

EECS 570

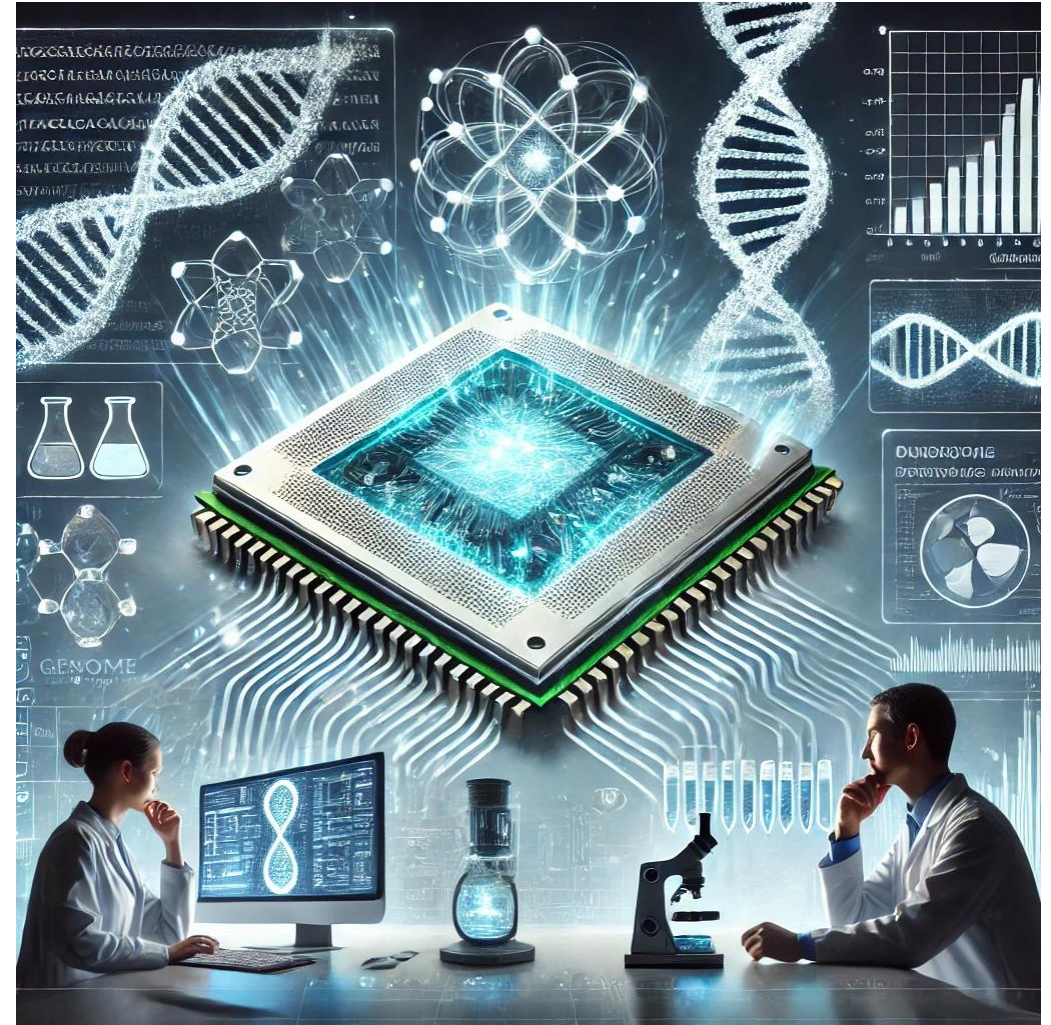
Lecture 18

Genomics Accelerators

Winter 2025

Prof. Satish Narayanasamy

<http://www.eecs.umich.edu/courses/eecs570/>



Team – Part of University of Michigan Precision Health Initiative



Reetu Das
Assoc. Professor, UM

Expertise:
Computer Architecture



Satish Narayanasamy
Professor, UM

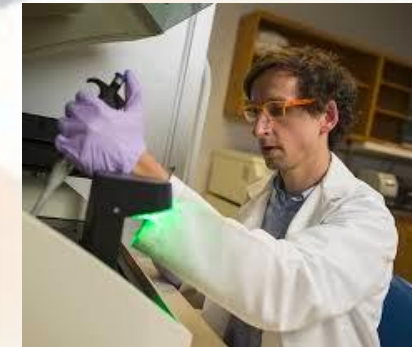
Expertise:
Parallel Architecture and Systems



David Blaauw
Professor, UM,
Expertise: VLSI Design



Robert Dickson
MD, UM
Expertise:
**Pulmonary and
Critical Care Medicine**



Carl Koschmann
MD, UM
Expertise:
**Pediatric
Hematology/Oncology**

“Discover the genetic, lifestyle and environmental factors that influence a population’s health and provides personalized solutions that allow individuals to improve their health and wellness.”



PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Work from Awesome Group of Fantastic Students!!



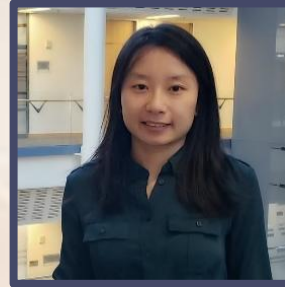
Arun
Subramaniyan



Daichi Fujiki



Jack Wadden



Xiao Wu



Timothy Dunn



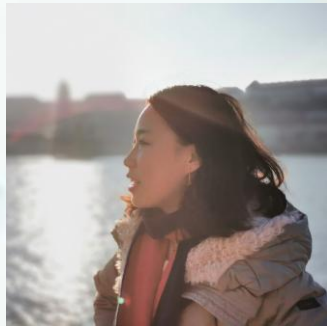
Hari Sadasivan



Yufeng Gu



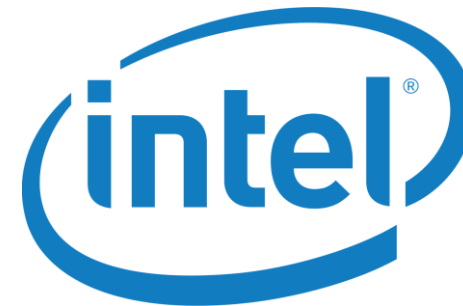
Jonah Rosenblum



Joy Dong

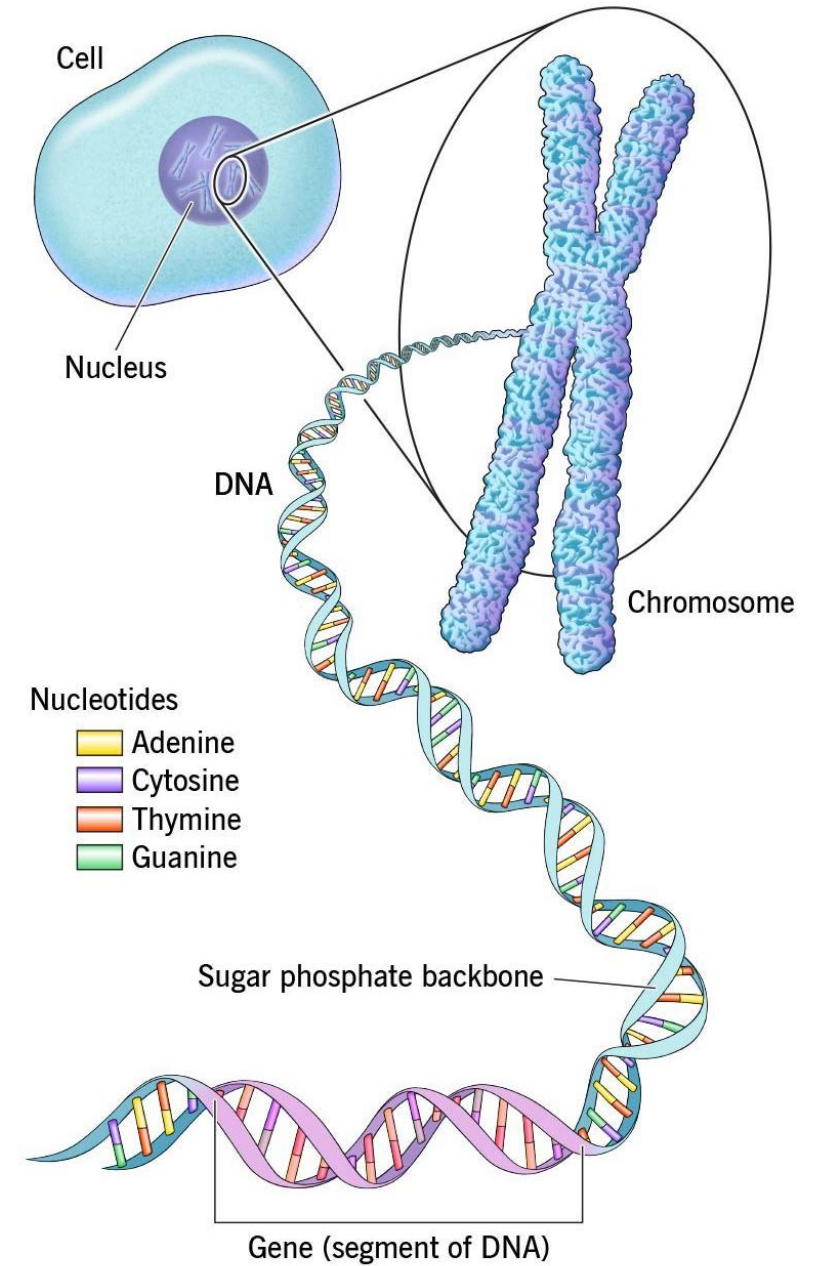
“Discover the genetic, lifestyle and environmental factors that influence a population’s health and provides personalized solutions that allow individuals to improve their health and wellness.”

Institutional partners

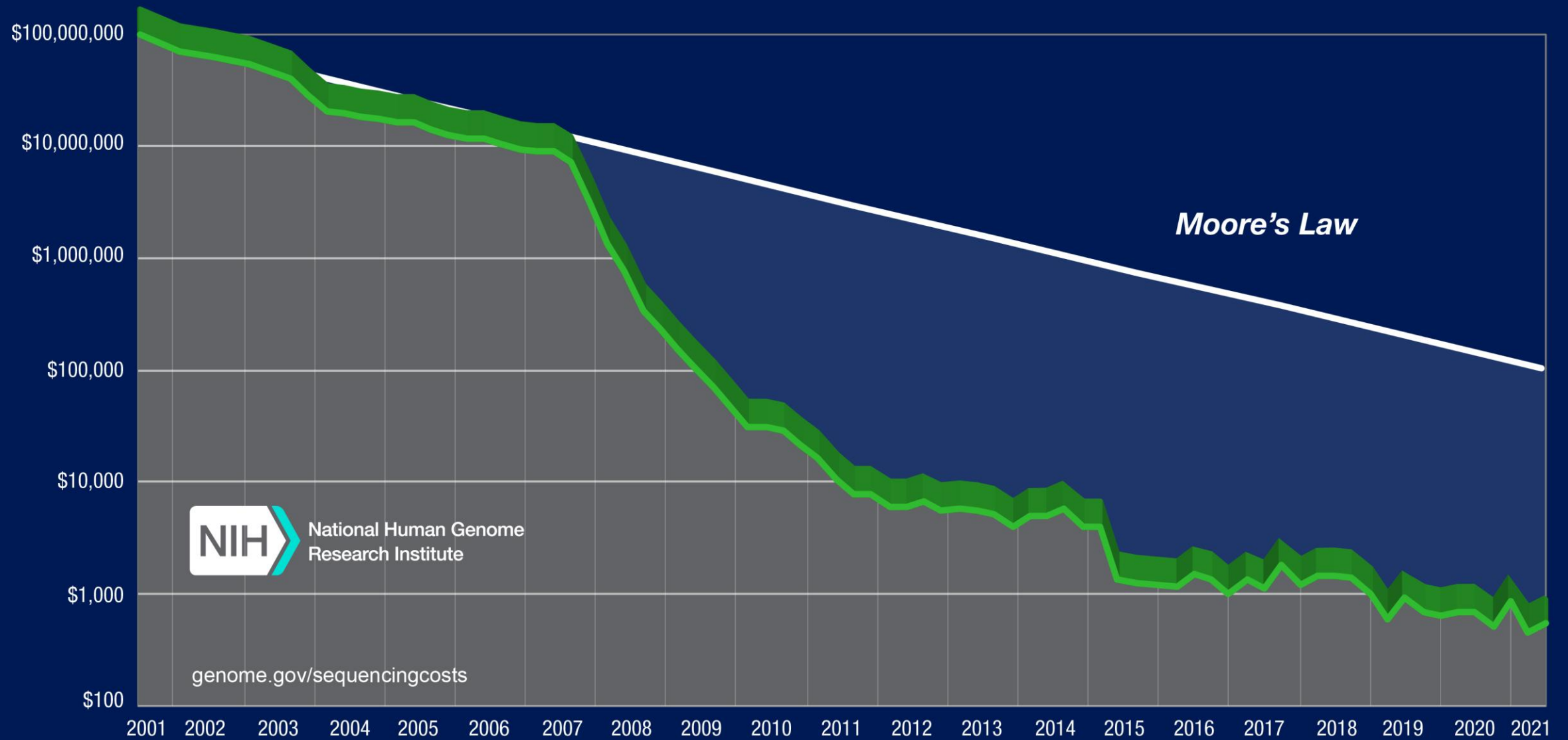


DNA Sequencing

CAGAGCTATCTAGCGACTATTATATCGTATATAGC



Cost per Human Genome



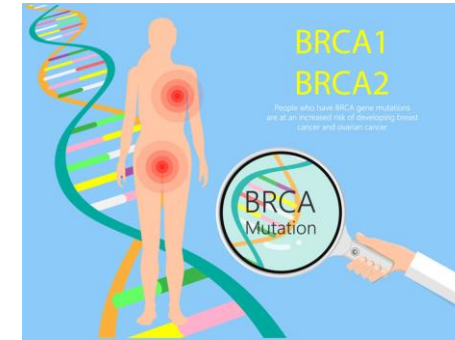
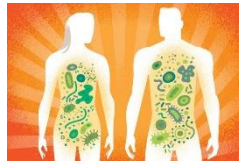
Exploding Applications

Pathogen
detection

Pandemic
prevention

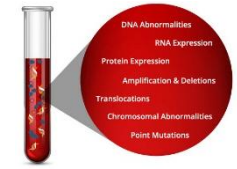


Antibiotic
resistance



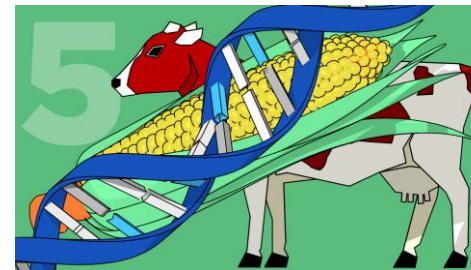
Precision health

Cancer



Liquid Biopsy

Human
WGS



Agriculture



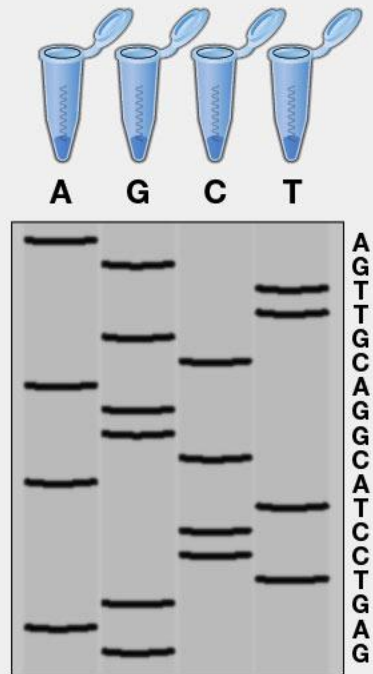
Food Safety

DNA sequencing by synthesis

Polymerase-based DNA sequencing

Sanger DNA sequencing

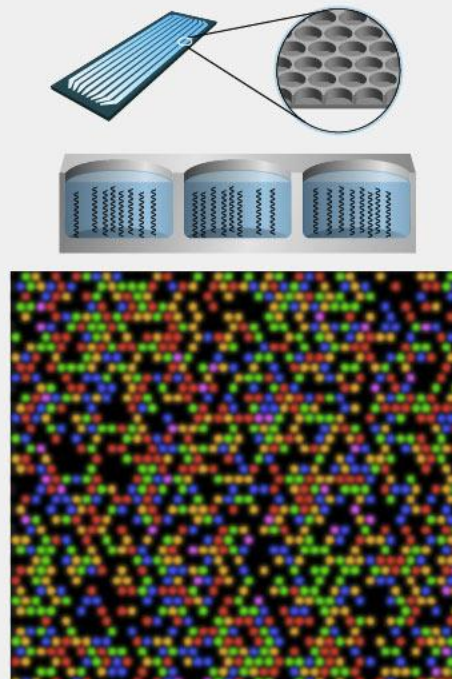
Sequence 500 - 700 DNA bases per reaction
16 reactions per gel



Sequence 10,000 DNA bases per gel

Massively parallel DNA sequencing

Sequence 100 - 5,000 DNA bases per reaction
10 thousand to 10 billion reactions per slide

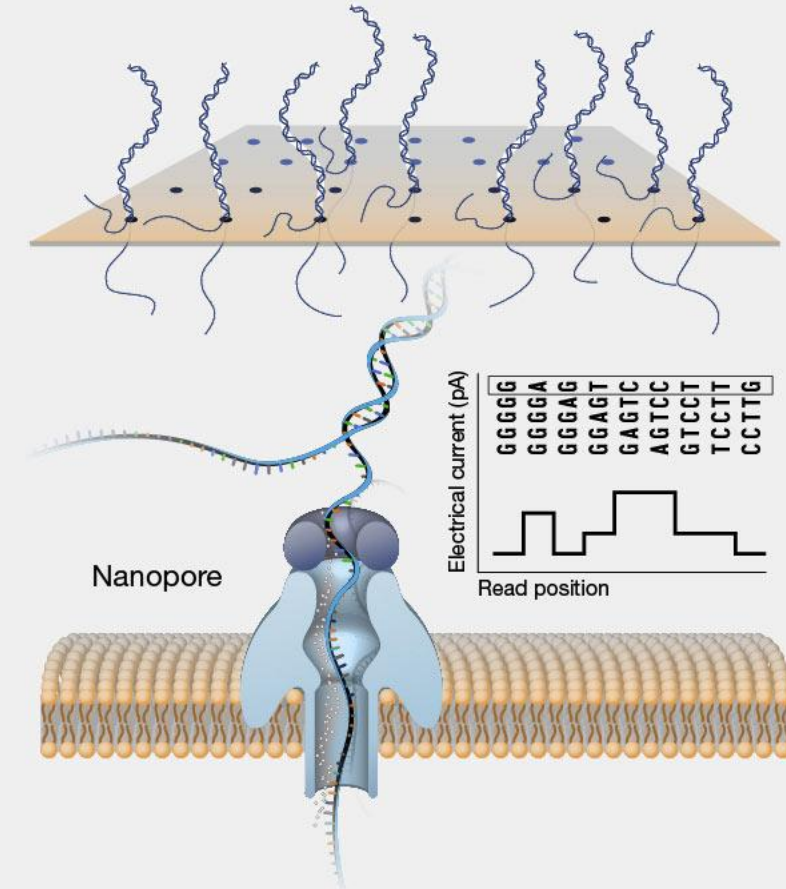


Sequence 2 trillion DNA bases per slide

Single molecule DNA sequencing

Nanopore DNA sequencing

Sequence 10 thousand to 4 million DNA bases per pore
40,000 - 250,000 pores per device



