Exam Review

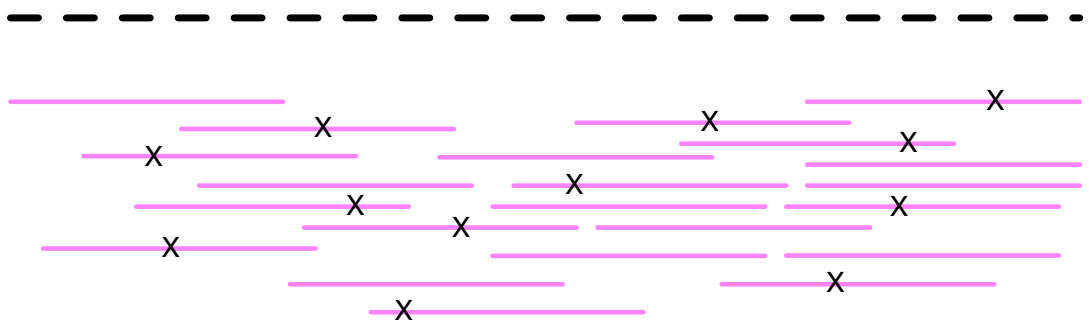
DISCLAIMER: This is a non-exhaustive list
Also, remember the things I told you Id ask again, from exam 1

Lecture 30
Nov 16, 2016

Announcements

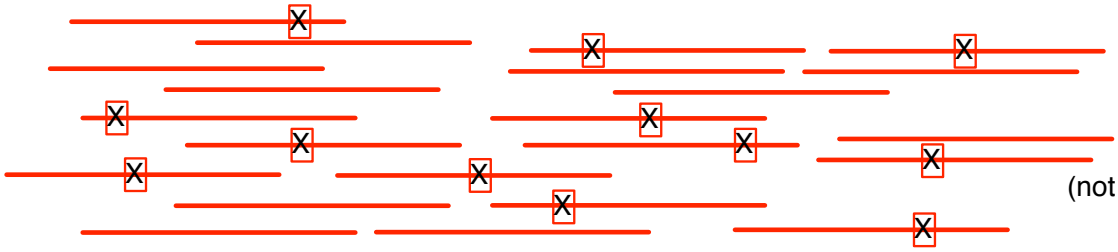
Lecture 14

Digital Normalization



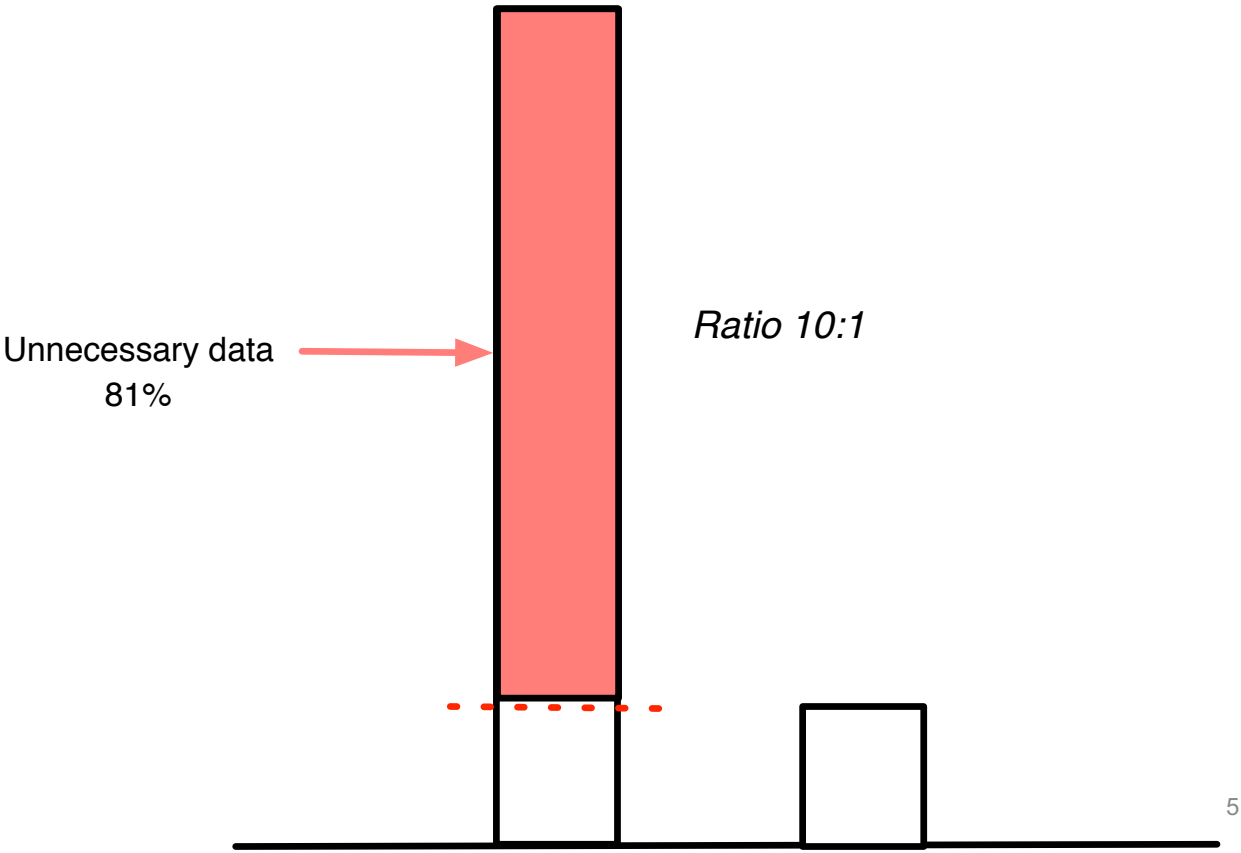
Reads
(randomly sequenced)

```
for read in dataset:
   if estimated_coverage(read) < C:
      accept(read)
   else:
      discard(read)
```

