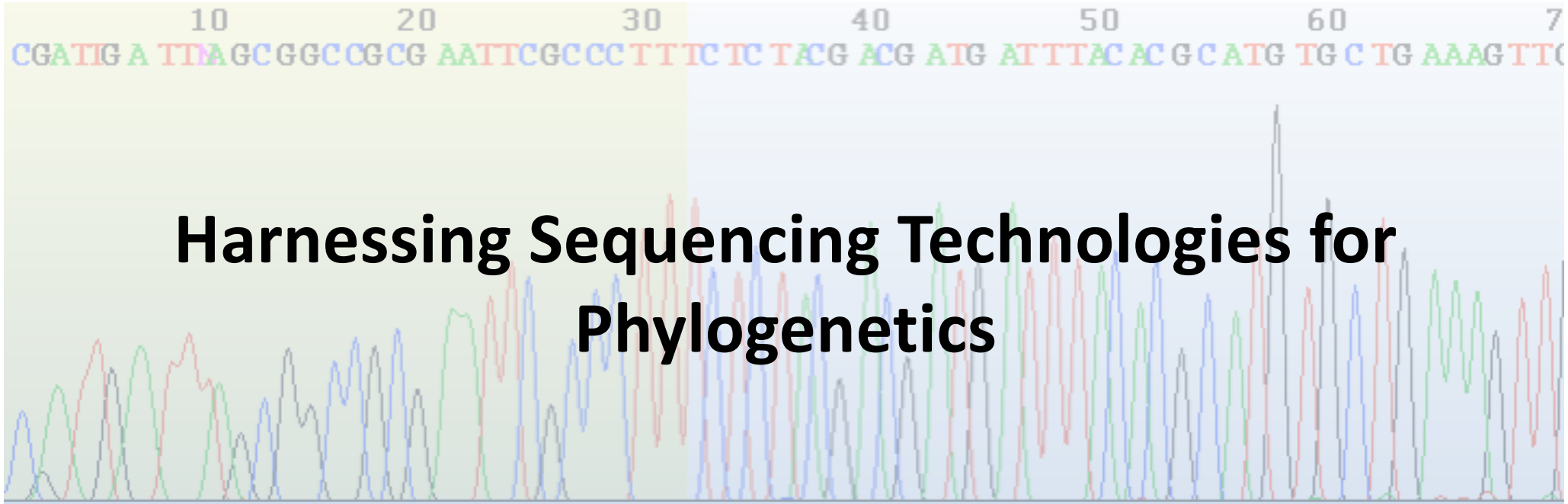


Harnessing Sequencing Technologies for Phylogenetics

Workshop in Applied Phylogenetics
March 12, 2017

Joanna C. Chiu
jcchiu@ucdavis.edu
Department of Entomology and Nematology
UC Davis

Outline

(1) Sequencing technologies : From Sanger sequencing to Oxford Nanopore Minion

(2) Types of NGS data for phylogenetics

Thanks to Brian Moore and Jonathan Eisen for ideas on this presentation!

GENOMICS

Sequencing all life captivates biologists

Project would read genomes of more than a million eukaryotes—just for starters

By **Elizabeth Pennisi**, *in Washington, D.C.*

Earth BioGenome Project (EBP)

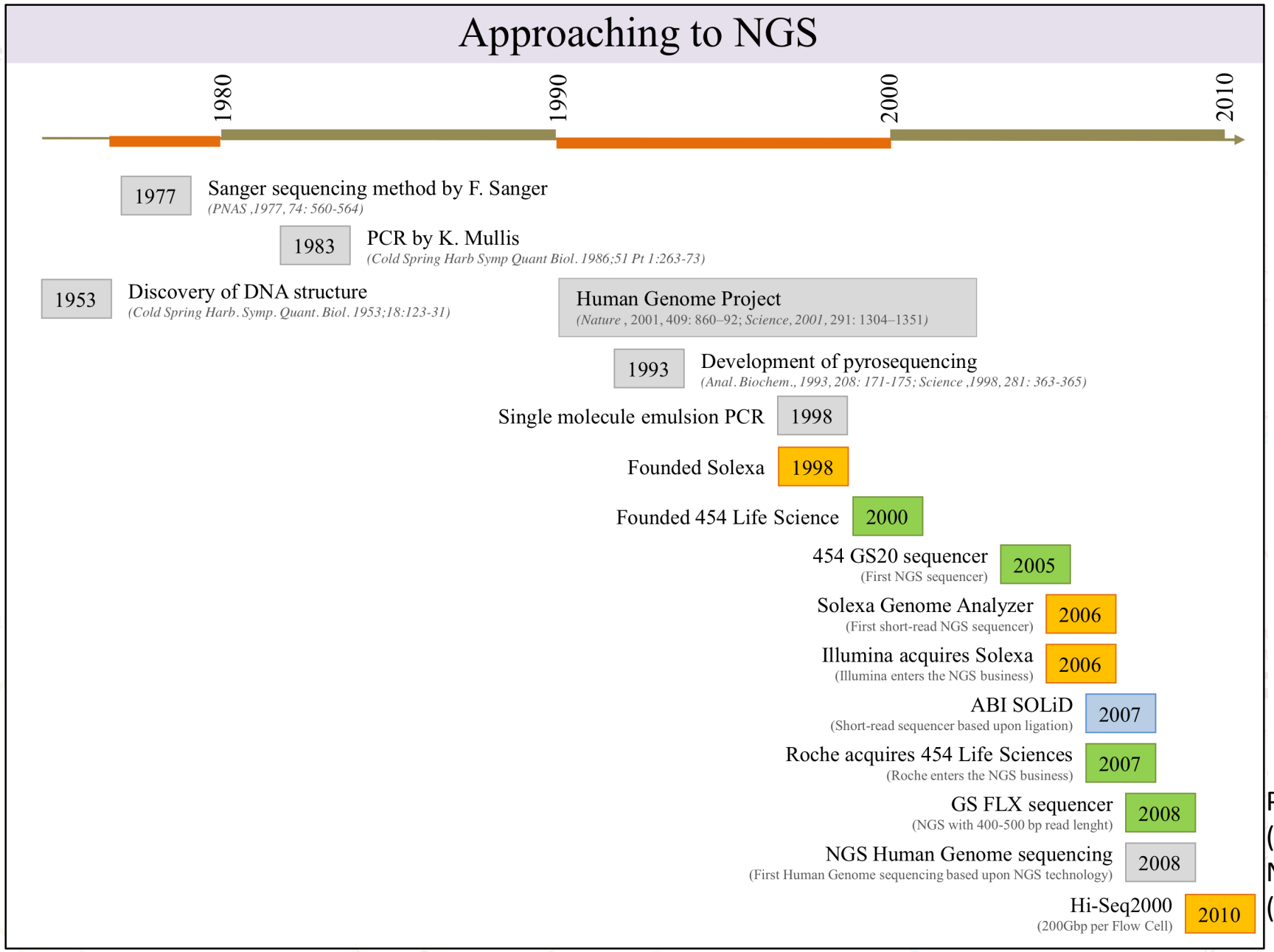
A head start

The Earth BioGenome Project could coordinate the efforts below and others that are already sequencing broad swaths of the planet's life.

PROJECT	YEAR STARTED	SEQUENCING GOAL	NUMBER SEQUENCED
G10K	2009	9478 vertebrate genera	100
i5K	2011	5000 arthropods	30
GIGA	2013	7000 marine invertebrates	60
GAGA	2016	All 300 ant genera	25
B10K	2016	All 10,500 bird species	300
AOCC	2013	101 African food crops	22

Pennisi (Science, March 3, 2017)

Timeline for the development of sequencing technologies

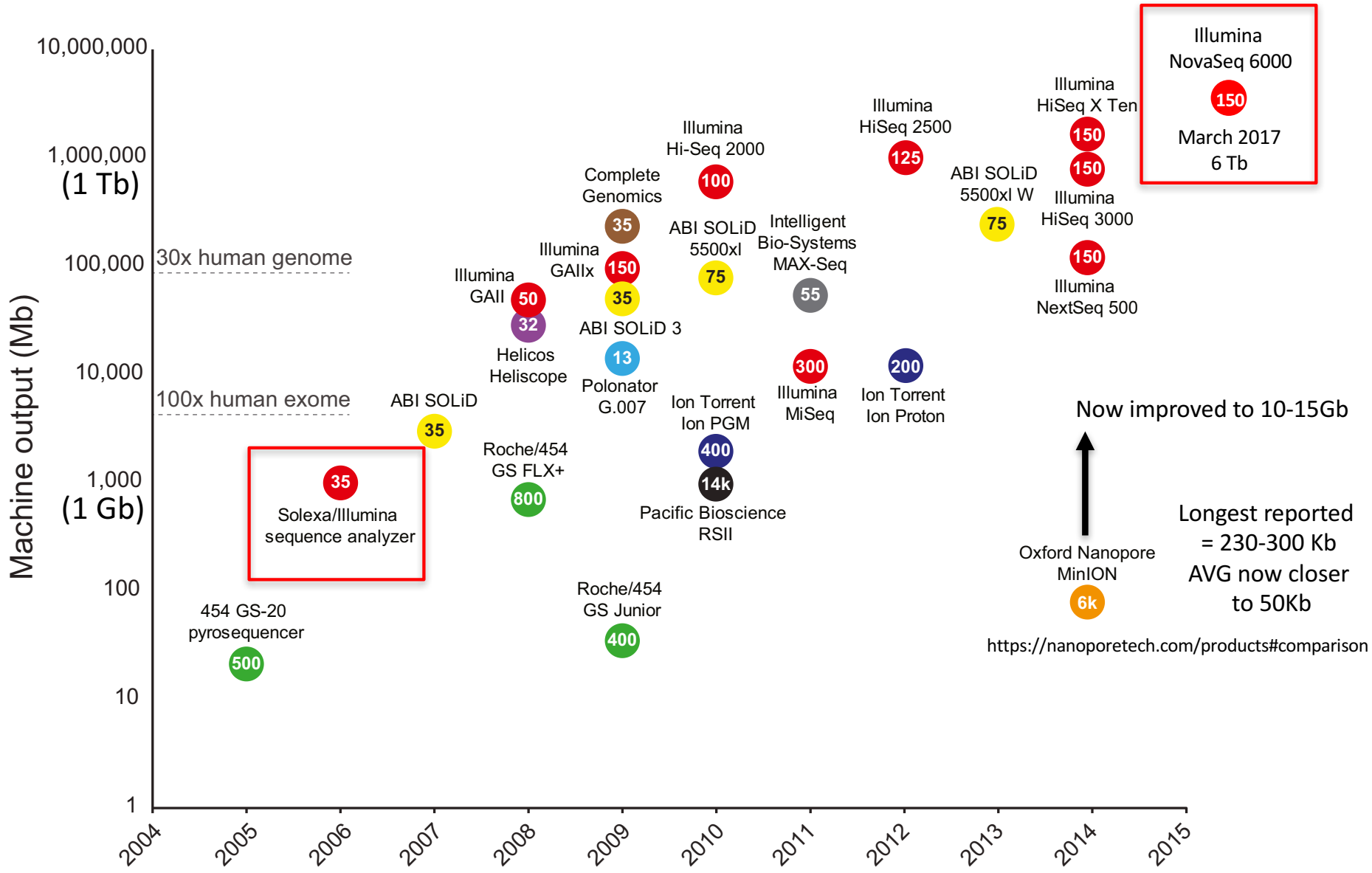


PacBio (2011)

Nanopore (5/2015)

From: Slideshare presentation of Cosentino Cristian and Jonathan Eisen

Timeline and Comparison of output for sequencing platforms



Numbers inside circles = read-length in bp

Reuter et al. (Mol Cell 2015)

Work by Maxam & Gilbert and Sanger started it all

The Nobel Prize in Chemistry 1980
Paul Berg, Walter Gilbert, Frederick Sanger

The Nobel Prize in Chemistry 1980

Paul Berg

Walter Gilbert

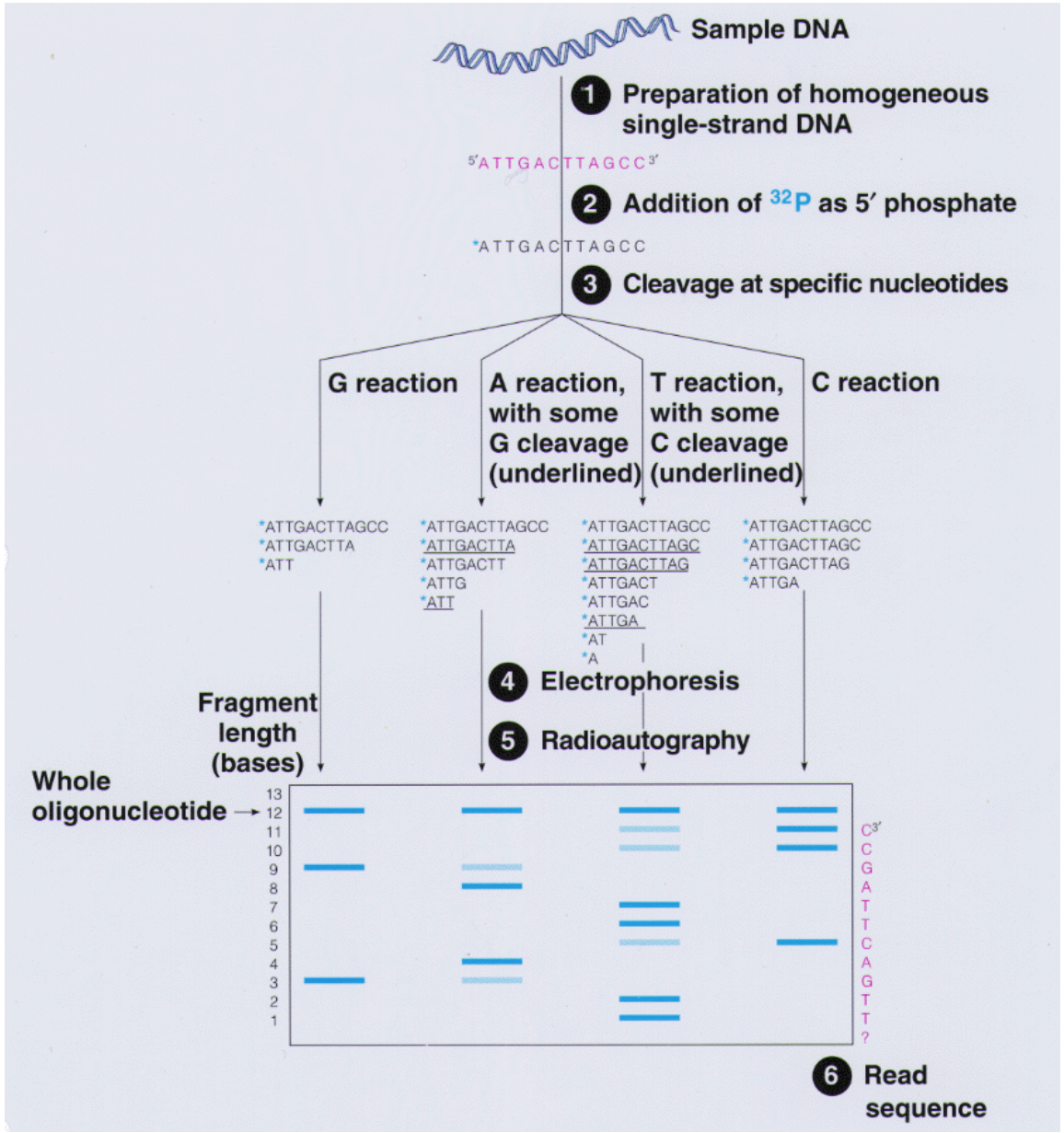
Frederick Sanger

The Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg *"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"*, the other half jointly to Walter Gilbert and Frederick Sanger *"for their contributions concerning the determination of base sequences in nucleic acids"*.

http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/

Slide from Jonathan Eisen (UCD Bodega Bay Workshop in Applied Phylogenetics 2015)

Maxam-Gilbert Sequencing



Mathews and van Holde (Biochemistry 4th ed.)

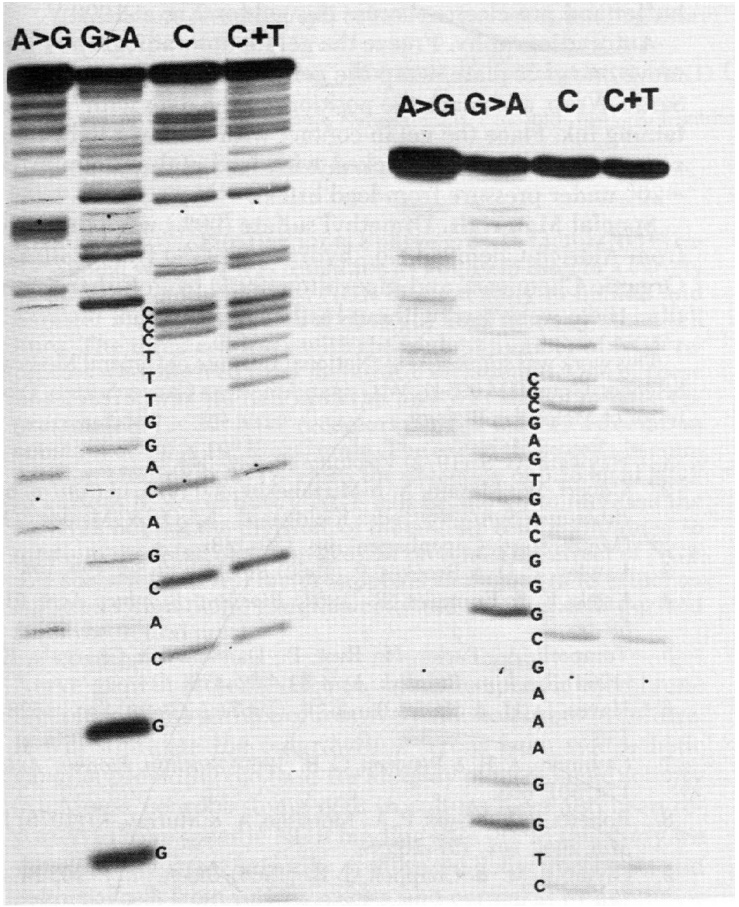
A new method for sequencing DNA

(DNA chemistry/dimethyl sulfate cleavage/hydrazine/piperidine)

ALLAN M. MAXAM AND WALTER GILBERT

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Contributed by Walter Gilbert, December 9, 1976



Maxam and Gilbert (PNAS 1977)

Capacity = For 40 cm gel, 100 bp

Sanger Sequencing

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

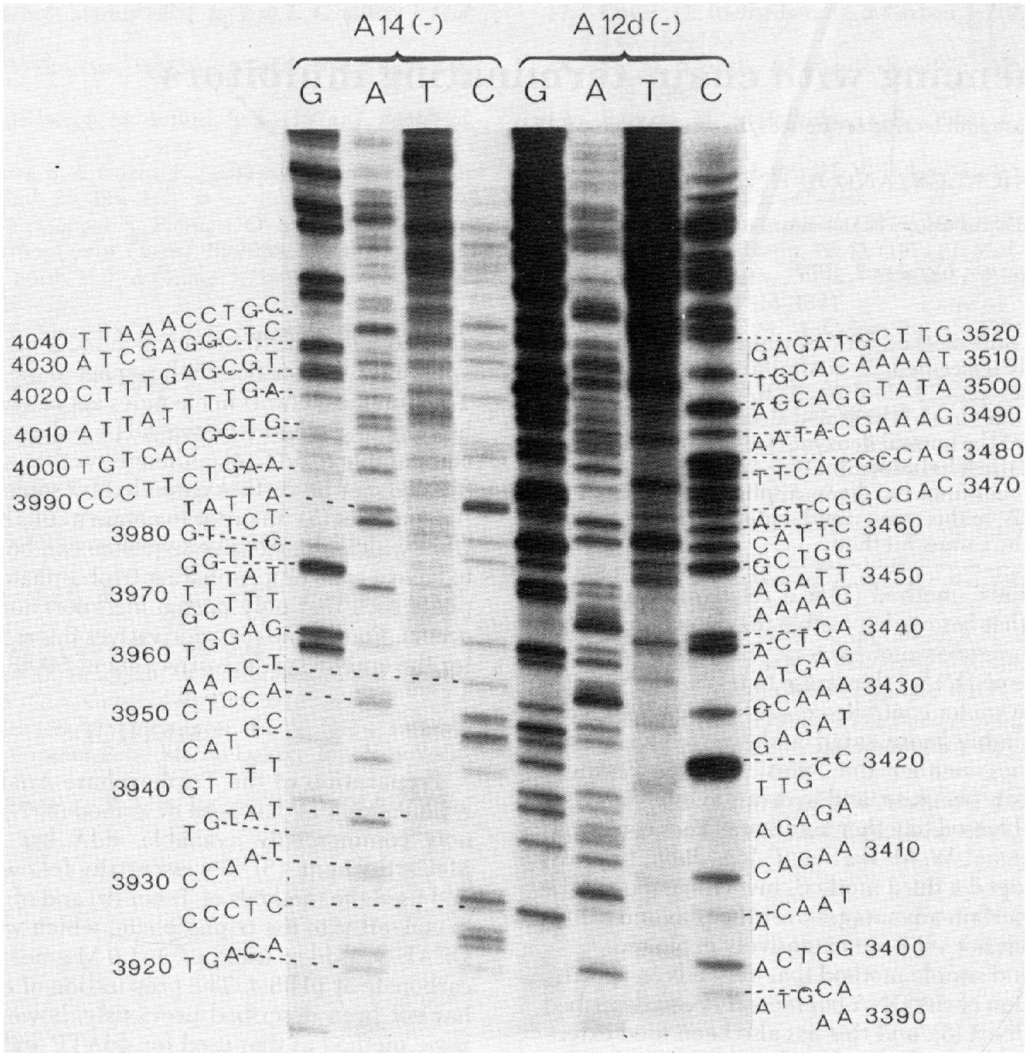
(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

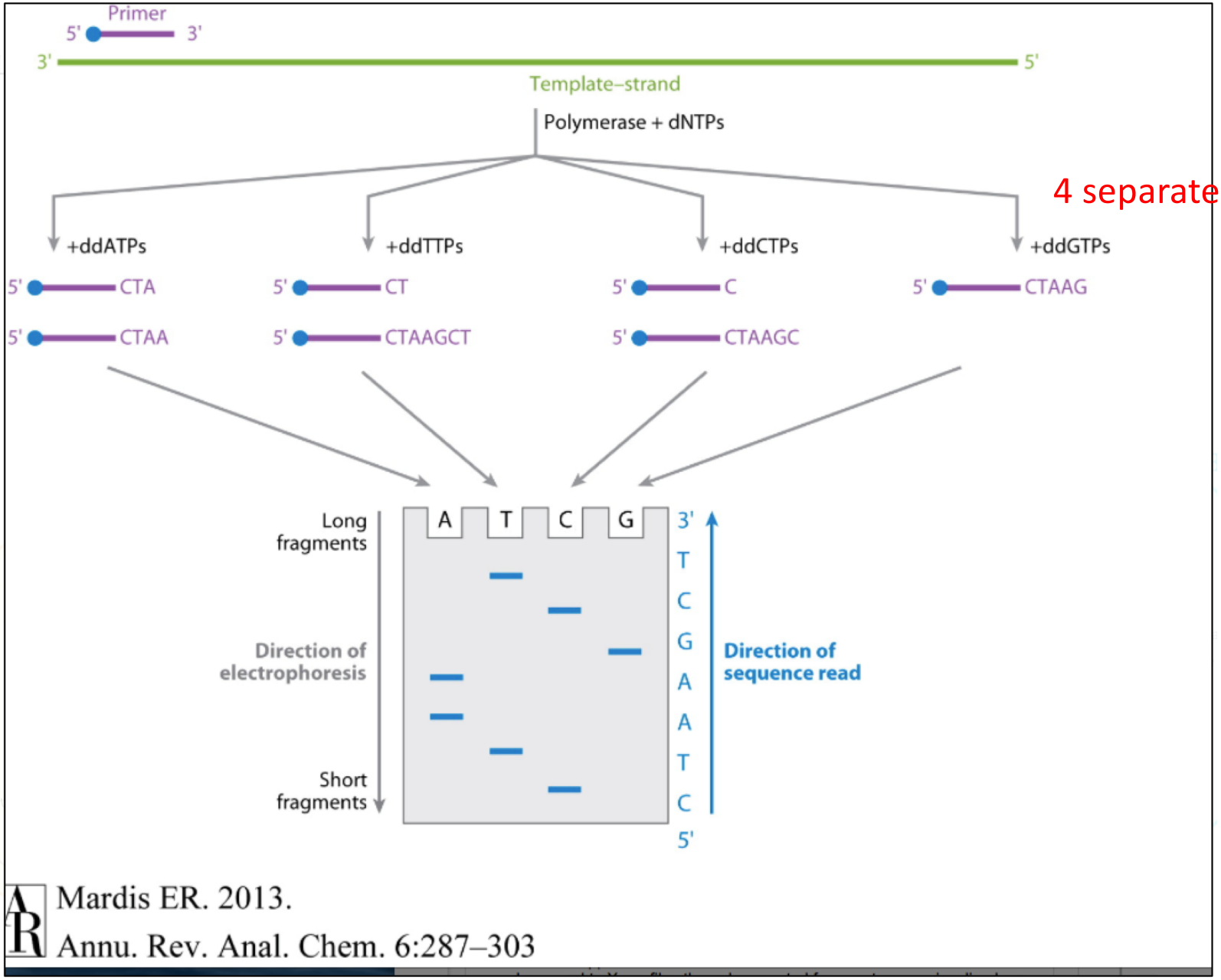
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

