

Accelerated and Accurate Alignment of Short Reads in High Throughput Next Generation Sequencing [NGS] Platforms



Santhi Natarajan (Research Scholar, CeNSE)

Advisors : Prof. Debnath Pal (CDS), Prof. S. K. Nandy (CDS)

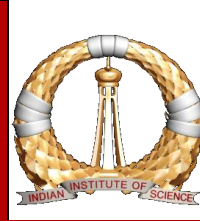
Indian Institute of Science, Bangalore



- **Genome: The Secret Code of Life**
- **Short Read Mapping: What we have?**
- **Short Read Mapping: What we need?**
- **Our Architecture**
- **Our Prototypes: Results and Scalability Analysis**
- **Putting it Altogether.....**

“It is far more important to know what person the disease has than what disease the person has.” – Hippocrates

Genome: The Secret Code of Life

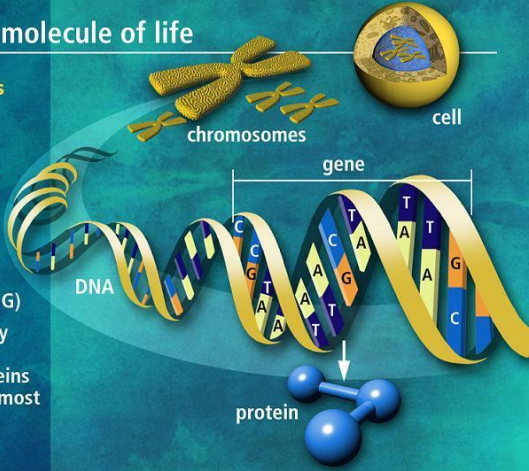


DNA the molecule of life

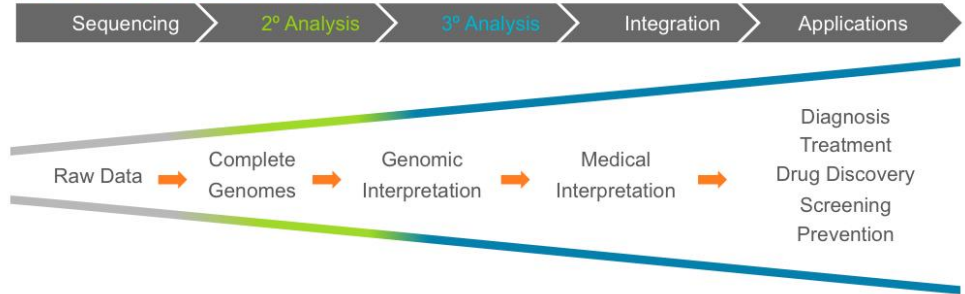
Trillions of cells

Each cell:

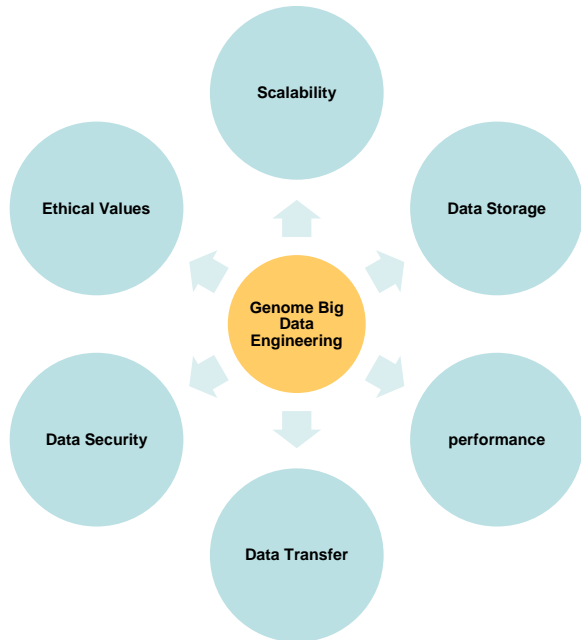
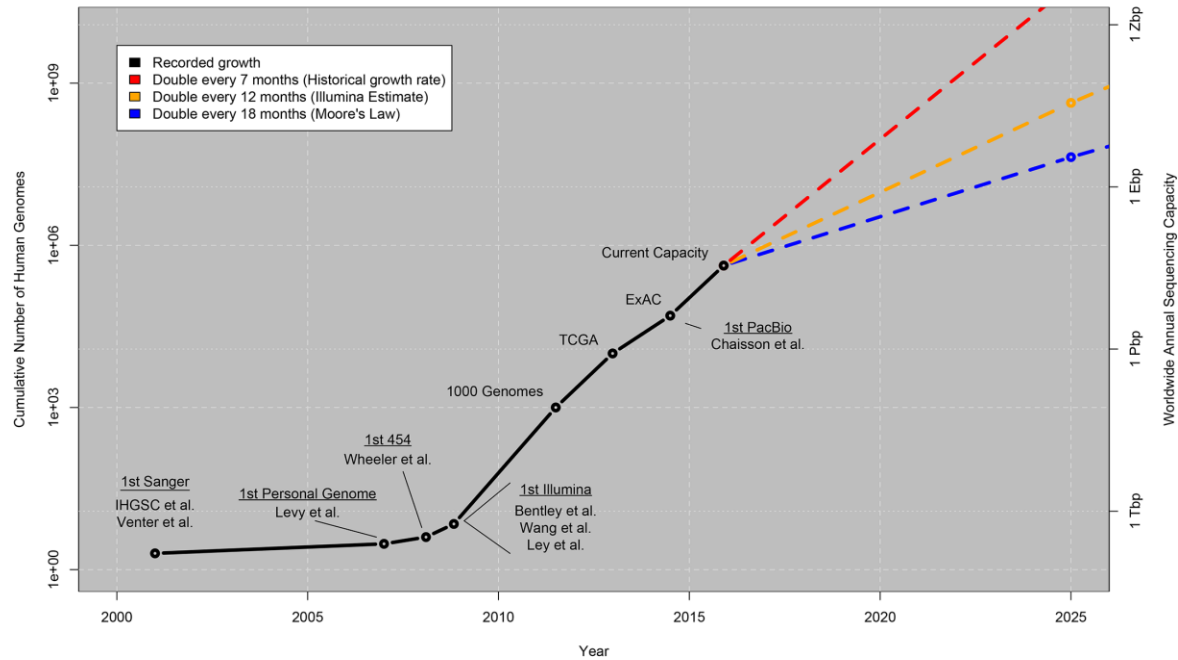
- 46 human chromosomes
- 2 meters of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- Approximately 30,000 genes code for proteins that perform most life functions



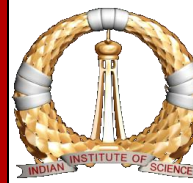
Genomics Landscape



Growth of DNA Sequencing



Short Read Mapping (SRM): What we have ?

