Sequencing Technology

Michael C. Zody, Ph.D.
Scott A Handley, Ph.D.

Workshop on Genomics
Cesky Krumlov
January 8, 2024

Logistics

- Introduction
- Please feel free to ask questions at any point
- Slides will be posted on workshop website
- Break/finish at about 60 minutes

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing

What is a read? What is a library?

- Definition of “read”: A single sequence from one fragment in the sequencing library (one cluster, bead, *etc*.)
- If generating paired reads, then 2 reads derived from each fragment in the library
- Definition of “library”: A collection of DNA fragments that have been prepared to be sequenced
- Definition of “coverage”: The number of reads spanning a particular base in the genome

Where does library come from?

› Cell. 1978 Oct;15(2):687-701. doi: 10.1016/0092-8674(78)90036-3.

The isolation of structural genes from libraries of eucaryotic DNA

T Maniatis, R C Hardison, E Lacy, J Lauer, C O'Connell, D Quon, G K Sim, A Efstratiadis

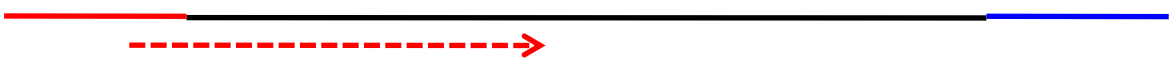
PMID: 719759 DOI: 10.1016/0092-8674(78)90036-3

Abstract

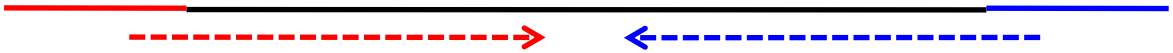
We present a procedure for eucaryotic structural gene isolation which involves the construction and screening of cloned libraries of genomic DNA. Large random DNA fragments are joined to phage lambda vectors by using synthetic DNA linkers. The recombinant molecules are packaged into viable phage particles in vitro and amplified to establish a permanent library. We isolated structural genes together with their associated sequences from three libraries constructed from Drosophila, silkmoth and rabbit genomic DNA. In particular, we obtained a large number of phage recombinants bearing the chorion gene sequence from the silkmoth library and several independent clones of beta-globin genes from the rabbit library. Restriction mapping and hybridization studies reveal the presence of closely linked beta-globin genes.

Types of reads

- Fragment reads (come from fragment libraries)
  - Single read in one direction from a fragment

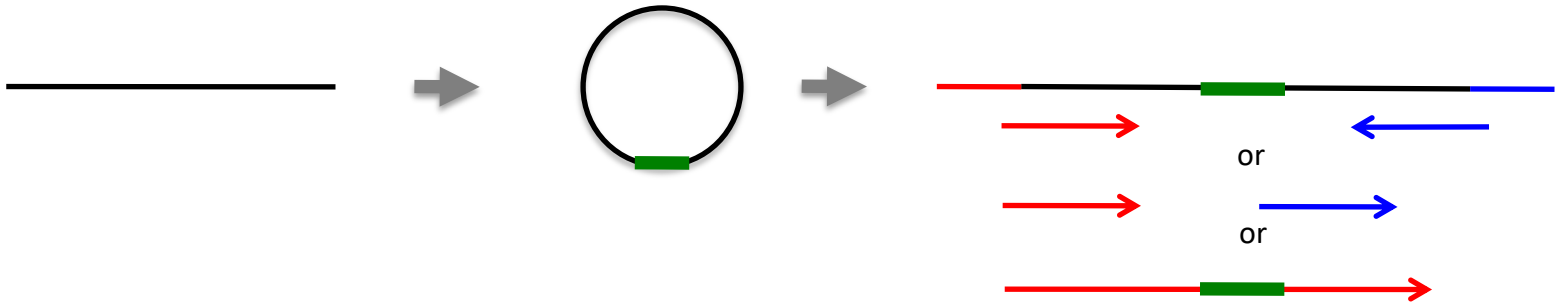


- Paired end reads (come from fragment libraries)
  - Two reads from opposite ends of the same fragment
  - Reads point towards each other



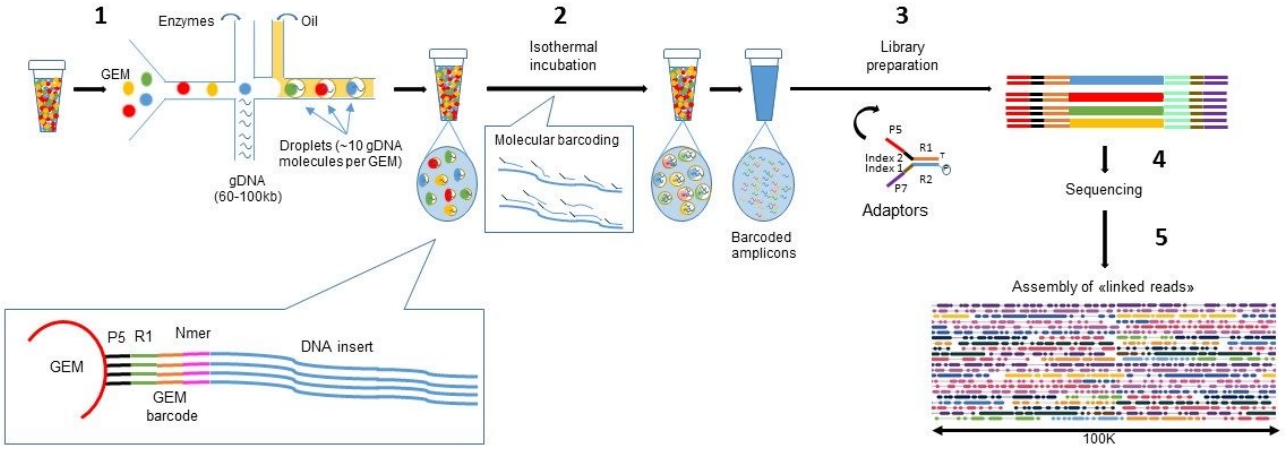
Types of reads

- Mate Pair Reads (come from Jumping Libraries)
  - Long fragment of DNA is circularized
  - Junction is captured (e.g., by biotinylated adapter)
  - Remainder is cleaved (many methods)
  - Ends are sequenced
  - Read orientations depend on the exact method



Types of reads

- Linked reads (several methods)
  - Long (10-100kb) DNA isolated
  - DNA is labeled (barcode, mutation, etc.)
  - Read pairs generated from specific long fragments
  - Sequence normal read pairs
  - Can use reads normally for alignment/assembly
  - In some methods, can assemble the long fragment



Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing

Sanger Sequencing (1977)

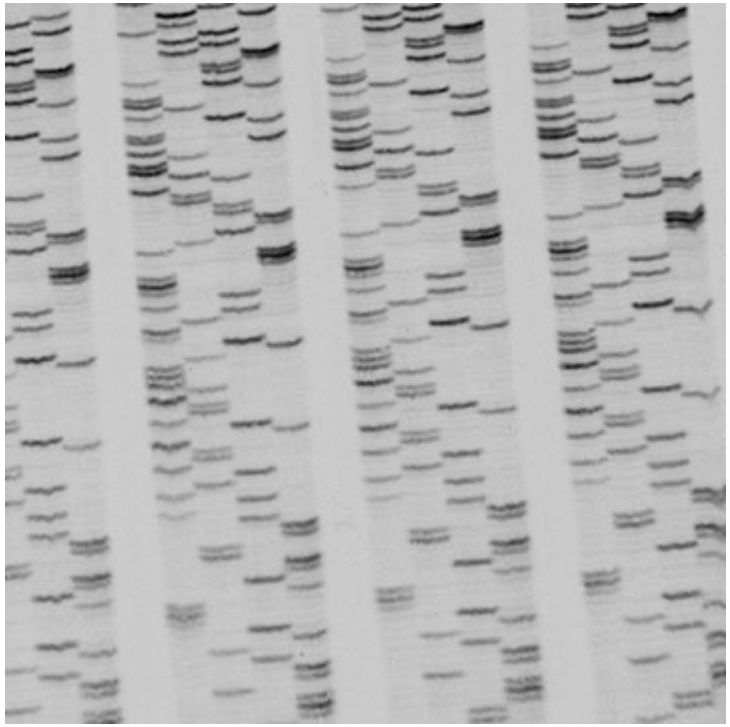
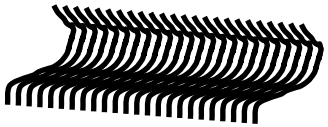
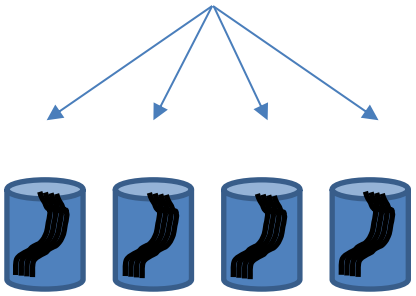


Image credit: https://unlockinglifescode.org/timeline/11

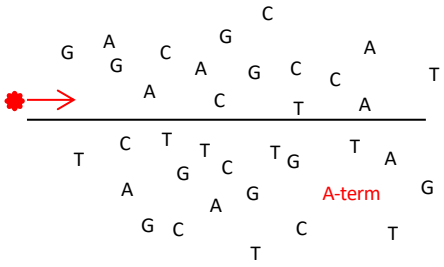
How Sanger Sequencing Works



Single-stranded DNA



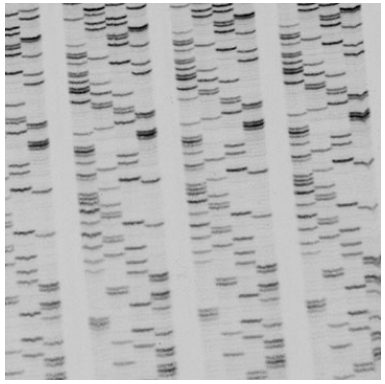
Split into four reactions



Add labeled primer and nucleotides, one of which is sometimes terminated

Allow the strands to extend

A	T	A	T	A	T
G	C	G	C	G	C
C	G	C	G	C	G
T	A	T	A	T	A
G	C	G	+	G	C
A	T	A		A	T
C	G	C		C	G
T	A	T		T	A
A	T	A		A	+
G	C	G		G	
A	T	A		A	
T	A	T		T	
C	+	C		C	
G		G		G	



