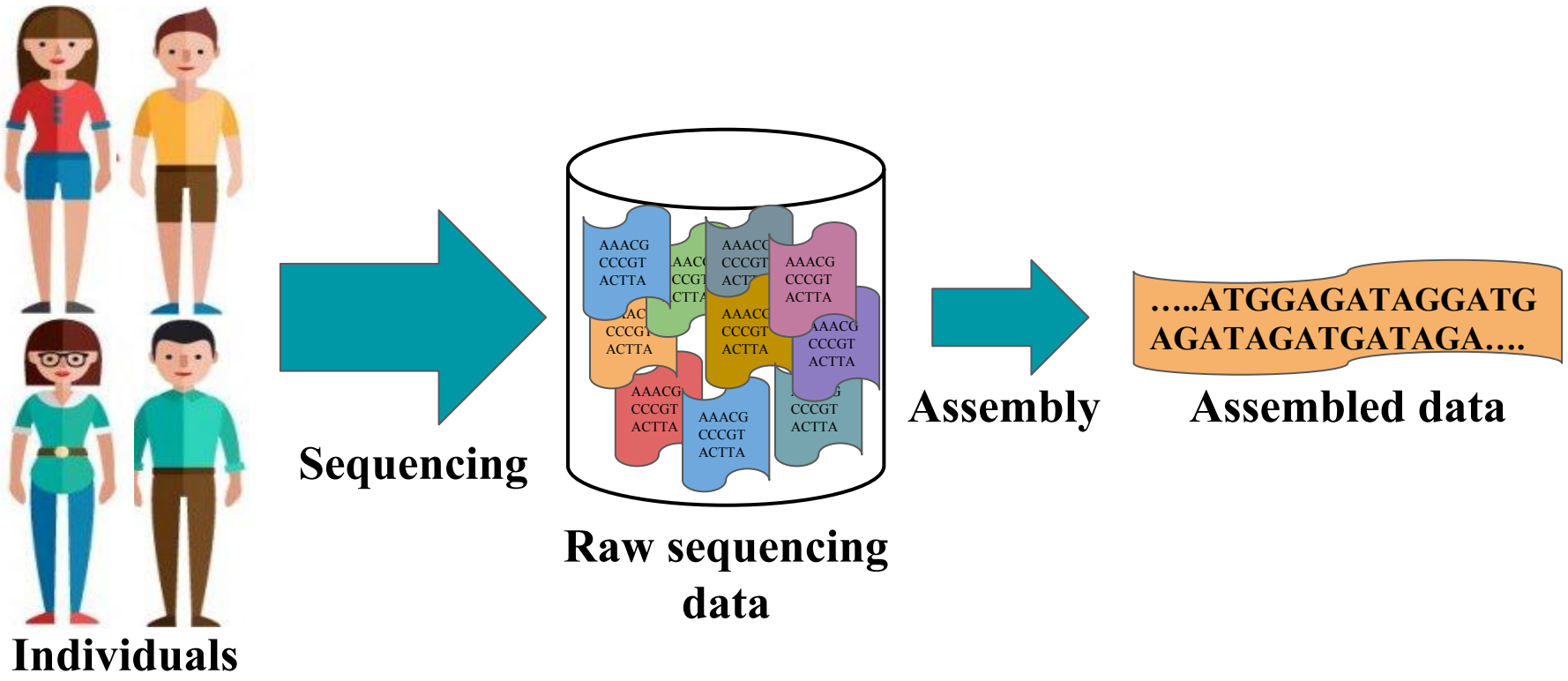
Mantis: A Fast, Small, and Exact Large-Scale Sequence-Search Index

Prashant Pandey, Fatemeh Almodaresi, Michael A. Bender,
Michael Ferdman, Rob Johnson*, Rob Patro
Stony Brook University, NY *VMware Research USA

A huge amount of information is available only in raw sequencing data



- Assembled data is hugely lossy**.** A lot of **variability information is lost during assembly**.
- And a lot of raw sequencing data never gets assembled at all.

The ability to perform searches on raw sequencing data would enable us to answer lots of questions

Q: What if I find a new putative disease-related transcript, and want to see if it appeared in other biological samples?

Q : What if I discover a new fusion event in a particular cancer subtype and want to know if it is common among samples with this subtype?

Q: What if I find an unexpected bacterial contaminant in my data; which other samples might contain this?

The ability to perform searches on raw sequencing data would enable us to answer lots of questions

Q: What if I find a new putative disease-related transcript, and want to see if it appeared in other biological samples?

Q : What if I discover a new fusion event in a particular cancer subtype and want to know if it is common among samples with this subtype?

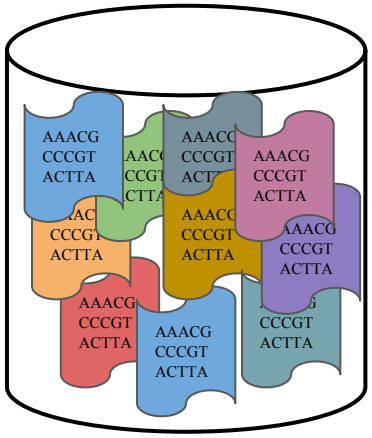
Q: What if I find an unexpected bacterial contaminant in my data; which other samples might contain this?

A: I need to search through tons of raw sequencing data.

Current tools (i.e., BLAST) can't answer diversity questions easily

Individuals

Sequencing

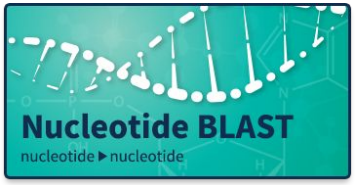


Raw sequencing data

Assembly

…..ATGGAGATAGGATG
AGATAGATGATAGA….

Assembled data



Current tools (i.e., BLAST) can't answer diversity questions easily

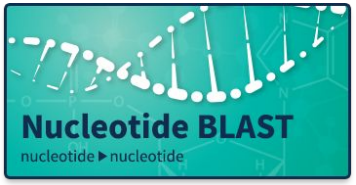
Individuals

Raw sequencing data

GATG
GA....