Sequence upwards of 200 billion DNA bases per device

DNA Sequencing: *Long reads are the future*

Chromosome: 50 to 300 Million bases

CAGAGCTATCTAGCGACTATTATATCGTATATAGCCTATTATATCGTATATAGCTTATATCGTATATAGC

Short Reads: 100 - 1,000 bases

- inexpensive
- currently dominate the market
- 99.9% accurate

TAATATCG

illumina®



Illumina NovaSeq
6000, 2021

3 Tbases/per day

\$10-35 per Gb

Long Reads: 1,000 - 1,000,000 bases

- more expensive
- niche industry applications
- 90% -> 99.9% accurate

AGCCTATTATATCGTATATAGCTTATATCGTATAT

Oxford
NANOPORE
Technologies



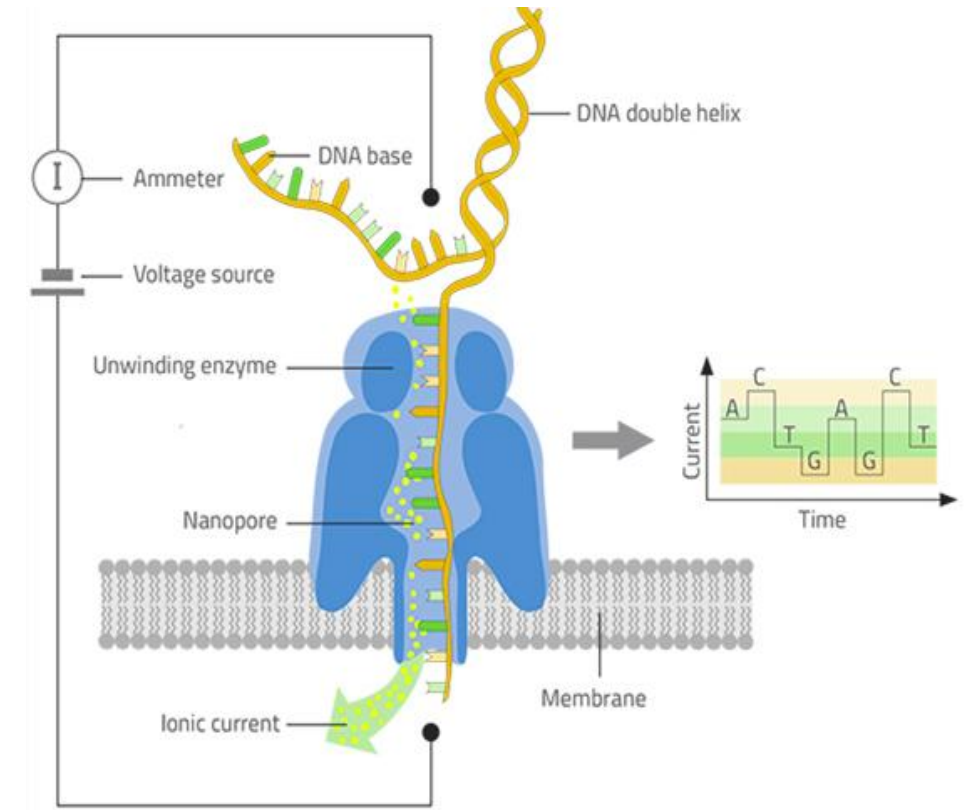
400 bases/sec
per flow-cell

PacBio

\$30-90 per Gigabase

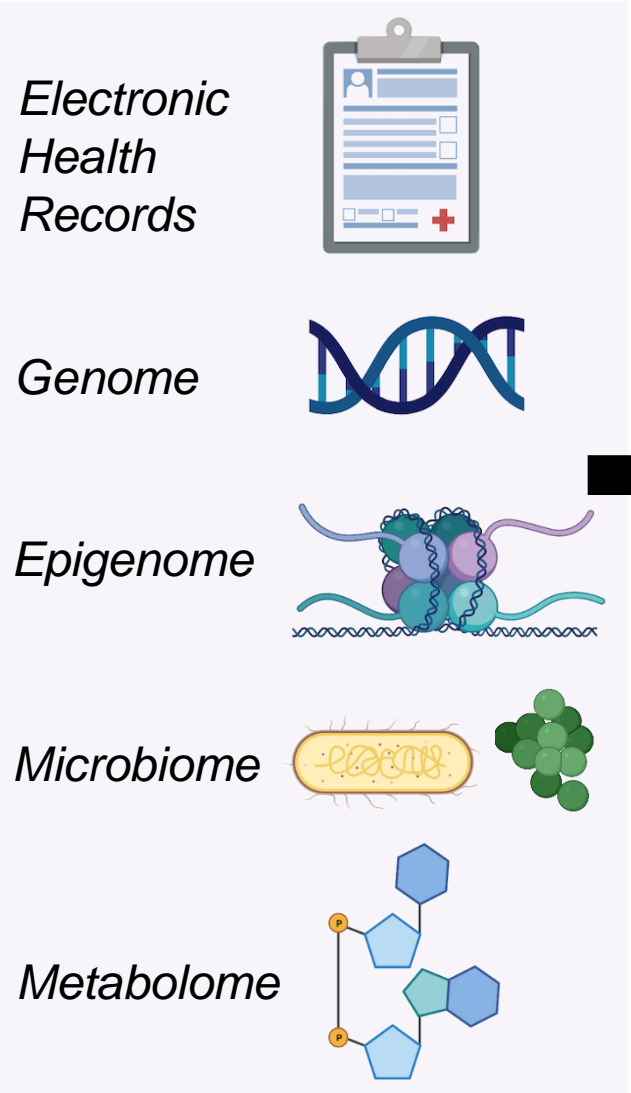
Oxford Nanopore Sequencers

10

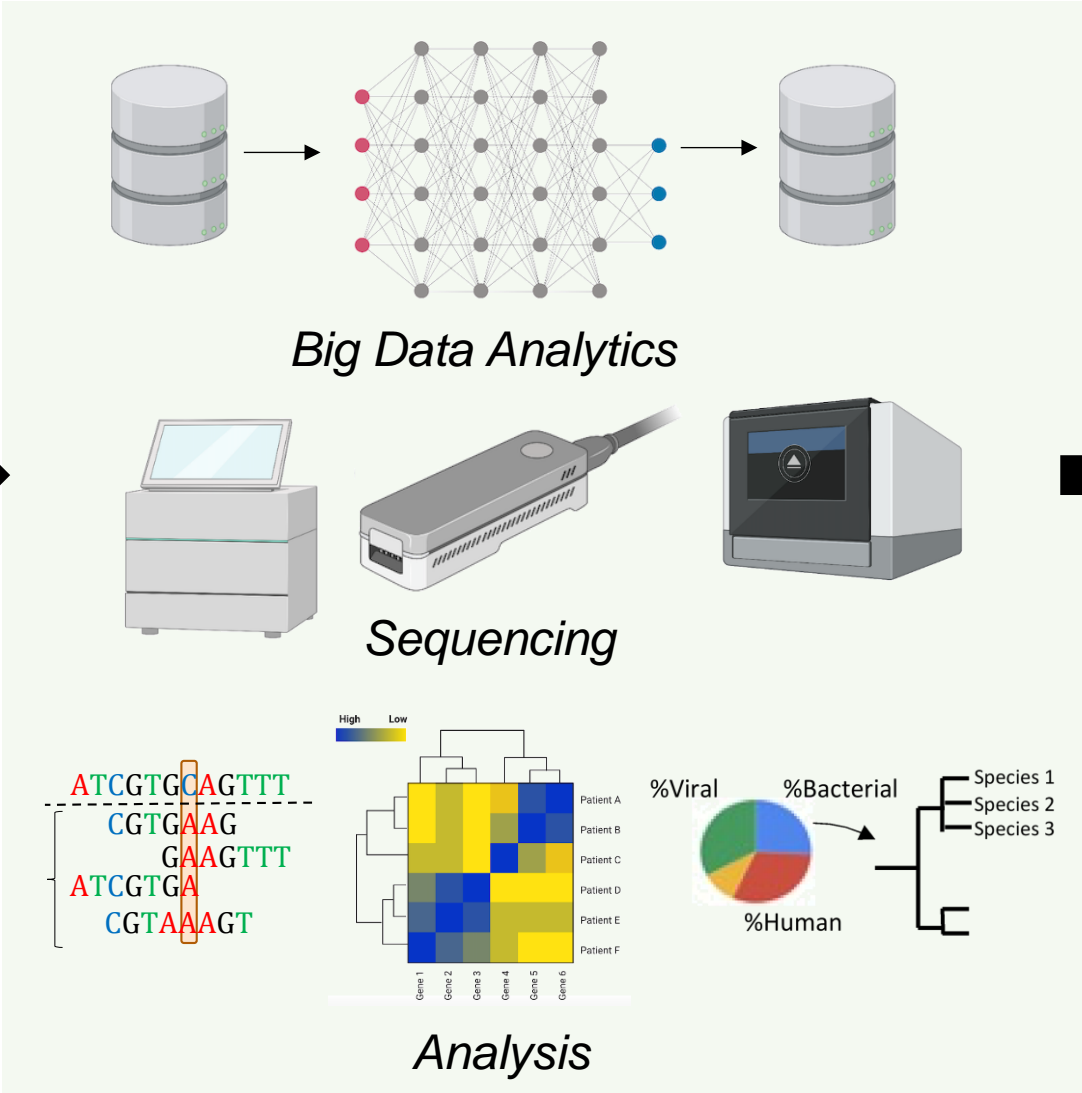


K. Goepfrich and K. Judge, "Decoding DNA with a pocket-sized sequencer," 2018.

Precision Health Platform



Data Sources



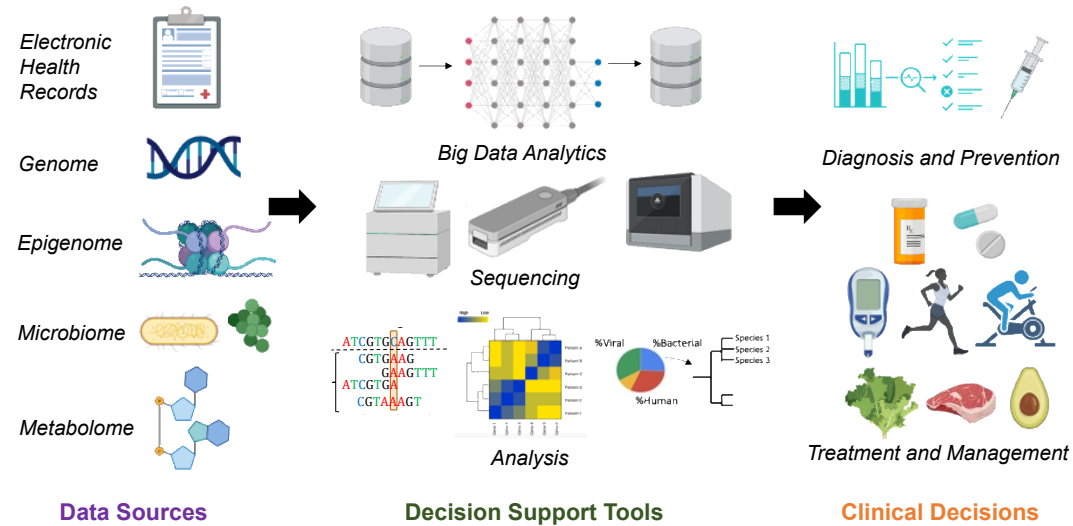
Decision Support Tools



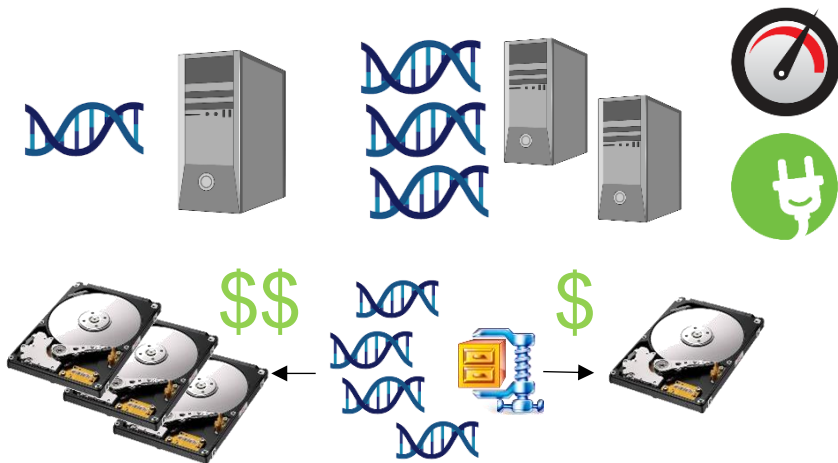
Clinical Decisions

Computing System Design Considerations

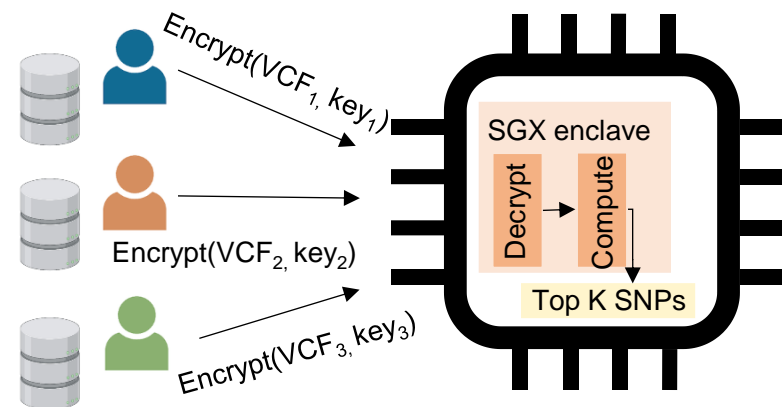
12



Efficiency



Security and Privacy



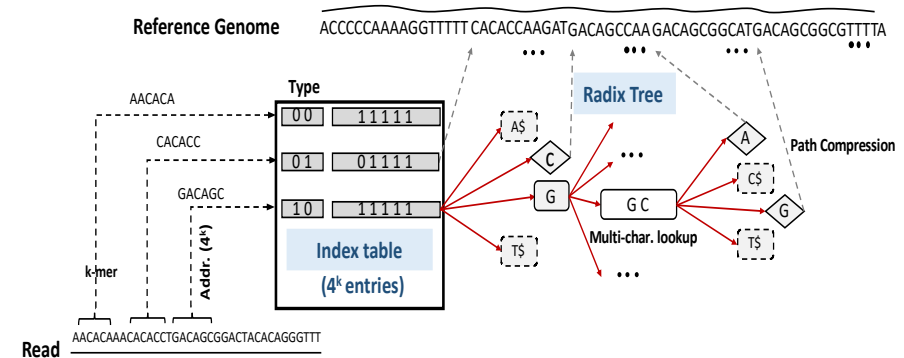
Homomorphic encryption, Intel SGX

Form Factor



CS Challenges and opportunities

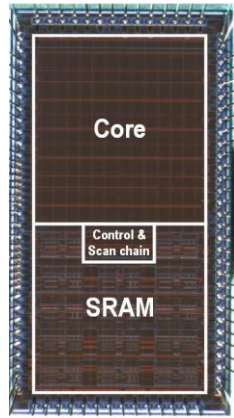
Abundant data parallelism



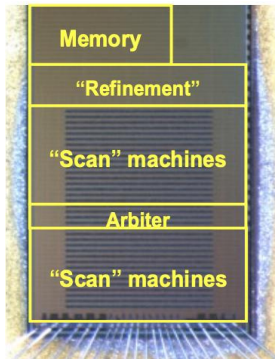
Irregular memory accesses; Memory bandwidth bound

Diverse constantly evolving kernels

Highlights: Custom computing solutions for genomics



SillaX ASIC
fabricated
(55nm)



Pruning
pairHMM ASIC
(40nm)



aws

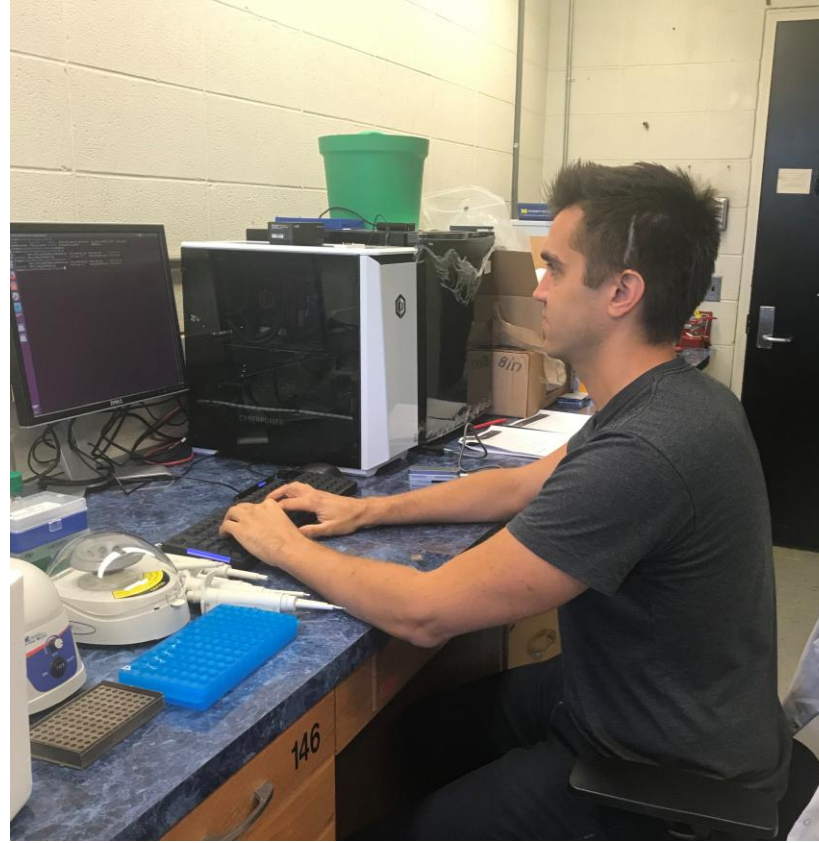
Whole Genome Sequencing
(WGS)

Pathogen detection

Intra-operative cancer
diagnosis




Privacy
using trusted hardware

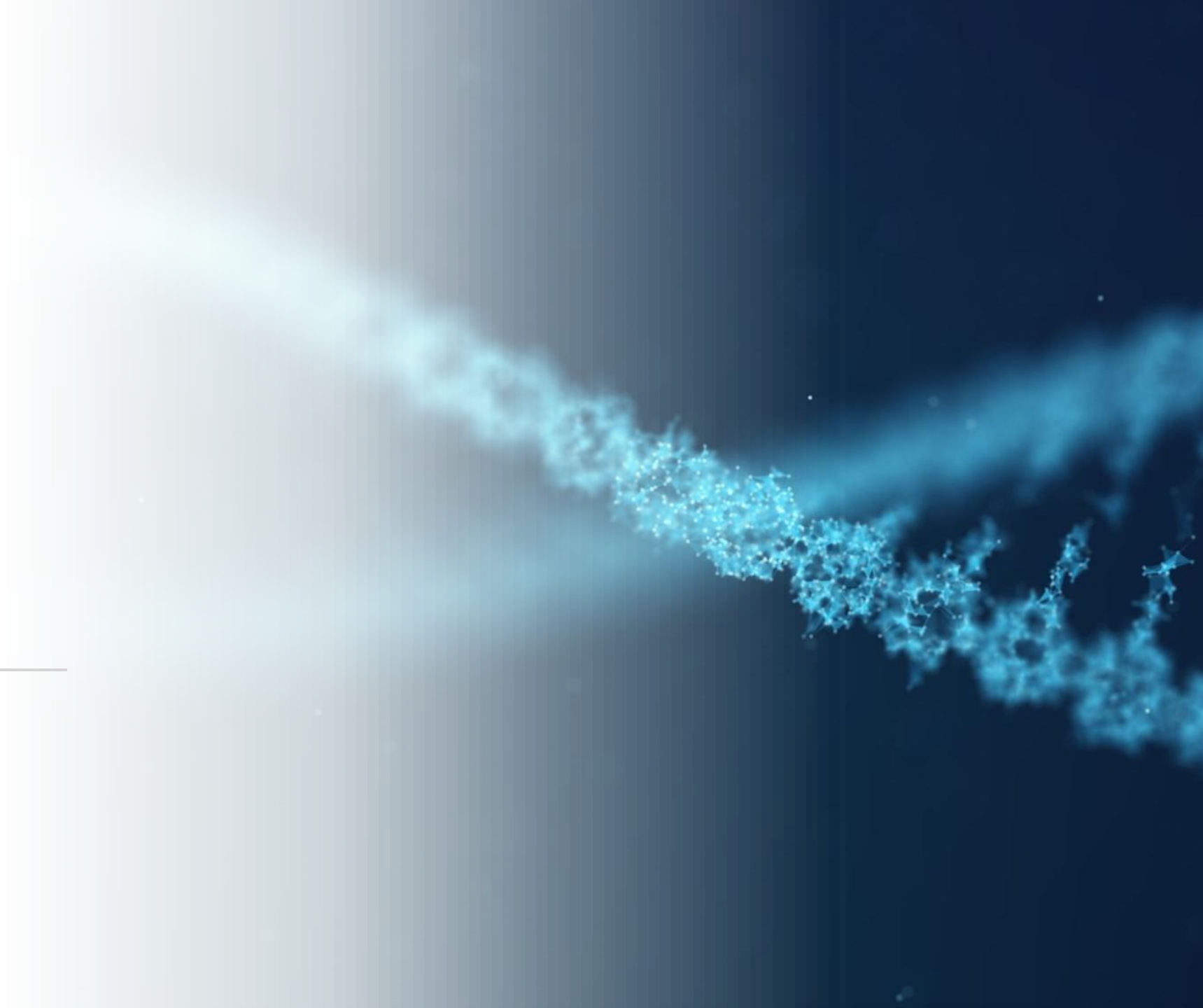


Nanopore Sequencing Lab at UM EECS

- Biosafety Level -2 Certification for tissue and RNA work
- Standard molecular biology equipment
- Small -20C freezer
- Enables tight coupling of informatics with nanopore sequencer



Whole Genome Sequencing



Acceleration Study: Whole Genome Sequencing



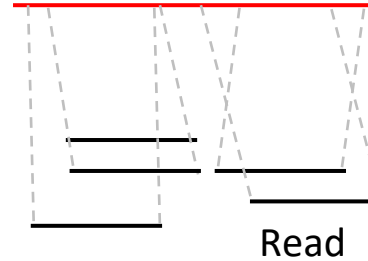
Human Genome
6 G bases



Sequenced reads
(~billions)

ATCGTGCAGT
GTGCATCTAC
CAGTACATCG
ATCGTGCTAC

Reference genome



Read Alignment

Reference genome

ATCGTGCAGTTT
CGTGAAG
GAAGTTT
ATCGTGA
CGTAAAGT
Aligned reads

Variant Calling



Diagnosis

Time (hr) 0 5 10 15

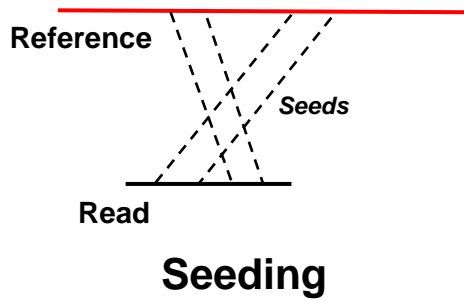
Baseline
m5.8xlarge
32 vCPUs

Seeding
Seed Extension
Sorting / Mark Duplicates
pairHMM
Other

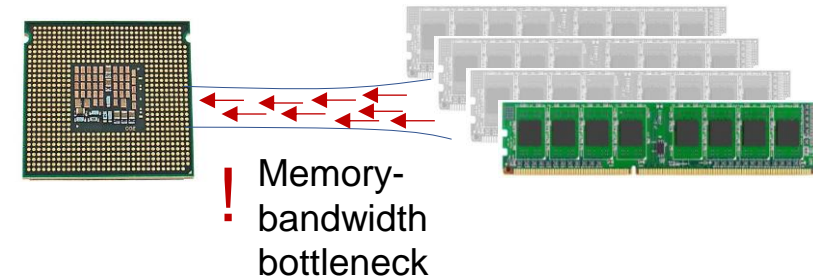
13.9 hr (445 CPU hrs)

Seeding: Memory Bandwidth Bottleneck

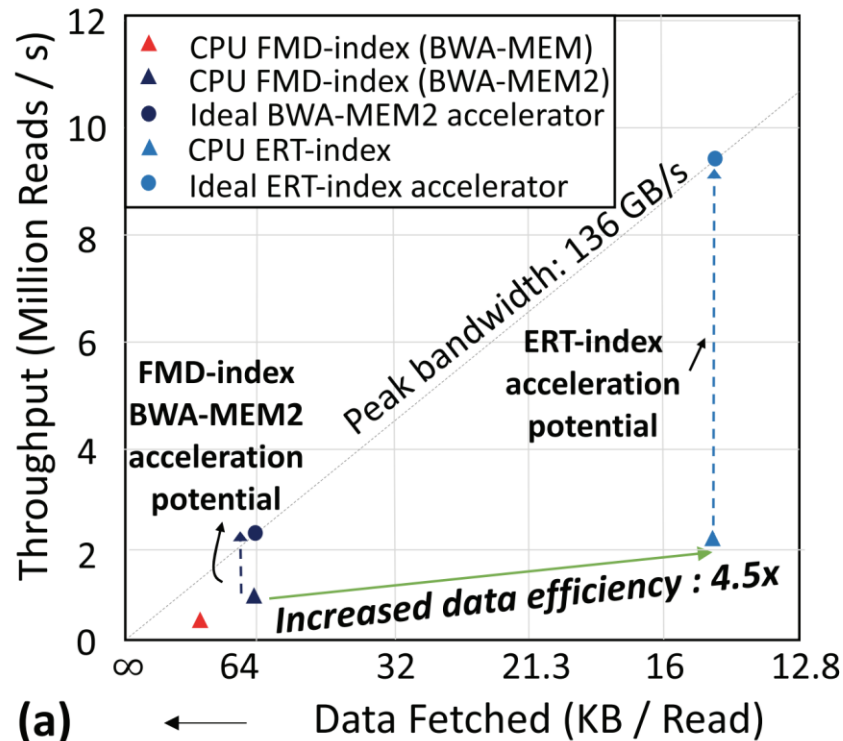
Problem



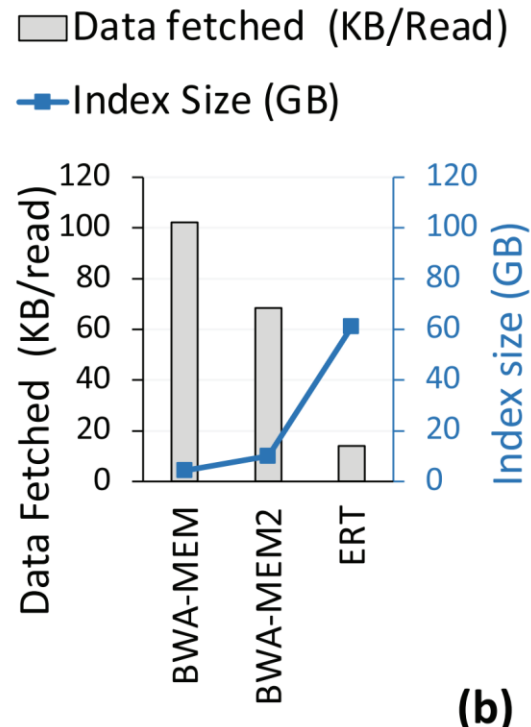
FM-index → widely used seeding data structure



FM-index
4.2 GB
human



(a)

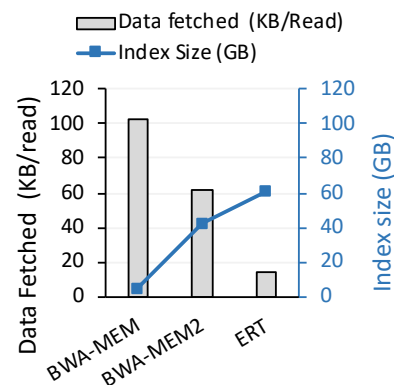
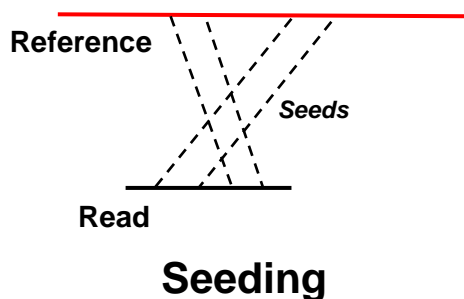


(b)

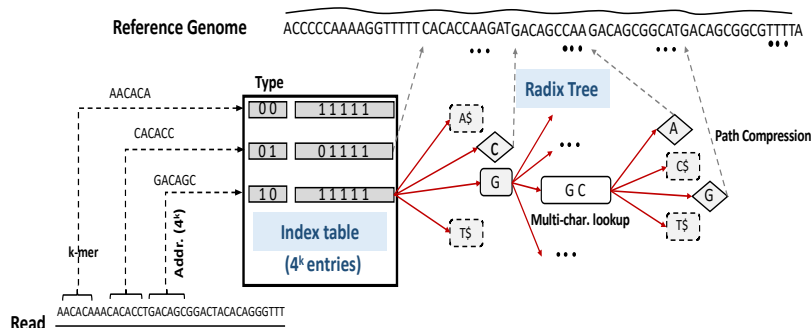
[Subramanian et al. ISCA'21]

Seeding: ERT

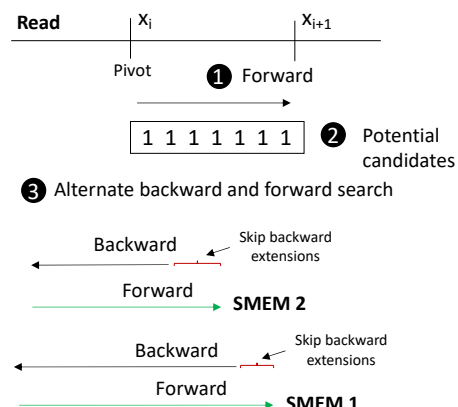
Problem



Our Solution



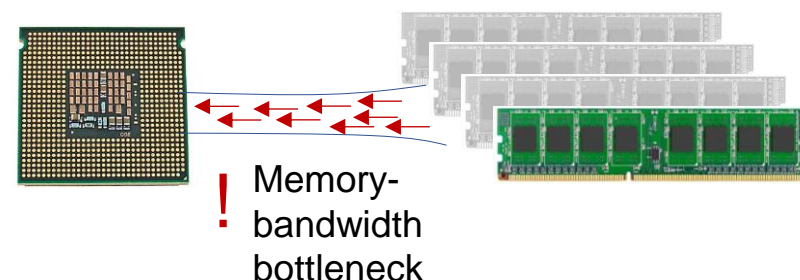
Bandwidth-efficient data structure



Bandwidth-efficient search algorithm

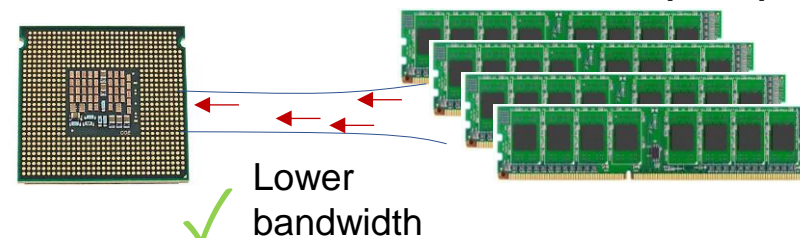
[ISCA'21]

FM-index → widely used seeding data structure



FM-index
4.2 GB
human

Enumerated Radix Tree (ERT)



ERT
~60 GB
human

Trades-off memory capacity
for memory bandwidth

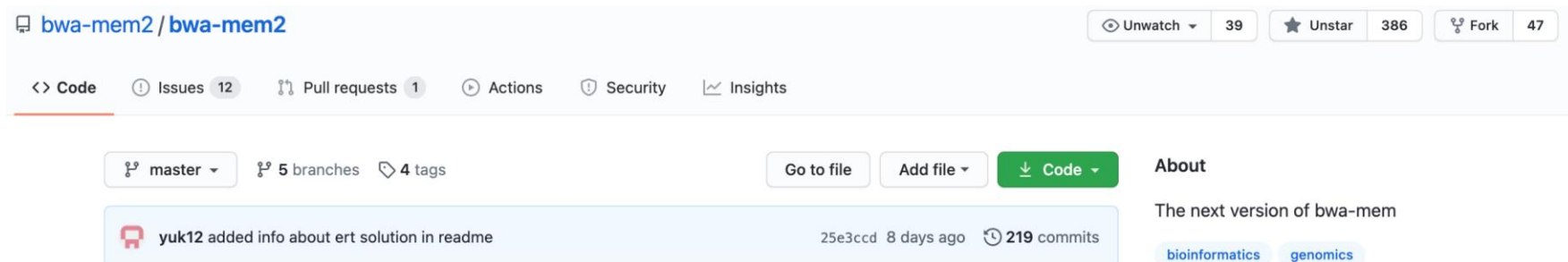
Results



2.3x over BWA-MEM2
with SeedEx

Open-source: <https://github.com/bwa-mem2/bwa-mem2/tree/ert>

ERT software integration with Broad Institute / Intel's BWA-MEM 2



Commit	Author	Message	Time	Commits
25e3ccd	yuk12	added info about ert solution in readme	8 days ago	219

bwa-mem2 seeding speedup with Enumerated Radix Trees (Code in ert branch)

The ert branch of bwa-mem2 repository contains codebase of enumerated radix tree based acceleration of bwa-mem2. The ert code is built on the top of bwa-mem2 (thanks to the hard work by @arun-sub). The following are the highlights of the ert based bwa-mem2 tool:

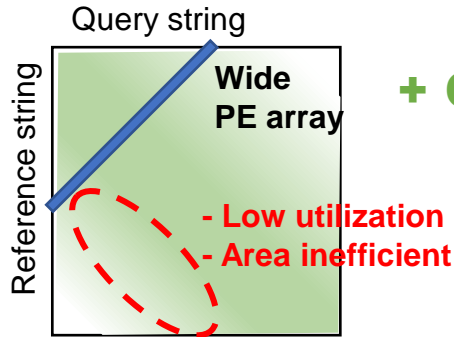
1. Exact same output as bwa-mem2
2. The tool has two additional flags to enable the use of ert solution (for index creation and mapping), else it runs in vanilla bwa-mem2 mode
3. It uses 1 additional flag to create ert index (different from bwa-mem2 index) and 1 additional flag for using that ert index (please see the readme of ert branch)
4. The ert solution is 10% - 30% faster (tested on above machine configuration) in comparison to vanilla bwa-mem2 -- users are advised to use option `-K 1000000` to see the speedups
5. The memory footprint of the ert index is ~60GB
6. The code is present in ert branch: <https://github.com/bwa-mem2/bwa-mem2/tree/ert>

BWA-MEM is the gold standard read aligner used worldwide

Read Alignment: SeedEx

[MICRO'20]

