Genome Assembly

Michael Schatz
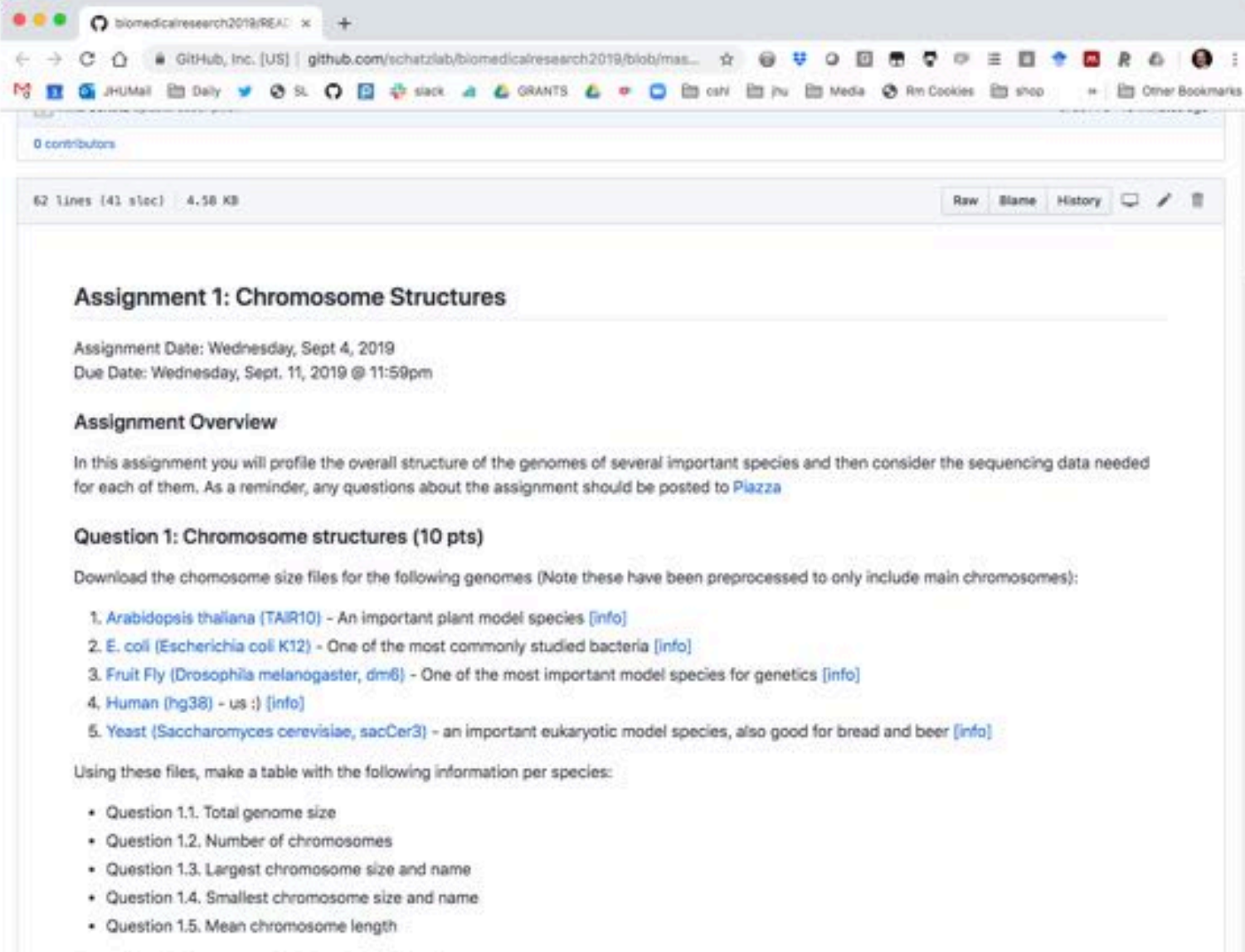
Assignment 1: Chromosome Structures
Due Wed Sept 11 @ 11:59pm

https://github.com/schatzlab/biomedicalresearch2019

Part 1: Recap

Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

>3Tbp / day
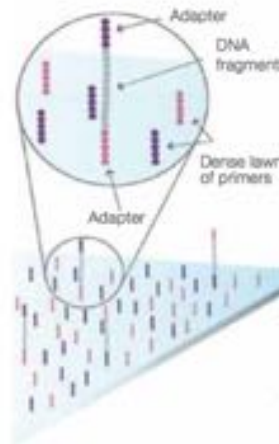
1. Attach
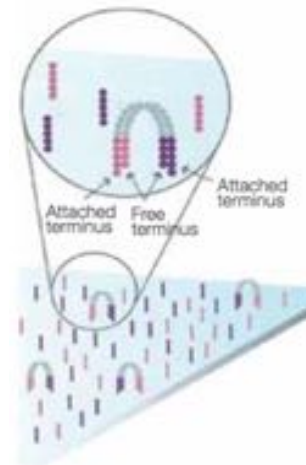
2. Amplify
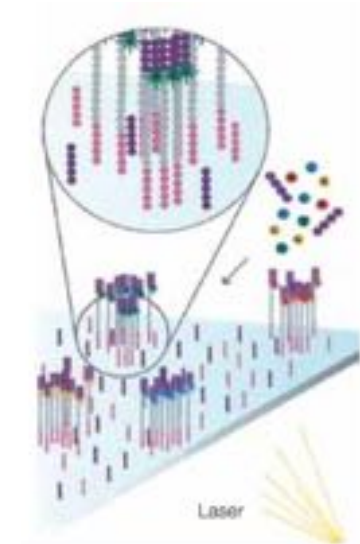
3. Image
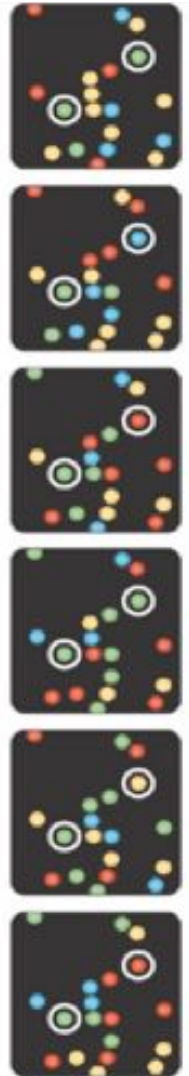
Metzker (2010) Nature Reviews Genetics 11:31-46
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Typical sequencing coverage
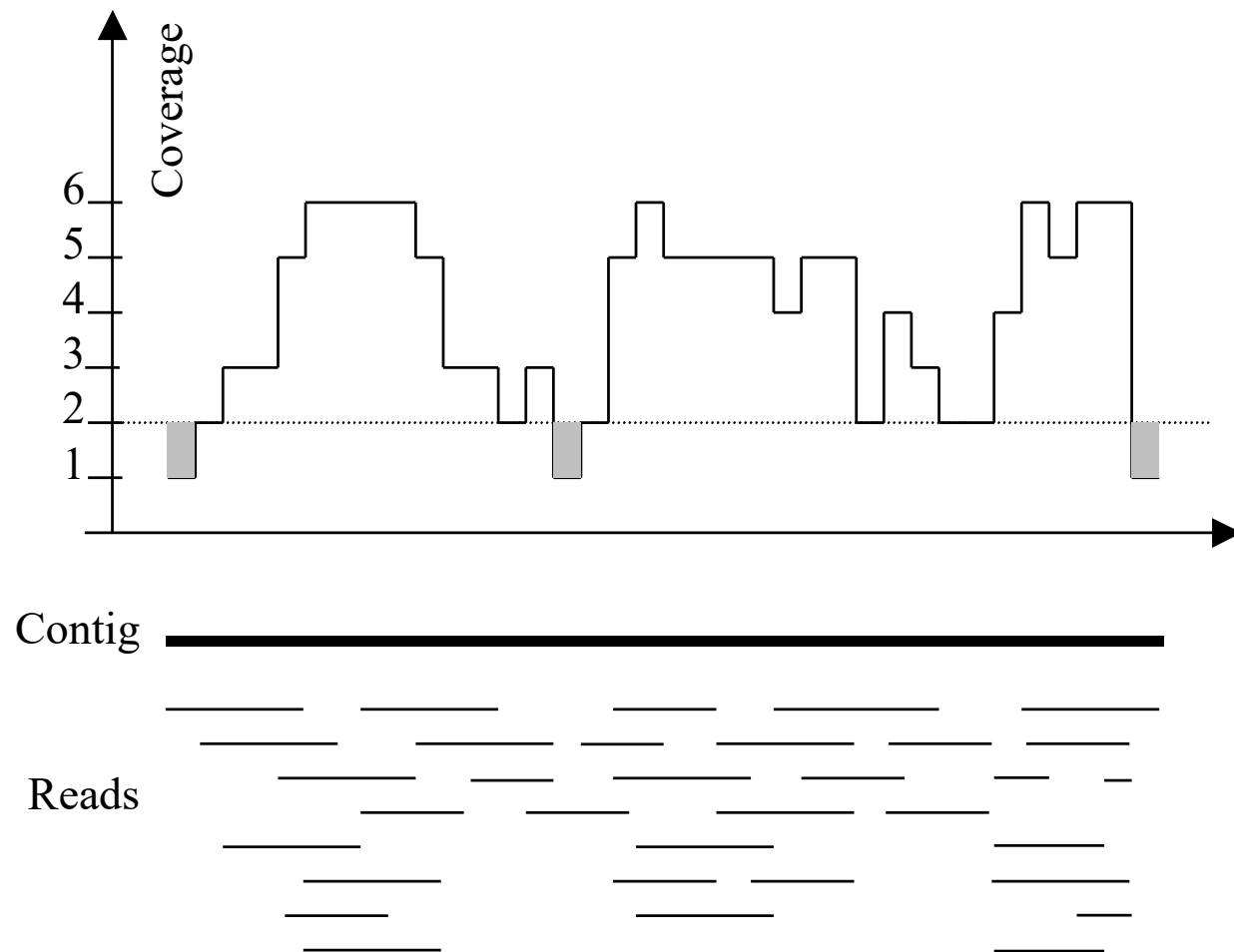
Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs $1