Automation of Sanger (1986)

- Replacement of radioactive label with laser-excitable fluorescent dyes
- Allowed all four nucleotides to be run in a single lane of a gel
- Base sequence could be read off with a camera as the fluorescing strands passed a certain point near the end of the gel
- Signal from each lane could be converted to a nucleotide sequence by a computational process called base calling

Fluorescent slab gel image

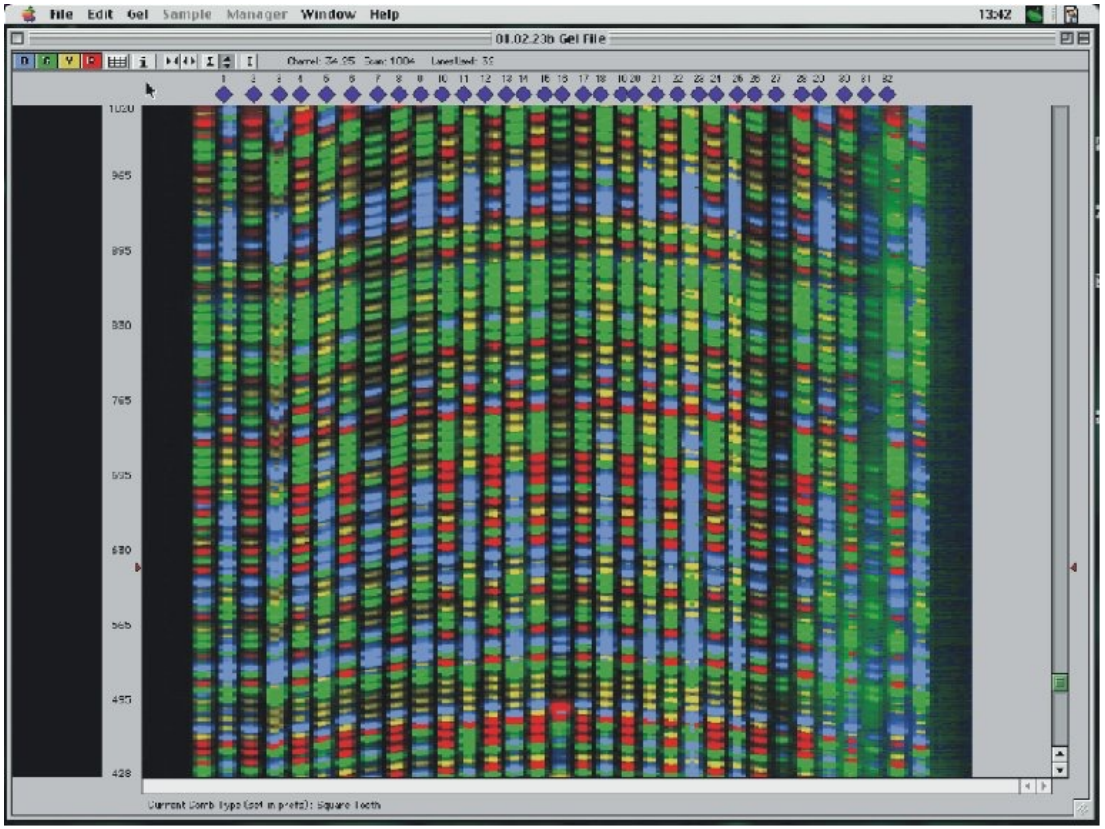


Image credit: http://www.mun.ca/biology/scarr/How_it_works.htm

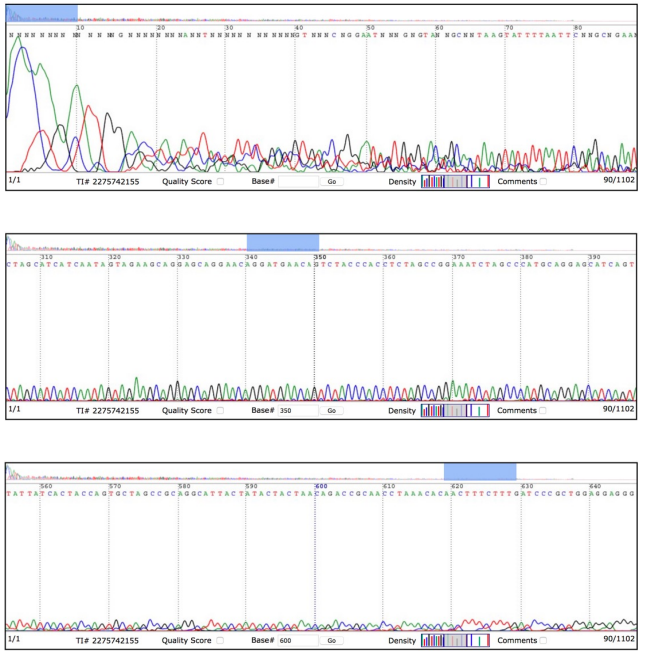
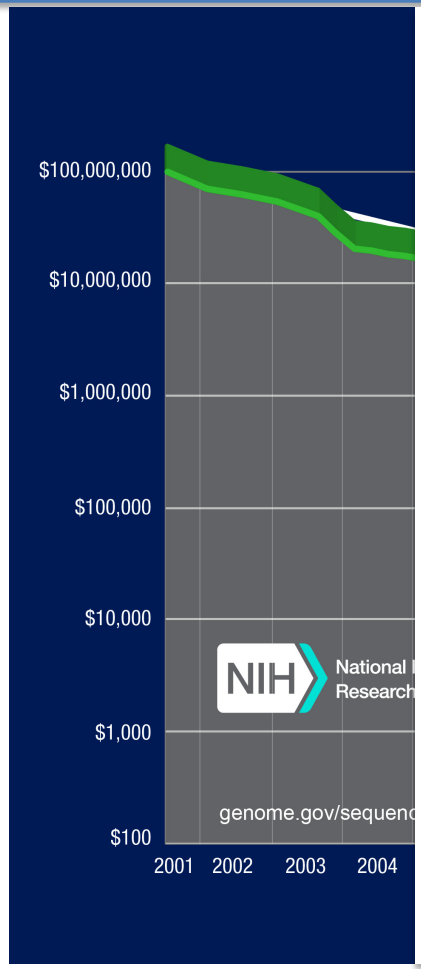


Image credit: NCBI Trace Archive

Capillary Gel Sequencing (1998)

- Replacement of 2D slab gels with an array of enclosed capillaries
- Cleaner signal processing
- Fully automated loading
- Faster run times

Cost Curve for Sanger Gel Sequencing



Other Early Technologies

- Maxam-Gilbert sequencing
- LI-COR
- Molecular Dynamics MegaBACE
- Pyrosequencing
- Mass spectrometry

Statistics of Apex Capillary Sequencing

- 96 reads per run
- 700-1000 bases per read
- Very high base accuracy over most of the read length (<1/100,000)
- ~$1 per read
- ~1 run per hour
- ~2 million bases per machine per day
- Large sequencing centers could do a single mammalian genome to assembly depth in about 2-3 months

Limits on Sanger Gel Performance

- Tradeoff between loss of signal due to diffusion and loss of resolution at high voltage or short gel length
- Longer gels/capillaries or lower voltages provide better separation of short to medium fragments
- Longer gel run times mean more diffusion of fragments in the gel, which blurs adjacent signal and spreads peaks
- The maximal high quality read length is around 1000 bp

454 Sequencing

- First “Next Generation” or massively parallel technique
- Based on pyrosequencing
- Emulsion PCR DNA prep on beads
- Beads loaded into a picotiter plate for sequencing

