Variant Calling

Michael Schatz

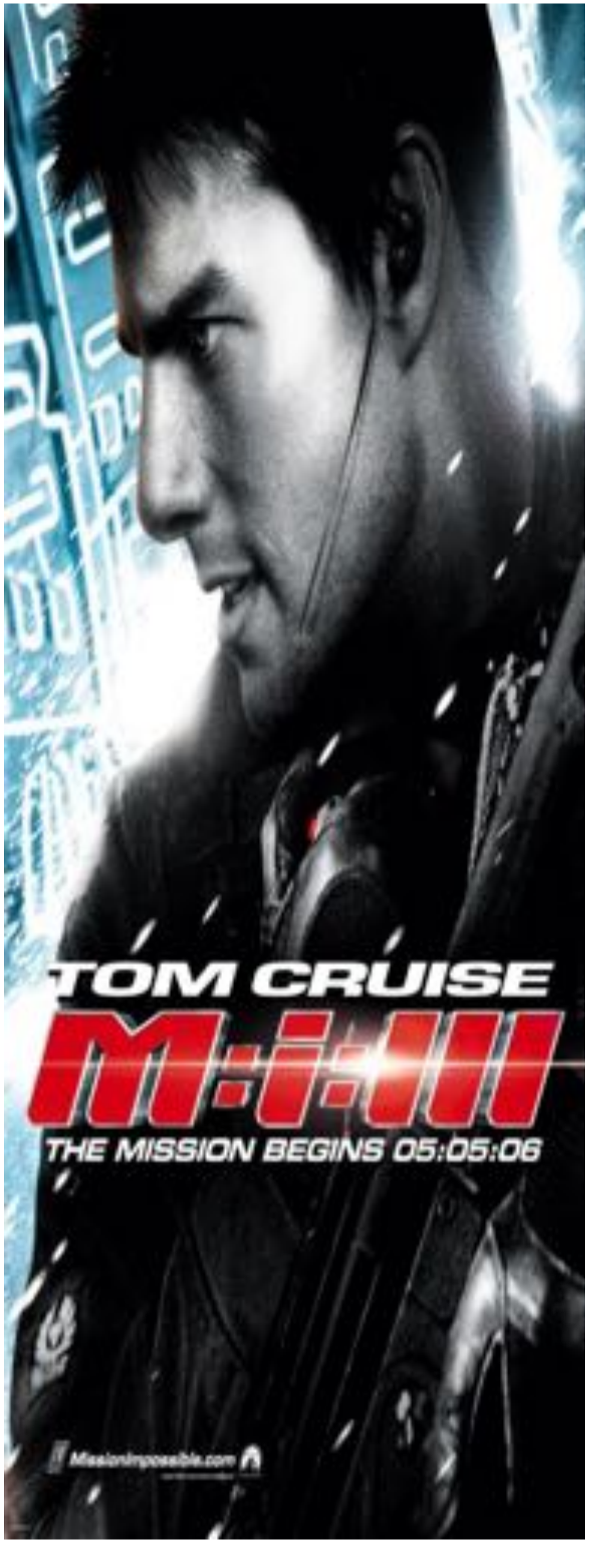
Feb 20, 2018

Lecture 7: Applied Comparative Genomics

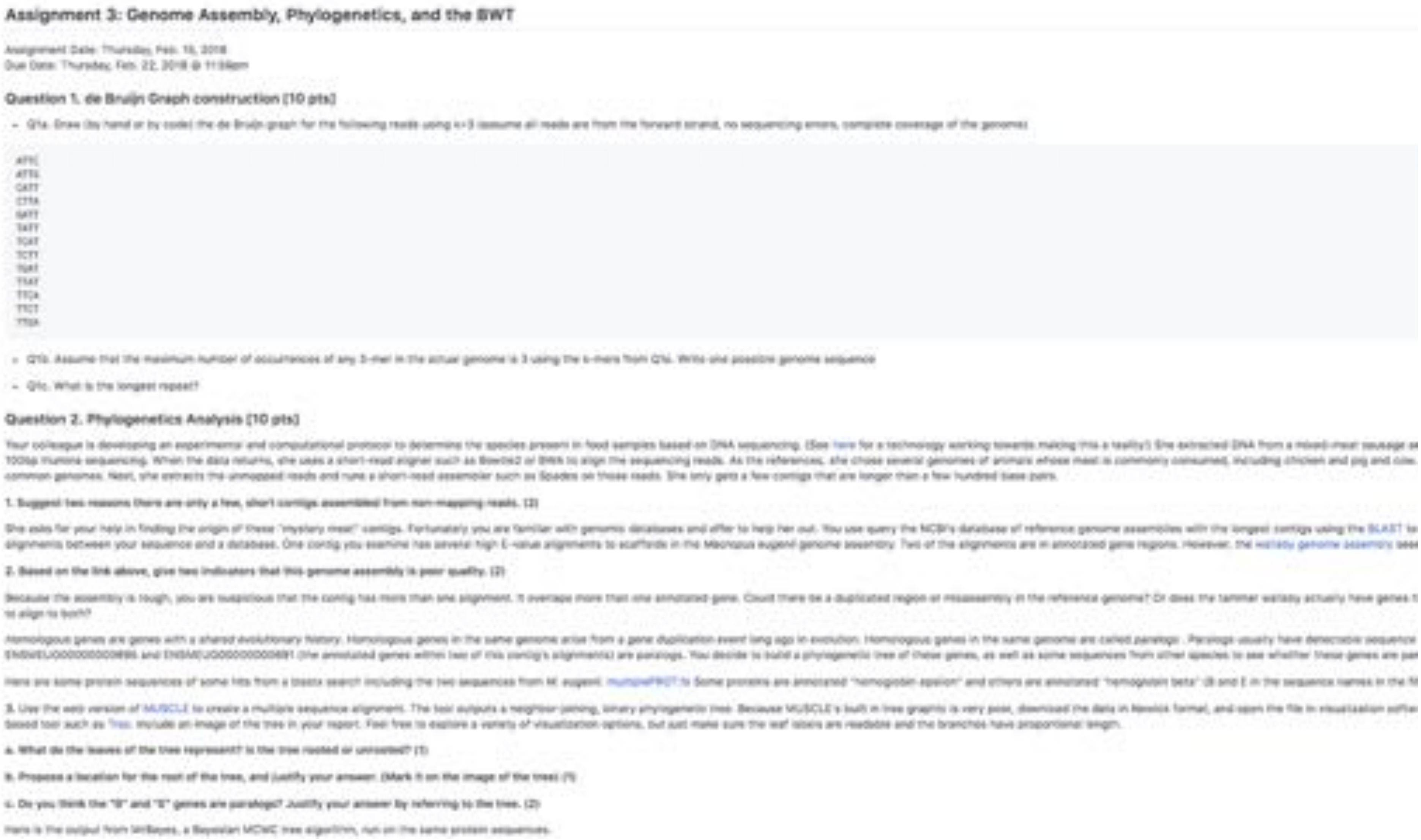
Mission Impossible

1. ***Setup VirtualBox***
2. ***Initialize Tools***
3. ***Download Reference Genome & Reads***
4. ***Decode the secret message***
    1. *Estimate coverage, check read quality*
    2. *Check kmer distribution*
    3. *Assemble the reads with spades*
    4. *Align to reference with MUMmer*
    5. *Extract foreign sequence*
    6. *dna-encode.pl -d*

https://github.com/schatzlab/appliedgenomics2018/blob/master/assignments/assignment2/README.md



