Efficient de novo Assembly of Eleven Human Genomes Using PromethION Sequencing and a Novel Nanopore Toolkit

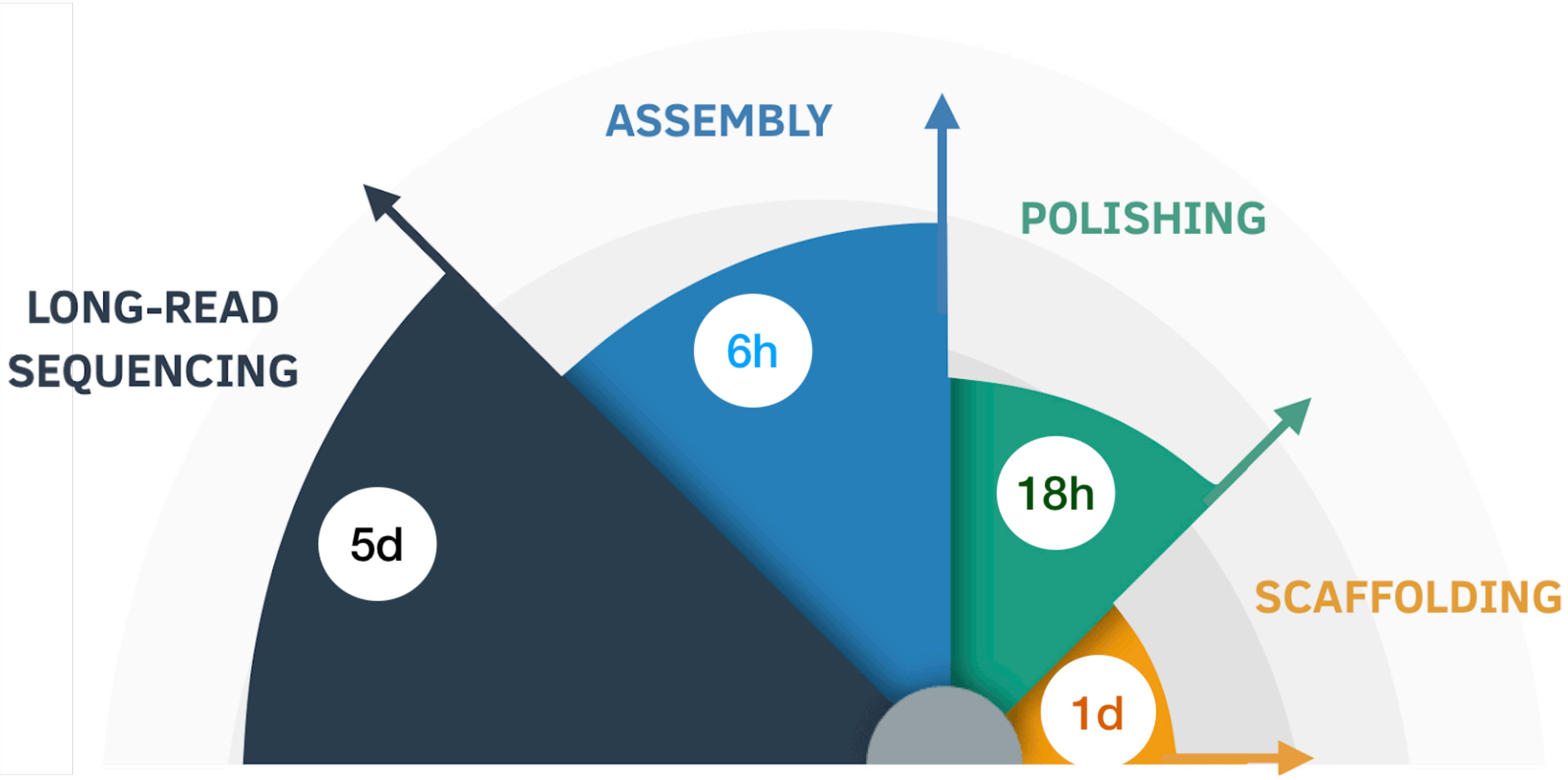
Miten Jain, Hugh Olsen

UC Santa Cruz

1 March 2020

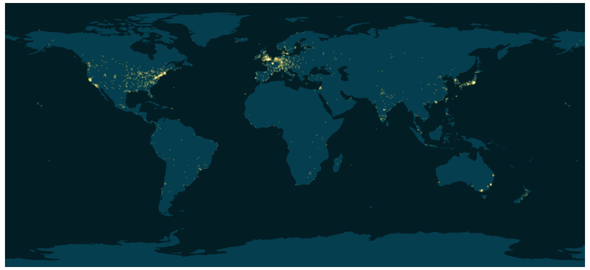
@mitenjain

MESSAGE: Reference-quality human genome in ~7 days using nanopore + HiC

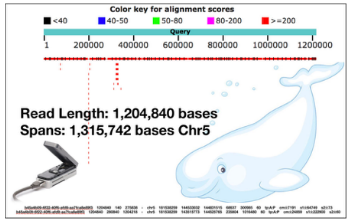


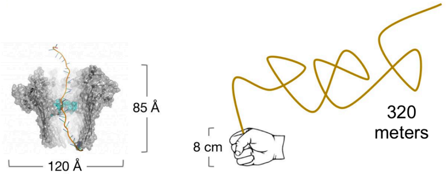
Genomics Using Nanopore Devices

1. Mobile



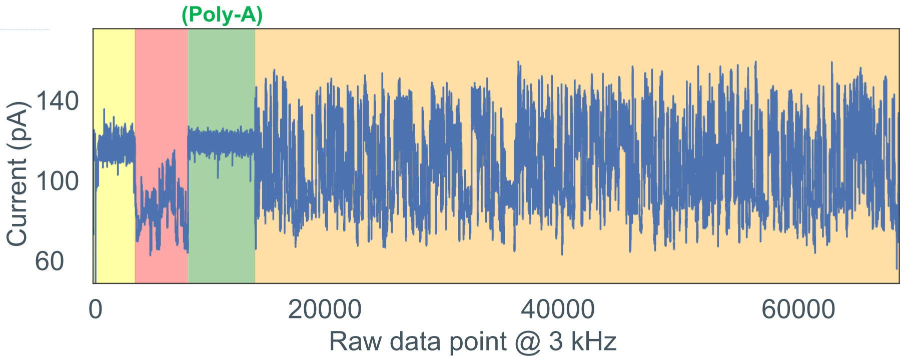
2. Long read lengths



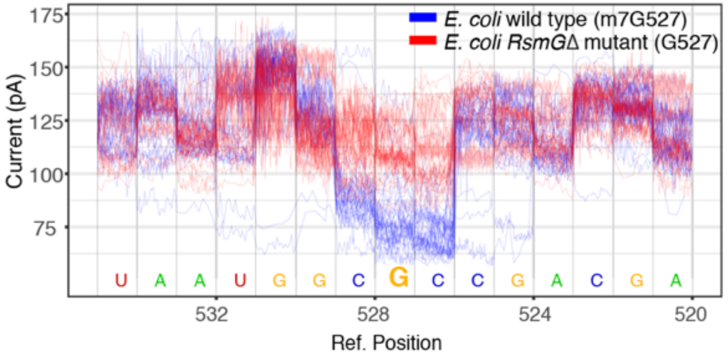


*Assuming 3.4 Å per bp, 1 Mbp = 3,400,000 Å = 40,000x height of the pore

3. Direct readouts of RNA



4. RNA/DNA base modification detection



What’s the bottleneck in capturing genome variation?

