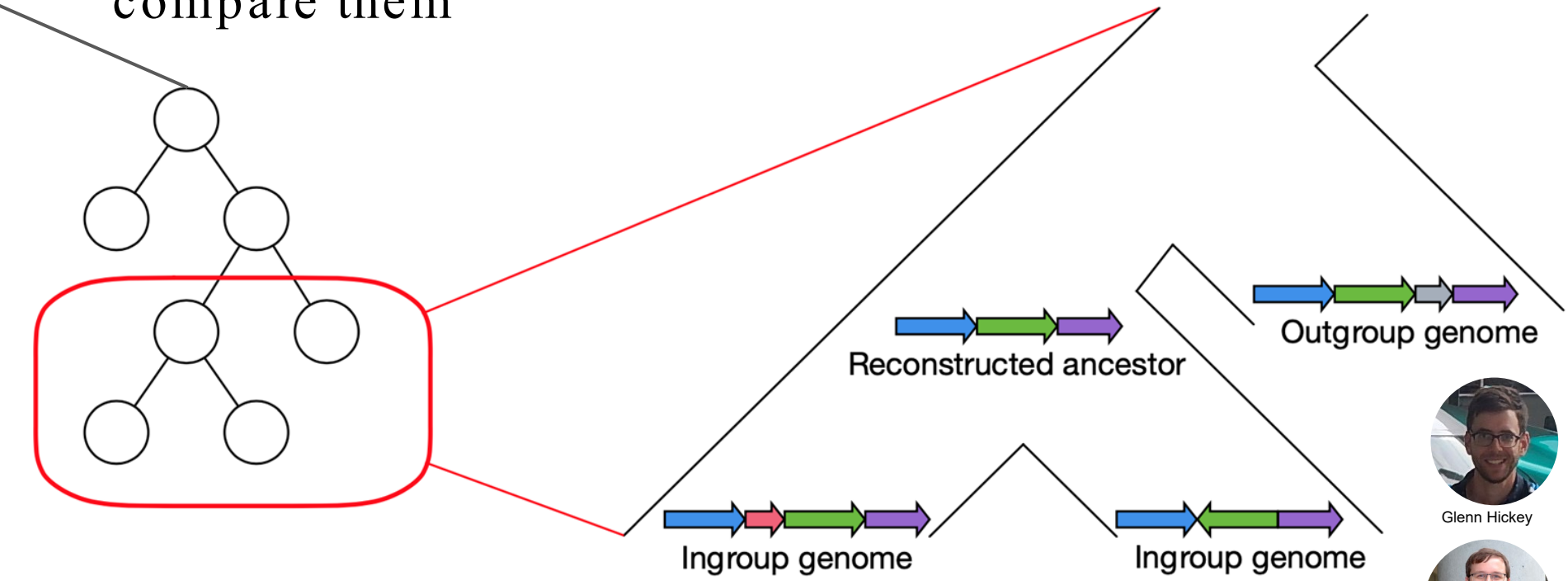
Furthering our understanding of human genetic variation: the human pangenome reference project second release

Benedict Paten, Professor, Biomolecular Engineering
Associate Director, UC Santa Cruz Genomics Institute

Santa Cruz Breakwater Lighthouse, photo courtesy Kishwar Shafin

UNIVERSITY OF CALIFORNIA
SANTA CRUZ | Genomics Institute

To understand genomes you need to compare them



Glenn Hickey


Joel Armstrong

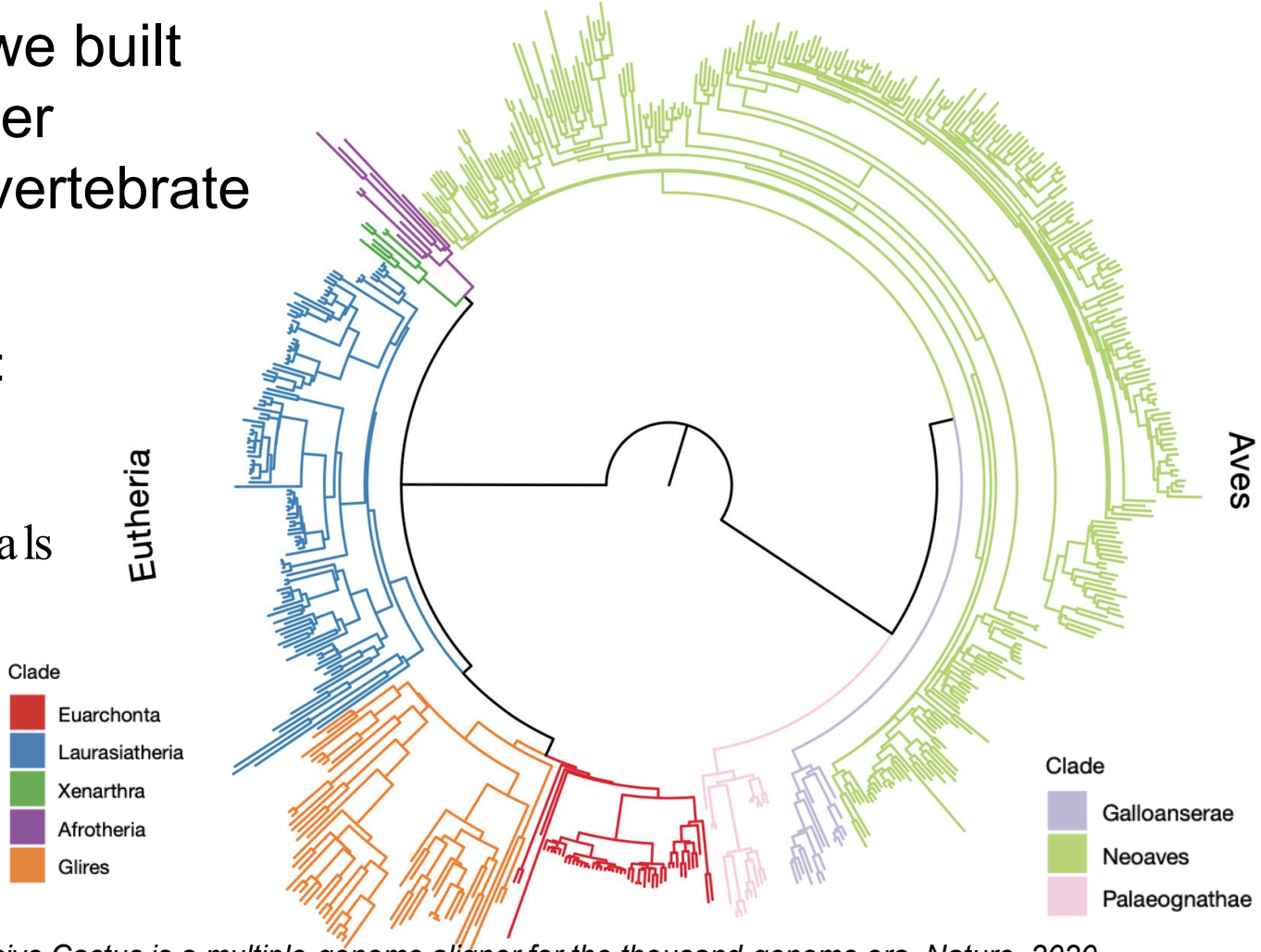
Armstrong J., et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era, Nature, 2020

With Cactus we built the largest ever alignment of vertebrate genomes

605 genomes:

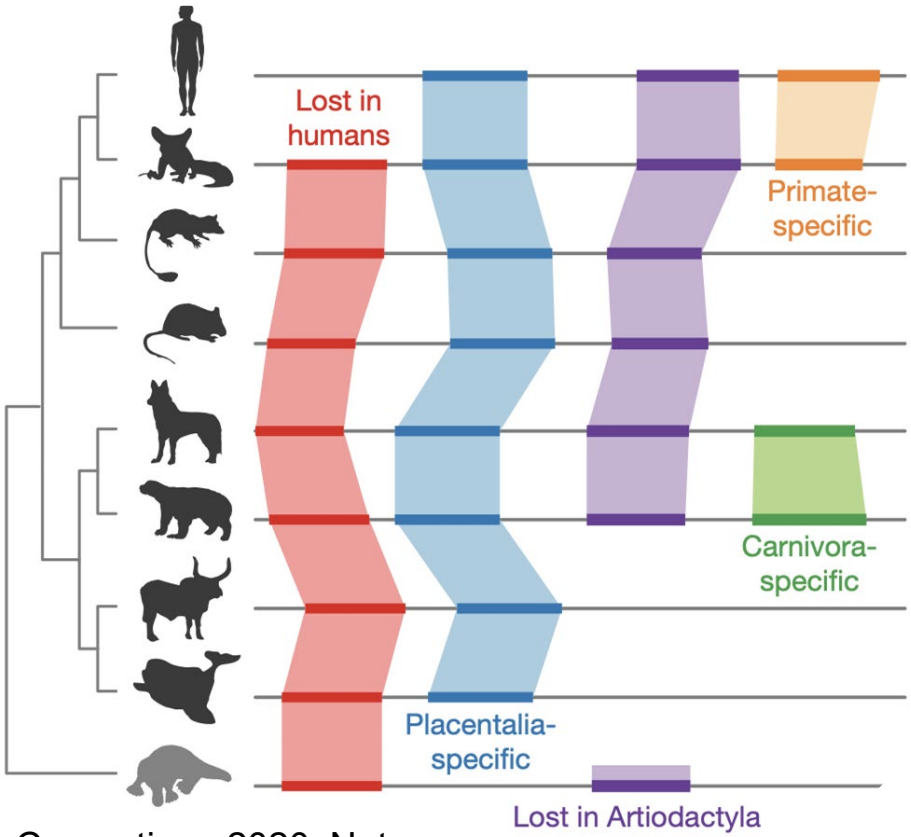
363 birds

242 mammals



Armstrong J., et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era, Nature, 2020

Cactus Makes Reference Free Genome Alignments



Adapted from Zoonomia Consortium, 2020, Nature

Coverage Vs. (Chicken-referenced) MultiZ

Scale: Aligning Everything

There are an estimated 8.7 million species on earth!

It took 2.1 million core hours to align 605 species.

We are very focused on further simplifying and accelerating Cactus

EARTH BIOGENOME PROJECT

VGP VERTEBRATE GENOMES PROJECT

A PROJECT OF THE G10K CONSORTIUM

Alignment	No. of genomes	Total bases	Instance-hours	Core-hours	Common ancestor size
Zoonomia	242	669 billion	68,166	1.9 million	1.73 Gb
B10K	363	400 billion	5,302	0.2 million	1.13 Gb
Combined	605	1.07 trillion	73,692	2.1 million	181 Mb

* https://www.nature.com/articles/news.2011.498

MAF

- We (genome aligners) all use MAF - a human readable, column based alignment format
- MAF has problems:
  - Block fragmentation
  - Verbosity
  - No standard for compression, indexing
- As we scale to thousands of genomes, let’s rethink MAF

(A): MAF (602 bytes)

```
##maf version=1 scoring=N/A

a
s       dog.chr6     437451  11      +       593897  CCCGTCAGTGT
s       human.chr6   446327  11      +       601863  TCCGCCAAGGT
s       mouse.chr6   460751  11      +       636262  TTCATCAGAGT
s       rat.chr6     470339  11      +       647215  TTCATTAGGGT

a
s       cow.chr6     445326  8       +       602619  TTTTCCCA
s       dog.chr6     437462  8       +       593897  TT-TTCCG
s       human.chr6   446338  8       +       601863  TTCTTCCG
s       mouse.chr6   460762  8       +       636262  TTTTACCG
s       rat.chr6     470355  8       +       647215  TTTTACCG
```

TAF

- We propose TAF
- TAF:
  - Is still human readable (roughly)
  - Does not fragment - no blocks
  - Is less verbose
  - Has indexing and compression builtin - you can random access on .gz files
  - Supports per column tag annotations
  - Has a C, Python and CLI called Taffy*, with Pytorch utils and viz scripts
- Future work: integrate with GFA to represent non-linearities

https://github.com/ComparativeGenomicsToolkit/taffy

(A): MAF (602 bytes)

```
##maf version=1 scoring=N/A

a
s       dog.chr6     437451  11      +       593897  CCCGTCAGTGT
s       human.chr6   446327  11      +       601863  TCCGCCAAGGT
s       mouse.chr6   460751  11      +       636262  TTCATCAGAGT
s       rat.chr6     470339  11      +       647215  TTCATTAGGGT

a
s       cow.chr6     445326  8       +       602619  TTTTCCCA
s       dog.chr6     437462  8       +       593897  TT-TTCCG
s       human.chr6   446338  8       +       601863  TTCTTCCG
s       mouse.chr6   460762  8       +       636262  TTTTACCG
s       rat.chr6     470355  8       +       647215  TTTTACCG
```

(B) TAF (265 bytes)

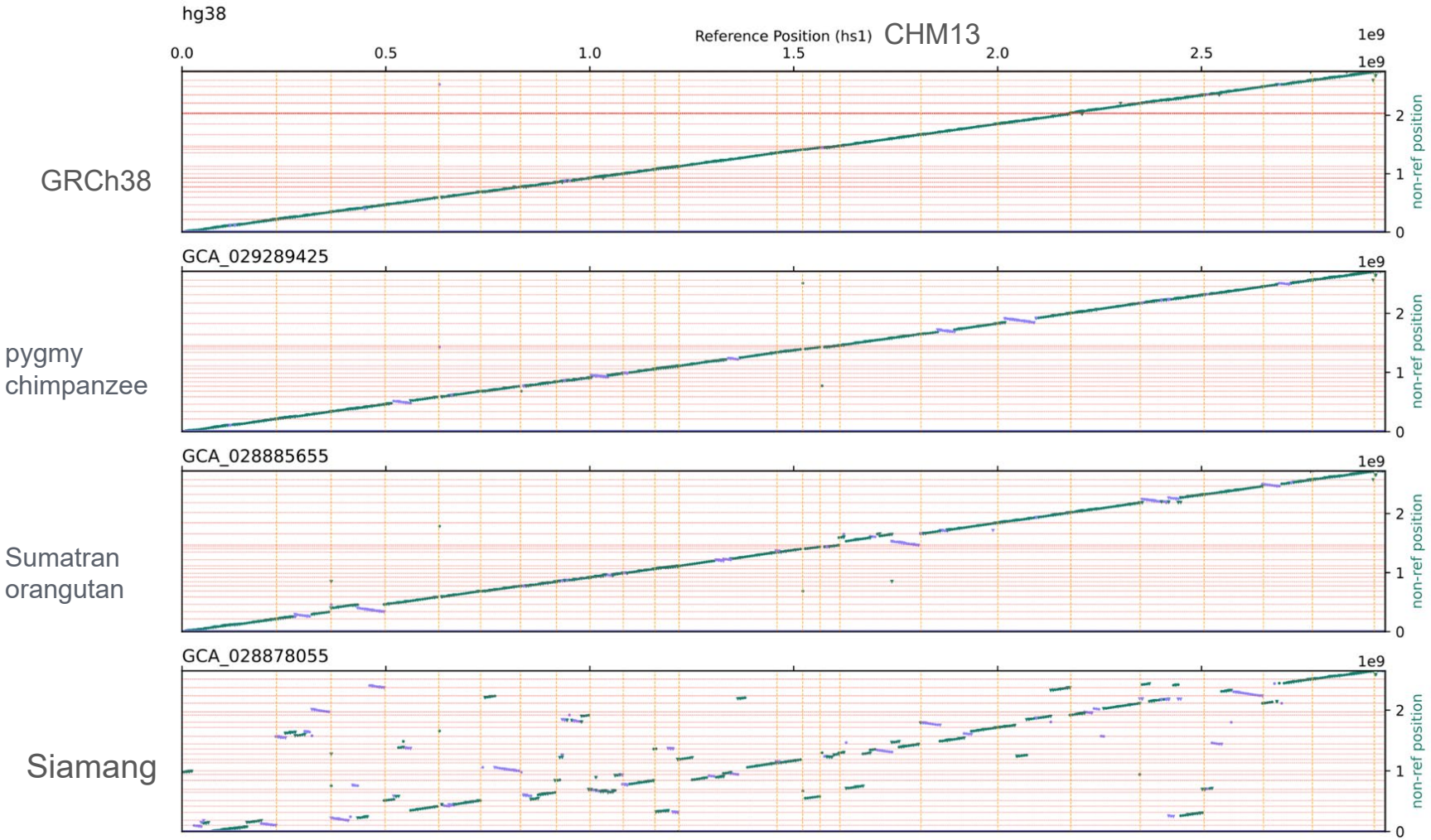
```
#taf version:1 scoring:N/A
CTTT ; i 0 dog.chr6 437451 + i 1 human.chr6 446327 + i 2 mouse.chr6
460751 + i 3 rat.chr6 470339 11
CCTT
CCCC
GGAA
CCCT
AAAA
GAGG
TGAG
GGGG
TTTT
TTTTT ; i 0 cow.chr6 445326 + g 4 5
TTTTT
T-CTT
TTTTT
CTTAA
CCCCC
CCCCC
AGGGG
```

TAF

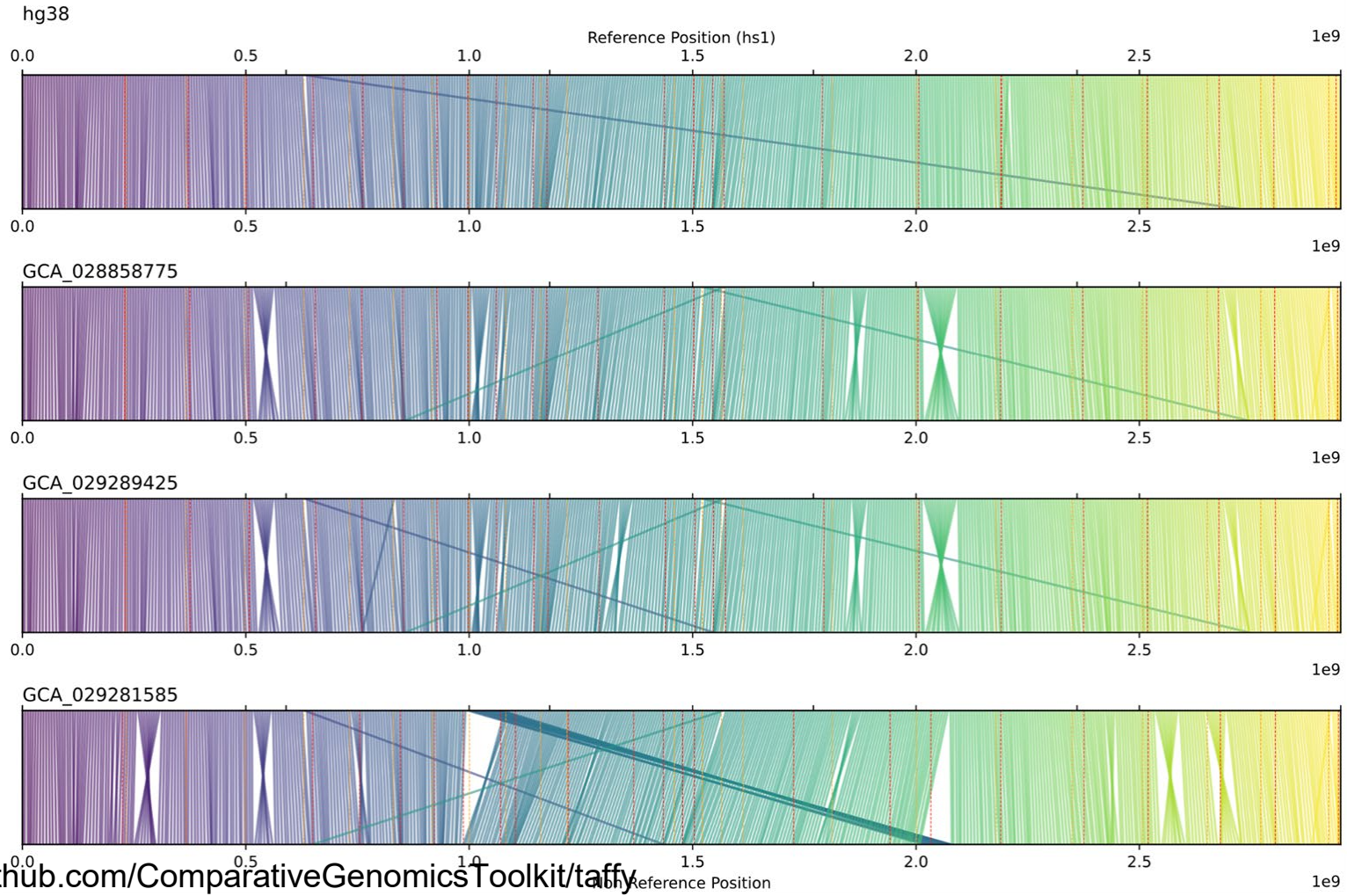
File suffix	Size (Terabytes)	Compression relative .maf
.maf	7,817	1.0x
.maf.gz	2.098	3.7x
.taf	1.075	7.3x
.taf.gz	.296	26.4x

Table 1: MAF vs. TAF file size for a 447-way mammalian genome alignment.

https://github.com/ComparativeGenomicsToolkit/taffy



https://github.com/ComparativeGenomicsToolkit/taffy



https://github.com/ComparativeGenomicsToolkit/taffy

Human Reference Genome

The current human reference genome (GRCh38) is the cornerstone of human genomics

It is a proxy to a universal coordinate system for human genetics

It originally cost $3B and took an act of congress

Released in 2001, it has been refined over 20 years

Originally it was built to represent the euchromatin, it is still incomplete..

