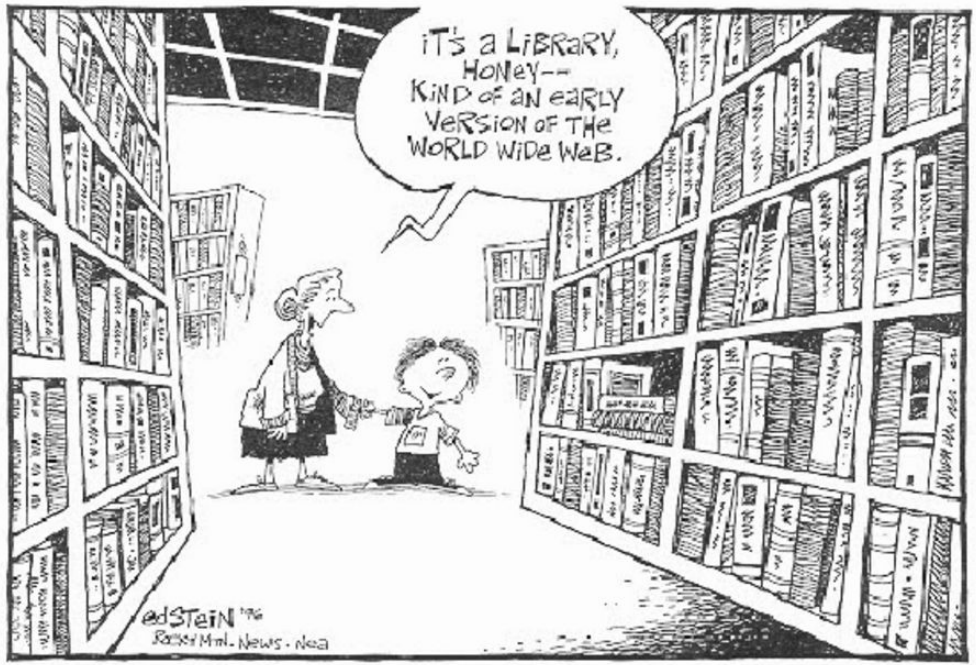
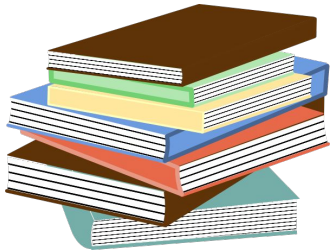
Comp 555 - BioAlgorithms - Spring 2022



- Problem set #3 is due 3/8
- Midterm is a week from Thursday on 3/3

Multi-String BWTs

MSBWT



A BWT of a ***string collection*** instead of just a single string

- Earliest: Mantaci et al. (2005), used concatenation approach
- Bauer et al. (2011) - proposed version we will discuss today

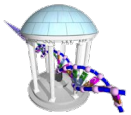
Analogy:

- Instead of searching for a substring within a single book, search every book of a library
  - Each book has its own text, suffix array, and end-of-text delimiter
  - Searching allows us to find how many times a substring appears and in which texts

Bioinformatics?

- Search all genomes? You could, but that's not the main application.
- Search multiple chromosomes of an organism?
  You should, but even that is not the killer app

Naive Construction



- Create all rotations for all strings in the collection
- Sort all rotations together (Suffix Array)
- Store the predecessor of each suffix
- Strings are “cyclic”
- The predecessor is always from the same string
- Impossible to “jump” from one string to another
- Strings can have different lengths

String1	String2
ACCA$	CAAA$
CCA$A	AAA$C
CA$AC	AA$CA
A$ACC	A$CAA
$ACCA	$CAAA

(Unsorted suffixes)

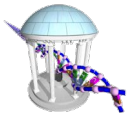
Sorted
$ACCA
$CAAA
A$ACC
A$CAA
AA$CA
AAA$C
ACCA$
CA$AC
CAAA$
CCA$A

(Merged and sorted)

MSBWT
A
A
C
A
A
C
$
C
$
A

(multi-string BWT
Note the 2 '$'s)

