

Bioinformatics

CS300

Genome Sequencing and Assembly
Chapter 8

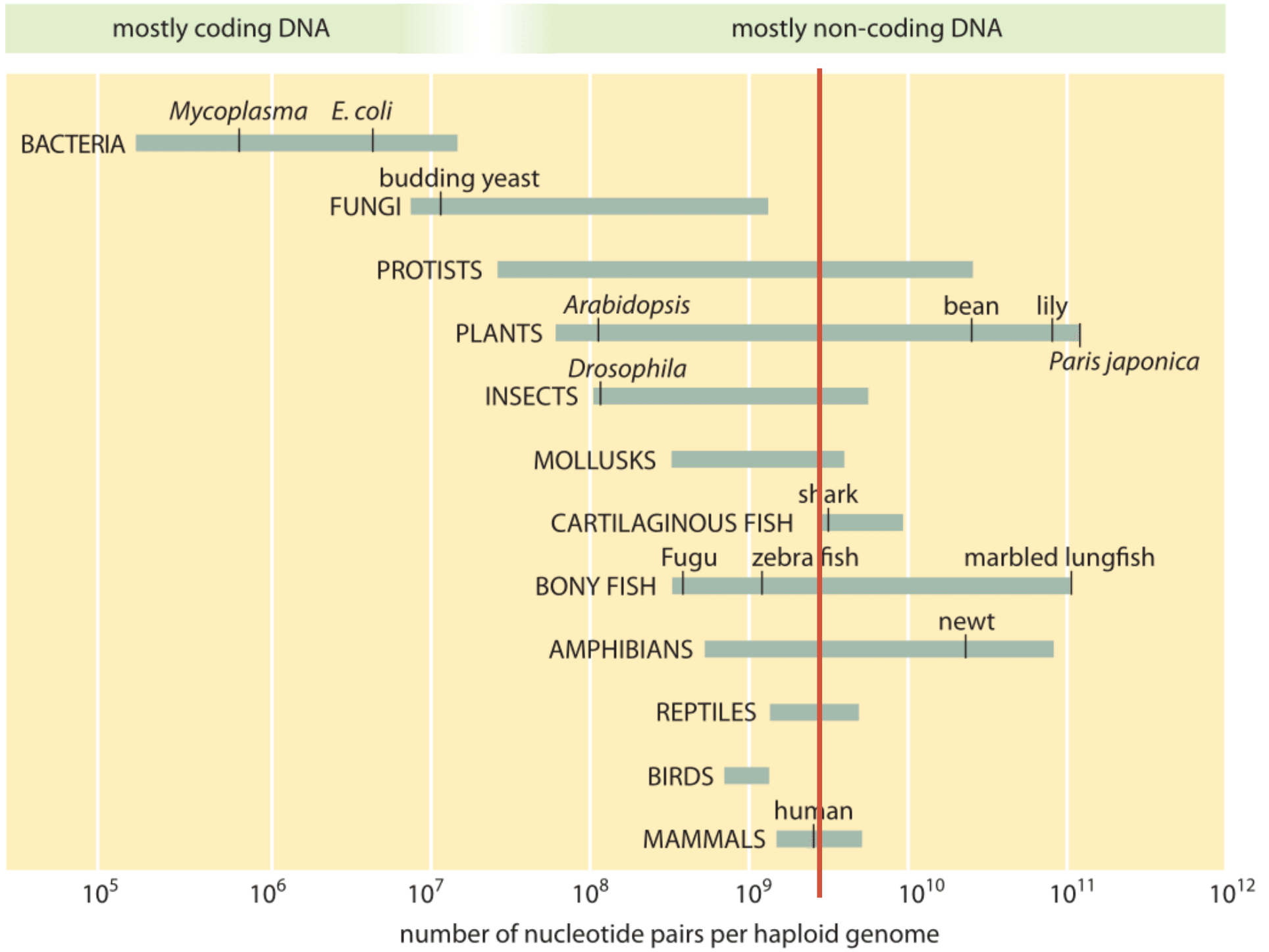
Spring 2021

Oliver BONHAM-CARTER




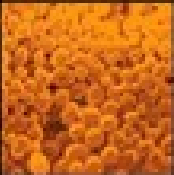

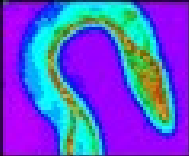



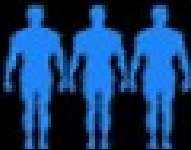
What is a Genome?

- An organism's complete set of DNA, including all of its genes, regulatory regions, non-coding regions, etc.
- An organism's complete set of genetic instructions





What Is In a Genome?