The First Complete, Haploid Human Genome

20 years after the human genome Karen Miga (UCSC), Adam Phillippy (NHGRI), et al. released the first complete assembly of a haploid human genome, T2T-CHM13

TELOMERE-TO-TELOMERE CONSORTIUM

Missing Polymorphic Sequence

- There are >100 megabases of commonly polymorphic euchromatic sequence missing from any individual reference
- As a result, no single reference assembly, even a complete one, is optimal for all people, because any reference creates a bias away from the missing sequence

Reference bias is an observational bias, aka streetlamp effect: it is harder to find something not in the reference.

Call to Action: A Human Pangenome Reference

HUMAN PANGENOME

NIH
NHGRI

- Better representation of sequence diversity in the human population (>350 diverse humans)
- Comprehensive, public map of genome variation
- New reference data structure and nucleate and foster a new ecosystem of pangenome tools

First Release: A Draft Human Pangenome Reference*

HUMAN PANGENOME

NIH
NHGRI

- 47 phased, diploid genome assemblies (~1/7th of final cohort)
- Pangenome alignments, annotations
- Pangenome tools and applications

* Liao, Asri, Ebler, et al. A Draft Human Pangenome Reference, Nature, 2023

Human Pangenome

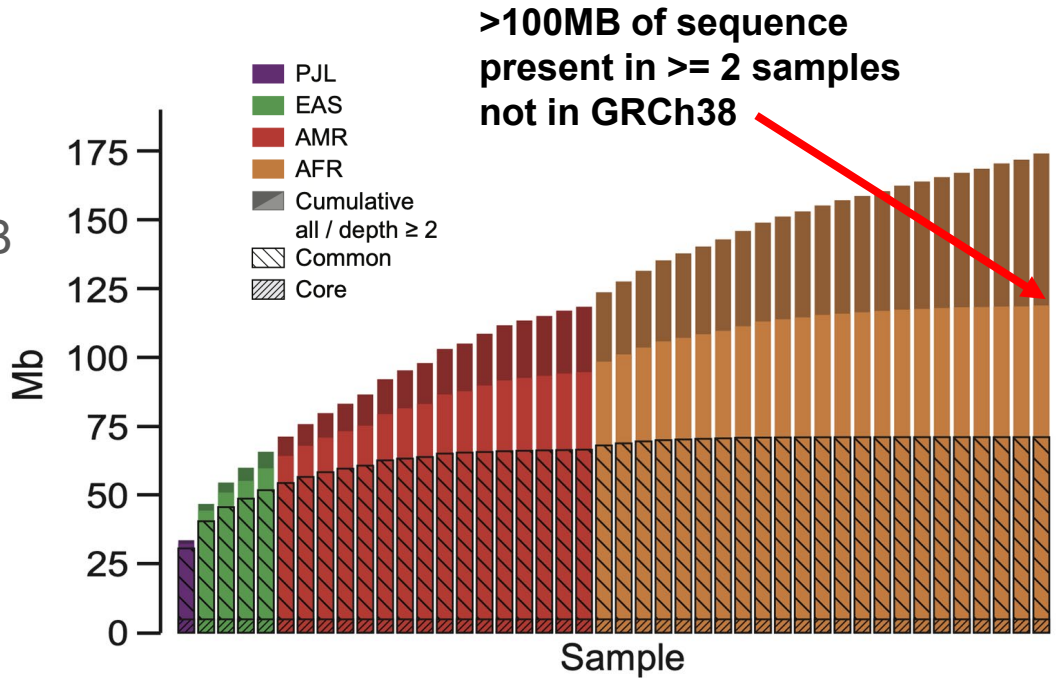
Defined by three As:

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

Goal: provide a comprehensive view of common human variation

Pangenome: Adding Common, Polymorphic Sequence To The Reference

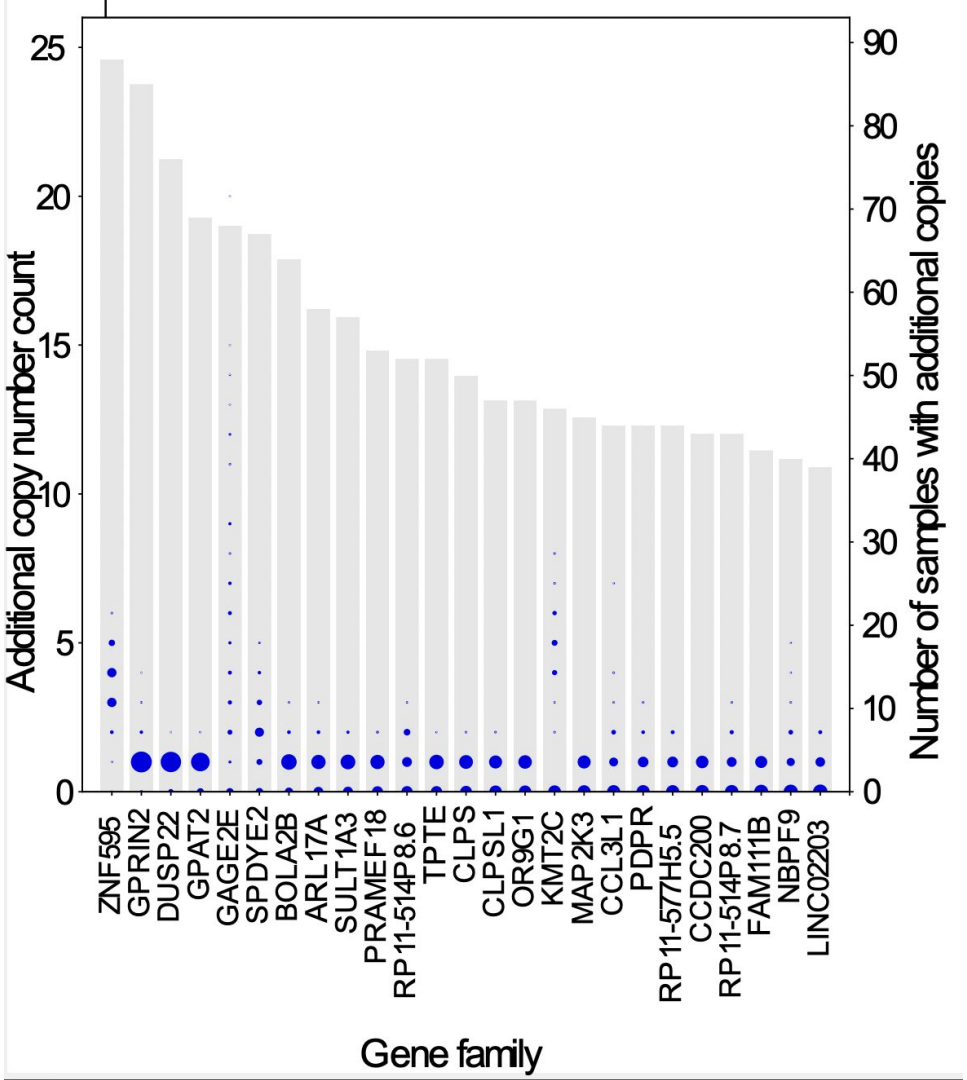
- T2T-CHM13 adds ~200MB of (principally) heterochromatin to reference (6-7%)
- Draft pangenome adds >100MB of common, polymorphic euchromatin (3-4%) (and a **lot** more heterochromatin)
- 0.6-4.4 Mb of additional genic sequences per haplotype compared to GRCh38 (38 gene CNVs/haplotype)



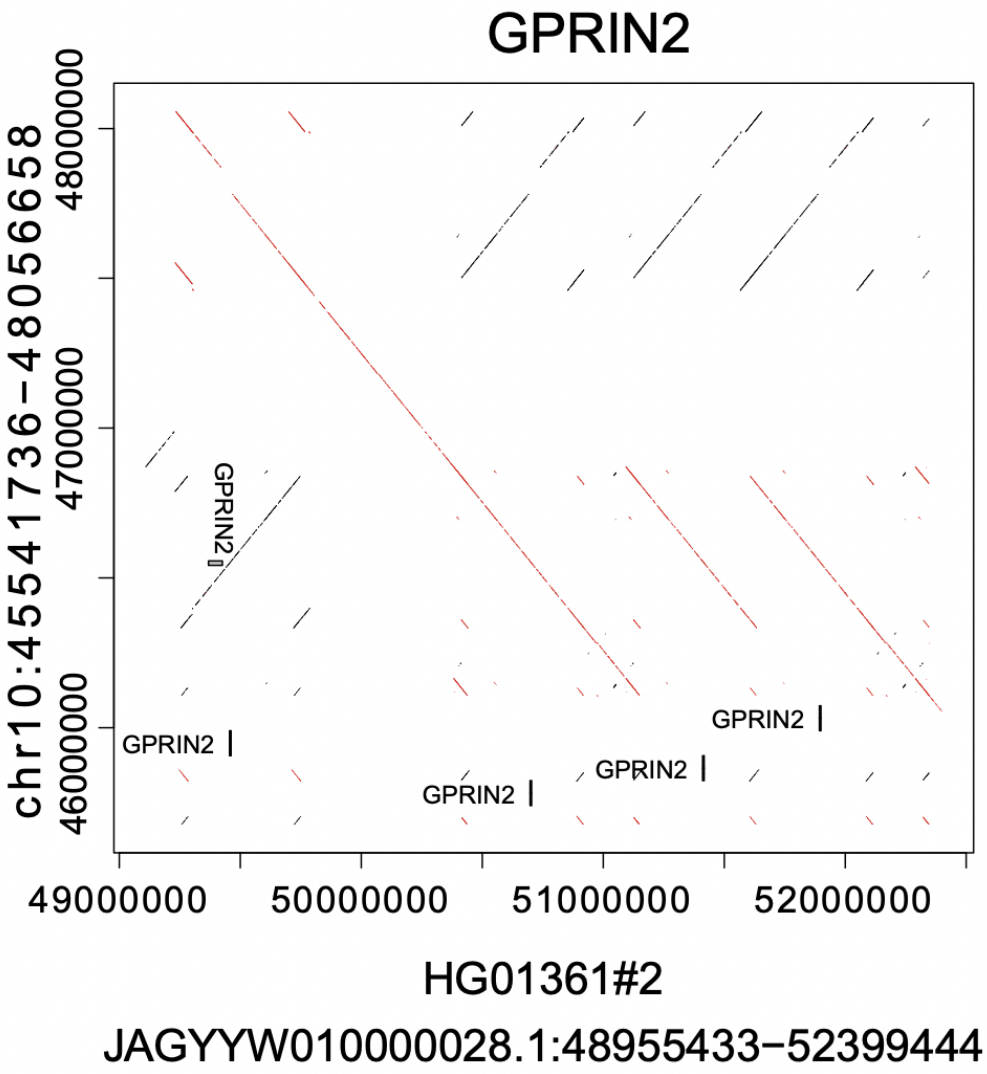
Haplotype Resolved Gene Duplications

Grey bars: number of samples with additional copies

Blue dots: CNVs per haplotype - size of dot proportional to # haplotypes



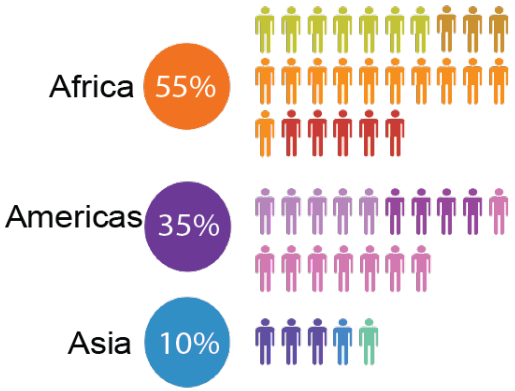
Credit Marina Haukness, Mark Chaisson



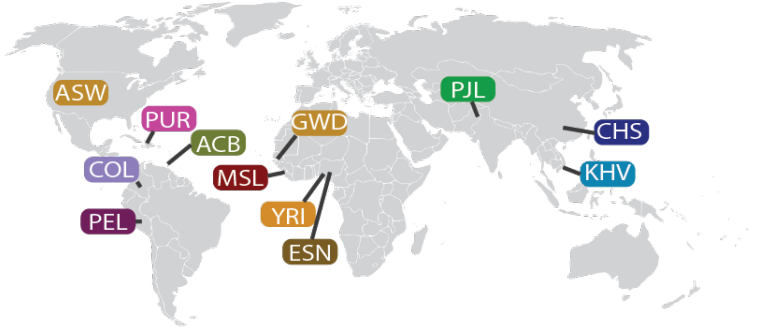
Credit Mark Chaisson

Population Representation and Sampling: Draft Selection

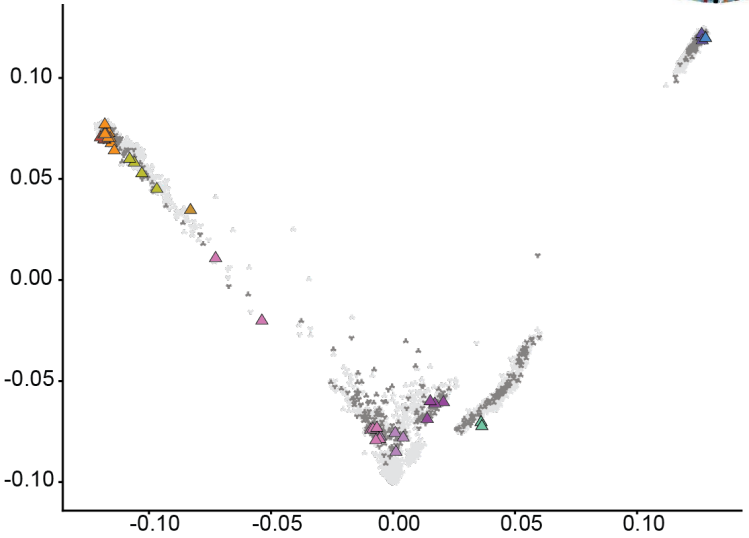
1000 Genomes Consortium Recruitment



African Caribbean in Barbados (ACB)
Esan in Nigeria (ESN)
African Ancestry in Southwest US (ASW)
Gambian in Western Division (GWD)
Yoruba in Ibadan, Nigeria (YRI)
Mende in Sierra Leone (MSL)
Colombian in Medellin, Colombia (CLM)
Peruvian in Lima, Peru (PEL)
Puerto Rican in Puerto Rico (PUR)
Han Chinese South (CHS)
Kinh in Ho Chi Minh City, Vietnam (KHV)
Punjabi in Lahore, Pakistan (PJL)

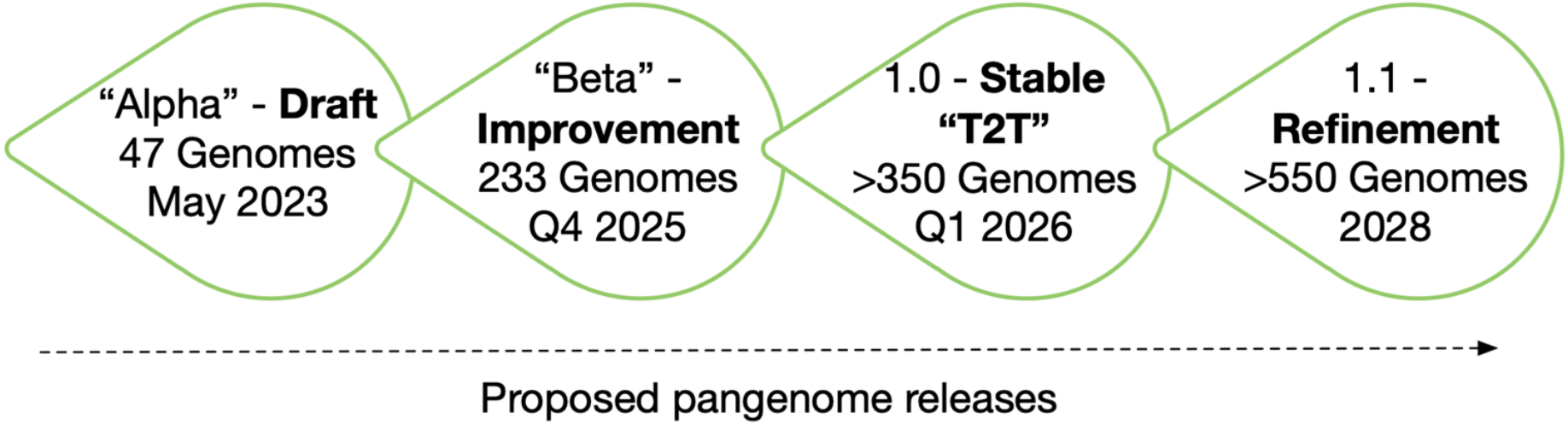


Initial Sampling Efforts:



- ☑ Cover genetic and geographic diversity
- ☑ Availability of low passage cell lines
- ☑ Availability of trios/parental data.

Soon: Beta release



New "beta" human pangenome release coming summer 2025!

Sequencing/assemblies available now from: humanpangenome.org

De novo assembly quantum leaps

- New sequencing technologies are leading to a dramatic improvement in contiguous assembly
- Haplotype resolution is now essential
- Simultaneously, computational efficiency of *de novo* assembly is being dramatically improved
- T2T will shortly be the standard

> Nat Biotechnol. 2023 Oct;41(10):1474-1482. doi: 10.1038/s41587-023-01662-6.
Epub 2023 Feb 16.

