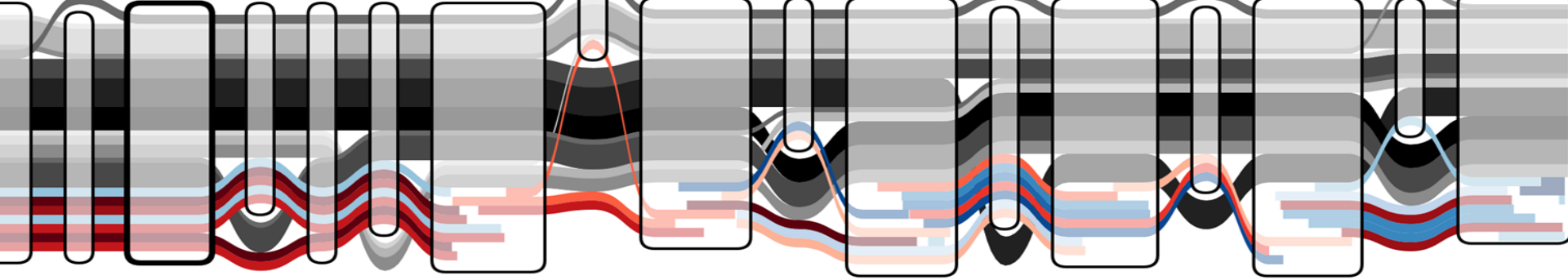


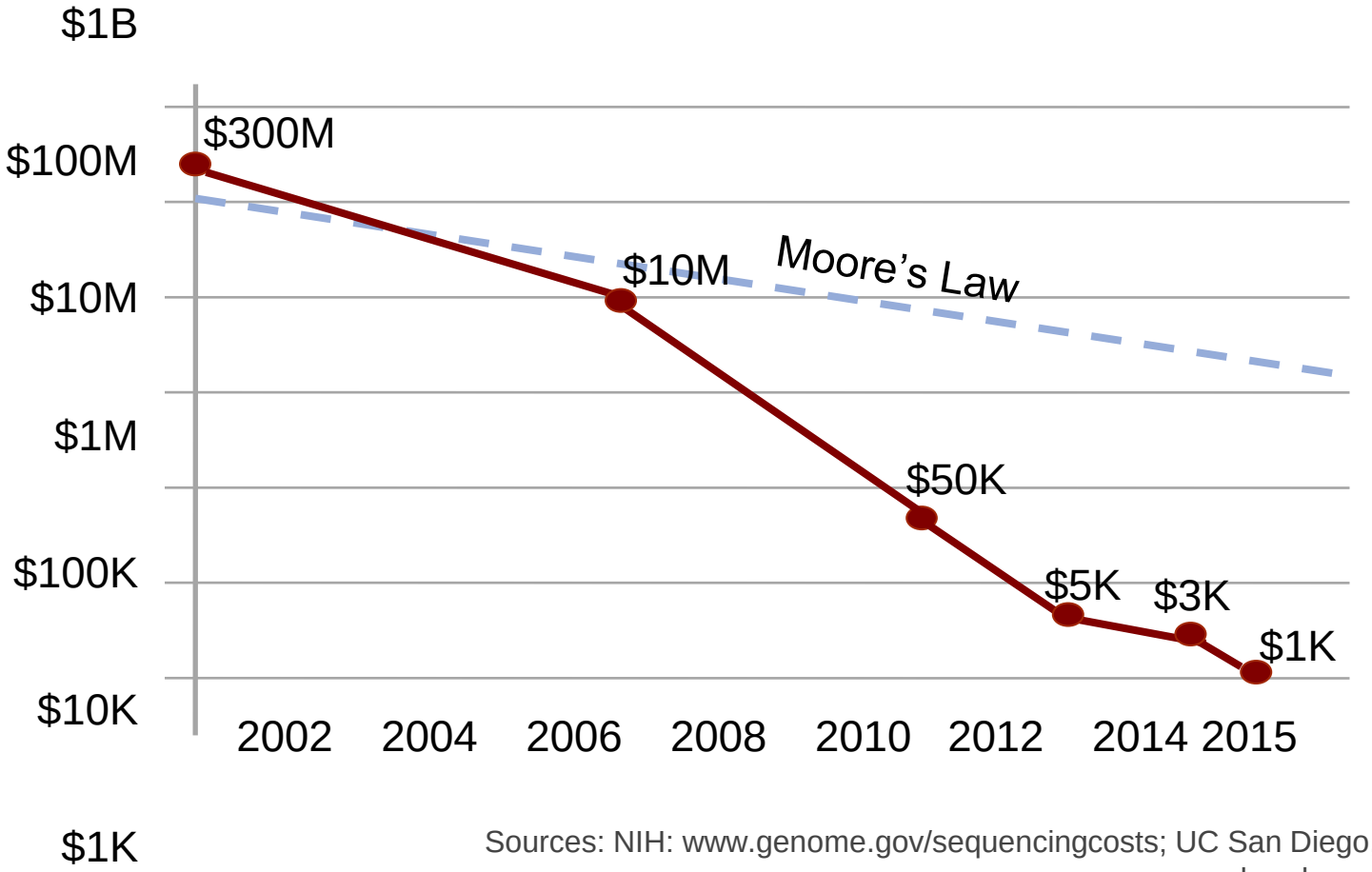
Building the human pangenome

Benedict Paten - UC Santa Cruz Genomics Institute

bpaten@ucsc.edu

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Now the $1,000 individual genome is here… but



Sources: NIH: www.genome.gov/sequencingcosts; UC San Diego, 1/14/14: Illumina breaks genome cost barrier

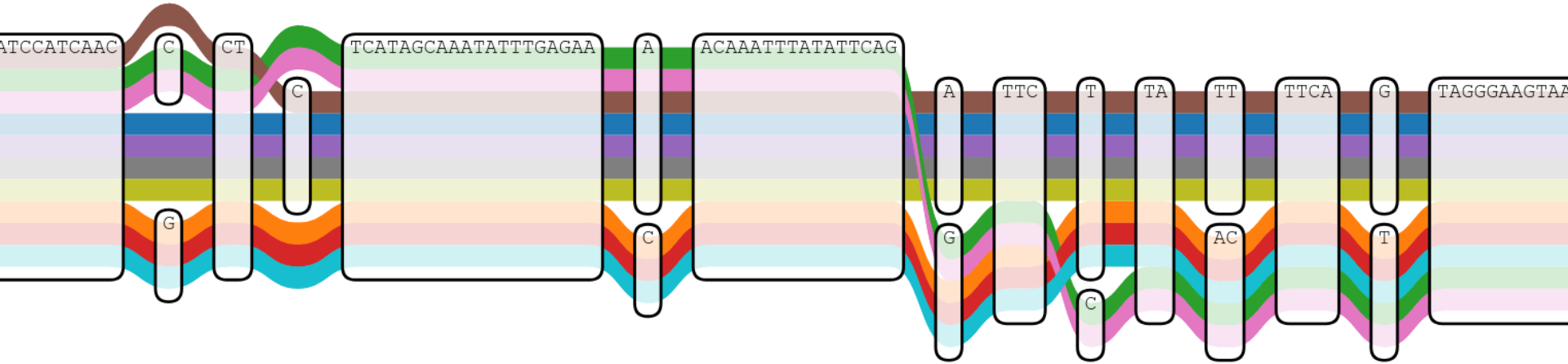
All variants are currently detected relative to a single human reference genome. A typical person is not the reference.

A typical person has

- Avg. of **5 million** isolated single DNA base variations **different** from the reference (out of 3 billion)
- Avg. of **20 million** DNA bases in large segments of DNA that are **not present** in the same form in the reference genome
- Many of these variants not currently assayed accurately: reference allele bias

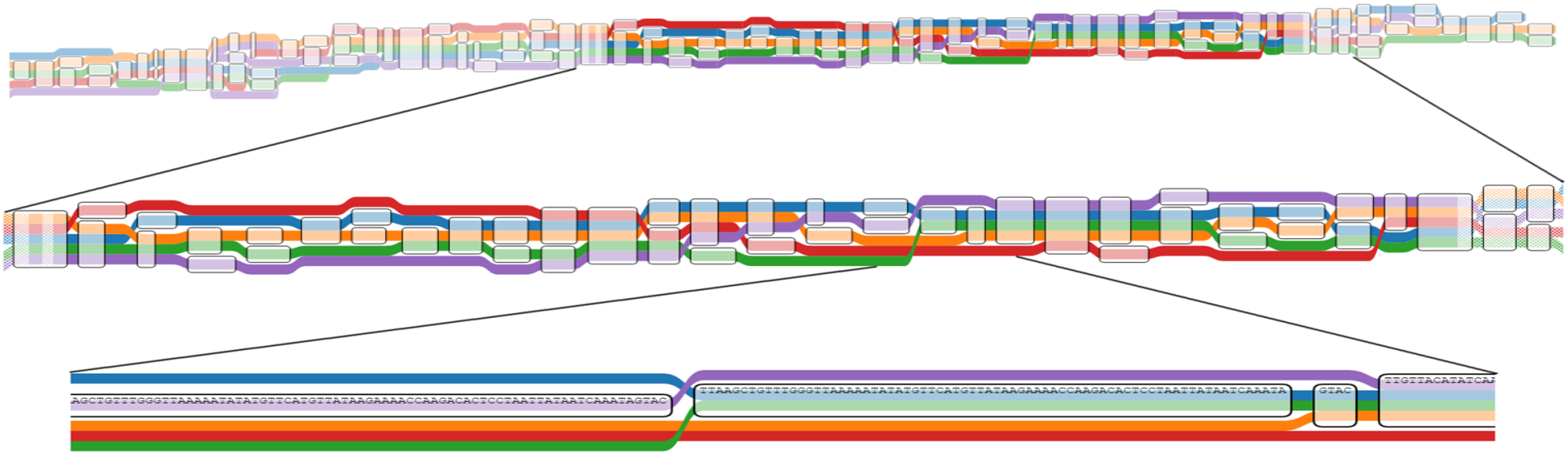
Vision - The Human Pangenome

Instead - imagine mapping to a reference structure that contains all common variation: a pangenome graph



This Talk

- Part 1: How do we make long-read reference quality assembly efficient and routine, so that we can create the genomes for the human pangenome
- Part 2: How do we build the pangenome and use it?



Genome assembly bottlenecks

- Need for revolution in generation of high-quality genomes to ensure all variation is captured, bottlenecks:
  - Sequencing cost for high quality
  - Sequencing speed for high quality
  - Scalable and cheaper informatics

Solution

- Nanopore 100kb+ sequencing
- Scalable algorithms and informatics

Article | OPEN | Published: 29 January 2018

Nanopore sequencing and assembly of a human genome with ultra-long reads

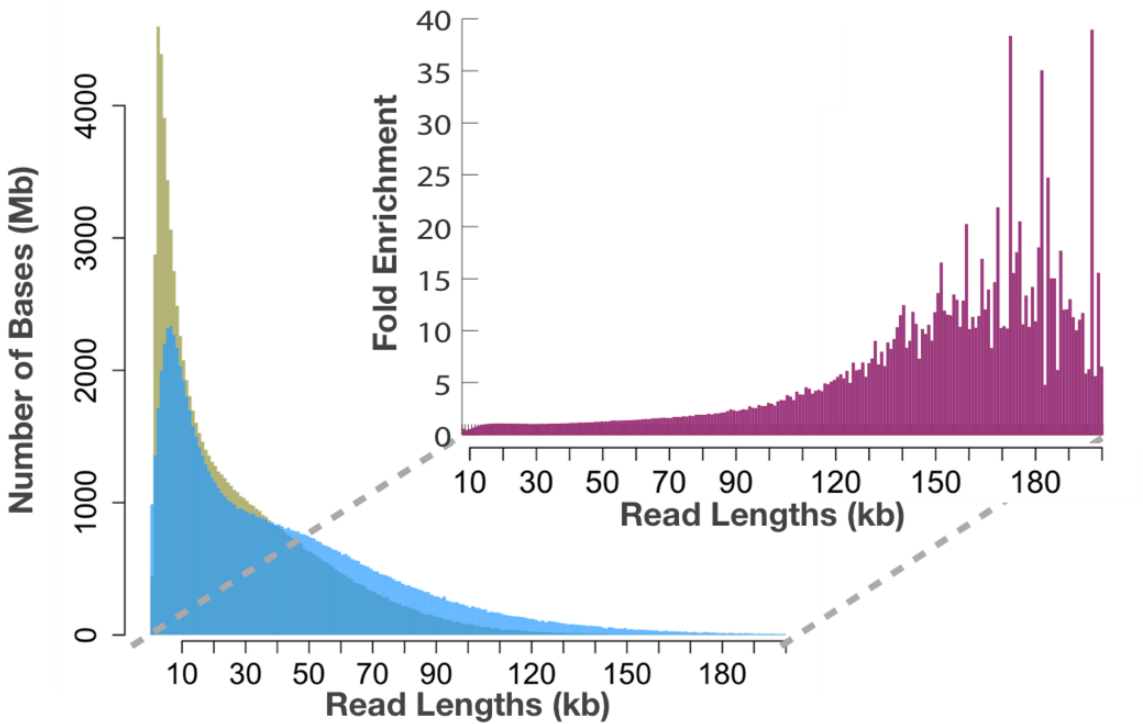
Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman ✉ & Matthew Loose ✉ - Show fewer authors

Nature Biotechnology **36**, 338–345 (2018) | Download Citation ↓

Nanopore sequencing

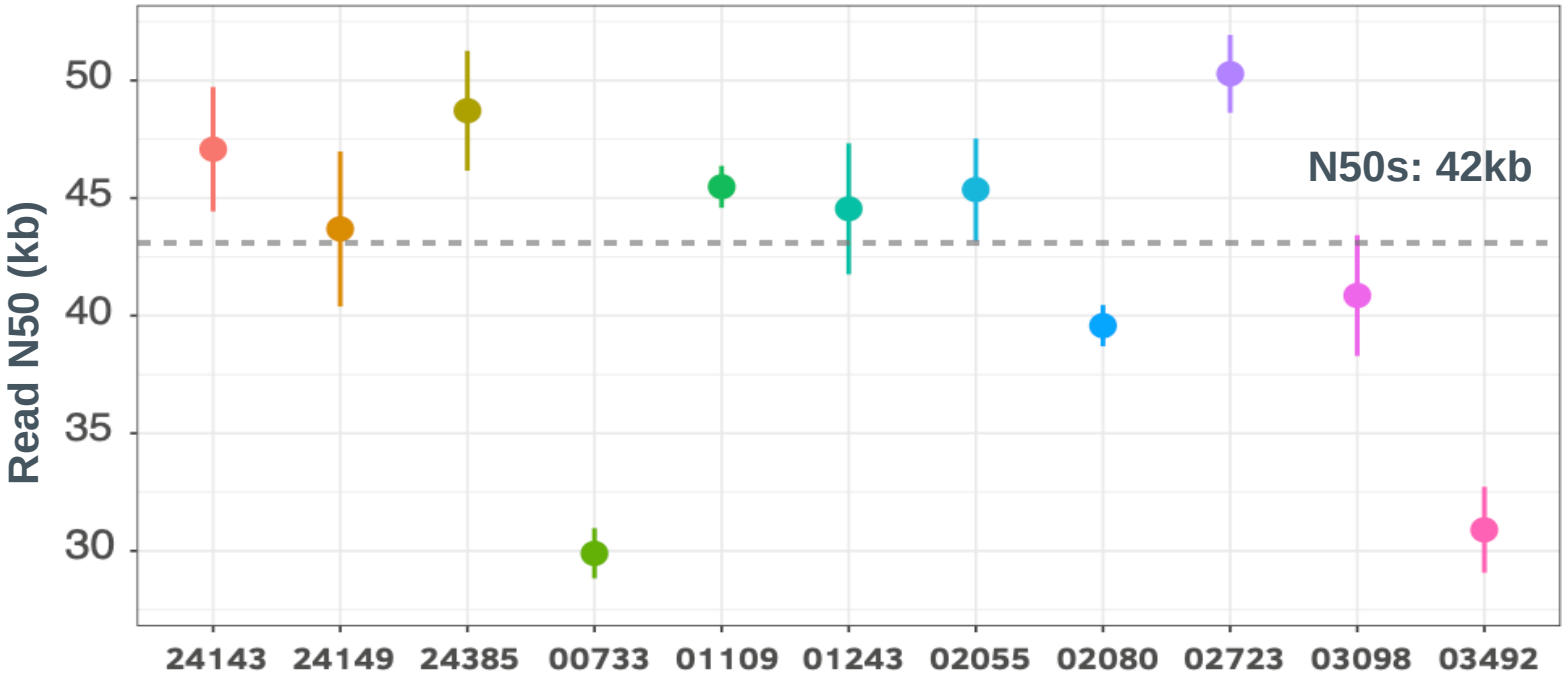
Data acquisition for 11 genomes in 9 days
(>60x total coverage)