- Developed in 1977
- Chain termination using ddNTPs
- First sequenced genome: Bacteriophage ϕX174 (5,368bp)



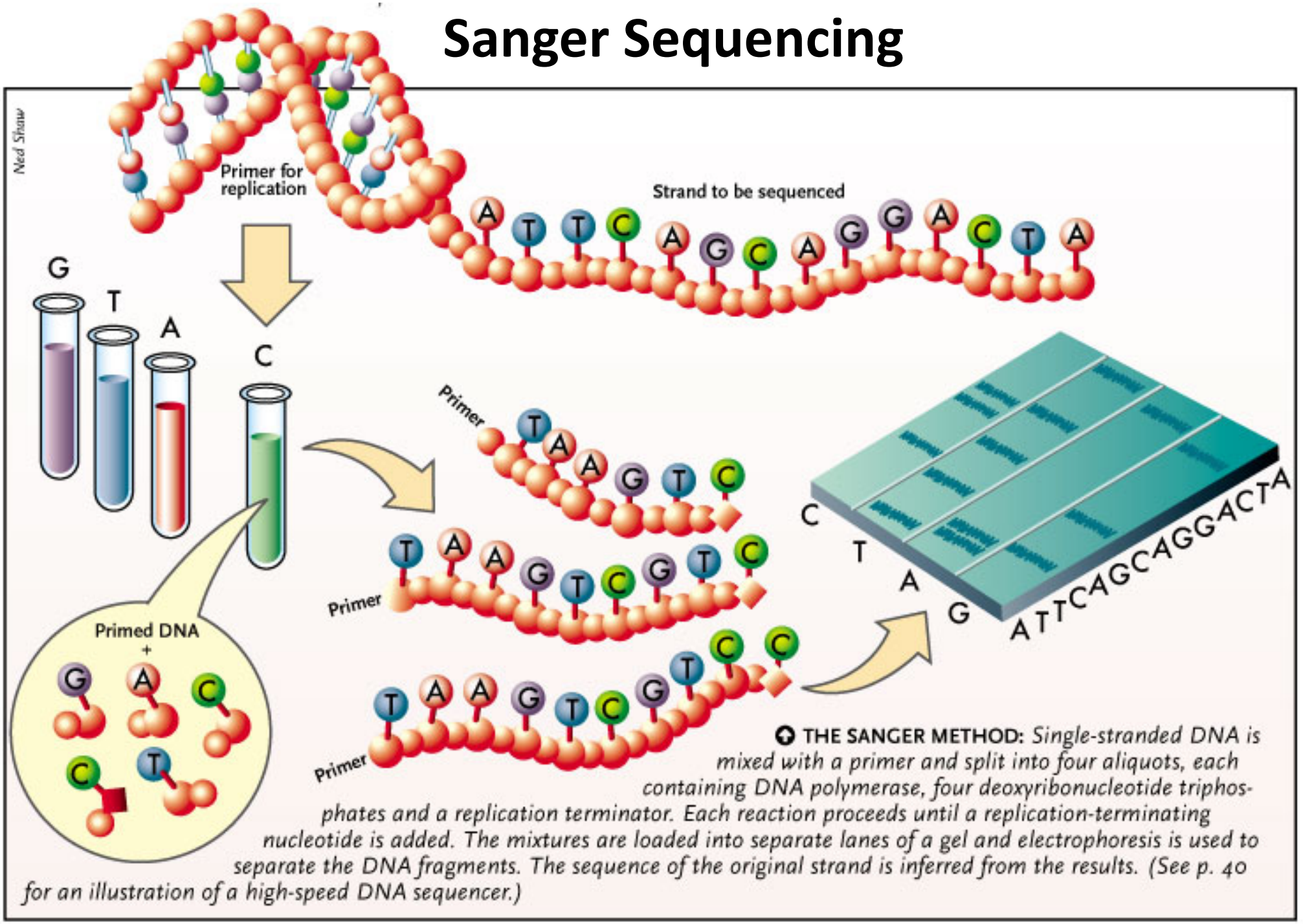
Sanger Sequencing



4 separate reactions

Mardis ER. 2013.
Annu. Rev. Anal. Chem. 6:287–303

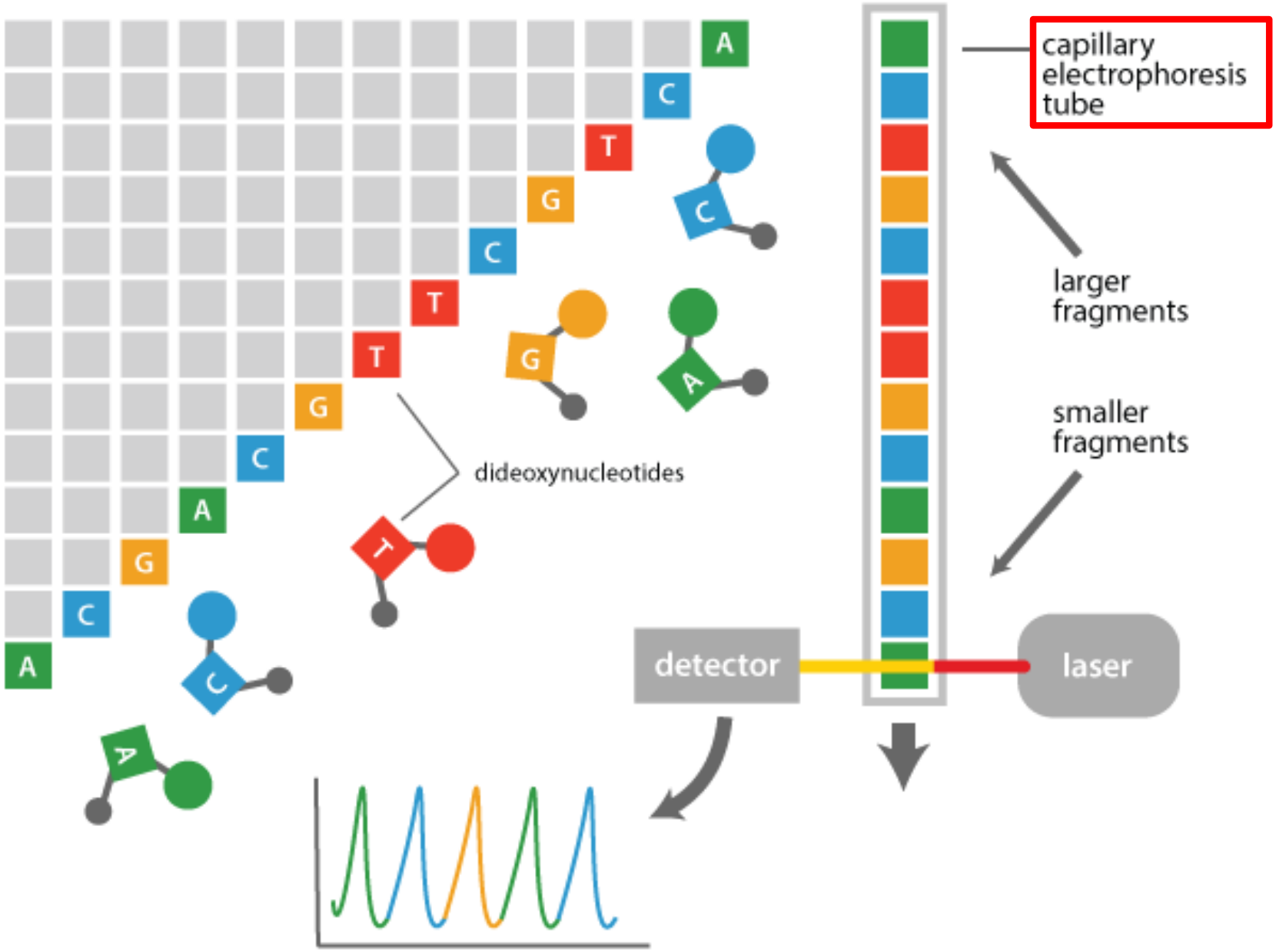
Sanger Sequencing



THE SANGER METHOD: *Single-stranded DNA is mixed with a primer and split into four aliquots, each containing DNA polymerase, four deoxyribonucleotide triphosphates and a replication terminator. Each reaction proceeds until a replication-terminating nucleotide is added. The mixtures are loaded into separate lanes of a gel and electrophoresis is used to separate the DNA fragments. The sequence of the original strand is inferred from the results. (See p. 40 for an illustration of a high-speed DNA sequencer.)*

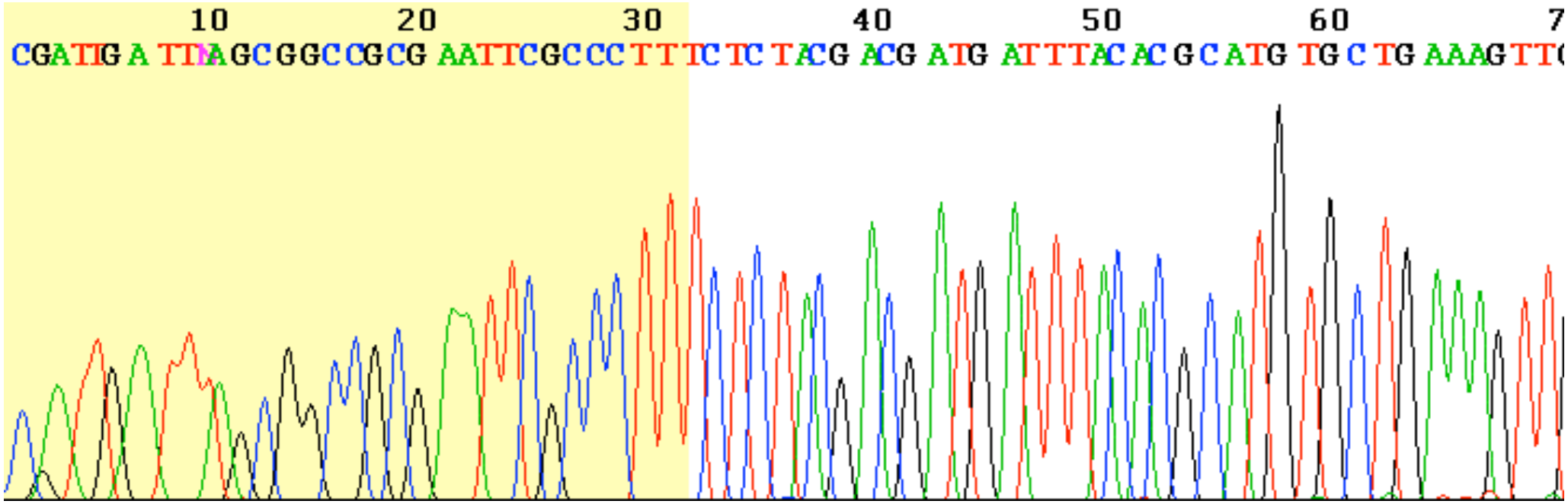
The Scientist 2004, 18(18):44

Sanger Sequencing – Automation I (e.g. ABI sequencers)



https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_introduction.php

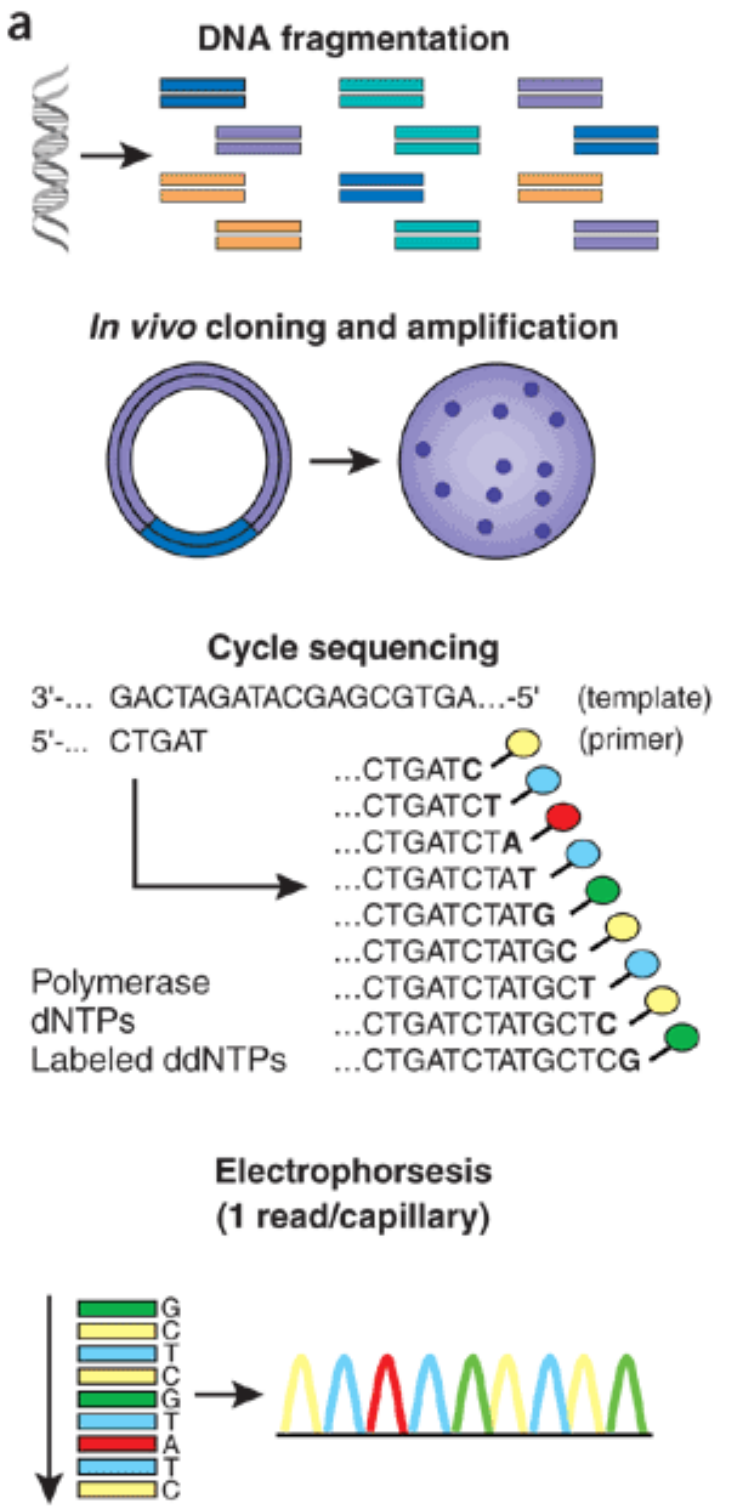
Sanger Sequencing Trace/ Chromatograms



Up to 800 – 1000 bases per read

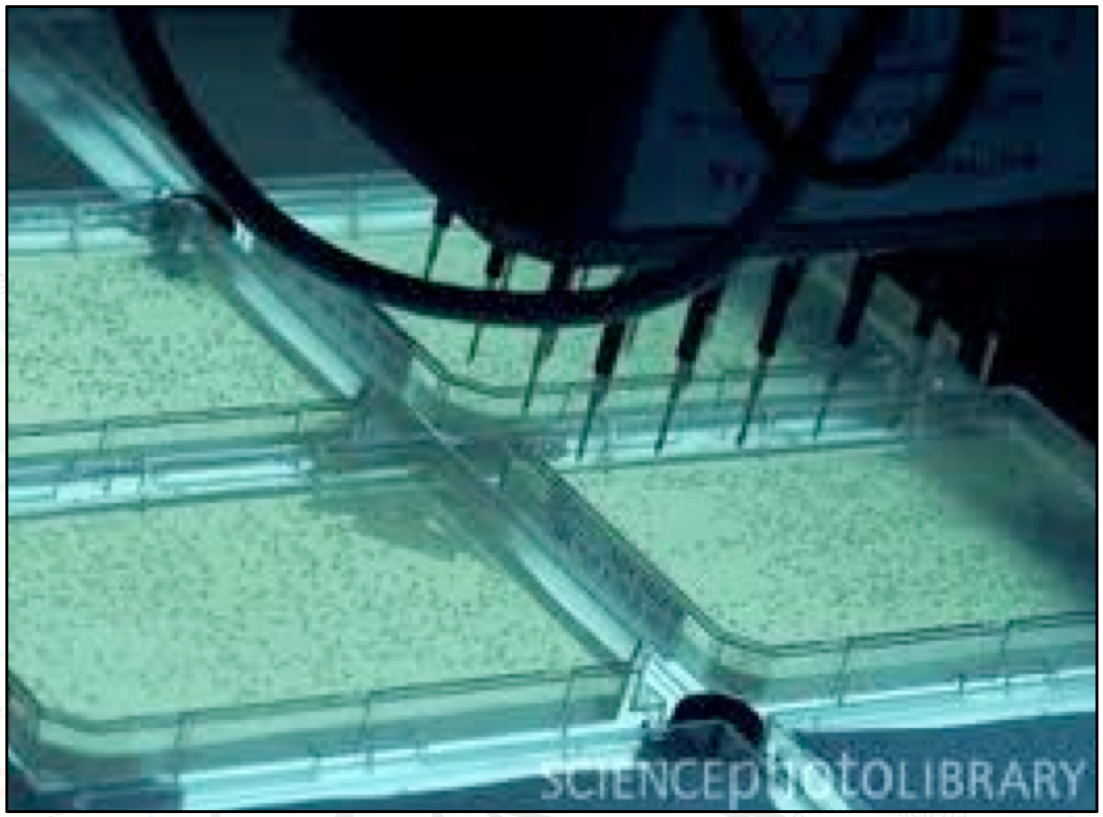
High Throughput Sanger Shotgun Sequencing

Not really high throughput if it's manual



Nature Biotechnology 26, 1135 - 1145 (2008)

Sanger Sequencing – Automation II



Slide from Jonathan Eisen (UCD Bodega Bay Workshop in Applied Phylogenetics 2015)

Automated Sanger Sequencing Highlights

- 1991: ESTs by C. Venter
- 1995: *H. influenzae* shotgun genome
- 1996: Yeast, archaeal genomes
- 1998: 1st animal genome – *C. elegans*
- 1999: *Drosophila melanogaster* shotgun genome
- 2000: *Arabidopsis thaliana* genome
- 2000: Human genome (Lander et al. 2001)
  ($2.7 billion in FY 1991 – genome.gov)

Slide from Jonathan Eisen (UCD Bodega Bay Workshop in Applied Phylogenetics 2015)

Next-generation sequencing (NGS)

Sequence by Synthesis (SBS) Pyrosequencing Sequence by Ligation

Slide from Jonathan Eisen (UCD Bodega Bay Workshop in Applied Phylogenetics 2015)

Next-generation sequencing (NGS) – paradigm shift but still relies on PCR

2 Clonal Amplification

Increase input materials

DNA need to attach to a substrate for clonal amplification

Bridge PCR

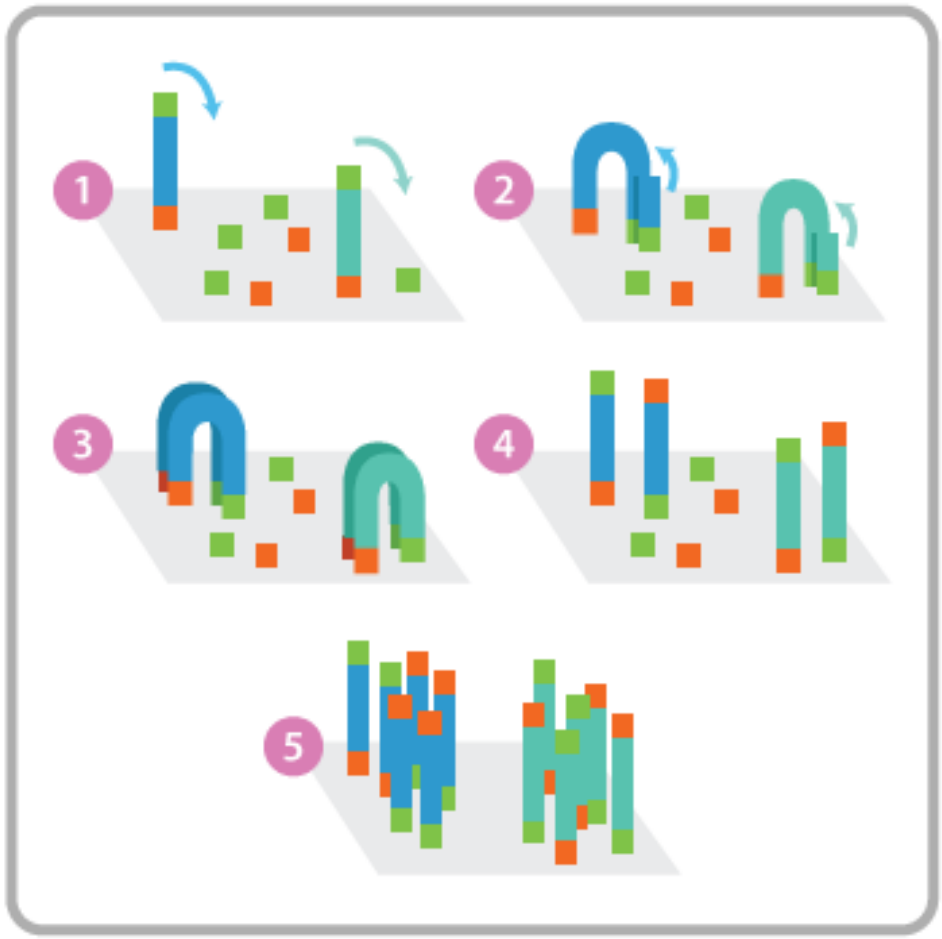
Illumina

Emulsion PCR

Roche 454
ABI SOLiD

https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_introduction.php

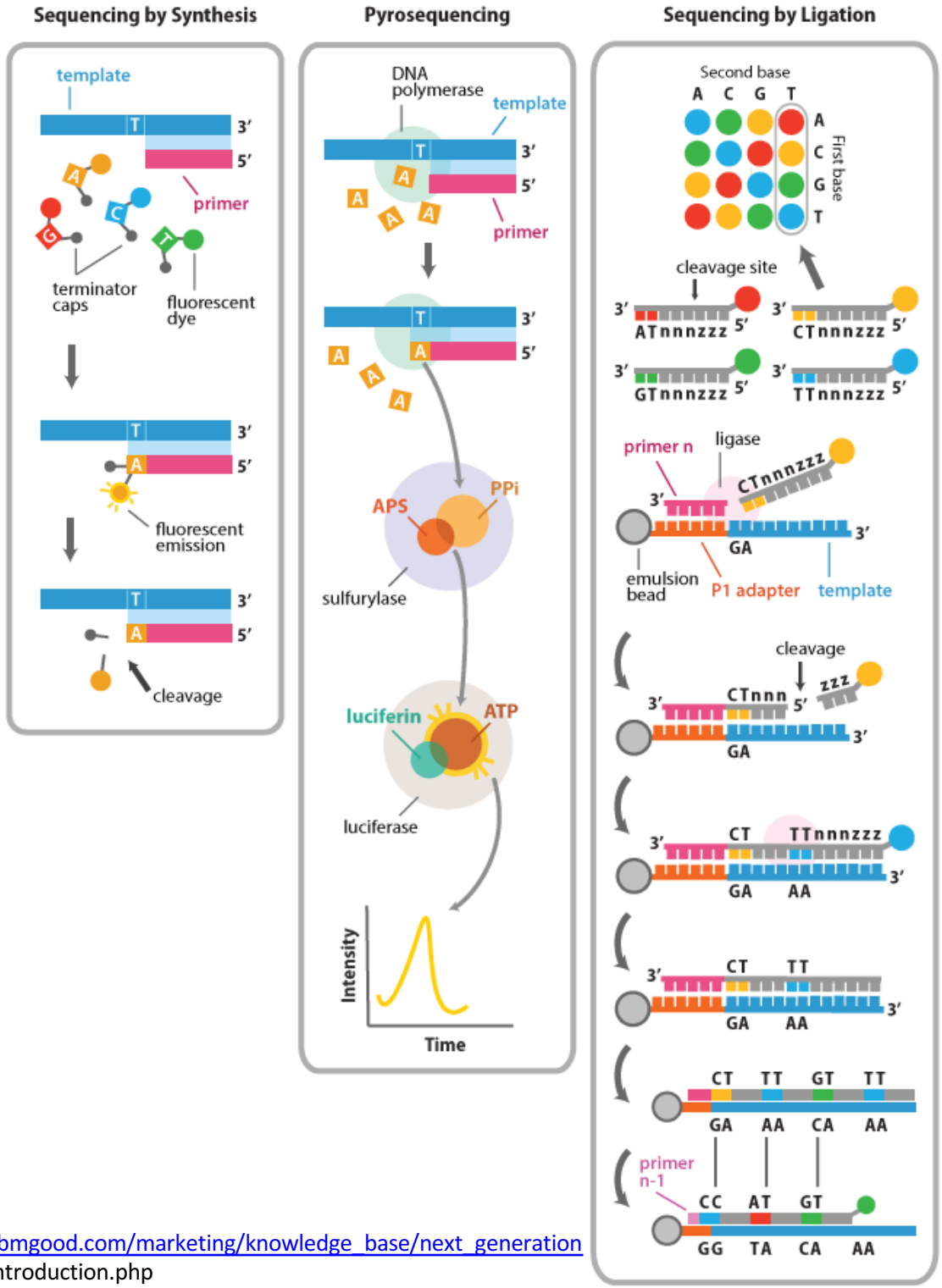
Clonal amplification step



Illumina

Roche 454
ABI SOLiD

https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_introduction.php



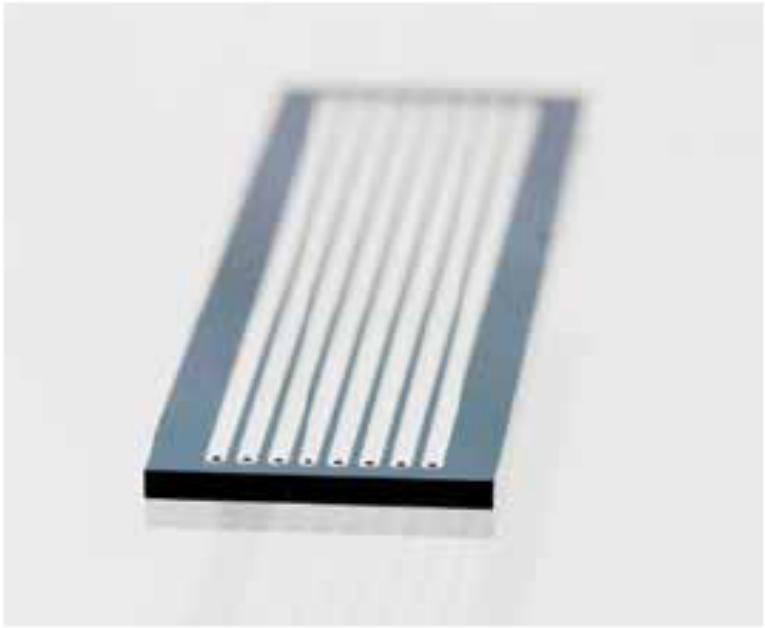
https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_introduction.php

