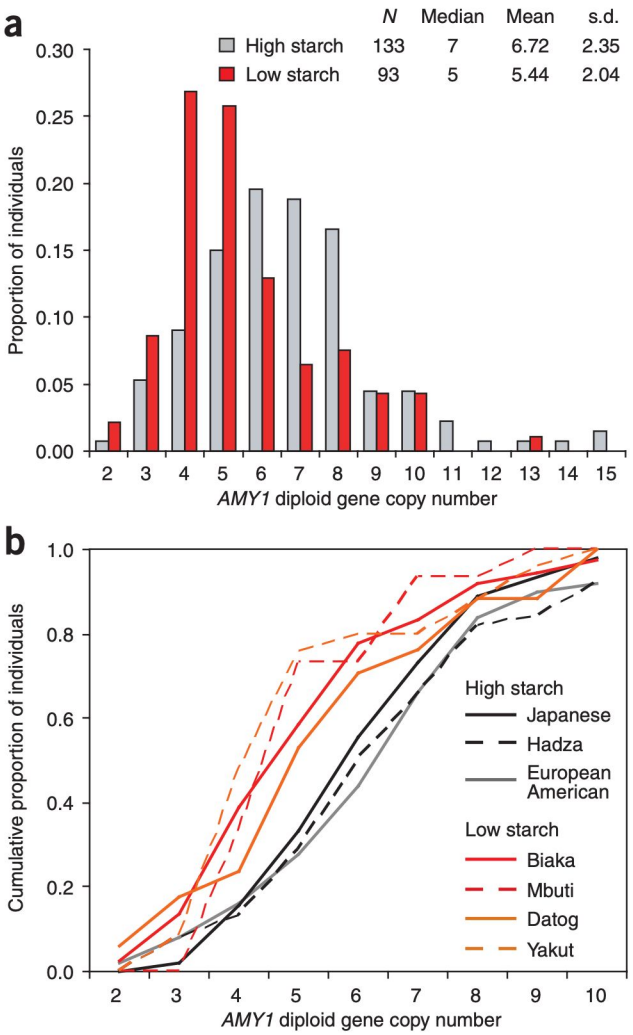
Amylase is a multi-copy gene family



Across human populations, diet correlates with amylase copy number

Diet and the evolution of human amylase gene copy number variation

George H Perry[1,2], Nathaniel J Dominy[3], Katrina G Claw[1,4], Arthur S Lee[2], Heike Fiegler[5], Richard Redon[5], John Werner[4], Fernando A Villanea[3], Joanna L Mountain[6], Rajeev Misra[4], Nigel P Carter[5], Charles Lee[2,7,8] & Anne C Stone[1,8]



Across mammals, amylase copy correlates with diet

Independent amylase gene copy number bursts correlate with dietary preferences in mammals

Petar Pajic[1,2], Pavlos Pavlidis[3], Kirsten Dean[1], Lubov Neznanova[2], Rose-Anne Romano[2], Danielle Garneau[4], Erin Daugherity[5], Anja Globig[6], Stefan Ruhl[2]*, Omer Gokcumen[1]*

But, no evidence for selection in humans!

FADS1 and the Timing of Human Adaptation to Agriculture

Sara Mathieson[1] and Iain Mathieson*[,2]
[1]Department of Computer Science, Swarthmore College, Swarthmore, PA
[2]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
***Corresponding author:** E-mail: mathi@pennmedicine.upenn.edu.
Associate editor: Evelyne Heyer

Abstract

Variation at the *FADS1/FADS2* gene cluster is functionally associated with differences in lipid metabolism and is often hypothesized to reflect adaptation to an agricultural diet. Here, we test the evidence for this relationship using both modern and ancient DNA data. We show that almost all the inhabitants of Europe carried the ancestral allele until the derived allele was introduced ~8,500 years ago by Early Neolithic farming populations. However, we also show that it was not under strong selection in these populations. We find that this allele, and other proposed agricultural adaptations at *LCT/MCM6* and *SLC22A4*, were not strongly selected until much later, perhaps as late as the Bronze Age. Similarly, increased copy number variation at the salivary amylase gene *AMY1* is not linked to the development of agriculture although, in this case, the putative adaptation precedes the agricultural transition. Our analysis shows that selection at the *FADS* locus was not tightly linked to the initial introduction of agriculture and the Neolithic transition. Further, it suggests that the strongest signals of recent human adaptation in Europe did not coincide with the Neolithic transition but with more recent changes in environment, diet, or efficiency of selection due to increases in effective population size.

Key words: **Human evolution, selection, ancient DNA, agriculture, diet.**

Selective sweep on human amylase genes postdates the split with Neanderthals

Charlotte E. Inchley[1], Cynthia D. A. Larbey[1], Nzar A. A. Shwan[2,3], Luca Pagani[1,4], Lauri Saag[4], Tiago Antão[5], Guy Jacobs[6], Georgi Hudjashov[4,7], Ene Metspalu[4], Mario Mitt[8,9], Christina A. Eichstaedt[1,10], Boris Malyarchuk[11], Miroslava Derenko[11], Joseph Wee[12], Syafiq Abdullah[13], François-Xavier Ricaut[14], Maru Mormina[15], Reedik Mägi[8], Richard Villems[4,16,17], Mait Metspalu[4], Martin K. Jones[1], John A. L. Armour[2] & Toomas Kivisild[1,4]

Discussion

In this study we have analysed genetic regions surrounding the human AMY cluster for evidence of natural selection and we have found: that human populations within and outside Africa are characterized by unusually low genetic diversity in the flanks of amylase locus relative to other genetic loci genome-wide; a young coalescent date postdating the human-Neanderthal population split; a significant Tajima's D signal in Africans; and the lack of strong signal of recent positive selection in human population groups we studied. These results are generally in line with Middle Pleistocene[9] rather than Holocene[1] selection at the *AMY* locus although the significantly

2015

And, conflicting GWAS results!

LETTERS

nature genetics

Low copy number of the salivary amylase gene predisposes to obesity

Mario Falchi[1,39,40], Julia Sarah El-Sayed Moustafa[1,39], Petros Takousis[1], Francesco Pesce[1,2], Amélie Bonnefond[3–6], Johanna C Andersson-Assarsson[1,7,8], Peter H Sudmant[9], Rajkumar Dorajoo[1,10], Mashael Nedham Al-Shafai[1,11], Leonardo Bottolo[12], Erdal Ozdemir[1], Hon-Cheong So[13], Robert W Davies[14], Alexandre Patrice[6,15,16], Robert Dent[17], Massimo Mangino[18], Pirro G Hysi[18], Aurélie Dechaume[3,4,6], Marlène Huyvaert[3,4,6], Jane Skinner[19], Marie Pigeyre[4,6,15,16], Robert Caiazzo[4,6,15,16], Violeta Raverdy[6,15,16], Emmanuel Vaillant[3,4,6], Sarah Field[20], Beverley Balkau[21,22], Michel Marre[23,24], Sophie Visvikis-Siest[25], Jacques Weill[26], Odile Poulain-Godefroy[3,4,6], Peter Jacobson[7,8], Lars Sjostrom[7,8], Christopher J Hammond[18], Panos Deloukas[20,27,28], Pak Chung Sham[13], Ruth McPherson[29,30], Jeannette Lee[31], E Shyong Tai[31–33], Robert Sladek[34–36], Lena M S Carlsson[7,8], Andrew Walley[1,37], Evan E Eichler[9,38], Francois Pattou[4,6,15,16], Timothy D Spector[18,40] & Philippe Froguel[1,3–6,40]

2014

Low AMY associated with obesity

LETTERS

nature genetics

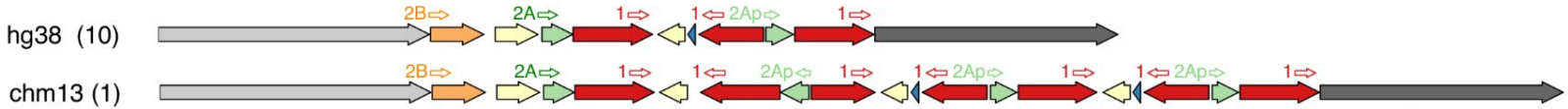
Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity

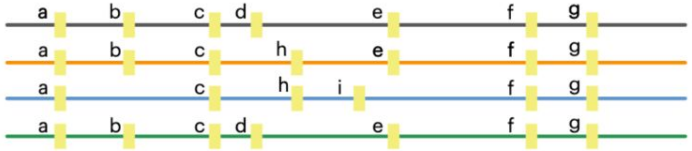
Christina L Usher[1], Robert E Handsaker[1–3], Tõnu Esko[1,2,4–6], Marcus A Tuke[7], Michael N Weedon[7], Alex R Hastie[8], Han Cao[8], Jennifer E Moon[1,2,4,5], Seva Kashin[2,3], Christian Fuchsberger[9,10], Andres Metspalu[6,11], Carlos N Pato[12], Michele T Pato[12], Mark I McCarthy[13–15], Michael Boehnke[9,10], David M Altshuler[1,2,16], Timothy M Frayling[7], Joel N Hirschhorn[1,2,4,5] & Steven A McCarroll[1–3]

2015

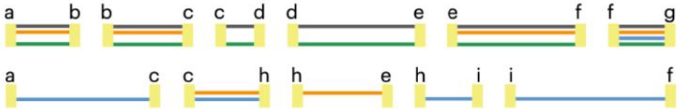
Low AMY **NOT** associated with obesity

Human amylase copy number diversity

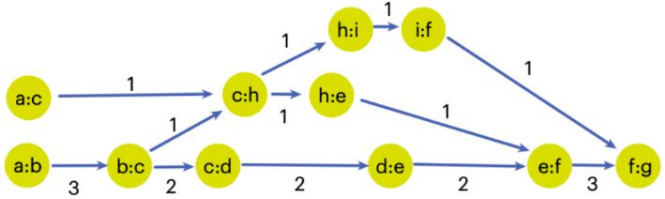




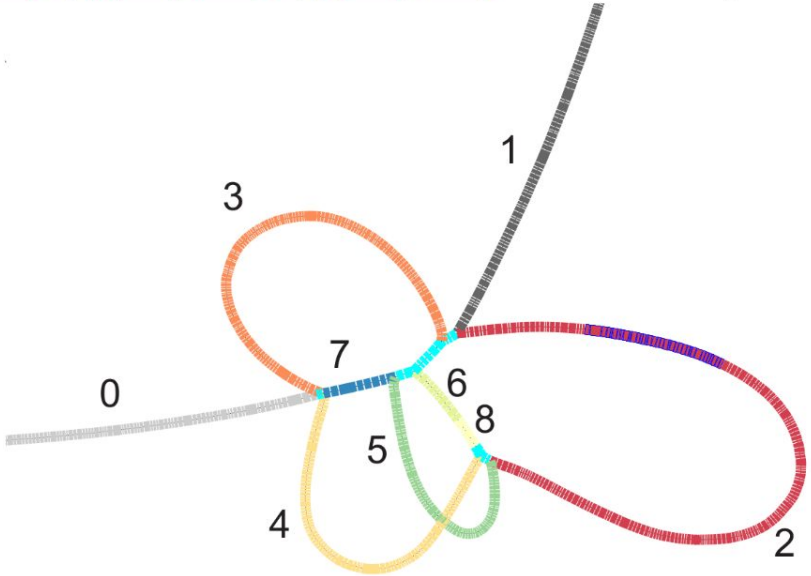
Sequence with minimizer anchors



Graph vertex: A set of sequences with shared minimizer anchors at both ends



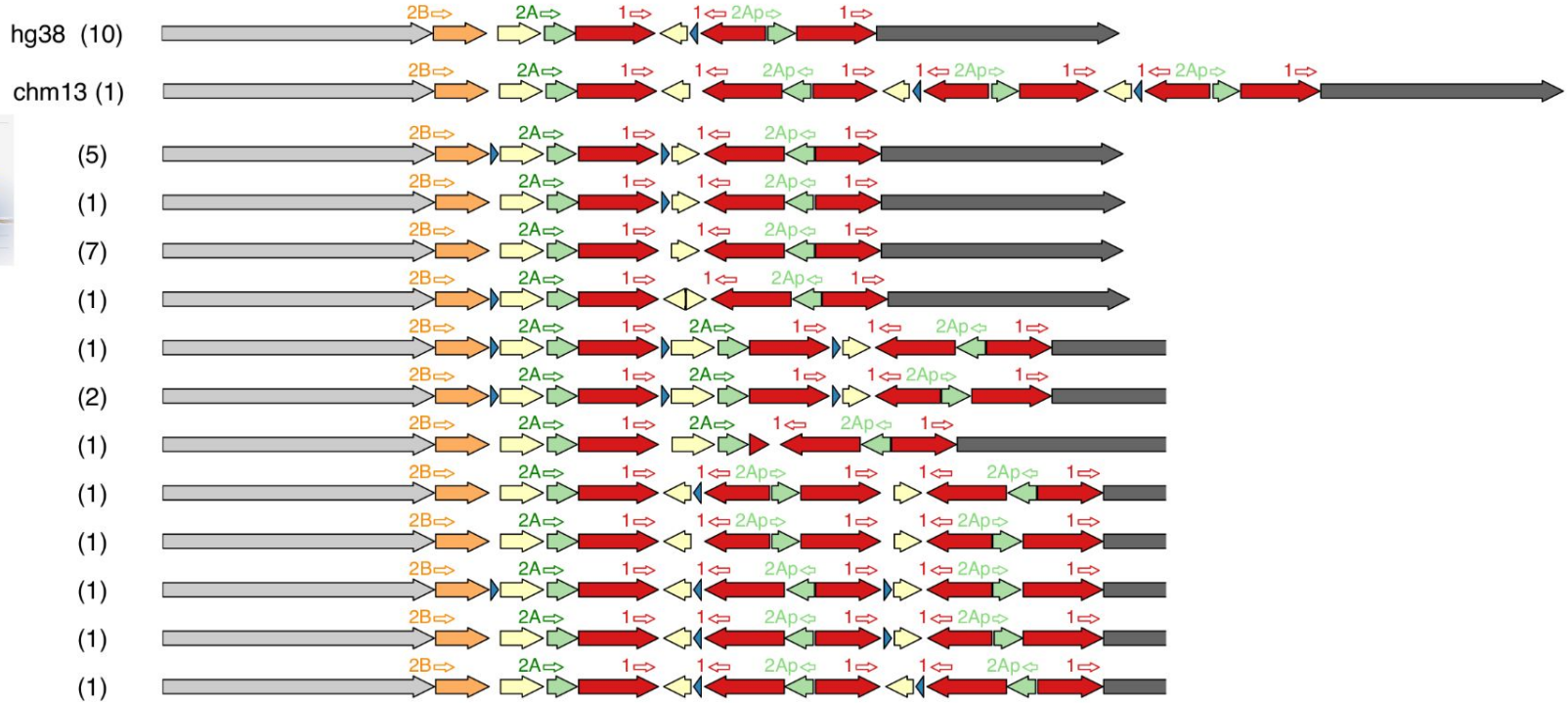
Induced MAP-graph

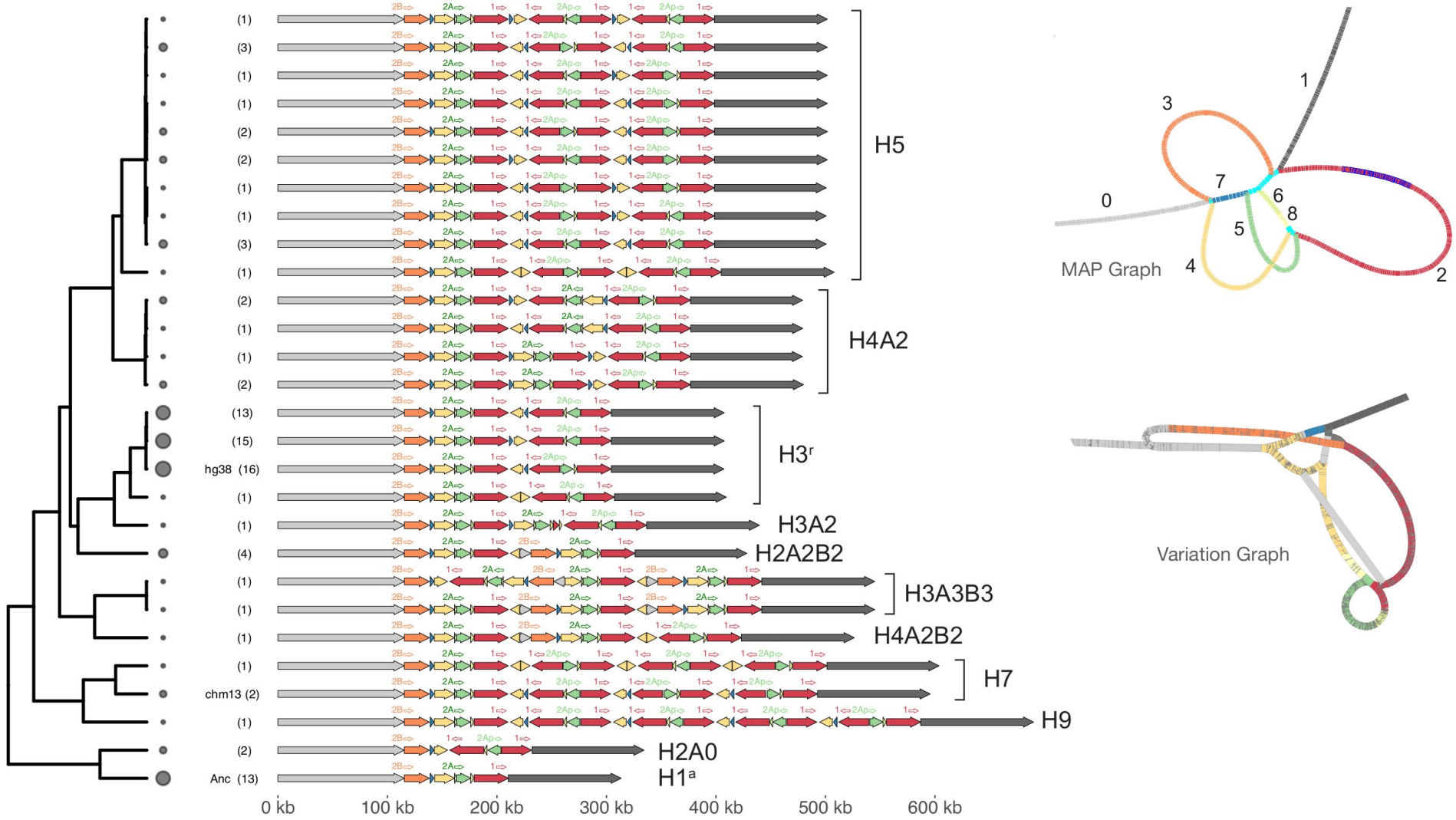


MAP graph (PGR-TK)

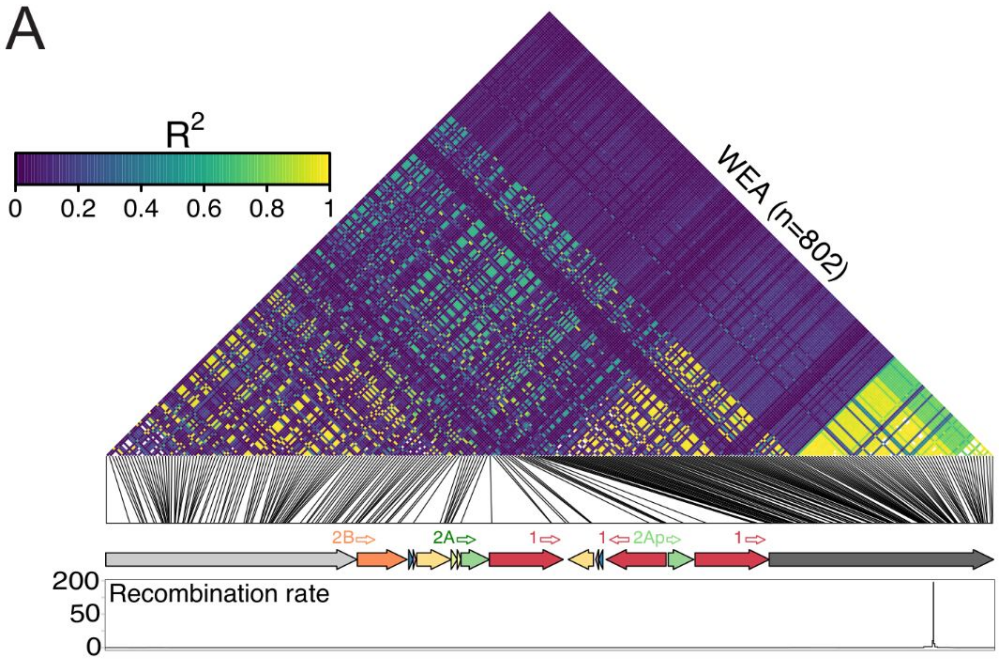
https://doi.org/10.1038/s41592-023-01914-y

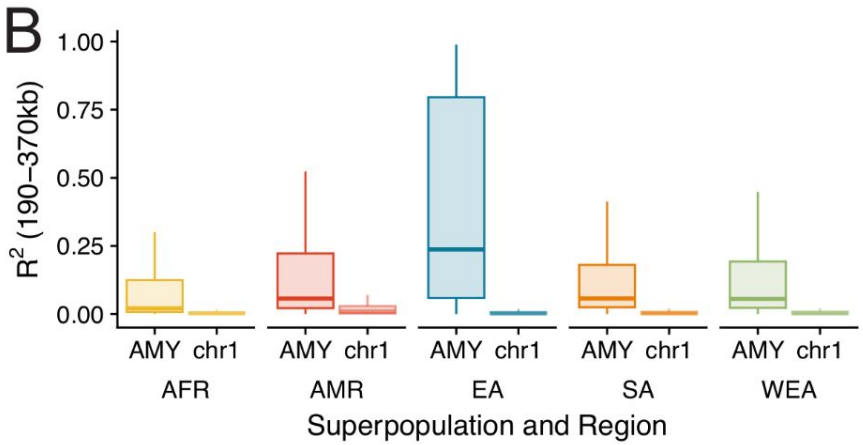
Human amylase copy number diversity

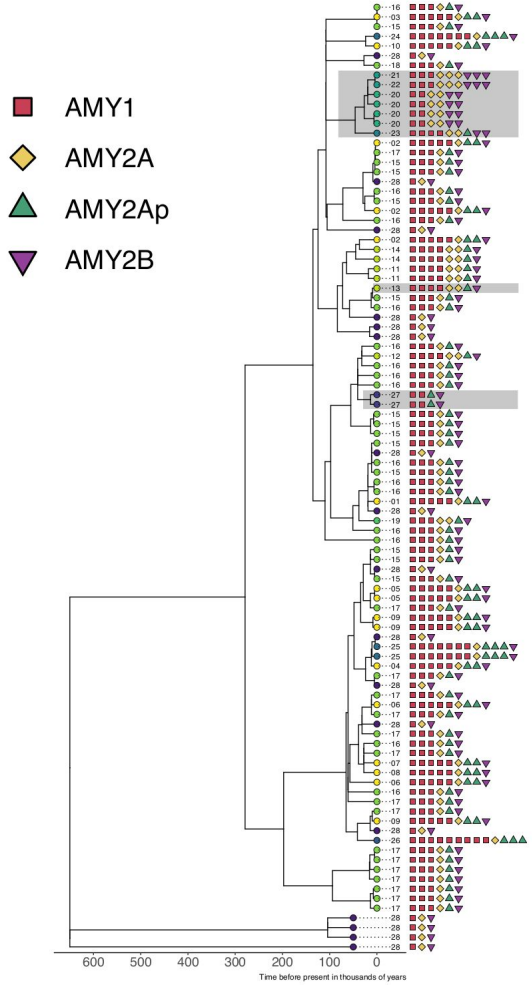



(1)
(3)
(1)
(1)
(2)
(2)
(1)
(1)
(3)
(1)
H5
(2)
(1)
(1)
(2)
H4A2
(13)
(15)
hg38 (16)
(1)
H3r
(1)
H3A2
(4)
H2A2B2
(1)
(1)
H3A3B3
(1)
H4A2B2
(1)
chm13 (2)
H7
(1)
H9
(2)
H2A0
Anc (13)
H1a
0 kb
100 kb
200 kb
300 kb
400 kb
500 kb
600 kb
MAP Graph
Variation Graph