Assignment 3: Due Thursday Feb 22



Assignment 3: Genome Assembly, Phylogenetics, and the BWT

Assignment Date: Thursday, Feb. 15, 2018
Due Date: Thursday, Feb. 22, 2018 @ 11:59pm

Question 1. de Bruijn Graph construction [10 pts]

- Q1a. Draw (by hand or by code) the de Bruijn graph for the following reads using k=3 (assume all reads are from the forward strand, no sequencing errors, complete coverage of the genome)

```
ATTC
ATTG
CATT
CTTA
GATT
TATT
TCAT
TCTT
TGAT
TTAT
TTCA
TTCT
TTGA
```

- Q1b. Assume that the maximum number of occurrences of any 3-mer in the actual genome is 3 using the k-mers from Q1a. Write one possible genome sequence
- Q1c. What is the longest repeat?

Question 2. Phylogenetics Analysis [10 pts]

Your colleague is developing an experimental and computational protocol to determine the species present in food samples based on DNA sequencing. (See here for a technology working towards making this a reality!) She extracted DNA from a mixed-meat sausage a
100bp Illumina sequencing. When the data returns, she uses a short-read aligner such as Bowtie2 or BWA to align the sequencing reads. As the references, she chose several genomes of animals whose meat is commonly consumed, including chicken and pig and cow.
common genomes. Next, she extracts the unmapped reads and runs a short-read assembler such as Spades on those reads. She only gets a few contigs that are longer than a few hundred base pairs.

1. Suggest two reasons there are only a few, short contigs assembled from non-mapping reads. (2)

She asks for your help in finding the origin of these "mystery meat" contigs. Fortunately you are familiar with genomic databases and offer to help her out. You use query the NCBI's database of reference genome assemblies with the longest contigs using the BLAST to
alignments between your sequence and a database. One contig you examine has several high E-value alignments to scaffolds in the Macropus eugenii genome assembly. Two of the alignments are in annotated gene regions. However, the wallaby genome assembly see

2. Based on the link above, give two indicators that this genome assembly is poor quality. (2)

Because the assembly is rough, you are suspicious that the contig has more than one alignment. It overlaps more than one annotated gene. Could there be a duplicated region or misassembly in the reference genome? Or does the tammar wallaby actually have genes t
to align to both?

Homologous genes are genes with a shared evolutionary history. Homologous genes in the same genome arise from a gene duplication event long ago in evolution. Homologous genes in the same genome are called paralogs . Paralogs usually have detectable sequence
ENSMEUG00000000898 and ENSMEUG00000000891 (the annotated genes within two of this contig's alignments) are paralogs. You decide to build a phylogenetic tree of these genes, as well as some sequences from other species to see whether these genes are pa

Here are some protein sequences of some hits from a blastx search including the two sequences from M. eugenii: mysteryPROT.fa Some proteins are annotated "hemoglobin epsilon" and others are annotated "hemoglobin beta" (B and E in the sequence names in the fi

3. Use the web version of MUSCLE to create a multiple sequence alignment. The tool outputs a neighbor-joining, binary phylogenetic tree. Because MUSCLE's built in tree graphic is very poor, download the data in Newick format, and open the file in visualization softw
based tool such as Trex. Include an image of the tree in your report. Feel free to explore a variety of visualization options, but just make sure the leaf labels are readable and the branches have proportional length.

a. What do the leaves of the tree represent? Is the tree rooted or unrooted? (1)

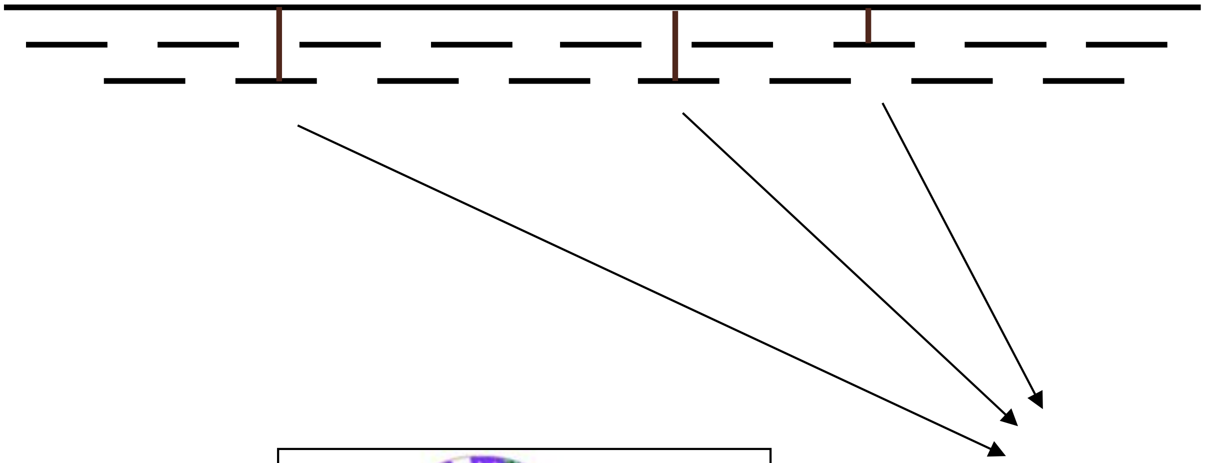
b. Propose a location for the root of the tree, and justify your answer. (Mark it on the image of the tree) (1)