ata

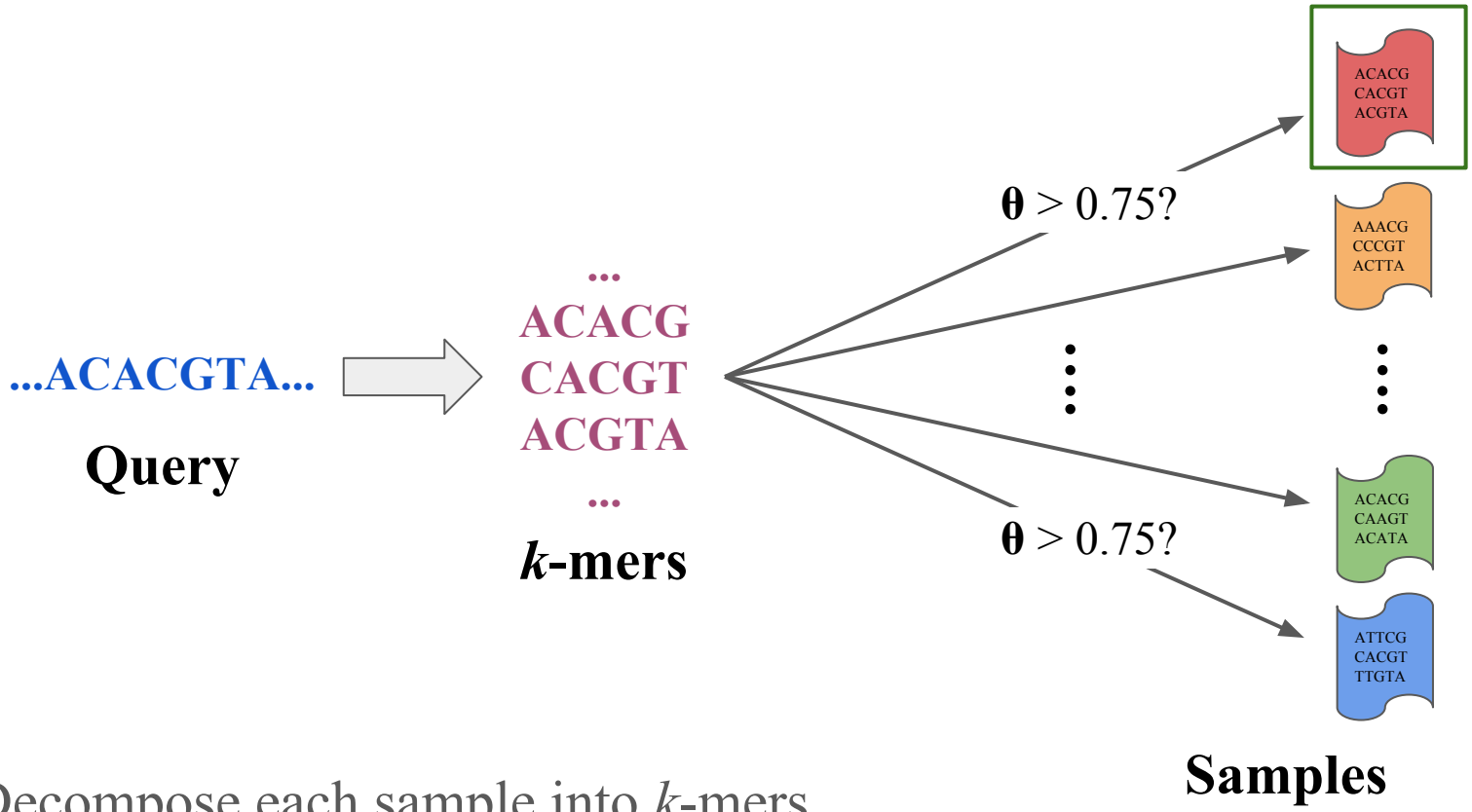
This renders what is otherwise an immensely valuable public resource **largely inert**.



Our Answer: Mantis

- A system to **index and search through large collections of raw sequencing samples**.
- Mantis uses **new data structures** to enable:
  - **6X faster index construction** than the state-of-the-art.
  - **6X-100X faster searches** than the state-of-the-art.
  - **20% smaller** index size.
  - **Exact results**, i.e., no false-positives or -negatives.
- Mantis is also a **colored de Bruijn graph:**
  - **Fast graph traversal**
  - **Topological analyses**

The sample discovery problem [Solomon and Kingsford]



- Decompose each sample into k-mers.
- If more than θ-fraction k-mers from a query appear in an sample then there is a high chance that query appears in that sample.

Existing tools for sample discovery problem

- **SBT:** Solomon and Kingsford 2016
- **SSBT:** Solomon and Kingsford 2017
- **AllSome SBT:** Sun et al. 2017

Existing tools for sample discovery problem

- **SBT:** Solomon and Kingsford 2016
- **SSBT:** Solomon and Kingsford 2017
- **AllSome SBT:** Sun et al. 2017

- All these tools use Bloom filters to represent *k*-mer content of samples.

Existing tools for sample discovery problem

- **SBT:** Solomon and Kingsford 2016
- **SSBT:** Solomon and Kingsford 2017
- **AllSome SBT:** Sun et al. 2017

- All these tools use Bloom filters to represent *k*-mer content of samples.
- Using Bloom filter saves a lot of space but results contain false-positives.

Existing tools for sample discovery problem

- **SBT:** Solomon and Kingsford 2016
- **SSBT:** Solomon and Kingsford 2017
- **AllSome SBT:** Sun et al. 2017

- All these tools use Bloom filters to represent *k*-mer content of samples.
- Using Bloom filter saves a lot of space but results contain false-positives.
- Also, all these tools have to work around the limitations of Bloom filters.

Mantis: A fundamentally different technique

Input Samples

S1	S2	S3	S4
	ACTG	ACTG	
ACTT			
		CTTG	CTTG
	TTTC	TTTC	
	GCGT	GCGT	GCGT
	AGCC	AGCC	

Map: *k*-mers to Samples

k-mer	Samples
ACTG	S2, S3
ACTT	S1
CTTG	S3, S4
TTTC	S2, S3
GCGT	S2, S3, S4
AGCC	S2, S3

- We want to map *k*-mers to the samples in which they appear.

Mantis: A fundamentally different technique

Input Samples

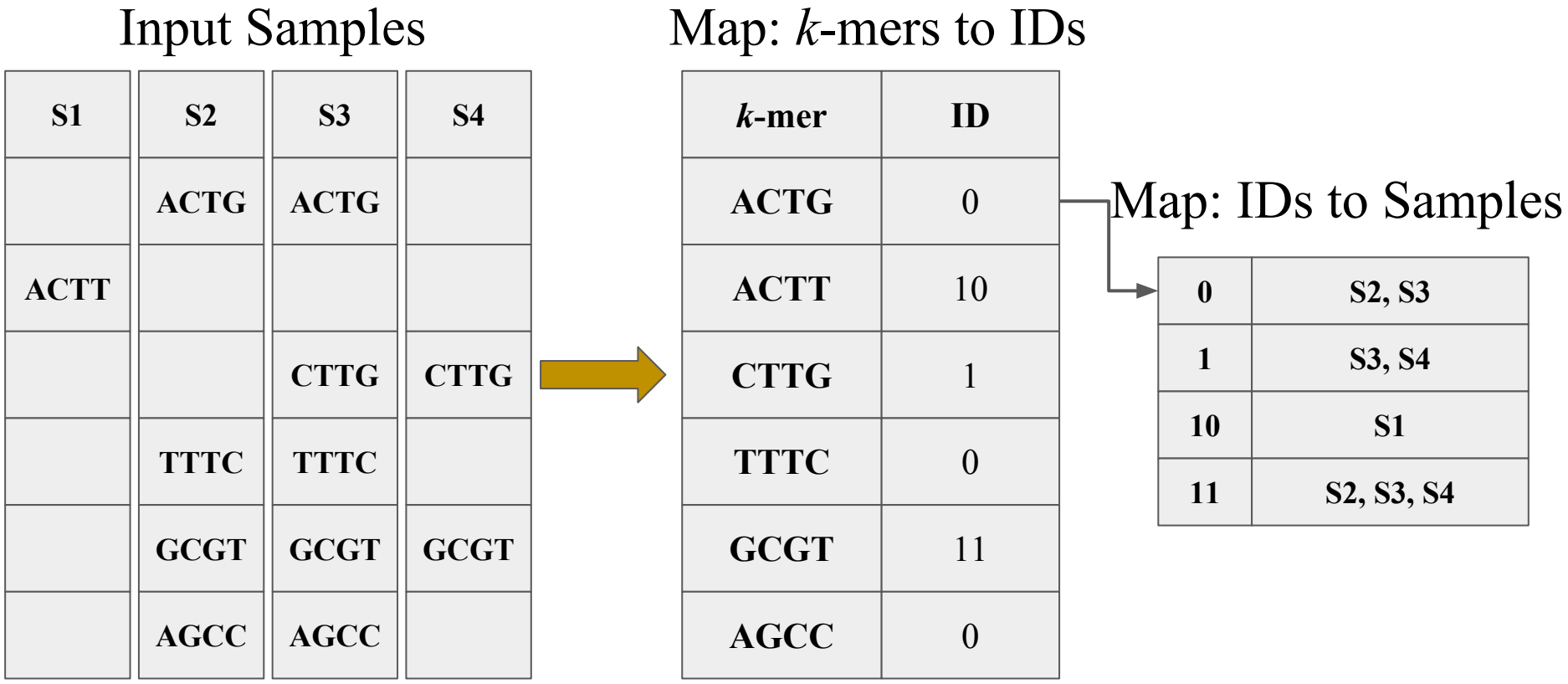
S1	S2	S3	S4
	ACTG	ACTG	
ACTT			
		CTTG	CTTG
	TTTC	TTTC	
	GCGT	GCGT	GCGT
	AGCC	AGCC	

