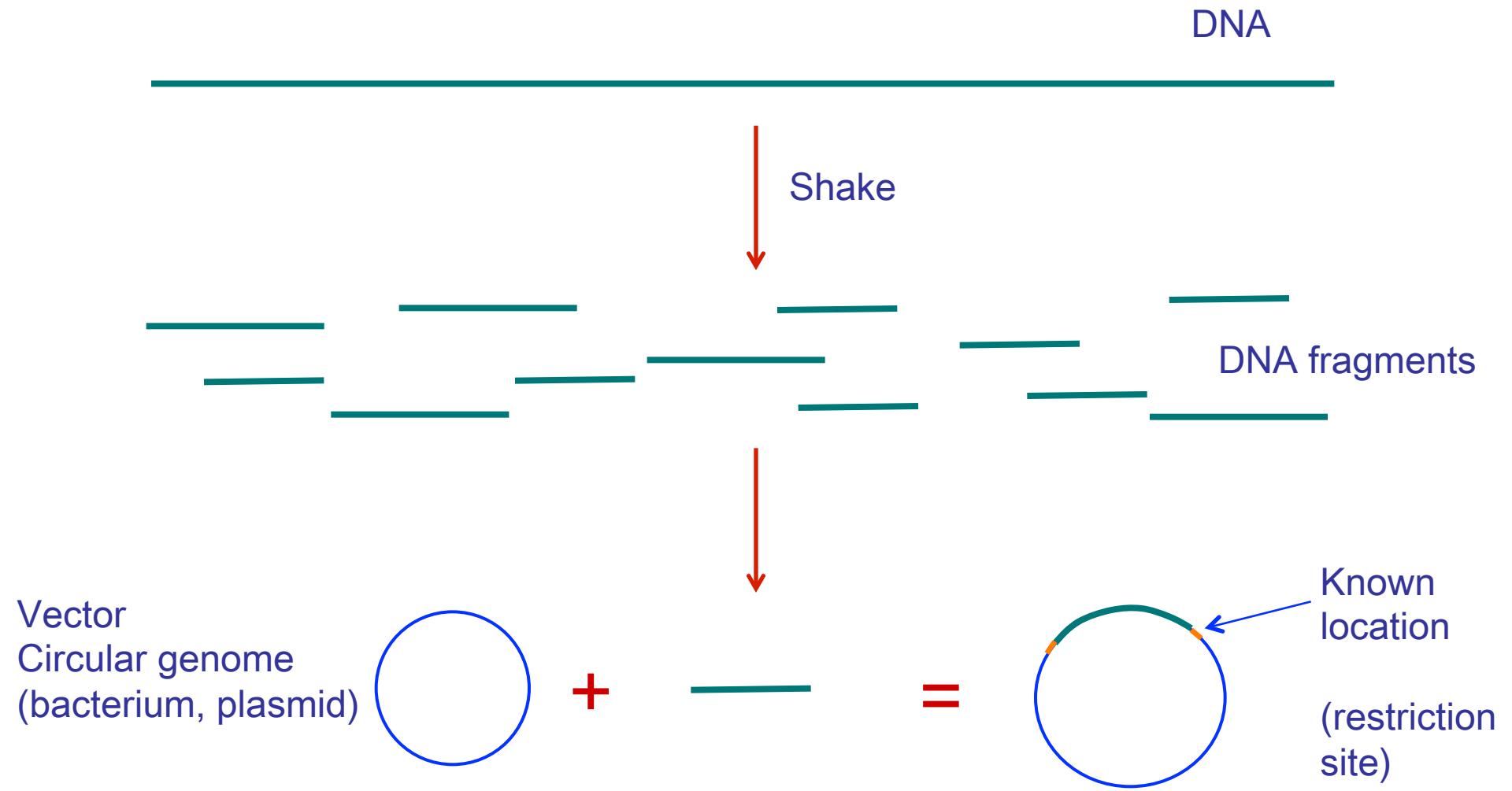
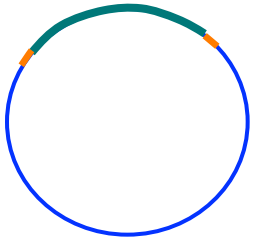
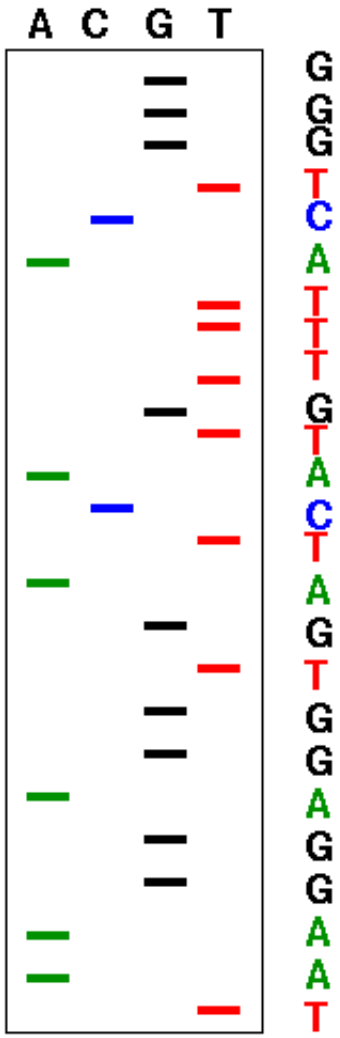
Ancient sequencing technology – Sanger Vectors

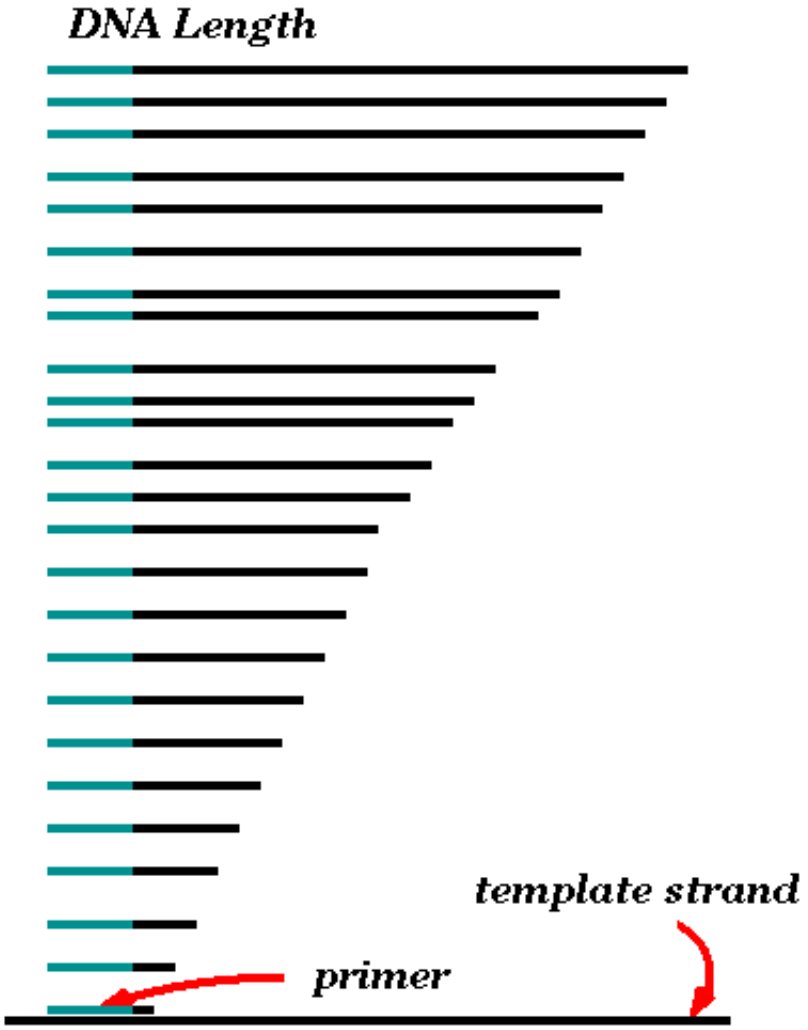


Ancient sequencing technology – Sanger Gel Electrophoresis



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





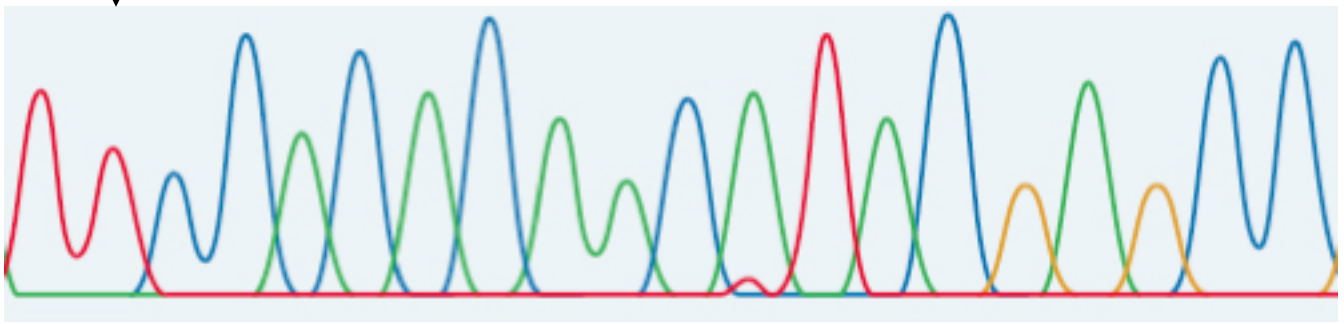
Fluorescent Sanger sequencing trace

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)

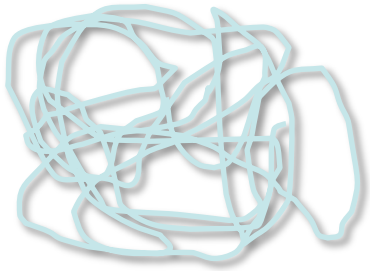
Trace



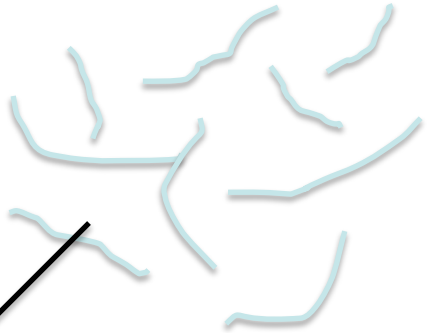
Recombinant DNA: Genes and Genomes.
3rd Edition (Dec06). WH Freeman Press.

Slide Credit: Arend Sidow

Making a Library (present)



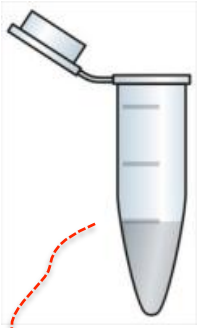
shear to ~500 bases



put on linkers

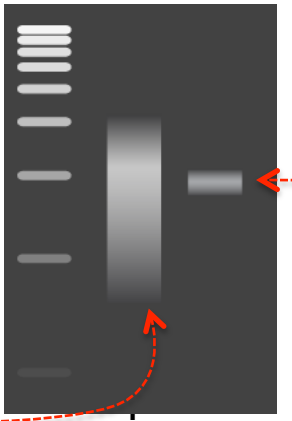
eventual forward and reverse sequence

Left handle: amplification, sequencing | "Insert" | Right handle: amplification, sequencing



