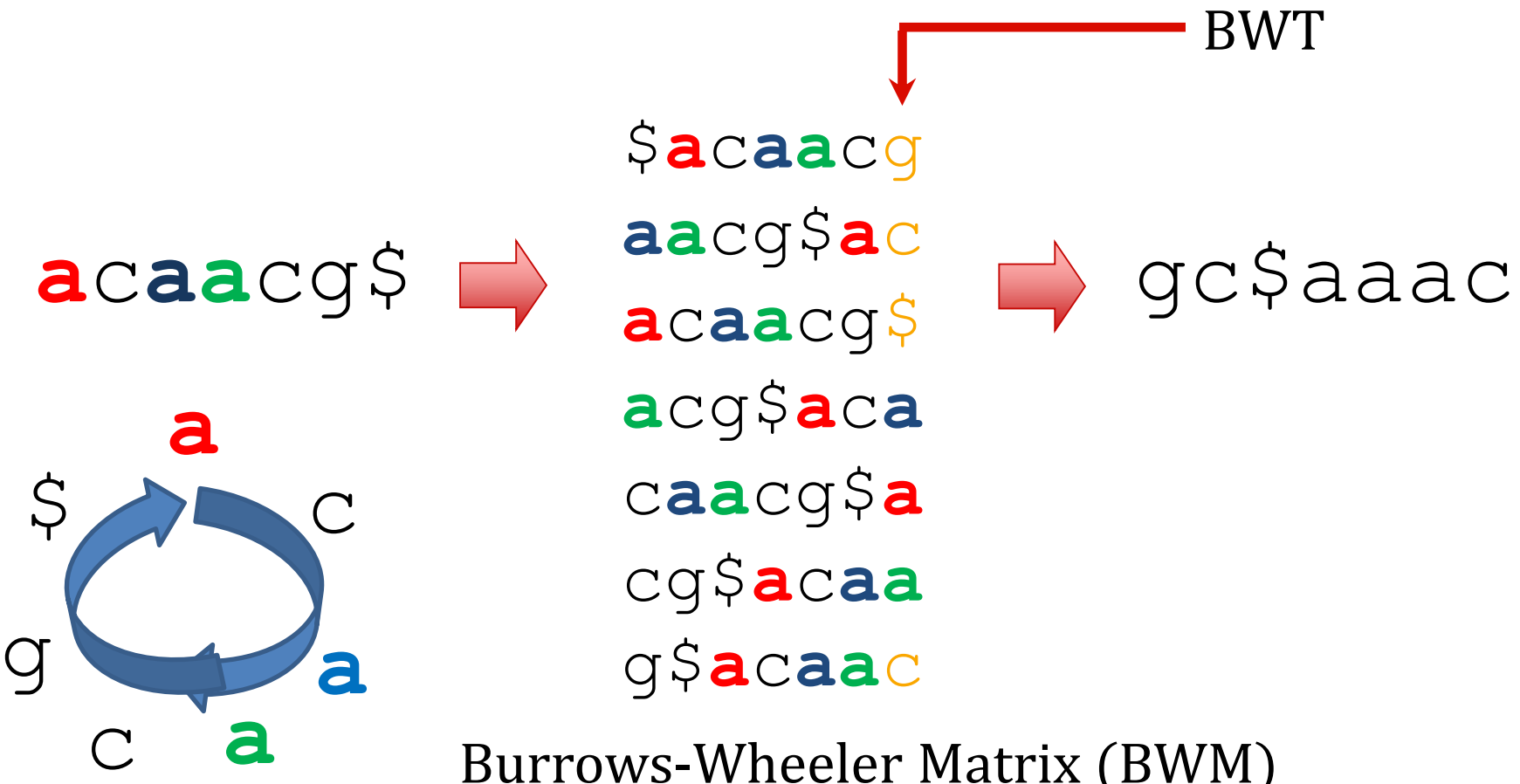
Alignment of Short Reads: Burrows-Wheeler Transform (BWT)

Ahmet Sacan

Slides adapted from:
https://www.coursera.org/learn/algorithms-on-strings
https://ocw.mit.edu/courses/biology/7-91j-foundations-of-computational-and-systems-biology-spring-2014/video-lectures/lecture-5-library-complexity-and-short-read-alignment-mapping/
http://www.cs.utsa.edu/~jruan/teaching/cs5263_spring_2015/slides/slide5_shortreadmapping.ppt

Burrows-Wheeler Transform (BWT)



Burrows-Wheeler Matrix (BWM)

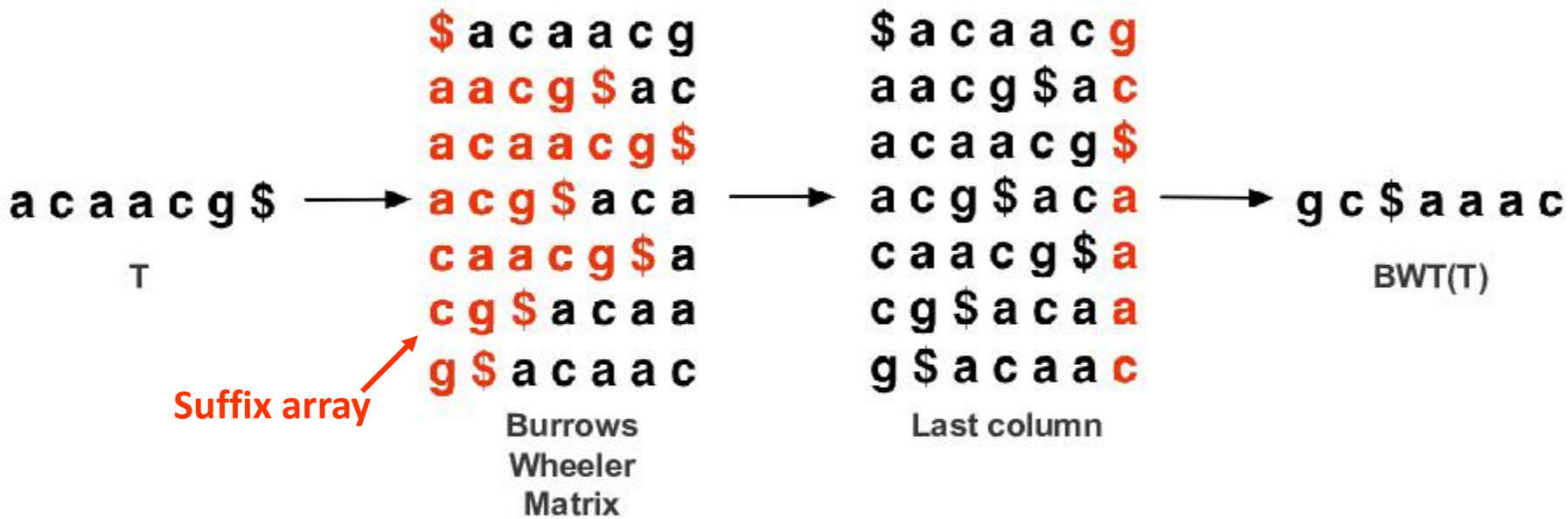
Applying BWT to the Double Helix Paper by Watson&Crick

```
nd Corey (1).  They kindly made their manuscript availa ..... a
nd criticism, especially on interatomic distances.  We  ..... a
nd cytosine.  The sequence of bases on a single chain d ..... a
nd experimentally (3,4) that the ratio of the amounts o ..... u
nd for this reason we shall not comment on it.  We wish ..... a
nd guanine (purine) with cytosine (pyrimidine).  In oth ..... a
nd ideas of Dr.  M. H. F. Wilkins, Dr.  R. E. Franklin  ..... a
nd its water content is rather high.  At lower water co ..... a
nd pyrimidine bases.  The planes of the bases are perpe ..... a
nd stereochemical arguments.  It has not escaped our no ..... a
nd that only specific pairs of bases can bond together  ..... u
nd the atoms near it is close to Furberg's 'standard co ..... a
nd the bases on the inside, linked together by hydrogen ..... a
nd the bases on the outside.  In our opinion, this stru ..... a
nd the other a pyrimidine for bonding to occur.  The hy ..... a
nd the phosphates on the outside.  The configuration of ..... a
nd the ration of guanine to cytosine, are always very c ..... a
nd the same axis (see diagram).  We have made the usual ..... u
nd their co-workers at King's College, London.  One of  ..... a
```

