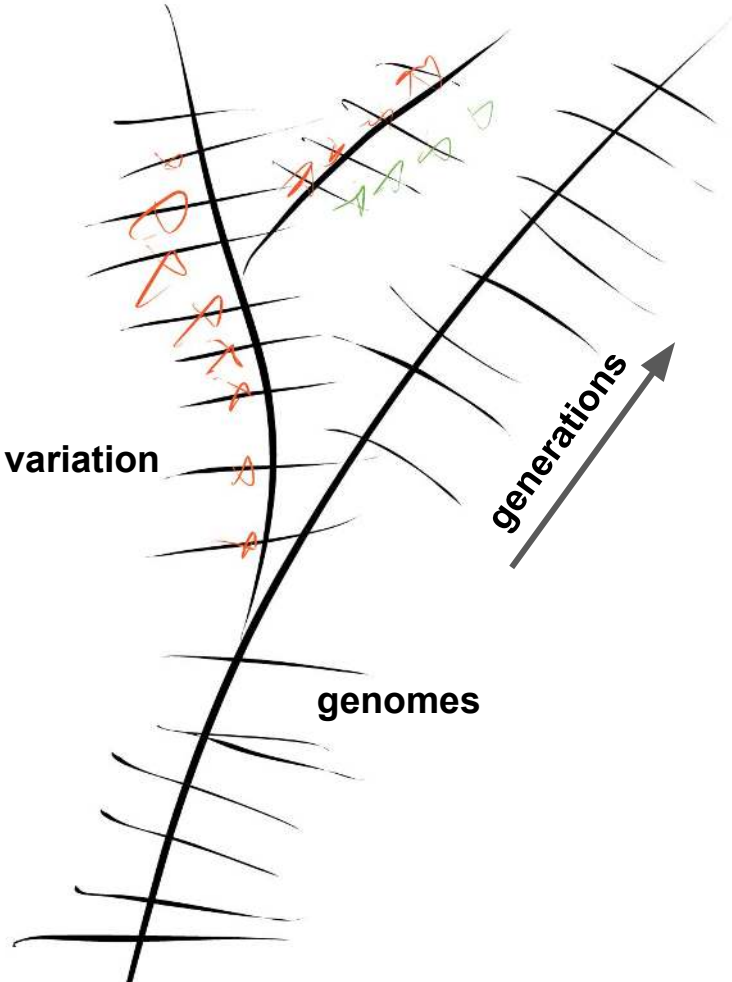
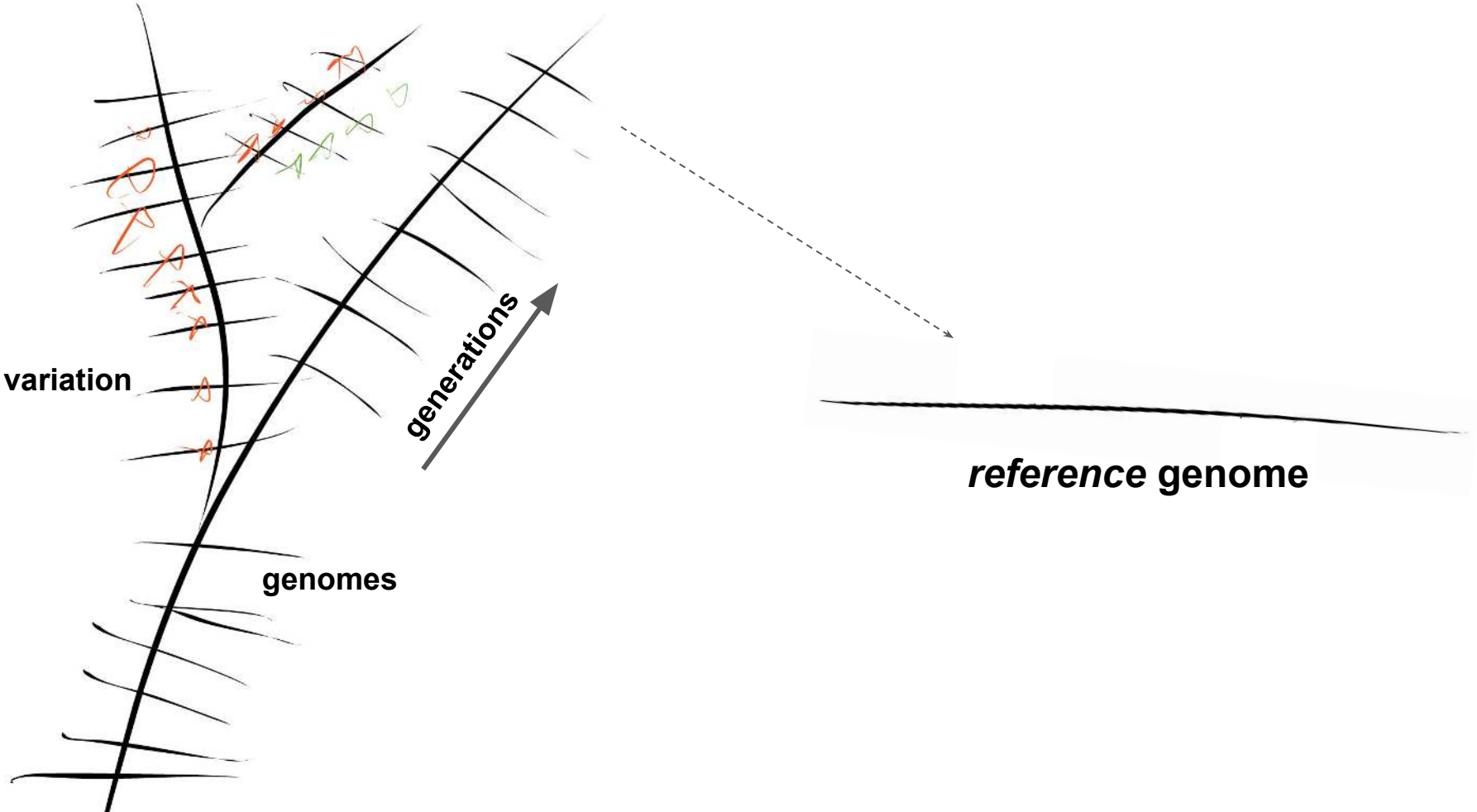
Building and understanding pangenomes

Erik Garrison

University of Tennessee (UTHSC), Memphis

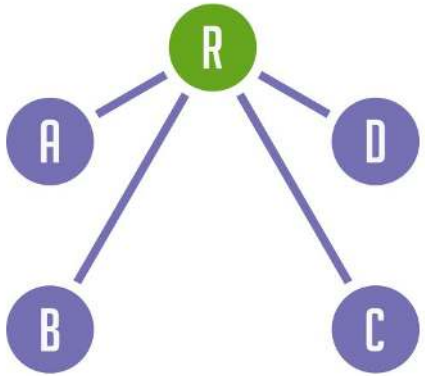
@Workshop on Genomics, Český Krumlov
May 20, 2023

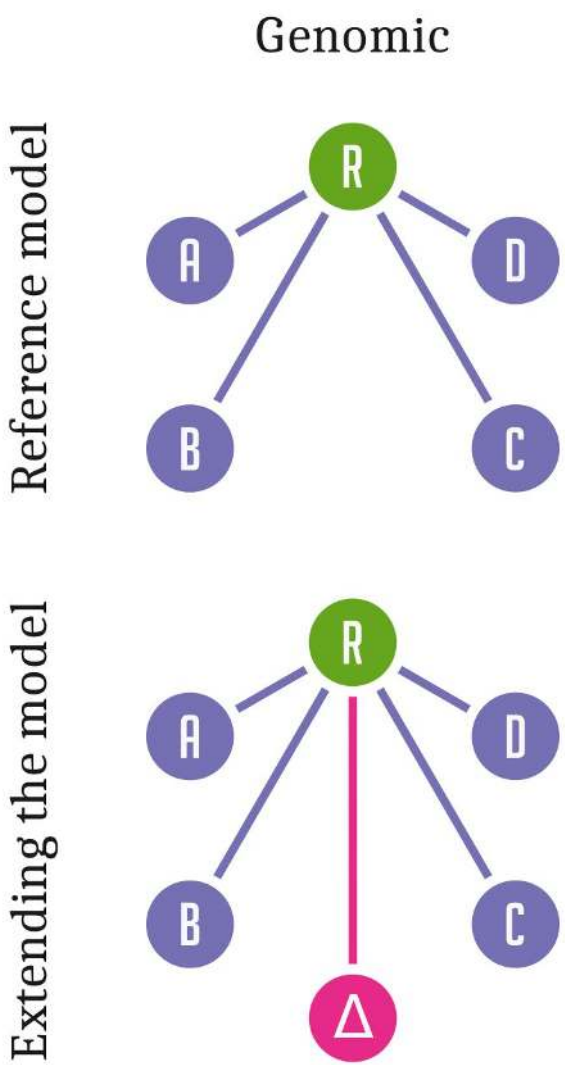

variation
generations
genomes


variation
generations
genomes
reference genome

Genomic

Reference model

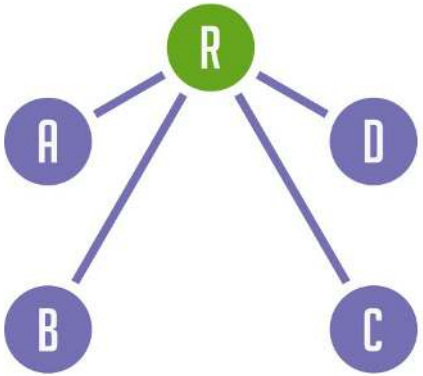




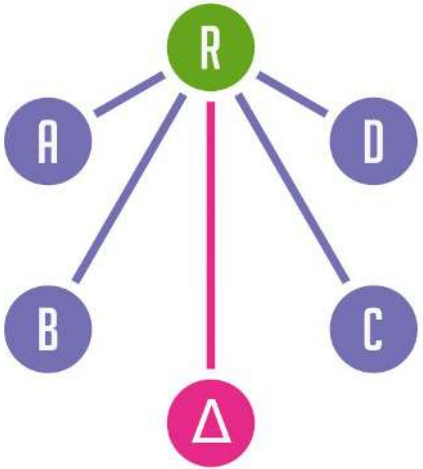
Δ: new genome; R: reference genome.
Figure from Eizenga et al., 2020.

Genomic

Reference model



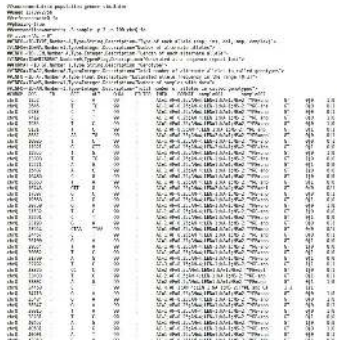
Extending the model



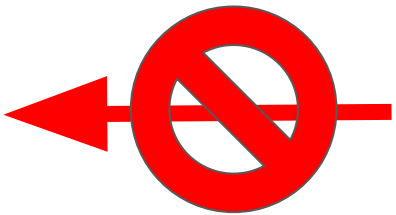
Genome (*FASTA*)

Reads (*FASTQ*)

alignment and variant calling



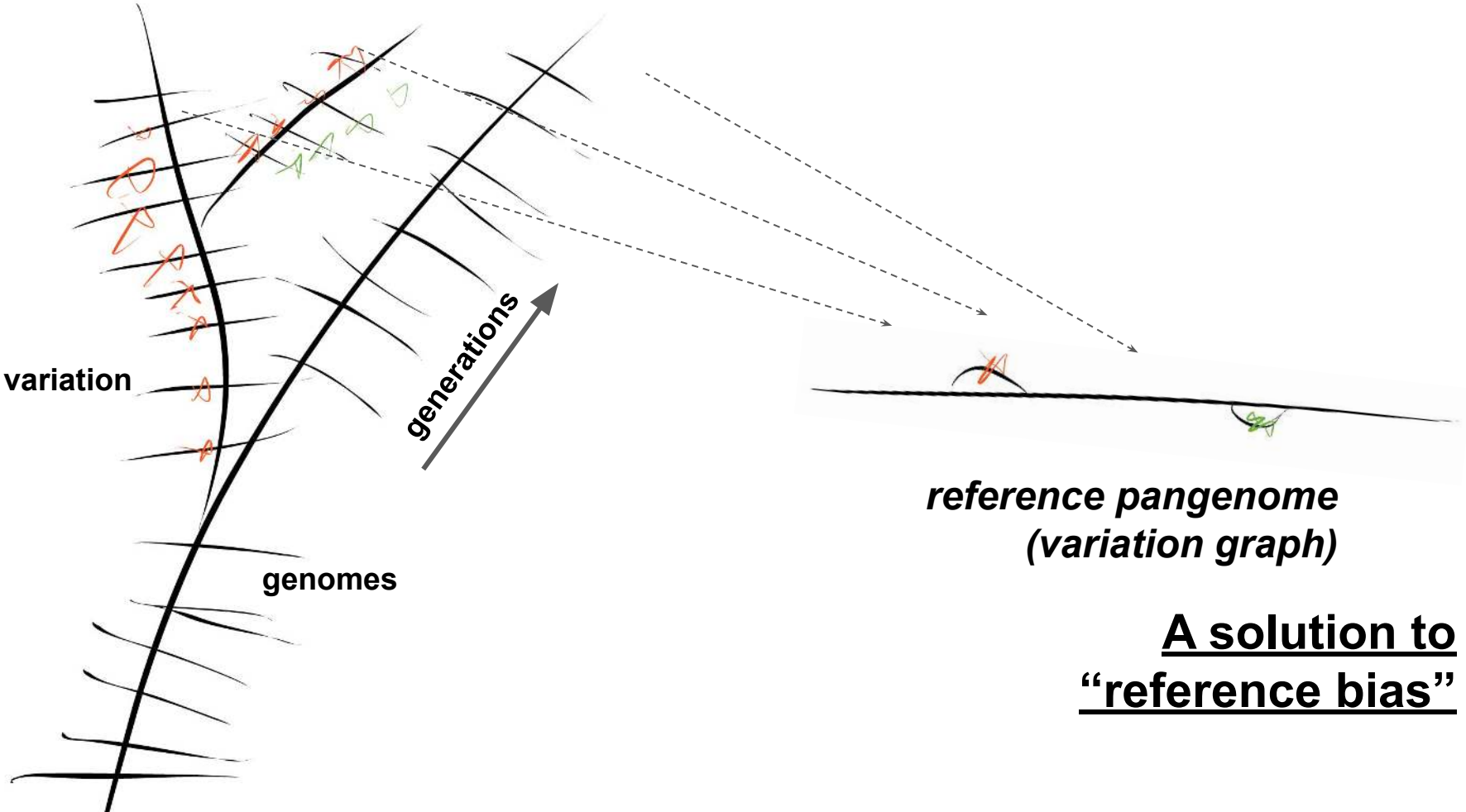
Variation (*VCF*)



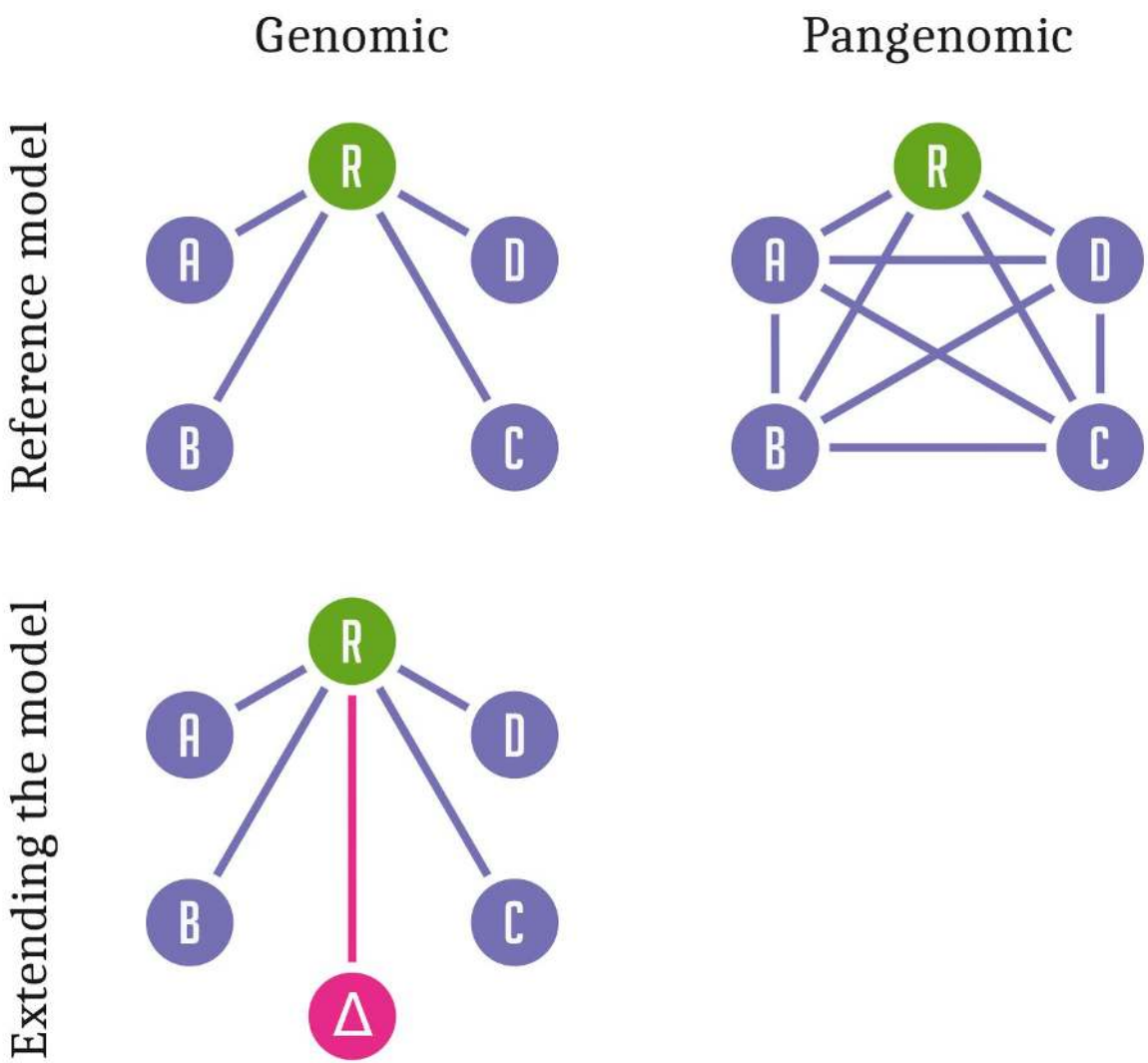
We cannot update a linear reference sequence

Genome (*FASTA*)

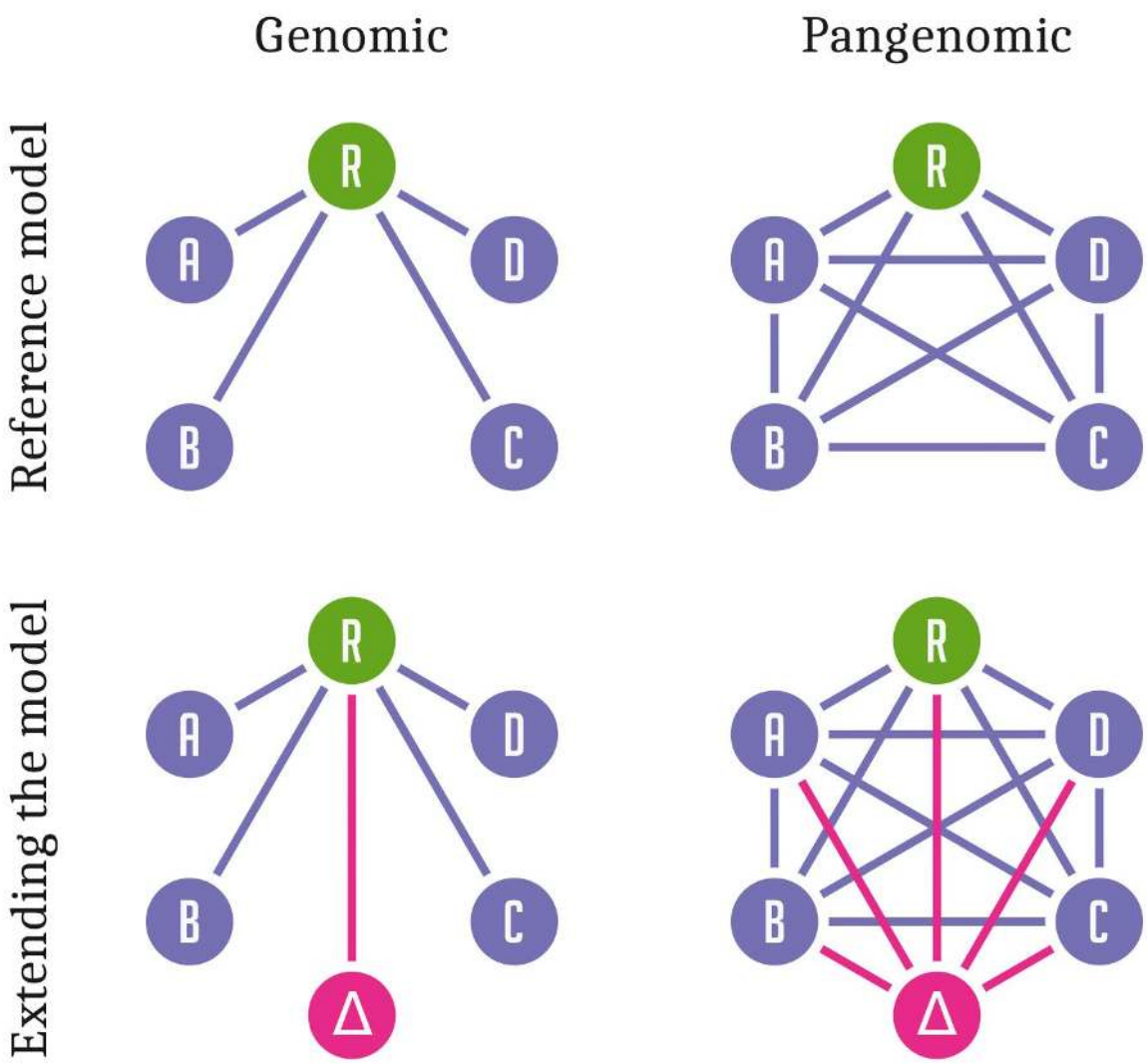
Variation (*VCF*)



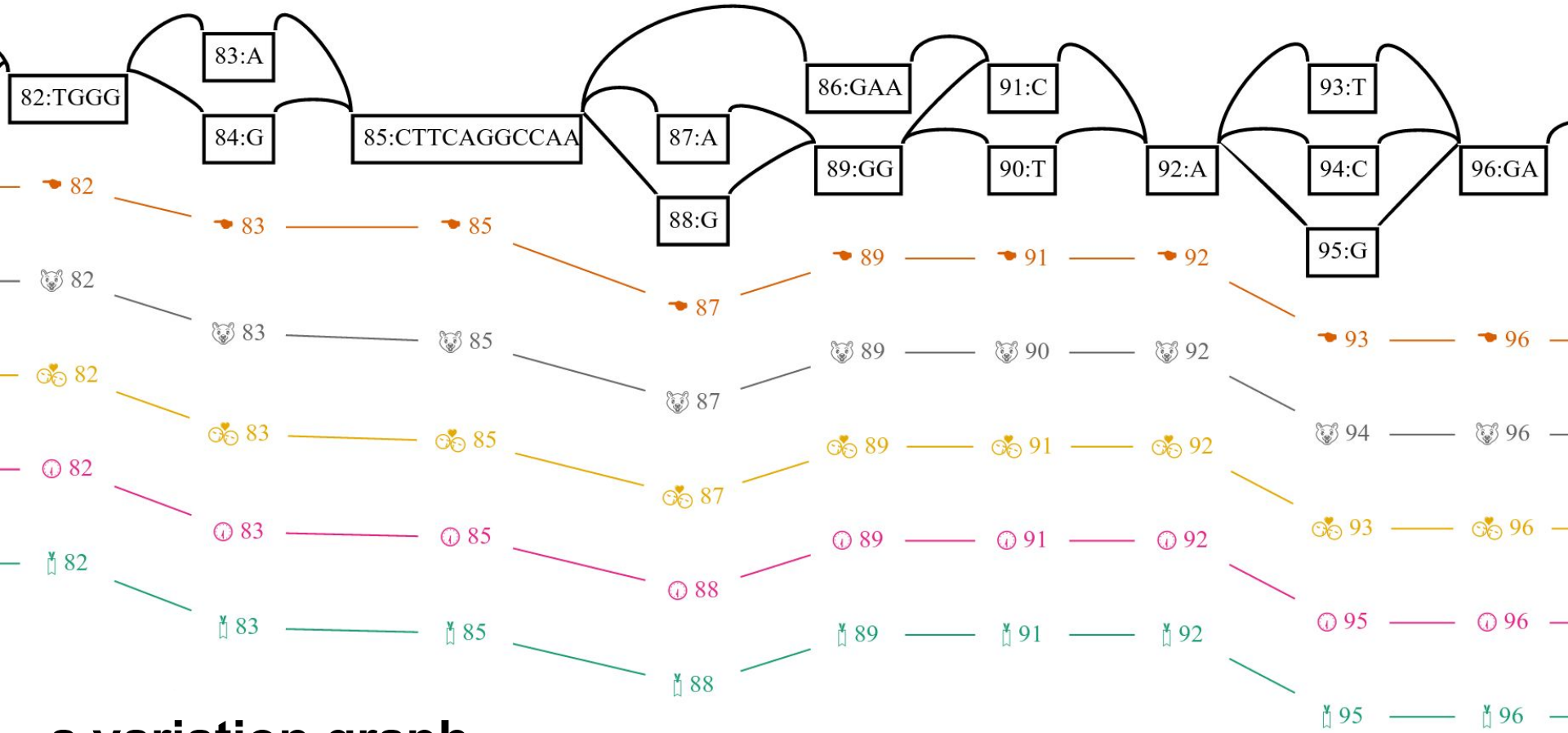
A solution to "reference bias"



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

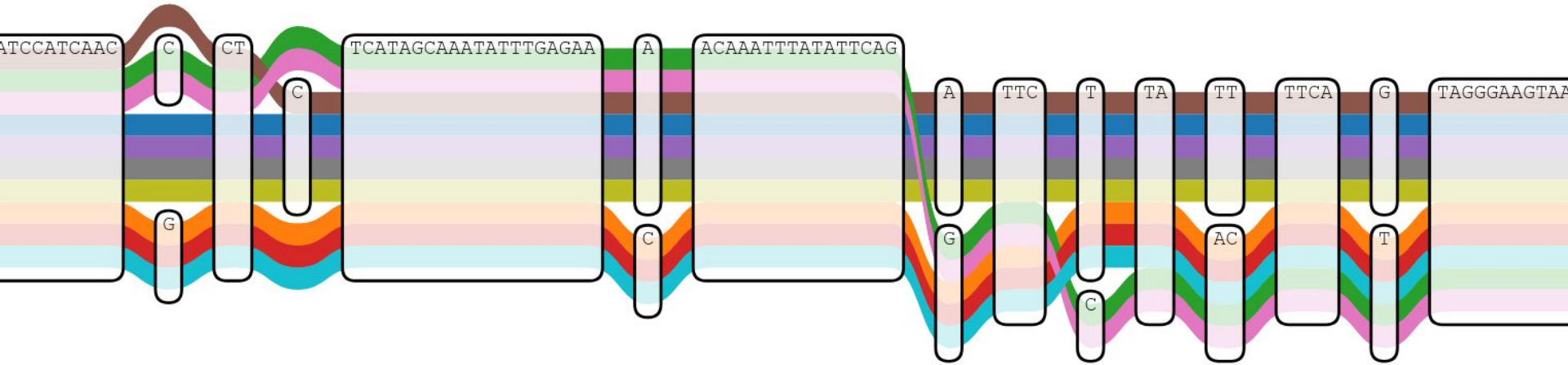


Δ: new genome;
R: reference genome.
Figure from
Eizenga et al., 2020.

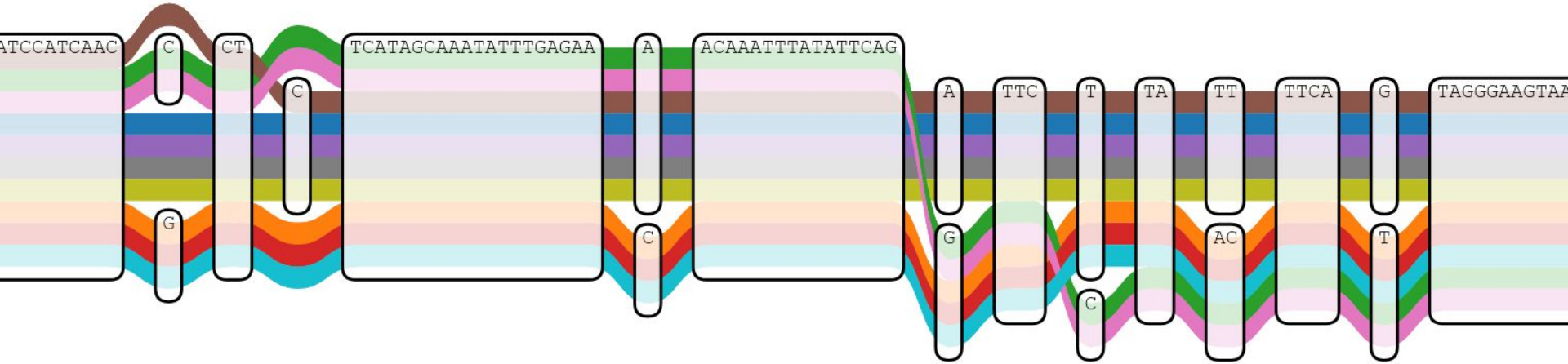


a variation graph

Variation graphs answer a key problem in bioinformatics:

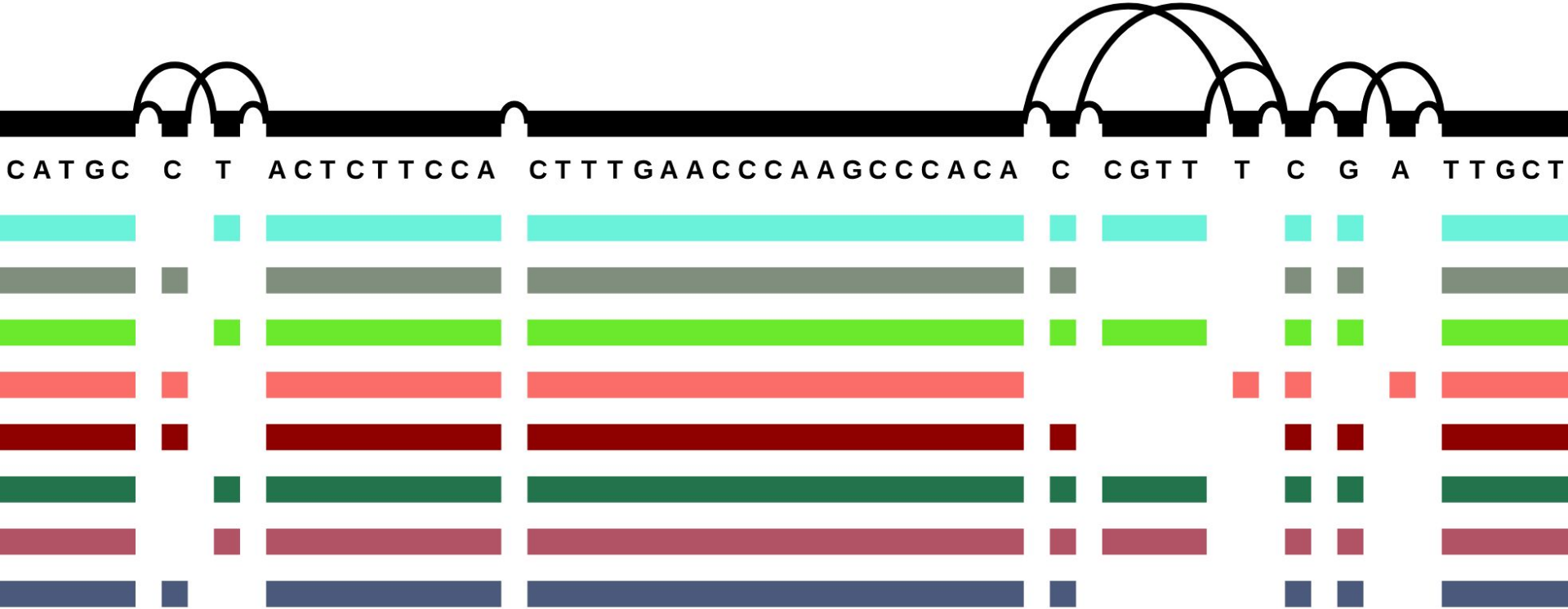


Variation graphs answer a key problem in bioinformatics:

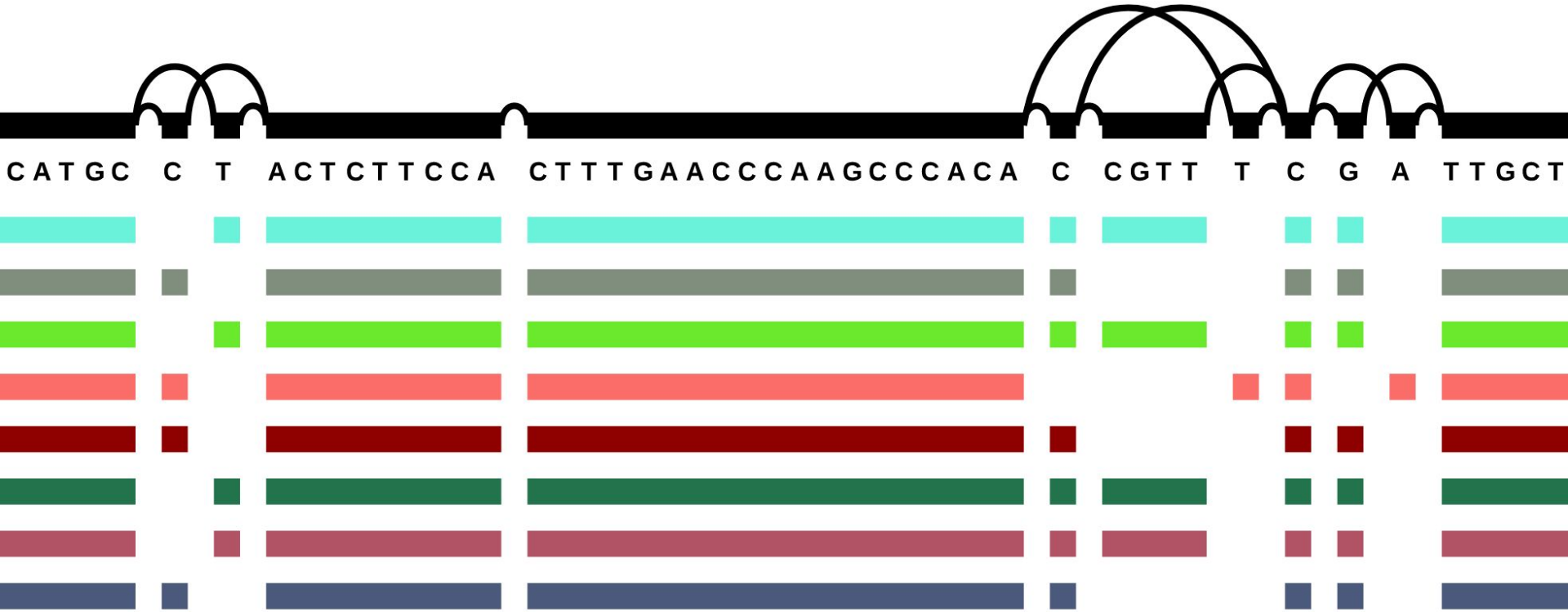


How to represent *both* sequences and *any* kind of variation between them.

variation graphs are *pangenome* models



variation graphs are *pangenome* models



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

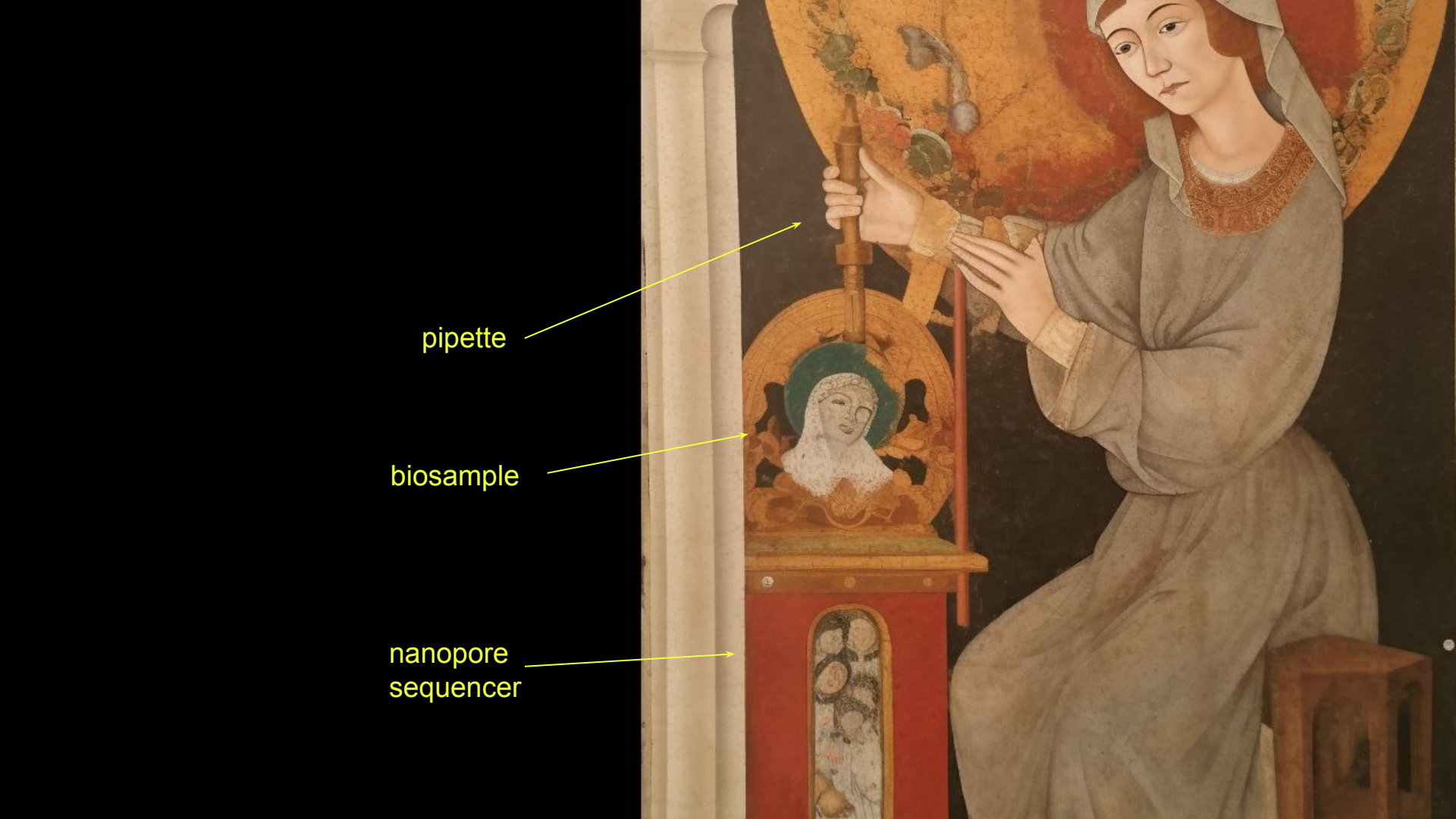
“New” ideas often have a long history

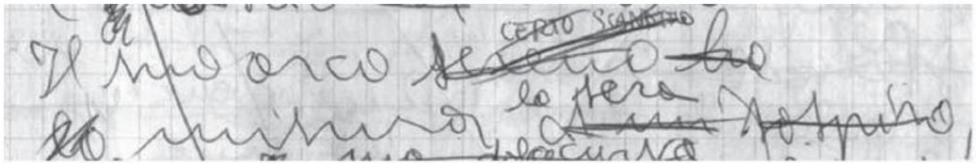
This all seems cool and “new” but ideas are rarely that.

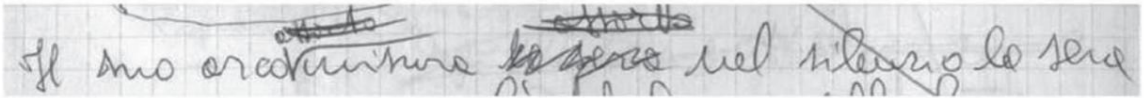
Pangenomes and variation graphs have a long† history.

