بنام خدا

DNA Sequencing

سید ابوالفضل مطهری

Sequencing Problem

Sequencing Problem

$$S = X_1 X_2 X_3 X_4 \ldots X_{G-1} X_G$$

Sequencing Problem

$$S = X_1 X_2 X_3 X_4 \ldots X_{G-1} X_G$$

- finding the sequence

Sequencing Problem

$$S = X_1 X_2 X_3 X_4 \ldots X_{G-1} X_G$$

- finding the sequence
- or finding some parts of the sequence

Sequencing Problem

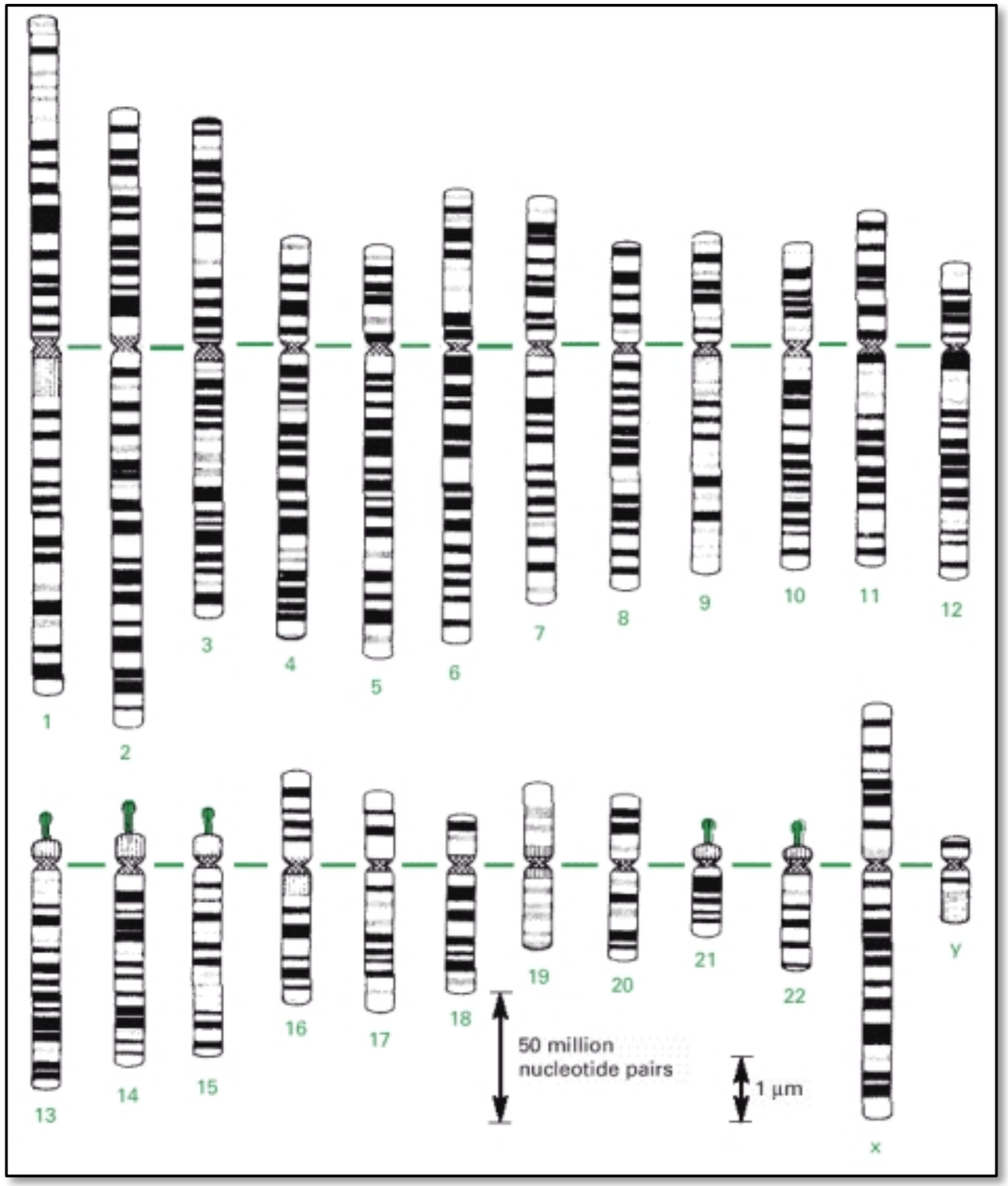
$$S = X_1 X_2 X_3 X_4 \ldots X_{G-1} X_G$$

- finding the sequence
- or finding some parts of the sequence

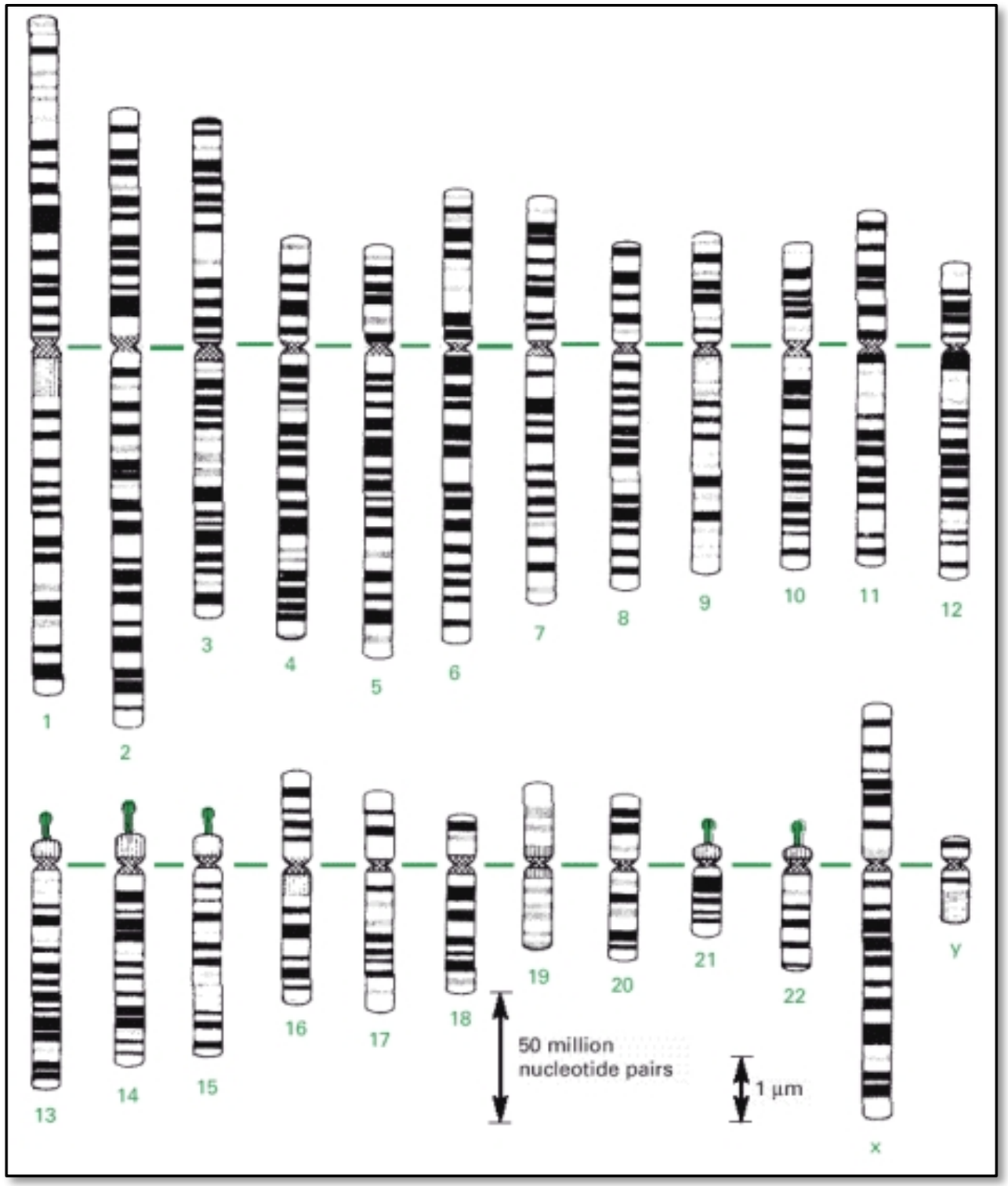
DNA Sequencing

sequence is the genome

Genome

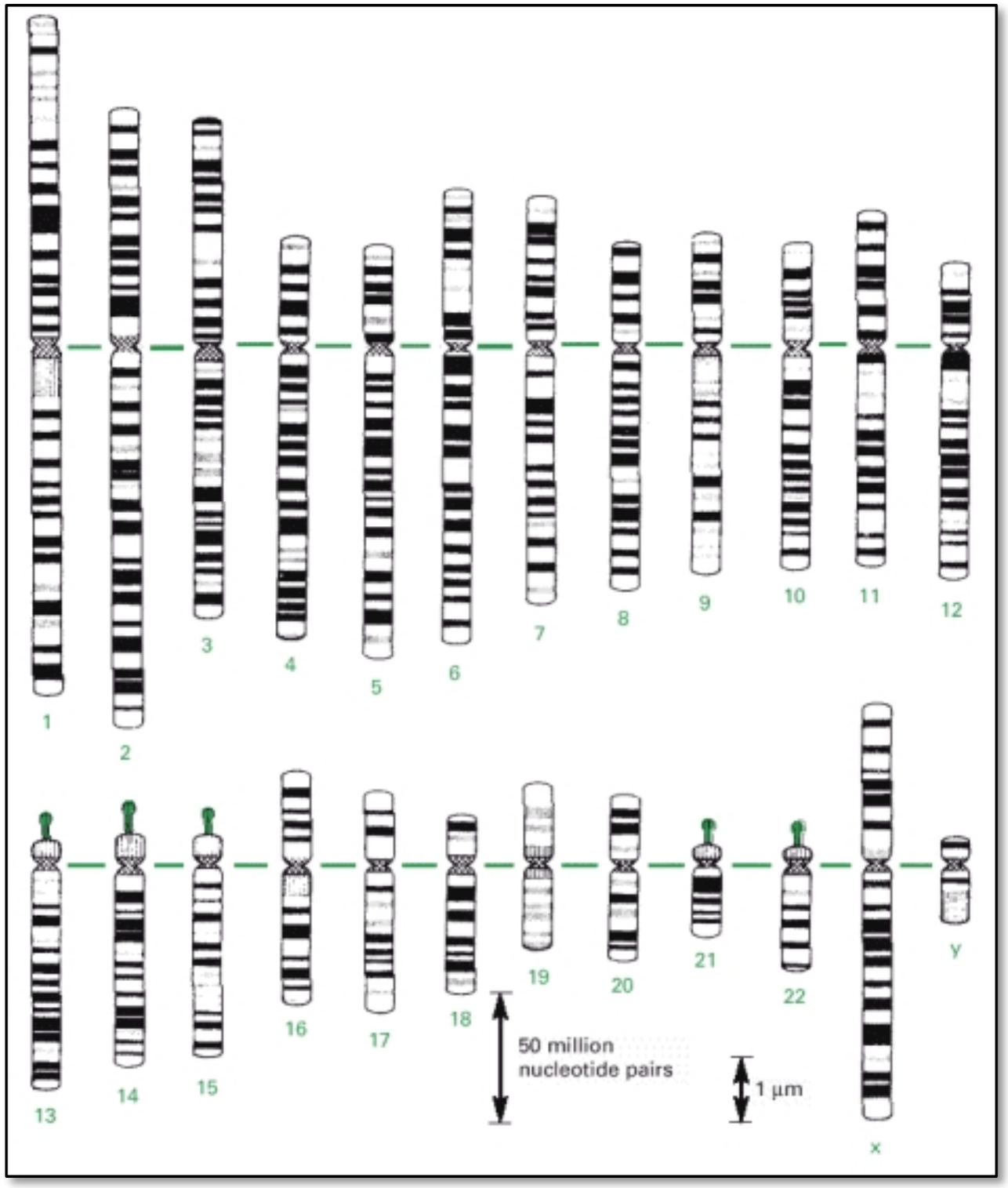

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
x
y
50 million
nucleotide pairs
1 µm

Genome



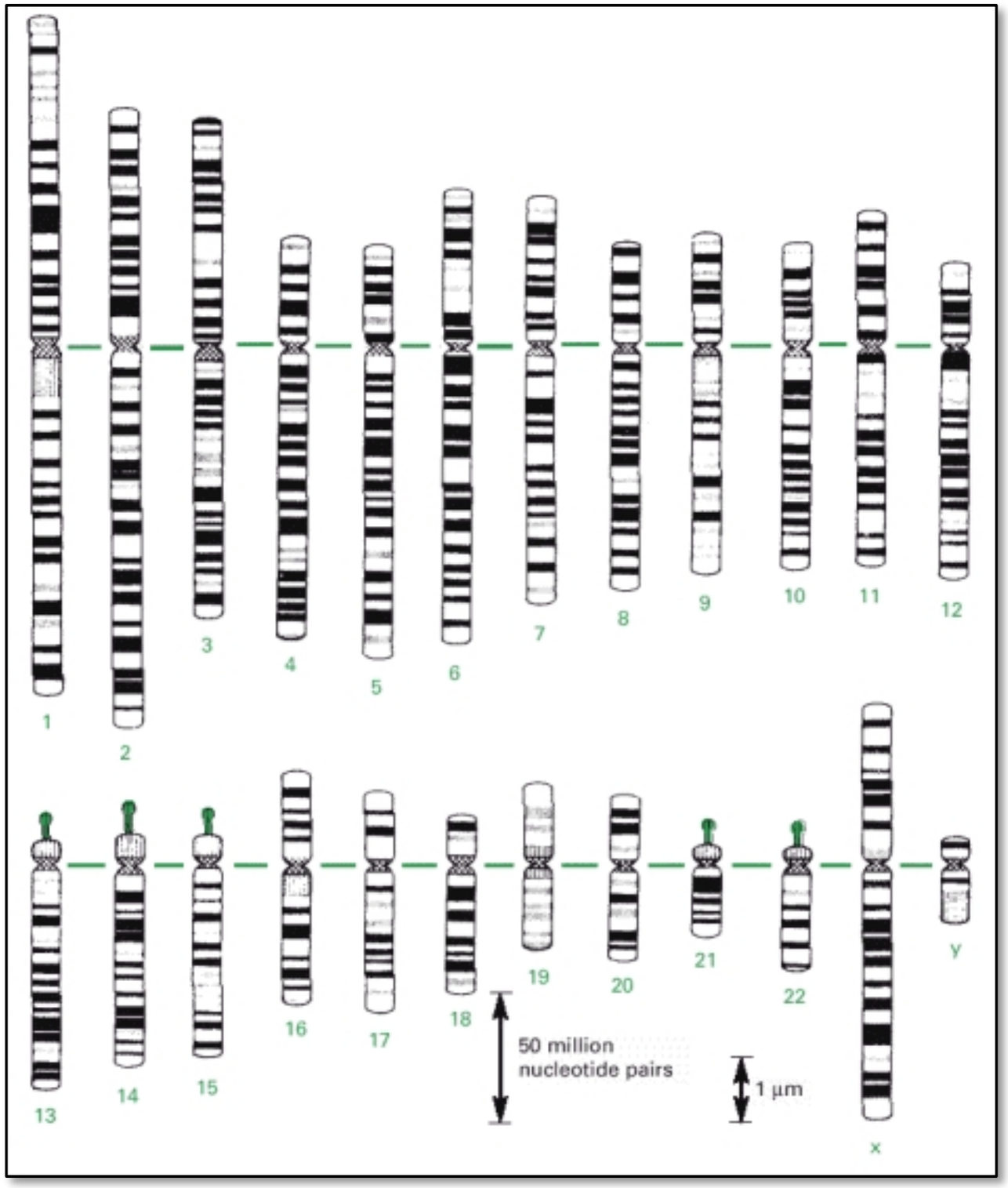
- Human Genome consists of 23 chromosomes

Genome



- Human Genome consists of 23 chromosomes
- Each chromosome contains a DNA sequence

Genome



- Human Genome consists of 23 chromosomes
- Each chromosome contains a DNA sequence
- The total length is

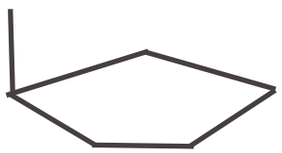
$$G = 3.2 \times 10^9$$

Genome

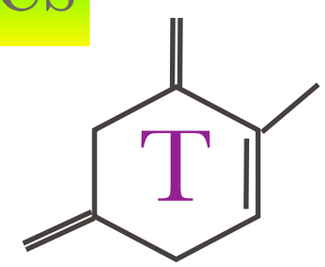
DexoyriboNucleic Acid (DNA)

Phosphate

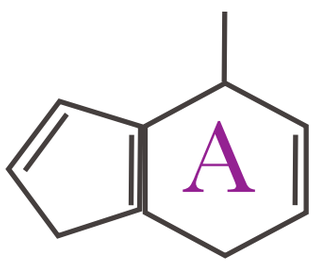
Sugar



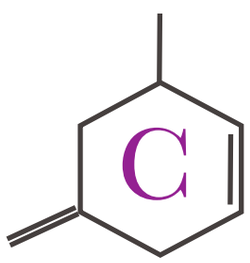
Bases



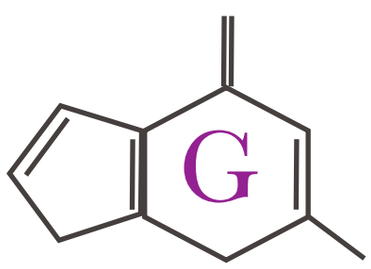
Thymine



Adenine



Cytosine

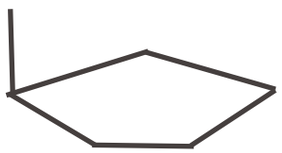


Guanine

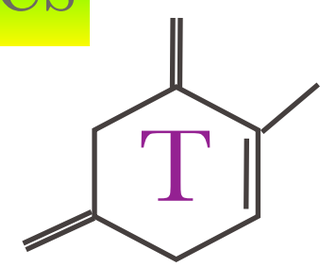
DexoyriboNucleic Acid (DNA)

Phosphate

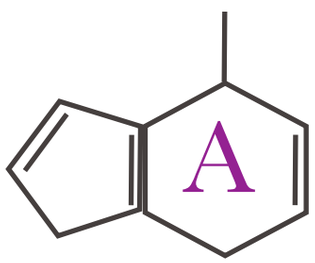
Sugar



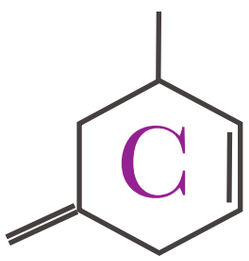
Bases



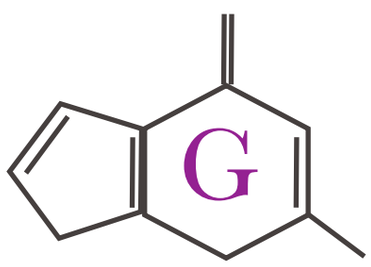
Thymine



Adenine

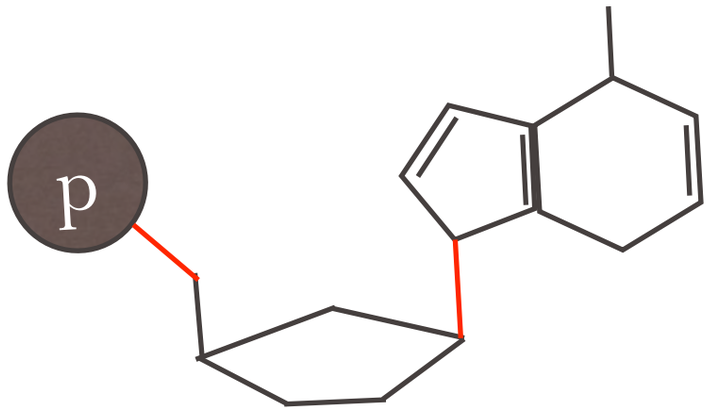


Cytosine



Guanine

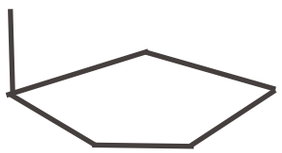
Covalent



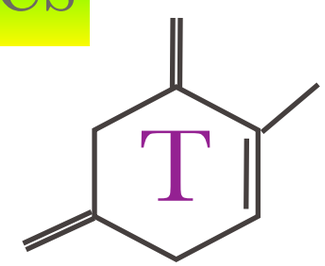
DexoyriboNucleic Acid (DNA)