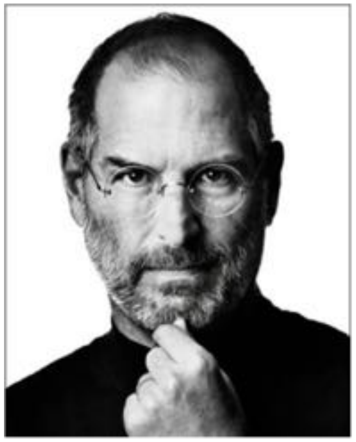
c. Do you think the "B" and "E" genes are paralogs? Justify your answer by referring to the tree. (2)

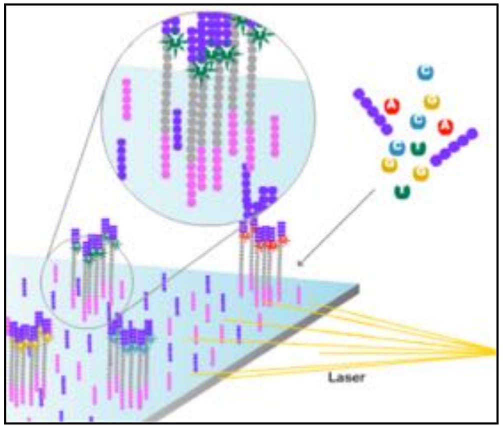
Here is the output from MrBayes, a Bayesian MCMC tree algorithm, run on the same protein sequences.

Personal Genomics

How does your genome compare to the reference?







Heart Disease

Cancer

Creates magical technology

Binary Search Analysis of Suffix Arrays

- Binary Search

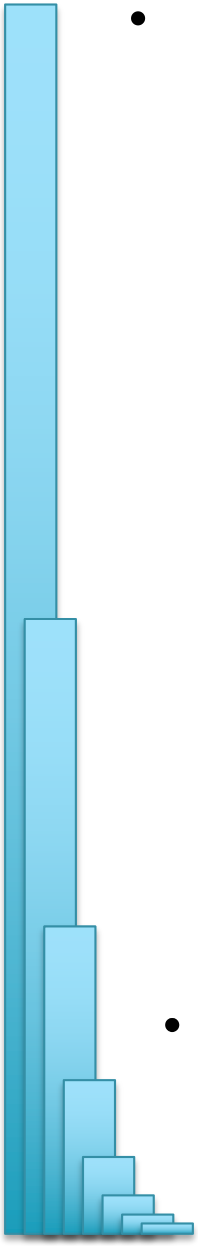
```
Initialize search range to entire list
    mid = (hi+lo)/2; middle = suffix[mid]
    if query matches middle: done
    else if query < middle: pick low range
    else if query > middle: pick hi range
Repeat until done or empty range                [WHEN?]
```

- Analysis
  - More complicated method
  - How many times do we repeat?
    - How many times can it cut the range in half?
    - Find smallest x such that: $n/(2^x) \leq 1$; $x = \lg_2(n)$ [32]

- Total Runtime: O(m lg n)
  - More complicated, but much faster!
  - Looking up a query loops 32 times instead of 3B

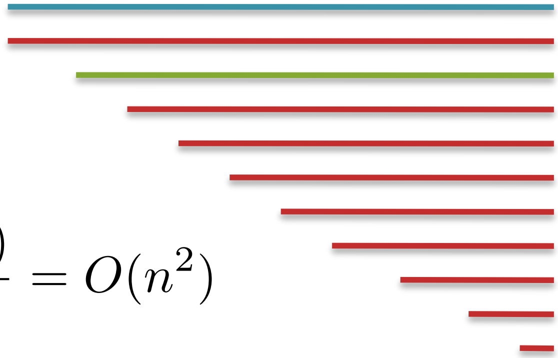
[How long does it take to search 6B or 24B nucleotides?]

Suffix Array Construction



- How can we store the suffix array?

  [How many characters are in all suffixes combined?]



$$S = 1 + 2 + 3 + \cdots + n = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = O(n^2)$$

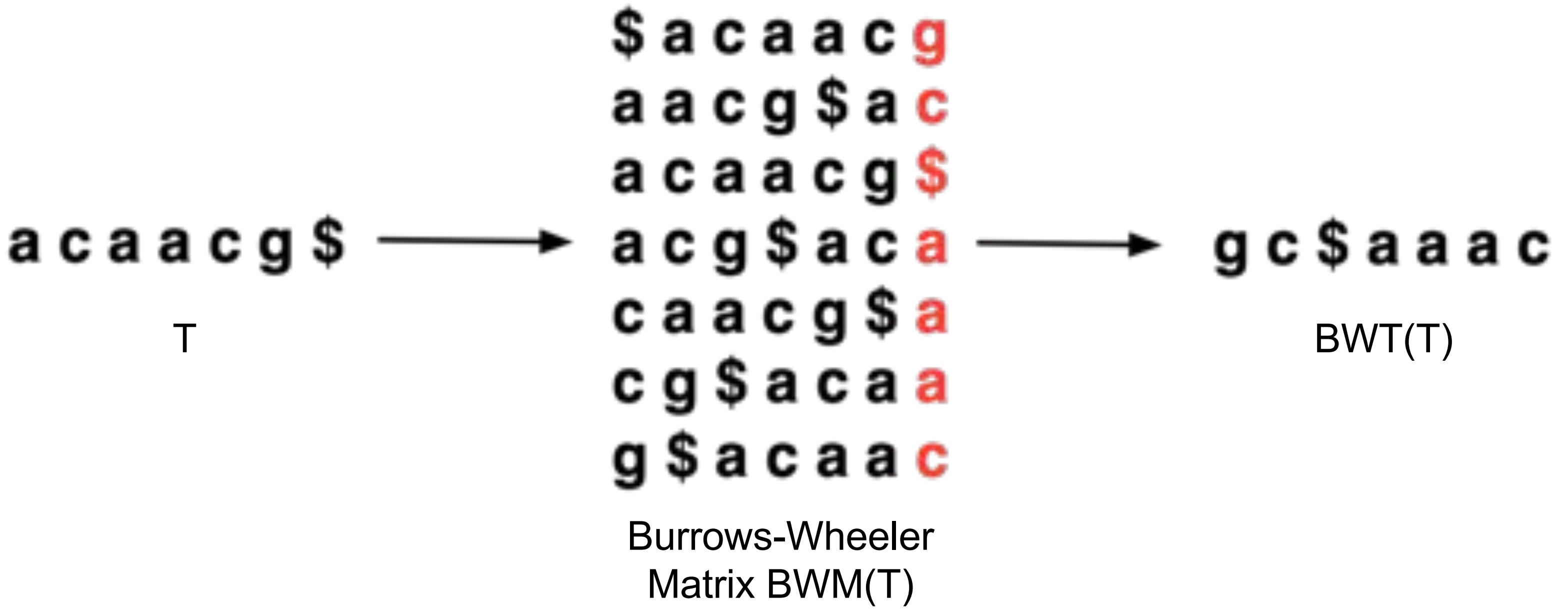
  - Hopeless to explicitly store 4.5 billion billion characters
  - Instead use implicit representation
    - Keep 1 copy of the genome, and a list of sorted offsets
    - Storing 3 billion offsets fits on a server (12GB)
- Searching the array is very fast, but it takes time to construct
  - This time will be amortized over many, many searches
  - Run it once "overnight" and save it away for all future queries