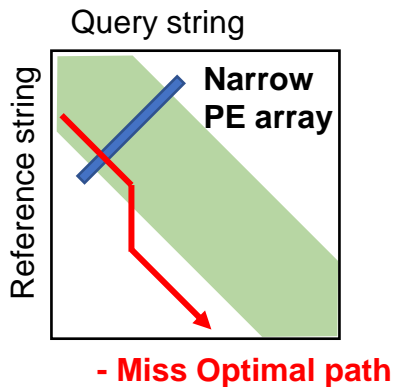
Full-band implementation



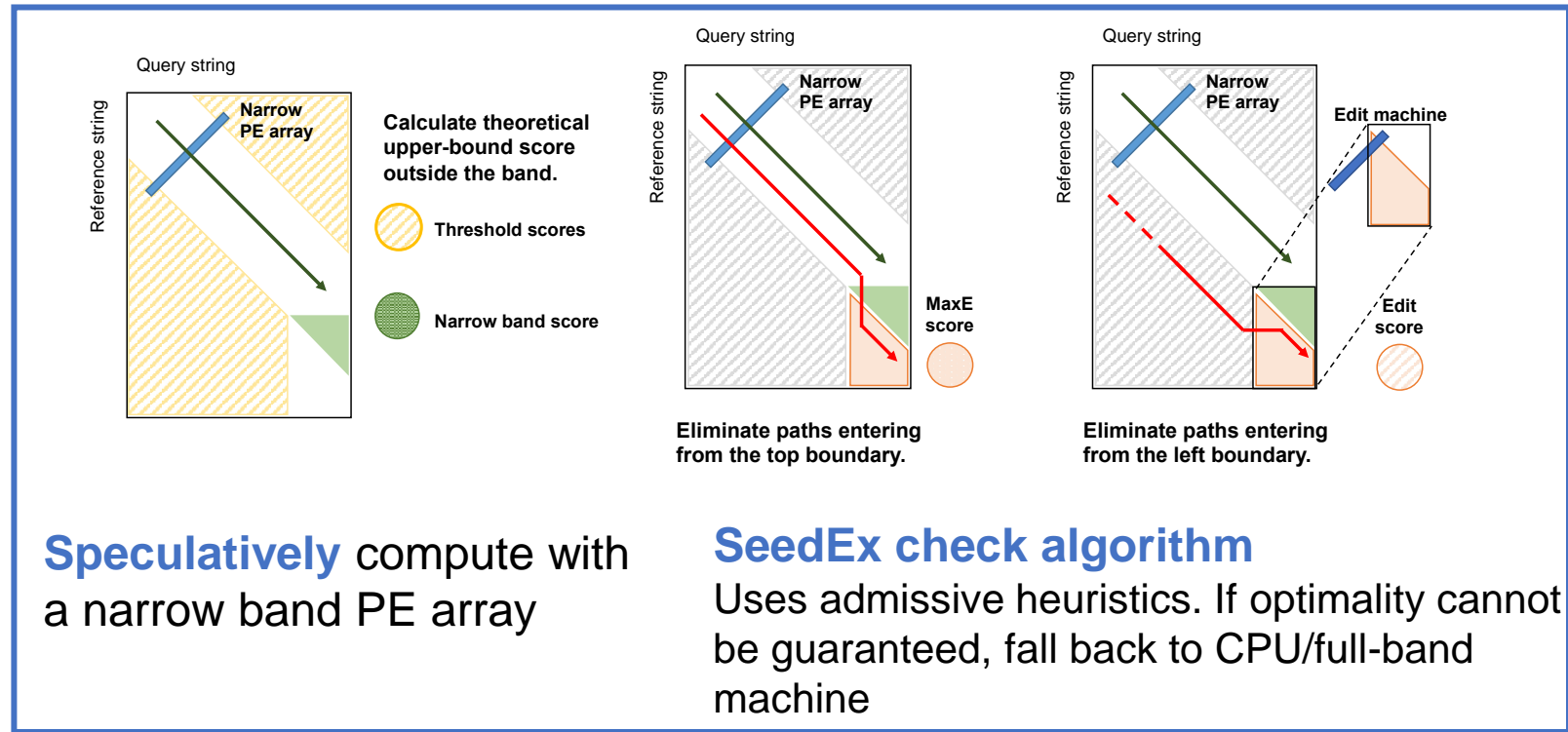
+ Optimality



Banded implementation



+ Area Efficiency



Speculatively compute with a narrow band PE array

SeedEx check algorithm

Uses admissive heuristics. If optimality cannot be guaranteed, fall back to CPU/full-band machine

Accuracy



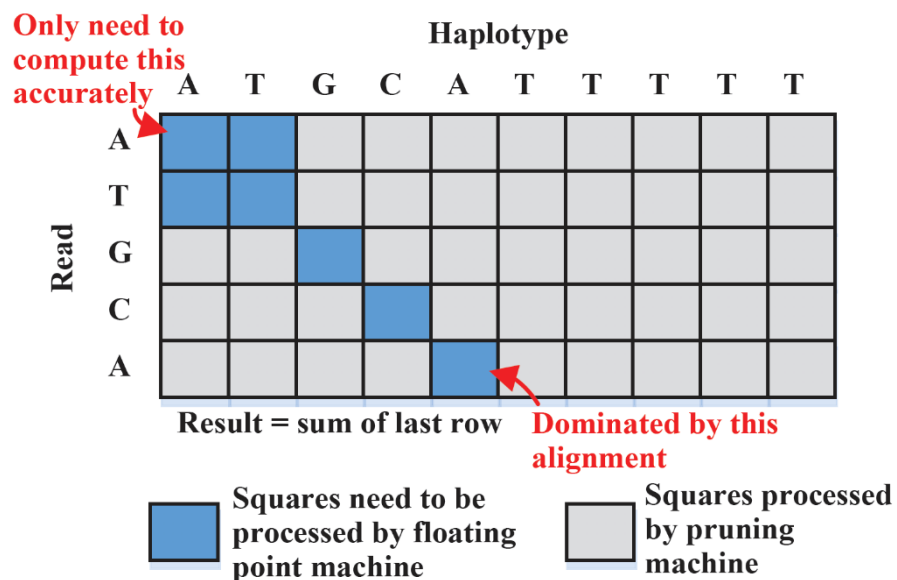
100% equivalent results on AWS cloud FPGA when integrated with BWA-MEM software

6x

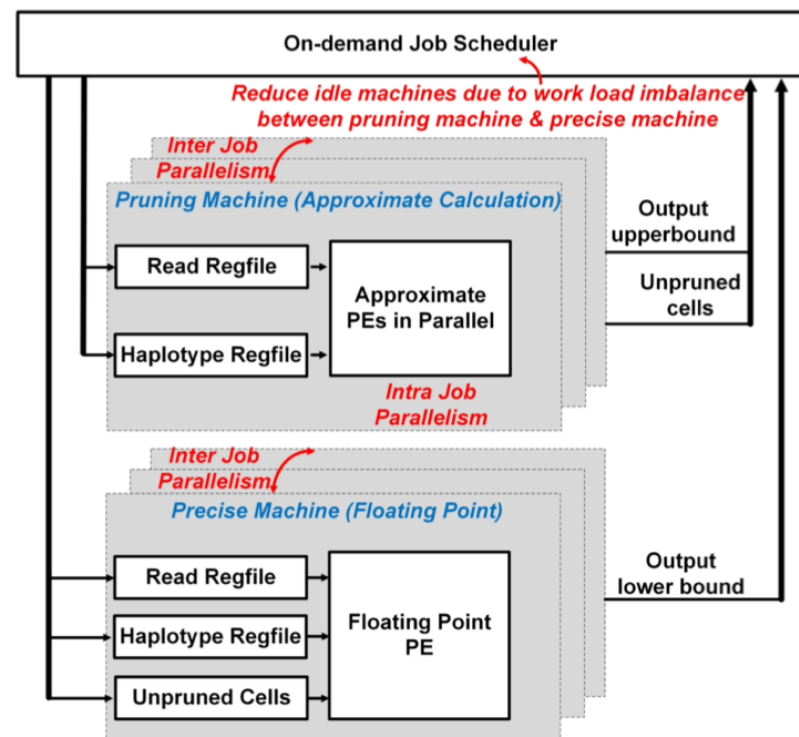
higher throughput over banded Smith-Waterman FPGA ($w = 101$) for same area

Variant Calling: pairHMM Acceleration

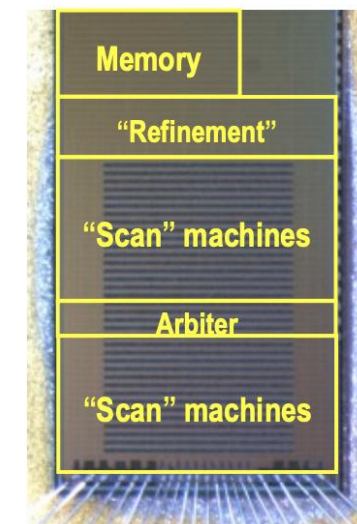
[VLSI Circuits'20]



Pruning Algorithm



Accelerator Architecture



Pruning pairHMM ASIC (40nm)

Bit equivalent output

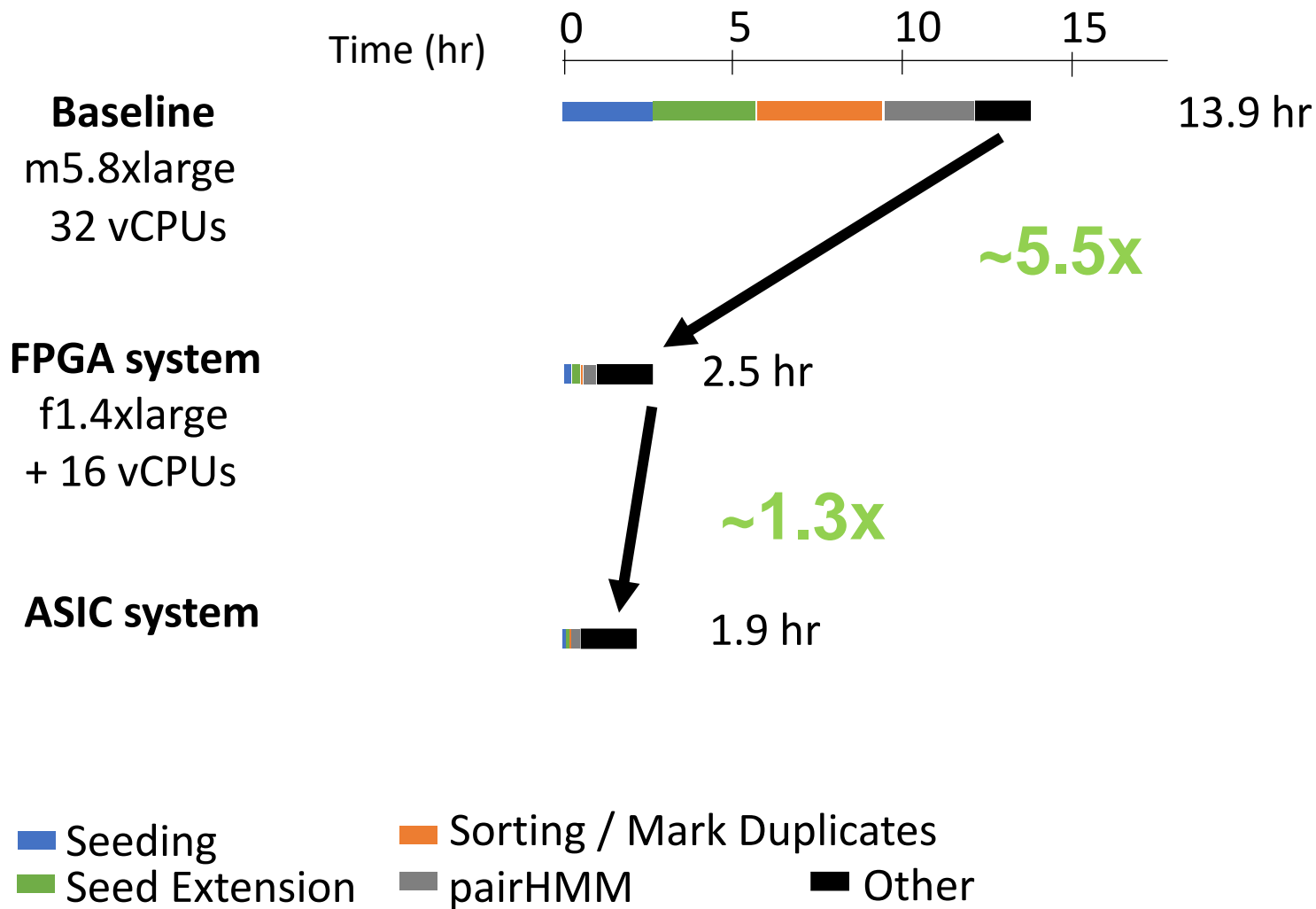
43x

fewer cells computed in precise floating point

8.3x

higher throughput (GCUPS) than floating-point ASIC of the same area

Summary: Accelerating Short-Read WGS



ERT integrated with
Heng Li's
BWA-MEM2 open-source

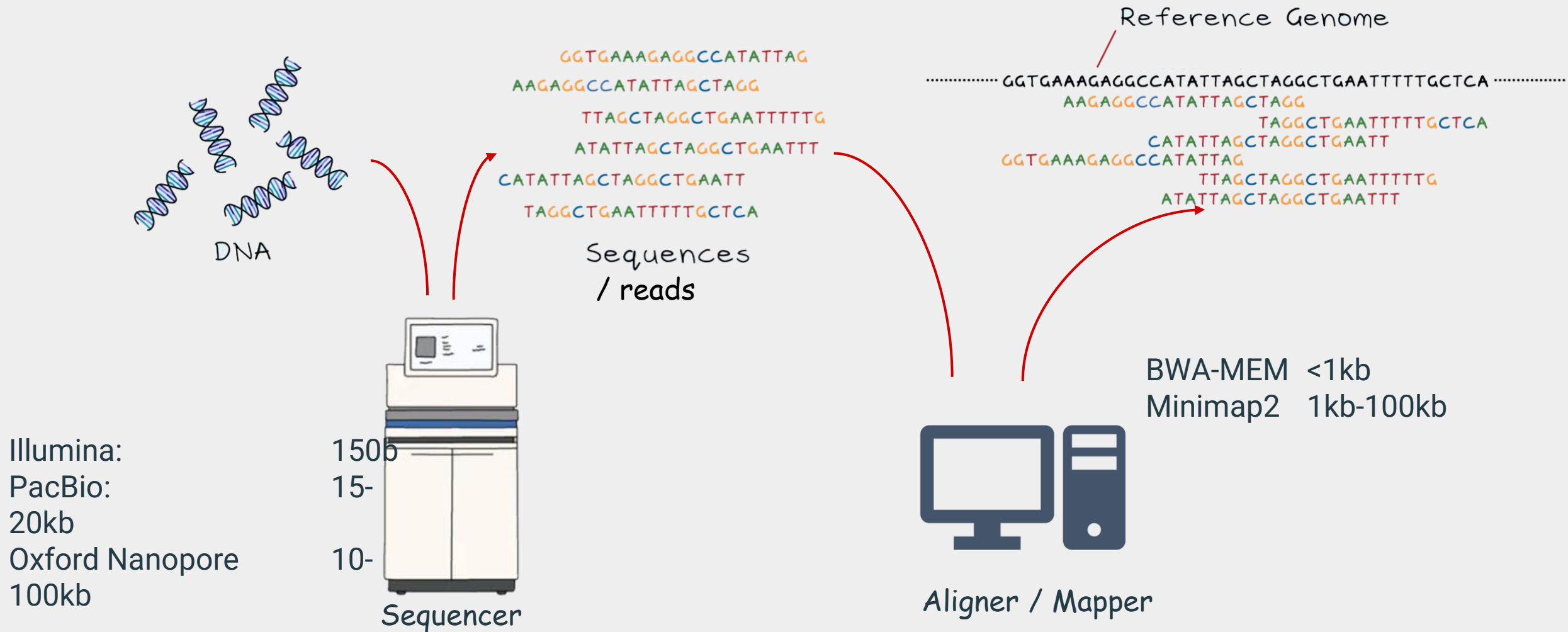
@BioSys Workshop'24

mm2-gb: GPU Accelerated Minimap2 for Long Read DNA Mapping

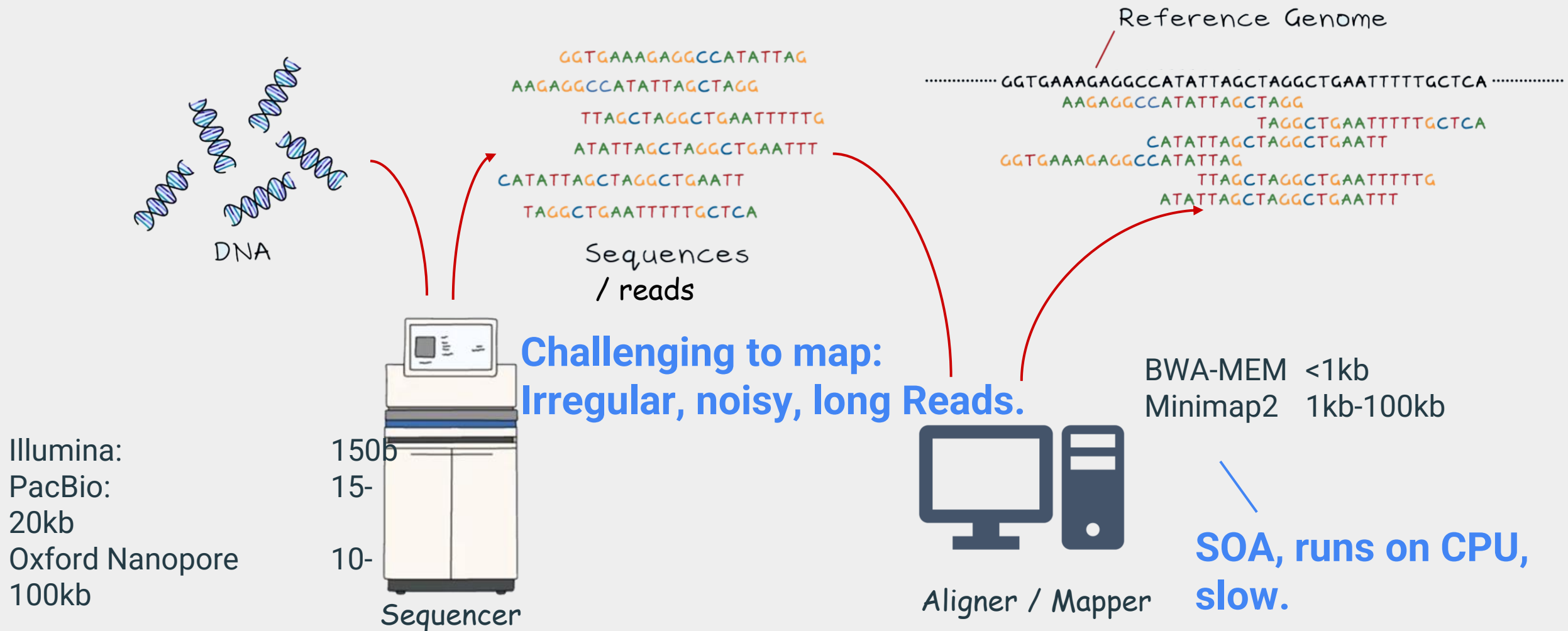
Juechu Dong*¹, Xueshen Liu*¹, Harisankar Sadasivan², Sriranjani
Sitaraman², Satish Narayanasamy¹

1. University of Michigan 2. AMD Inc.

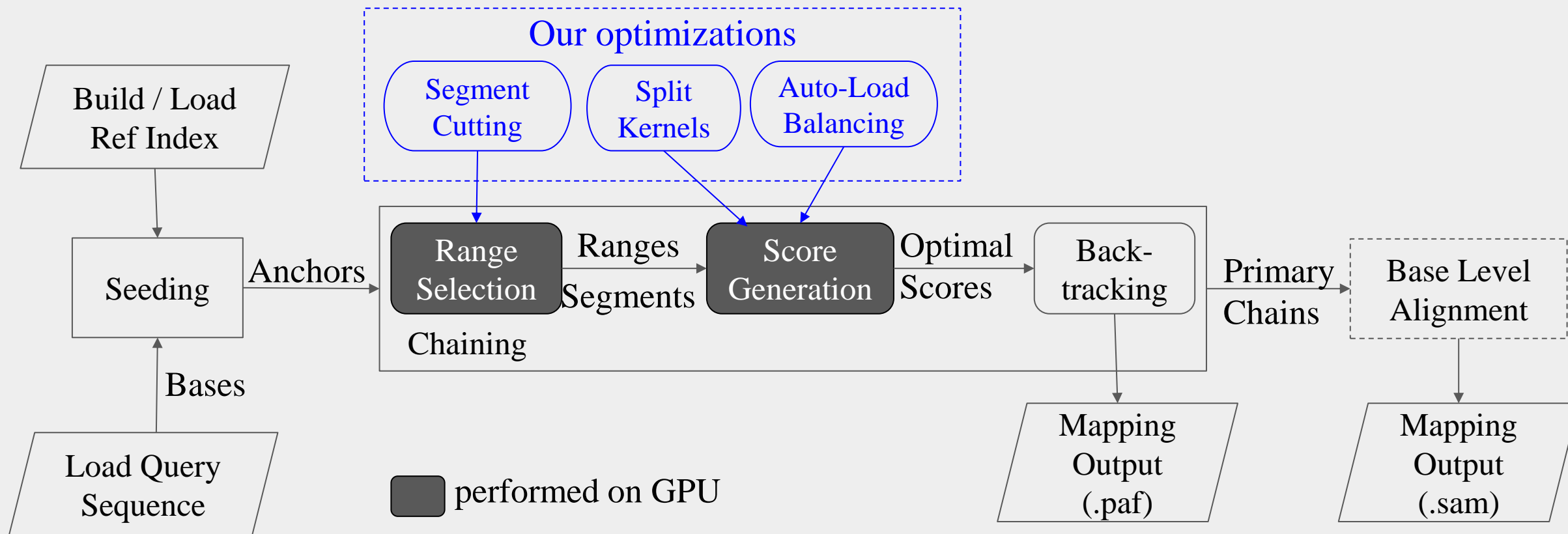
Long Read Mapping is slow



Long Read Mapping is slow

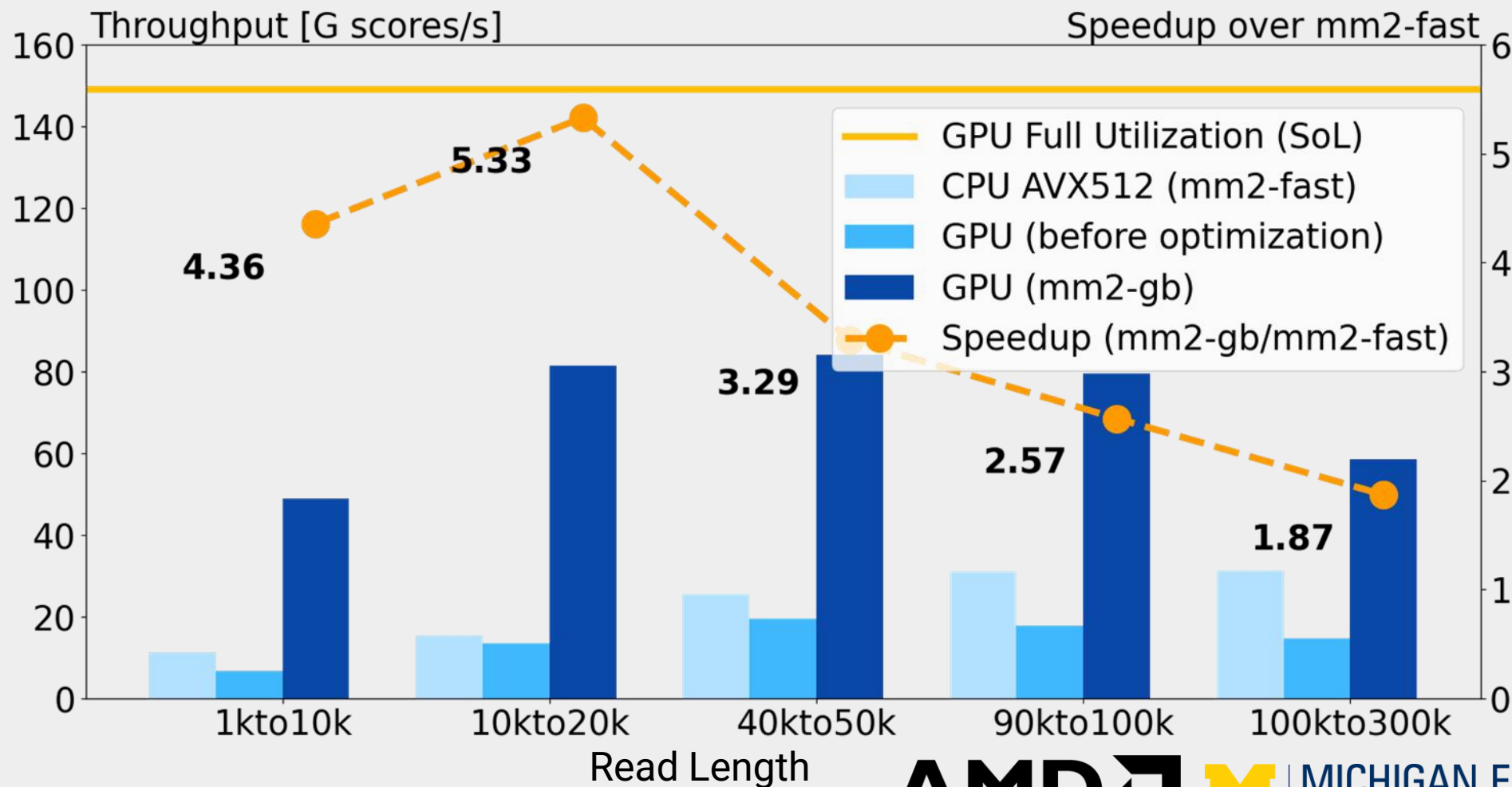


Accelerating *minimap2* on GPU



mm2-gb offers 5.33x faster chaining

No accuracy loss
Open sourced



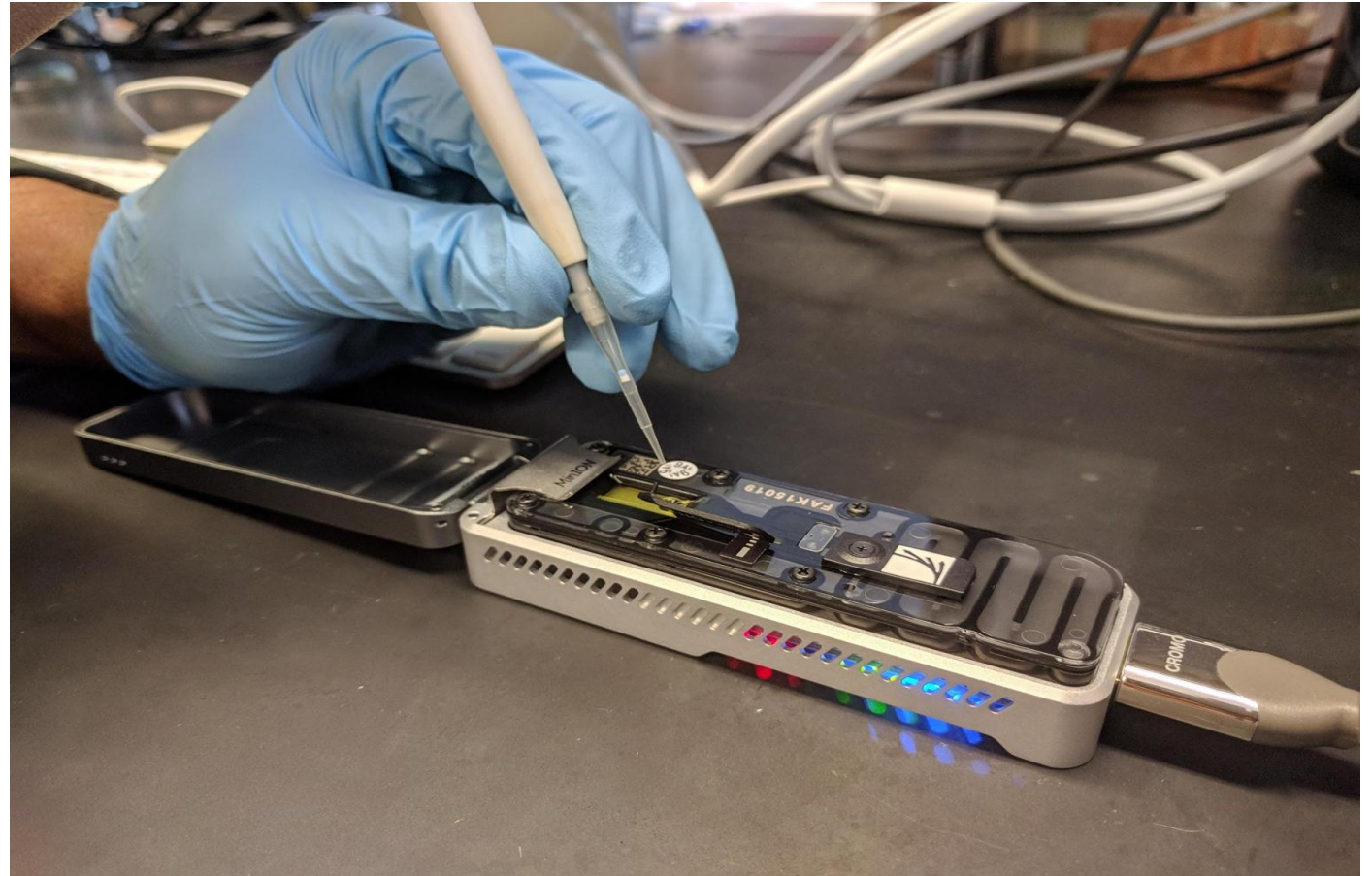
Real-Time Pathogen Detection

Dunn et al. MICRO 2021

ACM/IEEE MICRO

Top Picks Award Honorable Mention

Artifact badges



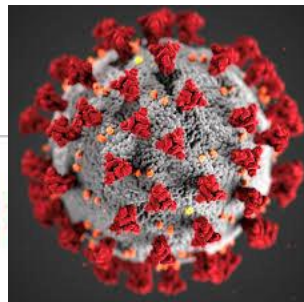
Viral Pandemics & Rise of Superbugs



Coronavirus Cases:
30,862,212
Deaths:
561,225



Coronaviruses
SARS, MERS and 2019-nCoV



The New York Times

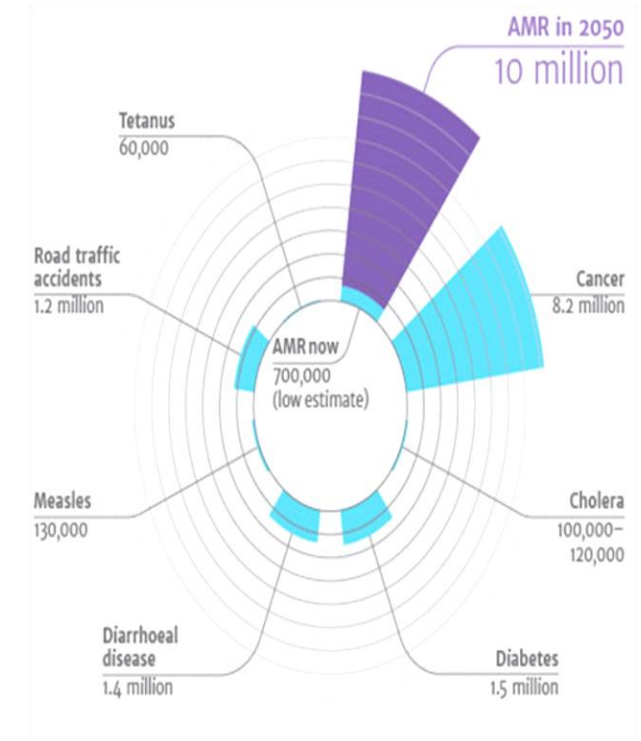
DEADLY GERMS, LOST CURES

A Mysterious Infection, Spanning the Globe in a Climate of Secrecy

The rise of *Candida auris* embodies a serious and growing public health threat: drug-resistant germs.



