Succinct Colored De Bruijn Graphs (VARI)

Muggli *et al*

Rachel Sherman
Genomics Journal Club
June 29, 2017

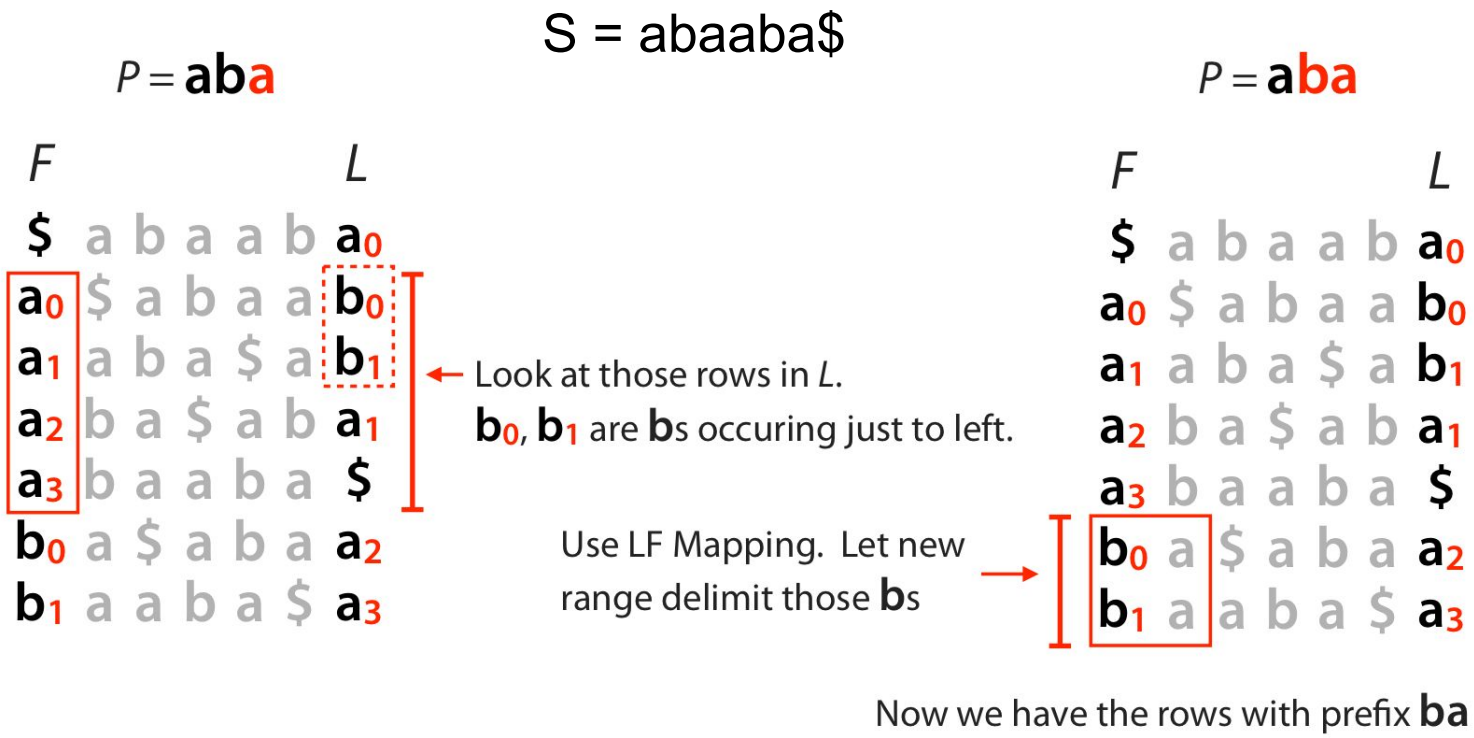
Why (Colored) Graph Genomes?

