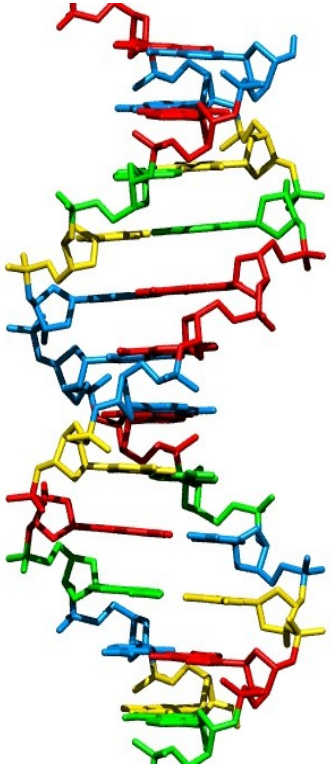
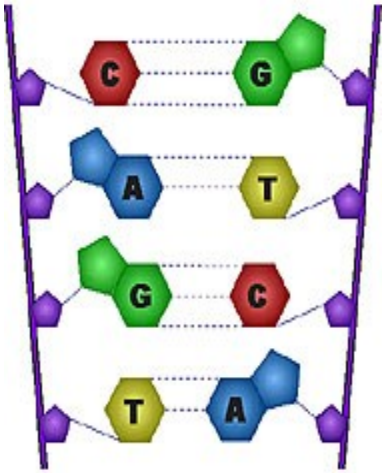
DNA Sequencing

DNA sequencing

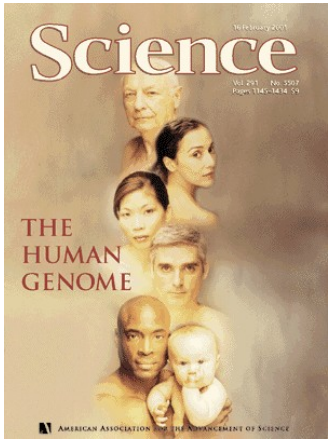
How we obtain the sequence of nucleotides of a species





…ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT…

Human Genome Project



1990: Start

2000: Bill Clinton: *"most important scientific discovery in the 20th century"*

2001: Draft

2003: Finished

3 billion basepairs

$3 billion

now what?

Which representative of the species?

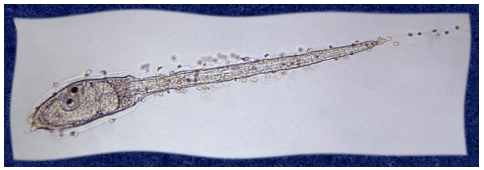
Which human?

Answer one:

Answer two: it doesn't matter

Polymorphism rate: number of letter changes between two different members of a species

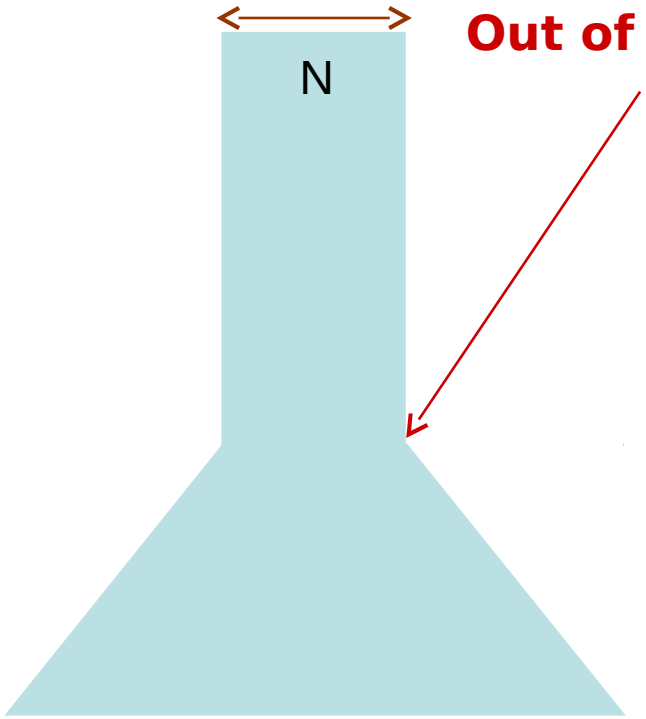
Humans: ~1/1,000

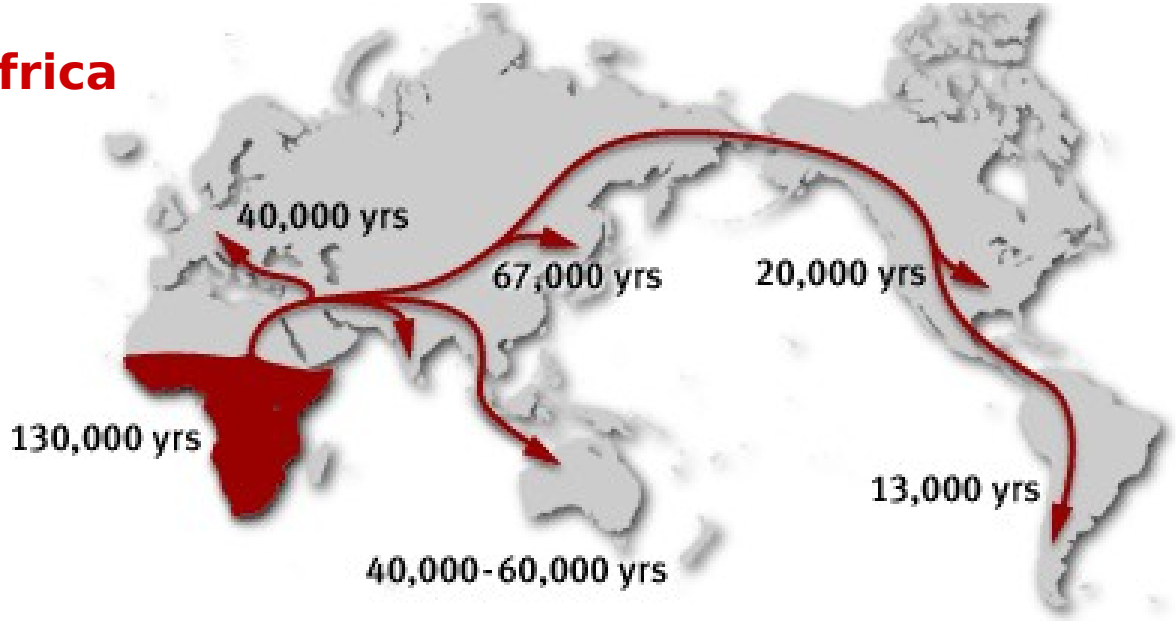


Other organisms have much higher polymorphism rates

- Population size!

Why humans are so similar





Heterozygosity: H

$H = 4Nu/(1 + 4Nu)$

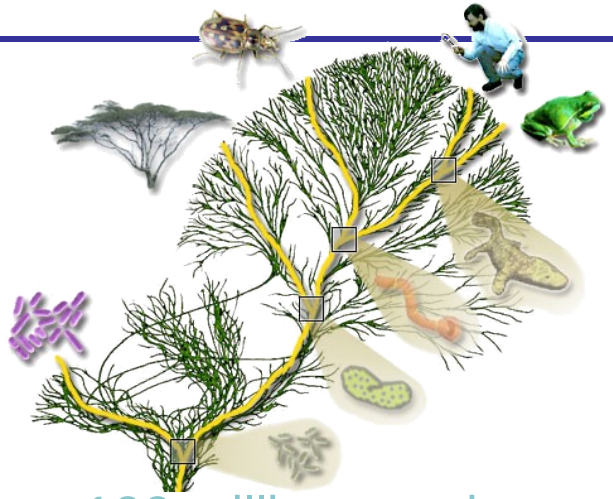
$u \sim 10^{-8}$, $N \sim 10^{4}$

$\Rightarrow H \sim 4\times10^{-4}$

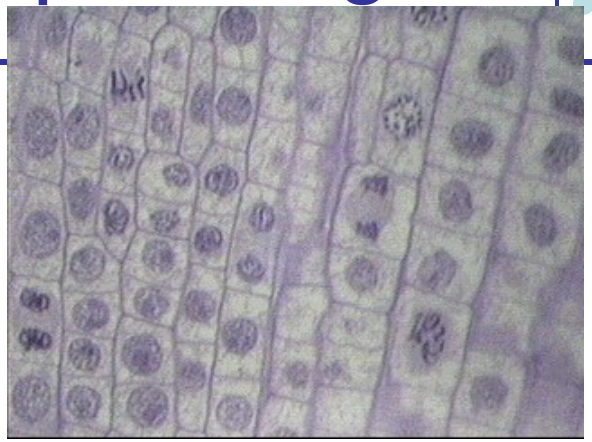
A small population that interbred reduced the genetic variation