PCR to obtain preparative quantities

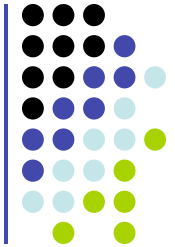
size selection on preparative gel



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

Slide Credit: Arend Sidow

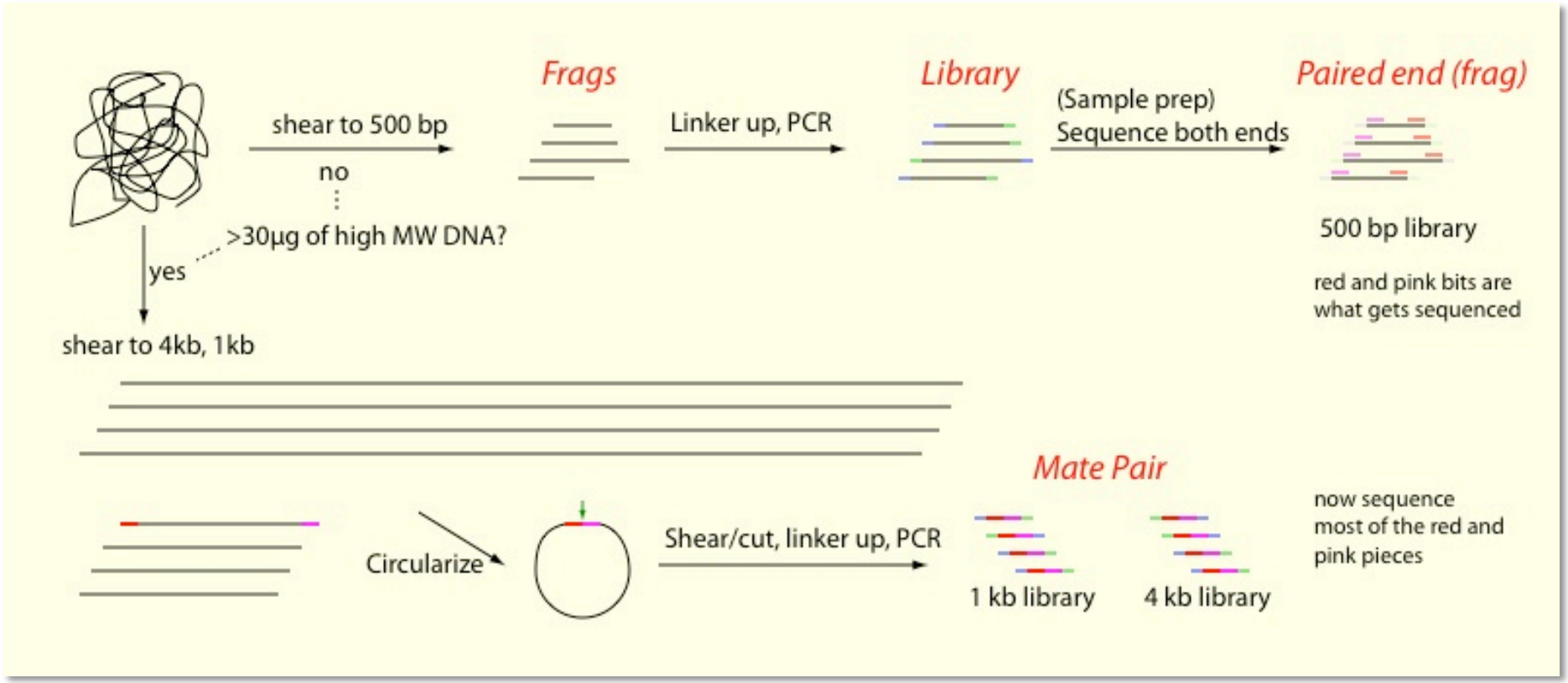
Library



- 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

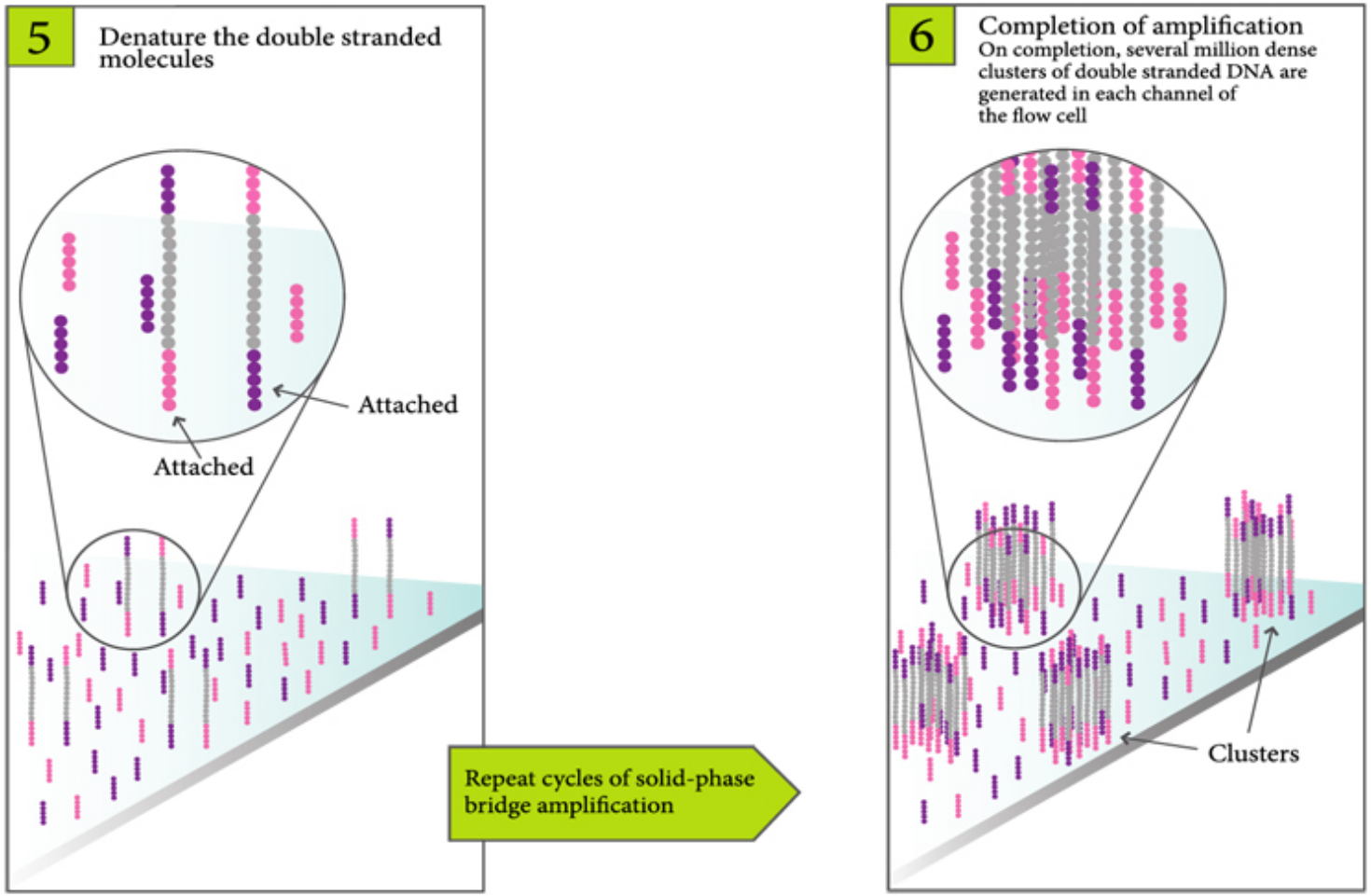
Fragment vs Mate pair ('jumping')



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

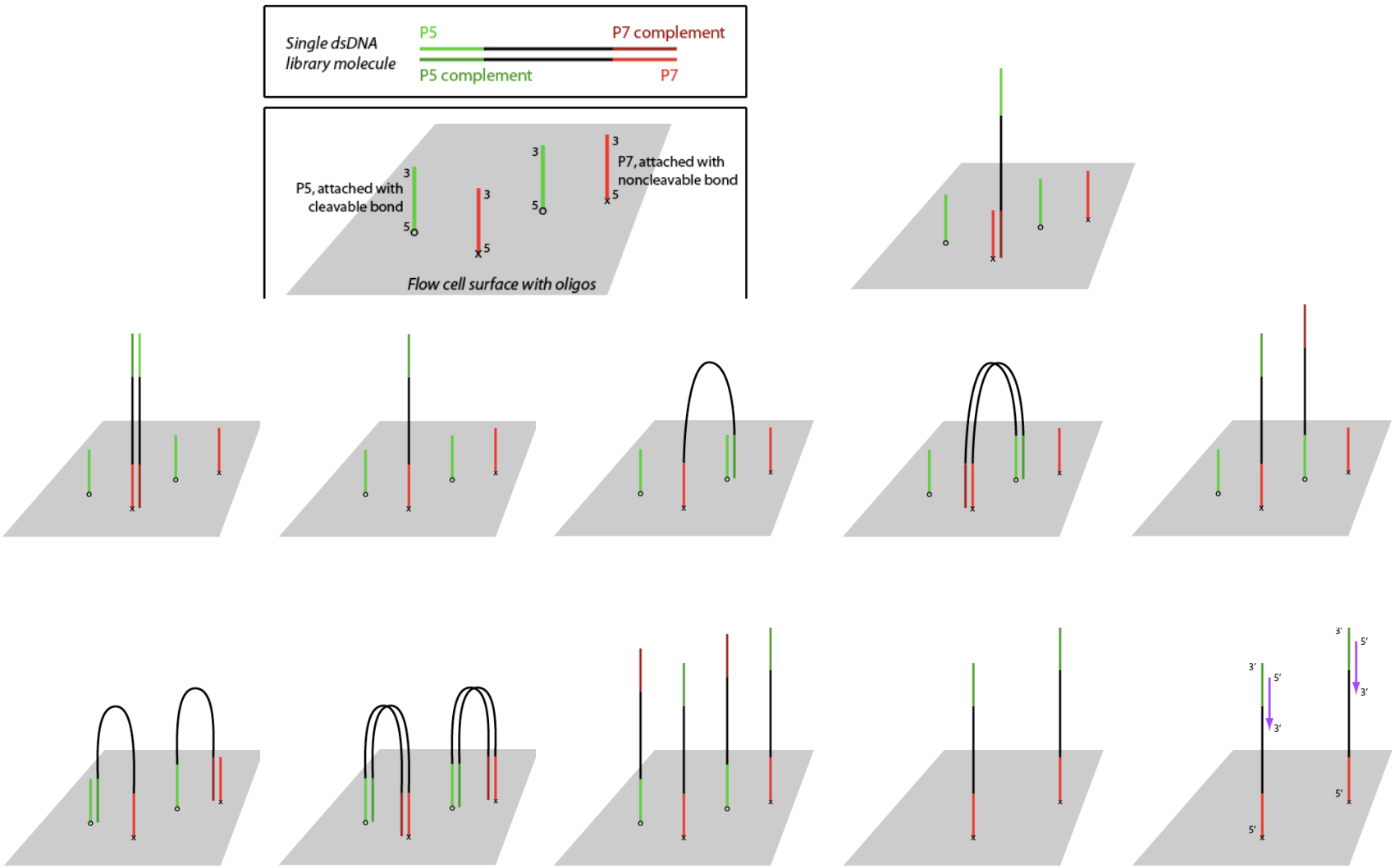
Slide Credit: Arend Sidow

Illumina cluster concept



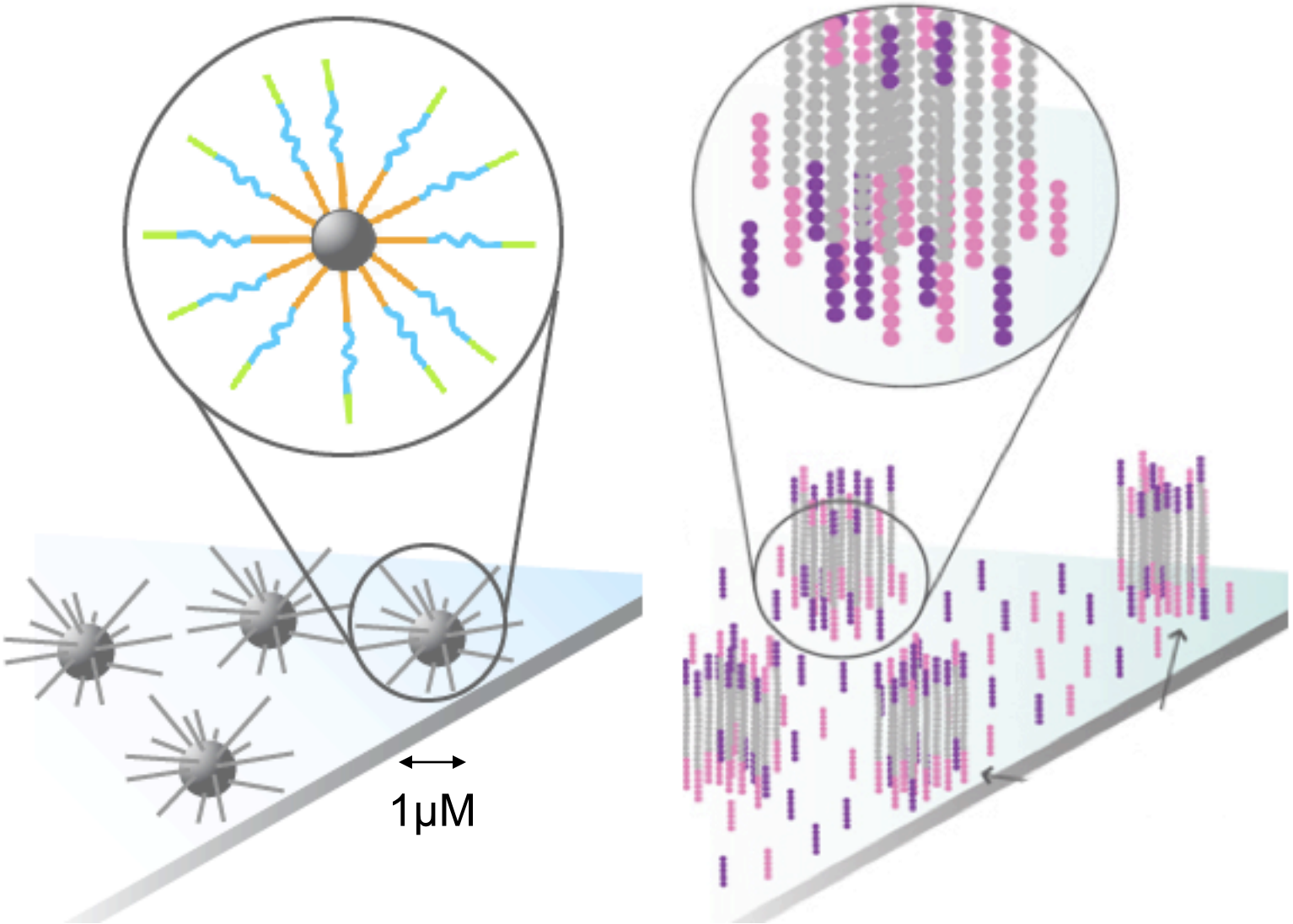
Slide Credit: Arend Sidow

Cluster generation ('bridge amplification')