Pos
6
13
8
3
10
15
7
14
2
9
5
12
1
4
11

TGATTACAGATTACC

Burrows-Wheeler Transform

- Permutation of the characters in a text



acaacg$ → BWM(T) → gc$aaac

T

```
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
```

Burrows-Wheeler Matrix BWM(T)

gc$aaac

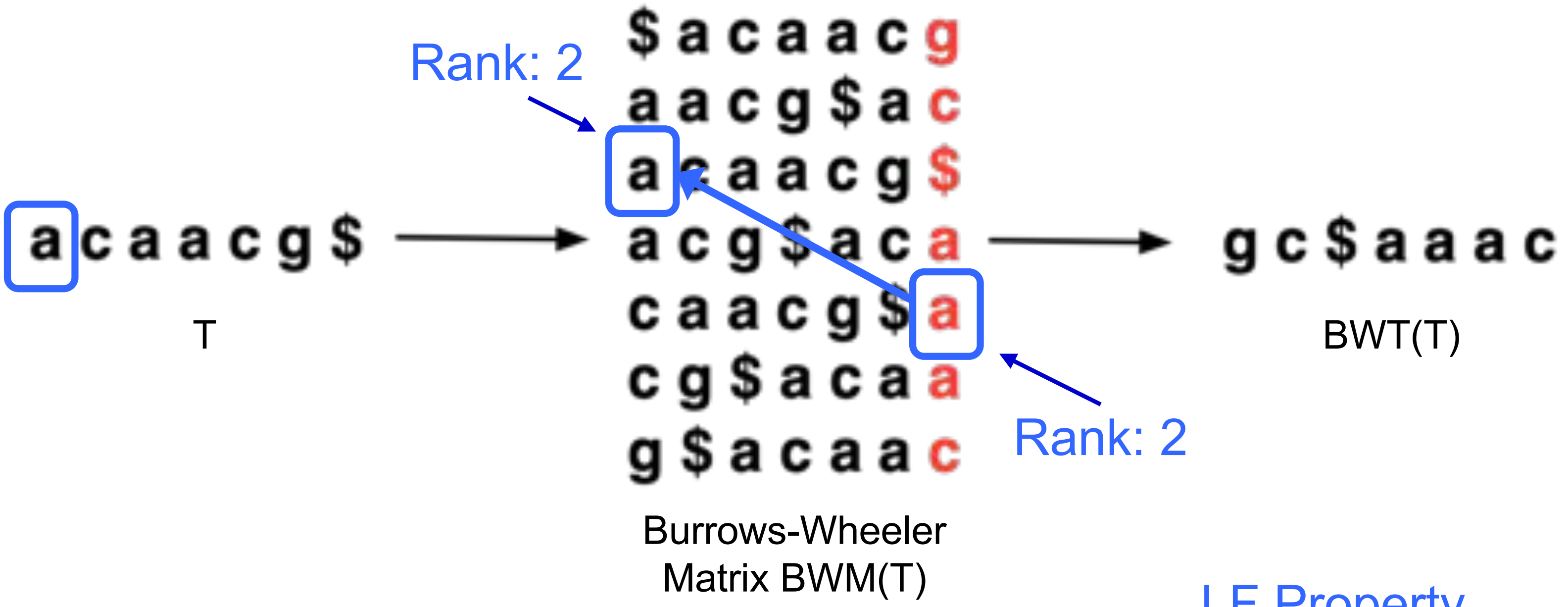
BWT(T)

- 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



Rank: 2

```
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
```

Rank: 2

acaacg$

T

gc$aaac

BWT(T)

Burrows-Wheeler Matrix BWM(T)

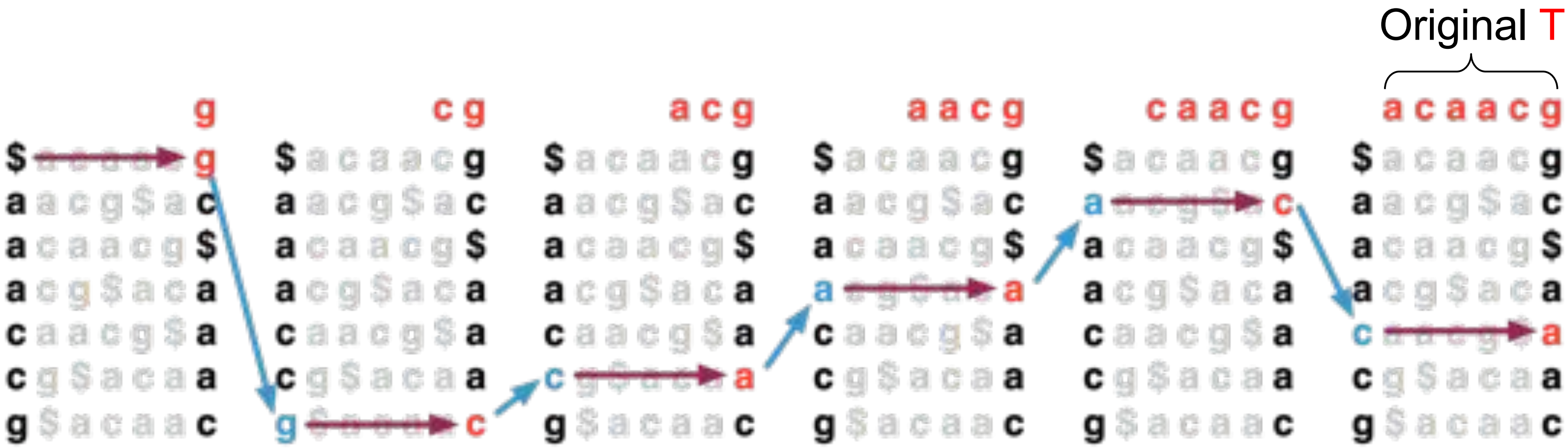
LF Property implicitly encodes Suffix Array

