Systems for Precision Health

Reetu Das

Associate Professor, EECS Department

University of Michigan

KEEP CALM AND GO BLUE

What is Precision Health?

> *an emerging approach for disease treatment and prevention that takes into account individual variability in **genes**, **environment**, and **lifestyle** for each person*

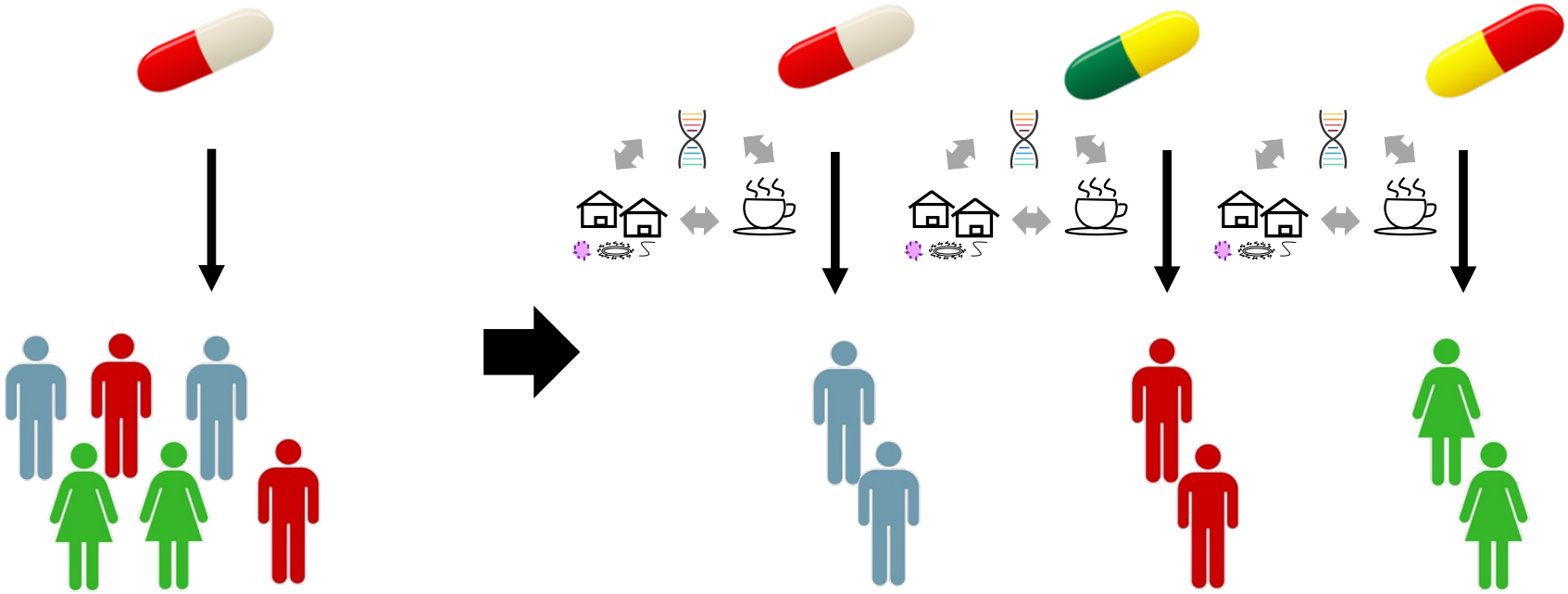
"Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?"

- President Obama, January 30, 2015

All of Us
THE FUTURE OF HEALTH BEGINS WITH YOU

NIH National Institutes of Health

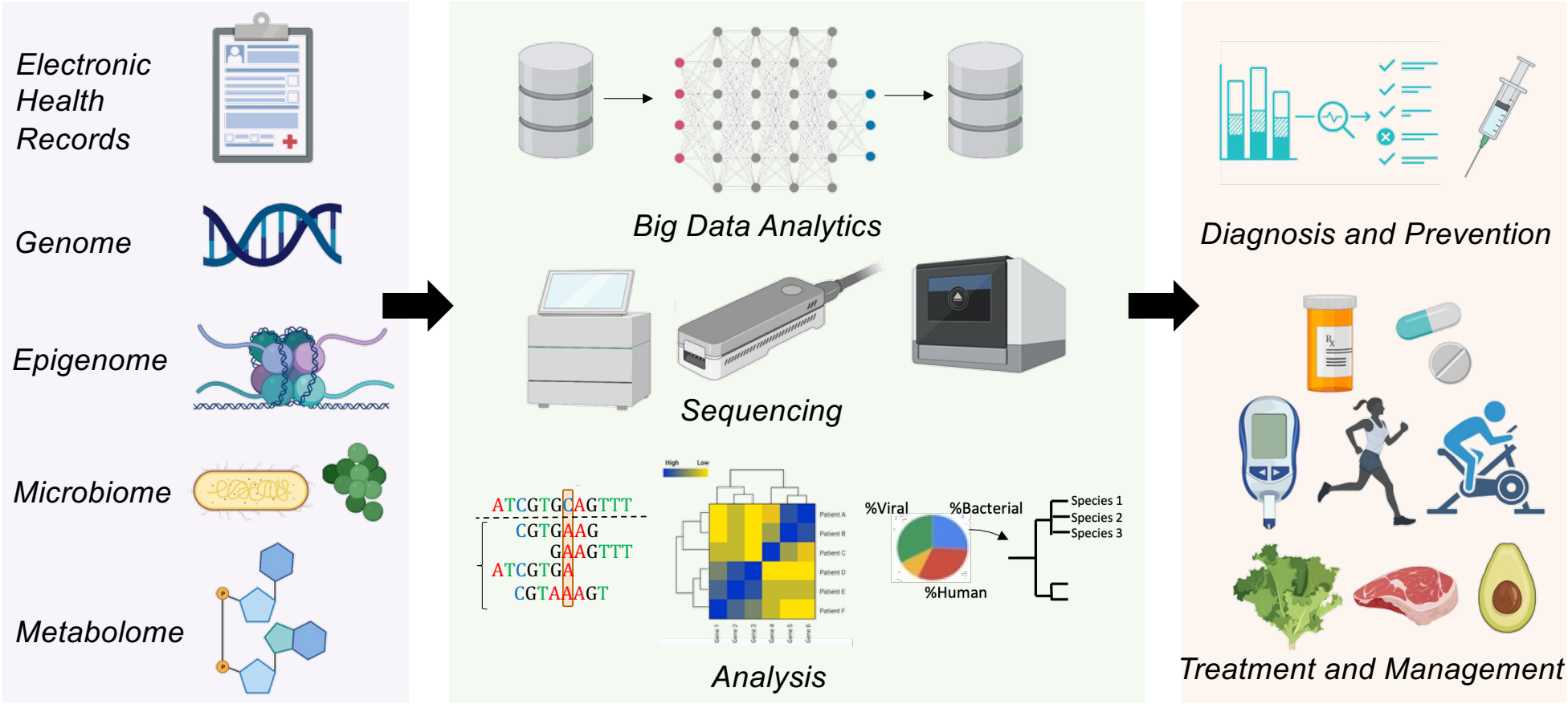
What is Precision Health?



One Size Fits All X

Precision Medicine ✓

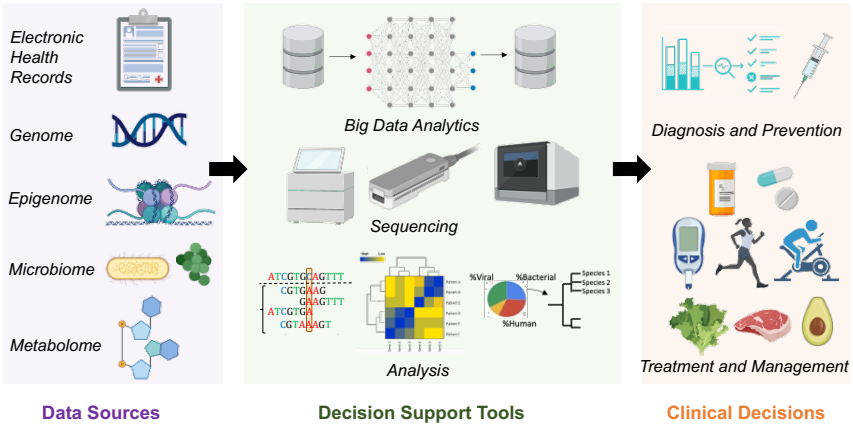
Precision Health Platform



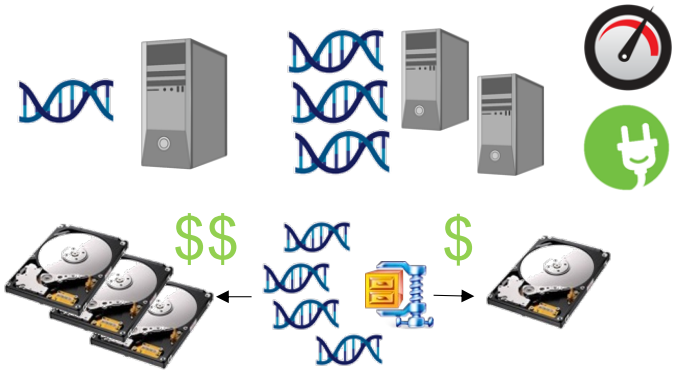
Data Sources | **Decision Support Tools** | **Clinical Decisions**