MSBWT's FM-index



Identical Definition

Identical Usage

- Find k-mer “CA”
- Initialize to full range (")
- lo, hi = 0, 10
- Find occurrences of 'A'
  - lo = Offset['A'] + FMindex[lo]['A'] = 2 + 0 = 2
  - hi = Offset['A'] + FMindex[hi]['A'] = 2 + 5 = 7
- Find occurrences of “CA”
  - lo = Offset['C'] + FMindex[lo]['C'] = 7 + 0 = 7
  - hi = Offset['C'] + FMindex[hi]['C'] = 7 + 2 = 9
- Searching and extracting suffixes are identical to a BWT

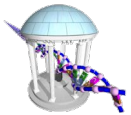
```
String1
ACCA$
CCA$A
CA$AC
A$ACC
$ACCA

String2
CAAA$
AAA$C
AA$CA
A$CAA
$CAAA
```

Sorted	MSBWT
$ACCA	A
$CAAA	A
A$ACC	C
A$CAA	A
AA$CA	A
AAA$C	C
ACCA$	$
CA$AC	C
CAAA$	$
CCA$A	A

FM-index

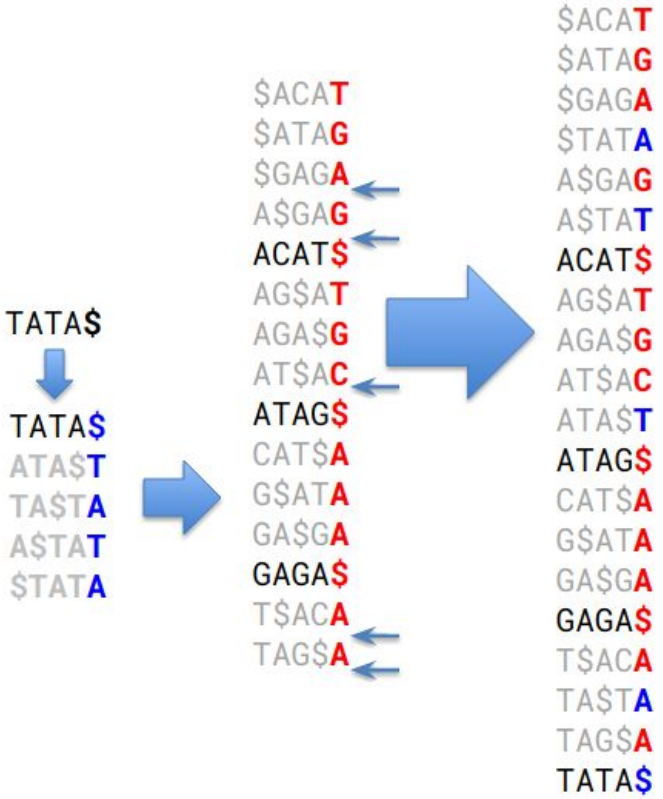
	$	A	C
0:	0	0	0
1:	0	1	0
2:	0	2	0
3:	0	2	1
4:	0	3	1
5:	0	4	1
6:	0	4	2
7:	1	4	2
8:	1	4	3
9:	2	4	3
10:	2	5	3
Offset:	0	2	7

Incremental MSBWT Construction

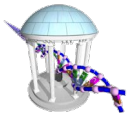


- A key tool missing from the BWTs toolbox--
  adding new strings to an existing msBWT

- You could reconstruct the suffix array of the msBWT using suffix(i, fmindex) for all i, and then insert the suffixes of the new string.
- Variant of find(); Find the insertion point of new string's j^{th} suffix, s_j
- Add last character to msBWT
- Update the FMindex



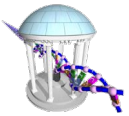
Our original BWT code



In [8]:

```
def FMIndex(bwt):
    fm = [{c: 0 for c in bwt}]
    for c in bwt:
        row = {symbol: count + 1 if (symbol == c) else count for symbol, count in fm[-1].items()}
        fm.append(row)
    offset = {}
    N = 0
    for symbol in sorted(row.keys()):
        offset[symbol] = N
        N += row[symbol]
    return fm, offset

def recoverSuffix(i, BWT, FMIndex, Offset):
    suffix = ''
    c = BWT[i]
    predec = Offset[c] + FMIndex[i][c]
    suffix = c + suffix
    while (predec != i):
        c = BWT[predec]
        predec = Offset[c] + FMIndex[predec][c]
        suffix = c + suffix
    return suffix

def findBWT(pattern, FMIndex, Offset):
    lo = 0
    hi = len(FMIndex) - 1
    for symbol in reversed(pattern):
        lo = Offset[symbol] + FMIndex[lo][symbol]
        hi = Offset[symbol] + FMIndex[hi][symbol]
    return lo, hi
```