- Need for hundreds of high-quality reference genomes
- Sequencing cost
- Sequencing speed
- Scalable and cheaper informatics

Solution

- Nanopore 100kb+ sequencing

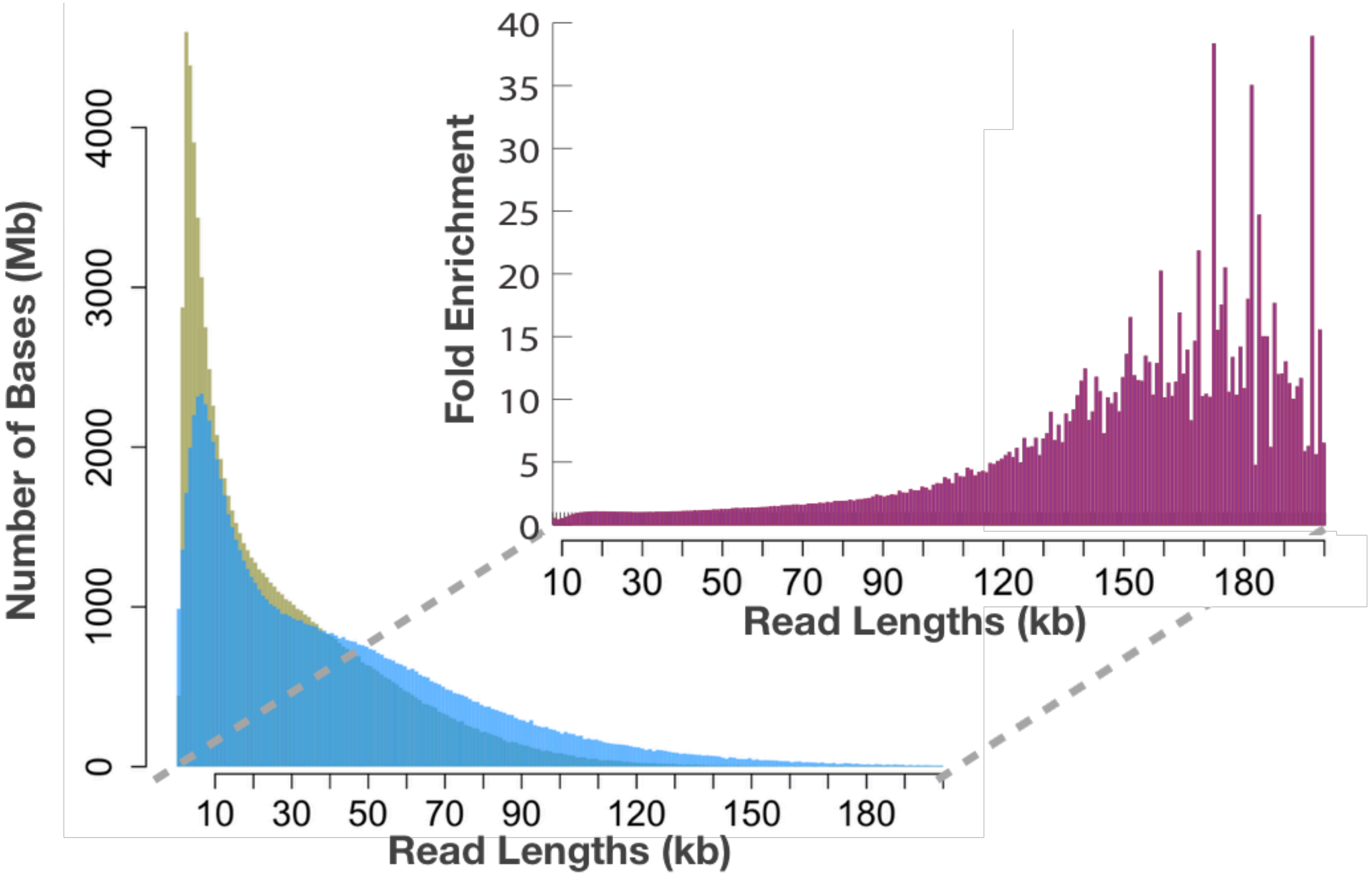
- Scalable assembly and polishing

  tools

Nanopore 100kb+ sequencing

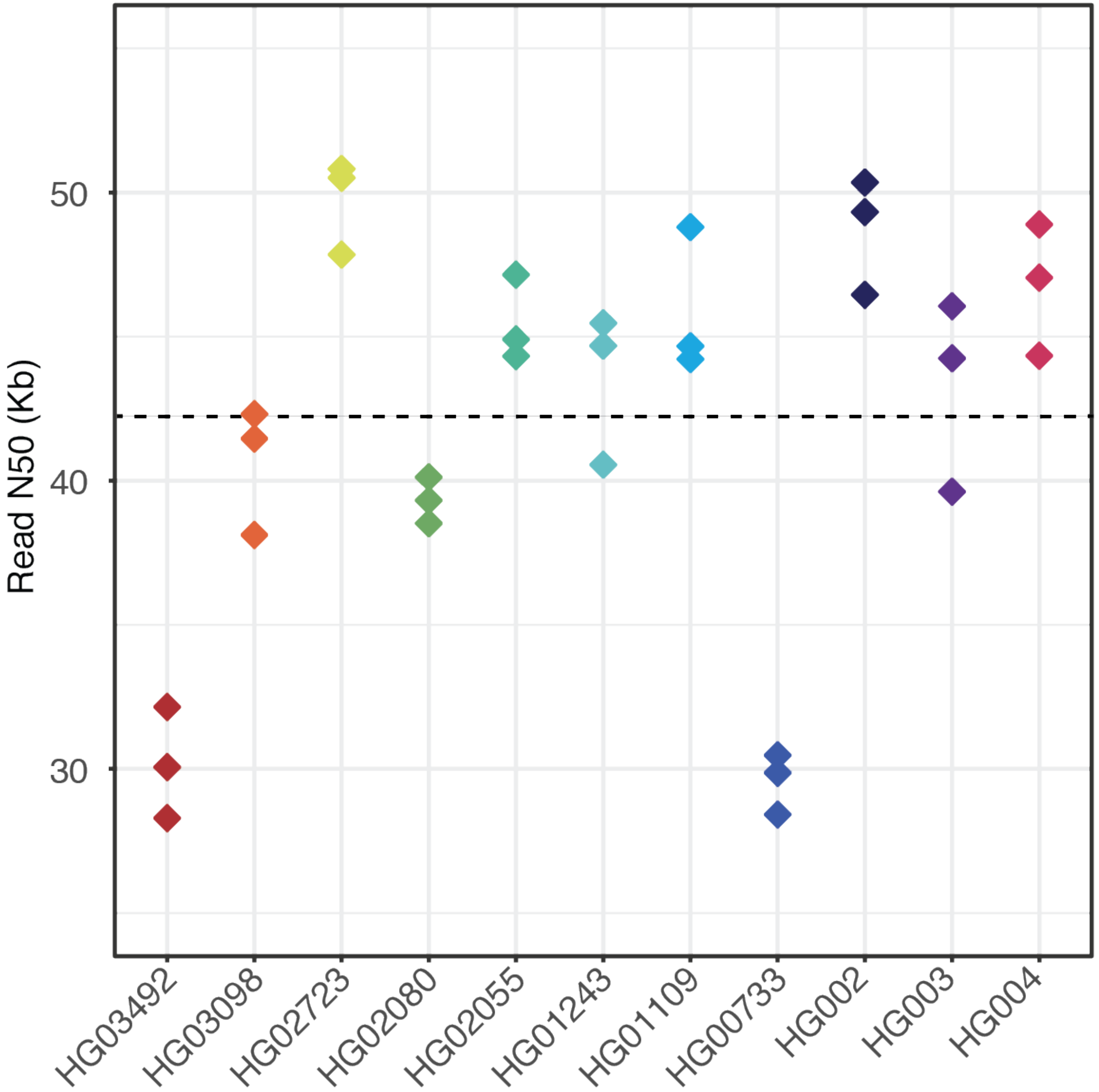
Data acquisition for 11 genomes in 9 days (>60x total coverage)

