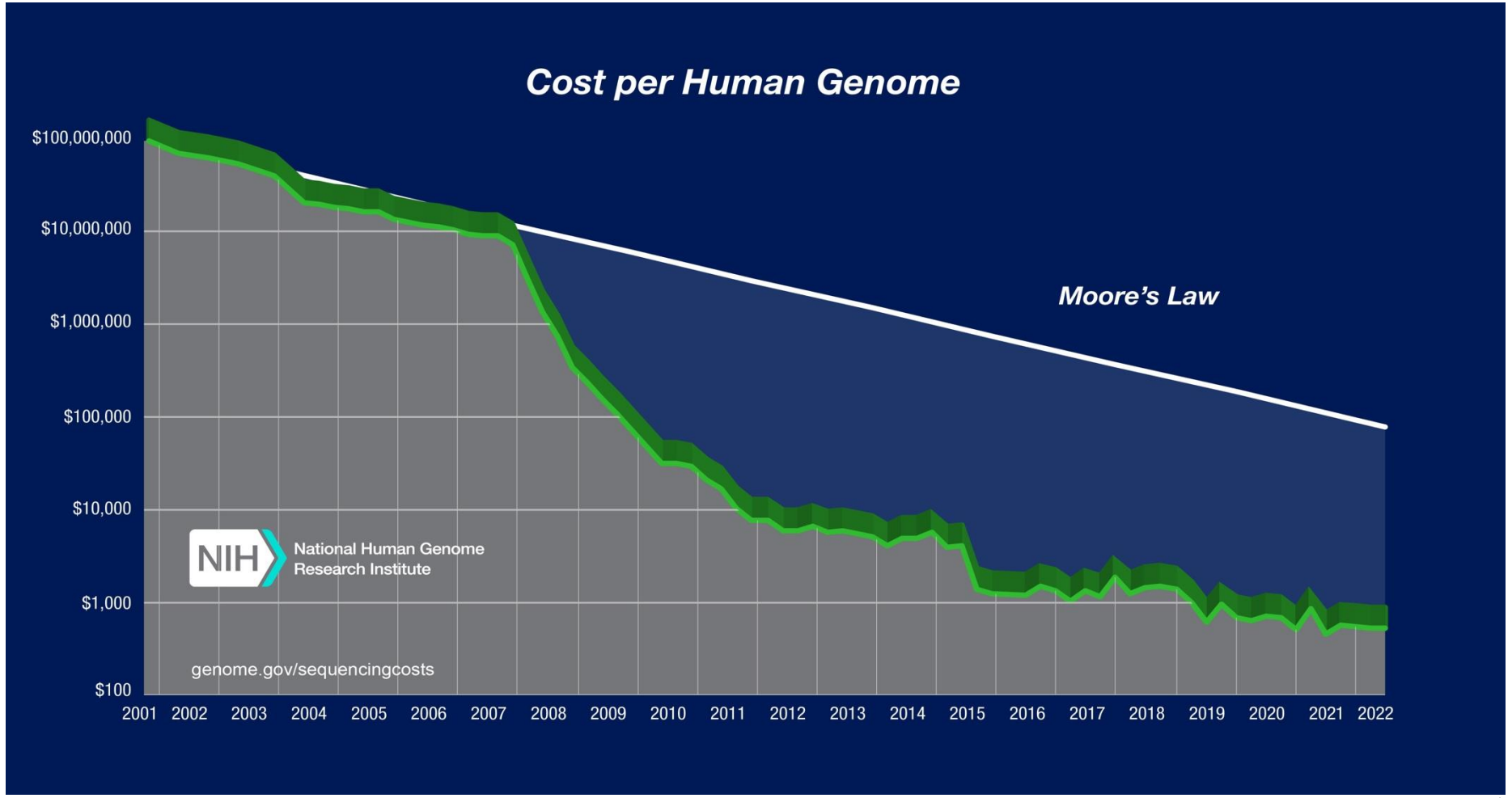
Architects of the Genome Era

Bridging Biology and Compute at Scale

Sneha Goenka

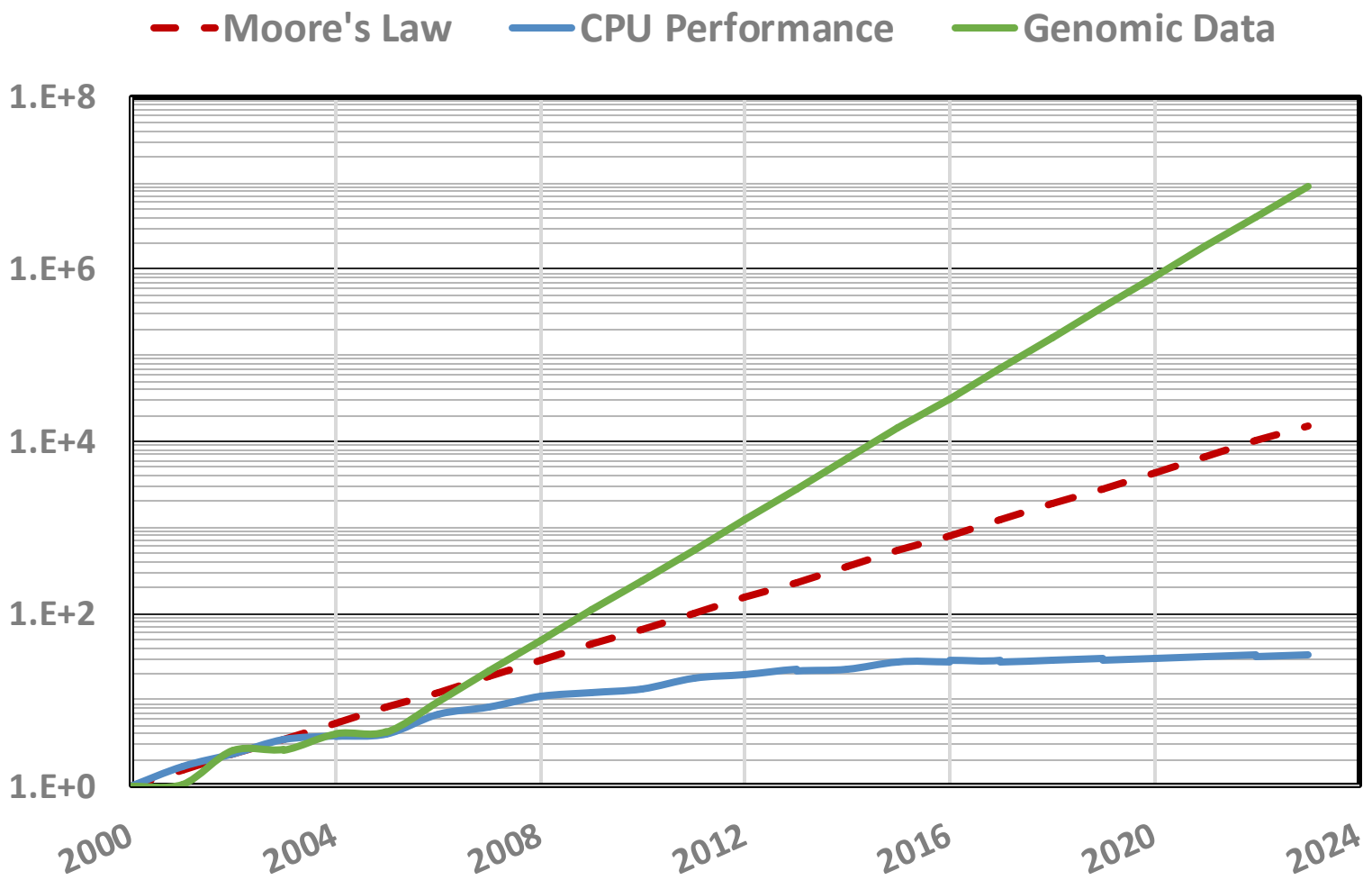
Princeton University

Past 2 decades witnessed significantly reducing cost of sequencing

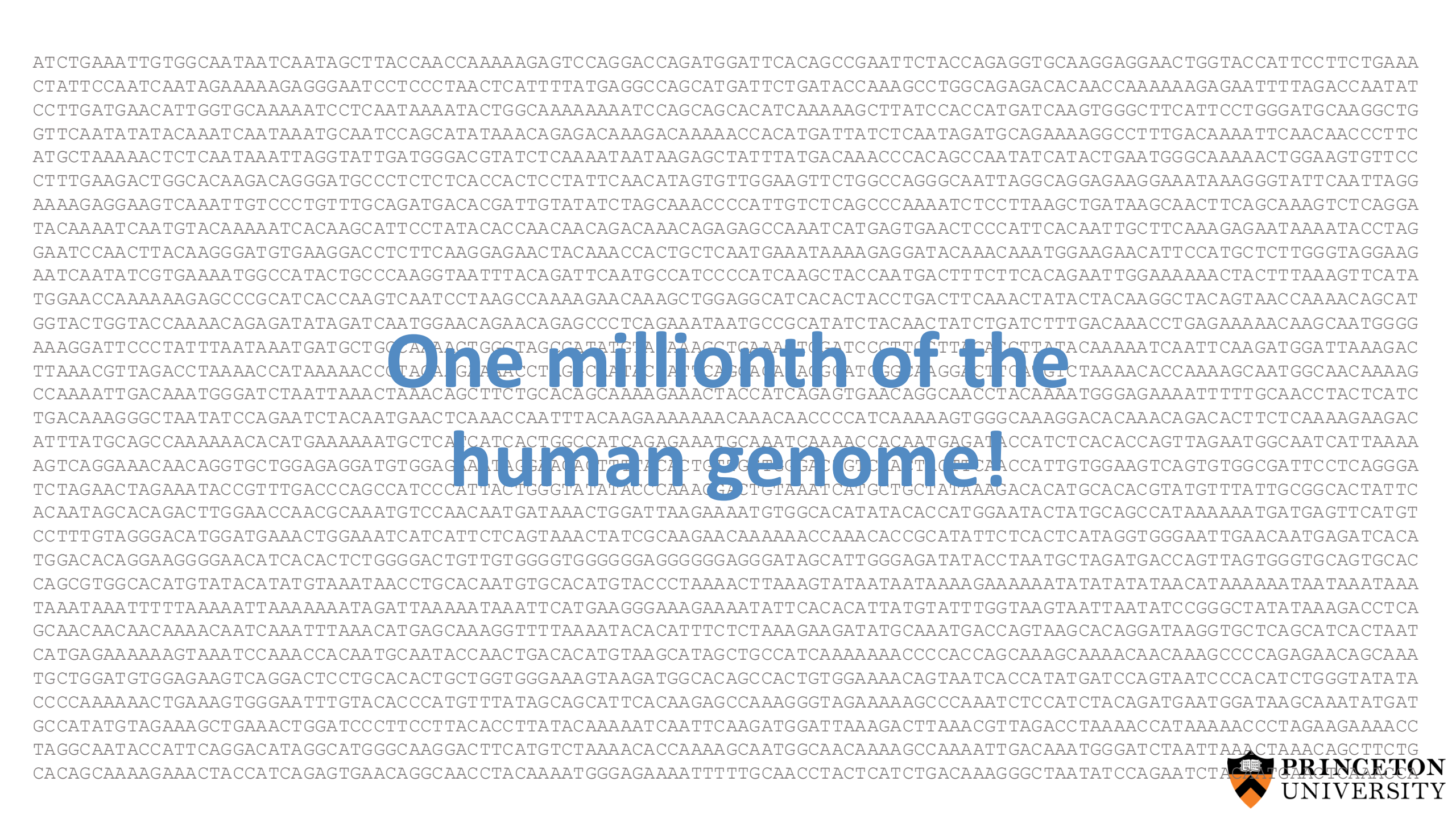


PRINCETON UNIVERSITY

Genomics growth rate >>> CPU performance



PRINCETON UNIVERSITY



PRINCETON UNIVERSITY

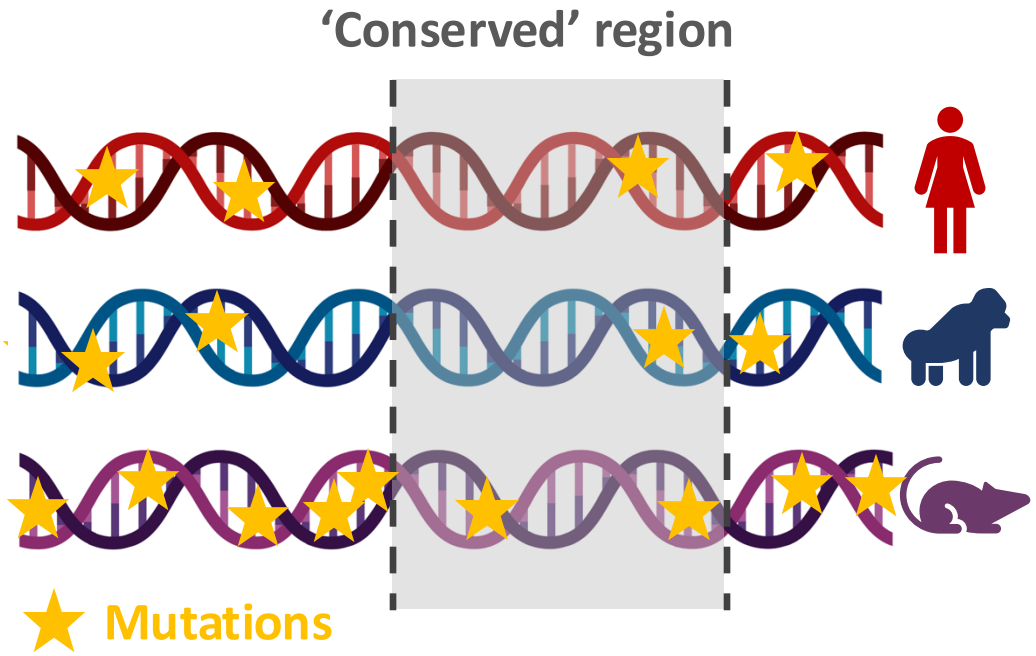
Comparative genomics enables genome interpretation

Article | Published: 25 February 2021

A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity

Chimp Genome Helps Scientists Learn More About Human DNA

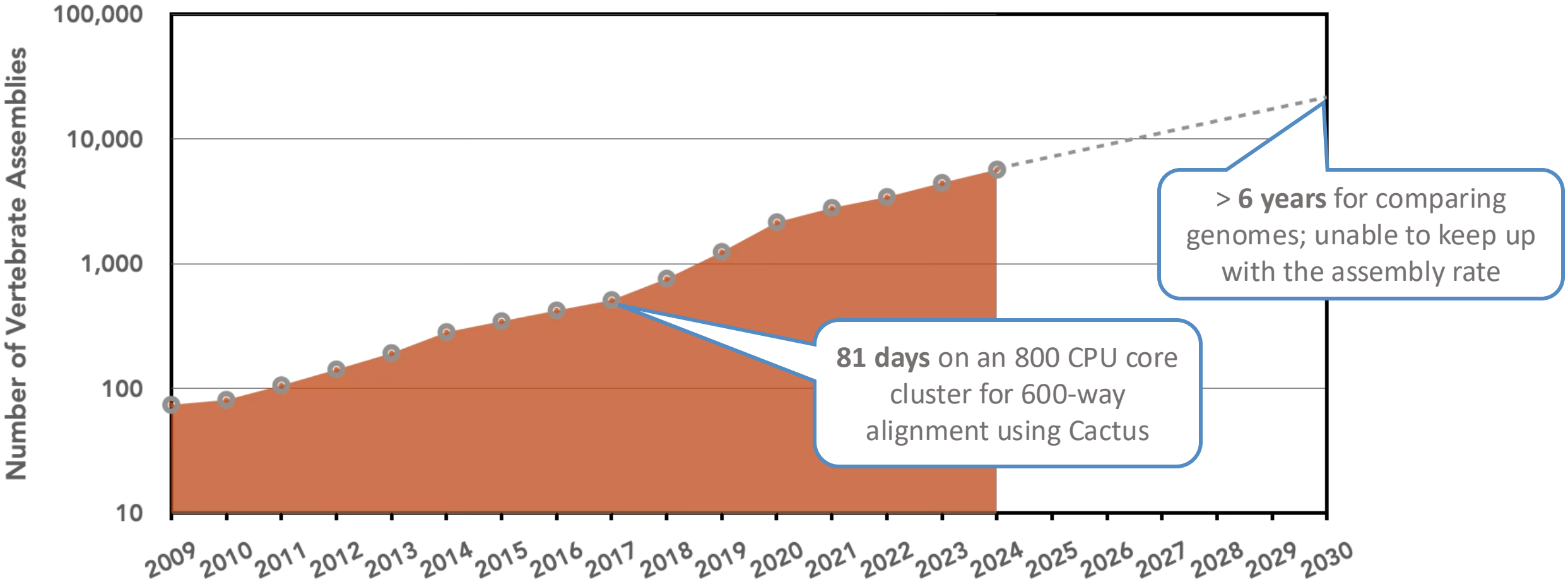
News Published: September 8, 2005



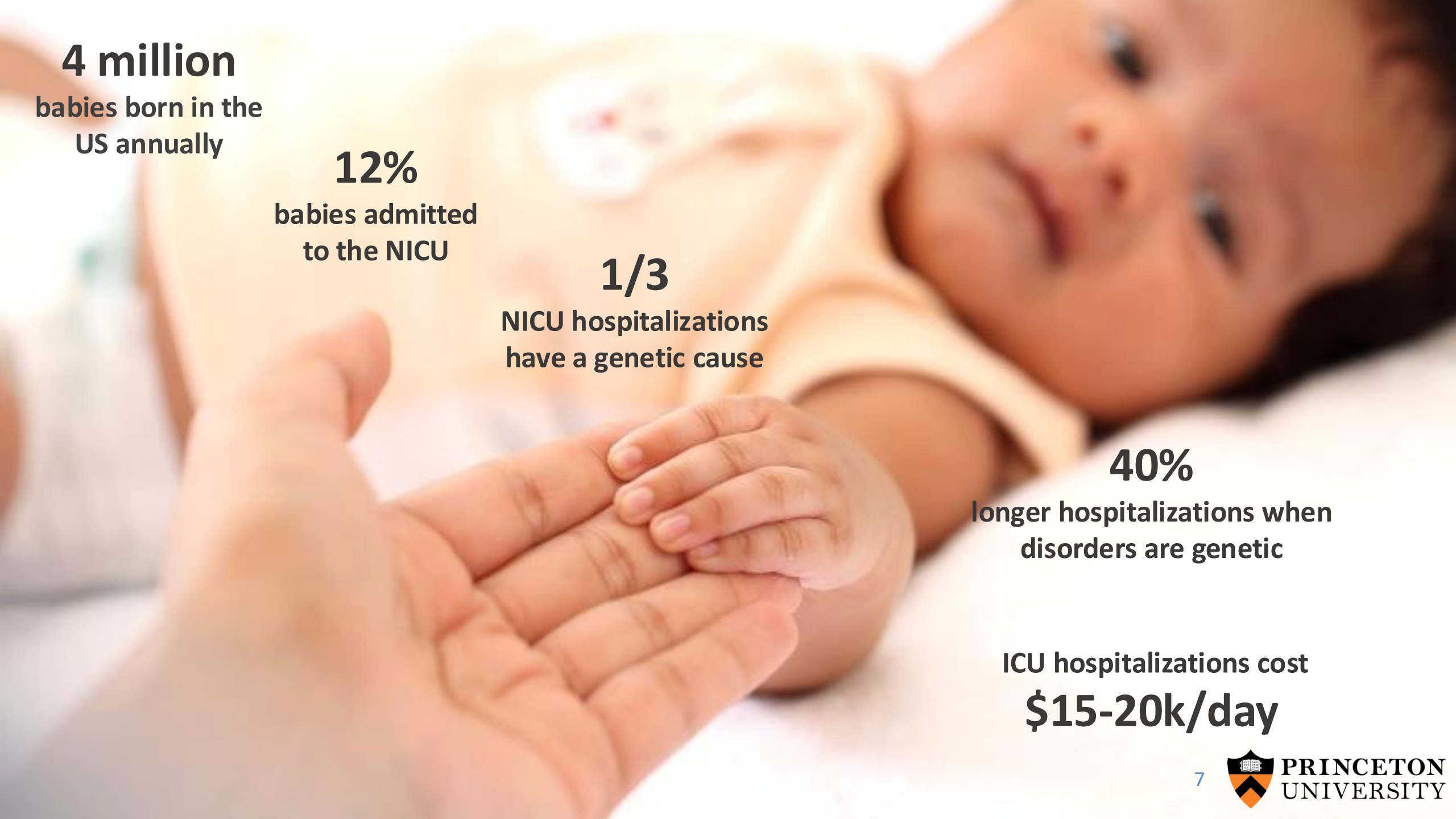
Prediction of functional elements

PRINCETON UNIVERSITY

Thousand-genome era is already here



PRINCETON UNIVERSITY



4 million
babies born in the US annually

12%
babies admitted to the NICU

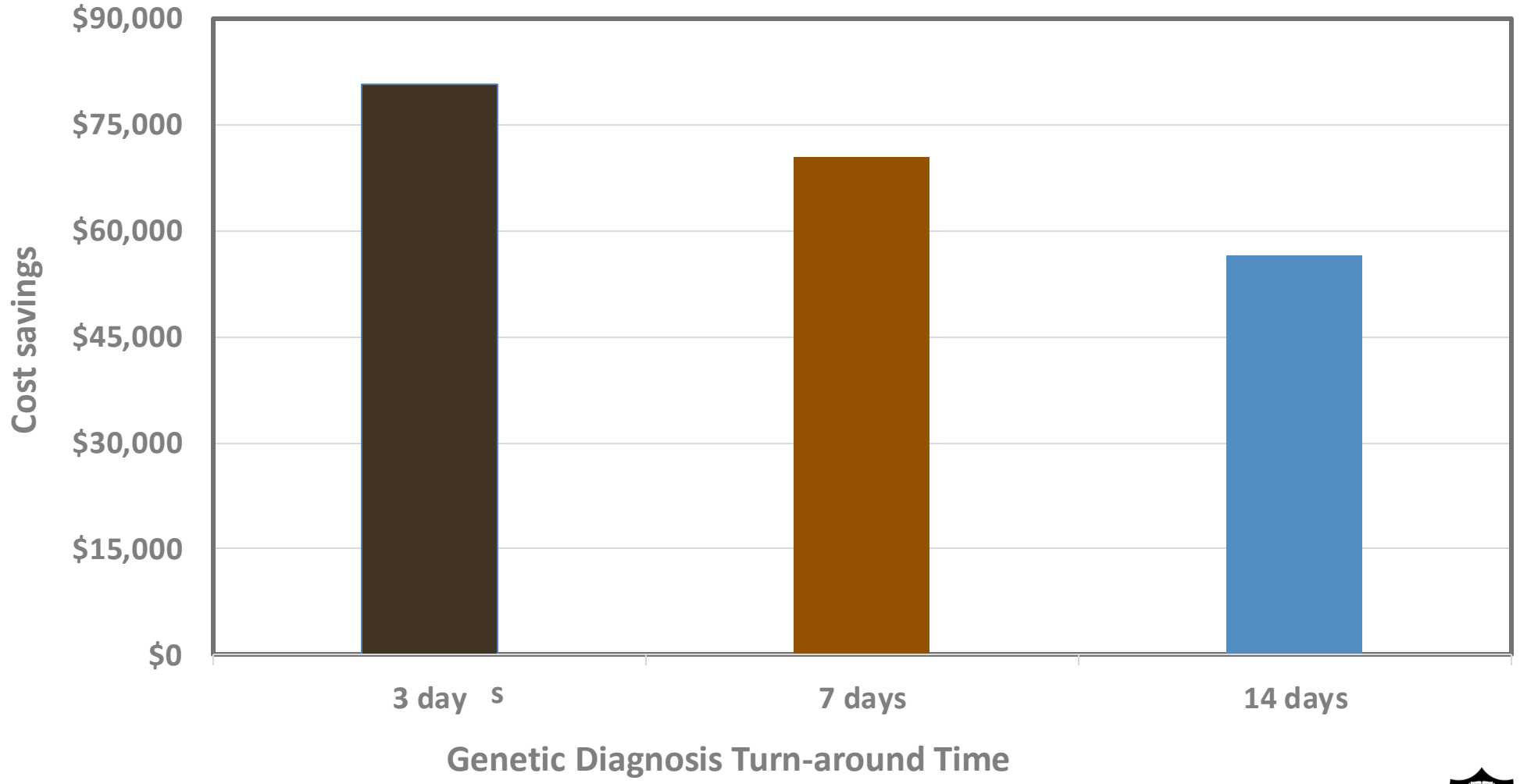
1/3
NICU hospitalizations have a genetic cause

