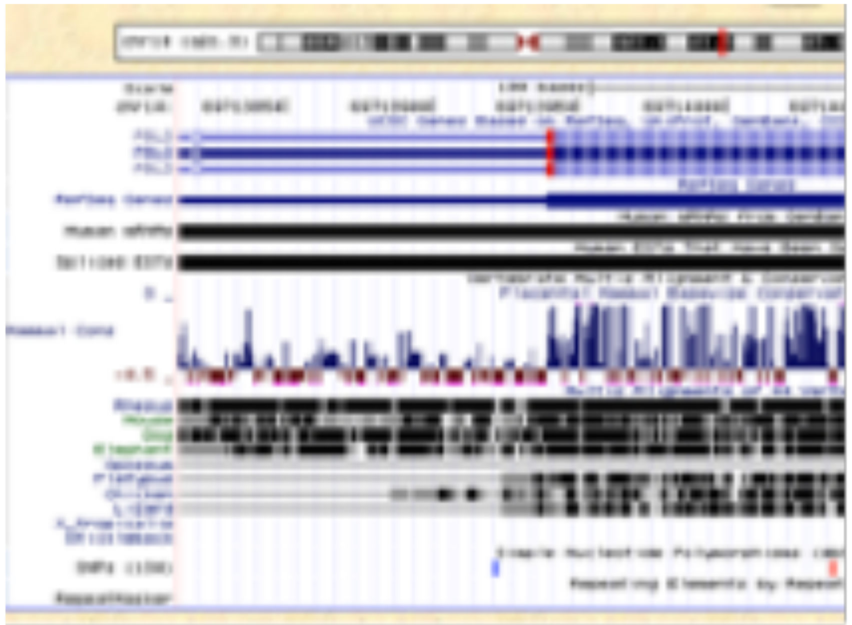
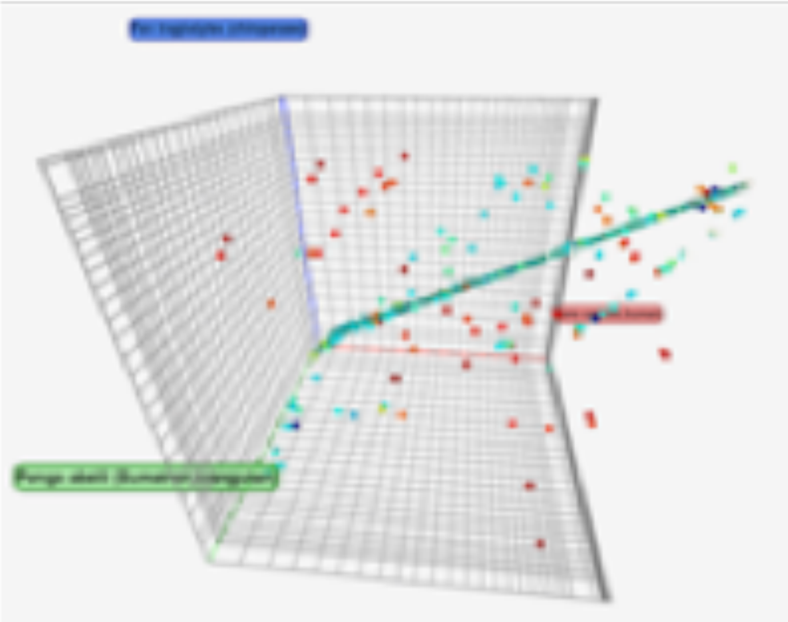
Computational Genomics

The Carbon and Clarke Formula
and
It's Relationship to
Genome Assembly





[27] Selection of Specific Clones from Colony Banks by Suppression or Complementation Tests

By Louise Clarke and John Carbon

METHODS IN ENZYMOLOGY, VOL. 68

ISBN 0-12-181968-X

We have determined the transformant colony bank size needed to obtain a plasmid collection representing 90–99% of the *E. coli* or yeast genome as follows.[2] Given a preparation of cell DNA fragmented to a size such that each fragment represents a fraction (f) of the total genome, the probability (p) that a given unique DNA sequence is present in a collection of N transformant colonies is given by the expression