7x enrichment of reads >100kb using Circulomics SRE



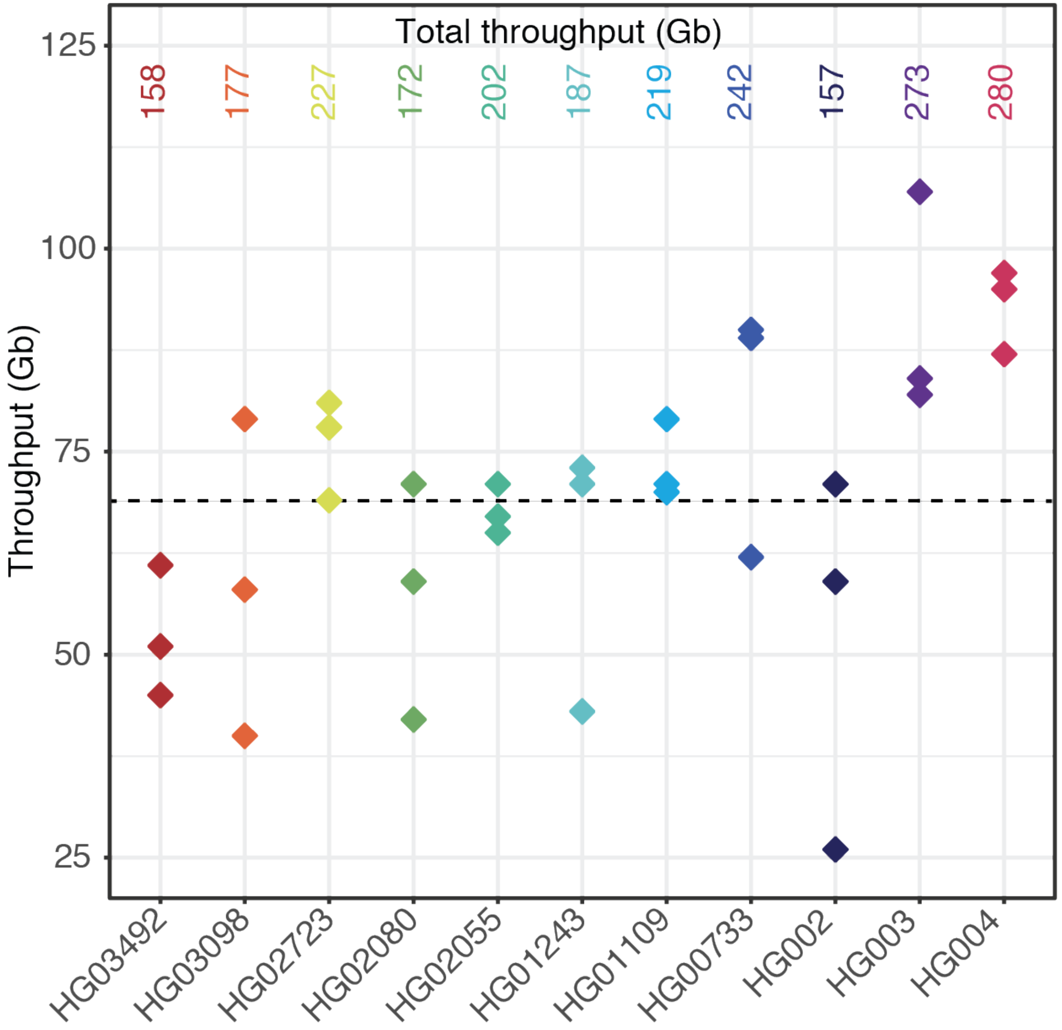
Short Read Eliminator Kit (https://www.circulomics.com)