Represent populations

Help identify structural similarity/difference that may be missed by reference based approaches

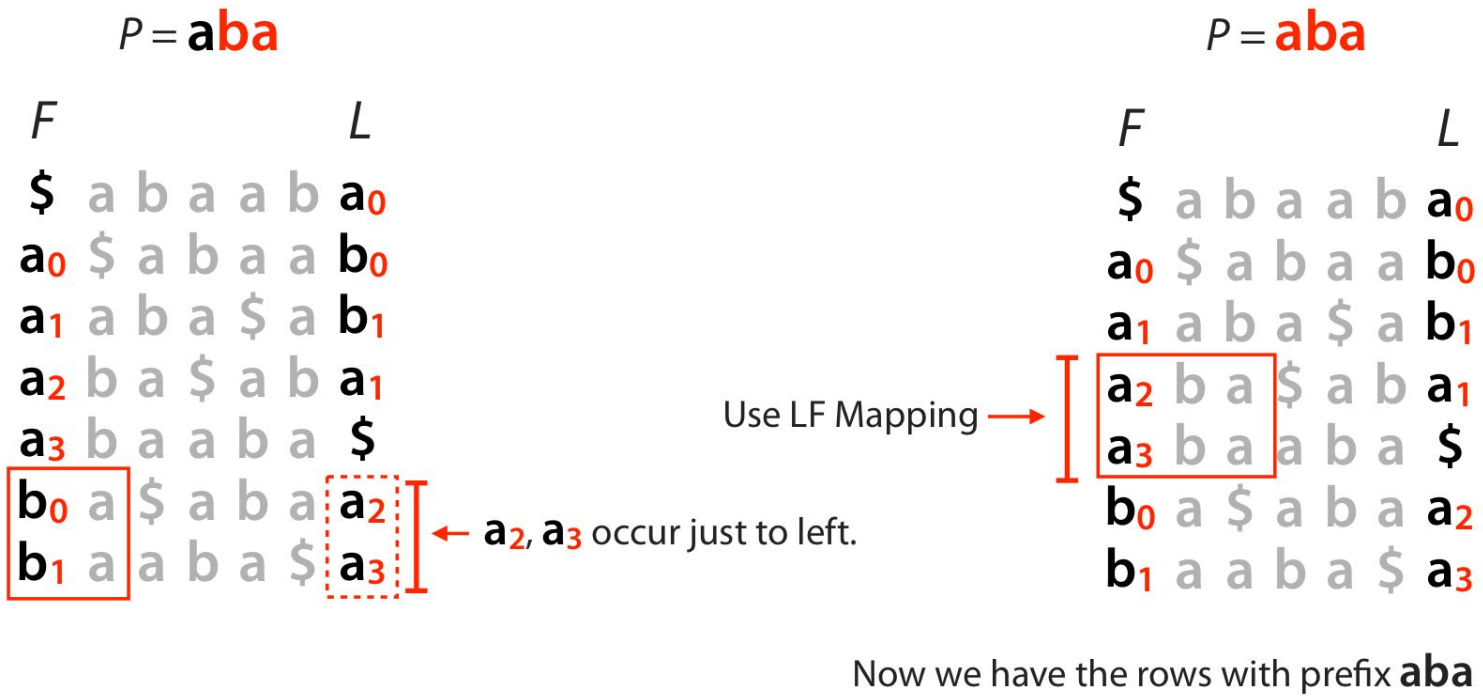
More “compressed” and easier to compare than separate graphs for each individual in a population/experiment