	Organism	Number of genes in the genome
	<i>Mycoplasma genitalium</i>	517
	<i>Saccharomyces cerevisiae</i>	6,275
	<i>Arabidopsis thaliana</i>	~ 20,000
	<i>Caenorhabditis elegans</i>	19,099
	<i>Haemophilus influenzae</i>	1,743
	<i>Drosophila melanogaster</i>	13,601
	<i>Neisseria meningitidis</i>	2,158
	<i>Homo sapiens</i>	20,000–25,000



Genome Projects

- Goals:
 - Determine complete genome sequence of an organism
 - Annotate (*exhibit*) protein-coding genes and other important genome-encoded features



Genome Projects

- Projects:
 - Over 15,000 [genome projects](#) in progress or completed

Genome Information by organism

[Download Reports from FTP site](#)

Overview [30649] Eukaryotes [4874] Prokaryotes [118997] Viruses [7497] Plasmids [10401] Organelles [10835]

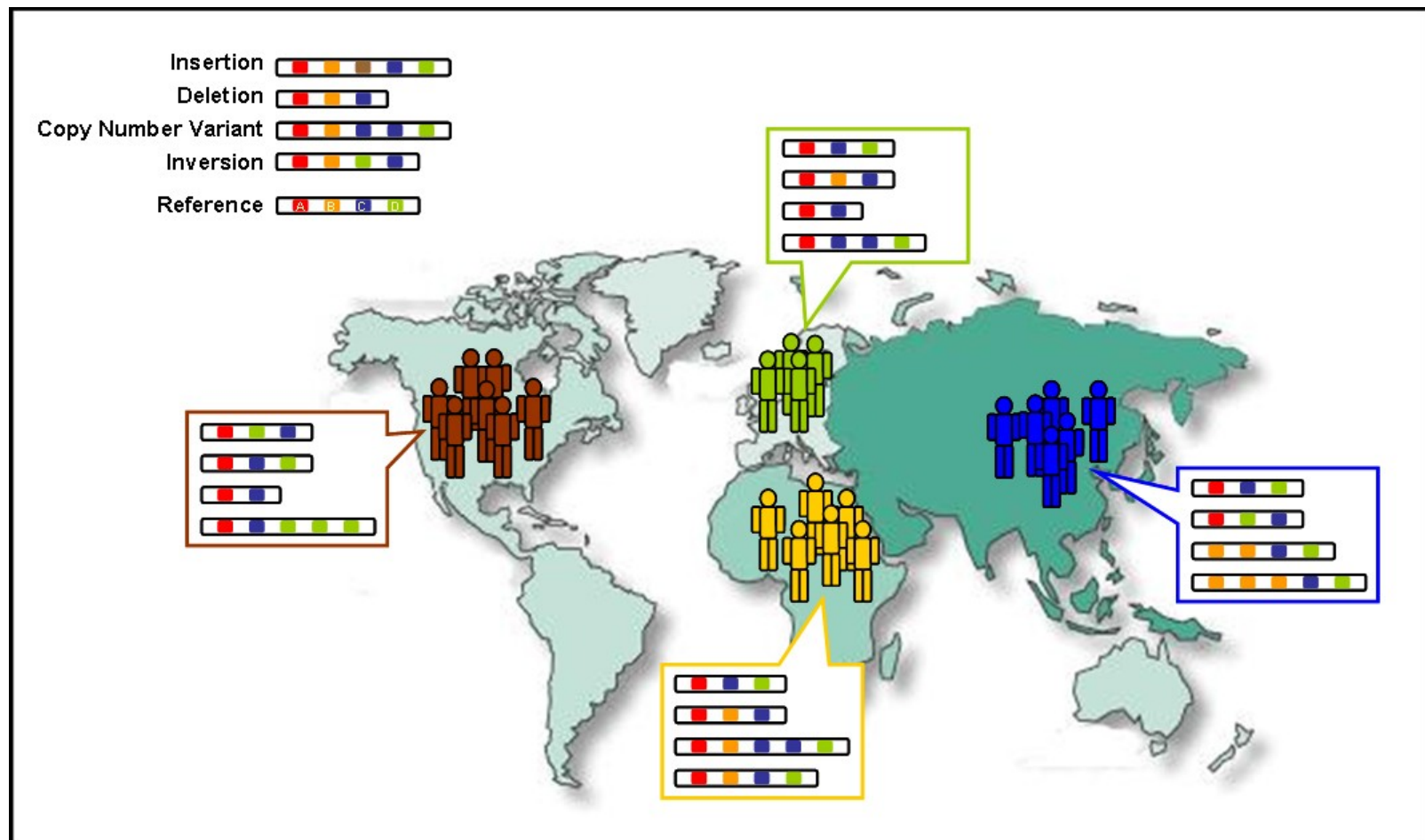
[Download selected records](#)

