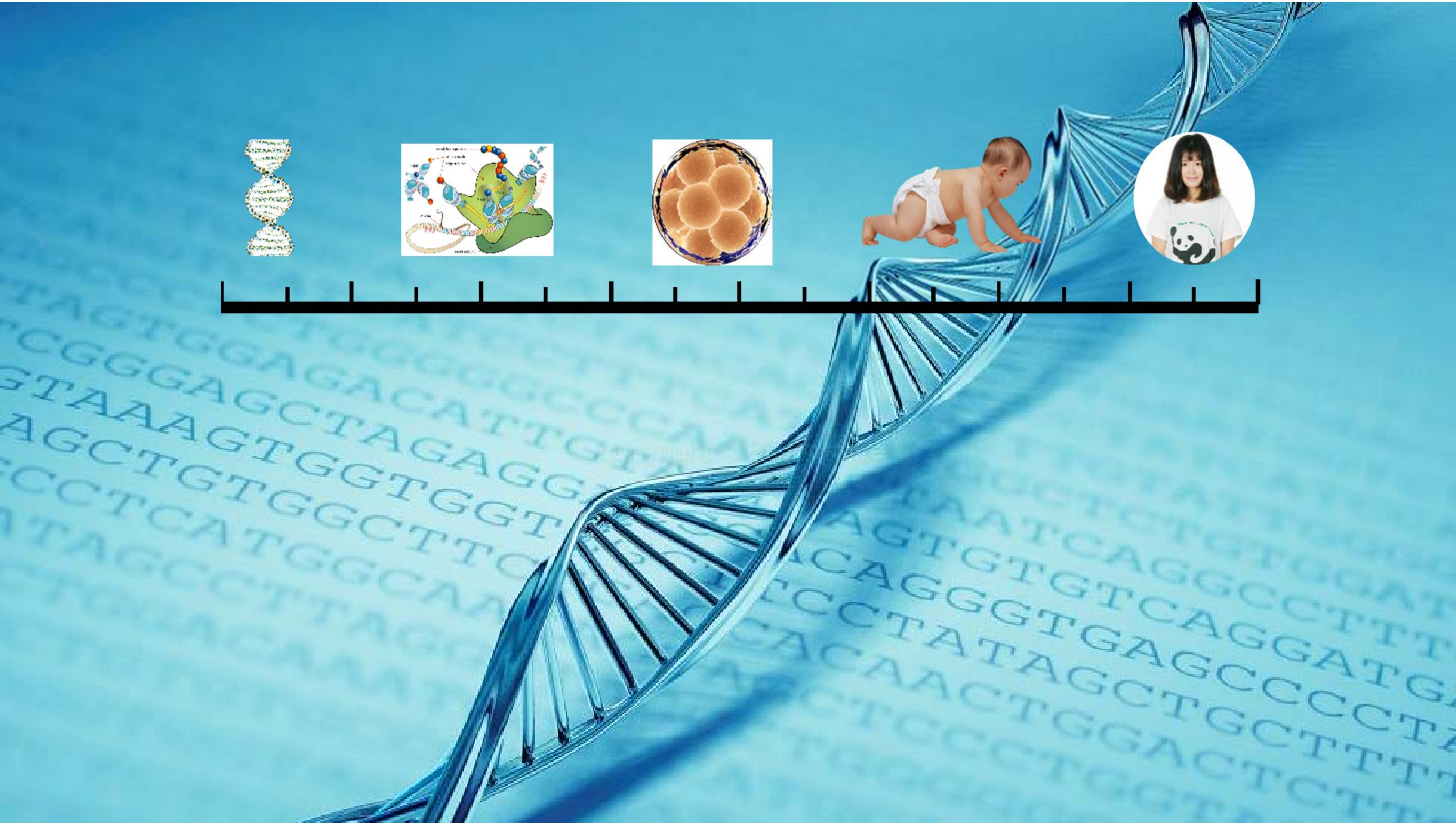
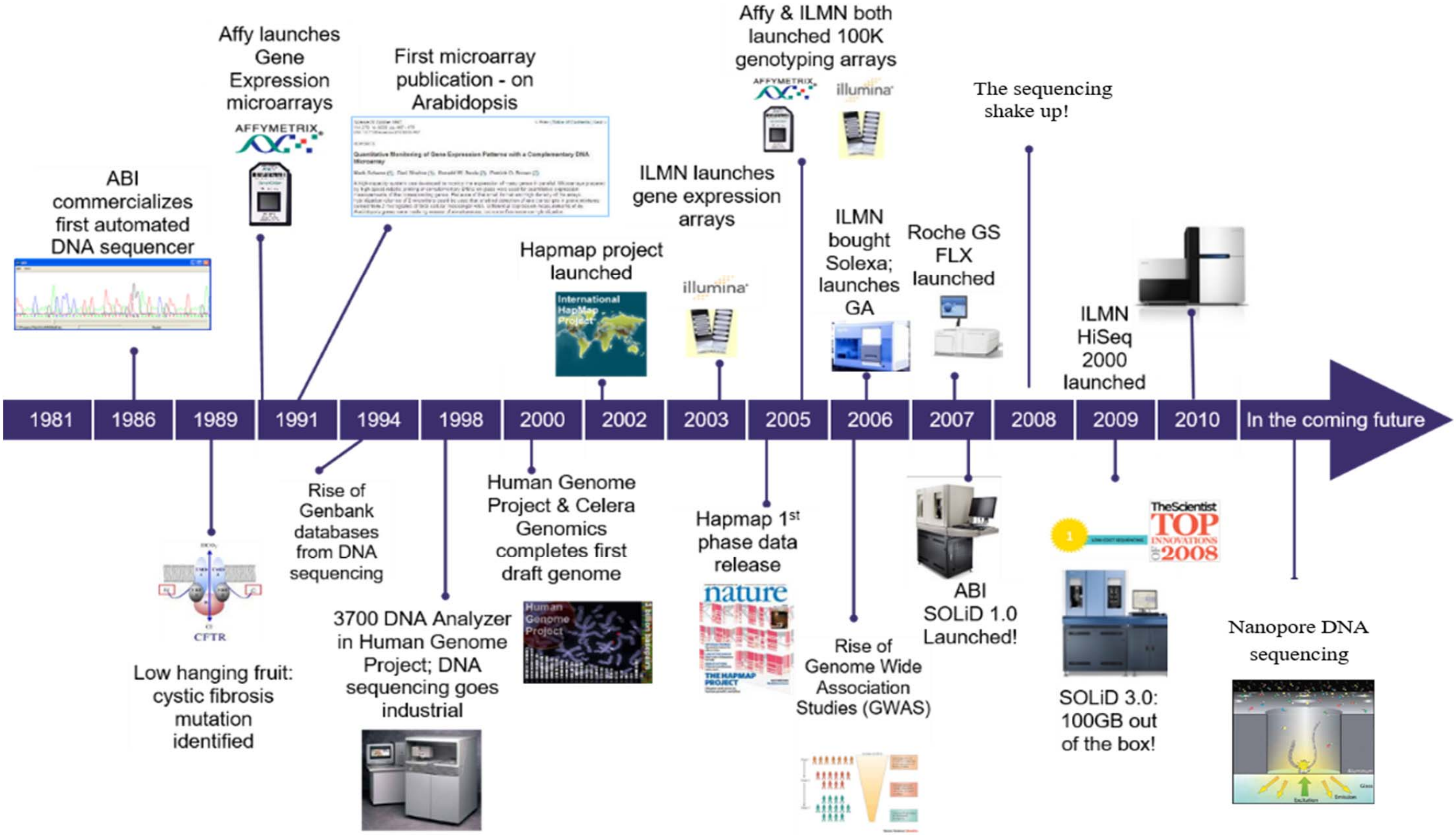
Accurate and Scalable Query Over Large RNA-seq Experiments

Jinze Liu
Department of Computer Science
University of Kentucky,
Oct 17th , 2017




AFFYMETRIX
illumina
International HapMap Project
Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray
CFTR
Human Genome Project
nature
THE HAPMAP PROJECT
TheScientist TOP INNOVATIONS OF 2008

1981 | 1986 | 1989 | 1991 | 1994 | 1998 | 2000 | 2002 | 2003 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | In the coming future

ABI commercializes first automated DNA sequencer

Low hanging fruit: cystic fibrosis mutation identified

Affy launches Gene Expression microarrays

Rise of Genbank databases from DNA sequencing

First microarray publication - on Arabidopsis

3700 DNA Analyzer in Human Genome Project; DNA sequencing goes industrial

Human Genome Project & Celera Genomics completes first draft genome

Hapmap project launched

ILMN launches gene expression arrays

Affy & ILMN both launched 100K genotyping arrays

Hapmap 1st phase data release

ILMN bought Solexa; launches GA

Rise of Genome Wide Association Studies (GWAS)

Roche GS FLX launched

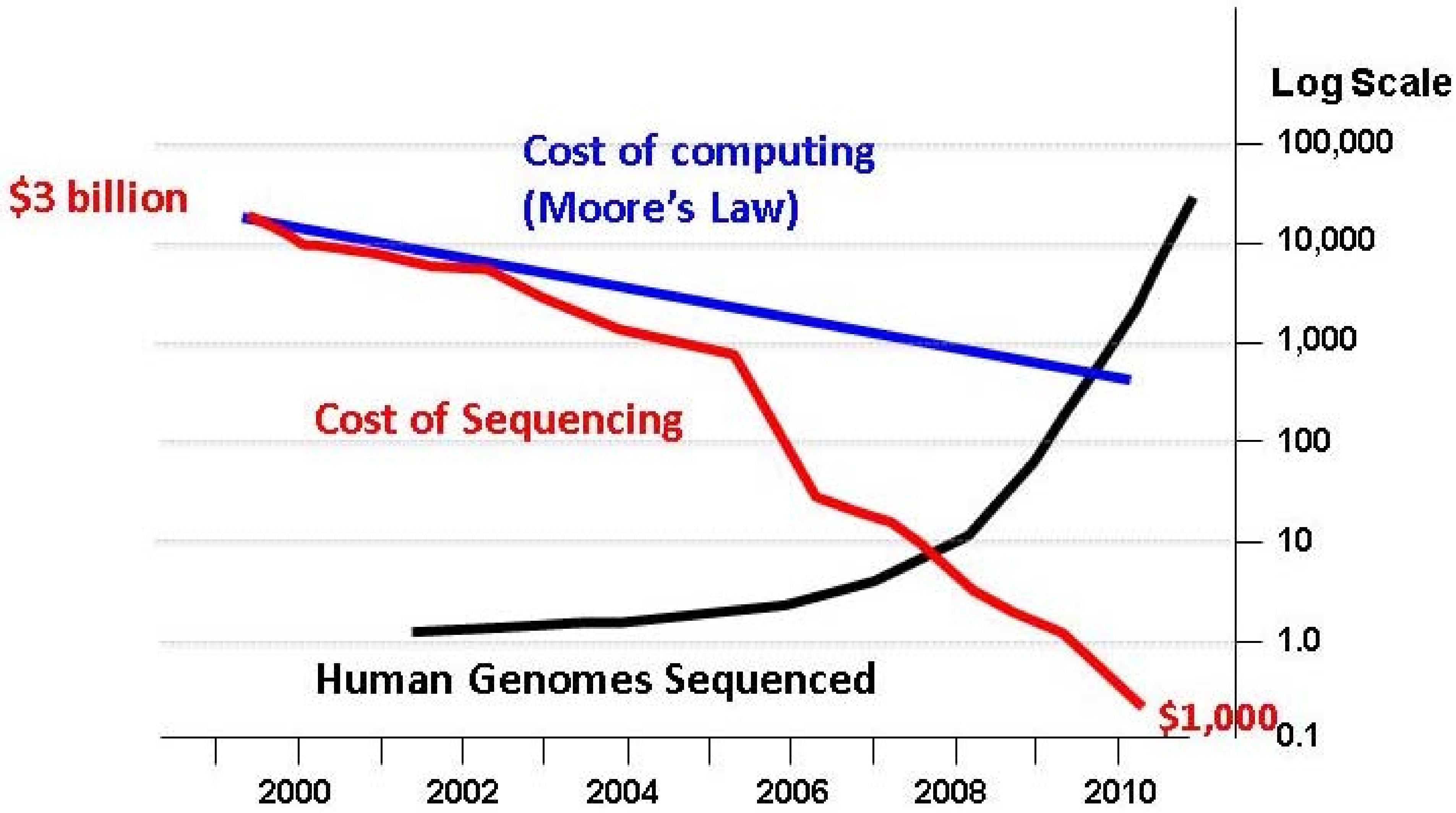
ABI SOLiD 1.0 Launched!

