

# Read Alignment

Michael Schatz

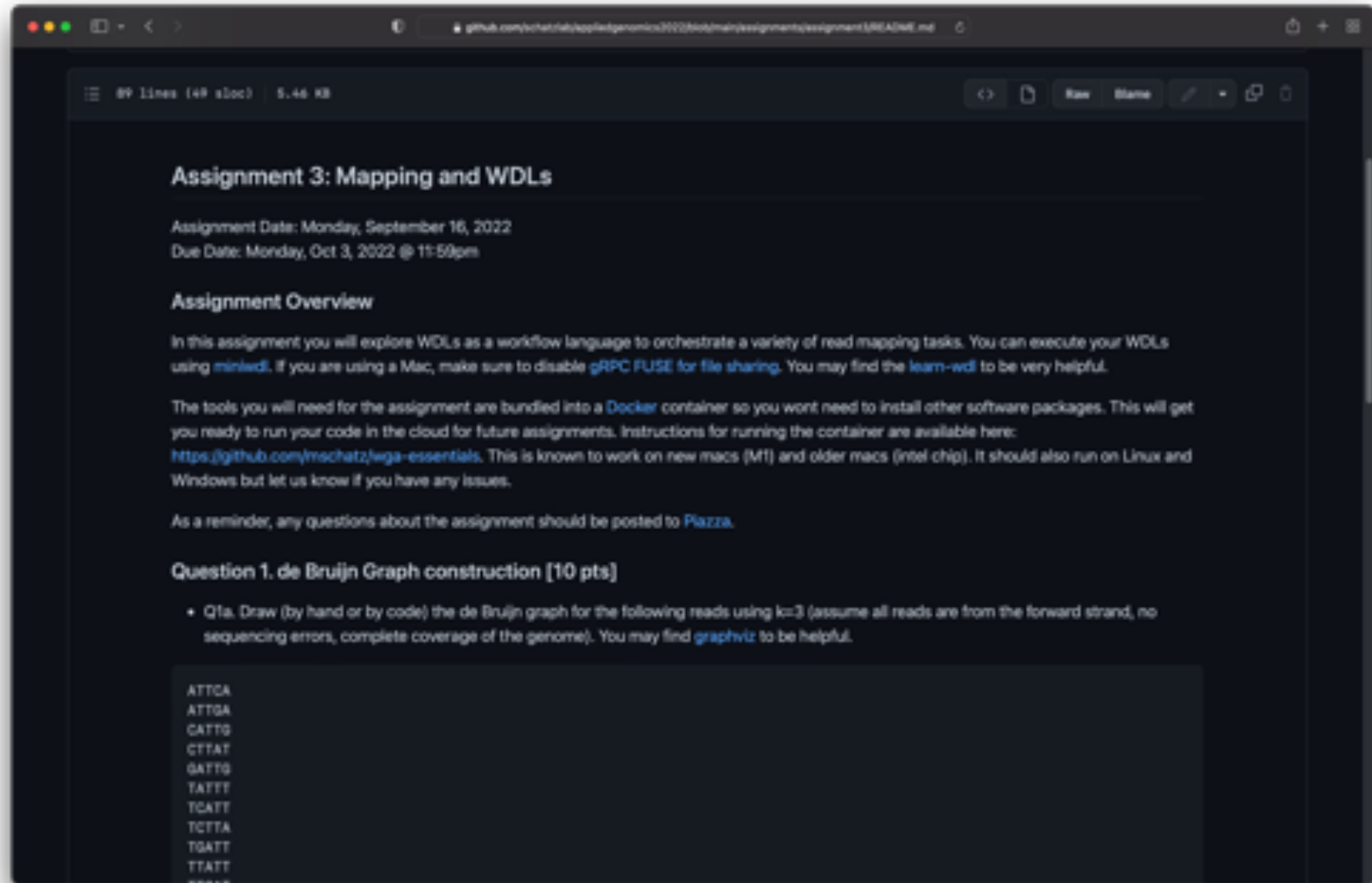
Sept 28, 2022

Lecture 9: Applied Comparative Genomics



# Assignment 3: Mapping and WDL

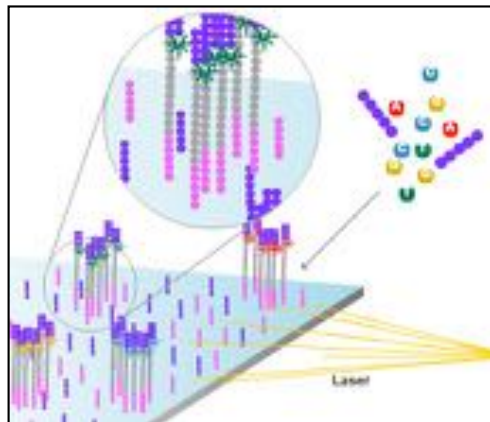
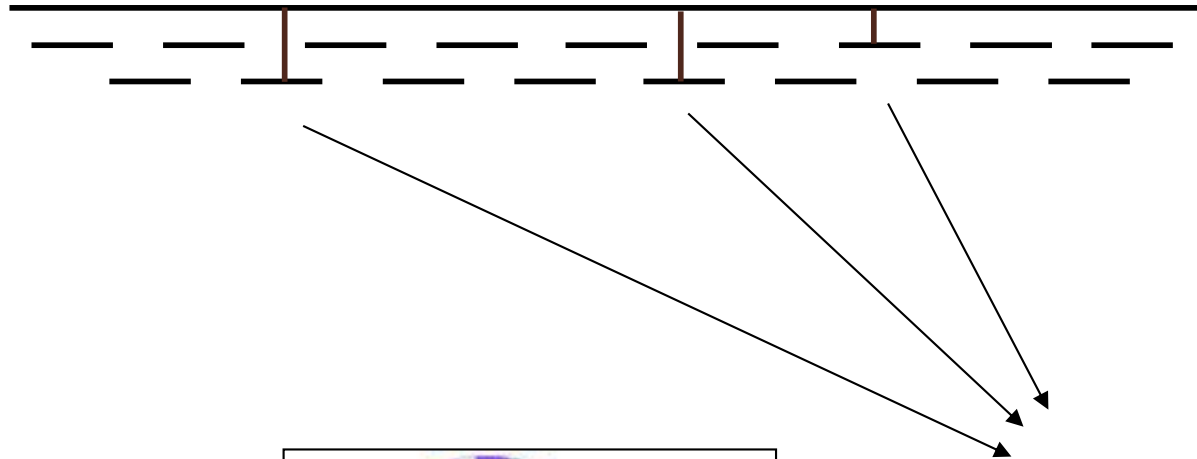
## Due Monday Oct 3 by 11:59pm



<https://github.com/schatzlab/appliedgenomics2022/tree/main/assignments/assignment3>  
Check Piazza for questions!

# Personal Genomics

How does your genome compare to the reference?



