Bioinformatics

CS300

Genome Sequencing and Assembly

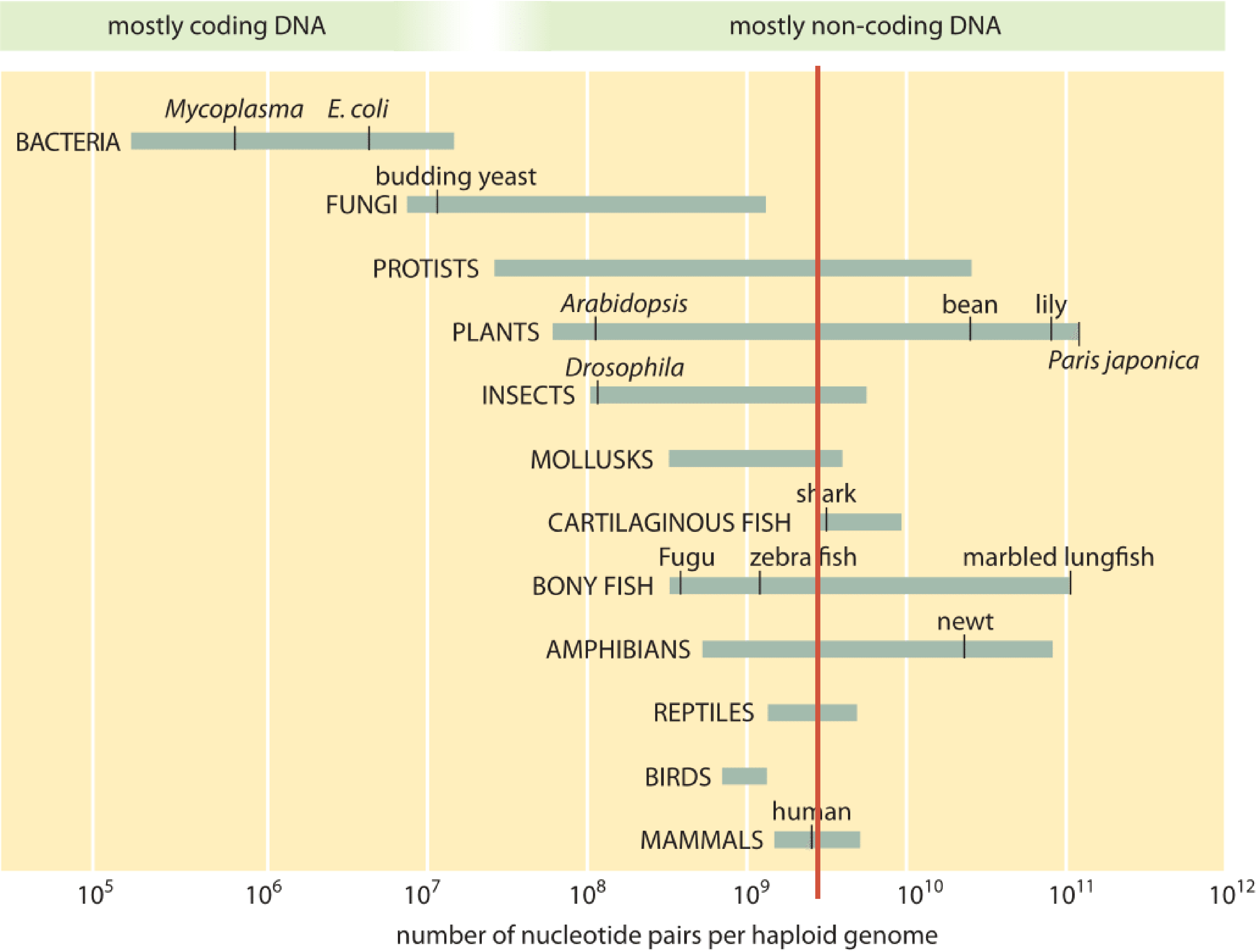
Fall 2019

Oliver BONHAM-CARTER

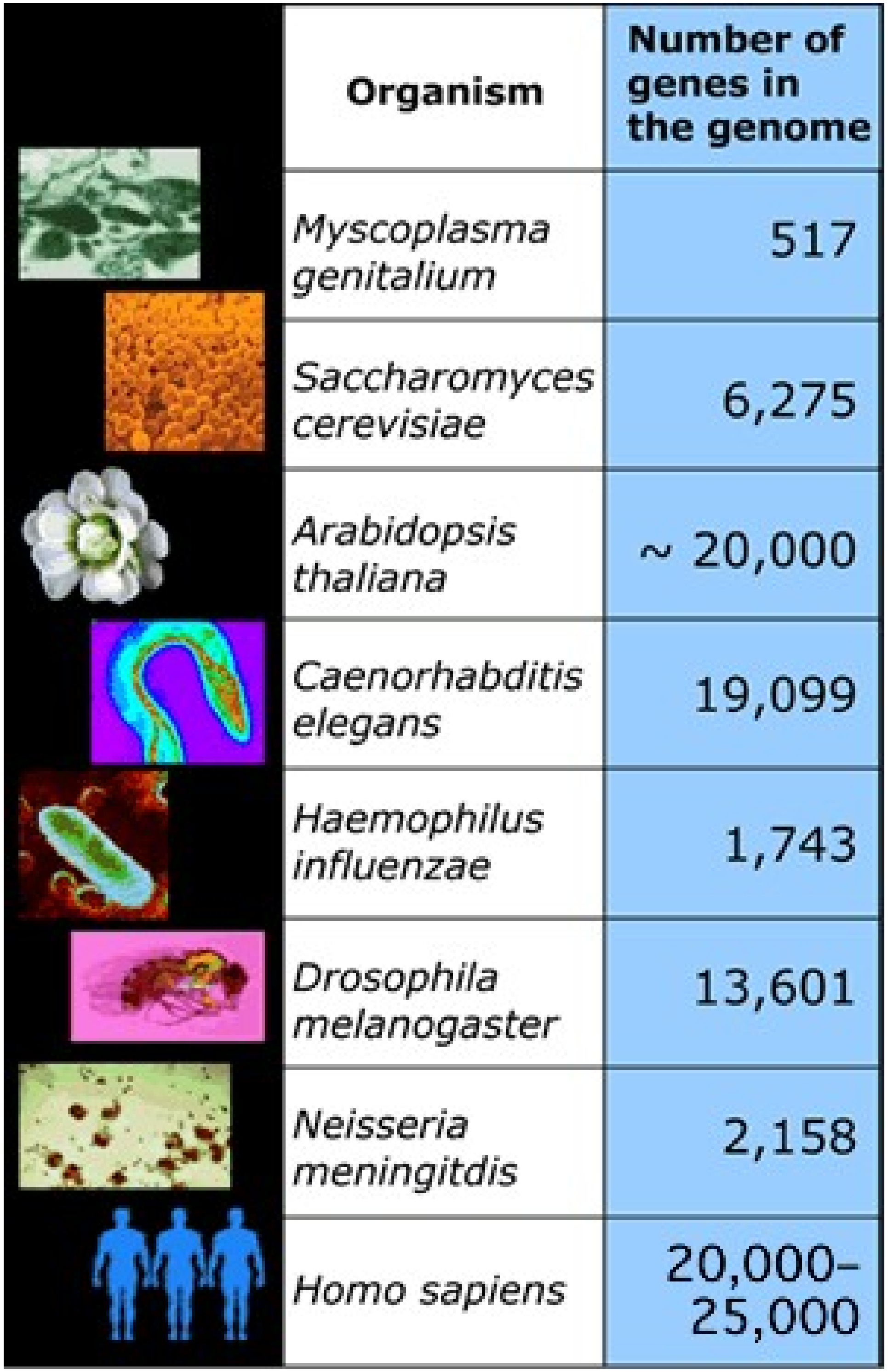
What is a Genome?

