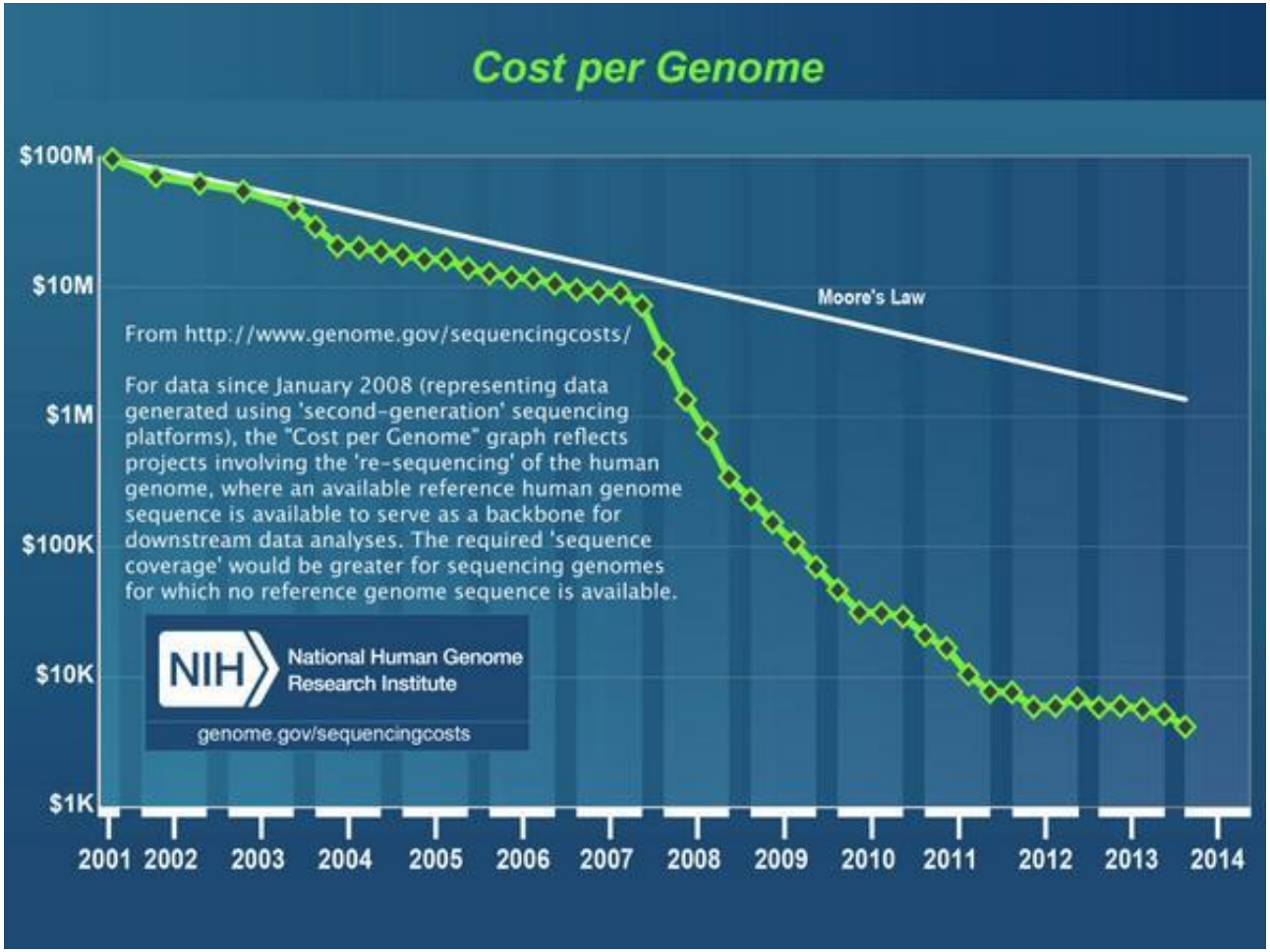
High Throughput Sequencing Technologies

J Fass
UCD Genome Center Bioinformatics Core
Monday March 23, 2015

Sequencing Explosion

www.genome.gov/sequencingcosts
http://t.co/Ka5cVGhdqo



Sequencing Explosion

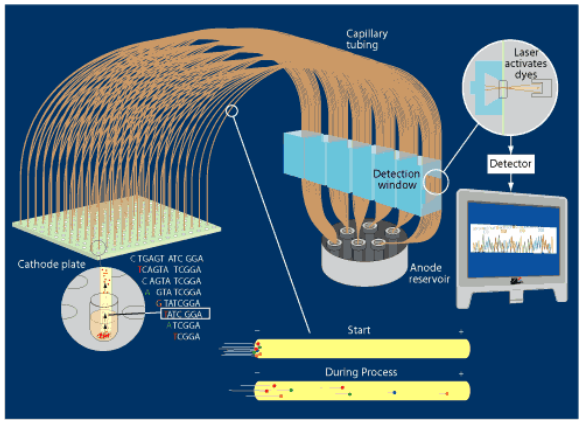
adapted from Mardis 2011 Nature 470:198

Current Sequencing Technologies

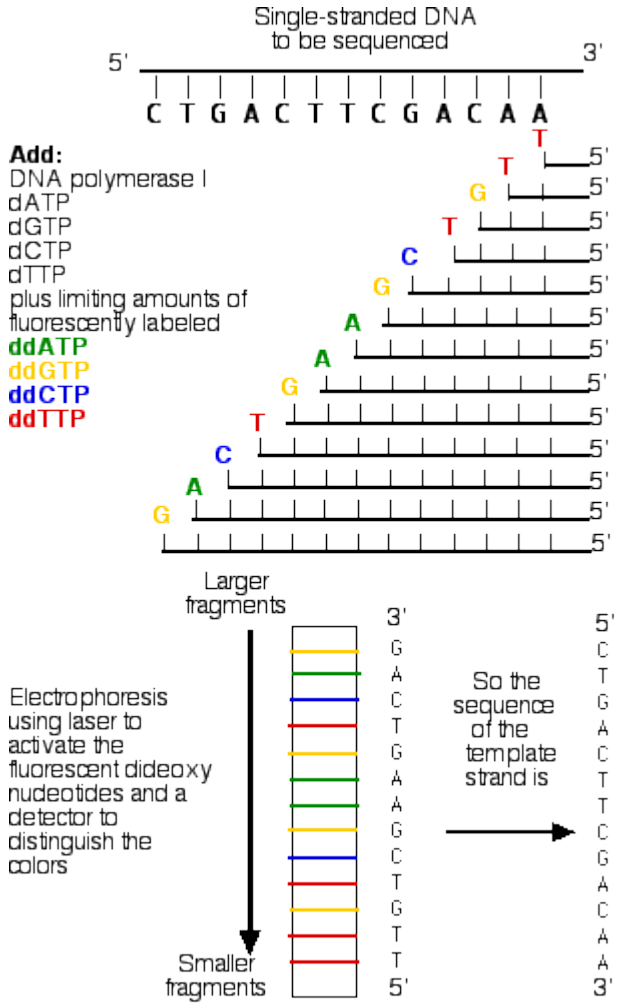
- Sanger
- *(Roche) 454*
- Illumina
- *SOLiD*
- PacBio
- *Ion Torrent*
- *Complete Genomics … now owned by BGI*
- *(Illumina) Moleculo = "TruSeq Synthetic Long Reads"*
- Oxford Nanopore
- *BioNano Genomics*
- 10X Genomics
- Dovetail Genomics

Sanger Sequencing

- ddNTP's (with fluorescent labels) incorporated (along with unlabeled dNTP's) in amplification step, resulting in some molecules terminated *at every position*
- Gel / capillary electrophoresis orders molecules by length
- Fluorescent label (color) indicates terminal base identity at each position
- Read colors, in order, to derive sequence



http://www.jgi.doe.gov/sequencing/education/how/how_10.html



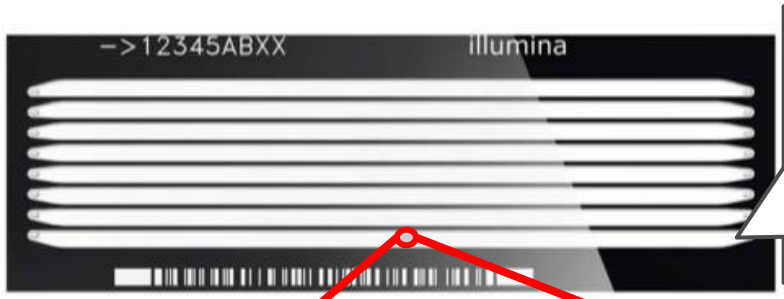
http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/DNAsequencing.html

Illumina

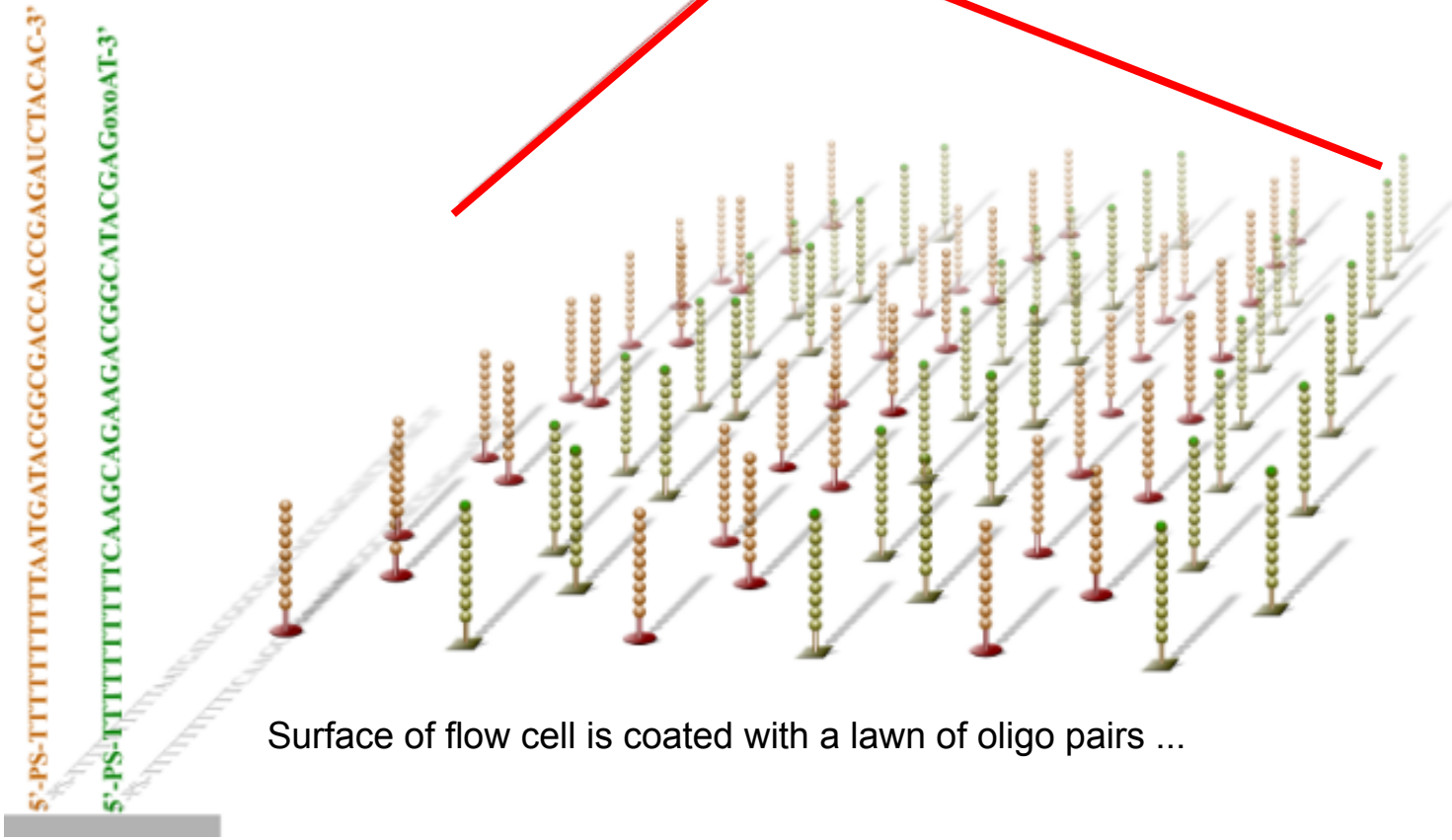
Illumina MiSeq

Illumina HiSeq 2000 / 2500

Illumina



8 "Lanes"
(two flow cells per HiSeq)
(one-lane FC for MiSeq)

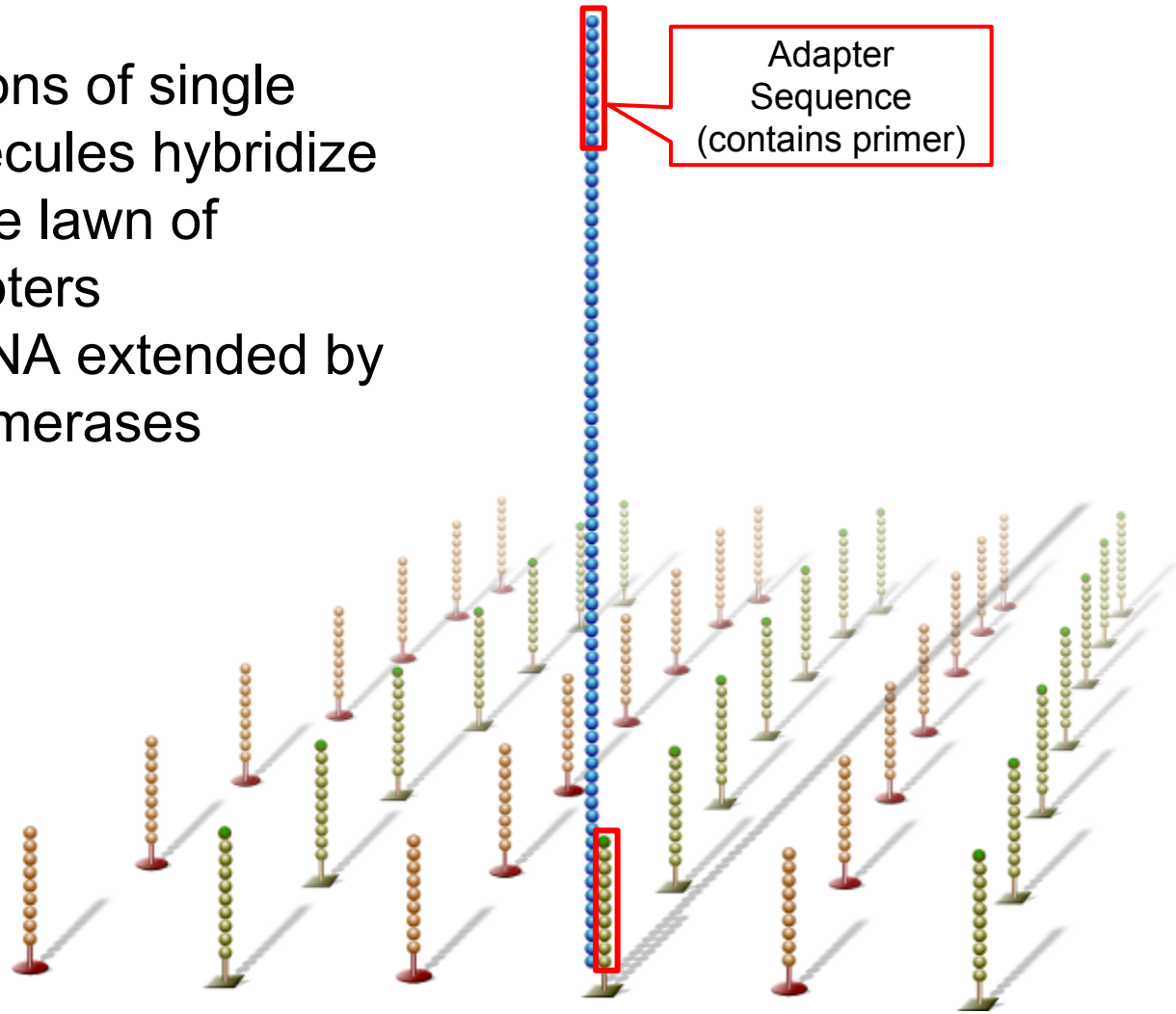


Surface of flow cell is coated with a lawn of oligo pairs ...