Heart Disease

Cancer

Presidential smile

# Brute Force Analysis

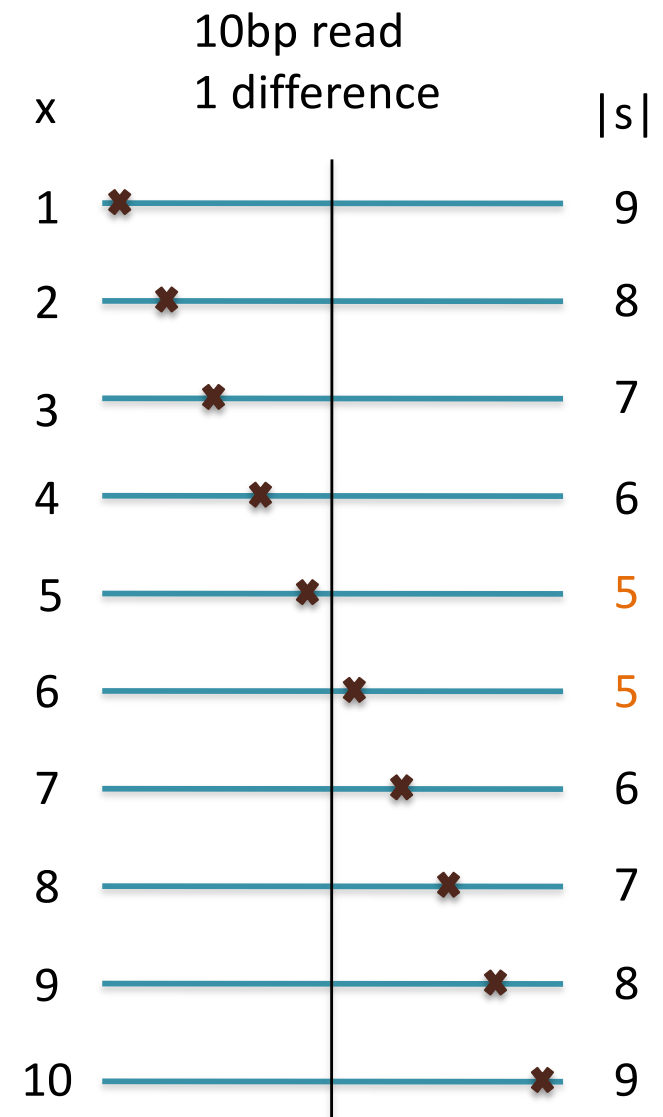


- Brute Force:
  - At every possible offset in the genome:
    - Do all of the characters of the query match?
- Analysis
  - Simple, easy to understand
  - Genome length =  $n$  [3B]
  - Query length =  $m$  [7]
  - Comparisons:  $(n-m+1) * m$  [21B]
- Overall runtime:  $O(nm)$ 
  - [How long would it take if we double the genome size, read length?]
  - [How long would it take if we double both?]

# Seed-and-Extend Alignment

Theorem: An alignment of a sequence of length  $m$  with at most  $k$  differences **must** contain an exact match at least  $s = m / (k + 1)$  bp long  
(Baeza-Yates and Perleberg, 1996)

- Proof: Pigeonhole principle
  - 1 pigeon can't fill 2 holes
- Seed-and-extend search
  - Use an index to rapidly find short exact alignments to seed longer in-exact alignments
    - BLAST, MUMmer, Bowtie, BWA, SOAP, ...
  - Specificity of the depends on seed length
    - Guaranteed sensitivity for  $k$  differences
    - Also finds some (but not all) lower quality alignments <- heuristic



# Searching the Index

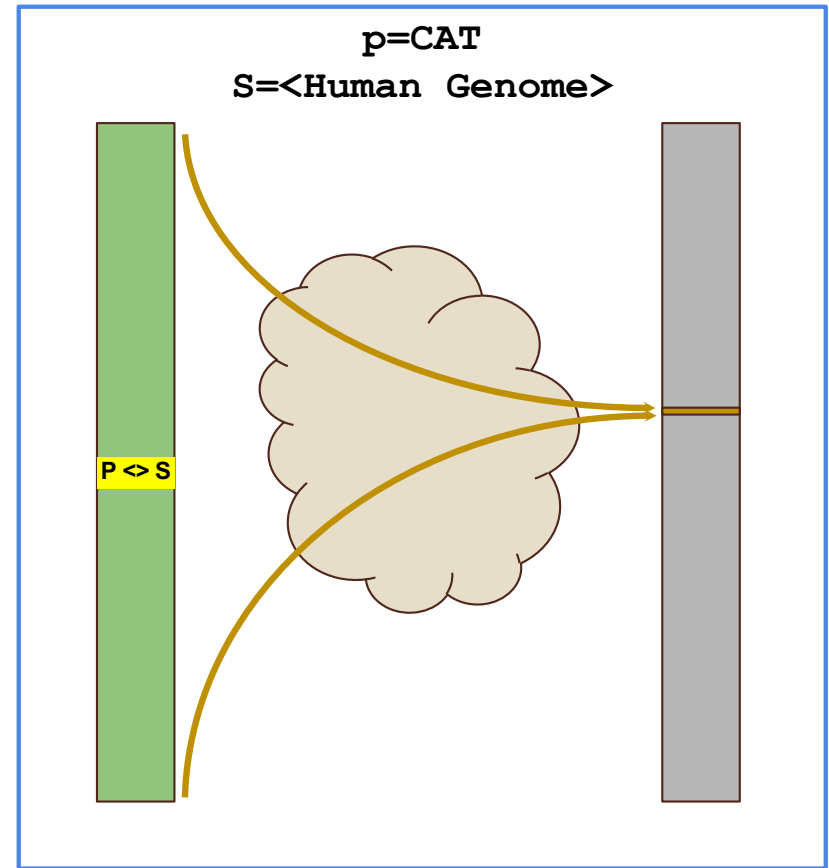
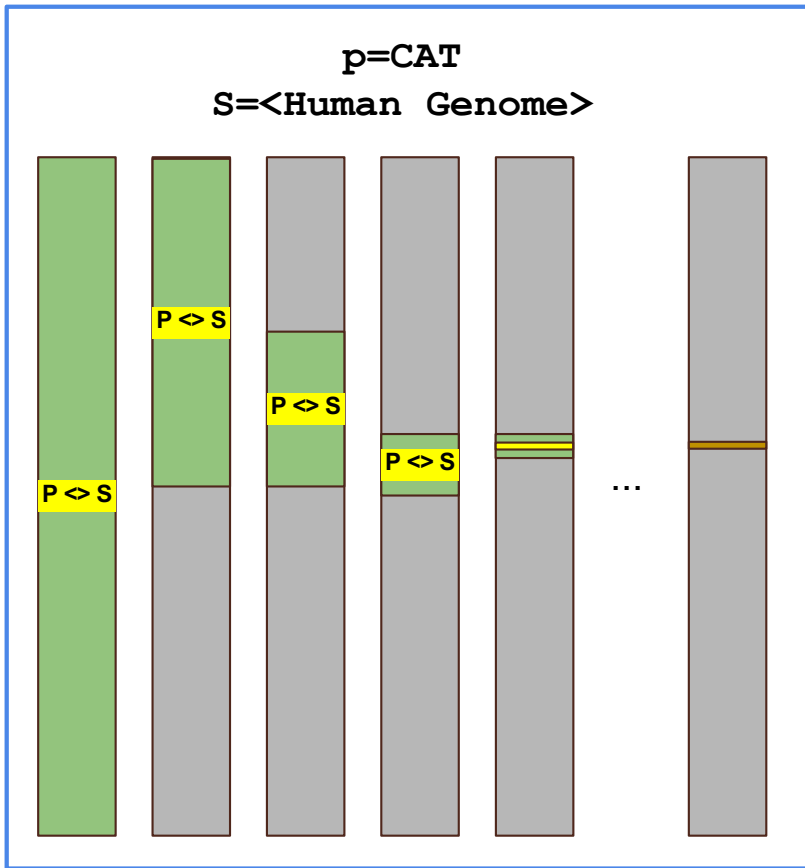
- Strategy 2: Binary search
  - Compare to the middle, refine as higher or lower
- Searching for GATTACA
  - $Lo = 1; Hi = 15; Mid = (1+15)/2 = 8$
  - Middle = Suffix[8] = CC  
=> Higher:  $Lo = Mid + 1$
  - $Lo = 9; Hi = 15; Mid = (9+15)/2 = 12$
  - Middle = Suffix[12] = TACC

