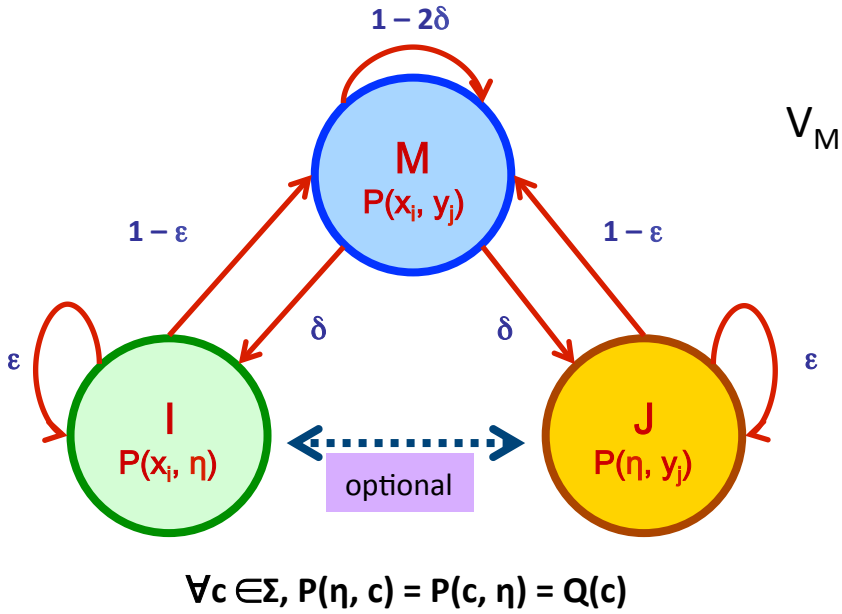
Review: Pair HMMs

- Consider this special case:



$\forall c \in \Sigma, P(\eta, c) = P(c, \eta) = Q(c)$

$$V_M(i, j) = P(x_i, y_j) \max \begin{cases} (1 - 2\delta)\, V_M(i - 1, j - 1) \\ (1 - \varepsilon)\, V_I(i - 1, j - 1) \\ (1 - \varepsilon)\, V_J(i - 1, j - 1) \end{cases}$$

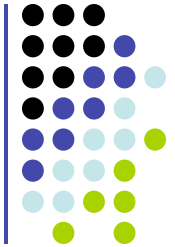
$$V_I(i, j) = Q(x_i) \max \begin{cases} \delta\, V_M(i - 1, j) \\ \varepsilon\, V_I(i - 1, j) \end{cases}$$

$$V_J(i, j) = Q(y_j) \max \begin{cases} \delta\, V_M(i, j - 1) \\ \varepsilon\, V_J(i, j - 1) \end{cases}$$

- Similar for **forward/backward** algorithms
  - (see Durbin et al for details)

QUESTION: What's the computational complexity of DP?

Connection to NW with affine gaps



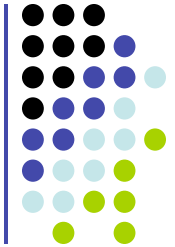
$$V_M(i, j) = \frac{P(x_i, y_j)}{Q(x_i)\,Q(y_j)} \max \begin{cases} (1 - 2\delta)\, V_M(i - 1, j - 1) \\ (1 - \varepsilon)\, V_I(i - 1, j - 1) \\ (1 - \varepsilon)\, V_J(i - 1, j - 1) \end{cases}$$

$$V_I(i, j) = \max \begin{cases} \delta\, V_M(i - 1, j) \\ \varepsilon\, V_I(i - 1, j) \end{cases}$$

$$V_J(i, j) = \max \begin{cases} \delta\, V_M(i, j - 1) \\ \varepsilon\, V_J(i, j - 1) \end{cases}$$

- Account for the extra terms “along the way.”

Connection to NW with affine gaps



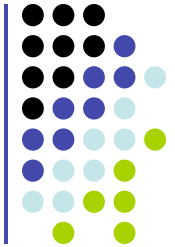
$$\log V_M(i, j) = \log \frac{P(x_i, y_i)}{Q(x_i)\, Q(y_j)} + \max \begin{cases} \log (1 - 2\delta) + \log V_M(i - 1, j - 1) \\ \log (1 - \varepsilon) + \log V_I(i - 1, j - 1) \\ \log (1 - \varepsilon) + \log V_J(i - 1, j - 1) \end{cases}$$

$$\log V_I(i, j) = \max \begin{cases} \log \delta + \log V_M(i - 1, j) \\ \log \varepsilon + \log V_I(i - 1, j) \end{cases}$$

$$\log V_J(i, j) = \max \begin{cases} \log \delta + \log V_M(i, j - 1) \\ \log \varepsilon + \log V_J(i, j - 1) \end{cases}$$

- Take logs, and ignore a couple terms.

Connection to NW with affine gaps

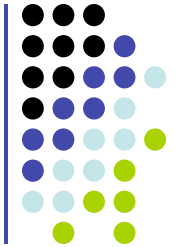


$$M(i, j) = S(x_i, y_j) + \max \begin{cases} M(i-1, j-1) \\ I(i-1, j-1) \\ J(i-1, j-1) \end{cases}$$

$$I(i, j) = \max \begin{cases} d + M(i-1, j) \\ e + I(i-1, j) \end{cases}$$

$$J(i, j) = \max \begin{cases} d + M(i, j-1) \\ e + J(i, j-1) \end{cases}$$

- Rename!

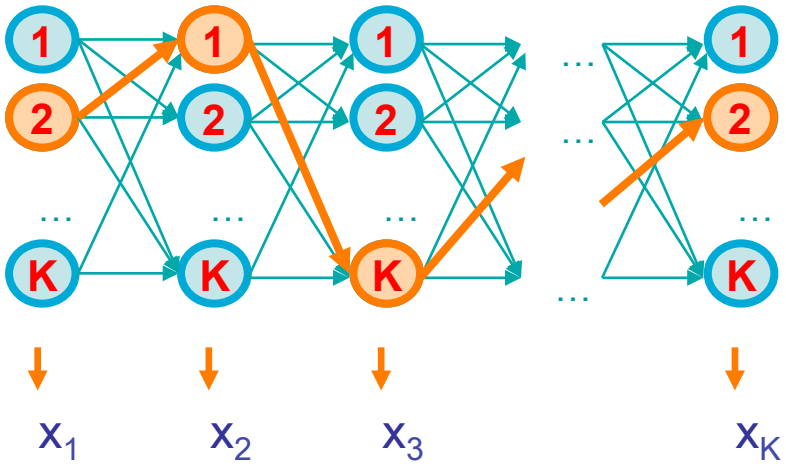


Conditional random fields

Recall Likelihood P(x, π)

$P(x, \pi) = P(x_1, \ldots, x_N, \pi_1, \ldots\ldots, \pi_N) =$

$a_{0\pi 1}\, a_{\pi 1 \pi 2} \ldots\ldots a_{\pi N-1 \pi N}\, e_{\pi 1}(x_1) \ldots\ldots e_{\pi N}(x_N)$



- Enumerate all parameters a_{ij} and $e_i(b)$; n params

$a_{0Fair} : \theta_1$; $a_{0Loaded} : \theta_2$; … $e_{Loaded}(6) = \theta_{18}$

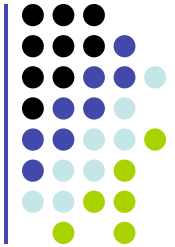
- Count the # of times each parameter j = 1, …, n occurs

F(j, x, π) = # parameter θ_j occurs in (x, π)

- *(call **F(.,.,.)** the **feature counts**)* Then,

$$P(x, \pi) = \prod_{j=1\ldots n} \theta_j^{F(j, x, \pi)} = \exp\left[\sum_{j=1\ldots n} \log(\theta_j) \times F(j, x, \pi)\right]$$

Conditional random fields - Recap



- **Definition**

$$P(\pi \mid x) = \frac{\exp(\sum_{i=1 \ldots |x|} w^T F(\pi_i, \pi_{i-1}, x, i))}{\sum_{\pi'} \exp(\sum_{i=1 \ldots |x|} w^T F(\pi'_i, \pi'_{i-1}, x, i))}$$

partition coefficient

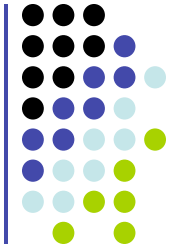
where

F : (state, state, observations, index) → $\mathbf{R}^n$ "local feature mapping"

$w \in \mathbf{R}^n$ "parameter vector"

  - Summation over all possible state sequences $\pi'_1 \ldots \pi'_{|x|}$
  - $a^T b$ for vectors $a, b \in \mathbf{R}^n$ denotes inner product, $\sum_{i=1 \ldots n} a_i b_i$

Relationship with HMMs

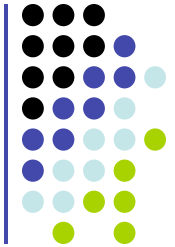


- For each component w_j, define F_j to be a 0/1 indicator variable of whether the j^{th} parameter should be included in scoring x, π at position i:

$$w = \begin{bmatrix} \log a_0(1) \\ \dots \\ \log a_0(K) \\ \log a_{11} \\ \dots \\ \log a_{KK} \\ \log e_1(b_1) \\ \dots \\ \log e_K(b_M) \end{bmatrix} \in \mathbf{R}^n \qquad F(\pi_i, \pi_{i-1}, x, i) = \begin{bmatrix} 1\{i = 1 \wedge \pi_{i-1} = 1\} \\ \dots \\ 1\{i = 1 \wedge \pi_{i-1} = K\} \\ 1\{\pi_{i-1} = 1 \wedge \pi_i = 1\} \\ \dots \\ 1\{\pi_{i-1} = K \wedge \pi_i = K\} \\ 1\{x_i = b_1 \wedge \pi_i = 1\} \\ \dots \\ 1\{x_i = b_M \wedge \pi_i = K\} \end{bmatrix} \in \mathbf{R}^n$$

- Then, $\log P(x, \pi) = \sum_{i=1 \dots |x|} w^T F(\pi_i, \pi_{i-1}, x, i)$

CRFS ≥ HMMs (continued)



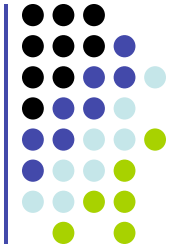
- In an HMM, our features were of the form

$$F(\pi_i, \pi_{i-1}, x, i) = F(\pi_i, \pi_{i-1}, x_i, i)$$

  - i.e., when scoring position i in the sequence, feature only considered the emission x_i at position i.
  - Cannot look at other positions (e.g., x_{i-1}, x_{i+1}) since that would involve "emitting" a character more than once – double-counting of probability

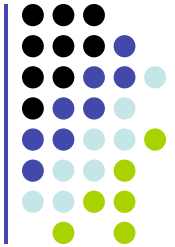
- CRFs don't have this restriction
  - Why? Because CRFs don't attempt to model the observations x!

