

# Burrows-Wheeler Transform and Suffix Arrays

Pavel Pevzner

Department of Computer Science and Engineering  
University of California at San Diego

**Algorithms on Strings**  
**Data Structures and Algorithms**

This slide deck is incomplete.  
For the complete set of frames,  
please see our videos in the  
[Algorithms on Strings](#) course on [Coursera](#)  
([Algorithms and Data Structures](#) Specialization)

# Outline

- **Burrows-Wheeler Transform**
- Inverting Burrows-Wheeler Transform
- Using BWT for Pattern Matching
- Suffix Arrays
- Approximate Pattern Matching

# Text Compression by Run-Length Encoding

- **Run-length encoding** compresses a run of  $n$  identical symbols:

*Text*

GGGGGGGGGGCCCCCCCCCCCCAAAAAATTTTTTTTTTTTTTTTTTCCCCCG

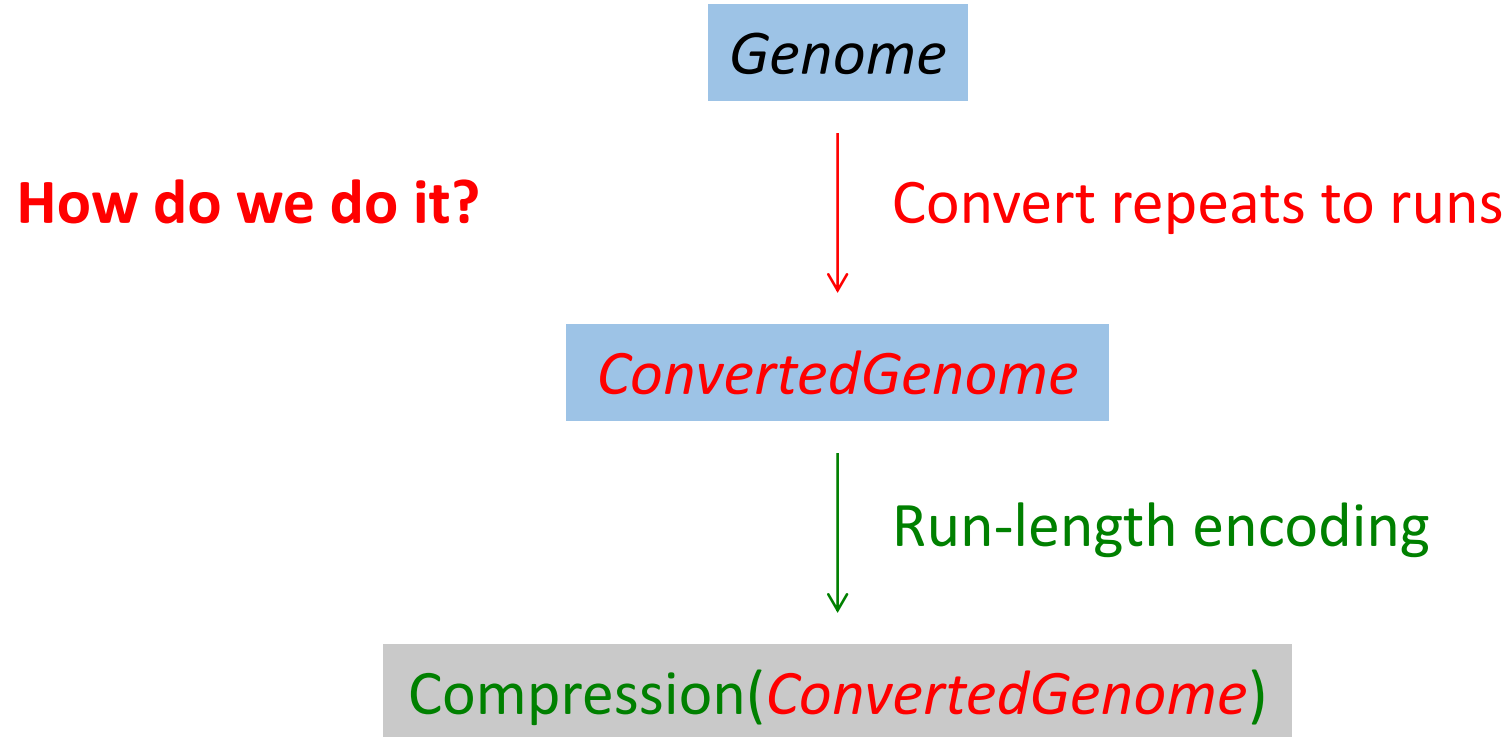
↓

10G11C7A15T5C1G

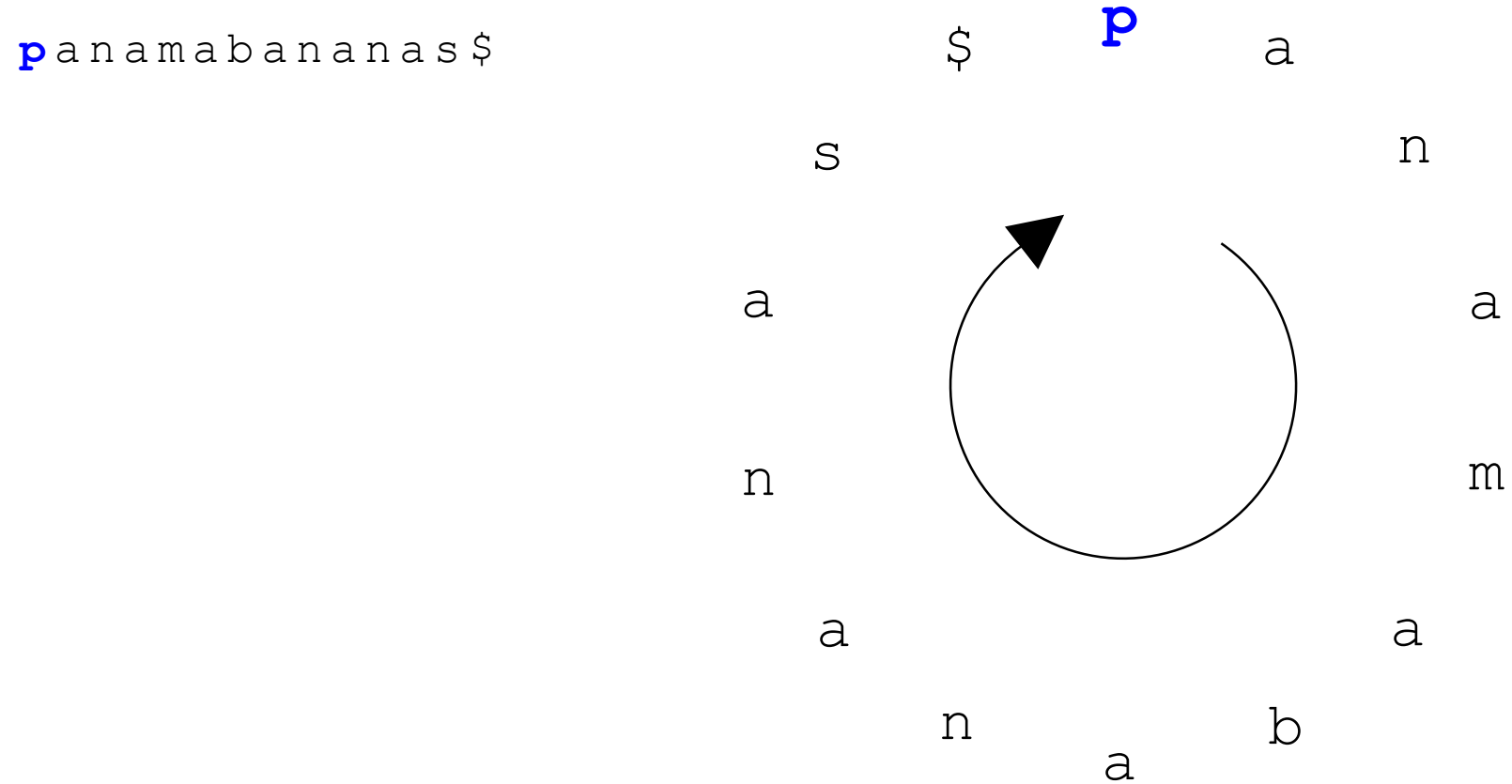
- genomes don't have lots of runs... but they do have lots of repeats:

ACTGACCGAACTGAGTATCCGACTGAACTGATCAGTACTGACATTGC

# Idea: Converting Repeats to Runs

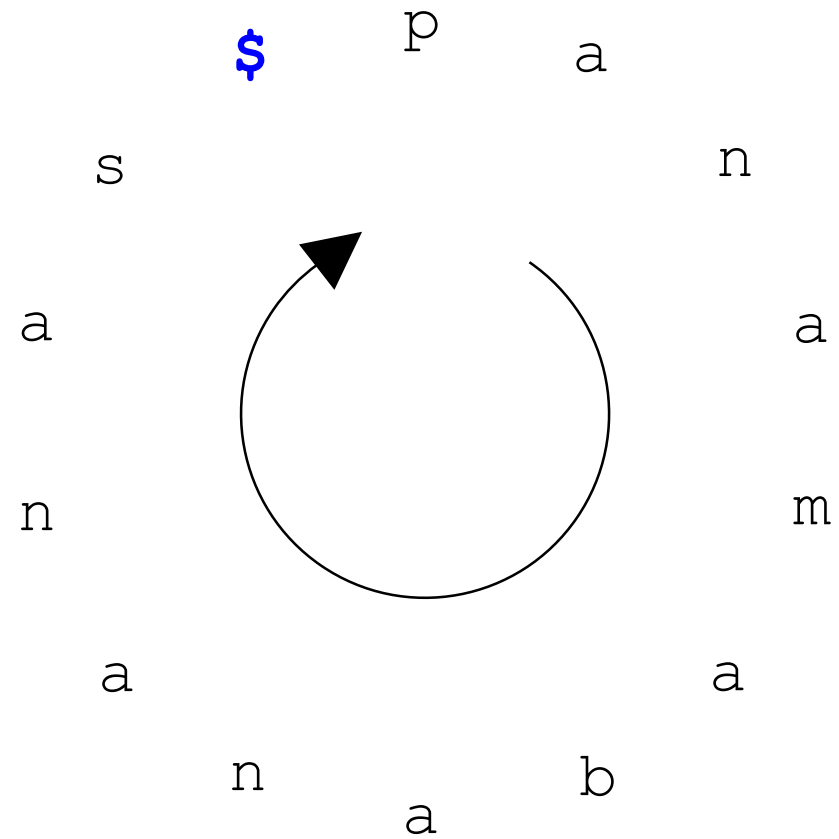


# Forming All Cyclic Rotations of *Text*



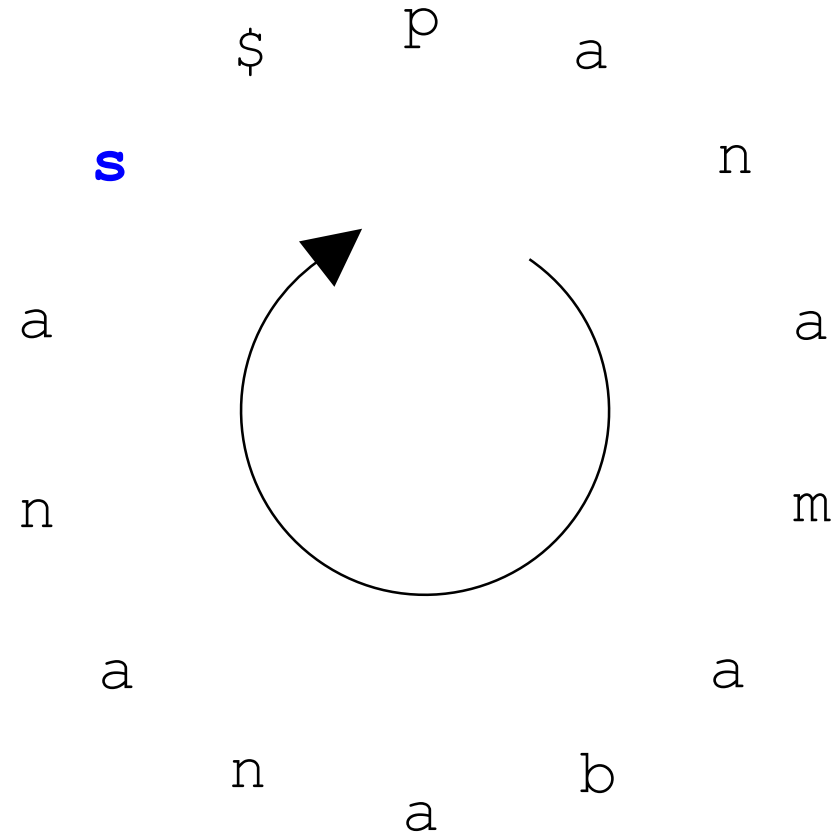
# Cyclic Rotations

panamabananas\$  
**\$panamabananas**



# Cyclic Rotations

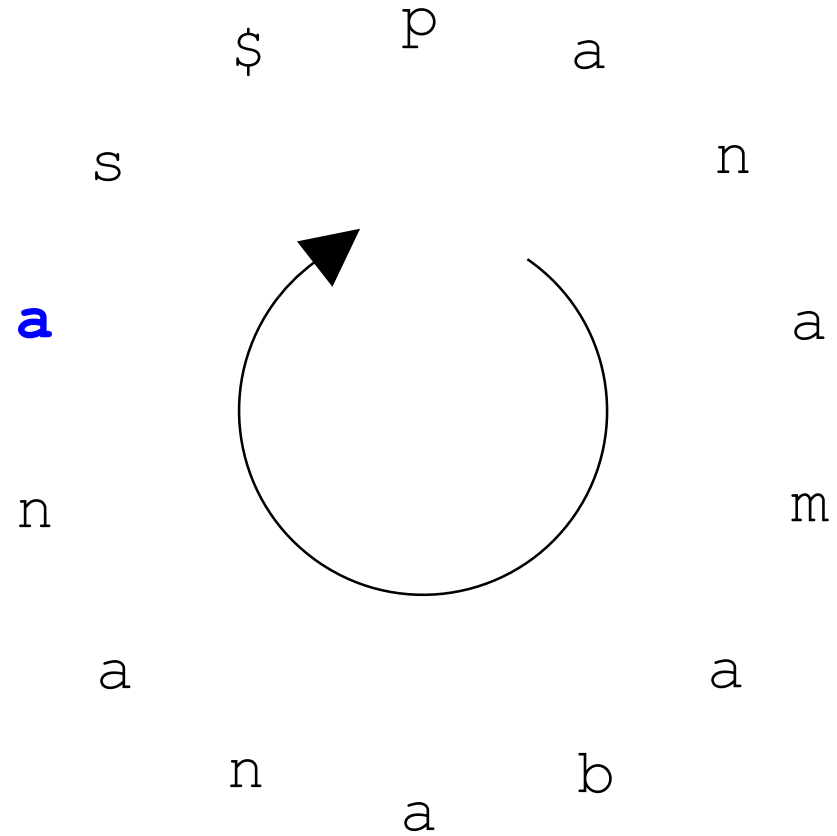
panamabananas\$  
\$panamabananas  
**s\$panamabanana**





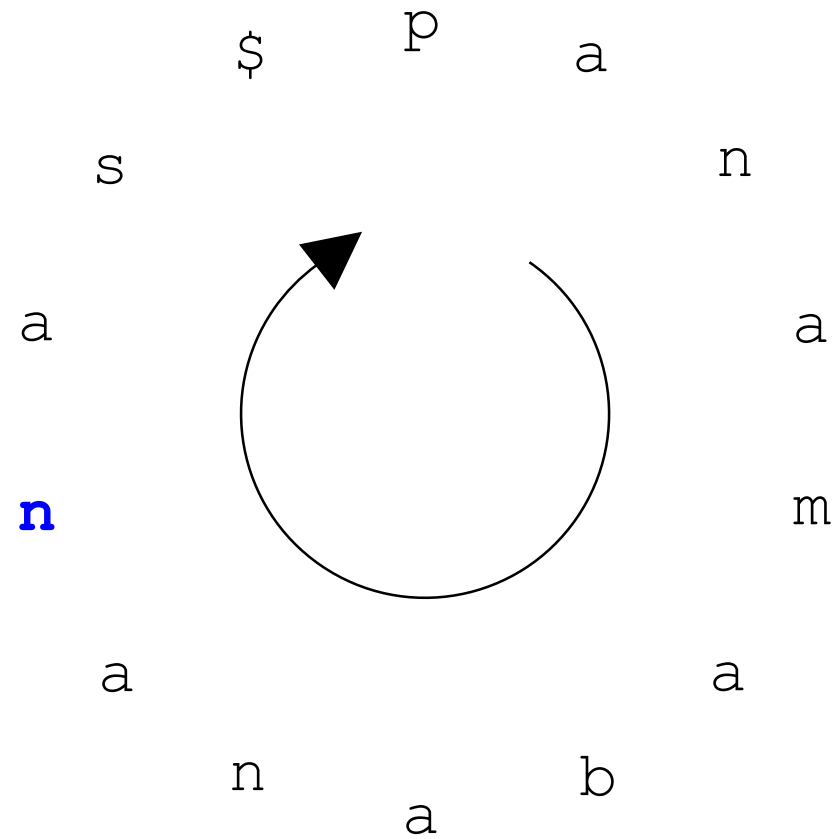
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
**a**s\$panamabanana



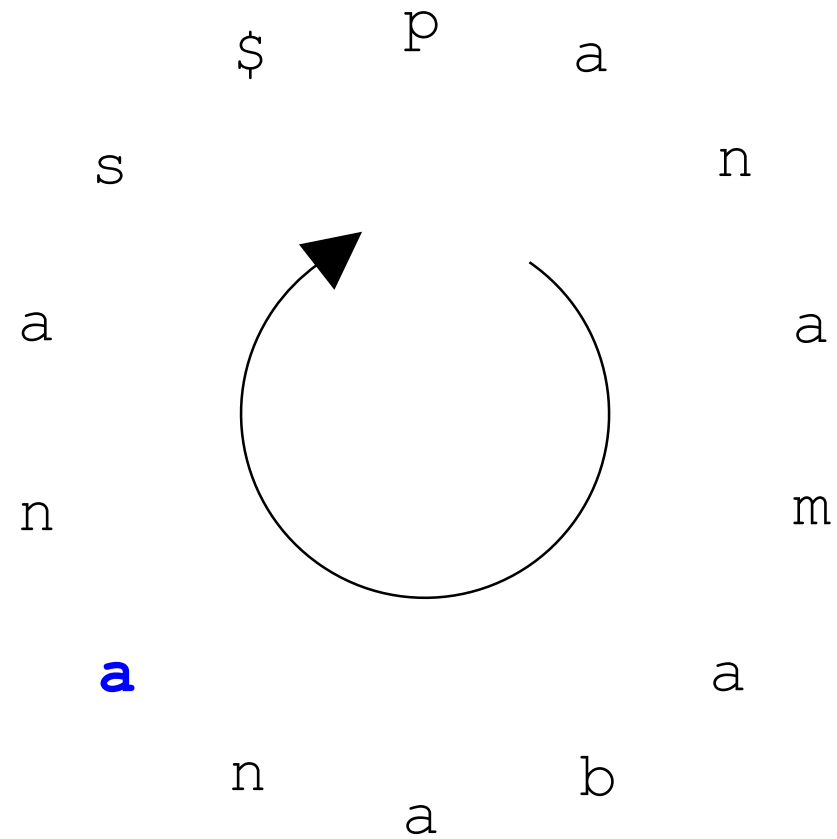
# Cyclic Rotations

```
panamabananas$  
$panamabananas  
s$panamabanana  
as$panamabanana  
nas$panamabana
```



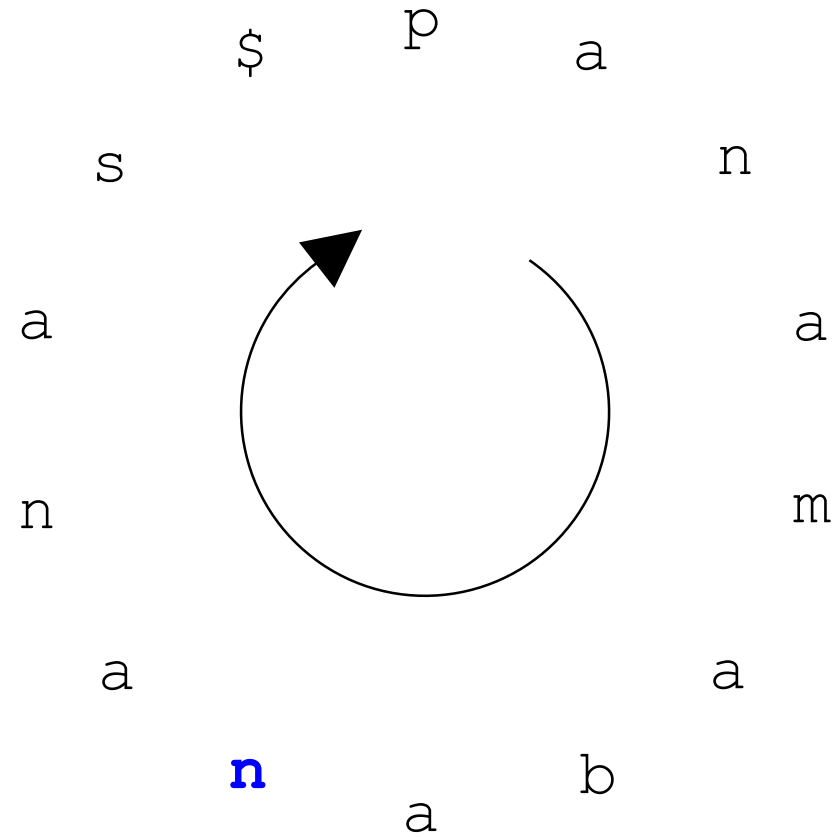
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
**anas\$panamaban**



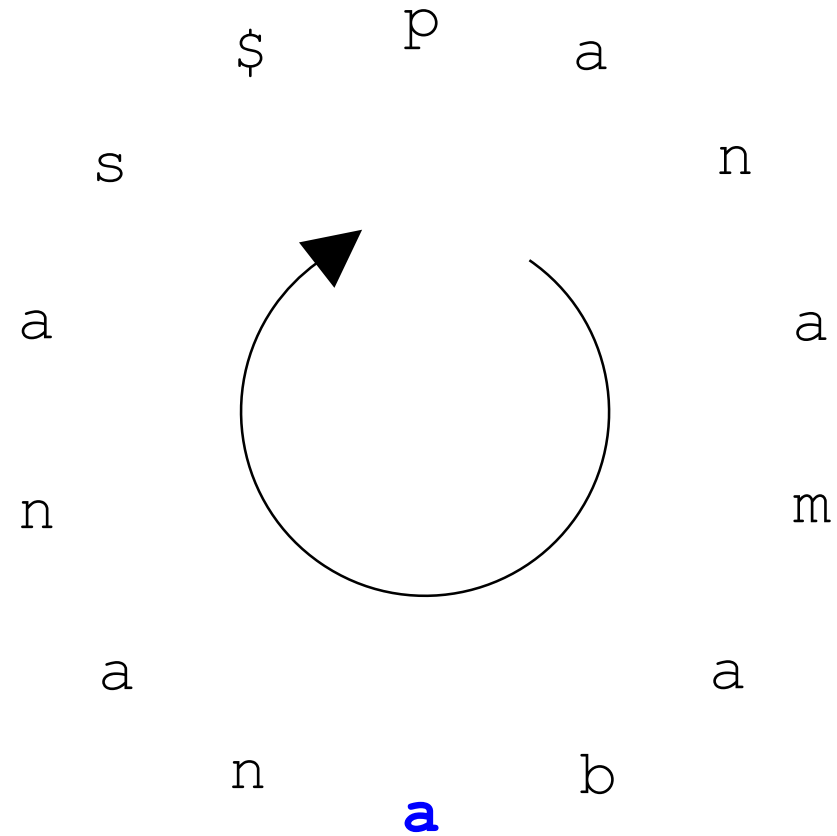
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamaban  
**n**anas\$panamaba



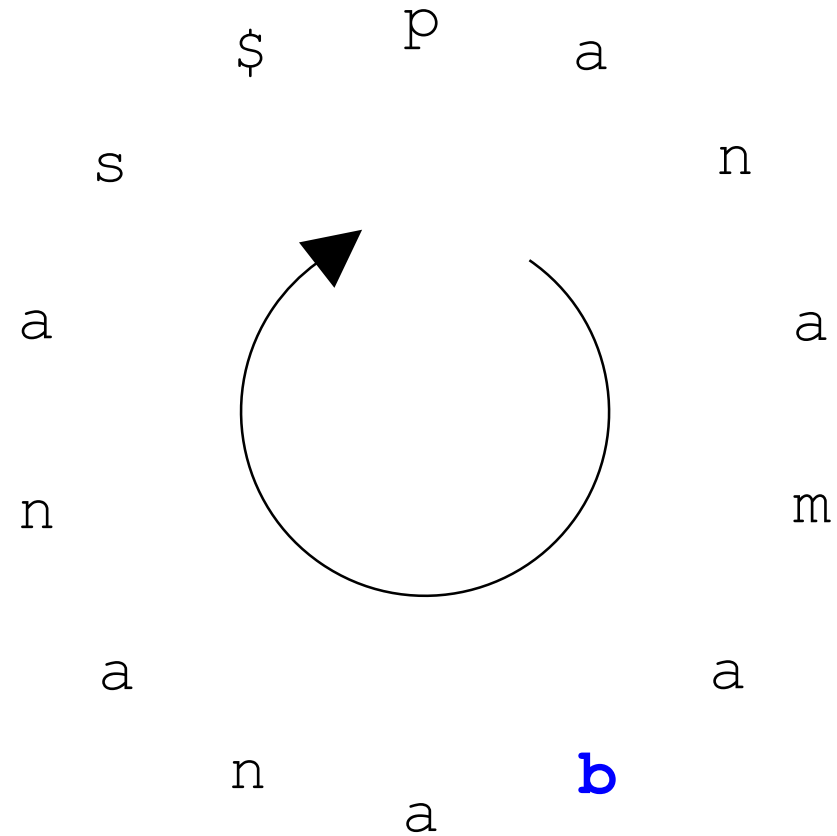
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamabana  
nanas\$panamaba  
**ananas\$panamab**



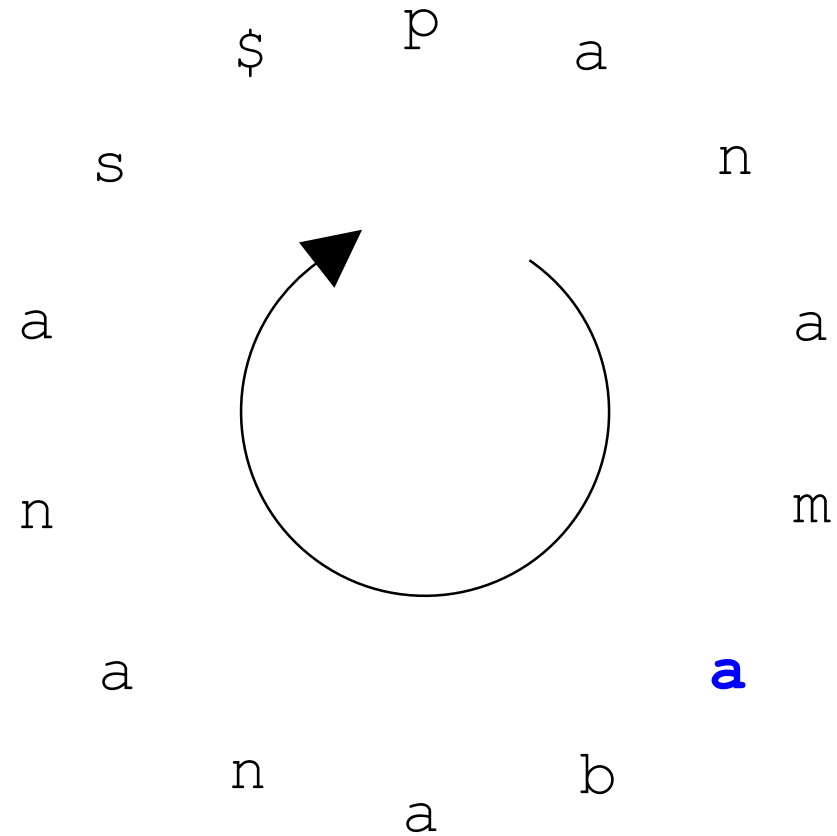
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamabana  
nanas\$panamaba  
ananas\$panamab  
**bananas\$panama**



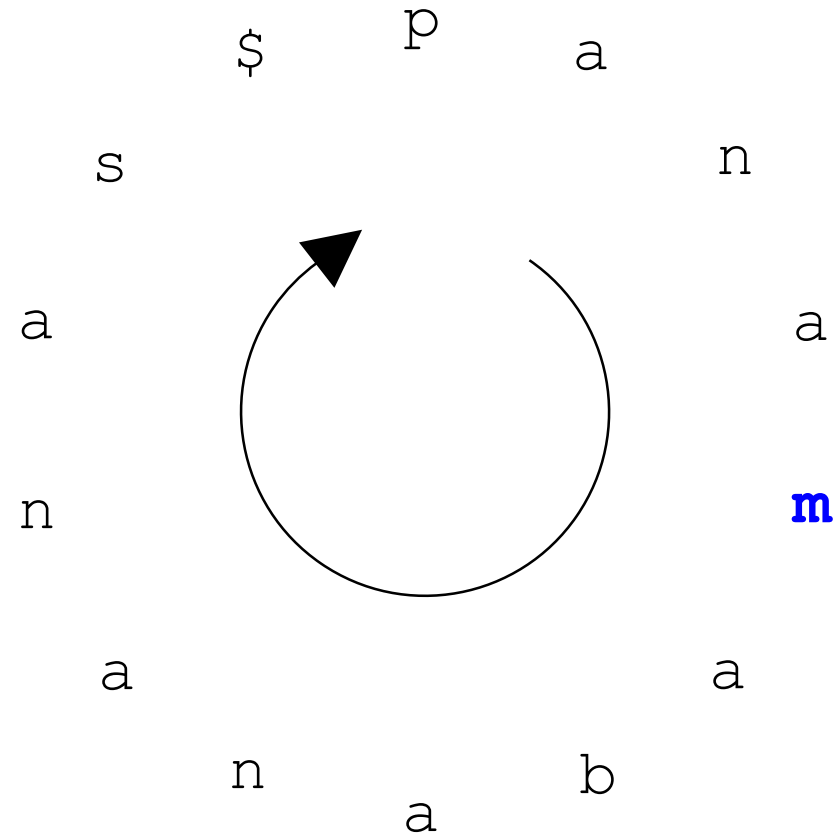
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamabana  
nanas\$panamaba  
ananas\$panamab  
bananas\$panama  
**a**bananas\$panam



# Cyclic Rotations

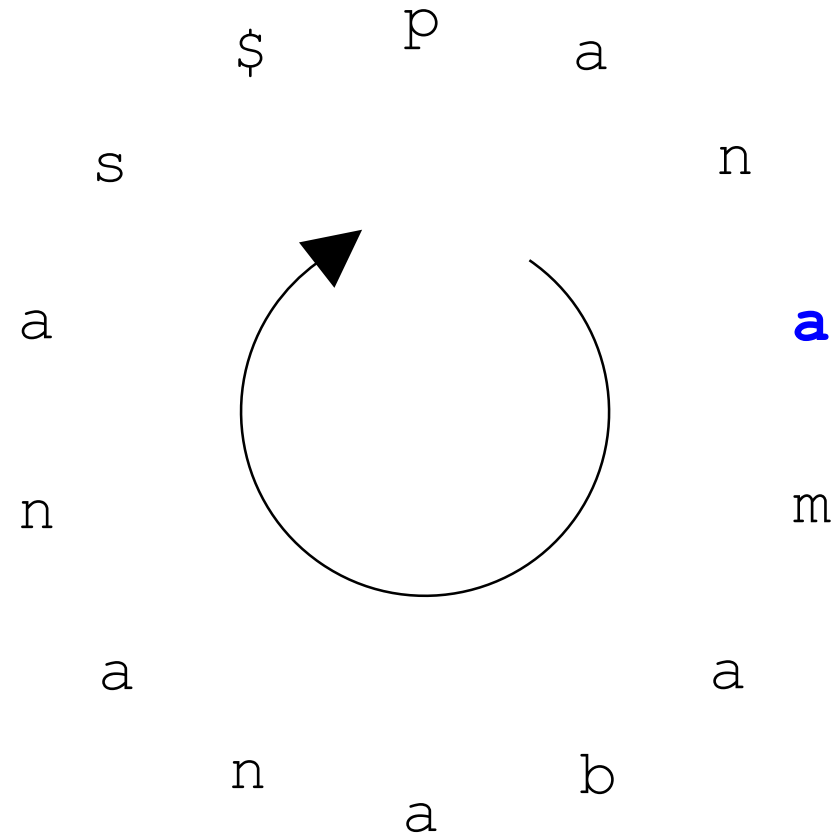
panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamabana  
nanas\$panamaba  
ananas\$panamab  
bananas\$panama  
abananas\$panam  
**mabananas\$pana**