"and" is a frequent repeat in English texts

The Burrows-Wheeler Transform is a reversible representation with handy properties

- Sort all the possible rotations of original string



- Once BWT(T) is built, *all else shown here is discarded*
  - Matrix will be shown for illustration only

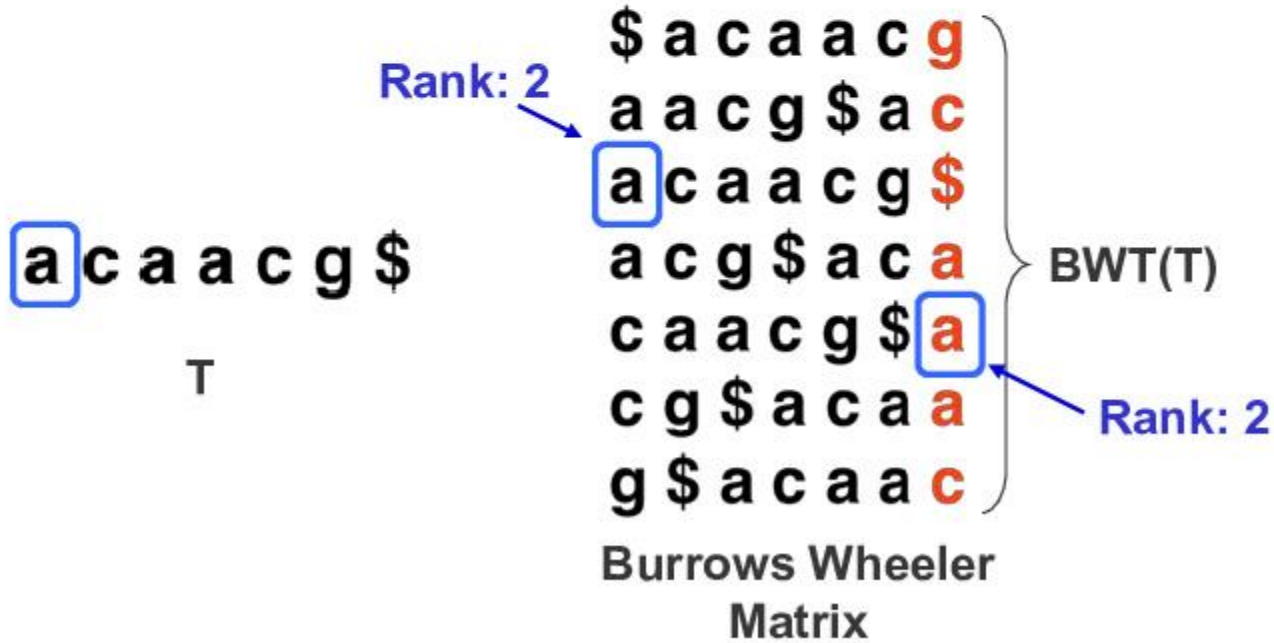
Burrows M, Wheeler DJ: **A block sorting lossless data compression algorithm**. *Digital Equipment Corporation, Palo Alto, CA* 1994, Technical Report 124; 1994

Recovering the original string

- If the BWT(s) is: "gc$aaac", what is the original string s ?

A text occurrence has the same rank in the first and last columns

- When we rotate left and sort, the first character retains its rank. Thus the same text occurrence of a character has the same rank in the **L**ast and **F**irst columns.



Courtesy of Ben Langmead. Used with permission.

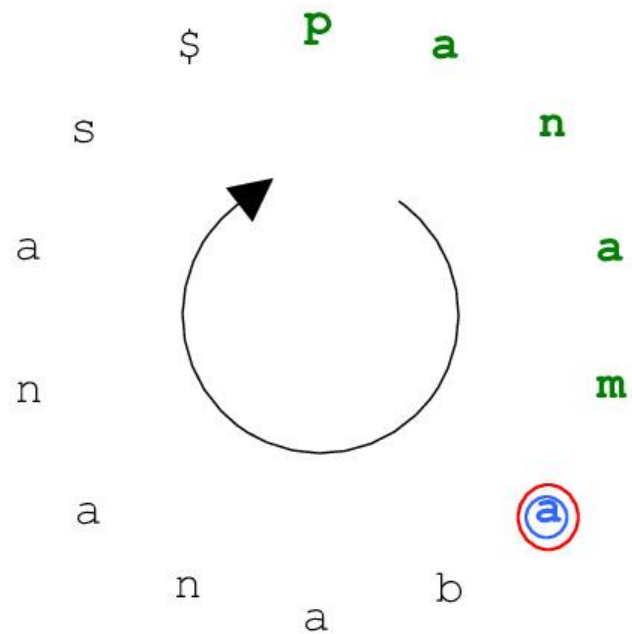
http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

