HOT CHIPS

Hot Chips Poster Session

A Fully Integrated Genetic Variant Discovery SoC for Next-Generation Sequencing

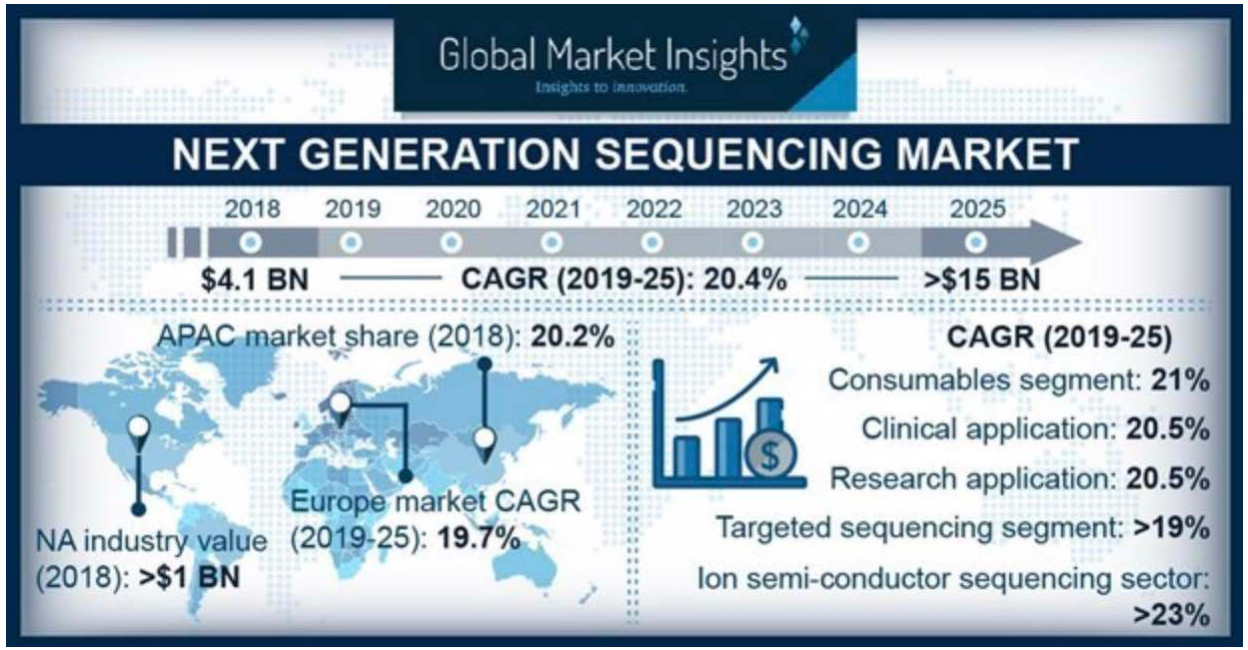
Y.-C. Wu[1], Y.-L. Chen[1], C.-H. Yang[1], C.-H. Lee[2], C.-Y. Yu[3], N.-S. Chang[3], L.-C. Chen[3],
J.-R. Chang[3], C.-P. Lin[3], H.-L. Chen[3], C.-S. Chen[3], J.-H. Hung[2], C.-H. Yang[1]
[1]National Taiwan University, [2]National Chiao Tung University, [3]Taiwan Semiconductor Research Institute

Abstract

This work presents the first dedicated system-on-chip (SoC) that supports full data analysis workflow for next-generation sequencing (NGS) data analysis. The SoC implements four major steps: preprocessing, short-read mapping, haplotype calling, and variant calling. This work achieves comparable precision as the software solutions by adopting the sBWT and Genome Analysis ToolKit Haplotype Caller (GATK-HC) algorithms. The multitask sorting engine (MTSE) and the dynamic programming processing engine (DPPE) are proposed for essential computing tasks for NGS data analysis. An Synopsys ARC processor is integrated to facilitate IO interfaces, including DDR3 and SD card protocols. Fabricated in a TSMC 28nm technology, the chip area is 12mm^2. It dissipates 975mW at a clock frequency of 400MHz from a 0.9V supply. The SoC can complete full analysis for whole human genome in 37 minutes. Compared to an optimized GPU solution, this work is 66× faster, and achieves more than 15,000× and 3,000× higher in energy and area efficiency. The prototype system demonstrates the analysis in real-time for on-site DNA sequencing.