454 Sequencer

Emulsion PCR

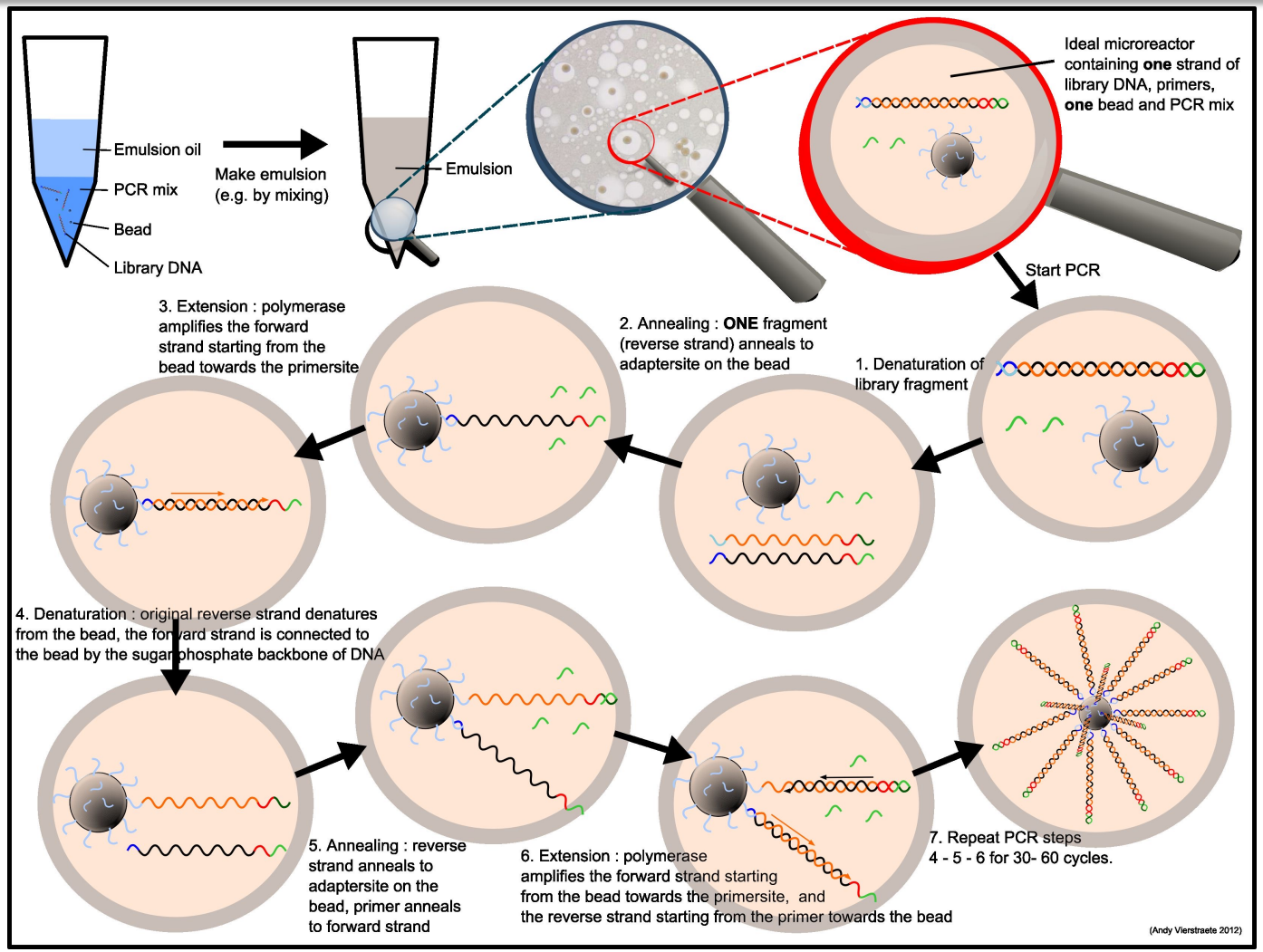
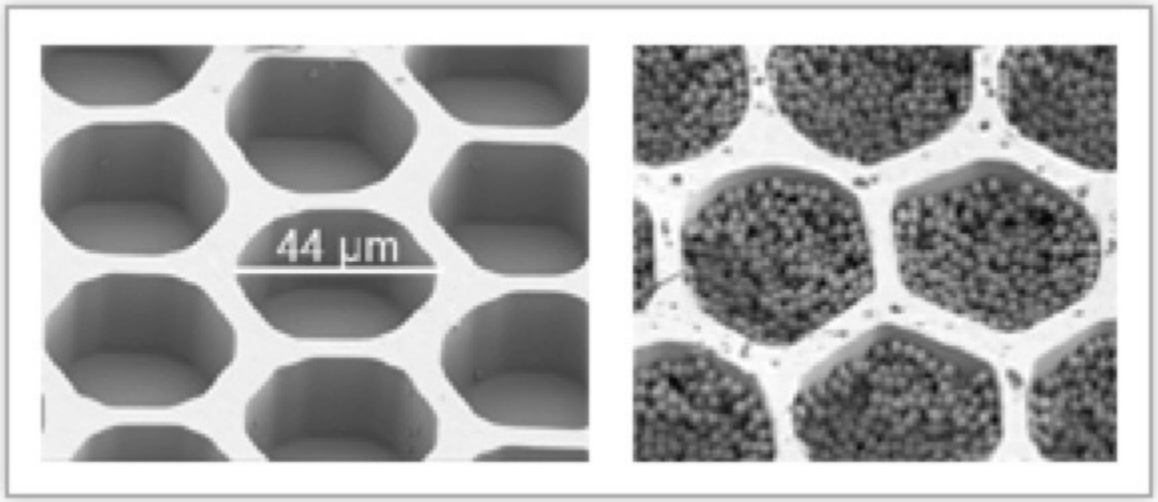


Image credit:
https://users.ugent.be/~avierstr/nextgen/nextgen.html

454 Picotiter Plate



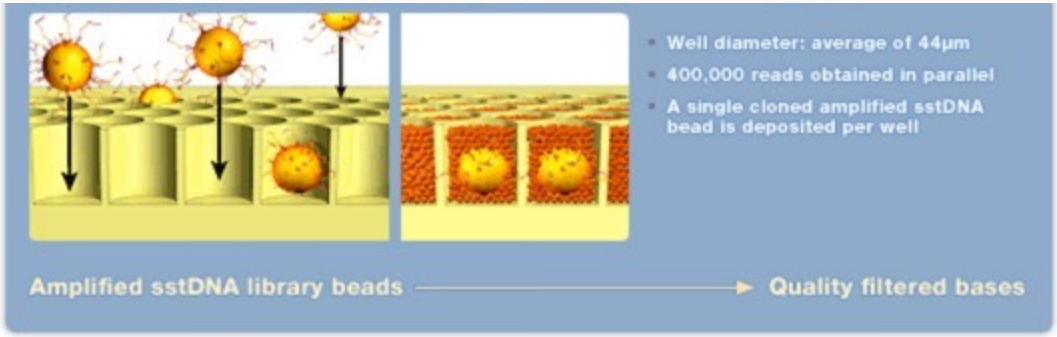
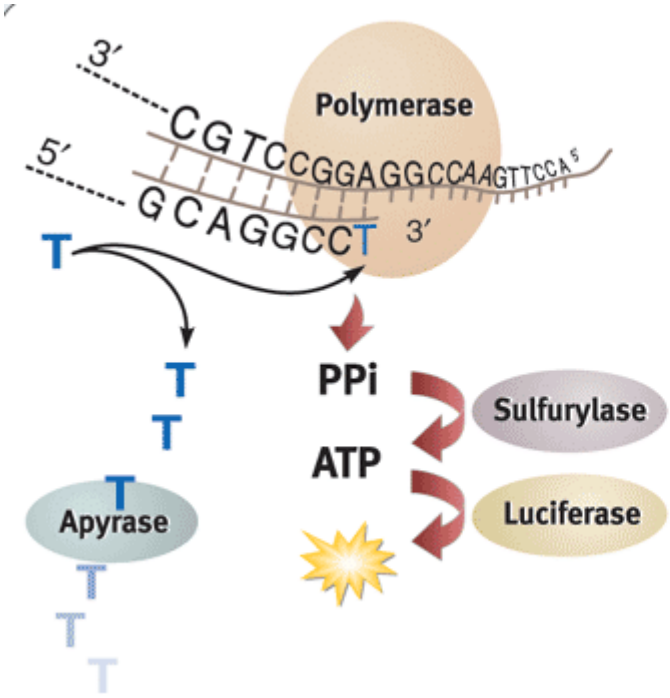
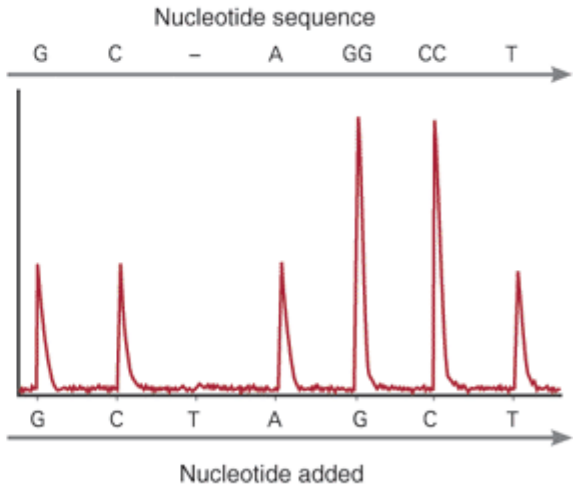


Image credit: http://www.mbio.ncsu.edu/MB451/lectureModules/molecularEcology/molecularSurveys/454/454.html