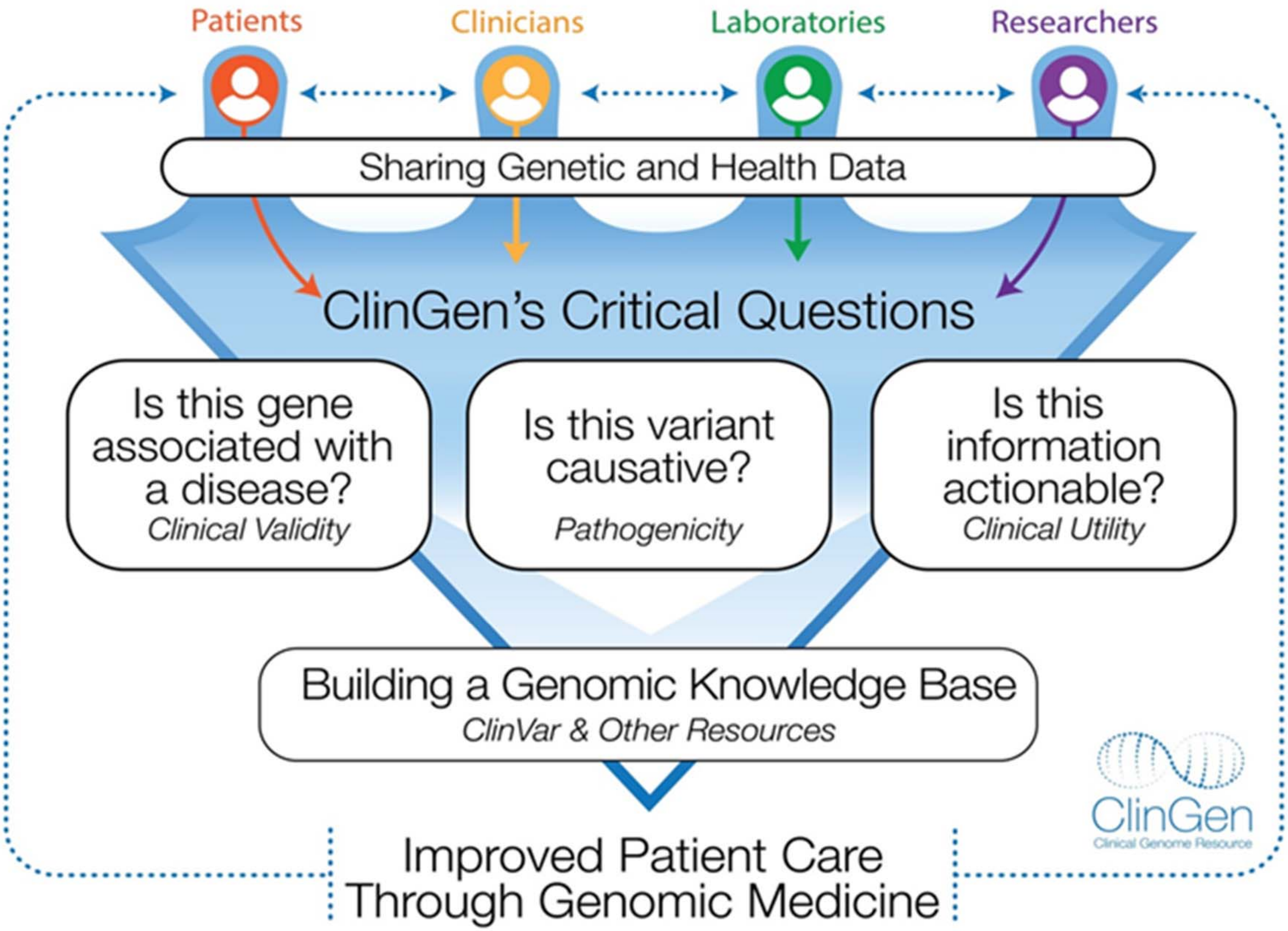
The sequencing shake up!

SOLiD 3.0: 100GB out of the box!

ILMN HiSeq 2000 launched

Nanopore DNA sequencing


Log Scale
100,000
10,000
1,000
100
10
1.0
0.1
$3 billion
Cost of computing
(Moore's Law)
Cost of Sequencing
Human Genomes Sequenced
$1,000
2000
2002
2004
2006
2008
2010



Patients

Clinicians

Laboratories

Researchers

Sharing Genetic and Health Data

ClinGen's Critical Questions

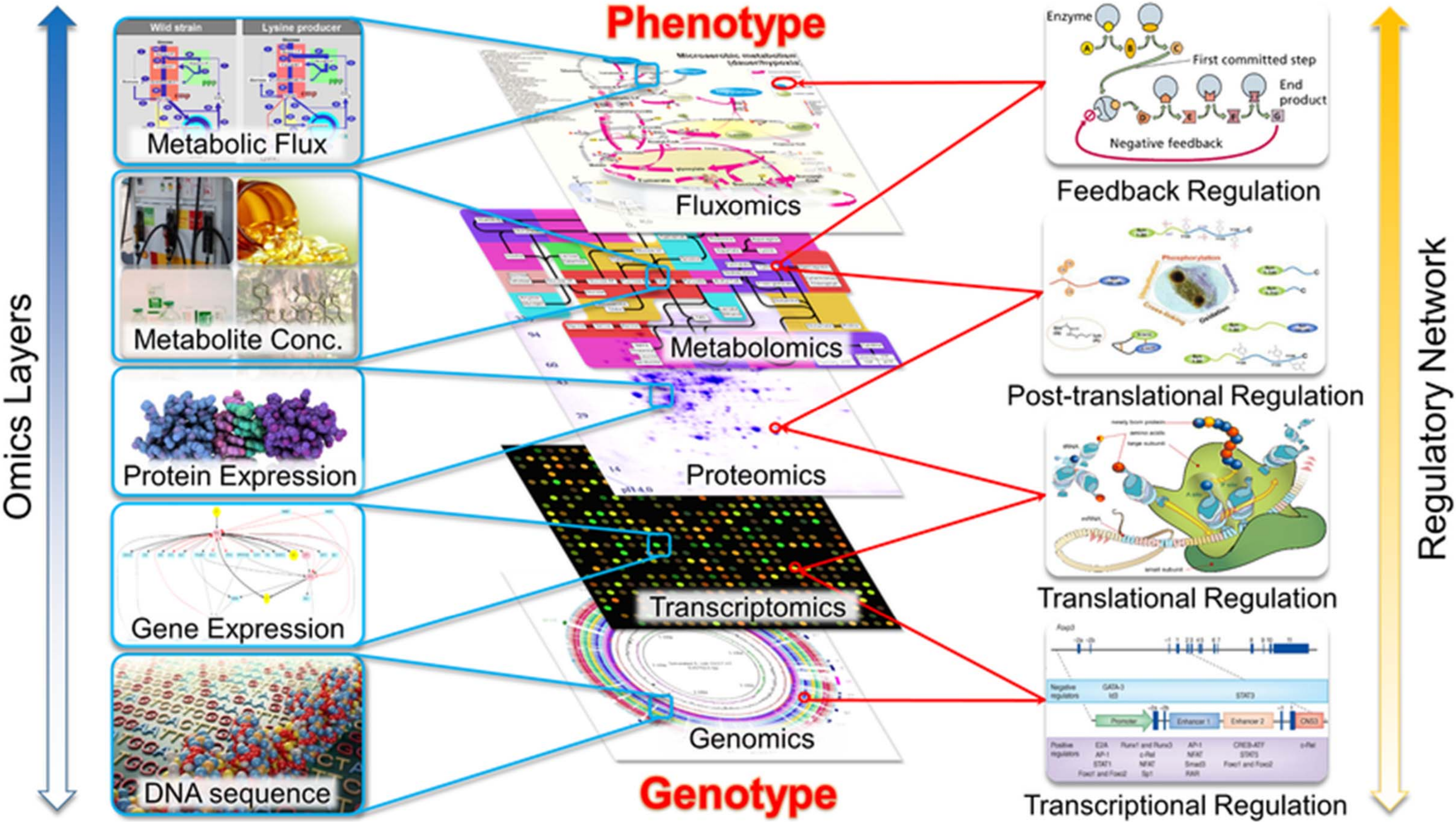
Is this gene associated with a disease?
Clinical Validity

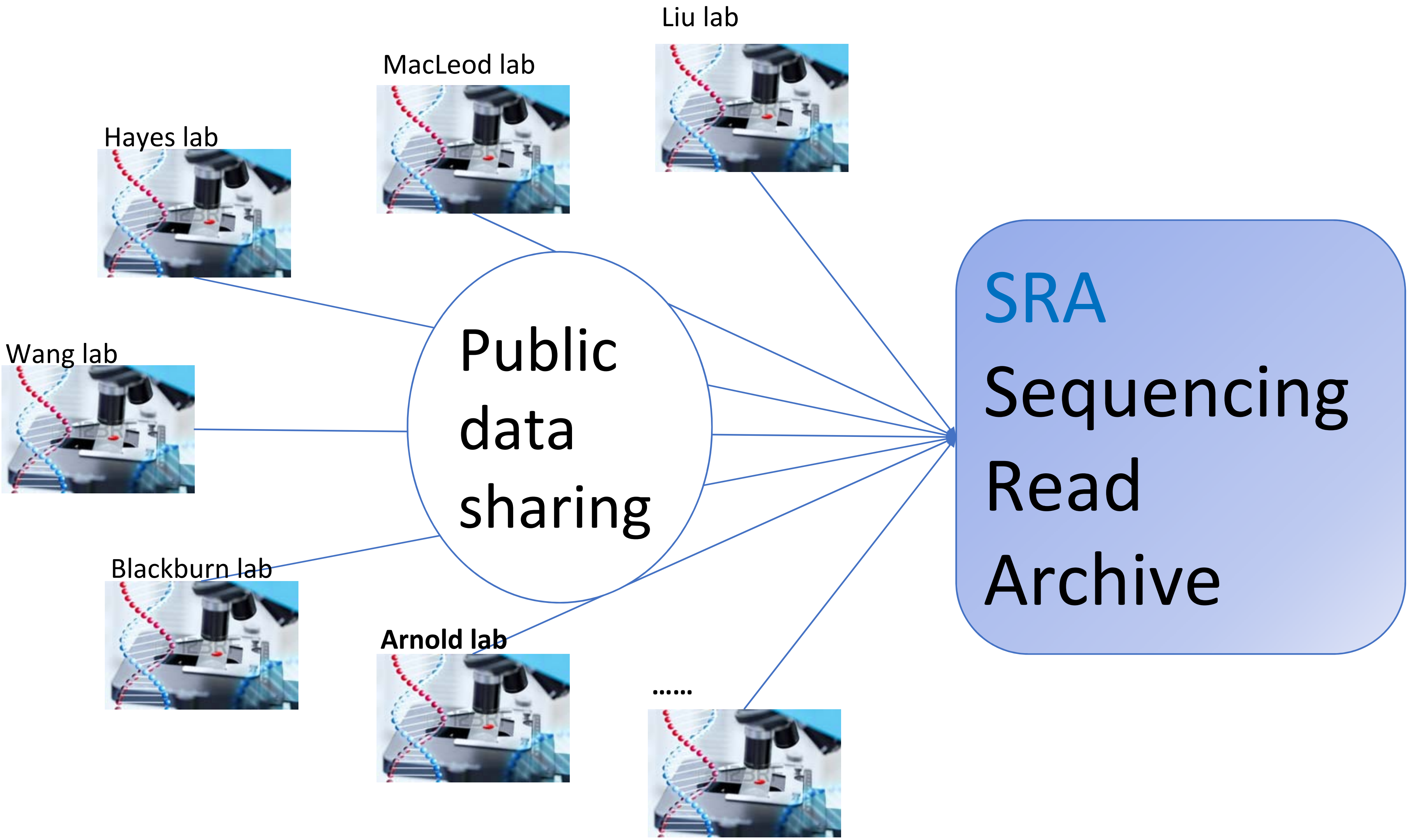
Is this variant causative?
Pathogenicity