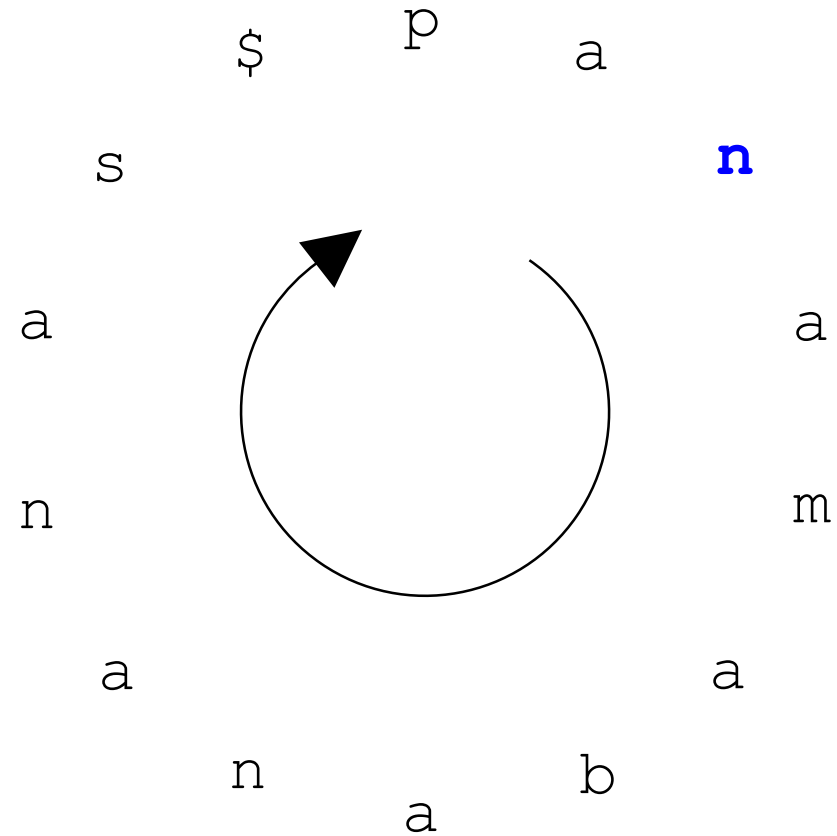
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamabana  
nanas\$panamaba  
ananas\$panamab  
bananas\$panama  
abananas\$panam  
mabananas\$pana  
**amabananas\$pan**



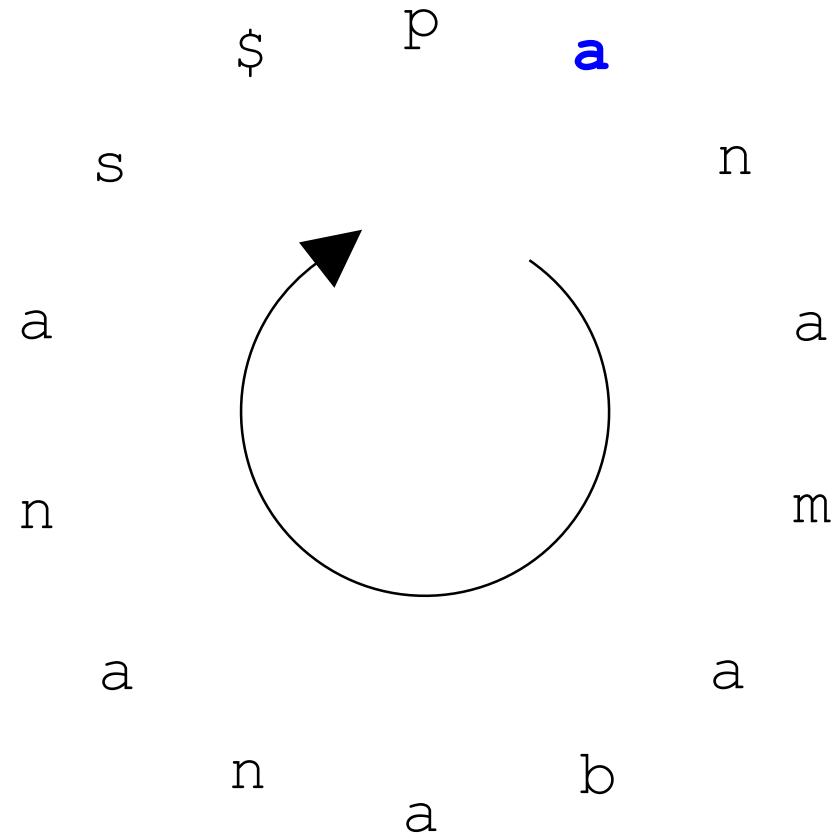
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamaban  
nanas\$panamaba  
ananas\$panamab  
bananas\$panama  
abananas\$panam  
mabananas\$pana  
amabananas\$pan  
**n**amabananas\$pa



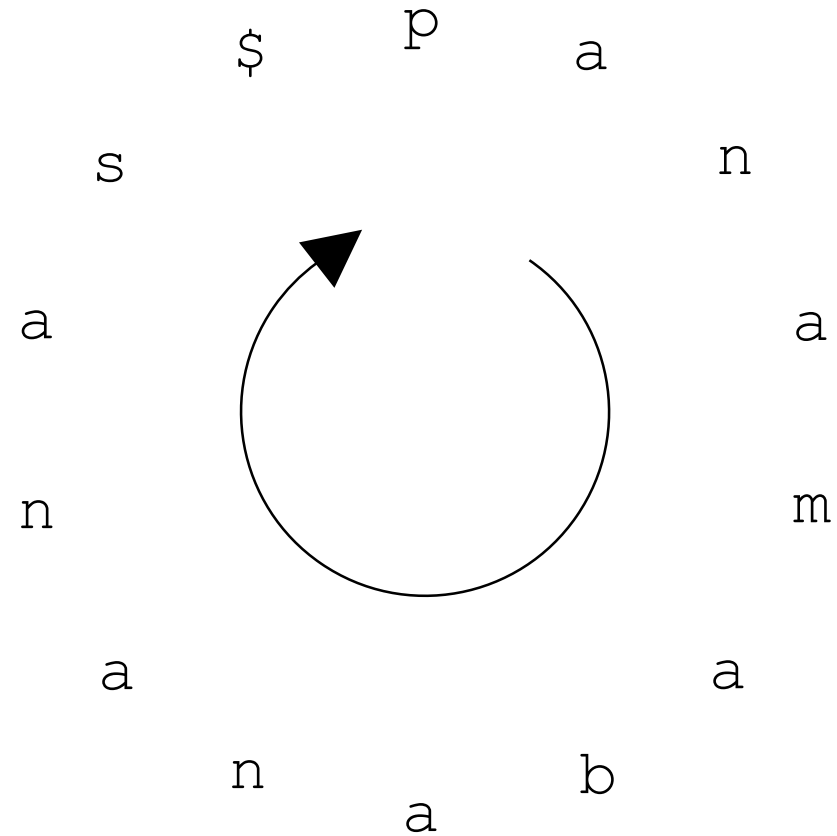
# Cyclic Rotations

panamabananas\$  
\$panamabananas  
s\$panamabanana  
as\$panamabanana  
nas\$panamabana  
anas\$panamaba  
nanas\$panamaba  
ananas\$panamab  
bananas\$panama  
abananas\$panam  
mabananas\$pana  
amabananas\$pan  
namabananas\$pa  
**a****n****a****m****a****b****a****n****a****n****a****s****\$****p**




# Cyclic Rotations

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



# Sorting Cyclic Rotations

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



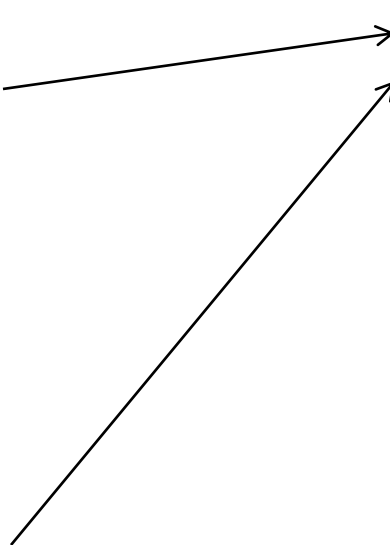
\$panamabananas

Sort the strings  
lexicographically  
(\$ comes first)

# Sorting Cyclic Rotations

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

\$panamabananas  
abananas\$panam



Sort the strings  
lexicographically  
(\$ comes first)

# Sorting Cyclic Rotations

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

\$panamabananas  
**a**bananas\$panam  
**am**abananas\$pan

Sort the strings  
lexicographically  
(\$ comes first)

# Sorting Cyclic Rotations

panamabananas\$	→	<b>\$</b> panamabananas
\$panamabananas	→	<b>a</b> bananas\$panam
s\$panamabanana	→	<b>am</b> abananas\$pan
as\$panamabanana	→	<b>anam</b> abananas\$p
nas\$panamabana		
anas\$panamaban		
anas\$panamaba		
nanas\$panamaba		
ananas\$panamab		
ananas\$panamab		
bananas\$panama		
abananas\$panam		
mabananas\$pana		
amabananas\$pan		
namabananas\$pa		
anamabananas\$p		

Sort the strings  
lexicographically  
(\$ comes first)



# Sorting Cyclic Rotations

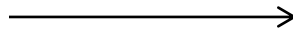
panamabananas\$  
\$panamabananas  
s\$panamabananas  
as\$panamabananas  
nas\$panamabananas  
anas\$panamabananas  
nanas\$panamabananas  
bananas\$panamabananas  
abanas\$panamabananas  
mabananas\$panamabananas  
amabananas\$panamabananas  
namabananas\$panamabananas  
anamabananas\$panamabananas

\$panamabananas  
**a**bananas\$panamabananas  
**am**abananas\$panamabananas  
**anam**abananas\$panamabananas  
**anan**as\$panamabananas

Sort the strings  
lexicographically  
(\$ comes first)

# Sorting Cyclic Rotations

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

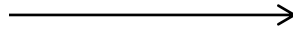


**\$**panamabananas  
**a**bananas\$panam  
**am**abananas\$pan  
**anam**abananas\$p  
**anan**as\$panamab  
**anas**\$panamaban  
**as**\$panamabanan  
**b**abananas\$panama  
**m**abananas\$pana  
**nam**abananas\$pa  
**nan**as\$panamaba  
**nas**\$panamabana  
**p**anamabananas\$  
**s**\$panamabanan

Sort the strings  
lexicographically  
(\$ comes first)

BWT(panamabananas\$)=smnbnbnaaaa\$a

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



\$panamabananas**s**  
abananas\$pana**m**  
amabananas\$pa**n**  
anamabananas\$**p**  
ananas\$panama**b**  
anas\$panamaba**n**  
as\$panamabana**n**  
bananas\$panam**a**  
mabananas\$pan**a**  
namabananas\$**p****a**  
nanas\$panamab**a**  
nas\$panamaban**a**  
panamabananas\$**s**  
s\$panamabanan**a**

All cyclic rotations of  
“panamabananas\$”

**Burrows-Wheeler Transform (BWT):**  
Last column = **smnpbnnaaaaa\$a**

BWT(p**an**a**m**a**b**a**n**a**s**\$)=s**m**n**p**b**n**n**a****a****a****a**\$**a**

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

\$panamabananas**s**  
 abananas\$pana**m**  
 amabananas\$pa**n**  
 anamabananas\$**p**  
 ananas\$panama**b**  
 anas\$panamaba**n**  
 as\$panamabana**n**  
 bananas\$panam**a**  
 mabananas\$pan**a**  
 namabananas\$**p****a**  
 nanas\$panamab**a**  
 nas\$panamaban**a**  
 panamabananas\$**s**  
 s\$panamabanan**a**

All cyclic rotations of  
“panamabananas\$”

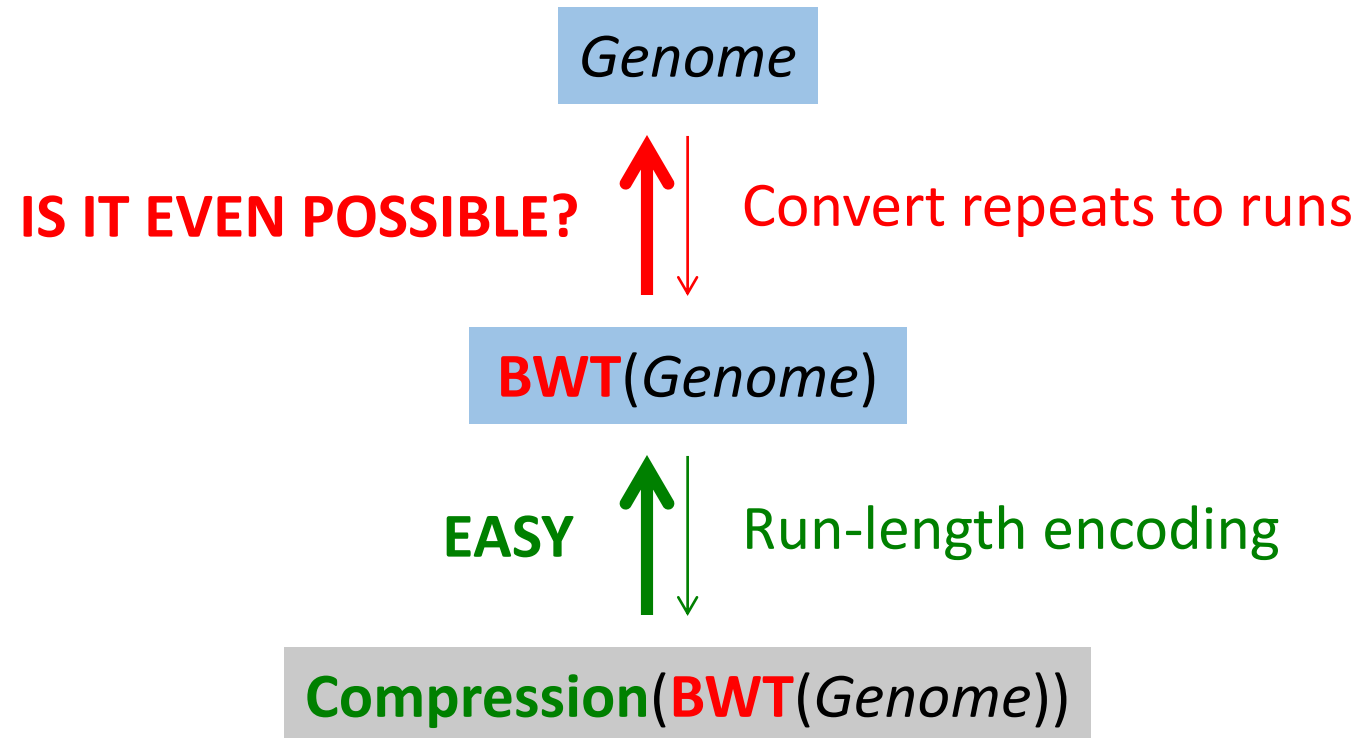
**Burrows-Wheeler Transform (BWT):**  
Last column = **smnpbnnaaaaaa\$a**

# 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

# Going Back From $\text{BWT}(\text{Genome})$ to $\text{Genome}$



# Outline

- Burrows-Wheeler Transform
- **Inverting Burrows-Wheeler Transform**
- Using BWT for Pattern Matching
- Suffix Arrays
- Approximate Pattern Matching

# Reconstructing banana from annb\$aa

\$ b a n a n **a**  
a \$ b a n a **n**  
a n a \$ b a **n**  
a n a n a \$ **b**  
b a n a n a \$  
n a \$ b a n **a**  
n a n a \$ b **a**



# Reconstructing banana

**\$** b a n a n a  
a **\$** b a n a n  
a n a **\$** b a n  
a n a n a **\$** b  
b a n a n a **\$**  
n a **\$** b a n a  
n a n a **\$** b a

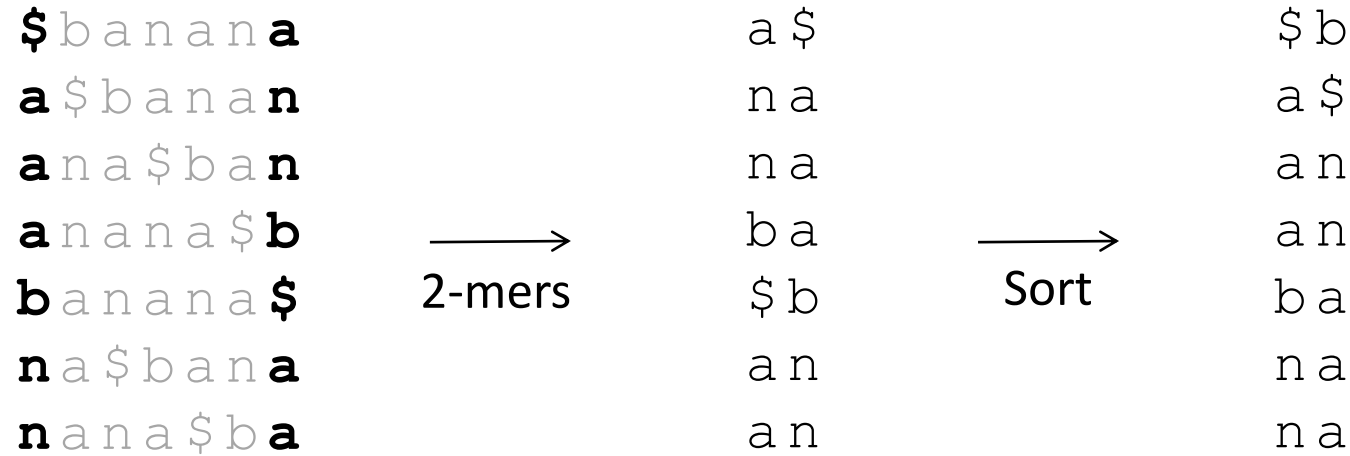
- Sorting all elements of “annb\$aa” gives first column of BWT matrix.

# Reconstructing banana

<b>\$</b> b a n a n a <b>a</b>		a \$
<b>a</b> \$ b a n a n <b>n</b>		n a
<b>a</b> n a \$ b a n <b>n</b>		n a
<b>a</b> n a n a \$ <b>b</b>	→	b a
<b>b</b> a n a n a \$	2-mers	\$ b
<b>n</b> a \$ b a n <b>a</b>		a n
<b>n</b> a n a \$ b <b>a</b>		a n

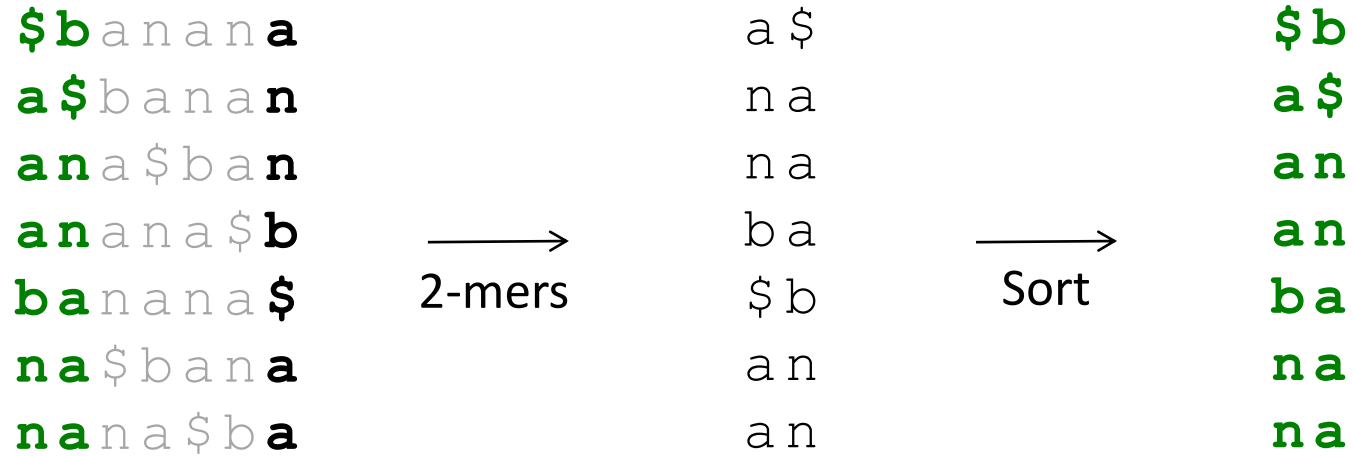
- We now know 2-mer composition of the circular string banana\$

# Reconstructing banana



- We now know 2-mer composition of the circular string banana\$
- Sorting gives us the first 2 columns of the matrix.

# Reconstructing banana



- We now know 2-mer composition of the circular string banana\$
- Sorting gives us the first 2 columns of the matrix.

# Reconstructing banana

**\$b** a n a n a

a **\$b** a n a n

a n a **\$b** a n

a n a n a **\$b**

b a n a n a **\$**

n a **\$b** a n a

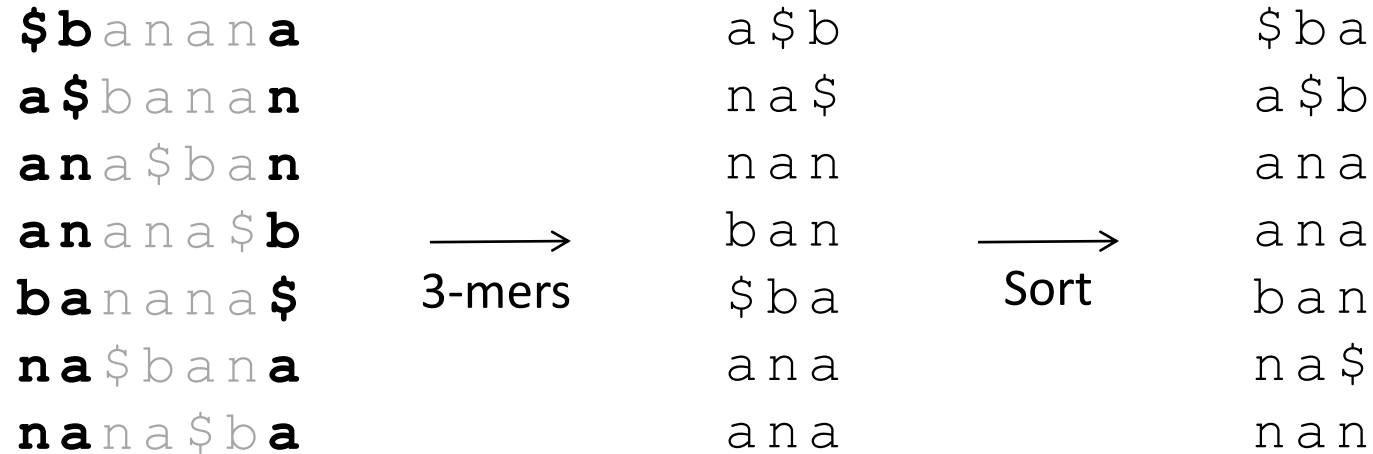
n a n a **\$b** a

# Reconstructing banana

<b>\$</b> b a n a n a		a \$ b
a <b>\$</b> b a n a n		n a \$
a n a <b>\$</b> b a n		n a n
a n a n a <b>\$</b> b	→	b a n
b a n a n a <b>\$</b>	3-mers	\$ b a
n a <b>\$</b> b a n a		a n a
n a n a <b>\$</b> b a		a n a

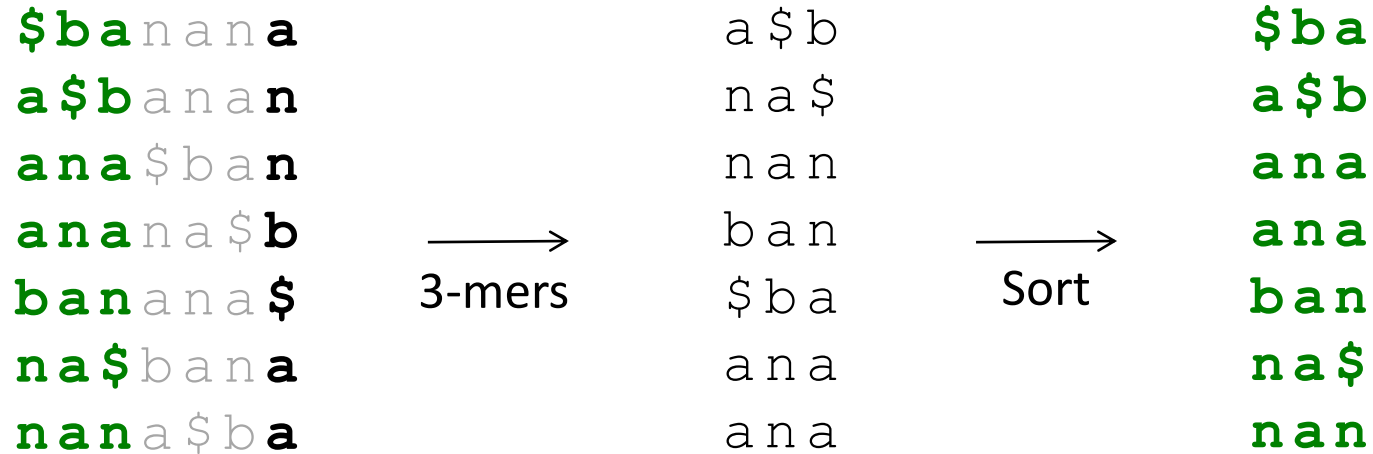
- We now know 3-mer composition of the circular string banana\$

# Reconstructing banana



- We now know 3-mer composition of the circular string banana\$
- Sorting gives us the first 3 columns of the matrix.

# Reconstructing banana



- We now know 3-mer composition of the circular string banana\$
- Sorting gives us the first 3 columns of the matrix.



# Reconstructing banana

**\$b**anana

a**\$b**anan

ana**\$b**an

ana**\$b**ana

ban**\$b**ana

na**\$b**ana

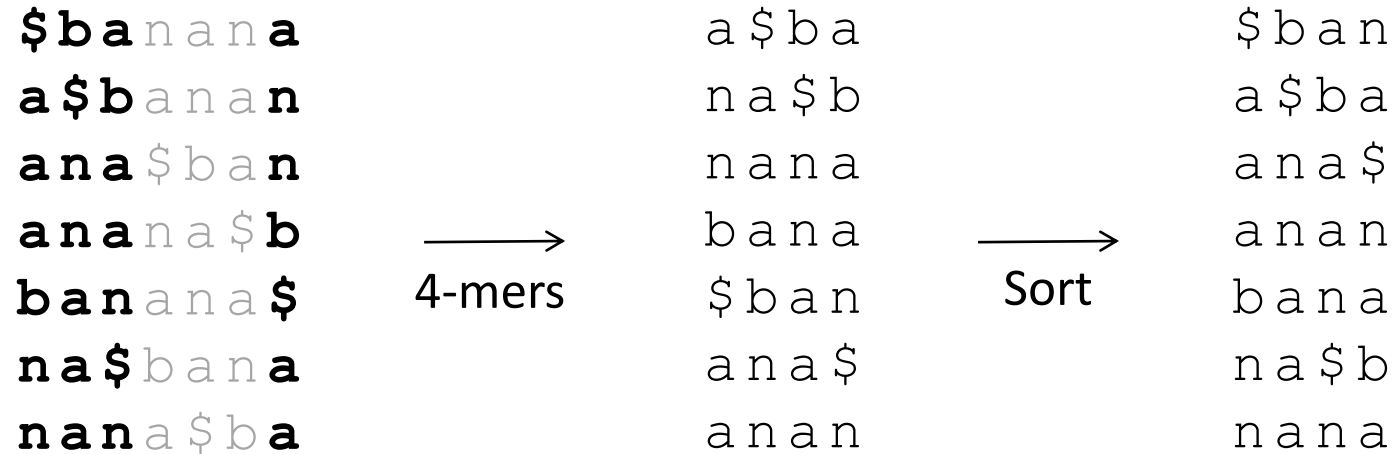
nan**\$b**ana

# Reconstructing banana

<b>\$b</b> anana		a\$ba
<b>a\$b</b> anan		na\$b
<b>ana</b> \$ban		nana
<b>ana</b> na\$b	→	bana
<b>ban</b> ana\$	4-mers	\$ban
<b>na</b> \$bana		ana\$
<b>nan</b> a\$b		anan

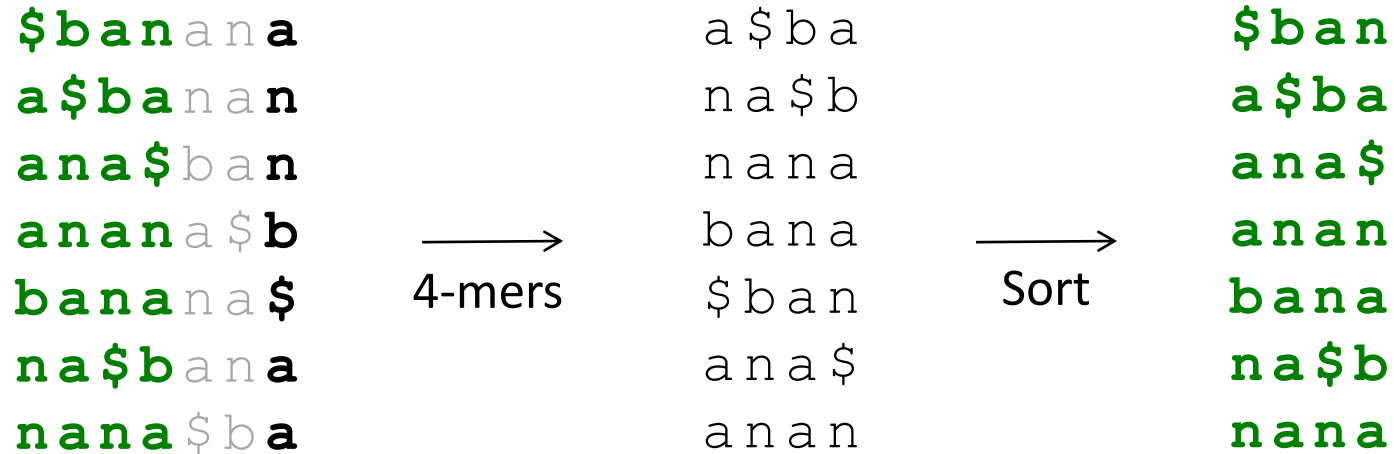
- We now know 4-mer composition of the circular string banana\$

# Reconstructing banana



- We now know 4-mer composition of the circular string banana\$
- Sorting gives us the first 4 columns of the matrix.

# Reconstructing banana



- We now know 4-mer composition of the circular string banana\$
- Sorting gives us the first 4 columns of the matrix.