3 basic questions for CRFs



- **Evaluation:** Given a sequence of observations x and a sequence of states π, compute $P(\pi \mid x)$
- **Decoding:** Given a sequence of observations x, compute the maximum probability sequence of states $\pi_{ML} = \arg\max_{\pi} P(\pi \mid x)$
- **Learning:** Given a CRF with unspecified parameters w, compute the parameters that maximize the likelihood of π given x, i.e., $w_{ML} = \arg\max_{w} P(\pi \mid x, w)$

Viterbi for CRFs



- Note that:

$$\text{argmax}_\pi P(\pi \mid x) = \text{argmax}_\pi \frac{\exp(\sum_{i=1 \ldots |x|} w^T F(\pi_i, \pi_{i-1}, x, i))}{\sum_{\pi'} \exp(\sum_{i=1 \ldots |x|} w^T F(\pi'_i, \pi'_{i-1}, x, i))}$$

$$= \arg\max_\pi \exp(\sum_{i=1 \ldots |x|} w^T F(\pi_i, \pi_{i-1}, x, i))$$

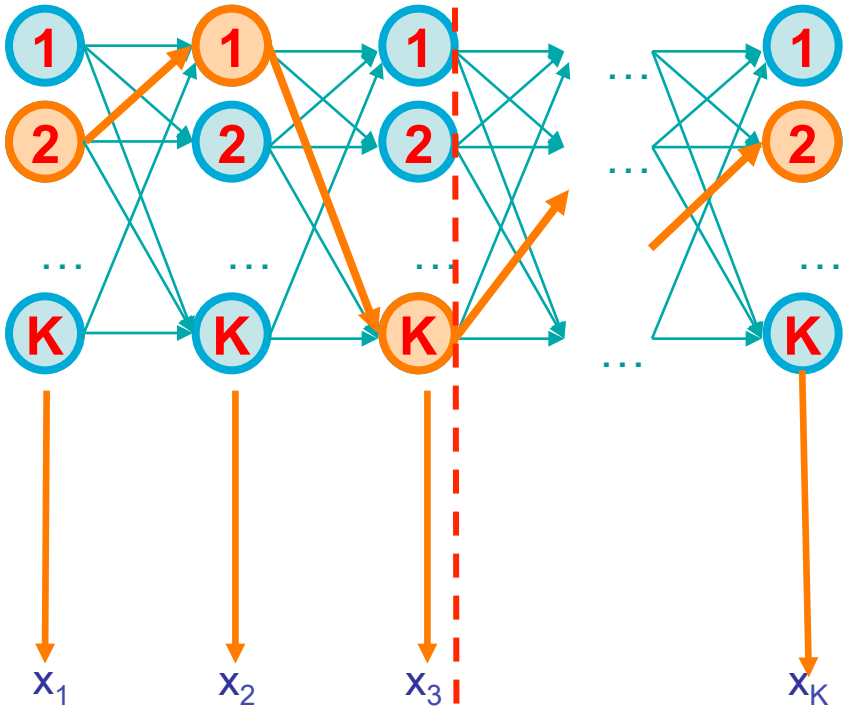
$$= \arg\max_\pi \sum_{i=1 \ldots |x|} w^T F(\pi_i, \pi_{i-1}, x, i)$$

- We can derive the following recurrence:

$$V_k(i) = \max_j [\, w^T F(k, j, x, i) + V_j(i-1) \,]$$

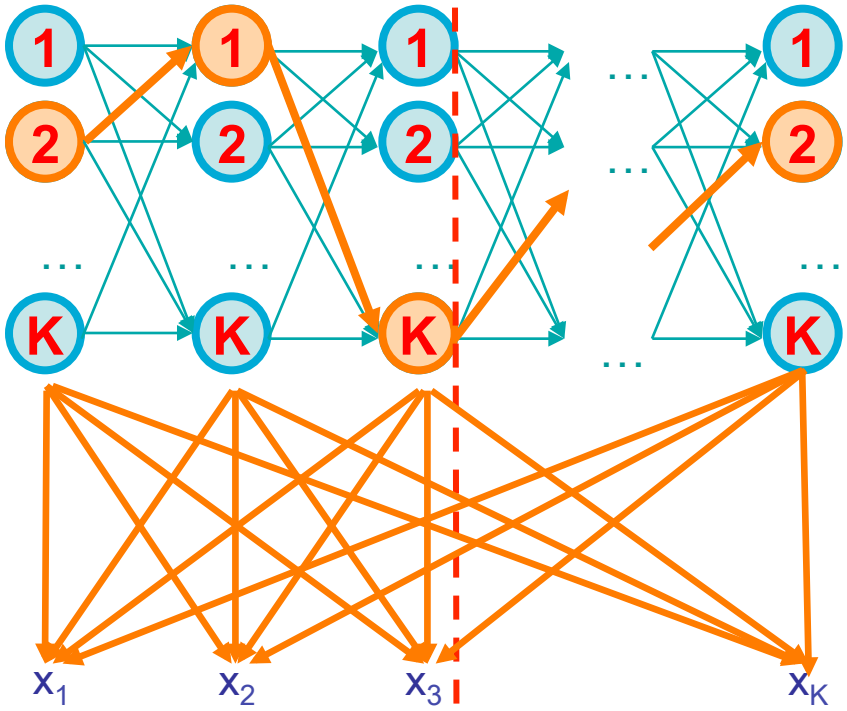
- **Notes:**
  - Even though the features may depend on arbitrary positions in x, x is constant. DP depends only on knowing the previous state
  - Computing the partition function (denominator) can be done by a similar adaptation of the forward/backward algorithms

Viterbi for CRFs



Given that we end up in state k at step i, maximize score to the left and right

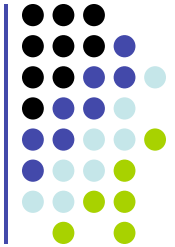
Viterbi for CRFs



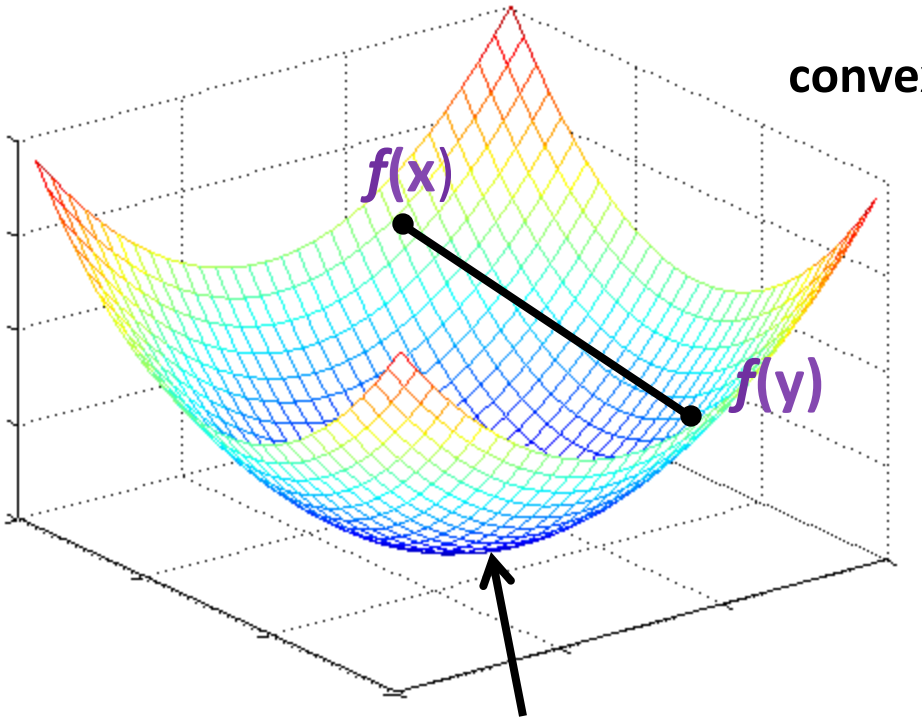
Given that we end up in state k at step i, maximize score to the left and right

X is fixed:
=> parse to the left of step i, given we end in state k, does not affect parse to the right of step i

Learning CRFs

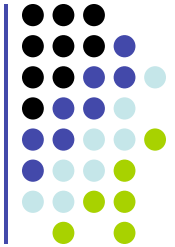


- Key observation: $-\log P(\pi \mid x, w)$ is a differentiable, **convex** function of w



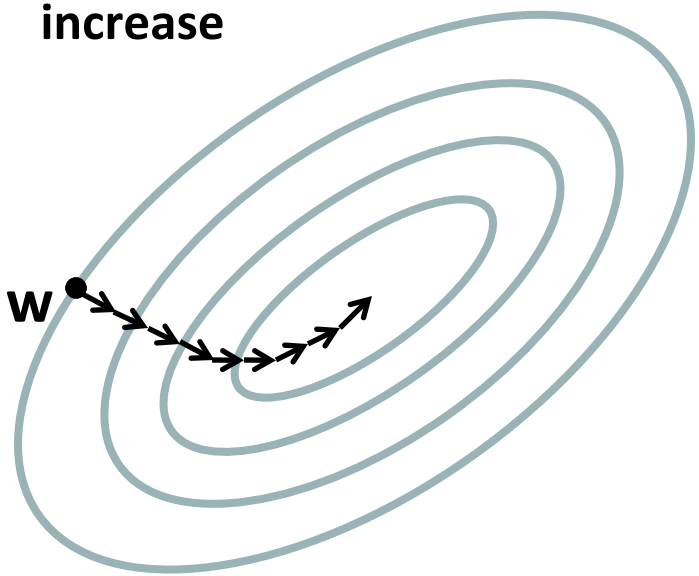
Any local minimum is a global minimum.

Learning CRFs (continued)

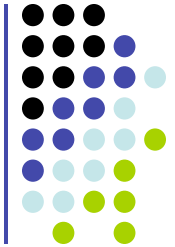


- Compute partial derivative of log $P(\pi \mid x, w)$ with respect to each parameter w_j, and use the gradient ascent learning rule:

Gradient points in the direction of greatest function increase



The CRF gradient



- It turns out that

$$(\partial/\partial w_j) \log P(\pi \mid x, w) = F_j(x, \pi) - E_{\pi' \sim P(\pi' \mid x, w)} [\, F_j(x, \pi') \,]$$

correct value for jth feature

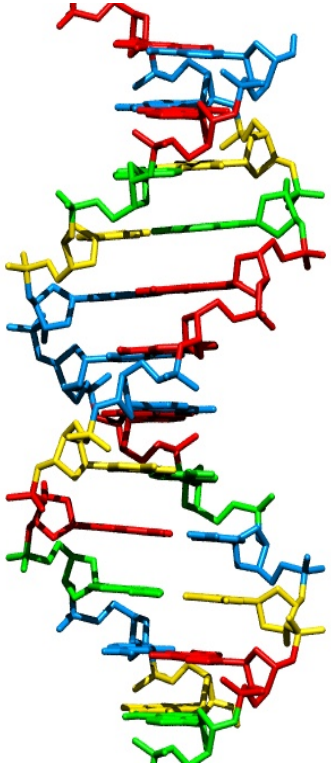
expected value for jth feature (given the current parameters)

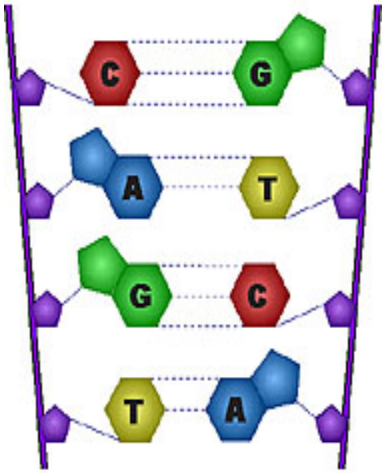
- This has a very nice interpretation:
  - We increase parameters for which the correct feature values are greater than the predicted feature values
  - We decrease parameters for which the correct feature values are less than the predicted feature values
- This moves probability mass from incorrect parses to correct parses

DNA Sequencing

DNA sequencing

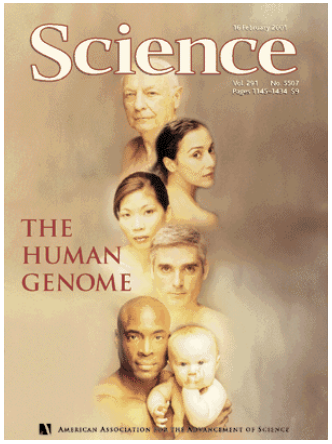
How we obtain the sequence of nucleotides of a species





...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?