Is this information actionable?
Clinical Utility

Building a Genomic Knowledge Base
ClinVar & Other Resources

Improved Patient Care
Through Genomic Medicine


Omics Layers
Phenotype
Wild strain
Lysine producer
Metabolic Flux
Fluxomics
Metabolite Conc.
Metabolomics
Protein Expression
Proteomics
Gene Expression
Transcriptomics
DNA sequence
Genomics
Genotype
Enzyme
First committed step
End product
Negative feedback
Feedback Regulation
Phosphorylation
Post-translational Regulation
Translational Regulation
Negative regulators
Promoter
Enhancer 1
Enhancer 2
Positive regulators
Transcriptional Regulation
Regulatory Network



Hayes lab

MacLeod lab

Liu lab

Wang lab

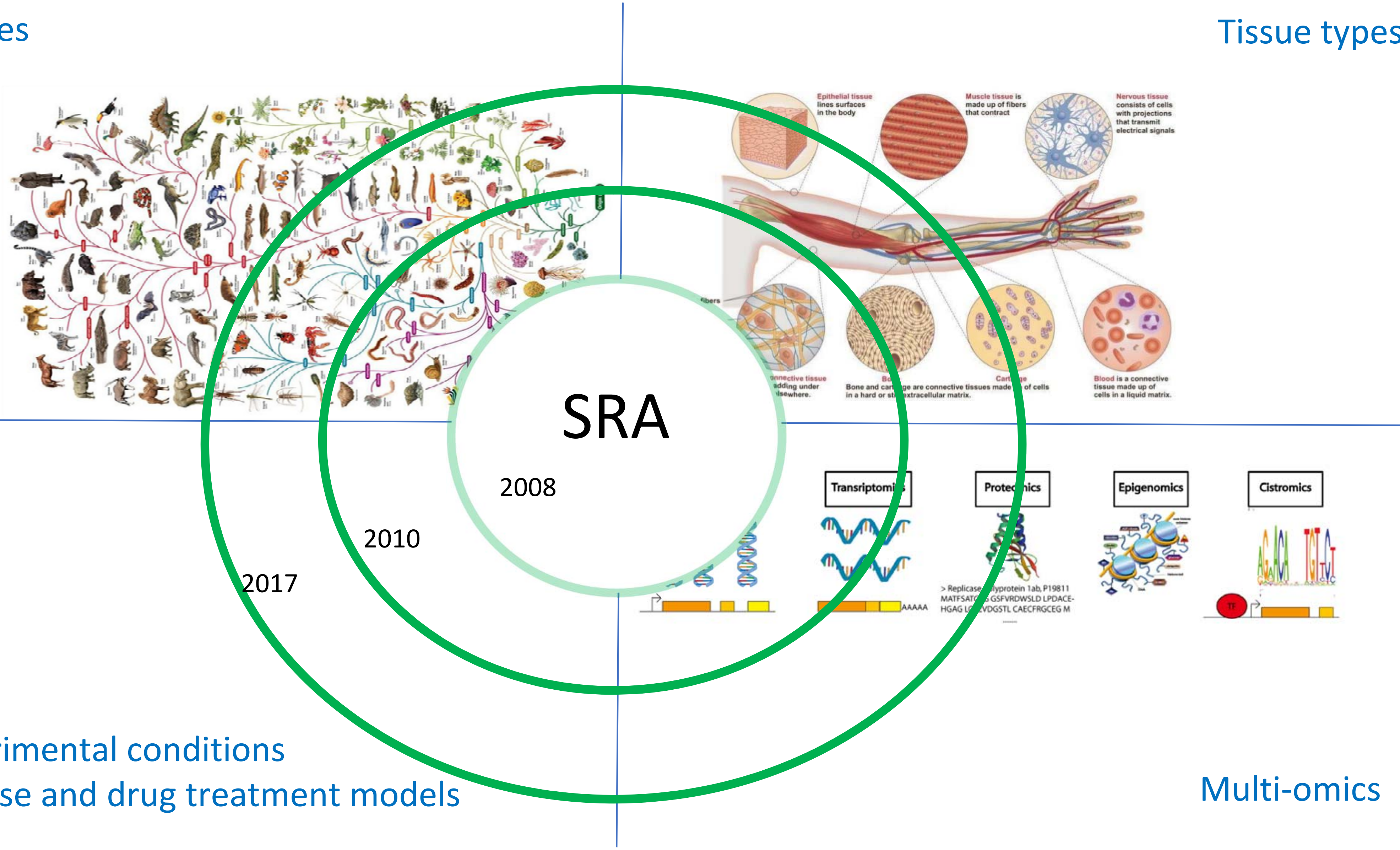
Blackburn lab

Arnold lab

……

Public data sharing

SRA Sequencing Read Archive



Species

Tissue types

SRA

2008

2010

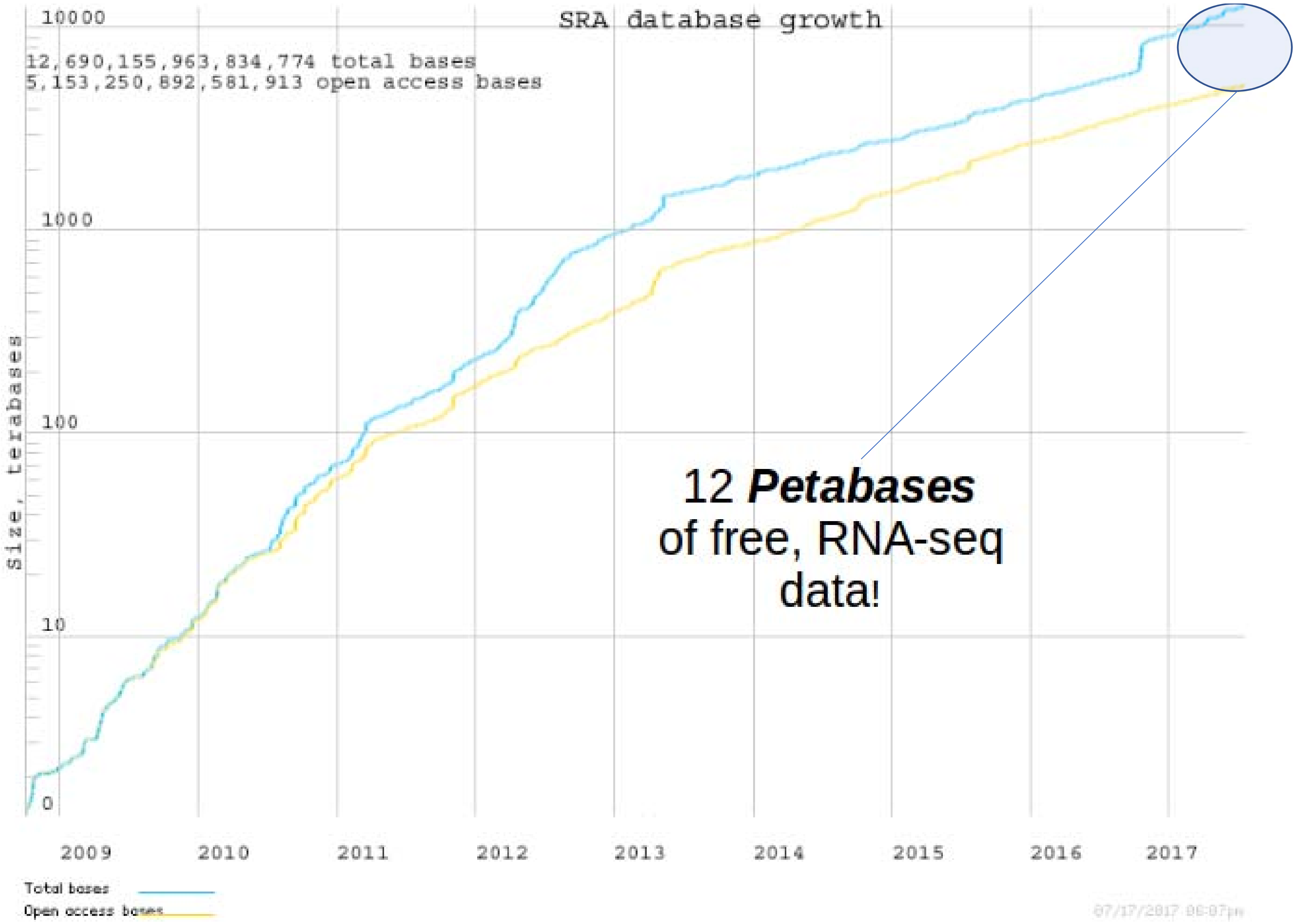
2017

Experimental conditions

Disease and drug treatment models

Multi-omics

Epithelial tissue lines surfaces in the body
Muscle tissue is made up of fibers that contract
Nervous tissue consists of cells with projections that transmit electrical signals
Connective tissue
Bone and cartilage are connective tissues made of cells in a hard or stiff extracellular matrix.
Blood is a connective tissue made up of cells in a liquid matrix.
Transriptomics
Proteomics
Epigenomics
Cistromics
AAAAA
> Replicase polyprotein 1ab, P19811
MATFSATGSGSFVRDWSLD LPDACE-
HGAG LGEVDGSTL CAECFRGCEG M
TF



12 ***Petabases***
of free, RNA-seq
data!

- -Free
- -Information-rich
- -Only get bigger
- -Not searchable
- -Difficult to reuse

Computational algorithms to enable efficient sequence search over large scale sequencing data

What kind of query? Why important?

- Basic query:
  - Given a biological sequence, what experimental samples contain it?

- Many questions can be answered with such a query
  - Retroactive study on patients with a newly discovered actionable mutation.
  - Tissue specificity of a novel transcript
  - Reannotation of a new reference genome
    - Currently annotating the latest version of horse genome

Can we download the data and process?

- A single sequencing data is about 10G-20G.
- Downloading it take 1 to 2 hours depending on your internet speed.
- Analyzing one such dataset typically takes 4 or more hours with well setup bioinformatics pipelines.
- Currently sequencing pipeline is developed for analyzing one file at a time, and it is reference-dependent.

- This may work with 10 samples.
- But does not scale with dozens, 100s, 1000s, or even more.

