

Partitioned Interleaved Bloom Filter Using Optane DC Persistent Memory

**Enrico Seiler, Freie Universität Berlin, Max Planck Institute for
Molecular Genetics**
(Knut Reinert, Svenja Mehringer)

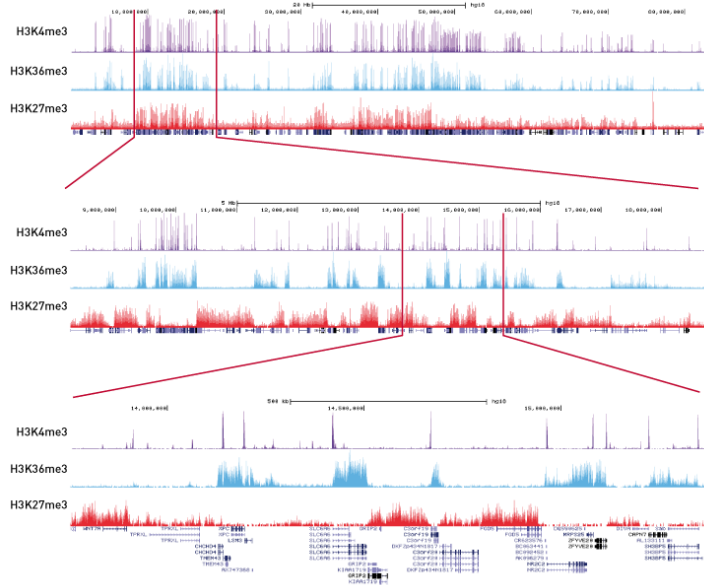


Next generation sequencing

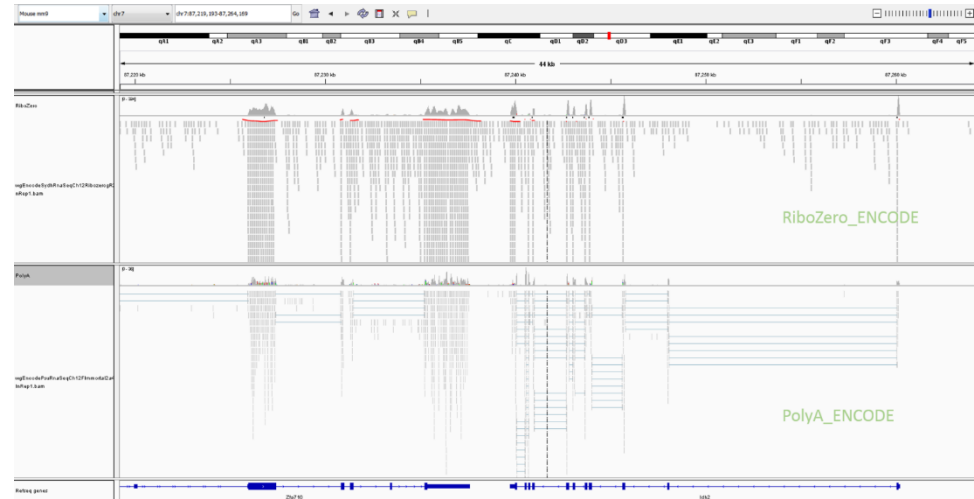


Current genomic analysis

NGS applications



ChIP-Seq



RNA-Seq

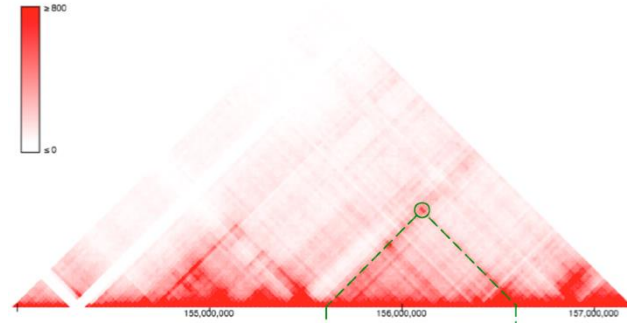
NGS applications

Assembly: hg19 @ resolution: 25 kb
Location:
chr7:153975000-157200000
Name: SHH
RefSeq: NM_000193
Ensembl (Gene): ENSG00000164690