7x enrichment of reads >100kb using Circulomics SRE



Short Read Eliminator Kit (https://www.circulomics.com)

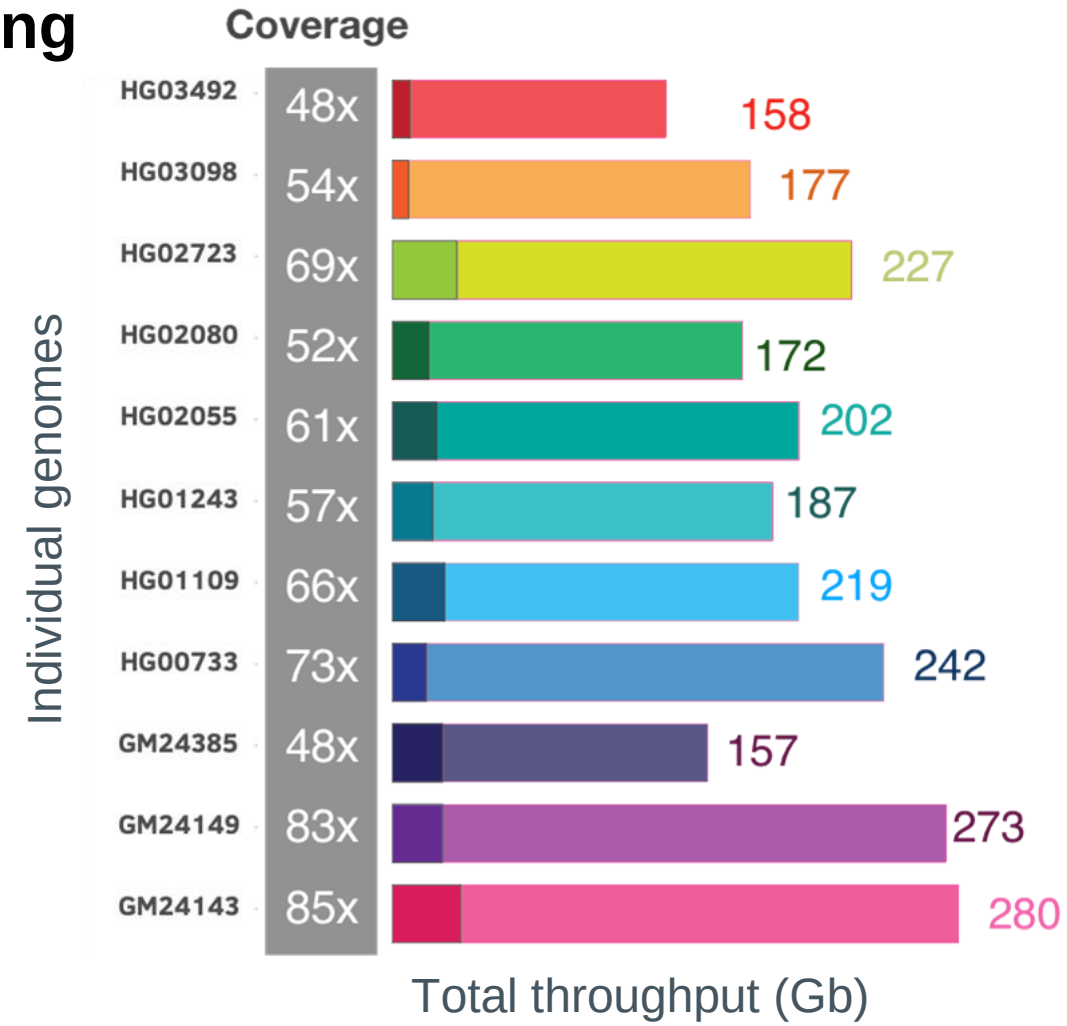
Read N50 improvement is reproducible



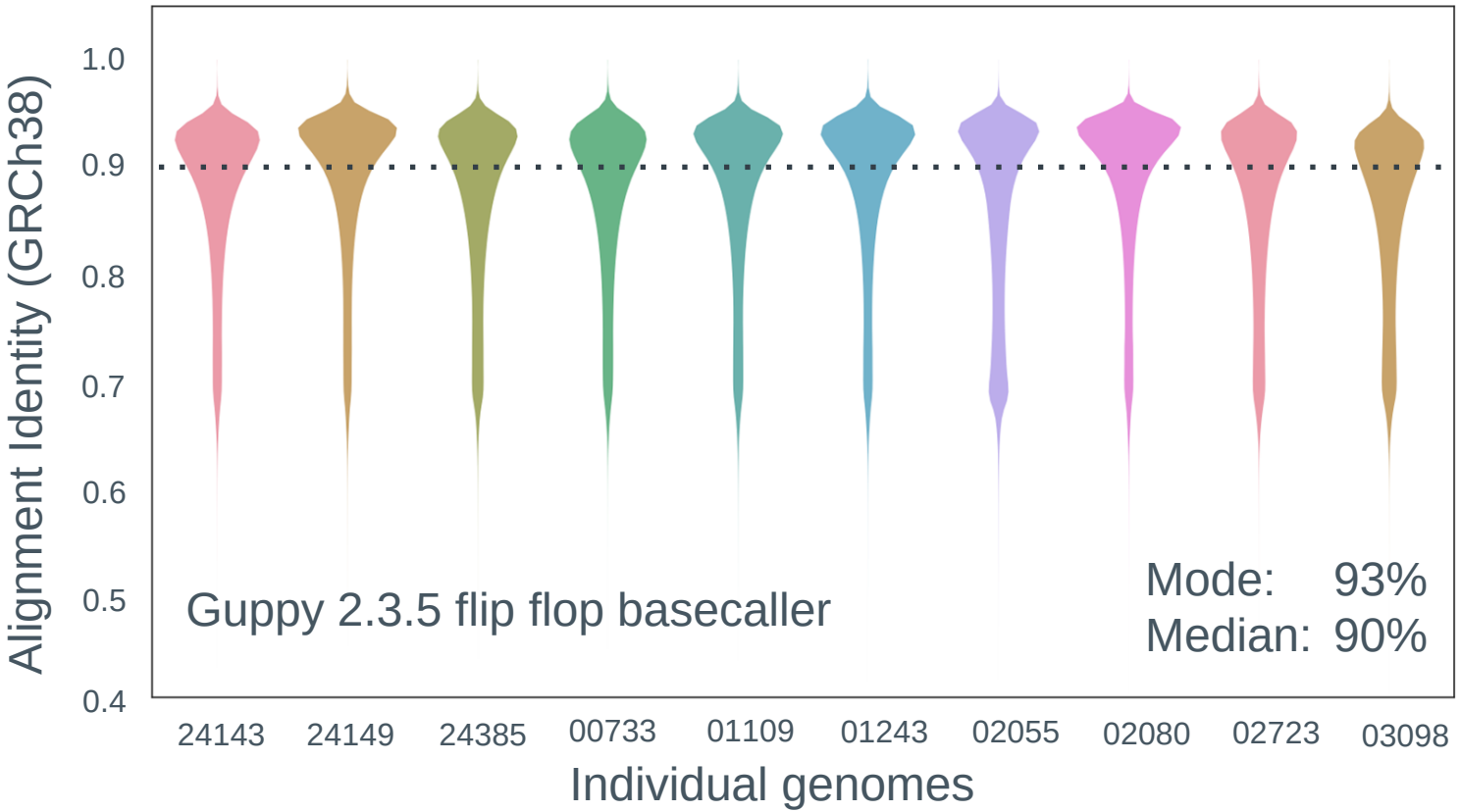
Individual genomes

https://github.com/human-pangenomics/hpgp-data

PromethION sequencing throughput



Median alignment identity is 90%



Alignment identity = matches / (matches + mismatches + insertions + deletions)

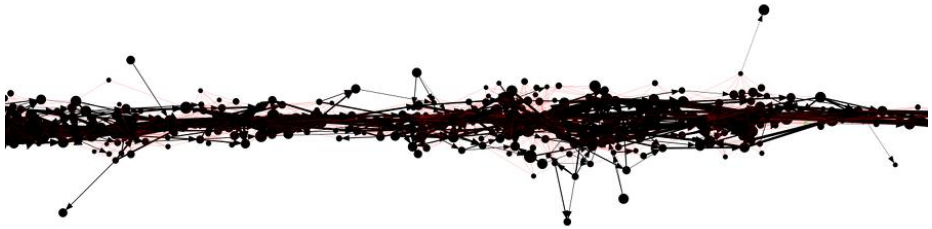
Scalable assembly and polishing tools

Pipeline


Sequencing/
Basecalling
Assembly
amazon
web services
Polishing
Scaffolding
Guppy
Flip-Flop
Shasta
Margin-
Polish
HELEN
HiRise
Phasing
FINISHED
ASSEMBLY

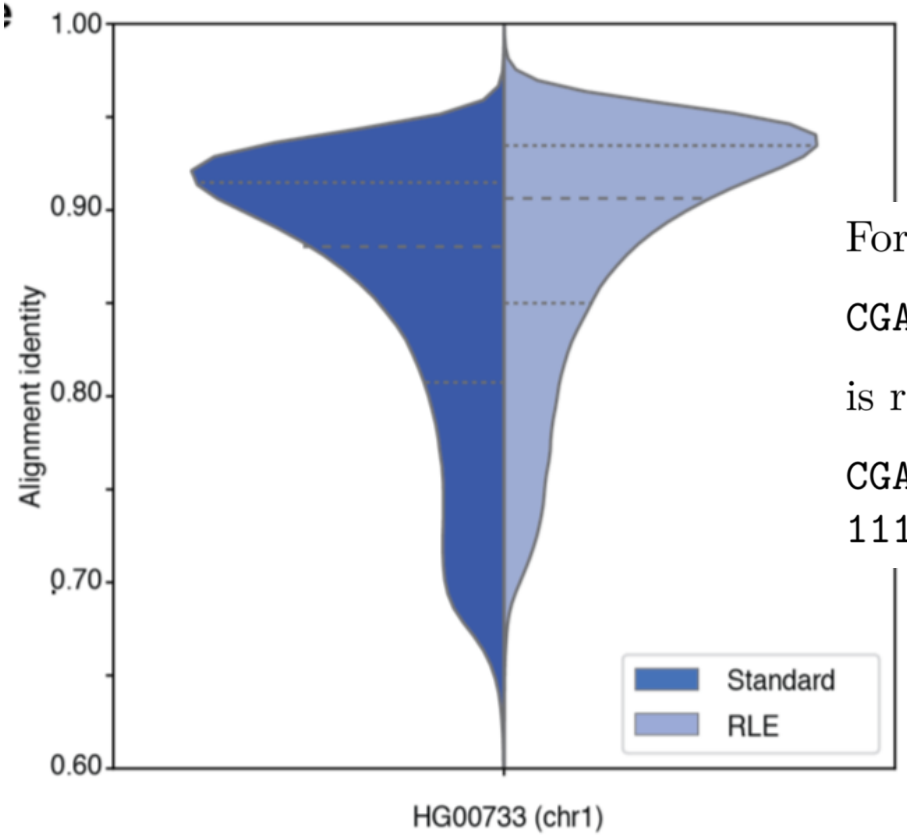
Shasta – a nanopore *de novo* long read assembler

- New *de novo* assembler tailored for long reads and parameterized for ONT data - principally developed by Paolo Carnevali at CZI
- Beautiful new algorithms (https://chanzuckerberg.github.io/shasta/ComputationalMethods.html)
    - Use run-length encoding (RLE) throughout to compress homopolymer confusion - the dominant source of error in ONT reads
    - Uses novel high-cardinality marker space representation for super efficient overlap alignment
    - Does everything in memory (requires 1.5TB of memory for 60x human)
    - Outputs GFA, intent for whole pipeline to use GFA to represent ambiguities



https://github.com/chanzuckerberg/shasta

Run Length Encoding (RLE)



For example, the following read

`CGATTTAAGTTA`

is represented as follows using run-length encoding:

```
CGATAGTA
11132121
```

Marker Representation

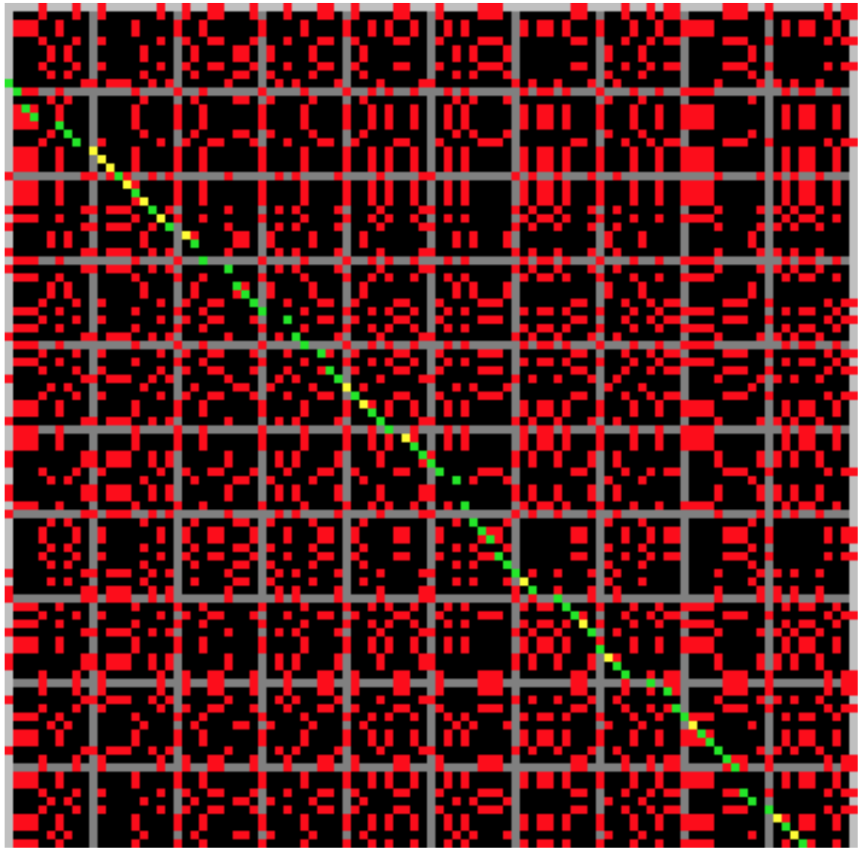
Consider now the following portion of a read in run-length representation (here, the repeat counts are irrelevant and so they are omitted):