Representation of data – word-doc format

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Representation of sequencing data – *k*-mers

```
ATTCGAAGCTAG
ATTCG
 TTCGA
  TCGAA
   CGAAG
    GAAGC
     AAGCT
      AGCTA
       GCTAG
```

Sequence Containment Query

SRA3214

Kmer	Frequency
ATTCG	32
TCGAA	546
TTCGA	111
…	…
AAGCT	1311
GCTAG	56
GGCAA	37
AGCTA	3
CGAAG	19

Transcript Query to Kmer

```
ATTCGAAGCTAG
ATTCG
 TTCGA
  TCGAA
   CGAAG
    GAAGC
     AAGCT
      AGCTA
       GCTAG
```

Transcript Present in SRA3214?

$$\frac{Kmer\ Matches}{|Transcript|} = \frac{7}{8} \geq\ .85$$

- Raw RNA-seq experiments represented by their *k*-mer content
- Transcript 'present' in experiment if the proportion of *k*-mer matches greater than the given threshold, θ

Challenges Toward Answering the Sequence Coverage Query

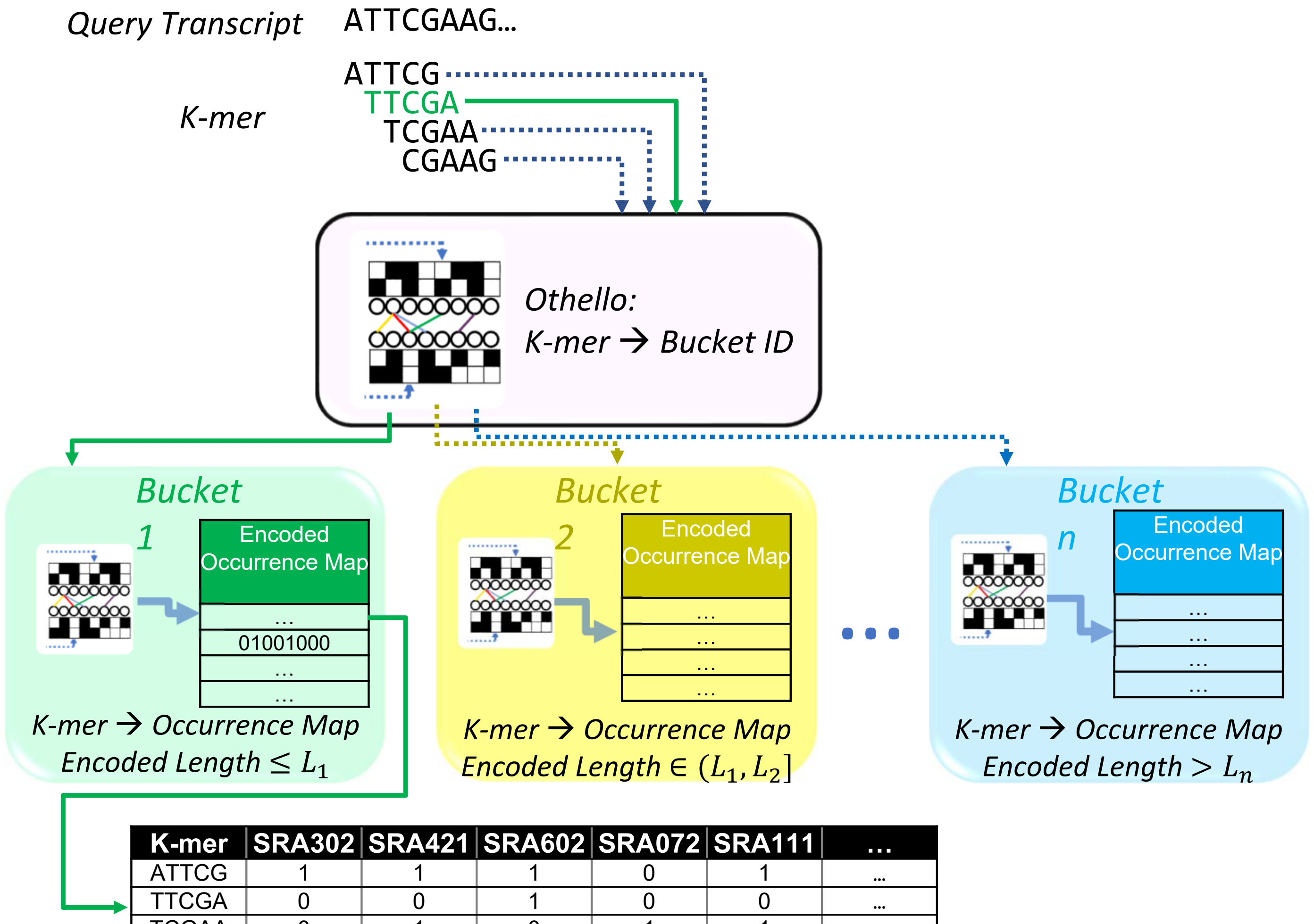
$\sim 10^3\ Samples$

$\sim 10^9\ kmers$

Kmer	SRA302	SRA421	⋯	SRA072	SRA111
TTCGA	1	1	⋯	0	1
GCAGG	1	0	⋯	0	0
AAAGT	0	1	⋯	1	1
⋮	⋮	⋮	⋱	⋮	⋮
CACTT	0	1	⋯	1	0
CGTGC	1	1	⋯	1	1

- Focus on retrieving individual *k*-mer occurrence across all samples
- Index to facilitate fast access to *k*-mer occurrence information is key
- Must scale to *thousands* of samples and potentially *billions* of *k*-mers
  - Regular hashtable with k-mers as key will not be efficient in both memory and speed

SeqOthello



Query Transcript ATTCGAAG...

K-mer

ATTCG
TTCGA
TCGAA
CGAAG

SeqOthello Level 1

Othello:
K-mer → Bucket ID

SeqOthello Level 2

Bucket 1
Encoded Occurrence Map
…
01001000
…
…
K-mer → Occurrence Map
Encoded Length $\leq L_1$

Bucket 2
Encoded Occurrence Map
…
…
…
…
K-mer → Occurrence Map
Encoded Length $\in (L_1, L_2]$

• • •

Bucket n
Encoded Occurrence Map
…
…
…
…
K-mer → Occurrence Map
Encoded Length $> L_n$

Query Result

K-mer	SRA302	SRA421	SRA602	SRA072	SRA111	…
ATTCG	1	1	1	0	1	…
TTCGA	0	0	1	0	0	…
TCGAA	0	1	0	1	1	…
CGAAG	0	0	1	1	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Summary	10%	30%	90%	80%	90%	…

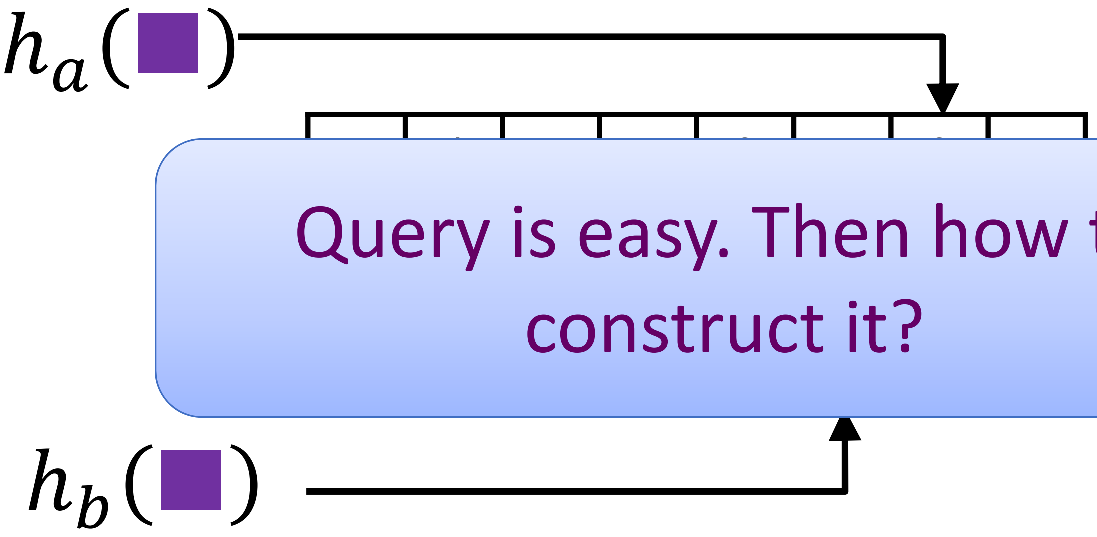
Othello - Fast *k*-mer search

- Basic version: Classifies keys to two sets X and Y
  - Equivalent to key lookups for a 1-bit value
- Query result
  - $\tau(k) = 0 \Leftrightarrow k \in X$
  - $\tau(k) = 1 \Leftrightarrow k \in Y$
- Advanced version: Classifies keys to 2^l sets
  - Equivalent to key lookups for a l-bit value

Othello Query Structure

- Two bitmaps a, b with size m_a, m_b $(m_a + m_b < 4n)$

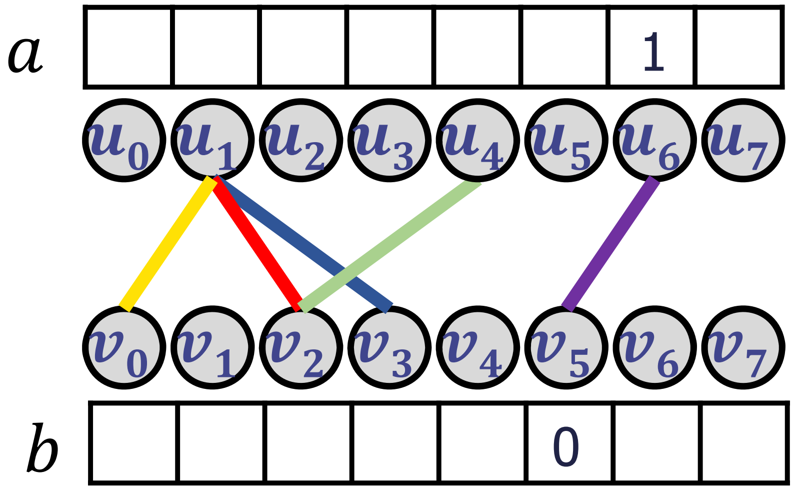
n is # of keys



Query is easy. Then how to construct it?

