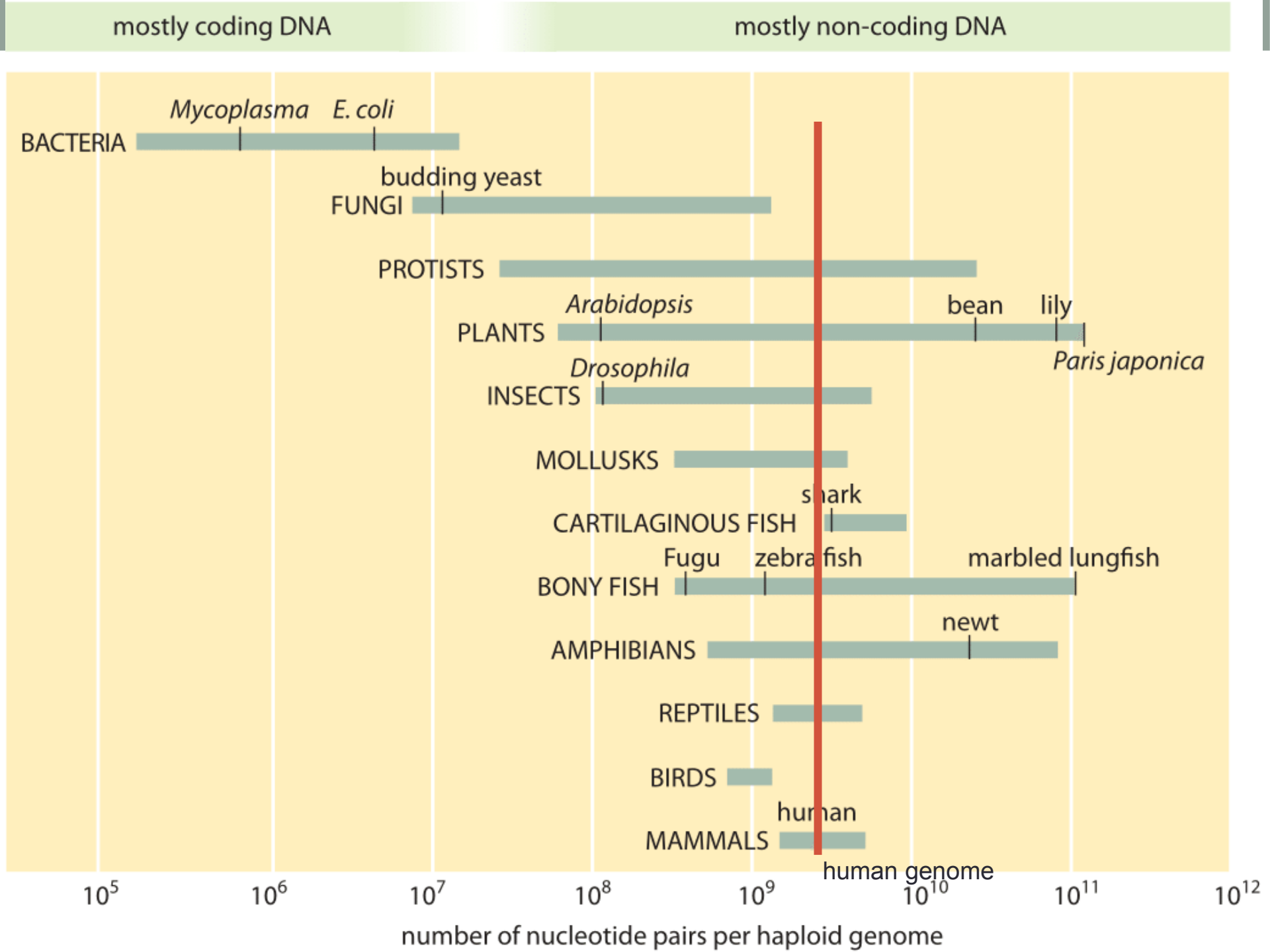
GENOME SEQUENCING AND ASSEMBLY

BIO 300/CMPSC 300

Spring 2016

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
zebrafish
marbled lungfish
BONY FISH
newt
AMPHIBIANS
REPTILES
BIRDS
human
MAMMALS
human genome
10^5
10^6
10^7
10^8
10^9
10^{10}
10^{11}
10^{12}
number of nucleotide pairs per haploid 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