Illumina

Cluster Generation:
Hybridize Fragments & Extend

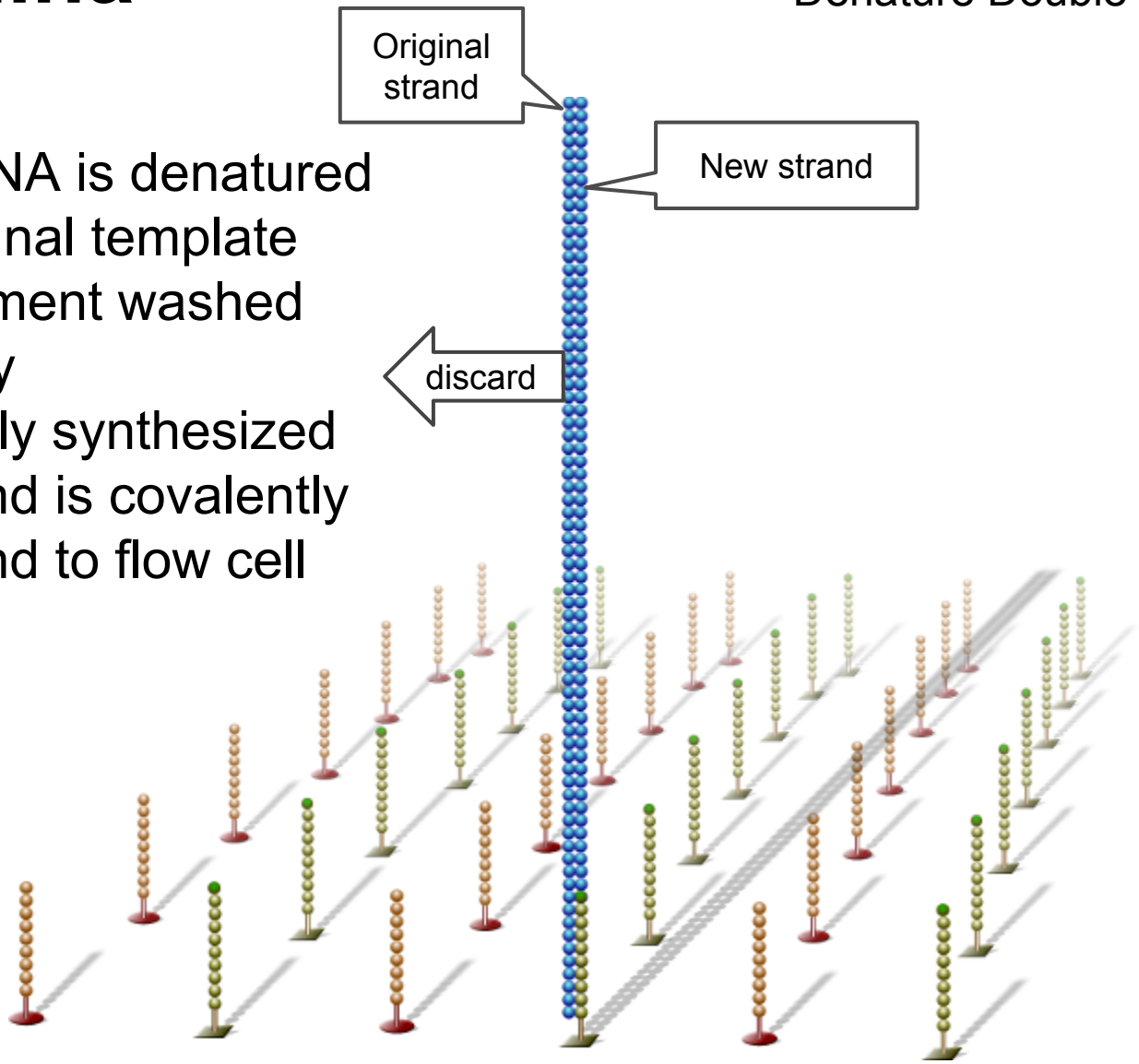
- Millions of single molecules hybridize to the lawn of adapters
- dsDNA extended by polymerases



Illumina

Cluster Generation:
Denature Double-stranded DNA

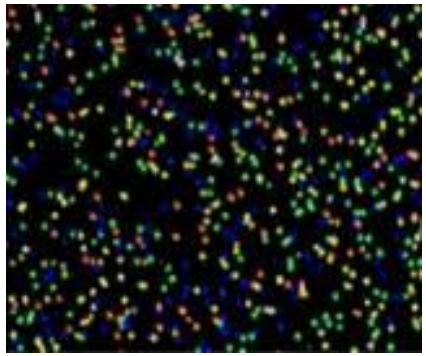
- dsDNA is denatured
- Original template fragment washed away
- Newly synthesized strand is covalently bound to flow cell



Illumina

Cluster Generation:
Covalently-Bound, Randomly Dispersed Single Molecules

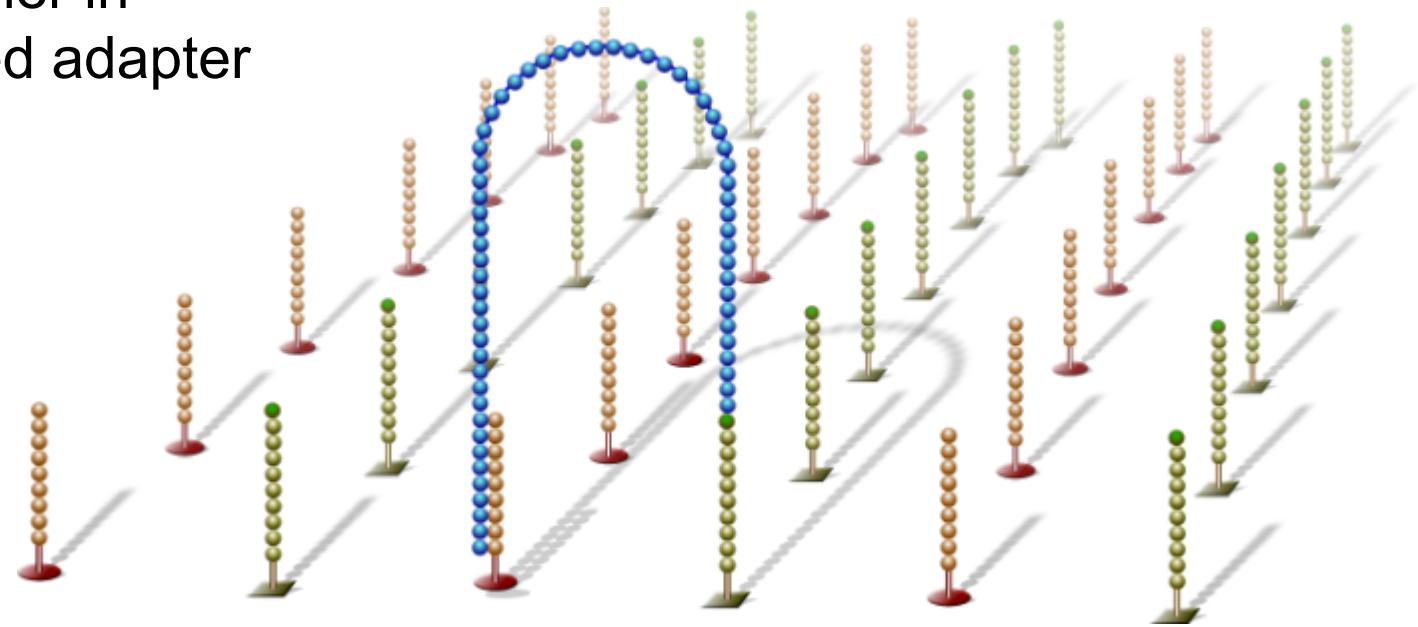
- Resulting covalently-bound DNA fragments are bound to the flow cell surface in a random pattern



Illumina

Cluster Generation:
Bridge Amplification

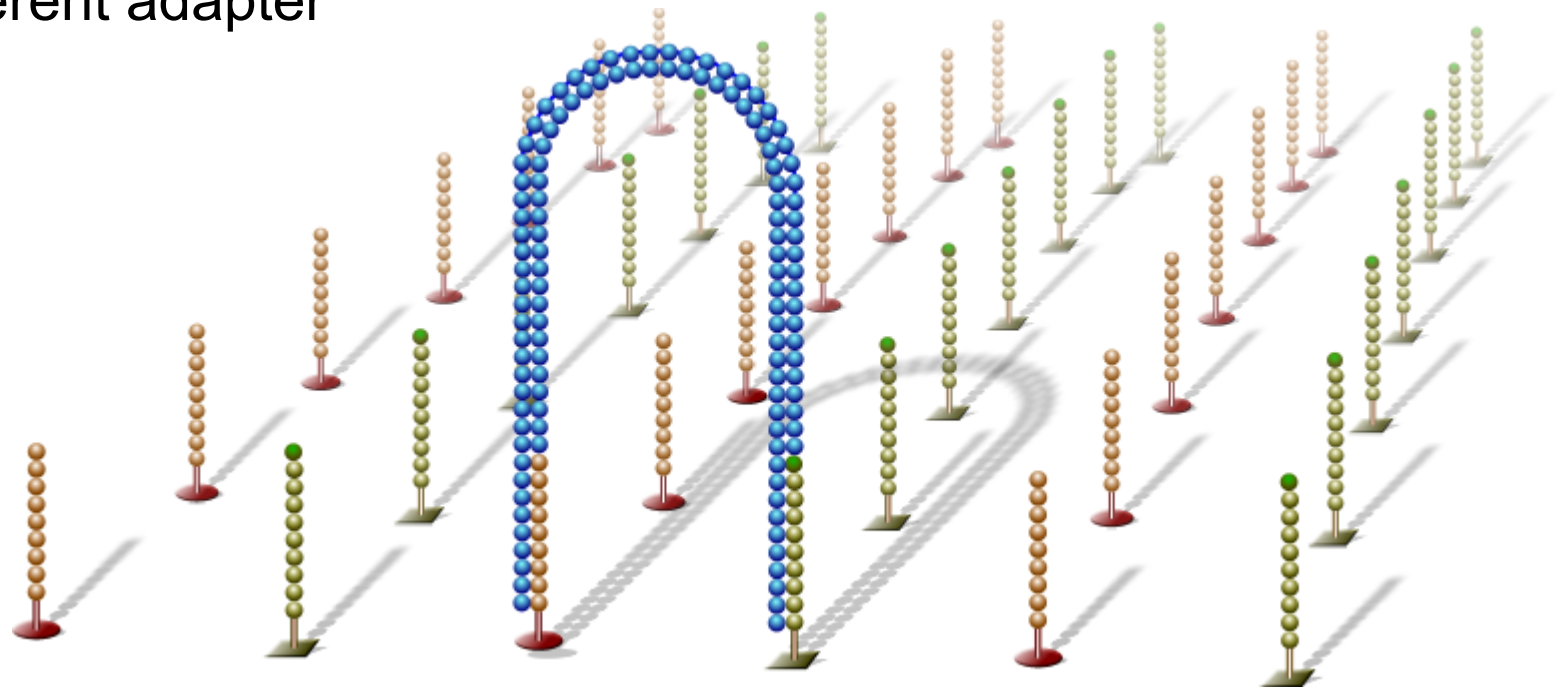
- Single-strand flops over to hybridize to adjacent adapter, forming a bridge
- dsDNA synthesized from primer in hybridized adapter



Illumina

Cluster Generation:
Bridge Amplification

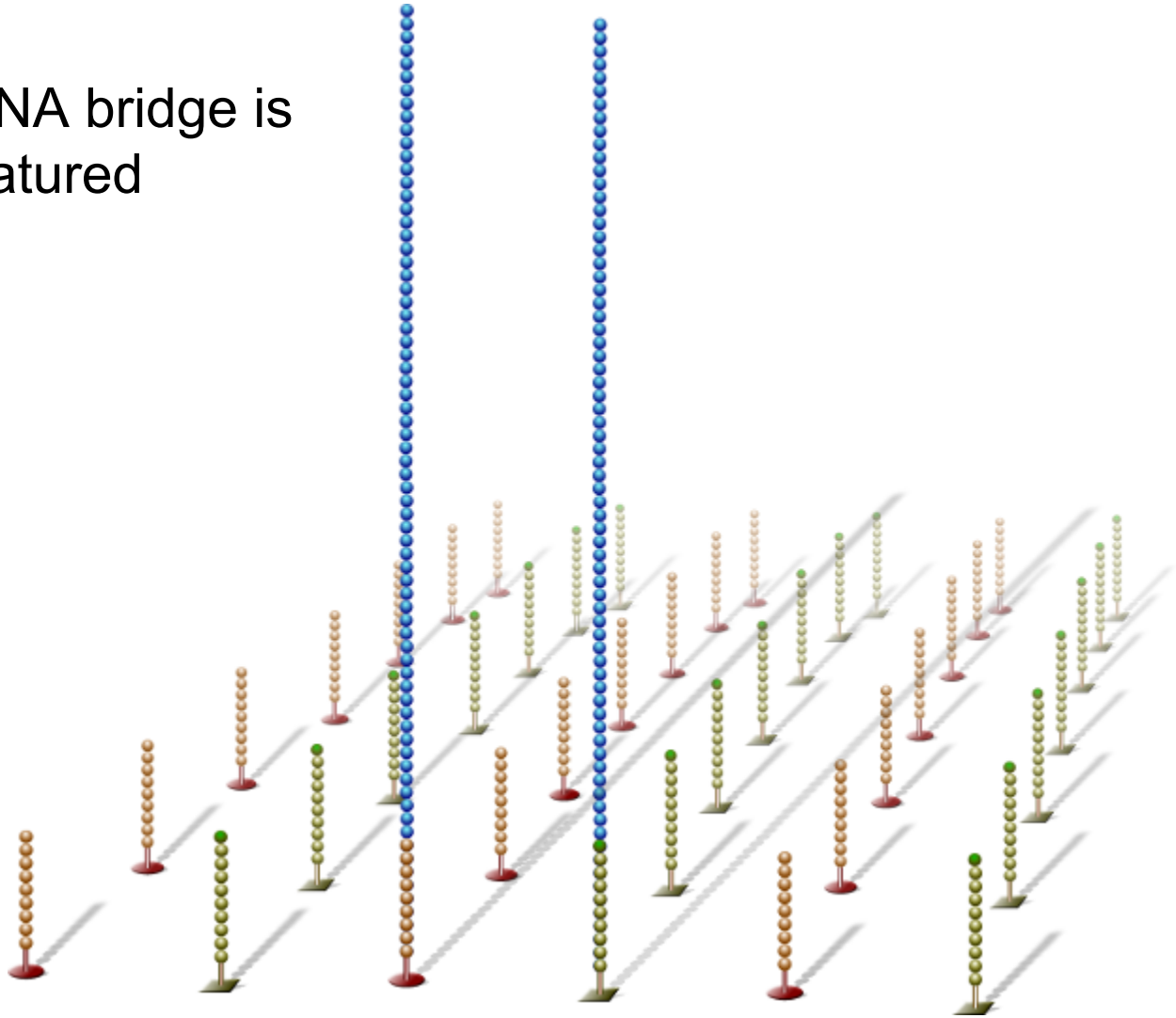
- dsDNA bridge now formed
- each strand covalently bound to different adapter



Illumina

Cluster Generation:
Bridge Amplification

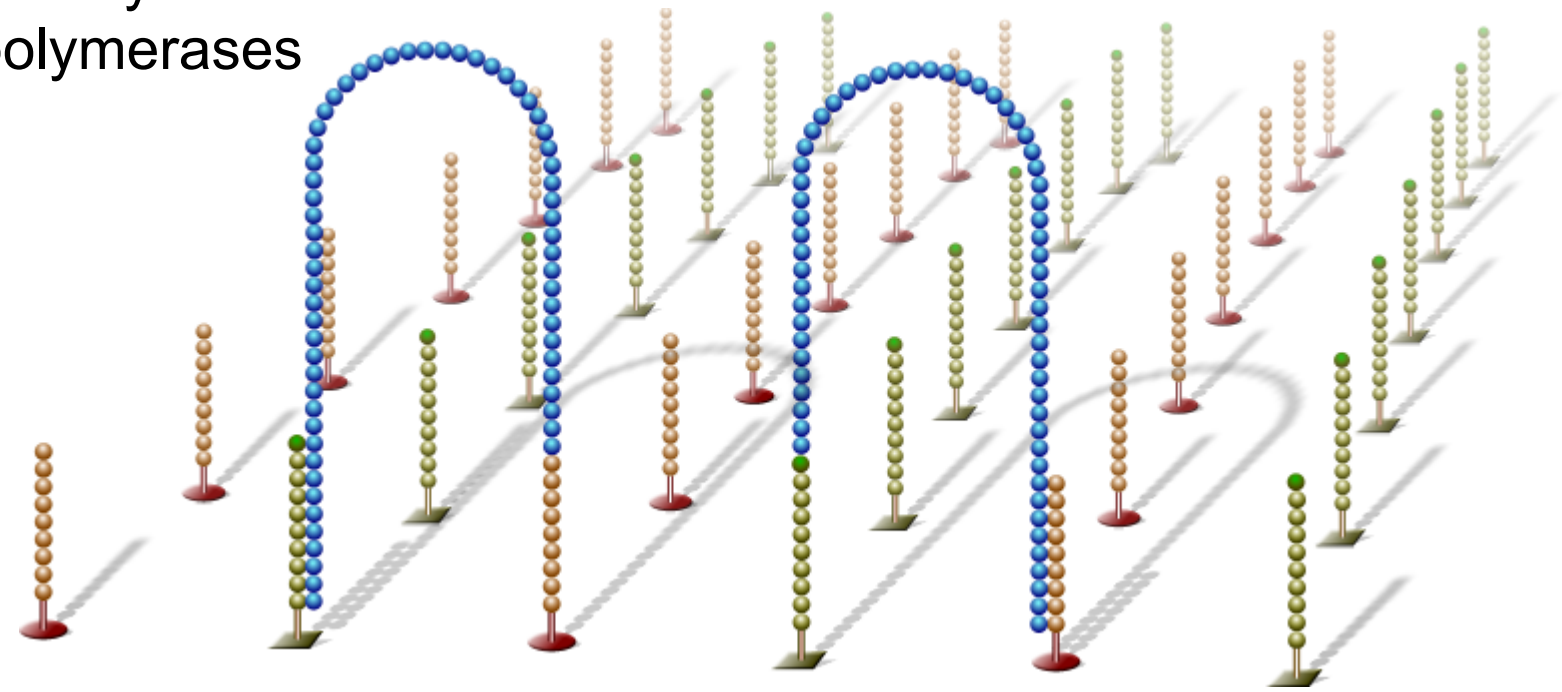
- dsDNA bridge is denatured



Illumina

Cluster Generation:
Bridge Amplification

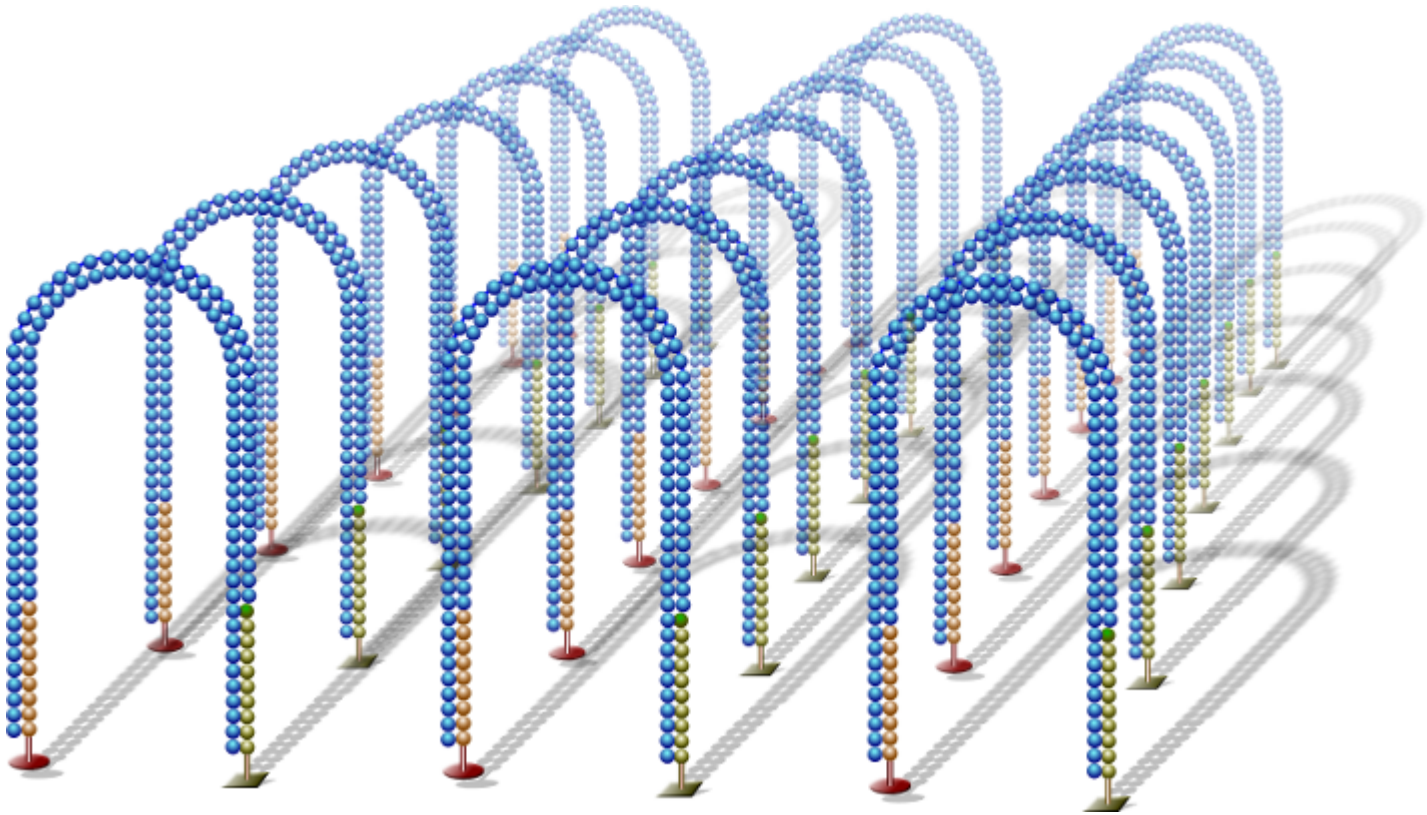
- Single strands flop over to hybridize to adjacent adapters, forming bridges
- dsDNA synthesized by polymerases