$$P = 1 - (1-f)^N$$

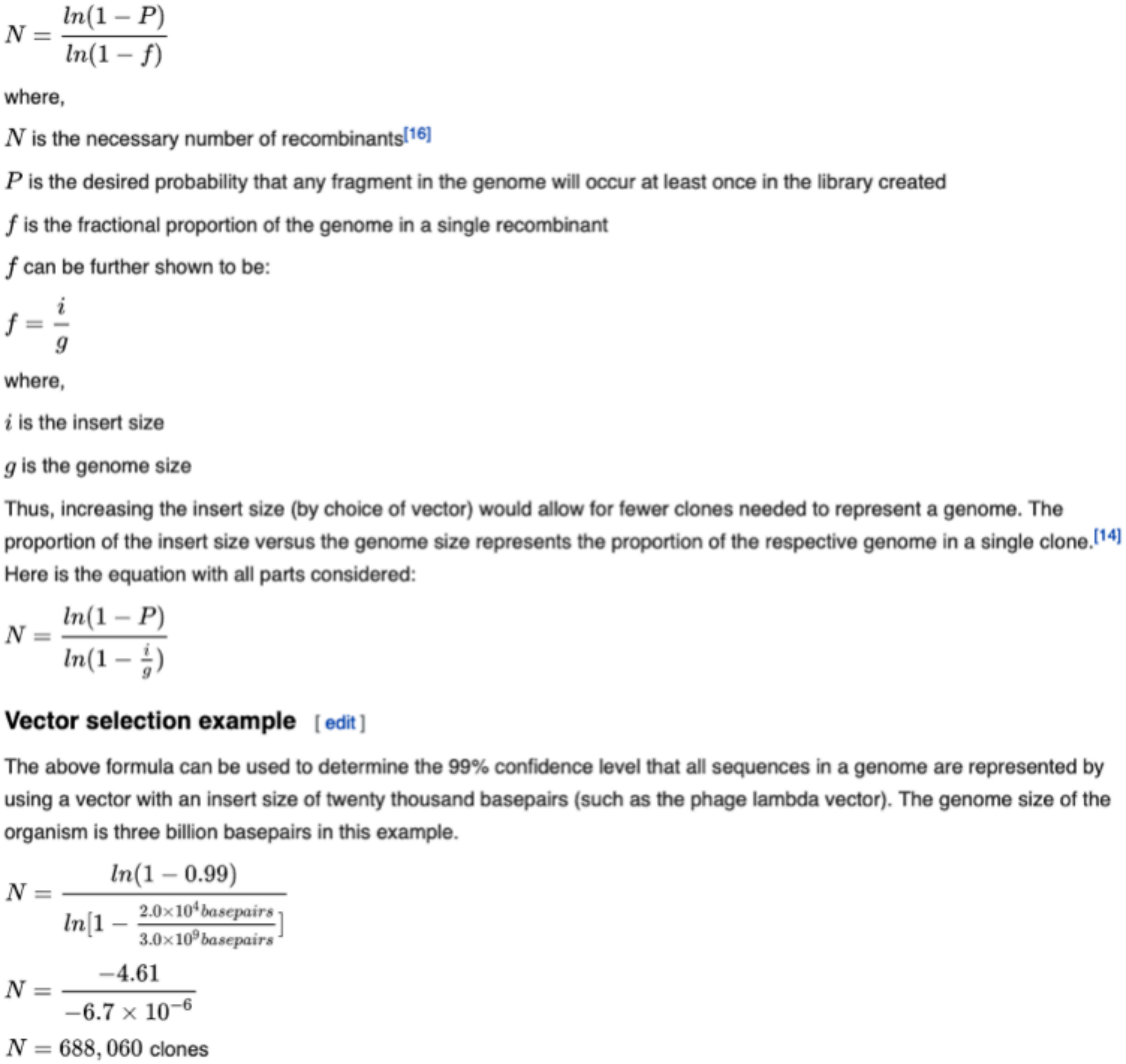
or

$$N = \ln (1\text{-}P)/\ln (1\text{-}f)$$

A sample calculation for *E. coli* (genome size, 2.7 × 10^9 daltons) for $P = 0.99$ is

$$N = \frac{\ln (1\text{-}0.99)}{\ln [1 - (8.5 \times 10^6/2.7 \times 10^9)]} = 1437$$

Thus, using a preparation of DNA randomly sheared to an average size of 8.5 × 10^6 daltons for the construction of annealed hybrid circular DNA, a colony bank of only about 1400 transformants for *E. coli* or 5400 transformants for yeast is adequate to give a probability of 99% that any *E. coli* or yeast gene will be on a hybrid plasmid in one of the clones.



$$N = \frac{ln(1-P)}{ln(1-f)}$$

where,

N is the necessary number of recombinants[16]

P is the desired probability that any fragment in the genome will occur at least once in the library created

f is the fractional proportion of the genome in a single recombinant

f can be further shown to be:

$$f = \frac{i}{g}$$

where,

i is the insert size

g is the genome size

Thus, increasing the insert size (by choice of vector) would allow for fewer clones needed to represent a genome. The proportion of the insert size versus the genome size represents the proportion of the respective genome in a single clone.[14] Here is the equation with all parts considered:

$$N = \frac{ln(1-P)}{ln(1-\frac{i}{g})}$$

Vector selection example [edit]

The above formula can be used to determine the 99% confidence level that all sequences in a genome are represented by using a vector with an insert size of twenty thousand basepairs (such as the phage lambda vector). The genome size of the organism is three billion basepairs in this example.

$$N = \frac{ln(1-0.99)}{ln[1-\frac{2.0\times10^4 basepairs}{3.0\times10^9 basepairs}]}$$

$$N = \frac{-4.61}{-6.7\times10^{-6}}$$

$N = 688,060$ clones

How many clones you need to analyze to have a 0.98 Probability of representation for a given gene when you construct a library that, on average, has inserts that are 4 kbp long from a genome that is 4.5 Mbp long?