Demonstration Video:
https://www.youtube.com/playlist?list=PLm0H1AD1FScj4c7tWiOUaTkENqCjrmvCT

Next-Generation Sequencing (NGS)



- **NGS**
  - **Growing market value (>$15 BN by 2025)**
  - **Extensive applications: non-invasive prenatal test, quick virus test, cancer screening, targeted medicine, biotechnology research**

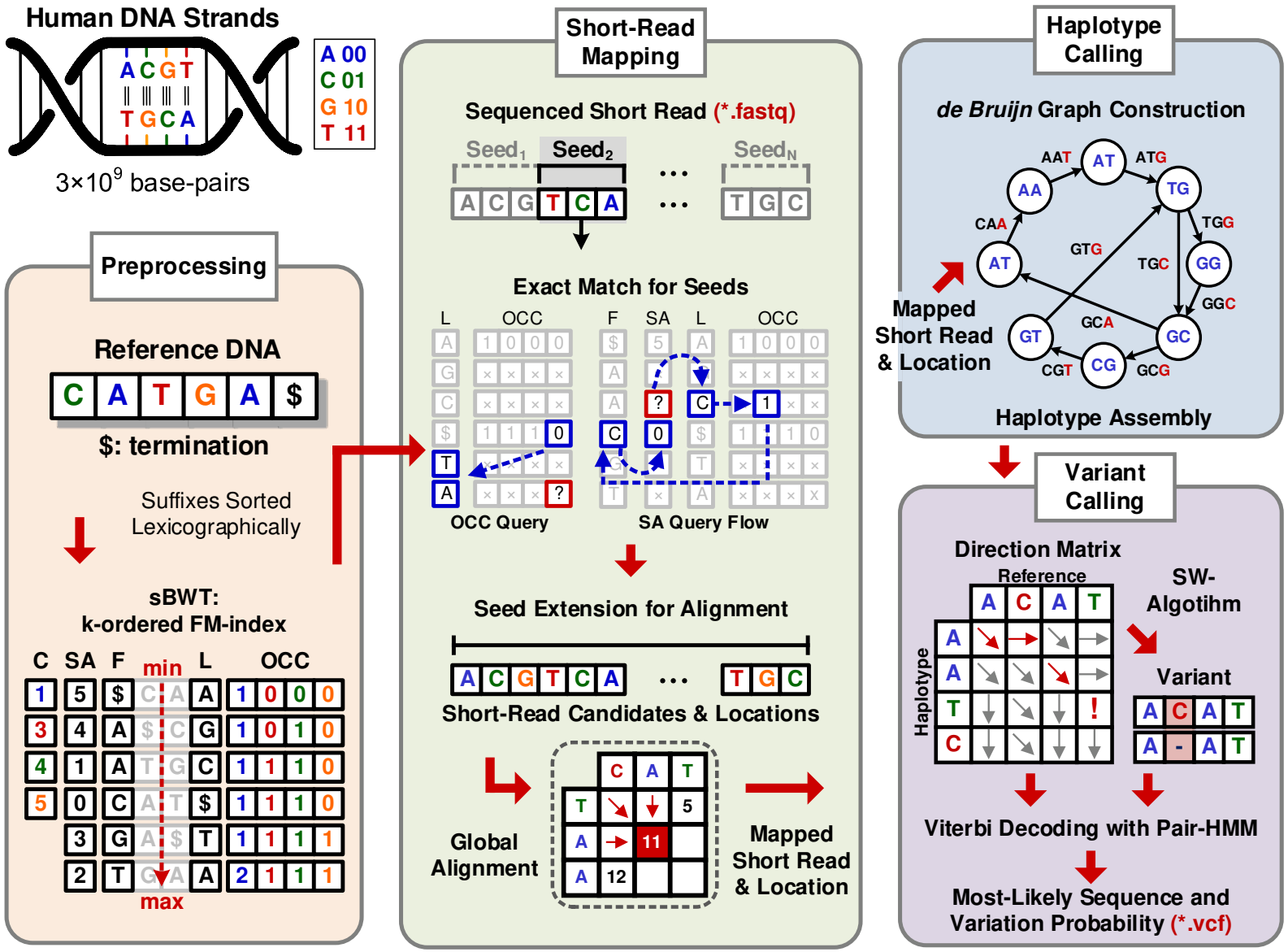
NGS Data Analysis

- **Length of DNA sequence: 3 billion (base pair)**