Illumina

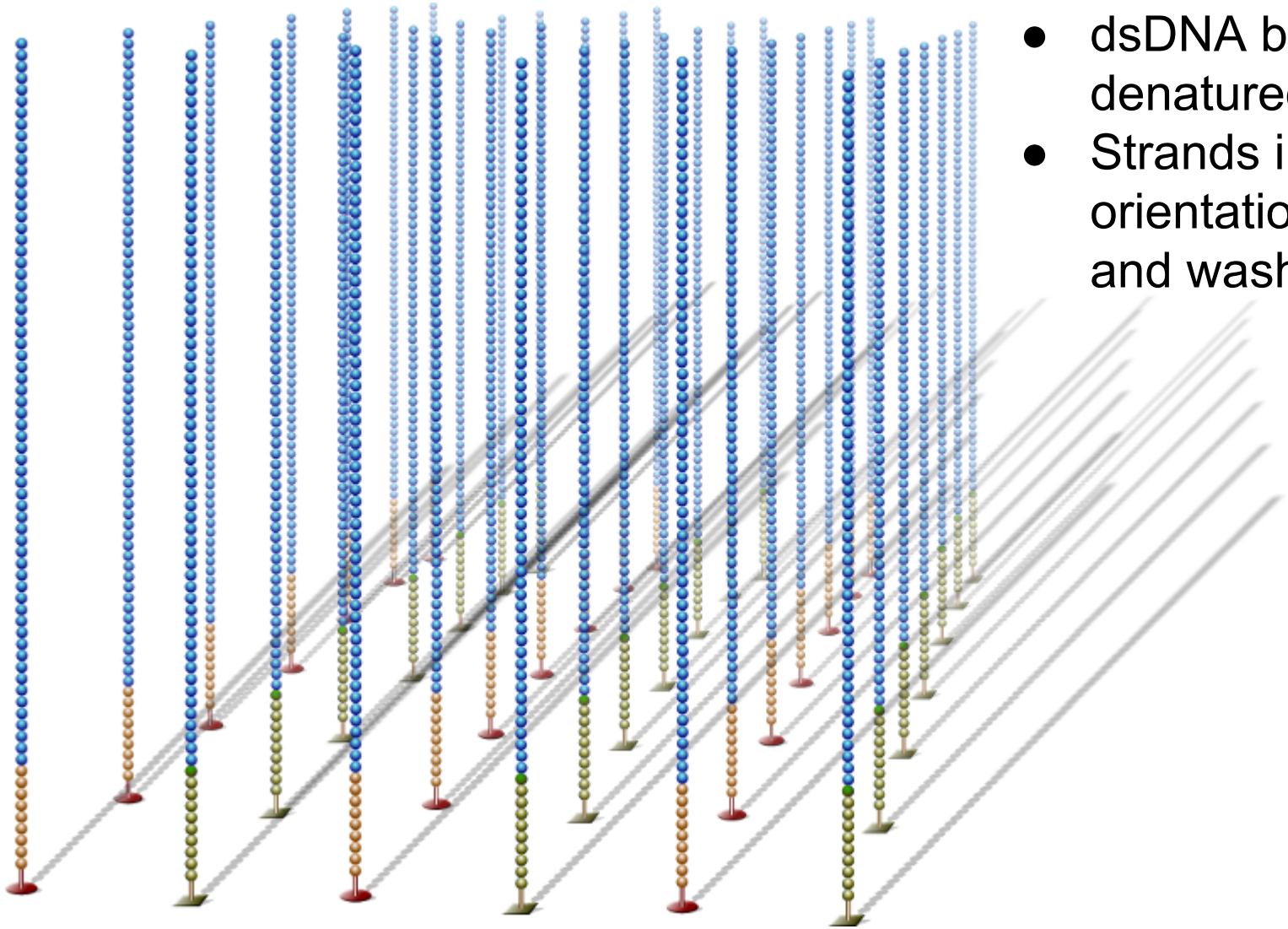
Cluster Generation:
Bridge Amplification

- Bridge amplification cycles repeated many times



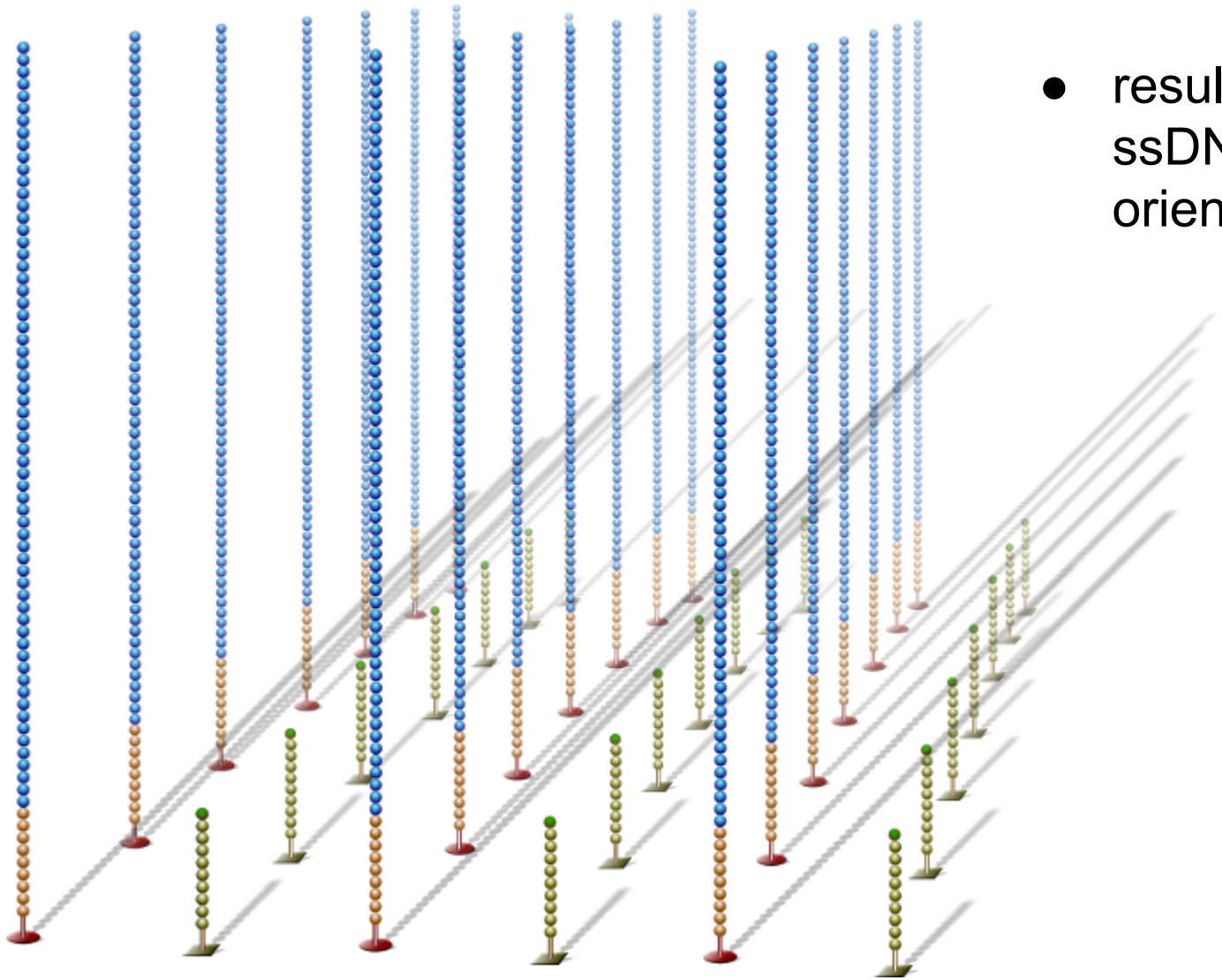
Illumina

Cluster Generation



- dsDNA bridges denatured
- Strands in one of the orientations cleaved and washed away

