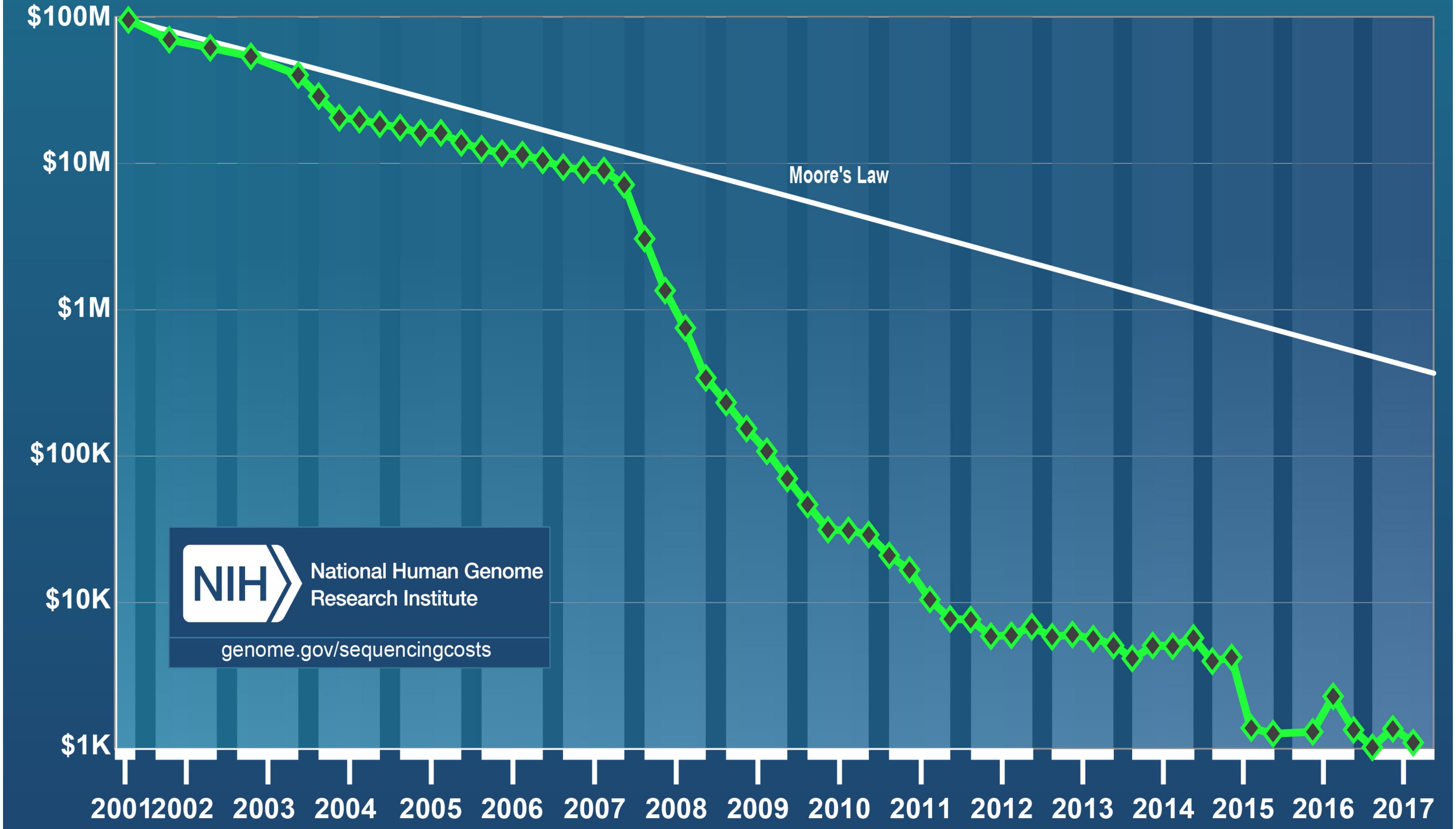
Hardware-Enabled Biology

AACBB Workshop
February 16, 2019

Bill Dally
Chief Scientist and SVP of Research, NVIDIA Corporation
Professor (Research), Stanford University

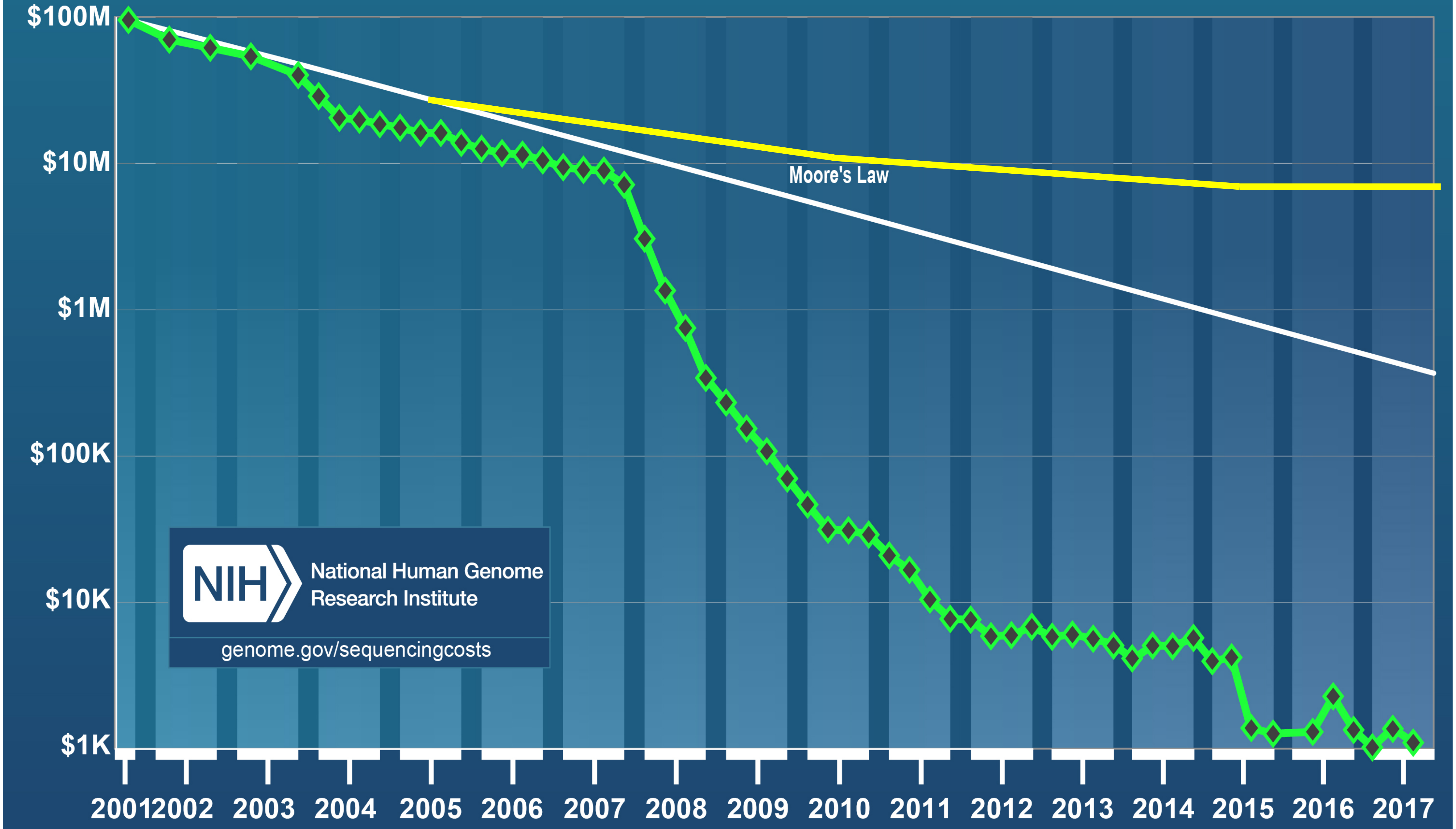
Sequence Data is Growing Exponentially

Computation Isn’t



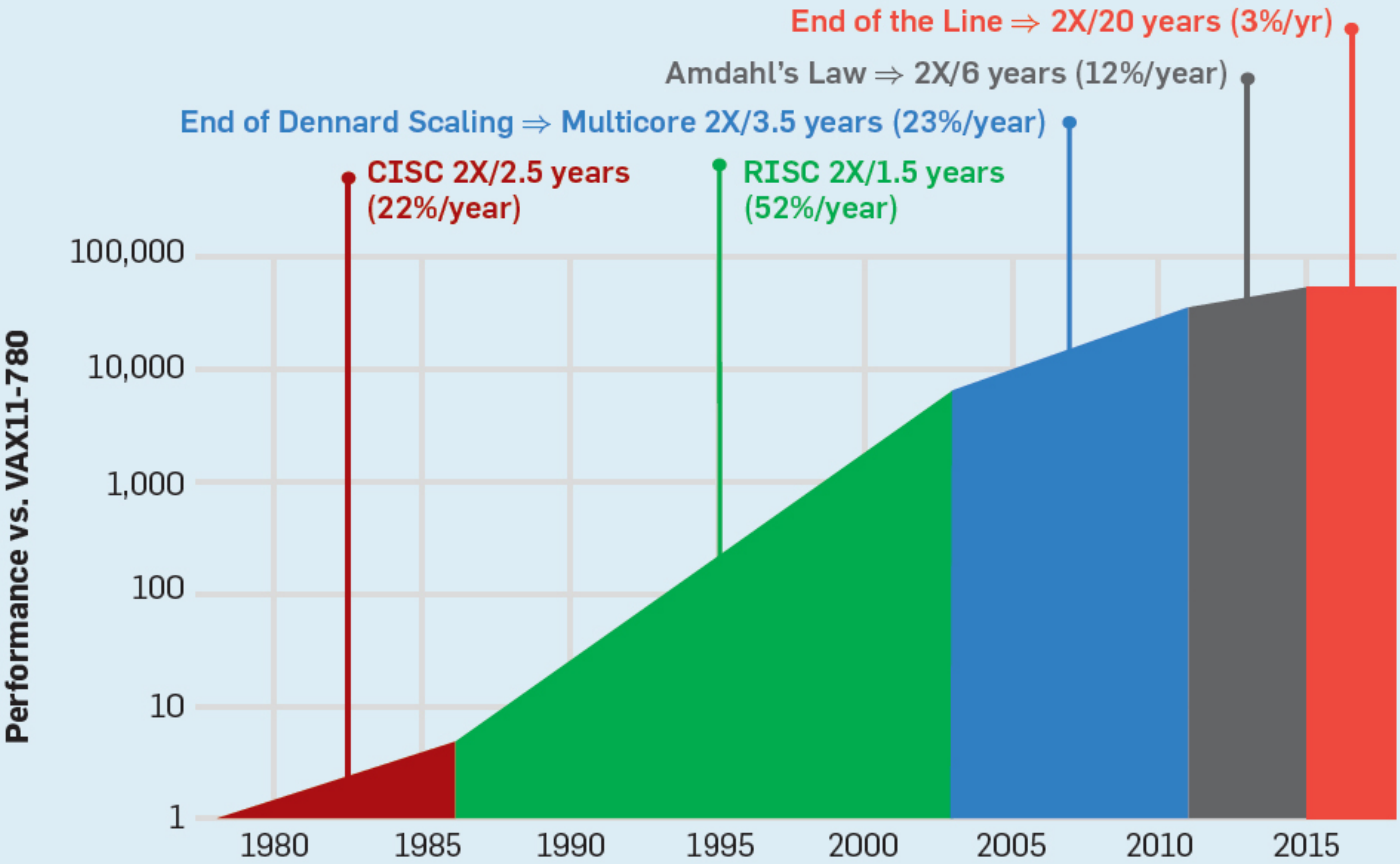
NIH National Human Genome Research Institute

genome.gov/sequencingcosts



NIH National Human Genome Research Institute

genome.gov/sequencingcosts



John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018

Cost To

- Sequence a human genome - $1k today (short reads, 30x coverage)
  - $3k for long reads (10x coverage)
  - $100 soon
- Perform reference-based assembly of it - $15 (short reads)
- Perform de-novo assembly of it - $10k (long reads)

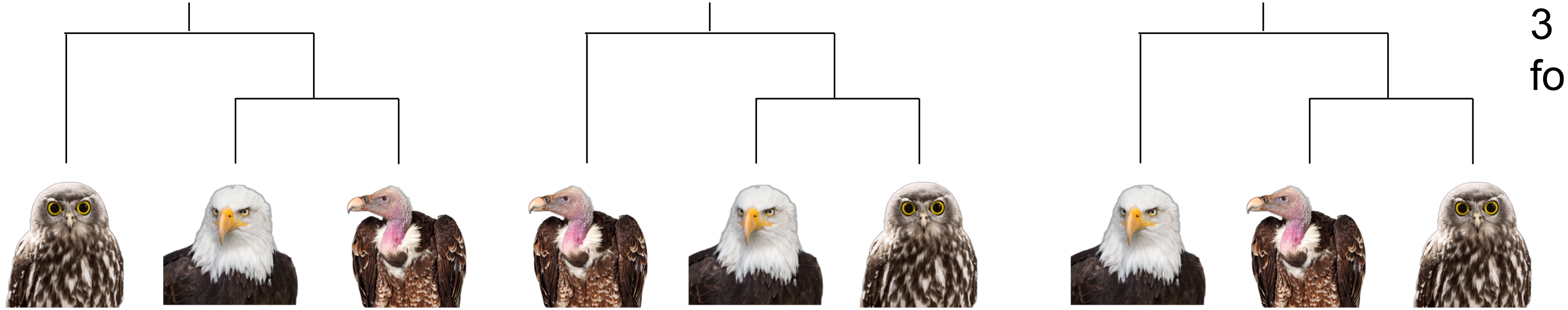
Computation is a growing fraction of genomics cost (scaling slower than sequencing)

Computation cost already dominates some tasks (e.g., de-novo assembly).

https://hpcbio.illinois.edu/services-and-fees

Many Demanding Computational Problems

Phylogenomics: Inferring phylogenetic relationships from genomes



3 possible trees
for 3 bird species

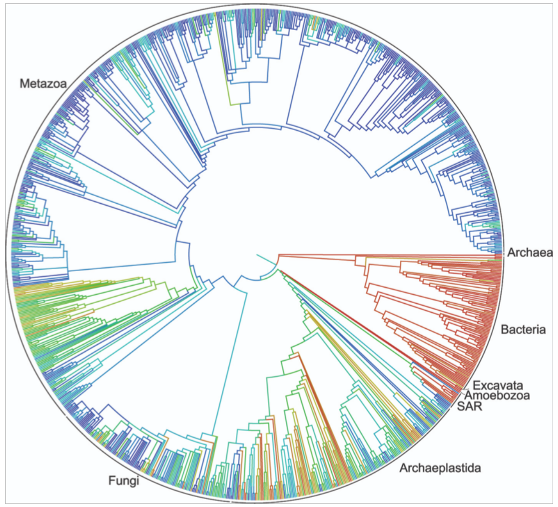
# species	# rooted trees
3	3
6	945
9	2.0×10^{6}
30	4.9×10^{38}
2.3×10^{6}	???

270 CPU years required for solving the topology of 48 birds [Jarvis et al, Science 2014]

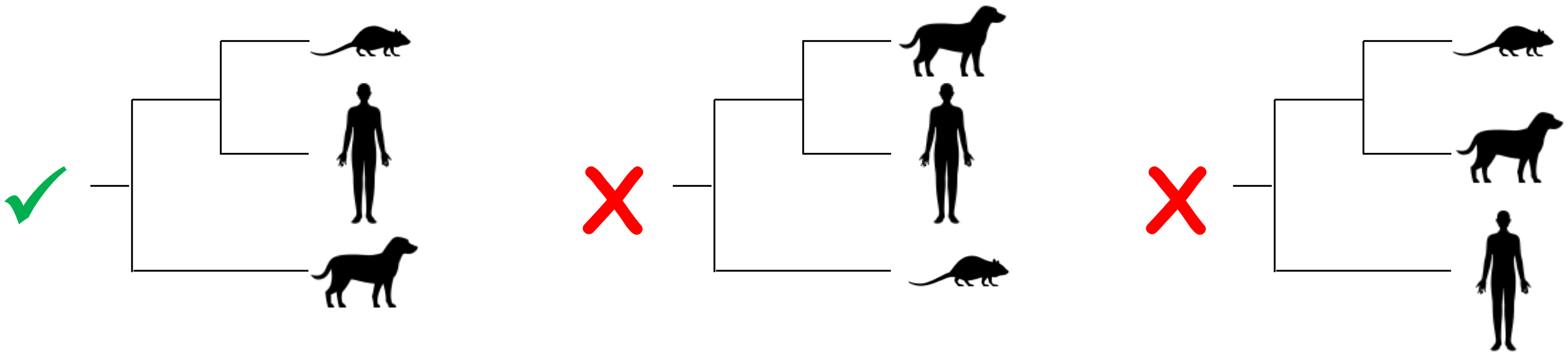
Open questions

1. What is the tree of life for ~2.3 million extant species?
2. What is the best method to infer this tree from genomes?

Extant Tree of life has 2.3 million species!
OpenTreeOfLife.org



Phylogenomics: Inferring phylogenetic relationships from genomes



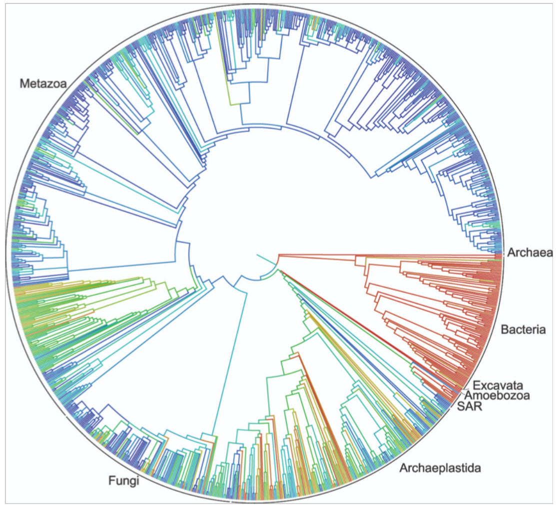
This topology was “resolved” only in 2007 [Cannarozzi et al] with the help genomic data