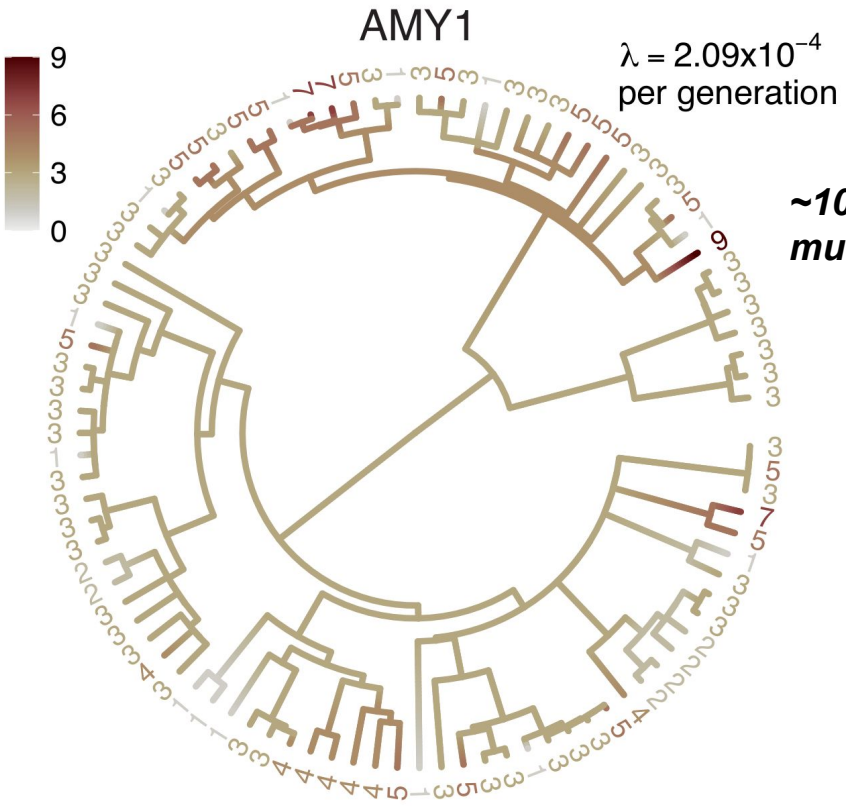
Strong linkage disequilibrium block across AMY locus







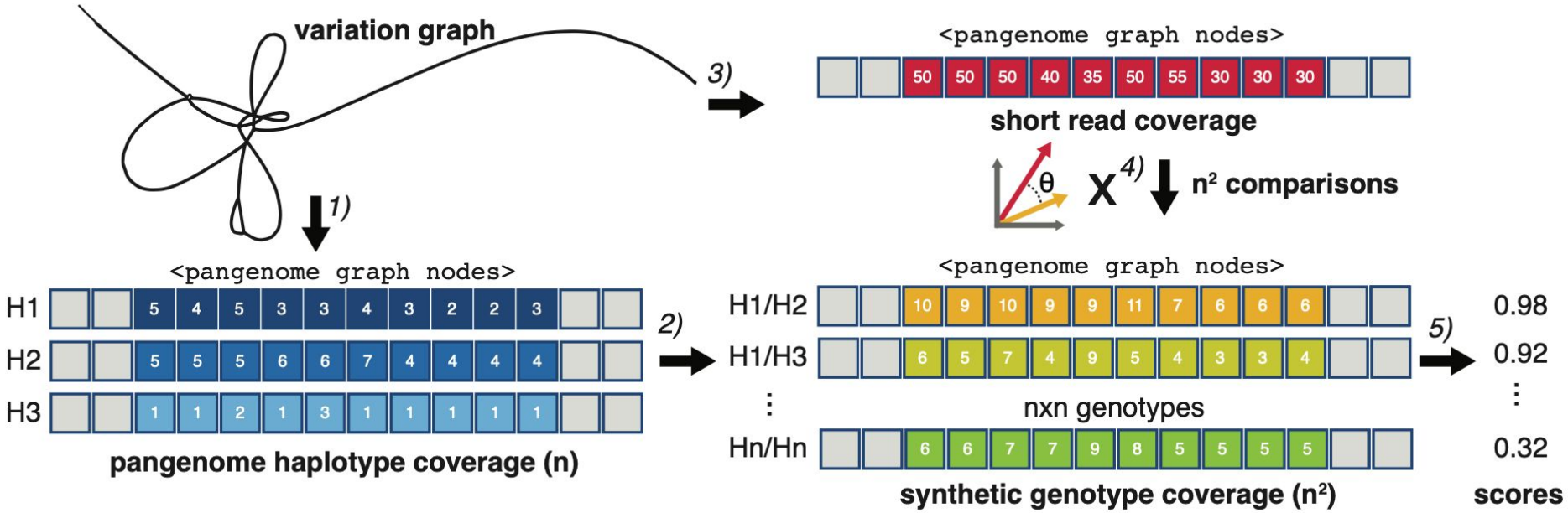
tree from left-flanking SNPs



~10,000X SNP mutation rate!

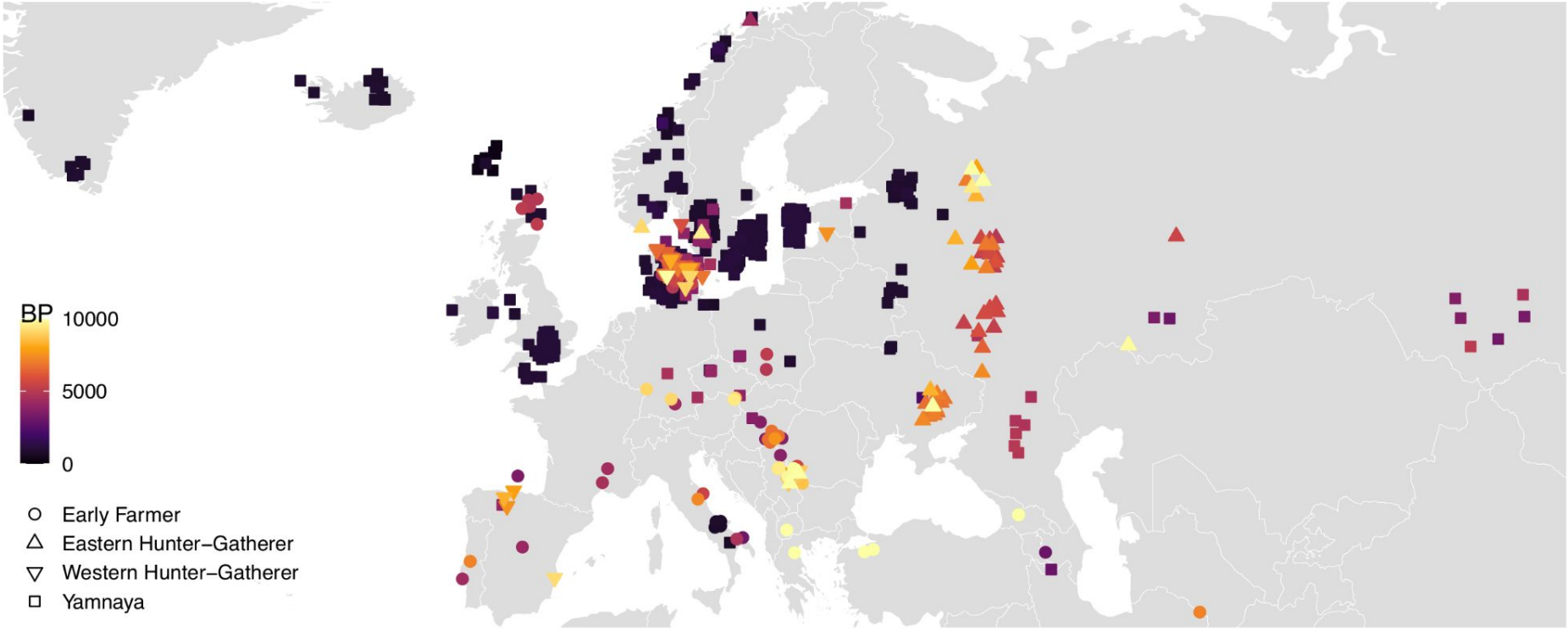
Because structural mutation rate is high, haplotypes with different lengths and configurations have emerged on the same background! *This breaks GWAS and selection scans.*

Deconvolving haplotypes from short reads



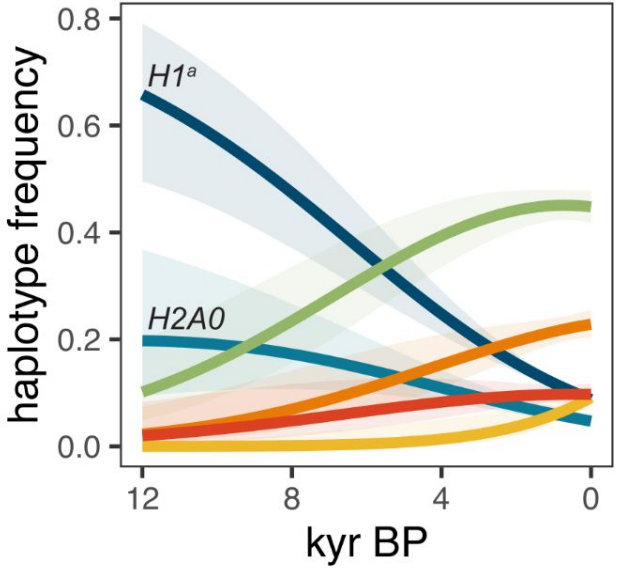
cross-validation with ddPCR, copy number, and hold-one-out experiments shows ~95% accuracy

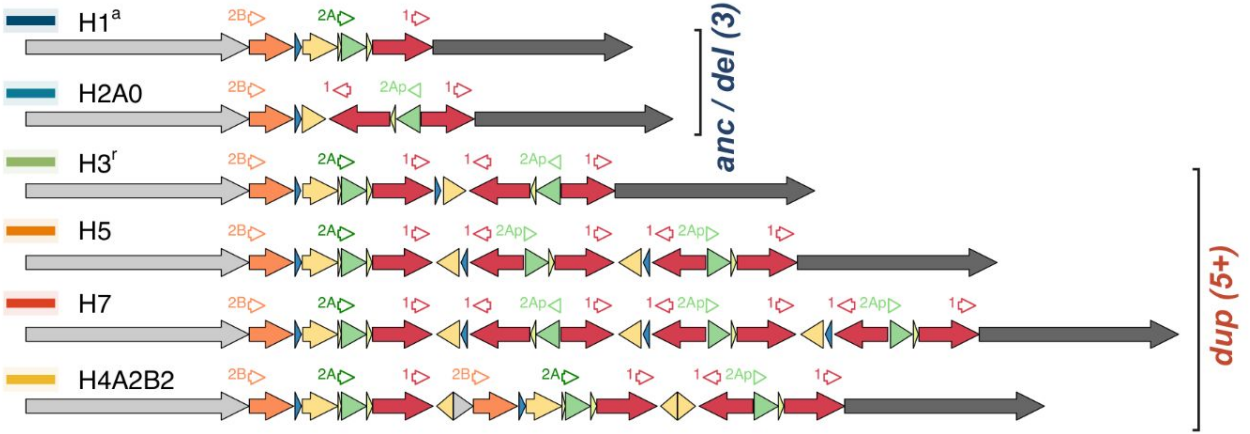
Recent evolution from 534 ancient European genomes



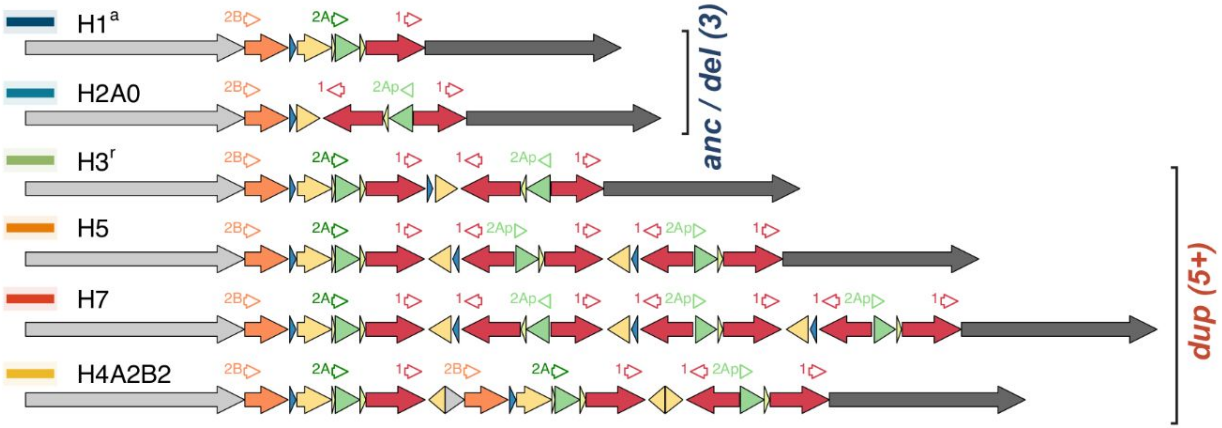
Data from Allentoft et al 2023 - in press

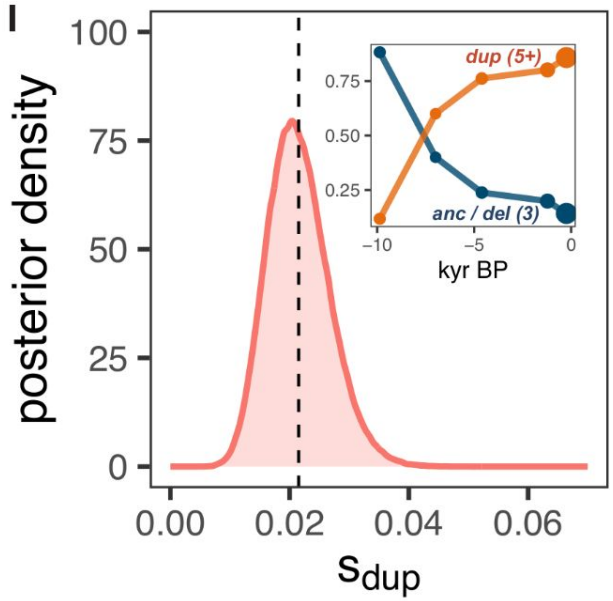
Recent evolution of human amylase copy number diversity





Evidence for selection of high-copy amylase haplotypes





There *is* selection at amylase in humans for haplotypes with more AMY1 copies.

Selection coefficient of 0.02 is equivalent to selection at lactase!

https://doi.org/10.1101/2020.11.17.387761

The human pangenome explains acrocentric evolution

https://doi.org/10.1126/science.abj6987

SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, ARIEL GERSHMAN, SERGEY AGANEZOV, SAVANNAH J. HOYT, MARK DIEKHANS, GLENNIS A. LOGSDON, MICHAEL ALONGE, STYLIANOS E. ANTONARAKIS, MATTHEW BORCHERS, GERARD G. BOUFFARD, SHELISE Y. BROOKS, GINA V. CALDAS, NAE-CHYUN CHEN, HAOYU CHENG, CHEN-SHAN CHIN, WILLIAM CHOW, LEONARDO G. DE LIMA, PHILIP C. DISHUCK, RICHARD DURBIN, TATIANA DVORKINA, IAN T. FIDDES, GIULIO FORMENTI, ROBERT S. FULTON, ARKARACHAI FUNGTAMMASAN, ERIK GARRISON, PATRICK G. S. GRADY, TINA A. GRAVES-LINDSAY, IRA M. HALL, NANCY F. HANSEN, GABRIELLE A. HARTLEY, MARINA HAUKNESS, KERSTIN HOWE, MICHAEL W. HUNKAPILLER, CHIRAG JAIN, MITEN JAIN, ERICH D. JARVIS, PETER KERPEDJIEV, MELANIE KIRSCHE, MIKHAIL KOLMOGOROV, JONAS KORLACH, MILINN KREMITZKI, HENG LI, VALERIE V. MADURO, TOBIAS MARSCHALL, ANN M. MCCARTNEY, JENNIFER MCDANIEL, DANNY E. MILLER, JAMES C. MULLIKIN, EUGENE W. MYERS, NATHAN D. OLSON, BENEDICT PATEN, PAUL PELUSO, PAVEL A. PEVZNER, DAVID PORUBSKY, TAMARA POTAPOVA, EVGENY I. ROGAEV, JEFFREY A. ROSENFELD, STEVEN L. SALZBERG, VALERIE A. SCHNEIDER, FRITZ J. SEDLAZECK, KISHWAR SHAFIN, COLIN J. SHEW, ALAINA SHUMATE, YING SIMS, ARIAN F. A. SMIT, DANIELA C. SOTO, IVAN SOVIĆ, JESSICA M. STORER, AARON STREETS, BETH A. SULLIVAN, FRANÇOISE THIBAUD-NISSEN, JAMES TORRANCE, JUSTIN WAGNER, BRIAN P. WALENZ, AARON WENGER, JONATHAN M. D. WOOD, CHUNLIN XIAO, STEPHANIE M. YAN, ALICE C. YOUNG, SAMANTHA ZARATE, URVASHI SURTI, RAJIV C. MCCOY, MEGAN Y. DENNIS, IVAN A. ALEXANDROV, JENNIFER L. GERTON, RACHEL J. O'NEILL, WINSTON TIMP, JUSTIN M. ZOOK, MICHAEL C. SCHATZ, EVAN E. EICHLER, KAREN H. MIGA, AND ADAM M. PHILLIPPY

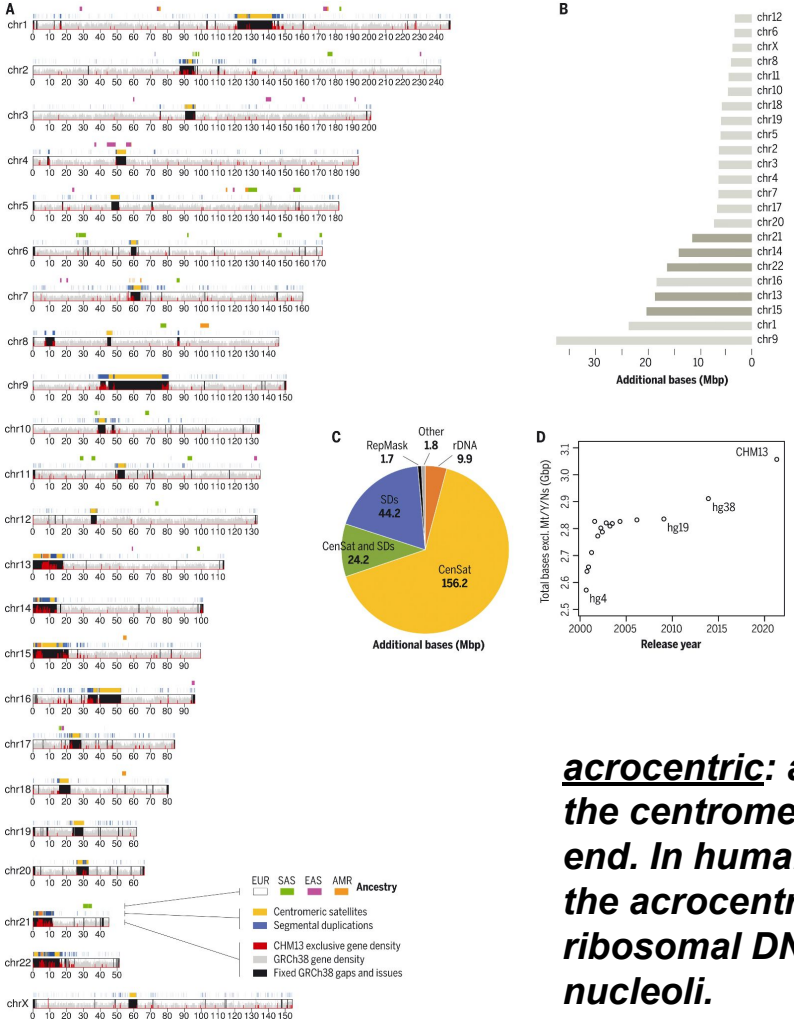
fewer

Authors Info & Affiliations

SCIENCE • 31 Mar 2022 • Vol 376, Issue 6588 • pp. 44-53 • DOI: 10.1126/science.abj6987

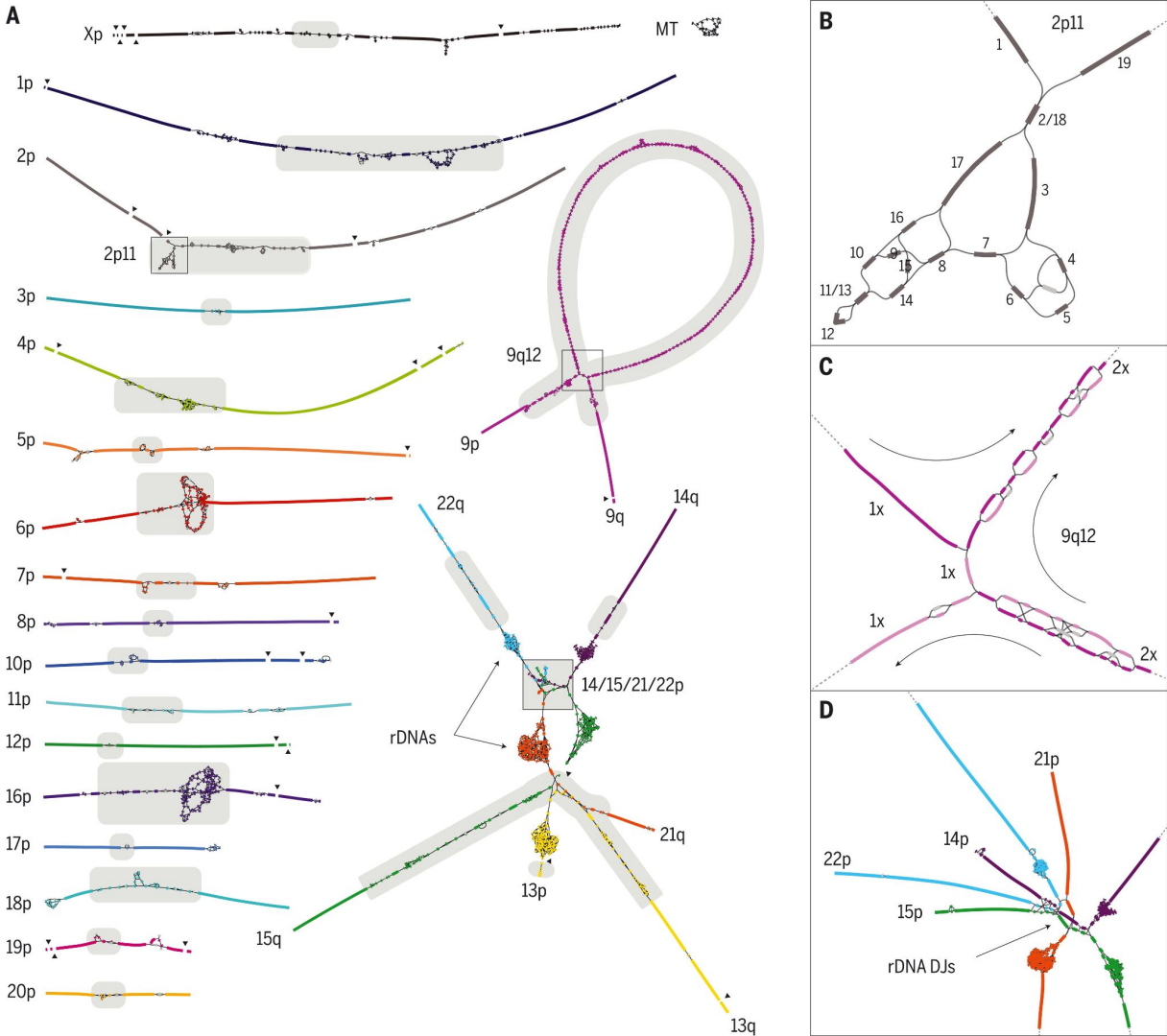
T2T-CHM13 fills 8% of the reference which was incomplete

All of the acrocentric p-arms were assembled for the first time!