LF Mapping Refresher



Slides from Ben Langmead

LF Mapping Refresher

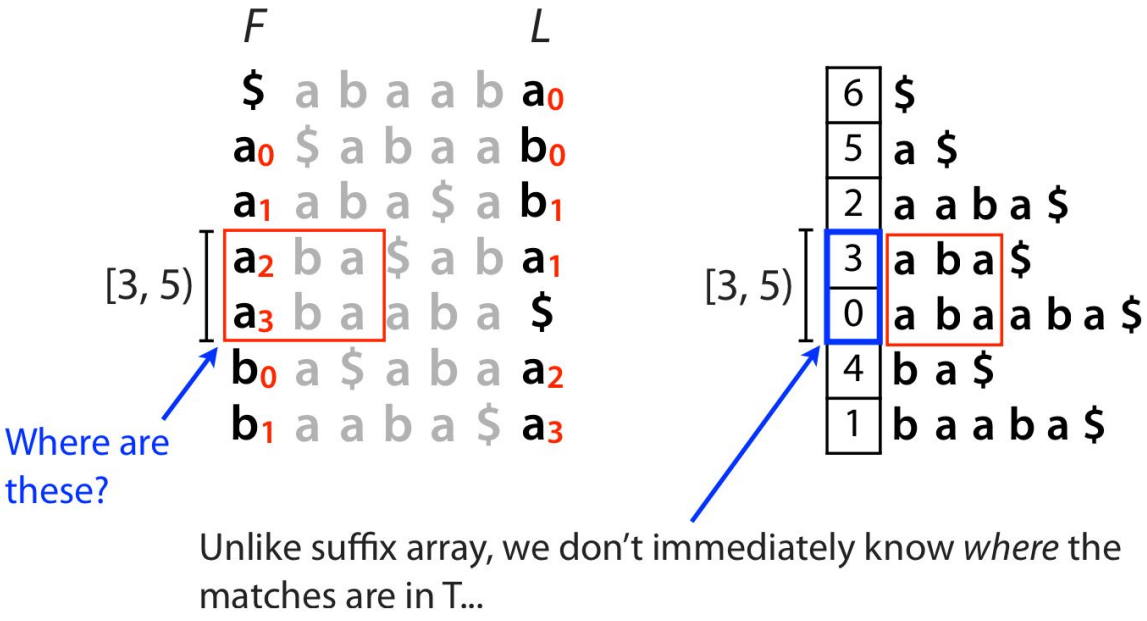


Slides from Ben Langmead

LF Mapping Refresher

$P = \mathbf{aba}$

Got the same range, [3, 5), we would have got from suffix array



Slides from Ben Langmead

LF Mapping Refresher

(1) Scanning for preceding character is slow

(2) Storing ranks takes too much space

```
def reverseBwt(bw):
    """ Make T from BWT(T) """
    ranks, tots = rankBwt(bw)
    first = firstCol(tots)
    rowi = 0
    t = "$"
    while bw[rowi] != '$':
        c = bw[rowi]
        t = c + t
        rowi = first[c][0] + ranks[rowi]
    return t
```