Out of Africa ~ 40,000 years ago

There is never “enough” sequencing

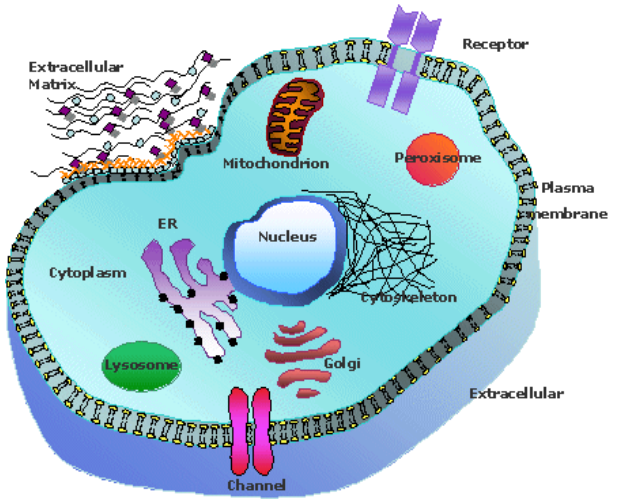


100 million species



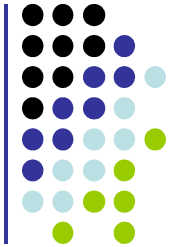
Somatic mutations
(e.g., HIV, cancer)

7 billion individuals



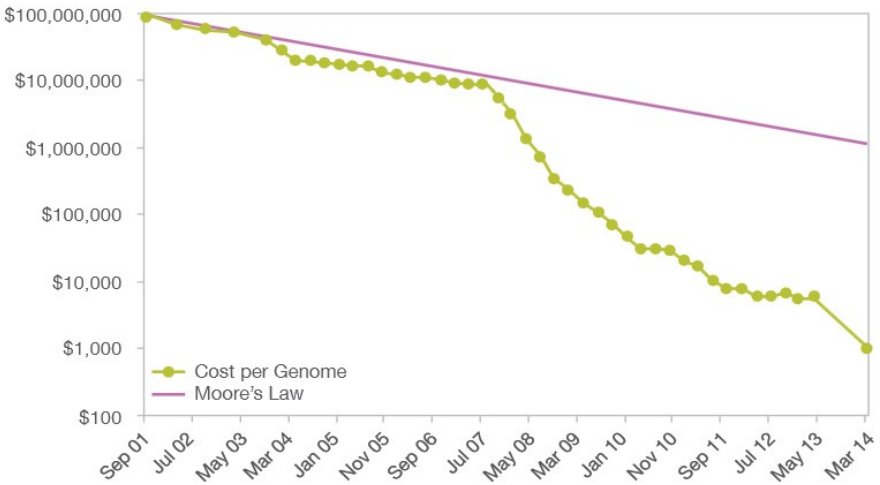
Sequencing is a
functional assay

Sequencing Growth



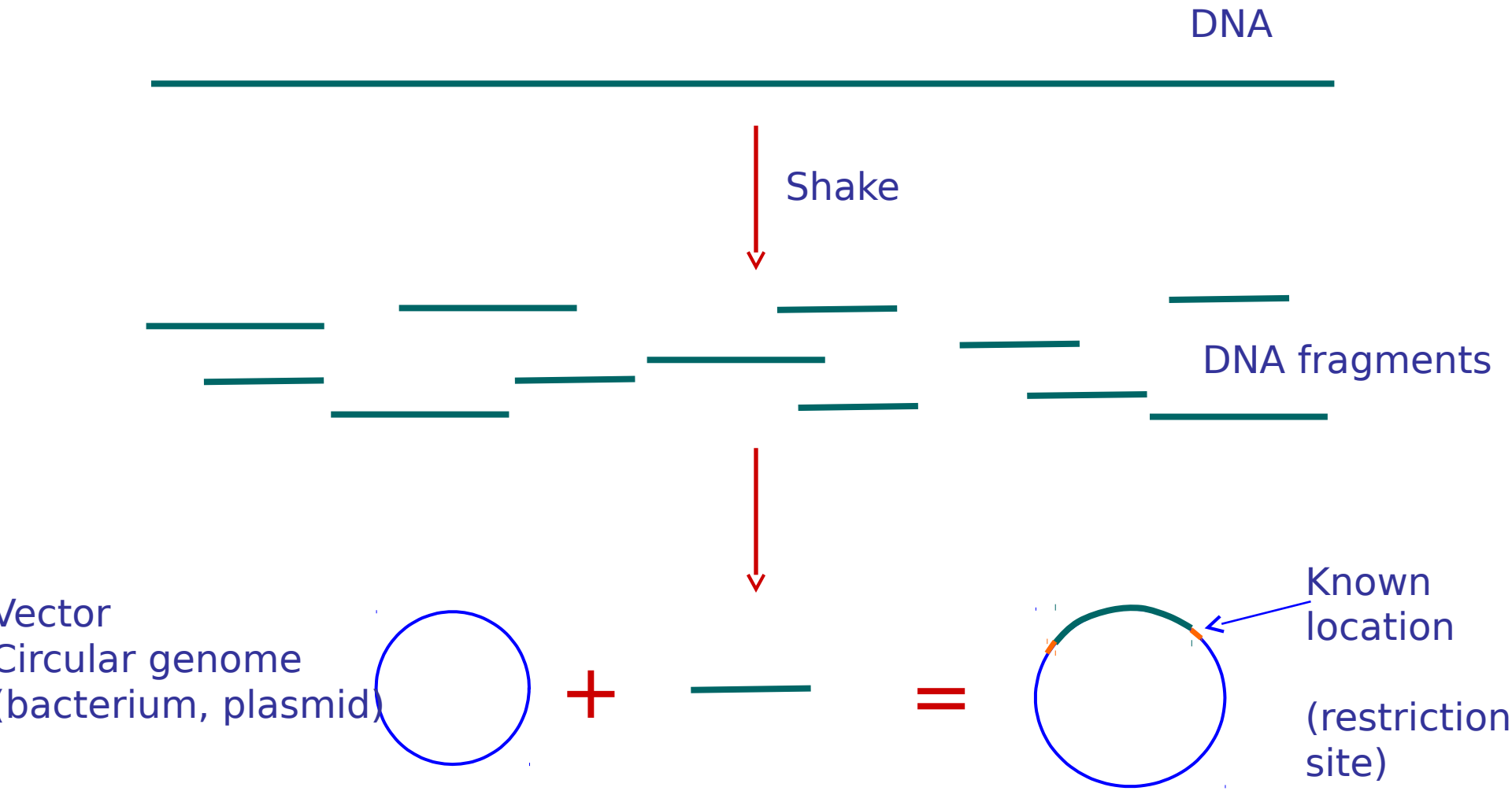
Cost of one human genome

- 2004: $30,000,000
- 2008: $100,000
- 2010: $10,000
- **2014: "$1,000" (???)**
- ???: $300

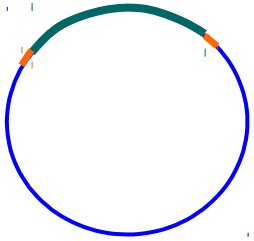


How much would you pay for a smartphone?