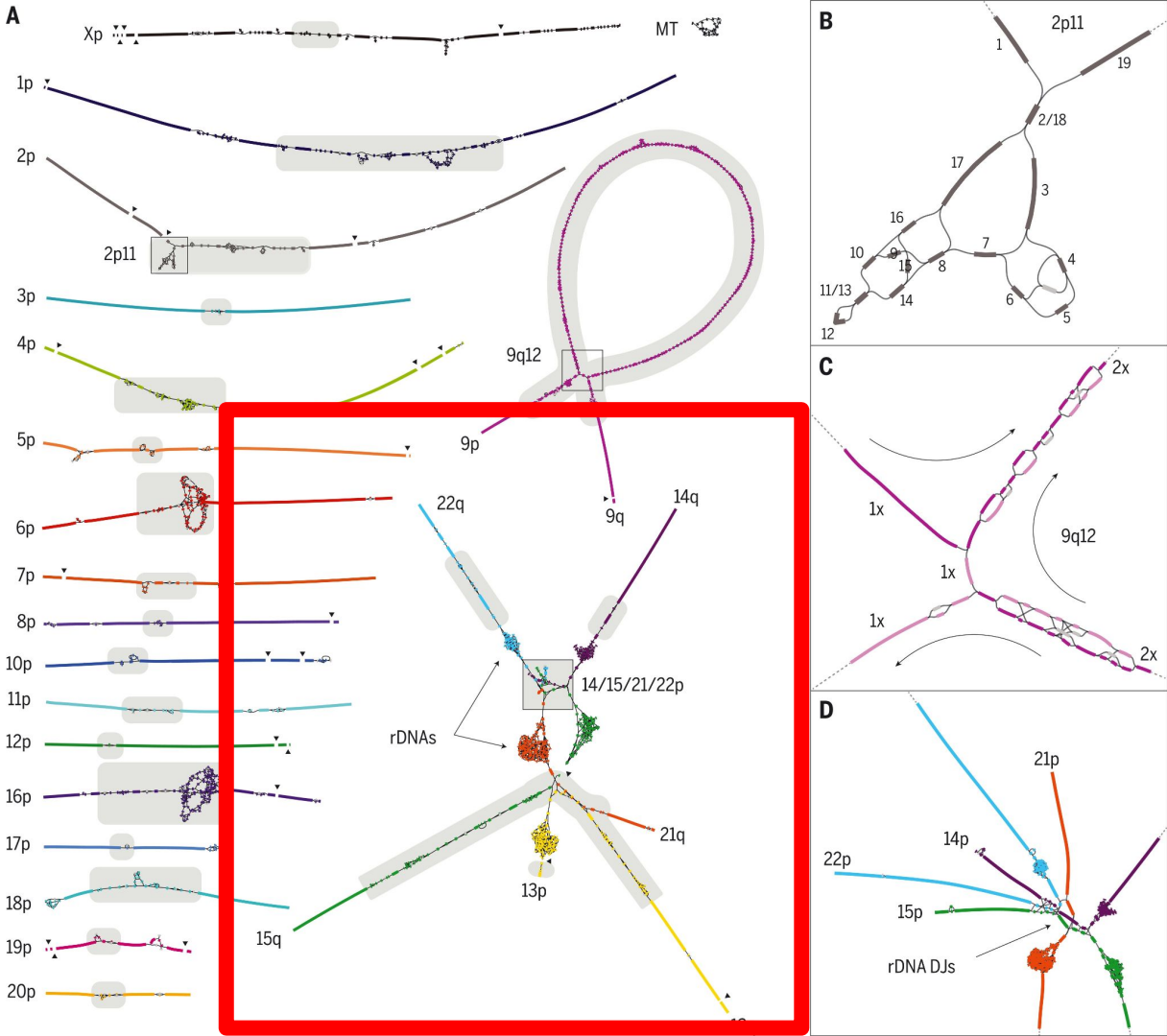


<u>acrocentric</u>: a chromosome where the centromere is almost at one end. In humans the short arms of the acrocentrics are the location of ribosomal DNA and organize the nucleoli.

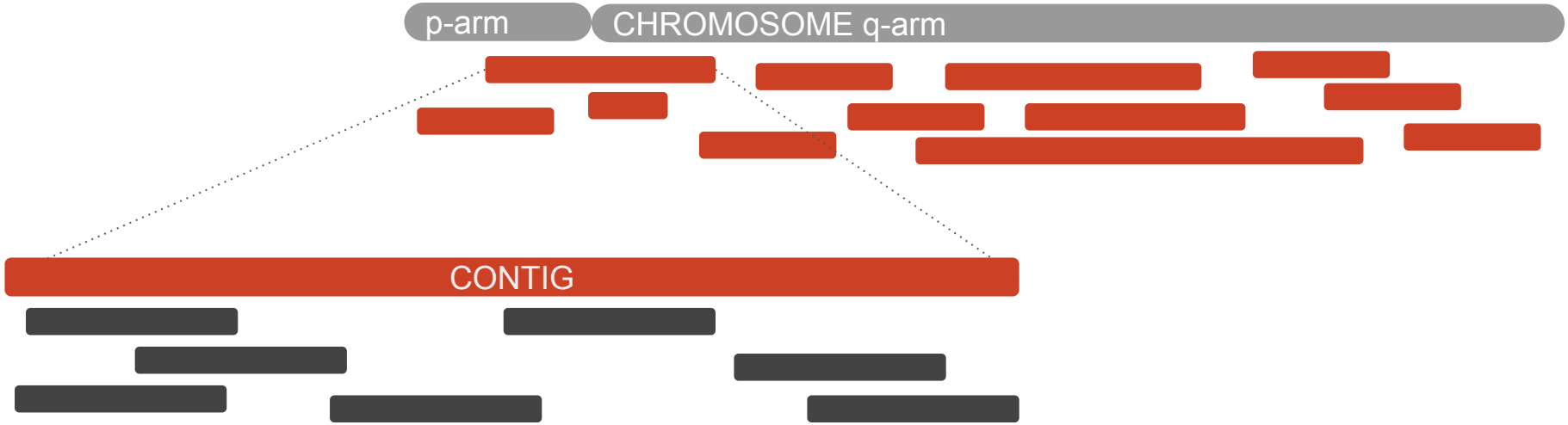
Revealing new mysteries…



Revealing new mysteries…



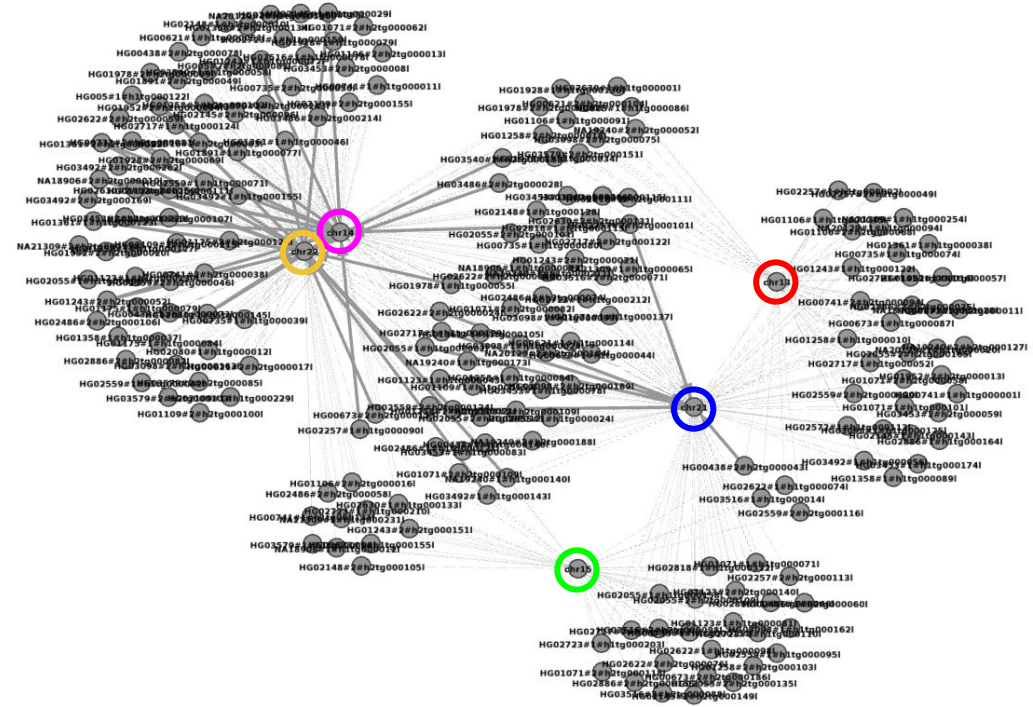
Assembly jargon: *Contigs*



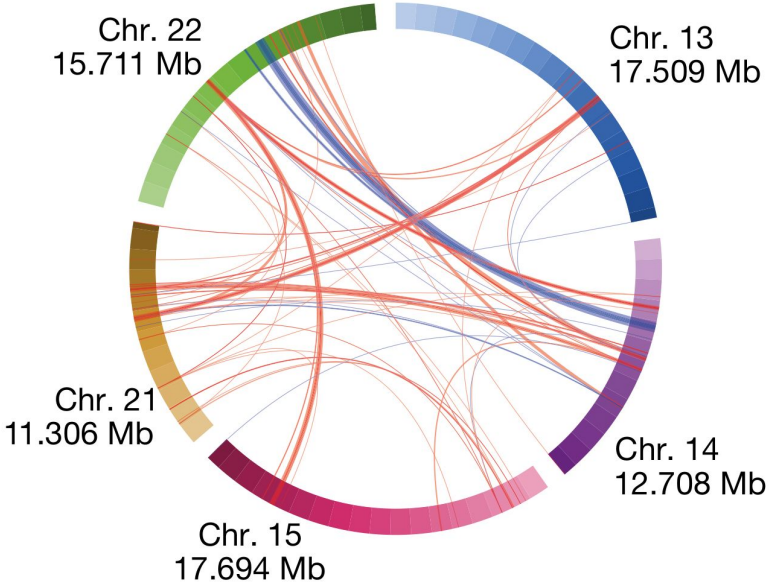

```
contigs_are_assembled_from_sequencing_reads
  ntigs_ar          d_from_sequenc
     gs_are_c                     cing_reads
contig     _assemble          uencing_r
```

https://github.com/lh3/minimap2/blob/67dd906a80988dddacc8c551623fdc75b0c12dd2/misc/paftools.js#L2605-L2719

HPRC acrocentric “misjoins”

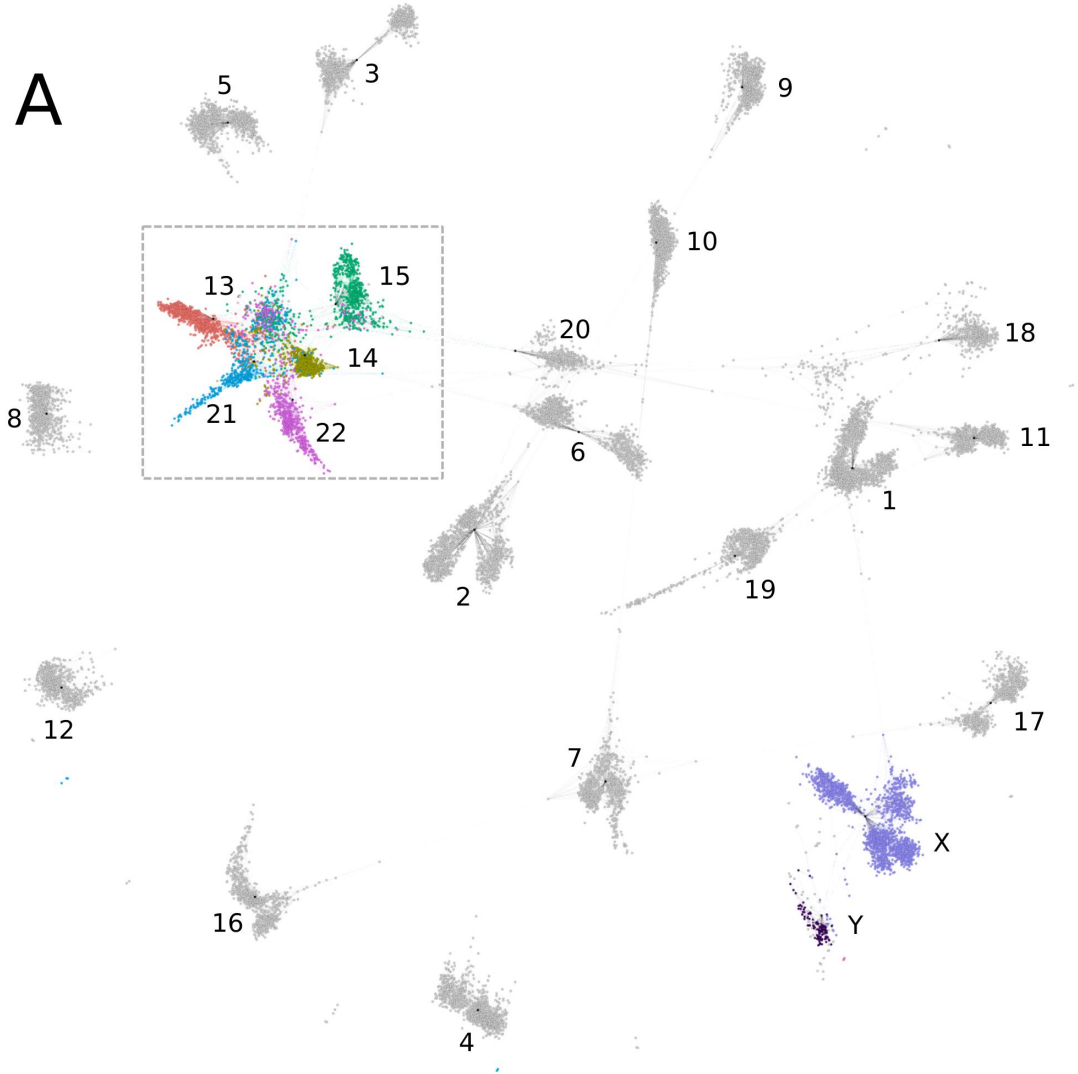


Alignment graph of the misjoins. Every node is a contig and every edge represents the number of mapping between nodes. Alignment graph obtained with pafnet and visualized with gephi. Color code: **chr13**, **chr14**, **chr15**, **chr21**, **chr22**.



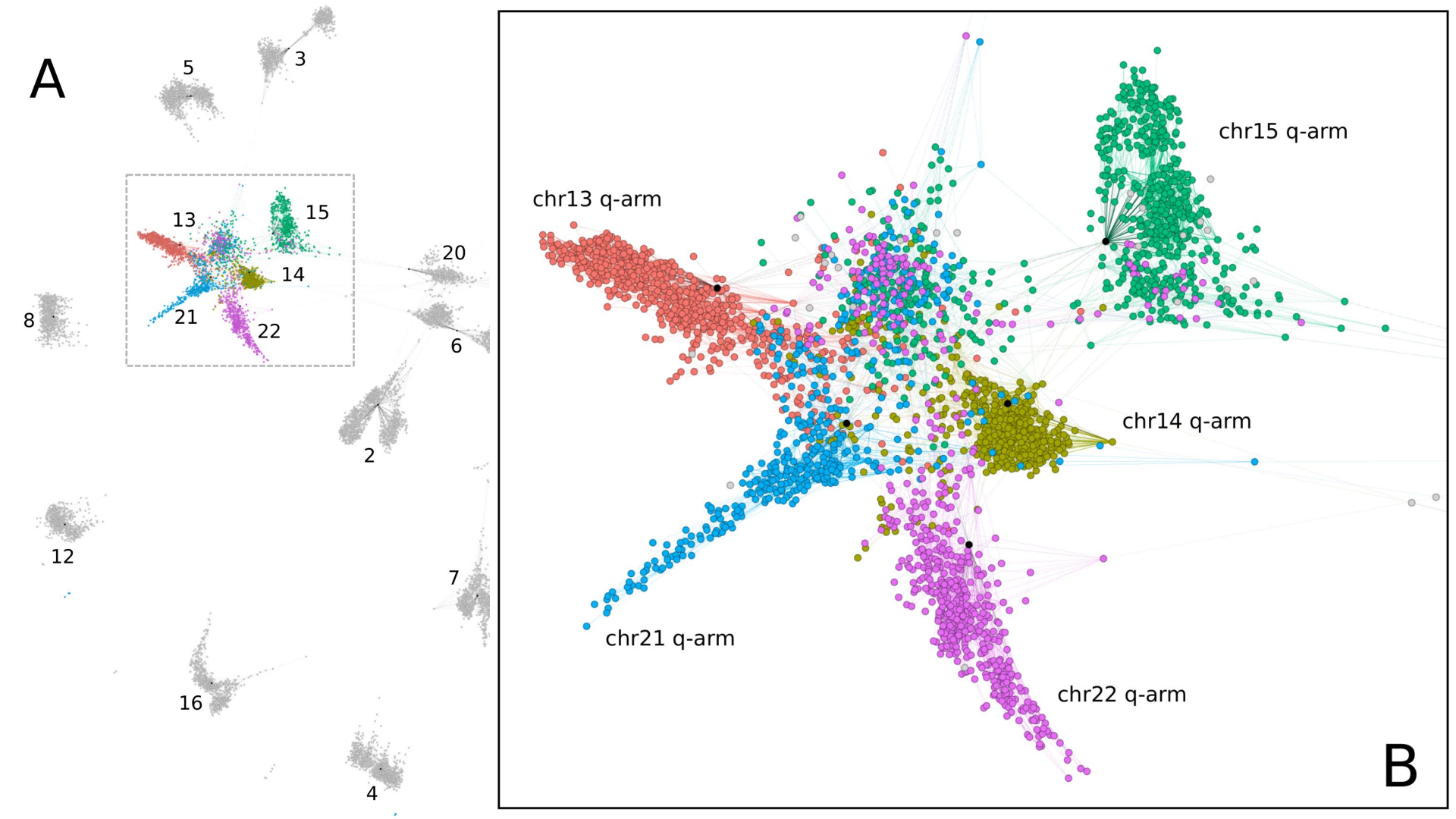
From HPRC main paper.

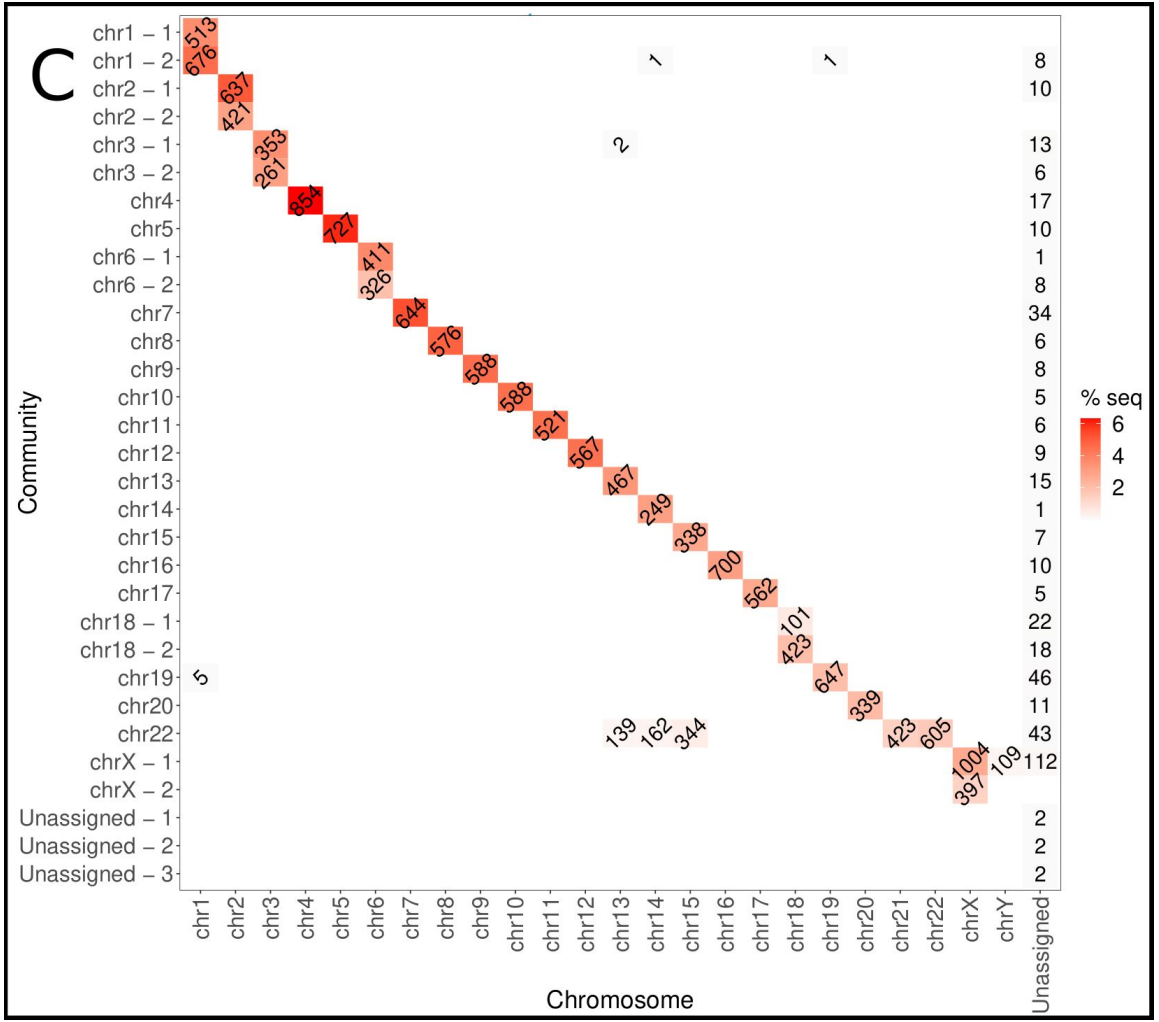
* With the exception of assembly errors in one haplotype (HG02080 paternal).



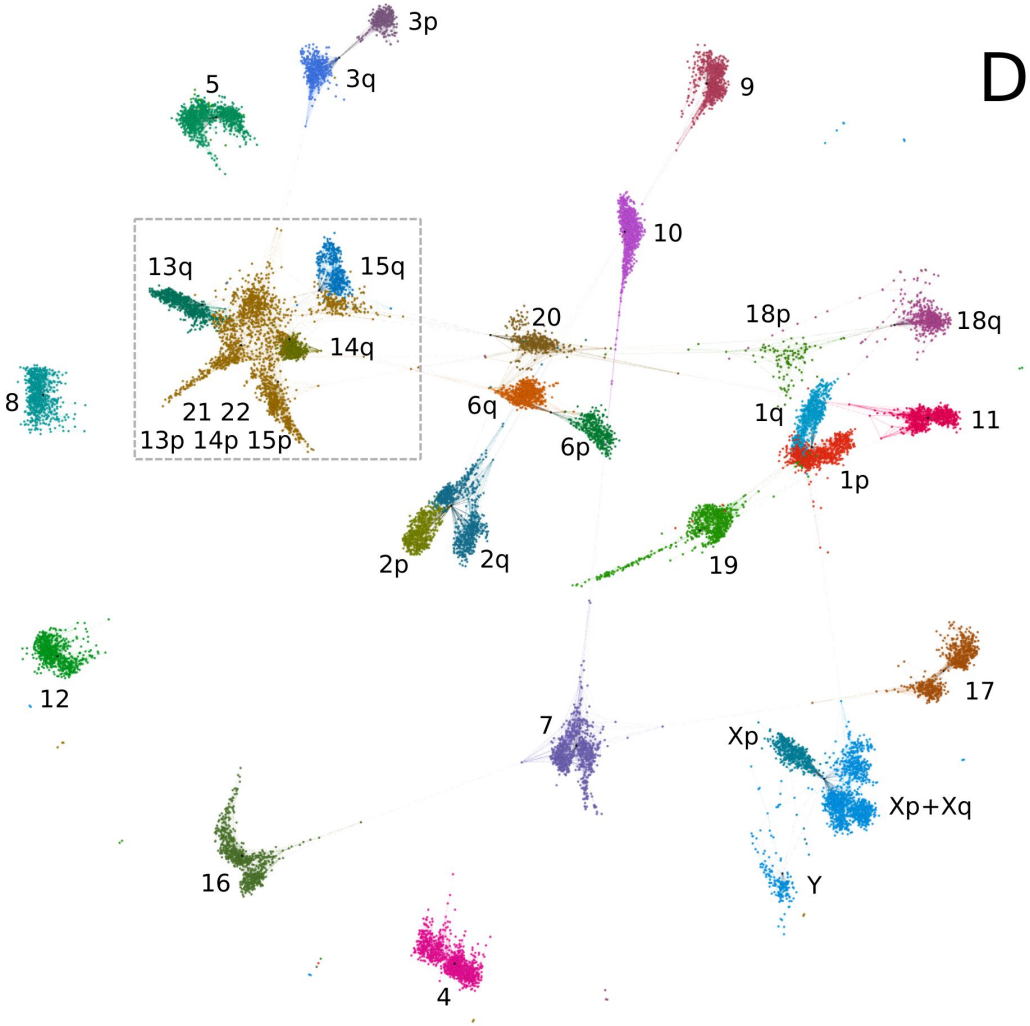
Chromosome communities in the HPRC

An all-vs-all mapping graph for the HPRC contigs >1mbp.


A
5
3
13
15
20
14
8
21
22
6
2
12
7
16
4
B
chr15 q-arm
chr13 q-arm
chr14 q-arm
chr21 q-arm
chr22 q-arm



Leiden community detection



Leiden community detection

Labeling the layout with community assignments.

We decided to take a closer look, focusing on the best assemblies in these regions.