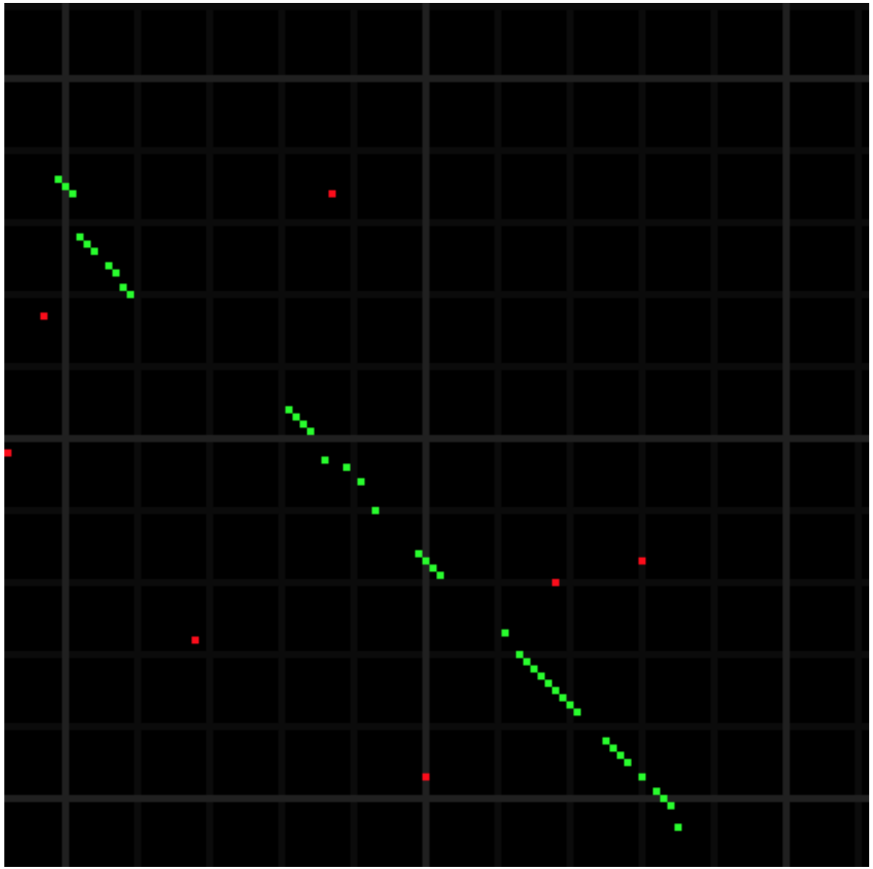
```
CGACACGTATGCGCACGCTGCGCTCTGCAGC
 GAC      TGC    CGC        TGC
           CGC     TGC      GCA
            GCA      CGC
```

Occurrences of the k-mers defined in the table above are shown and define the markers in this read. Note that markers can overlap. Using the marker ids defined in the table above, we can summarize the sequence of this read portion as follows:

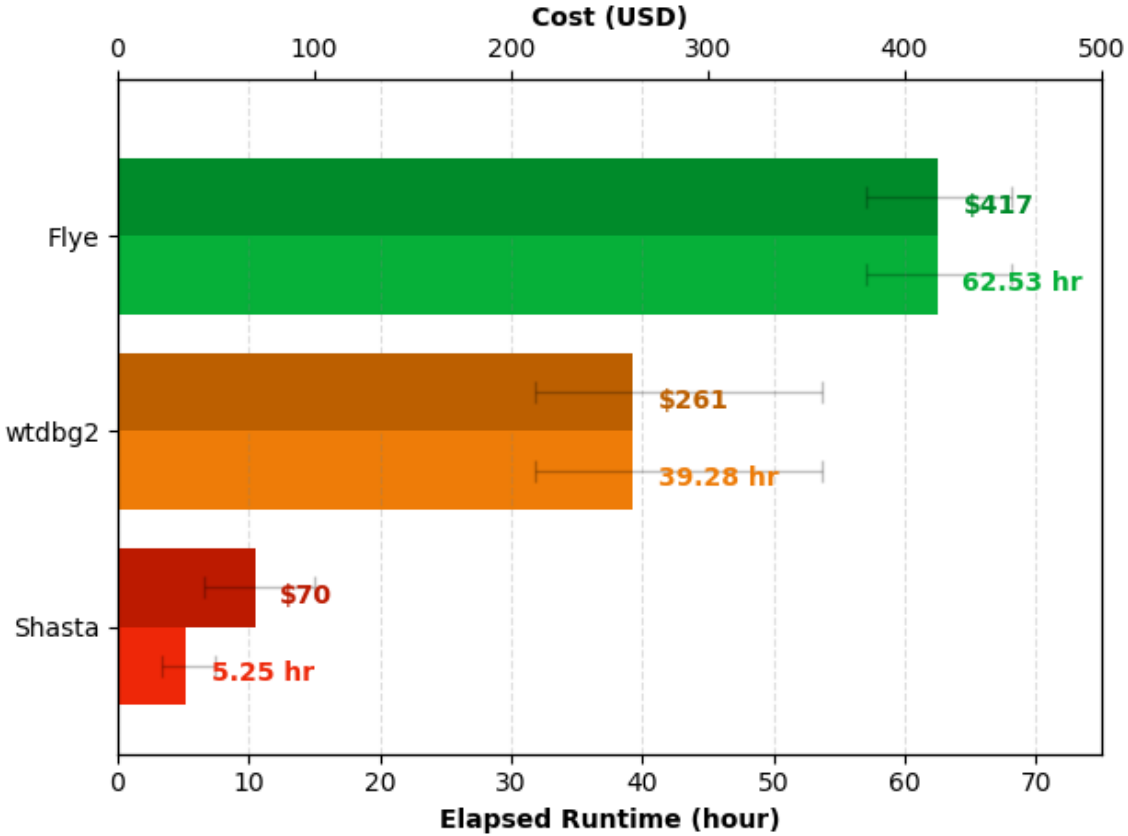
```
2 0 3 1 3 0 3 0 1
```

Marker Representation





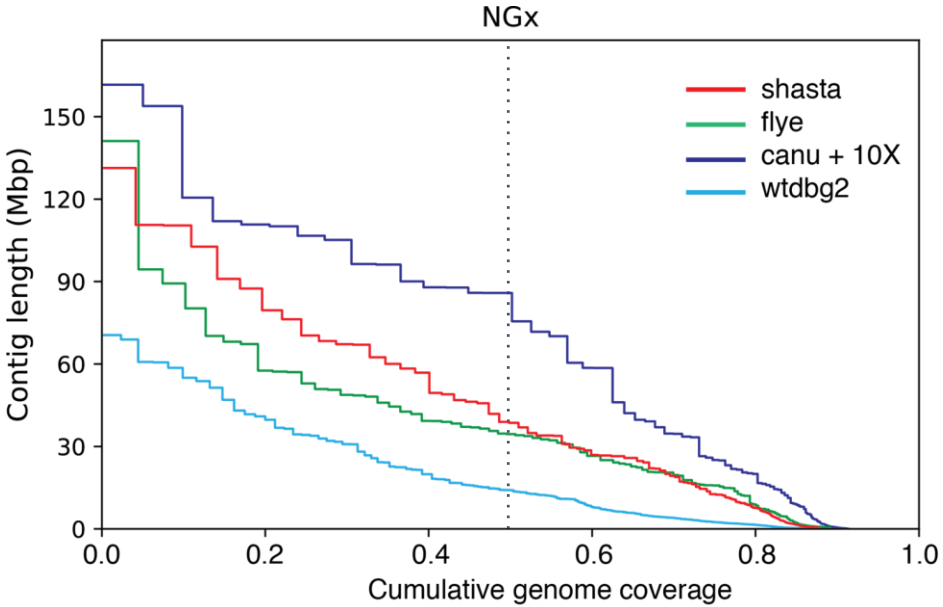
Assembly at a fraction of time and cost

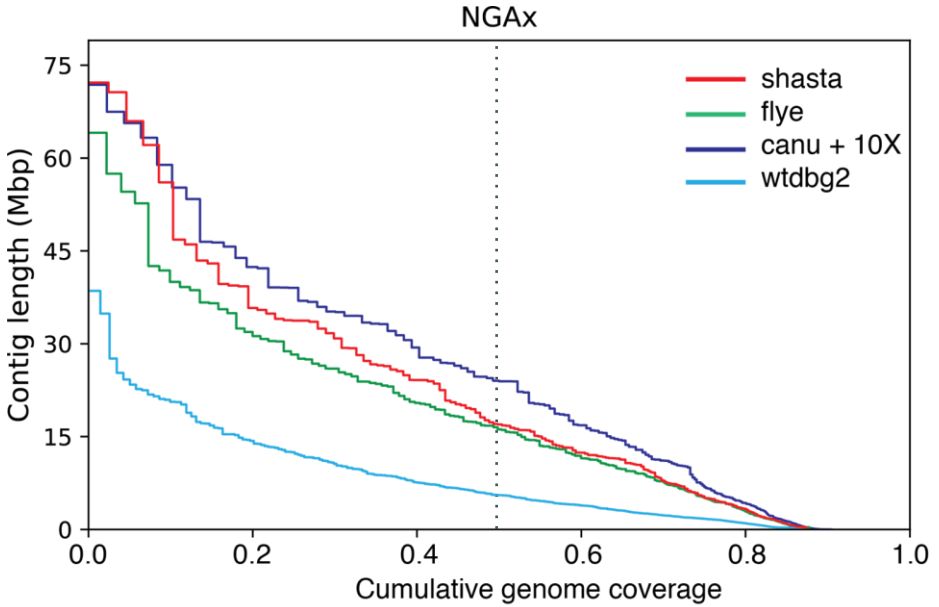


Shasta GPU Acceleration

	CPU*	GPU*
Instance	r5.16xlarge	g3.16xlarge
Config	(64vCPUs, 512GB)	(64vCPUs, 488GiB, 4 M60 GPUs)
Instance cost	$4.032/hr	$4.45/hr
Assembly time	159m	131m (**~1.2x faster**)
Assembly cost	$10.68	$9.71 (**~9% cheaper**)
Total length	2.773 Gb	2.775 Gb
N50	21.74 Mb	21.94 Mb
# misassemblies	866	801
# local misassemblies	955	886
# mismatches per 100 kbp	209.63	210.78
# indels per 100 kbp	418.60	420.08

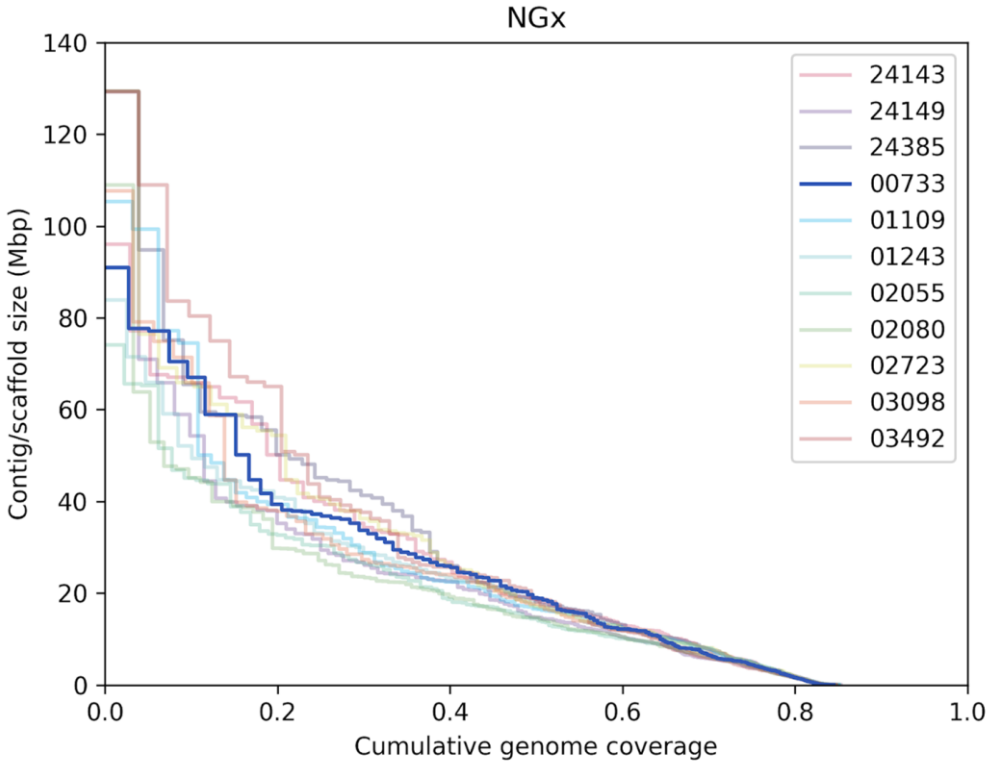
Comparable contig NG50 and lower misassemblies