Pyrosequencing





454 Output: the Flowgram

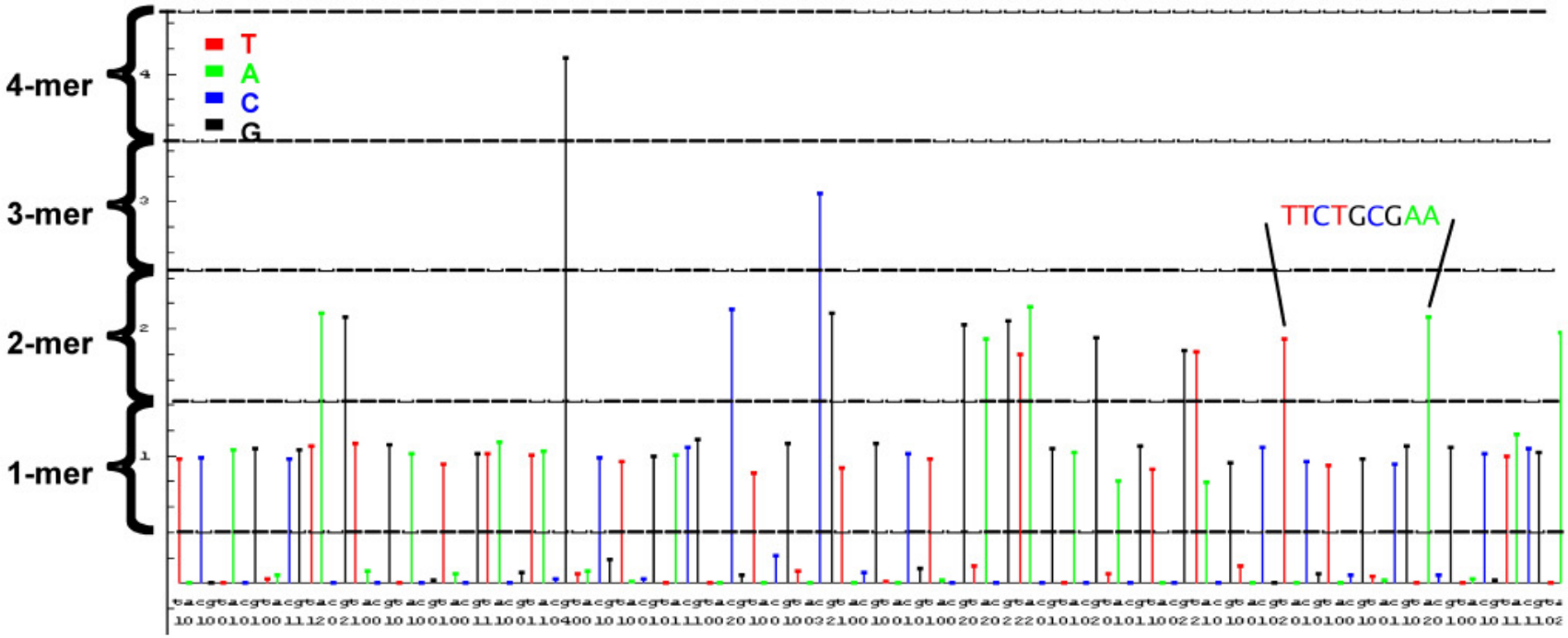
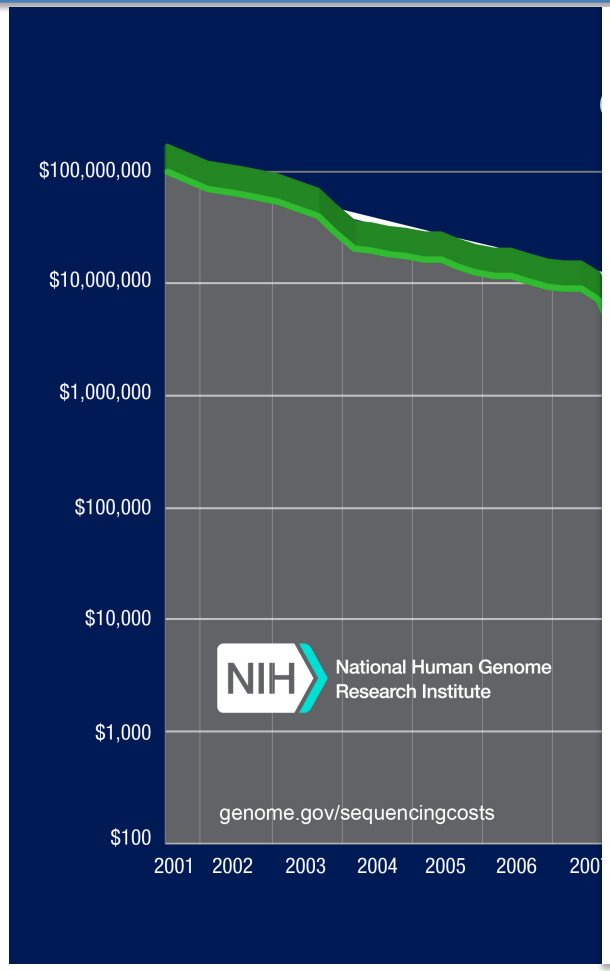


Image credit: https://contig.wordpress.com/2010/10/28/newbler-input-i-the-sff-file/

Cost Curve for 454 Sequencing



Statistics of Apex 454 Sequencing

- 1 million reads per run
- 400-500 bases per read (750?)
- High error rate (~1.5%), very motif dependent (homopolymers)
- Cost several thousand dollars per run
- ~10 hours per run
- ~1 billion bases per machine per day

Limits on 454 Performance

- Failure to accurately read homopolymers or sequences near homopolymers; physical limit on ability to read the full incorporation
- Loss of signal over time
- Signal/noise degradation due to asynchrony of extending strands
- Length of fragment that could be amplified on bead in emPCR
- No ability to sequence the second strand of DNA or do non-contiguous reads

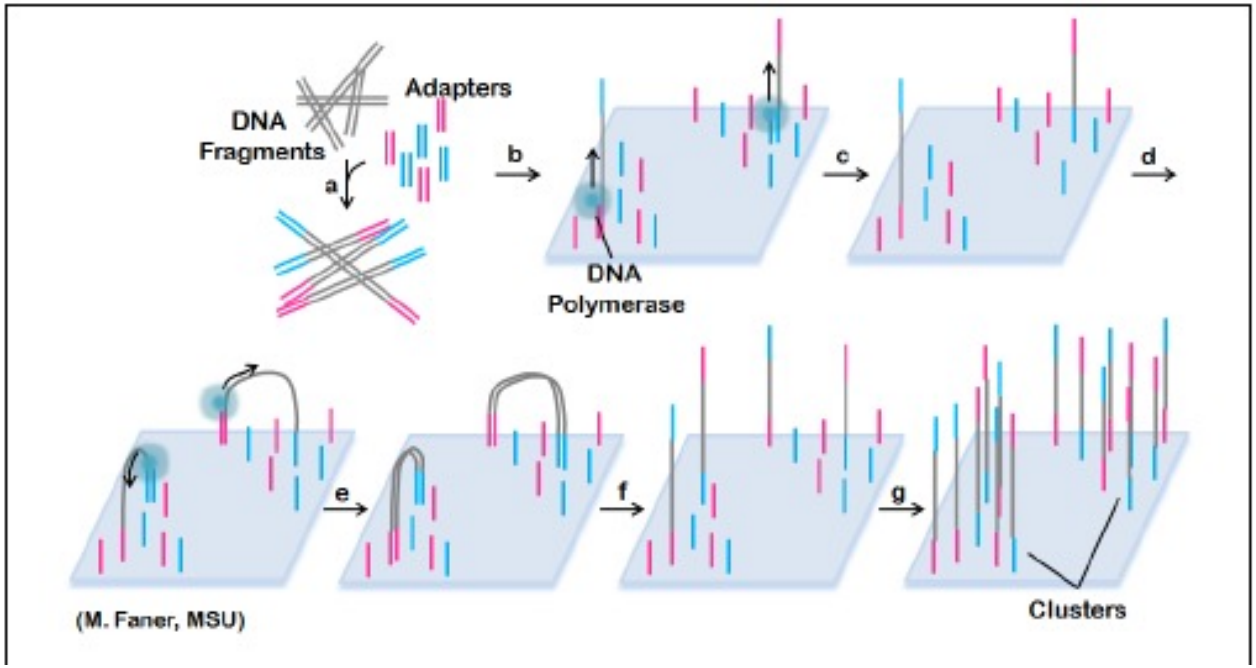
Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing

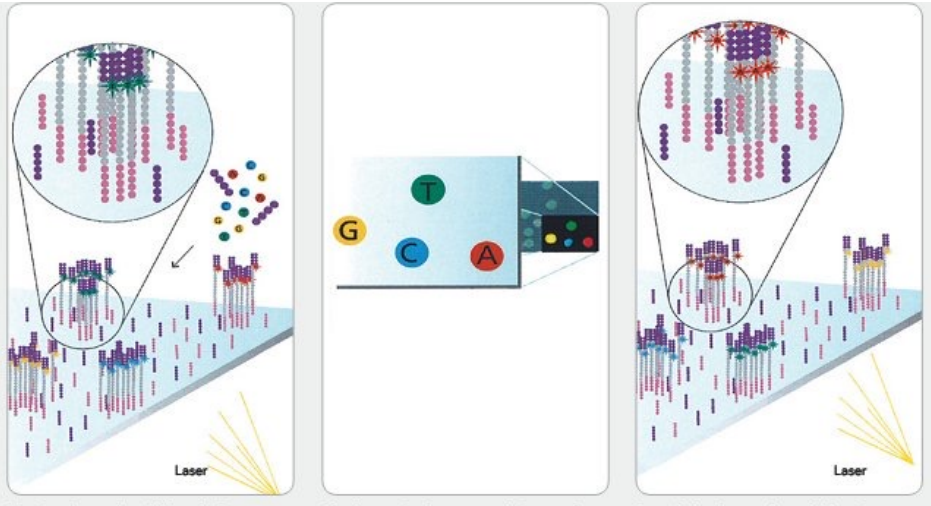
Illumina (Solexa) Sequencing

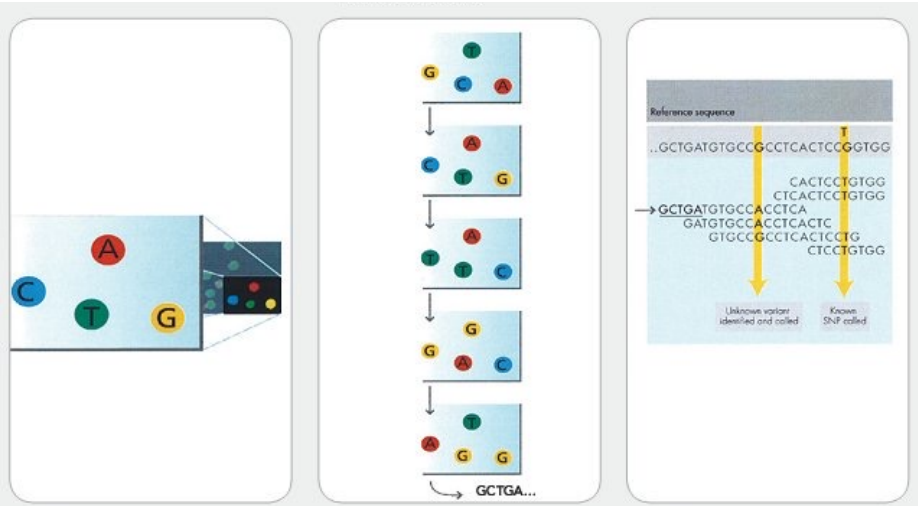
- Sequencing of DNA strands amplified *in situ* on a glass slide
- Use reversible terminators to sequence one base at a time