Credits: Created from BioRender.com

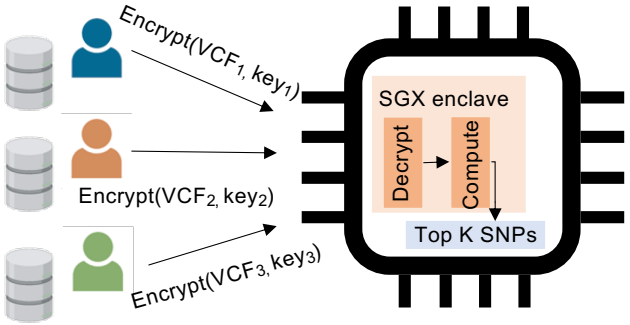
System Design Considerations



Efficiency



Security and Privacy



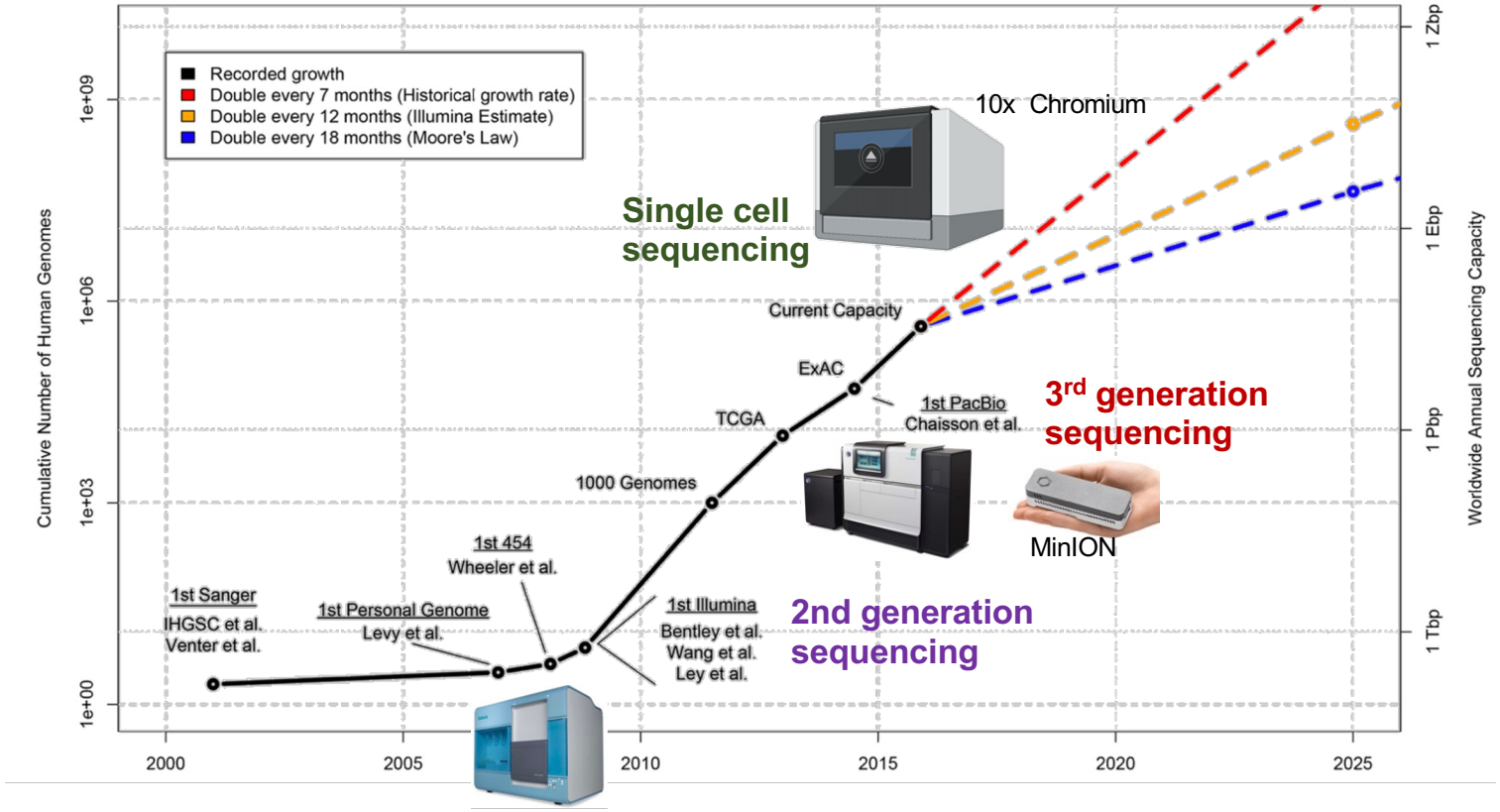
Homomorphic encryption, Intel SGX

Form Factor

Credits: Created from BioRender.com

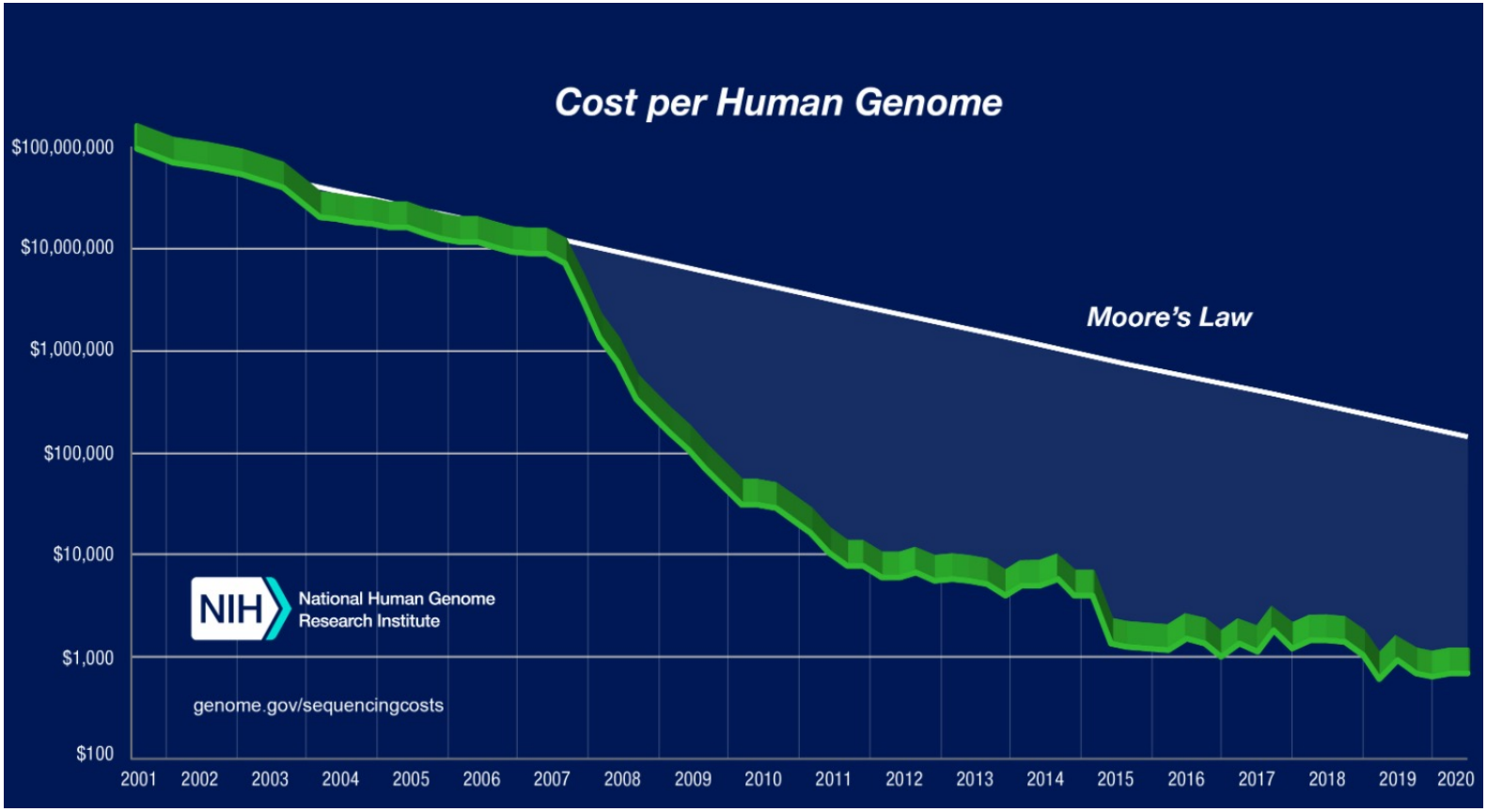

Sequencing is Key Ingredient
of Precision Health

Exponential Growth in Genome Sequencing

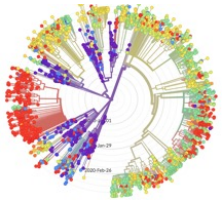


Credits: [Stephens et al . PLOS Bio, 2015] [Illumina] [Oxford Nanopore] [10x Chromium][Biorender.com]

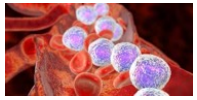
Sequencing Costs have Plummeted



Exploding Sequencing Applications



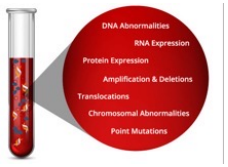
Cancer treatment



Microbiome

In operating room sequencing

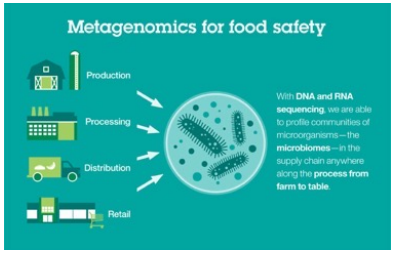
Liquid Biopsy



Consumer Genotyping

23andMe

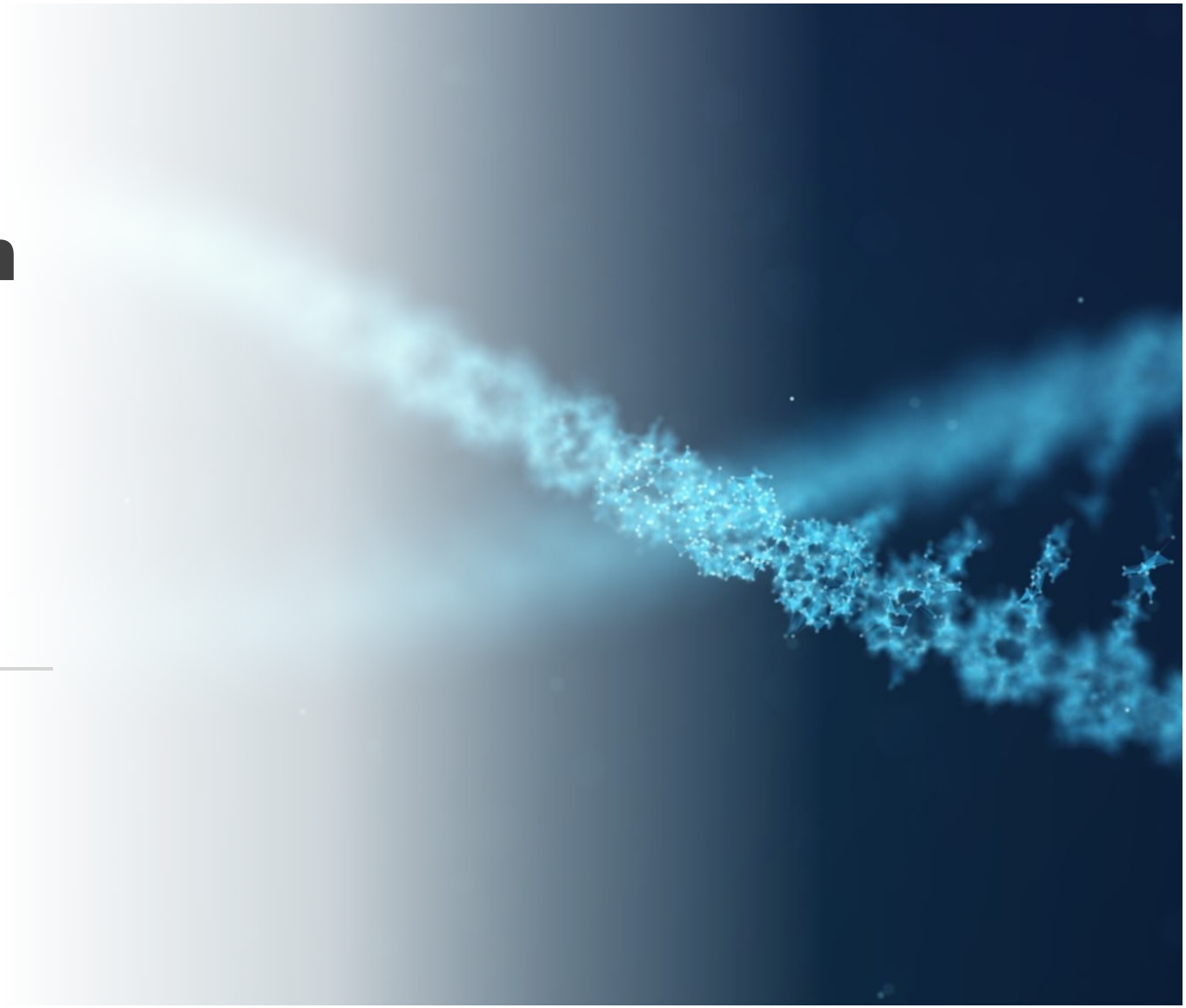
Agricultural sequencing



Food Safety

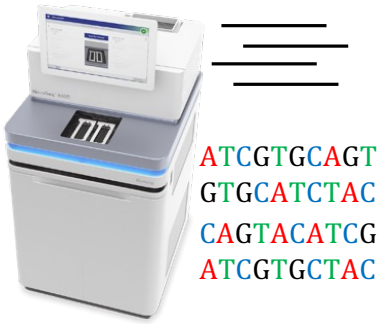
Portable Pathogen detector

Acceleration Study – Whole Genome Sequencing

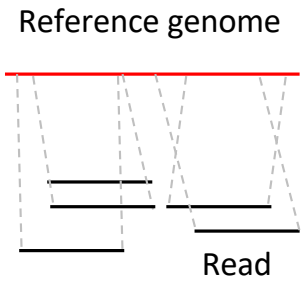


Acceleration Study: Whole Genome Sequencing

Human Genome
6 G bases

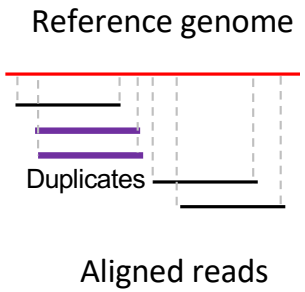


Sequenced reads
(~billions)