Cyclic Array Sequencing

Each method uses different chemistry

Illumina/Solexa

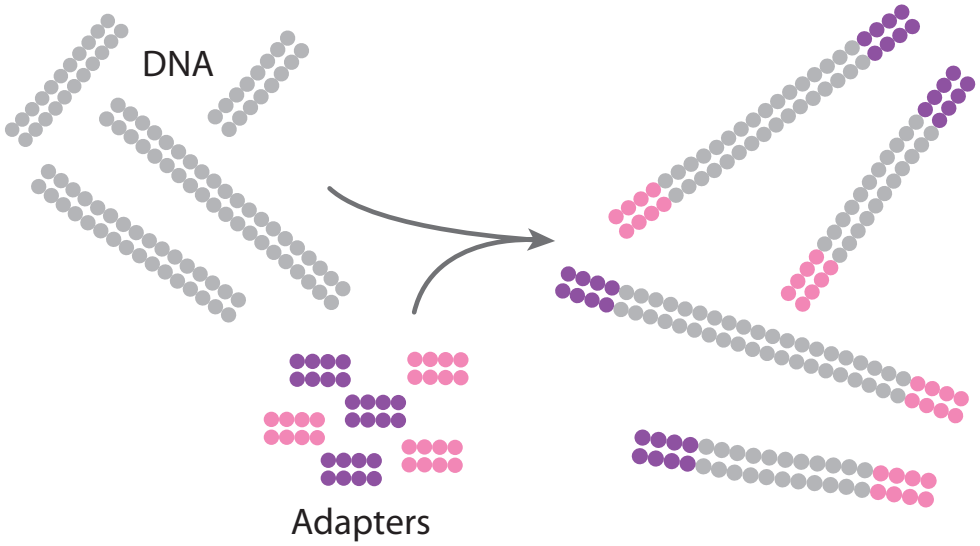
Illumina HiSeq



8-lane flow cell

Source: www.illumina.com

Prepare Genomic DNA Sample – Library prep



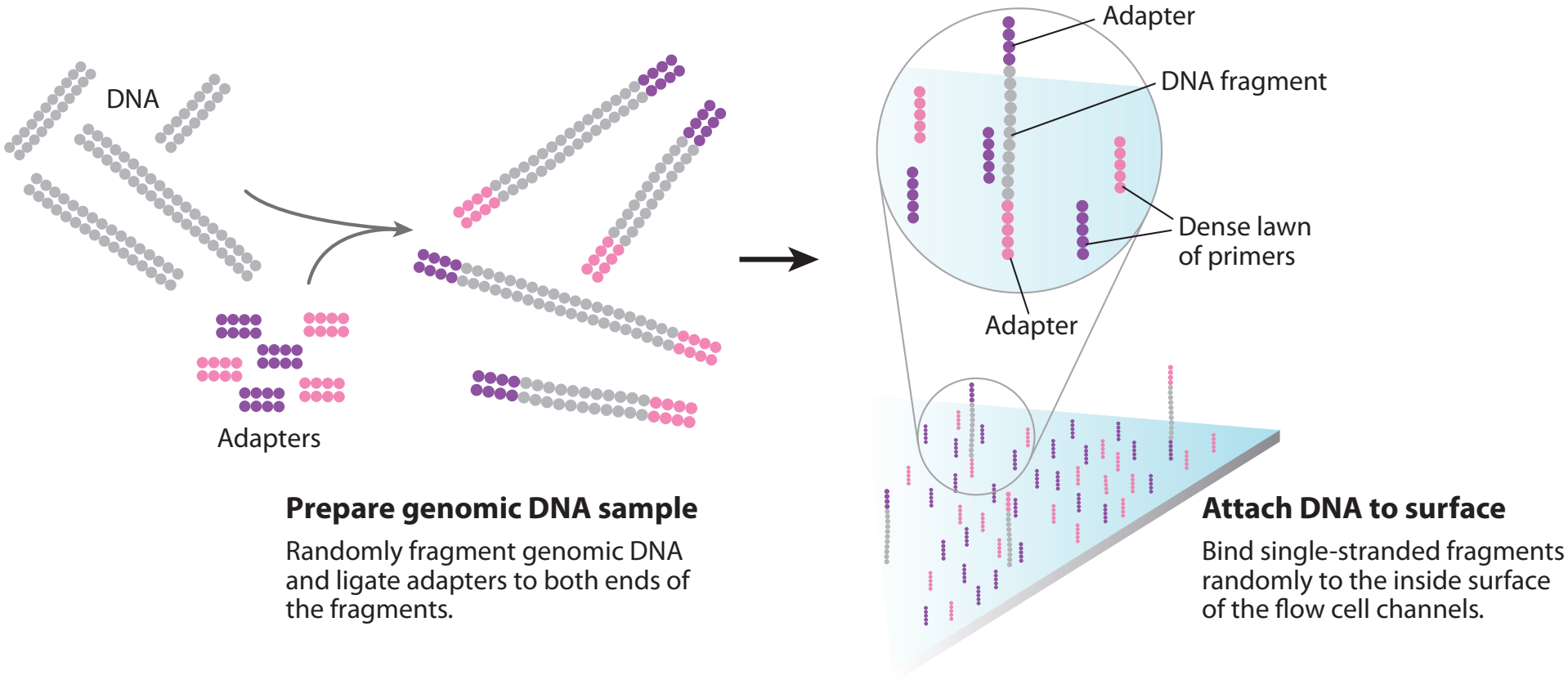
Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

- Randomly fragment genomic DNA and ligate adapters to both end of fragments
- Size selection

E. Mardis (Annu. Rev. Genomics Hum. Genet 2008)

Load samples into Flow Cell



Prepare genomic DNA sample

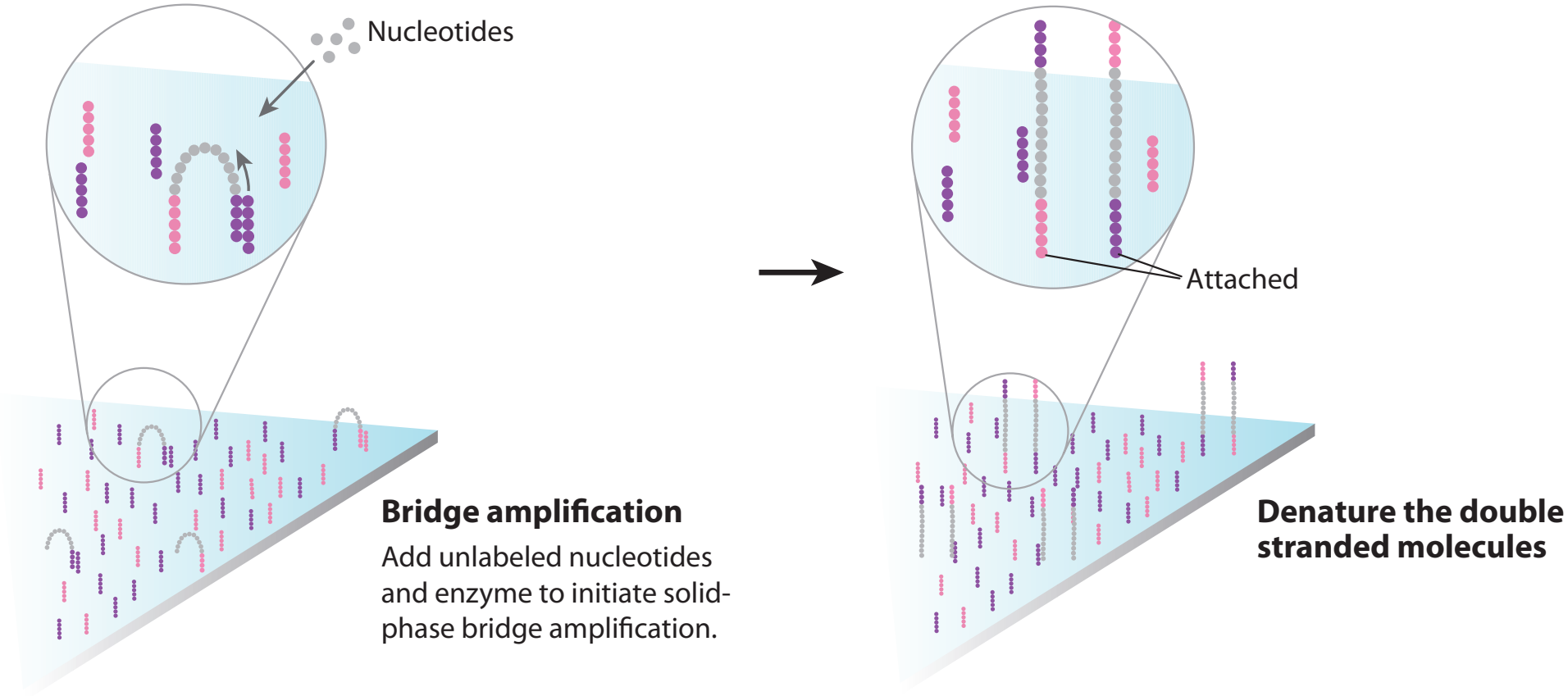
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

E. Mardis (Annu. Rev. Genomics Hum. Genet 2008)

Bridge amplification to generate clusters (amplify signal)



Bridge amplification

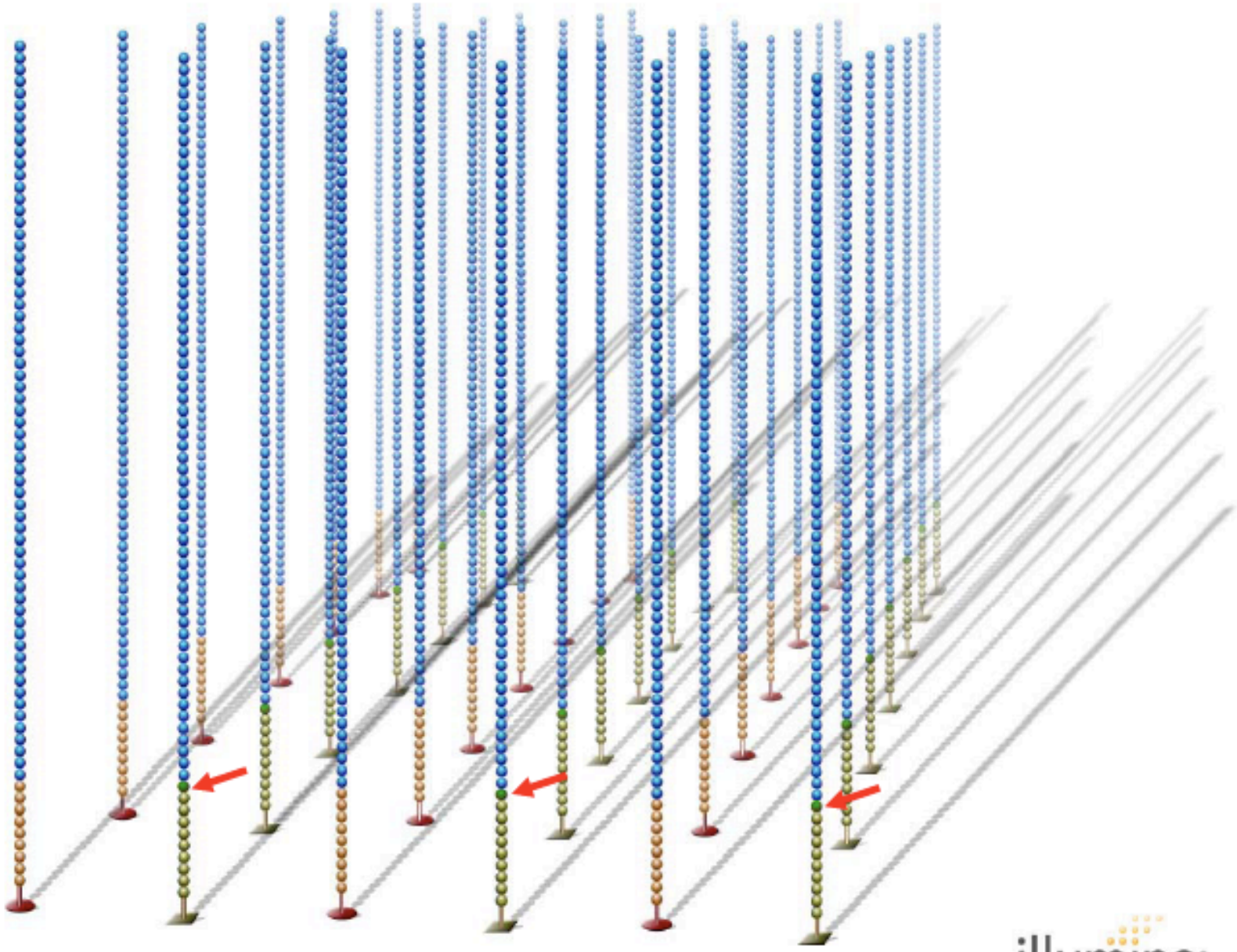
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules

E. Mardis (Annu. Rev. Genomics Hum. Genet 2008)

Cluster Generation

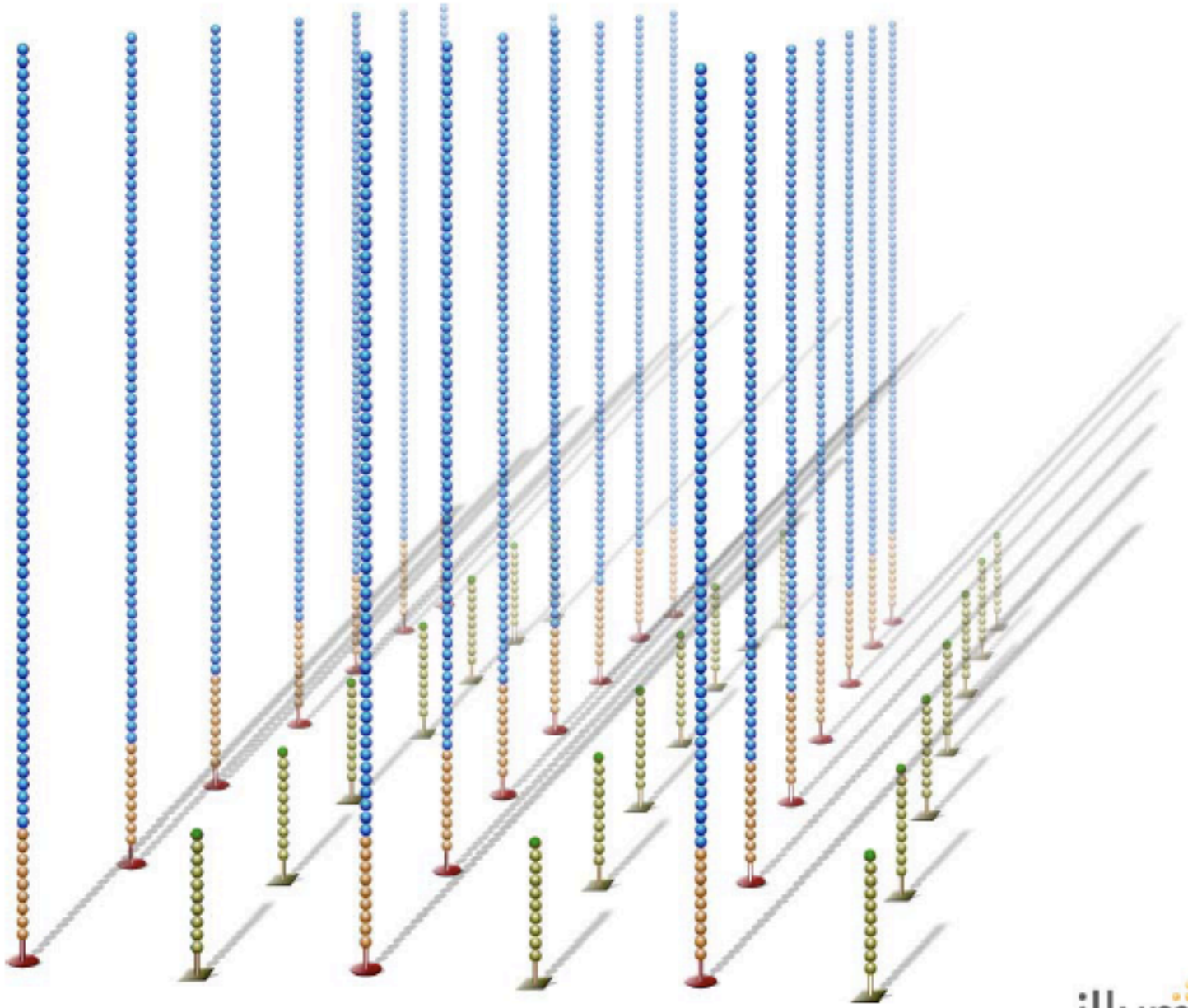
- dsDNA bridges denatured
- Reverse strands cleaved and washed away



illumina

Cluster Generation

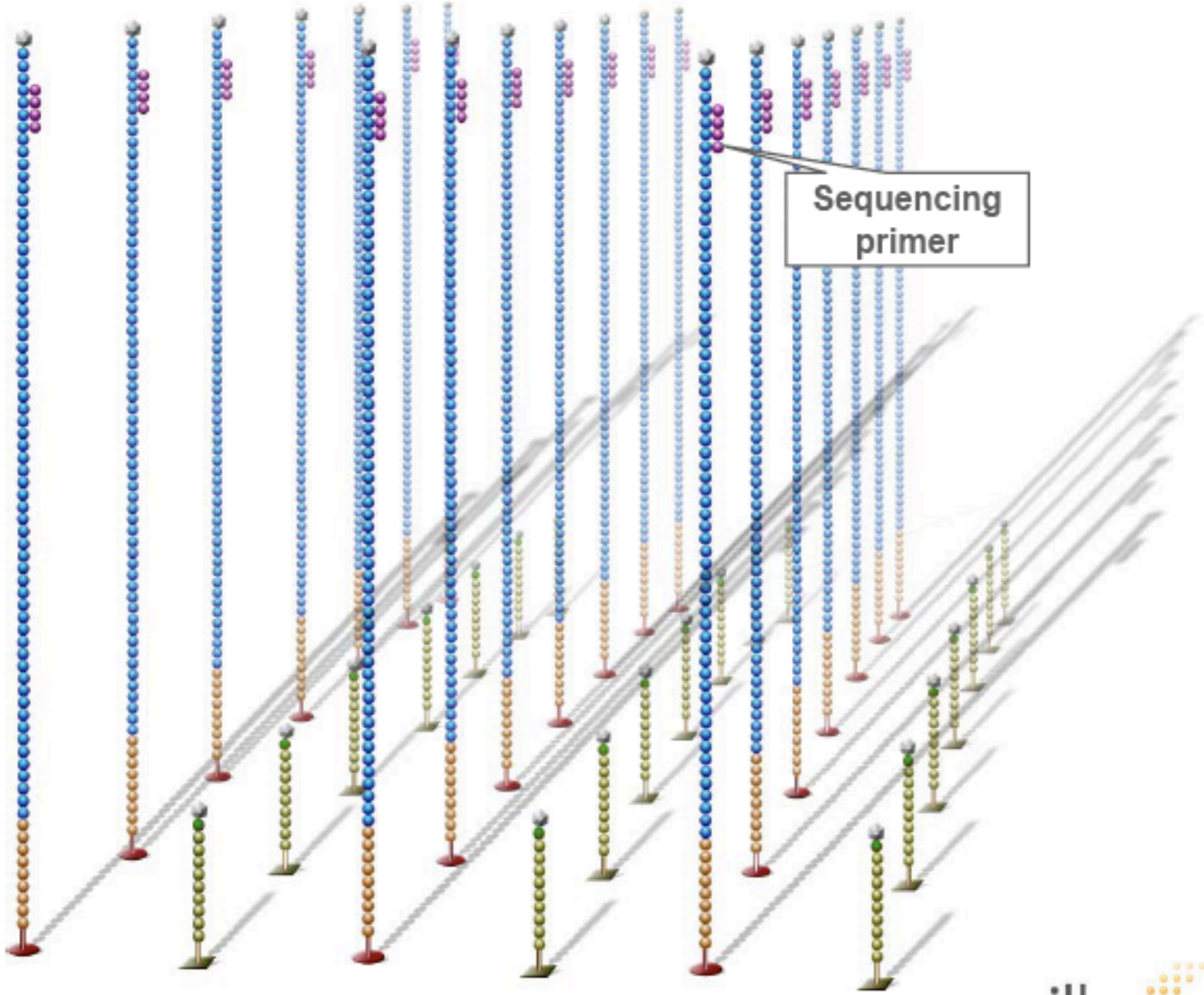
- Leaving a cluster with forward strands only