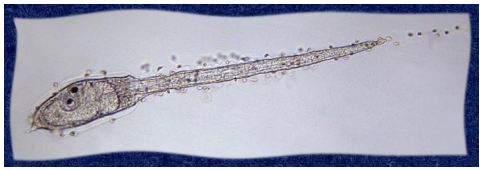
Which human?

Answer one:

Answer two: it doesn't matter

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

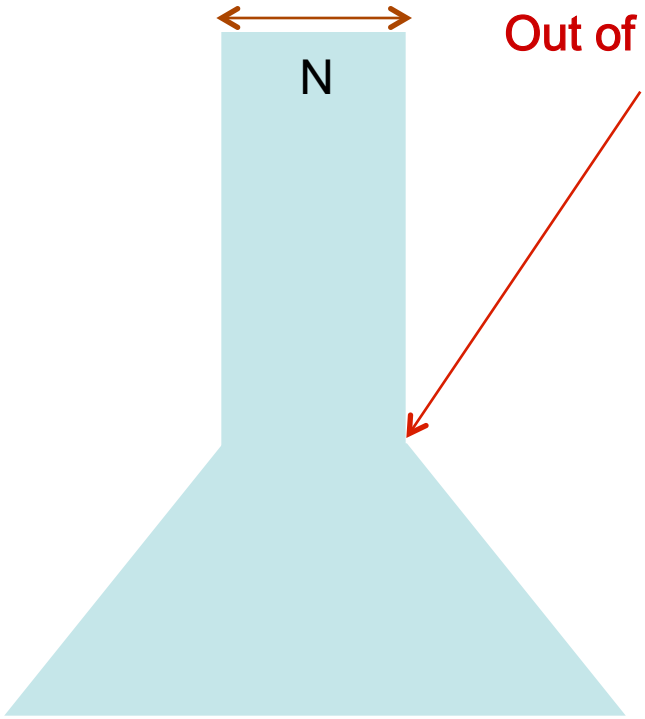
Humans: ~1/1,000



Other organisms have much higher polymorphism rates

- Population size!

Why humans are so similar



Heterozygosity: H

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

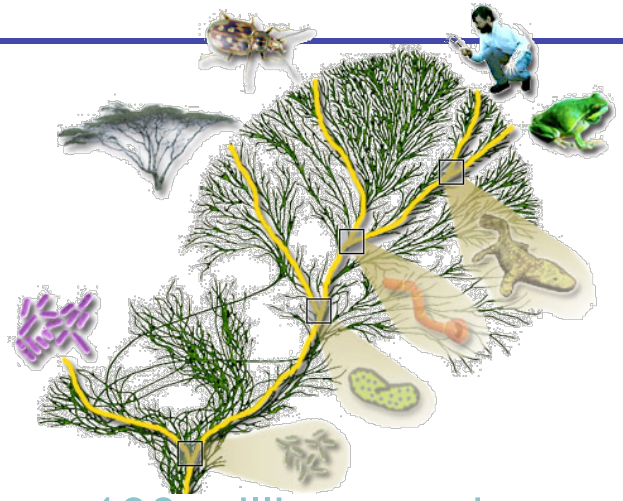
$u \sim 10^{-8}$, $N \sim 10^{4}$

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

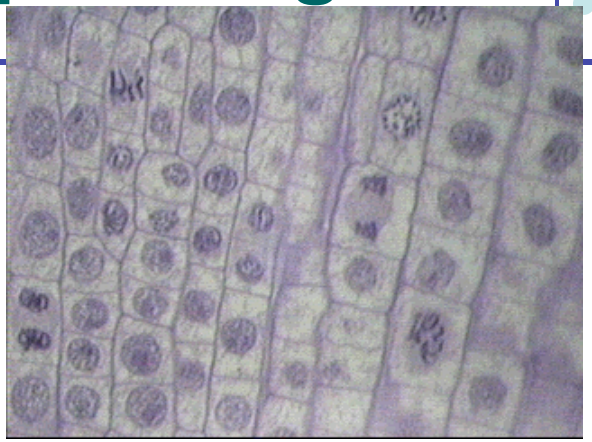
A small population that interbred reduced the genetic variation

Out of Africa ~ 40,000 years ago

There is never “enough” sequencing

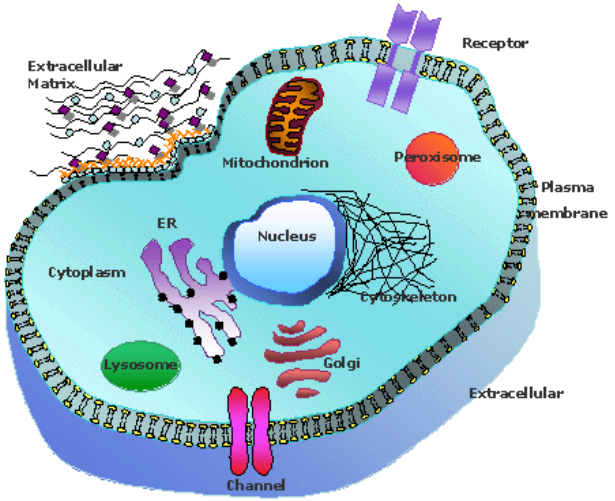


100 million species



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

7 billion individuals

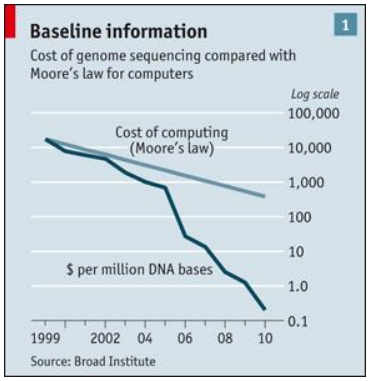


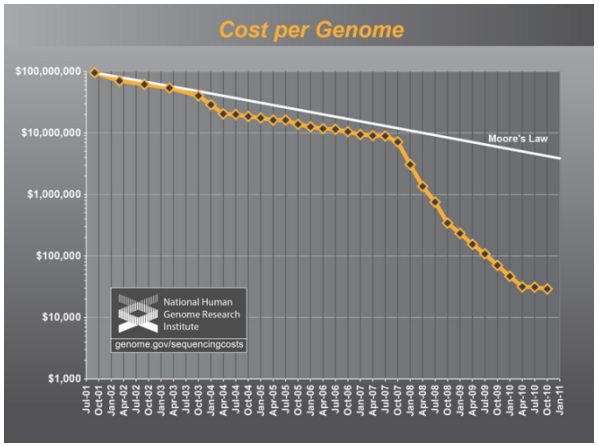
Sequencing is a
functional assay

Sequencing Growth

Cost of one human genome

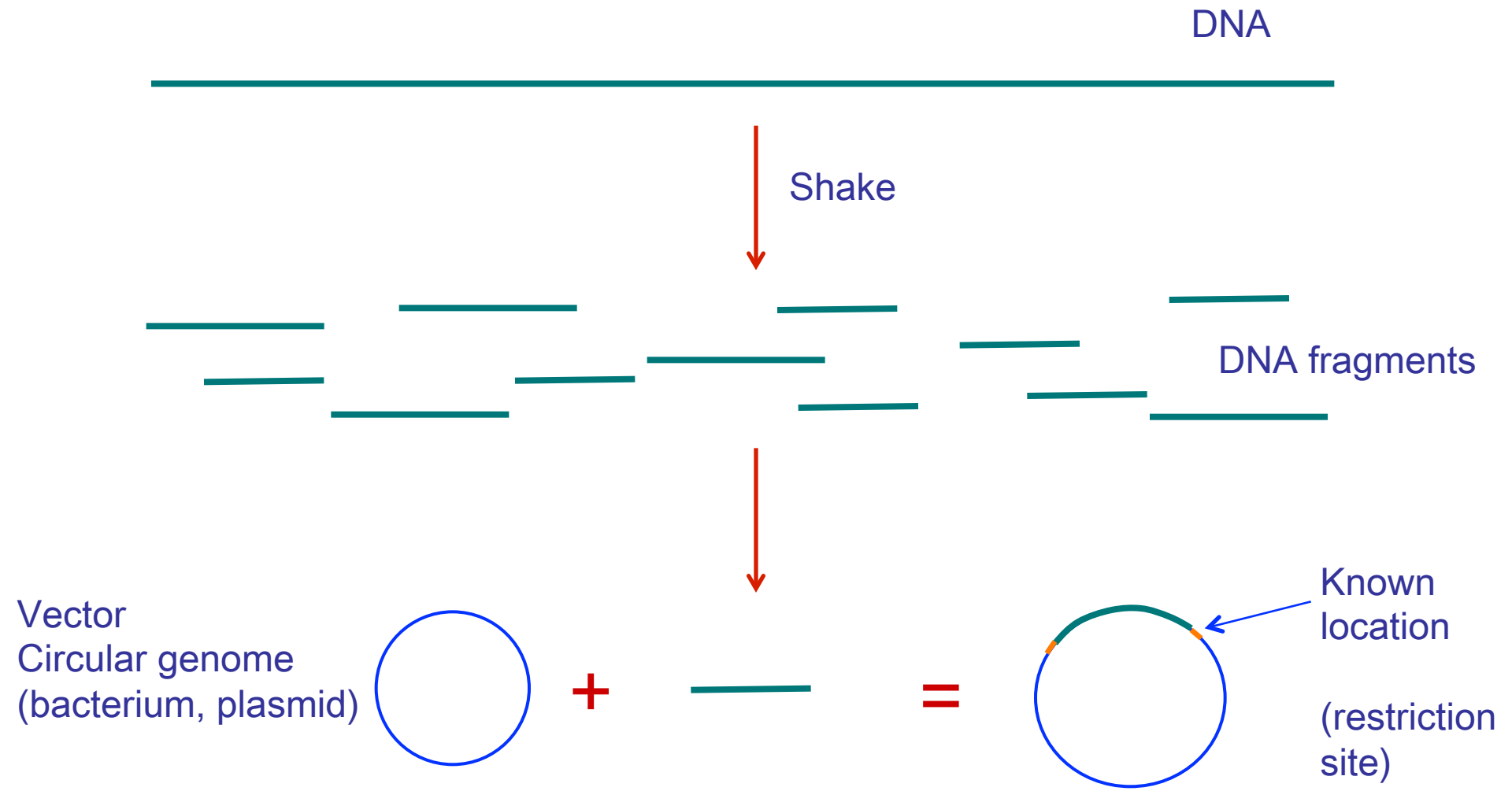
- 2004: $30,000,000
- 2008: $100,000
- 2010: $10,000
- **2015: $1,000**
- ???: $300



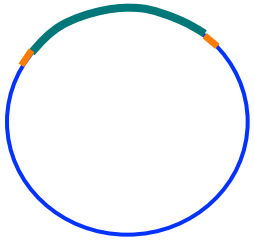


How much would you pay for a smartphone?