Phosphate

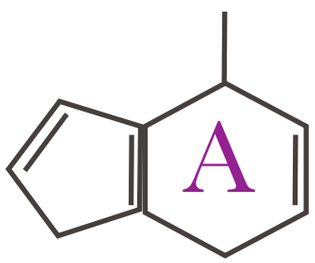
Sugar



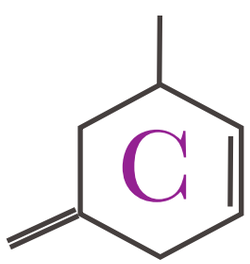
Bases



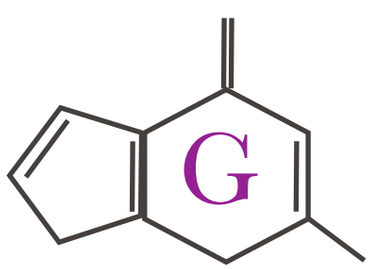
Thymine



Adenine

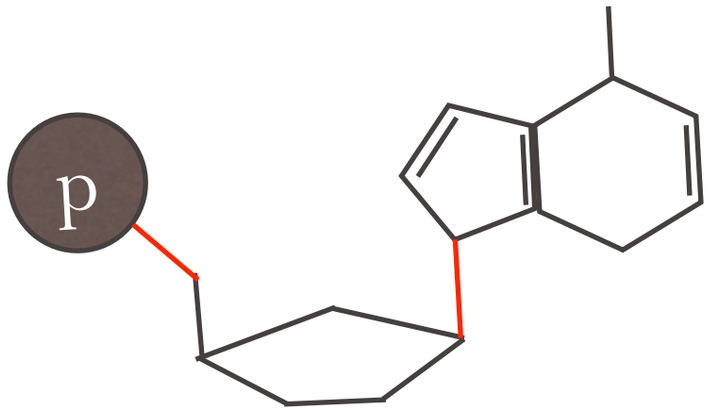


Cytosine

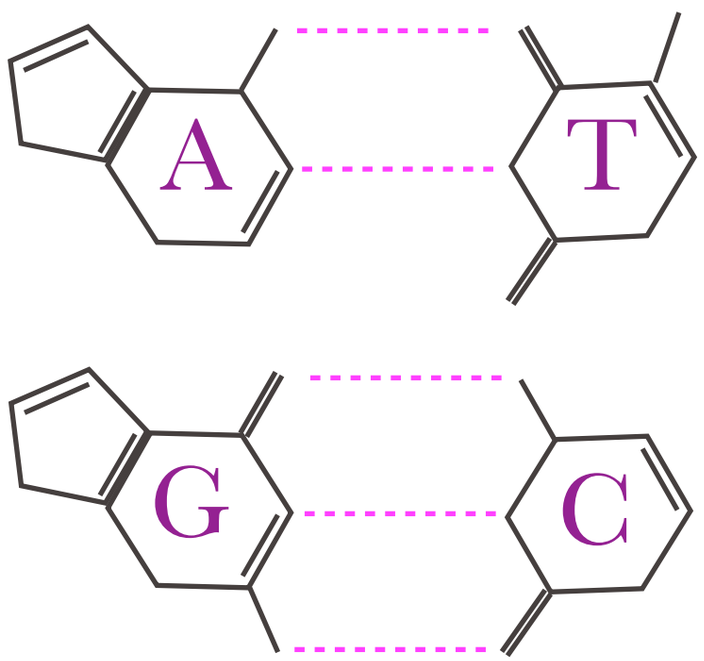


Guanine

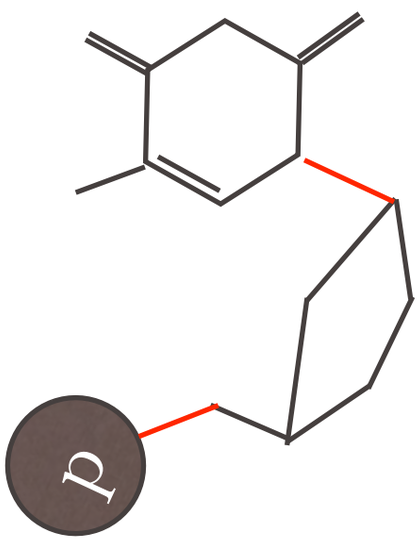
Covalent



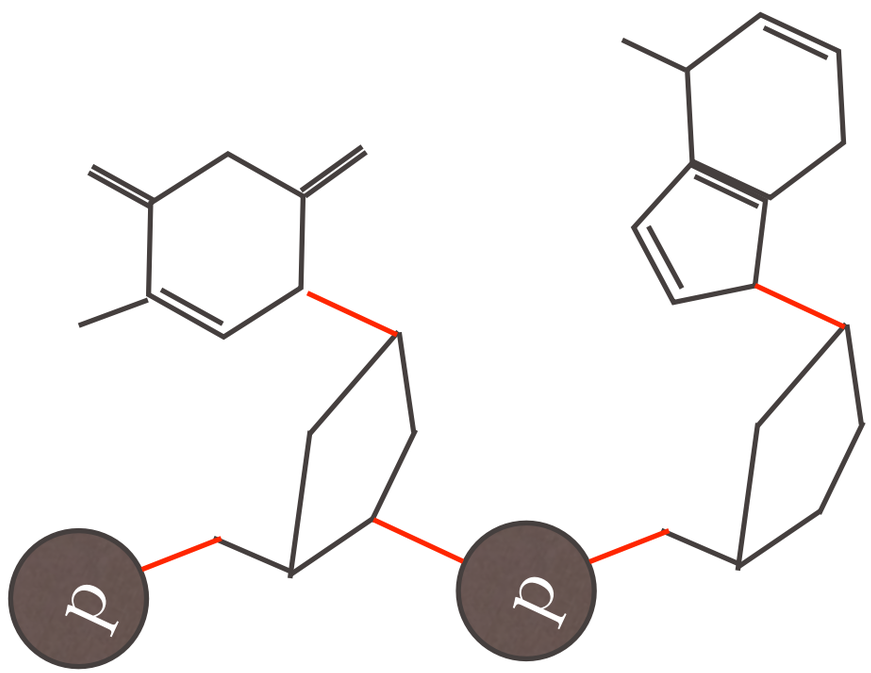
Hydrogen



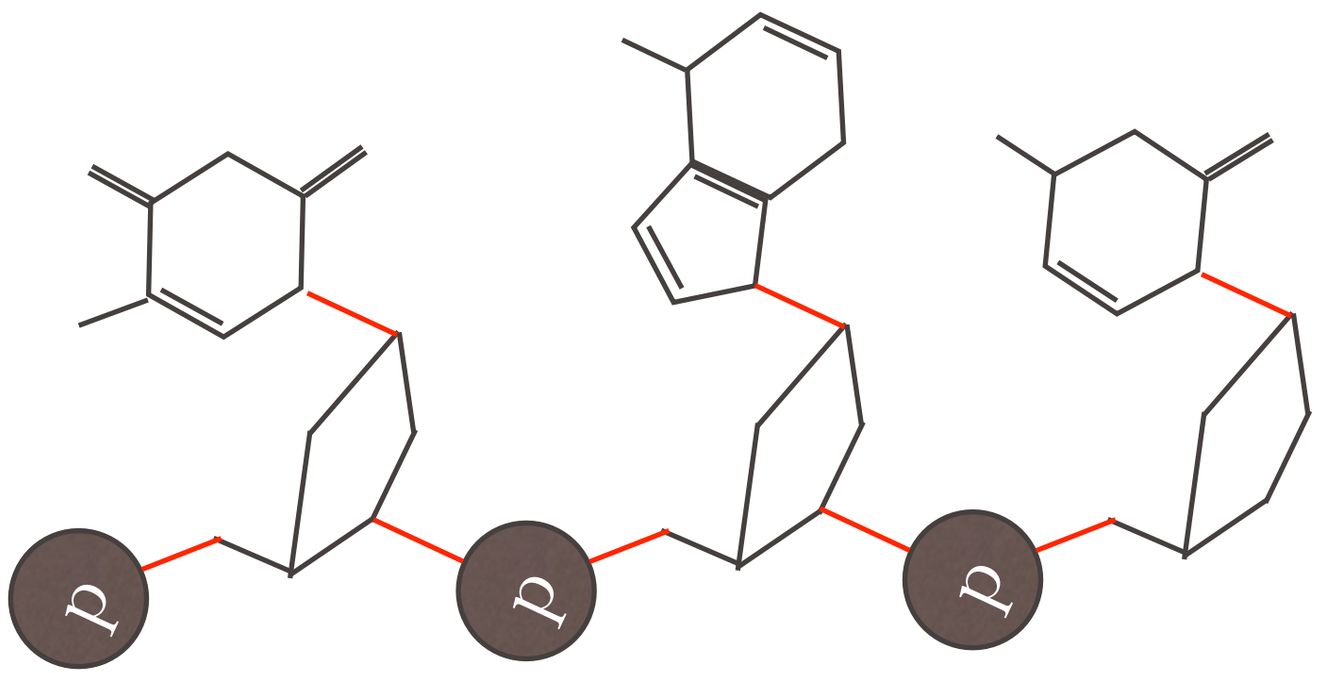
Single Stranded DNA



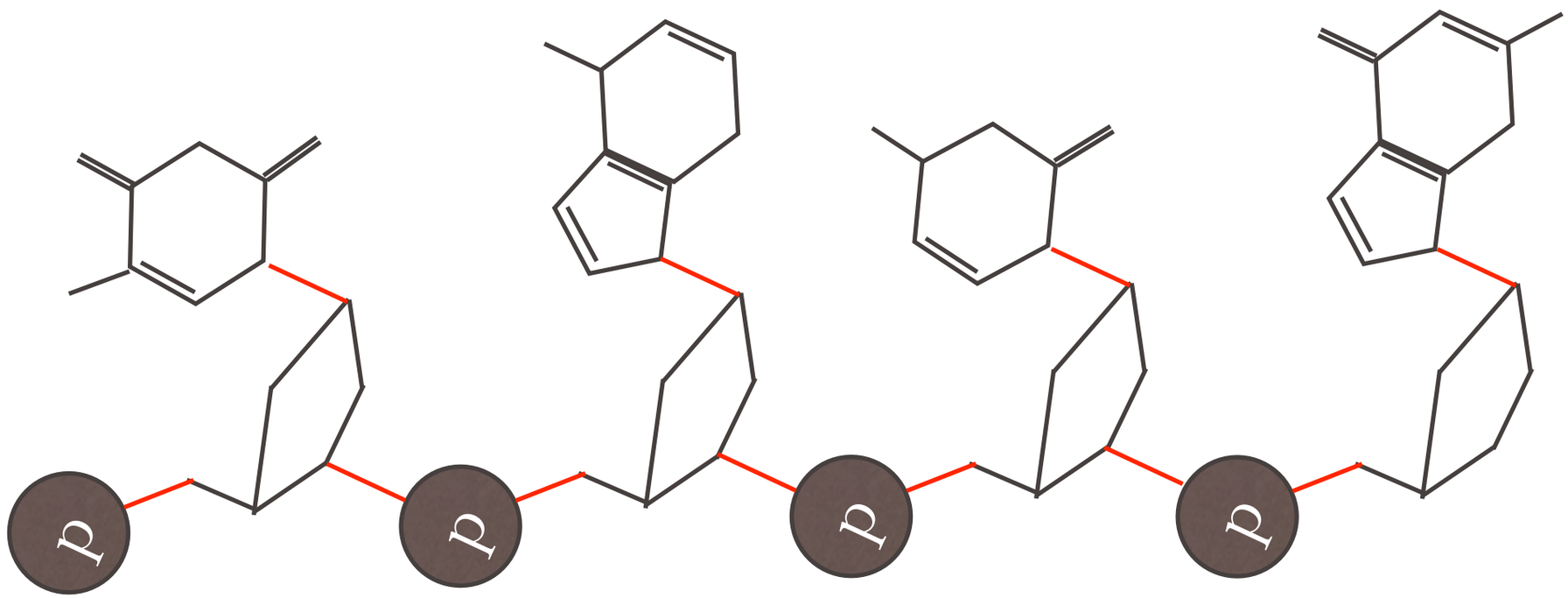
Single Stranded DNA



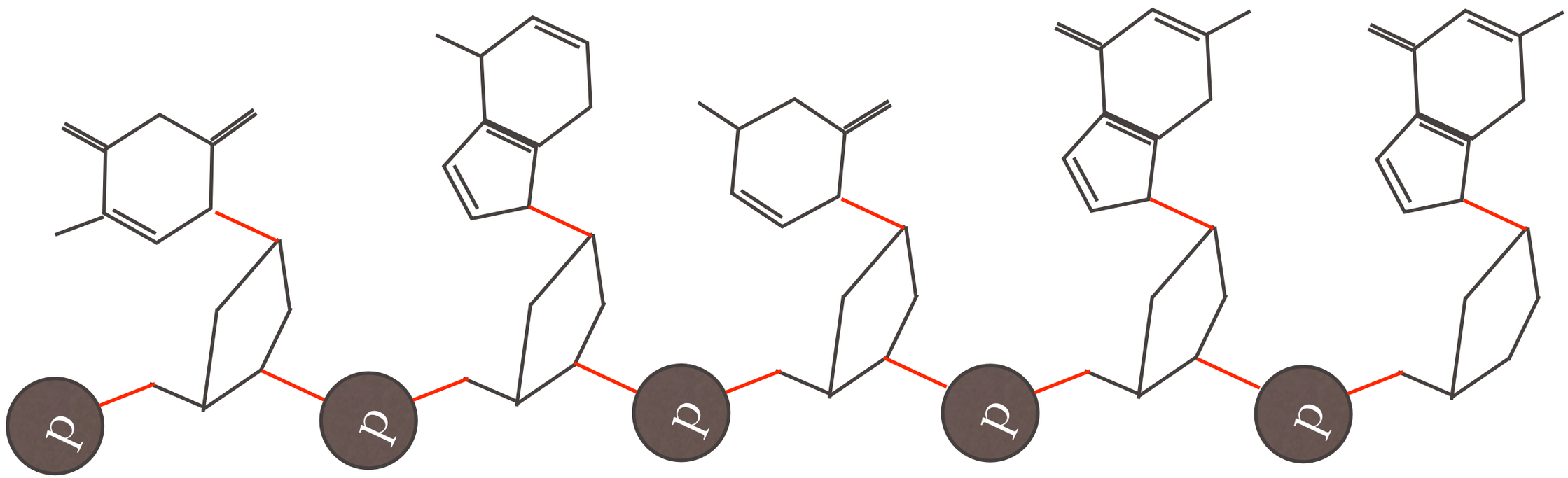
Single Stranded DNA



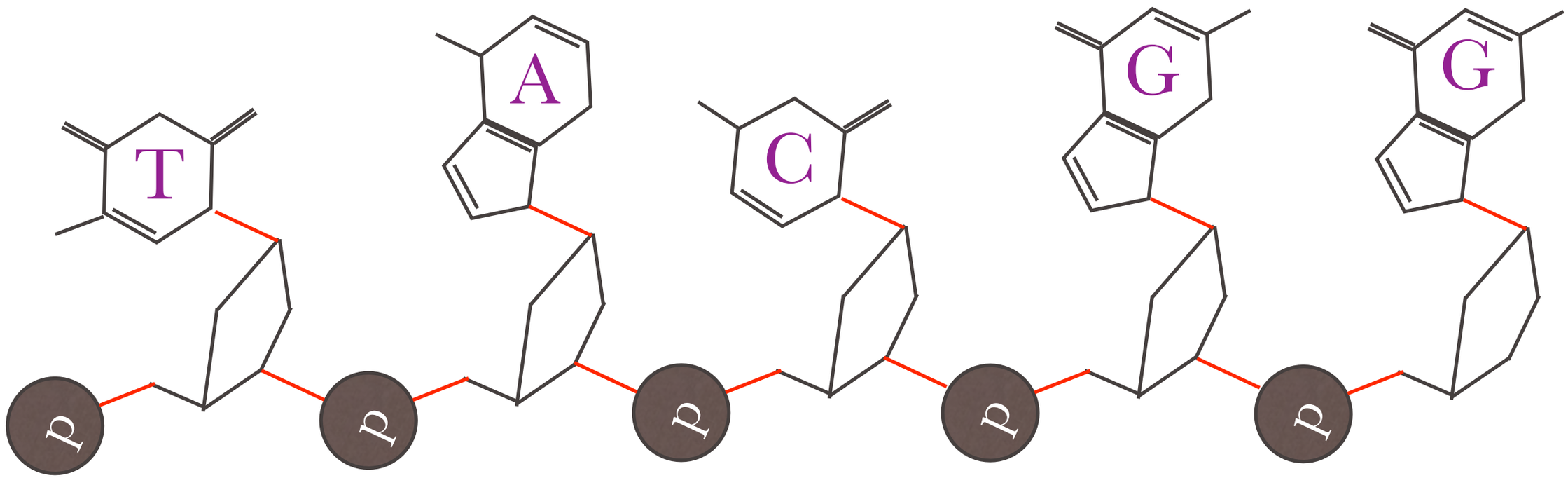
Single Stranded DNA



Single Stranded DNA



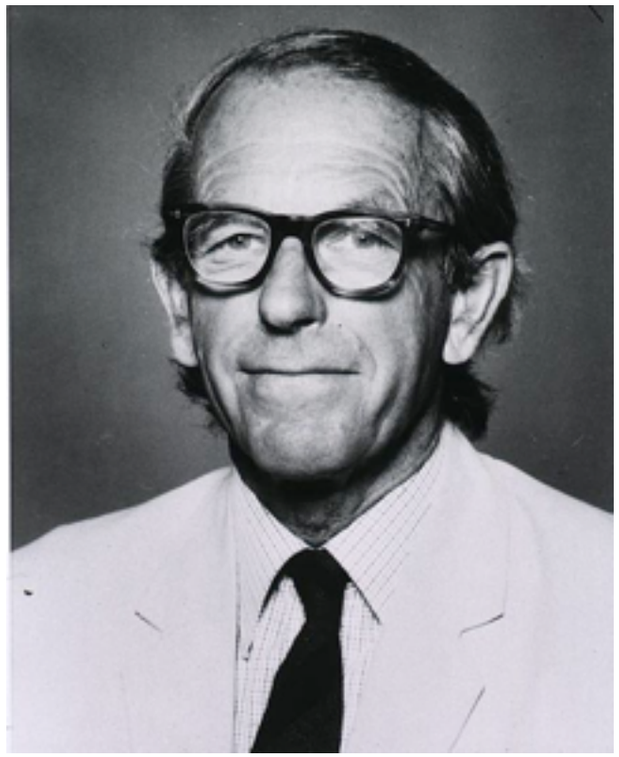
Single Stranded DNA



TACGG.....

Fred Sanger

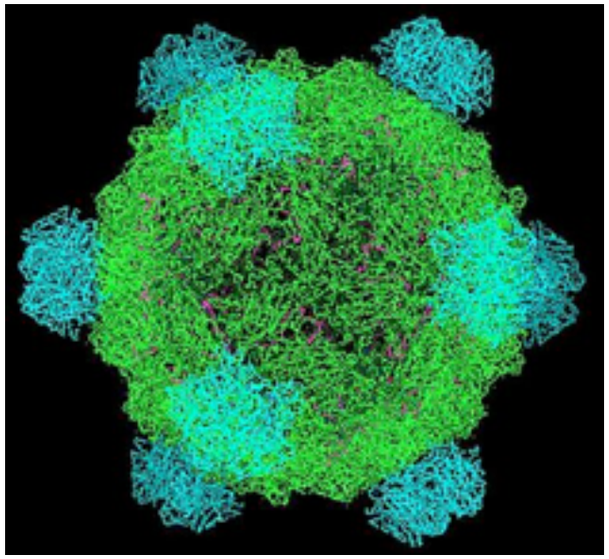
- Won the Nobel Prize for Chemistry twice
  Sequencing Insulin
  DNA Sequencing



1918-2013

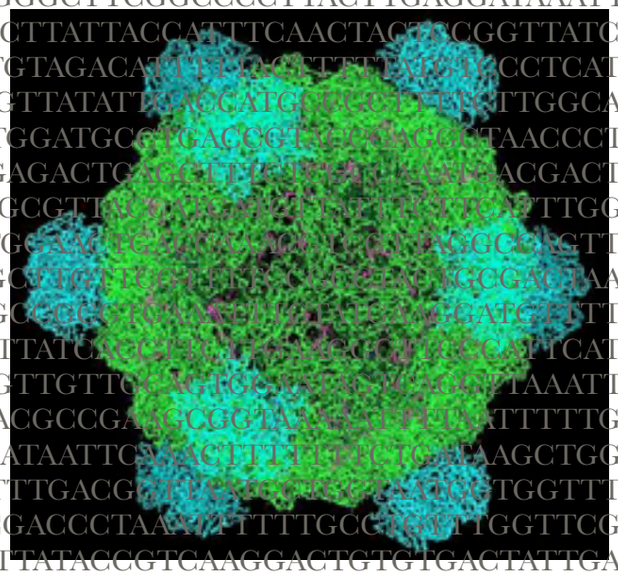
Phi X 174 Bactriophage

- First Genome to be Sequenced
- Genome Size: 5386 bp
- Fred Sanger (1977)



PhiX174 Bactriophage

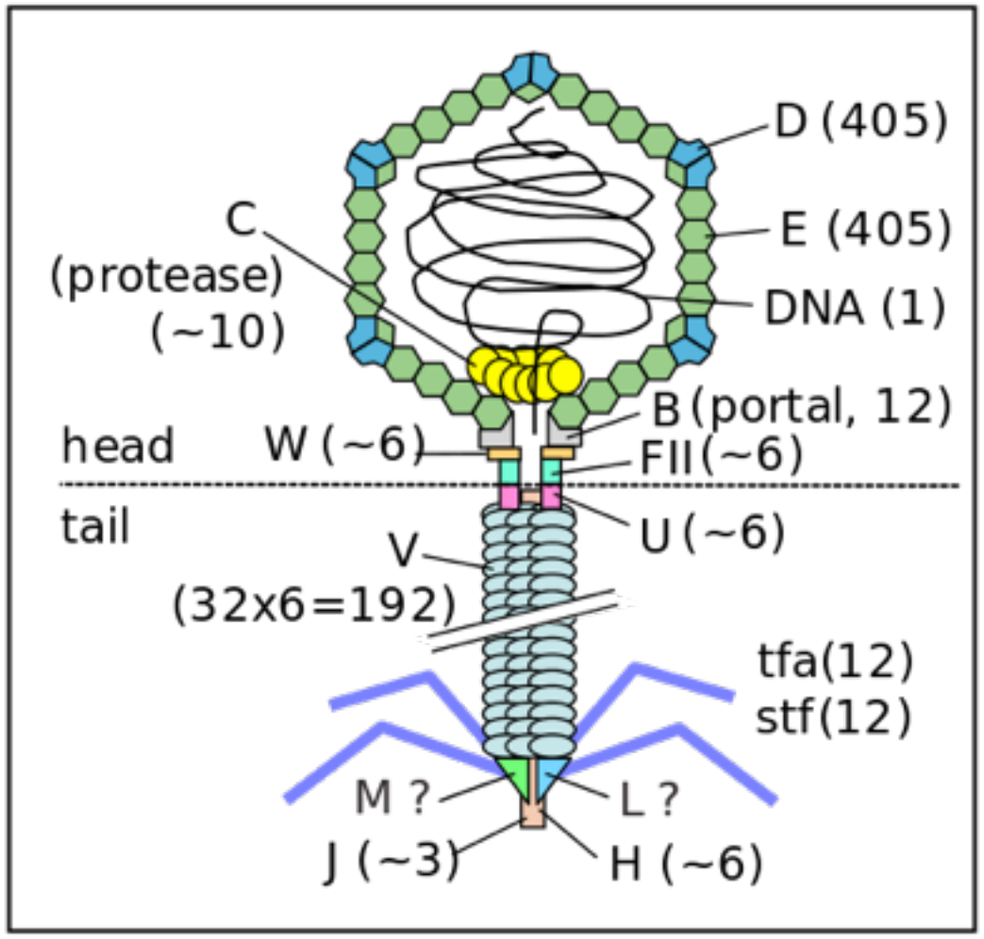
- First Genome to be Sequenced
- Genome Size: 5386 bp
- Fred Sanger (1977)



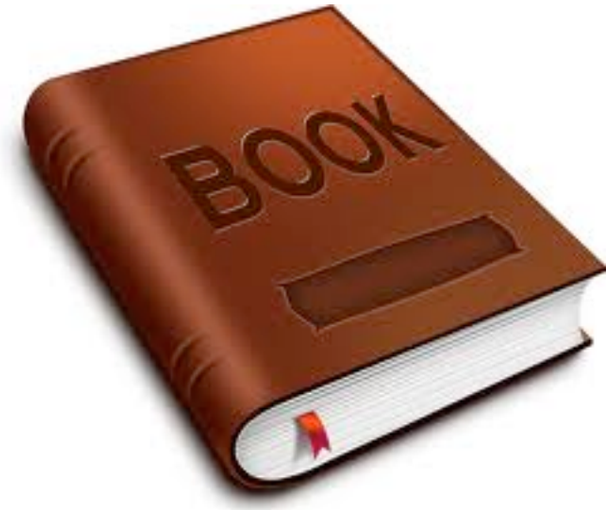
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTTGATAAAGCAGGAATTACTACTGCTTGTTTACGAATTAAATCGA
AGTGGACTGCTGGCGGAAAATGAGAAAATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGATTCTGTCAAAAACTGACGC
GTTGGATGAGGAGAAGTGGCTTAATATGCTTGTCACGTTCGTCAAGGACTGGTTTGATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATTTTAAAAGAG
CGTGGATTACTATCTGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCGTACGTTTCCAGACCGCTTTGGCCTCTATTAAGCT
CATTCAGGCTTCTGCCGTTTTGGATTTACCCAAGATGATTTCCATTTTCTGACGTCAAGAAAGTTTGATTGGTATCTGACGCTCTCCTGCTCGTCGCTGCGTTGAGG
CTTGCGTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCCTGCTCCTGTTGAGTTTATTGCTGCCGTCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGG
CCTGTCTCATCATGGAAGGCGCTGAATTTACGGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCGTGTACGCGCAGGA
AACACTGACGTTCTTACTGACGCAGAAGAAAACGTGCGTCAAAAATTACGTGCGGAAGGAGTGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTC
CGCAGCCGTTGCGAGGTACTAAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTTCGGCCCCTTACTTGAGGATAAATTA
TGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCATCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTGCCGGTTATCG
CTGGCGACTCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTACTGTAGACATTTTCTCTTTTTATGTCCTCATC
GTCACCTTTATGCTGAACAGTGGATTAAGTTCATGAAGGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGCTTATATTCACGATGTCGCTTTTCTTGGCAC
GATTAAGCCTGATACCGATAAATCCTATGGATTCTTTCACGCTATTCAATCTTATTAACACTATTTTAAAGCGCCGTGGATGCCTGACCGTCTGCAGCGTTAACCCTA
ATGAGCTTAATCAAGATGATGCTCGTTATGGTTTCCGTTGCTGCCATCTCAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGAGCTTTGTTGTAGTTGACGACTT
CTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCTAATTTGCATACTGACCAAGAACGTGATTACTTCATGCAGCGTTTCTGTGTGTGTTTTGTTTTTGGA
GGTAAACTCTTAAGCCGTGACAACGTTCGTTACTTGGAATGCCTCTAATCTCTGGGCATCTGGCTATGATGTTGATGGAAGTTGATGACCATTCGTTACGGCCGGTTT
TCTGGTCGTGTTCAACAGACCTATAAACATTCTGTGCCGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTTCGTTTTTCCGTTTTCCGAAAA
GAGATTCAGTACGTTAAGGCTAAAGGTGCTTTGACTTATACGGATATTGCTGGCGACCCTGTTTTGTATGGCAACTTGCCGCGCGTGATTTTTTGTTTTTGAAGATGTTTTC
CGTTCTGATTGATTCGCTAACAACTTGAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCAGCCTGTTTCGCGTTTCATTCATT
CAGGAACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCGCCACCATGATTATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGGCGATGTGCGTTAAATTT
AATGTGACCGTTTATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTATAACGCCGATGCCGGTATGATTTTTGTTTTTGC
CGCTGAGGGGTTGACCAAGCGAAGCGCGGTAGGTTTTCTGCTTAGGAGTTTAATCATGTTTCAGACTTTTATTTCTCGCCATAATTCAAACTTTTTTTCTGCTTAAGCTGGT
TCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTATATTTTGATAGTTTGACGCTTTGAGCGCTGGATCGGTGGTTTT
CTTCATTGCATTCAGATGGATACATCTGTCAACGCCGCTAATCAGGTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATCTTTTGCCTGCTTTGGTTCGC
TTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTGAATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGAC
GTCCTTCCCCGTACGCCGGGCAATAACGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGTTTCGCTGAATCAGGTTATTAAAGAG
ATTATTTGTCTCCAGCCACTTAAGTGAGGTGATTTATGTTTGGTGCTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTTTGGAGGCGGTC
AAAAAGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCC
TGATGAGGCCGCCCCTAGTTTTGTTTCTGGTGCTATGGCTAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGATAAGTTGCTTG
ATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTATCTTGCTGCTGCATTTCCTGAGCTTAATGCTTGGGAGCGTGCTGGTGCTGAT
GCTTCCTCTGCTGGTATGGTTGACGCCGGATTTGAGAATCAAAAAGAGCTTACTAAAATGCAACTGGACAATCAGAAAGAGATTGCCGAGATGCAAAATGAGACTCAAAAAG
AGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGACCAGGTATATGCACAAAATGAGATGCTTGCTTATCAACAGAAGGAGTCTACTGCTCGCGTTGCGT
CTATTATGGAAAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCAAGCTCAAACGGCTGGTCAGTATTTTACCAATGACCAAATCAA
AGAAATGACTCGCAAGGTTAGTGCTGAGGTTGACTTAGTTCATCAGCAAACGCAGAATCAGCGGTATGGCTCTTCTCATATTGGCGCTACTGCAAAGGATATTTCTAATGTC
GTCACTGATGCTGCTTCTGGTGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGATACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTA
ATTTGTCTAGGAAATAACCGTCAGGATTGACACCCTCCCAATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTTCTTATTACCCTTCTGAATGTCA
CGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTATTGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTTCCATAAGCAGATGGAT
AACCGCATCAAGCTCTTGGAAGAGATTCTGTCTTTTCGTATGCAGGGCGTTGAGTTCGATAATGGTGATATGTATGTTGACGGCCATAAGGCTGCTTCTGACGTTCGTGAT
GAGTTTGTATCTGTTACTGAGAAGTTAATGGATGAATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGGGACGAAA
AATGGTTTTTAGAGAACGAGAAGACGGTTACGCAGTTTTGCCGCAAGCTGGCTGCTGAACGCCCTCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTA
TTAAGGATGAGTGTTCAAGATTGCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTTGCTATTCAGCGTTTGATGAATGCAATGCGACAGGCTCATGCTGATGGTTGGT
TTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTTATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGC
CGAGGGTCGCAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTGAGTATGGTACAGCTAATGGCCGTCTTCATTTCCATGCGGTGCACTTTATGCG
GACACTTCCTACAGGTAGCGTTGACCCTAATTTTGGTCGTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCATCGC
AGTTCGCTACACGCAGGACGCTTTTTCACGTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAGCCGCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGG
CTAAATACGTTAACAAAAAGTCAGATATGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACTTCCCAAGAAGCTGTT
CAGAATCAGAATGAGCCGCAACTTCGGGATGAAAATGCTCACAATGACAAATCTGTCCACGGAGTGCTTAATCCAACTTACCAAGCTGGGTTACGACGCGACGCCGTTCAA
CCAGATATTGAAGCAGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGCGGCGCAACCTGTGACGACAAATCTGCTCAAATTTAT
GCGCGCTTCGATAAAAATGATTGGCGTATCCAACCTGCA

Lambda Phage

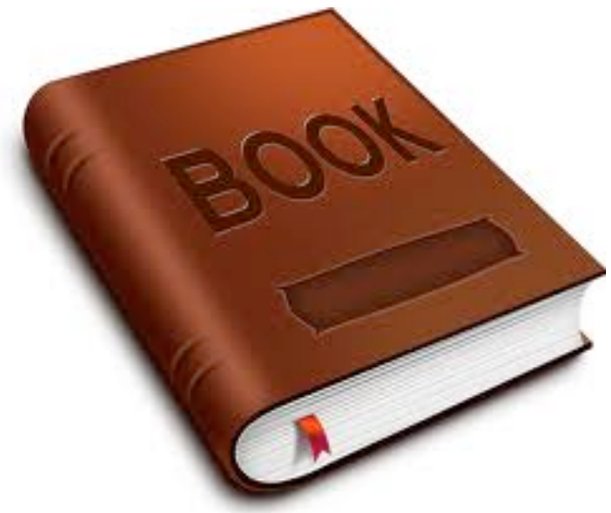
- Genome Size: 48,502 bp
- Fred Sanger, et al (1982)
- Method: Shotgun Sequencing