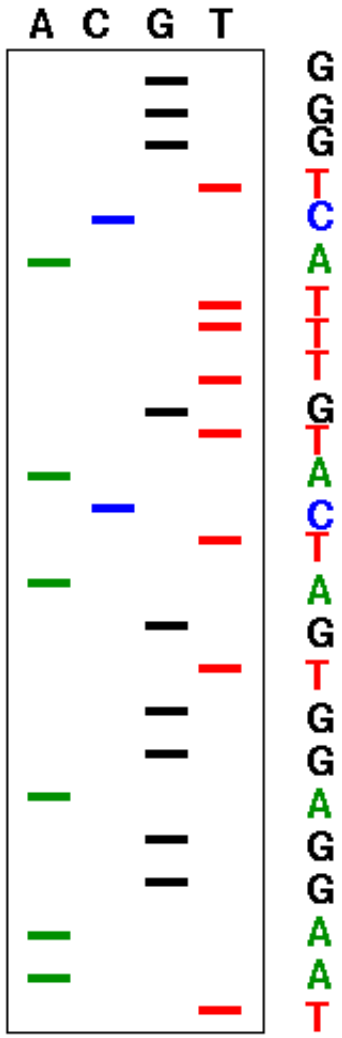
Ancient sequencing technology – Sanger Vectors

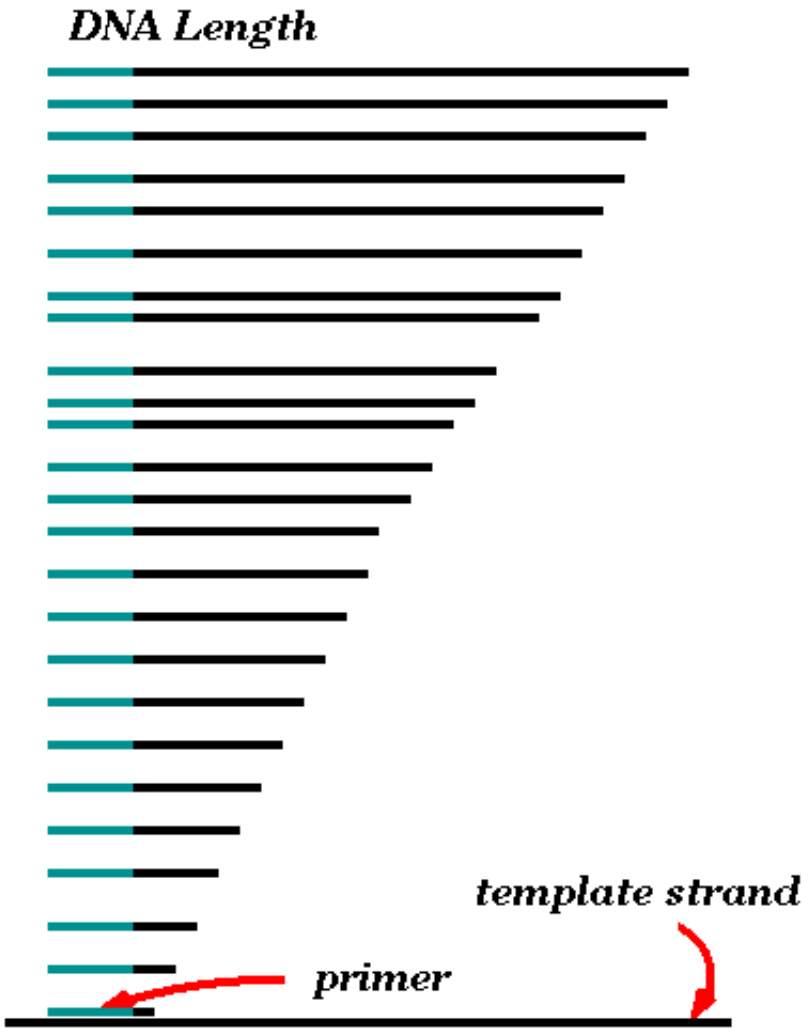


Ancient sequencing technology – Sanger Gel Electrophoresis



1. Start at primer (restriction site)
2. Grow DNA chain
3. Include dideoxynucleoside (modified a, c, g, t)
4. Stops reaction at all possible points
5. Separate products with length, using gel electrophoresis





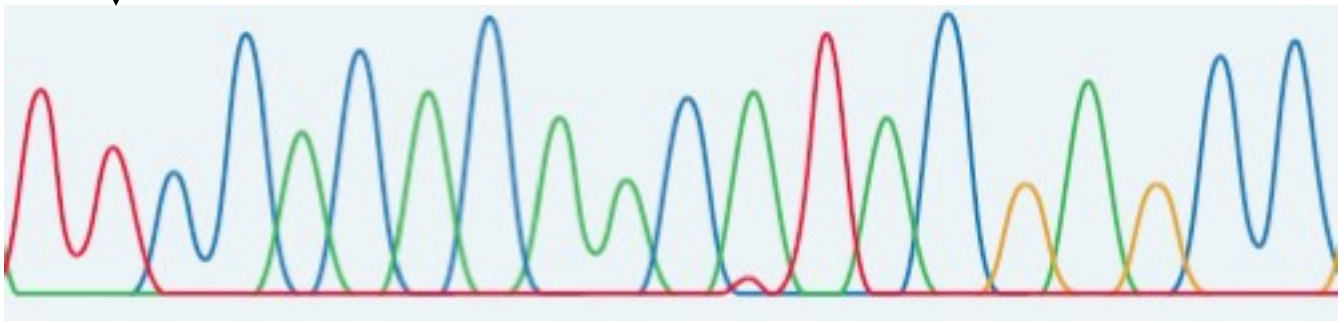
Fluorescent Sanger sequencing trace

Lane signal

(Real fluorescent signals from a lane/capillary are much uglier than this).

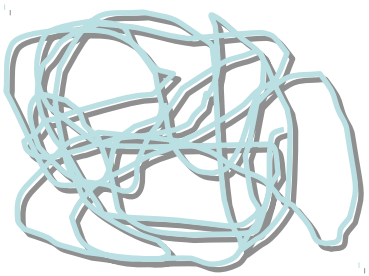
A bunch of magic to boost signal/noise, correct for dye-effects, mobility differences, etc, generates the 'final' trace (for each capillary of the run)

Trace

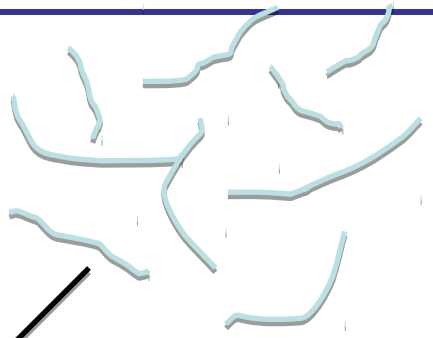


Slide Credit: Arend Sidow

Making a Library (present)



shear to ~500 bases



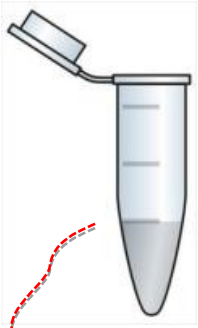
put on linkers

eventual forward and reverse sequence

Left handle: amplification, sequencing

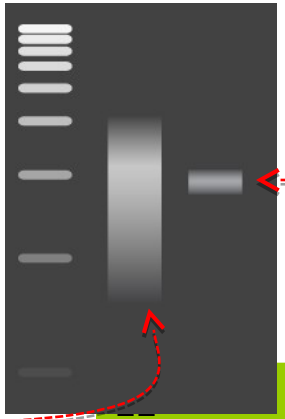
"Insert"