- 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


mostly coding DNA
mostly non-coding DNA
Mycoplasma
E. coli
BACTERIA
budding yeast
FUNGI
PROTISTS
Arabidopsis
bean
lily
PLANTS
Paris japonica
Drosophila
INSECTS
MOLLUSKS
shark
CARTILAGINOUS FISH
Fugu
zebra fish
marbled lungfish
BONY FISH
newt
AMPHIBIANS
REPTILES
BIRDS
human
MAMMALS
10^5
10^6
10^7
10^8
10^9
10^{10}
10^{11}
10^{12}
number of nucleotide pairs per haploid genome

What Is In a Genome?



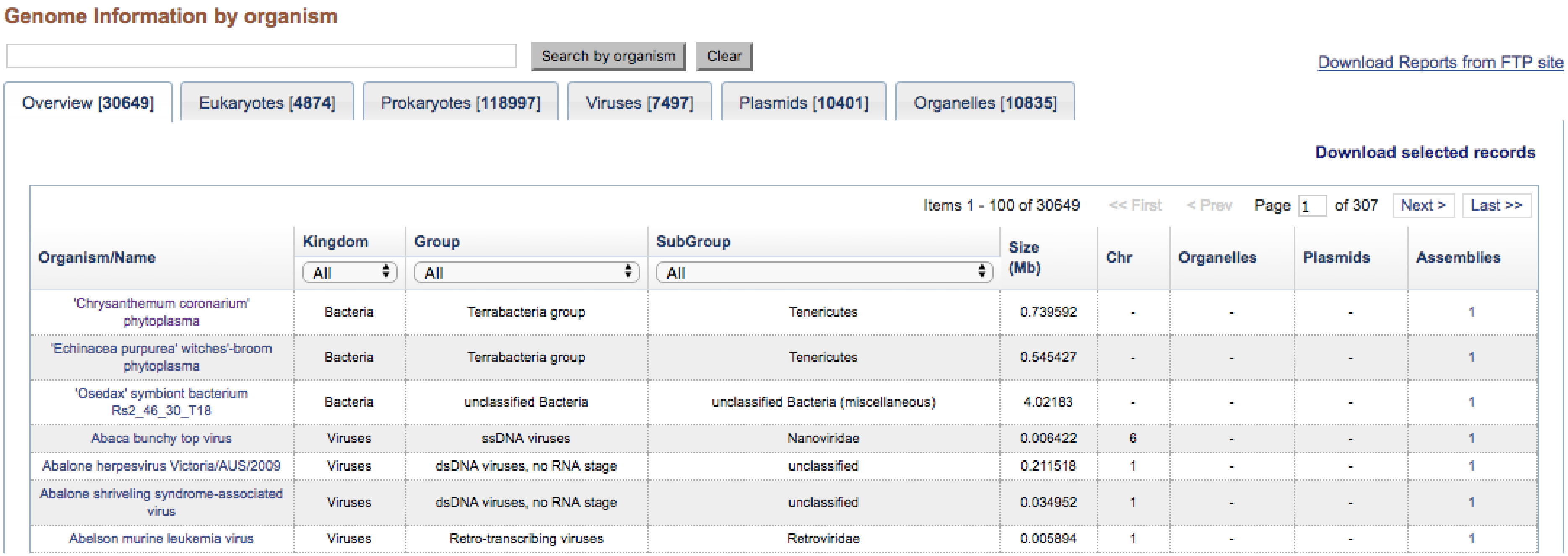
Organism	Number of genes in the genome
Myscoplasma genitalium	517
Saccharomyces cerevisiae	6,275
Arabidopsis thaliana	~ 20,000
Caenorhabditis elegans	19,099
Haemophilus influenzae	1,743
Drosophila melanogaster	13,601
Neisseria meningitdis	2,158
Homo sapiens	20,000–25,000

Genome Projects

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

Genome Projects

- Projects:
  - Over 15,000 genome projects in progress or completed



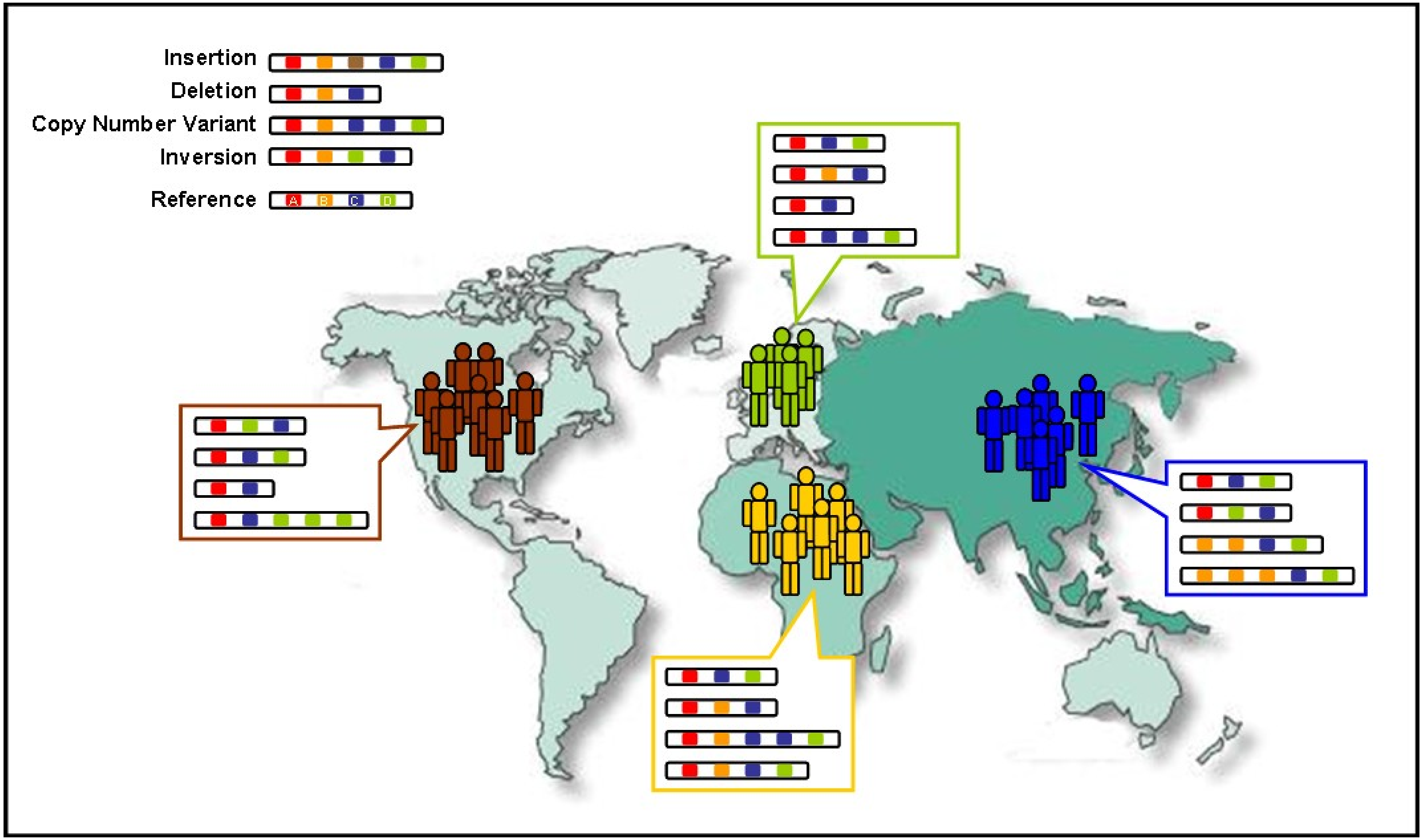
Genome Information by organism

Organism/Name	Kingdom All	Group All	SubGroup 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