Read N50 improvement is reproducible

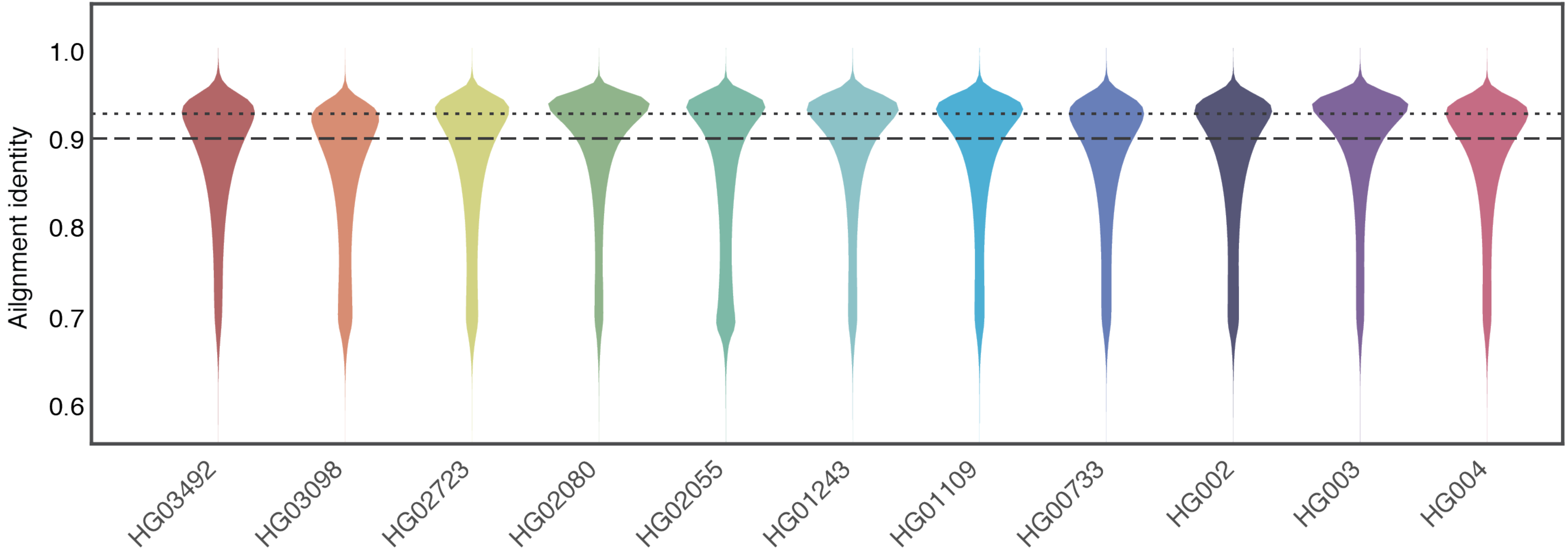


N50s: 42kb

PromethION sequencing throughput

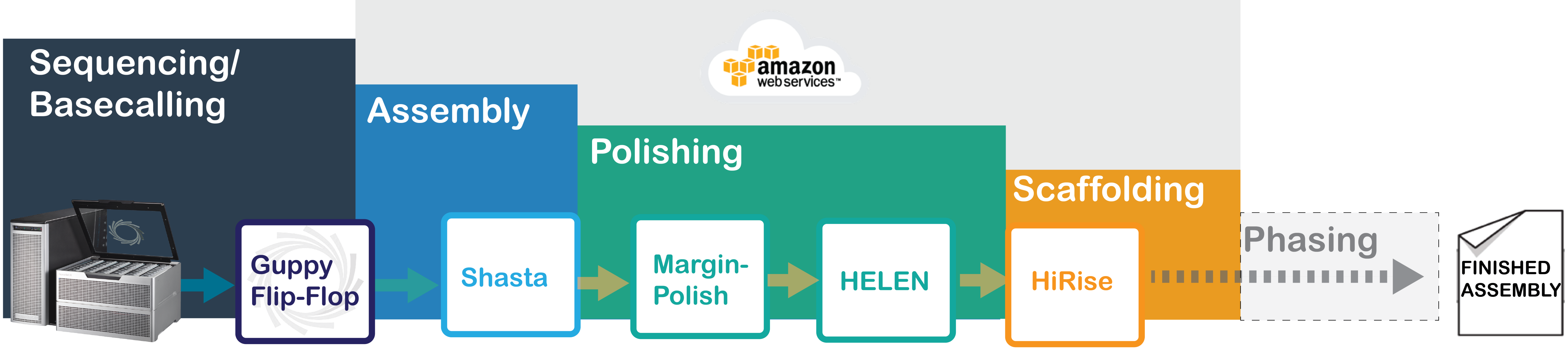


Median alignment identity is 90%



Scalable assembly and polishing tools

Pipeline



Sequencing/ Basecalling

Assembly

Polishing

Scaffolding

Phasing

Guppy Flip-Flop

Shasta

Margin-Polish

HELEN

HiRise

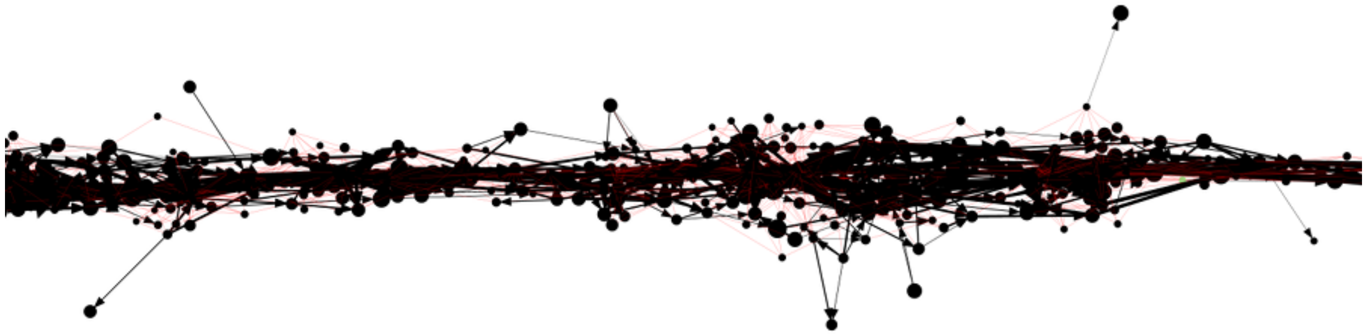
FINISHED ASSEMBLY

Shasta – a nanopore *de novo* long read assembler

CHAN ZUCKERBERG INITIATIVE

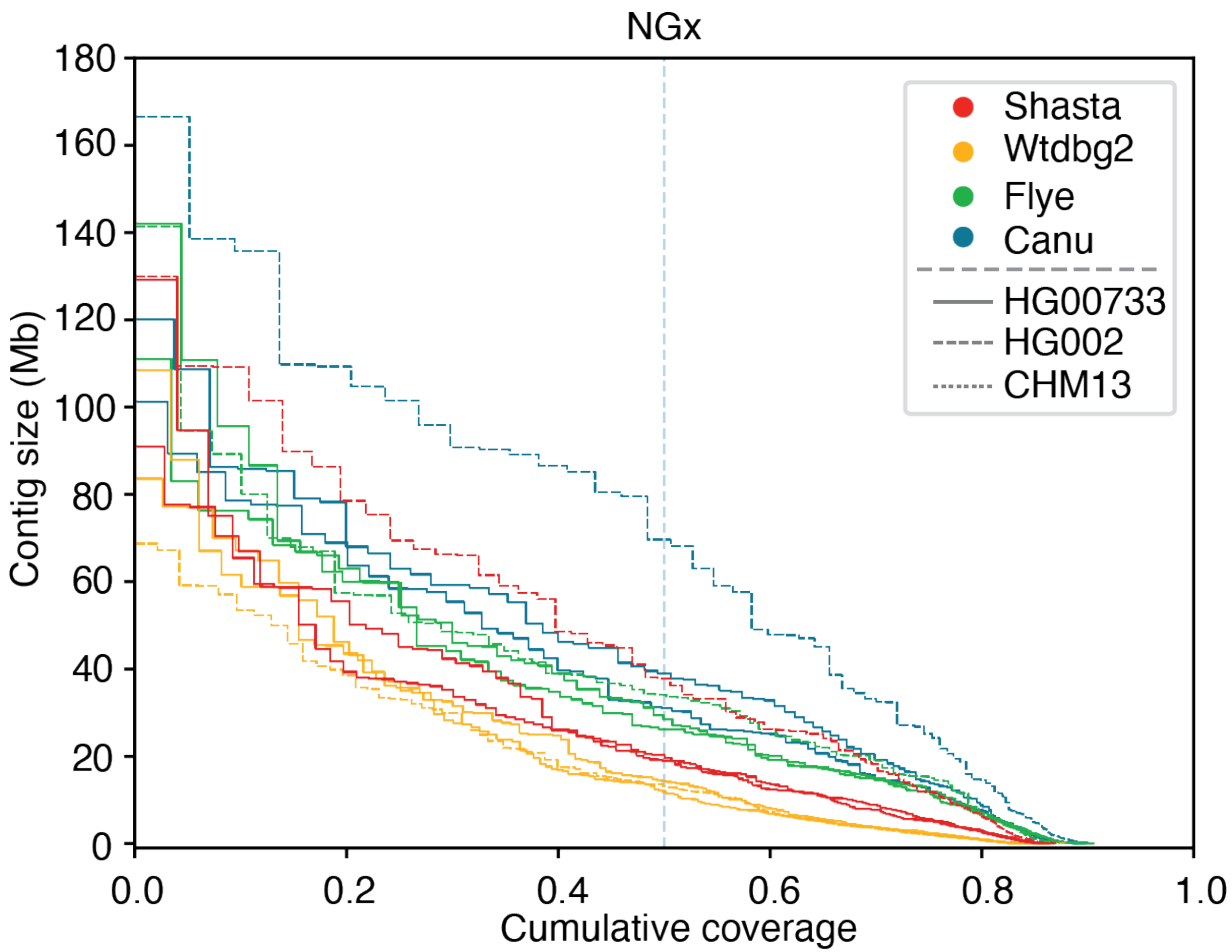
UC SANTA CRUZ

- Extremely fast (can assemble a human genome in < 6 hours on a single node)
- Uses run-length representation of read sequence
- Uses a memory-based marker representation for fast compute