Bridge Amplification



Reversible Terminator Sequencing

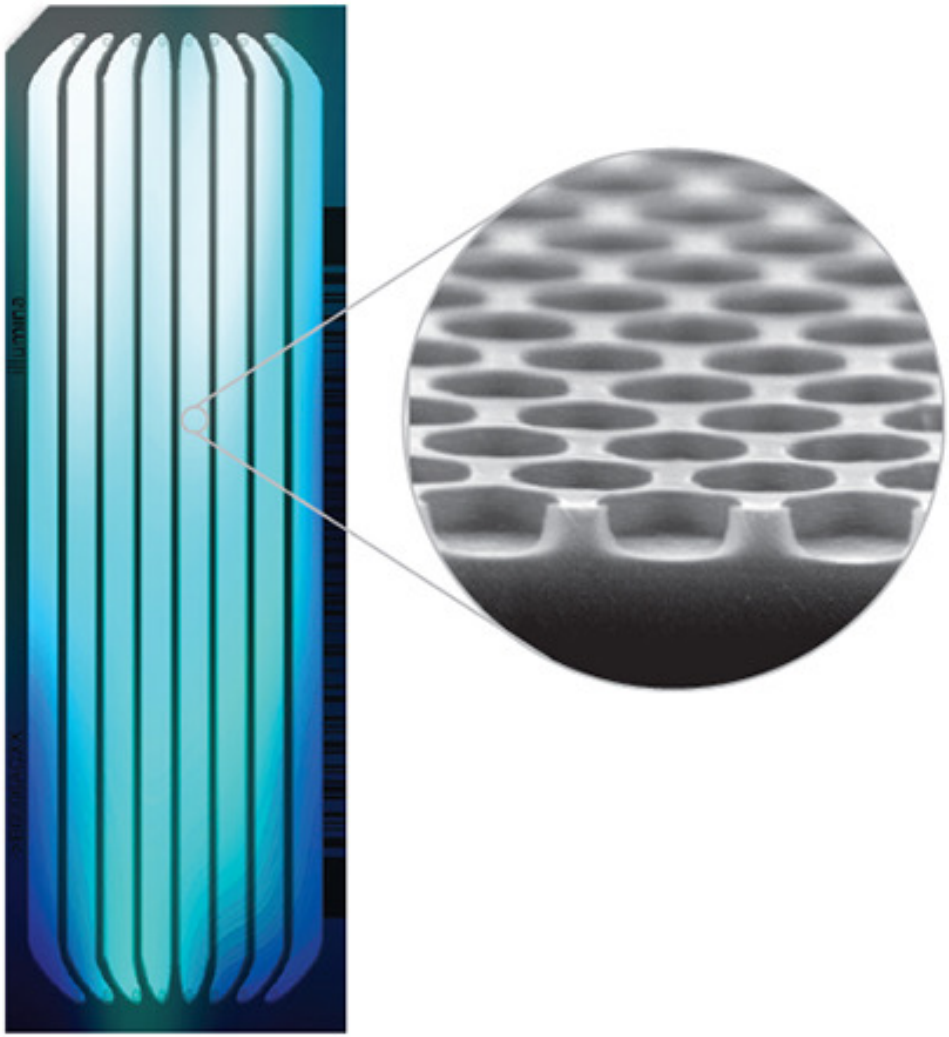




Recent Changes in Illumina

- Patterned flowcells
- Exclusion amplification
- 2 color chemistry

Patterned Flowcells

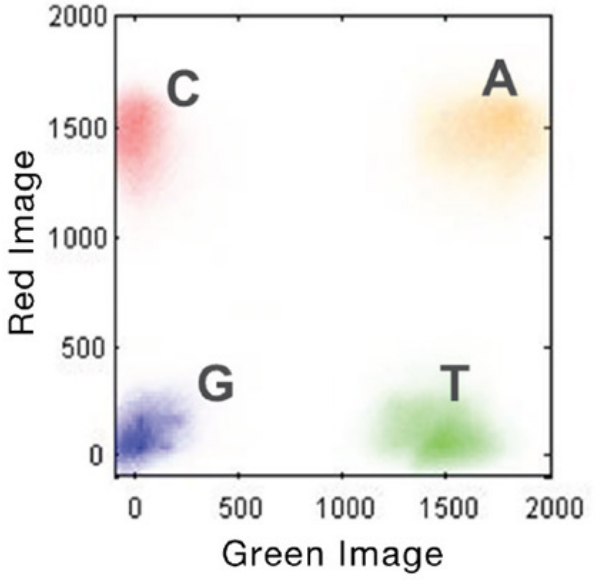


Exclusion amplification

- No more bridge PCR on patterned flowcells
- Fragments rapidly amplify as soon as they arrive at the patterned spot
  - Prevents a second fragment from amplifying there
  - Allows overloading to maximally fill flowcell
  - (Not perfect)
- Exact method is not described (see patent)
- Results in “proximal duplicates” or “pad hops”
- Problems with “index switching”

2 color chemistry

- One base (A) labeled with 2 colors
- One base (G) unlabeled
- Allows faster image scanning
- Dead clusters look like runs of G
  - Mostly do not align (in human)
  - ”Supplemental” alignments



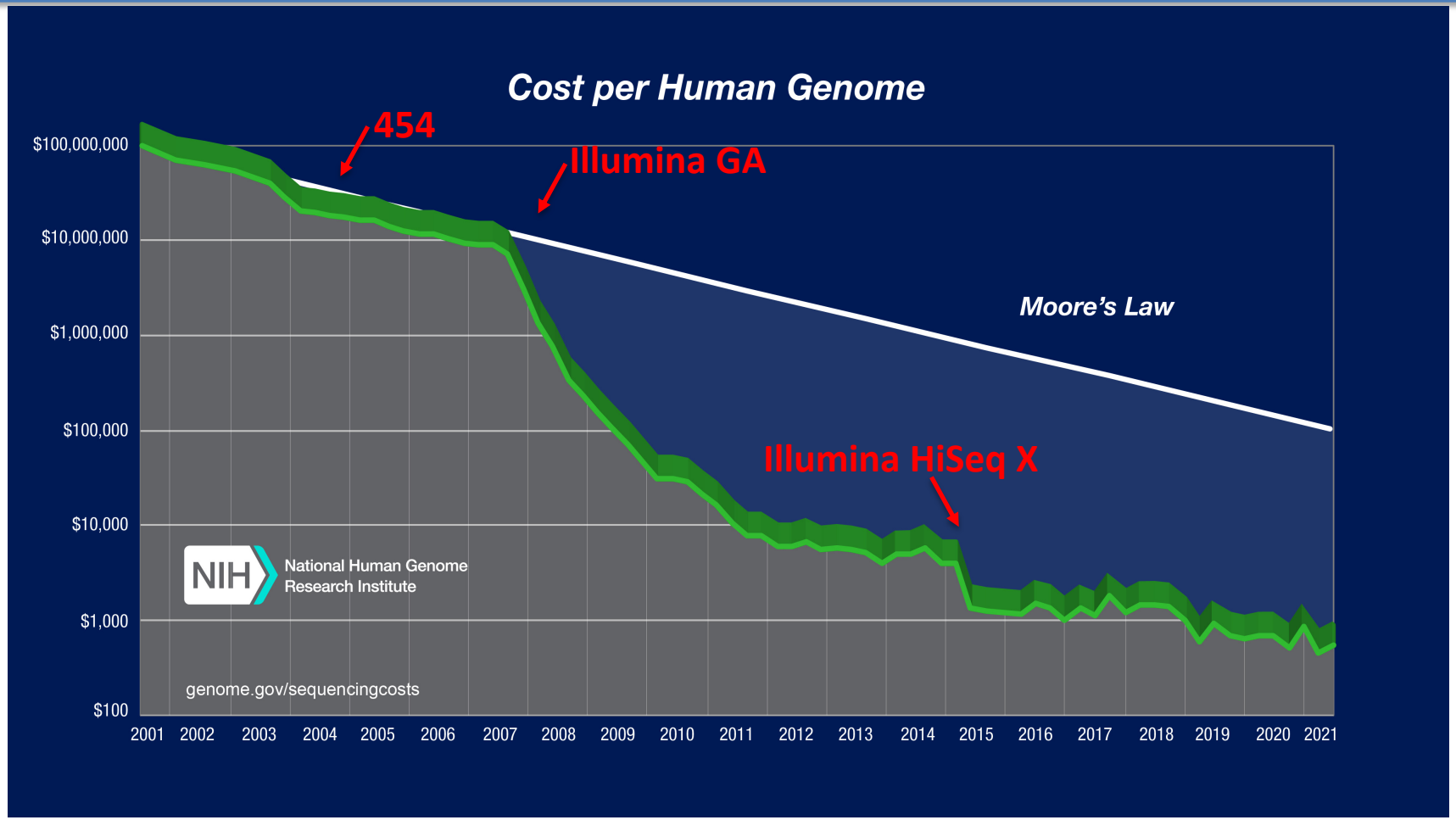
Statistics of Apex (so far) Illumina

- 25 billion read pairs per run
- 300 bases per read pair
- Relatively low error (<1%), some context dependency
- Cost ~$12,500 per run (for NovaSeq X Plus on largest flow cell size)
- ~2 days per run
- ~4 trillion bases per machine per day

Limits on Illumina Performance

- Loss of signal over time
- Signal/noise degradation due to asynchrony of extending strands
- Viability of sequencing reagents over the course of a run
- Length of fragment that could be amplified into clusters on the slide

Cost Curve for Illumina Sequencing



Moore's law is the observation that the number of transistors in an integrated circuit (IC) doubles about every two years. Moore's law is an observation and projection of a historical trend. Rather than a law of physics, it is an empirical relationship linked to gains from experience in production.