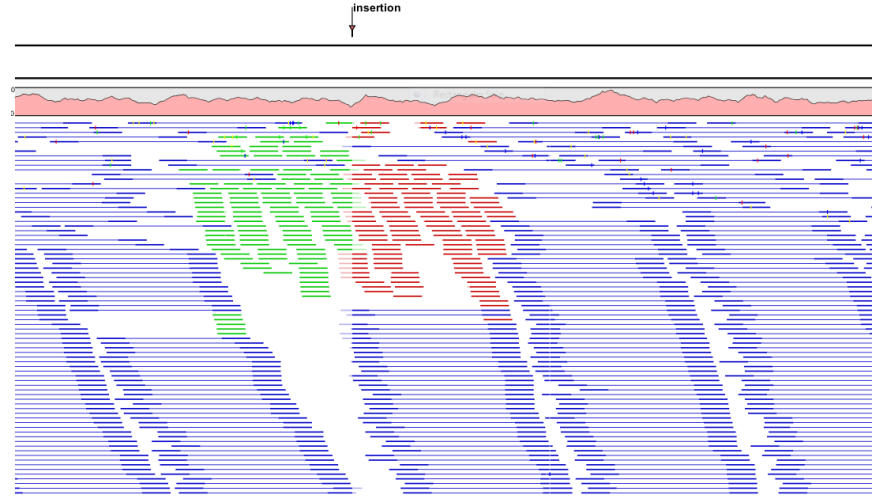
[Click to check its expression in ENCODE](#)