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

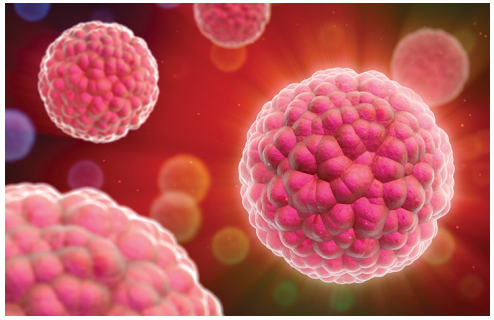
- Projects:
  - Over 15,000 genome projects in progress or completed

Genome Projects

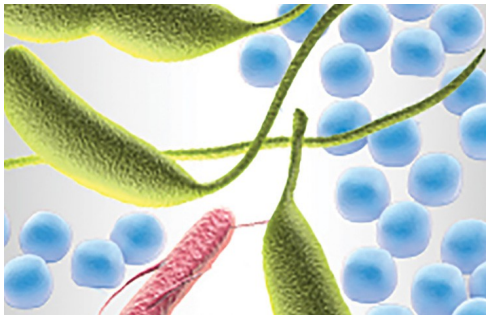
- Goals:
  - Determine complete genome sequence of an organism
  - Annotate protein-coding genes and other important genome-encoded features