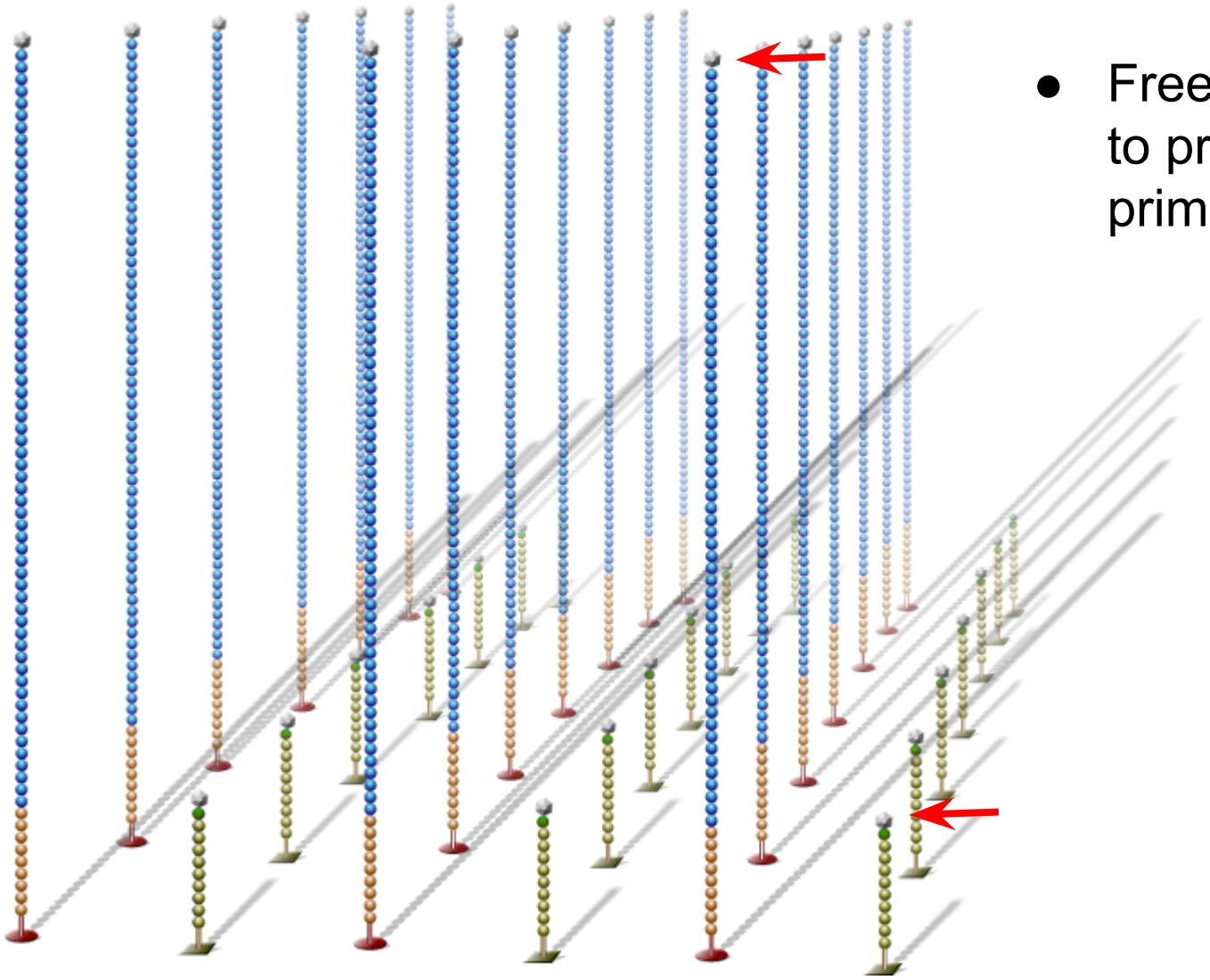
Illumina



- resulting cluster has ssDNA in only one orientation

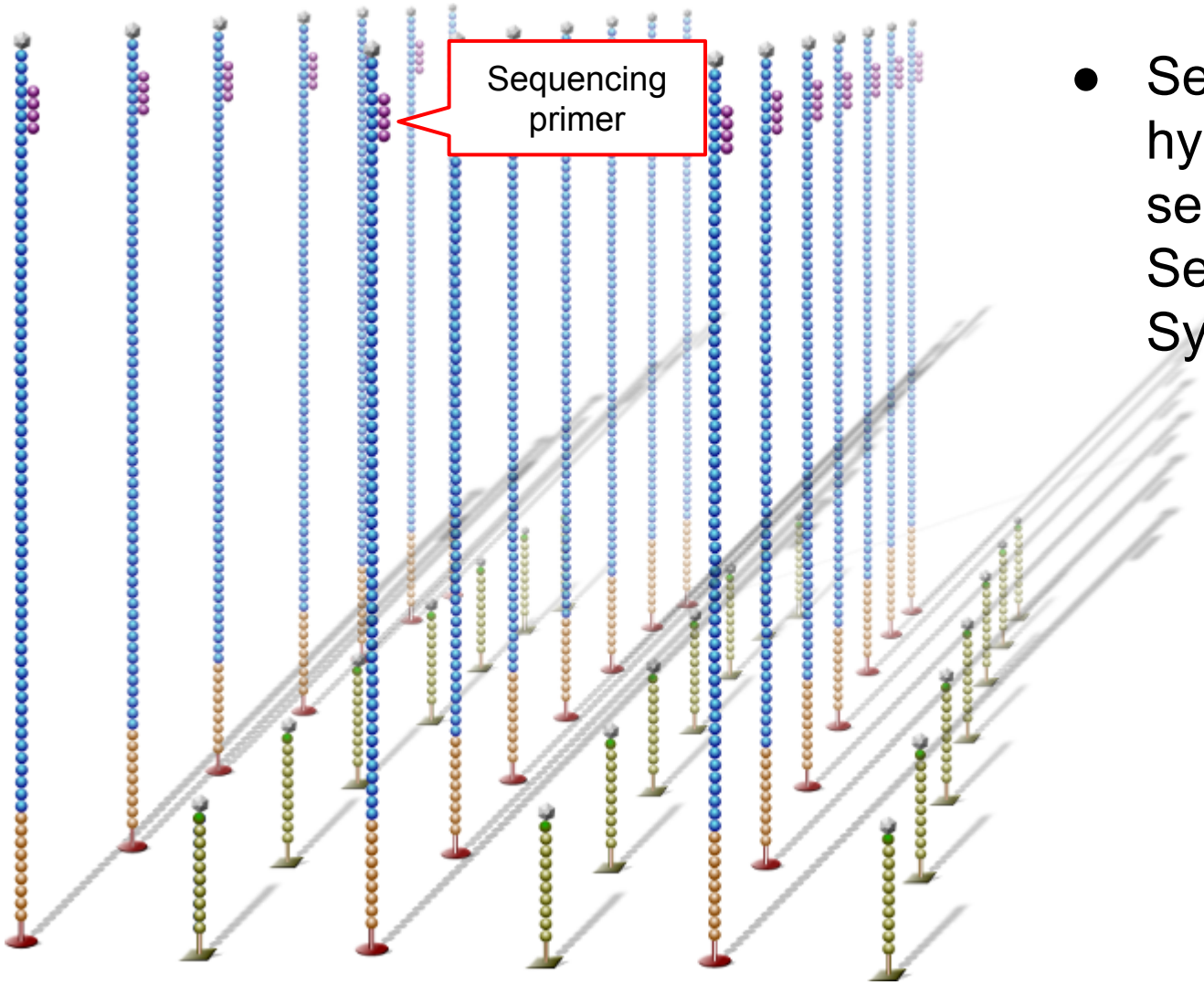
Illumina

Cluster Generation



- Free 3'-ends blocked to prevent unwanted priming

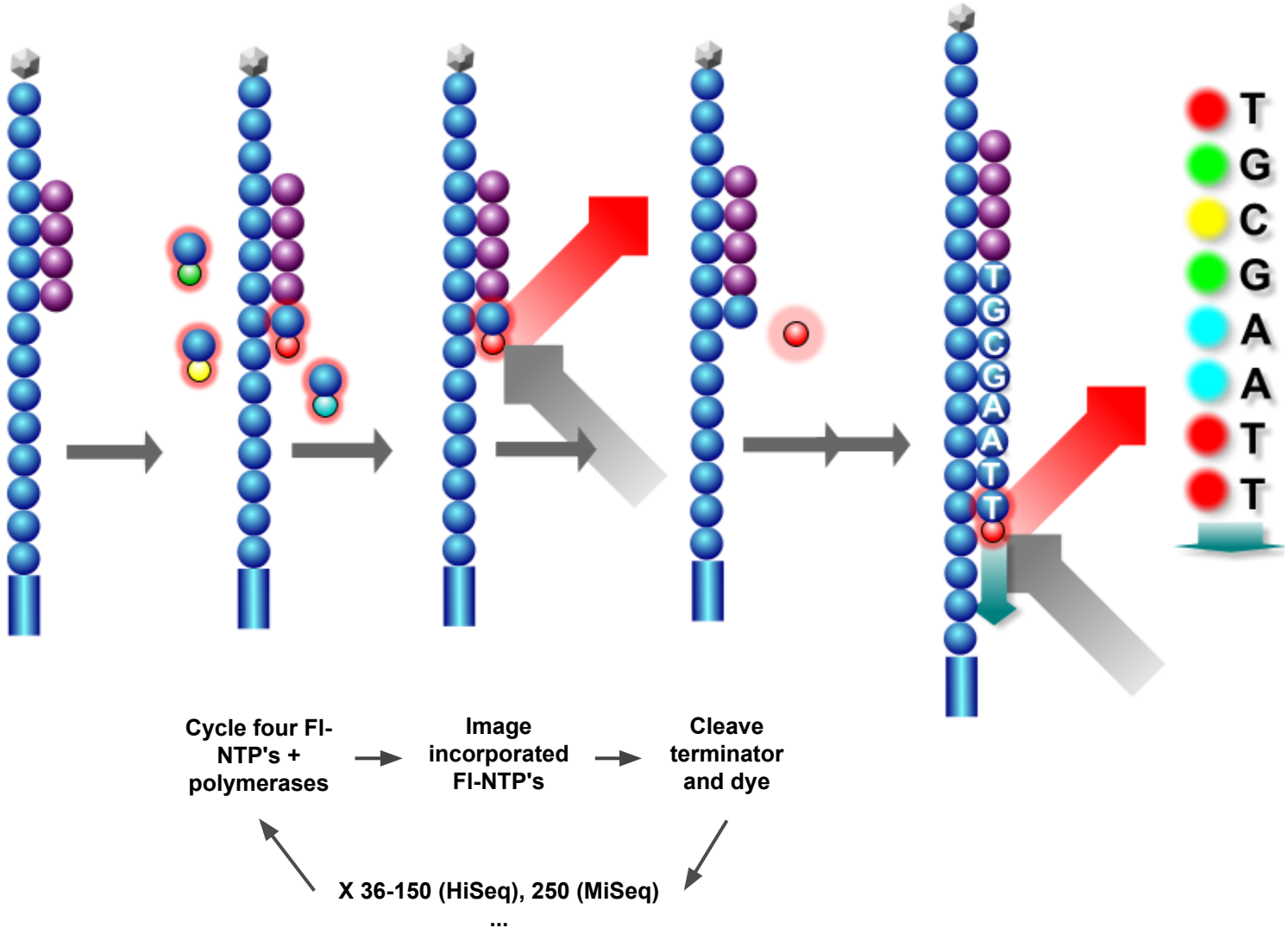
Illumina



- Sequencing primer is hybridized to adapter sequence, starting Sequencing By Synthesis

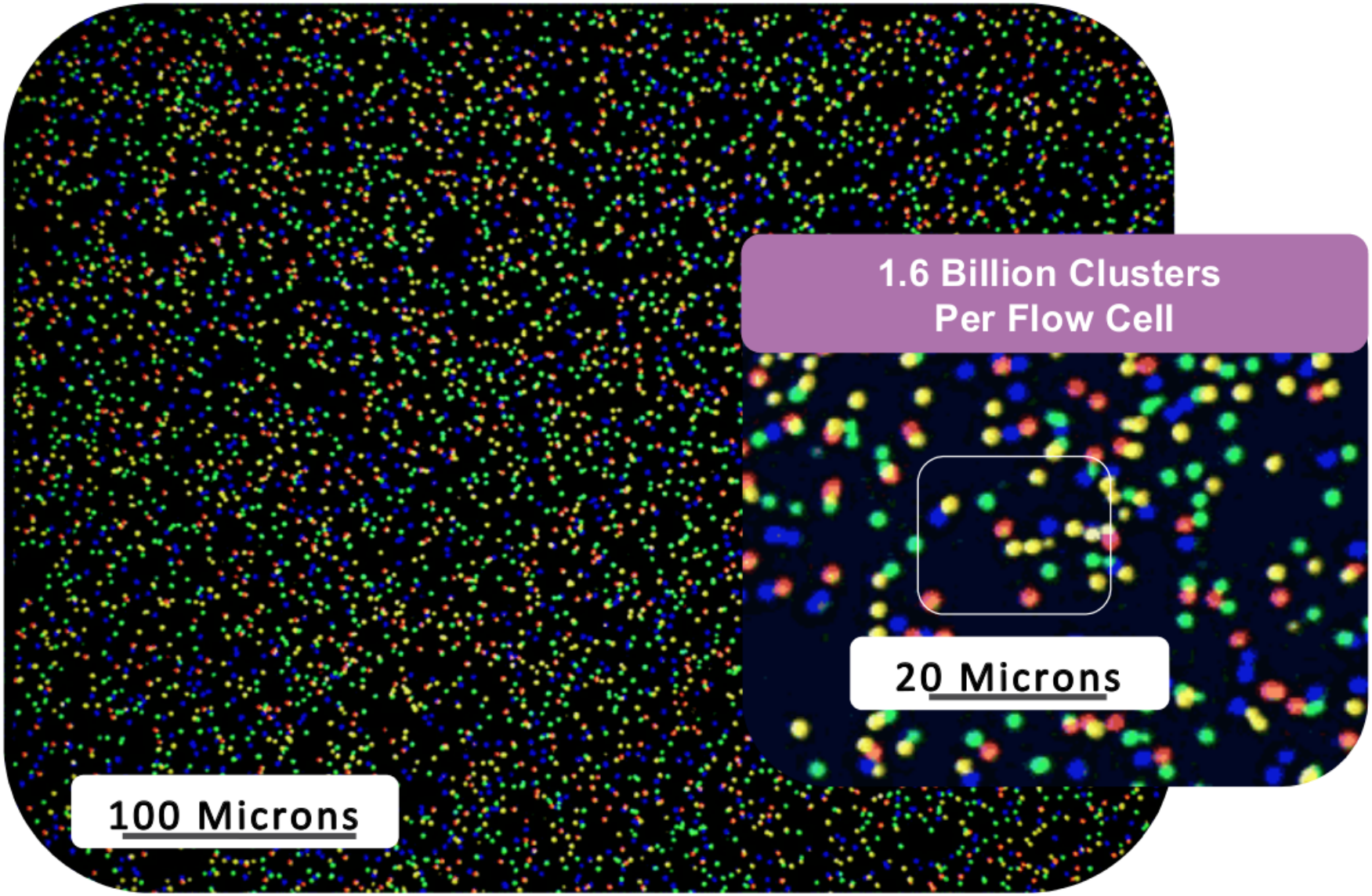
Illumina

Sequencing By Synthesis

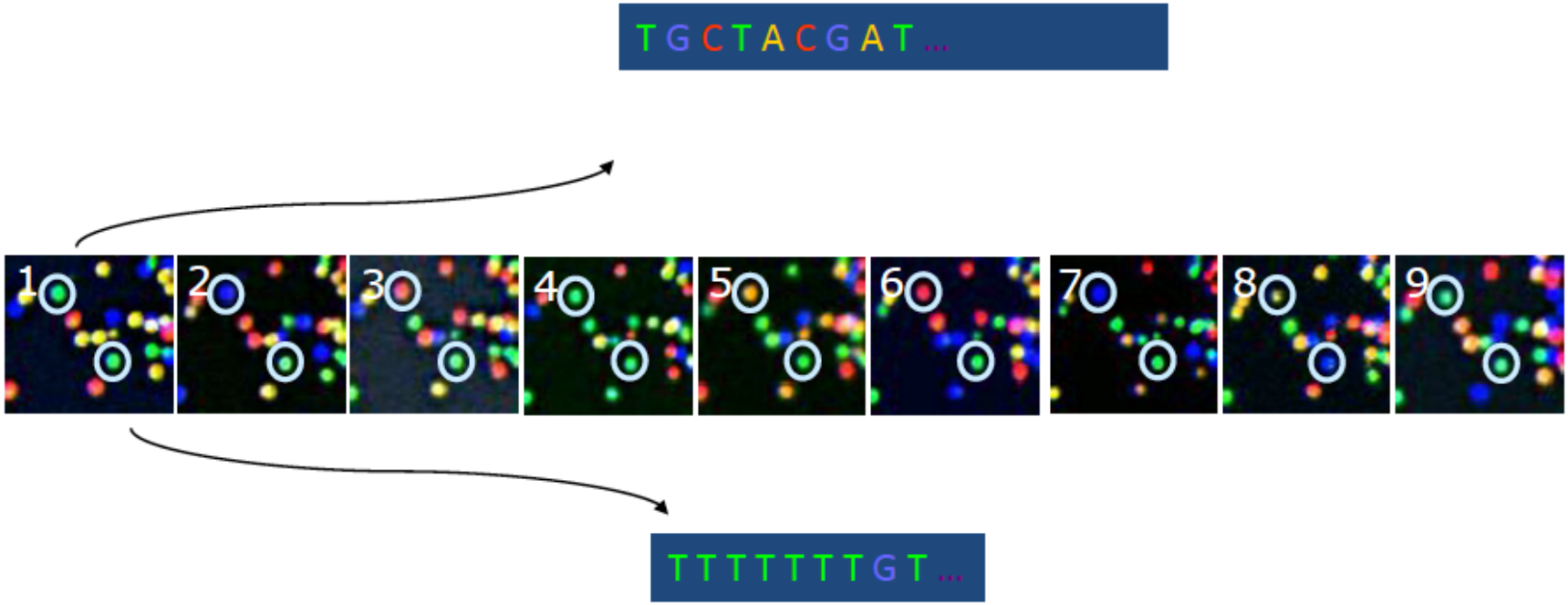


Illumina

Sequencing By Synthesis



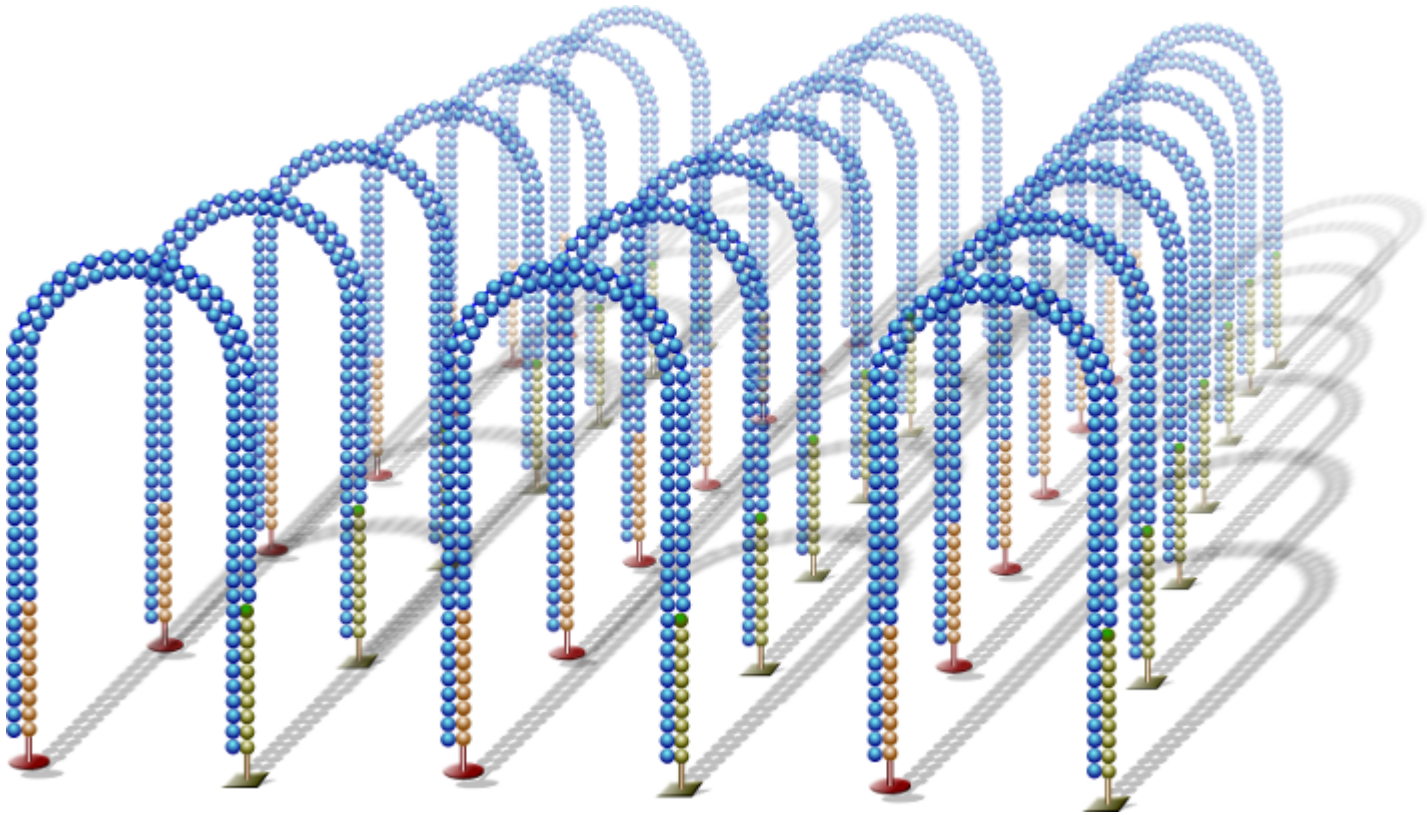
Illumina



The identity of each base of a cluster is read off from sequential images.