(†for genomics)

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


pipette
biosample
nanopore
sequencer





e lo illumina.Il suo arco misura
nel silenzio la sera

e lo illumina.Il suo arco
misura nel silenzio la sera

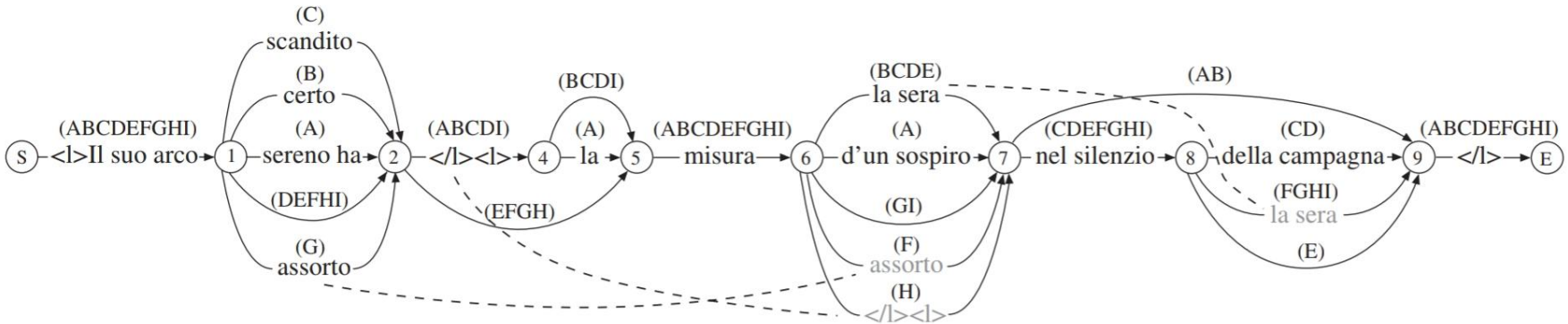
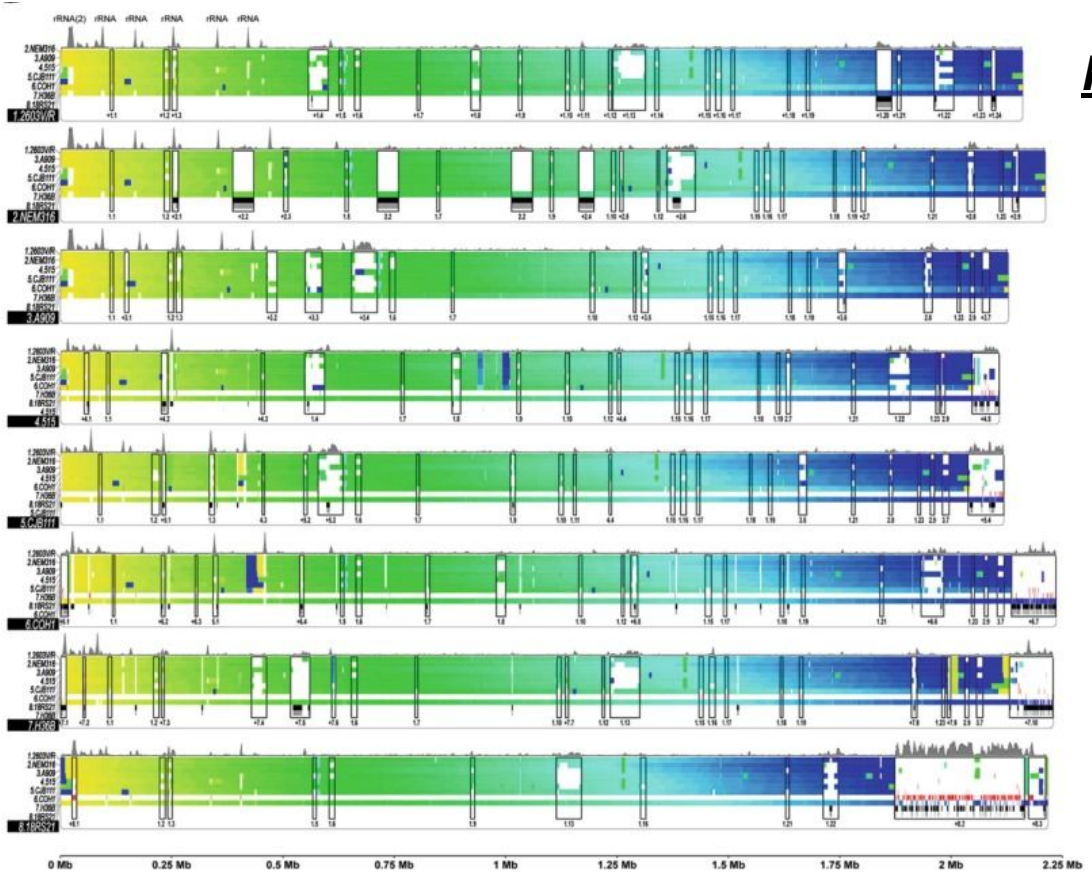


Fig. 5. A variant graph.

variation graphs: not a new idea!

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



Group B Streptococcus assemblies from 2002

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

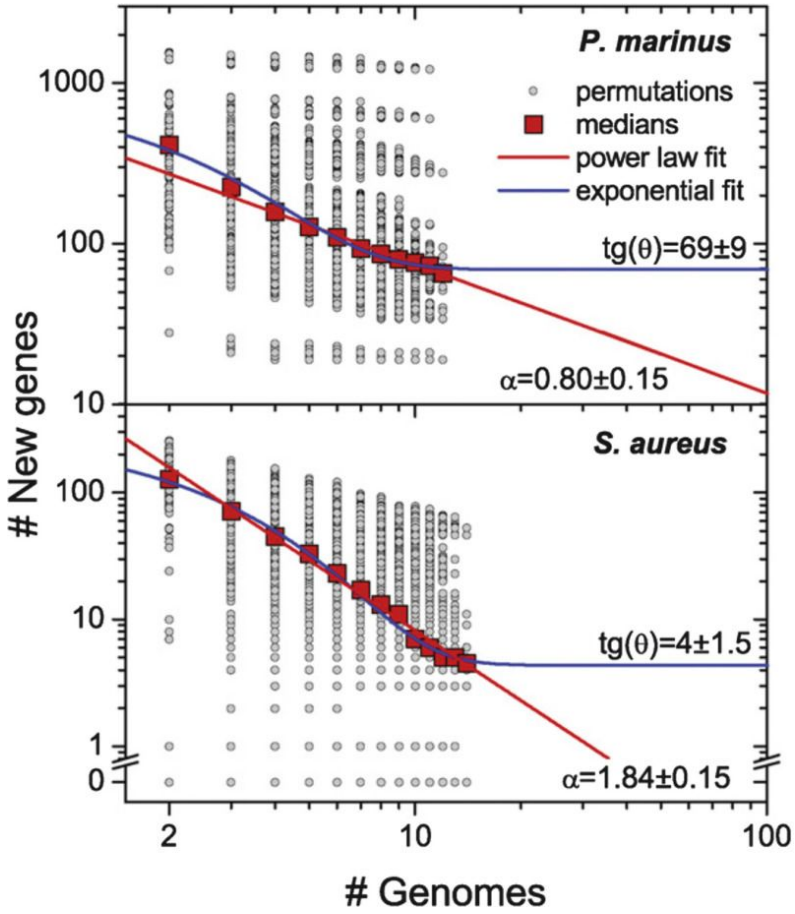
Pangenome: not a new concept

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”.



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.

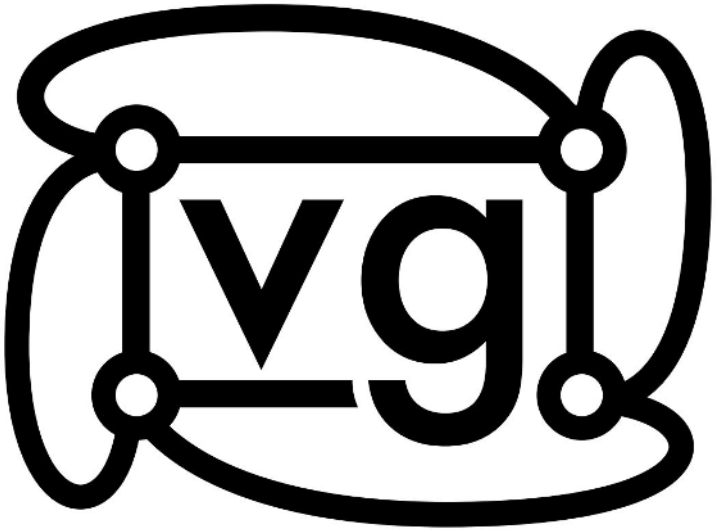
https://doi.org/10.1007/978-3-030-38281-0

Pangenome research timeline

2000-2010s: counting genes

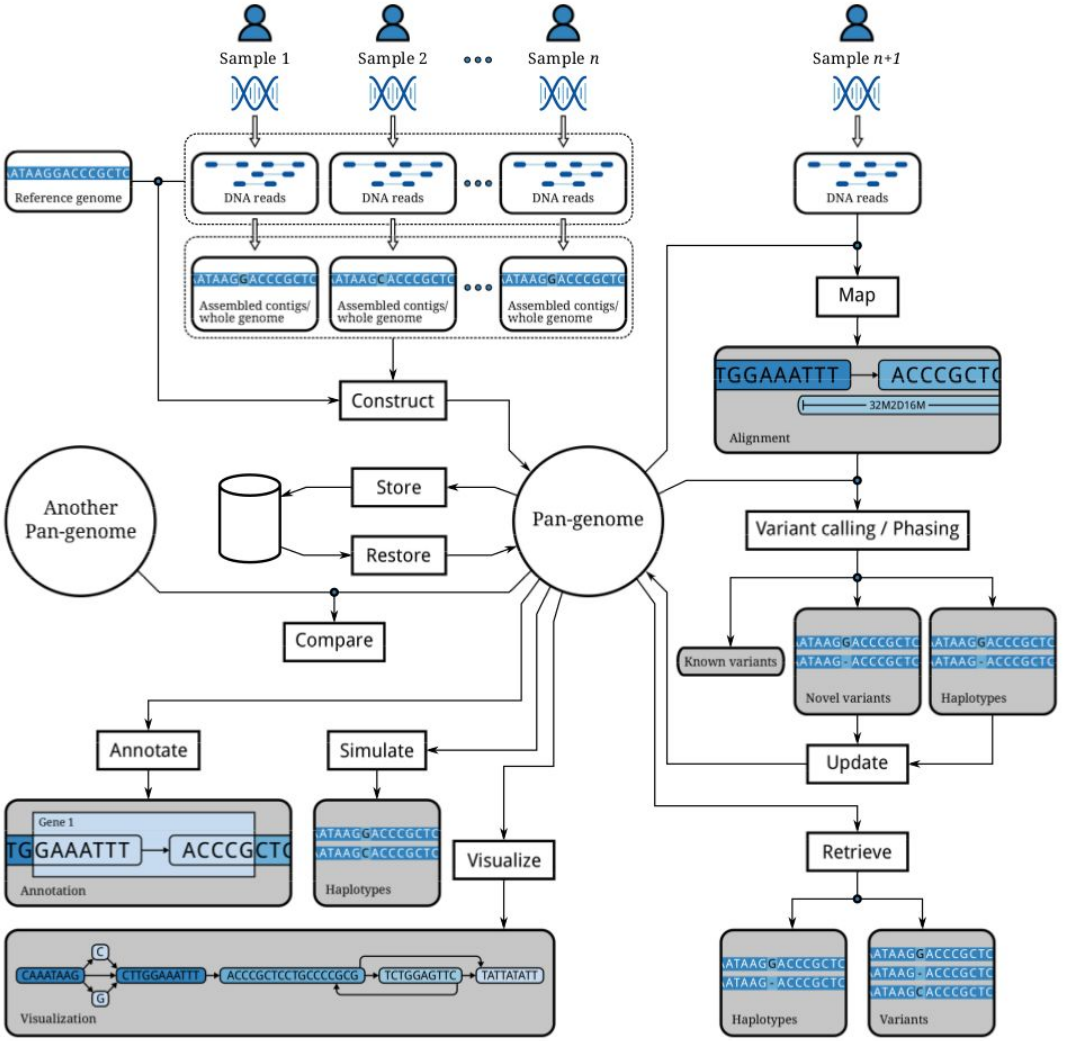
~2015: let’s take it to the sequence level (genome graphs)

2020s: complete assemblies (T2T pangenomes)



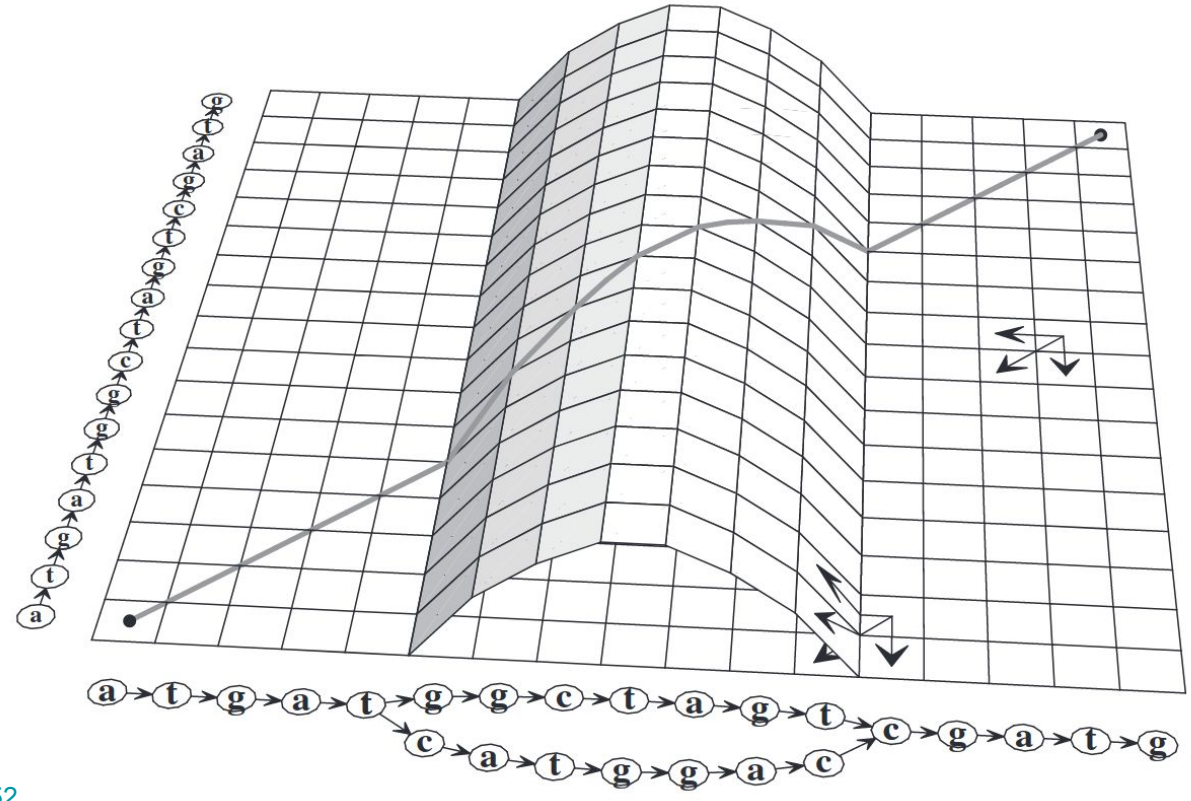
github.com/vgteam/vg

“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

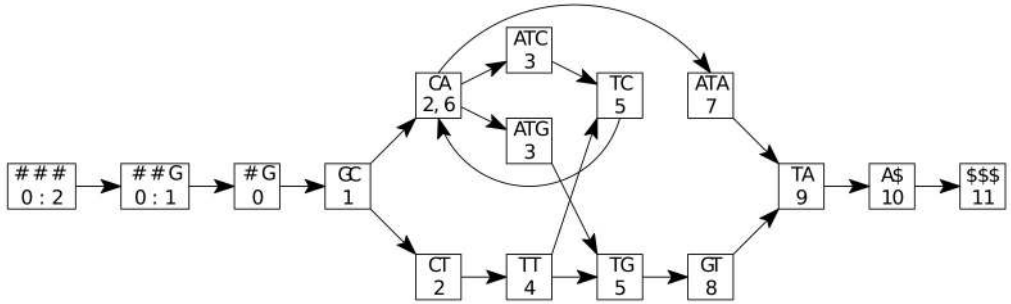
And the FM-index?

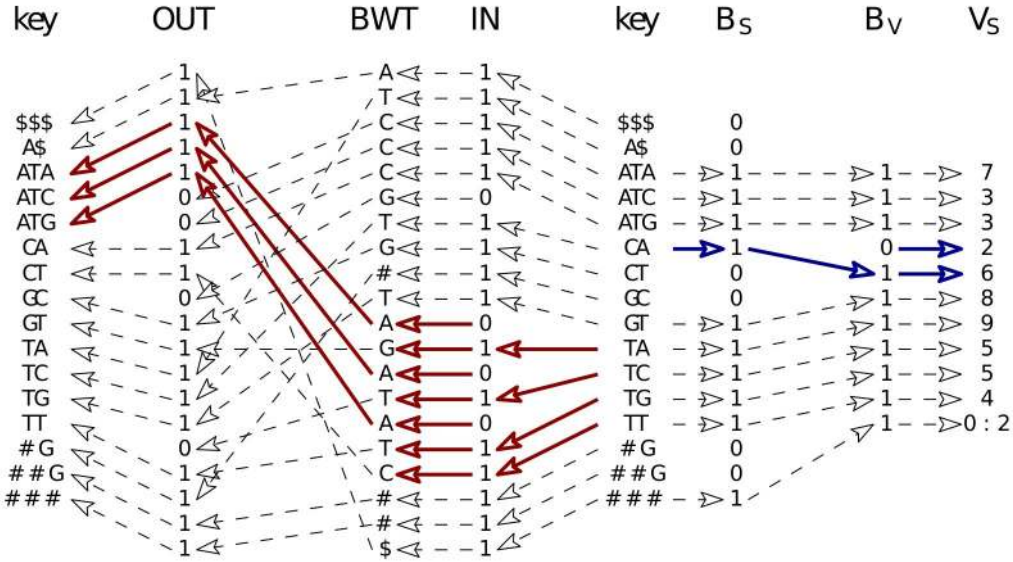
Jouni Sirén generalized the FM-index to work on a transformation of the variation graph (technically a de Bruijn graph with k=256).

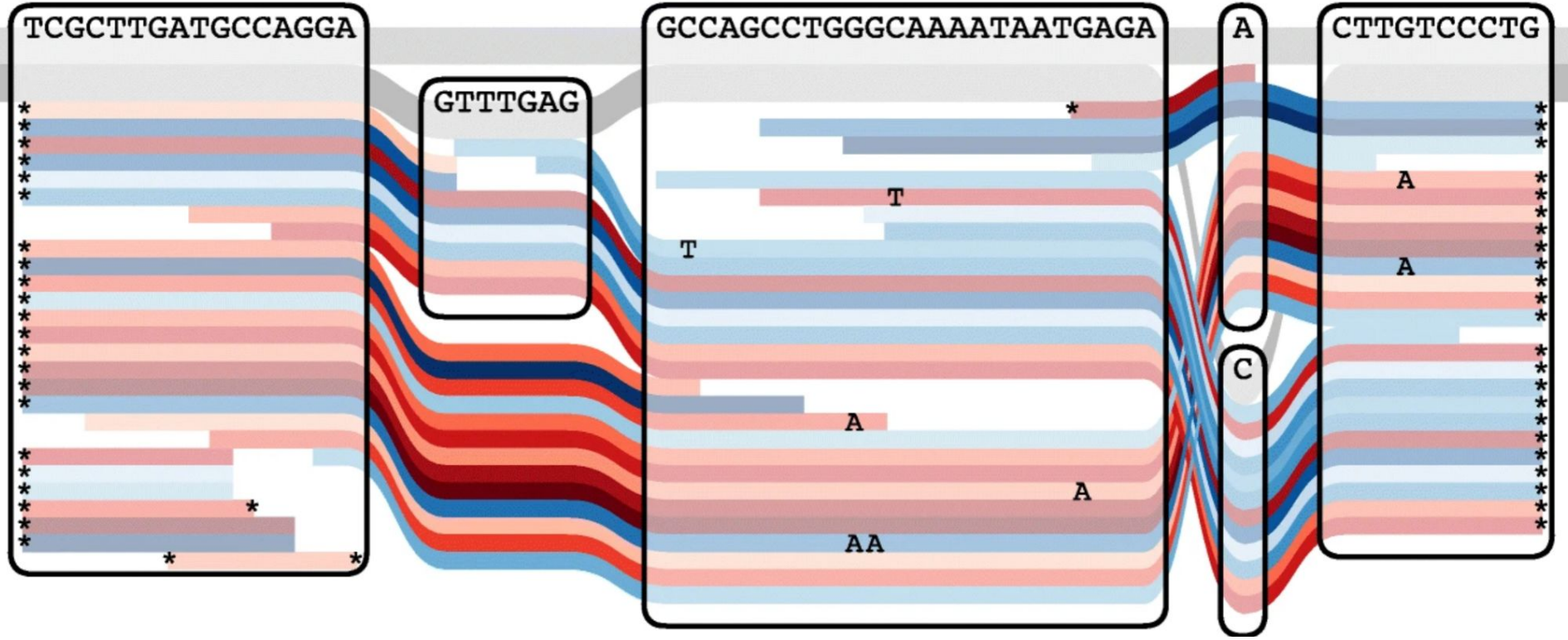
GCSA2 →

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

This seeds alignment to the graph.



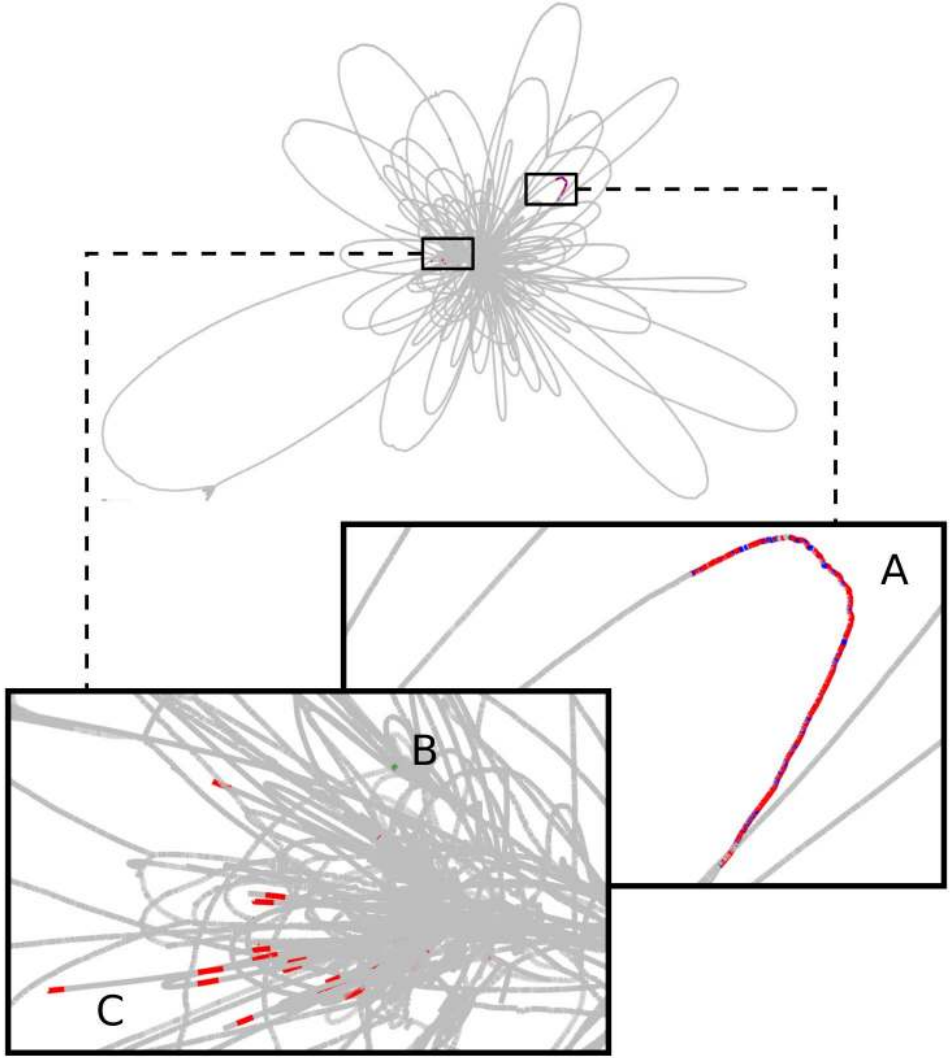




reads aligned to a variation graph

and long reads too!

a pacbio read vs. a yeast graph:



nature biotechnology

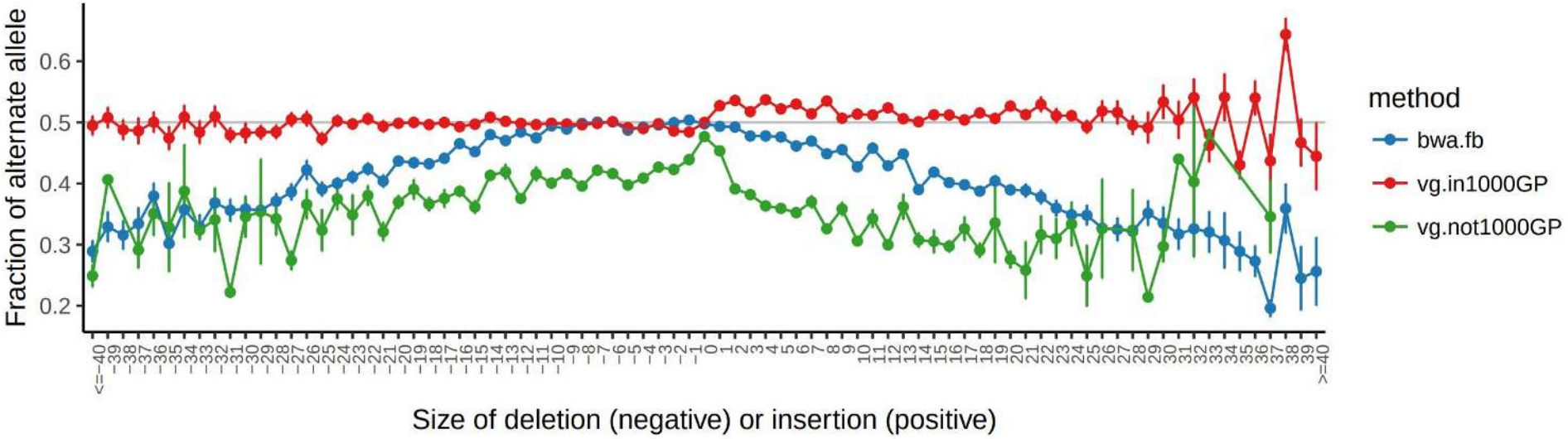
Letter | Published: 20 August 2018

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*



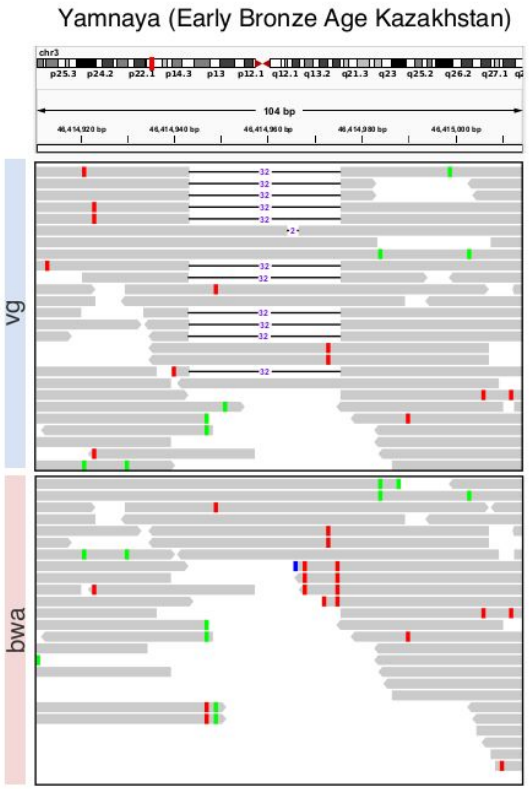
50x 2x150bp Illumina sequencing of HG002

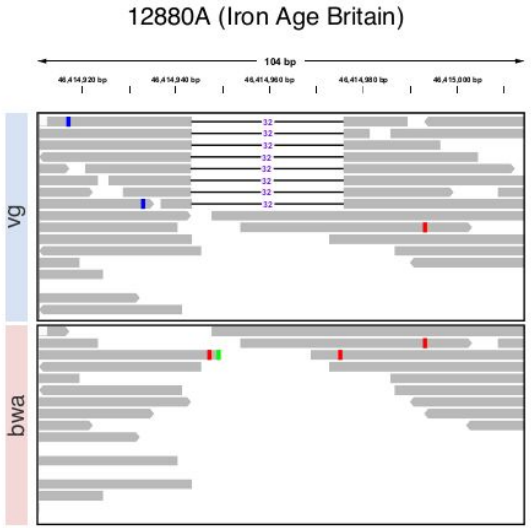
Genome Biology

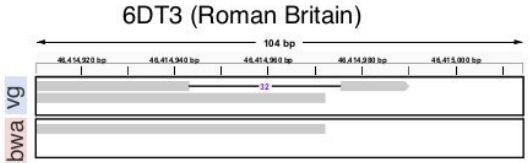
Research | Open Access | Published: 17 September 2020

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







Using variation graphs to observe CCR5-delta in ancient samples

Rui Martiniano

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

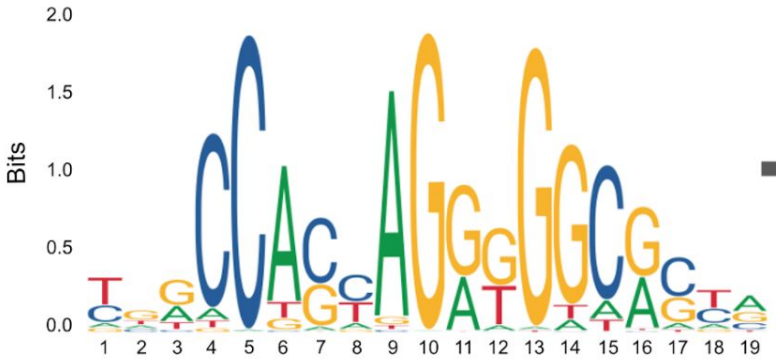
RESEARCH ARTICLE

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)

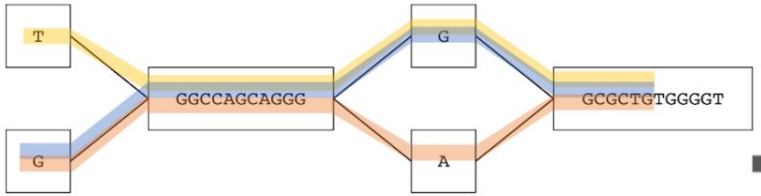


Processed motif PWM

S1	G	G	G	C	C	A	G	C	A	G	G	G	G	C	G	C	T	G
S2	T	G	G	C	C	A	G	C	A	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

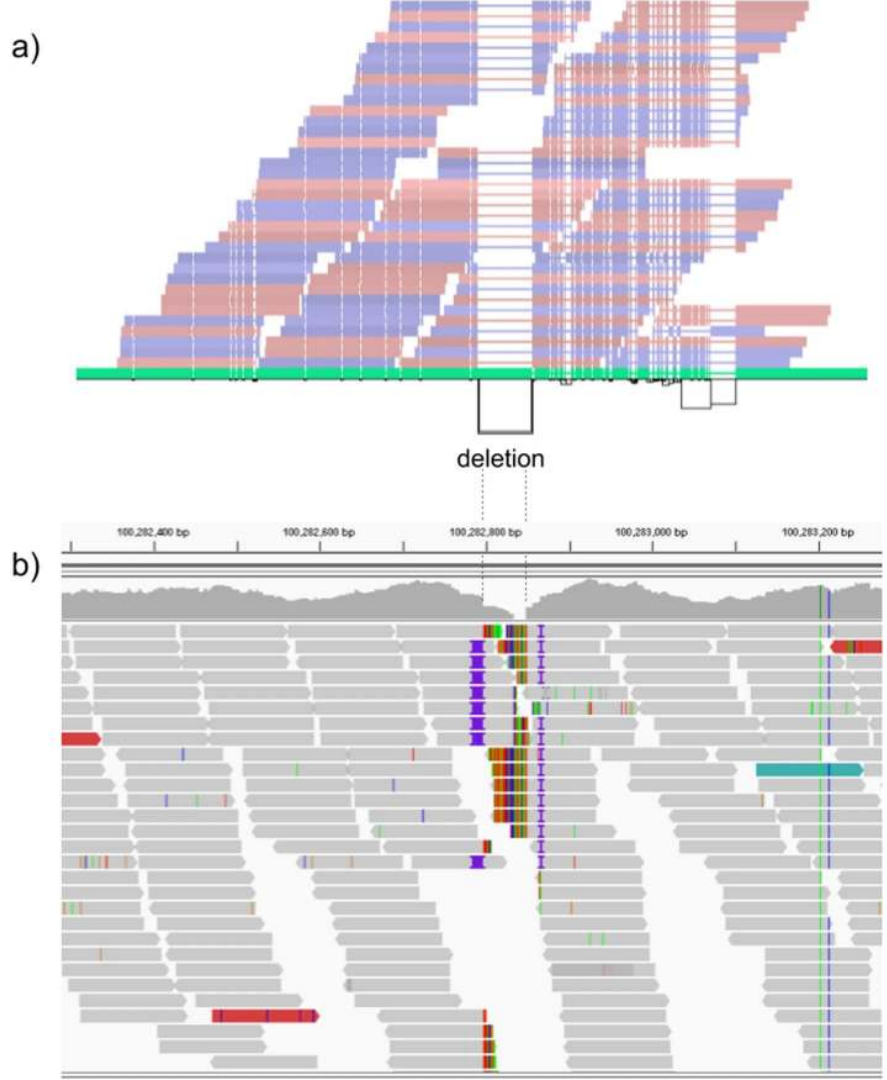
Genome Biology

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



reads vs. graph

reads vs. GRCh38

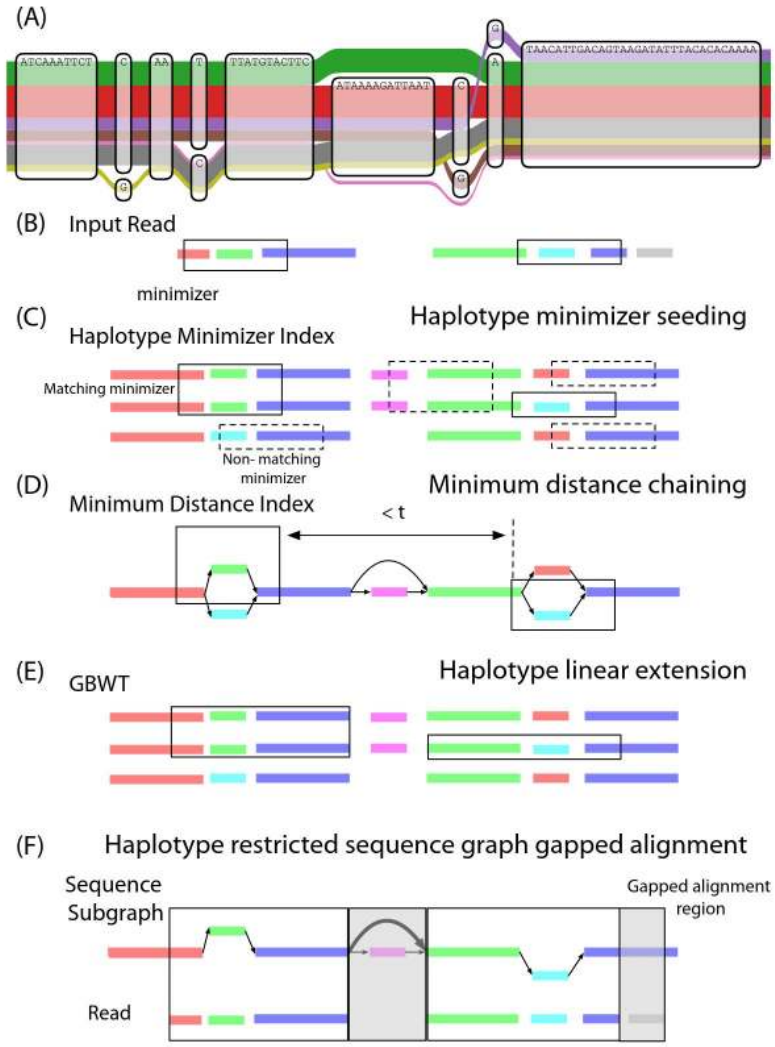
RESEARCH ARTICLE | GENOMICS

Pangenomics enables genotyping of known structural variants in 5202 diverse genomes

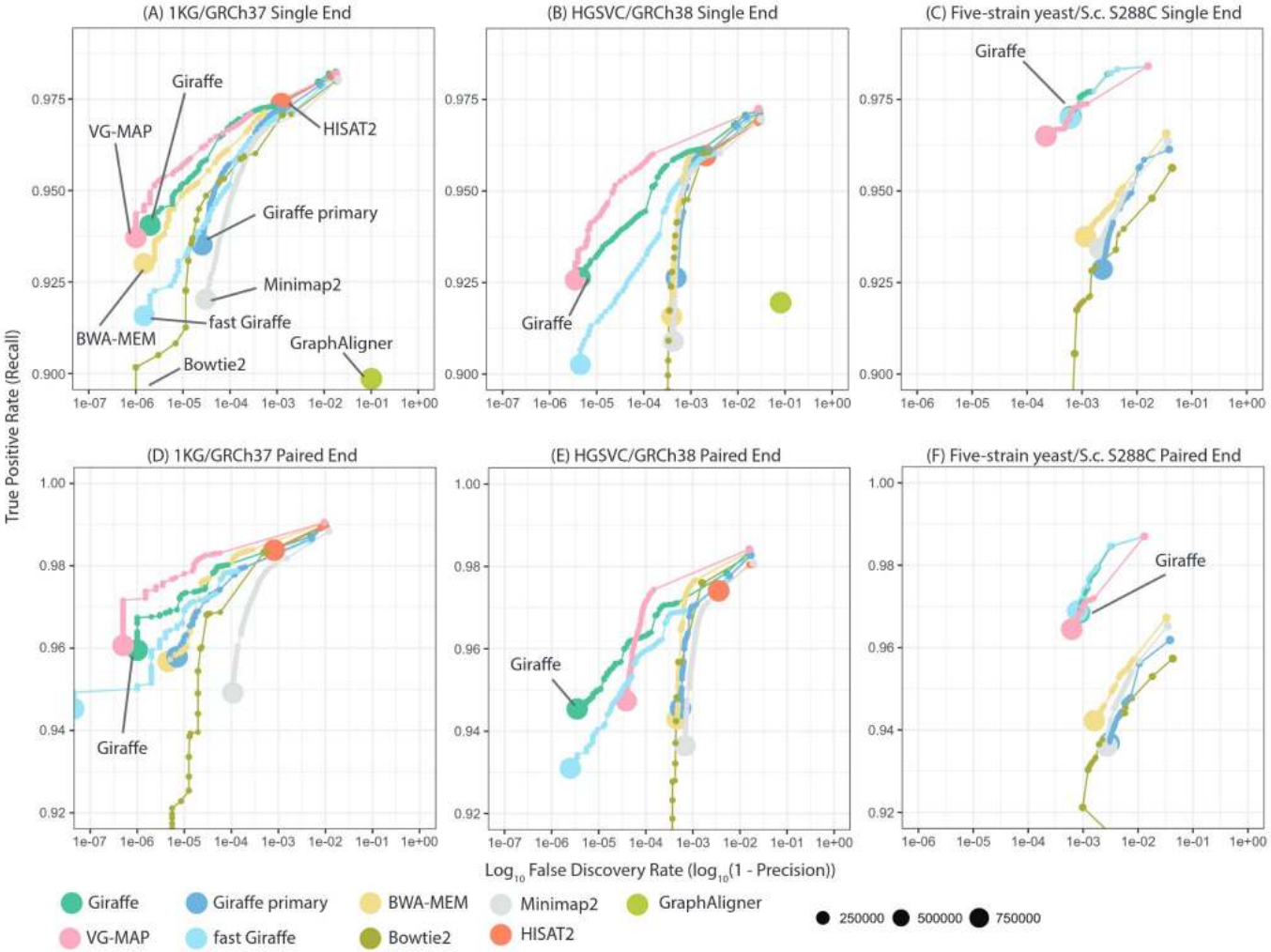
JOUNI SIRÉN, JEAN MONLONG, XIAN CHANG, ADAM M. NOVAK, JORDAN M. EIZENGA, CHARLES MARKELLO, JONAS A. SIBBESEN, GLENN HICKEY, PI-CHUAN CHANG, ANDREW CARROLL, NAMRATA GUPTA, STACEY GABRIEL, THOMAS W. BLACKWELL, AAKROSH RATAN, KENT D. TAYLOR, STEPHEN S. RICH, JEROME I. ROTTER, DAVID HAUSSLER, ERIK GARRISON,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