40%
longer hospitalizations when disorders are genetic

ICU hospitalizations cost
$15-20k/day

PRINCETON UNIVERSITY

Faster genetic diagnosis has significant impact

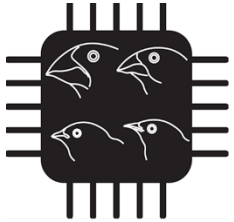


PRINCETON UNIVERSITY

Enabling biologists to leverage sequencing

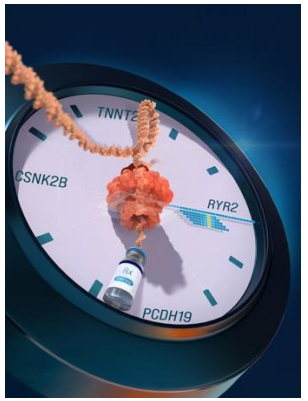
Comparative Genomics

SegAlign: A Scalable GPU-based whole genome aligner
[Supercomputing Conference]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments
[HPCA]

Clinical Genomics



An ultra-rapid workflow for clinical whole genome sequencing
[New England Journal of Medicine Nature Biotechnology]

PRINCETON UNIVERSITY

Fragmented Acceleration in Genomics

SquiggleFilter: An Accelerator for Portable Virus Detection

Tim Dunn* timdunn@umich.edu University of Michigan Ann Arbor, MI, USA

Harisankar Sadasivan* hariss@umich.edu University of Michigan Ann Arbor, MI, USA

Jack Wadden jackwadden@gmail.com University of Michigan Ann Arbor, MI, USA

Kush Goliya kgoliya@umich.edu University of Michigan Ann Arbor, MI, USA

Home / Magazines / IEEE Micro / 2021.04

IEEE Micro

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

Jul.-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

CiMBA: Accelerating Genome Sequencing Through On-Device Basecalling via Compute-in-Memory

William Andrew Simon, *Member, IEEE*, Irem Boybat, *Member, IEEE*, Riselda Kodra, *Student Member, IEEE*, Elena Ferro, *Student Member, IEEE*, Gagandeep Singh, Mohammed Alser, *Member, IEEE*, Shubham Jain, *Member, IEEE*, Hsinyu Tsai, *Senior Member, IEEE*, Geoffrey W. Burr, *Fellow, IEEE*, Onur Mutlu, *Fellow, IEEE*, and Abu Sebastian, *Fellow, IEEE*

SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping

Damla Senol Cali[1] Konstantinos Kanellopoulos[2] Joël Lindegger[2] Zülal Bingöl[3] Gurpreet S. Kalsi[4] Ziyi Zuo[5] Can Firtina[2] Meryem Banu Cavlak[2] Jeremie Kim[2] Nika Mansouri Ghiasi[2] Gagandeep Singh[2] Juan Gómez-Luna[2] Nour Almadhoun Alserr[2] Mohammed Alser[2] Sreenivas Subramoney[4] Can Alkan[3] Saugata Ghose[6] Onur Mutlu[2]

PRINCETON UNIVERSITY

Fragmented Acceleration in Genomics

SquiggleFilter: An Accelerator for Portable Virus Detection

CiMBA: Accelerating Genome Sequencing Through On-Device Basecalling via Compute-in-Memory

Most accelerators lack software frameworks —and when they exist, they're built in isolation with no shared ecosystem

Intensive Applications

Jul.-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

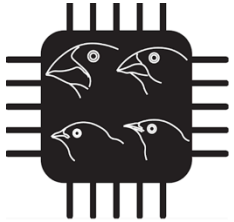
Damla Senol Cali[1] Konstantinos Kanellopoulos[2] Joël Lindegger[2] Zülal Bingöl[3]
Gurpreet S. Kalsi[4] Ziyi Zuo[5] Can Firtina[2] Meryem Banu Cavlak[2] Jeremie Kim[2]
Nika Mansouri Ghiasi[2] Gagandeep Singh[2] Juan Gómez-Luna[2] Nour Almadhoun Alserr[2]
Mohammed Alser[2] Sreenivas Subramoney[4] Can Alkan[3] Saugata Ghose[6] Onur Mutlu[2]

PRINCETON UNIVERSITY

Enabling biologists to leverage sequencing

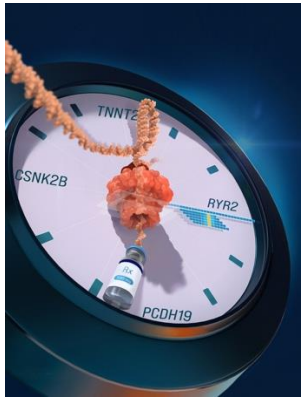
Comparative Genomics

SegAlign: A Scalable GPU-based whole genome aligner
[Supercomputing Conference]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments
[HPCA]

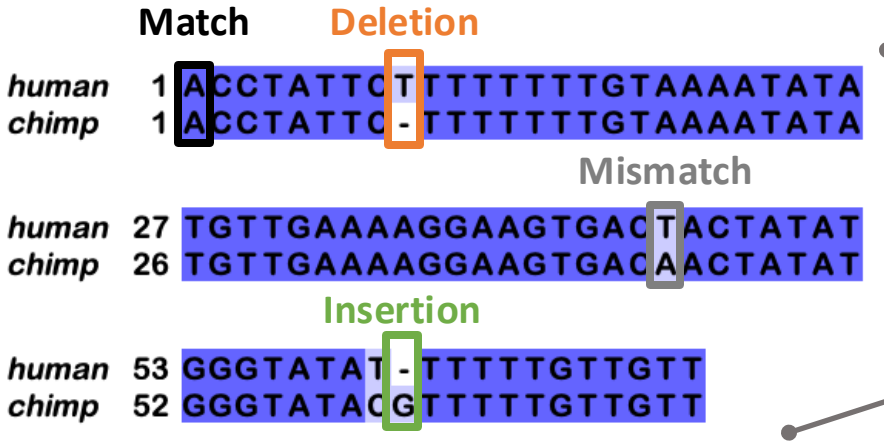
Clinical Genomics

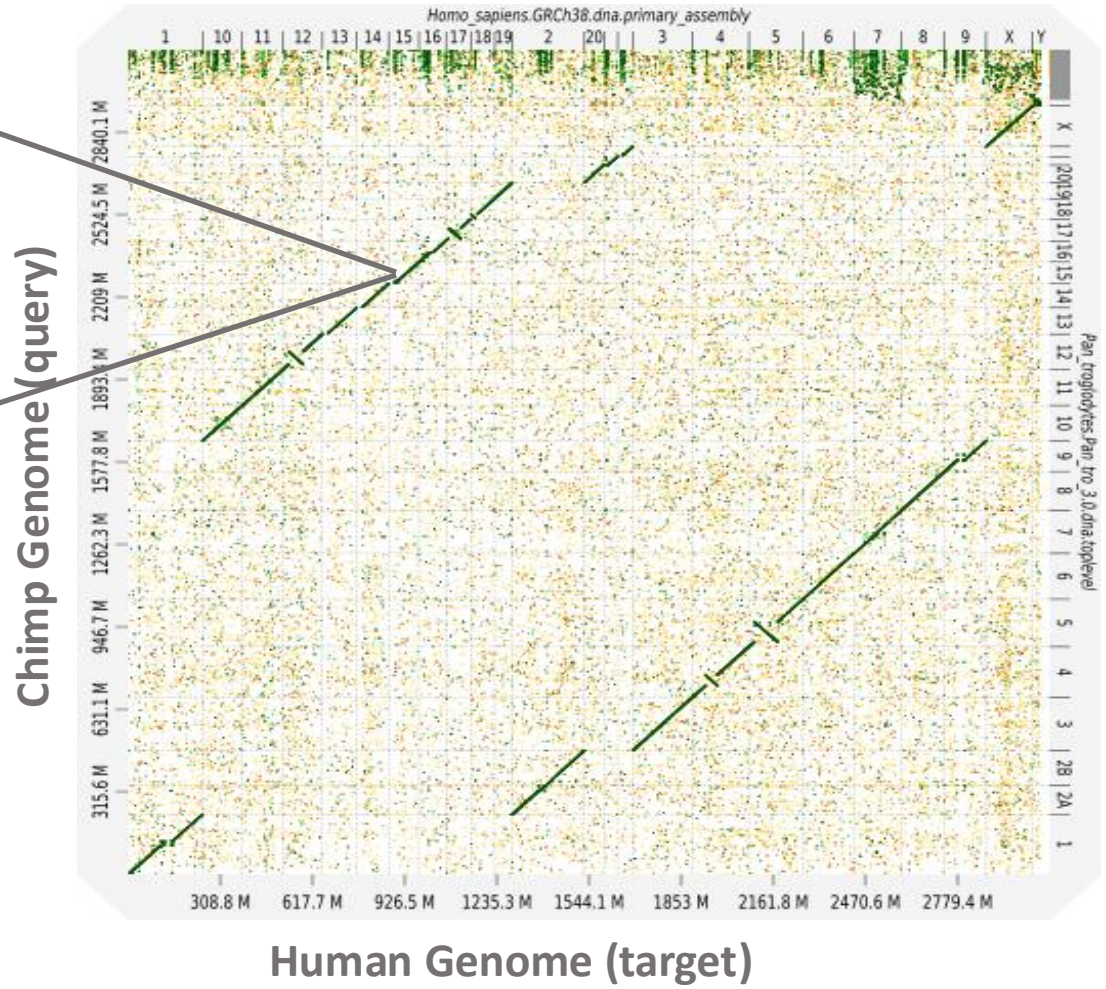


An ultra-rapid workflow for clinical whole genome sequencing
[New England Journal of Medicine Nature Biotechnology]

PRINCETON UNIVERSITY

Understanding whole genome alignment





LASTZ is the state-of-the art whole genome aligner, based on the ***seed-filter-extend*** algorithm

Cabanettes et al. PeerJ6:e4958(2018)

PRINCETON UNIVERSITY

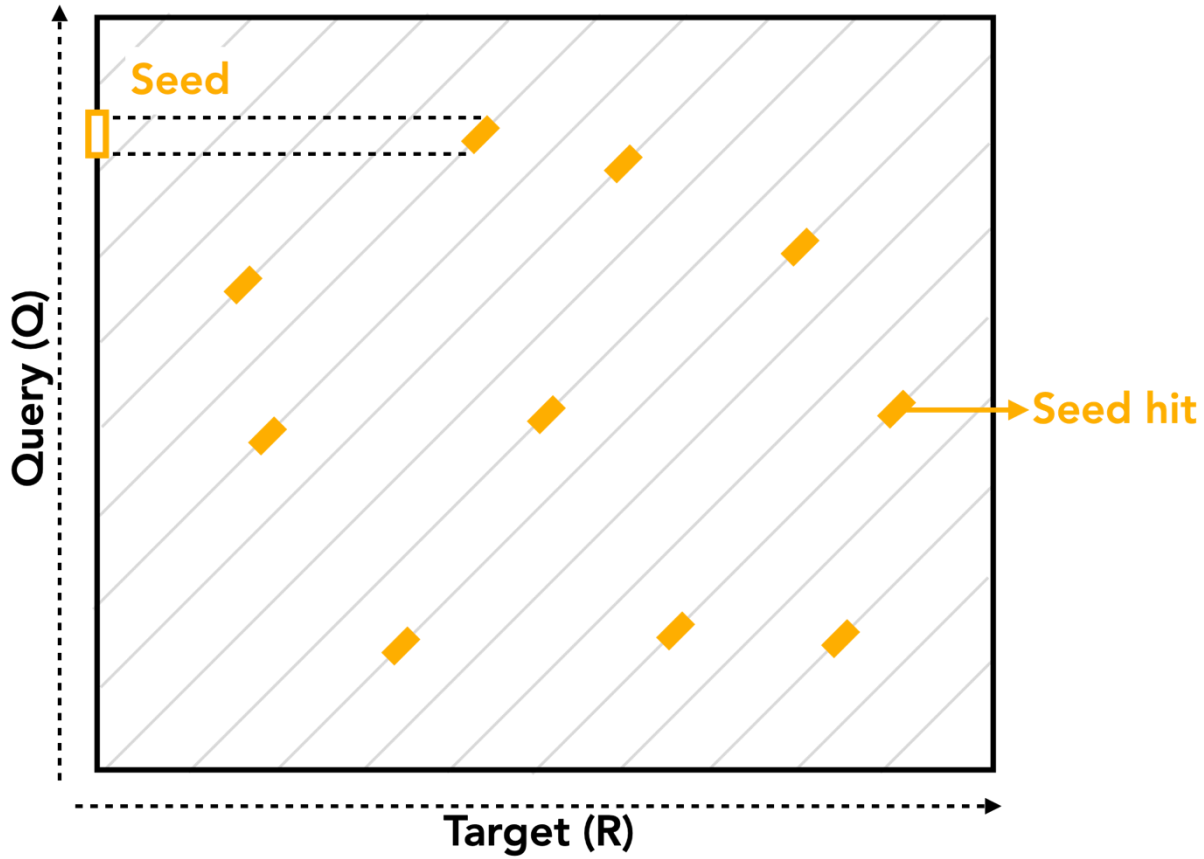
Seeding finds small, local matching base-pairs

Seed hit

R ...CTTGGGTATTCCGTA...

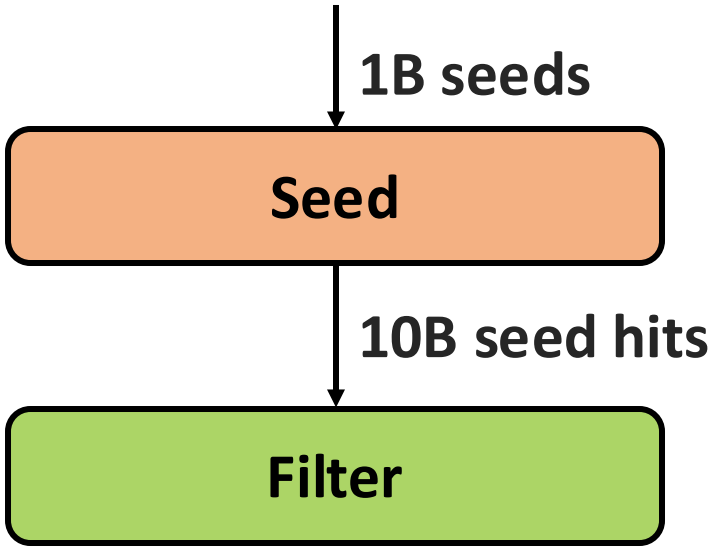
Q ...CTTGGGTATTCCTTA...