# species	# rooted trees
3	3
6	945
9	2.0×10^6
30	4.9×10^{38}
2.3×10^6	???

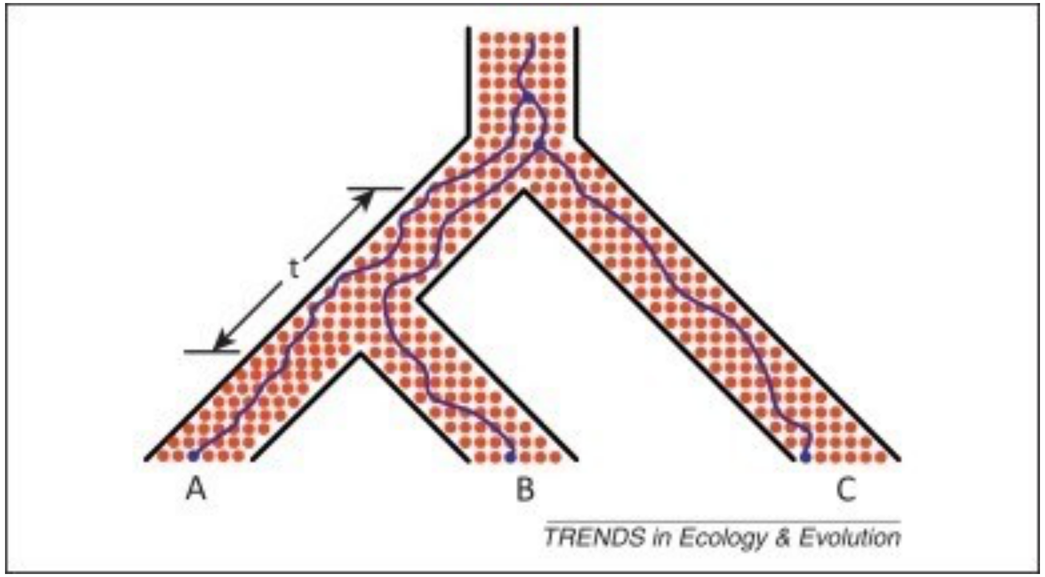
270 CPU years required for solving the topology of 48 birds [Jarvis et al, Science 2014]

Open questions

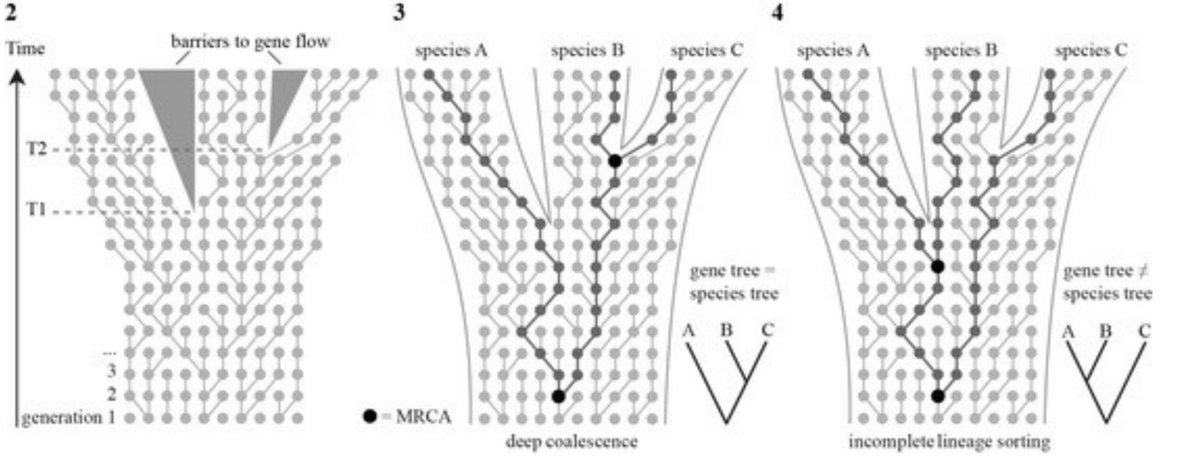
1. What is the tree of life for ~2.3 million extant species?
2. What is the best method to infer this tree from genomes?



Not Really a Tree – Incomplete Lineage Sorting



Luak Nakhleh, Trends in Ecology and Evolution 2003



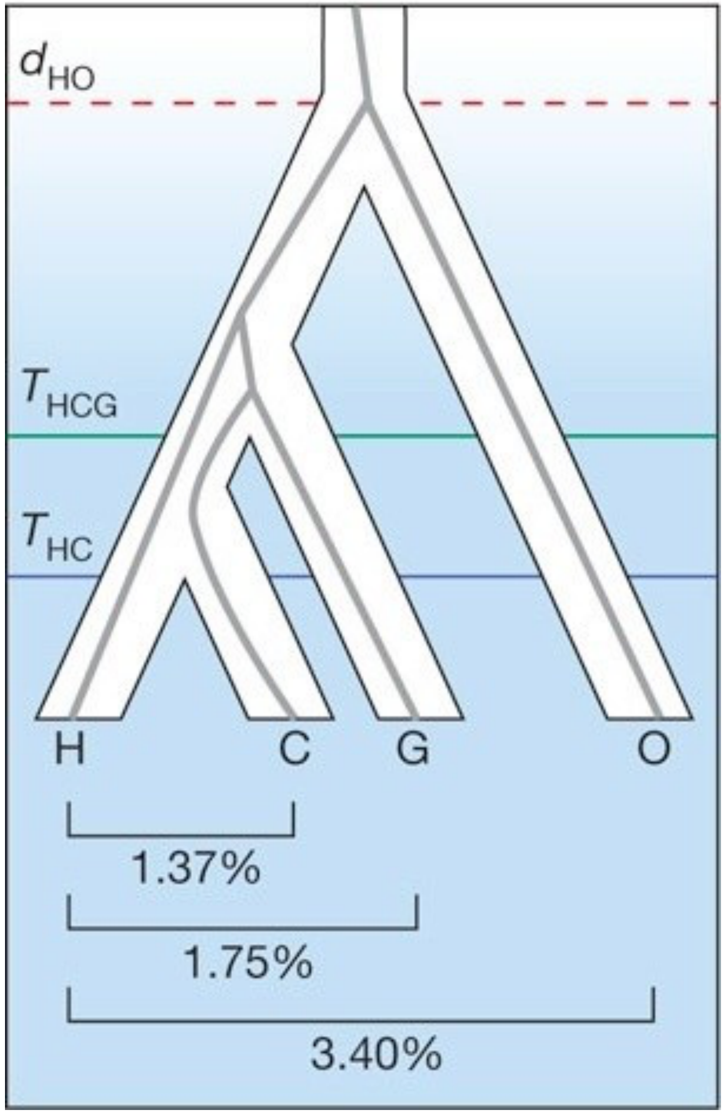
Frederik Leliaert, European Journal of Phycology, 2014

Deep coalescence

Have to go far back in time for genes to “coalesce”

Gene can split before speciation

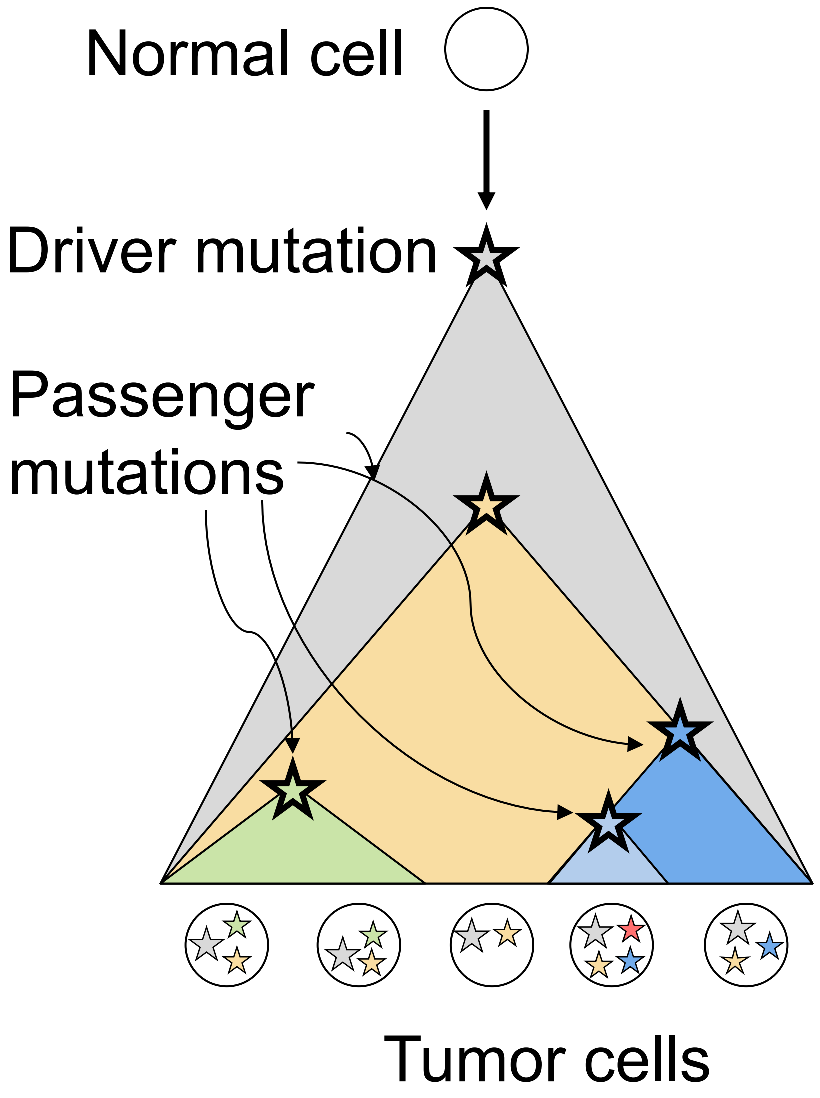
Human-Chip-Gorilla-Orangutan



Gene Genealogy different than Species Phylogeny for 25% of genome

https://www.dailykos.com/stories/2016/6/10/1534820/-Incomplete-Lineage-Sorting-and-a-Non-Tree-View-of-Life

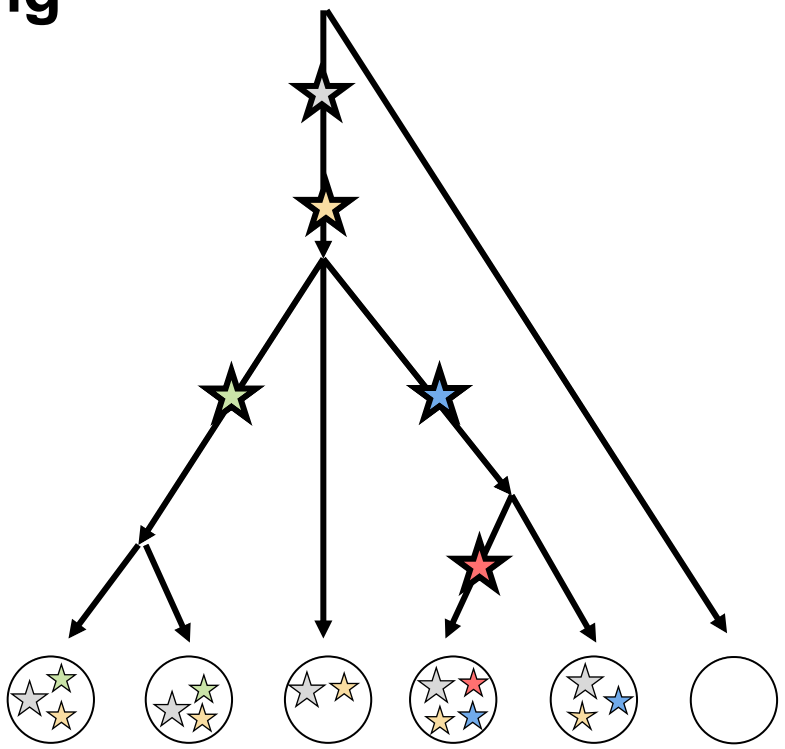
Identifying driver mutations in cancer



Single-cell sequencing

★	1	1	1	1	1
★	1	1	1	1	1
★	1	1	0	0	0
★	0	0	0	1	1
★	0	0	0	1	0

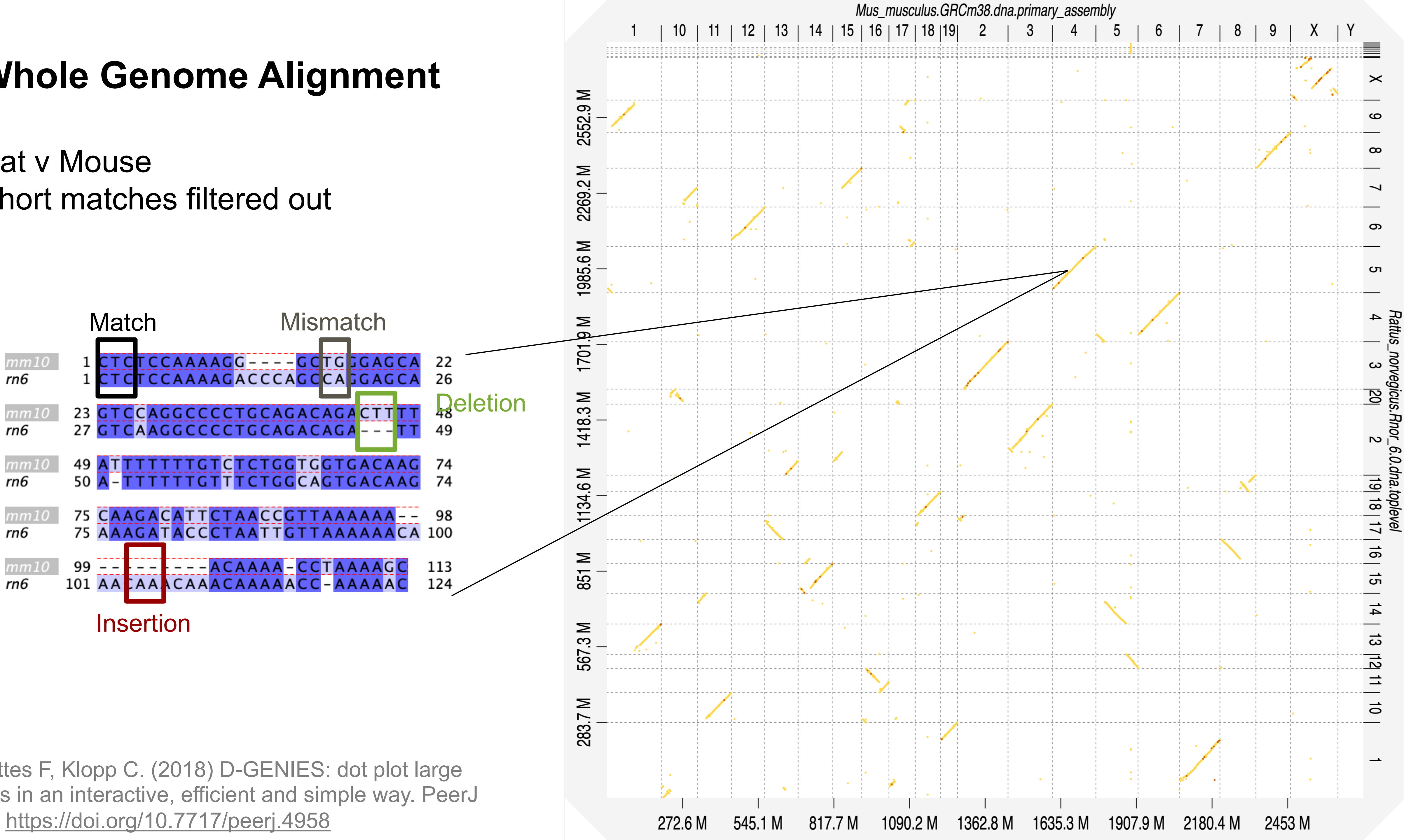
Tumor phylogeny



Inspired from [Jahn et al, Genome Biol. 2016]

Whole Genome Alignment

Rat v Mouse
Short matches filtered out

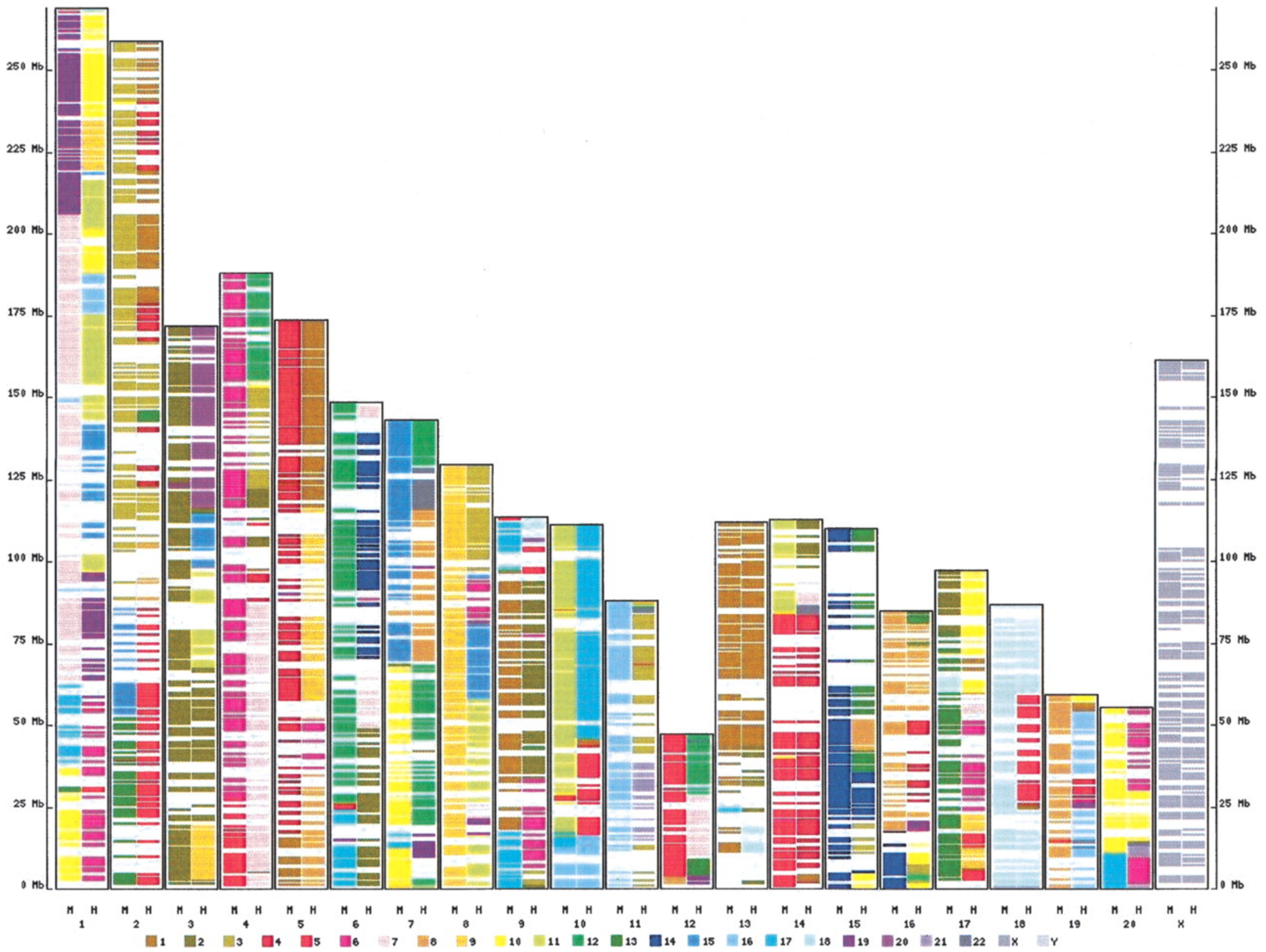


```
mm10   1 CTCTCCAAAAGG----GCTGGGAGCA  22
rn6    1 CTCTCCAAAAGACCCAGCCAGGAGCA  26

mm10  23 GTCCAGGCCCCTGCAGACAGACTTTT  48
rn6   27 GTCAAGGCCCCTGCAGACAGA---TT  49

mm10  49 ATTTTTTTGTCTCTGGTGGTGACAAG  74
rn6   50 A-TTTTTTGTTTCTGGCAGTGACAAG  74

mm10  75 CAAGACATTCTAACCGTTAAAAAA--  98
rn6   75 AAAGATACCCTAATTGTTAAAAAACA 100

mm10  99 ---------ACAAAA-CCTAAAAGC 113
rn6  101 AACAAACAAACAAAAACC-AAAAAC 124
```