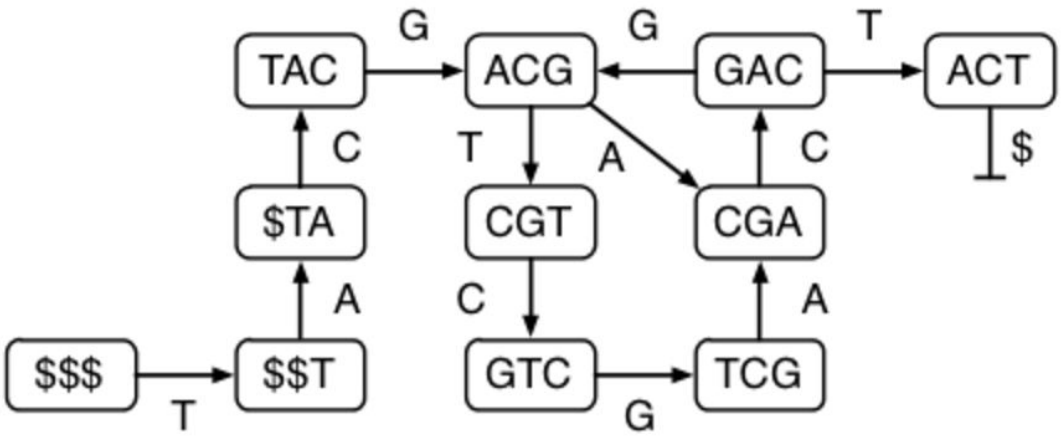
m integers

(3) Need way to find where matches occur in *T:*

Slides from Ben Langmead

Succinct De Bruijn Graph

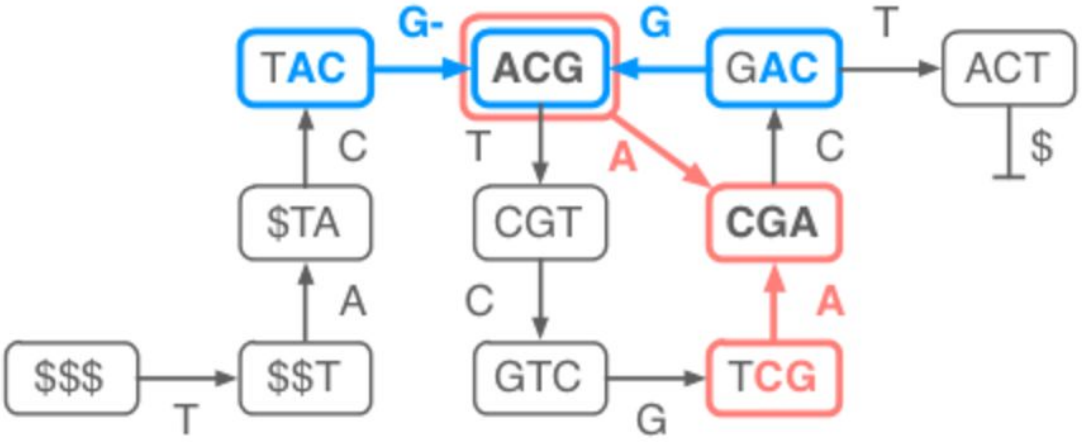
S = TACGACGTCGACT



Node	W
$ $ $	T
C G A	C
$ T A	C
G A C	G
G A C	T
T A C	G-
G T C	G
A C G	A
A C G	T
T C G	A-
$ $ T	A
A C T	$
C G T	C

Slides from Alex Bowe

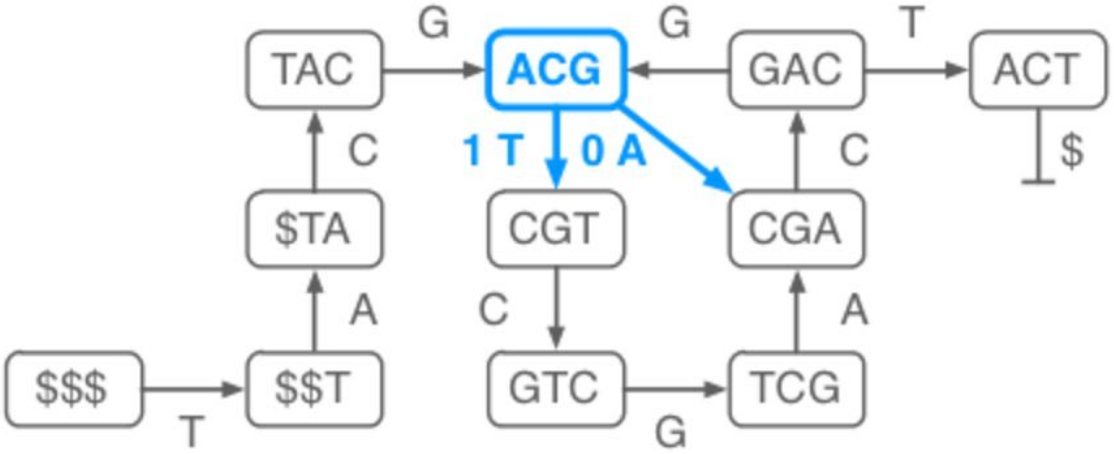