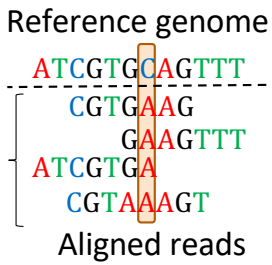
Read Alignment

BWA-MEM



Sorting +
Duplicate Marking

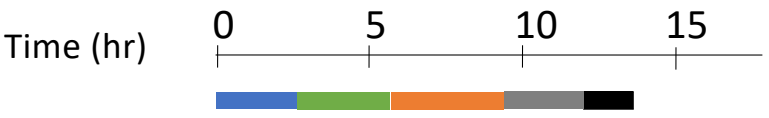
Picard SortSam,
MarkDuplicates



Variant Calling

GATK
Haplotype Caller

Diagnosis

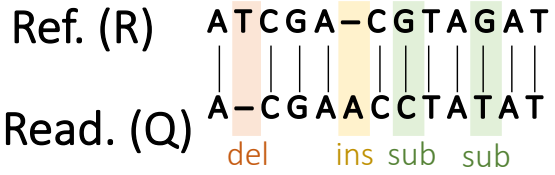


Baseline
m5.8xlarge
32 vCPUs

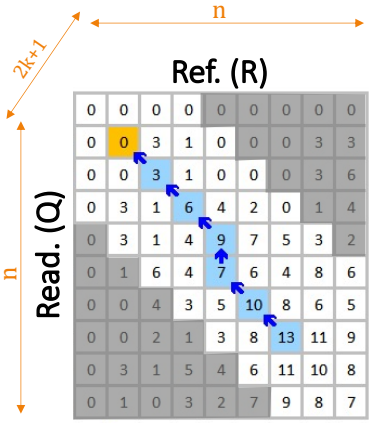
13.9 hr (i.e. 445 CPU hrs)

Seeding
Seed Extension
Sorting / Mark Duplicates
pairHMM
Other

Read Alignment: GenAx

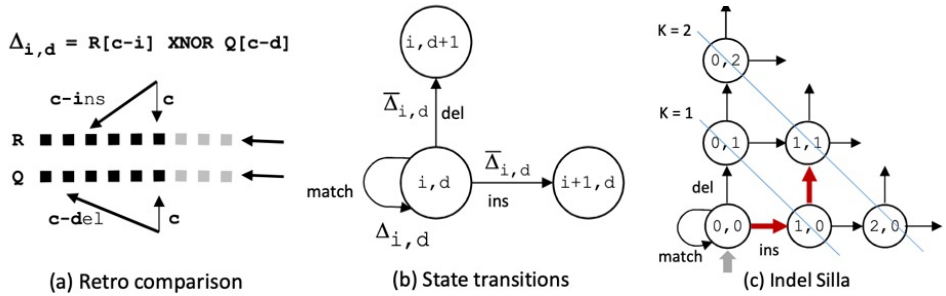


Approximate string matching



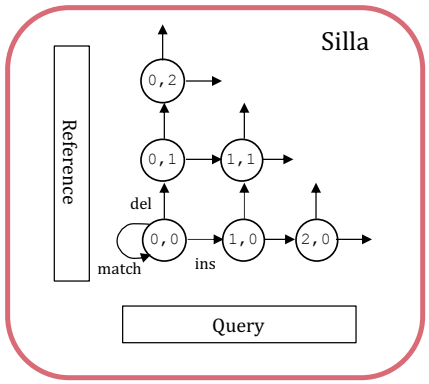
$O(kn)$

Banded Smith-Waterman



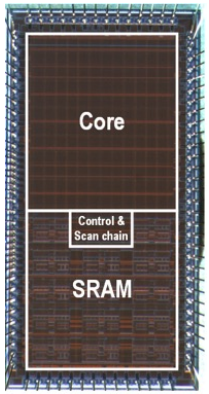
$O(k^2)$

String Independent Local Levenshtein Automata (Silla)



- In-place Traceback
- Affine gap Scoring
- Composability

SillaX hardware accelerator

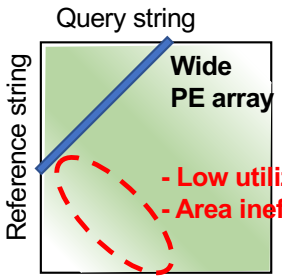


SillaX ASIC fabricated (55nm)

63x faster than 56-thread CPU SeqAn for 100bp reads

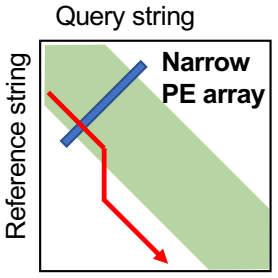
Read Alignment: SeedEx

Full-band implementation



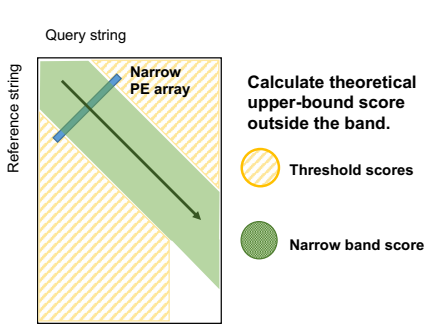
+ Optimality

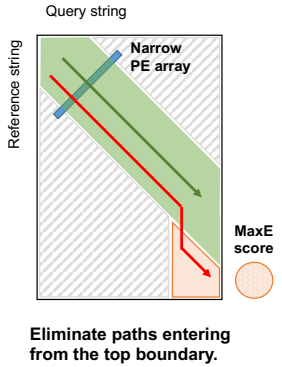
Banded implementation

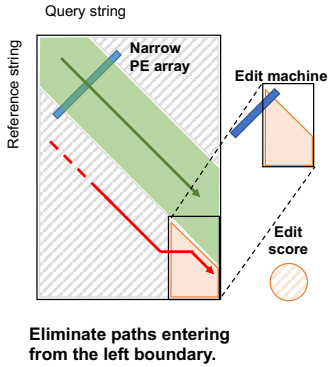


+ Area Efficiency

SeedEx







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

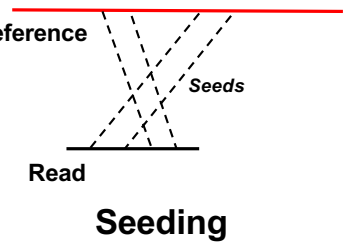
100% equivalent results on AWS cloud FPGA when integrated with BWA-MEM software

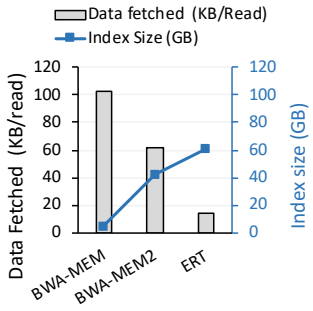
2.3x smaller than banded Smith Waterman core (w = 41 + edit machine)

6x higher throughput over banded Smith-Waterman FPGA (w = 101) for same area

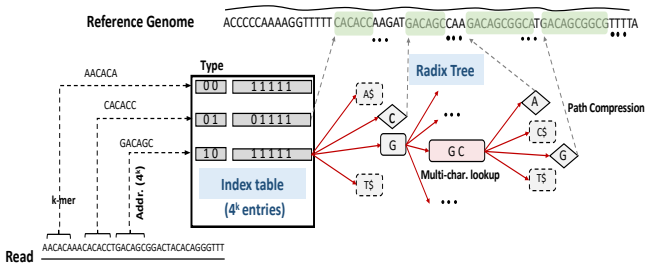
Read Alignment: ERT

Problem

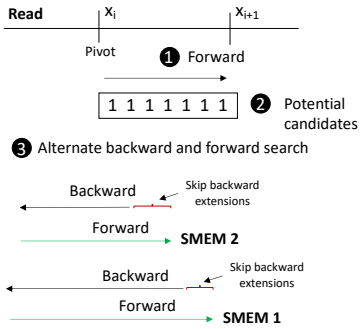




Our Solution

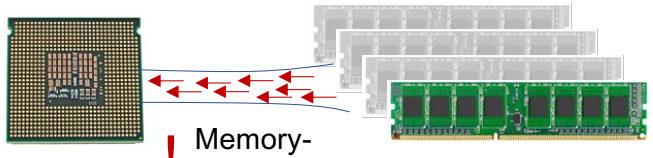


Bandwidth-efficient data structure



Bandwidth-efficient search algorithm

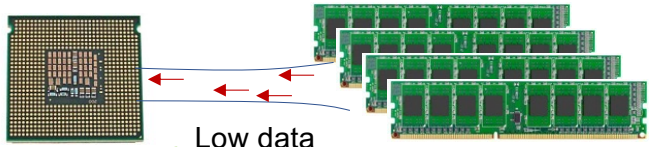
FM-index → widely used seeding data structure



! Memory-bandwidth bottleneck

FM-index
4.2 GB
human

Enumerated Radix Tree (ERT)