Workflow

HPRC assemblies

https://github.com/human-pangenomics/HPP_Year1_Assemblies

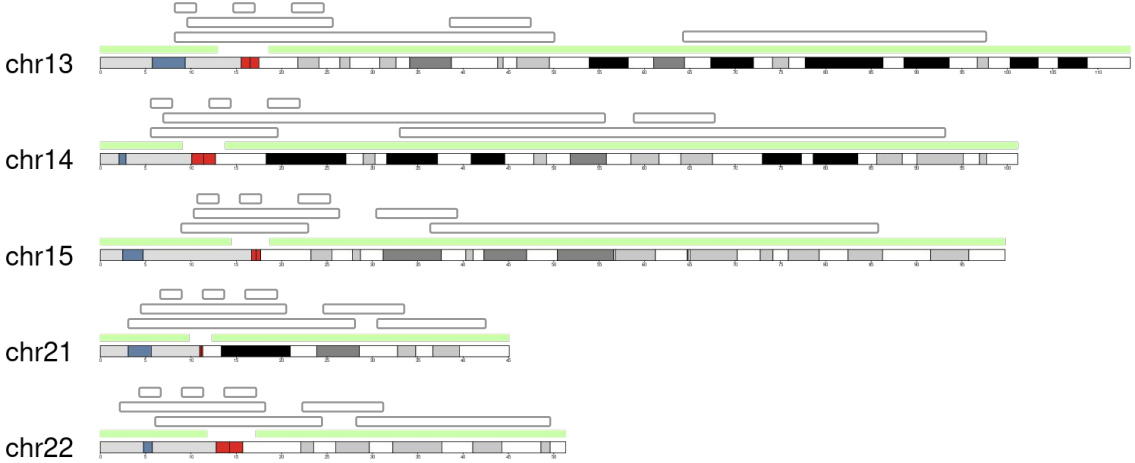
We decided to take a closer look, focusing on the best assemblies in these regions.

Workflow

HPRC assemblies

https://github.com/human-pangenomics/HPP_Year1_Assemblies

Mapping against
the whole CHM13

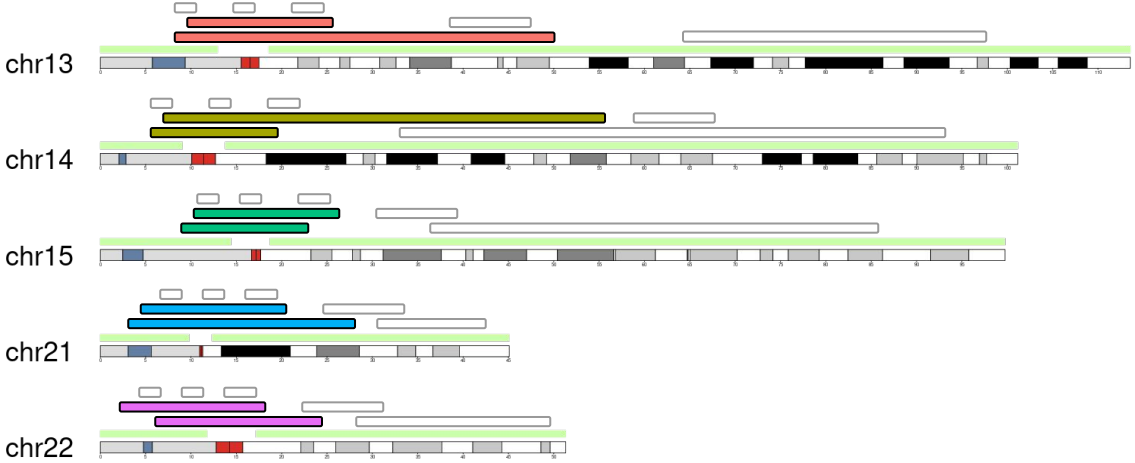


Workflow

HPRC assemblies

https://github.com/human-pangenomics/HPP_Year1_Assemblies

Mapping against the whole CHM13



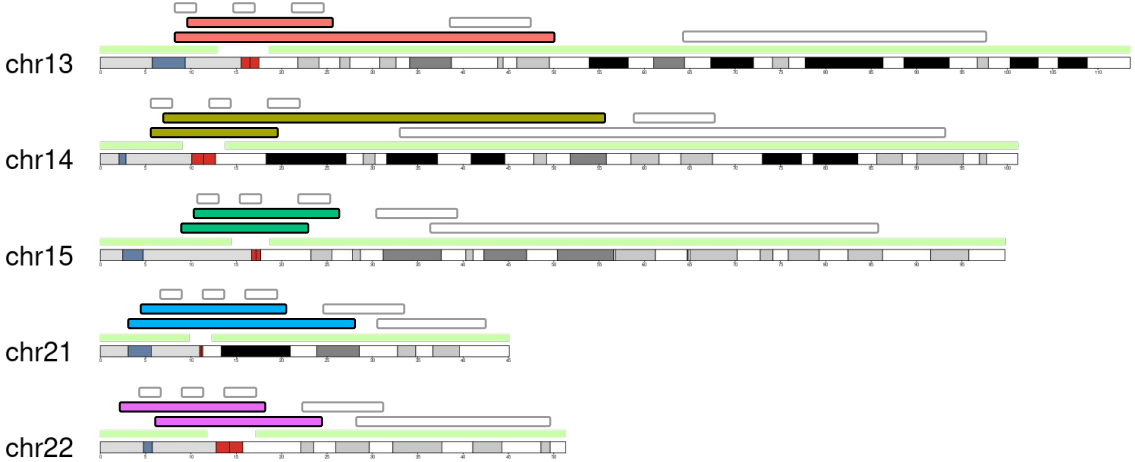
Acrocentric contigs covering (+/- 1Mbp) both the p and q arms (pq-contigs)

Workflow

HPRC assemblies

https://github.com/human-pangenomics/HPP_Year1_Assemblies

Mapping against the whole CHM13

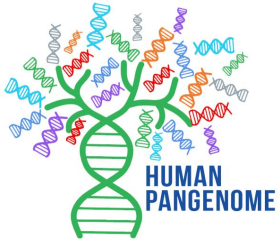


Acrocentric contigs covering (+/- 1Mbp) both the p and q arms (pq-contigs)

+ HG002 contigs >= 300kbps which map to acrocentrics

PanGenome Graph Builder (PGGB)

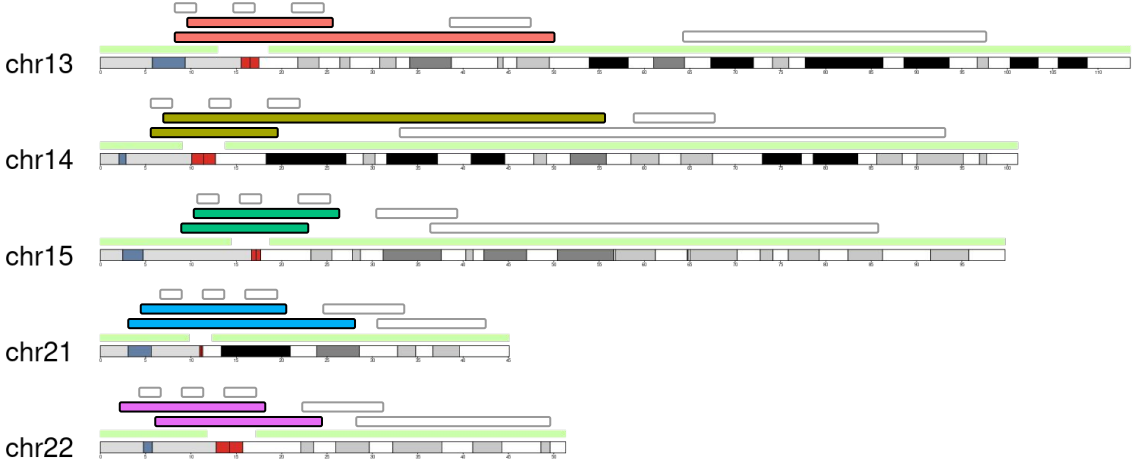
Workflow



HPRC assemblies

https://github.com/human-pangenomics/HPP_Year1_Assemblies

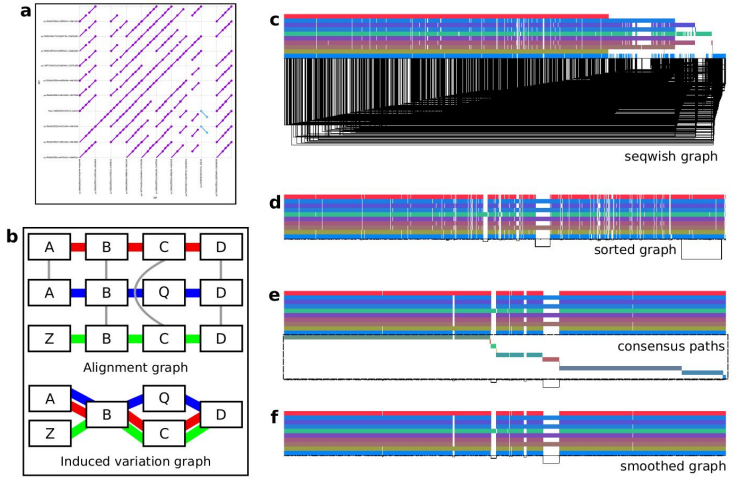
Mapping against
the whole CHM13



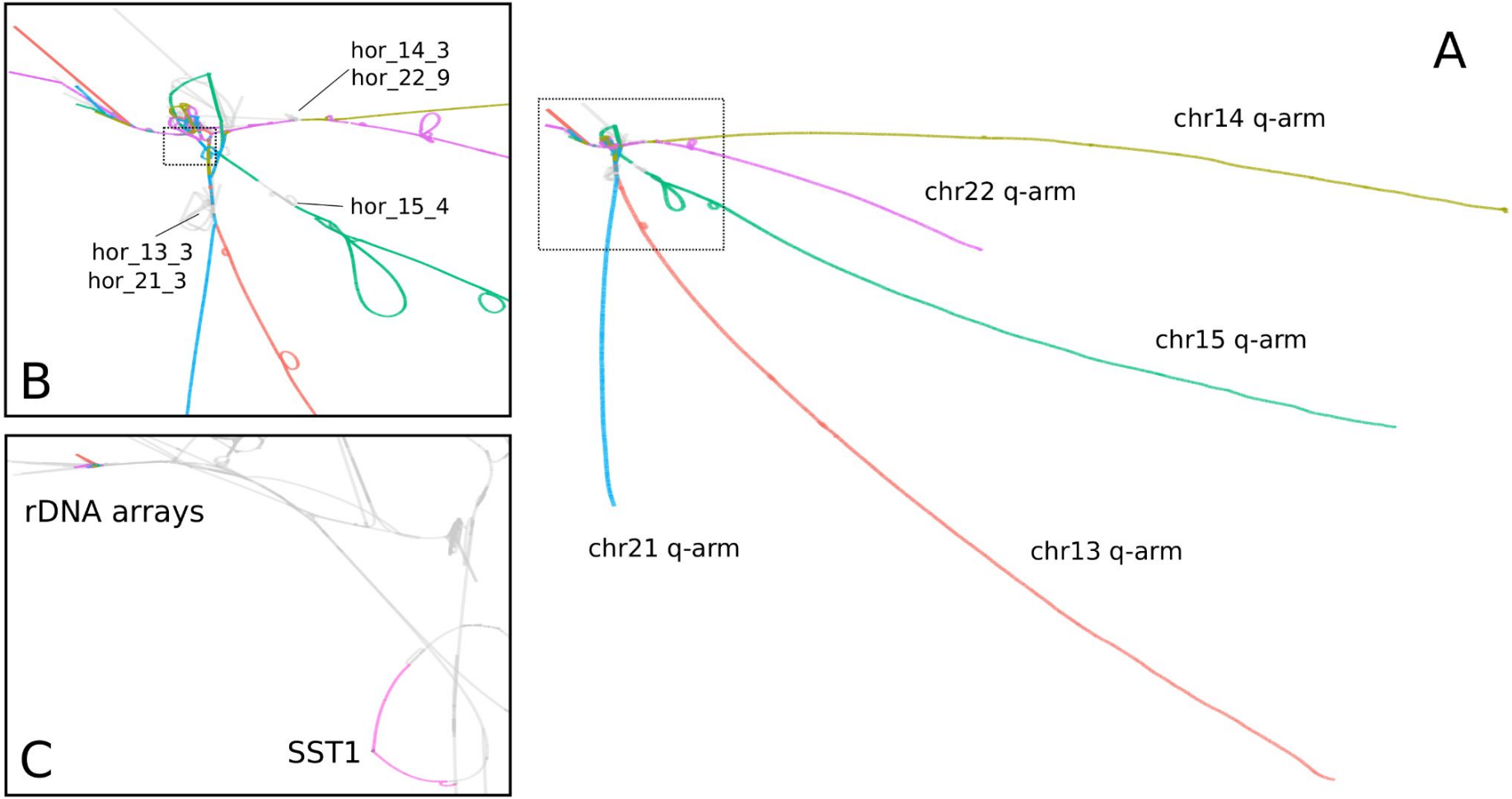
Acrocentric contigs covering (+/- 1Mbp) both the p and q arms (pq-contigs)

+ HG002 contigs >= 300kbps
which map to acrocentrics

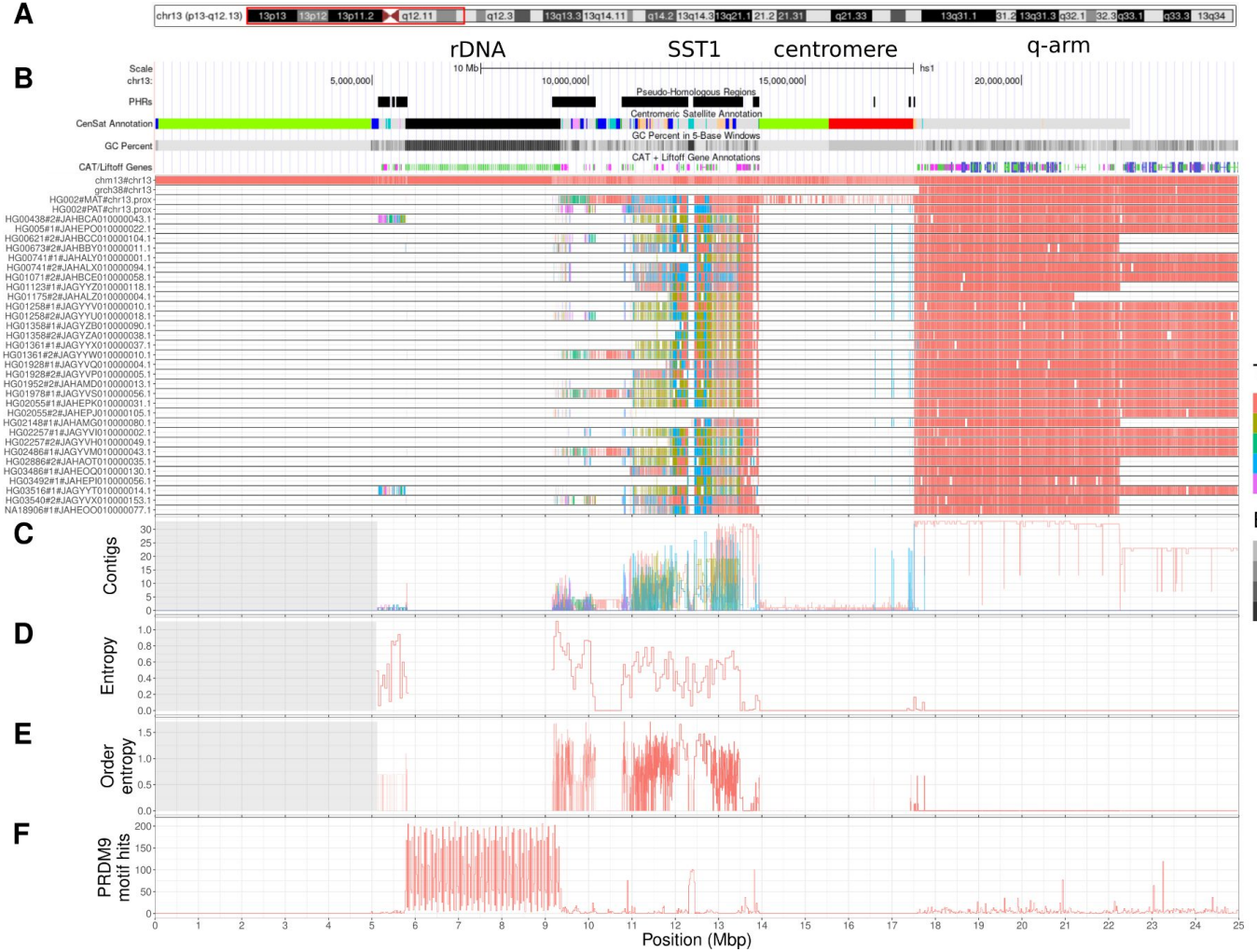
PanGenome Graph
Builder (PGGB)



https://github.com/pangenome/pggb


A
chr14 q-arm
chr22 q-arm
chr15 q-arm
chr21 q-arm
chr13 q-arm
B
hor_14_3
hor_22_9
hor_15_4
hor_13_3
hor_21_3
C
rDNA arrays
SST1

Untangling the pangenome graph



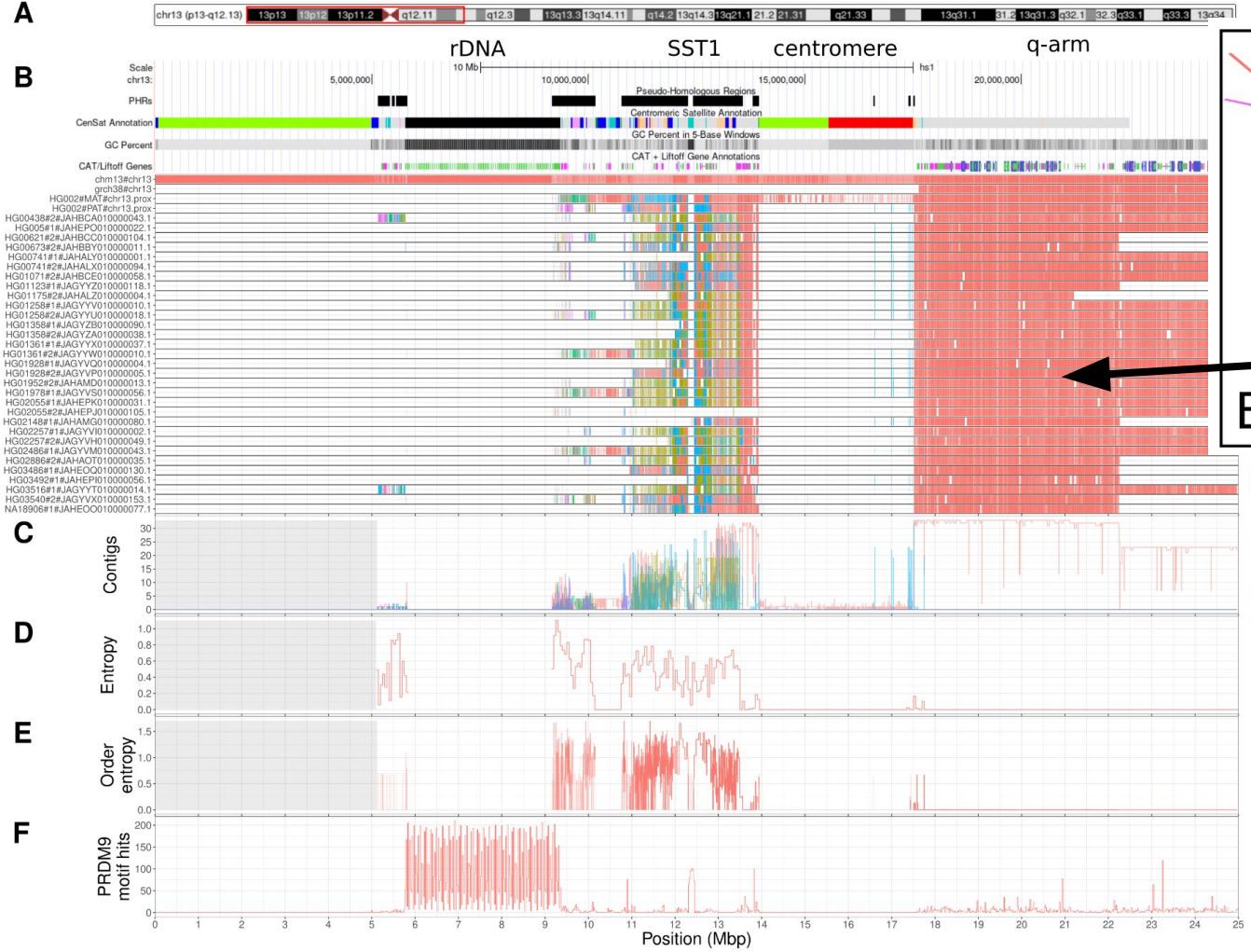
Target

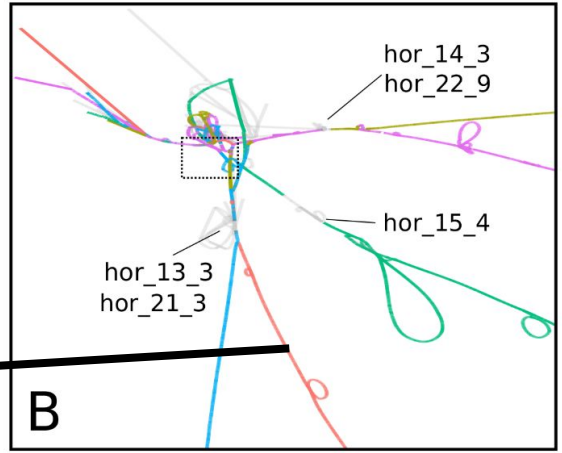
- chm13#chr13
- chm13#chr14
- chm13#chr15
- chm13#chr21
- chm13#chr22

Estimated identity

- 0.925
- 0.950
- 0.975
- 1.000

We look into the graph from the perspective of **chromosome 13**. Full information from pangenome plus reference annotations.

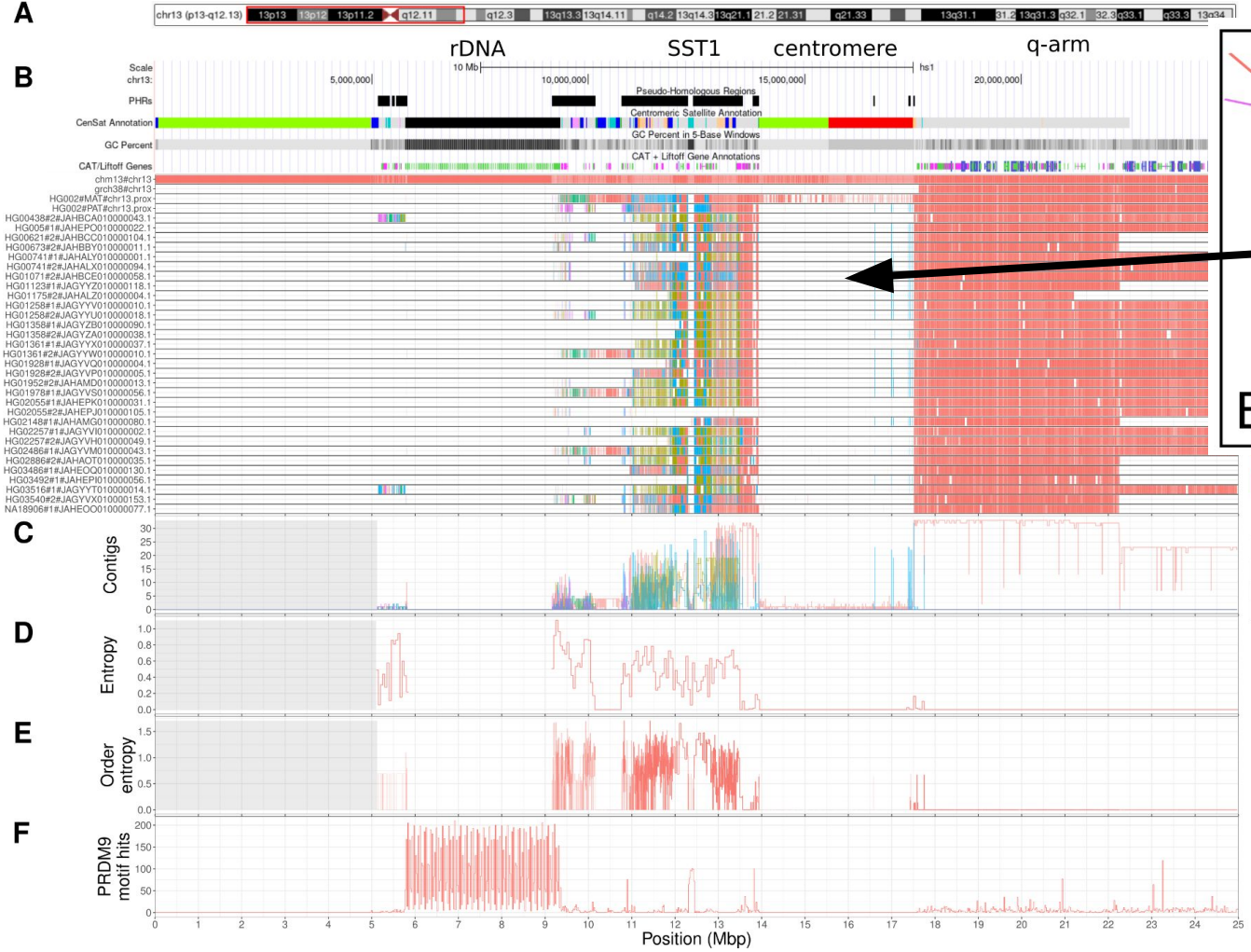


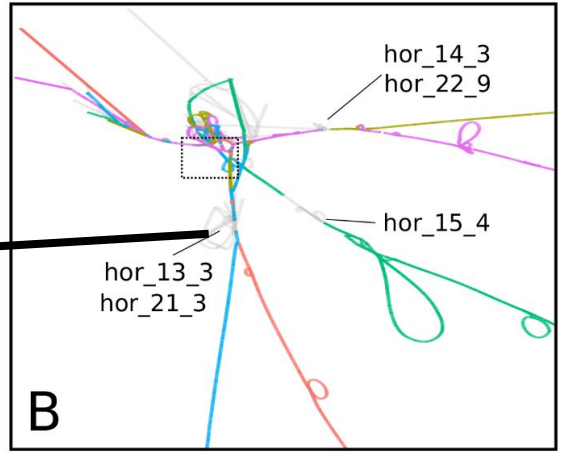


chm13#chr21
chm13#chr22

Estimated identity
0.925
0.950
0.975
1.000

We look into the graph from the perspective of **chromosome 13**. Full information from pangenome plus reference annotations.