Download Image

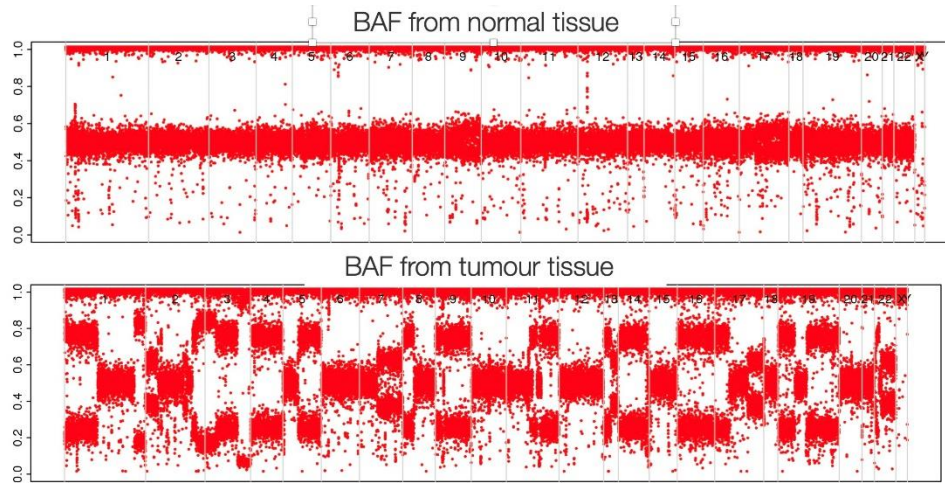


HiC-Seq

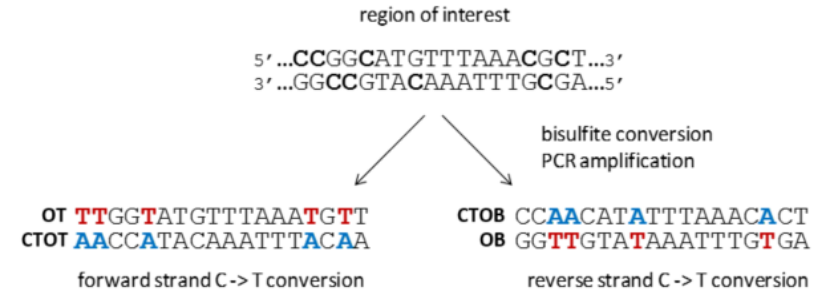


Variation mapping

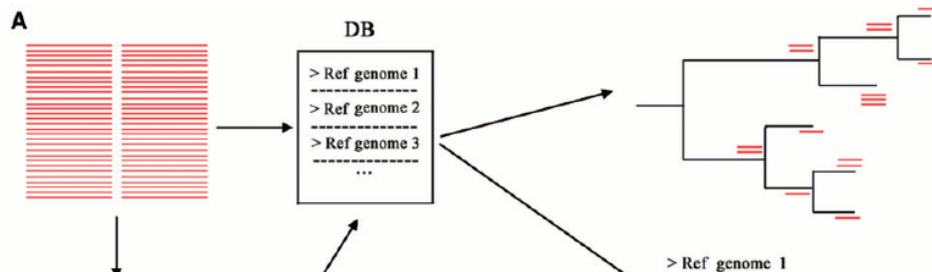
NGS applications



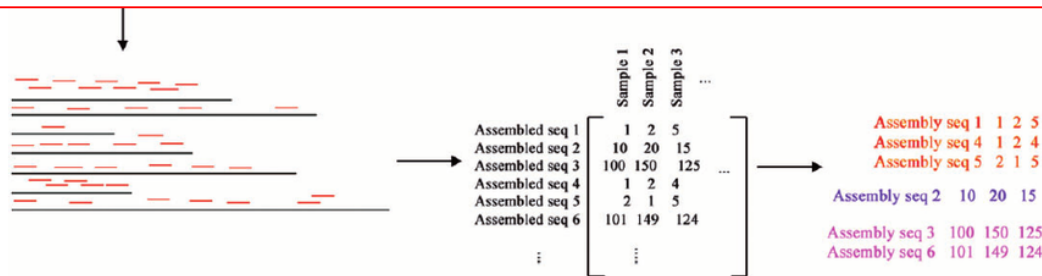
Biallelic frequencies (cancer signatures)



Bisulphite mapping



Basically **ALL** methods **search** NGS reads approximately in a **reference database**



Metagenomic mapping

Methods for searching

There are two main approaches:

A) Alignment based -> build compressed suffix array and map read
---- Data structure not dynamic, not practical to build for over 30 GiB

B) Alignment free -> built k-mer lookup table and compare k-mer composition of reference and read

---- Also grows large for over 10 GiB and large k
+++ Faster to build than CSA

=> Currently most approaches use k-mer based methods

How many k-mers are there?

Sequencing earth?

Published online 23 August 2011 | Nature | doi:10.1038/news.2011.498
Corrected online: 24 August 2011

News

Number of species on Earth tagged at 8.7 million

Most precise estimate yet suggests more than 80% of species still undiscovered.

Lee Sweetlove

10^7 species x 10^8 genome size =>
earth genome has **10^{15} bps**

Let's assume we can divide the earth genome
(or the part we have) in about 1000 subsets

Then each subset is ~1 Tbp

1 Tbp of sequence contains too many
different k-mers for $k > 19$
e.g. a 30GiB metagenomics data set
contains $\sim 10^{10}$ different 19-mers

=> We need to partition the sequence into bins

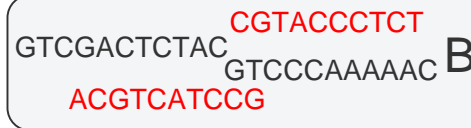
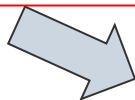
Effective text size and text ratio

The **reference** can be a genome (or many) or raw data files.
Important are:

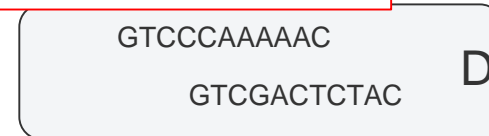
* the **effective text size** $N_e(k)$ (k is k -mer size, e.g. 20)