✓ Low data requirement

ERT
~60 GB
human

1 Trades-off memory capacity for memory bandwidth to improve seeding performance

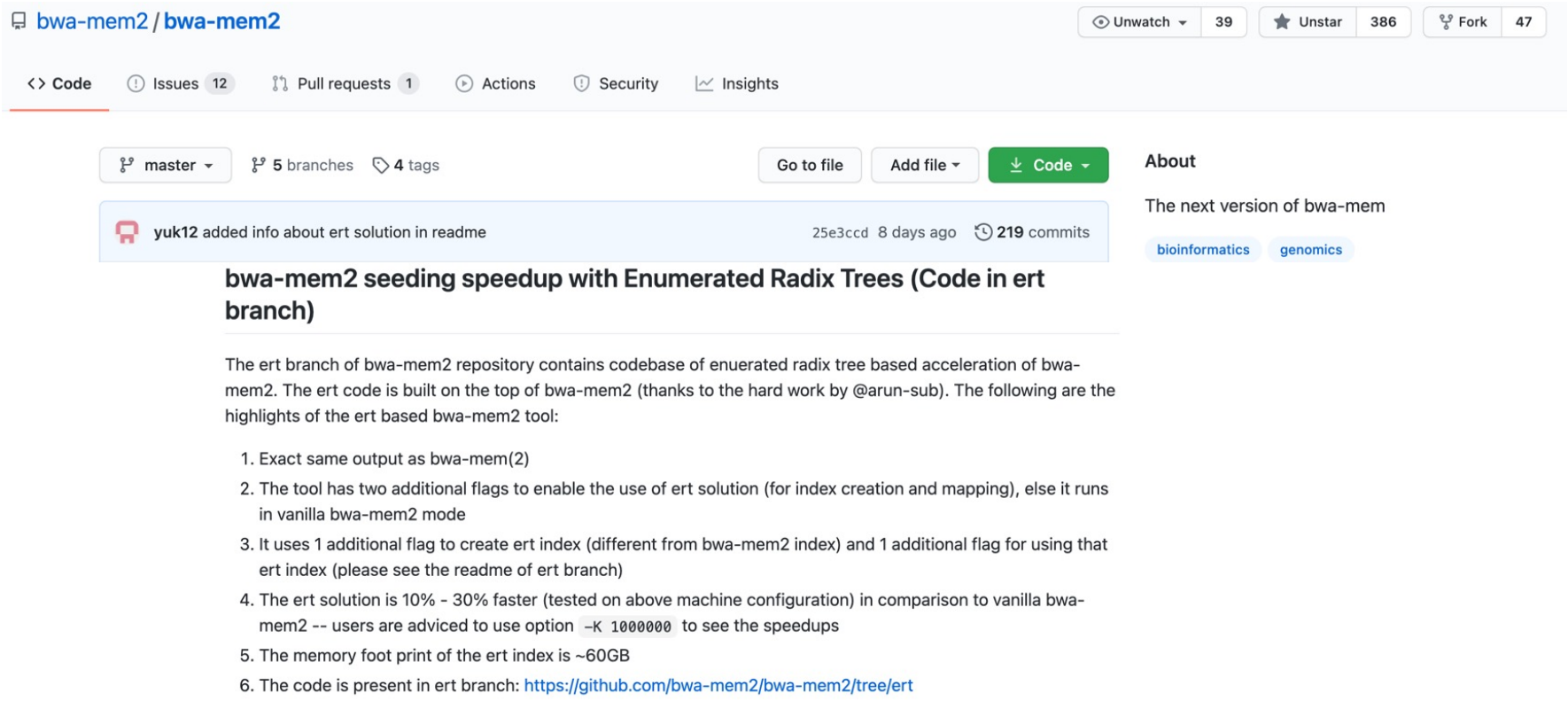
2 Supports multi-character lookup with index table and customized radix tree. Used to implement optimized longest match search algorithm in BWA-MEM

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



bwa-mem2 / bwa-mem2

Unwatch 39 | Unstar 386 | Fork 47

Code | Issues 12 | Pull requests 1 | Actions | Security | Insights

master | 5 branches | 4 tags | Go to file | Add file | Code

yuk12 added info about ert solution in readme — 25e3ccd 8 days ago — 219 commits

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

The ert branch of bwa-mem2 repository contains codebase of enuerated 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-mem(2)
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 adviced to use option `-K 1000000` to see the speedups
5. The memory foot print of the ert index is ~60GB
6. The code is present in ert branch: https://github.com/bwa-mem2/bwa-mem2/tree/ert

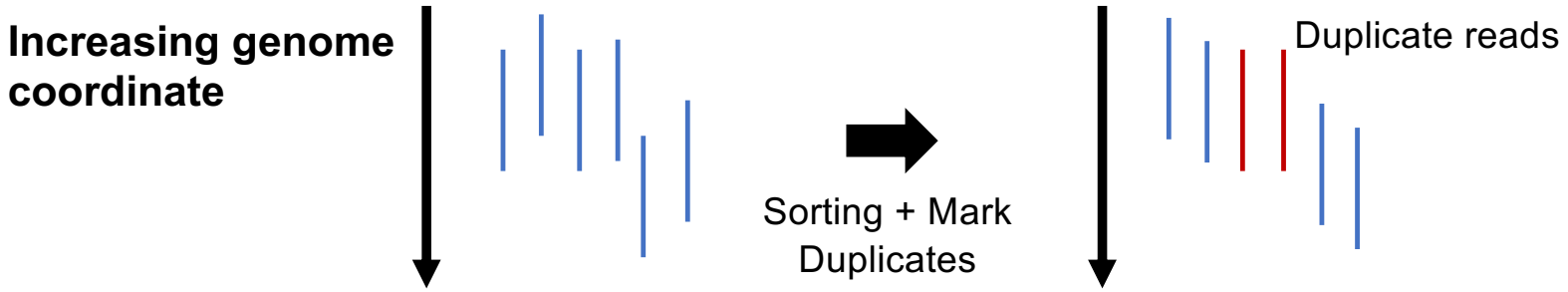
About

The next version of bwa-mem

bioinformatics | genomics

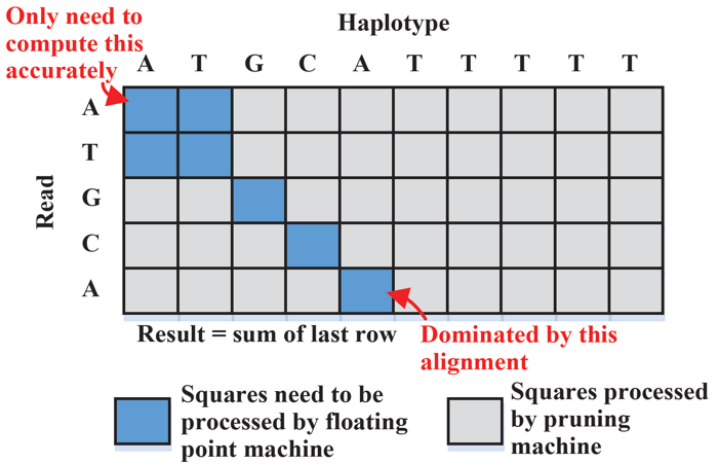
BWA-MEM is the de-facto genomics read alignment tool used by researchers and practitioners worldwide

Sorting/Duplicate Marking Optimizations

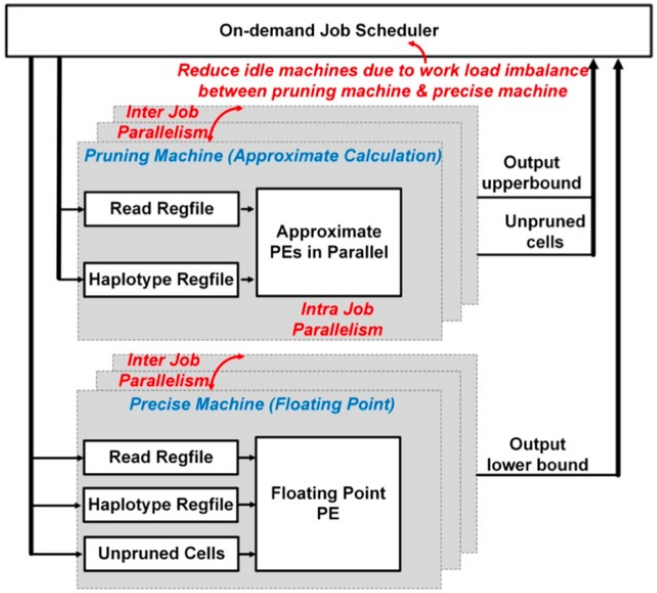


- I/O bandwidth bound. Optimized counting sort based multi-thread CPU implementation
- Same results as Picard SortSam and Picard MarkDuplicates
- Runtime: +3 min for 50x coverage WGS alignments (56 thread CPU)
  Memory: ~75 GB memory

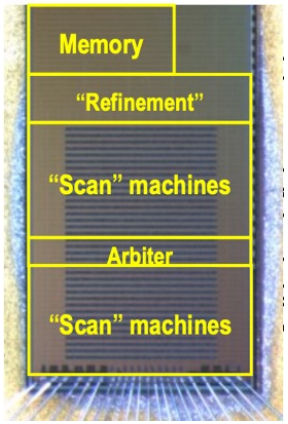
Variant Calling: pairHMM Acceleration



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

Why Accuracy Matters?