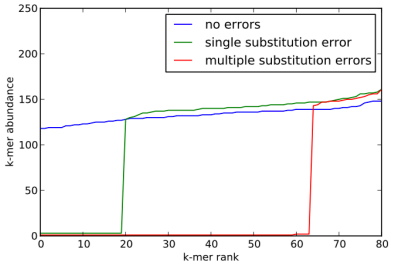


Redundant reads
(not needed for assembly)

Digital Normalization



Digital Normalization



Median kmer abundance

0 error: 32,33,34,34,35,36,40
1 error: 1,1,1,32,24,24,40
>1 error: 1,1,1,1,1,32,34

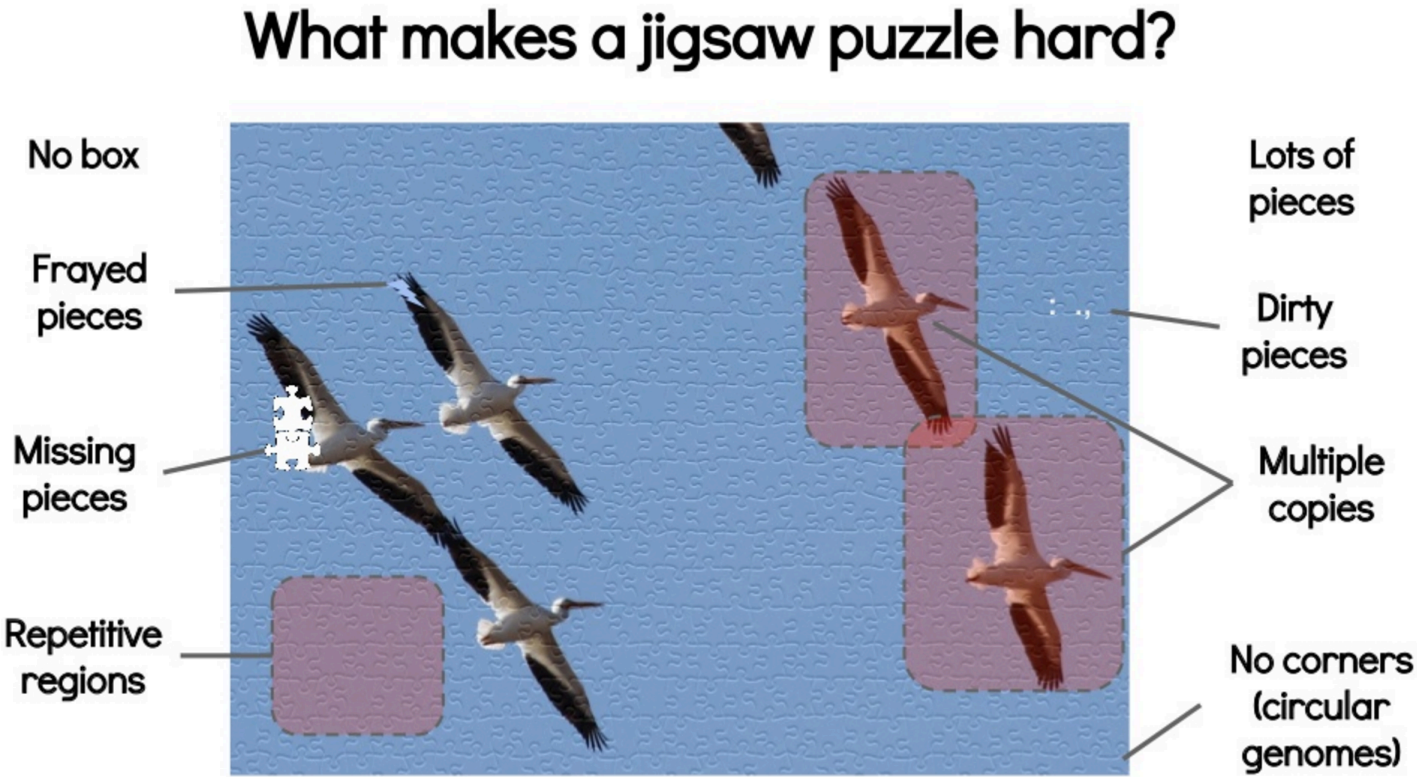
Lecture 16

Assemble a genome? General Strategies