Right handle: amplification, sequencing



PCR to obtain preparative quantities

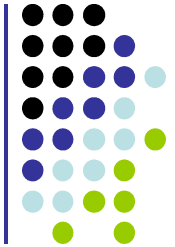
size selection on preparative gel



Final library (~600 bp incl linkers) after size selection

Slide Credit: Arend Sidow

Library

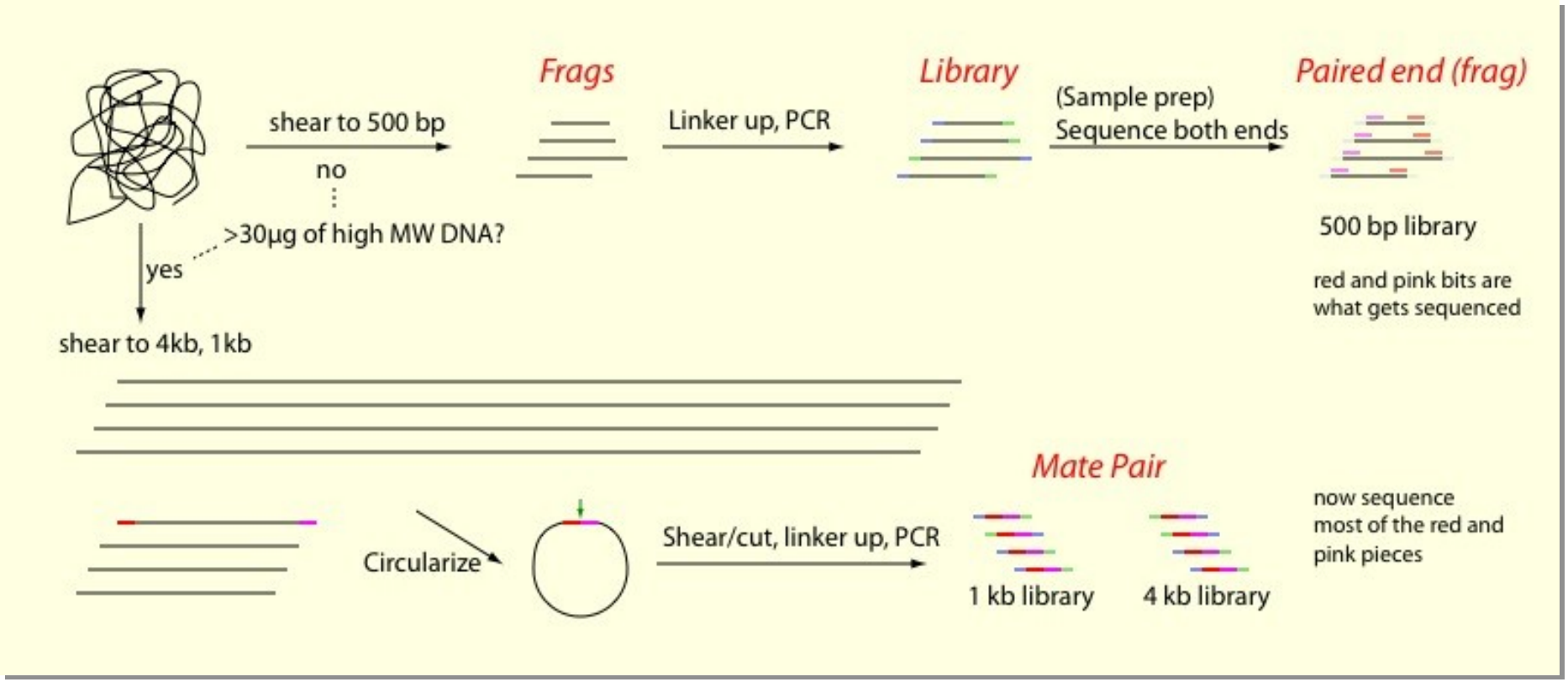


- Library is a massively complex mix of -initially- individual, unique fragments

- Library amplification mildly amplifies each fragment to retain the complexity of the mix while obtaining preparative amounts
  - (how many-fold do 10 cycles of PCR amplify the sample?)

Slide Credit: Arend Sidow

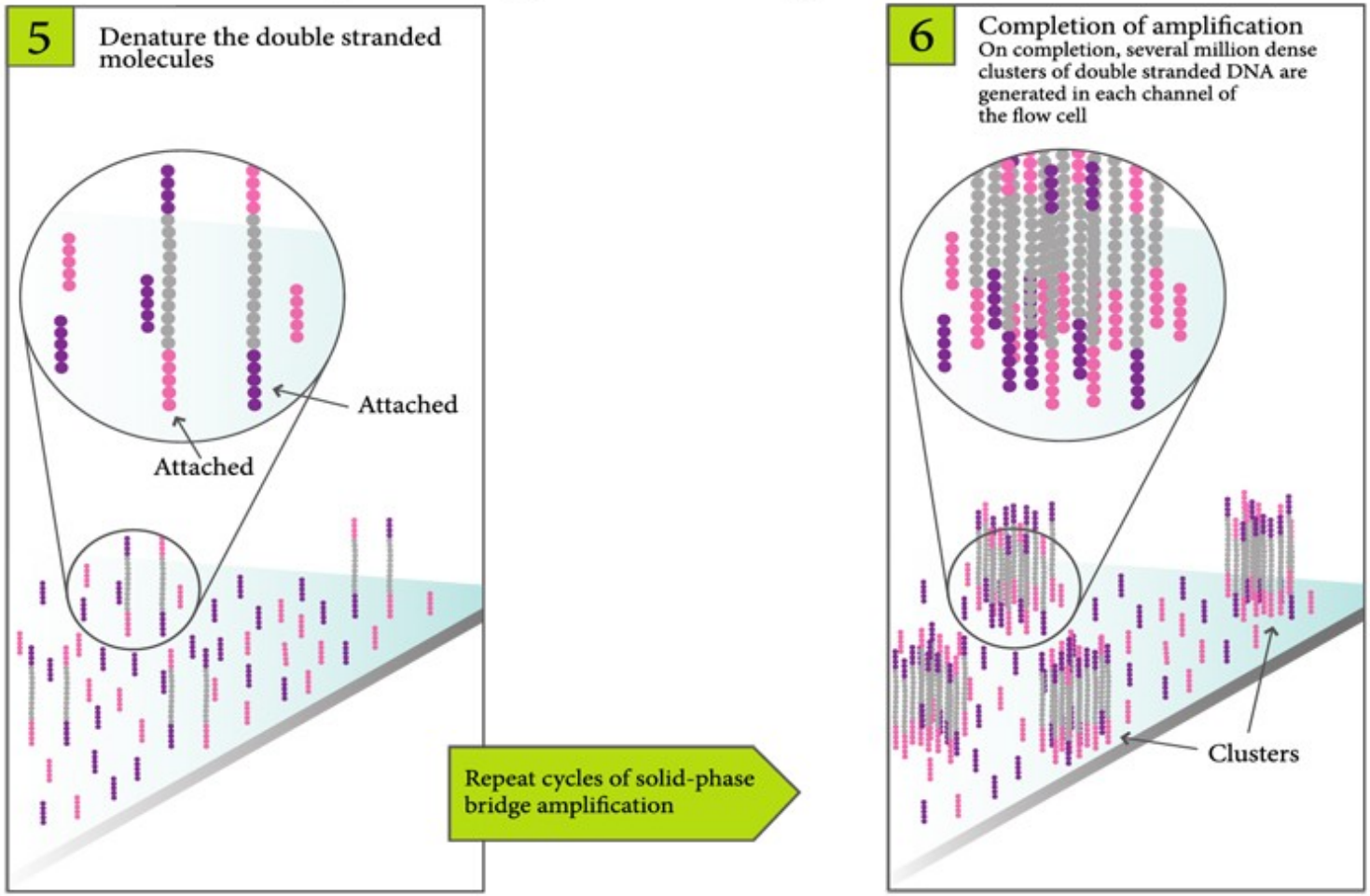
Fragment vs Mate pair (‘jumping’)



(Illumina has new kits/methods with which mate pair libraries can be built with less material)

