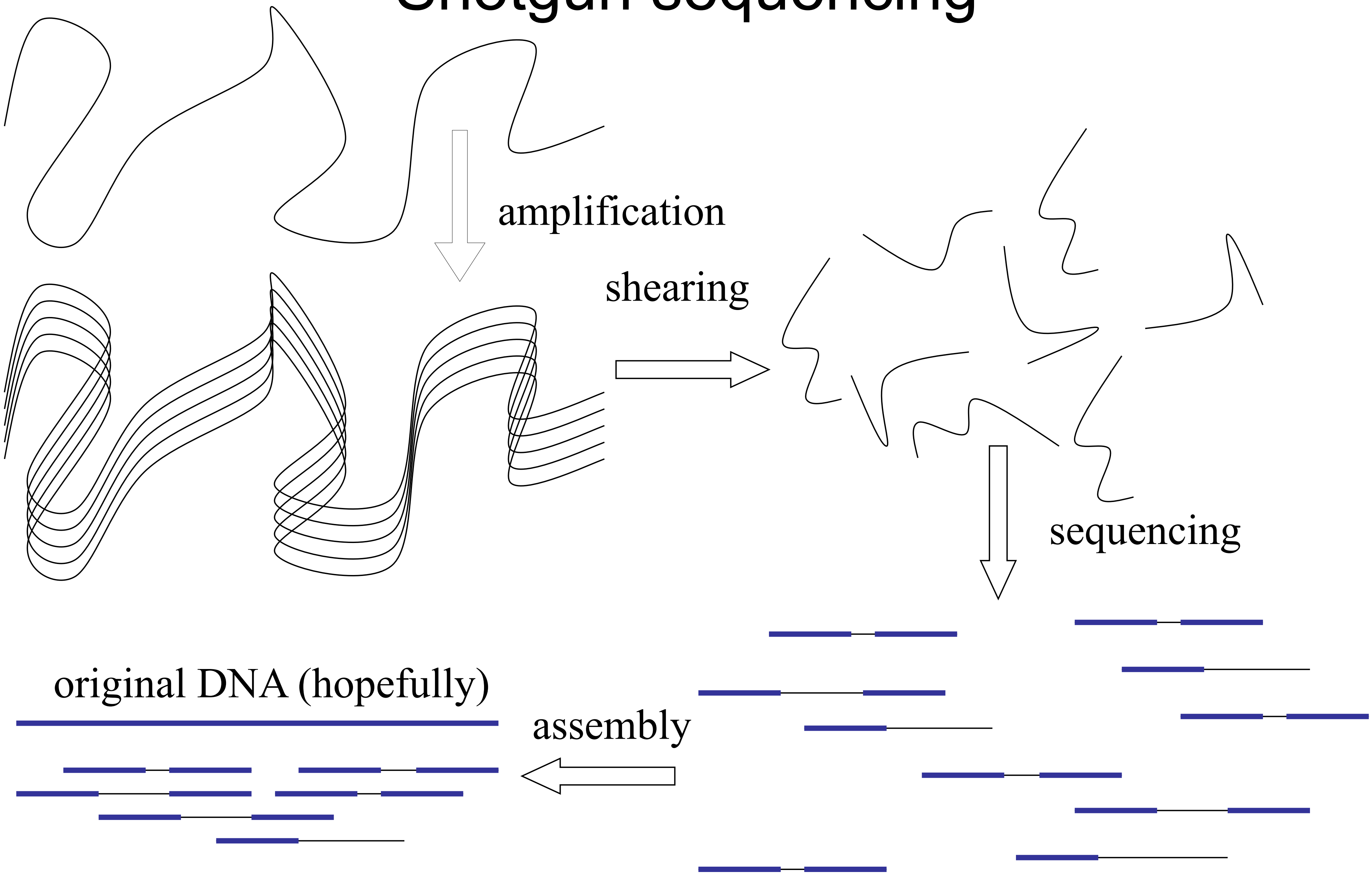