- 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

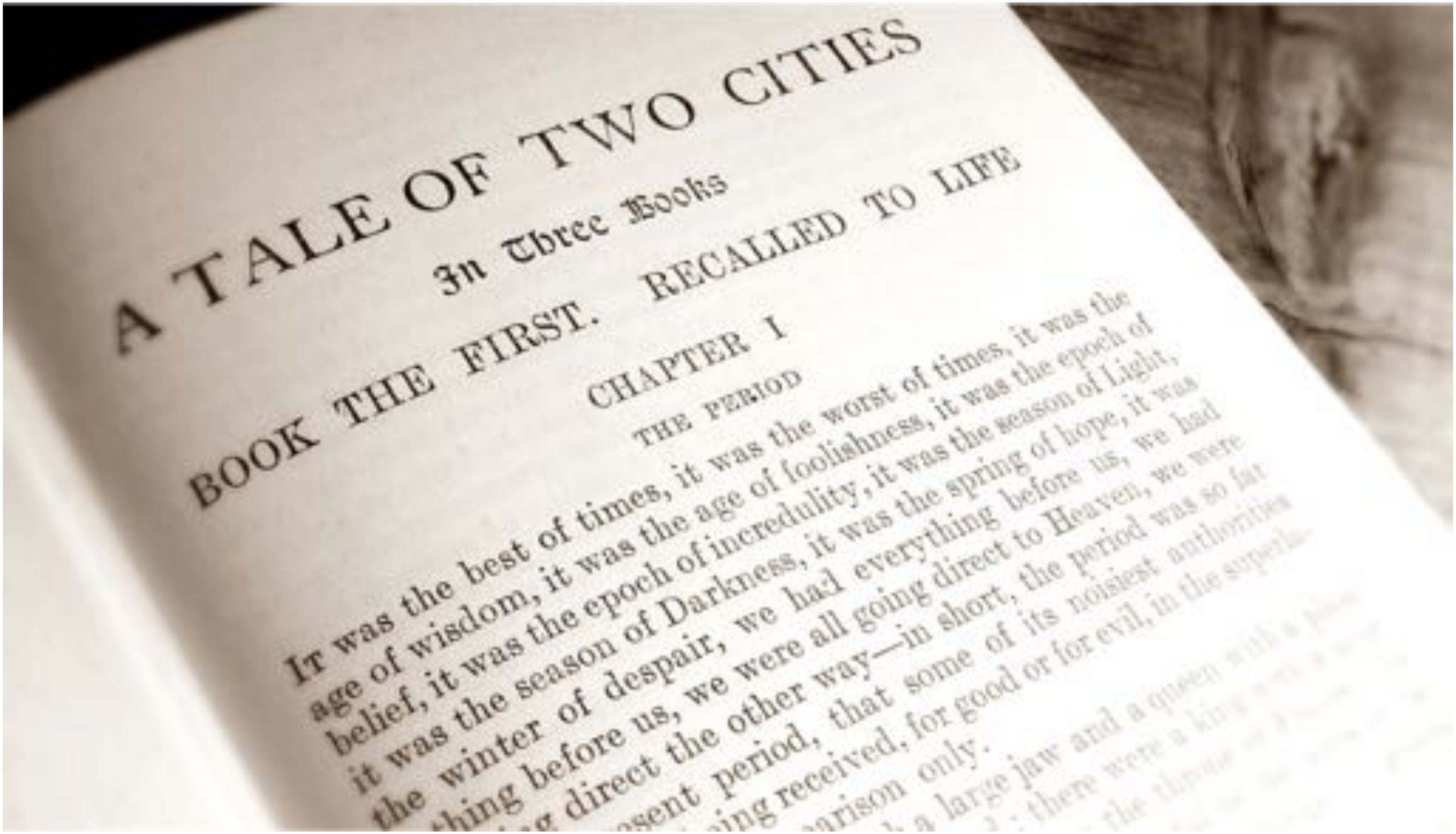
Burrows-Wheeler Transform

- 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



A TALE OF TWO CITIES

In Three Books

BOOK THE FIRST. RECALLED TO LIFE

CHAPTER I

THE PERIOD

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 was the season of Darkness, it was the spring of hope, it was
the winter of despair, we had everything before us, we had
thing before us, we were all going direct to Heaven, we were
g direct the other way—in short, the period was so far
sent period, that some of its noisiest authorities
ing received, for good or for evil, in the superla
arison only.
a large jaw and a queen with a
; there were a king with a
the throne of

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
- GAAAAAAAATTACA => 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__$_yfofeeeetggsfefefggeedrofr,llreef-,fs,,,,,,,
,,nfrsdnnhereghettedndeteegeenstee,ssssst,esssnssffteedttttttttttr,,
,,eeefehh__p__fpDwwwwwwwwwwweehl_ew_____eoo_neeeoaaeoo____sephhrrhvh
hwwegmghhhhhhkrrwwhhssHrrrvtrribbdbcbvs__thwwpppvmmirdnnib__eoooooo
oooooo____eennnnnnaai___ecc__tttttttttttttttttts_tsgltsLlvtt___hhoor
e_wrraddwlors_________r__lteirillre_ouaanooiioeooooiiihkiiiiiio__iei
tsppioi____________ggnodsc_sss_gfhf_fffhwh_nsmo__uee_sioooaeeeeoo_ii
cgppeeaoaeooeesseuutetaaaaaaaaaaai__ei_in__aaie_eeerei_hrsssnacciiIi
iiiiiisn_______________oyoui__a_iiids__aiiaee_____________________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__$_yfofeeeetggsfefefggeedrofr,llreef-,fs,,,,,,,
,,nfrsdnnhereghettedndeteegeenstee,ssssst,esssnssffteedttttttttttr,,
,,eeefehh__p__fpDwwwwwwwwwwweehl_ew_____eoo_neeeoaaeoo____sephhrrhvh
hwwegmghhhhhhhkrrwwhhssHrrrvtrribbdbcbvs__thwwpppvmmirdnnib__eoooooo
oooooo____eennnnnnaai___ecc__tttttttttttttttttts_tsgltsLlvtt___hhoor
e_wrraddwlors_________r__lteirillre_ouaanooiioeooooiiihkiiiiiio__iei
tsppioi____________ggnodsc_sss_gfhf_fffhwh_nsmo__uee_sioooaeeeeoo_ii
cgppeeaoaeooeesseuutetaaaaaaaaaaai__ei_in__aaie_eeerei_hrsssnacciiIi
iiiiiisn____                                              _______tlar
```

Why does the BWT tend to make runs in english text?

Run Length Encoding

bwt[614]:

