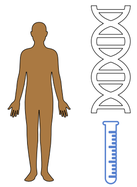
Deploying deep neural network with ONNX for efficient genome analysis with nanopore sequencing

Kishwar Shafin

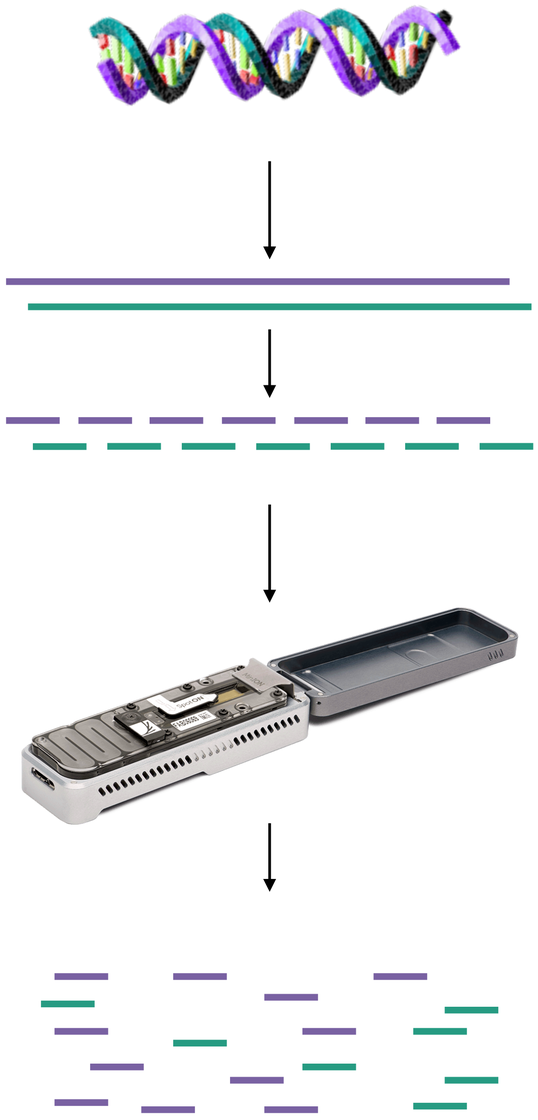
UNIVERSITY OF CALIFORNIA
SANTA CRUZ | Genomics Institute