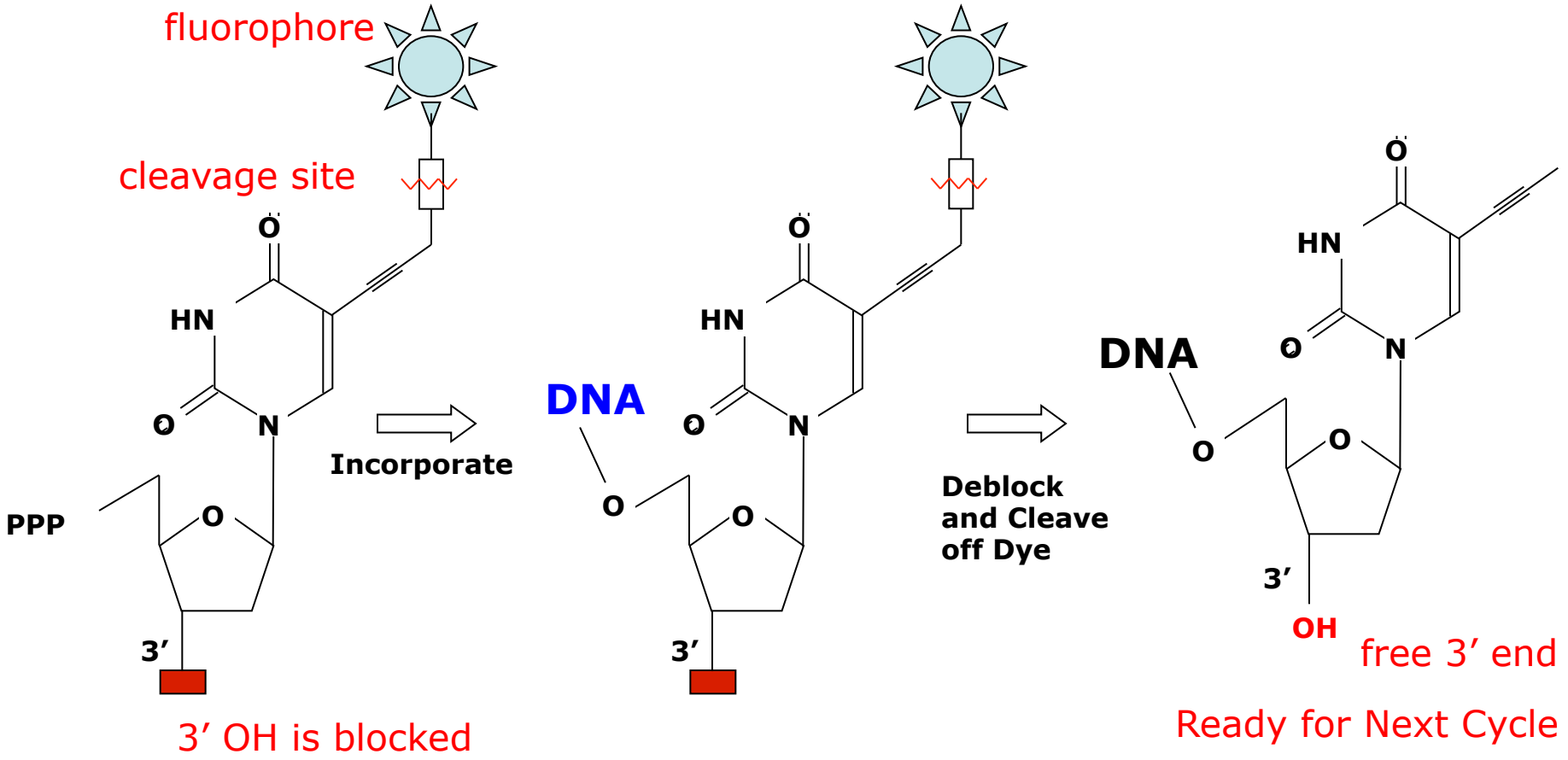
Slide Credit: Arend Sidow

Clonally Amplified Molecules on Flow Cell



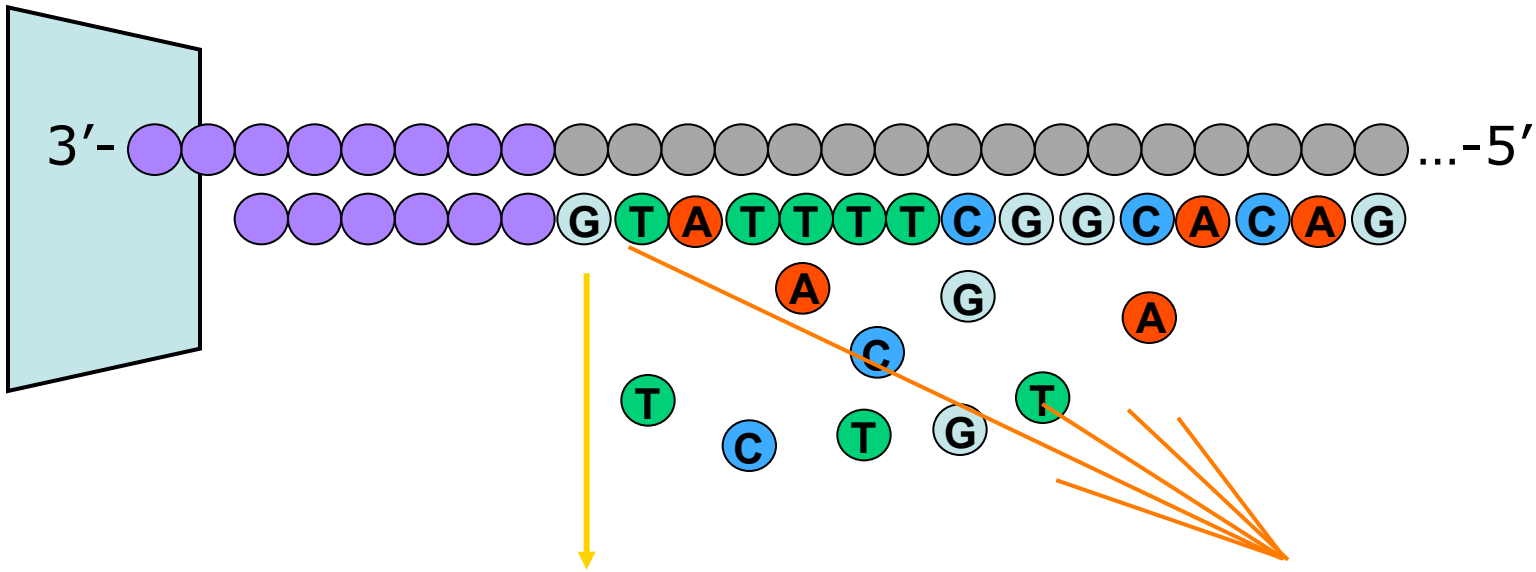
Slide Credit: Arend Sidow

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

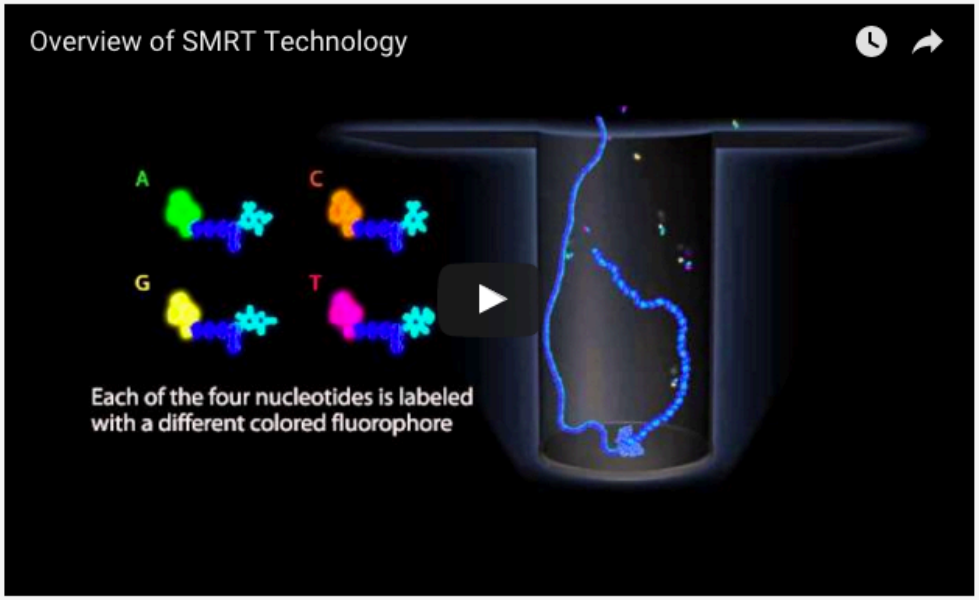
Pacific Biosciences SMRT technology

The SMRT Sequencing advantage

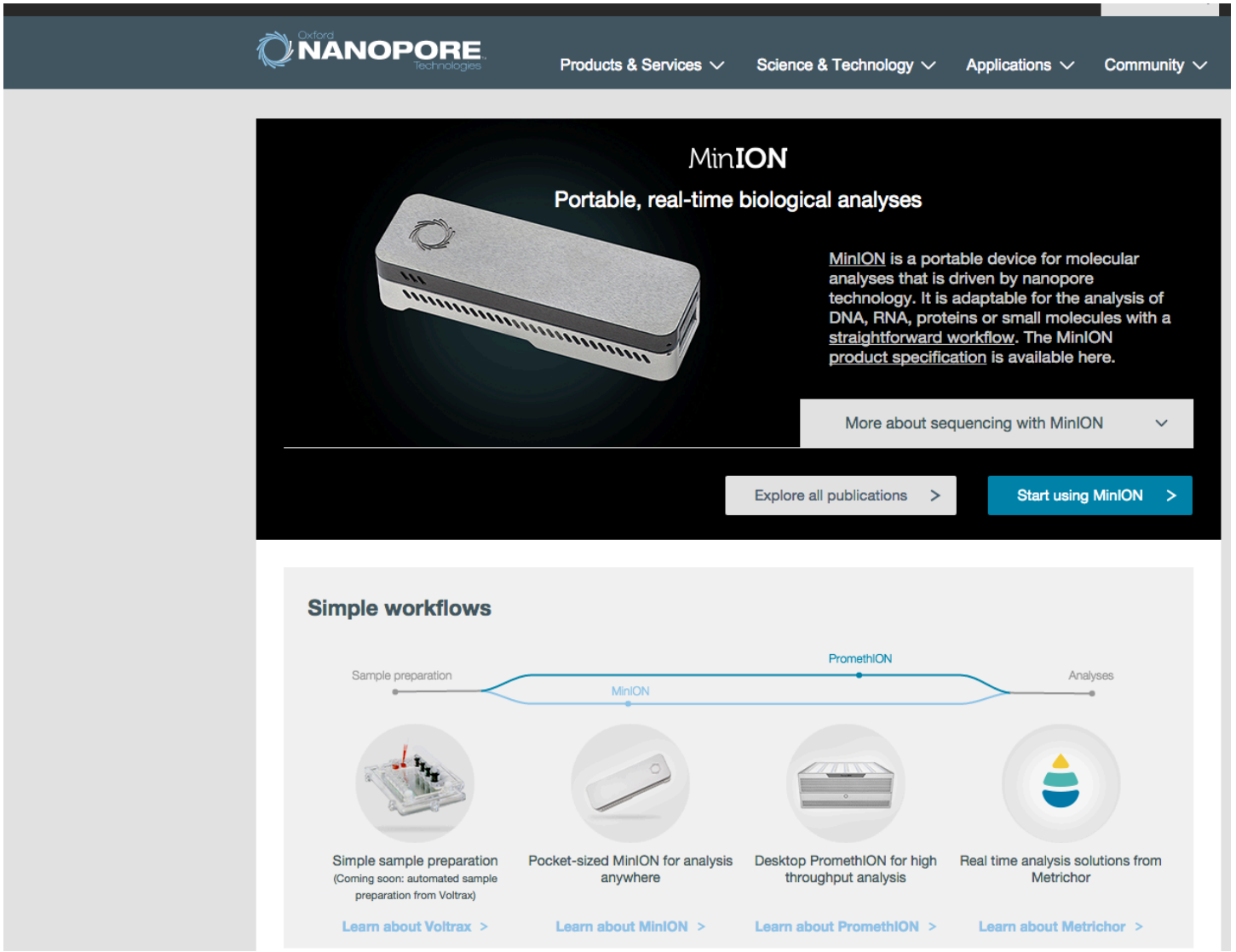
SMRT Sequencing is ideal for a variety of research applications and offers many benefits, including:

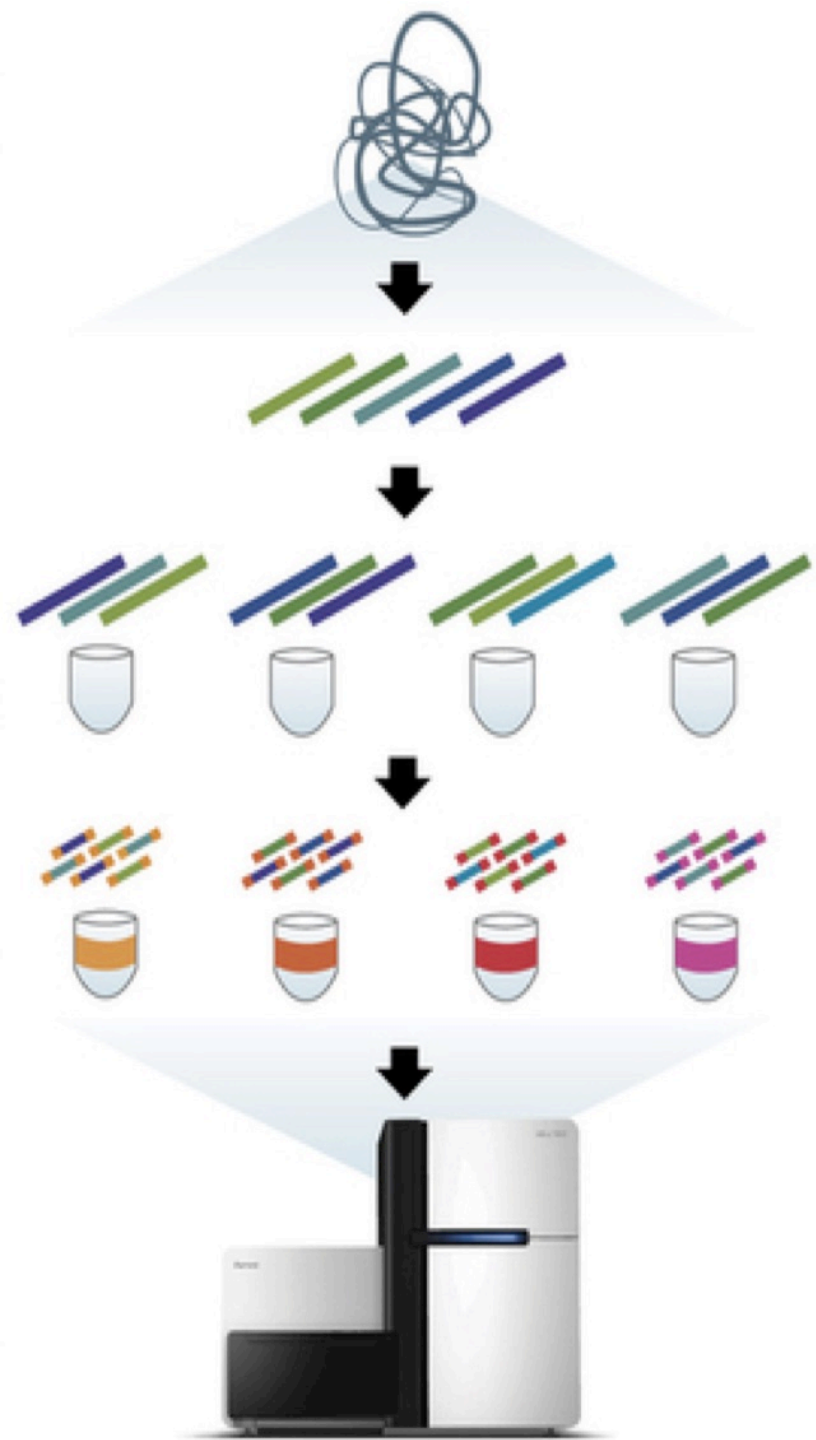
- Longest average read lengths
- Highest consensus accuracy
- Uniform coverage
- Simultaneous epigenetic characterization
- Single-molecule resolution

An overview of SMRT Sequencing



Oxford Nanopore





Moleculo Overview

1. Sample DNA is sheared into fragments of about 10 kbp

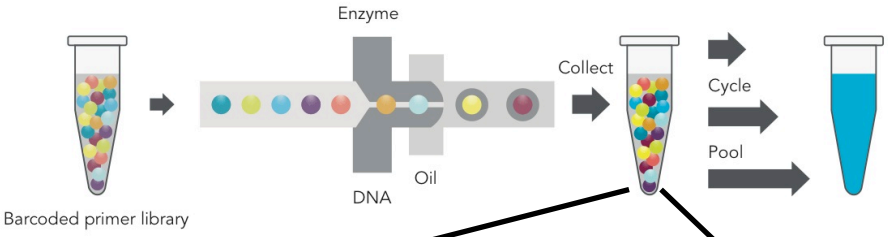
2. Fragments are diluted and placed into 384 wells

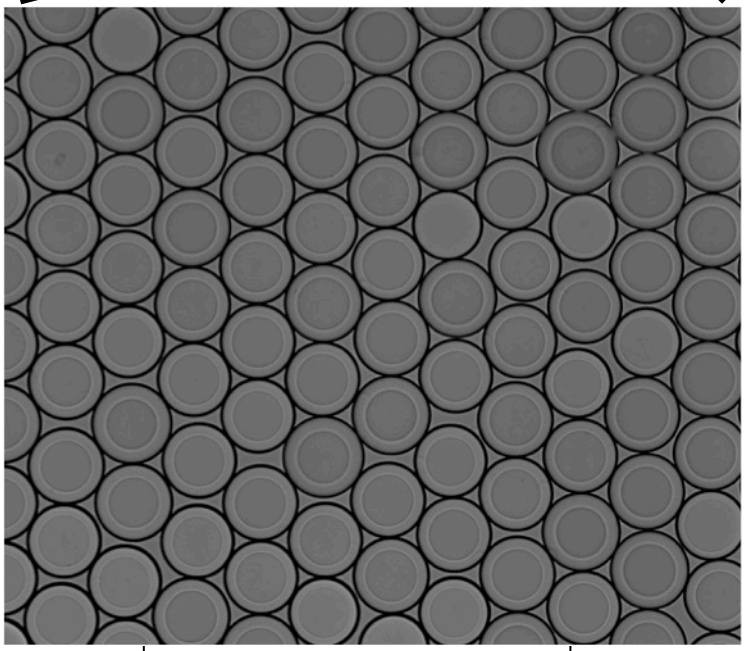
3. Fragments are amplified through long-range PCR, cut into short fragments and barcoded

4. Short fragments are pooled together and sequenced

10x System

Massively Parallel Partitioning

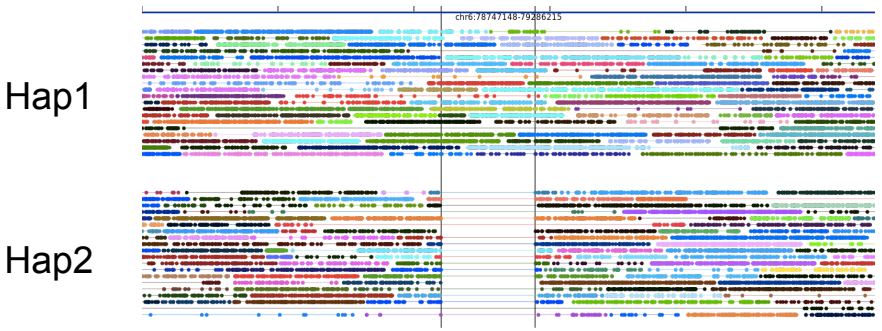




X 200,000+

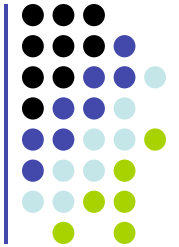
10X Instrument & Reagents

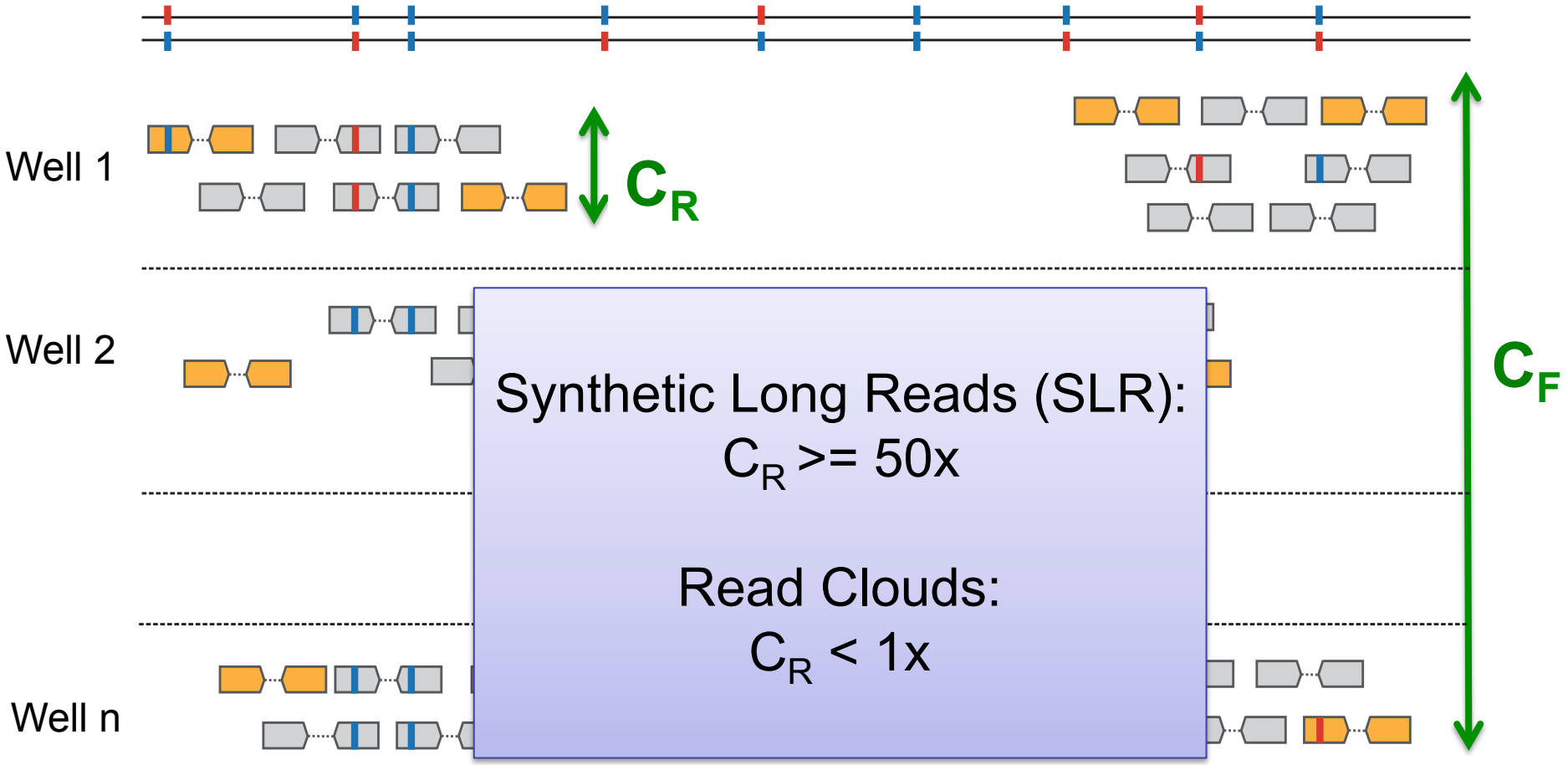
Read Clouds (“linked reads”)