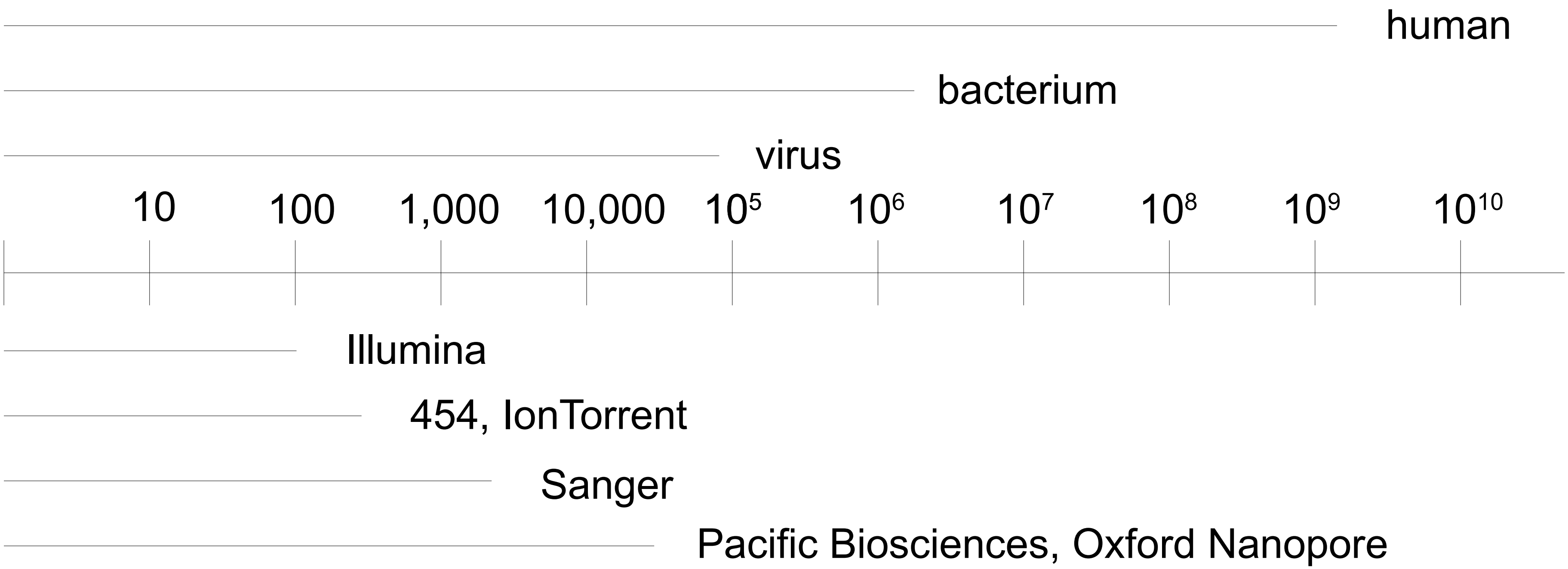
Introduction to Shotgun Sequencing