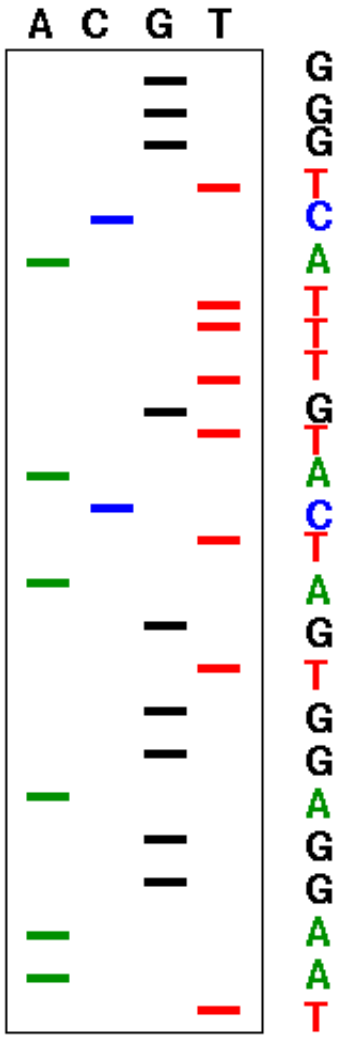
Ancient sequencing technology – Sanger Vectors

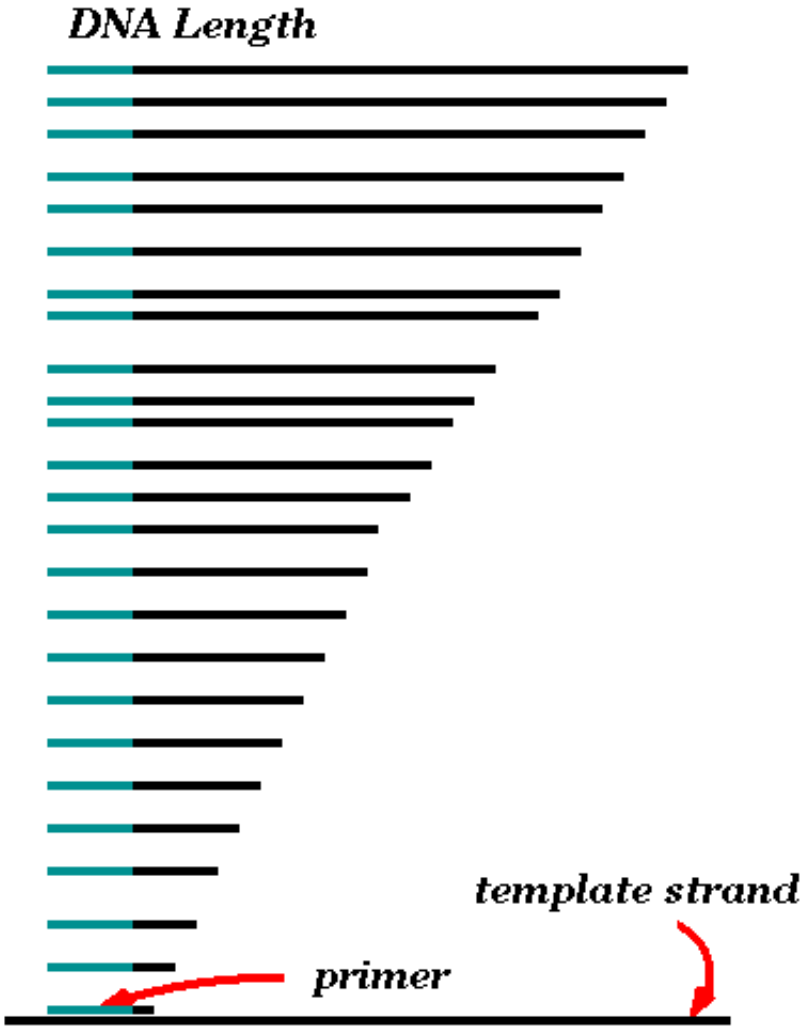


Ancient sequencing technology – Sanger Gel Electrophoresis



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





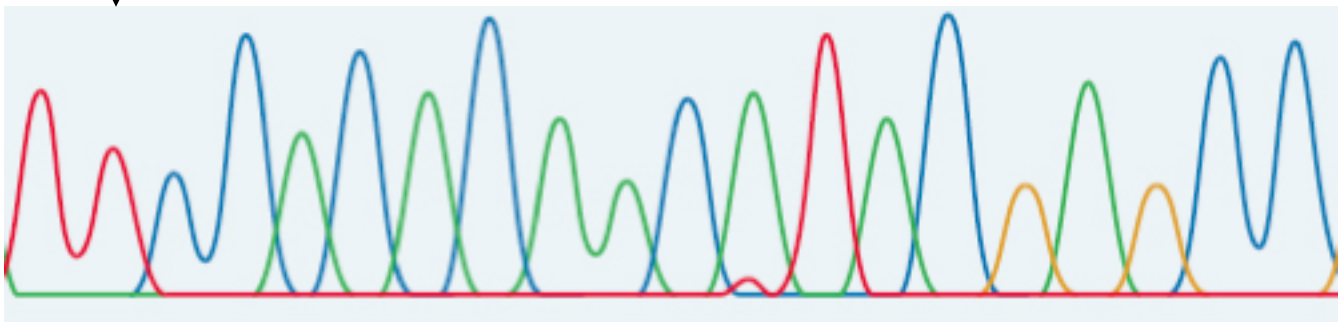
Fluorescent Sanger sequencing trace

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)

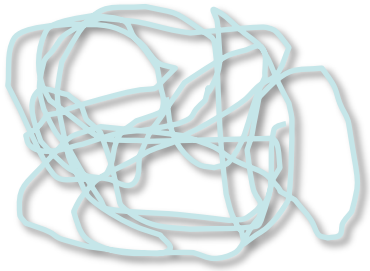
Trace



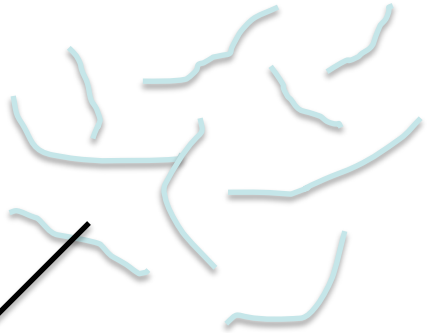
Recombinant DNA: Genes and Genomes.
3rd Edition (Dec06). WH Freeman Press.

Slide Credit: Arend Sidow

Making a Library (present)