Human ~ Human
99.9%

Human ~ Chimpanzee
96%

Human ~ Cat
90%

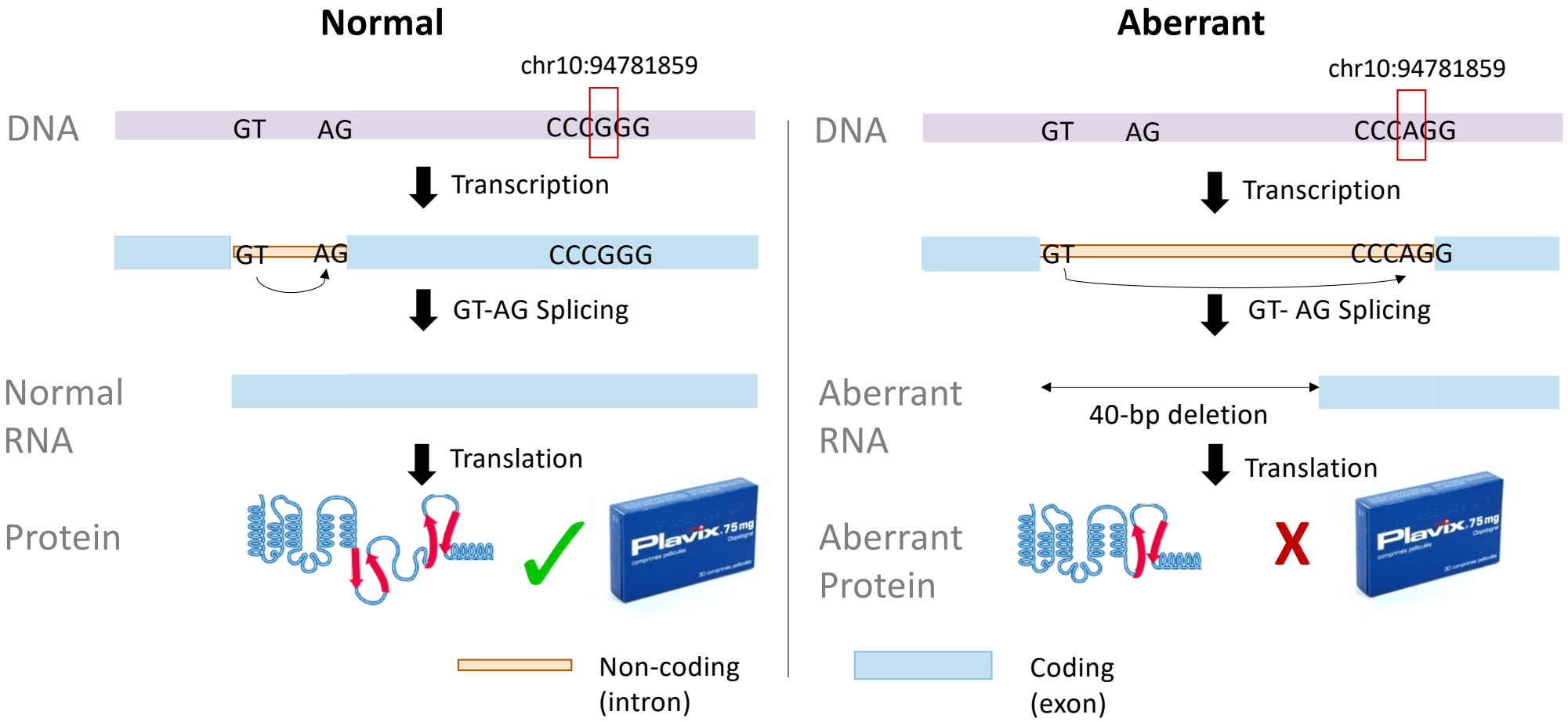
Human ~ Cow
80%

Human ~ Banana
50-60%

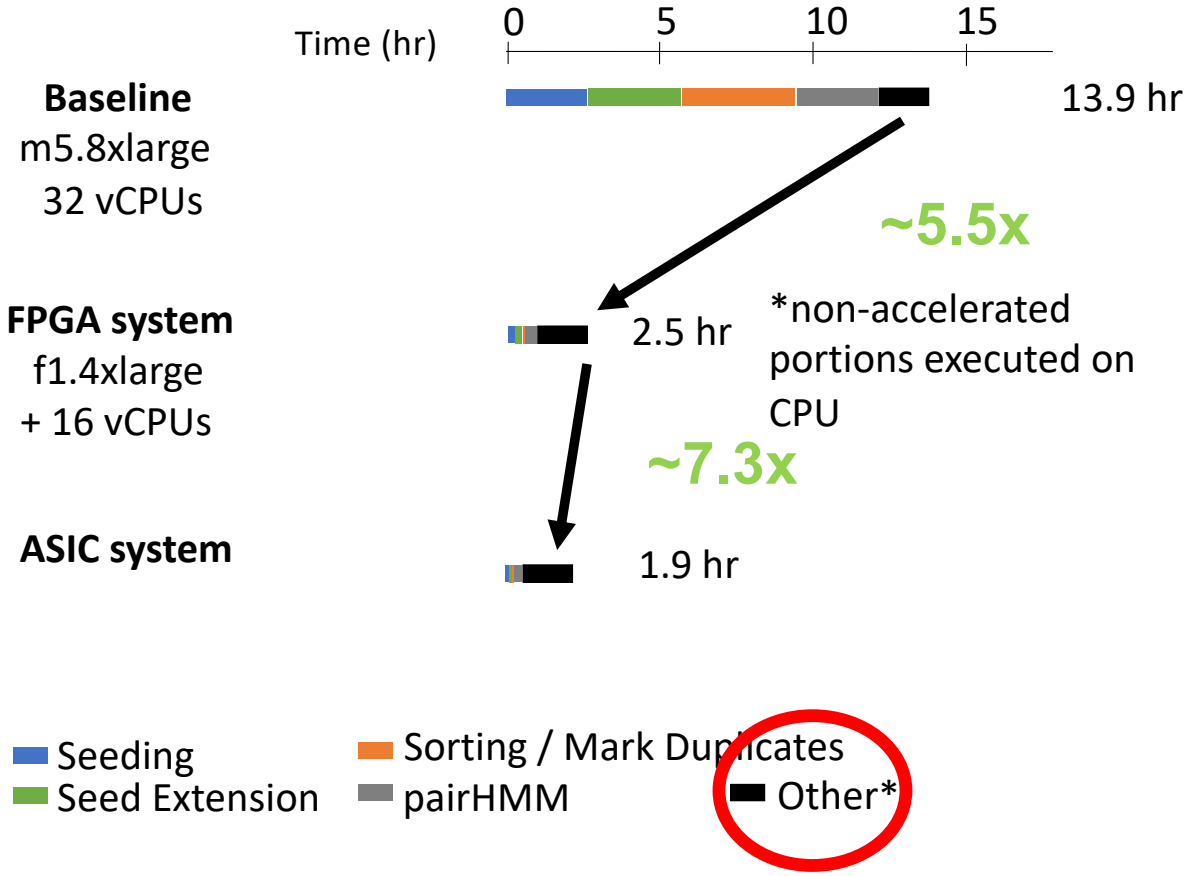
Slide credit: Onur Mutlu,"Accelerating Genome Analysis: A Primer on an Ongoing Journey"

Effect of G->A variant in the CYP2C19 gene

CYP2C19 involved in metabolism of > 10% commonly prescribed drugs

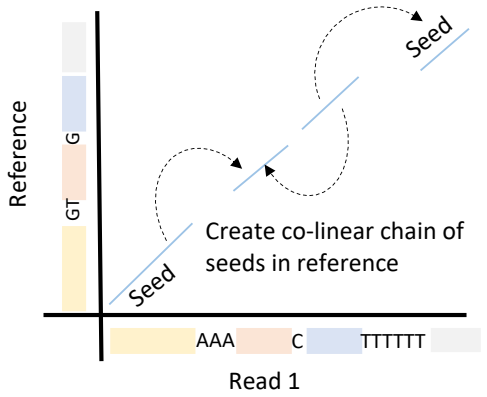


Acceleration Study: Whole Genome Sequencing

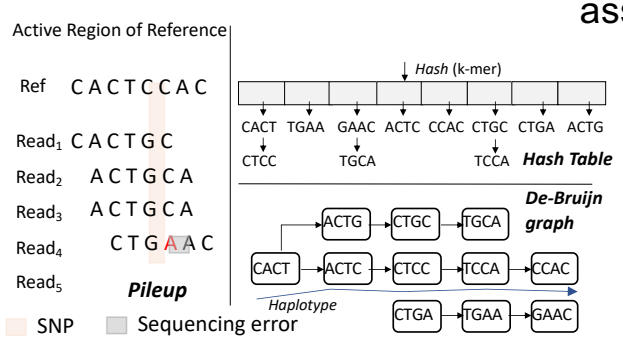


Other Acceleration Candidates

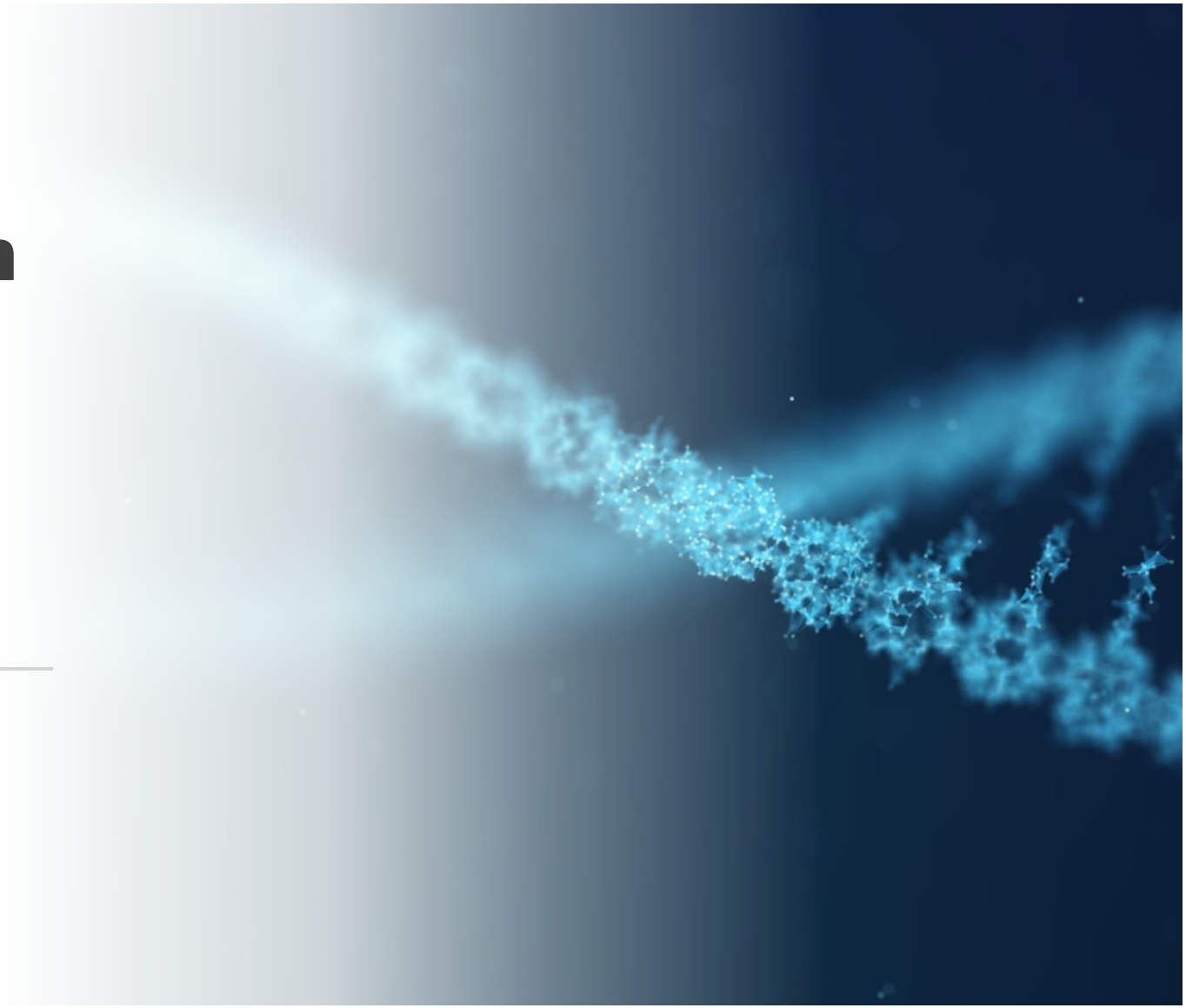
Chaining



De-Bruijn graph assembly



Acceleration Study – Ultra Rapid Cancer Diagnosis



Sequencing Technologies: Evolution

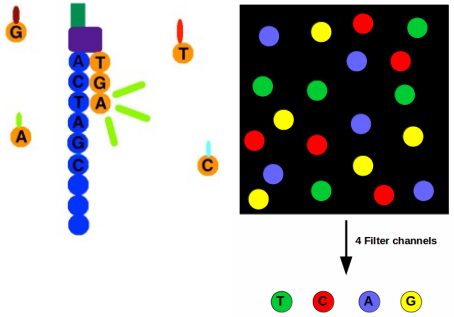
Illumina Sequencing by Synthesis

Illumina Genome Analyzer, 2005

1 Gbases/per day

Illumina NovaSeq 6000, 2021

3 Tbases/per day

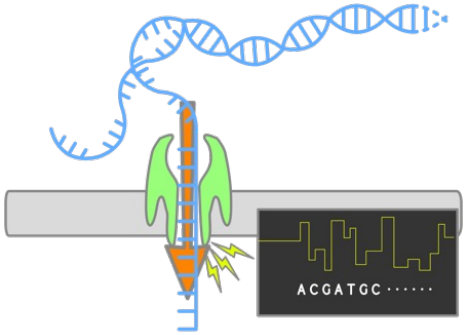


Read length: 100-350bp

Per base inaccuracy: 0.1%

1000x increase in sequencing machine throughput

Nanopore Sequencing



Read length: 1kb-1Mbp

Per base inaccuracy: 1-15%

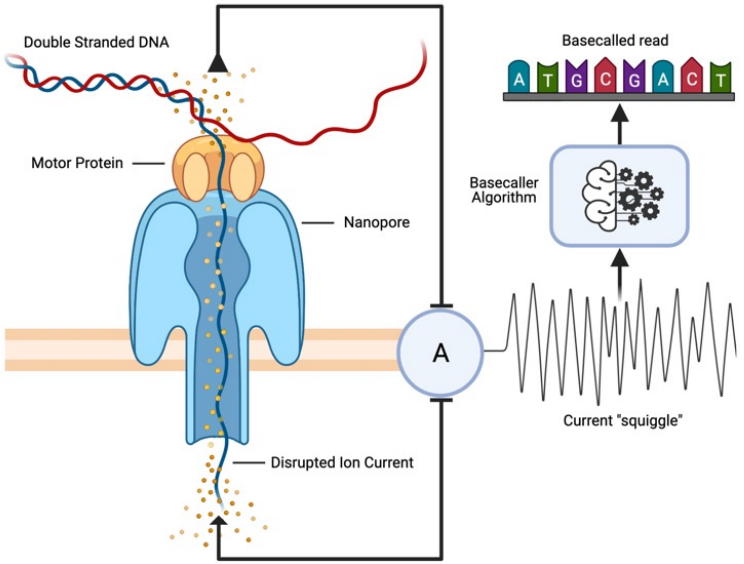
1000x increase in sequencing fragment length