# Reconstructing banana

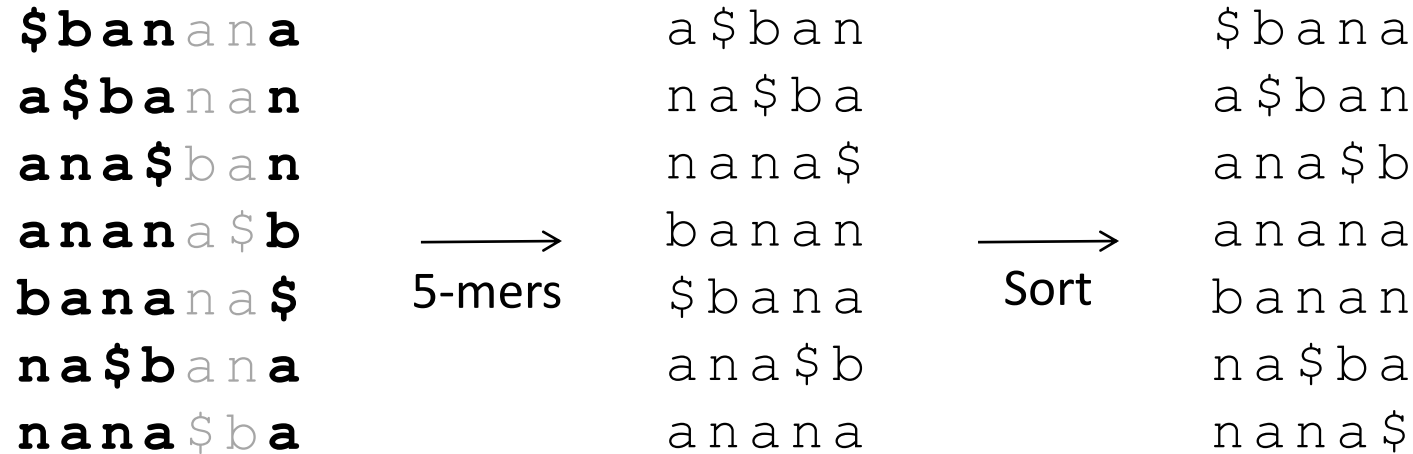
\$banaa  
a\$banan  
ana\$ban  
anana\$b  
banana\$  
na\$bana  
nana\$ba

# Reconstructing banana

<b>\$</b> banan <b>a</b>		a\$ban
<b>a</b> \$banan <b>n</b>		na\$b
<b>ana</b> \$ban		nana\$
<b>anan</b> a\$b	→	banan
<b>ban</b> ana\$	5-mers	\$bana
<b>na</b> \$bana		ana\$b
<b>nana</b> \$ba		anana

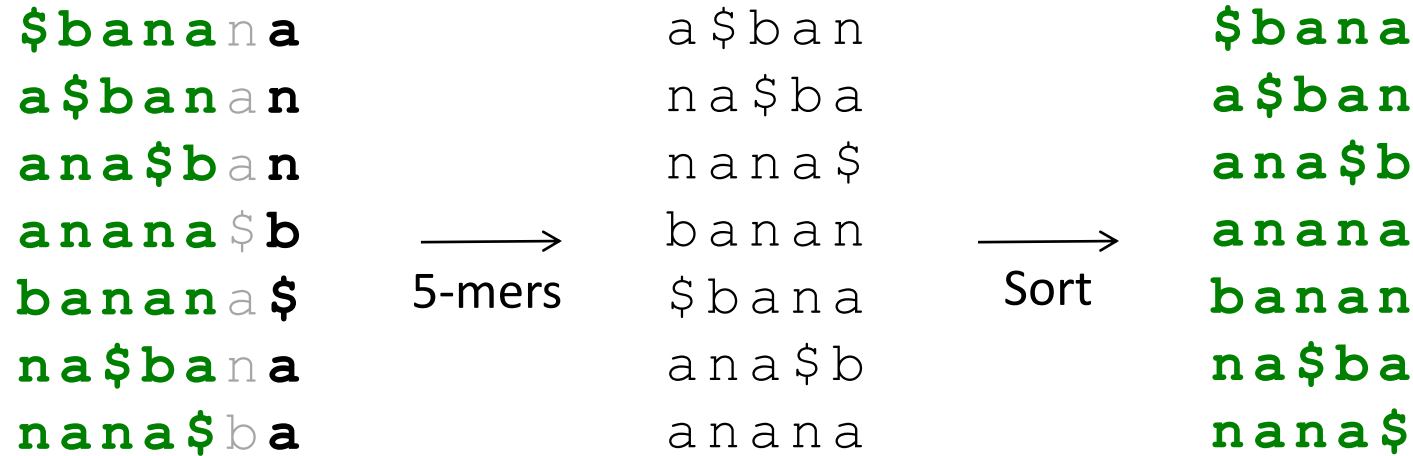
- We now know 5-mer composition of the circular string banana\$

# Reconstructing banana



- We now know 5-mer composition of the circular string banana\$
- Sorting gives us the first 5 columns of the matrix.

# Reconstructing banana



- We now know 5-mer composition of the circular string `banana$`
- Sorting gives us the first 5 columns of the matrix.



# Reconstructing banana

\$bana n a  
a \$ban a n  
ana \$b a n  
anana \$b  
banan a \$  
na \$ba n a  
nana \$b a

# Reconstructing banana

\$bana	n	a	
a\$bana	n		
ana\$b	a	n	
anana\$b			
banana	\$		
na\$b	a	n	a
nana\$b	a		

→ 6-mers

a\$bana
na\$bana
nana\$b
banana
\$banan
ana\$ba
anana\$

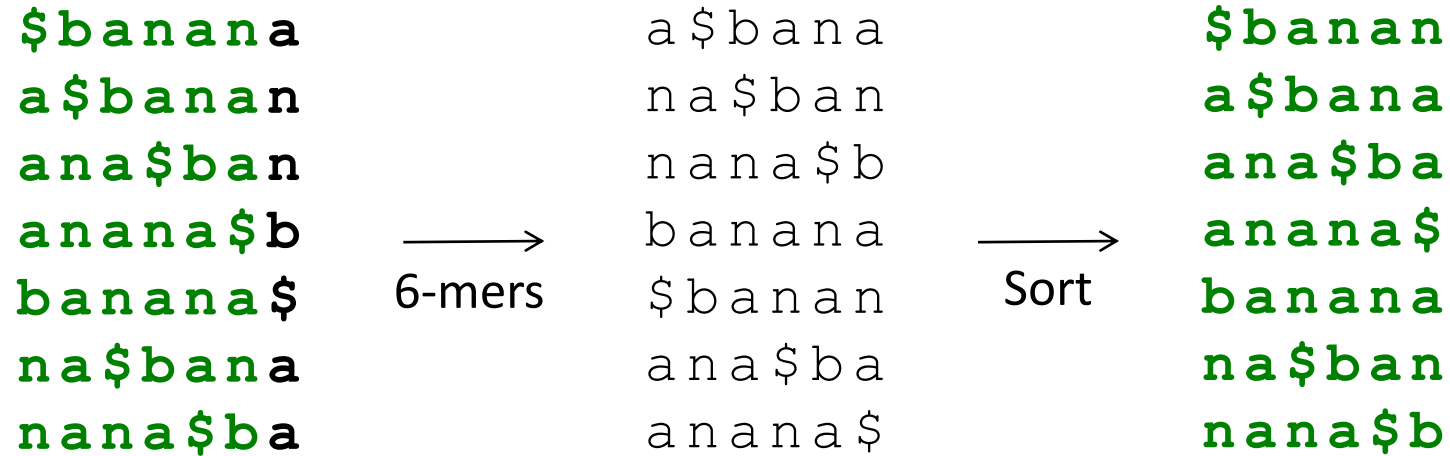
- We now know 6-mer composition of the circular string banana\$

# Reconstructing banana

\$bana <b>n</b> a		a\$bana		\$banan
a\$bana <b>n</b>		na\$bana		a\$bana
ana\$b <b>a</b> n		nana\$b		ana\$b
anana <b>\$</b> b	→	banana	→	anana\$
banan <b>a</b> \$	6-mers	\$banan	Sort	banana
na\$b <b>a</b> n		ana\$ba		na\$ban
nana <b>\$</b> ba		anana\$		nana\$b

- We now know 6-mer composition of the circular string banana\$
- Sorting gives us the first 6 columns of the matrix.

# Reconstructing banana



- We now know 6-mer composition of the circular string banana\$
- Sorting gives us the first 6 columns of the matrix.

# Reconstructing banana

\$banana  
a\$banan  
ana\$ban  
anana\$b  
banana\$  
na\$bana  
nana\$ba

- We now know the entire matrix!

# Reconstructing banana

**\$banana**

a \$ b a n a n

a n a \$ b a n

a n a n a \$ b

b a n a n a \$

n a \$ b a n a

n a n a \$ b a

- We now know the entire matrix!
- Symbols in the first row (after \$) spell **banana**.

# More Memory Issues

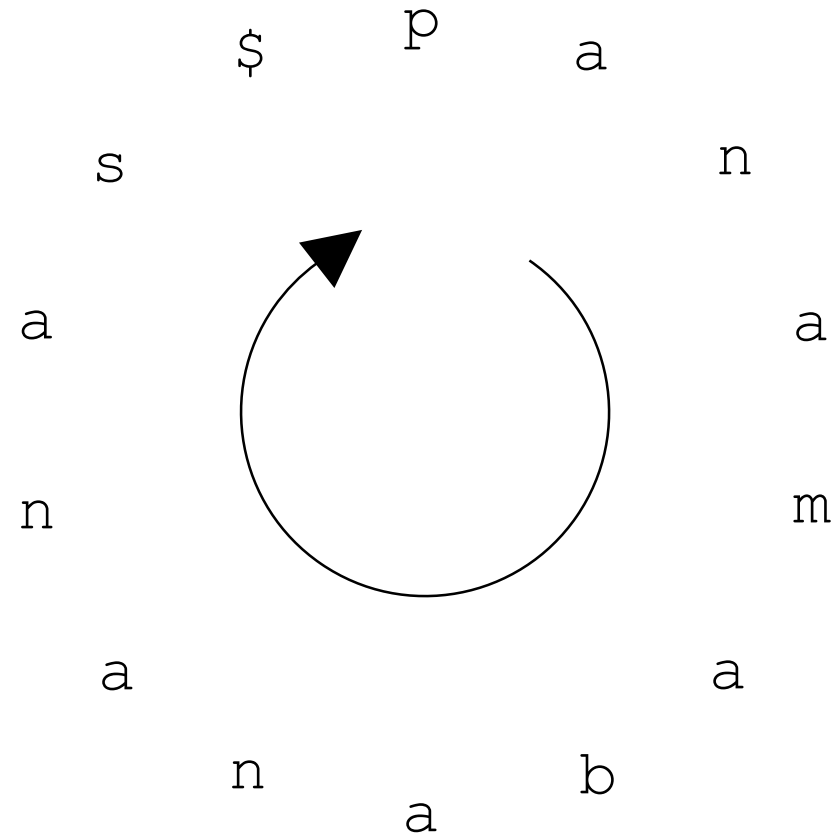
- Reconstructing *Text* from  $BWT(Text)$  required us to store  $|Text|$  cyclic rotations of  $|Text|$ .

```
$b a n a n a  
a $b a n a n  
a n a $b a n  
a n a n a $b  
b a n a n a $  
n a $b a n a  
n a n a $b a
```

- Can we invert  $BWT(Text)$  with less space and without  $|Text|$  rounds of sorting?

# A Strange Observation

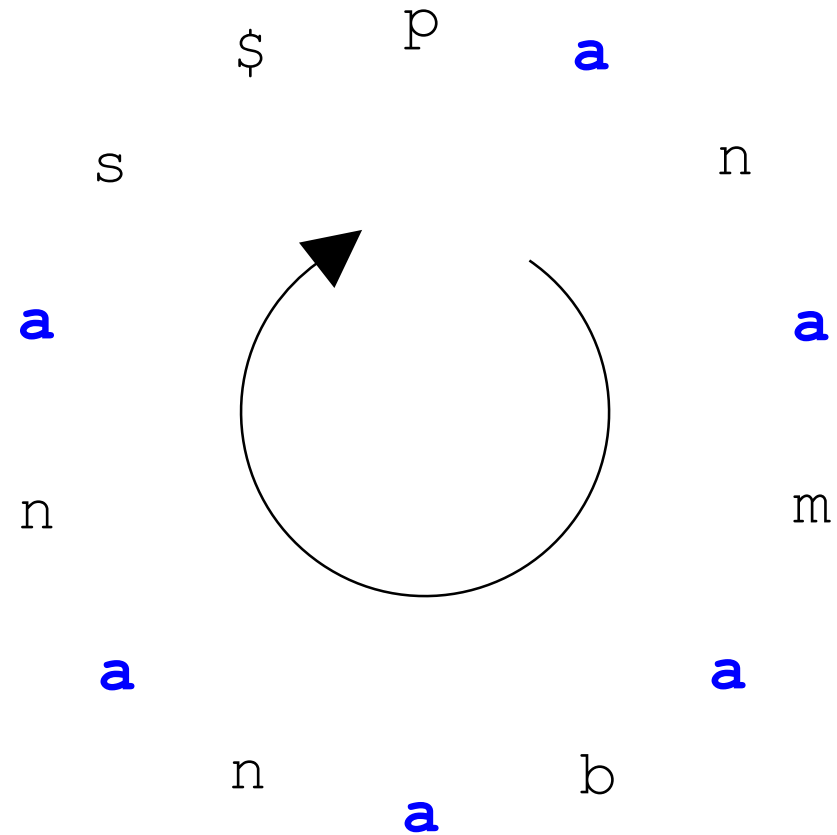
\$panamabananas  
abananas\$panam  
amabananas\$pan  
anamabananas\$p  
ananas\$panamab  
anas\$panamaban  
as\$panamabanan  
bananas\$panama  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabanana





# A Strange Observation

\$panamabananas  
**a**bananas\$panam  
**a**mabananas\$pan  
**a**namabananas\$p  
**a**nanas\$panamab  
**a**nas\$panamaban  
**a**s\$panamabanan  
bananas\$panam**a**  
mabananas\$pan**a**  
namabananas\$p**a**  
nanas\$panamab**a**  
nas\$panamaban**a**  
panamabananas\$  
s\$panamabanan**a**



# A Strange Observation

Where

is first

“a”

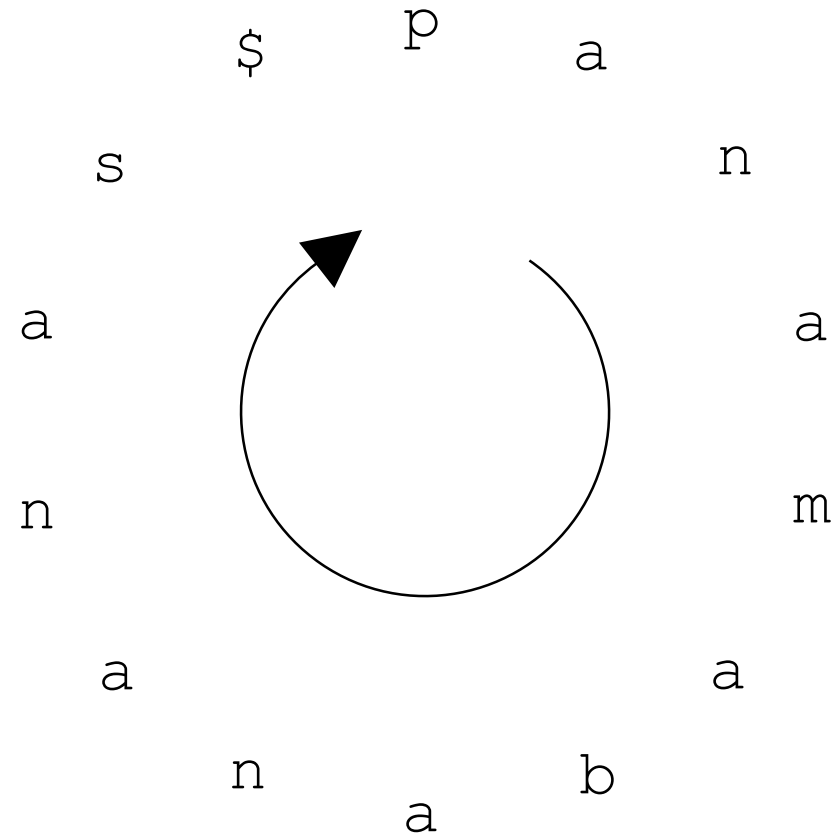
hiding

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



# 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\$panam(a  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabannan

Where

is first

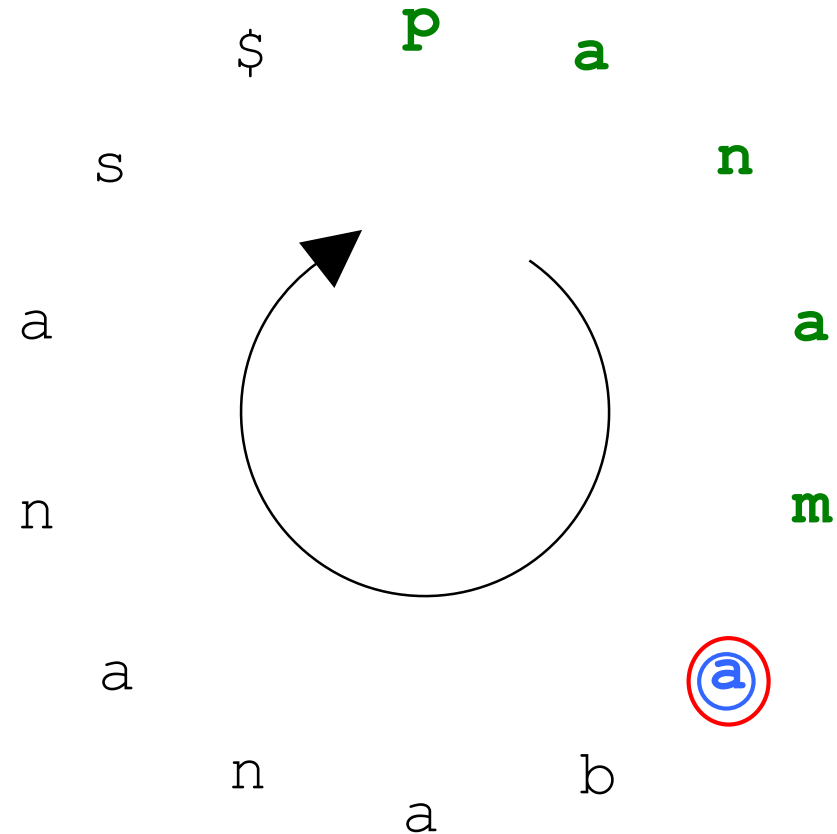
"a"

hiding

inside

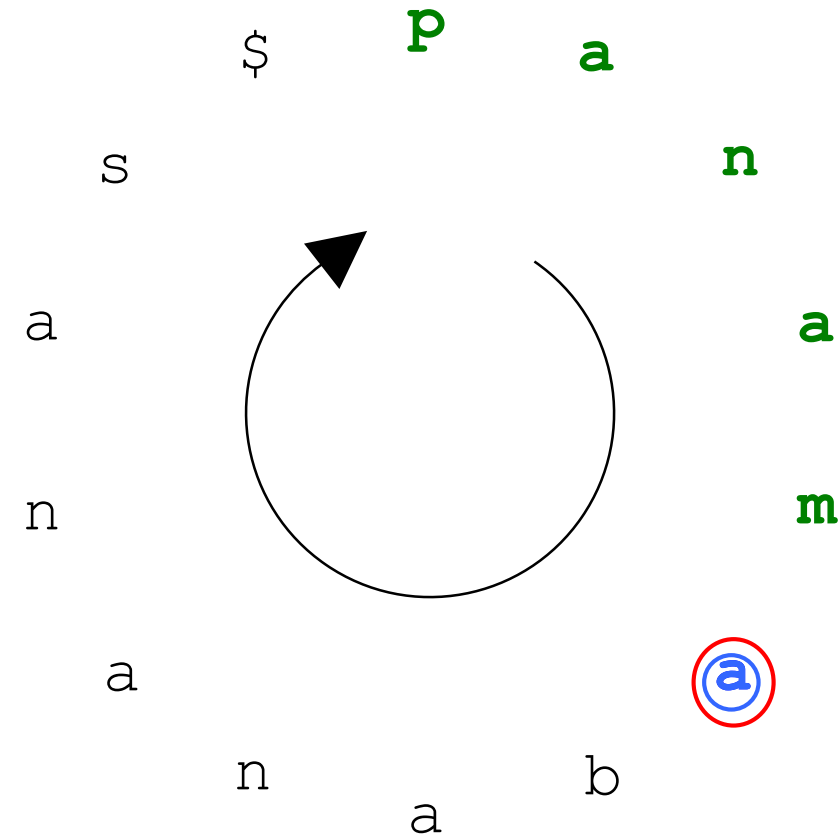
the

circle?



# They Are Hiding at the Same Position!

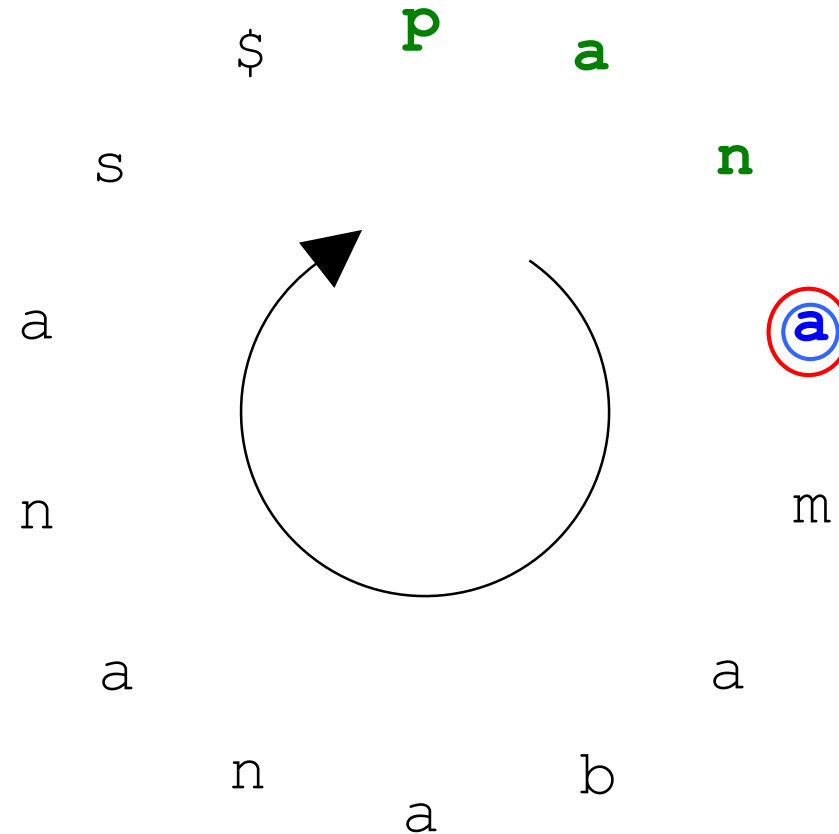
\$panamabananas  
a bananas \$panam  
amabananas \$pan  
anamabananas \$p  
ananas \$panamab  
anas \$panamaban  
as \$panamaban  
bananas \$panam  
mabananas \$pana  
namabananas \$pa  
nanas \$panamaba  
nas \$panamabana  
panamabananas \$  
s \$panamabana



1<sup>st</sup> a in *FirstColumn* and 1<sup>st</sup> a in *LastColumn*  
are hiding at the same position along the cycle!

# Another Strange Observation

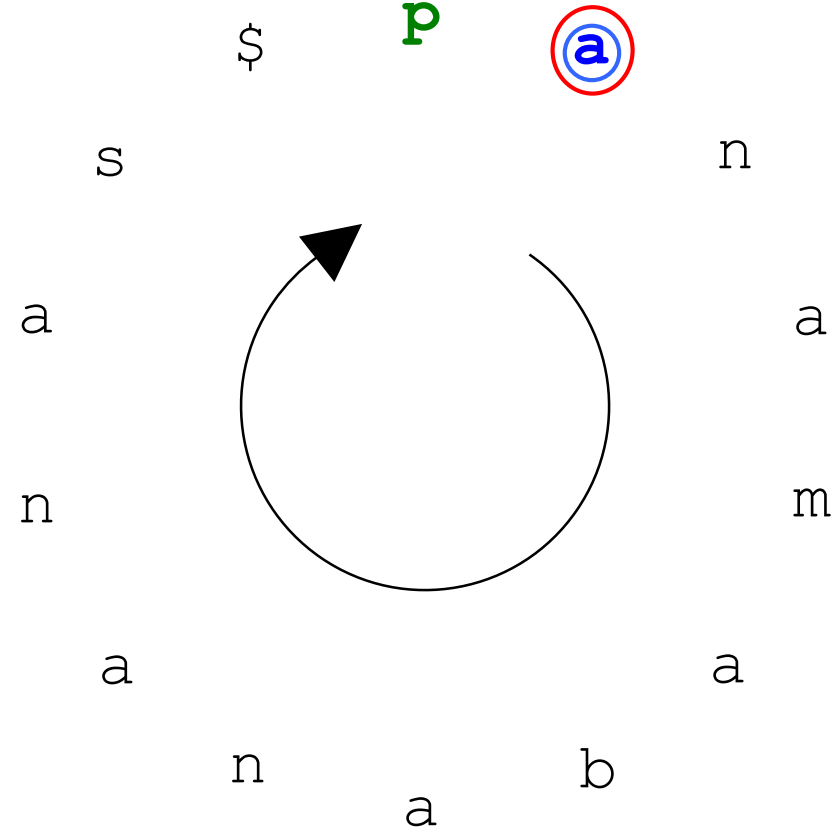
\$panamabananas  
abananas\$panam  
**a**mabananas\$**pan**  
anamabananas\$p  
ananas\$panamab  
anas\$panamaban  
as\$panamabanan  
bananas\$panama  
mabananas\$**pana****a**  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabana



2<sup>nd</sup> **a** in *FirstColumn* and 2<sup>nd</sup> **a** in *LastColumn*  
are hiding at the same position along the cycle!

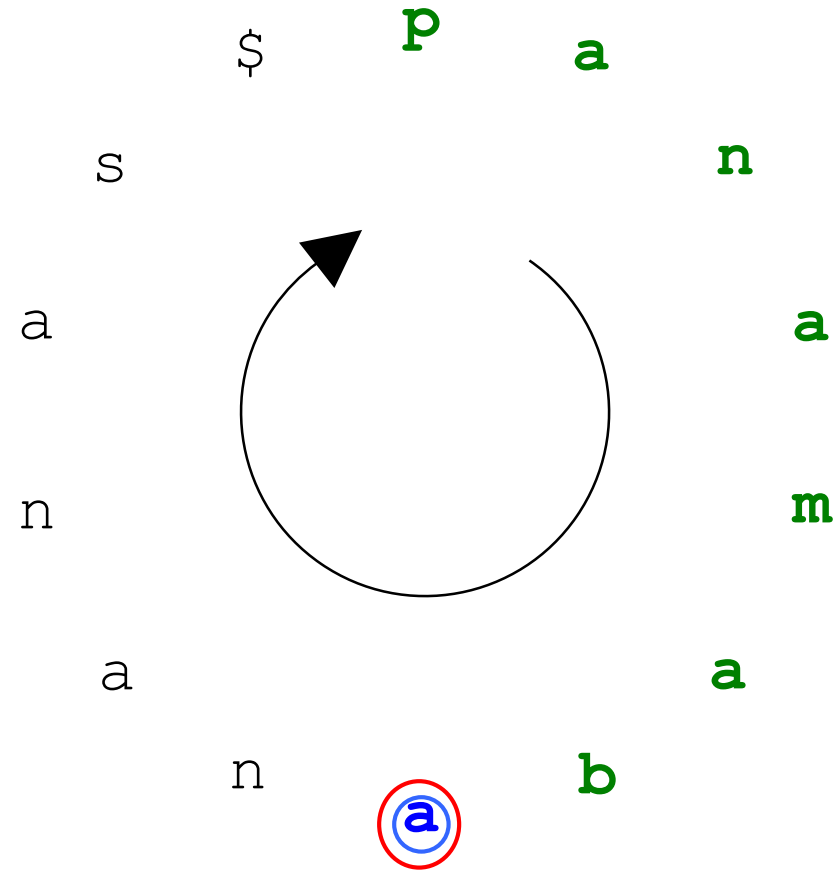
# Another Strange Observation

\$panamabananas  
abananas\$panam  
amabananas\$pan  
**a**namabananas\$**p**  
ananas\$panamab  
anas\$panamaban  
as\$panamabanan  
bananas\$panama  
mabananas\$pana  
namabananas\$**p****a**  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabana



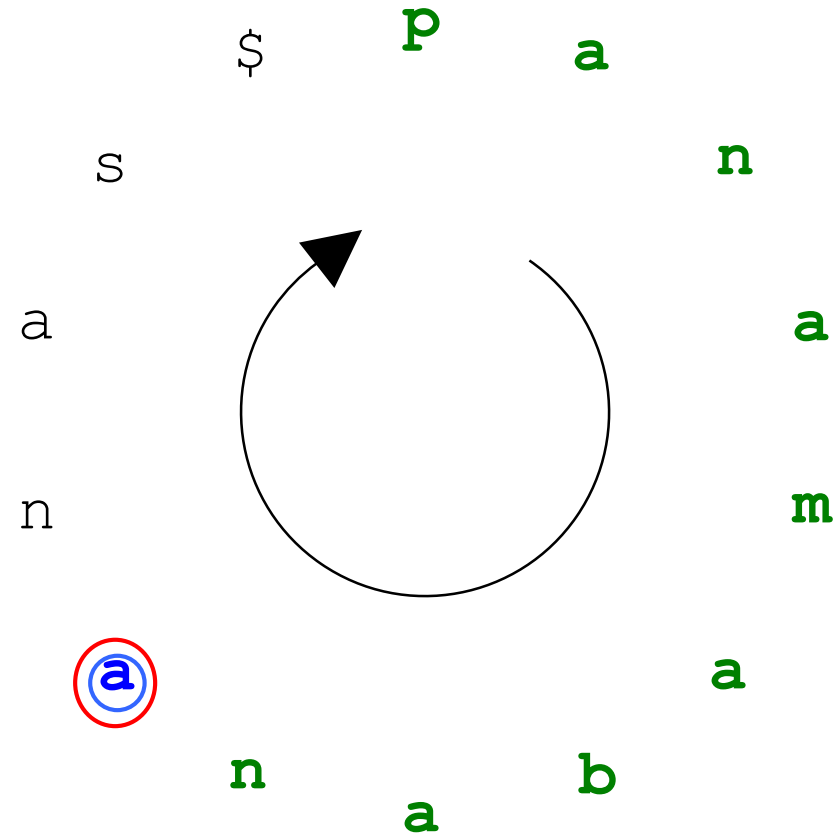
# Another Strange Observation

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



# Another Strange Observation

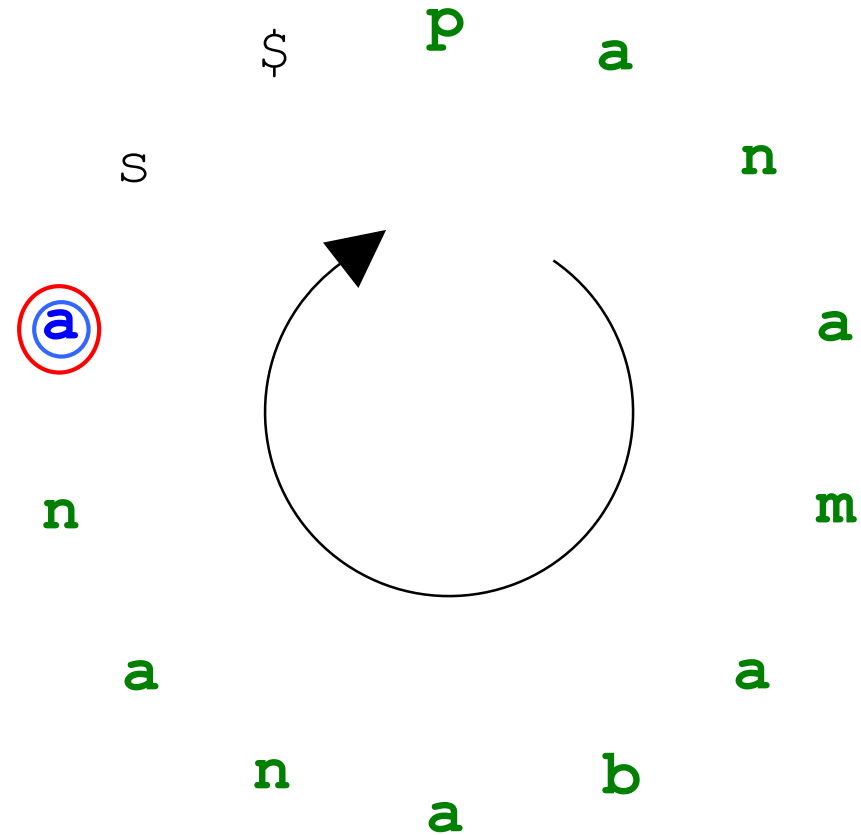
\$panamabananas  
abananas\$panam  
amabananas\$pan  
anamabananas\$p  
ananas\$panamab  
**anas\$panamaban**  
as\$panamabanan  
bananas\$panama  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$**panamabana**  
panamabananas\$  
s\$panamabana