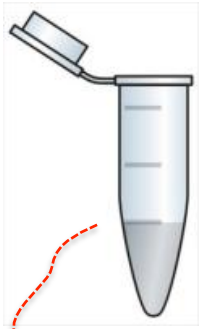


shear to ~500 bases



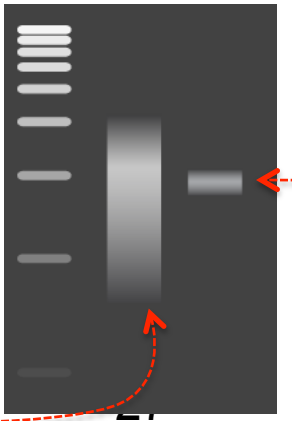
put on linkers

eventual forward and reverse sequence



PCR to obtain preparative quantities

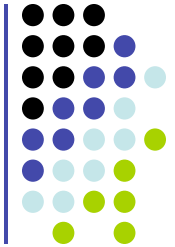
size selection on preparative gel



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

Slide Credit: Arend Sidow

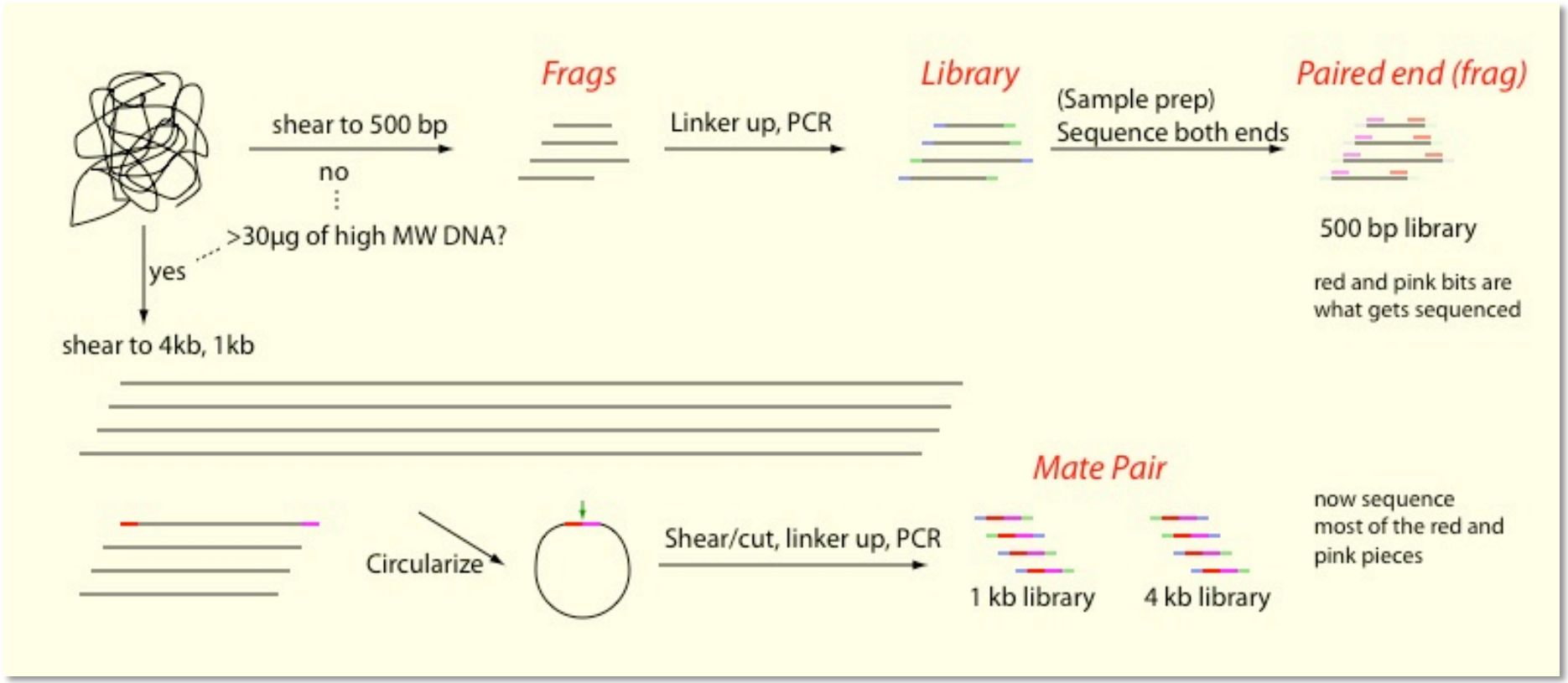
Library



- 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

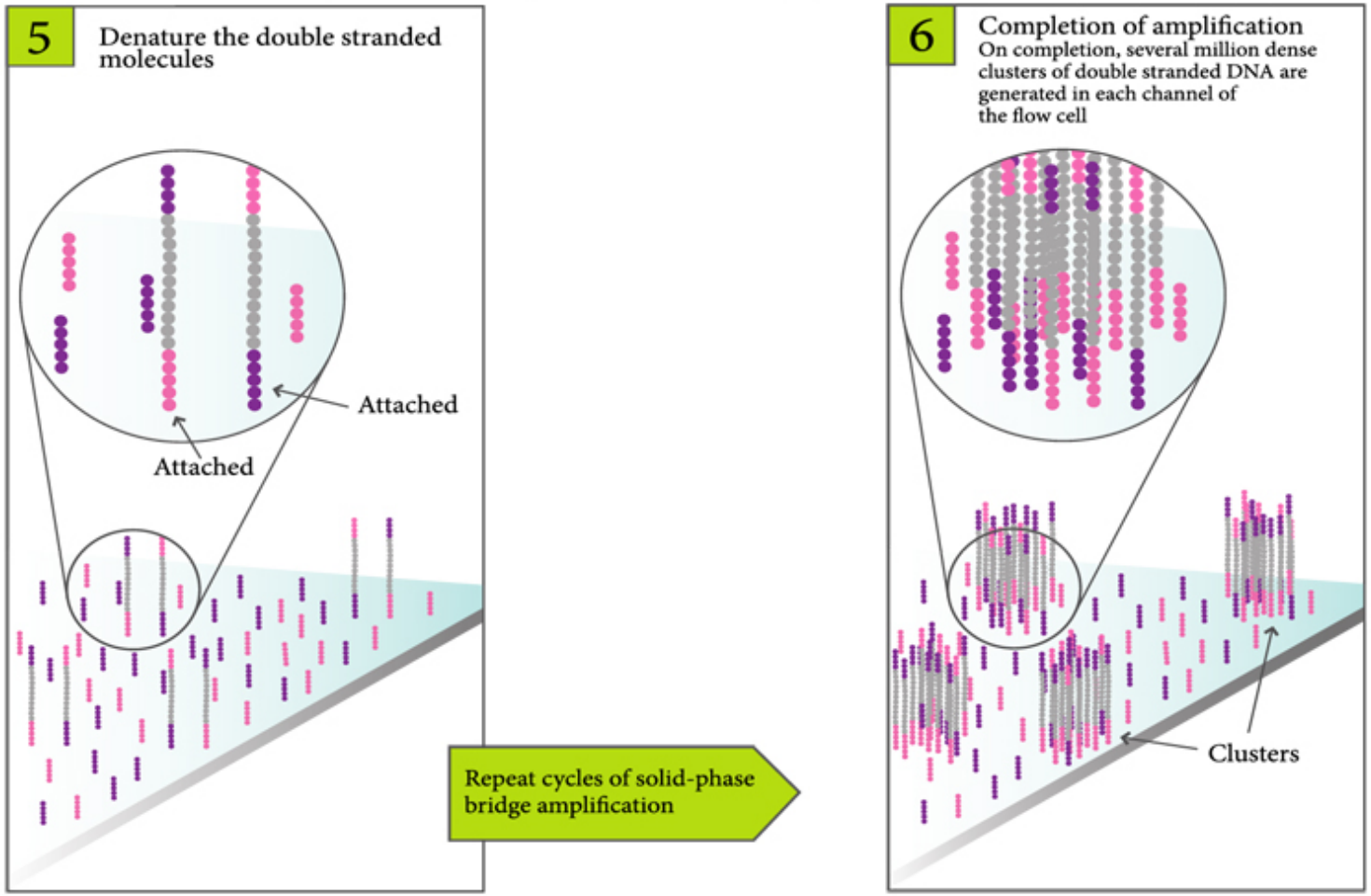
Fragment vs Mate pair ('jumping')



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

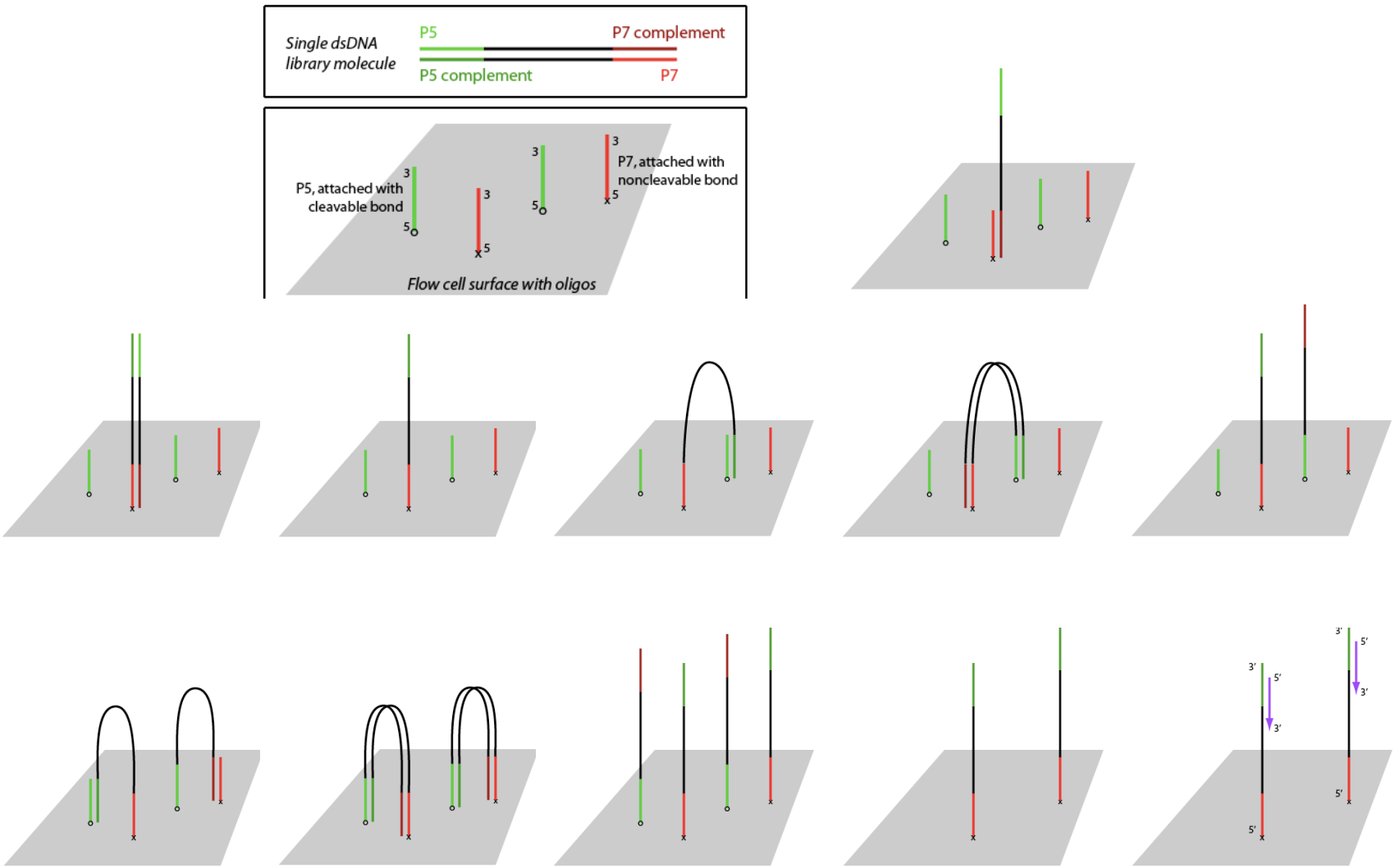
Slide Credit: Arend Sidow

Illumina cluster concept



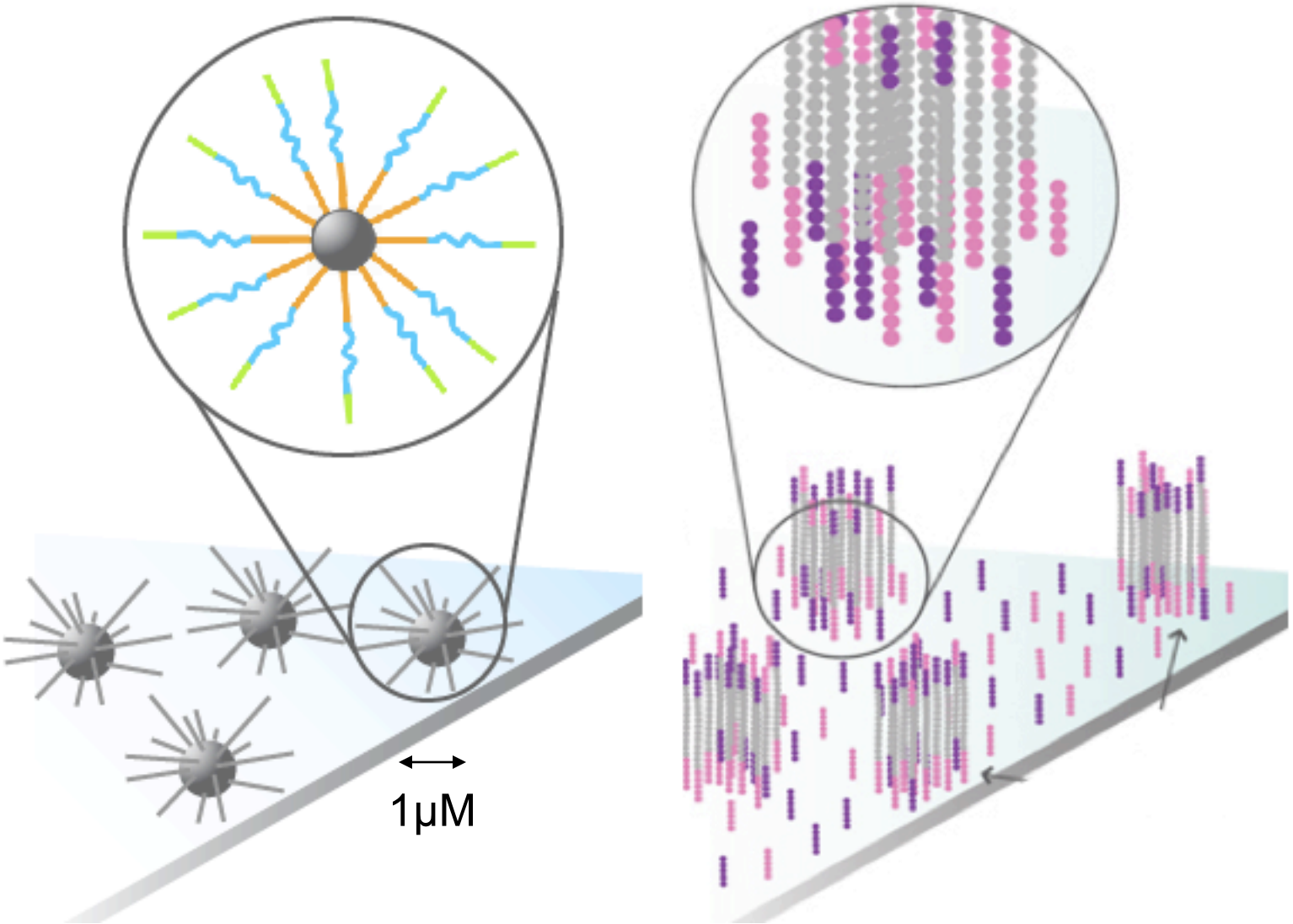
Slide Credit: Arend Sidow

Cluster generation ('bridge amplification')