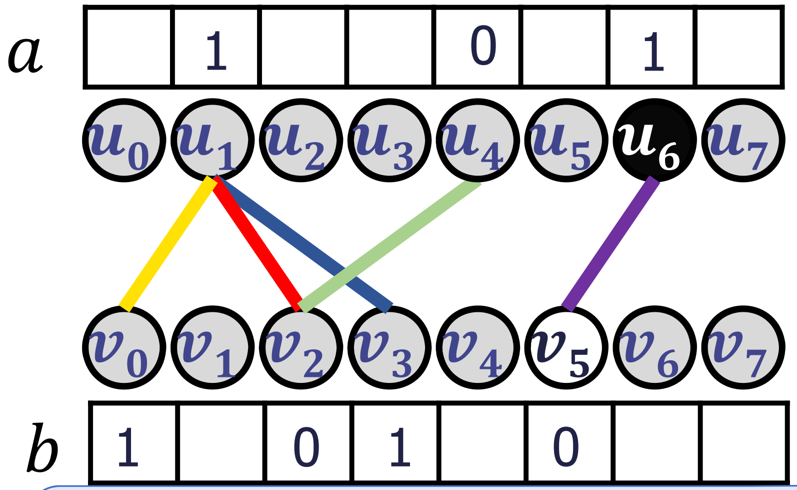
🟪 is in set *Y*

Othello Control Structure: Compute Bitmap


a
1
u_0 u_1 u_2 u_3 u_4 u_5 u_6 u_7
v_0 v_1 v_2 v_3 v_4 v_5 v_6 v_7
b
0

k	$h_a(k)$	$h_b(k)$	set
🟪	6	5	Y
🟨	1	0	
🟥	1	2	
🟦	1	3	
🟩	4	2	

Compute Bitmap

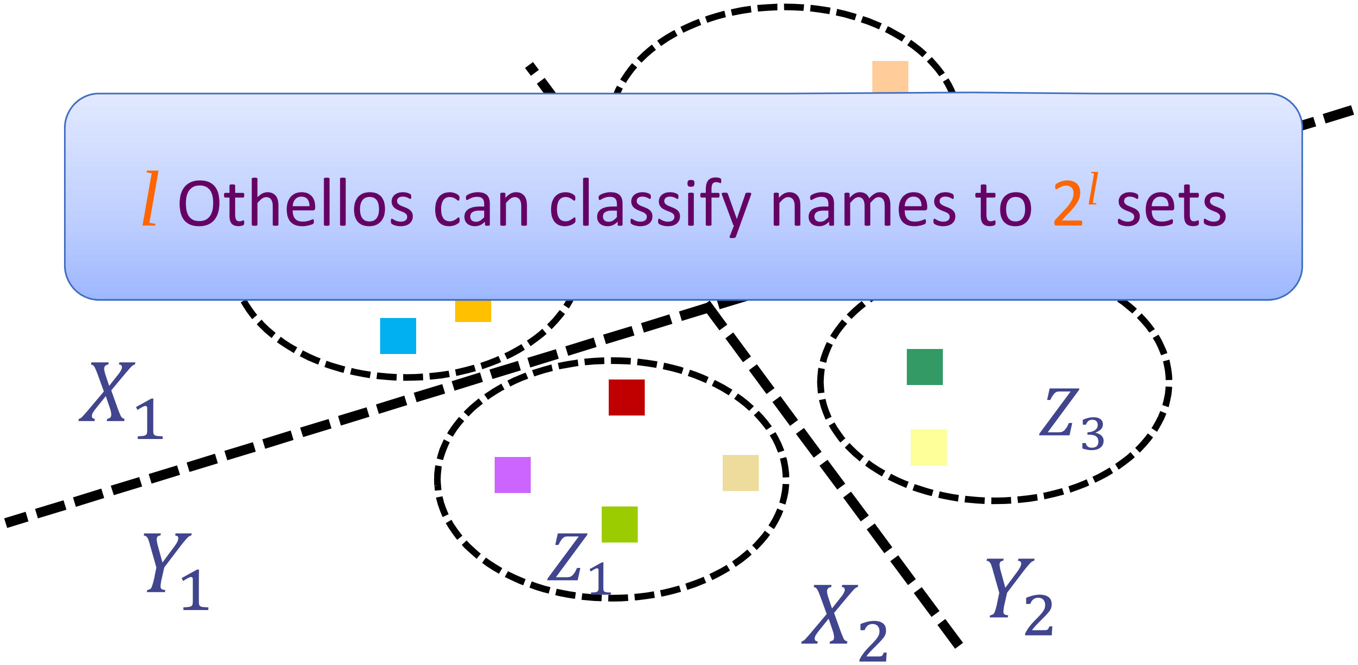


k	$h_a(k)$	$h_b(k)$	set
🟪	6	5	Y
🟨	1	0	X
🟥	1	2	Y
🟦	1	3	X
🟩	4	2	X

If *G* is acyclic, easy to find a coloring plan

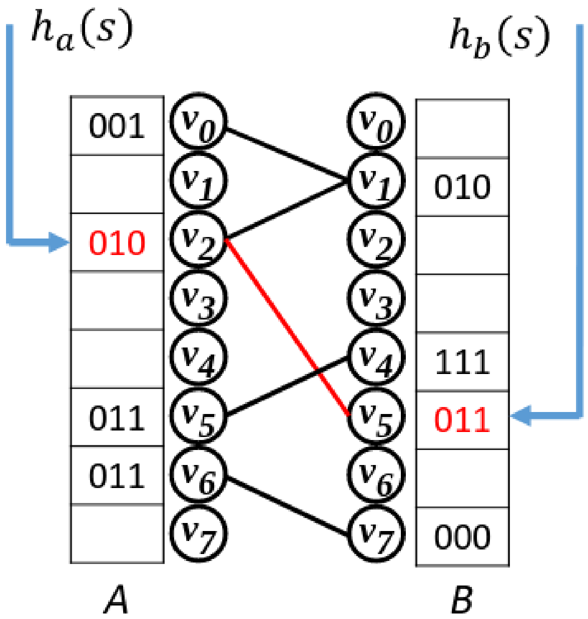
l-Othello functionality

- Classifies names into 2^l sets: $Z_0, Z_1, \cdots, Z_{2^l-1}$



l Othellos can classify names to 2^l sets

l-Othello: Fast and Memory Efficient Hashing Classifier

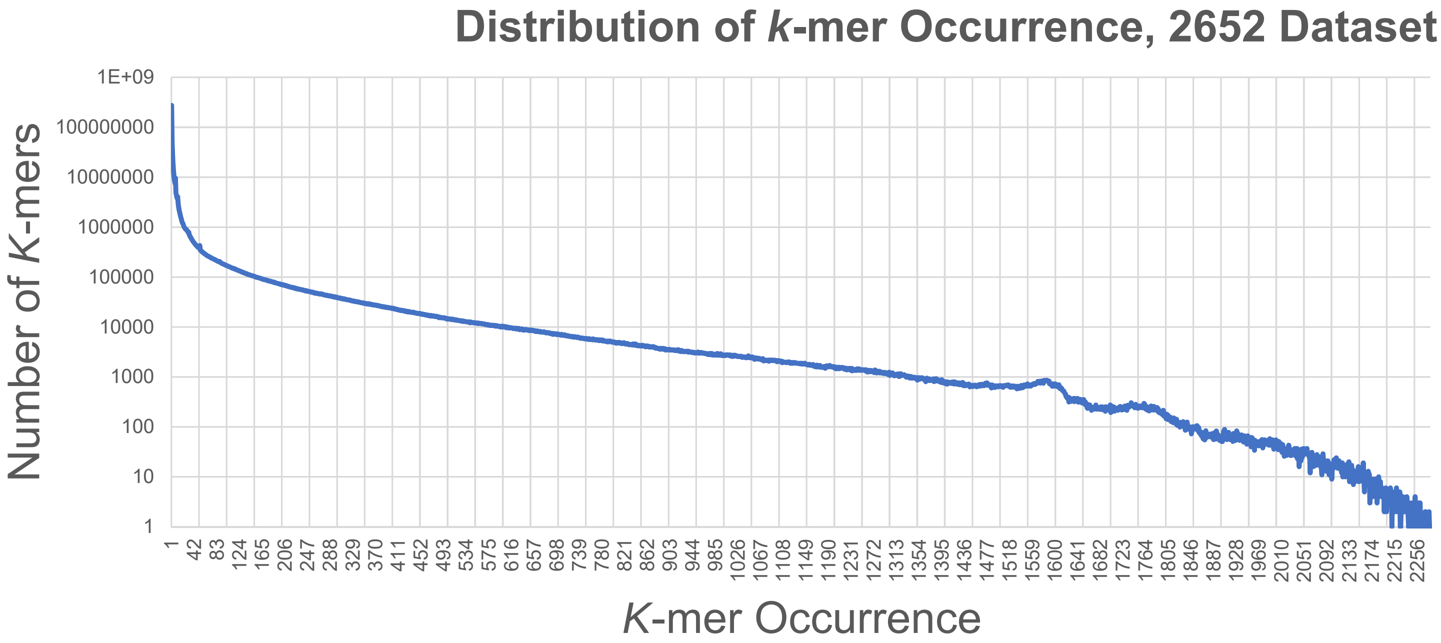


$$\tau(s) = A[h_a(s)] \oplus B[h_b(s)]$$

$$\begin{array}{rr} & 010 \\ \oplus & 011 \\ \hline & 001 \end{array}$$

- Time Complexity: Construction → O(*n*), Query → O(1)
- Memory Complexity: Construction → [2.67*n*, 4*n*], Query → 4*ln*
- Supports indexing at both levels of SeqOthello

Compression using matrix sparsity



- Approximately 90% of *k*-mers occur in less than 1% samples (Based on datasets used in the experiment)

SeqOthello: A ‘Horizontal’ Partitioning Approach

***k*-mer Occurrence Map**

Kmer	SRA302	SRA421	SRA602	SRA072	SRA111
TTCGA	1	1	1	0	1
GCAGG	0	0	1	0	0
AAAGT	0	1	0	1	1
CCTGA	0	0	1	1	0
CACTT	0	1	0	1	0
CGTGC	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮
TCGCT	1	0	1	1	1
GTAAC	0	0	0	0	1
AGGAA	0	0	0	1	0
TTTTC	1	0	0	0	0
CAAAG	1	0	1	1	0

- Leverage sparse distribution by partitioning *k*-mers into frequency bins

SeqOthello: A ‘Horizontal’ Partitioning Approach

***k*-mer Occurrence Map**

Kmer	SRA302	SRA421	SRA602	SRA072	SRA111
TTCGA	1	1	1	0	1
GCAGG	0	0	1	0	0
AAAGT	0	1	0	1	1
CCTGA	0	0	1	1	0
CACTT	0	1	0	1	0
CGTGC	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮
TCGCT	1	0	1	1	1
GTAAC	0	0	0	0	1
AGGAA	0	0	0	1	0
TTTTC	1	0	0	0	0
CAAAG	1	0	1	1	0

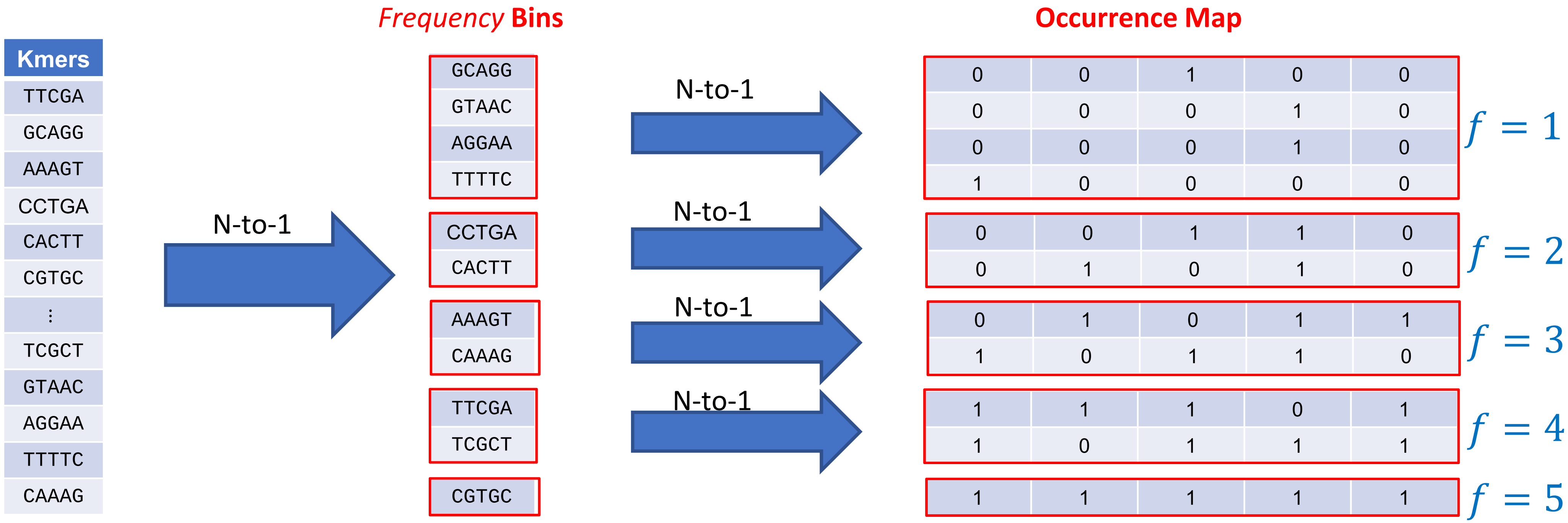
Sort by frequency

Occurrence Map

GCAGG	0	0	1	0	0	$f = 1$
GTAAC	0	0	0	1	0	
AGGAA	0	0	0	1	0	
TTTTC	1	0	0	0	0	
CCTGA	0	0	1	1	0	$f = 2$
CACTT	0	1	0	1	0	
AAAGT	0	1	0	1	1	$f = 3$
CAAAG	1	0	1	1	0	
TTCGA	1	1	1	0	1	$f = 4$
TCGCT	1	0	1	1	1	
CGTGC	1	1	1	1	1	$f = 5$