#	Sequence	Pos
1	ACAGATTACC...	6
2	ACC...	13
3	AGATTACC...	8
4	ATTACAGATTACC...	3
5	ATTACC...	10
6	C...	15
7	CAGATTACC...	7
8	CC...	14
9	GATTACAGATTACC...	2
10	GATTACC...	9
11	TACAGATTACC...	5
12	TACC...	12
13	TGATTACAGATTACC...	1
14	TTACAGATTACC...	4
15	TTACC...	11

Lo  
→

Hi  
→

# Sapling: Accelerating Suffix Array Queries with Learned Data Models



***What if instead of a slow algorithmic approach to find the correct rows, we could somehow quickly guess/predict the correct rows?***

Kirsche, M, Das, A, Schatz, MC (2020) Bioinformatics  
doi: <https://doi.org/10.1093/bioinformatics/btaa911>



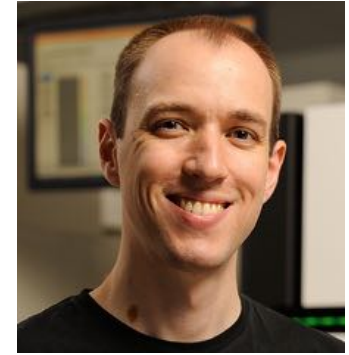
## Part 2: Burrows Wheeler Transform



# Algorithmic challenge

How can we combine the speed of a suffix array  $O(m + \lg(n))$  (or even  $O(m)$ ) with the size of a brute force analysis ( $n$  bytes)?

What would such an index look like?

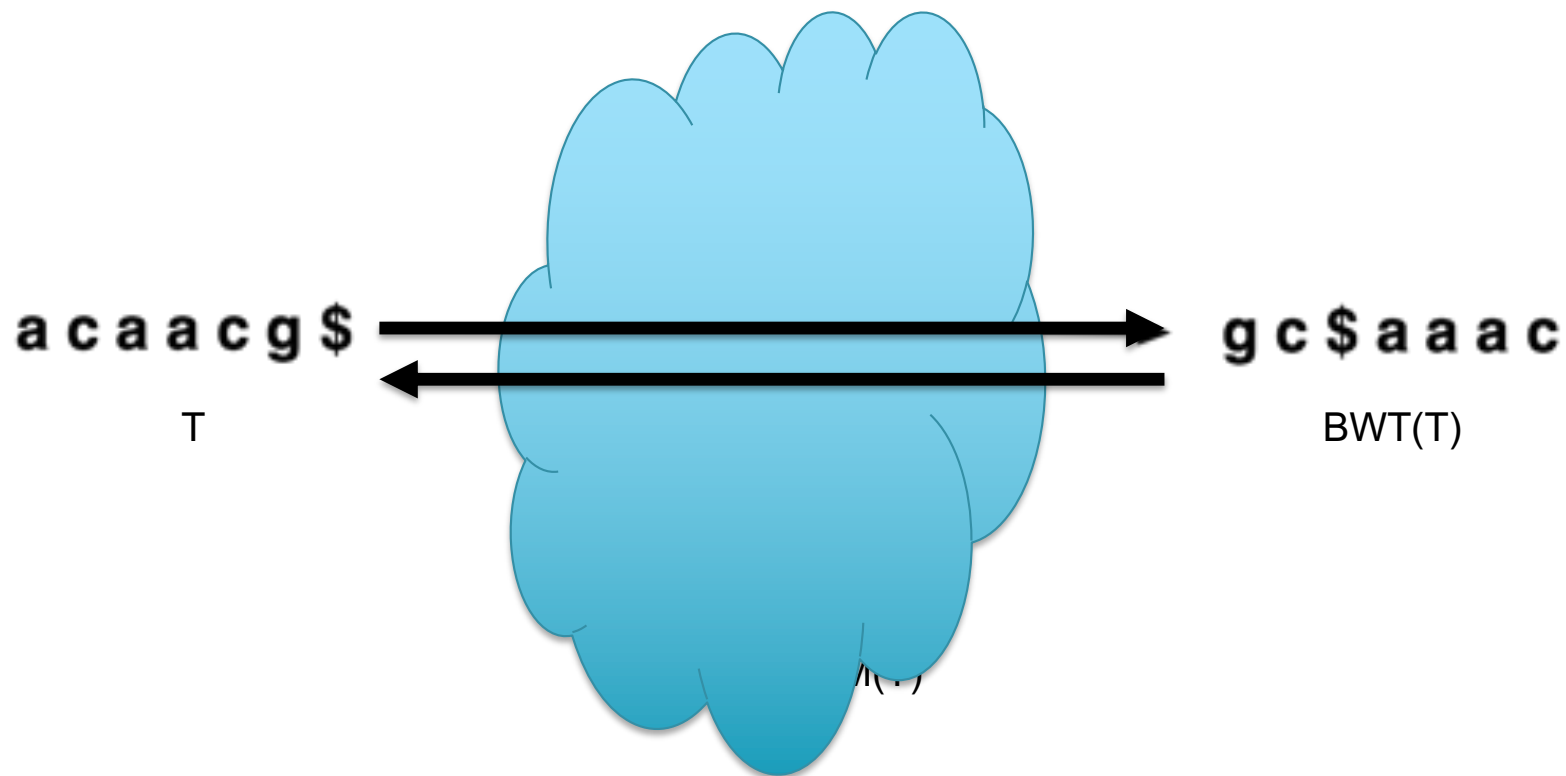


# Bowtie: Ultrafast and memory efficient alignment of short DNA sequences to the human genome

Slides Courtesy of Ben Langmead

# Burrows-Wheeler Transform

- Reversible permutation of the characters in a text



**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

# Burrows-Wheeler Transform

- Permutation of the characters in a text

a c a a c g \$ →

T

**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

# Burrows-Wheeler Transform

- Permutation of the characters in a text

a c a a c g \$

T

→

a c a a c g \$  
c a a c g \$ a  
a a c g \$ a c  
a c g \$ a c a  
c g \$ a c a a  
g \$ a c a a c  
\$ a c a a c g