Phased 60Kb deletion

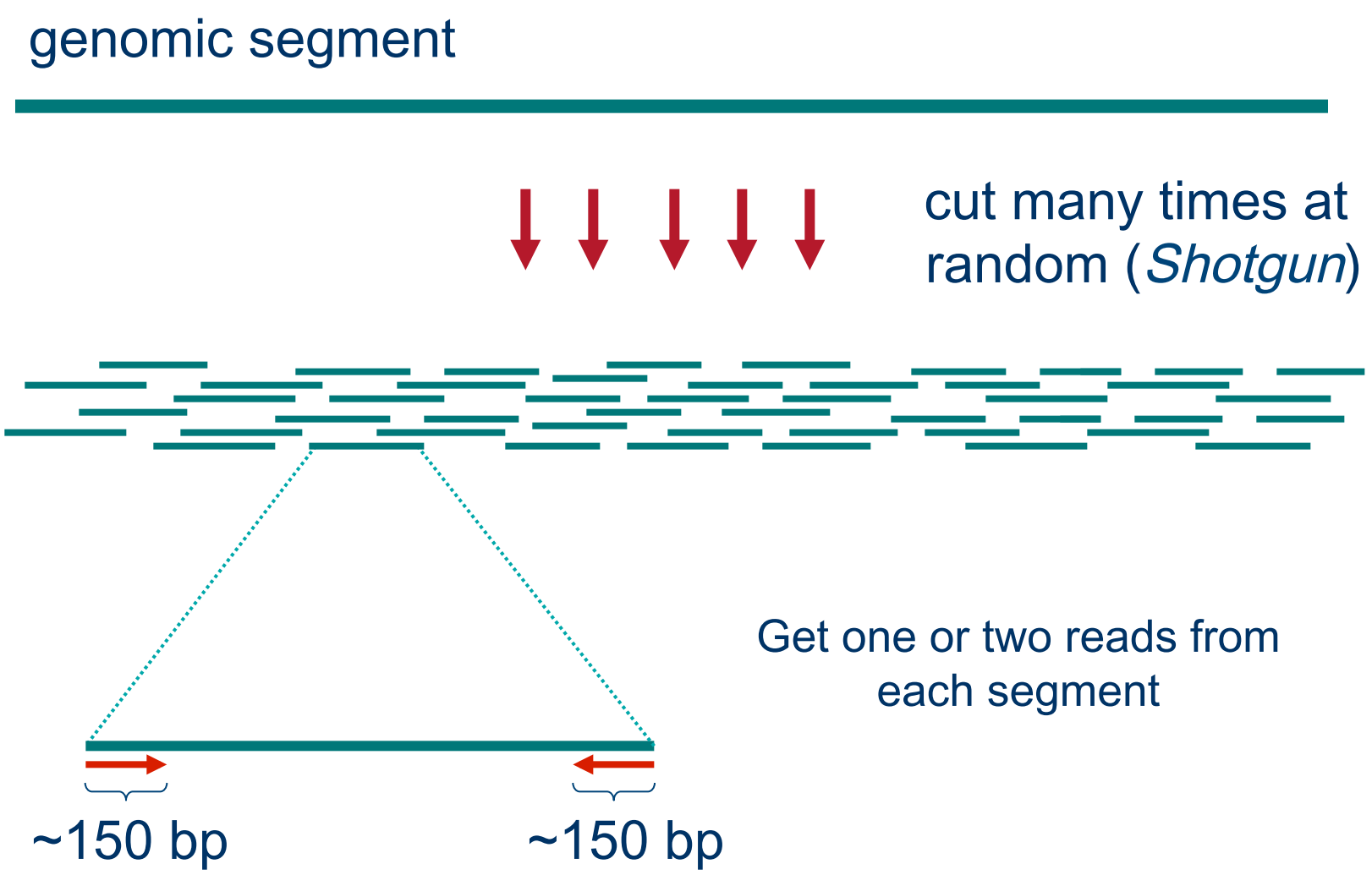
Read Clouds





Coverage = $C_F C_R$

Shotgun Sequencing

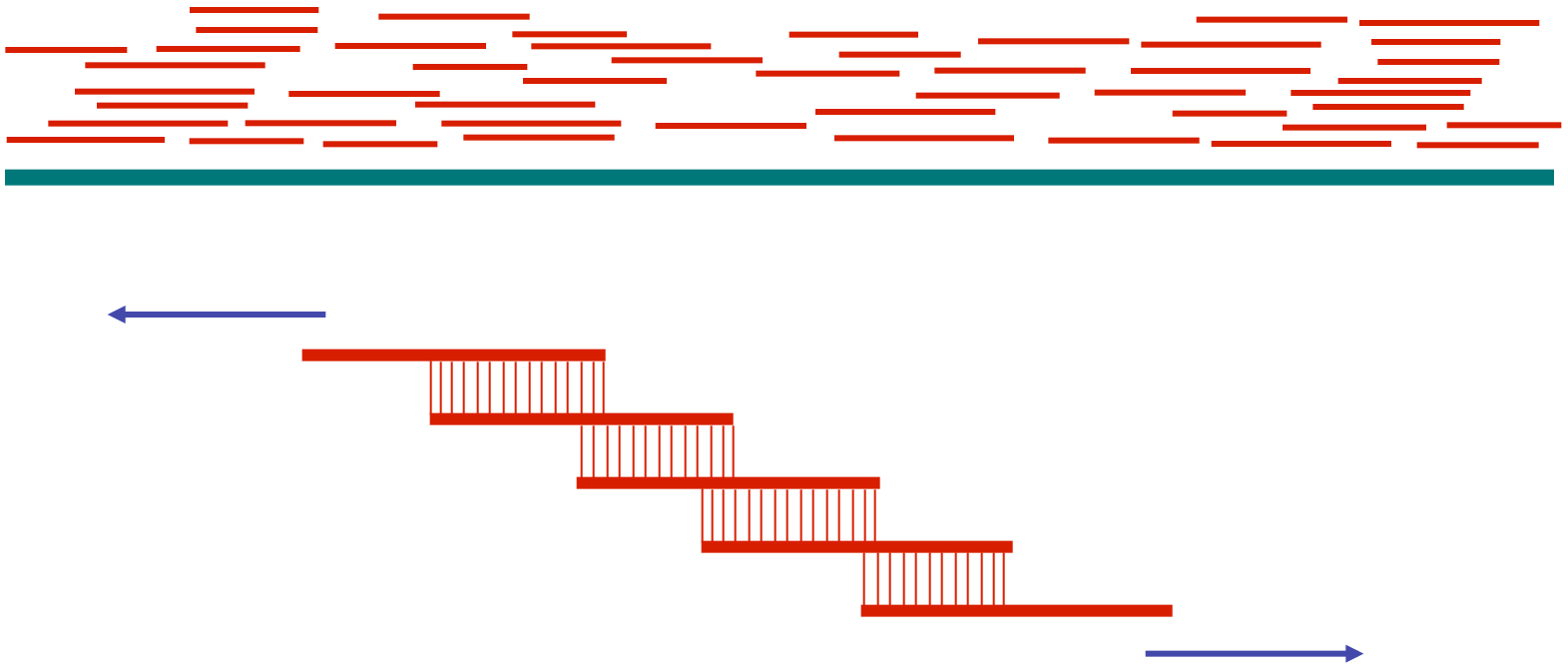


Two main assembly problems

- 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

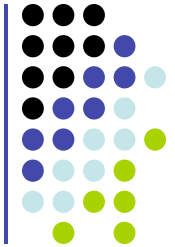
Length of genomic segment: **G**
Number of reads: **N**
Length of each read: **L**

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

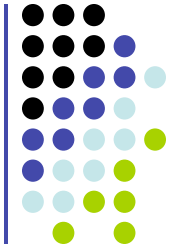
Sequencing and Fragment Assembly

$3x10^9$ nucleotides

50% of human DNA is compose

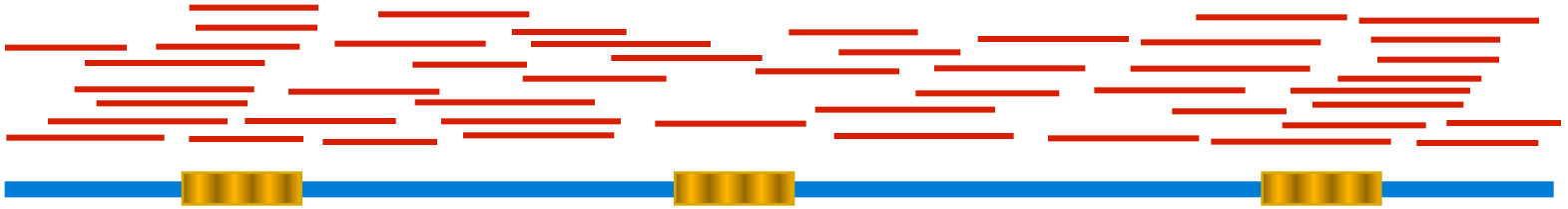
Error!
Glued together two distant regions

What can we do about repeats?



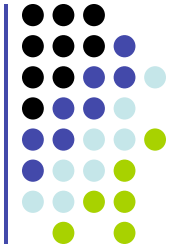
Two main approaches:

- Cluster the reads



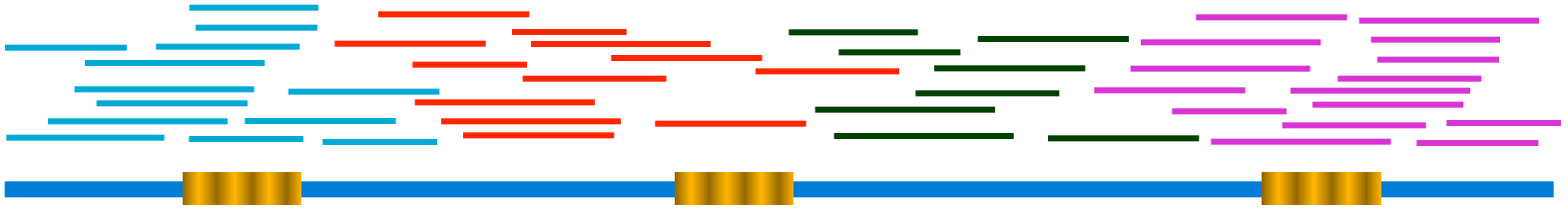
- Link the reads

What can we do about repeats?



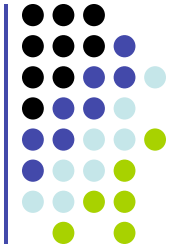
Two main approaches:

- Cluster the reads



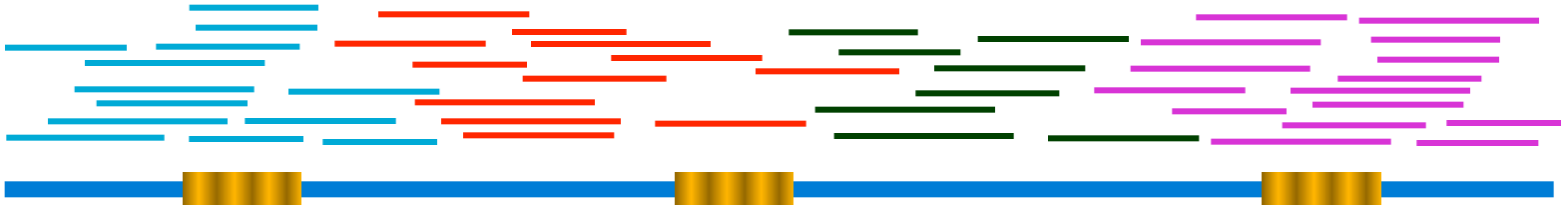
- Link the reads

What can we do about repeats?