	shasta	flye	canu + 10X	wtdbg2
Number of misassemblies	**1160**	5580	6093	4164

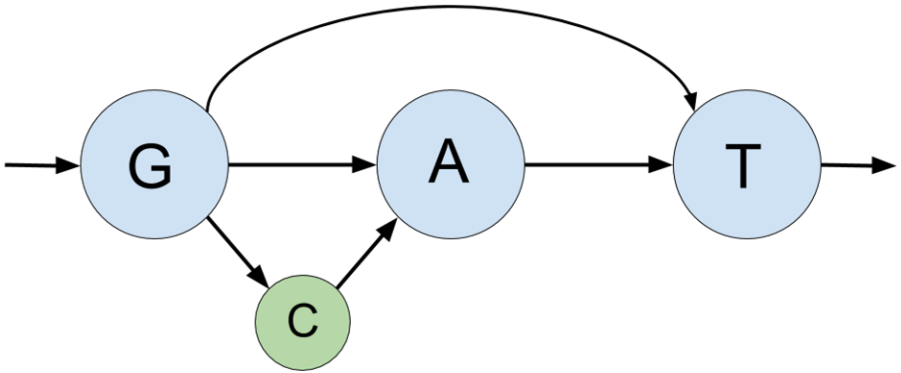
Shasta assemblies are reproducible



Median contig NG50 = 23 Mb

Two-step polishing of assemblies

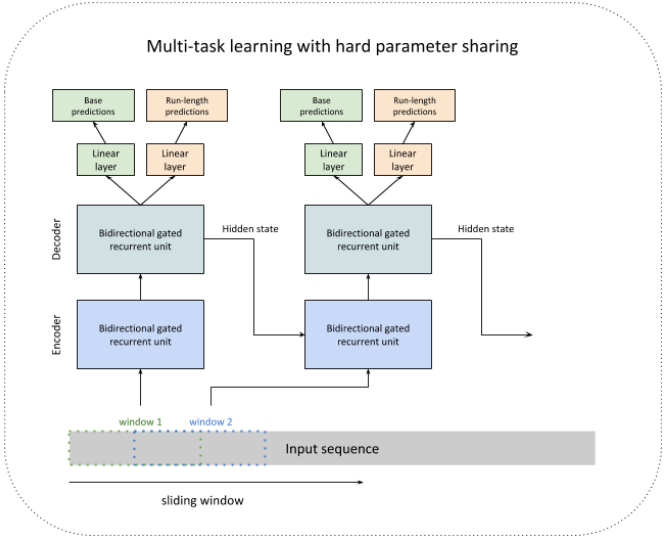
1. MarginPolish



A graph-based alignment polisher

https://github.com/UCSC-nanopore-cgl/marginPolish

2. HELEN

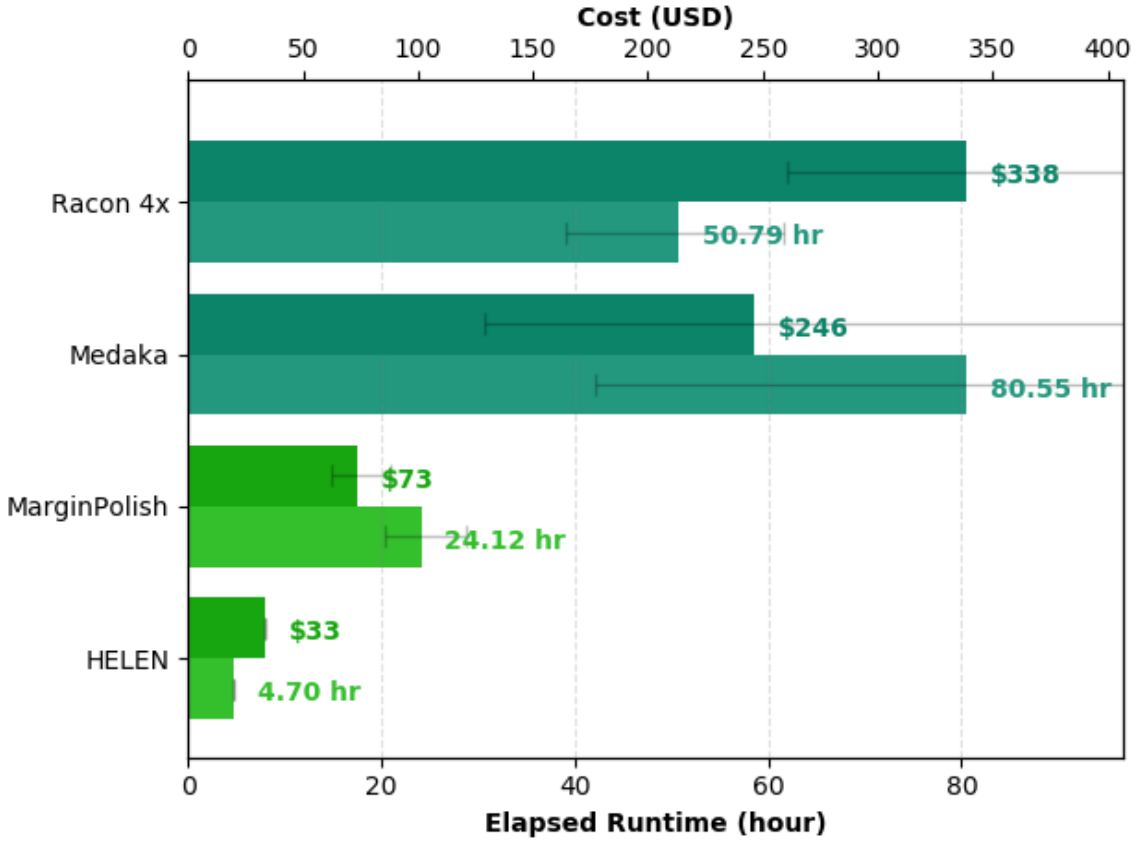


A DNN-based consensus sequence polisher

https://github.com/kishwarshafin/helen

UC SANTA CRUZ

Polishing at a fraction of time and cost

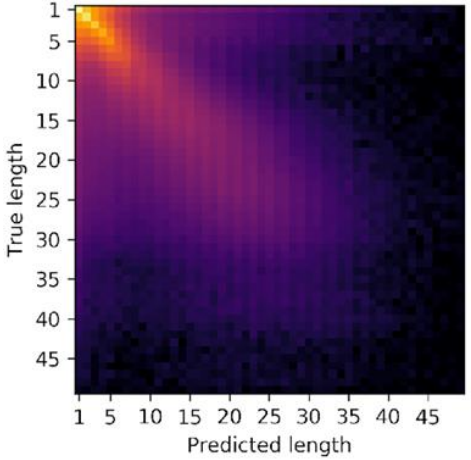


MarginPolish and HELEN outperform other polishers

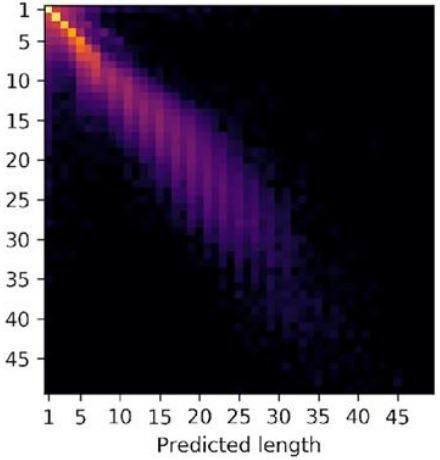
Assembler	Polisher	Diploid (HG00733)	Haploid (CHM13)
Shasta	-	98.78%	99.37%
	$Racon_{4x}$	99.16%	99.50%
	$Racon_{4x}$+ Medaka	99.42%	99.58%
	MarginPolish	99.41%	99.62%
	MarginPolish + HELEN	**99.47%**	**99.70%**

Improvements in homopolymer length predictions

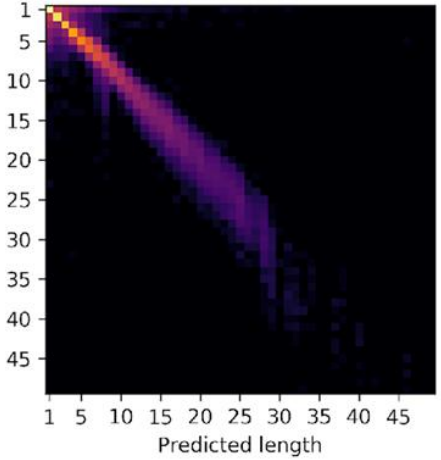
Guppy basecaller



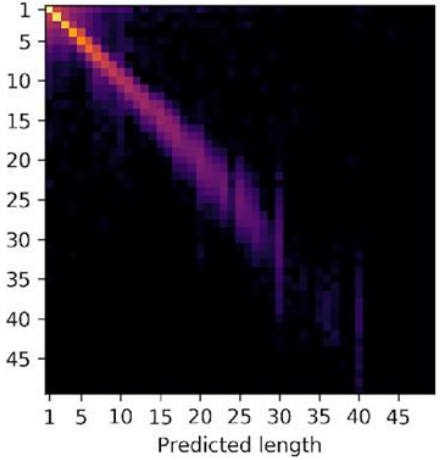
Shasta



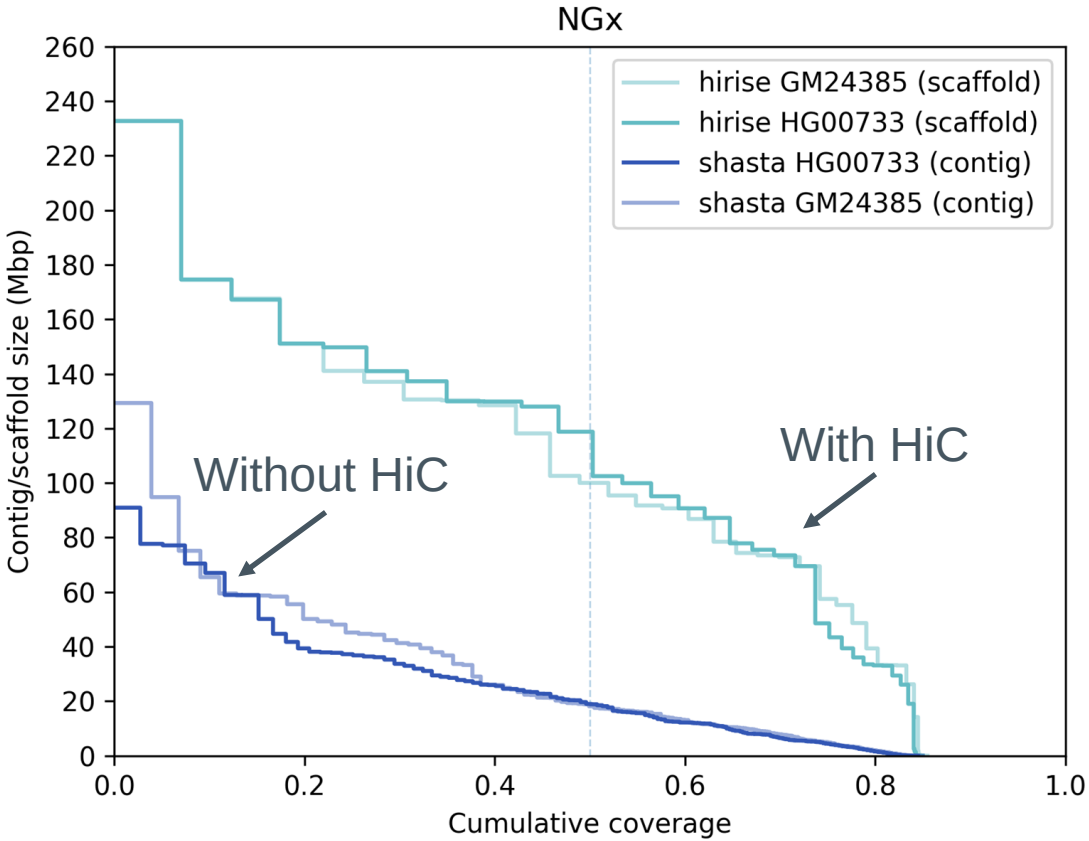
Shasta
+ MarginPolish



Shasta
+ MarginPolish
+ HELEN

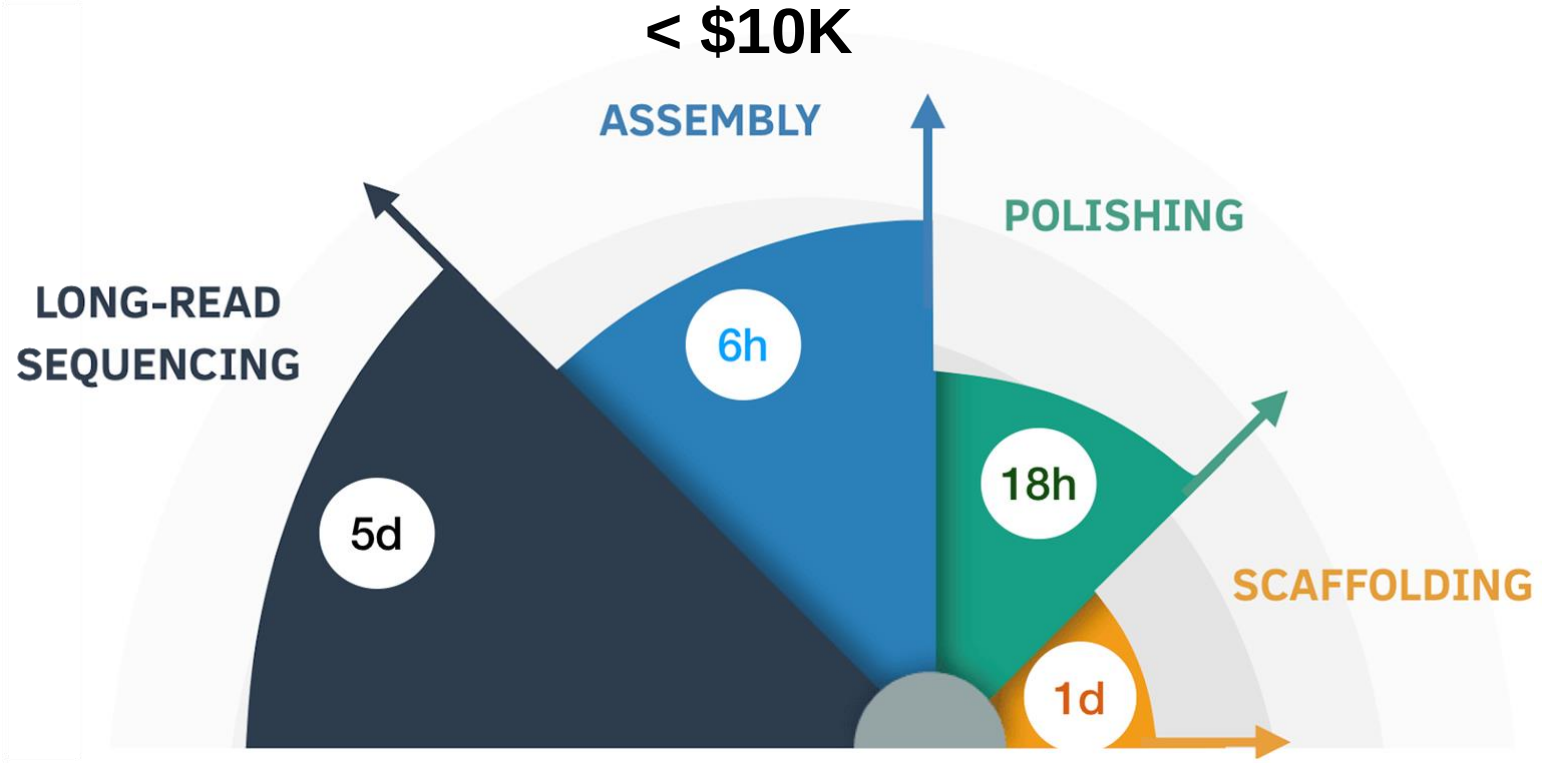


Chromosome-level scaffolding using HiC data



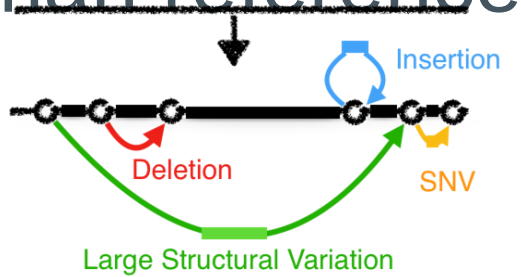
Near term future

The near future: A reference-quality human-scale genome in ~7 days for < $10K



Key next steps

- Faster basecalling (ONT)
- Haplotype phasing (UCSC, CZI)
- Exploring real-time applications
- Integrating into human reference pan-genomes



Acknowledgements

David Haussler

Ed Green

Sofie Salama

Mark Akeson

Kristof Tigyi

Nicholas Maurer

Yatish Turakhia

Kishwar Shafin

Marina Haukness

Trevor Pesout

Colleen Bosworth

Karen Miga

Ryan Lorig-Roach

Miten Jain

Hugh Olsen

Adam Phillippy (NHGRI)

Fritz Sedlazeck (Baylor)

Adam Novak

Glenn Hickey

Jordan Eizenga

Erik Garrison

Jean Monlong

Xian Chang

Daniel Garalde

Rosemary Dokos

Simon Mayes

Chris Seymour

Chris Wright

David Stoddart

Dan Turner

Sidney Bell

Charlotte Weaver

Michael Barrientos

Ryan King

Bruce Martin

Phil Smoot

Cori Bargmann

Kelvin Liu

Duncan Kilburn

Mapping everybody's genome to one reference genome creates significant bias

- Mapping is biased against variation
- Structural variants particularly hard to map
- Risk some genetic variants from other subpopulation groups inaccurately represented
- Bias is unacceptable for global biomedicine