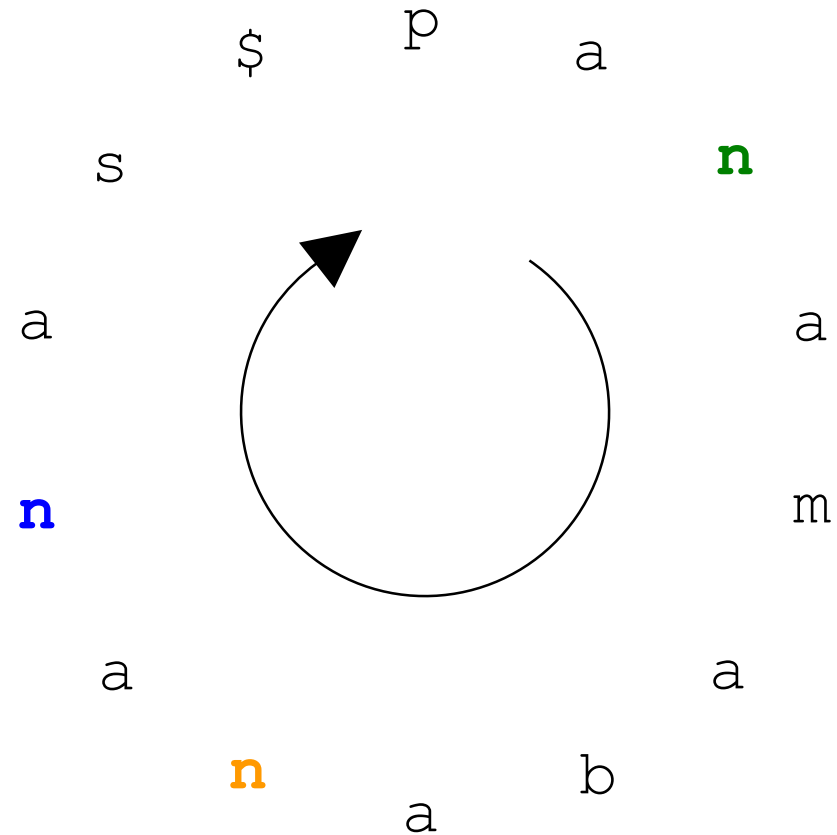
# Another Strange Observation

\$panamabananas  
abananas\$panam  
amabananas\$pan  
anamabananas\$p  
ananas\$panamab  
anas\$panamaban  
**a**s\$**p****a****n****a****m****a****b****a****n****a****n**  
bananas\$panama  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$**p****a****n****a****m****a****b****a****n****a****n****a**



# Another Strange Observation

\$panamabananas  
abananas\$panam  
amabananas\$pa**n**  
anamabananas\$p  
ananas\$panamab  
anas\$panamaba**n**  
as\$panamabana**n**  
bananas\$panama  
mabananas\$pana  
**n**amabananas\$pa  
**n**anas\$panamaba  
**n**as\$panamabana  
panamabananas\$  
s\$panamabanana



# Is It True in General?

```
$panamabananas  
1 abananas$panam  
2 amabananas$pan  
3 anabananas$p  
4 ananas$panamab  
5 anas$panamaban  
6 as$panamabanan  
bananas$panama  
mabananas$pana  
namabananas$pa  
nanas$panamaba  
nas$panamabana  
panamabananas$  
s$panamabanna
```

These strings are sorted

# Is It True in General?

\$panamabananas  
1 a bananas\$panam  
2 amabananas\$pan  
3 anabananas\$p  
4 ananas\$panamab  
5 anas\$panamaban  
6 as\$panamabanan  
bananas\$panama  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabannan

Chop off a

bananas\$panam  
mabananas\$pan  
namabananas\$p  
nanas\$panamab  
nas\$panamaban  
s\$panamabanan

These strings are sorted

# Is It True in General?

\$panamabananas  
1 a bananas\$panam  
2 amabananas\$pan  
3 anabananas\$p  
4 ananas\$panamab  
5 as\$panamaban  
6 as\$panamabanan  
bananas\$panama  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabannan

Chop off a

bananas\$panam  
mabananas\$pan  
namabananas\$p  
nanas\$panamab  
nas\$panamaban  
s\$panamabanan

Still  
sorted

These strings are sorted

# Is It True in General?

\$panamabananas  
1 **a**bananas\$panam  
2 **a**mabananas\$pan  
3 **a**namabananas\$p  
4 **a**nanas\$panamab  
5 **a**nas\$panamaban  
6 **a**s\$panamabanan  
bananas\$panama  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabanan

These strings are sorted

Chop off **a**

bananas\$panam  
mabananas\$pan  
namabananas\$p  
nanas\$panamab  
nas\$panamaban  
s\$panamabanan

Still  
sorted

Add **a**  
to end

bananas\$panam**a**  
mabananas\$pan**a**  
namabananas\$p**a**  
nanas\$panamab**a**  
nas\$panamaban**a**  
s\$panamabanan**a**

# Is It True in General?

\$panamabananas  
1 **a**bananas\$panam  
2 **a**mabananas\$pan  
3 **a**namabananas\$p  
4 **a**nanas\$panamab  
5 **a**nas\$panamaban  
6 **a**s\$panamabanan  
bananas\$panama  
mabananas\$pana  
namabananas\$pa  
nanas\$panamaba  
nas\$panamabana  
panamabananas\$  
s\$panamabannan

These strings are sorted

Chop off **a**

bananas\$panam  
mabananas\$pan  
namabananas\$p  
nanas\$panamab  
nas\$panamaban  
s\$panamabanan

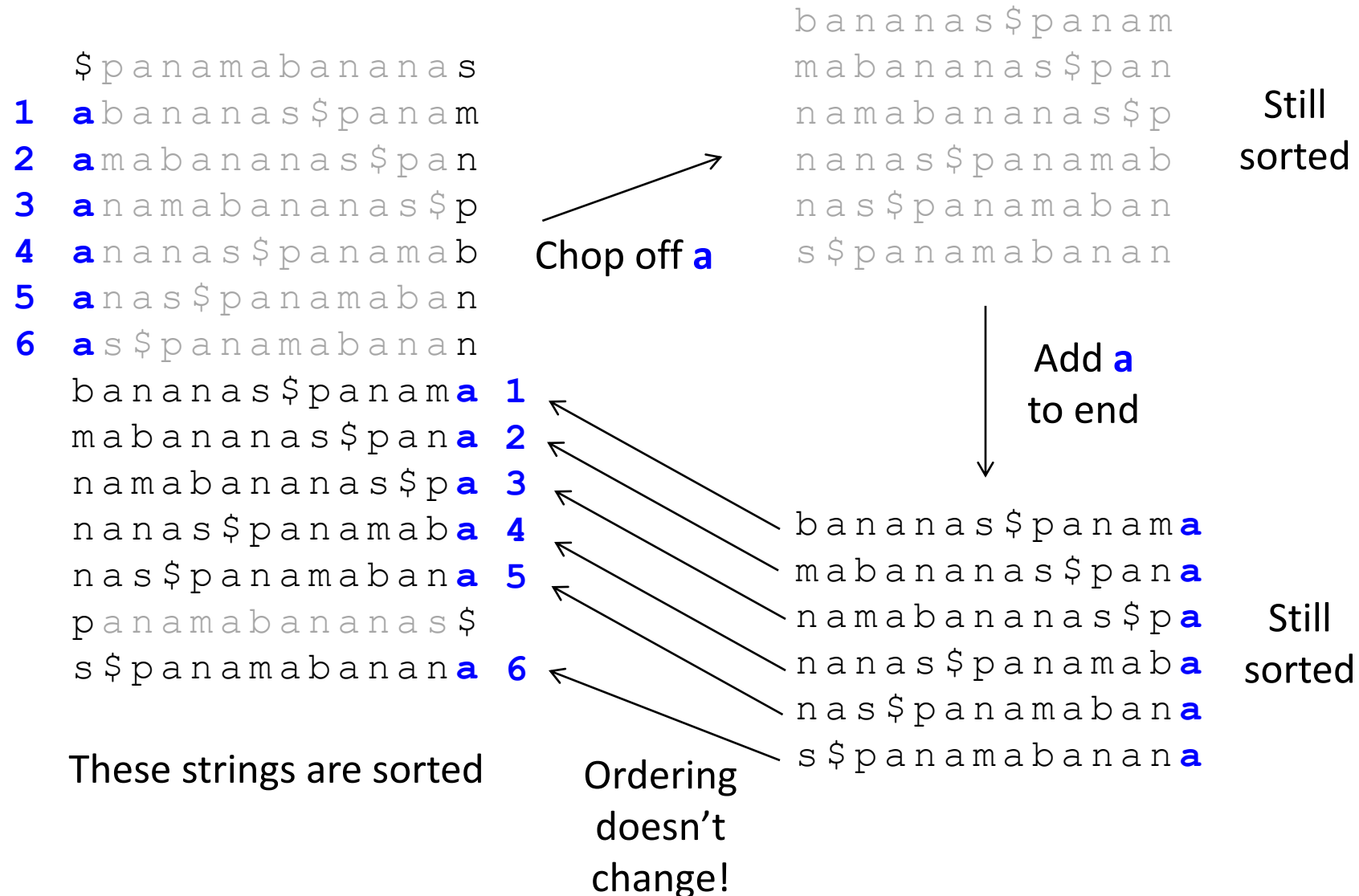
Still  
sorted

Add **a**  
to end

bananas\$panam**a**  
mabananas\$pan**a**  
namabananas\$p**a**  
nanas\$panamab**a**  
nas\$panamaban**a**  
s\$panamabanan**a**

Still  
sorted

# Is It True in General?





# First-Last Property

- the  $k$ -th occurrence of *symbol* in ***FirstColumn***
- and the  $k$ -th occurrence of *symbol* in ***LastColumn***
- correspond to appearance of *symbol* at the same position in *Text*.

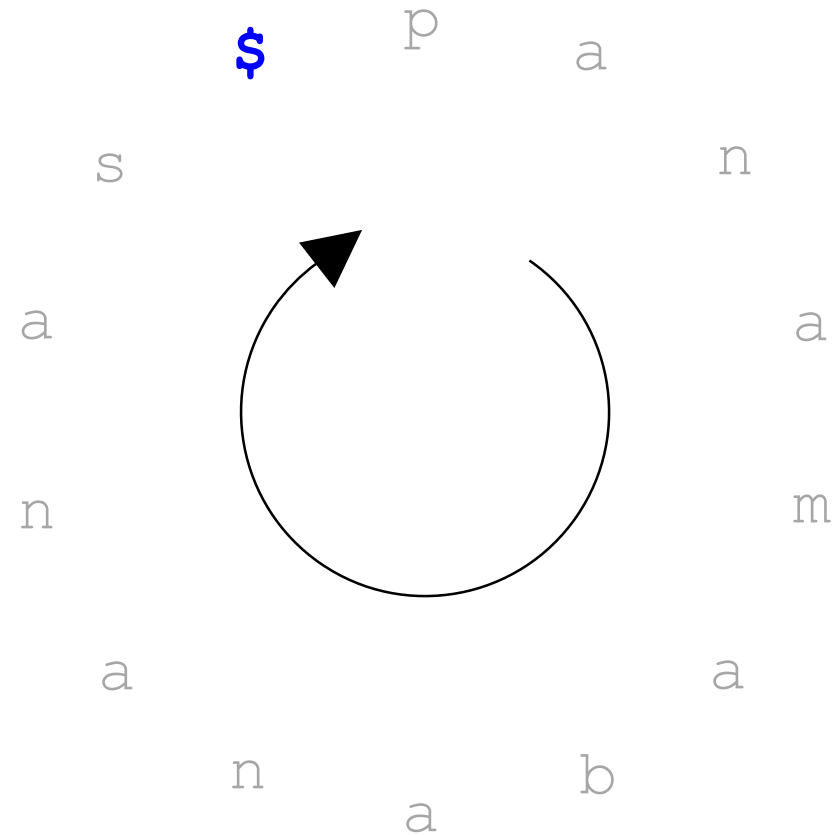
•

$p_1 a_3 n_1 a_2 m_1 a_1 b_1 a_4 n_2 a_5 n_3 a_6 s_1 \$_1$

$\$_1$ panamabanana $s_1$   
 $a_1$ bananas\$panam $m_1$   
 $a_2$ mabananas\$pan $n_1$   
 $a_3$ namabananas\$ $p_1$   
 $a_4$ nanas\$panamab $b_1$   
 $a_5$ nas\$panamaban $n_2$   
 $a_6$ s\$panamabanan $n_3$   
 $b_1$ ananas\$panama $a_1$   
 $m_1$ abananas\$pana $a_2$   
 $n_1$ amabananas\$pa $a_3$   
 $n_2$ anas\$panamaba $a_4$   
 $n_3$ as\$panamabana $a_5$   
 $p_1$ anamabananas $\$_1$   
 $s_1$ \$panamabanana $a_6$

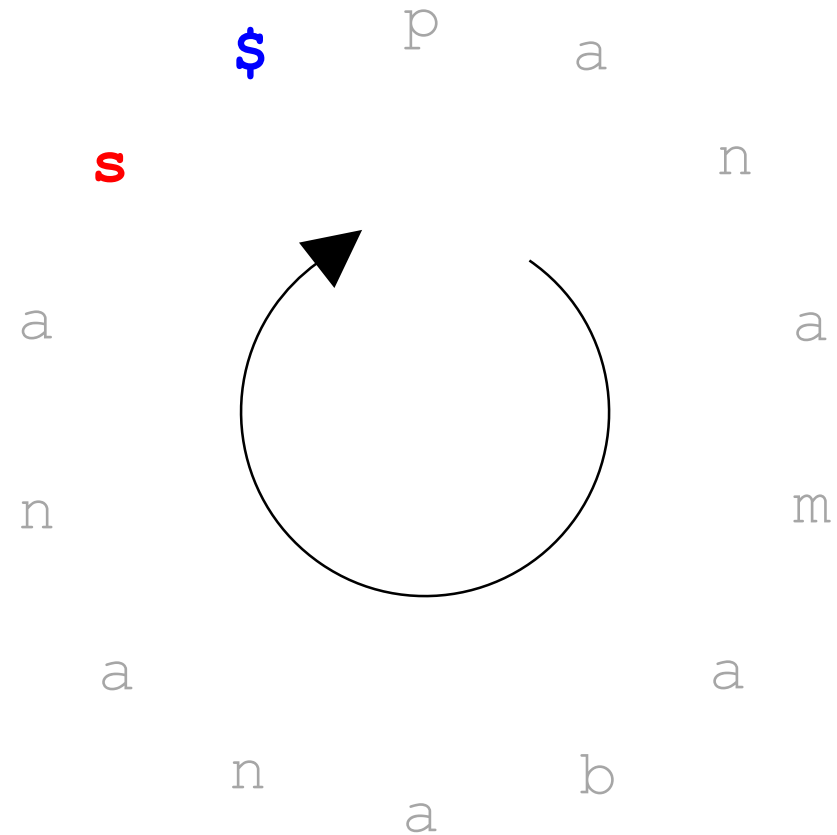
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



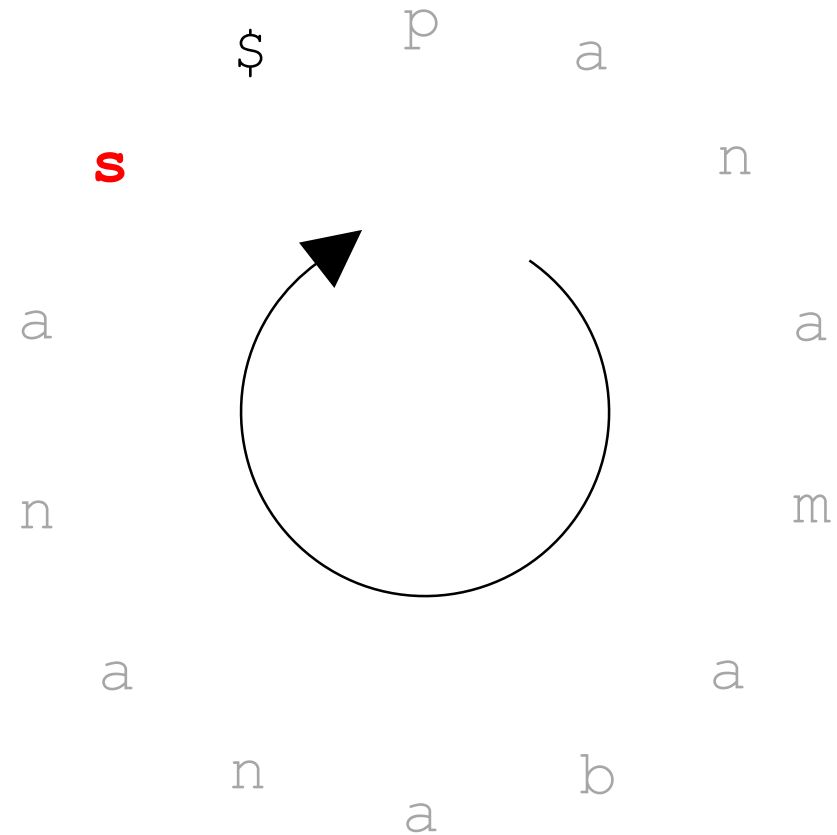
# Inverting BWT Again

**\$**<sub>1</sub>panamabanana**s**<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



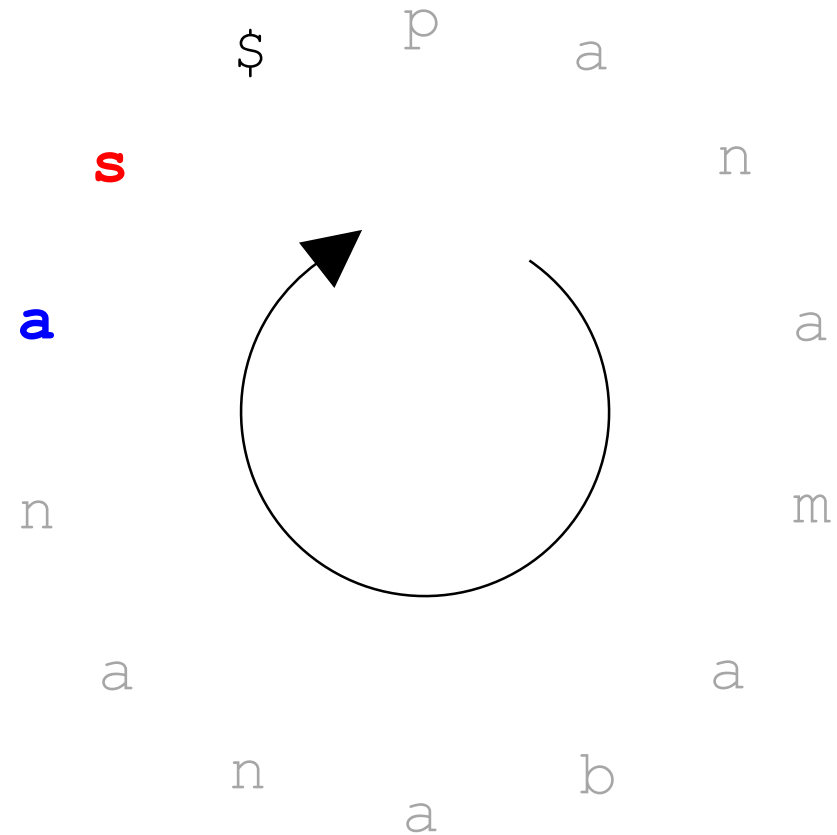
# Inverting BWT Again

\$<sub>1</sub>panamabanana **s**<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamabanana<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
**s**<sub>1</sub>\$panamabanana<sub>6</sub>



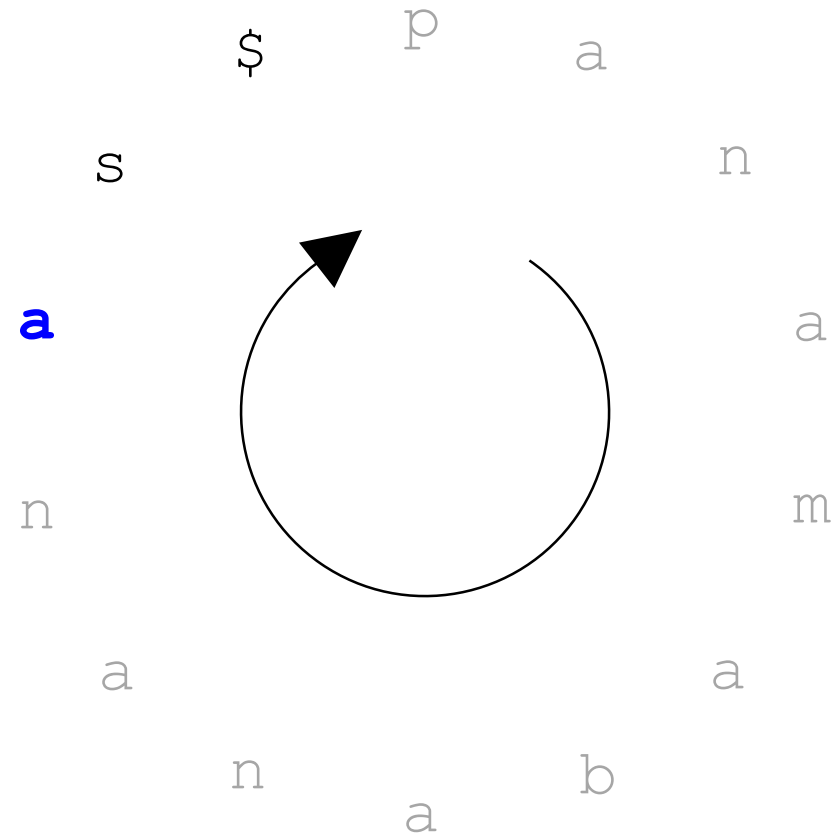
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
**s**<sub>1</sub>\$panamabanana**a**<sub>6</sub>



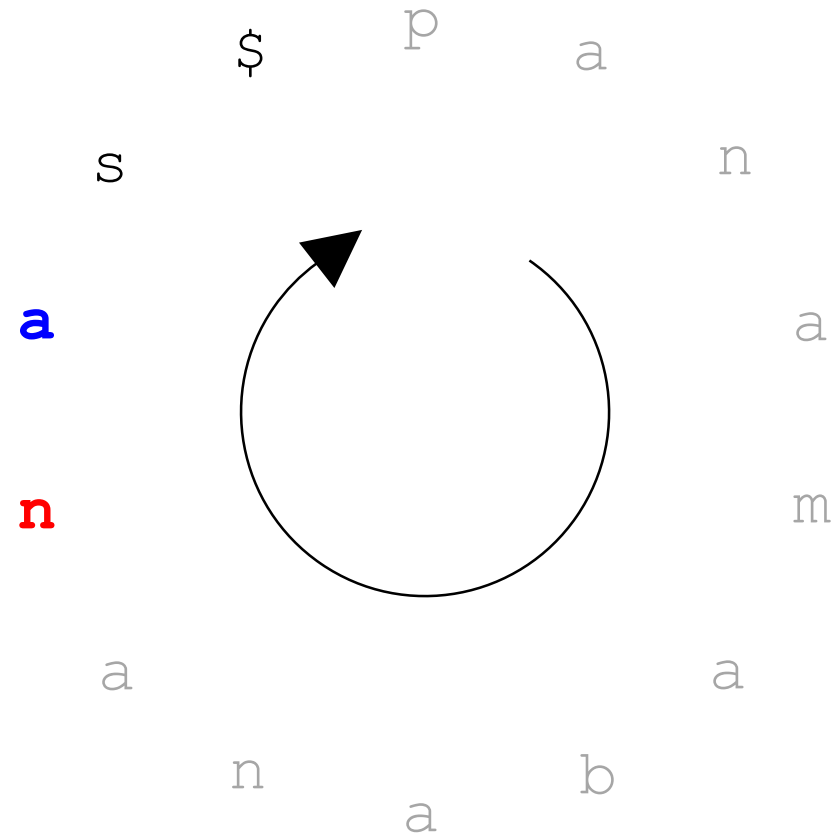
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
**a<sub>6</sub>**s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamaban**a<sub>6</sub>**



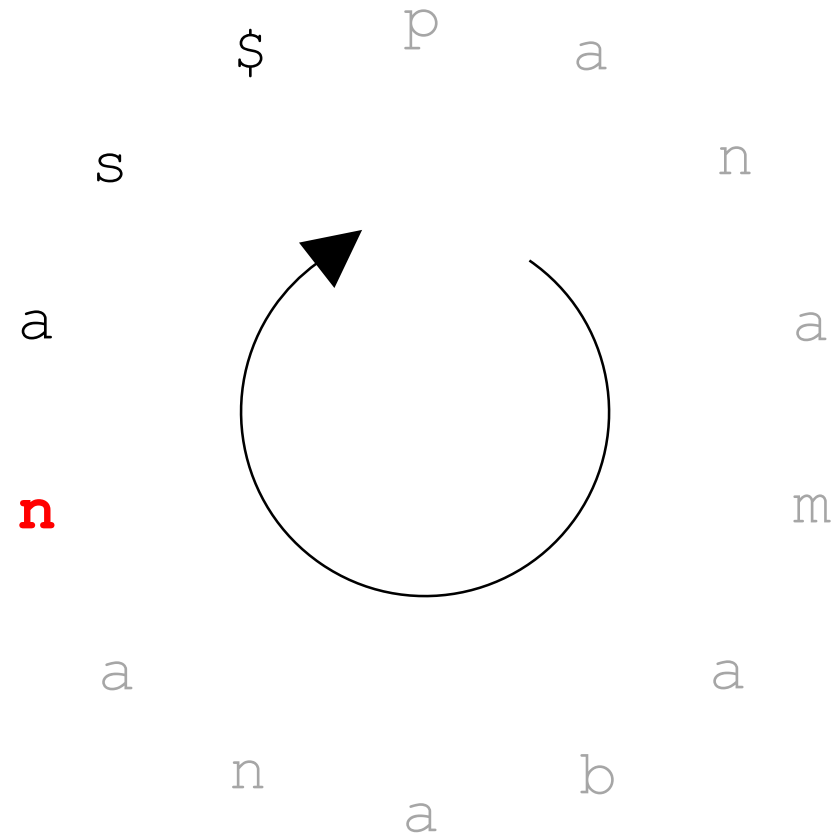
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
**a**<sub>6</sub>s\$panamabana**n**<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanaa<sub>6</sub>



# Inverting BWT Again

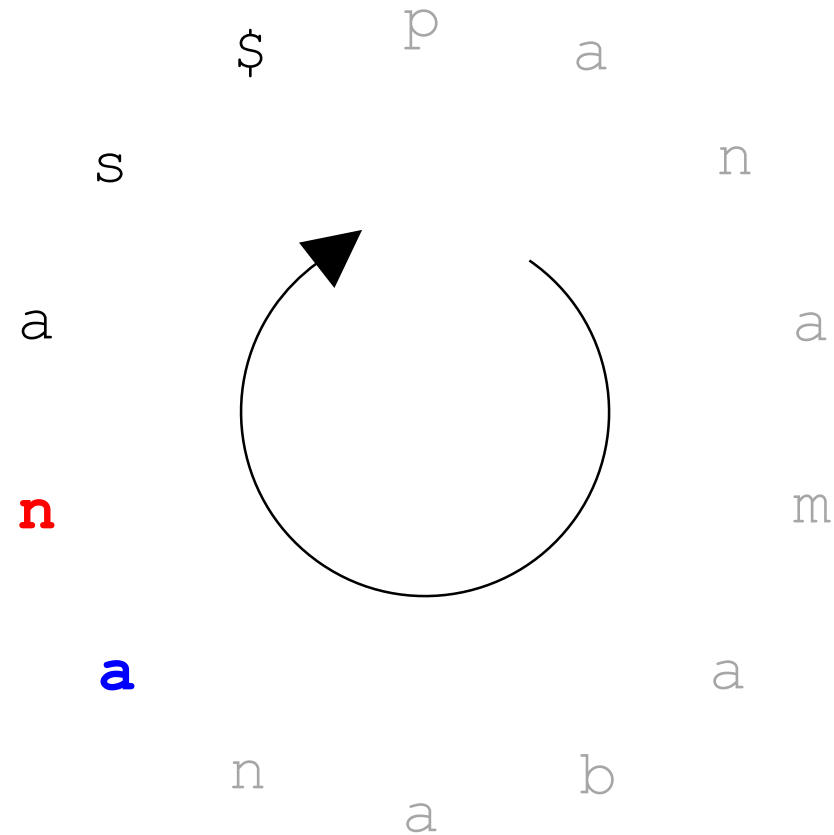
\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamabana**n**<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
**n**<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>





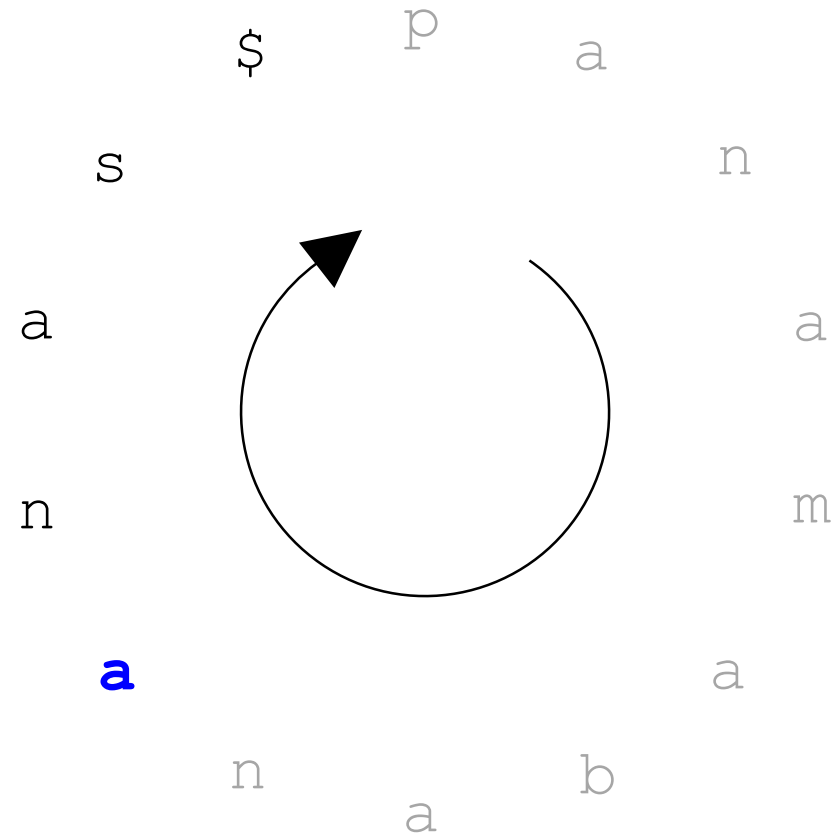
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
**n**<sub>3</sub>as\$panamaban**a**<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



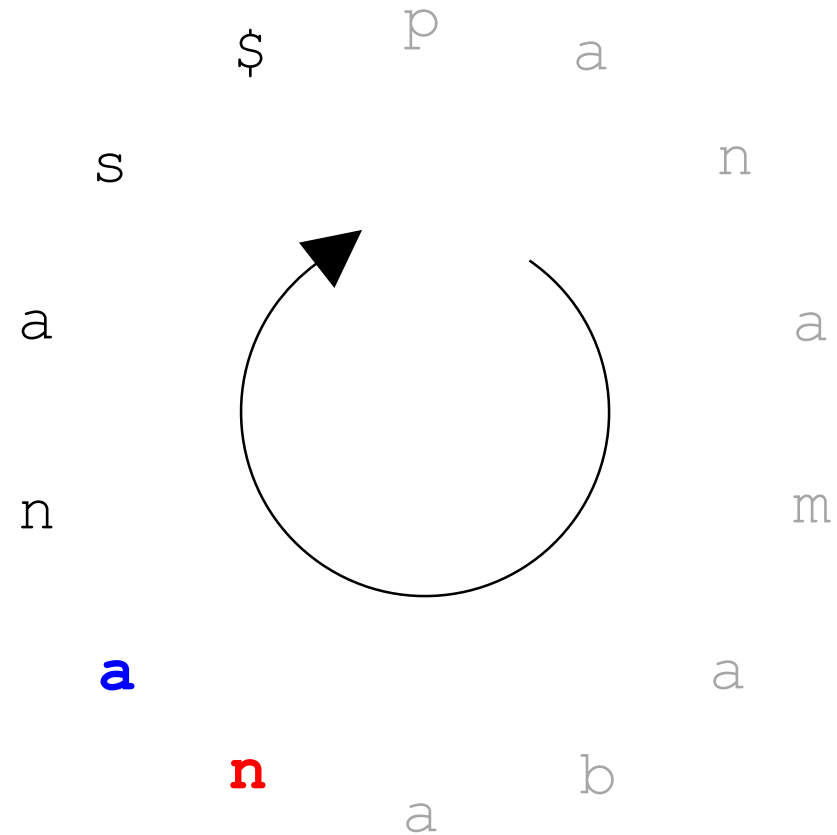
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
**a**<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamaban**a**<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