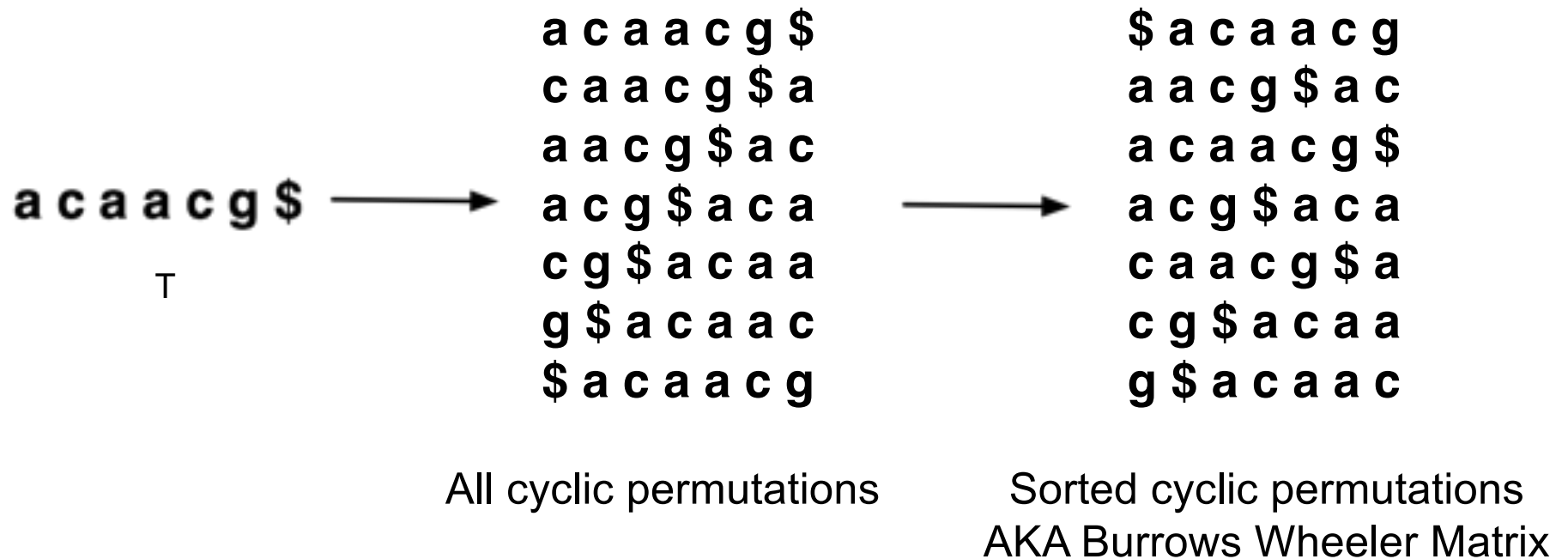
All cyclic permutations

**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

# Burrows-Wheeler Transform

- Permutation of the characters in a text



**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

# Burrows-Wheeler Transform

- Permutation of the characters in a text

`a c a a c g $`  $\longrightarrow$ 

\$	a	c	a	a	c	g
a	a	c	g	\$	a	c
a	c	a	a	c	g	\$
a	c	g	\$	a	c	a
c	a	a	c	g	\$	a
c	g	\$	a	c	a	a
g	\$	a	c	a	a	c

T

Sorted cyclic permutations  
AKA Burrows Wheeler Matrix

**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

# Burrows-Wheeler Transform

- Permutation of the characters in a text

$a\ c\ a\ a\ c\ g\ \$$   $\longrightarrow$

\$	a	c	a	a	c	g
a	a	c	g	\$	a	c
a	c	a	a	c	g	\$
a	c	g	\$	a	c	a
c	a	a	c	g	\$	a
c	g	\$	a	c	a	a
g	\$	a	c	a	a	c

Sorted cyclic permutations  
AKA Burrows Wheeler Matrix

Last Column = Burrows Wheeler Transform

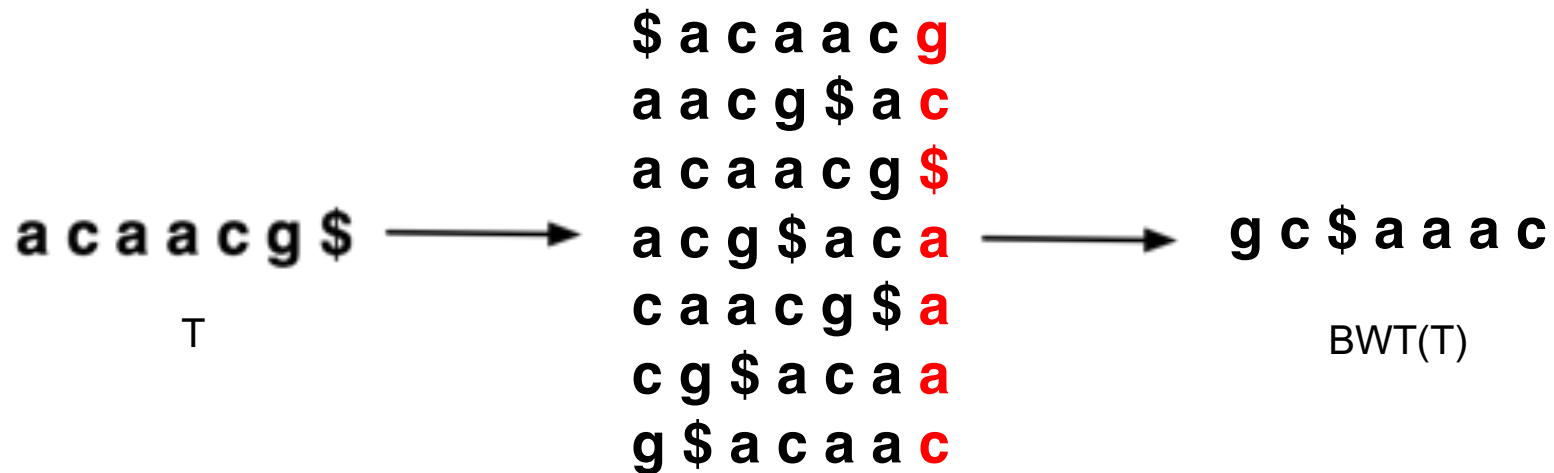
**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124



