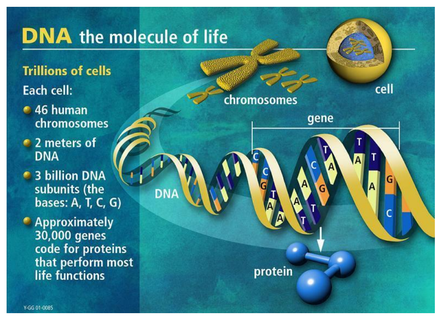
A Scalable Hardware Accelerator Model for Accurate Alignment of Short Reads from Next Generation Sequencing Platforms

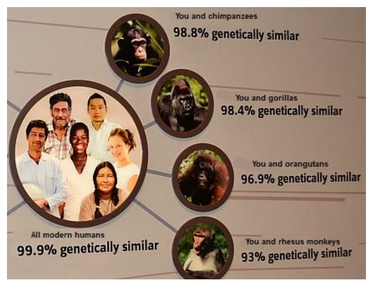
Santhi Natarajan, CADLab, CDS
Supervisors: Prof. Debnath Pal, Prof. S. K. Nandy

Genome: The Secret Code of Life

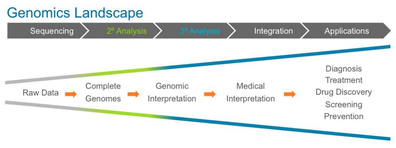


	Genome size (base pairs)	Chromosome number (*n*)
Amoeba dubia	670,000,000,000	Several hundred
Trumpet lily (*Lilium longiflorum*)	90,000,000,000	12
Mouse (*Mus musculus*)	3,454,200,000	20
Human (*Homo sapiens*)	3,200,000,000	23
Carp (*Cyprinus carpio*)	1,700,000,000	49
Chicken (*Gallus gallus*)	1,200,000,000	39
Housefly (*Musca domestica*)	900,000,000	6
Tomato (*Lycopersicon esculentum*)	655,000,000	12

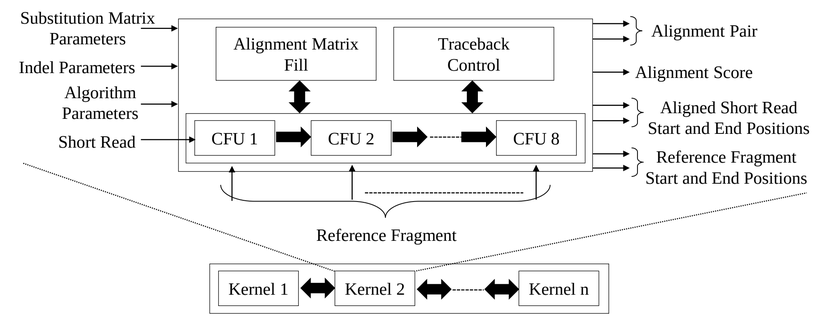
What's in a number? Genome size and chromosome number seem unrelated to complexity. The figures above are for haploid genomes — most cells are diploid (2*n*), carrying two copies of each chromosome.