Shotgun Sequencing

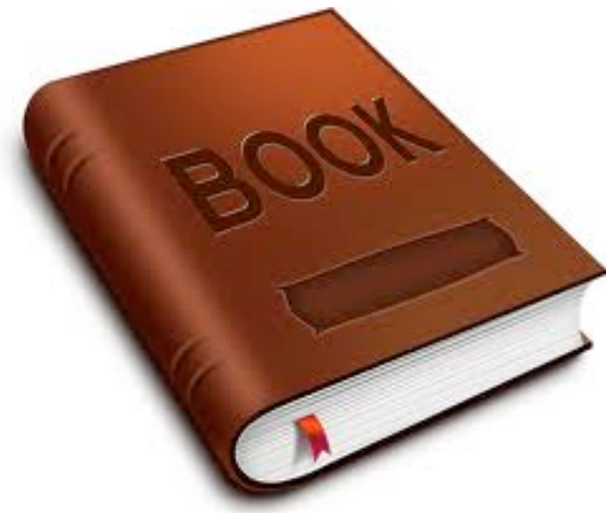


Shotgun Sequencing



Reads

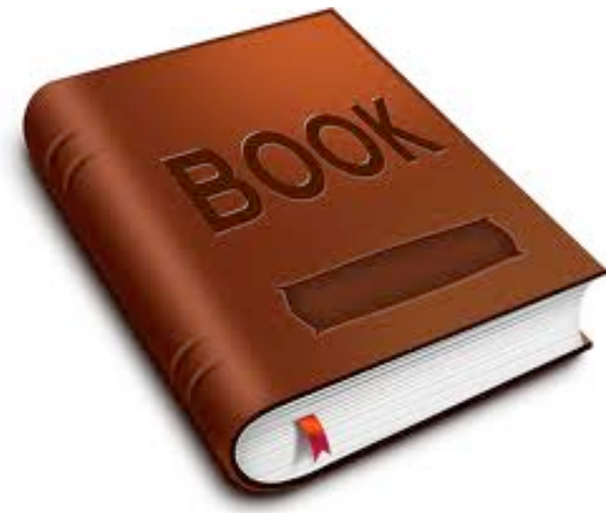
Shotgun Sequencing



Reads

Hearken to the reed-flute

Shotgun Sequencing



Reads

banishment from its home

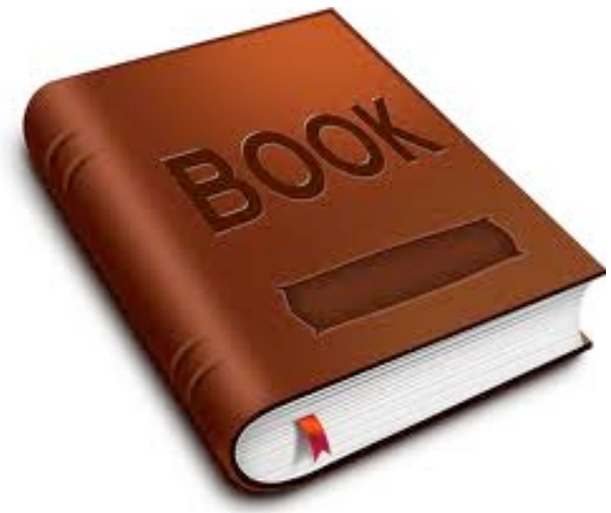
Lamenting its banishment from

it complains, Lamenting its

Hearken to the reed-flute

reed-flute, how it complains,

Shotgun Sequencing



Reads

banishment from its home

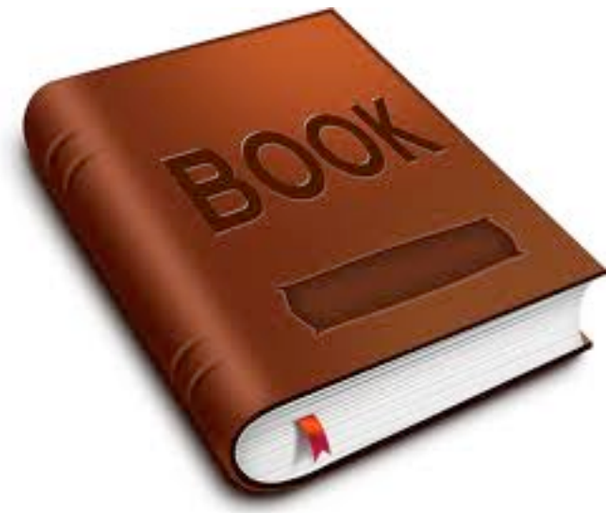
Lamenting its banishment from

it complains, Lamenting its

reed-flute, how it complains,

Hearken to the reed-flute

Shotgun Sequencing



Reads

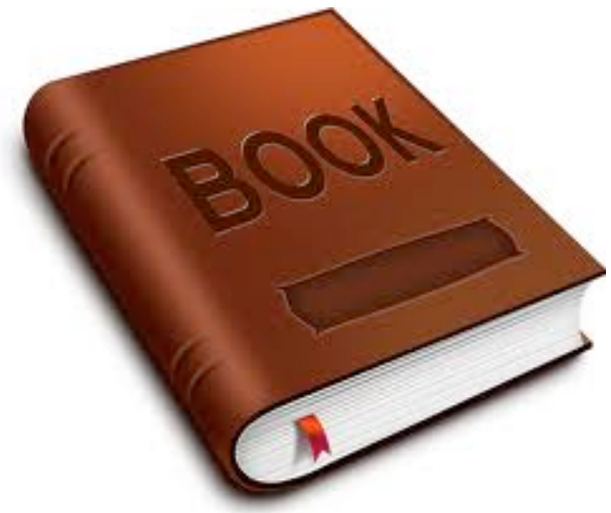
banishment from its home

Lamenting its banishment from

it complains, Lamenting its

Hearken to the reed-flute, how it complains,

Shotgun Sequencing



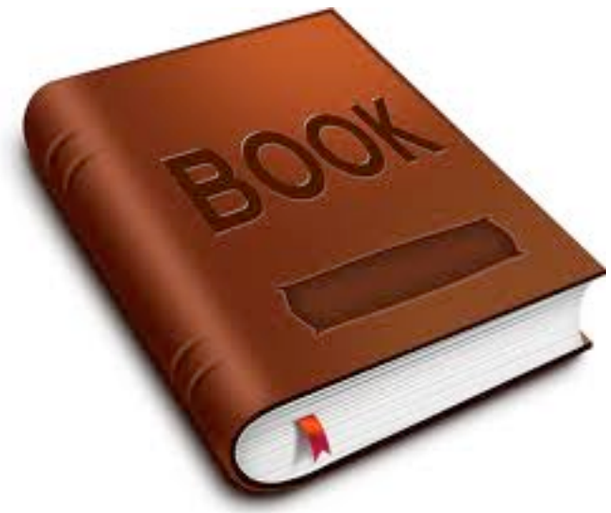
Reads

banishment from its home

Lamenting its banishment from

Hearken to the reed-flute, how it complains, Lamenting its

Shotgun Sequencing

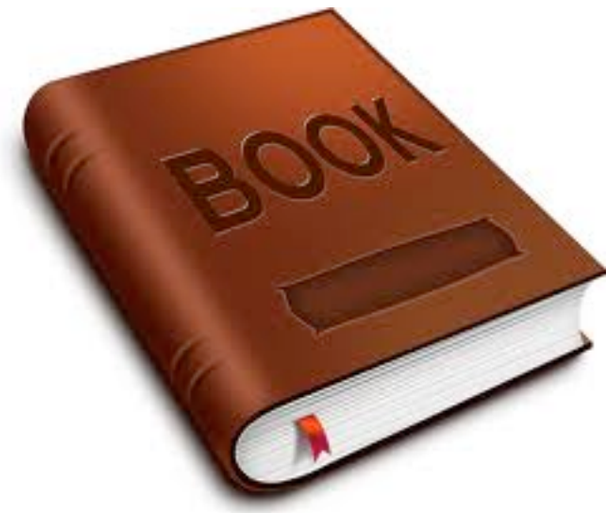


Reads

banishment from its home

Hearken to the reed-flute, how it complains, **Lamenting its** banishment from

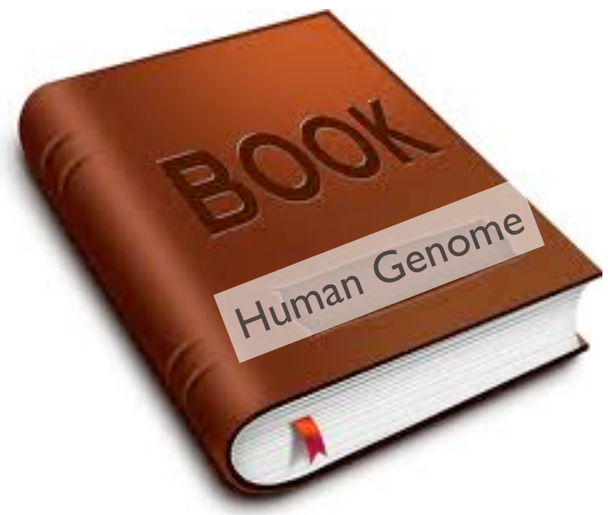
Shotgun Sequencing



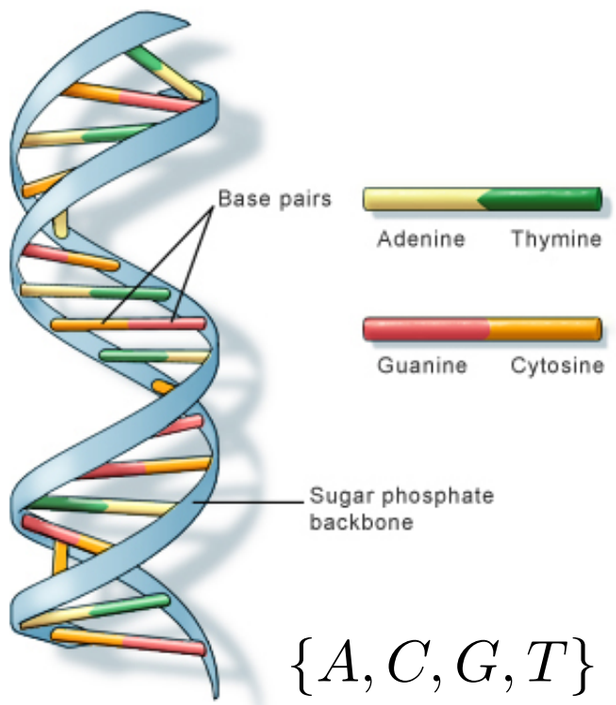
Reads

Hearken to the reed-flute, how it complains, **Lamenting its banishment from** its home

DNA Shotgun Sequencing



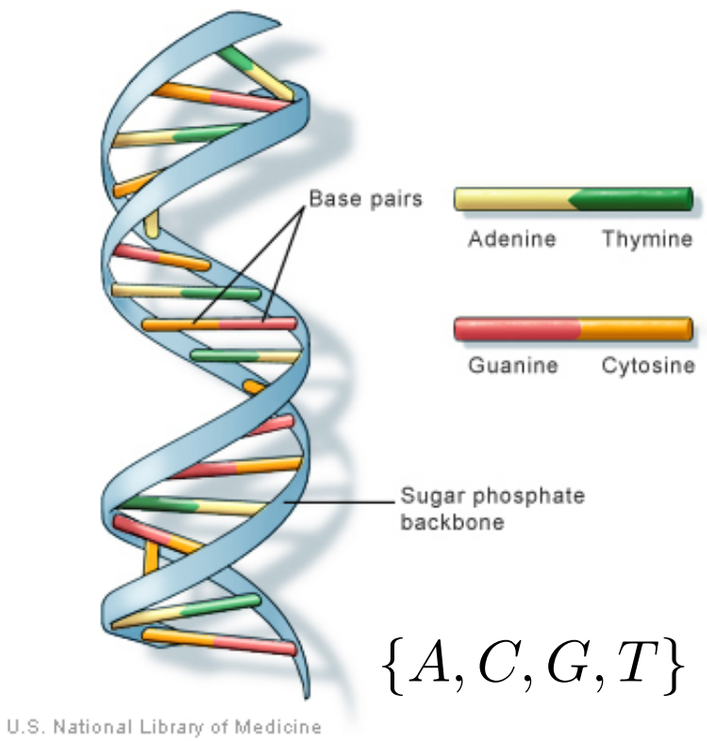
DNA Shotgun Sequencing



U.S. National Library of Medicine

$$\{A, C, G, T\}$$

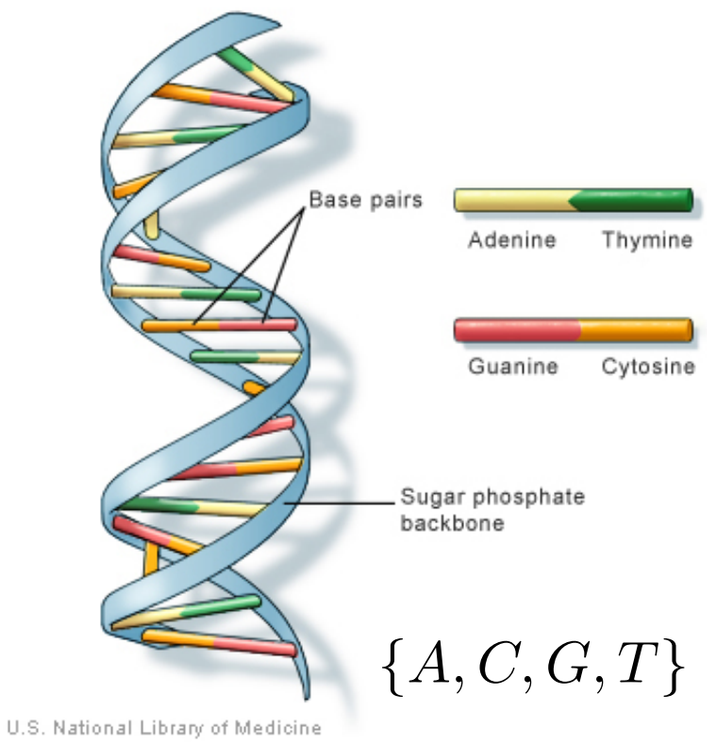
DNA Shotgun Sequencing



$\{A, C, G, T\}$

Reads

DNA Shotgun Sequencing

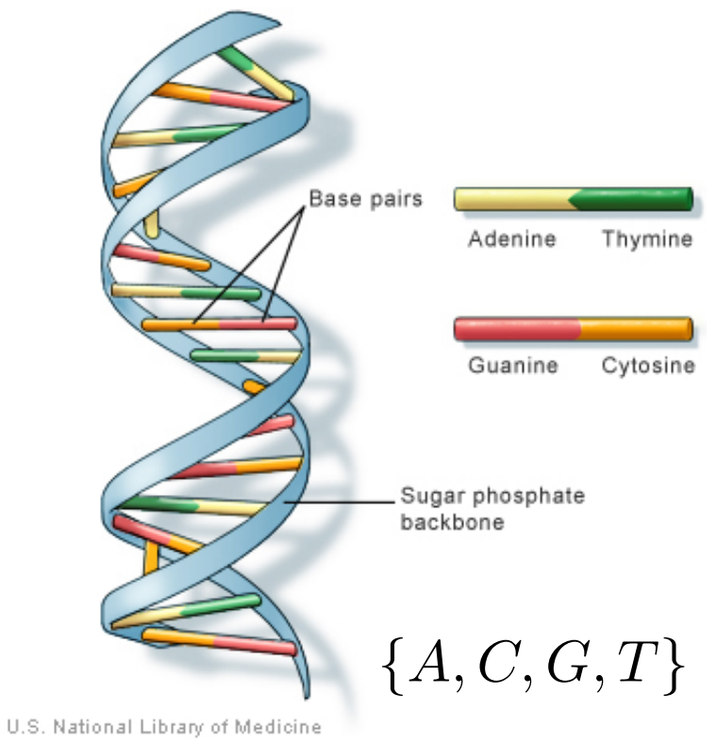


$\{A, C, G, T\}$

π

Reads

DNA Shotgun Sequencing



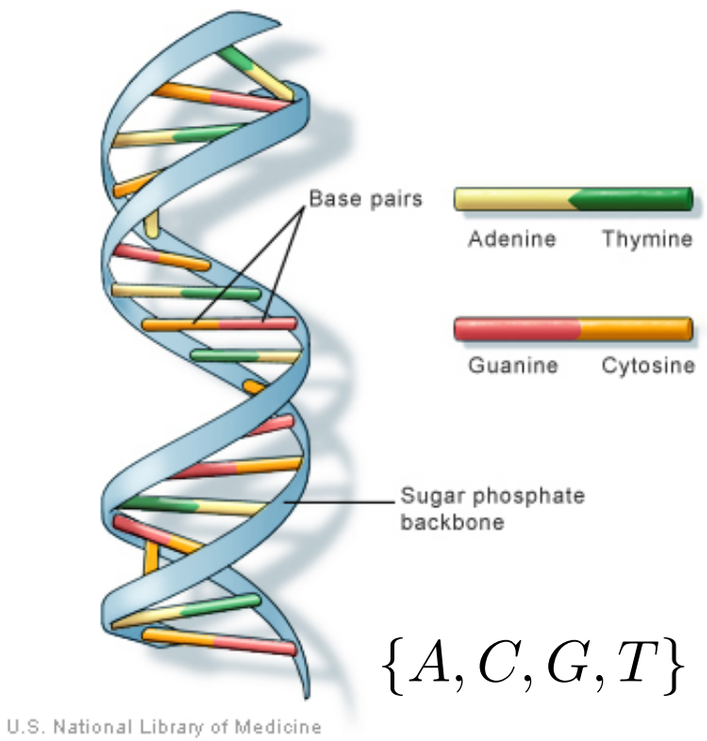
$\{A, C, G, T\}$

ACCGTCCTTAAGG

π

Reads

DNA Shotgun Sequencing



$\{A, C, G, T\}$

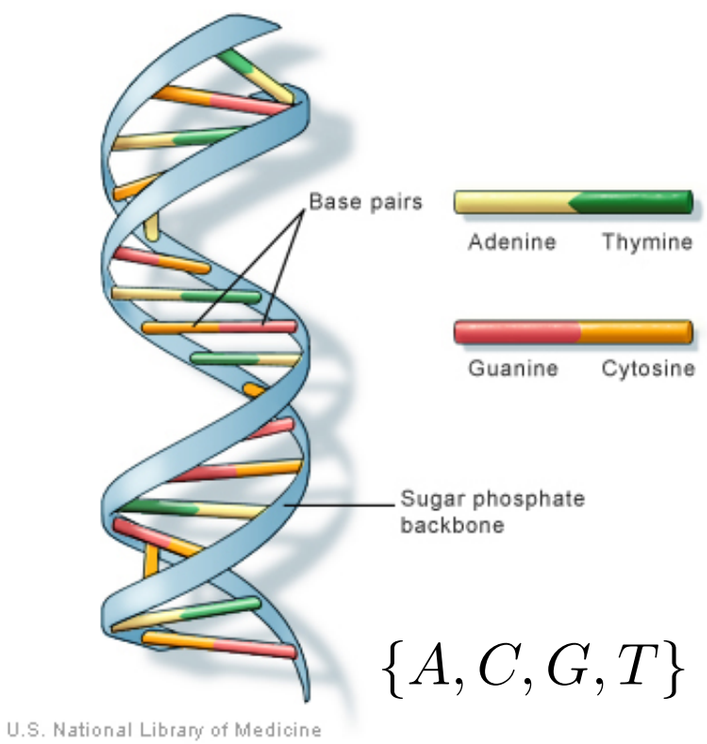
ACCGTCCTTAAGG

π

ACCGTCGTTAAGG

Reads

DNA Shotgun Sequencing



$\{A, C, G, T\}$

ACCGTCCTTAAGG

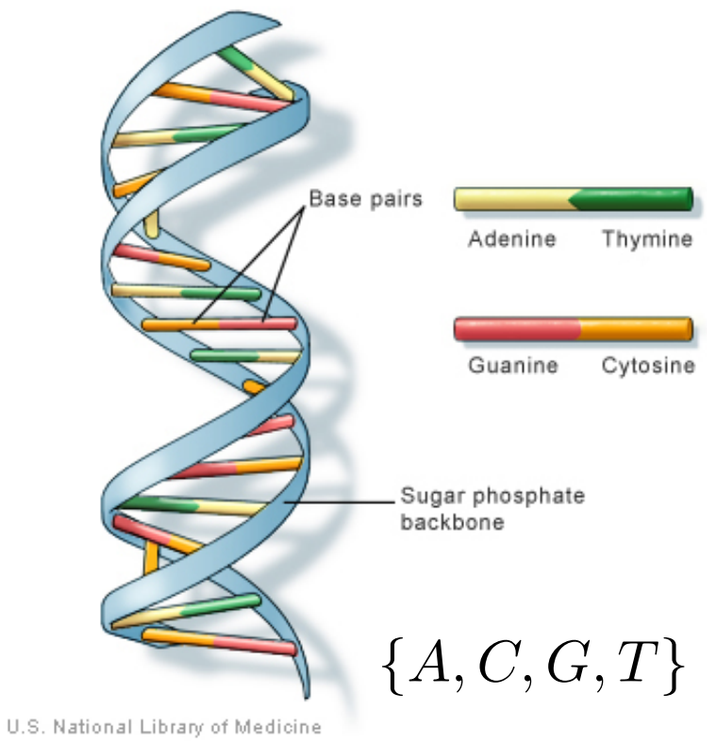
π

ACCGTCGTTAAGG

Reads

ACCGTCGTTAAGG

DNA Shotgun Sequencing



$\{A, C, G, T\}$

π

Reads

ACCGTCGTTAAGG

GTTTCCTAAGTGG

ACTGCTTCGTCCTA

ACTGCTTCGTCCTA

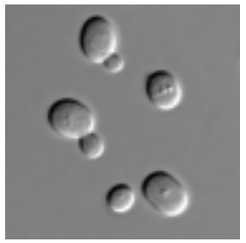
CCTTAAGCGGTTC

Starting Big Projects

Starting Big Projects

1989

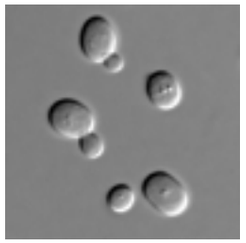
- Genome of s. cerevisiae
- Size of Genome: 12 Mb
- 16 Chromosomes
- Finished 1996



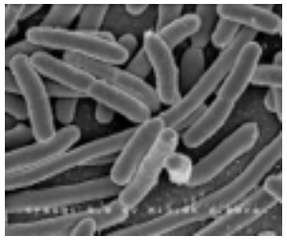
Starting Big Projects

1989

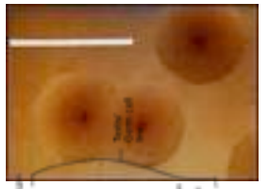
- Genome of s. cerevisiae
- Size of Genome: 12 Mb
- 16 Chromosomes
- Finished 1996

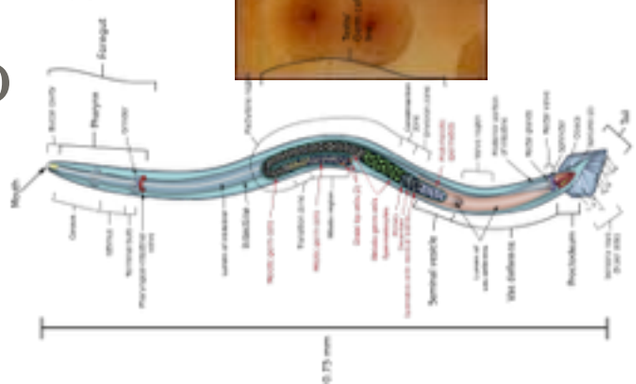


1990



- E. Coli: 4.6 Mb
- M. Capricolum: 1 Mb
- C. elegans: 100 Mb

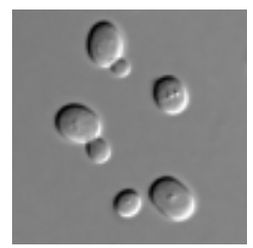




Starting Big Projects

1989

- Genome of s. cerevisiae
- Size of Genome: 12 Mb
- 16 Chromosomes
- Finished 1996

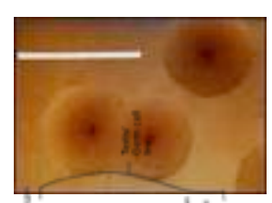


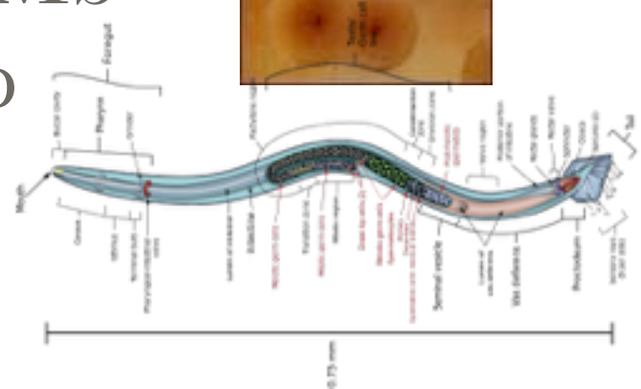
1990

- Human Genome Project (HGP)
- Size of Genome: 3.2 Gb
- 23 Chromosomes
- Finished 2003