Overview of genome analysis



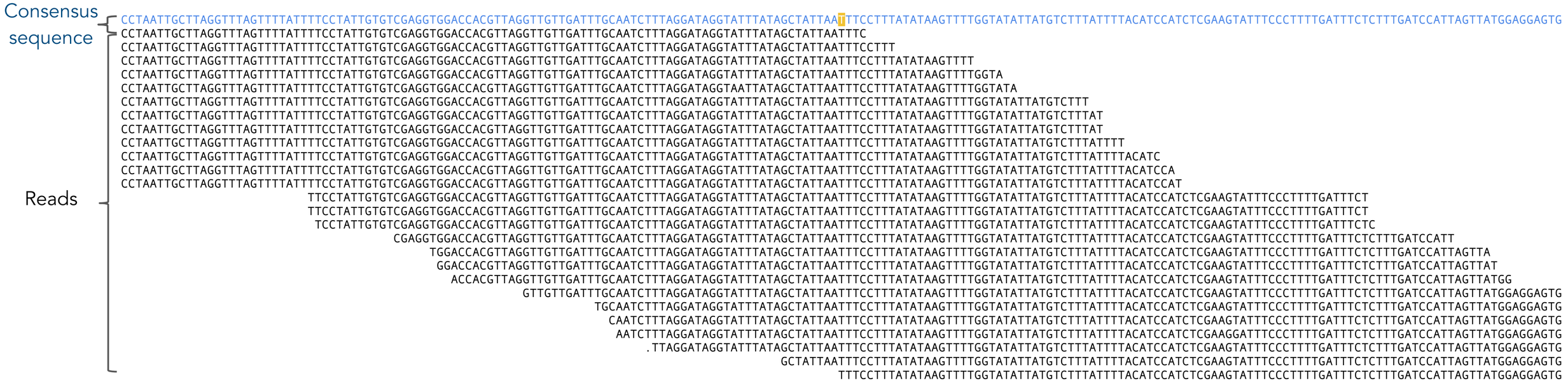
Sample collection

Library Preparation



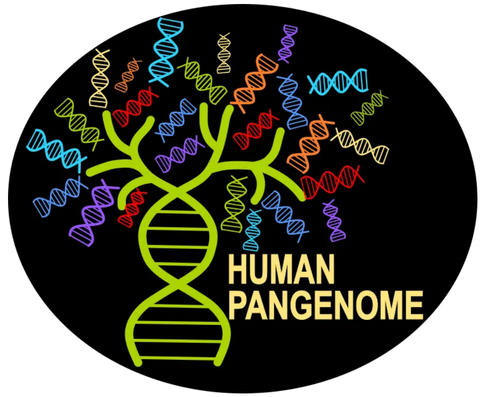
ACGTTACGTTATTCAGTTT

de novo assembly



Global effort

T2T
TELOMERE-TO-TELOMERE CONSORTIUM



Human Pangenome Reference Consortium

VGP
VERTEBRATE
GENOMES
PROJECT
A PROJECT OF THE G10K CONSORTIUM

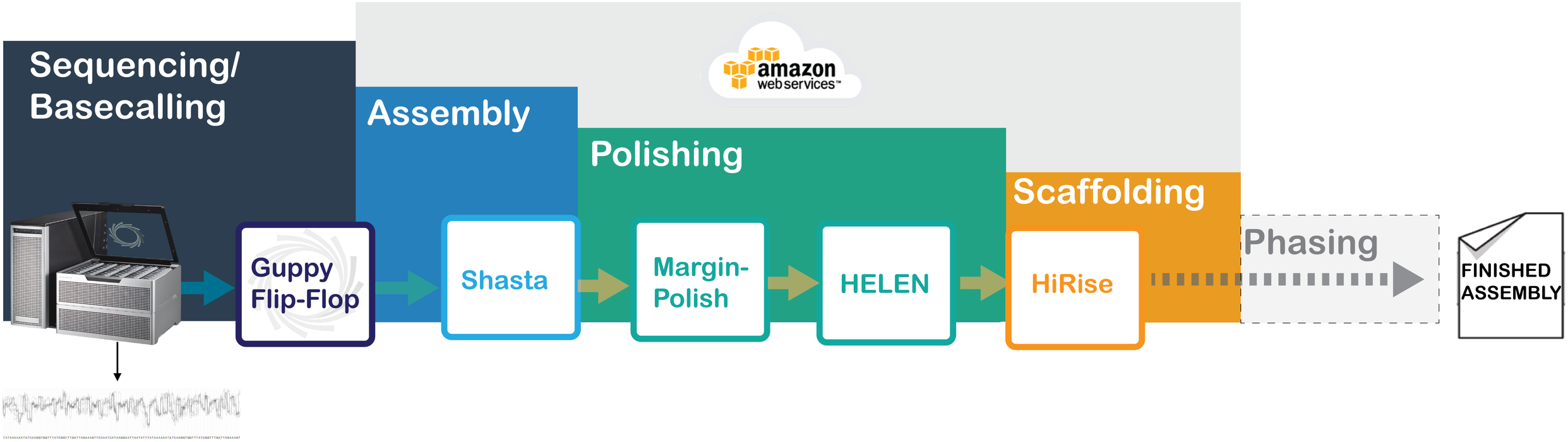

Darwin
TREE
of
LIFE

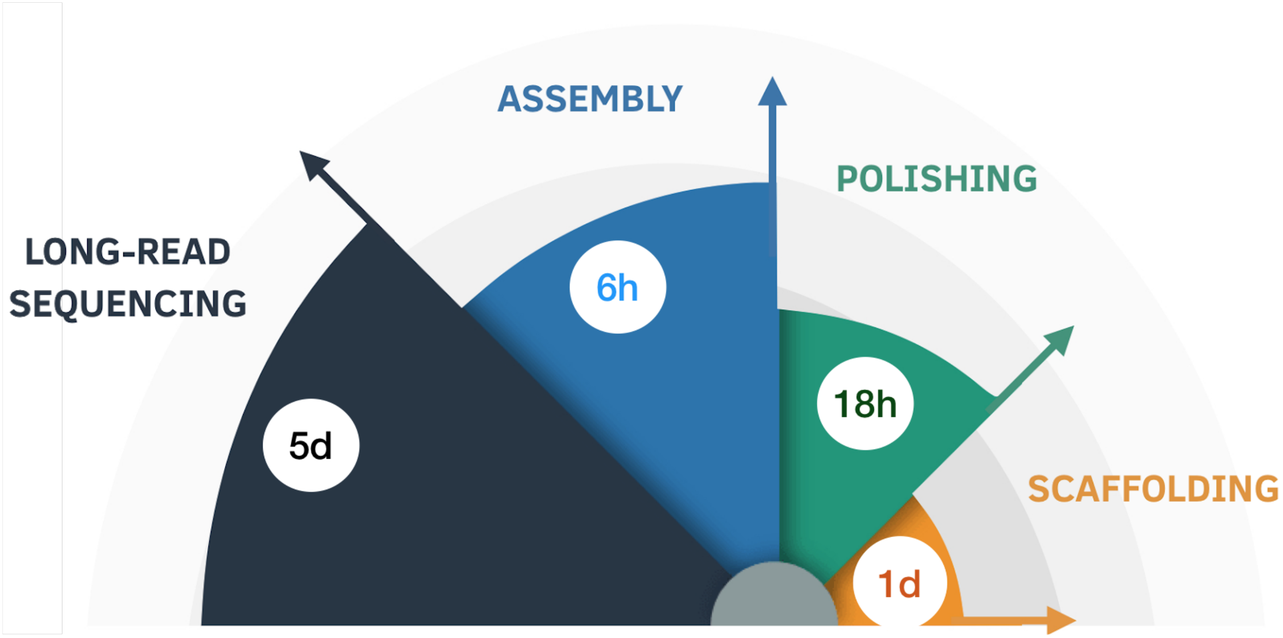
Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes.