* the

- A metagenomic data set of 30 GiB has $N(19)=10^{10}$
- Partitioned in 1000 bins the $r(19)$ is 1.2 (not 1000)



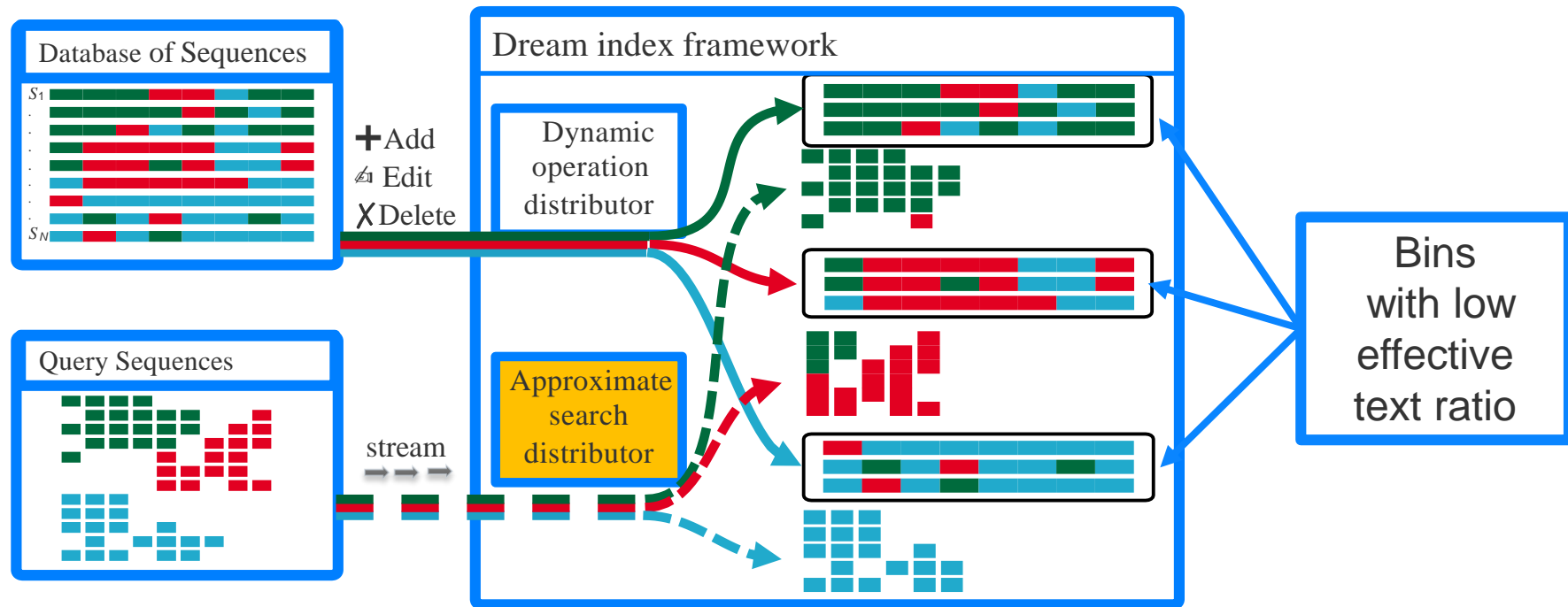
$$r(k) = 2$$



$$r(k) = 1$$

DREAM Index framework

DREAM = Dynamic seaRchable pArallel coMpressed index



Can we search a NGS read
in a database of size
1 TiB partitioned into bins?