Items 1 - 100 of 30649 << First < Prev Page 1 of 307 Next > Last >>

Organism/Name	Kingdom <input type="button" value="All"/>	Group <input type="button" value="All"/>	SubGroup <input type="button" value="All"/>	Size (Mb)	Chr	Organelles	Plasmids	Assemblies
'Chrysanthemum coronarium' phytoplasma	Bacteria	Terrabacteria group	Tenericutes	0.739592	-	-	-	1
'Echinacea purpurea' witches'-broom phytoplasma	Bacteria	Terrabacteria group	Tenericutes	0.545427	-	-	-	1
'Osedax' symbiont bacterium Rs2_46_30_T18	Bacteria	unclassified Bacteria	unclassified Bacteria (miscellaneous)	4.02183	-	-	-	1
Abaca bunchy top virus	Viruses	ssDNA viruses	Nanoviridae	0.006422	6	-	-	1
Abalone herpesvirus Victoria/AUS/2009	Viruses	dsDNA viruses, no RNA stage	unclassified	0.211518	1	-	-	1
Abalone shriveling syndrome-associated virus	Viruses	dsDNA viruses, no RNA stage	unclassified	0.034952	1	-	-	1
Abelson murine leukemia virus	Viruses	Retro-transcribing viruses	Retroviridae	0.005894	1	-	-	1

Genome Projects

- Contrast genetic material of populations to determine ancestry



Genome Projects: Data

- Annotations: gene locations for protein products in sequences.

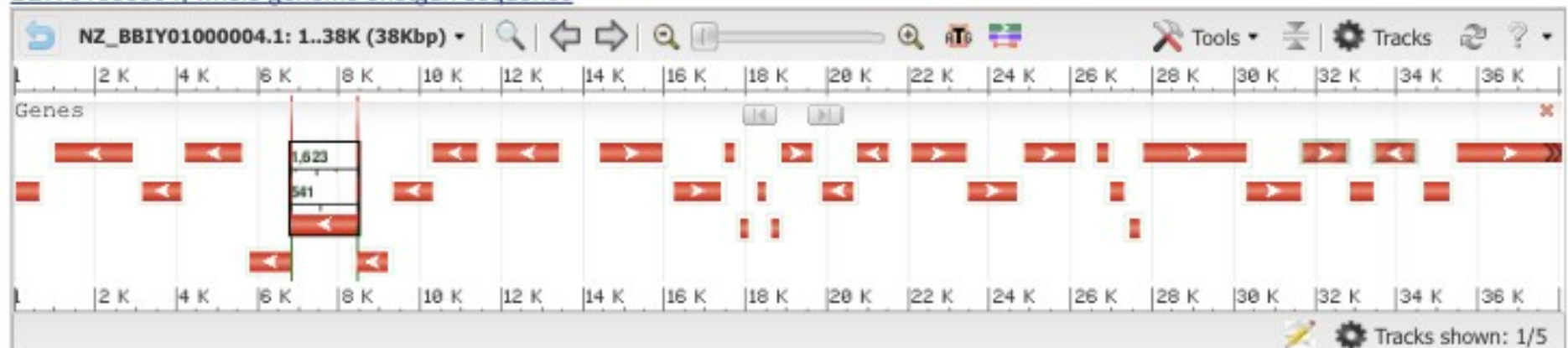
Genome Assembly Annotation

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	tRNA	Other RNA	Gene	Pseudogene
	master WGS	NZ_BBIY000000000.1	BBIY000000000.1	0.74	27.6	901	27	-	928	-

Genome Region

'Chrysanthemum coronarium' phytoplasma strain OY-V
BBIY01000004, whole genome shotgun sequence

Go to nucleotide: [Graphics](#) [FASTA](#) [GenBank](#)



<https://www.ncbi.nlm.nih.gov/genome/browse/>

Genome Projects: Data

- Protein
meta
data

'Chrysanthemum coronarium' phytoplasma strain OY-V
BBIY01000004, whole genome shotgun sequence

Go to nucleotide

NZ_BBIY01000004.1: 1..38K (38Kbp)

Genes

WP_042067579.1

CDS: WP_042067579.1
Title: sugar ABC transporter substrate-binding protein
Location: complement(6,823..8,445)
[Length]
Span: 1,623
Product: 540
[Qualifiers]
inference: COORDINATES: similar to AA
sequence:RefSeq:WP_011161091.1

Download: [WP_042067579.1](#)

Links & Tools

BLAST Genomic: [NZ_BBIY01000004.1 \(6,823..8,445\)](#)
BLAST Protein: [WP_042067579.1](#)
BLINK Results: [WP_042067579.1](#)
FASTA View: [NZ_BBIY01000004.1 \(6,823..8,445\)](#), [WP_042067579.1](#)
GenBank View: [NZ_BBIY01000004.1 \(6,823..8,445\)](#), [WP_042067579.1](#)
Graphical View: [WP_042067579.1](#)

Run Blast

are here: NCBI > Genomes & Maps >

SETTING STARTED

BI Education
BI Help Manual
BI Handbook
ining & Tutorials
bmit Data

Genetic Variation

- Having diverse genetic information helps to spot genetic conditions in organisms
- Find *Genetic drift*: a random fluctuation in the population frequency of a trait
 - Occurring in descendant generations from a particular organism
 - Are evolutionary pressures causing a change in a species? Can we compare species from two different environments to learn about this drift?