Match
Mismatch
Deletion
Insertion
Mus_musculus.GRCm38.dna.primary_assembly
1 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | X | Y
272.6 M
545.1 M
817.7 M
1090.2 M
1362.8 M
1635.3 M
1907.9 M
2180.4 M
2453 M
2552.9 M
2269.2 M
1985.6 M
1701.9 M
1418.3 M
1134.6 M
851 M
567.3 M
283.7 M
Rattus_norvegicus.Rnor_6.0.dna.toplevel
X | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 20 | 2 | 19 | 18 | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 1

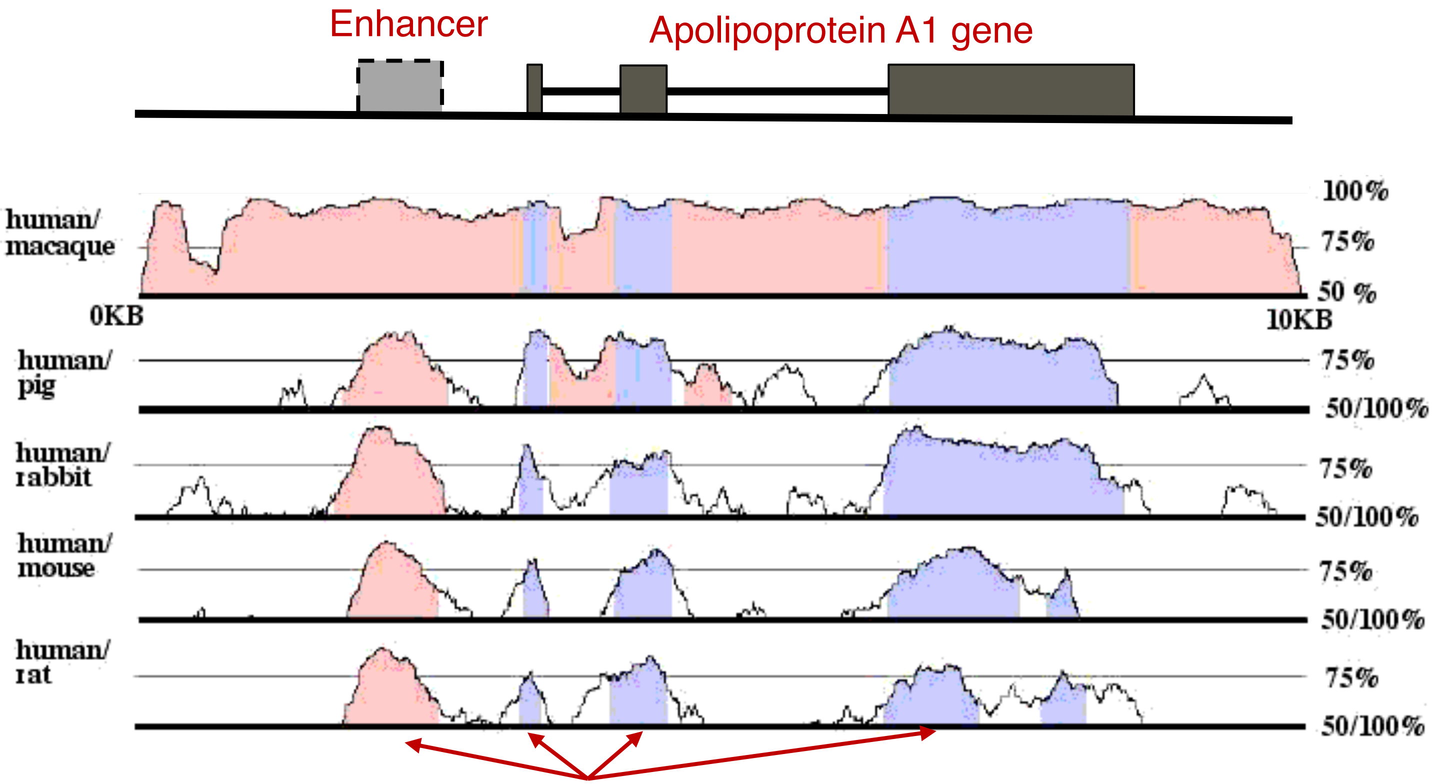
Cabanettes F, Klopp C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958 https://doi.org/10.7717/peerj.4958

Exon-based map of conserved synteny between the rat, human, and mouse genomes.



Michael Brudno et al. Genome Res. 2004;14:685-692

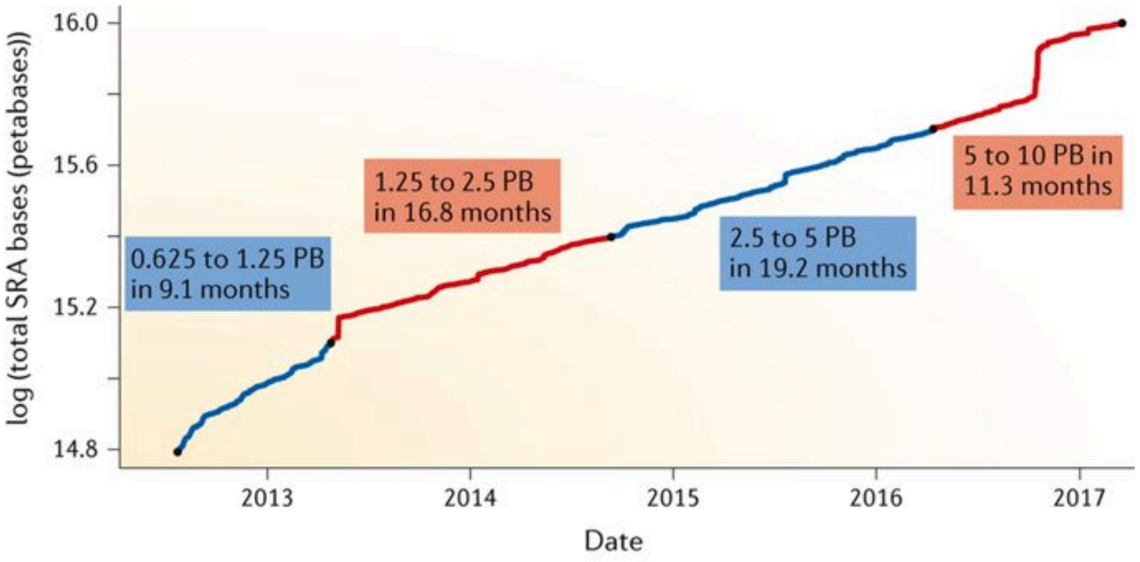
Whole Genome Alignment



Regions with sequence conservation

(Mayor et al. , 2000)

Memory and storage

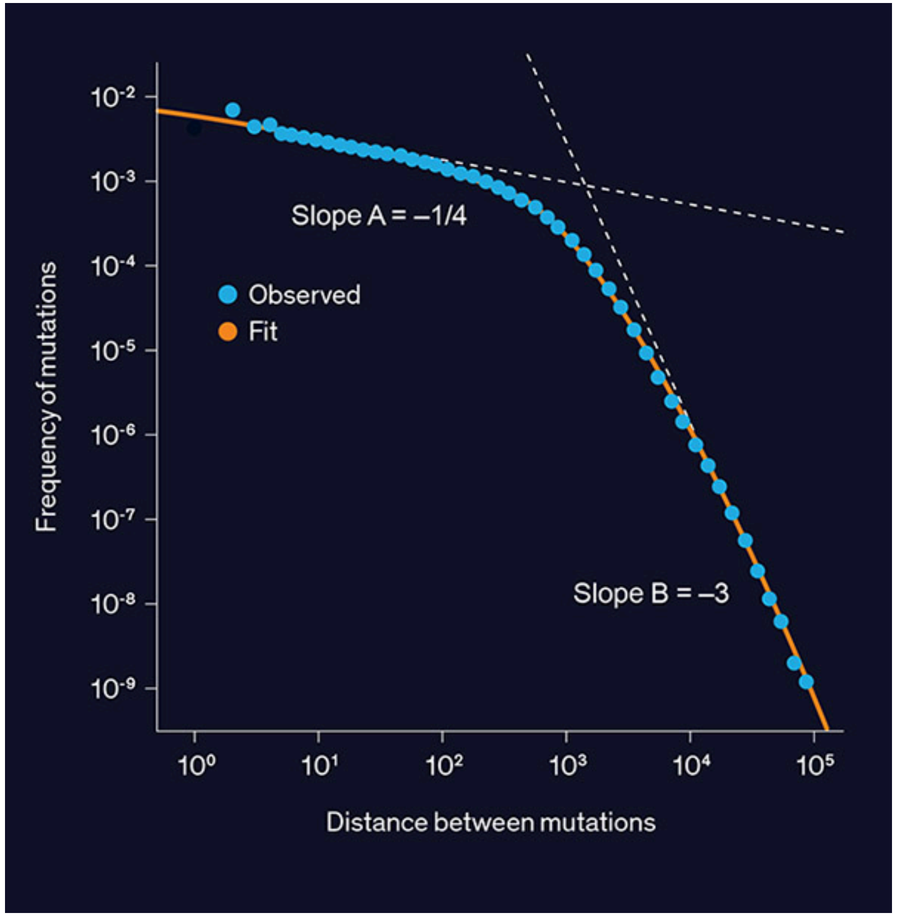


Nature Reviews | Genetics

- Genomic data doubling roughly every 14 months since 2013
- Exabyte of genomic data per year from 2025, surpassing Youtube and Astronomy
- **Open questions**

1. How and where to store genomic data?
2. How to enable secure data sharing?
3. How to enable exabyte scale processing of genomic data?

Genome compression



"Double power law" distribution => compressibility of variation data

[Pavlichin et al, Bioinformatics 2013]

- In general, genomic data is highly compressible

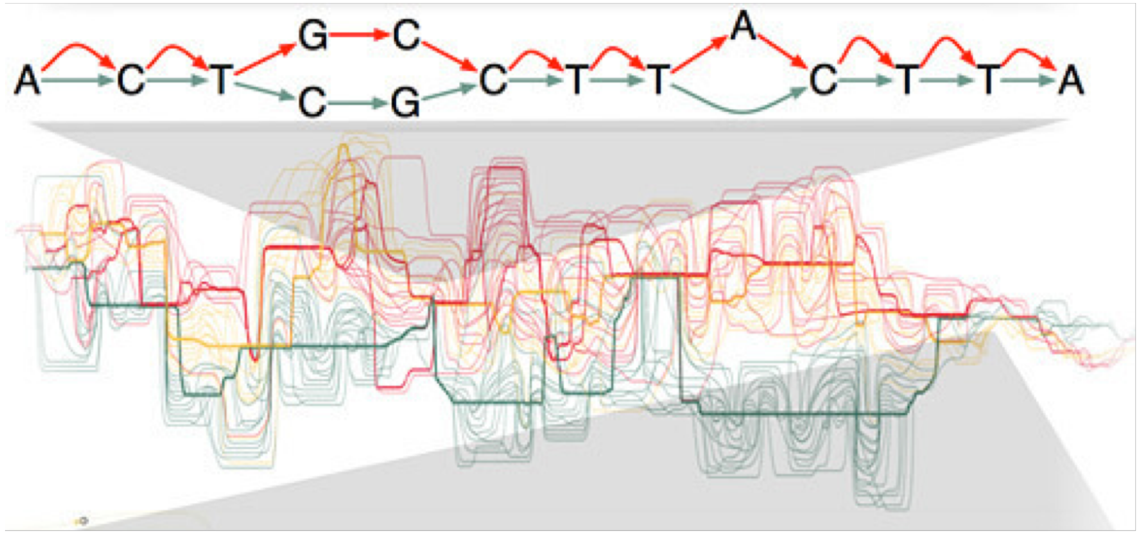
- **Open questions:**

1. How to enable lossless compression with a high compression rate?
2. How to enable lossy compression without affecting informatics?
3. How to enable fast compute on compressed data?

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

MPEG

Genome graphs



- Graphs as a way to represent common human genomic variation
- More representative - minimizes bias to a single reference
- More informative than a single "profile"
- **Open questions:**

1. How to build a genome graph?
2. How to align sequencing reads to a genome graph accurately?

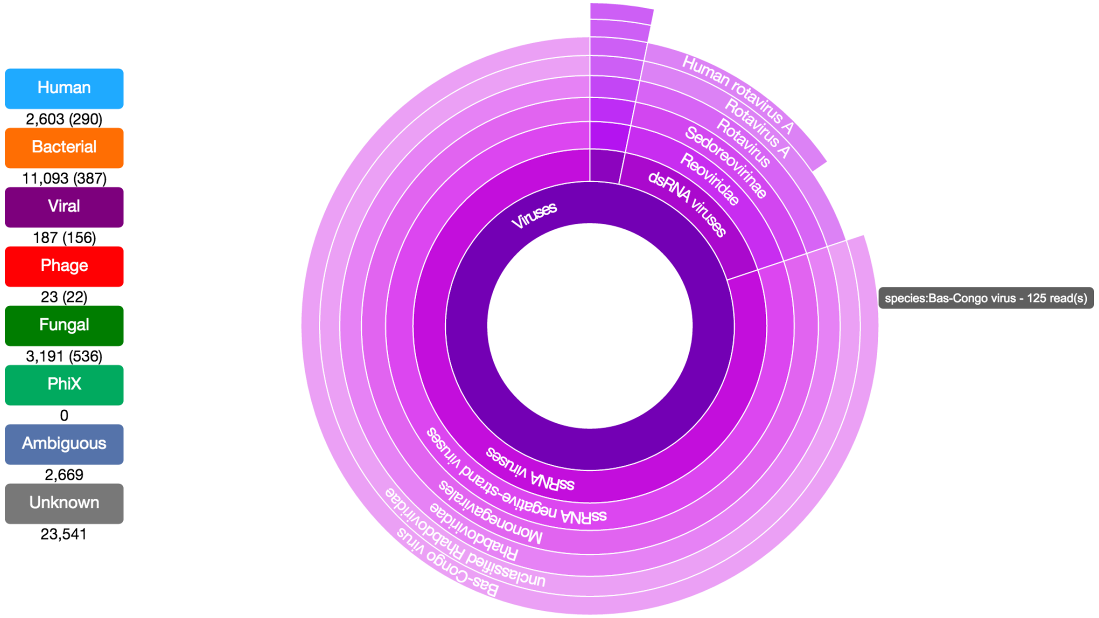
Metagenomics and liquid biopsy

- Sequence reads from a environment sample (human gut, soil etc)

- Build a taxonomic profile of species (bacteria, virus, fungal, human, etc.) from reads

- Applications

1. Infectious disease (Karius Inc.)
2. Discover new natural products (Radiant Genomics)
3. Microbiome analysis and therapeutics (MicroBiome Therapeutics)

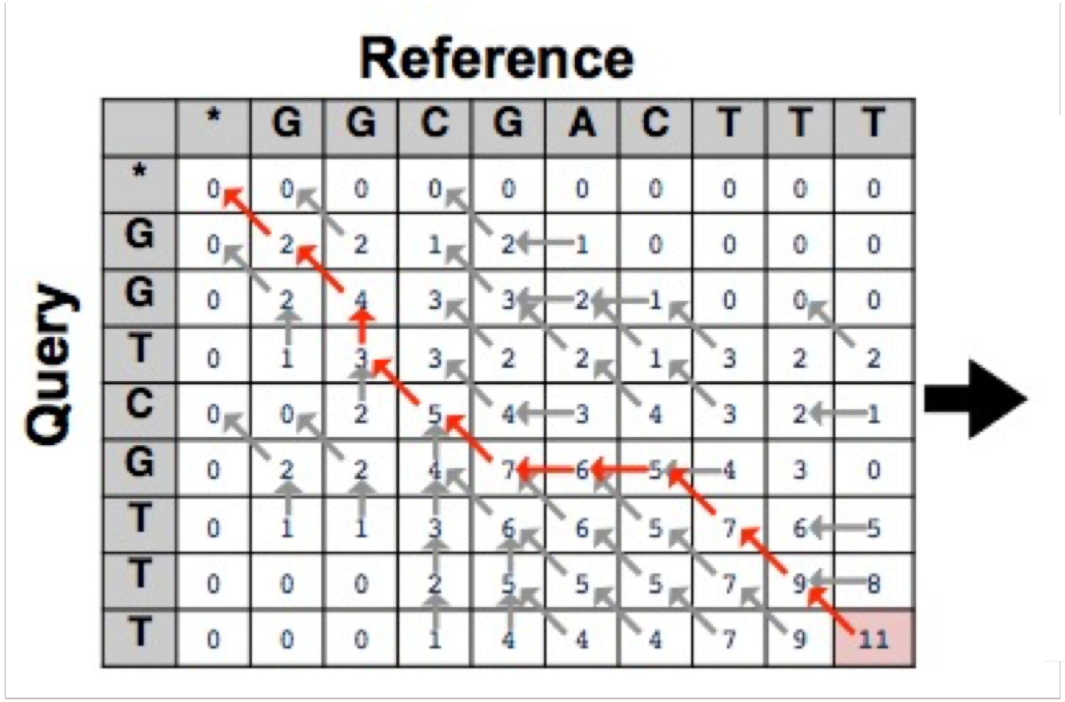


[taxonomer.iobio.io]

Specialized Operations

Orders of Magnitude Speedup & Efficiency

Specialized Operations



	*	G	G	C	G	A	C	T	T	T
*	0	0	0	0	0	0	0	0	0	0
G	0	2	2	1	2	1	0	0	0	0
G	0	2	4	3	3	2	1	0	0	0
T	0	1	3	3	2	2	1	3	2	2
C	0	0	2	5	4	3	4	3	2	1
G	0	2	2	4	7	6	5	4	3	0
T	0	1	1	3	6	6	5	7	6	5
T	0	0	0	2	5	5	5	7	9	8
T	0	0	0	1	4	4	4	7	9	11

$$
\begin{aligned}
I(i,j) &= \max\{H(i,j-1)-o,\ I(i,j-1)-e\} \\
D(i,j) &= max\{H(i-1,j)-o,\ D(i-1,j)-e\} \\
H(i,j) &= max \begin{cases} 0 \\ I(i,j) \\ D(i,j) \\ H(i-1,j-1)+W(r_i,q_j) \end{cases}
\end{aligned}
$$

Dynamic programming for gene sequence alignment (Smith-Waterman)

On 14nm CPU
35 ALU ops, 15 load/store
37 cycles
81nJ

On 40nm Special Unit
1 cycle (37x speedup)
3.1pJ (26,000x efficiency)
300fJ for logic (remainder is memory)