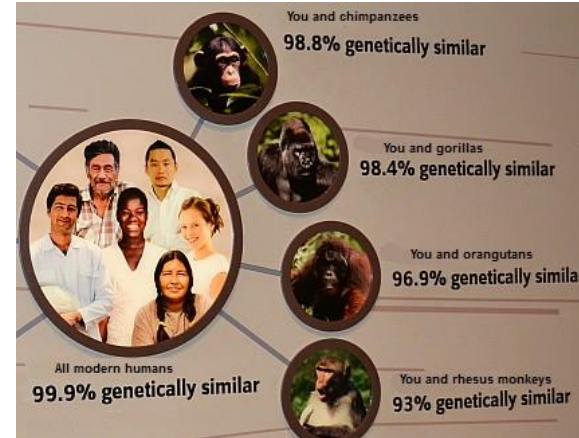
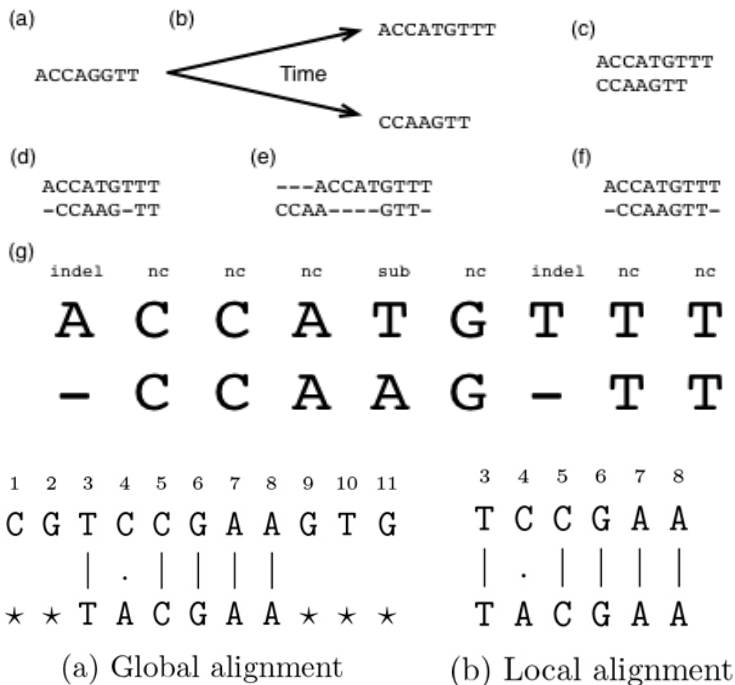


```

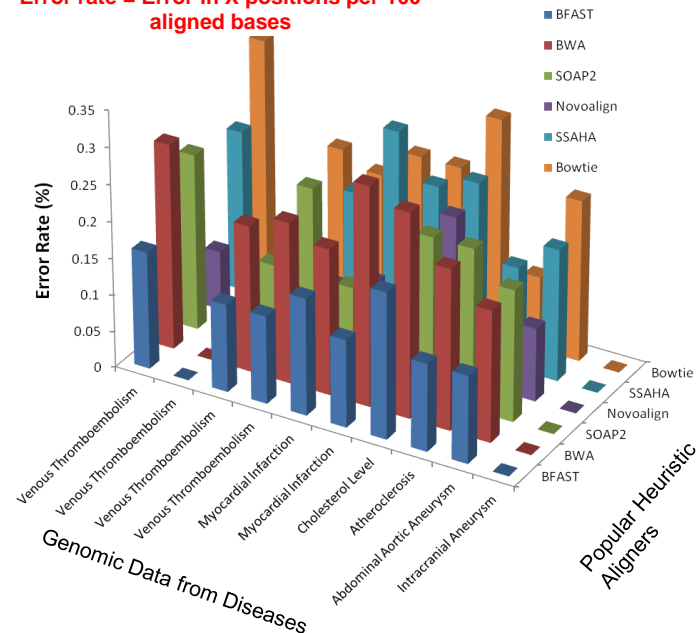
GTAATCAGCTACCTTTATCCCTAGCCCCCTGCGCCCCGCGCCCTTGGCA
GTATGGTGAGGTCTGCTCTGGTCCCTCGCCACCATGTACGTGAGC
GTACCTAGTCTCGTGGCCACTCCGGCGGCCTCAACTGGCGC
GTACCCGGACTACGGTGGTTACCACGTGGCGGCCGCGCTGTGC
TCCAGGGCCATCCTGGCCACCAGTACGGCGCCCTCTCCGC
TGGGCGCTGCCAGCCAACCGGTAGCCACCGCCTCAATGTT
GCTAGCGAGCCCGCGAATACCACGACACATCACCCGCATCATCA
GCTCTGCGCTCCGGATTGCTGCAGACGCTCAACCCGGCCCTCCAI
CAGCTGCTCCAGCGGCAGCGCCGAAACCTTTTGAATGGA
GAAAGCAAGGTATGCGGTGCTGGGCGCGCCCTGTCGCC
TAGACAGGACAGGAGAAGGACAAGGGGAGAAATAGGGGTG
GTATATTTGTGCTCGACCCCAAGTTTGTAGGGCAGAAGAGT
ACCTCGTTGGCAACTCGGAGCCCTCGGAAGCTCCCGTAGTGC
TAGCAGAATCGCTCCTTAGTTGGCTGCCACGACCTTCTACCCI
GGGTGATTGGATCCCTTGGCCGAGCTACTCACTTAGCTACCCACTCT
    
```



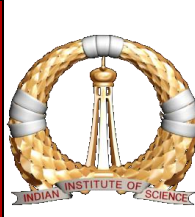
Approximate String Matching



Error rate = Error in X positions per 100 aligned bases



SRM: What we need?