Inserting a new BWT into an existing msBWT



In [9]:

```
# Constructing a multistring BWT one suffix at a time
# first let's take a look at the implicit suffix array
bwt = "TGAG$TGC$AAA$AA"
fm, off = FMIndex(bwt)
for i in range(len(bwt)):
    print("%2d: %s" % (i, recoverSuffix(i, bwt, fm, off)))
print()

# New string to include
new = "TATA$"
inserts = []
for i in range(len(new)):
    rotation = new[i:]+new[:i]
    l, h = findBWT(rotation,fm,off)
    inserts.append((h,new[i-1],rotation))
print(inserts)

# Insert into original BWT in reverse order
for i, c, rot in sorted(inserts, reverse=True):
    bwt = bwt[:i] + c + bwt[i:]

# Look at result
print(bwt)
print()
fm, off = FMIndex(bwt)
for i in range(len(bwt)):
    print("%2d: %s" % (i, recoverSuffix(i, bwt, fm, off)))
```

```
 0: $ACAT
 1: $ATAG
 2: $GAGA
 3: A$GAG
 4: ACAT$
 5: AG$AT
 6: AGA$G
 7: AT$AC
 8: ATAG$
 9: CAT$A
10: G$ATA
11: GA$GA
12: GAGA$
13: T$ACA
14: TAG$A
```

1. **How many strings are in this msBWT?**
2. **What strings are in this msBWT?**
3. **How could you have figured this out from the msBWT alone?**

Before and After

```
[(15, '$', 'TATA$'), (8, 'T', 'ATA$T'), (14, 'A', 'TA$TA'), (4, 'T', 'A$TAT'), (3, 'A', '$TATA')]
TGAAGT$TGCT$AAA$AAA$
```

```
                0: $ACAT
                1: $ATAG
$TATA           2: $GAGA
A$TAT           3: A$GAG
                4: ACAT$
                5: AG$AT
                6: AGA$G
ATA$T           7: AT$AC
                8: ATAG$
                9: CAT$A
               10: G$ATA
               11: GA$GA
               12: GAGA$
TA$TA          13: T$ACA
TATA$          14: TAG$A
```

```
 0: $ACAT
 1: $ATAG
 2: $GAGA
 3: $TATA
 4: A$GAG
 5: A$TAT
 6: ACAT$
 7: AG$AT
 8: AGA$G
 9: AT$AC
10: ATA$T
11: ATAG$
12: CAT$A
13: G$ATA
14: GA$GA
15: GAGA$
16: T$ACA
17: TA$TA
18: TAG$A
19: TATA$
```

Problem: Add the string "TATA$" to an existing msBWT.

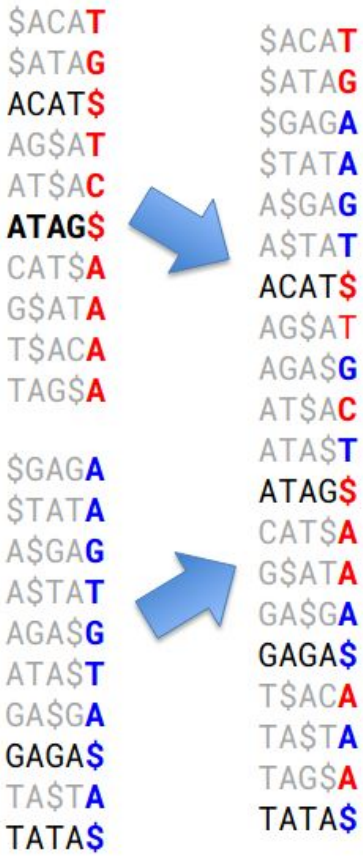
Implies 5 new suffixes: TATA$, ATA$T, TA$TA, A$TAT, $TATA

Insert these into existing msBWT. Use BWT search to find all, all insert positions first, then insert from highest to lowest index.

That seems a little tricky...