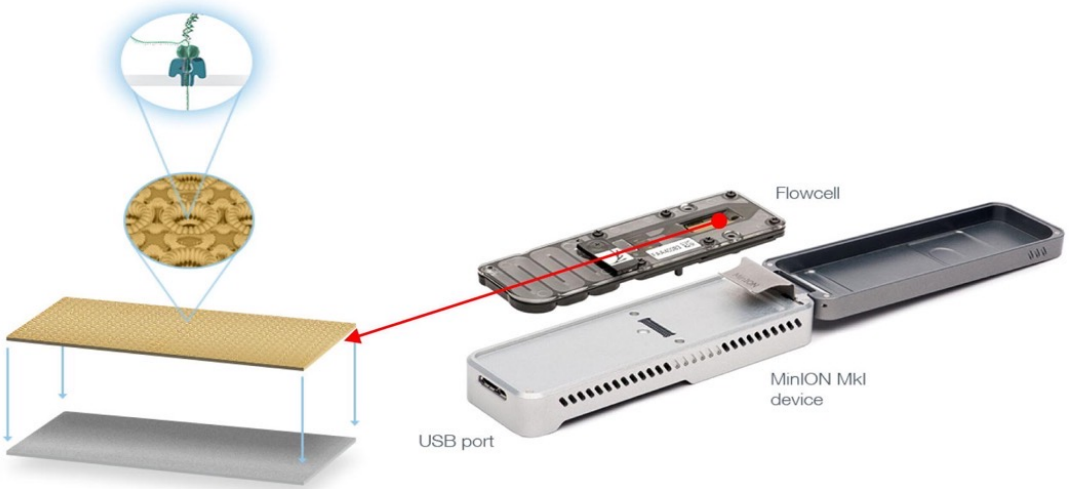
! 10 - 100x increase in sequencing error rate

Credits: Illumina, DataBase Center for Life Science (DBCLS), https://doi.org/10.7875/togopic.2020.01, Wkipedia DMLapato https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina

Nanopore Sequencing is poised to revolutionize molecular diagnostics

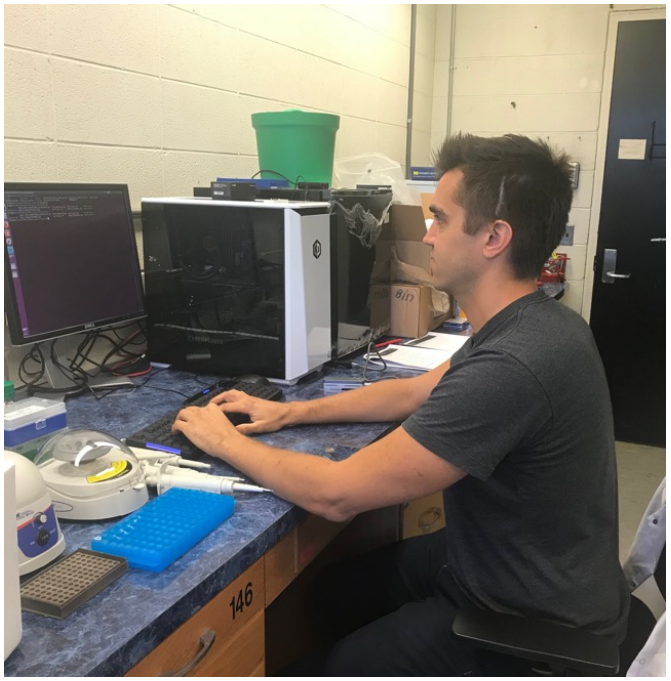
- Nanopore sequencing feeds DNA strands through a biological pore in a membrane
- Current disruptions across the membrane are recorded
- Current disruptions correspond to individual DNA base-pairs (A, T, G, C)

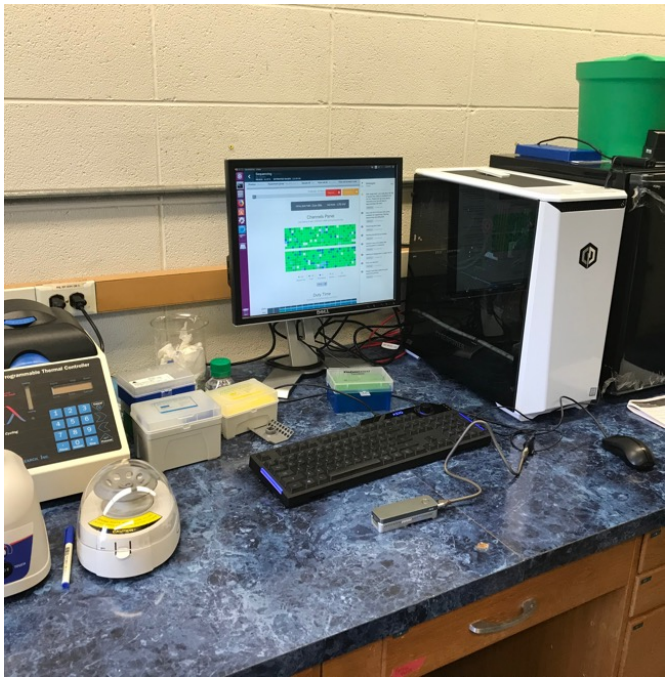




https://www.sciencedirect.com/science/article/pii/S1672022916301309

- Thousands of parallel pores are embedded into a “flowcell”
- Flowcells are run via a hand-held, USB-powered device called a MinION





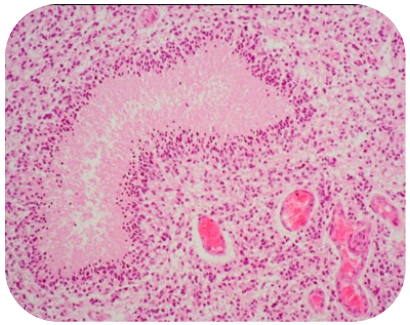
Nanopore Sequencing Lab at UM EECS

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

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

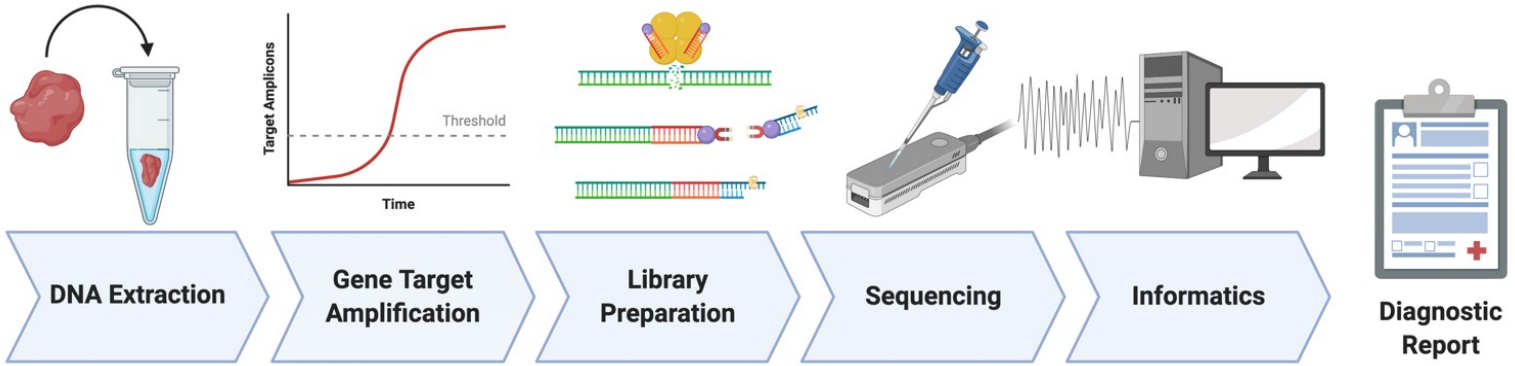
The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary

David N. Louis[1] · Arie Perry[2] · Guido Reifenberger[3,4] · Andreas von Deimling[4,5] · Dominique Figarella-Branger[6] · Webster K. Cavenee[7] · Hiroko Ohgaki[8] · Otmar D. Wiestler[9] · Paul Kleihues[10] · David W. Ellison[11]

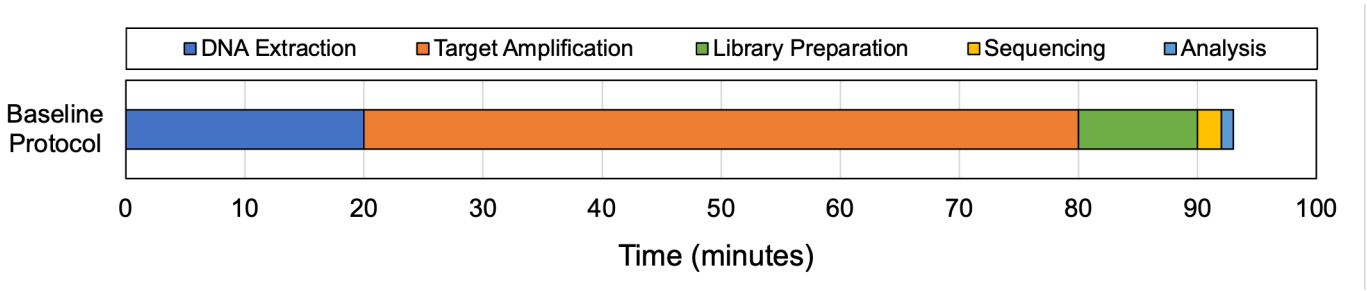
"For the first time, the WHO classification of CNS tumors ***uses molecular parameters*** in addition to histology to define many tumor entities, thus formulating a concept for how CNS tumor diagnoses should be structured in the molecular era."

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

How does a sequencing-based molecular diagnostic work?