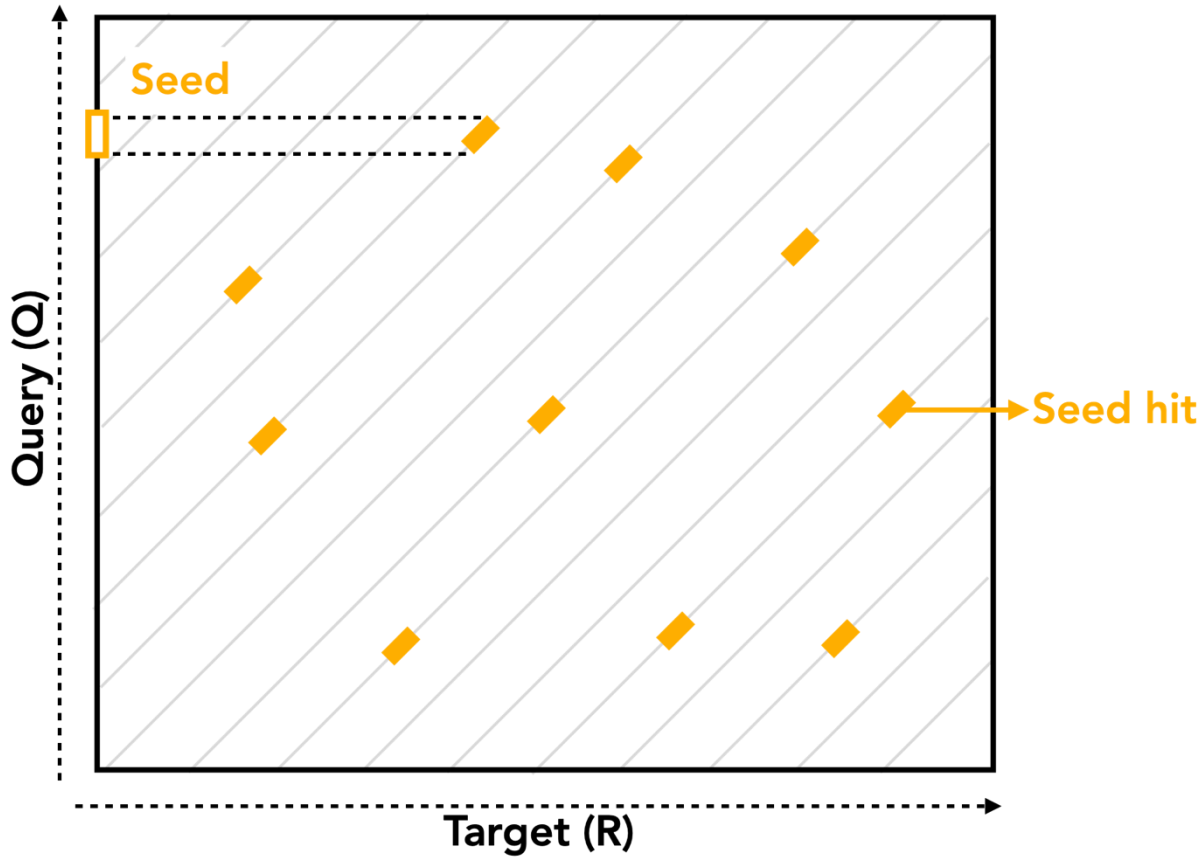
Seed



PRINCETON UNIVERSITY

Seeding finds small, local matching base-pairs

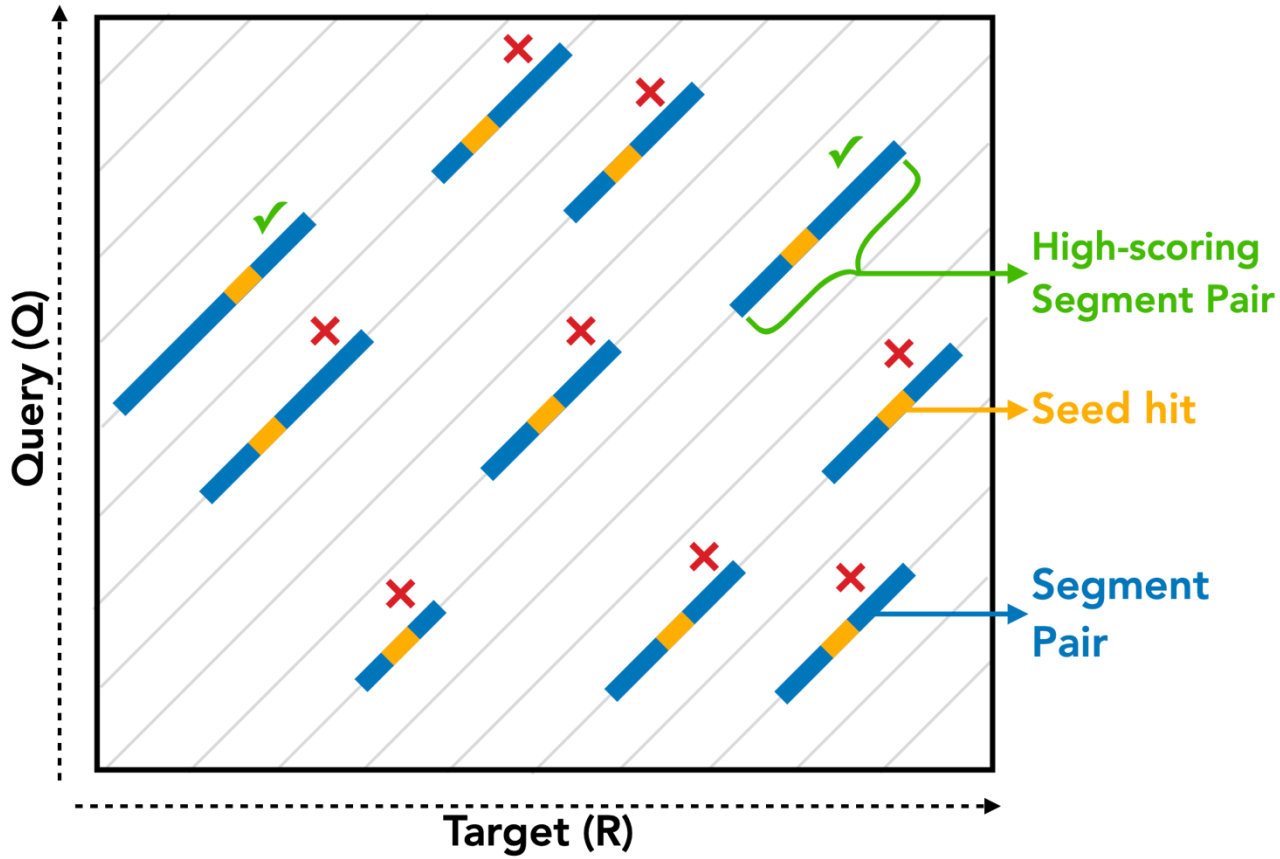




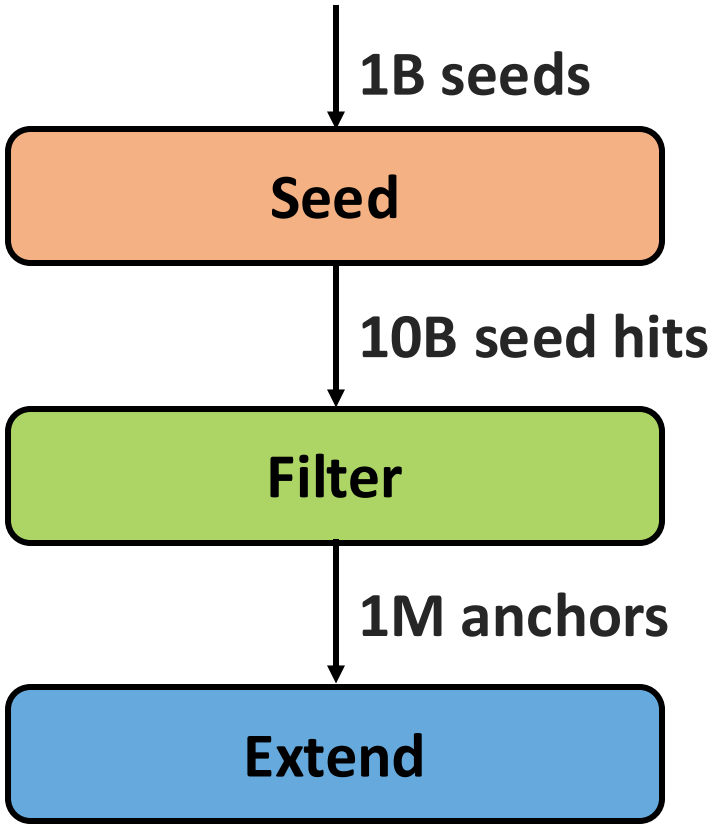
PRINCETON UNIVERSITY

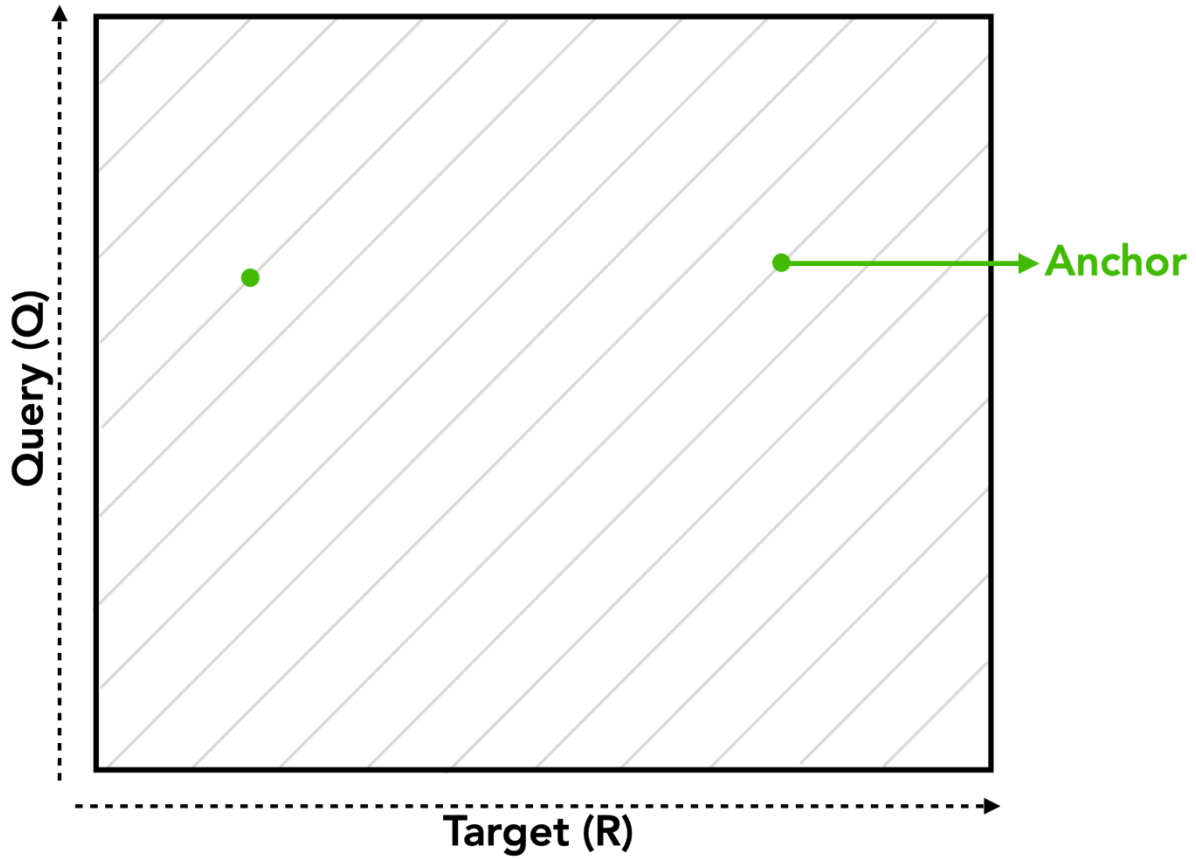
Filtering aligns ~100bp around seed hits

R	A	A	G	T	C	A	A	T
Q	A	T	G	T	A	T	T	C
	2	-1	2	2	-1	-1	-1	-1
Cumulative Score	2	1	3	5	4	3	2	1
Max Score	2	2	3	5	5	5	5	5
Score Difference	0	1	0	0	1	2	3	4



High-scoring Segment Pair reduced to Anchor



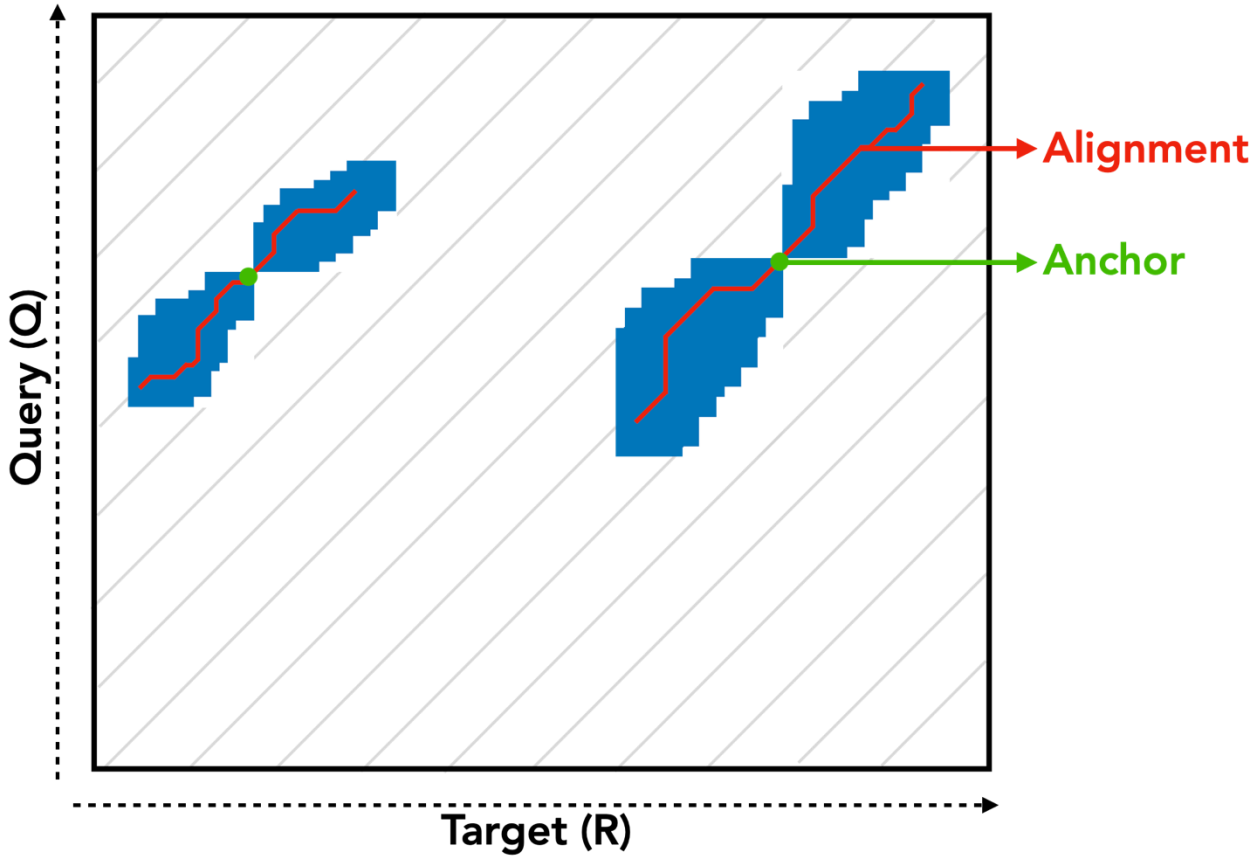


PRINCETON UNIVERSITY

Extension results in the final alignments

Dynamic Programming Equations

$$
\begin{aligned}
I(i,j) &= \max\{H(i,j-1)-o,\ I(i,j-1)-e\} \\
D(i,j) &= max\{H(i-1,j)-o,\ D(i-1,j)-e\} \\
H(i,j) &= max\begin{cases} 0 \\ I(i,j) \\ D(i,j) \\ H(i-1,j-1)+W(r_i,q_j) \end{cases}
\end{aligned}
$$

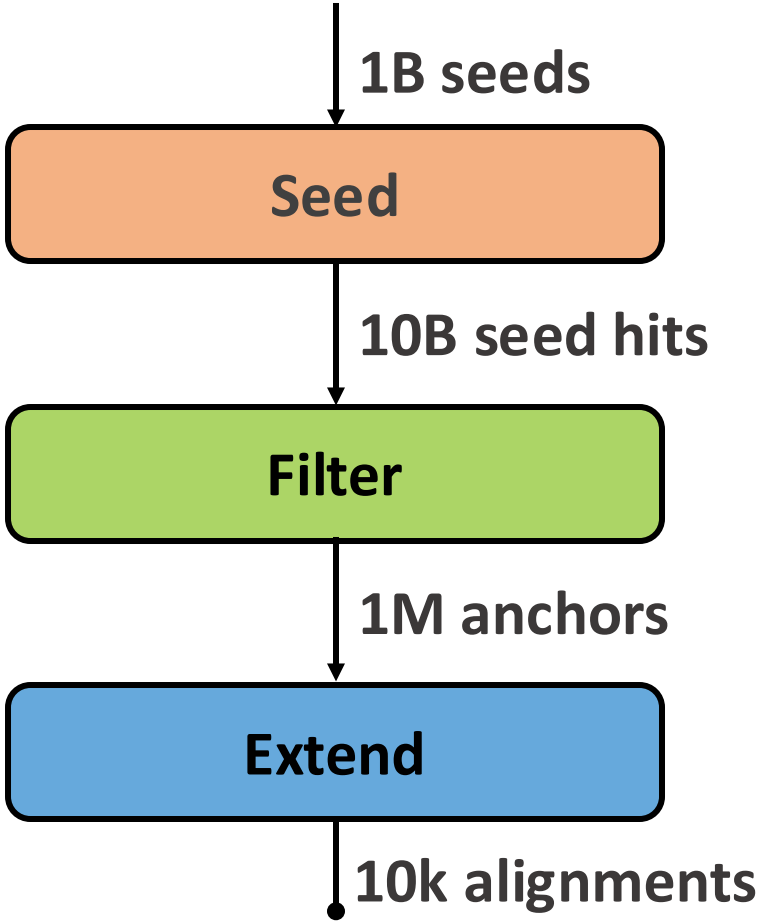


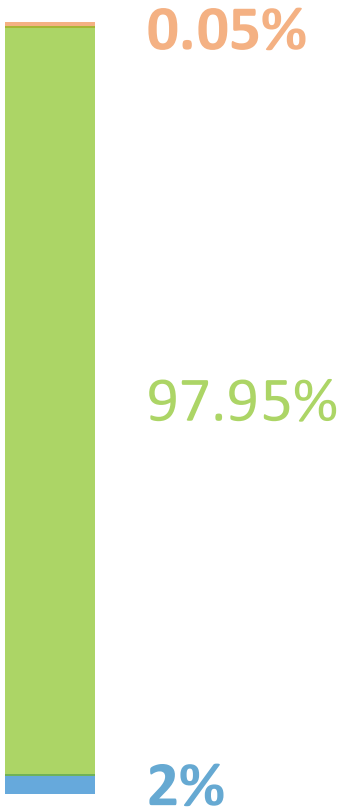
Alignment

```
human     1 AGGTAGCAAGGGGACAGGAG----------GGGCC
mouse     1 AGGCAGGAGGGGGACAGGAAACAGTCTGCAGAGGC

human    26 AGGAGGGGACAGGAG-TGGCCAGGAGTGGCCAGGA
mouse    36 AGGAGGGGGCAGGAAACAGCCTGCAGGGGT-AGGA

human    60 GGGGGCAGG
mouse    70 GGGGGCAGG
```


PRINCETON UNIVERSITY

Filtering stage dominates the runtime





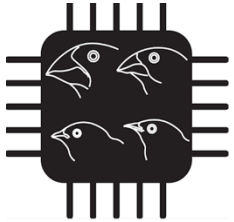
Runtime distribution per stage (LASTZ)

PRINCETON UNIVERSITY

Enabling biologists to leverage sequencing

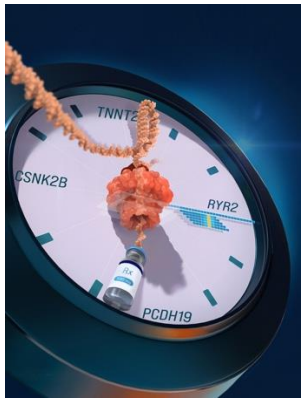
Comparative Genomics

SegAlign: A Scalable GPU-based whole genome aligner
[Supercomputing Conference]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments
[HPCA]

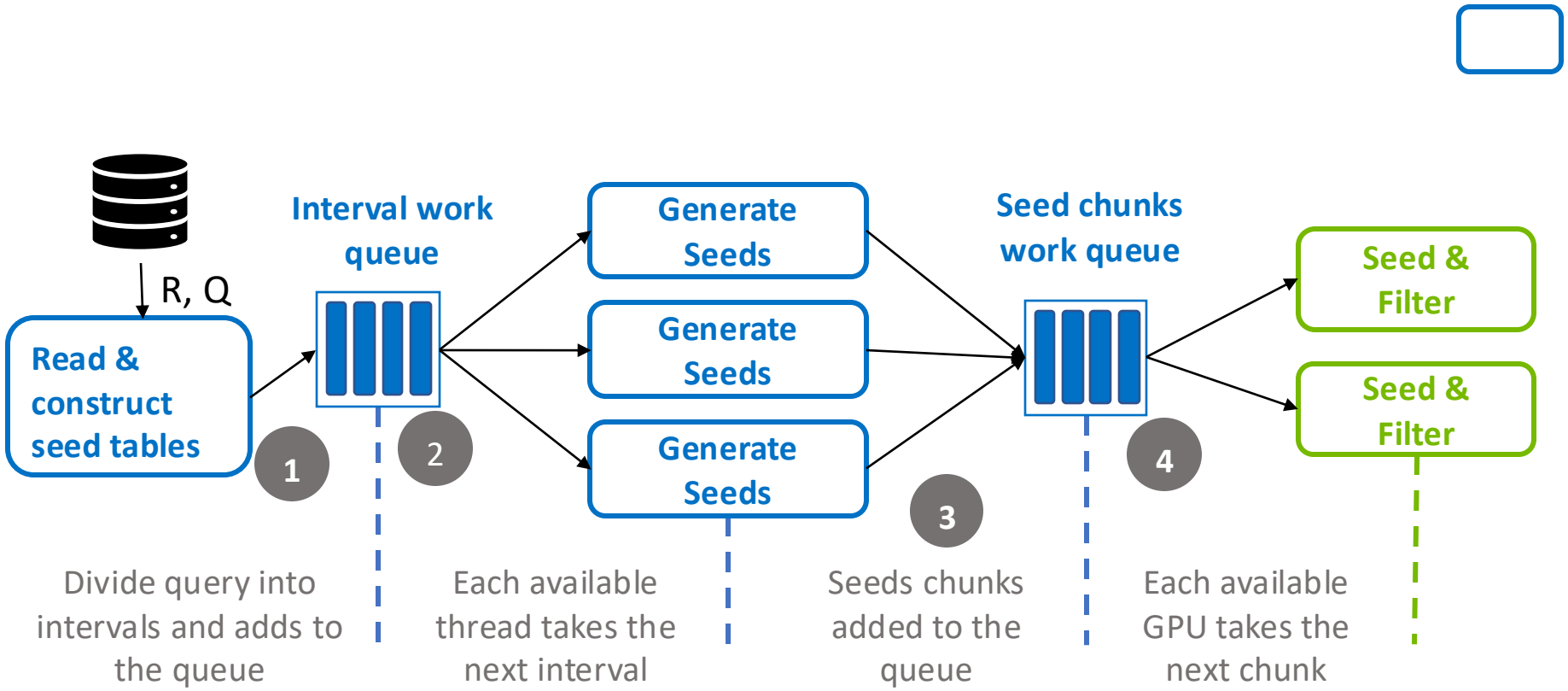
Clinical Genomics



An ultra-rapid workflow for clinical whole genome sequencing
[New England Journal of Medicine Nature Biotechnology]

PRINCETON UNIVERSITY

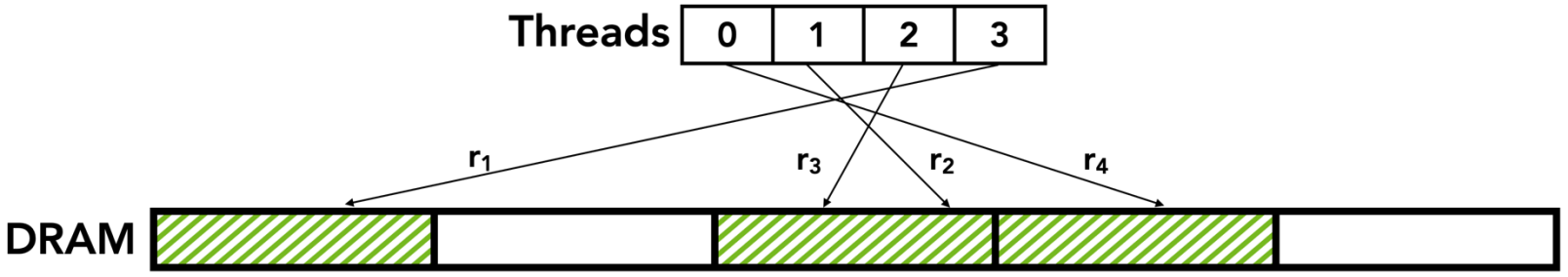
SegAlign system for single chromosome pair



PRINCETON UNIVERSITY

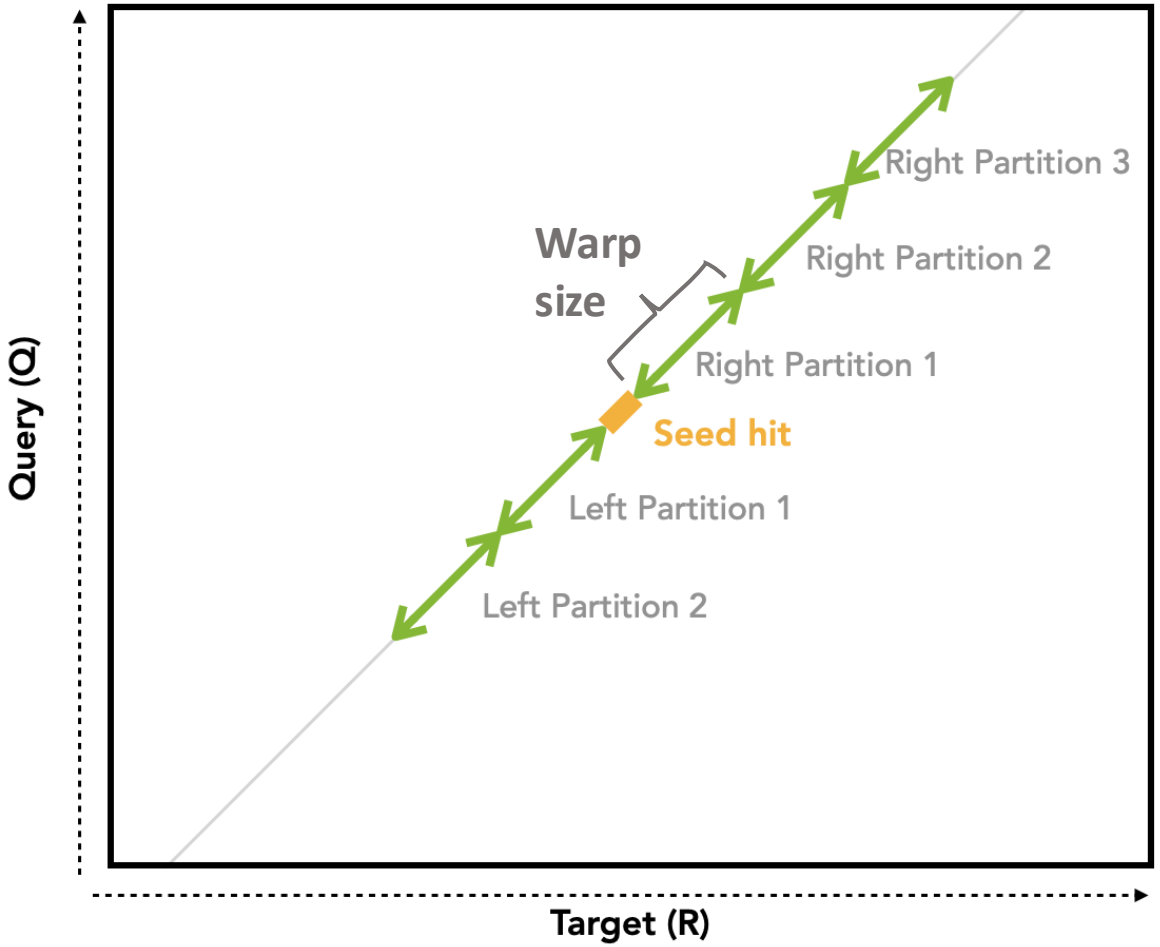
Naïve approach allocates 1 seed hit/thread