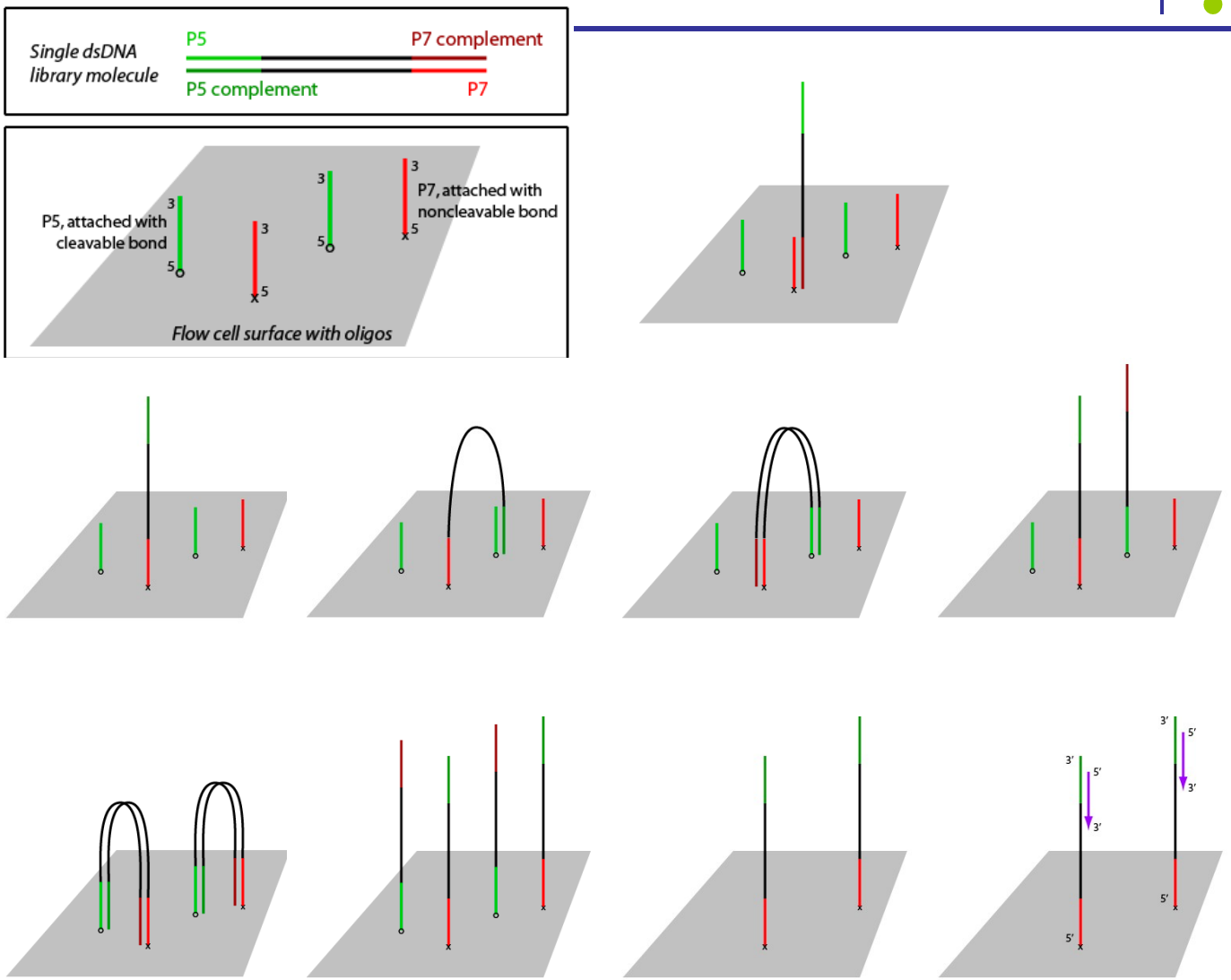
Slide Credit: Arend Sidow

Illumina cluster concept



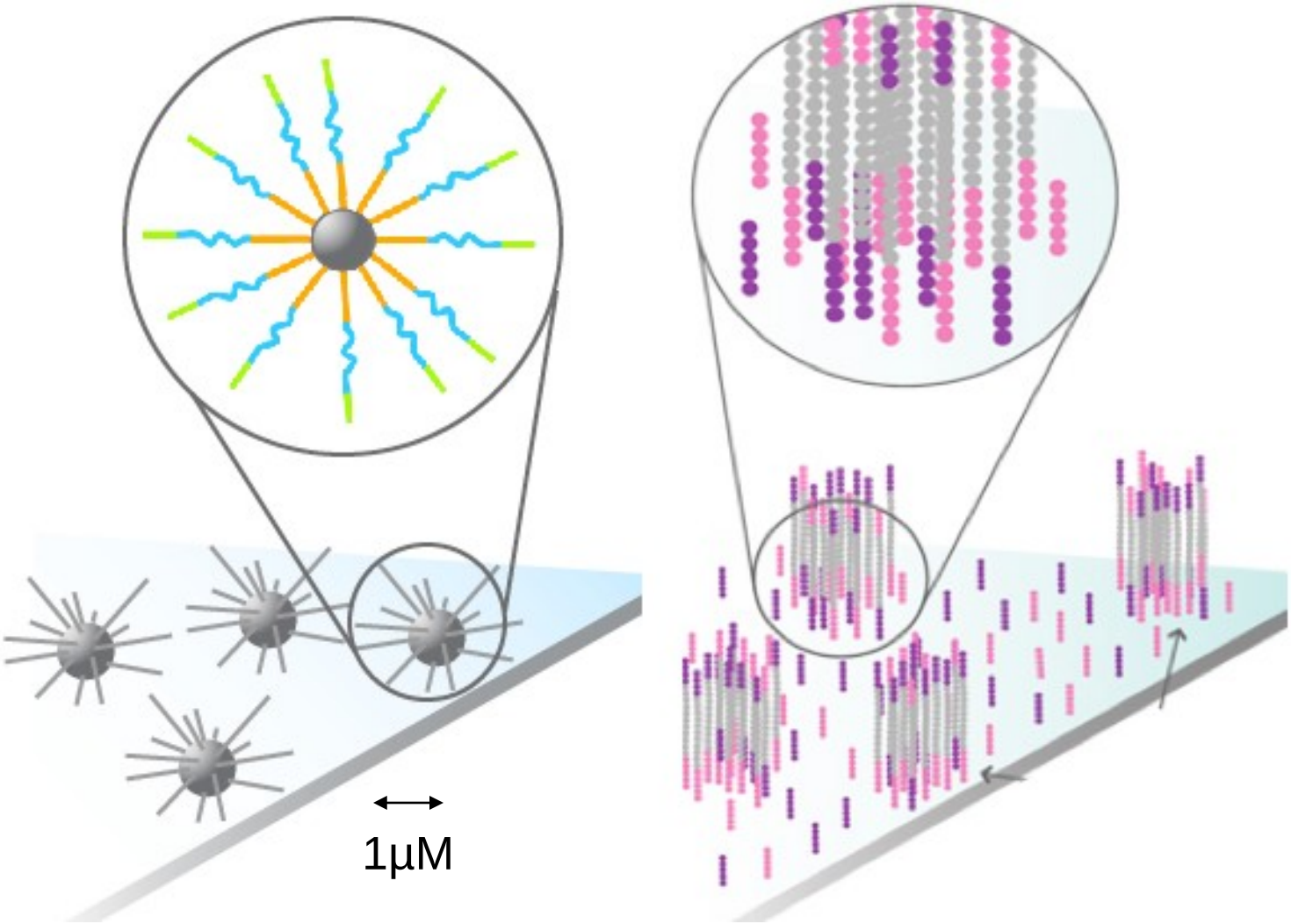
Slide Credit: Arend Sidow

Cluster generation ('bridge amplification')