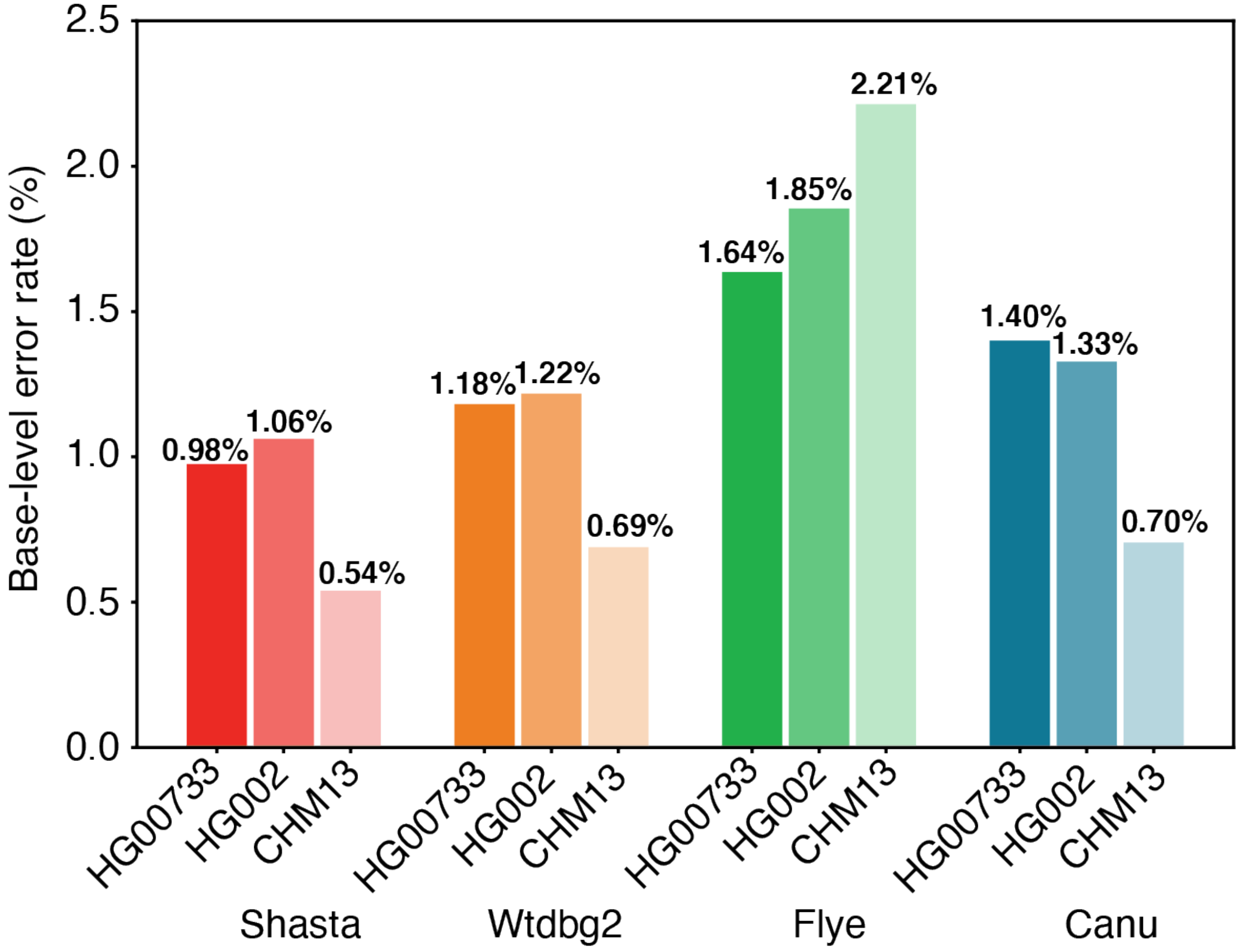
- https://github.com/chanzuckerberg/shasta