Accelerator Design is Guided by Cost

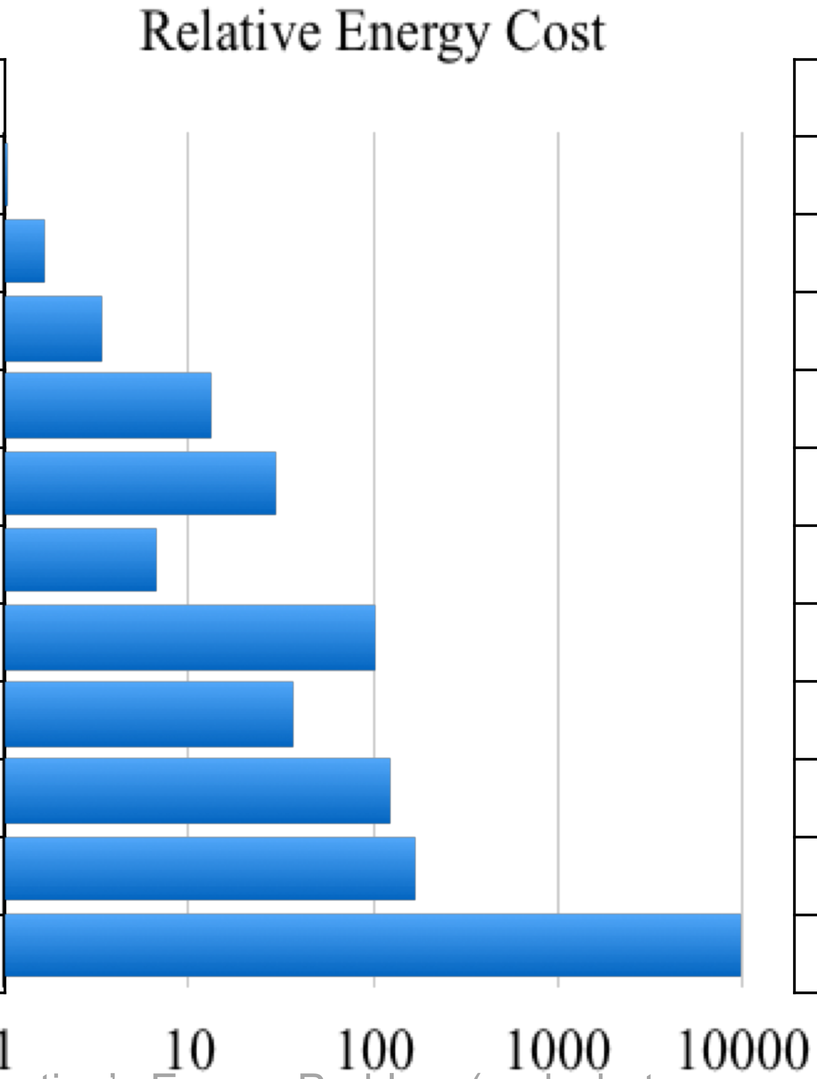
Arithmetic is Free
(particularly low-precision)

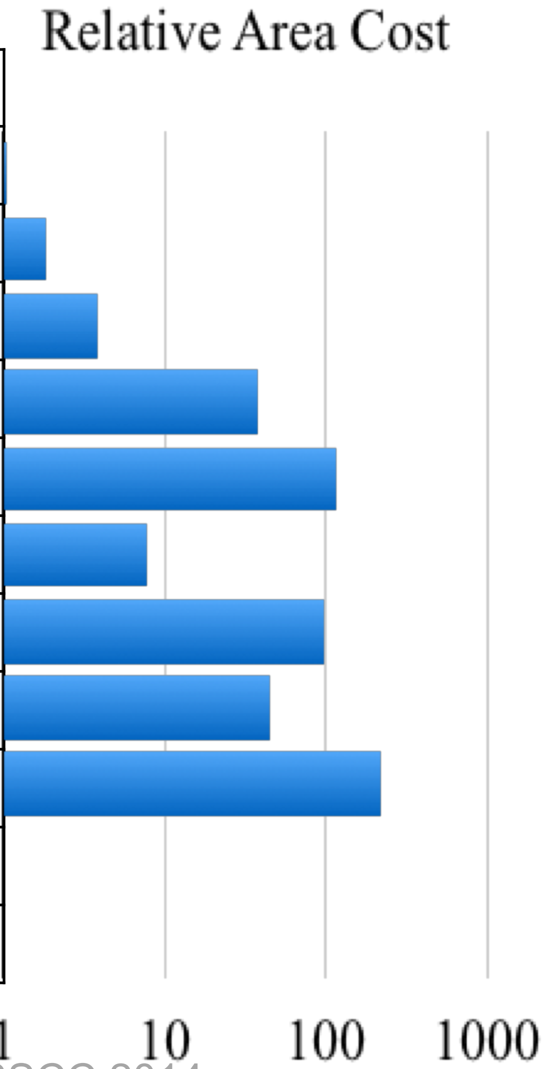
Memory is expensive

Communication is prohibitively expensive

Need to Understand Cost of Operations And Communication

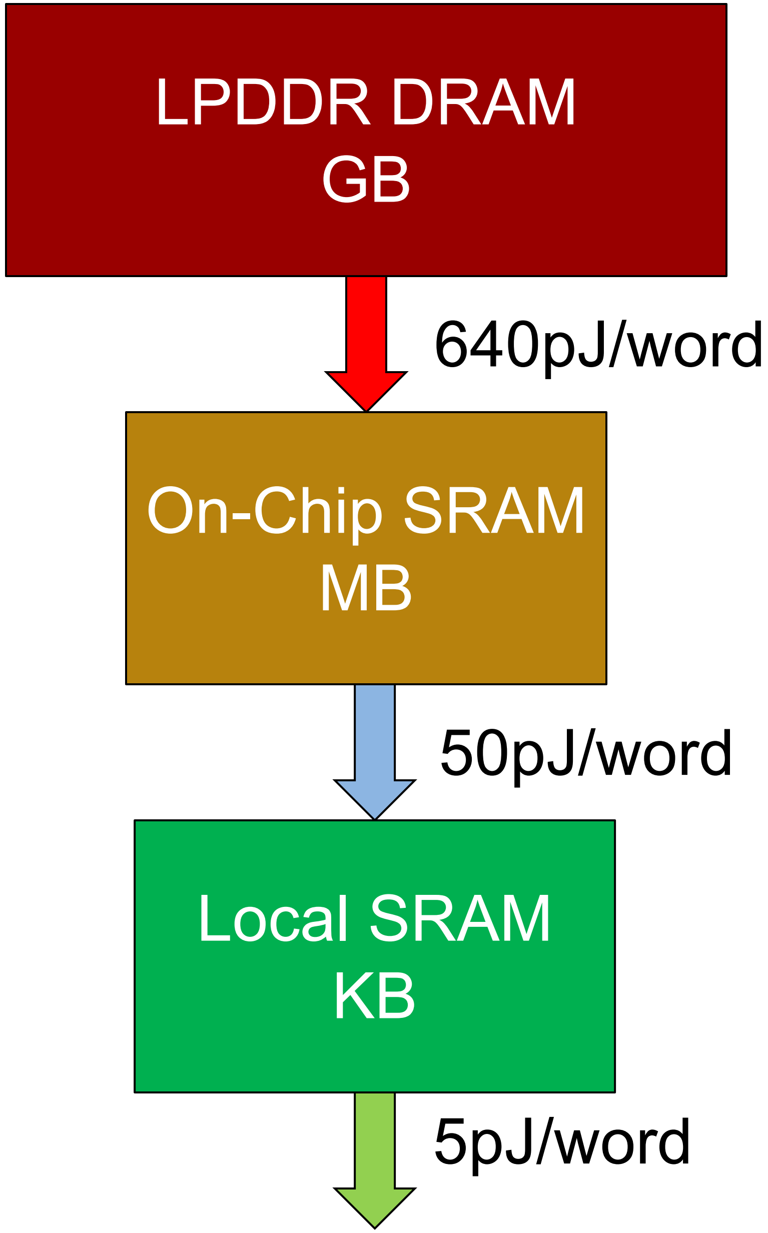
Operation:	Energy (pJ)	Area (μm^2)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A





Energy numbers are from Mark Horowitz “Computing’s Energy Problem (and what we can do about it)”, ISSCC 2014
Area numbers are from synthesized result using Design Compiler under TSMC 45nm tech node. FP units used DesignWare Library.

Communication is Expensive, Be Small, Be Local



LPDDR DRAM
GB

640pJ/word

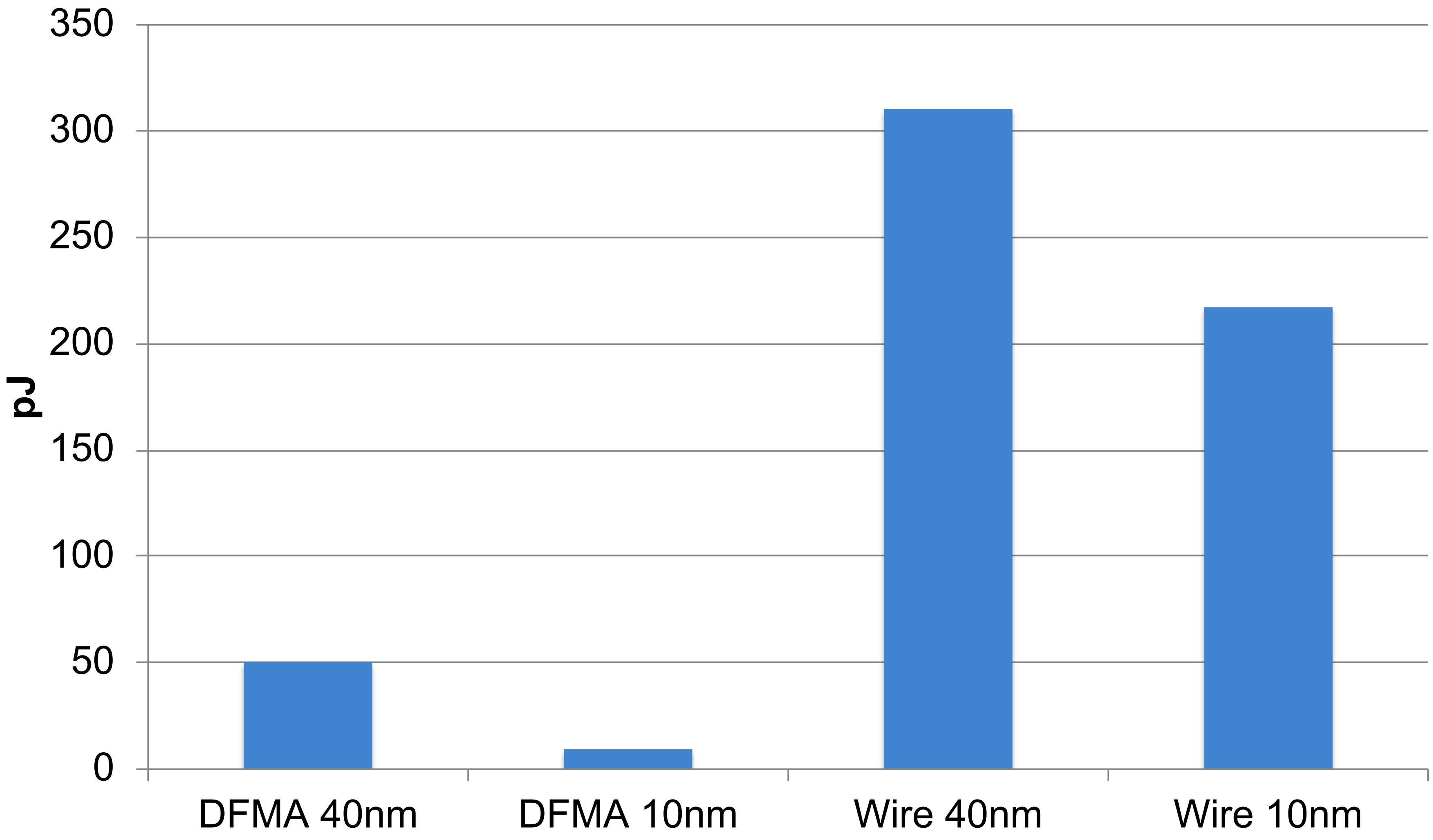
On-Chip SRAM
MB

50pJ/word

Local SRAM
KB

5pJ/word

Scaling of Communication

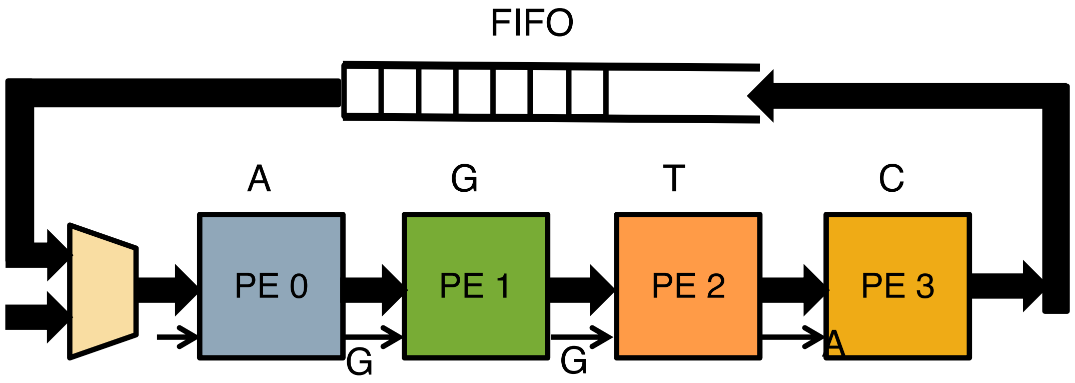


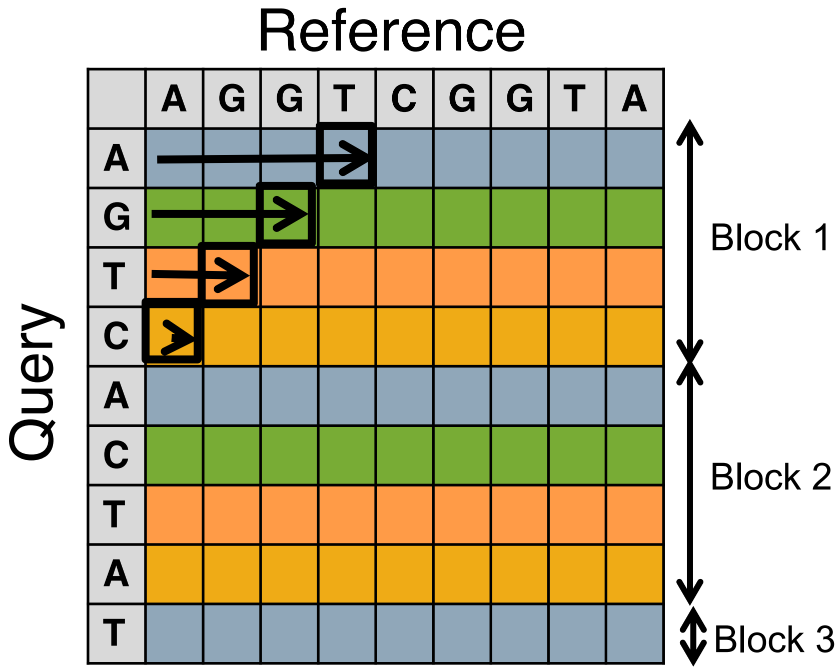
Keckler et al. Micro 2011.

Most Speedup Comes from Parallelism

Enabled by Specialization

Inner-Loop Parallelism
Systolic Array to Compute DP Matrix





Tile Size (T) = 9

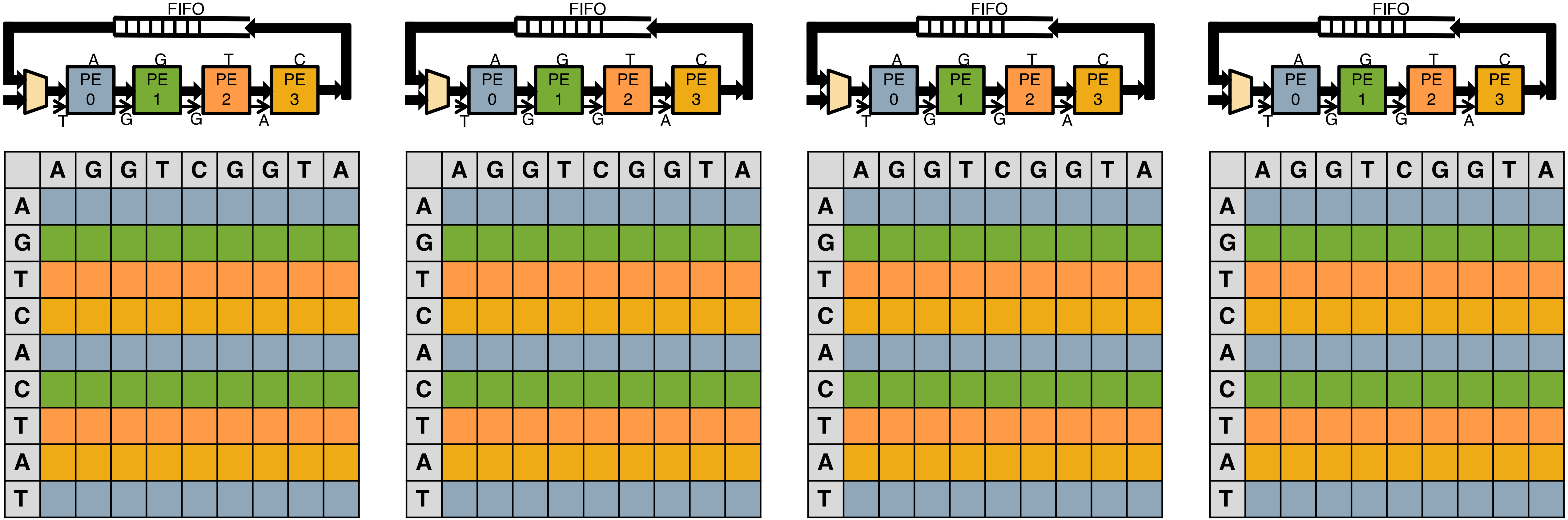
Darwin has 64 PEs per array

Communication: One-Way Nearest Neighbor

Synchronization: Lockstep

Memory: Store Traceback Pointer

Outer-Loop Parallelism
Compute Many DP Arrays at Once



Darwin has 64 arrays

Comm & Sync – Master/Slave

Memory – Distribute problems – Read back traceback

Speedup for GACT

- Specialization 37x
- Inner-Loop Parallelism 63x
- Outer-Loop Parallelism 64x
- Total ~ 150,000x

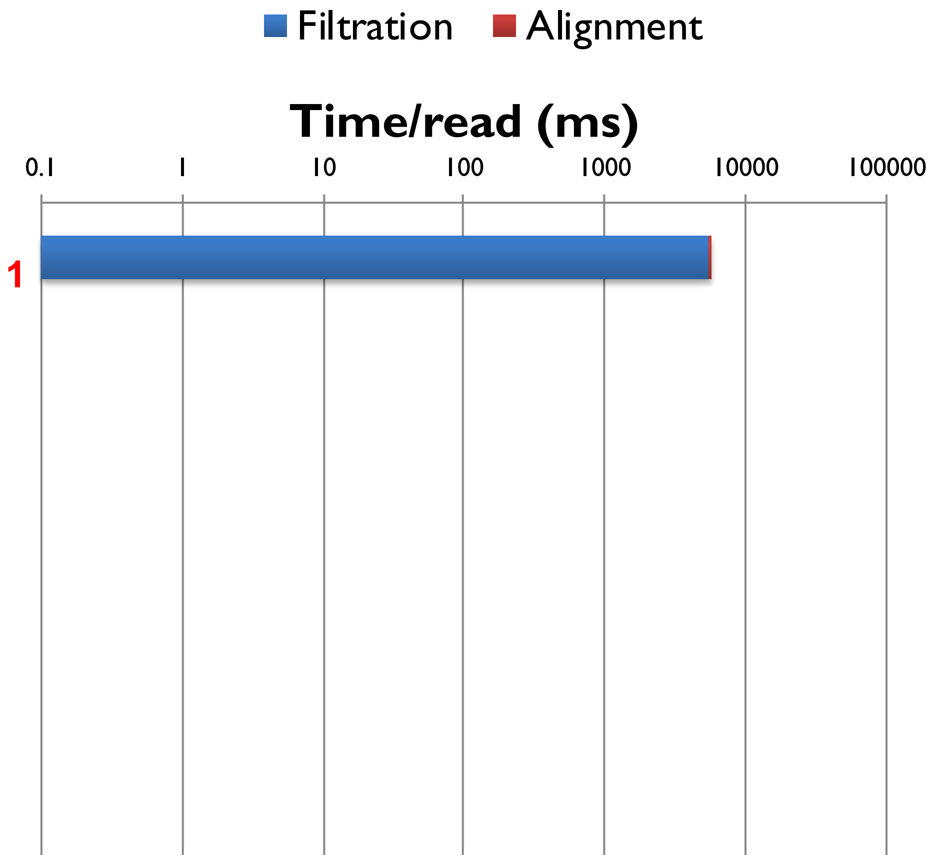
- Darwin speedup is 15,000x because filtering doesn't speed up as much as alignment.