- Leverage sparse distribution by partitioning *k*-mers into frequency bins
- Horizontal compression of occurrence information
- Improved search locality

SeqOthello: Two Tiered Search Structure



Kmers
TTCGA
GCAGG
AAAGT
CCTGA
CACTT
CGTGC
⋮
TCGCT
GTAAC
AGGAA
TTTTC
CAAAG

N-to-1

Frequency **Bins**

GCAGG
GTAAC
AGGAA
TTTTC

CCTGA
CACTT

AAAGT
CAAAG

TTCGA
TCGCT

CGTGC

N-to-1

N-to-1

N-to-1

N-to-1

Occurrence Map

0	0	1	0	0
0	0	0	1	0
0	0	0	1	0
1	0	0	0	0

$f = 1$

0	0	1	1	0
0	1	0	1	0

$f = 2$

0	1	0	1	1
1	0	1	1	0

$f = 3$

1	1	1	0	1
1	0	1	1	1

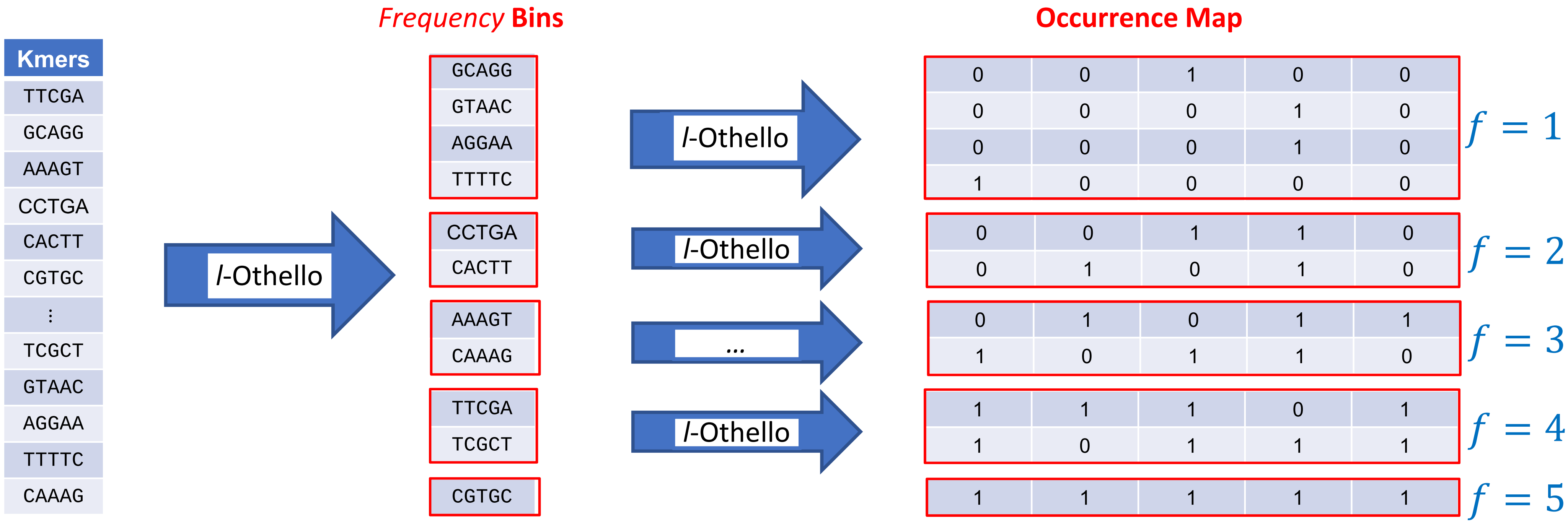
$f = 4$

1	1	1	1	1

$f = 5$

- Tier 1: N-to-1 mapping from constituent *k*-mers to *frequency bins*
- Tier 2: N-to-1 mapping from *k*-mers to their occurrence maps

SeqOthello: Two-Tiered Search Structure



Kmers
TTCGA
GCAGG
AAAGT
CCTGA
CACTT
CGTGC
⋮
TCGCT
GTAAC
AGGAA
TTTTC
CAAAG

Frequency **Bins**

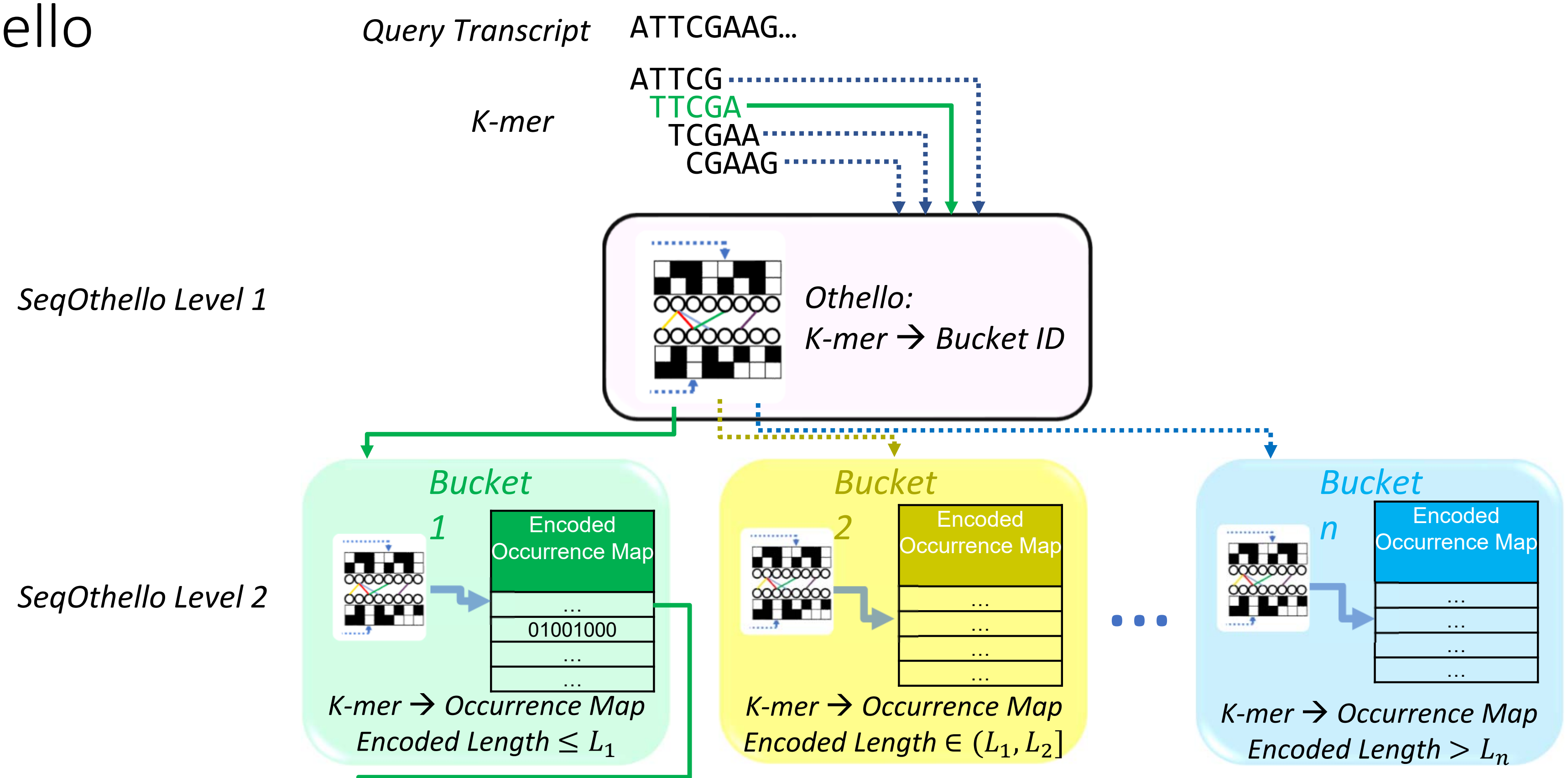
GCAGG GTAAC AGGAA TTTTC
CCTGA CACTT
AAAGT CAAAG
TTCGA TCGCT
CGTGC

Occurrence Map

0	0	1	0	0	$f=1$
0	0	0	1	0	
0	0	0	1	0	
1	0	0	0	0	
0	0	1	1	0	$f=2$
0	1	0	1	0	
0	1	0	1	1	$f=3$
1	0	1	1	0	
1	1	1	0	1	$f=4$
1	0	1	1	1	
1	1	1	1	1	$f=5$

- N-to-1 mapping is achieved by multi-class hashing classifier *l*-othello

SeqOthello



Query Transcript ATTCGAAG…

K-mer

ATTCG
TTCGA
TCGAA
CGAAG

SeqOthello Level 1

Othello:
K-mer → Bucket ID

SeqOthello Level 2

Bucket 1

Encoded Occurrence Map
…
01001000
…
…

K-mer → Occurrence Map
Encoded Length $\leq L_1$

Bucket 2

Encoded Occurrence Map
…
…
…
…

K-mer → Occurrence Map
Encoded Length $\in (L_1, L_2]$

…

Bucket n

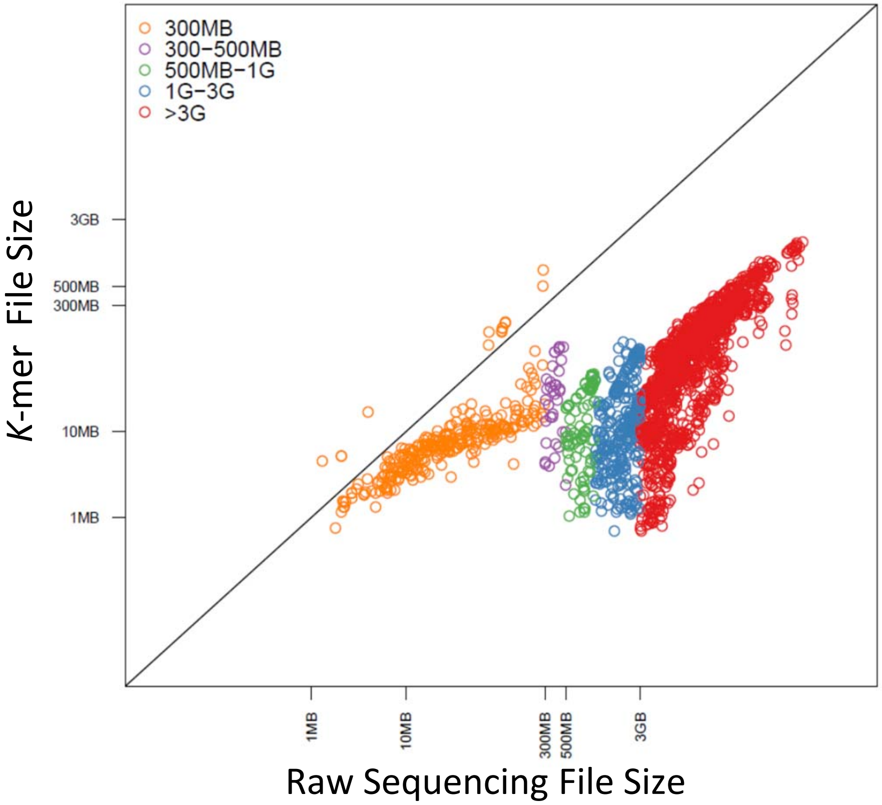
Encoded Occurrence Map
…
…
…
…

K-mer → Occurrence Map
Encoded Length $> L_n$

Query Result

K-mer	SRA302	SRA421	SRA602	SRA072	SRA111	…
ATTCG	1	1	1	0	1	…
TTCGA	0	0	1	0	0	…
TCGAA	0	1	0	1	1	…
CGAAG	0	0	1	1	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Summary	10%	30%	90%	80%	90%	…