If the genome is 10 Mbp, should we sequence 100k 100bp reads?

8x sequencing

Balls in Bins
Total balls: 8000

Histogram of balls in each bin
Total balls: 8000 Empty bins: 1
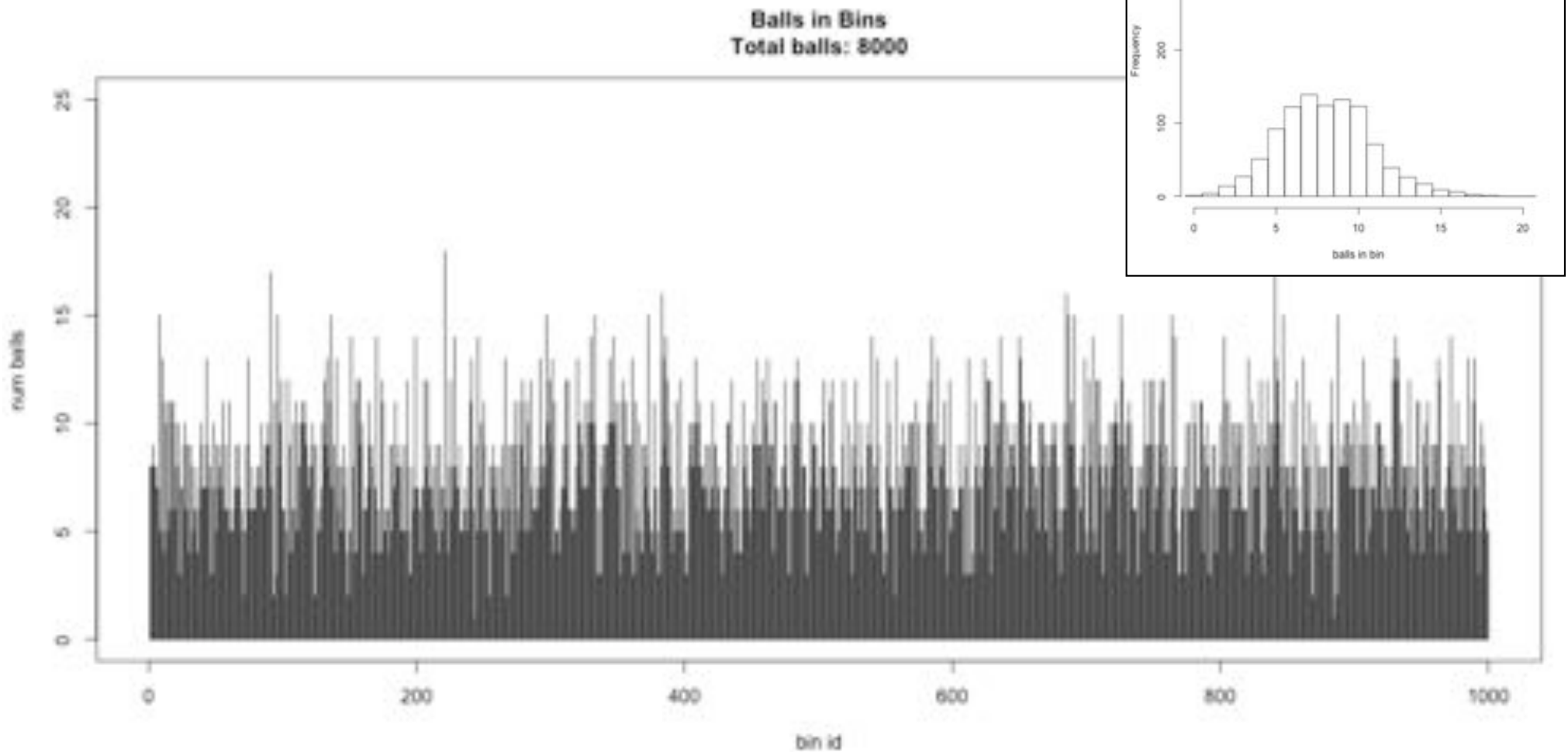
Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

Formulation comes from the limit of the binomial equation

Resembles a normal distribution, but over the positive values, and with only a single parameter.

Key properties:
- *The standard deviation is the square root of the mean.*
- *For mean > 5, well approximated by a normal distribution*