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

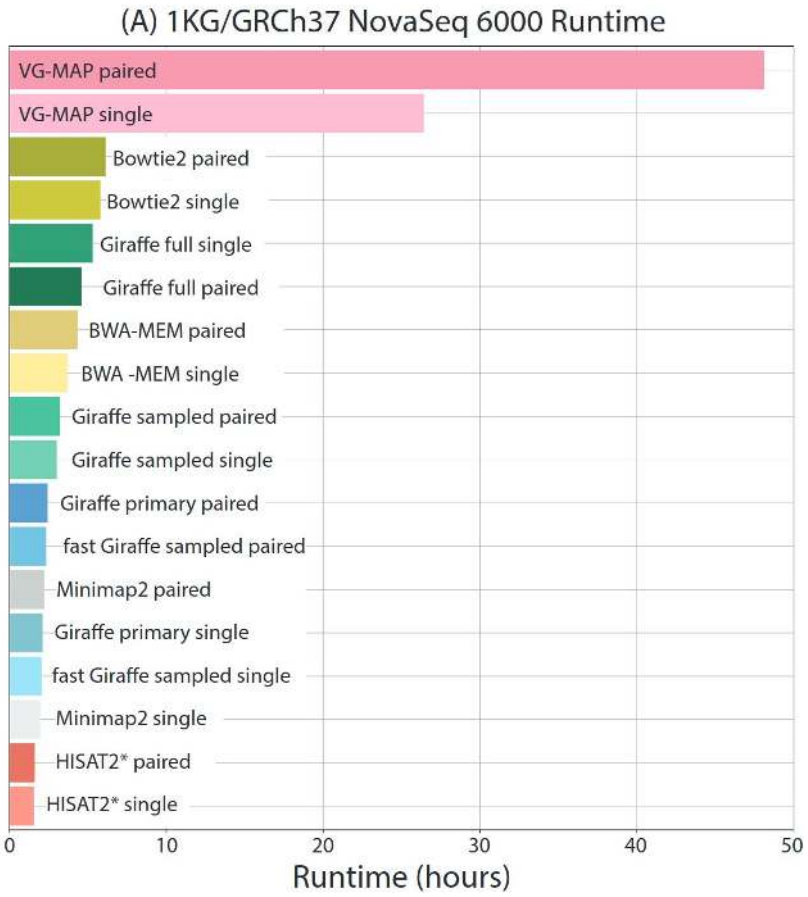
vg giraffe: approach

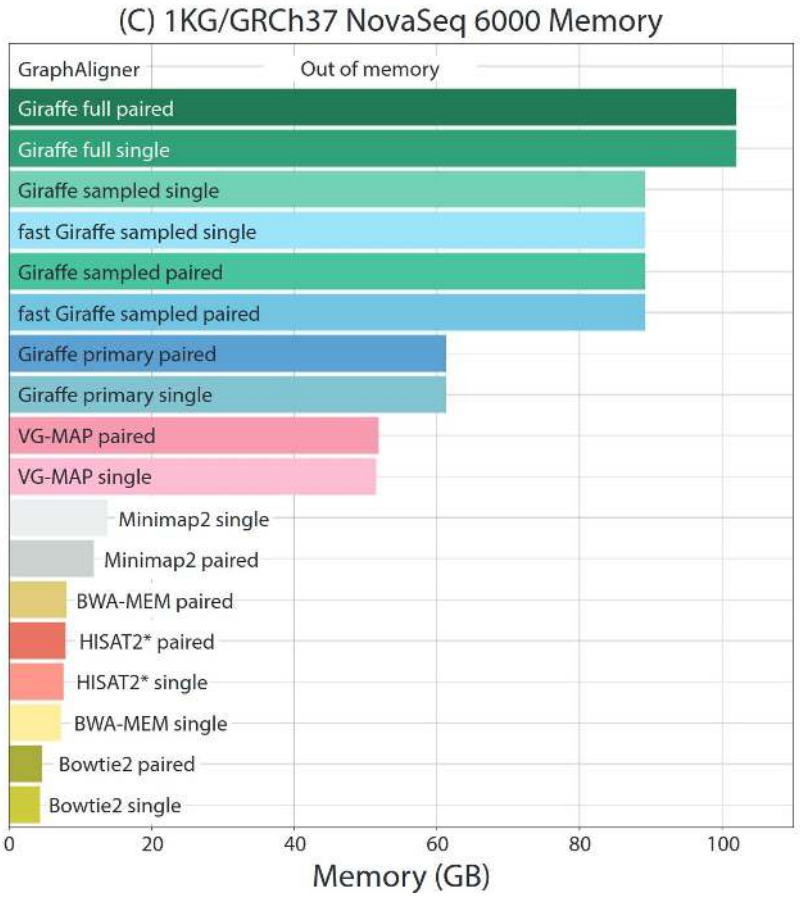


vg giraffe is accurate enough

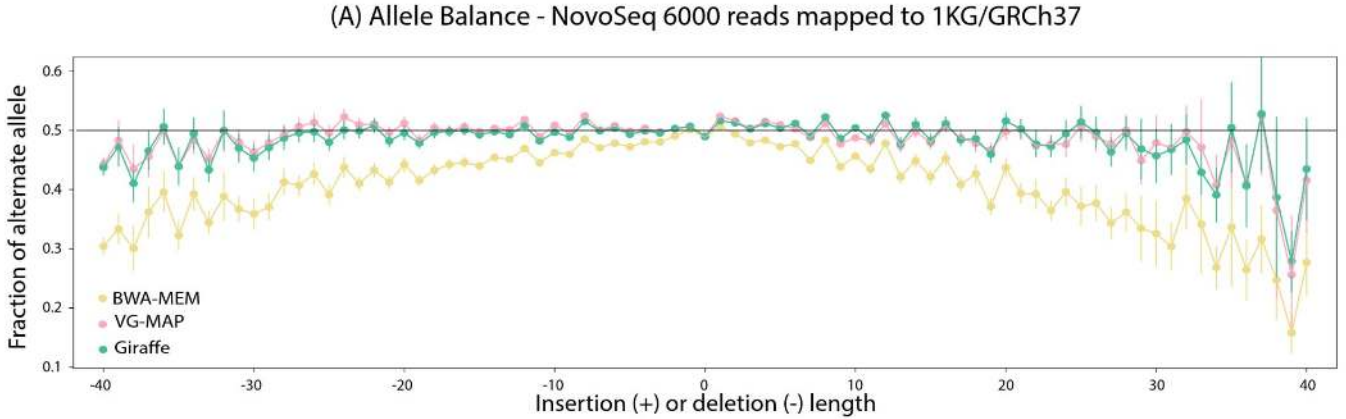


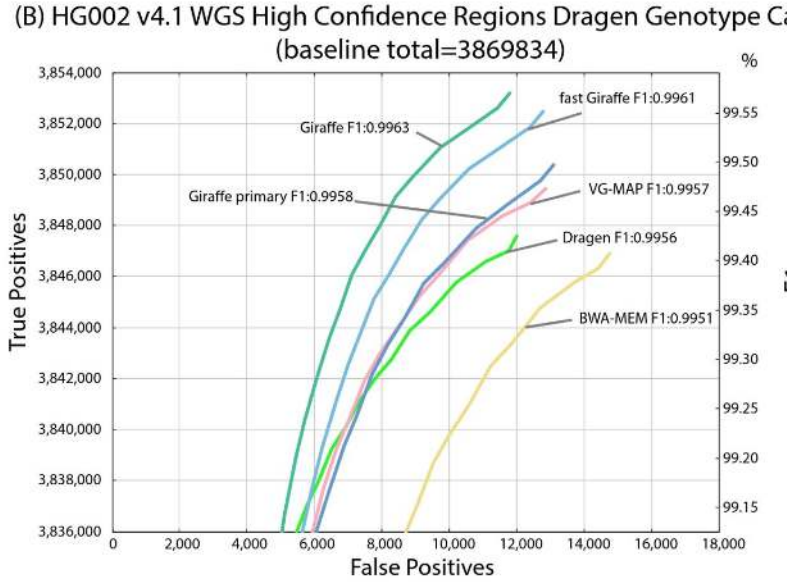
vg giraffe is *very fast*

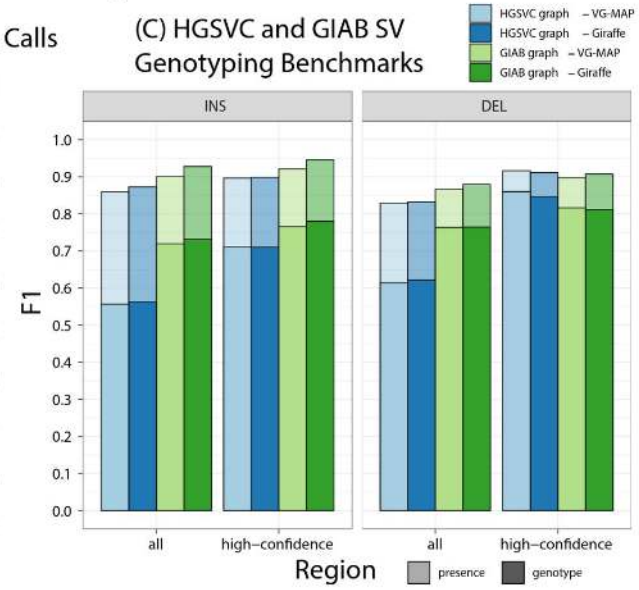




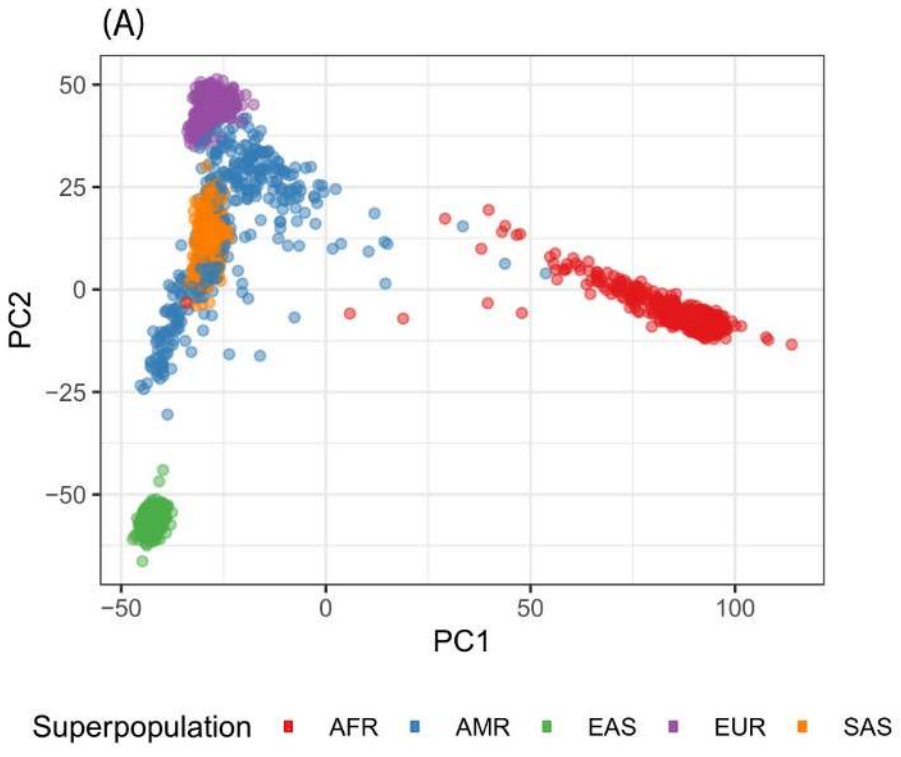
vg giraffe improves variant calling

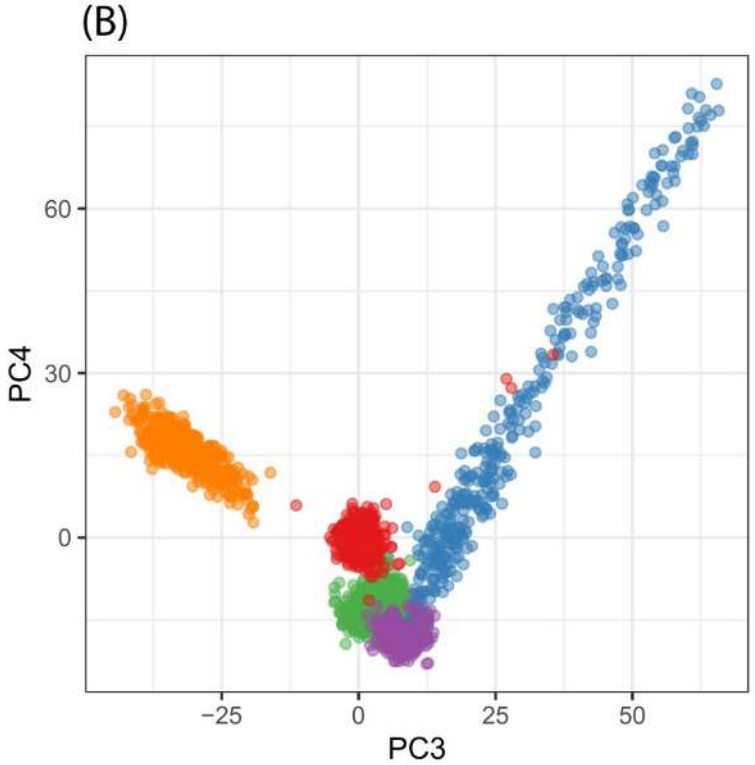






vg giraffe lets us scale: PCA from SVs in 5k genomes





Building the human pangenome

Article

A draft human pangenome reference

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

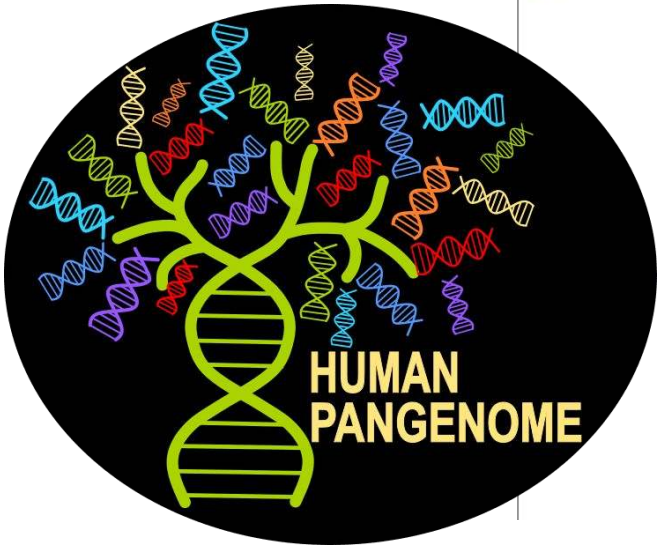
Received: 9 July 2022

Accepted: 28 February 2023

Published online: 10 May 2023

Open access

Check for updates

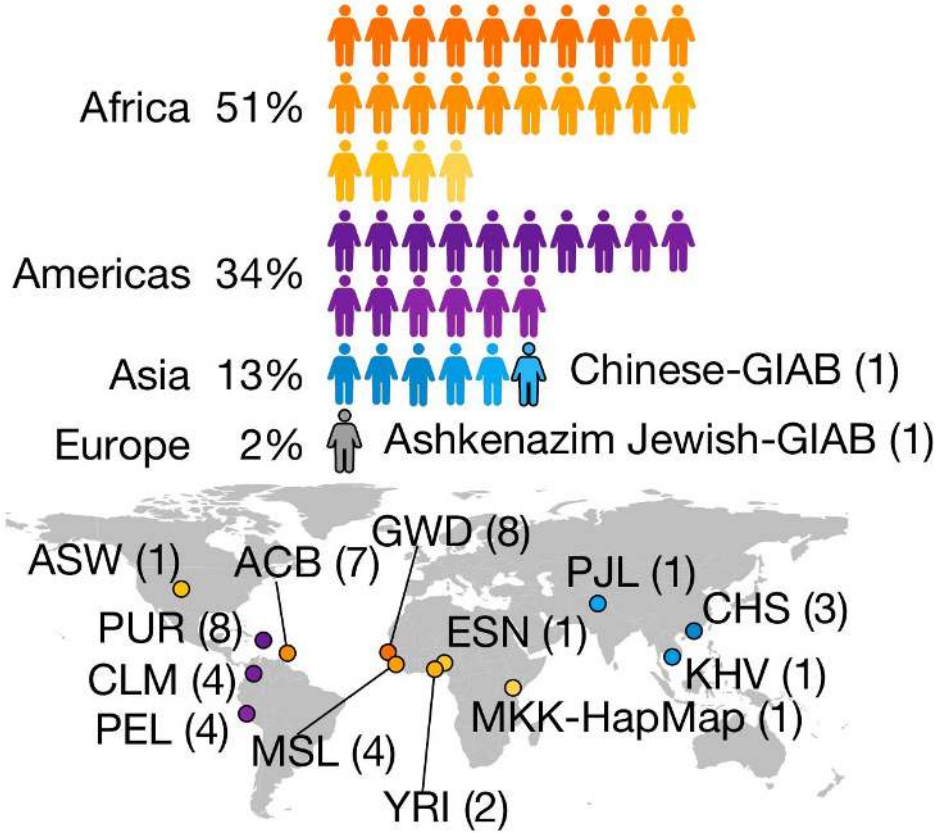


Wen-Wei Liao[1,2,3,60], Mobin Asri[4,60], Jana Ebler[5,6,60], Daniel Doerr[5,6], Marina Haukness[4], Glenn Hickey[4], Shuangjia Lu[1,2], Julian K. Lucas[4], Jean Monlong[4], Haley J. Abel[7], Silvia Buonaiuto[8], Xian H. Chang[4], Haoyu Cheng[9,10], Justin Chu[9], Vincenza Colonna[8,11], Jordan M. Eizenga[4], 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]✉

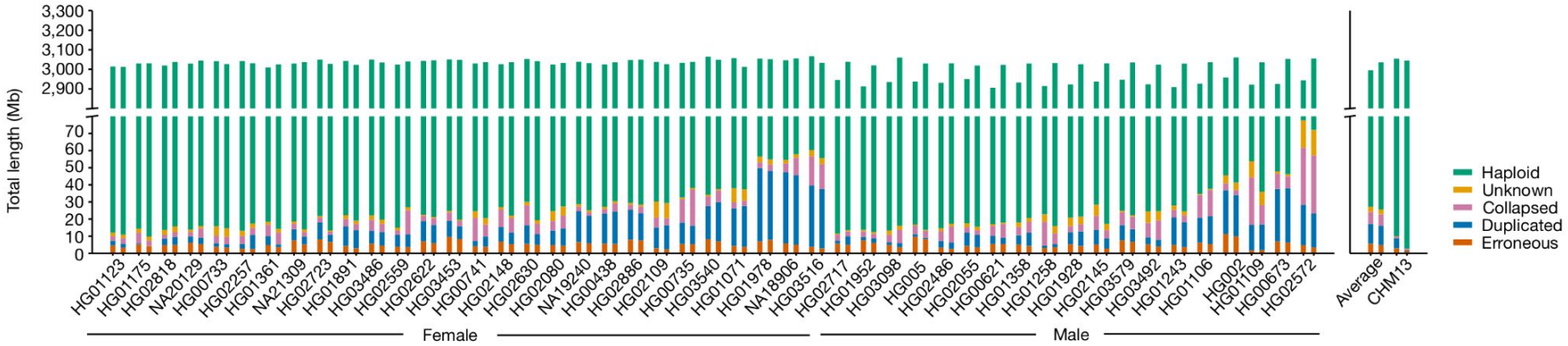
Draft pangenome composition

Sample selection was constrained by:

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



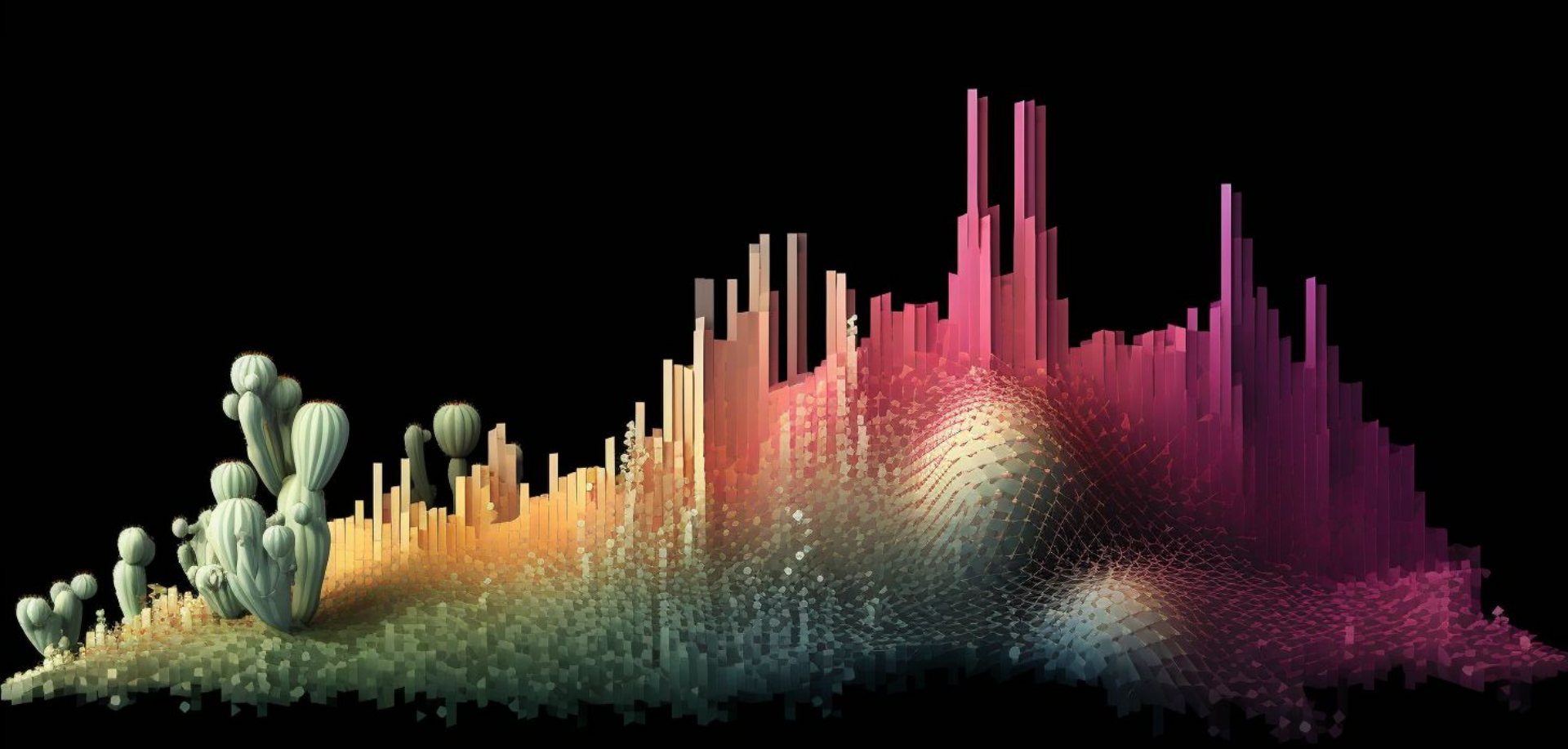
Amazing assemblies approach reference quality



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 7 pangenome (reference) graphs…

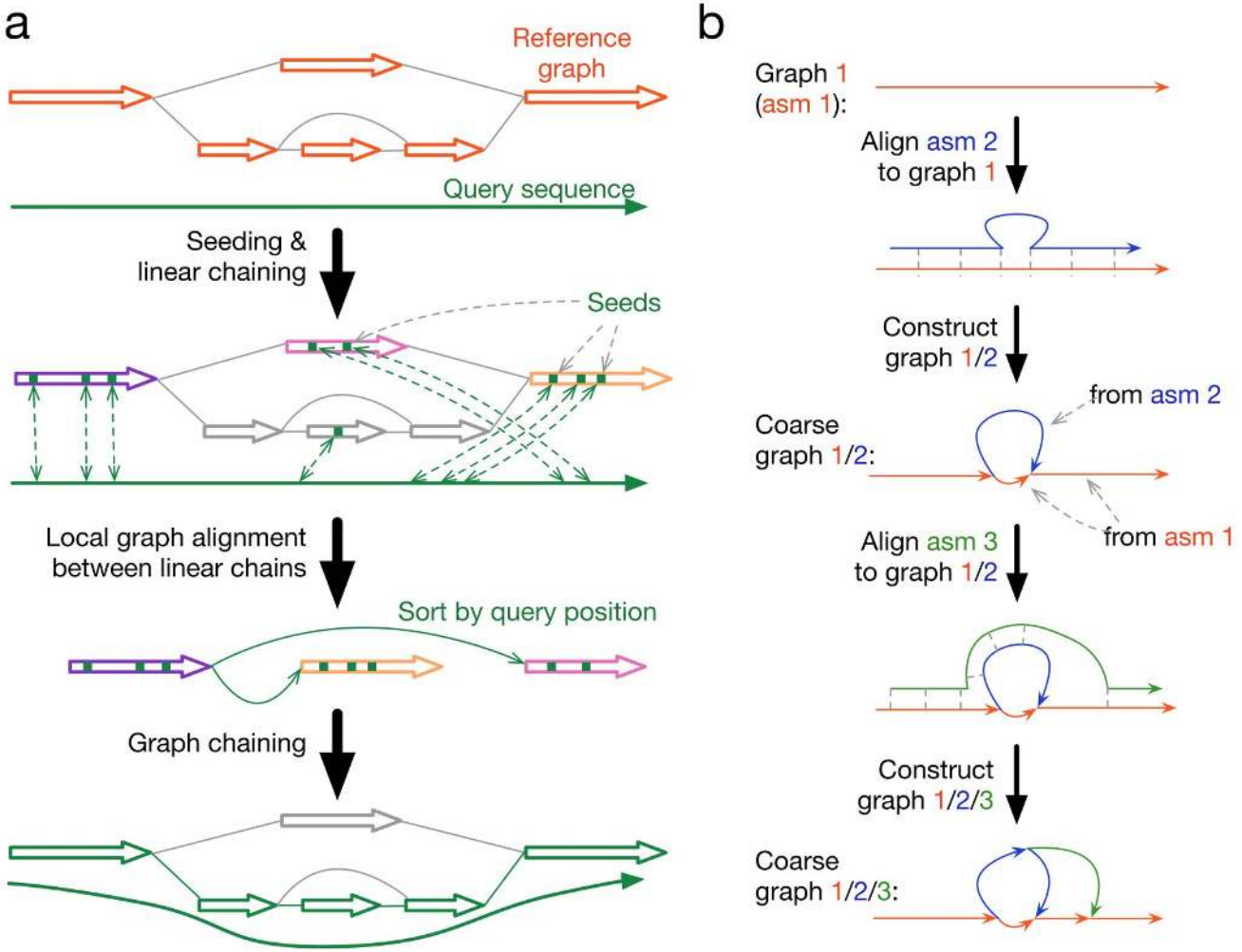
Minigraph

Heng Li (author of bwa, samtools, minimap2, miniprot, and many other tools).

Idea: build the graph from the reference, progressively.

Only SVs > 50bp.

Uses a model similar to POA, but capable of dealing with structural variation, and using the minimizer chaining concept from minimap2.



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

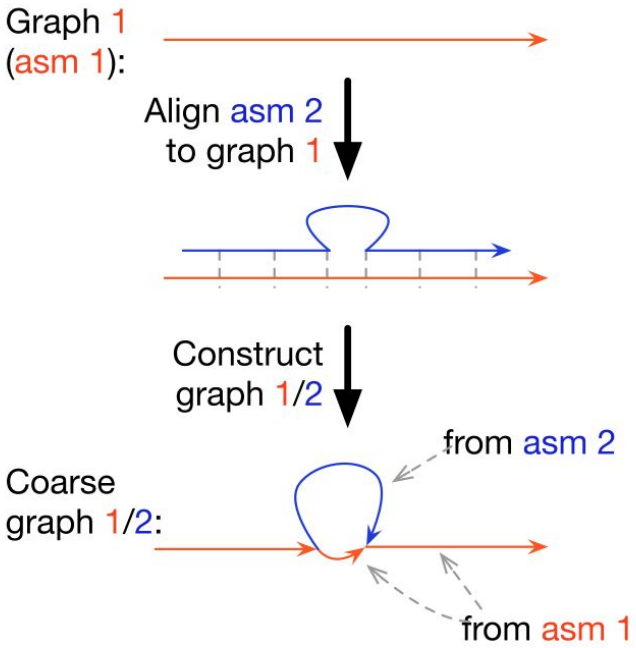
Minigraph

Heng Li (author of bwa, samtools, minimap2, miniprot, and many other tools).

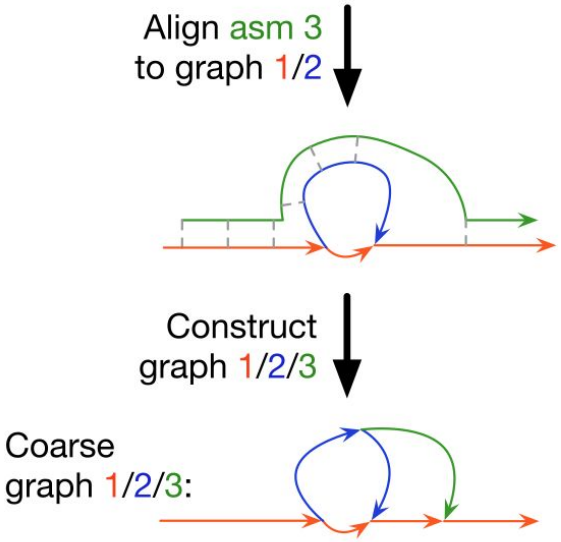
Idea: build the graph from the reference, progressively.

Only SVs > 50bp.

Uses a model similar to POA, but capable of dealing with structural variation, and using the minimizer chaining concept from minimap2.



for each new assembly…



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

Minigraph-Cactus

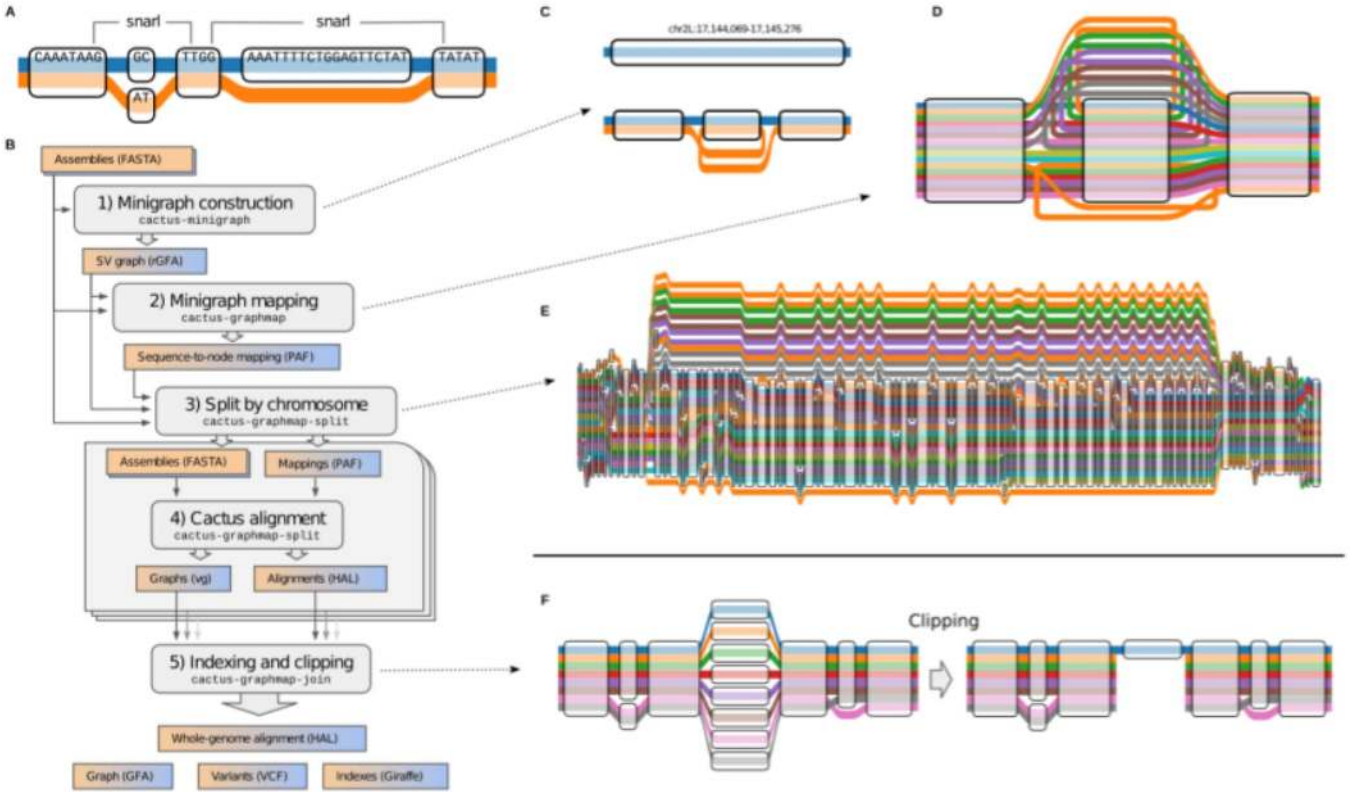
https://doi.org/10.1038/s41587-023-01793-w

Add base-level variation onto the minigraph.

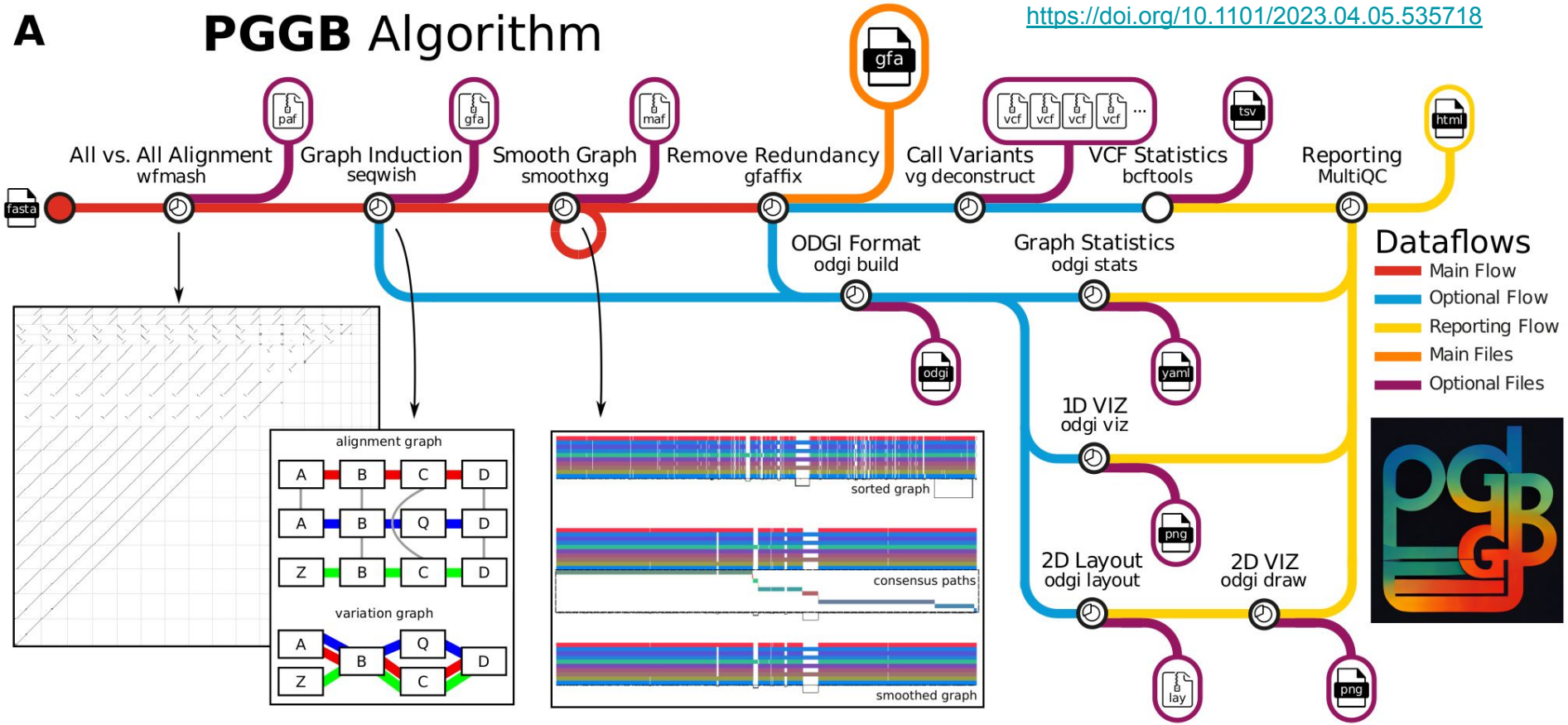
Uses tooling from Cactus to drive local alignment across pieces of the graph.

Makes a variation graph, but the graph has the structure of the minigraph model with some differences.

Clipping is used to make the graph easy to align to with vg giraffe.



https://doi.org/10.1101/2023.04.05.535718



“Hard-mode” pangenomes.

All-to-all = quadratic. ^That’s an *exascale* matrix of chr6 in all great apes.

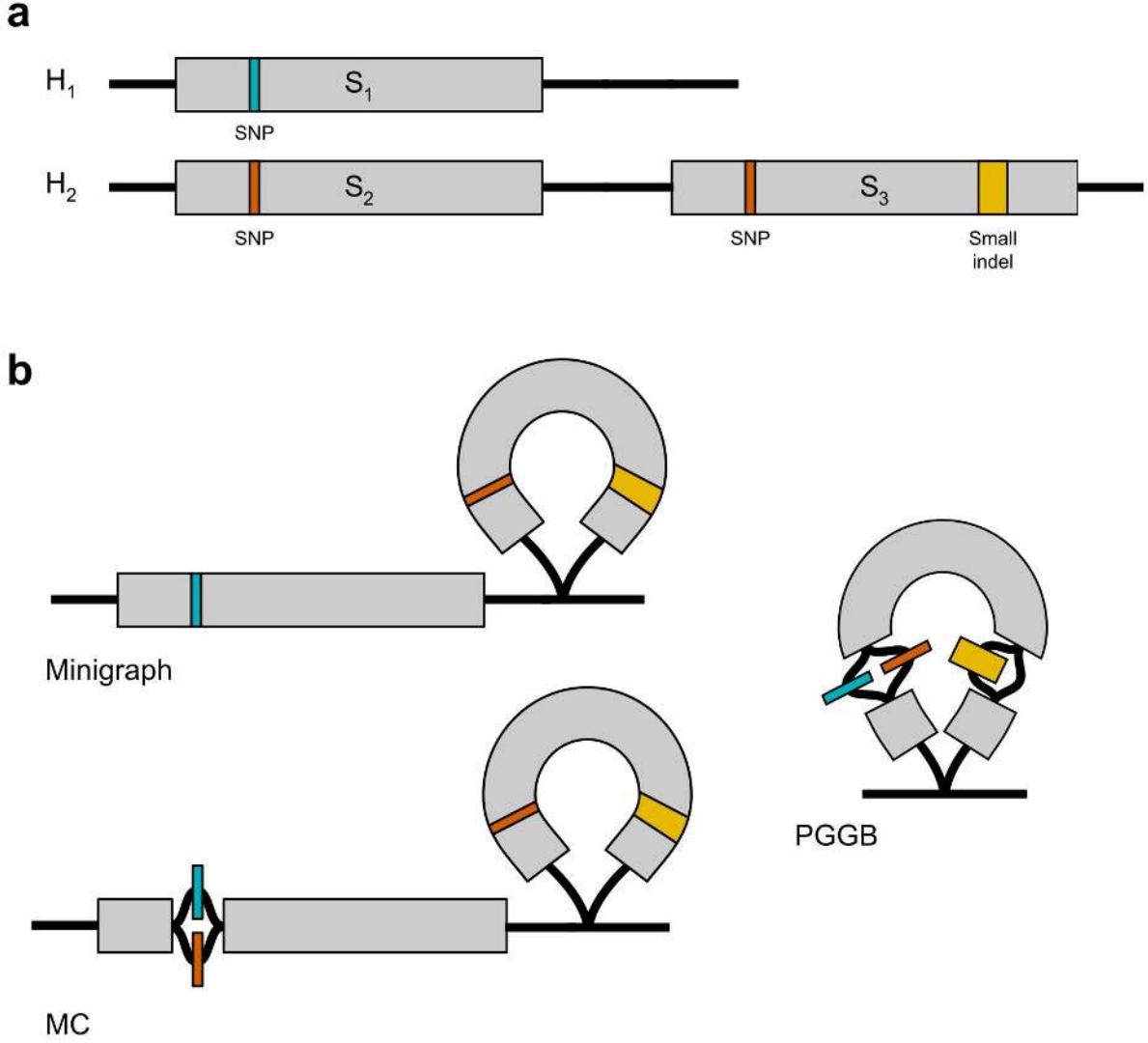
Key conceptual differences between HPRC pangenome construction methods

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

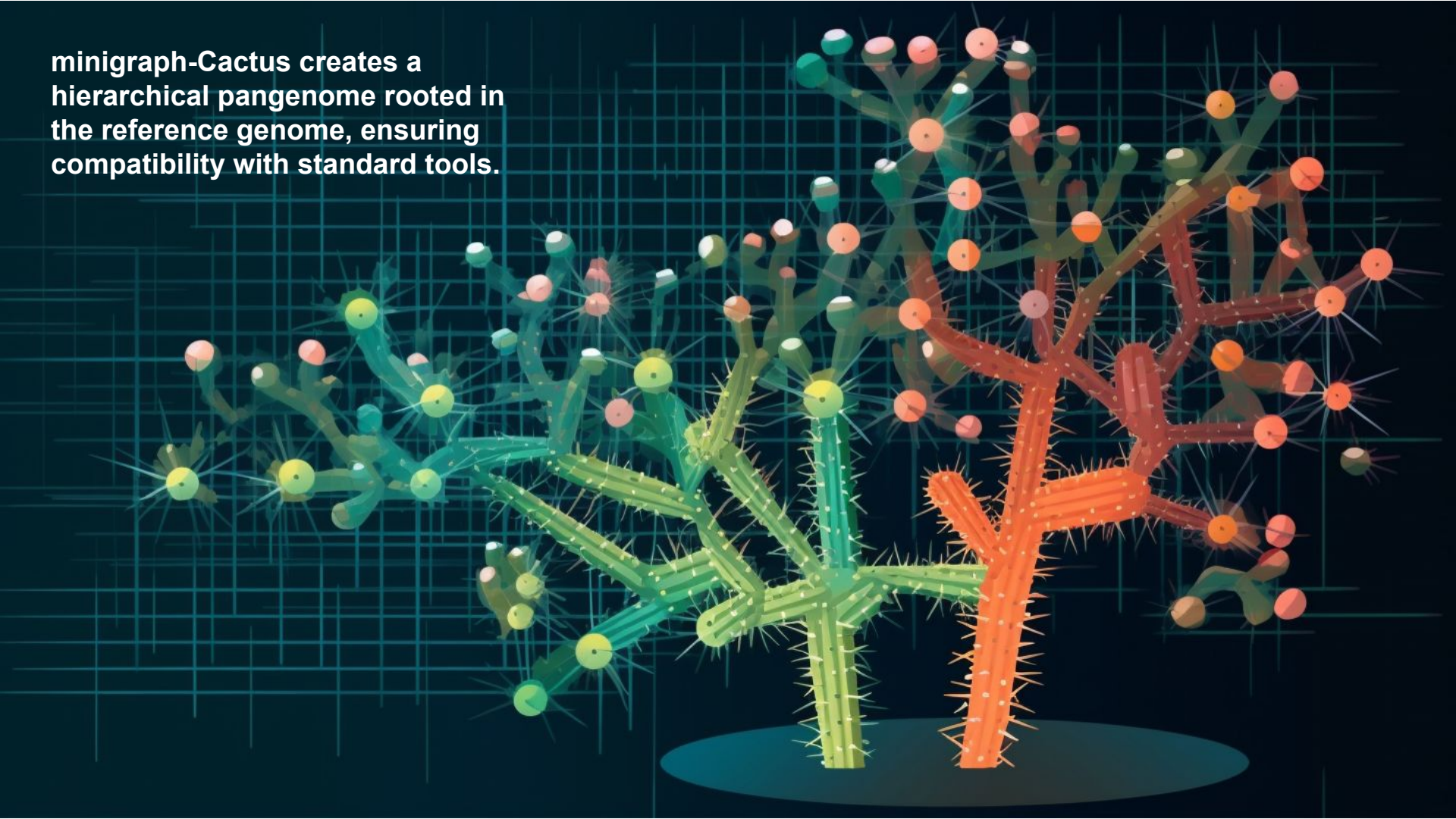
minigraph-cactus: add SNPs, clean up the breakpoints, useful for alignment, one reference.