Genome size	Unlimited $$	Typical
>10Mb		
10Mb - 100Mb		
> 100 Mb		

Lecture 18

Assemble a genome?



Other Assembly Terms

Unitig

Contig

scaffold

Lecture 19

First law of assembly

If a suffix of read A is similar to a prefix of read B...

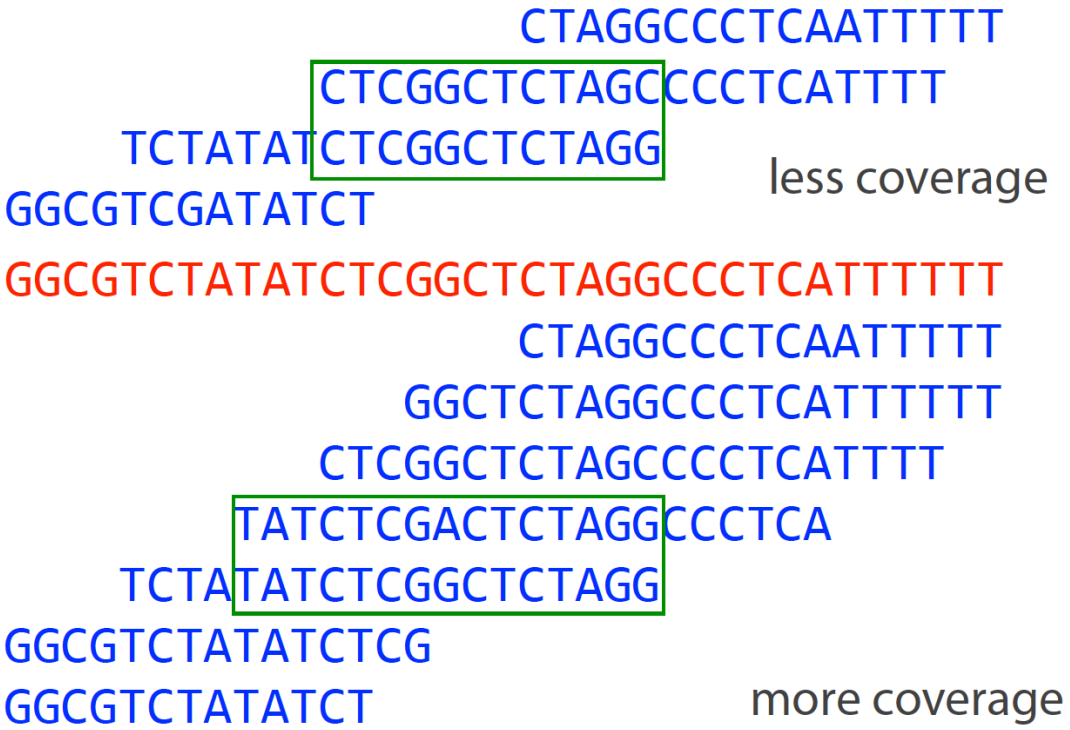
```
TCTATATCTCGGCTCTAGG
       ||||||| |||||||
       TATCTCGACTCTAGGCC
```