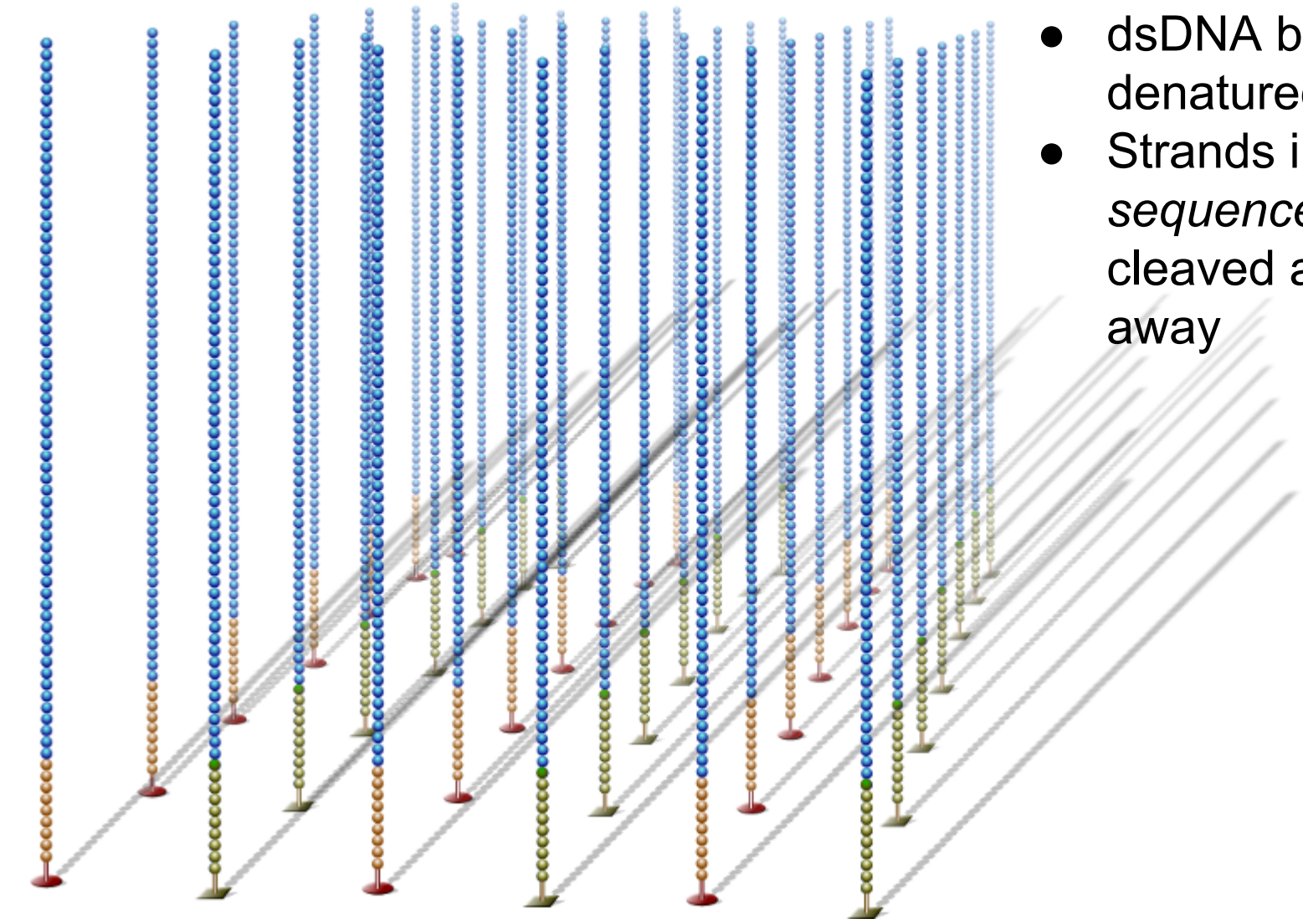
Illumina

- Bridge amplification to generate strands with opposite orientation



Illumina

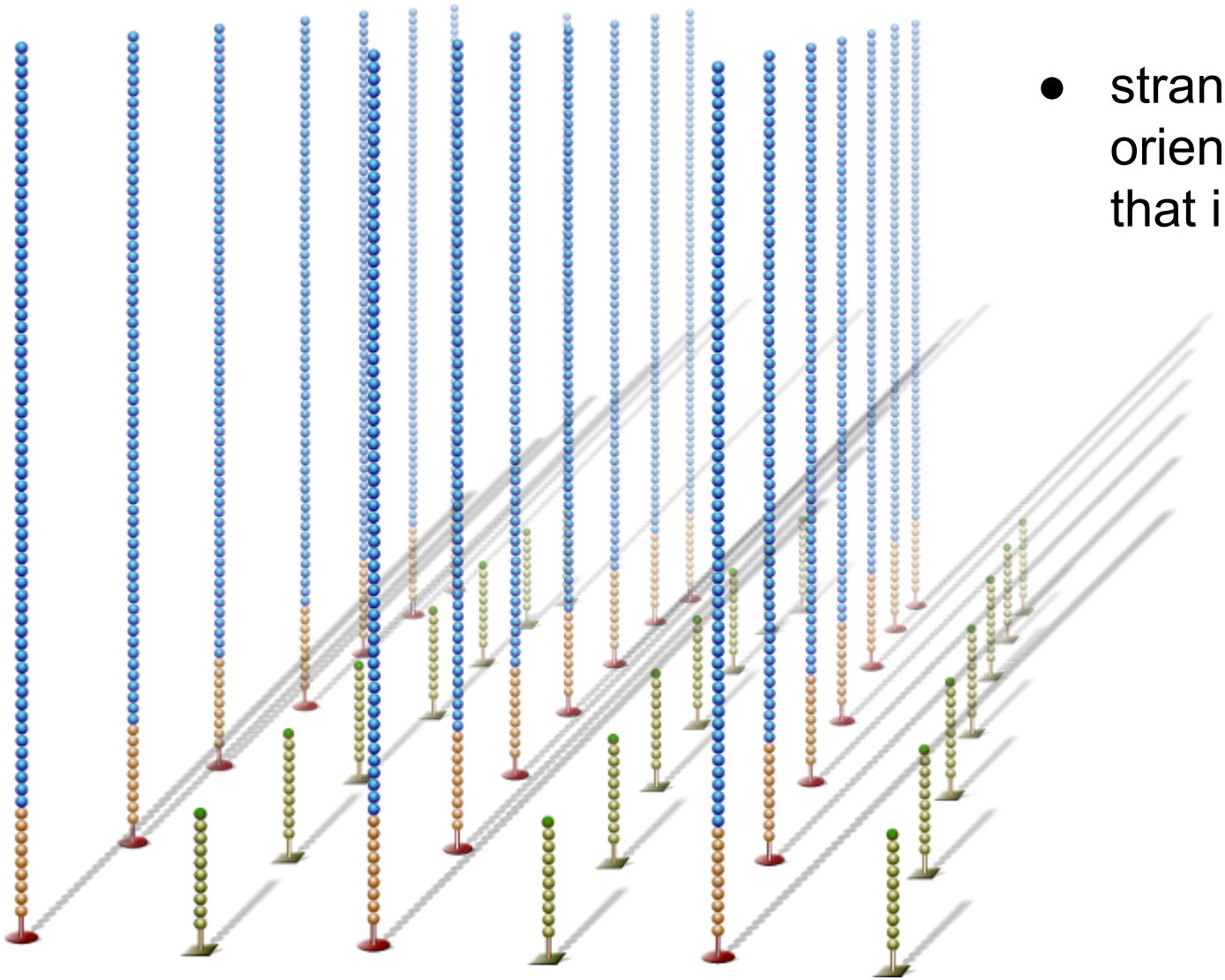
Paired-end sequencing



- dsDNA bridges denatured
- Strands in *already sequenced* orientation cleaved and washed away

Illumina

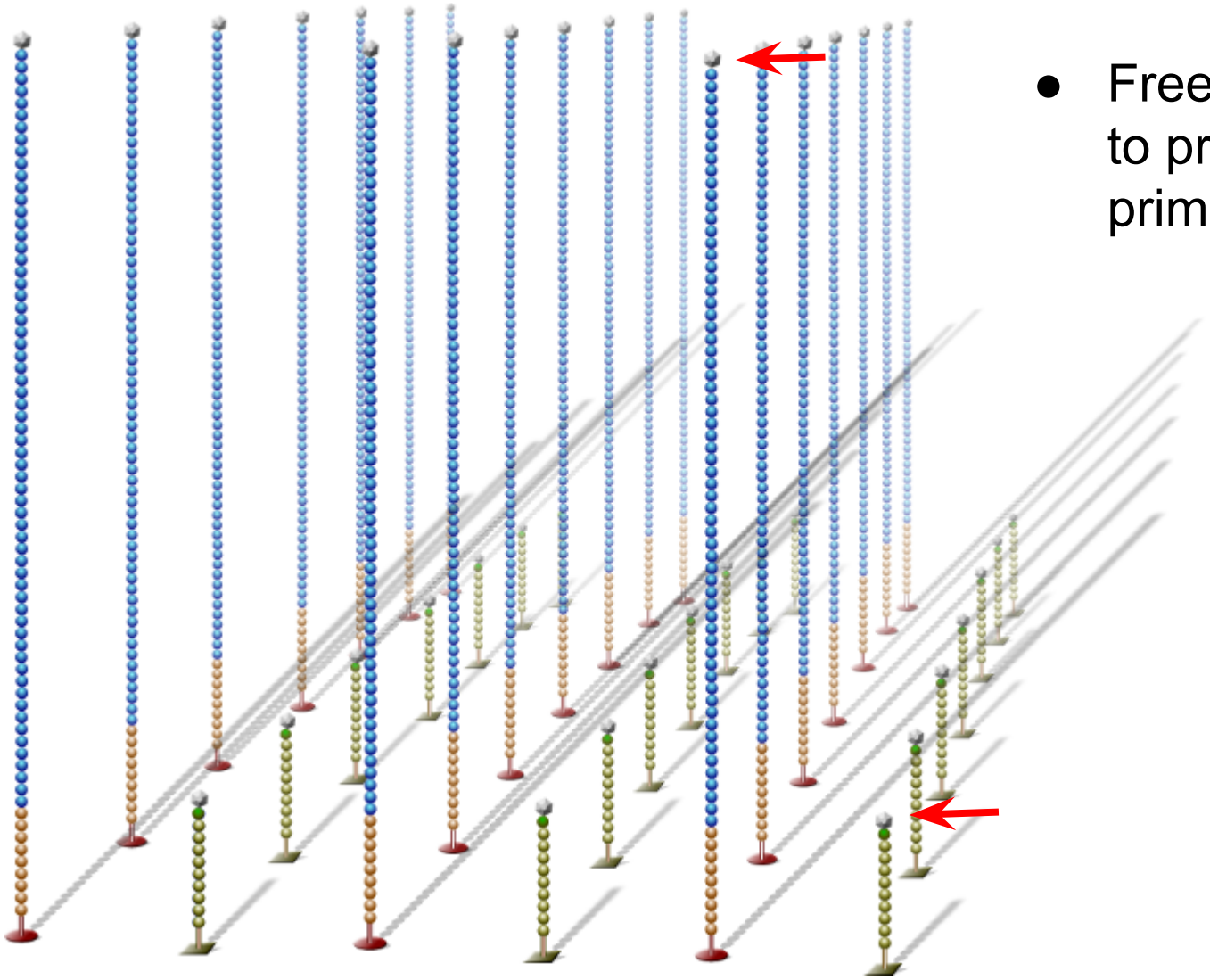
Paired-end sequencing



- strands with uniform orientation, *opposite* that in first read

Illumina

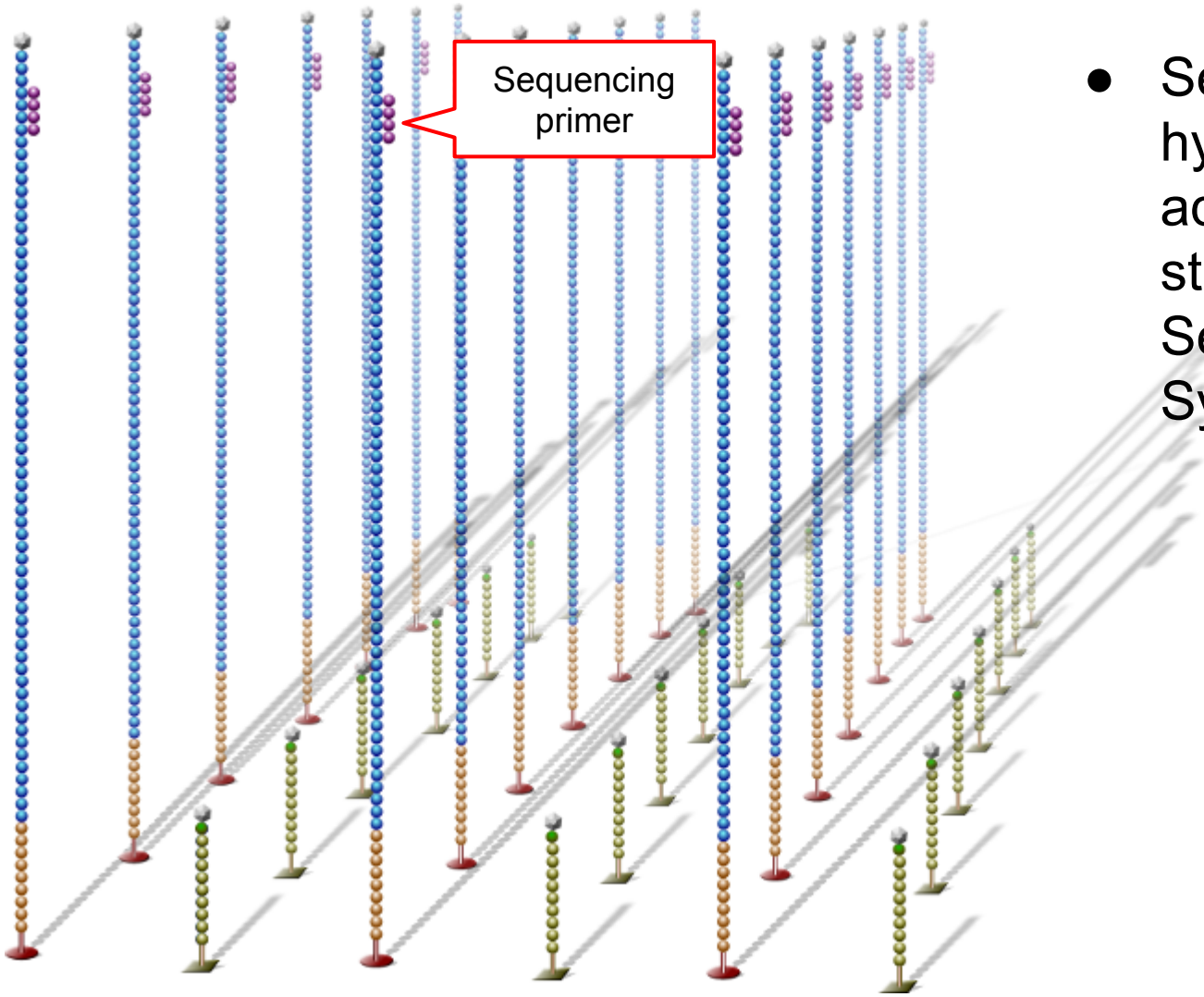
Paired-end sequencing



- Free 3'-ends blocked to prevent unwanted priming

Illumina

Paired-end sequencing



- Sequencing primer is hybridized to other adapter sequence, starting second read's Sequencing By Synthesis

Illumina

HiSeq 2000 stats:

- Dual surface imaging
- Fast scanning and imaging
- Two flow cells (in sequence)
- Initially: 200 Gbp per run
- Currently: up to 1Tbp per run
- Run time
  - 7-8 days (100bp PE)
  - 1-2 day “rapid mode”
- 25 Gbp / day
- 2 billion paired-end reads (130-180 million clusters per lane)
- < $5k per human genome
- < $100 per transcriptome

Illumina

http://systems.illumina.com/systems/sequencing.ilmn

Sequencing systems for every lab, application, and scale of study.

From the power of the HiSeq X to the speed of MiSeq, Illumina has the sequencer that's just right for you.

MiSeq
Focused power. Speed and simplicity for targeted and small genome sequencing.

NextSeq 500
Flexible power. Speed and simplicity for everyday genomics.

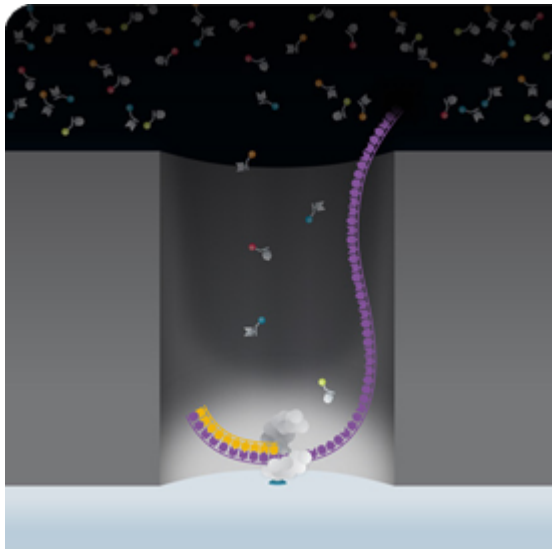
HiSeq 2500
Production power. Power and efficiency for large-scale genomics.

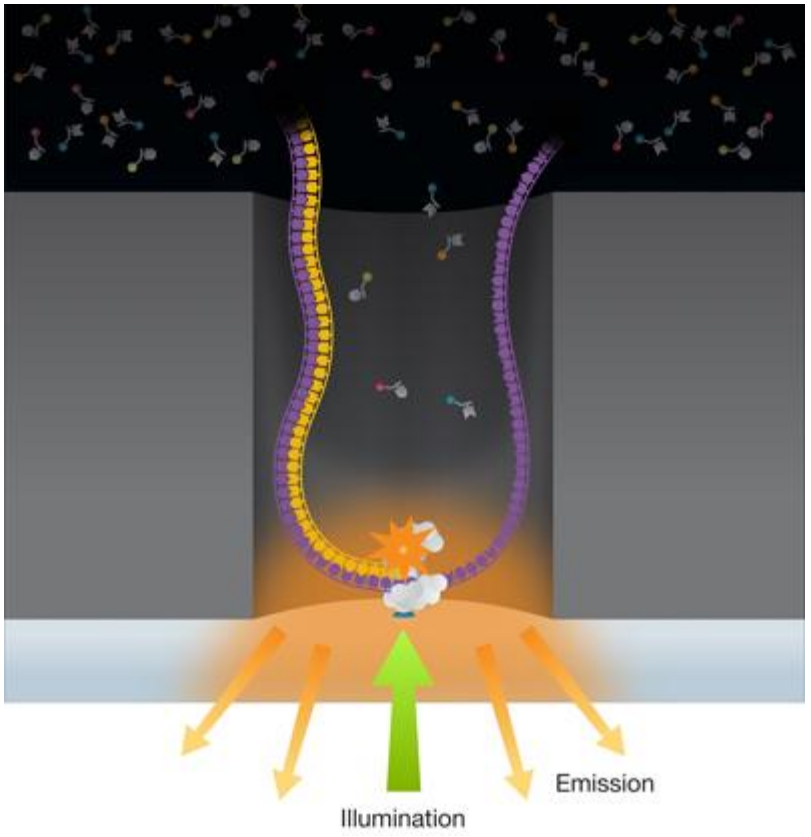
HiSeq X*
Population power. $1,000 human genome and extreme throughput for population-scale sequencing.