Korean reference genome project

***De novo* assembly and phasing of a Korean human genome**

Jeong-Sun Seo et al. 2016

Danish reference genome project

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

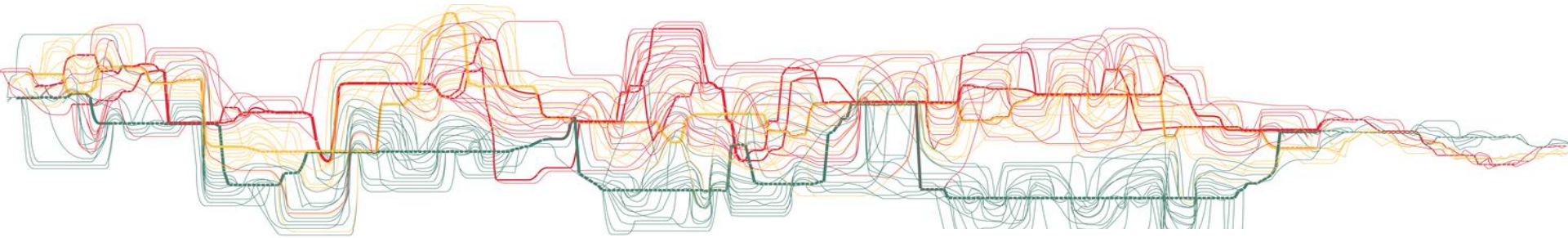
Lasse Maretty et al. 2017

…

Human Pangenome Project

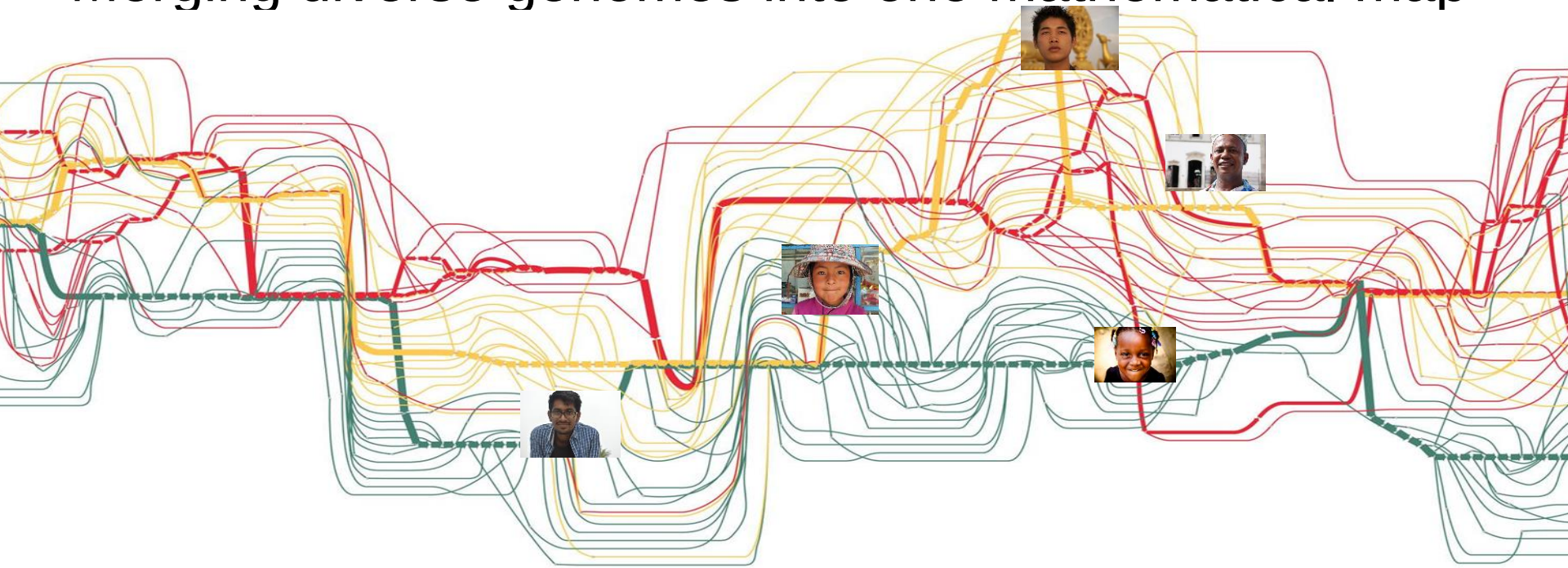
Goals:

- Develop next generation human genetic reference that includes known variation from all human ethnic populations
- Build the software required to switch biomedicine over to using this new human genetic reference



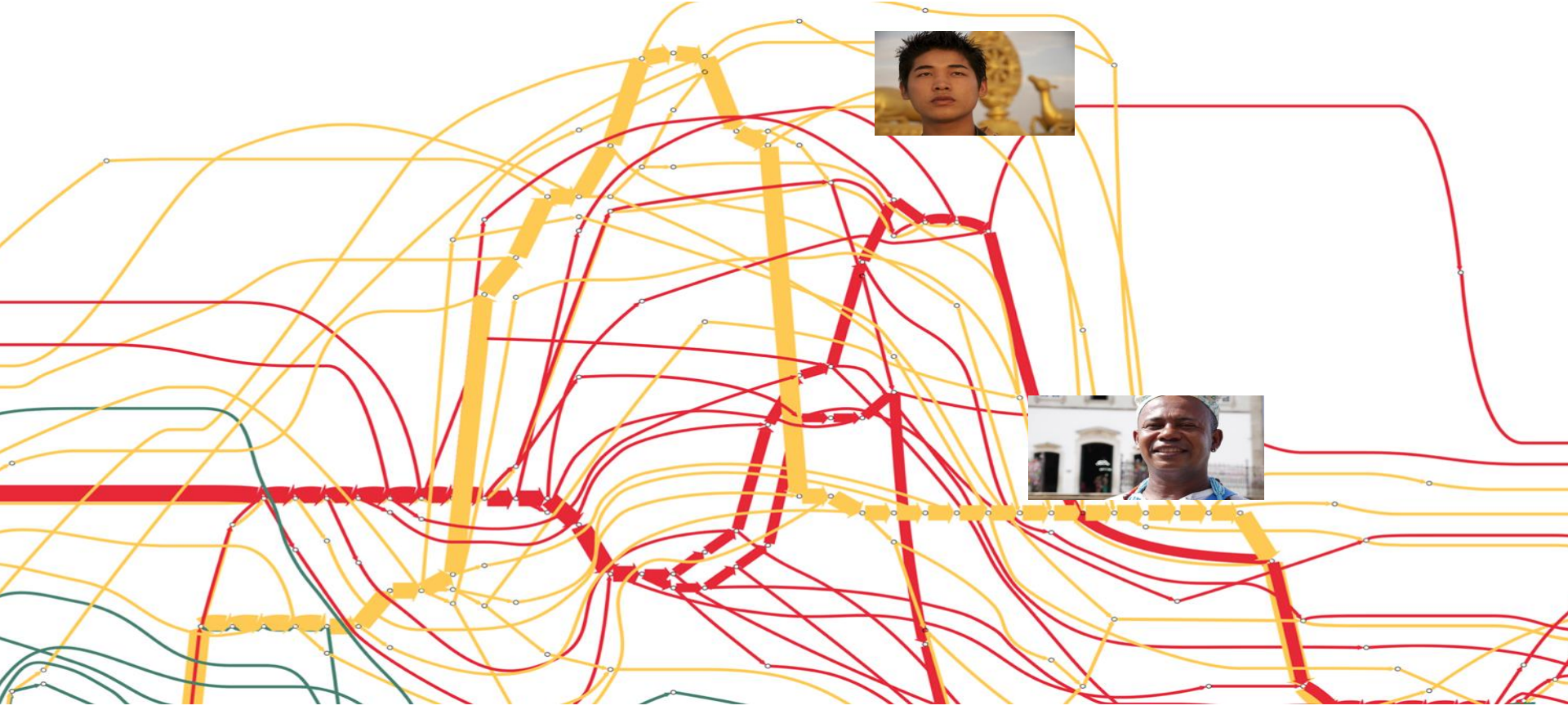
CREDIT: Kiran Garimella and Benedict Paten

Merging diverse genomes into one mathematical map

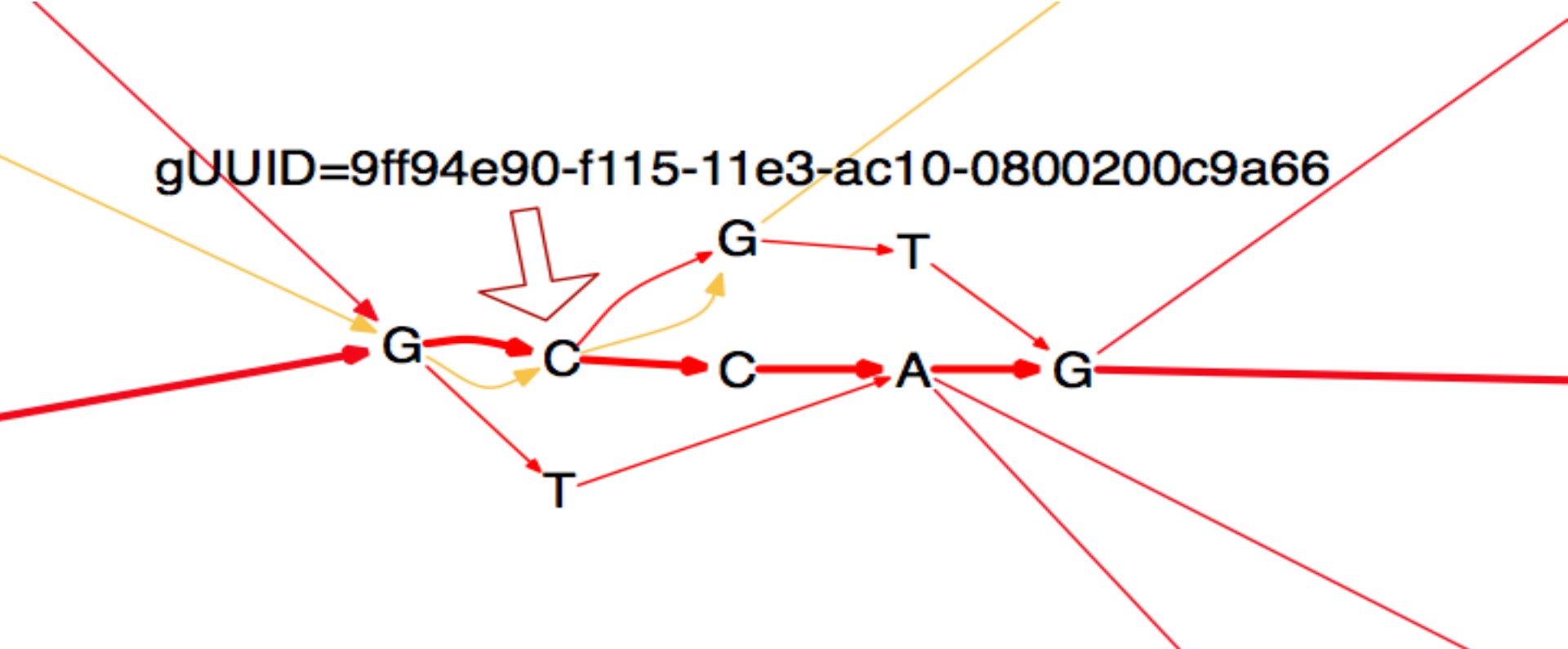


The major histocompatibility complex: Kiran Garimella and Benedict Paten

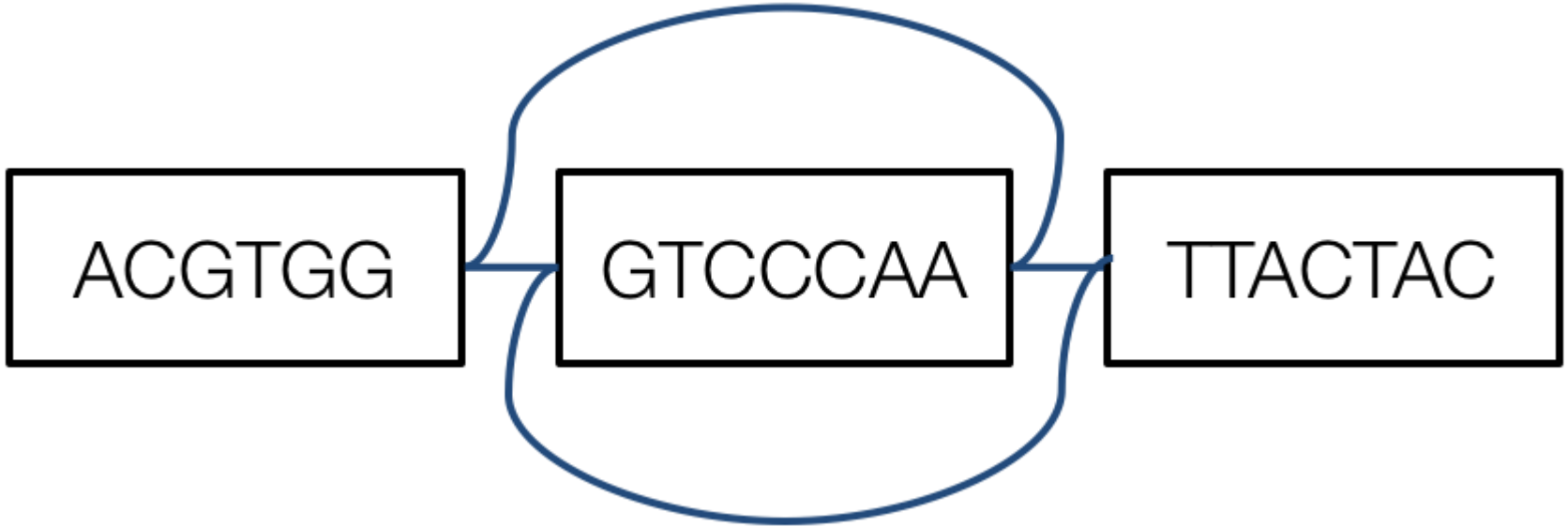
Zooming in, you start to see structure of local genetic variants



At base level, we assign unique identifiers to genetic variants to enable precision



Variation Graphs – The Essentials

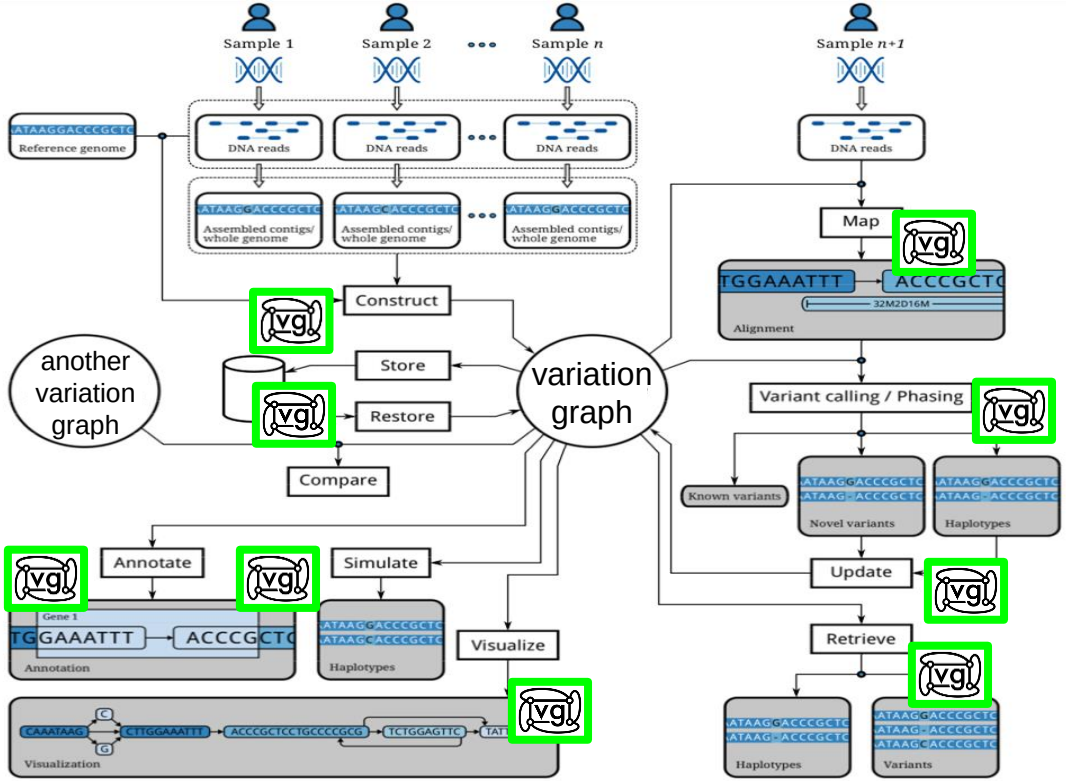


Joins can connect either side of a sequence (bidirected edges)