Map: *k*-mers to Samples

k-mer	Samples
ACTG	S2, S3
ACTT	S1
CTTG	S3, S4
TTTC	S2, S3
GCGT	S2, S3, S4
AGCC	S2, S3

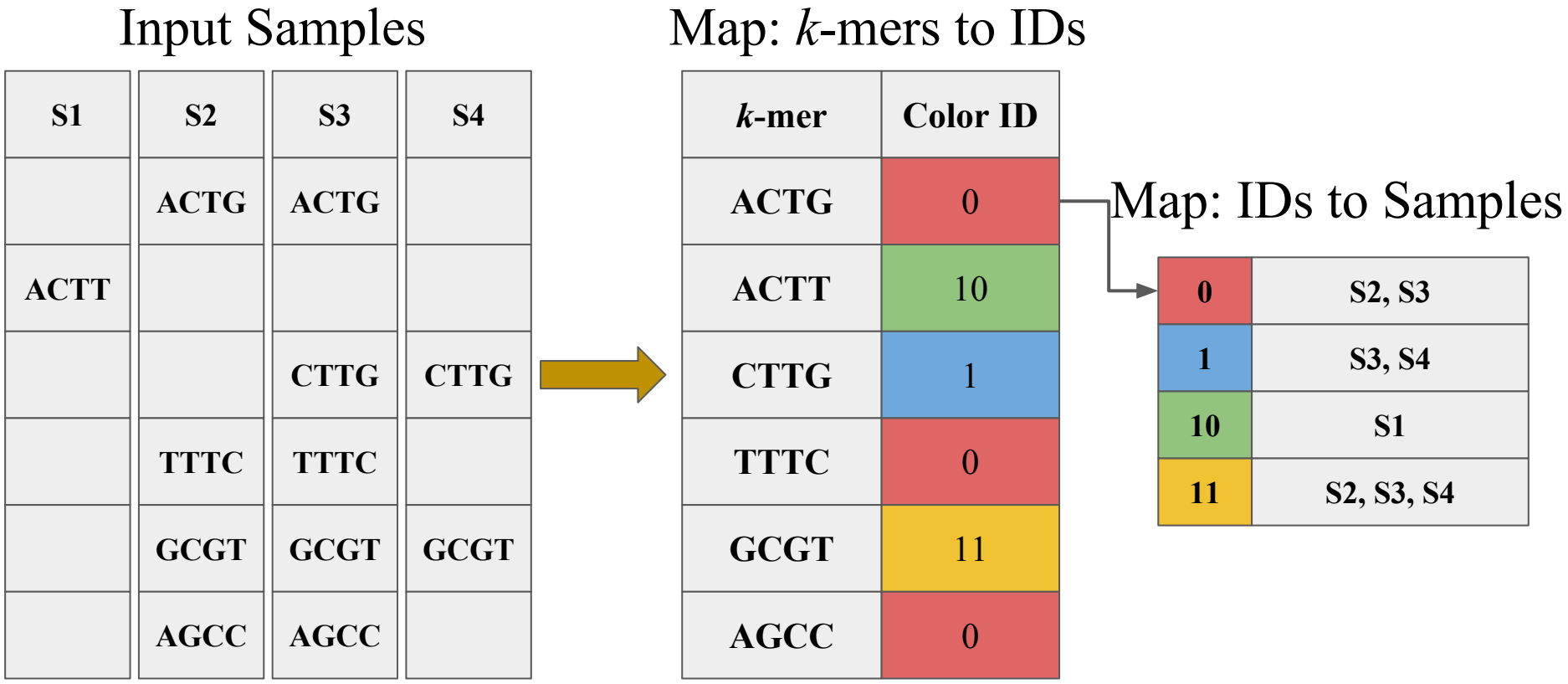
- There is an inherent redundancy in this this design.