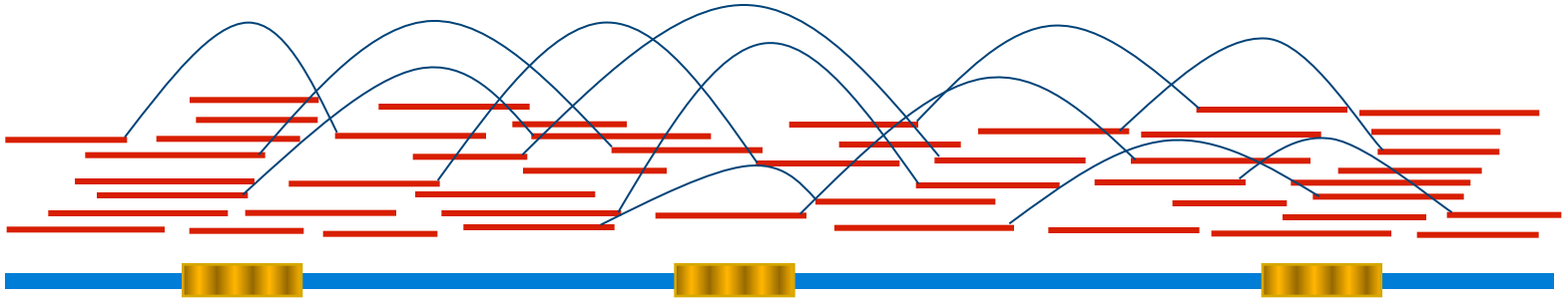


Two main approaches:

- Cluster the reads

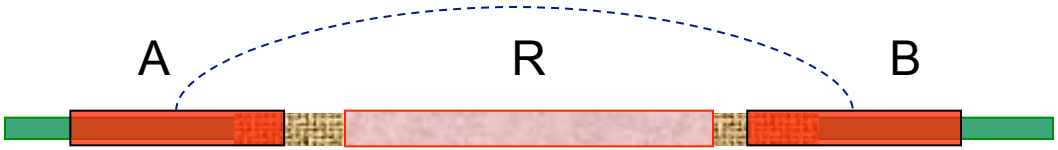


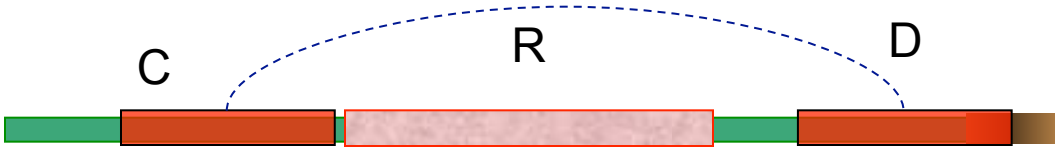
- Link the reads



Sequencing and Fragment Assembly

$3x10^9$ nucleotides





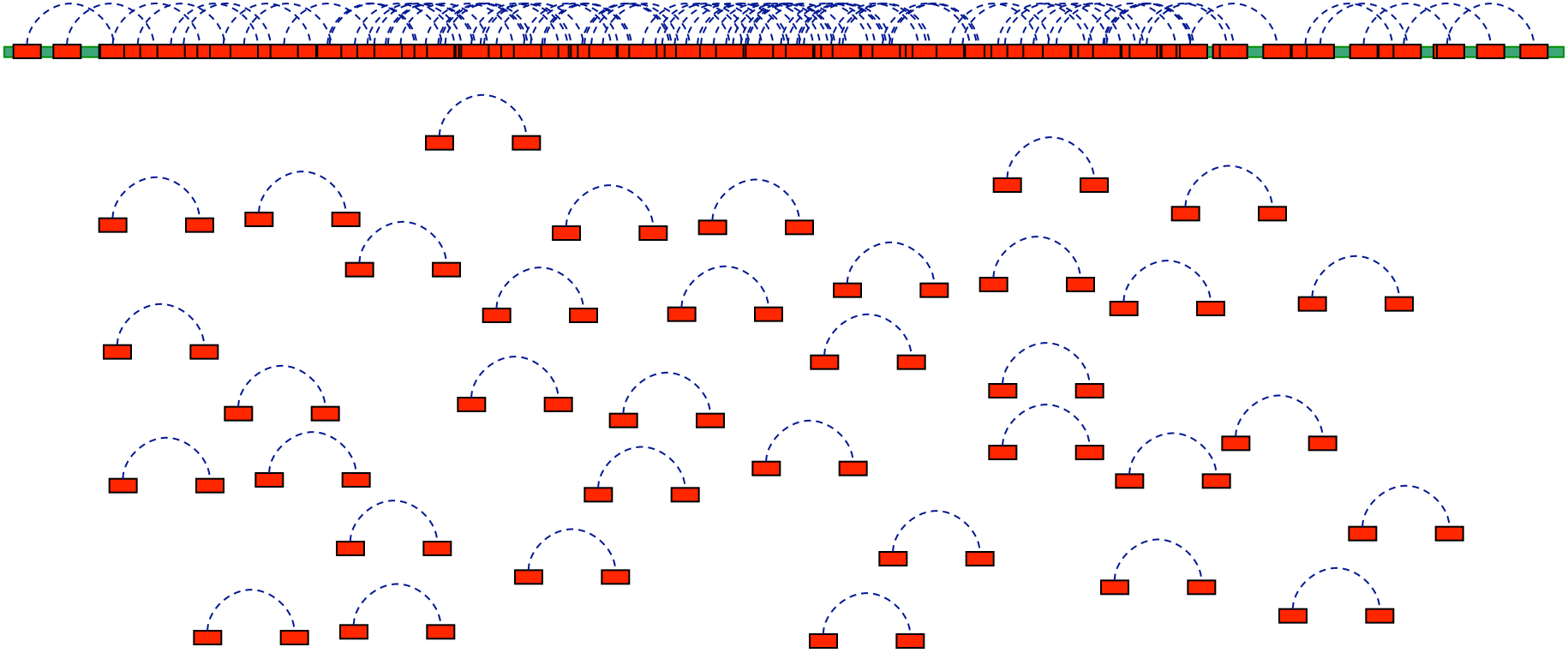
ARB, CRD

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

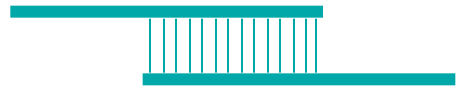
read a 500-900 long word that comes out of sequencer

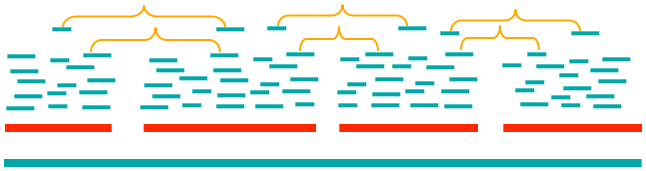
mate pair a pair of reads from two ends of the same insert fragment

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

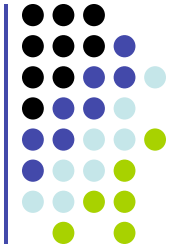
consensus sequene sequence derived from the multiple alignment of reads in a contig





..ACGATTACAATAGGTT..

1. Find Overlapping Reads

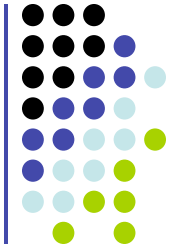


```
aaactgcagtacggatct
aaactgcag
 aactgcagt
...
         gtacggatct
          tacggatct
gggcccaaactgcagtac
gggcccaaa
 ggcccaaac
...
         actgcagta
          ctgcagtac
gtacggatctactacaca
gtacggatc
 tacggatct
...
         ctactacac
          tactacaca
```