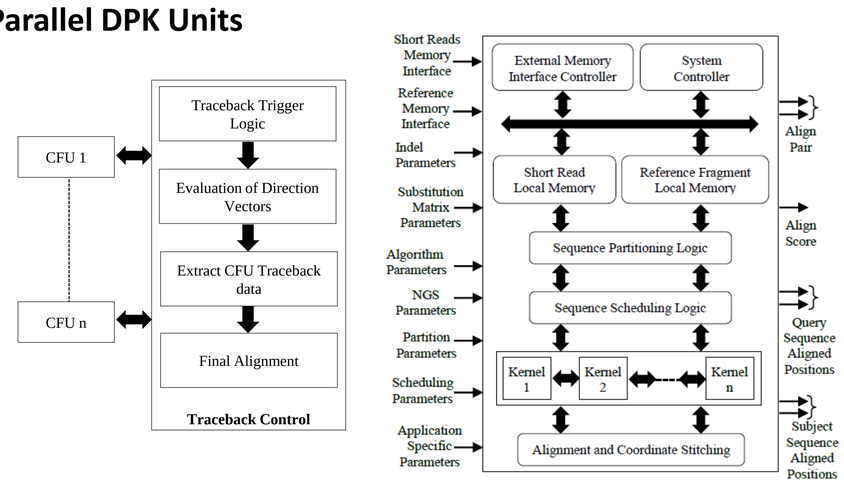
Genomics Landscape



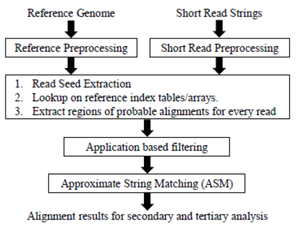
Dynamic Programming Kernel (DPK)

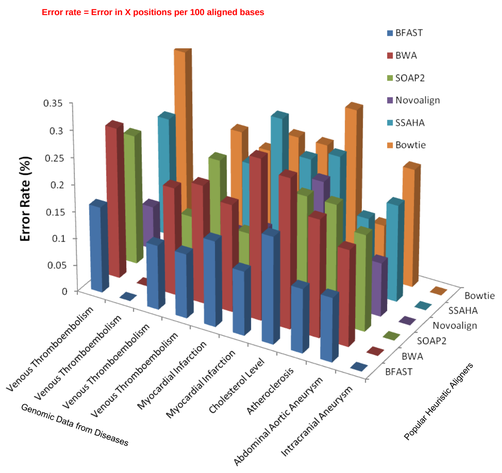


Parallel DPK Units



Short Read Mapping (SRM) Workflow and Heuristic Rates





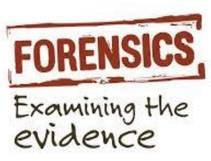
Applications

OEMs

Large Research Groups

Medical research groups and hospitals

Small Research Groups and Diagnostic Labs



Individuals like clinicians

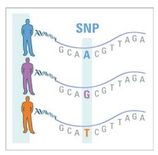
Problem Statement

Accuracy with Precision

Cost-effective

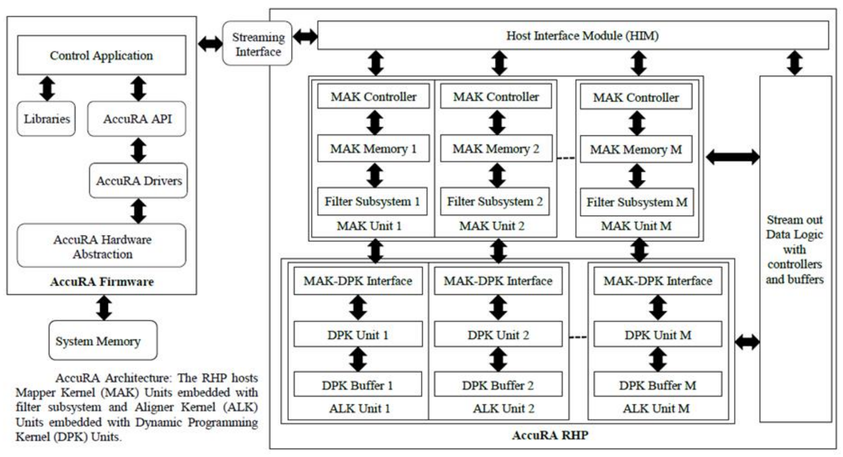
Problem Statement:

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?



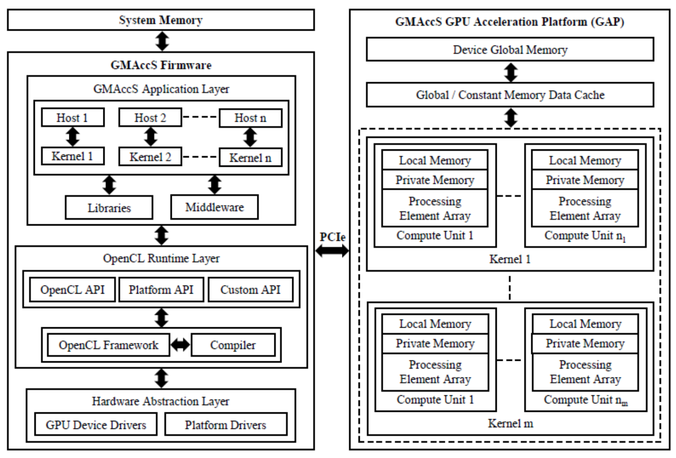
Resolution

AccuRA: SRM on Scalable Reconfigurable Accelerators



AccuRA Architecture: The RHP hosts Mapper Kernel (MAK) Units embedded with filter subsystem and Aligner Kernel (ALK) Units embedded with Dynamic Programming Kernel (DPK) Units.