- Target amplification uses the Polymerase Chain Reaction (PCR) to exponentially amplify a region of the genome
- PCR exponentially amplifies a small cancer-relevant gene target that might contain a mutation
- Amplified targets can then be sequenced to determine if a mutation is present



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

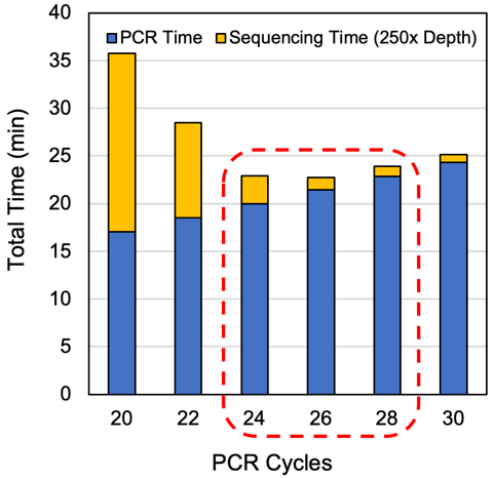
$$T_{total} = T_{amp} + T_{seq}$$

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

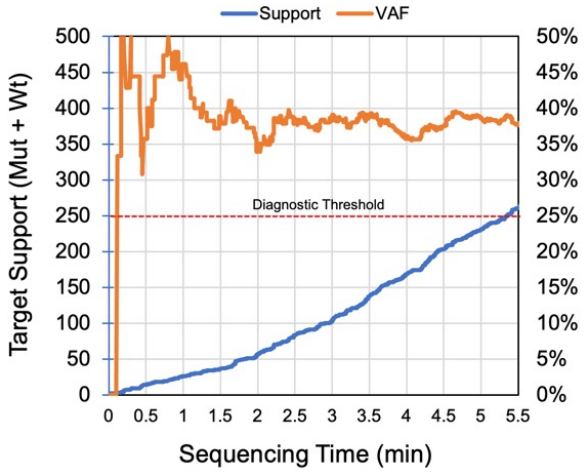
$$F_{target} = \frac{2^{N_{cycle}}}{2^{N_{cycle}} + N_{background}}$$

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

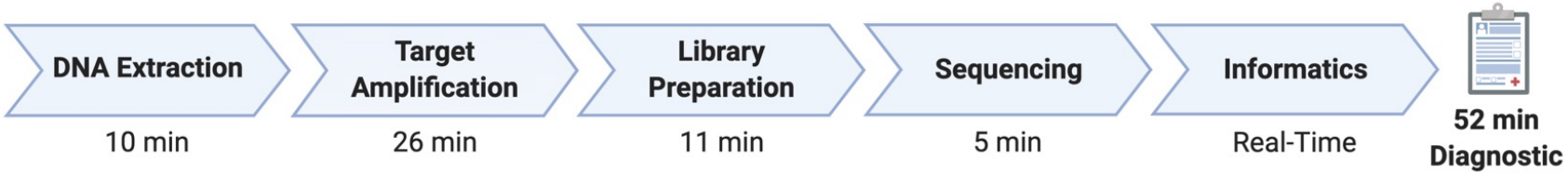
2) Augment model with experimentally derived parameters



3) Run diagnostic with final optimal parameters



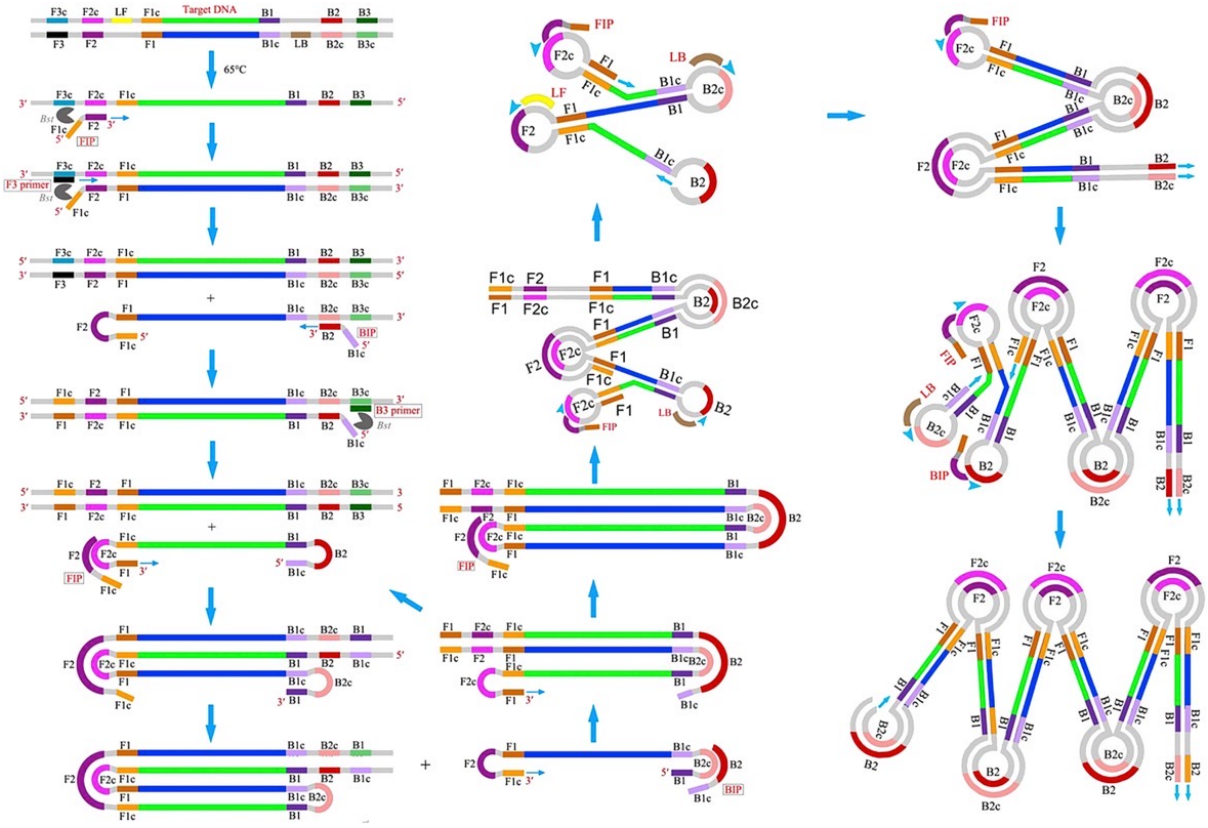
Co-optimization allowed for a world-first demonstration of a sub-1 hour sequencing-based diagnostic



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

N=6 concatemer

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

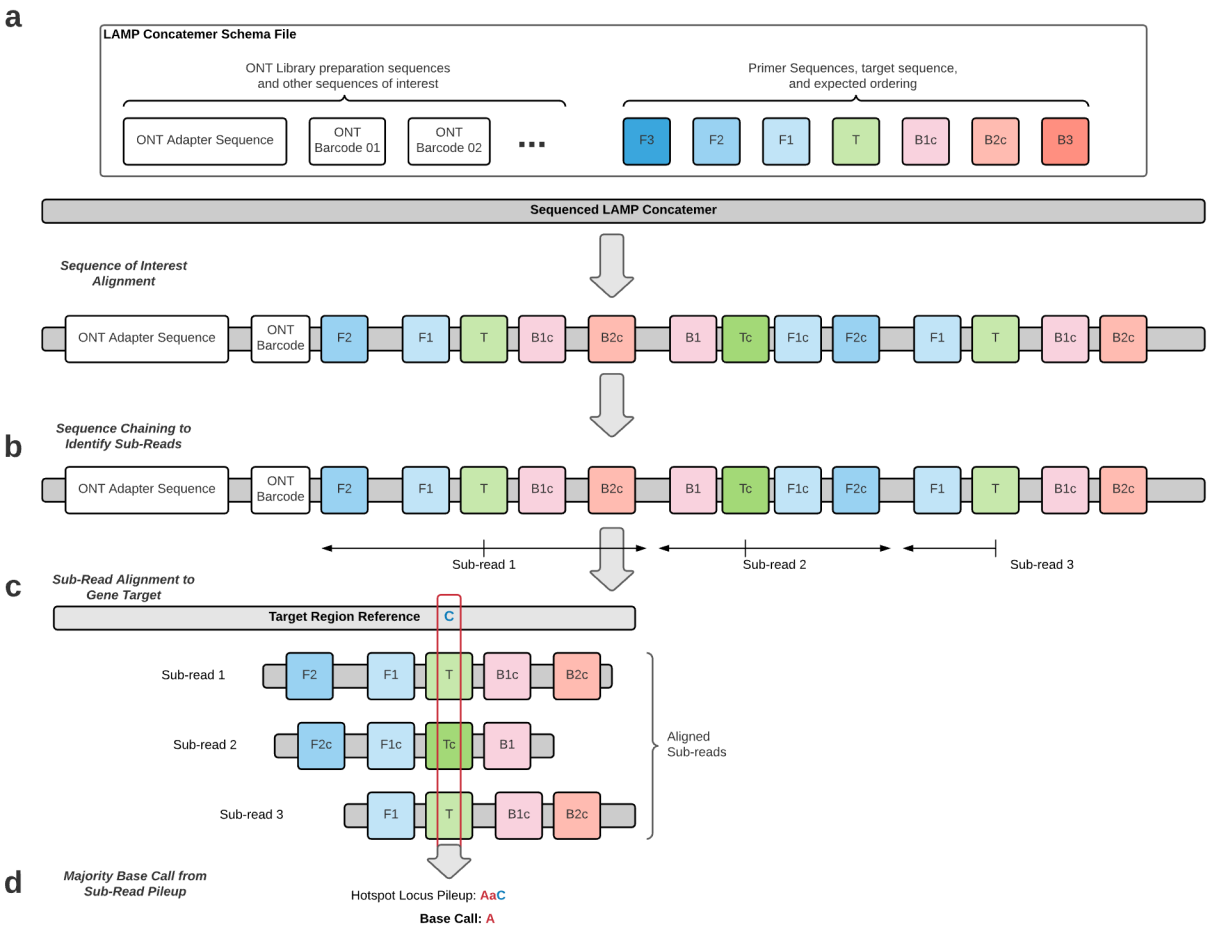
We leverage LAMP's rapid amplification and redundant information to further reduce diagnostic time

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

LAMPrey identifies concatemer “sub-reads” in noisy amplicons

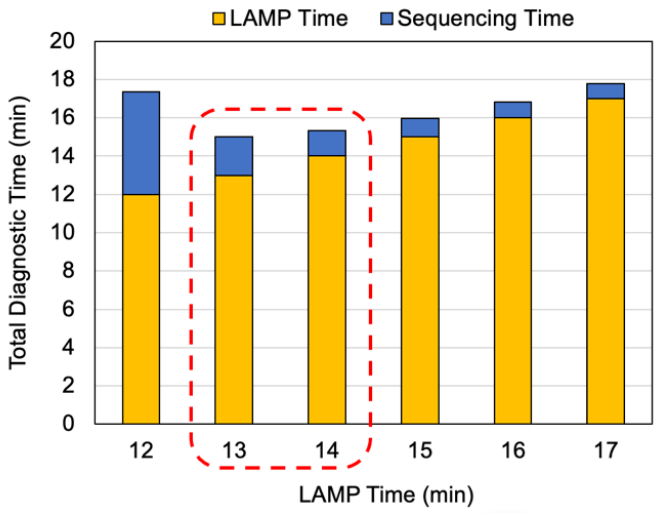
LAMPrey is able to recover about 50% more information than traditional informatics tools

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

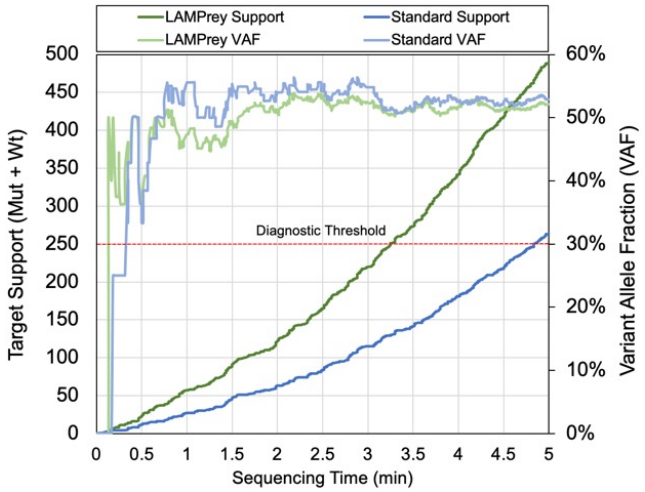


LAMPrey + Threshold Sequencing = <30min Sequencing-based Diagnostic

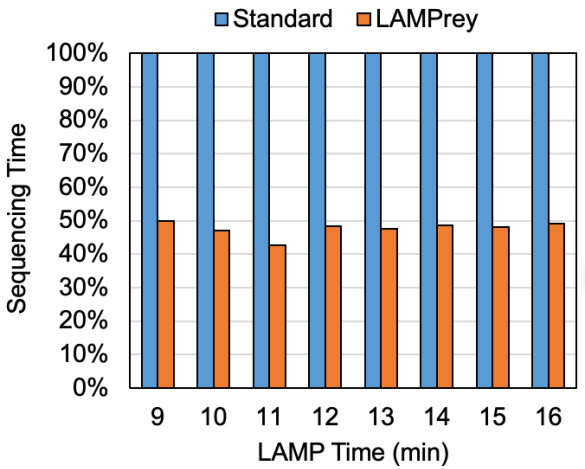
Experimentally informed LAMP diagnostic model