Evaluation 2652 Human Datasets from SRA



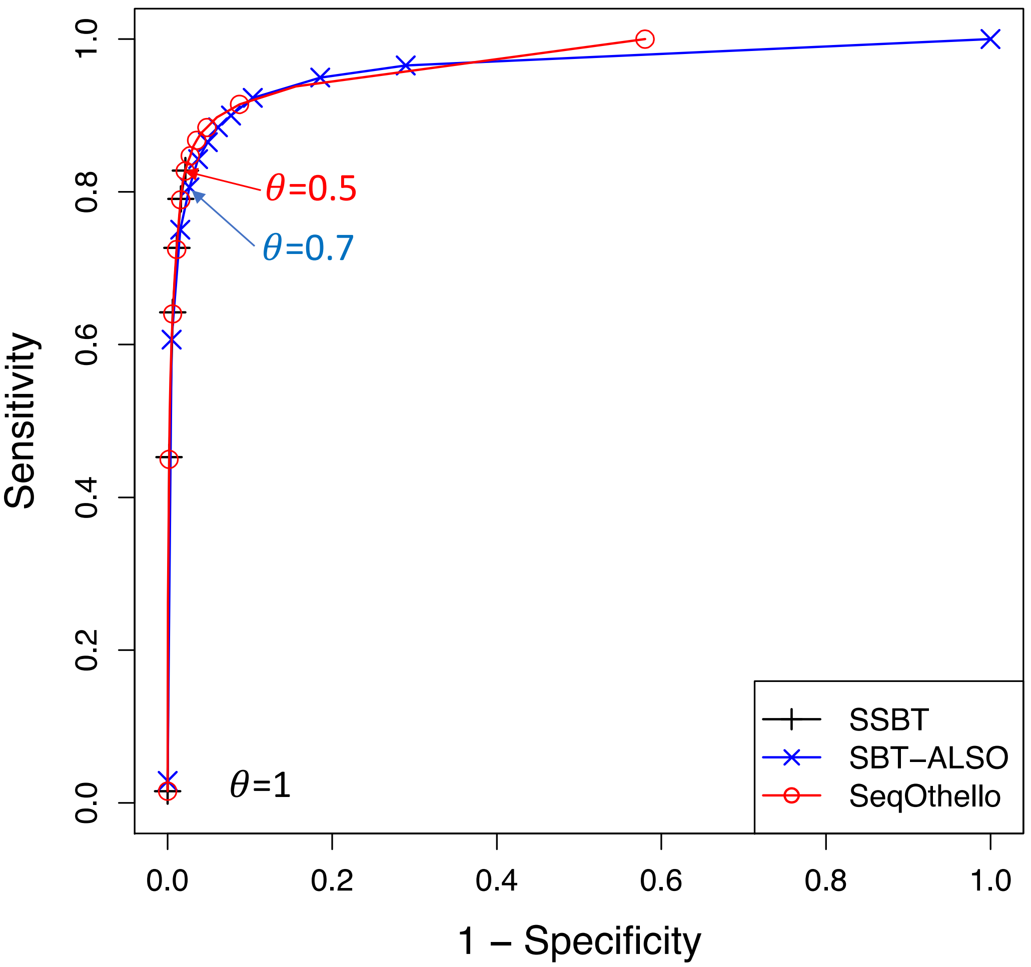
- RNA-Seq Datasets[1]: 2,652 sequences extracted from human brain, blood, and breast tissue Same set of samples used to evaluate SBT and variants
- Query Transcripts: GENCODE v25, includes 198,093 transcripts

Comparison configuration on sequence ontainment query

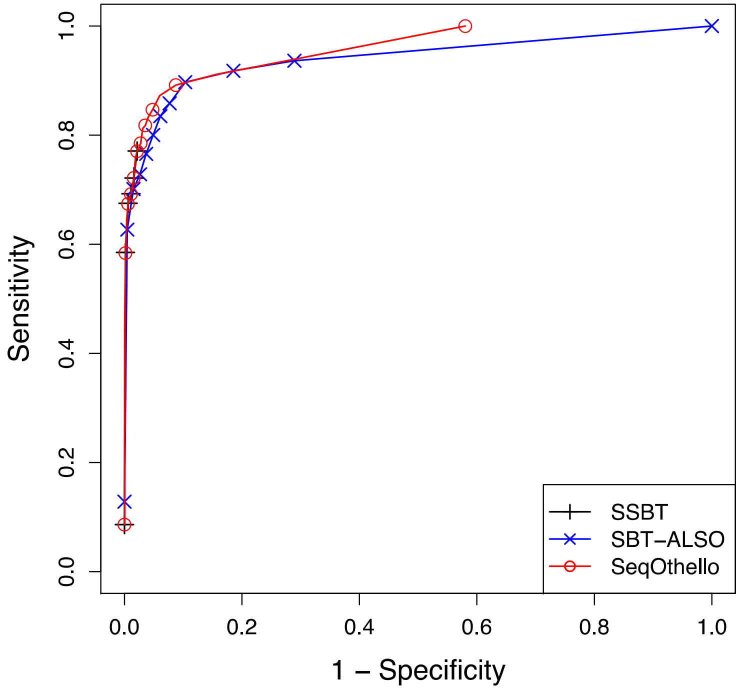
- SeqOthello
  - in development
  - k-mers are extracted from raw fastq datasets using Jellyfish following the k-mer count threshold applied in original SBT paper[1]
- SSBT[2]
  - Downloaded on Jun 16, 2017
  - Same k-mers as in SeqOthello
- SBT-ALSO[3]
  - Downloaded in April, 2017
  - Constructed using the bloom filters downloaded from SBT-SK software and data following instruction in the original paper.
- Platform:
  - 32 cores (4 x Intel E5-4640 2.4 GHz 8 cores)
  - 512GB memory
  - 4T NLSAS disk Linux OS (RHEL)

Accuracy of Query

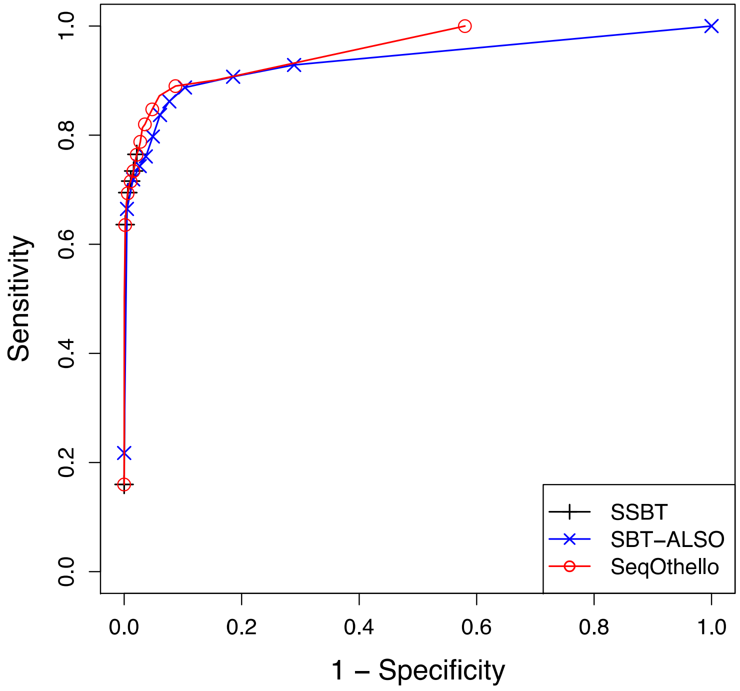
- Dataset – D200
  - 200 experiments, a representative subset of 2,652 experiments
- Transcripts
  - 34,608 transcripts, representing the entire gene set
- Ground Truth
  - Positive: Expression > 100 TPM
  - Negative: Expression < 1 TPM
  - Expression is estimated by Sailfish



More ROCs with different TPM cutoffs



Expression >500 TPM



Expression >1000 TPM

Construction Performance for D200

	Build Map (min)	Peak Mem (GB)	Disk Space (GB)	Index (GB)
SeqOthello	**10.3**	10.3	**48**	**2.5**
SSBT	121.3	**5.6**	141	1.0
SBT-ALSO	47.6	7.0	294	15.0

Query Performance for D200 of 34k Transcripts

Tools	θ	Memory (GB)	Time (min)
SeqOthello	-	**4.5**	**2.3**
SBT-ALSO	0.7	12.1	41.7
	0.8	11.7	29.1
	0.9	10.9	25.0
	1.0	7.6	3.6
SSBT	0.7	1.6	276.9
	0.8	1.6	220.7
	0.9	1.6	140.4
	1.0	1.6	6.3

SeqOthello will output all k-mers present in a query transcript, irregardless of θ.

Construction Performance for D2652

	k-mer Prep (Days)	Build Map (Hours)	Peak Mem (GB)	Disk Space (TB)	Index (GB)
SeqOthello	**3.4**	**2.1**	30	**0.4**	**9**
SSBT	4.8	18.3	**6**	1.8	6.1
SBT-ALSO	-	7.3	39	3.8	177

Query Performance for D2652 with 190k queries

Tools	Thread	θ	**Memory (GB)**	**Time (Hour)**
SeqOthello	1	**-**	22	**1.36**
	4	-	24	**0.7**
	8	-	28	**0.4**
SBT-ALSO	-	0.8	67	12.5
	-	0.9	63	9.9
SSBT	-	0.9	4.8	>4 days

Conclusion

- Rich sequencing data have been accumulated in many biological communities and public data repositories such as SRA. Without a search capability, access to these large datasets is severely limited.
- SeqOthello facilitates on-demand sequence search across large sequencing dataset, leveraging a two-tiered indexing structure leveraging a fast hashing classifier.
  - Great Compression – 9G for over 20T raw RNA-seq data.
  - Fast speed – Query 190k transcripts over 2652 datasets in less than 30 minutes.
  - Can be easily deployed to cloud to increase the search space.