Accuracy with Precision

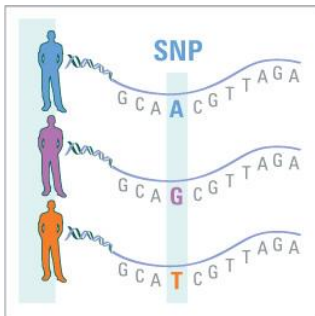


Problem Statement:

If the reference genome is very large, and if there are billions of short reads, how accurately can we align the reads to the genome, guaranteeing precision and performance?



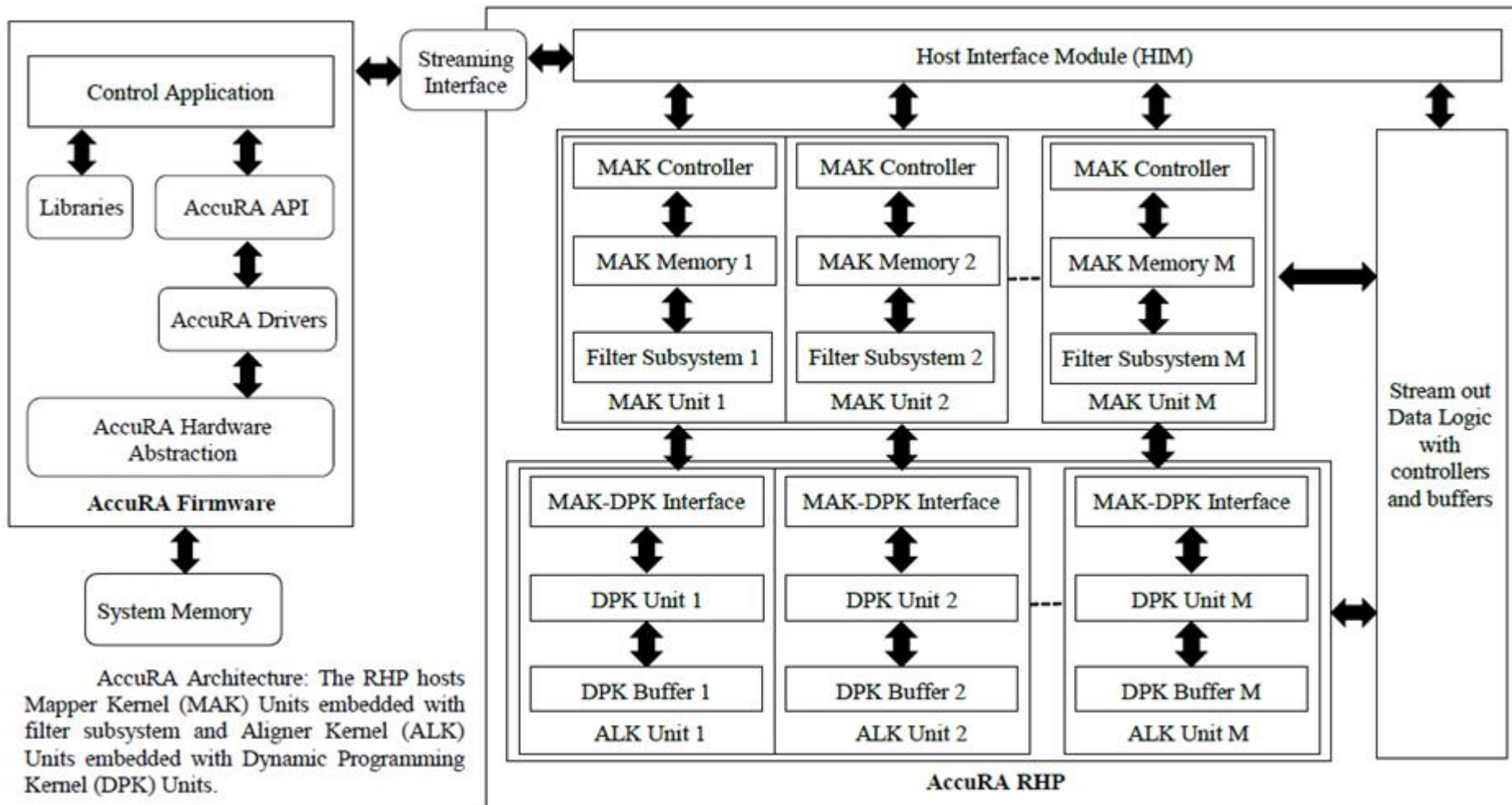
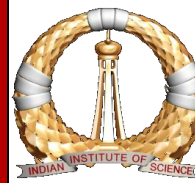
Cost-effective