```
.dlmssftysesdtrsns_y__$_yfofeeeetggsfefefggeedrofr,llreef-,fs,,,,,,,
,,nfrsdnnhereghettedndeteegeenstee,ssssst,esssnssffteedttttttttttr,,
,,eeefehh__p__fpDwwwwwwwwwwweehl_ew_____eoo_neeeoaaeoo____sephhrrhvh
hwwegmghhhhhhhkrrwwhhssHrrrvtrribbdbcbvs__thwwpppvmmirdnnib__eoooooo
oooooo____eennnnnnaai___ecc__tttttttttttttttttts_tsgltsLlvtt___hhoor
e_wrraddwlors_________r__lteirillre_ouaanooiioeooooiiihkiiiiiio__iei
tsppioi____________ggnodsc_sss_gfhf_fffhwh_nsmo__uee_sioooaeeeeoo_ii
cgppeeaoaeooeesseuutetaaaaaaaaaaai__ei_in__aaie_eeerei_hrsssnacciiIi
iiiiiisn_______________oyoui__a_iiids__aiiaee_____________________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_tsgltsLlvt2_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__$_yfofeeeetggsfefefggeedrofr,llreef-,fs,,,,,,,
,,nfrsdnnhereghettedndeteegeenstee,ssssst,esssnssffteedttttttttttr,,
,,eeefehh__p__fpDwwwwwwwwwwweehl_ew_____eoo_neeeoaaeoo____sephhrrhvh
hwwegmghhhhhhhkrrwwhhssHrrrvtrribbdbcbvs__thwwpppvmmirdnnib__eoooooo
oooooo____eennnnnnaai___ecc__ttttttttttttttttttts_tsgltsLlvtt___hhoor
e_wraddwlors_________r__lteirillre_ouaanooiioeooooiiihkiiiiiio__iei
tsppioi____________ggnodsc_sss_gfhf_fffhwh_nsmo__uee_sioooaeeeeoo_ii
cgppeeaoaeooeesseuutetaaaaaaaaaaai__ei_in__aaie_eeerei_hrsssnacciiIi
iiiiiisn_______________oyoui__a_iiids__aiiaee_____________________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_tsgltsLlvt2_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

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_i2cgp2e2aoaeo2e2s2eu2teta11i_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_i2cgp2e2aoaeo2e2s2eu2teta11i_2ei_in_2a
2ie_e3rei_
```

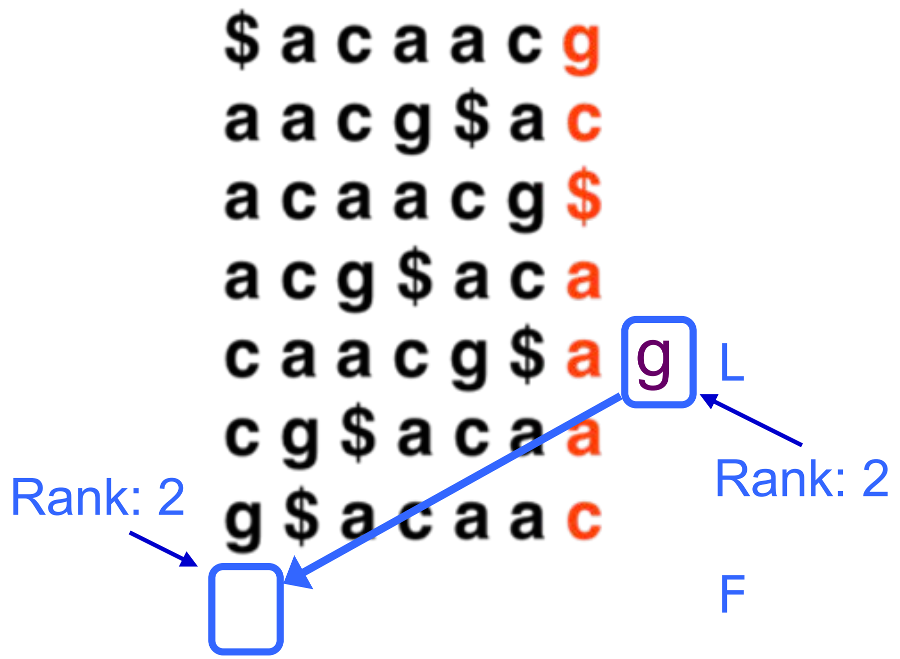
Saved 614-464 = 150 bytes (24%) with zero loss of information!

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:

LFc(5, g) = 8