Mantis: A fundamentally different technique



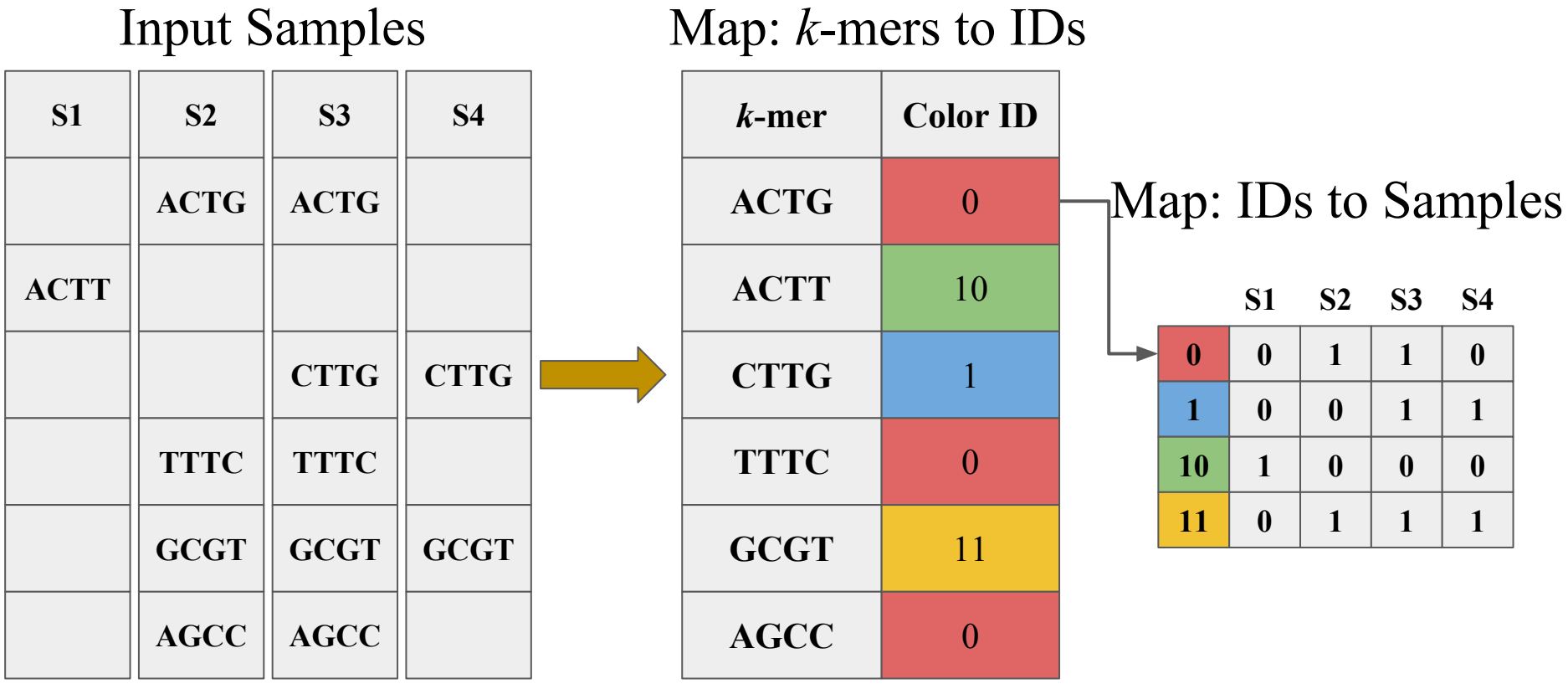
- We add another layer of indirection from IDs to sets of samples.

Mantis: A fundamentally different technique



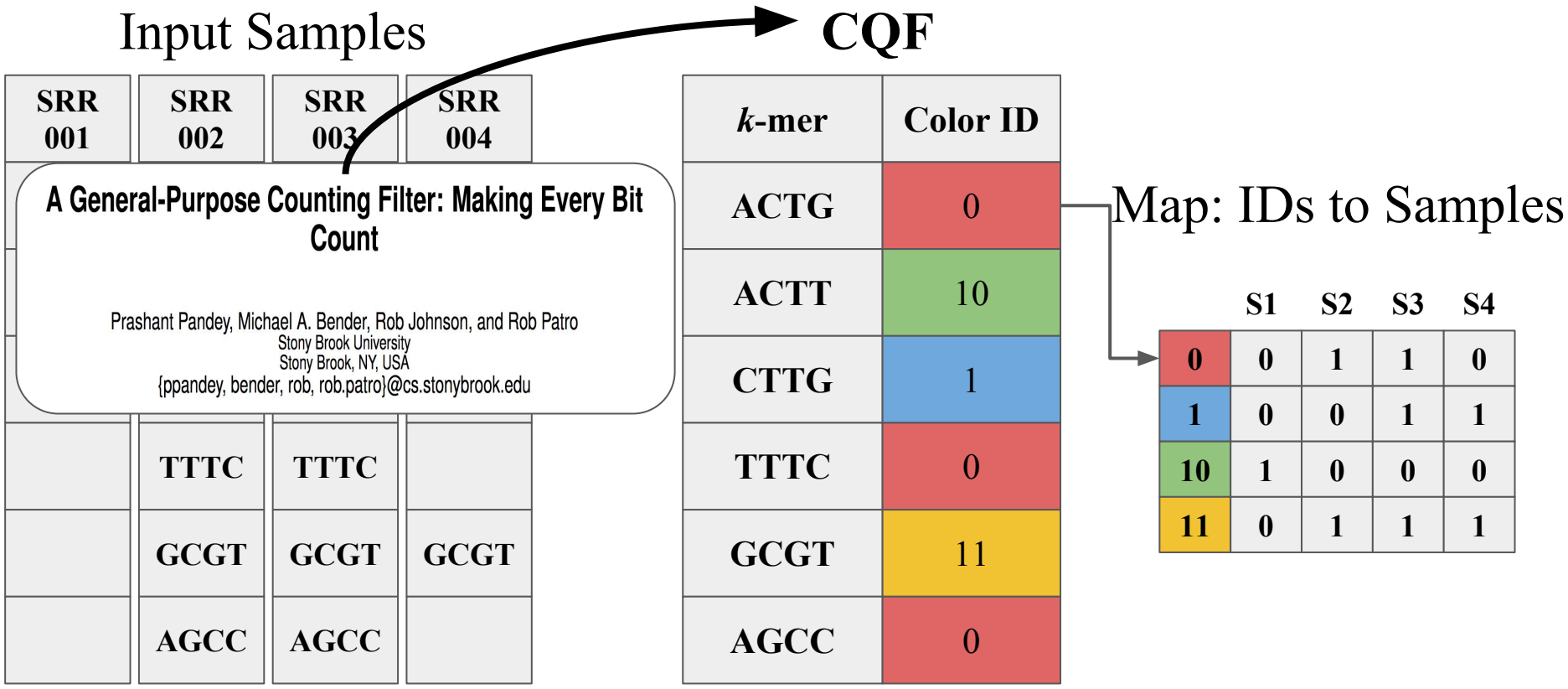
- We call each set a color class.