Poisson Calculations

The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution

The probability a base is not sequenced is given by:

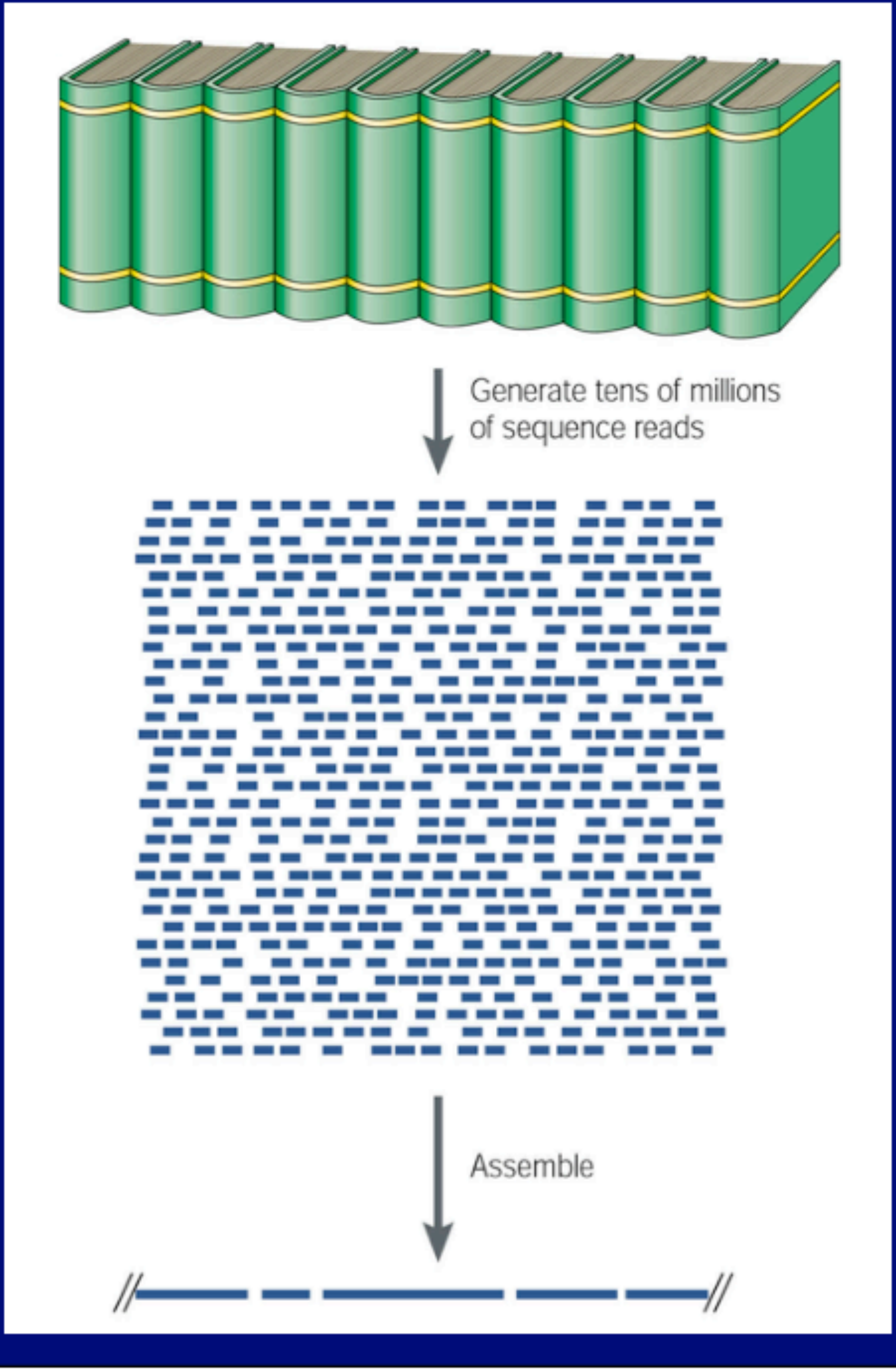
$$P_0=e^{-c}$$

Where:

- c = fold sequence coverage (c=LN/G),
- LN = # bases sequenced, i.e. L = average sequencing read length and N = # reads
- G = target sequence length
- e = 2.718 (e=2.718281828459)