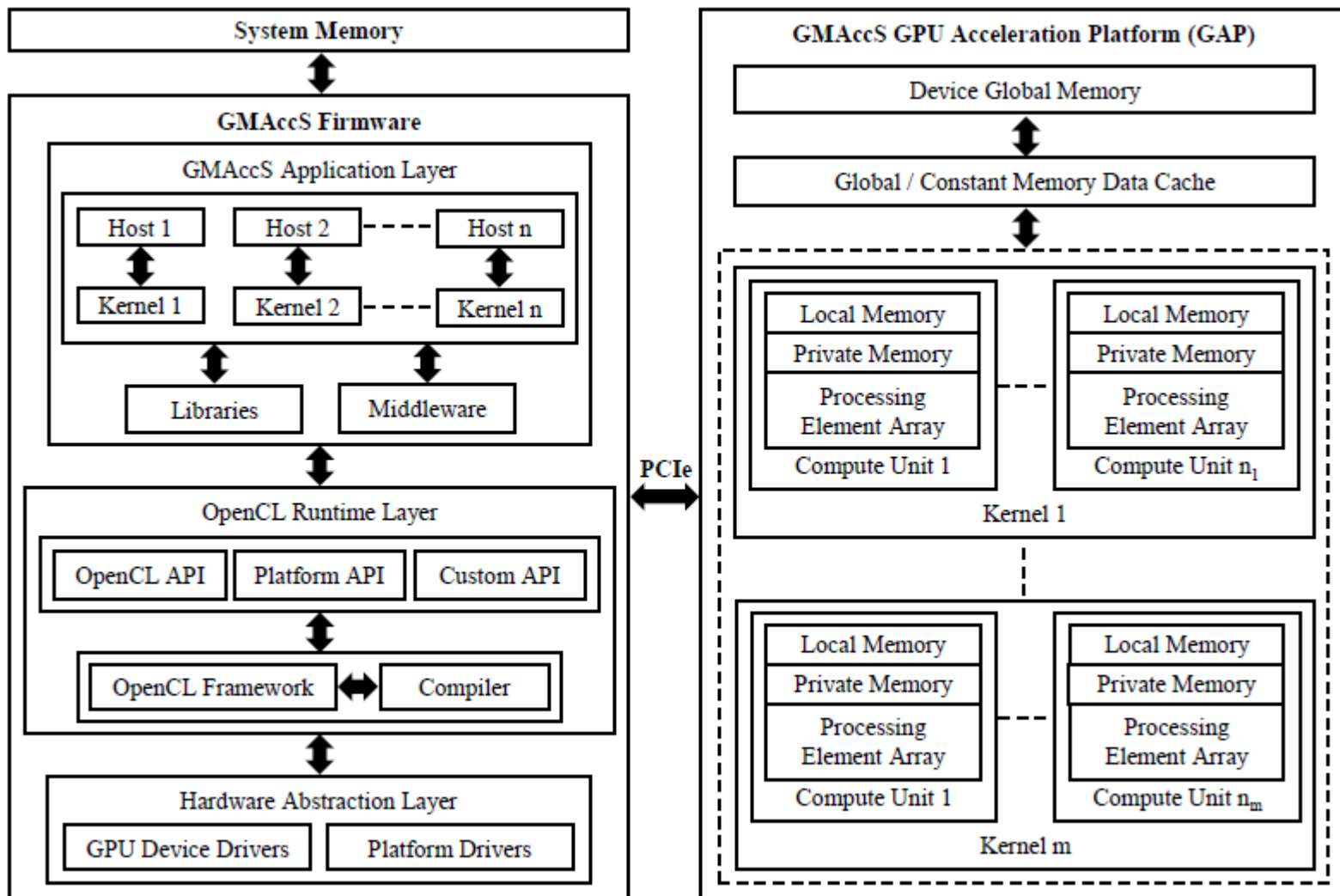
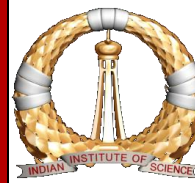
Resolution