illumina

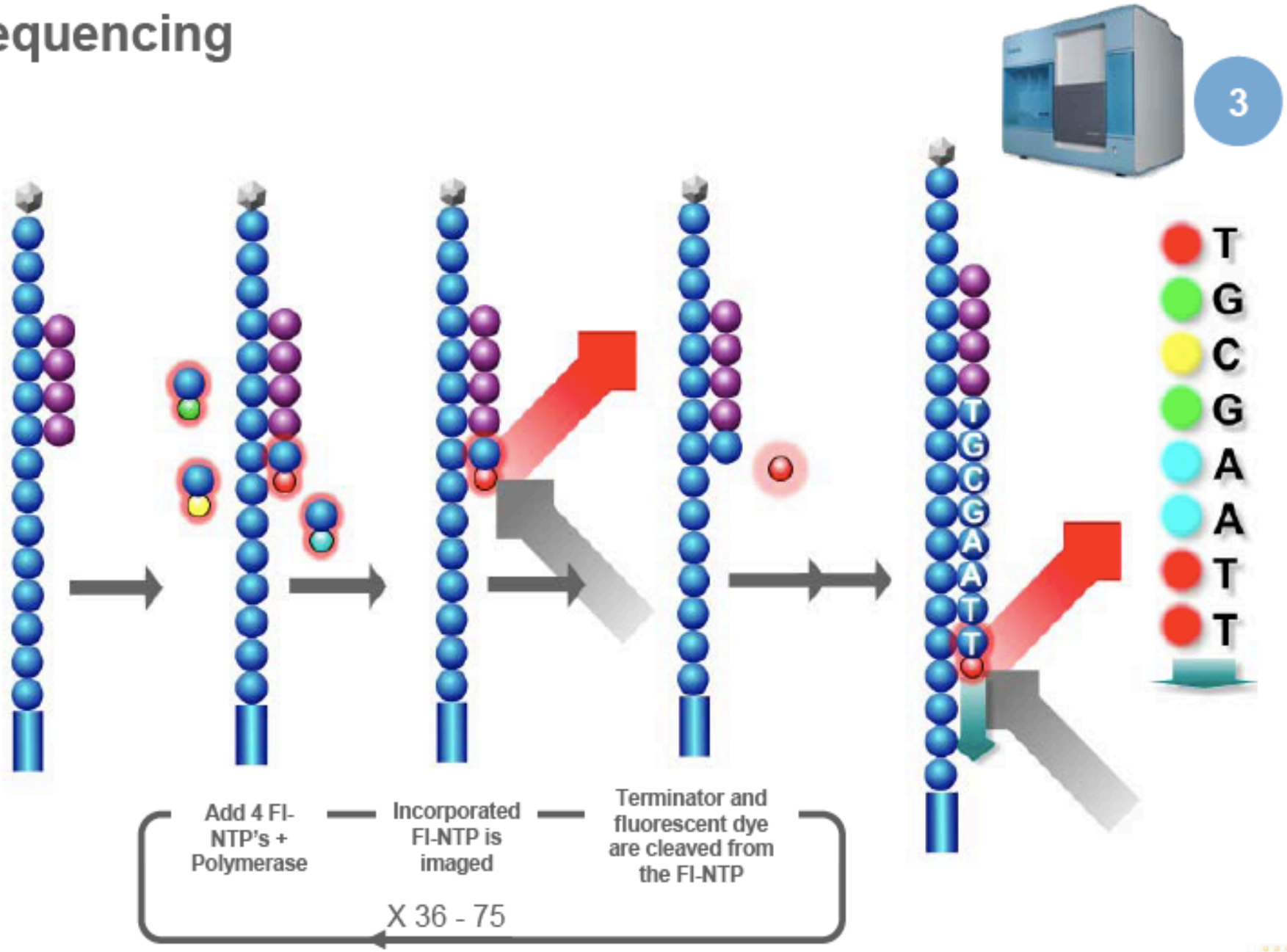
Genome Analyzer Sequencing reaction

- Sequencing primer is hybridized to adapter sequence



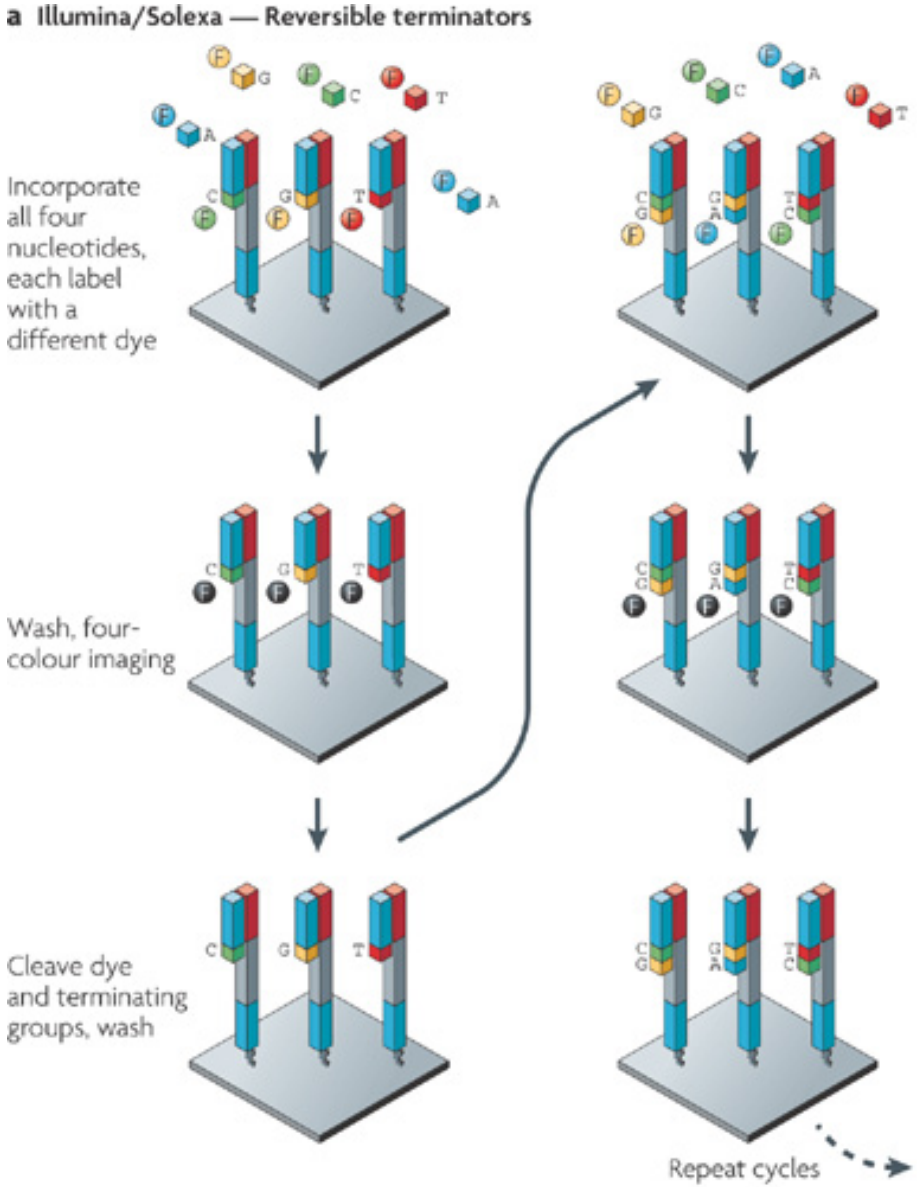
illumina

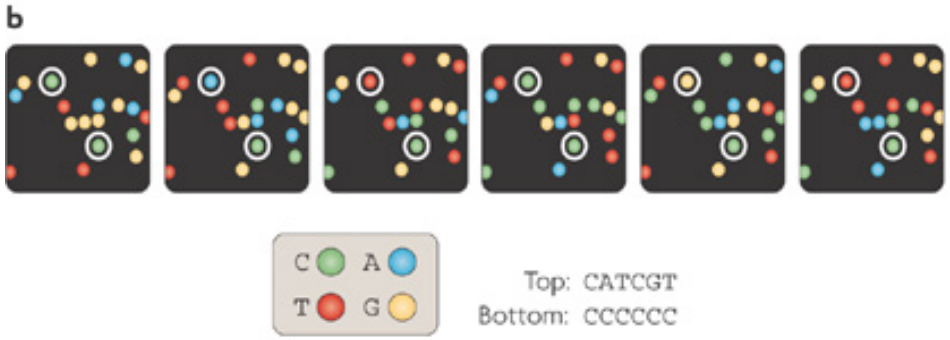
Sequencing



illumina

Sequencing over Multiple Chemistry Cycles





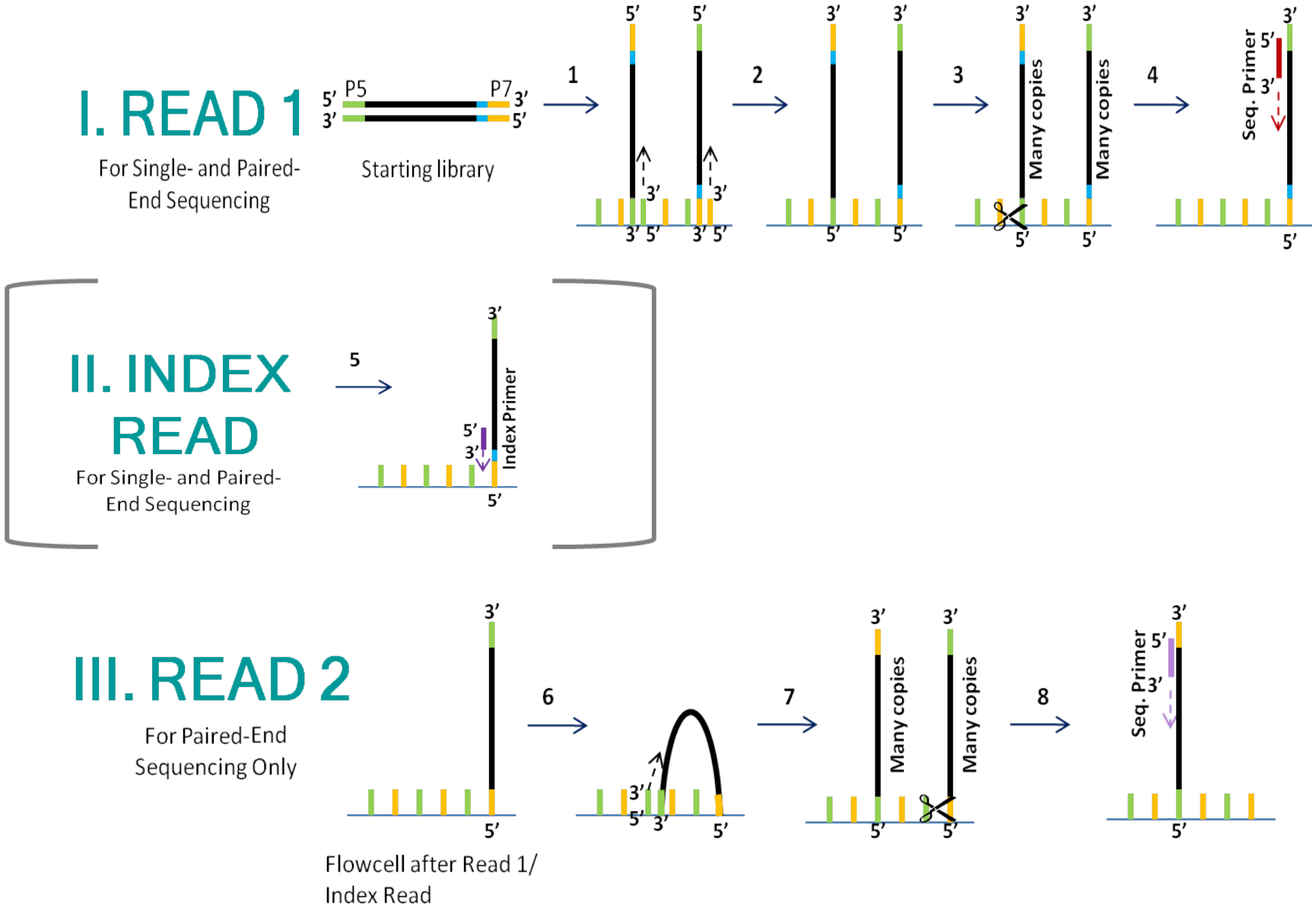
** Generate lots of image files

Metzker, Nat. Rev. Genet, 11:31-46

Multiplexing is easy – barcode and pool your libraries

- Allow sequencing of multiple samples per lane
- Add 4-6 base barcode/index for sample identification
- Single index vs Dual-index (pooling more samples per lane)
- Reads are “de-multiplexed” before assembly

Single-end vs. Paired-end sequencing



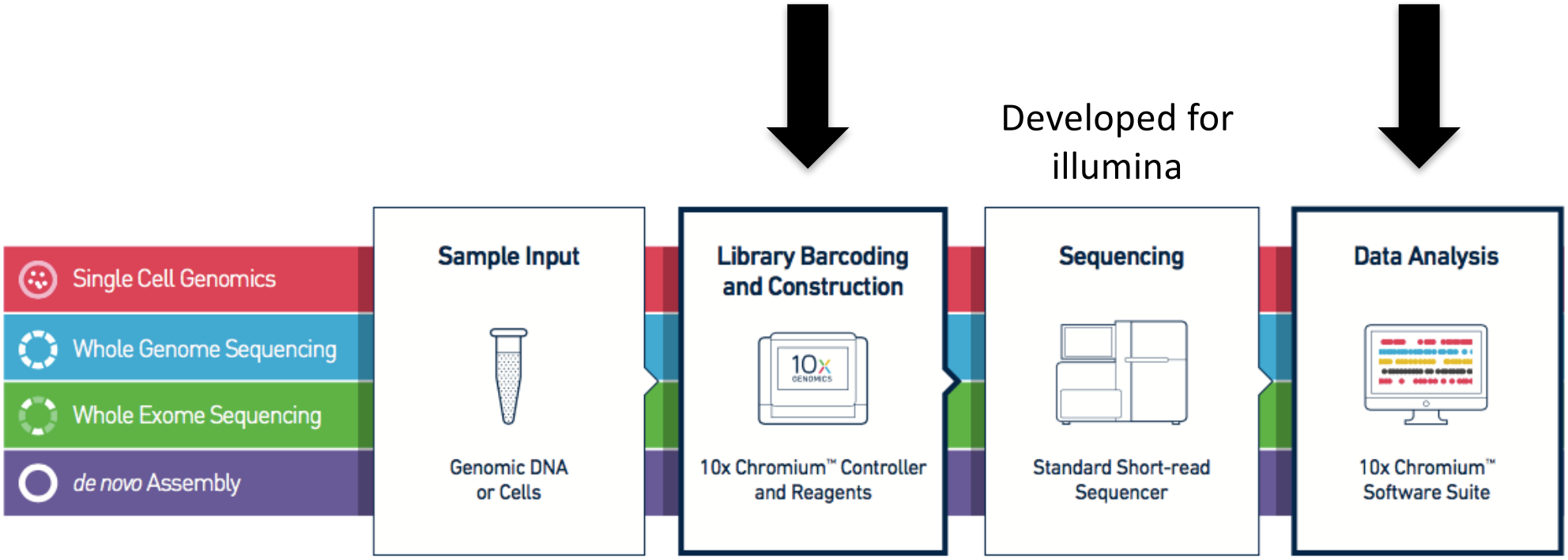
http://nextgen.mgh.harvard.edu/IlluminaChemistry.html

Illumina – Model Specifications

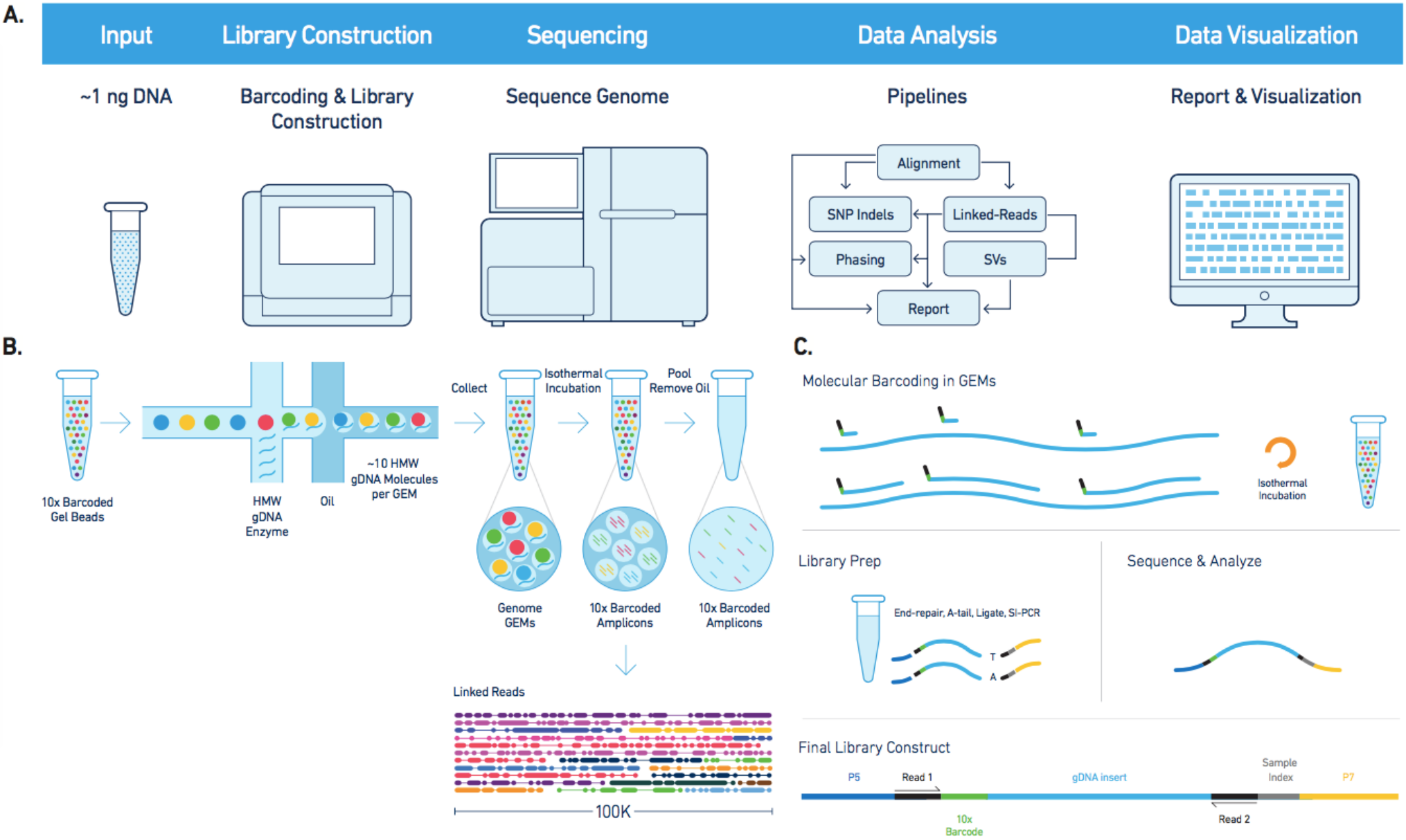
	MiSeq	NextSeq 500		HiSeq 2500		HiSeq 3000	HiSeq 4000	HiSeq X	NovaSeq
Run Mode	N/A	Mid-Output	High-Output	Rapid Run	High-Output	N/A	N/A		
Flow Cells Per Rnu	1	1	1	1 or 2	1 or 2	1	1 or 2		
Output Range	0.3-15 Gb	20-39 Gb	30-120 Gb	10-300 Gb	50-1000 Gb	125-750 Gb	125-1500 Gb	**1800 Gb**	**6000Gb**
Run Time	5-55 hrs	15-26 hrs	12-30 hrs	7-60 hrs	<1-6 days	<1-3.5 days	<1-3.5 days	**< 3 days**	**19-40 hrs**
Reads per Flow Cell	25million	130 million	400 million	300 million	2 billion	2.5 billion	2.5 billion	**3 billion**	**10 billion**
Maximum Read Length	2 x 300bp	2 x 150bp	2 x 150bp	2 x 250bp	2 x 125bp	2 x 150bp	2 x 150bp	**2X150bp**	**2X150bp**

10x GENOMICS®

Developed for illumina



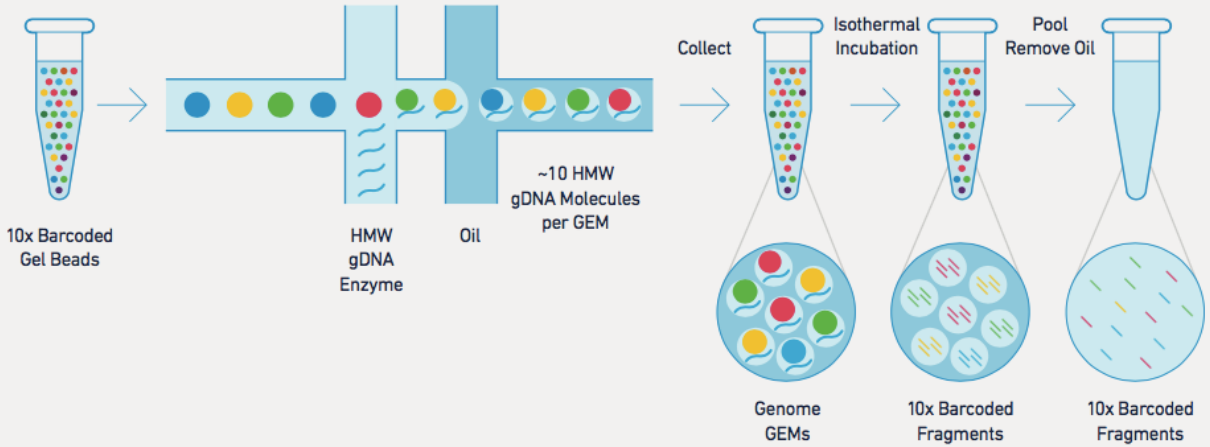
New ways to improve assembly even with Illumina short reads



10X genomics LIT00003 Rev B Chromium Genome Solution Application Note

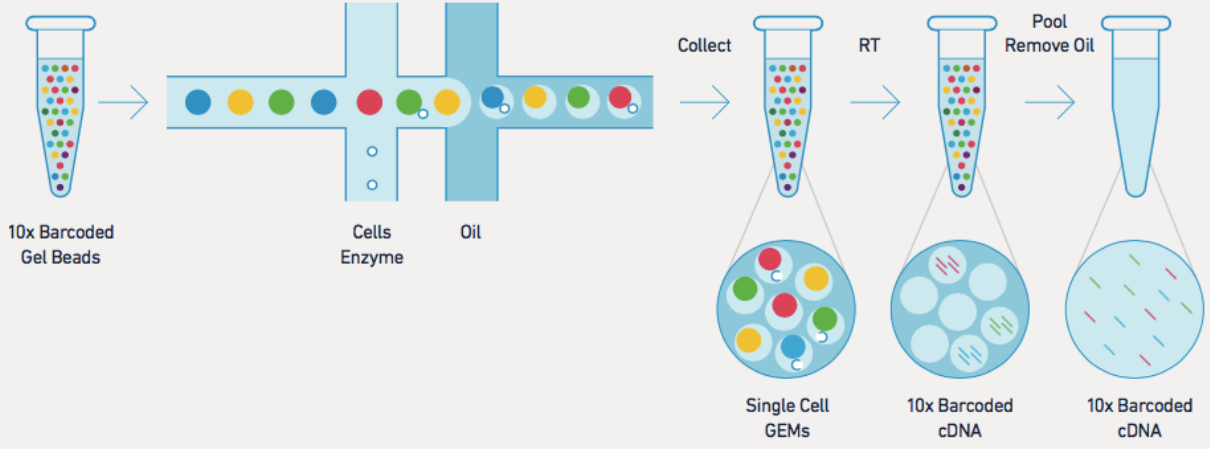
GemCode™ Technology for Partitioning High Molecular Weight DNA

Develop high-quality libraries from as little as 1 ng of genomic DNA, ready for standard short-read sequencing.

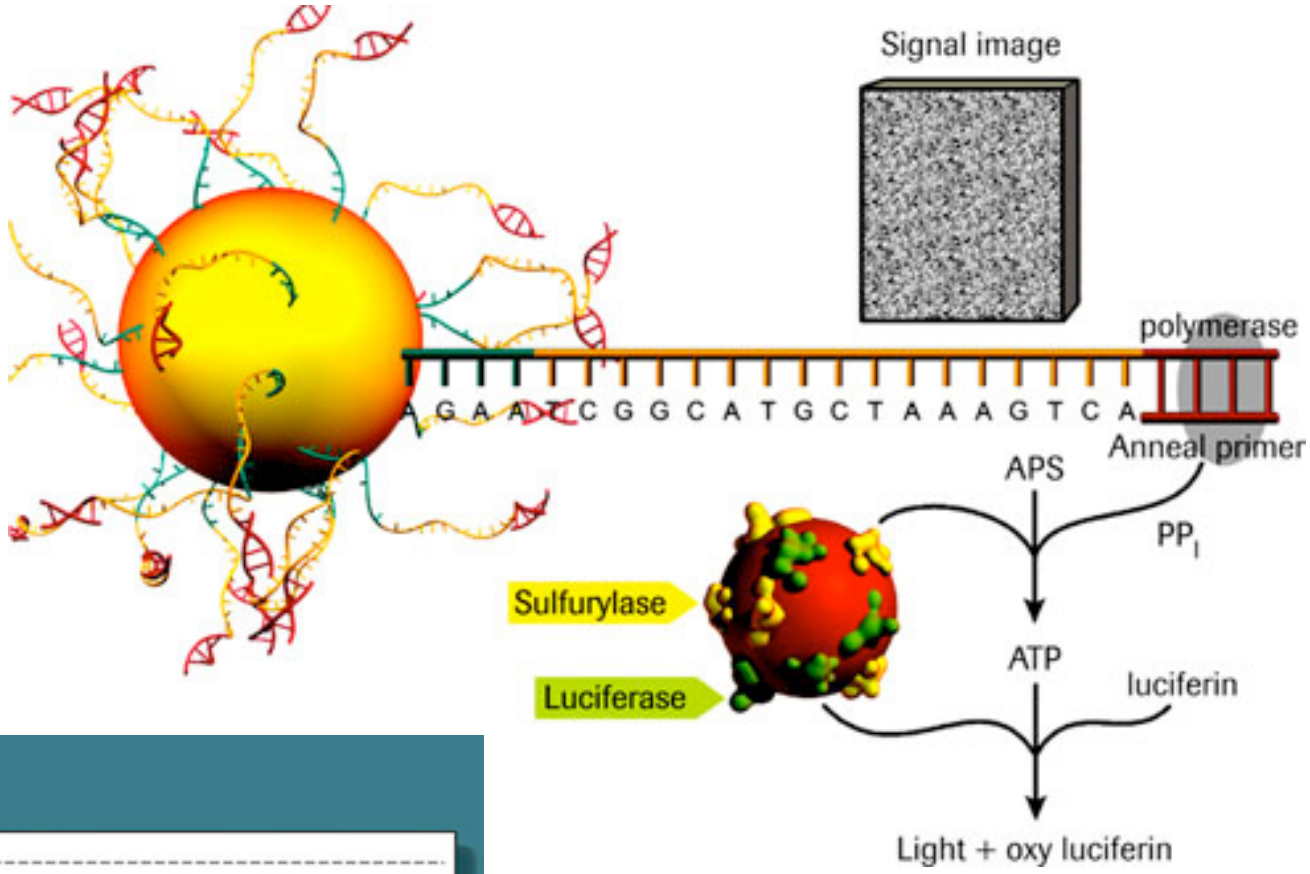


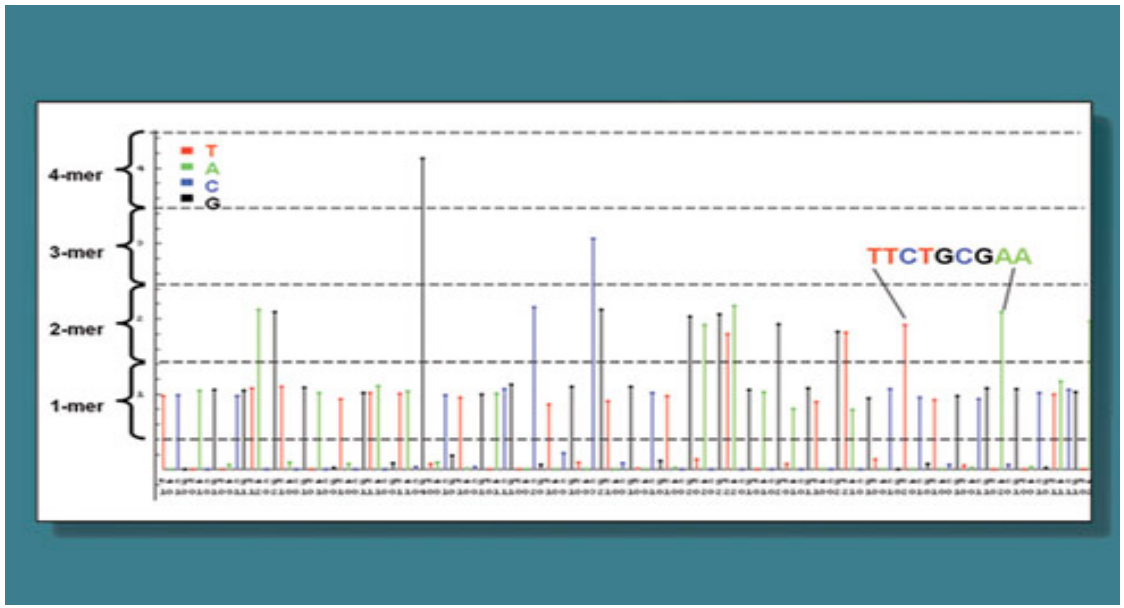
GemCode™ Technology for Single Cell Partitioning

Utilize an efficient droplet-based system to encapsulate up to 100-80,000+ cells in a single 10-minute run.