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

Genome Projects

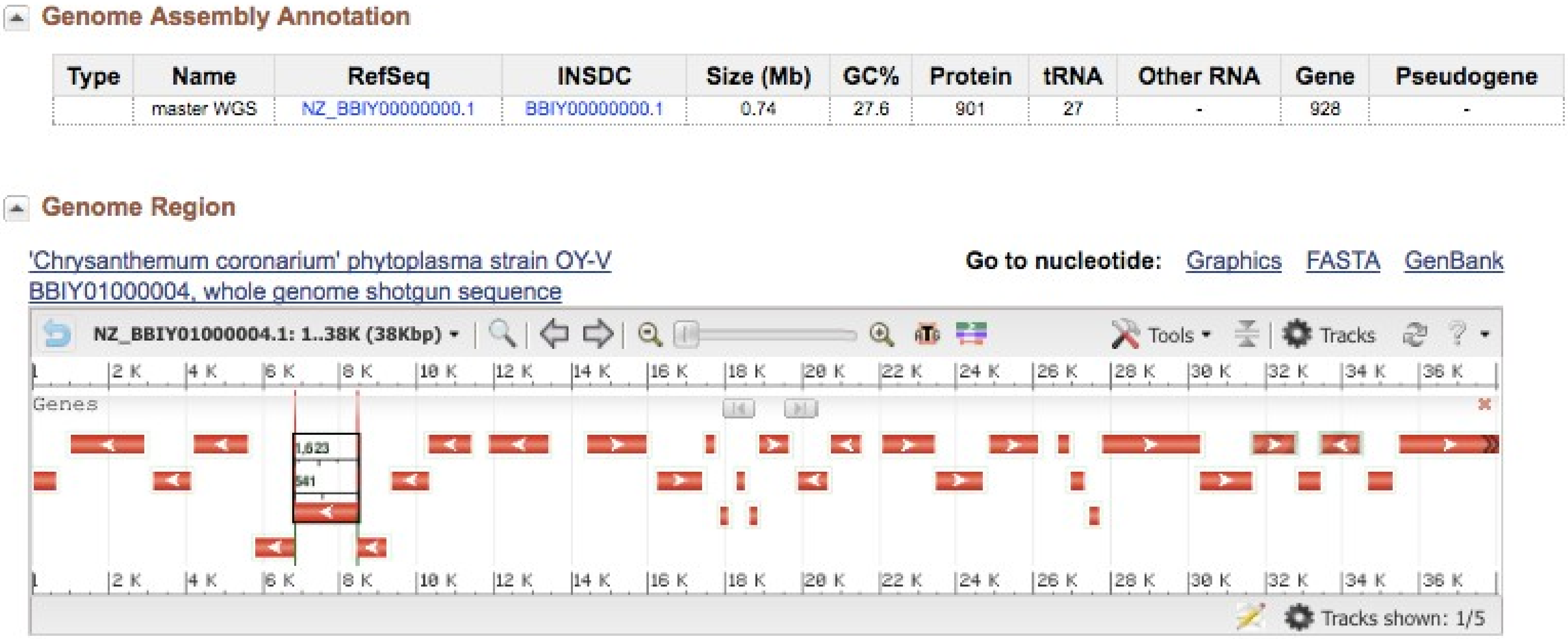
- Contrast genetic material of populations to determine ancestry



https://en.wikipedia.org/wiki/1000_Genomes_Project#Human_genetic_variation

Genome Projects: Data

- Locate genes for proteins in sequences.



Genome Assembly Annotation

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	tRNA	Other RNA	Gene	Pseudogene
	master WGS	NZ_BBIY00000000.1	BBIY00000000.1	0.74	27.6	901	27	-	928	-

Genome Region

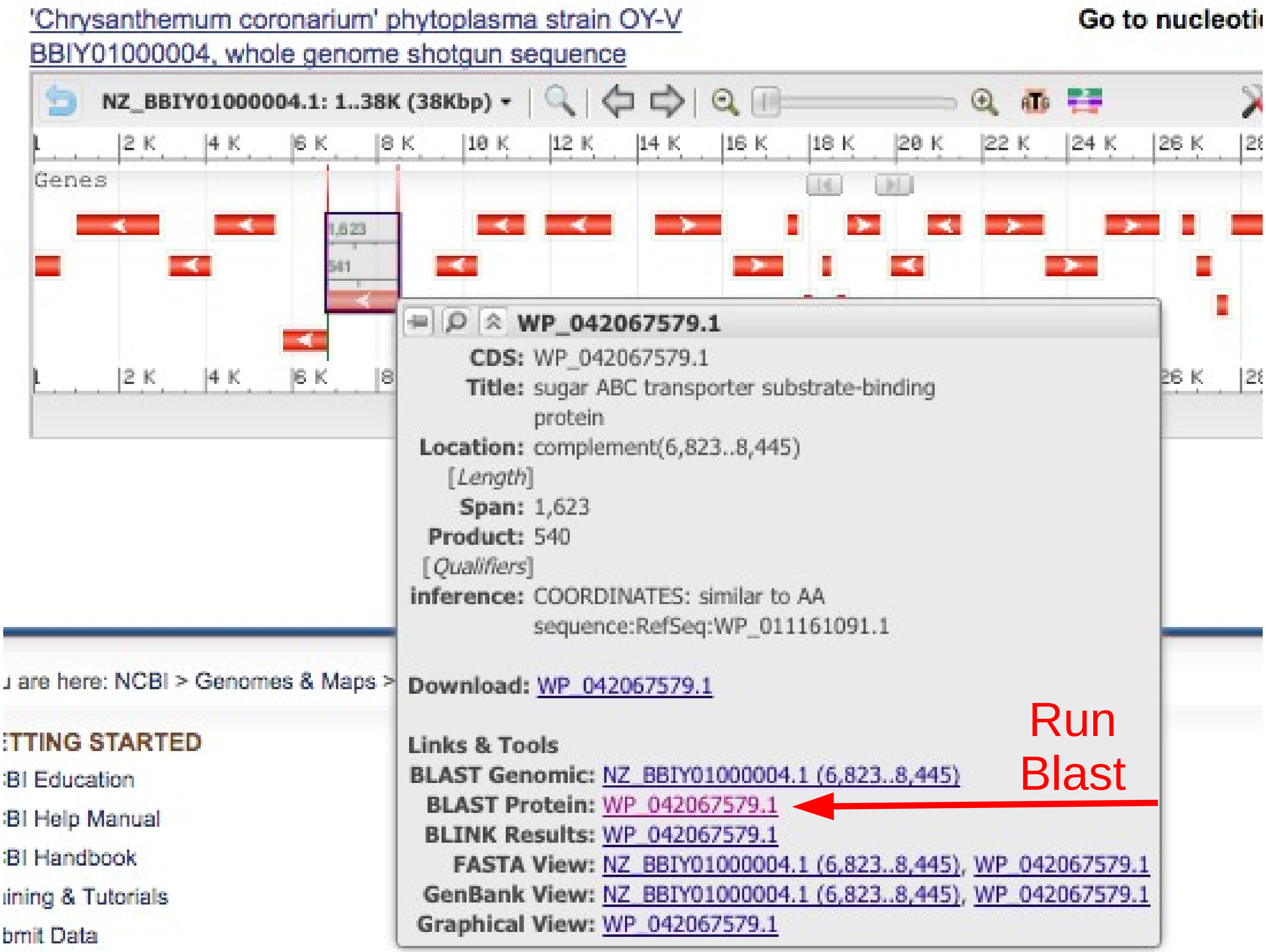
'Chrysanthemum coronarium' phytoplasma strain OY-V BBIY01000004, whole genome shotgun sequence