Merging msBWTs

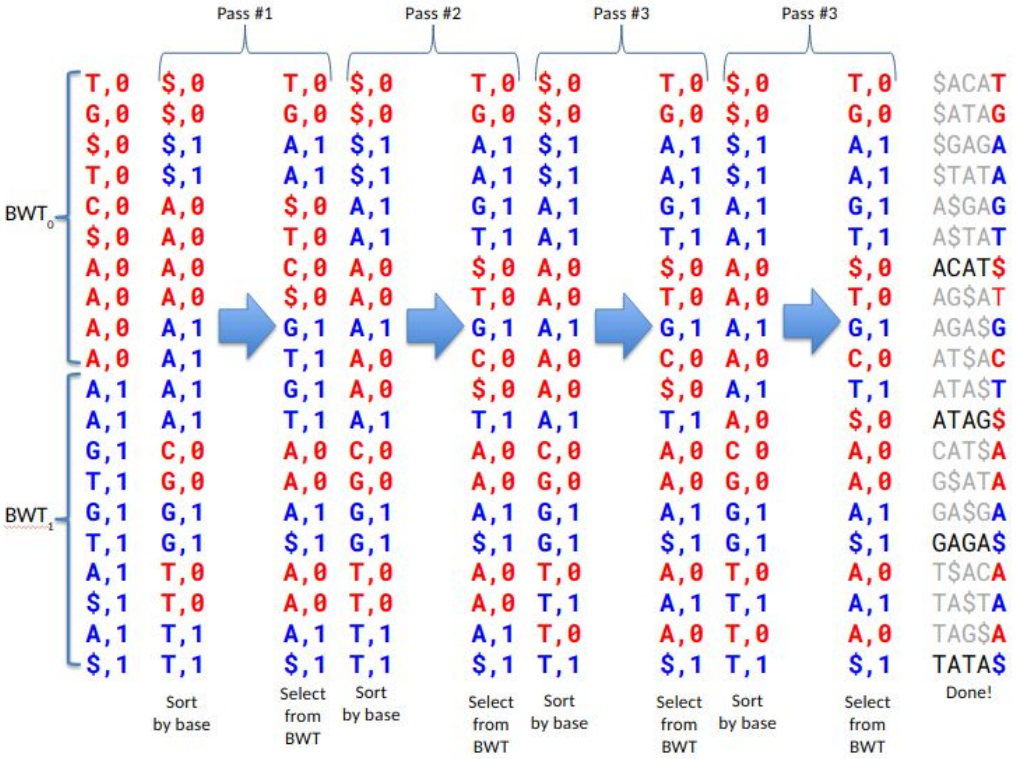
- BETTER YET! Rather than inserting new strings, build a BWT of the new strings and merge the new and old BWTs
- Suffixes of BTWs are already sorted
- BTWs are interleaved
- In the worse case (ties) the entire suffix must be considered, but general the longest common prefix of suffixes is smaller
- Minimal overhead
- Well suited for divide an conquer approaches (like merge sort)
- Easy to merge multiple data sets!
- Compression improves!



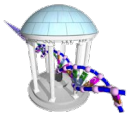
Merging Steps

msBWT merging alternates between sorting and interleaving

1. Consider the BWTs as a tuple of (character, BWTid) pairs
2. Sort these tuples
3. Based on the BWTids after the sort, select a new character For each tuple from the original msBWTs
4. Repeat from Step 2 until the sort is stable
5. The resulting characters are the merged msBWT
6. Number of passes is proportional to largest LCP value.



In Python

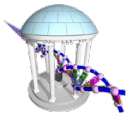


In [12]:

```
def mergeBWT(bwt1, bwt2):
    interleave = [(c, 0) for c in bwt1] + [(c, 1) for c in bwt2]
    passes = min(len(bwt1), len(bwt2))
    for p in range(passes):
        i, j = 0, 0
        nextInterleave = []
        for c, k in sorted(interleave, key=lambda x: x[0]):
            if (k == 0):
                b = bwt1[i]
                i += 1
            else:
                b = bwt2[j]
                j += 1
            nextInterleave.append((b, k))
        if (nextInterleave == interleave):
            break
        interleave = nextInterleave
    return ''.join([c for c, k in interleave])

bwt1 = "TG$TC$AAAA"
bwt2 = "AAGTGTA$A$"
bwt12 = mergeBWT(bwt1, bwt2)
print(bwt12)
FM, Offset = FMIndex(bwt12)
for i in range(len(bwt12)):
    j = (i>>2)+(i&3)*(len(bwt12)//4)
    print("%2d: %s" % (j, recoverSuffix(j,bwt12,FM,Offset)), "\n" if (i % 4 == 3) else "", end='')
```