1. Considerably varying seed hit positions -> inefficient uncoalesced memory accesses within a warp
   [Warp - basic unit for scheduling execution and memory accesses]



2. Divergent branches within a warp due to the dynamic X-drop condition for each thread

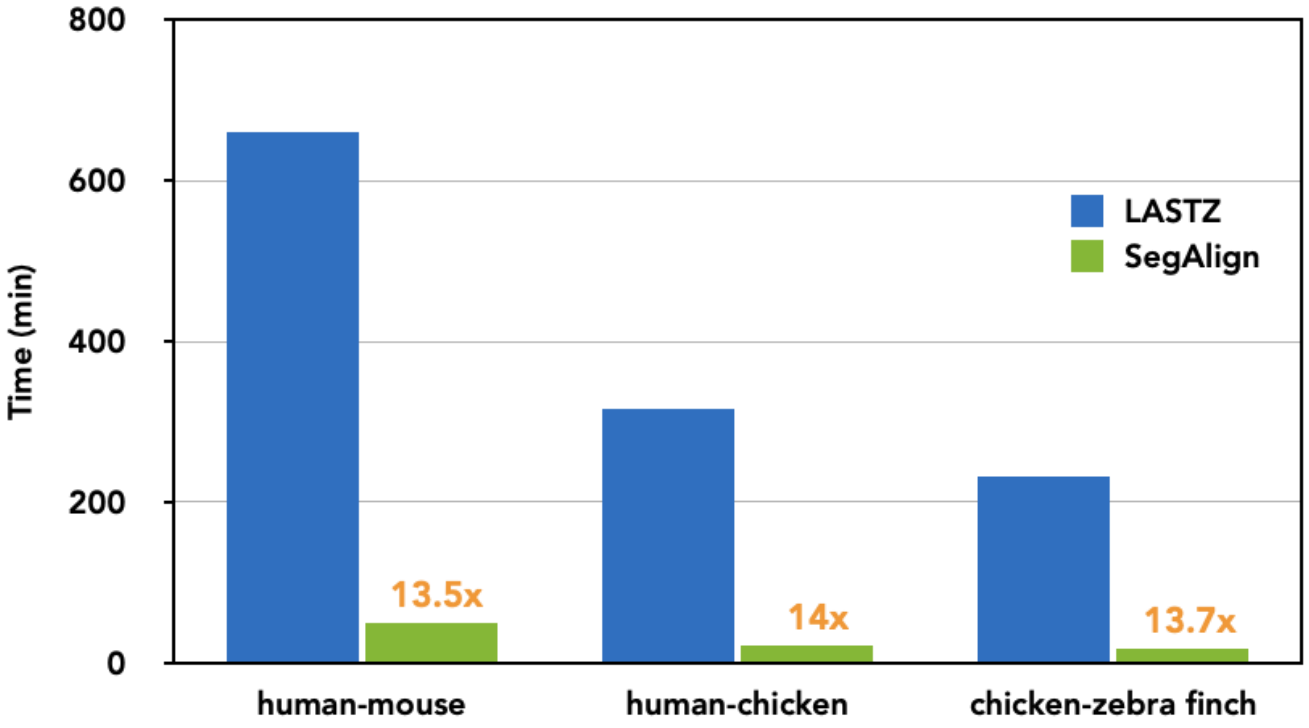
SegAlign allocates 1 seed hit/warp

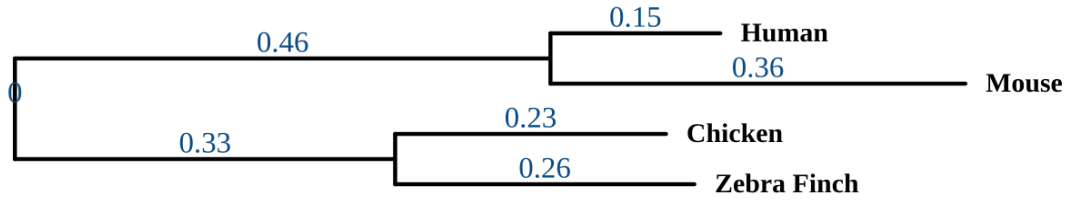


1. Efficient bandwidth gains with coalesced memory accesses

2. Exploiting data locality within each partition using parallel prefix scan

13.5x-14x speedup at ~2x cost improvement





	HW config	Cost/hour
LASTZ	96 vCPUs	$4.08
SegAlign	8 V100 GPU 64 vCPUs	$24.48

PRINCETON UNIVERSITY

SegAlign's Ungapped extension kernel now in NVIDIA GenomeWorks library

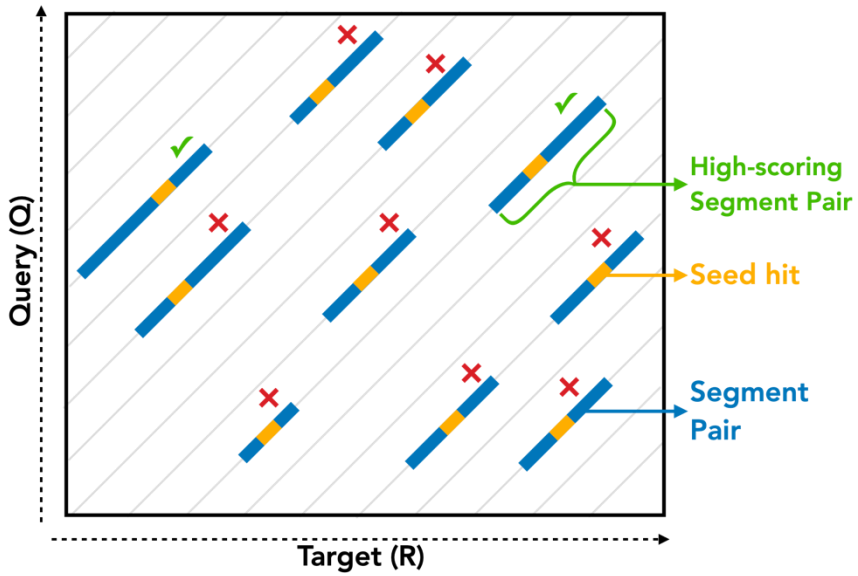
https://github.com/clara-parabricks/GenomeWorks

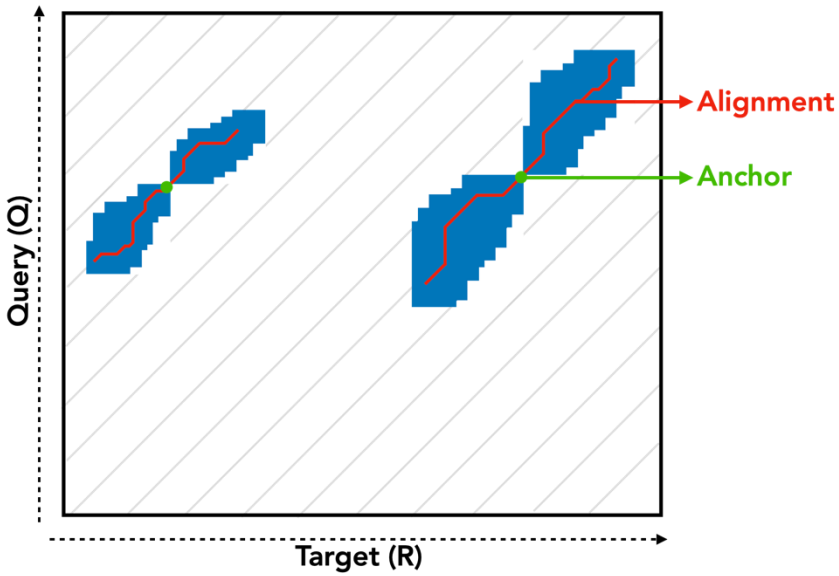
GenomeWorks

Overview

GenomeWorks is a GPU-accelerated library for biological sequence analysis. This section provides a brief overview of the different components of GenomeWorks. For more detailed API documentation please refer to the documentation.

Seed-Ungapped Filter-Extend isn't sensitive enough





Improved search heuristics find 20 000 new alignments between human and mouse genomes

Martin C. Frith ✉, Laurent Noé

Nucleic Acids Research, Volume 42, Issue 7, 1 April 2014, Page e59,
https://doi.org/10.1093/nar/gku104
Published: 31 January 2014 **Article history ▾**

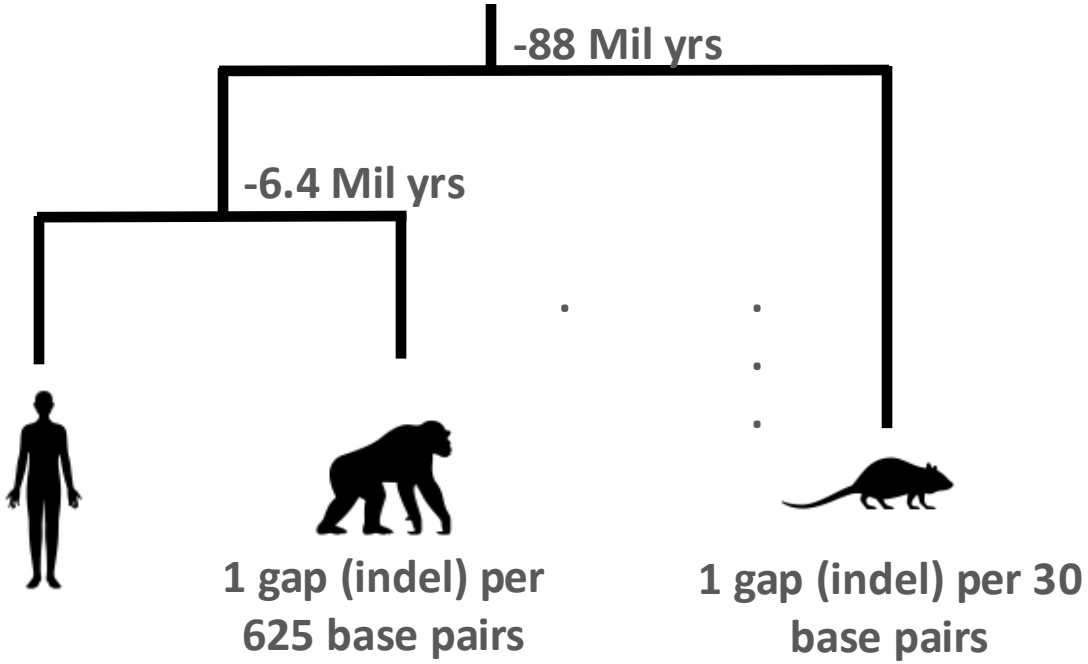
Increased alignment sensitivity improves the usage of genome alignments for comparative gene annotation

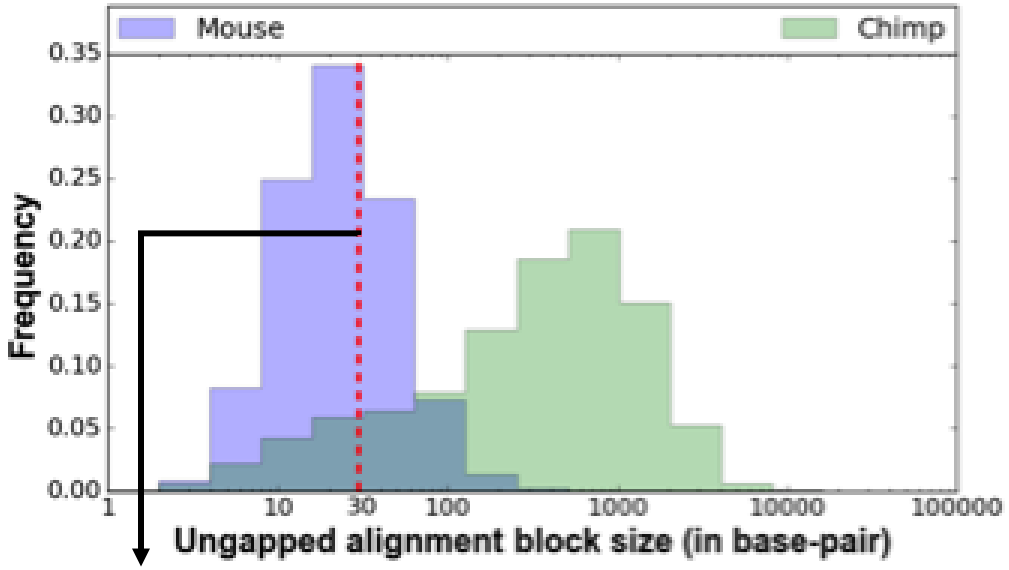
Virag Sharma, Michael Hiller ✉

Nucleic Acids Research, Volume 45, Issue 14, 21 August 2017, Pages 8369–8377,
https://doi.org/10.1093/nar/gkx554
Published: 21 June 2017 **Article history ▾**

PRINCETON UNIVERSITY

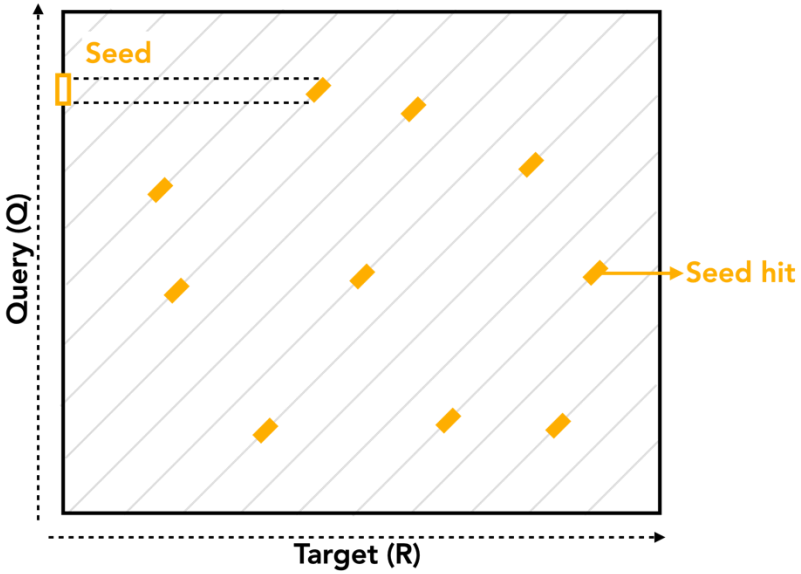
Increasing indel frequency => increasing need for gapped filtering

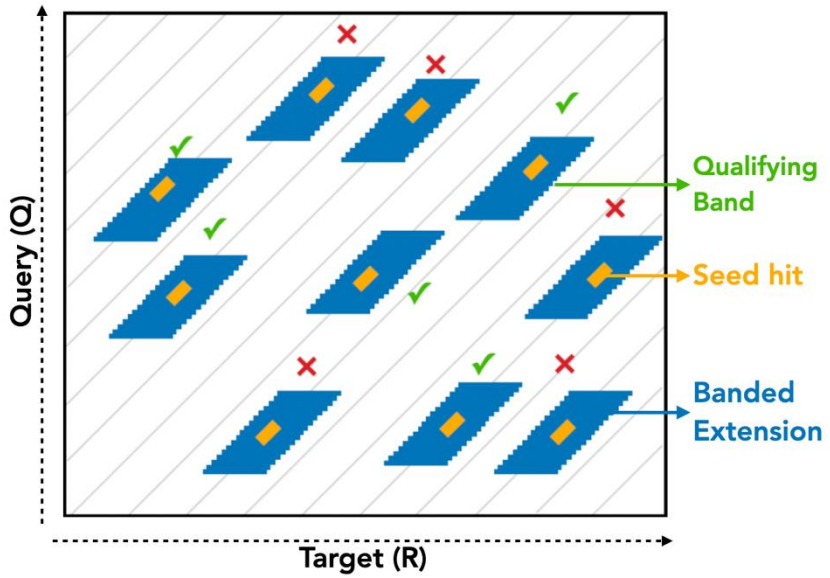


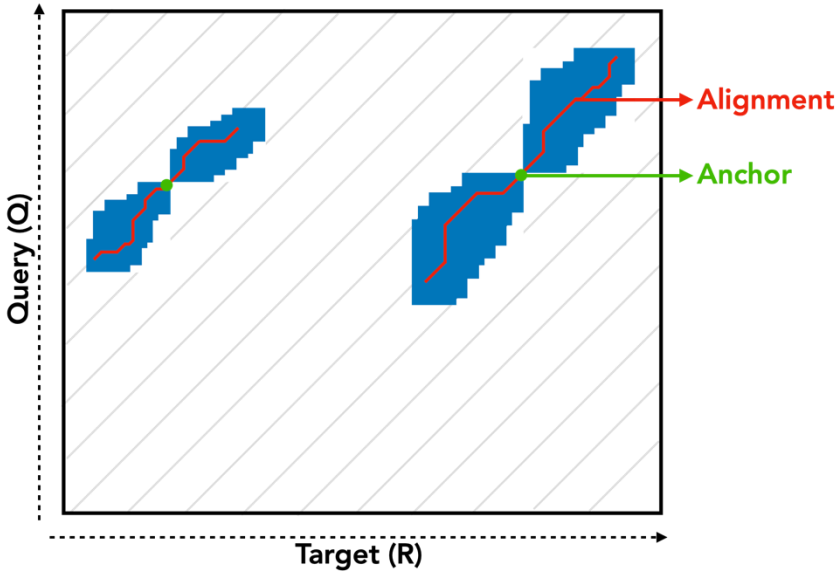


PRINCETON UNIVERSITY

Seed-Gapped Filter-Extend







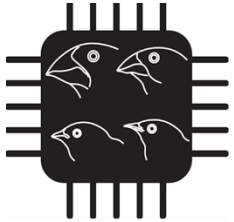
Replacing ungapped filtering by gapped filtering slows down the software by 200x!

PRINCETON UNIVERSITY

Enabling biologists to leverage sequencing

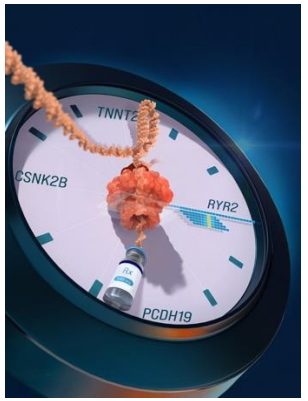
Comparative Genomics

SegAlign: A Scalable GPU-based whole genome aligner
[Supercomputing Conference]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments
[HPCA]

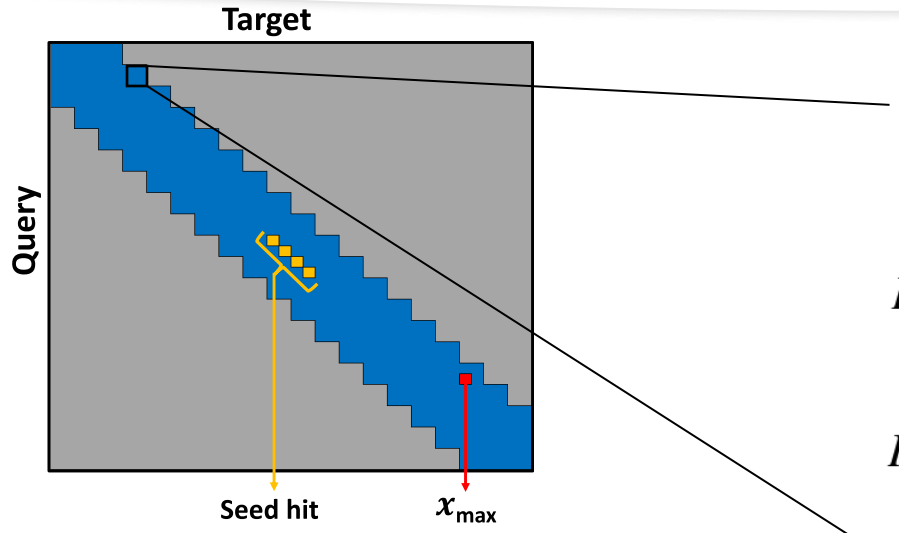
Clinical Genomics



An ultra-rapid workflow for clinical whole genome sequencing
[New England Journal of Medicine Nature Biotechnology]

PRINCETON UNIVERSITY

Specialized Operations

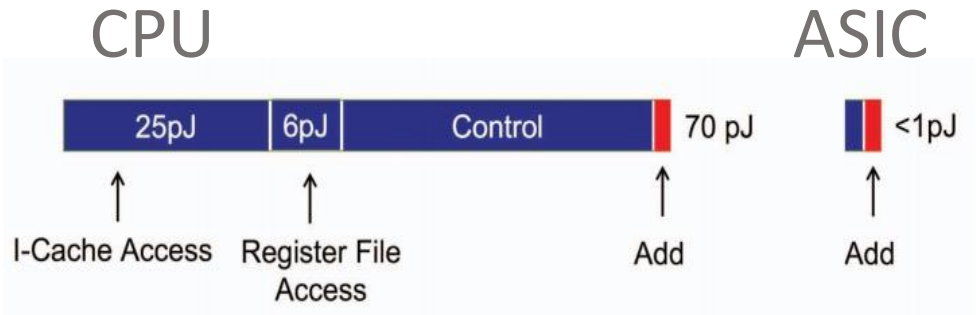


Per cell computation

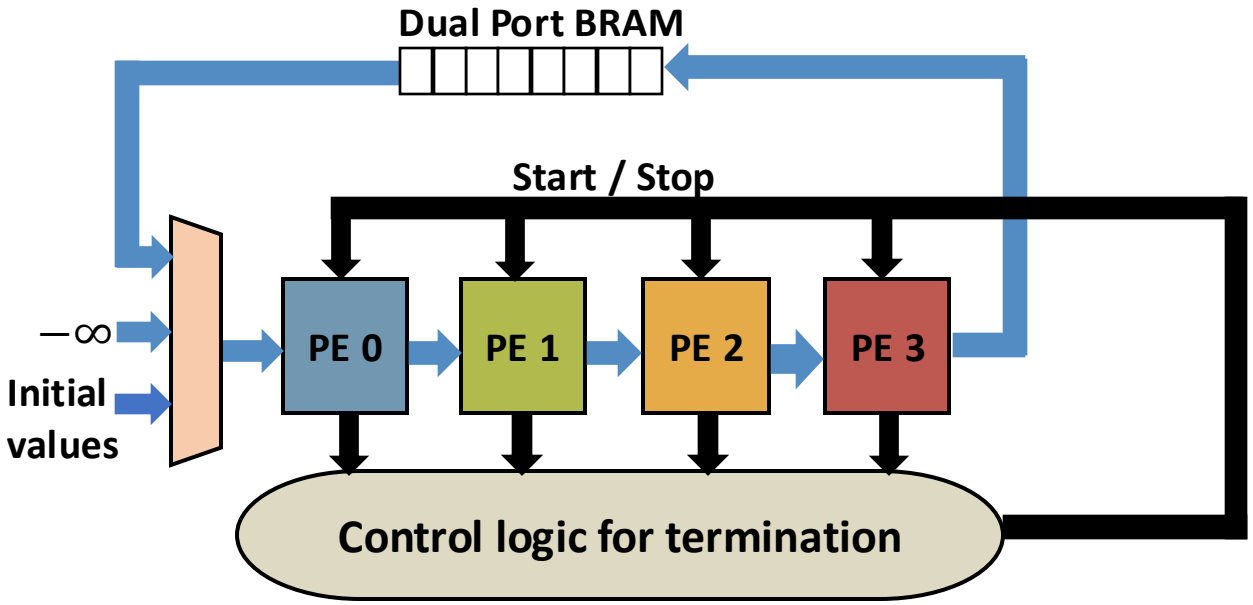
$$
\begin{aligned}
I(i,j) &= \max\{H(i,j-1)-o,\ I(i,j-1)-e\} \\
D(i,j) &= max\{H(i-1,j)-o,\ D(i-1,j)-e\} \\
H(i,j) &= max\begin{cases} 0 \\ I(i,j) \\ D(i,j) \\ H(i-1,j-1)+W(r_i,q_j) \end{cases}
\end{aligned}
$$