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

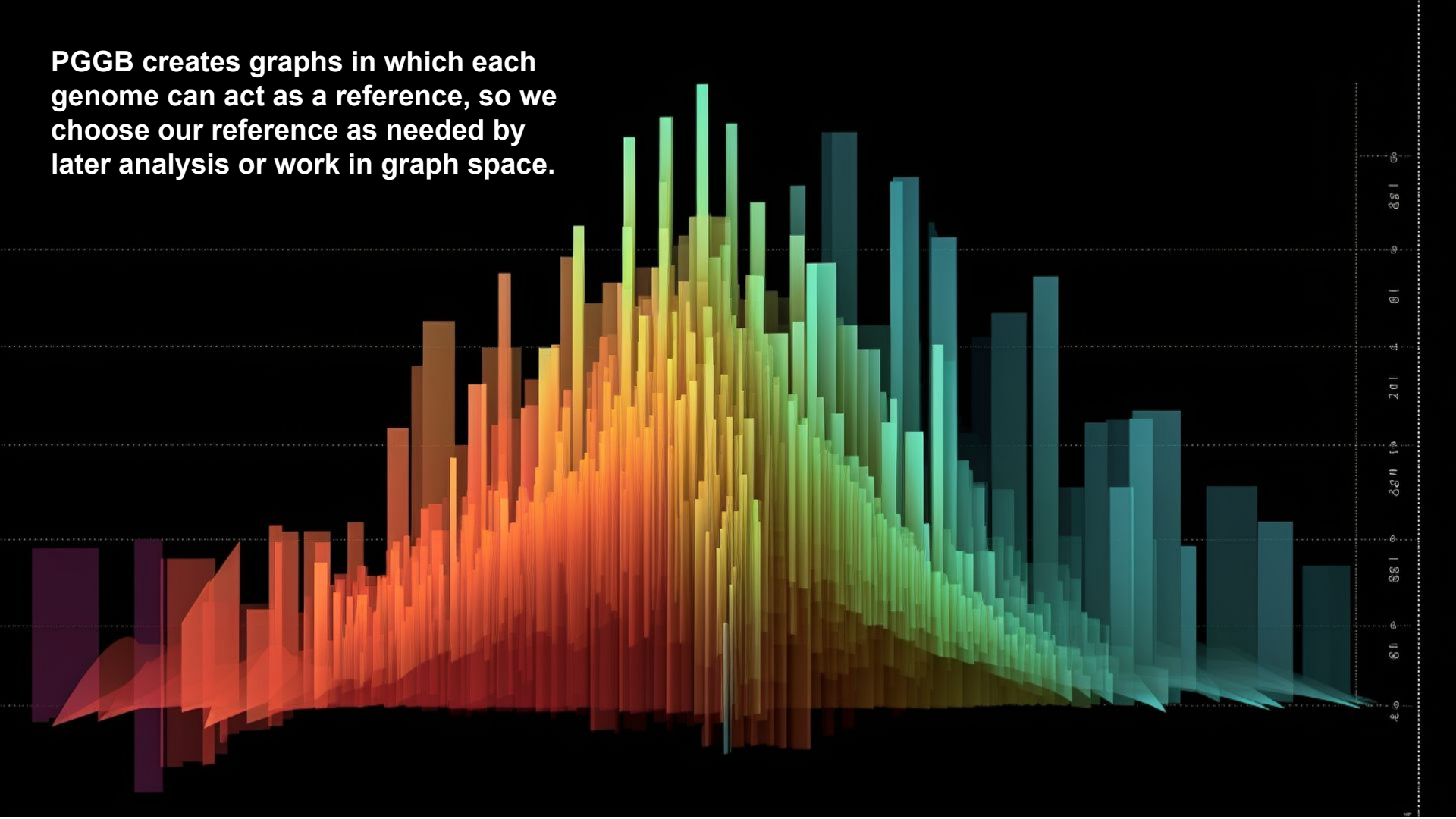
“Collapse” in high-copy repeats →



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

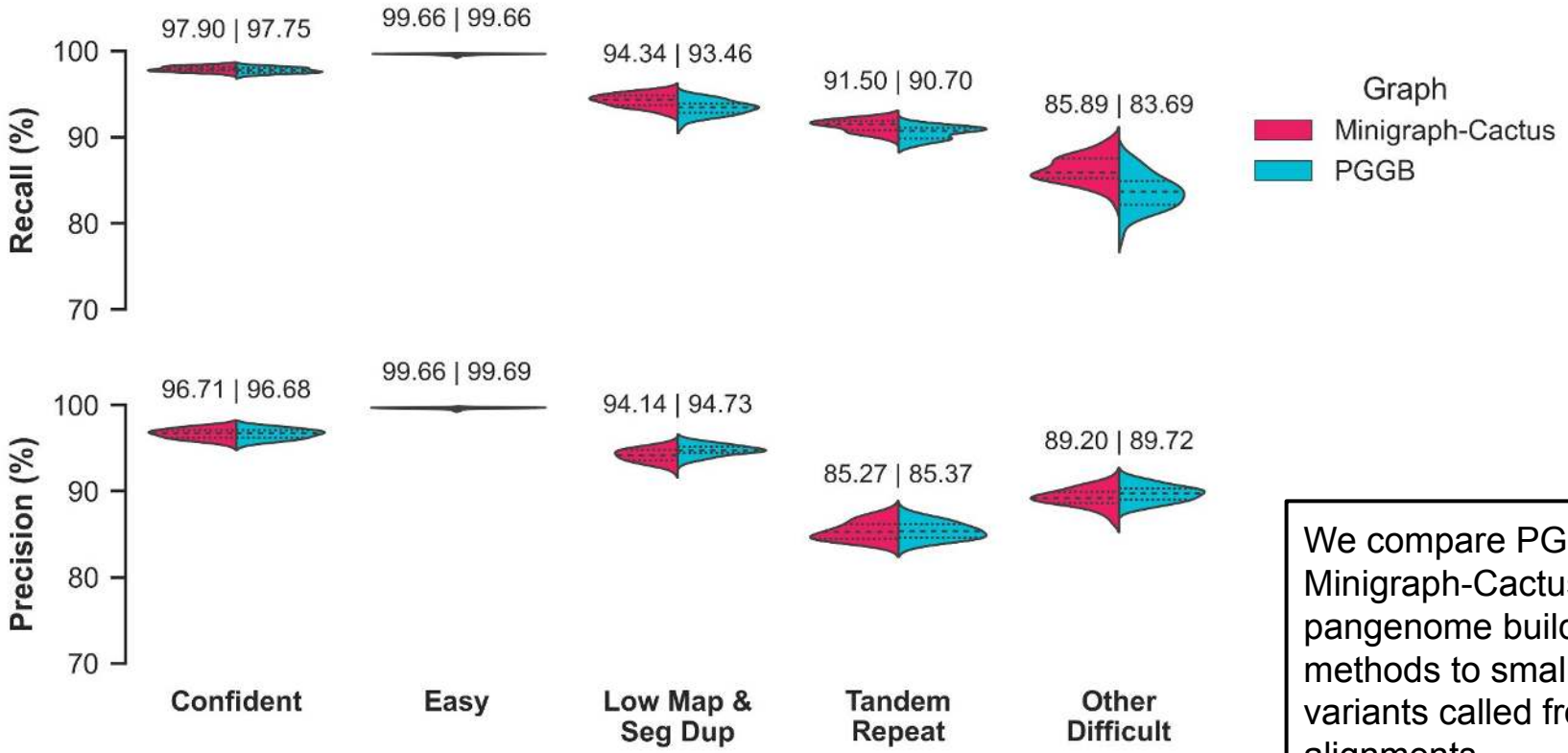


PGGB creates graphs in which each genome can act as a reference, so we choose our reference as needed by later analysis or work in graph space.



Multiple graph building methods show consistent quality

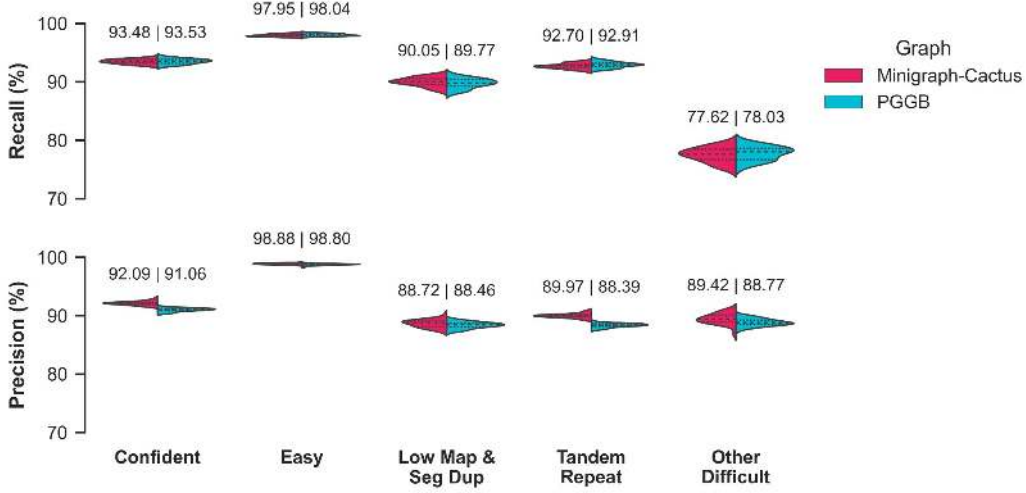
* variants extracted from graphs with vg deconstruct

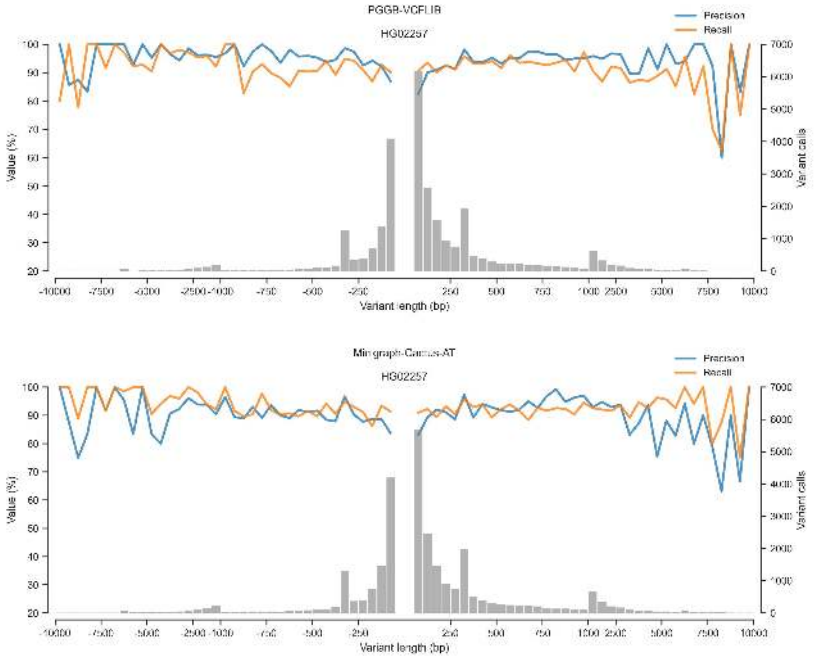


We compare PGGB and Minigraph-Cactus pangenome building methods to small variants called from HiFi alignments.

Wen-Wei Liao

The graphs accurately characterize structural variants





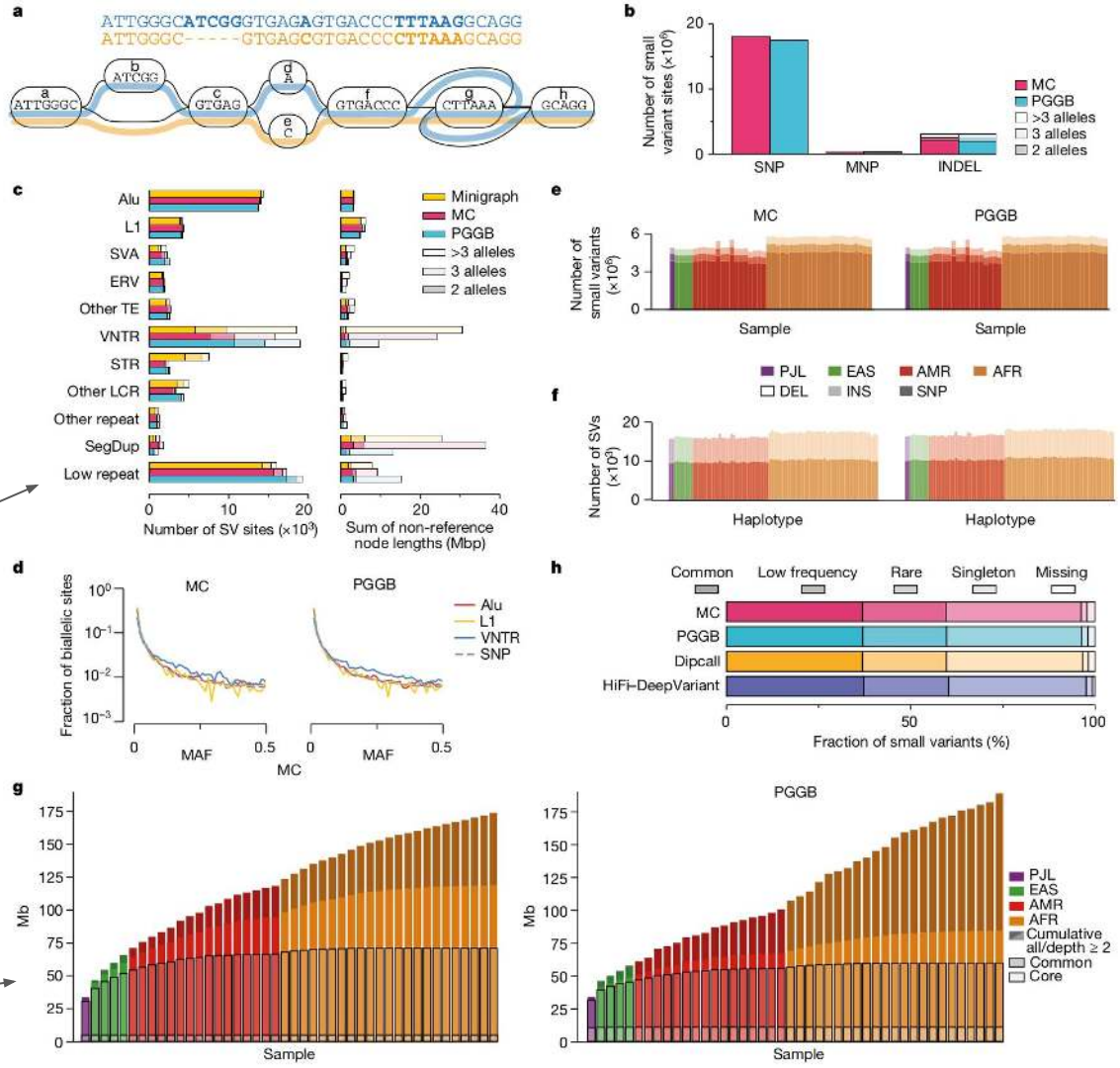
Comparing to consensus calls made by many reference-based SV callers.

Wen-Wei Liao

And they show very similar pictures of the pangenome

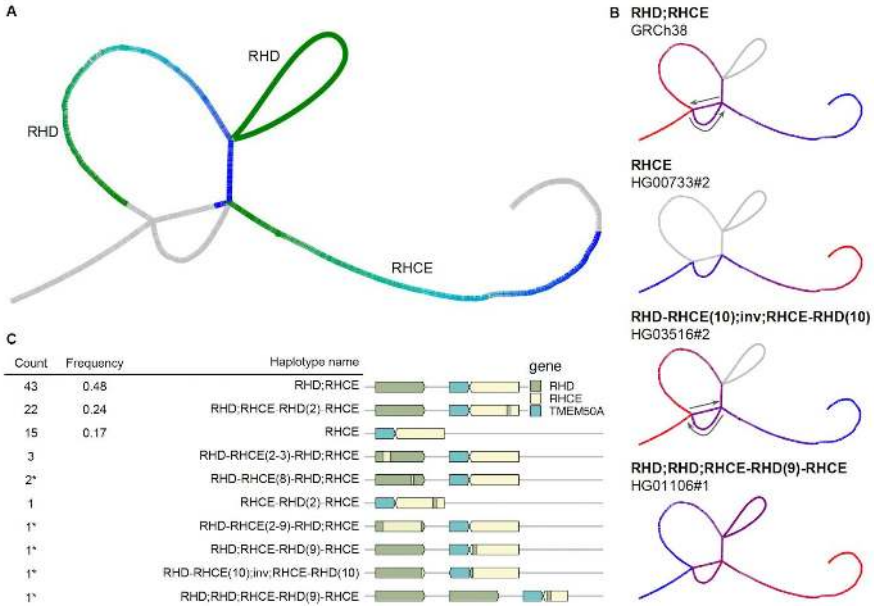
Major differences are in repetitive sequences, which tend to collapse in the PGGB graph and expand in the MC graph.

We also look at pangenome growth via permutations, a nod to old-school pangenomics (g).

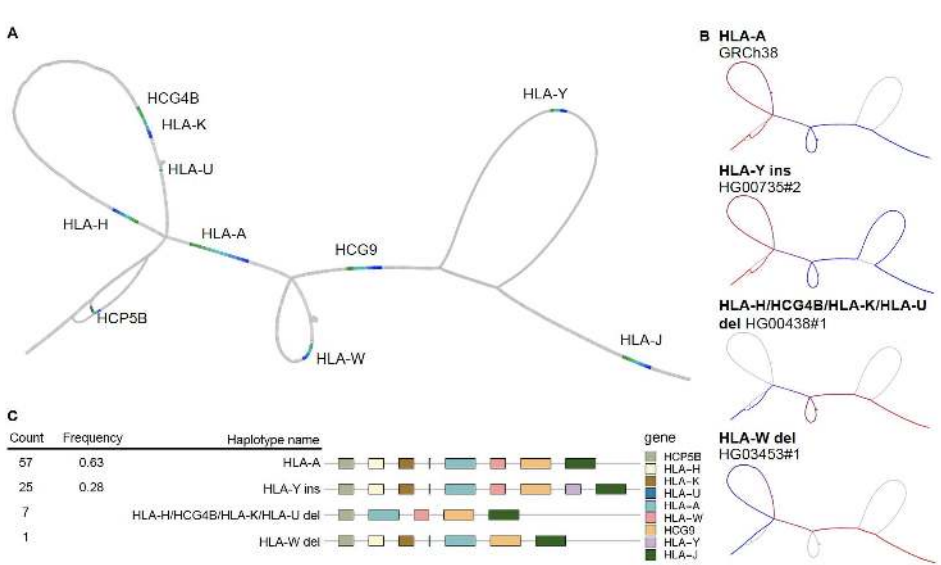


Understanding the human pangenome

Highly polymorphic medically-relevant loci



View of locus that defines the Rh blood group system in the *minigraph-cactus* graph.

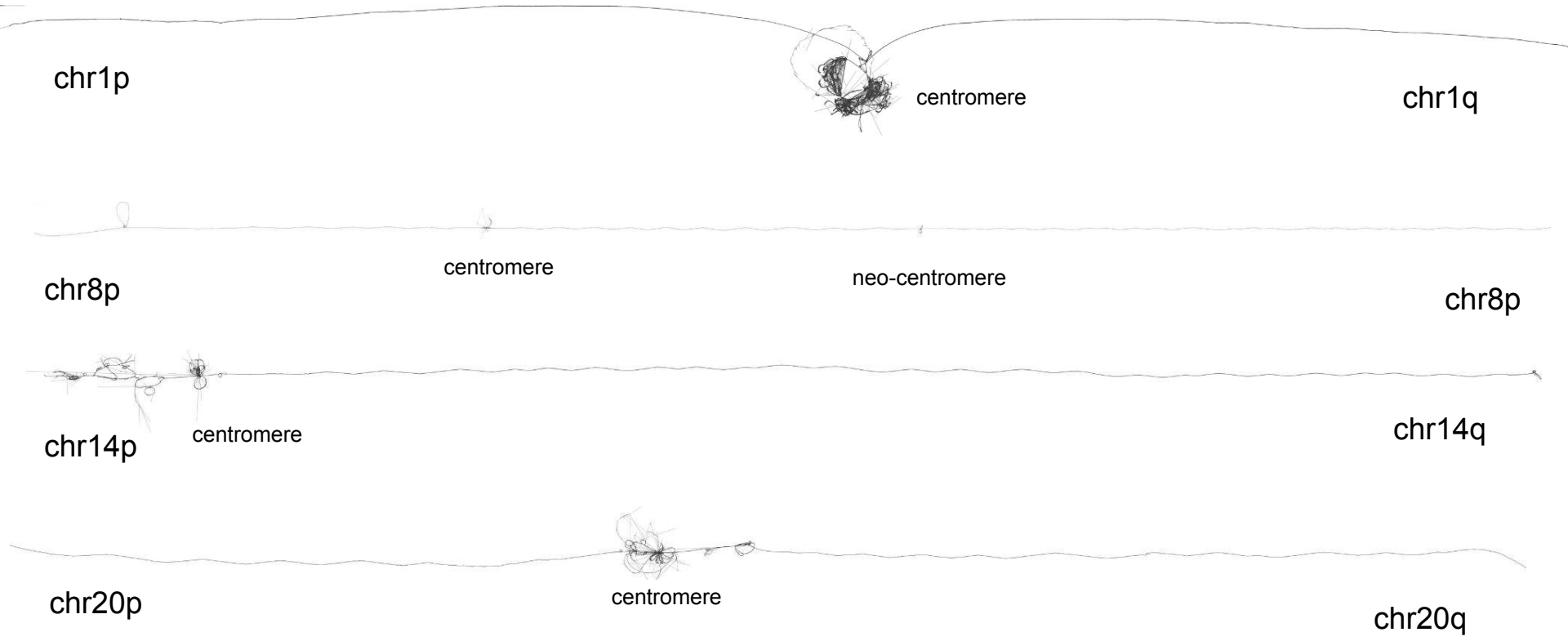


View of HLA/MHC in the *PGGB* graph

Shuangjia Lu

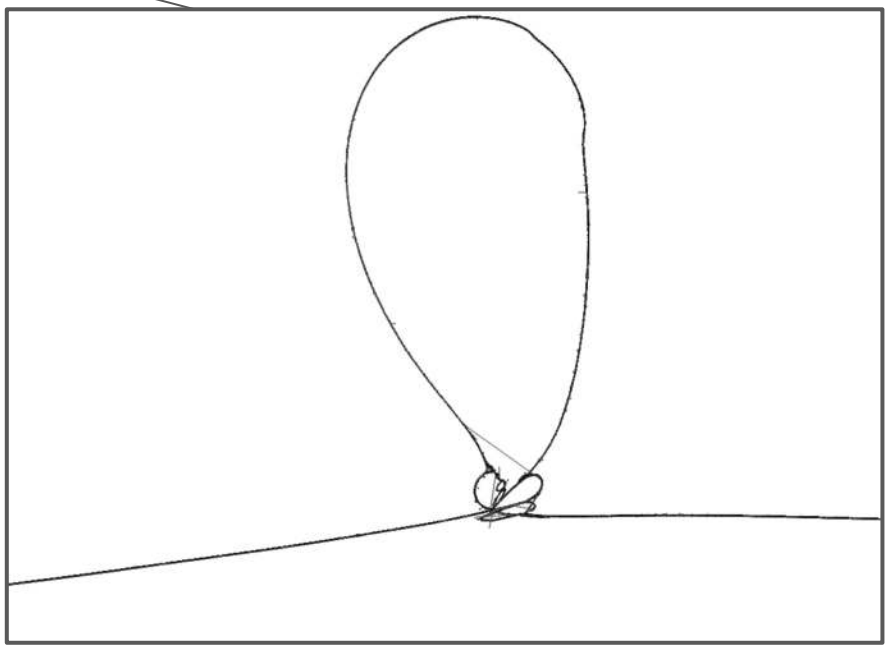
PGGB lets us look at *all* parts of the pangenome

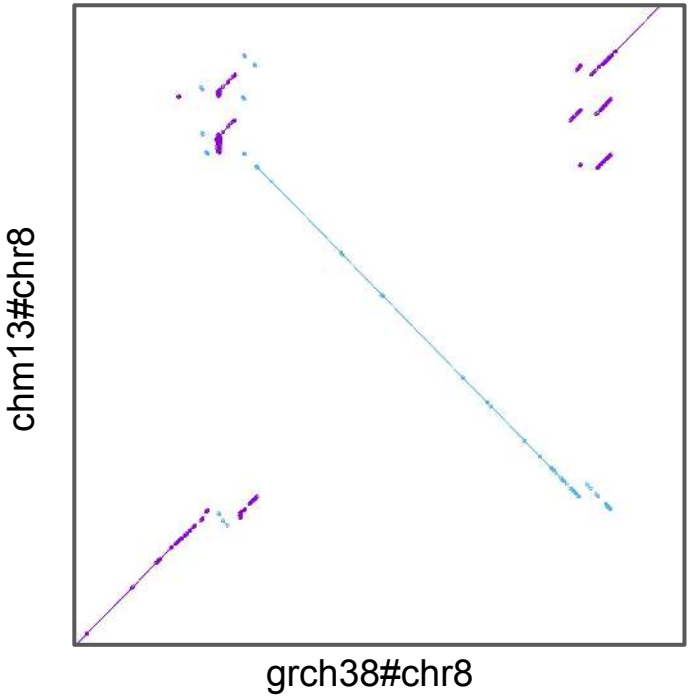
(learned 2D visualization of PGGB HPRC chromosomes)

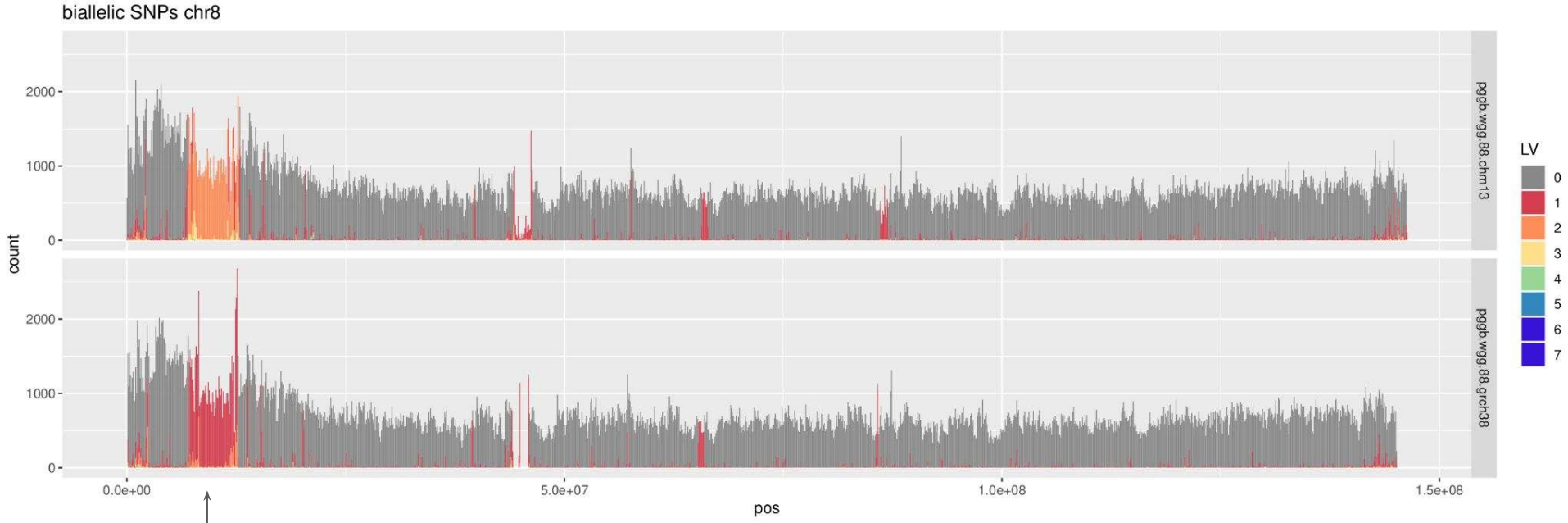


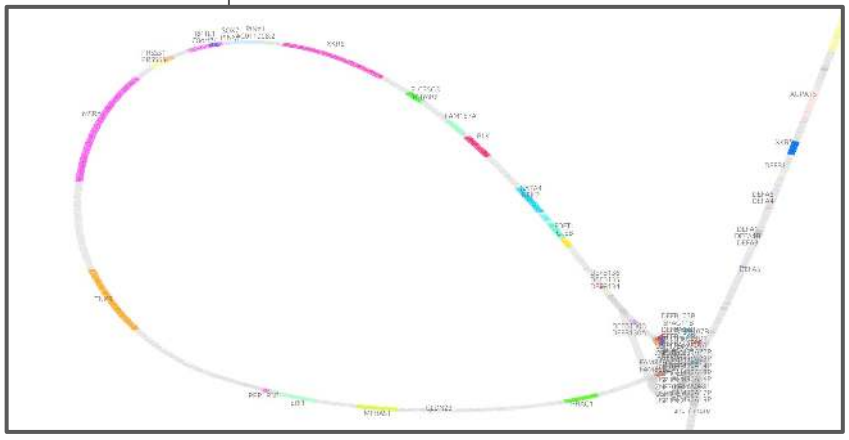
chr8's beta-defensin locus

5 Mbp polymorphic inversion flanked by clusters of beta-defensin (~antibiotic) genes with variable copy number







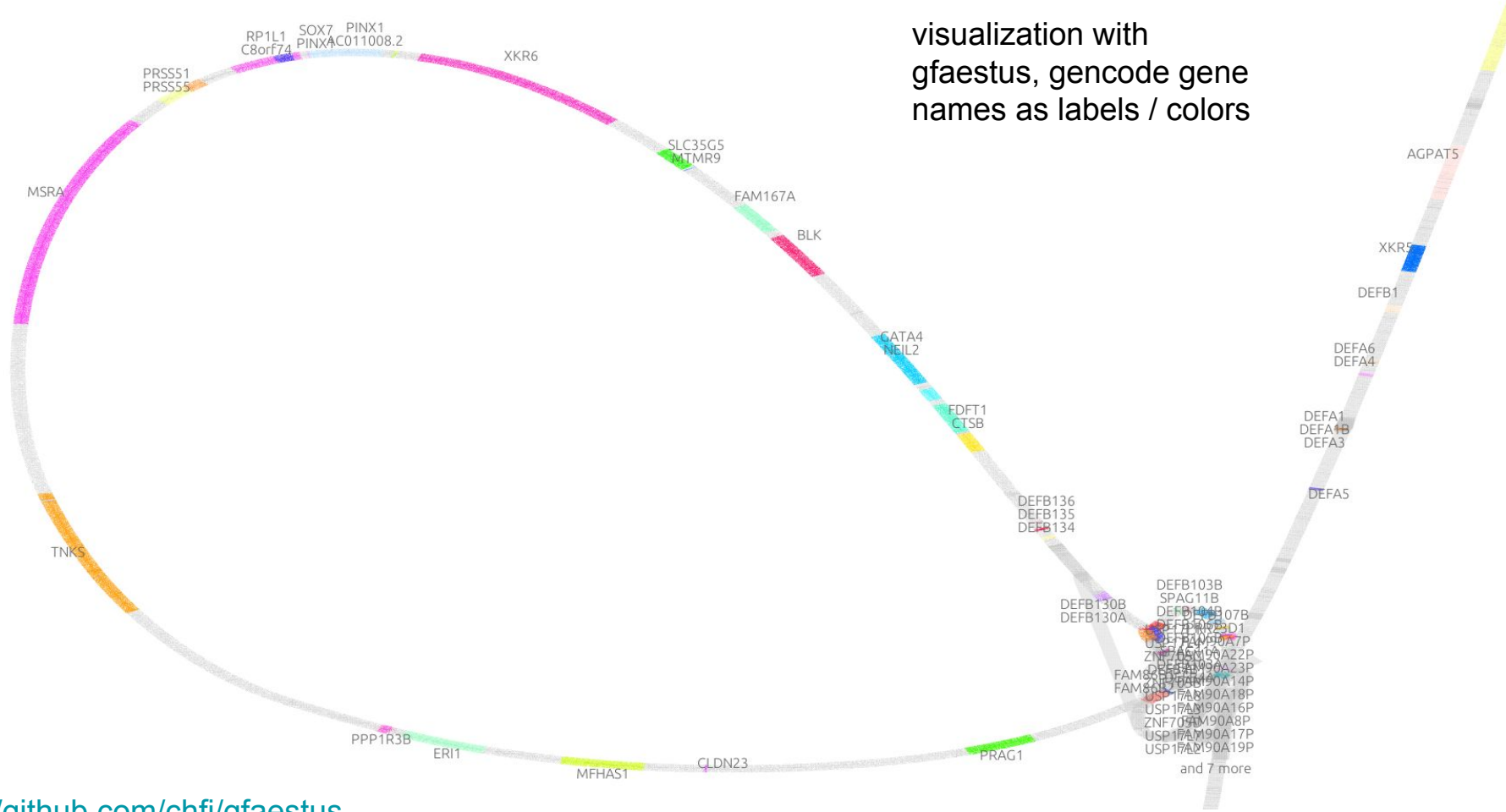


The p-arm of chr8 features a region with one of the highest rates of diversity in the HPRC pangenome. (Here we see 100kb bins.)

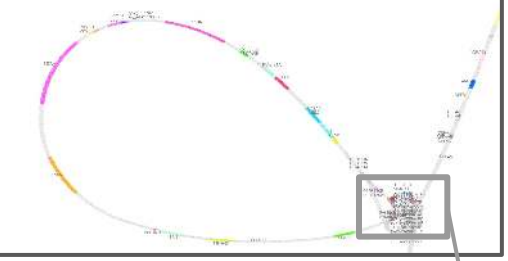
Regionally, it's higher than chr6's MHC, although the MHC has higher peaks of diversity in HLA genes.

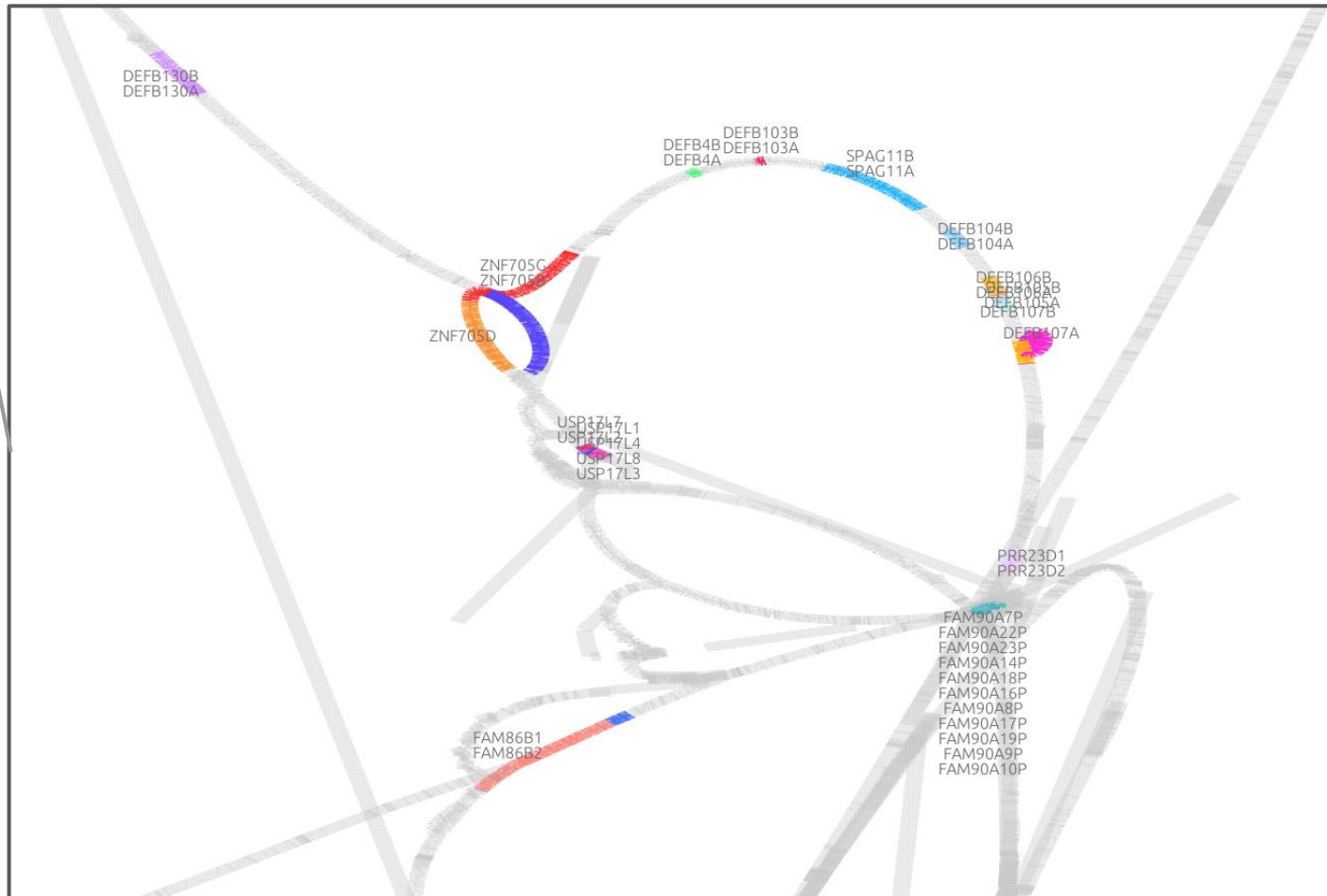
An implication is that this 4.7Mbp polymorphic inversion is driving diversifying processes.

visualization with gfaestus, gencode gene names as labels / colors



https://github.com/chfi/gfaestus

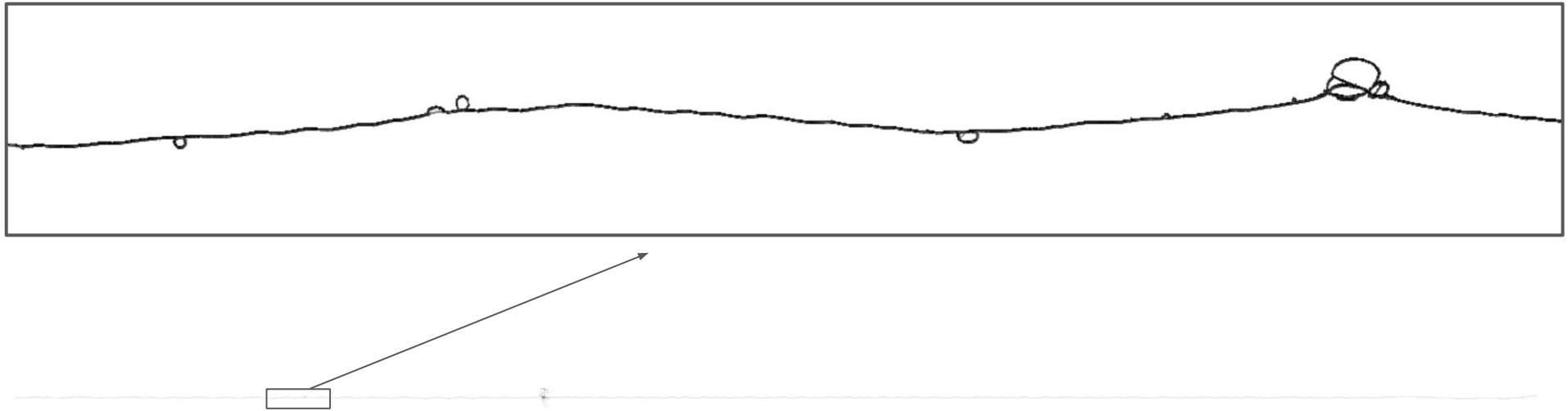




DEFB* genes are beta-defensins. The number of copies of this loop varies ~2-4 in different haplotypes.