Telomere-to-telomere assembly of diploid chromosomes with Verkko

Mikko Rautiainen [1], Sergey Nurk [1 2], Brian P Walenz [1], Glennis A Logsdon [3], David Porubsky [3], Arang Rhie [1], Evan E Eichler [3 4], Adam M Phillippy [5], Sergey Koren [6]

HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads

Sergey Nurk[1], Brian P Walenz[1], Arang Rhie[1], Mitchell R Vollger[2], Glennis A Logsdon[2], Robert Grothe[3], Karen H Miga[4], Evan E Eichler[5], Adam M Phillippy[1] and Sergey Koren[1,6]

Haplotype-resolved *de novo* assembly with phased assembly graphs

Haoyu Cheng[1,2], Gregory T Concepcion[3], Xiaowen Feng[1,2], Haowen Zhang[4], and Heng Li[1,2,*]

[1]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[3]Pacific Biosciences, Menlo Park, CA, USA
[4]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
[*]To whom correspondence should be addressed: hli@jimmy.harvard.edu

A fully phased accurate assembly of an individual human genome

David Porubsky, Peter Ebert, Peter A. Audano, Mitchell R. Vollger, William T. Harvey, Katherine M. Munson, Melanie Sorensen, Arvis Sulovari, Marina Haukness, Maryam Ghareghani,
Human Genome Structural Variation Consortium, Peter M. Lansdorp, Benedict Paten, Scott E. Devine, Ashley D. Sanders, Charles Lee, Mark J.P. Chaisson, Jan O. Korbel, Evan E. Eichler, Tobias Marschall

doi: https://doi.org/10.1101/855049

Article | Open Access | Published: 04 May 2020

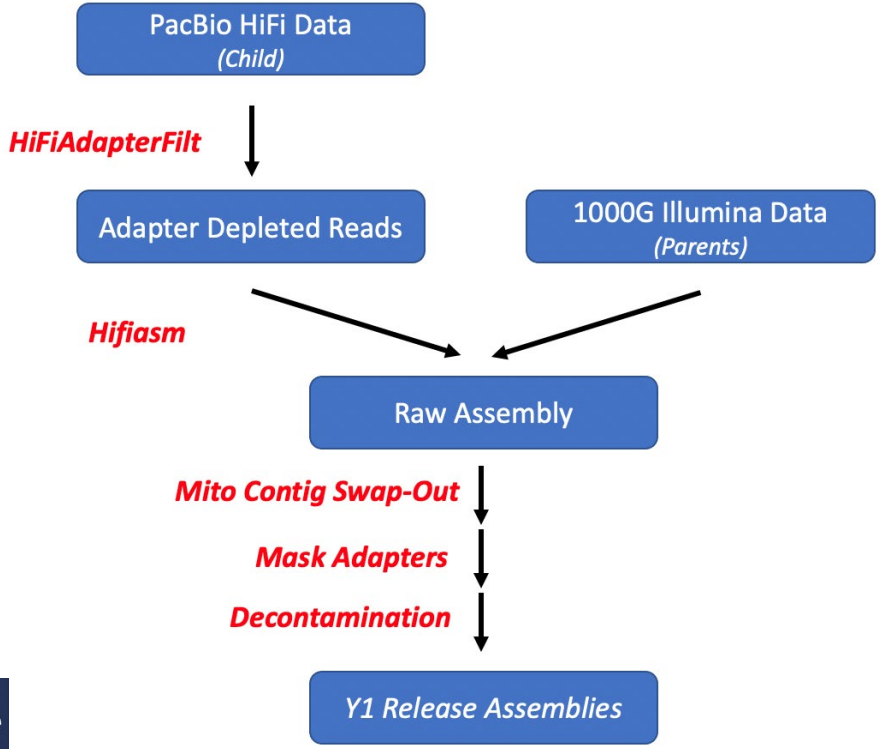
Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes

Kishwar Shafin, Trevor Pesout, [...] Benedict Paten

Nature Biotechnology **38**, 1044–1053(2020) | Cite this article

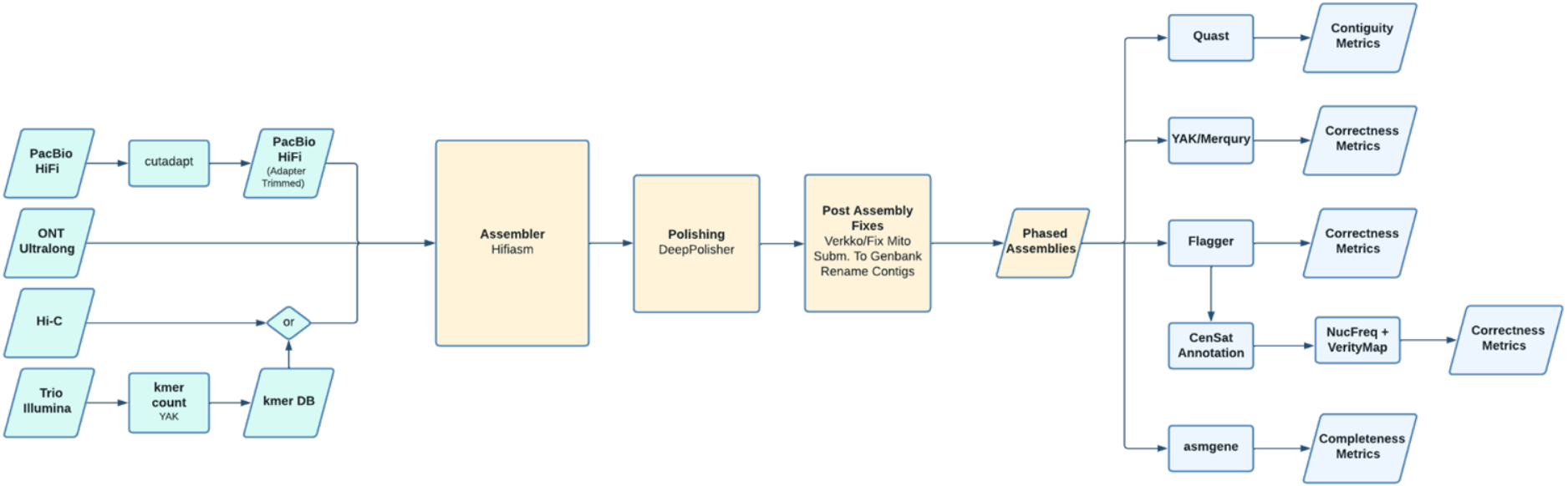
Draft Pangenome Diploid Assembly Pipeline

- Held bake off* to establish best of breed pipeline, using HG002 sample
- Picked trio-Hifiasm for contig assembly
- Used the AnVIL to assemble all samples based on reproducible, published pipelines
  - https://dockstore.org/organizations/HumanPangenome



* Jarvis E., et al. Automated assembly of high-quality diploid human reference genomes, Nature, Nov. 2022

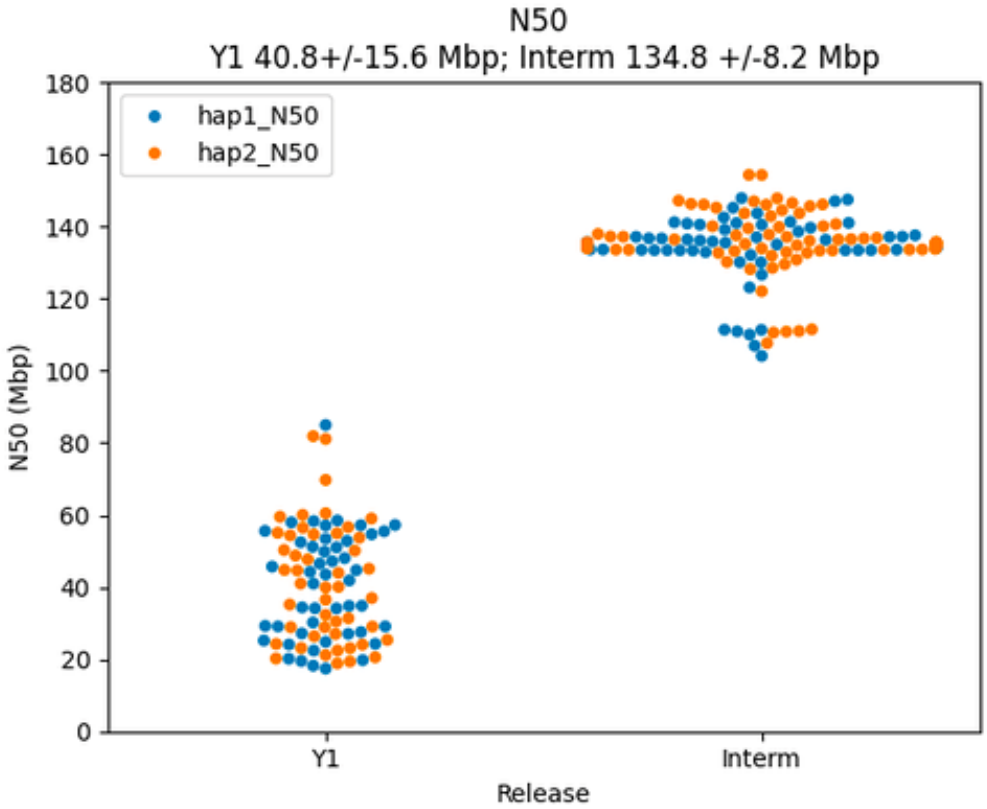
V2 Pangenome - Assembly Pipeline & QC



- Both Hifiasm and Verkko now integrate both HiFi/Duplex + ONT UL reads
- Hi-C or Trio Illumina used for long-range phasing
- Lots of QC!

Julian Lucas
& the HPRC Assembly Working Group

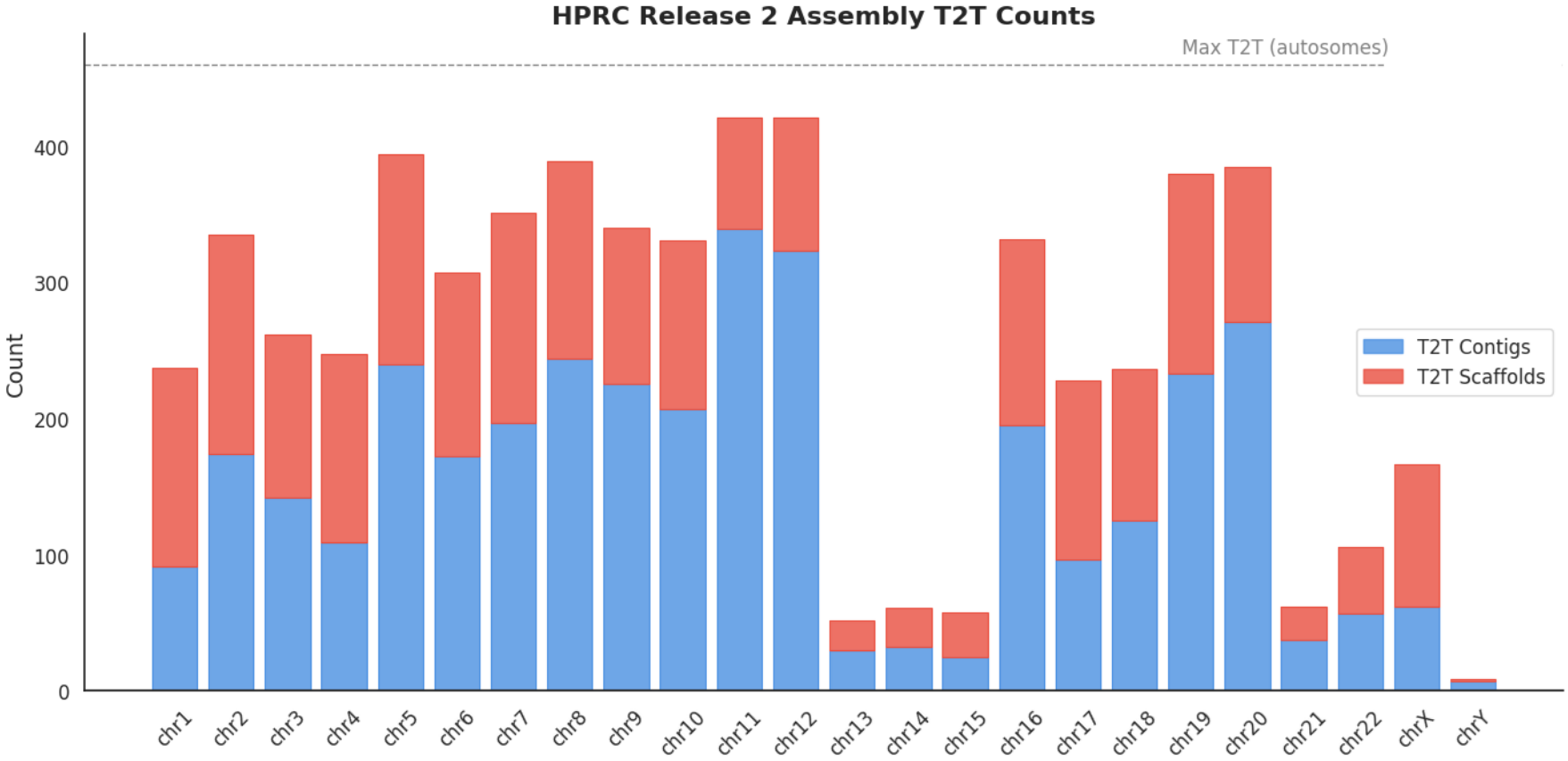
Beta (R2) Pangenome - AUN



(First 168 haplotypes / 340, pre-polishing)

Contiguity & Completeness

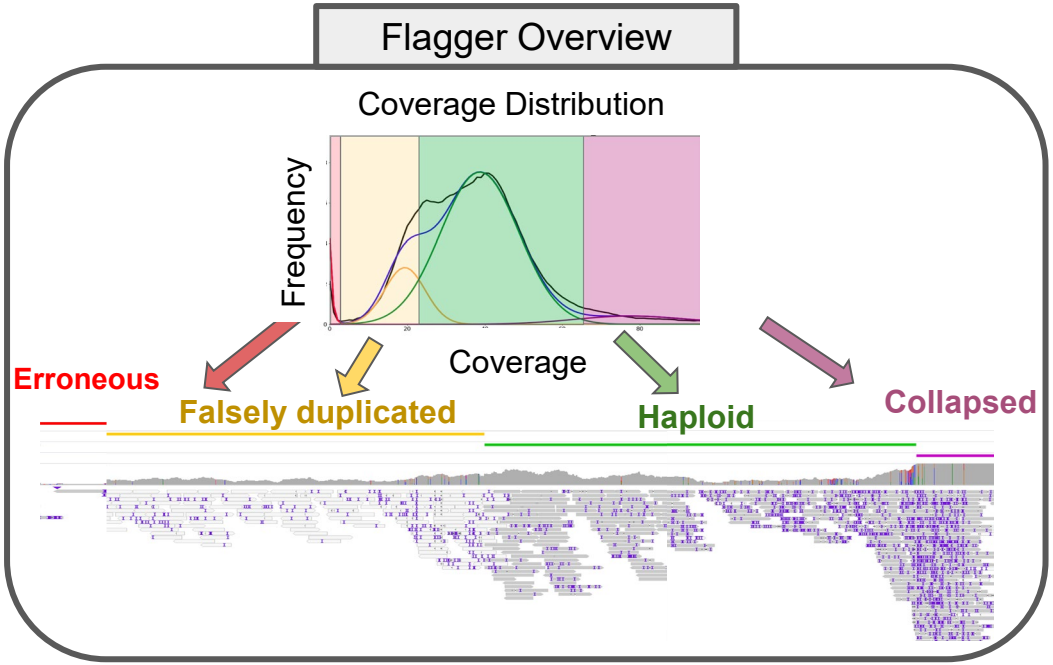
By Chromosome



Assembly QC: Flagger :
A read-mapping-based pipeline for assessing diploid assemblies

- Flagger takes **long reads (ONT or HiFI)** mapped to the diploid assembly in a haplotype-aware manner and finds read depth of coverages along the assembly.

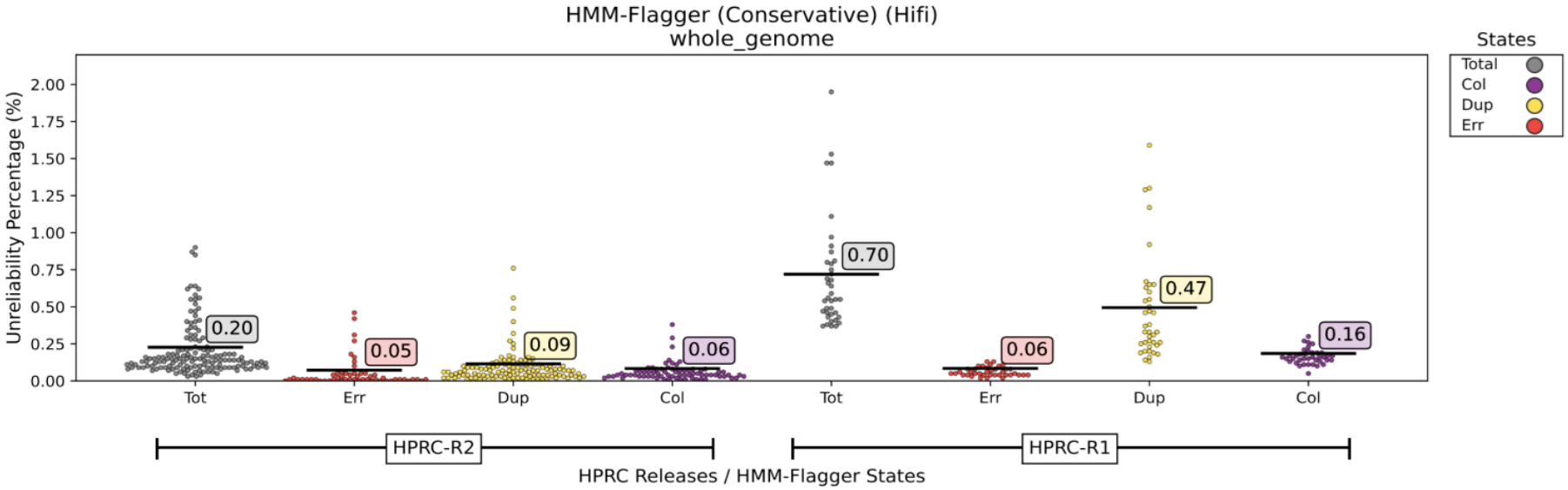
- It then uses a **Gaussian Mixture Model** to infer the coverage boundaries for
  - Well-assembled blocks (**Haploid**)
  - and 3 kinds of unreliable blocks which can be either
    - **Erroneous,**
    - **Falsely duplicated**
    - **Collapsed**



Mobin Asri

Correctness

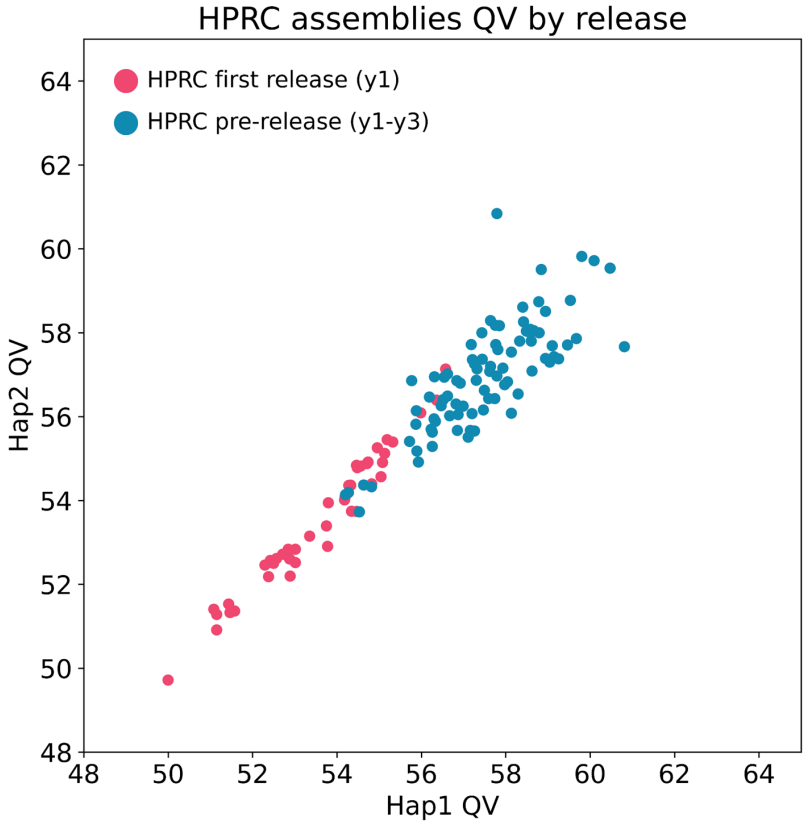
Release 2 Has Fewer Problem Regions Than Release 1