A Strange Observation

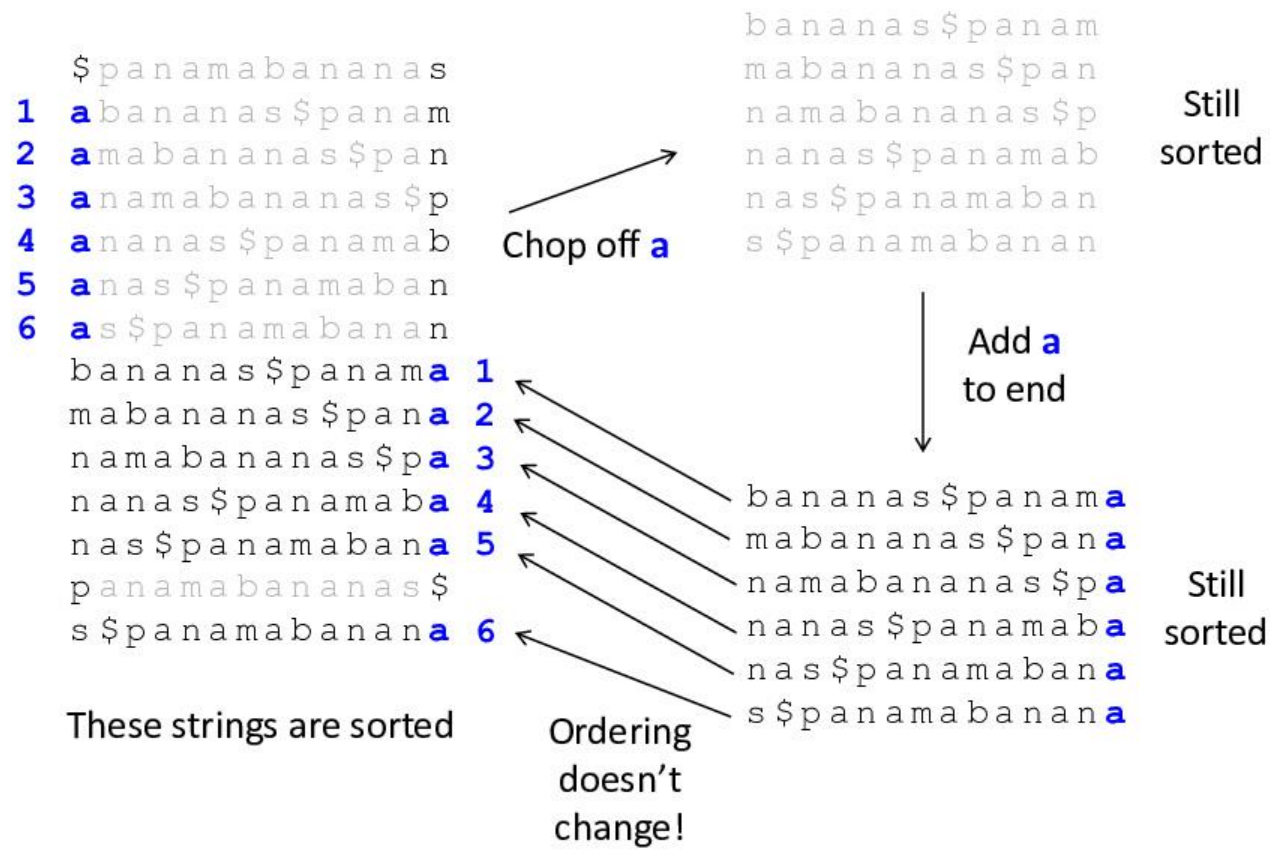
Where is first "a" hiding inside the circle?

$panamabananas
abananas$panam
amabananas$pan
anamabananas$p
ananas$panamab
anas$panamaban
as$panamabanan
bananas$panama
mabananas$pana
namabananas$pa
nanas$panamaba
nas$panamabana
panamabananas$
s$panamabanana

Where is first "a" hiding inside the circle?

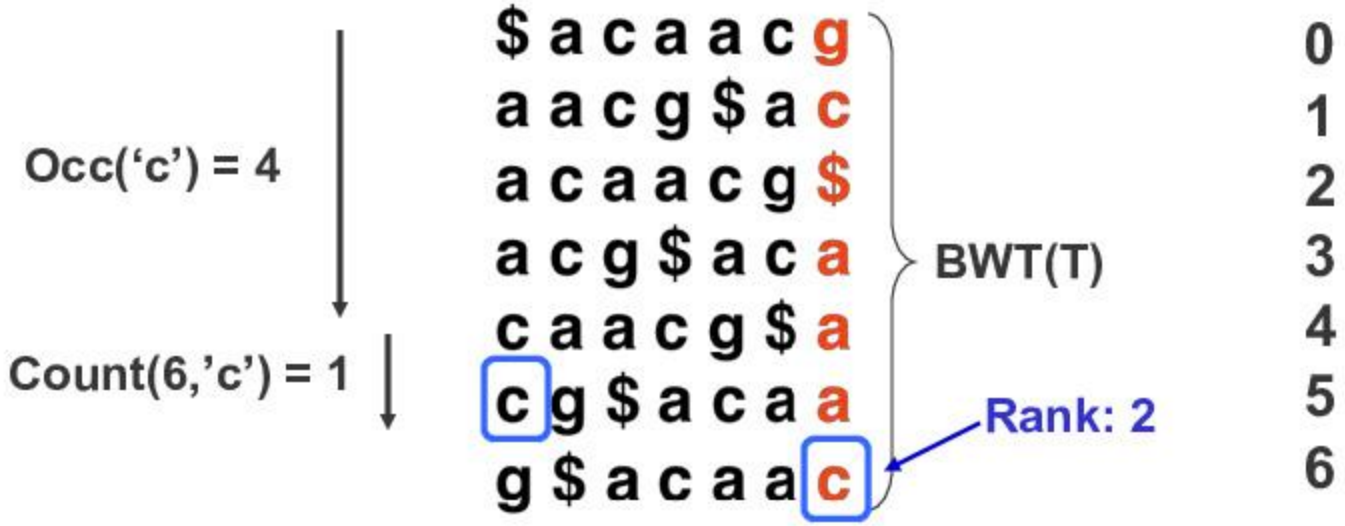


Is It True in General?



The Last to First (LF) function matches character and rank

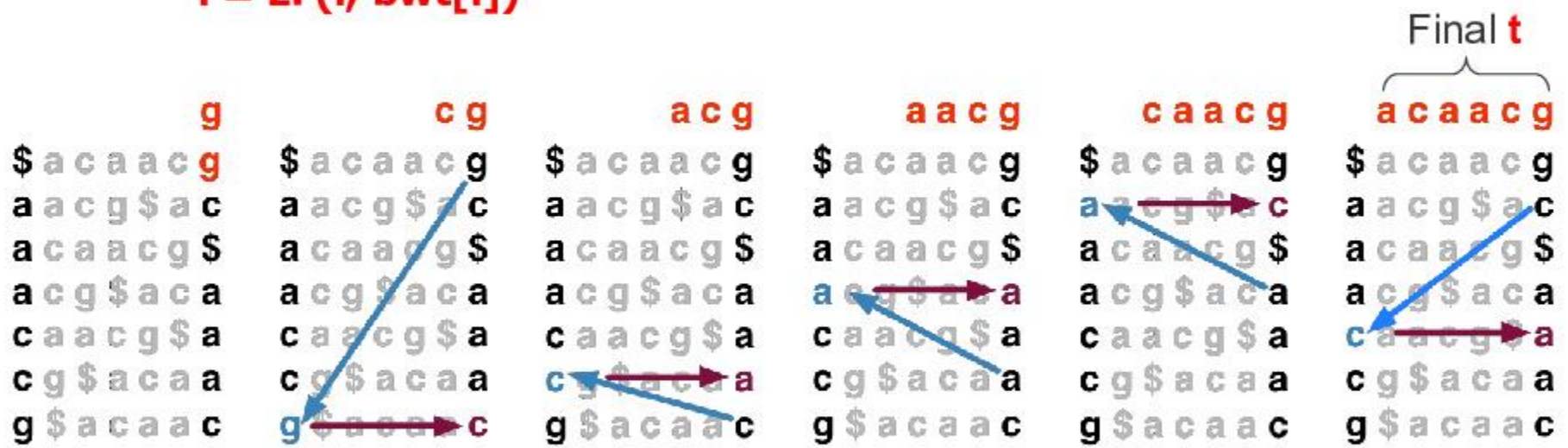
LF(6, 'c') = Occ('c') + Count(6,'c') = 5



- Occ(qc): First occurrence of the character type qc.
- Count(ind, qc): How many characters of type qc are before ind?