Detection By Comparison

As you already know!

- Genetic drift may have unusual consequences. Why are these even genes *out-there*?



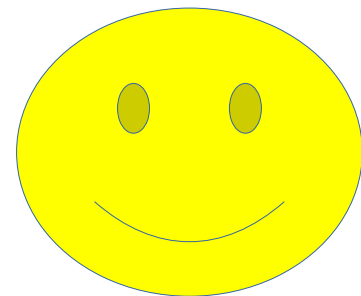
Hapsburg jaw



Ellis-Van Creveld syndrome, a sixth finger



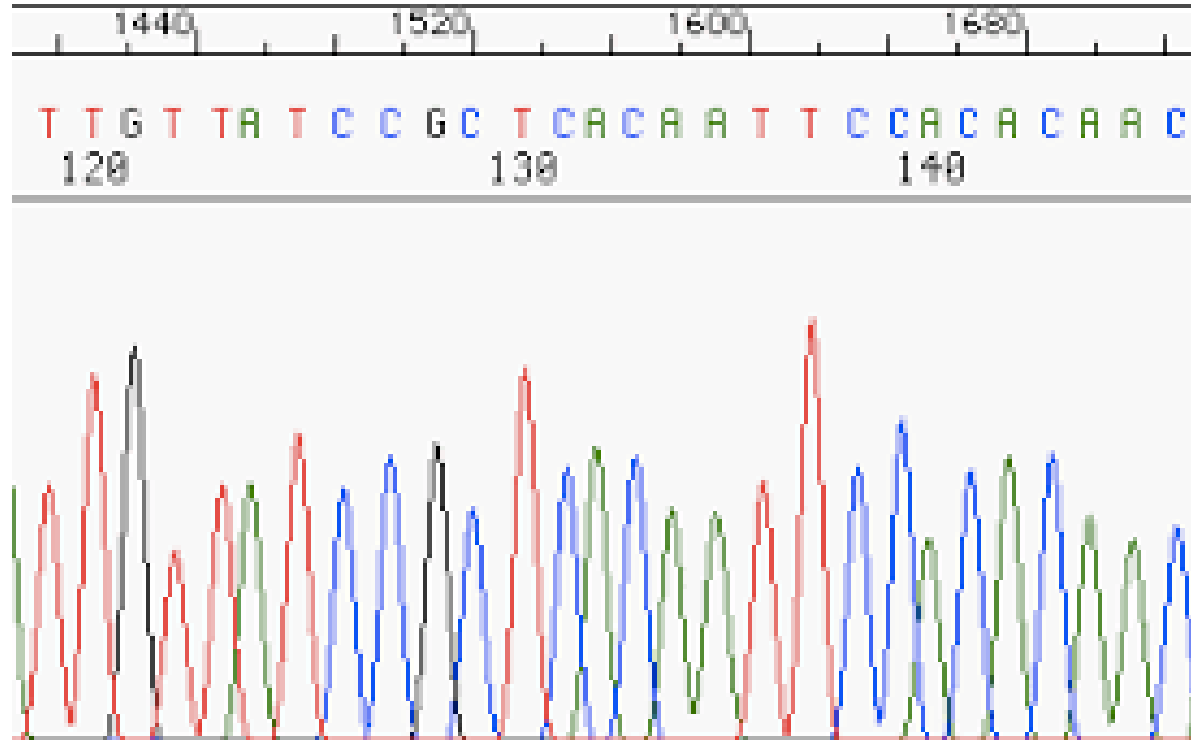
How do we
get data to learn
about drift??



Genome Sequencing:

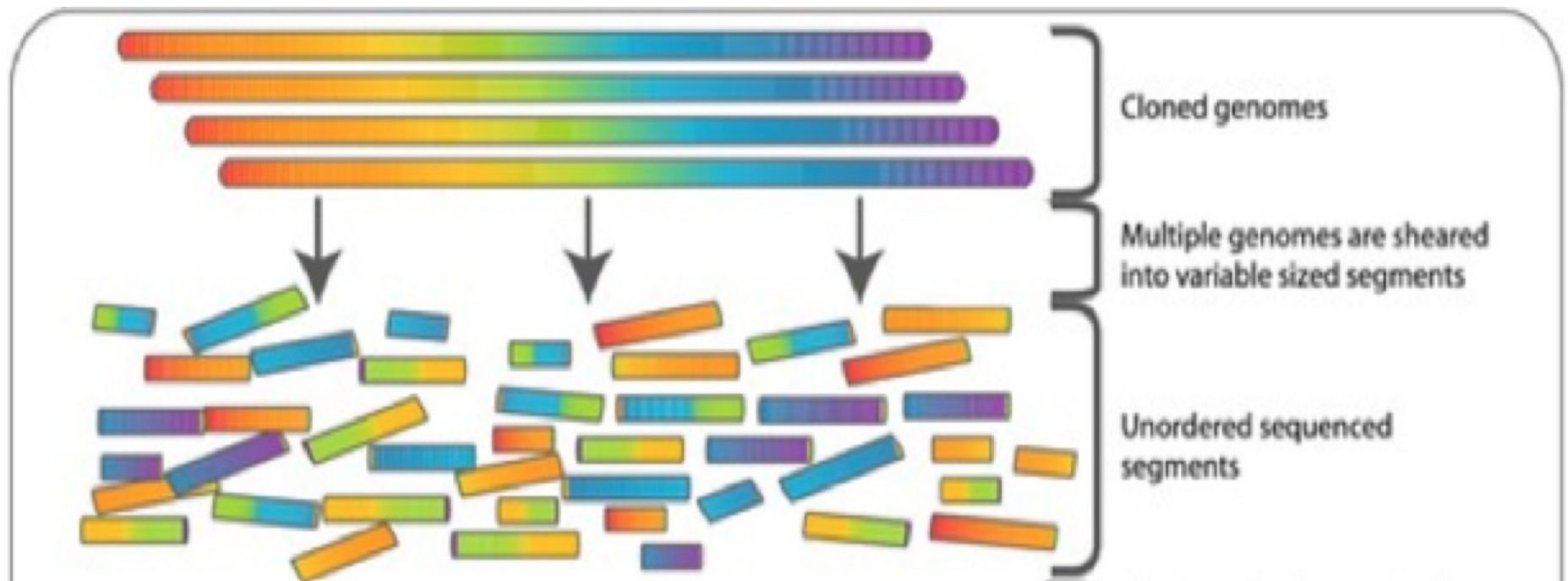
Getting genetic data (for analysis)

- Bases are recorded as little peaks
- Reads = Small segments of DNA from sequencer machine
- Contigs = Segments of partially combined reads



Genome Sequencing

- The technology works by “exploding” DNA into smaller, manageable pieces
- It recombines pieces (***Reads***) into bigger pieces (***Contigs***)
- And then bigger chunks are combined like a jigsaw puzzle




Shredded Book Reconstruction

- Imagine that Dickens has “accidentally” shredded his first printing of a Tale of Two Cities
- What can be done to re-create the manuscript?



- Dickens accidentally shreds first printing of Tale of Two Cities
 - first printing = 5 copies

It was the best of wisdom, best of times, was the worst wisdom, as was the best of wisdom, it was the, age...

A man with a balding head, wearing a red long-sleeved shirt, is shown from the chest up. He is sitting in a light-colored chair and has his right hand pressed against his face, covering his eyes and nose. His expression is one of distress or embarrassment. The background is a plain, light-colored wall.



Shredded Book Reconstruction

Dickens accidentally shreds first printing of Tale of Two Cities

- first printing = 5 copies
- shredding was random (can cut between different words in each copy)
- always 5 words per fragment

It was the best of times, it was the worst of times, it was the

It was the best of times, it was the worst of times, it was the

It was the best of times, it was the worst of times, it was the



Shredded Book Reconstruction

Dickens accidentally shreds first printing of Tale of Two Cities

- first printing = 5 copies
- shredding was random (can cut between different words in each copy)
- always 5 words per fragment

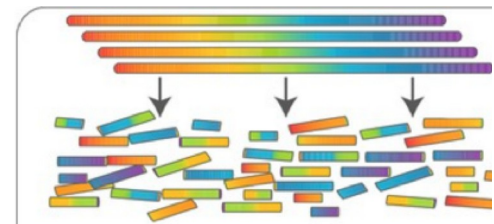
It was the best of times, it was the worst of times, it was the

It was the best of times, it was the worst of times, it was the

It was the best of times, it was the worst of times, it was the

5 copies x 138, 656 words/5 words per fragment = 138k fragments

All short fragments are mixed together



times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

the best of times, it

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

the best of times, it

best of times, it was

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

Assembly Parameter:
100% identify across
four words

the best of times, it

best of times, it was

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

was the best of times,

the best of times, it

best of times, it was

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

It was the best of

was the best of times,

the best of times, it

best of times, it was

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

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



Tale of Two Cities

Charles Dickens

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 nothing before us, we were all going direct to Heaven, we were all going direct the other 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.

Making sense of it all: We can already see how these words are coming together!