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Normal Approximation

99.7% of the data are within
3 standard deviations of the mean

95% within
2 standard deviations

68% within
1 standard deviation

$\mu - 3\sigma$ $\mu - 2\sigma$ $\mu - \sigma$ μ $\mu + \sigma$ $\mu + 2\sigma$ $\mu + 3\sigma$
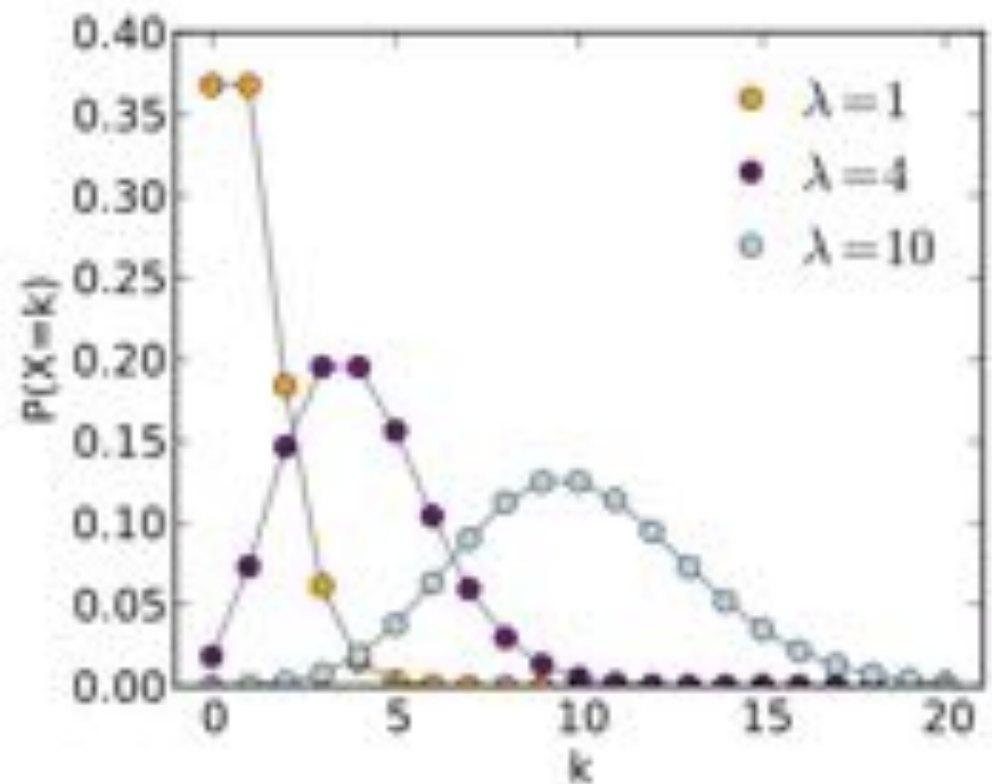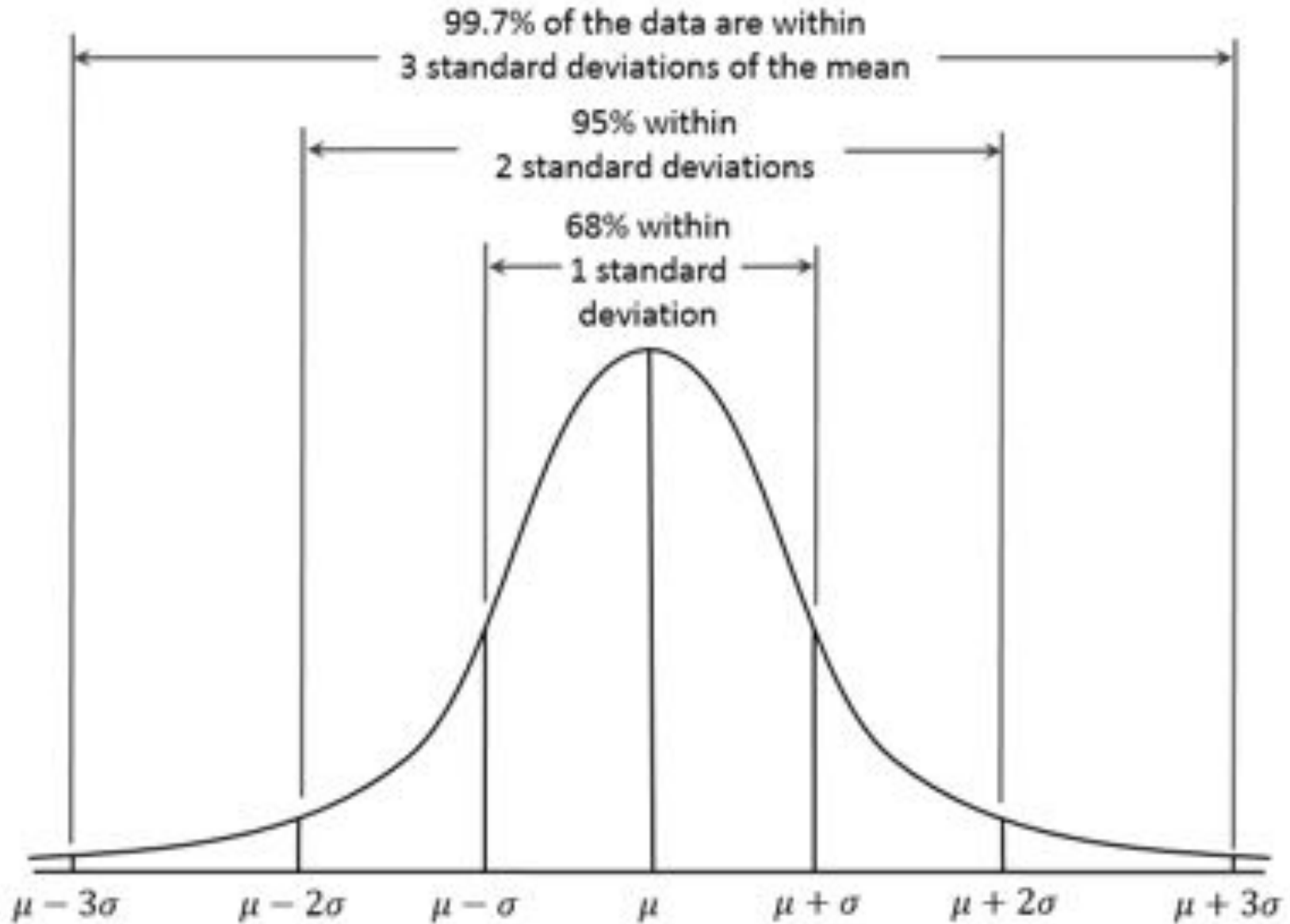
Can estimate Poisson distribution as a normal distribution when $\lambda > 10$

Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 120bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 120bp / read = 2M reads

Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 120bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 120bp / read = 2M reads

I want to sequence a 10Mbp genome so that
>97.5% of the genome has at least 24x coverage.
How many 120bp reads do I need?

Find X such that X-2*sqrt(X) = 24

36-2*sqrt(36) = 24

I need 10Mbp x 36x = 360Mbp of data
360Mbp / 120bp / read = 3M reads

Illumina Sequencing Summary

Advantages:

- Best throughput, accuracy and read length for any 2nd gen. sequencer

- Fast & robust library preparation

Disadvantages:

- Inherent limits to read length (practically, 150bp)

- Some runs are error prone

- Requires amplification, sequences a population of molecules

Illumina HiSeq
~3 billion paired 100bp reads
~600Gb, $10K, 8 days
(or "rapid run" ~90Gb in 1-2 days)

Illumina X Ten
~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome($$)
(or "rapid run" ~90Gb in 1-2 days)

Illumina NextSeq
One human genome in **<30 hours**

Ira Hall

Market Summary > Illumina, Inc.
NASDAQ: ILMN

270.46 USD -10.89 (3.87%) ↓
Sep 3, 2:37 PM EDT · Disclaimer

| 1 day | 5 days | 1 month | 6 months | YTD | 1 year | 5 years | Max |

19.58 USD Jul 28, 2000

Open	279.88	Div yield	-
High	280.36	Prev close	281.34
Low	270.24	52-wk high	380.76
Mkt cap	39.78B	52-wk low	268.62
P/E ratio	43.17		

→ Financial news, comparisons and more

Part 2: De novo genome assembly

Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of <u>A Tale of Two Cities</u>
  - Text printed on 5 long spools

- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction

It was the best of

age of wisdom, it was

best of times, it was

it was the age of

it was the age of

it was the worst of

of times, it was the

of times, it was the

of wisdom, it was the

the age of wisdom, it

the best of times, it

the worst of times, it

times, it was the age

times, it was the worst

was the age of wisdom,

was the age of foolishness,

was the best of times,

was the worst of times,

wisdom, it was the age

worst of times, it was

It was the best of

was the best of times,

the best of times, it

best of times, it was

of times, it was the

of times, it was the

times, it was the worst

times, it was the age

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

How long will it take to compute the overlaps?

de Bruijn Graph Construction

- $G_k = (V,E)$
  - V = Length-*k* sub-fragments
  - E = Directed edges between consecutive sub-fragments
    - Sub-fragments overlap by *k*-1 words

Fragments |f|=5   Sub-fragment *k*=4   Directed edges (overlap by *k*-1)