- **Amount of short reads: hundreds of millions**
- **Raw data to be processed: up to TB**
- **Analysis time: 3 days (optimized GPU)**

Technical Breakthrough

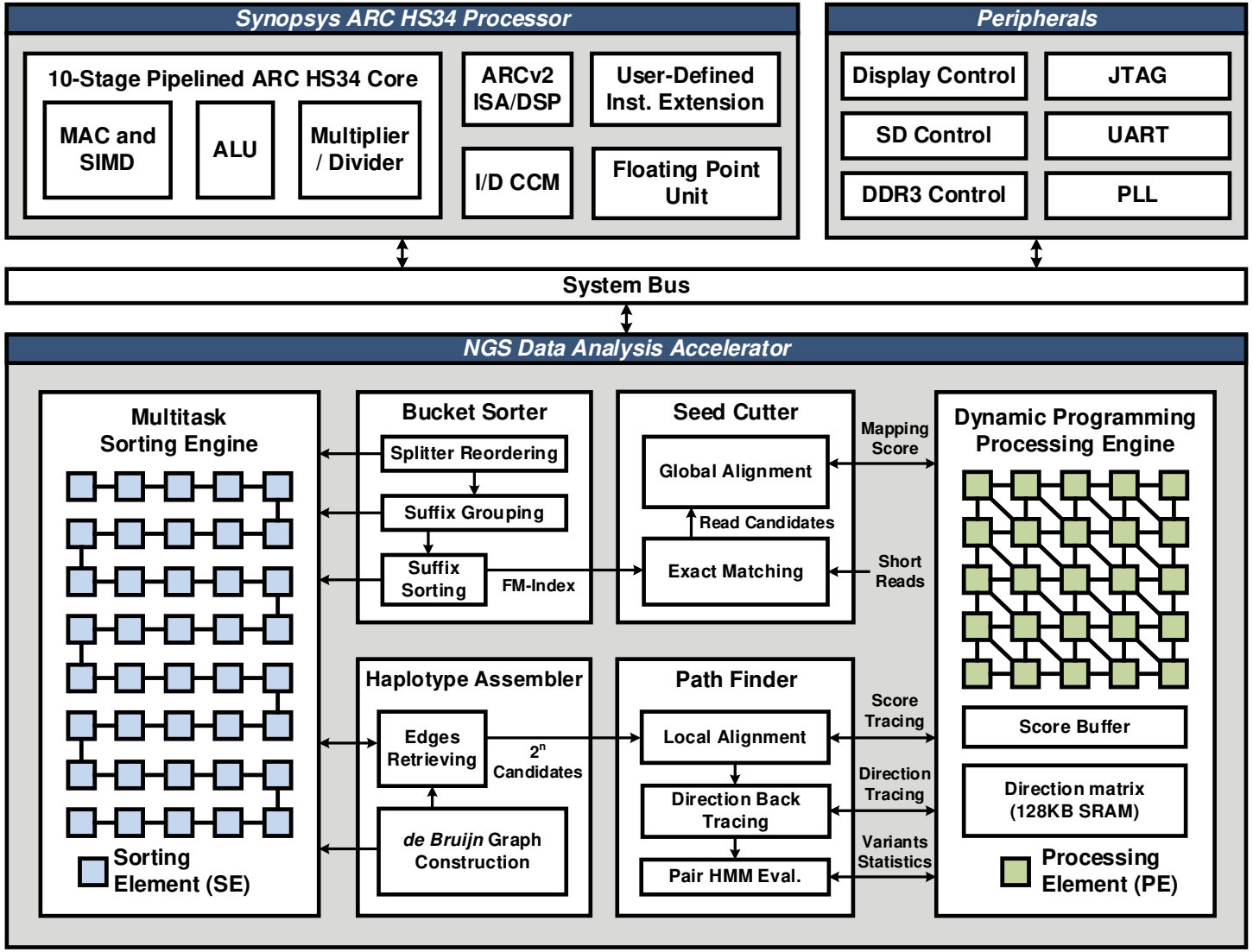
- **First dedicated SoC for entire NGS data analysis workflow**
- **Fast NGS data analysis**
  - **3 days (GPU) → 40 mins (this work)**
- **Low power dissipation: < 1W**
- **GUI for real-time status monitoring**