Walk Left to Invert BWT (recover original)

```
i = 0
t = ""
while bwt[i] != '$':
    t = bwt[i] + t
    i = LF(i, bwt[i])
```



Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

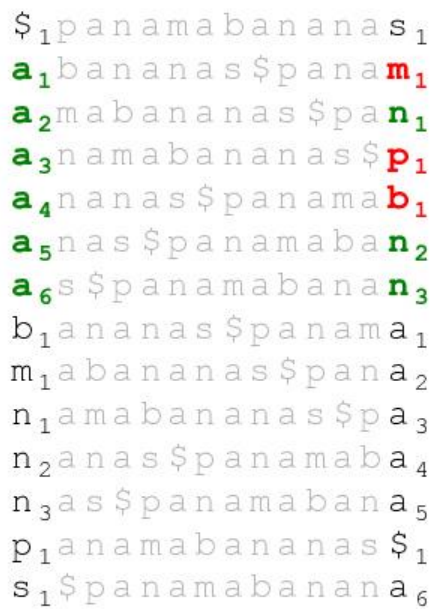
Find **ana** in **panamabananas**

Lets Start by Matching the Last Symbol (**a**)

- Searching for an**a** in panamabananas

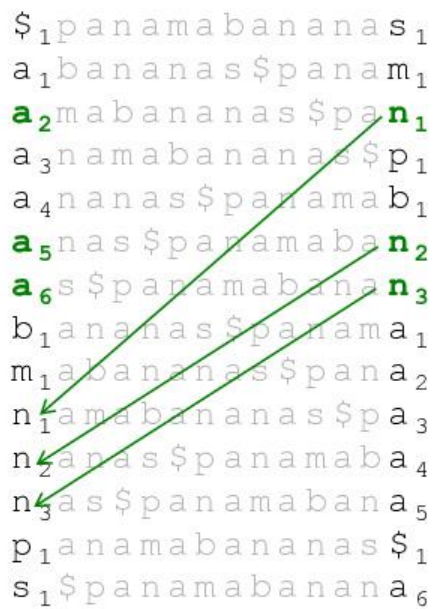
Matching the Last Two Symbols (na)

- Searching for a**na** in panamabananas



Three Matches of **na** Found!

- Searching for a**na** in panamabananas

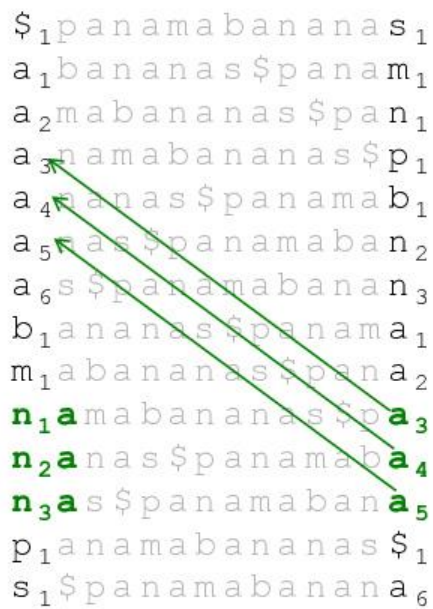


Three Matches of **na** Found!

- Searching for a**na** in panamabananas

Matching **ana**

- Searching for **ana** in panamabananas



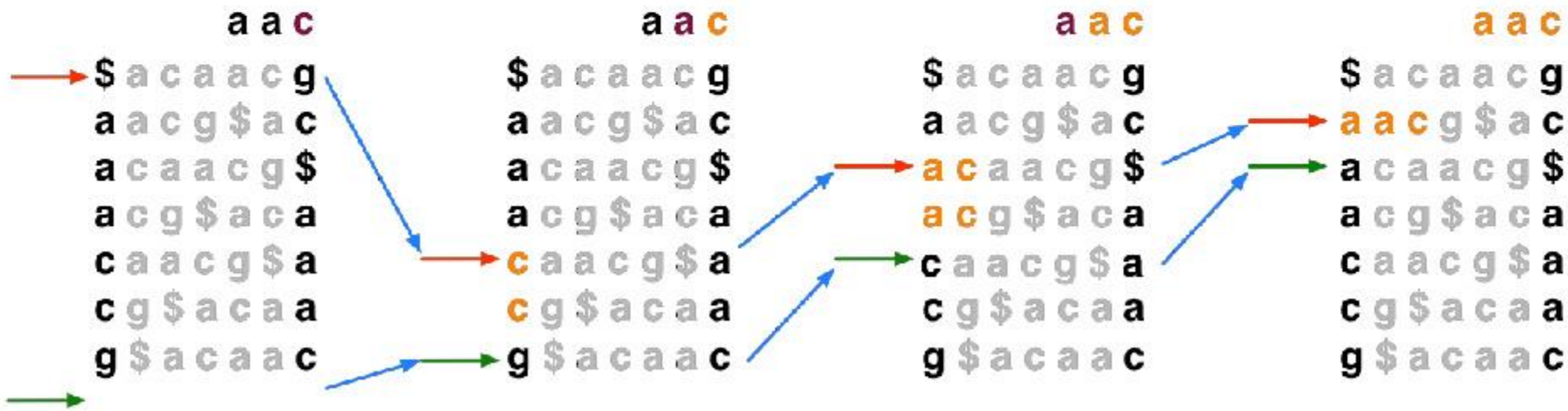
Three Matches of **ana** Found!