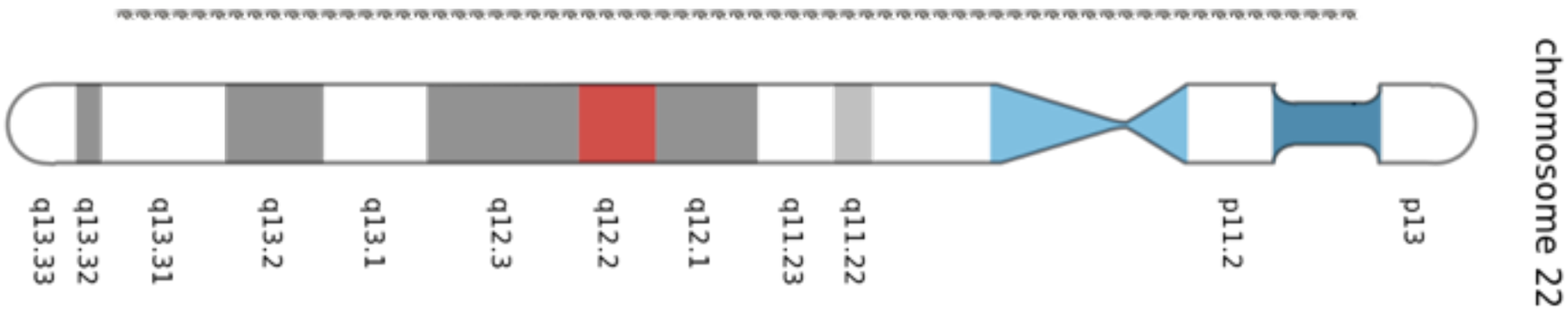
1990

- E. Coli: 4.6 Mb
- M. Capricolum: 1 Mb
- C. elegans: 100 Mb

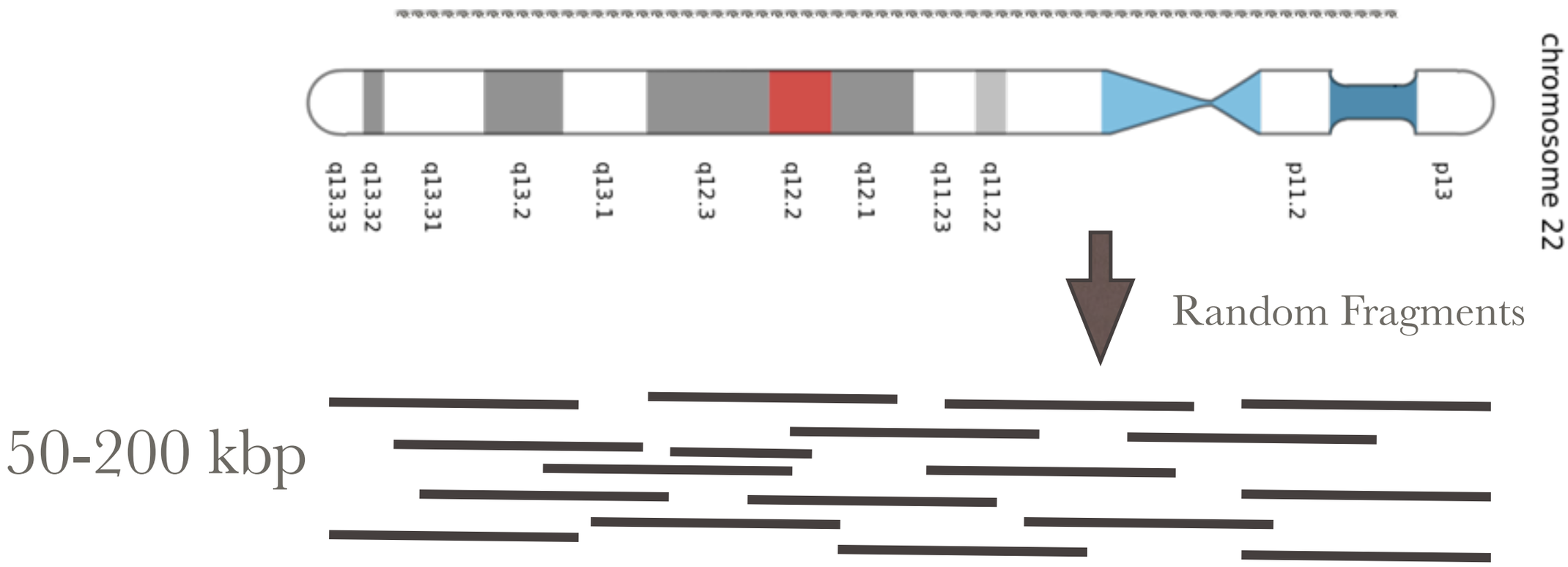




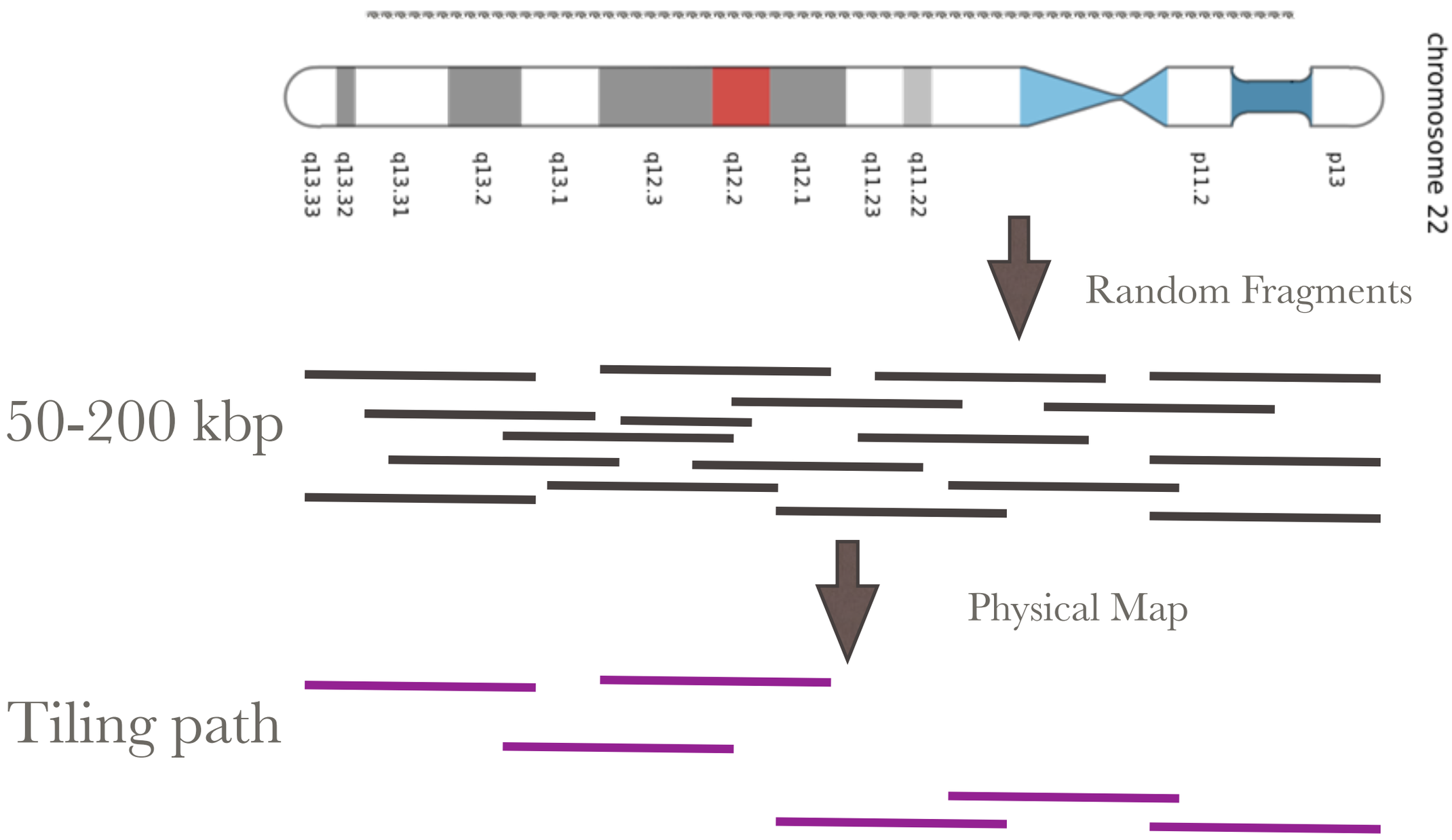
Strategy: Hierarchical SG



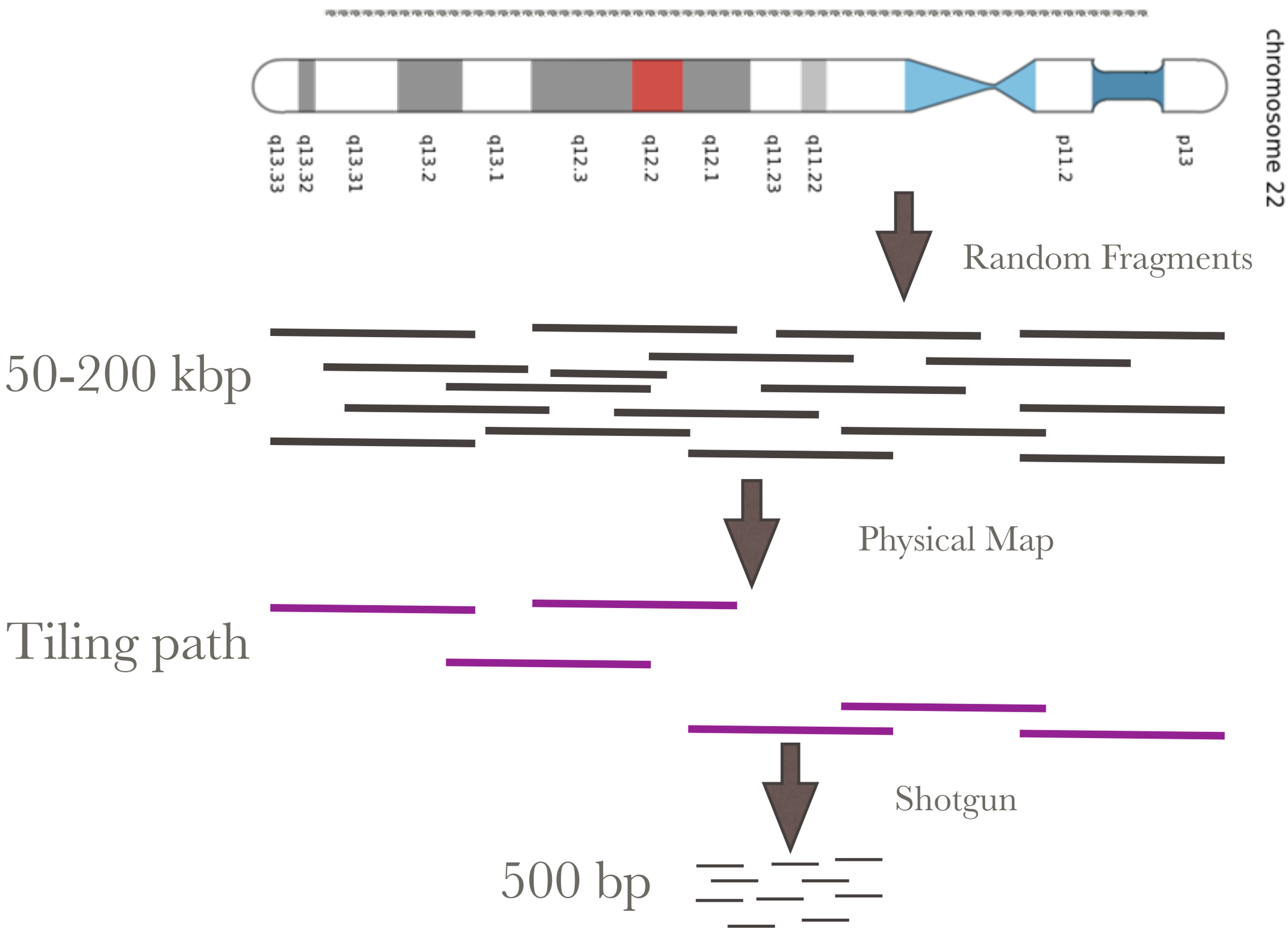
Strategy: Hierarchical SG



Strategy: Hierarchical SG

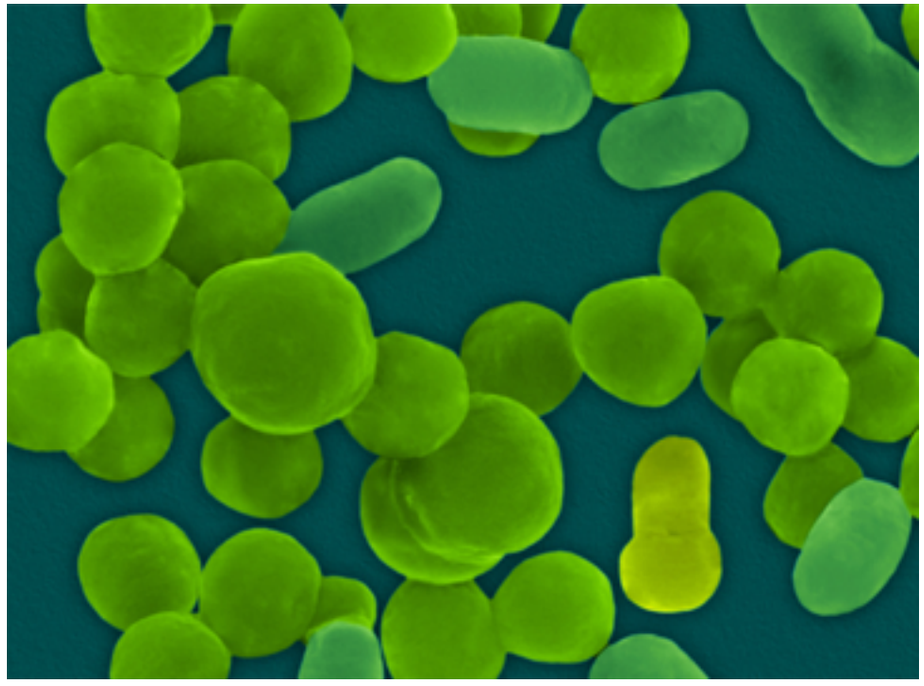


Strategy: Hierarchical SG

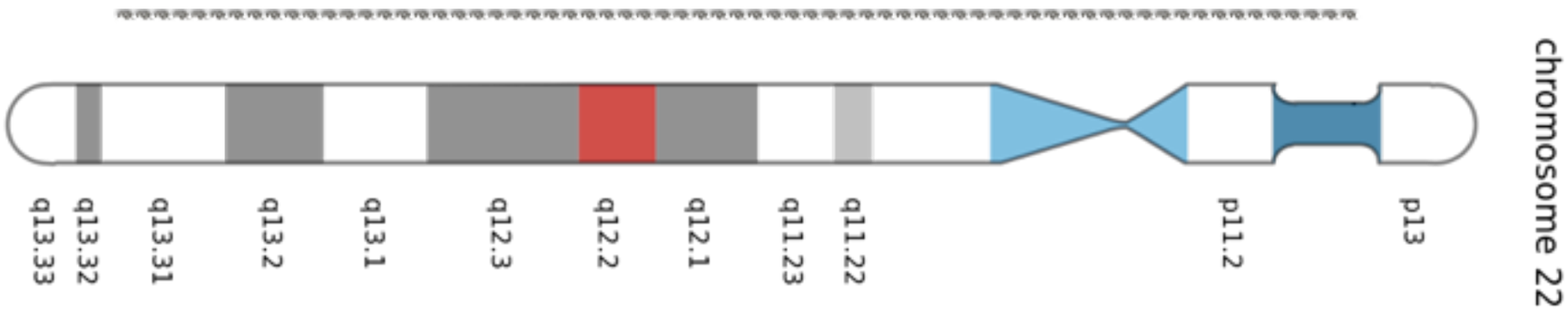


H Influenzae

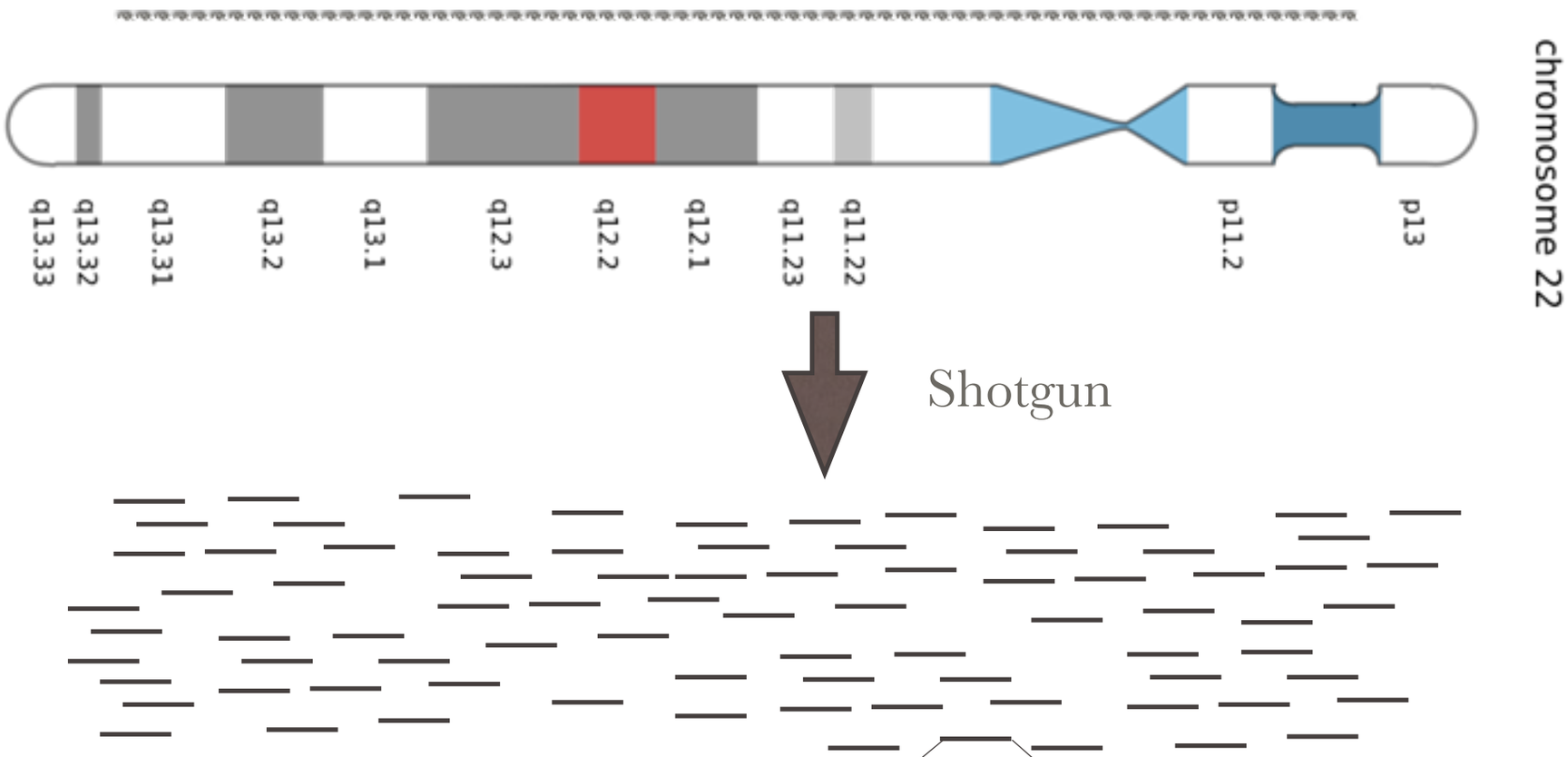
- The first sequenced free-living organism
- Genome Size: 1.8 Mb
- Strategy: Whole Genome Shotgun Sequencing
- The Institute for Genomic Research TIGR
- Finished 1995
- Done in 13 months
- Cost: 50 cents per base