| It was the best of |

| was the best of times |

| It was the best | | was the best of |

| was the best of | | the best of times |

| It was the best |

| was the best of |

| the best of times |

– Overlaps between fragments are implicitly computed

de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

de Bruijn Graph Assembly

It was the best

was the best of

the best of times,

best of times, it

of times, it was

times, it was the

it was the worst

was the worst of

the worst of times,

worst of times, it

it was the age

was the age of

the age of foolishness

the age of wisdom,

age of wisdom, it

of wisdom, it was

wisdom, it was the

After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly

It was the best of times, it

of times, it was the

it was the worst of times, it

it was the age of

the age of foolishness

the age of wisdom, it was the

After graph construction, try to simplify the graph as much as possible

The full tale

… it was the best of times it was the worst of times …

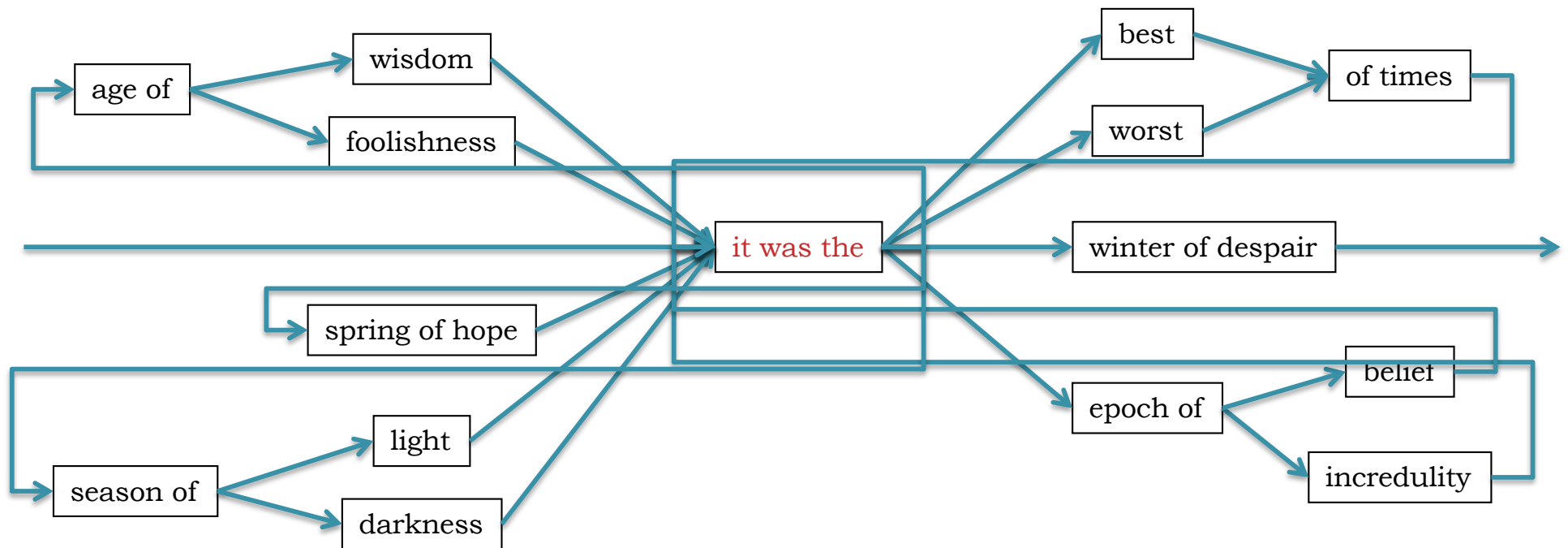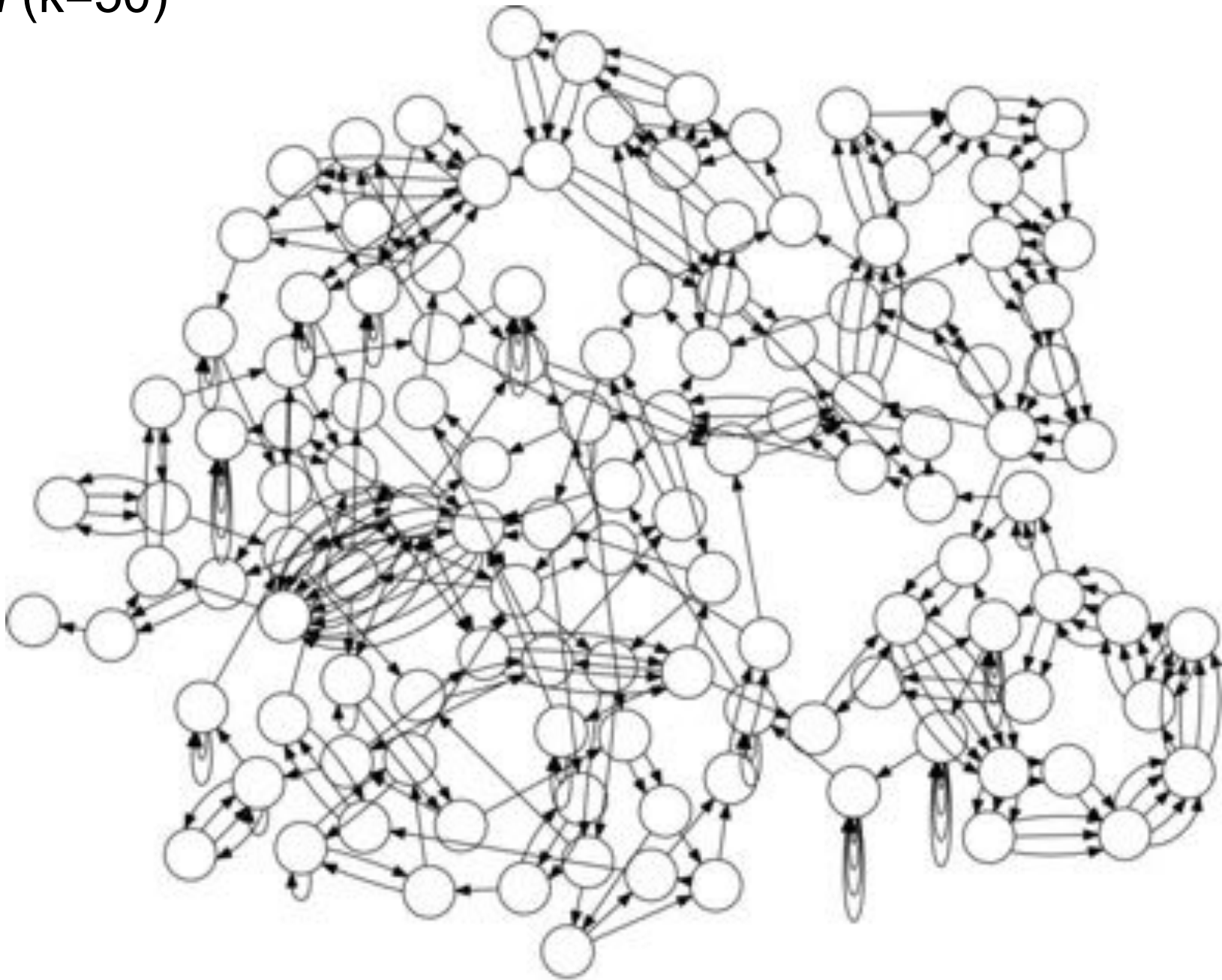… it was the age of wisdom it was the age of foolishness …

… 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 winder of despair …

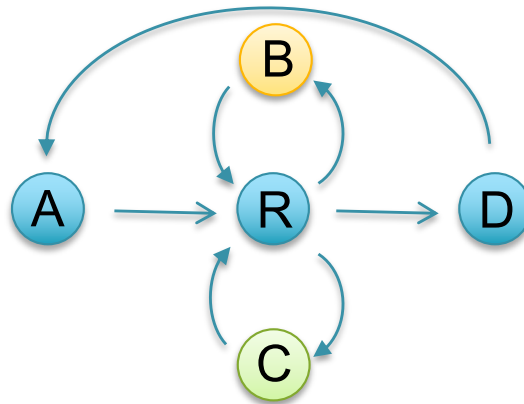# E. coli (k=50)

Counting Eulerian Cycles

ARBRCRD
or
ARCRBRD

Generally an exponential number of compatible sequences

– Value computed by application of the BEST theorem (Hutchinson, 1975)

$$W(G,t) = (\det L)\left\{\prod_{u \in V}(r_u - 1)!\right\}\left\{\prod_{(u,v) \in E} a_{uv}!\right\}^{-1}$$
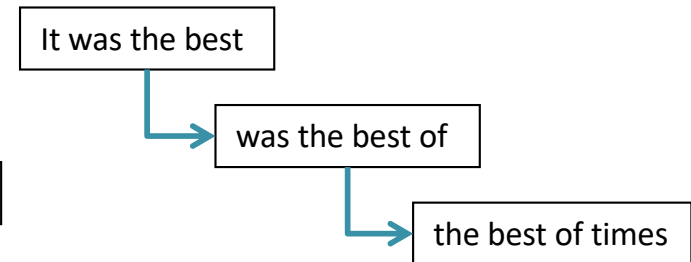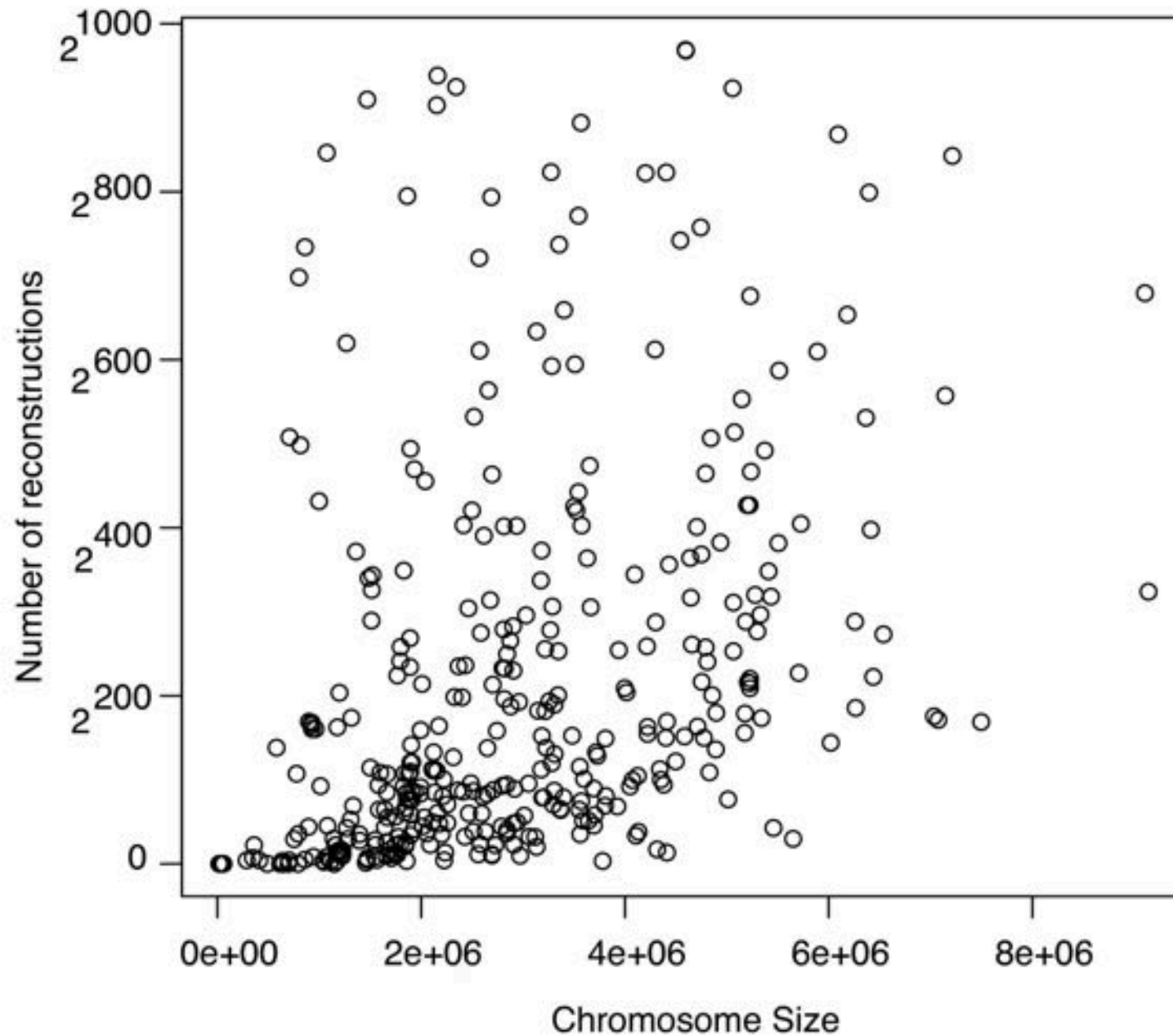
L = n x n matrix with r_u-a_{uu} along the diagonal and -a_{uv} in entry uv