NGS Data Analysis Workflow



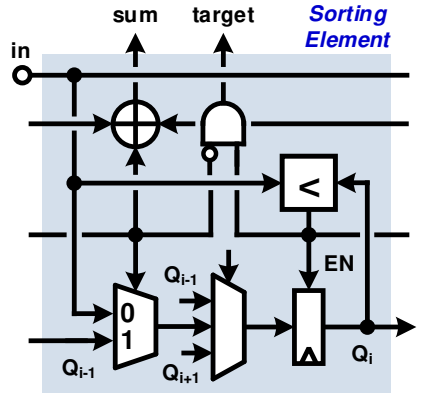
- **NGS Data Analysis Steps**
  - Preprocessing
  - Short-Read Mapping
  - Haplotype Calling
  - Variant Calling
- **Low Memory Requirement**
  - sBWT Algorithm
- **High Precision**
  - Genome Analysis Toolkit Haplotype Caller (GATK HC)
- **Standard File Format**
  - Input: Raw Sequenced Data
  - Output: Discovered Variants

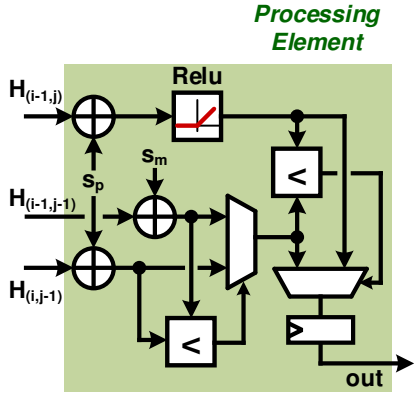
System Architecture and Design Optimization



Key Features

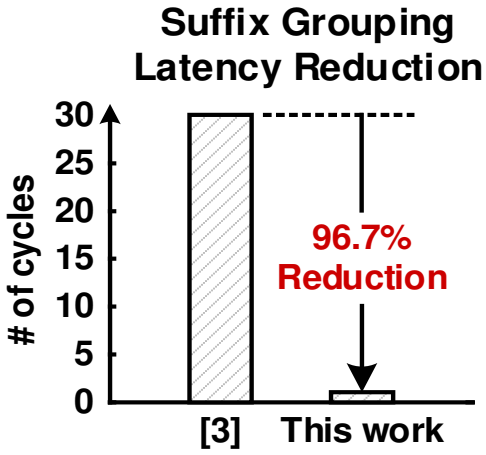
- **Heterogeneous Computing Platform**
  - Embedded Synopsys ARC HS34 processor
  - Dedicated NGS Data Analysis Accelerator
- **Multitask Sorting Engine**
  - Cascaded and parallel architecture for low latency data movement
  - Hardware sharing for area reduction
- **Dynamic Programming Processing Engine**
  - Support global, semi-global, local alignment
  - Area reduction: 1-dimentinal PE array