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

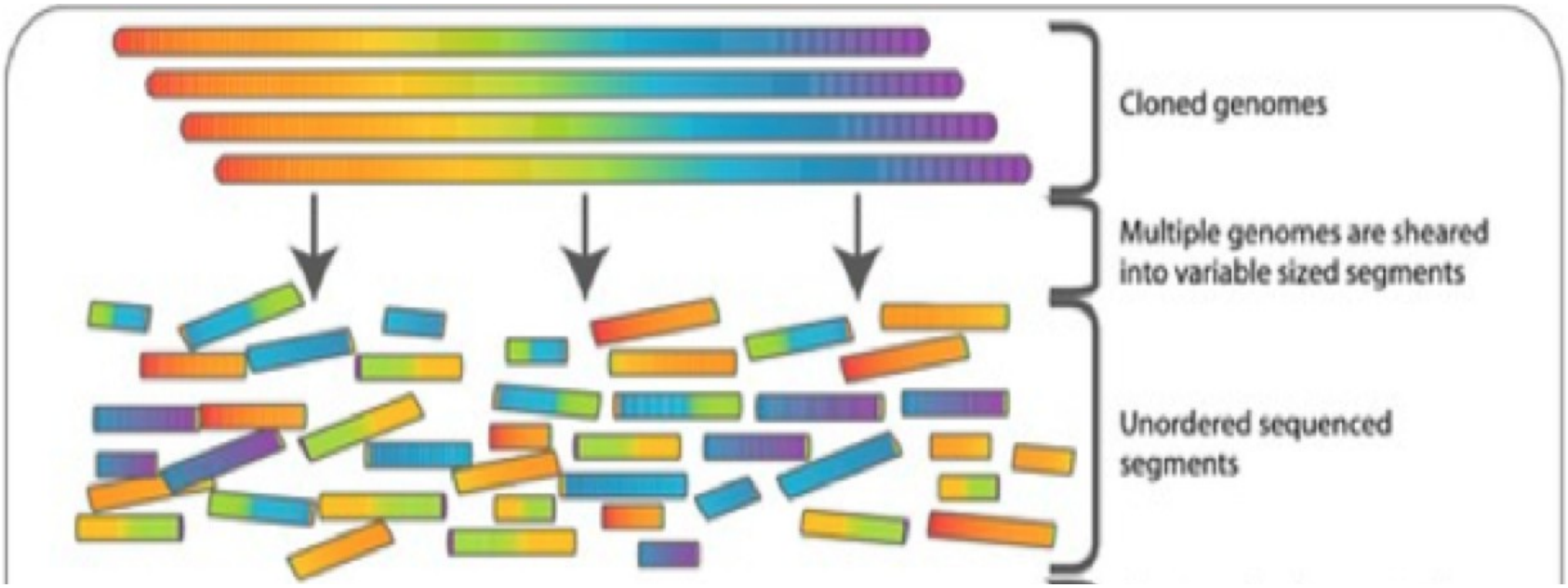
Cancer Genomics



Shotgun Metagenomics



Genome Sequencing

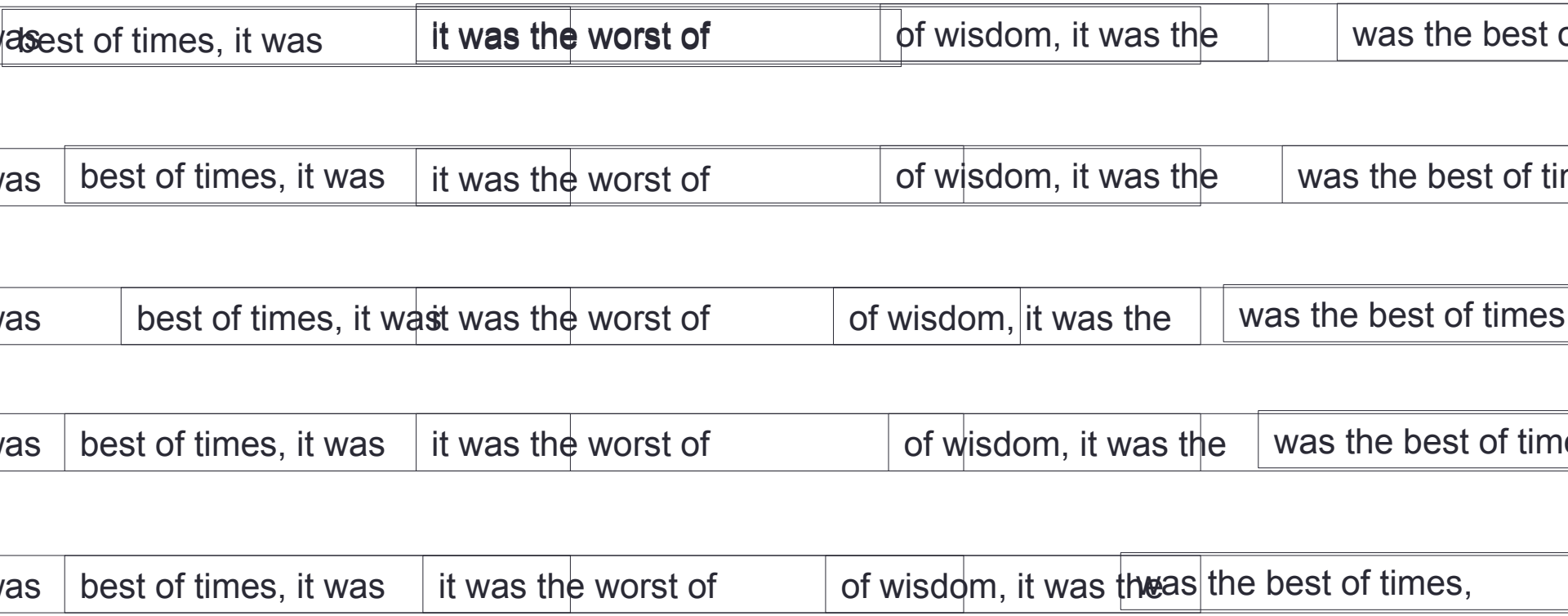


Tale of Two Cities – Charles Dickens

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.