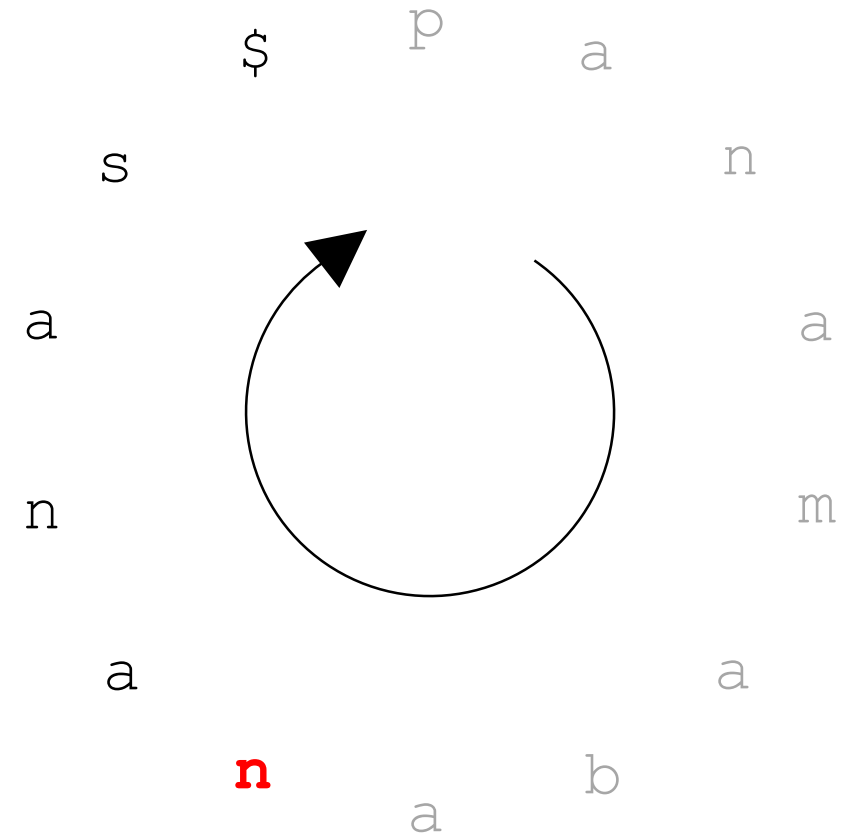
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
**a**<sub>5</sub>nas\$panamaba**n**<sub>2</sub>  
a<sub>6</sub>s\$panamabanana<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



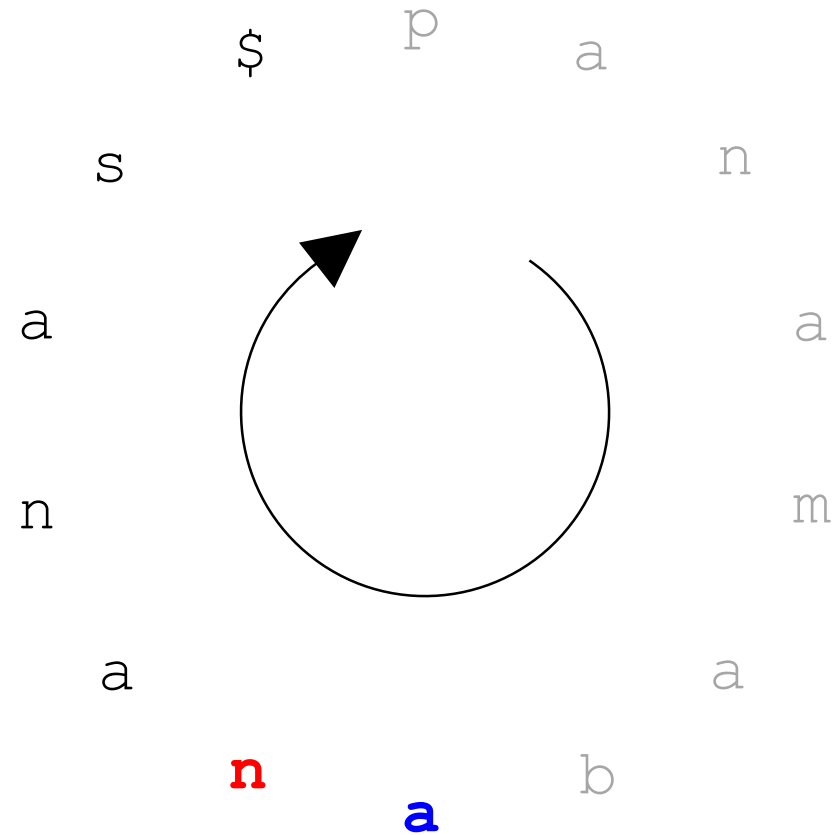
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaba**n**<sub>2</sub>  
a<sub>6</sub>s\$panamabanan<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
**n**<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

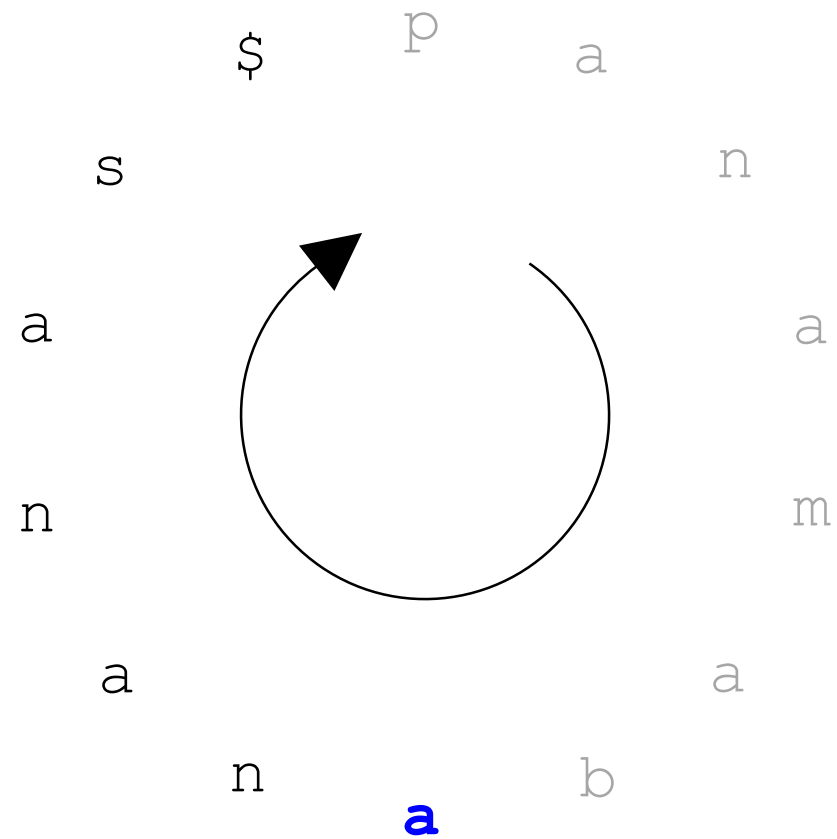


# Inverting BWT Again

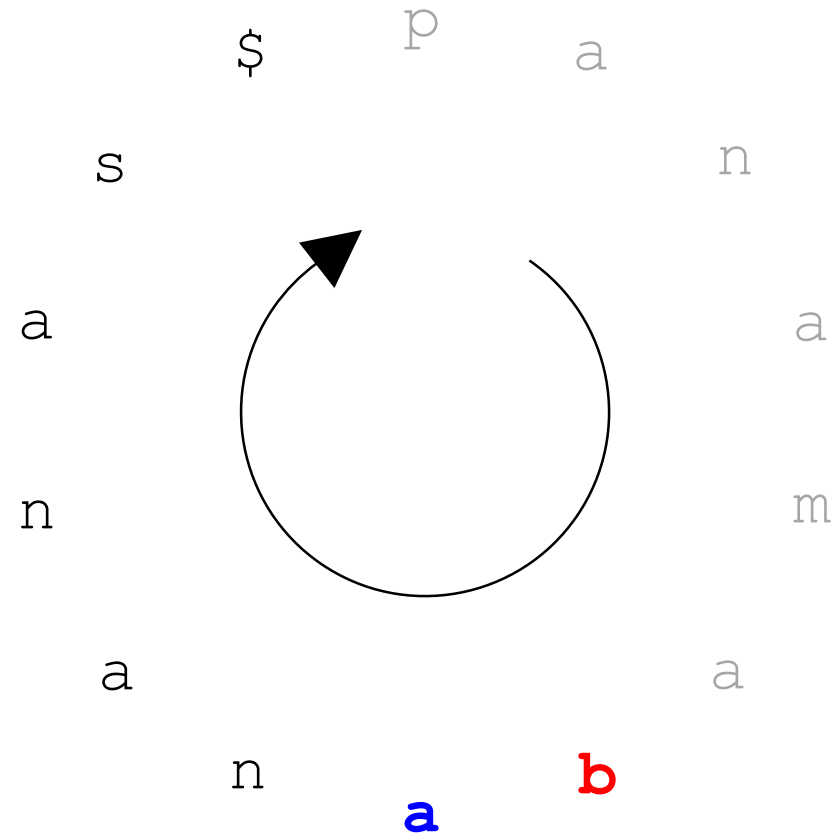
\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
**n**<sub>2</sub>anas\$panamab**a**<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



$\$$ <sub>1</sub> p a n a m a b a n a n a s  $s$ <sub>1</sub>  
a<sub>1</sub> b a n a n a s  $\$$  p a n a m  $m$ <sub>1</sub>  
a<sub>2</sub> m a b a n a n a s  $\$$  p a n  $n$ <sub>1</sub>  
a<sub>3</sub> n a m a b a n a n a s  $\$$  p  $p$ <sub>1</sub>  
**a**<sub>4</sub> n a n a s  $\$$  p a n a m a b  $b$ <sub>1</sub>  
a<sub>5</sub> n a s  $\$$  p a n a m a b a n  $n$ <sub>2</sub>  
a<sub>6</sub> s  $\$$  p a n a m a b a n a n  $n$ <sub>3</sub>  
b<sub>1</sub> a n a n a s  $\$$  p a n a m a  $a$ <sub>1</sub>  
m<sub>1</sub> a b a n a n a s  $\$$  p a n a  $a$ <sub>2</sub>  
n<sub>1</sub> a m a b a n a n a s  $\$$  p a  $a$ <sub>3</sub>  
n<sub>2</sub> a n a s  $\$$  p a n a m a b **a**<sub>4</sub>  
n<sub>3</sub> a s  $\$$  p a n a m a b a n a  $a$ <sub>5</sub>  
p<sub>1</sub> a n a m a b a n a n a s  $\$$   $s$ <sub>1</sub>  
s<sub>1</sub>  $\$$  p a n a m a b a n a n a  $a$ <sub>6</sub>

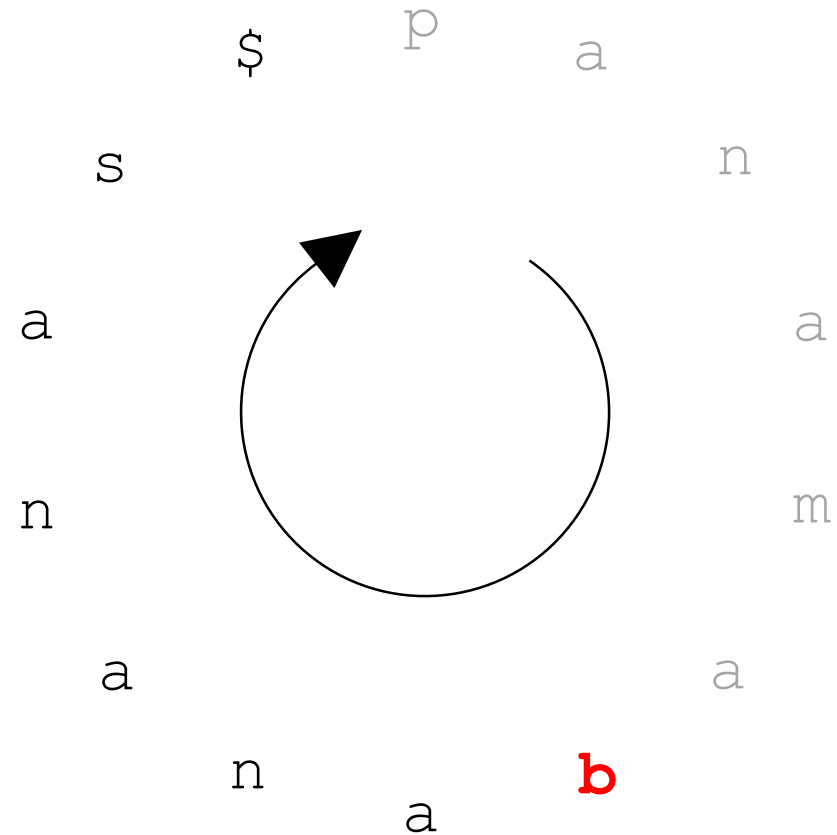


$\$$ <sub>1</sub> p a n a m a b a n a n a s  $s$ <sub>1</sub>  
a<sub>1</sub> b a n a n a s  $\$$  p a n a m  $m$ <sub>1</sub>  
a<sub>2</sub> m a b a n a n a s  $\$$  p a n  $n$ <sub>1</sub>  
a<sub>3</sub> n a m a b a n a n a s  $\$$  p  $p$ <sub>1</sub>  
**a**<sub>4</sub> n a n a s  $\$$  p a n a m a **b**<sub>1</sub>  
a<sub>5</sub> n a s  $\$$  p a n a m a b a n  $n$ <sub>2</sub>  
a<sub>6</sub> s  $\$$  p a n a m a b a n a n  $n$ <sub>3</sub>  
b<sub>1</sub> a n a n a s  $\$$  p a n a m a  $a$ <sub>1</sub>  
m<sub>1</sub> a b a n a n a s  $\$$  p a n a  $a$ <sub>2</sub>  
n<sub>1</sub> a m a b a n a n a s  $\$$  p a  $a$ <sub>3</sub>  
n<sub>2</sub> a n a s  $\$$  p a n a m a b a  $a$ <sub>4</sub>  
n<sub>3</sub> a s  $\$$  p a n a m a b a n a  $a$ <sub>5</sub>  
p<sub>1</sub> a n a m a b a n a n a s  $\$$   $s$ <sub>1</sub>  
s<sub>1</sub>  $\$$  p a n a m a b a n a n a  $a$ <sub>6</sub>



# Inverting BWT Again

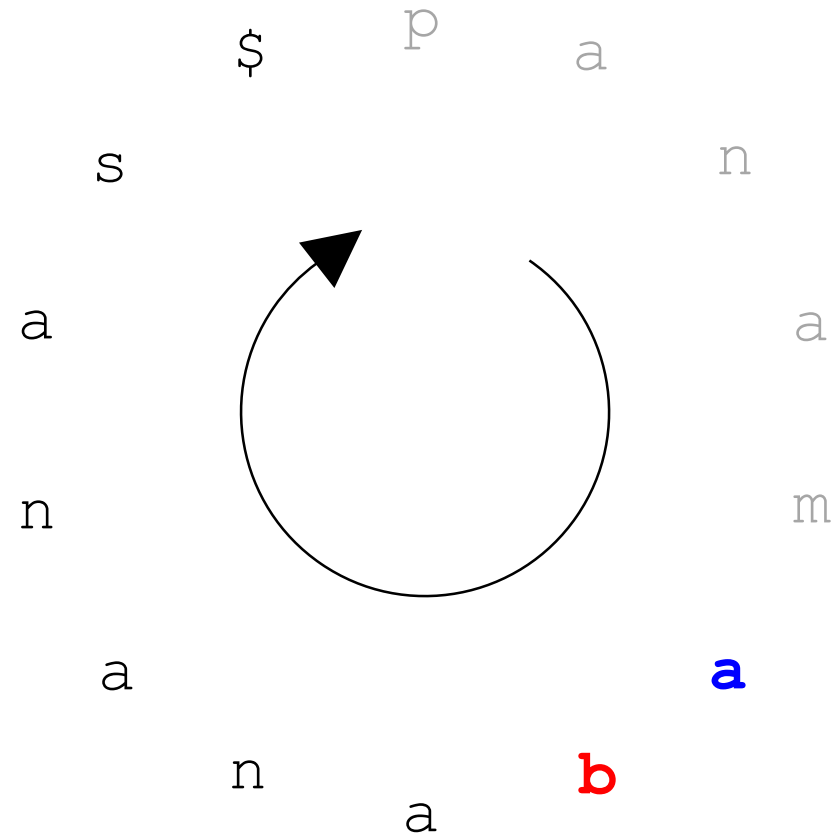
\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panama**b**<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
**b**<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>





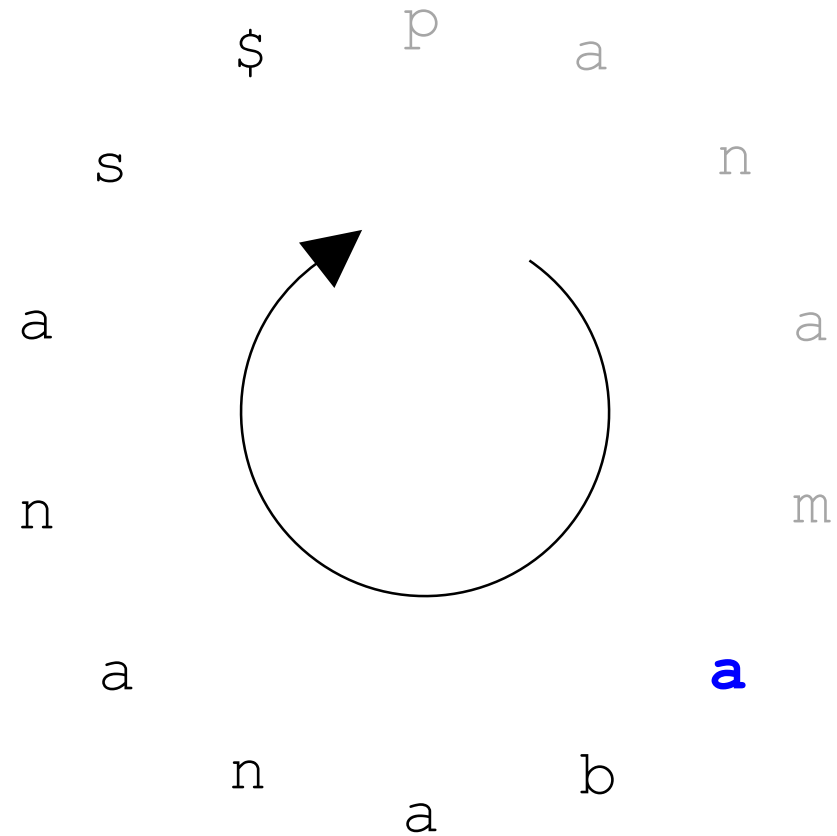
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
**b**<sub>1</sub>ananas\$panam**a**<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



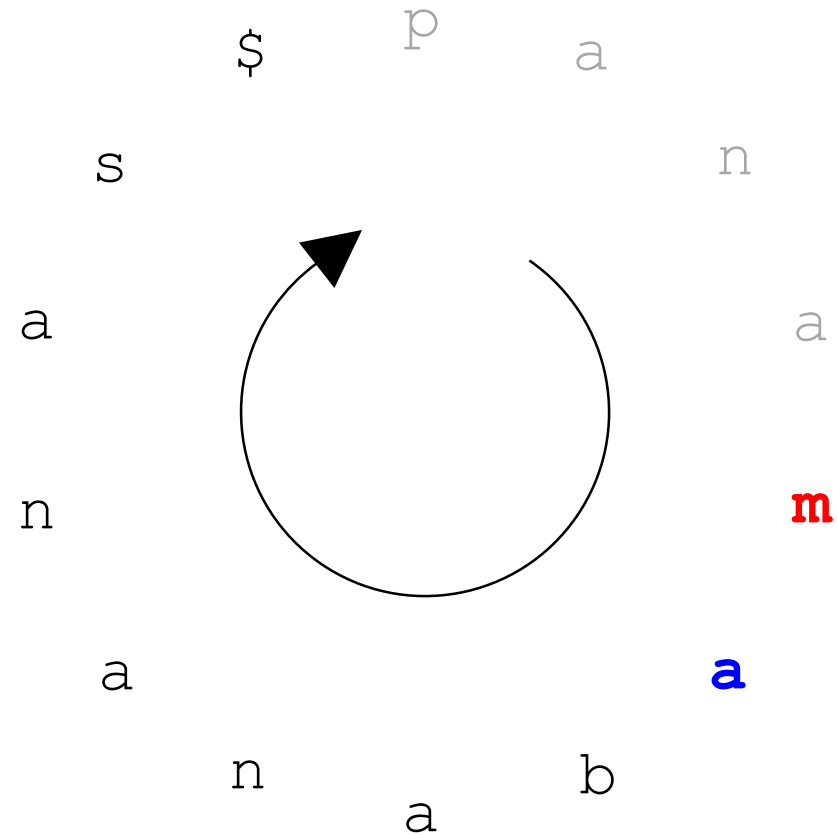
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
**a**<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panam**a**<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



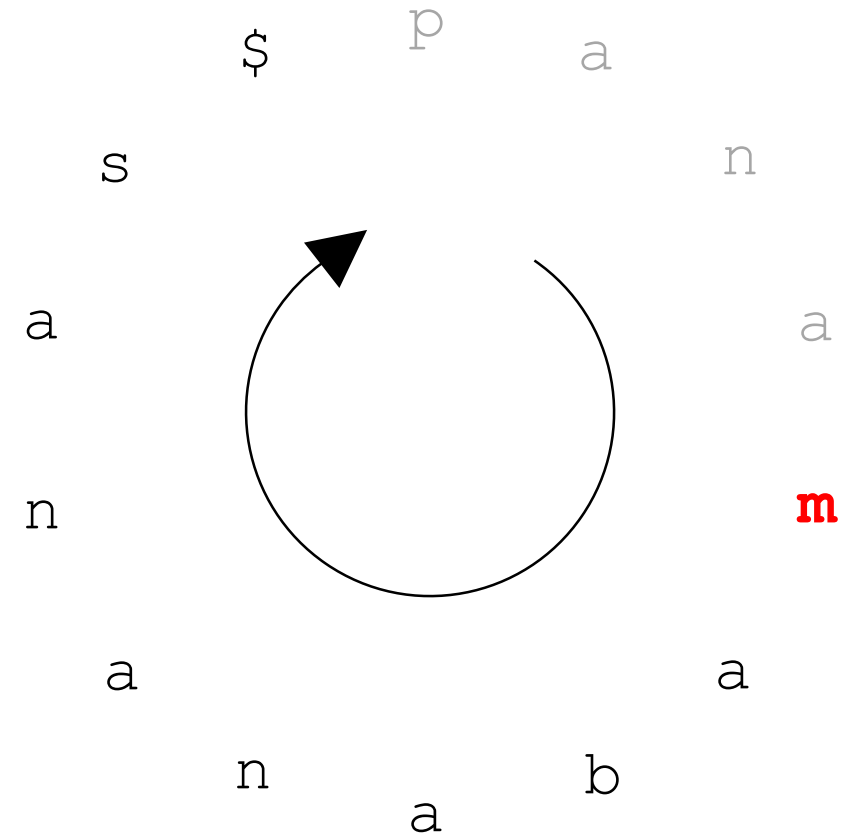
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
**a**<sub>1</sub>bananas\$pana**m**<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



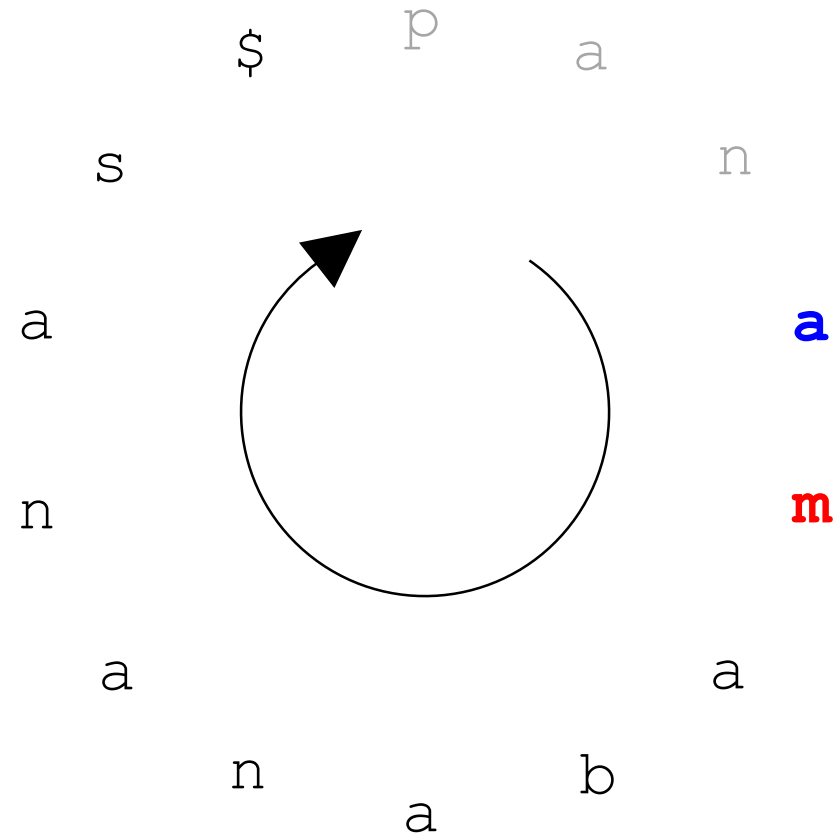
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$pana**m**<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
**m**<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



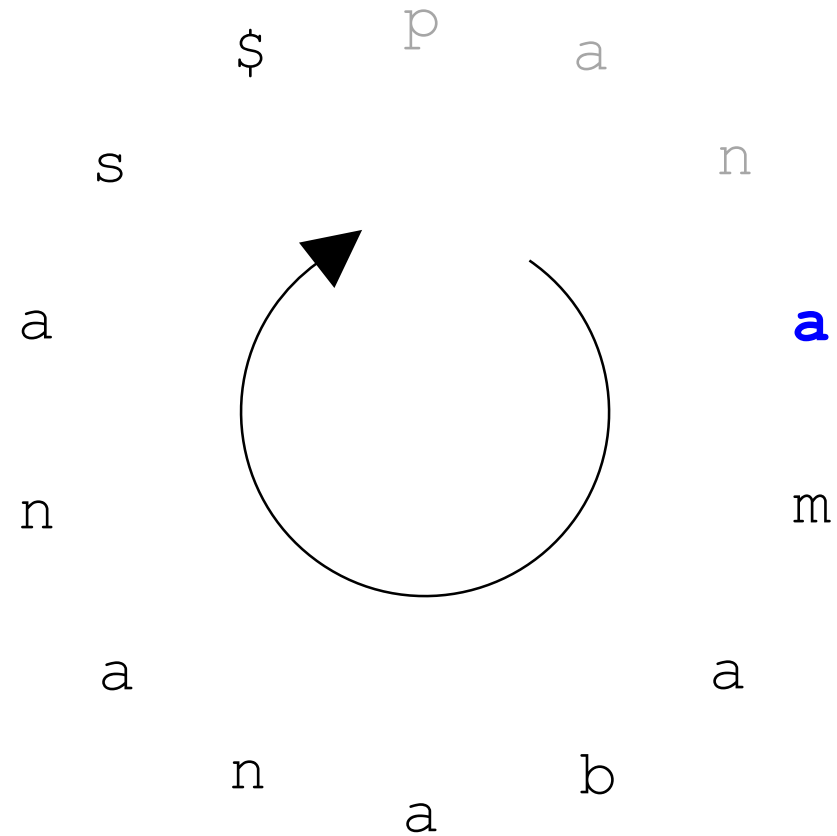
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
**m**<sub>1</sub>abananas\$pan**a**<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



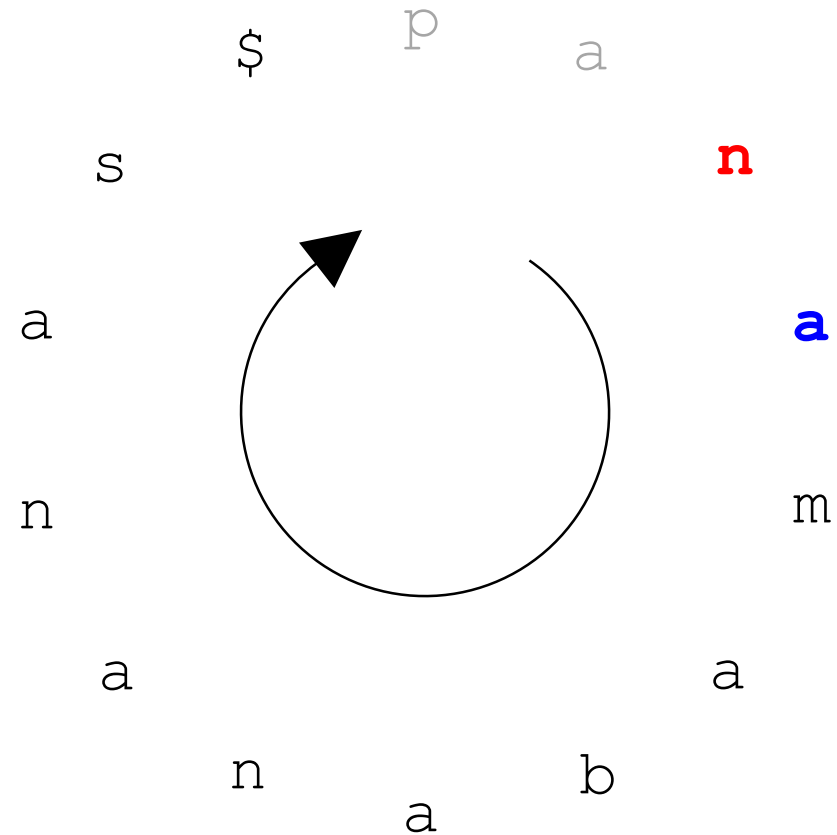
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
**a**<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamabanan<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pan**a**<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



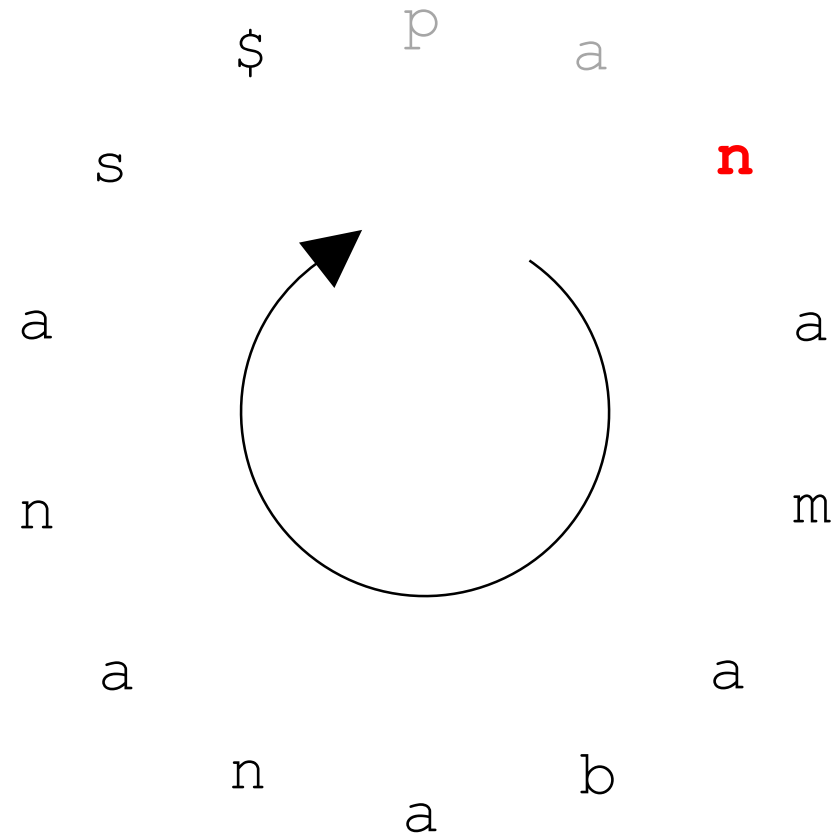
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
**a**<sub>2</sub>mabananas\$pa**n**<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