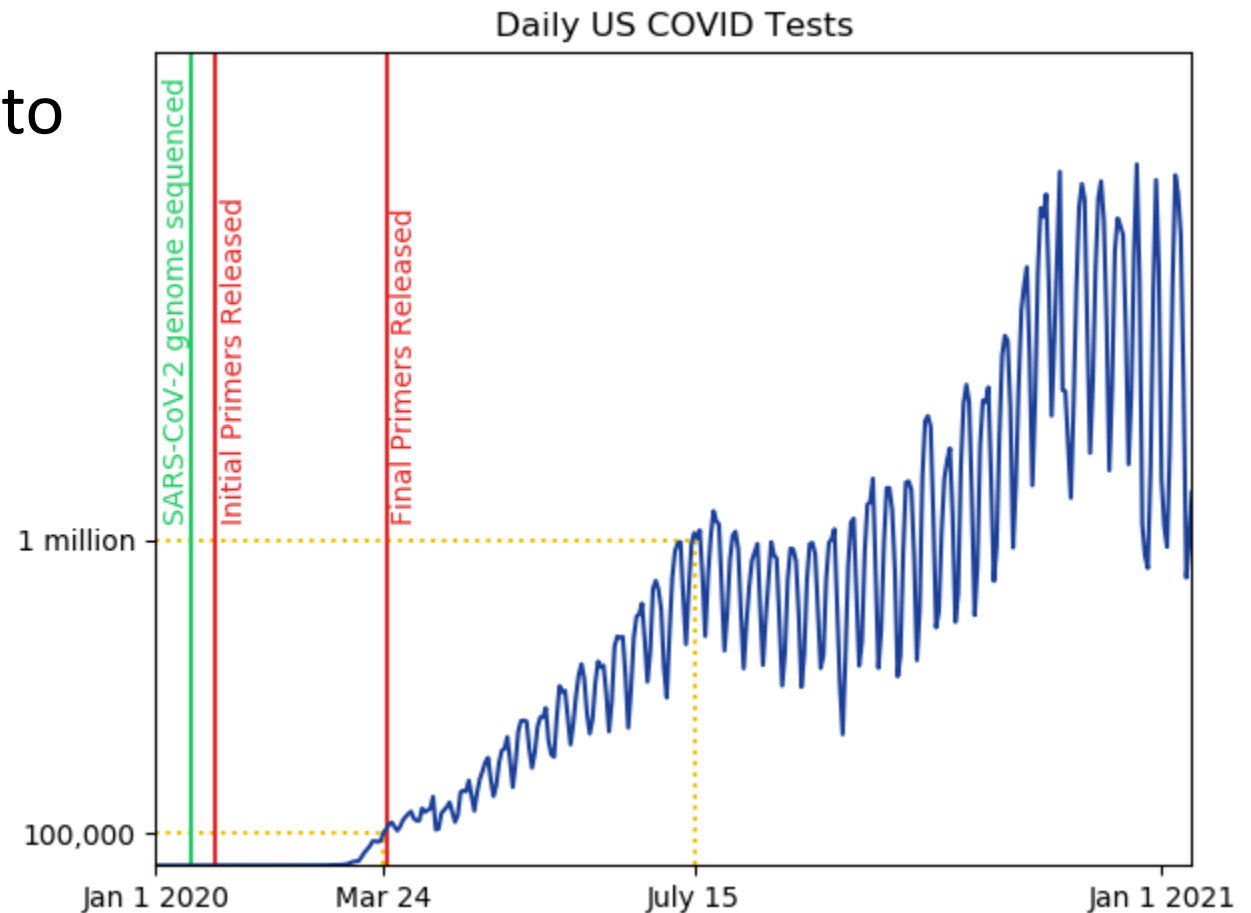
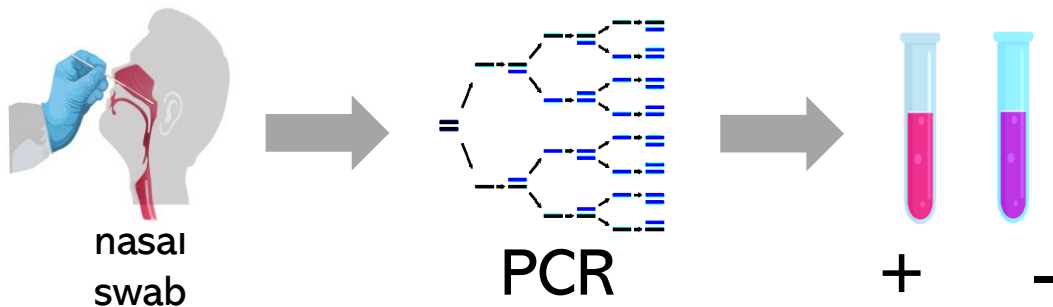
Superbugs will kill more than cancer by 2050
[2019 UN report: "No Time To Wait"](#)



It Took Months For Mass COVID Testing Capabilities

Custom primers time-consuming to

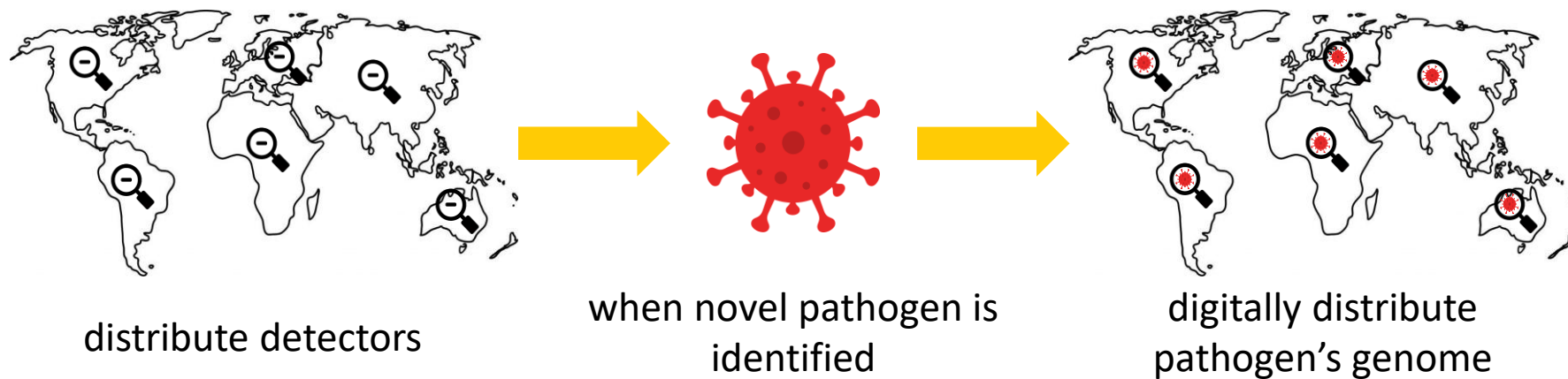
- Design
- Verify
- Manufacture
- Distribute



How can we be ready for the next pandemic?

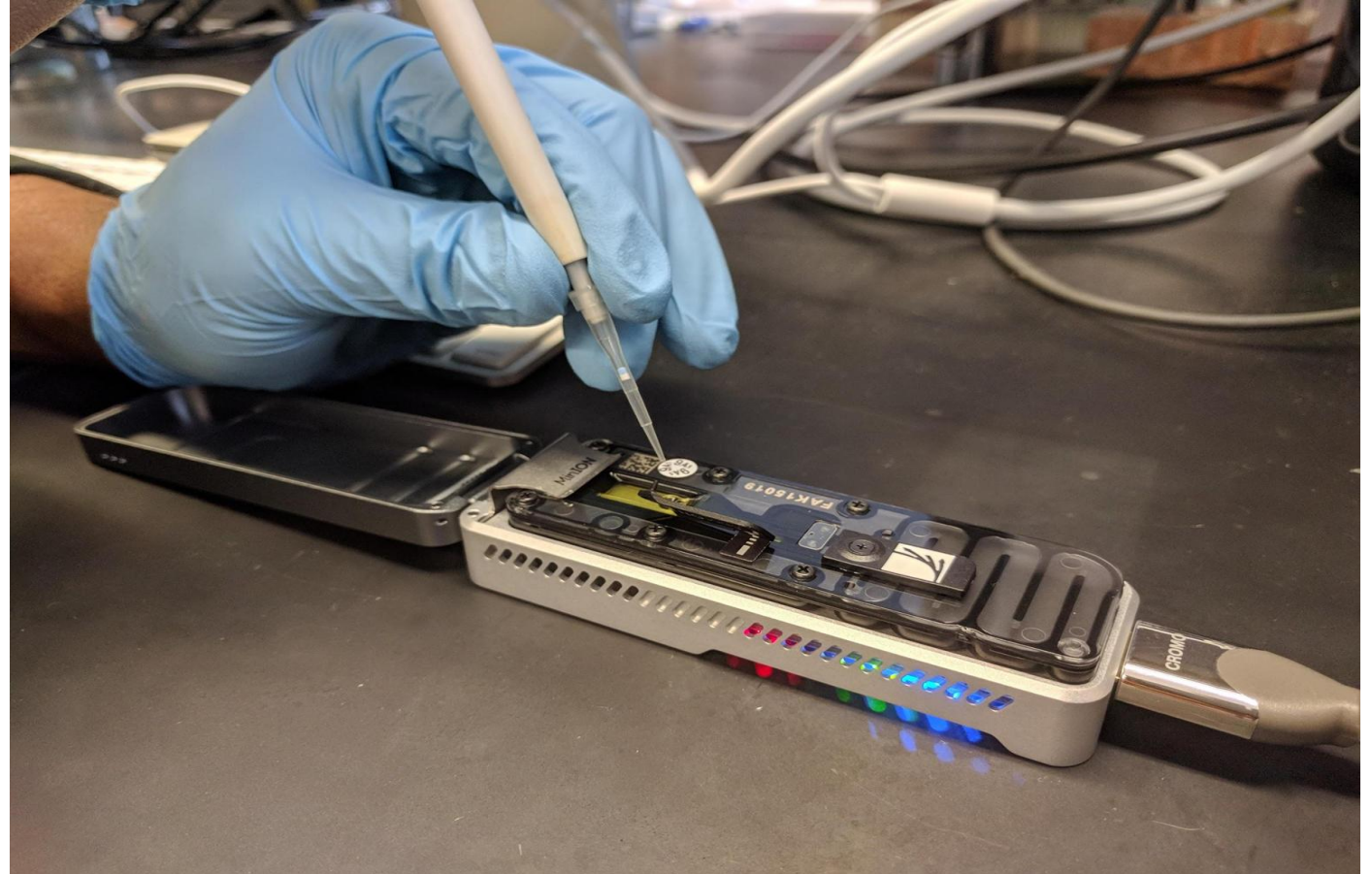
Portable Virus Detector

- Digitally programmable using the target virus's genome

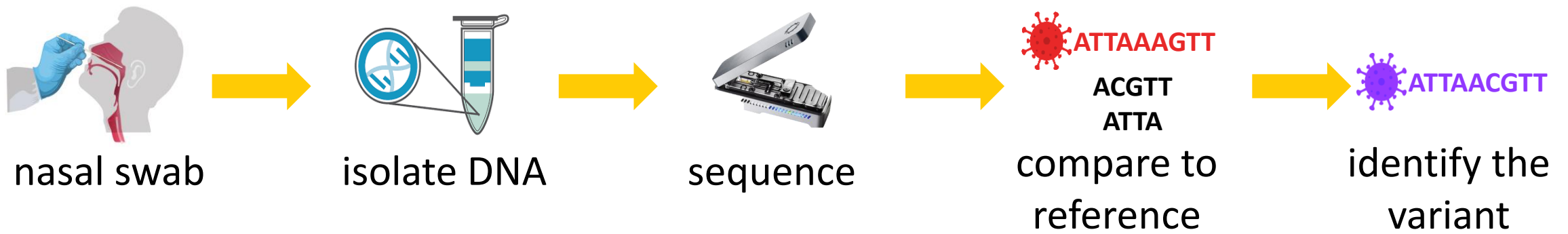


MinION: Portable Nanopore Sequencer

- Recent-to-market
- Portable
- Fast (real-time)
 - 512 sequencing channels
 - 450 bases per second, each
- Relatively Low Cost
- Long Reads



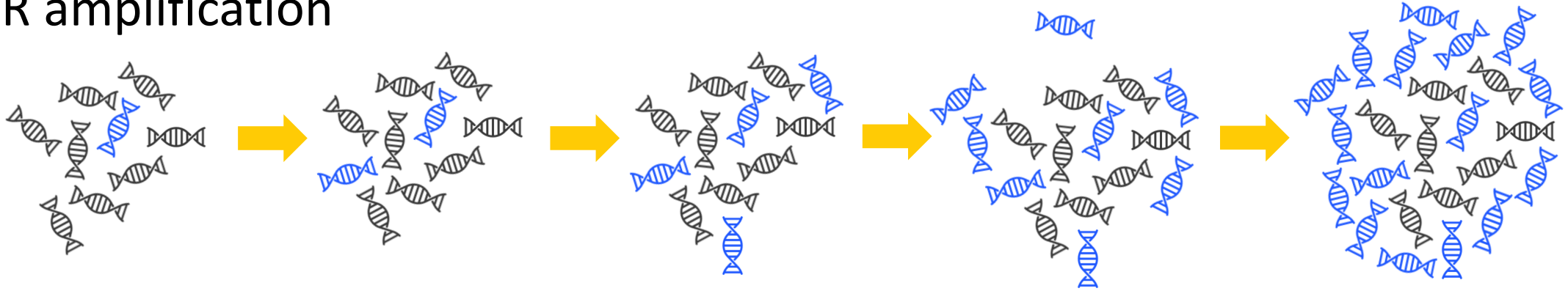
Portable Virus Detection



Problem: >99% of a sample is non-viral DNA

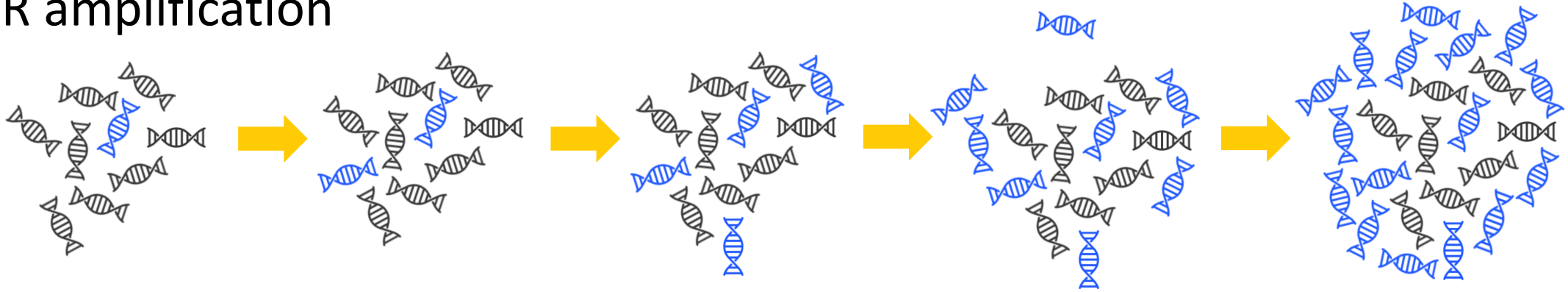
Problem: >99% of a sample is non-viral DNA

- PCR amplification

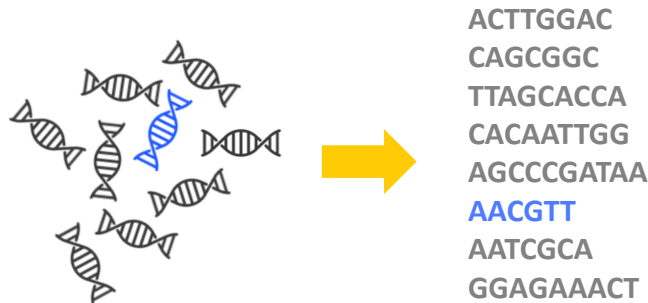


Problem: >99% of a sample is non-viral DNA

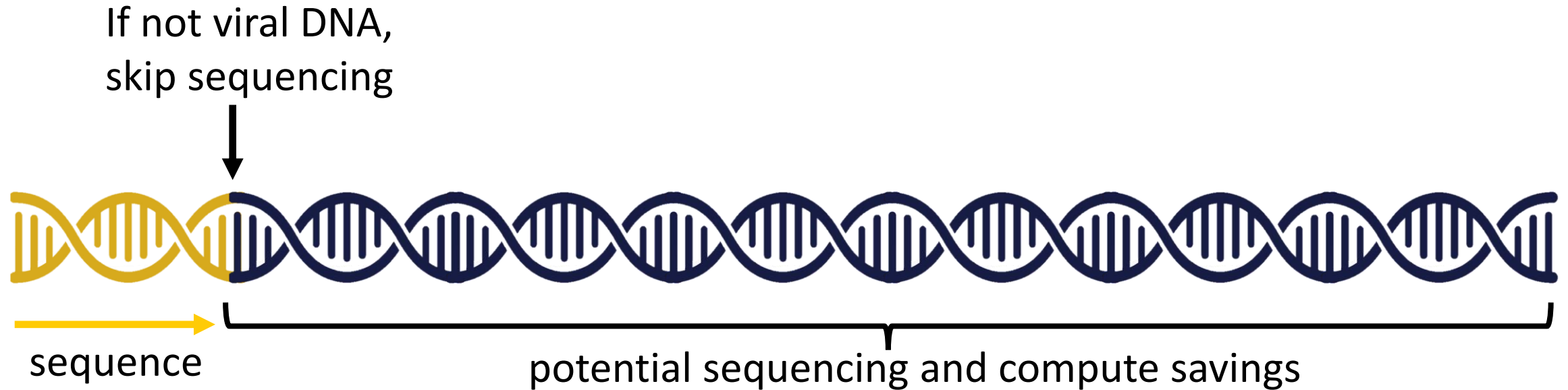
- PCR amplification



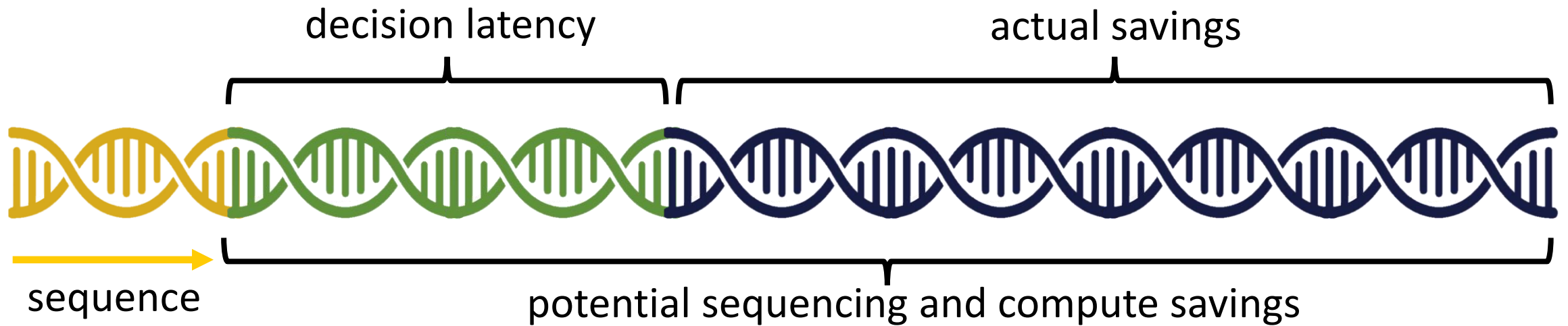
- Sequencing



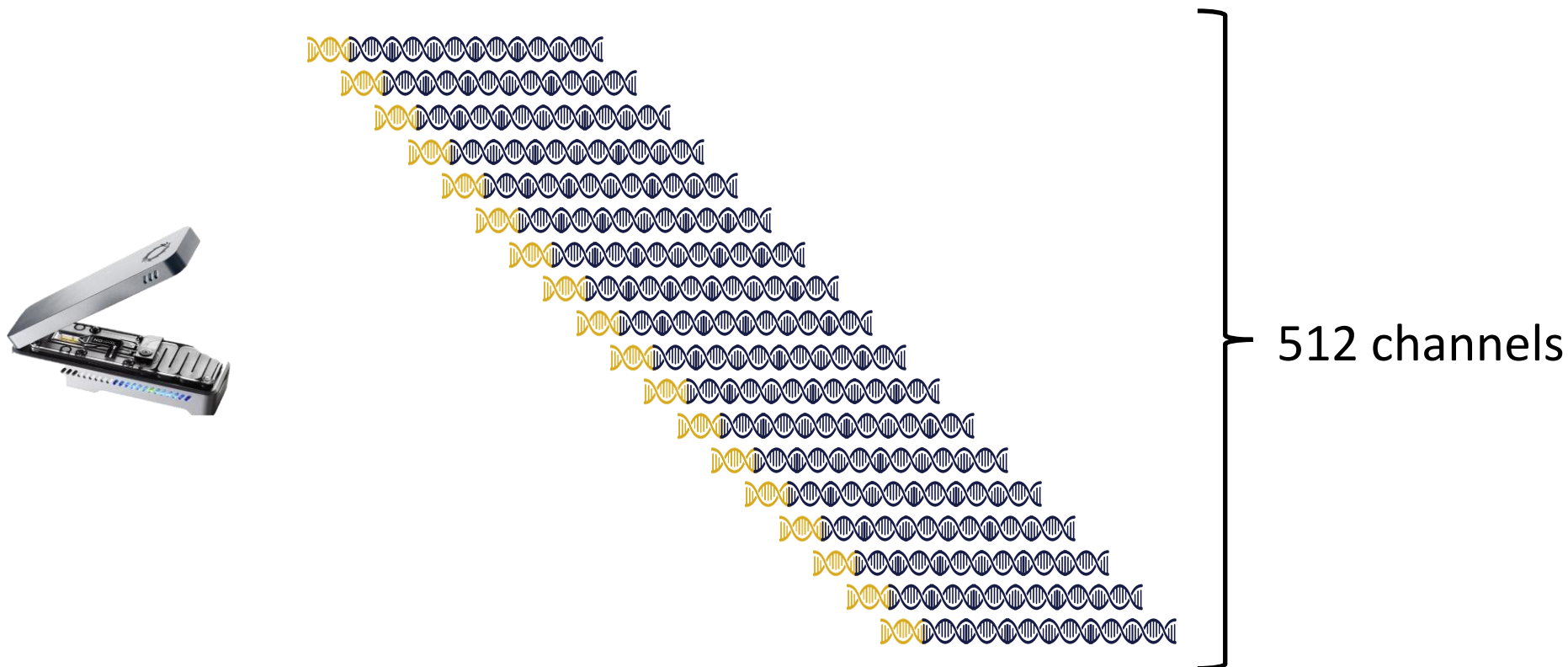
Solution: Read Until - skip sequencing non-viral reads