Specialization Provides Efficiency

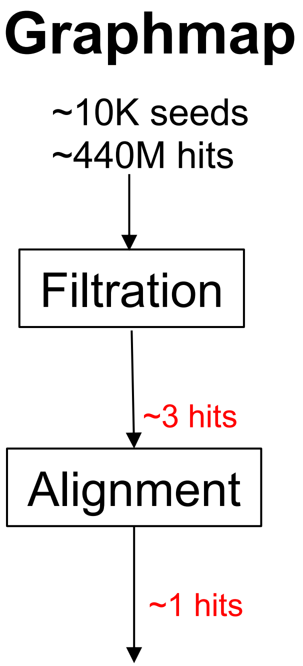
Parallelism Converts Efficiency to Speedup

The Algorithm often Has to Change

Algorithm-Architecture Co-Design for Darwin
Start with Graphmap



1. Graphmap (software)



Graphmap

~10K seeds
~440M hits

Filtration

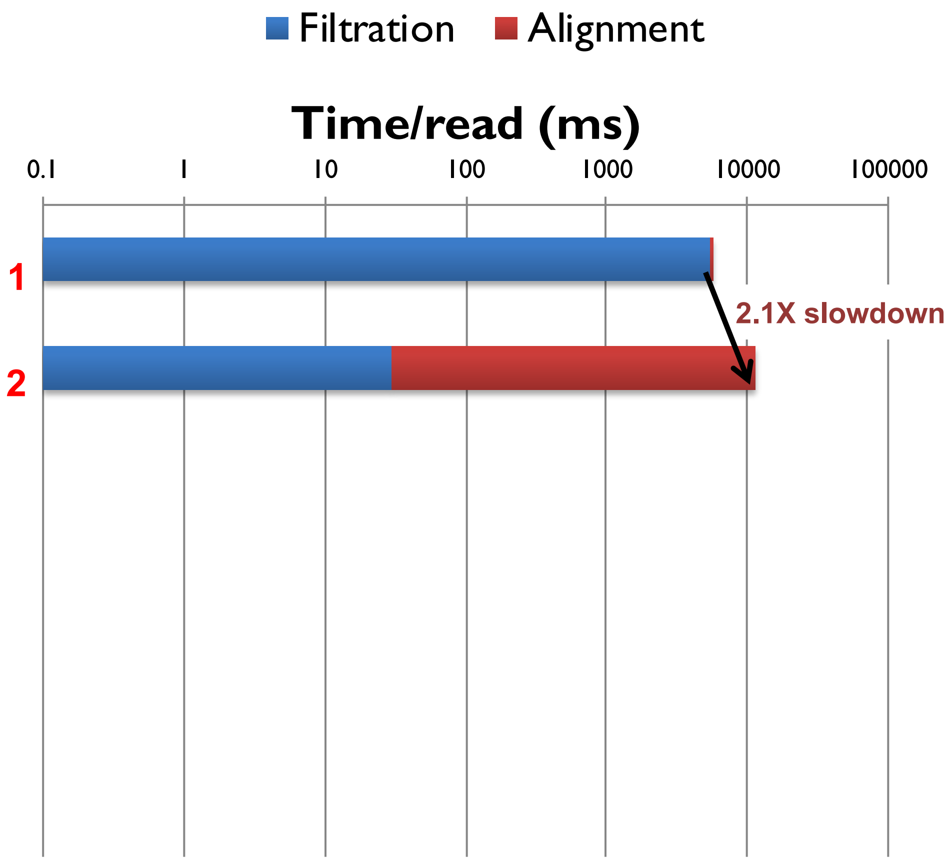
~3 hits

Alignment

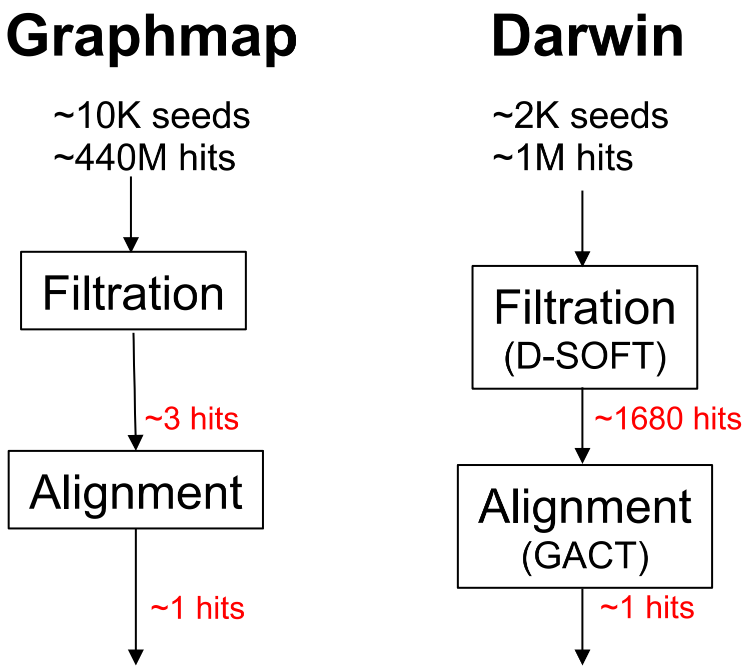
~1 hits

Yatish Turakhia, Gill Bejerano, and William J. Dally. "Darwin: A Genomics Co-processor Provides up to 15,000 X Acceleration on Long Read Assembly." ASPLOS 2018.

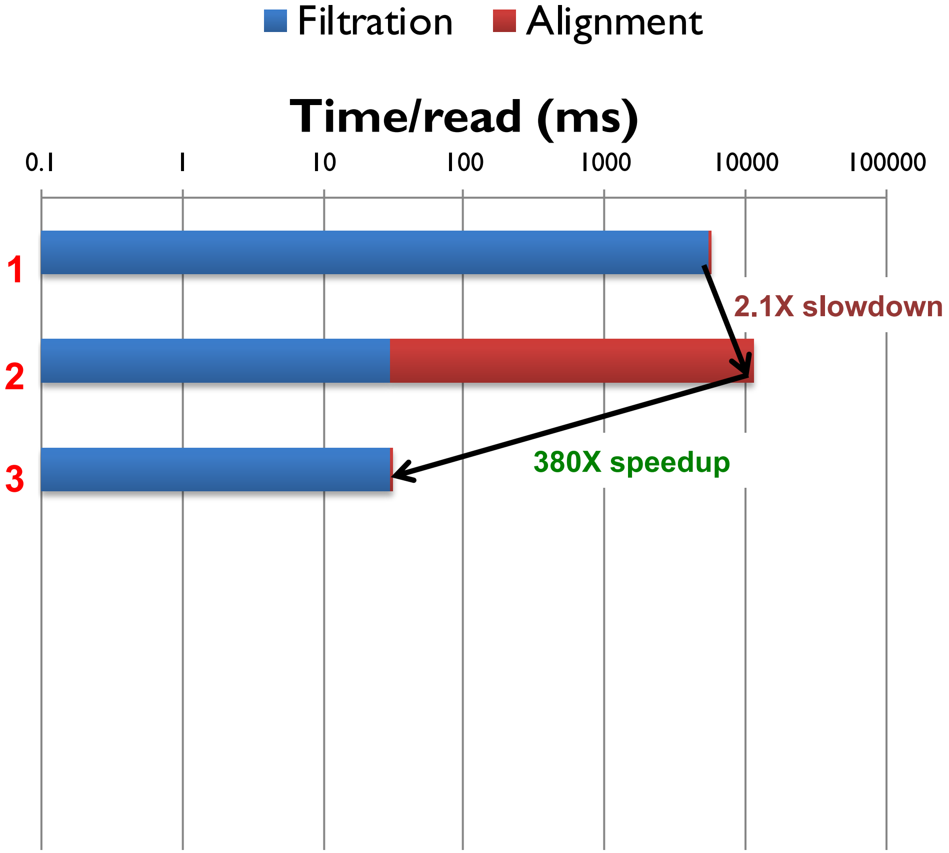
Algorithm-Architecture Co-Design for Darwin
Replace Graphmap with Hardware-Friendly Algorithms
Speed up Filtering by 100x, but 2.1x Slowdown Overall



1. Graphmap (software)
2. Replace by D-SOFT and GACT (software)



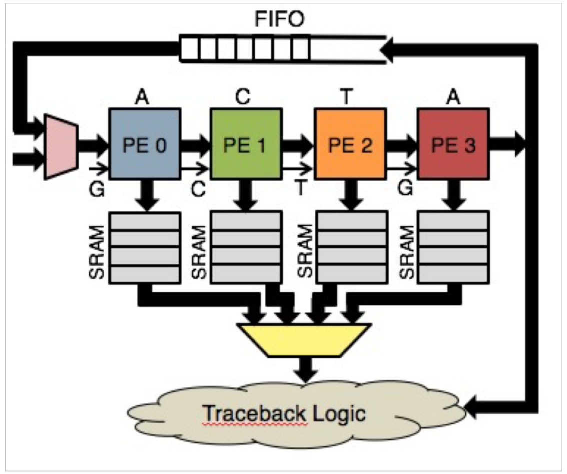
Algorithm-Hardware Co-Design for Darwin Accelerate Alighment – 380x Speedup



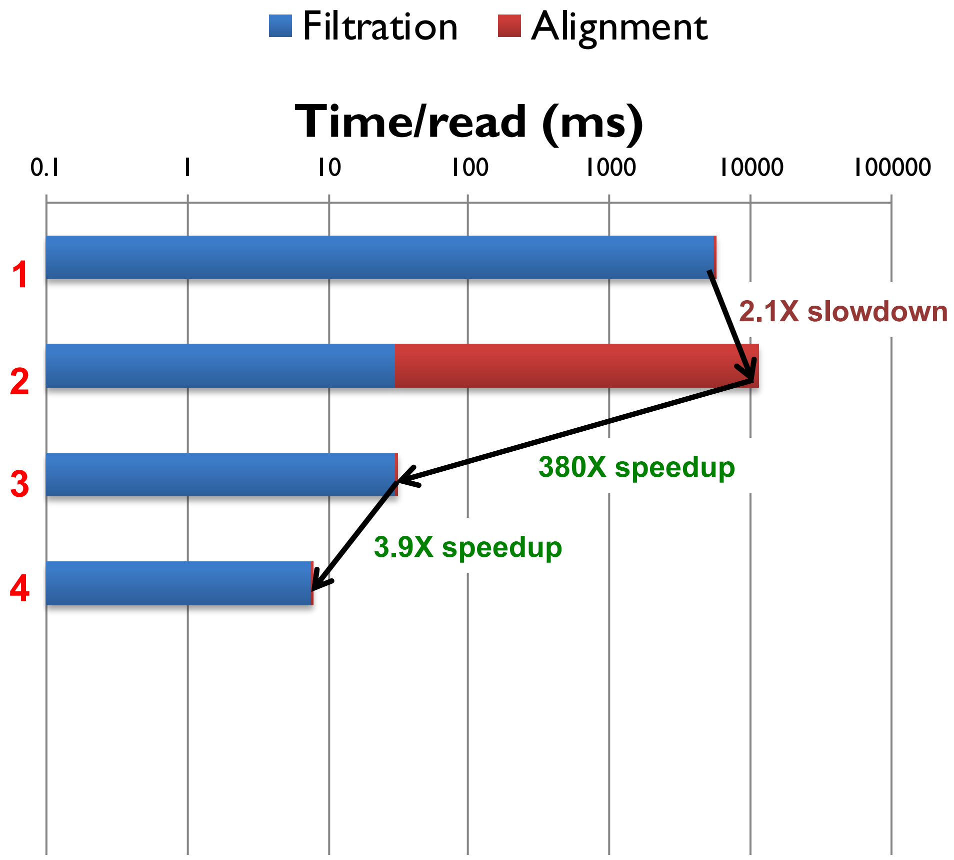
2.1X slowdown

380X speedup

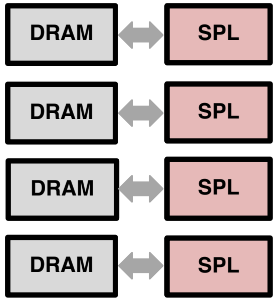
1. Graphmap (software)
2. Replace by D-SOFT and GACT (software)
3. GACT hardware-acceleration



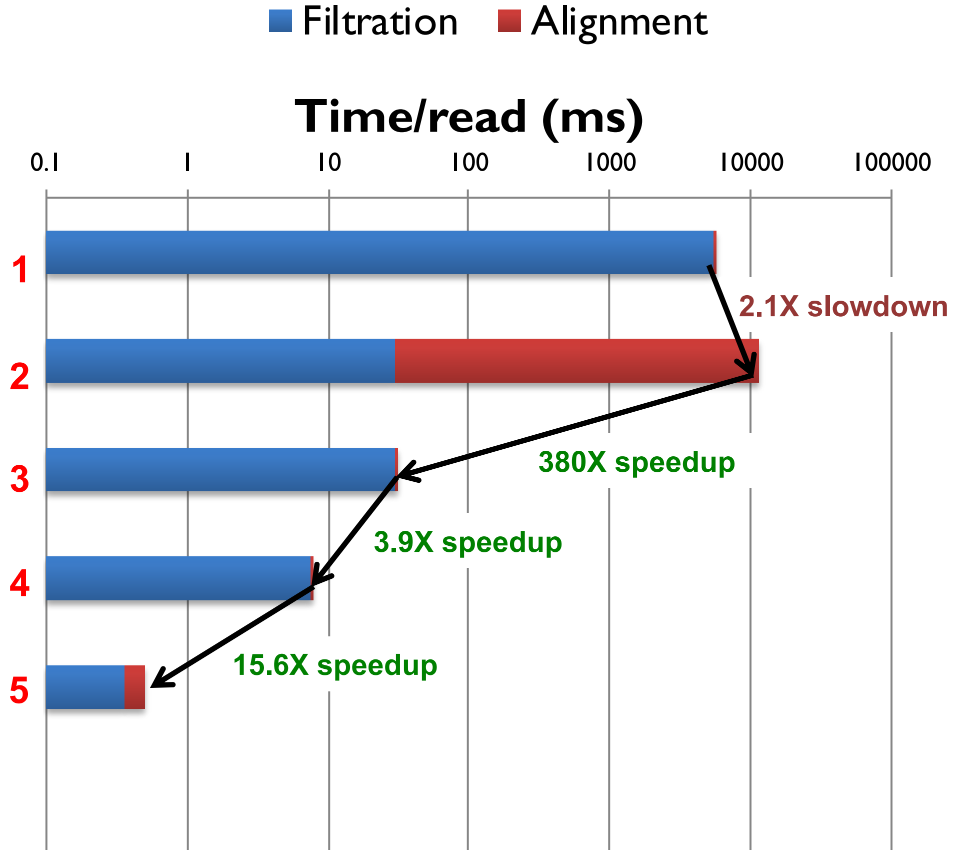
Algorithm-Hardware Co-Design for Darwin
4x Memory Parallelism – 3.9x Speeedup



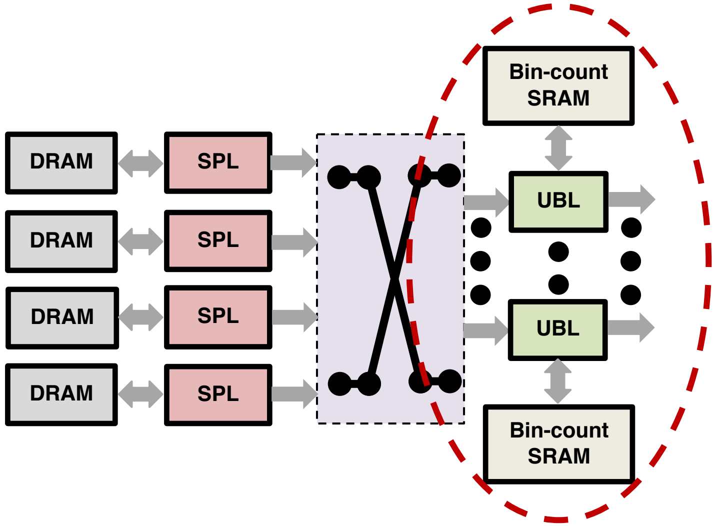
1. Graphmap (software)
2. Replace by D-SOFT and GACT (software)
3. GACT hardware-acceleration
4. Four DRAM channels for D-SOFT