Kishwar Shafin et al.

Nature Biotechnology, Accepted March 26, 2020

UNIVERSITY OF CALIFORNIA
SANTA CRUZ | Genomics Institute

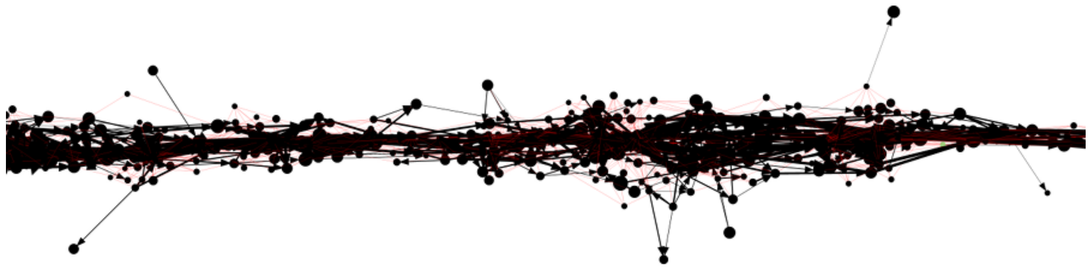




UNIVERSITY OF CALIFORNIA
SANTA CRUZ | Genomics Institute

Shasta assembler

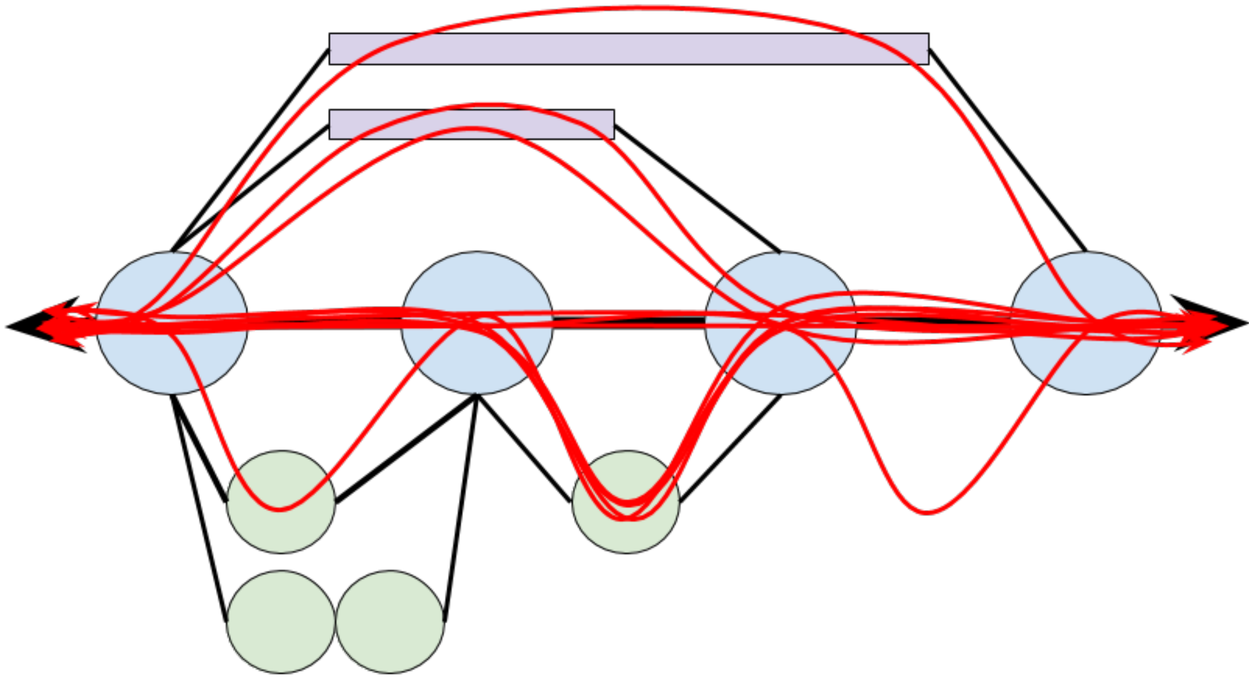
```
 560       570       580       590       600       610       620       630       640       650       660
.|....+....|....+....|....+....|....+....|....+....|....+....|....+....|....+....|....+....|....+....|....+.
1122111132111221311111131211111131211121111211111311112111221111111212111241522531412221112111111111311111
ATATCATCGCATGATCTGAGTACAGCTGTGACTATCACTCATATCAGACTACTGACATGTGATACTCATAGTGCTATACTACTAGTCAGTCTATGTGTATGTGTGATA
 TATCATCGCA  ATCTGAGTAC                          GACTACTGAC          TCATAGTGCT      CTAGTCAGTC ATGTGTATGT
        GCATGATCTG                                    TGACATGTGA    CATAGTGCTA           AGTCTATGTG
                 CTGAGTACAG                                                                    TGTGTATGTG
                  TGAGTACAGC                                                                    GTGTATGTGT
```