Challenge: Requires low latency computing



Challenge: Requires high throughput computing



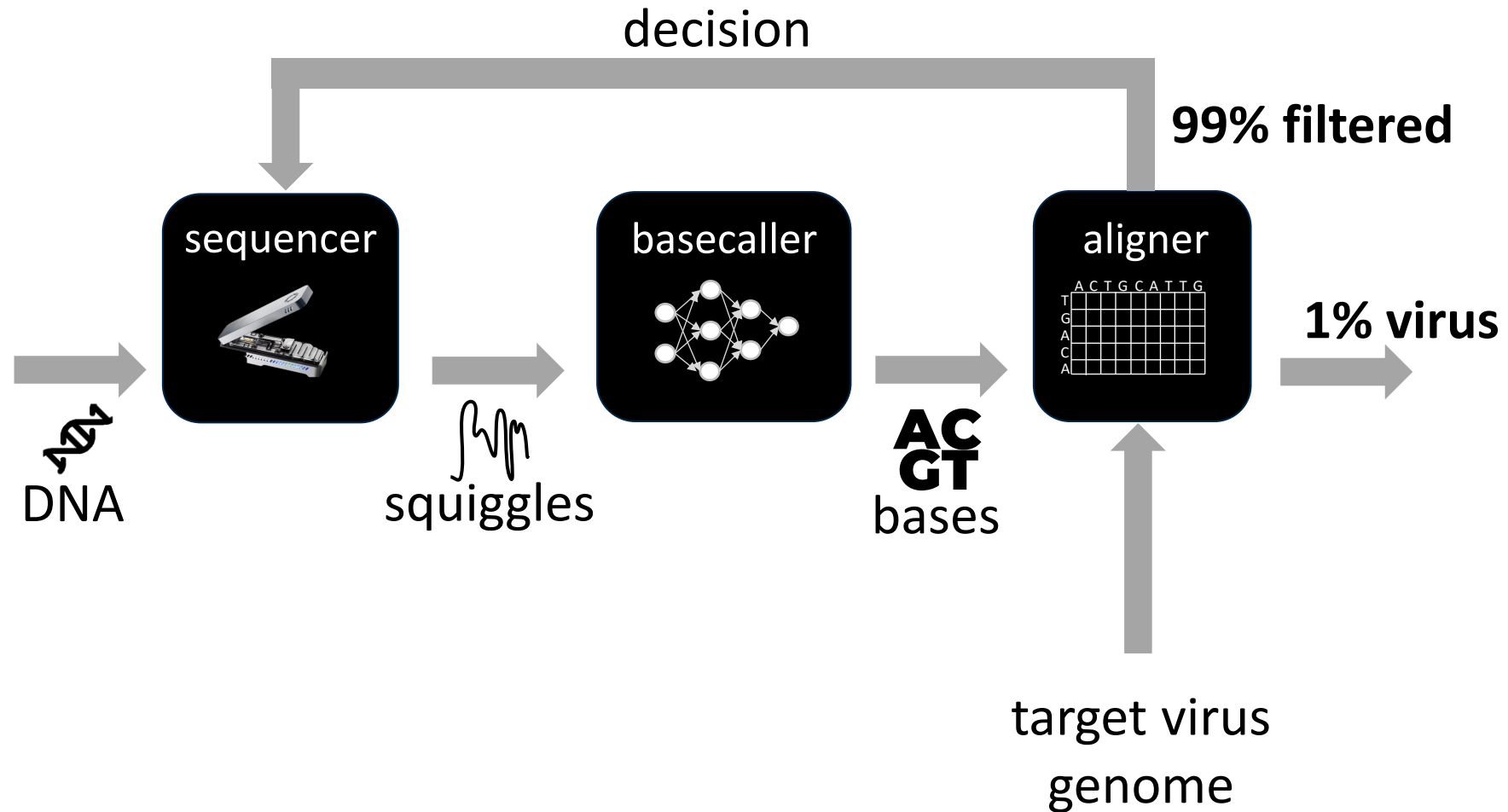
Challenge: Portability



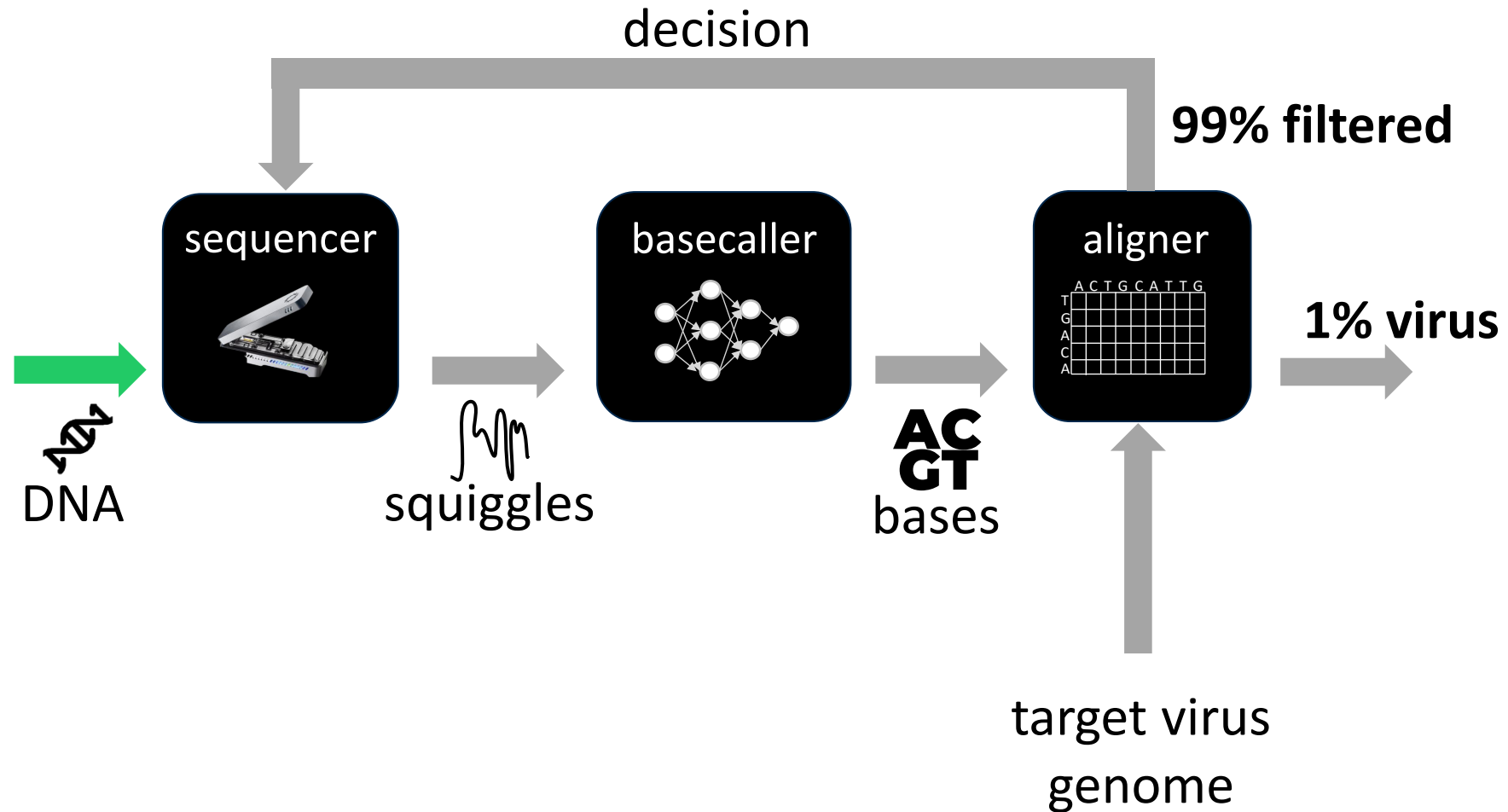
Problem: No compute capability

Goal: Efficient data analysis for portable detection

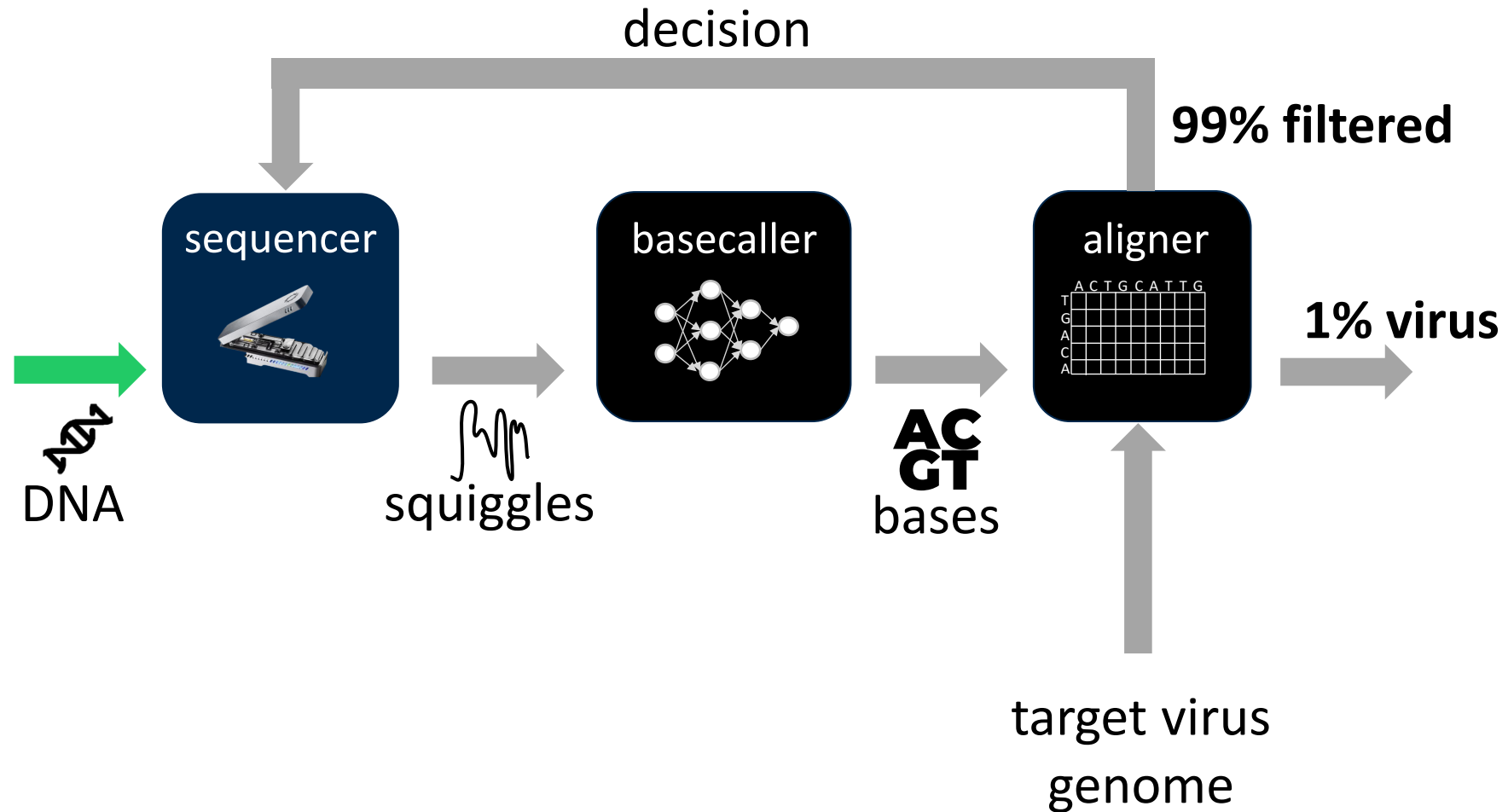
Read Until pipeline



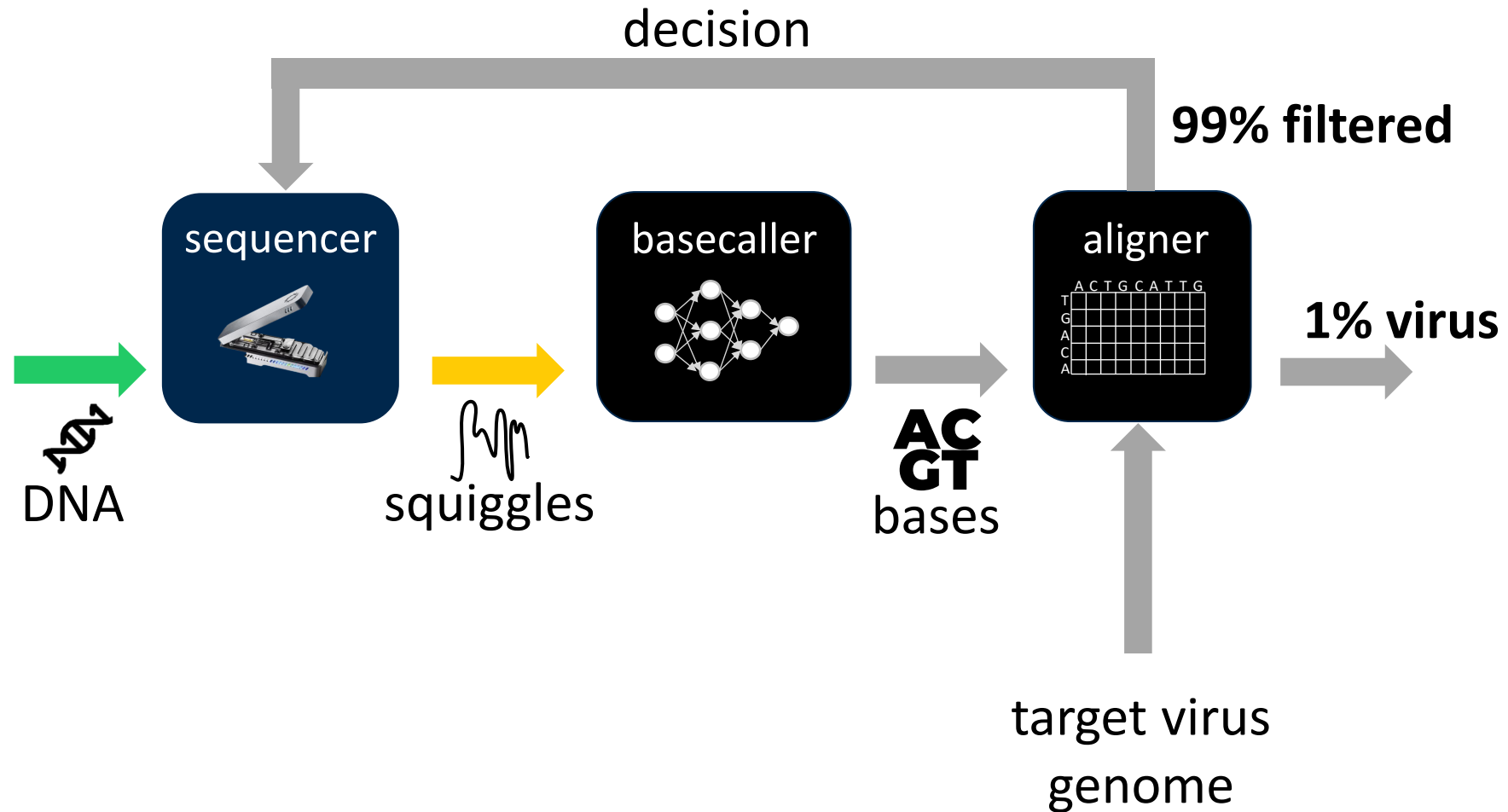
Read Until pipeline



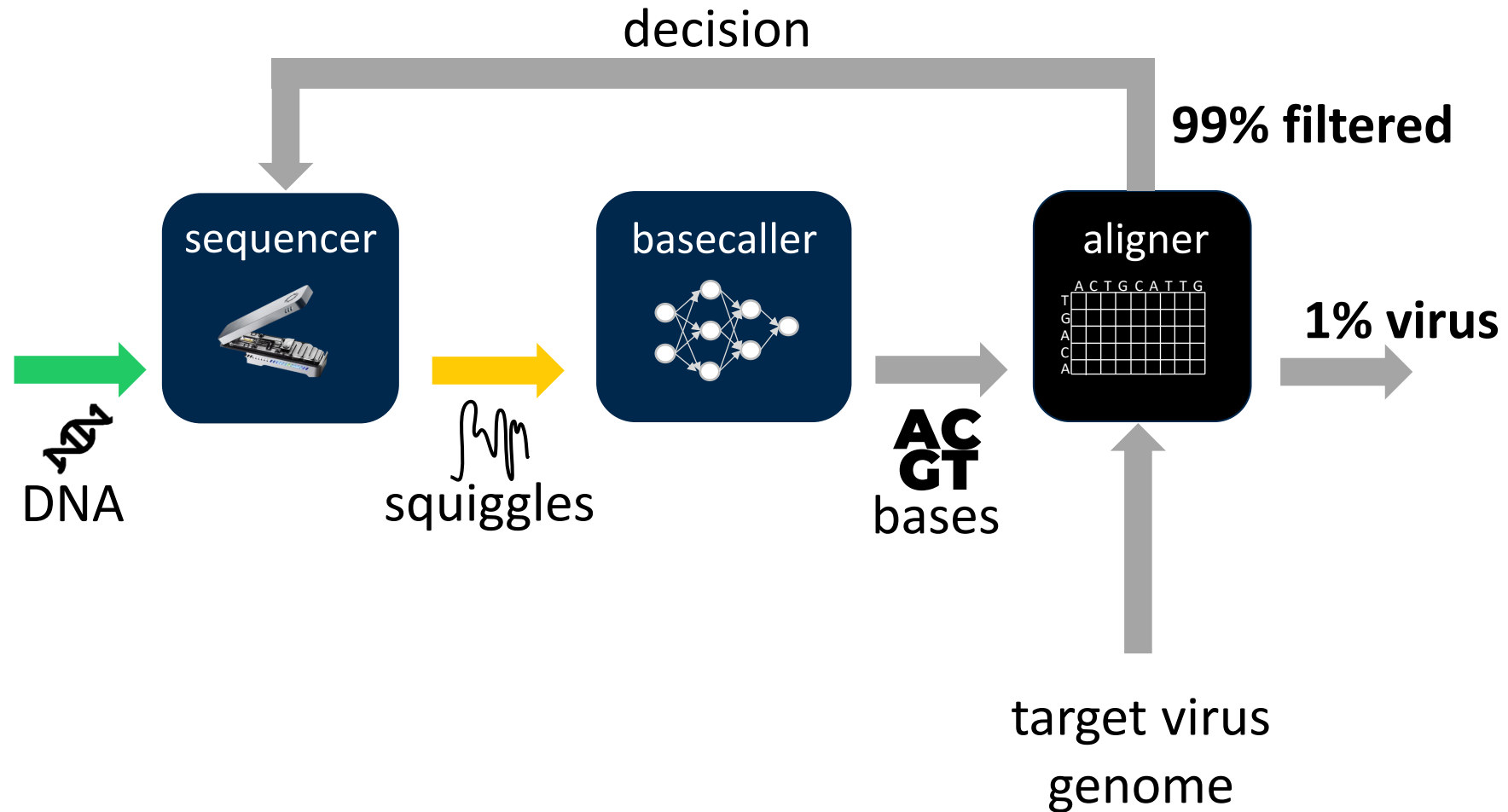
Read Until pipeline



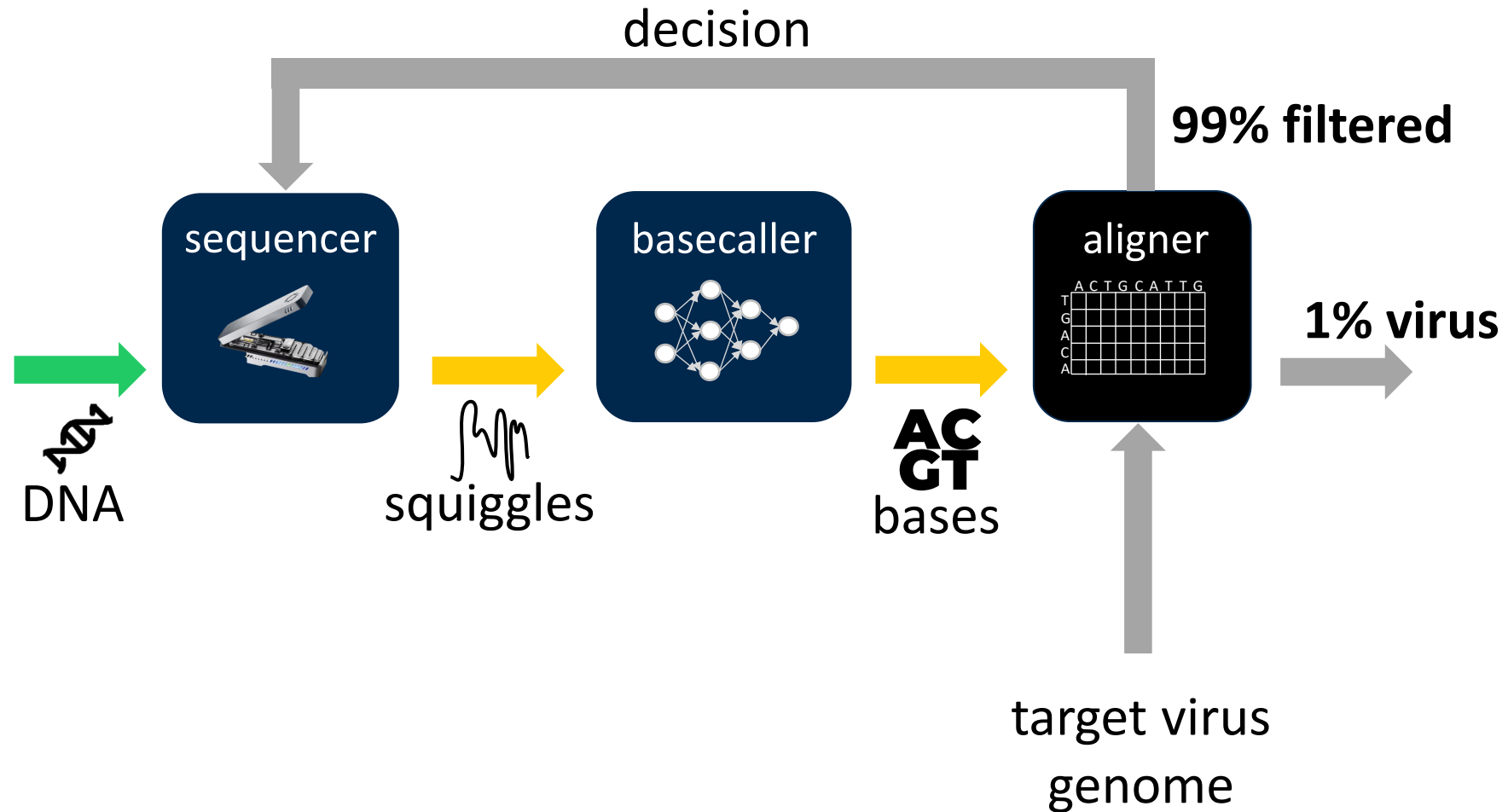
Read Until pipeline



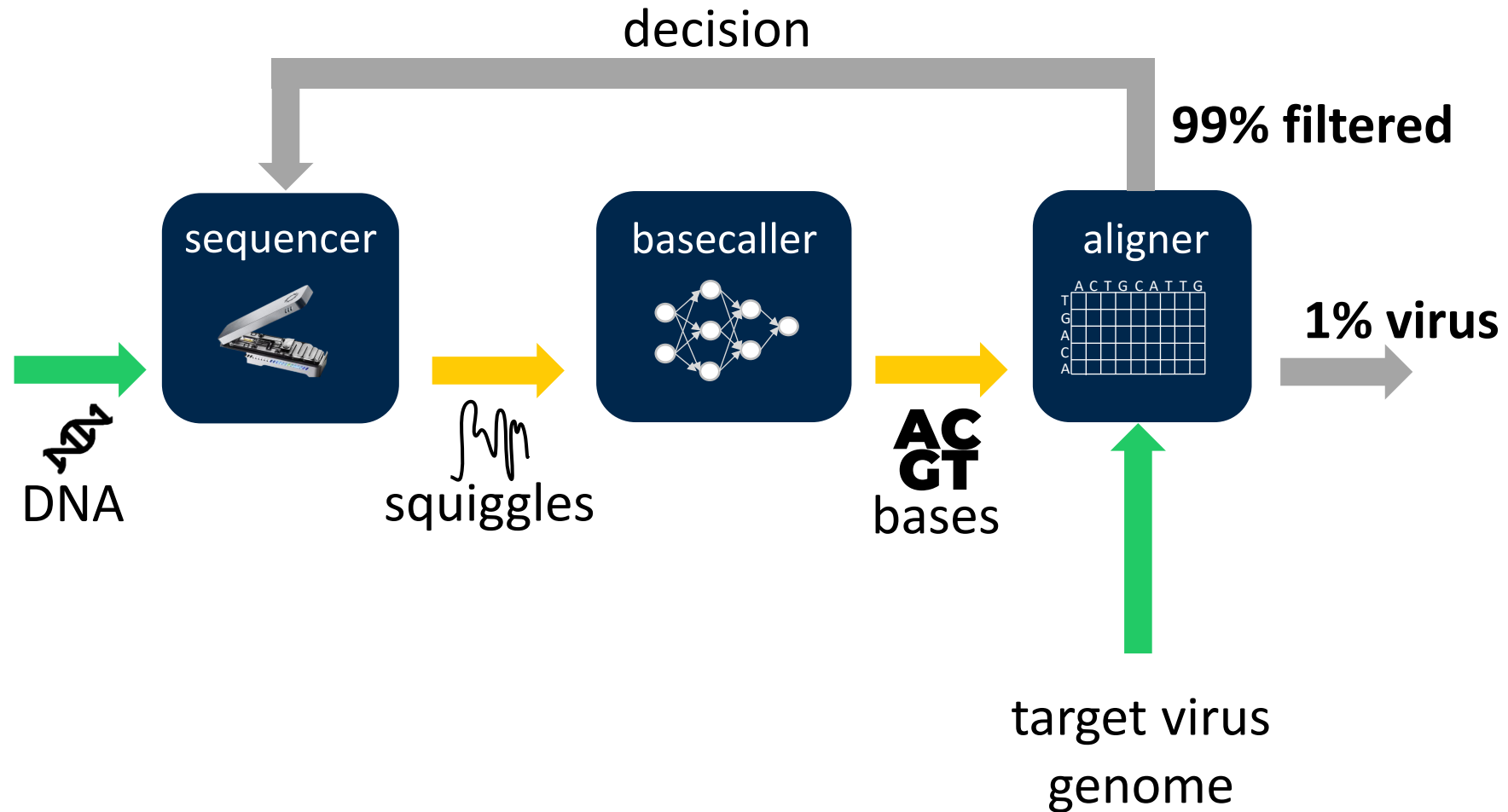
Read Until pipeline



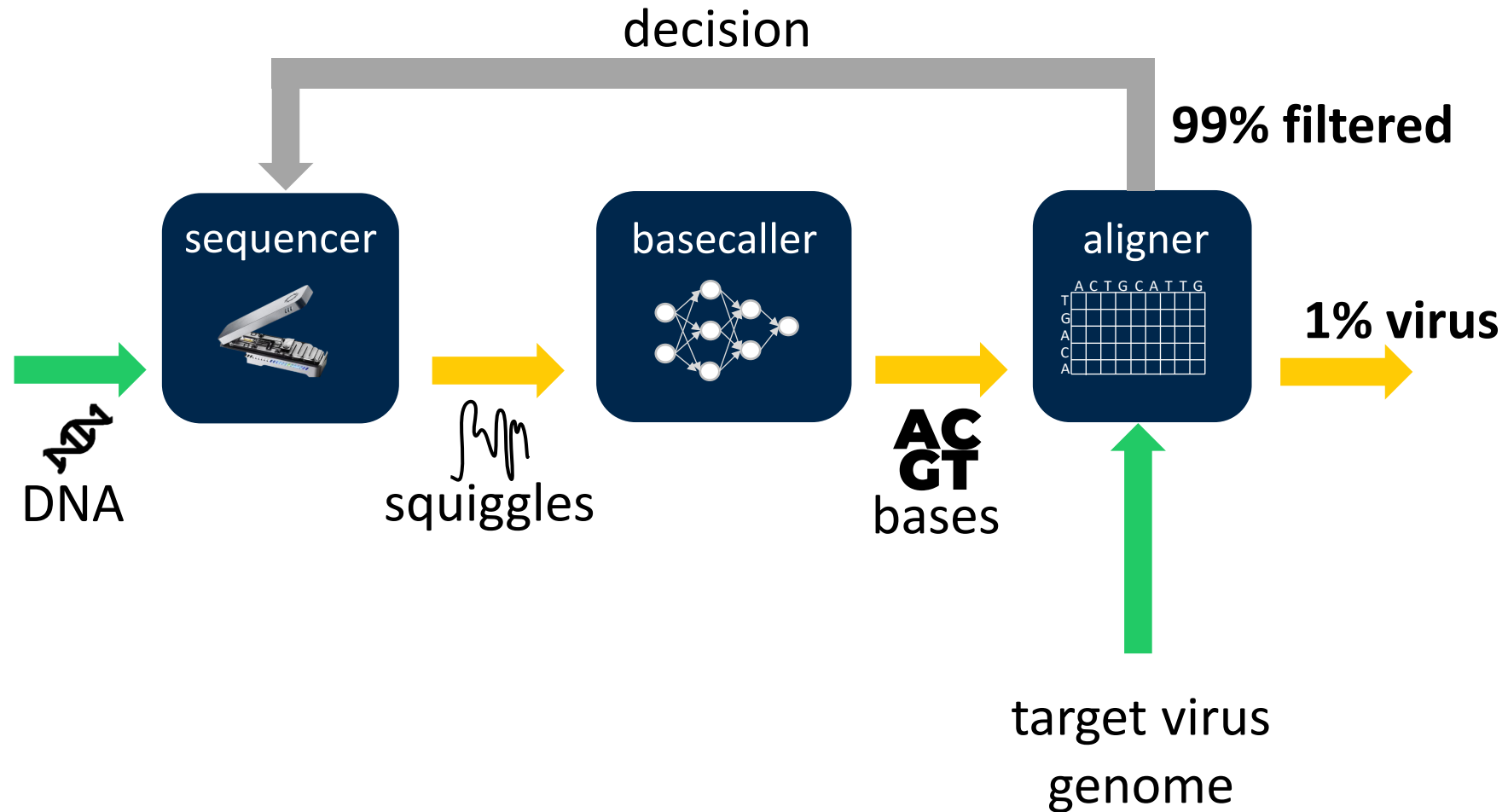
Read Until pipeline



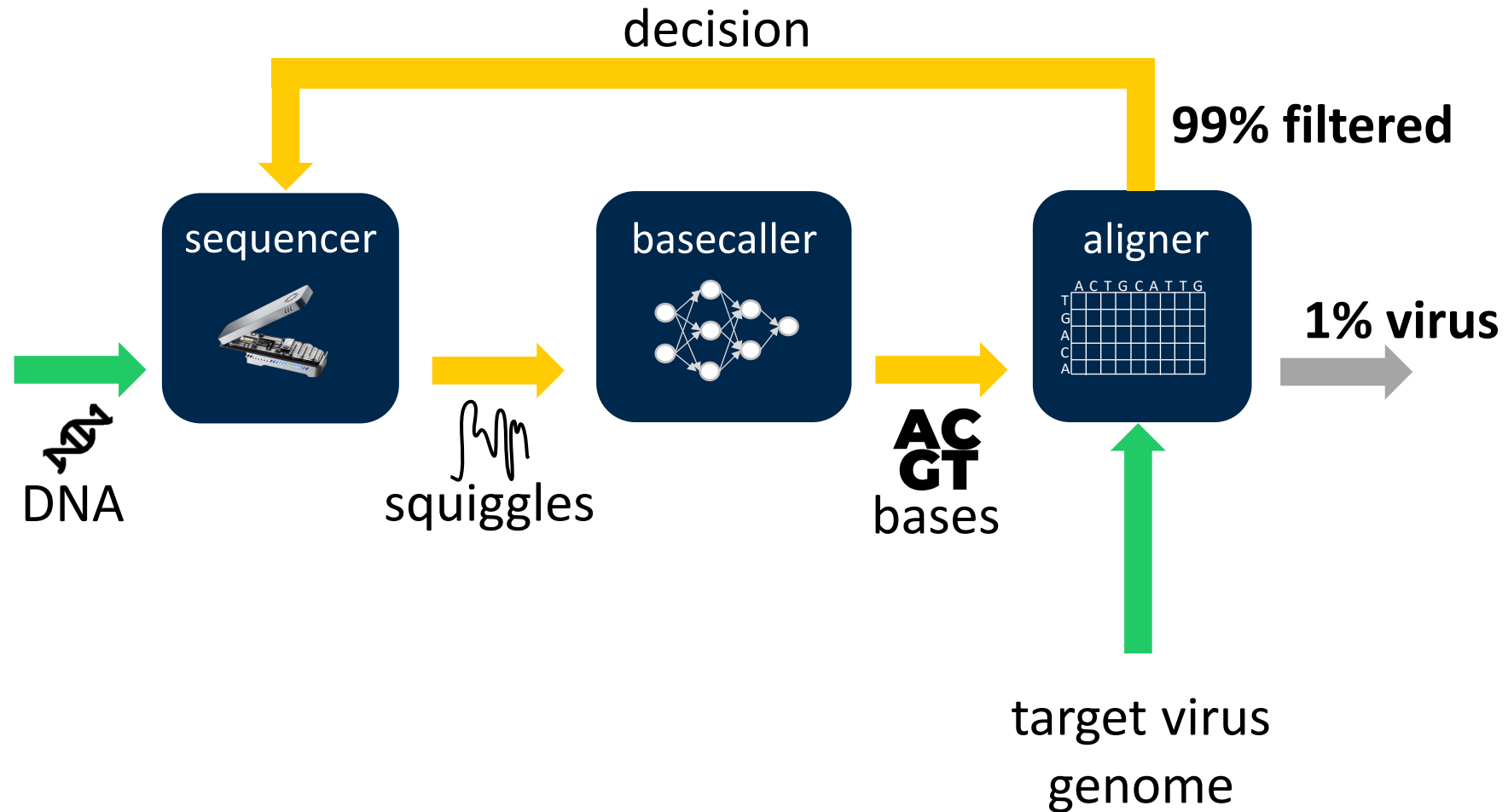
Read Until pipeline



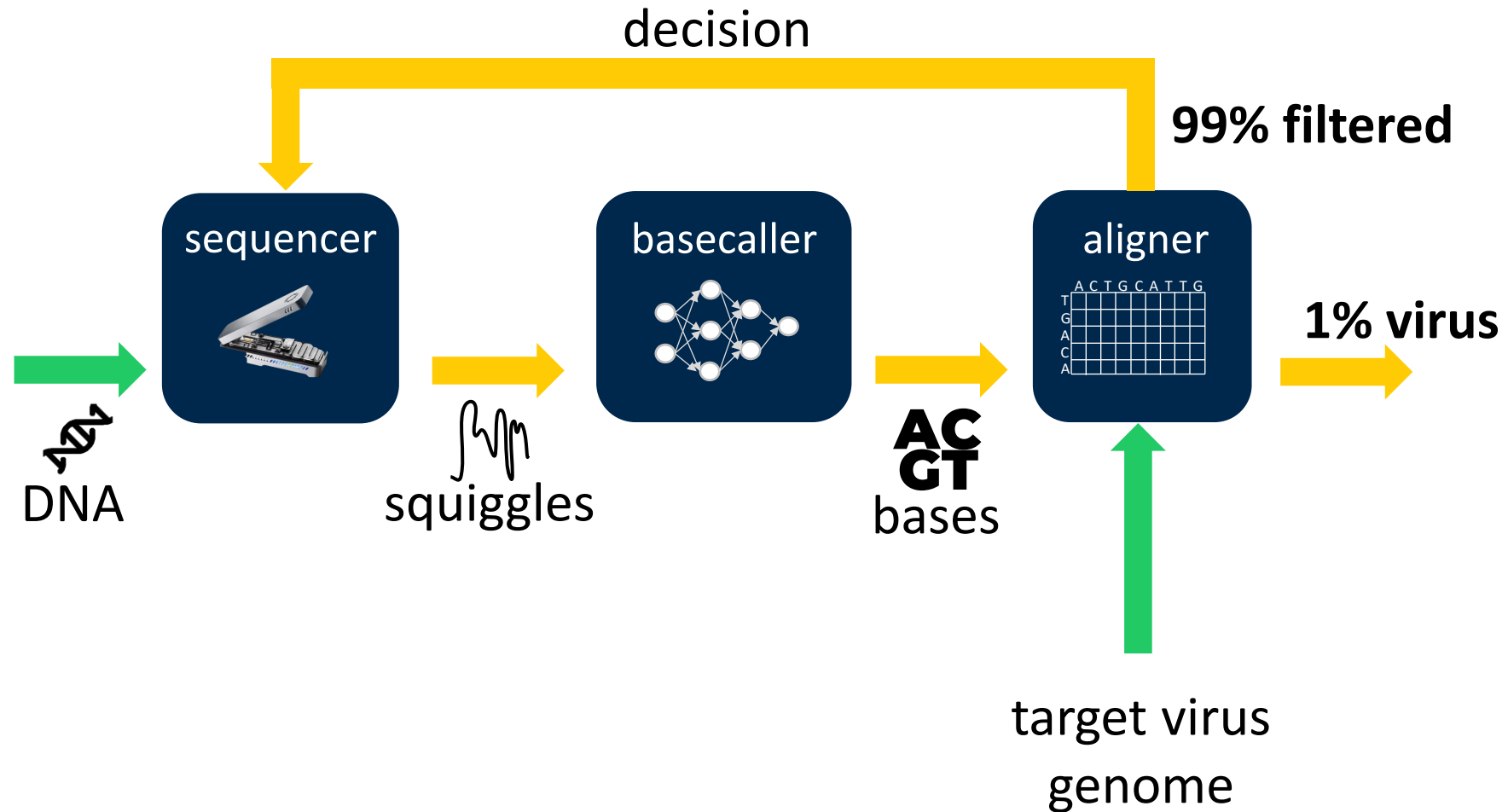
Read Until pipeline



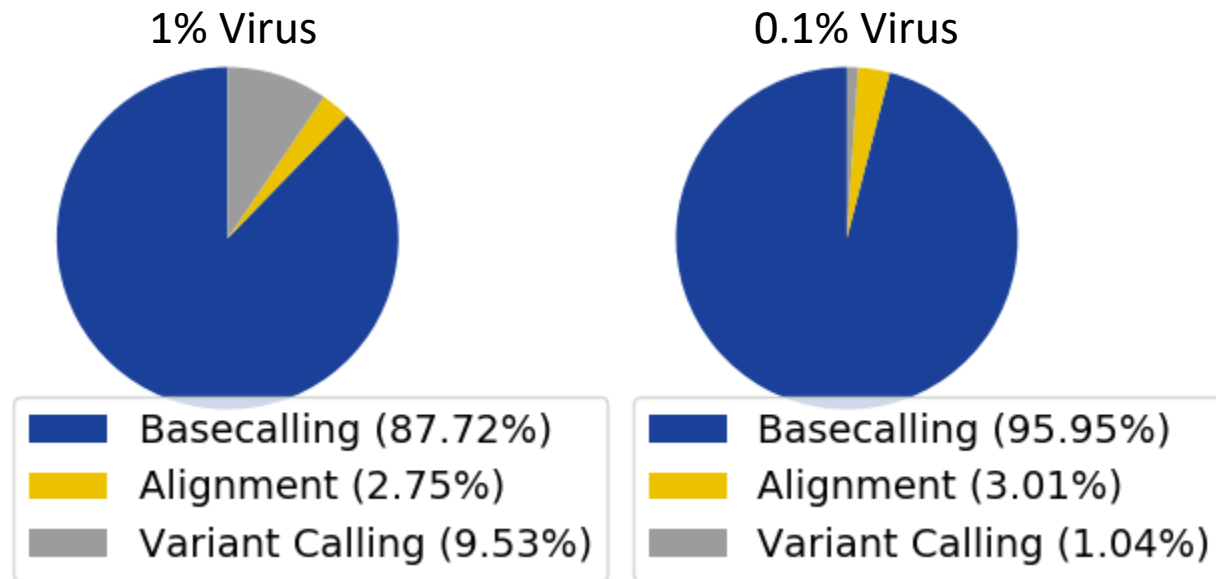
Read Until pipeline



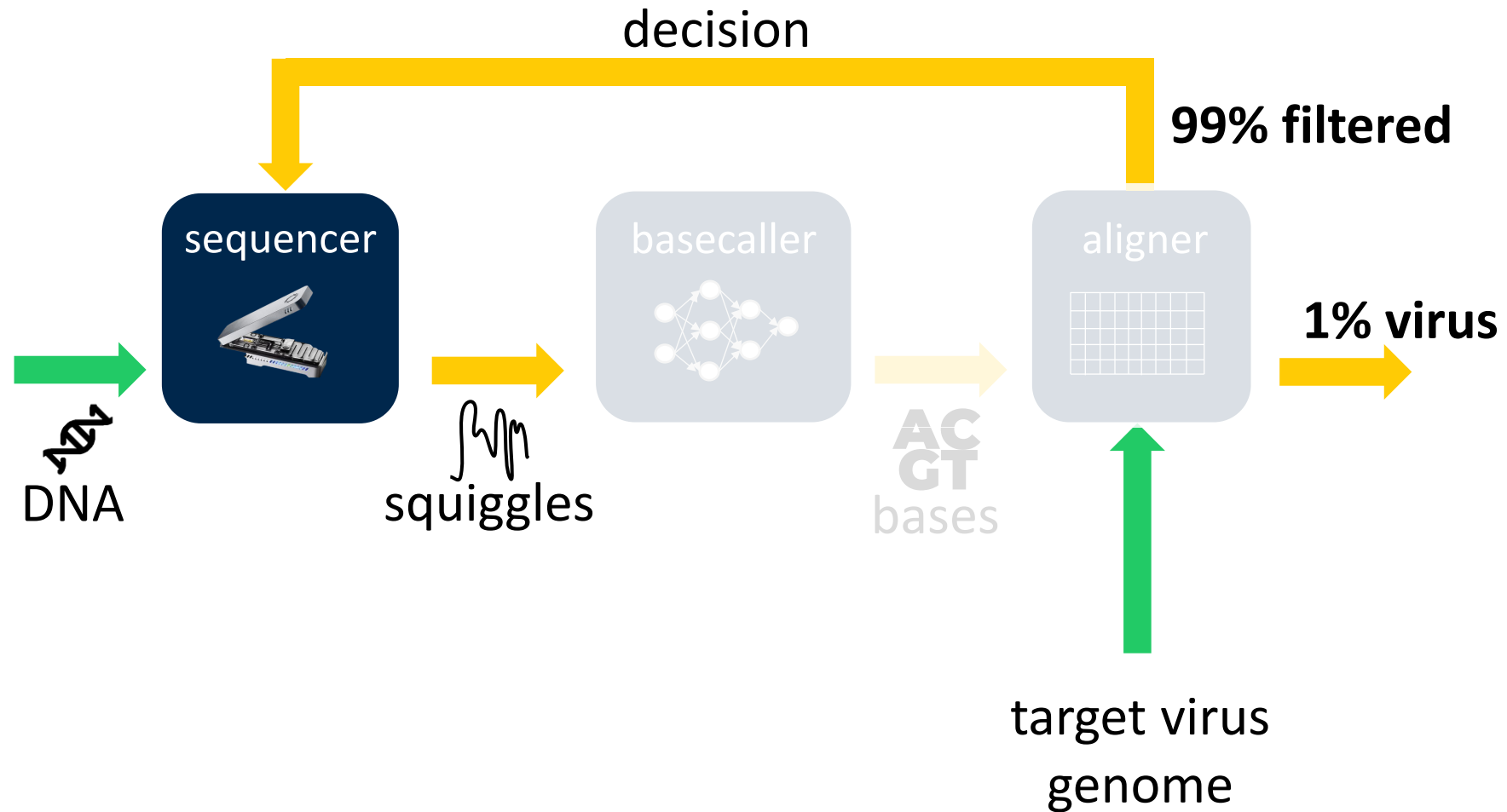
Read Until pipeline



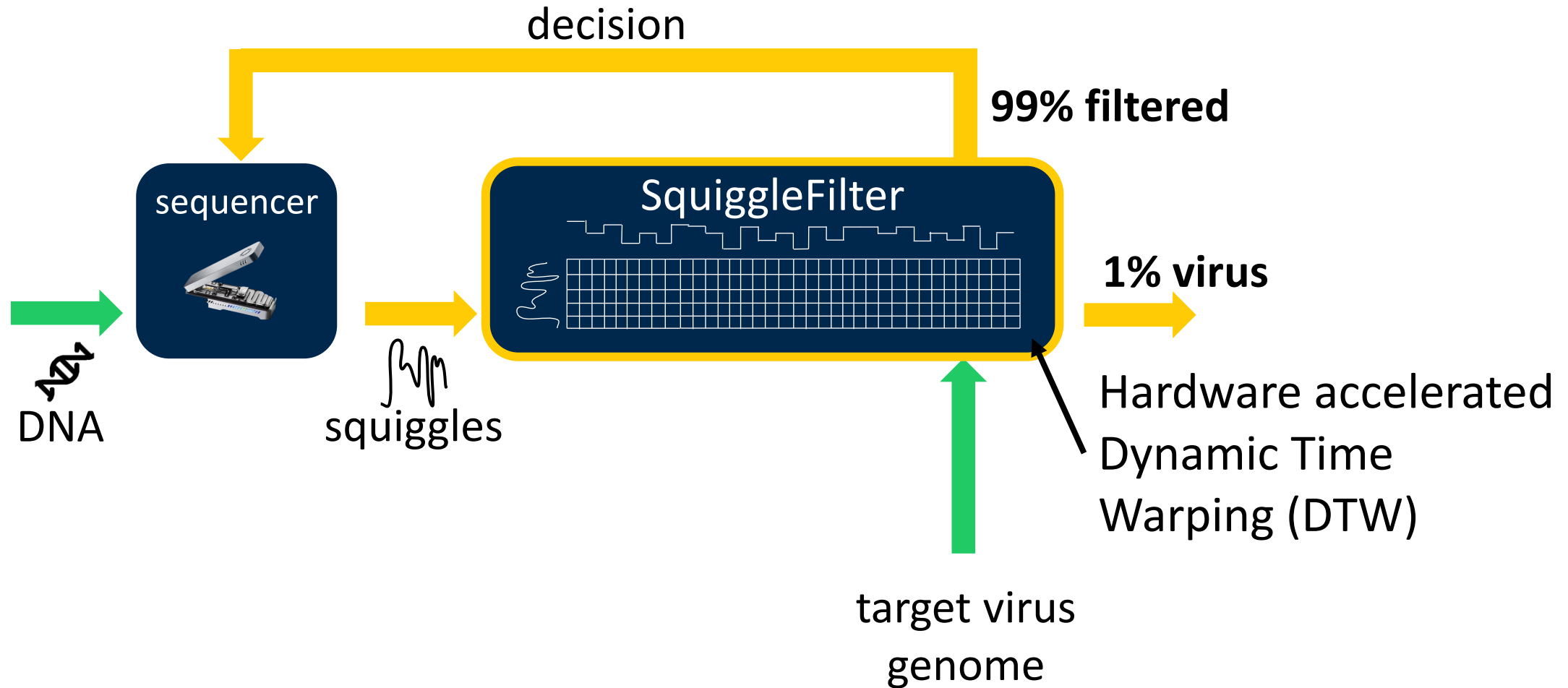
Problem: Basecalling is compute-intensive