454 Pyrosequencing Chemistry and Base Calling





Release of pyrophosphate turns into light

Source: my454.com

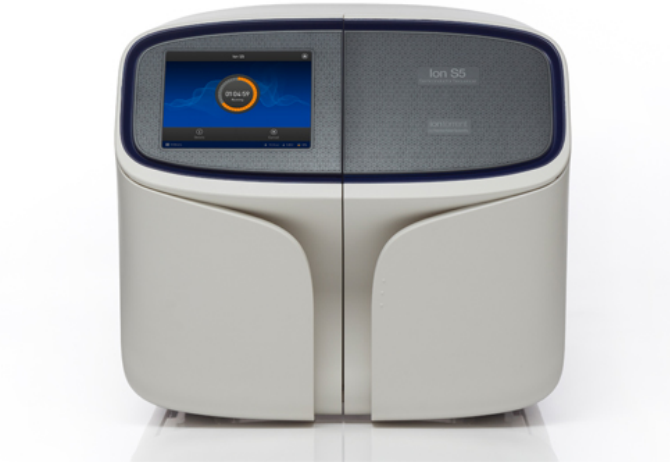
Ion Torrent: Ion Chip Non-Optical Sequencing

- Thermo Fisher
- Natural nucleotides
- Leverage semiconductor manufacturing techniques
- Non-optical but still requires amplification

Ion PGM System

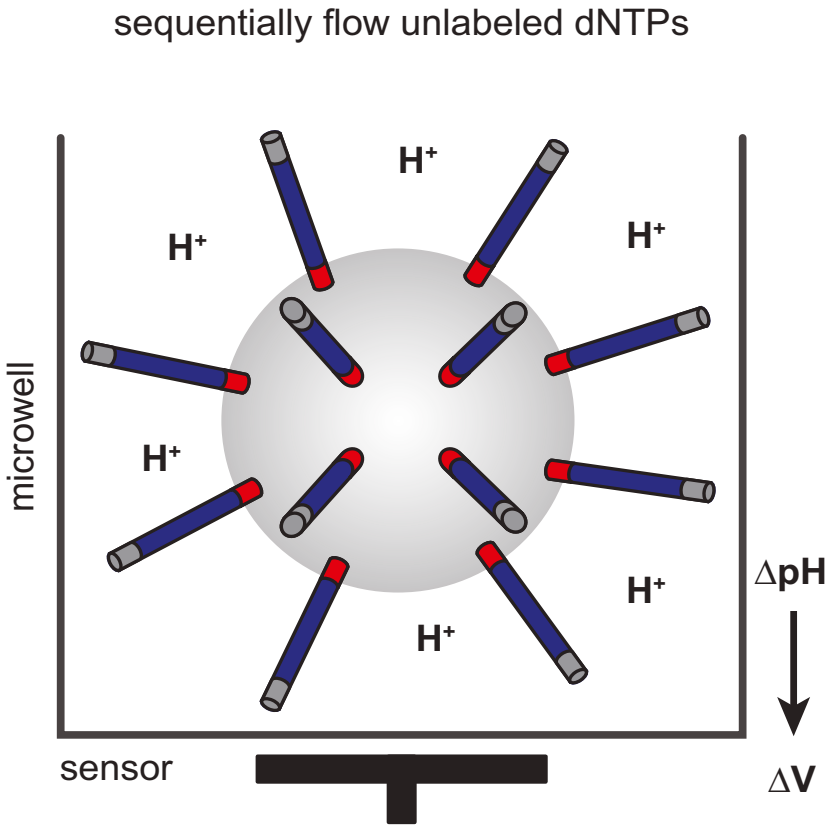
Ion Chef (library prep)

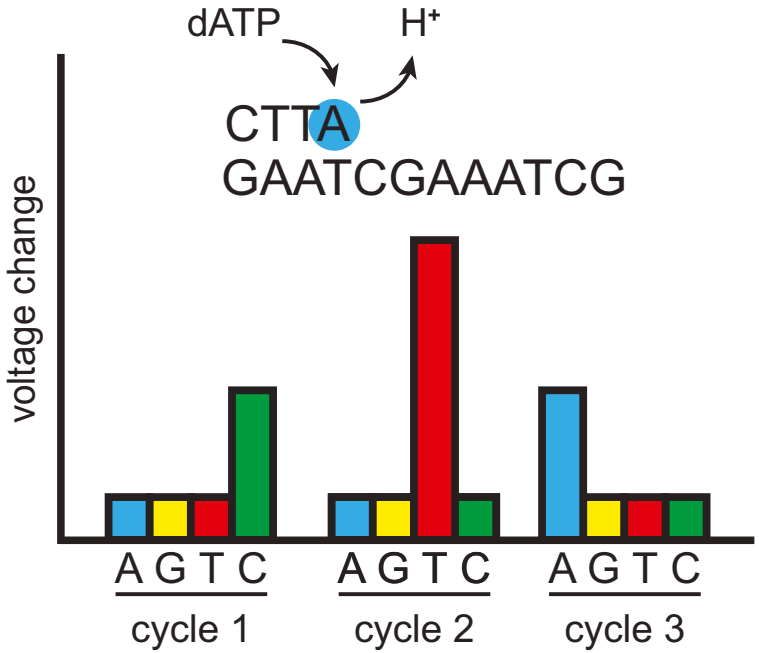
Ion Proton



Ion S5

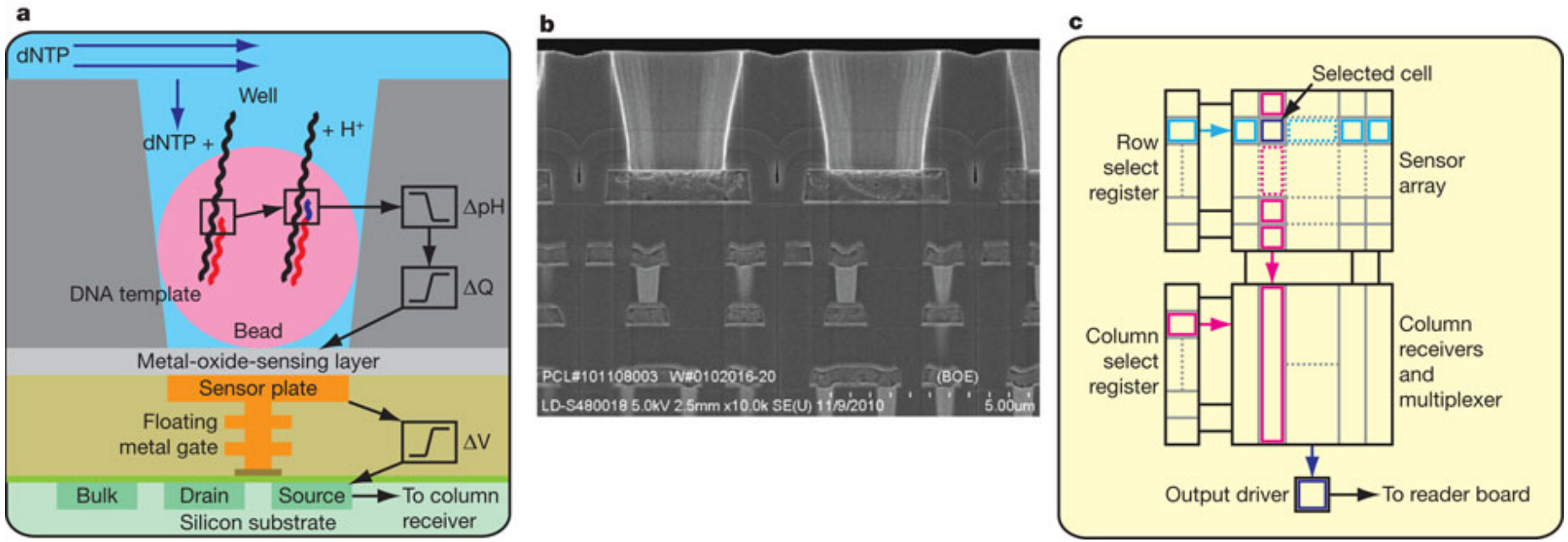
Sequencing by detecting pH change (release of proton during extension)





Reuter et al. (Mol Cell 2015)

Sensor, Well and Chip Architecture



JM Rothberg *et al. Nature* **475**, 348-352 (2011) doi:10.1038/nature10242

3rd Generation Sequencing

- Longer read lengths, easier assembly
- Single molecule sequencing – amplification not necessary, can handle regions with high levels of repeat sequences

Pacific Biosciences
(PacBio RS II)

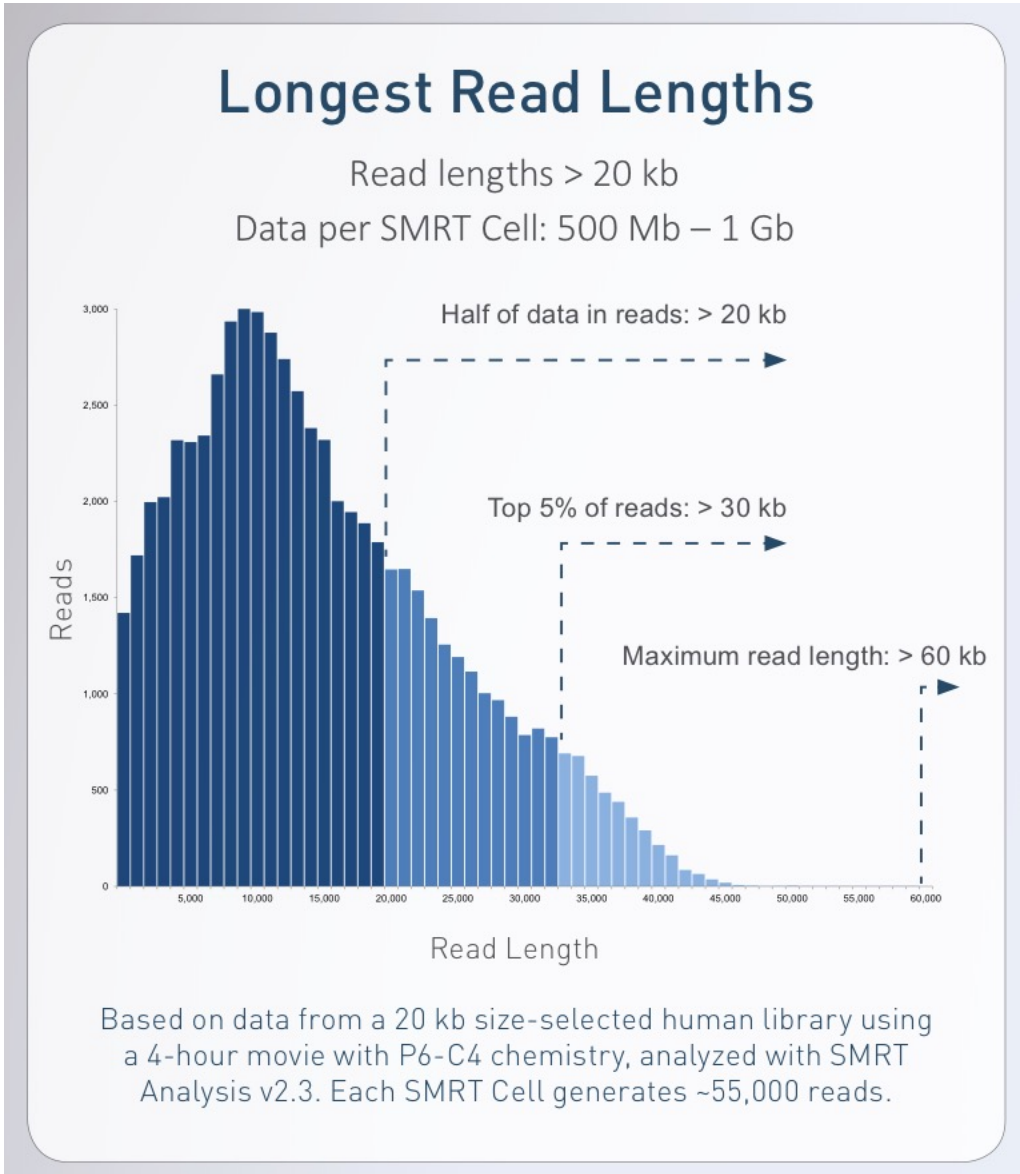
Pacific Biosciences
(Sequel)

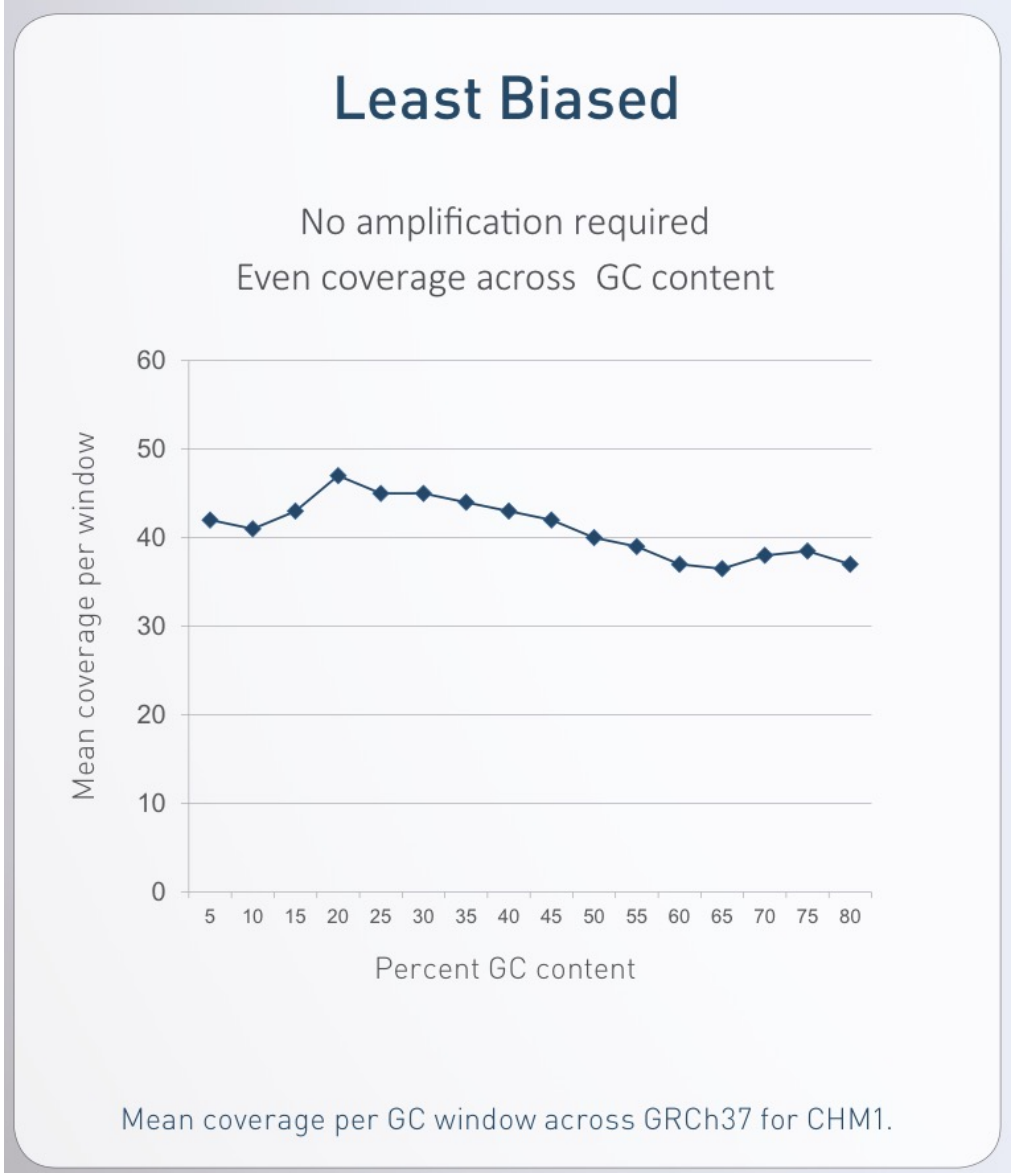
Oxford Nanopore
(MinION) - $1000

Oxford Nanopore
(PromethION)