Acknowledgement

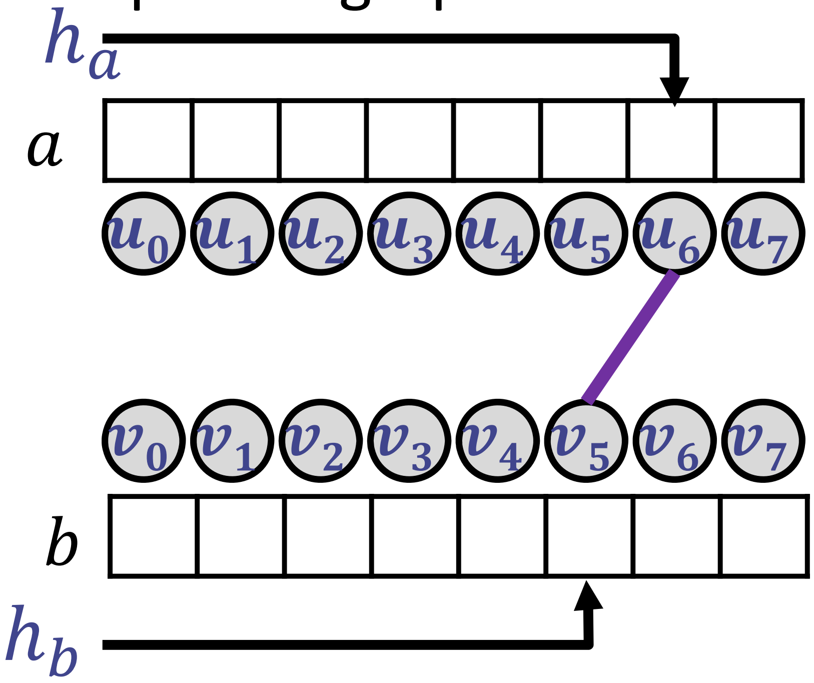
- Department of Computer Science at Uky
  - Ye Yu
  - Jinpeng Liu
  - Xinan Liu
  - Eamonn Magner

- CCS & HPC

- Department of Computer Science at UCSC
  - Qian Chen

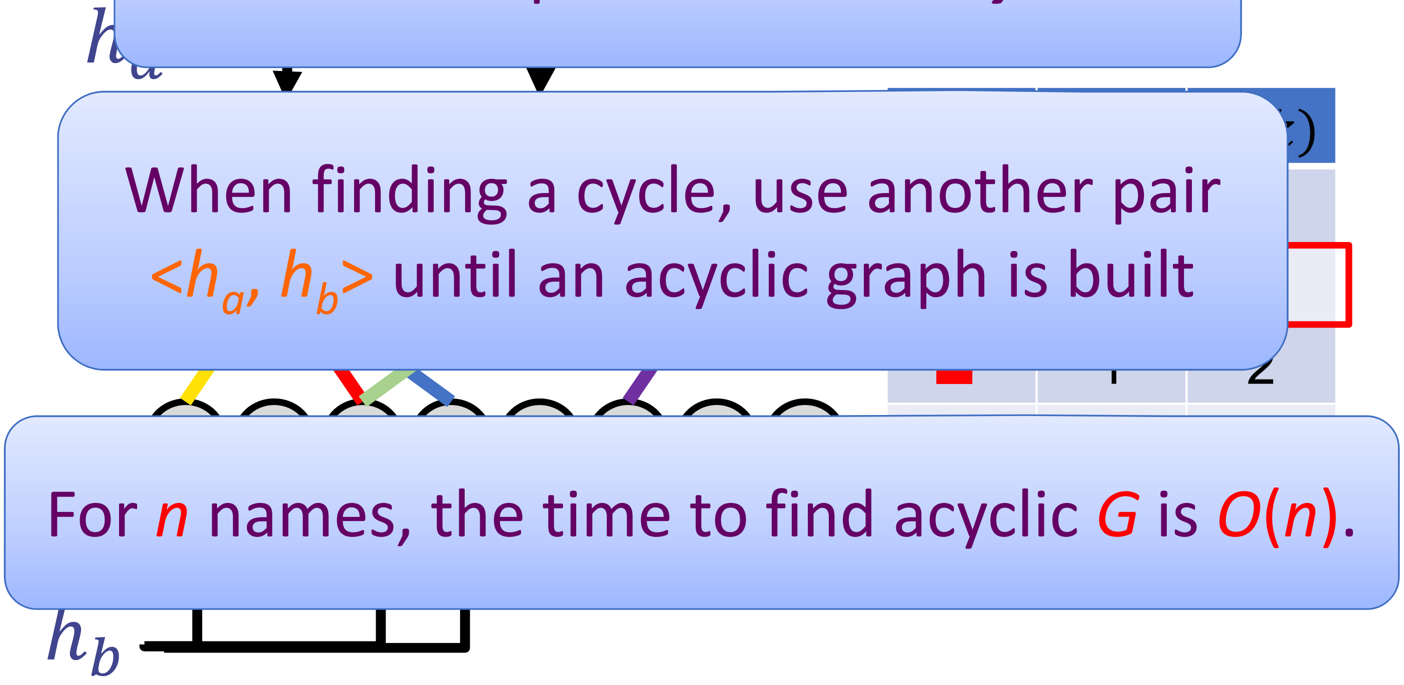
Othello Control Structure: Construct

- G: acyclic bipartite graph



k	$h_a(k)$	$h_b(k)$
🟪	6	5

Othello Control Structure: Construct

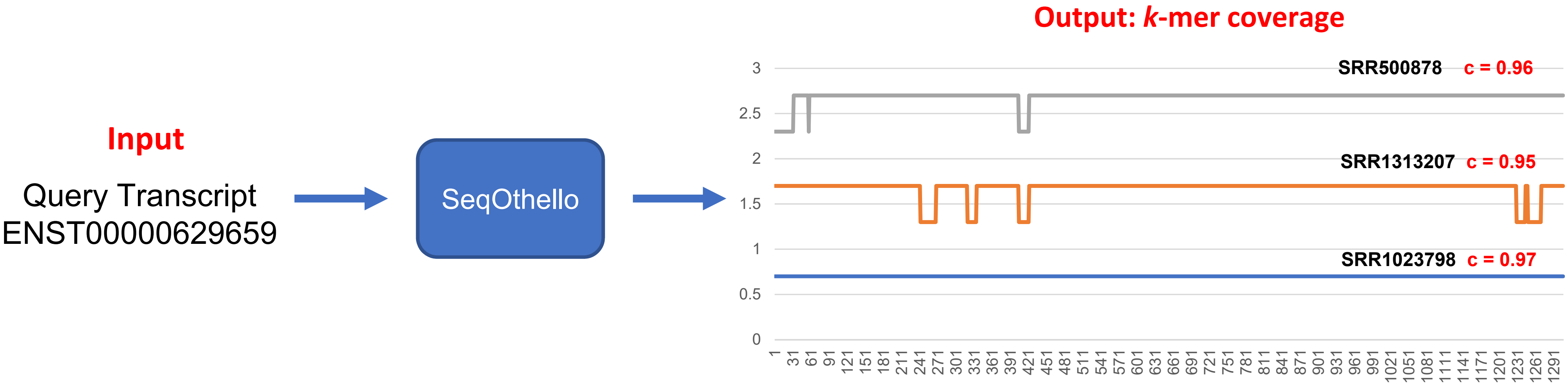


Othello requires G to be ***acyclic***.

When finding a cycle, use another pair $\langle h_a, h_b \rangle$ until an acyclic graph is built

For n names, the time to find acyclic G is $O(n)$.

Sequence Coverage Query



Input

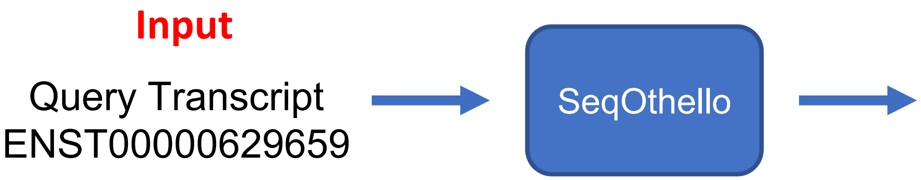
Query Transcript
ENST00000629659

SeqOthello

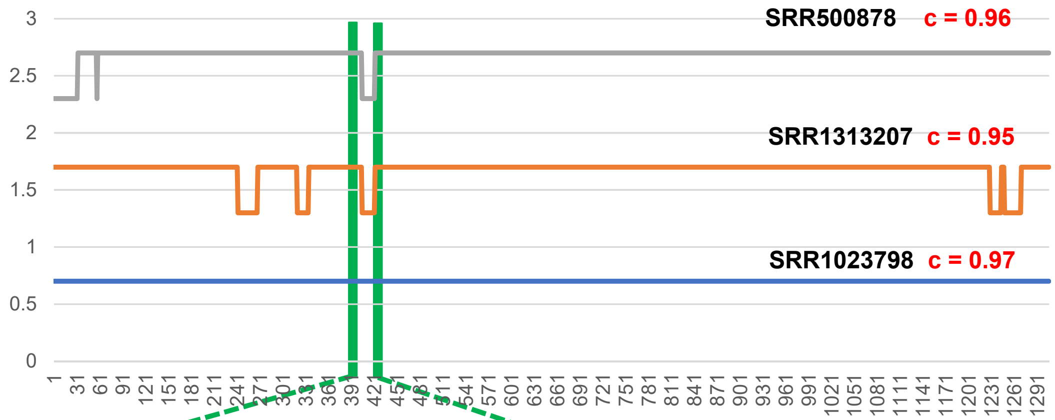
Output: *k*-mer coverage

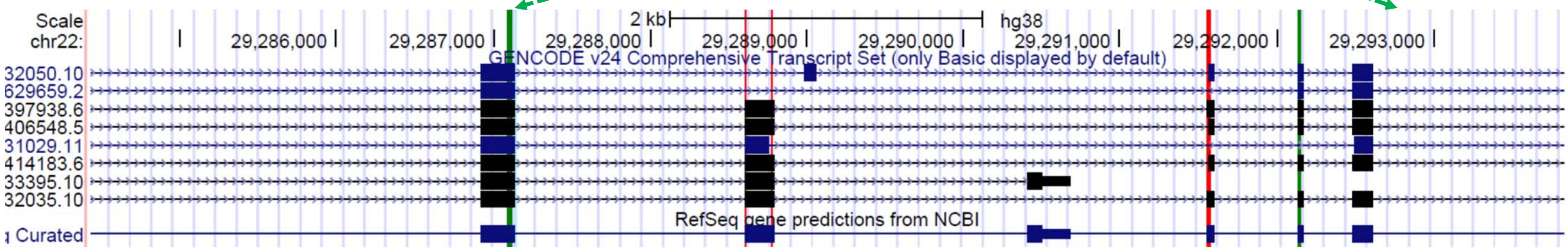
- Provides base-level k-mer coverage information with biological meanings
  - Not dependent on θ
  - differentiate alternative isoforms
  - detect mutations and regions possibly containing noise
  - an assessment of relative expression values

Sequence Coverage Query



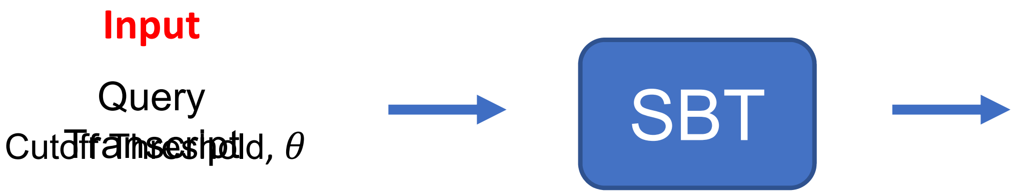
Output: *k*-mer coverage





Missing junction k-mers indicates absence of the query transcript even with high overall k-mer coverage. In fact, the transcript expressed below 1TPM in all three samples

SBT and the Sequence Containment Query



Output: Transcript Present in Sample?

θ	SRA30	SRA42	SRA95
0.0	Yes	Yes	Yes
.25	Yes	Yes	No
.50	No	Yes	No
.75	No	Yes	No
1.0	No	No	No

- Returns only sequencing experiments containing a significant portion (> θ) of query sequence Setting θ is difficult
  - How much k-mers are used to construct the index
  - How much mismatches one would like to tolerate
  - How long is the transcript
- Further processing requires downloading and reanalyzing raw data