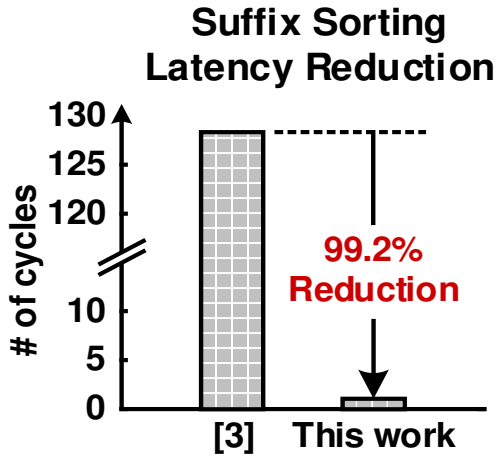


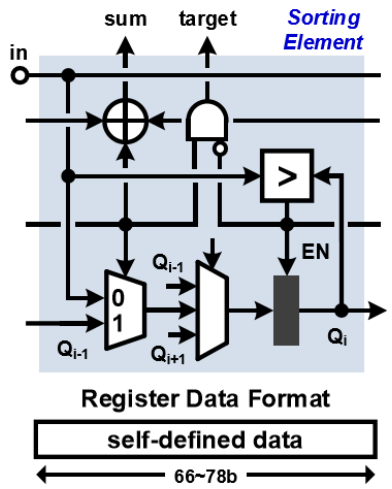


Multitask Sorting Engine (MTSE)

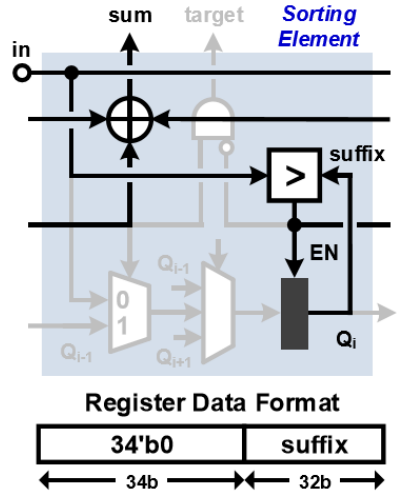
- **Supporting Functions**
  - Preprocessing: grouping, sorting
  - Haplotype Calling: *de Bruijn* graph, assembly
- **Key Features**
  - Cascaded parallel architecture
  - 1-cycle latency
  - Self-defined data format