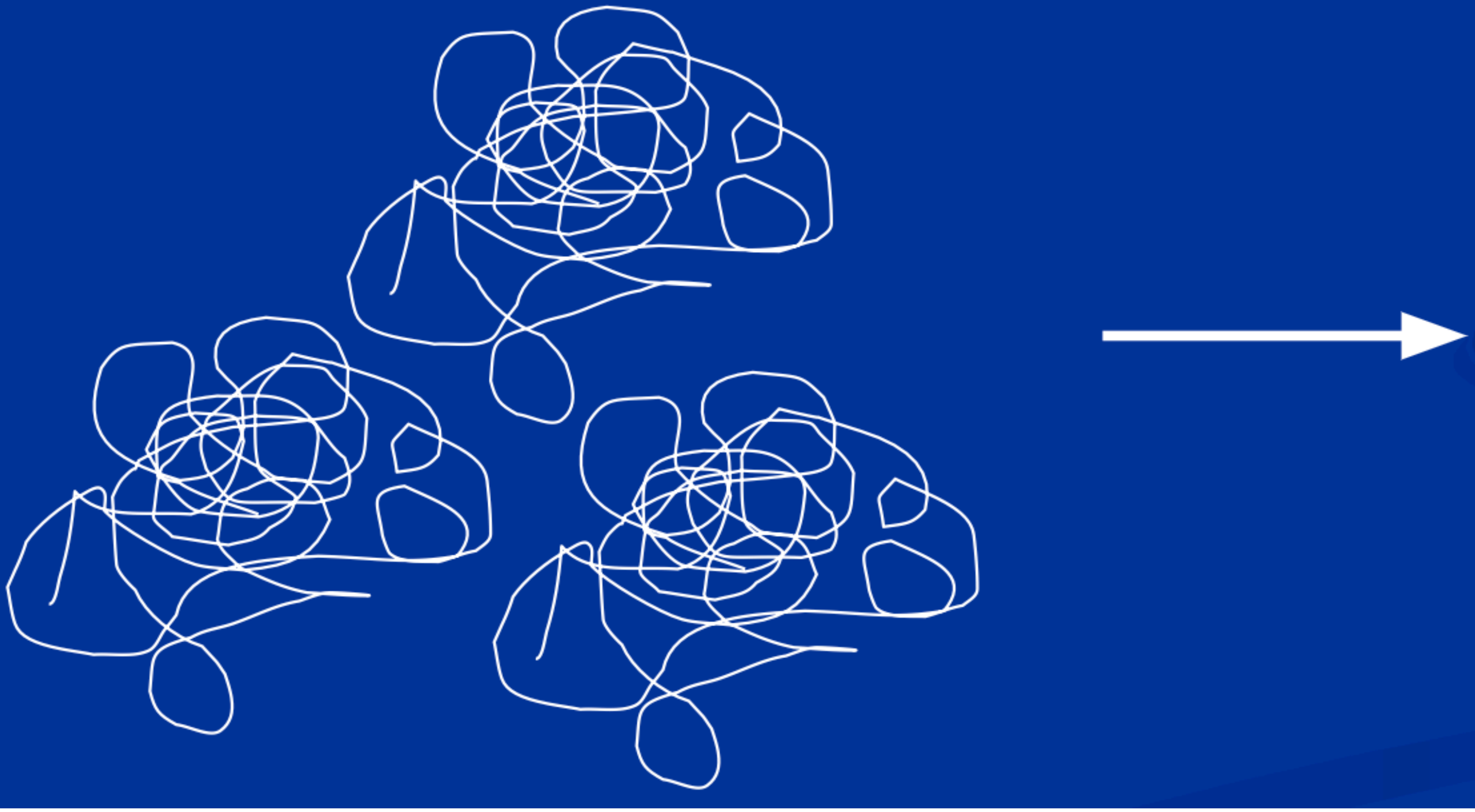
Fold Coverage	$P_0=e^{-c}$	% not sequenced	% sequenced
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