r_u = $d^+(u)$+1 if u=t, or $d^+(u)$ otherwise

a_{uv} = multiplicity of edge from u to v

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics.*
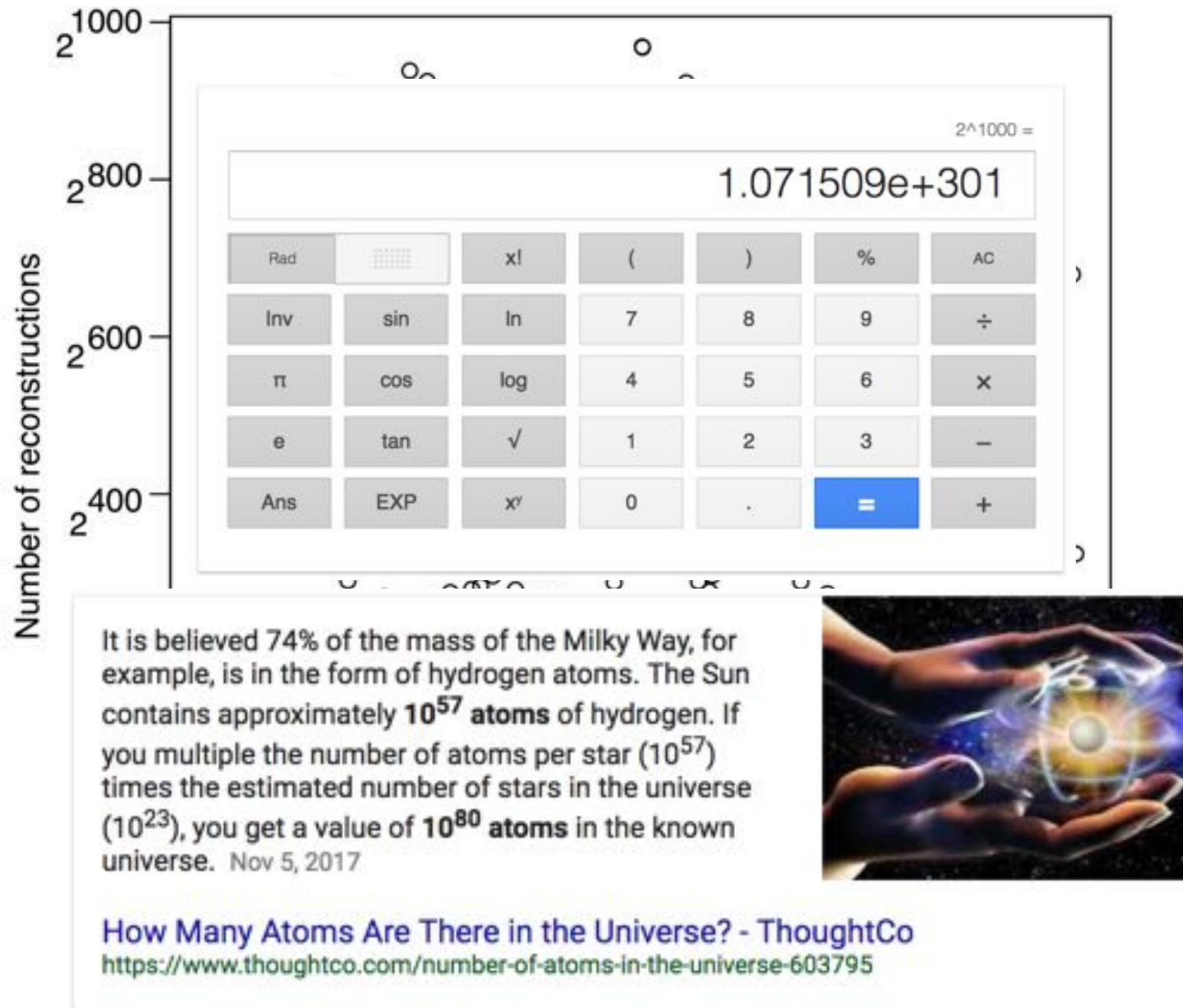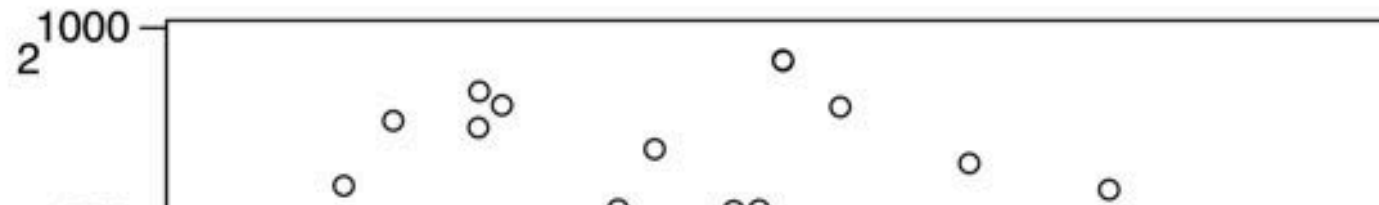
Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics.*

- *Finding possible assemblies is easy!*

- *However, there is an ~~astronomical~~ genomical number of possible paths!*

- *Hopeless to figure out the whole genome/chromosome, figure out the parts that you can*