Go to nucleotide: Graphics FASTA GenBank

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

Genome Projects: Data

- Protein metadata



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

Go to nucleoti

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

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

ı are here: NCBI > Genomes & Maps >

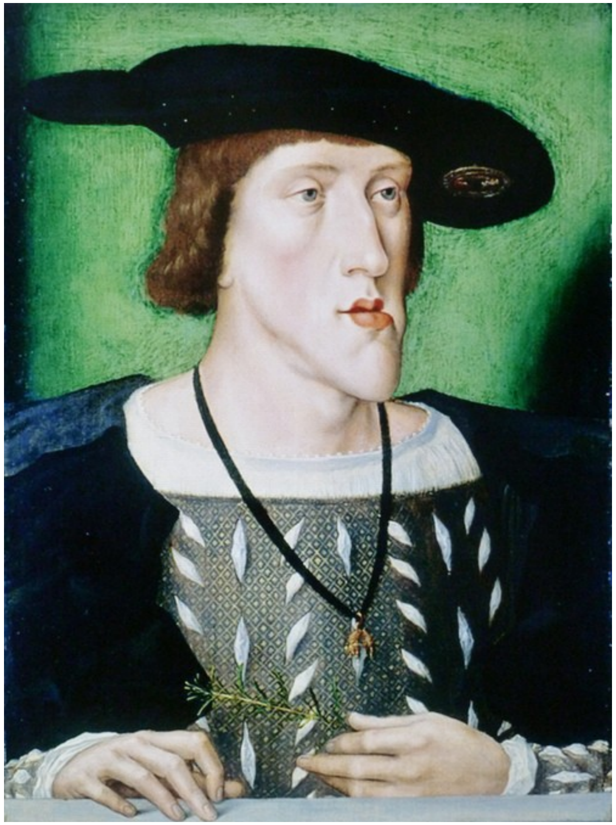
ETTING STARTED
BI Education
BI Help Manual
BI Handbook
ining & Tutorials
bmit Data

Human Genetic Variation

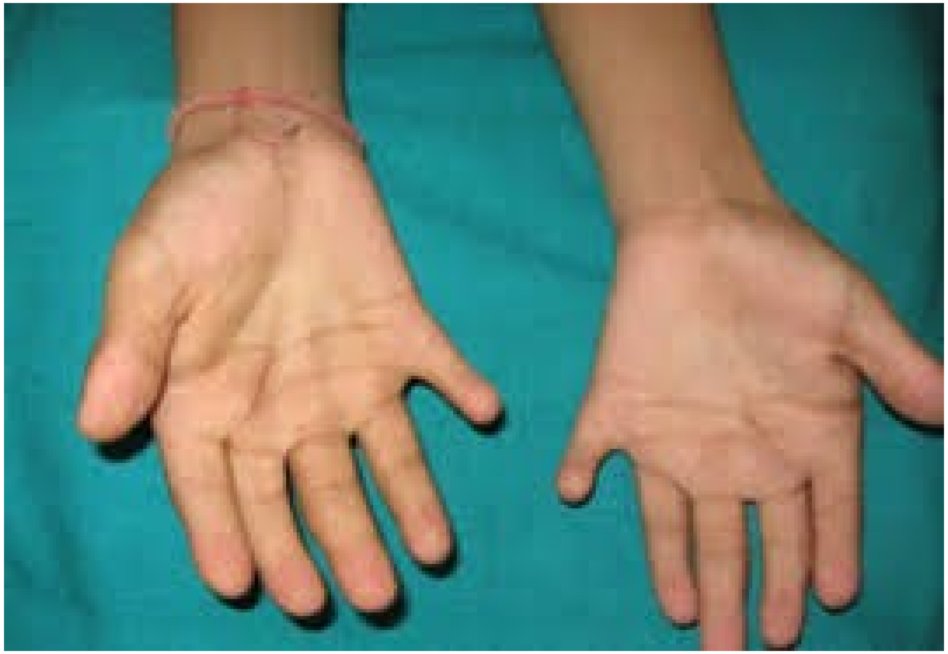
- Having diverse human genetic information helps to spot genetic conditions
- Genetic drift: a random fluctuation in the population frequency of a trait
  - Occurring in subsequent generations and would result in the loss of all variation in the absence of external influence

Detection By Comparison

- Early detection of genetic problems by being able to compare genomes to some “wild-type” genome.