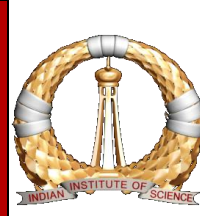
AccuRA: Accurate Alignment of Short Reads on Scalable Reconfigurable Accelerators



GMAccS: A Scalable GPGPU Model for Accurate Alignment of Short Reads

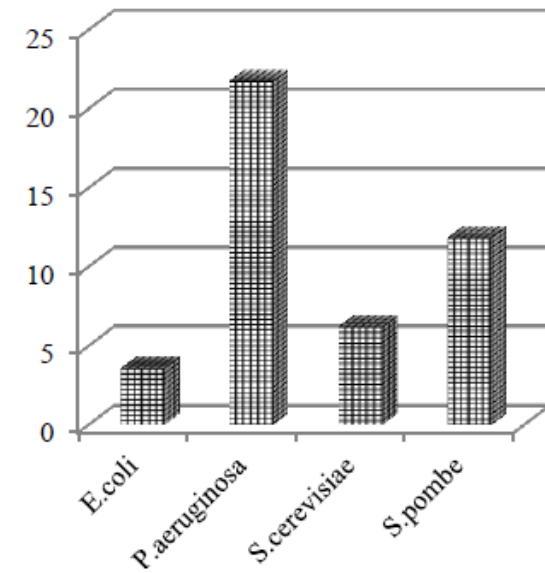


AccuRA Prototype: Performance and Scalability Analysis



Feature	E. coli	P. aeruginosa	S. cerevisiae	S. pombe
Read Size	502	600	50	70
No. of Reads	596100	1245456	6632993	7545533
Subread length	32	32	32	32
Total subreads	10133700	24909120	13265986	22636599
Num Maps	51351473	319544370	91371619	174206729
GC Content (%)	50.79	66.56	38.15	36.05

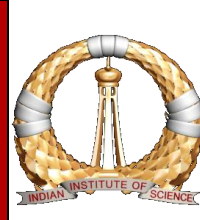
Alignment Time (s)