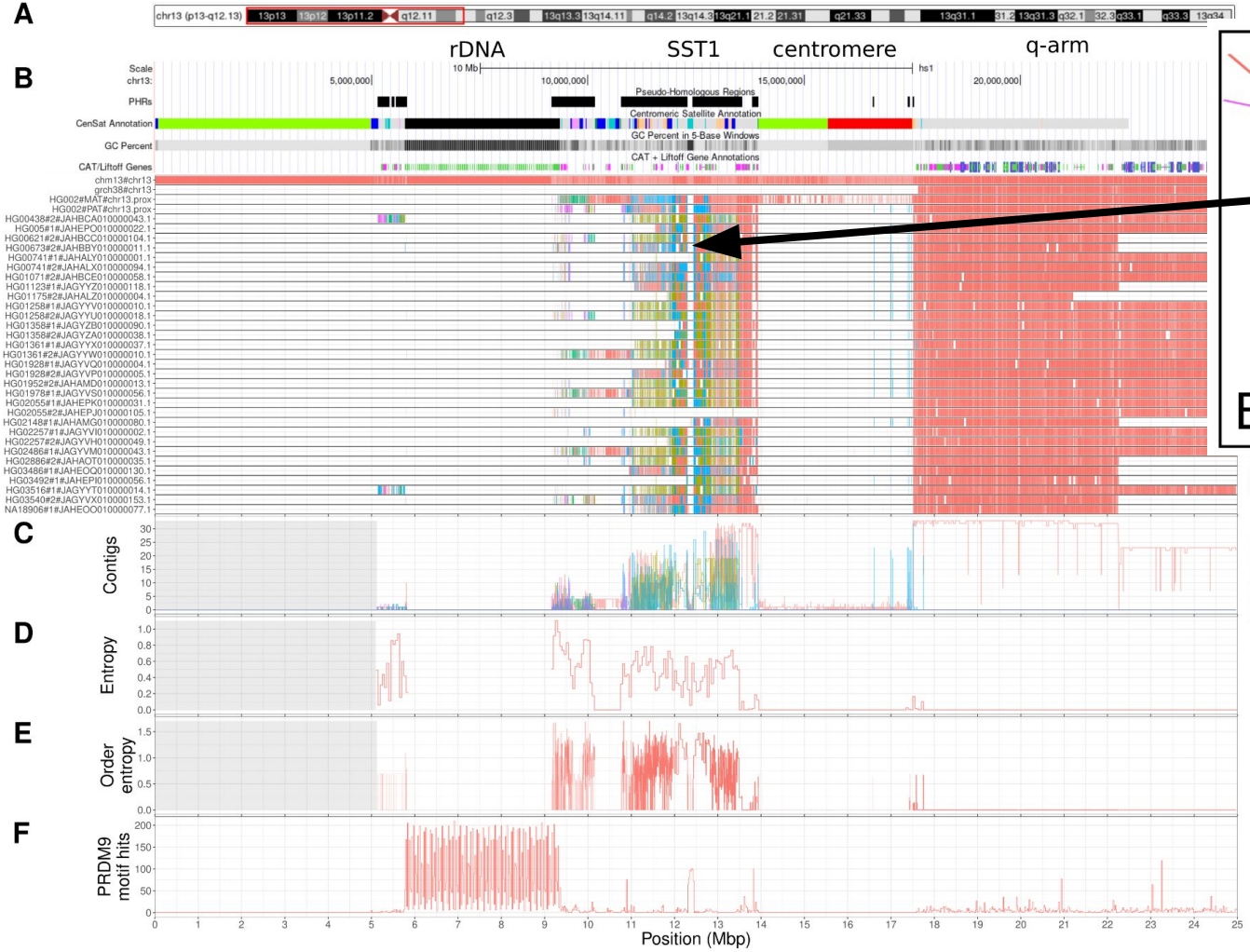


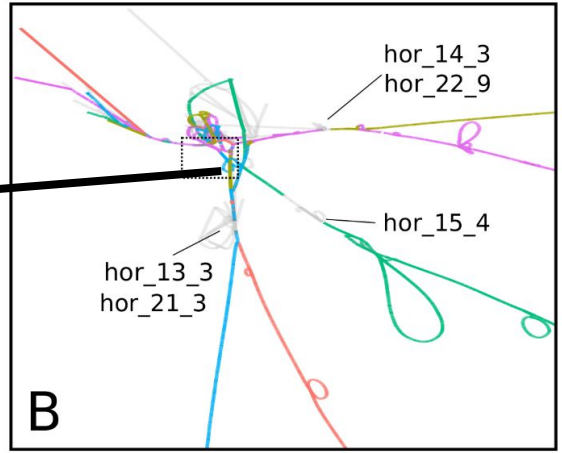


chm13#chr21
chm13#chr22

Estimated identity
0.925
0.950
0.975
1.000

We look into the graph from the perspective of **chromosome 13**. Full information from pangenome plus reference annotations.

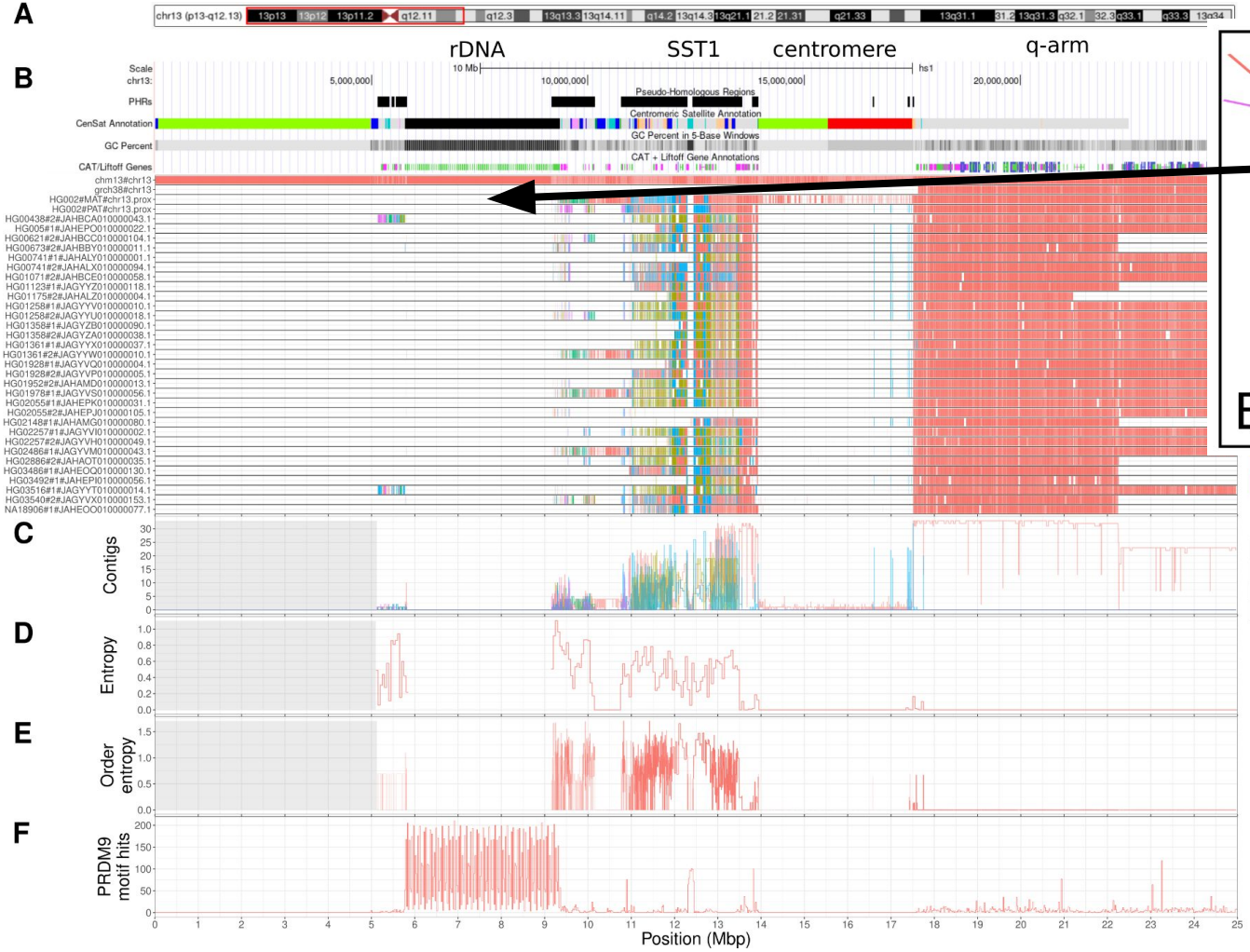


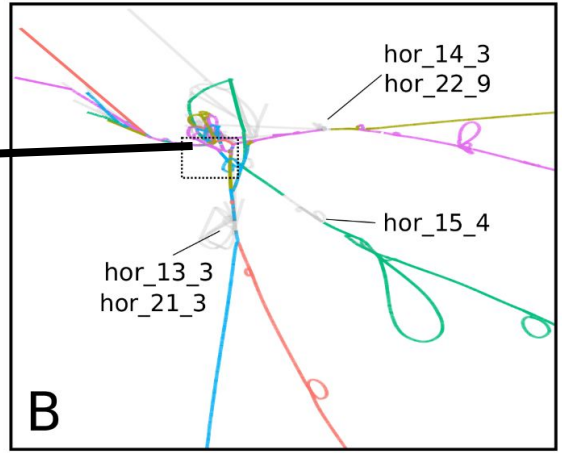


chm13#chr21
chm13#chr22

Estimated identity
0.925
0.950
0.975
1.000

We look into the graph from the perspective of **chromosome 13**. Full information from pangenome plus reference annotations.

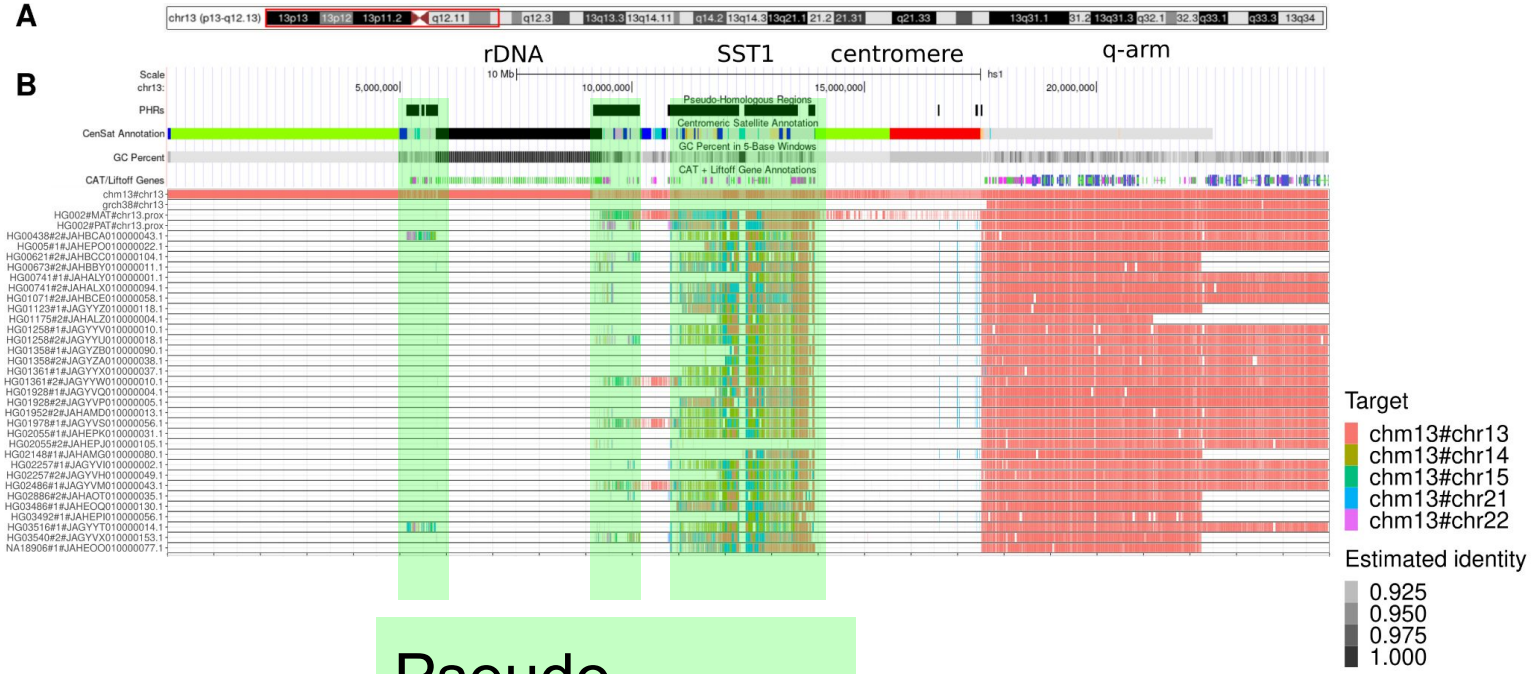




chm13#chr21
chm13#chr22

Estimated identity
0.925
0.950
0.975
1.000

We look into the graph from the perspective of **chromosome 13**. Full information from pangenome plus reference annotations.



Pseudo
Homologous
Regions

chromosome 14

Target
- chm13#chr13
- chm13#chr14
- chm13#chr15
- chm13#chr21
- chm13#chr22

Estimated identity
- 0.925
- 0.950
- 0.975
- 1.000

chromosome 14

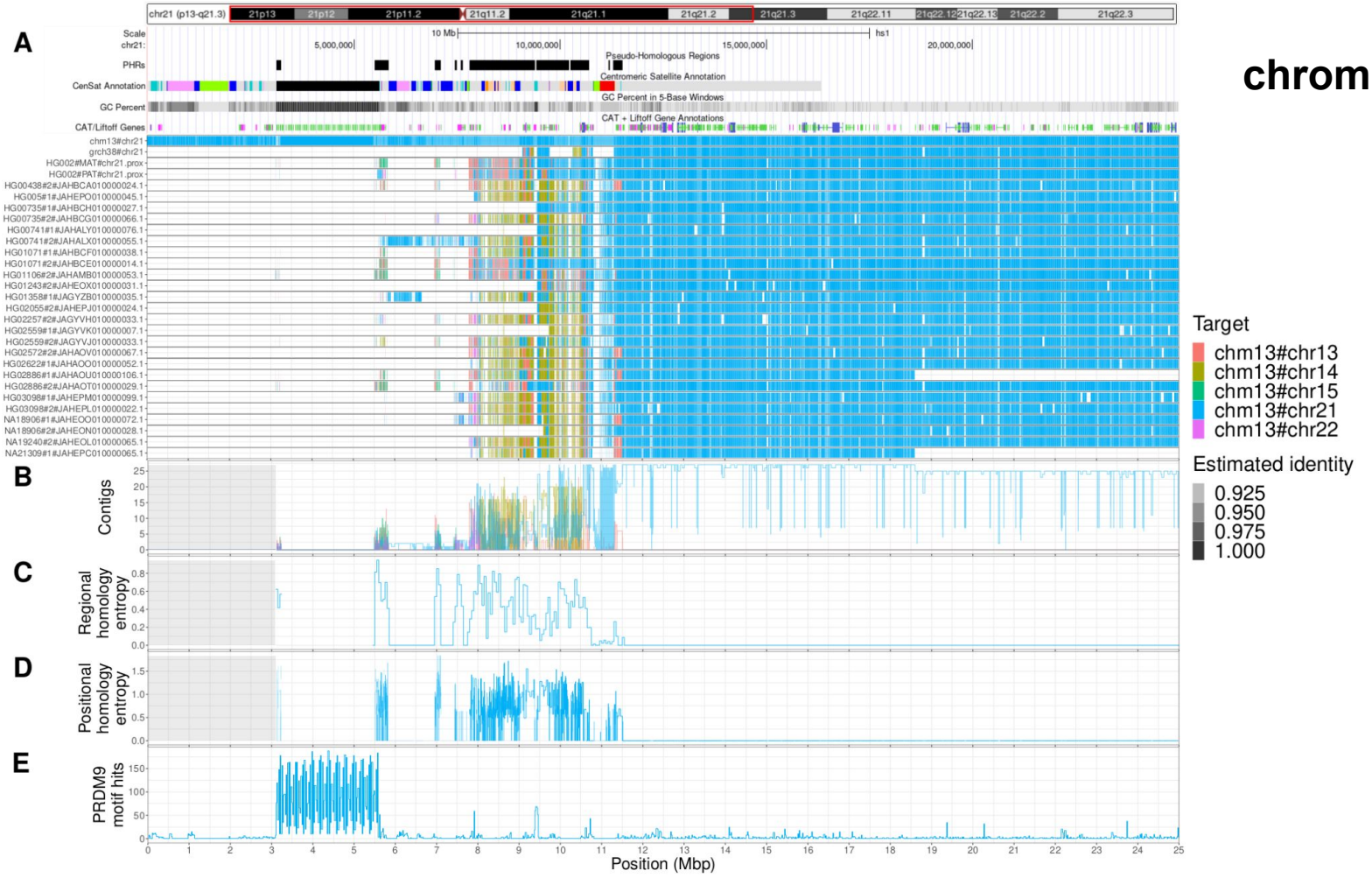
Target
- chm13#chr13
- chm13#chr14
- chm13#chr15
- chm13#chr21
- chm13#chr22

Estimated identity
- 0.925
- 0.950
- 0.975
- 1.000

chromosome 15

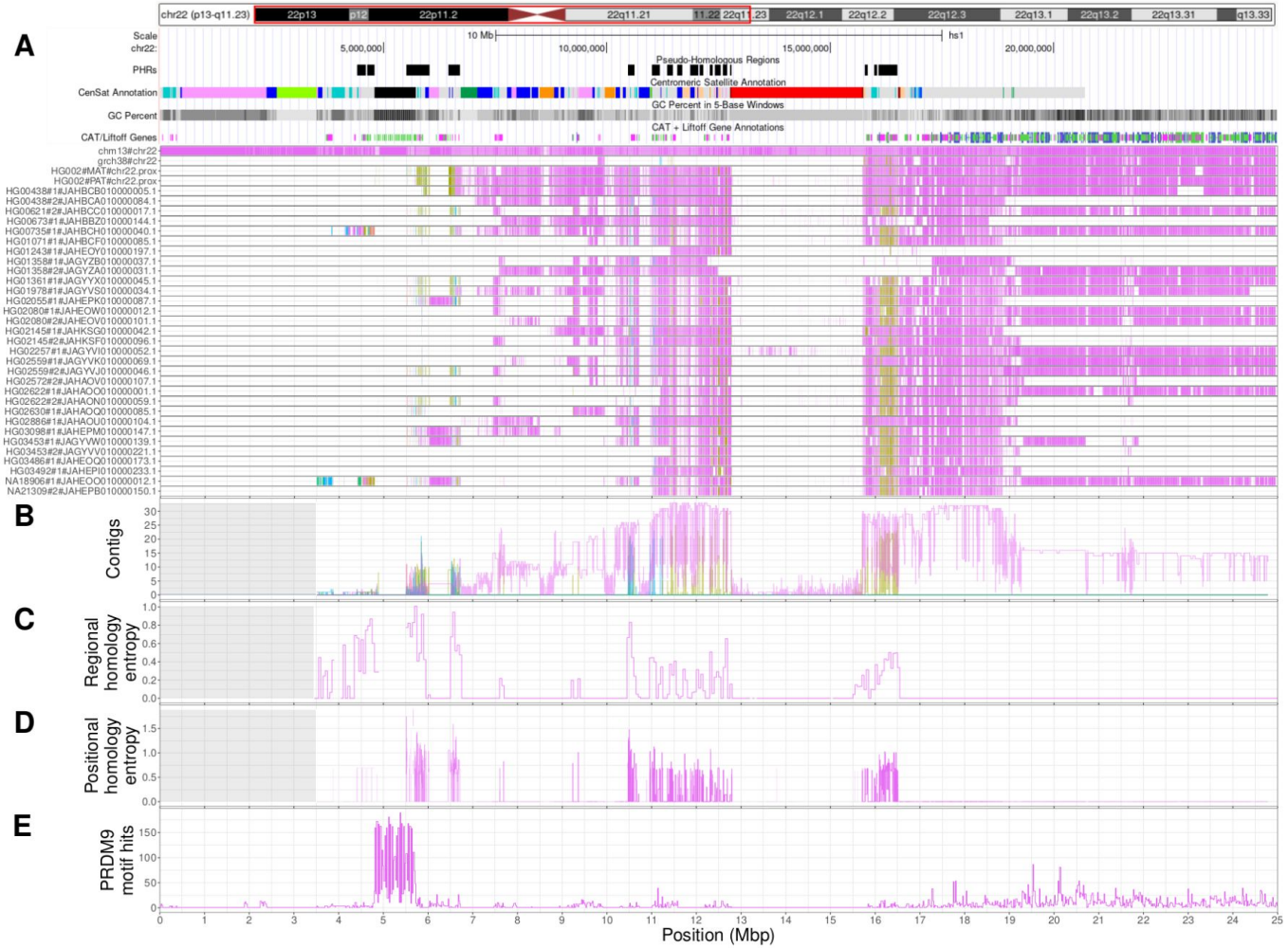
chromosome 15

chromosome 21



chromosome 21

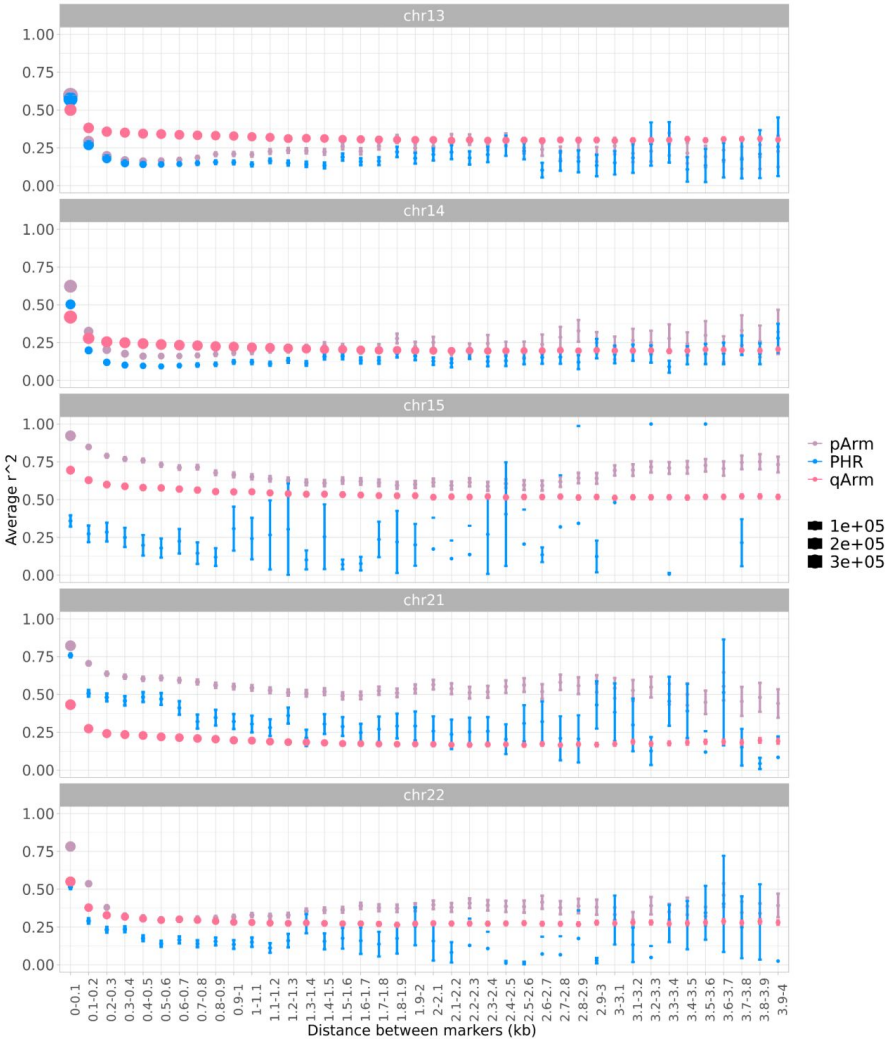
chromosome 22



Target
- chm13#chr13
- chm13#chr14
- chm13#chr15
- chm13#chr21
- chm13#chr22

Estimated identity
- 0.925
- 0.950
- 0.975
- 1.000

chromosome 22



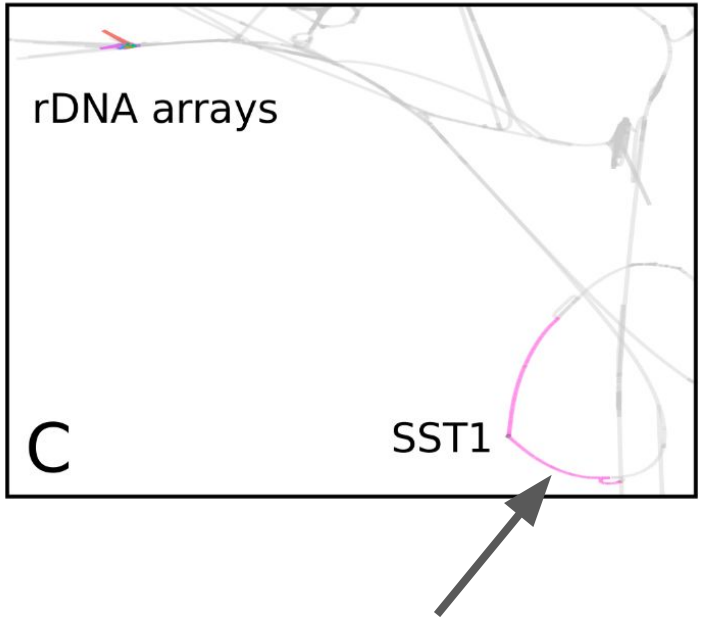
LD decays faster in pseudo-homologous regions than elsewhere in the p-arms or in the q-arms.