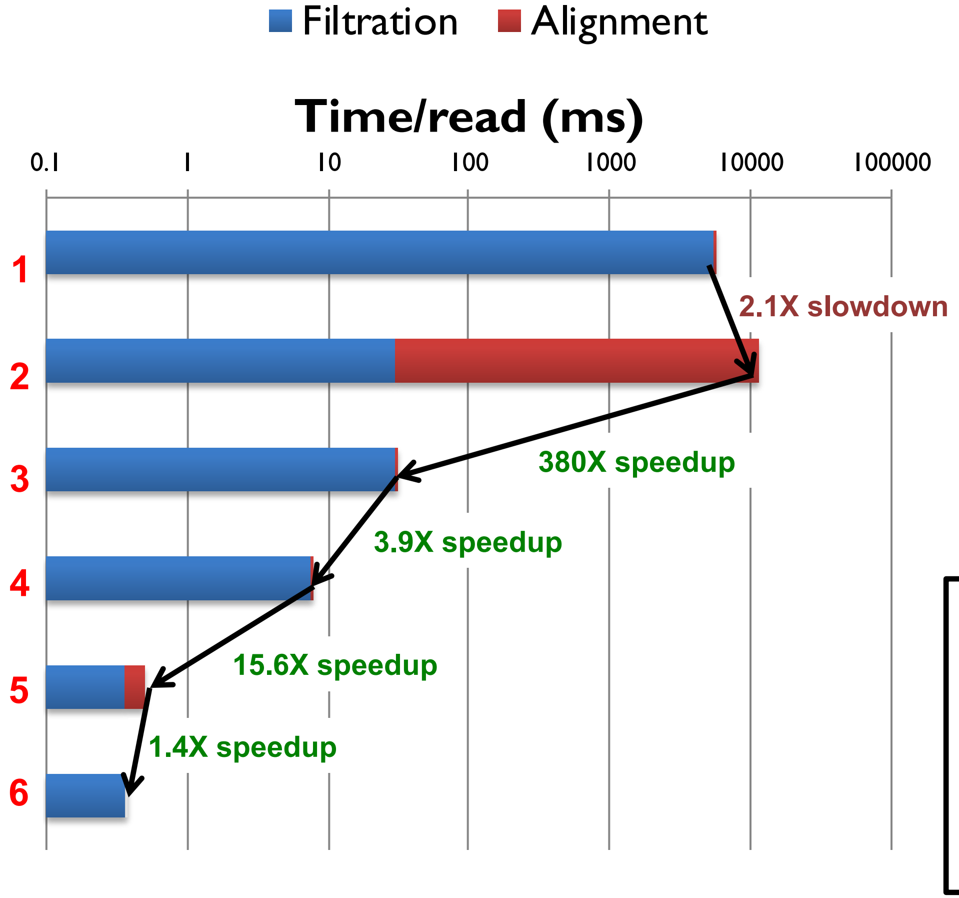
Algorithm-Hardware Co-Design for Darwin
Specialized Memory for D-Soft Bin Updates – 15.6x Speedup



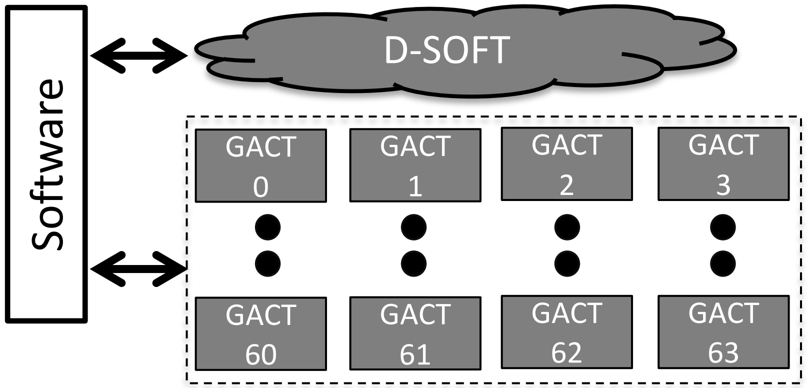
1. Graphmap (software)
2. Replace by D-SOFT and GACT (software)
3. GACT hardware-acceleration
4. Four DRAM channels for D-SOFT
5. Move bin updates in D-SOFT to SRAM (ASIC)



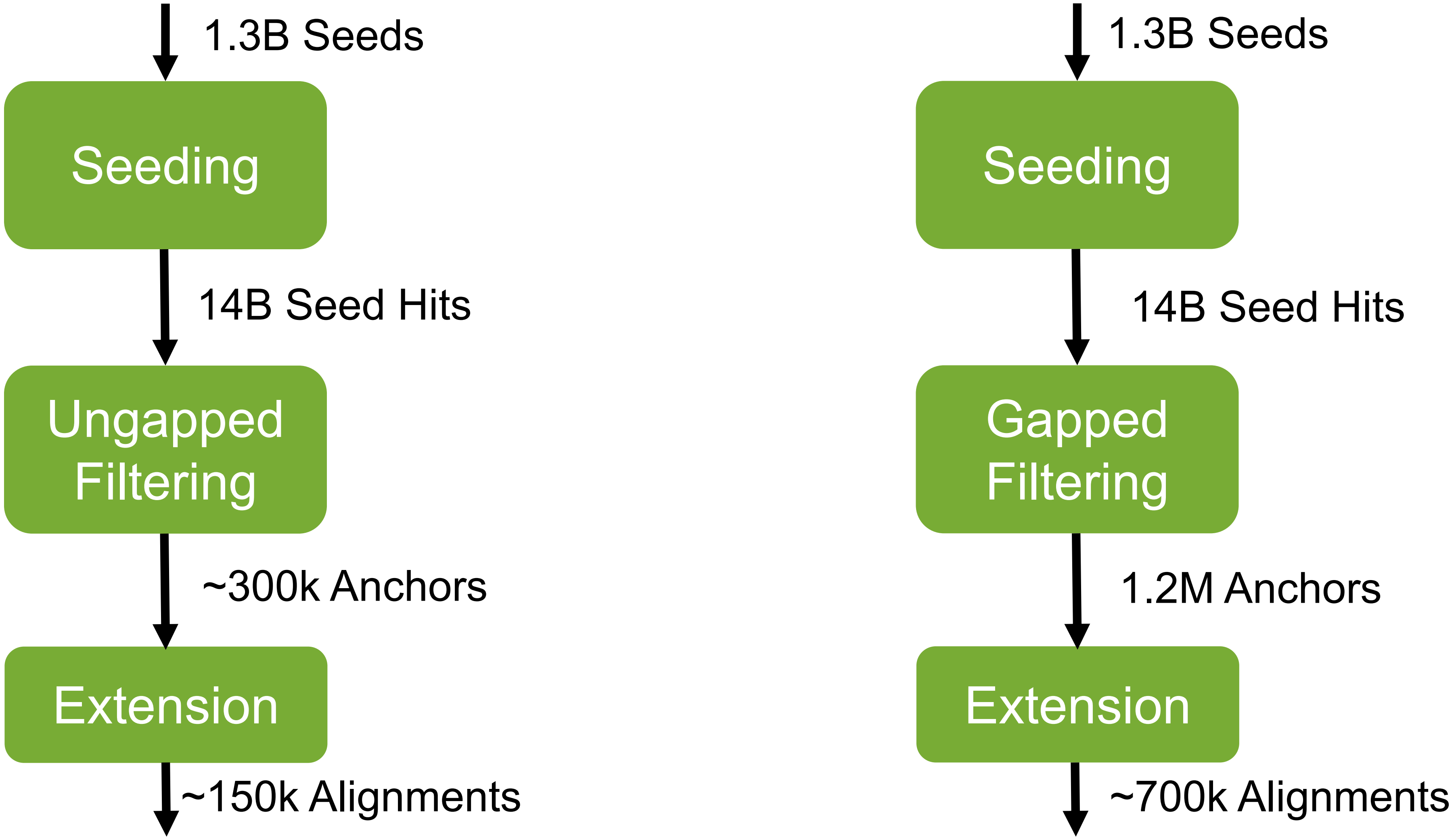
Algorithm-Hardware Co-Design for Darwin
Pipeline D-Soft and GACT – now completely D-Soft limited – 1.4x
Overall 15,000x



1. Graphmap (software)
2. Replace by D-SOFT and GACT (software)
3. GACT hardware-acceleration
4. Four DRAM channels for D-SOFT
5. Move bin updates in D-SOFT to SRAM (ASIC)
6. Pipeline D-SOFT and GACT



Algorithm and Hardware Co-Design for Darwin-WGA



1.3B Seeds

Seeding

14B Seed Hits

Ungapped Filtering

~300k Anchors

Extension

~150k Alignments

1.3B Seeds

Seeding

14B Seed Hits

Gapped Filtering

1.2M Anchors

Extension

~700k Alignments

Yatish Turakhia*, Sneha D. Goenka*, Gill Bejerano, and William J. Dally. "Darwin-WGA: A Co-processor Provides Increased Sensitivity in Whole Genome Alignments with High Speedup" HPCA 2019.

Memory Dominates

Memory dominates power and area

Darwin: ASIC overview

Darwin

		Configuration	Area (mm^2) (40nm TSMC)	Power (W) (40nm TSMC)
GACT	Logic	64 x (64PE array)	17.6	1.04
	Memory	64 x (64PE x 2KB/PE)	68.0	3.36
D-SOFT	Logic	2xSPL + NoC + 16xUBL	6.2	0.41
	Bin-count SRAM	16 banks x 4MB/bank	300.8	7.84
	NZ-bin SRAM	16 x 256KB	19.5	0.96
DRAM	LPDDR4-2400	4 x 32GB	-	1.64
TOTAL			**412.1**	**15.25**

Power and Area dominated by memory

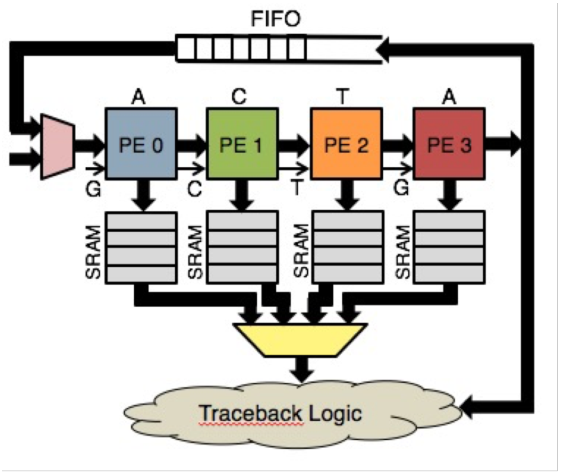
GACT: 79% Area, 76% Power

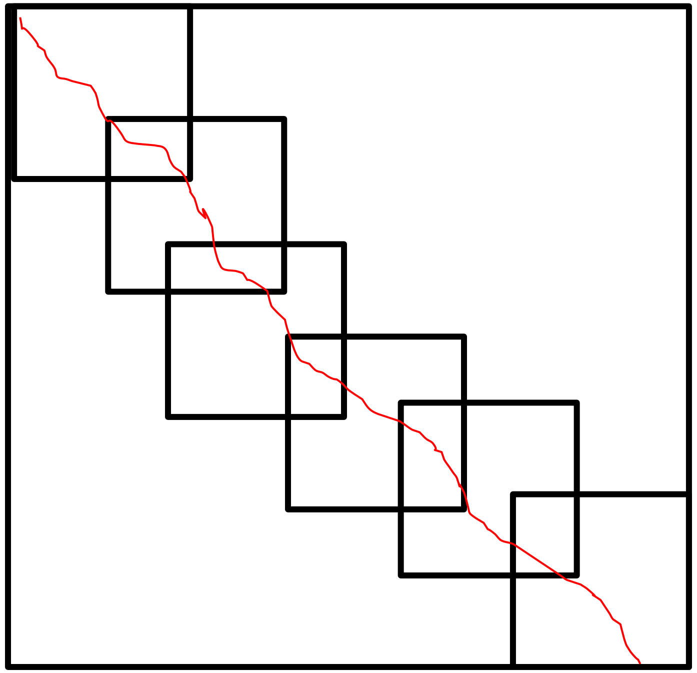
D-SOFT: 98% Area, 96% Power

Algorithms must be optimized to use memory efficiently

GACT Alignment

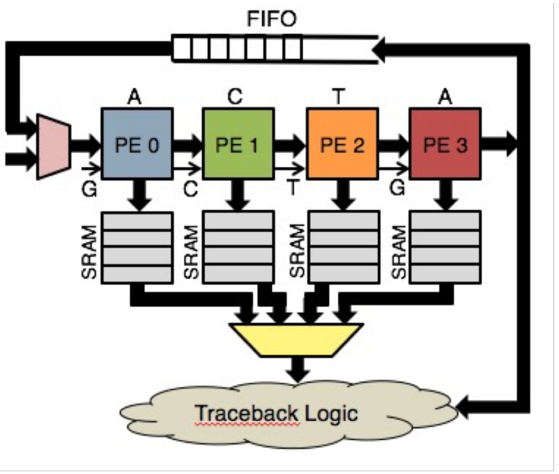
- 15M Reads, 10k bases each, ~2k hits each
  - ~300T Alignments to be done
  - Additional parallelism within each alignment
- But long reads have large (10M) memory footprint
- Solution: GACT (Tiling)

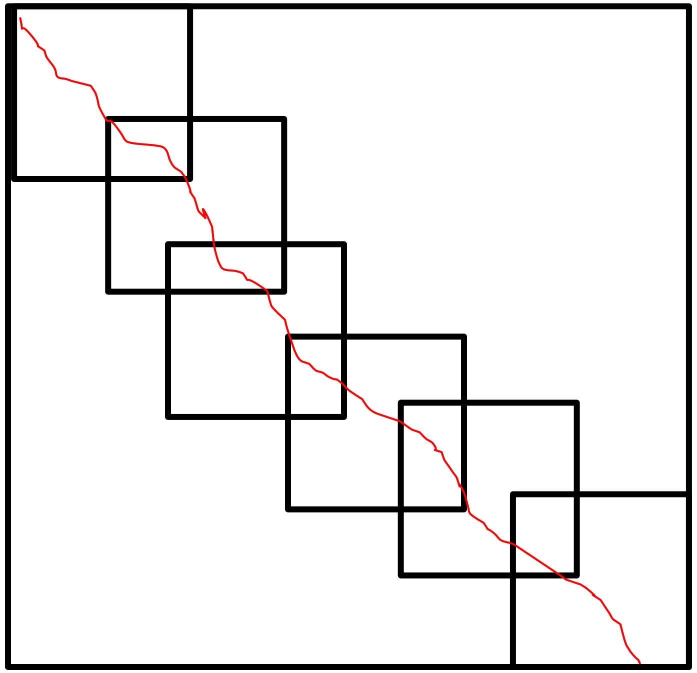




GACT Alignment

- 15M Reads, 10k bases each, ~2k hits each
  - ~300T Alignments to be done
  - Additional parallelism within each alignment
- But long reads have large (10M) memory footprint
- Solution: GACT (Tiling)



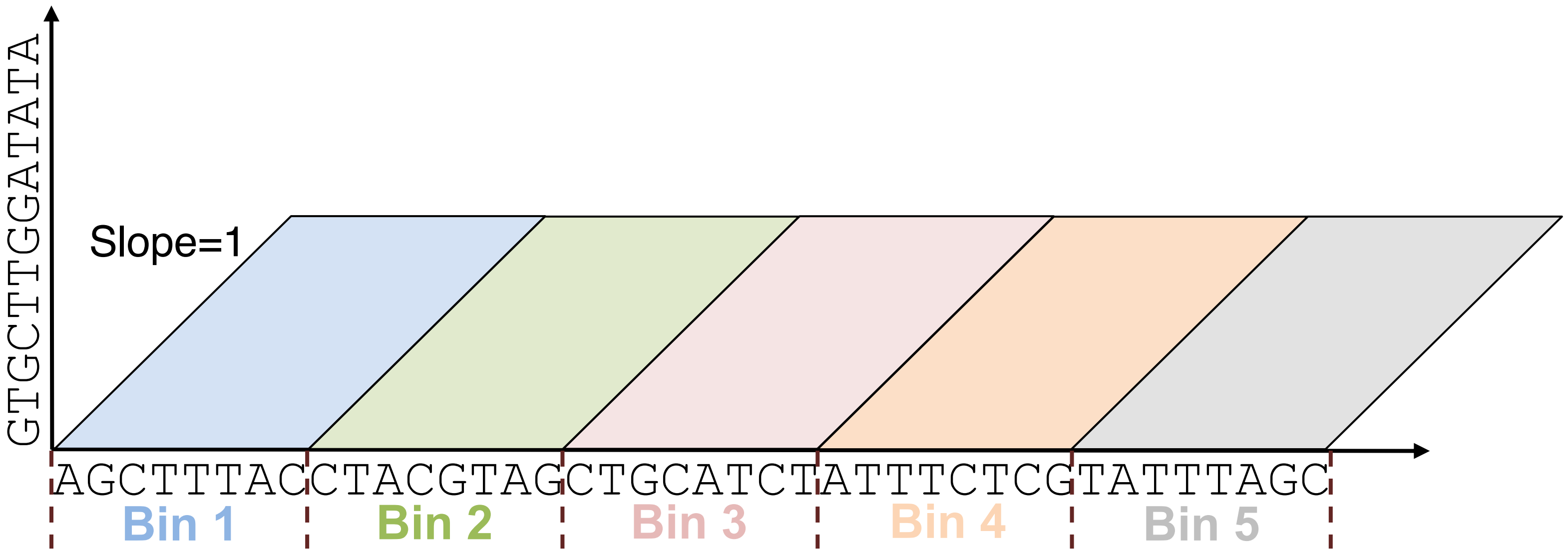


Darwin GACT hardware
4k PEs - 64 PEs per Array x 64 Arrays
~50 operations per cycle per PE
200k operations per cycle
Specialized memory
150,000x speedup vs CPU