...then A and B might *overlap* in the genome

```
          TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
          TATCTCGACTCTAGGCC
```

Second law of assembly

More coverage leads to more and longer overlaps

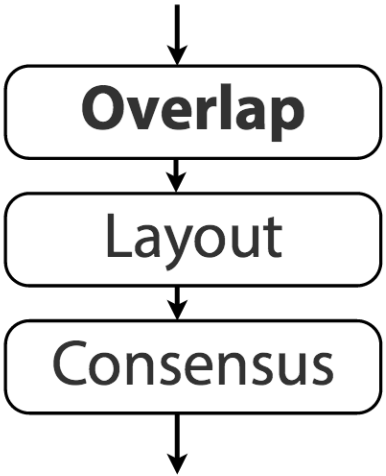


Assembly

- OLC Assembly: Characteristics

Assembly

- OLC Assembly



Build overlap graph

Bundle stretches of the overlap graph into *contigs*

Pick most likely nucleotide sequence for each contig

Lecture 20

Assembly – De Bruijn

Hamiltonian Path Problem

Eulerian Path Problem

Lecture 21

Genome Assembly

Lecture 21
Oct 21, 2016

Assembly – De Bruijn

GATTACAGTTCA

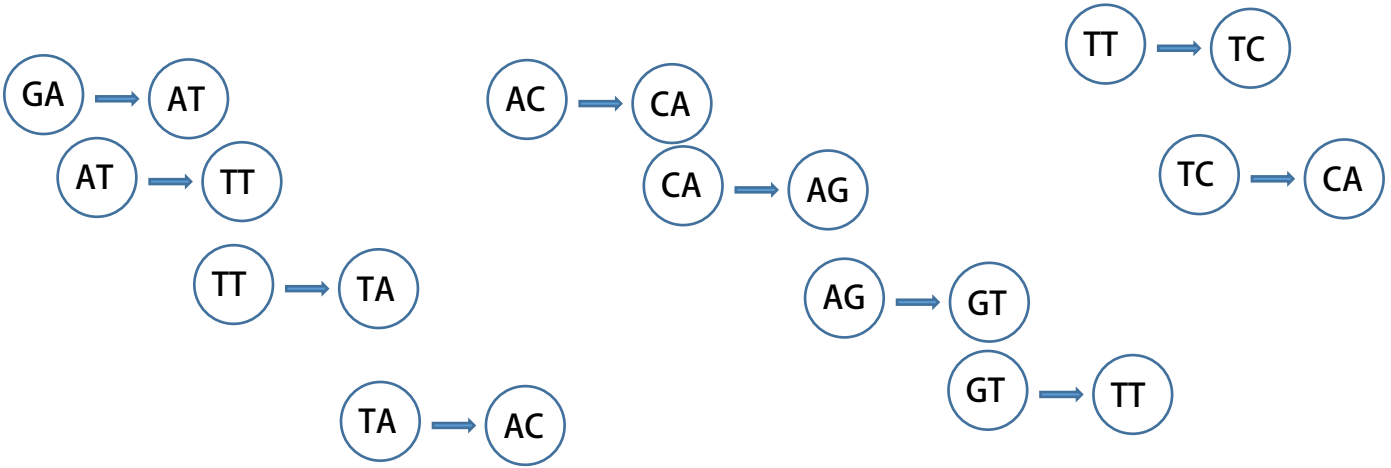
Assembly – De Bruijn

```
GATTAC
GAT
  ATT
    TTA
      TAC
```

```
ACAGTTCA
ACA
  CAG
    AGT
      GTT
        TTC
         TCA
```