- Searching for **ana** in panamabananas

Exact Matching with FM Index

```
q = "aac"
top = 0
bot = len(bwt)
for qc in reverse(q):
   top = LF(top, qc)
   bot = LF(bot, qc)
```

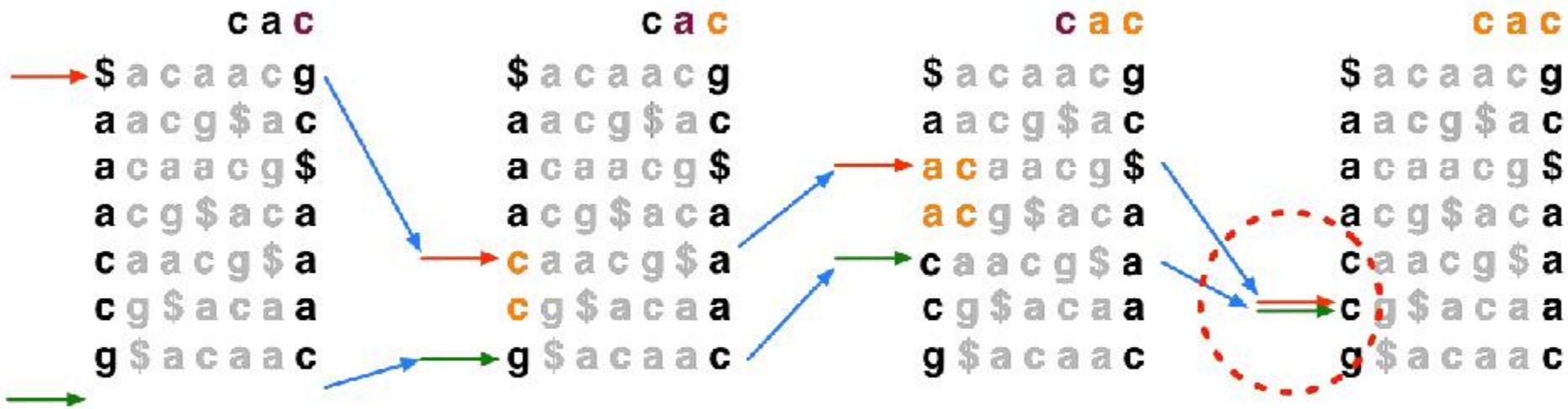
In each iteration top & bot delimit the range of rows beginning with progressively longer suffixes of q



Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

Exact Matching with FM Index

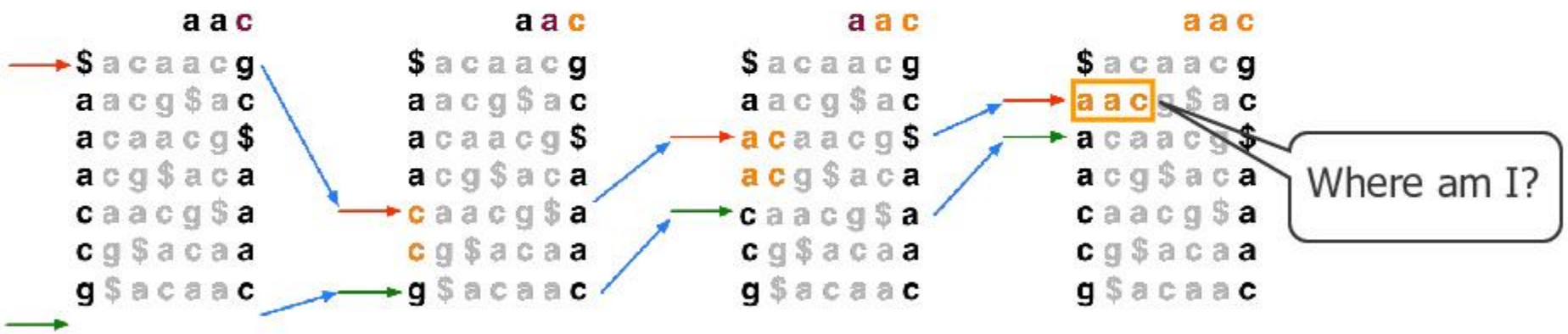


- If range becomes empty (**top** = **bot**) the query suffix (and therefore the query) does not occur in the text

Courtesy of Ben Langmead. Used with permission.

Rows to Reference Positions

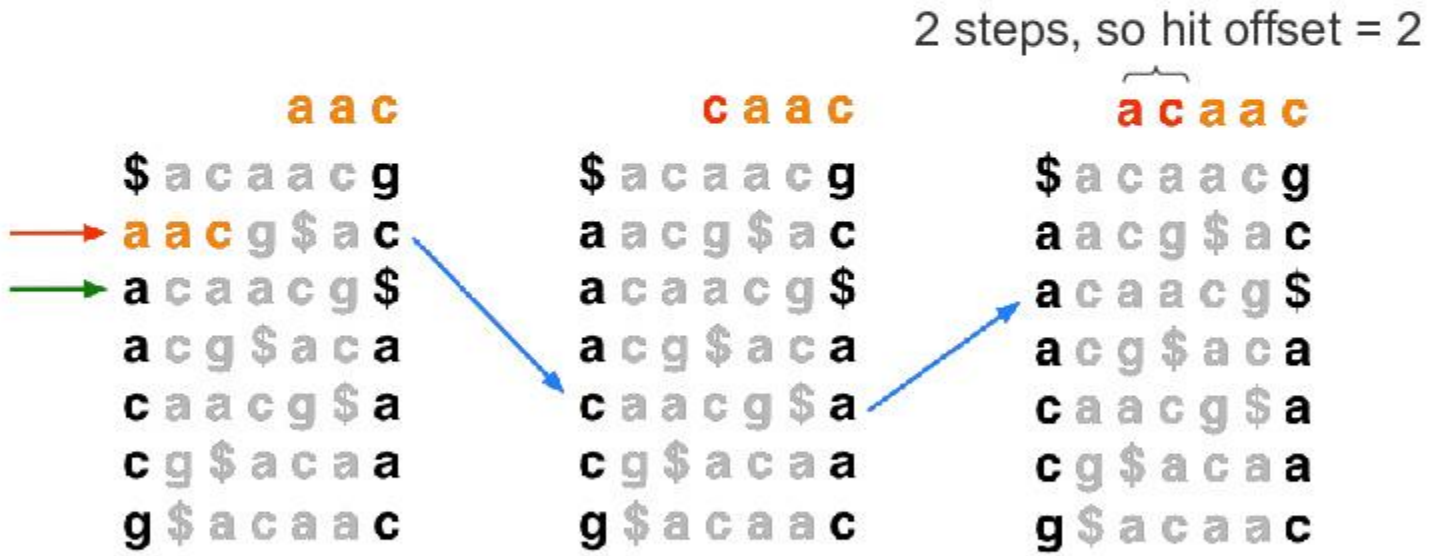
- Once we know a row contains a legal alignment, how do we determine its position in the reference?



Courtesy of Ben Langmead. Used with permission.