# Inverting BWT Again

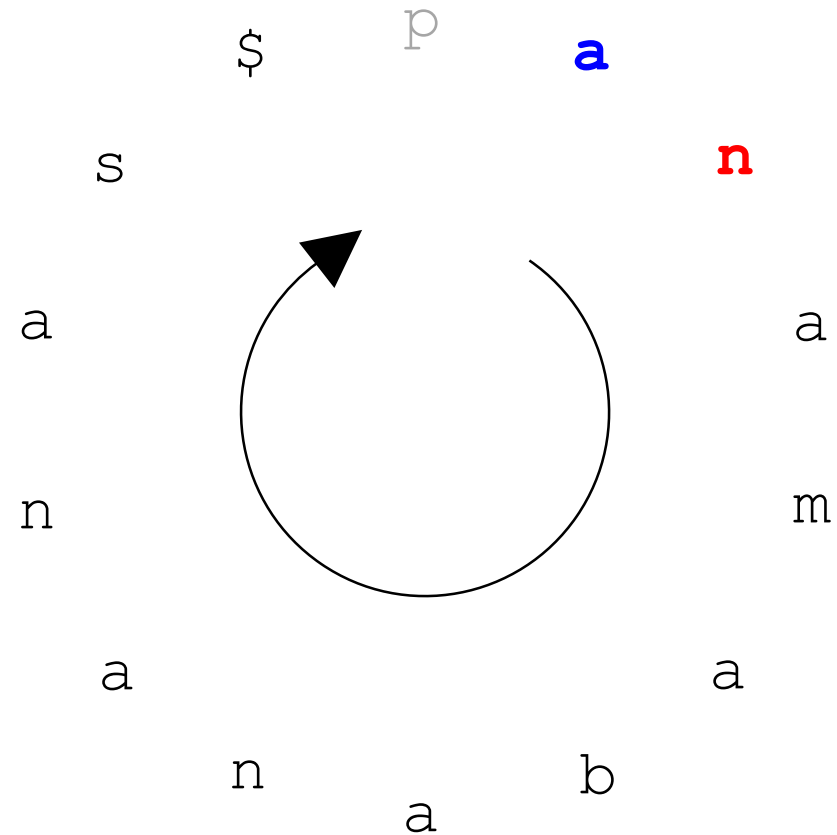
\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan**n**<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamabanan<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
**n**<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>





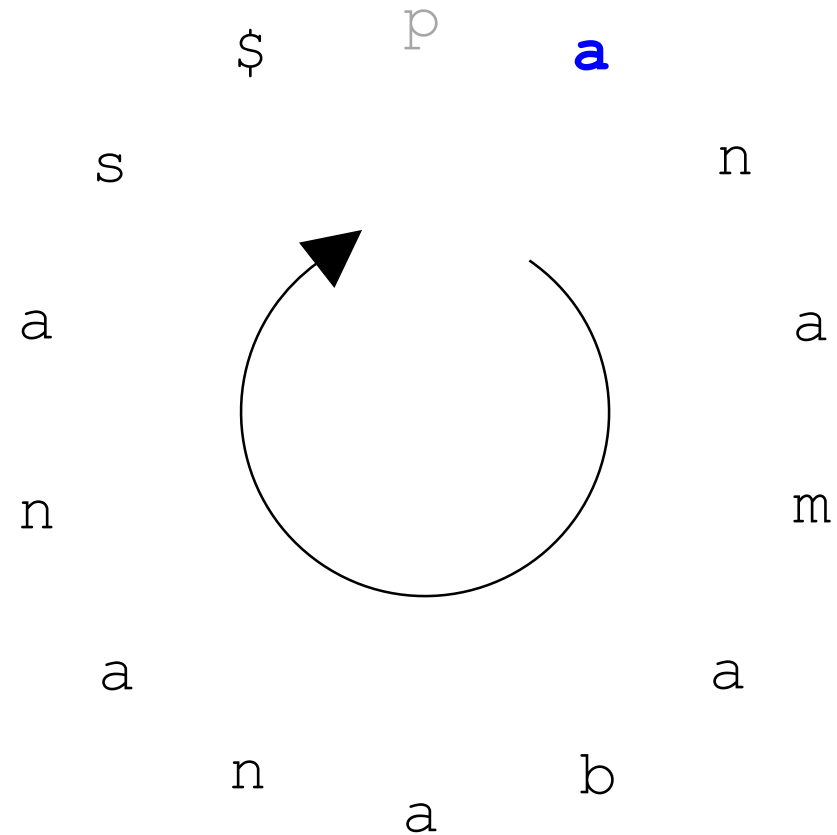
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
**n**<sub>1</sub>amabananas\$p**a**<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



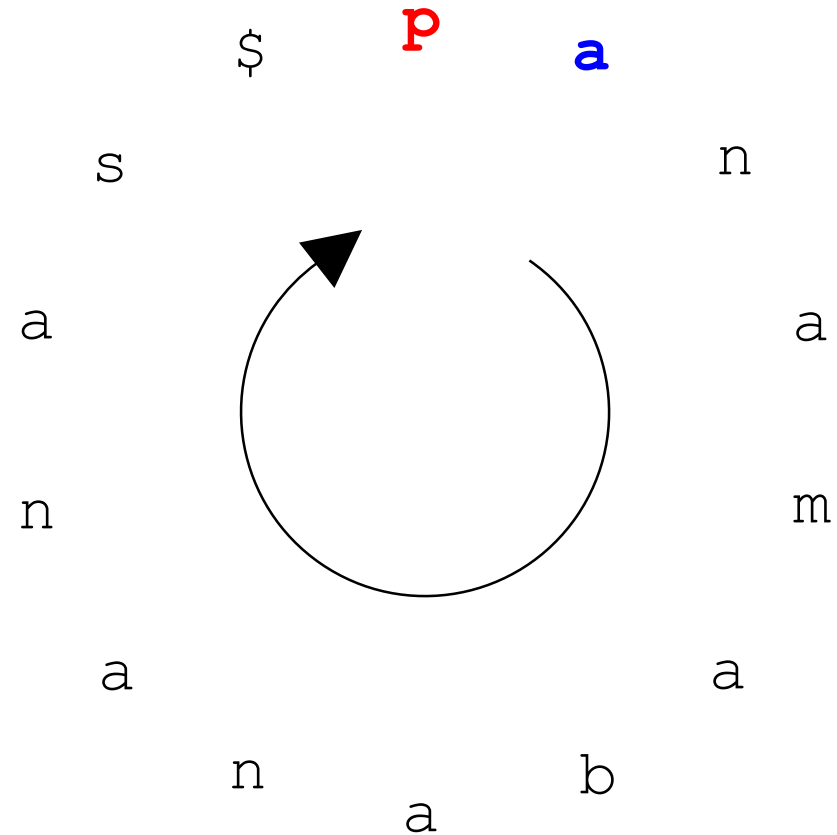
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a**<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamabanan<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$p<sub>1</sub>**a**<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



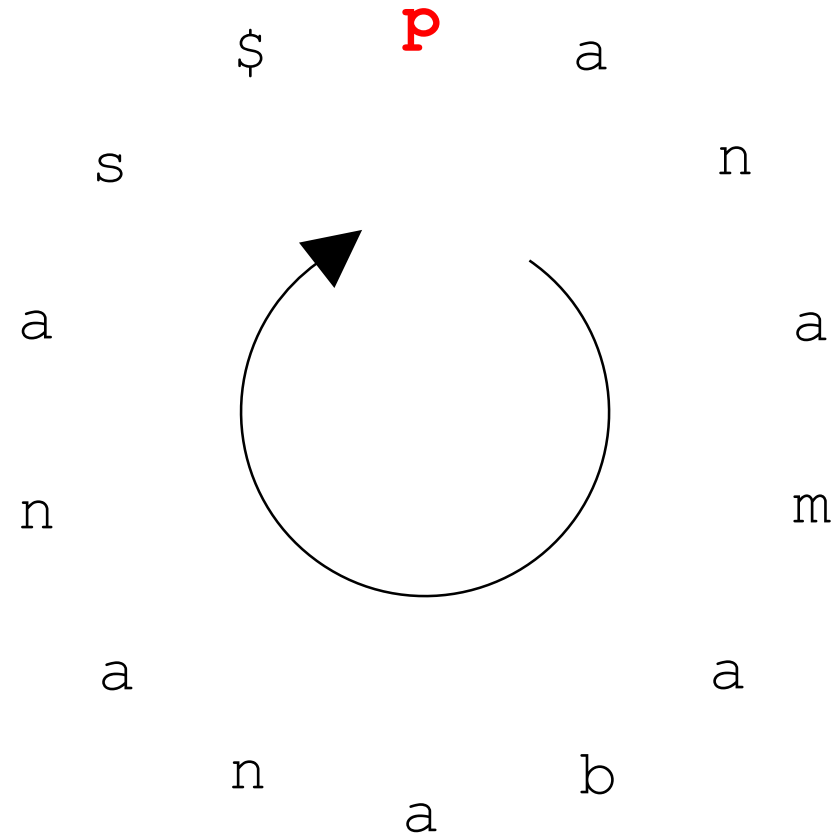
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a**<sub>3</sub>namabananas\$**p**<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



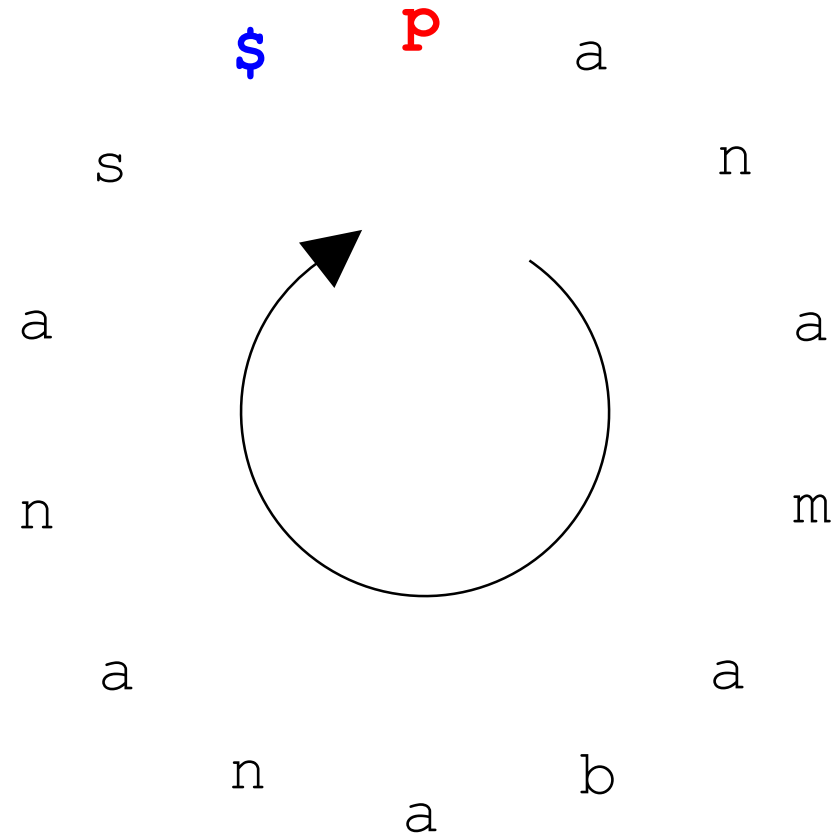
# Inverting BWT Again

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$**p<sub>1</sub>**  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
**p<sub>1</sub>**anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

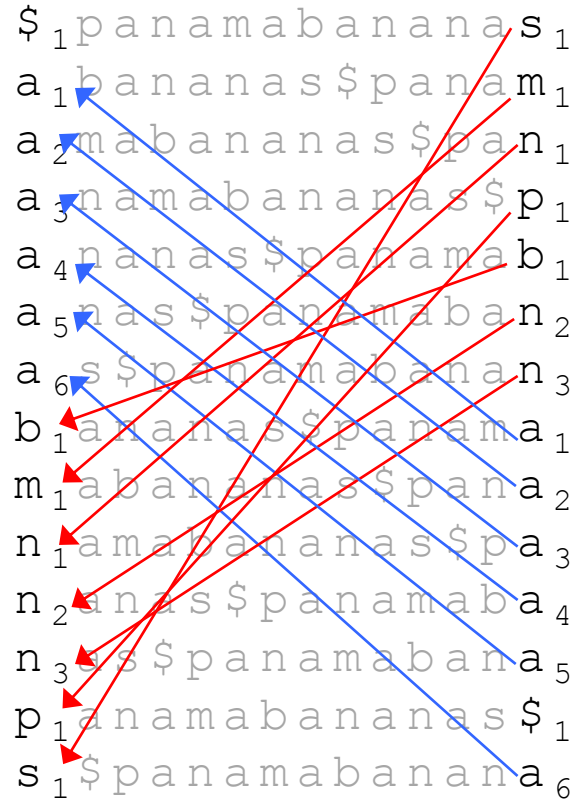


# We Are Done!

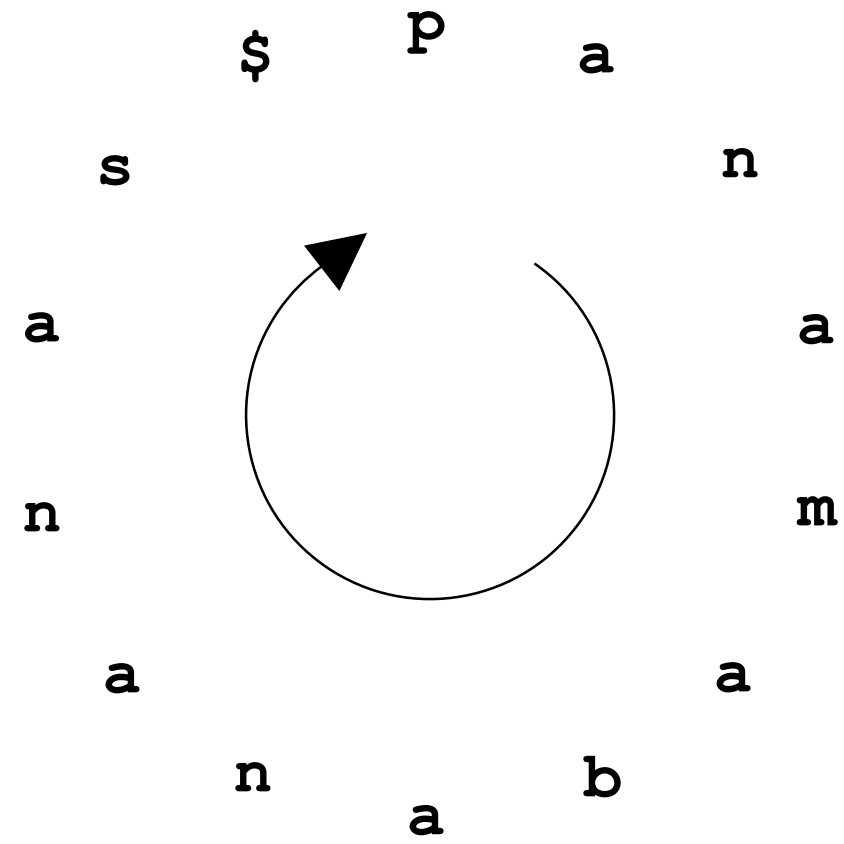
\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
**p**<sub>1</sub>anamabananas**\$**<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>



# This Was Fast!



- Memory:  $2|Text|$
- Time:  $O(|Text|)$



# Outline

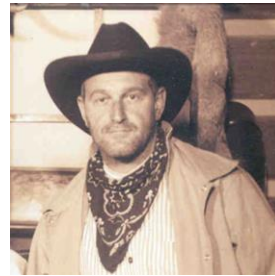
- Burrows-Wheeler Transform
- Inverting Burrows-Wheeler Transform
- **Using BWT for Pattern Matching**
- Suffix Arrays
- Approximate Pattern Matching

# Back to Pattern Matching

- Suffix Tree Pattern Matching:
  - Runtime:  $O(|Text| + |Patterns|)$
  - Memory:  $20 \cdot |Text|$

For human genome:

- $|Text| \approx 3 \cdot 10^9$
- Can we use  $BWT(Text)$  to design a more memory efficient linear-time algorithm for Multiple Pattern Matching?





# Finding Pattern Matches Using BWT

- Searching for **ana** in **p****ana**mab**anana**s

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```

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

- Searching for an **a** in panamabananas

\$<sub>1</sub>panamabananas<sub>1</sub>  
**a**<sub>1</sub>bananas\$panam<sub>1</sub>  
**a**<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a**<sub>3</sub>namabananas\$p<sub>1</sub>  
**a**<sub>4</sub>nanas\$panamab<sub>1</sub>  
**a**<sub>5</sub>nas\$panamaban<sub>2</sub>  
**a**<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

# Matching the Last Two Symbols (**na**)

- Searching for a**na** in panamabananas

\$<sub>1</sub>panamabananas<sub>1</sub>  
**a**<sub>1</sub>bananas\$pana**m**<sub>1</sub>  
**a**<sub>2</sub>mabananas\$pa**n**<sub>1</sub>  
**a**<sub>3</sub>namabananas\$**p**<sub>1</sub>  
**a**<sub>4</sub>nanas\$panama**b**<sub>1</sub>  
**a**<sub>5</sub>nas\$panamaba**n**<sub>2</sub>  
**a**<sub>6</sub>s\$panamabana**n**<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabana<sub>6</sub>

# Three Matches of **na** Found!

- Searching for a**na** in panamabananas

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
**a**<sub>2</sub>mabananas\$pa**n**<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
**a**<sub>5</sub>nas\$panamaba**n**<sub>2</sub>  
**a**<sub>6</sub>s\$panamabana**n**<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabana<sub>6</sub>

Three green arrows point from the bolded 'na' substrings in the grid to the positions of the 'na' in the original string 'panamabananas':  
- Arrow 1: From **a**<sub>2</sub>mabananas\$pa**n**<sub>1</sub> to the 'na' at index 4.  
- Arrow 2: From **a**<sub>5</sub>nas\$panamaba**n**<sub>2</sub> to the 'na' at index 9.  
- Arrow 3: From **a**<sub>6</sub>s\$panamabana**n**<sub>3</sub> to the 'na' at index 14.

# Three Matches of **na** Found!

- Searching for a**na** in panamabananas

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pa**n**<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaba**n**<sub>2</sub>  
a<sub>6</sub>s\$panamabana**n**<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
**n**<sub>1</sub>amabananas\$pa<sub>3</sub>  
**n**<sub>2</sub>anas\$panamaba<sub>4</sub>  
**n**<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabana<sub>6</sub>

# Three Matches of **na** Found!

- Searching for a**na** in panamabananas

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
**n<sub>1</sub>a**mabananas\$pa<sub>3</sub>  
**n<sub>2</sub>a**nas\$panamaba<sub>4</sub>  
**n<sub>3</sub>a**s\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

# Matching **ana**

- Searching for **ana** in panamabananas

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
a<sub>3</sub>namabananas\$p<sub>1</sub>  
a<sub>4</sub>nanas\$panamab<sub>1</sub>  
a<sub>5</sub>nas\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
**n<sub>1</sub>a**mabananas\$p<sub>1</sub>**a<sub>3</sub>**  
**n<sub>2</sub>a**nas\$panamab<sub>1</sub>**a<sub>4</sub>**  
**n<sub>3</sub>a**s\$panamaban<sub>2</sub>**a<sub>5</sub>**  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

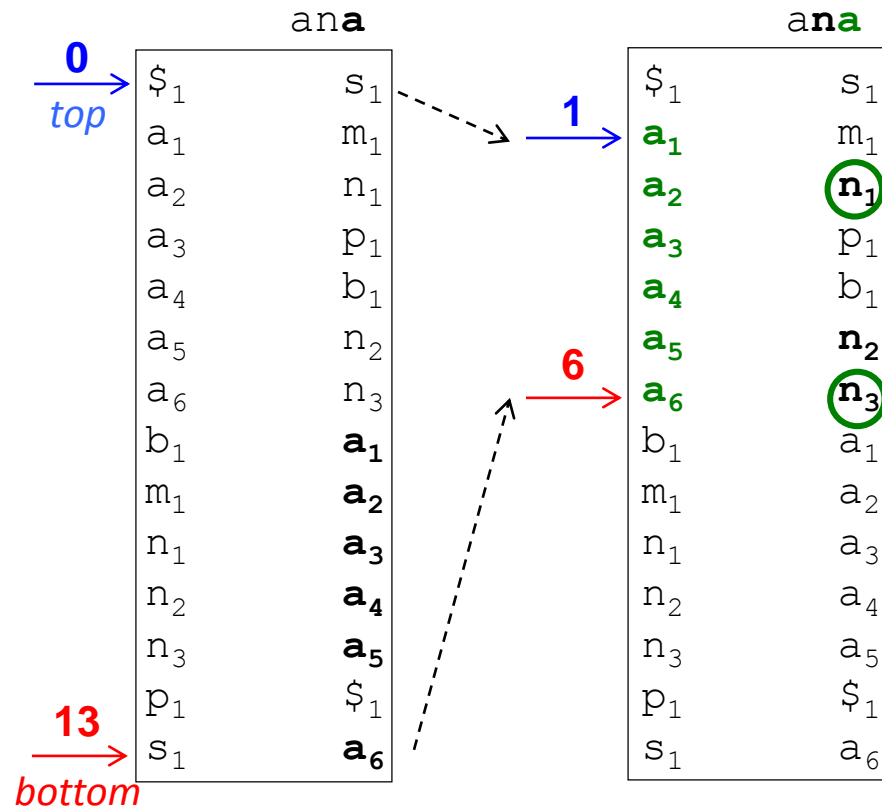
# Three Matches of **ana** Found!

- Searching for **ana** in panamabananas

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```



# Searching for **ana** using *top* and *bottom* pointers



*topIndex*  $\leftarrow$  first position of symbol among positions from *top* to *bottom* in *LastColumn*

*bottomIndex*  $\leftarrow$  last position of symbol among positions from *top* to *bottom* in *LastColumn*

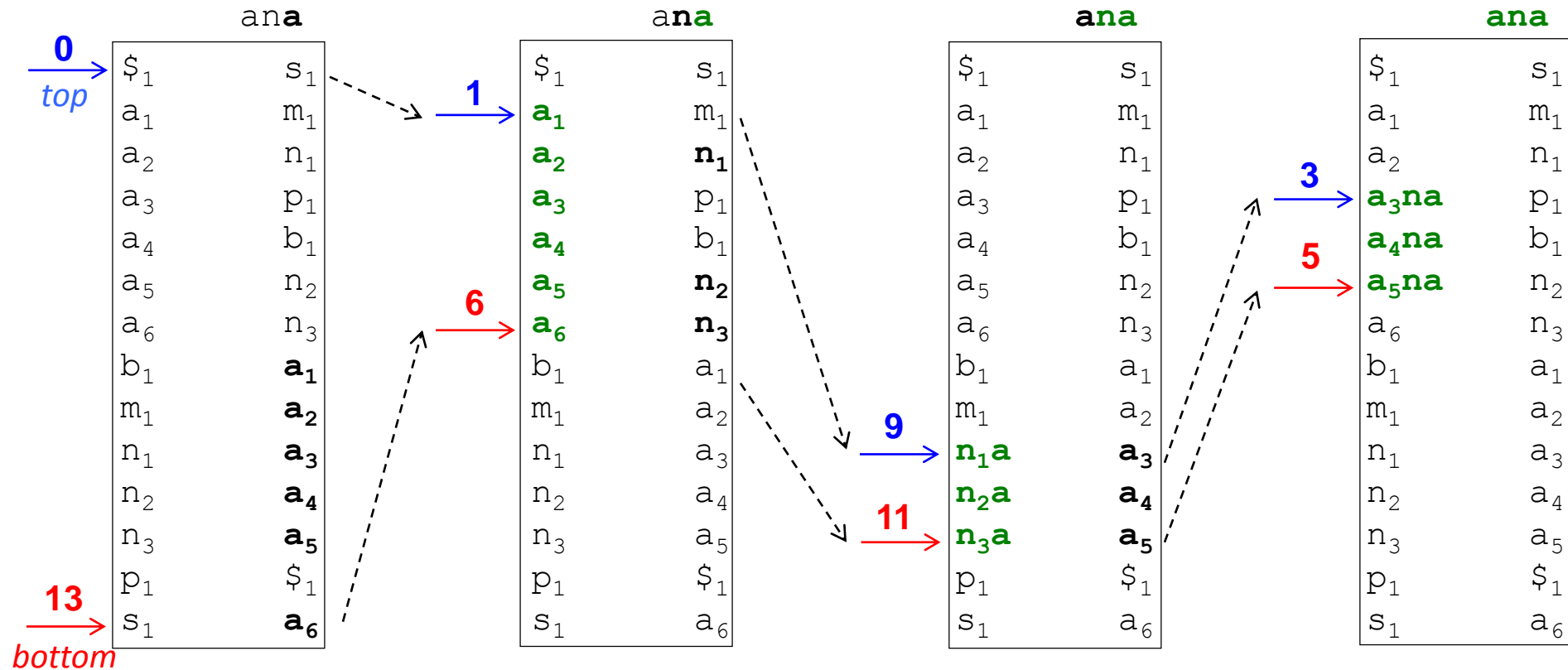
# BWMatching

```
BWMATCHING(FirstColumn, LastColumn, Pattern, LASTTOFIRST)  
  top  $\leftarrow$  0  
  bottom  $\leftarrow$  |LastColumn| - 1  
  while top  $\leq$  bottom  
    if Pattern is nonempty  
      symbol  $\leftarrow$  last letter in Pattern  
      remove last letter from Pattern  
      if positions from top to bottom in LastColumn contain symbol  
        topIndex  $\leftarrow$  first position of symbol among positions from top to bottom  
          in LastColumn  
        bottomIndex  $\leftarrow$  last position of symbol among positions from top to  
          bottom in LastColumn  
        top  $\leftarrow$  LASTTOFIRST(topIndex)  
        bottom  $\leftarrow$  LASTTOFIRST(bottomIndex)  
      else  
        return 0  
    else  
      return bottom - top + 1
```

Given a symbol at position *index* in *LastColumn*,  
**LastToFirst**(*index*) defines the position of this symbol in *FirstColumn*

# BWMatching is slow:

it analyzes every symbol from *top* to *bottom* in each step!



if positions from *top* to *bottom* in *LastColumn* contain symbol  
*topIndex* ← first position of symbol among positions from *top* to *bottom* in *LastColumn*  
*bottomIndex* ← last position of symbol among positions from *top* to *bottom* in *LastColumn*

# Introducing *Count* Array

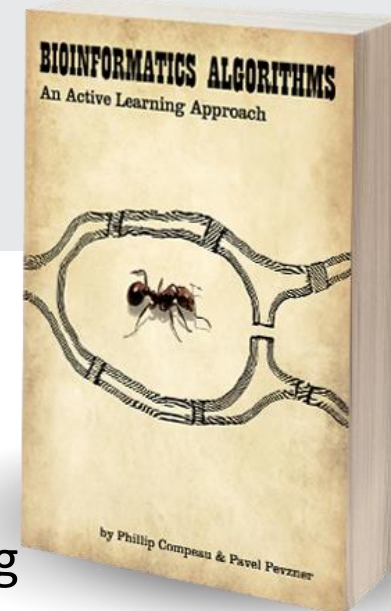
<i>i</i>	<i>FirstColumn</i>	<i>LastColumn</i>	LASTTOFIRST( <i>i</i> )	COUNT						
				\$	a	b	m	n	p	s
0	\$ <sub>1</sub>	s <sub>1</sub>	13	0	0	0	0	0	0	0
1	a <sub>1</sub>	m <sub>1</sub>	8	0	0	0	0	0	0	1
2	a <sub>2</sub>	n <sub>1</sub>	9	0	0	0	1	0	0	1
3	a <sub>3</sub>	p <sub>1</sub>	12	0	0	0	1	1	0	1
4	a <sub>4</sub>	b <sub>1</sub>	7	0	0	0	1	1	1	1
5	a <sub>5</sub>	n <sub>2</sub>	10	0	0	1	1	1	1	1
6	a <sub>6</sub>	n <sub>3</sub>	11	0	0	1	1	2	1	1
7	b <sub>1</sub>	a <sub>1</sub>	1	0	0	1	1	3	1	1
8	m <sub>1</sub>	a <sub>2</sub>	2	0	1	1	1	3	1	1
9	n <sub>1</sub>	a <sub>3</sub>	3	0	2	1	1	3	1	1
10	n <sub>2</sub>	a <sub>4</sub>	4	0	3	1	1	3	1	1
11	n <sub>3</sub>	a <sub>5</sub>	5	0	4	1	1	3	1	1
12	p <sub>1</sub>	\$ <sub>1</sub>	0	0	5	1	1	3	1	1
13	s <sub>1</sub>	a <sub>6</sub>	6	1	5	1	1	3	1	1
				1	6	1	1	3	1	1

*Count*<sub>*symbol*</sub>(*i*, *LastColumn*):

#occurrences of *symbol* in the first *i* positions of *LastColumn*

# BetterBWMatching

```
BETTERBWMATCHING(FIRSTOCCURRENCE, LastColumn, Pattern, COUNT)
  top  $\leftarrow$  0
  bottom  $\leftarrow$  |LastColumn| - 1
  while top  $\leq$  bottom
    if Pattern is nonempty
      symbol  $\leftarrow$  last letter in Pattern
      remove last letter from Pattern
      top  $\leftarrow$  FIRSTOCCURRENCE(symbol) + COUNTsymbol(top, LastColumn)
      bottom  $\leftarrow$  FIRSTOCCURRENCE(symbol) + COUNTsymbol(bottom + 1,
        LastColumn) - 1
    else
      return bottom - top + 1
  return
```



# Where Are the Matches?

- We know that **ana** occurs 3 times, but where does **ana** appear in *Text*???