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

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

The *repeats* pile up:
The actual placement
of each individual
fragment unknown

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

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 age

times, it was the worst

The repeats can
cause ambiguity
and prevent
proper assembly

times, it was the age

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

worst of times, it was

wisdom, it was the age

was the worst of times,

was the best of times,

was the age of foolishness,

was the age of wisdom

times, it was the worst

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

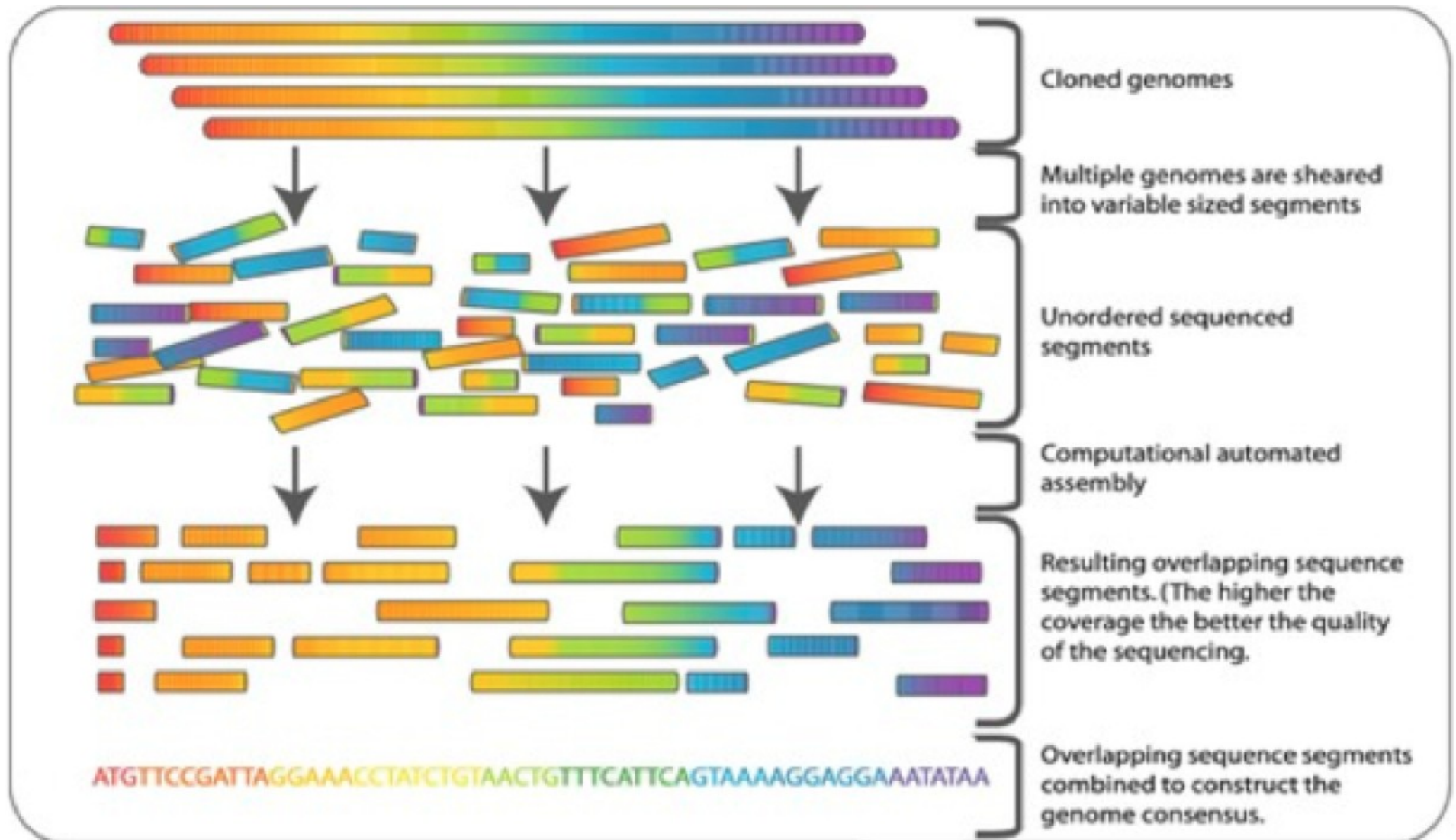
times, it was the age

times, it was the worst

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

Which word to use here?!

Summarizing Genome Sequencing

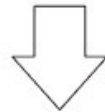




Coverage and Ordering

random short
sequence reads

TTTTACCACCTA
CGGACCAGA
CCATGG
AGACTTTTTTTACCAA
ATACCCATG
ATCGGA
GACCAGACTTTT
CCATACCCGA
CATGG
ACCTAAAT
ACCTAAT
CCCGACATACCGA

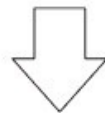


coverage

1 1 2 2 3 3 2 2 3 3 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 3 3 3 3 3 2 2 2 1 2 2 2 2 3 4 4 3 3 2