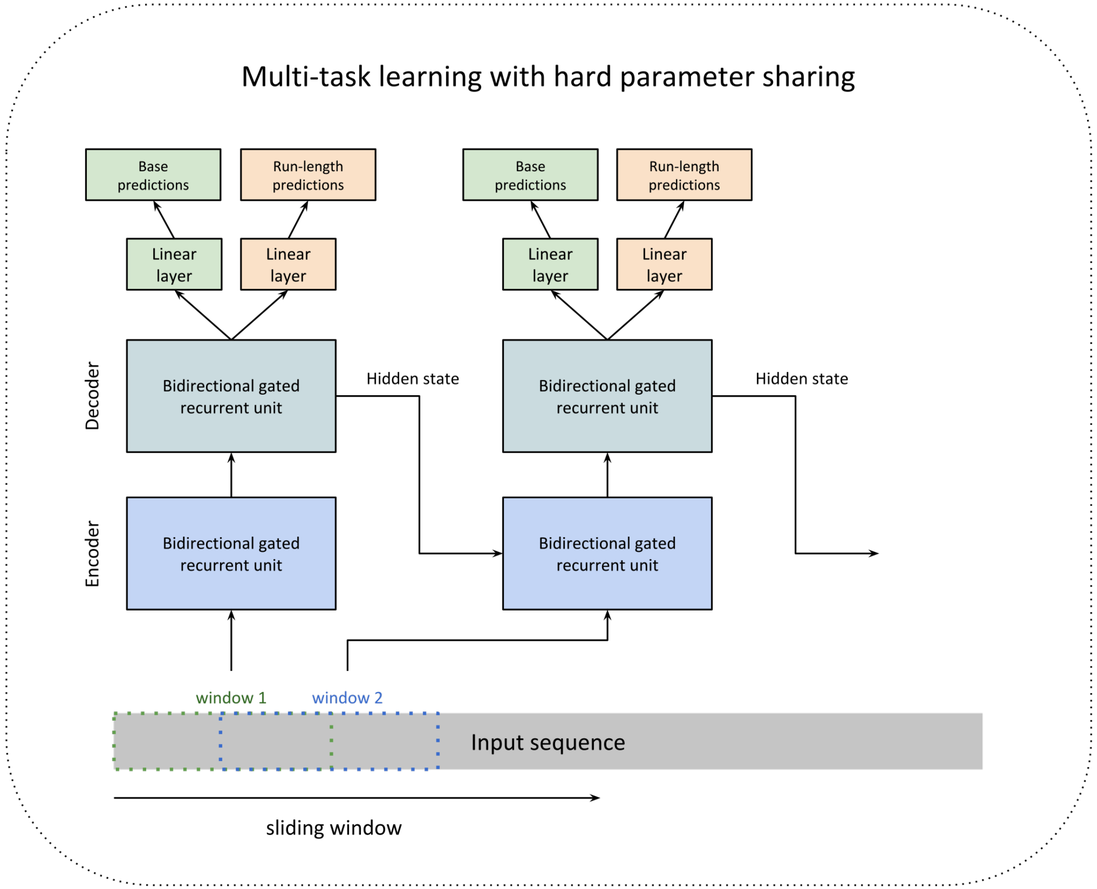
https://github.com/chanzuckerberg/shasta