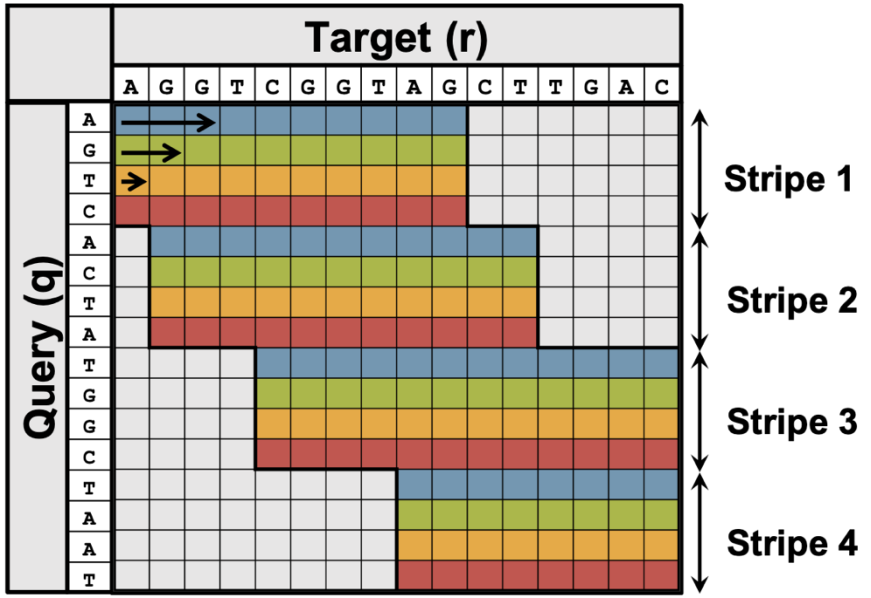
On 14nm CPU
35 ALU ops, 15 load/store
37 cycles
81nJ

On 40nm Special Unit
1 cycle *(37x speedup)*
3.1pJ *(26,000x efficiency)*
300fJ for logic *(remainder is memory)*



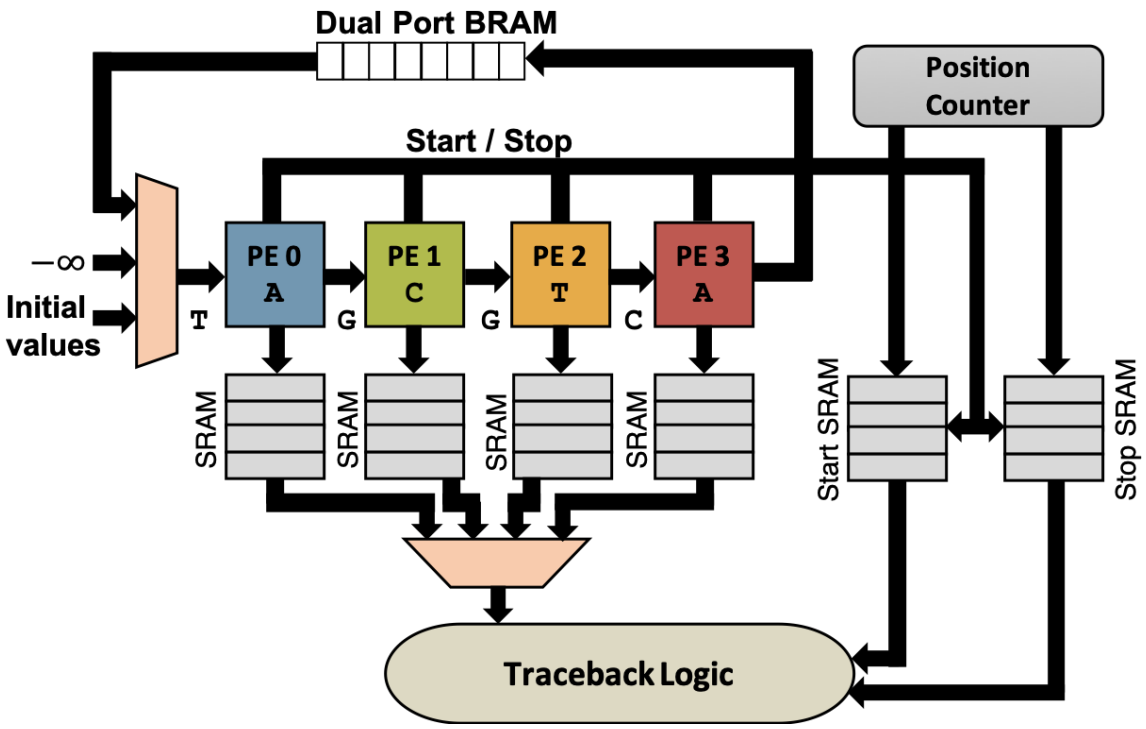
Exploiting inner-loop parallelism with systolic arrays

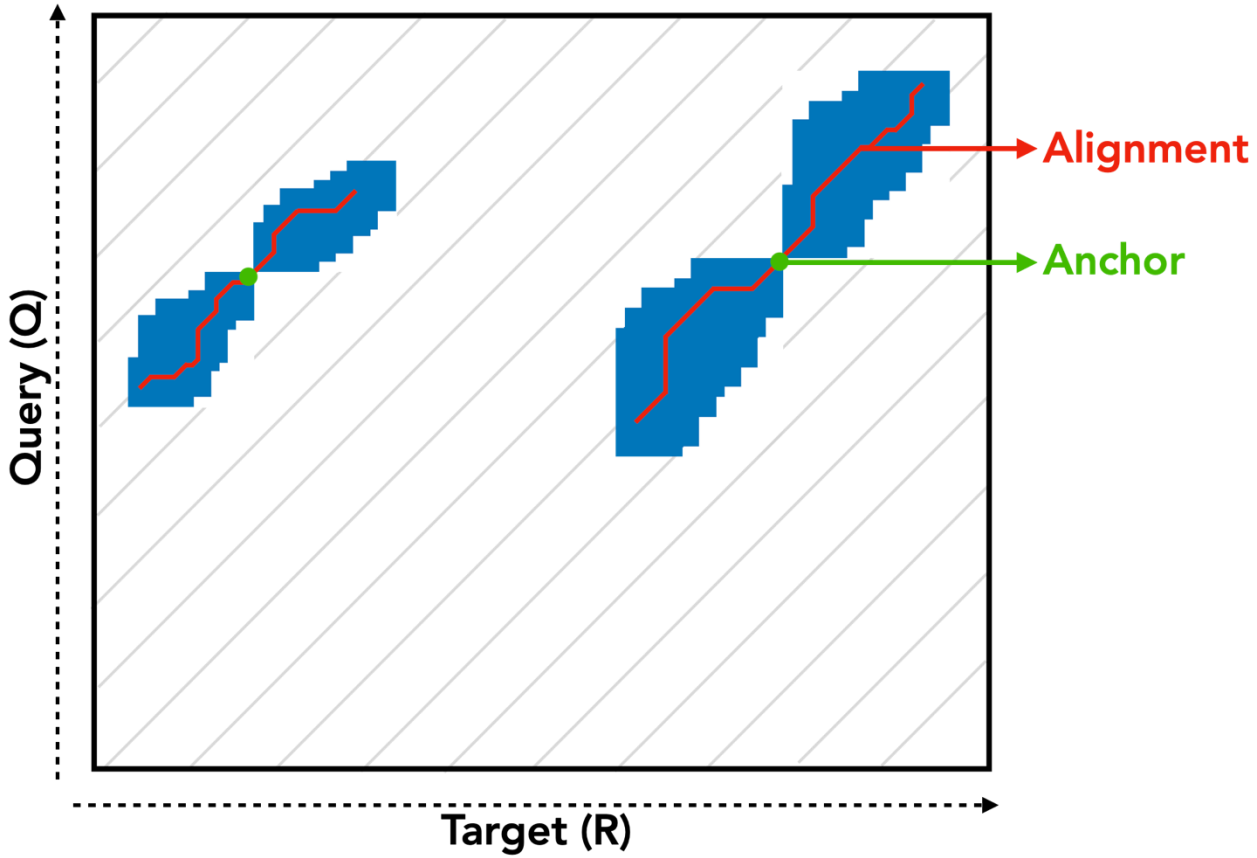




PRINCETON UNIVERSITY

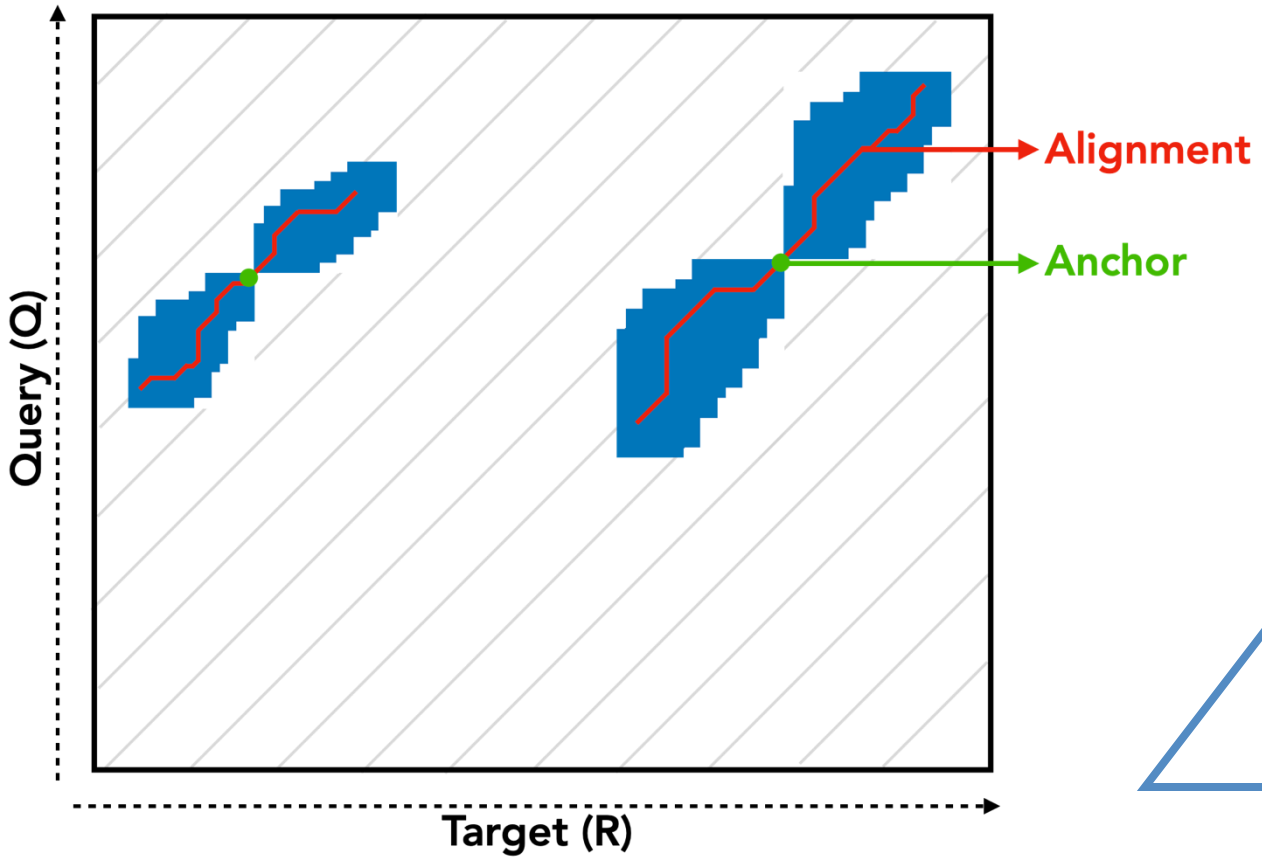
Can we adapt this architecture for extension?

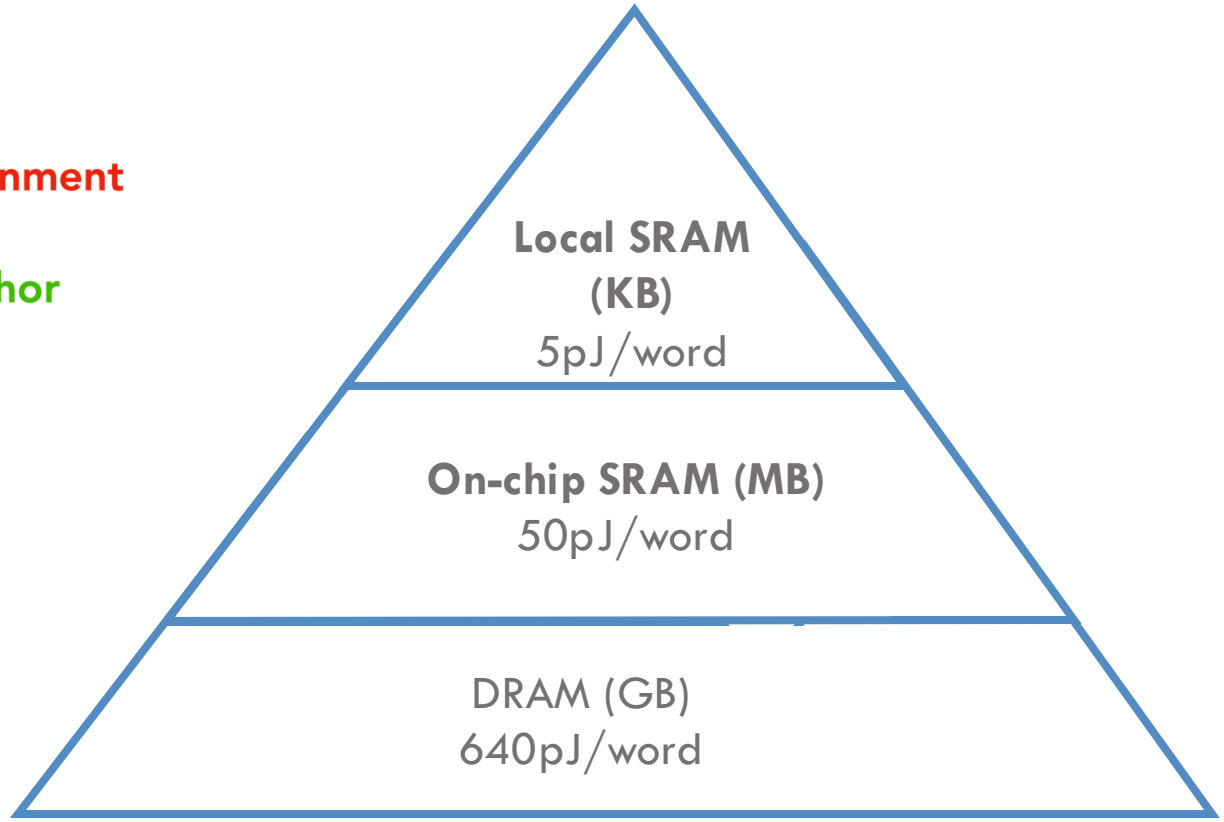




PRINCETON UNIVERSITY

Utilizing local memory – size is prohibitive for larger compute

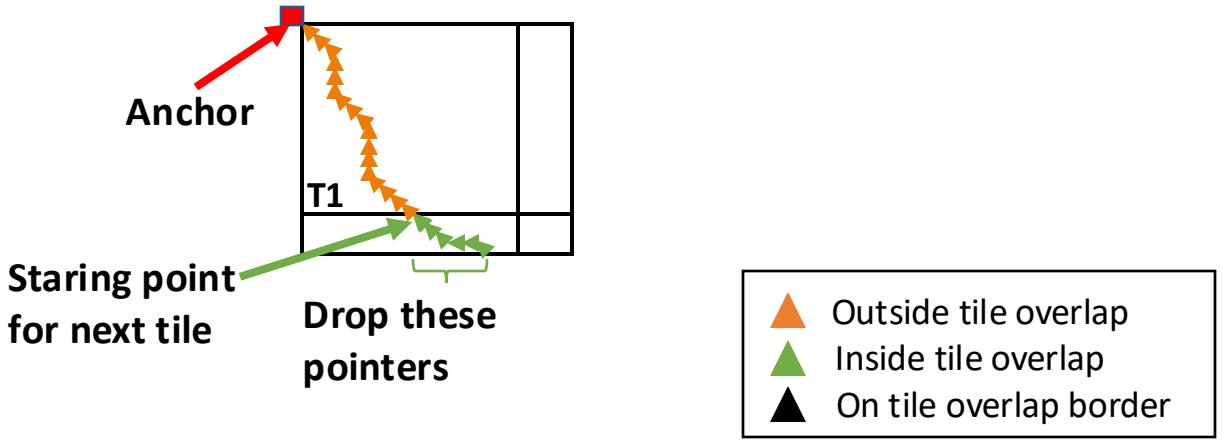




PRINCETON UNIVERSITY

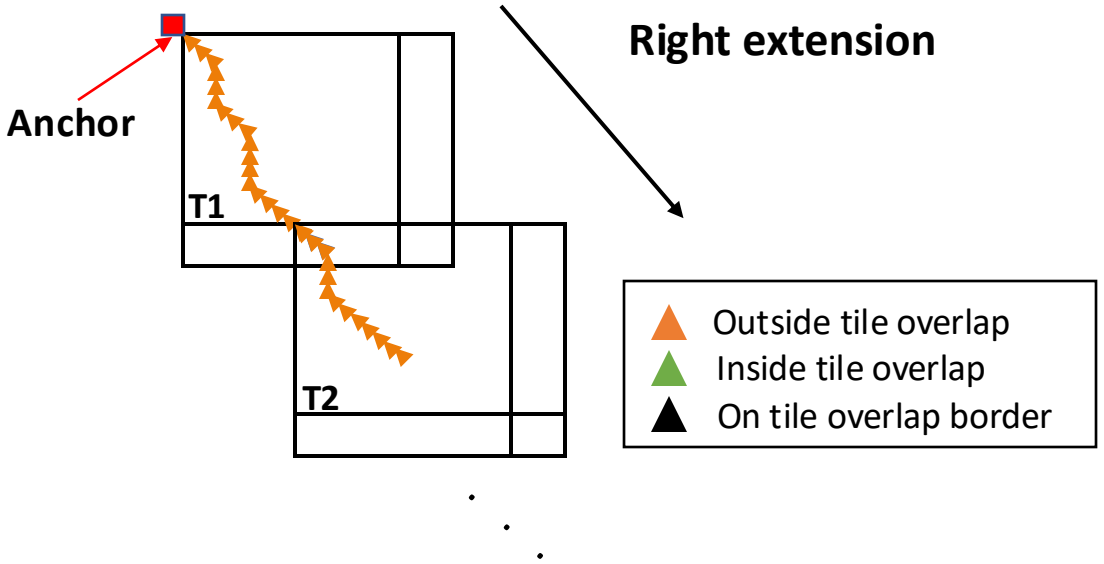
Overlapped extension uses constant on-chip memory

- Tiled (tile size T, overlap O) implementation inspired by GACT in Darwin
- Origin of the next tile lies at the intersection of the current traceback path with the overlap



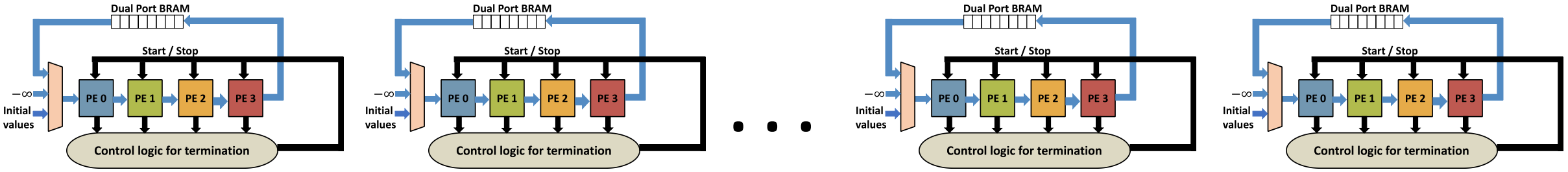
Overlapped extension uses constant on-chip memory

- Extension along a direction continues until a tile is encountered with a non-positive maximum score