Shasta assemblies are reproducible, with comparable contig NG50

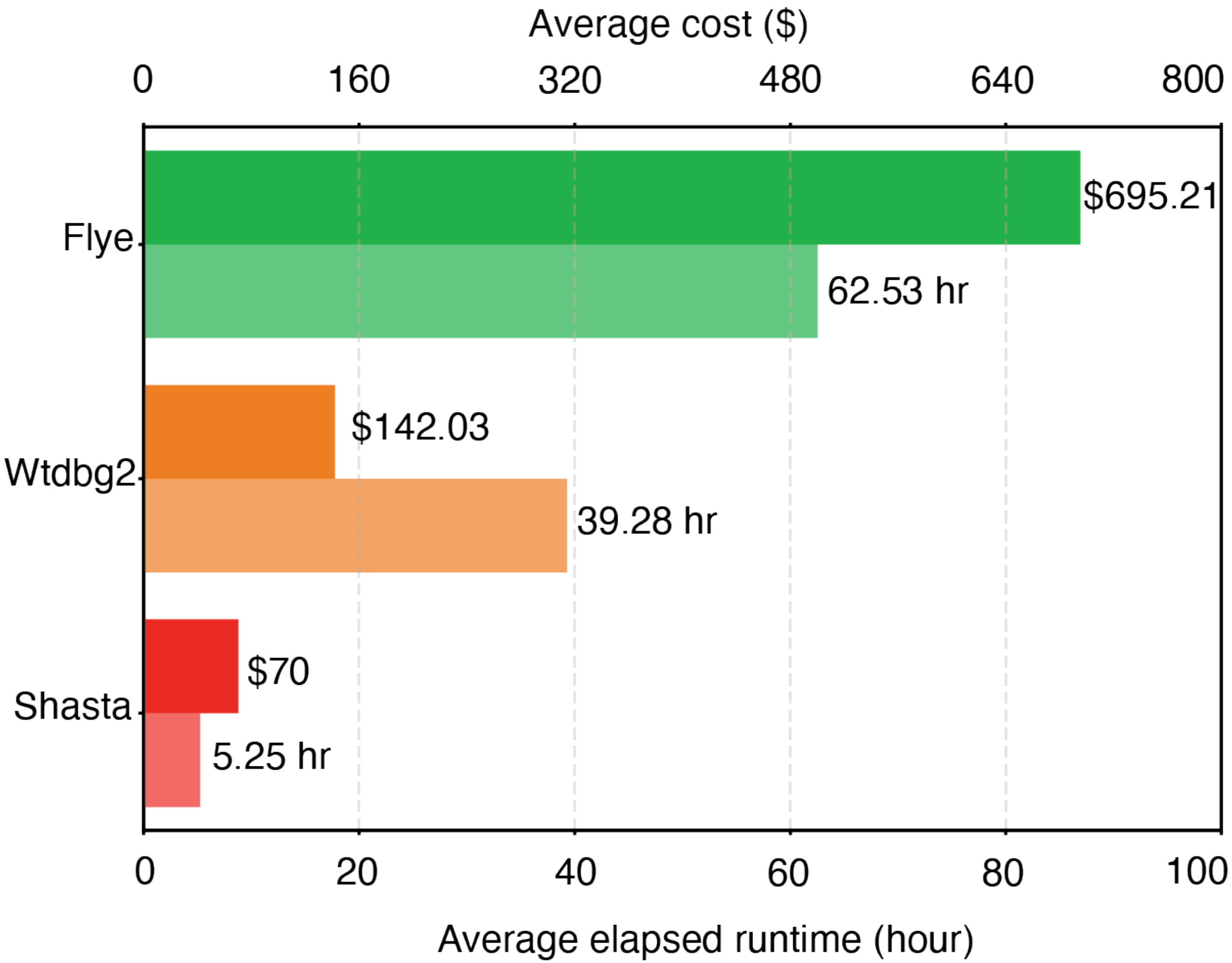


Median contig NG50 = 23 Mb

Shasta assemblies have higher accuracy

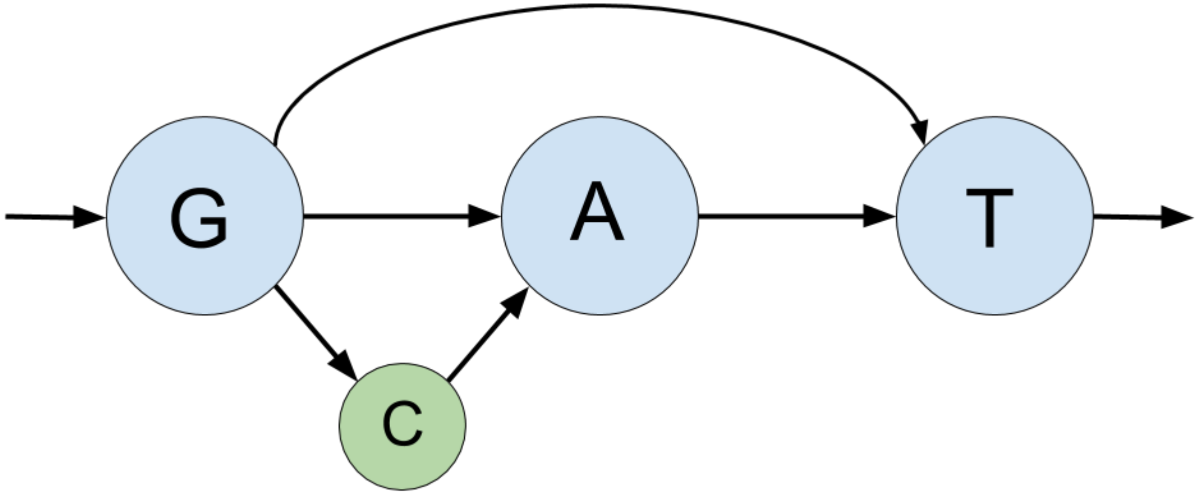


Assembly at a fraction of time and cost



Two-step polishing of assemblies

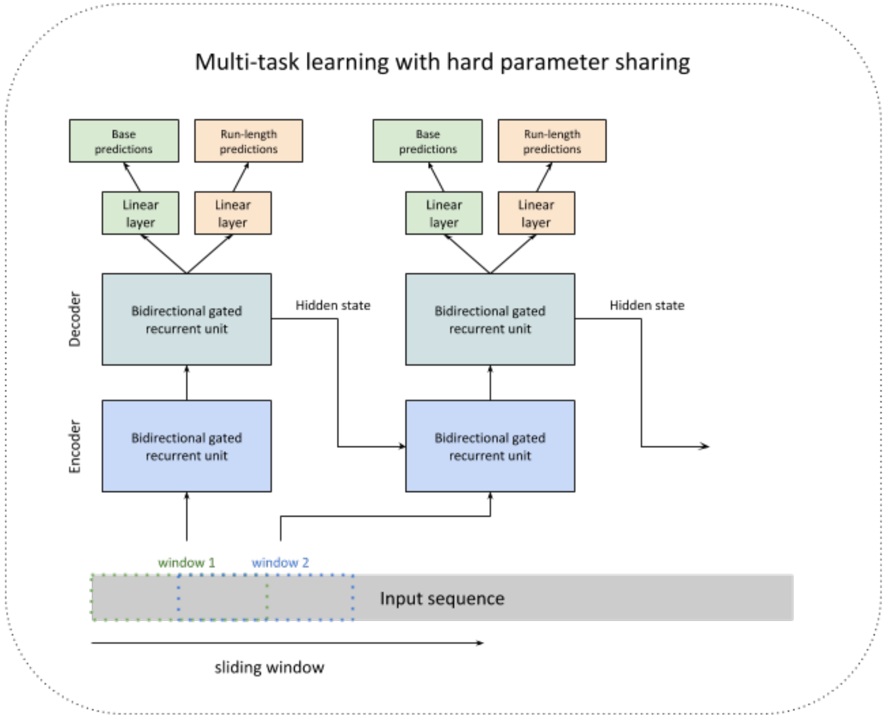
1. MarginPolish