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a<sub>3</sub>na**mabananas\$p<sub>1</sub>  
**a<sub>4</sub>na**nas\$panamab<sub>1</sub>  
**a<sub>5</sub>na**s\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

# Outline

- Burrows-Wheeler Transform
- Inverting Burrows-Wheeler Transform
- Using BWT for Pattern Matching
- **Suffix Arrays**
- Approximate Pattern Matching

# Where Are the Matches?

- **Suffix array** holds starting position of each suffix

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```



# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabananas\$

1	3	\$ <sub>1</sub>	panamabananas <sub>1</sub>
		a <sub>1</sub>	bananas\$panam <sub>1</sub>
		a <sub>2</sub>	mabananas\$pan <sub>1</sub>
		a <sub>3</sub>	namabananas\$p <sub>1</sub>
		a <sub>4</sub>	nanas\$panamab <sub>1</sub>
		a <sub>5</sub>	nas\$panamaban <sub>2</sub>
		a <sub>6</sub>	s\$panamabanan <sub>3</sub>
		b <sub>1</sub>	ananas\$panama <sub>1</sub>
		m <sub>1</sub>	abananas\$pana <sub>2</sub>
		n <sub>1</sub>	amabananas\$pa <sub>3</sub>
		n <sub>2</sub>	anas\$panamaba <sub>4</sub>
		n <sub>3</sub>	as\$panamabana <sub>5</sub>
		p <sub>1</sub>	anamabananas\$ <sub>1</sub>
		s <sub>1</sub>	\$panamabannana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panam**abananas**\$

1	3	\$ <sub>1</sub>	panamabananas <sub>1</sub>
	5	<b>a<sub>1</sub></b>	<b>abananas</b> \$panam <sub>1</sub>
		a <sub>2</sub>	mabananas\$pan <sub>1</sub>
		a <sub>3</sub>	namabananas\$p <sub>1</sub>
		a <sub>4</sub>	nanas\$panamab <sub>1</sub>
		a <sub>5</sub>	nas\$panamaban <sub>2</sub>
		a <sub>6</sub>	s\$panamaban <sub>3</sub>
		b <sub>1</sub>	ananas\$panama <sub>1</sub>
		m <sub>1</sub>	abananas\$pana <sub>2</sub>
		n <sub>1</sub>	amabananas\$pa <sub>3</sub>
		n <sub>2</sub>	anas\$panamaba <sub>4</sub>
		n <sub>3</sub>	as\$panamabana <sub>5</sub>
		p <sub>1</sub>	anamabananas\$_ <sub>1</sub>
		s <sub>1</sub>	\$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

pan**amabananas**\$

1	3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5		a <sub>1</sub> <b>bananas</b> \$panam <sub>1</sub>
3		a <sub>2</sub> <b>mabananas</b> \$pan <sub>1</sub>
		a <sub>3</sub> namabananas\$pa <sub>1</sub>
		a <sub>4</sub> nanas\$panamab <sub>1</sub>
		a <sub>5</sub> nas\$panamaban <sub>2</sub>
		a <sub>6</sub> s\$panamabanan <sub>3</sub>
		b <sub>1</sub> ananas\$panama <sub>1</sub>
		m <sub>1</sub> abananas\$pana <sub>2</sub>
		n <sub>1</sub> amabananas\$pa <sub>3</sub>
		n <sub>2</sub> anas\$panamaba <sub>4</sub>
		n <sub>3</sub> as\$panamabana <sub>5</sub>
		p <sub>1</sub> anamabananas\$ <sub>1</sub>
		s <sub>1</sub> \$panamabannana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

**panamabananas\$**

1 3	<b>\$</b> <sub>1</sub> panamabananas <sub>1</sub>
5	<b>a</b> <sub>1</sub> <b>bananas\$</b> panam <sub>1</sub>
3	<b>a</b> <sub>2</sub> <b>mabananas\$</b> pan <sub>1</sub>
1	<b>a</b> <sub>3</sub> <b>namabananas\$</b> p <sub>1</sub>
	a <sub>4</sub> nanas\$panamab <sub>1</sub>
	a <sub>5</sub> nas\$panamaban <sub>2</sub>
	a <sub>6</sub> s\$panamabanan <sub>3</sub>
	b <sub>1</sub> ananas\$panama <sub>1</sub>
	m <sub>1</sub> abananas\$pana <sub>2</sub>
	n <sub>1</sub> amabananas\$pa <sub>3</sub>
	n <sub>2</sub> anas\$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamab**ananas**\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
	a <sub>5</sub> nas\$panamaban <sub>2</sub>
	a <sub>6</sub> s\$panamabanan <sub>3</sub>
	b <sub>1</sub> ananas\$panama <sub>1</sub>
	m <sub>1</sub> abananas\$pana <sub>2</sub>
	n <sub>1</sub> amabananas\$pa <sub>3</sub>
	n <sub>2</sub> anas\$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$_ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamaban**anas**\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
	a <sub>6</sub> s\$panamaban <sub>3</sub>
	b <sub>1</sub> ananas\$panama <sub>1</sub>
	m <sub>1</sub> abananas\$pana <sub>2</sub>
	n <sub>1</sub> amabananas\$pa <sub>3</sub>
	n <sub>2</sub> anas\$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$_ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabananas\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
1 1	a <sub>6</sub> s\$panamaban <sub>3</sub>
	b <sub>1</sub> ananas\$panama <sub>1</sub>
	m <sub>1</sub> abananas\$pana <sub>2</sub>
	n <sub>1</sub> amabananas\$pa <sub>3</sub>
	n <sub>2</sub> anas\$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$_ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panama**bananas**\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
1 1	a <sub>6</sub> s\$panamaban <sub>3</sub>
6	b <sub>1</sub> ananas\$panama <sub>1</sub>
	m <sub>1</sub> abananas\$pana <sub>2</sub>
	n <sub>1</sub> amabananas\$pa <sub>3</sub>
	n <sub>2</sub> anas\$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$_ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>



# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panam**abananas**\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
1 1	a <sub>6</sub> s\$panamaban <sub>3</sub>
6	b <sub>1</sub> ananas\$panama <sub>1</sub>
4	m <sub>1</sub> abananas\$pana <sub>2</sub>
	n <sub>1</sub> amabananas\$pa <sub>3</sub>
	n <sub>2</sub> anas\$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$_ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabananas\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
1 1	a <sub>6</sub> s\$panamaban <sub>3</sub>
6	b <sub>1</sub> ananas\$panama <sub>1</sub>
4	m <sub>1</sub> abananas\$pana <sub>2</sub>
2	n <sub>1</sub> amabananas\$pa <sub>3</sub>
	n <sub>2</sub> anas\$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamab**ananas**\$

1 3	\$ <sub>1</sub> panamab <b>ananas</b> <sub>1</sub>
5	<b>a</b> <sub>1</sub> <b>bananas</b> \$panam <sub>1</sub>
3	<b>a</b> <sub>2</sub> <b>mabanas</b> \$pan <sub>1</sub>
1	<b>a</b> <sub>3</sub> <b>namabanas</b> \$p <sub>1</sub>
7	<b>a</b> <sub>4</sub> <b>nanas</b> \$panamab <sub>1</sub>
9	<b>a</b> <sub>5</sub> <b>nas</b> \$panamaban <sub>2</sub>
1 1	<b>a</b> <sub>6</sub> <b>s</b> \$panamabanan <sub>3</sub>
6	<b>b</b> <sub>1</sub> <b>ananas</b> \$panama <sub>1</sub>
4	<b>m</b> <sub>1</sub> <b>abanas</b> \$pana <sub>2</sub>
2	<b>n</b> <sub>1</sub> <b>amabanas</b> \$pa <sub>3</sub>
8	<b>n</b> <sub>2</sub> <b>anas</b> \$panamaba <sub>4</sub>
	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabanas\$ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabanan**as**\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
1 1	a <sub>6</sub> s\$panamabanan <sub>3</sub>
6	b <sub>1</sub> ananas\$panama <sub>1</sub>
4	m <sub>1</sub> abananas\$pana <sub>2</sub>
2	n <sub>1</sub> amabananas\$pa <sub>3</sub>
8	n <sub>2</sub> anas\$panamaba <sub>4</sub>
1 0	n <sub>3</sub> as\$panamabana <sub>5</sub>
	p <sub>1</sub> anamabananas\$ <sub>1</sub>
	s <sub>1</sub> \$panamabannana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

**panamabananas\$**

13	<b>\$<sub>1</sub></b> panamabananas <sub>1</sub>
5	<b>a<sub>1</sub></b> bananas\$panam <sub>1</sub>
3	<b>a<sub>2</sub></b> mabananas\$pan <sub>1</sub>
1	<b>a<sub>3</sub></b> namabananas\$p <sub>1</sub>
7	<b>a<sub>4</sub></b> nanas\$panamab <sub>1</sub>
9	<b>a<sub>5</sub></b> nas\$panamaban <sub>2</sub>
11	<b>a<sub>6</sub></b> s\$panamaban <sub>3</sub>
6	<b>b<sub>1</sub></b> ananas\$panama <sub>1</sub>
4	<b>m<sub>1</sub></b> abananas\$pana <sub>2</sub>
2	<b>n<sub>1</sub></b> amabananas\$pa <sub>3</sub>
8	<b>n<sub>2</sub></b> anas\$panamaba <sub>4</sub>
10	<b>n<sub>3</sub></b> as\$panamabana <sub>5</sub>
0	<b>p<sub>1</sub></b> anamabananas\$ <sub>1</sub>
	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabanana**s**\$

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
1 1	a <sub>6</sub> s\$panamaban <sub>3</sub>
6	b <sub>1</sub> ananas\$panama <sub>1</sub>
4	m <sub>1</sub> abananas\$pana <sub>2</sub>
2	n <sub>1</sub> amabananas\$pa <sub>3</sub>
8	n <sub>2</sub> anas\$panamaba <sub>4</sub>
1 0	n <sub>3</sub> as\$panamabana <sub>5</sub>
0	p <sub>1</sub> anamabananas\$_ <sub>1</sub>
1 2	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

1 3	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
1	a <sub>3</sub> namabananas\$p <sub>1</sub>
7	a <sub>4</sub> nanas\$panamab <sub>1</sub>
9	a <sub>5</sub> nas\$panamaban <sub>2</sub>
1 1	a <sub>6</sub> s\$panamaban <sub>3</sub>
6	b <sub>1</sub> ananas\$panama <sub>1</sub>
4	m <sub>1</sub> abananas\$pana <sub>2</sub>
2	n <sub>1</sub> amabananas\$pa <sub>3</sub>
8	n <sub>2</sub> anas\$panamaba <sub>4</sub>
1 0	n <sub>3</sub> as\$panamabana <sub>5</sub>
0	p <sub>1</sub> anamabananas\$ <sub>1</sub>
1 2	s <sub>1</sub> \$panamabanana <sub>6</sub>

# Using the Suffix Array to Find Matches

- Thus, **ana** occurs at positions **1, 7, 9**:

- `panamabananas$`



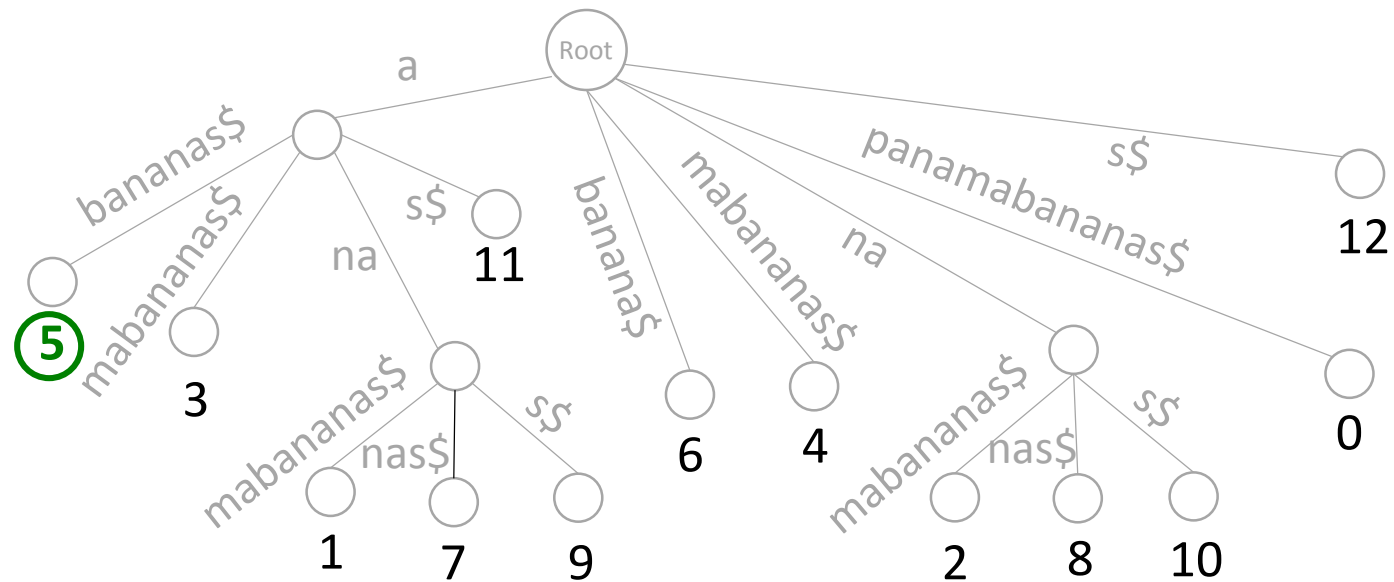
13	\$ <sub>1</sub> panamabananas <sub>1</sub>
5	a <sub>1</sub> bananas\$panam <sub>1</sub>
3	a <sub>2</sub> mabananas\$pan <sub>1</sub>
<b>1</b>	<b>a<sub>3</sub>na</b> mabananas\$p <sub>1</sub>
<b>7</b>	<b>a<sub>4</sub>na</b> nas\$panamab <sub>1</sub>
<b>9</b>	<b>a<sub>5</sub>na</b> s\$panamaban <sub>2</sub>
11	a <sub>6</sub> s\$panamaban <sub>3</sub>
6	b <sub>1</sub> ananas\$panama <sub>1</sub>
4	m <sub>1</sub> abananas\$pana <sub>2</sub>
2	n <sub>1</sub> amabananas\$pa <sub>3</sub>
8	n <sub>2</sub> anas\$panamaba <sub>4</sub>
10	n <sub>3</sub> as\$panamabana <sub>5</sub>
0	p <sub>1</sub> anamabananas\$ <sub>1</sub>
12	s <sub>1</sub> \$panamabanana <sub>6</sub>

**Naïve algorithm for constructing suffix array (sorting all suffixes of *Text*)**

$O(|Text| \cdot \log |Text|)$  comparisons

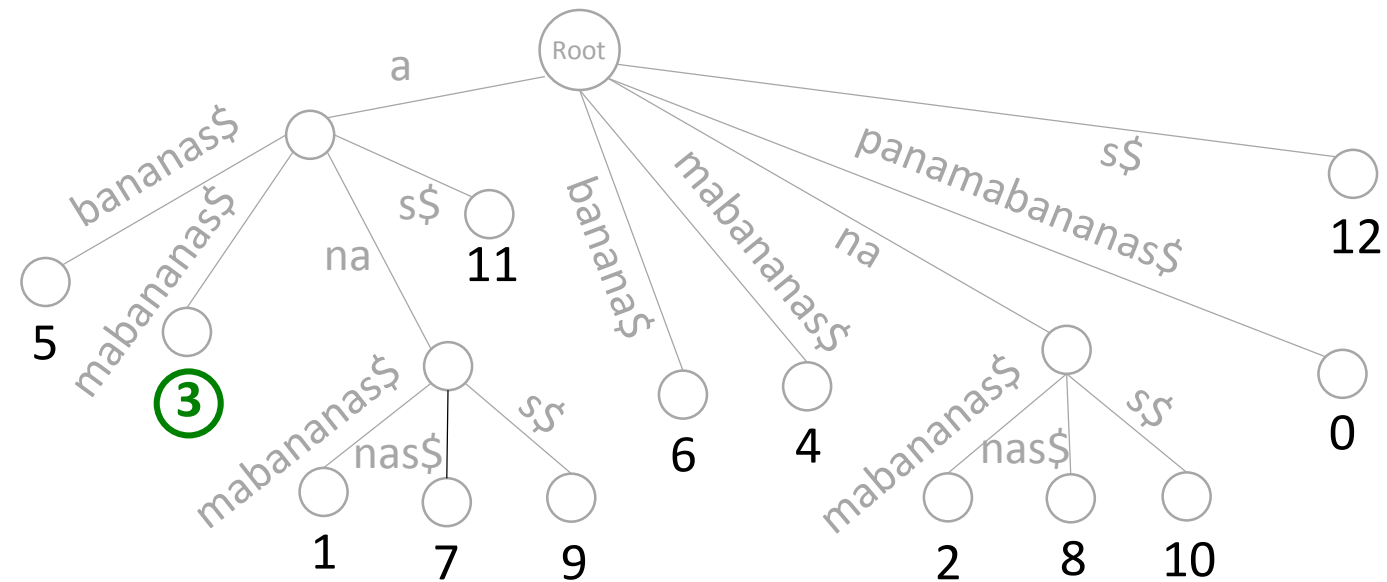


# From Suffix Tree to Suffix Array: Depth-First Traversal



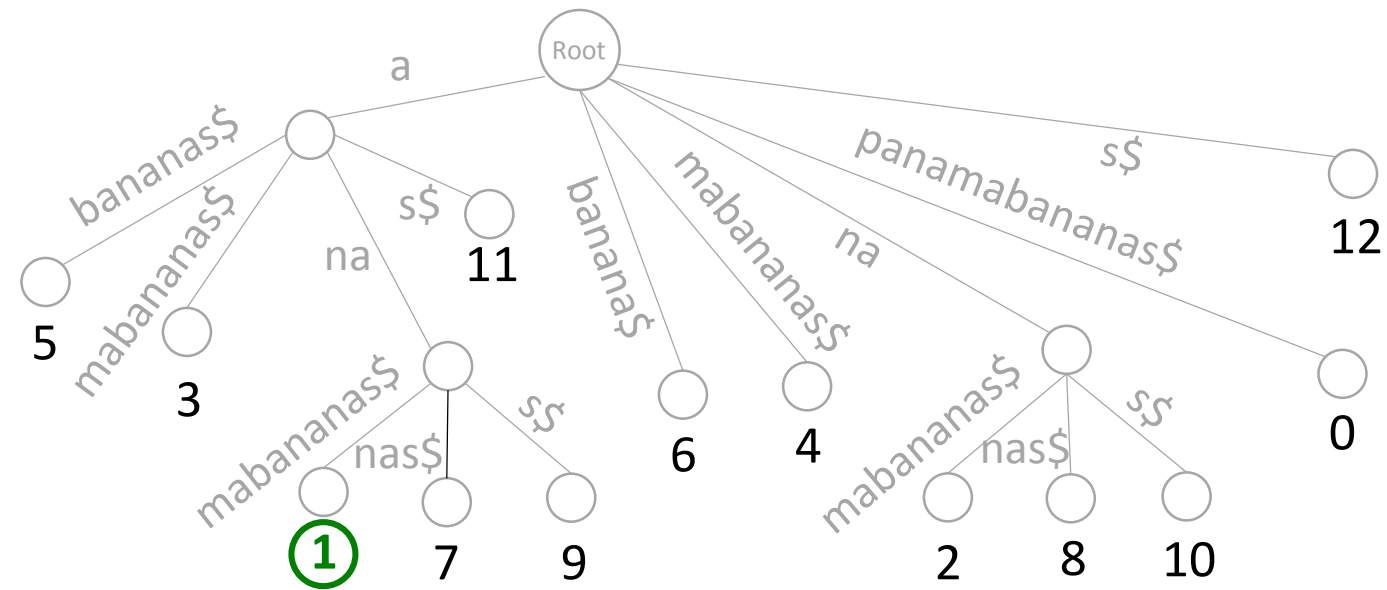
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

# From Suffix Tree to Suffix Array: Depth-First Traversal



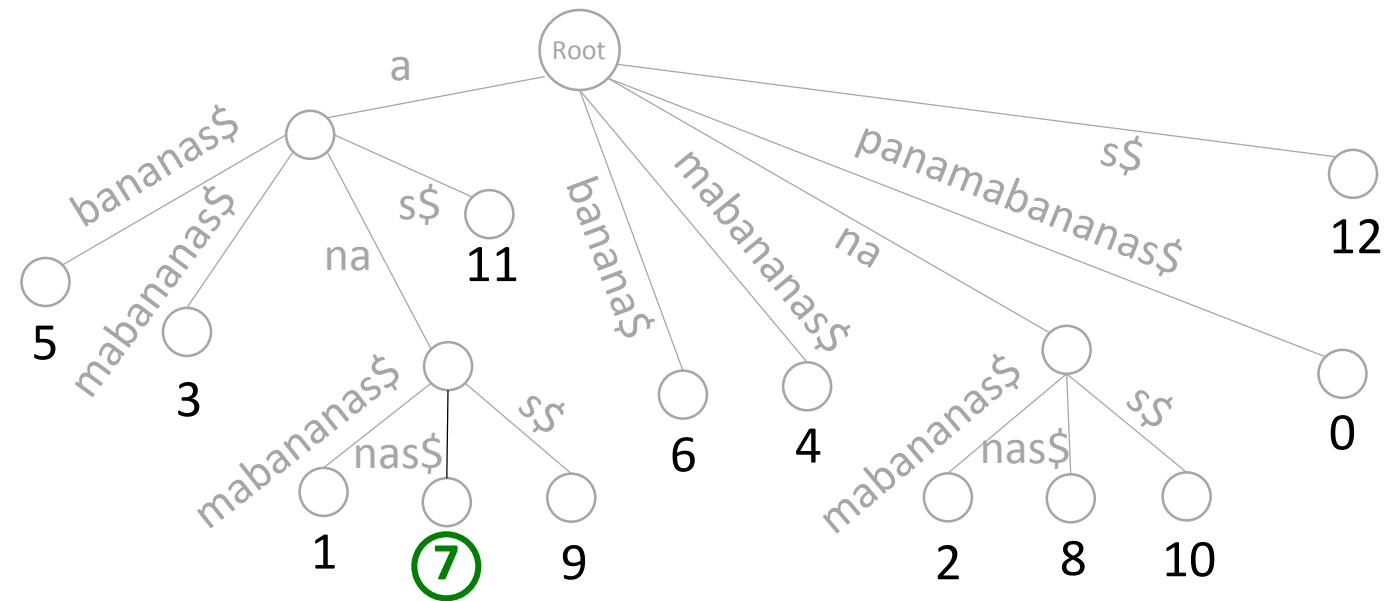
[13 5 **3** 1 7 9 11 6 4 2 8 10 0 12]

# From Suffix Tree to Suffix Array: Depth-First Traversal



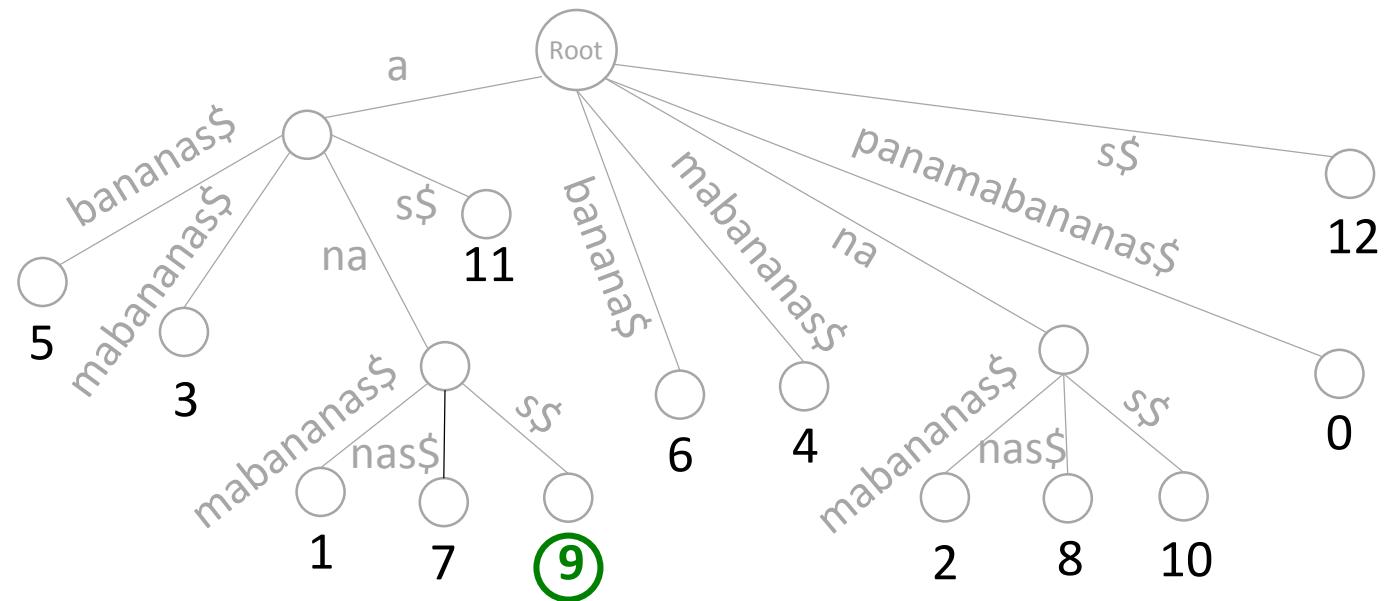
[13 5 3 **1** 7 9 11 6 4 2 8 10 0 12]

# From Suffix Tree to Suffix Array



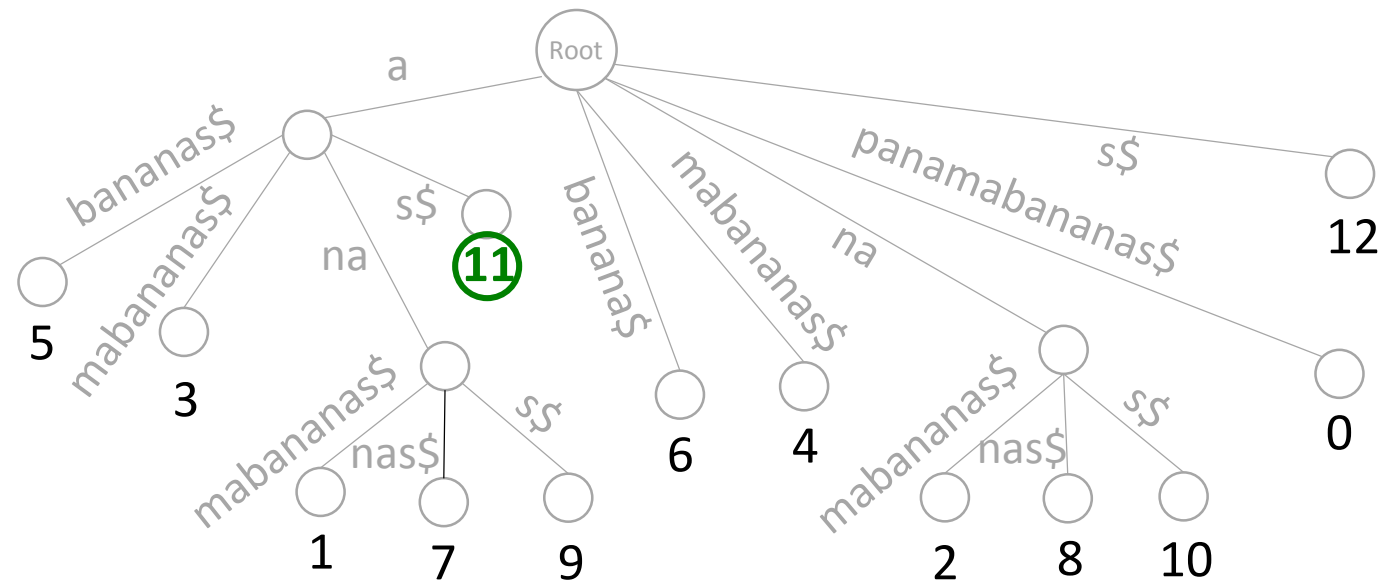
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

# From Suffix Tree to Suffix Array



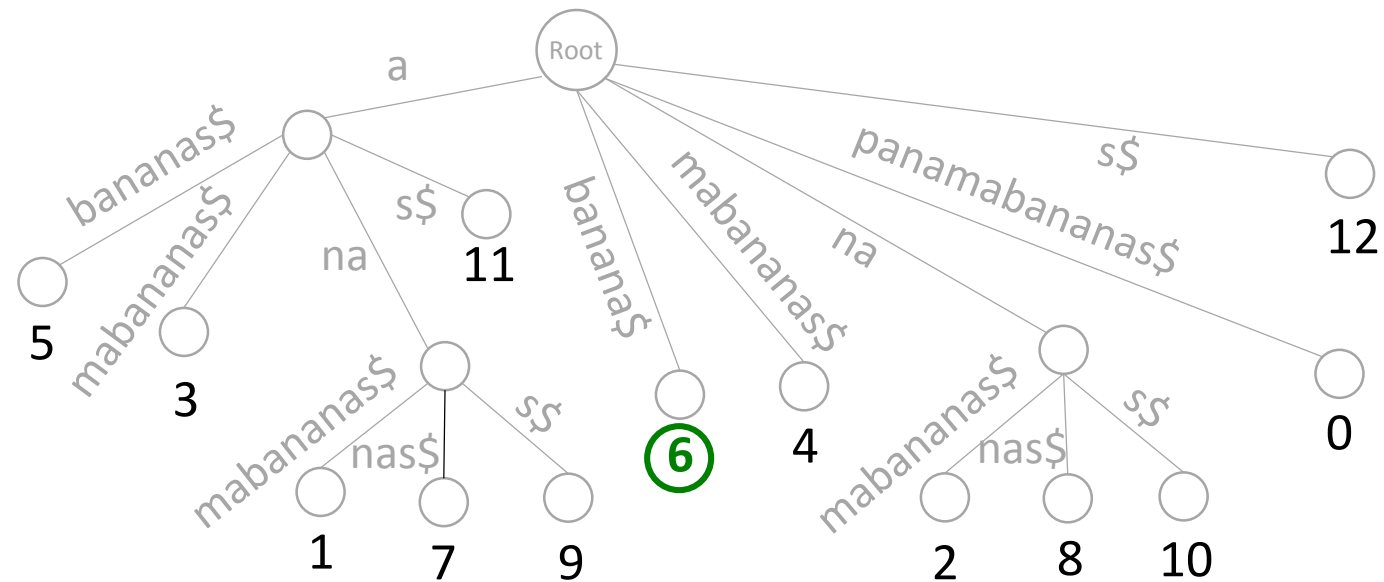
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

# From Suffix Tree to Suffix Array



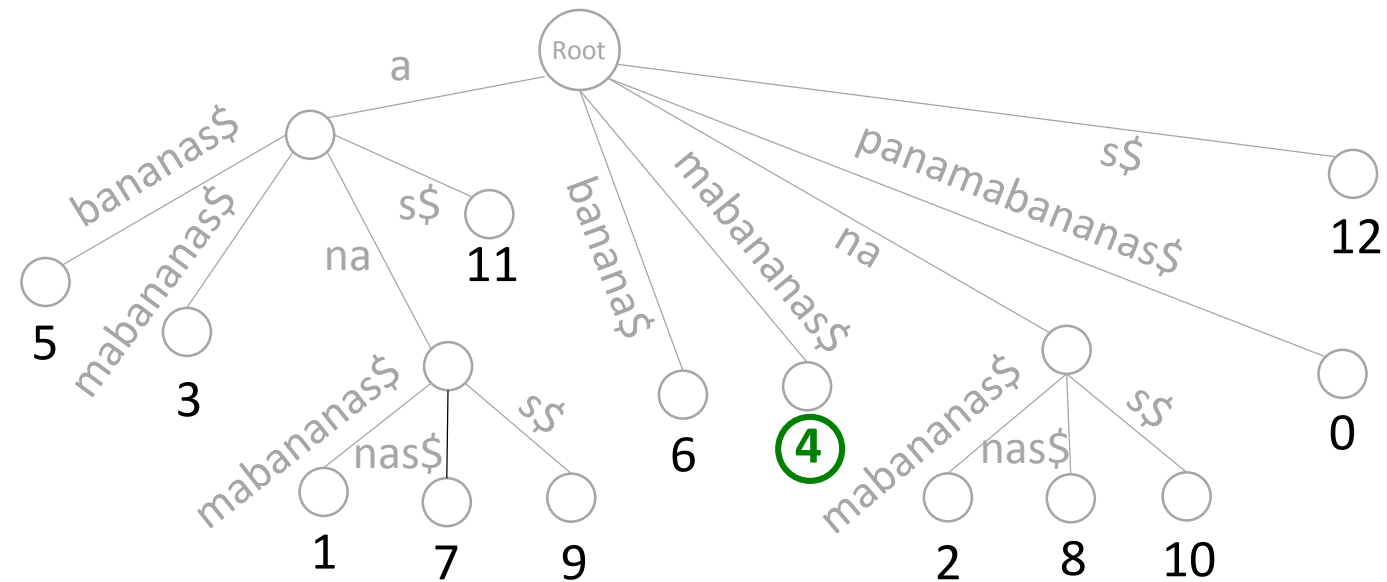
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

# From Suffix Tree to Suffix Array



[13 5 3 1 7 9 11 **6** 4 2 8 10 0 12]