FAM90A → a gene of unknown function!

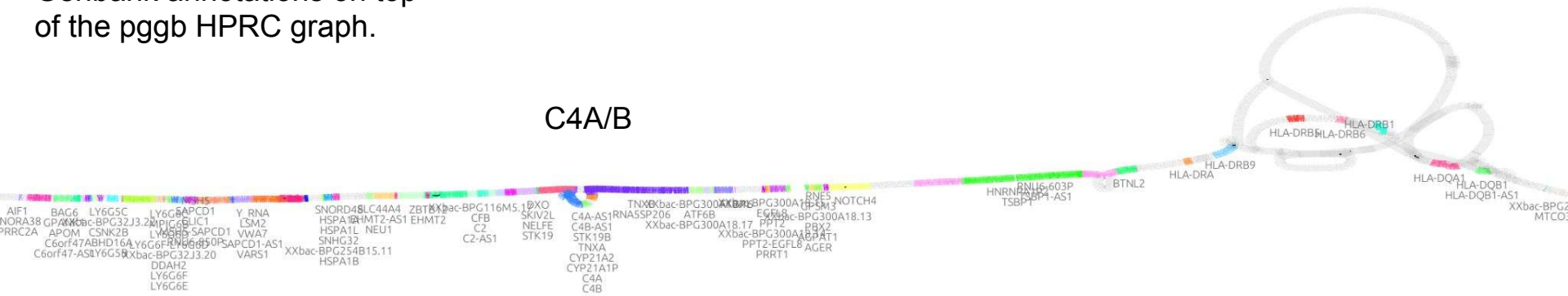
the MHC in chr6



chr6 pggb graph

C4A/B in pggb graph

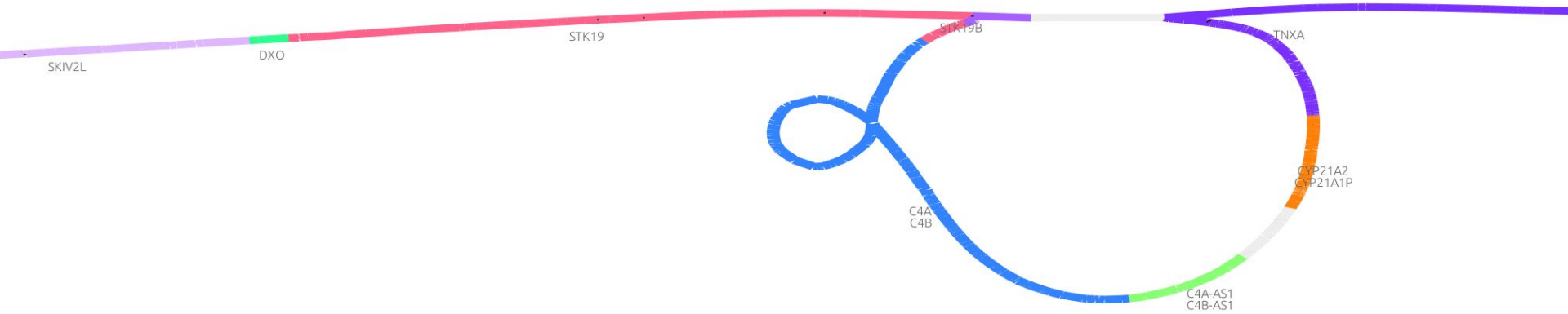
Genbank annotations on top of the pggb HPRC graph.



https://github.com/chfi/gfaestus

Christian Fischer (UTHSC)

C4A/B in pggb graph

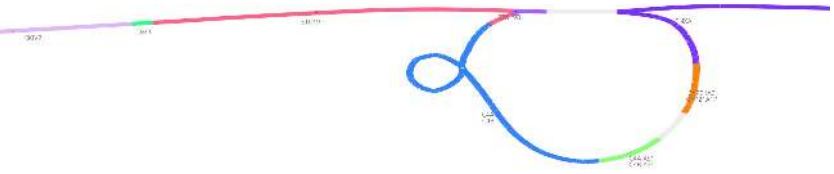


C4A/B

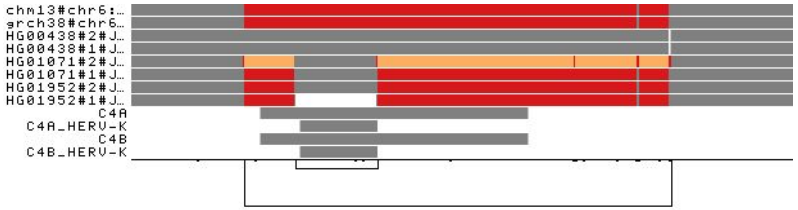
https://github.com/chfi/gfaestus

Christian Fischer (UTHSC)

We learn that genome evolution is often *nonlinear*



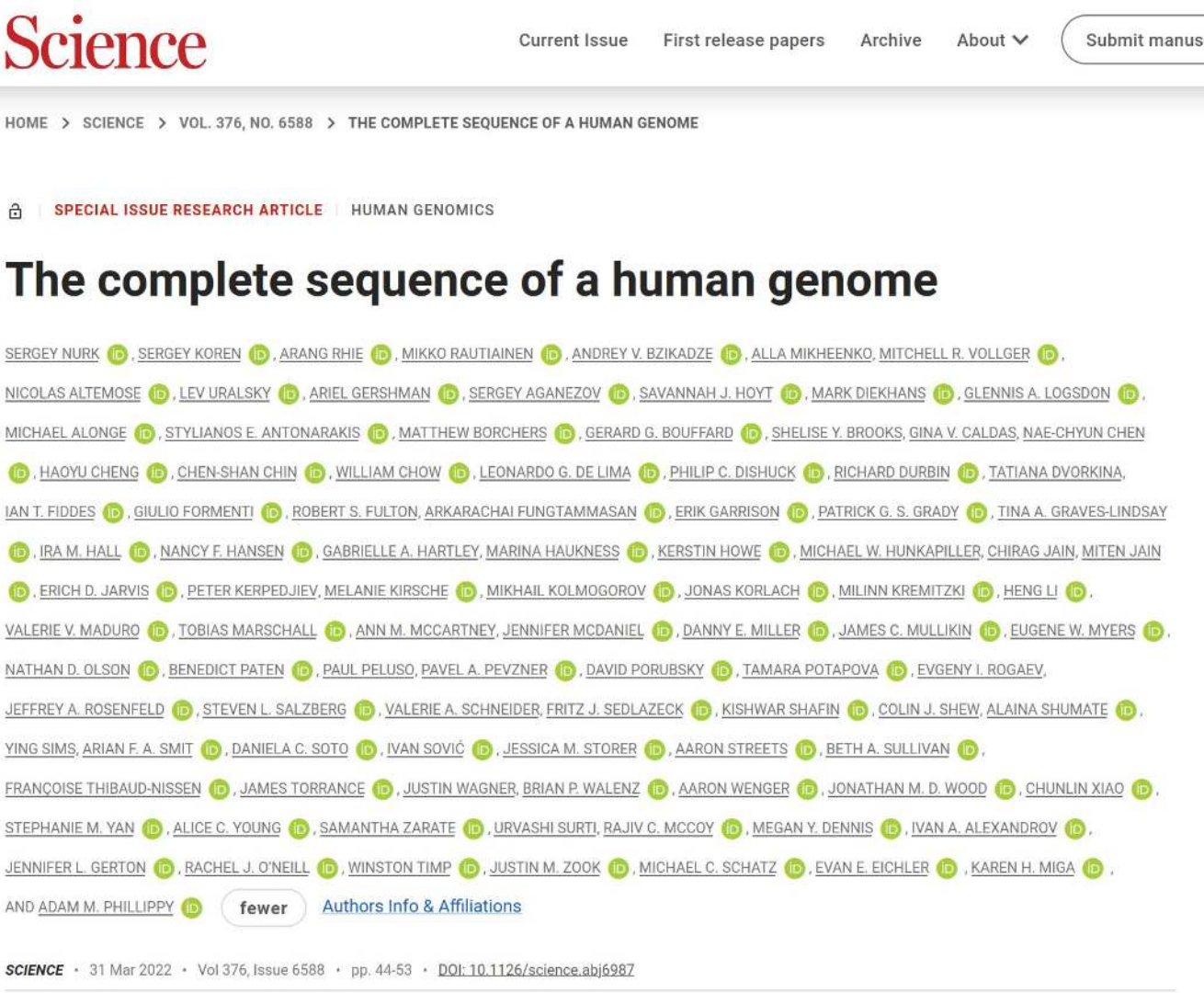
complement component 4 locus



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

The human pangenome illuminates the unknown

We finished the human genome



Science

Current Issue First release papers Archive About Submit manus

HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

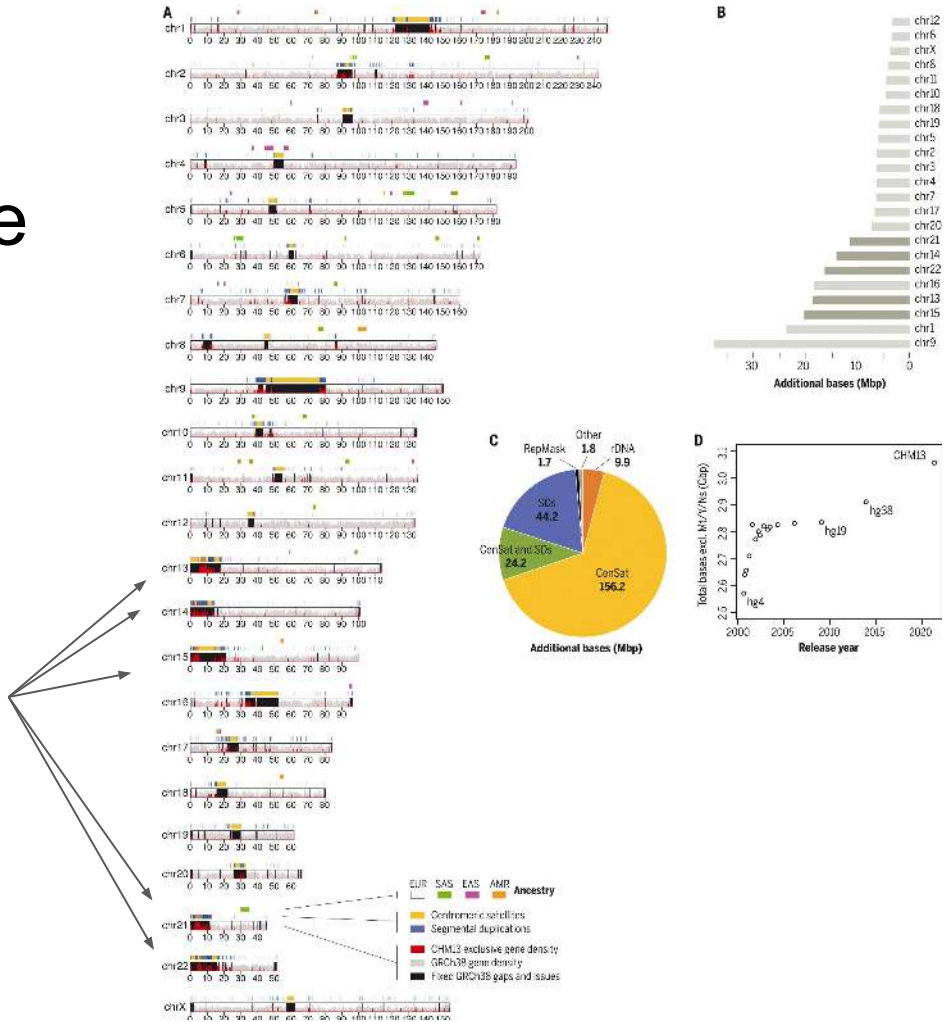
SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, ARIEL GERSHMAN, SERGEY AGANEZOV, SAVANNAH J. HOYT, MARK DIEKHANS, GLENNIS A. LOGSDON, MICHAEL ALONGE, STYLIANOS E. ANTONARAKIS, MATTHEW BORCHERS, GERARD G. 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

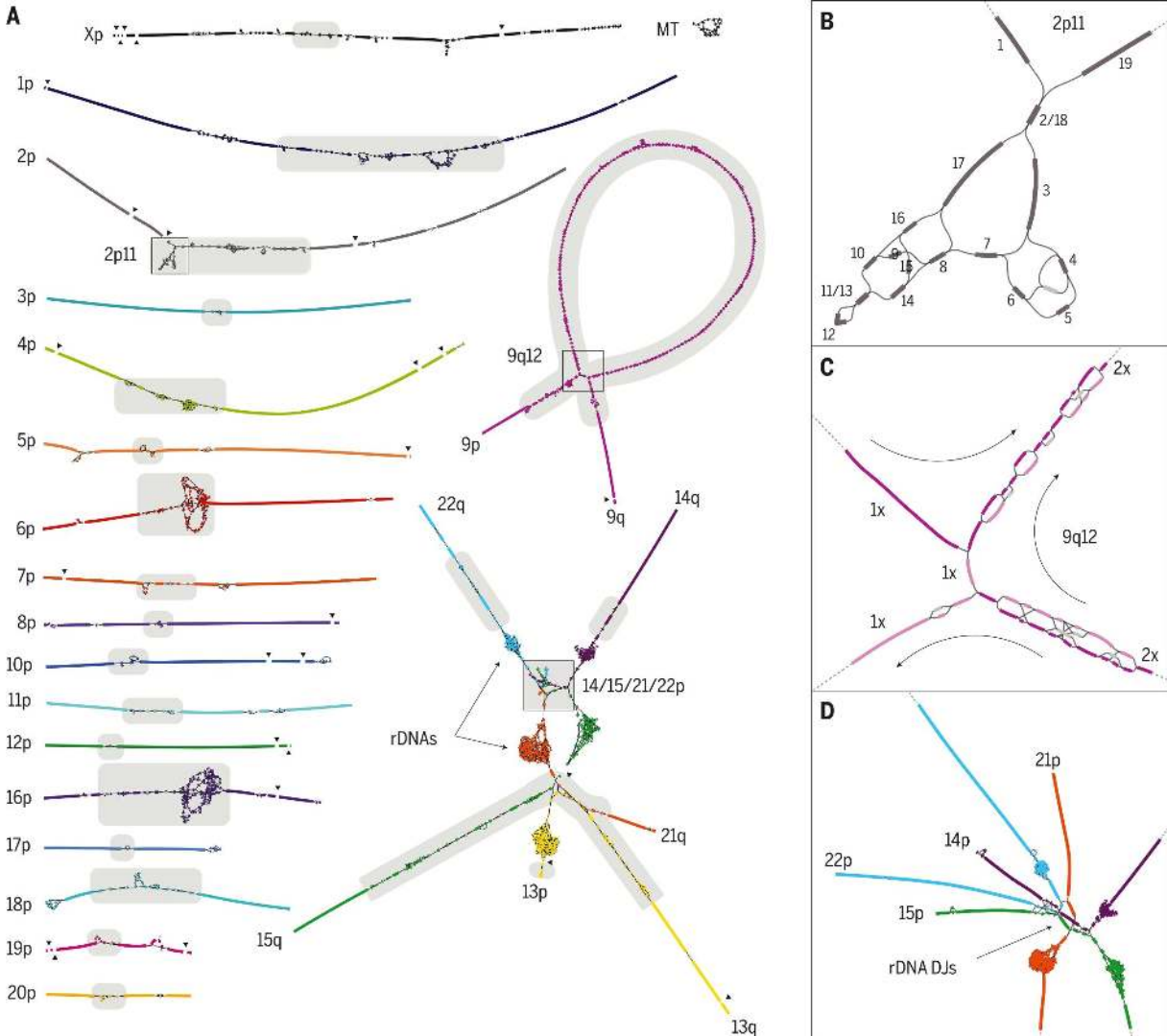
https://doi.org/10.1126/science.abj6987

Filling 8% of the reference which was incomplete

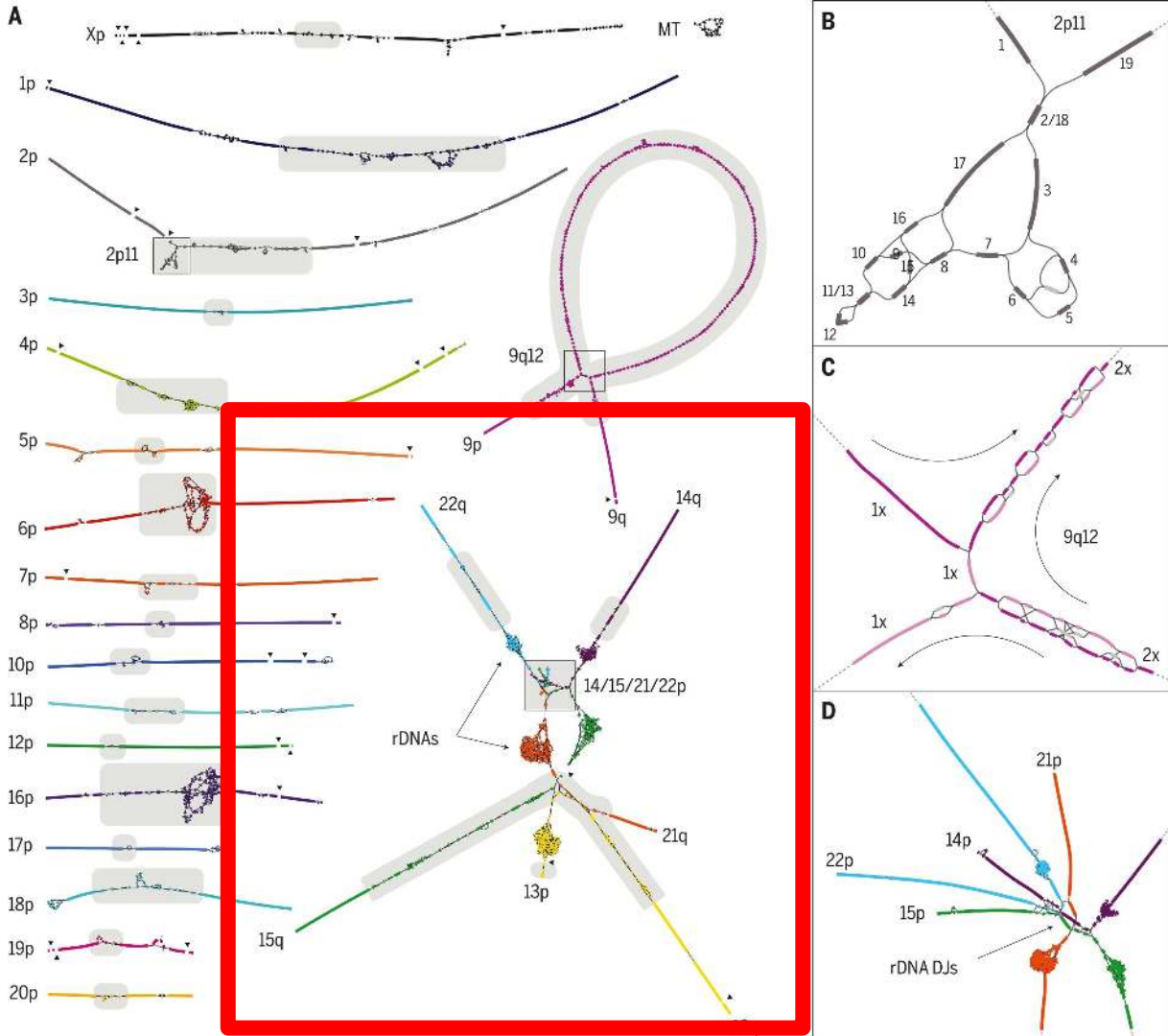
Of particular note, all of the acrocentric p-arms were assembled for the first time!



But there are new mysteries…

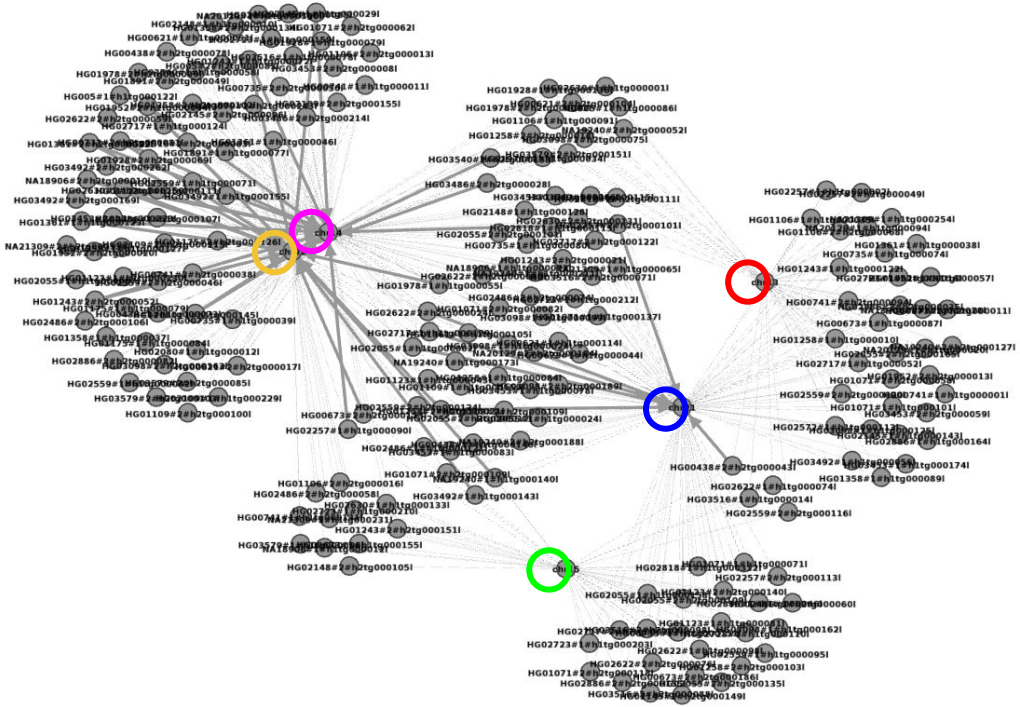


But there are new mysteries…

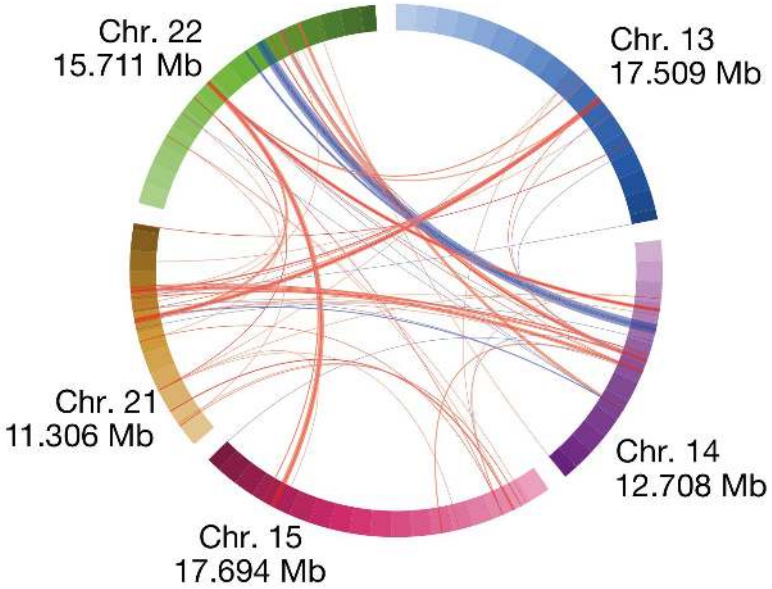


HPRC “misjoin” identification

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

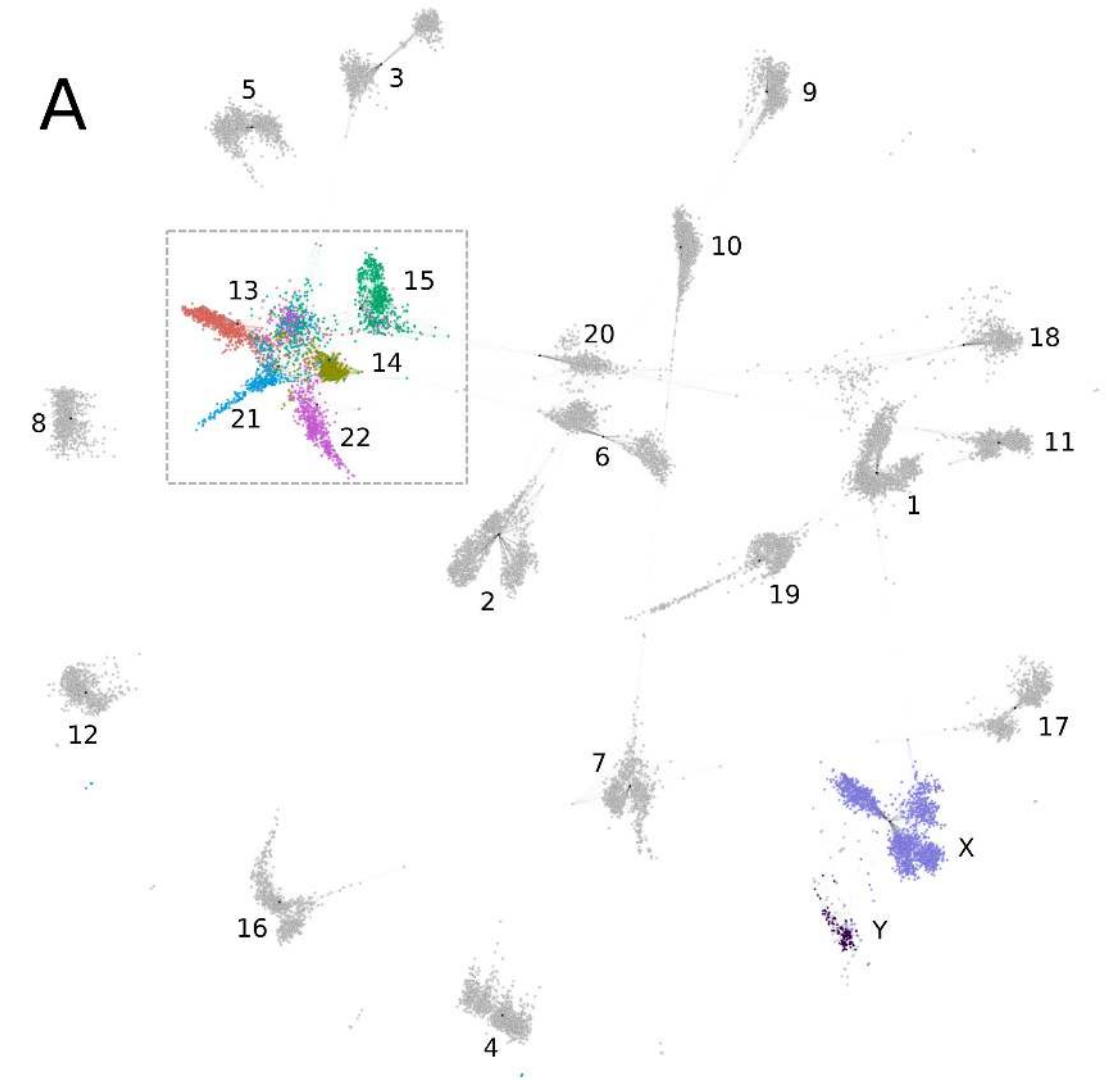


Alignment graph of the misjoins. Every node is a contig and every edge represents the number of mapping between nodes. Alignment graph obtained with pafnet and visualized with gephi. Color code: **chr13**, **chr14**, **chr15**, **chr21**, **chr22**.



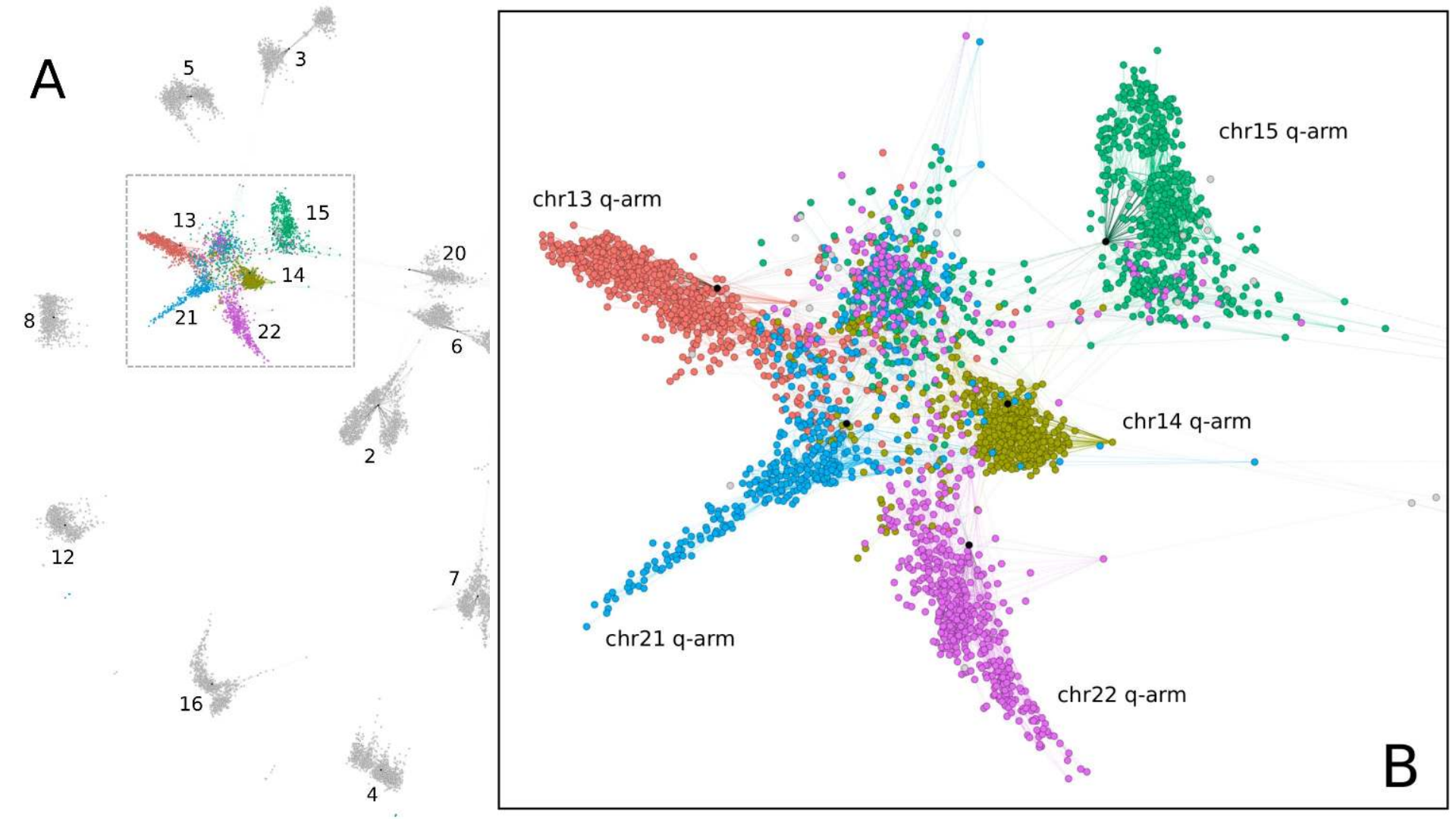
From HPRC main paper.

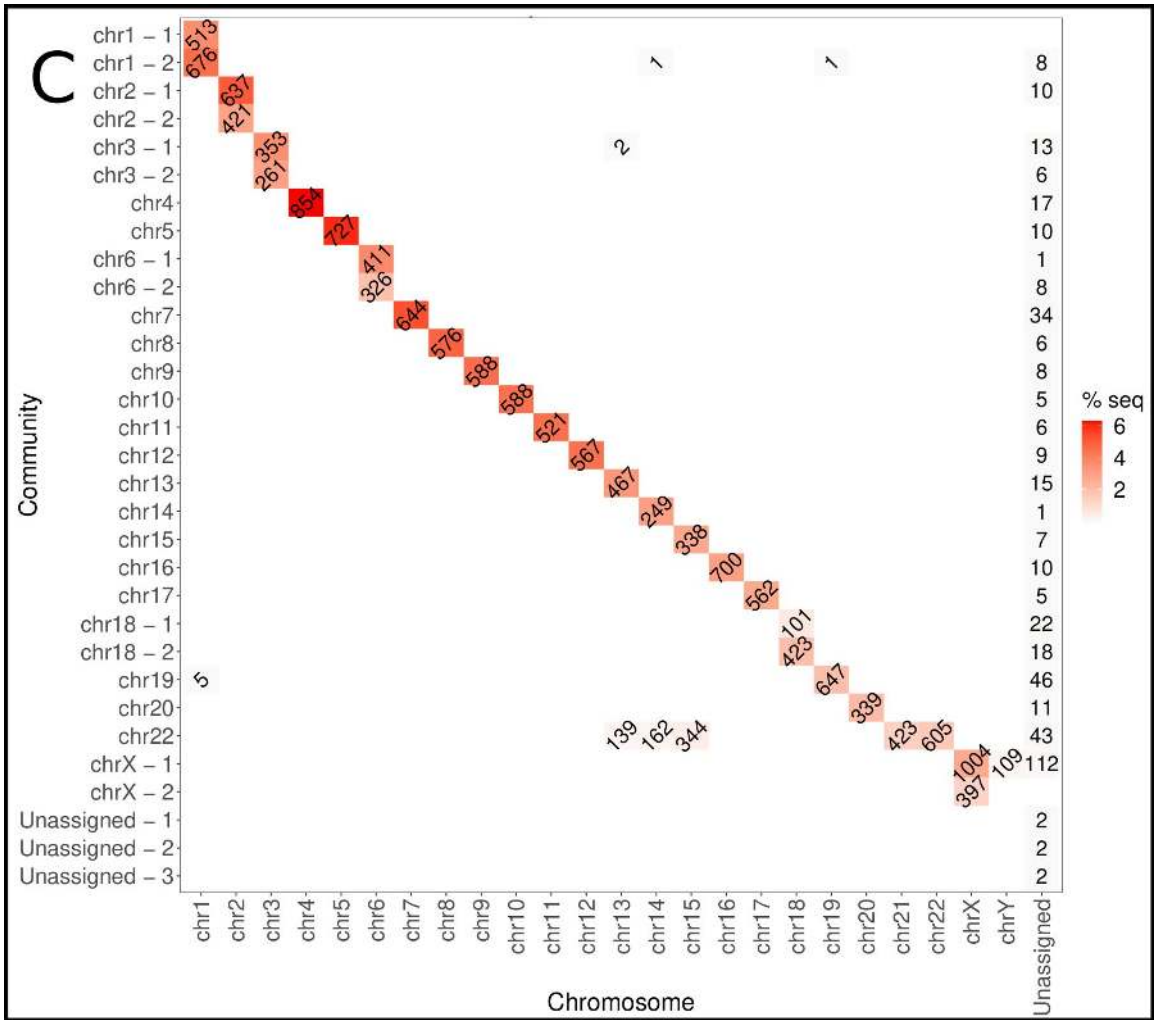
* With the exception of assembly errors in one haplotype (HG02080 paternal).



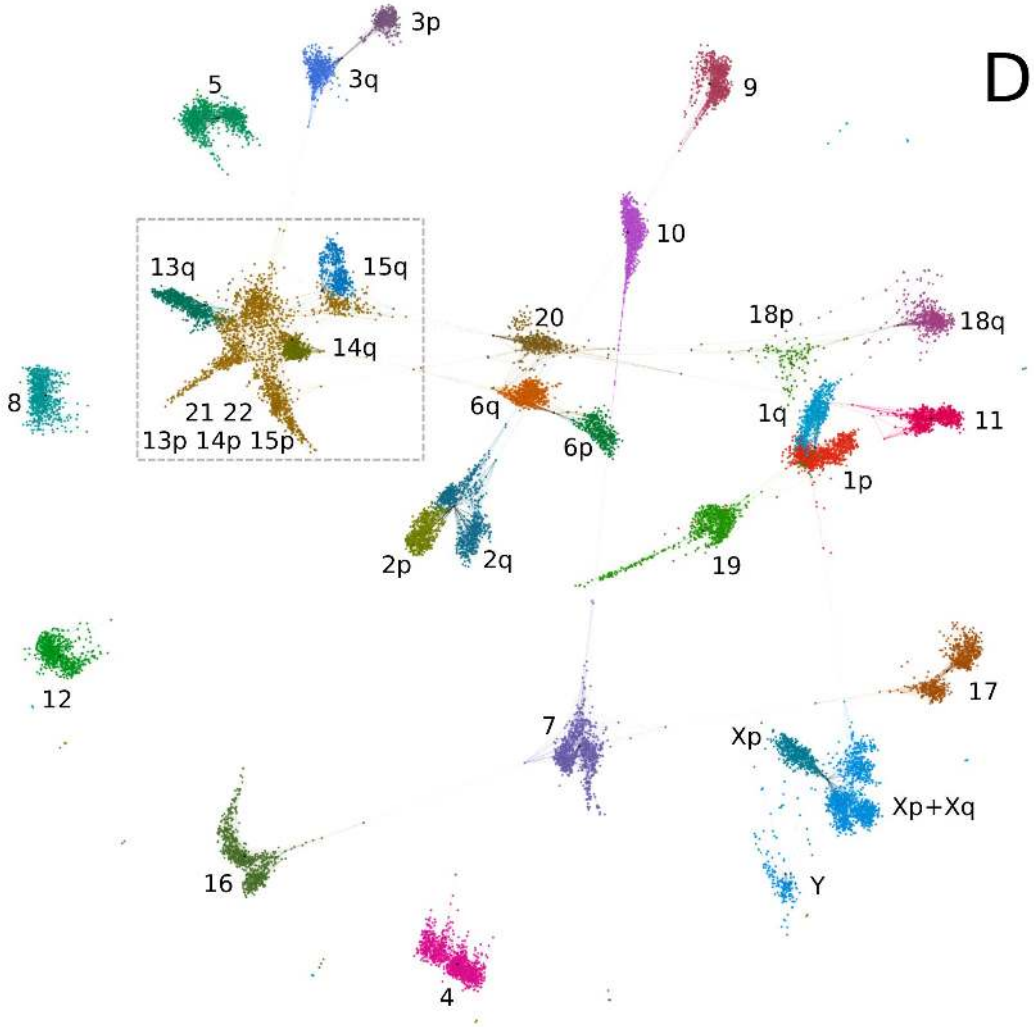
Chromosome communities in the HPRC

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


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



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

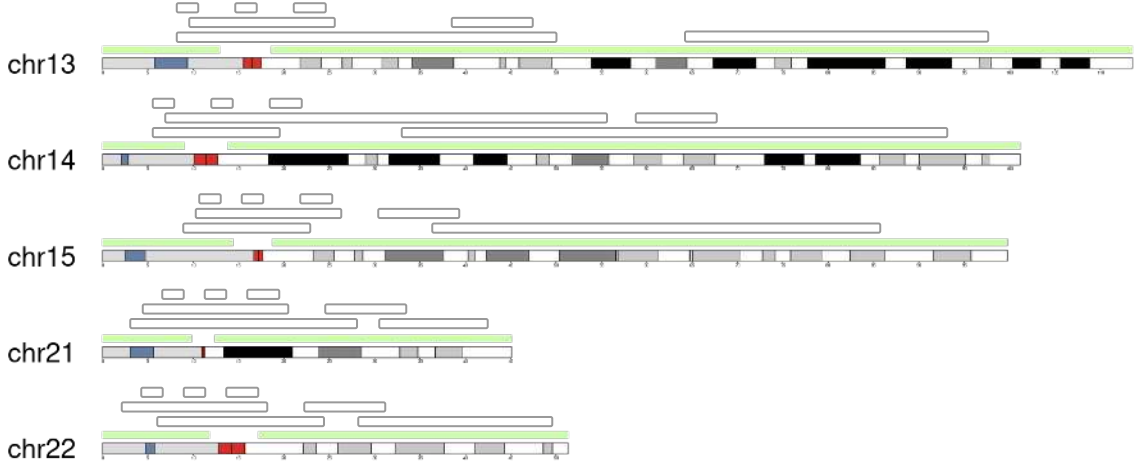
We decided to take a closer look, focusing on the best assemblies in these regions.