Black bars : Average across samples

Figure from Mobin

Base-level errors are found in even the highest quality assemblies



HPRC first release:
Average QV: **53.57**
(1 error per every 227,509 bases)

HPRC pre-release:
Average QV: **57.23**
(1 error per every 528,445 bases)

Google Health

Kishwar Shafin

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

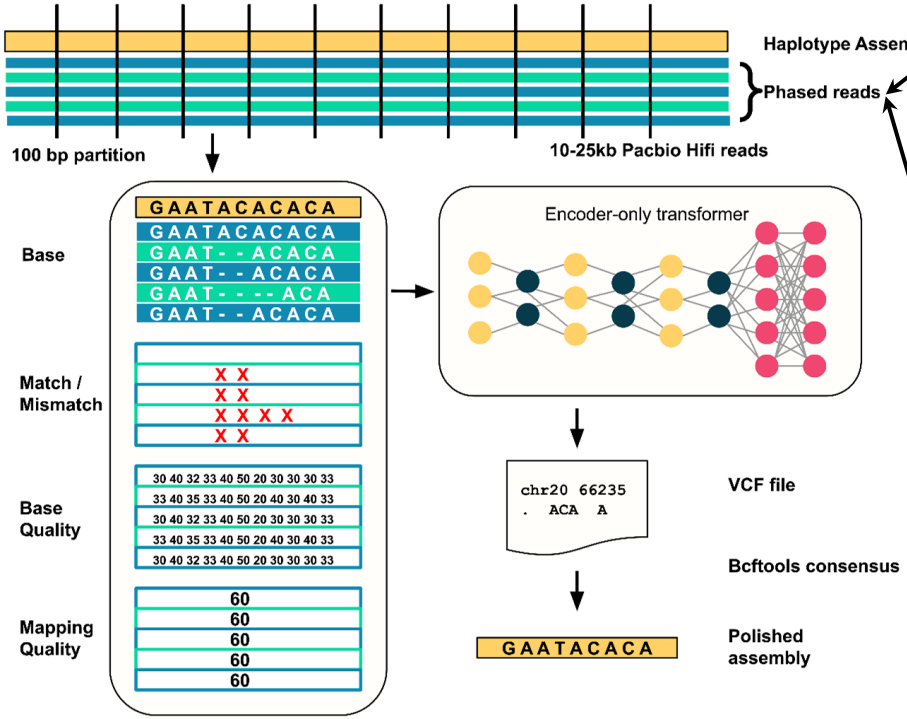
Mira Mastoras

Mobin Asri

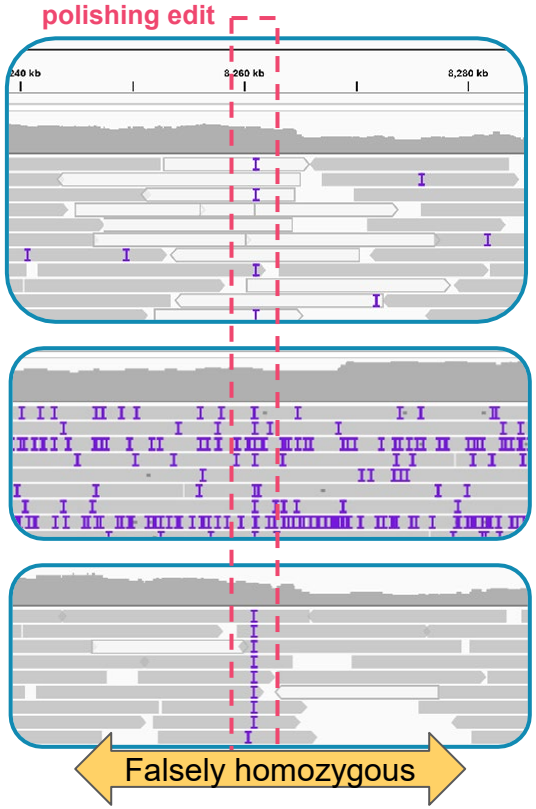
*QV calculated by **Yak** using Illumina kmers of size **31**

New methods are required to polish these residual assembly errors

DeepPolisher: an encoder-only transformer for sequence prediction



PHARAOH: leveraging ONT UL data to phase HiFi alignments

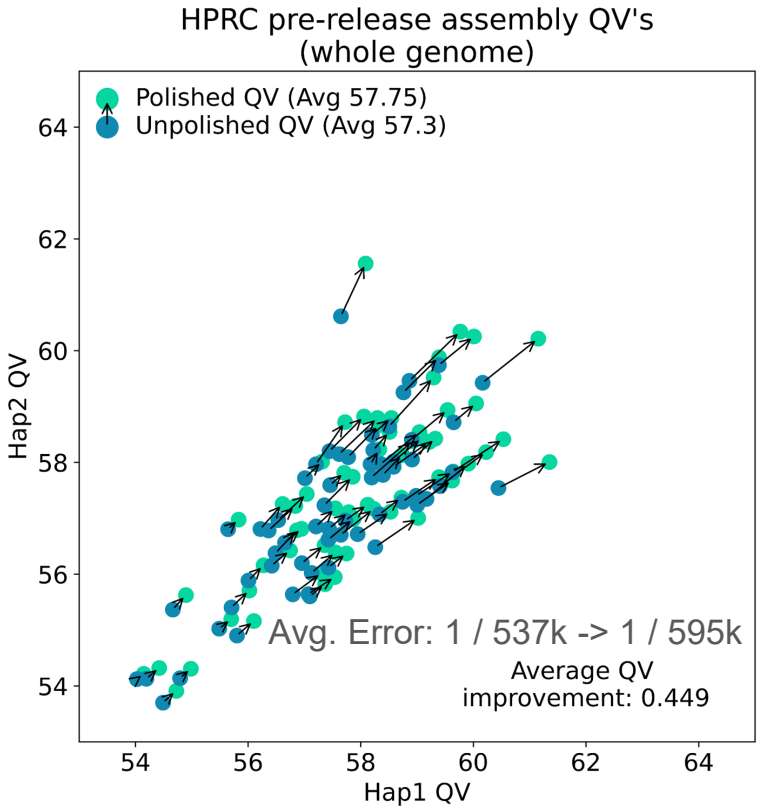


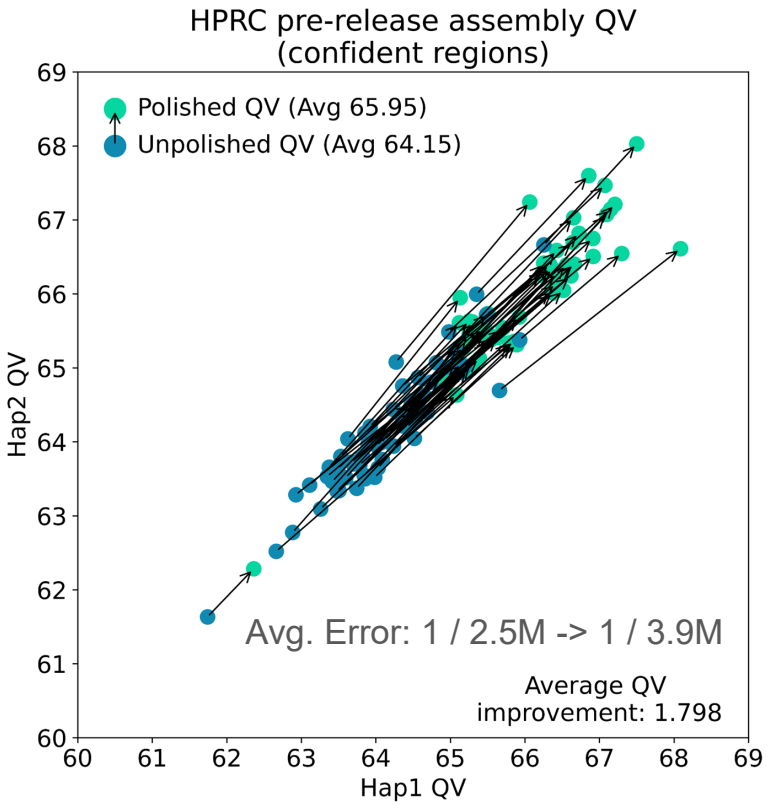
Without PHARAOH, errors inside false homozygous blocks > 30kb are randomly assigned a haplotype

ONT UL can span across homozygous blocks and phase these edits

After PHARAOH, edit is correctly phased

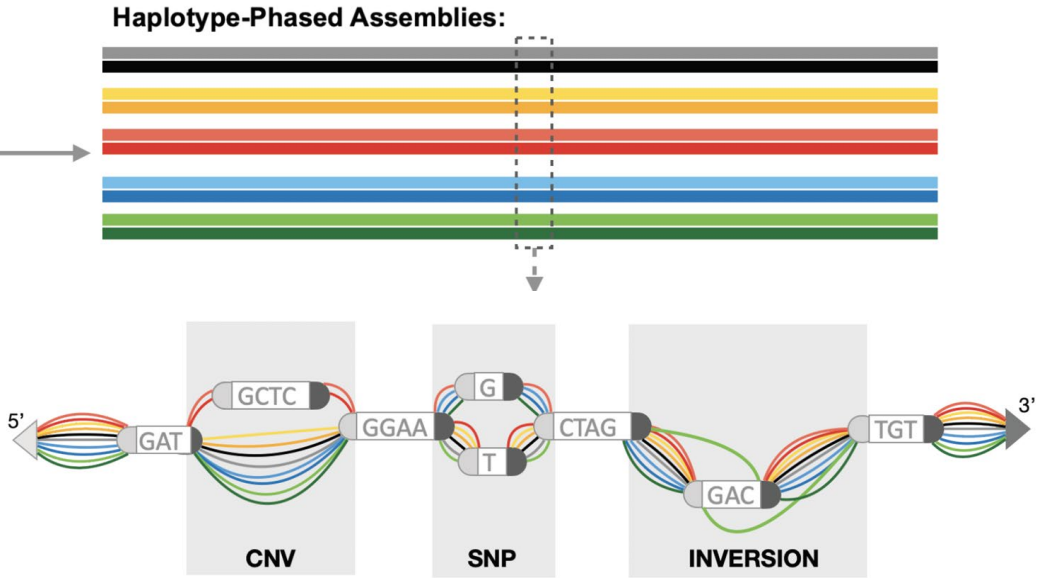
DeepPolisher produces substantial QV improvement for the next release of HPRC assemblies





*QV calculated by **Yak** using Illumina kmers of size **31**

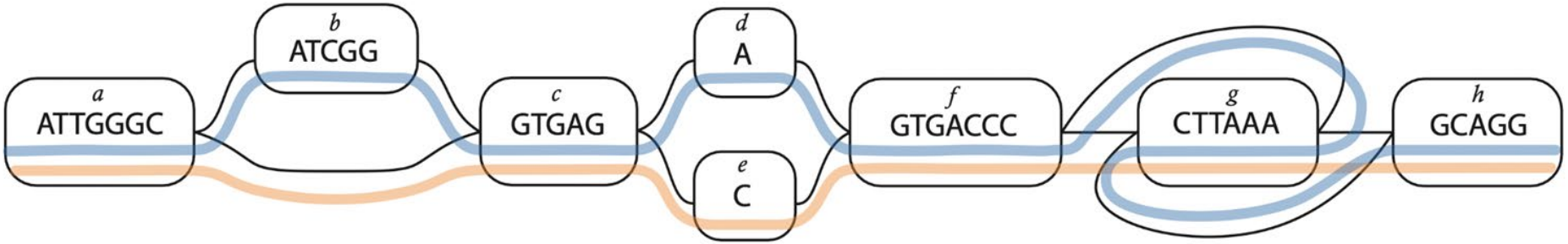
Human Pangenome Reference: Assemblies + Alignment + Annotation



Genome Alignments Using Genome Graphs

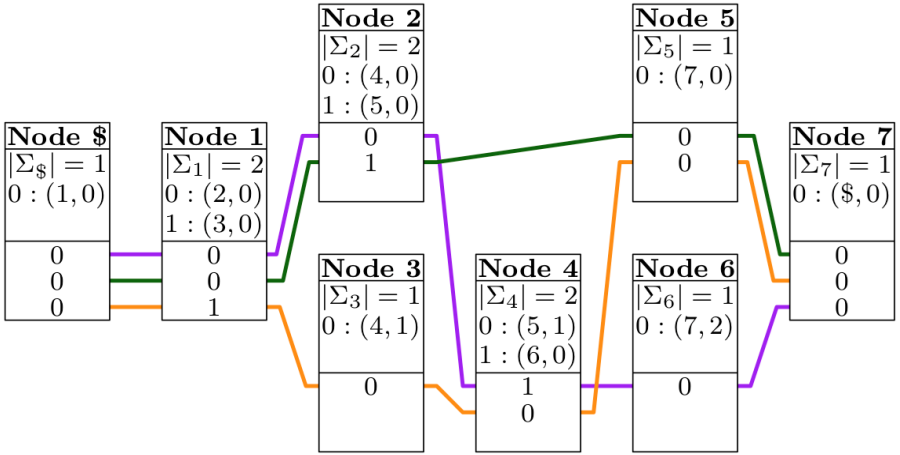
ATTGGGC**ATCGG**GTGAG**A**GTGACCC**TTTAAG**GCAGG

ATTGGGC - - - - - GTGAG**C**GTGACCC**CTTAAA**GCAGG



GBZ File Format

- Graph structure compresses common sequence:
- But *paths* (need 1 per contig per haplotype) are still expensive to store in GFA
- We propose GBZ, a binary graph format that compresses common *subpaths,* builds on GBW*T* data structure
- For 90 haplotypes:
  - `fasta  ~270G`
  - `fasta.gz  ~70G`
  - `gfa  45G`
  - `gfa.gz  11G`
  - `gbz  3G`



(gfa/gbz stats for minigraph/cactus graph)

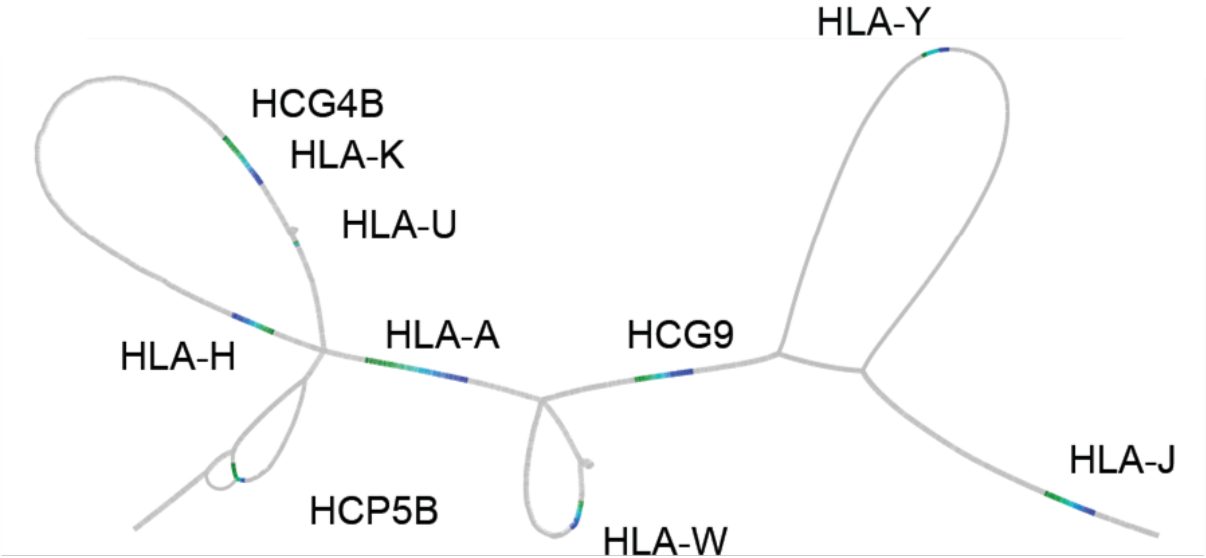
Work by Jouni Siren

Why bother with pangenome alignments? (two arguments)

Q: “Why do I need to use a pangenome alignment?”; aka, “Why can’t I just use a population specific reference?”

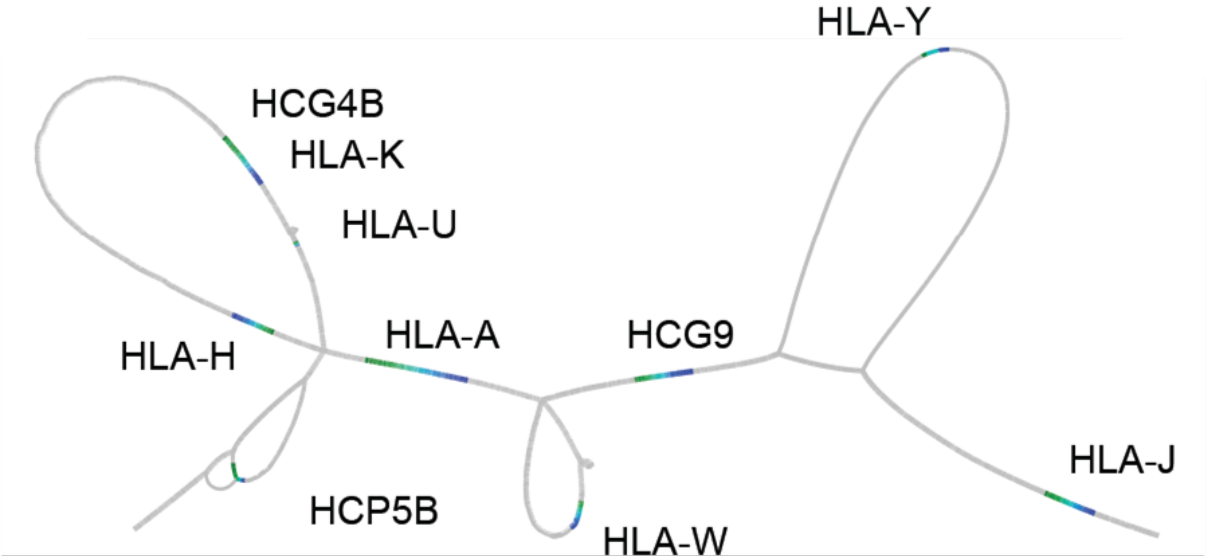
- Previous work (1000 Genomes, HapMap, etc.) show:
  - Common variants in most populations are mostly global (because they are old)
  - Common variant frequencies don't generally vary that much between most populations (see F_{ST} estimates)
  - Most variant alleles are common: 96-99% of alleles in a sample
- The upshot:
  - Any two randomly selected genomes from any significant population differ by millions of variants, a population specific reference will therefore generally have only a small effect
  - In contrast, a pangenome of just a few hundred diverse individuals will represent the large majority of alleles in any human sample