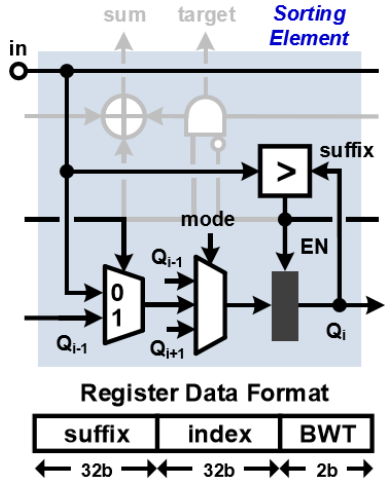




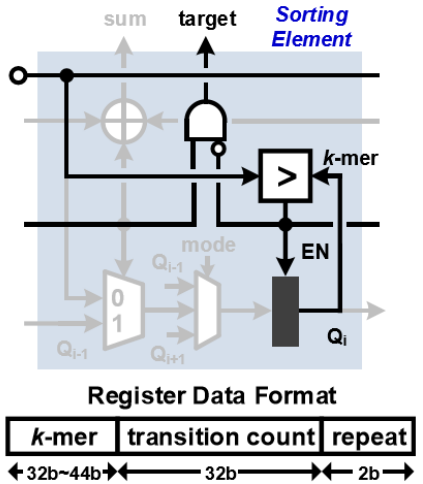
Overall Architecture



Grouping



Sorting

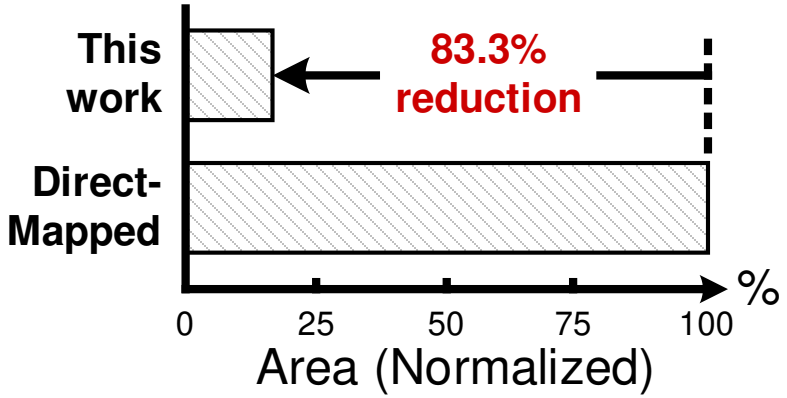


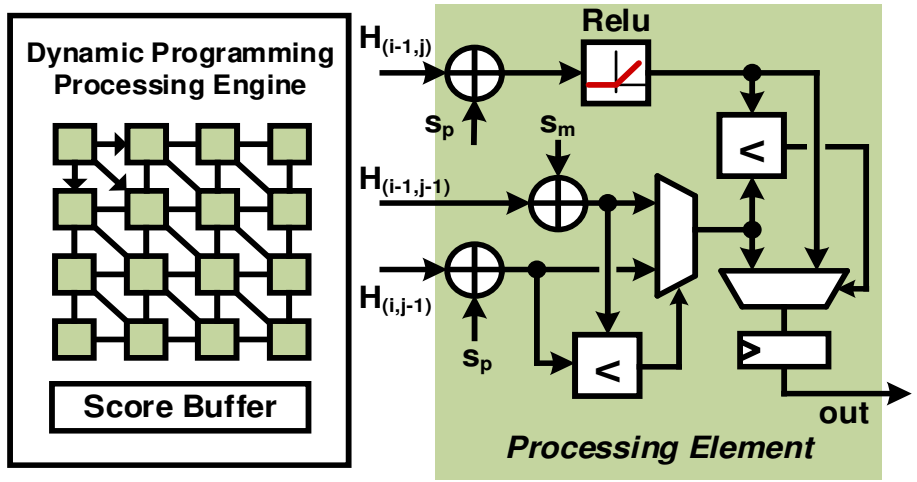
de Bruijn Graph

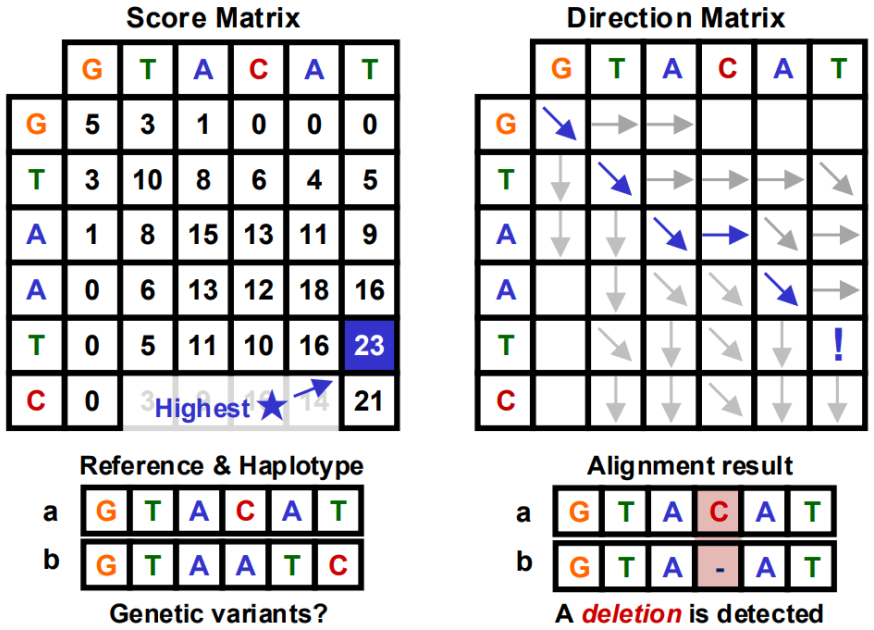
Dynamic Programming Processing Engine (DPPE)

- **Supporting Functions**
  - Short-Read Mapping: global alignment
  - Variant Calling: semi-global alignment
  - Viterbi Algorithm: local alignment