Mihai Pop

Shotgun sequencing



Why shotgun sequencing?



human

bacterium

virus

10 100 1,000 10,000 10^5 10^6 10^7 10^8 10^9 10^{10}

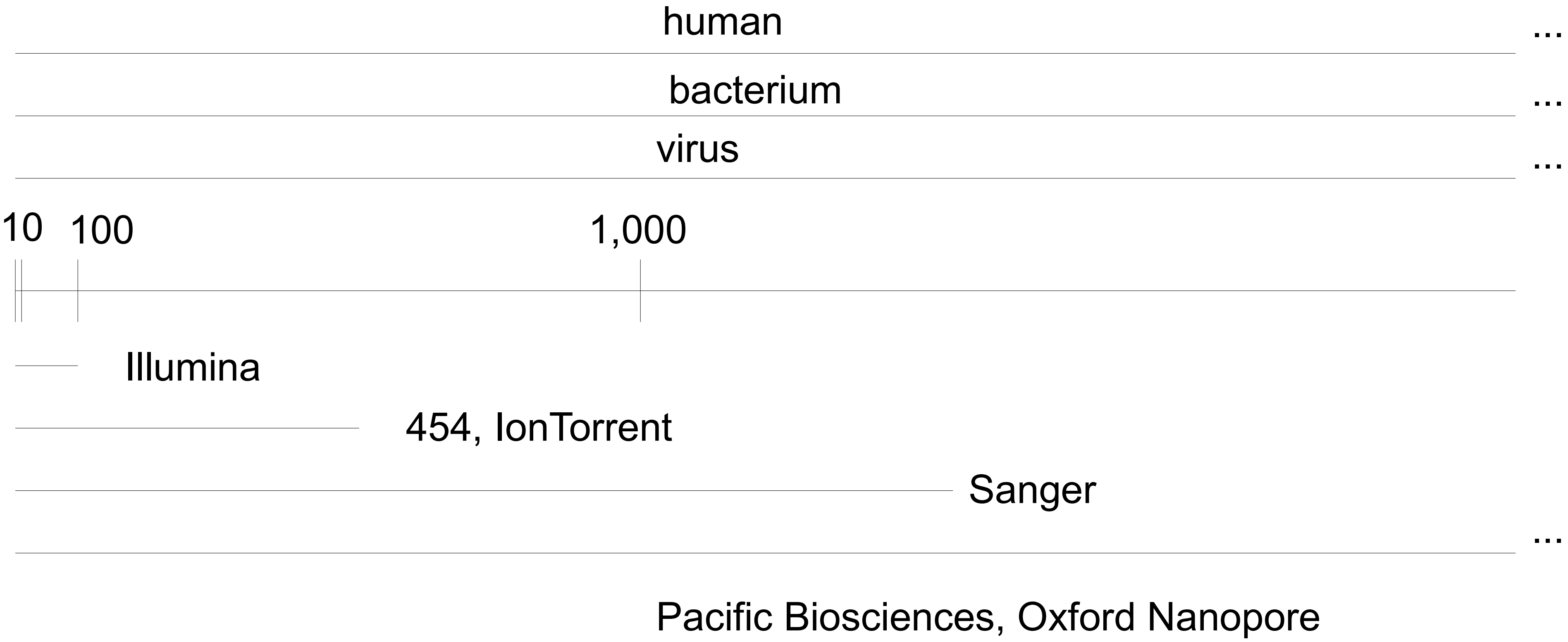
Illumina

454, IonTorrent

Sanger

Pacific Biosciences, Oxford Nanopore

Why shotgun sequencing?



human ...

bacterium ...

virus ...