Workflow

HPRC assemblies

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

Mapping against the whole CHM13

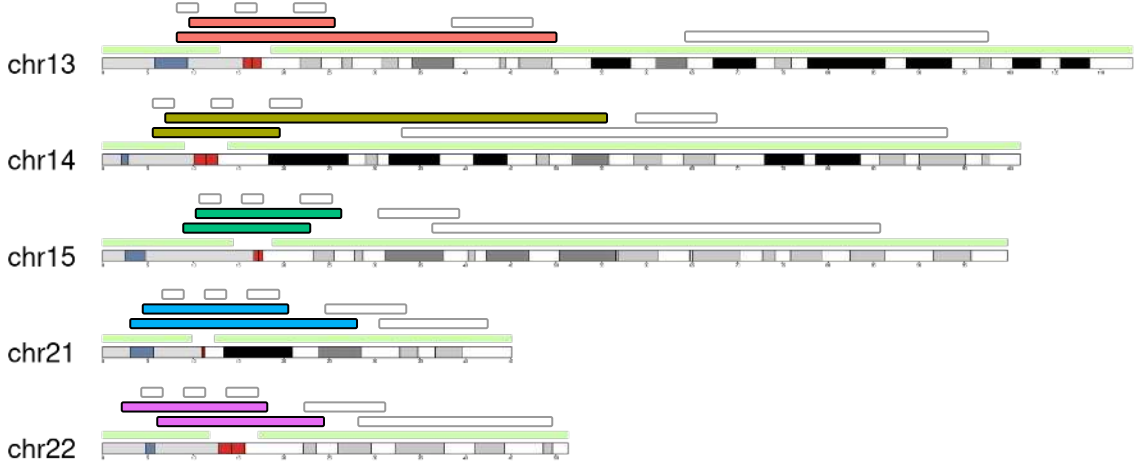


Workflow

HPRC assemblies

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

Mapping against the whole CHM13



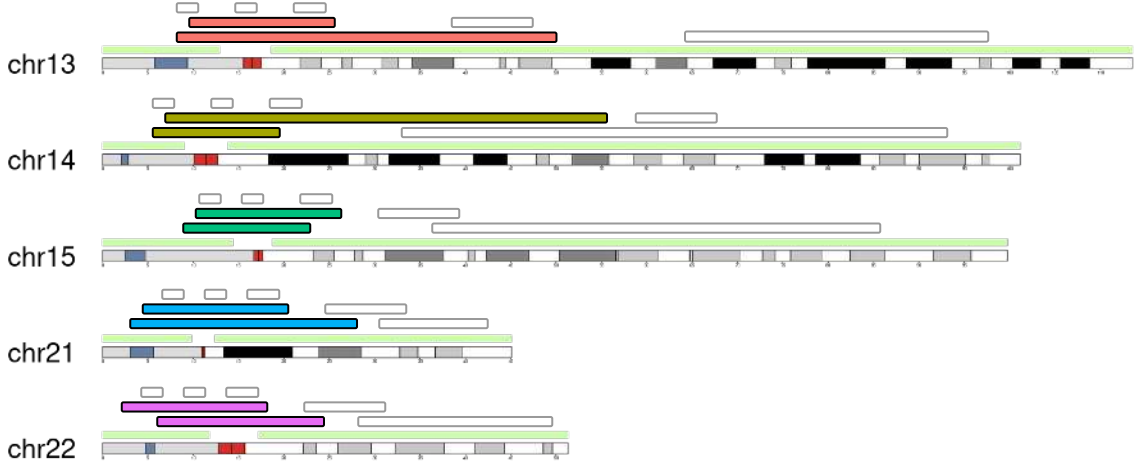
Acrocentric contigs covering (+/- 1Mbp) both the p and q arms (pq-contigs)

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

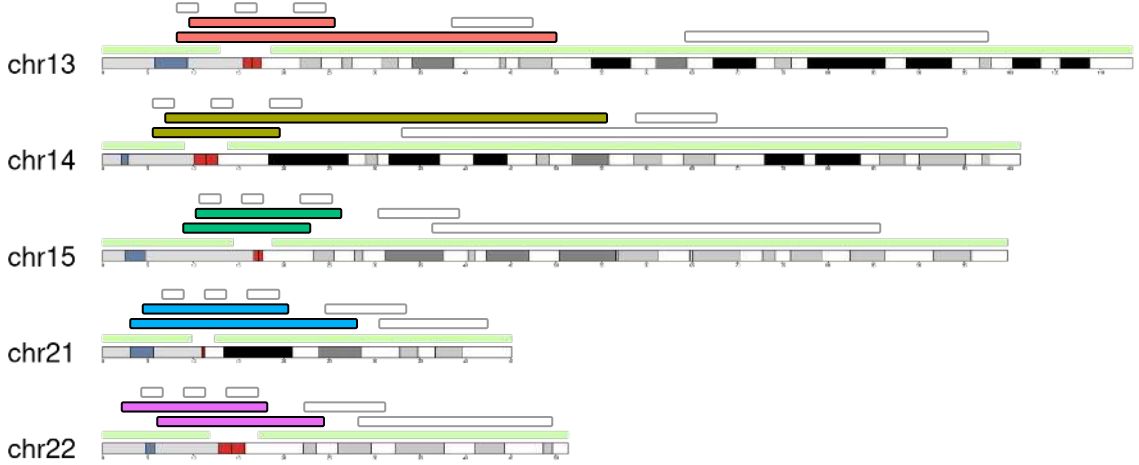
PanGenome Graph Builder (PGGB)

Workflow

HPRC assemblies

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

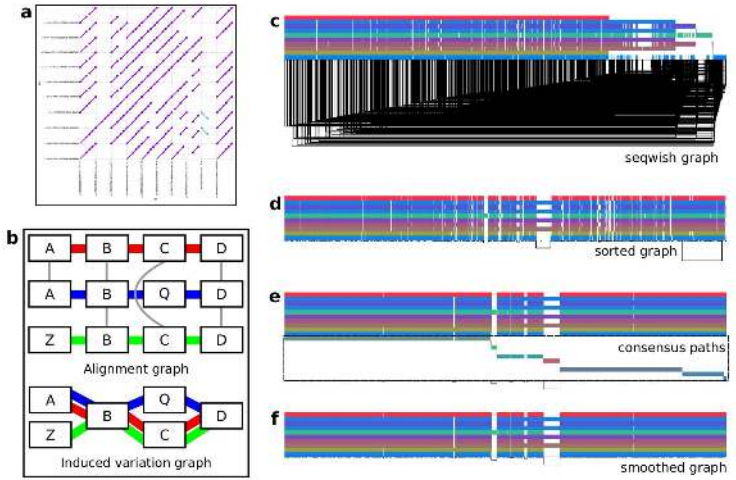
Mapping against the whole CHM13



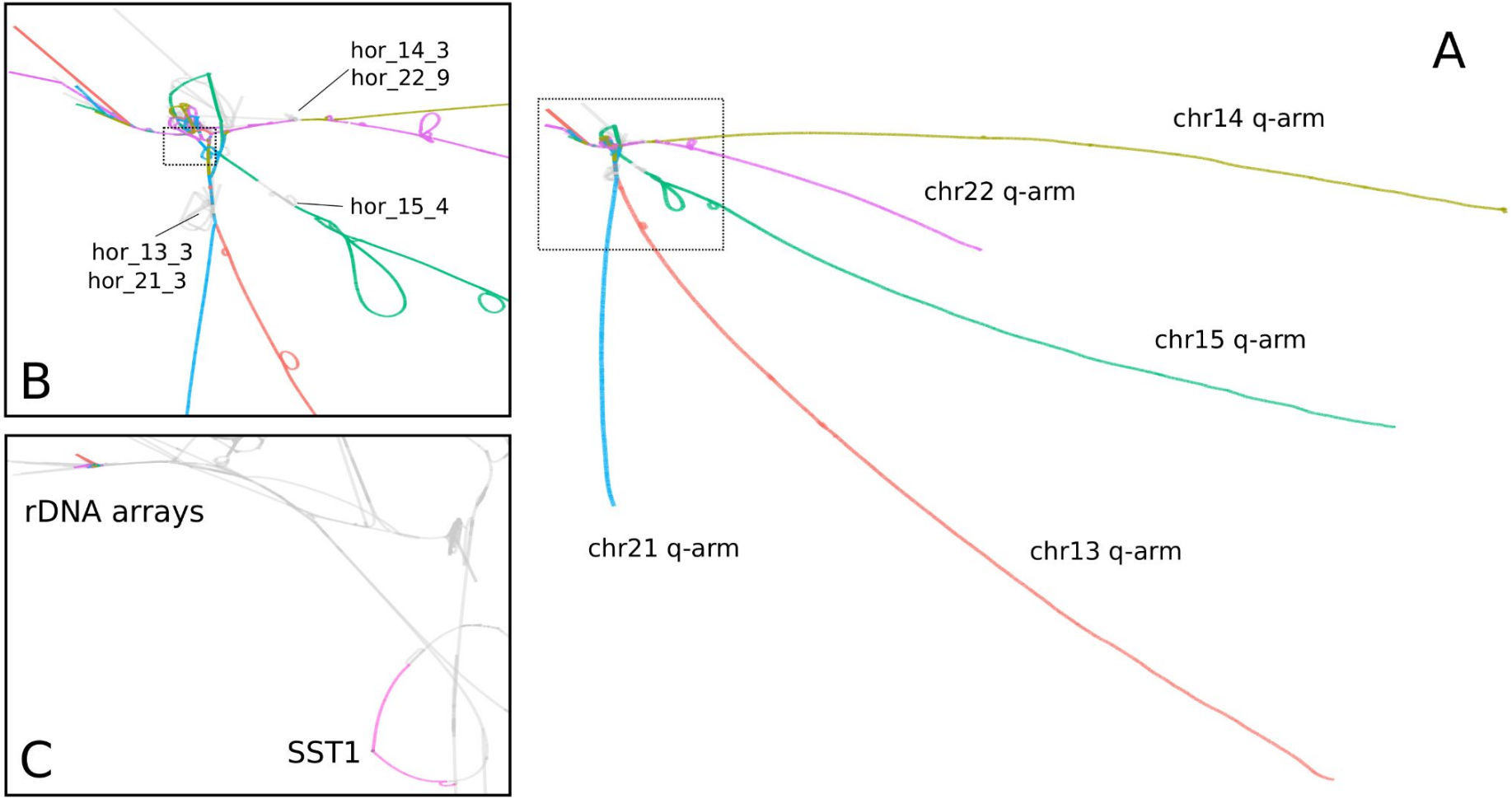
Acrocentric contigs covering (+/- 1Mbp) both the p and q arms (pq-contigs)

+ HG002 contigs >= 300kbps which map to acrocentrics

PanGenome Graph Builder (PGGB)



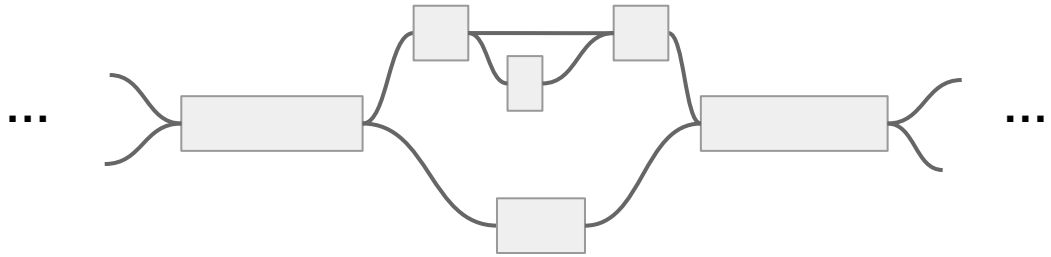
https://github.com/pangenome/pggb


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 recombination in the acrocentric short arm(s)

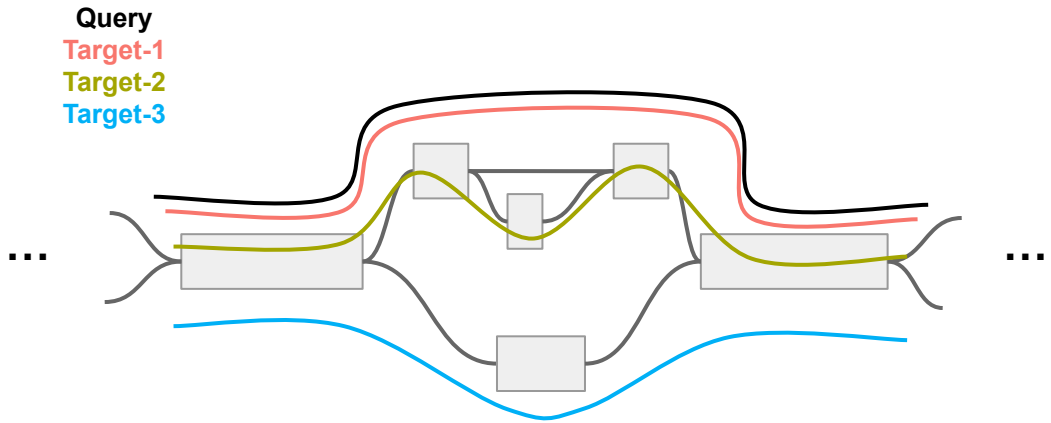
Pangenome untangling

Untangling extracts pairwise alignments from variation graphs.



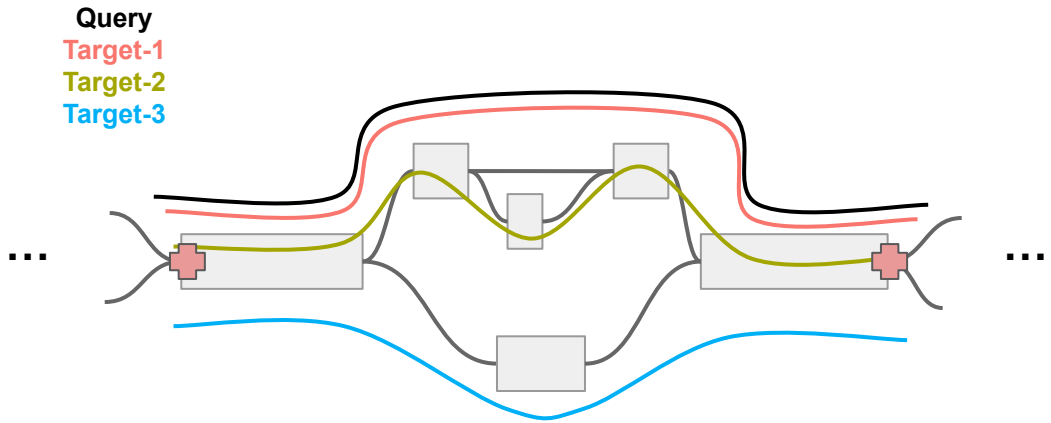
Pangenome untangling

Untangling extracts pairwise alignments from variation graphs.



Pangenome untangling

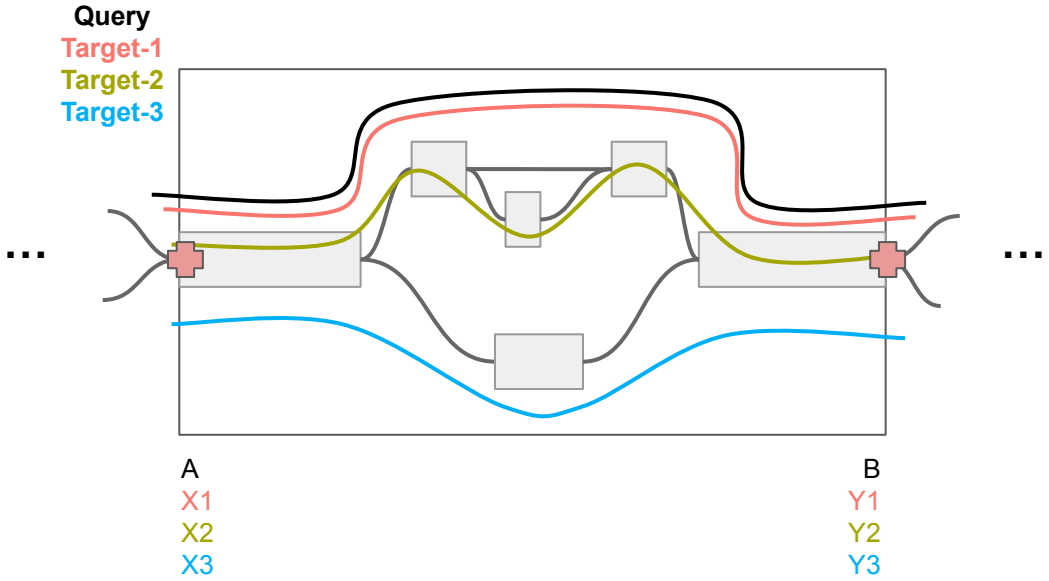
Untangling extracts pairwise alignments from variation graphs.



Identify cut points in the graph

Pangenome untangling

Untangling extracts pairwise alignments from variation graphs.

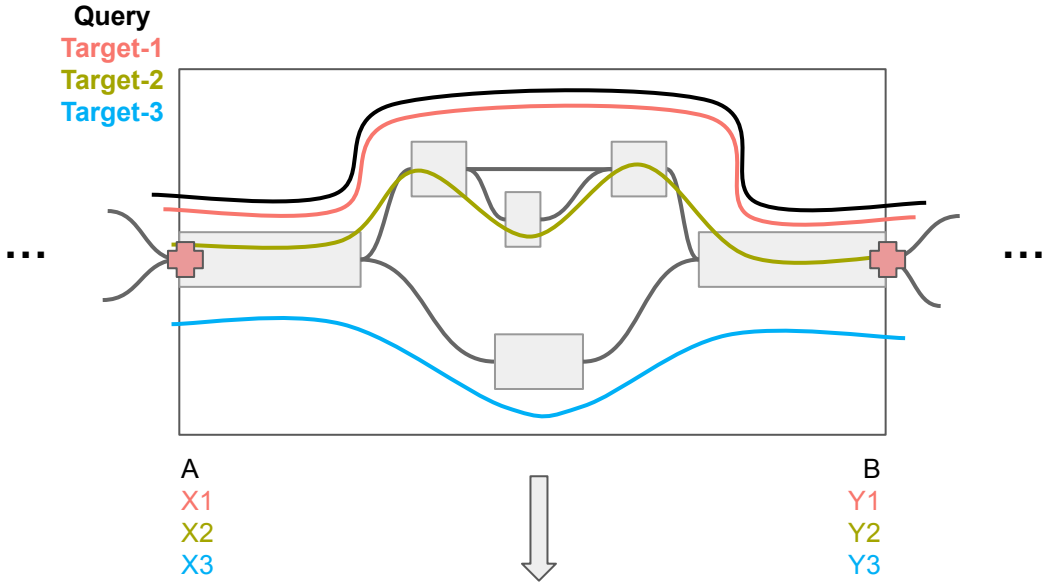


Identify cut points in the graph

Define segment boundaries

Pangenome untangling

Untangling extracts pairwise alignments from variation graphs.



query	start	end	target	start	end	jaccard	rank
Query	A	B	Target-1	X1	Y1	1	1
Query	A	B	Target-2	X2	Y2	0.95	2
Query	A	B	Target-3	X3	Y3	0.7	3

Identify cut points in the graph

Define segment boundaries

Compare segments

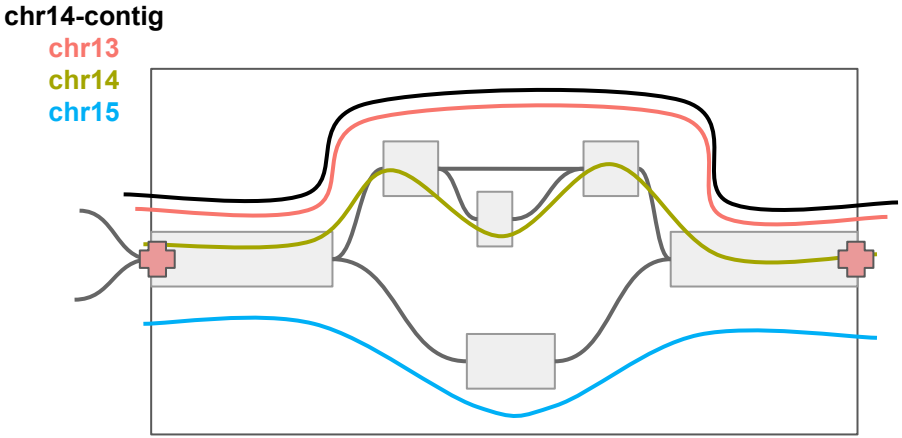
with jaccard in node (allele) space

Grounding the untangle against CHM13

We first pick a consistent set of query contig segments.

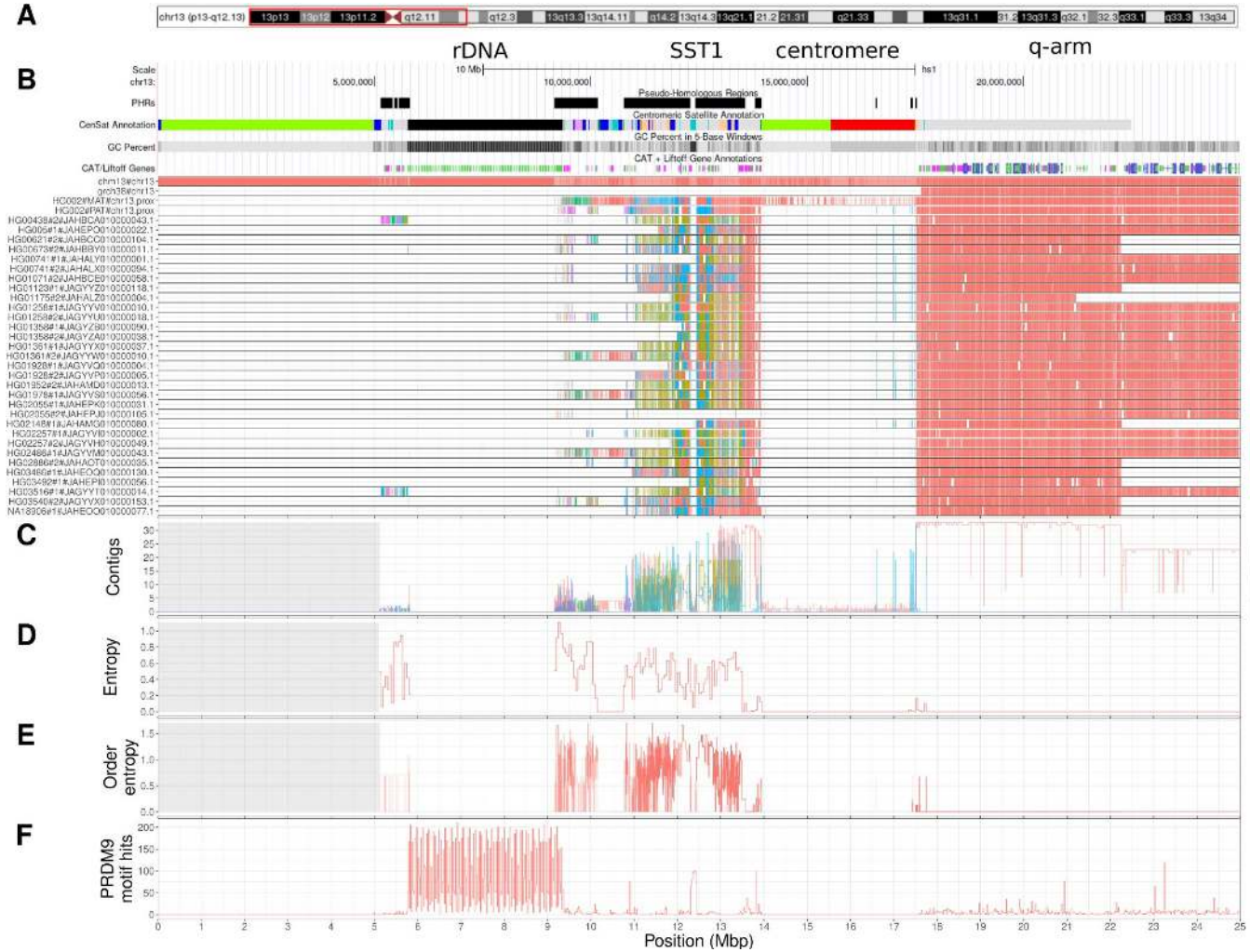
For each, we find its best mapping against each acrocentric chromosome.

We then merge this (grounding) with a “multiple” untangling result where each query contig segment picks the best *N* target (reference) segments.



query	start	end	target	start	end	jaccard	rank
chr14-contig	A	B	chr14	X2	Y2	0.95	1

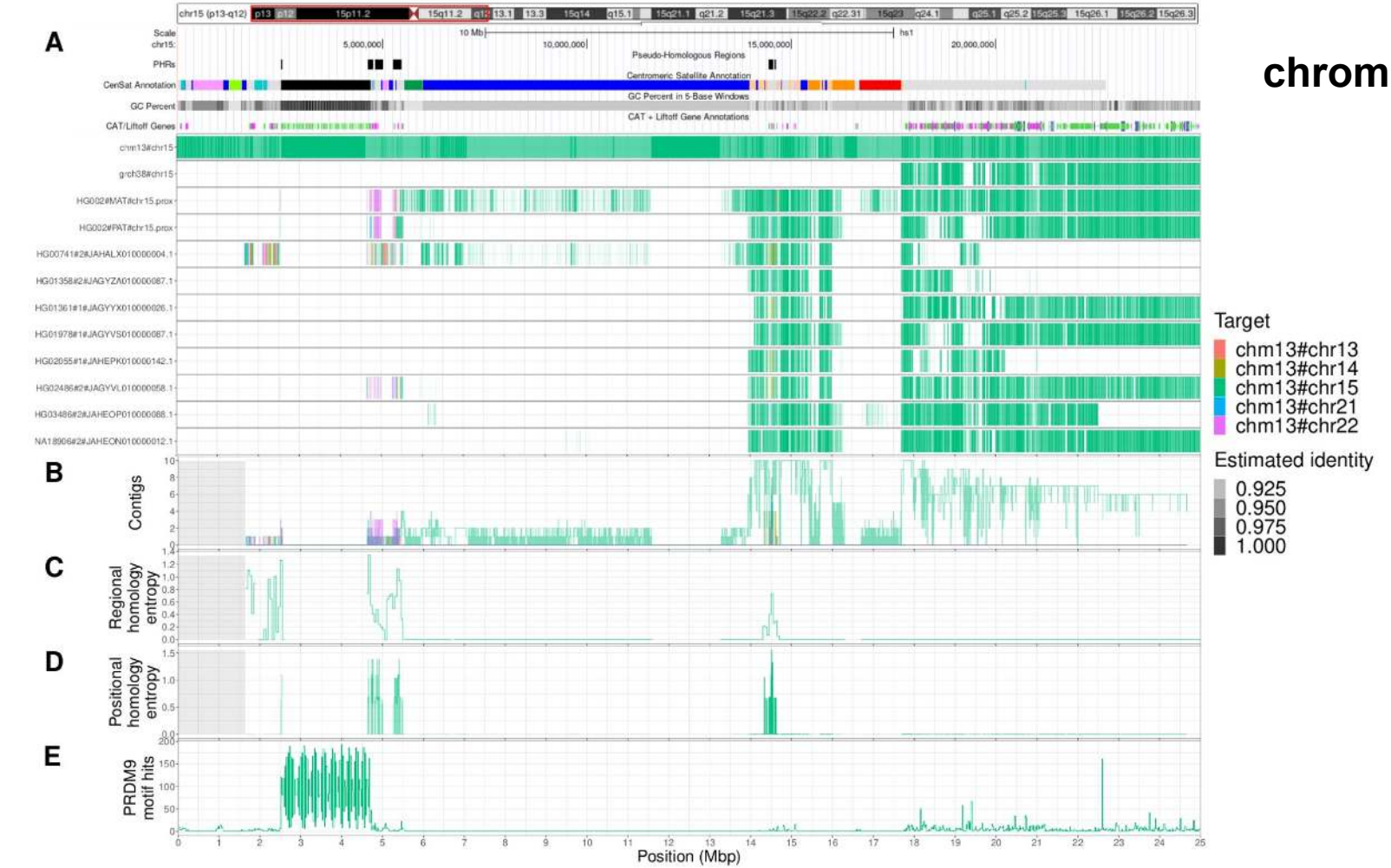
query	start	end	target	start	end	jaccard	rank
chr14-contig	A	B	chr13	X1	Y1	1	1
chr14-contig	A	B	chr14	X2	Y2	0.95	2
chr14-contig	A	B	chr15	X3	Y3	0.7	3



Untangling the pangenome graph

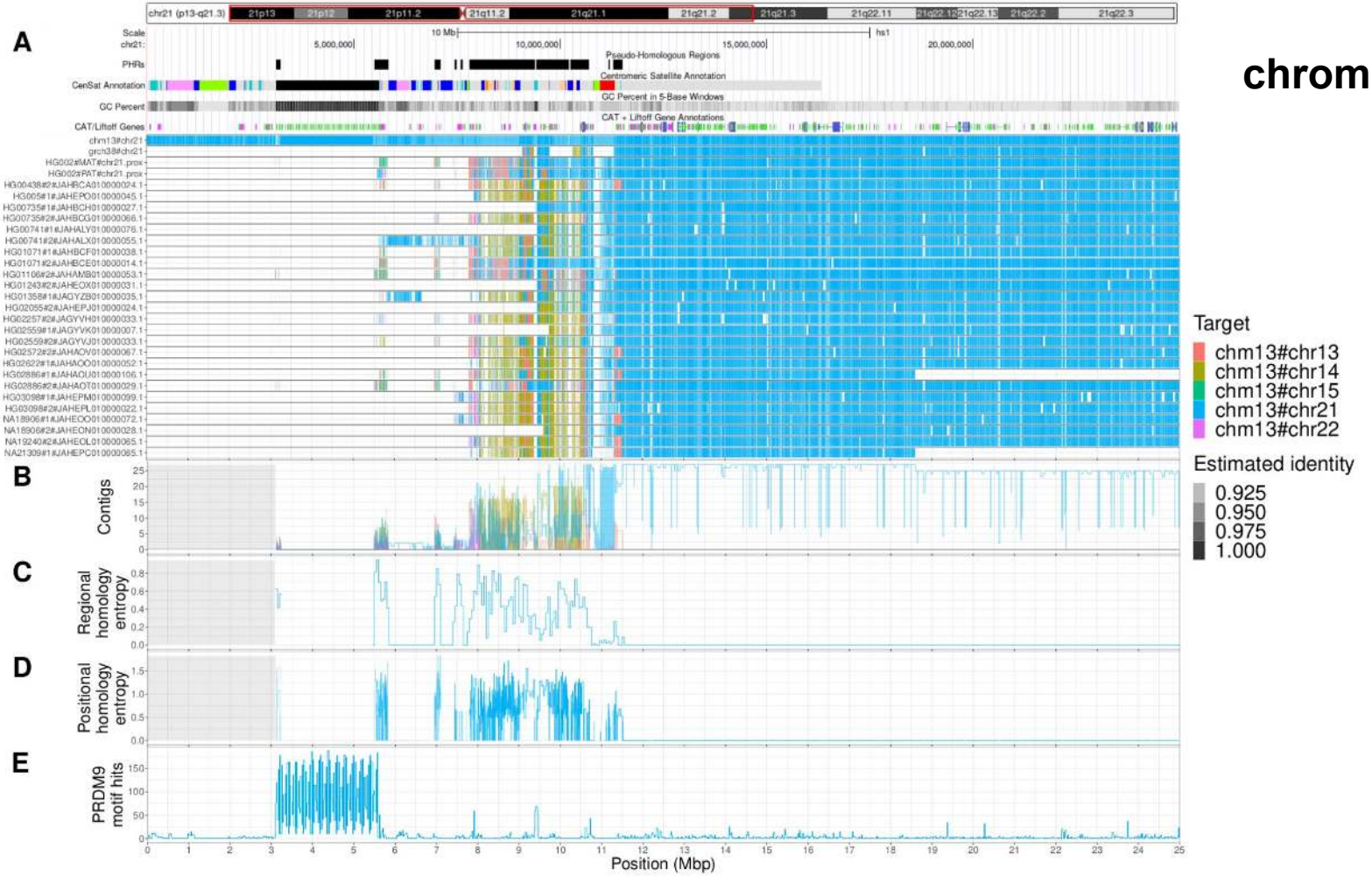
We look into the graph from the perspective of chromosome 13. Full information from pangenome plus reference annotations.

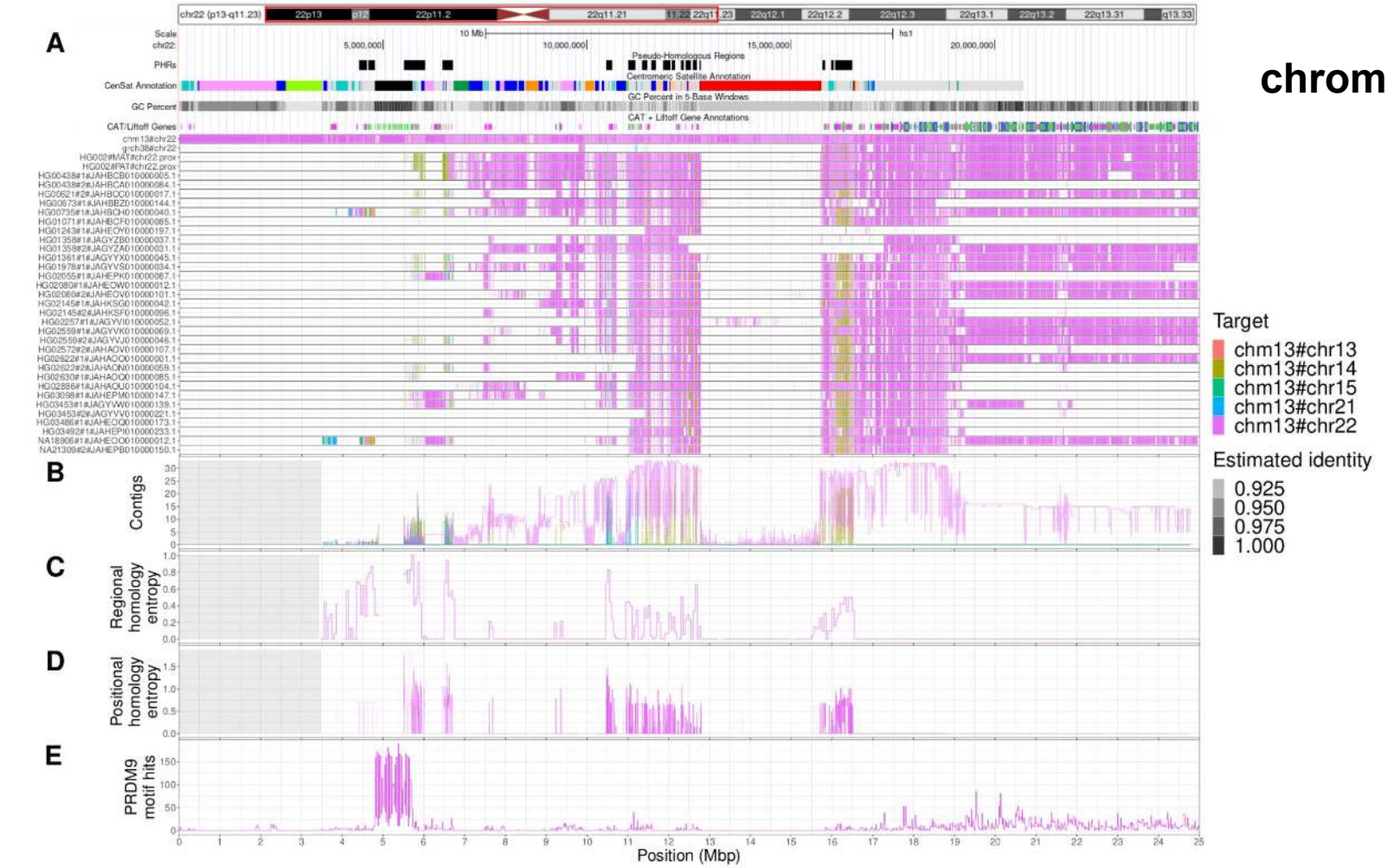
chromosome 14



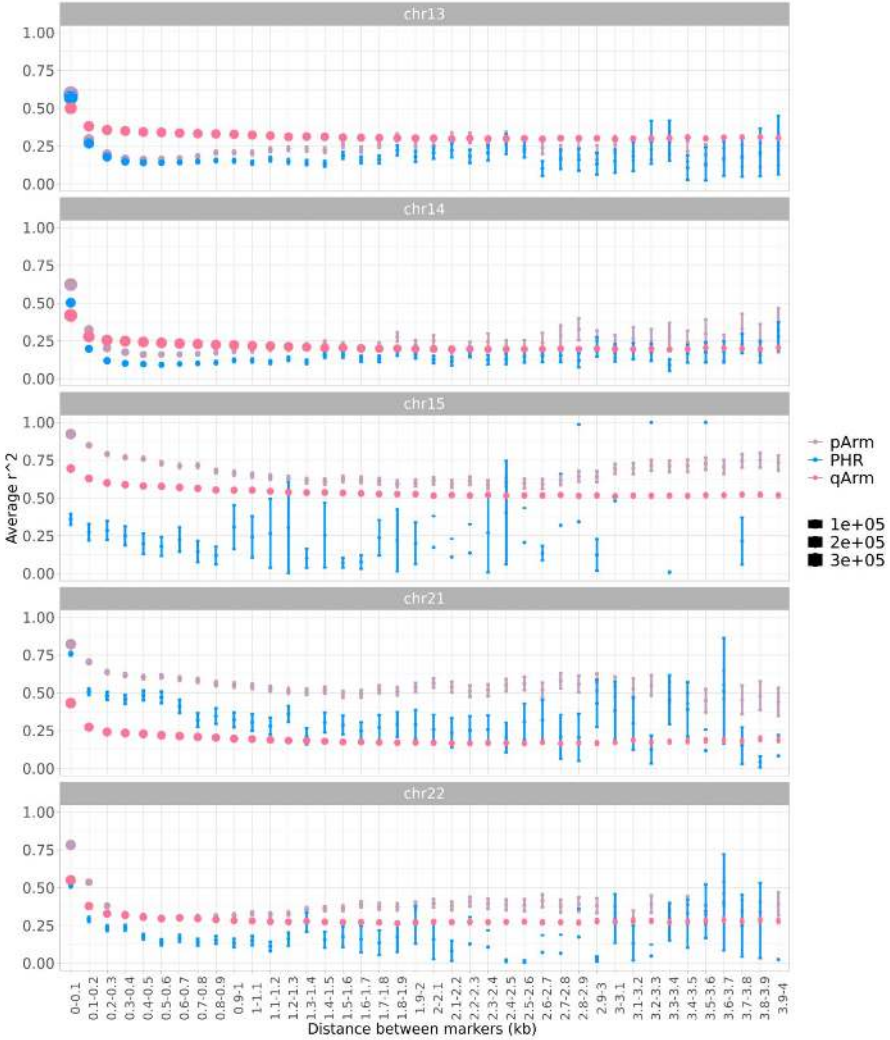
chromosome 15

chromosome 21





chromosome 22



Linkage disequilibrium (LD) describes correlation between pairs of SNPs.

Higher LD (correlation) means lower rates of recombination.

We express LD in terms of its “decay” with distance. That’s what we see in this plot.