Oxford Nanopore
(SmidgION) – use
with smartphone

PacBIO -SMRT™ Sequencing



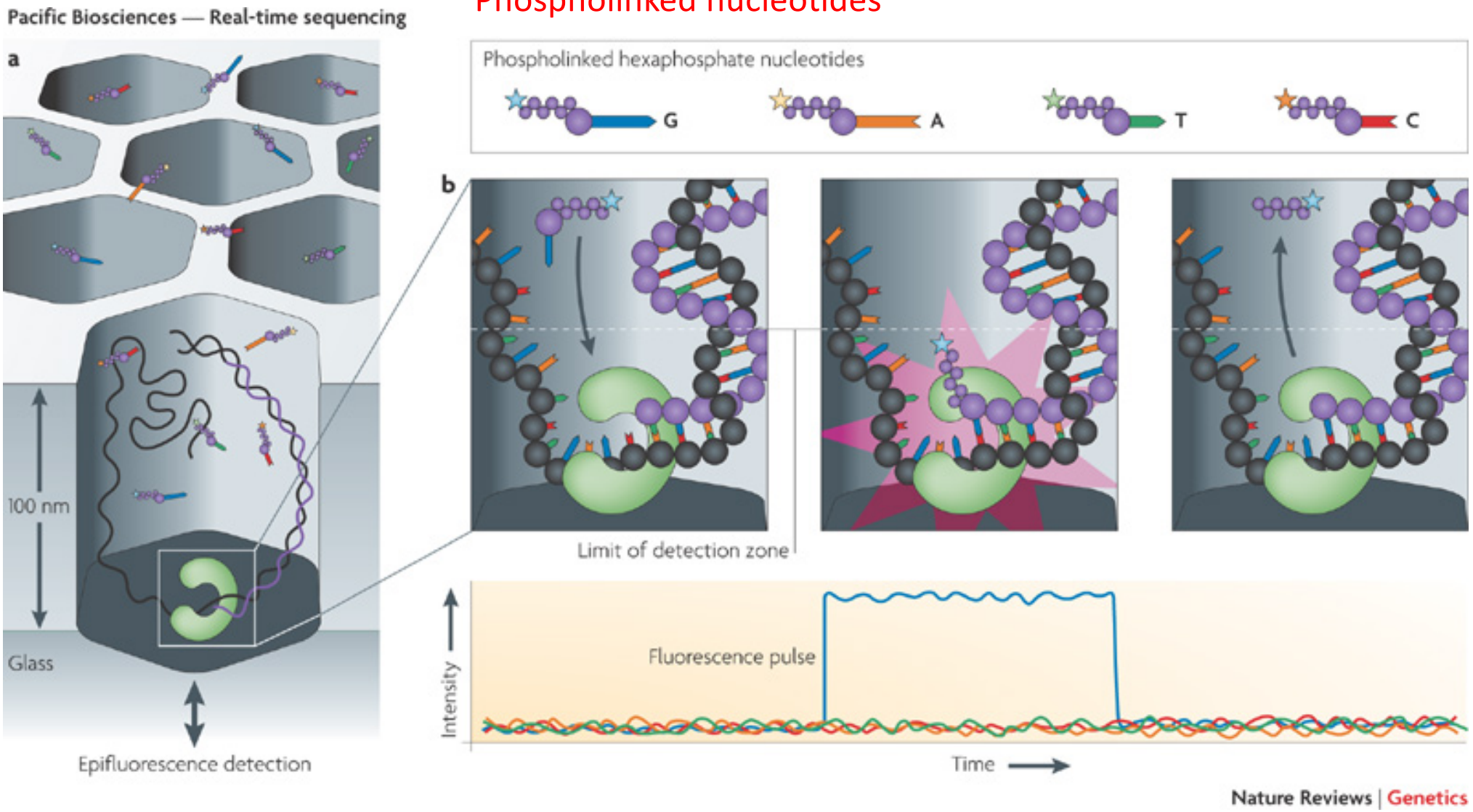


Regions with lots of repeats = easier assembly

From: PacBio_RSII_Brochure

Key Innovations for PacBIO -SMRT™ Sequencing

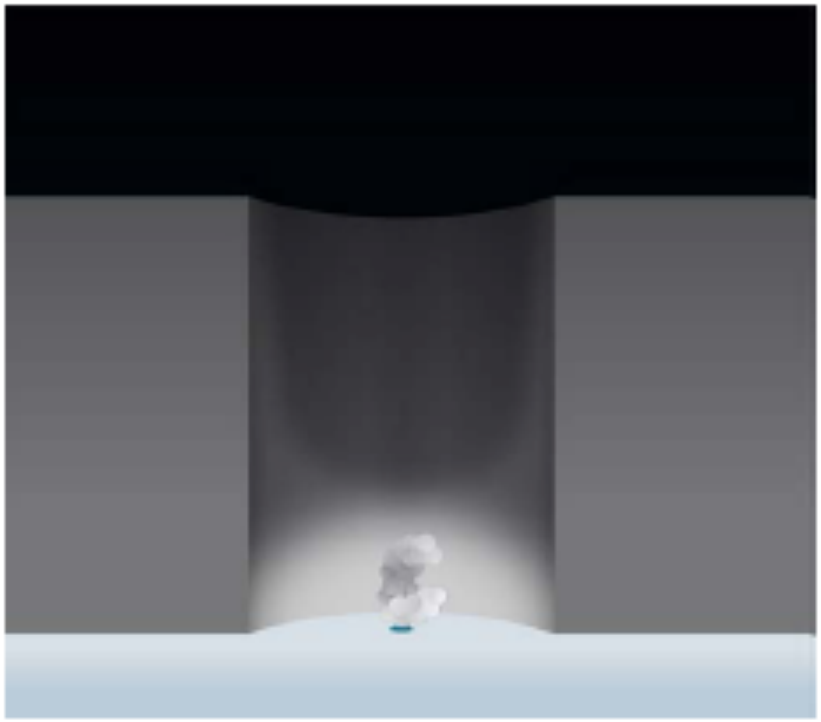
Phospholinked nucleotides



ZMWs = Zero-mode Waveguides

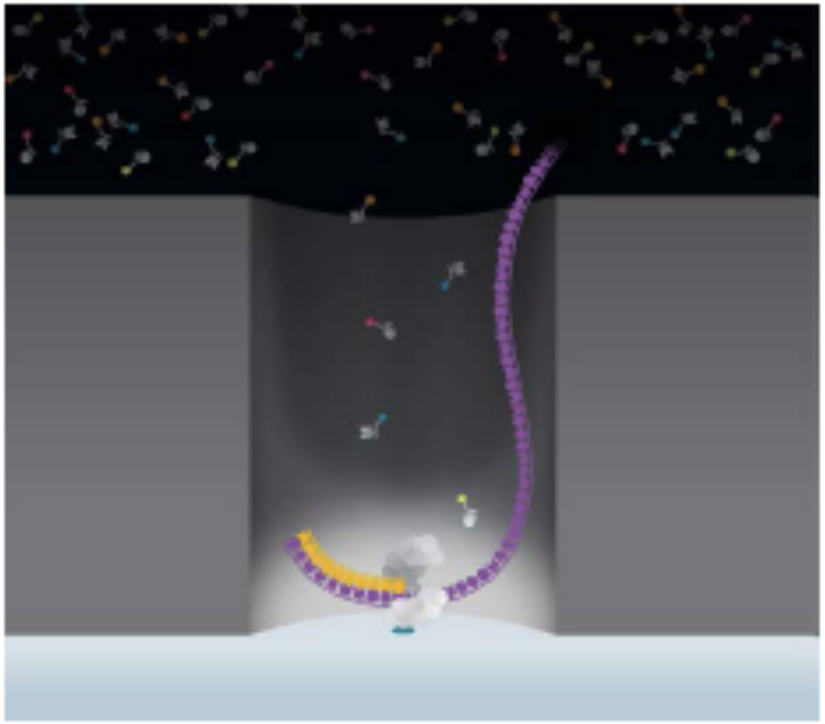
Metzker, Nat. Rev. Genet, 11:31-46

DNA Polymerase as a Sequencing Engine



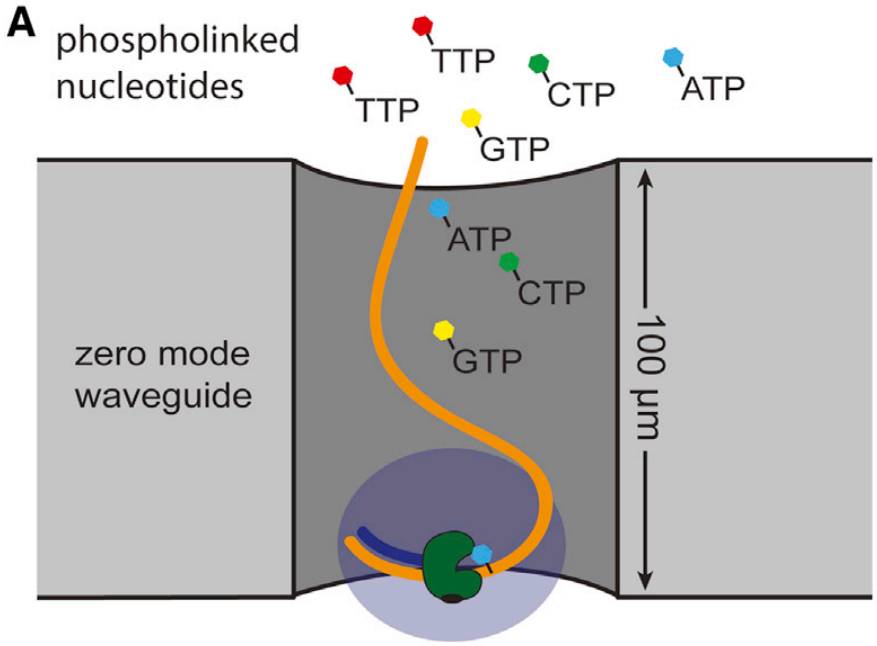
ZMW with DNA polymerase

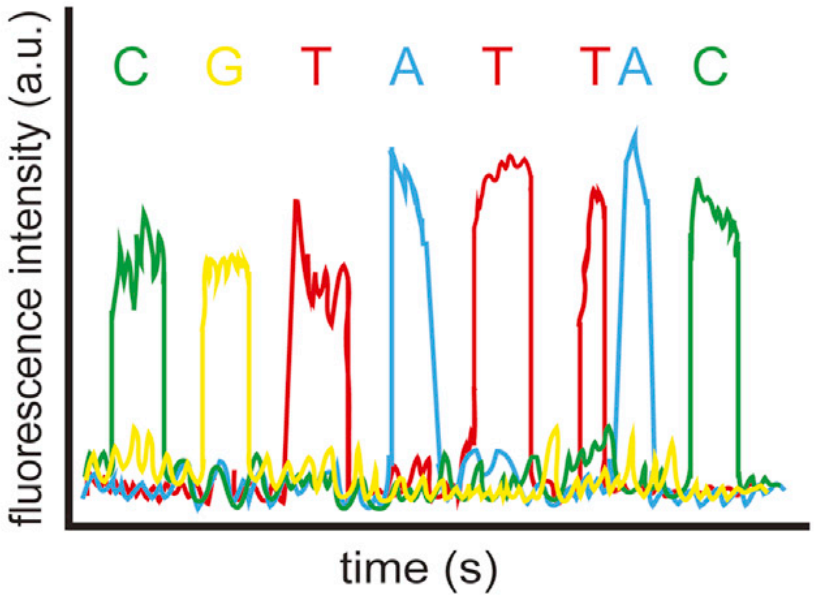
DNA polymerase is immobilized



ZMW with DNA polymerase and phospholinked nucleotides

DNA Polymerase Processive Synthesis with Phospholinked Nucleotides



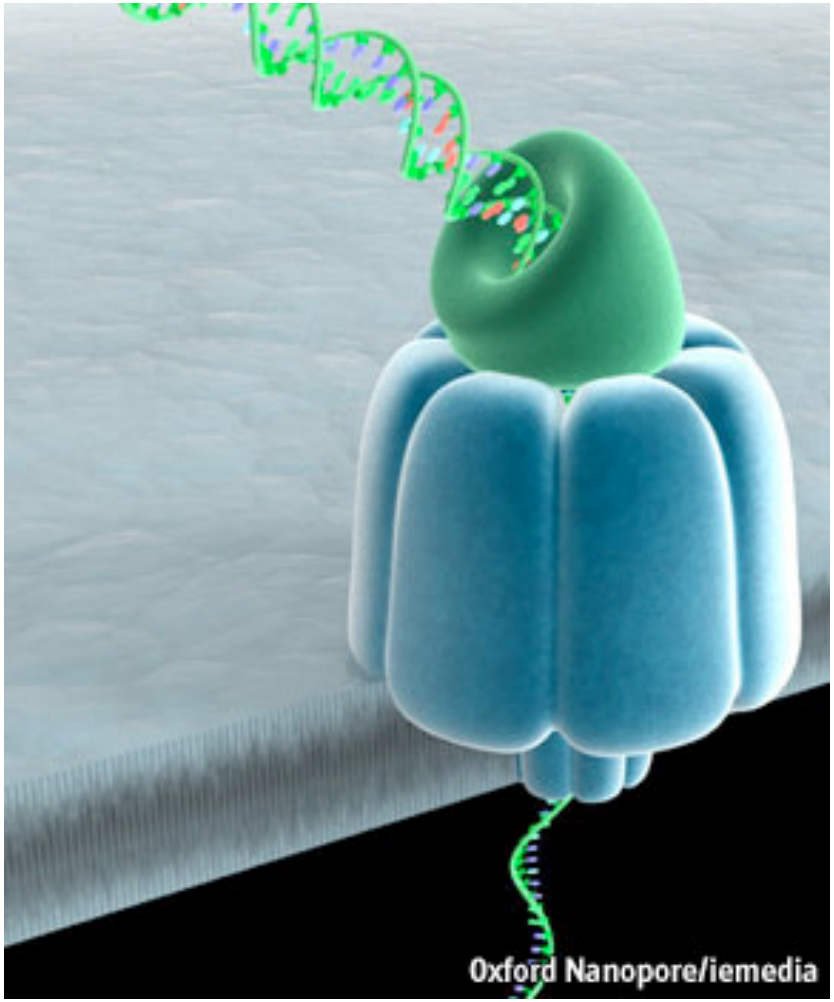


Base-specific fluorescence and DNA sequence can be detected in real-time

Reuter et al. (Mol Cell 2015)

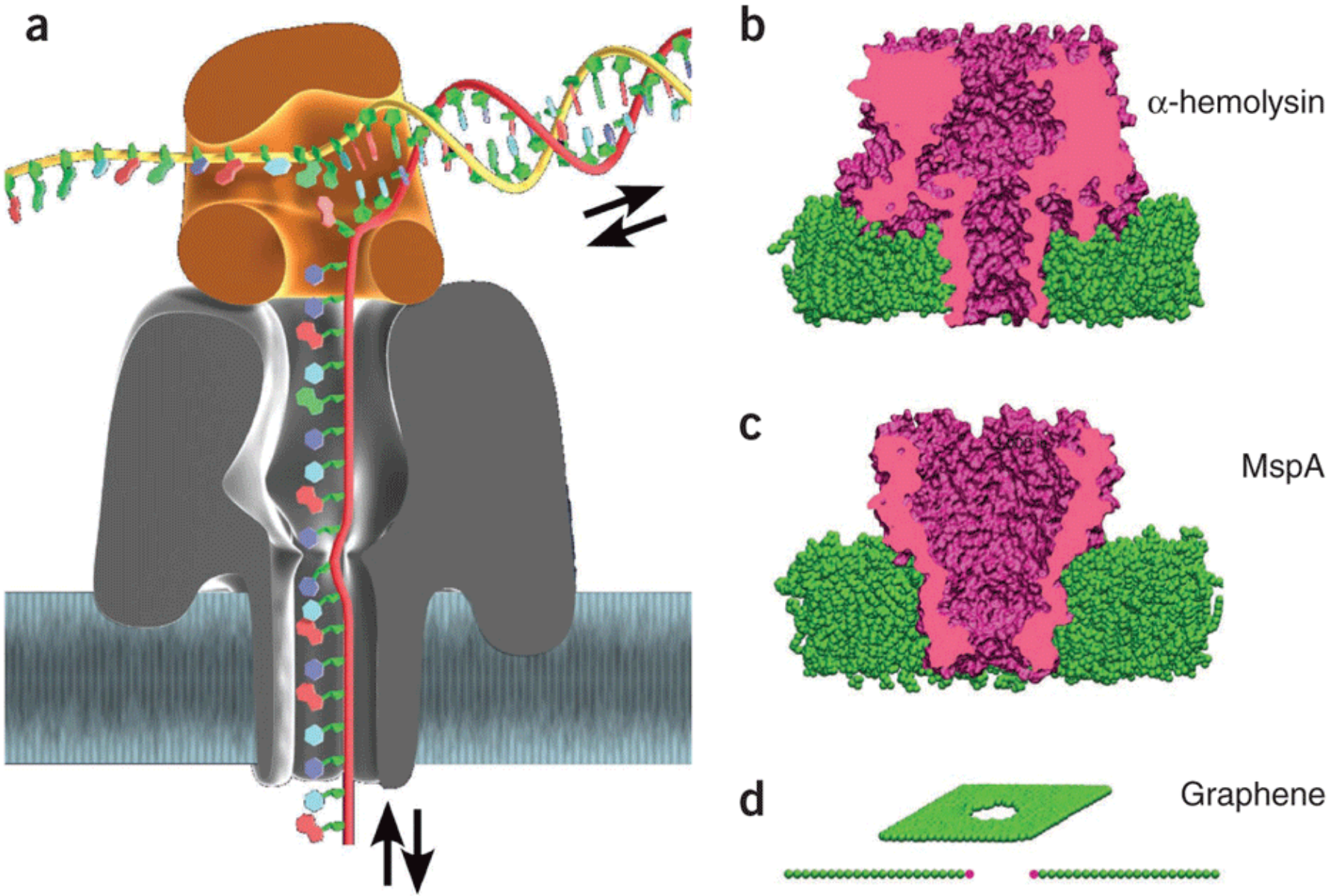
Nanopore Sequencing

- Oxford Nanopore
- First commercial product = MinION



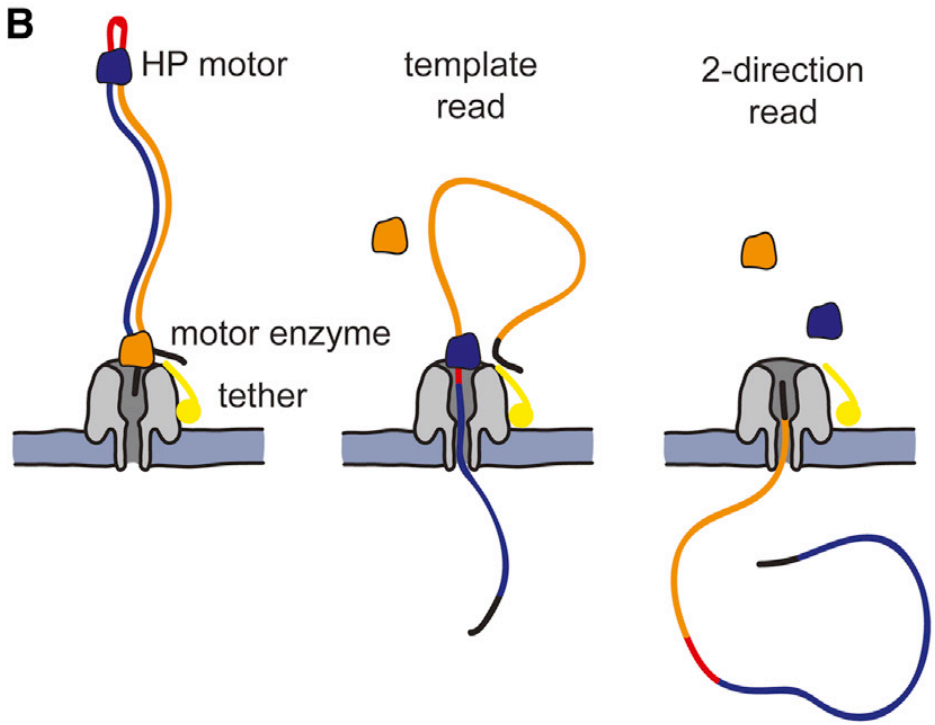
http://www.medgadget.com/

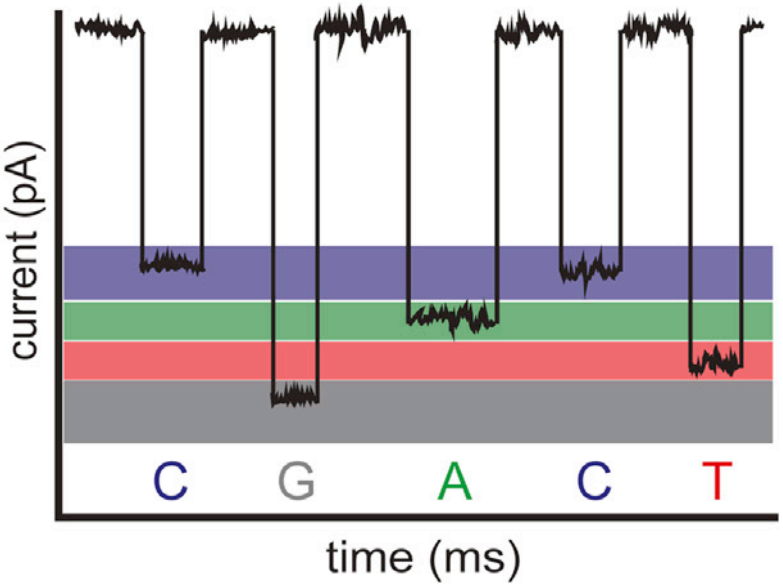
Different Nanopore Base-Readers



Schneider & Dekker (Nat Methods 2012)

Nanopore Sequencing



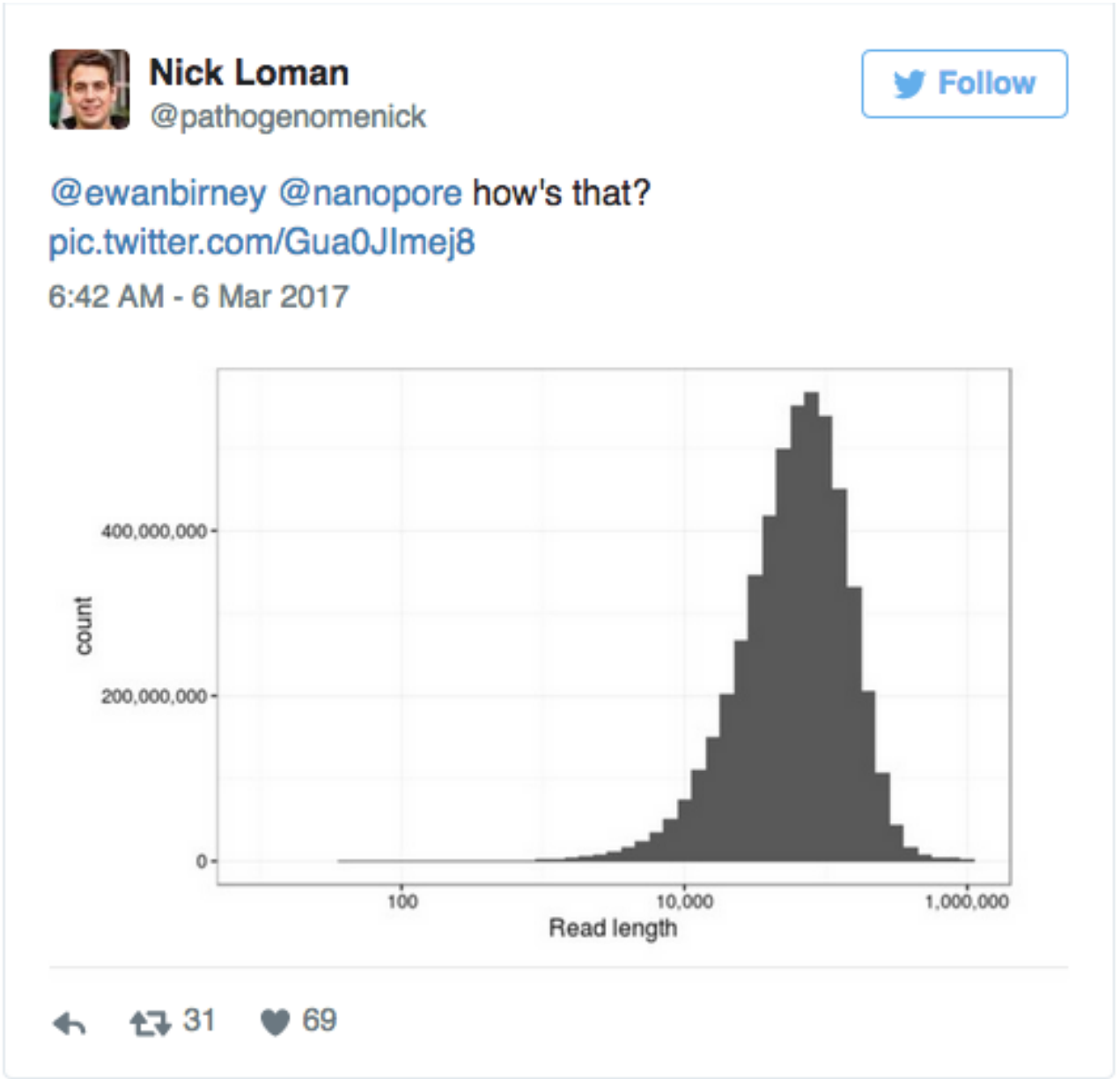


Library prep (~1.5 hours):

- DNA fragmentation
- Ligation of 2 adaptors for attachment to motor enzyme and HP motor
- HP adaptor (red) enables the sequencing of both DNA strands (blue and yellow)

Reuter et al. (Mol Cell 2015)

How long are we talking about?



http://omicsomics.blogspot.com/2017/03/minion-leviathan-reads-update.html

Start using MinION for $1000

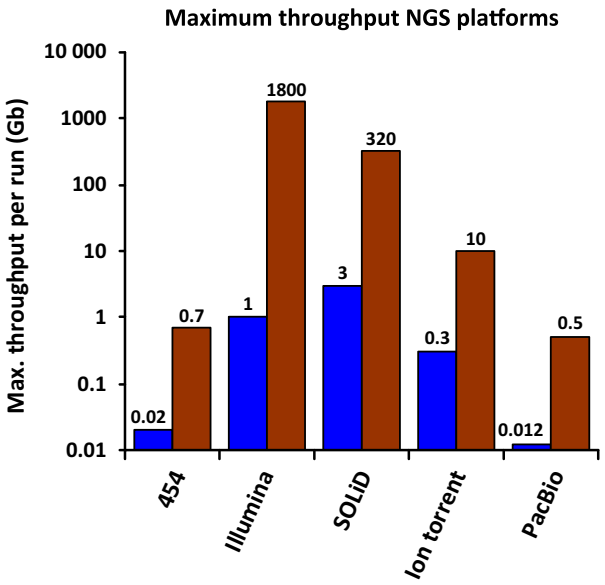
MinION

- Pocket-sized, portable device for biological analysis
- Up to 512 nanopore channels
- Simple 10-minute sample prep available
- Real-time analysis for rapid, efficient workflows
- Adaptable to direct DNA or RNA sequencing

Start using MinION

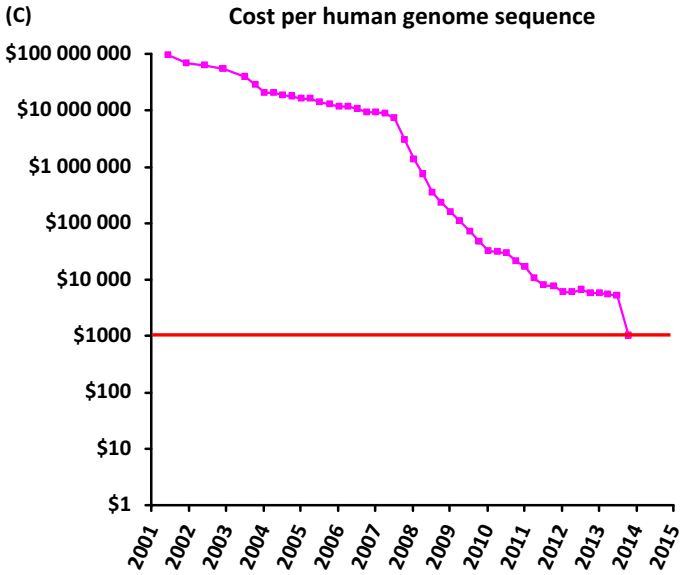
https://nanoporetech.com/products

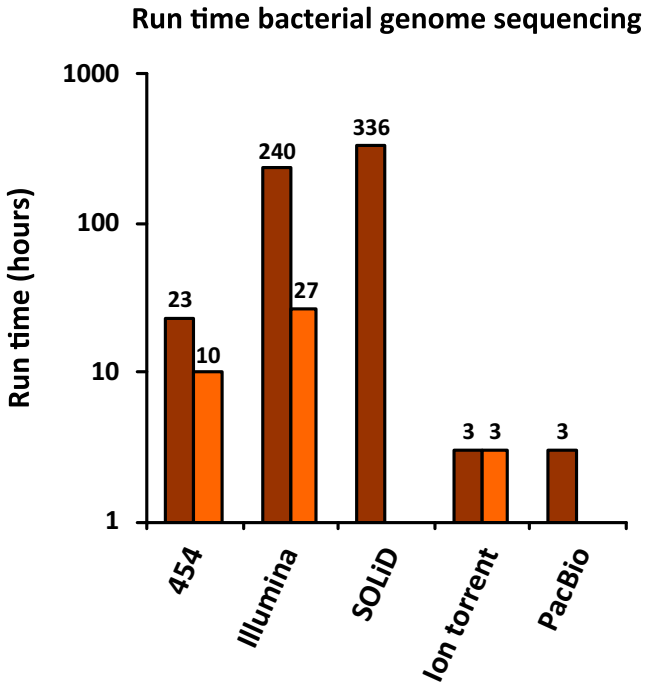
What to consider when using sequencing platforms?



Throughput
Read length

Cost





Run time

Others??