This pattern is consistent with higher recombination rates and/or effective population size in these regions.

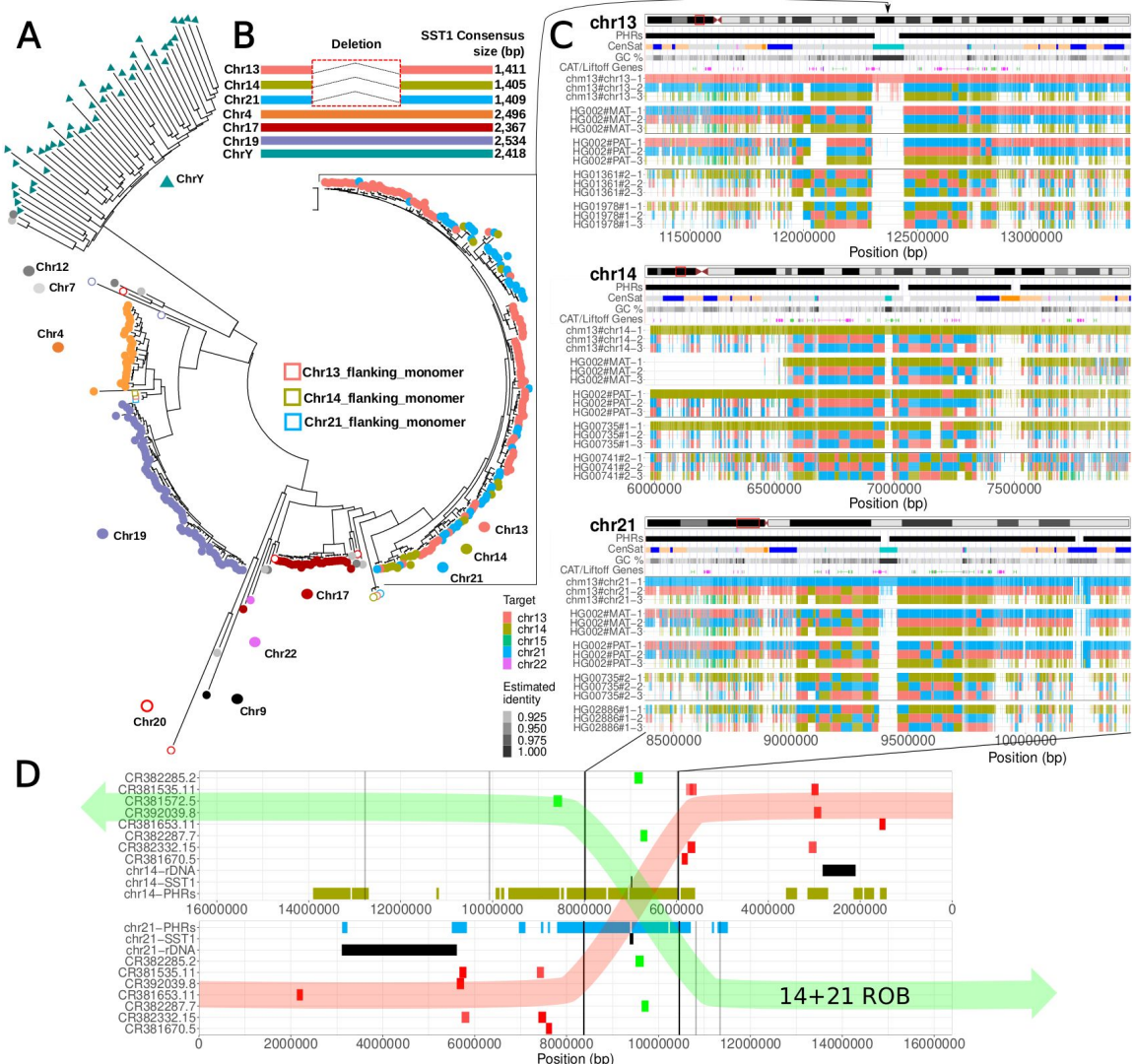
Silvia Buonaiuto, Vincenza Colonna

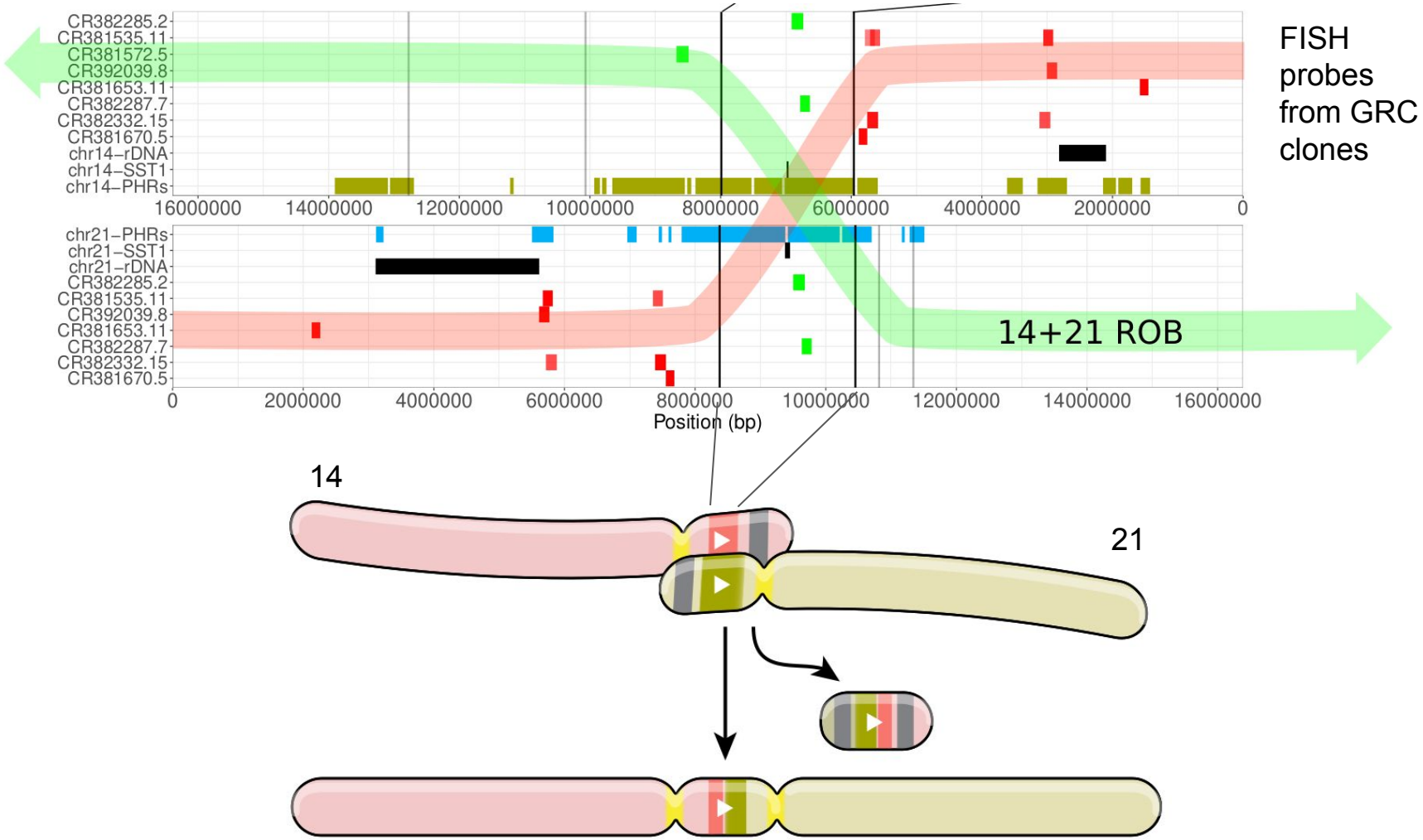
chromosome 13/14/21

Pseudo-homologous regions



The SST1-linked PHR is the site of the most intense signals of recombination between heterologous chromosomes.




CR382285.2
CR381535.11
CR381572.5
CR392039.8
CR381653.11
CR382287.7
CR382332.15
CR381670.5
chr14–rDNA
chr14–SST1
chr14–PHRs
16000000
14000000
12000000
10000000
8000000
6000000
4000000
2000000
0
chr21–PHRs
chr21–SST1
chr21–rDNA
CR382285.2
CR381535.11
CR392039.8
CR381653.11
CR382287.7
CR382332.15
CR381670.5
0
2000000
4000000
6000000
8000000
10000000
12000000
14000000
16000000
Position (bp)
FISH probes from GRC clones
14+21 ROB
14
21

Recombination between heterologous chromosomes

The high level of homology of the acrocentric chromosomes is likely due to **recombination between heterologous chromosomes!**

High-quality *de novo* assemblies and pangenomic approaches thus shed light on the most difficult regions of the human genomes.

This answers questions that arose in the early era of cytogenetics, ~50 years ago.

Volume 16 Number 4 1988 Nucleic Acids Research

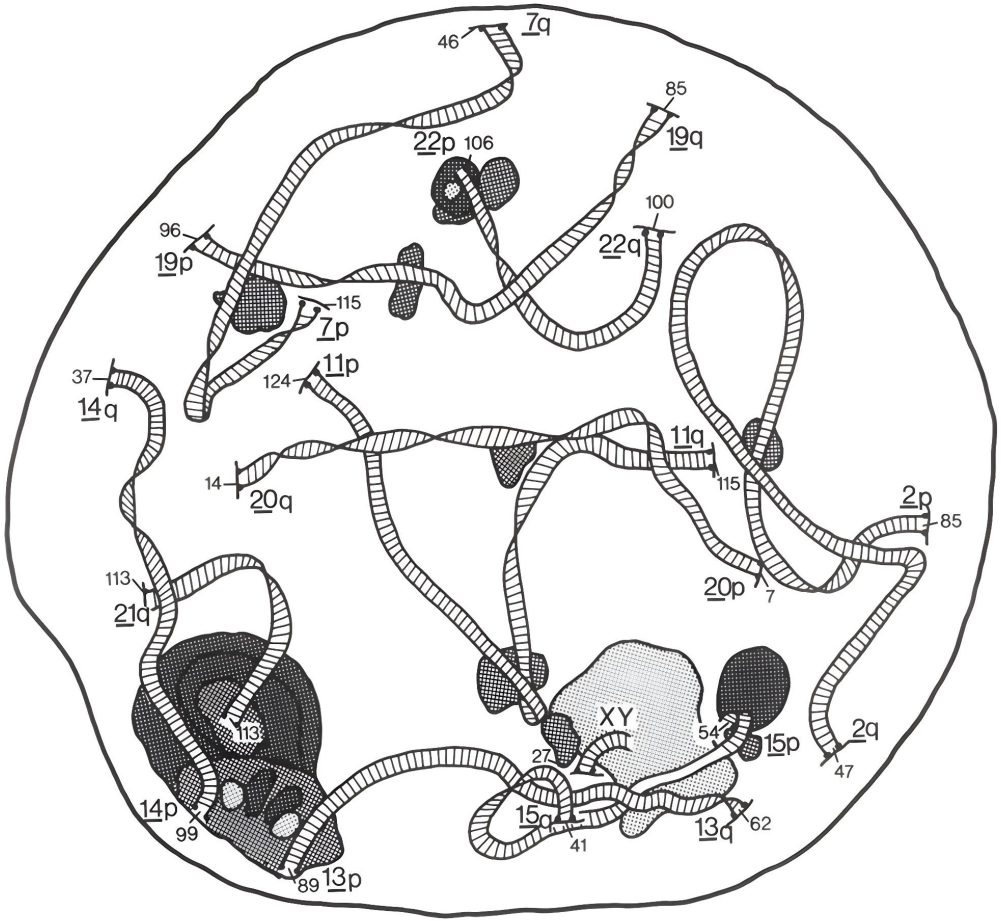
Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations

K.H.Choo*, B.Vissel, R.Brown, R.G.Filby and E.Earle

ABSTRACT

We report a new subfamily of alpha satellite DNA (pTRA-2) which is found on all the human acrocentric chromosomes. The alphoid nature of the cloned DNA was established by partial sequencing. Southern analysis of restriction enzyme-digested DNA fragments from mouse/human hybrid cells containing only human chromosome 21 showed that the predominant higher-order repeating unit for pTRA-2 is a 3.9 kb structure. Analysis of a "consensus" in situ hybridisation profile derived from 13 normal individuals revealed the localisation of 73% of all centromeric autoradiographic grains over the five acrocentric chromosomes, with the following distribution: 20.4%, 21.5%, 17.1%, 7.3% and 6.5% on chromosomes 13, 14, 21, 15 and 22 respectively. An average of 1.4% of grains was found on the centromere of each of the remaining 19 nonacrocentric chromosomes. These results indicate the presence of a common subfamily of alpha satellite DNA on the five acrocentric chromosomes and suggest an evolutionary process consistent with recombination exchange of sequences between the nonhomologues. The results further suggests that such exchanges are more selective for chromosomes 13, 14 and 21 than for chromosomes 15 and 22. The possible role of centromeric alpha satellite DNA in the aetiology of 13q14q and 14q21q Robertsonian translocations involving the common and nonrandom association of chromosomes 13 and 14, and 14 and 21 is discussed.

Chroo et al., 1988.



HUMAN MEIOSIS I. THE HUMAN PACHYTENE KARYOTYPE ANALYZED BY THREE DIMENSIONAL RECONSTRUCTION OF THE SYNAPTONEMAL COMPLEX

by

PREBEN BACH HOLM

and

SØREN WILKEN RASMUSSEN

Department of Physiology, Carlsberg Laboratory
Gamle Carlsberg Vej 10, DK-2500 Copenhagen, Valby

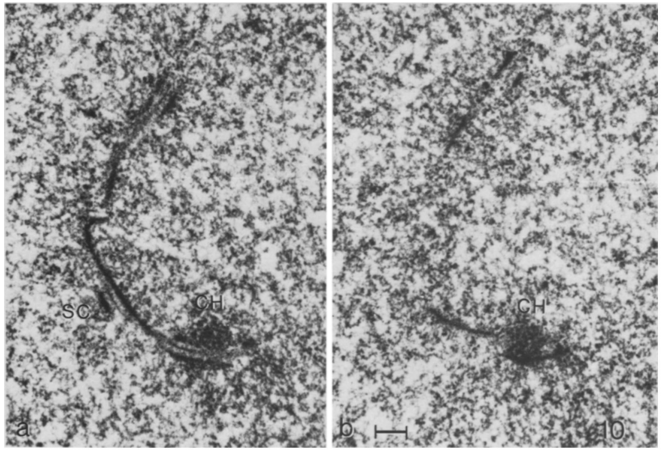


Figure 10. Two consecutive sections through the centromeric heterochromatin of a bivalent at early pachytene. The synaptonemal complex (SC) passes unaltered through the centromeric heterochromatin (CH). (Bar = 0.2 μm)

https://doi.org/10.1007/BF02910454

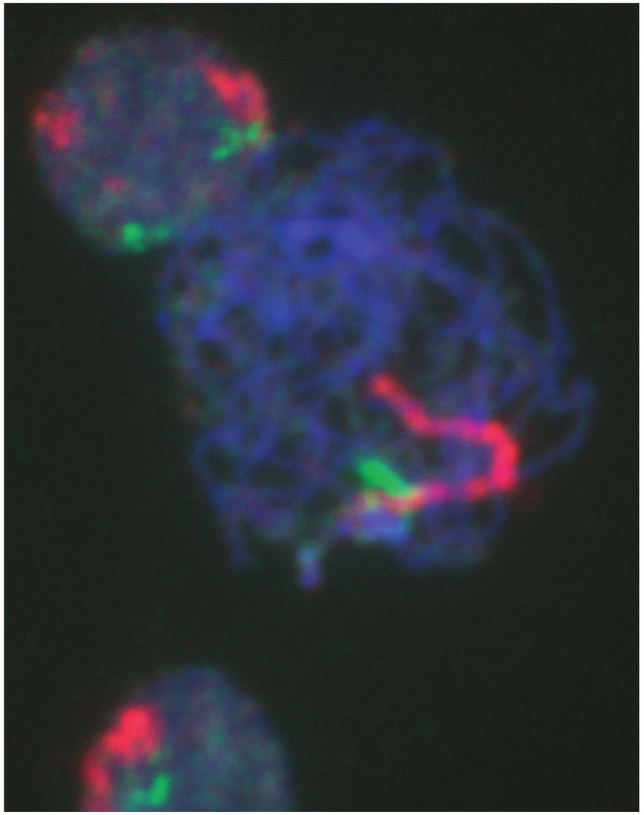


Figure The pachytene nucleus with 2 hybridization signals is oriented in a linear fashion. The *red signal* represents chromosome 14, and the *green signal* represents chromosome 21.

https://doi.org/10.1016/j.ajog.2004.02.062

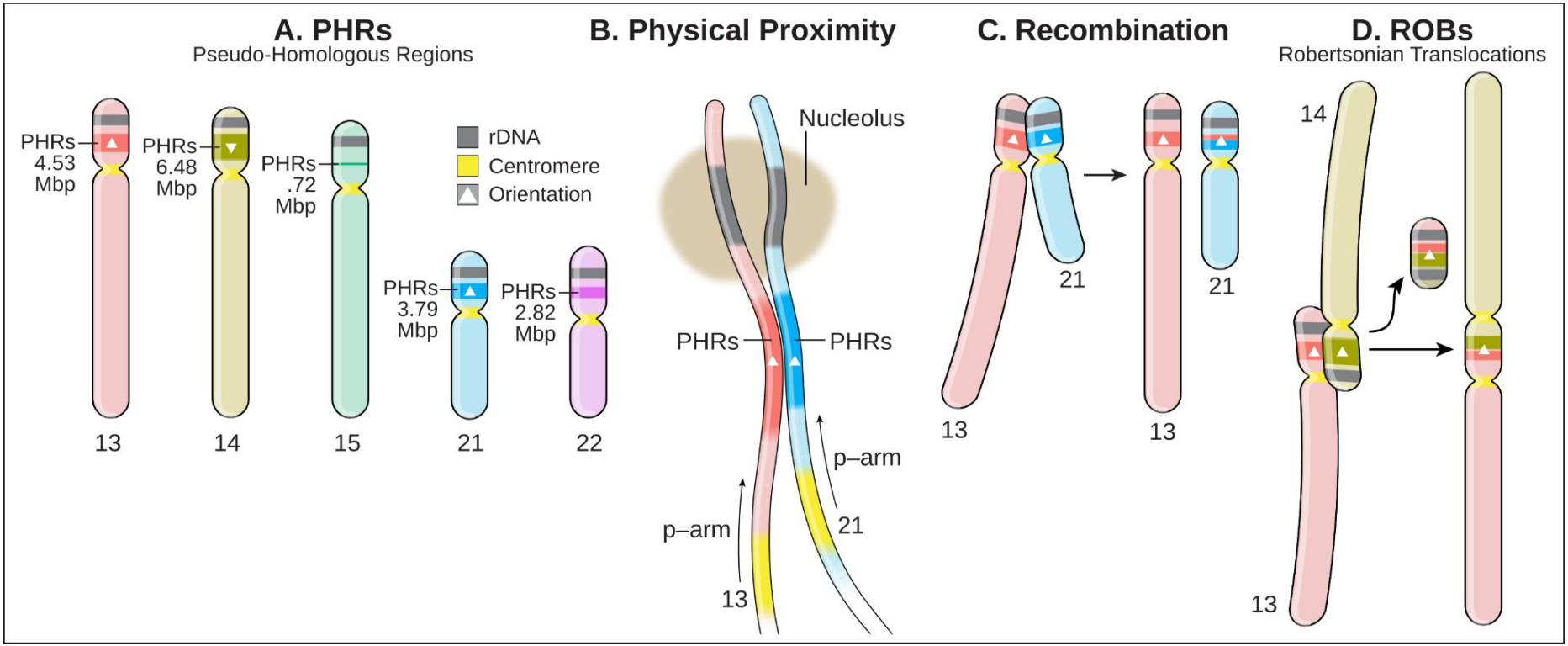
TRANSACTIONS OF THE 70TH ANNUAL MEETING OF THE PACIFIC COAST OBSTETRICIANS AND GYNECOLOGICAL SOCIETY | VOLUME 190, ISSUE 6, P1781-1785, JUNE 01, 2004

FISHing for acrocentric associations between chromosomes 14 and 21 in human oogenesis

Edith Y Cheng, MD • Theresa Naluai-Cecchini, BA

observed 300 cells to see a single putative 14:21 synapse!

Pseudo-homologous regions (PHRs)



Article

Recombination between heterologous human acrocentric chromosomes

https://doi.org/10.1038/s41586-023-05976-y

Received: 15 August 2022

Accepted: 17 March 2023

Published online: 10 May 2023

Open access

Check for updates

Andrea Guarracino[1,2], Silvia Buonaiuto[3], Leonardo Gomes de Lima[4], Tamara Potapova[4], Arang Rhie[5], Sergey Koren[5], Boris Rubinstein[4], Christian Fischer[1], Human Pangenome Reference Consortium*, Jennifer L. Gerton[4], Adam M. Phillippy[5], Vincenza Colonna[1,3] & Erik Garrison[1] ✉

The short arms of the human acrocentric chromosomes 13, 14, 15, 21 and 22 (SAACs) share large homologous regions, including ribosomal DNA repeats and extended segmental duplications[1,2]. Although the resolution of these regions in the first complete assembly of a human genome—the Telomere-to-Telomere Consortium's CHM13 assembly (T2T-CHM13)—provided a model of their homology[3], it remained unclear whether these patterns were ancestral or maintained by ongoing recombination exchange. Here we show that acrocentric chromosomes contain pseudo-homologous regions (PHRs) indicative of recombination between non-homologous sequences. Utilizing an all-to-all comparison of the human pangenome from the Human Pangenome Reference Consortium[4] (HPRC), we find that contigs from all of the SAACs form a community. A variation graph[5] constructed from centromere-spanning acrocentric contigs indicates the presence of regions in which most contigs appear nearly identical between heterologous acrocentric chromosomes in T2T-CHM13. Except on chromosome 15, we observe faster decay of linkage disequilibrium in the pseudo-homologous regions than in the corresponding short and long arms, indicating higher rates of recombination[6,7]. The pseudo-homologous regions include sequences that have previously been shown to lie at the breakpoint of Robertsonian translocations[8], and their arrangement is compatible with crossover in inverted duplications on chromosomes 13, 14 and 21. The ubiquity of signals of recombination between heterologous acrocentric chromosomes seen in the HPRC draft pangenome suggests that these shared sequences form the basis for recurrent Robertsonian translocations, providing sequence and population-based confirmation of hypotheses first developed from cytogenetic studies 50 years ago[9].