Key applications	Small genome, amplicon, and targeted gene panel sequencing.	Everyday genome, exome, transcriptome sequencing, and more.		Production-scale genome, exome, transcriptome sequencing, and more.		Population-scale human whole-genome sequencing.
…						
Output range	0.3-15 Gb	20-39 Gb	30-120 Gb	10-180 Gb	50-1000 Gb	1.6-1.8 Tb
Run time	5-55 hours	15-26 hours	12-30 hours	7-40 hours	< 1 day - 6 days	< 3 days
Reads per flow cell†	25 Million‡	130 Million	400 Million	300 Million	2 Billion	3 Billion
Maximum read length	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 125 bp	2 × 150 bp

PacBio

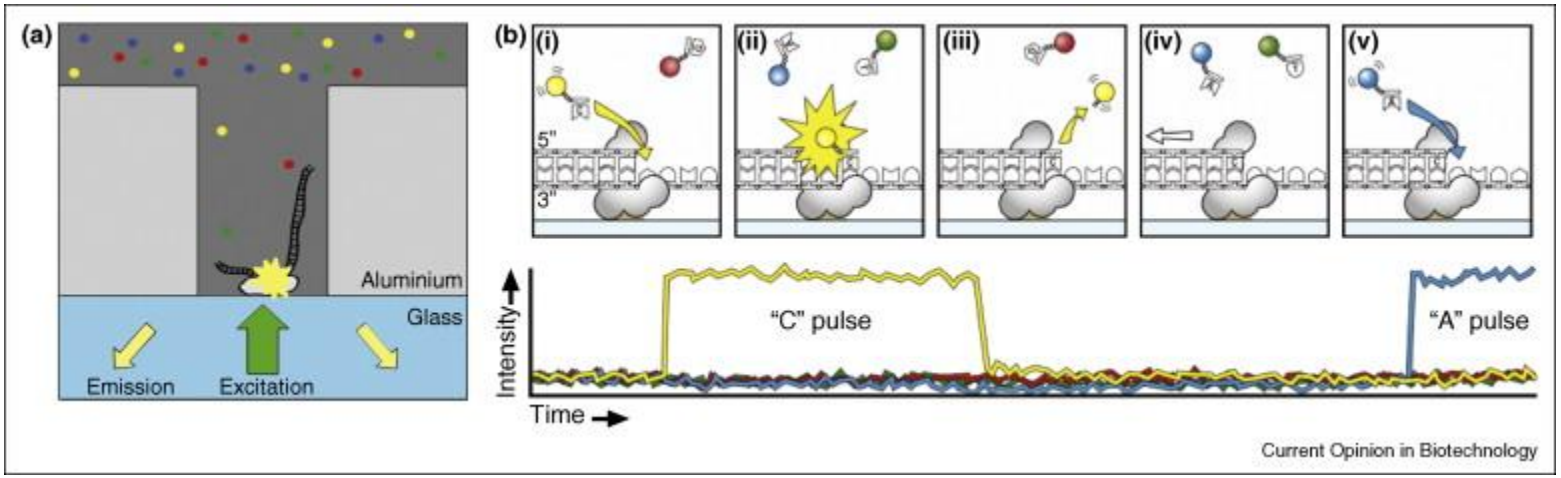
PacBio RS (Real-time Sequencer)





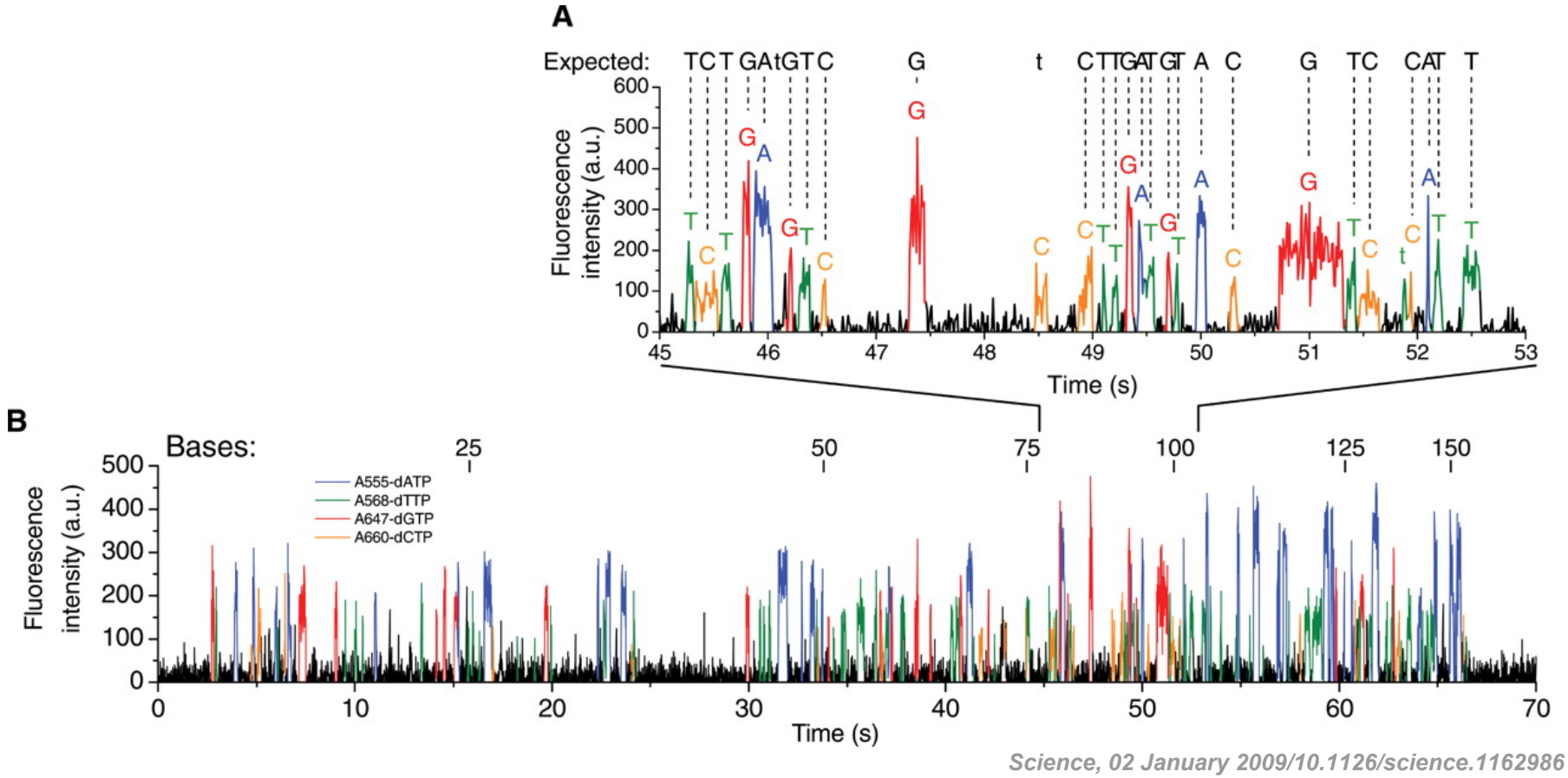
Polymerase / DNA complex adhered to bottom of imaging well (**Z**ero **M**ode **W**aveguide) ... evanescent wave illuminates tiny volume around polymerase.

PacBio RS (Real-time Sequencer)



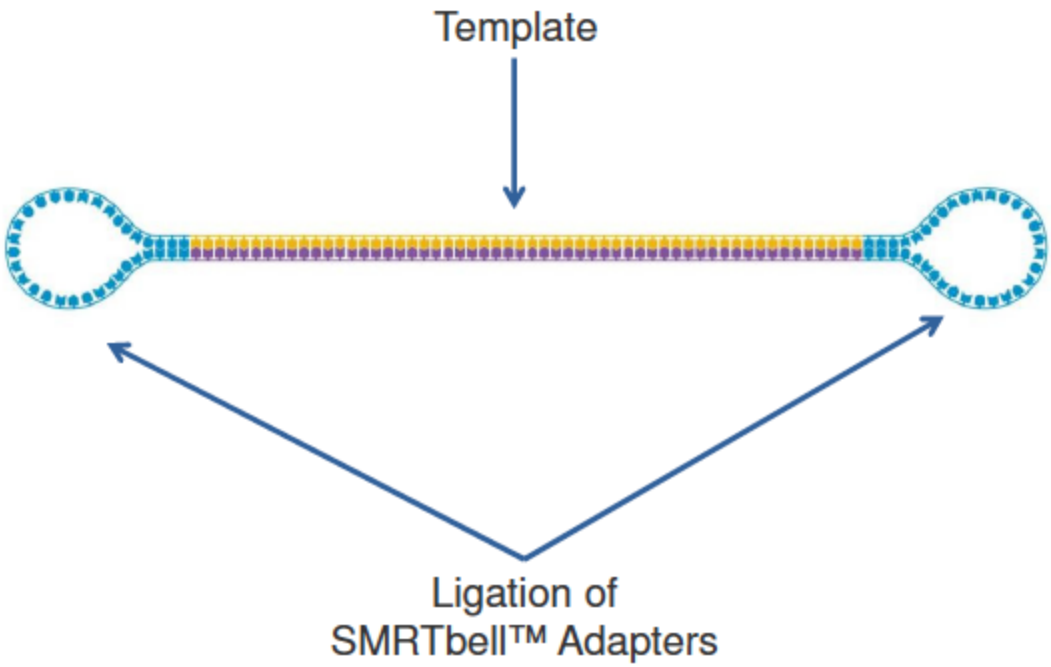
Fluorescently-tagged nucleotides are only seen (for an *appreciable* amount of time) when associated with polymerase. Persistent time in the excitation volume can be recognized as a "pulse."

PacBio RS (Real-time Sequencer)



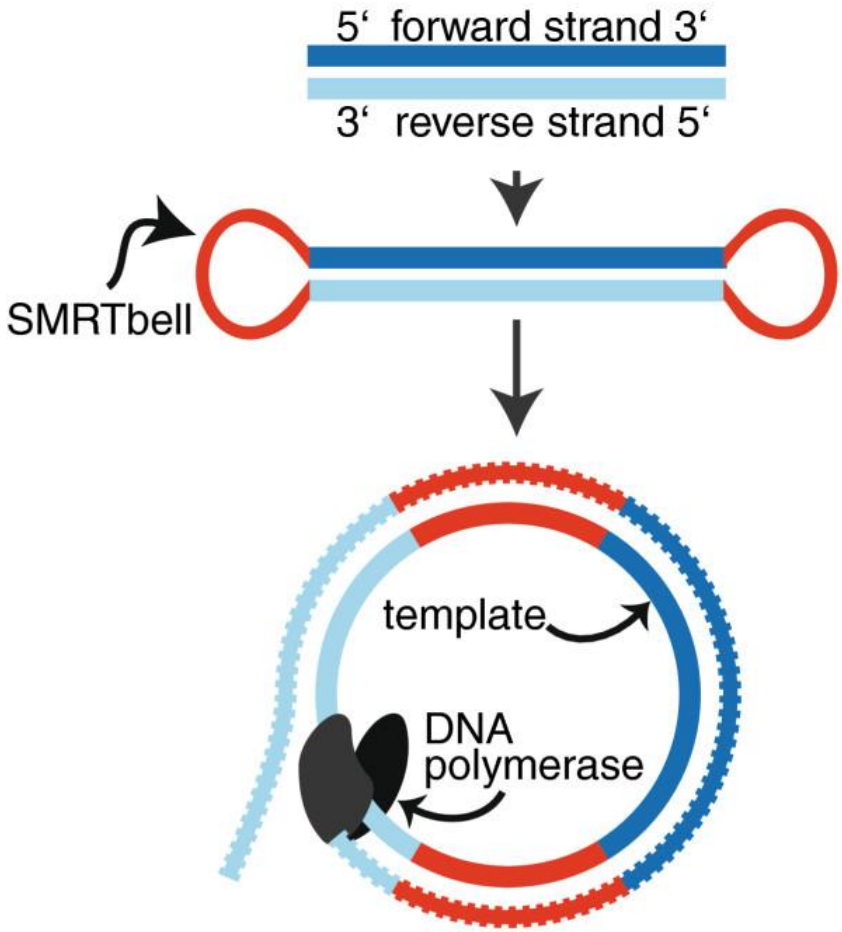
Science, 02 January 2009/10.1126/science.1162986

PacBio SMRTbell Construct



PacBio Sequencing

PacBio Sequencing



PacBio Sequencing

Standard Sequencing for Continuous Long Reads (CLR)

Large Insert Sizes (>2kb)

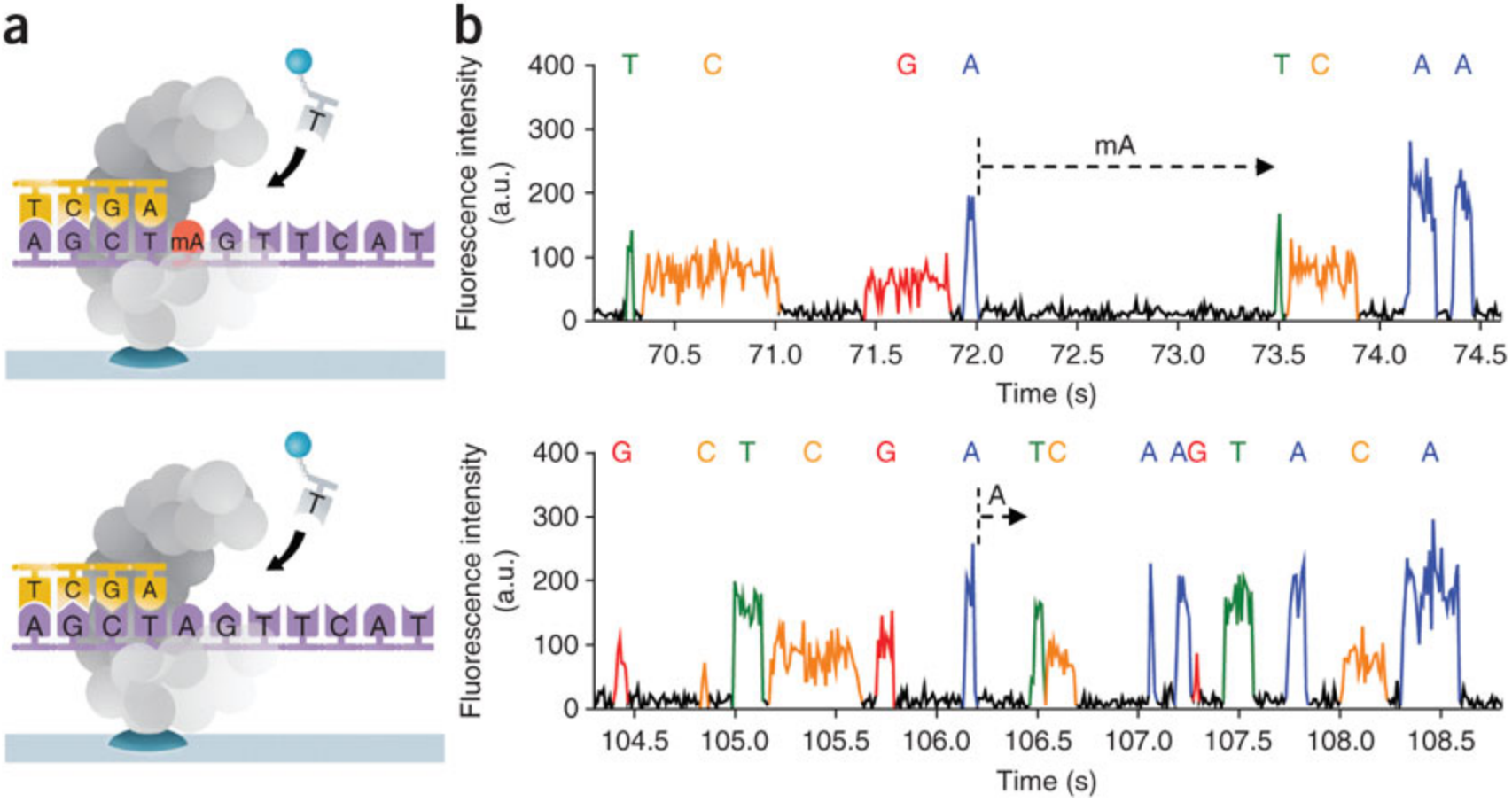
Generates one pass on each molecule sequenced

Circular Consensus Sequencing (CCS)

Small Insert Sizes (250 bp – 2 kb)

Generates multiple passes on each molecule sequenced

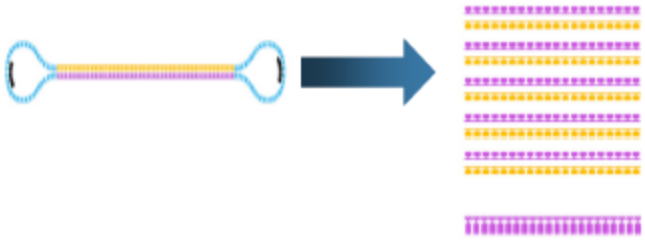
PacBio detection of modified bases



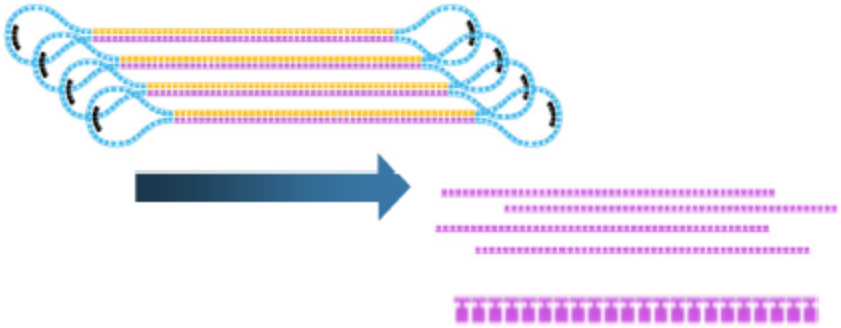
Movie → trace → pulse timing can reveal nucleotide modification, e.g. N6-methyladenosine

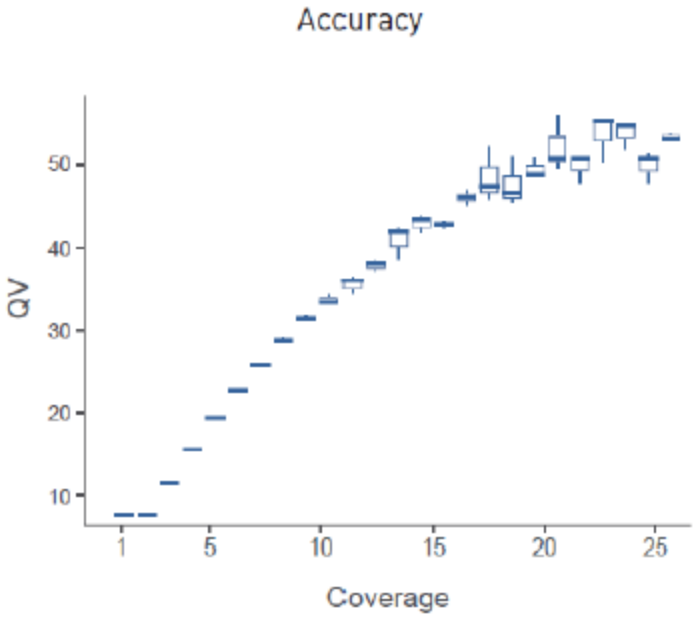
PacBio accuracy

Single-Molecule CCS:



Multi-Molecule Consensus:





Accuracy boost with more coverage

Oxford Nanopore

Oxford Nanopore
MinION

Oxford Nanopore
PromethION

Oxford Nanopore
GridION

Oxford Nanopore

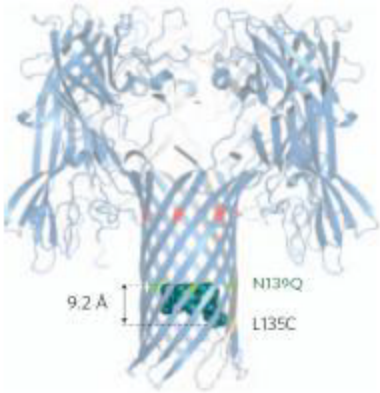
nature nanotechnology

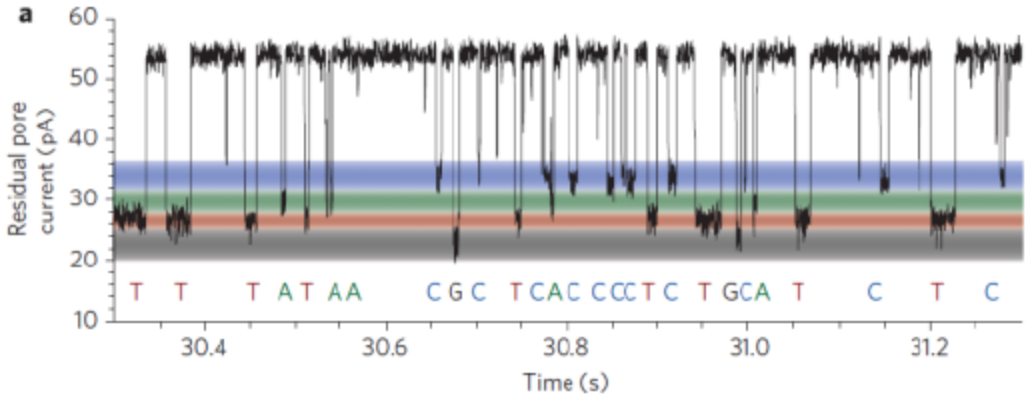
ARTICLES

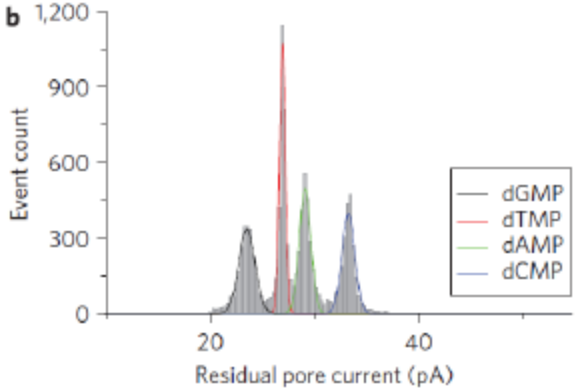
PUBLISHED ONLINE: XX XX 2009 | DOI: 10.1038/NNANO.2009.12

Continuous base identification for single-molecule nanopore DNA sequencing

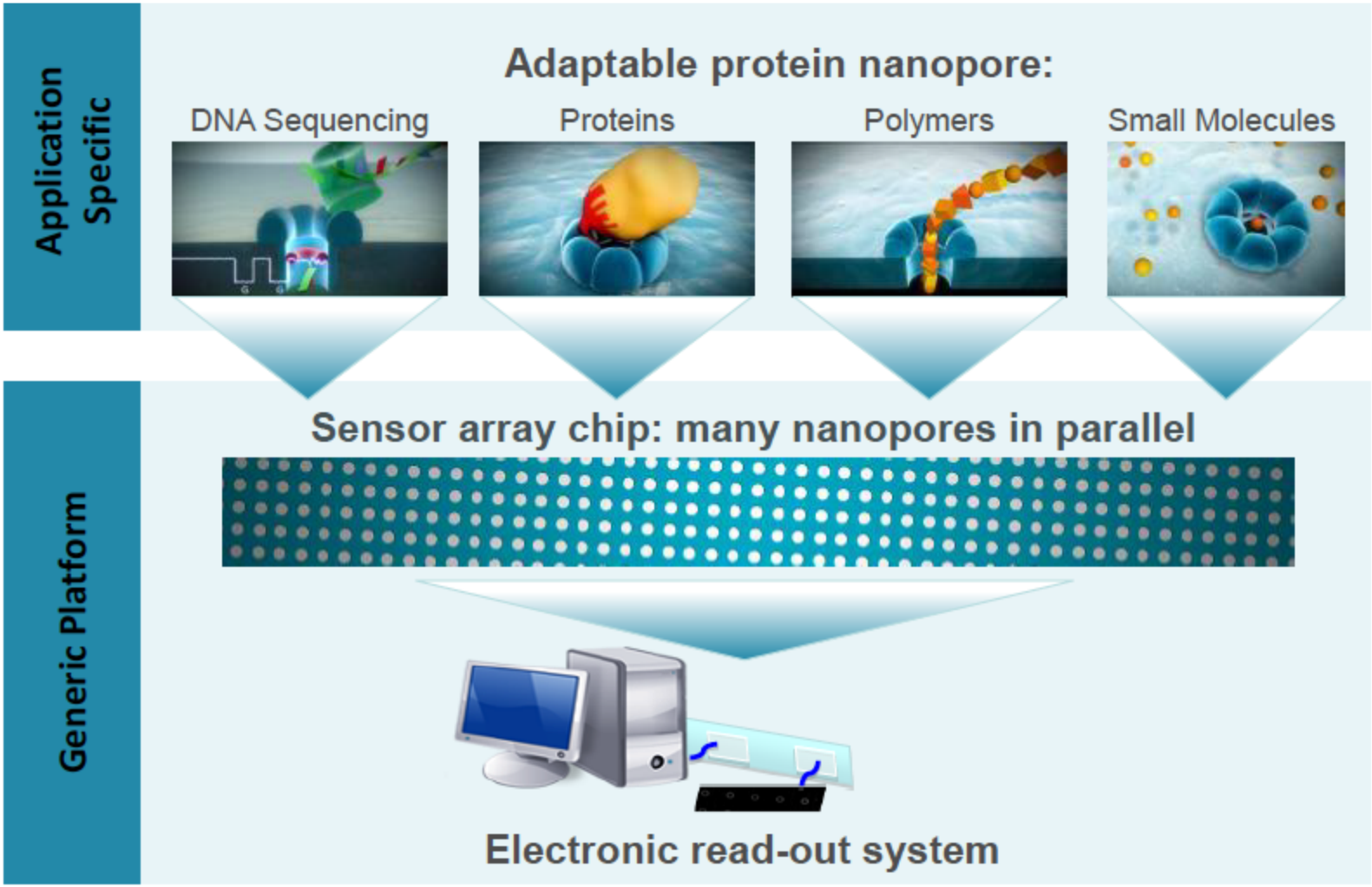
James Clarke[1], Hai-Chen Wu[2], Lakmal Jayasinghe[1,2], Alpesh Patel[1], Stuart Reid[1] and Hagan Bayley[2]*



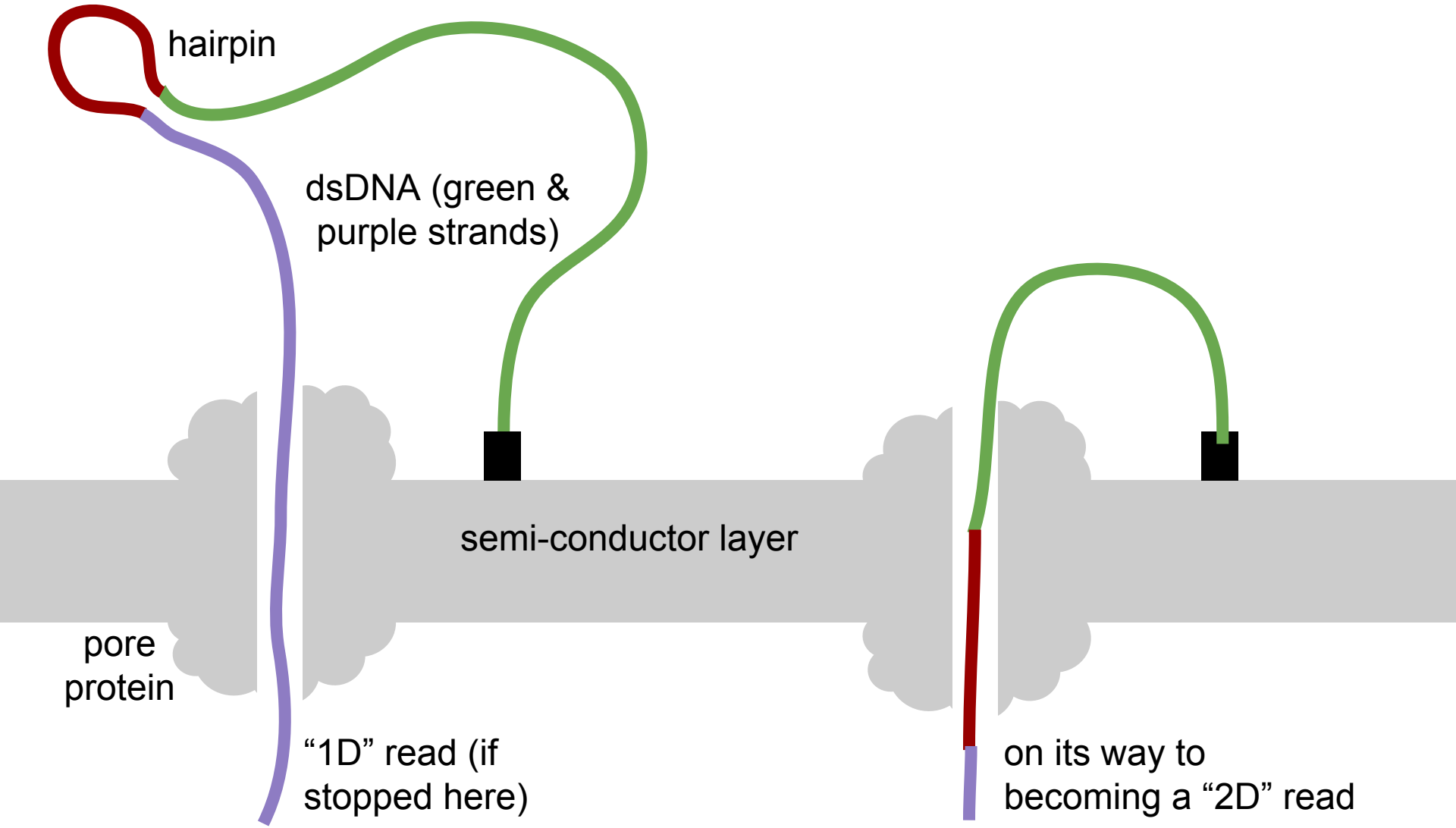




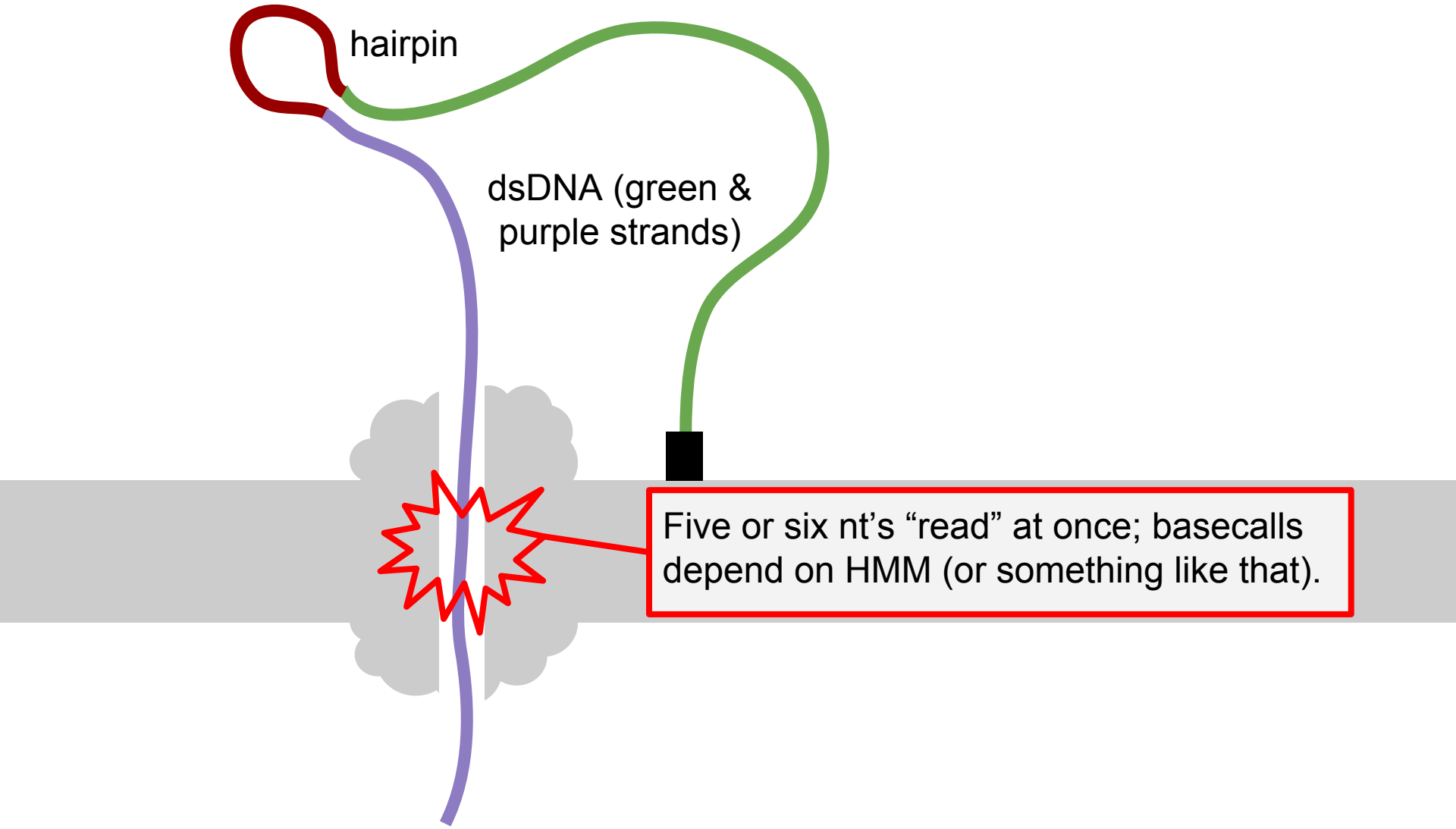
Oxford Nanopore



Oxford Nanopore



Oxford Nanopore



Oxford Nanopore

Whole genome shotgun sequence data of *E coli* MG1655 strain *published* on GigaDB by Nick Loman.

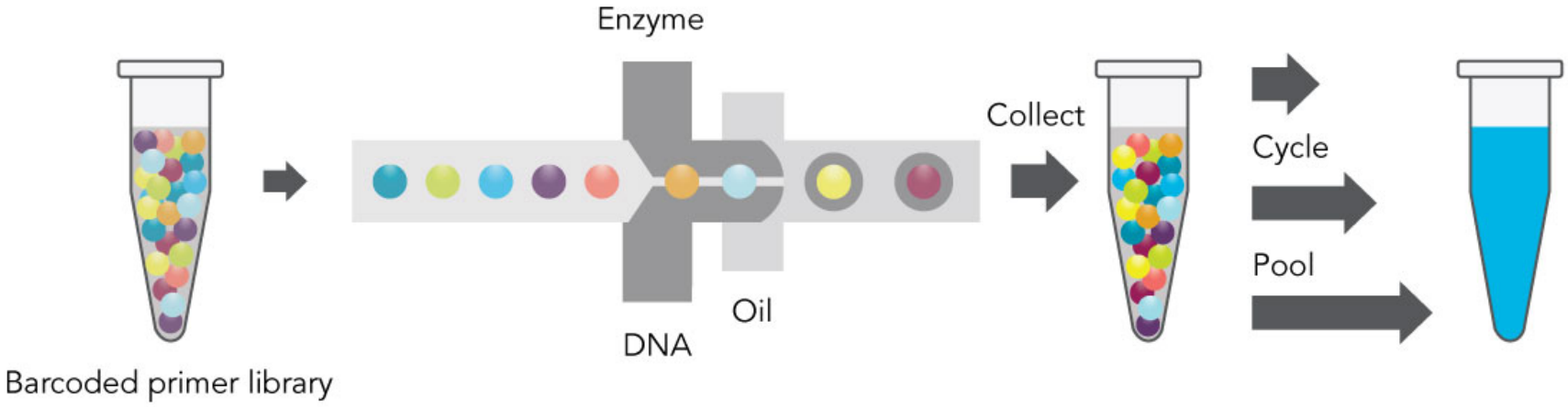
Mean read length ~6 Kbp
Max read length ~40 Kbp

Errors largely indels … error rate comparable to PacBio *in some hands!*

stay tuned! …

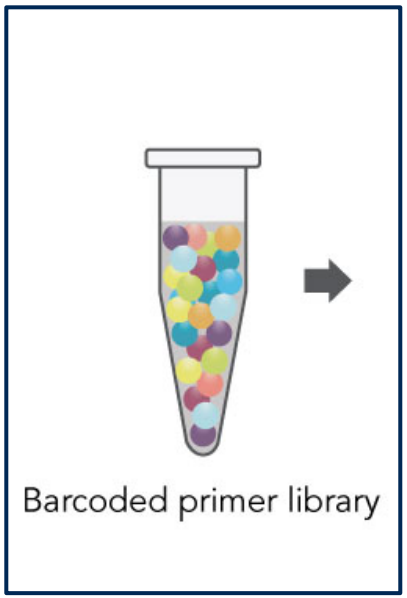
10X Genomics GemCode Platform

10X Genomics GemCode Platform



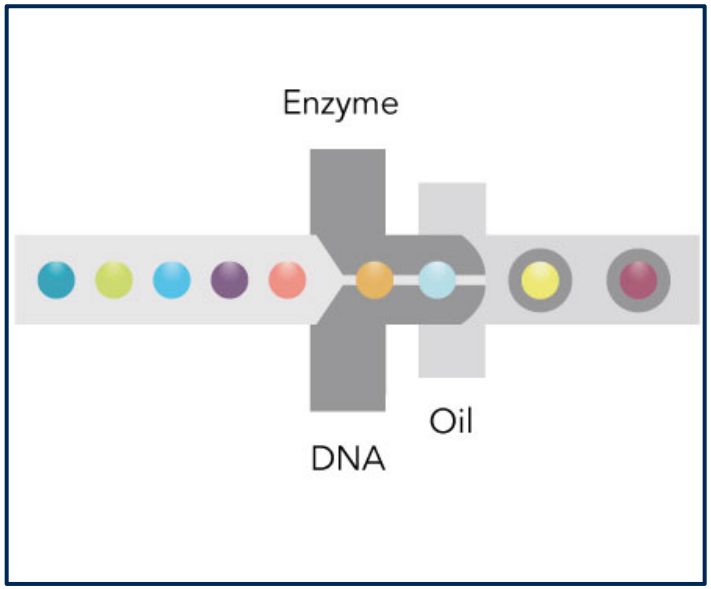
Produces up to 750k sets of “linked reads” - each pair of reads with the same barcode came from the same long DNA molecule, allowing haplotype resolution (*video*).