Other Next Generation Technologies

- SOLiD
  - Ligation rather than polymerase based
  - Used redundant base sampling with error correction (“color space”) to enhance error rate (<0.1%), but made analysis very challenging
  - Short reads, limited second read capability
- Ion Torrent
  - Like 454 (emPCR, well-based sequencing)
  - Uses direct measurement of pH changes with base addition (”post-light”)
- Helicos
  - Like Illumina but with single molecules

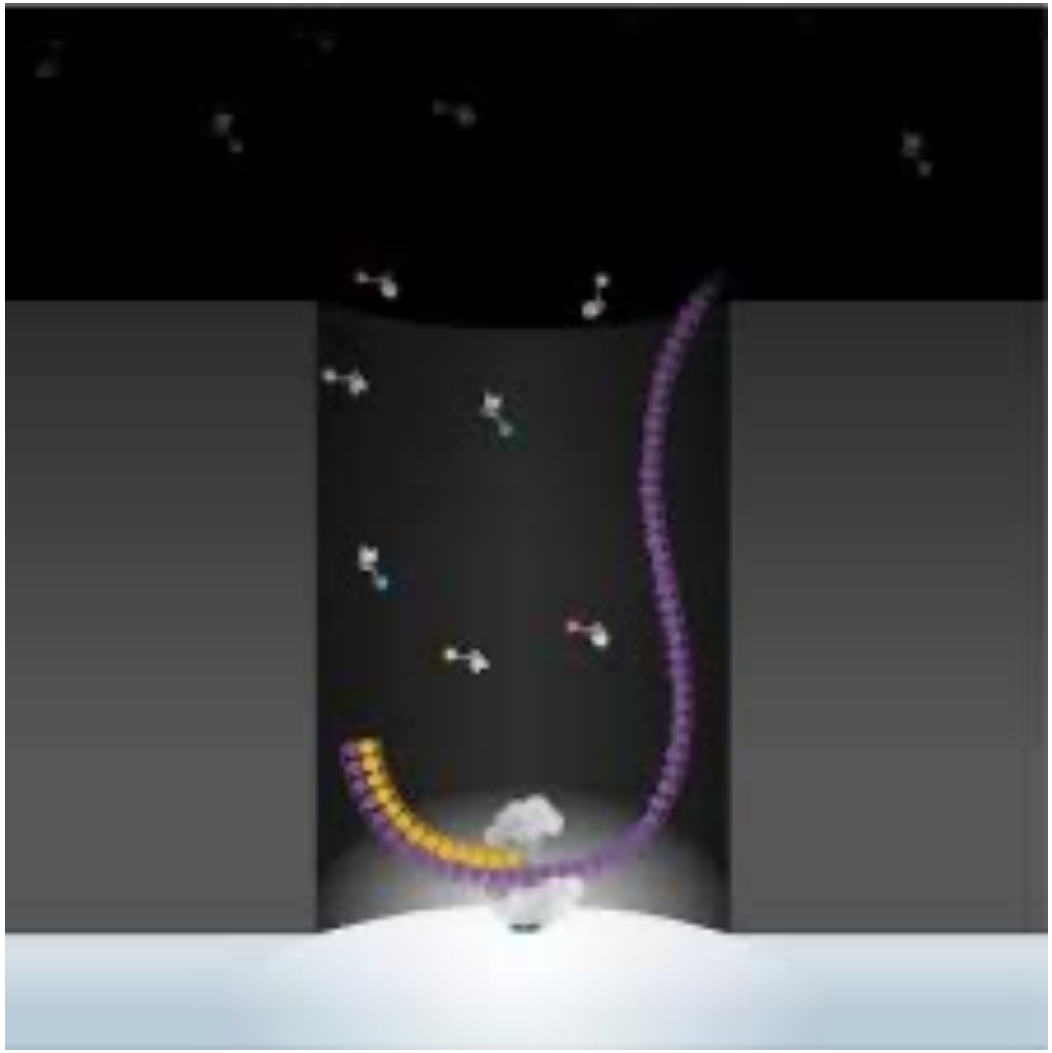
New Second Generation Methods

- MGI
- Singular
- Element
- Ultima

Single Molecule (Third Generation) Methods

- Pacific Biosciences SMRT (Single Molecule Real Time) sequencing
  - Uses a polymerase anchored in a zero-mode waveguide
  - Images all wells at the same time in real time with digital video
  - Interprets bases by the light signal visible at incorporation
  - Very large instrument
- Oxford Nanopore Technologies
  - Uses protein nanopores in synthetic membrane to thread DNA through
  - Current sensors measure change in fluid current flow through pore to differentiate groups of multiple bases as they occupy the pore
  - Very small instrument (attaches to compute like a USB drive)

PacBio SMRT

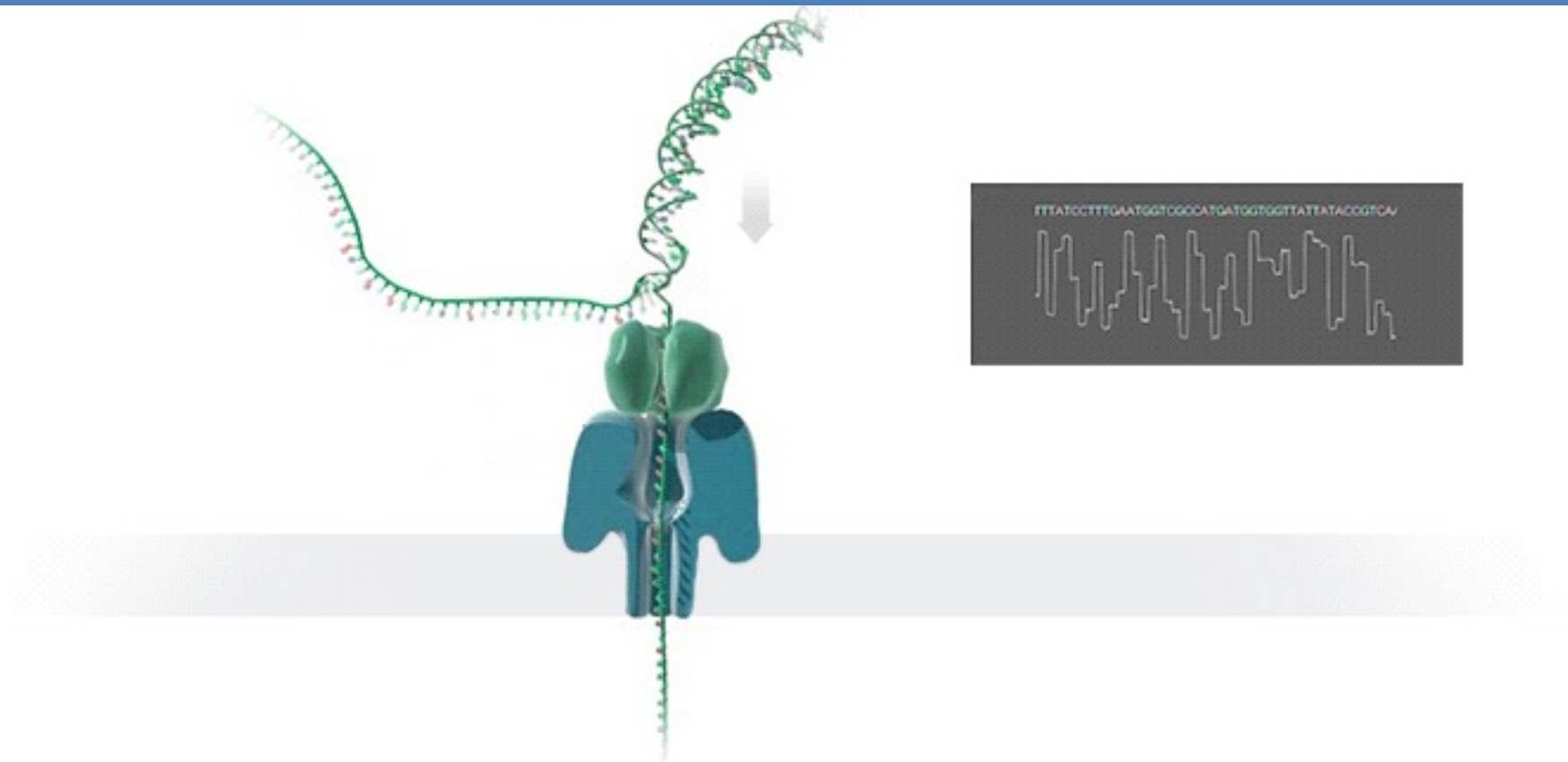


https://www.youtube.com/watch?v=WMZmG00uhwU&feature=youtu.be

PacBio “Hi-Fi” Reads

- Sized libraries, 10-20kb long
- Generate circular consensus on these
- Can read each linear piece 7-10 times
- Generates very high accuracy long reads
- Can assemble easily and even distinguish highly similar repeats

Oxford Nanopore



Current State of Oxford Nanopore

- Very long reads (100,000+, >1 Mbp)
- High error rates (10%+), non random
- Can read both strands to improve accuracy (<1% error, but not all reads)
- Low throughput
- Run times variable (few hours to 3+ days)
- Higher costs per base than Illumina (2-fold)
- Reads end when pores die

Limitations of Single Molecule Techniques

- Single molecule means no redundancy, so error rates will be high unless the same molecule can be read more than once
- Methods of detection are hard to massively parallelize
- Currently, these techniques actually require large amounts of DNA
- Getting very long reads requires very high quality input DNA
- Data processing and base calling is more expensive

Course Outline

- Terminology
- History of Sequencing
- Current Sequencing Technologies
- Prepping DNA for Sequencing