Rows to Reference Positions

- Naïve solution 1: Use "walk-left" to walk back to the beginning of the text; number of steps = offset of hit

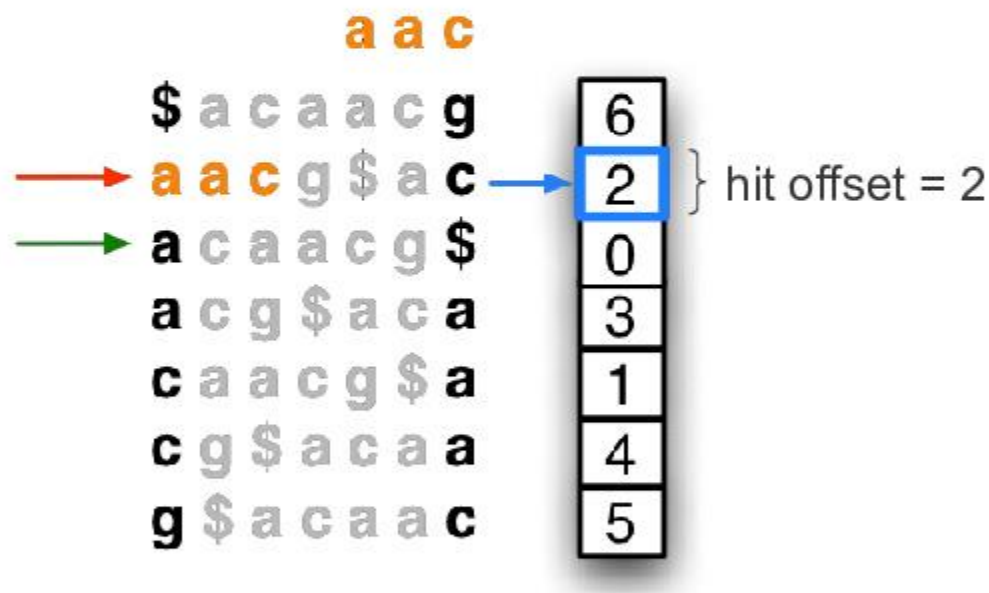


- Linear in length of text in general – too slow

Courtesy of Ben Langmead. Used with permission.

Rows to Reference Positions

- Naïve solution 2: Keep whole suffix array in memory. Finding reference position is a lookup in the array.



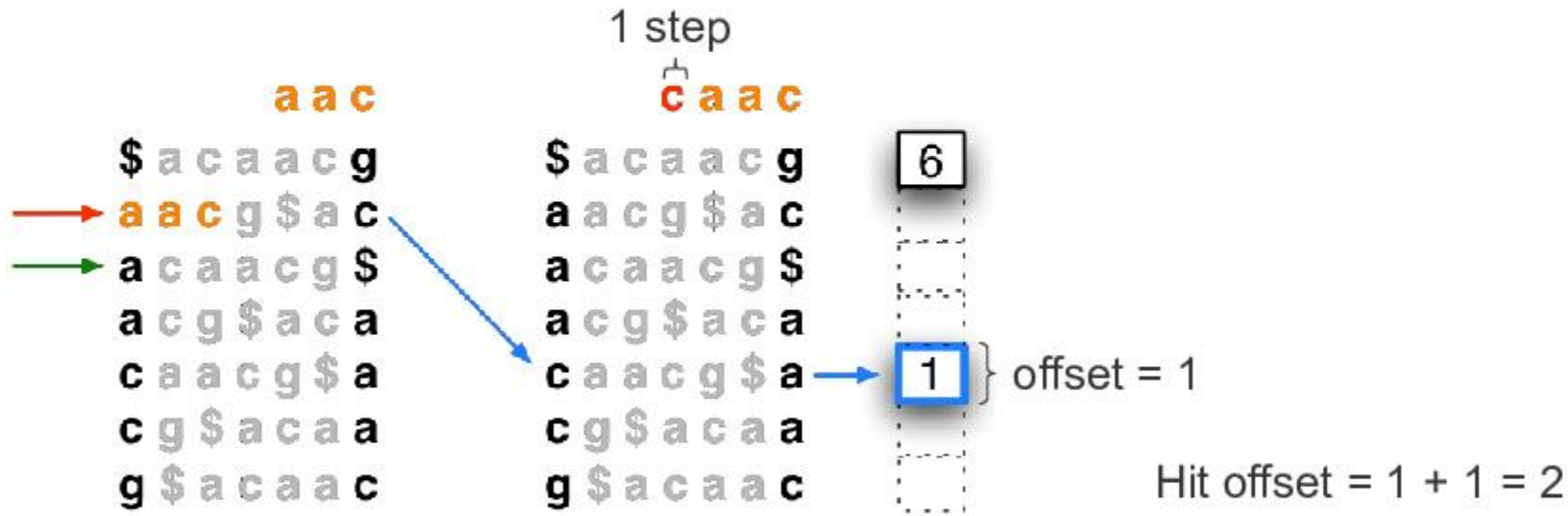
- Suffix array is ~12 gigabytes for human – too big

Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

Rows to Reference Positions

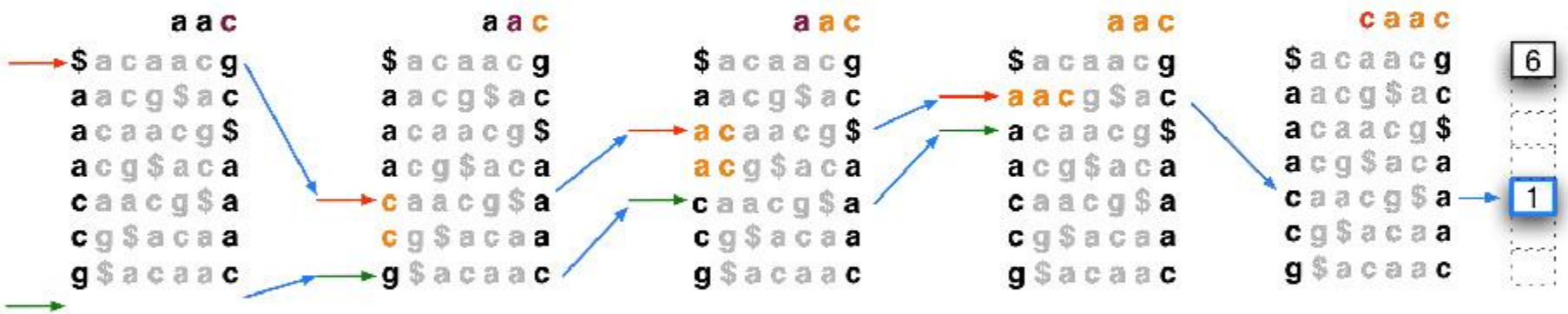
- Hybrid solution: Store *sample* of suffix array; "walk left" to next sampled ("marked") row to the left
  - Due to Ferragina and Manzini



- Bowtie marks every 32nd row by default (configurable)

Courtesy of Ben Langmead. Used with permission.