Whole-Genome Shotgun Sequencing



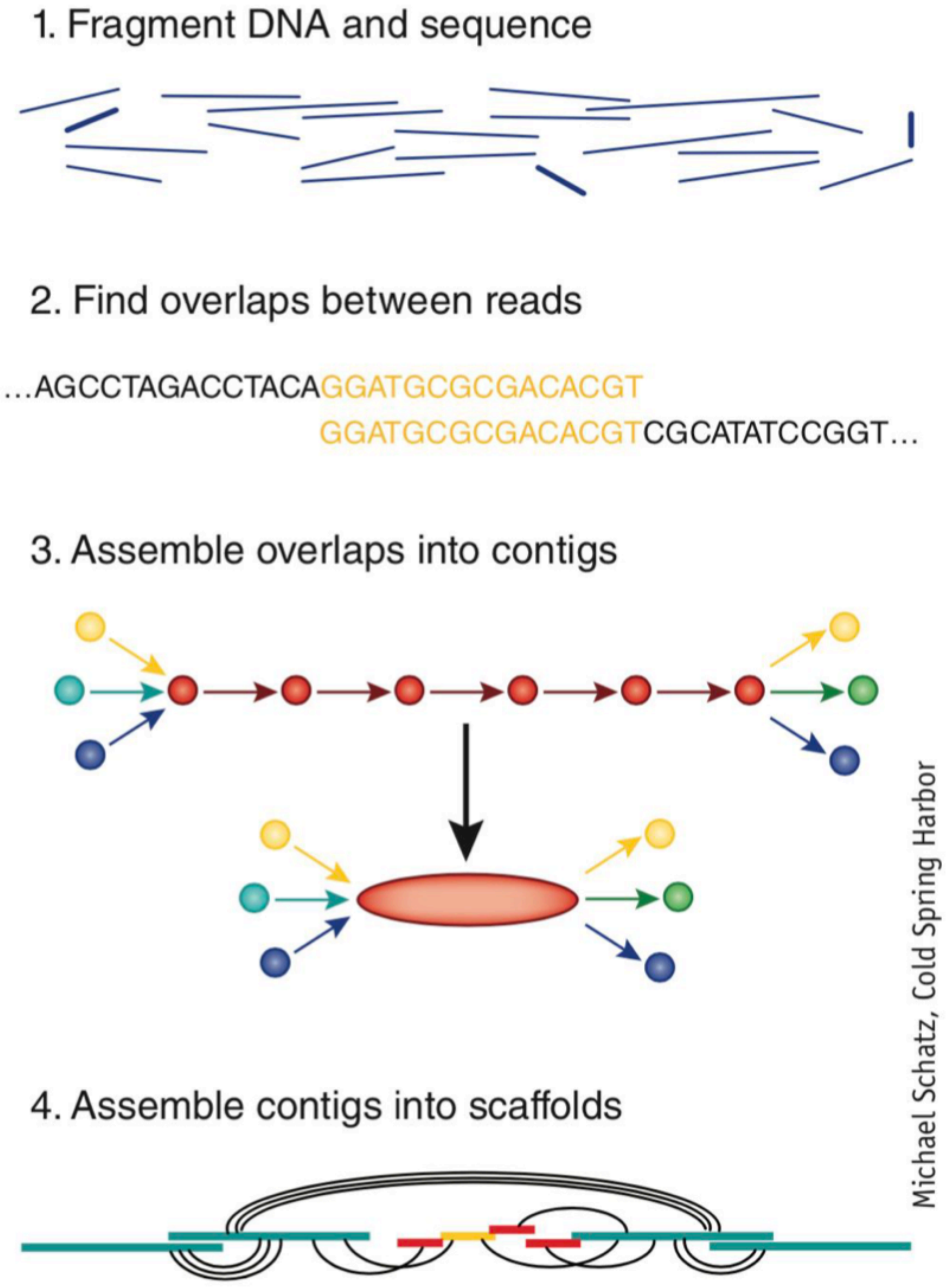
WGS sequencing

Multiple copies of DNA



Fragments of DNA

No information is retained on which part of the DNA the fragments came from



1. Fragment DNA and sequence

2. Find overlaps between reads

...AGCCTAGACCTACAGGATGCGCGACACGT

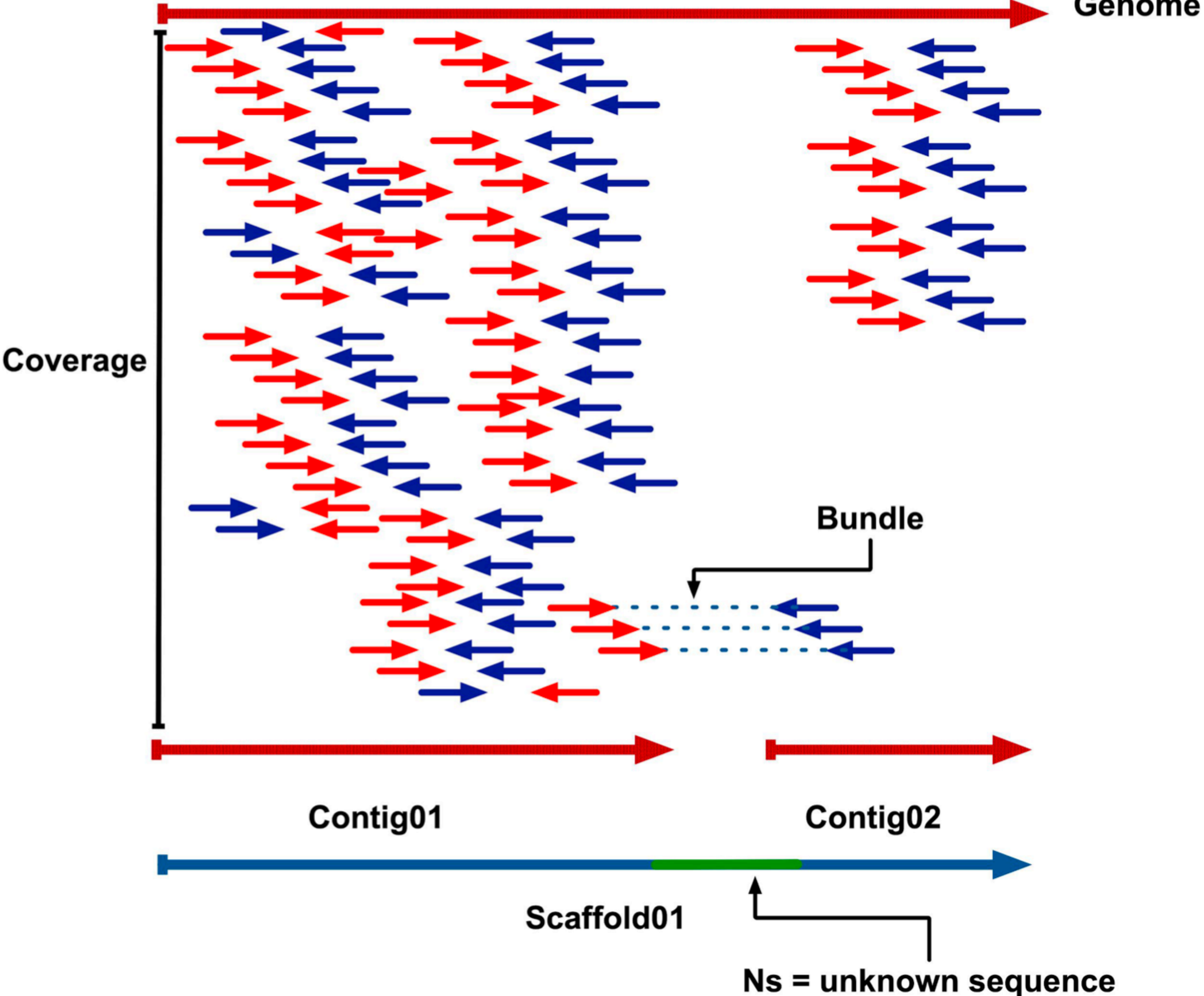
GGATGCGCGACACGTCGCATATCCGGT...