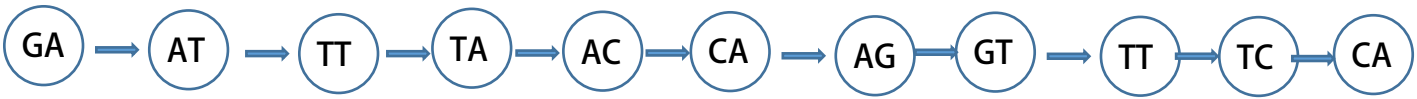
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



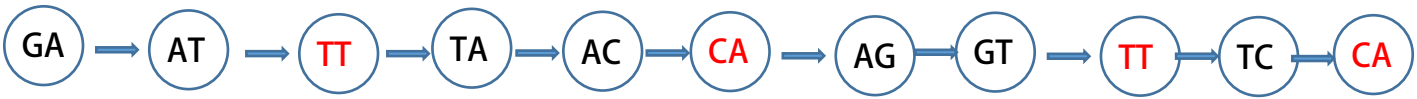
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



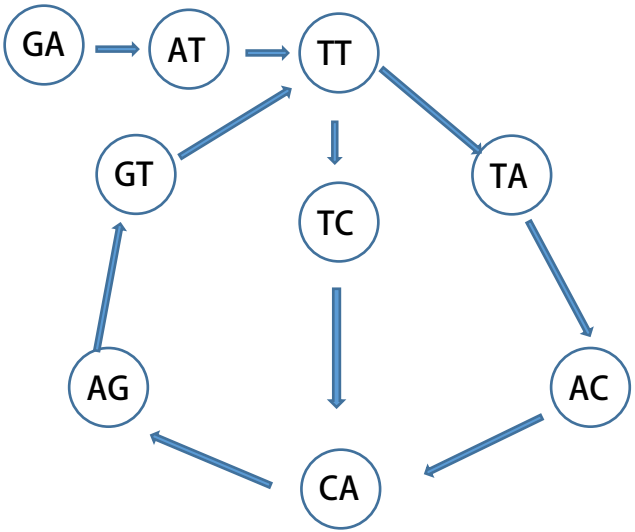
Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



Assembly – De Bruijn

GAT ATT TTA TAC ACA CAG AGT GTT TTC TCA



Lecture 23

Evaluating genome assembies

Lecture 24

Transcriptome Assembly

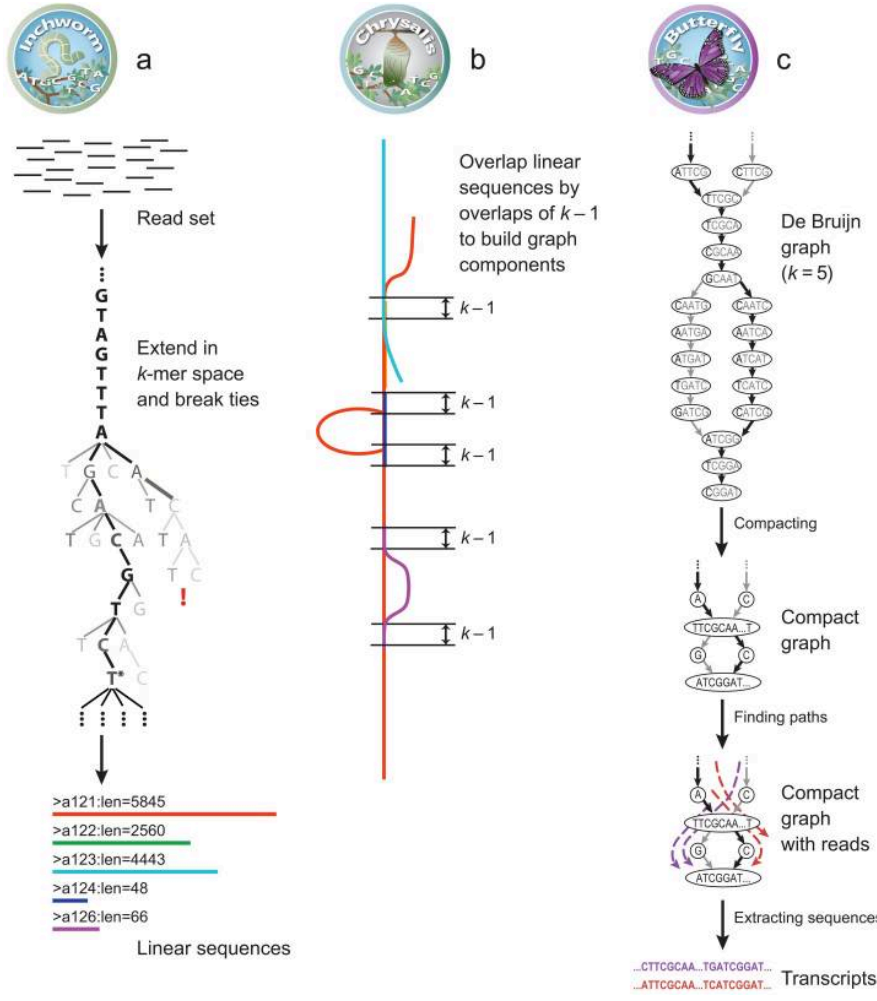
Why easier than genome assembly?