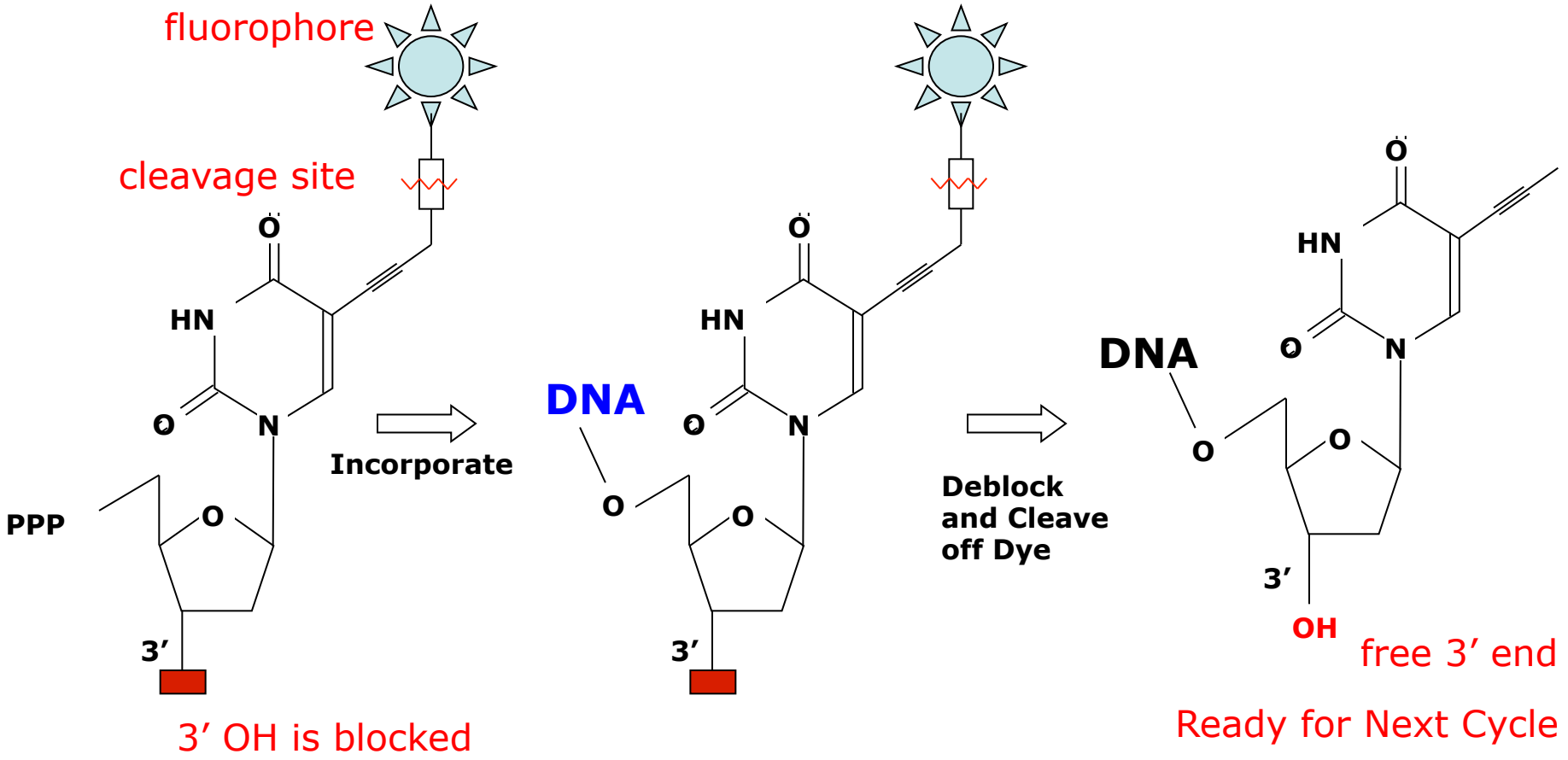
Slide Credit: Arend Sidow

Clonally Amplified Molecules on Flow Cell



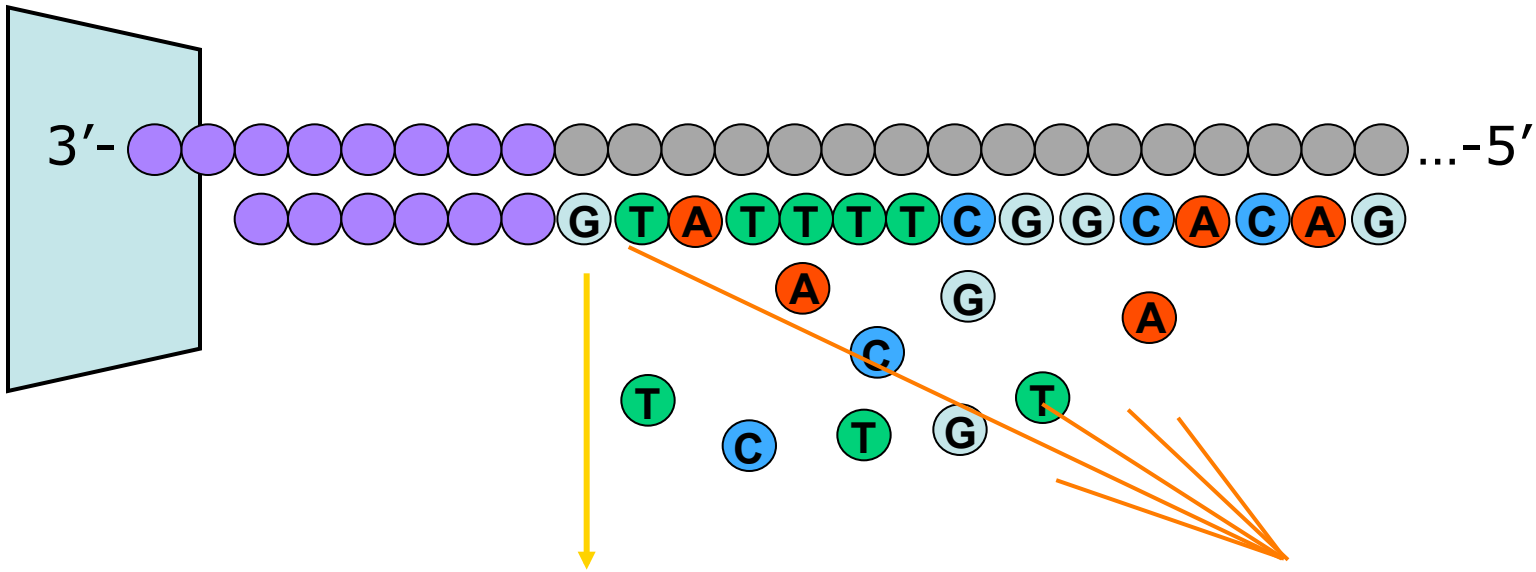
Slide Credit: Arend Sidow

Reversible Terminators



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

Sequencing power for every scale.

Find the sequencing system that's right for your lab.

Compare key specifications across the whole portfolio of Illumina sequencing systems. Understand the differences between the MiniSeq, MiSeq, NextSeq, HiSeq, and HiSeq X Series.

	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	• MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics	• HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

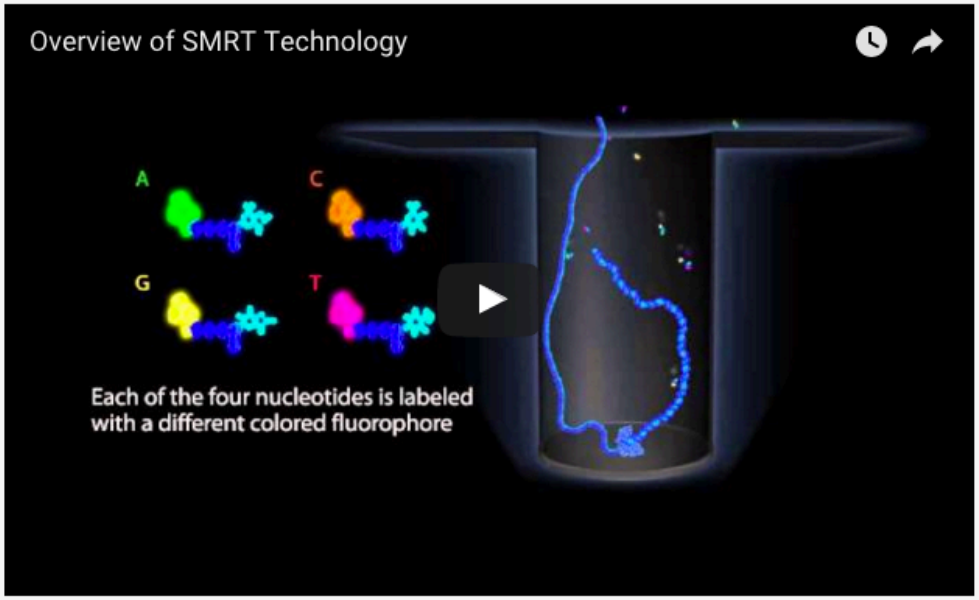
Pacific Biosciences SMRT technology

The SMRT Sequencing advantage

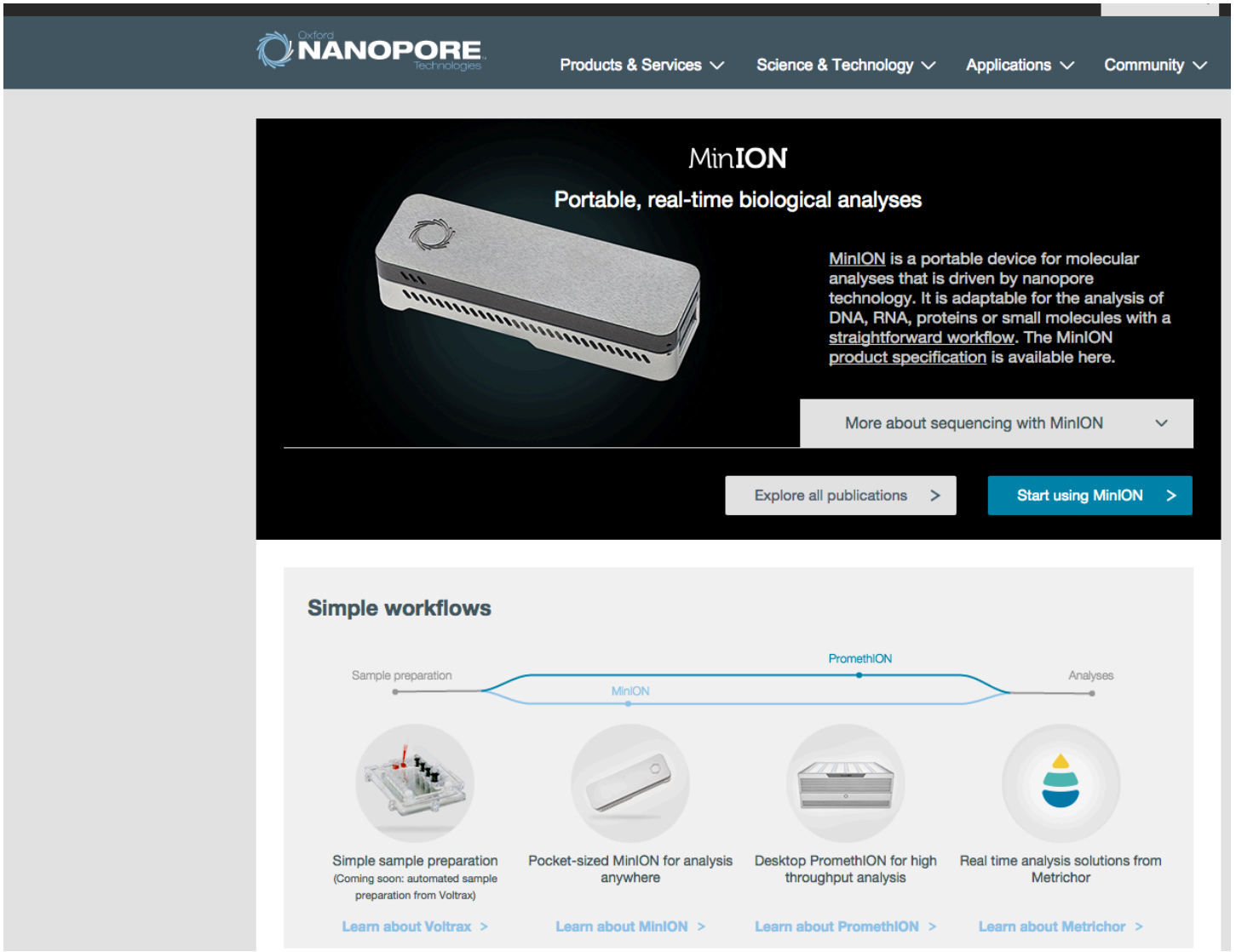
SMRT Sequencing is ideal for a variety of research applications and offers many benefits, including:

- Longest average read lengths
- Highest consensus accuracy
- Uniform coverage
- Simultaneous epigenetic characterization
- Single-molecule resolution

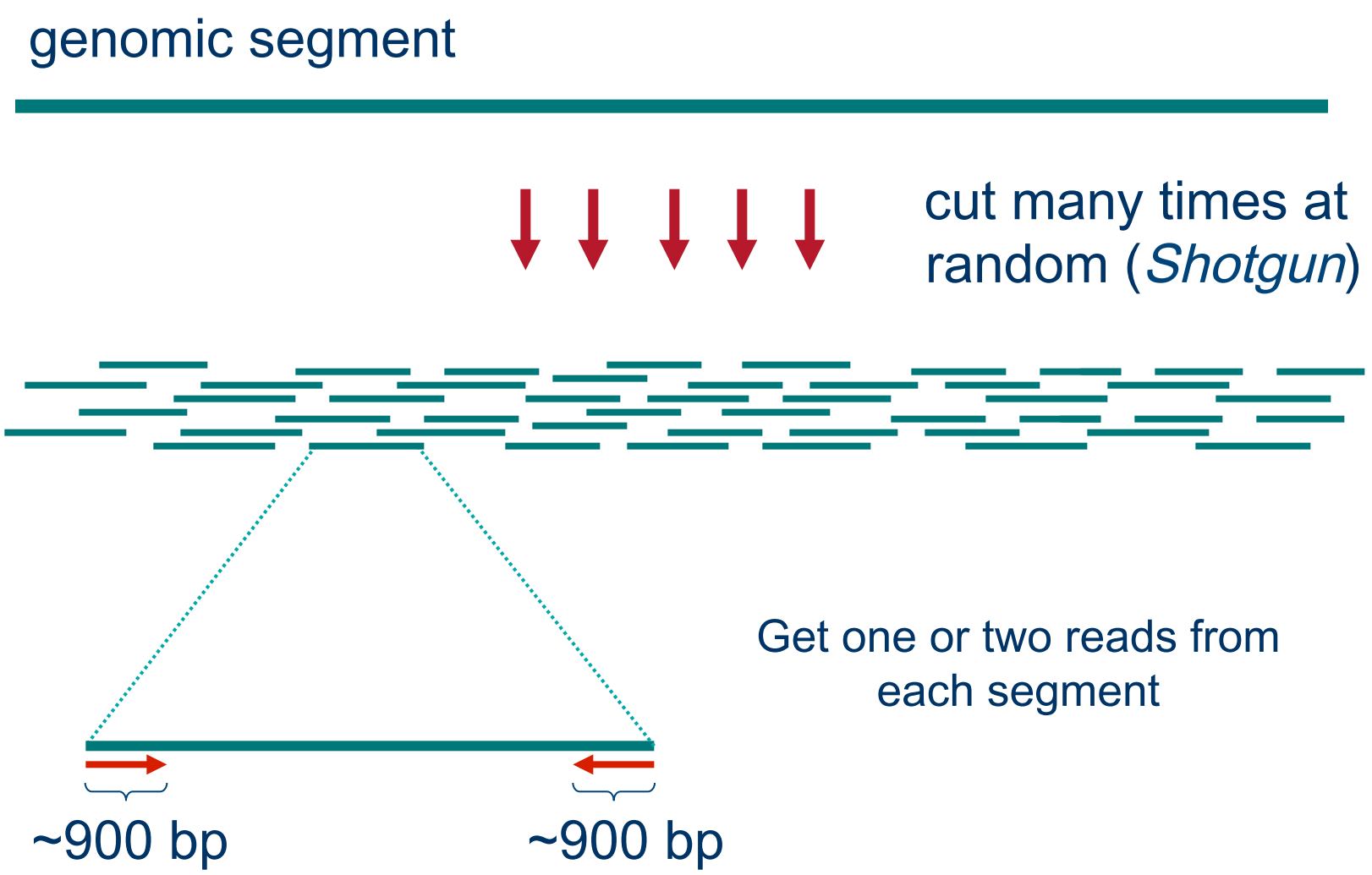
An overview of SMRT Sequencing



Oxford Nanopore



Method to sequence longer regions

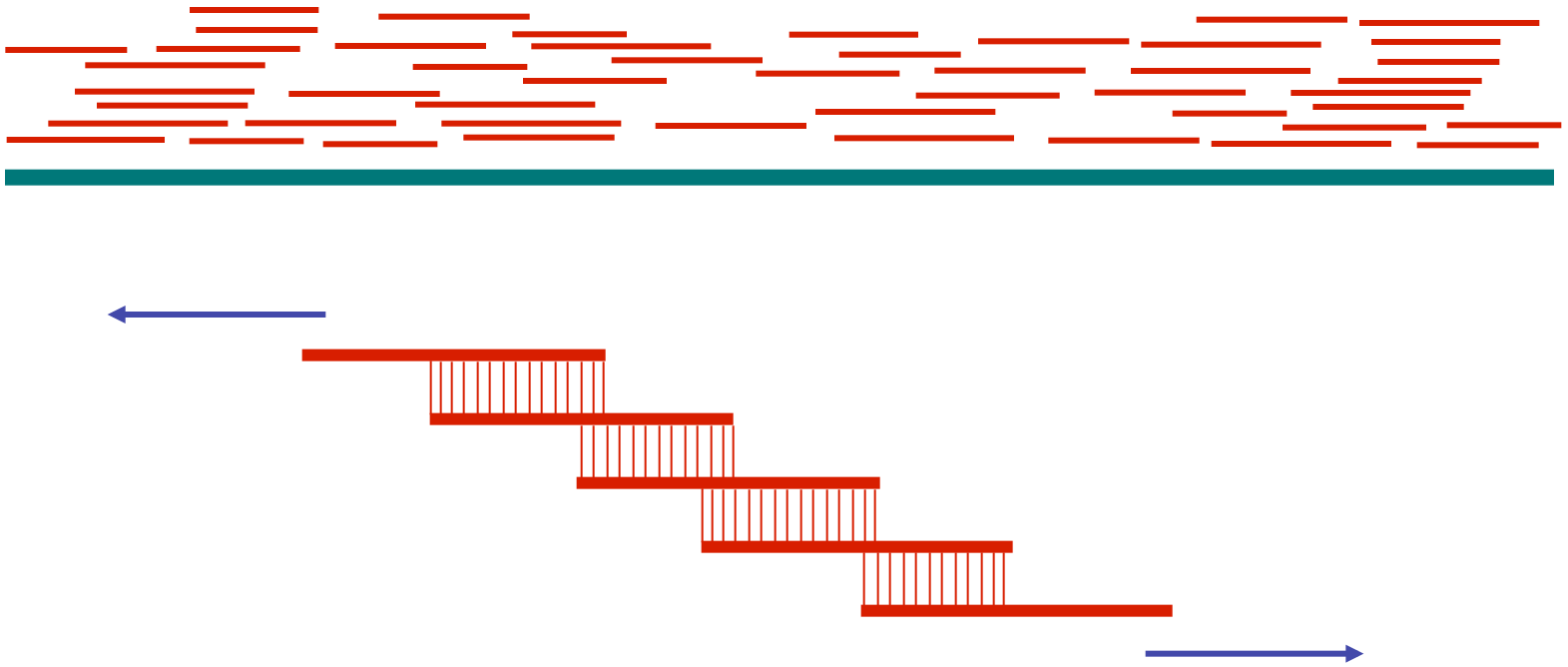


Two main assembly problems

- 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

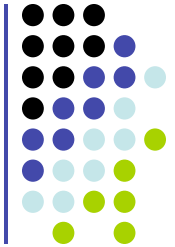
Length of genomic segment: **G**
Number of reads: **N**
Length of each read: **L**

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

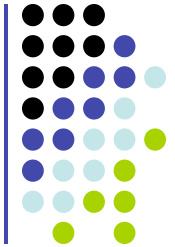
Sequencing and Fragment Assembly

$3x10^9$ nucleotides

50% of human DNA is compose

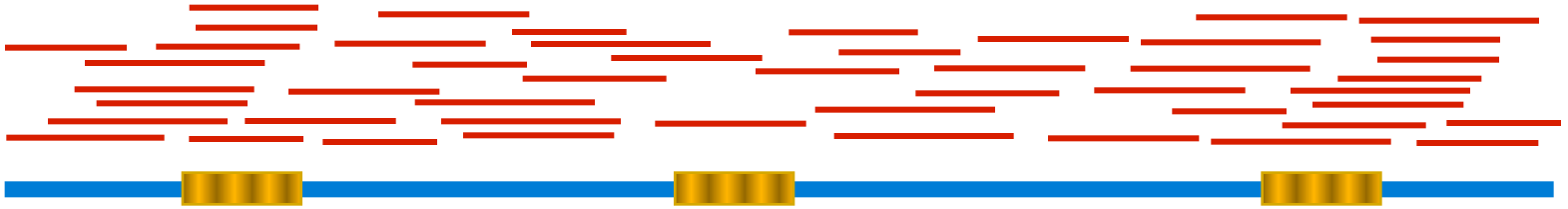
Error!
Glued together two distant regions

What can we do about repeats?



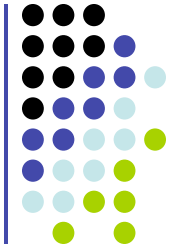
Two main approaches:

- Cluster the reads



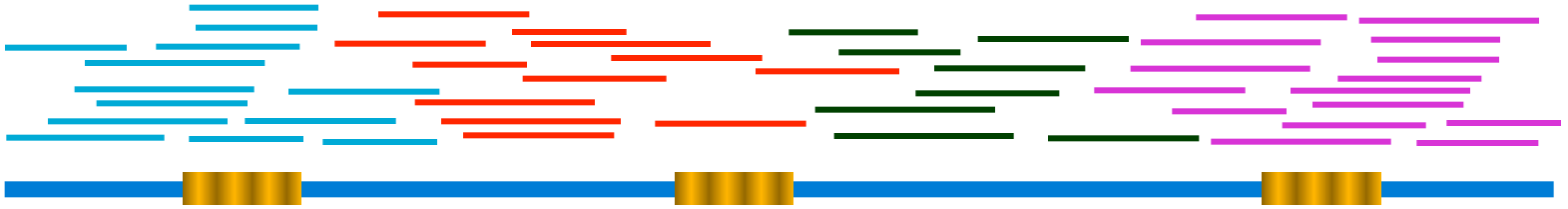
- Link the reads

What can we do about repeats?