Mantis: A fundamentally different technique



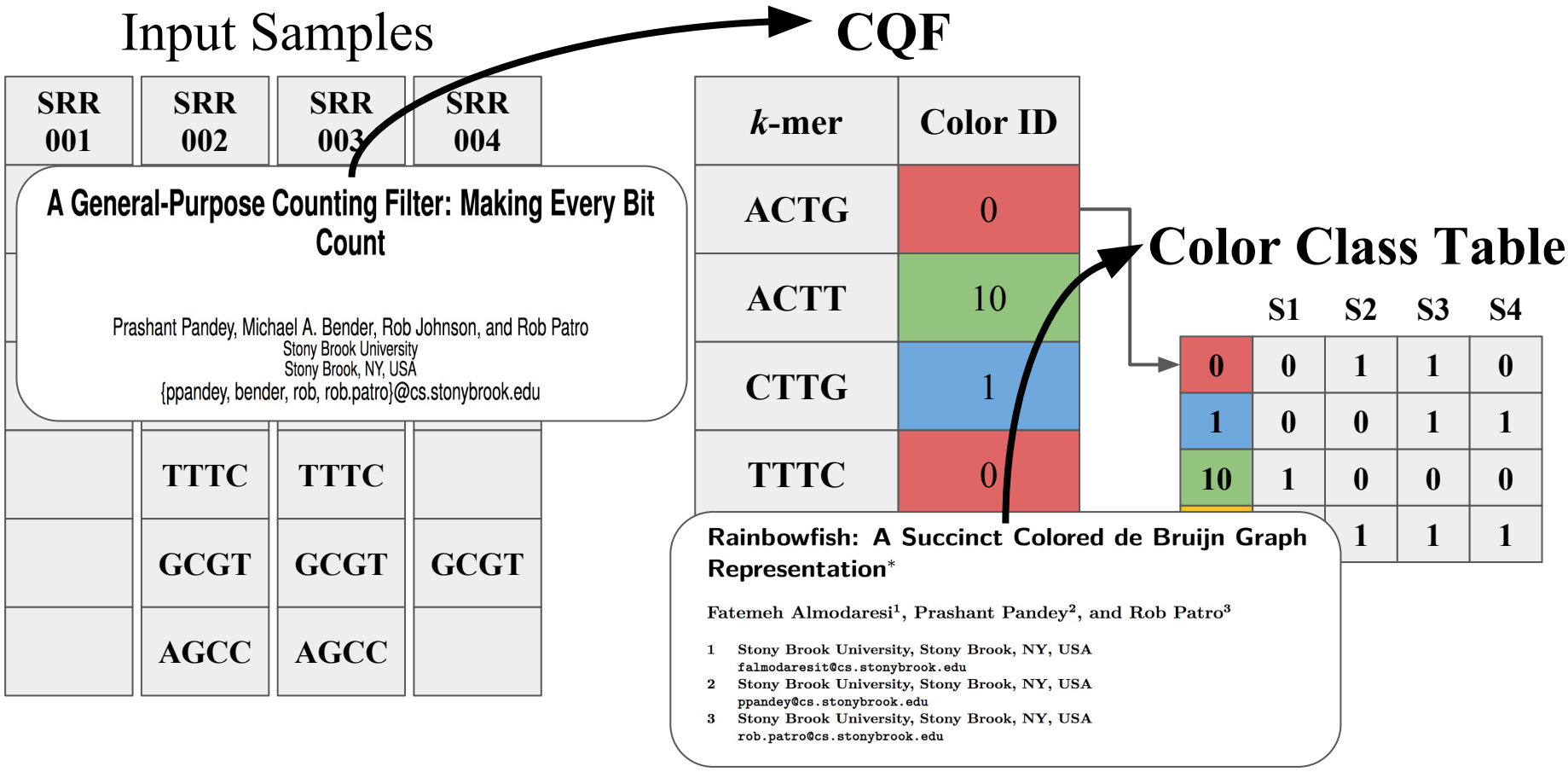
- We store sets of samples as bit vectors.

Mantis: A fundamentally different technique



- We use the CQF to map *k*-mers to color-class IDs.

Mantis: A fundamentally different technique



- We use Rainbowfish technique to map color-class IDs to sets of samples.

Counting quotient filter (CQF)

- A replacement for the (counting) Bloom filter.
- Space and computationally efficient.
- Uses variable-sized counters to handle skewed data sets efficiently.

$$\text{CQF space} \leq \text{BF space} + O\left(\sum_{x \in S} \log c(x)\right)$$

Asymptotically optimal

~~Counting~~ quotient filter (CQF)
Colored

- A replacement for the (counting) Bloom filter.
- Space and computationally efficient.
- U
  e

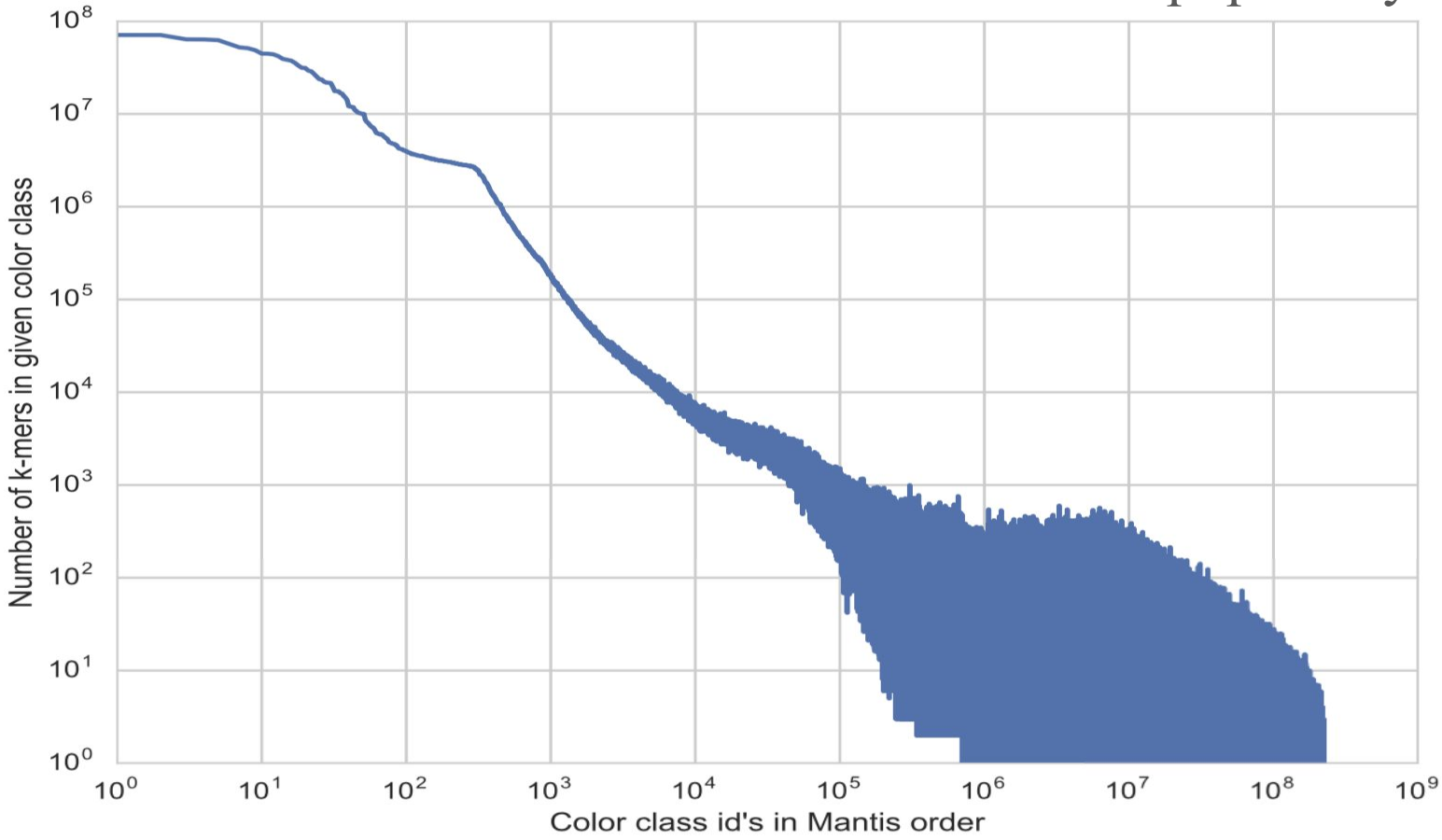
We repurpose variable-sized counters in the CQF to store color-class IDs.