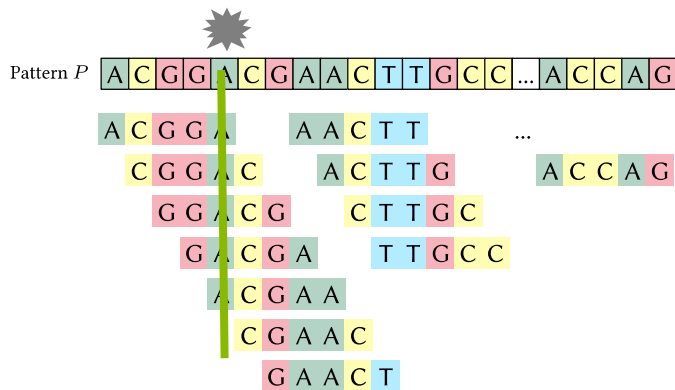
K-mer counting lemma

For a given k and number of errors e , there are $k_p = |P| - k + 1$ many k -mers in pattern P and an approximate occurrence of P in T has to share at least $t = (k_p - k \cdot e)$ k -mers

Total number of k -mers

An error can destroy at most k k -mers
 e errors can destroy $k \cdot e$ k -mers

Threshold t is:
 $t = |P| - e \cdot k - k + 1$



Reducing num. of relevant k-mers

a) (4,4,####)-minimizer

ACGTCGACGTTTAG

ACGT GACG aatc

CGTC ACGT

b) (8,4,####)-minimizer

ACGTCGACGTTTAG

ACGT----

---CGAC

c) (8,3,#. #. #)-minimizer

ACGTCGACGTTTAG

A.G.C---

---C.A.G

We can reduce the amount of relevant k-mers by a factor of 2-8

--caaa-

-caaa--

11 ungapped 4-mers

-g.a.a.--

g.a.a.---

5 ungapped 4-mers

5 gapped 3-mers

Use minimizers and adapt counting lemma

In which of the bins is a given, relevant k-mer ?

Bin	1	2	3	.	.			1023	1024
Binning bitvector	1	0	0	0	0	1	0	1	1

Cannot be done with k-mer index:
For 10^{10} 20-mers and 1024 bins it takes
2-4 TiB main memory

Bloom Filter

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	...	n
0	0	0	0	1	1	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	1	1	1	0	...	0

- A bit-vector of size n
- h different hash functions $\{H_1, H_2, \dots, H_h\}$
- $H_i : \mathbb{N} \rightarrow [1, \dots, n]$
- To add a ***k-mer*** we simply set h bit positions
- During lookup, we expect all of the h bit positions to be set
- Collisions are allowed in the expense of **false positive** answers
- To keep FP low you need enough space

$$P_{fp} = \left(1 - \left(1 - \frac{1}{n} \right)^{h \cdot m} \right)^h$$

Where:

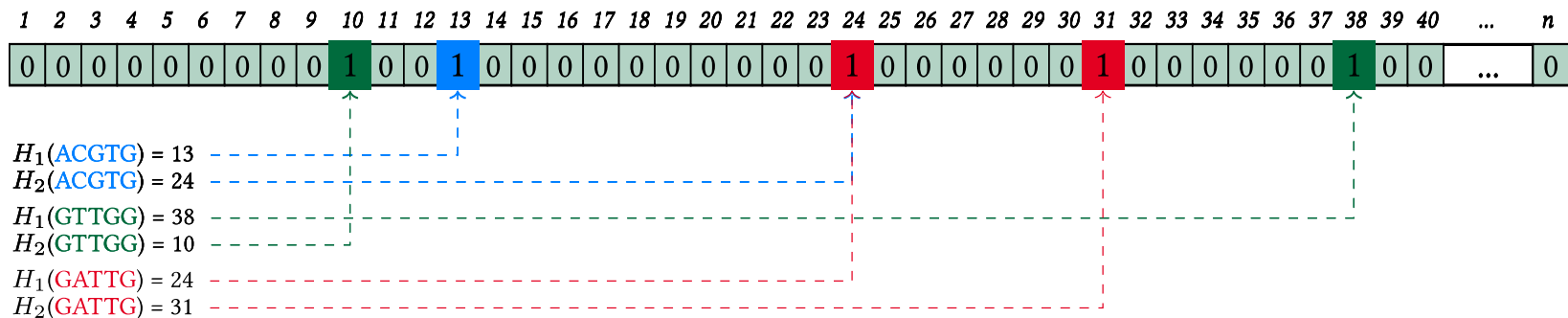
m = the number of elements **added**

Bloom Filter

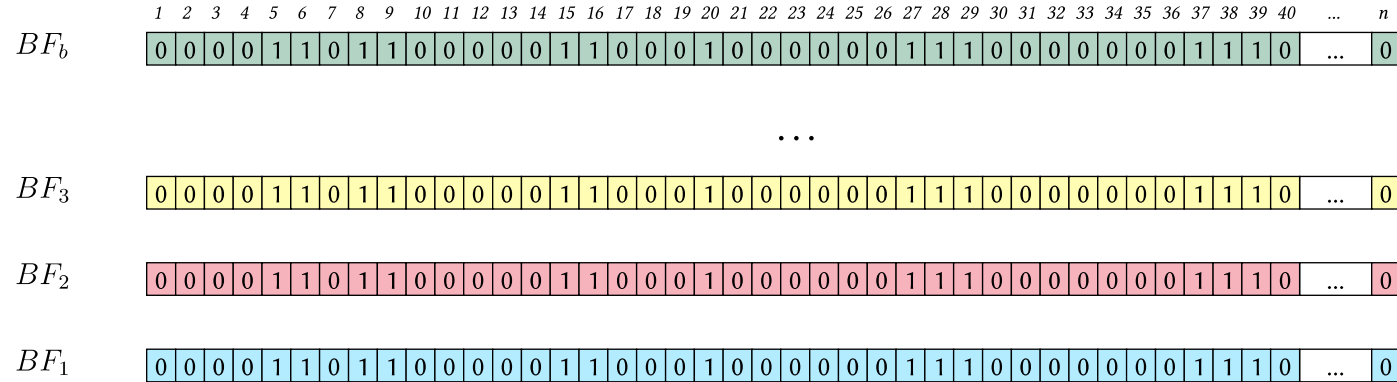
E.g. Add k-mers

ACGTG GTTGG GATTG

2 hash functions: $\{H_1, H_2\}$

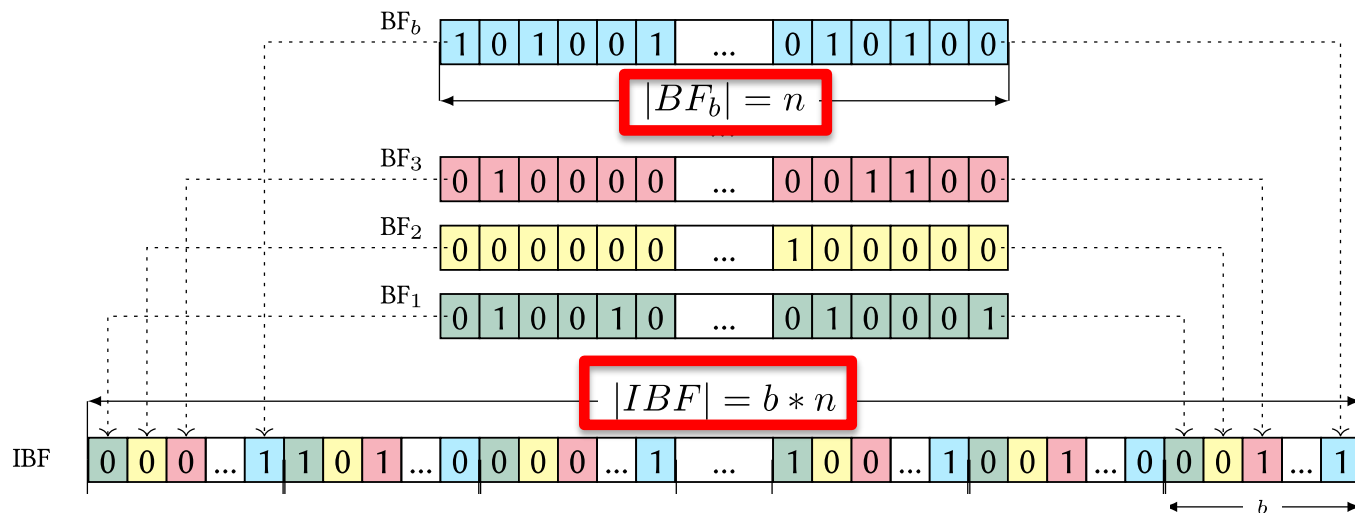


Bloom Filter for bins



- We need to check every read against each Bloom filter

Interleaved Bloom Filter (IBF)



To add a k-mer (**K**) from bin **j**

We set position

$$\mathbf{b} * H_i(\mathbf{K}) + \mathbf{j}$$

To check in which bins a k-mer is present, we use the bit blocks

$$[H_i(\mathbf{K}) * \mathbf{b}, H_i(\mathbf{K}) * (\mathbf{b} + 1))$$

This is the binning bitvector!

Increment counters

	Bin	1	2	3	.	.			1023	1024
Binning bitvector		1	0	0	0	0	1	0	1	1

	Bin	1	2	3	.	.			1023	1024
Counters (16 bit)		1	12	0	2	0	2	0	12	54

We have to increment a counter if a bit is set.
This operation is accelerated using vectorized computing
`lzcnt_u64` or `lzcnt_epi64` (AVX512)
or `mask_add_epi16` (AVX512)
or precomputed tables

Experiments

4 GiB random DNA divided in 64, 256, 1024, 8192 bins

Each “bin genome” is mutated and 16 times replicated to simulate subspecies for a total of 64 GiB of sequence with $N(19) = 4.3 \text{ G}$



1 million Illumina reads sampled uniformly from bins and counting lemma filter applied