Paolo Carnevali, CZI

Margin-Polish

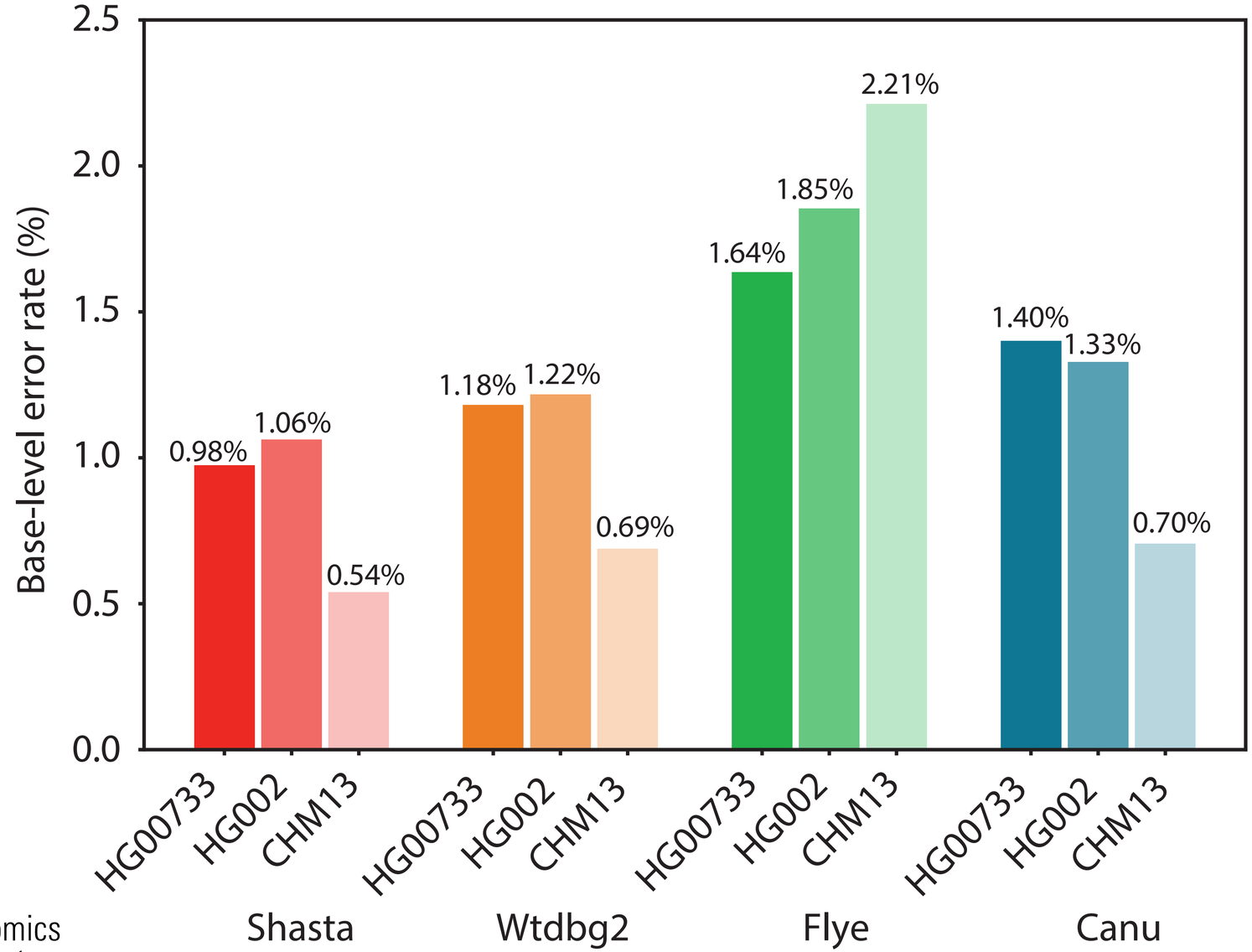


H.E.L.E.N.