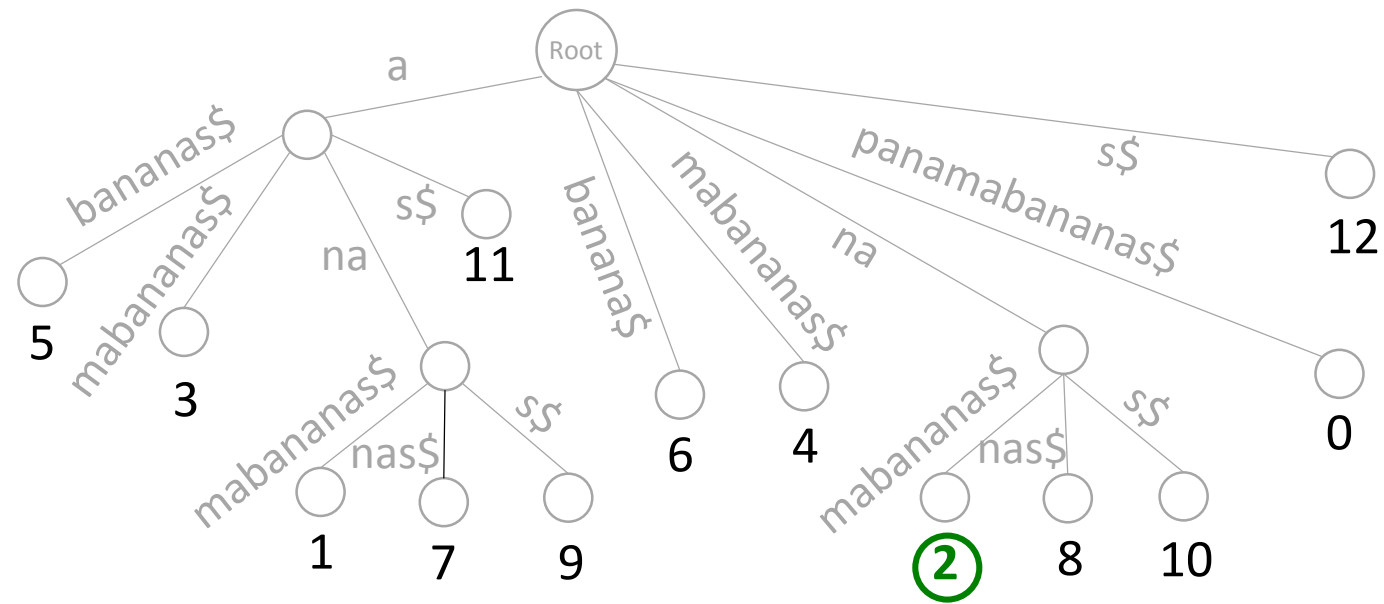
# From Suffix Tree to Suffix Array



[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

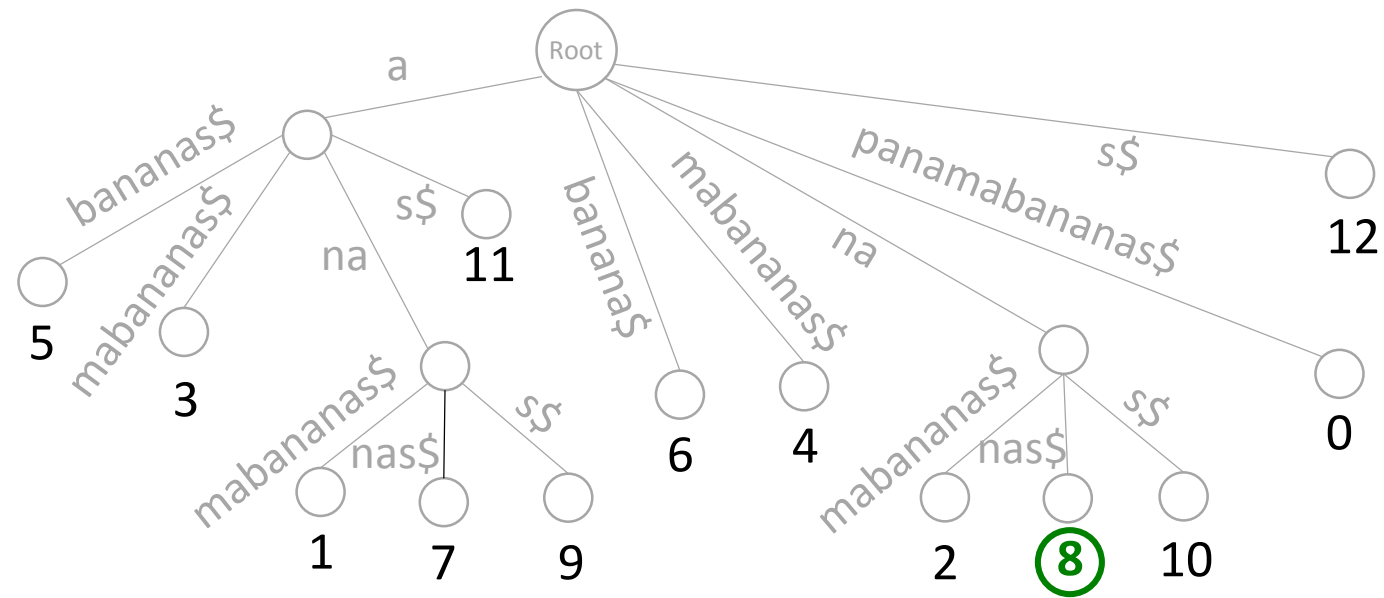


# From Suffix Tree to Suffix Array



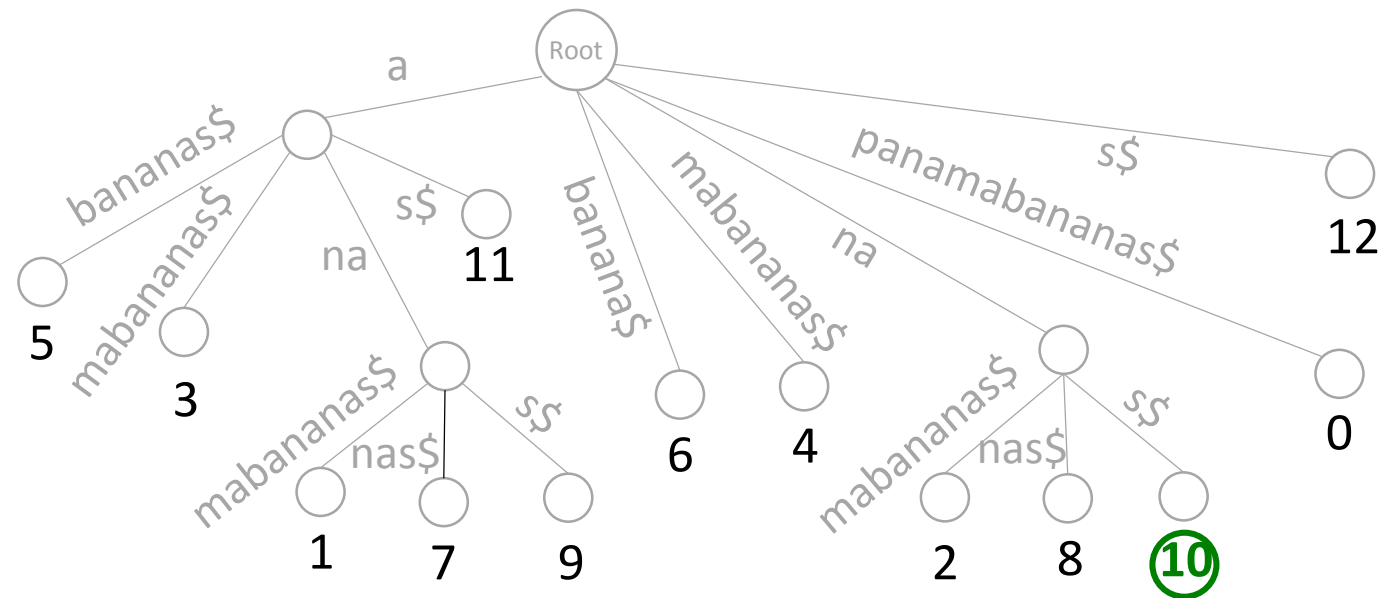
[13 5 3 1 7 9 11 6 4 **2** 8 10 0 12]

# From Suffix Tree to Suffix Array



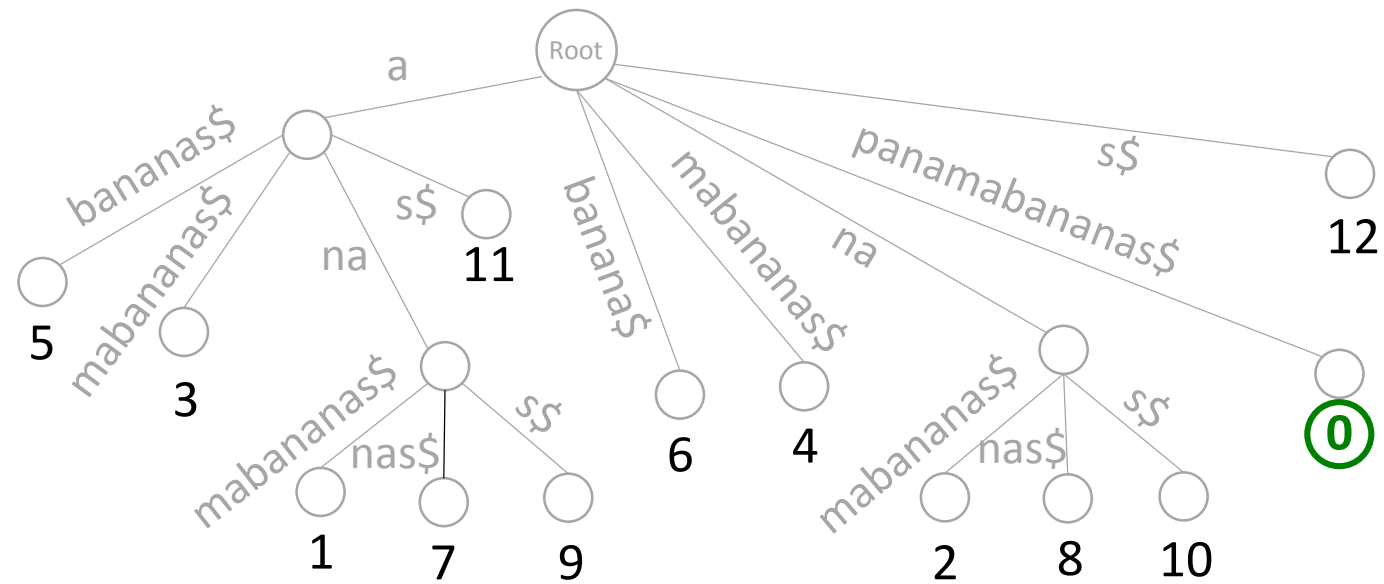
[13 5 3 1 7 9 11 6 4 2 **8** 10 0 12]

# From Suffix Tree to Suffix Array



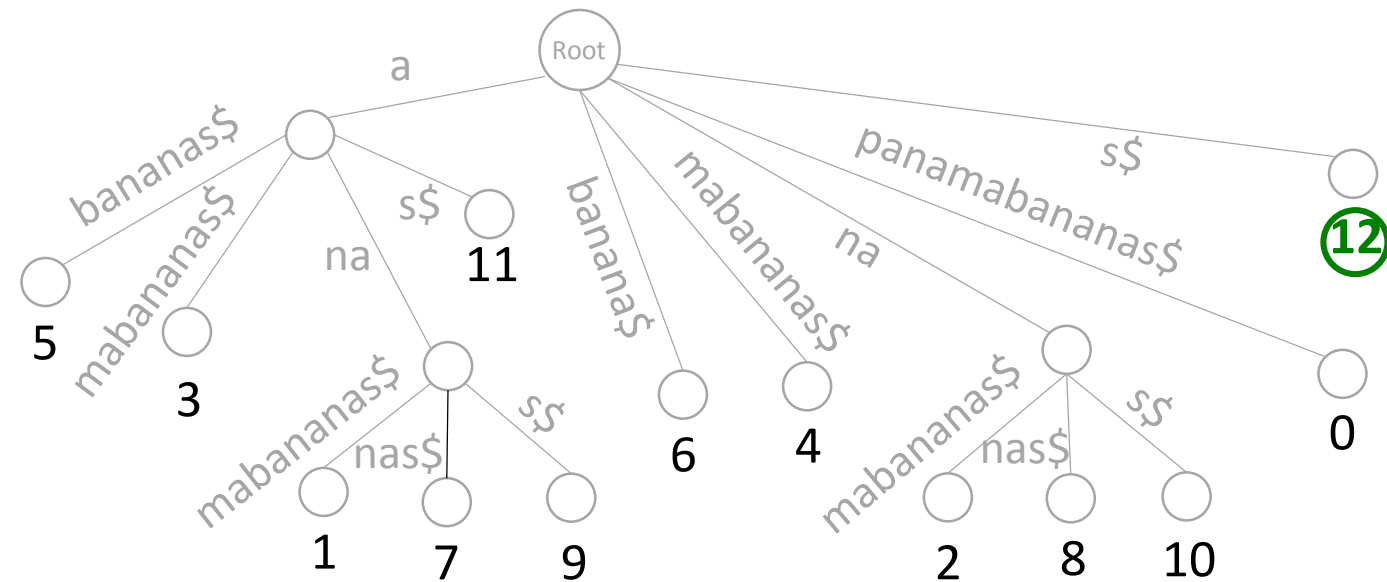
[13 5 3 1 7 9 11 6 4 2 8 **10** 0 12]

# From Suffix Tree to Suffix Array



[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

# From Suffix Tree to Suffix Array

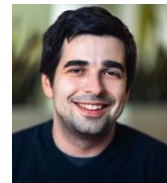


[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

# Constructing Suffix Array

- Depth-first traversal of suffix tree
  - $O(|Text|)$  time and  $\sim 20 \cdot |Text|$  space
- Manber-Myers algorithm (1990):
  - $O(|Text|)$  time and  $\sim 4 \cdot |Text|$  space
- But memory footprint is still large for human genome!

We will learn how to quickly construct suffix array  
without relying on suffix tree later in this course



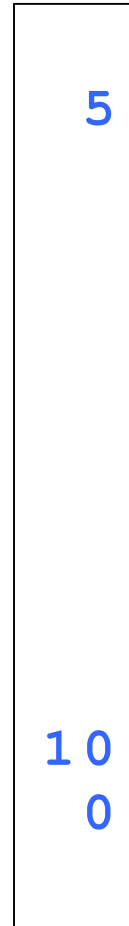
# Reducing Memory Footprint for Suffix Array

- Can we store only a fraction of the suffix array but still do fast pattern matching?

1	3
	5
	3
	1
	7
	9
1	1
	6
	4
	2
	8
1	0
	0
1	2

# Reducing Memory Footprint for Suffix Array

- Can we store only a fraction of the suffix array but still do fast pattern matching?
- Partial suffix array  $\text{SuffixArray}_K(\text{Text})$  only contains values that are multiples of some integer  $K$





# Using the Suffix Array to Find Matches

	suffix array
$s_1$ panamabananas $s_1$	1 3
$a_1$ bananas\$panam $s_1$	5
$a_2$ mabananas\$pan $s_1$	3
<b><math>a_3</math> na</b> mabananas\$pan $s_1$	<b>1</b>
<b><math>a_4</math> na</b> nas\$panamab $s_1$	<b>7</b>
<b><math>a_5</math> na</b> s\$panamaban $s_2$	<b>9</b>
$a_6$ s\$panamaban $s_3$	1 1
$b_1$ ananas\$panama $s_1$	6
$m_1$ abananas\$pana $s_2$	4
$n_1$ amabananas\$pa $s_3$	2
$n_2$ anas\$panamaba $s_4$	8
$n_3$ as\$panamabana $s_5$	1 0
$p_1$ anamabananas\$ $s_1$	0
$s_1$ \$panamabana $s_6$	1 2

# Using the Partial Suffix Array to Find Matches

**partial  
suffix  
array**

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a<sub>3</sub>na**mabananas\$p<sub>1</sub>  
**a<sub>4</sub>na**nas\$panamab<sub>1</sub>  
**a<sub>5</sub>na**s\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

**5**

**10**

**0**

# Using the Partial Suffix Array to Find Matches

partial  
suffix  
array

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a<sub>3</sub>na**mabananas\$p<sub>1</sub>  
**a<sub>4</sub>na**nas\$panamab<sub>1</sub>  
**a<sub>5</sub>na**s\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

5

Where are these **ana** prefixes located in *Text*???

10

0

# Focus on **a<sub>4</sub>na**

partial  
suffix  
array

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a<sub>3</sub>**namabananas\$p<sub>1</sub>  
**a<sub>4</sub>na**nas\$panama**b<sub>1</sub>**  
**a<sub>5</sub>na**s\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

Where is **a<sub>4</sub>na**?

5

10

0

# Focus on **b<sub>1</sub>ana**

partial  
suffix  
array

\$<sub>1</sub>panamabananas<sub>1</sub>  
a<sub>1</sub>bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a**<sub>3</sub>namabananas\$p<sub>1</sub>  
**a**<sub>4</sub>**n**anas\$panama**b**<sub>1</sub>  
**a**<sub>5</sub>**n**a\$panamaban<sub>2</sub>  
a<sub>6</sub>s\$panamaban<sub>3</sub>  
**b**<sub>1</sub>**a****n**aanas\$panama**a**<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

5

Where is **b<sub>1</sub>ana**?

10

0

# Focus on $a_1bana$

partial  
suffix  
array

\$<sub>1</sub>panamabananas<sub>1</sub>  
 $a_1$  $b$  $a$  $n$  $a$  $n$  $a$  $s$ \$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
 $a_3$ namabananas\$p<sub>1</sub>  
 $a_4$  $n$  $a$  $n$  $a$  $s$ \$panama $b_1$   
 $a_5$  $n$  $a$  $s$ \$panamaban<sub>2</sub>  
a<sub>6</sub> $s$ \$panamaban<sub>3</sub>  
 $b_1$  $a$  $n$  $a$  $n$  $a$  $s$ \$panam $a_1$   
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabanana<sub>6</sub>

Where is  $a_1bana$ ?

5

10

0

# Partial suffix array reveals position of **a<sub>1</sub>bana**

\$<sub>1</sub> panamabananas<sub>1</sub>  
**a<sub>1</sub>**bananas\$panam<sub>1</sub>  
a<sub>2</sub>mabananas\$pan<sub>1</sub>  
**a<sub>3</sub>**namabananas\$p<sub>1</sub>  
**a<sub>4</sub>**nanas\$panama**b<sub>1</sub>**  
**a<sub>5</sub>**nas\$panamaban<sub>2</sub>  
a<sub>6</sub>\$spanamaban<sub>3</sub>  
**b<sub>1</sub>**ananas\$panam**a<sub>1</sub>**  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabana<sub>6</sub>

**a<sub>1</sub>bana** is at position 5

**a<sub>4</sub>na** is at position 7

**b<sub>1</sub>ana** is at position 6

partial  
suffix  
array

5

7

6

1 0

0

# Outline

- Burrows-Wheeler Transform
- Inverting Burrows-Wheeler Transform
- Using BWT for Pattern Matching
- Suffix Arrays
- **Approximate Pattern Matching**



# Returning to Search for Mutations

- **Approximate Pattern Matching Problem:**
  - **Input:** A string *Pattern*, a string *Text*, and an integer  $d$ .
  - **Output:** All positions in *Text* where the string *Pattern* appears as a substring with at most  $d$  mismatches.

# Revealing Mutations by Analyzing **Billions** of Reads

- **Multiple Approximate Pattern Matching Problem**
  - **Input:** A **set** of strings *Patterns*, a string *Text*, and an integer  $d$ .
  - **Output:** All positions in *Text* where a string from *Patterns* appears as a substring with at most  $d$  mismatches.

# BWT Saves the Day Again

- searching for ana in panamabananas

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```

# BWT Saves the Day Again

- searching for an **a** in panamabananas

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```

# BWT Saves the Day Again

- searching for **ana** in panamabananas

\$<sub>1</sub>panamabananas<sub>1</sub>  
**a**<sub>1</sub>bananas\$pana**m**<sub>1</sub>  
**a**<sub>2</sub>mabananas\$pa**n**<sub>1</sub>  
**a**<sub>3</sub>namabananas\$**p**<sub>1</sub>  
**a**<sub>4</sub>nanas\$panama**b**<sub>1</sub>  
**a**<sub>5</sub>nas\$panamaba**n**<sub>2</sub>  
**a**<sub>6</sub>s\$panamabana**n**<sub>3</sub>  
b<sub>1</sub>ananas\$panama<sub>1</sub>  
m<sub>1</sub>abananas\$pana<sub>2</sub>  
n<sub>1</sub>amabananas\$pa<sub>3</sub>  
n<sub>2</sub>anas\$panamaba<sub>4</sub>  
n<sub>3</sub>as\$panamabana<sub>5</sub>  
p<sub>1</sub>anamabananas\$<sub>1</sub>  
s<sub>1</sub>\$panamabana<sub>6</sub>

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.

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamabanan3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanaa6
```

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
\$ <sub>1</sub> panamabananas <sub>1</sub>	
<b>a</b> <sub>1</sub> bananas\$pana <b>m</b> <sub>1</sub>	1
<b>a</b> <sub>2</sub> mabananas\$pa <b>n</b> <sub>1</sub>	0
<b>a</b> <sub>3</sub> namabananas\$ <b>p</b> <sub>1</sub>	1
<b>a</b> <sub>4</sub> nanas\$panama <b>b</b> <sub>1</sub>	1
<b>a</b> <sub>5</sub> nas\$panamaba <b>n</b> <sub>2</sub>	0
<b>a</b> <sub>6</sub> s\$panamabana <b>n</b> <sub>3</sub>	0
b <sub>1</sub> ananas\$panama <sub>1</sub>	
m <sub>1</sub> abananas\$pana <sub>2</sub>	
n <sub>1</sub> amabananas\$pa <sub>3</sub>	
n <sub>2</sub> anas\$panamaba <sub>4</sub>	
n <sub>3</sub> as\$panamabana <sub>5</sub>	
p <sub>1</sub> anamabananas\$ <sub>1</sub>	
s <sub>1</sub> \$panamabanana <sub>6</sub>	

# 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.

	# Mismatches
\$ <sub>1</sub> panamabananas <sub>1</sub>	
<b>a</b> <sub>1</sub> bananas\$pana <b>m</b> <sub>1</sub>	1
<b>a</b> <sub>2</sub> mabananas\$pa <b>n</b> <sub>1</sub>	0
<b>a</b> <sub>3</sub> namabananas\$ <b>p</b> <sub>1</sub>	1
<b>a</b> <sub>4</sub> nanas\$panama <b>b</b> <sub>1</sub>	1
<b>a</b> <sub>5</sub> nas\$panamaba <b>n</b> <sub>2</sub>	0
<b>a</b> <sub>6</sub> s\$panamabana <b>n</b> <sub>3</sub>	0
b <sub>1</sub> ananas\$panama <sub>1</sub>	
m <sub>1</sub> abananas\$pana <sub>2</sub>	
n <sub>1</sub> amabananas\$pa <sub>3</sub>	
n <sub>2</sub> anas\$panamaba <sub>4</sub>	
n <sub>3</sub> as\$panamabana <sub>5</sub>	
p <sub>1</sub> anamabananas\$ <sub>1</sub>	
s <sub>1</sub> \$panamabanana <sub>6</sub>	



# 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.

	# Mismatches
\$ <sub>1</sub> panamabananas <sub>1</sub>	
a <sub>1</sub> bananas\$pana <b>m</b> <sub>1</sub>	1
a <sub>2</sub> mabananas\$pa <b>n</b> <sub>1</sub>	0
a <sub>3</sub> namabananas\$ <b>p</b> <sub>1</sub>	1
a <sub>4</sub> nanas\$panama <b>b</b> <sub>1</sub>	1
a <sub>5</sub> nas\$panamaba <b>n</b> <sub>2</sub>	0
a <sub>6</sub> s\$panamabana <b>n</b> <sub>3</sub>	0
<b>b</b> <sub>1</sub> <b>a</b> nanas\$panama <sub>1</sub>	
<b>m</b> <sub>1</sub> <b>a</b> bananas\$pana <sub>2</sub>	
<b>n</b> <sub>1</sub> <b>a</b> mabananas\$pa <sub>3</sub>	
<b>n</b> <sub>2</sub> <b>a</b> nas\$panamaba <sub>4</sub>	
<b>n</b> <sub>3</sub> <b>a</b> s\$panamabana <sub>5</sub>	
<b>p</b> <sub>1</sub> <b>a</b> namabananas\$ <sub>1</sub>	
s <sub>1</sub> \$panamabana <sub>6</sub>	

# 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.

	# Mismatches
\$ <sub>1</sub> panamabananas <sub>1</sub>	
a <sub>1</sub> bananas\$panam <sub>1</sub>	
a <sub>2</sub> mabananas\$pan <sub>1</sub>	
a <sub>3</sub> namabananas\$p <sub>1</sub>	
a <sub>4</sub> nanas\$panamab <sub>1</sub>	
a <sub>5</sub> nas\$panamaban <sub>2</sub>	
a <sub>6</sub> s\$panamaban <sub>3</sub>	
<b>b</b> <sub>1</sub> <b>a</b> nanas\$panama <sub>1</sub>	<b>1</b>
<b>m</b> <sub>1</sub> <b>a</b> bananas\$pana <sub>2</sub>	<b>1</b>
<b>n</b> <sub>1</sub> <b>a</b> mabananas\$pa <sub>3</sub>	0
<b>n</b> <sub>2</sub> <b>a</b> nas\$panamaba <sub>4</sub>	0
<b>n</b> <sub>3</sub> <b>a</b> s\$panamabana <sub>5</sub>	0
<b>p</b> <sub>1</sub> <b>a</b> namabananas\$ <sub>1</sub>	<b>1</b>
s <sub>1</sub> \$panamabanana <sub>6</sub>	

# BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

	# Mismatches
\$ <sub>1</sub> panamabananas <sub>1</sub>	
a <sub>1</sub> bananas\$panam <sub>1</sub>	
a <sub>2</sub> mabananas\$pan <sub>1</sub>	
a <sub>3</sub> namabananas\$p <sub>1</sub>	
a <sub>4</sub> nanas\$panamab <sub>1</sub>	
a <sub>5</sub> nas\$panamaban <sub>2</sub>	
a <sub>6</sub> s\$panamaban <sub>3</sub>	
<b>b</b> <sub>1</sub> <b>a</b> nanas\$panam <b>a</b> <sub>1</sub>	<b>1</b>
<b>m</b> <sub>1</sub> <b>a</b> bananas\$pan <b>a</b> <sub>2</sub>	<b>1</b>
<b>n</b> <sub>1</sub> <b>a</b> mabananas\$p <b>a</b> <sub>3</sub>	0
<b>n</b> <sub>2</sub> <b>a</b> nas\$panamab <b>a</b> <sub>4</sub>	0
<b>n</b> <sub>3</sub> <b>a</b> s\$panamaban <b>a</b> <sub>5</sub>	0
<b>p</b> <sub>1</sub> <b>a</b> namabananas\$ <b>s</b> <sub>1</sub>	<b>2</b>
s <sub>1</sub> \$panamabanana <sub>6</sub>	

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

# Five Approximate Matches Found!

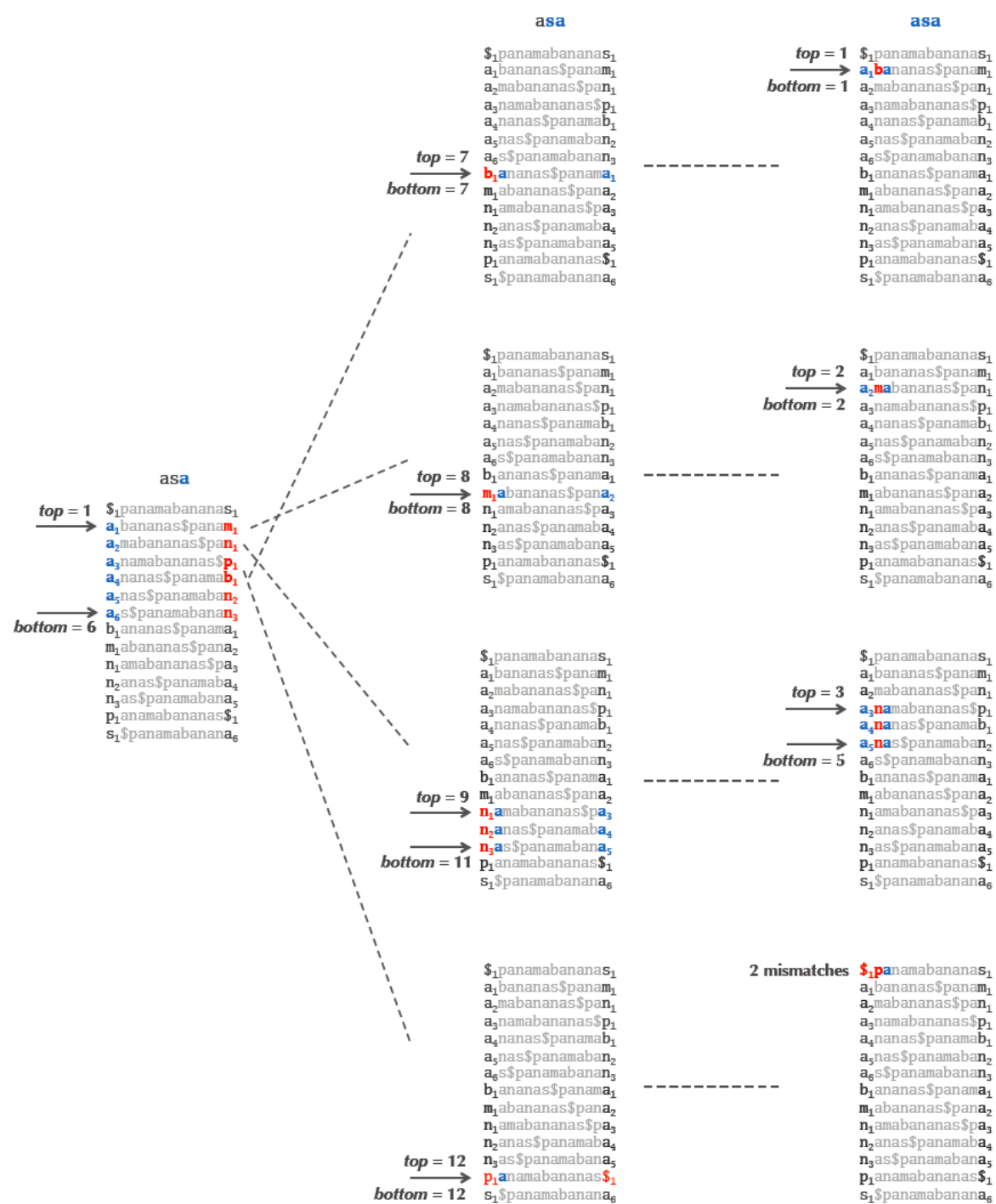
- searching for **ana** in panamabananas

	# Mismatches
\$ <sub>1</sub> panamabananas <sub>1</sub>	
<b>a</b> <sub>1</sub> <b>b</b> ananas\$panam <sub>1</sub>	1
<b>a</b> <sub>2</sub> <b>m</b> abananas\$pan <sub>1</sub>	1
<b>a</b> <sub>3</sub> <b>n</b> amabananas\$p <sub>1</sub>	0
<b>a</b> <sub>4</sub> <b>n</b> anas\$panamab <sub>1</sub>	0
<b>a</b> <sub>5</sub> <b>n</b> as\$panamaban <sub>2</sub>	0
a <sub>6</sub> s\$panamaban <sub>3</sub>	
b <sub>1</sub> ananas\$panam <b>a</b> <sub>1</sub>	
m <sub>1</sub> abananas\$pan <b>a</b> <sub>2</sub>	
n <sub>1</sub> amabananas\$p <b>a</b> <sub>3</sub>	
n <sub>2</sub> anas\$panamab <b>a</b> <sub>4</sub>	
n <sub>3</sub> as\$panamaban <b>a</b> <sub>5</sub>	
p <sub>1</sub> anamabananas\$ <b>s</b> <sub>1</sub>	
s <sub>1</sub> \$panamabanana <sub>6</sub>	

# Where Are The Matches?

- searching for **ana** in panamabananas

Suffix Array	
\$ <sub>1</sub> panamabananas <sub>1</sub>	
<b>a</b> <sub>1</sub> <b>b</b> ananas\$panam <sub>1</sub>	5
<b>a</b> <sub>2</sub> <b>m</b> abananas\$pan <sub>1</sub>	3
<b>a</b> <sub>3</sub> <b>n</b> amabananas\$p <sub>1</sub>	1
<b>a</b> <sub>4</sub> <b>n</b> anas\$panamab <sub>1</sub>	7
<b>a</b> <sub>5</sub> <b>n</b> as\$panamaban <sub>2</sub>	9
a <sub>6</sub> s\$panamaban <sub>3</sub>	
b <sub>1</sub> ananas\$panama <sub>1</sub>	
m <sub>1</sub> abananas\$pana <sub>2</sub>	
n <sub>1</sub> amabananas\$pa <sub>3</sub>	
n <sub>2</sub> anas\$panamaba <sub>4</sub>	
n <sub>3</sub> as\$panamabana <sub>5</sub>	
p <sub>1</sub> anamabananas\$ <sub>1</sub>	
s <sub>1</sub> \$panamabanana <sub>6</sub>	



In reality, approximate pattern matching with BWT is more complex (we omitted various details)