A graph-based alignment polisher

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

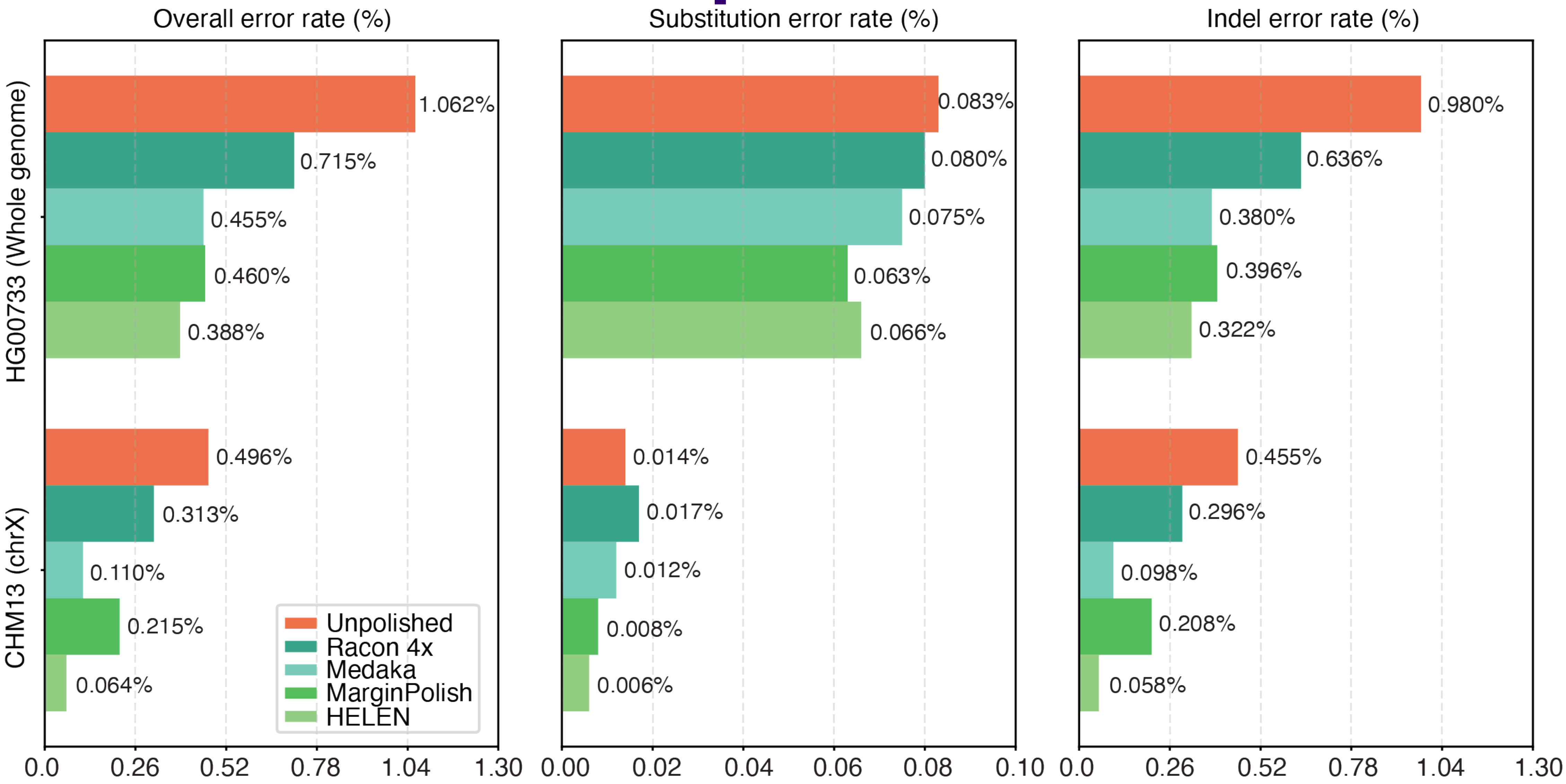
2. HELEN



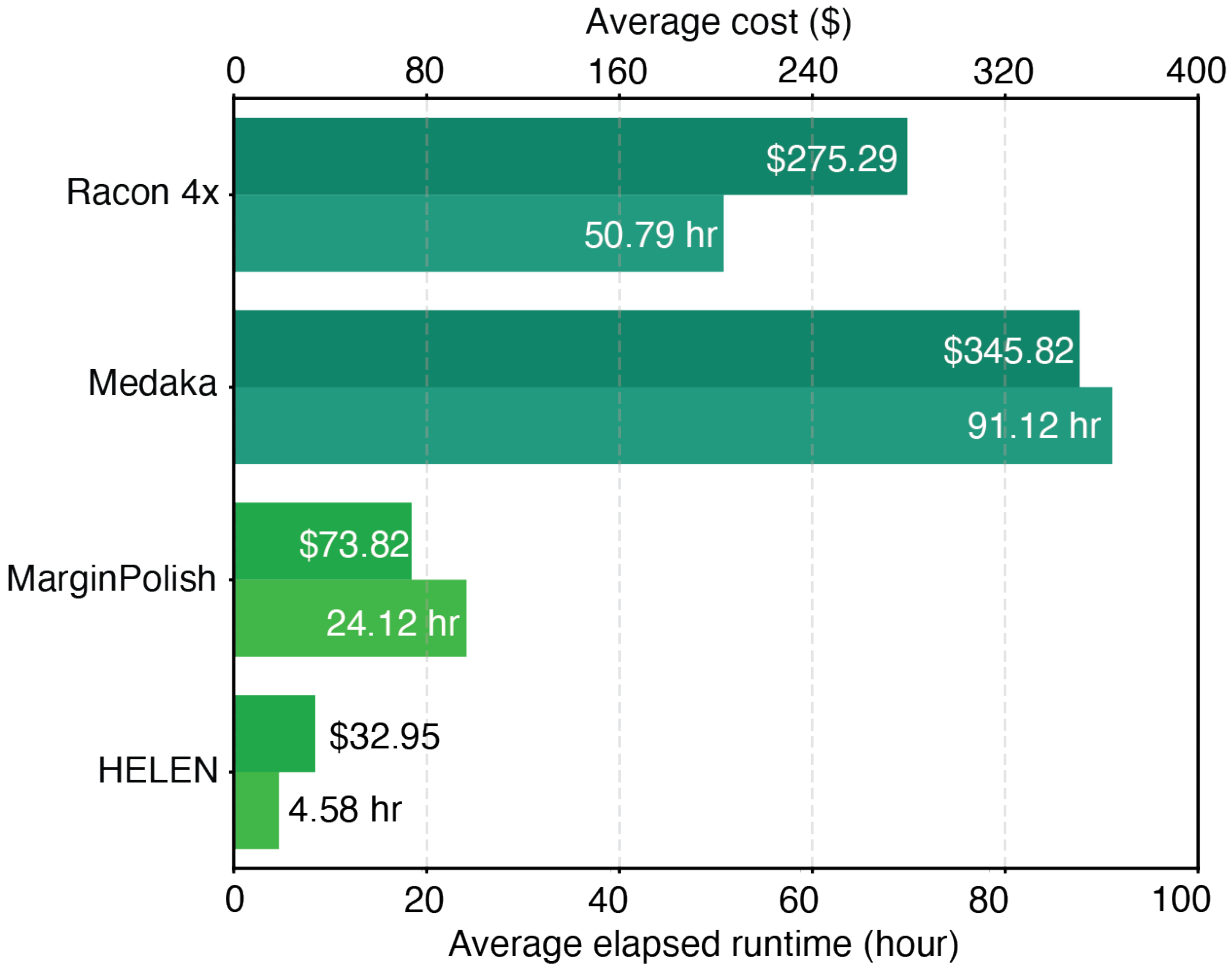
A DNN-based consensus sequence polisher

https://github.com/kishwarshafin/helen

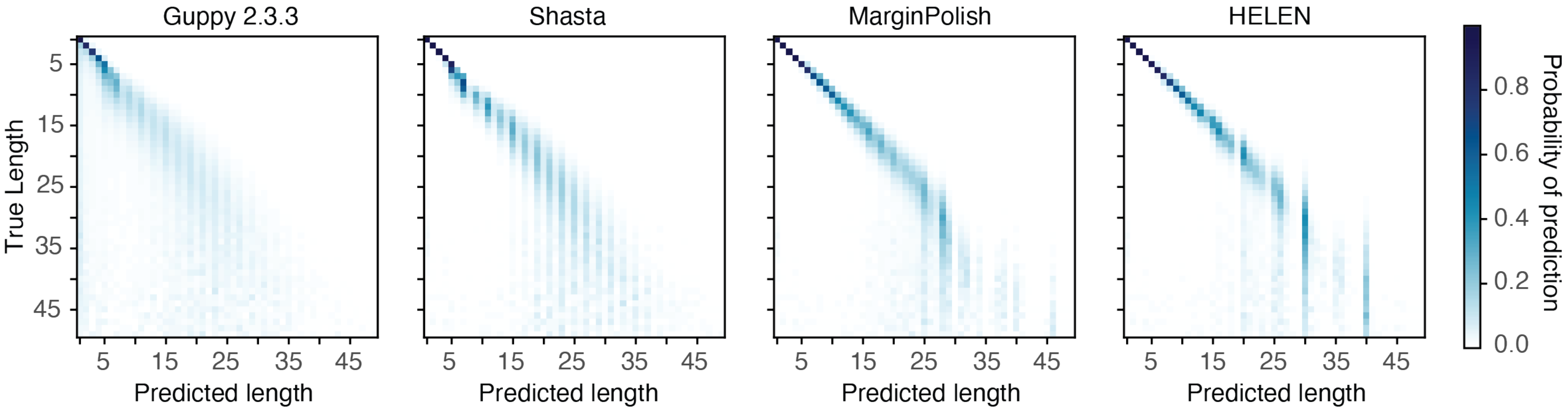
MarginPolish and HELEN outperform other polishers