- **Design Techniques**
  - 2D systolic array → 1D cascaded architecture
  - Hardware Shared



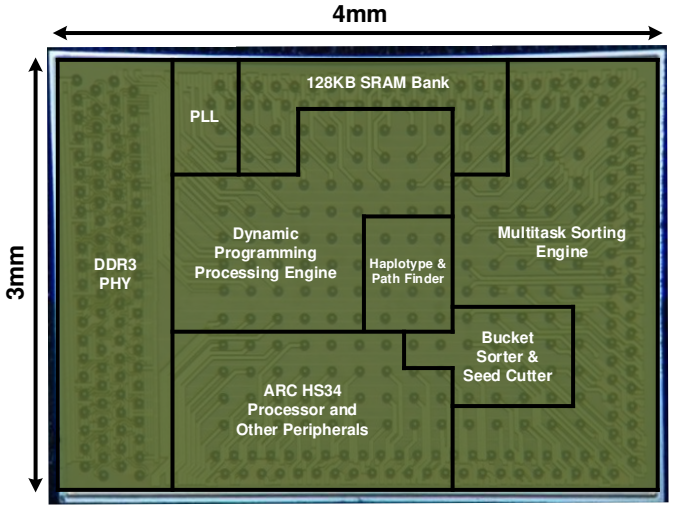


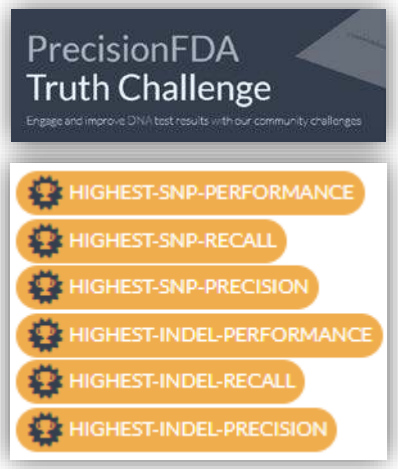


Chip Implementation and Performance Comparison

Core Area		3.63 x 2.63 mm^2
Gate Count		19.5 M
Core V_{DD}		0.9 V
Embedded SRAM		448 KB
Analysis Time	Chromosome 10	1.2 mins
	Whole genome	37 mins
Precision*		99.6%

*verified with PrecisionFDA Truth benchmark [1]





	NVIDIA GK110B [2]	This Work [3]
Technology	28nm	28nm
Chip Area [mm^2]	561	12
Clock Freq. [MHz]	705	400
Power [W]	235	0.975
External Memory	12GB GDDR5	1GB DDR3
Analysis Time [s]	4687	71 **66x ↓**
Energy Efficiency [task(s)/J]	9.1 x 10^{-7}	1.4 x 10^{-2} **>15,000x ↑**
Area Efficiency [task(s)/s/mm^2]	3.8 x 10^{-7}	1.2 x 10^{-3} **>3,000x ↑**