AGACTTTTTTTACCAA CCATACCCGA CCATGG
ATCGGA TTTTACCAACCTA CCCGACATACCGA
GACCAGACTTTT ACCTAAAT ATACC CATGG
CGGACCAGA AATCCATA ATACCCATG

assembly of
overlapping
fragments



assembled
contig sequence

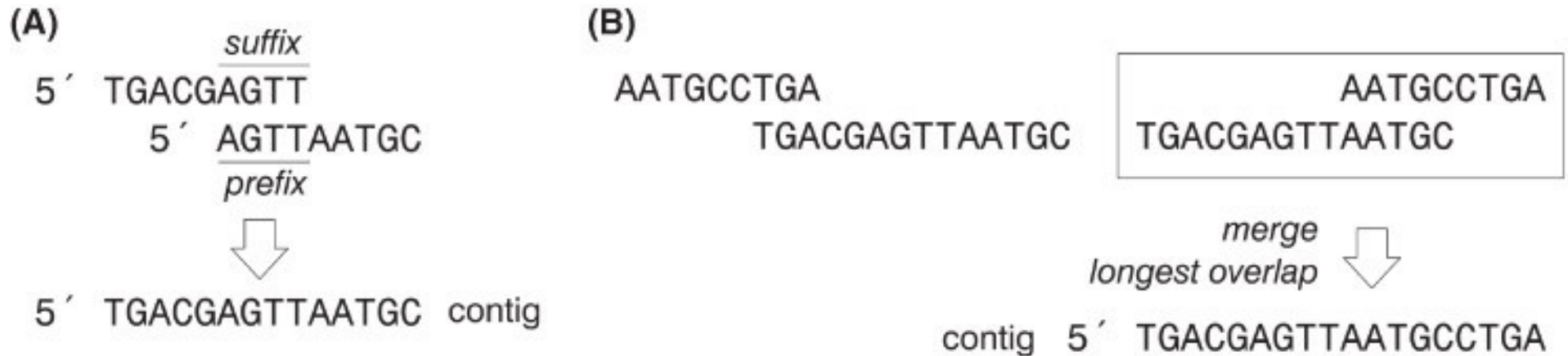
ATCGGACCAGACTTTTTTTACCAACCTAAATCCATACCCGACATACCCATGG



Finding the Largest Overlap

- Consider the assembly of two fragments:
 - If there is more than one overlap, choose the **longest** overlap
 - Assume the sequences are not identical
 - Assume neither sequence is a substring of the other
 - The longest **possible** overlap is length of the shorter sequence minus a character (to determine placement in the larger sequence)

Algorithm to Find Overlaps



1. Start with reads; s1 and s2
2. n = size of the smallest sequence – 1
3. Compare n suffix/prefix characters from s1 with n prefix/suffix characters s2
4. Count matching bases in the prospective overlap region. If the number of matches = n , then the largest overlap is found
5. If the number of matches $< n$, $n = n-1$
6. If $n = 0$ then no overlap, go to step 3



Assembling Contigs

Table 8.3 Overlaps for a hypothetical set of sequence reads.