On-Chip Memory
Cost per Bit is 10-100x Commodity DRAM

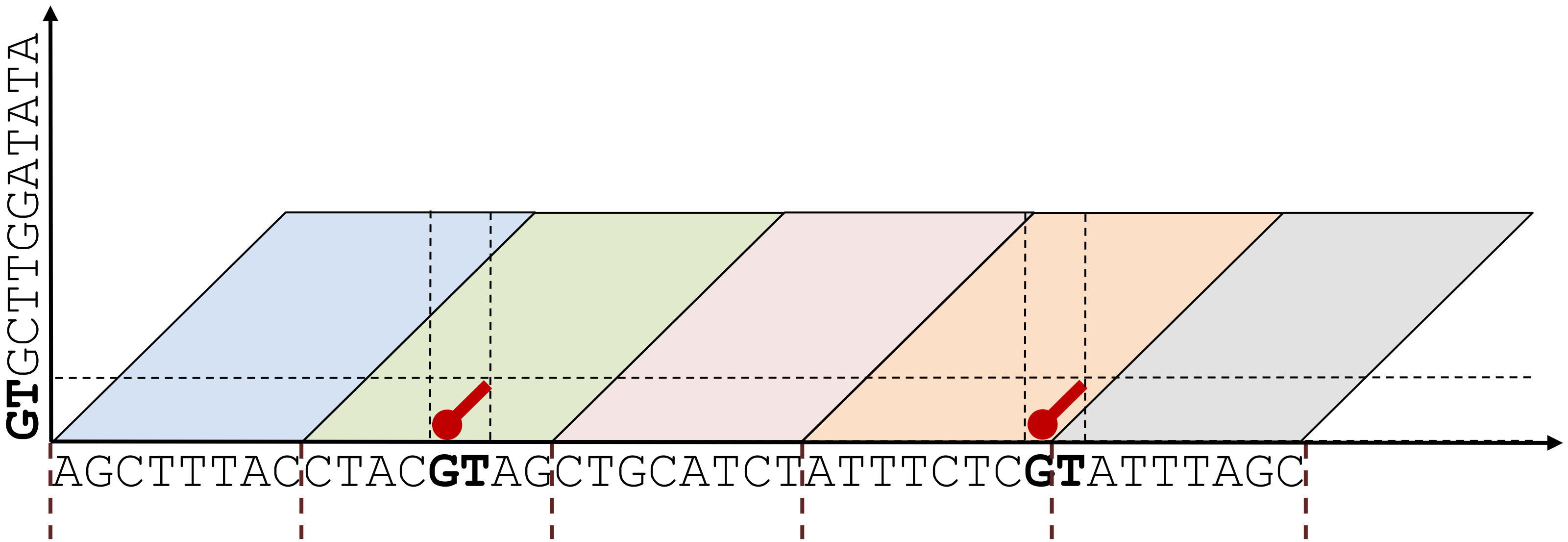
And It’s Often Less Expensive

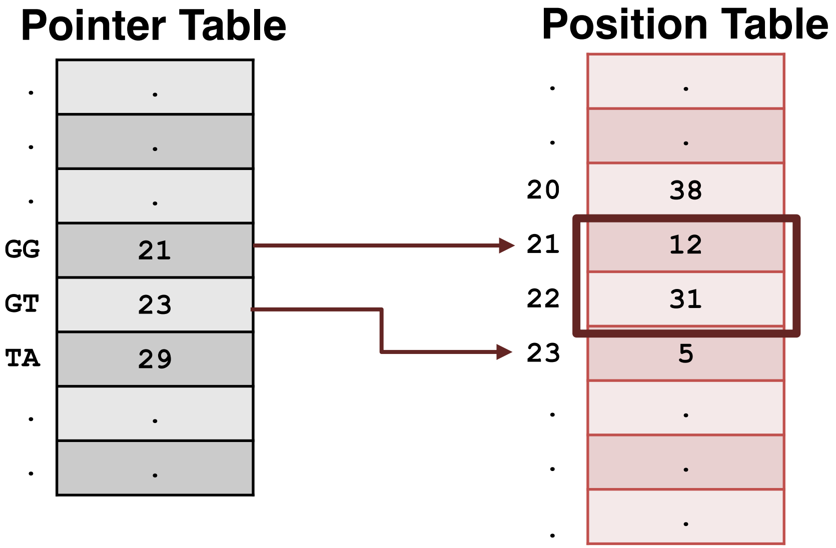
D-SOFT: Algorithm Overview



Bin count (bases)	Last hit offset
0	-inf
0	-inf
0	-inf
0	-inf
0	-inf

D-SOFT: Algorithm Overview





Pointer Table

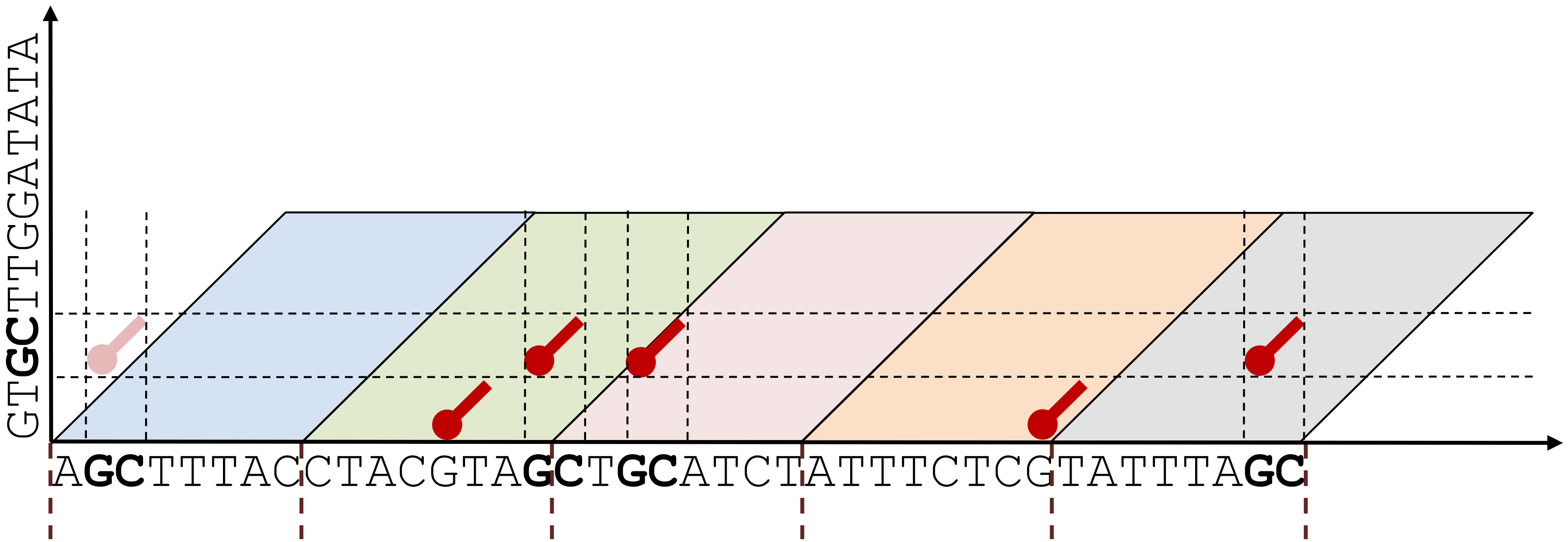
.	.
.	.
.	.
GG	21
GT	23
TA	29
.	.
.	.

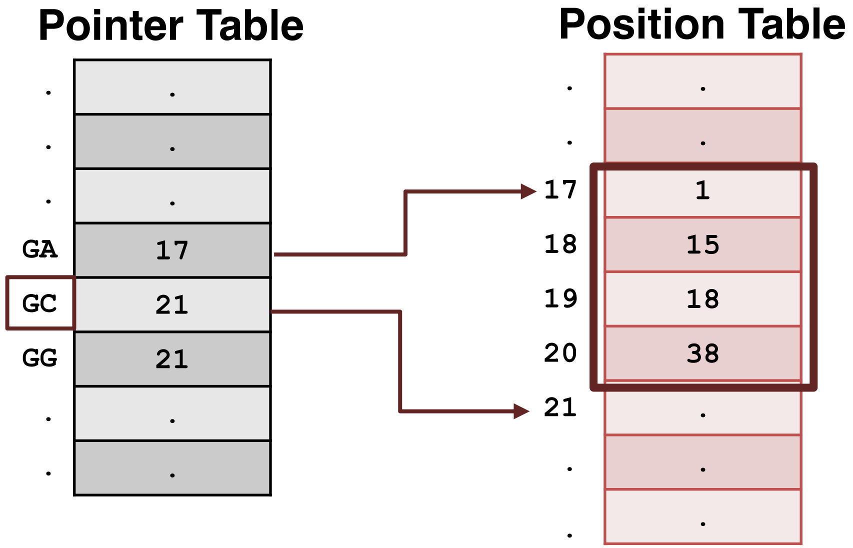
Position Table

.	.
.	.
20	38
21	12
22	31
23	5
.	.
.	.
.	.

Bin count (bases)	Last hit offset
0	-inf
2	0
0	-inf
2	0
0	-inf

D-SOFT: Algorithm Overview





Pointer Table

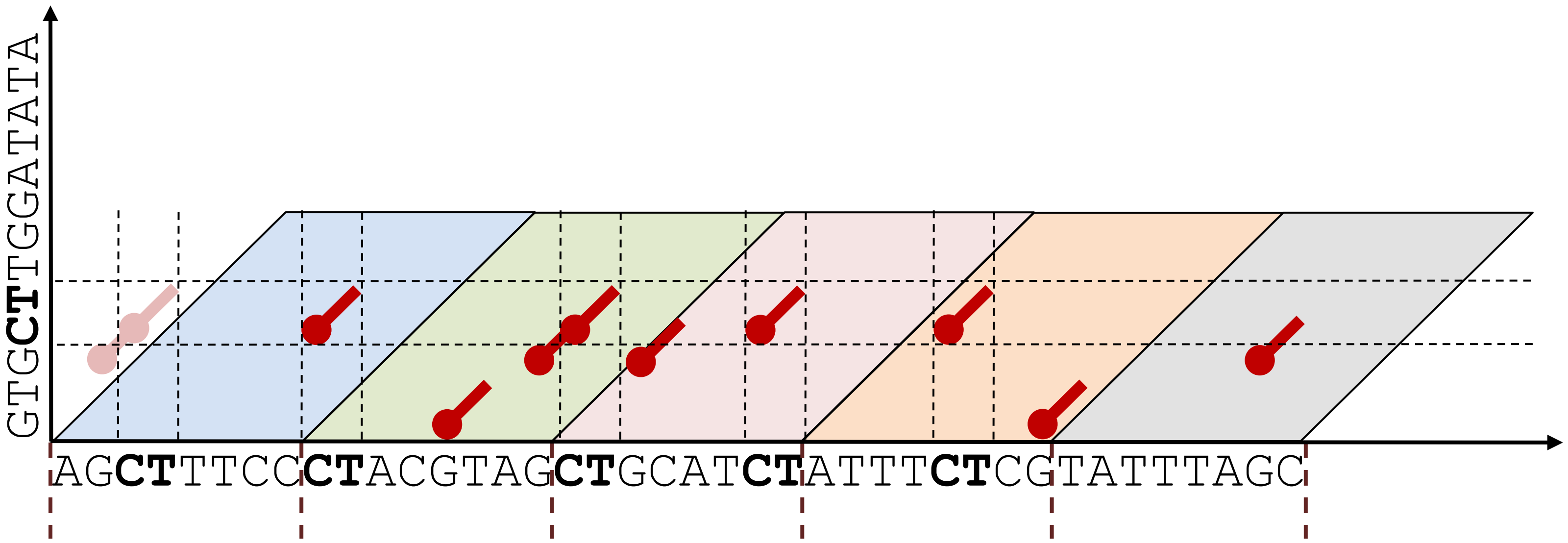
.	.
.	.
.	.
GA	17
GC	21
GG	21
.	.
.	.

Position Table

.	.
.	.
17	1
18	15
19	18
20	38
21	.
.	.
.	.

Bin count (bases)	Last hit offset
0	-inf
4	2
2	2
2	0
2	2

D-SOFT: Algorithm Overview


GTG**CT**TGGATATA
AG**CT**TTCC**CT**ACGTAG**CT**GCAT**CT**ATTT**CT**CGTATTTAGC

Pointer Table

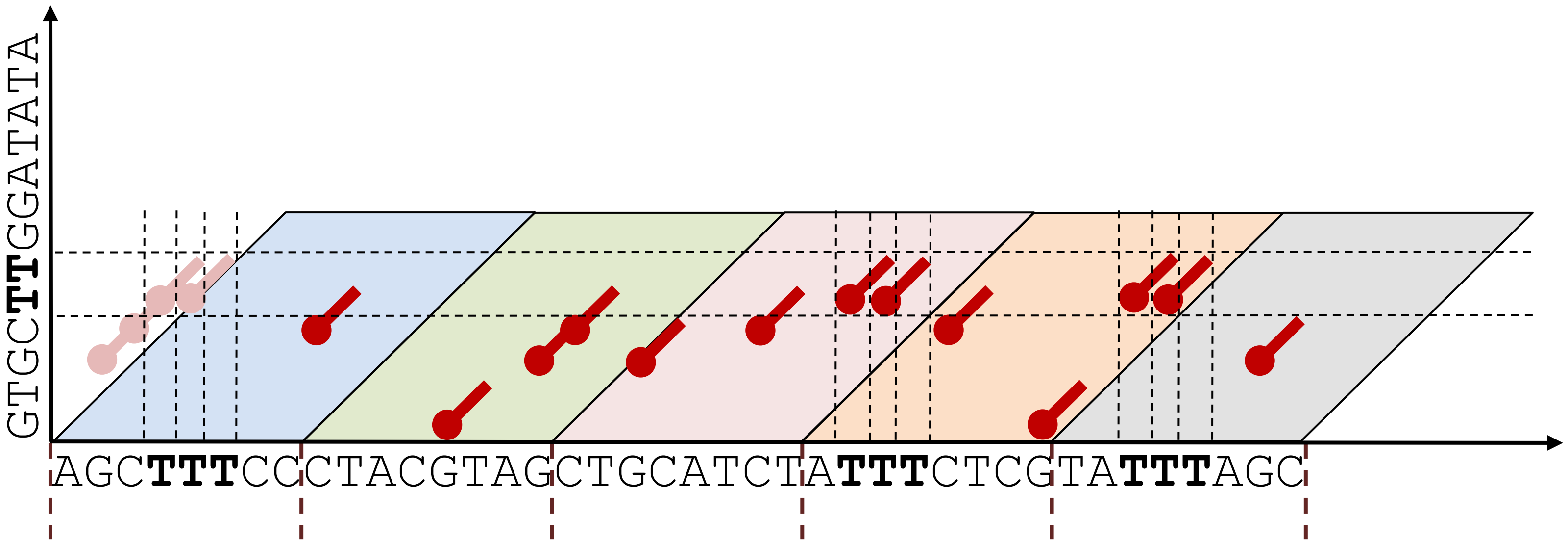
.	.
.	.
CG	12
CT	17
GA	17
.	.
.	.
.	.

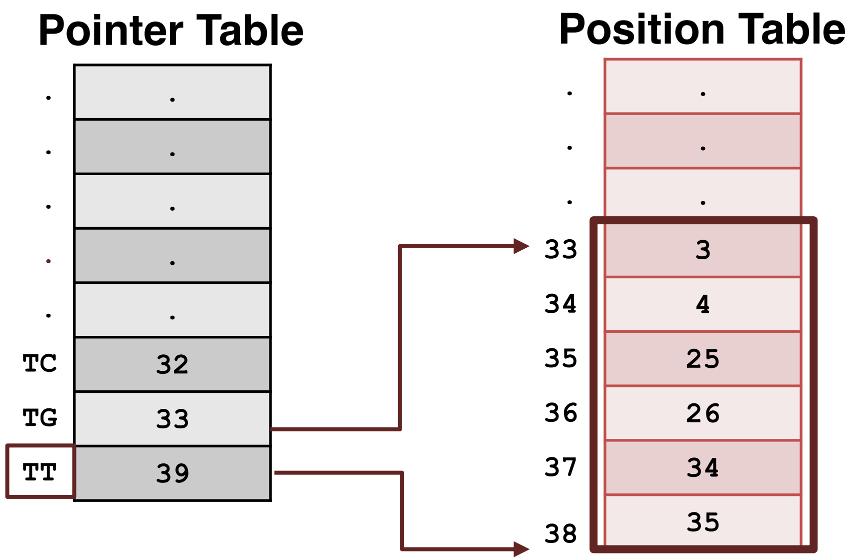
Position Table

.	.
12	2
13	8
14	16
15	22
16	28
17	1
.	.
.	.

Bin count (bases)	Last hit offset
2	3
5	3
3	3
4	3
2	2

D-SOFT: Algorithm Overview





Pointer Table

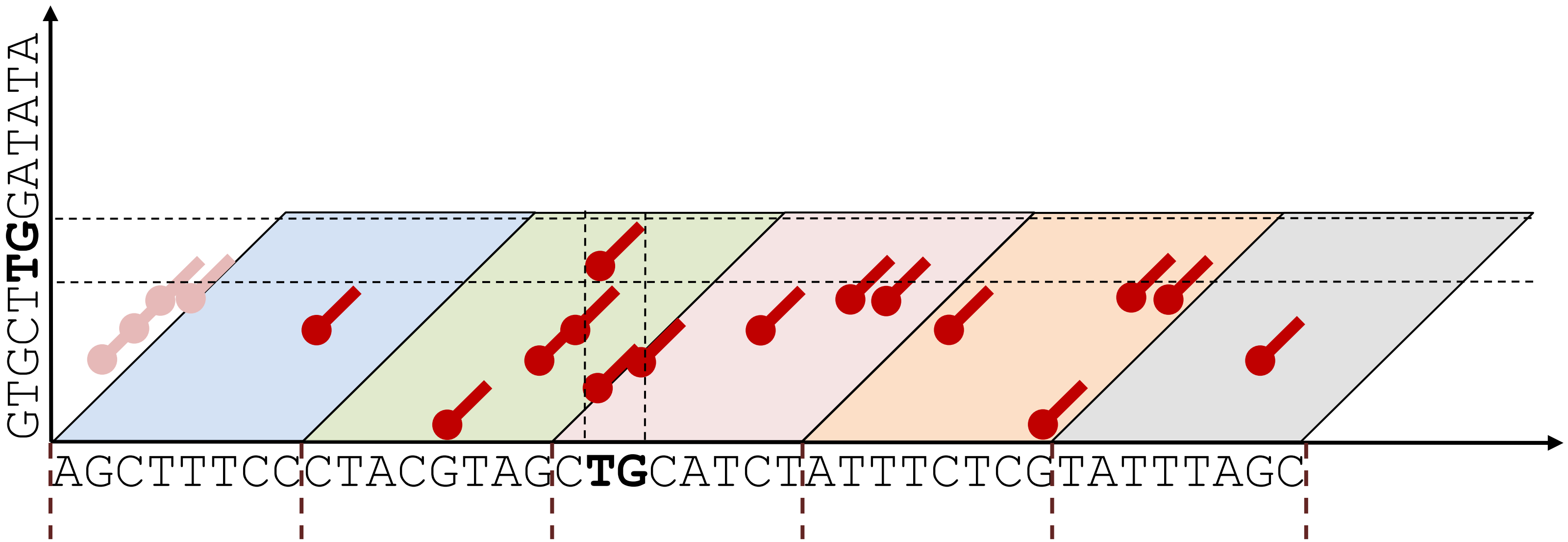
.	.
.	.
.	.
.	.
.	.
TC	32
TG	33
TT	39

Position Table

.	.
.	.
.	.
33	3
34	4
35	25
36	26
37	34
38	35

Bin count (bases)	Last hit offset
2	3
5	3
4	4
5	4
2	2

D-SOFT: Algorithm Overview



Pointer Table

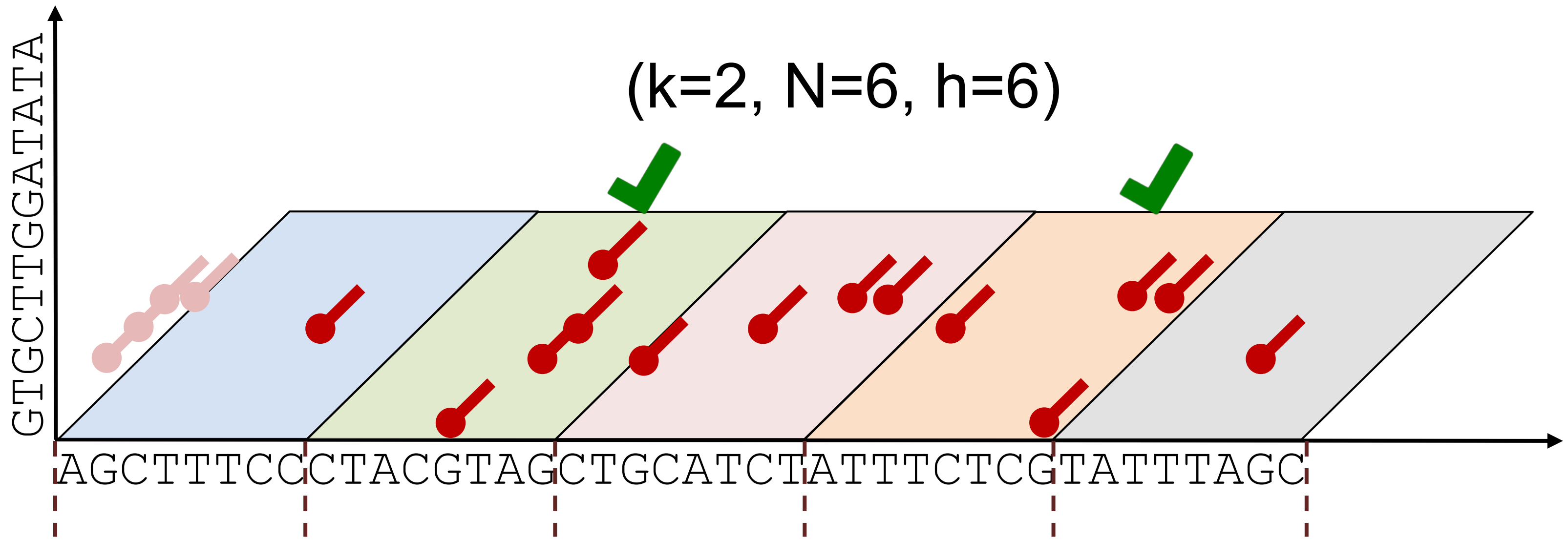
.	.
.	.
.	.
.	.
.	.
TC	32
TG	33
TT	39

Position Table

.	.
.	.
.	.
.	.
32	17
33	3
34	4
.	.
.	.