```
$ 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   g   L
c g $ a c a a
g $ a c a a c
                    F
```

Rank: 2

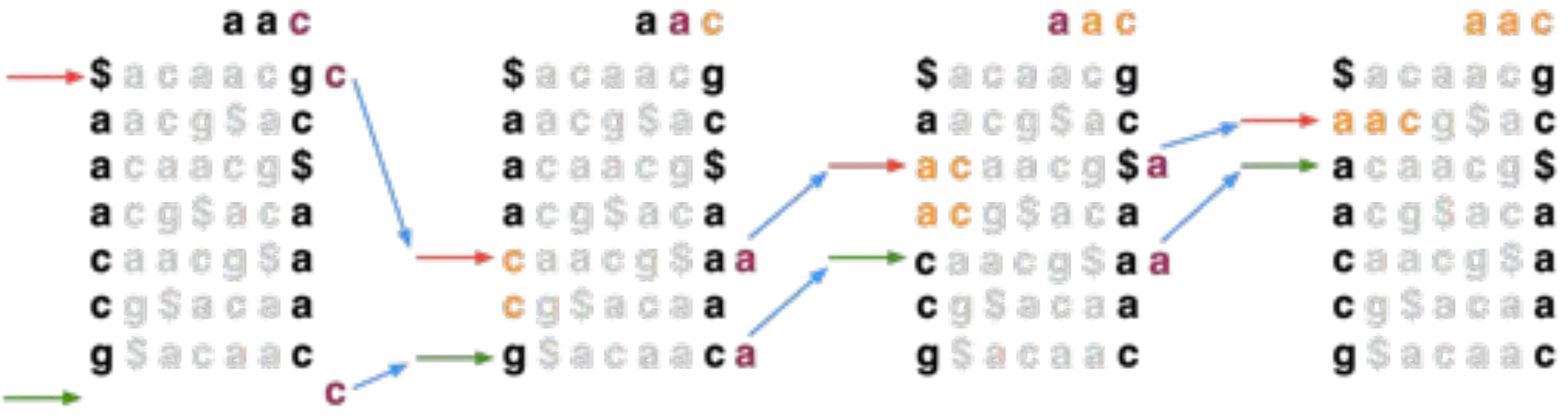
Rank: 2

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]

In-exact alignment

- Where is GATTACA *approximately* in the human genome?
  - And how do we efficiently find them?

- It depends…
  - Define 'approximately'
    - Hamming Distance, Edit distance, or Sequence Similarity
    - Ungapped vs Gapped vs Affine Gaps
    - Global vs Local
    - All positions or the single 'best'?

  - Efficiency depends on the data characteristics & goals
    - Smith-Waterman: Exhaustive search for optimal alignments
    - BLAST: Hash-table based homology searches
    - Bowtie: BWT alignment for short read mapping

Searching for GATTACA

- Where is GATTACA *approximately* in the human genome?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
G	A	T	T	A	C	A									

Match Score: 1/7

Searching for GATTACA

- Where is GATTACA *approximately* in the human genome?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
	G	A	T	T	A	C	A								

Match Score: 7/7

Searching for GATTACA

- Where is GATTACA *approximately* in the human genome?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
		G	A	T	T	A	C	A	…						

Match Score: 1/7

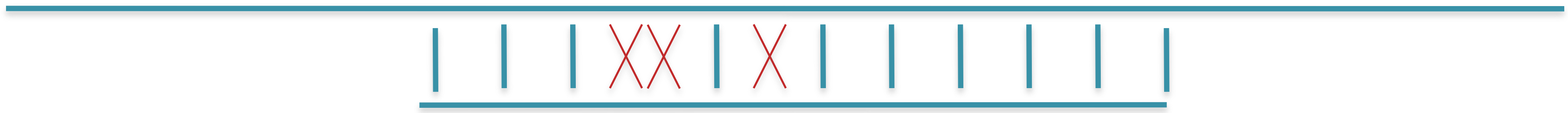
Searching for GATTACA

- Where is GATTACA *approximately* in the human genome?

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
								G	A	T	T	A	C	A	

Match Score: 6/7 <- We may be very interested in these imperfect matches
Especially if there are no perfect end-to-end matches

Hamming Distance



- How many characters are different between the 2 strings?
  - Minimum number of substitutions required to change transform A into B

- Traditionally defined for end-to-end comparisons
  - Here end-to-end (global) for query, partial (local) for reference

- Find all occurrences of GATTACA with Hamming Distance ≤ 1
- Find all occurrences with minimal Hamming Distance

[What is the running time of a brute force approach?]

Edit Distance

		A	C	A	C	A	C	T	A
	0	1	2	3	4	5	6	7	8
A	1	0	1	2	3	4	5	6	7
G	2	1	1	2	3	4	5	6	7
C	3	2	1	2	2	3	4	5	6
A	4	3	2	1	2	2	3	4	5
C	5	4	3	2	1	2	2	3	4
A	6	5	4	3	2	1	2	3	3
C	7	6	5	4	3	2	1	2	3
A	8	7	6	5	4	3	2	2	2

D[AGCACACA,ACACACTA] = 2

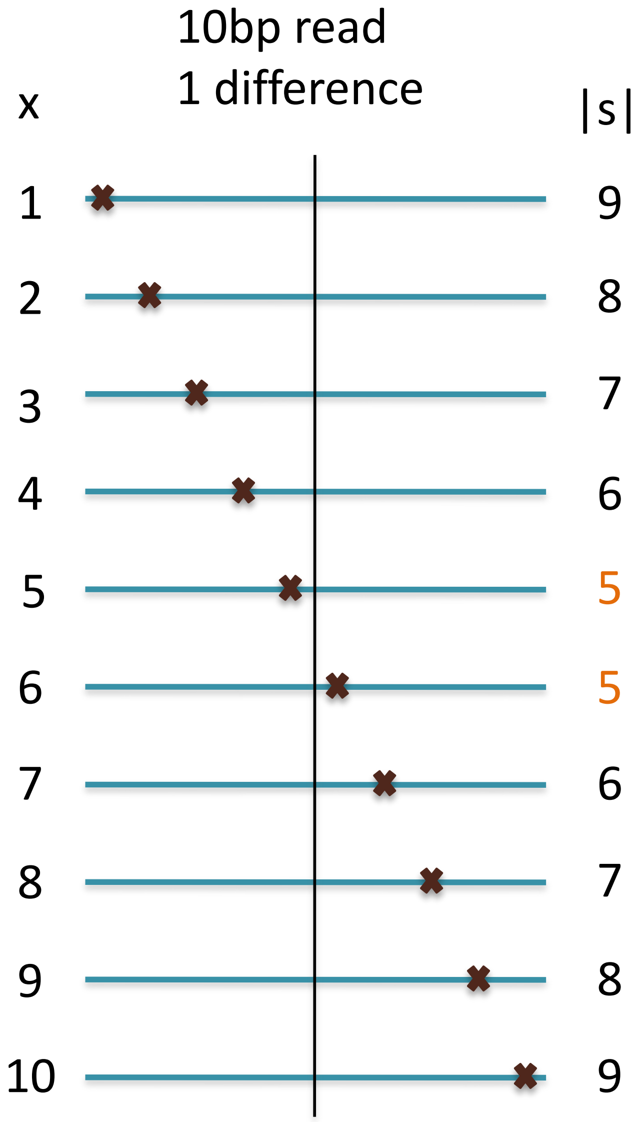
```
AGCACAC-A
|*|||||*|
A-CACACTA
```

[Can we do it any better?]

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)

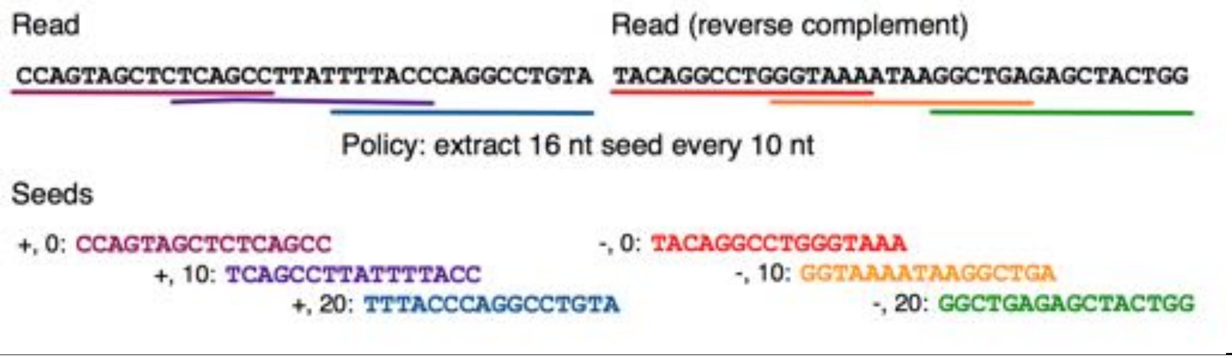


- Proof: Pigeonhole principle
  - 1 pigeon can't fill 2 holes

- 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

Algorithm Overview

1. Split read into segments



Read

CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA

Read (reverse complement)

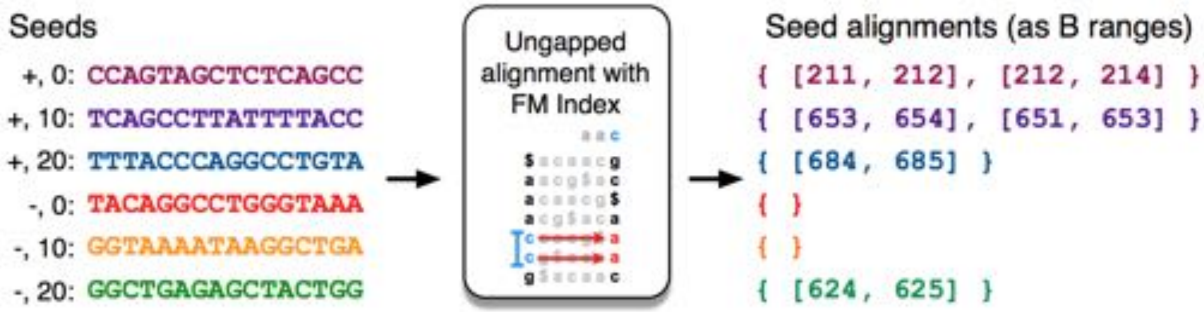
TACAGGCCTGGGTAAAATAAGGCTGAGAGCTACTGG

Policy: extract 16 nt seed every 10 nt

Seeds