Thanks!

to you, and…

Andrea Guarracino (pggb, wfmash, seqwish, odgi, chromosome communities)
Simon Heumos (pggb, odgi)
Flavia Villani (pggb, applications to mouse, popgen)
Njagi Mwaniki (wfmash, WFA applications)
Santiago Marco-Sola (WFA, wfmash)
Pjotr Prins (vcflib, vcfwave)
Richard Durbin (PhD guidance)
Nicole Soranzo (support)
Benedict Paten (vgteam)
Hao Chen (rat, mouse)
Zhigui Bao (plant applications)
Lorenzo Tattini (yeast pangenomes)
Enza Colonna (applications to mouse, popgen)
Nadia Pisanti (algorithms)
Luca Pinello (applications)
Peter Sudmant (primate pangenomes)
Robert Williams (guidance)

HPRC pangenomes working group and many others

funders:

NLnet
NSF
NIH (NIDA)

Amylase project!

Alessandro Raveane (Human Technopole)
Davide Bolognini (Human Technopole)
Peter Sudmant (Berkeley)
Joana Rocha (Berkeley)
Andrea Guarracino (UTHSC)
Alma Halgren (Berkeley)
Jason Chin (GeneDX)
Nicholas Lou (Berkeley)

TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOME

THE GENOME INSTITUTE at Washington University

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Genome Sciences UNIVERSITY OF WASHINGTON

The Rockefeller University 1901

NCBI National Center for Biotechnology Information

CORIELL INSTITUTE FOR MEDICAL RESEARCH DECODING THE GENOME

EMBL-EBI

HARVARD MEDICAL SCHOOL

Icahn School of Medicine at Mount Sinai

sanger institute

Yale University

We would like to acknowledge the National Genome Research Institute (NHGRI) for funding the following grants which are in support of creating the human pangenome reference: 1U41HG010972, 1U01HG010971, 1U01HG010961, 1U01HG010973, 1U01HG010963, and the Human Pangenome Reference Consortium (https://humanpangenome.org/)

PacBio illumina Google Health aws

NIH National Human Genome Research Institute

Cantata Bio

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

NIST

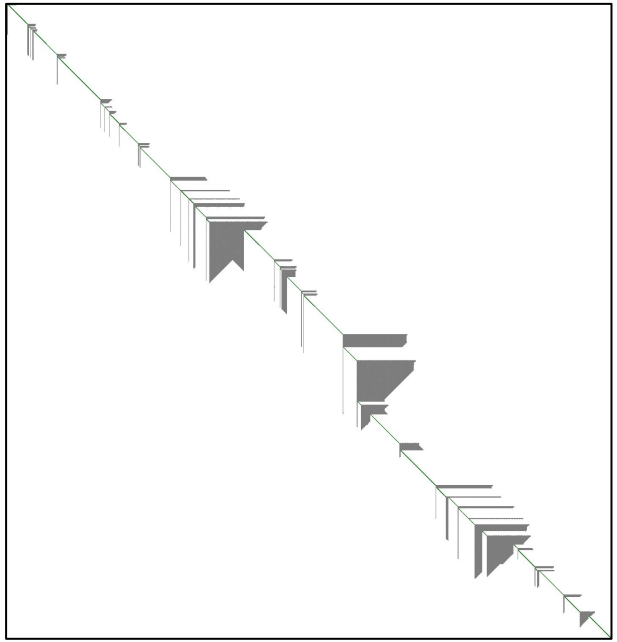
Practical!

Let’s build some pangenome variation graphs with **pggb**!

First: a deeper dive into how the method works.

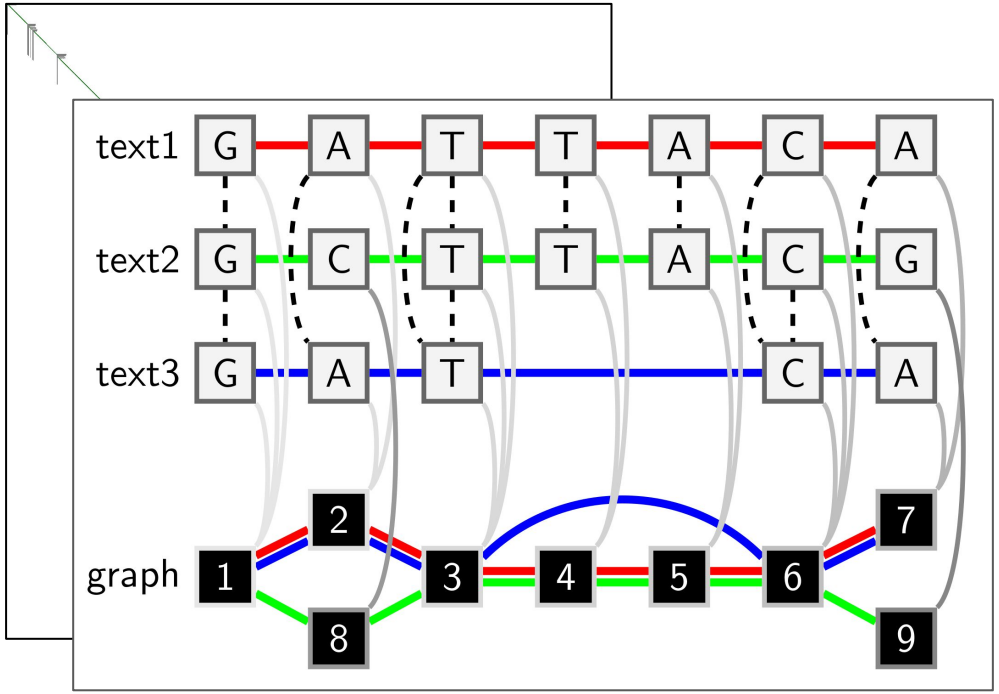
Then: we’ll work through small examples to learn how to drive it.

PanGenome Graph Builder



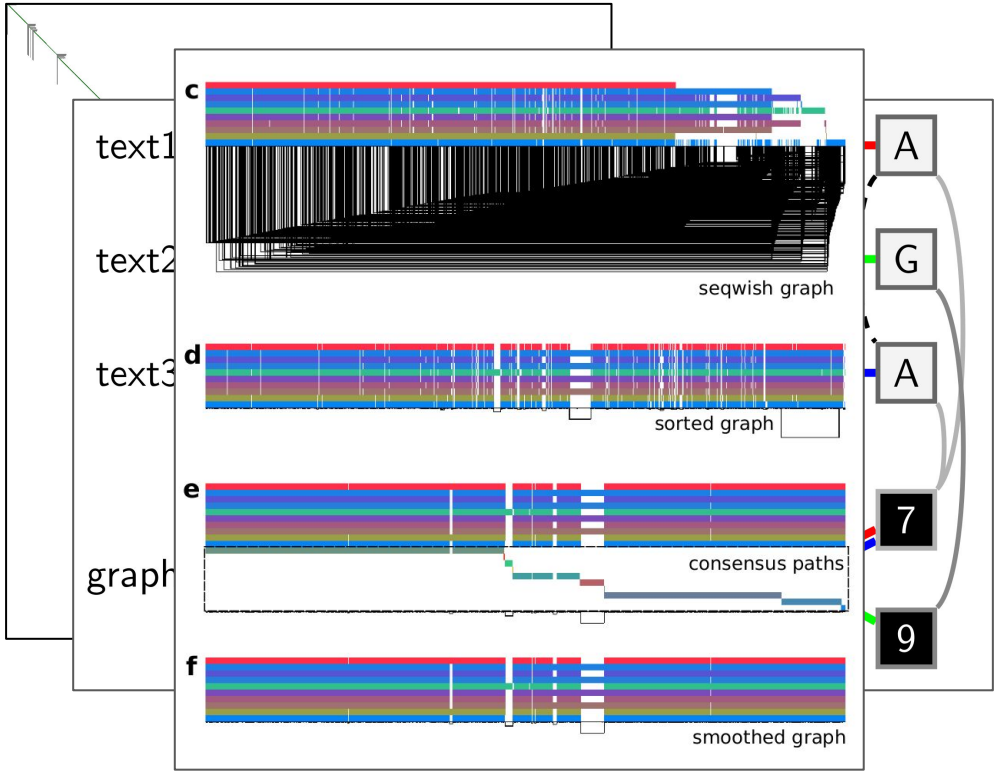
wfmash (biWFA)

PanGenome Graph Builder



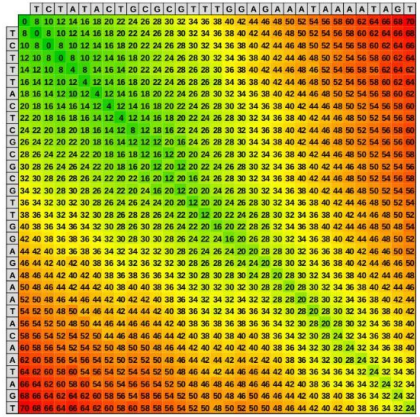
seqwish (unbiased graph builder)

PanGenome Graph Builder

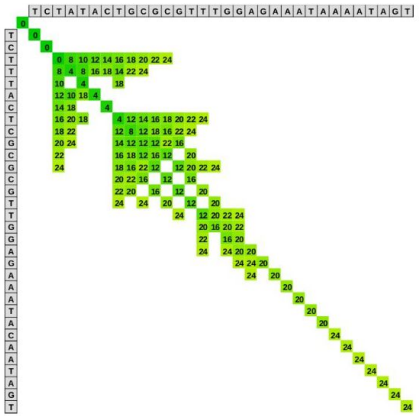


smoothxg (graph normalization)

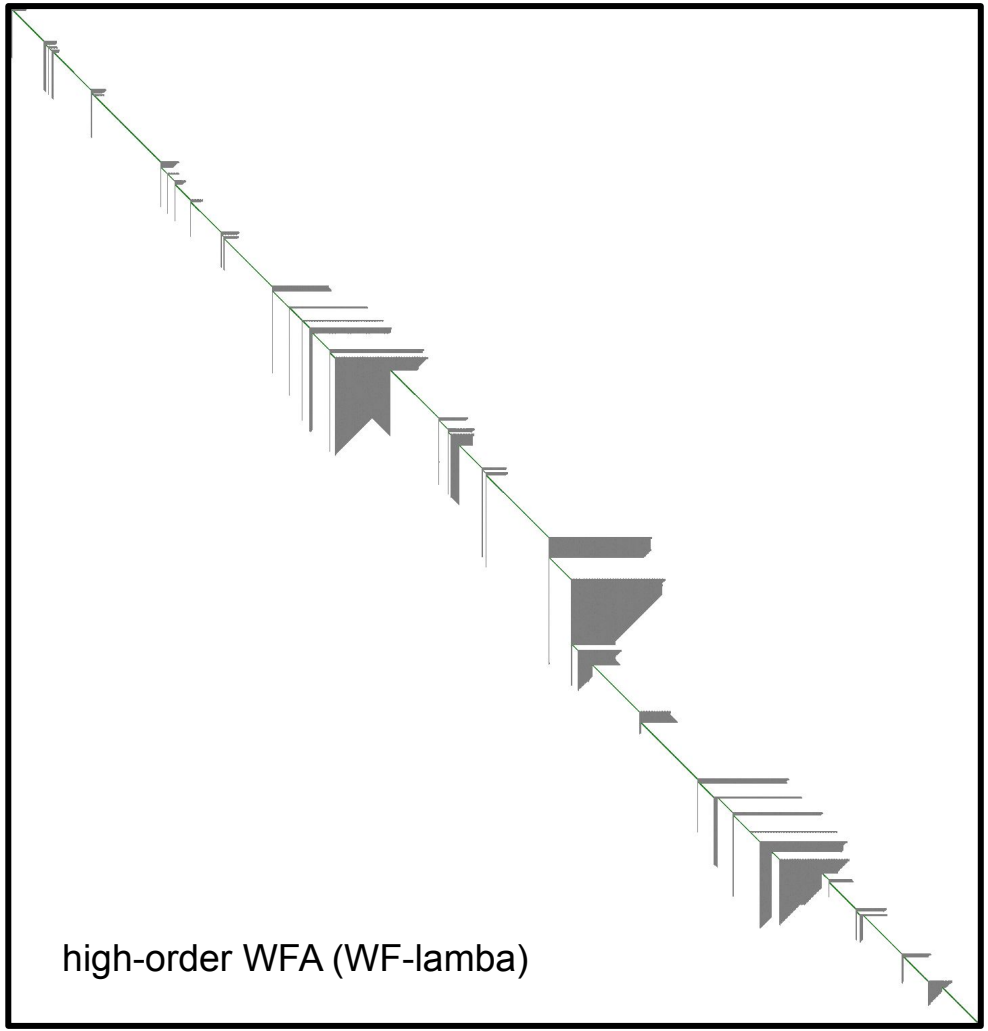
wfmash makes initial alignments



Needleman-Wunsch

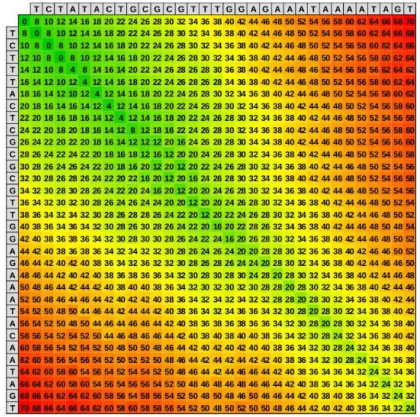


the wavefront algorithm (WFA)

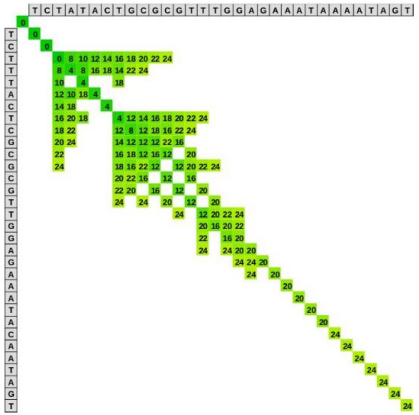


high-order WFA (WF-lamba)

wfmash makes initial alignments



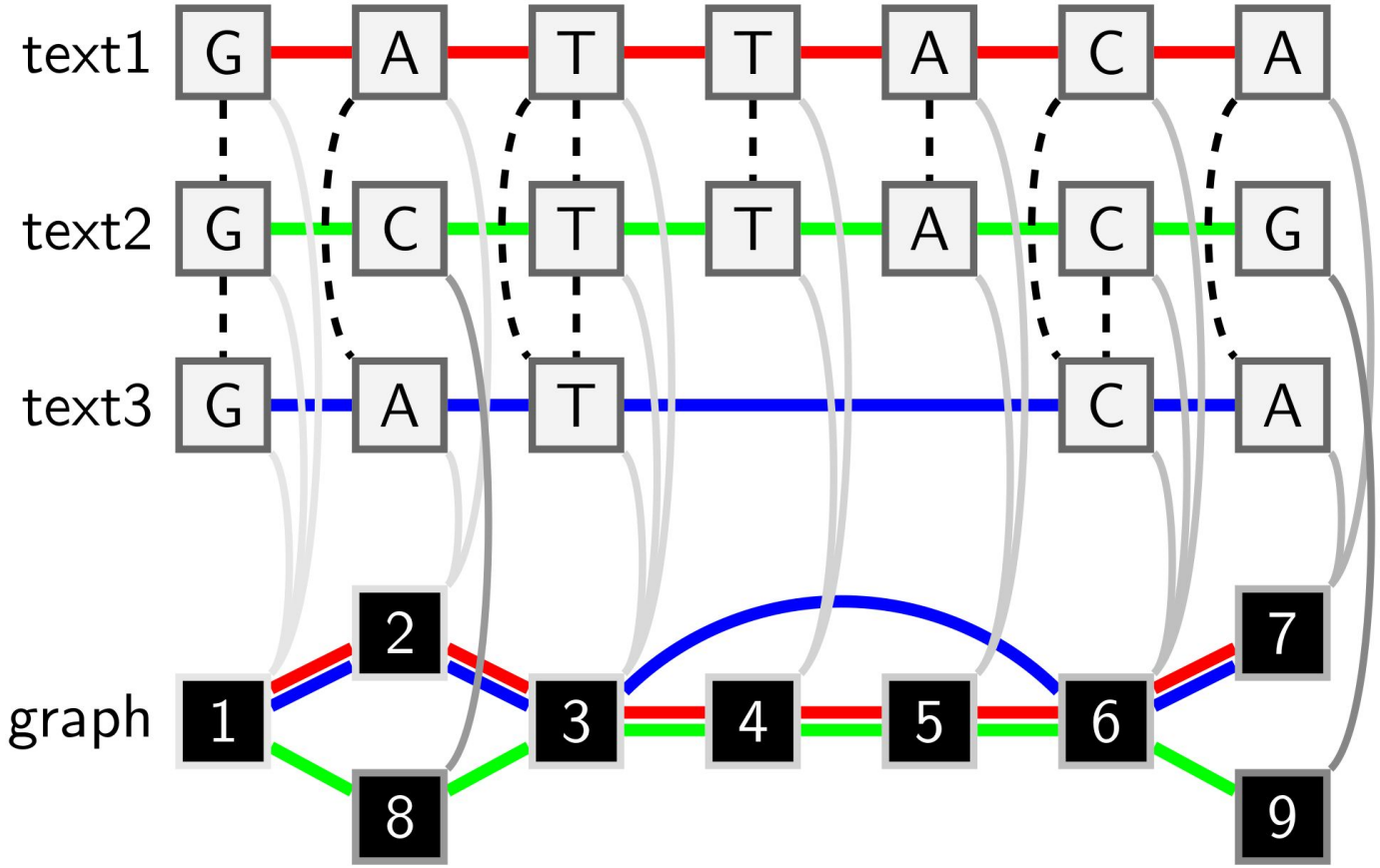
Needleman-Wunsch



the wavefront algorithm (WFA)

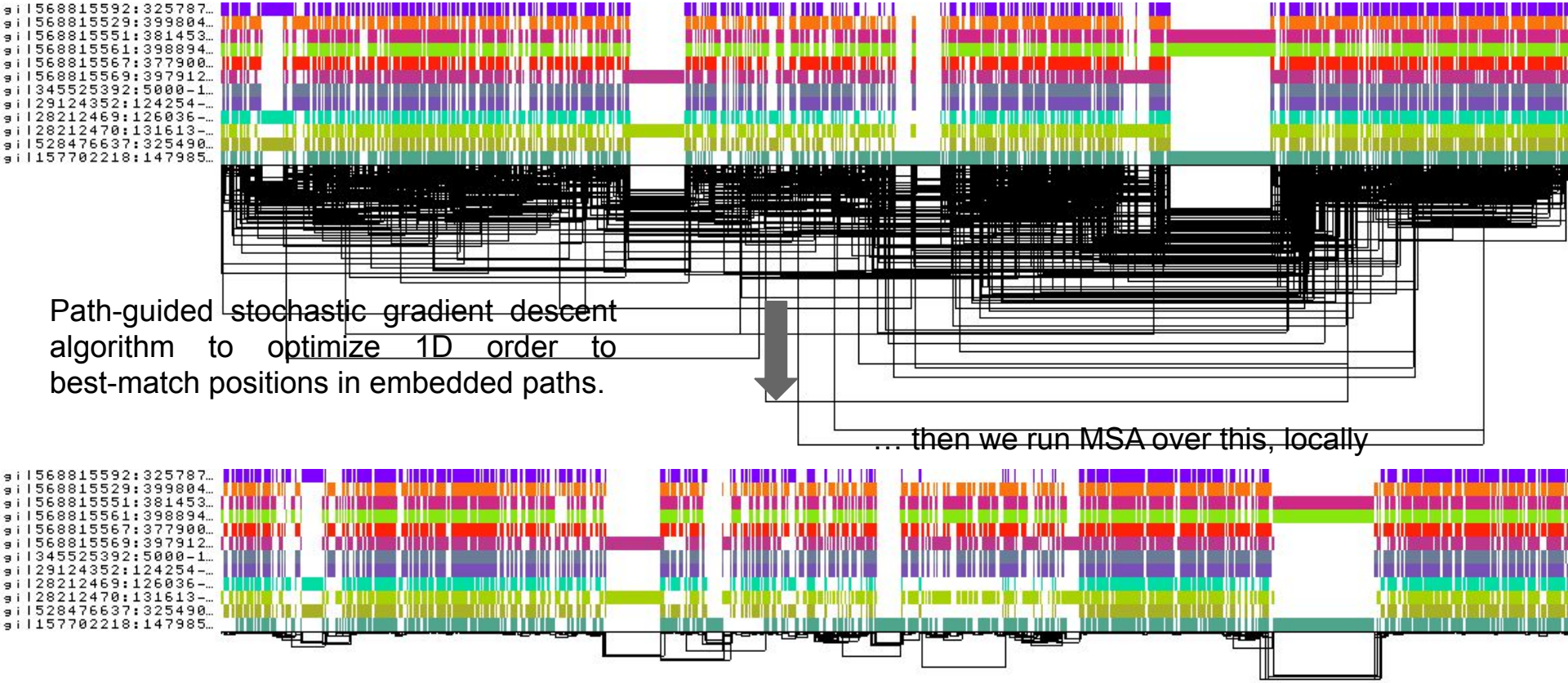
high-order *bidirectional* WFA (BiWFλ)

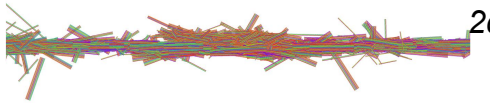
seqwish builds the graph



smoothxg organizes & normalizes the graph

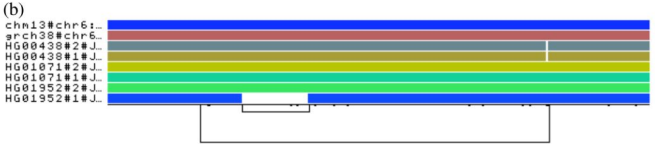
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



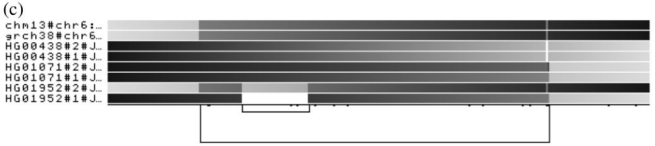


2d layout

identity

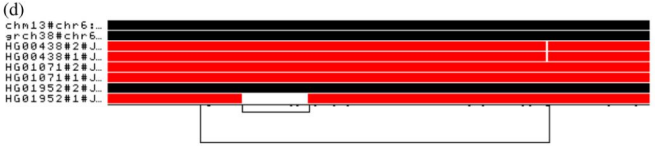


position

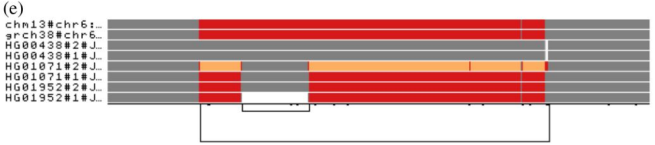


(f)

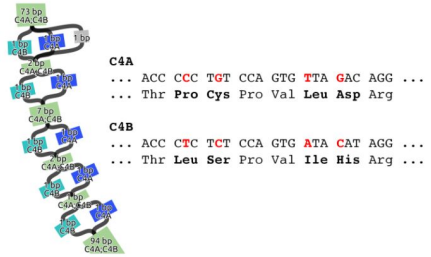
orientation



copy number variation



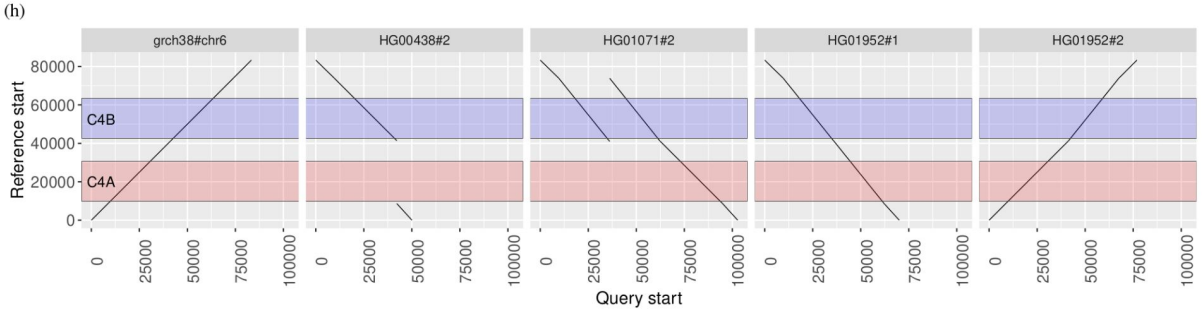
(g)



ODGI is meant to be a basic toolkit for interacting with pangenome graphs.