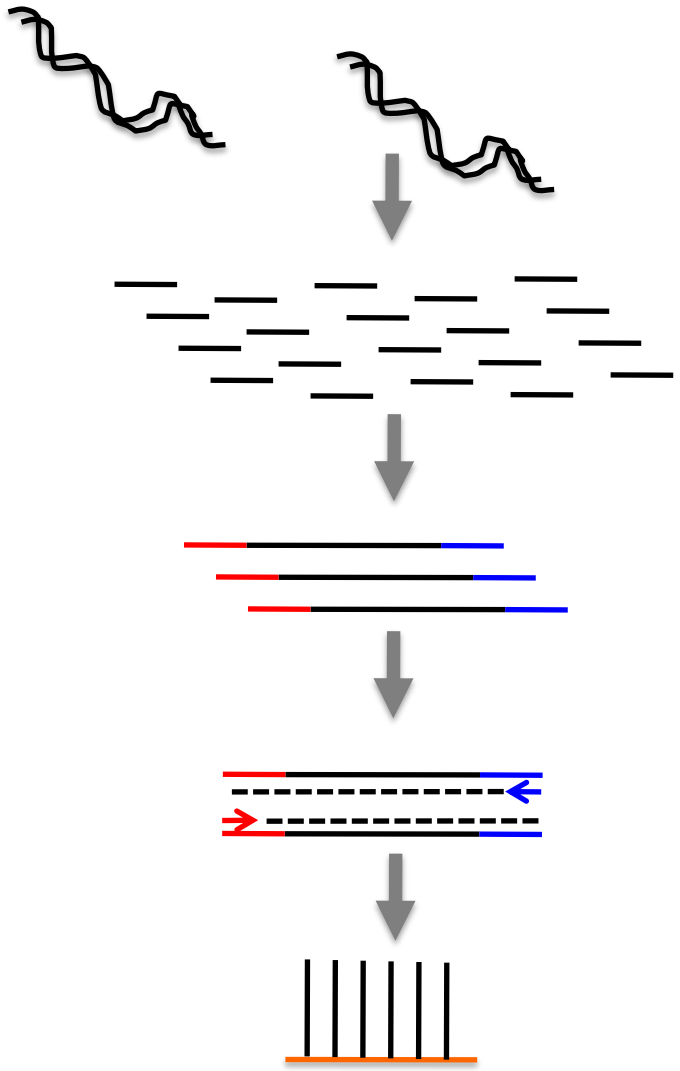
Prepping DNA for Sequencing

- Steps of library construction and sequencing
- Making Fragment libraries (to generate fragment or paired end reads)
- Making Jumping libraries (to generate mate pair reads)
- Pooling with or without barcoding
- Possible artifacts of library construction
  - PCR-based artifacts
  - Sequencing of primers, adapters, and tags

Steps of Library Construction

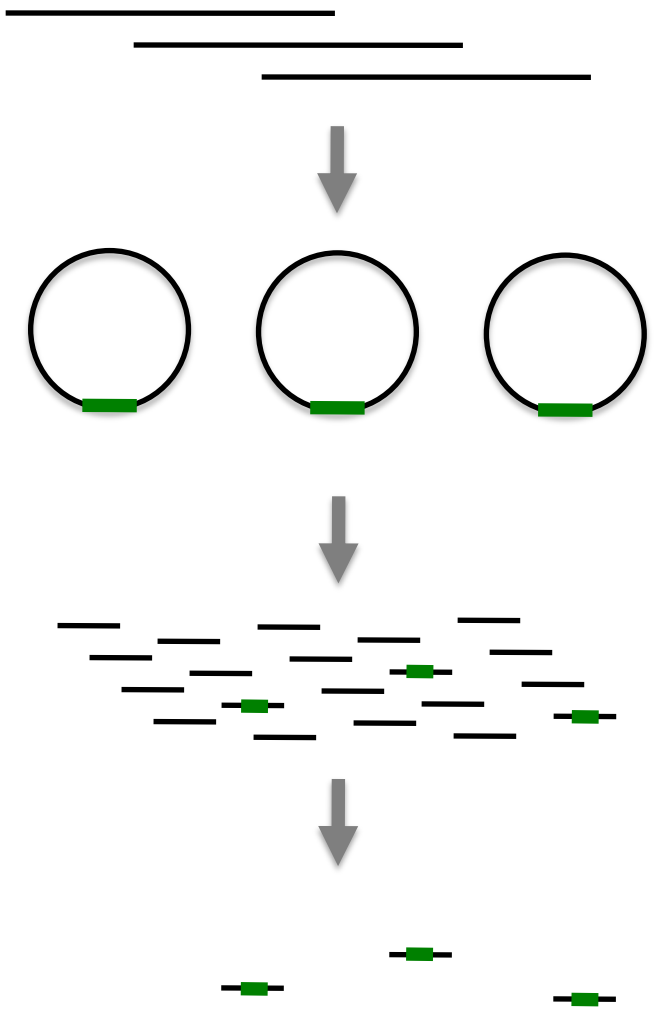
- Add adapters containing:
  - Barcodes (for multiplexing)
  - Sequencing primers
  - Amplification primers
  - Sequence for substrate attachment
- Amplify fragments by universal PCR
- Optionally pool barcoded libraries

Steps of Fragment Library Construction



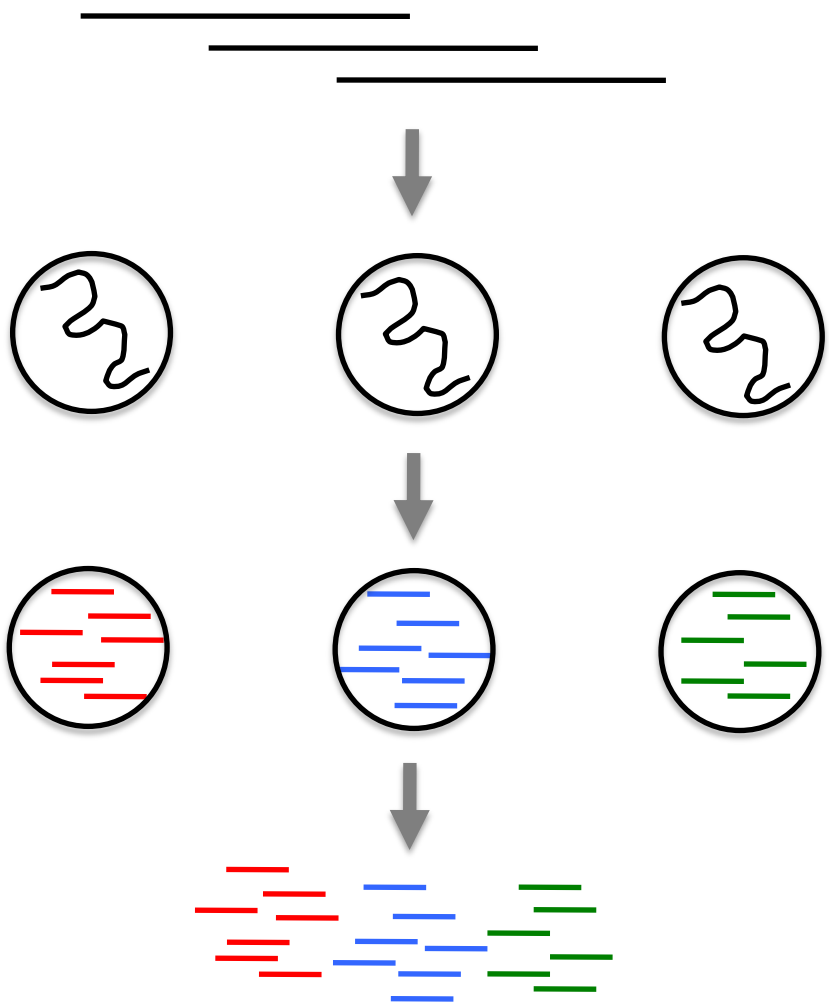
- Extract DNA
- Fragment and possibly size select (300-600 bp)
- Add adapters
- Amplify
- Select single molecules
- Amplify in clusters/beads

Steps of Jumping Library Construction



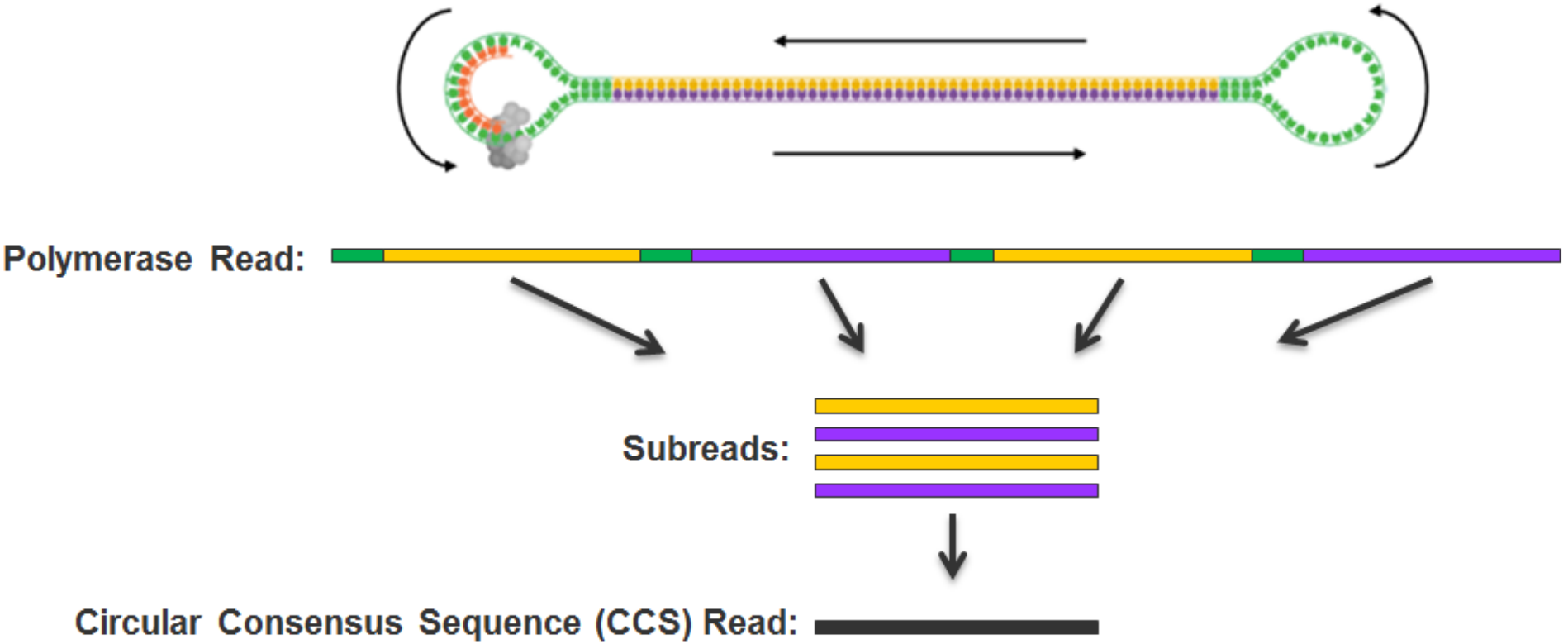
- Extract DNA, fragment and size select (2-40 kb)
- Circularize with labeled adapters
- Fragment and size select (300-600 bp)
- Select fragments containing labeled adapters
- Proceed as for fragment library