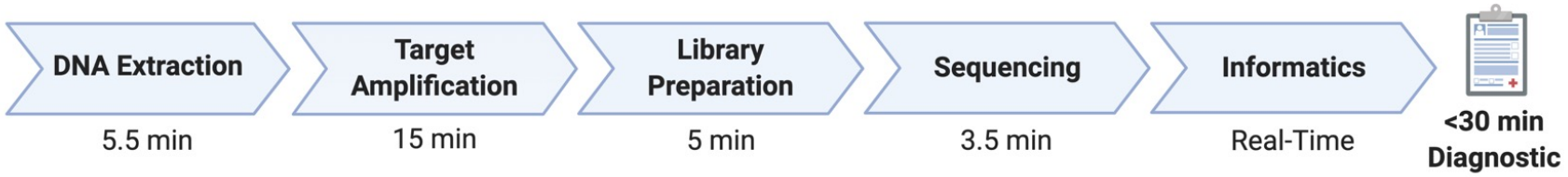
Final LAMP diagnostic result



LAMPrey benefit

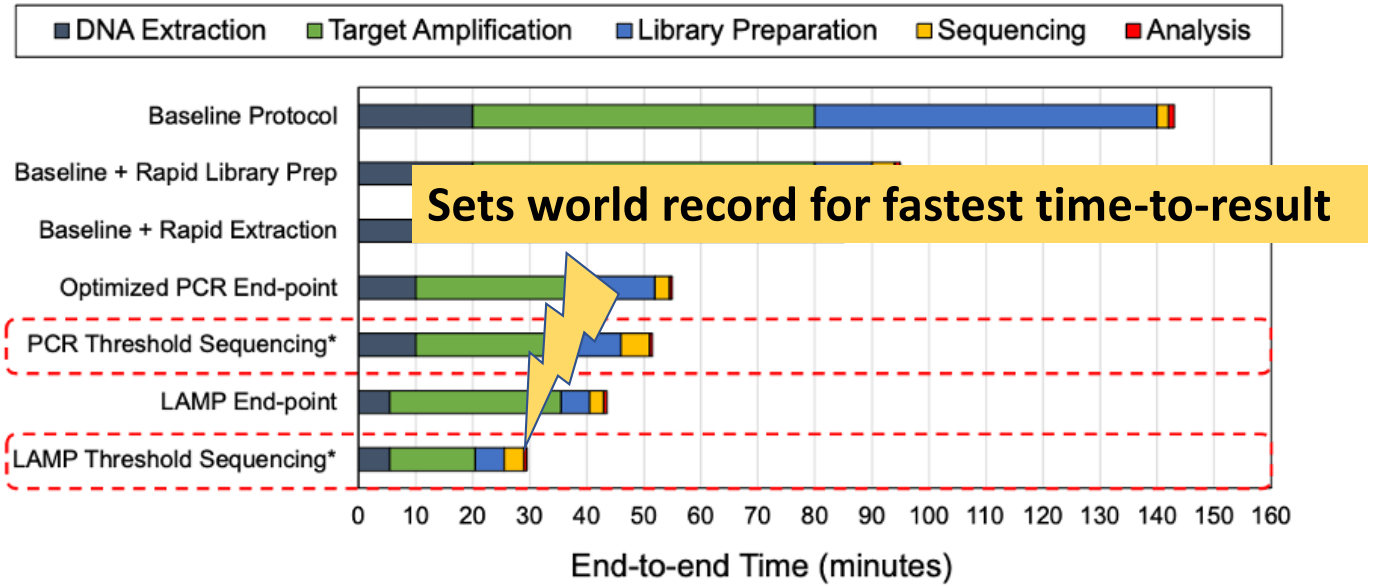


LAMPrey and other optimizations allowed for a world-first demonstration of a sub-30 minute sequencing-based diagnostic

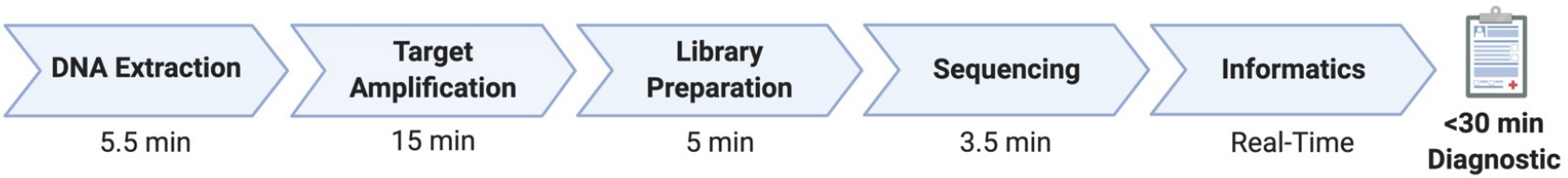


Open source: https//www.github.com/jackwadden/lamprey

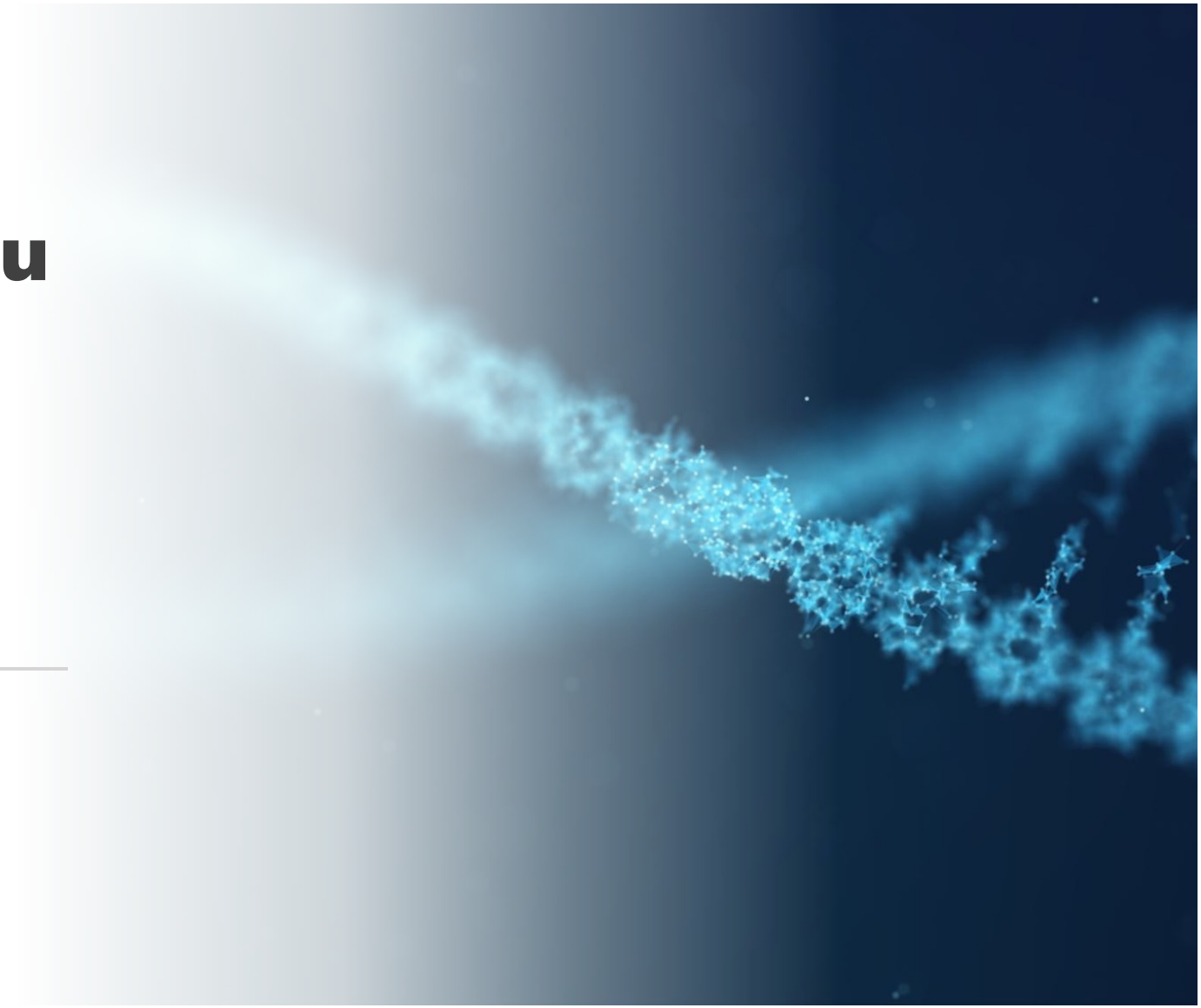
LAMPrey + Threshold Sequencing = <30min Sequencing-based Diagnostic

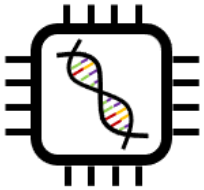


LAMPrey and other optimizations allowed for a world-first demonstration of a sub-30 minute sequencing-based diagnostic

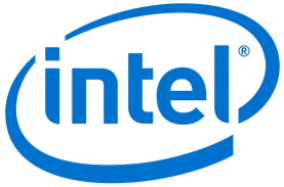


How Can You Kick-Start Precision Health Research?





GenomicsBench



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

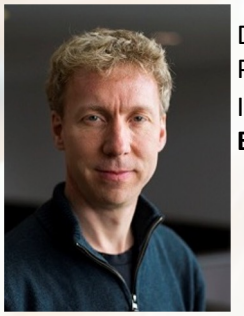
Includes both short and long read analysis algorithms

Small/large input datasets

Team – Part of University of Michigan Precision Health Initiative

Reetu Das
Assoc. Professor, UM
Sloan Fellow
ISCA and MICRO Hall of fame
Expertise: Systems

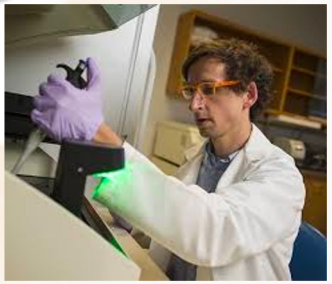
Satish Narayanasamy
Professor, UM
NSF CAREER
ISCA and ASPLOS Hall of fame
Expertise: Systems



David Blaauw
Professor, UM,
IEEE Fellow
Expertise: VLSI Design

Jenna Wiens
Assoc. Professor, UM
MIT TR 35 under 35
Expertise: Machine Learning

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."

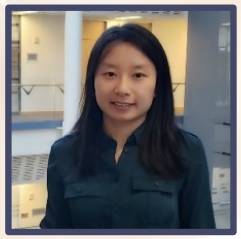
PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Work from Awesome Group of Fantastic Students!!

Arun Subramaniyan

Daichi Fujiki

Jack Wadden



Xiao Wu

Timothy Dunn

Hari Sadasivan

Yufeng Gu

"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

Thank You!

Reetu Das

Associate Professor

EECS Department