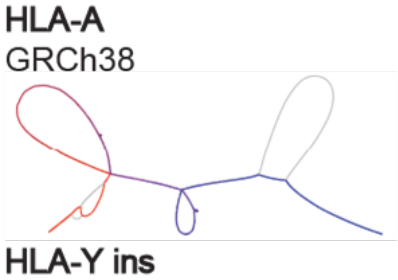
Why one ref doesn’t work - HLA-A



Credit: Shuangjia Lu

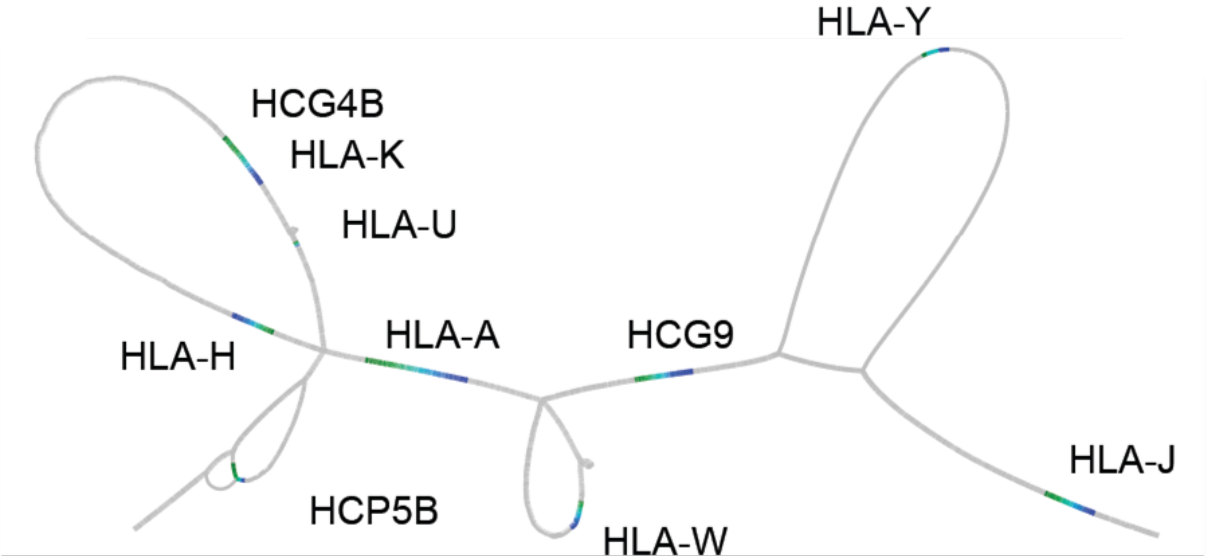
Why one ref doesn't work - HLA-A

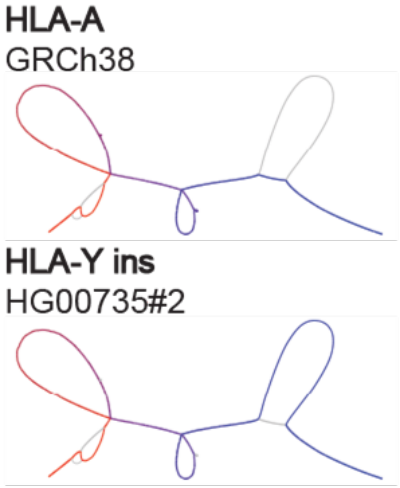




Credit: Shuangjia Lu

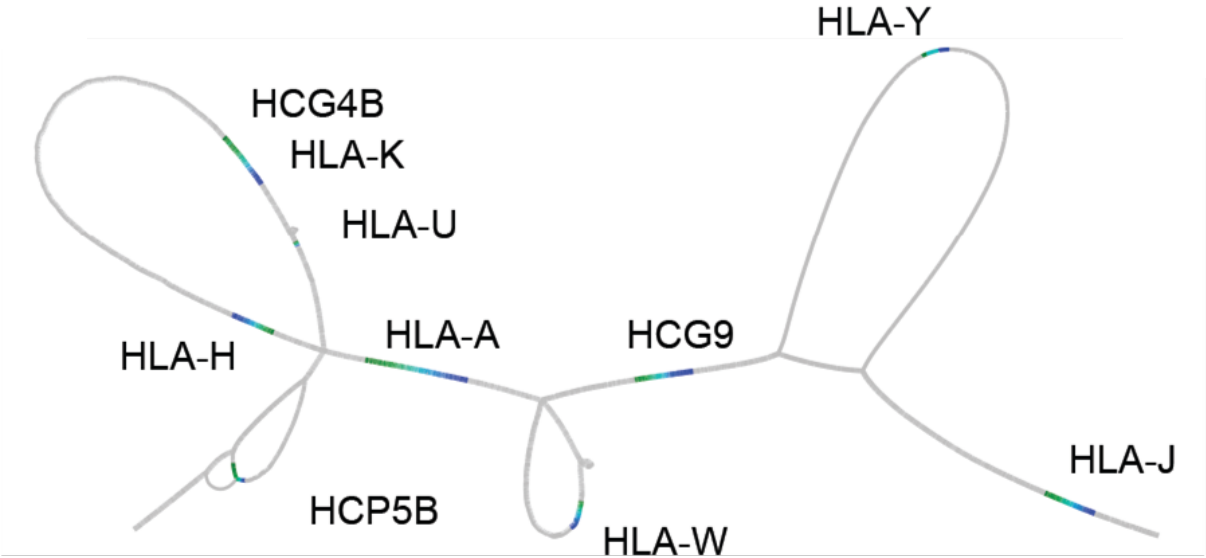
Why one ref doesn't work - HLA-A





Credit: Shuangjia Lu

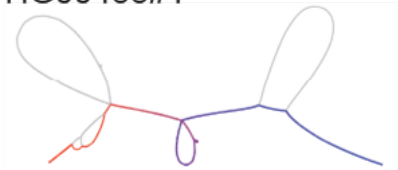
Why one ref doesn't work - HLA-A



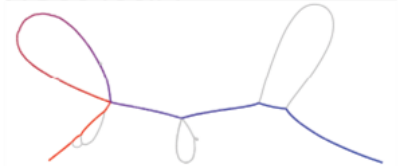
HLA-A
GRCh38

HLA-Y ins
HG00735#2

HLA-H/HCG4B/HLA-K/HLA-U del
HG00438#1

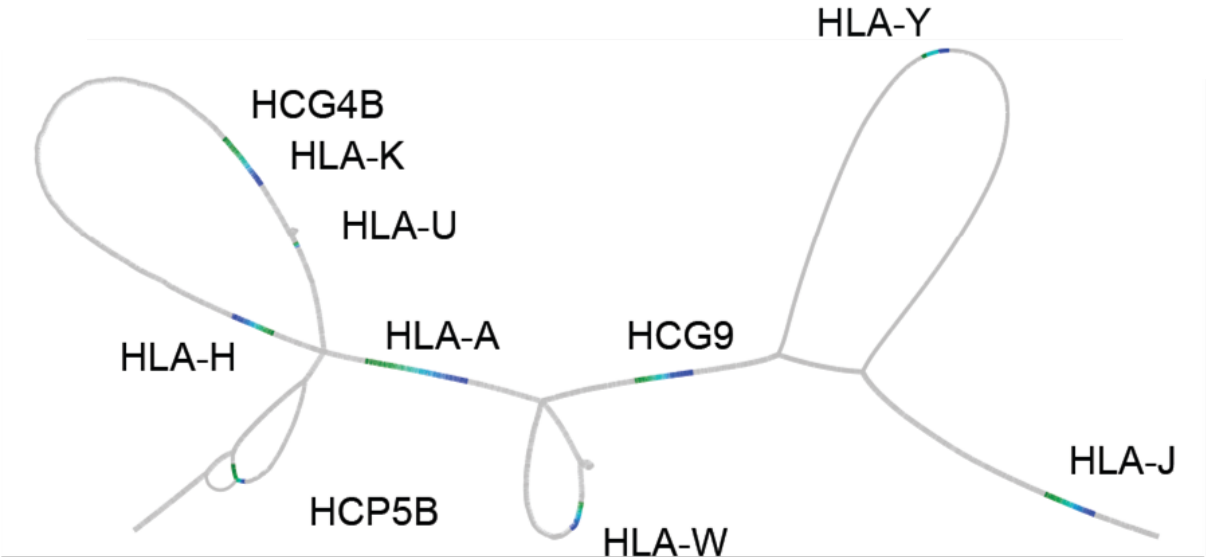


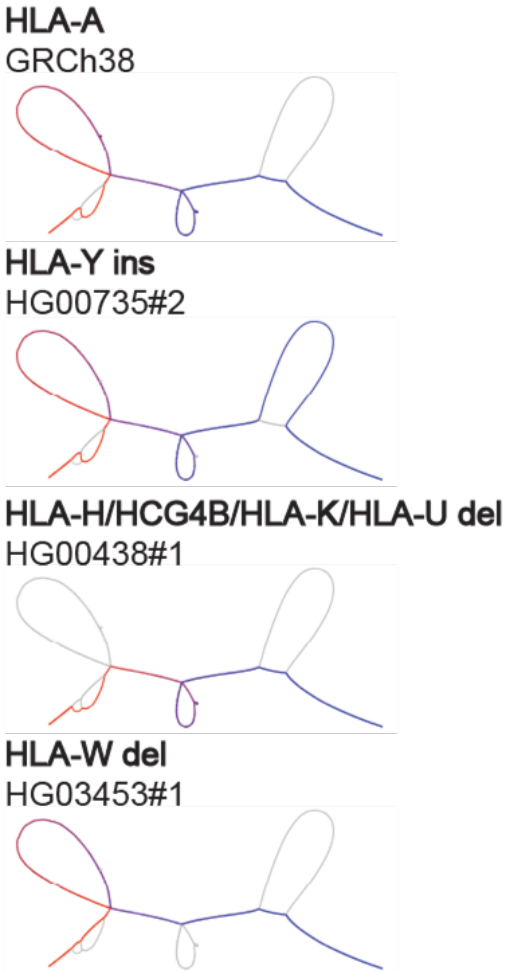
HLA-W del
HG03453#1



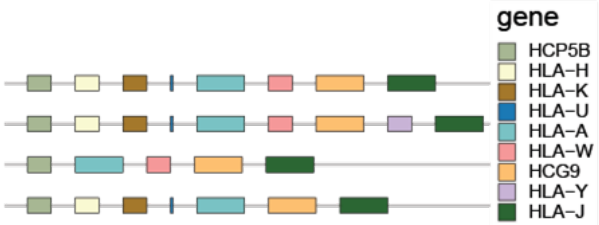
Credit: Shuangjia Lu

Why one ref doesn't work - HLA-A





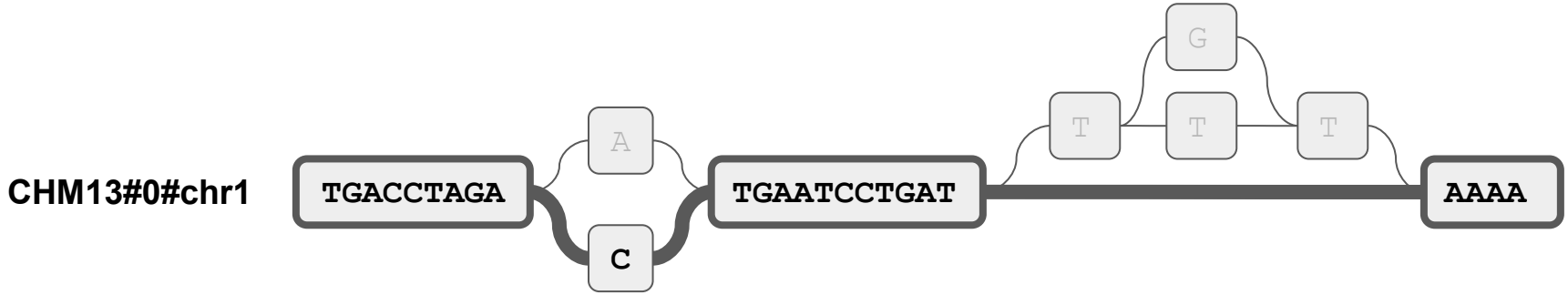
Count	Frequency	Haplotype name
57	0.63	HLA-A
25	0.28	HLA-Y ins
7	0.08	HLA-H/HCG4B/HLA-K/HLA-U del
1	0.01	HLA-W del



Credit: Shuangjia Lu

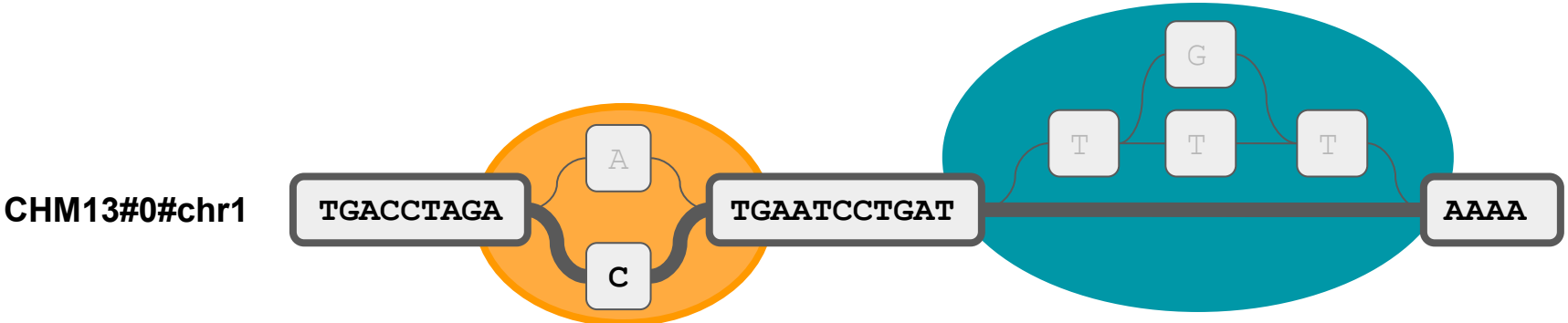
Coordinates and Bubbles

- Standard reference genomes included in HPRC graphs to serve as backbone / coordinate system
  - CHM13, GRCh38
  - Effective as genetic similarity results in largely “linear” graph structure



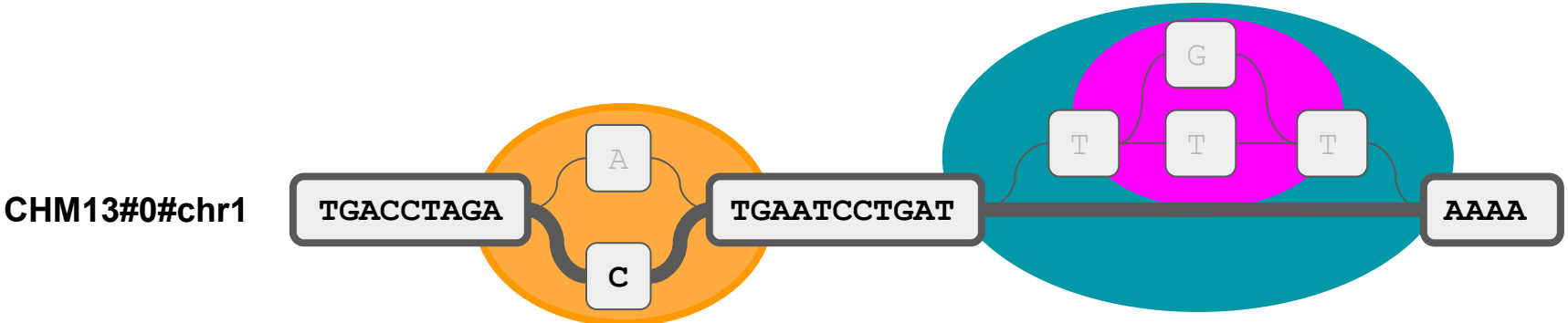
Coordinates and Bubbles

- Standard reference genomes included in HPRC graphs to serve as backbone / coordinate system
  - CHM13, GRCh38
  - Effective as genetic similarity results in largely “linear” graph structures
- Bubbles (aka snarls) are minimal sites of variation in graph



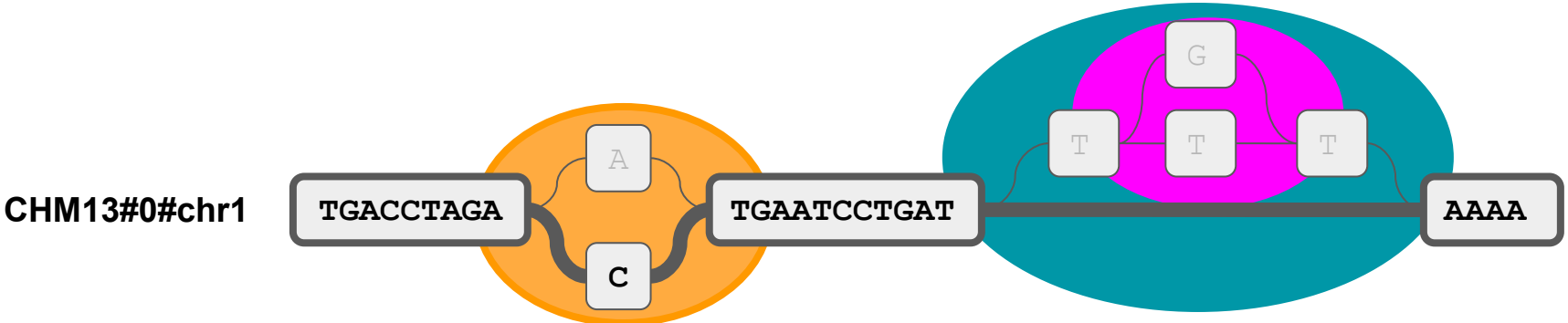
Coordinates and Bubbles

- Standard reference genomes included in HPRC graphs to serve as backbone / coordinate system
  - CHM13, GRCh38
  - Effective as genetic similarity results in largely “linear” graph structures
- Bubbles (aka snarls) are minimal sites of variation in graph
- Bubbles can be nested (but not otherwise overlap)

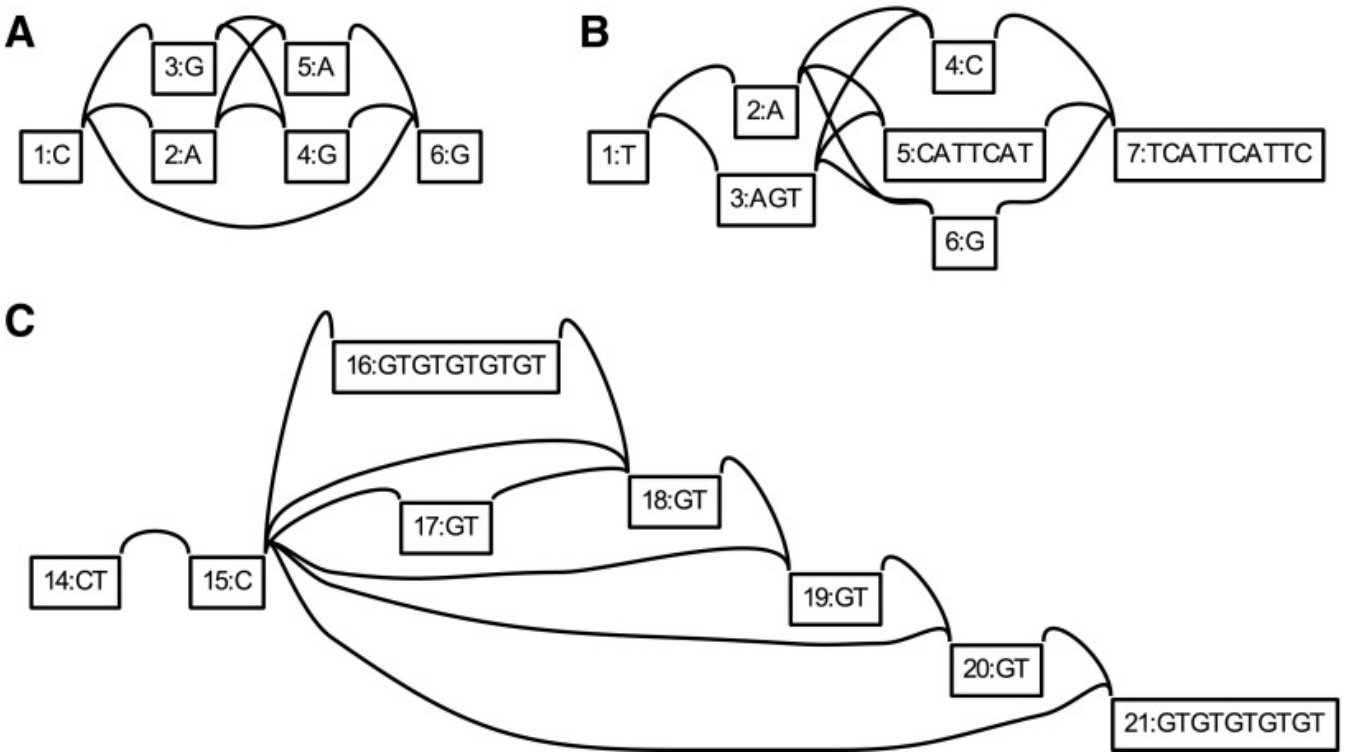


Coordinates and Bubbles

- Originally devised in context of *de novo* assembly (e.g. Zerbino, Birney, 2011)
- Linear time algorithms for bubble detection started for directed graphs (e.g. Onodera, et al., 2013)
- Generalized to bidirected graphs and related to cactus graph decomposition (Paten, et al. 2017)
- Several alternative (simpler) algorithms now proposed (e.g. Mwaniki et al., 2024, Li et al, 2024)