# Burrows-Wheeler Transform

- Permutation of the characters in a text



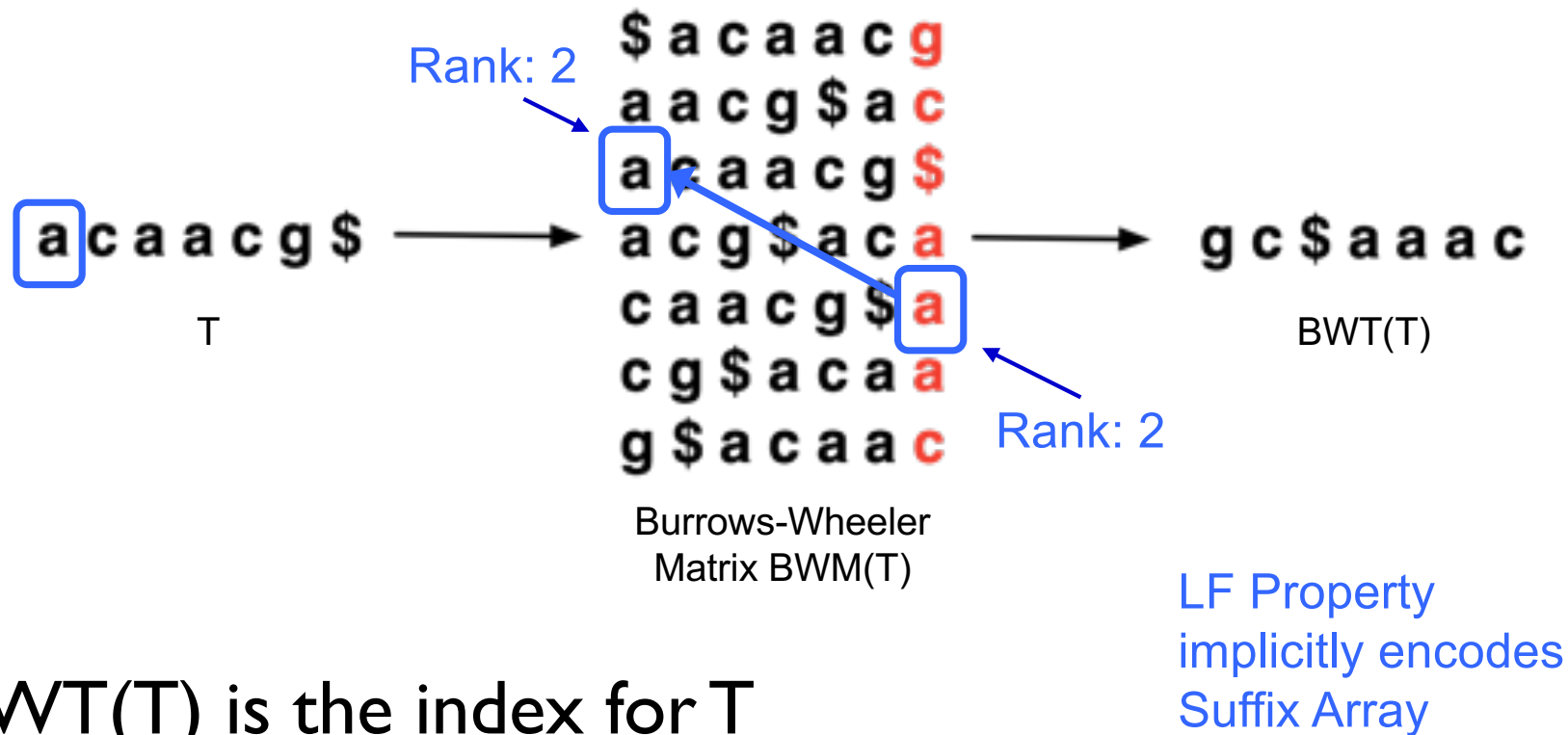
BWT(T) is the index for T

**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

# Burrows-Wheeler Transform

- Reversible permutation of the characters in a text



- BWT( $T$ ) is the index for  $T$

**A block sorting lossless data compression algorithm.**

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

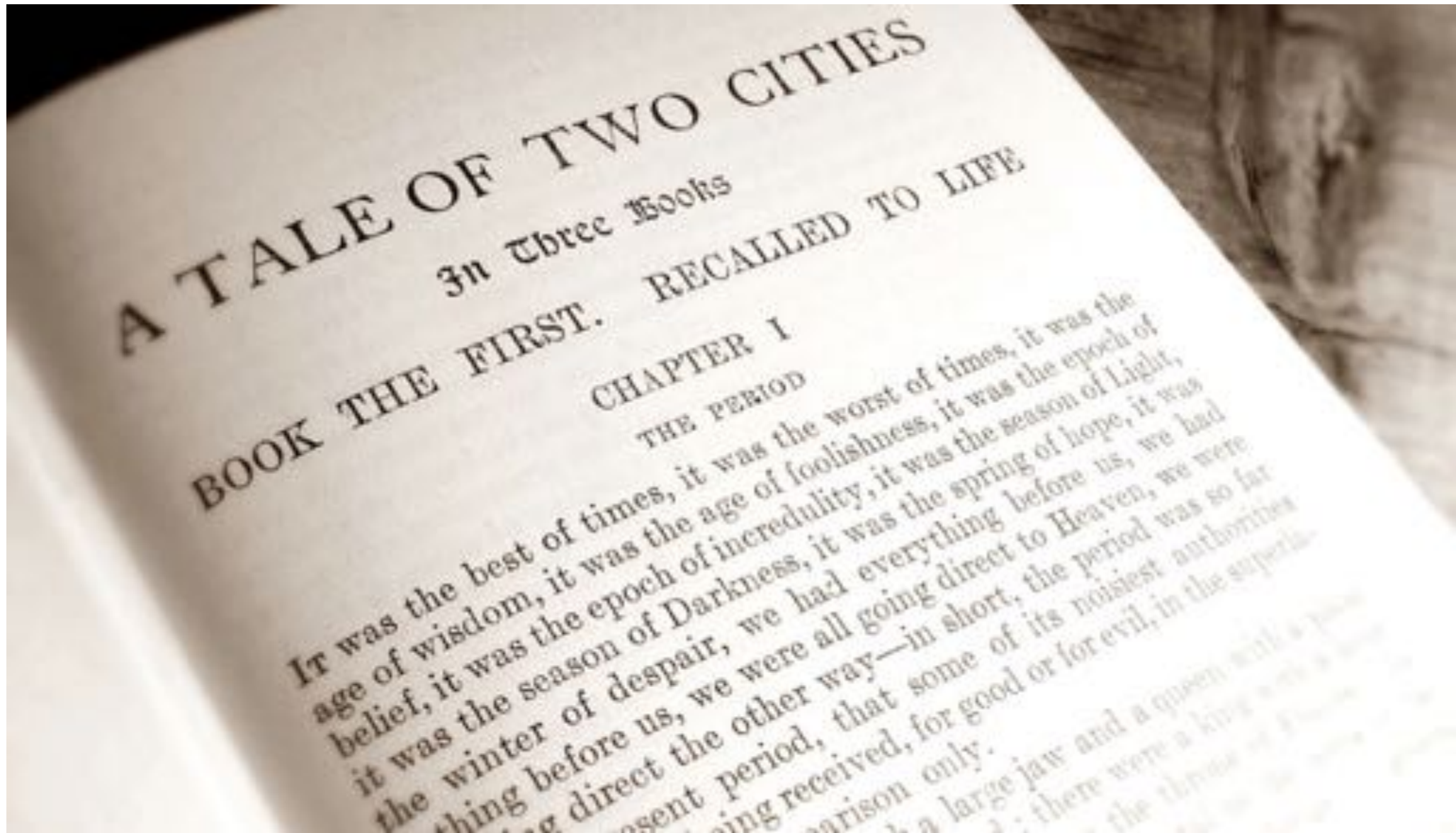
# Burrows-Wheeler Transform

- Recreating T from BWT(T)
  - Start in the first row and apply **LF** repeatedly, accumulating predecessors along the way



[Decode this BWT string: ACTGA\$TTA ]

# Run Length Encoding



# Run Length Encoding

**ref[614]:**

It\_was\_the\_best\_of\_times,\_it\_was\_the\_worst\_of\_times,\_it\_was\_the\_age\_of\_wisdom,\_it\_was\_the\_age\_of\_foolishness,\_it\_was\_the\_epoch\_of\_belief,\_it\_was\_the\_epoch\_of\_incredulity,\_it\_was\_the\_season\_of\_Light,\_it\_wa\_s\_the\_season\_of\_Darkness,\_it\_was\_the\_spring\_of\_hope,\_it\_was\_the\_wint\_er\_of\_despair,\_we\_had\_everything\_before\_us,\_we\_had\_nothing\_before\_us,\_we\_were\_all\_going\_direct\_to\_Heaven,\_we\_were\_all\_going\_direct\_the\_o ther\_way\_-\_in\_short,\_the\_period\_was\_so\_far\_like\_the\_present\_period,\_that\_some\_of\_its\_noisiest\_authorities\_insisted\_on\_its\_being\_received,\_for\_good\_or\_for\_evil,\_in\_the\_superlative\_degree\_of\_comparison\_only.\$

## ***Run Length Encoding:***

- Replace a “run” of a character X with a single X followed by the length of the run
- GAAAAAAAAATTACA => GA8T2ACA (reverse is also easy to implement)
- If your text contains numbers, then you will need to use a (slightly) more sophisticated encoding

# Run Length Encoding

**ref[614]:**

It\_was\_the\_best\_of\_times,\_it\_was\_the\_worst\_of\_times,\_it\_was\_the\_age\_of\_wisdom,\_it\_was\_the\_age\_of\_foolishness,\_it\_was\_the\_epoch\_of\_belief,\_it\_was\_the\_epoch\_of\_incredulity,\_it\_was\_the\_season\_of\_Light,\_it\_wa\_s\_the\_season\_of\_Darkness,\_it\_was\_the\_spring\_of\_hope,\_it\_was\_the\_wint\_er\_of\_despair,\_we\_had\_everything\_before\_us,\_we\_had\_nothing\_before\_us,\_we\_were\_all\_going\_direct\_to\_Heaven,\_we\_were\_all\_going\_direct\_the\_o ther\_way\_-\_in\_short,\_the\_period\_was\_so\_far\_like\_the\_present\_period,\_that\_some\_of\_its\_noisiest\_authorities\_insisted\_on\_its\_being\_received,\_for\_good\_or\_for\_evil,\_in\_the\_superlative\_degree\_of\_comparison\_only.\$