Bin count (bases)	Last hit offset
2	3
7	5
4	4
5	4
2	2

D-SOFT: Algorithm Overview



Parameters:

k: seed size

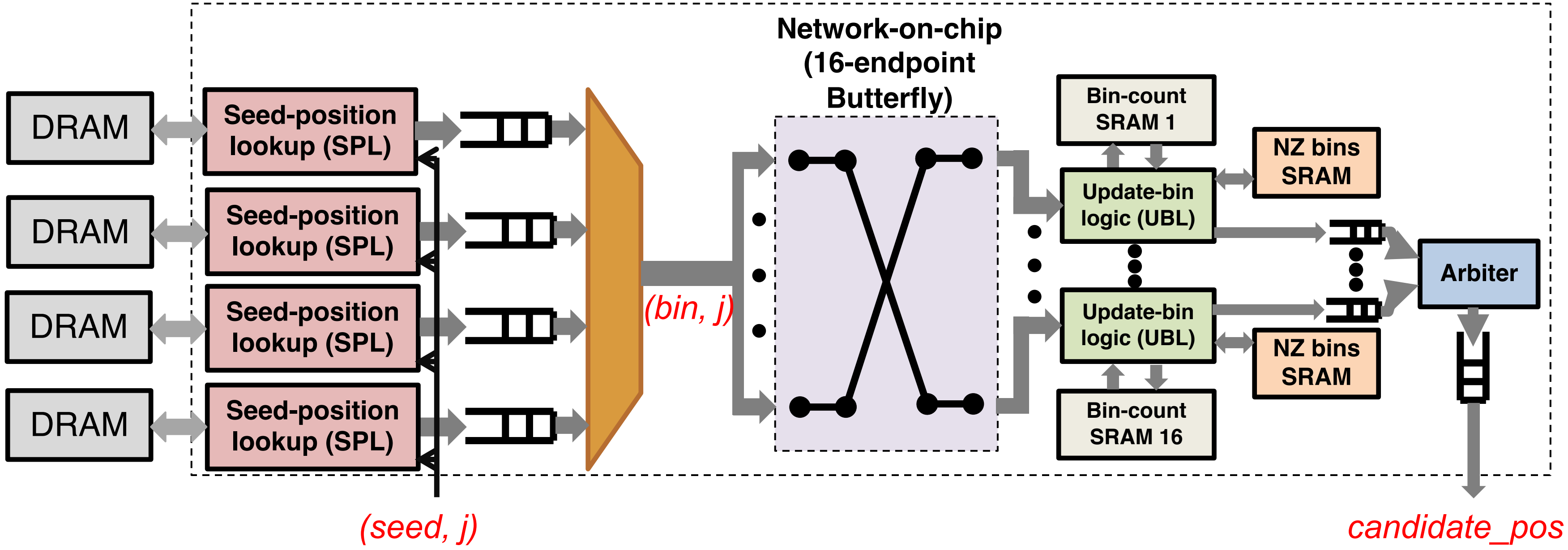
N: number of seeds

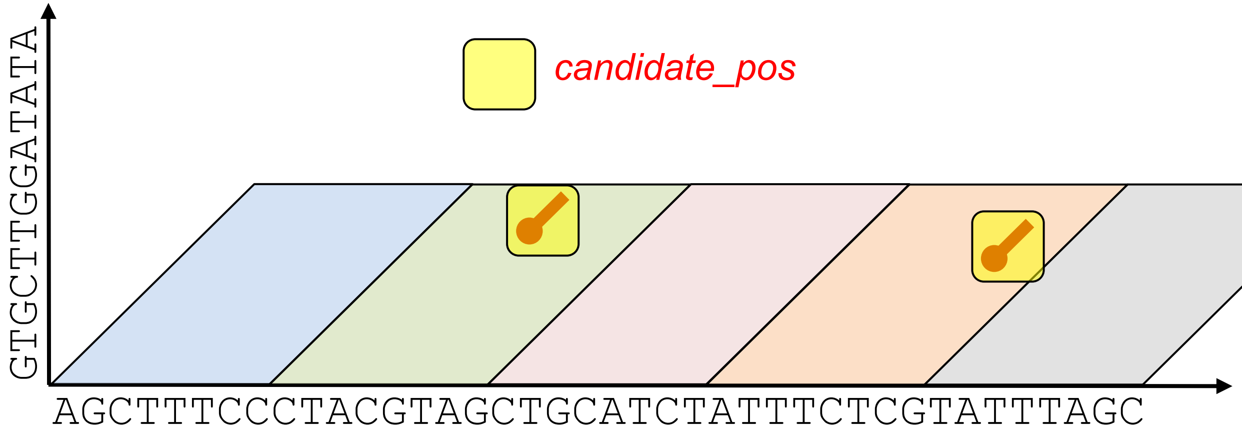
h: threshold on non-overlapping bases

B: bin size (number of bases, fixed to 128)

	Bin count (bases)	Last hit offset
	2	3
✓	7	5
	4	4
✓	5	4
	2	2

D-SOFT: Hardware-acceleration





Cost has a Time Component

$$C = T(B_1N_1 + B_2N_2 + \dots + P)$$

	T	B_1	N_1	B_2	N_2	C
Darwin Filter	1	100	64M	1	128G	134G
All DRAM	15.6			1	128G	1,997G

Platforms for Acceleration

GPUs Provide:

- High-Bandwidth, Hierarchical **Memory** System
  - Can be configured to match application

- Programmable **Control** and **Operand Delivery**

- Simple places to bolt on **Domain-Specific Hardware**
  - As instructions or memory clients

Volta V100

21B xtors | TSMC 12nm FFN | 815mm^2

5,120 CUDA cores

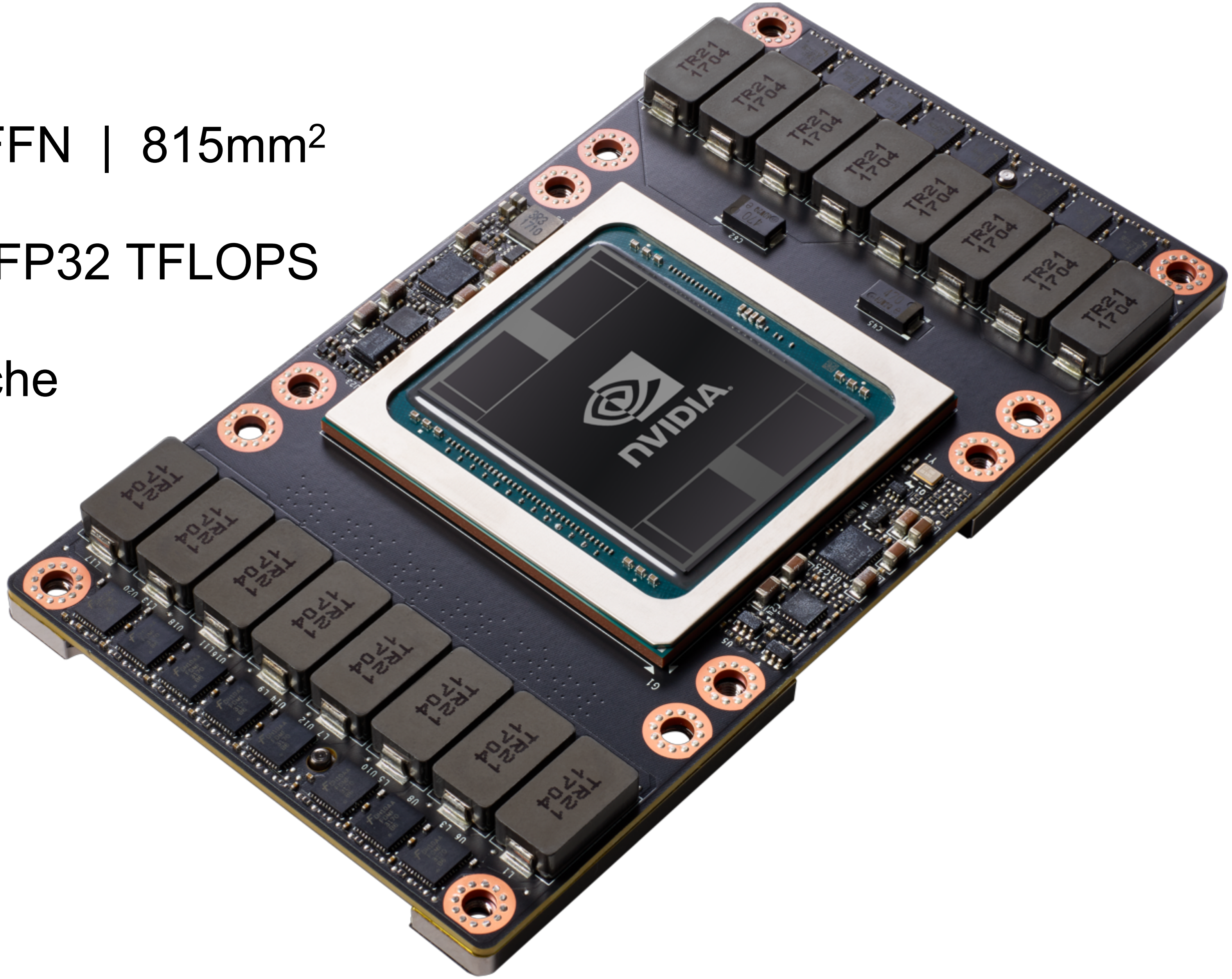
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS

125 Tensor TFLOPS

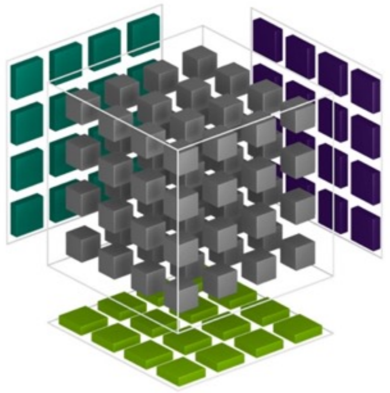
20MB SM RF | 16MB Cache

32GB HBM2 @ 900 GB/s

300 GB/s NVLink



Tensor Core



$$D = \underset{\text{FP16}}{\begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix}} \underset{\text{FP16}}{\begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix}} + \underset{\text{FP16 or FP32}}{\begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}}$$

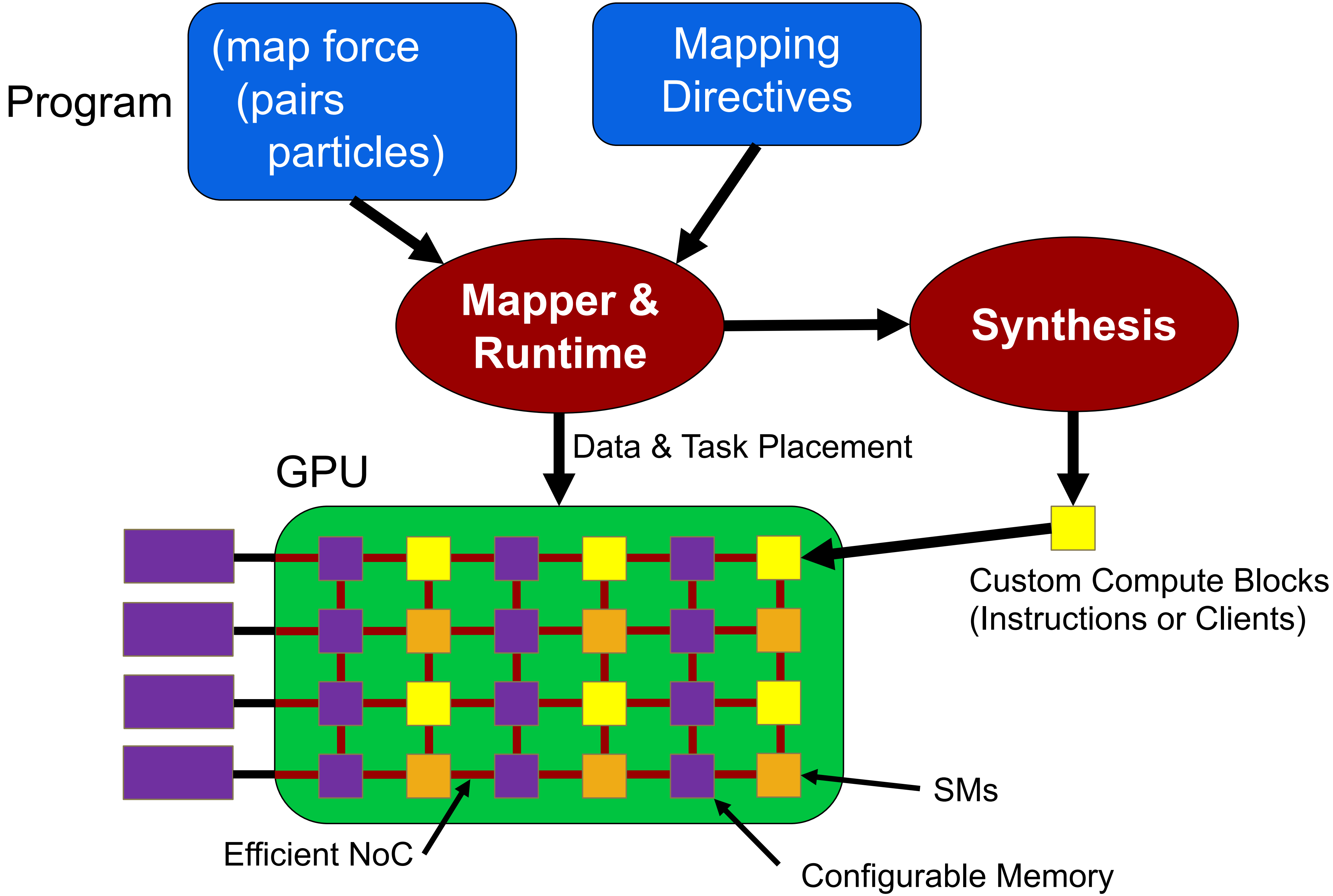
$$D = AB + C$$

Specialized Instructions Amortize Overhead

Operation	Ops	Energy**	Overhead*
HFMA	2	1.5pJ	2000%
HDP4A	8	6.0pJ	500%
HMMA	128	110pJ	27%

*Overhead is instruction fetch, decode, and operand fetch – 30pJ

**Energy numbers from 45nm process



Program

(map force (pairs particles))

Mapping Directives

Mapper & Runtime

Synthesis

Data & Task Placement

GPU

Custom Compute Blocks (Instructions or Clients)

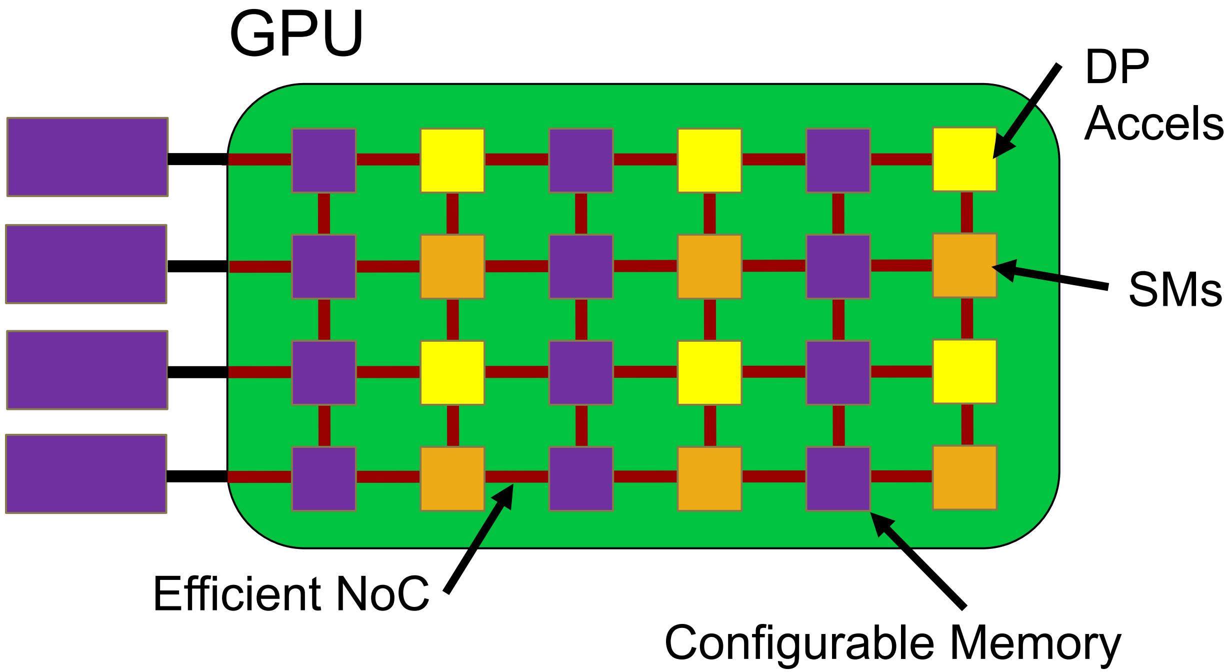
SMs

Efficient NoC

Configurable Memory

Toward a General Bio-Informatics Accelerator

- GPU Substrate
  - Optimized memory subsystem for accessing seed tables
  - SMs update bins in local memory for filtering



- General Dynamic Programming Accelerator
  - Variable alphabet (bases, amino acids,…)
  - Gapped or ungapped filtering or extension
  - GACT-X
  - Arbitrary cost function
  - Supports genome graphs
  - Subset of arrays have traceback memory
- Can do
  - Reference-guided assembly
  - De-novo assembly
  - Whole genome alignment
  - Multiple-sequence alignment
  - Others…

Conclusion

Summary

- **Sequencing** technology is **scaling**, compute performance isn’t
- Many **compelling problems** in bioinformatics
  - Phylogenomics
  - Driver mutation for cancer
  - Metagenomics
- Problems have **enormous complexity** (270 CPU years to solve birds)
- Specialized hardware is needed
  - **Specialization** provides **efficiency**
  - **parallelization** provides **performance**
  - **Memory** dominates
  - Algorithm/Hardware **co-design** required
- **GPUs** provide a **platform** for acceleration
  - Can support a **general bioinformatics accelerator**