3. Assemble overlaps into contigs

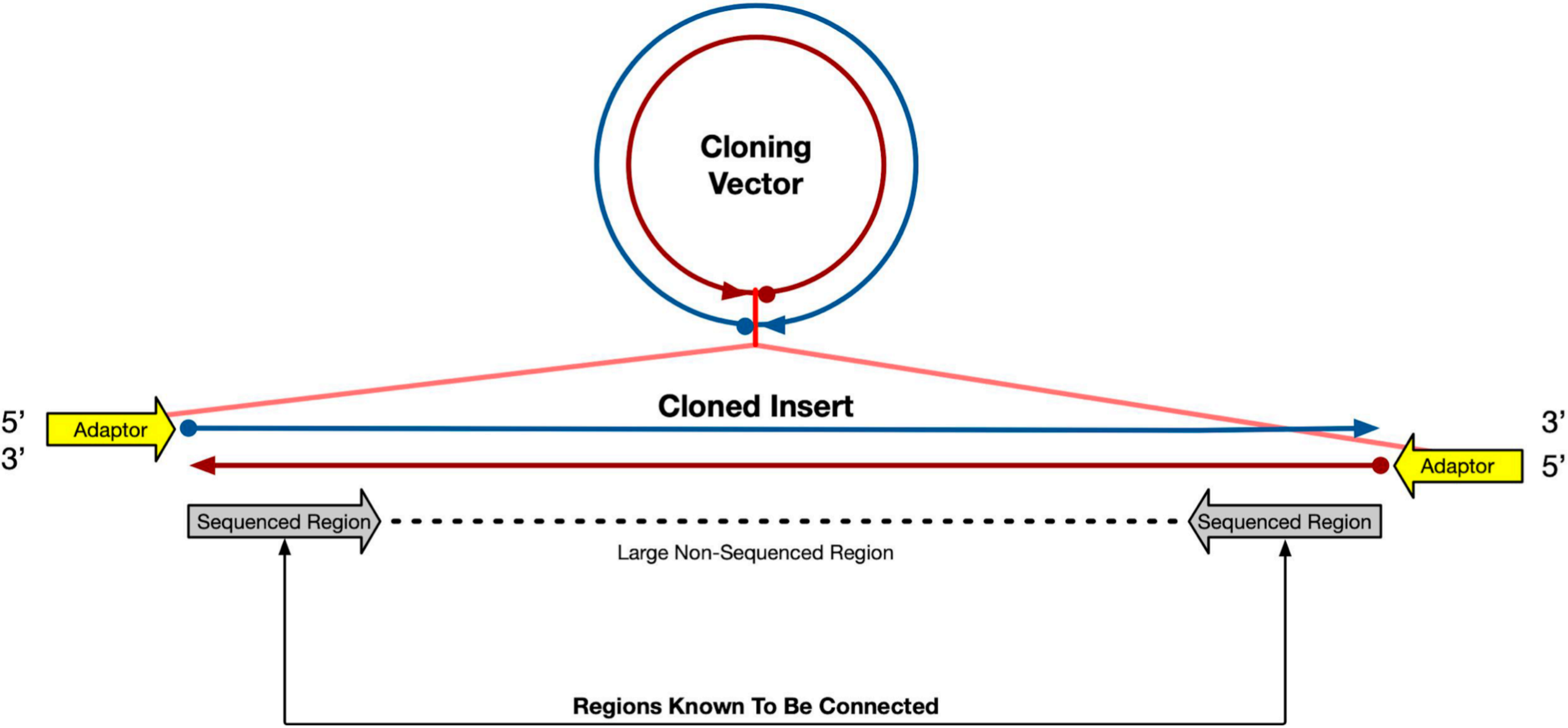
4. Assemble contigs into scaffolds

Michael Schatz, Cold Spring Harbor

Genome assembly stitches together a genome from short sequenced pieces of DNA.


Genome
Coverage
Bundle
Contig01
Contig02
Scaffold01
Ns = unknown sequence

Long-Cloned-DNA Fragment



Cloning Vector

Cloned Insert

5'

3'

Adaptor

3'

5'