Put It All Together

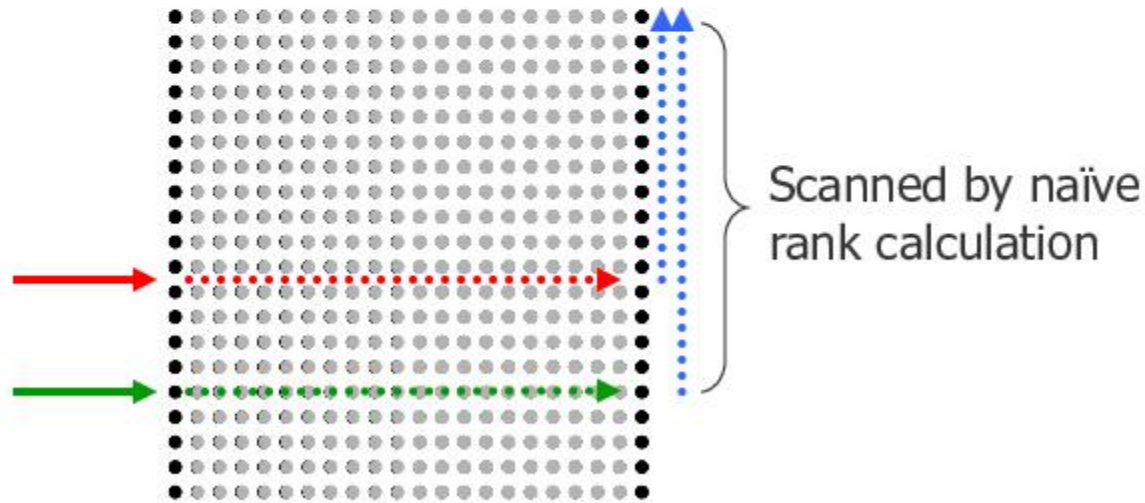


- Algorithm concludes: “aac” occurs at offset 2 in “acaacg”

Courtesy of Ben Langmead. Used with permission.

The FM index makes LF fast

- **LF**(i, **qc**) must determine the *rank* of **qc** in row i
- Naïve way: count occurrences of **qc** in all previous rows
  - This **LF**(i, **qc**) is linear in length of text – too slow

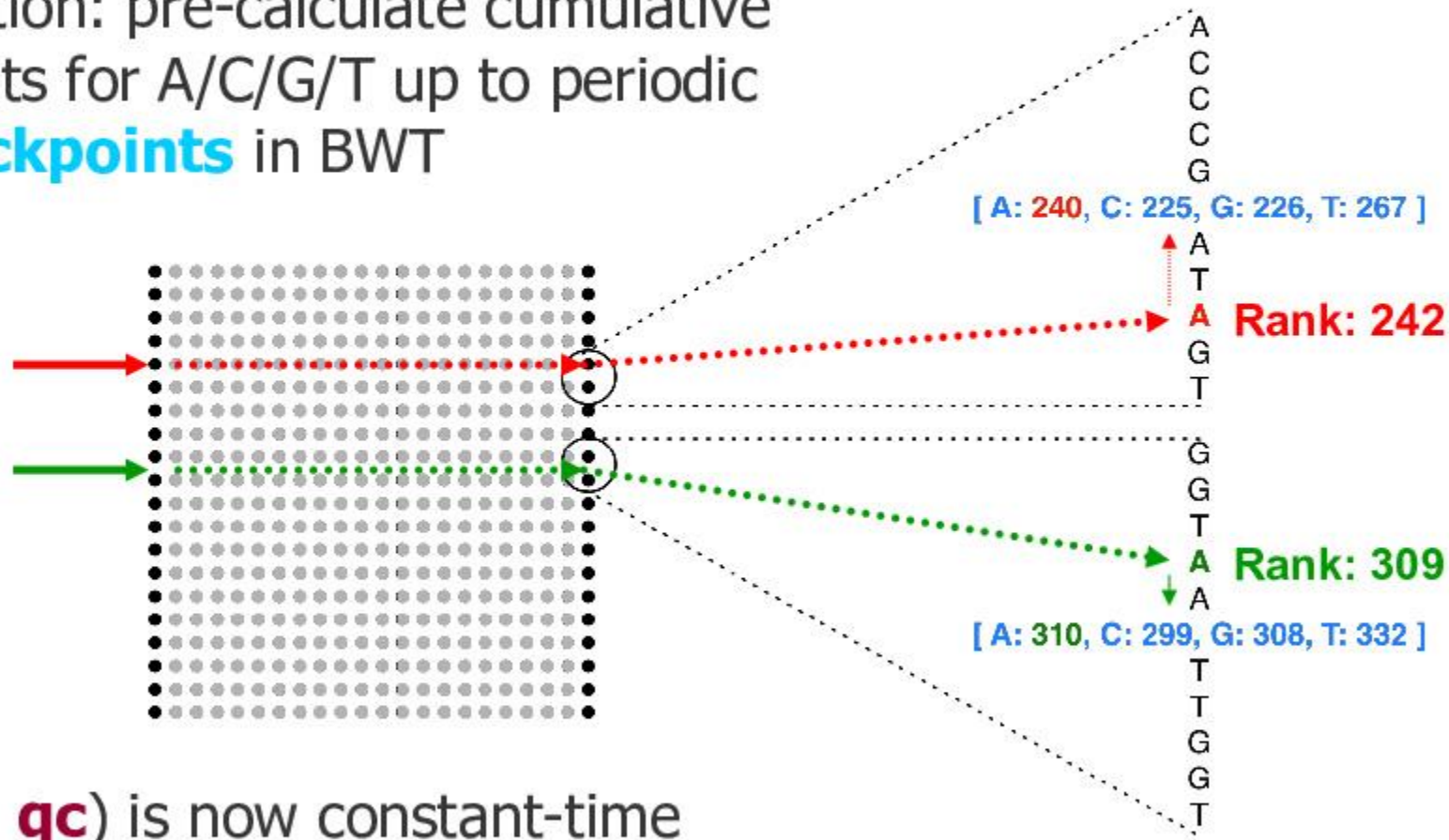


Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

A Full-text Minute-size (FM) index makes LF constant time

- Solution: pre-calculate cumulative counts for A/C/G/T up to periodic **checkpoints** in BWT



- **LF**(i, **qc**) is now constant-time
  (if space between checkpoints is considered constant)

Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

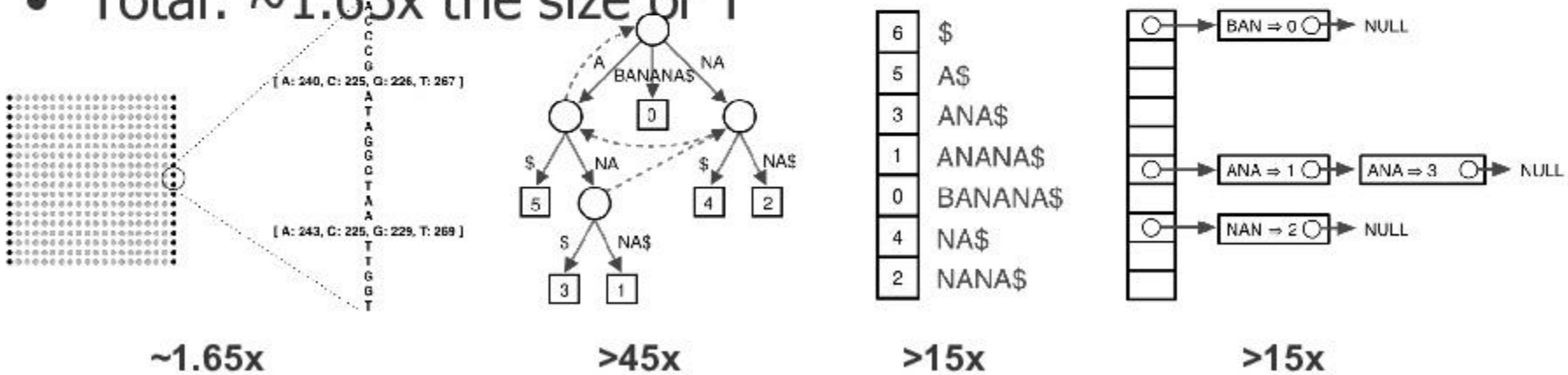
An FM Index is Small

- Entire FM Index on DNA reference consists of:
  - BWT (same size as T)
  - Checkpoints (~15% size of T)
  - Suffix array sample (~50% size of T)
- Total: ~1.65x the size of T

Assuming 2-bit-per-base encoding and no compression, as in Bowtie

Assuming a 16-byte checkpoint every 448 characters, as in Bowtie

Assuming Bowtie defaults for suffix-array sampling rate, etc



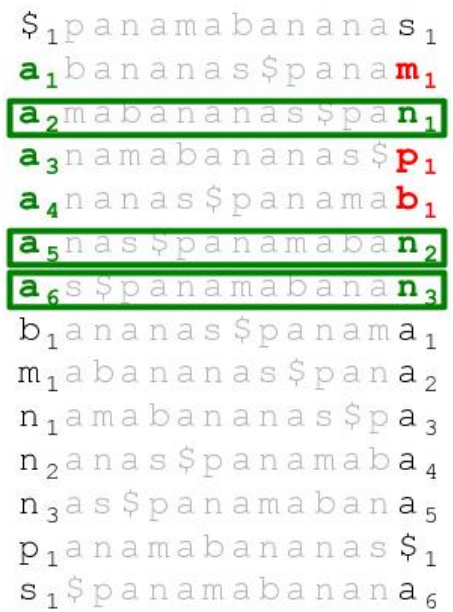
~1.65x >45x >15x >15x

Courtesy of Ben Langmead. Used with permission.

http://www.cbcb.umd.edu/~langmead/NCBI_Nov2008.ppt

How about **Approximate** Matching ?

- searching for a**na** in panamabananas

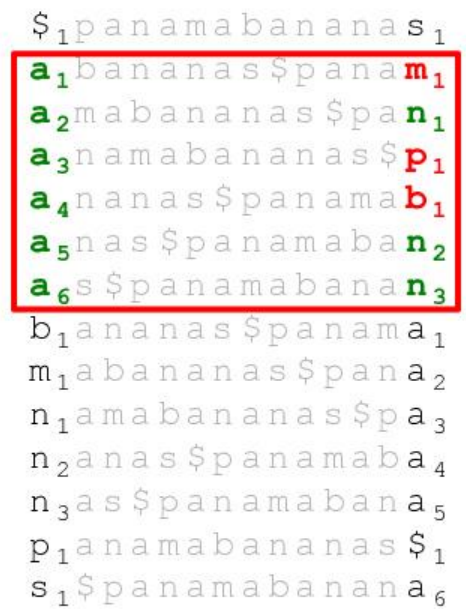


Exact matching

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

To allow for 1 mismatch, we need to analyze the rows ending in red letters as well.

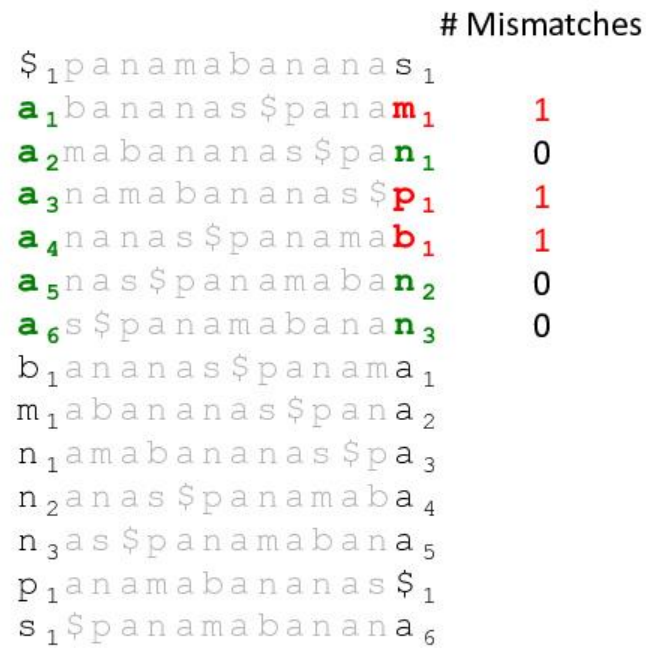


Approximate matching with at most 1 mismatch

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

To allow for 1 mismatch, we need to analyze the rows ending in red letters as well.



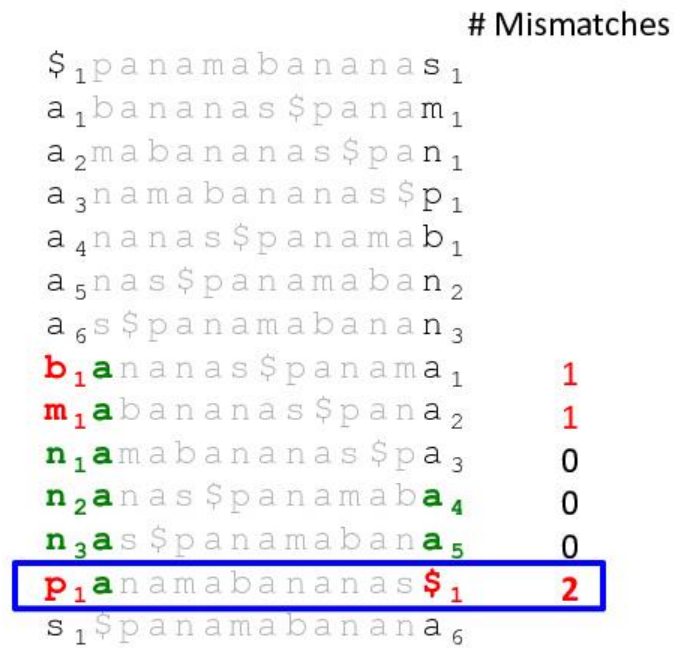
Mismatches
1
0
1
1
0
0

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

Now we analyze all rows with at most 1 mismatch using the First-Last property.

This row results in a 2nd mismatch (the $), so we discard it.



Five Approximate Matches Found!

- searching for **ana** in panamabananas