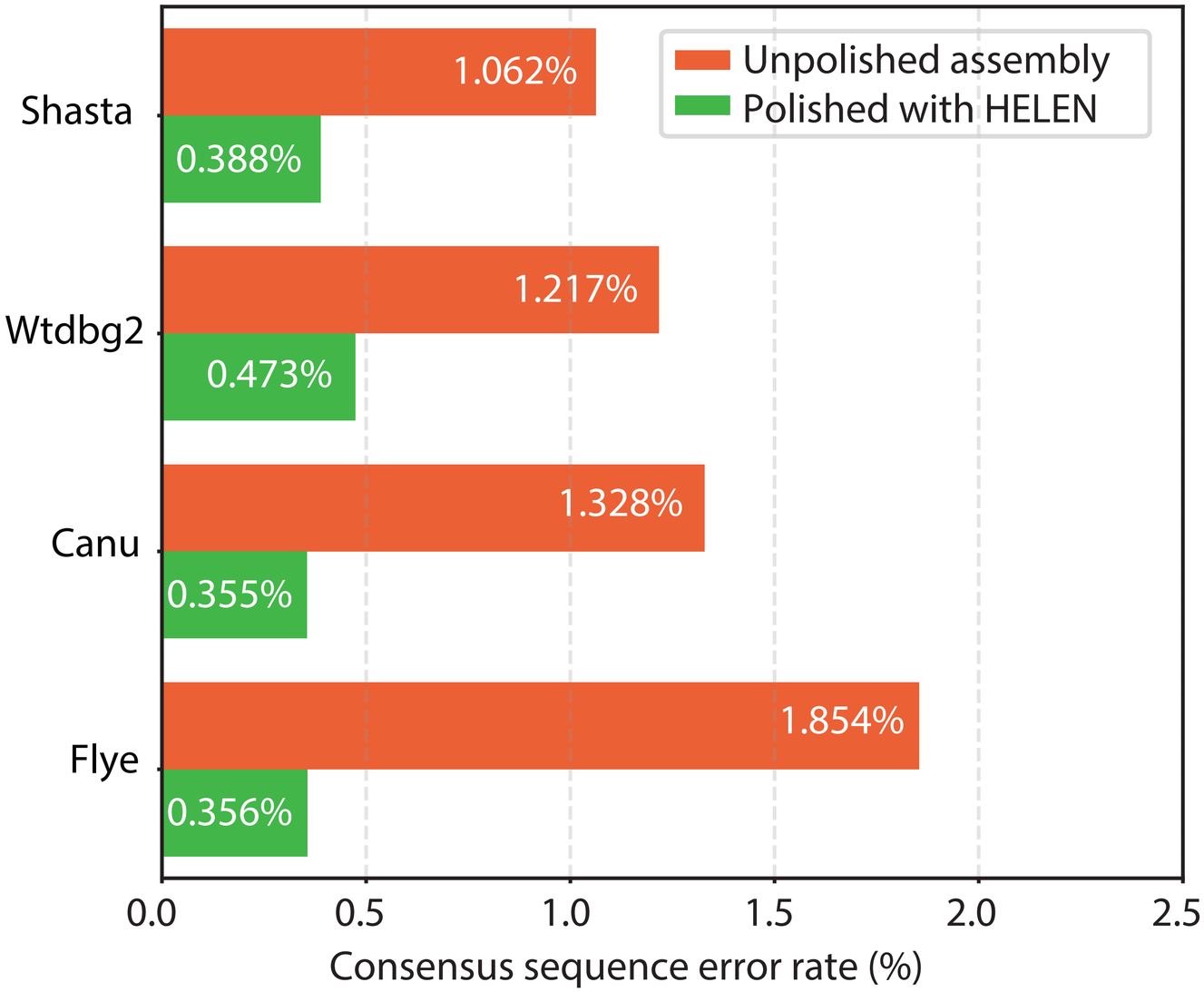


https://github.com/kishwarshafin/helen

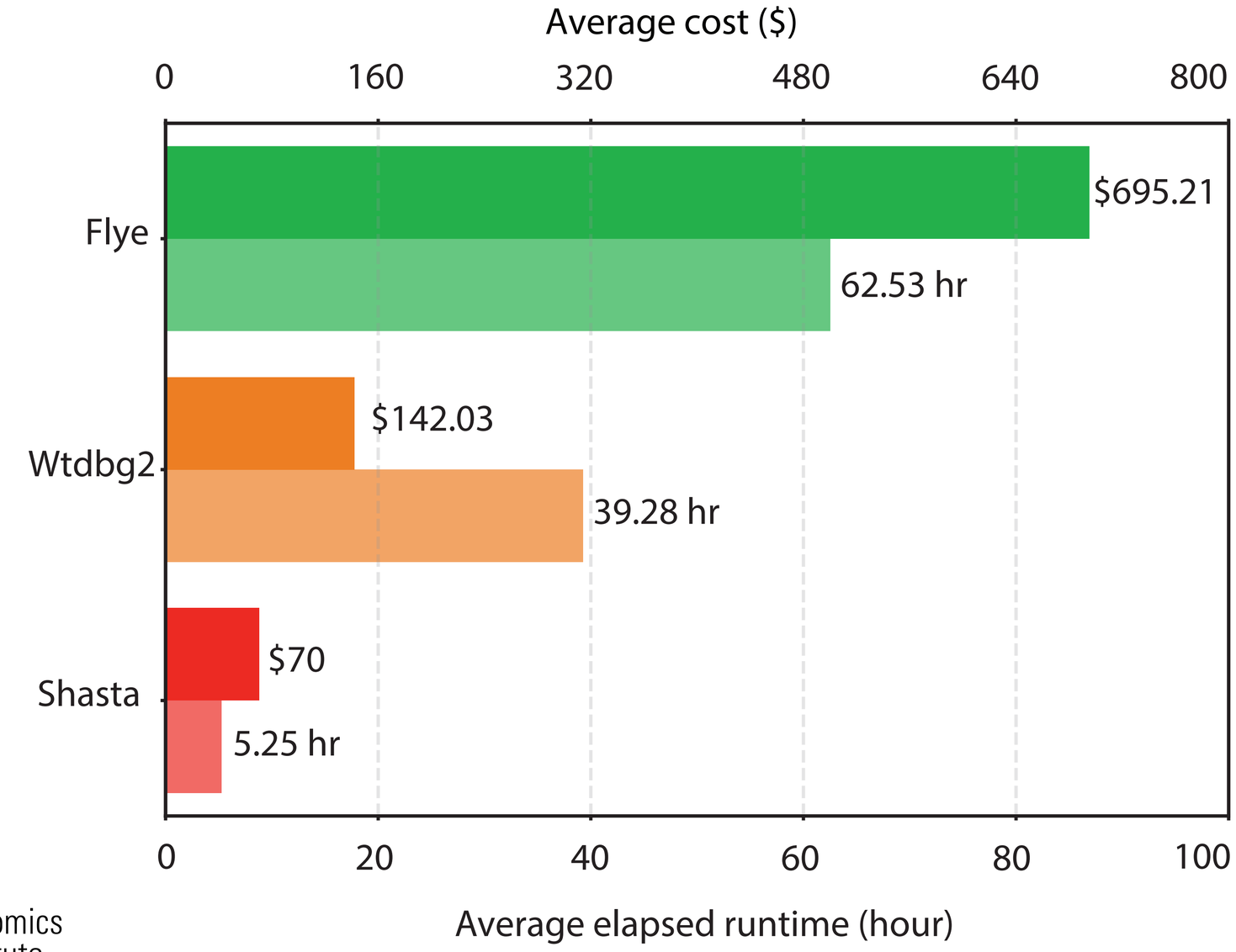
Base-level accuracy



Base-level accuracy improvement

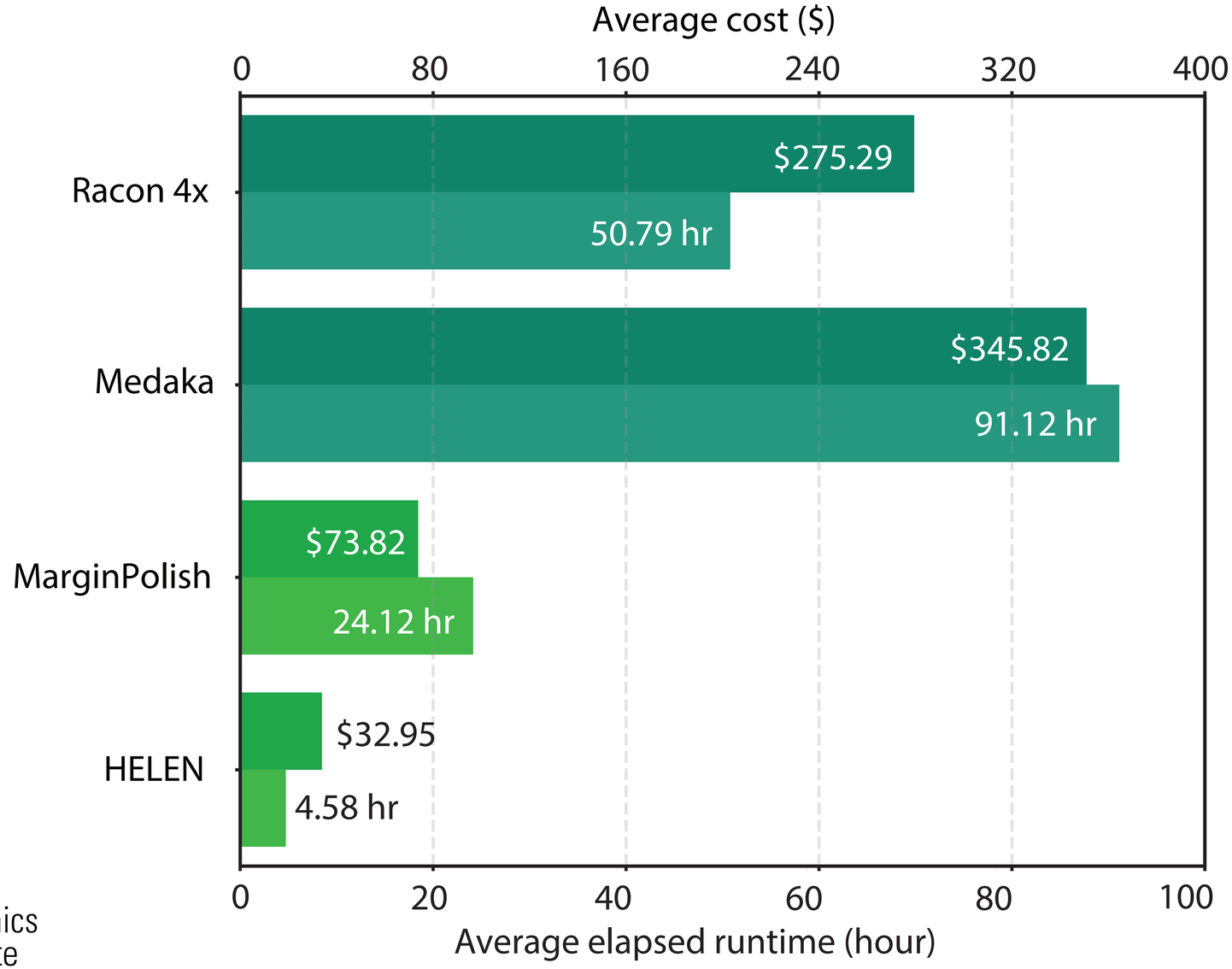


Shasta run-time



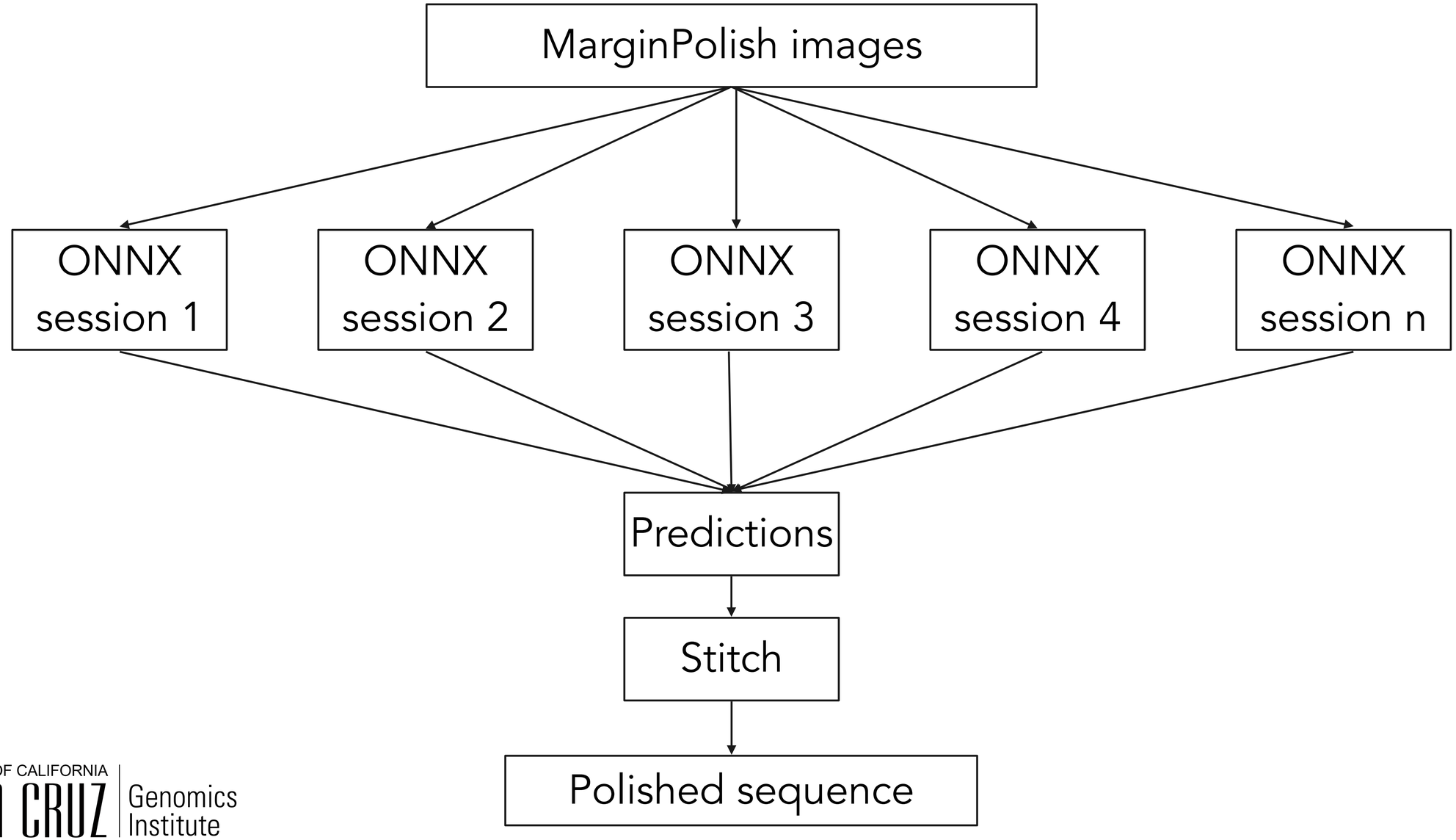
UNIVERSITY OF CALIFORNIA
SANTA CRUZ | Genomics Institute

Run-time analysis



UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

ONNX based HELEN CPU model deployment



Run-time improvement after introducing ONNX runtime

Subset	Expected genome size	Previous release (Wall-clock time)	Current release (Wall-clock time)