Adaptor

Sequenced Region

Sequenced Region

Large Non-Sequenced Region

Regions Known To Be Connected

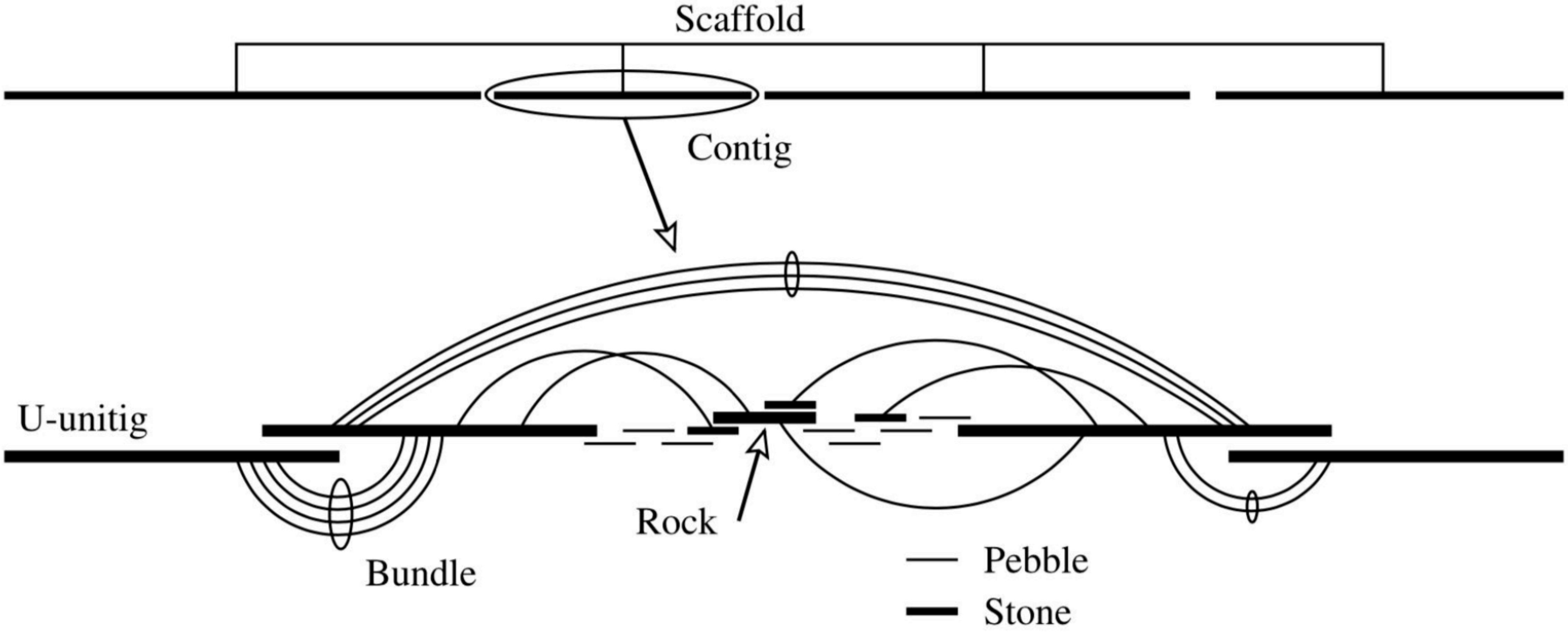
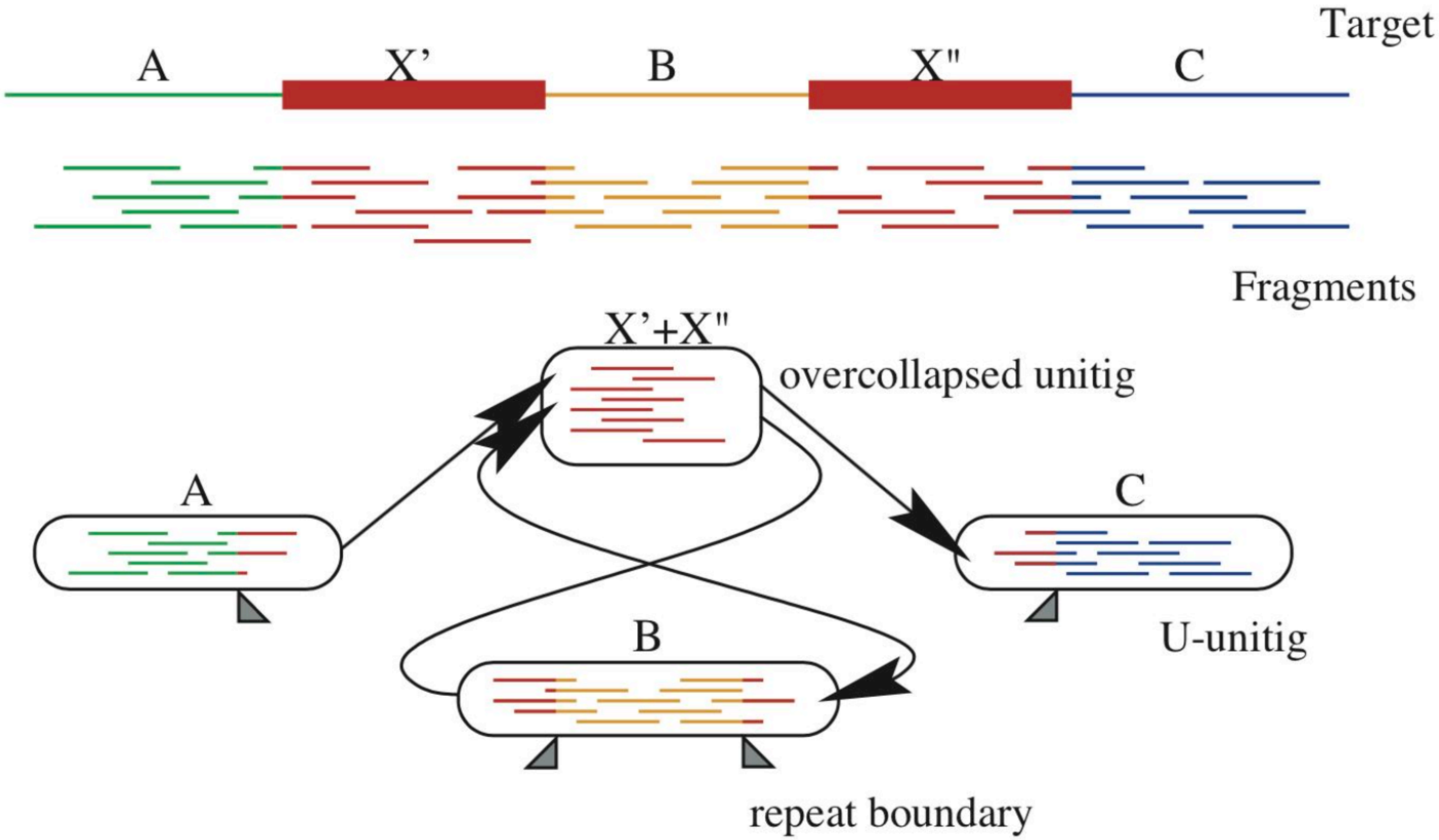