**rle(ref)[614]:**

It\_was\_the\_best\_of\_times,\_it\_was\_the\_worst\_of\_times,\_it\_was\_the\_age\_of\_wisdom,\_it\_was\_the\_age\_of\_fo2lishnes2,\_it\_was\_the\_epoch\_of\_belief,\_it\_was\_the\_epoch\_of\_incredulity,\_it\_was\_the\_season\_of\_Light,\_it\_wa\_s\_the\_season\_of\_Darknes2,\_it\_was\_the\_spring\_of\_hope,\_it\_was\_the\_wint\_er\_of\_despair,\_we\_had\_everything\_before\_us,\_we\_had\_nothing\_before\_us,\_we\_were\_al2\_going\_direct\_to\_Heaven,\_we\_were\_al2\_going\_direct\_the\_o ther\_way\_-\_in\_short,\_the\_period\_was\_so\_far\_like\_the\_present\_period,\_that\_some\_of\_its\_noisiest\_authorities\_insisted\_on\_its\_being\_received,\_for\_go2d\_or\_for\_evil,\_in\_the\_superlative\_degre2\_of\_comparison\_only.\$

# Run Length Encoding

**ref[614]:**

It\_was\_the\_best\_of\_times,\_it\_was\_the\_worst\_of\_times,\_it\_was\_the\_age\_of\_wisdom,\_it\_was\_the\_age\_of\_foolishness,\_it\_was\_the\_epoch\_of\_belief,\_it\_was\_the\_epoch\_of\_incredulity,\_it\_was\_the\_season\_of\_Light,\_it\_wa\_s\_the\_season\_of\_Darkness,\_it\_was\_the\_spring\_of\_hope,\_it\_was\_the\_wint\_er\_of\_despair,\_we\_had\_everything\_before\_us,\_we\_had\_nothing\_before\_us,\_we\_were\_all\_going\_direct\_to\_Heaven,\_we\_were\_all\_going\_direct\_the\_o ther\_way\_-\_in\_short,\_the\_period\_was\_so\_far\_like\_the\_present\_period,\_that\_some\_of\_its\_noisiest\_authorities\_insisted\_on\_its\_being\_received,\_for\_good\_or\_for\_evil,\_in\_the\_superlative\_degree\_of\_comparison\_only.\$

**bwt[614]:**

.dlmssftysesdtrsns\_y\_\_\$\_yfofeeeeetggsfefefggedrofr,llreef-,fs,,,,,,,,, ,nfrsdnnhereghettedndeteegenstee,sssst,esssnssffteedtttttttttttr,, ,eeefehh\_\_p\_\_fpDwwwwwwwwweehl\_ew\_\_\_\_\_eoo\_neeeoaaeoo\_\_\_\_sephhrrrhvh hwwegmghhhhhhhkrrwwhhssHrrrvtrribdbcbvs\_\_thwwpppvmmirdnnib\_\_eoooooo oooooo\_\_\_\_eenennnnnaai\_\_ecc\_\_ttttttttttttttttttts\_tsgltsLlvtt\_\_hhoor e\_wrraddwlors\_\_\_\_\_r\_\_lteirillre\_ouaanooiioeooooiihkiiiiiiio\_\_iei tsppioi\_\_\_\_\_ggnodsc\_sss\_gfhf\_fffhwh\_nsmo\_\_uee\_sioooaeeeeoo\_ii cgppeeaoaeooeesseuutetaaaaaaaaaaaai\_\_ei\_in\_\_aaie\_eereei\_hrsssnacciiIi iiiiiisn\_\_\_\_\_oyoui\_\_a\_iids\_\_aiaae\_\_\_\_\_tlar

# Run Length Encoding

**ref[614]:**

It\_was\_the\_best\_of\_times,\_it\_was\_the\_worst\_of\_times,\_it\_was\_the\_age\_of\_wisdom,\_it\_was\_the\_age\_of\_foolishness,\_it\_was\_the\_epoch\_of\_belief,\_it\_was\_the\_epoch\_of\_incredulity,\_it\_was\_the\_season\_of\_Light,\_it\_wa\_s\_the\_season\_of\_Darkness,\_it\_was\_the\_spring\_of\_hope,\_it\_was\_the\_wint\_er\_of\_despair,\_we\_had\_everything\_before\_us,\_we\_had\_nothing\_before\_us,\_we\_were\_all\_going\_direct\_to\_Heaven,\_we\_were\_all\_going\_direct\_the\_o ther\_way\_-\_in\_short,\_the\_period\_was\_so\_far\_like\_the\_present\_period,\_that\_some\_of\_its\_noisiest\_authorities\_insisted\_on\_its\_being\_received,\_for\_good\_or\_for\_evil,\_in\_the\_superlative\_degree\_of\_comparison\_only.\$

**bwt[614]:**

.dlmssftysesdtrsns\_y\_\_\$\_yfofeeeeetggsfefefggeedrofr,llreef-,fs,,,,,,,,, ,nfrsdnnhereghettedndeteegenstee,sssst,esssnssffteedtttttttttttr,, ,eeefehh\_\_p\_\_fpDwwwwwwwwweehl\_ew\_\_\_\_\_eoo\_neeeoaaeoo\_\_\_\_sephhrrrhvh hwwegmghhhhhhhkrrwwhhssHrrrvtrribdbcbvs\_\_thwwpppvmmirdnnib\_\_eoooooo oooooo\_\_\_\_eenennnnnaai\_\_ecc\_\_ttttttttttttttttttts\_tsgltsLlvtt\_\_hhoor e\_wrraddwlors\_\_\_\_\_r\_\_lteirillre\_ouaanooiioeooooiihkiiiiiiio\_\_iei tsppioi\_\_\_\_\_ggnodsc\_sss\_gfhf\_fffhwh\_nsmo\_\_uee\_sioooaeeeeoo\_ii cgppeeaoaeooeesseuutetaaaaaaaaaaaai\_\_ei\_in\_\_aaie\_eereei\_hrsssnacciiIi iiiiiisn\_\_\_\_\_ Why does the BWT tend to make runs in english text? \_\_\_\_\_tlar



# Run Length Encoding

**bwt[614]:**

```
.dlmssftysesdtrsns_y__$_yfofeeeeetggsfefefggedrofr,llreef-,fs,,,,,,,,,  
,,nfrsdnnhereghettedndeteegenstee,sssst,esssnssffteedtttttttttttr,,  
,,eeefehh__p__fpDwwwwwwwwweehl_ew_____eoo_neeeoaaeoo____sephhrrhvh  
hwwegmghhhhhhhkrrwwhhssHrrrvtrribbdbcbvs__thwwpppvmmirdnnib__eoooooo  
oooooo____eenennnnnaai__ecc__ttttttttttttttttttts_tsgltsLlvtt__hhoor  
e_wrraddwlors_____r__lteirillre_ouaanooiioeooooiiihkiiiiio__iei  
tsppioi_____ggnodsc_sss_gfhf_fffhwh_nsmo__uee_sioooaeeeeoo_ii  
cgppeeaoaeooesseuutetaaaaaaaaaaaaaai__ei_in__aaie_eeerei_hrsssnacciIi  
iiiiisn_____oyoui__a_iids__aiaee_____tlar
```

**rle(bwt)[464]:**

```
.dlms2ftysesdtrsns_y_2$_yfofe4tg2sfefefg2e2drofr,l2re2f-,fs,9nfrsdn2  
hereghet2edndete2ge2nste2,s5t,es3ns2f2te2dt10r,4e3feh2_2p_2fpDw11e2h  
l_ew_5eo2_ne3oa2eo2_4seph2r2hvh2w2egmgh7kr2w2h2s2Hr3vtr2ib2dbcbvs_2t  
hw2p3vm2irdn2ib_2eo12_4e2n6a2i_3ec2_2t18s_tsgltsLlvtt2_3h2o2re_wr2ad2  
wlors_9r_2lteiril2re_oua2no2i2oeo4i3hki6o_2ieitsp2ioi_12g2nodsc_s3_g  
fhf_f3hwh_nsmo_2ue2_sio3ae4o2_i2cgp2e2aoaeo2e2s2eu2teta11i_2ei_in_2a  
2ie_e3rei_hrs3nac2i2Ii7sn_15oyoui_2a_i3ds_2ai2ae2_21tlar
```