Van Dijk et al (Trends Genet 2014)

Types of Data for Molecular Phylogenetics

A few mitochondrial or nuclear genes

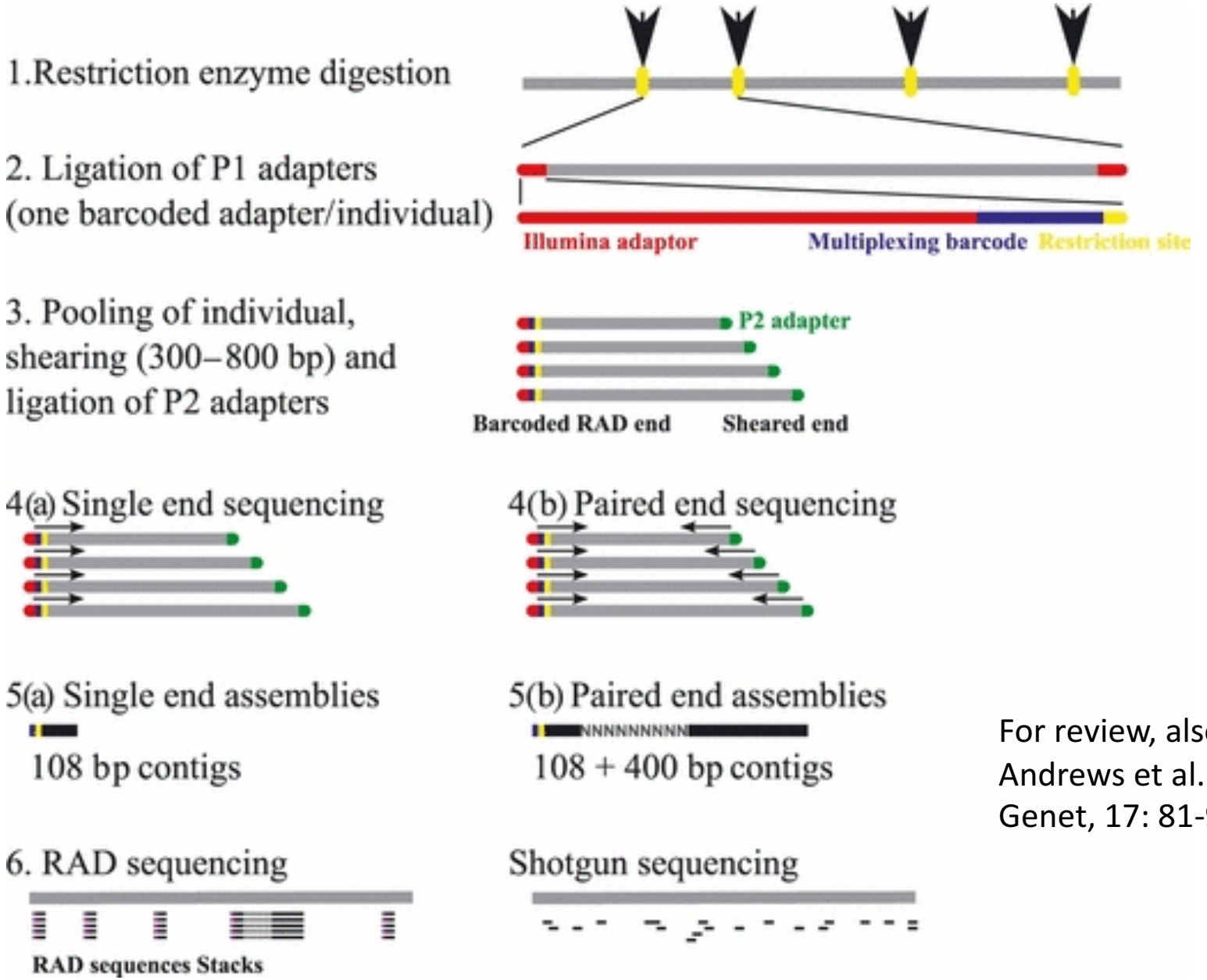
Maybe somewhere in the middle?
Reduced Representation sequencing

Whole Genomes

Reduced Representation Sequencing for Molecular Phylogenetics

- Transcriptome Sequencing (RNA)
- Restriction Site Associated DNA (RAD) Sequencing
- Ultra Conserved Element (UCE) Sequencing
- Anchored Hybrid Enrichment (AHE) Sequencing

Restriction Site Associated DNA (RAD) Sequencing



For review, also see:
Andrews et al. (Nat Rev Genet, 17: 81-92, 2016)

Rowe et al. (Mol Ecol, 2011)

RAD-seq for phylogenomics

16 species of Lake Victoria cichlid fish
(HiSeq 2000, 1 ug DNA)

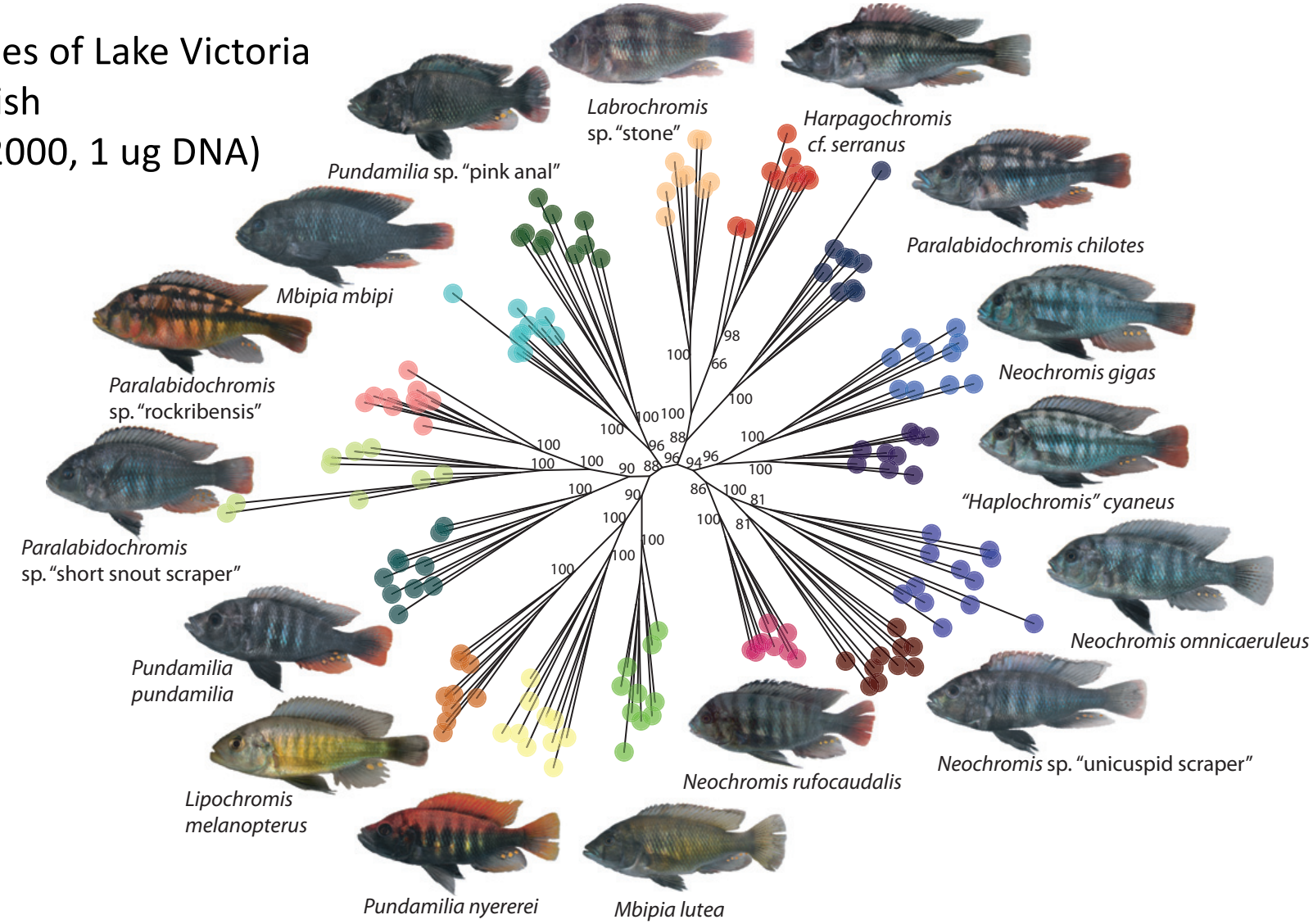


Fig. 3 The phylogeny produced based on the largest supermatrix analysed, which contains a minimum of 15 individuals out of the total 156 with sequence data per locus ('min individuals 15'; Table 1g).

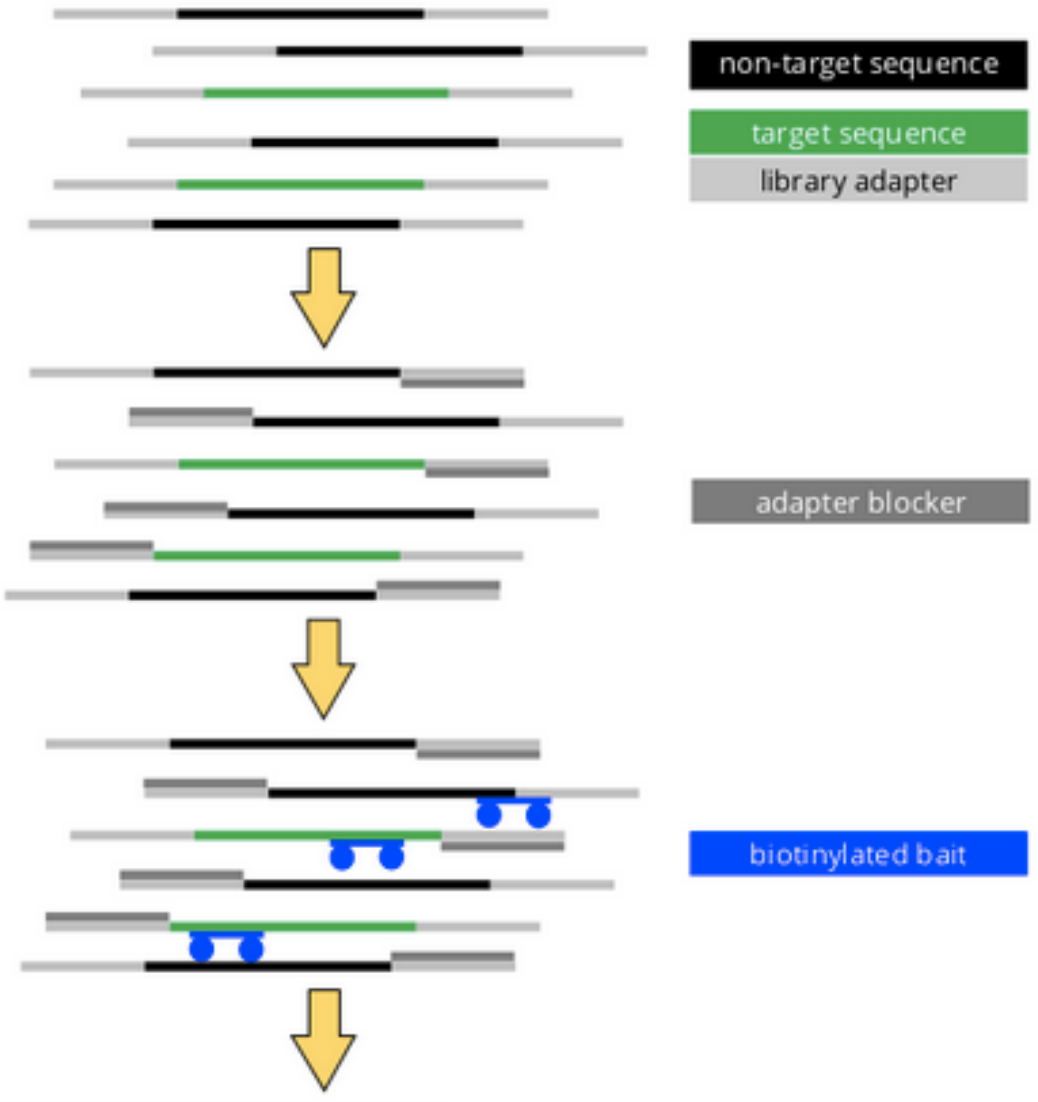
Wagner et al. (Mol Ecol, 22:787-798, 2013)

Probe capture and library enrichment (UCE and AHE)

1) DNA sequencing library is heat-denatured in the presence of adapter-specific blocking oligonucleotides

2) Library and blockers are dropped to the hybridization temperature, allowing blockers to hybridize to the library adapters

3) Biotinylated RNA baits are introduced and allowed to hybridize to targets for several hours



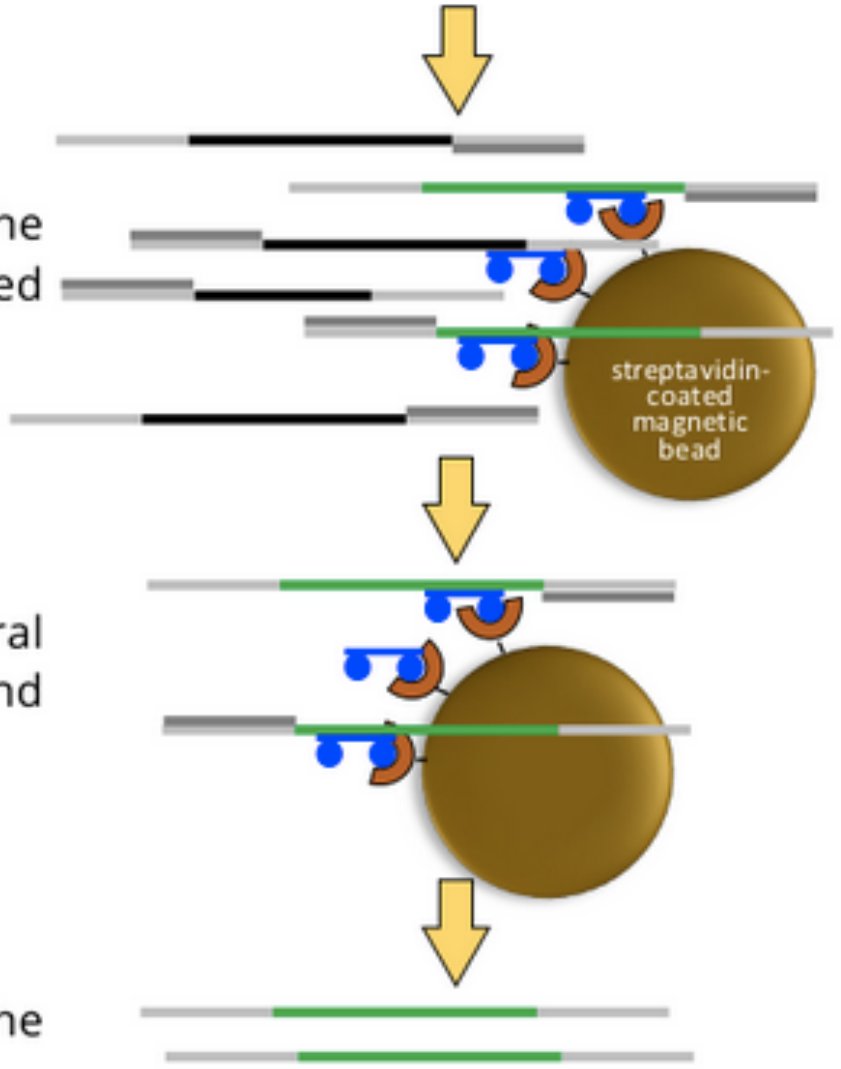
http://www.mycroarray.com/mybaits/mybaits-technology.html

Probe capture and library enrichment

4) Bait-target hybrids are pulled out of the solution with streptavidin-coated magnetic beads

5) Beads are stringently washed several times to remove non-hybridized and nonspecifically-hybridized molecules

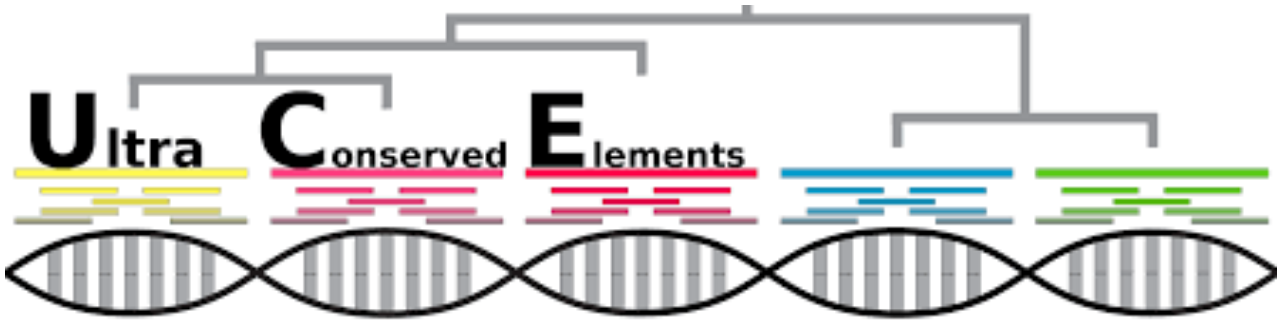
6) Captured DNA library is released from the beads and amplified



**** Design Custom Probe Sets**

http://www.mycroarray.com/mybaits/mybaits-technology.html

Ultra Conserved Element (UCE) for Phylogenetics



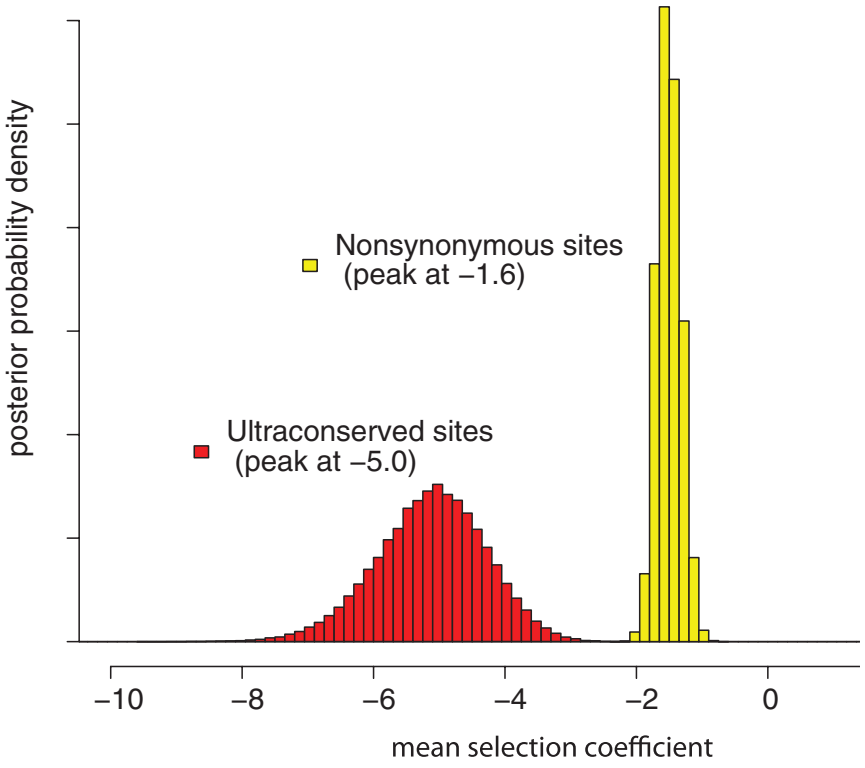
- 481 regions perfectly conserved over 200bp or more between human, mouse, and rat
- 20-fold fewer SNPs than the human average
- Most are in non-coding regions, but some in exonic regions
- May be enhancer sequences, control gene expression
- The exonic UCEs tend to overlap with alternative spliced exons

Bejerano et al. (Science 2004)

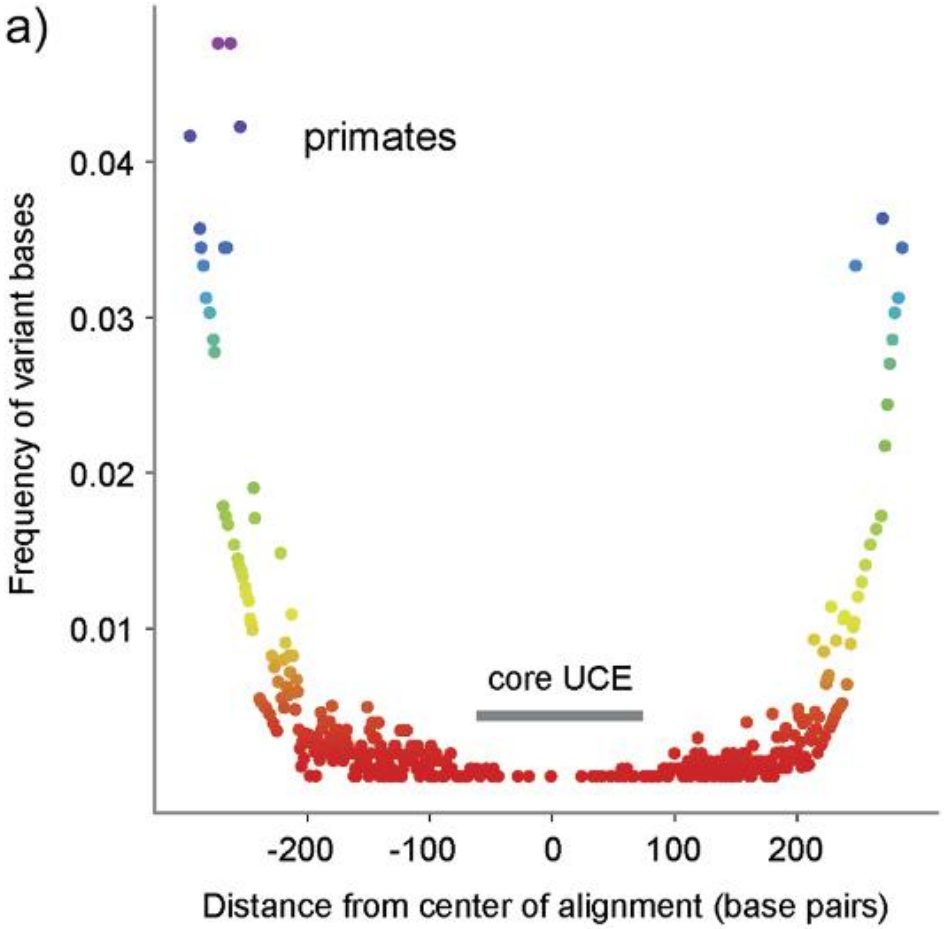
Human Genome Ultraconserved Elements Are Ultraselected

Sol Katzman,[1]* Andrew D. Kern,[2]* Gill Bejerano,[2]† Ginger Fewell,[3] Lucinda Fulton,[3]
Richard K. Wilson,[3] Sofie R. Salama,[2,4] David Haussler[1,2,4]‡

Science, 317:915 (2007)

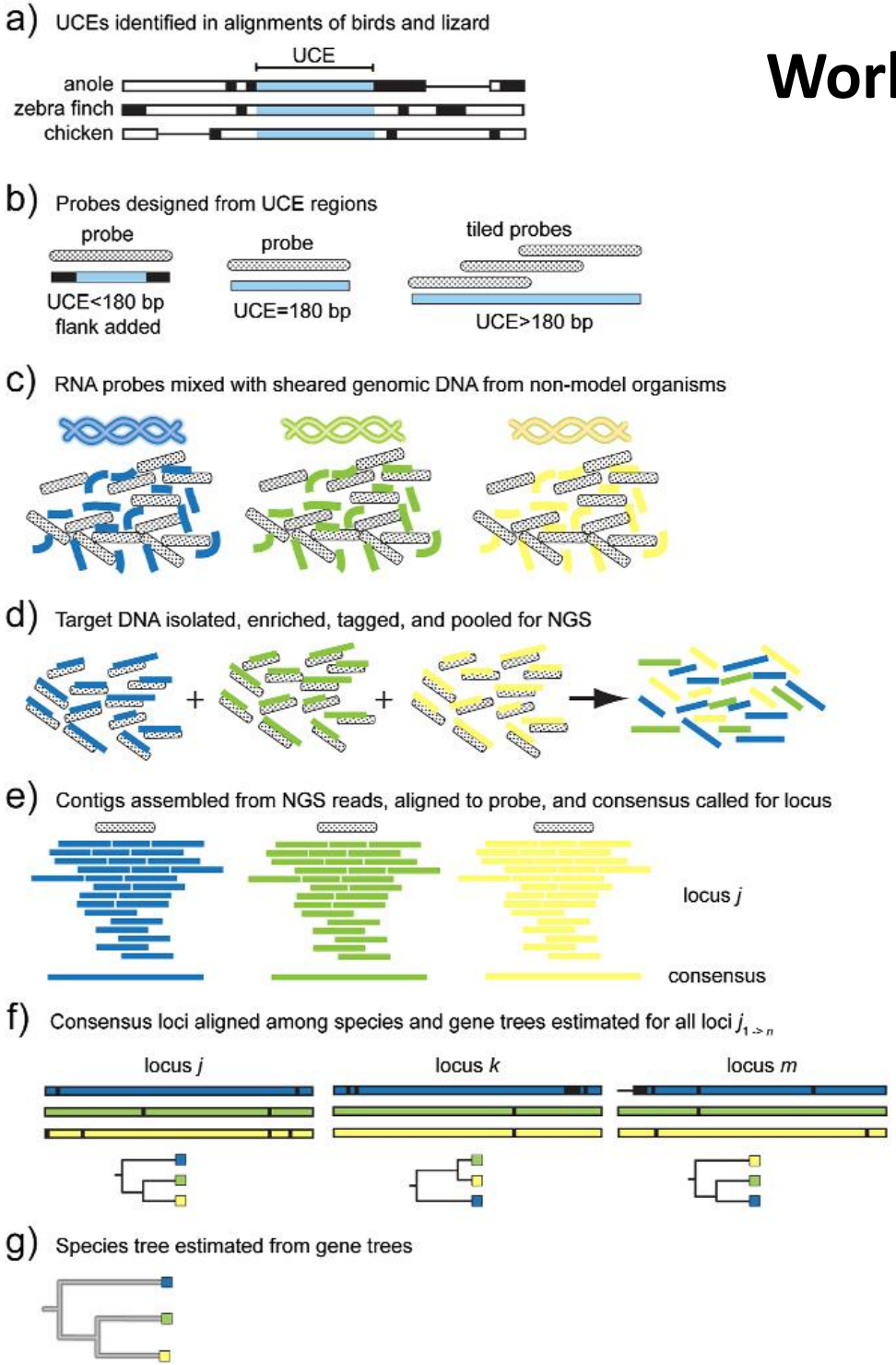


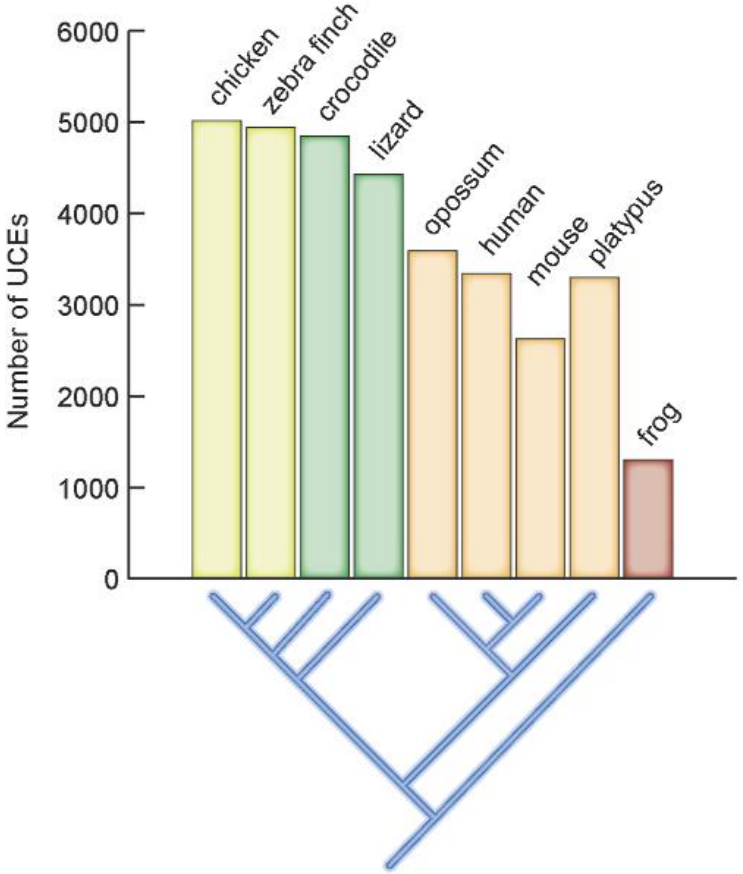
Data for the human genome



Data for primates
(Faircloth et al. Syst Biol 61(5): 717-726, 2012)