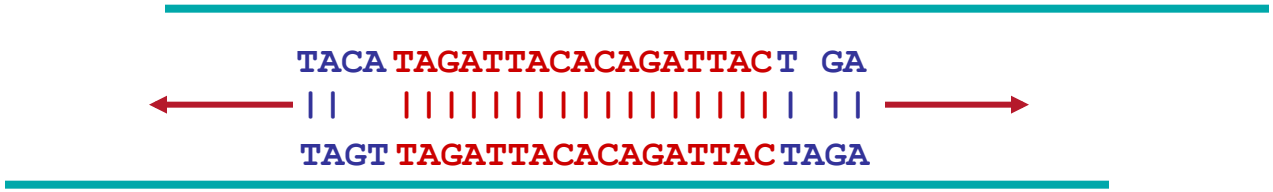
```
(read, pos., word, orient.)
aaactgcag
aactgcagt
actgcagta
...
gtacggatc
tacggatct
gggcccaaa
ggcccaaac
gcccaaact
...
actgcagta
ctgcagtac
gtacggatc
tacggatct
acggatcta
...
ctactacac
tactacaca
```

```
(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

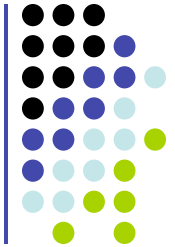


- Find pairs of reads sharing a k-mer, k ~ 24
- Extend to full alignment – throw away if not >98% similar



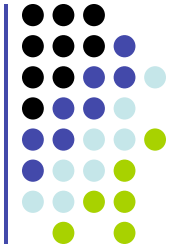
- Caveat: repeats
  - A k-mer that occurs N times, causes $O(N^2)$ read/read comparisons
  - ALU k-mers could cause up to $1{,}000{,}000^2$ comparisons
- Solution:
  - Discard all k-mers that occur “too often”
    - Set cutoff to balance sensitivity/speed tradeoff, according to genome at hand and computing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

1. Find Overlapping Reads



Create local multiple alignments from the overlapping reads

1. Find Overlapping Reads



- Correct errors using multiple alignment

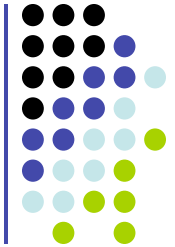
insert A

replace T with C

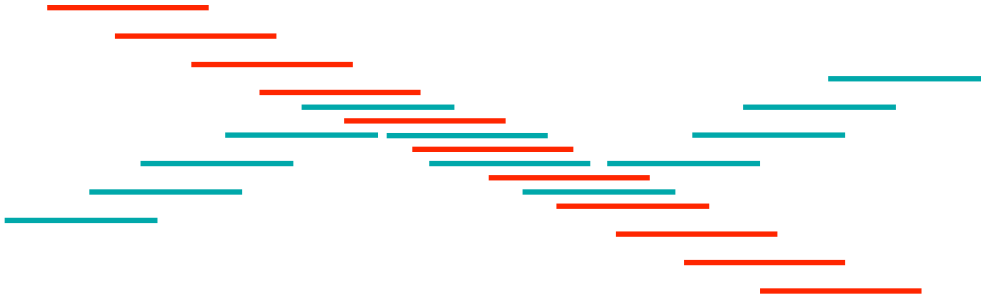
correlated errors—
probably caused by repeats
⇒ disentangle overlaps

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

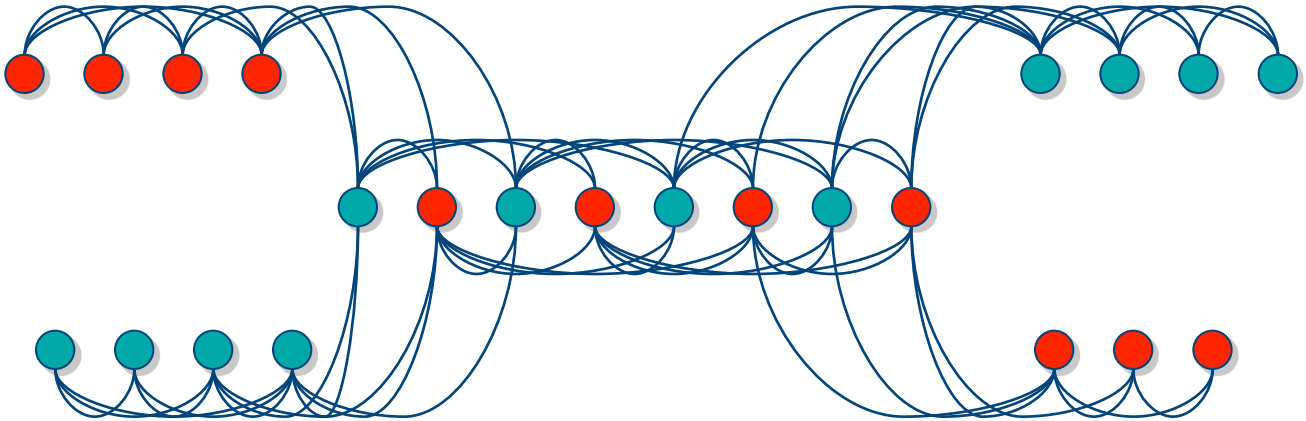
2. Merge Reads into Contigs



- Overlap graph:
  - Nodes: reads r_1…..r_n
  - Edges: overlaps (r_i, r_j, shift, orientation, score)

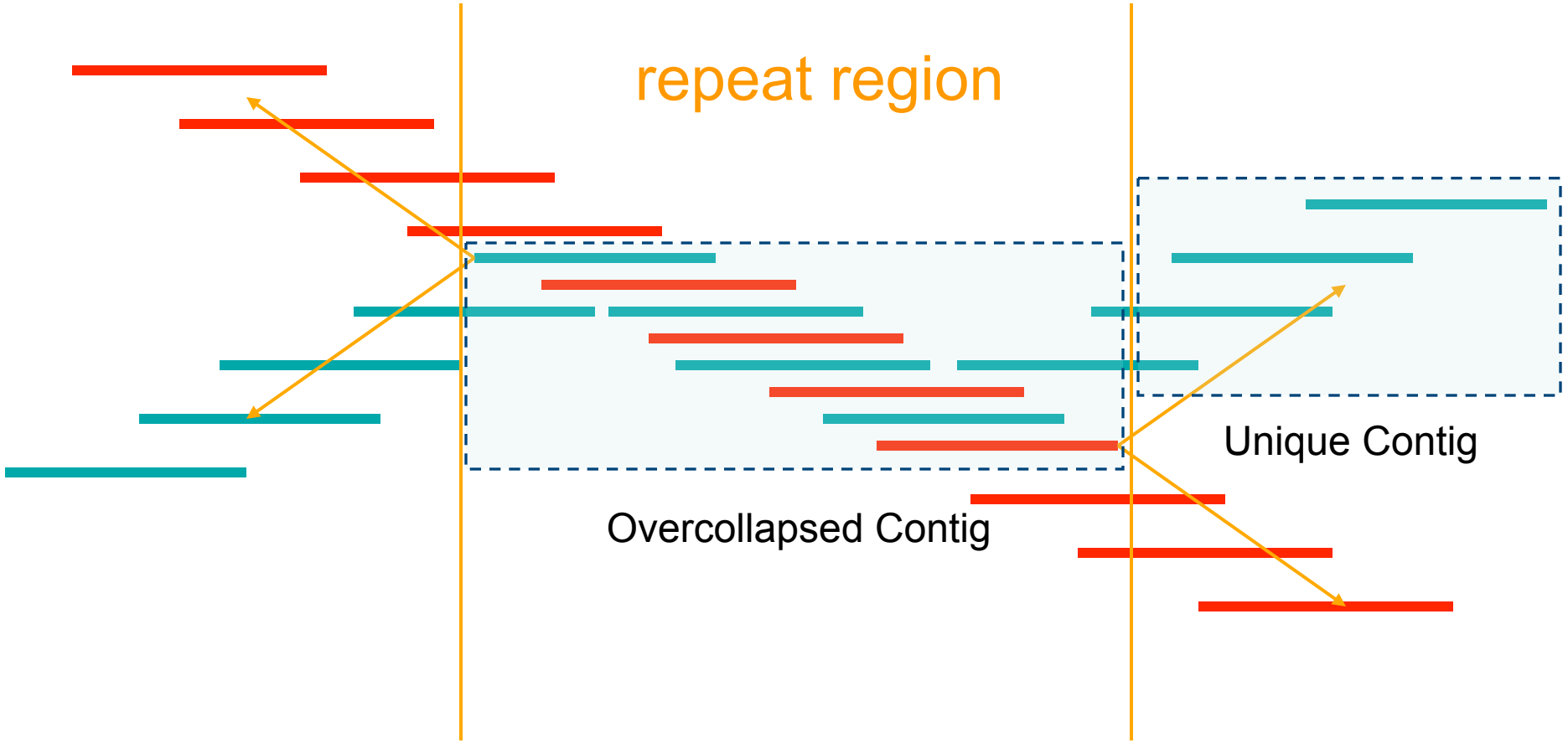


Reads that come from two regions of the genome (blue and red) that contain the same repeat



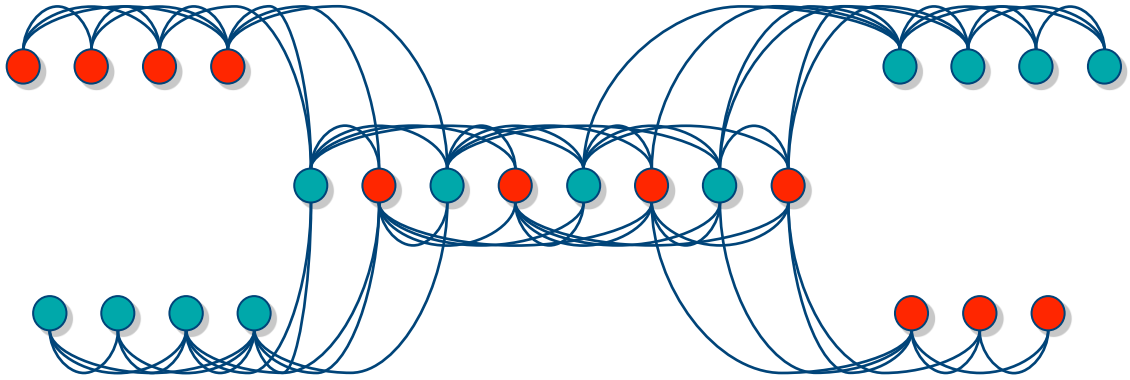
Note:
of course, we don't know the "color" of these nodes

2. Merge Reads into Contigs

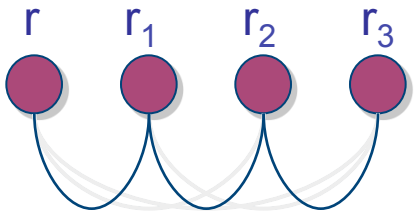


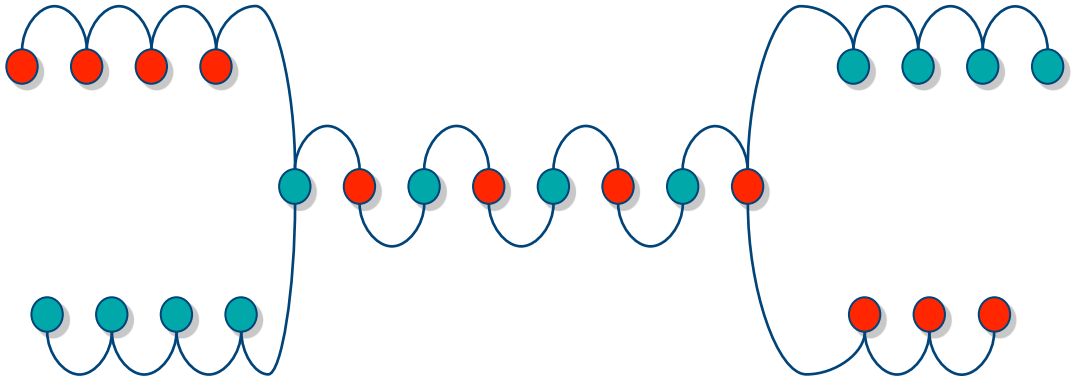
We want to merge reads up to potential repeat boundaries

2. Merge Reads into Contigs

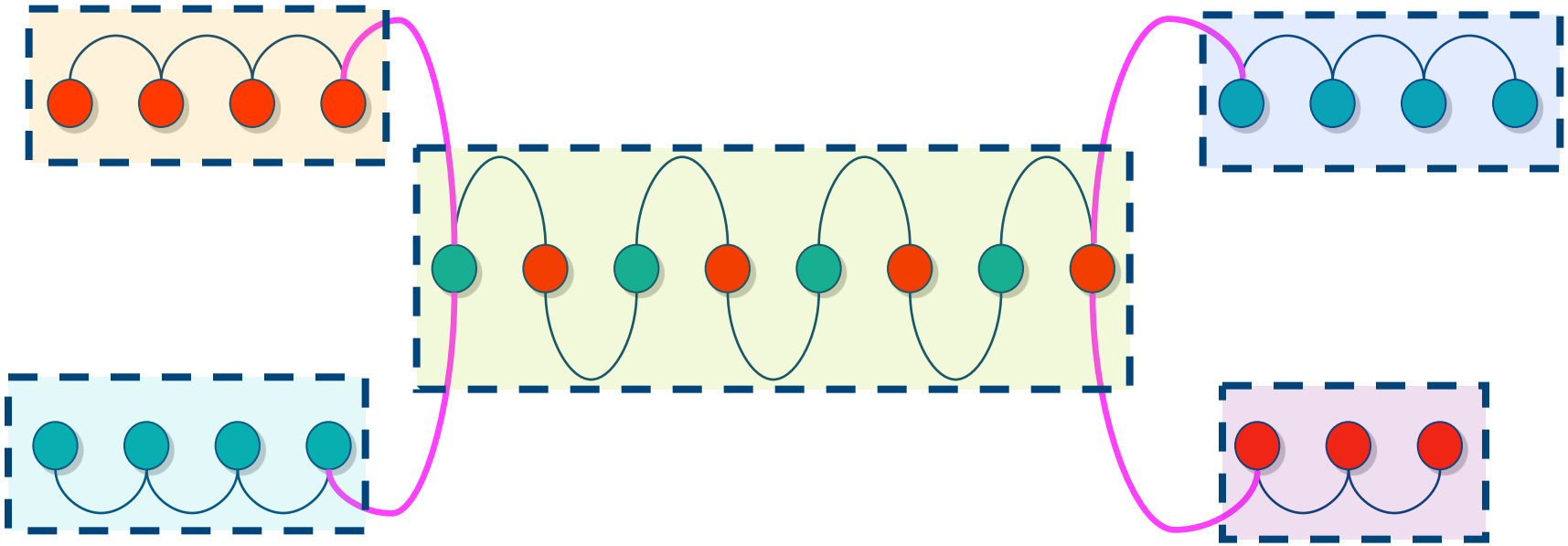


- Remove transitively inferable overlaps
  - If read r overlaps to the right reads r_1, r_2, and r_1 overlaps r_2, then (r, r_2) can be inferred by (r, r_1) and (r_1, r_2)

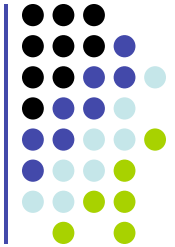




2. Merge Reads into Contigs



Repeats, errors, and contig lengths



- Repeats shorter than read length are easily resolved
  - Read that spans across a repeat disambiguates order of flanking regions
- Repeats with more base pair diffs than sequencing error rate are OK