Transcriptome Assembly

Why harder than genome assembly?

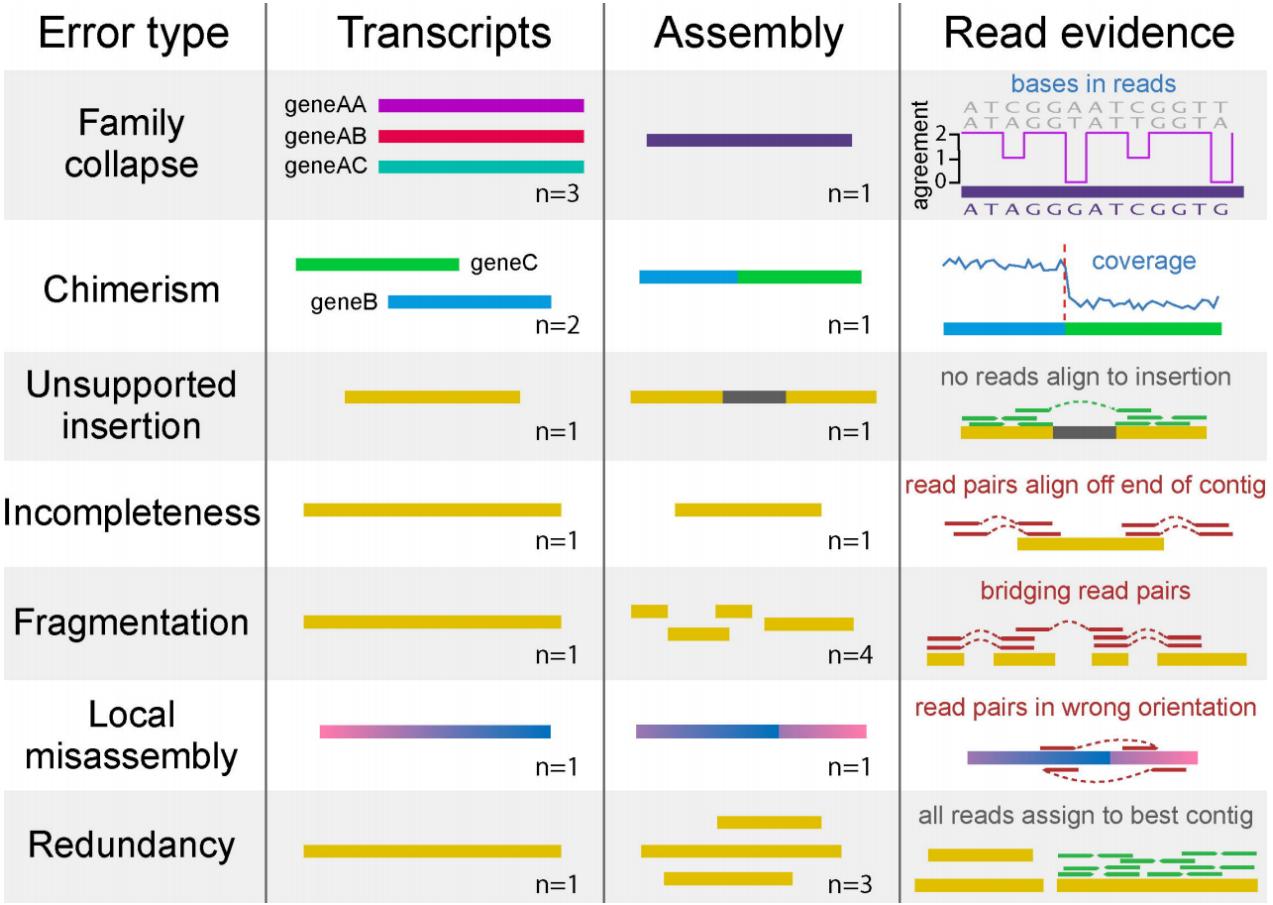
Transcriptome Assembly

Trinity



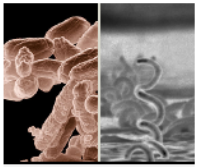
Lecture 26

Transcriptome Assembly – evaluation

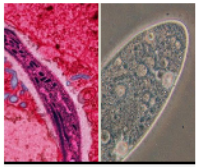


Transcriptome Assembly – evaluation – BUSCO

Datasets ***(Beta versions, updated sets and additional lineages coming soon)***


Bacteria sets


Eukaryota sets


Protists sets


Metazoa sets


Fungi sets


Plants set

Download all datasets

Image credits

Lecture 27

How to assess “abundance”

RPKM — Reads per kilobase per million mapped reads

FPKM — Fragments per kilobase per million mapped reads

Don’t use these measures, TPM measures the “same thing”, but in a better way.