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics.*

Contig N50

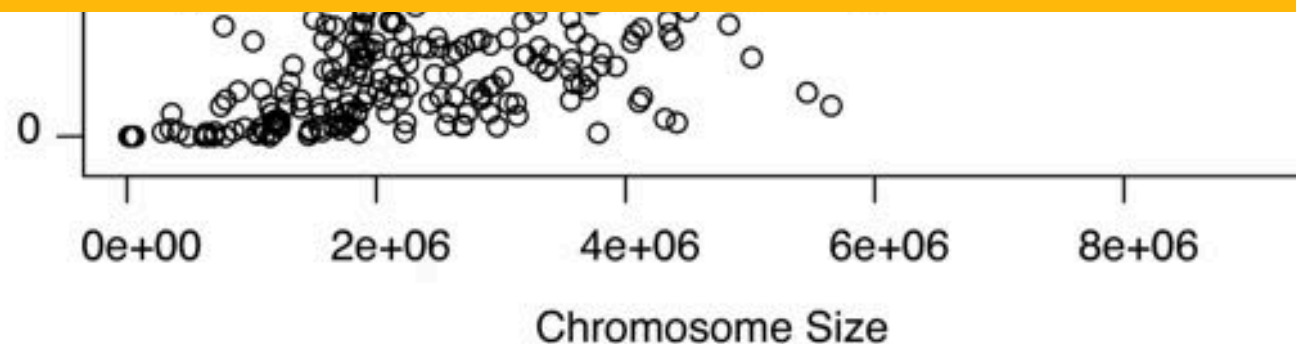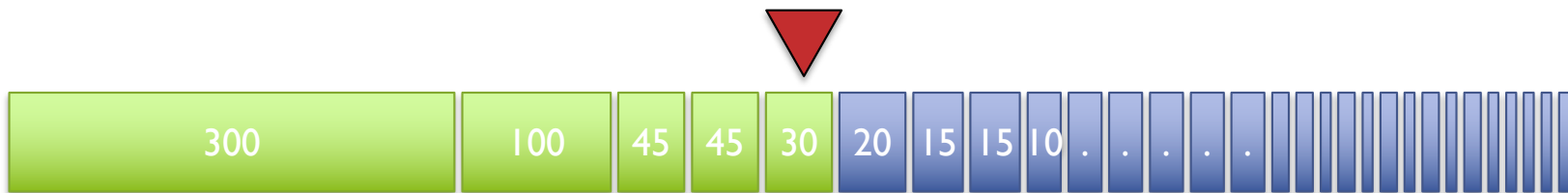Def: 50% of the genome is in contigs as large as the N50 value

Better N50s improves the analysis in every dimension

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

Just be careful of N50 inflation!

- *A very very very bad assembler in 1 line of bash:*
- *cat *.reads.fa > genome.fa*

N50 size = 3 kbp

Pop Quiz 1

Assemble these reads using a de Bruijn graph approach (k=3):

ATTA

GATT

TACA

TTAC

Pop Quiz 1

Assemble these reads using a de Bruijn graph approach (k=3):
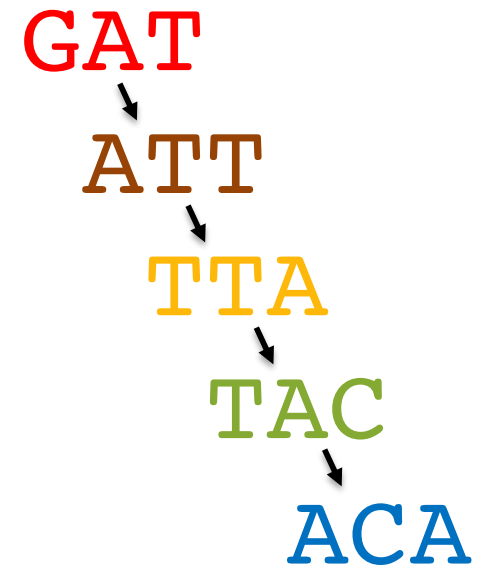
```
ATTA: ATT -> TTA
GATT: GAT -> ATT
TACA: TAC -> ACA
TTAC: TTA -> TAC
```

Pop Quiz 1

Assemble these reads using a de Bruijn graph approach (k=3):

ATTA: ATT -> TTA

GATT: GAT -> ATT

TACA: TAC -> ACA

TTAC: TTA -> TAC

GAT
↓
ATT
↓
TTA
↓
TAC
↓
ACA

GATTACA

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

ACGA

ACGT

ATAC

CGAC

CGTA

GACG

GTAT

TACG

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~

ACGT

ATAC

CGAC

CGTA

GACG

GTAT

TACG

CGA

ACG

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~

~~ACGT~~

ATAC

CGAC

CGTA

GACG

GTAT

TACG

```
        CGA
       ↗
ACG
       ↘
        CGT
```

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~

~~ACGT~~

ATAC

~~CGAC~~

CGTA

GACG

GTAT

TACG

                                    GAC
                                   ↗
                              CGA
                             ↗
                        ACG
                             ↘
                              CGT

Pop Quiz 2

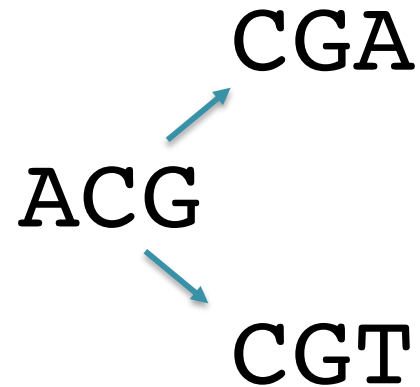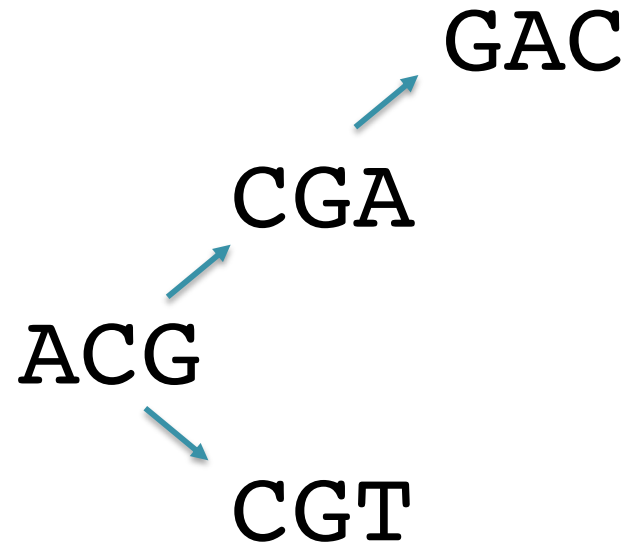Assemble these reads using a de Bruijn graph approach (k=3):
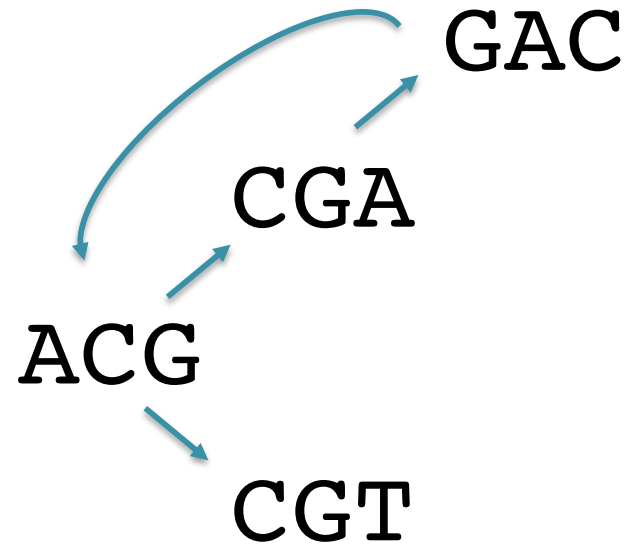
~~ACGA~~