  - We throw overlaps between two reads in different copies of the repeat
- To make the genome **appear** less repetitive, try to:
  - Increase read length
  - Decrease sequencing error rate

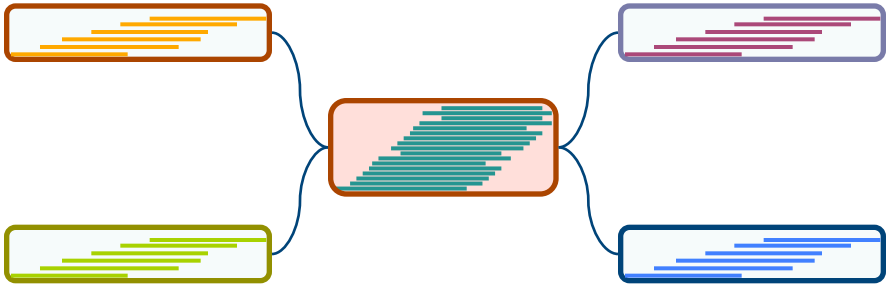
Role of error correction:

Discards up to 98% of single-letter sequencing errors
decreases error rate
⇒ decreases effective repeat content
⇒ increases contig length

3. Link Contigs into Supercontigs

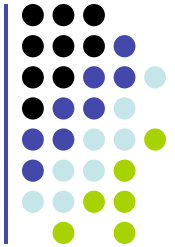
Normal density

Too dense
⇒ Overcollapsed



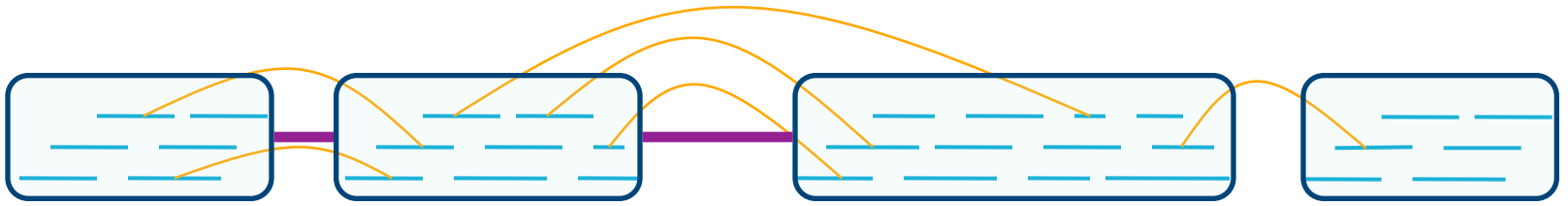
Inconsistent links
⇒ Overcollapsed?

3. Link Contigs into Supercontigs



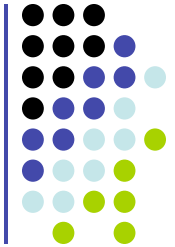
Find all links between unique contigs

Connect contigs incrementally, if ≥ 2 forward-reverse links



supercontig
(aka *scaffold*)

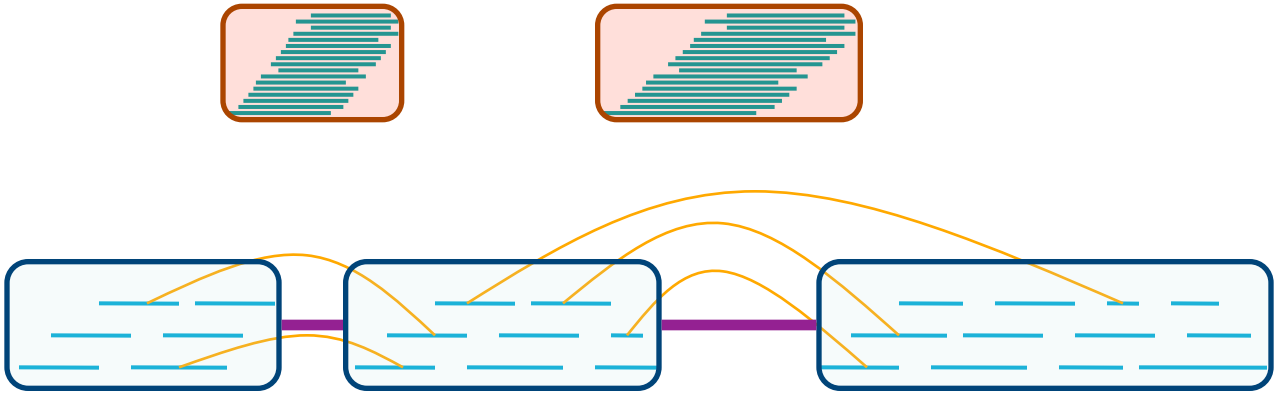
3. Link Contigs into Supercontigs



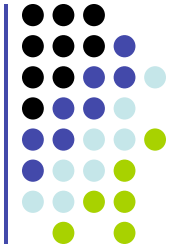
Fill gaps in supercontigs with paths of repeat contigs

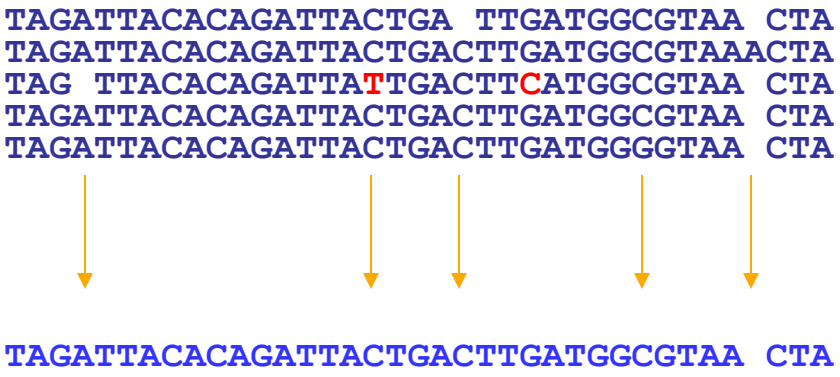
Complex algorithmic step

- Exponential number of paths
- Forward-reverse links



4. Derive Consensus Sequence





Derive multiple alignment from pairwise read alignments

Derive each consensus base by weighted voting

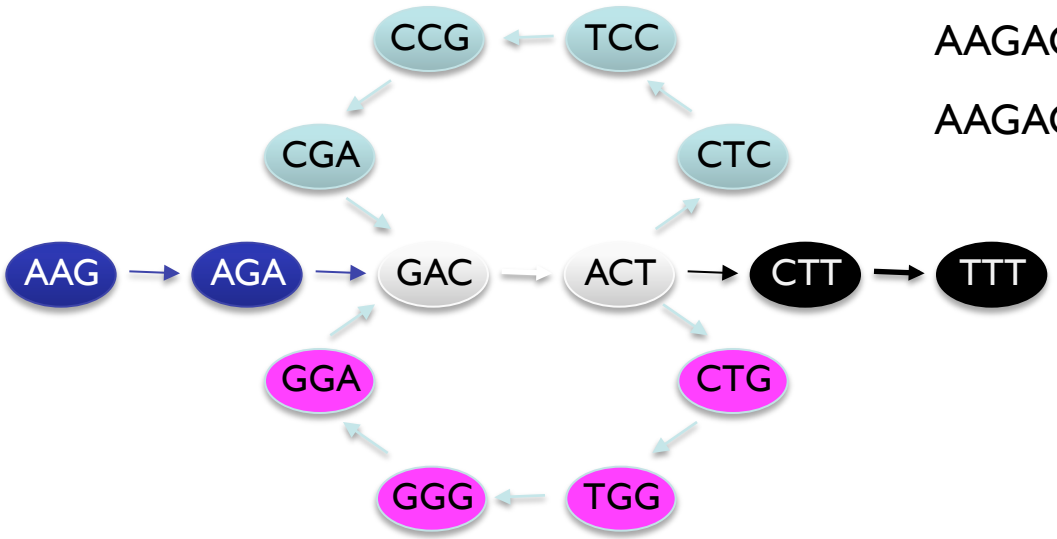
(**Alternative:** take maximum-quality letter)

De Brujin Graph formulation

- Given sequence $x_1 \ldots x_N$, k-mer length k,
  Graph of 4^k vertices,
  Edges between words with (k-1)-long overlap

Reads

AAGA
ACTT
ACTC
ACTG
AGAG
CCGA
CGAC
CTCC
CTGG
CTTT
…



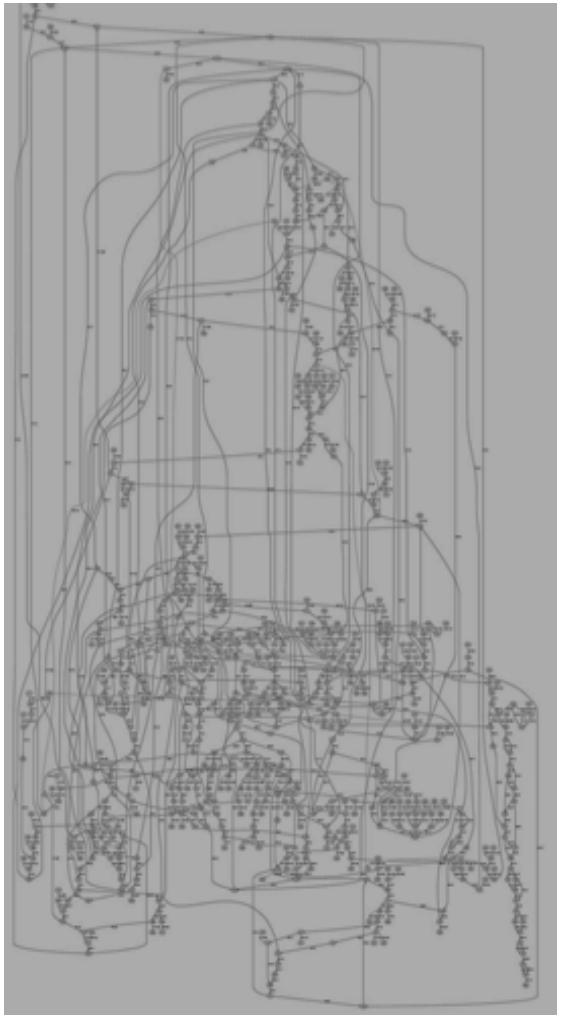
Potential Genomes

AAGACTCCGACTGGGACTTT

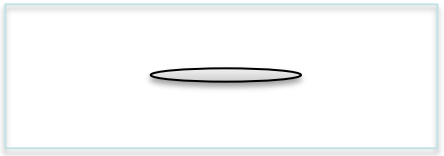
AAGACTGGGACTCCGACTTT

Slide by Michael Schatz

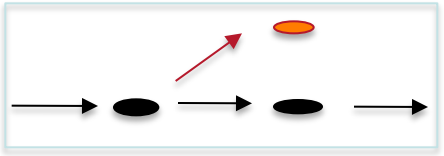
Node Types



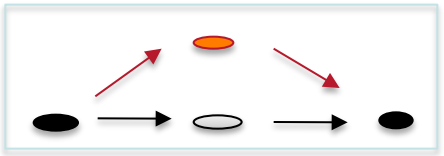
(Chaisson, 2009)



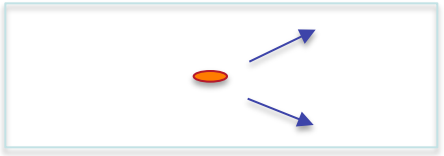
Isolated nodes (10%)



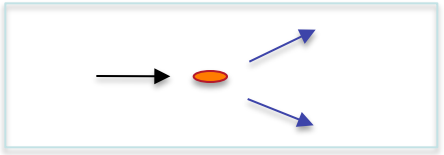
Tips (46%)



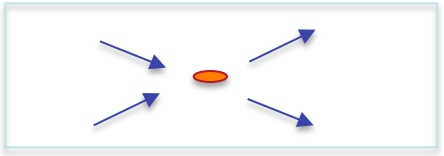
Bubbles/Non-branch (9%)



Dead Ends (.2%)



Half Branch (25%)

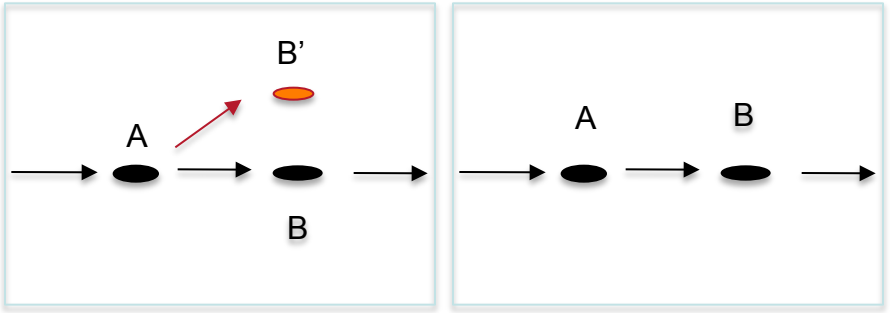


Full Branch (10%)

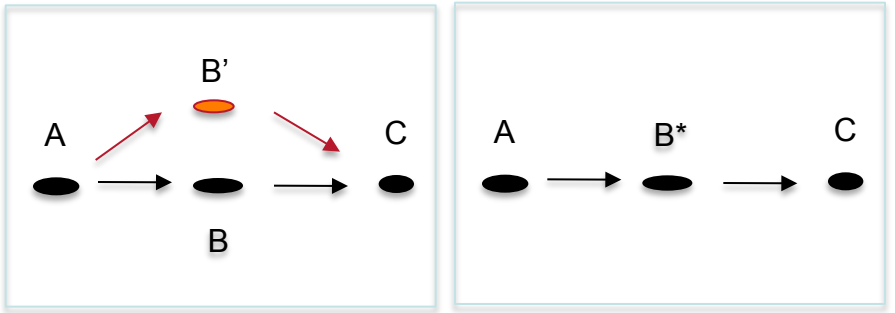
Slide by Michael Schatz

Error Correction

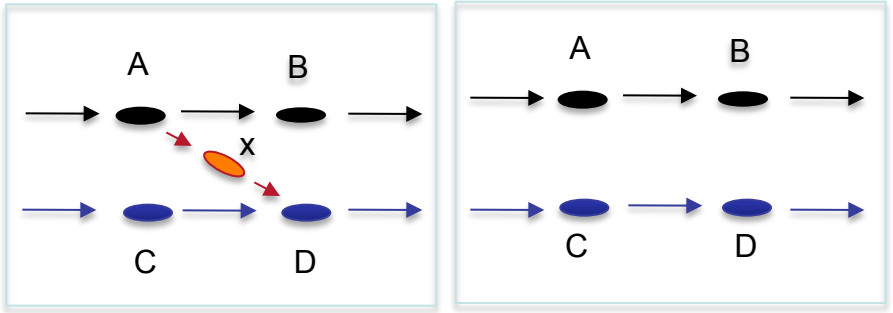
- Errors at end of read
  - Trim off ‘dead-end’ tips



- Errors in middle of read
  - Pop Bubbles



- Chimeric Edges
  - Clip short, low coverage nodes



Slide by Michael Schatz

De Brujin Graph formulation