CQF space ≤ BF space + $O(\sum_{x \in S} \log c(x))$

Asymptotically optimal

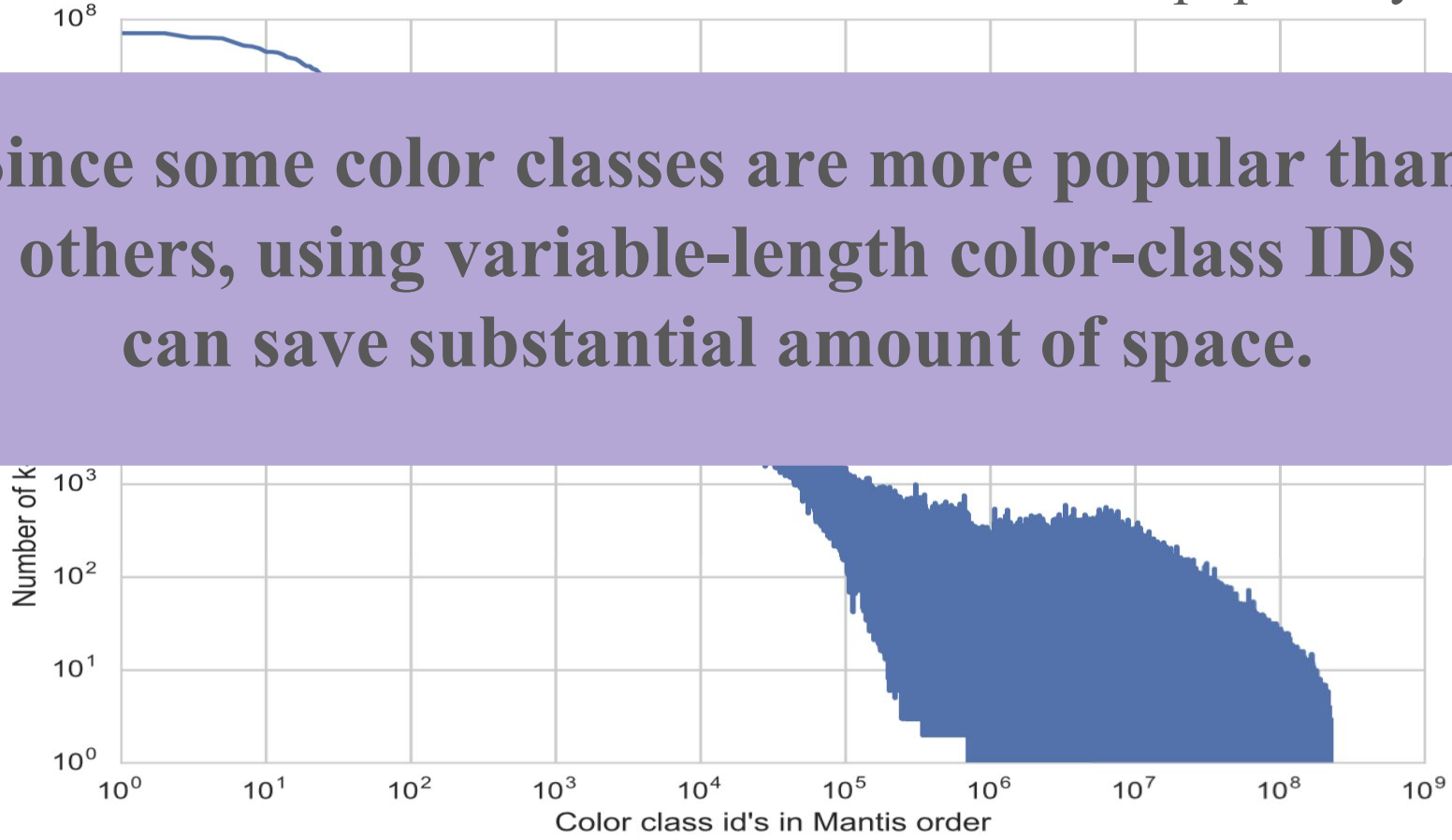
The CQF stores variable-sized color IDs efficiently

The distribution of IDs of color classes and their popularity.



The CQF stores variable-sized color IDs efficiently

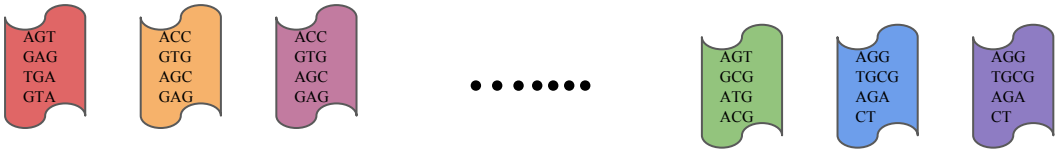
The distribution of IDs of color classes and their popularity.



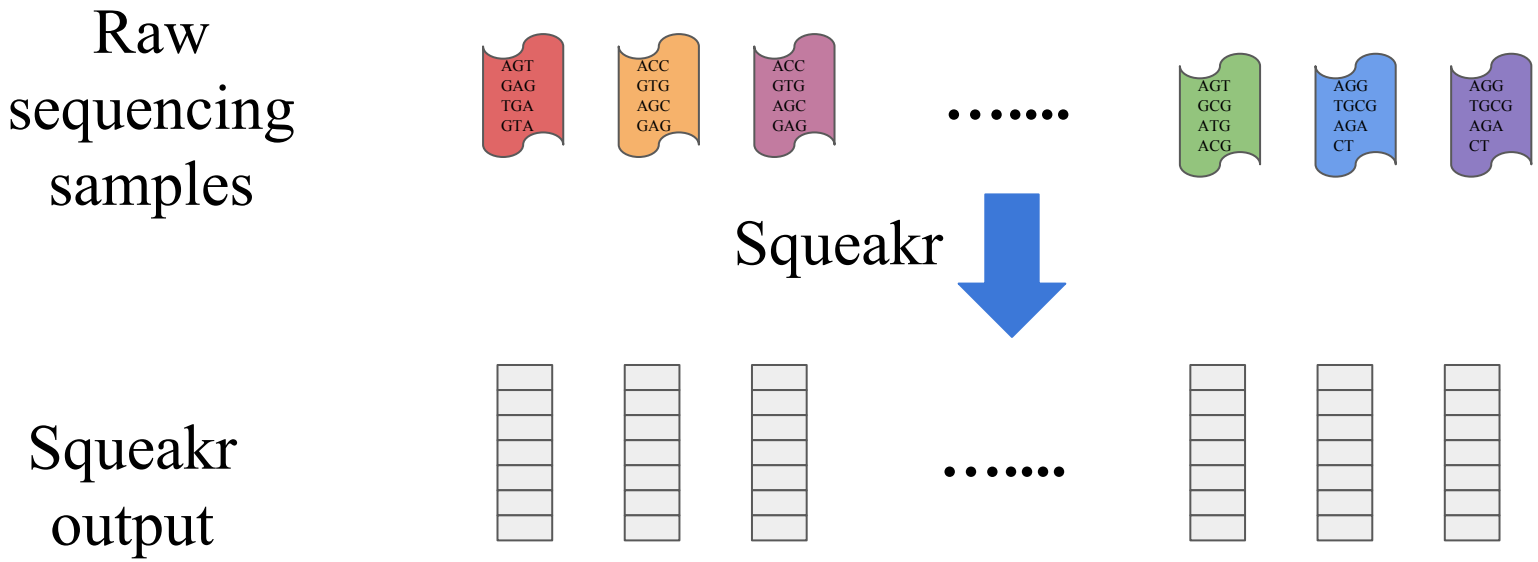
Since some color classes are more popular than others, using variable-length color-class IDs can save substantial amount of space.