LD decays faster in pseudo-homologous regions than elsewhere in the p-arms or in the q-arms. (chr15 lacks data)

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

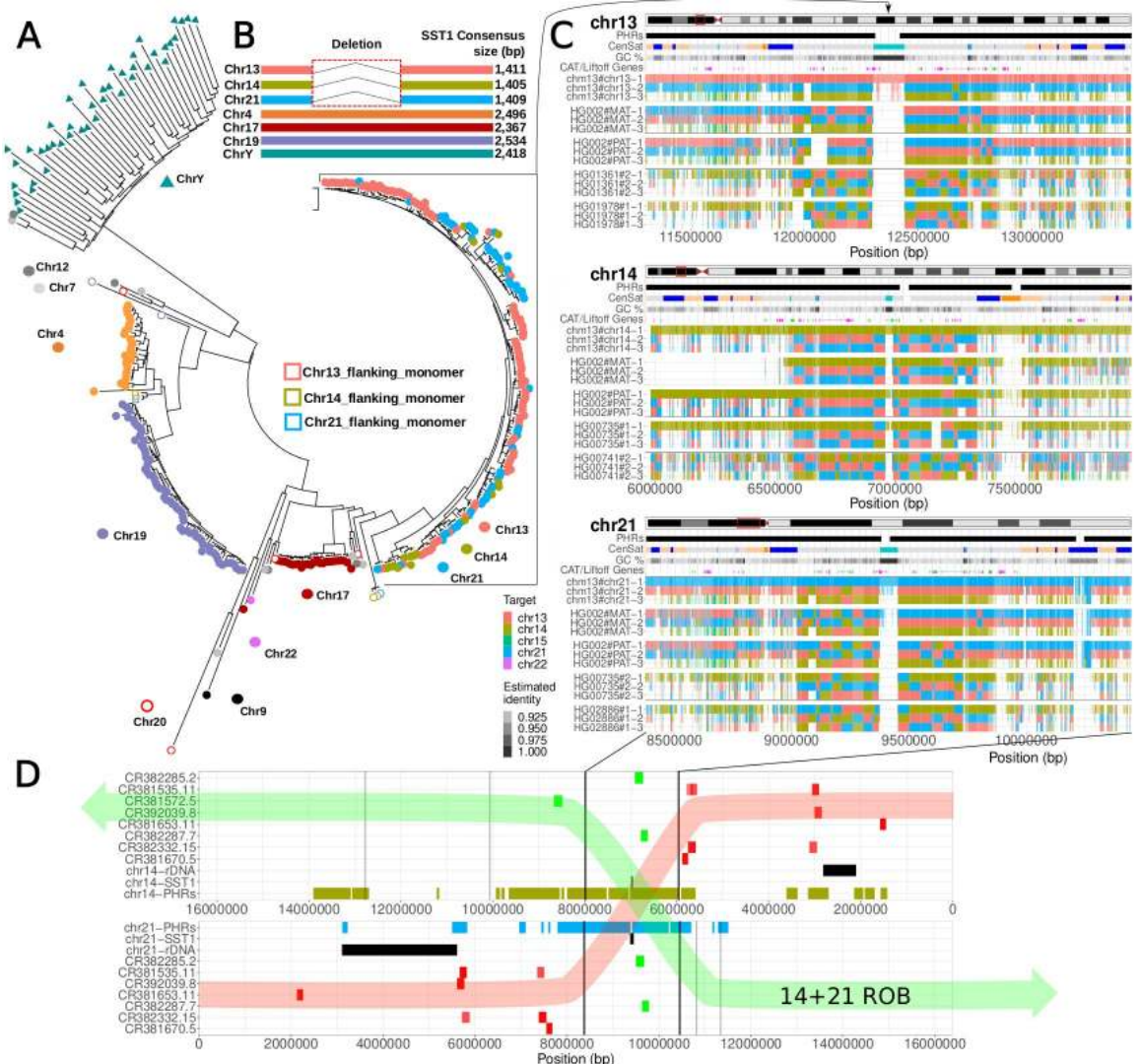
Silvia Buonaiuto, Vincenza Colonna

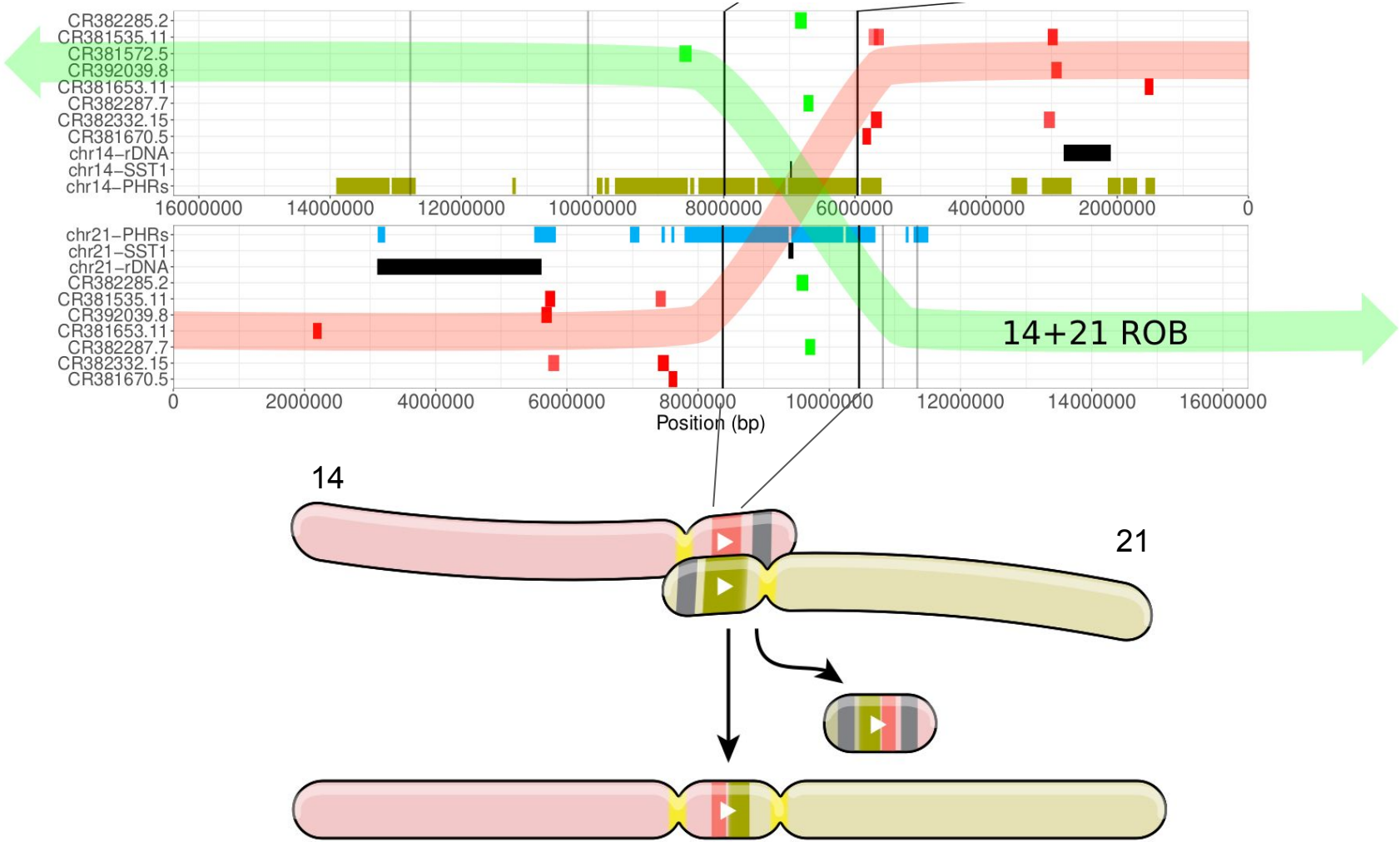
chromosome 13/14/21
Pseudo-homologous regions

The SST1 array is a GC-rich satellite (DNA repeat) derived from Alu elements.

It lies at the center of pseudo-homologous regions (PHRs) on 3 acrocentric chromosomes.

This is where we see recombination occurring in Robertsonian translocations.




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
14+21 ROB
0
2000000
4000000
6000000
8000000
10000000
12000000
14000000
16000000
Position (bp)
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

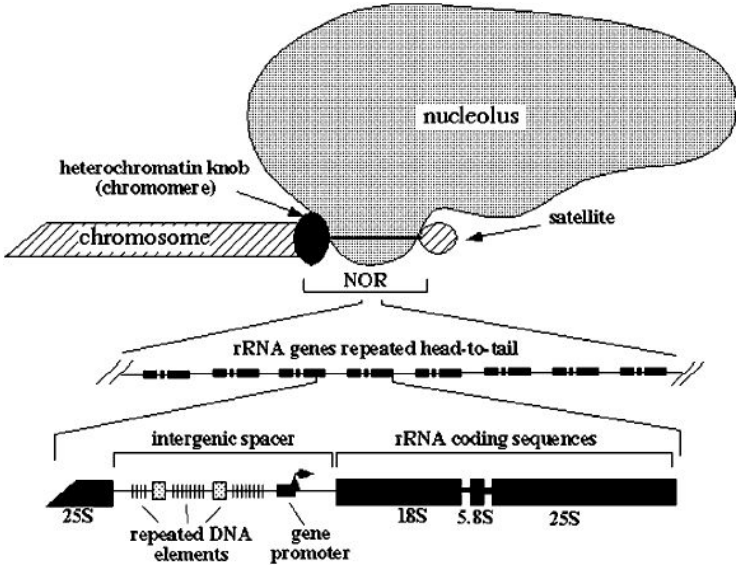
K.H.Choo*, B.Vissel, R.Brown, R.G.Filby and E.Earle

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.

distal region of NOR



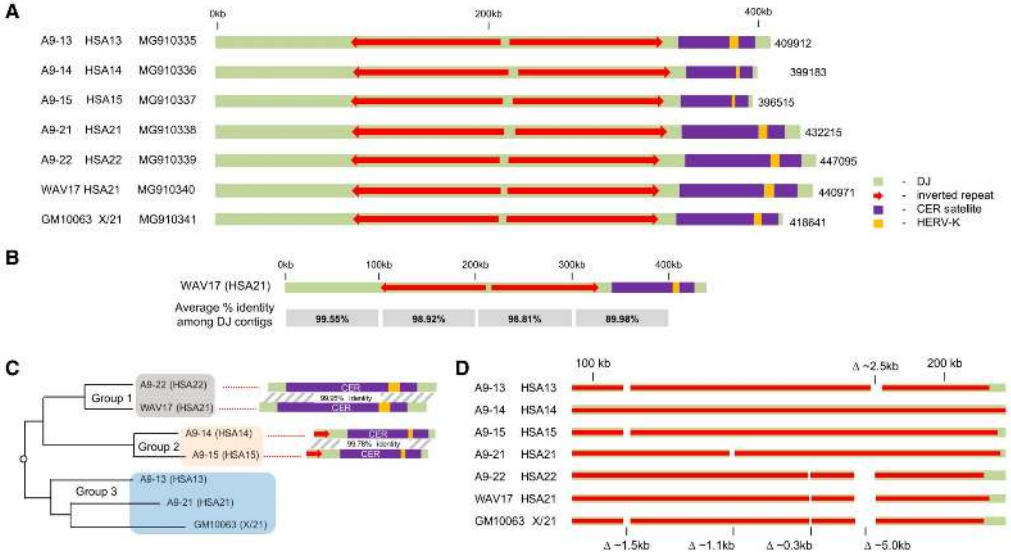
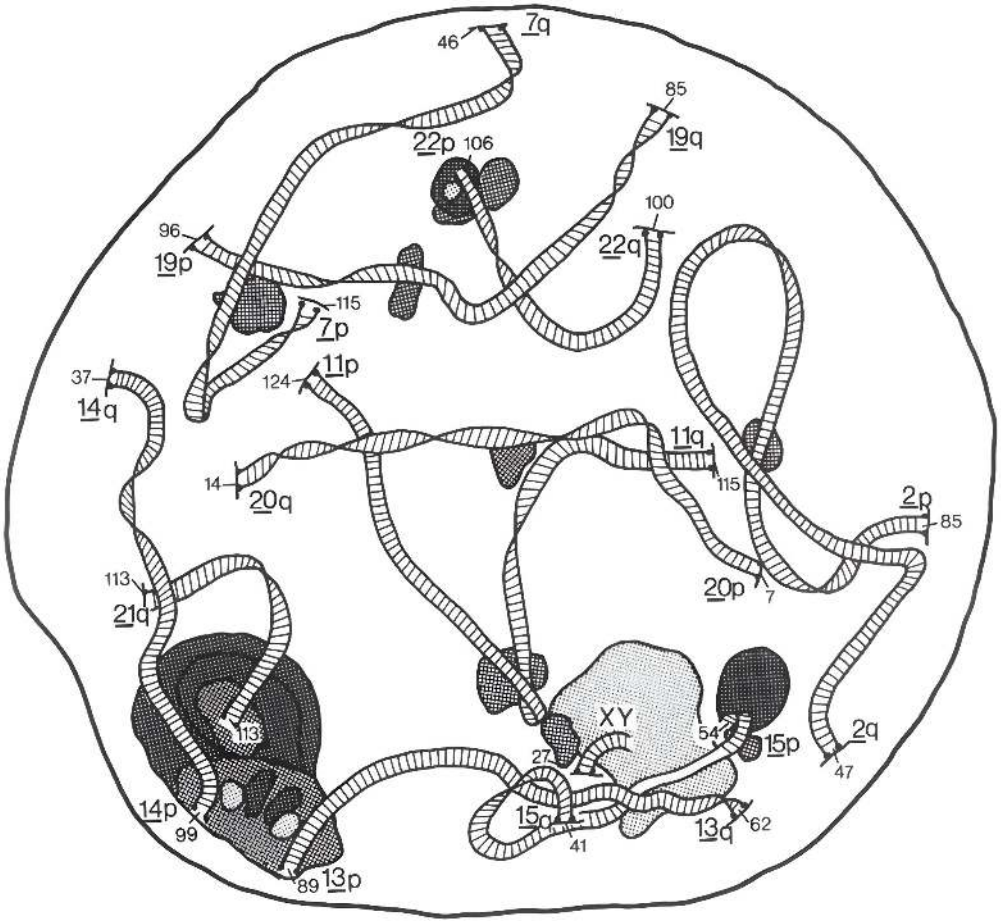


Figure 2. Sequencing of DJ regions from individual human acrocentric chromosomes. (*A*) Schematic representations of DJ contigs. Hybrid and chromosomal identities are shown on the *left*, together with GenBank accession numbers. (*B*) The average percentage identity of 100-kb blocks, among all seven DJ contigs is shown *below* the WAV17 (HSA21) contig. (*C*) Alignment of DJ contigs demonstrating that they can be clustered into three groups, based on the sequence of their distal ends. The percentage identity of the most distal sequences within group 1 and 2 members is shown schematically on the *right*. (*D*) A schematic representation of indel distribution on the left arm inverted repeat. Hybrid and chromosomal identities are shown on the *left* (see Supplemental Fig. S7B for sequence alignments at break points).

https://doi.org/10.1101/gad.331892.119



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

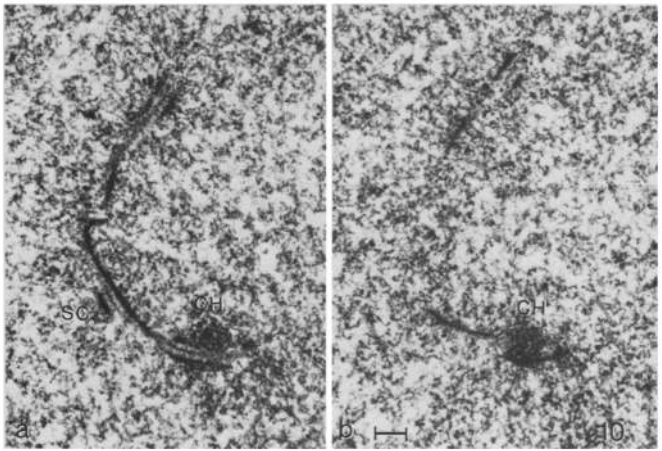


Figure 10. Two consecutive sections through the centromeric heterochromatin of a bivalent at early pachytene. The synaptonemal complex (SC) passes unaltered through the centromeric heterochromatin (CH). (Bar = 0.2 μm)

https://doi.org/10.1007/BF02910454

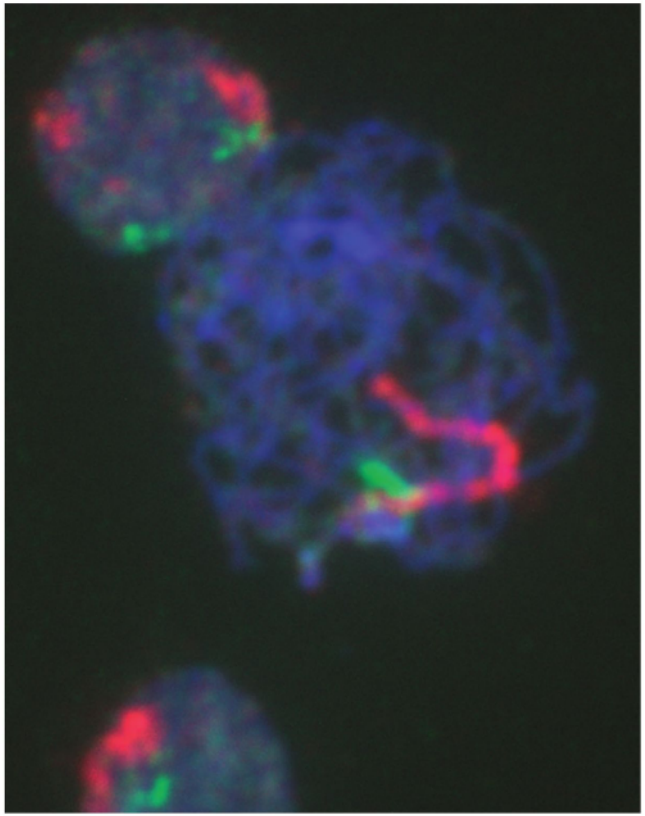


Figure The pachytene nucleus with 2 hybridization signals is oriented in a linear fashion. The *red signal* represents chromosome 14, and the *green signal* represents chromosome 21.

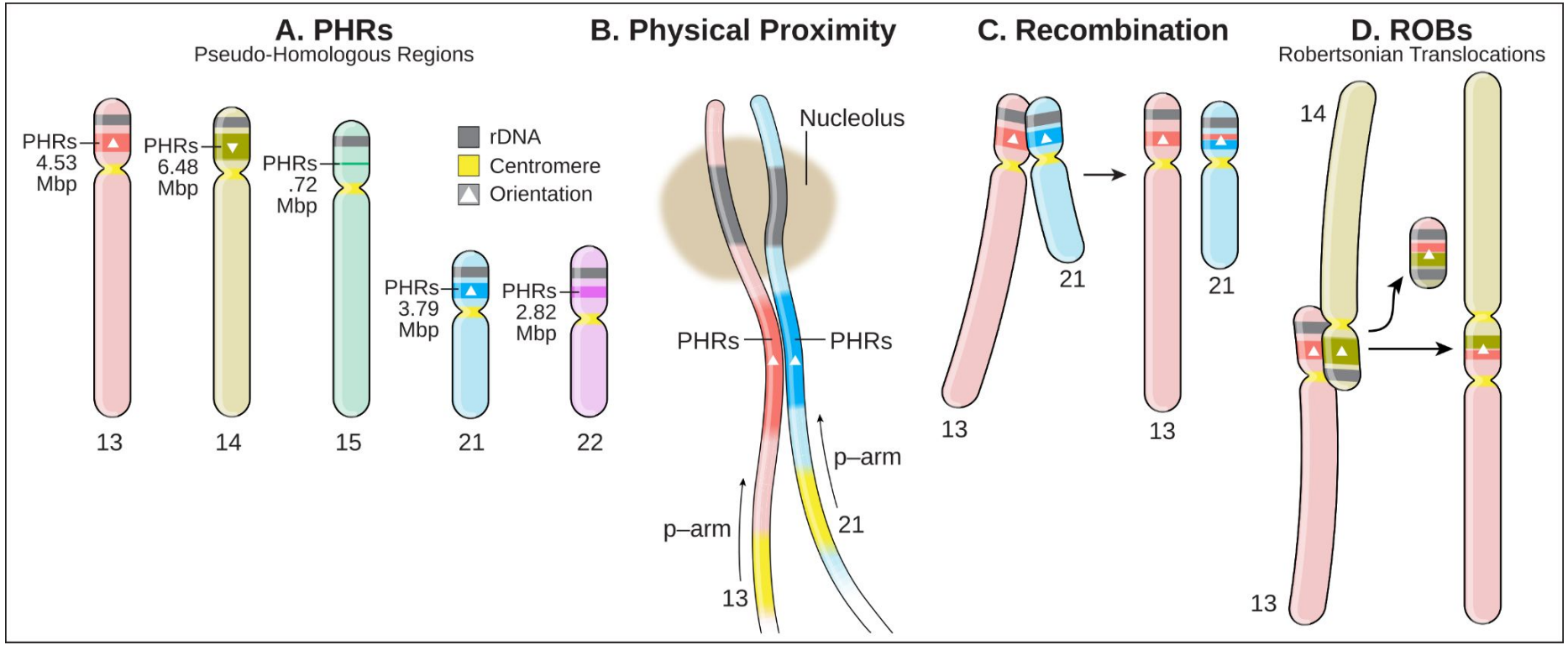
TRANSACTIONS OF THE 70TH ANNUAL MEETING OF THE PACIFIC COAST OBSTETRICIANS AND GYNECOLOGICAL SOCIETY | VOLUME 190, ISSUE 6, P1781-1785, JUNE 01, 2004

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

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

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

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].

Practical!

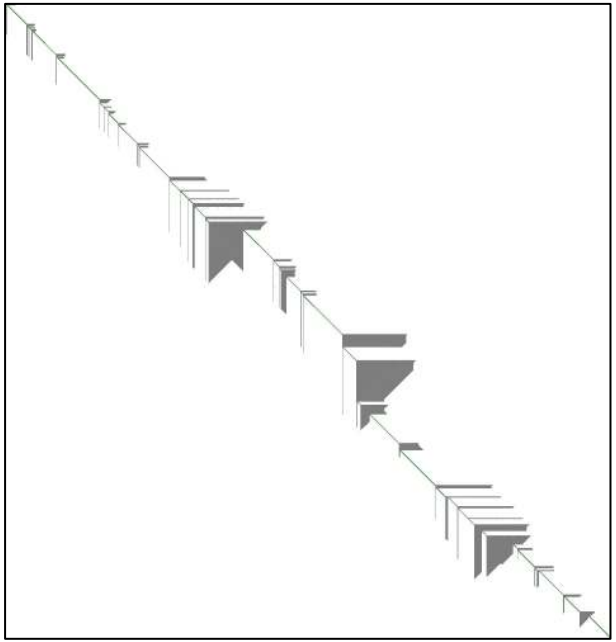
Let’s build some pangenome variation graphs with **pggb**!

First: a deeper dive into how the method works.

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

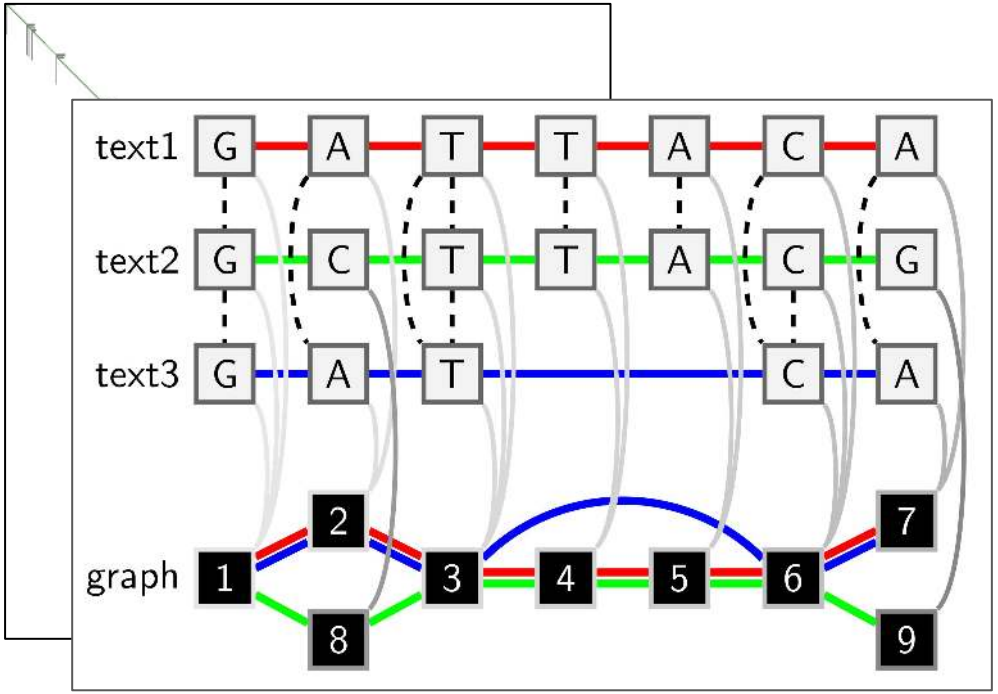

PGP

PanGenome Graph Builder



wfmash (biWFA)

PanGenome Graph Builder

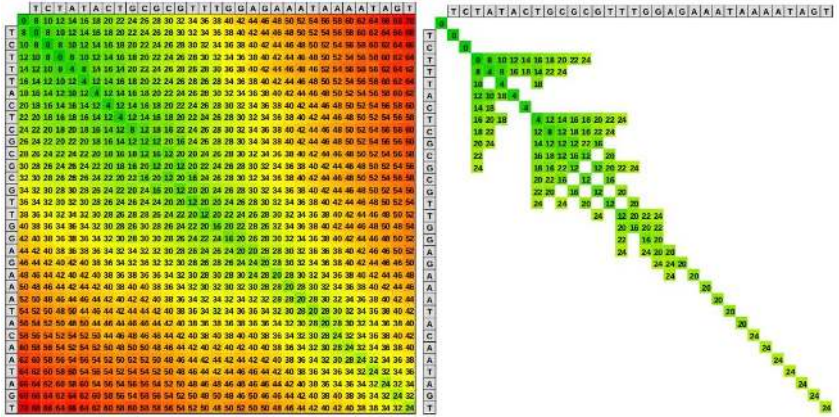


seqwish (unbiased graph builder)

PanGenome Graph Builder

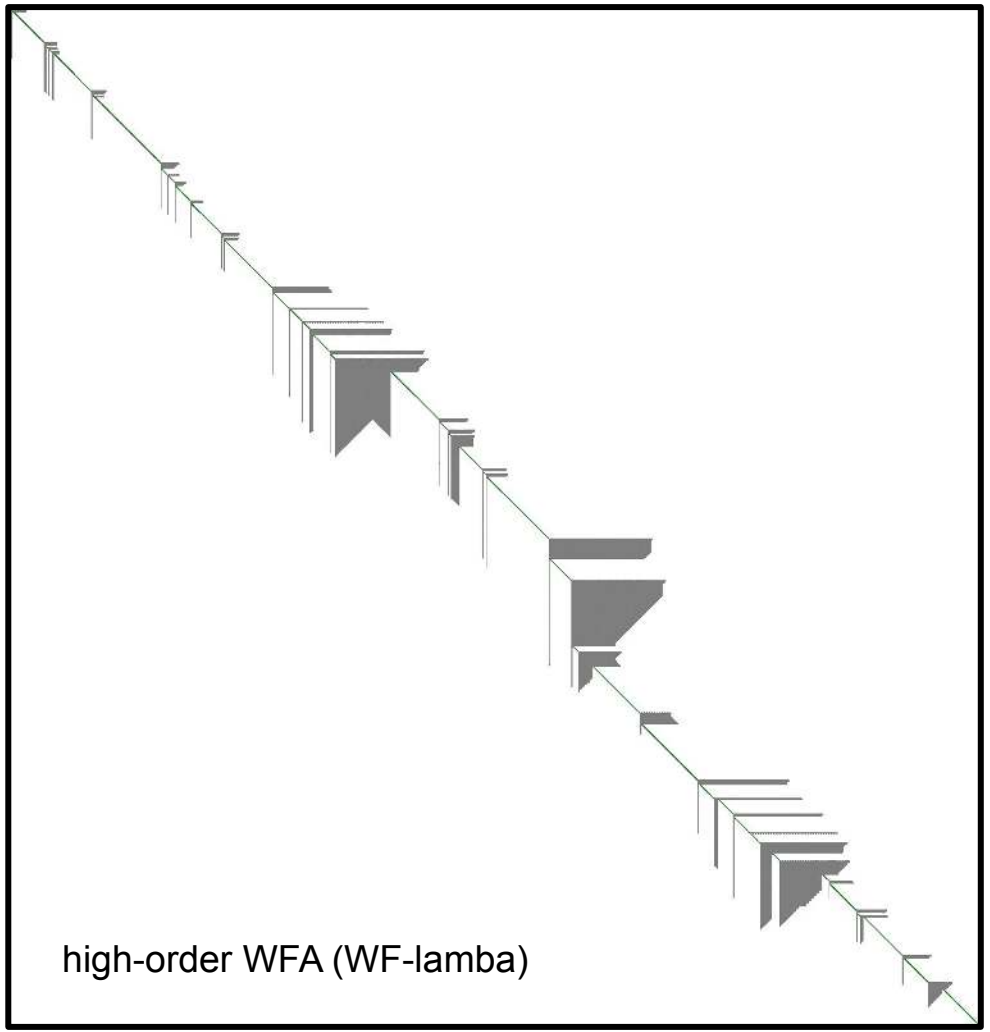
smoothxg (graph normalization)

wfmash makes initial alignments



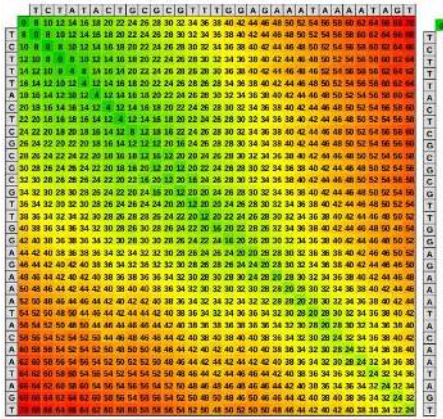
Needleman-Wunsch

the wavefront algorithm (WFA)

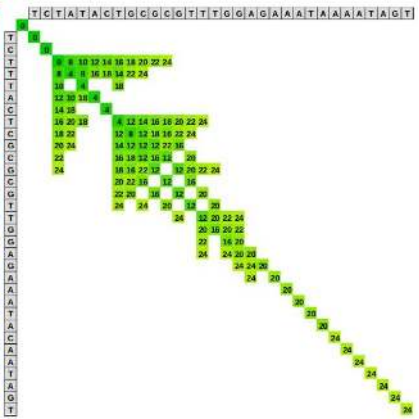


high-order WFA (WF-lamba)

wfmash makes initial alignments



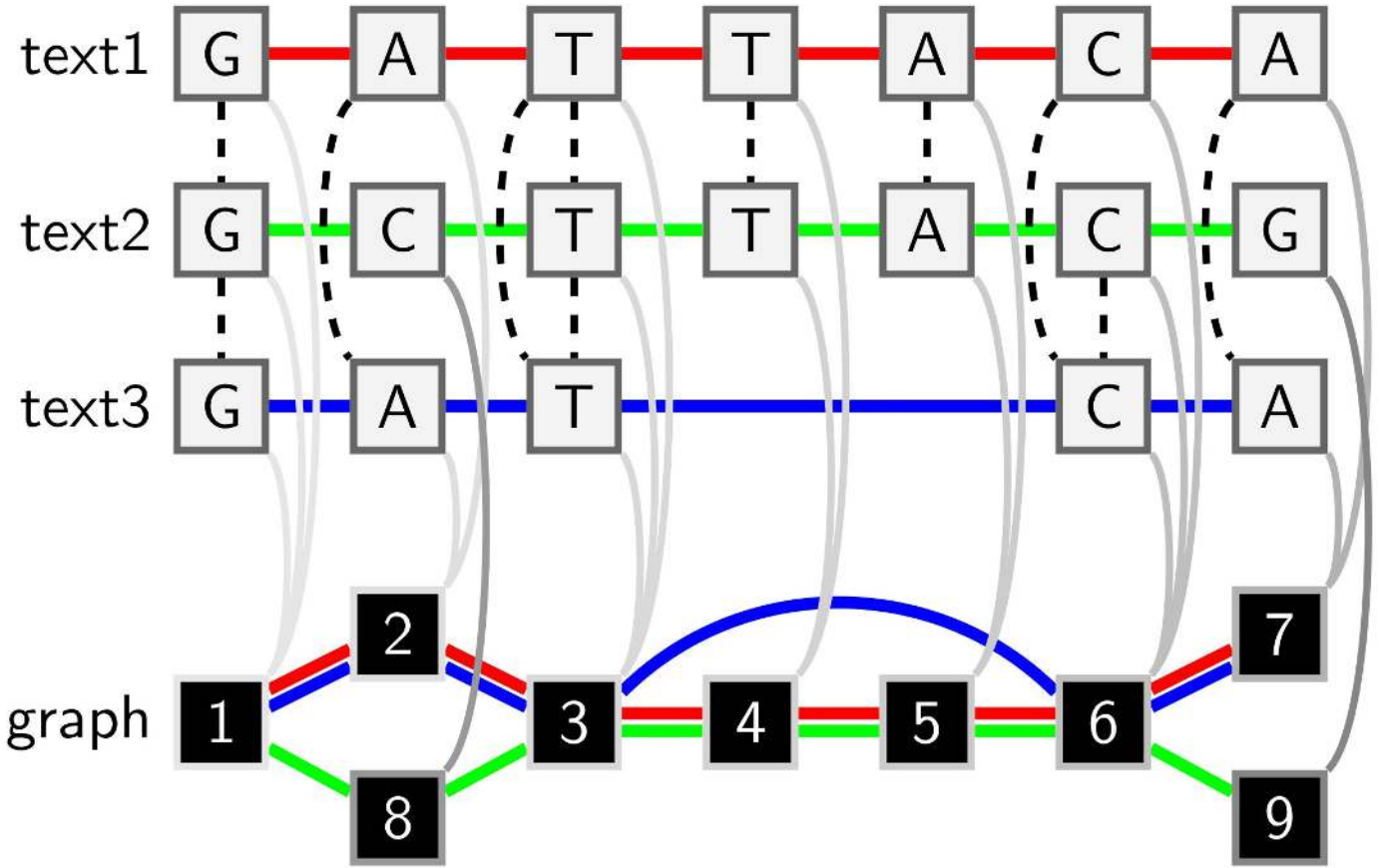
Needleman-Wunsch



the wavefront algorithm (WFA)

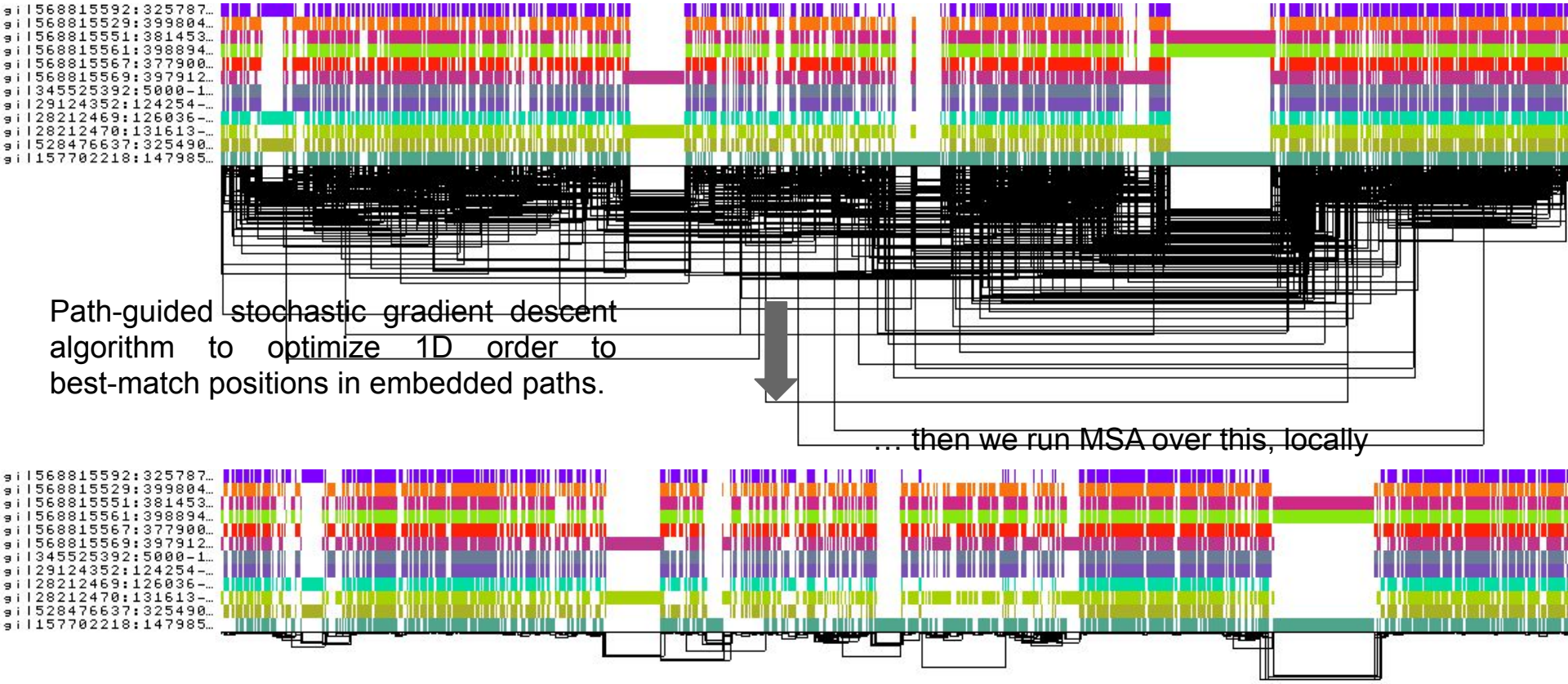
high-order *bidirectional* WFA (BiWFλ)

seqwish builds the graph



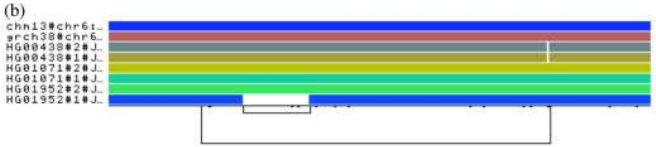
smoothxg organizes & normalizes the graph

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

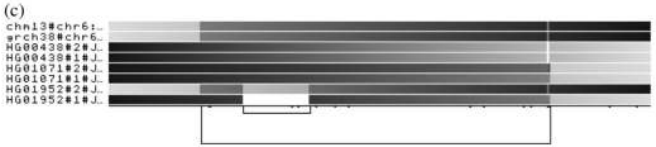


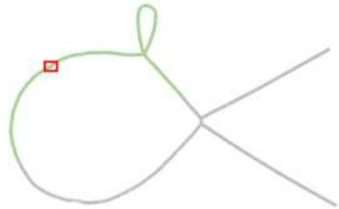
2d layout

identity



position

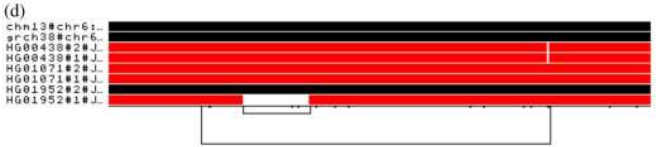




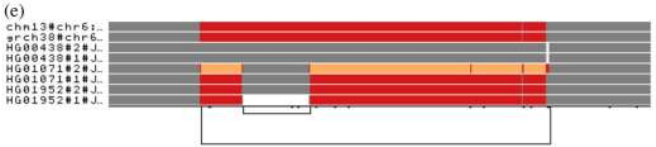
ODGI is meant to be a basic toolkit for interacting with pangenome graphs.