Whole Genome Shotgun Sequencing



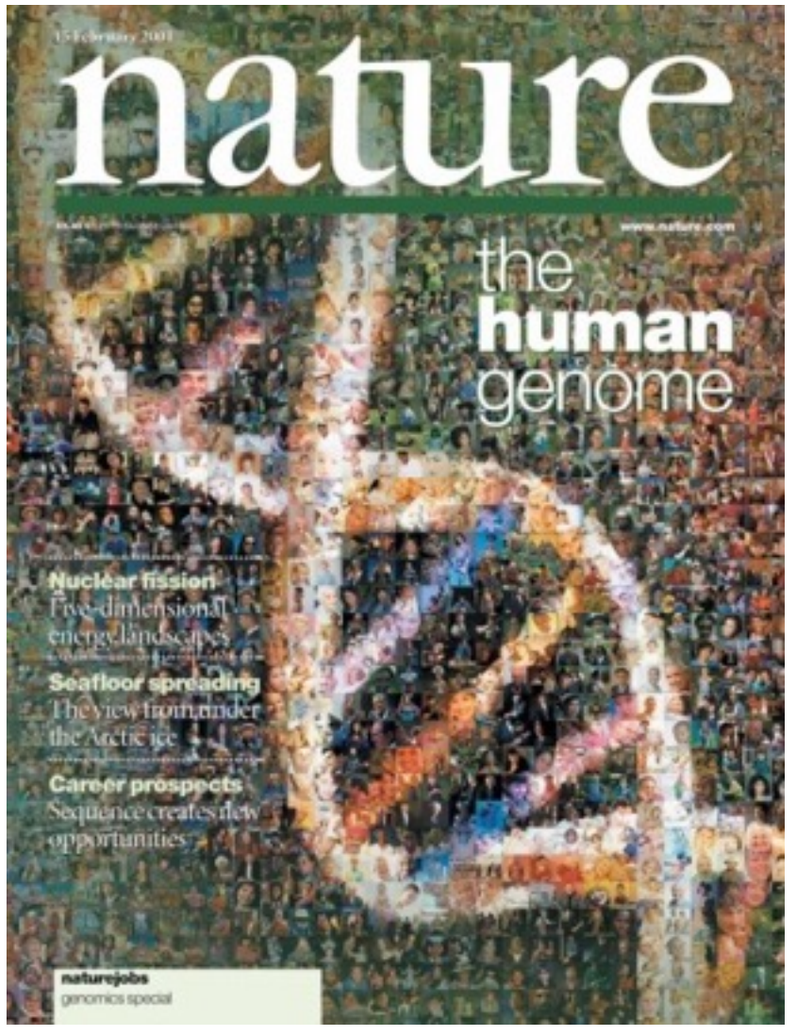
Whole Genome Shotgun Sequencing



Number of Reads: N
Read Length: L

ATCCGGTACGGATCGTACA

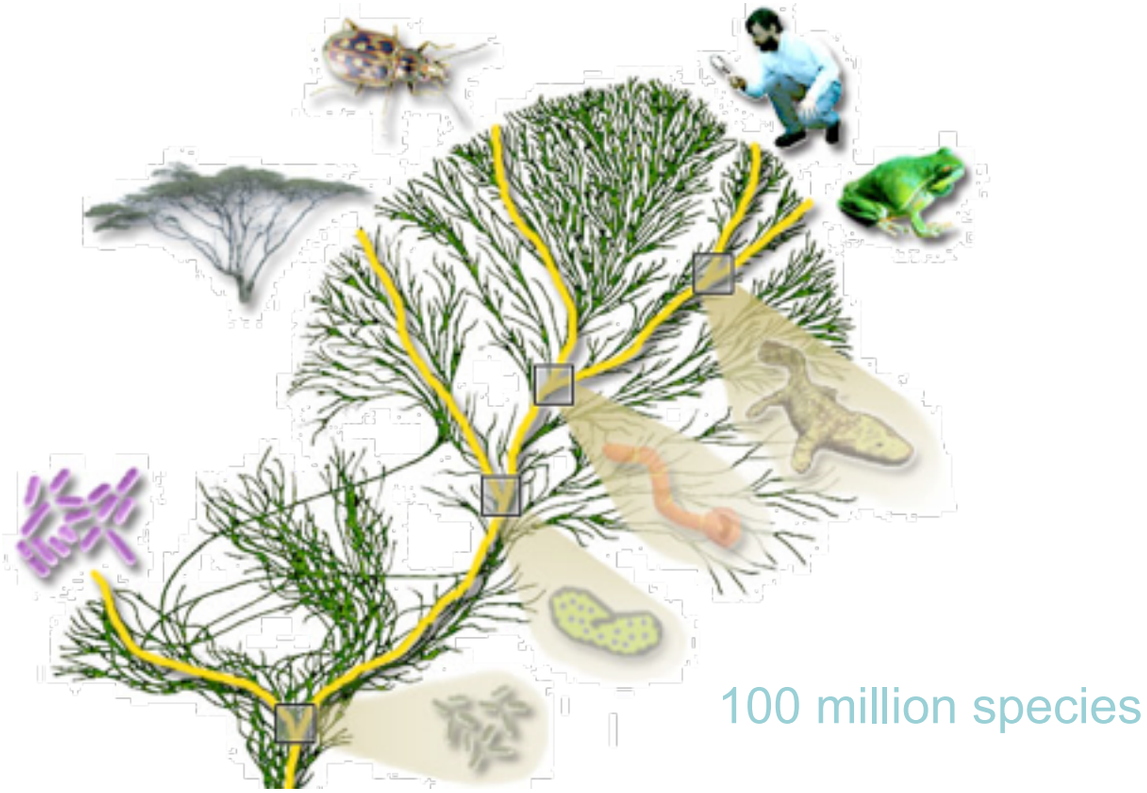
HGP FINISHED



There is Never Enough Sequencing

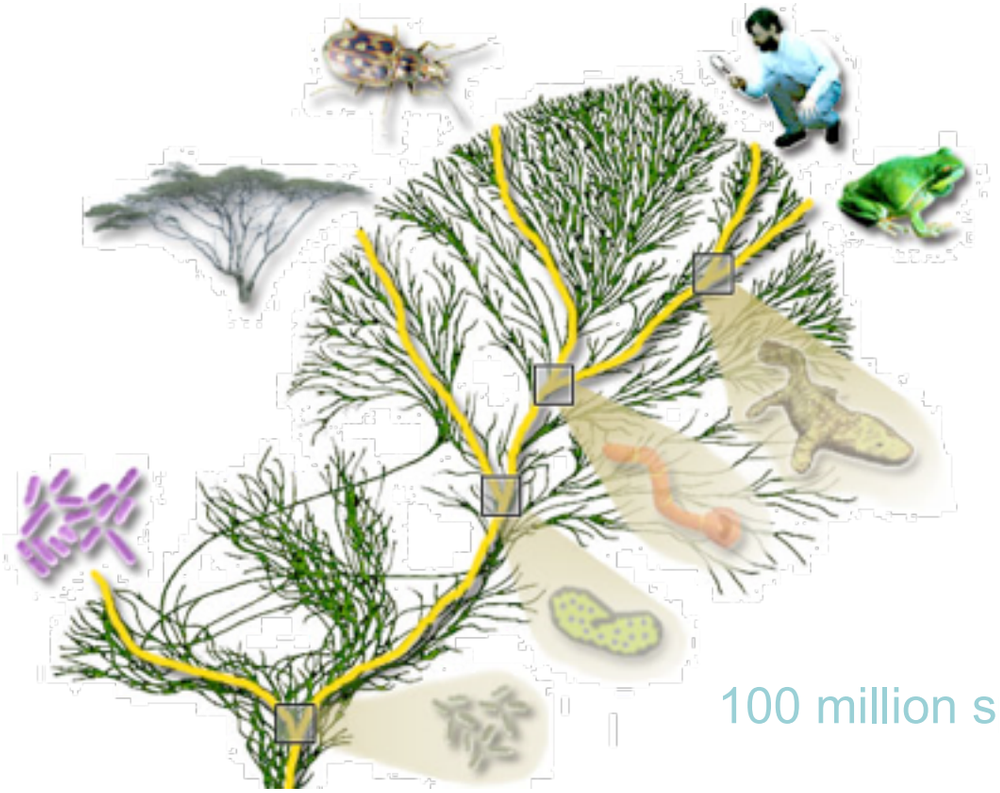
courtesy: Batzoglou

There is Never Enough Sequencing



courtesy: Batzoglou

There is Never Enough Sequencing

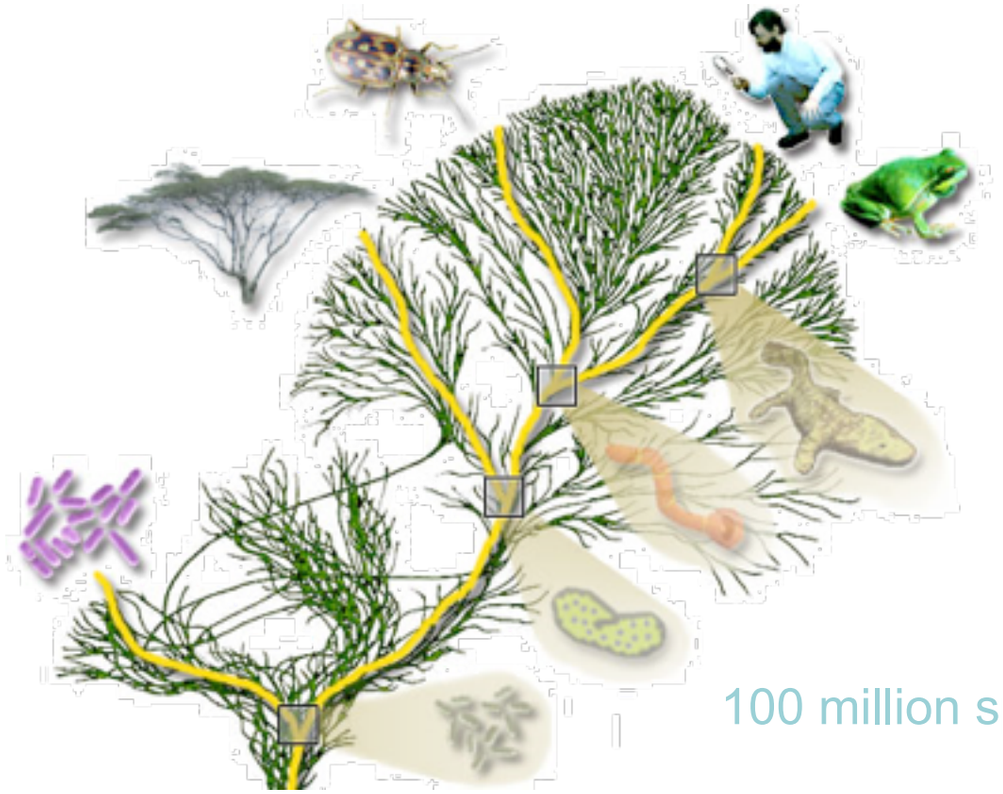


100 million species

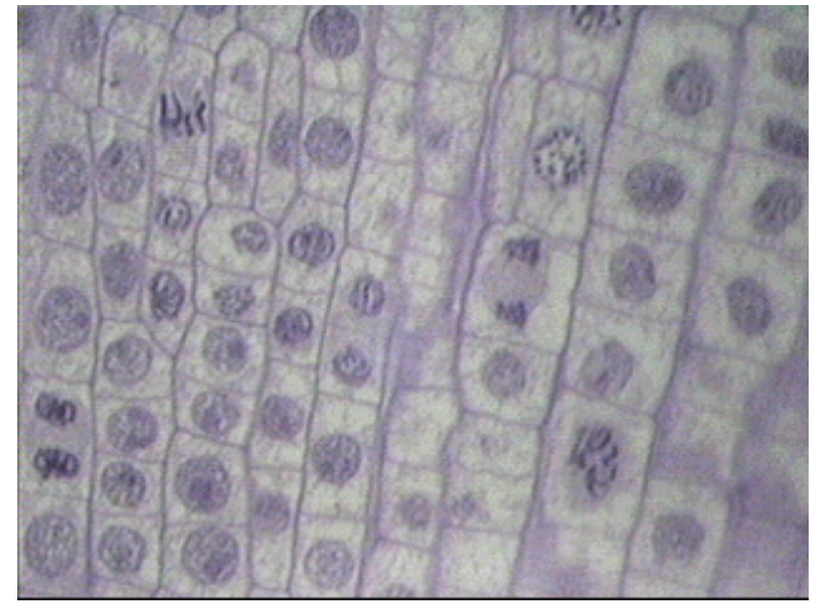
7 billion individuals

courtesy: Batzoglou

There is Never Enough Sequencing



100 million species



Somatic mutations
(e.g., HIV, cancer)

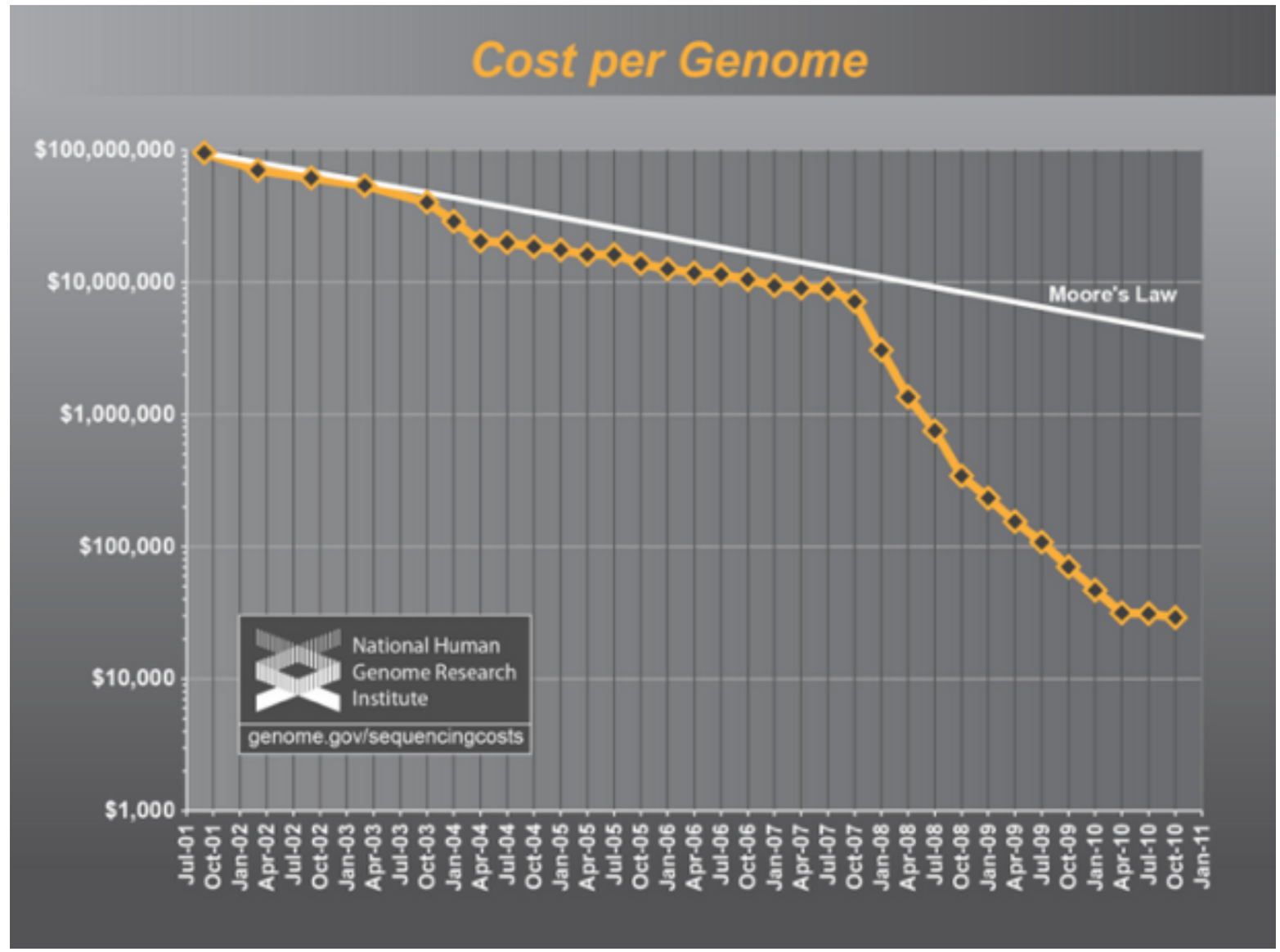
7 billion individuals

courtesy: Batzoglou

Sequencing Growth

<u>Cost of one human genome</u>

- 2004: $30,000,000
- 2008: $100,000
- 2010: $10,000
- **2013: (today) $4,000**
- ???: $300



courtesy: Batzoglou

Next Generation Sequencing

- Very High Throughput
- Higher error rate
- Shorter Length
- Very Cheap

Computational Challenges

courtesy: Batzoglou

Computational Challenges

- How to assemble the reads?

courtesy: Batzoglou

Computational Challenges

- How to assemble the reads?
- For given genome size G, what are the requirements

courtesy: Batzoglou

Computational Challenges

- How to assemble the reads?
- For given genome size *G*, what are the requirements
  - Number of Reads: N

courtesy: Batzoglou

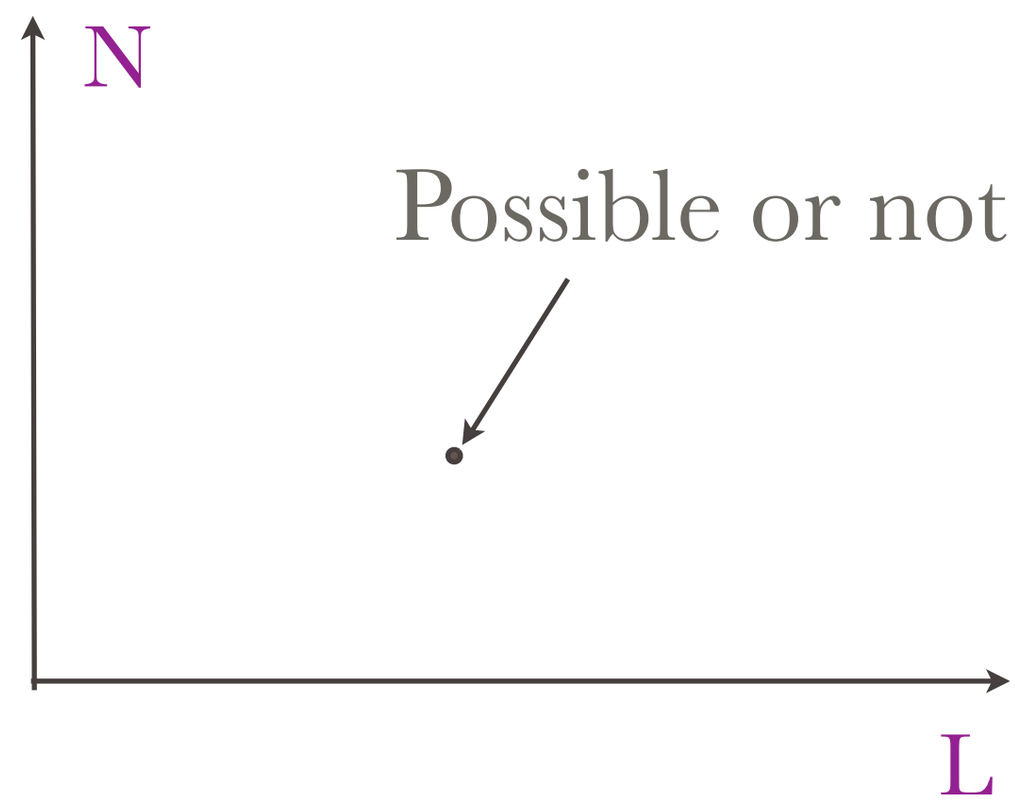
Computational Challenges

- How to assemble the reads?
- For given genome size *G*, what are the requirements
  - Number of Reads: N
  - Length of Reads: L

courtesy: Batzoglou

Computational Challenges

- How to assemble the reads?
- For given genome size *G*, what are the requirements
  - Number of Reads: N
  - Length of Reads: L



courtesy: Batzoglou

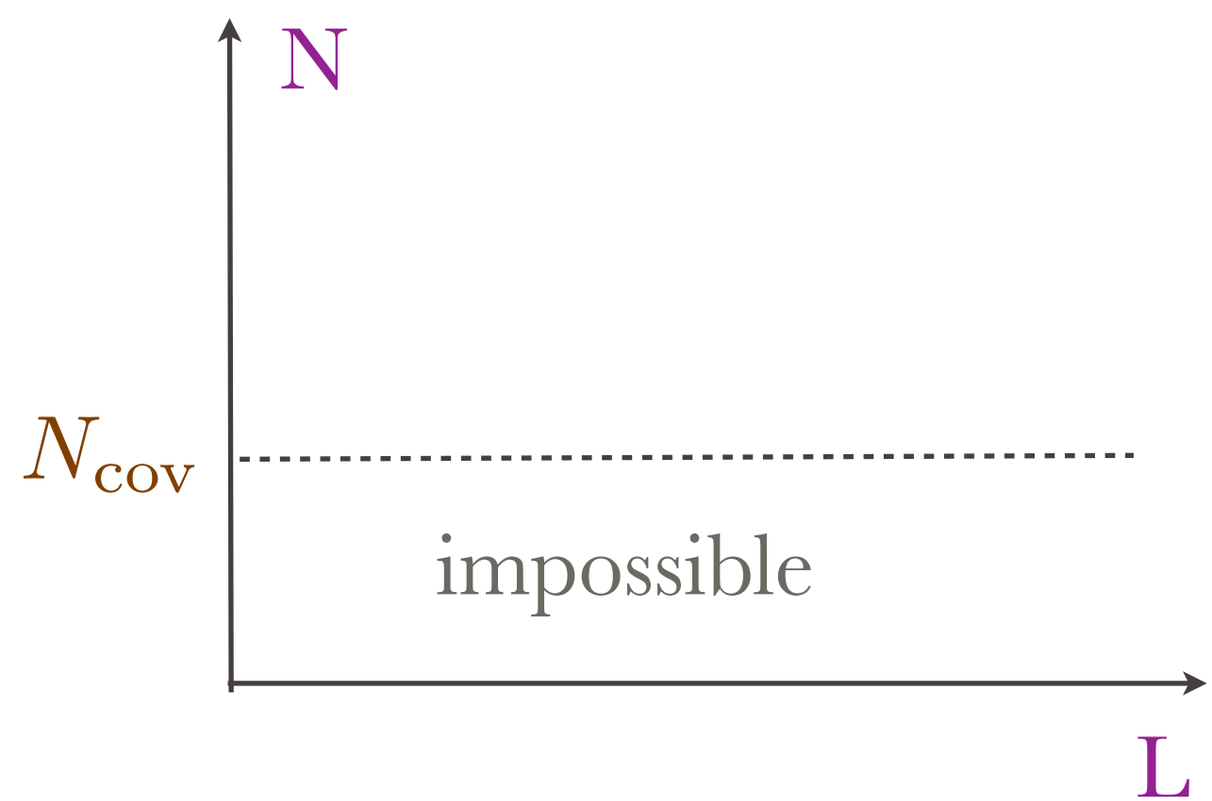
Coverage Condition

Coverage Condition

N_{cov} → is the minimum number of reads required to cover the sequence

Coverage Condition

N_{cov} → is the minimum number of reads required to cover the sequence



Repeat Condition

courtesy: Tse

Repeat Condition

- L-spectrum is given

Repeat Condition

- L-spectrum is given

ATAGCCTAGCGAT

Repeat Condition

- L-spectrum is given

ATAGCCTAGCGAT
ATAGC
TAGCC
AGCCT
GCCTA
CCTAG
CTAGC
TAGCG
AGCGA
GCGAT

Repeat Condition

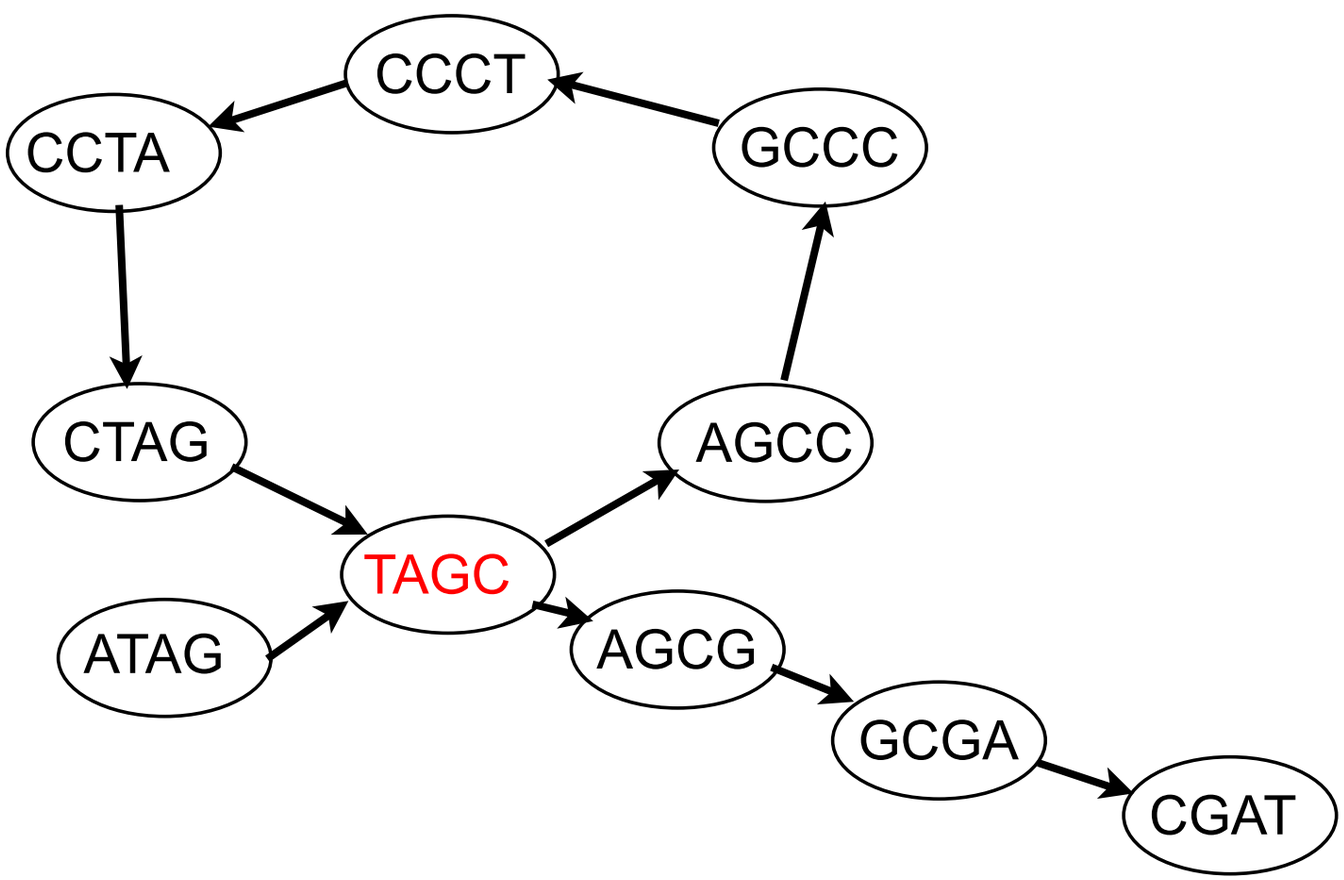
de Bruijn Graph

- Add a node for each (L-1)- mer in a read
- Add edges for adjacent (L-1)-mers

Repeat Condition

de Bruijn Graph

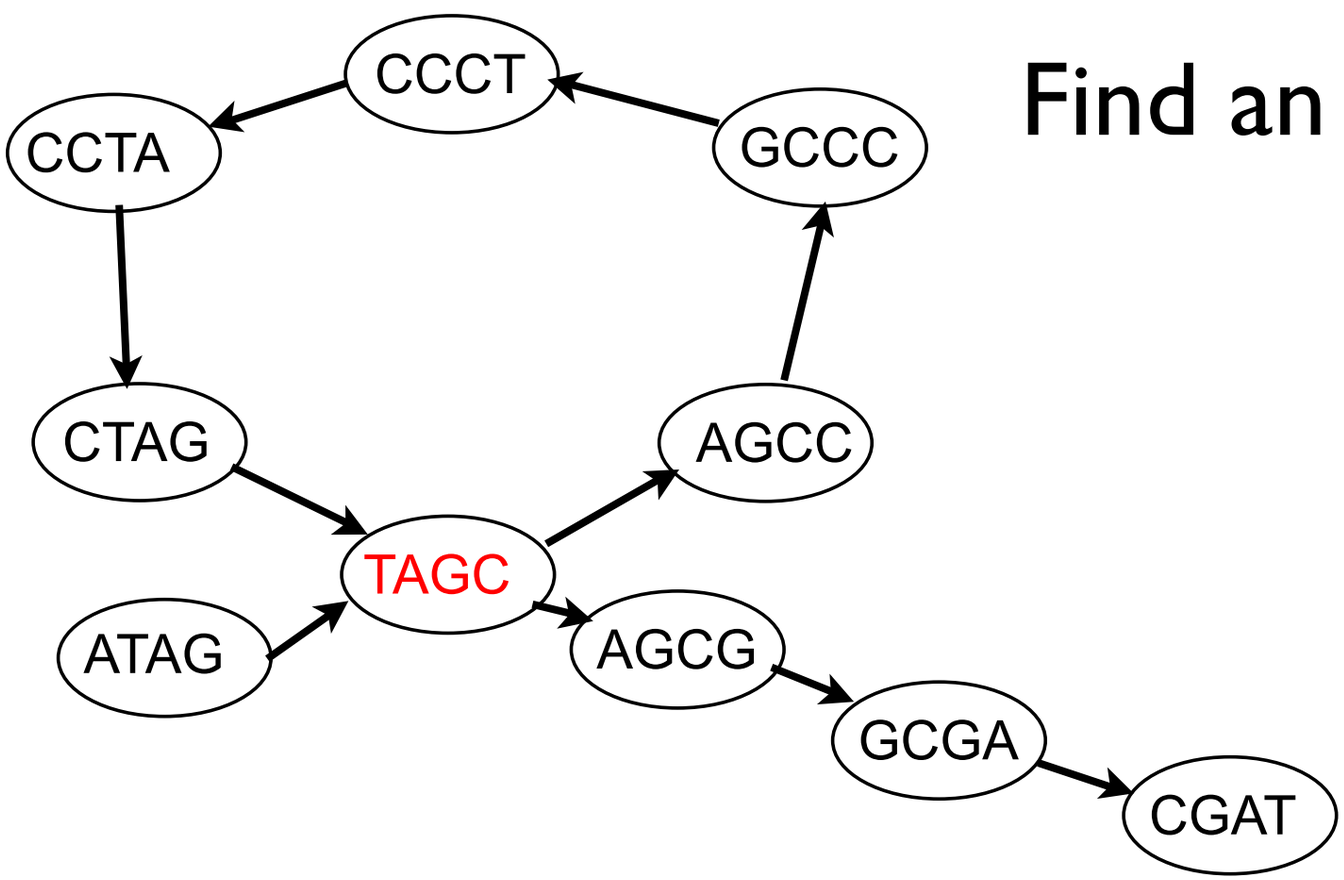
- Add a node for each (L-1)- mer in a read
- Add edges for adjacent (L-1)-mers



Repeat Condition

de Bruijn Graph

- Add a node for each (L-1)- mer in a read
- Add edges for adjacent (L-1)-mers

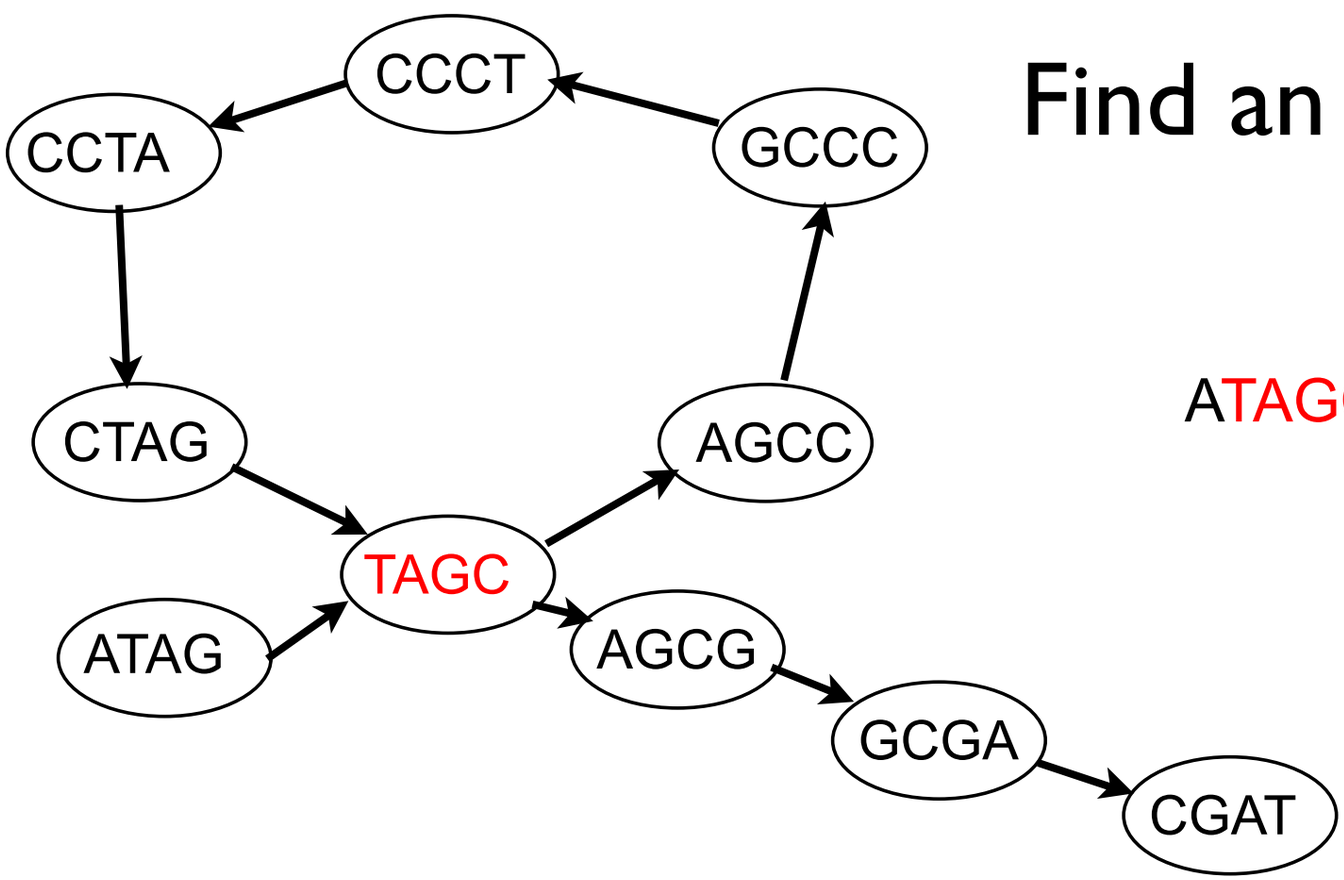


Find an Eulerian Path

Repeat Condition

de Bruijn Graph

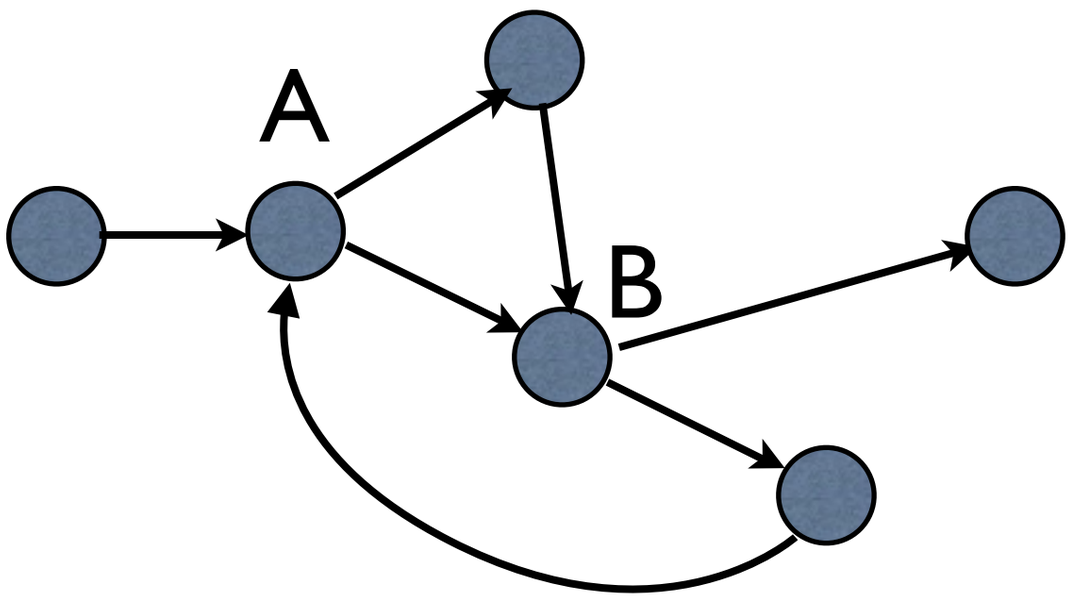
- Add a node for each (L-1)- mer in a read
- Add edges for adjacent (L-1)-mers