```
+, 0: CCAGTAGCTCTCAGCC                          -, 0: TACAGGCCTGGGTAAA
        +, 10: TCAGCCTTATTTTACC                          -, 10: GGTAAAATAAGGCTGA
                 +, 20: TTTACCCAGGCCTGTA                          -, 20: GGCTGAGAGCTACTGG
```

2. Lookup each segment and prioritize



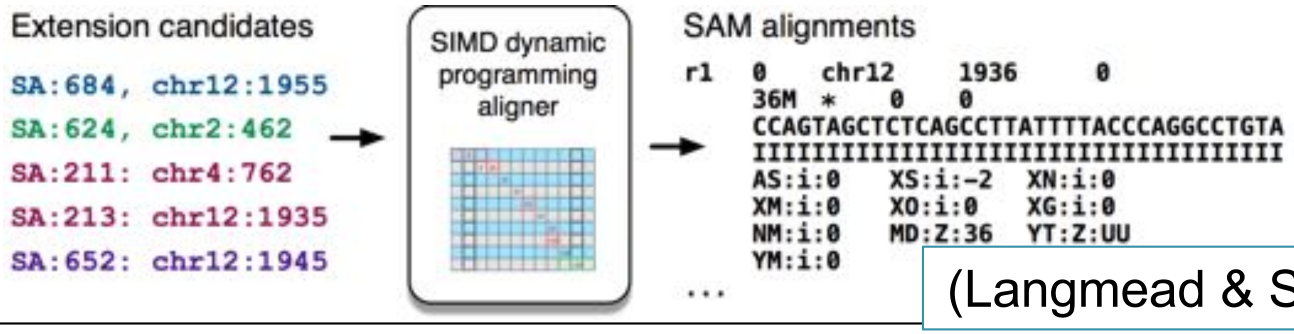
Seeds

```
+, 0: CCAGTAGCTCTCAGCC
+, 10: TCAGCCTTATTTTACC
+, 20: TTTACCCAGGCCTGTA
-, 0: TACAGGCCTGGGTAAA
-, 10: GGTAAAATAAGGCTGA
-, 20: GGCTGAGAGCTACTGG
```

Seed alignments (as B ranges)

```
{ [211, 212], [212, 214] }
{ [653, 654], [651, 653] }
{ [684, 685] }
{ }
{ }
{ [624, 625] }
```

3. Evaluate end-to-end match



Extension candidates

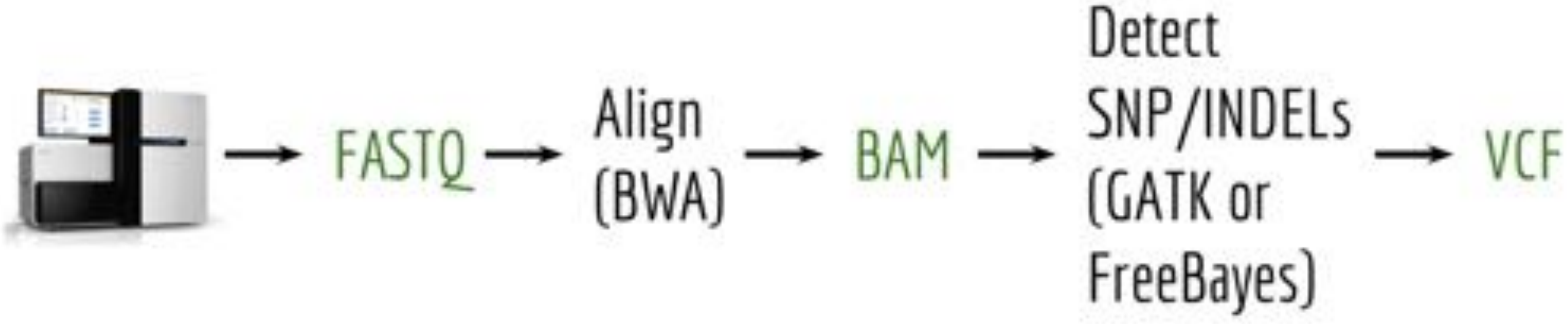
```
SA:684, chr12:1955
SA:624, chr2:462
SA:211: chr4:762
SA:213: chr12:1935
SA:652: chr12:1945
```

SAM alignments

```
r1   0    chr12    1936    0
     36M  *    0    0
     CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA
     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
     AS:i:0    XS:i:-2   XN:i:0
     XM:i:0    XO:i:0    XG:i:0
     NM:i:0    MD:Z:36   YT:Z:UU
     YM:i:0
...
```

(Langmead & Salzberg, 2012)

Variant Calling Overview


FASTQ
Align (BWA)
BAM
Detect SNP/INDELs (GATK or FreeBayes)
VCF