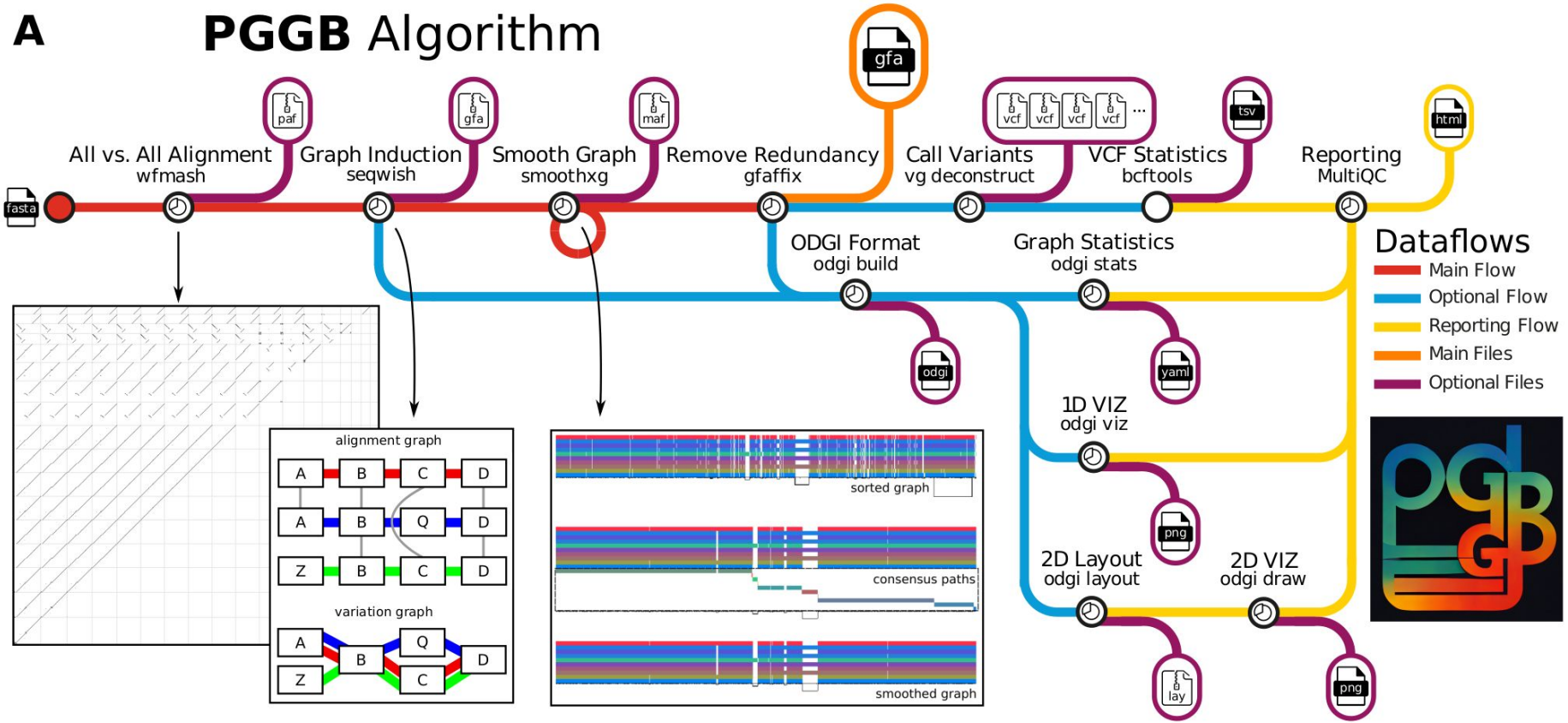
It uses the embedded genomes as references.

odgi helps us understand the pangenome



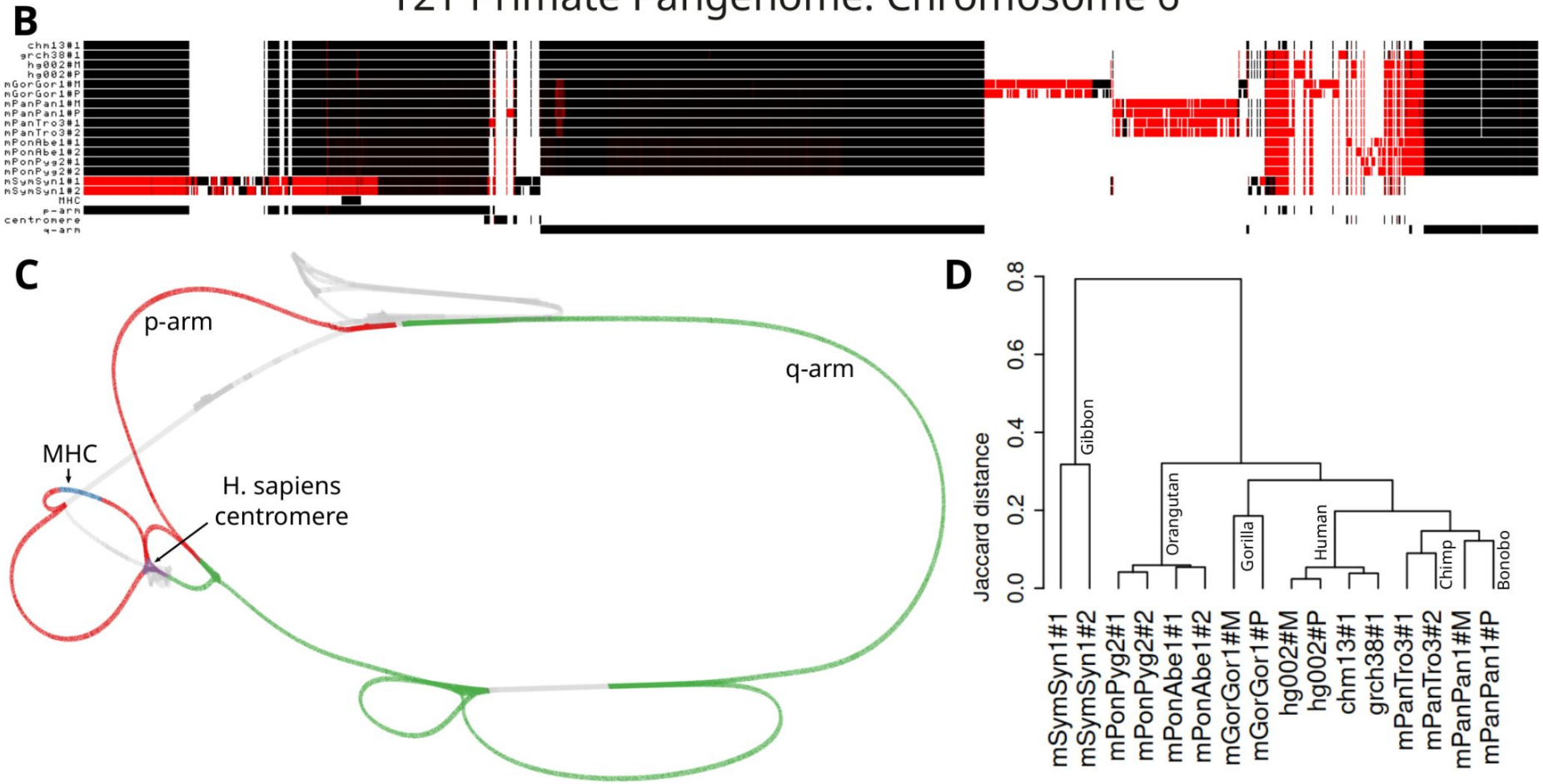
https://www.biorxiv.org/content/10.1101/2021.11.10.467921v1

Putting it all together!



https://doi.org/10.1101/2023.04.05.535718

T2T Primate Pangenome: Chromosome 6

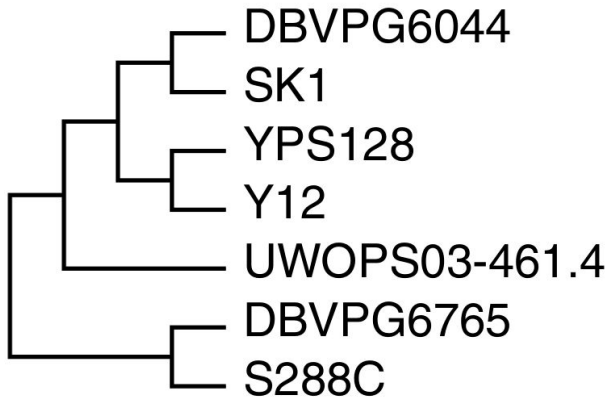


Test material today

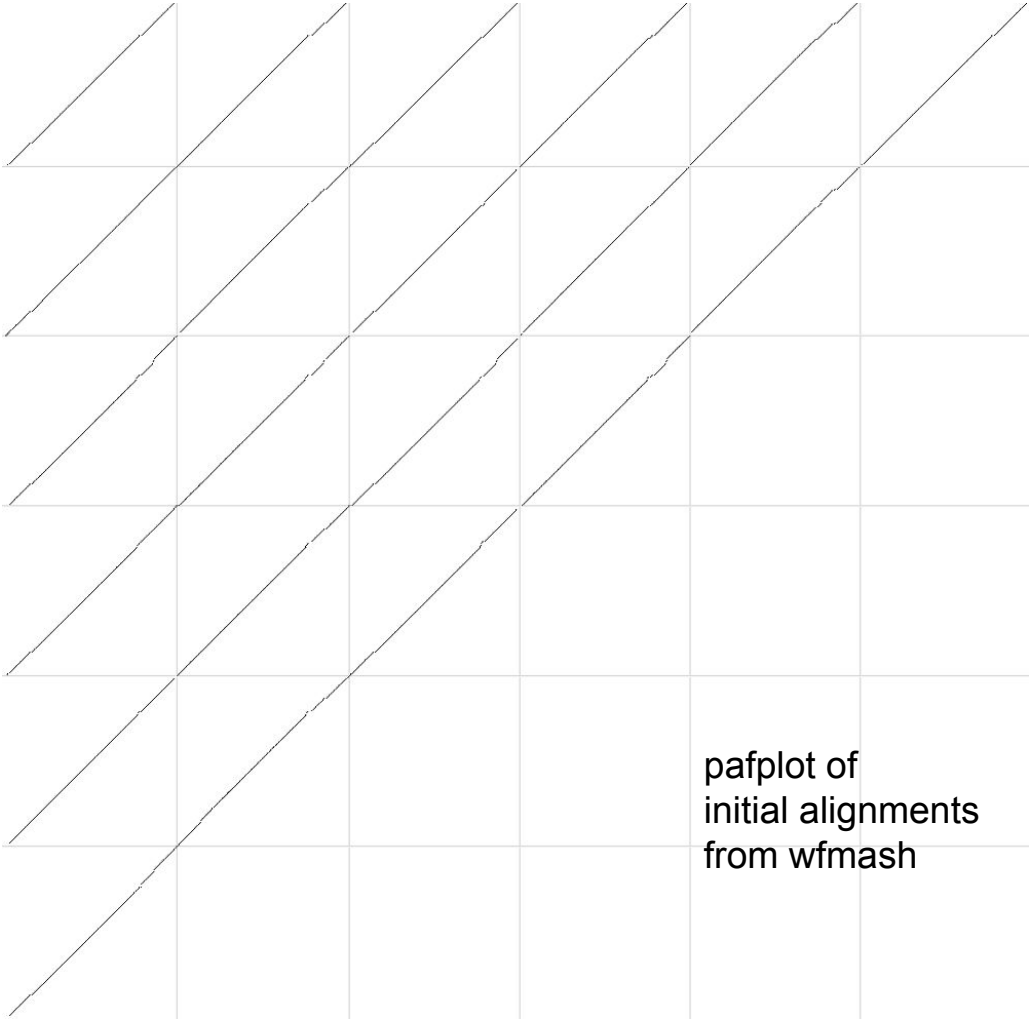
1. A few genes from HLA-D (MHC class II) in humans — getting started
   a. https://github.com/pangenome/pggb-workshop/tree/evomics2024
2. Yeast chromosome 6 — scaling up
   a. ~/workshop_materials/pangenomics/cerevisiae.chrV.fa
   b. you will want to apply samtools faidx to this… pggb will warn you
3. Whole yeast chromosomes — looking at chromosome variation
   a. ~/workshop_materials/pangenomics/cerevisiae.pan.fa.gz

Example: yeast chromosome 6

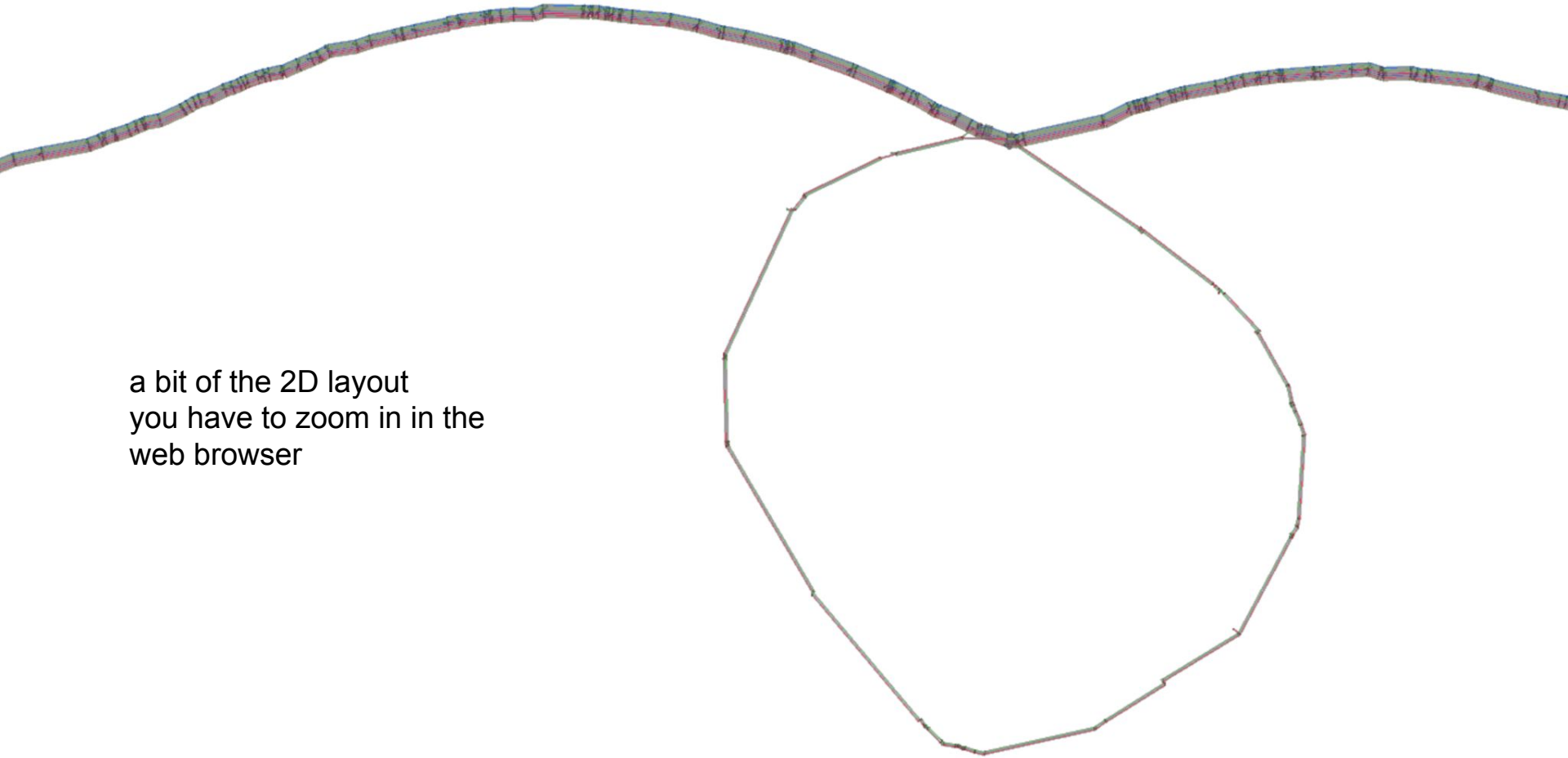
Yue, JX., Li, J., Aigrain, L. et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat Genet 49, 913–924 (2017). https://doi.org/10.1038/ng.3847



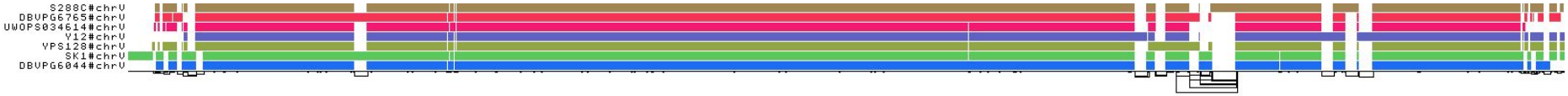
Cladogram of the *S.c.* clade



pafplot of initial alignments from wfmash

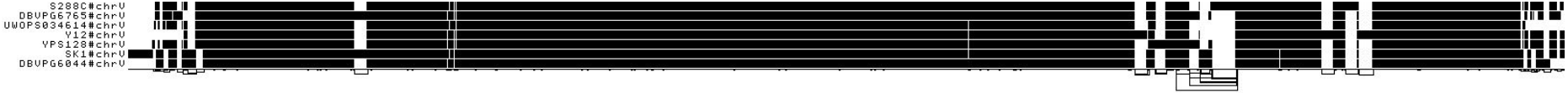


a bit of the 2D layout
you have to zoom in in the
web browser

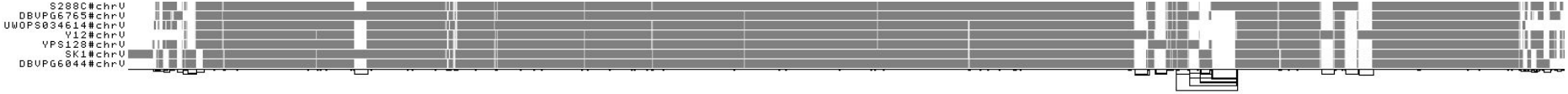


path view

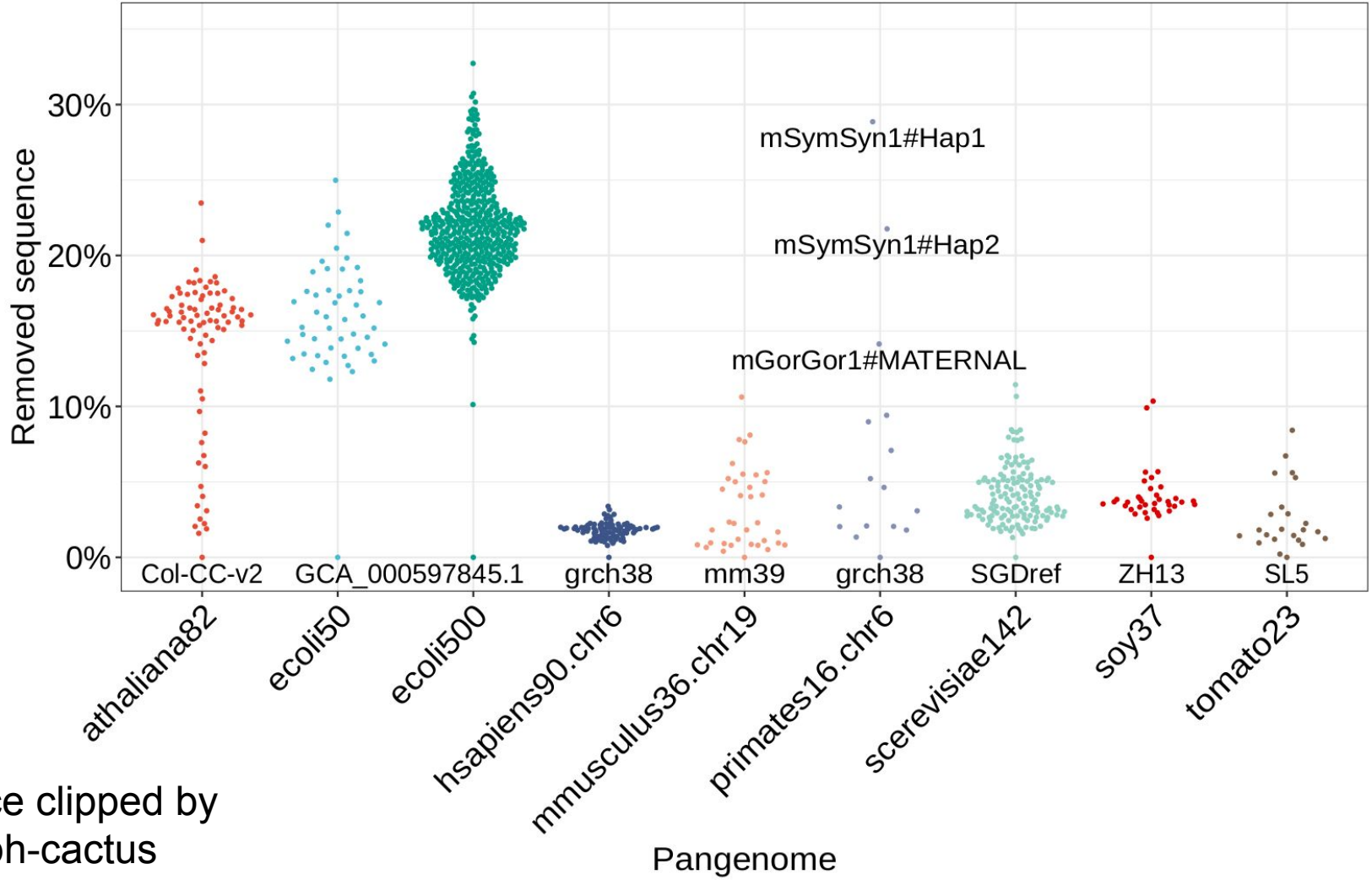
position view



inversion view

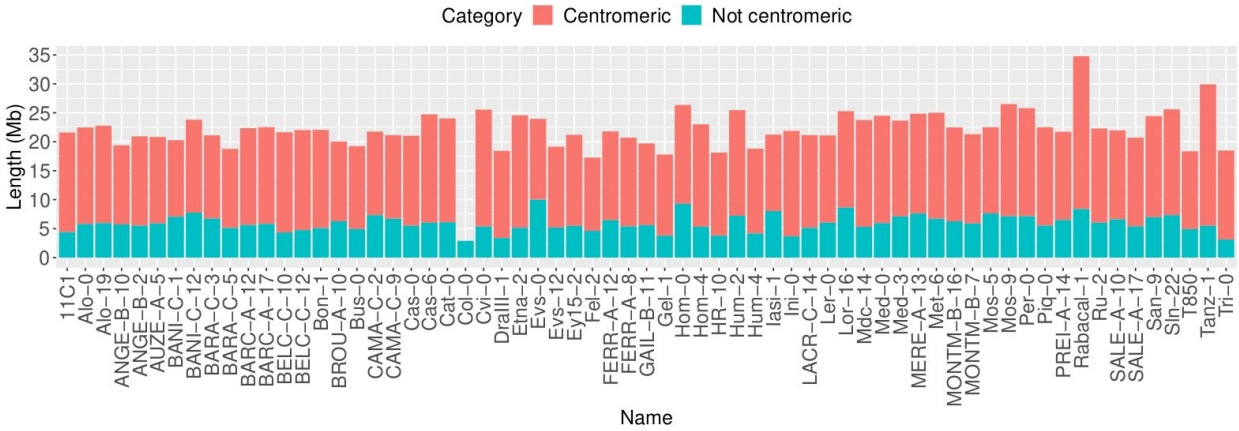


copy number view



sequence clipped by minigraph-cactus

annotation of clipped sequences in minigraph-cactus for *A. thaliana* pangenome



(b)