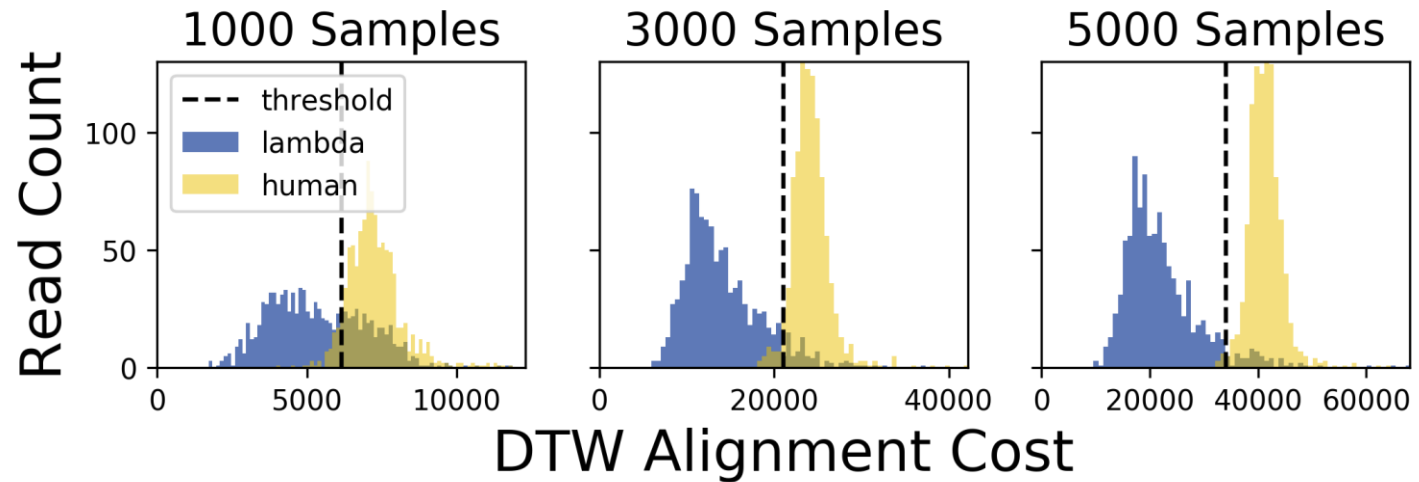
Idea: Skip basecalling, align squiggles



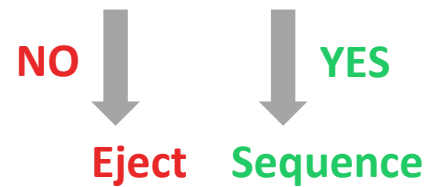
Contribution: Accelerated squiggle-level non-viral filter



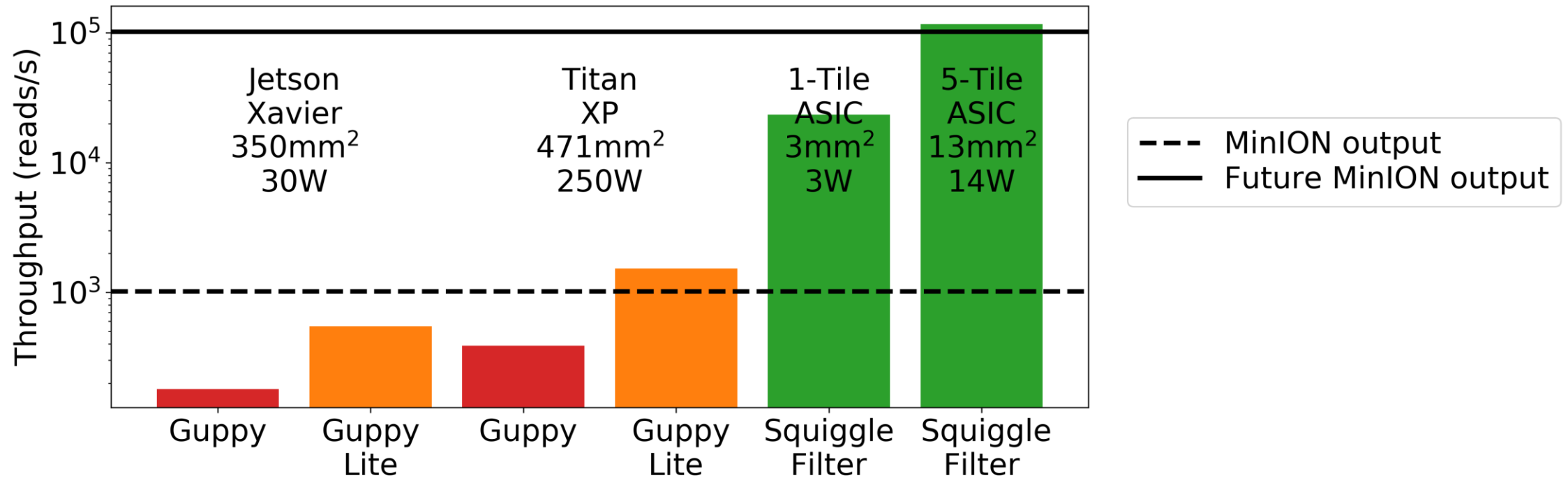
Filtering: Alignment Cost Threshold



Read Until: alignment cost < threshold?



Results: SquiggleFilter Throughput

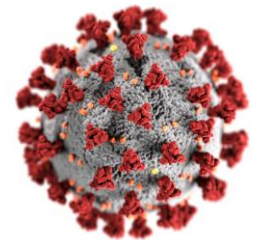


Sequence pathogens < 30 min

2,725x higher throughput/area

SARS-CoV-2

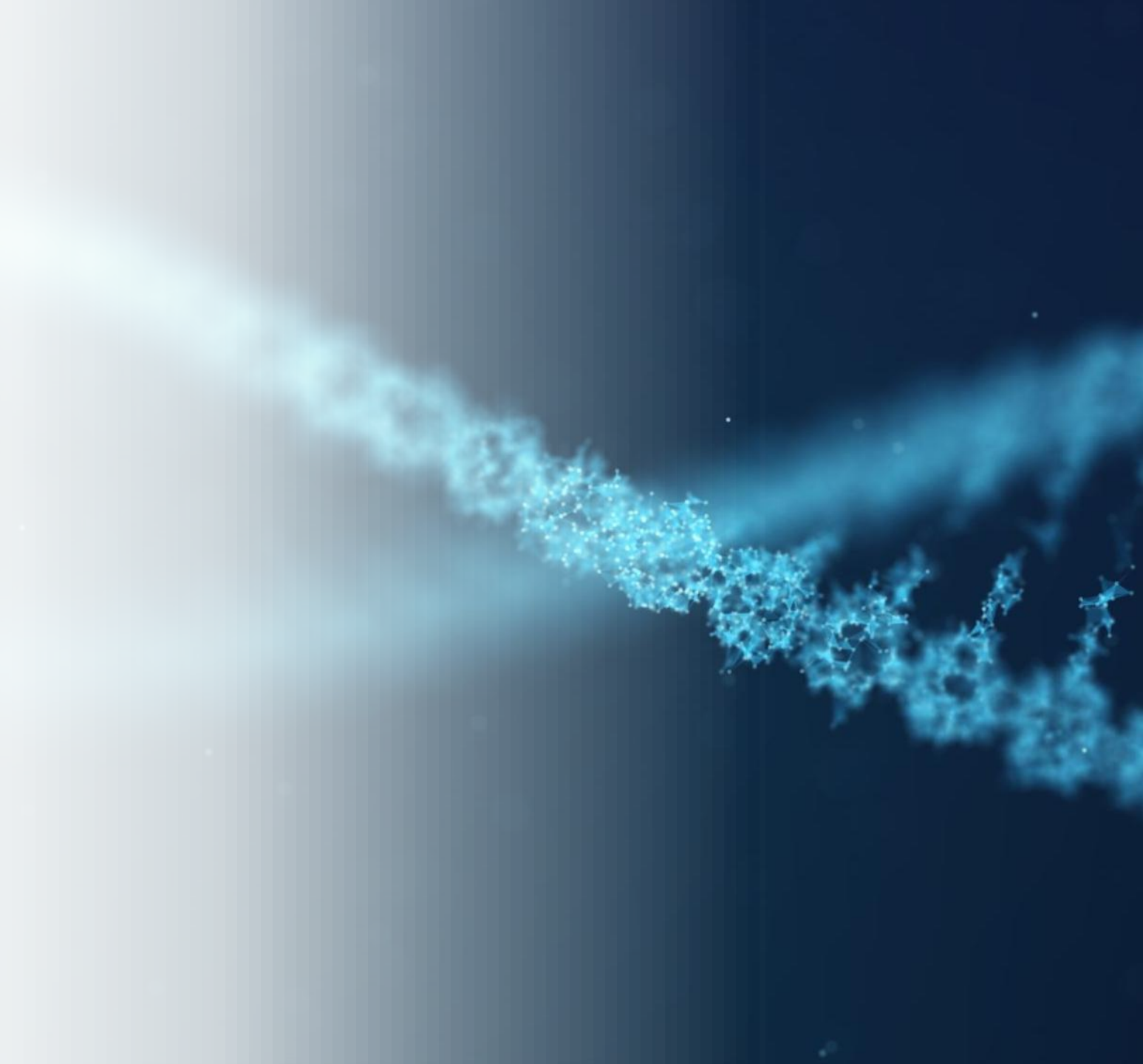
Reference: 29,970 bases
NCBI Database





Ultra Rapid Cancer Diagnosis

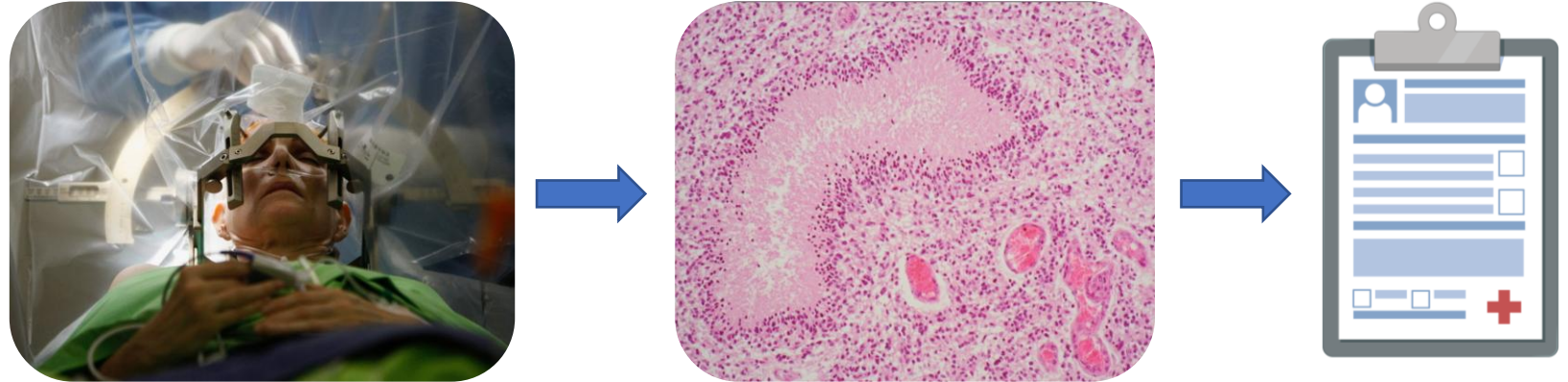
Wadden et al.
Communications Biology 2022



Intra-operative sequencing for accurate cancer diagnostics

- Intra-operative histology can help guide surgical decision making and combine surgeries
- Histology is subjective, and does not contain molecular information
- Genetic information is becoming increasingly important for diagnosis and targeted, personalized treatment!