Succinct De Bruijn Graph



Node	W
$ $ $	T
C G A	C
$ T A	C
G A C	G
G A C	T
T A C	G-
G T C	G
A C G	A
A C G	T
T C G	A-
$ $ T	A
A C T	$
C G T	C

Slides from Alex Bowe

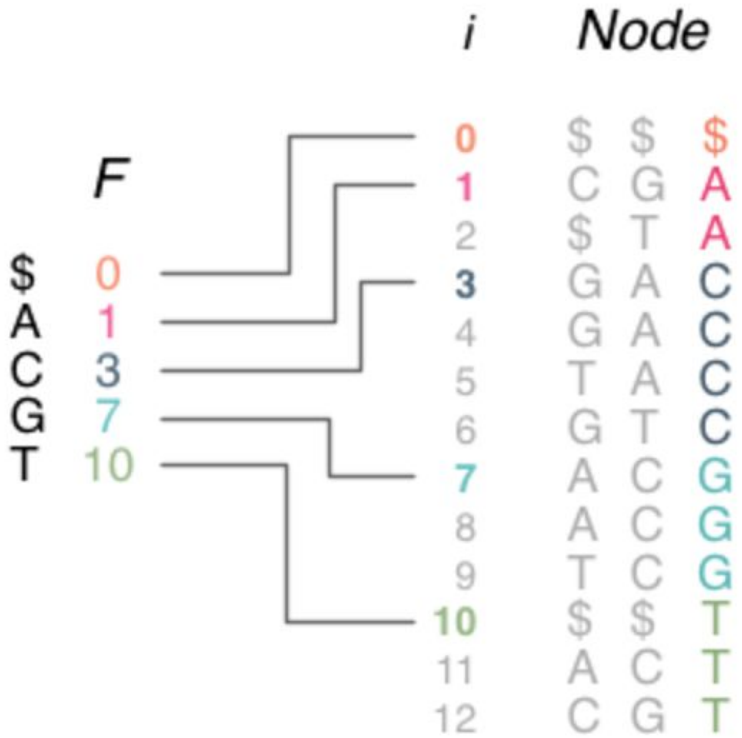
Succinct De Bruijn Graph



L	Node	W
1	$ $ $	T
1	C G A	C
1	$ T A	C
0	G A C	G
1	G A C	T
1	T A C	G-
1	G T C	G
0	A C G	A
1	A C G	T
1	T C G	A-
1	$ $ T	A
1	A C T	$
1	C G T	C