10X Genomics GemCode Platform

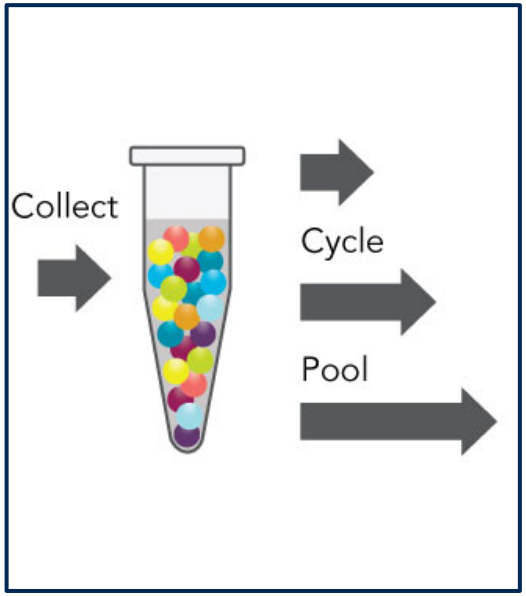


Each gel bead contains many copies of the same barcode, in molecules ready for Illumina library preparation.

Microfluidics encapsulates each bead in a reagent bubble in an oil stream; each bubble ideally contains one long DNA fragment (depends on correct titration).

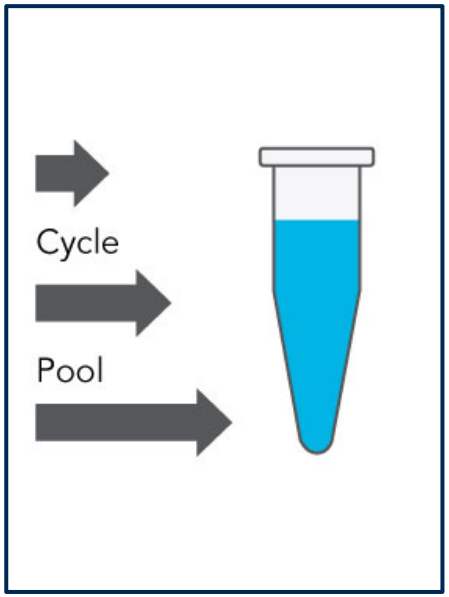


10X Genomics GemCode Platform

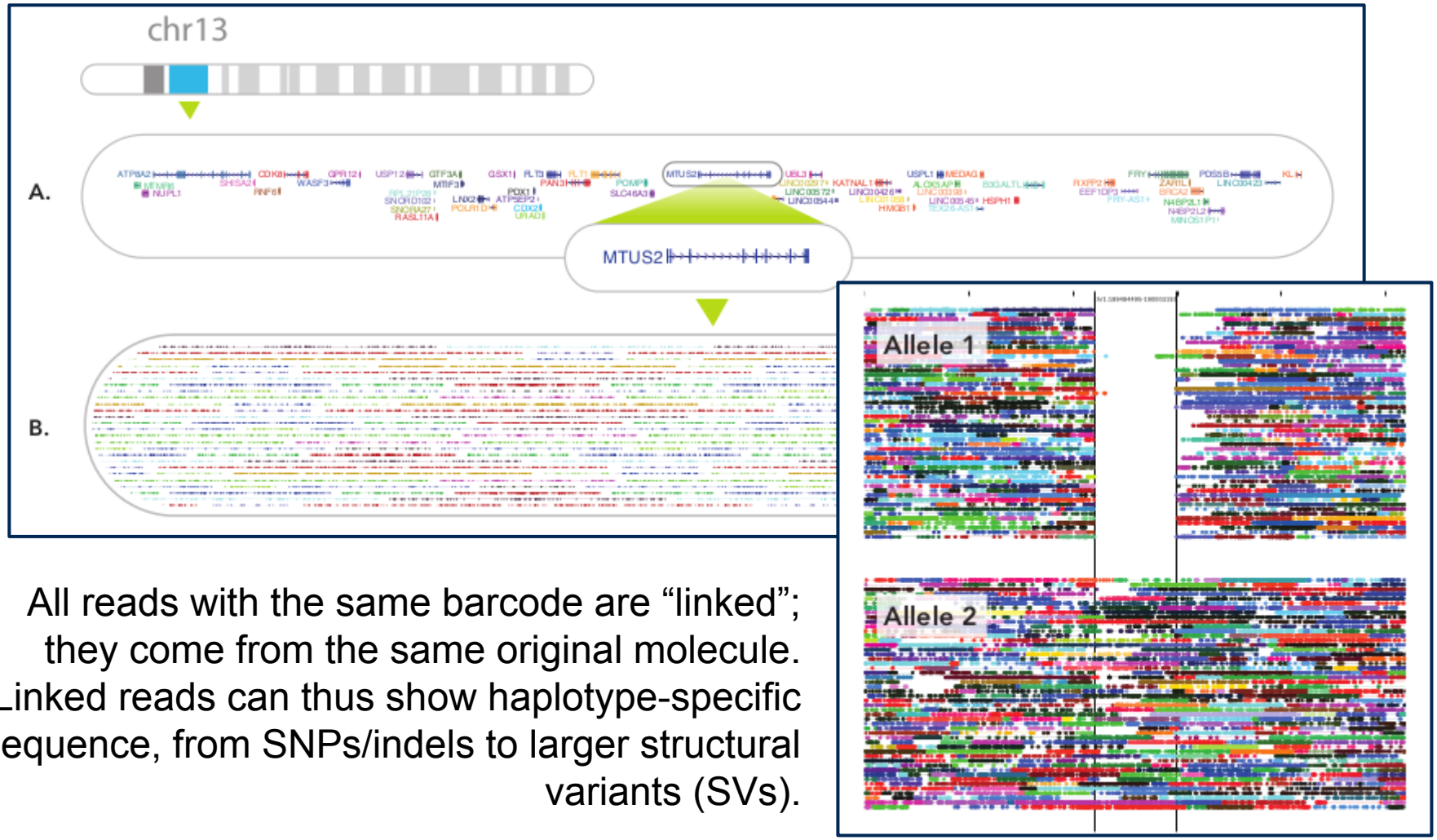


Bead / reagent bubbles are collectively temperature cycled, priming short barcoded PCR fragments from random locations on the trapped long DNA fragments. Each PCR fragment is ready for Illumina library preparation.

After fragment generation, bubbles are burst and Illumina library preparation is completed. Pooled, barcoded fragments are then sequenced on an Illumina instrument.



10X Genomics GemCode Platform



All reads with the same barcode are “linked”; they come from the same original molecule. Linked reads can thus show haplotype-specific sequence, from SNPs/indels to larger structural variants (SVs).

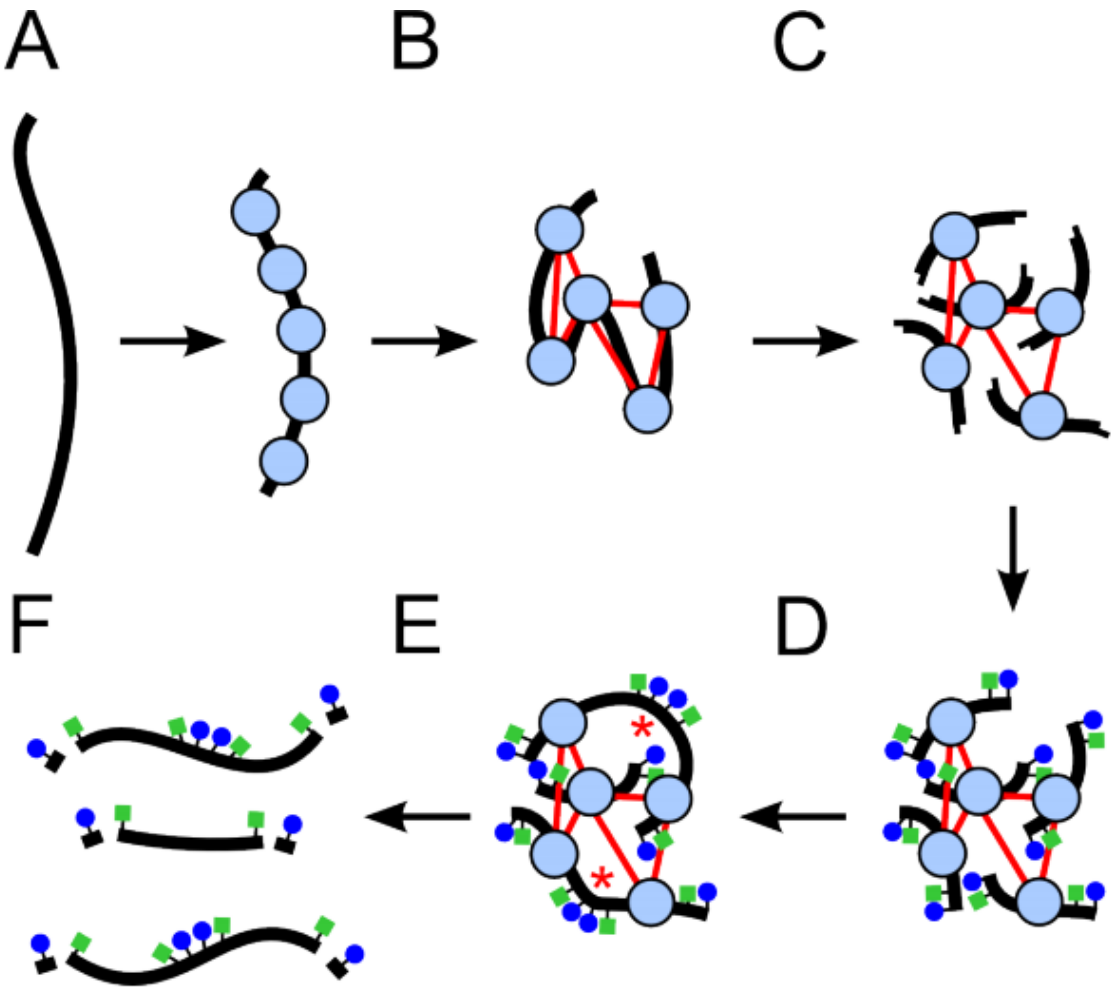
Dovetail Genomics

Uses Hi-C method *in vitro*, to capture clean profiles of chimeric reads with frequency that falls off ~logarithmically with distance.

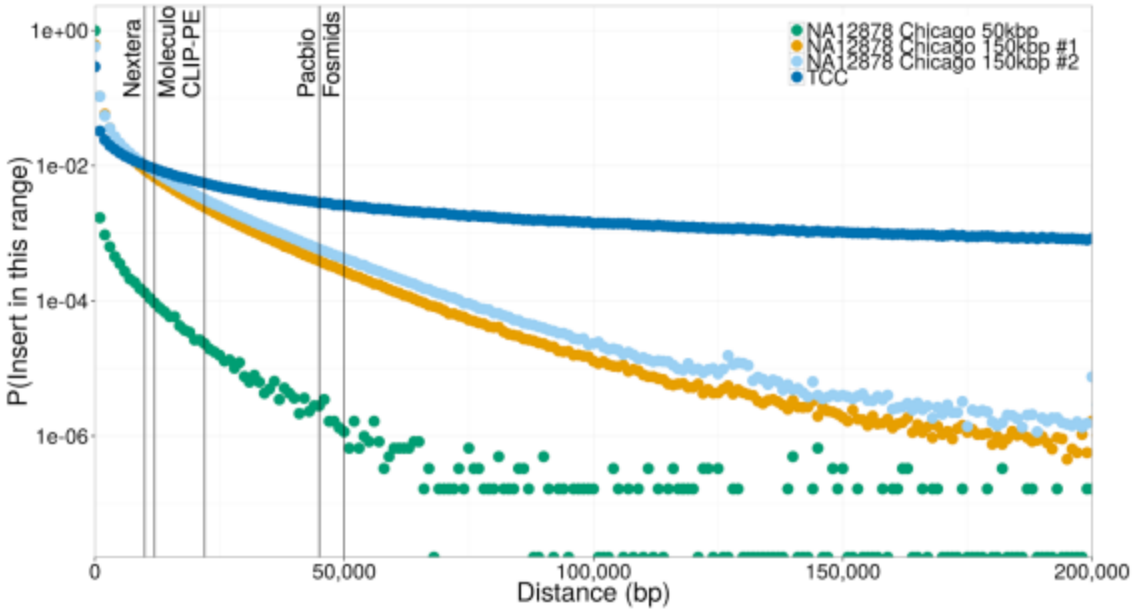
Dovetail Genomics

Long, naked DNA fragments, *in vitro*, are combined with engineered histones to coil DNA cleanly (without *in vivo* “signal”). DNA is cross-linked, cut, ends filled in, and ligated to create chimeric molecules, which are then enriched and go into Illumina sequencing.

Dovetail Genomics



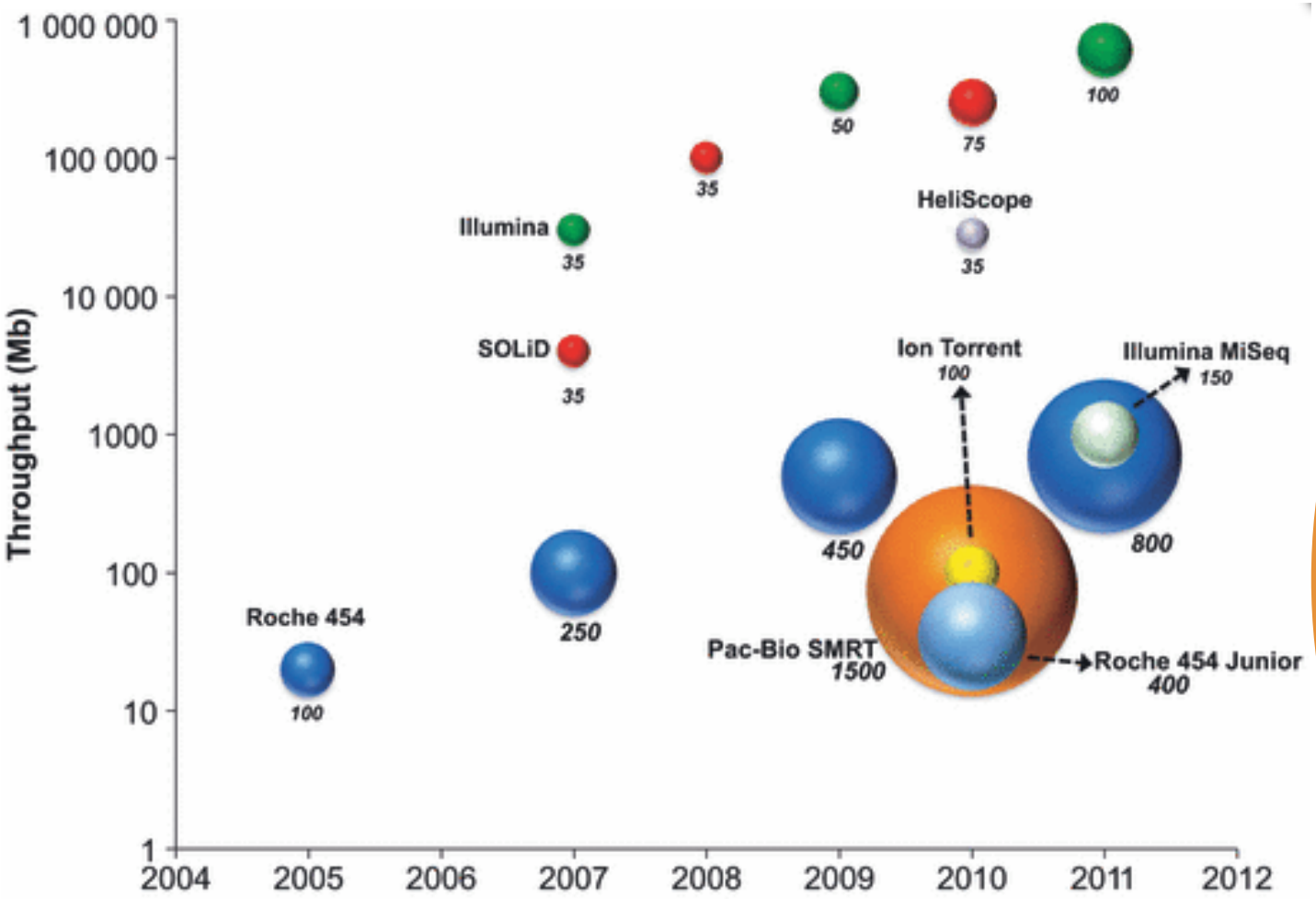
Dovetail Genomics



cHiCago libraries assess distances well beyond other technologies. Combined with Illumina contigs, (DiscoDove = DISCOVAR + Dovetail, or MERACULOUS + HiRise) scaffolds of 10s of Mbps have been achieved.

... Oxford Nanopore?

Sequencing Explosion



PacBio RS II

2013

adapted from Shokralla 2012 Molecular Ecology 21:1794

Tech Comparison




Feature	HiSeq2000	MiSeq	PacBio RS	Roche/454 FLX+
Number of reads	187 m/lane	15-18 m/lane	~40 K reads/SMRT cell	900-1500k/PTP
Read length	2 x 100 bp	2 x 250 bp	~ 3-10kb (120 min movie)	600-800 bp
Yield per run (PF data)	~37.5 Gb	~8.5 Gb	~Up to 0.2 Gb	~0.9 Gb
Pricing per run	$2,040	$1,179	$250	$6,800
Pricing per Gb	$54	$138	$1,250	$7,555
In Development	2 x 150 bp	2x300 bp	0.5G, 1Gb, 2Gb	???

Ryan Kim, ~Dec. 2012

Tech Comparison

- Non-technology considerations
  - error modes related to application
  - single-molecule preferred?
    - novel isoforms
    - haplotype resolution (phasing)
    - base modification
- Local expertise
  - library prep
  - secondary analysis
- Availability / turnaround time