Shredded Book Reconstruction

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



Shredded Book Reconstruction

- Dickens accidently 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 accidently 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

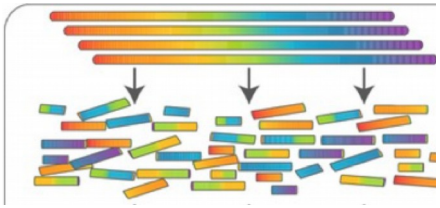
It was the best of times, it was the worst of times, it was the

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

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.

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

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

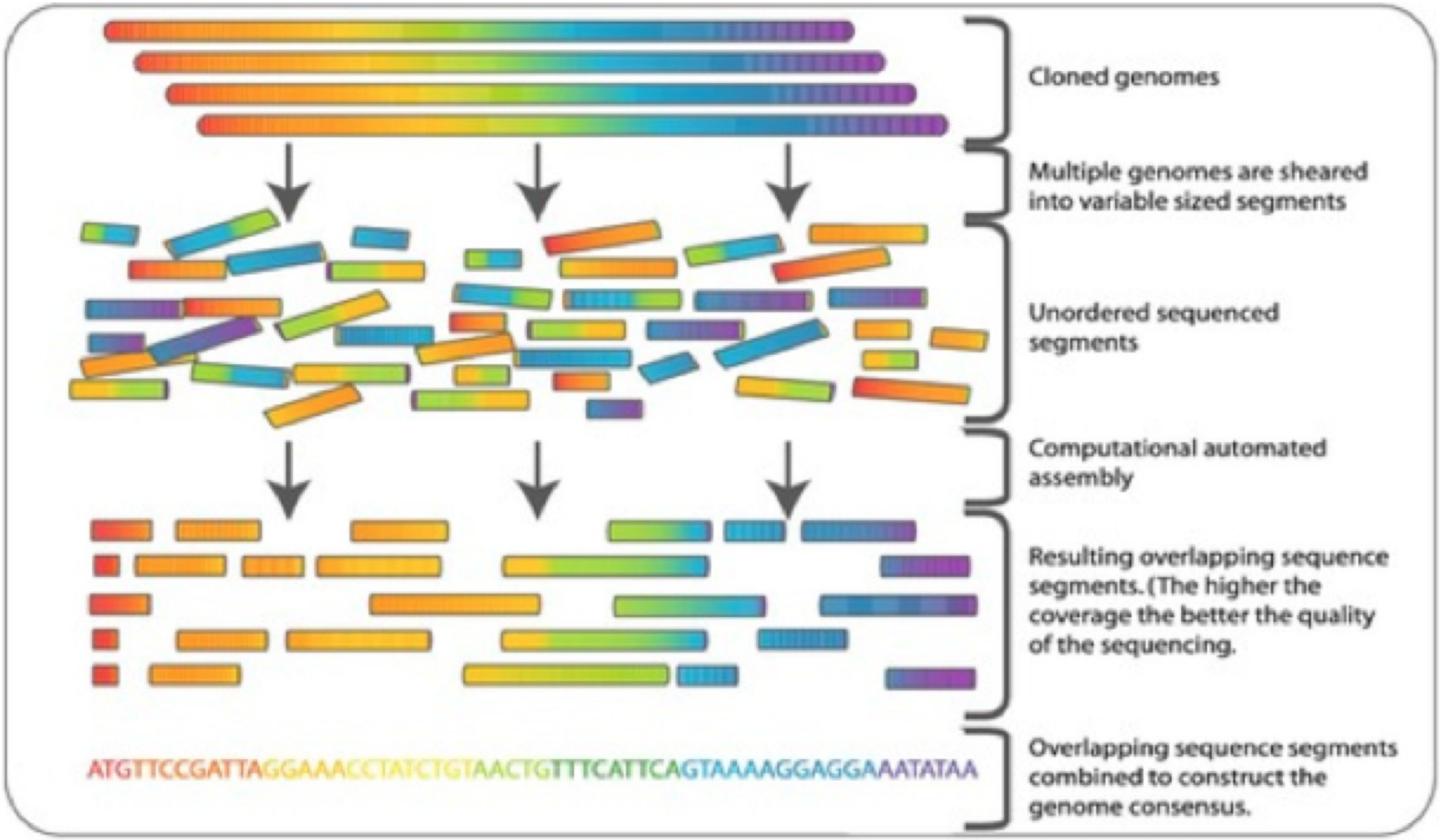
times, it was the age

times, it was the worst

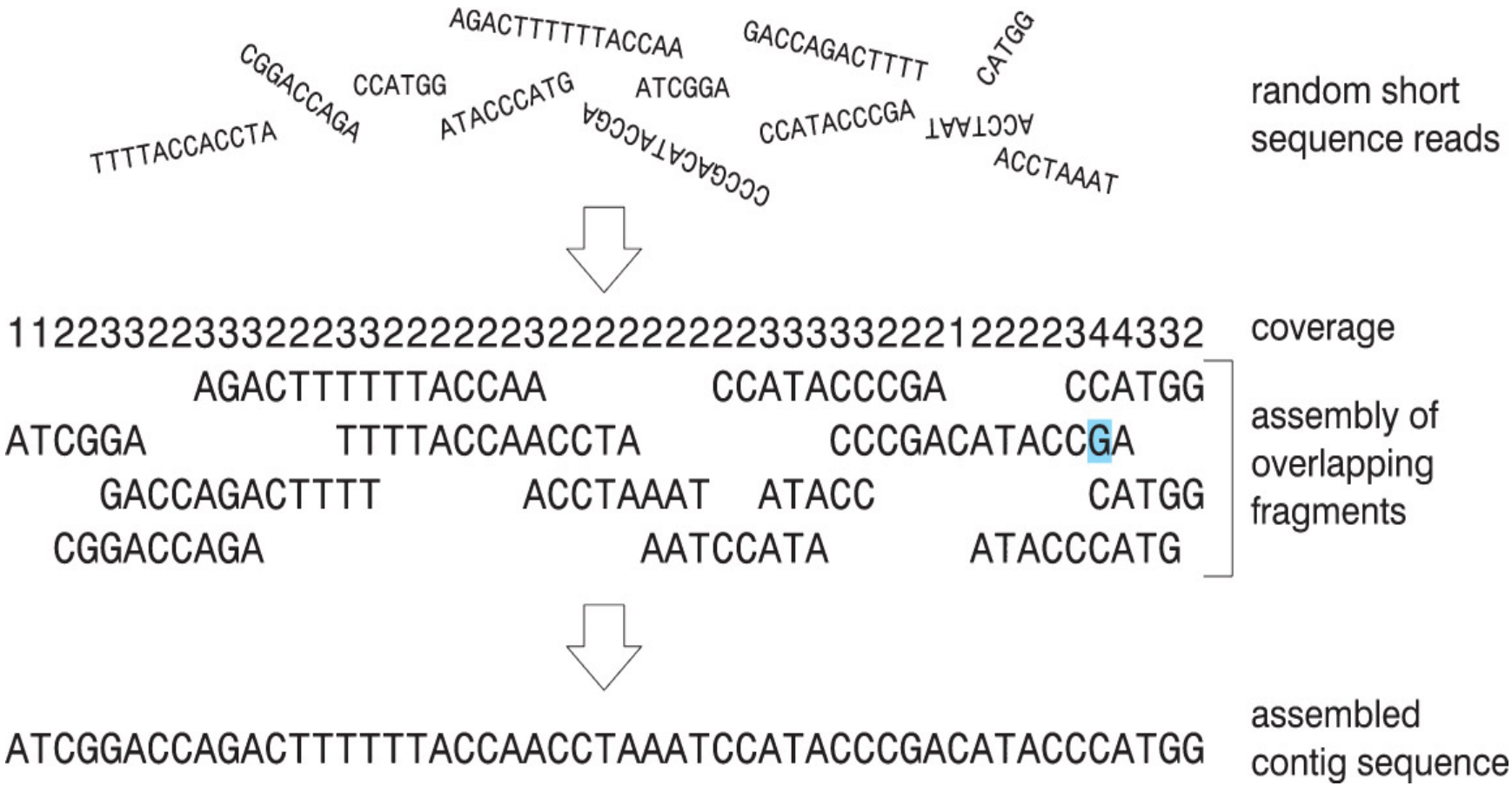
It was the best of times, it was the [age/worst]