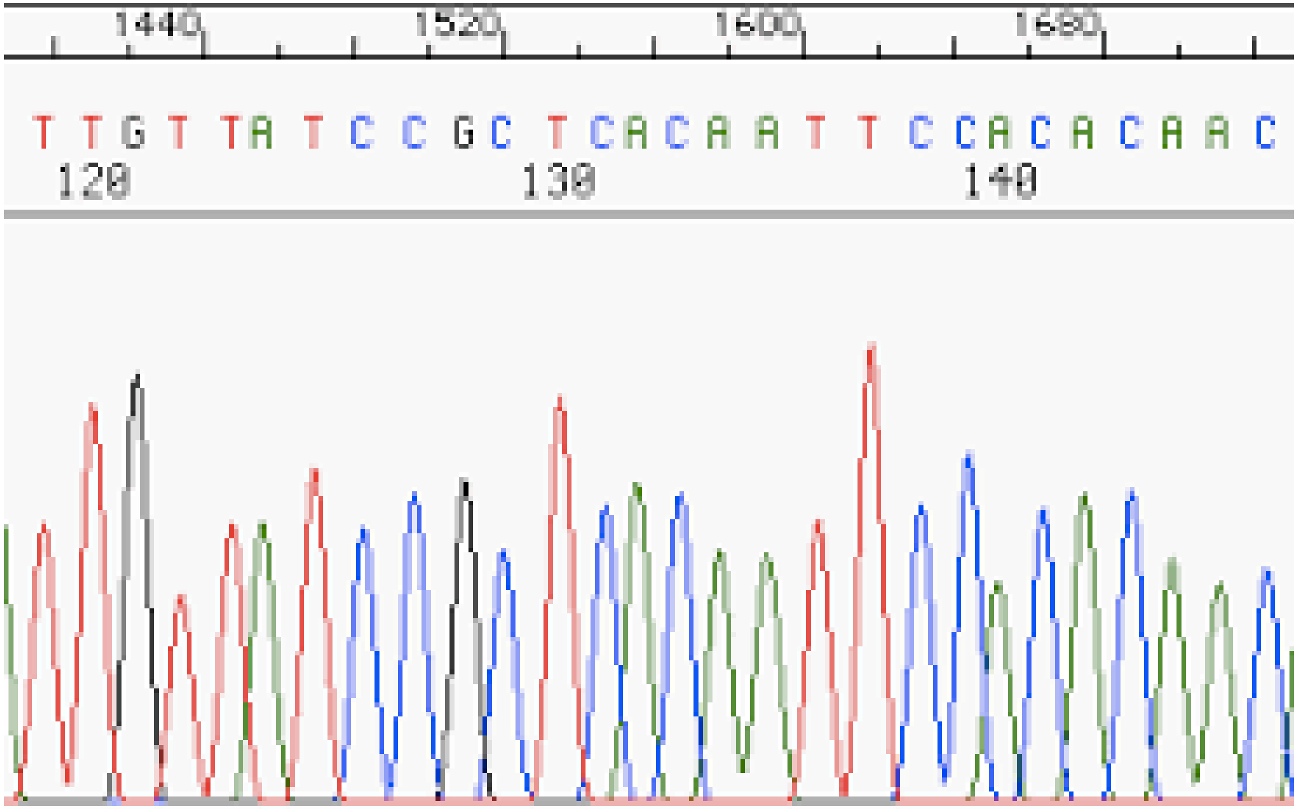
Hapsburg jaw



Ellis-Van Creveld syndrome, a sixth finger

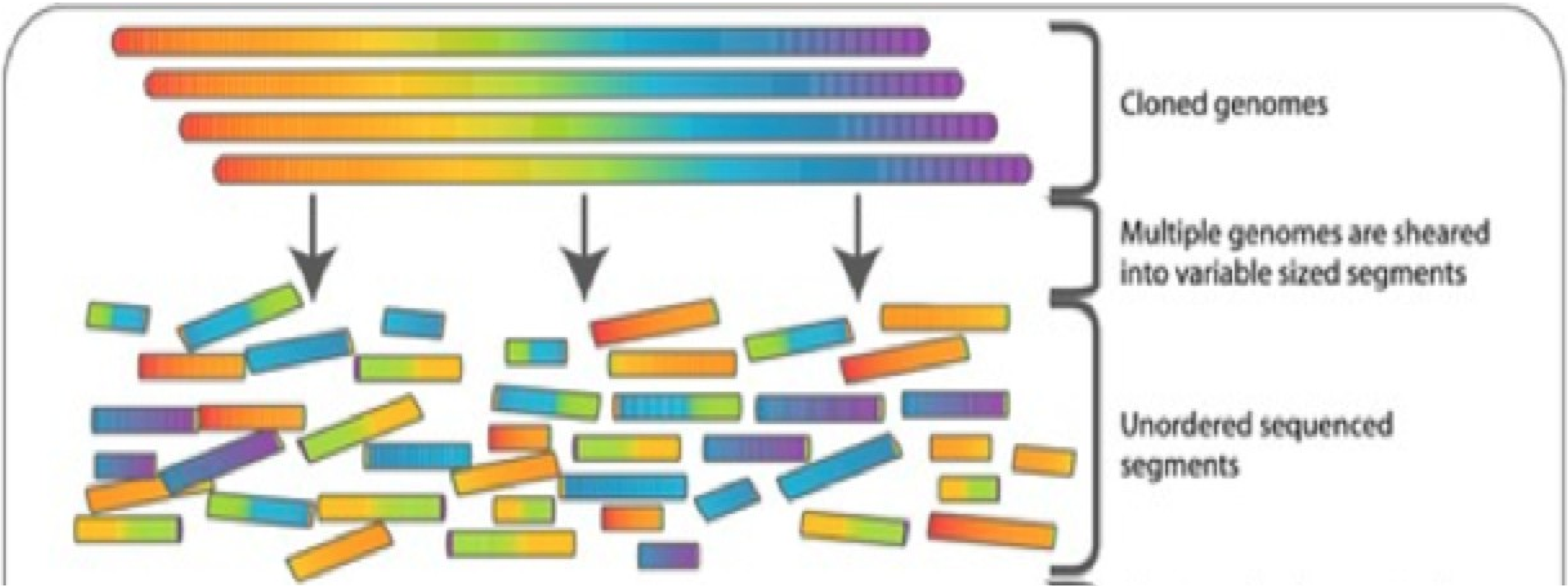
Genome Sequencing

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



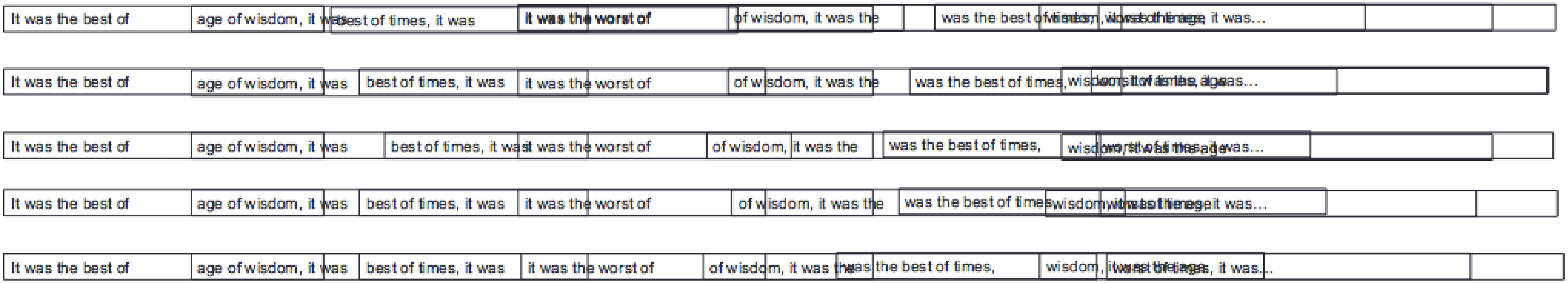
Genome Sequencing

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


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments

Shredded Book Reconstruction

- Dickens accidentally shreds first printing of Tale of Two Cities


It was the best of | age of wisdom, it was | best of times, it was | it was the worst of | of wisdom, it was the | was the best of | times, | wisdom, | worst of | the age, | it was…
It was the best of | age of wisdom, it was | best of times, it was | it was the worst of | of wisdom, it was the | was the best of times, | wisdom, | worst of times, | it was the age | it was…
It was the best of | age of wisdom, it was | best of times, it was | it was the worst of | of wisdom, it was the | was the best of times, | wisdom, it was the age | worst of times, it was…
It was the best of | age of wisdom, it was | best of times, it was | it was the worst of | of wisdom, it was the | was the best of times | wisdom, | worst of times, | it was the age | it was…
It was the best of | age of wisdom, it was | best of times, it was | it was the worst of | of wisdom, it was the | was the best of times, | wisdom, | it was the age | worst 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

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

the best of times, it

best of times, it was

Assembly Parameter:
100% identify across 4 words

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

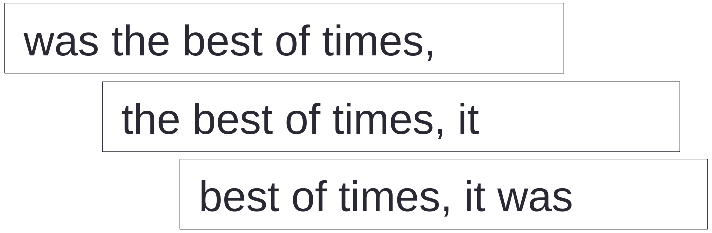
was the worst of times,

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

Assembly Parameter:
100% identify across 4 words

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

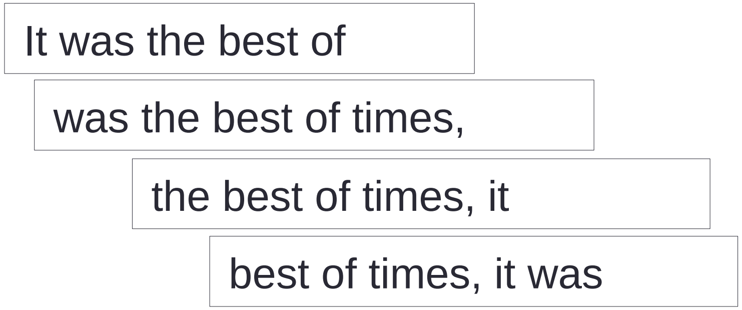
was the worst of times,

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

Assembly Parameter:
100% identify across 4 words

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.