10 100 1,000

Illumina

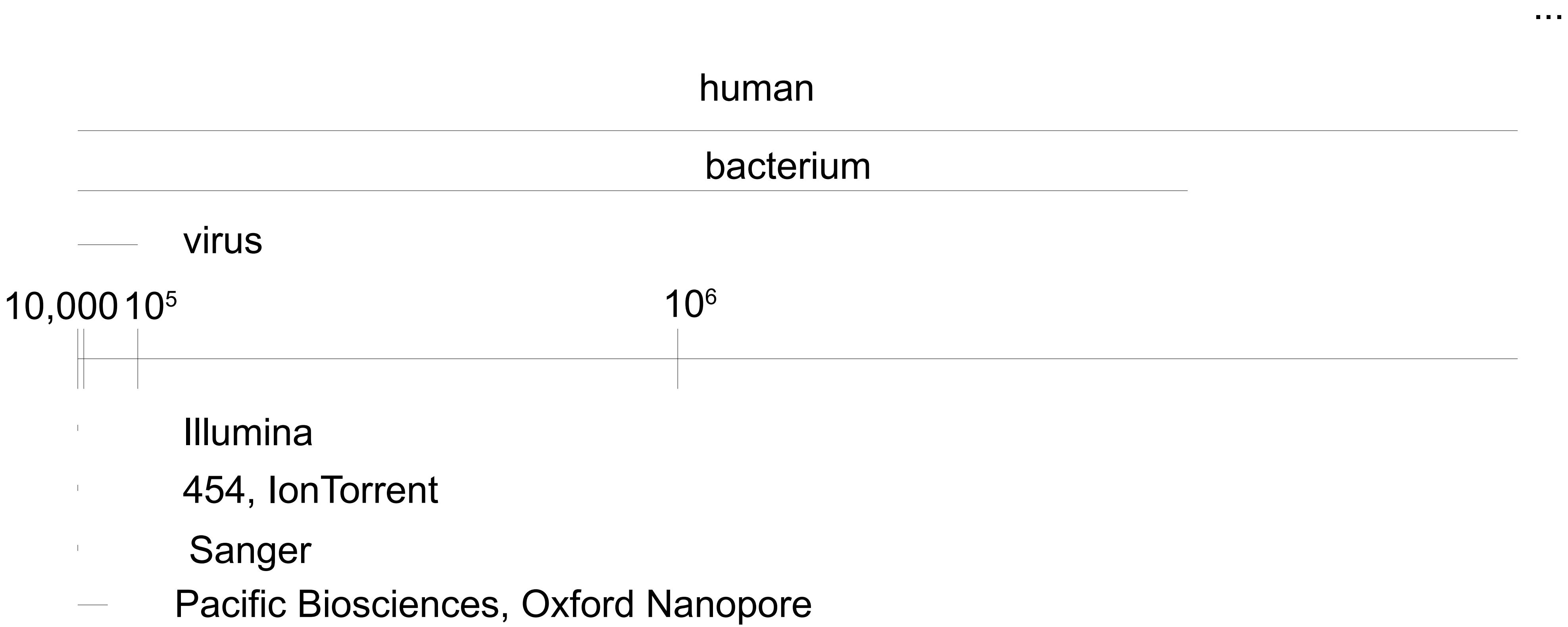
454, IonTorrent

Sanger

...

Pacific Biosciences, Oxford Nanopore

Why shotgun sequencing?



...

human

bacterium

virus

10,000 10^5 10^6

Illumina

454, IonTorrent

Sanger

Pacific Biosciences, Oxford Nanopore

Stop and Think!

- Why is it necessary to have multiple copies of the original DNA?

Stop and Think!

- Why is it necessary to have multiple copies of the original DNA?

- If adjacent DNA fragments do not share any information, it is impossible to reconstruct the original order.