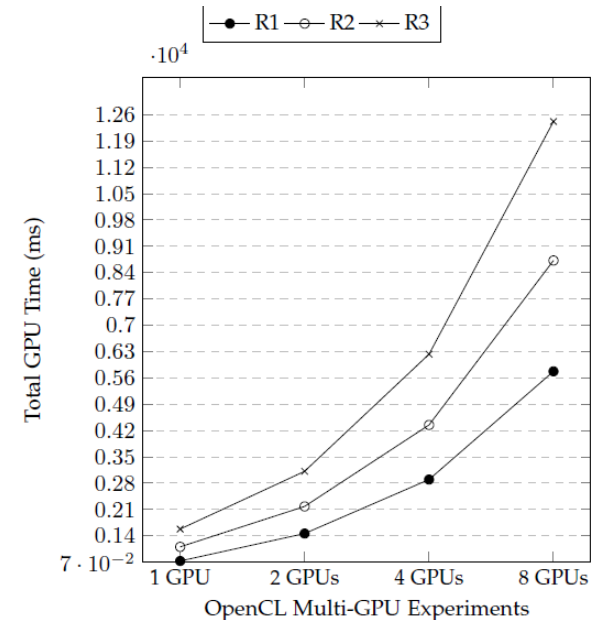
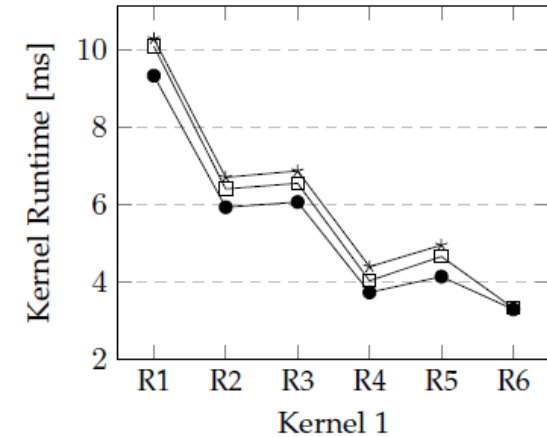
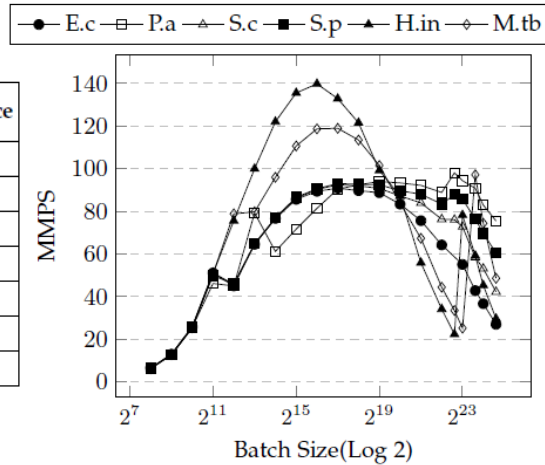
No. of Units, N	Filter Operations, N*K	GMPS	Cell Updates, N*C	GCUPS
8	3720	74.4	8192	20.48
16	7440	148.8	16384	40.96
32	14880	297.6	32768	81.92
64	29760	595.2	65536	163.84

Read Data Sets	SRR1559281	SRR1559282	SRR1559283
No. of Reads	142992687	146386600	144082500
No. of Pairs	5067156377	4898853334	5061571327
Alignment Time(s)	6214.239978	5962.010015	6066.540026
Alignment Time (min)	103.5706663	99.36683358	101.1090004

GMaccS Prototype: Performance for Small Genome Benchmarks

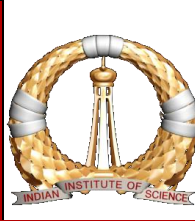


Genome	Optimal BS	Optimal NDR	Max. MMPS	Dominance
S.cerevisiae	131072	16	92.4637	AT
S.pombe	131072	16	93.008	AT
H.influenzae	65536	32	139.77	AT
E.coli	131072	16	90.8085	AT ≈ GC
P.aeruginosa	6418880	64	98.0067	GC
M.tuberculosis	131072	32	118.901	GC
B.pertussis	262144	128	538.47	GC