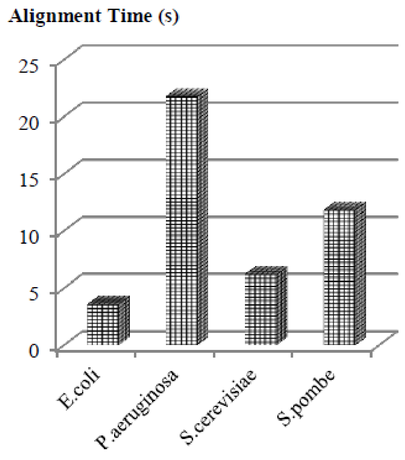
GMAccS: A Scalable GPGPU Model for Accurate Alignment of Short Reads



Dynamic Programming Recursive Model

$$a = a_1 \dots a_M \quad (1a)$$
$$b = b_1 \dots b_N \quad (1b)$$
$$\sum = \{y_1, y_2, \dots y_t\} \quad (1c)$$
$$a' = a'_1 a'_2 \dots a'_l \quad (2a)$$
$$b' = b'_1 b'_2 \dots b'_l \quad (2b)$$
$$\sum' = \sum \cup \{-\} \quad (2c)$$
$$Max\{M, N\} \le l \le M + N \quad (3)$$
$$\gamma(g) = -d - (g-1)e \quad (4)$$
$$D(i,j) = max \begin{cases} D(i\text{-}1)(j\text{-}1) + s(x_i, y_j) \\ I(i\text{-}1)(j\text{-}1) + s(x_i, y_j) \end{cases} \quad (5a)$$
$$D(i,j) = max \begin{cases} D(i, j\text{-}1) - d \\ I(i, j\text{-}1) - e \\ D(i\text{-}1, j) - d \\ I(i\text{-}1, j) - e \end{cases} \quad (5b)$$

AccuRA Performance for Small Genome Benchmarks

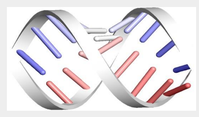


AccuRA Performance for Human Genome Benchmarks

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

Patents Filed:

1. HARDWARE ACCELERATOR FOR ALIGNMENT OF SHORT READS IN SEQUENCING PLATFORMS
2. MAPPING OF SHORT READS IN SEQUENCING PLATFORMS
3. DATA STREAMING IN HARDWARE ACCELERATOR FOR ALIGNMENT OF SHORT READS



Endless Life: Commercial Venture

References:

1. S. Batzoglou, "The many faces of sequence alignment," Brief. Bioinform. 6 (1), pp. 6-22, 2005.
2. M. Baker, "Next-generation sequencing: adjusting to data overload," Nature Methods, vol. 7 no. 7, pp. 495 - 499, July 2010.
3. H. Li, R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," Bioinformatics, vol. 25 no. 14, pp. 1754-1760, May 2009.
4. S. Schbath, V. Martin, M. Zytnicki, J. Fayolle, V. Loux, JF. Gibrat, "Mapping Reads on a Genomic Sequence: an Algorithmic Overwiew and a Practical Comparative Analysis," J. Comput. Biol. 19(6), pp. 796-813, June 2012.
5. T.F. Smith, M.S. Waterman, "Identification of common molecular subsequences," J. Mol. Biol. 147, pp. 195-197, 1981.