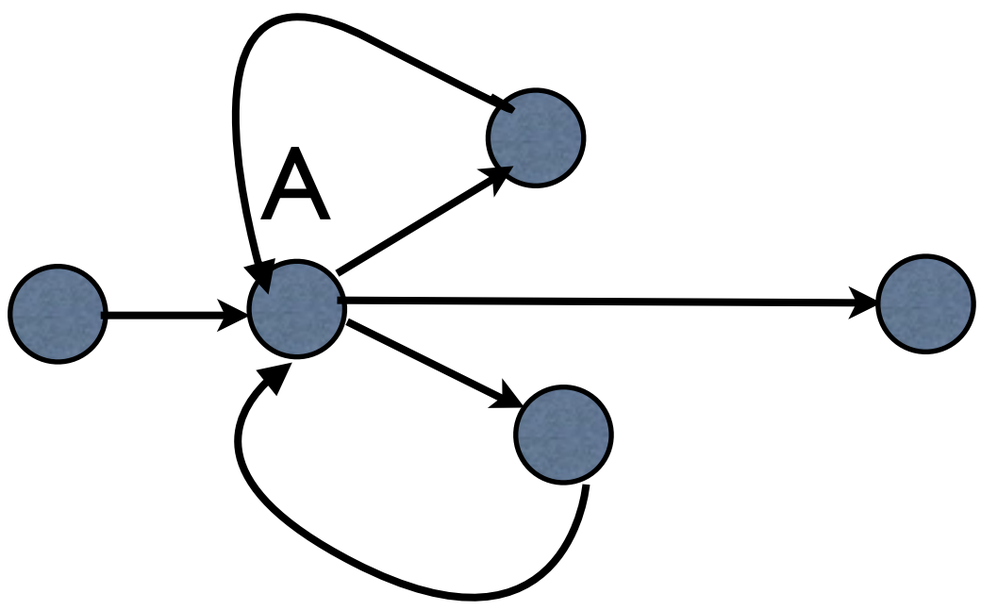


Find an Eulerian Path

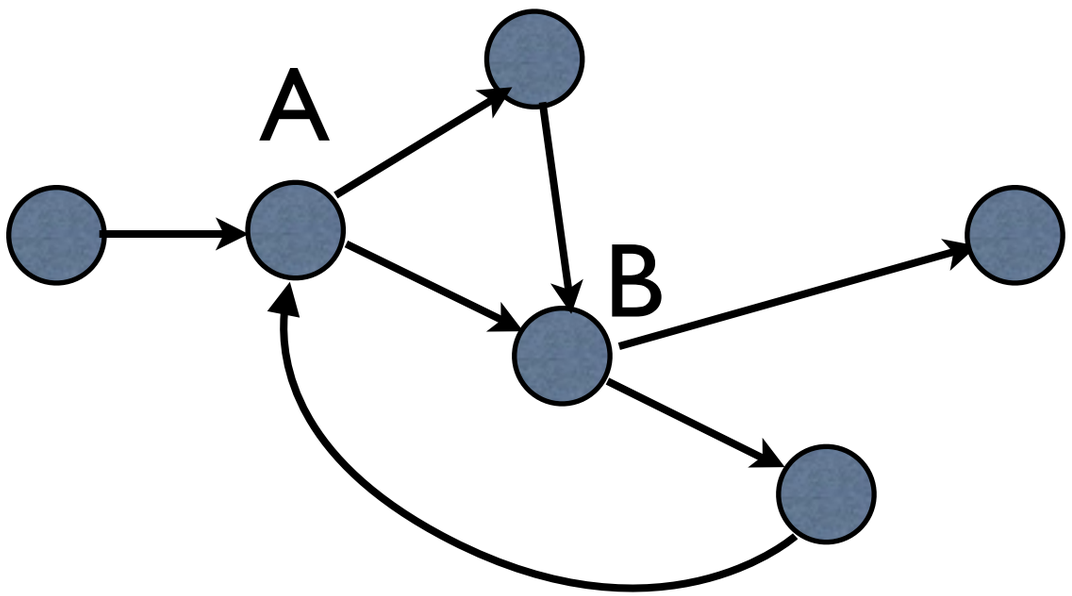
ATAGCCCTAGCGAT

Repeat Condition

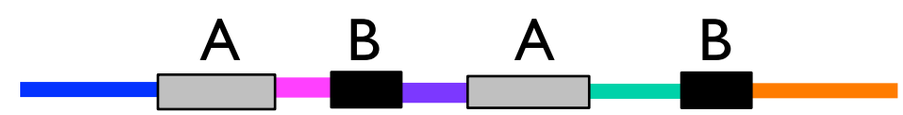


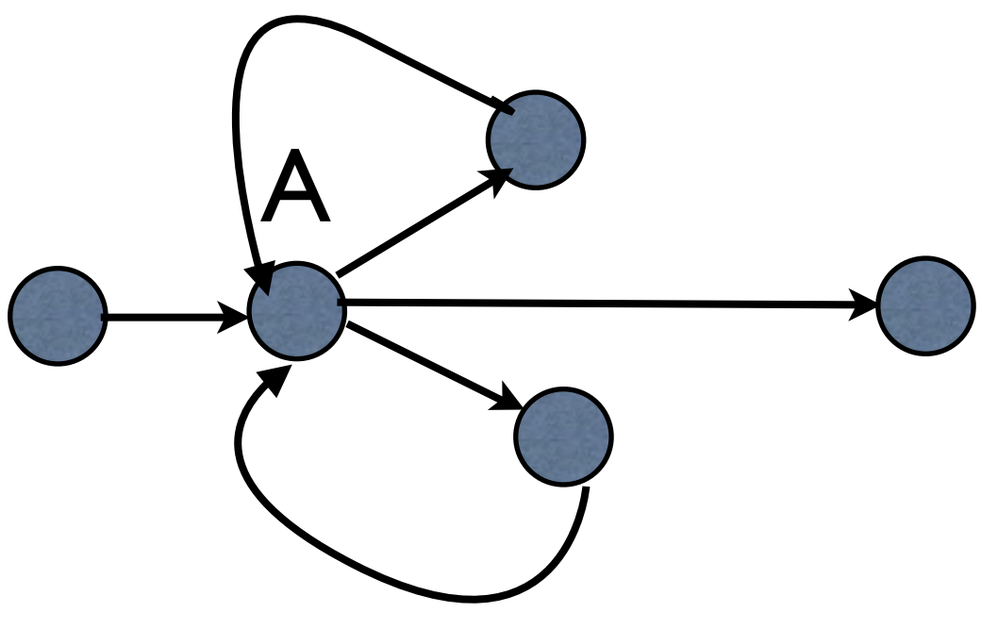


Repeat Condition

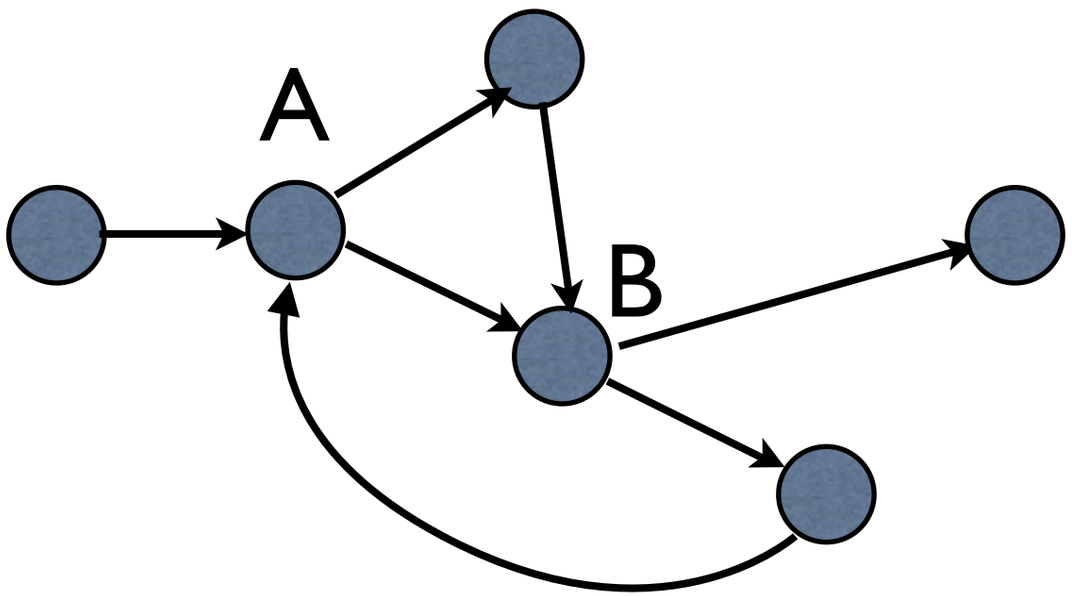


Interleaved repeats

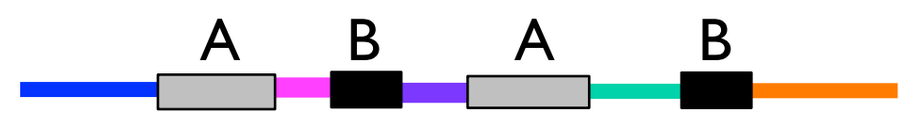


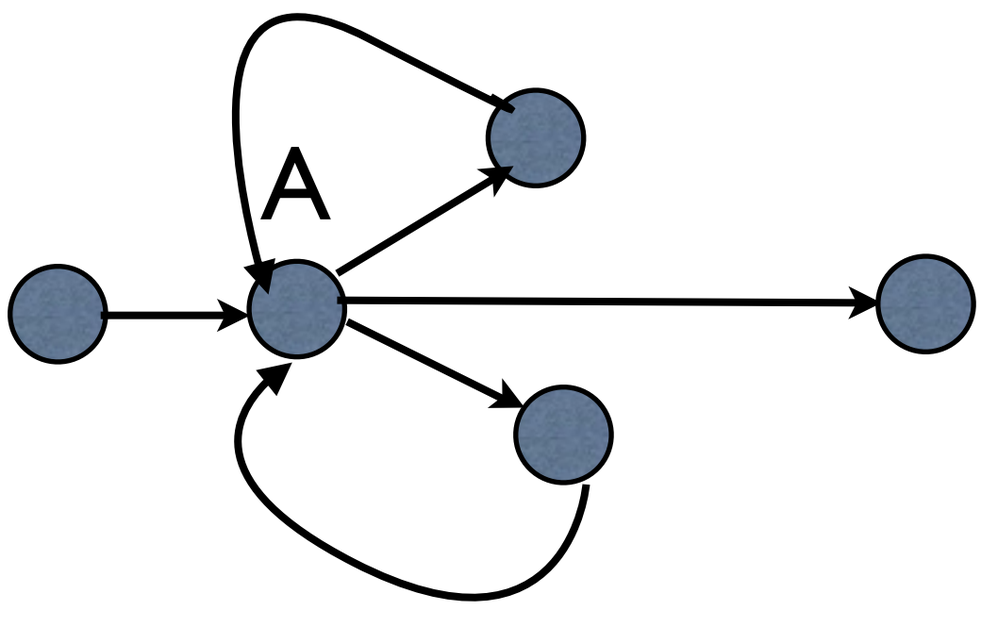


Repeat Condition

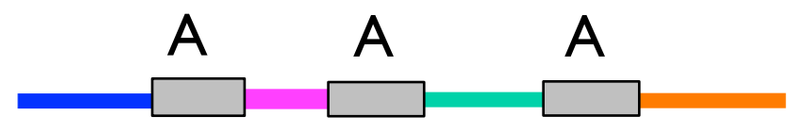


Interleaved repeats

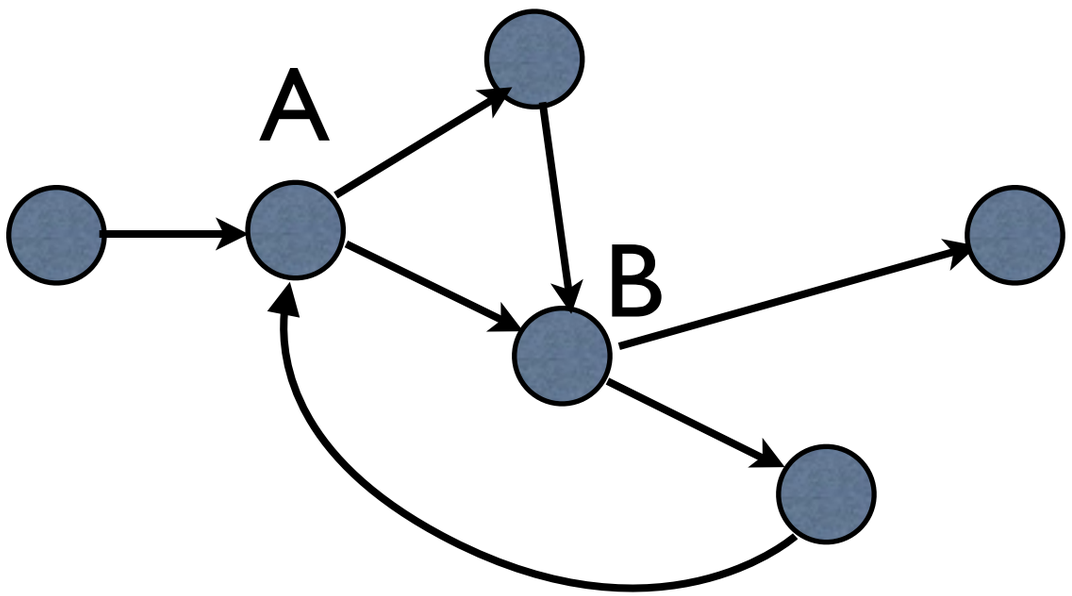




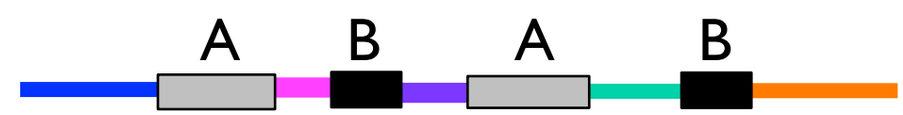
Triple repeat

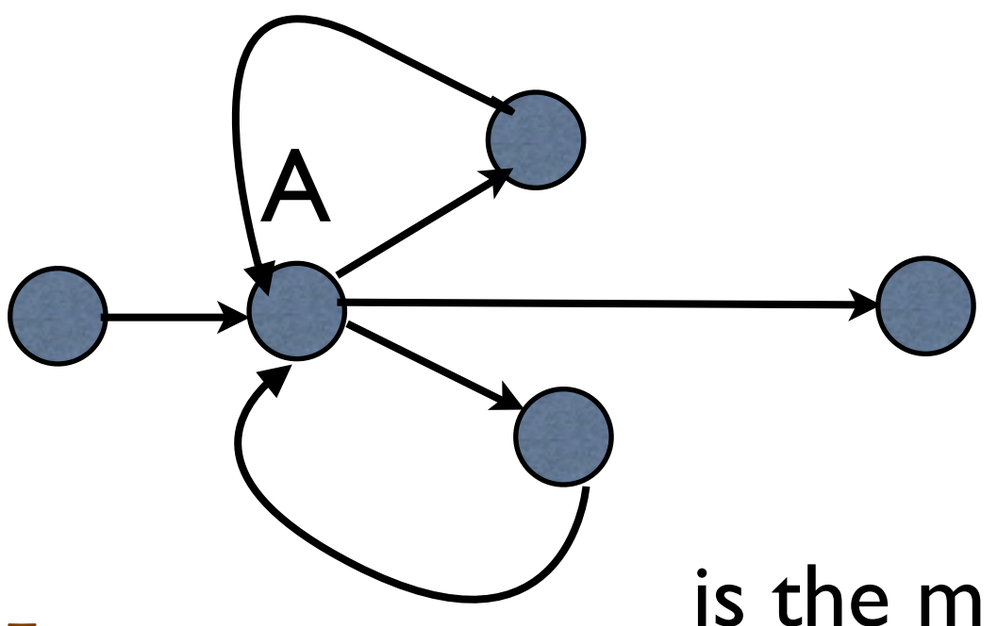


Repeat Condition

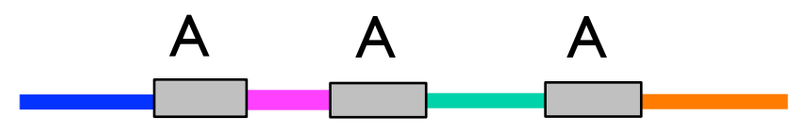


Interleaved repeats





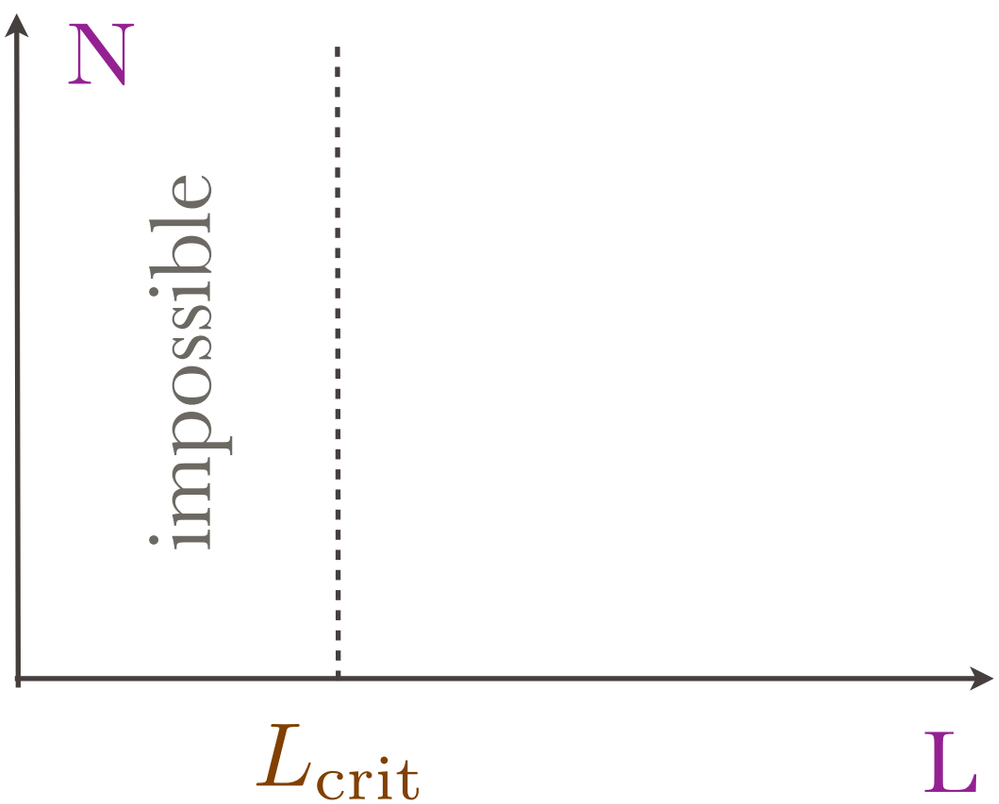
Triple repeat



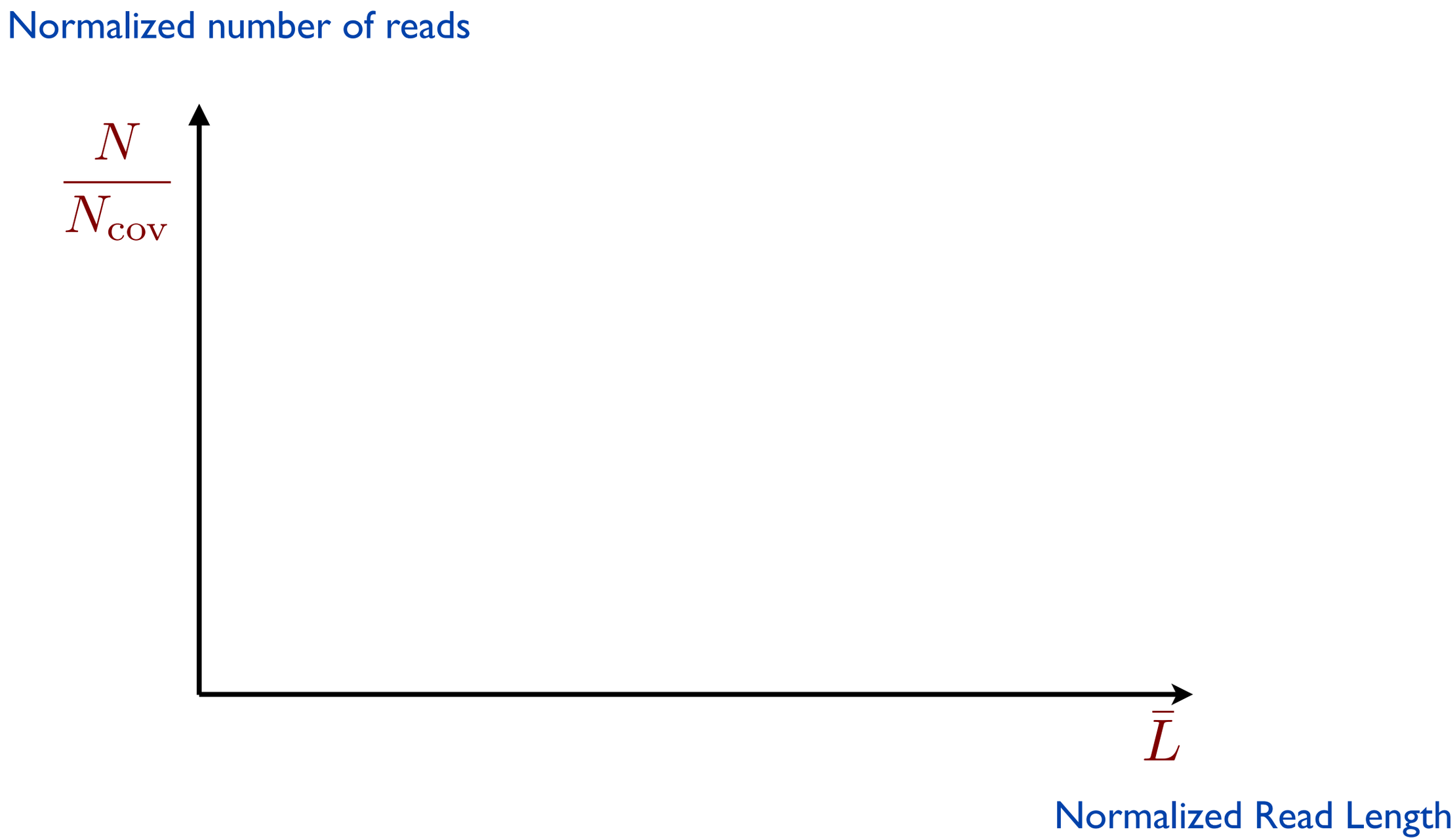
L_{crit} is the minimum read length where there exists a unique solution for the problem