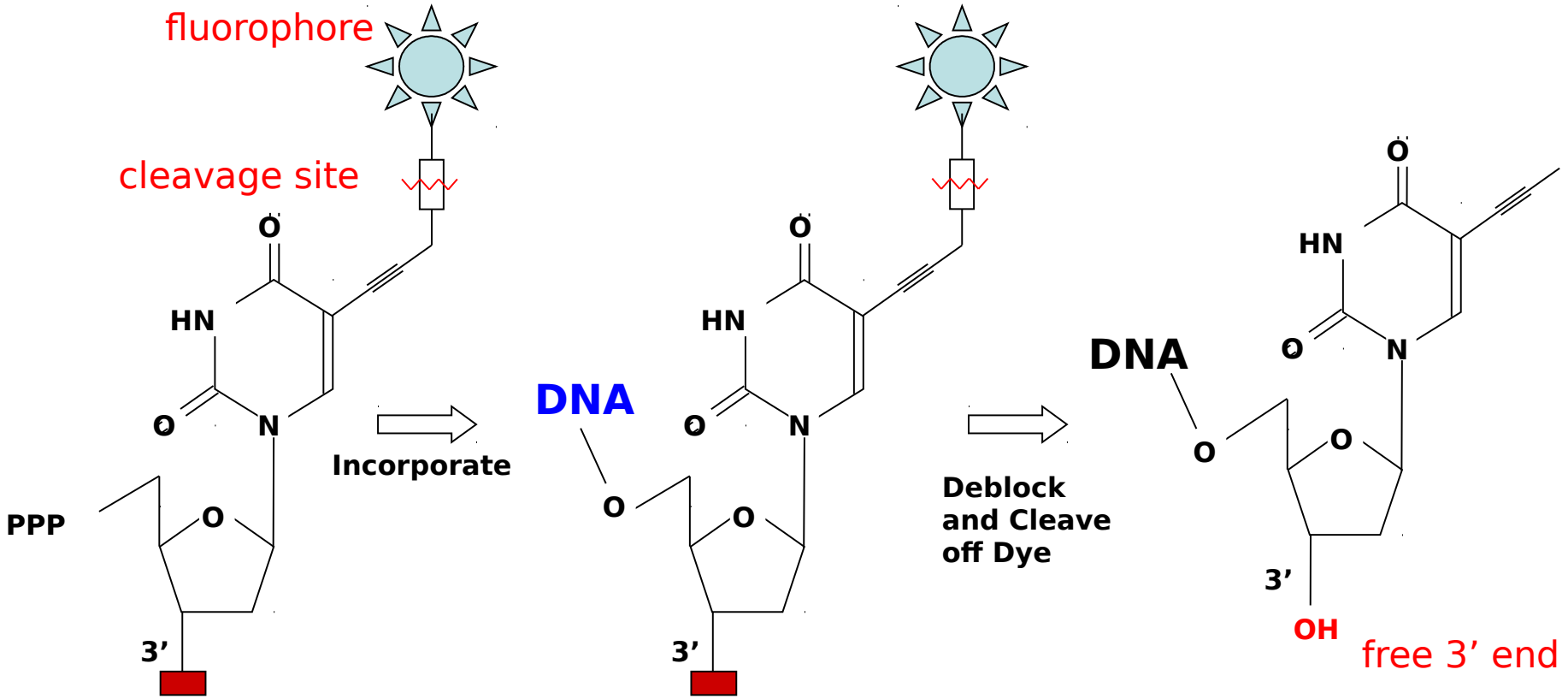
Slide Credit: Arend Sidow

Clonally Amplified Molecules on Flow Cell



Slide Credit: Arend Sidow

Illumina Sequencing: Reversible Terminators



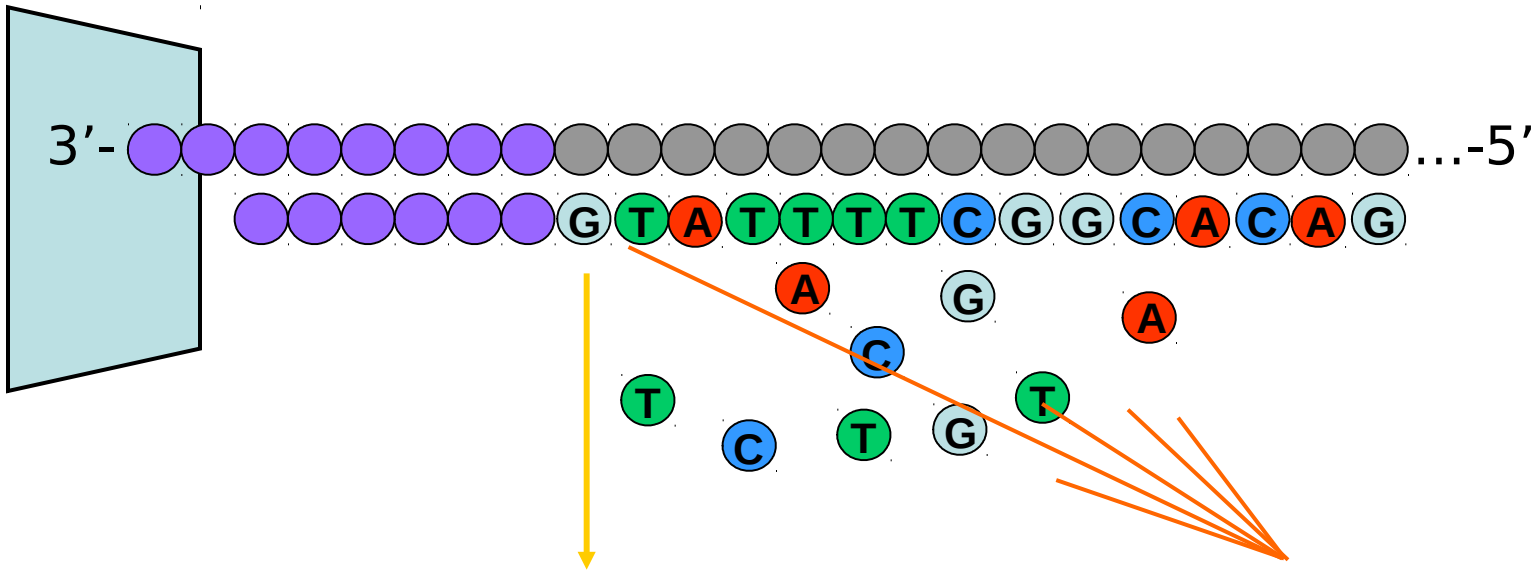
3' OH is blocked

Detection

Ready for Next Cycle

Slide Credit: Arend Sidow

Sequencing by Synthesis, One Base at a Time



Cycle 1: Add sequencing reagents

First base incorporated

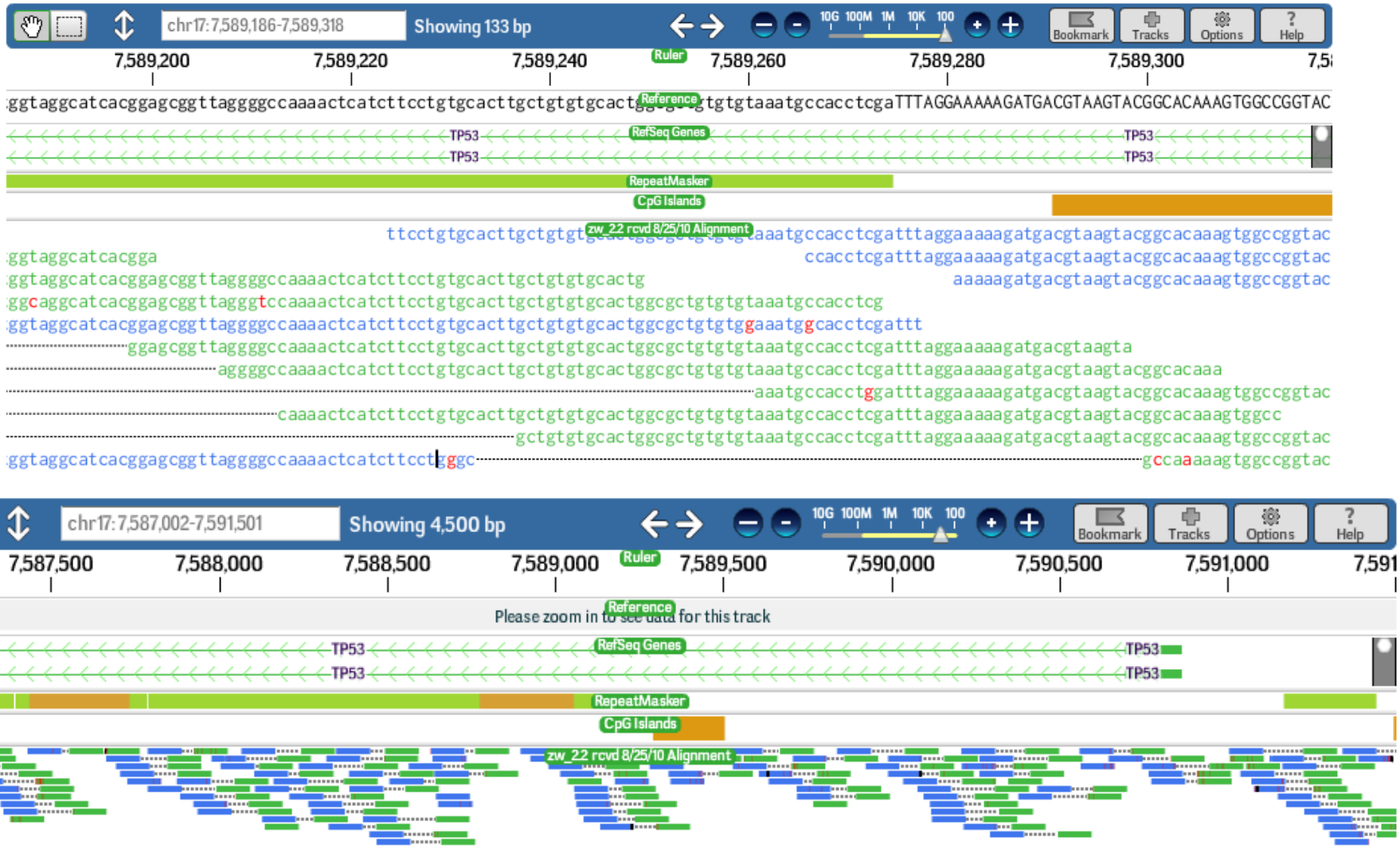
Remove unincorporated bases

Detect signal

Cycle 2-n: Add sequencing reagents and repeat

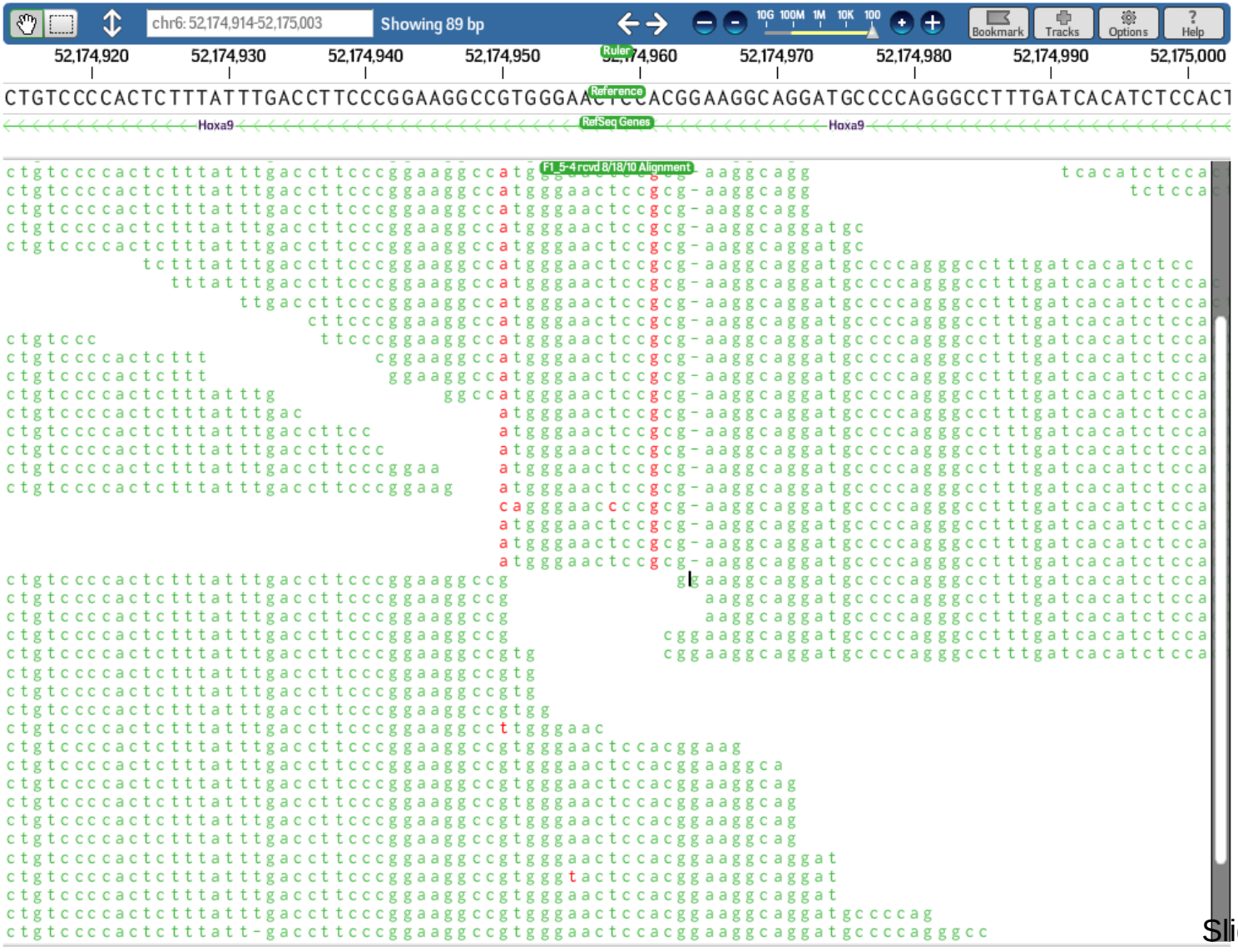
Slide Credit: Arend Sidow

Read Mapping



Slide Credit: Arend Sidow

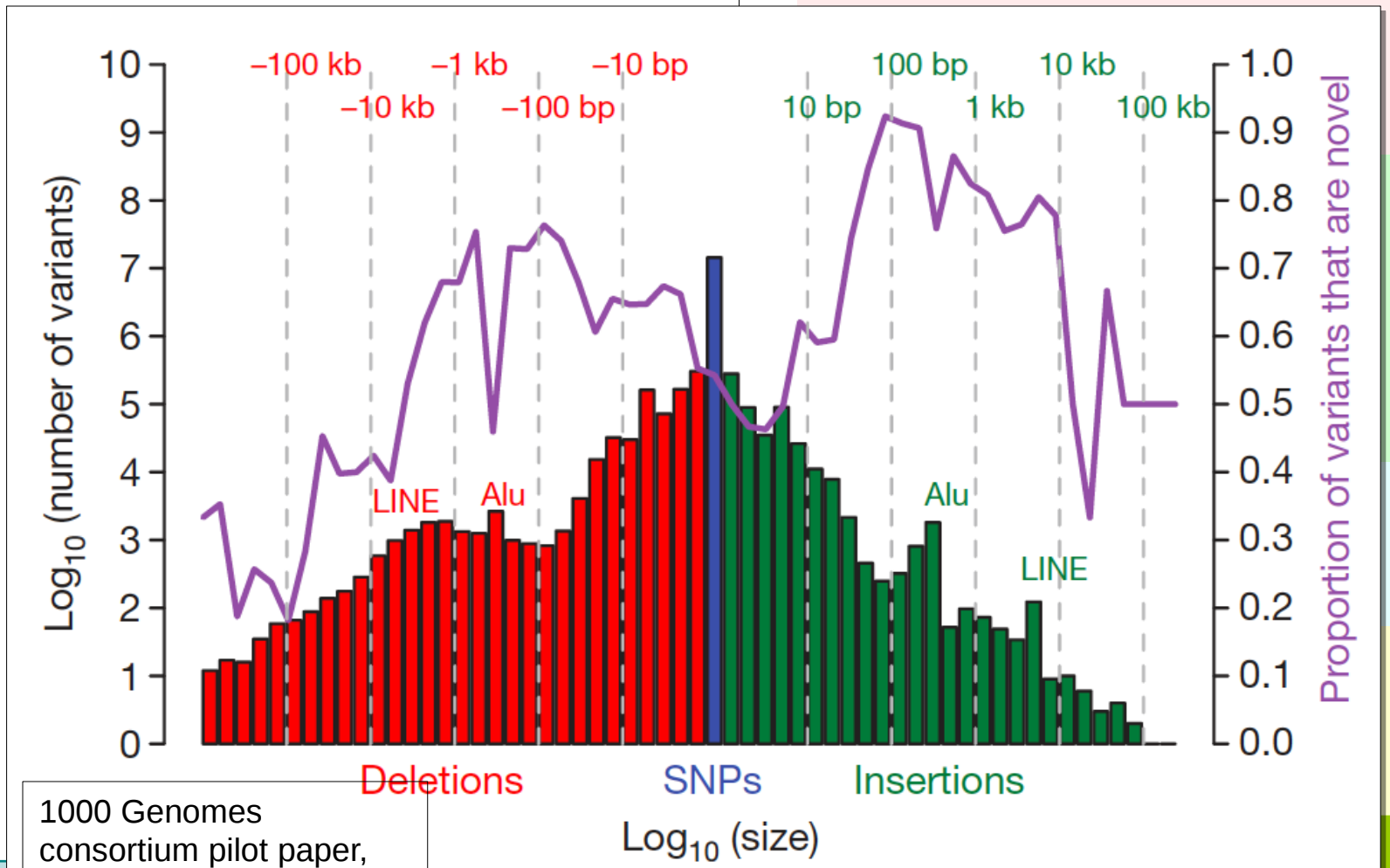
Variation Discovery



Slide Credit: Arend Sidow

Amount of variation – types of lesions

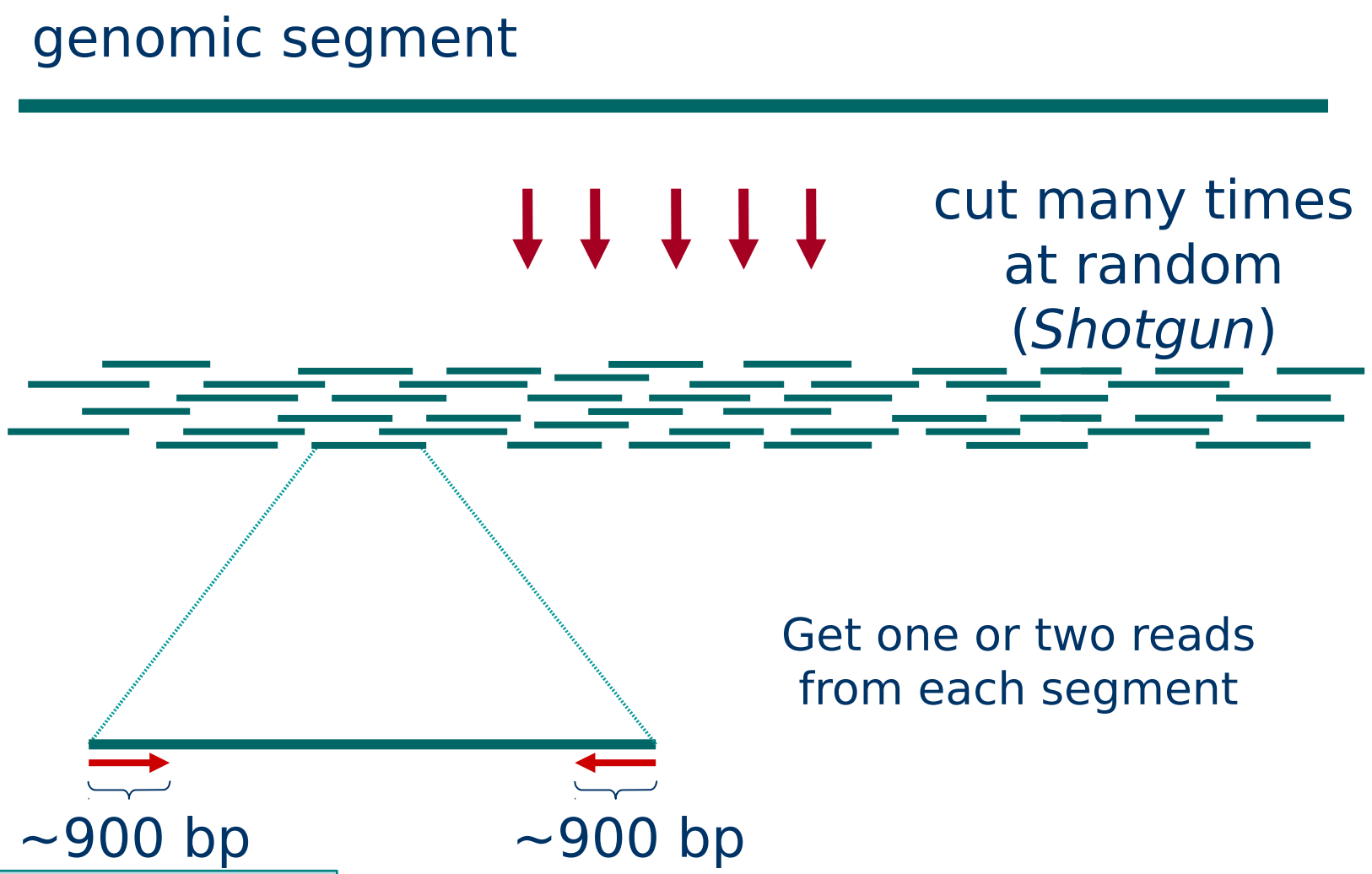
Mutation Types



1000 Genomes consortium pilot paper, Nature, 2010

Slide Credit: Arend Sidow