# Run Length Encoding

**bwt[614]:**

```
.dlmssftysesdtrsns_y__$_yfofeeeeetggsfefefggeedrofr,llreef-,fs,,,,,,,,,  
,,nfrsdnnherghettedndeteegenstee,sssst,esssnssffteedtttttttttr,,  
,,eeefehh__p__fpDwwwwwwwwweehl_ew_____eoo_neeeoaaeo_____sephhrrhvh  
hwwegmghhhhhhhkrrwwhhssHrrrvtrribdbcbvs__thwwpppvmmirdnnib__eooooo  
ooooo_____eennnnnnaai__ecc__ttttttttttttttts_tsgltsLlvtt__hhoor  
e_wrraddwlors_____r__lteirillre_ouaanooiioeooooiihkiiiiio__iei  
tsppioi_____ggnodsc_sss_gfhf_fffhwh_nsmo__uee_sioooaeeeeoo__ii  
cgppeeaoaeooesseuutetaaaaaaaaai__ei_in__aaie_eeerei_hrsssnacciIi  
iiiiisn_____oyoui__a_iids__aiaee_____tlar
```

**rle(bwt)[464]:**

```
.dlms2ftysesdtrsns_y_2$_yfofe4tg2sfefefg2e2drofr,l2re2f-,fs,9nfrsdn2  
herghet2edndete2ge2nste2,s5t,es3ns2f2te2dt10r,4e3feh2_2p_2fpDw11e2h  
l_ew_5eo2_ne3oa2eo2_4seph2r2hvh2w2egmgh7kr2w2h2s2Hr3vtr2ib2dbcbvs_2t  
hw2p3vm2irdn2ib_2eo12_4e2n6a2i_3ec2_2t18s_tsgltsLlvtt2_3h2o2re_wr2ad2  
wlors_9r_2lteiril2re_oua2no2i2oeo4i3hki6o_2ieitsp2ioi_12g2nodsc_s3_g  
fhf_f3hwh_nsmo_2ue2_sio3ae4o2_i2cgp2e2aoaeo2e2s2eu2tet11i_2ei_in_2a  
2ie_e3rei_hrs3nac2i2Ii7sn_15oyoui_2a_i3ds_2ai2ae2_21tlar
```

# Run Length Encoding

**ref[614]:**

It\_was\_the\_best\_of\_times,\_it\_was\_the\_worst\_of\_times,\_it\_was\_the\_age\_of\_wisdom,\_it\_was\_the\_age\_of\_foolishness,\_it\_was\_the\_epoch\_of\_belief,\_it\_was\_the\_epoch\_of\_incredulity,\_it\_was\_the\_season\_of\_Light,\_it\_wa\_s\_the\_season\_of\_Darkness,\_it\_was\_the\_spring\_of\_hope,\_it\_was\_the\_wint\_er\_of\_despair,\_we\_had\_everything\_before\_us,\_we\_had\_nothing\_before\_us,\_we\_were\_all\_going\_direct\_to\_Heaven,\_we\_were\_all\_going\_direct\_the\_o ther\_way\_-\_in\_short,\_the\_period\_was\_so\_far\_like\_the\_present\_period,\_that\_some\_of\_its\_noisiest\_authorities\_insisted\_on\_its\_being\_received,\_for\_good\_or\_for\_evil,\_in\_the\_superlative\_degree\_of\_comparison\_only.\$

**rle(bwt)[464]:**

.dlms2ftysesdtrsns\_y\_2\$\_yfofe4tg2sfefefg2e2drofr,l2re2f-,fs,9nfrsdn2 hereghet2edndete2ge2nste2,s5t,es3ns2f2te2dt10r,4e3feh2\_2p\_2fpDw11e2h l\_ew\_5eo2\_ne3oa2eo2\_4seph2r2hvh2w2egmgh7kr2w2h2s2Hr3vtr2ib2dbcbvs\_2t hw2p3vm2irdn2ib\_2eo12\_4e2n6a2i\_3ec2\_2t18s\_tsgltsLlvt2\_3h2o2re\_wr2ad2 wlors\_9r\_2lteiril2re\_oua2no2i2oeo4i3hki6o\_2ieitsp2ioi\_12g2nodsc\_s3\_g fhf\_f3hwh\_nsmo\_2ue2\_sio3ae4o2\_i2cgp2e2aoaao2e2s2eu2tet11i\_2ei\_in\_2a 2ie\_e3rei\_hrs3nac2i2Ii7sn\_15oyoui\_2a\_i3ds\_2ai2ae2\_21tlar

# Run Length Encoding

**ref[614]:**

It\_was\_the\_best\_of\_times,\_it\_was\_the\_worst\_of\_times,\_it\_was\_the\_age\_of\_wisdom,\_it\_was\_the\_age\_of\_foolishness,\_it\_was\_the\_epoch\_of\_belief,\_it\_was\_the\_epoch\_of\_incredulity,\_it\_was\_the\_season\_of\_Light,\_it\_wa\_s\_the\_season\_of\_Darkness,\_it\_was\_the\_spring\_of\_hope,\_it\_was\_the\_wint\_er\_of\_despair,\_we\_had\_everything\_before\_us,\_we\_had\_nothing\_before\_us,\_we\_were\_all\_going\_direct\_to\_Heaven,\_we\_were\_all\_going\_direct\_the\_o ther\_way\_-\_in\_short,\_the\_period\_was\_so\_far\_like\_the\_present\_period,\_that\_some\_of\_its\_noisiest\_authorities\_insisted\_on\_its\_being\_received,\_for\_good\_or\_for\_evil,\_in\_the\_superlative\_degree\_of\_comparison\_only.\$

**rle(bwt)[464]:**

.dlms2ftysesdtrsns\_y\_2\$\_yfofe4tg2sfefefg2e2drofr,l2re2f-,fs,9nfrsdn2 hereghet2edndete2ge2nste2,s5t,es3ns2f2te2dt10r,4e3feh2\_2p\_2fpDw11e2h l\_ew\_5eo2\_ne3oa2eo2\_4seph2r2hvh2w2egmgh7kr2w2h2s2Hr3vtr2ib2dbcbvs\_2t hw2p3vm2irdn2ib\_2eo12\_4e2n6a2i\_3ec2\_2t18s\_tsgltsLlvt2\_3h2o2re\_wr2ad2 wlors\_9r\_2lteiril2re\_oua2no2i2oeo4i3hki6o\_2ieitsp2ioi\_12g2nodsc\_s3\_g fhf\_f3hwh\_nsmo\_2ue2\_sio3ae4o2\_i2cgp2e2aoaeo2e2s2eu2tet11i\_2ei\_in\_2a 2ie\_e3rei

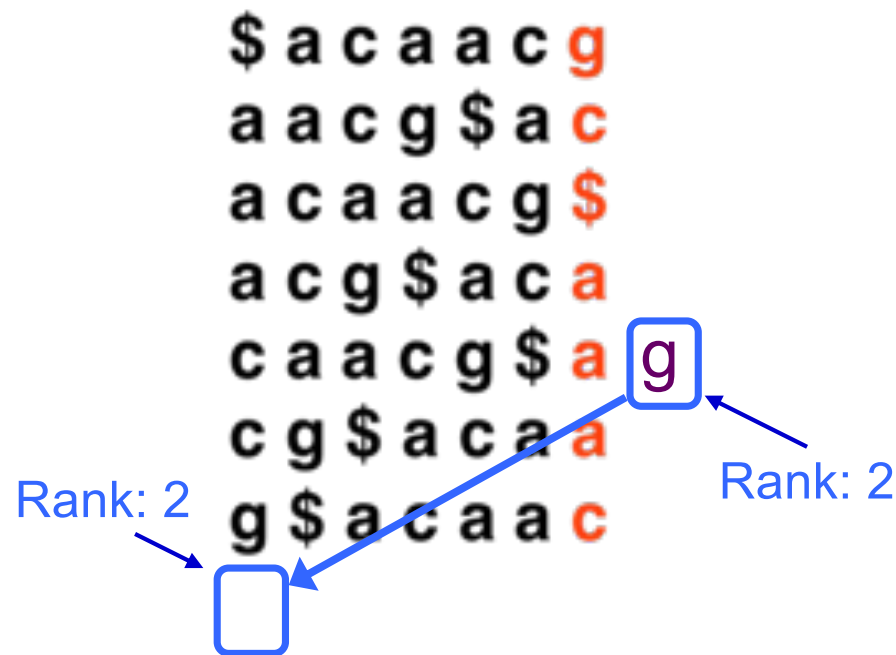
Saved 614-464 = 150 bytes (24%) with zero loss of information!