Repeat Condition

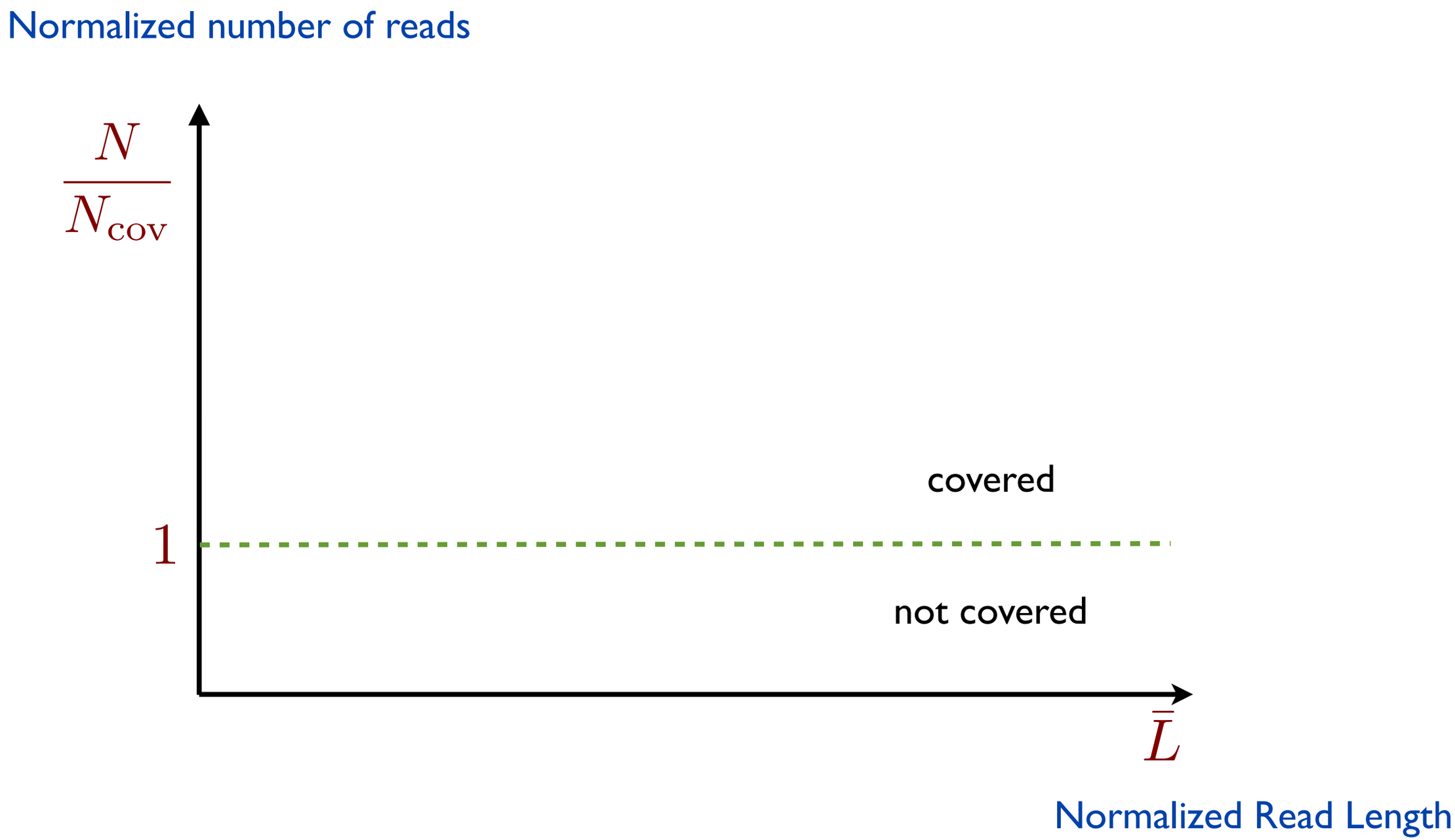
$$L > L_{\text{crit}}$$



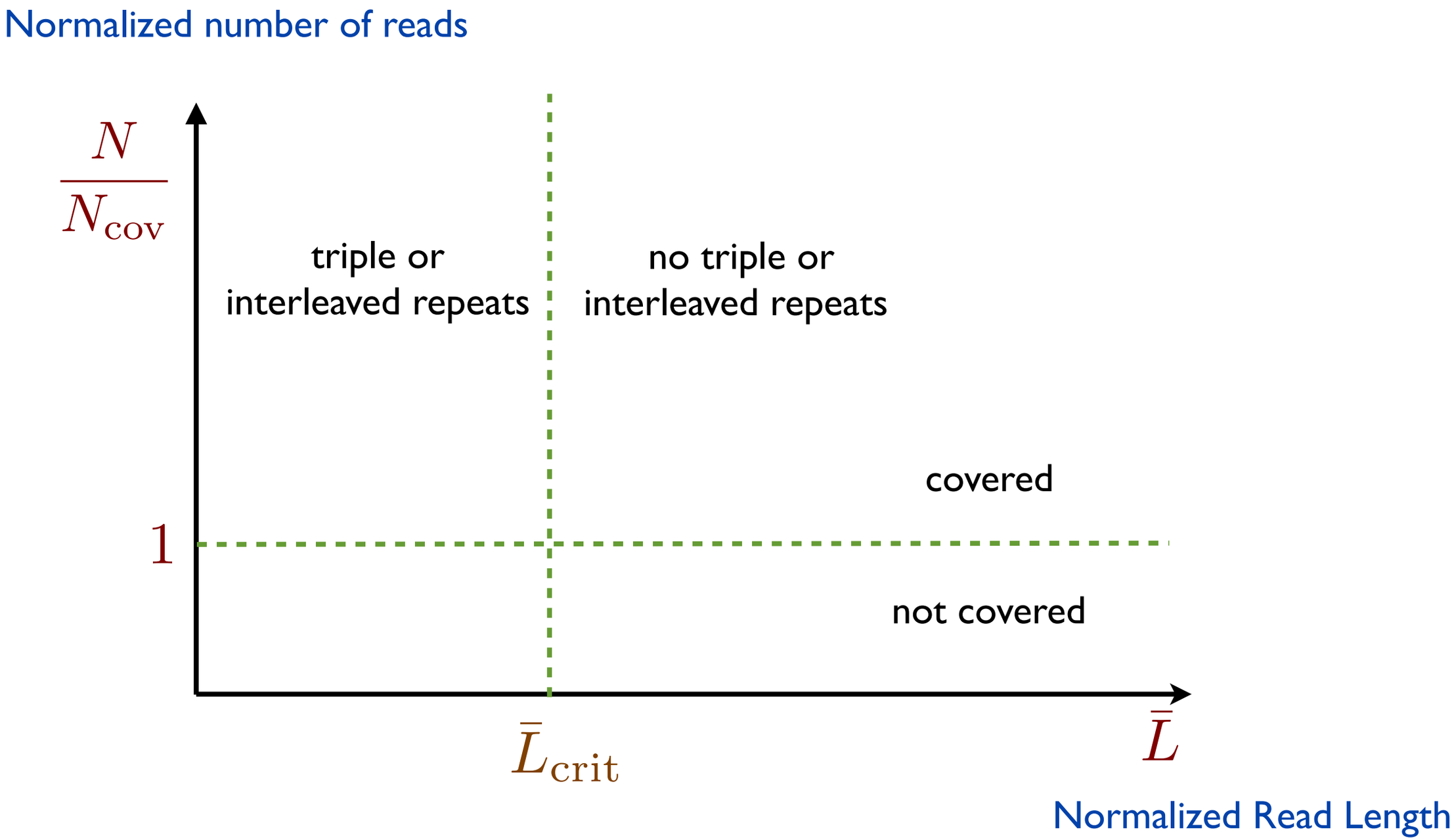
Putting together



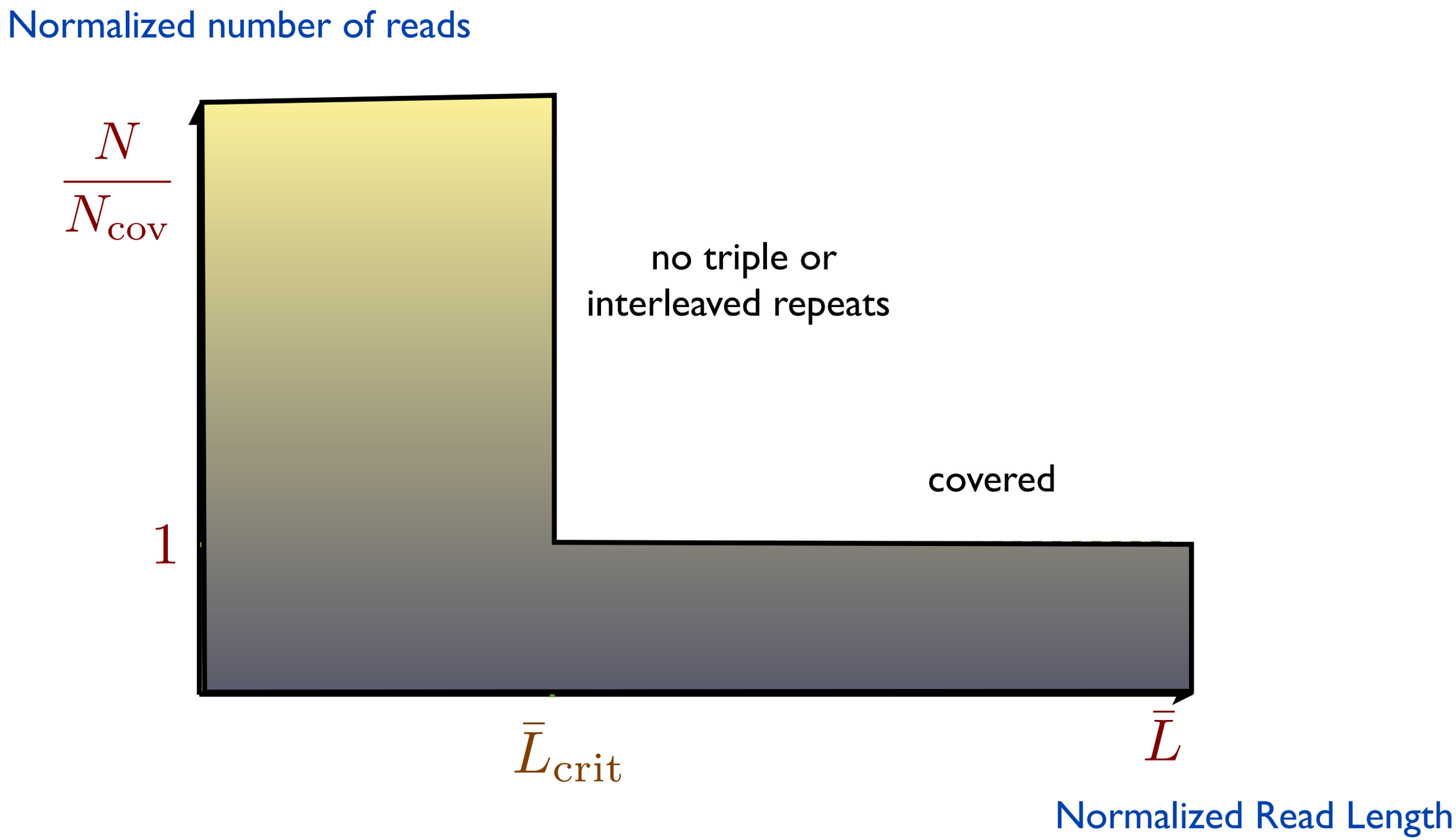
Putting together



Putting together

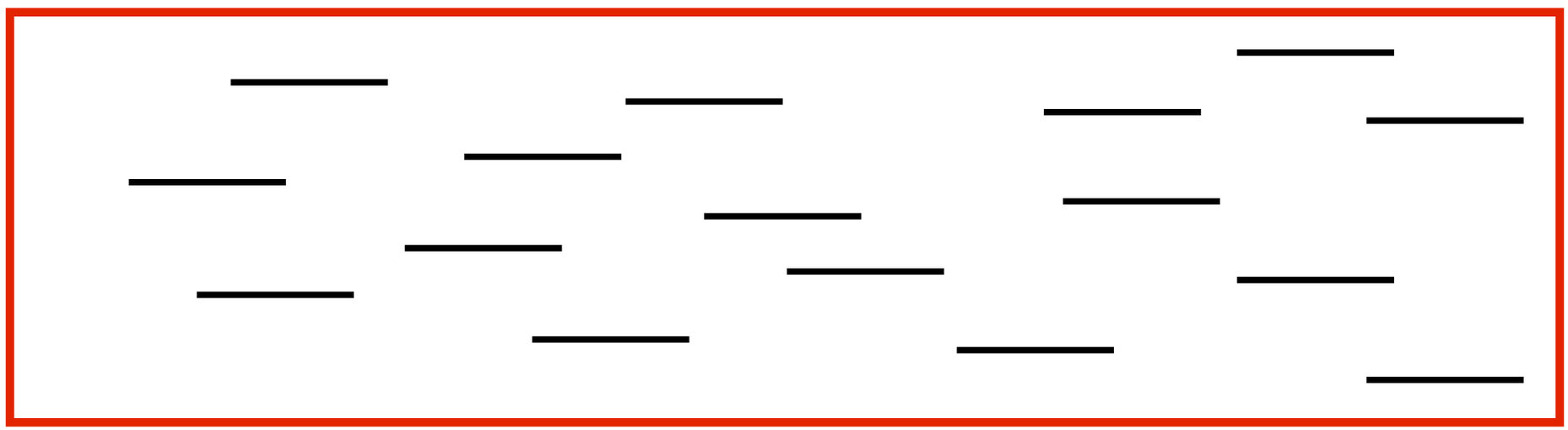


Putting together



what about the rest?

The Greedy Algorithm



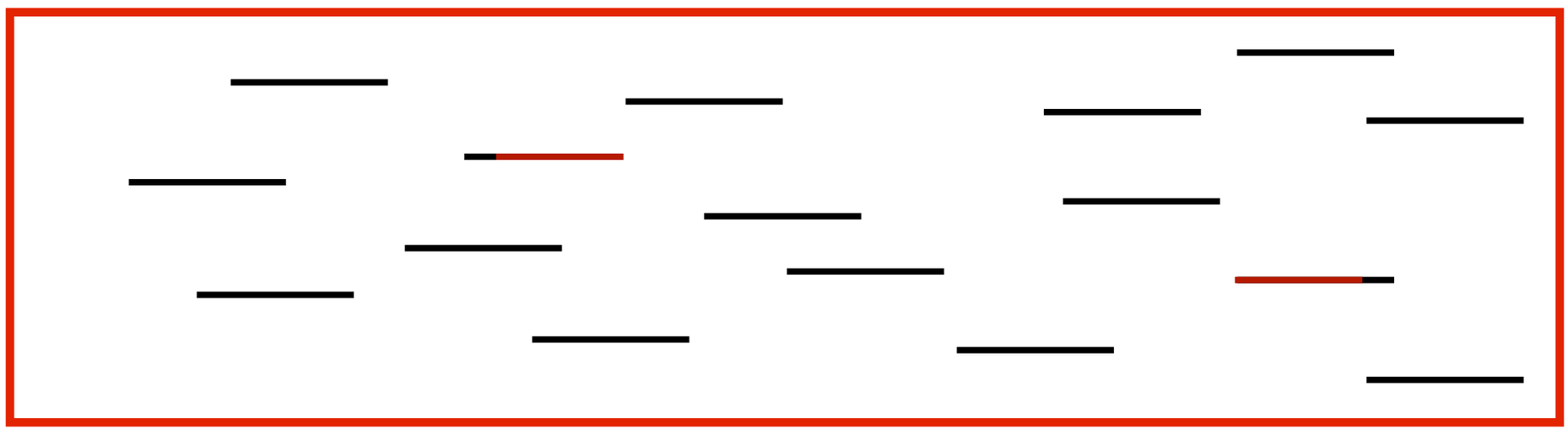
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



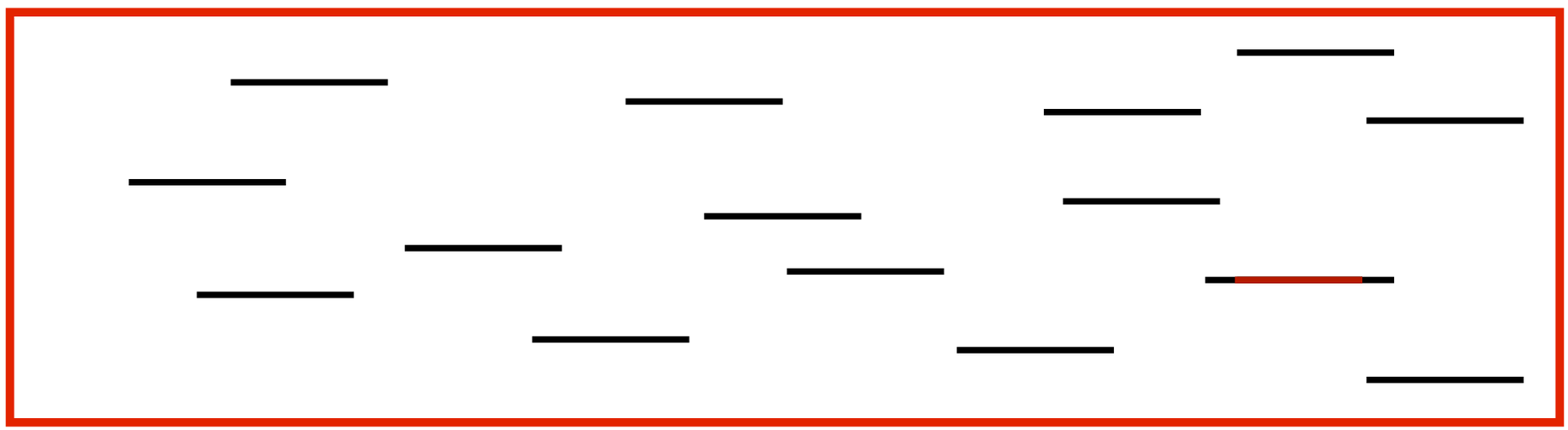
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



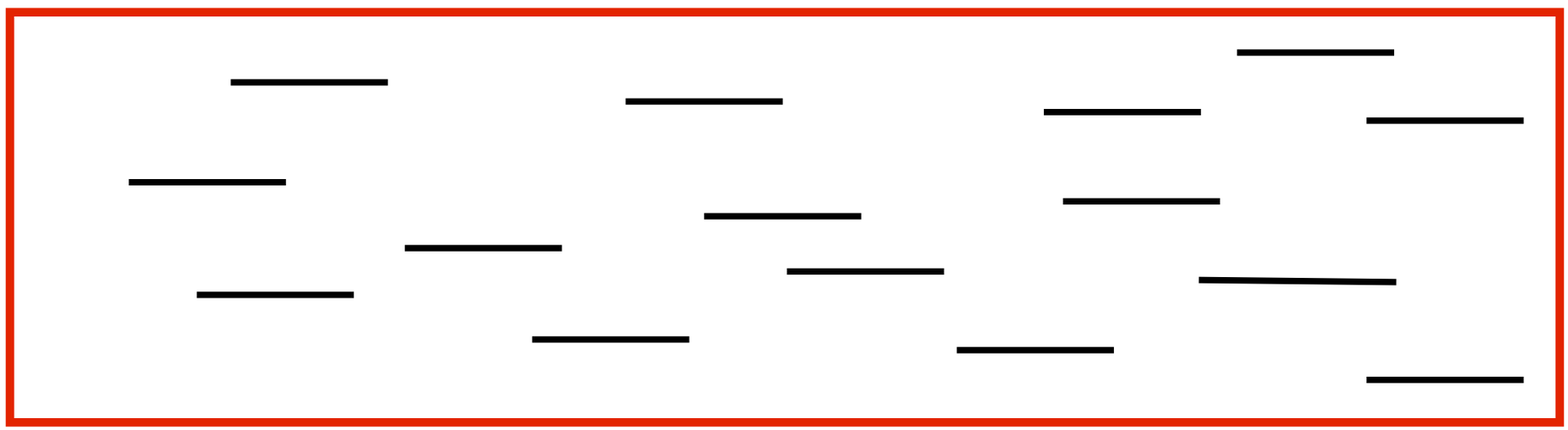
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



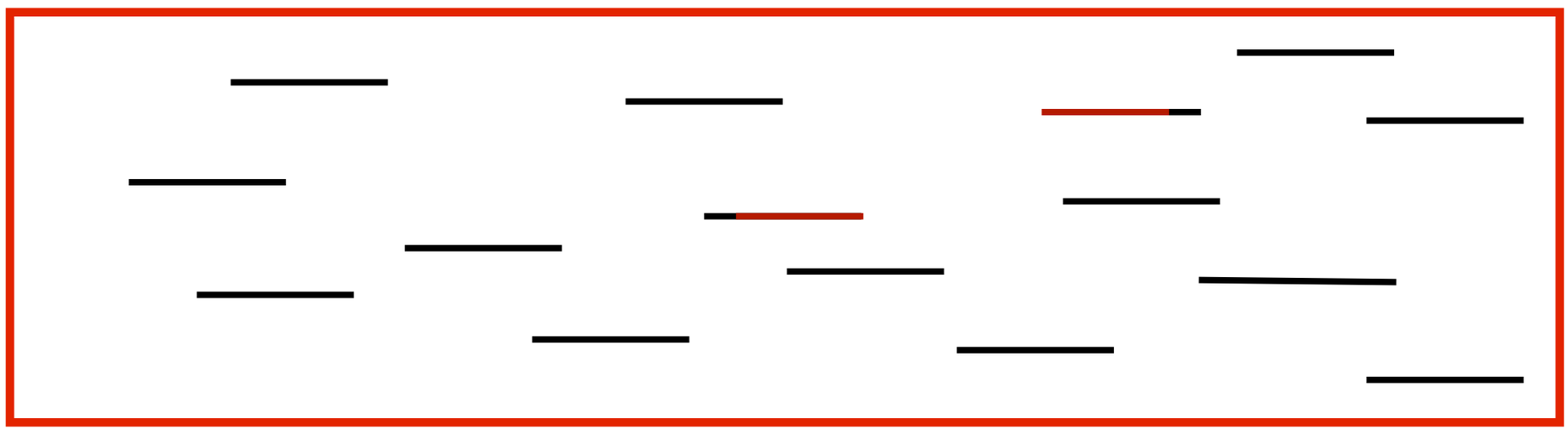
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



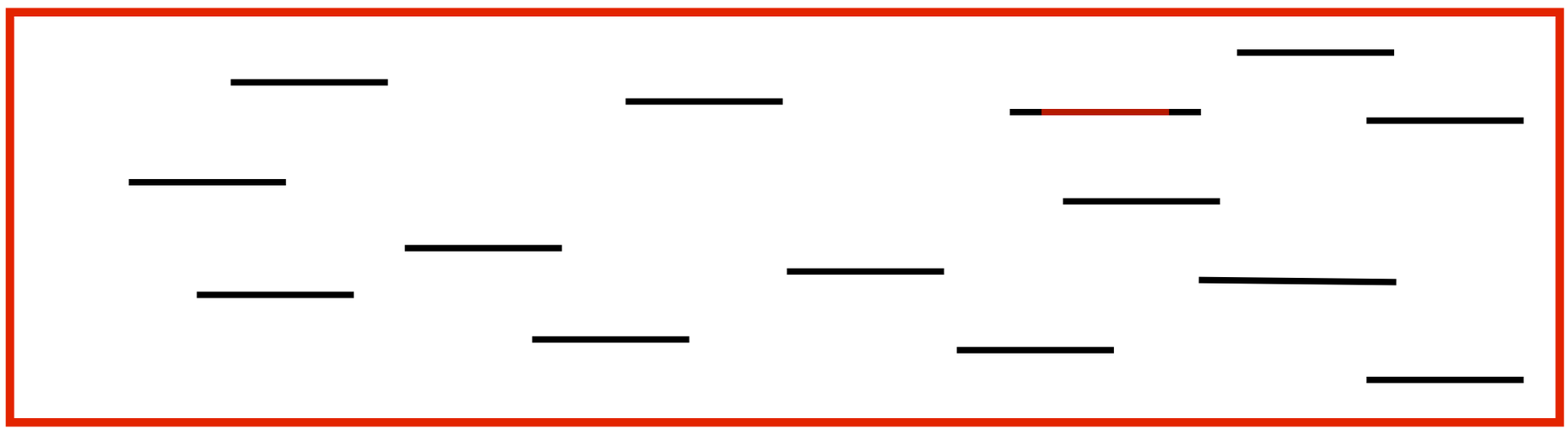
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



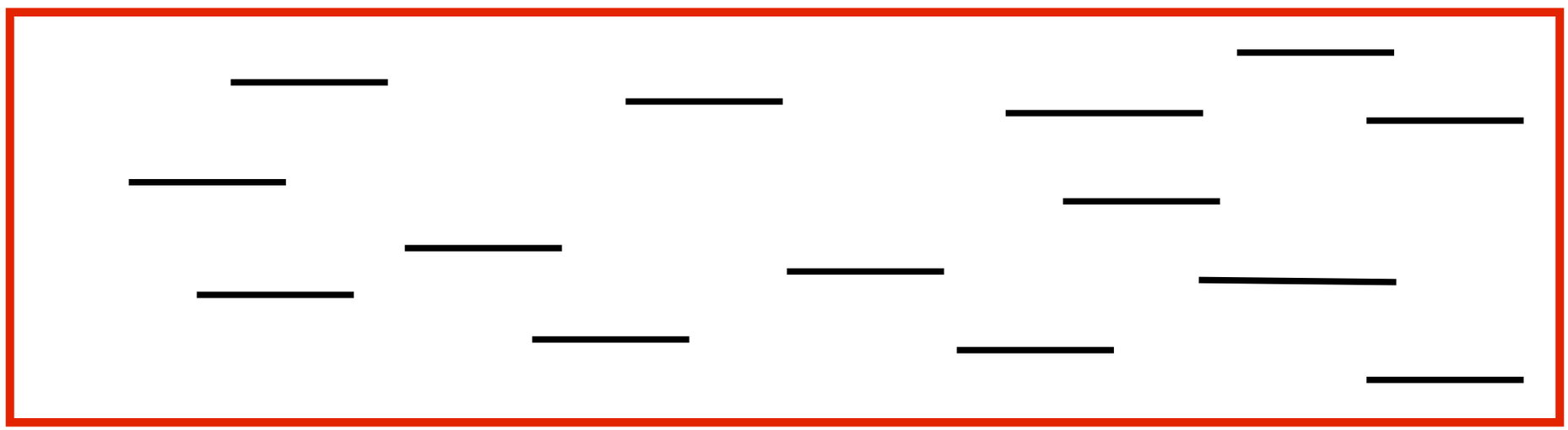
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



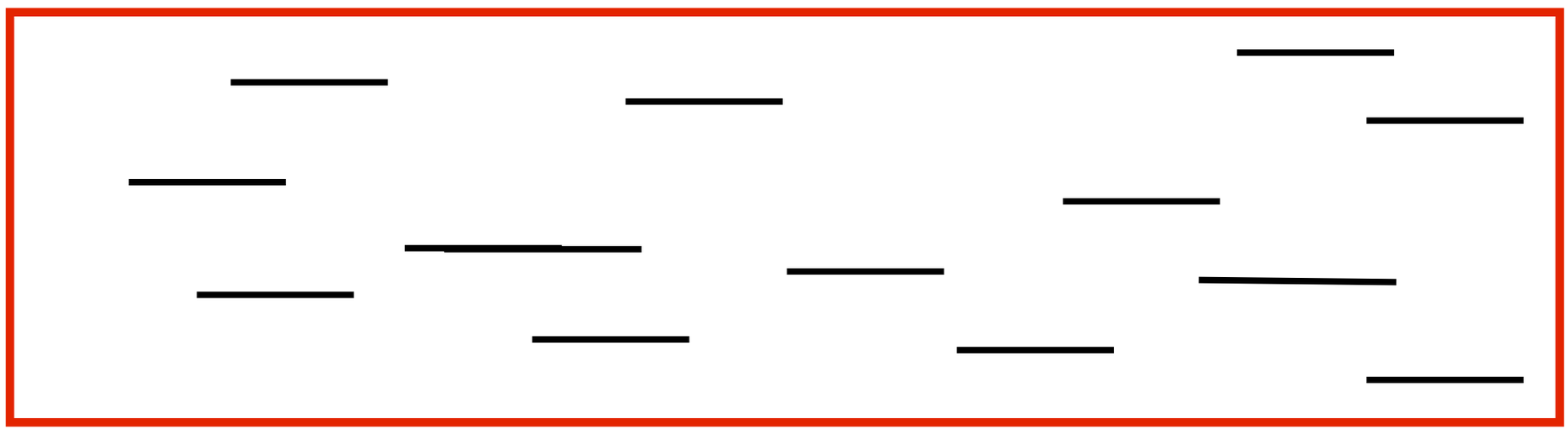
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