Walks encode DNA strings, with side of entry determining strand

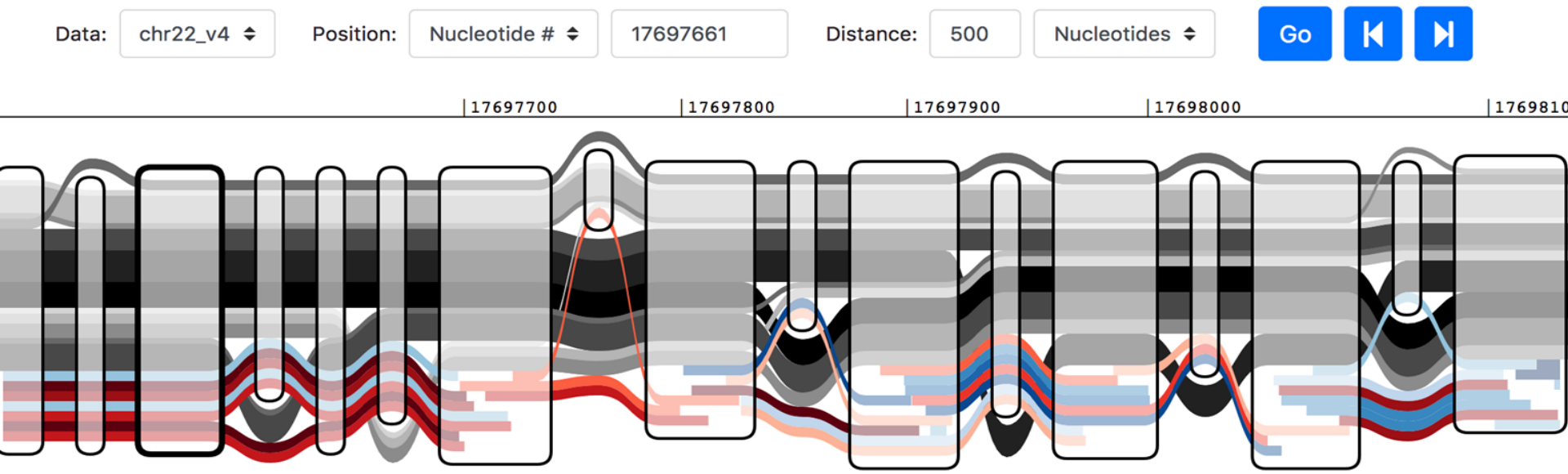
The VG group is building a software ecosystem for pangenomics

- Addresses all essential operations on genome graphs



https://github.com/vgteam/vg
doi.org/10.1101/234856

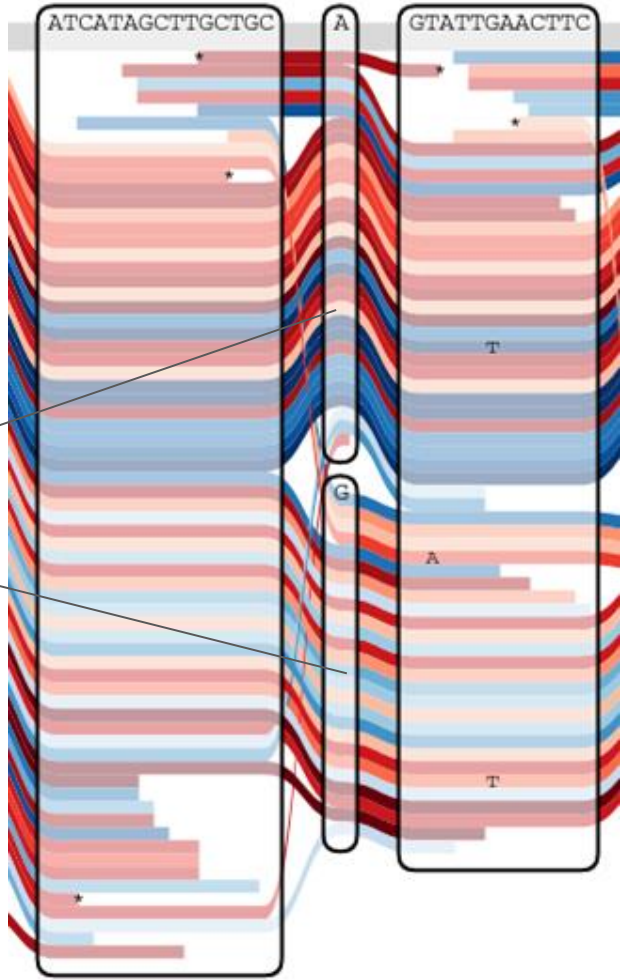
The first human genome variation map combines information from 1000 human genomes



View of genomes (gray to black) in an actual genome map, and DNA sequencing reads (colored worms) from a newly sequenced individual mapped to it

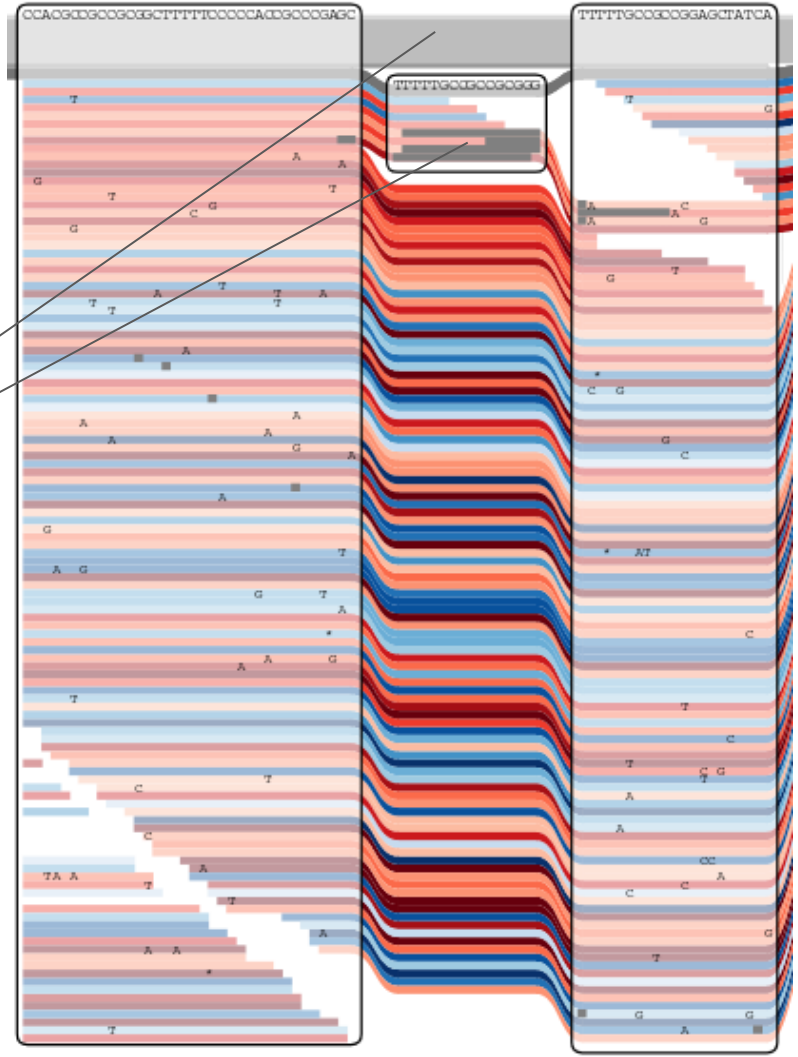
Genome Graph Models Naturally Represent All Variant Types

Substitution



Genome Graph Models Naturally Represent All Variant Types

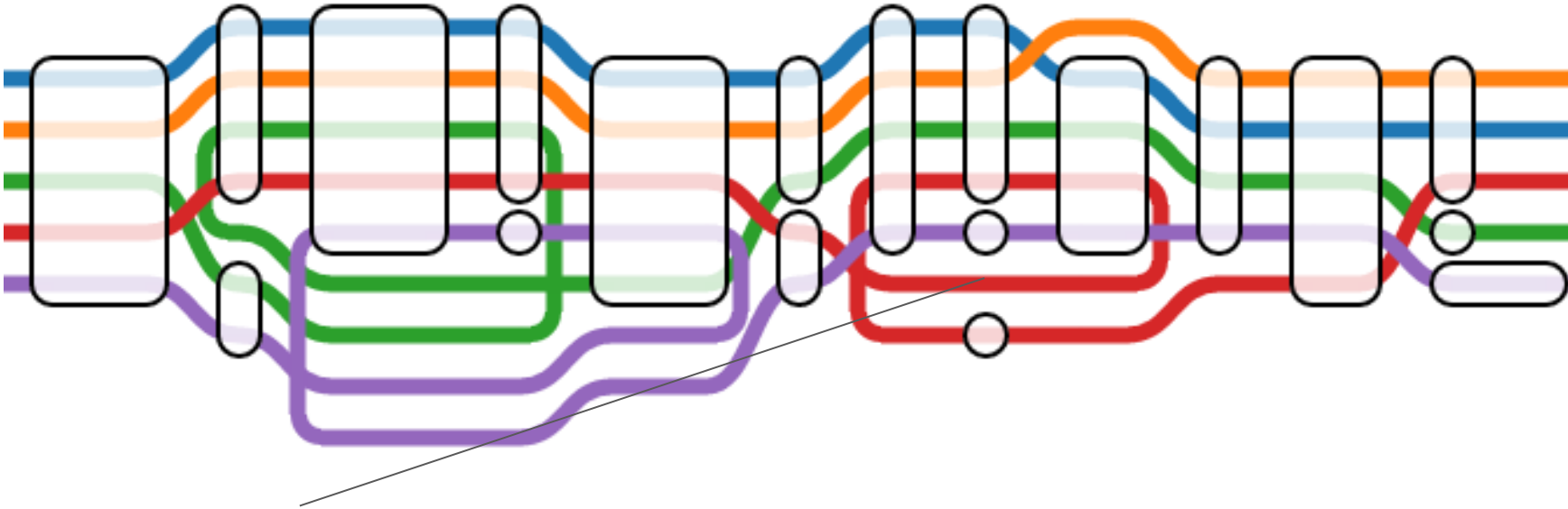
Insertion or deletion



Genome Graph Models Naturally Represent All Variant Types

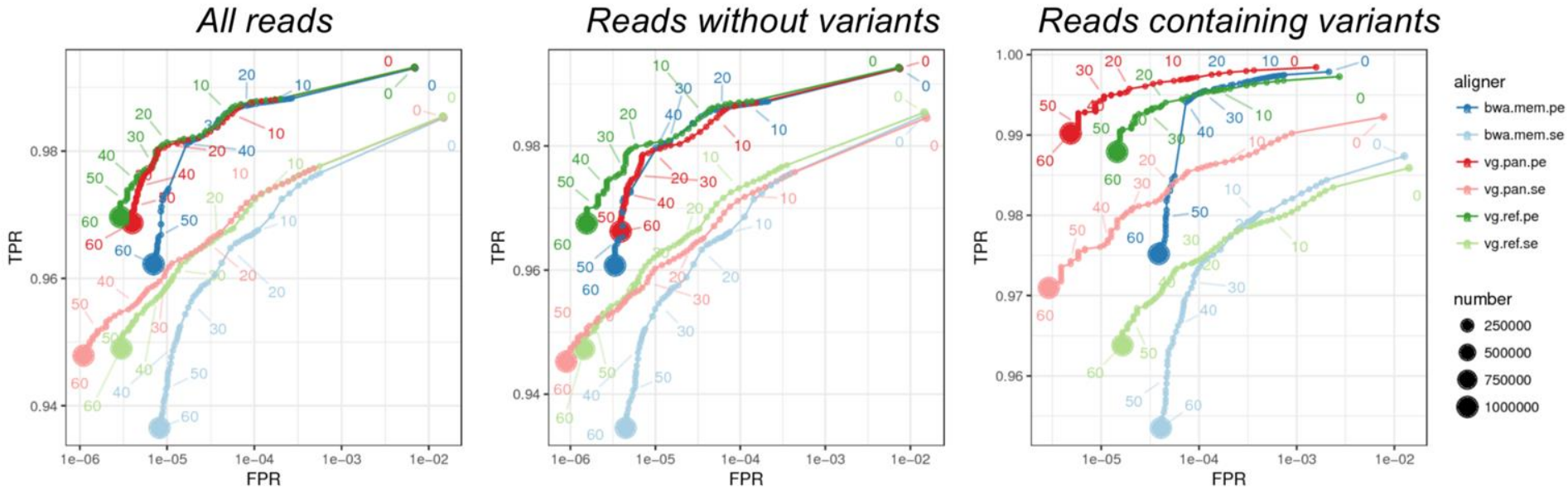
Duplication (top path traverses same nodes multiple times)

Genome Graph Models Naturally Represent All Variant Types



Inversion (red path traverses reverse complement)

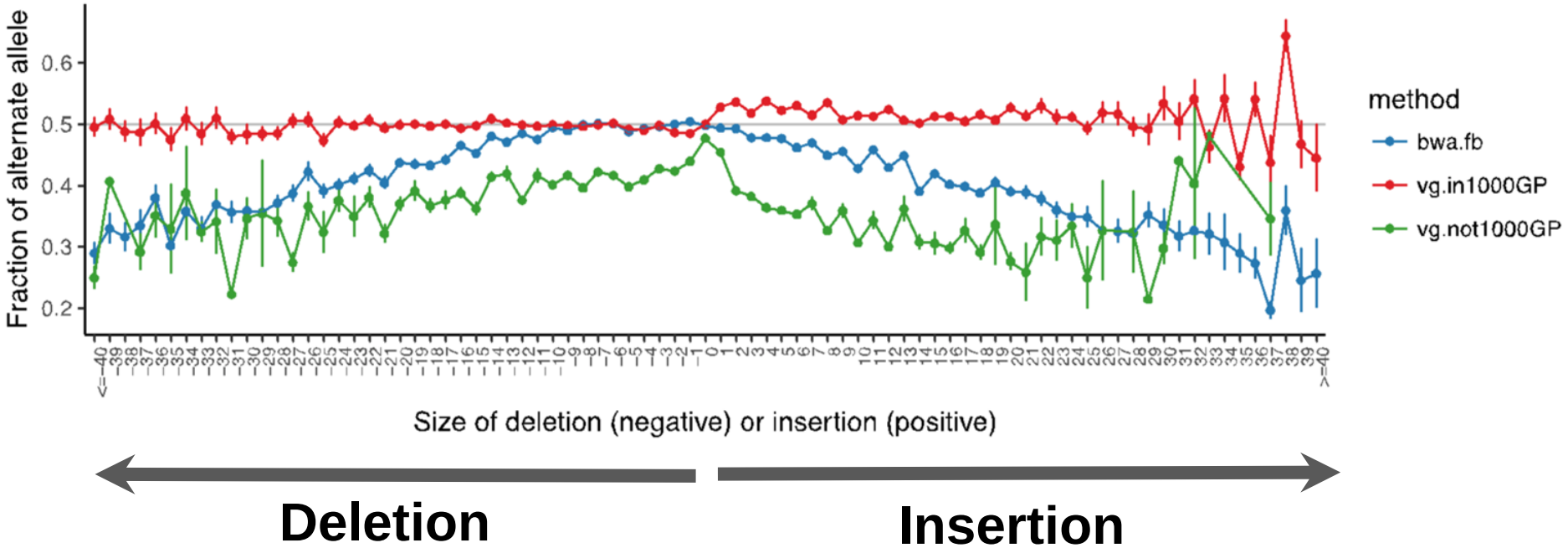
Human Read Mapping with VG



- Simulation study to GRCh38 / Graph using 1000 Genomes (80 Million Variants)
- 10 million read pairs (2x150mers)
- ROC stratified by MAPQ
- Reads sampled from Ashkenazi Jewish sample not in 1000 Genomes