Filter time, space, FP for IBF

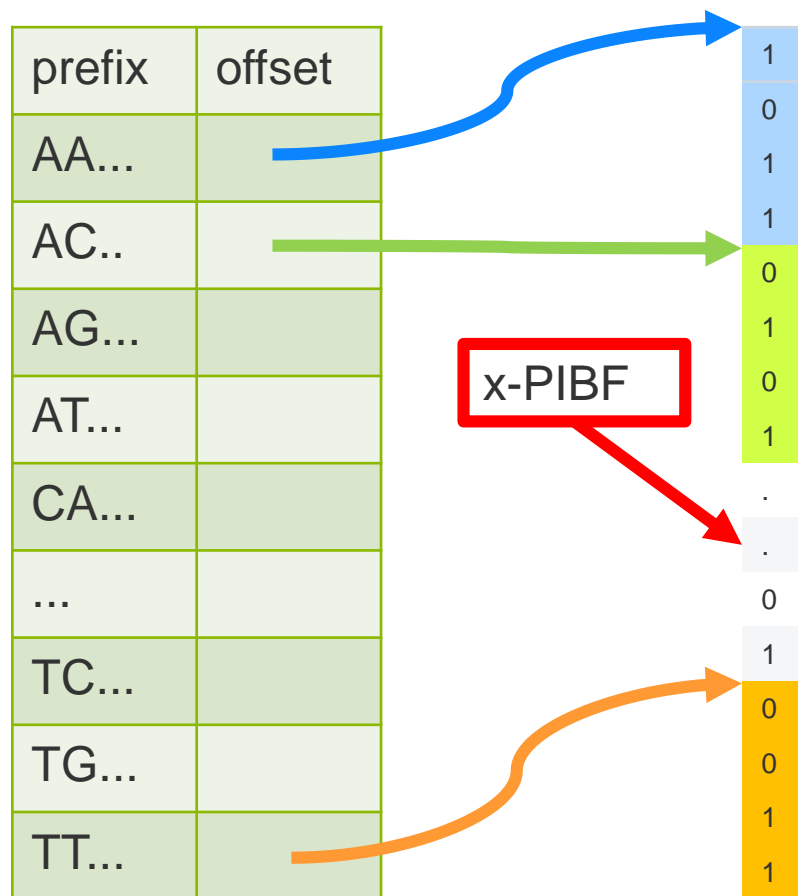
Size IBF

Construction time in min

search time in sec

b		64				1024			
k=19		Construct	Search	Space	FP	Construct	Search	Space	FP
1GB	1-PIBF	16.78	1.1	1024	66 * 10 ⁶	16.31	5.76	1024	1 * 10 ⁹
	1-PIBFc	18.01	10.17	3039	66 * 10 ⁶	17.73	92.19	3039	1 * 10 ⁹
2GB	1-PIBF	17	0.98	2048	16,647	16.85	3.78	2048	252,948
	1-PIBFc	16.46	9.84	4943	16,647	18.5	81.29	4943	252,948
4GB	1-PIBF	18.1	1.01	4096	192	18	3.17	4096	0
	1-PIBFc	20.04	8.7	6677	192	20.11	54	6676	0
8GB	1-PIBF	20.83	1.15	8192	189	17.81	2.82	8192	0
	1-PIBFc	23.8	7.93	8471	189	20.59	36.92	8471	0
16GB	1-PIBF	21.31	1.37	16384	189	20.42	2.85	16384	0
	1-PIBFc	27.15	6	6973	189	25.85	18.46	6972	0

Secondary memory



We have to scan a read x times if we have an x -partitioned IBF

b		1024			
k=19		Construct	Search	Space	FP
16GB	1-PIBF	26.54	4.79	16384	0
	1-PIBFc	20.56	29.53	10221	0
	2-PIBF	27.53	6.9	8192	0
	2-PIBFc	32.83	32.01	5111	0
	4-PIBF	31.91	11.49	4096	0
	4-PIBFc	42.32	36.42	2554	0
	8-PIBF	46.41	20.75	2048	0
	8-PIBFc	57.73	44.66	1278	0

Intel Optane DC PMM

2 basic modes:

- use as filesystem
- use as memory extension

3 scenarios:

- use as extended memory, but do not utilize
- use as filesystem
- use as extended memory, utilize

Run times on an 1 TiB 8-PIBF using a machine with

Cascade Lake CPU with 96 threads, 192 GiB DRAM, 1.5 TiB Optane NVRAM

1, 5 and 10 batches of 1 million Illumina reads

Only extend memory

# Batches	I/O Time	Query Time	Speedup
1 D	30:13	1:10	1/1
5 D	3:31:05	3:40	1/1
10 D	7:25:16	7:46	1/1
1 N	41:40	1:10	0.7/1
5 N	2:43:38	3:46	1.3/1
10 N	4:40:51	11:53	1.6/0.7

Use as filesystem

# Batches	I/O Time	Query Time	Speedup
1 D	30:13	1:10	1/1
5 D	3:31:05	3:40	1/1
10 D	7:25:16	7:46	1/1
1 N	13:48	1:10	2.2/1
5 N	1:11:13	3:59	3/0.9
10 N	2:27:06	7:48	3/1

Utilize extended memory

# Batches	I/O Time	Query Time	Speedup
1 D	30:13	1:10	1/1
5 D	3:31:05	3:40	1/1
10 D	7:25:16	7:46	1/1
1 N	36:50	0:26	0.8/2.7
5 N	42:20	2:29	5/1.5
10 N	41:53	3:52	10.6/2

Earth genome?

We can reduce the representative k-mer set
of 100

Assuming conjunction with an X-PIBF enables
k-mer analysis of data sets

Extrapolating with an even larger effective text size
(e.g. 6 TiB PMM would allow an

- 256 effective text size of ~ 1TiB (search 1 sec)
- 128 effective text size of ~ 1TiB (search 8 sec)
- 64 effective text size of ~ 1TiB (search 2 sec)
- 32 GiB 4-partitioned, comp. IBF (search 8-36 sec)



- Please follow us under:
 -  **@SeqAnLib**
 - www.seqan.de