Assembly Parameter:
100% identify across 4 words

Genome Sequencing



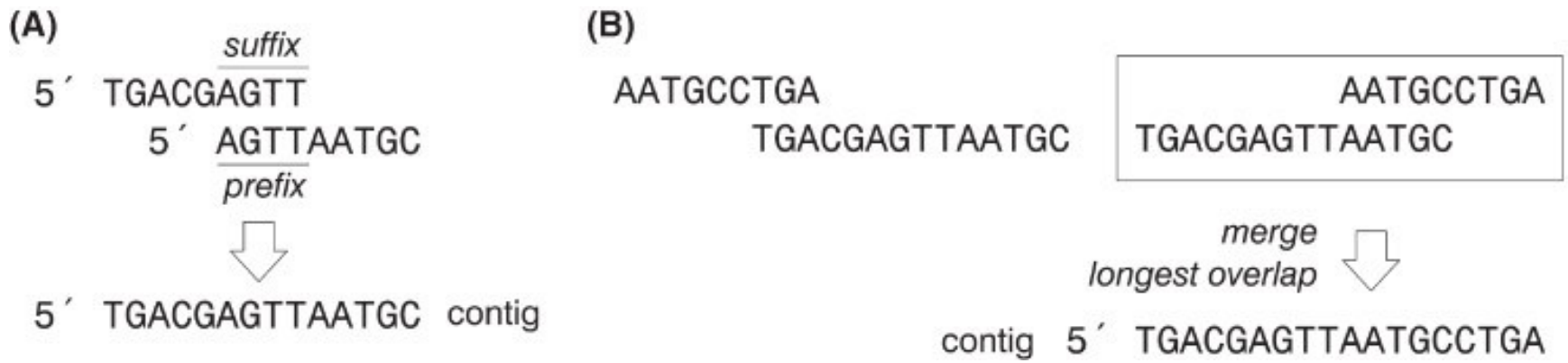
Coverage


AGACTTTTTTACCAA
GACCAGACTTTT
CATGG
CGGACCAGA
CCATGG
ATACCCATG
ATCGGA
TTTTACCACCTA
CCCGACATACCGA
CCATACCCGA
ACCTAAAT
random short sequence reads
112233223332223322222232222222223333322212222344332
coverage
AGACTTTTTTACCAA
CCATACCCGA
CCATGG
ATCGGA
TTTTACCAACCTA
CCCGACATACCGA
GACCAGACTTTT
ACCTAAAT
ATACC
CATGG
CGGACCAGA
AATCCATA
ATACCCATG
assembly of overlapping fragments
ATCGGACCAGACTTTTTTACCAACCTAAATCCATACCCGACATACCCATGG
assembled contig sequence

Finding the Largest Overlap

- <u>Consider two fragment assembly:</u>
  - 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-1

Finding the Largest Overlap



1. Start with 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, found the largest overlap
5. If the number of matches < n, n = n-1 If n=0 – no overlap, go to step 3

Finding the Largest Overlap

- Removing assumptions of identical subsequences and substrings:
  - set the initial **n** to the length of the sequence rather than the length of the sequence – 1.

Dealing with Noisy Sequencing Data

- Sequencing errors
- Ambiguities leading to incorrect base-calling

- Modify the algorithm so that the overlap exceeds some threshold value (instead of being perfect match)
  - Check if the number of matching bases is **threshold value x n**
  - With the **threshold value** being between 0 and 1

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)

Assembling a Contig: graph representation