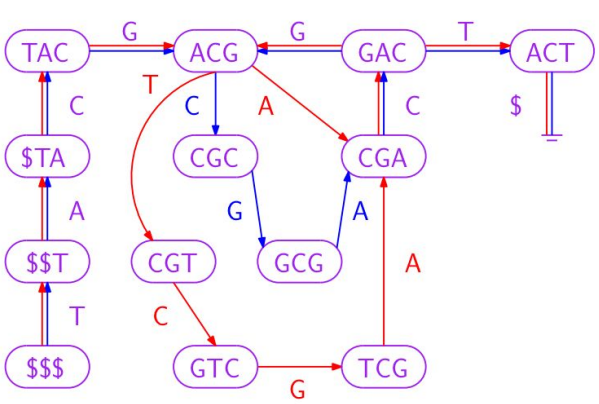
Slides from Alex Bowe

Succinct De Bruijn Graph



Slides from Alex Bowe

BOSS Representation



0	$$$	T	1	CGC	G	1	$$T	A
3	CGA	C	1	GTC	G	1	ACT	$
1	$TA	C	2	ACG	A C T	1	CGT	C
1	GAC	G T	1	GCG	A			
1	TAC	G	1	TCG	A			

$$\text{EBWT}(G) = \texttt{TCCGTGGGACTAAA\$C}$$

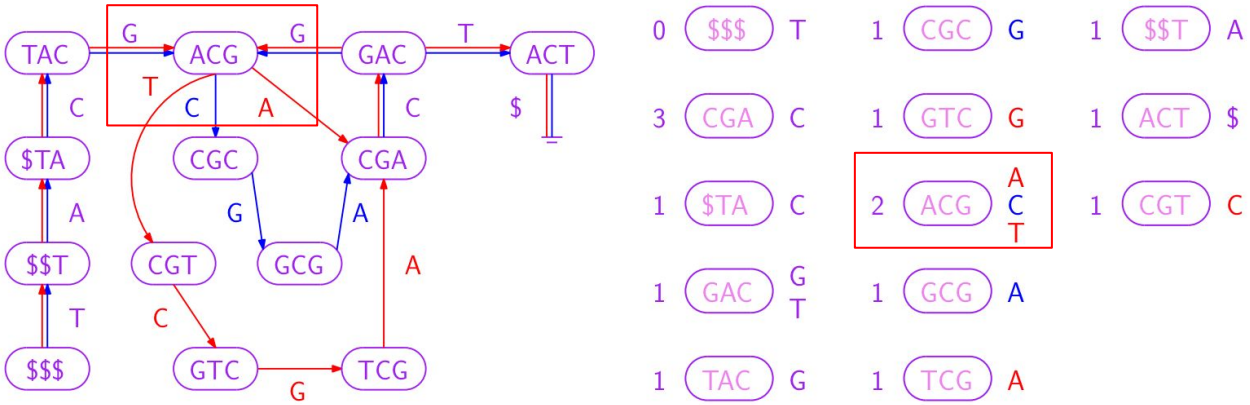
$$B_F = \texttt{001111110111111}$$

$$B_L = \texttt{1110111100111111}$$

$$C^{\mathrm{T}} = \texttt{0000001001010000}$$
$$\texttt{0000000110101001}$$

Muggsi et al, Figure 1

BOSS Representation



$$\text{EBWT}(G) = \texttt{TCCGTGGGACTAAA\$C}$$

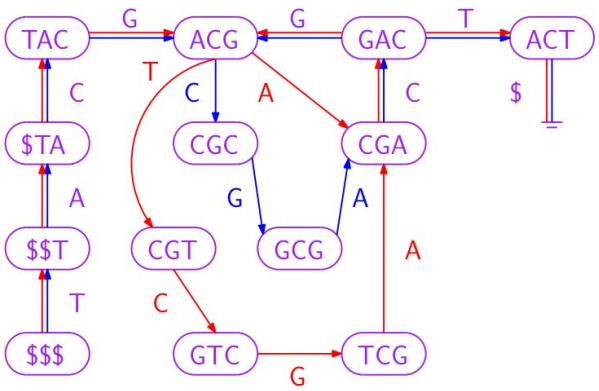
$$B_F = \texttt{0011111110111111}$$

$$B_L = \texttt{1110111100111111}$$

$$C^{\mathrm{T}} = \texttt{0000001001010000}$$
$$\texttt{0000000110101001}$$

Muggsi et al, Figure 1

BOSS Representation



0	$$$	T	1	CGC	G	1	$$T	A
3	CGA	C	1	GTC	G	1	ACT	$
1	$TA	C	2	ACG	A C T	1	CGT	C
1	GAC	G T	1	GCG	A			
1	TAC	G	1	TCG	A			

$$\text{EBWT}(G) = \texttt{TCCGTGGGACTAAA\$C}$$

$$B_F = 001111110111111$$

$$B_L = 11101111001111111$$

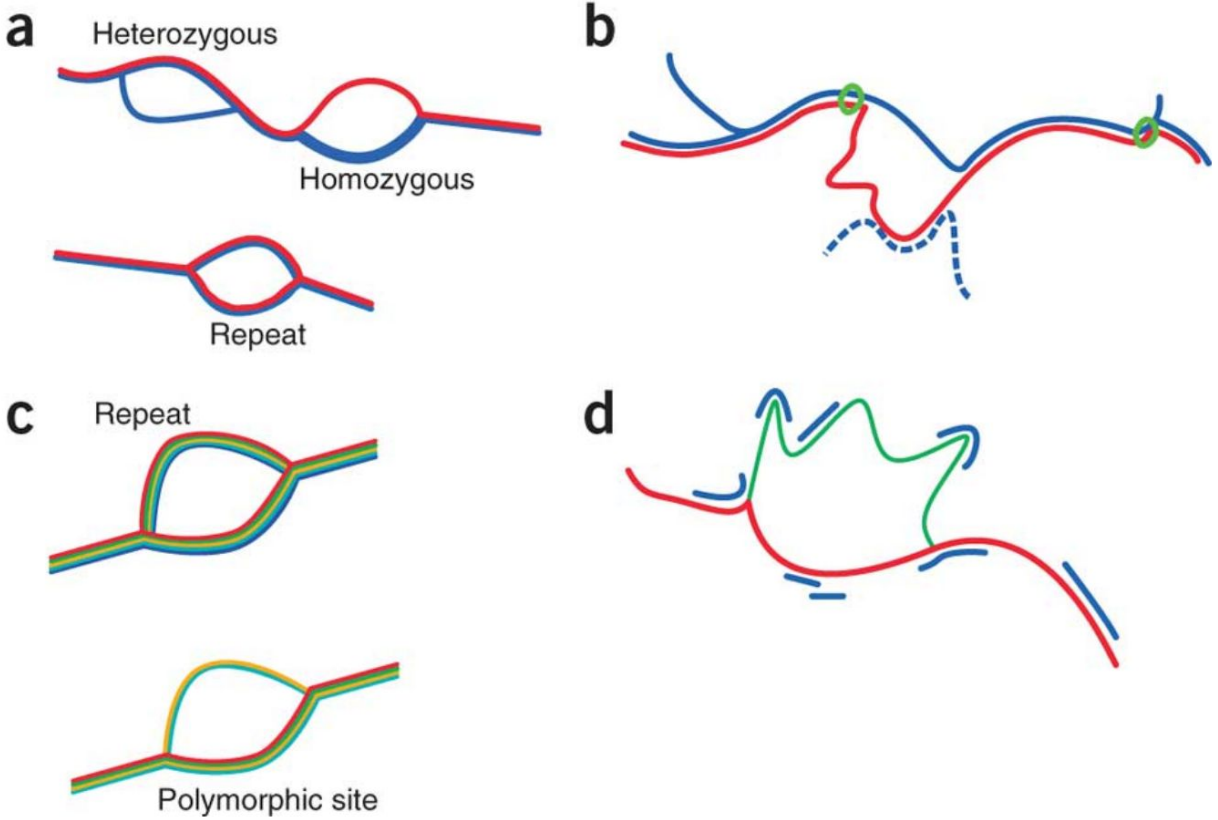
$$C^{\mathrm{T}} = 0000001001010000$$

$$0000000110101001$$

VARI Algorithm

1. Get union of k-mers for all sequences (KMC2), track “colors” (sequences of origin)
2. Construct succinct de Bruijn graph representation
    a. Compress where possible!
3. Traversal (Bubble Calling and Path Divergence)

Types of Bubbles/Divergence



	CORTEX		KMC2		VARI-dBG			VARI-C	
Dataset	**CPU time**	**Mem.**	**CPU time**	**Mem.**	**CPU time**	**Int. Mem.**	**Ext. Mem.**	**CPU time**	**Mem.**
Plants	2h 25m 27s	109,579	19m 50s	4,335	1h 34m 37s	5,388	156,504	3m 09s	3,528
E. coli (k=32)	N/A	N/A	3h 15m 40s	104	9h 30m 11s	126,777	319,328	53m 54s	42,043
E. coli (k=48)	N/A	N/A	4h 35m 29s	149	10h 47m 46s	128,077	427,460	1h 02m 07s	42,100
