Systems for Precision Health

Reetu Das
Associate Professor, EECS Department
University of Michigan
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
What is Precision Health?

One Size Fits All  ❌ Precision Medicine  ✓
Precision Health Platform

Data Sources

- Electronic Health Records
- Genome
- Epigenome
- Microbiome
- Metabolome

Decision Support Tools

- Big Data Analytics
- Sequencing
- Analysis

Clinical Decisions

- Diagnosis and Prevention
- Treatment and Management

Credits: Created from BioRender.com
System Design Considerations

Data Sources
- Electronic Health Records
- Genome
- Epigenome
- Microbiome
- Metabolome

Decision Support Tools

Clinical Decisions
- Diagnosis and Prevention
- Treatment and Management

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

Agricultural sequencing

Portable Pathogen detector

Food Safety
Acceleration Study – Whole Genome Sequencing
**Acceleration Study: Whole Genome Sequencing**

**Human Genome**  
6 G bases

**Sequenced reads**  
(~billions)

<table>
<thead>
<tr>
<th>Reference genome</th>
<th>Reference genome</th>
<th>Reference genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATCGTGCAGT</td>
<td>ATCGTGCTACT</td>
<td>ATCGTGCTACT</td>
</tr>
<tr>
<td>GTGCACTCTAC</td>
<td>TCTGCTAAAGT</td>
<td>ATCGTGCTACT</td>
</tr>
<tr>
<td>CAGTACATCG</td>
<td>GAAGTTTT</td>
<td>CAGTACATCG</td>
</tr>
<tr>
<td>ATCGTGCTACT</td>
<td>ATCGTGCTACT</td>
<td>ATCGTGCTACT</td>
</tr>
</tbody>
</table>

**Read Alignment**  
BWA-MEM

**Sorting + Duplicate Marking**  
Picard SortSam, MarkDuplicates

**Variant Calling**  
GATK Haplotype Caller

**Diagnosis**

**Time (hr)**  
0  5  10  15

Baseline  
m5.8xlarge  
32 vCPUs

13.9 hr (i.e. 445 CPU hrs)
Read Alignment: GenAx

Approximate string matching

Ref. (R) ATCGA−CGTAGAT
Read. (Q) A−CGAACCTATAT

Banded Smith-Waterman

O(kn)

String Independent Local Levenshtein Automata (Silla)

O(k^2)

SillaX ASIC fabricated (55nm)

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

SillaX hardware accelerator

In-place Traceback

Affine gap Scoring

Composability
Read Alignment: SeedEx

Full-band implementation

- Low utilization
- Area inefficient

Banded implementation

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

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-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

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 ~ Chimpanzee  
96%

Human ~ Cat  
90%

Human ~ Cow  
80%

Human ~ Banana  
50-60%

Human ~ Human  
99.9%

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

Normal

DNA

| chr10:94781859
| GT | AG | CCCGGG |

| GT | AG | CCCGGG |

Transcription

| GT | AG | CCCGGG |

GT-AG Splicing

Normal RNA

| 40-bp deletion |

Normal Protein

| Non-coding (intron) |

Aberrant

DNA

| chr10:94781859
| GT | AG | CCCAGG |

| GT | AG | CCCAGG |

Transcription

| GT | AG | CCCAGG |

GT-AG Splicing

Aberrant RNA

| Coding (exon) |

Aberrant Protein

| X |

Plavix 75 mg
**Baseline**
- m5.8xlarge
- 32 vCPUs

**FPGA system**
- f1.4xlarge
- + 16 vCPUs

**ASIC system**
- ~5.5x
- ~7.3x

---

**Other Acceleration Candidates**
- FPGA system
- ASIC system

---

**Seeding**
- Sorting / Mark Duplicates
- PairHMM
- Other*

---

**Active Region of Reference**
- Ref: C A C T C C A C
- Hash Table
- De-Bruijn graph assembly

---

**De-Bruijn graph assembly**
- Hash (k-mer)
- Haplotype
- SNP
- Sequencing error
Acceleration Study – Ultra Rapid Cancer Diagnosis
Sequencing Technologies: Evolution

**Illumina Sequencing by Synthesis**

- **Illumina Genome Analyzer, 2005**
- **Illumina NovaSeq 6000, 2021**
- Read length: 100-350bp
- Per base inaccuracy: 0.1%
- 1 Gbases/per day
- 3 Tbases/per day
- **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](https://doi.org/10.7875/togopic.2020.01), Wikipedia DMLapato [https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina](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)
- 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!

"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

\[ T_{\text{total}} = T_{\text{amp}} + T_{\text{seq}} \]

\[ T_{\text{amp}} = T_{\text{init}} + T_{\text{cycle}} \times N_{\text{cycle}} + T_{\text{final}} \]

\[ F_{\text{target}} = \frac{2^{N_{\text{cycle}}}}{2^{N_{\text{cycle}}} + N_{\text{background}}} \]

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

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

DNA Extraction 10 min
Target Amplification 26 min
Library Preparation 11 min
Sequencing 5 min
Informatics Real-Time
52 min Diagnostic

but target amplification is still a large bottleneck...
Loop-Mediated Isothermal Amplification (LAMP) Technology

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

DNA Extraction  Target Amplification  Library Preparation  Sequencing  Informatics
5.5 min  15 min  5 min  3.5 min  Real-Time  <30 min Diagnostic

Open source: https://www.github.com/jackwadden/lamprey
LAMPrey + Threshold Sequencing = <30min Sequencing-based Diagnostic

Sets world record for fastest time-to-result

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
“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.”
Work from Awesome Group of Fantastic Students!!

“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.”

Arun Subramaniyan  Daichi Fujiki  Jack Wadden  Xiao Wu  Timothy Dunn  Hari Sadasivan  Yufeng Gu
Thank You!

Reetu Das
Associate Professor
EECS Department