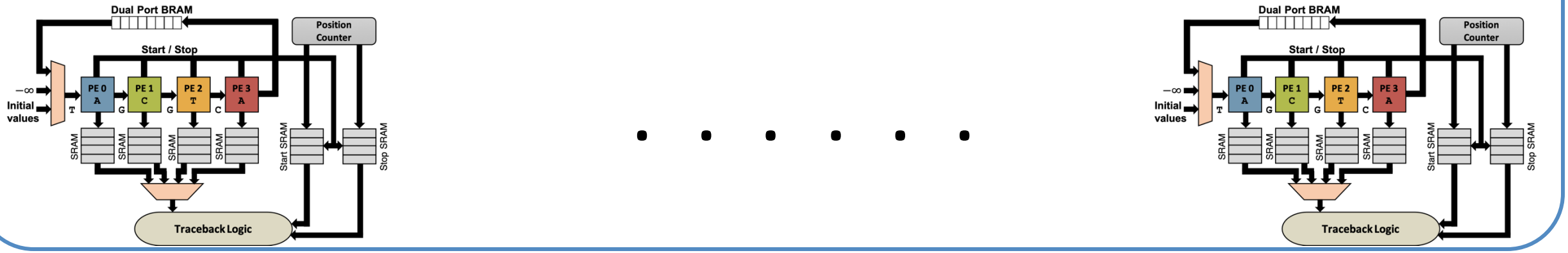


Banded + extension arrays = Outer-loop parallelism

Banded arrays

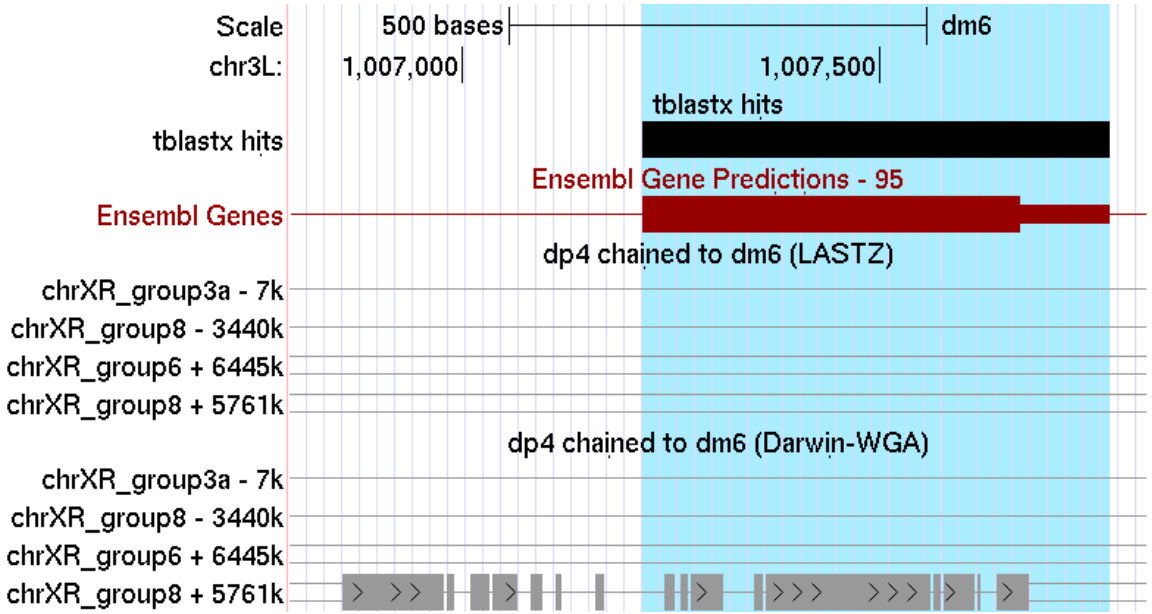


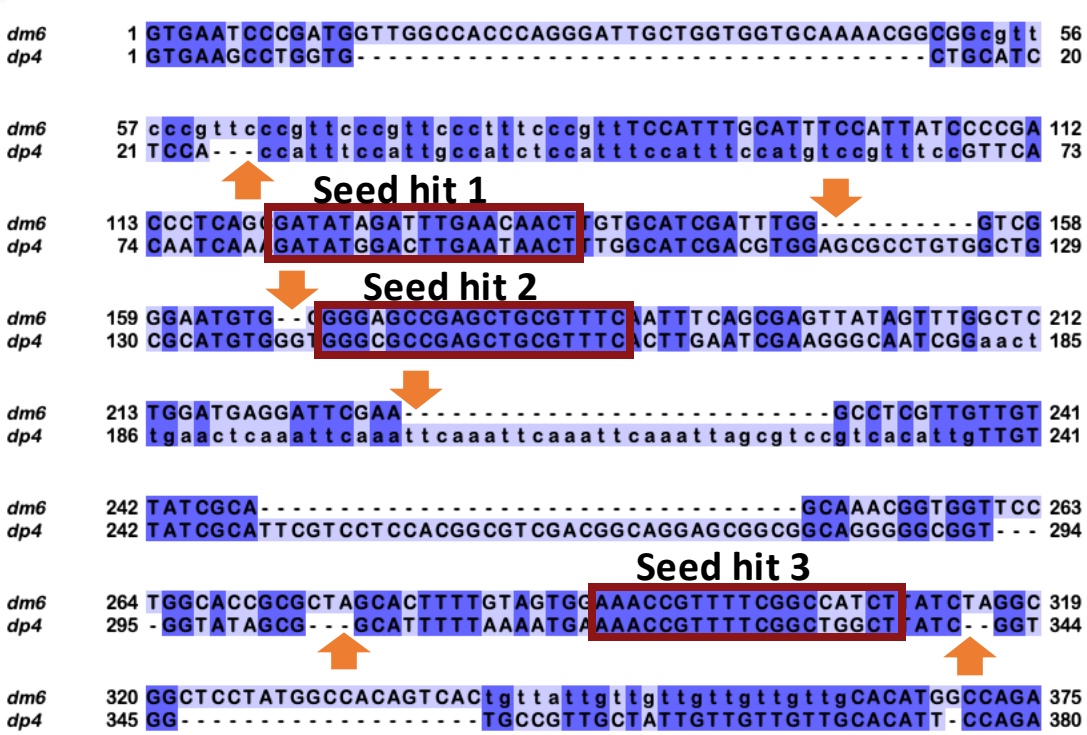
Extension arrays



PRINCETON UNIVERSITY

Darwin-WGA finds genes that LASTZ does not





Indels (shown by arrows) around each seed hit – dropped by ungapped filtering (*LASTZ*) but retained by gapped filtering (*Darwin-WGA*)

PRINCETON UNIVERSITY

Darwin-WGA is more sensitive than LASTZ

Species pair	Top-10 Alignment Chain Scores	Matching Base-pairs within Alignments	Number of Aligning Exons (protein-coding genes)
dm6-droSim1	+0.03%	1.25x	+0.20%
dm6-droYak2	+0.05%	1.41x	+0.09%
dm6-dp4	+1.86%	1.42x	+0.41%
ce11-cb4	+5.73%	3.12x	+2.70%

Represent orthologous sequences (derived from "speciation")

Represent paralagous sequences (derived from "duplication")

Represent functionally relevant orthologous sequences, under some selective pressure (at least in the target species)

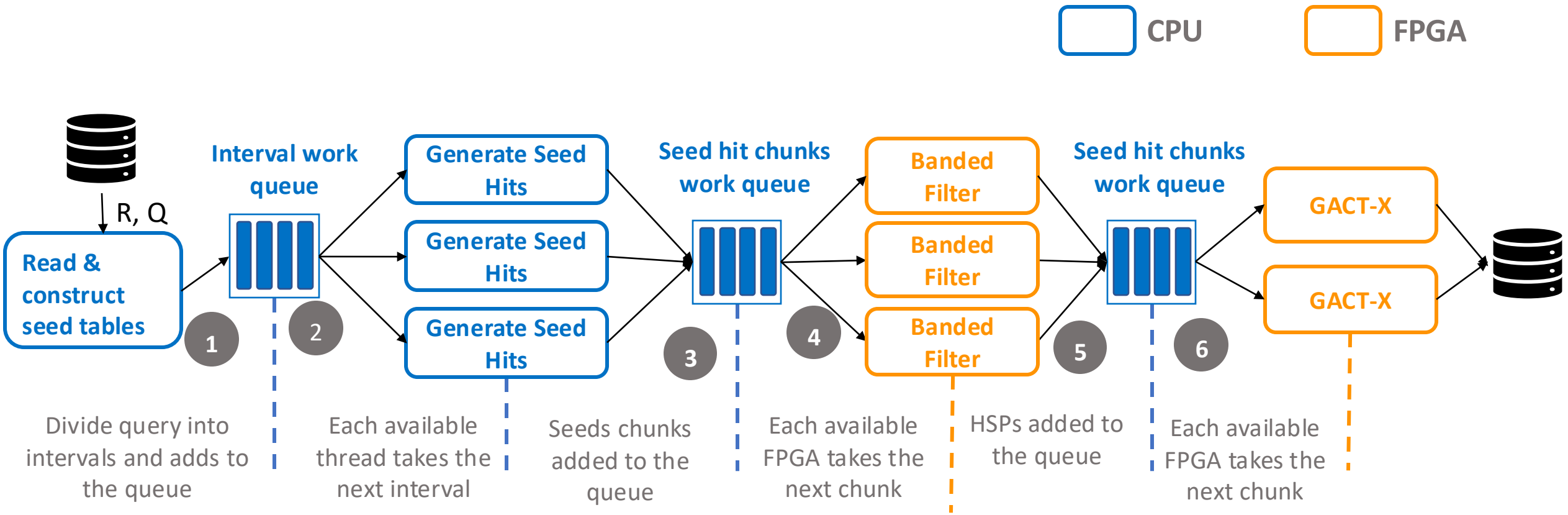
False positive rate (2-mer shuffled genome): 0.0007%

Darwin-WGA (FPGA) is 20x faster than iso-sensitive software

	LASTZ (CPU)	Iso-sensitive software
Darwin–WGA (FPGA)	0.1x (slowdown)	20x (speed and cost)

	HW config	Cost per hour
LASTZ	36 vCPUs	$1.59
Darwin-WGA (FPGA)	1 Xilinx Virtex Ultrascale+ FPGA + 8vCPUs	$1.65

While Darwin-WGA uses the same software framework, it features a different interface



PRINCETON UNIVERSITY

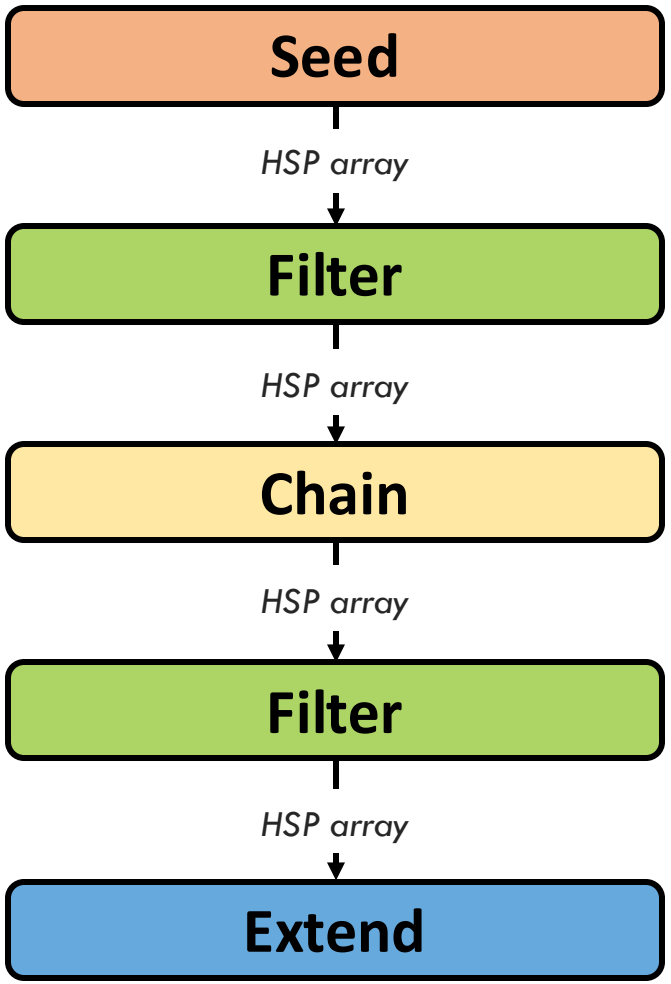
How about a new language?

Bioinformaticians focus on filters and recurrences.

We provide a domain-specific language to manage load balancing and scheduling.

Architects can then optimize these stages for performance.

Introducing FILTR



Pipeline: sequence of producer-consumer relationships as self-contained, reusable stages

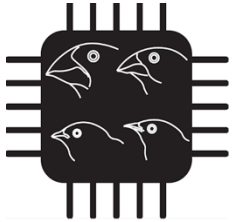
HSP: encapsulates intermediate hits/points along with metadata such as scores, positions, and alignment status

Dataflow control: built-in balancing policies or define custom strategies

Enabling biologists to leverage sequencing

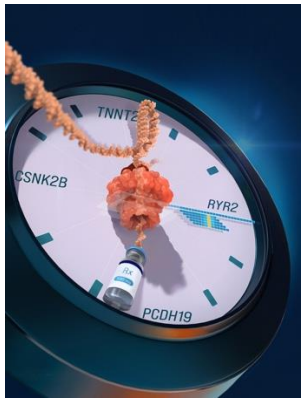
Comparative Genomics

SegAlign: A Scalable GPU-based whole genome aligner
[Supercomputing Conference]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments
[HPCA]

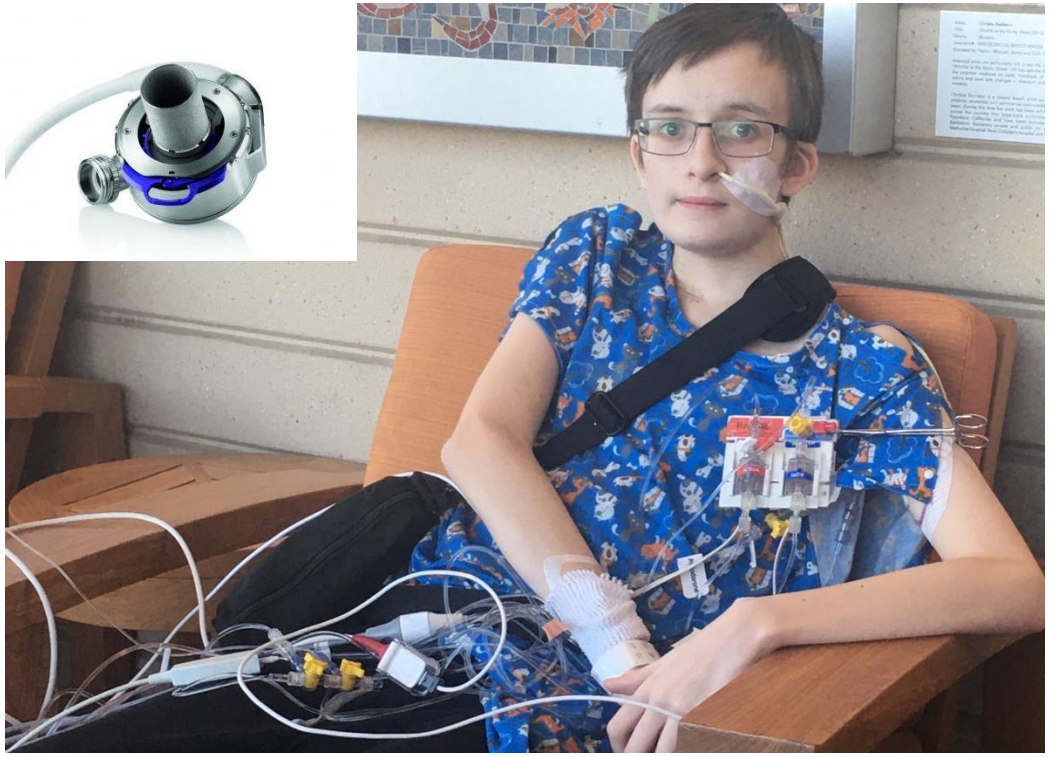
Clinical Genomics



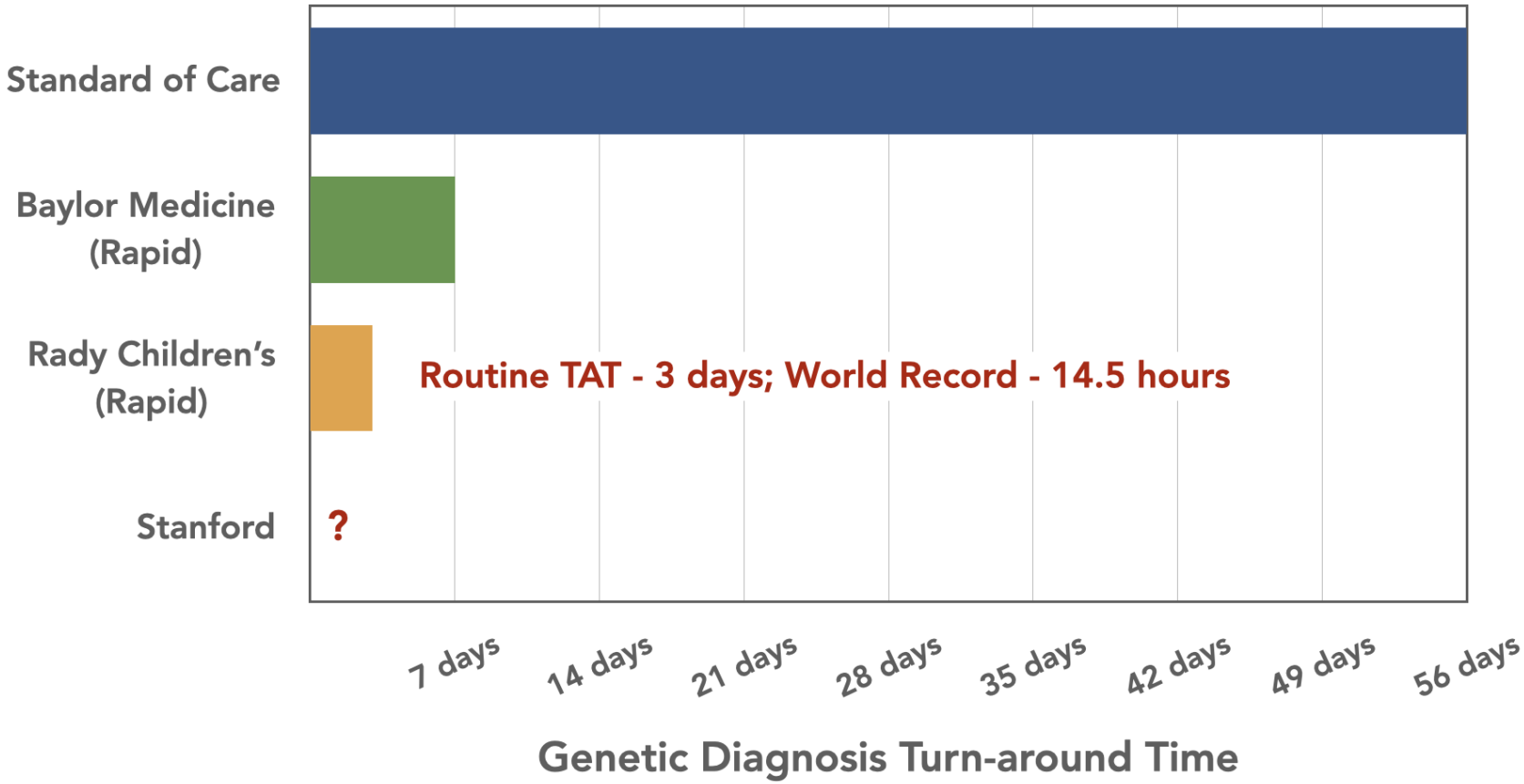
An ultra-rapid workflow for clinical whole genome sequencing
[New England Journal of Medicine Nature Biotechnology]

PRINCETON UNIVERSITY

- 13 years
- Two-week history of dry cough, decreased appetite, chest pain and severe fatigue
- On admission to hospital in Oregon: weak heart
- Rapid deterioration
- Airlifted to Stanford hospitals
- Placed on ECMO
- Possible causes included potentially reversible (myocarditis) and irreversible (genetic heart disease)
- How to make transplant decision?
  - Biopsy?
  - Genetic testing?