~~ACGT~~

ATAC

~~CGAC~~

CGTA

~~GACG~~

GTAT

TACG

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~

~~ACGT~~

ATAC

~~CGAC~~

~~CGTA~~

~~GACG~~

GTAT

TACG

GAC

CGA

ACG

CGT

GTA

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~

~~ACGT~~

ATAC

~~CGAC~~

~~CGTA~~

~~GACG~~

~~GTAT~~
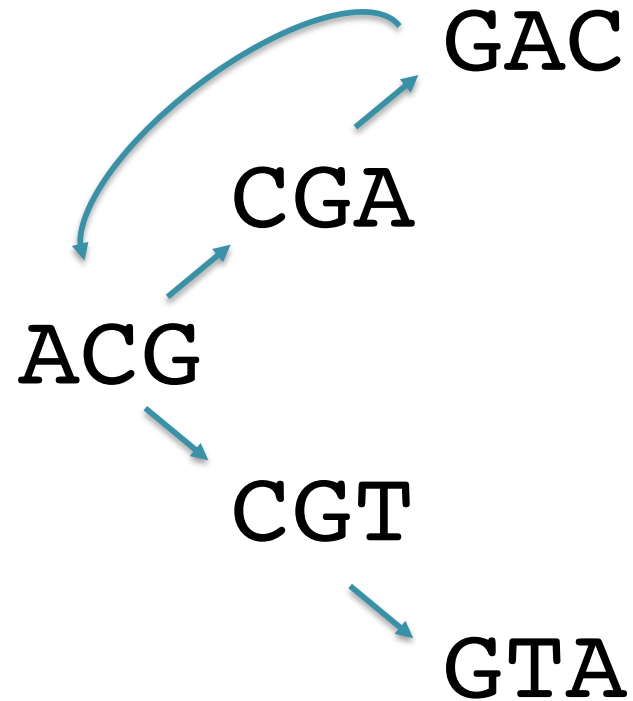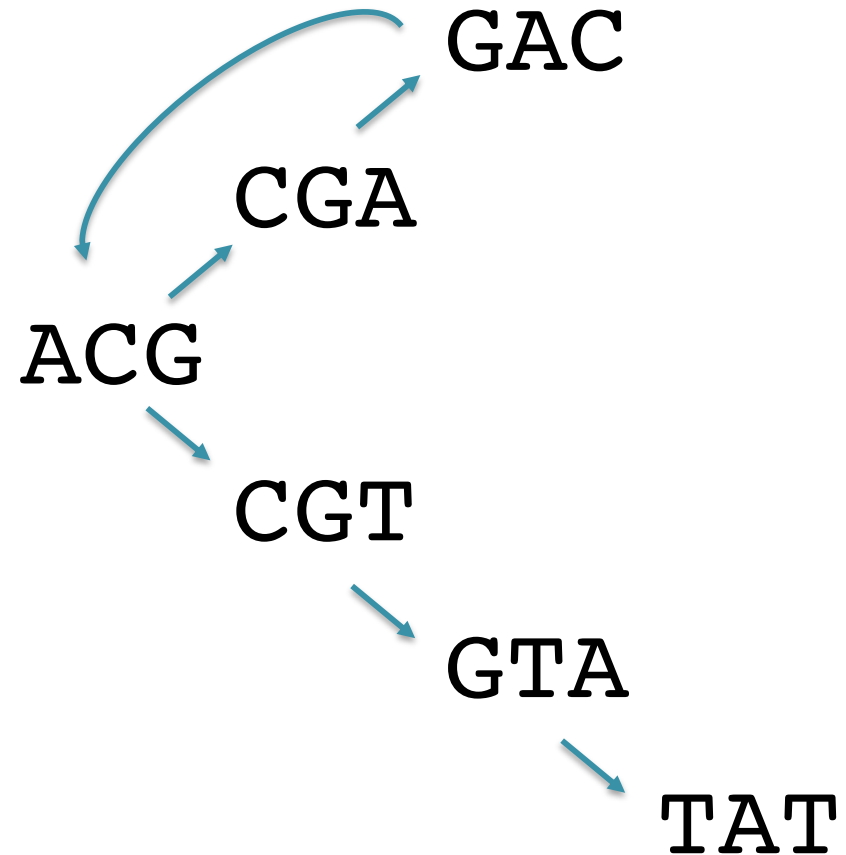
TACG

GAC

CGA

ACG

CGT

GTA

TAT

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):
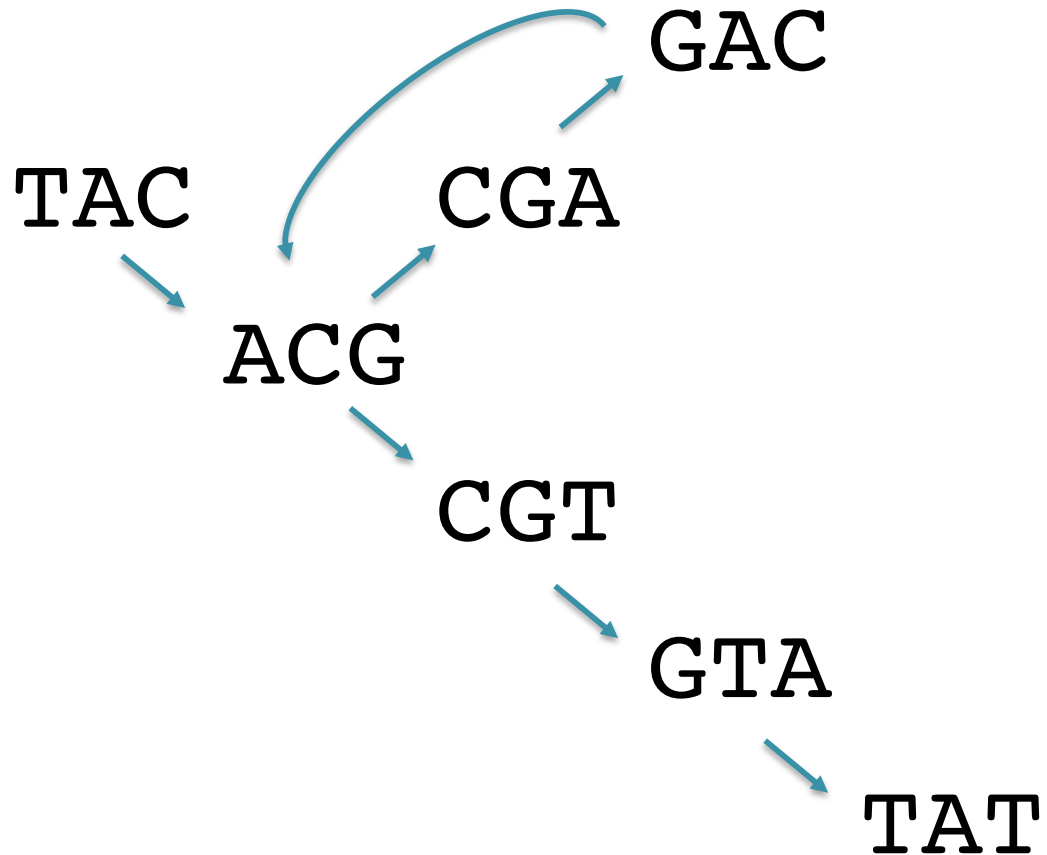
~~ACGA~~