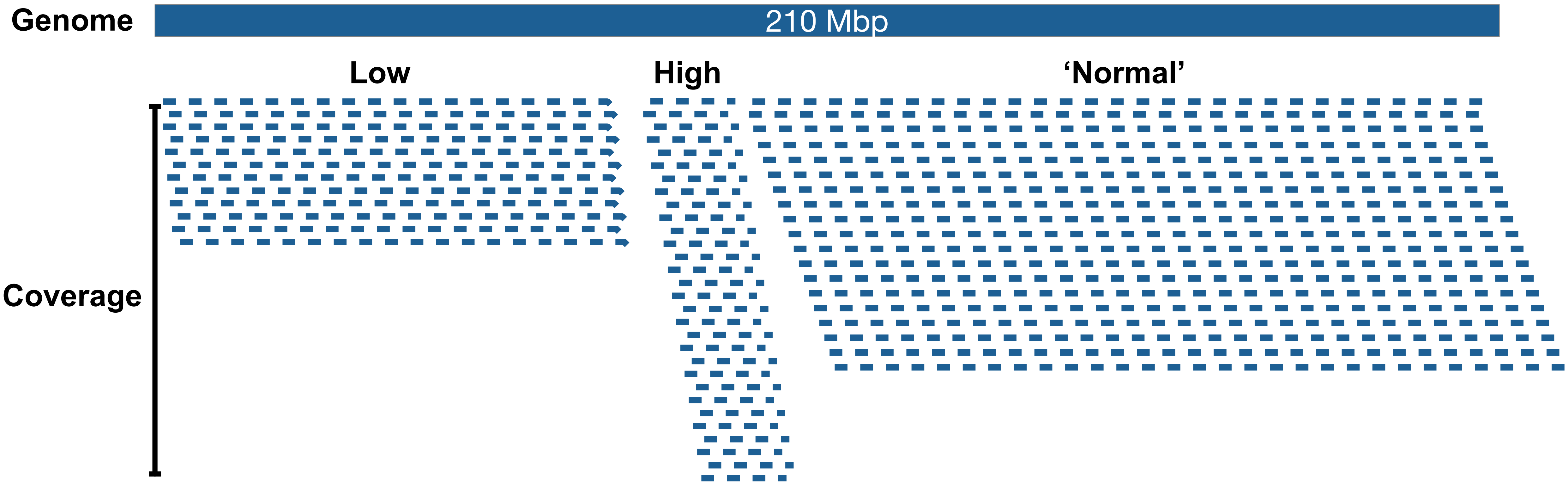
Fig. 4. Anatomy of a scaffold. A scaffold is a collection of ordered contigs with approximately known distances between them. Our contigs are built from U-unitigs that form a scaffold via bundles and then have a series of rocks, stones, and pebbles filled into the gaps between them (where possible).

A Whole-Genome Assembly of Drosophila Eugene W. Myers, et al. Science 287, 2196 (2000); DOI: 10.1126/science.287.5461.2196

Fig. 3. Unitigs and repeat boundaries. Consider the hypothetical genome consisting of three unique stretches *A*, *B*, and *C* with two nearly identical, interspersed copies, *X′* and *X″*, of a repeat element *X*. This results in the four unitigs and overlaps shown. As explained in the text the unitig *X′* + *X″* is overcollapsed, and the U-unitigs for regions *A*, *B*, and *C* have repeat boundaries indicating the tail portions that project into *X*.

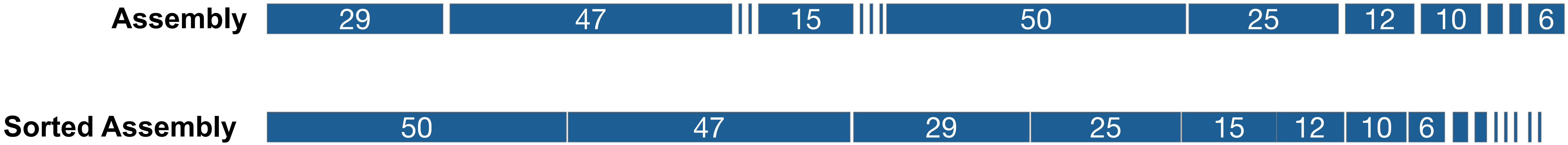


A Whole-Genome Assembly of Drosophila Eugene W. Myers, et al. Science 287, 2196 (2000); DOI: 10.1126/science.287.5461.2196



Coverage = N x L/G

Number of Reads (N)
Average Read Length (L)
Genome Length (G)



Genome Length: 210	N50 Genome Length: 210/2 = 105	GN50 = 29
Assembly Length: ~194	N50 Assembly Length: 194/2 = 97	N50 = 47

GN50 or N50: Reach the number that is making up 50% of your total Genome/Assembly length