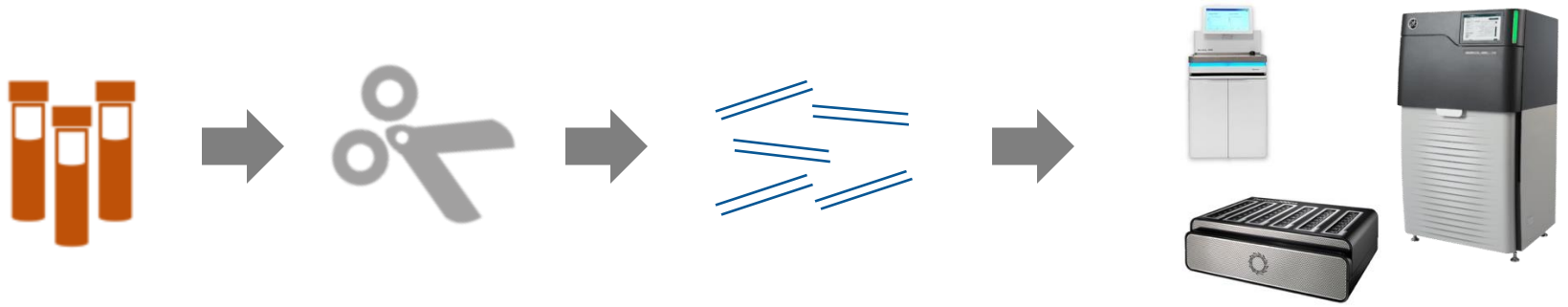


State-of-the-art turn-around times

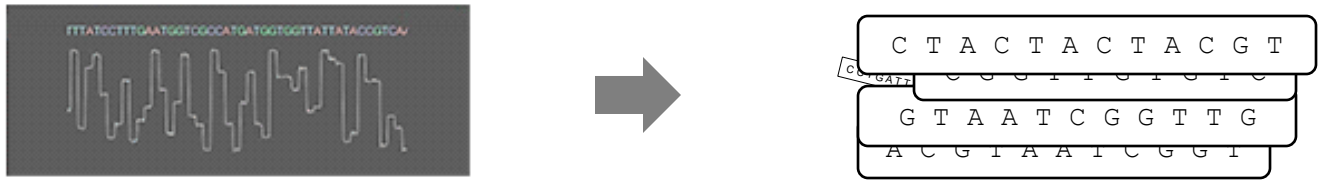


PRINCETON UNIVERSITY

DNA Sequencing

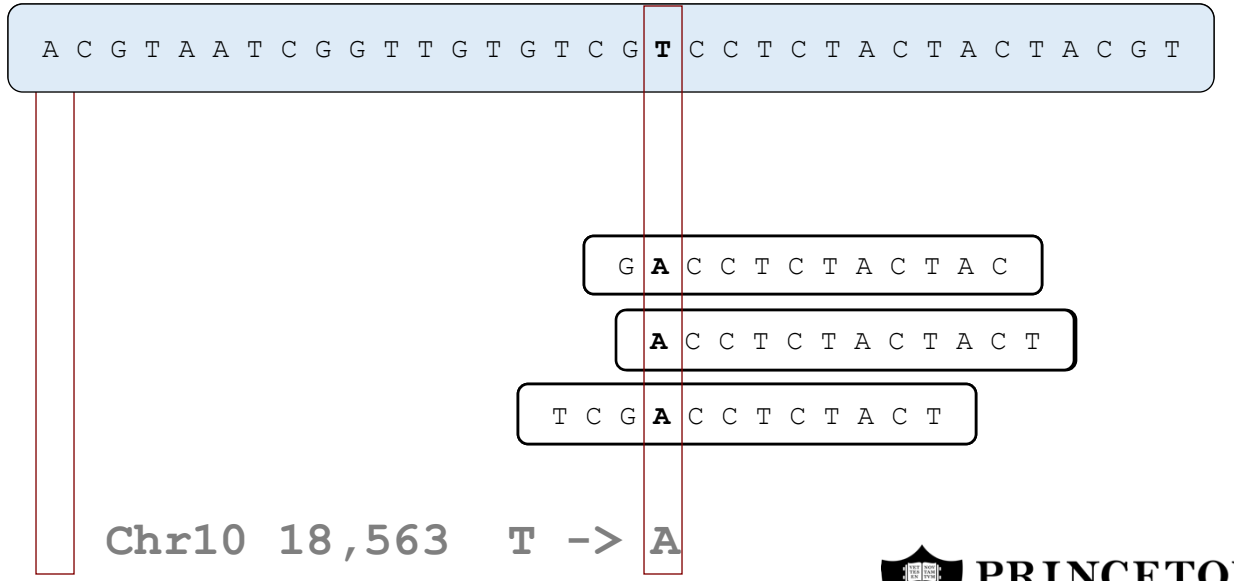


Base calling



Alignment

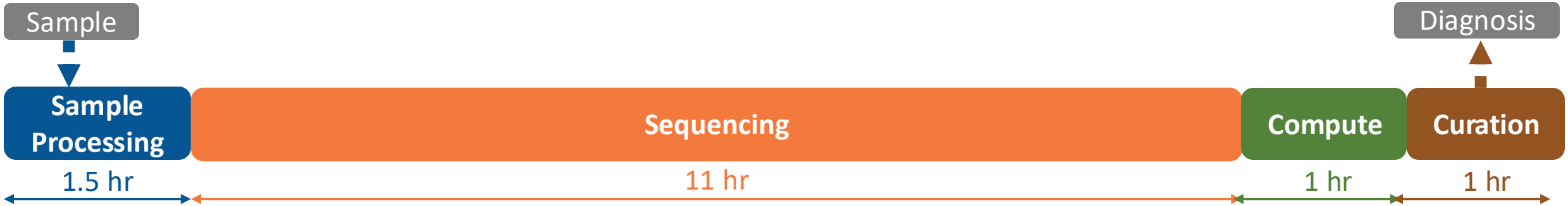
Human Reference Genome



Variant Discovery

PRINCETON UNIVERSITY

Existing fastest turn-around in 14.5 hours



Illumina HiSeqX

DRAGEN (FPGA)

PRINCETON UNIVERSITY

Compute dominates the new pipeline with 30-hour turnaround time

Nanopore PromethION

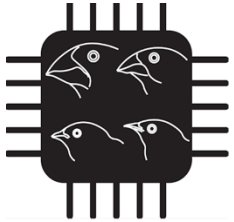
- Longer reads => more complicated computation
- Higher error rate => double the amount of sequencing
- No custom ASIC/FPGA

PRINCETON UNIVERSITY

Enabling biologists to leverage sequencing

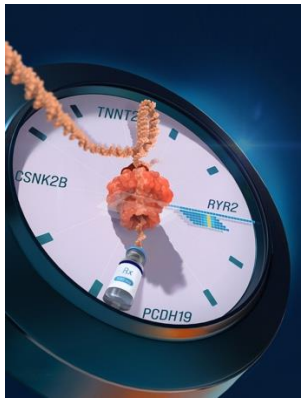
Comparative Genomics

SegAlign: A Scalable GPU-based whole genome aligner
[Supercomputing Conference]



Darwin-WGA: A fast and highly sensitive co-processor for whole genome alignments
[HPCA]

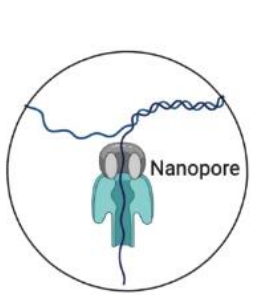
Clinical Genomics



An ultra-rapid workflow for clinical whole genome sequencing
[New England Journal of Medicine Nature Biotechnology]

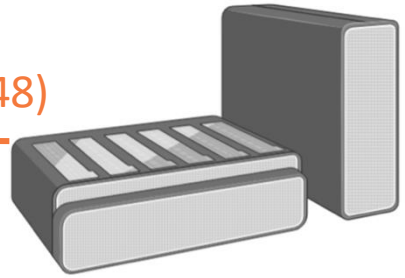
PRINCETON UNIVERSITY

Traditional computational pipeline

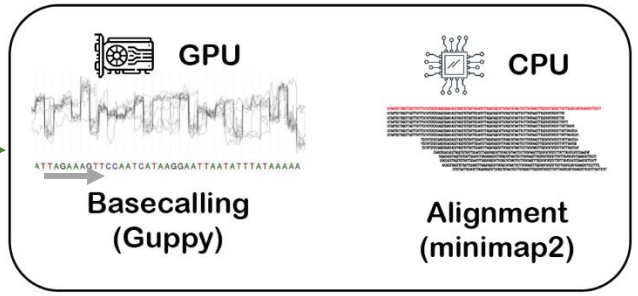


Sequencing

Sequencing Unit (PromethION 48)



Compute Tower (4x Tesla V100)



HG002
218 Gb
2.3 TB fast5

1.5 hr

Sequencing

Base calling

17.5 hr

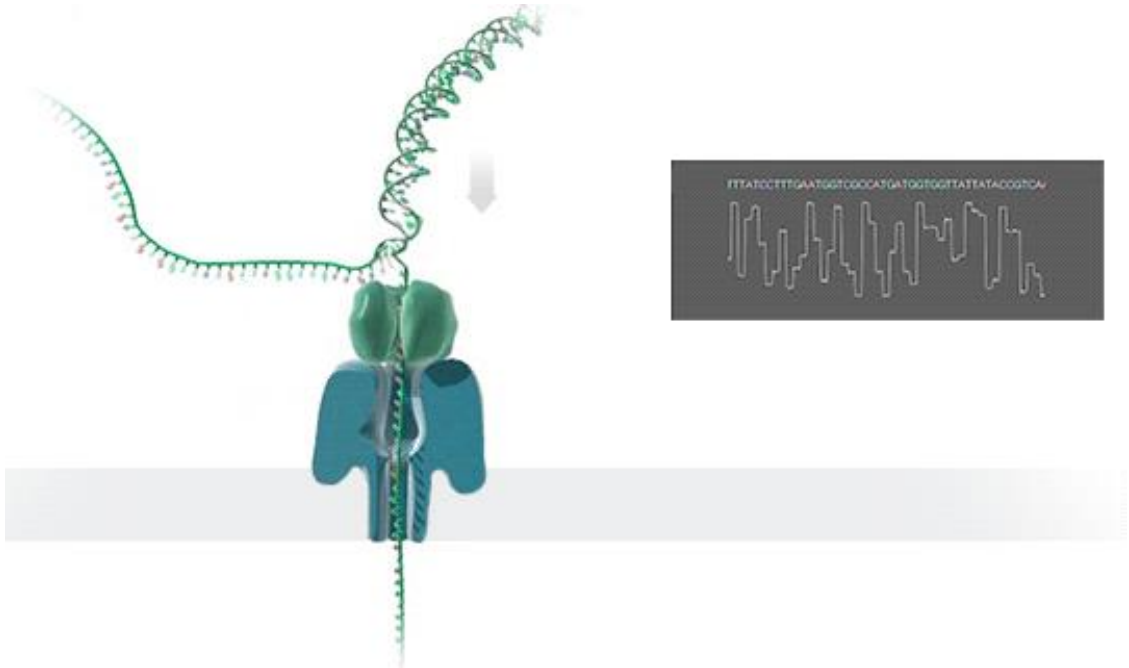
Alignment

2.5 hr

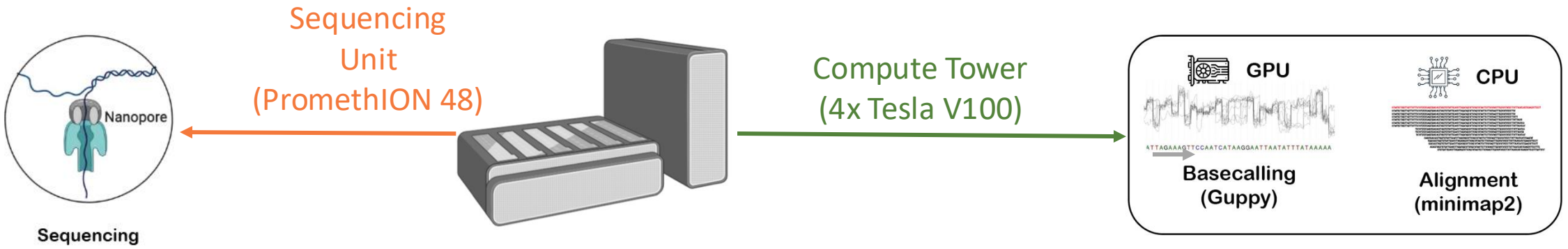
PRINCETON UNIVERSITY

Nanopore's "real-time" advantage

- Signal files are generated as soon as the strand passes through the nanopore
- Ideally, we can start base calling right away



Modified pipeline – overlap base calling and sequencing



HG002
218 Gb
2.3 TB fast5

1.5 hr

Sequencing

Base calling

Alignment

16 hr

2.5 hr

PRINCETON UNIVERSITY

Moving to cloud

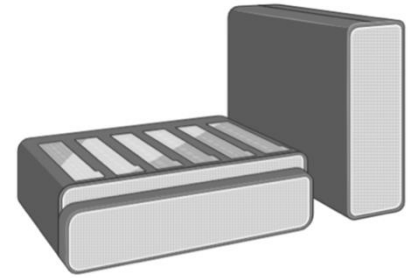
Challenge1: transfer TB data to cloud

- 2.3 TB of data in 1.5 hour = 3.4 Gbps
- Utilizing available bandwidth

Moving to cloud

Challenge1: transfer TB data to cloud

1. VBZ compression for raw signal file – 30% less file size
2. Optimize file size for
   (a) number of parallel uploads
   (b) latency overhead for each new file



Signal Files

Cloud Storage

PRINCETON UNIVERSITY

Moving to cloud

Near real time I/O

Challenge 2: Optimized distributed system

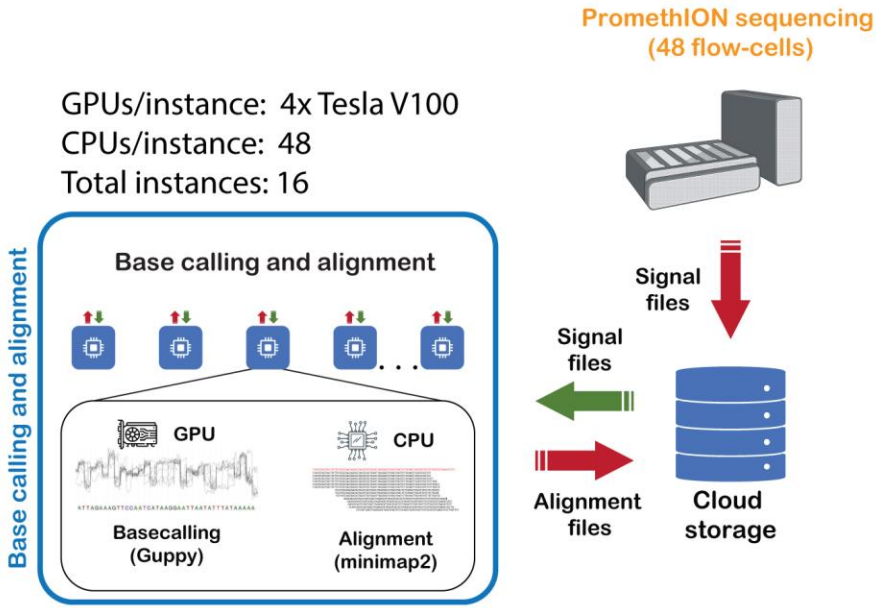
- Support streaming dataflow
- Minimize orchestration/inter-node communication
- Make sure all resources are fully utilized based on rate of data generation

Moving to cloud

Near real time I/O

Challenge 2: Optimized distributed system

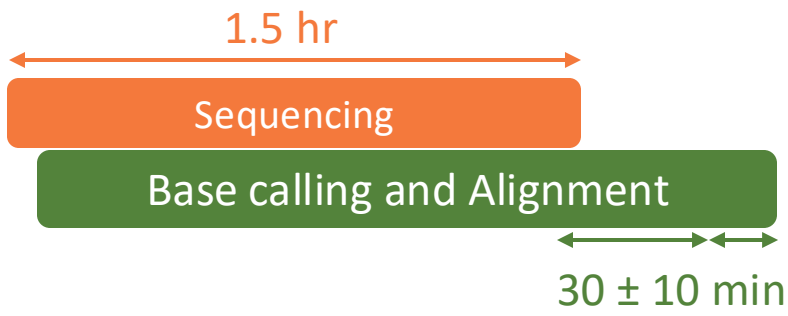
1. Analysis for specific set of flow cells assigned to each instance
2. Stateless pull architecture
3. Pipelining different compute stages



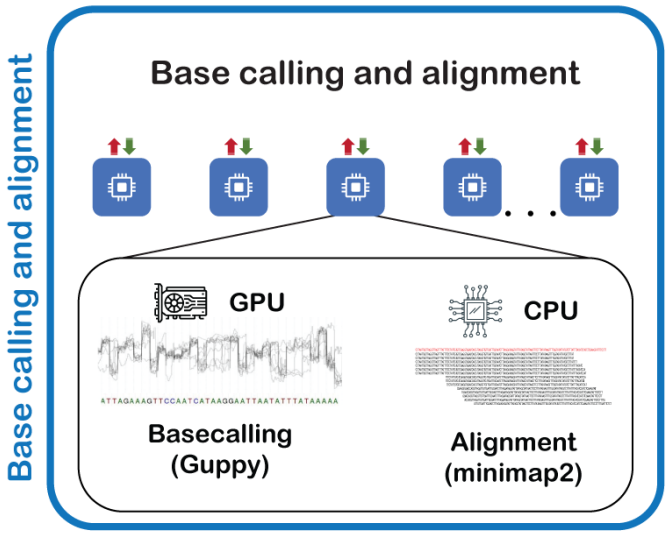
PRINCETON UNIVERSITY

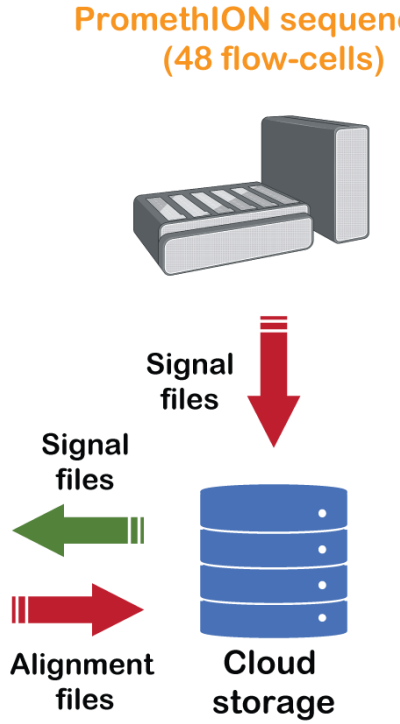
Near real time pipeline

HG002
218 Gb
2.3 TB fast5



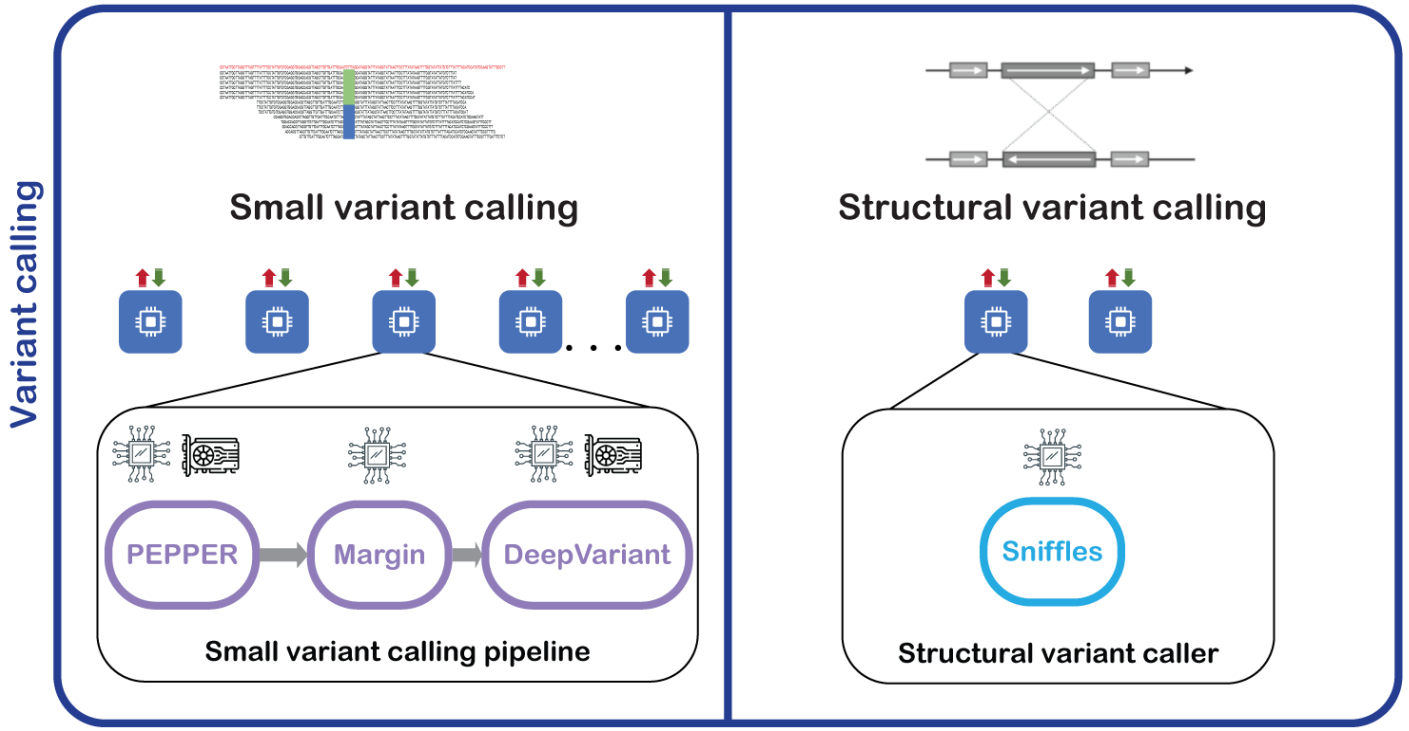
GPUs/instance: 4x Tesla V100
CPUs/instance: 48
Total instances: 16