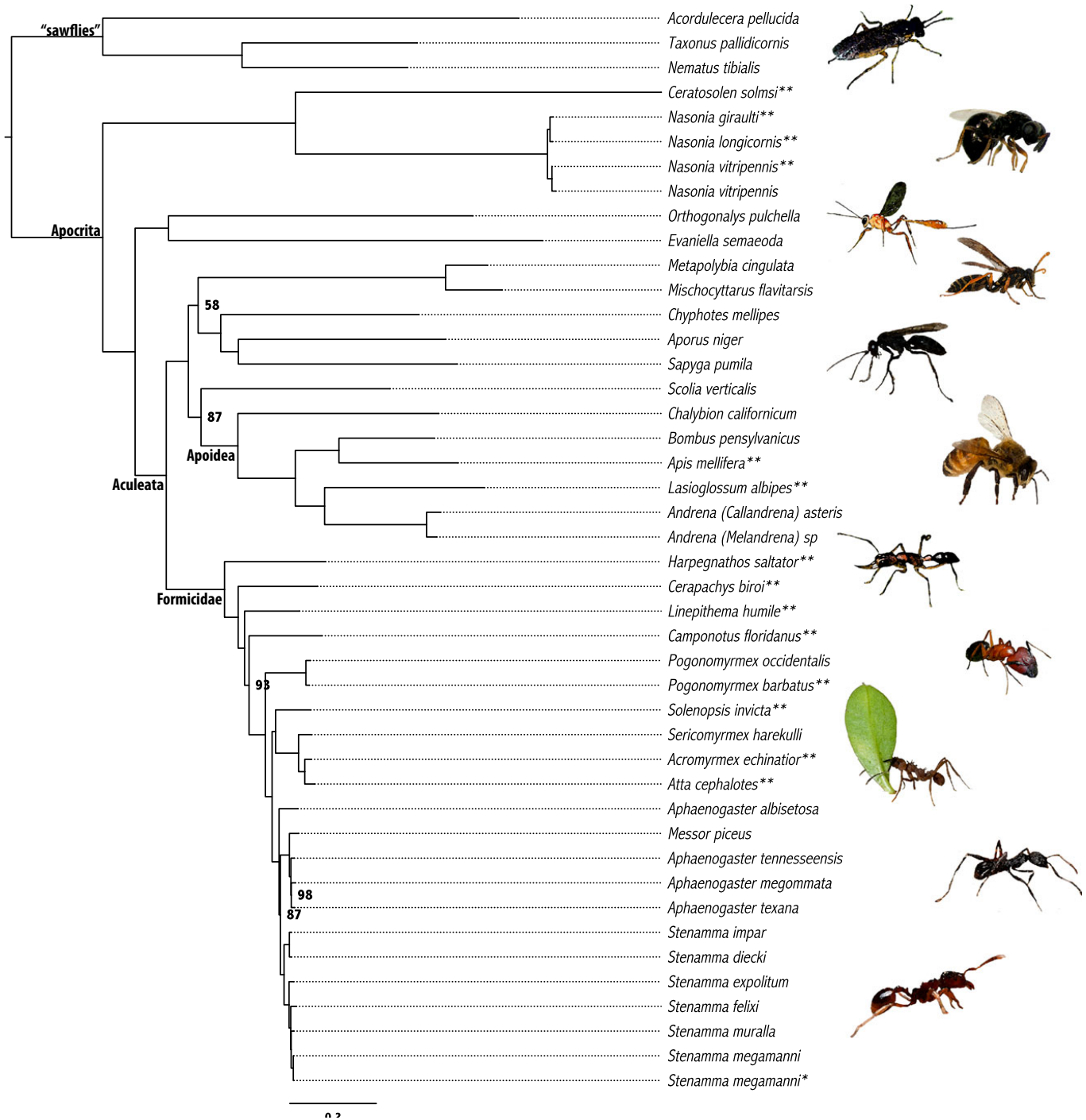
Workflow for UCE phylogenomics





Probe sets to resolve different levels of phylogenetic relationships need to be adjusted

Faircloth et al. Syst Biol 61(5): 717-726, 2012



UCEs were used to generate phylogeny for Hymenoptera

- Designed 1510 UCEs
- 2749 RNA probes
- Avg. 721 UCEs from 30 taxa
- 400 ng DNA input (avg)
- 2 runs of PE250 on MiSeq
- ~4X coverage

Faircloth et al. (Mol Ecol Res, 15:489-501, 2015)

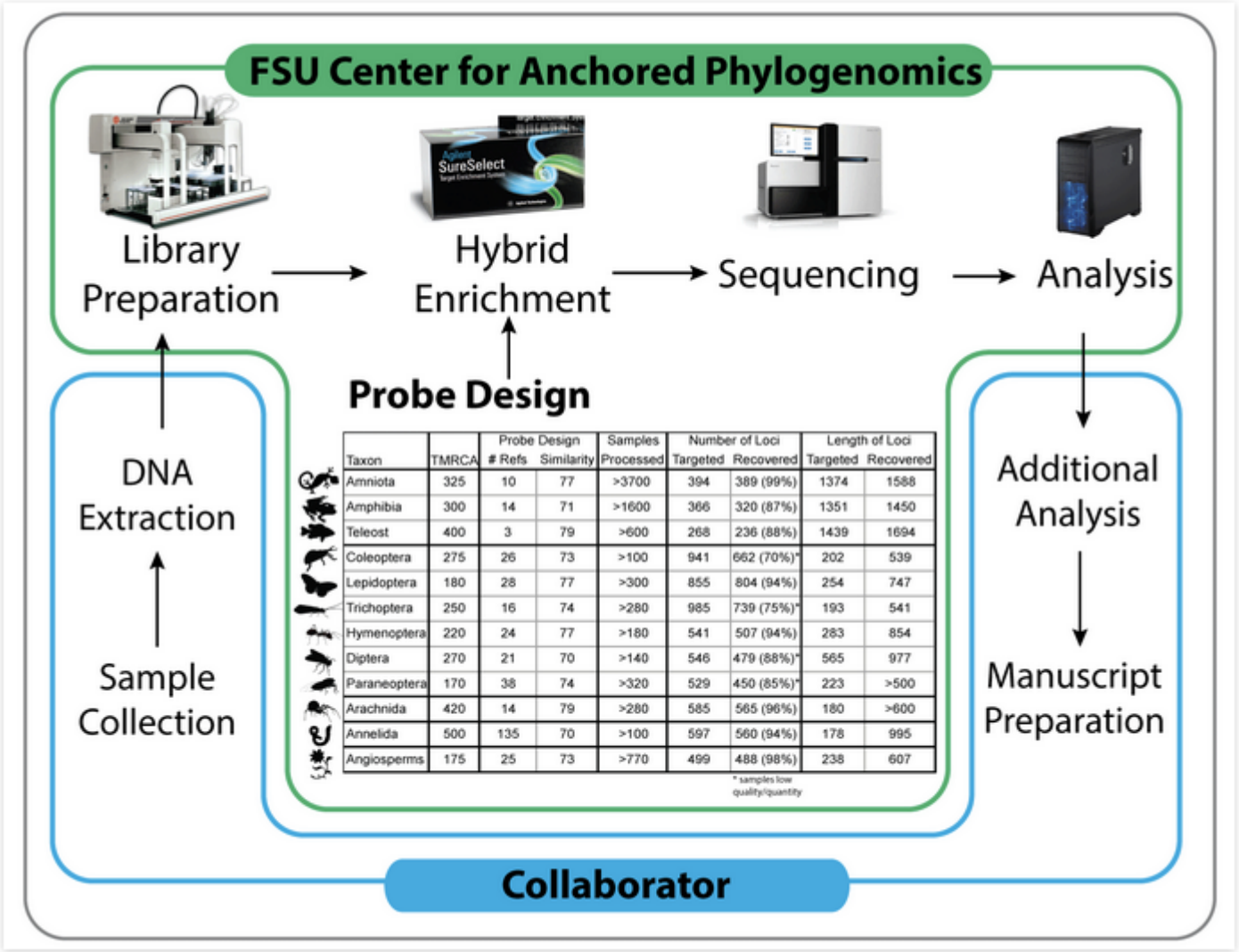
Anchored Hybrid Enrichment (AHE) for Phylogenetics

CENTER FOR
ANCHORED PHYLOGENOMICS
ACCELERATING THE RESOLUTION OF LIFE™

Agilent SureSelect Target Enrichment

Nuclear loci mostly designed from a few reference taxa

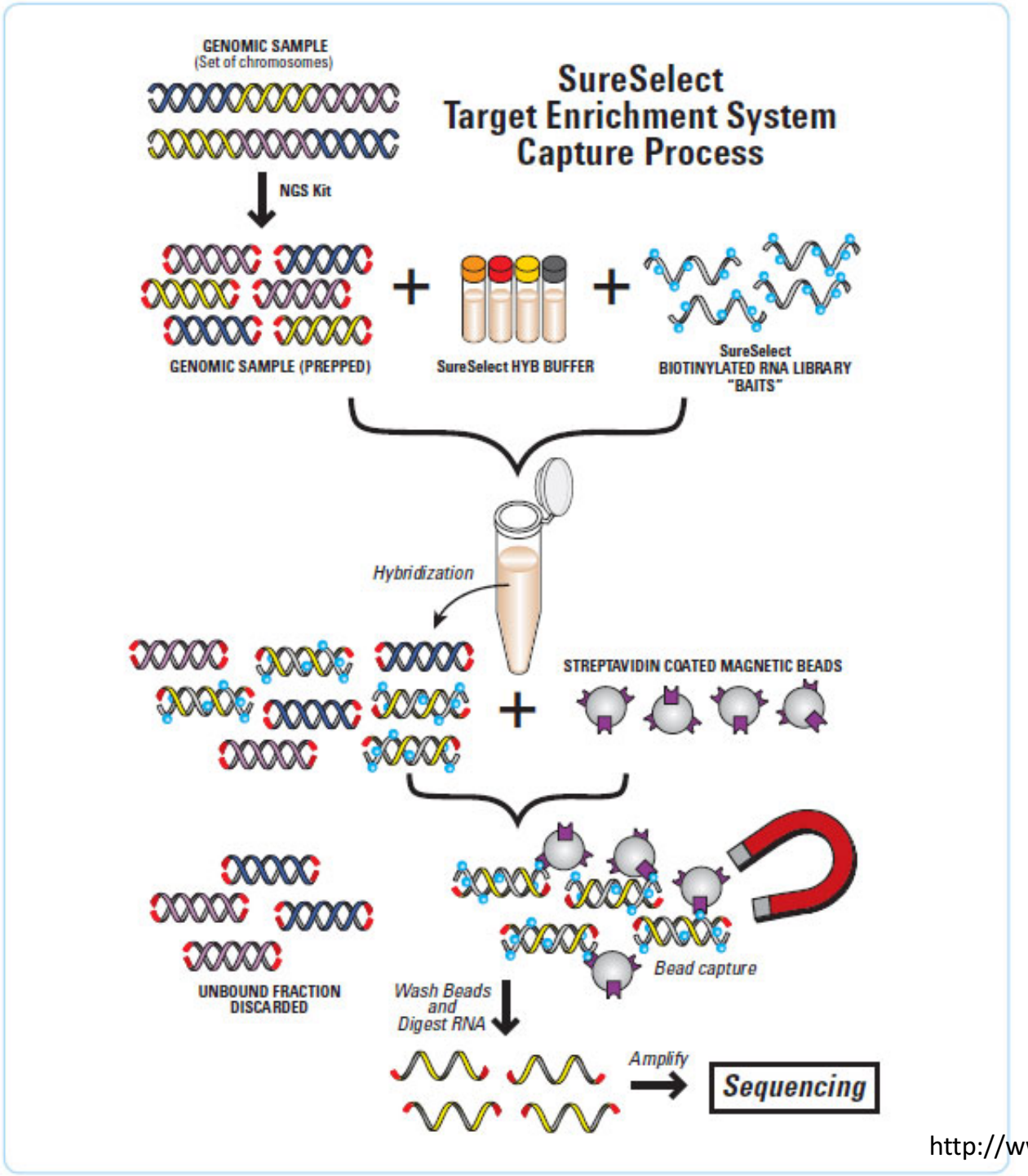
As low as 50ng DNA
But best 2ug DNA



Taxon	TMRCA	Probe Design # Refs	Probe Design Similarity	Samples Processed	Number of Loci Targeted	Number of Loci Recovered	Length of Loci Targeted	Length of Loci Recovered
Amniota	325	10	77	>3700	394	389 (99%)	1374	1588
Amphibia	300	14	71	>1600	366	320 (87%)	1351	1450
Teleost	400	3	79	>600	268	236 (88%)	1439	1694
Coleoptera	275	26	73	>100	941	662 (70%)*	202	539
Lepidoptera	180	28	77	>300	855	804 (94%)	254	747
Trichoptera	250	16	74	>280	985	739 (75%)*	193	541
Hymenoptera	220	24	77	>180	541	507 (94%)	283	854
Diptera	270	21	70	>140	546	479 (88%)*	565	977
Paraneoptera	170	38	74	>320	529	450 (85%)*	223	>500
Arachnida	420	14	79	>280	585	565 (96%)	180	>600
Annelida	500	135	70	>100	597	560 (94%)	178	995
Angiosperms	175	25	73	>770	499	488 (98%)	238	607

* samples low quality/quantity

http://anchoredphylogeny.com/



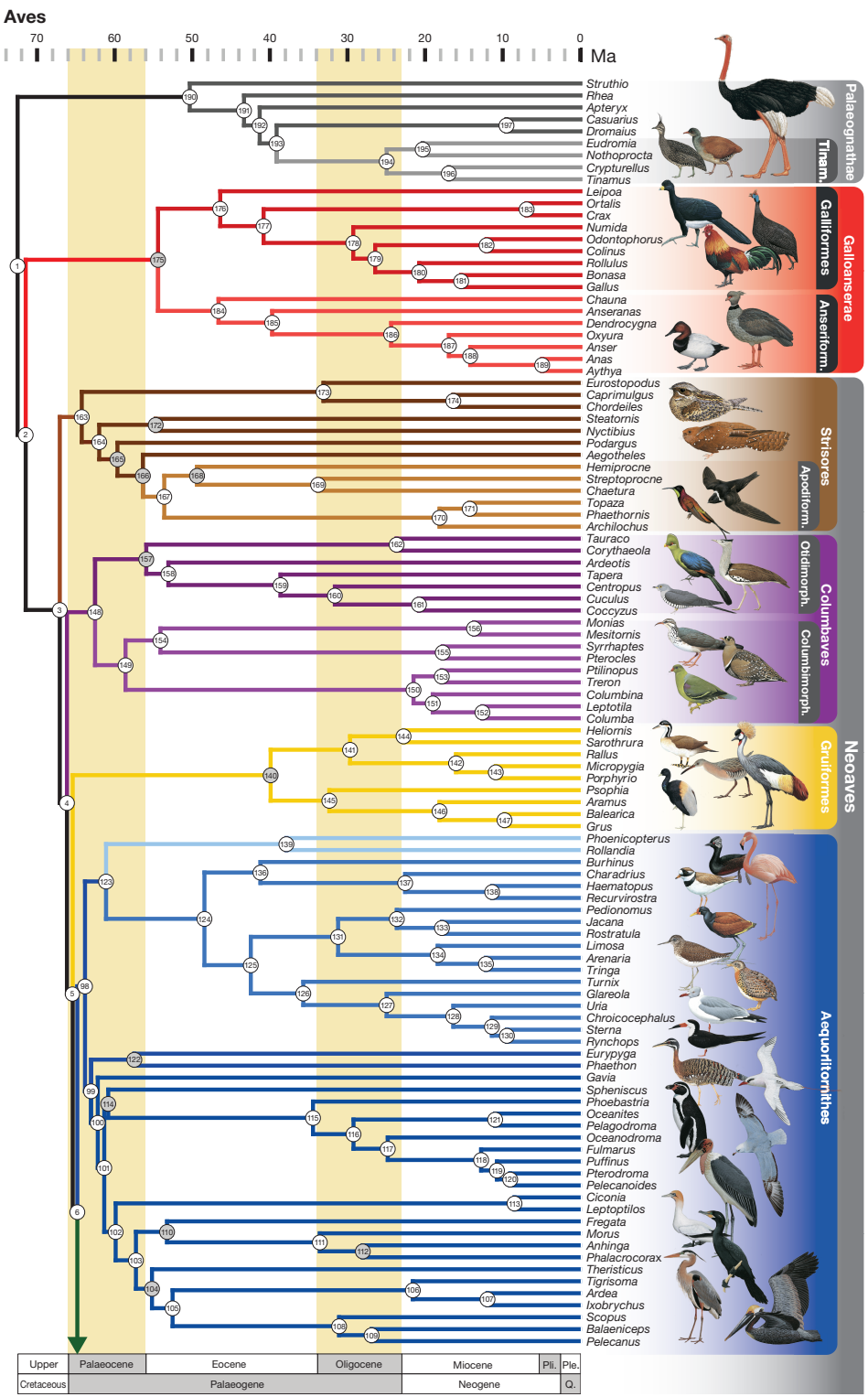
AHE target enrichment using Agilent SureSelect

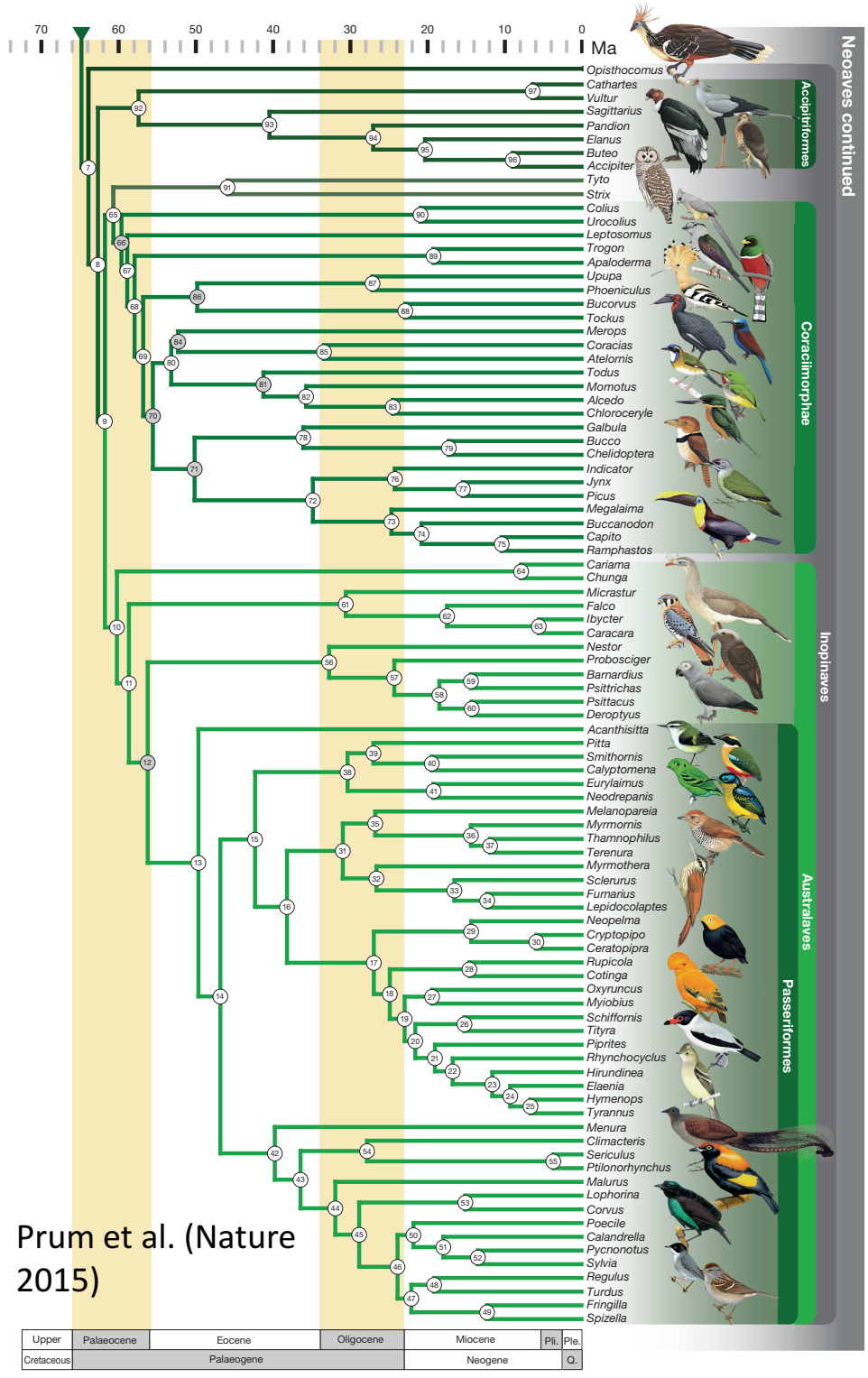
Originally used for exome capture

Bird data set (Prum et al. 2015):

- 394 vertebrate loci (use 259)
- ~1350 bp per locus
- HiSeq 2000 PE150, 4 lanes

http://www.genomics.agilent.com/article.jsp?pageId=3083





Prum et al. (Nature 2015)

Which method to use? (Caution)

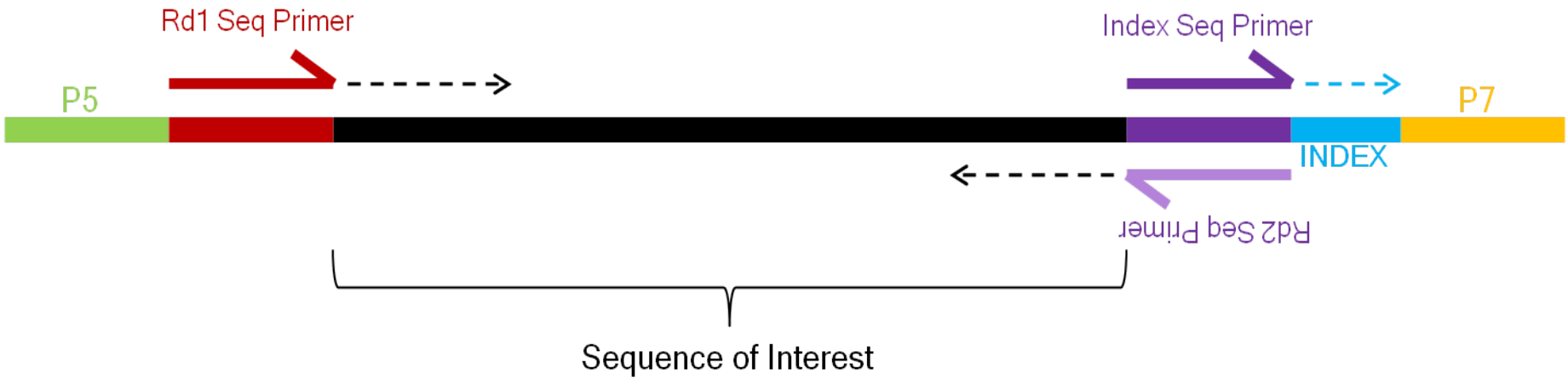
A few mitochondrial or nuclear genes

Reduced Representation sequencing
UCE, AHE, transcriptome, RADseq

** Method used could result in different Tree toplogy

Whole Genomes

STRUCTURE DETAILS



http://nextgen.mgh.harvard.edu/IlluminaChemistry.html