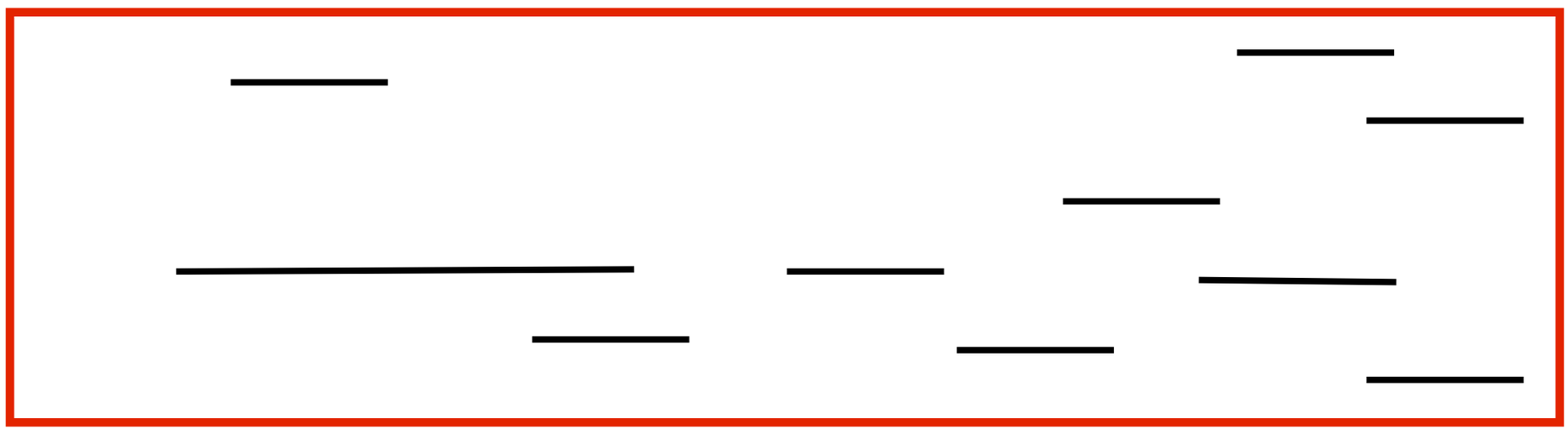
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



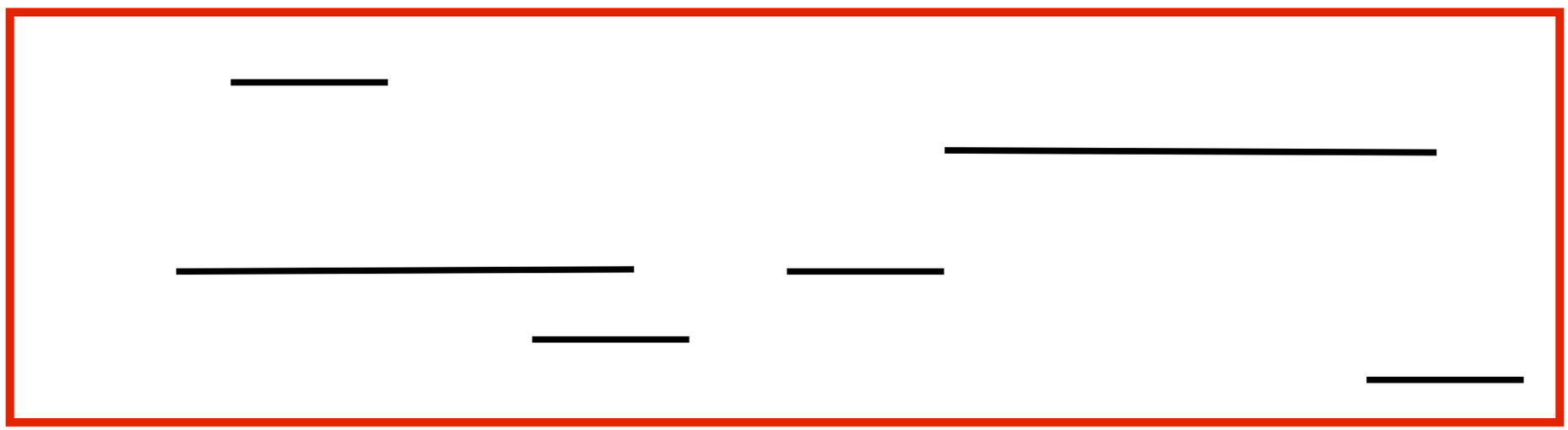
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

The Greedy Algorithm



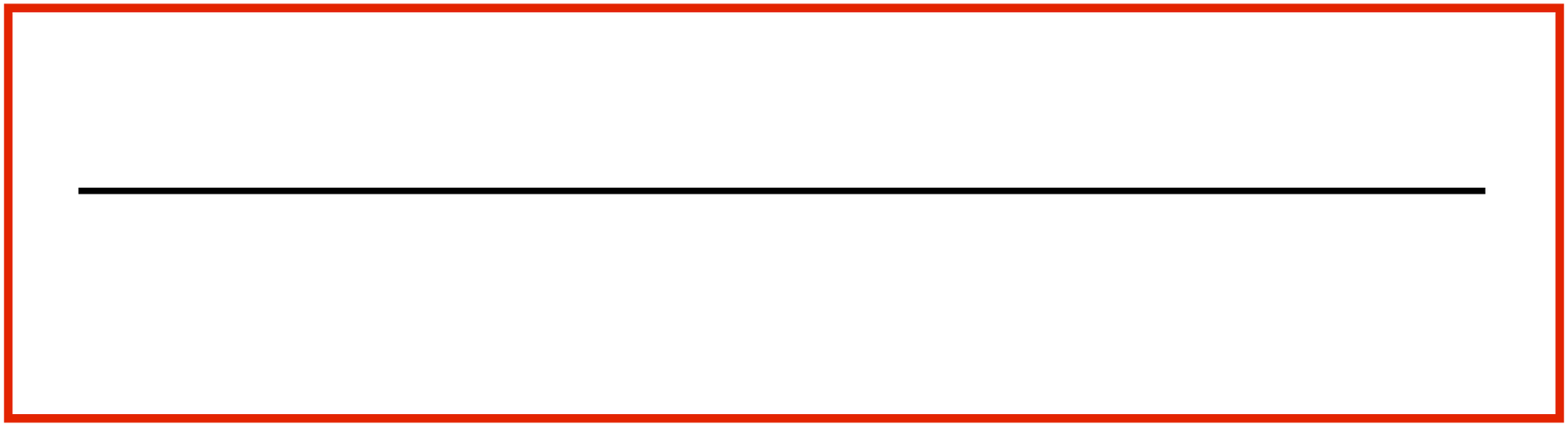
Algorithm

I. Find two reads with the largest overlap and merge them into one read.

II. Repeat I until only one contig remains.

what about the rest?

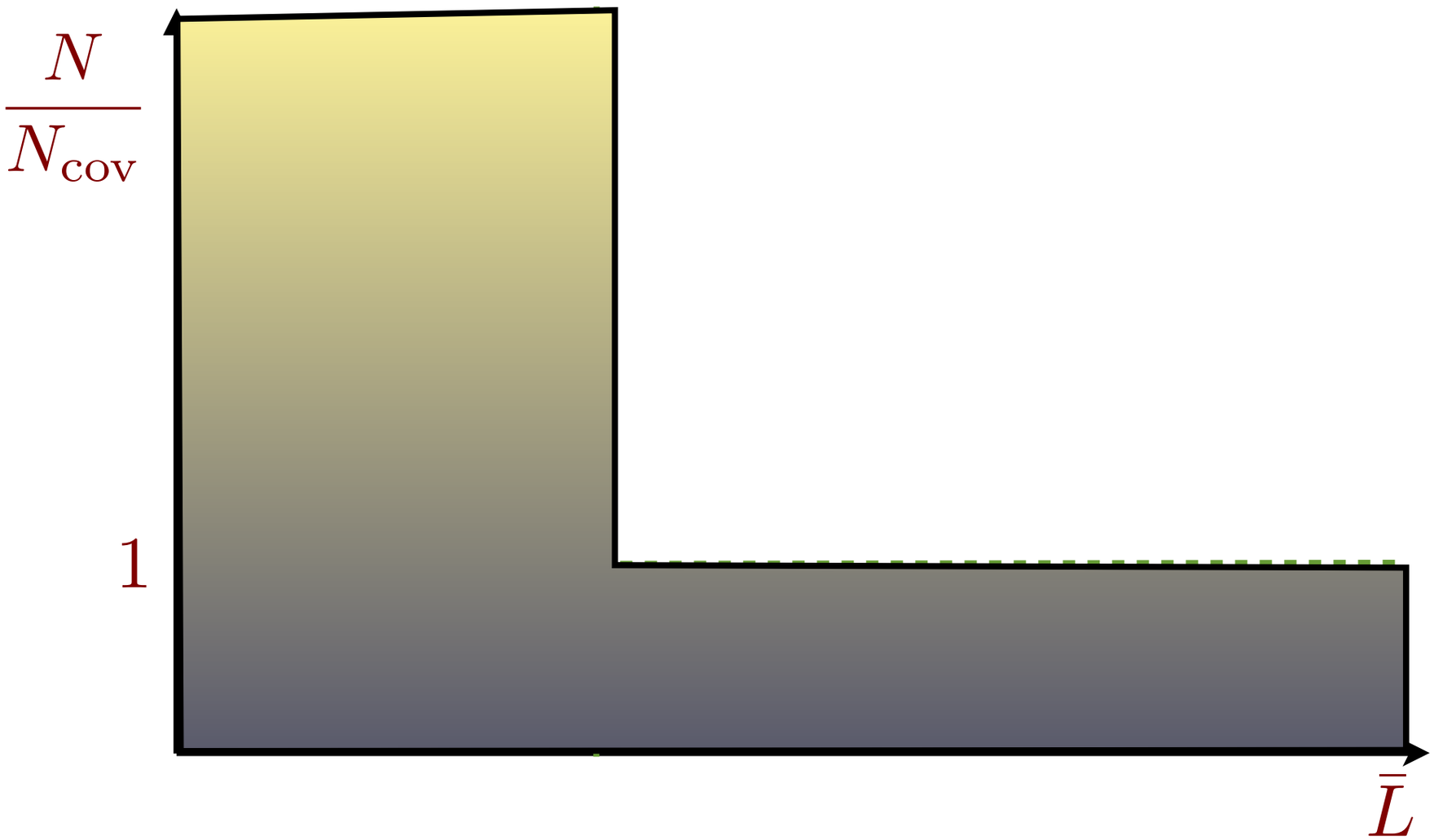
The Greedy Algorithm



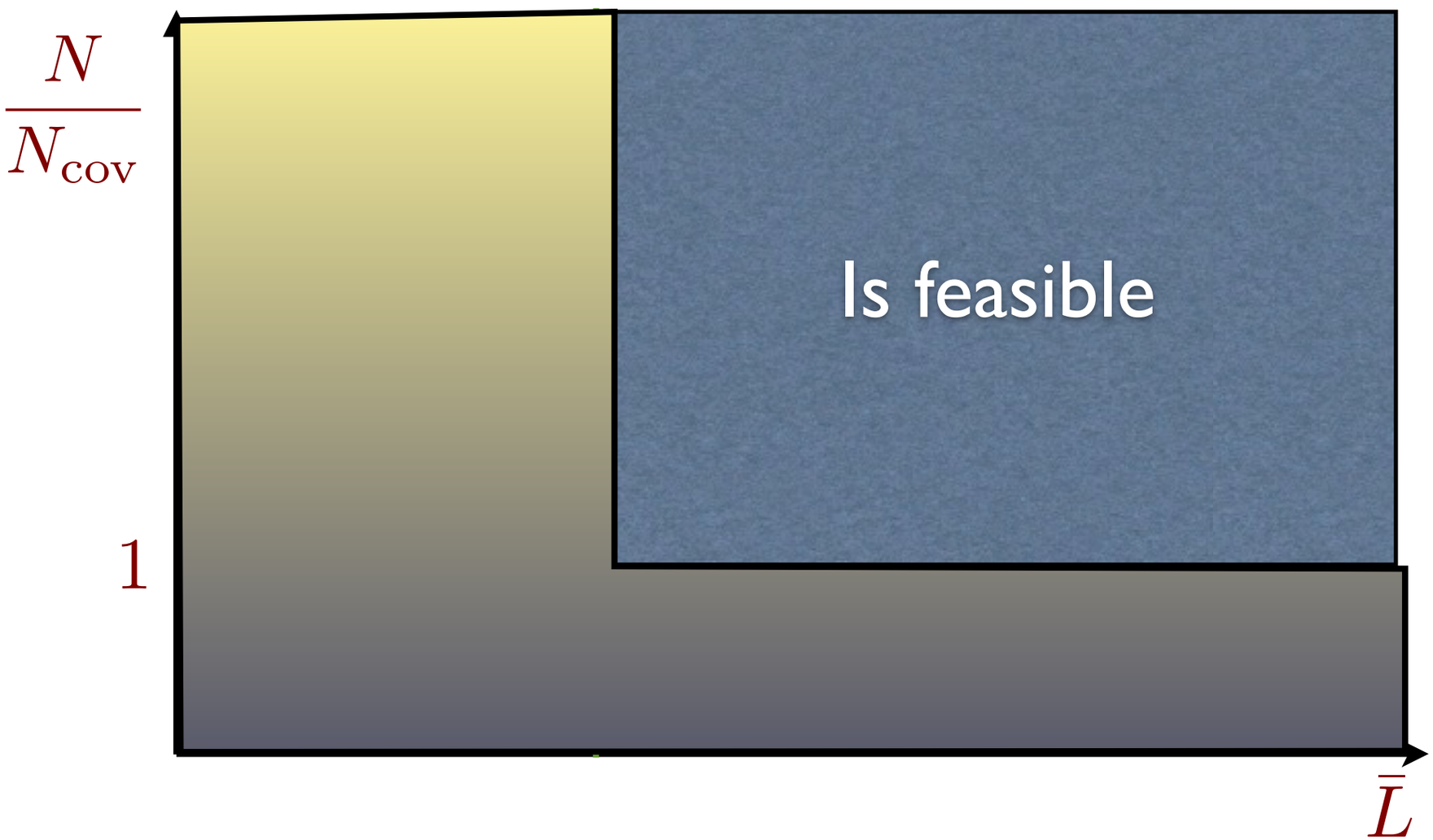
Algorithm

I. Find two reads with the largest overlap and merge them into one read.
II. Repeat I until only one contig remains.

RANDOM Sequence

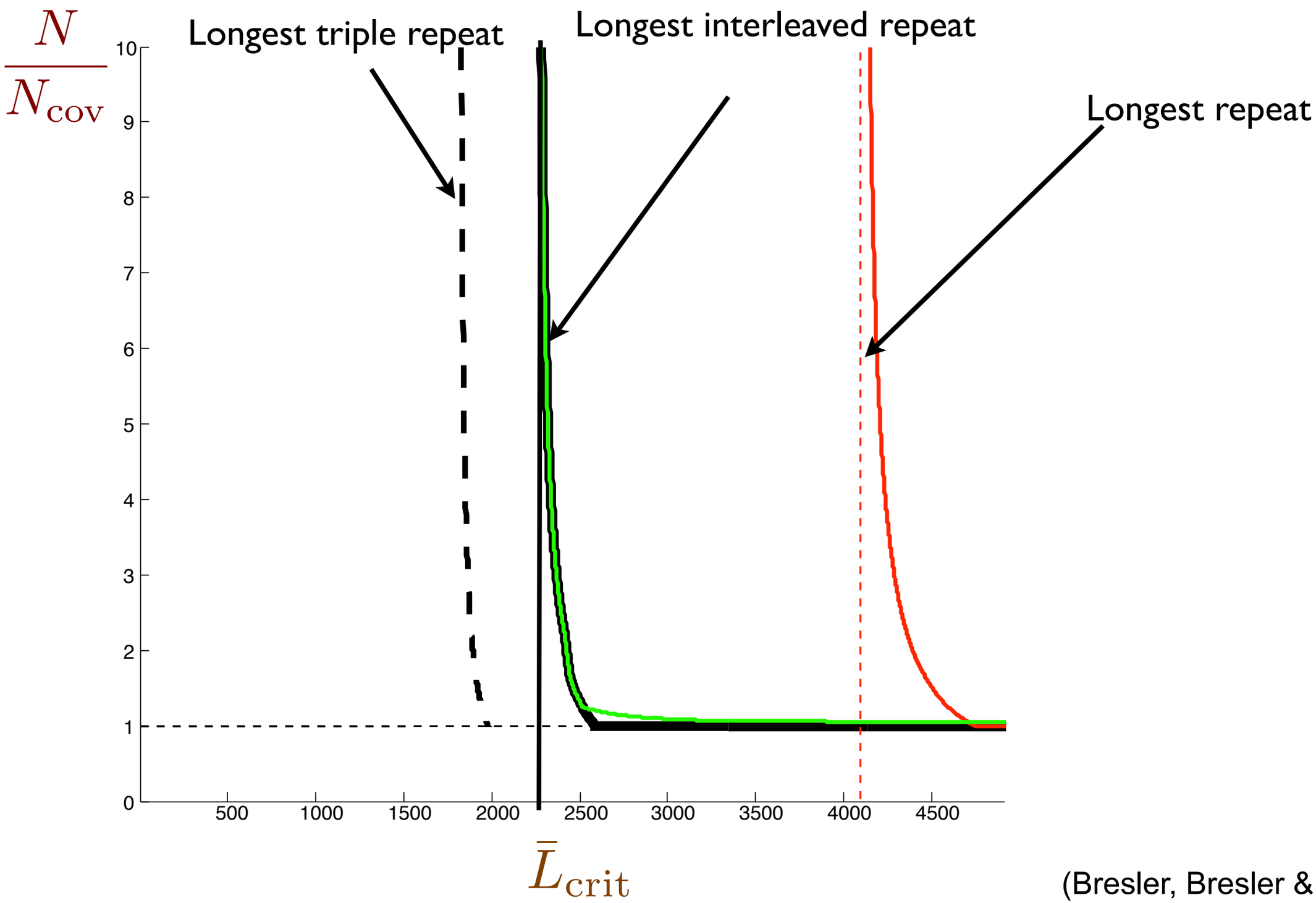


RANDOM Sequence



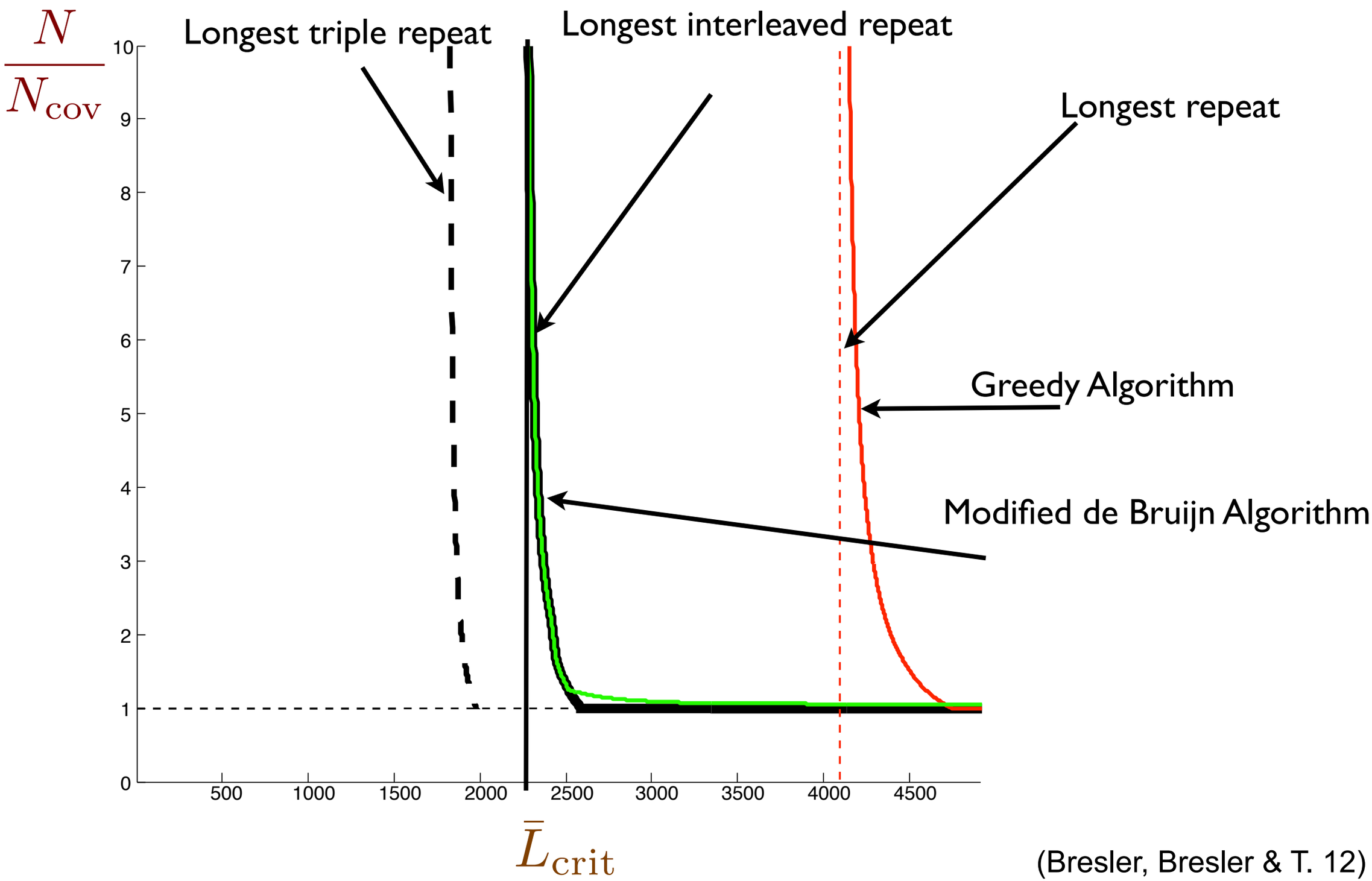
Motahari, Bresler, Tse (2012)

Human Chromosome 19



(Bresler, Bresler & T. 12)

Human Chromosome 19

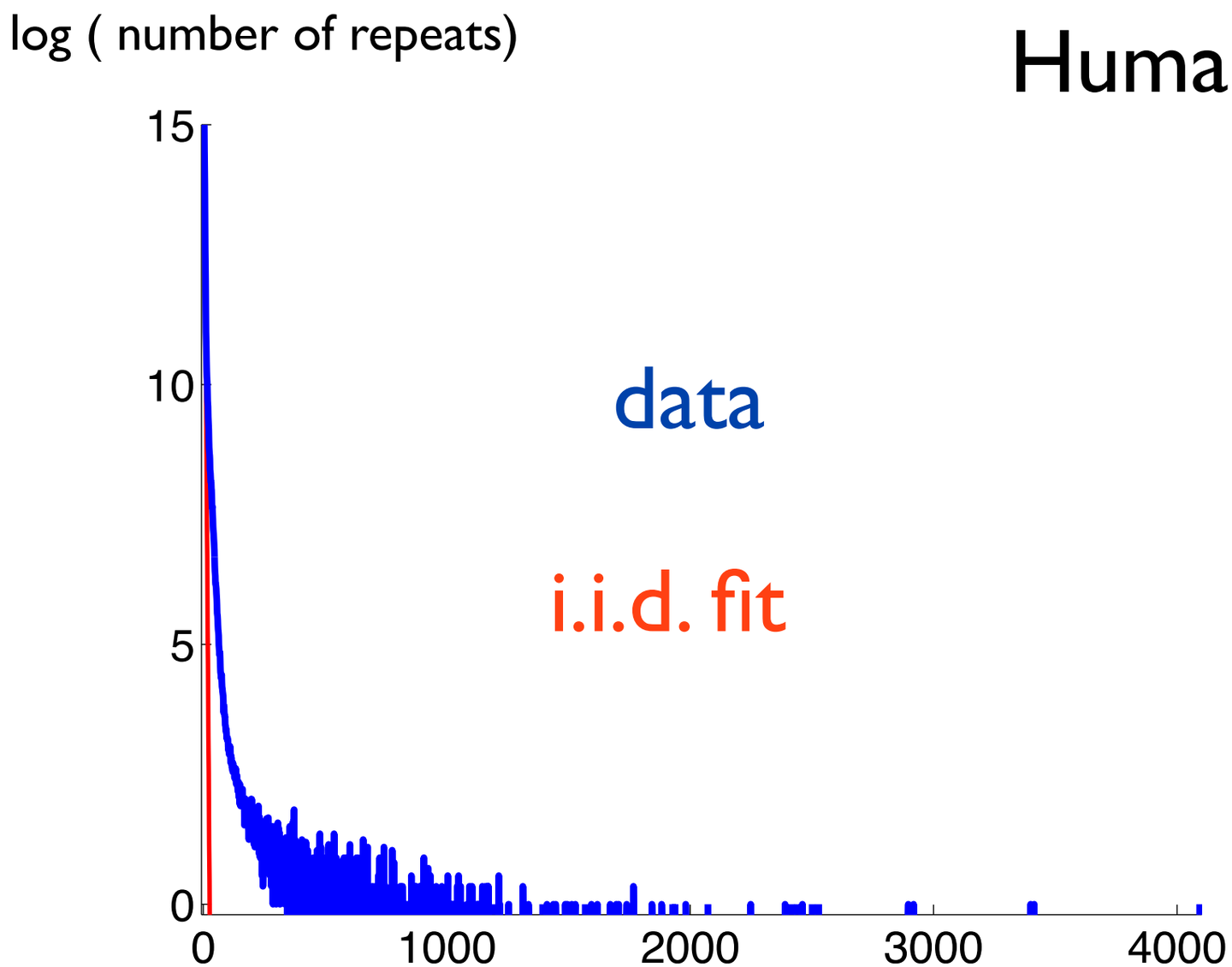


(Bresler, Bresler & T. 12)

WHY?

(Bresler, Bresler & T. 12)

WHY?



(Bresler, Bresler & T. 12)