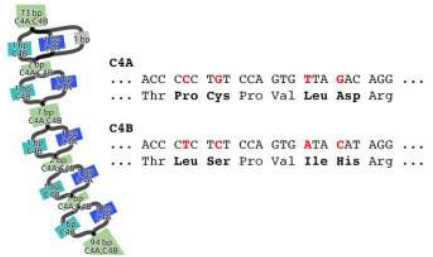
It uses the embedded genomes as references.

orientation

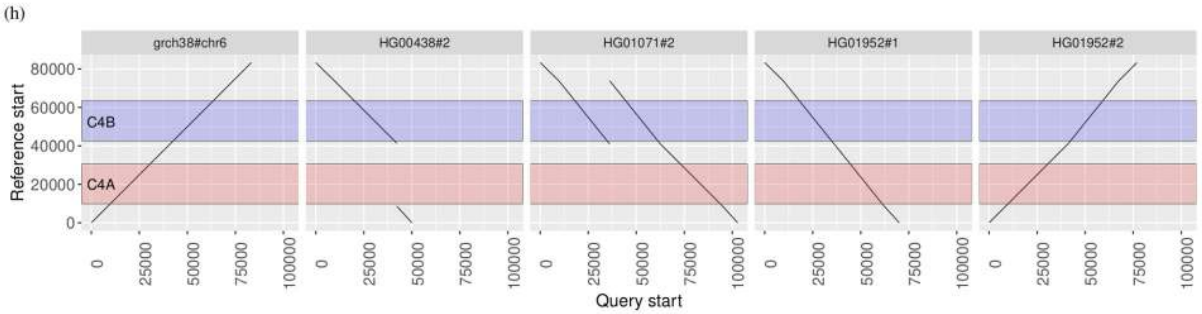


copy number variation



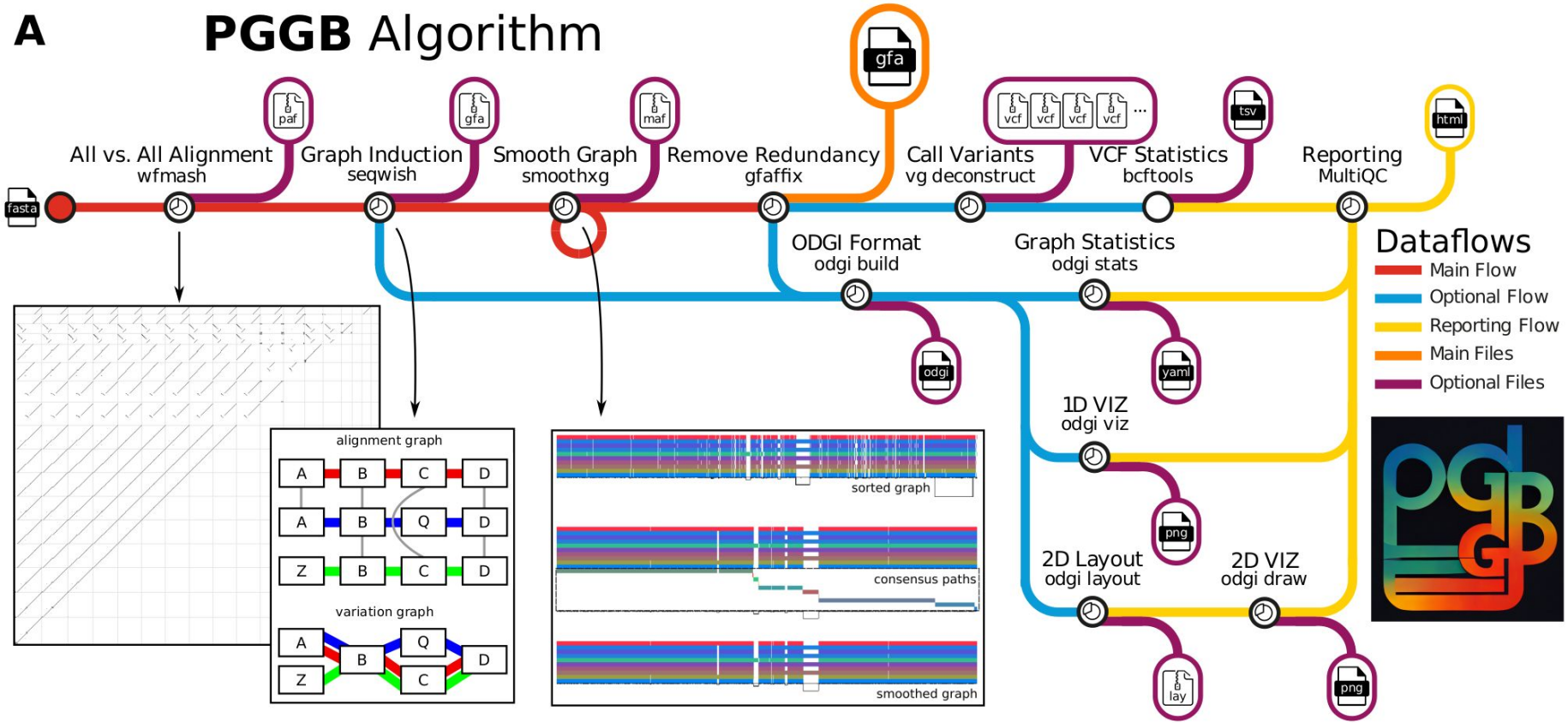


odgi helps us understand the pangenome



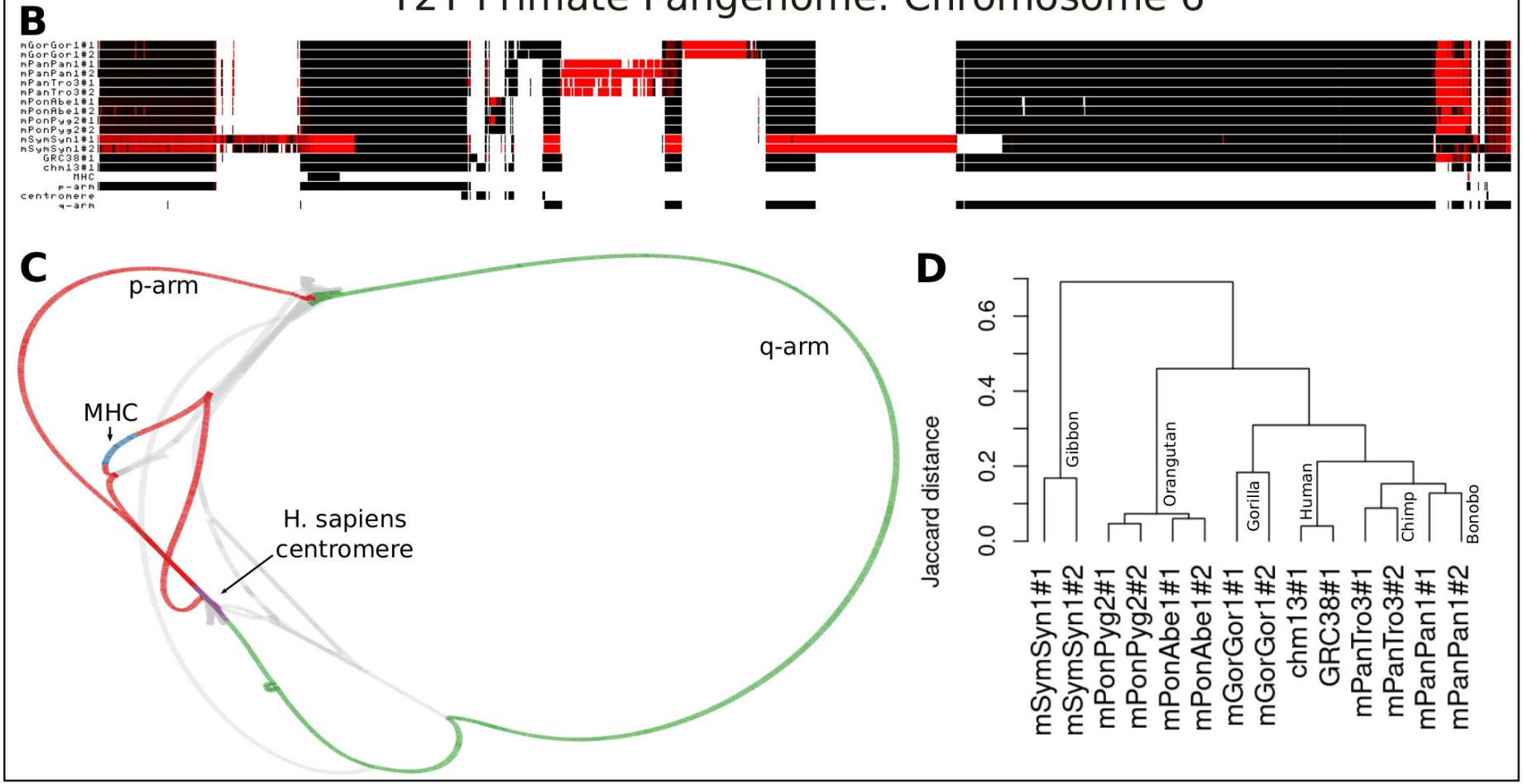
https://www.biorxiv.org/content/10.1101/2021.11.10.467921v1

Putting it all together!



https://doi.org/10.1101/2023.04.05.535718

T2T Primate Pangenome: Chromosome 6

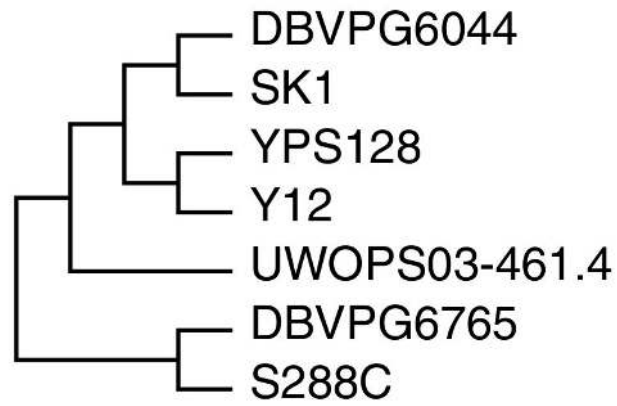


Test material today

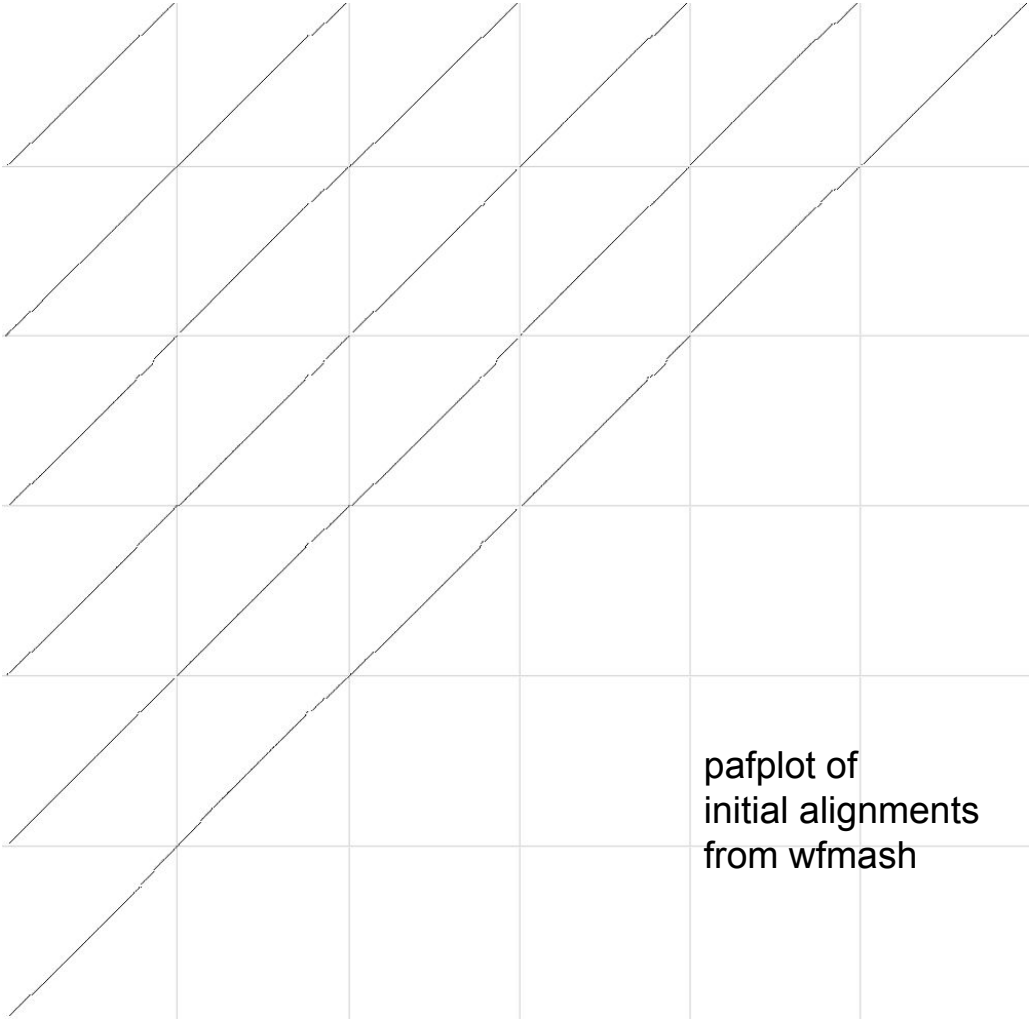
1. A few genes from HLA-D (MHC class II) in humans — getting started
    a. https://github.com/pangenome/hprc-workshop/tree/evomics2023
2. Yeast chromosome 6 — scaling up
    a. ~/workshop_materials/pangenomics/cerevisiae.chrV.fa
    b. you will want to apply samtools faidx to this… pggb will warn you
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

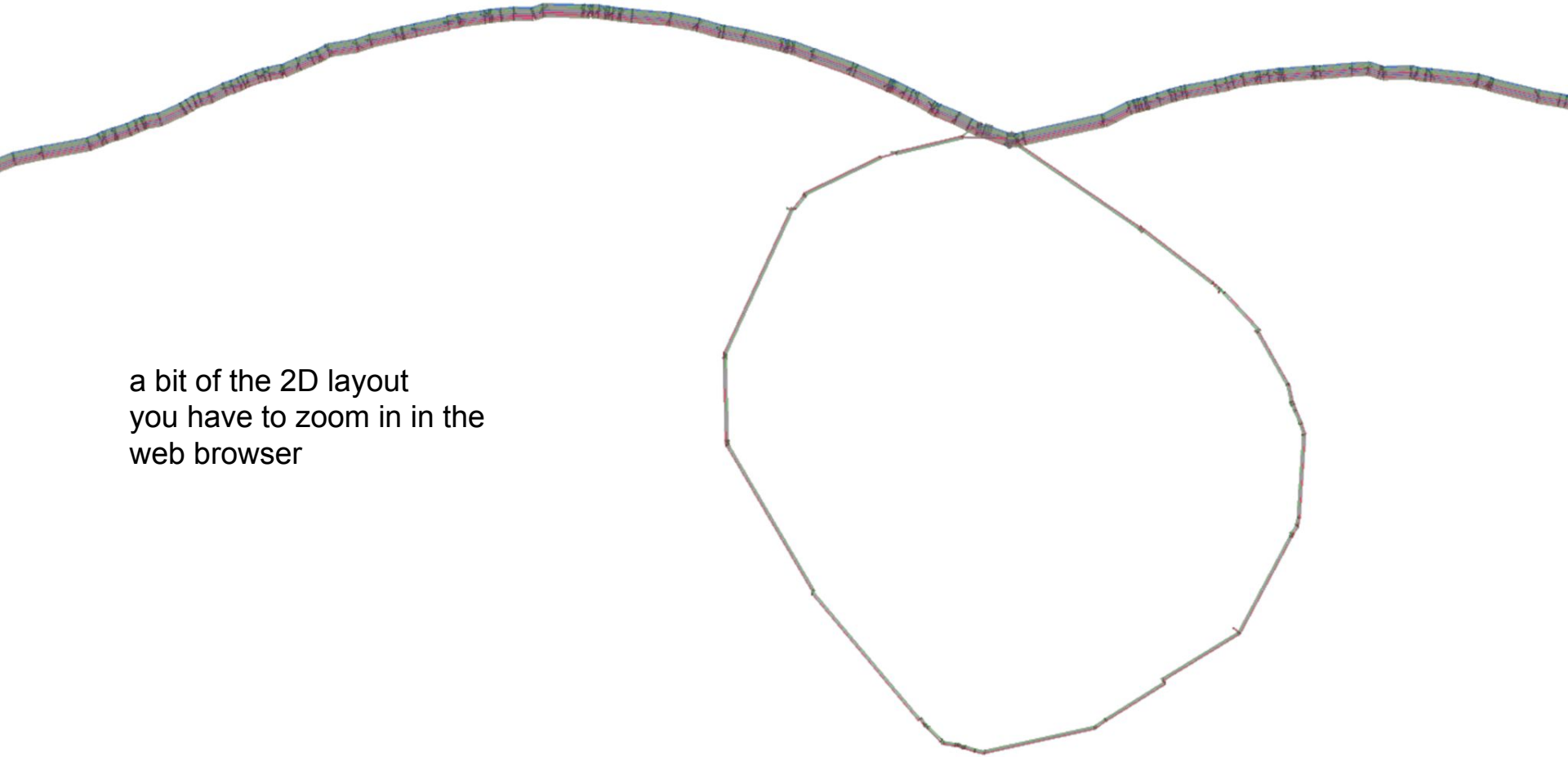


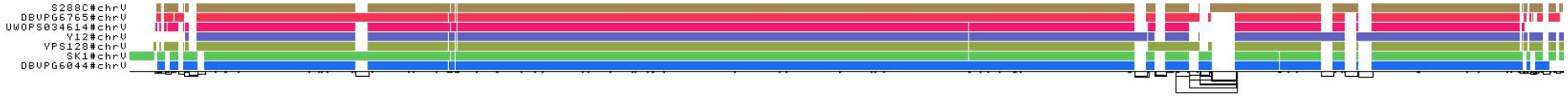
Cladogram of the *S.c.* clade



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

these are all quite boring here…

Conclusions

- Variation graphs solve a key conceptual problem in bioinformatics: joint representation of many genomes and their alignment.
- We can apply them to improve basic bioinformatic analyses.
- Building pangenomes benefits from an unbiased approach where many reference genomes are treated equivalently.
- Applying these methods to the human pangenome reveals patterns of recombination between heterologous acrocentric chromosomes.

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)

Recombination between 15p and 22p.

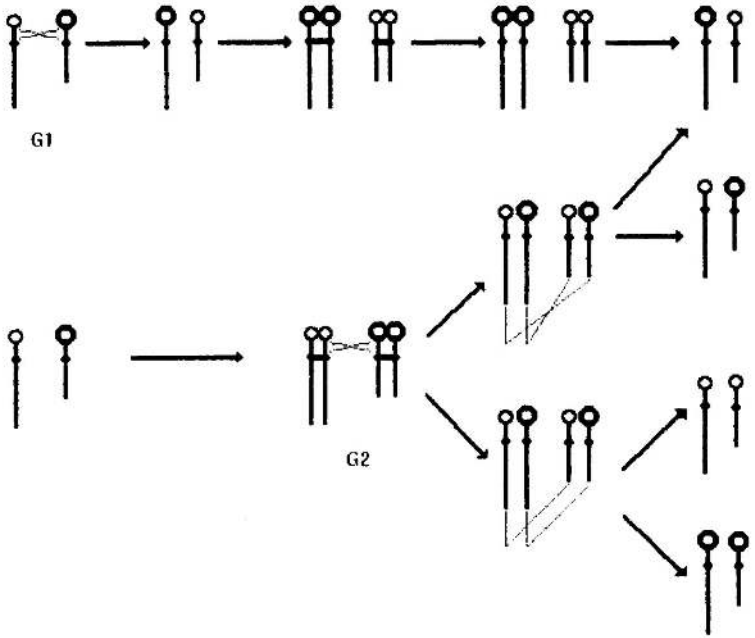
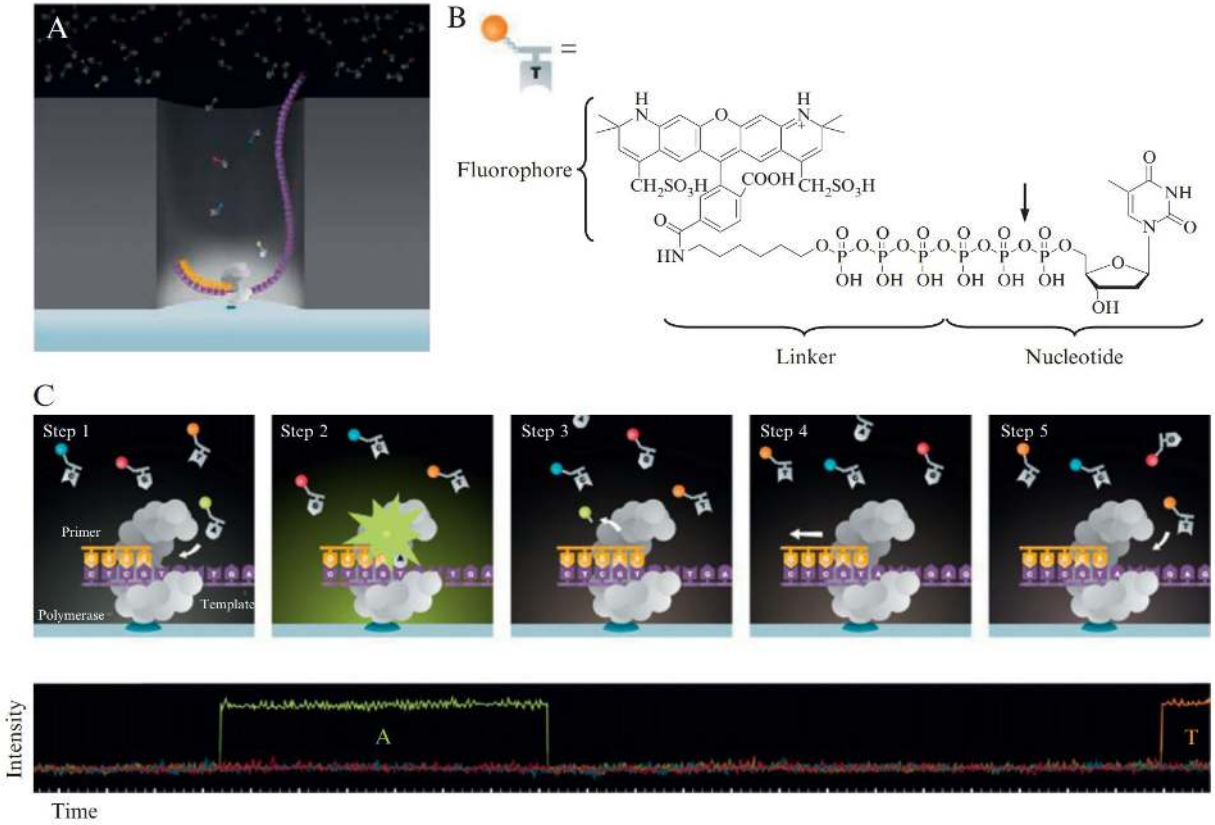


Fig. 3. Diagrammatic representation of how chromosome and chromatid exchanges may produce the different cell lines. Both a G_1 (a) and a G_2 (b) exchange may produce a shift of the polymorphism s+ from chromosome 22 to another acrocentric. The same G_2 exchange would result in absence (c) or presence (d) of the same polymorphism on chromosome 22 and on another acrocentric, if the alternative segregation occurs.

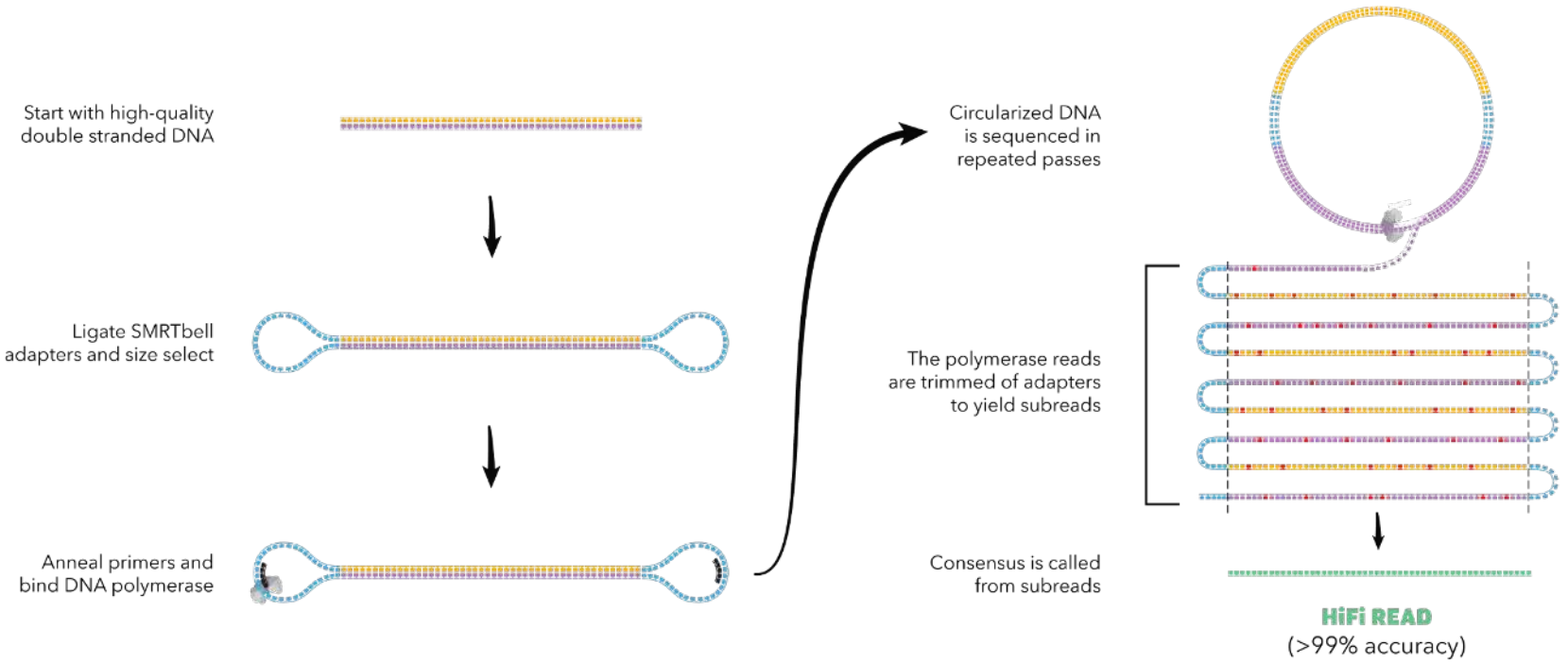
https://doi.org/10.1111/j.1469-1809.1996.tb01180.x

Zero-Mode Waveguide imaging of single polymerases



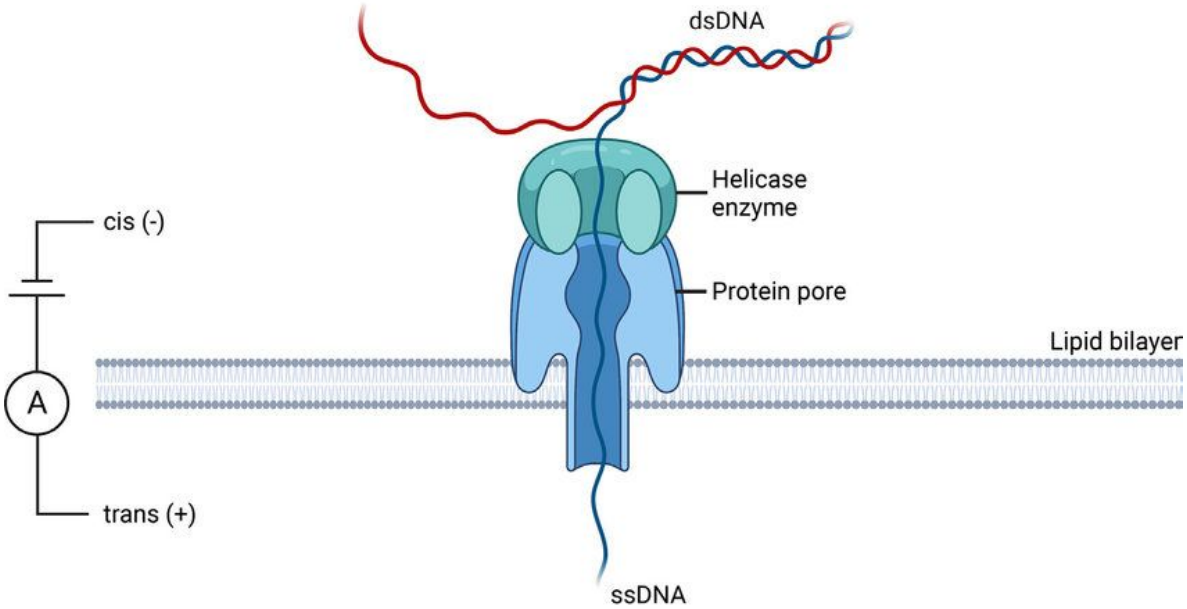
https://doi.org/10.1016/S0076-6879(10)72001-2

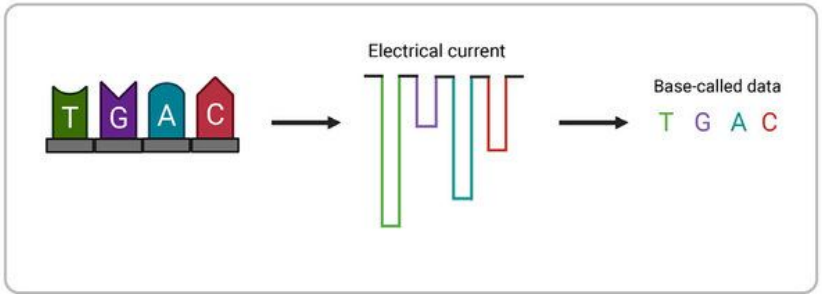
PacBio HiFi produces incredibly accurate long reads



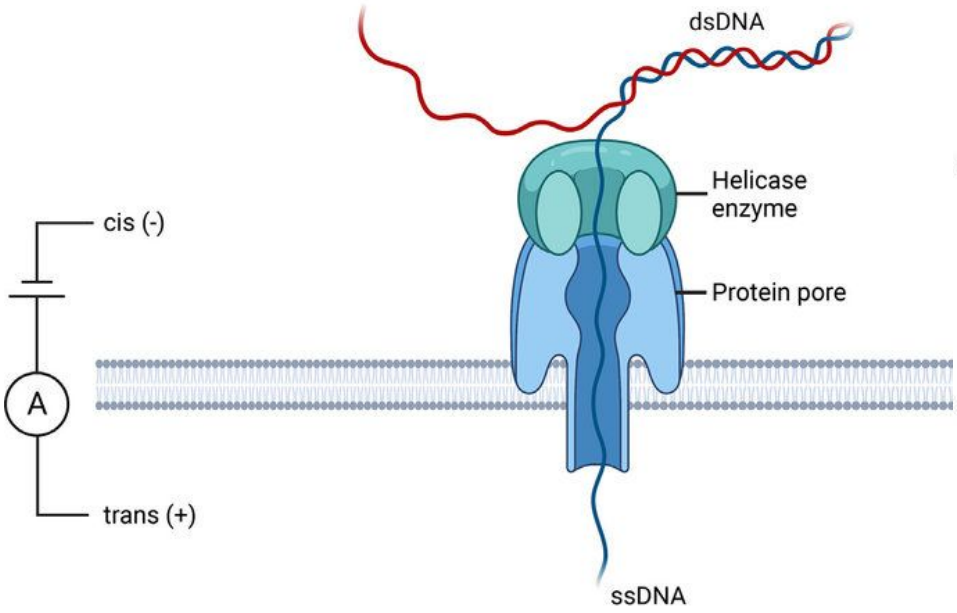
https://www.pacb.com/wp-content/uploads/How-to-get-HiFi-reads_v2.png

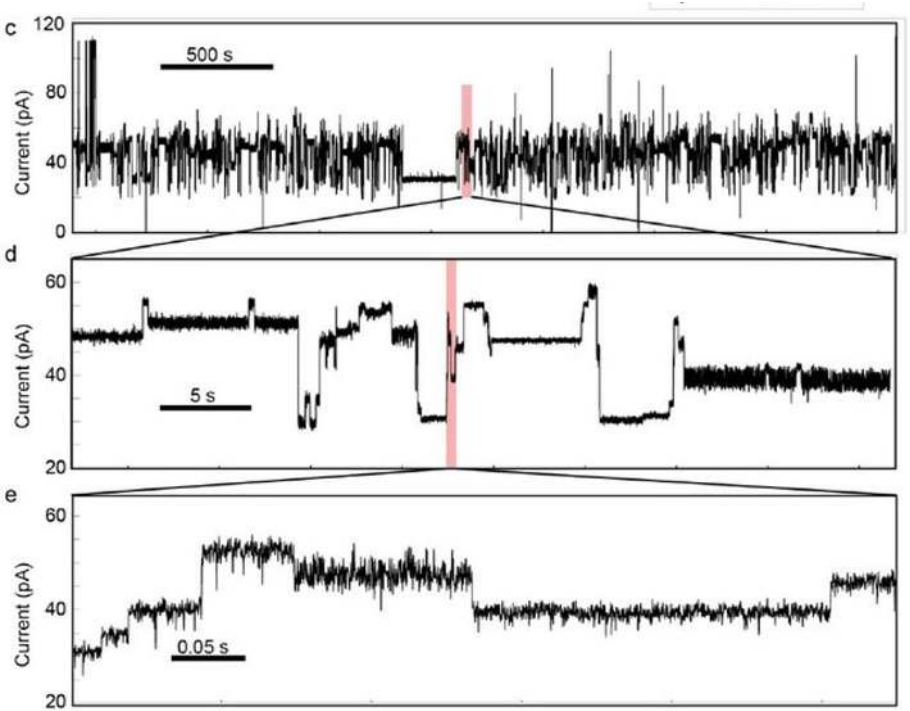
Nanopore sequencing

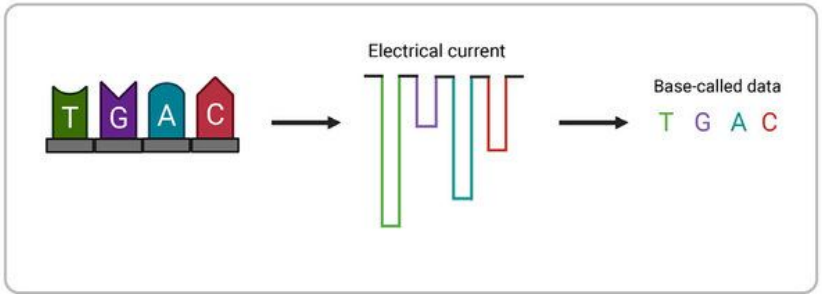




Nanopore sequencing





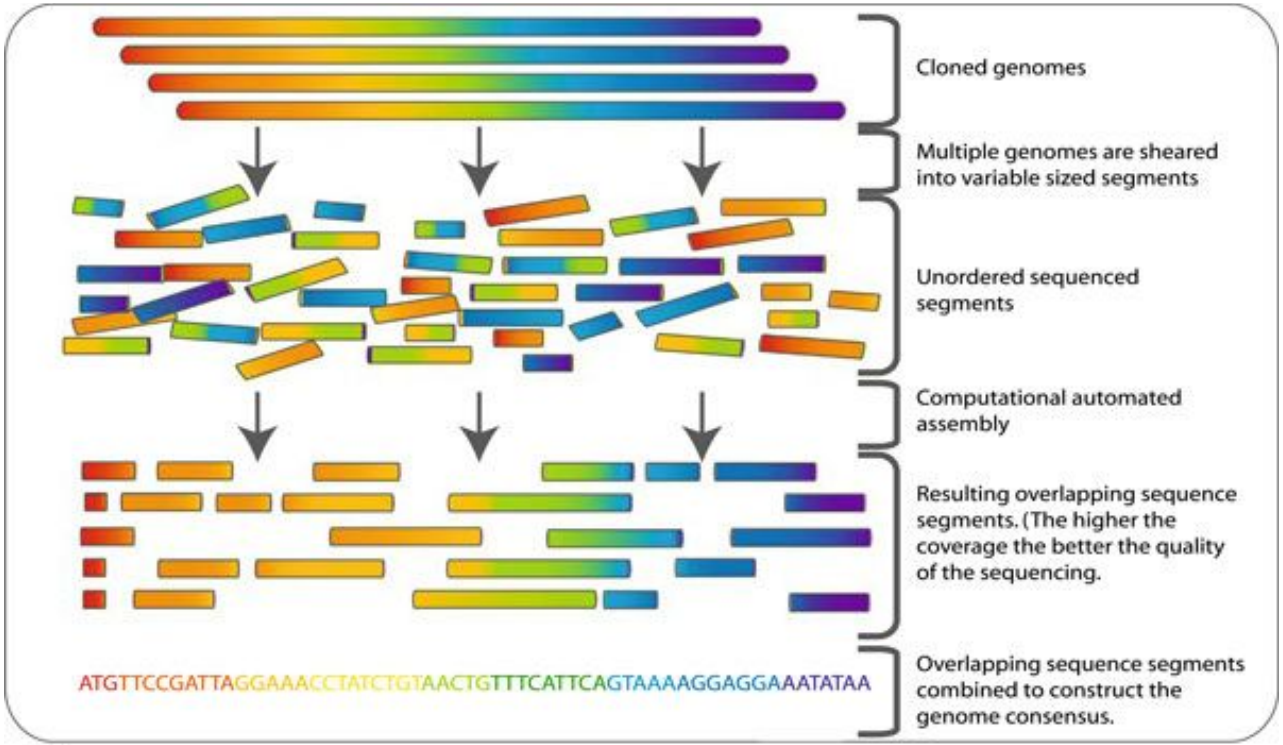


Reads are very noisy, but ***very*** long.

Preliminaries: Sequencing and assembling genomes

So you want to look at genomes?

We observe genomes by sequencing fragments of DNA and assembling them computationally.



https://doi.org/10.1007/s12575-009-9004-1

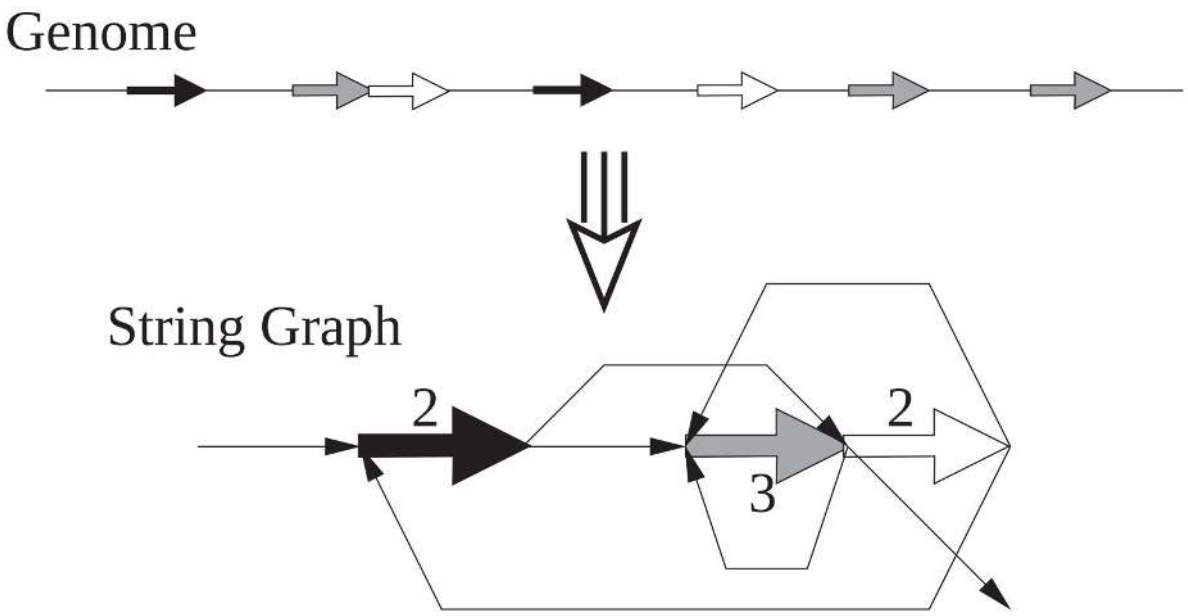
Biology: “Sorry, it’s not that simple”

Genomes evolve by copying. They are full of repeats.

Repeats make it hard to assemble genomes.

If our sequence reads are shorter than repeat sequences, the best we can do is the string graph, where repeats collapse.

<u>Before 2019</u>, *our sequence reads were shorter than repeats.*



Myers’ string graph
https://doi.org/10.1093/bioinformatics/bti1114

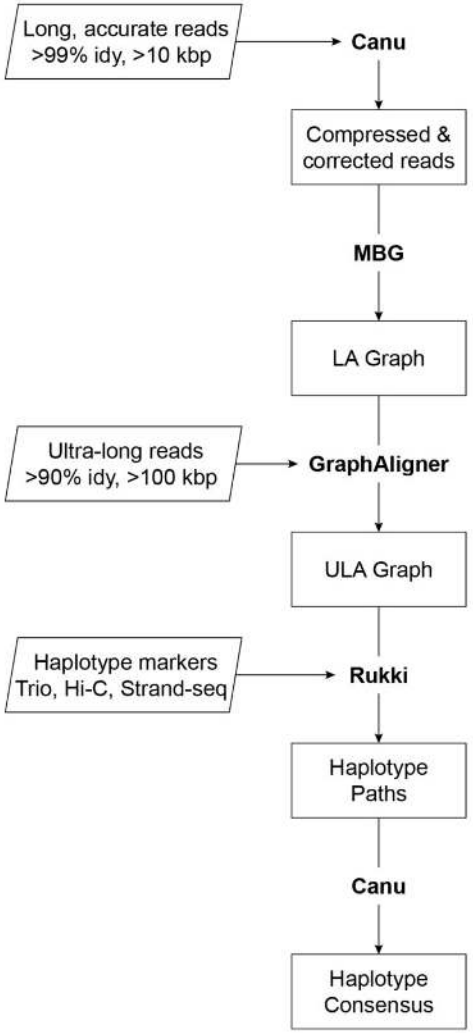
Preliminaries: Sequencing and assembling genomes ***completely!!***

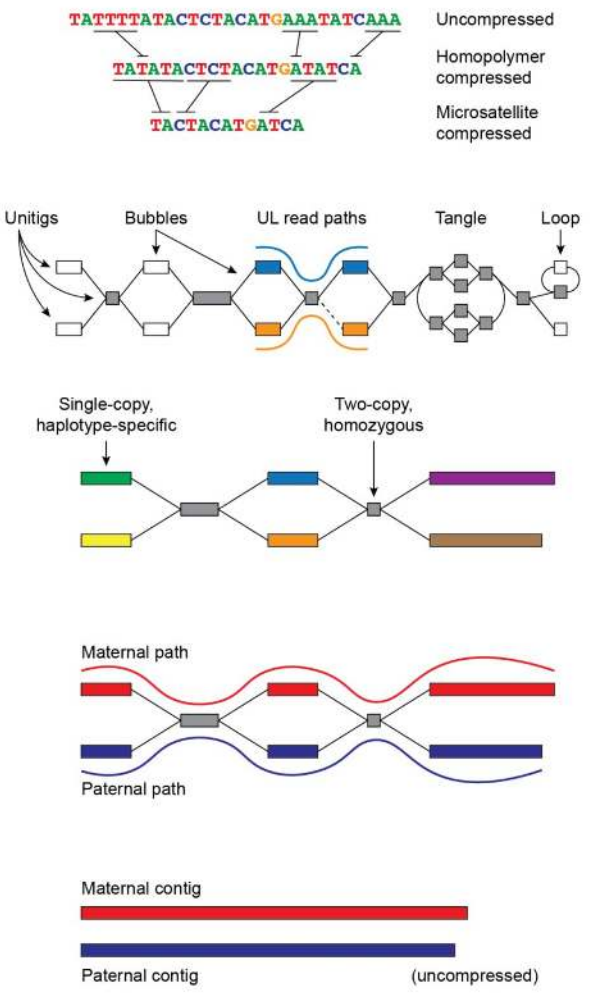
Verkko: automated diploid assembly

Building on top of 3rd gen sequencing, we can automatically generate haplotype resolved, highly accurate (QV>50 or 1/100,000 base error).

This depends on two new bioinformatic approaches.

- long-kmer de Bruijn graphs
- sequence to graph alignment (GraphAligner) *a pangenomic method





Verkko assembly graph for HG002

This diploid, haplotype resolved assembly was a first. The easy sequencing (only two types) will facilitate many more in the near future!