Repeats pile up – 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

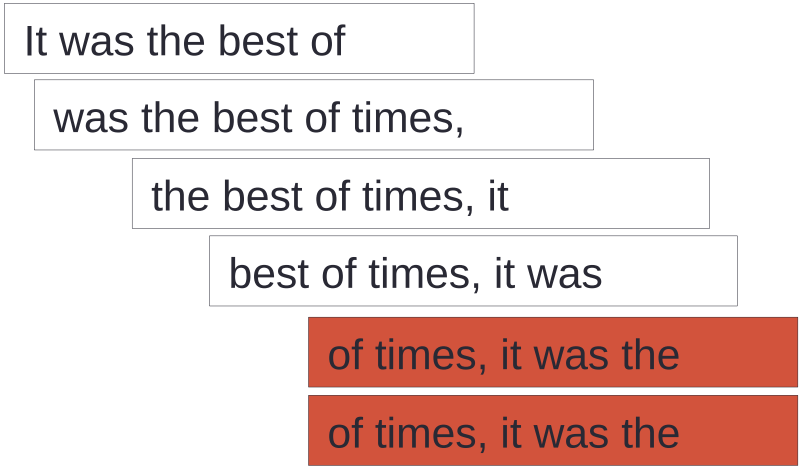
was the worst of times,

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

Assembly Parameter:
100% identify across 4 words

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,

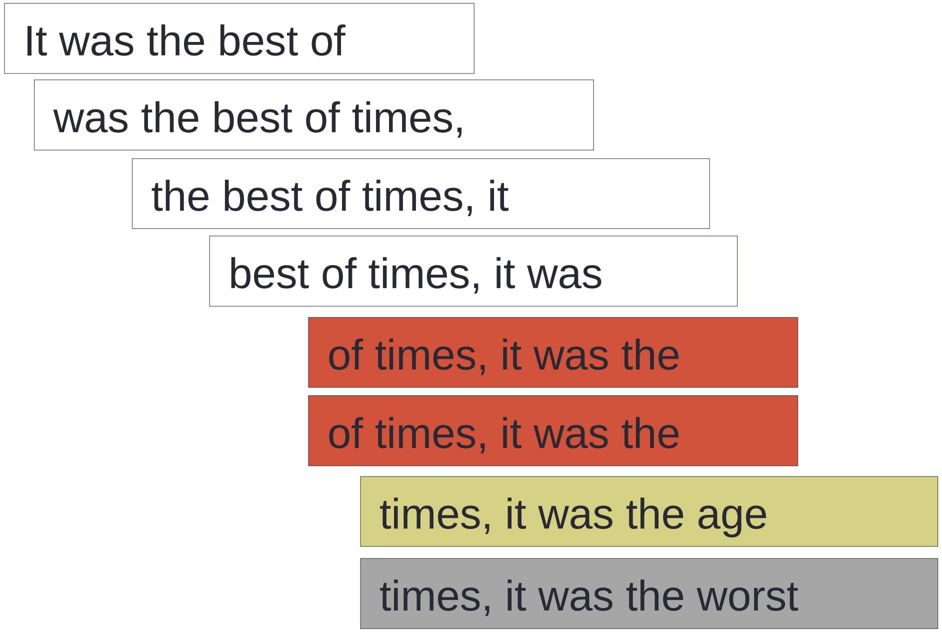
was the age of foolishness,

was the age of wisdom

times, it was the worst

Repeats pile up – actual placement of each individual fragment unknown

Repeats can cause ambiguity and prevent proper assembly



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

Assembly Parameter:
100% identify across 4 words

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,

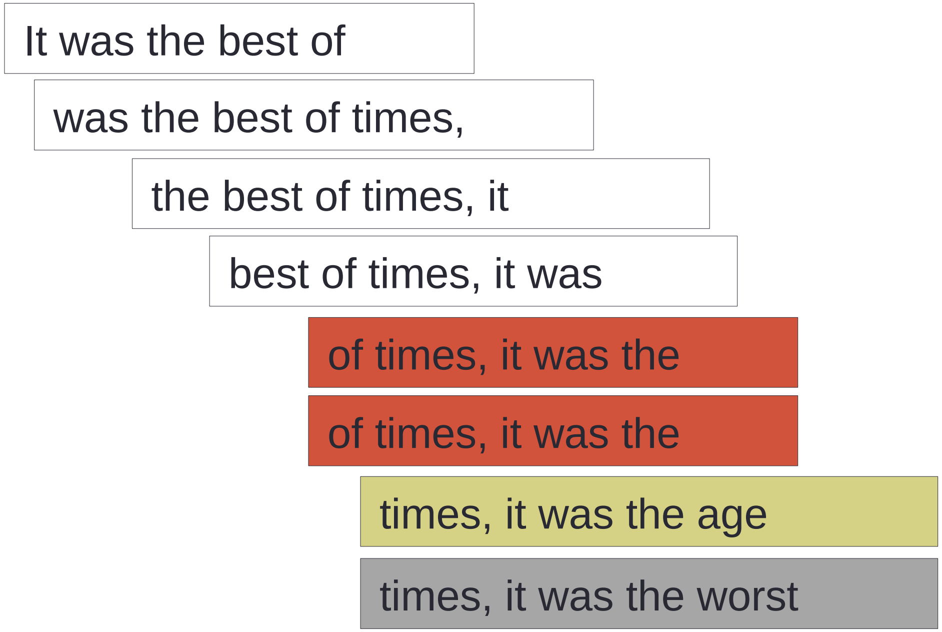
was the age of foolishness,

was the age of wisdom

times, it was the worst

Repeats pile up – actual placement of each individual fragment unknown

Repeats can cause ambiguity and prevent proper assembly