Garrison et al. bioxriv: doi.org/10.1101/234856

Human Read Mapping with VG - Indel Allele Balance



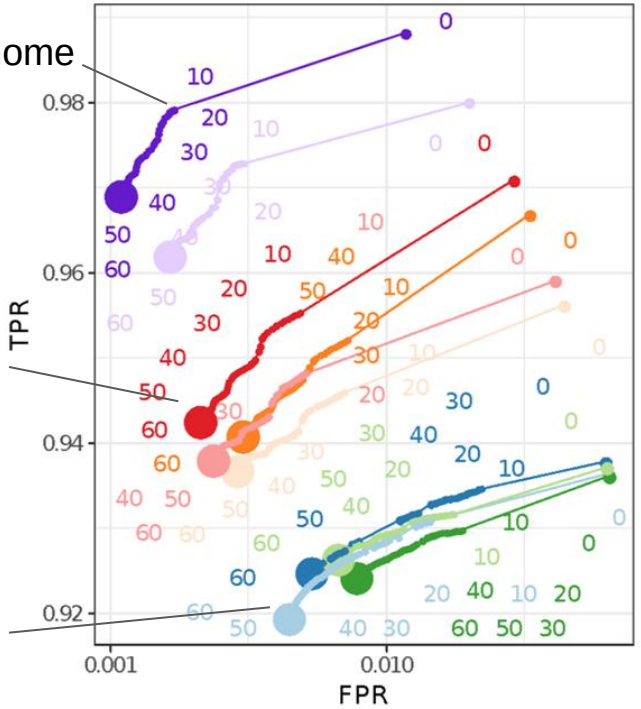
Garrison et al. bioxriv: doi.org/10.1101/234856

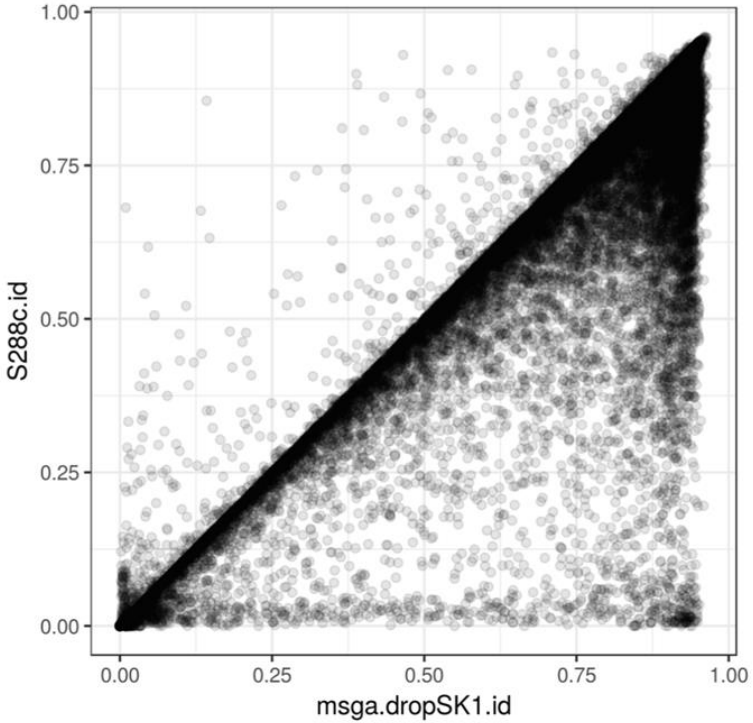
Yeast Mapping with VG - A More Polymorphic Example

Sample Genome

Pan genome

Reference genome





Garrison et al. bioxriv: doi.org/10.1101/234856

VG - Take Homes

- VG is practical for mapping human genome scale samples against graph with 80 Million point variants
- First tool to work with arbitrary graphs (cycles, copy number variants are possible)
- Provides interchange formats and many, many utilities

THANKS!

UC Santa Cruz

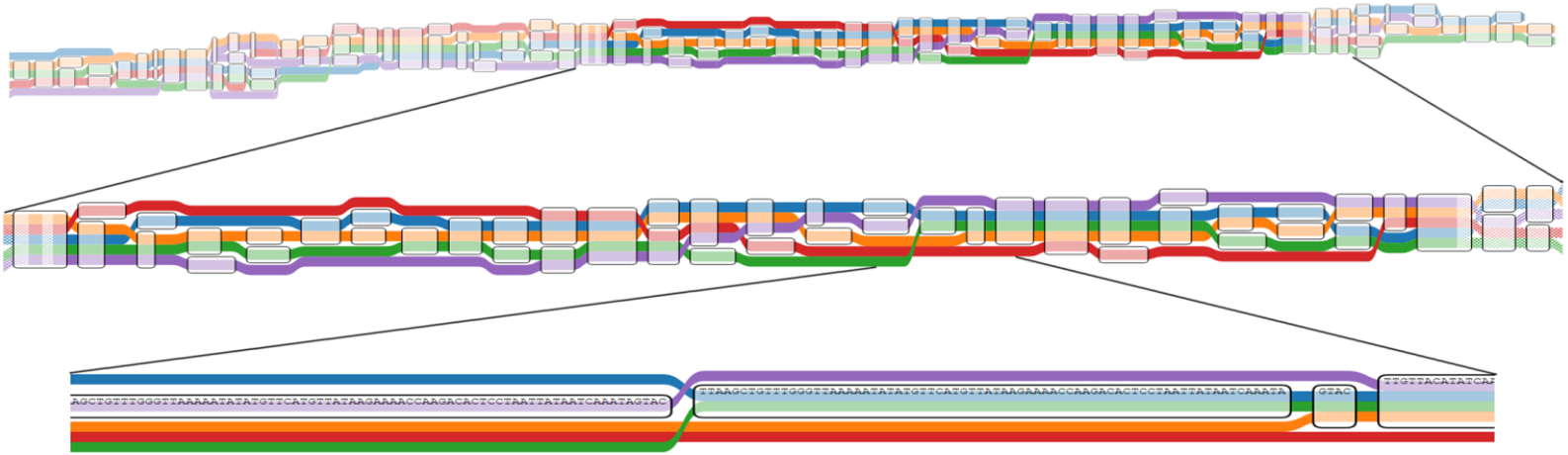
Adam Novak Wolfgang Beyer
Glenn Hickey Karen Miga
Yohei Rosen Jouni Siren
Jordan Eizenga Charles Markello
David Haussler Xian Chang
Yatish Turakhia

The Rest of Team VG
Erik Garrison
Richard Durbin
Eric Dawson
Mike Lin
(& many more)

GA4GH collaborators
Andres Kahles Heng Li
Ben Murray Stephen Keenan
Goran Rakocevic Gil McVean
Alex Dilthey (& many more)

Simons Foundation

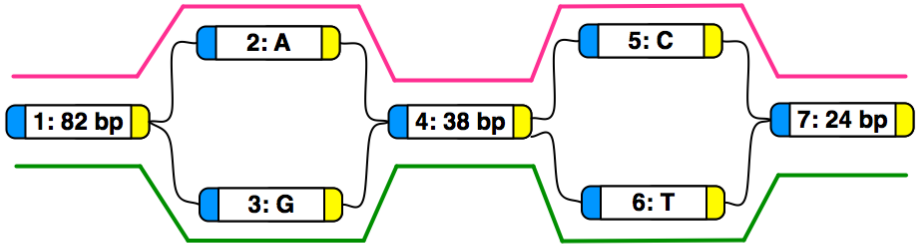
Summary



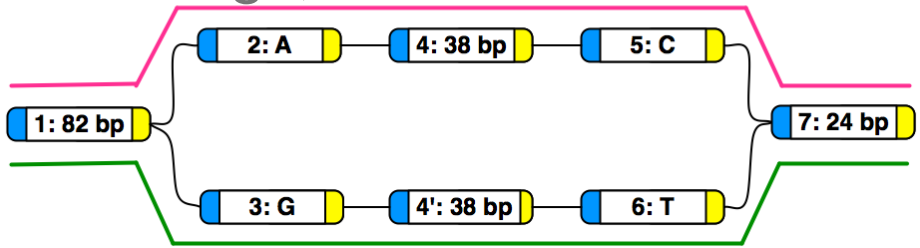
- Mapping is central to genomics, and reference genomes are perhaps the most important data structure in genomics
- With vg we can generalize reference genomes to reference genome graphs, and practically map to a population cohort instead, alleviating bias
- It’s not about replacing the reference with a graph, but with a population cohort

Embedding Haplotypes

- Genome graphs do not encode linkage



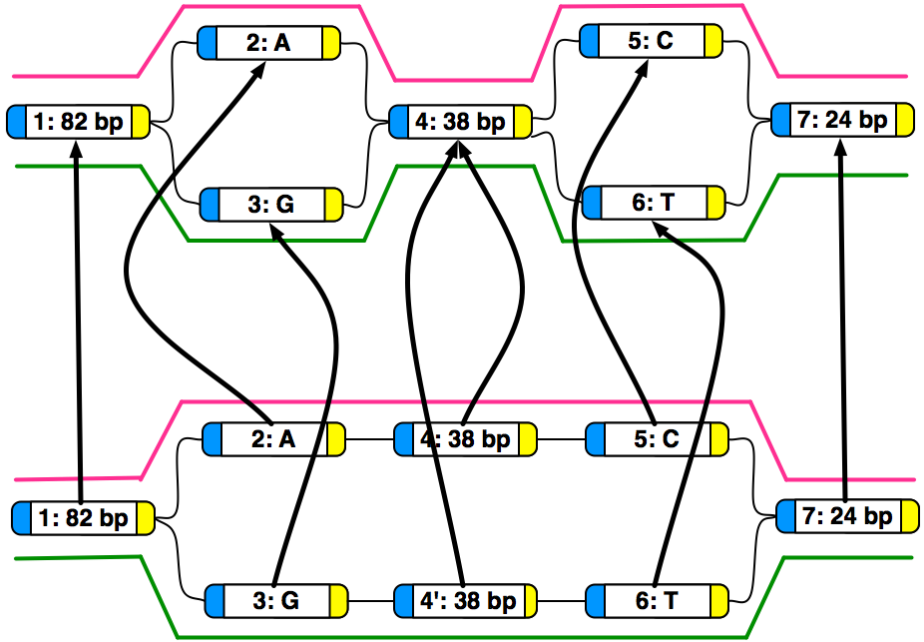
- To restrict linkage, natural solution is to duplicate paths:



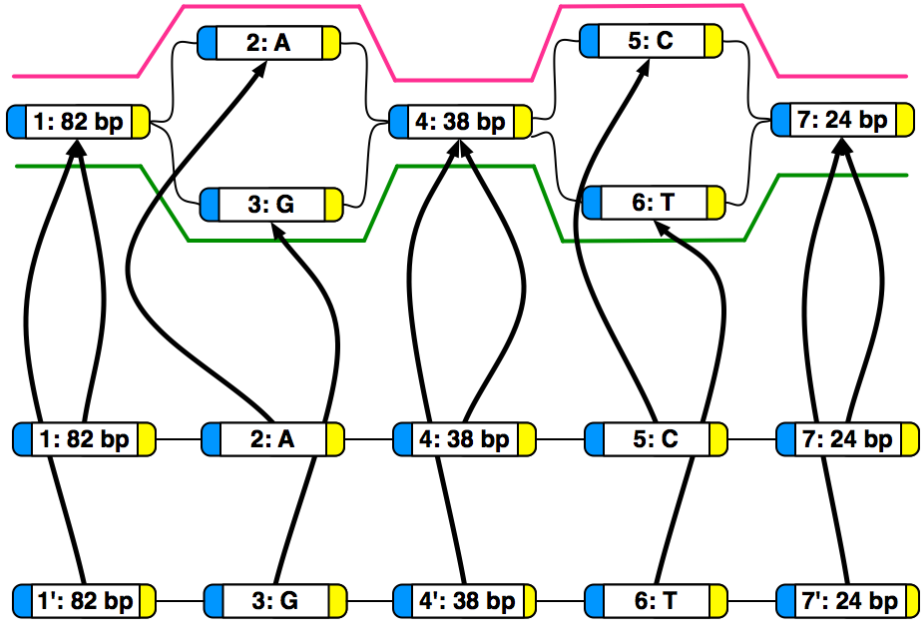
- But duplication creates mapping ambiguity

Embedding Haplotypes

- But note, there is a natural homomorphism (projection):



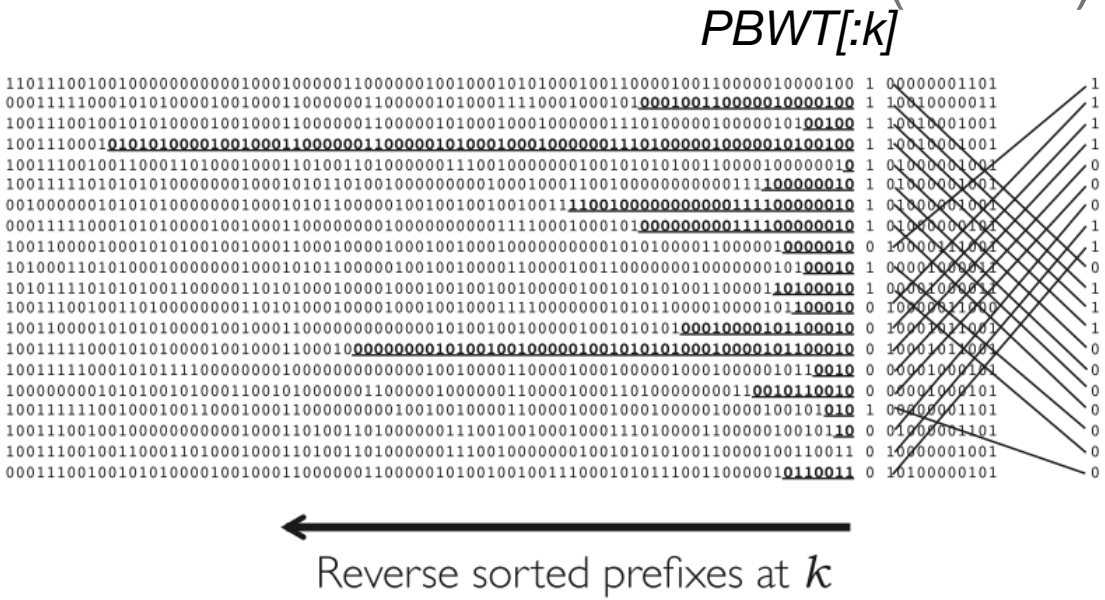
Embedding Haplotypes



- Instead maintain projection from haplotypes to graph:

Embedding Haplotypes

- The Positional Burrows Wheeler Transform (PBWT)

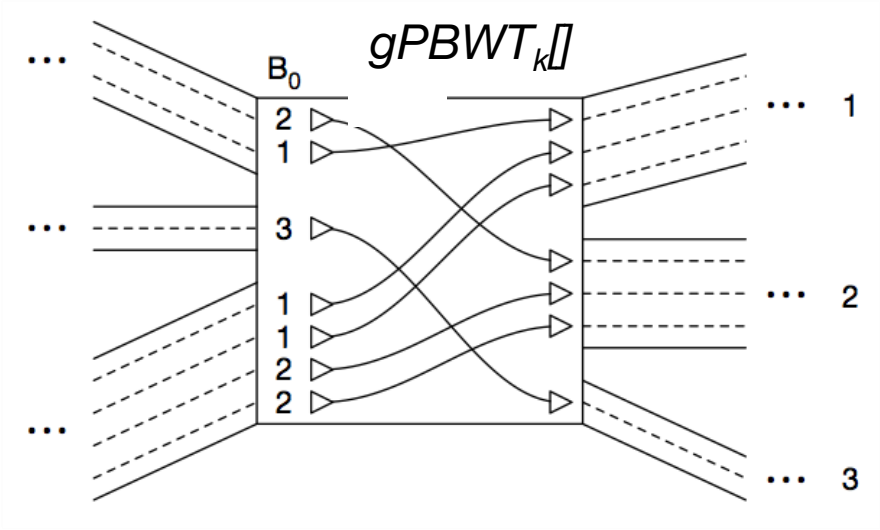


- Reversible, compressible, enables efficient indexed queries

Figure borrowed from *"Richard Durbin, Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT), Bioinformatics, 2014"*