Frozen Section Histology can return a diagnosis in ~20-40 min



REVIEW

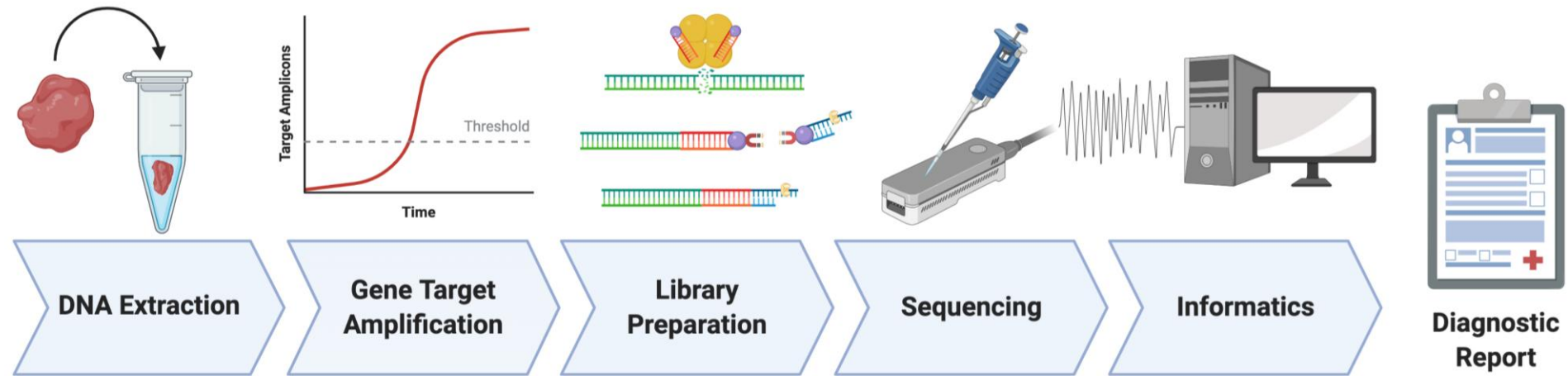
The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary

David N. Louis¹ · Arie Perry² · Guido Reifenberger^{3,4} · Andreas von Deimling^{4,5} · Dominique Figarella-Branger⁶ · Webster K. Cavenee⁷ · Hiroko Ohgaki⁸ · Otmar D. Wiestler⁹ · Paul Kleihues¹⁰ · David W. Ellison¹¹

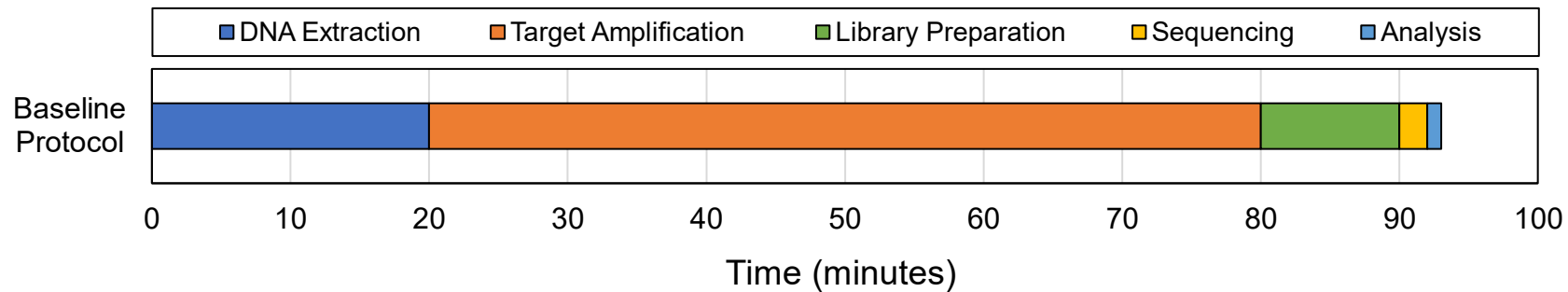
“For the first time, the WHO classification of CNS tumors *uses molecular parameters* in addition to histology to define many tumor entities, thus formulating a concept for how CNS tumor diagnoses should be structured in the molecular era.”

Can we sequence a tumor's DNA within the intra-operative time frame? (i.e. <1hr)

How does a sequencing-based molecular diagnostic work?



PCR amplifies a small cancer gene target
Amplified targets are sequenced to detect cancer mutation



Target amplification is the obvious bottleneck. How can we attack this?

Threshold Sequencing

Co-optimize amplification time and sequencing time to minimize time-to-result

1) Build a model to estimate total diagnostic time

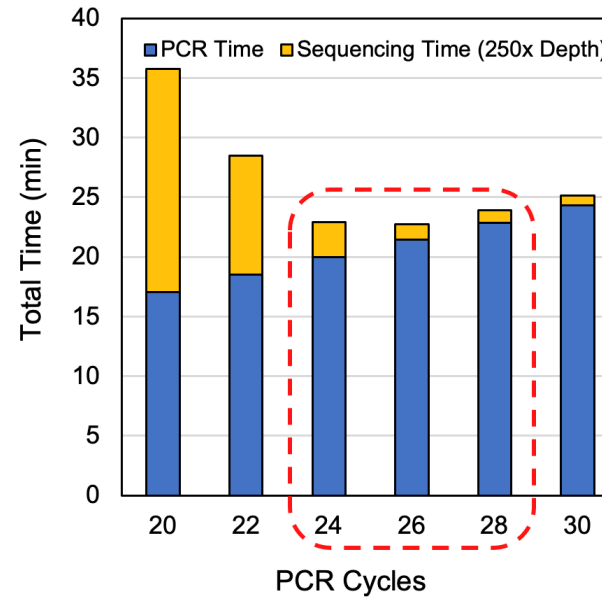
$$T_{total} = T_{amp} + T_{seq}$$

$$T_{amp} = T_{init} + T_{cycle} \times N_{cycle} + T_{final}$$

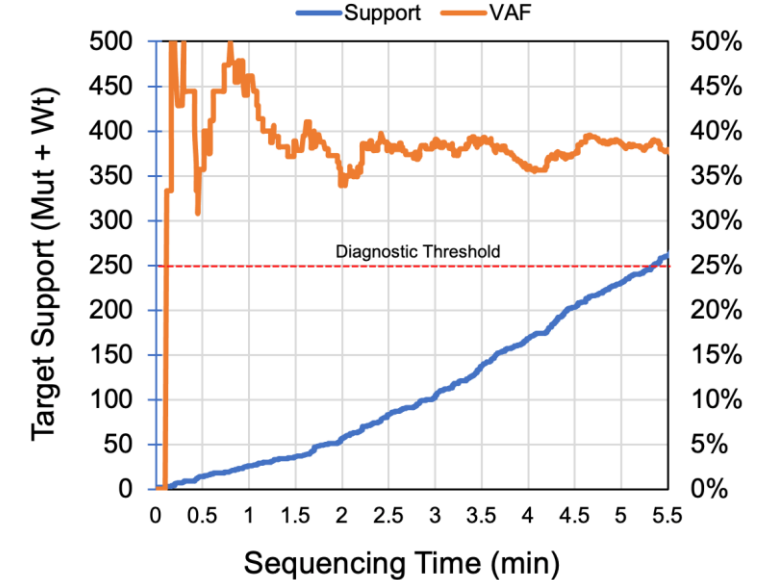
$$F_{target} = \frac{2^{N_{cycle}}}{2^{N_{cycle}} + N_{background}}$$

$$T_{seq} = N_{depth} \times \frac{1}{N_{pores} \times R_{sample} \times F_{target}}$$

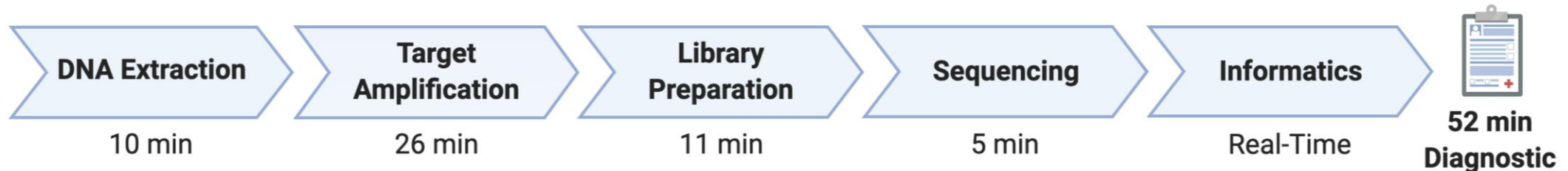
2) Augment model with experimentally derived parameters



3) Run diagnostic with final optimal parameters



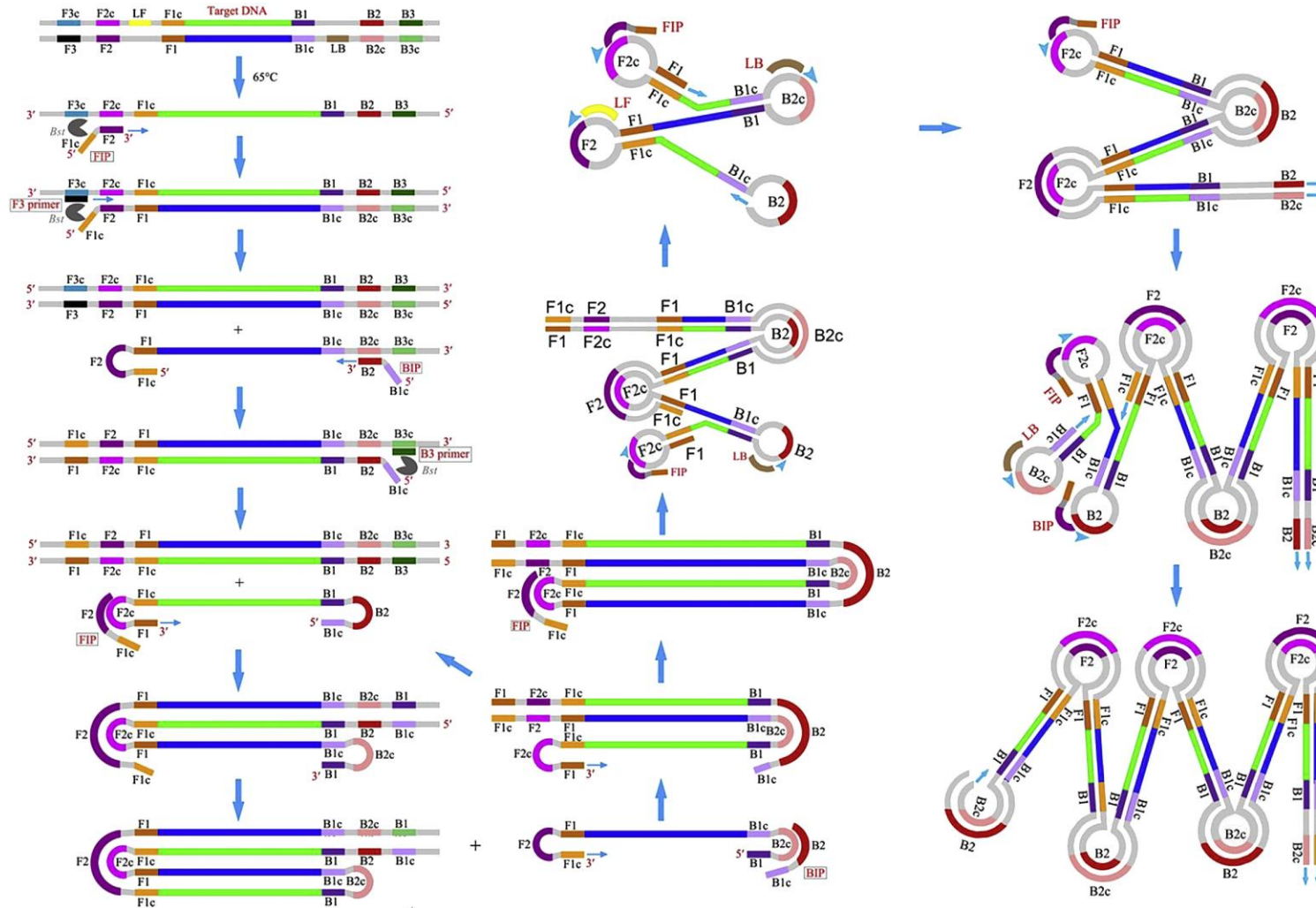
Co-optimization allowed for a world-first demonstration of a sub-1 hour sequencing-based diagnostic



but target amplification is still a large bottleneck...

Loop-Mediated Isothermal Amplification (LAMP) Technology

N=1 target



<https://doi.org/10.1016/j.trac.2019.01.015>

Benefits

- LAMP amplifies targets much more rapidly than PCR (14min vs 26min)
- LAMP generates concatemeric reads that contain redundant, and complementary information

Downsides

- Difficult to analyze and reason about complex product
- No LAMP specific bioinformatics tools

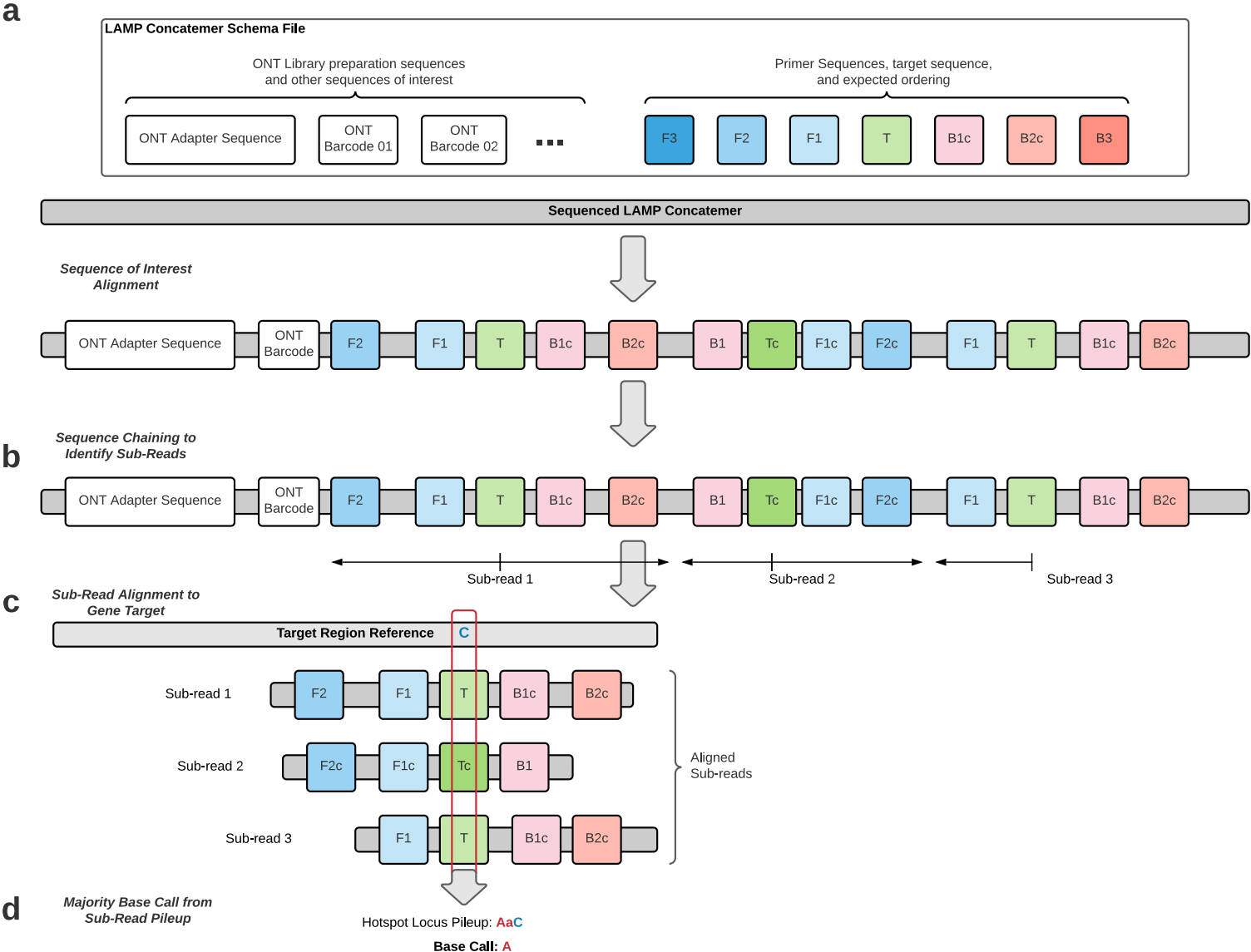
We leverage LAMP's rapid amplification and redundant information to further reduce diagnostic time

LAMPPrey: a new bioinformatics tool to analyze and “polish” LAMP concatemer product

LAMPPrey identifies concatemer “sub-reads” in noisy amplicons

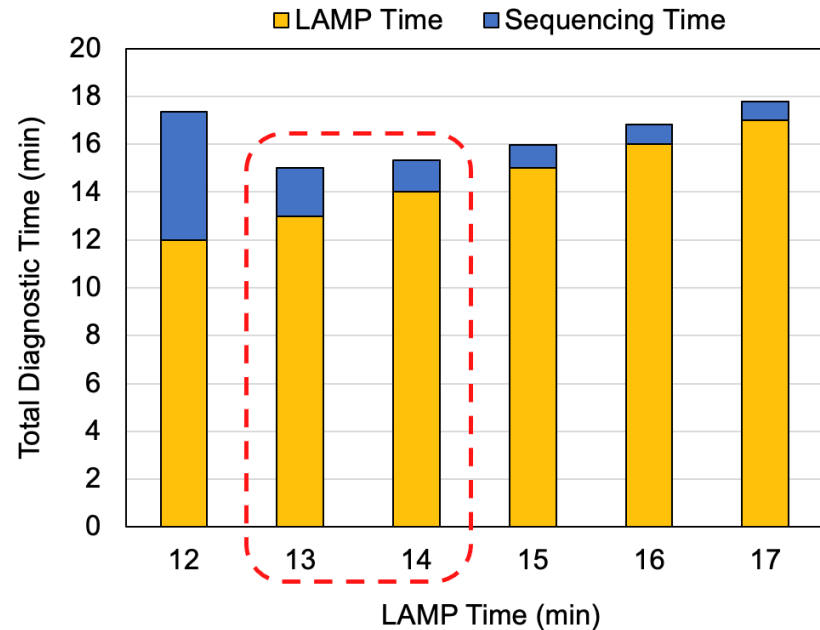
LAMPPrey is able to recover about 50% more information than traditional informatics tools

Information from each sub-read can be combined to form a more confident base call (polishing) resulting in a more rapid and accurate diagnostic

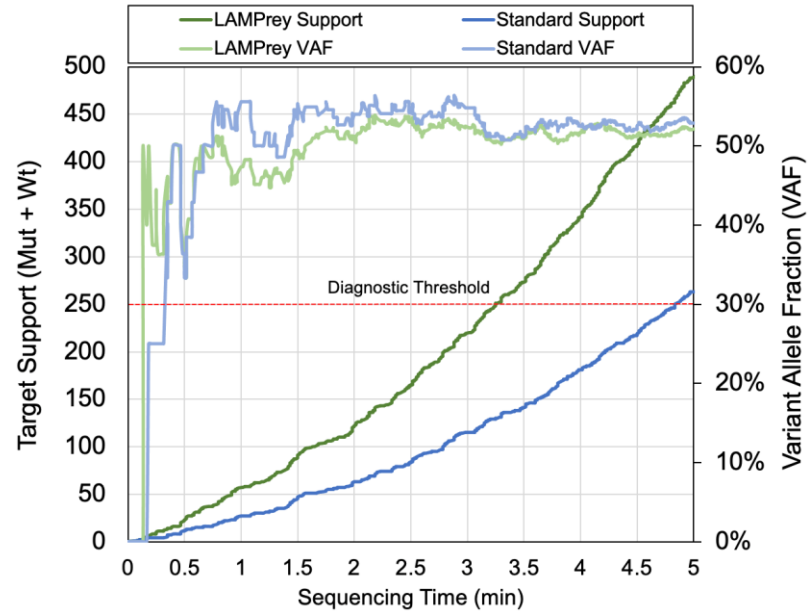


LAMPrey + Threshold Sequencing = <30min Sequencing-based Diagnostic

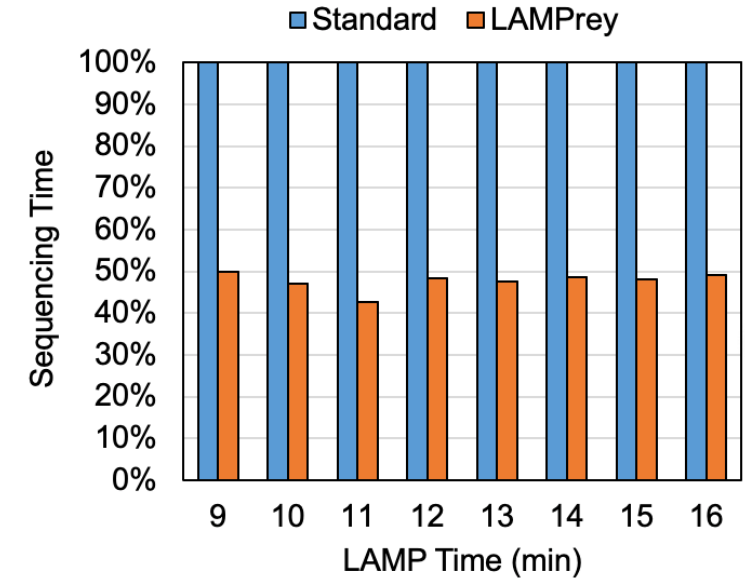
Experimentally informed
LAMP diagnostic model



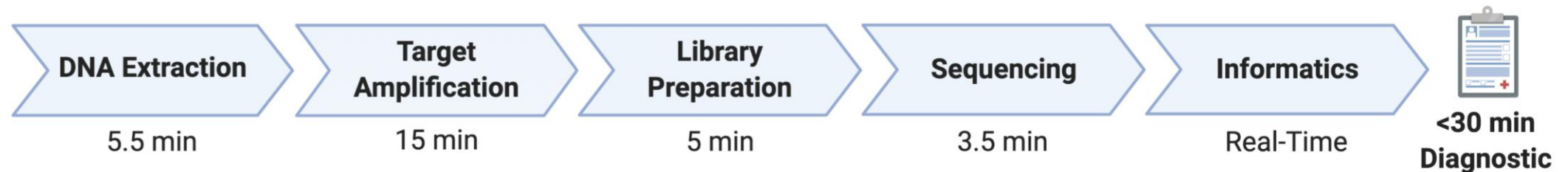
Final LAMP diagnostic result



LAMPrey benefit

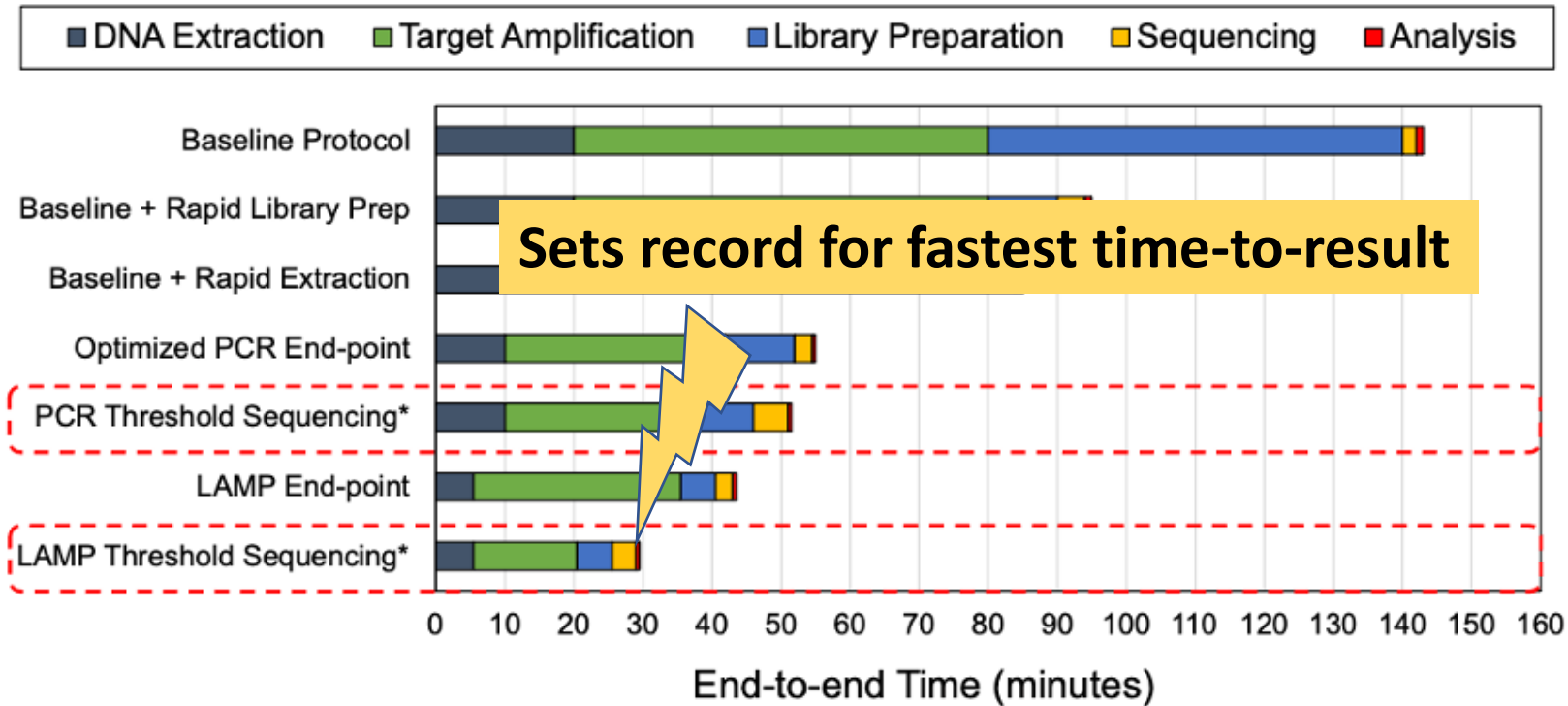


LAMPrey and other optimizations allowed for a world-first demonstration of a sub-30 minute sequencing-based diagnostic

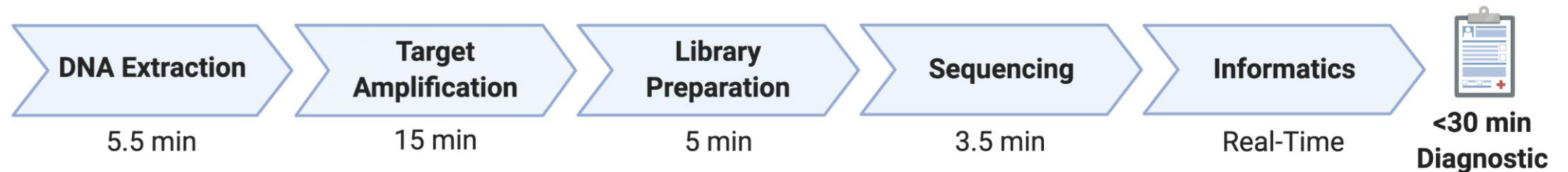


Open source: <https://www.github.com/jackwadden/lamprey>

LAMPrey + Threshold Sequencing = <30min Sequencing-based Diagnostic



LAMPrey and other optimizations allowed for a world-first demonstration of a sub-30 minute sequencing-based diagnostic





vcfDist – A new tool to analyze variant callers

*Dunn and Narayanasamy
Nature Communications 2023*

vcfdist: accurately benchmarking phased small variant calls in human genomes

Tim Dunn, Satish Narayanasamy; 2023

Which variant caller is better?