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

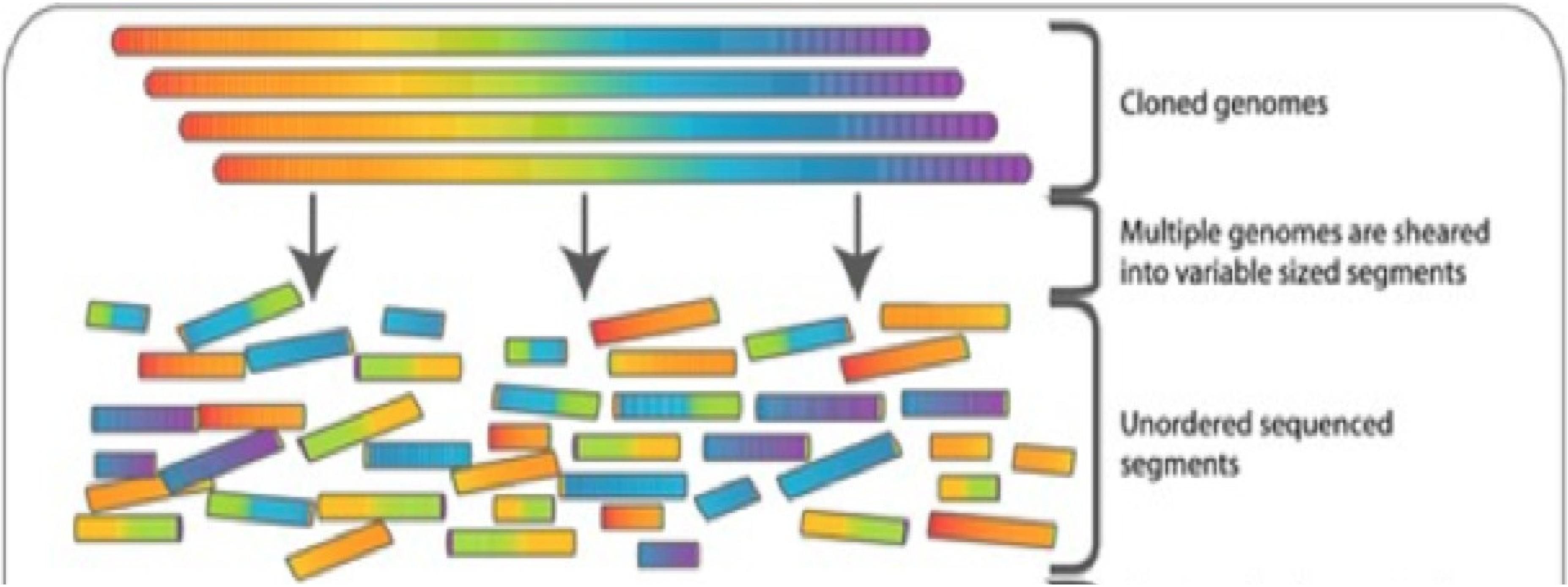
times, it was the age

times, it was the worst

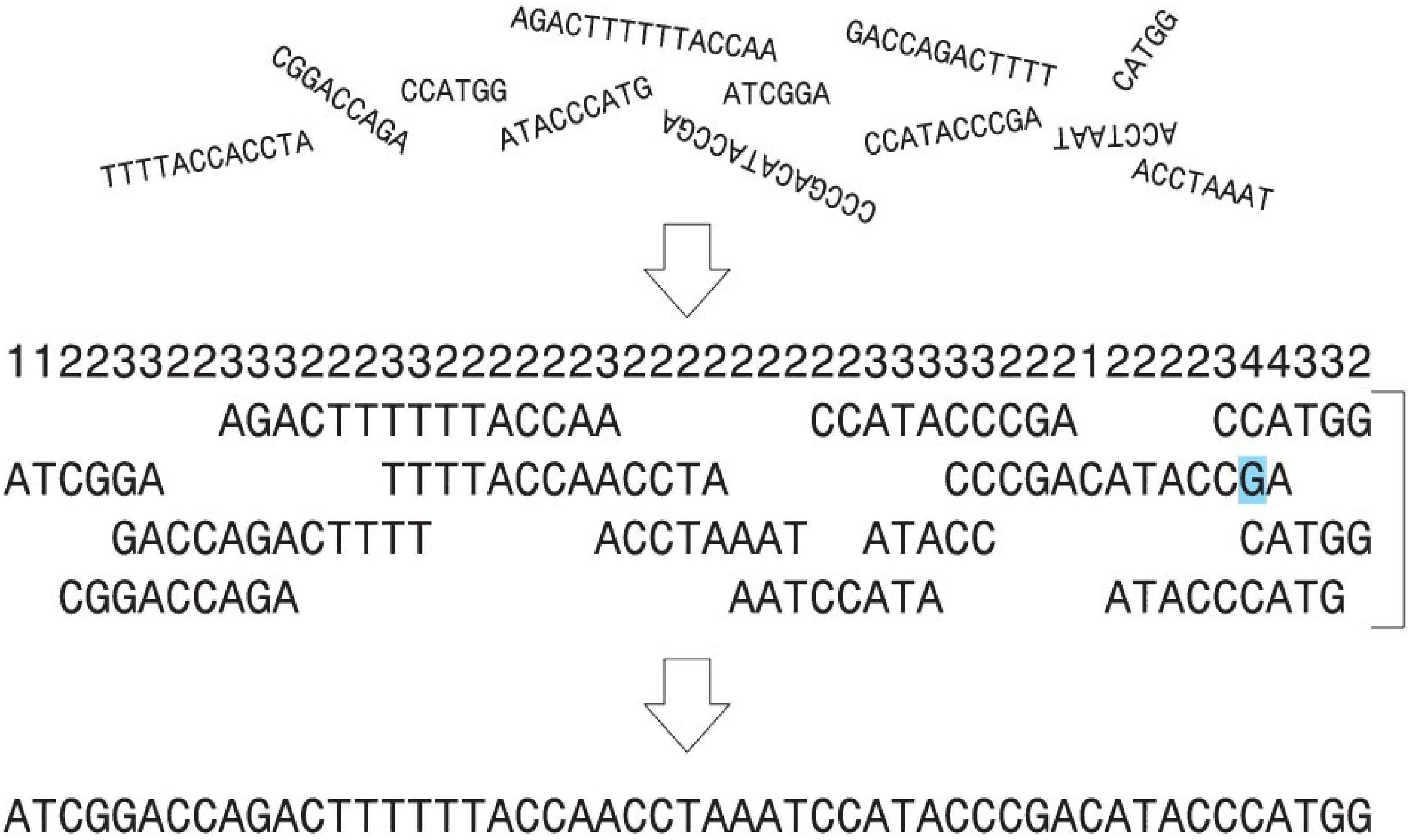
It was the best of times, it was the [age/worst]

Assembly Parameter:
100% identify across 4 words

Genome Sequencing


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments

Coverage



TTTTACCACCTA
CGGACCAGA
CCATGG
AGACTTTTTTACCAA
ATACCCATG
CCCGACATACCGA
ATCGGA
GACCAGACTTTT
CCATACCCGA
CATGG
TAATCCA
ACCTAAT

random short sequence reads

```
1122332233322233222222322222222223333222122222344332
         AGACTTTTTTACCAA           CCATACCCGA        CCATGG
ATCGGA             TTTTACCAACCTA             CCCGACATACCGA
    GACCAGACTTTT           ACCTAAT  ATACC                CATGG
  CGGACCAGA                            AATCCATA        ATACCCATG
```

coverage

assembly of overlapping fragments

```
ATCGGACCAGACTTTTTTACCAACCTAAATCCATACCCGACATACCCATGG
```