TPM — Transcripts per million

Useful for visualization / assessment etc.

(Estimated) Number of Reads

These are what are used (after normalization) for differential expression. Why can’t we use TPM?

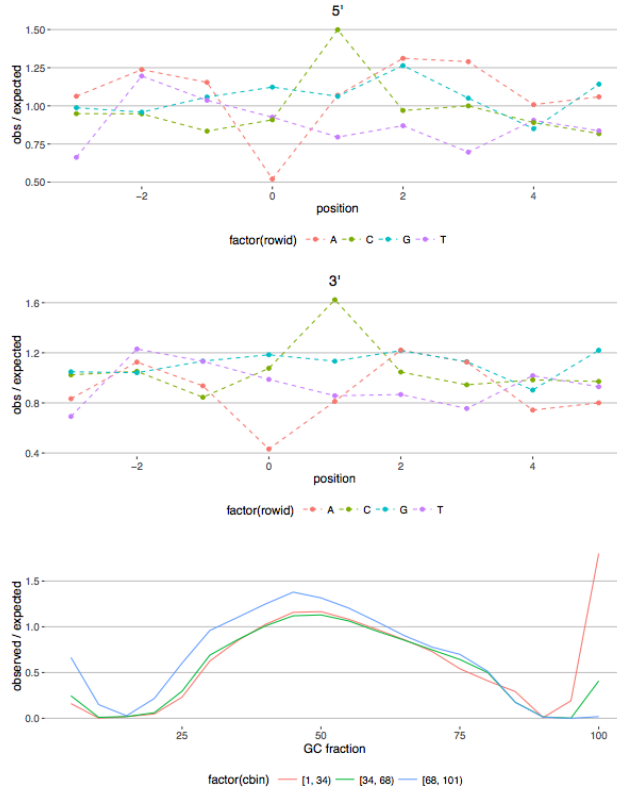
Biases abound in RNA-seq data

Biases in prep & sequencing can have a significant effect on the fragments we see.

Fragment gc-bias[1]—
The GC-content of the fragment affects the likelihood of sequencing

Sequence-specific bias[2]—
sequences surrounding fragment affect the likelihood of sequencing

Positional bias[2]—
fragments sequenced non-uniformly across the body of a transcript



1:Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation." bioRxiv (2015): 025767.

2:Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.