Method to sequence longer regions

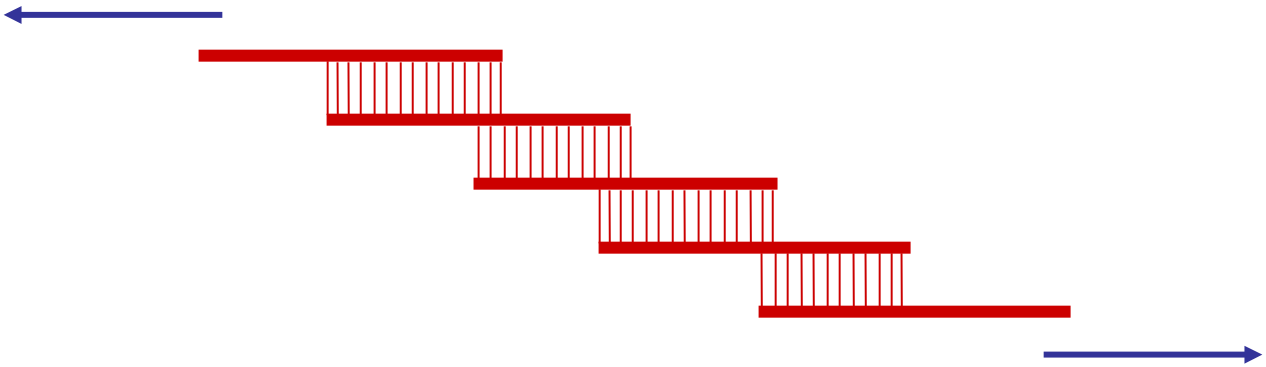


Two main assembly problems

- De Novo Assembly

- Resequencing

Reconstructing the Sequence (De Novo Assembly)



Cover region with high redundancy

Overlap & extend reads to reconstruct the original genomic region

Definition of Coverage

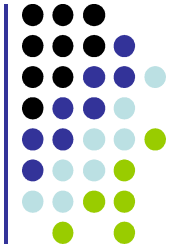
Length of genomic segment: **G**
Number of reads: **N**
Length of each read: **L**

Definition: Coverage **C = N L / G**

How much coverage is enough?

Lander-Waterman model: Prob[not covered bp] = e^{-C}
Assuming uniform distribution of reads, C=10 results in 1 gapped region /1,000,000 nucleotides

Repeats



Bacterial genomes: 5%
Mammals: 50%

Repeat types:

- **Low-Complexity DNA** (e.g. ATATATATACATA…)
- **Microsatellite repeats** $(a_1 \ldots a_k)^N$ where k ~ 3-6
  (e.g. CAGCAGTAGCAGCACCAG)
- **Transposons**
  - **SINE** (Short Interspersed Nuclear Elements)