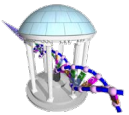
```
TGAAGT$TGCT$AAA$AAA$
 0: $ACAT  5: A$TAT 10: ATA$T 15: GAGA$
 1: $ATAG  6: ACAT$ 11: ATAG$ 16: T$ACA
 2: $GAGA  7: AG$AT 12: CAT$A 17: TA$TA
 3: $TATA  8: AGA$G 13: G$ATA 18: TAG$A
 4: A$GAG  9: AT$AC 14: GA$GA 19: TATA$
```

MSBWT Applications



- Instead of building a BWT of a reference genome, build a MSBWT of every sequenced reads
- Arbitrary exact-match k-mer queries
- O(k) time
- Enables fast searches/counting
- Recover an arbitrary read of length L from MSBWT
- O(L) time
- Enables extraction of user-selected reads

Compression of high-throughput sequencing



- Using Run-length encoding again
- Reasons we expect compression:
  - True genomic repeats: gene families, long repeats, etc.
  - Over-sampling: 30x coverage means we expect 30 copies of every k-mer pattern
- Sequencing errors may break up runs
- Technical errors may cause biases for or against a particular pattern
- Real Mouse DNA-seq:
  - 368654191 × 151 × 2 = ~112 Giga-bases
  - Compresses to ~15.3 GB using RLE (1.09 bits/base)
- Real Mouse RNA-seq:
  - ~8.9 Giga-bases
  - ~1.2 GB using RLE (1.05 bits/base)

K-mer Search & Read Extraction

Basic Use:

CC MRCAs: CC010_MRCA_2020
730,665,362 strings with 110,330,469,662 bases and index size of 14,976,208,773 bytes (1.09 bits per base)
Target: CTTGTCCTTGGAAGGAAGATATATG
Found 27 times (15 forward, 12 reverse-complemented)

Inconsistent read size: reported = 151, found = 151

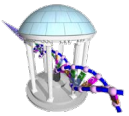
Shown below are all instances of the k-mer CTTGTCCTTGGAAGGAAGATATATG (red) or its reverse-complement (blue) within a read