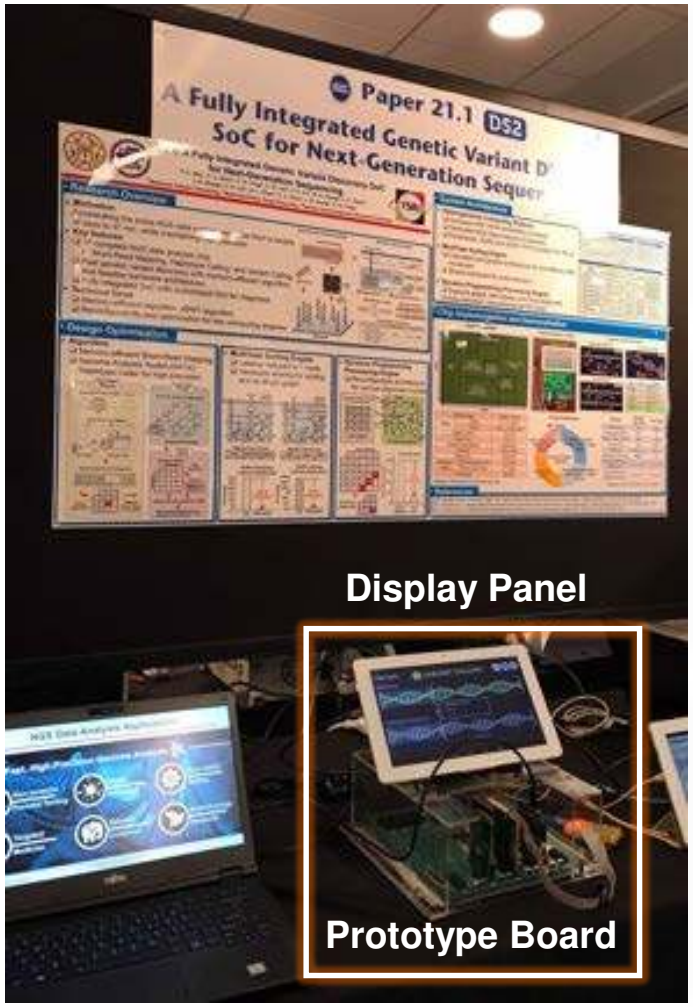
[1] https://precision.fda.gov/challenges

[2] S. Ren *et al.*, "GPU-Accelerated GATK Haplotype Caller with Load-Balanced Multi-Process Optimization," *IEEE Int. Conf. Bioinformatics Bioengineering*, 2017.

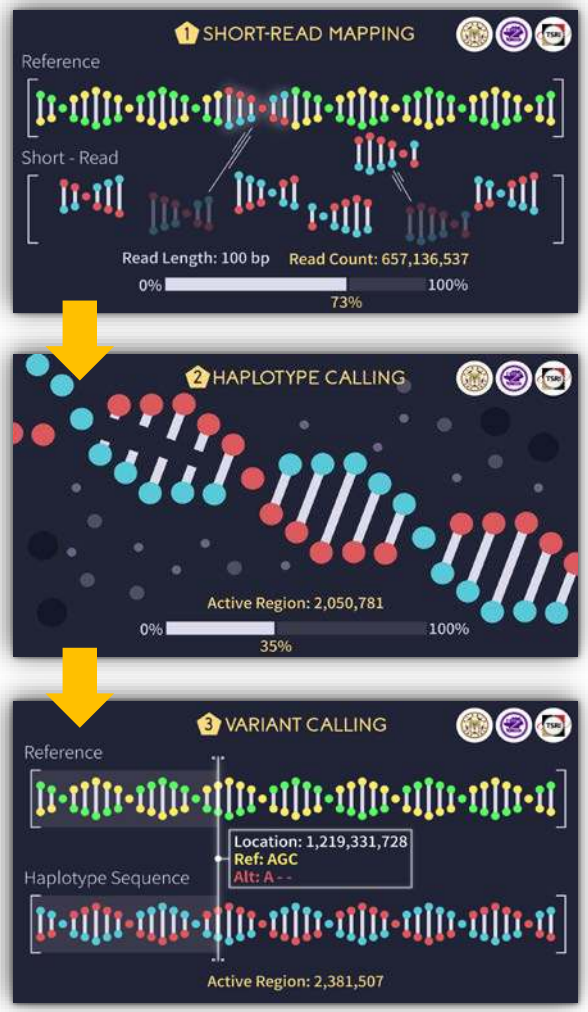
[3] Y.-C. Wu *et al.*, "A Fully Integrated Genetic Variant Discovery SoC for Next-Generation Sequencing," *ISSCC*, 2020.

System Demonstration

Demonstration Setup



GUI



Summary

- **First dedicated SoC for entire NGS data analysis workflow**
- **Optimized hardware architecture**
  - Multitask sorting engine
  - Dynamic programming processing engine
- **SoC integration in 28nm CMOS**
  - ARC processor, NGS data analysis accelerator, DRAM controller, PLL
  - <1W power consumption at 400MHz
- **Orders of magnitude improvement over GPU**
  - Processing time: 66x faster
  - Energy efficiency: >15,000x higher
  - Area efficiency: >3000x higher