Construction process in Mantis

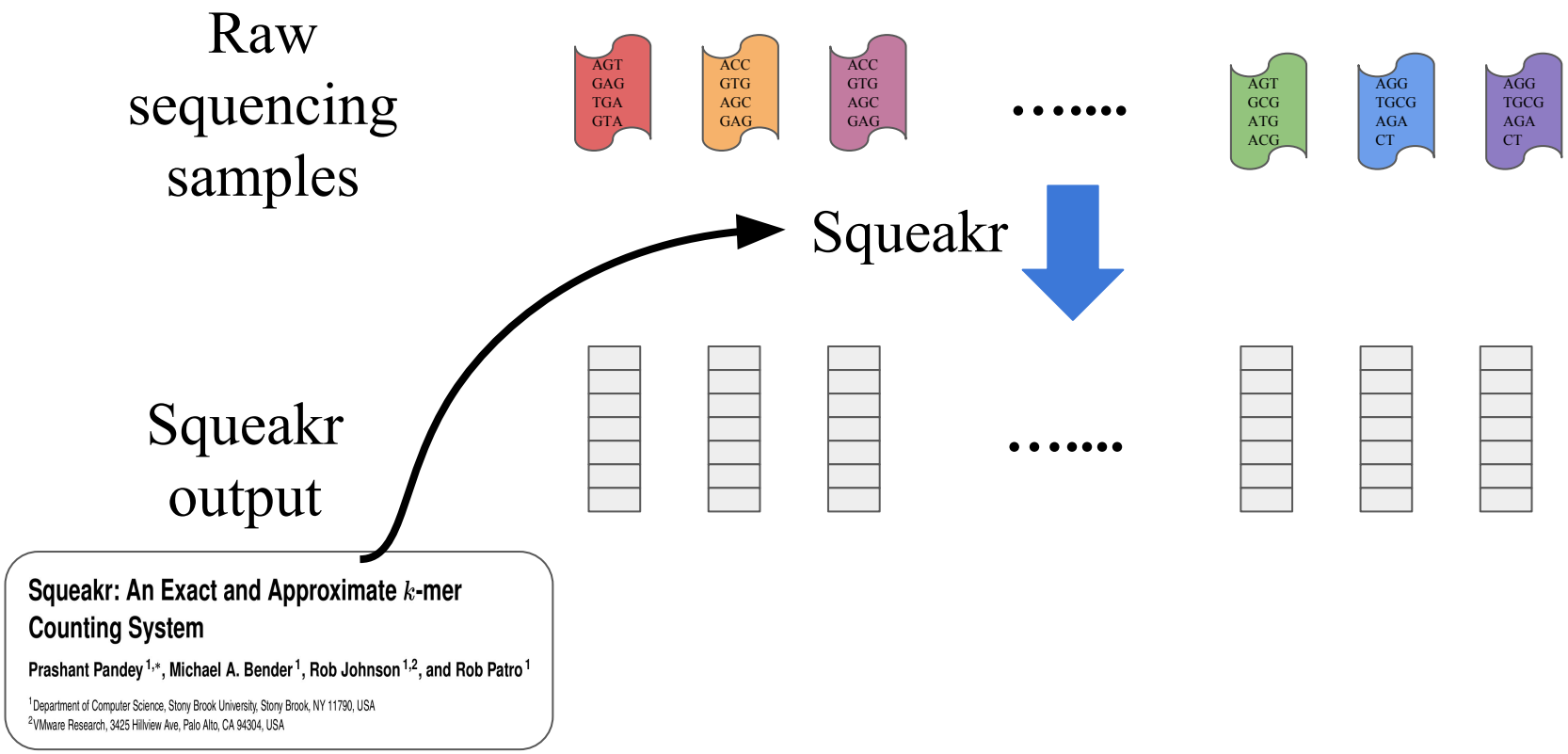
Raw sequencing samples



Construction process in Mantis

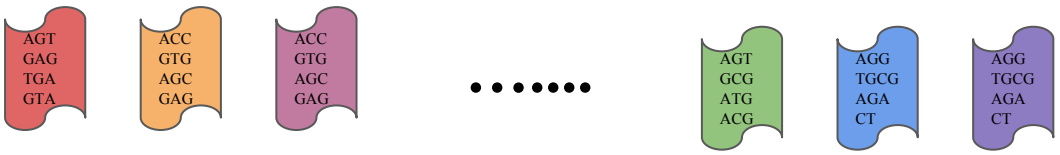


Construction process in Mantis

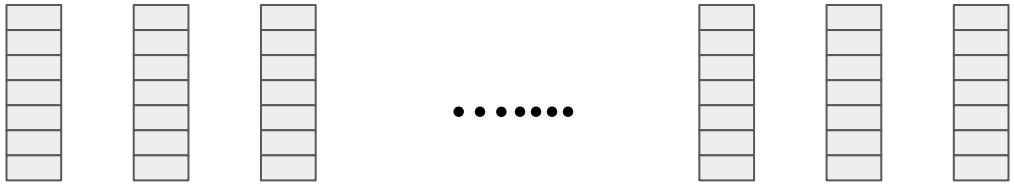


Construction process in Mantis

Raw sequencing samples



Squeakr



Squeakr output

Squeakr: An Exact and Approximate k-mer Counting System

Prashant Pandey [1,*], Michael A. Bender [1], Rob Johnson [1,2], and Rob Patro [1]

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY 11790, USA
[2]VMware Research, 3425 Hillview Ave, Palo Alto, CA 94304, USA