Common to save 50% to 90% on real world files with bzip2

# BWT Exact Matching

- **LFc**(r, c) does the same thing as **LF**(r) but it ignores r's actual final character and “pretends” it's c:

$$\text{LFc}(5, g) = 8$$

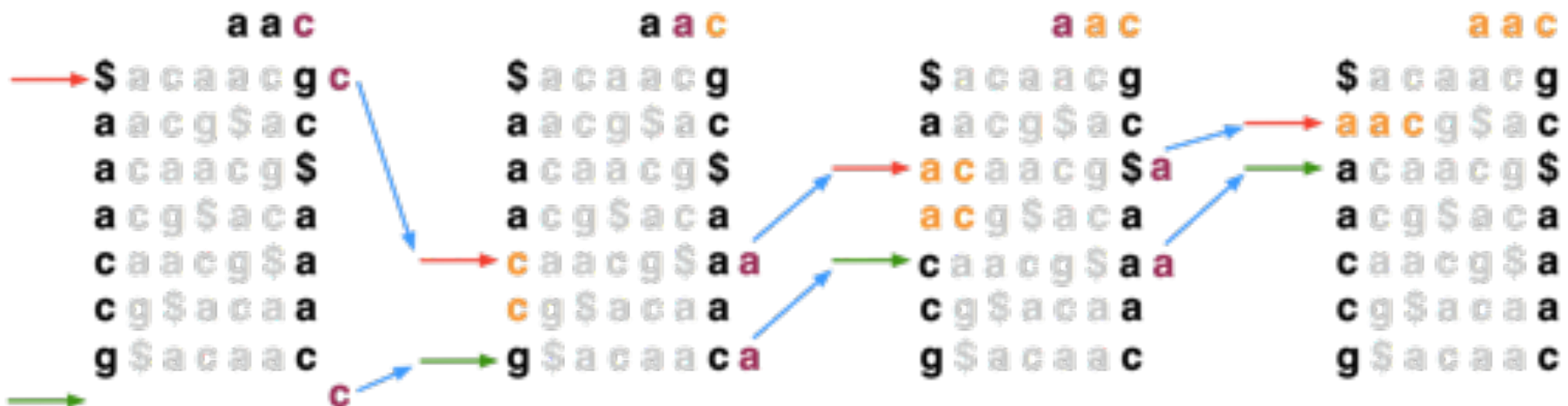


# BWT Exact Matching

- Start with a range, (**top**, **bot**) encompassing all rows and repeatedly apply **LFc**:

**top** = **LFc**(**top**, **qc**); **bot** = **LFc**(**bot**, **qc**)

**qc** = the next character to the left in the query

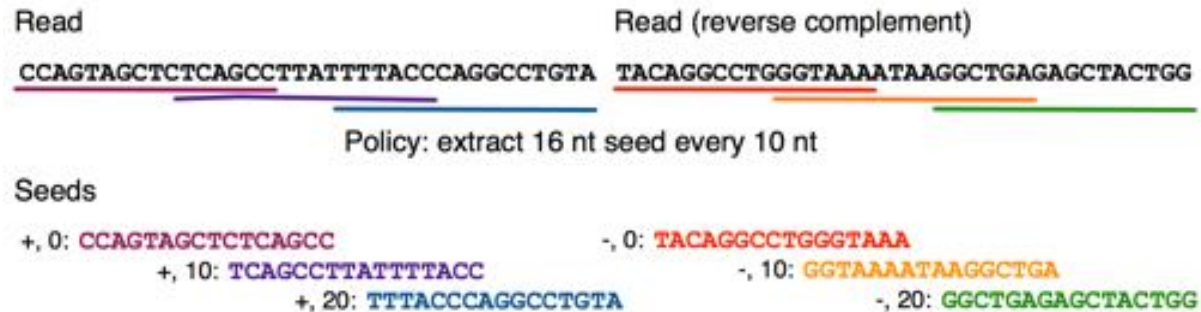


Ferragina P, Manzini G: Opportunistic data structures with applications. *FOCS. IEEE Computer Society; 2000.*

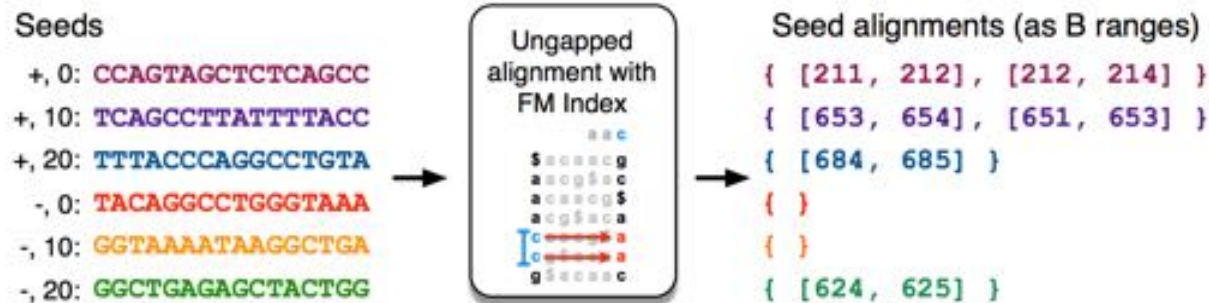
[Search for TTA this BWT string: ACTGA\$TTA ]

# Algorithm Overview

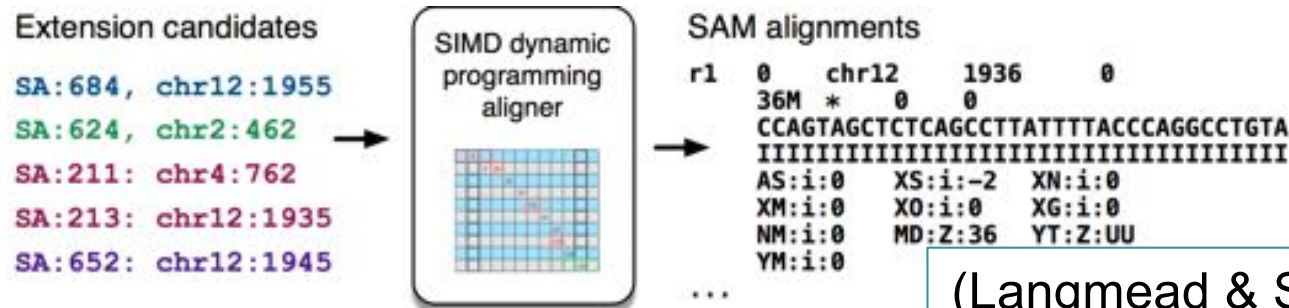
## 1. Split read into segments



## 2. Lookup each segment and prioritize



### 3. Evaluate end-to-end match



(Langmead & Salzberg, 2012)