Embedding Haplotypes

- The Graph Positional Burrows Wheeler Transform (gPBWT)



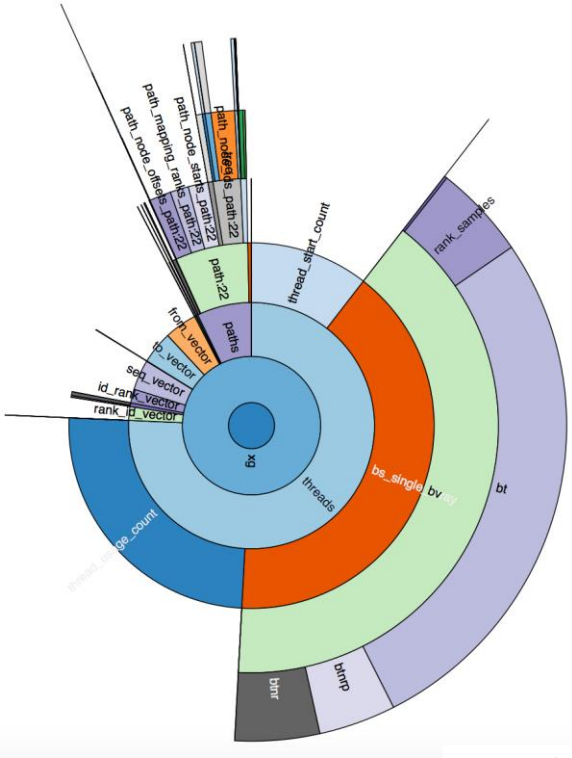
- Reversible, compressible, enables efficient indexed queries

From “Novak et al, A Graph Extension of the Positional Burrows-Wheeler Transform and its Applications (PBWT), WABI 2016”

gPBWT Performance

- Experiment:
  - chr22
  - 50,818,468 bp
  - 5004 Haplotypes

- Result:
  - 356 MB gPBWT + vg graph
  - **0.011 bits per base - 200x compression**
  - **~336 GB for whole genome w/80 million point variants @ 100,000 diploid genomes**



Novak *et al. Algorithms Mol Biol (2017) 12:18*
DOI 10.1186/s13015-017-0109-9

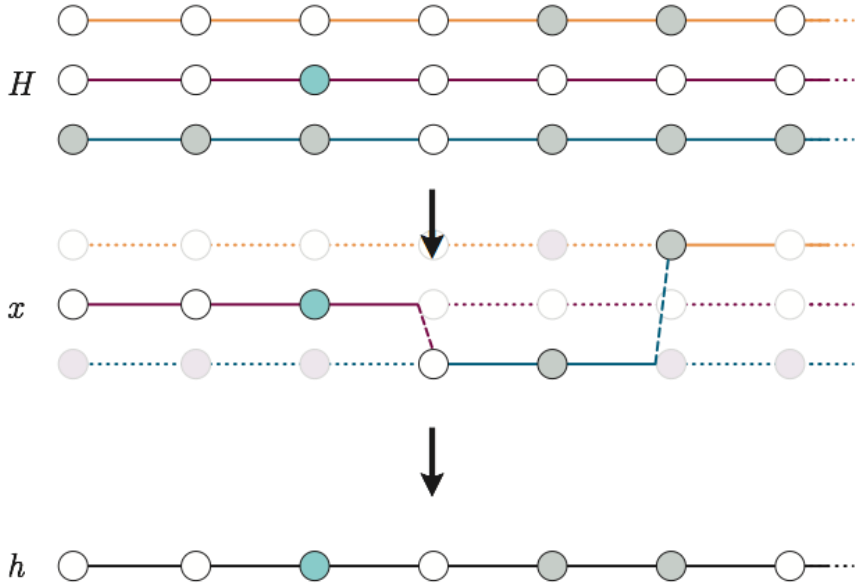
gPBWT → GBWT

- **Jouni Siren (now at UCSC!) showed gPBWT can be encoded as high cardinality alphabet BWT in which symbols in input strings represent nodes in VG graph**
- Call it Graph Burrows Wheeler Transform (GBWT)
- Implemented in VG:
  - Whole 1000 Genomes Graph construction on one machine in one day
  - Half space of gPBWT (14gb for entire index for 1000G)

Siren et al. https://arxiv.org/pdf/1805.03834.pdf

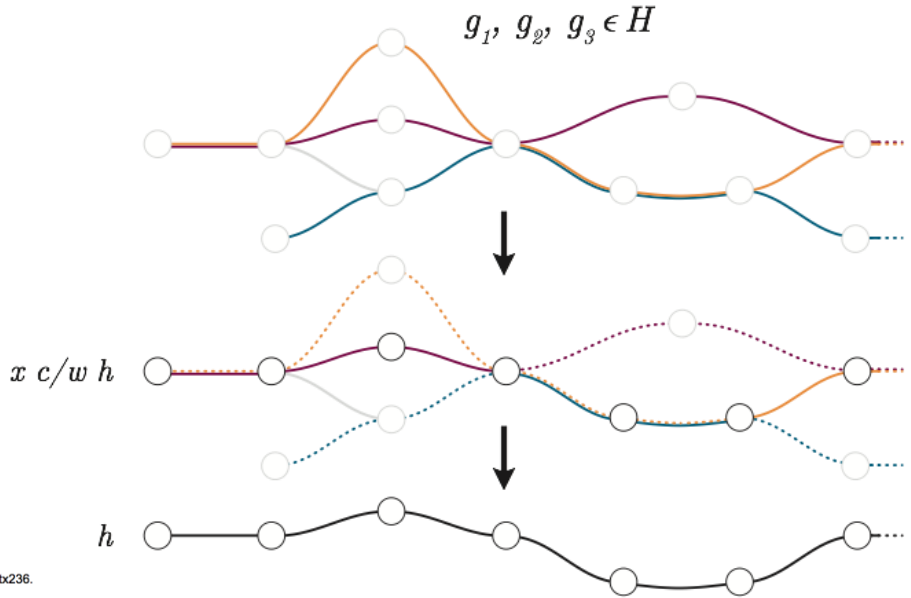
Haplotype Probabilities

- Li & Stephens: Efficiently compute P(h|H), where h is haplotype and H is population



Haplotype Probabilities

- Graph Li & Stephens: Efficiently compute P(x|H), where x is haplotype walk in a genome graph



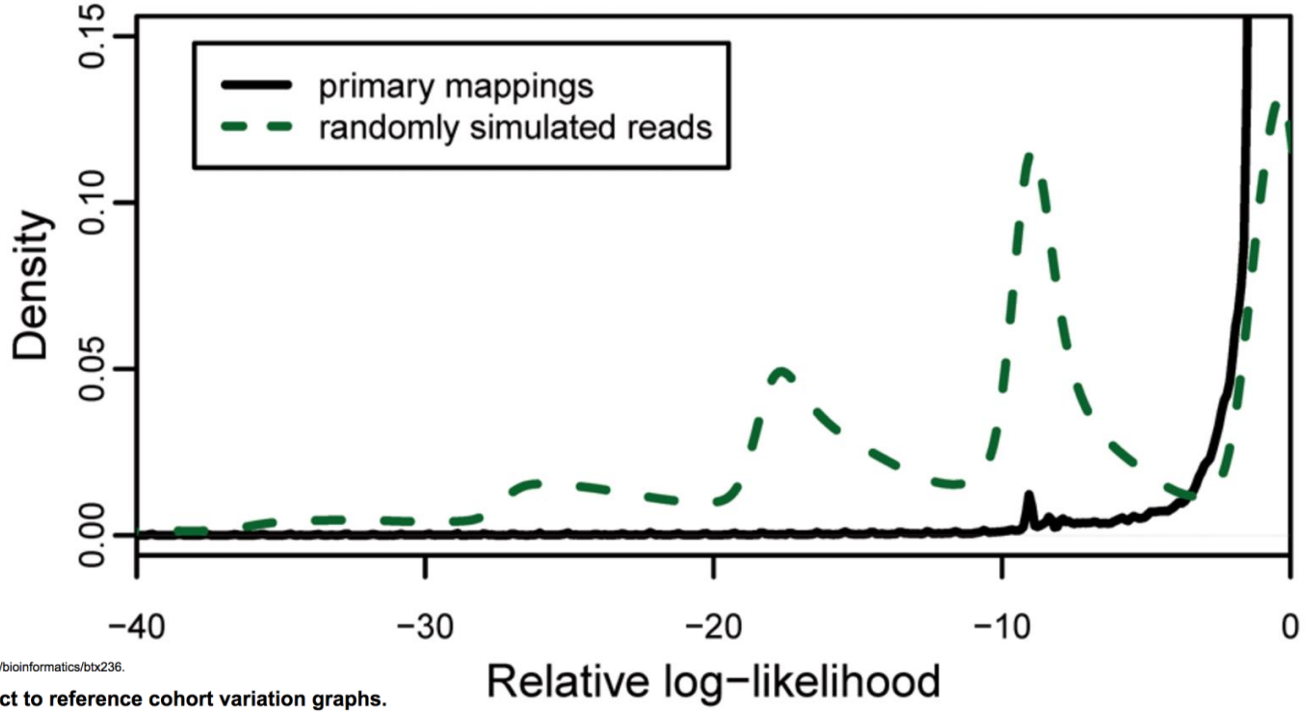
Bioinformatics. 2017 Jul 15;33(14):i118-i123. doi: 10.1093/bioinformatics/btx236.

Modelling haplotypes with respect to reference cohort variation graphs.

Rosen Y[1], Eizenga J[1], Paten B[1].

Haplotype Probabilities

- Applied to vg mapped reads:



Bioinformatics. 2017 Jul 15;33(14):i118-i123. doi: 10.1093/bioinformatics/btx236.

Modelling haplotypes with respect to reference cohort variation graphs.

Rosen Y[1], Eizenga J[1], Paten B[1].

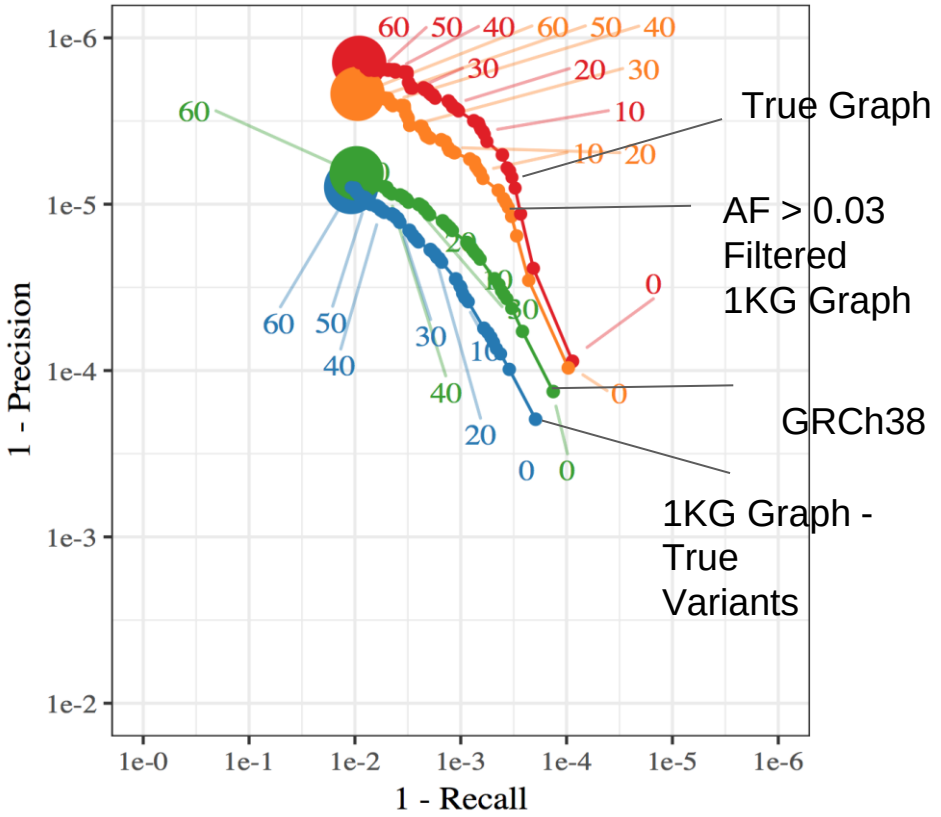
Richer Graphs: More Is Not Necessarily More

- Adding variation into a graph has both positive and negative effects on mapping
- From the HiSat folks:

FORGe: prioritizing variants for graph genomes

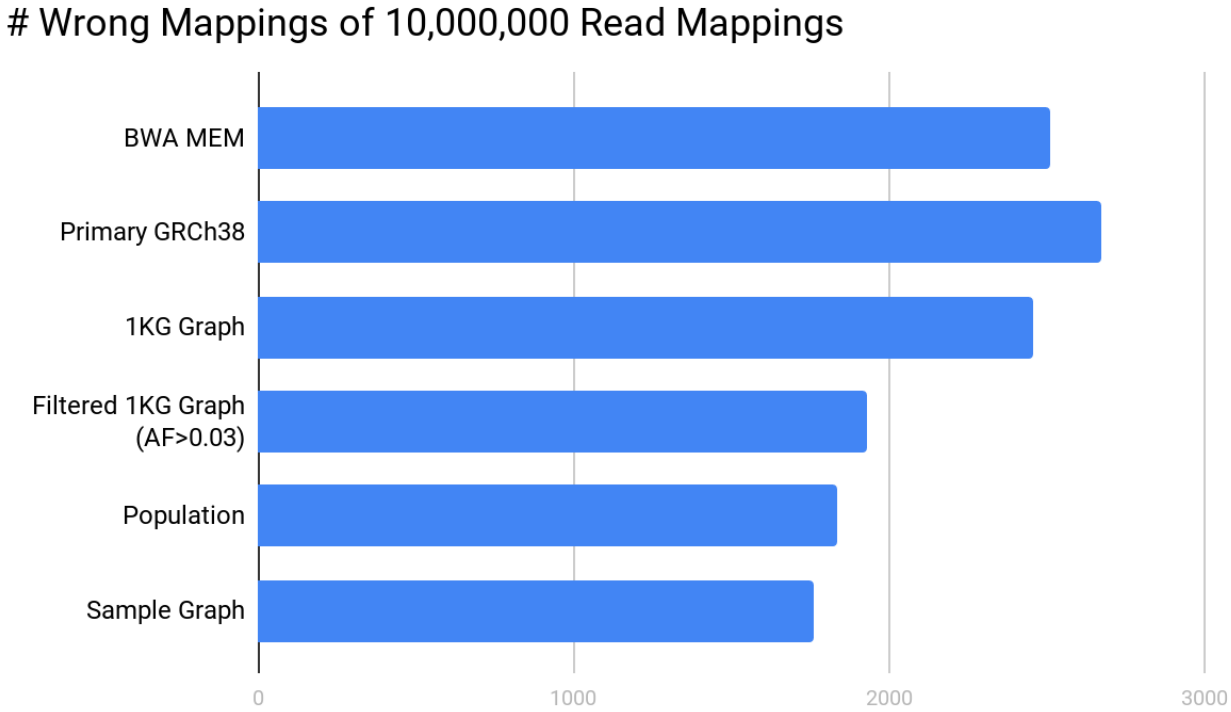
Jacob Pritt, Ben Langmead

doi: https://doi.org/10.1101/311720



Map to the population, not the graph

- **P(r | G) != P(r | H)**
- Accounting for haplotypes with all variants better than mapping to any graph
- ~30% fewer FP mappings relative to BWA



Workflow

Product	Version
Device	PromethION Alpha-Beta Flongle
Flow Cells	FLO-PRO002 FLO-FLG106
Kits	Ligation Sequencing Kit Circulomics SRE Puregene
Data analysis	Shasta MarginPolish HELEN minimap2

Shasta

MarginPolish

HELEN