E. coli (k=64)	N/A	N/A	5h 05m 27s	189	11h 21m 08s	127,523	522,576	1h 09m 07s	42,134
Beef safety	N/A	N/A	34h 04m 46s	11,688	82h 42m 48s	109,091	4,378,840	6h 44m 12s	217,705

Table 2. Data structure construction performance measurements. CPU time is user plus system time as reported by '/bin/time'. (Internal) memory is reported in megabytes and is the maximum resident set size. KMC2 includes both counting and sorting k-mers. VARI-dBG forms the k-mer union and builds the succinct de Bruijn graph. VARI-C compresses the color matrix.

			CORTEX		VARI	
Dataset	No. of k-mers	Colors	Memory	Time	Memory	Time
Plants (k=32)	1,709,427,823	4	100.93 GB	2h 18m	3.53 GB (sdBG=0.89 GB, sC=1.95 GB)	32h 39m
E. coli (k=32)	158,501,209	3,765	N/A	N/A	42.17 GB (sdBG=0.09 GB, sC=38.35 GB)	3h 57m
E. coli (k=48)	205,938,139	3,765	N/A	N/A	42.26 GB (sdBG=0.11 GB, sC=38.42 GB)	4h 38m
E. coli (k=64)	251,764,413	3,765	N/A	N/A	42.32 GB (sdBG=0.13 GB, sC=38.45 GB)	5h 28m
Beef safety (k=32)	40,995,794,366	88	N/A	N/A	245.54 GB (sdBG=27.08 GB, sC=200.34 GB)	N/A

Table 3. Comparison between the peak memory and time usage required to store all the k-mers and run bubble calling on the data in CORTEX and VARI. The peak memory is given in megabytes (MB) or gigabytes (GB). The running time is reported in seconds (s), minutes (m), and hours (h). The succinct de Bruijn graph and compressed color matrix components of the memory footprint are listed in parenthesis as sdBG and sC, respectively.

			Cortex		Vari	
Dataset	No. of k-mers	Colors	Memory	Time	Memory	Time
AMR genes and sample	9,348,365	55	7.08 GB	2m 55s	0.718 GB	29m 21s

Set of 54 antimicrobial resistance (AMR) genes and a simulated metagenomics sample containing seven of these 54 AMR genes and four AMR genes not contained in this set,

Cortex vs VARI

Time

Memory

Targeted vs general traversals

Disk Space???