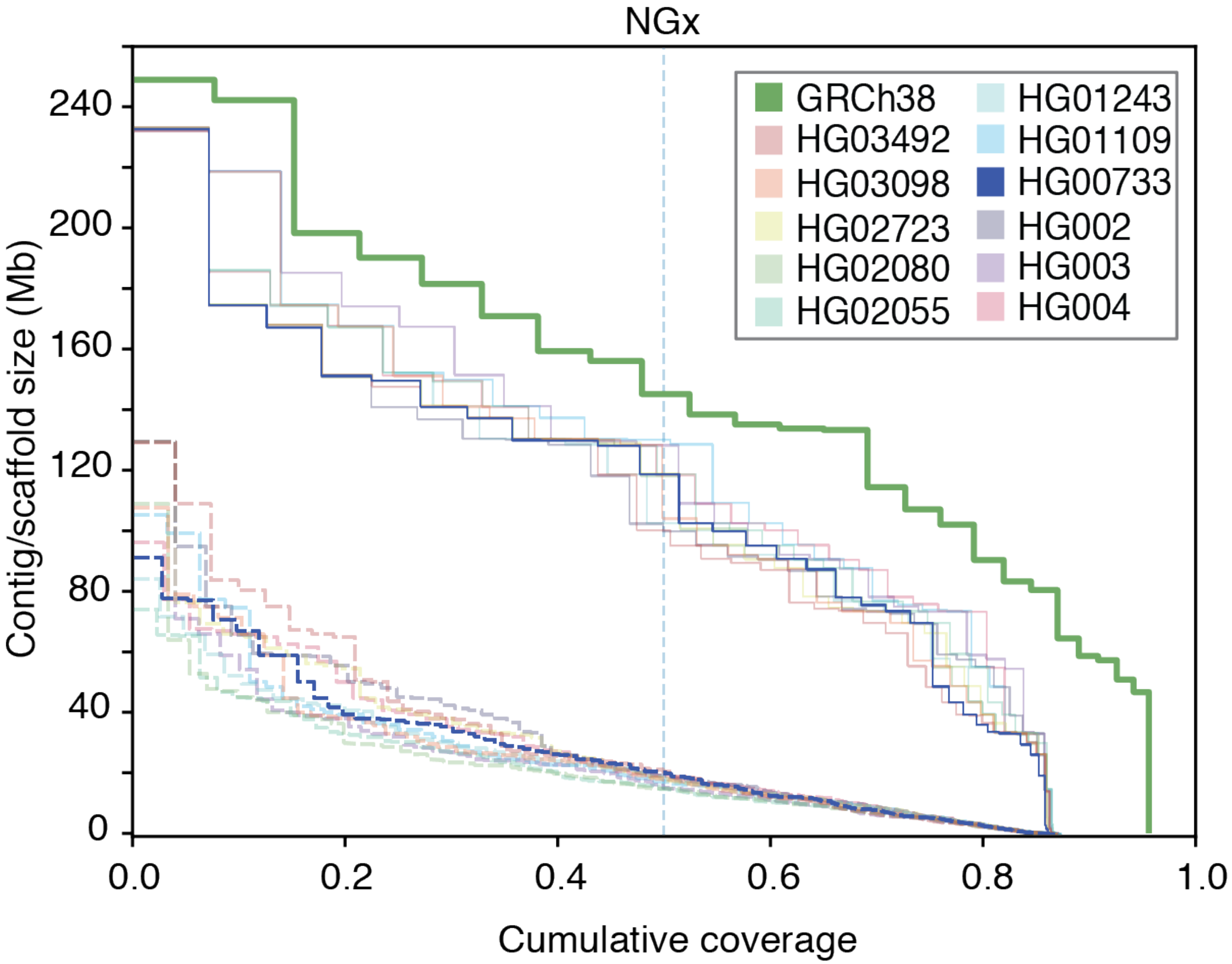
Polishing at a fraction of time and cost



Improvements in homopolymer length predictions



Chromosome-level scaffolding using HiC data



Key next steps

- Faster basecalling (ONT)
- Haplotype phasing (UCSC, CZI)

Shasta

MarginPolish

HELEN

Data resources

- Human Pangenome Reference Consortium
  - https://tinyurl.com/hpp-hg002
  - https://tinyurl.com/hpp-data
- Human Pangenomics
  - https://tinyurl.com/hpgp-data
- Nanopore DNA consortium
  - https://tinyurl.com/na12878-dna
- Nanopore RNA consortium
  - https://tinyurl.com/na12878-rna

Acknowledgements and Collaborators

U British Columbia

U Oregon

UCSF

UC Santa Cruz

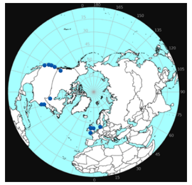
UCLA

Penn

Johns Hopkins

Daniel Garalde
Rosemary Dokos
Simon Mayes
Vania Costa
Jon Pugh
Chris Seymour
Chris Wright
David Stoddart
Dan Turner

NIH National Human Genome Research Institute



amazon web services

UC Santa Cruz

Hugh Olsen ✓
Robin Abu-Shumays
Vinay Poodari
Niki Thomas
Jenny Vo
Logan Mulroney
Ed Green
Holger Schmidt
Ali Yanik

Mark Akeson ✓
Karen Miga ✓
Benedict Paten ✓
David Haussler
Angela Brooks
Manny Ares
Jeremy Sanford
Holger Schmidt
Ali Yanik
Kristof Tigyi

Kishwar Shafin ✓
Sofie Salama
Marina Haukness
Trevor Pesout
Ryan Lorig-Roach
Chris Vollmers
Daniel Kim
Susan Carpenter
Ed Green
Ro Kamakaka

vg

Adam Novak
Glenn Hickey
Jordan Eizenga
Erik Garrison
Jean Monlong
Xian Chang

Adam Phillippy (NHGRI)
Fritz Sedlazeck (Baylor)
Sergey Koren (NHGRI)

circulomics

Kelvin Liu
Duncan Kilburn
Jeffrey Burke

OICR

U Birmingham

U Nottingham

Newcastle U

U Copenhagen

CHAN ZUCKERBERG INITIATIVE

Chan-Zuckerberg

Paolo Carnevali ✓

@mitenjain