Multi-way merge

Mantis

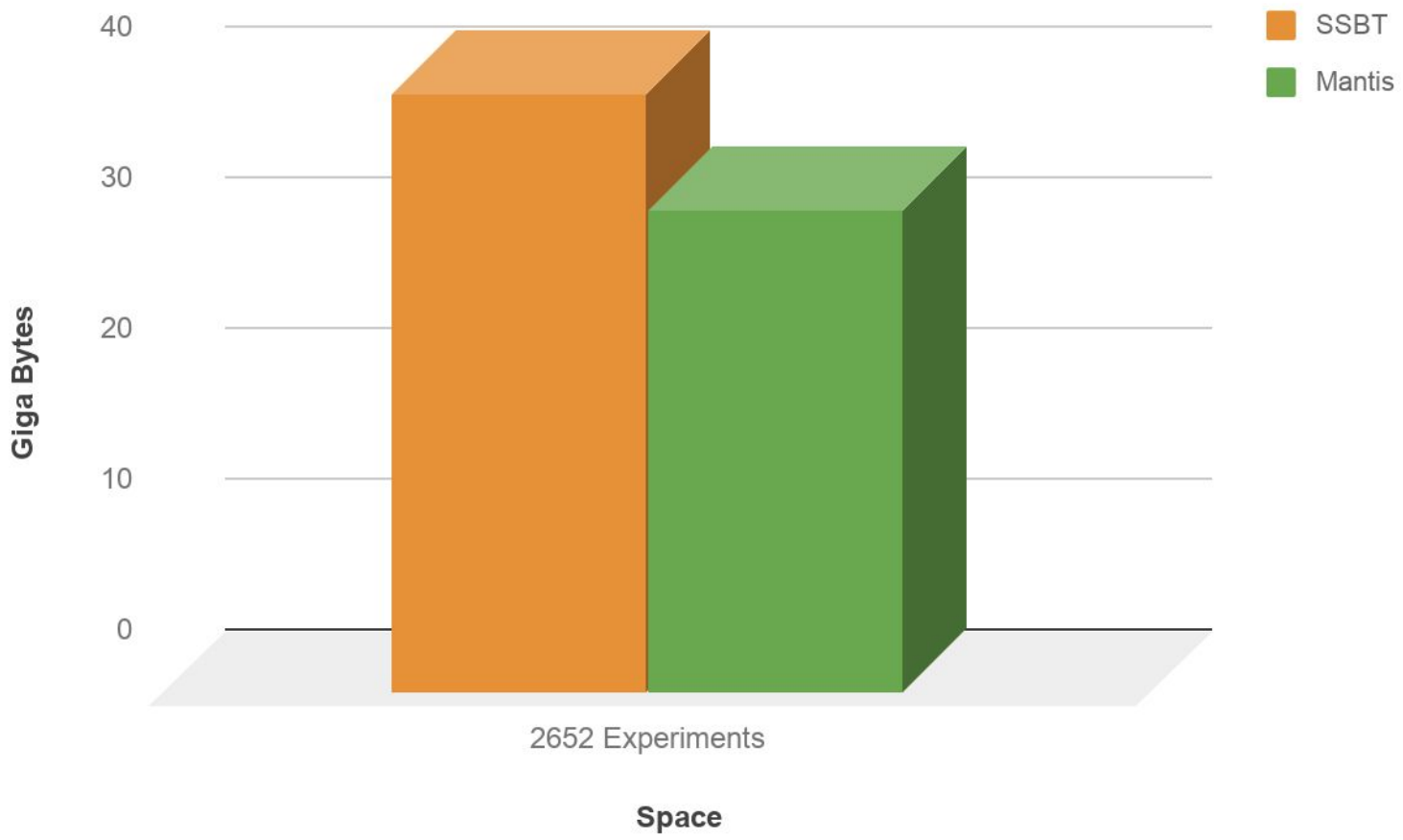
k-mer	Color Id
ACTG	0
ACTT	10
CTTG	1
TTTC	0
GCGT	11
AGCC	0

	S1	S2	S3	S4
0	0	1	1	0
1	0	0	1	1
10	1	0	0	0
11	0	1	1	1

Experimental setup

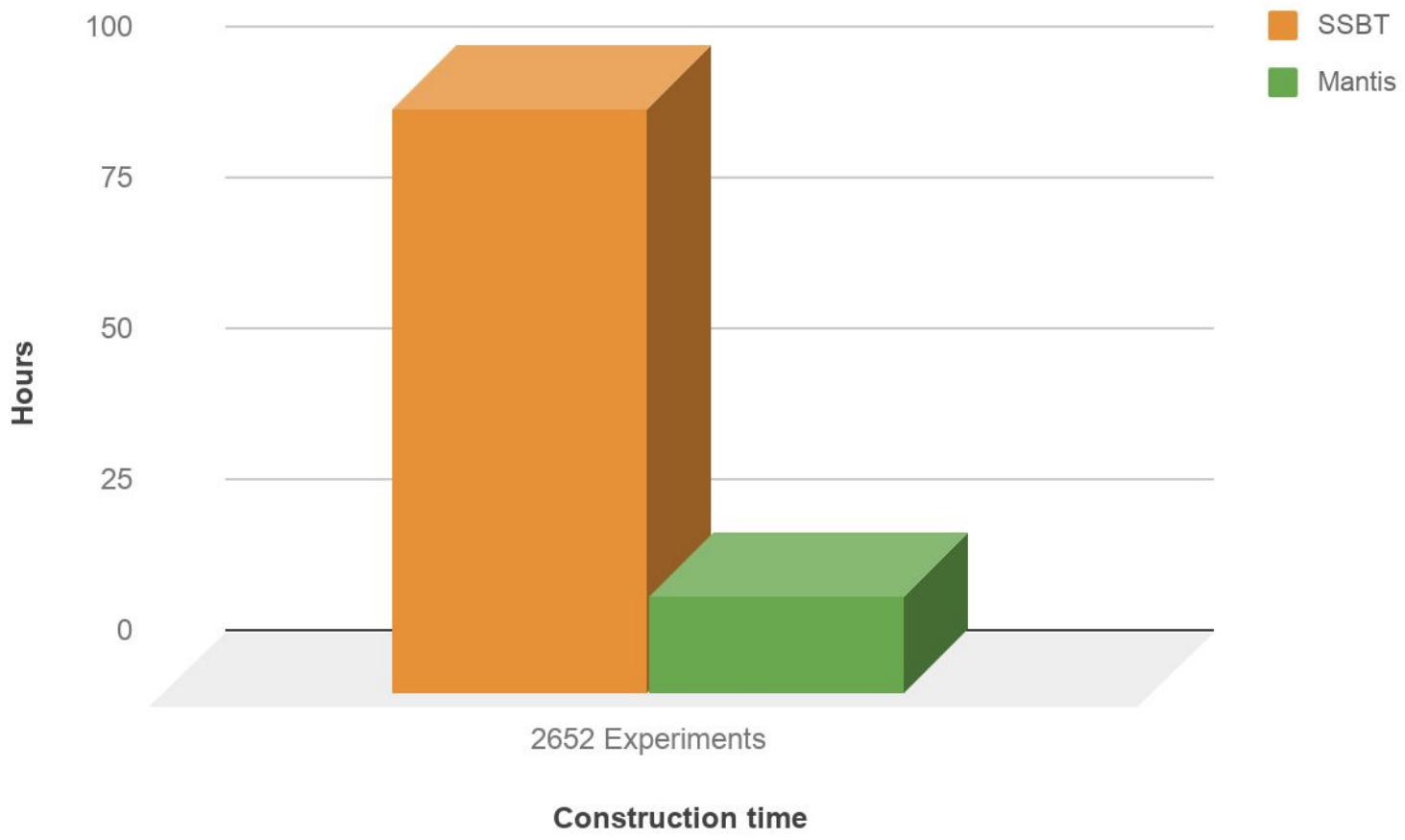
- Build index for 2652 samples of RNA-seq short-read sequencing runs of human blood, brain, and breast tissue.
- Compared with SSBT.
- Evaluation metrics:
  - Index size
  - Construction time
  - Query performance
  - Quality of results

Index size



Mantis is 20% smaller than the SSBT.

Construction time



Mantis is 6X faster than the SSBT for construction.

Query performance

Mantis is 6X-100X faster than the SSBT for queries.

Quality of results

θ = 0.9 for SSBT

Mantis is exact.

Conclusion

- Raw sequencing data archives are an untapped trove of information.
- Mantis makes it feasible to search these archives for new discoveries.
- Mantis outperforms prior systems by up to 100x.

Source code: https://github.com/splatlab/mantis

Fatemeh Almodaresi

Michael A. Bender

Michael Ferdman

Rob Johnson

Rob Patro

Source code: https://github.com/splatlab/mantis

Drowned in next generation sequencing data