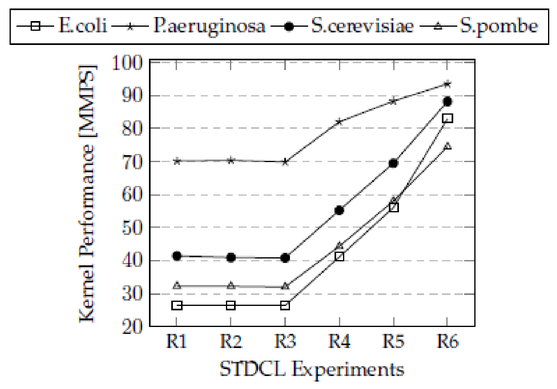
AccuRA Scalability in terms of number of CFU units

No. of Units, N	Batch Size, B	Time taken per batch (sec)	No. of Alignments	No. of Batches	Total Time (sec)
8	512	0.00062791	512	9896789.799	6214.23998
16	1024	0.00062791	1024	4948394.899	3107.11999
32	2048	0.00062791	2048	2474197.45	1553.55999
48	3072	0.00062791	3072	1649464.966	1035.70666

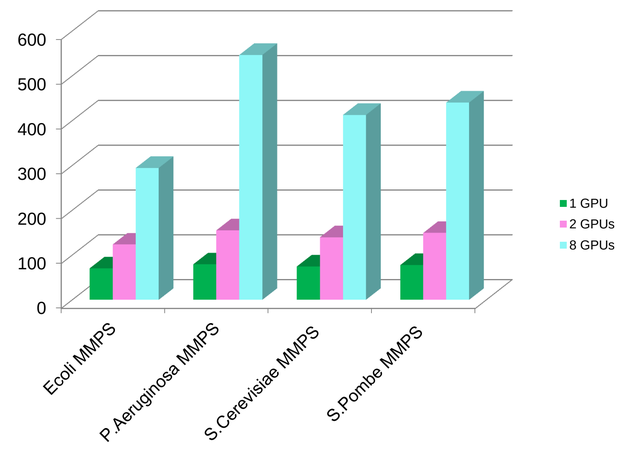
Scalability Versus Performance of MAK-DPK Units

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

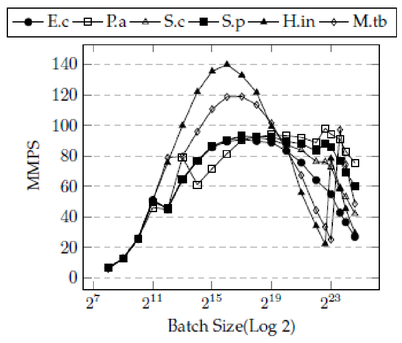
GMAccS Performance for Small Genome Benchmarks on Single GPU



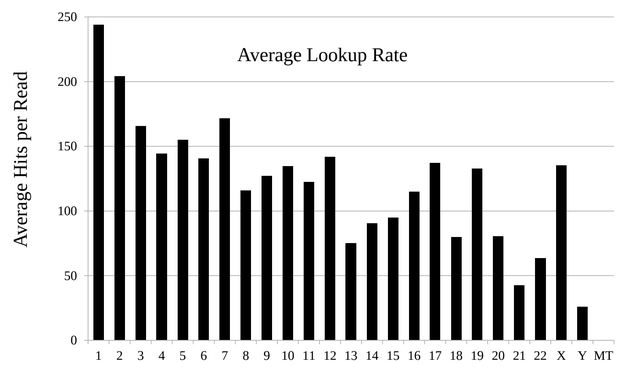
Scalability of GMAccS on SahasraT (Cray XC40)



GMAccS Kernel Perofrmance Optimization Efforts



Average Hit Rate per read in Human Genome



GMAccS Performance and Scalability with Human Genome Benchmarks on SahasraT (Cray XC40)

SRR Read Sets	P1 time (s)	P2 time 24 GPUs (s)
SRR1559289	351.02	17.18
SRR1559290	375.61	18.39
SRR1559291	2322.85	113.71
SRR1559292	2115.58	103.61
SRR1559293	2049.49	100.33
SRR1559294	2276.57	111.45
SRR1559295	1310.95	64.17
SRR1559296	1838.24	89.99
SRR1559297	417.75	20.45
SRR1559298	398.94	19.53
SRR1559281	2075.57	101.61
SRR1559282	2025.54	99.16
SRR1559283	1918.12	93.9
SRR1559284	1985.89	97.22