assembled contig sequence

Assembling a Contig

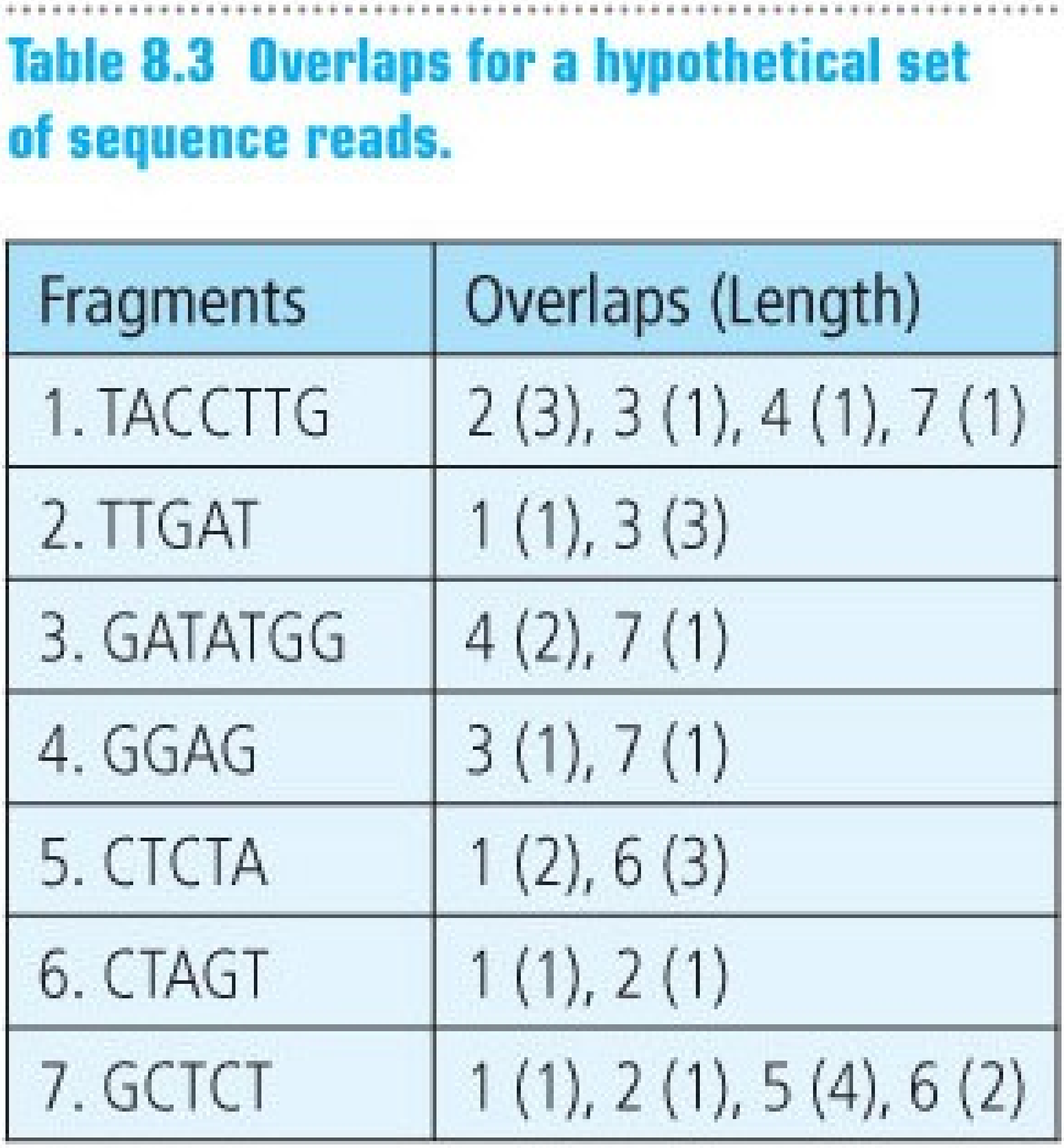
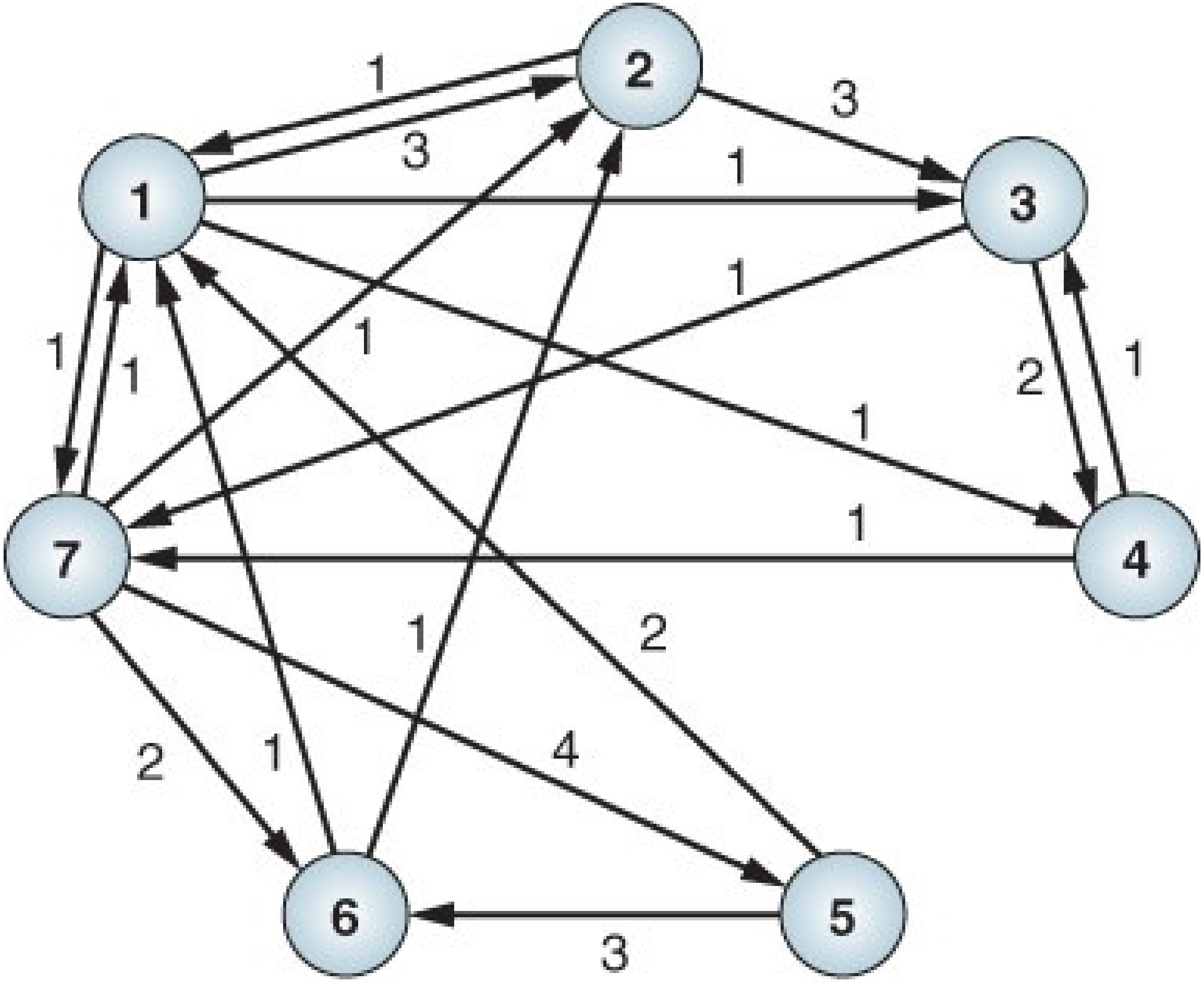


Table 8.3 Overlaps for a hypothetical set of sequence reads.

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

Assembling a Contig: graph representation



Your Turn to Investigate!!!

- Investigate the reads included in your sandbox (file: ***SH_reads.txt***).
- Questions:
  - What is the quote?
  - How did you determine this quote?

The Sherlock Holmes Museum
Souvenirs Books Antiques and Curios

GitHub Activity Repository:
https://classroom.github.com/a/DuBJW7yi
Due at 12:30 on 31 Oct. 2019
Create a directory: **act2**
Create work file: **act2/readWork.md**