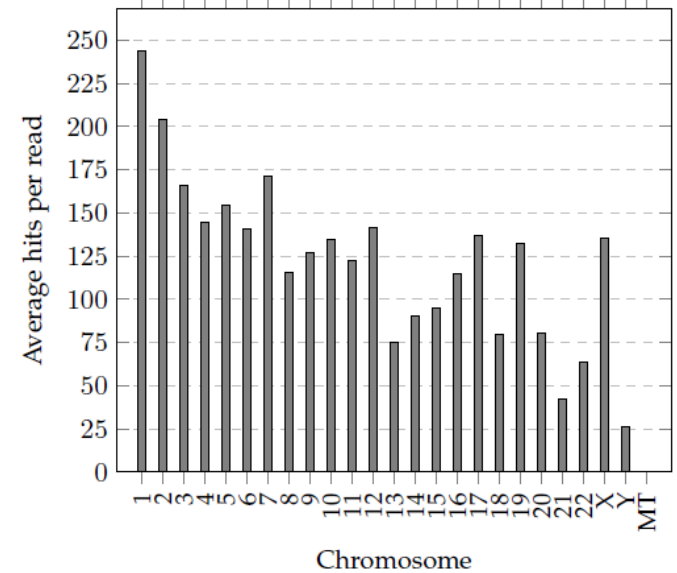


Run	Species	1 GPU	2 GPUs	4 GPUs	8 GPUs
R1	E.coli	32.49	65.49	130.22	261.3
	P.aeruginosa	67.78	135.86	272.53	544.72
	S.cerevisiae	44.38	89.52	178.66	358.28
	S.pombe	57.92	117.2	234.76	467.58
R2	E.coli	46.38	96.77	186.81	372.35
	P.aeruginosa	68.71	147.38	296.14	589.22
	S.cerevisiae	58.08	115.07	232.45	465.52
	S.pombe	67.13	132.52	269.61	538.03
R3	E.coli	70.16	139.63	281.68	563.12
	P.aeruginosa	79.19	158.88	318.74	635.02
	S.cerevisiae	74.5	150.2	299.78	601.05
	S.pombe	77.61	157.38	312.73	626.14

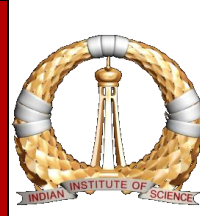
GMaccS Prototype: Performance and Scalability Analysis for Human Genome Benchmarks



Human Short Read Data Mapping					
	Read Set	Num_Reads	Num_Alignments	Time Taken P1 (s)	Time Taken P2 : 24 GPUs (s)
Child	81	146929886	2165293050	2075.570834	101.6067678
	82	143848074	2087047754	2025.539461	99.15701824
	83	144871968	2103835528	1918.123761	93.90011732
	84	142831237	2075868163	1985.887691	97.21585465
Father	89	27594045	386395924	351.0189056	17.18383874
	90	28019239	391775692	375.6061451	18.38729566
	91	169777482	2404021054	2322.850715	113.7114532
	92	168278483	2257855201	2115.577249	103.6086462
	93	168484341	2298703742	2049.489686	100.3294972
Mother	94	180827103	2612370550	2276.571873	111.4463701
	95	96741850	1429061823	1310.947561	64.17489068
	96	148849161	2179760293	1838.241985	89.98933444
	97	33028205	468013191	417.7479344	20.45016972
	98	33621893	474952706	398.9403287	19.52974641



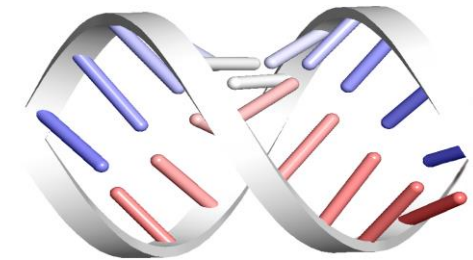
Putting it Altogether.....



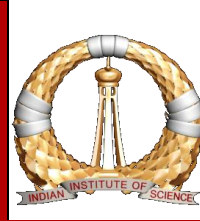
- **SRM Value Propositions**
- **Handling repeats and multi-read alignments**
- **Calculating the Delta: How insufficient are our reference genomes?**
- **Can SRM model be more domain aware?**
- **The big picture**
- **A peep beyond genomics**
- **The Road Ahead**



- **HARDWARE ACCELERATOR FOR ALIGNMENT OF SHORT READS IN SEQUENCING PLATFORMS**
- **MAPPING OF SHORT READS IN SEQUENCING PLATFORMS**
- **DATA STREAMING IN HARDWARE ACCELERATOR FOR ALIGNMENT OF SHORT READS**



Endless Life: Commercial Venture



Thank You