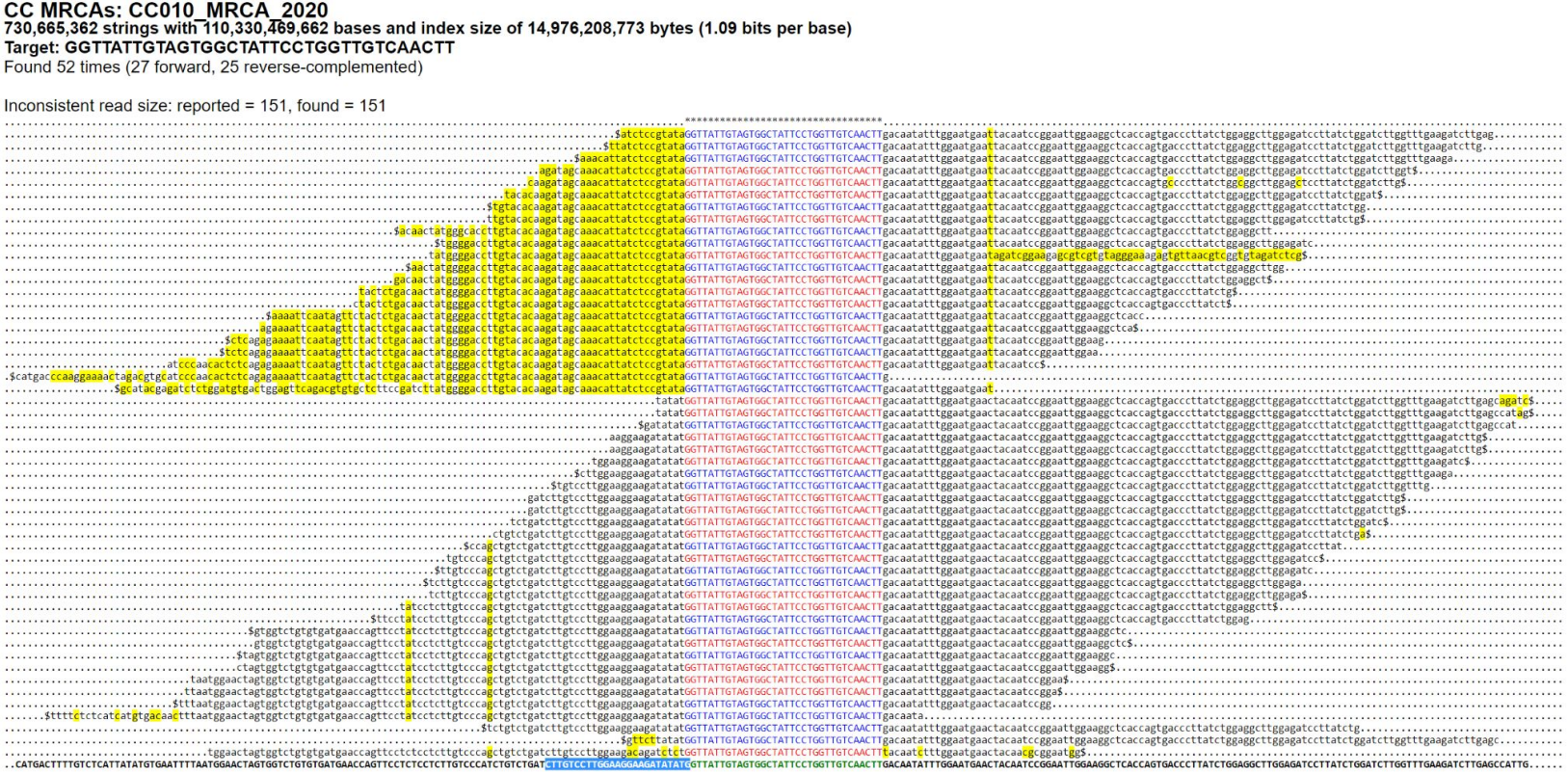
Green: query k-mer. Red: forward reads. Blue: reverse-complement reads. Yellow: sequencing errors

Shown below is the consensus base at each position

- Search for all reads with a given k-mer
- Extract all reads with that k-mer and its reverse-complement
- Build a consensus