Coordinates and Bubbles



Real sites (even small ones) are often complex

Missing Heterochromatin

Heterochromatic sequence contains the fastest changing regions of our genomes:

- Satellites, centromeres, telomeres, acrocentric short arms

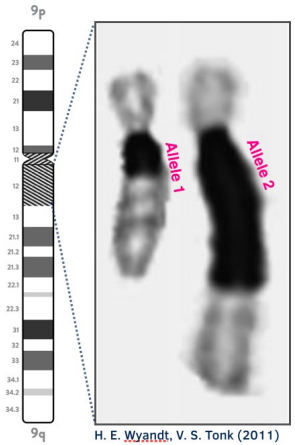
Our current pangenome alignments do not handle heterochromatin:

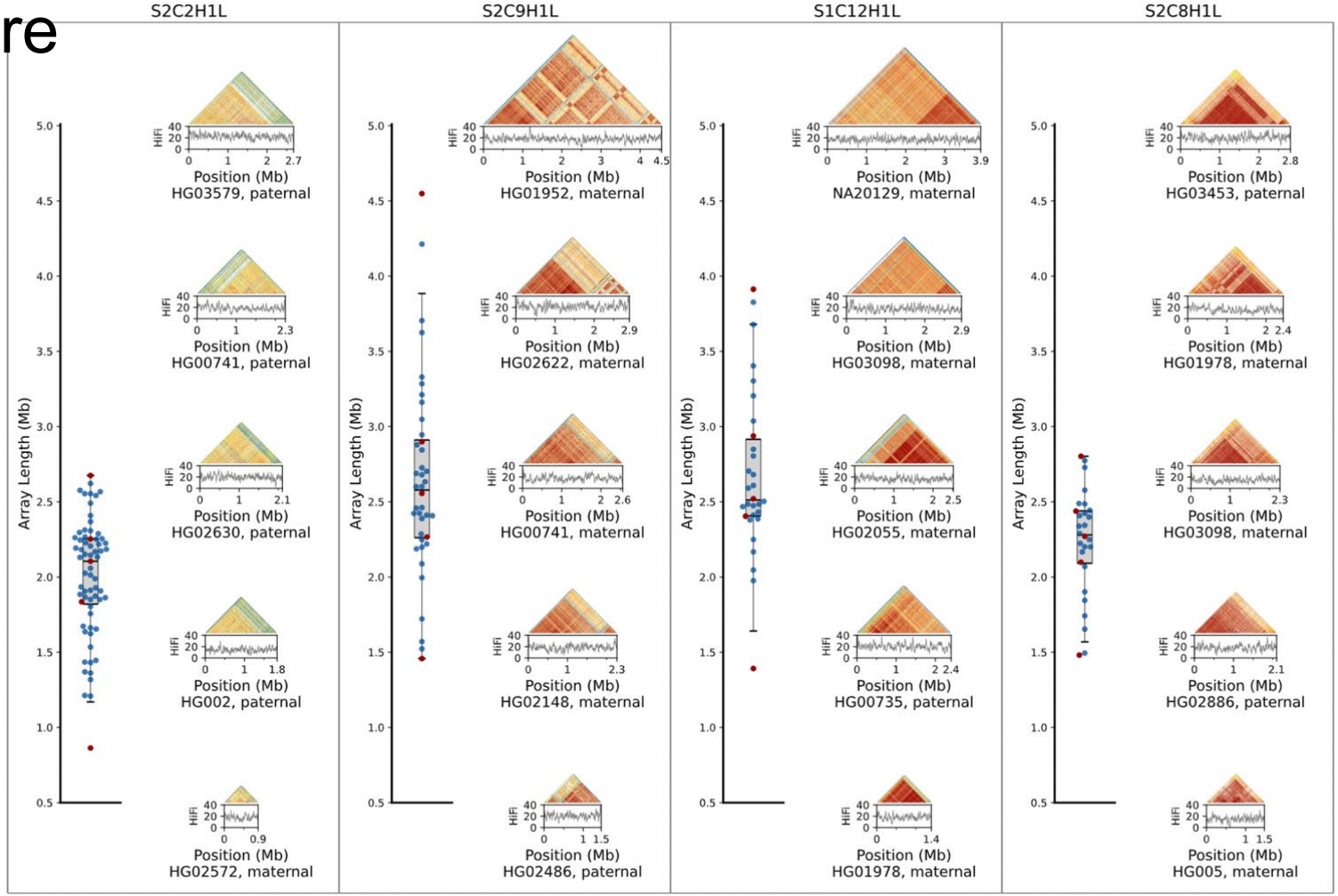
- Minigraph and Minigraph-Cactus clip out unalignable sequence
- PGGB encodes all the sequence, but can not align it reasonably

Graph	Nodes	Edges	Length (bp)
Minigraph	493,631	738,529	**3,365,688,482**
Minigraph-Cactus	85,591,995	118,409,526	**3,324,657,754**
PGGB	110,884,673	154,756,169	**8,415,267,572**

Challenge: We need to determine how and if this sequence can be aligned

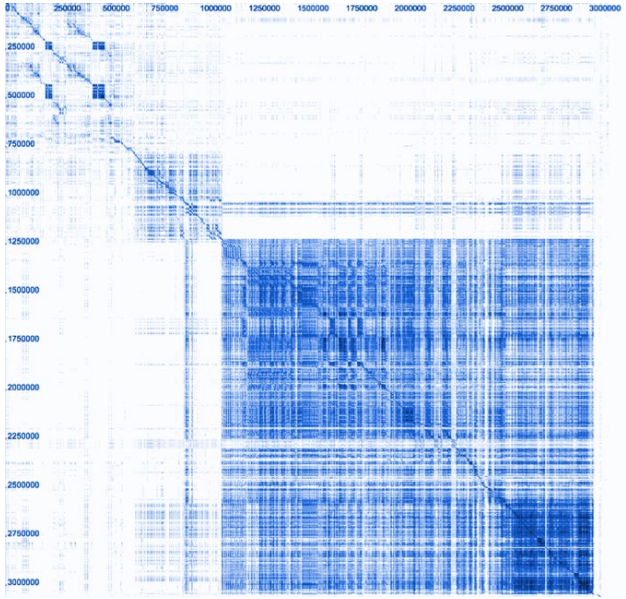
Centromere Variation





Credit: Mobin Asri, Karen Miga

Charting the last genomic wilderness: aligning centromeres*

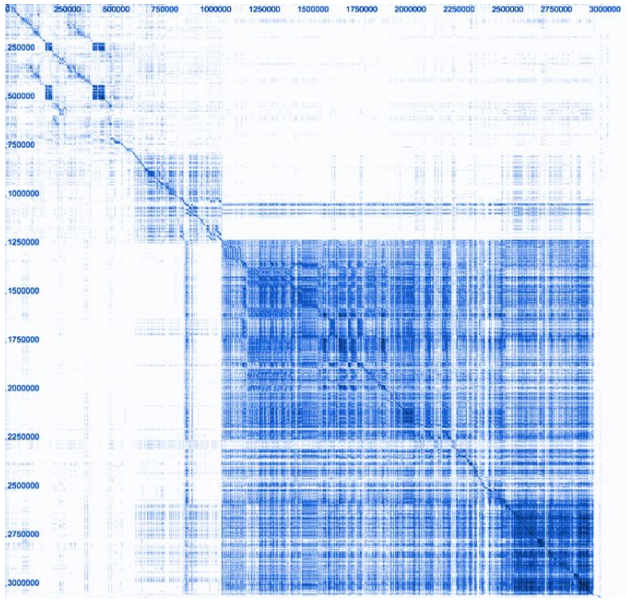


All matches

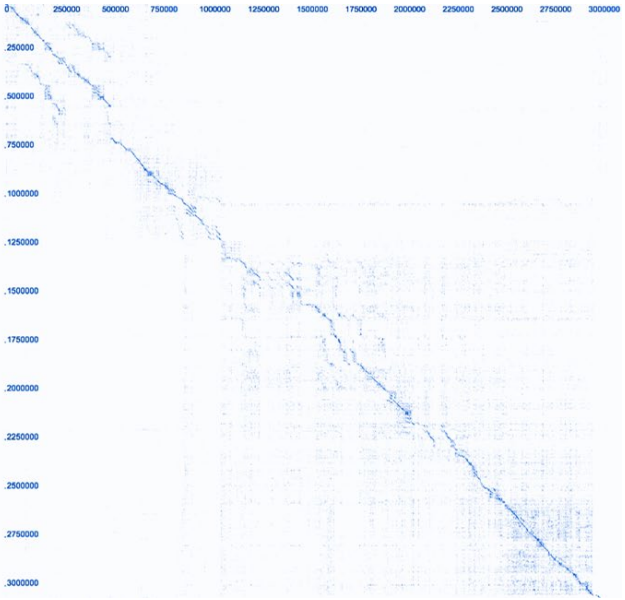
Example shows HG002 vs CHM13 chromosome X centromere HORs

* Prototype from Jordan Eizenga, inspired by UniAligner: a parameter-free framework for fast sequence alignment. Bzikadze AV, Pevzner PA. Nat Methods. 2023 Sep;20(9):1346-1354.

Charting the last genomic wilderness: aligning centromeres*



All matches

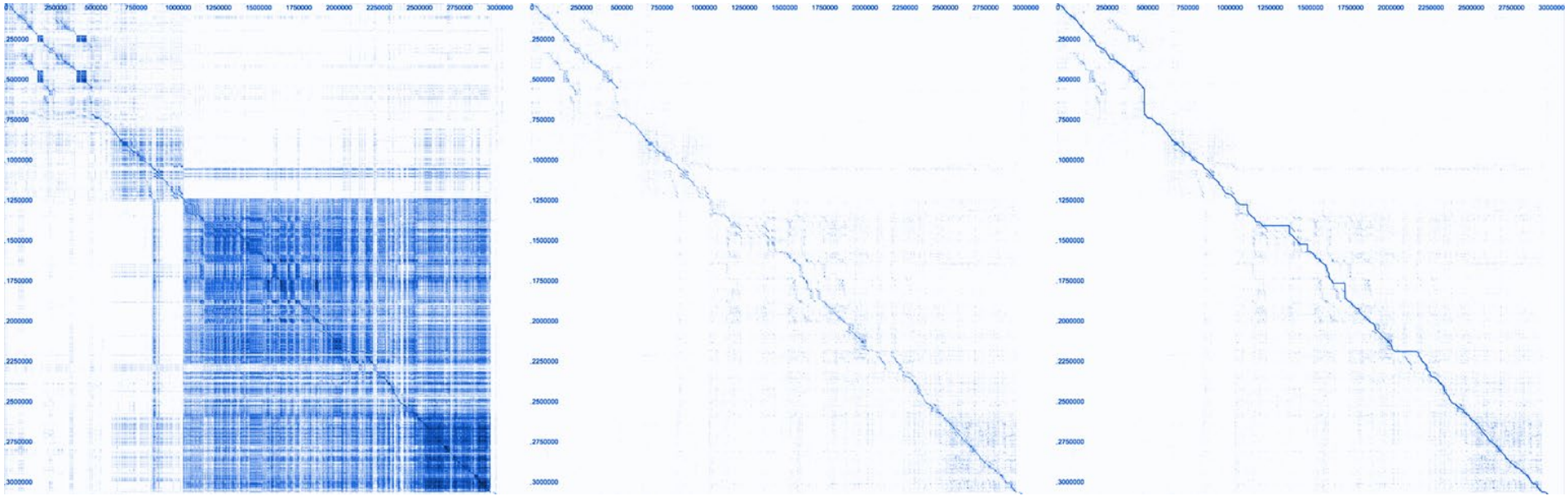


Colored by uniqueness

Example shows HG002 vs CHM13 chromosome X centromere HORs

* Prototype from Jordan Eizenga, inspired by UniAligner: a parameter-free framework for fast sequence alignment. Bzikadze AV, Pevzner PA. Nat Methods. 2023 Sep;20(9):1346-1354.

Charting the last genomic wilderness: aligning centromeres*



All matches | Colored by uniqueness | Unique Alignment

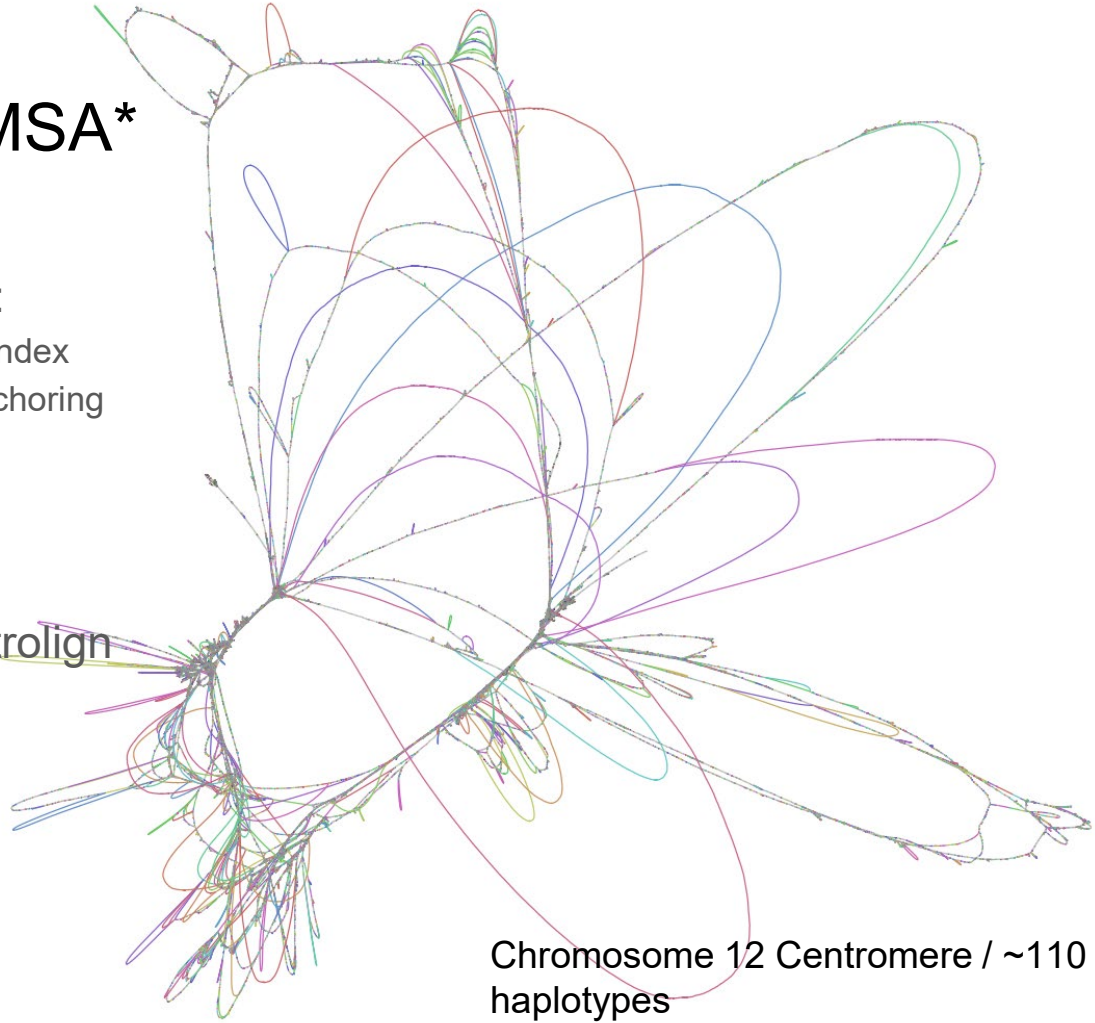
Example shows HG002 vs CHM13 chromosome X centromere HORs

* Prototype from Jordan Eizenga, inspired by UniAligner: a parameter-free framework for fast sequence alignment. Bzikadze AV, Pevzner PA. Nat Methods. 2023 Sep;20(9):1346-1354.