Many ways to represent adjacent variants (SNP <-> INDEL)

Challenge:

Same sequencing output (query) with different representation (w.r.t reference) yields different conclusions.

Solution:

Representation independent alignment for comparing variant caller accuracy

Reference	ACCCTTTTTTG		Query	ACCTTTG		Truth	ACCCTTTG	
Query VCF Representation 1			Query VCF Representation 2			Truth VCF		
POS	REF	ALT	POS	REF	ALT	POS	REF	ALT
3	CCTTT	C	1	AC	A	4	CTTT	C
			4	CTTT	C			

vcfeval Summary Statistics								
	TP	FP	FN	PP	Precision	Recall	F1	F1 Q-score
Query Repr. 1	0	1	1	0	0.00	0.00	0.00	0.00
Query Repr. 2	1	1	0	0	0.50	1.00	0.67	4.77

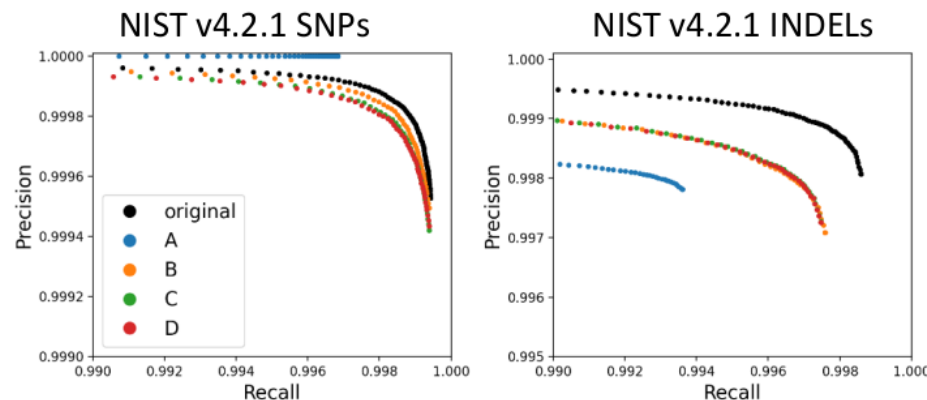
vcfdist Summary Statistics								
	TP	FP	FN	PP	Precision	Recall	F1	F1 Q-score
Query Repr. 1	0	0	0	1	0.67	0.67	0.67	4.77
Query Repr. 2	1	1	0	0	0.50	1.00	0.67	4.77



Result

 <https://github.com/timd1/vcfdist>

Before



Same sequencing output, but different representations (A, B, C, D)

Before:

accuracy is artificially dependent on variant representation

After:

vcfDist yields same accuracy for same output, Independent of variant representation

Users

NIST
National Institute of
Standards and Technology

PacBio





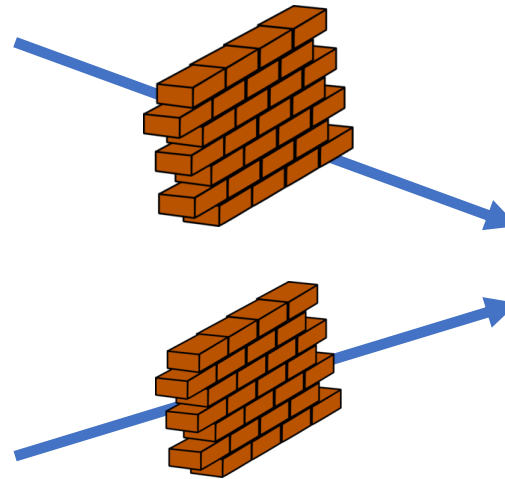
Privacy- preserving collaborative genomics



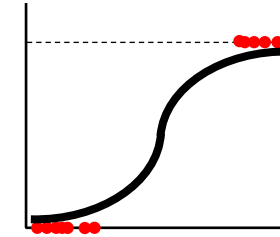
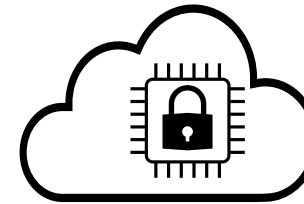
~70,000 patients



~10,000 patients



Privacy wall
(e.g., GDPR)



Powerful analysis on
aggregated large data sets

**Privacy-preserving
collaborative analysis
in cloud**

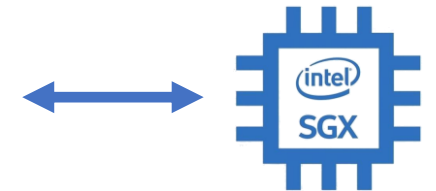
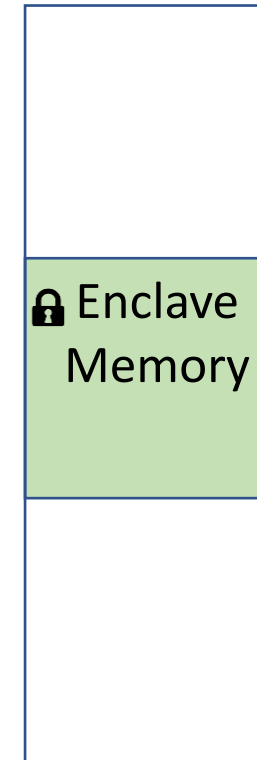
Small Enclave Memory

GWAS Dataset

	Patient A	Patient B	...
Var. 1			
Var. 2			
Var. 3			
...			
Diabetes	Yes	No	No
Smoker	No	No	Yes
...			

$\cong 100\text{s GB} \gg 100\text{-}200\text{ MB}$

Main Memory
(few GB)



Optimizations

Streaming

Batching

Data parallelism

Compression

Artifact

Privacy-preserving *Hail*

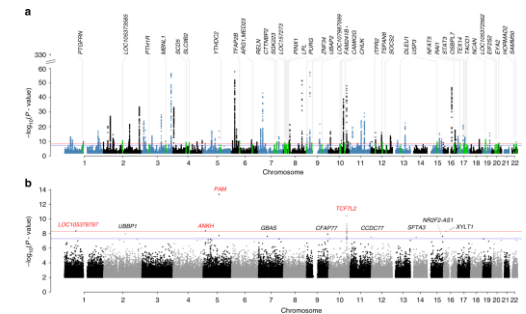
Supports linear and logistic regression

Open-source end-to-end GWAS system

Scales to >1000 cores on Azure

Efficient: < 1 min for a regression analysis

>4 million variants, 1 million patients, 12 cov. (\cong 150 GB)

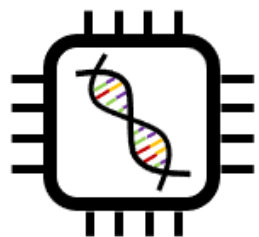


Genomic data (VCF)



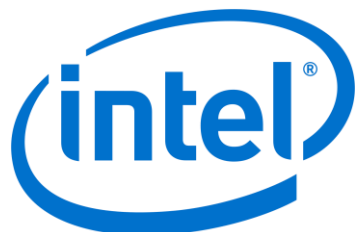
How Can You Kick-Start Genomics Research?





GenomicsBench

[ISPASS 21]



Open-source:

<https://github.com/arun-sub/genomicsbench>



12 computationally intensive kernels drawn from well maintained software tools



Covers the major steps of modern sequence analysis pipelines

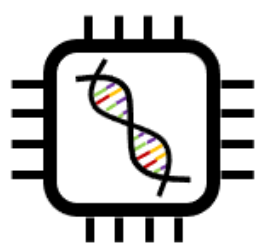


Includes both short and long read analysis algorithms

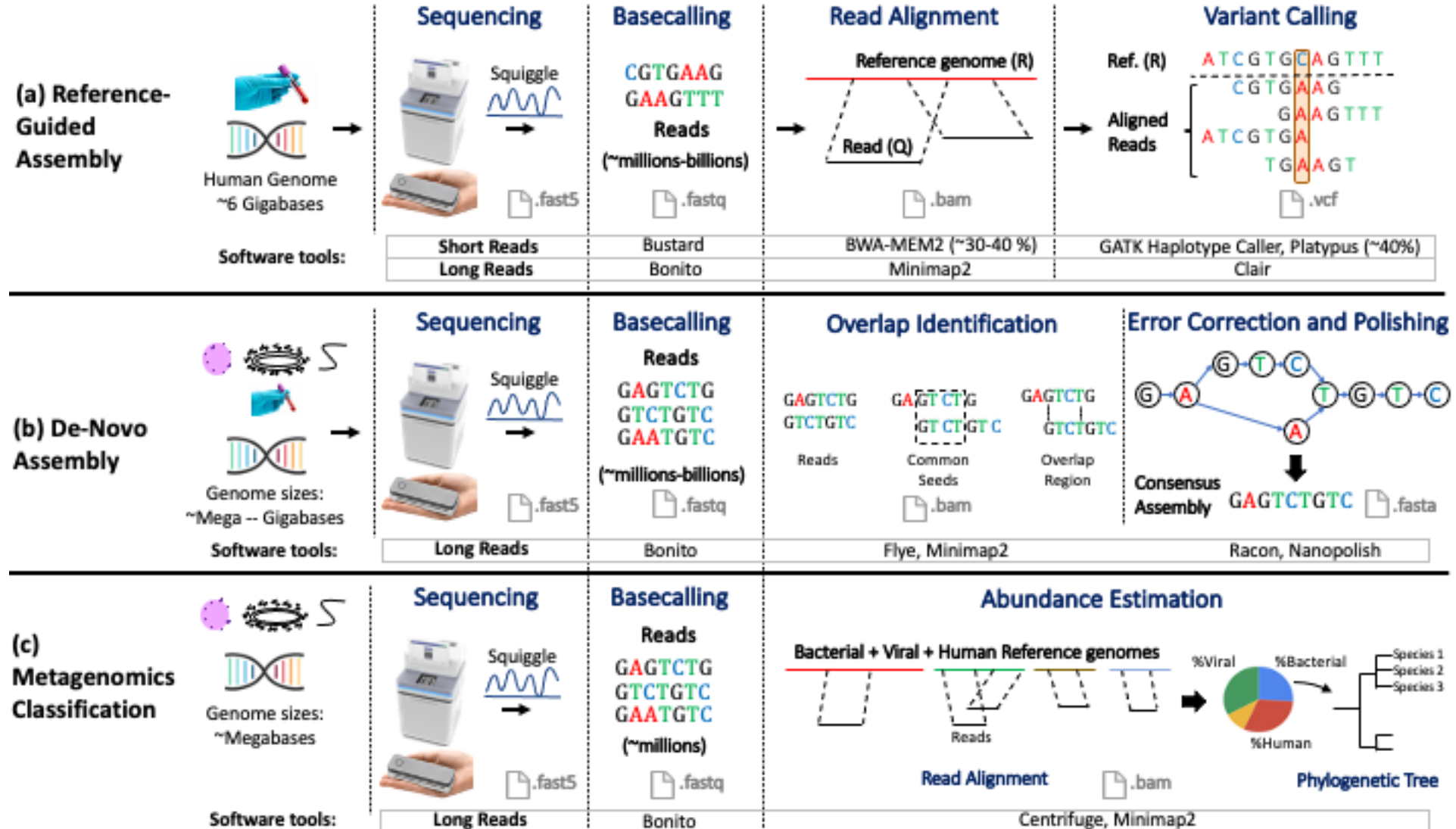


Small/large input datasets

How Benchmarks Leads to Ideas



GenomicsBench Pipelines



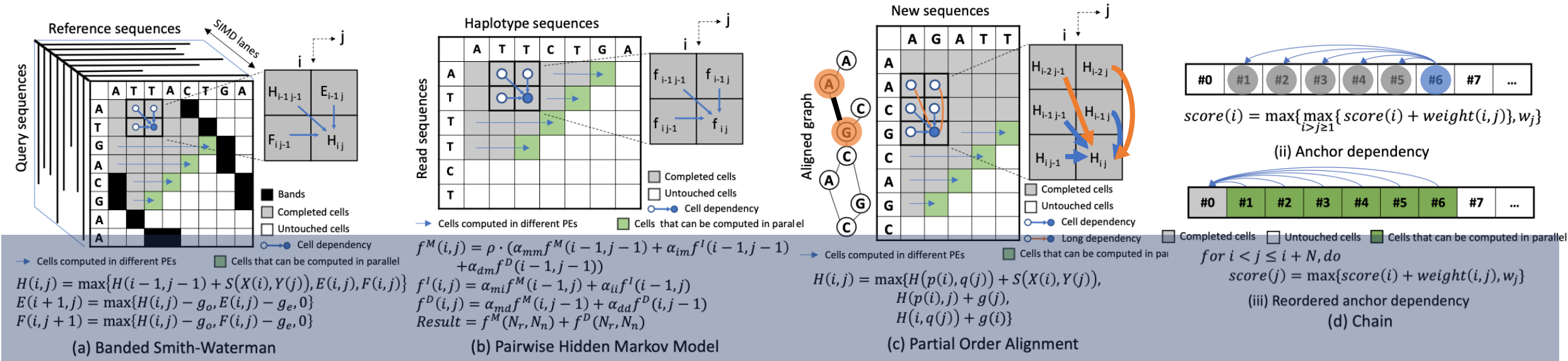
Dynamic Programming Kernels in GenomicsBench

Benchmark	Input Datatype	Applications	Chosen Tool	% Time Spent in Tool (single-thread)	Parallelism Motif
fmi	Short reads	Read Alignment Metagenomics Classification	BWA-MEM2	38%	Tree Traversal
bsw	Short reads	Read Alignment De-Novo Assembly	BWA-MEM2	31%	Dynamic Programming
dbg	Short reads	Variant Calling De-Novo Assembly	Platypus	65%	Graph Construction Hash Table
phmm	Short reads	Variant Calling Error Correction	GATK Haplotype Caller	70%	Dynamic Programming
chain	Long reads	De-Novo Assembly Read Alignment	Minimap2	47.4 %	Dynamic Programming (1D)
spoa	Long reads	Error Correction	Racon	75 %	Dynamic Programming Graph Construction
abea	Long reads	Basecalling Variant Calling	Nanopolish	71.4%	Dynamic Programming
grm	NA	Population Genomics	PLINK2	92.8 %	Dense Matrix Multiplication
nn-base	Long reads	Basecalling	Bonito	95 %	FP Matrix Multiplication

Dynamic programming is the fundamental algorithm in genome sequencing analysis and motivates a domain specific accelerator



Genomics DP Kernels



Kernel	Application	Dimension and Size	Dependency	Data Type
bsw	Read Alionment	2D ~120 × 60	Last 2 Wave-fronts	Int 8/16

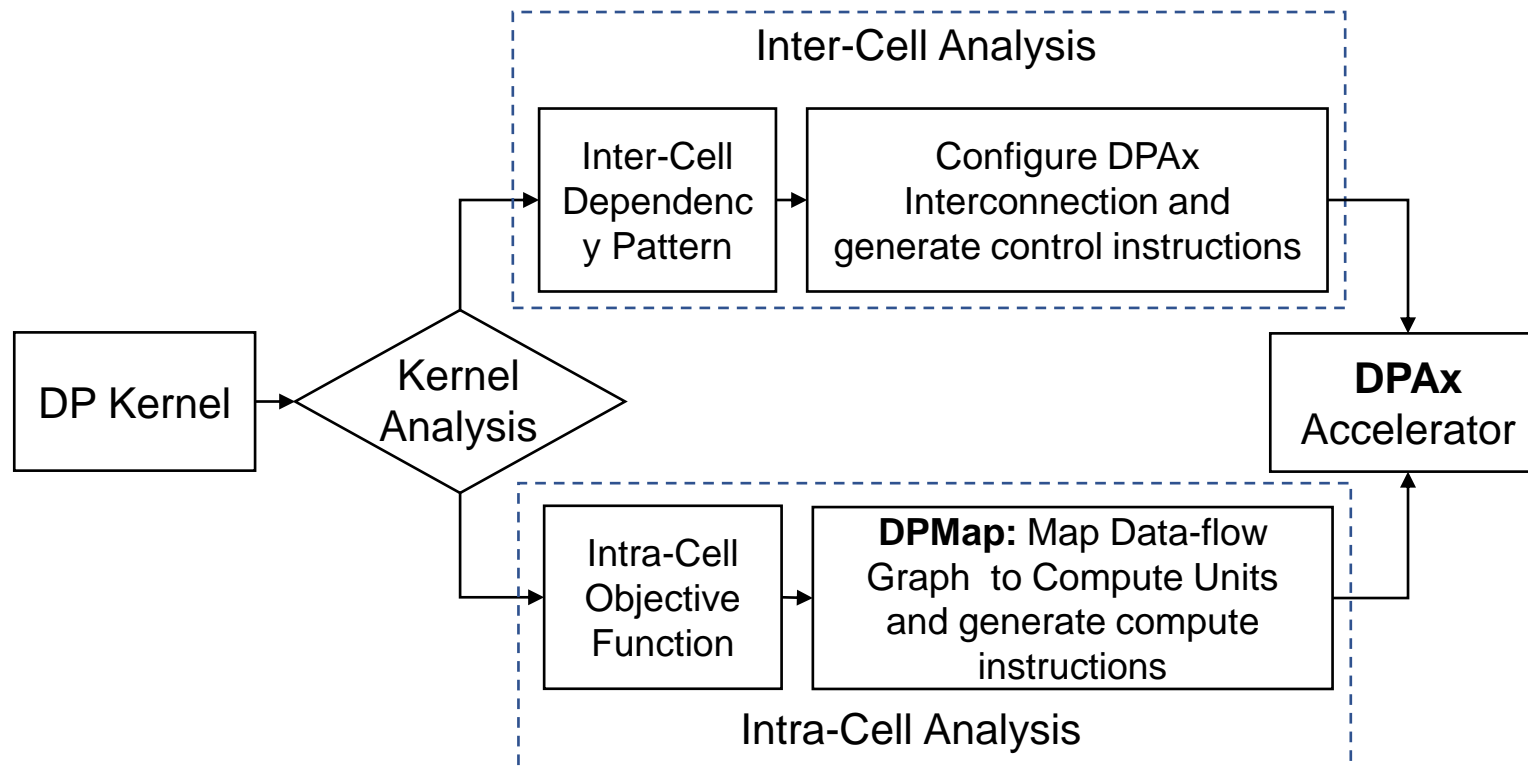
Similarity

Difference

Customization Programmability

GenDP: A Framework of Dynamic Programming Acceleration for Genome Sequencing Analysis

- **DPAx**: programmable dynamic programming (DP) accelerator.
- **DPMaP**: map the objective function of DP algorithm to DPAx accelerator.

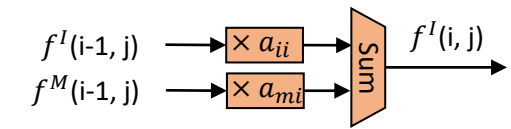
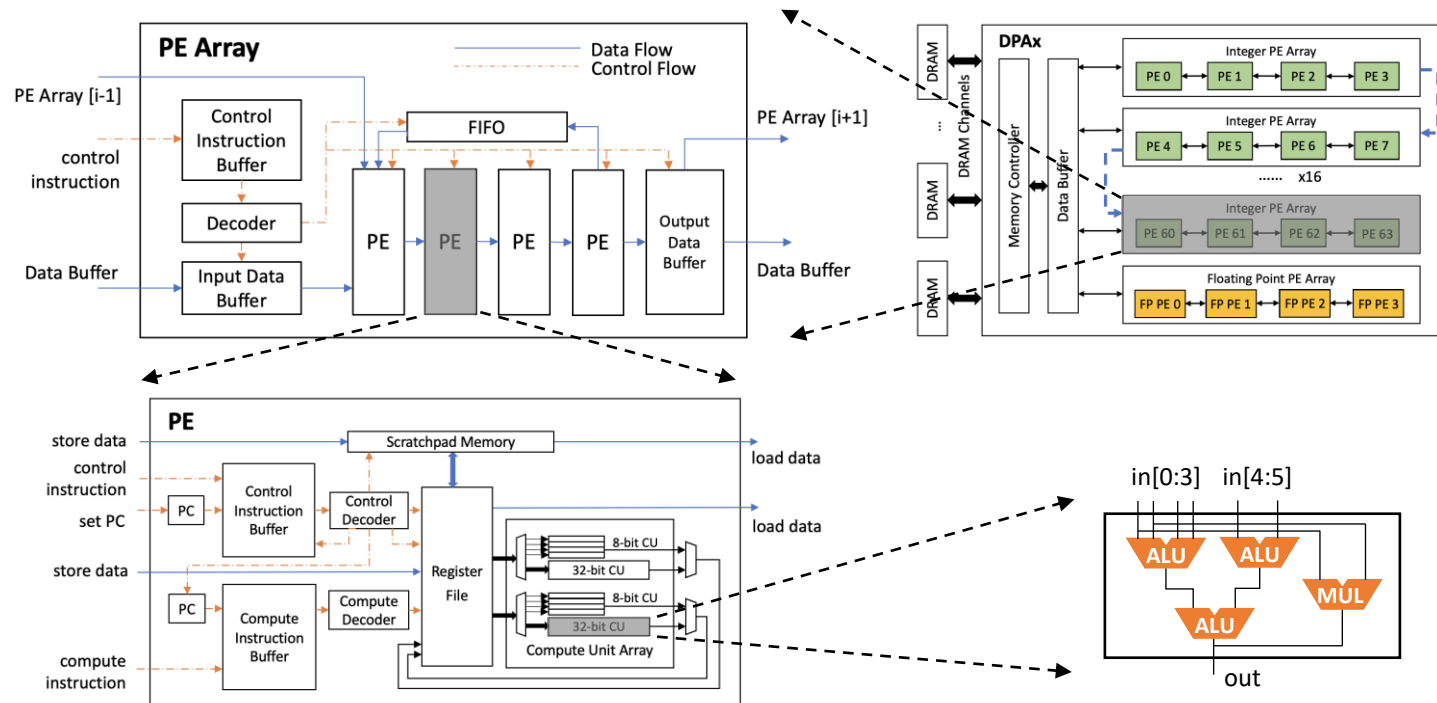


ISCA 2023

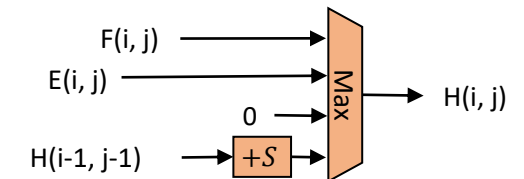
CACM
Research Highlights

Design Choice Take Away

- | | | |
|------------|---|---|
| Similarity | <ul style="list-style-type: none"> ➤ Local dependency ➤ Reduction tree data path | <ul style="list-style-type: none"> ✓ 1-Dimension systolic PE array with FIFO ✓ Compute unit – 2-level reduction tree |
| Difference | <ul style="list-style-type: none"> ➤ Precision requirement ➤ Dependency patterns ➤ Long dependency ➤ Objective func. and datapath | <ul style="list-style-type: none"> ✓ 16 Integer PE array (SIMD compute unit) and 1 FP PE array ✓ PE arrays could execute separately or combined ✓ Software managed scratchpad memory ✓ Custom ISA for control and computation |



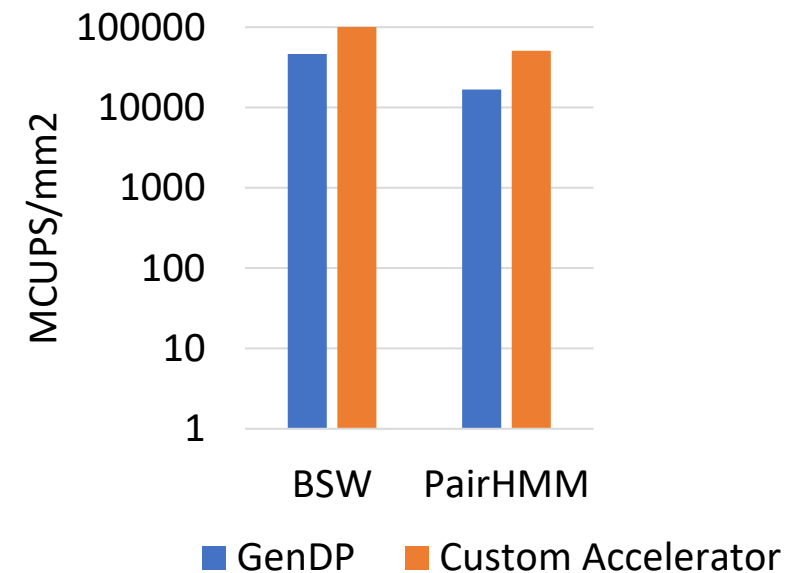
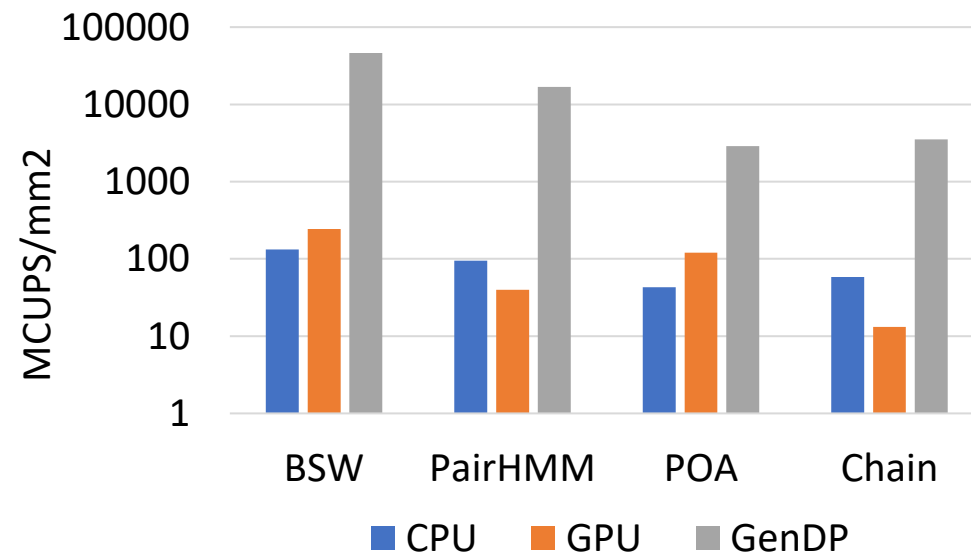
(a) Reduction Data-Flow in PairHMM

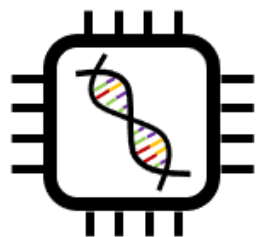


(b) Reduction Data-Flow in BSW

GenDP Performance

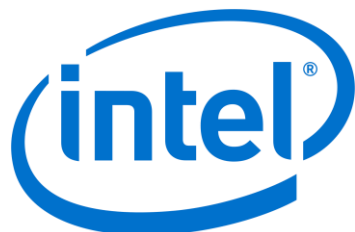
- Metrics: Throughput/Area – Million Cell Updates per Second/ mm^2 (MCUPS/ mm^2)
- GenDP achieves 157.8 \times throughput/ mm^2 over GPU
- GenDP has 2.8 \times slowdown when compared to custom accelerators
- Generality on DP algorithms in other domains
 - Dynamic time warping – speech recognition
 - Bellman-Ford – Robot motion planning





GenomicsBench

[ISPASS 21]



Open-source:

<https://github.com/arun-sub/genomicsbench>



12 computationally intensive kernels drawn from well maintained software tools



Covers the major steps of modern sequence analysis pipelines



Includes both short and long read analysis algorithms



Small/large input datasets

Acknowledgements



UM Medicine



Oxford Nanopore Technologies community

Kahn Foundation



National Science Foundation (NSF)