Human genome (HG00733)	3.2 Gb	40 hours	7 hours
Human genome (HG00733)	3.2 Gb	36 hours	6 hours
Human genome (CHM13)	3.2 Gb	42 hours	8 hours
Listeria monocytogenes (Microbial)	2.8 Mb	45 mins	6 mins
Bacillus subtilis (Microbial)	4.2 Mb	2 hours	12 mins
Salmonella entrica (Microbial)	5.1 Mb	3 hours	18 mins
Escherichia coli (Microbial)	4.6 Mb	2 hours	13 mins

Acknowledgements

UC SANTA CRUZ

David Haussler
Karen Miga
Ed Green
Sofie Salama
Hugh Olsen
Mark Akeson
Kristof Tigyi
Nicholas Maurer

Adam Phillippy (NHGRI)
Sergey Koren (NHGRI)
Justin Zook (NIST)
Fritz Sedlazeck (Baylor)

vg

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
Vania Costa

Paolo Carnevali
Sidney Bell
Charlotte Weaver
Michael Barrientos
Ryan King
Bruce Martin
Phil Smoot
Cori Bargmann

Google Health

Pi-Chuan Chang
Andrew Carroll
Sidharth Goel
Maria Nattestad
Howard Yang

VGP VERTEBRATE GENOMES PROJECT
A PROJECT OF THE G10K CONSORTIUM

Erich Jarvis
Chai Fungtammasan
Arang Rhie
+ Many More

amazon web services

NIH National Human Genome Research Institute

circulomics

Kelvin Liu
Duncan Kilburn

UNIVERSITY OF WASHINGTON 1861

Evan Eichler
Mitchell Vollger