    e.g., ALU: ~300-long, 10^6 copies
  - **LINE** (Long Interspersed Nuclear Elements)
    ~4000-long, 200,000 copies
  - **LTR retroposons** (Long Terminal Repeats (~700 bp) at each end)
    cousins of HIV
- **Gene Families** genes duplicate & then diverge (paralogs)
- **Recent duplications** ~100,000-long, very similar copies

Sequencing and Fragment Assembly

$3x10^9$ nucleotides

50% of human DNA is compose

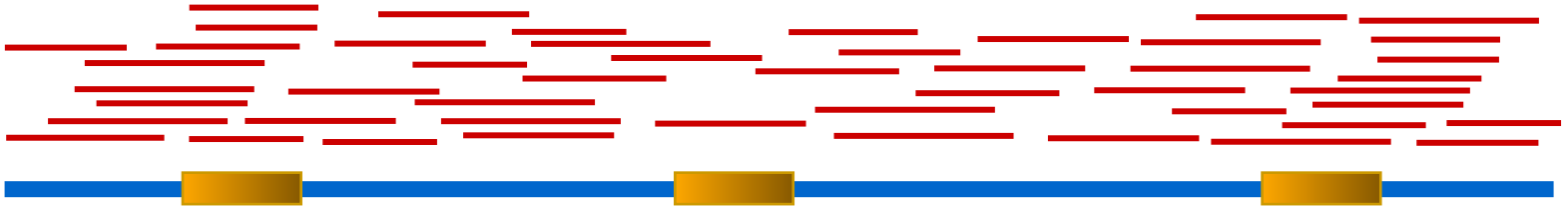
Error!

Glued together two distant regions

What can we do about repeats?

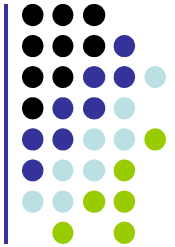
Two main approaches:

- Cluster the reads



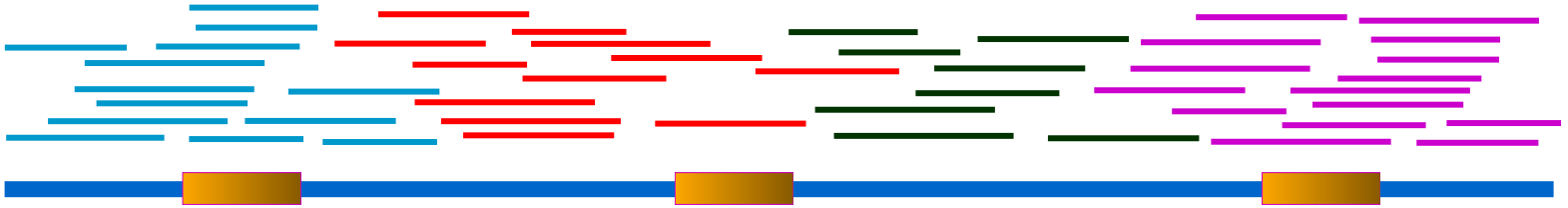
- Link the reads

What can we do about repeats?



Two main approaches:

- Cluster the reads

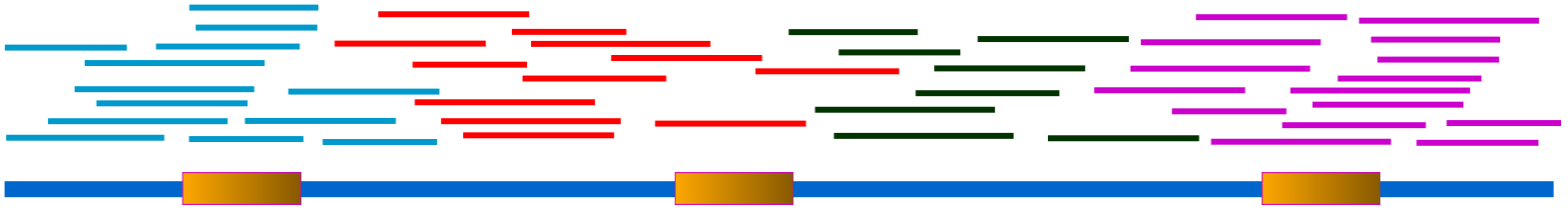


- Link the reads

What can we do about repeats?

Two main approaches:

- Cluster the reads



- Link the reads