The First Centromere MSA*

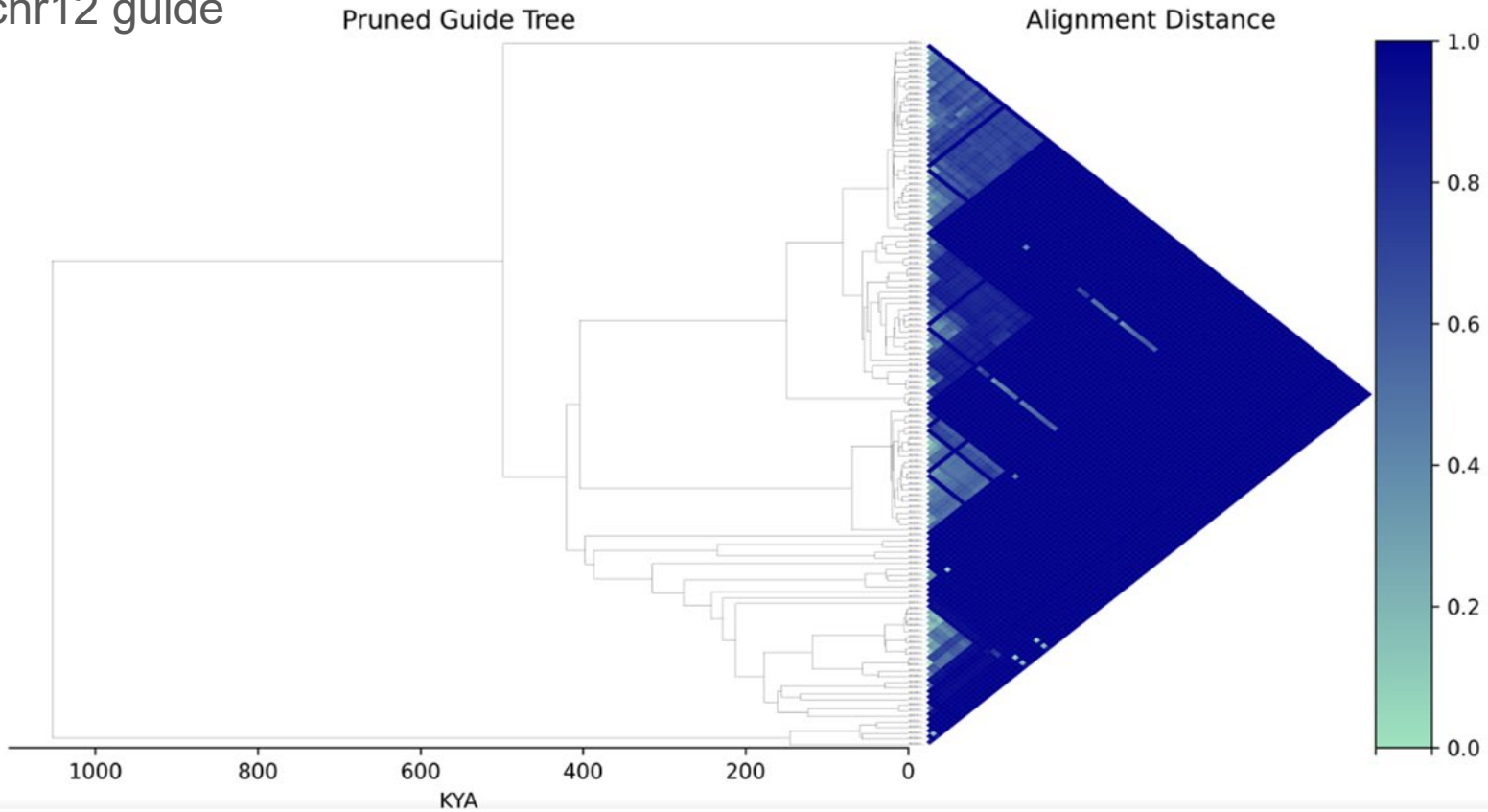
- Megabase-scale PO-POA
- Uniqueness Objective Function:
  - Match/count queries with a hybrid index
  - Sparse, affine-gap graph-graph anchoring
- Stitching between anchors with graph-graph WFA
- Centrolign https://github.com/jeizenga/centrolign

Dr. Jordan Eizenga



Chromosome 12 Centromere / ~110 haplotypes

Centrolign with current chr12 guide tree



Summary

A human pangenome is vital to ensuring all people are equally represented by the core reference structure that we all, as a community, use

The new “beta” pangenome assembly release is now availble, with >1.398 trillion bases of haplotype resolved, assembled sequence across 466 haploid genomes

While adoption of the pangenome will take time, pangenome methods are evolving fast and demonstrate promising applications right now

Adoption by the clinical community will happen as we create applications - better genome inference will be the start

Acknowledgements

U41HG010972
R01HG010485
U24HG010262
U01HG010961
OT2OD033761
U24HG011853
OT3HL142481

- **Xian Chang**
- **Jordan Eizenga**
- **Jouni Sirén**
- **Parsa Eskander**
- **Mobin Asri**
- **Adam Novak**
- Andrew Carroll
- Pi-Chuan Chang
- The Computational Genomics Lab
- Karen Miga & Miga Lab
- HPRC

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

National Center for Biotechnology Information NCBI

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

Cantata Bio

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

NIST

NIH National Human Genome Research Institute

HPRC Acknowledgements

Looking for a postdoc?, please email me: bpaten@ucsc.edu

Haley J. Abel, Lucinda L Antonacci-Fulton, **Mobin Asri**, Gunjan Baid, Carl A. Baker, Anastasiya Belyaeva, Konstantinos Billis, Guillaume Bourque, Silvia Buonaiuto, Andrew Carroll, **Mark JP Chaisson**, Pi-Chuan Chang, Xian H. Chang, Haoyu Cheng, Justin Chu, Sarah Cody, Vincenza Colonna, Daniel E. Cook, Omar E. Cornejo, Mark Diekhans, Daniel Doerr, Peter Ebert, **Jana Ebler**, Evan E. Eichler, **Jordan M. Eizenga**, Susan Fairley, Olivier Fedrigo, Adam L. Felsenfeld, Xiaowen Feng, Christian Fischer, Paul Flicek, Giulio Formenti, Adam Frankish, Robert S. Fulton, Yan Gao, Shilpa Garg, **Erik Garrison**, Carlos Garcia Giron, Richard E. Green, Cristian Groza, Andrea Guarracino, Leanne Haggerty, **Ira Hall**, William T Harvey, Marina Haukness, **David Haussler**, Simon Heumos, **Glenn Hickey**, Kendra Hoekzema, Thibaut Hourlier, Kerstin Howe, Miten Jain, Erich D. Jarvis, Hanlee P. Ji, Alexey Kolesnikov, Jan O. Korbel, Jennifer Kordosky, HoJoon Lee, Alexandra P. Lewis, **Heng Li**, **Wen-Wei Liao**, Shuangjia Lu, Tsung-Yu Lu, Julian K. Lucas, Hugo Magalhães, Santiago Marco-Sola, Pierre Marijon, Charles Markello, **Tobias Marschall**, Fergal J. Martin, Jennifer McDaniel, **Karen H. Miga**, Matthew W. Mitchell, **Jean Monlong**, Jacquelyn Mountcastle, Katherine M. Munson, Moses Njagi Mwaniki, Maria Nattestad, Adam M. Novak, Hugh E. Olsen, Nathan D. Olson, **Trevor Pesout**, Adam M. Phillippy, Alice B. Popejoy, David Porubsky, Pjotr Prins, Daniela Puiu, Allison A Regier, Arang Rhie, Samuel Sacco, Ashley D. Sanders, Valerie A. Schneider, Baergen I. Schultz, Kishwar Shafin, **Jonas A. Sibbesen**, **Jouni Sirén**, Michael W. Smith, Heidi J. Sofia, Ahmad N. Abou Tayoun, Françoise Thibaud-Nissen, Chad Tomlinson, Francesca Floriana Tricomi, Flavia Villani, Mitchell R. Vollger, Justin Wagner, Ting Wang, Jonathan M. D. Wood, Aleksey V. Zimin, Justin M. Zook

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

NIH National Human Genome Research Institute

Some new(ish) pangenome data structures

GBZ-base (or: Why not use a database?)

GBZ format is essentially:

- Set of records for (oriented) nodes:
  - List of outgoing edges.
  - BWT fragment for path visits.
  - Sequence.
- Index for finding records by node identifiers.

We could store this in a database:

```
CREATE TABLE Nodes (
        handle INTEGER PRIMARY KEY,
        edges BLOB NOT NULL,
        bwt BLOB NOT NULL,
        sequence BLOB NOT NULL)
```

HPRC v1.1 Minigraph–Cactus graph takes 3.06 GiB in GBZ format and 5.52 GiB as a SQLite database.

The database can be built in < 2 minutes on a laptop.

Extracting a 1000 bp context around a reference position typically takes < 10 milliseconds and a few megabytes of memory.

https://github.com/jltsiren/gbz-base

Work by Jouni Siren

GAF-base for Reads?

Store alignments using a simplified Nodes table for the paths and another table for the rest.

```
CREATE TABLE Alignments (
        handle INTEGER PRIMARY KEY,
        name TEXT NOT NULL,
        start_node INTEGER NOT NULL,
        numbers BLOB NOT NULL,
        quality BLOB,
        difference BLOB,
        pair BLOB)
```

Then we can select alignments by any node or subpath.

35x Novaseq 6000 reads (ERR3239454) mapped with Giraffe:

- GAM (vg internal format): 114 GiB
- GAF: 209 GiB
- Gzip-compressed GAF: 26 GiB
- Database with individually encoded alignments: 36 GiB
- Hypothetical binary format: 26 GiB
- BAM: 50 GiB
- CRAM: 10 GiB

Read paths compress as 3GB in GBWT

Further compression possible of quals, names, etc.

Work by Jouni Siren

Pangenome String Indexes

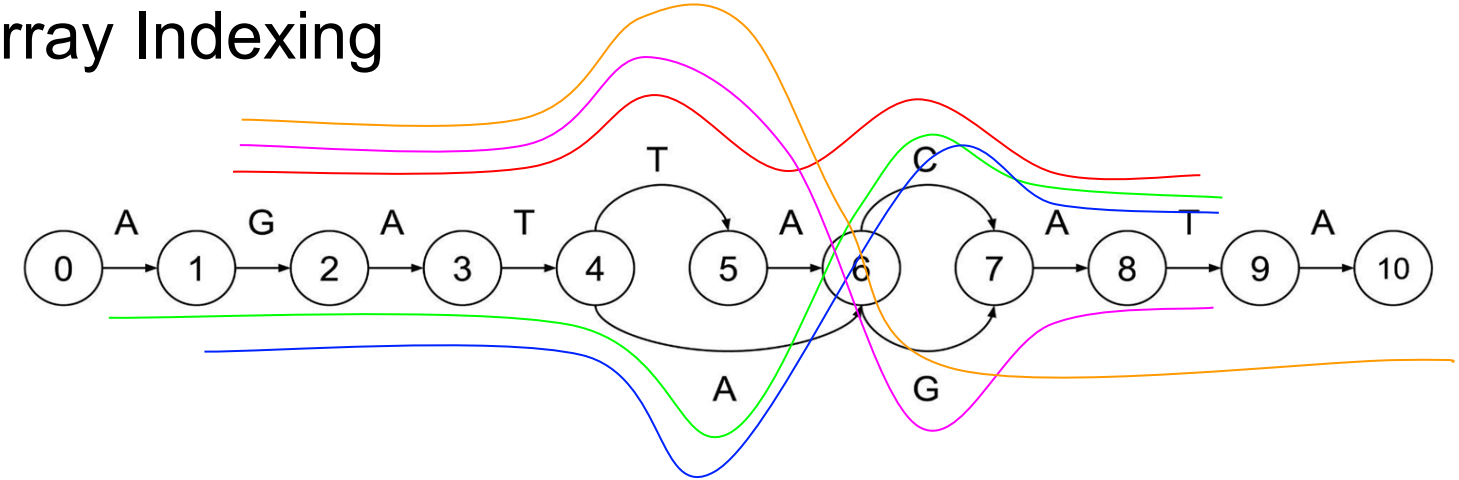
To map a substring to a pangenome:

- **k-mer / minimizer indexes** - space efficient, but fixed k makes it less useful for repetitive sequence
- **Sequence FM indexes** - locate instances on haplotypes, but then have a deduplication problem (same string may occur in many haplotypes)
- **Graph FM indexes** (e.g. GCSA) - work for De Bruijn graphs, otherwise finicky, forget underlying haplotype info

Q: For a haplotype to graph substring index, could we build an efficient map from a sequence FM index (of the haplotypes) to graph positions?

Enter the **Tag Array** - https://arxiv.org/abs/2411.15291 (Travis Gagie)

Tag Array Indexing



T = GATTACAT$AGATACAT$GATACAT$GATTAGAT$GATTAGATA$

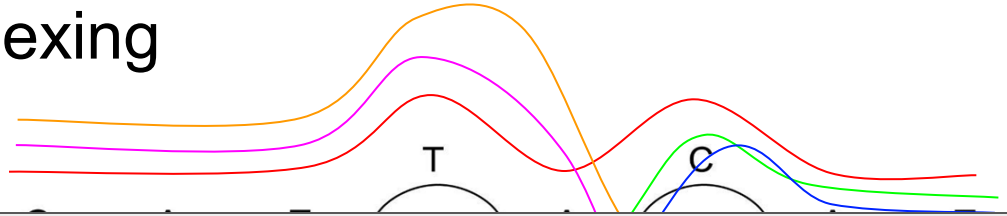
BWT = TTATTTTTTTT$CCCGGGGGGAAAAAA$$$$AAAAATAATTAAA

The Graph position of the corresponding BWT index

Tags = 99E99954455077777222226666611111888884334333

Slide courtesy of Parsa Eskandar

Tag Array Indexing



Travis's key insight: ".. a property has contextual locality if characters with similar contexts tend to have the same or similar values (``tags") of that property. … if we consider a repetitive text and such a property and the tags in their characters' BWT order, then the resulting string -- the text and property's **tag array** -- will be run-length compressible either directly or after some minor manipulation"

In this context: "the tag array of a genome graph is highly compressible"

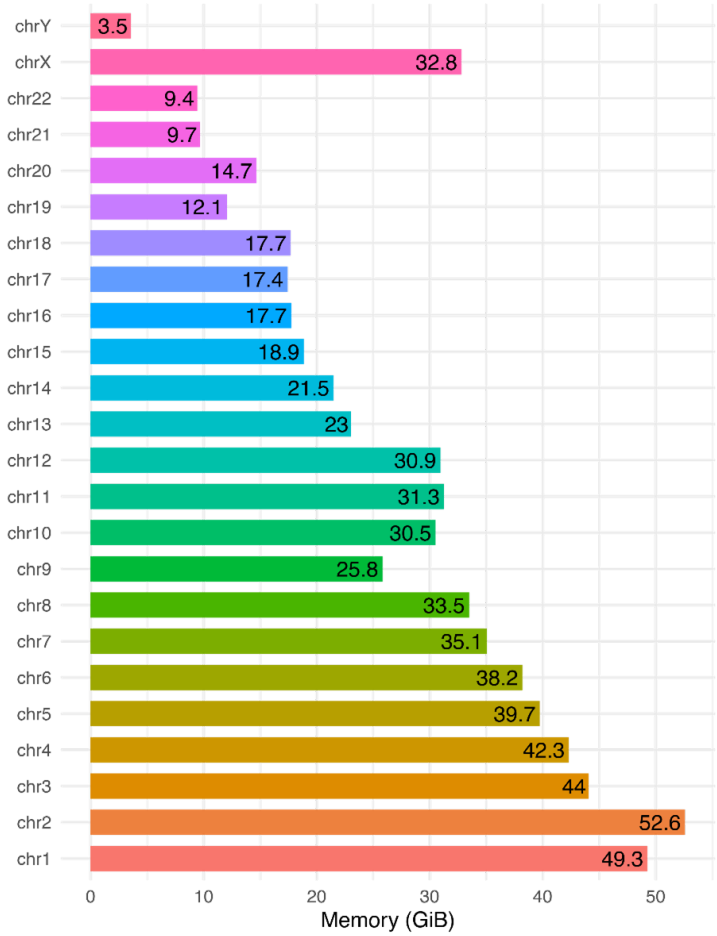
The Grap
the corresponding BWT
index

Tags = **99E99954455077777222226666611111888884334333**

Building the Tag Array Index

- Building Tag Arrays for each chromosome to reduce memory usage
- Merging per-chromosome arrays in later stages

Credit: Dr. Jouni Siren, Parsa Eskandar

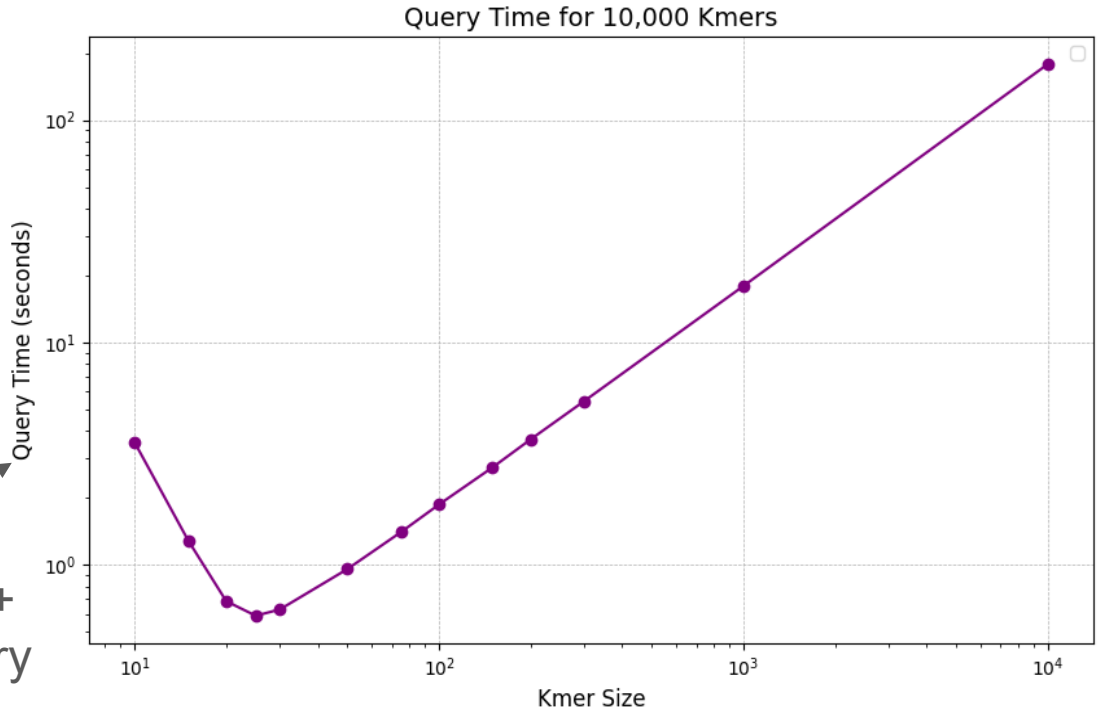


v1.1 HPRC

Query time

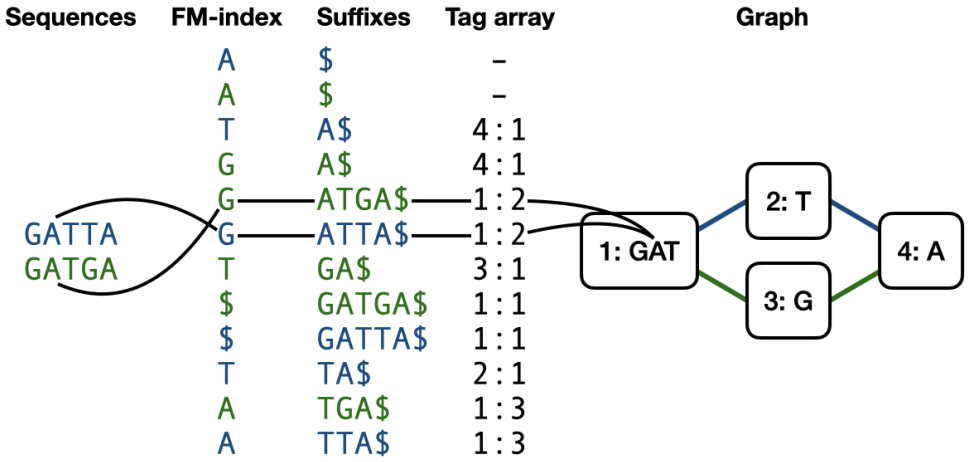
- Chr19, v1.1 HPRC
- 5.1 Gbases
- 1:01:58 using 16 threads
- 29.2GB construction memory
- 1.8 GB on disk

r-index query +
Tag array query



Credit: Dr. Jouni Siren, Parsa Eskandar

Tag Array Uses



Lossless Pangenome Indexing Using Tag Arrays, Parsa Eskandar, Benedict Paten, Jouni Sirén doi: https://doi.org/10.1101/2025.05.12.653561

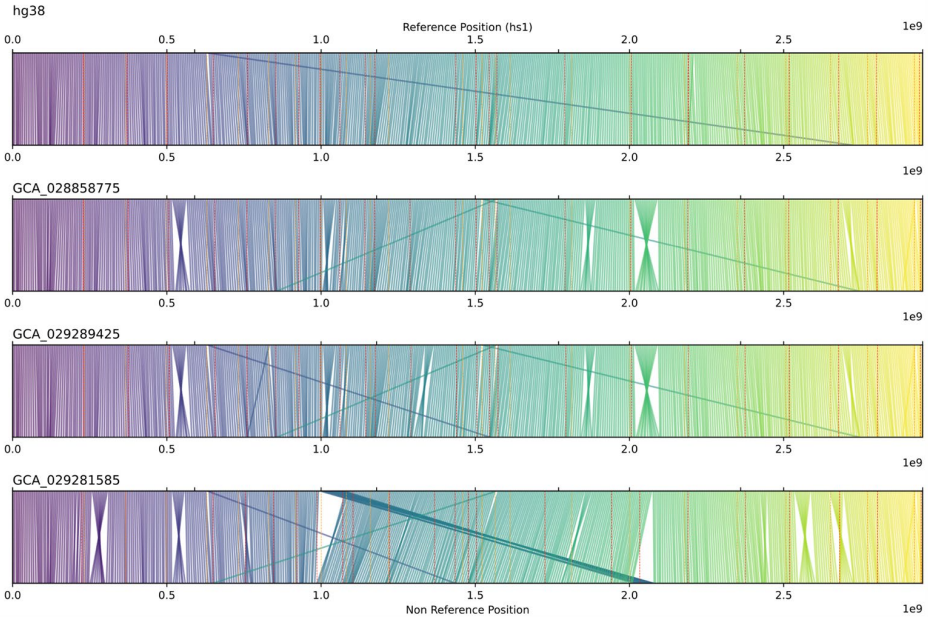
Article | Open access | Published: 09 April 2025