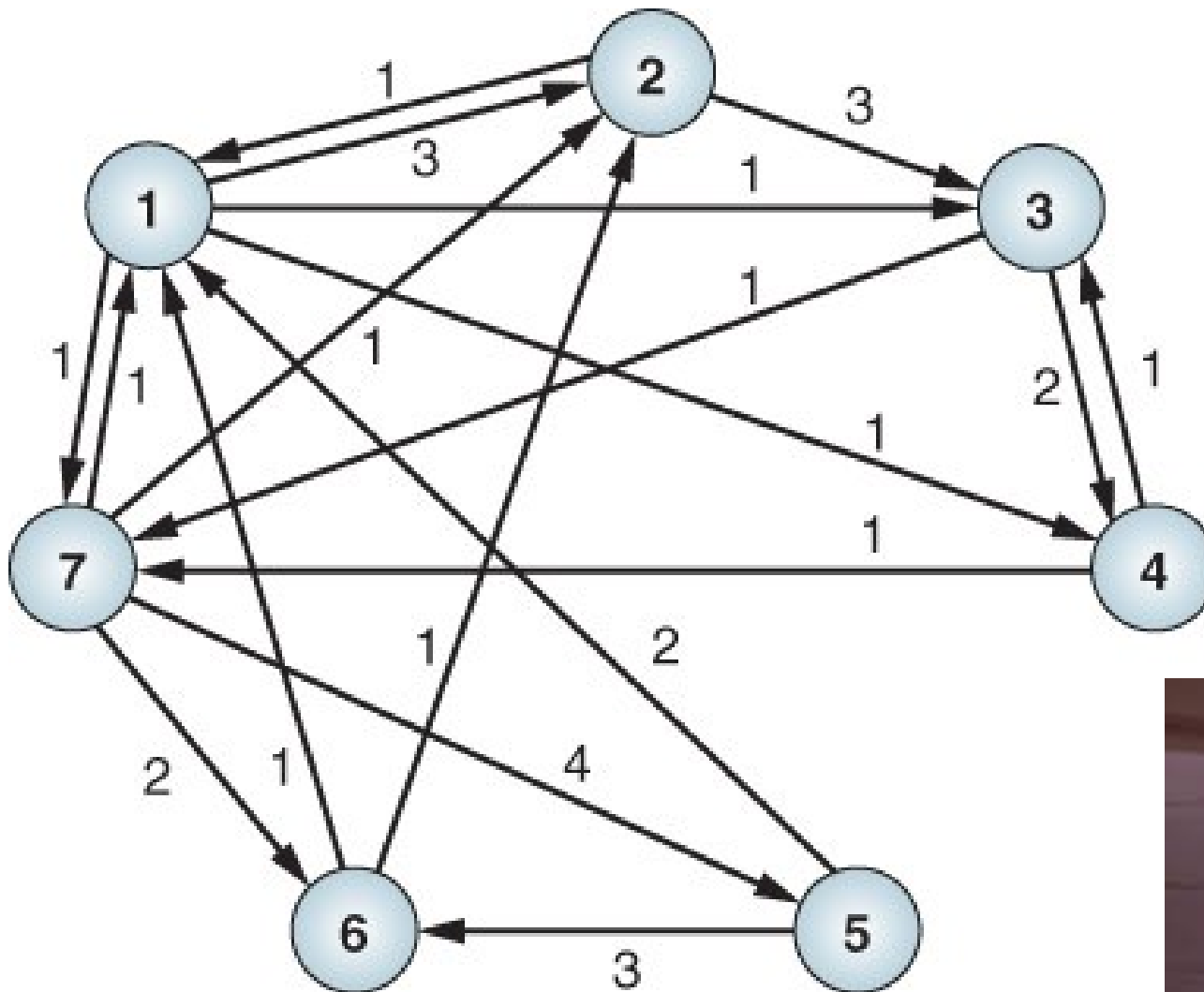
Fragments	Overlaps (Length)
1. TACCTTG	2 (3), 3 (1), 4 (1), 7 (1)
2. TTGAT	1 (1), 3 (3)
3. GATATGG	4 (2), 7 (1)
4. GGAG	3 (1), 7 (1)
5. CTCTA	1 (2), 6 (3)
6. CTAGT	1 (1), 2 (1)
7. GCTCT	1 (1), 2 (1), 5 (4), 6 (2)

For each sequence, we name an overlap with another sequence by number and number of overlaps.

Seq1: TACC**TTG**
Seq2: **TTG**AT

With_Seq (num of overlaps)
Ex: **2 (3)**
Seq 1 has three overlaps with Seq 2

Assembling a Contig: graph representation



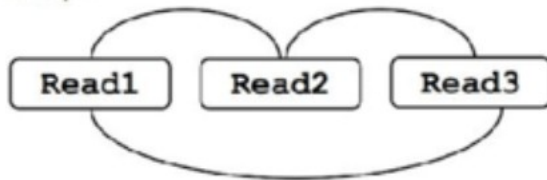
*Make
It
so!*



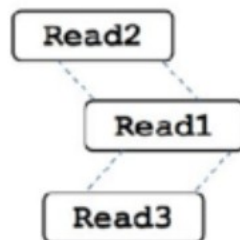
Two Basic Techniques

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

```

CGATTCTA
  TTCTAAGT
   GATTGTA
  -----
CGATTCTAAGT
    
```

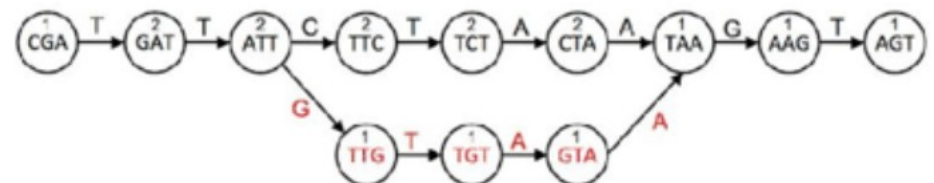
We just saw this one

(b) De Bruijn graph assembly

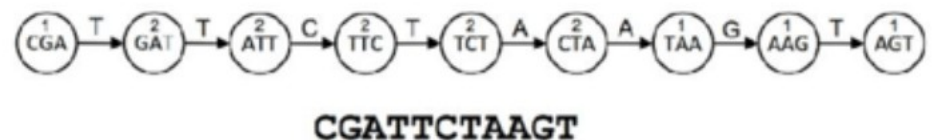
(i) Make kmers

Read1: TTCTAAGT	Read2: CGATTCTA	Read3: GATTGTA
Kmers: TTC	Kmers: CGA	Kmers: GAT
TCT	GAT	ATT
CTA	ATT	TTG
TAA	TTC	TGT
AAG	TCT	GTA
AGT	CTA	TAA

(ii) Build graph



(iii) Walk graph and output contigs



Same idea but we use *k*-mers here