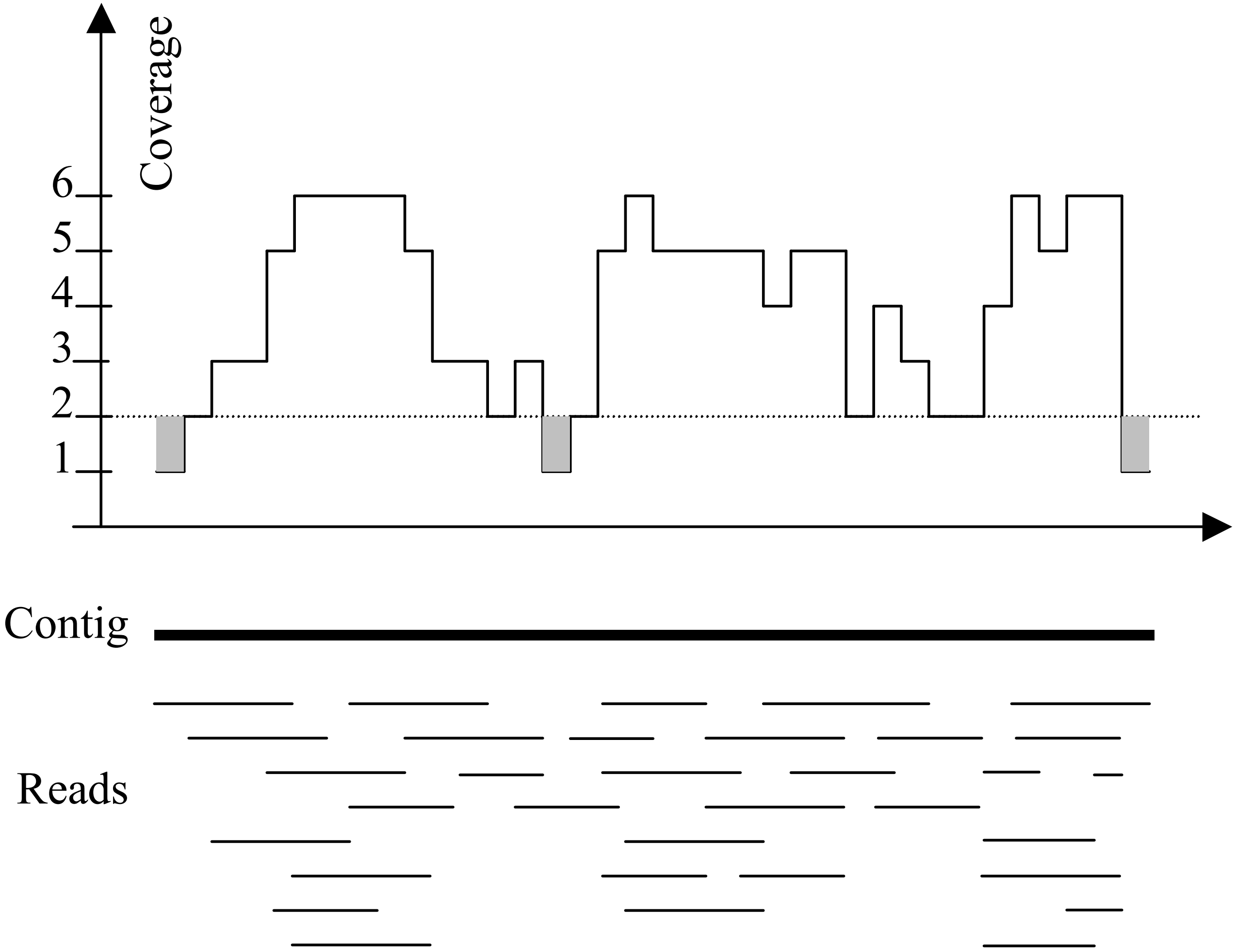
GCAACAT   TTCAGT   CCGCCGT ATCACAG

Is assembly even possible?

- If we randomly sequence will we ever cover every base in the genome?

- How much DNA do we need to sequence to cover every base in the genome?

Impact of randomness – non-uniform coverage

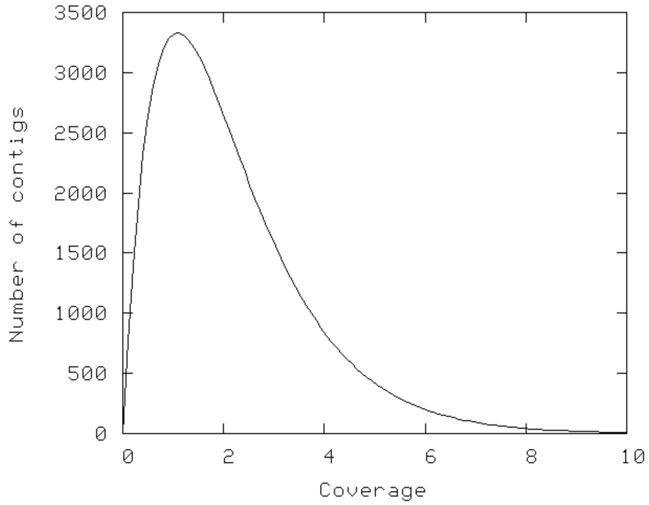


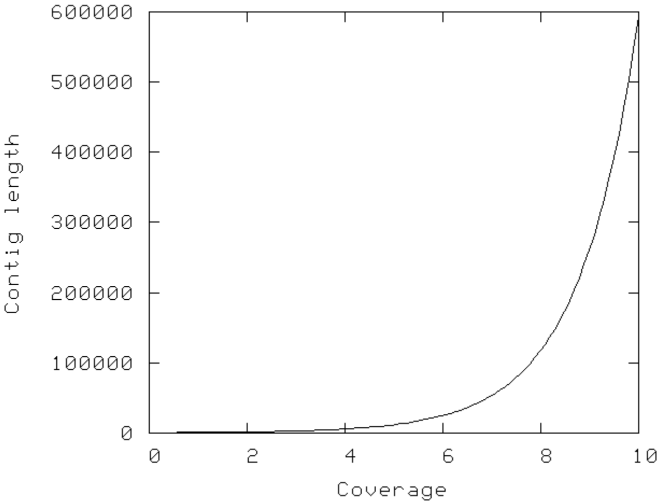
Imagine raindrops on a sidewalk

Lander-Waterman statistics

L = read length
T = minimum overlap
G = genome size
N = number of reads
c = coverage (NL / G)
$\sigma = 1 - T/L$

E(#islands) = $Ne^{-c\sigma}$
E(island size) = $L(e^{c\sigma} - 1) / c + 1 - \sigma$
contig = island with 2 or more reads





Next: assembly algorithms