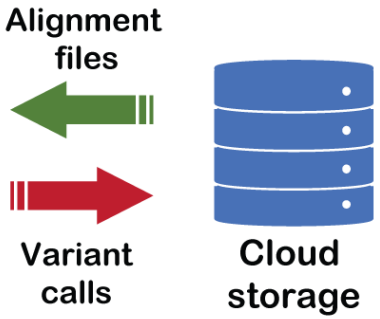




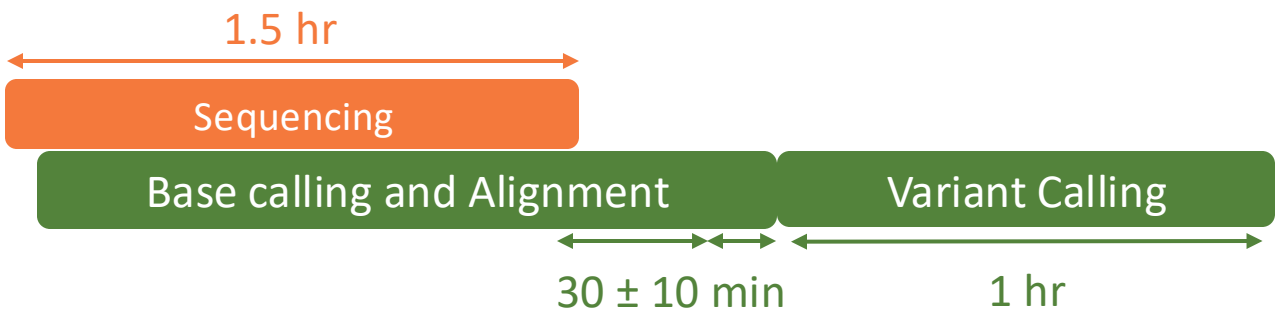
PRINCETON UNIVERSITY

Variant Calling



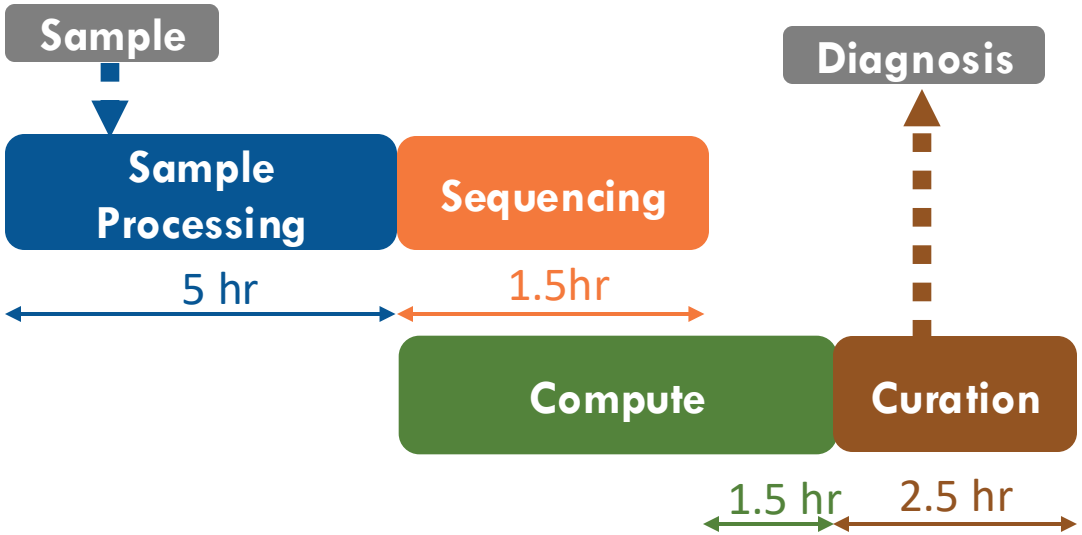


HG002
218 Gb
2.3 TB fast5



PRINCETON UNIVERSITY

10.5-hour compute-optimized workflow



PRINCETON UNIVERSITY

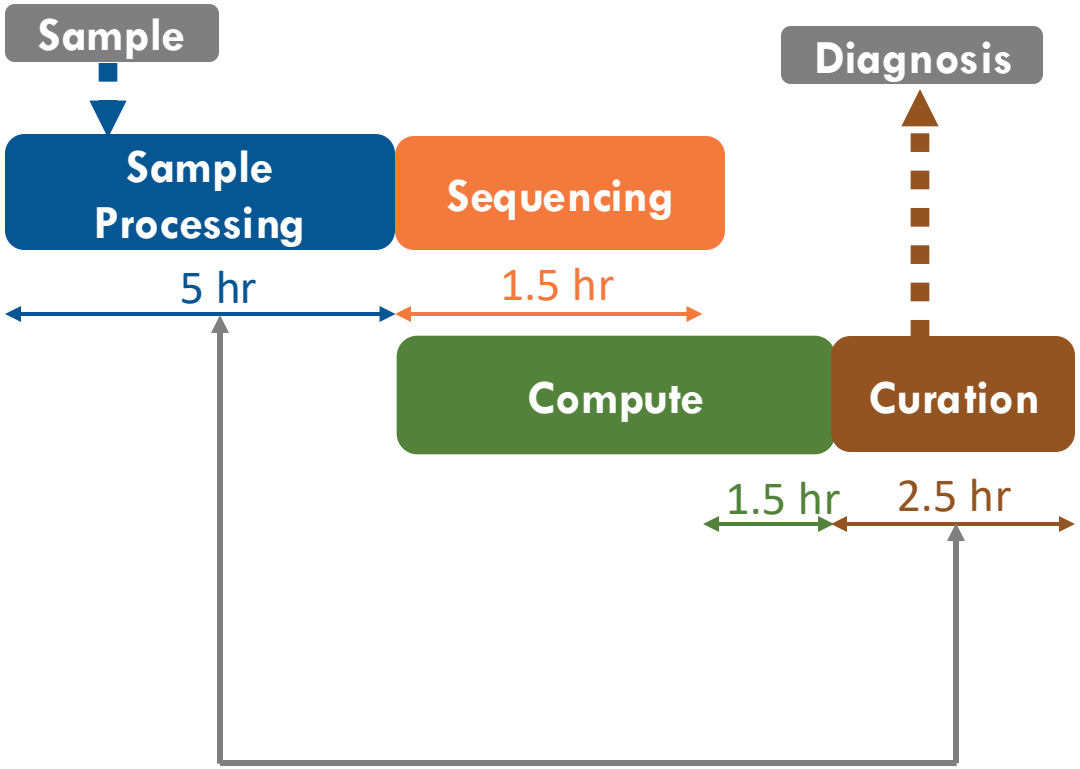
Diagnosis made in

11.3 hrs

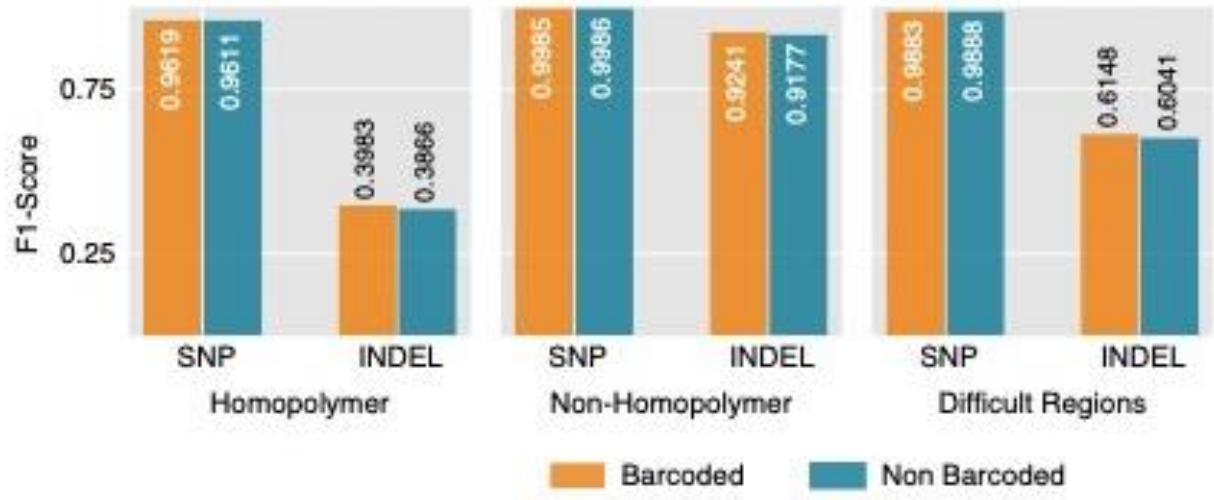
Genetic Dilated Cardiomyopathy

The patient was urgently listed for transplant and received a new heart 21 days later.

Co-design across stages

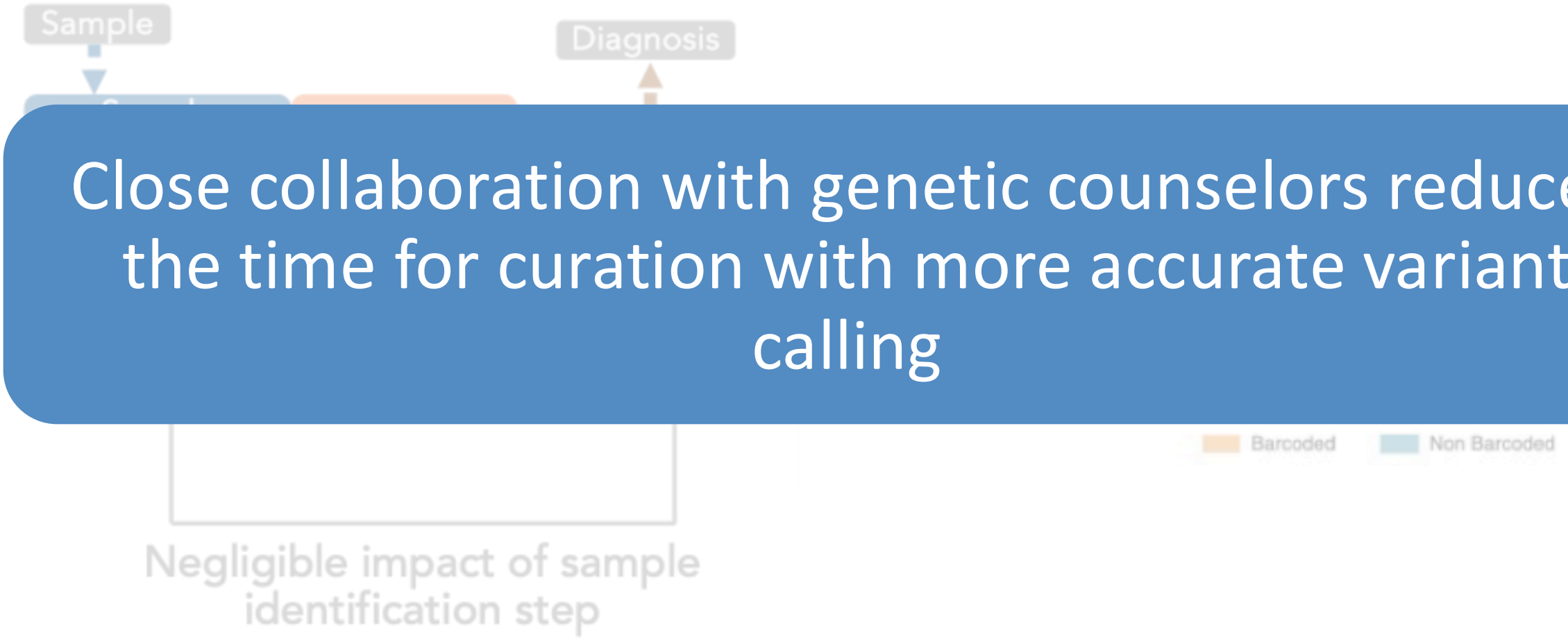


Negligible impact of sample identification step



PRINCETON UNIVERSITY

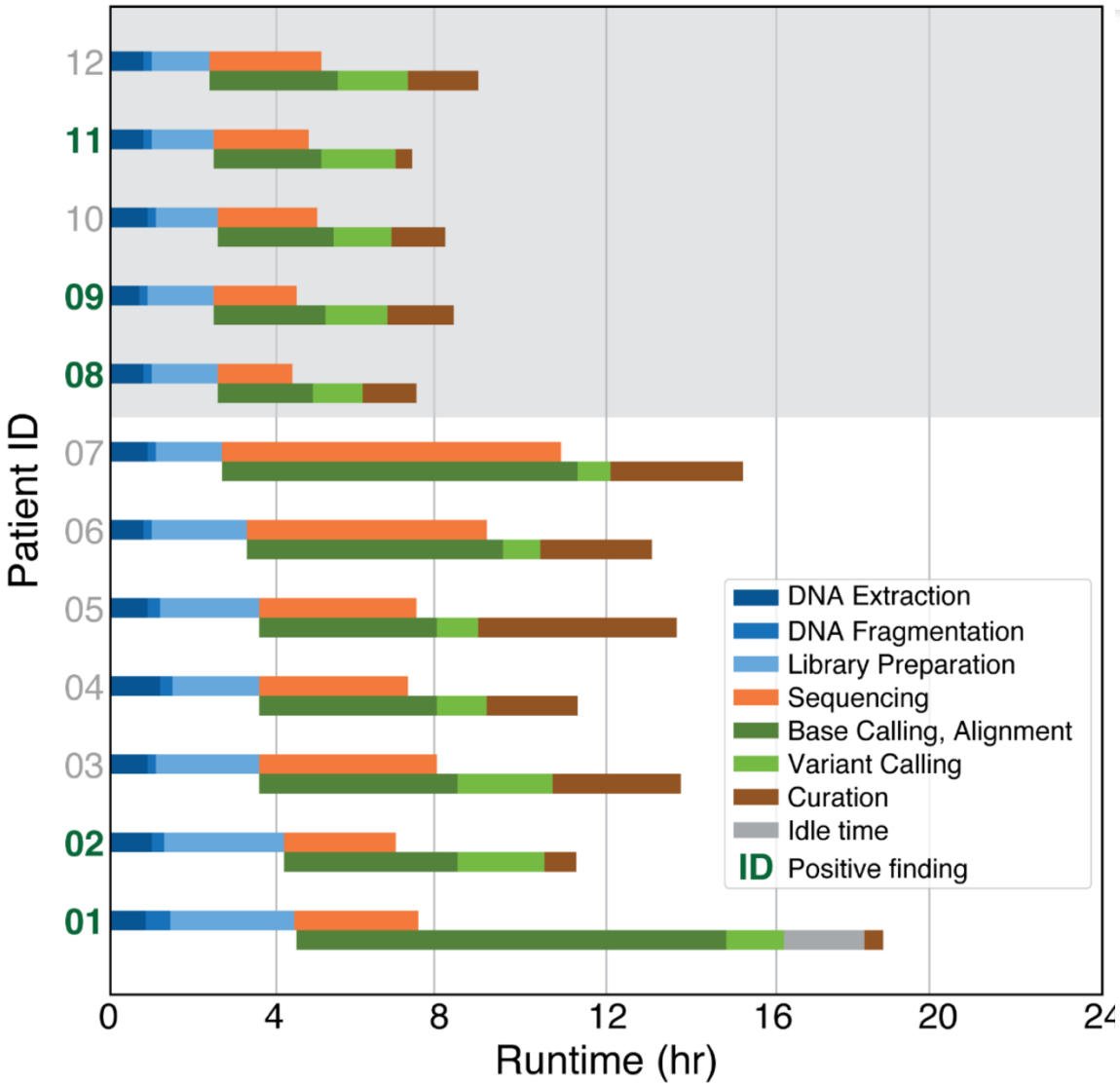
Co-design across stages

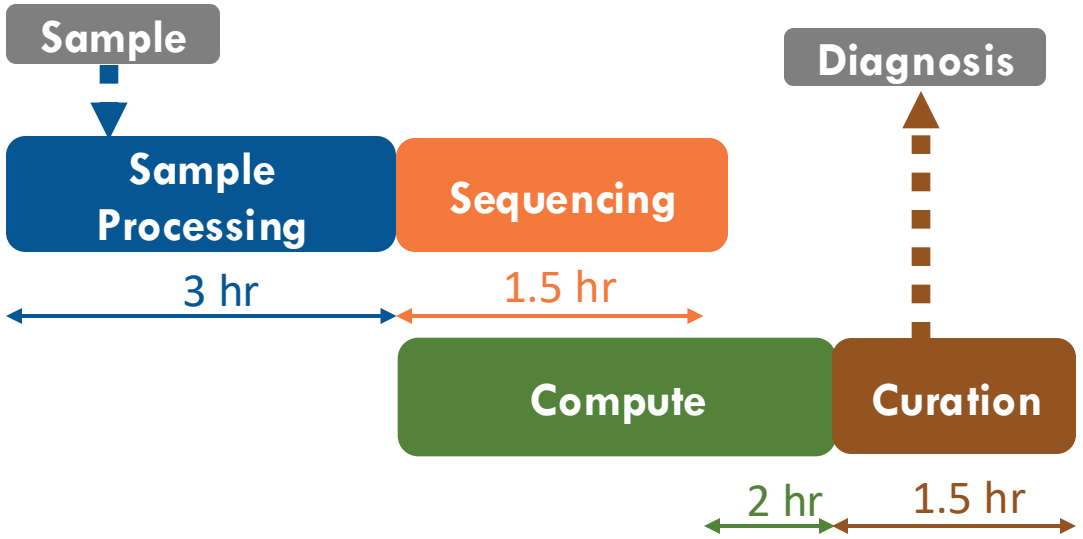


Close collaboration with genetic counselors reduces the time for curation with more accurate variant calling

PRINCETON UNIVERSITY

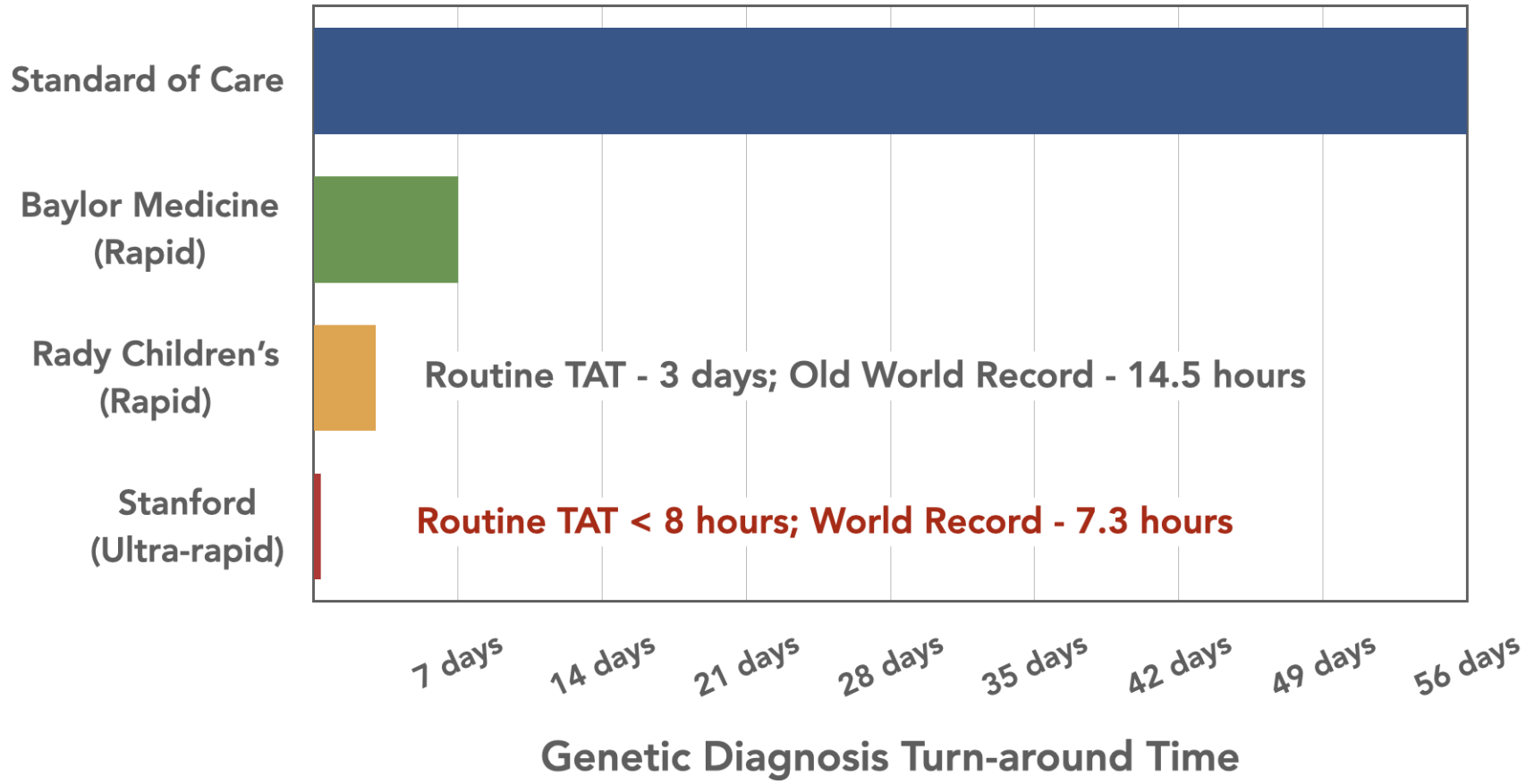
Ultra-rapid pipeline with 8 hour turnaround time





PRINCETON UNIVERSITY

Diagnosis in 8 hours.. Or less



PRINCETON UNIVERSITY

Conclusions

Performance in domain-specific design significantly depends on

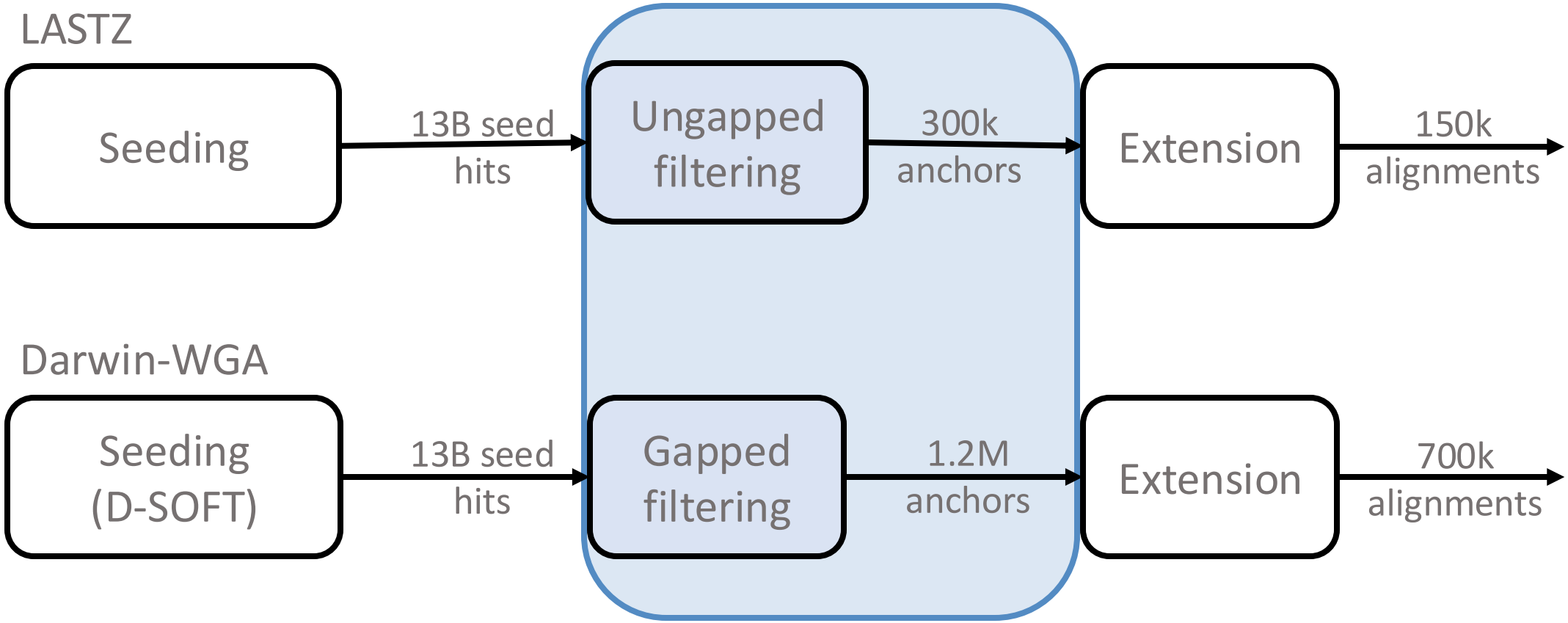
- Co-Design through integrated systems architecture and algorithmic re-design
- Deep understanding of the domain

Landscape for genomics as a domain

- Technological Agility
  Must keep pace with rapid tech and algorithmic evolution
- Hardware Ramifications
  GPUs/FPGAs are better platforms; Software framework are critical
- Community Dynamics
  Community engagement drives adoption of innovations

Thank you!

Workloads in LASTZ v/s Darwin-WGA



PRINCETON UNIVERSITY

ASIC

TSMC 40nm DC synthesis (not a chip prototype)

		Configuration	Area (mm^2)	Power (W)
Filtering	Logic	64 x (64PE array)	16.6	25.6
Extension	Logic	12 x (64PE array)	4.2	6.72
	Traceback SRAM	12 x (64PE x 16KB/PE)	15.1	7.92
DRAM	DDR4-2400R	4 x 32GB	-	3.10
TOTAL			**35.9**	**43.34**

~80% logic area, ~60% chip power

~40% chip area

Darwin-WGA is 2 orders of magnitude faster than iso-sensitive software

	LASTZ (CPU)	Iso-sensitive software
Darwin–WGA (FPGA)	0.1x (slowdown)	20x (speed and cost)
Darwin–WGA (ASIC)	1.5x	300x (1500x perf/Watt)

	HW config	Cost per hour
LASTZ	36 vCPUs	$1.59
Darwin-WGA (FPGA)	1 Xilinx Virtex Ultrascale+ FPGA + 8vCPUs	$1.65
Darwin-WGA (ASIC)	36 mm^2, 43 Watt, 40nm TSMC	