Complete sequencing of ape genomes

DongAhn Yoo, Arang Rhie, Prajna Hebbar, Francesca Antonacci, Glennis A. Logsdon, Steven J. Solar, Dmitry Antipov, Brandon D. Pickett, Yana Safonova, Francesco Montinaro, Yanting Luo, Joanna Malukiewicz, Jessica M. Storer, Jiadong Lin, Abigail N. Sequeira, Riley J. Mangan, Glenn Hickey, Graciela Monfort Anez, Parithi Balachandran, Anton Bankevich, Christine R. Beck, Arjun Biddanda, Matthew Borchers, Gerard G. Bouffard, ... Evan E. Eichler

+ Show authors

Nature **641**, 401–418 (2025) | Cite this article



Pangenome applications and algorithms

Pangenomes Power The Best Short-read Variant Calling Methods

- The best performing Illumina Dragen and Google DeepVariant methods are using pangenomes
- 5x reduction in errors vs. GATK achieved
- **This exemplifies the initial application of the pangenome: as a black box to improve key tasks**

Comprehensive and accurate genome analysis at scale using DRAGEN accelerated algorithms

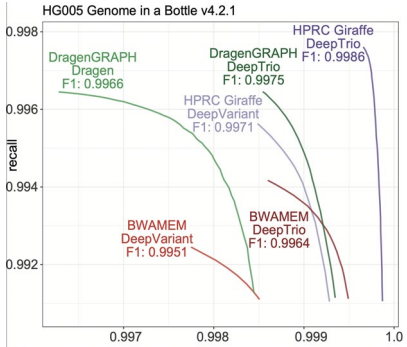
Sairam Behera, Severine Catreux, Massimiliano Rossi, Sean Truong, Zhuoyi Huang, Michael Ruehle, Arun Visvanath, Gavin Parnaby, Cooper Roddey, Vitor Onuchic, Daniel L Cameron, Adam English, Shyamal Mehtalia, James Han, Rami Mehio, Fritz J Sedlazeck

doi: https://doi.org/10.1101/2024.01.02.573821

Personalized Pangenome References

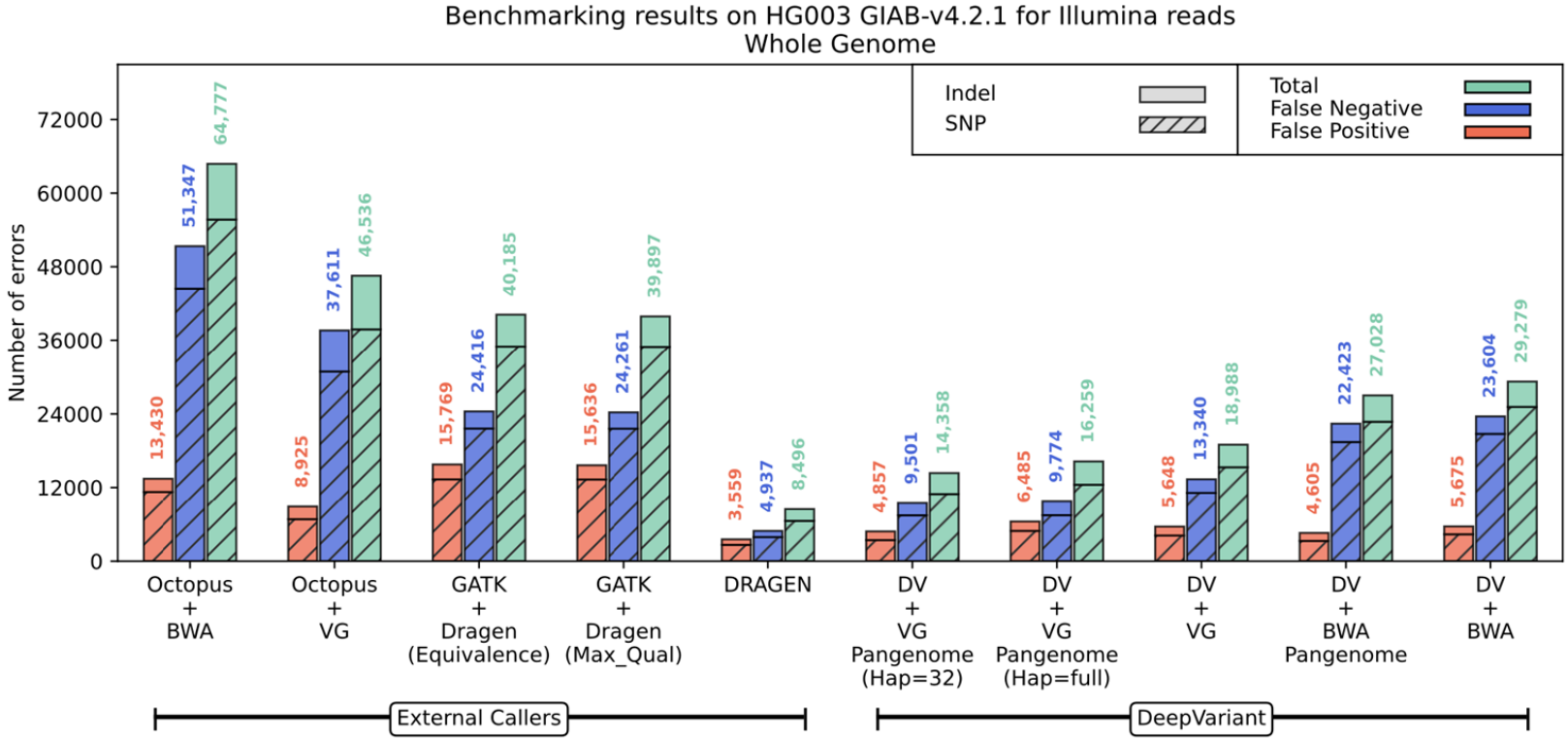
Jouni Sirén, Parsa Eskandar, Matteo Tommaso Ungaro, Glenn Hickey, Jordan M. Eizenga, Adam M. Novak, Xian Chang, Pi-Chuan Chang, Mikhail Kolmogorov, Andrew Carroll, Jean Monlong, Benedict Paten

doi: https://doi.org/10.1101/2023.12.13.571553



Credit: Jean Monlong, Google Health

Now Released: Pangenome Aware DeepVariant



Courtesy, Mobin Asri & Google Health

Now Released: Pangenome Aware DeepVariant

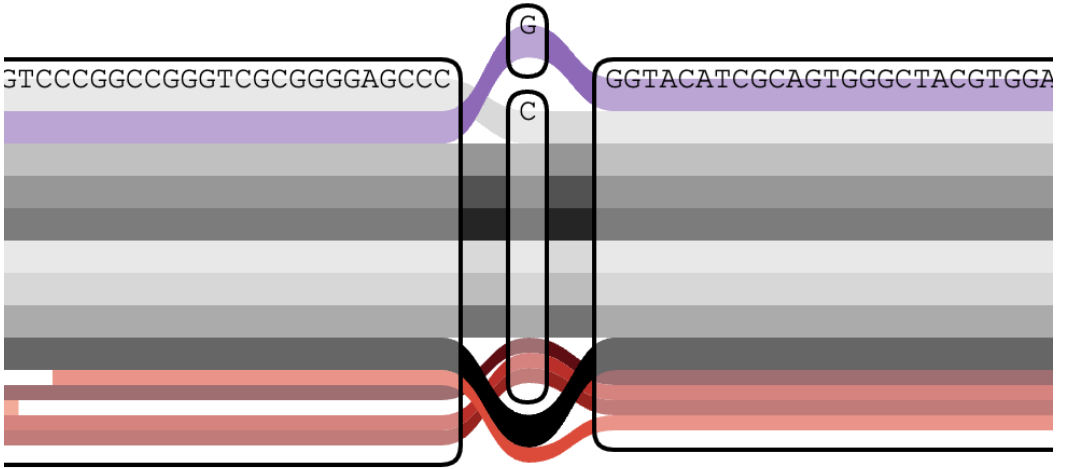
Method	Reads	Graph	Total Error
DRAGEN	**illumina-novaseq**		**74042**
VG + DV	**illumina-novaseq**	hprc1.1	**103827**
VG + pangenome-aware DV	**Element**	hprc1.1	**53348**

Q100 T2T Benchmark

Courtesy, Mobin Asri & Google Health

Personalized Pangenomes

- Most rare variation added to a pangenome will be absent from a given sample under study
- This rarer variation asks <u>like noise</u>, causing mismapped reads
- This problem gets worse as the pangenome scales



Consider the purple haplotype.
The sample (reads in red) probably don't contain it

Personalized Pangenomes

- (Dirty secret) To solve this problem, pangenome mapping pipelines throw away most of the variation in the pangenome!
- This is a waste: tens of thousands of those rarer variants will be in the sample under study

vg map: 1% threshold in the 1000GP graph
(Garrison et al.: Variation graph toolkit improves read mapping by representing genetic variation in the reference, 2018)

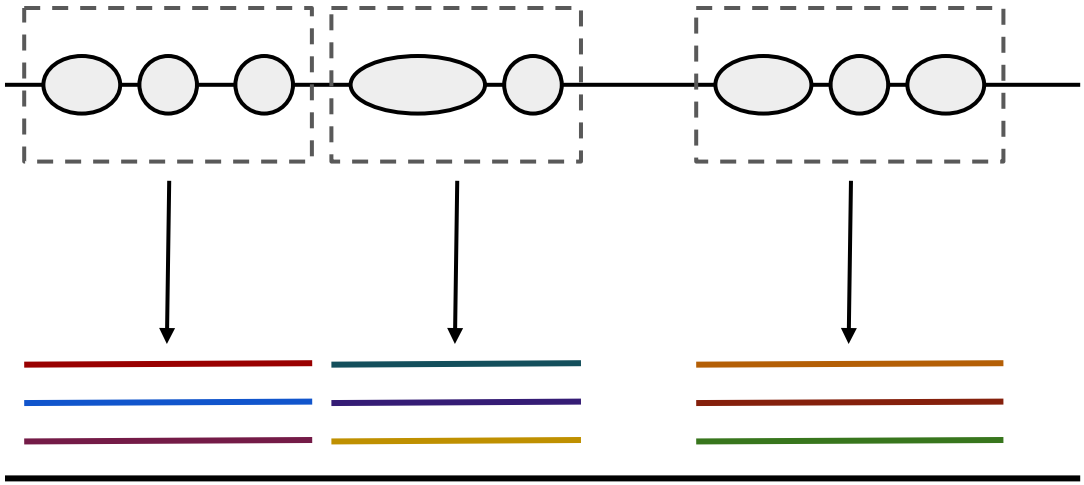
FORGe: consider both frequency and effect on repetitiveness
(Pritt et al.: FORGe: prioritizing variants for graph genomes, 2018)

Giraffe: 64 synthetic haplotypes based on proportional sampling of the 1000GP graph
(Sirén et al.: Pangenomics enables genotyping of known structural variants in 5202 diverse genomes, 2021)

Minigraph–Cactus: 10% threshold in the HPRC graph intended for Giraffe
(Hickey et al.: Pangenome graph construction from genome alignments with Minigraph-Cactus, 2023)

Personalized Pangenomes

- Solution: preprocess the pangenome to locally select only relevant haplotypes:
  - Uses kmers from sample
  - Picks haplotypes in each 10kb subgraph, (currently) free recombination
- Two selection modes:
  - M best haplotypes
  - Diploid sampling:
    - Optimal 2 from M



Credit: Dr. Jouni Siren

Personalized Pangenomes: Mapping Speed

BWA-MEM to GRCh38 (for comparison)

Mapping 7452 s

Filtered v1.1 graph

Mapping 4181 s

Diploid sampling (32 candidates from v1.1 graph) + reference

Mapping 3499 s

Index construction 976 s

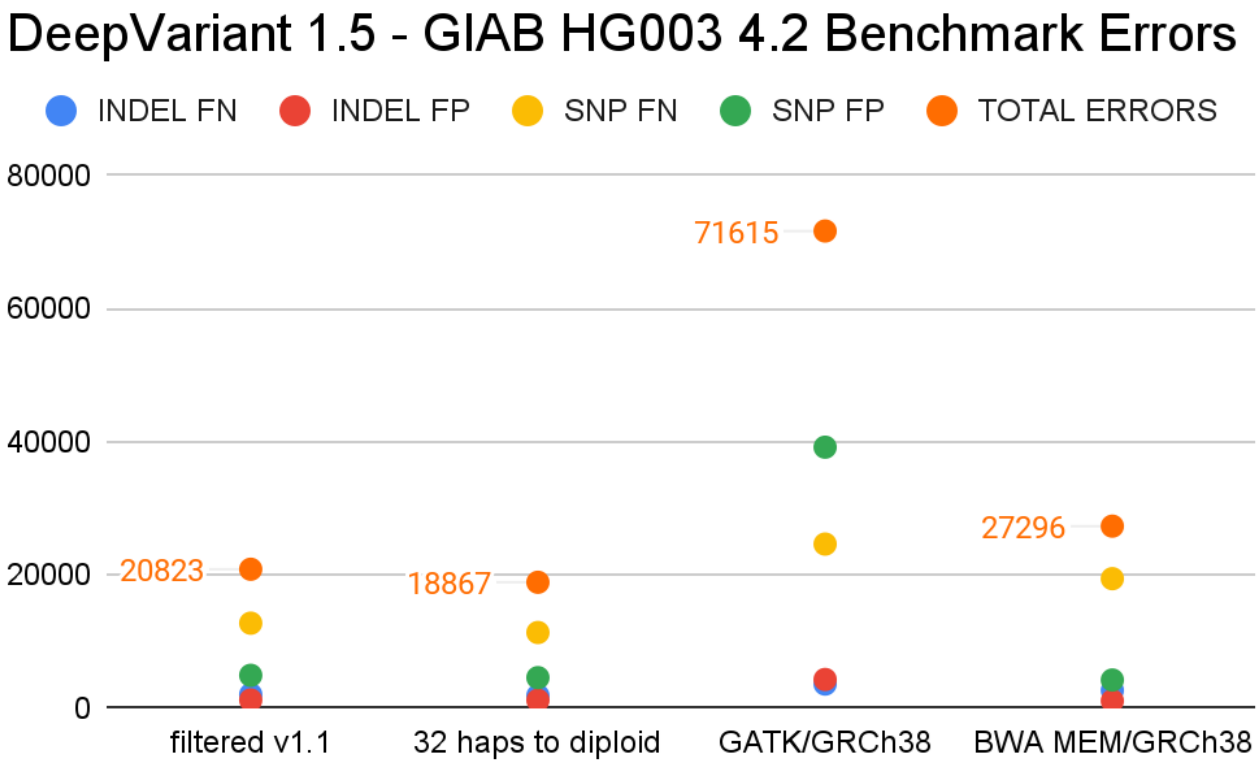
Sampling 299 s

Kmer counting 429 s

HG002, 30x NovaSeq reads
AWS i4i.16xlarge
KMC3 for kmer counting
(k = 29, w = 11) minimizers
Giraffe with 32 threads

Credit: Dr. Jouni Siren

Compared to linear-reference methods, personalized diploid graph has fewest small errors



Personalized diploid:

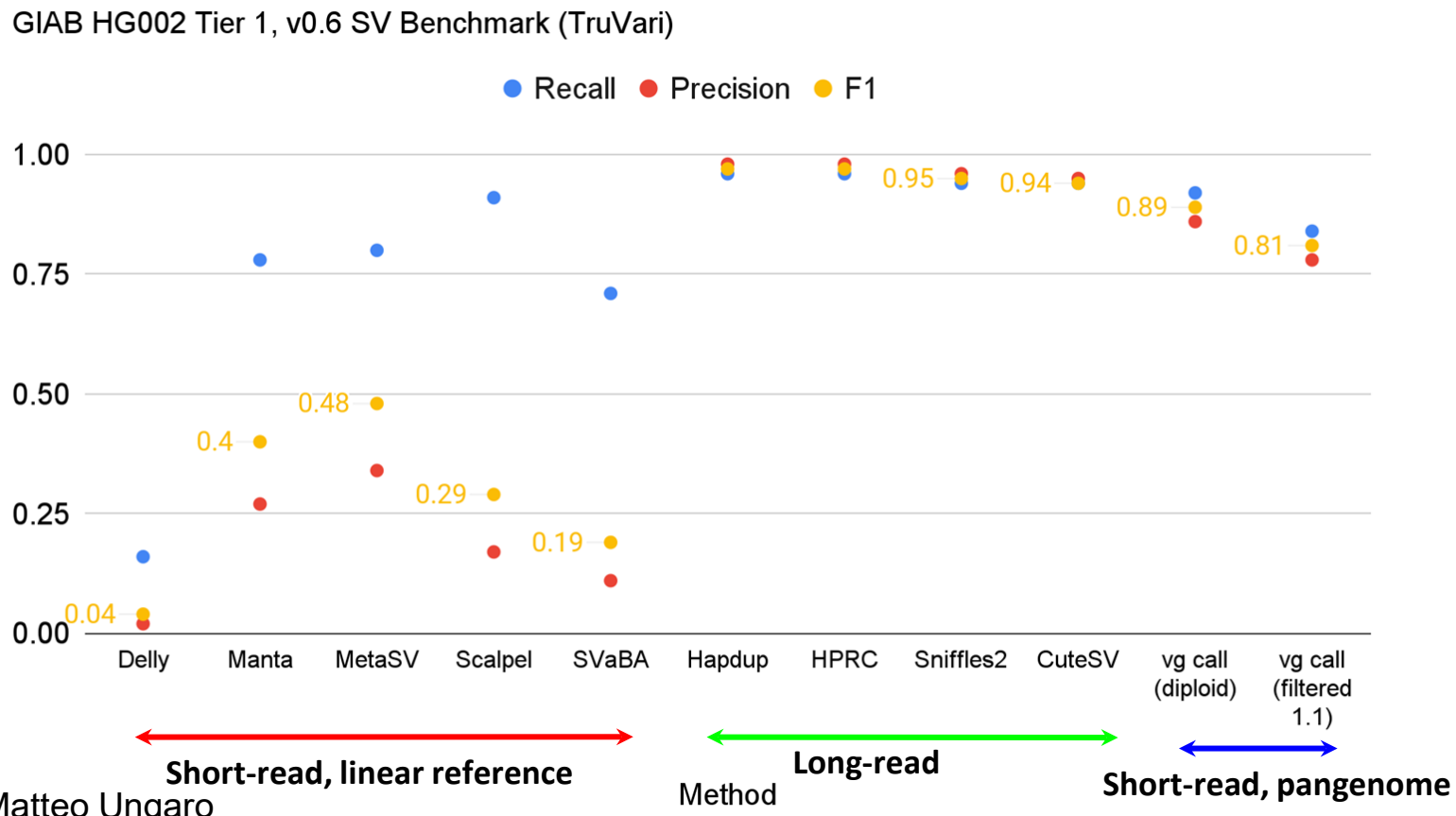
- GATK* - 379% more errors
- BWA MEM / GRCh38 - 45% more errors

Mapping to Minigraph–Cactus GRCh38 based graphs

HG003, 40x NovaSeq reads

Credit: Parsa Eskander

Personalized diploid makes short-read SV typing almost competitive with long-read discovery methods



Credit: Matteo Ungaro