~~ACGT~~

ATAC

~~CGAC~~

~~CGTA~~

~~GACG~~

~~GTAT~~

~~TACG~~

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):
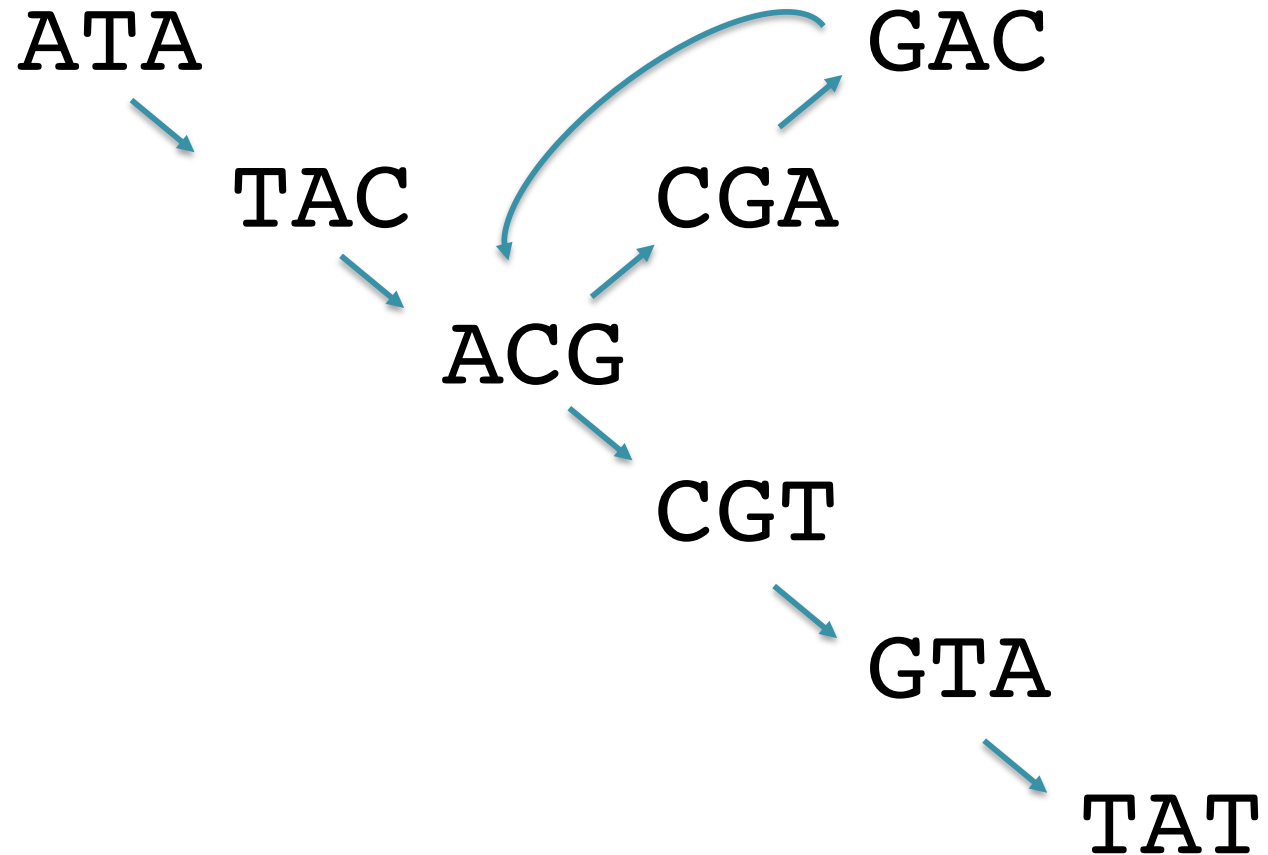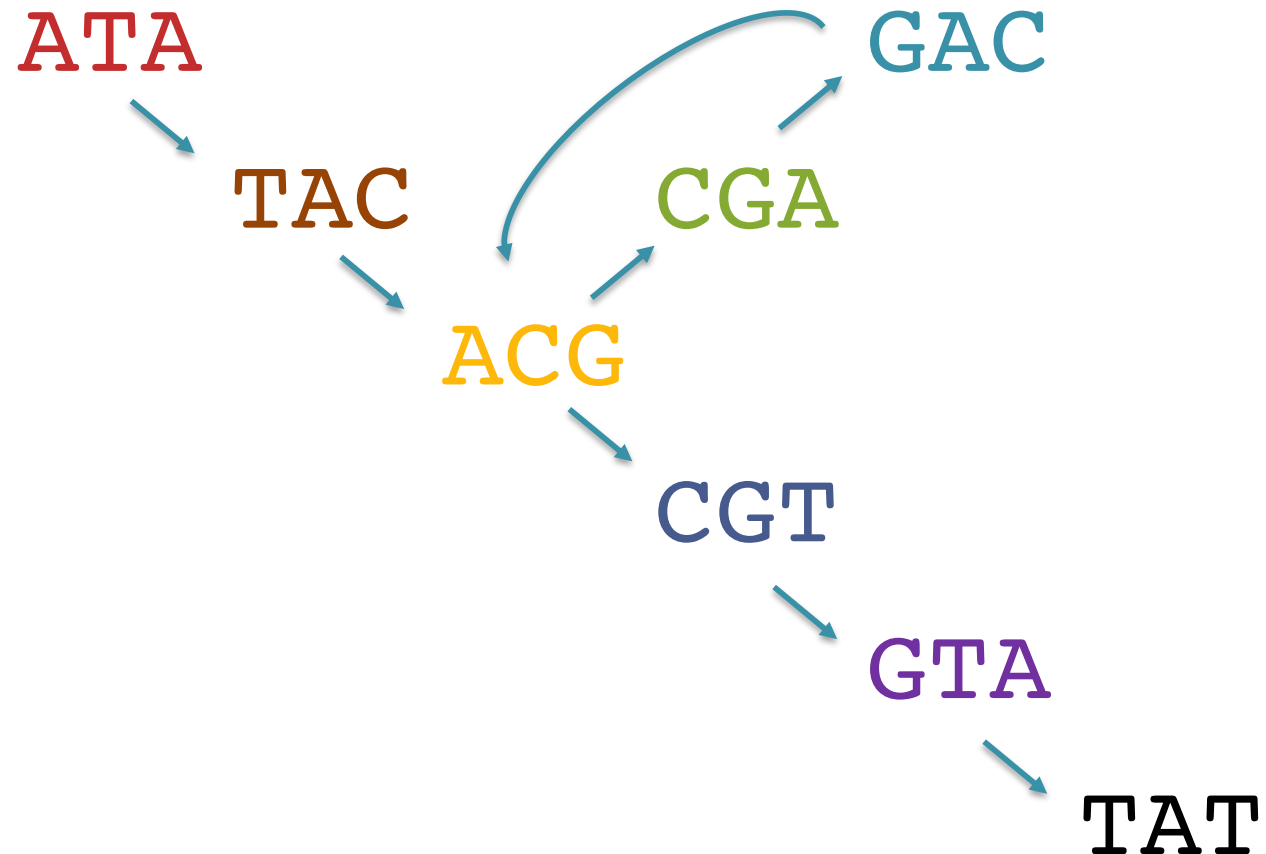
~~ACGA~~

~~ACGT~~

~~ATAC~~

~~CGAC~~

~~CGTA~~

~~GACG~~

~~GTAT~~

~~TACG~~

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
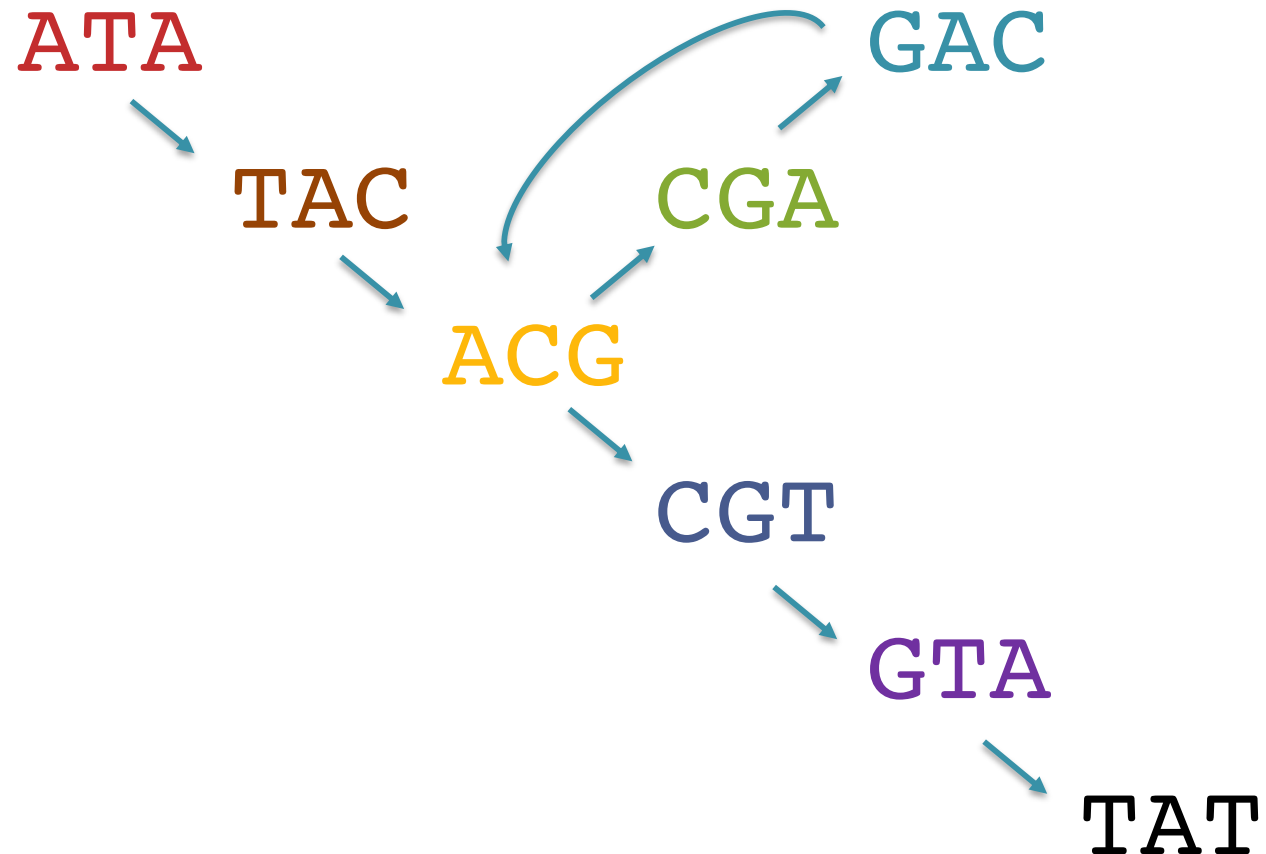~~TACG~~

ATA → TAC → ACG → GAC, CGA, CGT → GTA → TAT

ATACGACGTAT

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~
~~ACGT~~
~~ATAC~~
~~CGAC~~
~~CGTA~~
~~GACG~~
~~GTAT~~
~~TACG~~

ATA → TAC → ACG → CGA → GAC

ACG → CGT → GTA → TAT

Whats another possible genome?

ATACGACGTAT

Titus Brown
@ctitusbrown

Wow, this could double as life philosophy, too!

> Michael Schatz @mike_schatz
>
> Replying to @ZaminIqbal @nomad421 and 4 others
>
> Yep, very easy to find *a* path, very hard to find *the* path

11:40 AM - 22 Jan 2018

4 Retweets   17 Likes

2   4   17

Next Steps

1. Reflect on the magic and power of DNA ☺

2. Check out the course webpage

3. Register on Piazza & GradeScope

4. Work on Assignment 1

   1. Set up Dropbox for yourself!
   2. Get comfortable on the command line