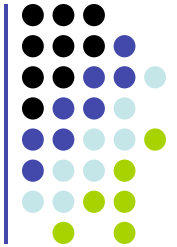
Two main approaches:

- Cluster the reads



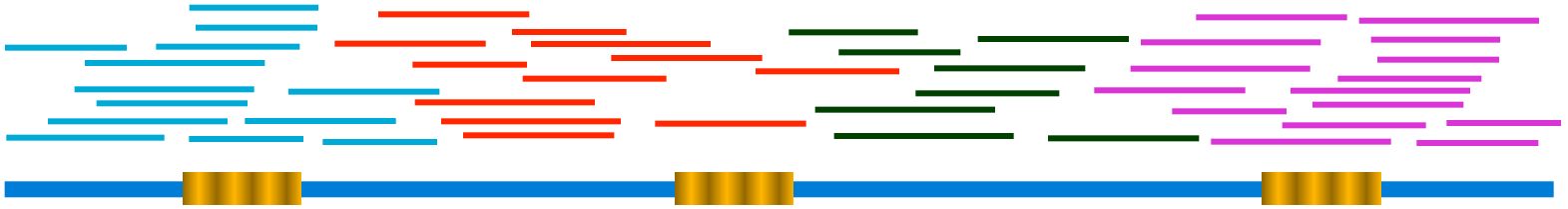
- Link the reads

What can we do about repeats?

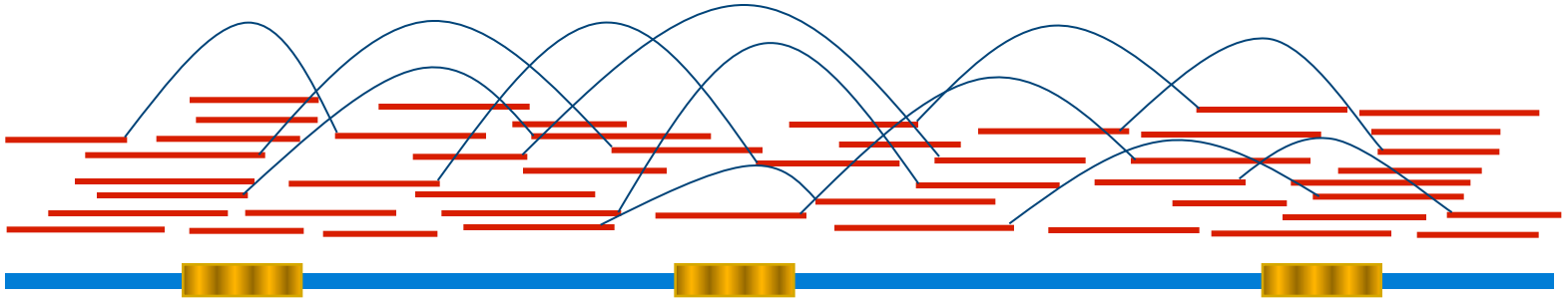


Two main approaches:

- Cluster the reads

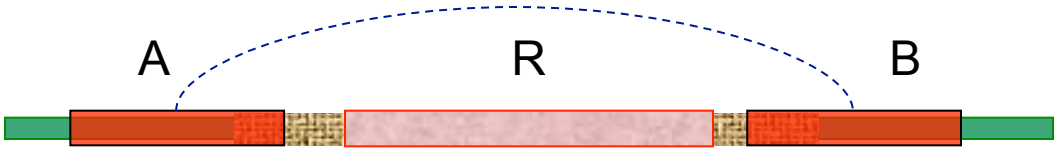


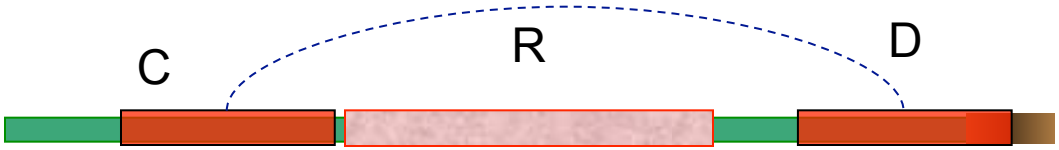
- Link the reads



Sequencing and Fragment Assembly

$3x10^9$ nucleotides





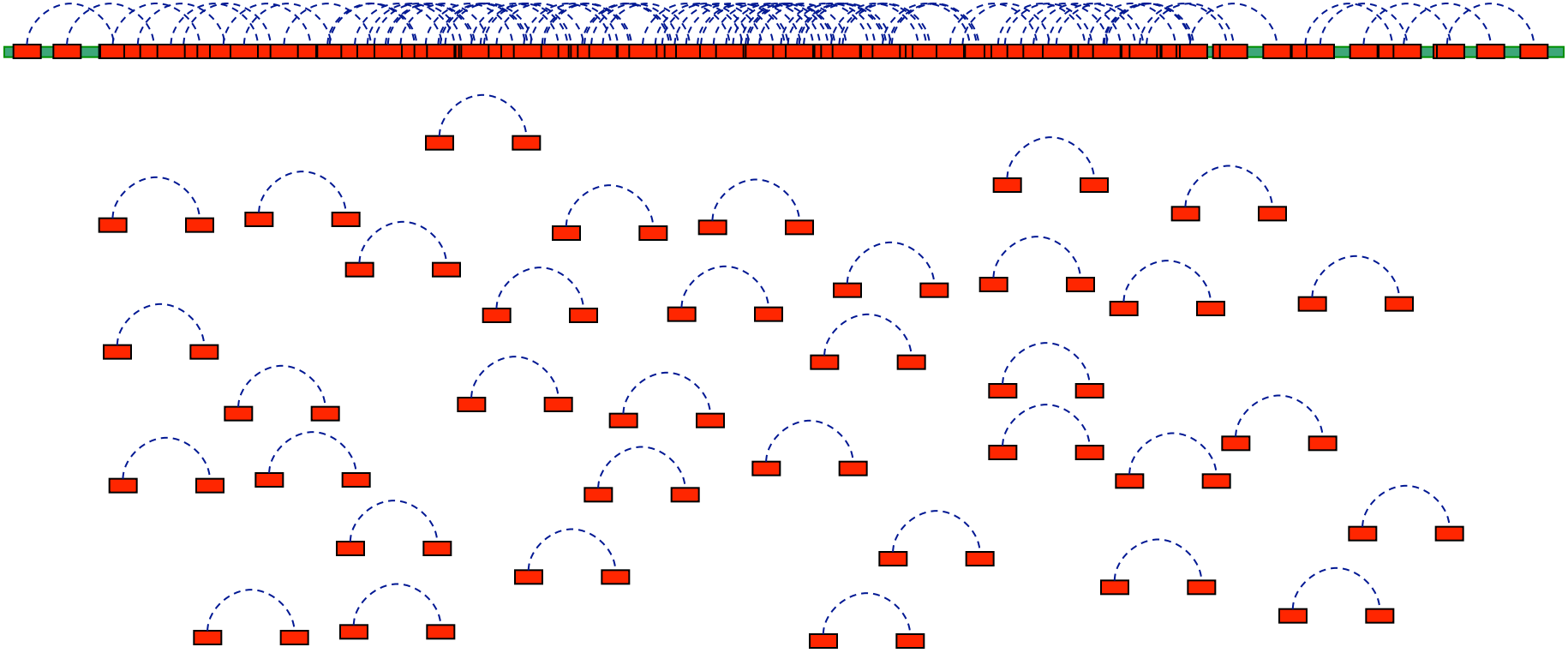
ARB, CRD

or

ARD, CRB ?

Sequencing and Fragment Assembly

$3x10^9$ nucleotides



Fragment Assembly
(in whole-genome shotgun sequencing)

Fragment Assembly

Steps to Assemble a Genome

Some Terminology

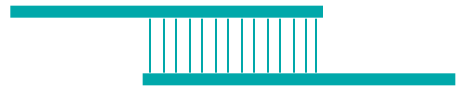
read a 500-900 long word that comes out of sequencer

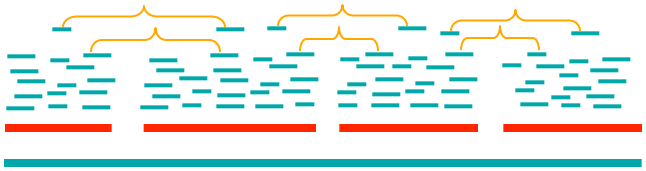
mate pair a pair of reads from two ends of the same insert fragment

contig a contiguous sequence formed by several overlapping reads with no gaps

supercontig (scaffold) an ordered and oriented set of contigs, usually by mate pairs

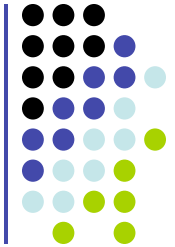
consensus sequene sequence derived from the multiple alignment of reads in a contig





..ACGATTACAATAGGTT..

1. Find Overlapping Reads

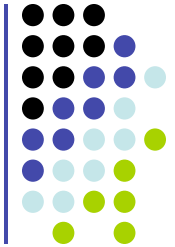


```
aaactgcagtacggatct
aaactgcag
 aactgcagt
...
         gtacggatct
          tacggatct
gggcccaaactgcagtac
gggcccaaa
 ggcccaaac
...
         actgcagta
          ctgcagtac
gtacggatctactacaca
gtacggatc
 tacggatct
...
         ctactacac
          tactacaca
```

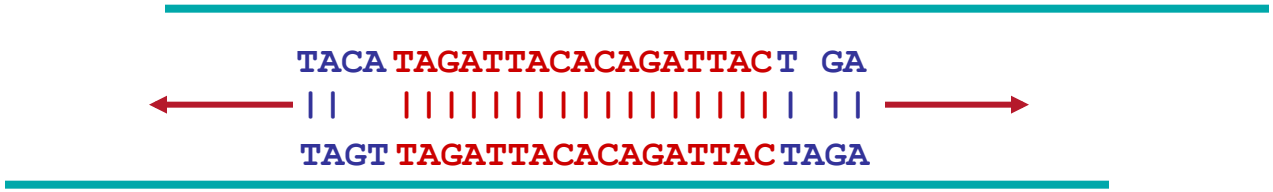
```
(read, pos., word, orient.)
aaactgcag
aactgcagt
actgcagta
...
gtacggatc
tacggatct
gggcccaaa
ggcccaaac
gcccaaact
...
actgcagta
ctgcagtac
gtacggatc
tacggatct
acggatcta
...
ctactacac
tactacaca
```

```
(word, read, orient., pos.)
aaactgcag
aactgcagt
acggatcta
actgcagta
actgcagta
cccaaactg
cggatctac
ctactacac
ctgcagtac
ctgcagtac
gcccaaact
ggcccaaac
gggcccaaa
gtacggatc
gtacggatc
tacggatct
tacggatct
tactacaca
```

1. Find Overlapping Reads

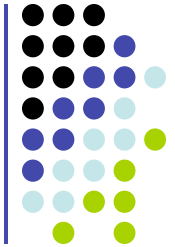


- Find pairs of reads sharing a k-mer, k ~ 24
- Extend to full alignment – throw away if not >98% similar



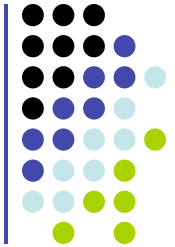
- Caveat: repeats
  - A k-mer that occurs N times, causes $O(N^2)$ read/read comparisons
  - ALU k-mers could cause up to $1{,}000{,}000^2$ comparisons
- Solution:
  - Discard all k-mers that occur “too often”
    - Set cutoff to balance sensitivity/speed tradeoff, according to genome at hand and computing resources available

1. Find Overlapping Reads



Create local multiple alignments from the overlapping reads

1. Find Overlapping Reads



- Correct errors using multiple alignment

insert A

replace T with C

correlated errors—
probably caused by repeats
⇒ disentangle overlaps

In practice, error correction removes up to 98% of the errors