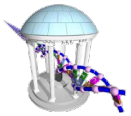
Moving over a little bit



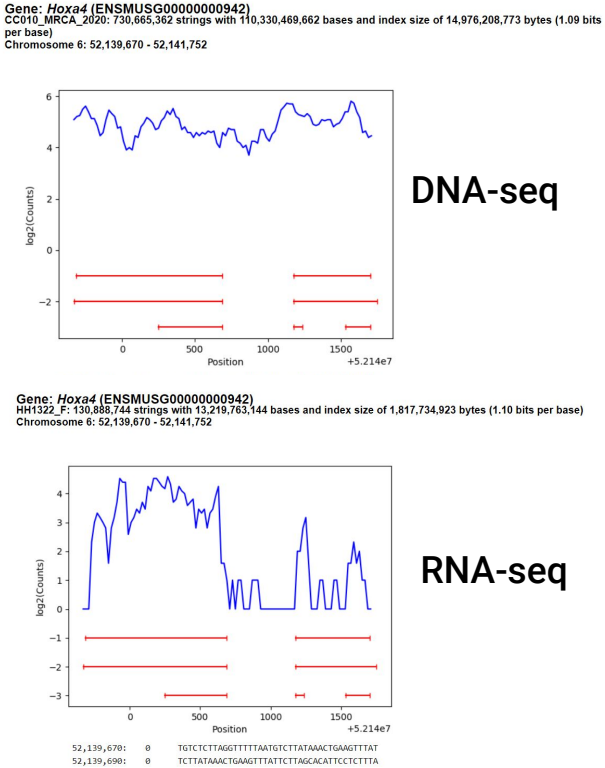


Moving over 24 bases shows a different story

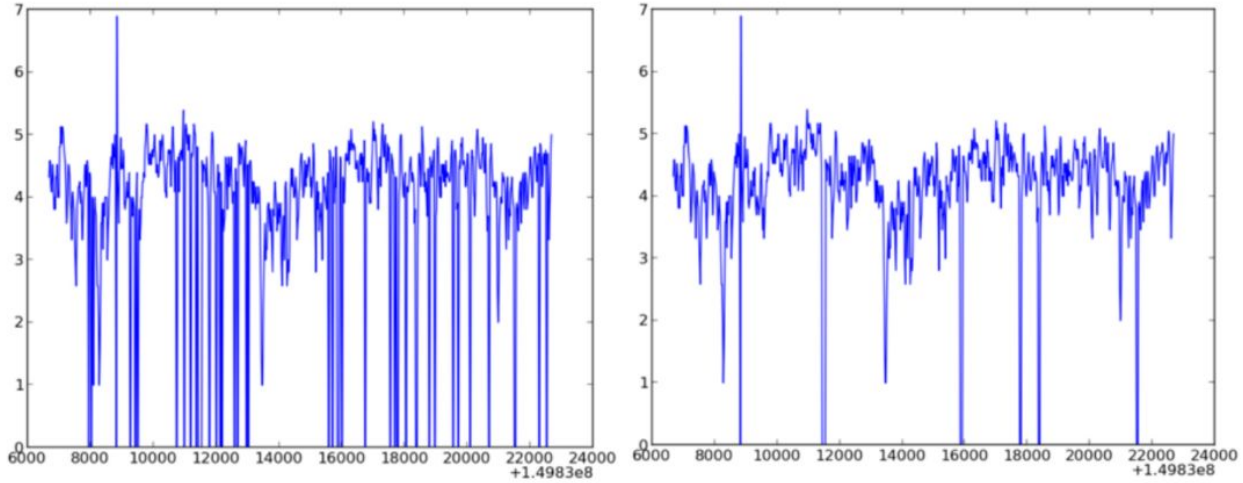
Reference-based Searches



- Given a reference genome and region of that genome
- Split reference into k-mers
- Count the abundance of each k-mer and plot
- Fast - O(k) time per k-mer
- Similar to a post-alignment pileup

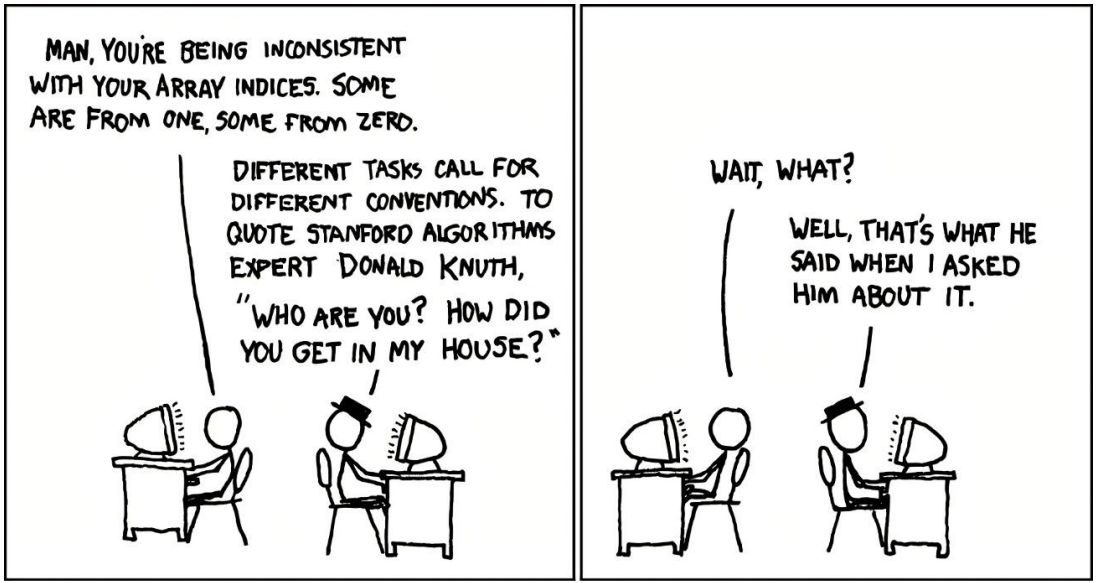


Iterative Reference Correction



CAST/EiJ DNA-seq for annotated gene *Igf2*

Next Time



We take a deep dive into Dynamic Programming