Steps of Linked Read Library Construction



- Extract DNA, fragment and size select (50+ kb)
- Isolate large fragments
- Fragment, barcode, and size select (300-600 bp)
- Pool and proceed as for fragment library

Steps of PacBio Library Construction



- Extract DNA, fragment and size select (50+ kb)
- Add hairpin adapters to both ends

Pooling with barcoding

- Unique DNA tags identify samples
- Allows multiple distinct samples on one run/lane
- Advantages:
  - Reduced cost of sequencing for small samples
  - Analysis is identical to unpooled data
- Disadvantages:
  - Some small throughput loss due to barcode fails
  - Data mis-assignment from bad barcode reads
  - Increased per sample cost for library construction

Pooling without barcoding

- Mix input DNA without identification
- No way to definitively separate data from different samples afterwards
- Advantages:
  - Single library prep for a number of samples
  - No yield lost to barcodes
- Disadvantages:
  - Loss of all individual associations
    - Loss of ability to use replicates!
  - No check on accuracy of pooling

PCR-based artifacts

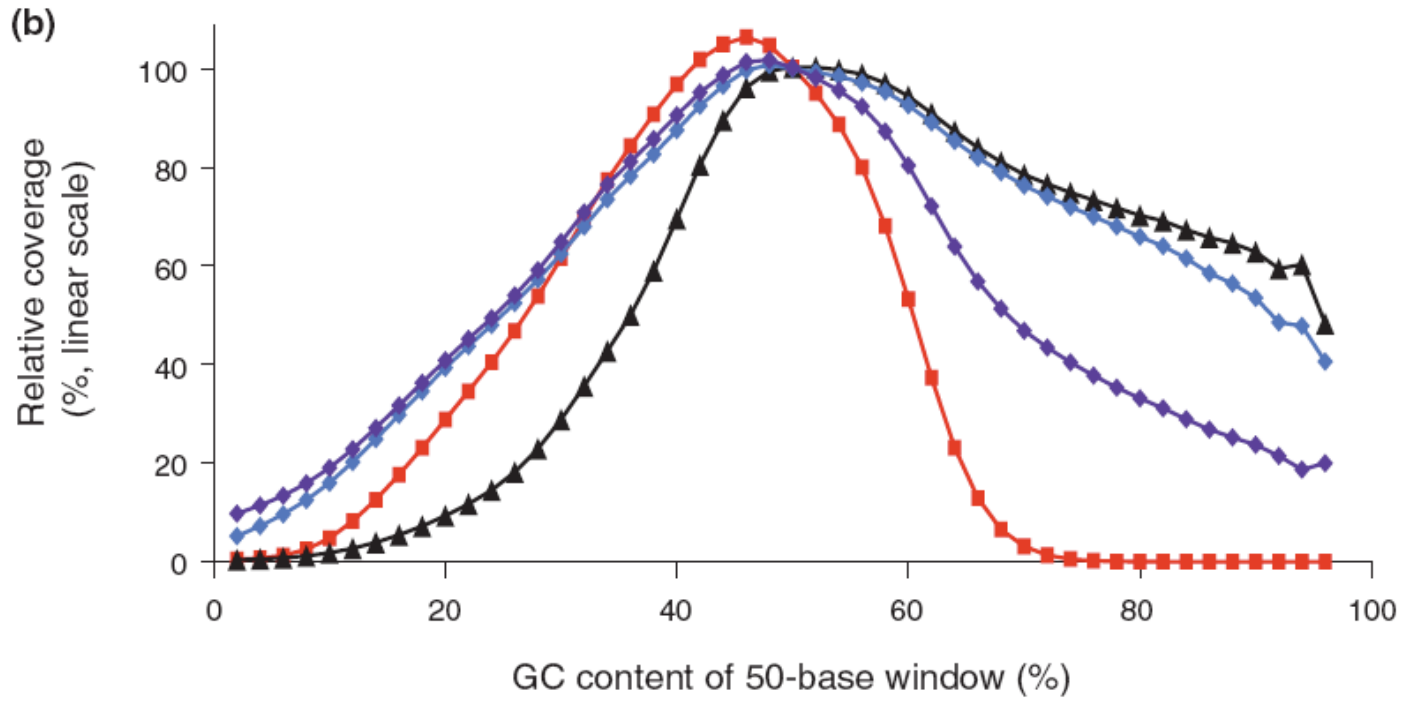
- Most libraries are PCR amplified during construction
- After library construction, single molecules are isolated and then amplified again for sequencing
- Errors from library construction PCR will not be detectable as sequencing errors
- Regions with secondary structure or extreme GC content:
  - Will amplify poorly and be underrepresented
  - May form small or weak clusters with poor sequence quality
- PCR may form chimeric sequences (especially in targeted designs)
- PCR amplification may result in duplicated sequences

PCR Errors: How Much PCR?

- You may be doing more PCR than you think
- Initial amplification of sample
- Targeting PCR
- Library amplification
- 100 rounds of PCR is equivalent to a 2 order of magnitude drop in polymerase accuracy

PCR-based artifacts: PCR bias

- Most PCR protocols work best for ~50% GC
- Extreme GC sequences are underrepresented



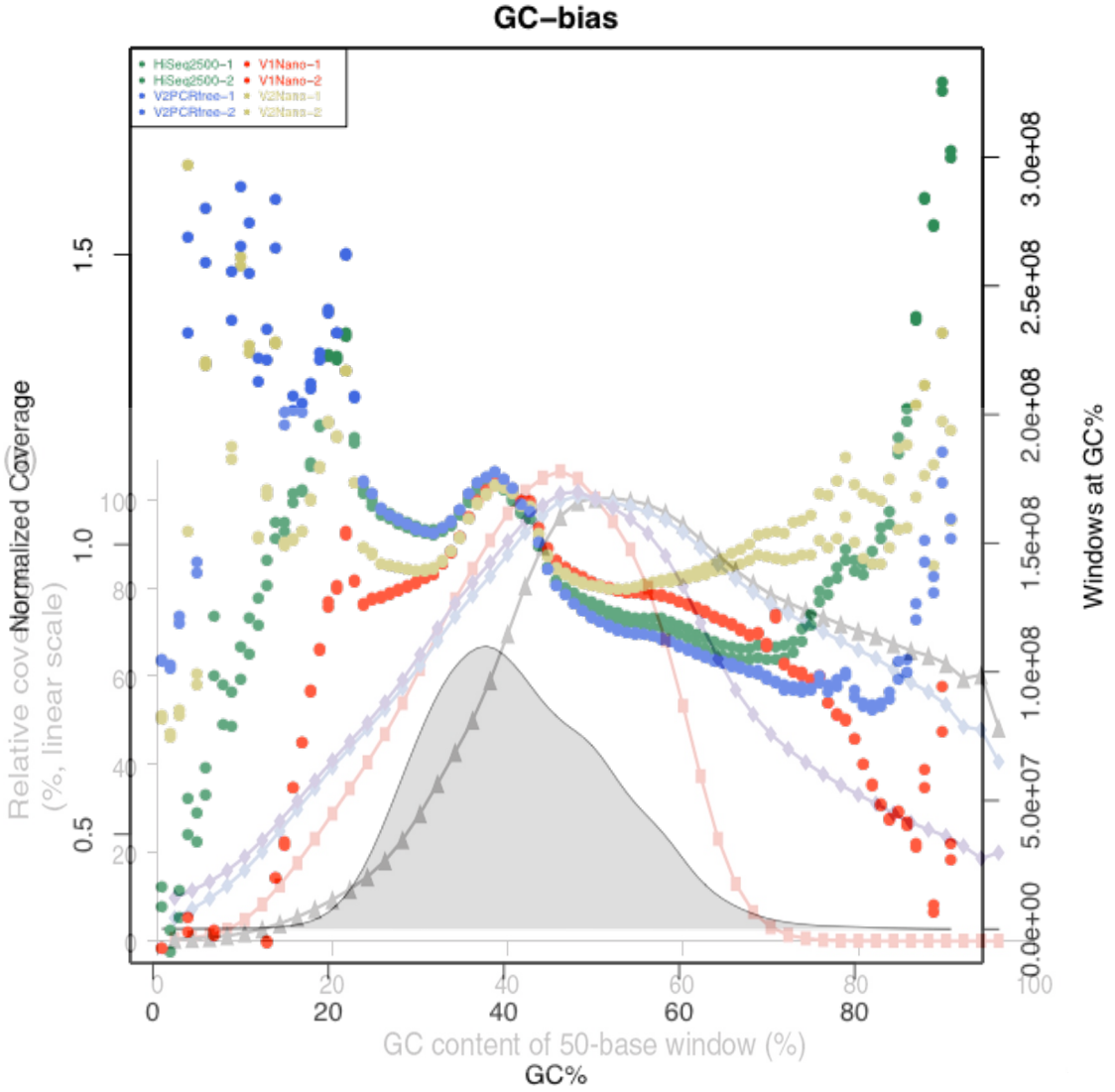
Red = standard PCR protocol
Other colors = modified PCR protocols

From Aird et al., Genome Biology *(2011)*

PCR-Free Libraries

- No PCR amplification in library construction
  - Not the same as no PCR, depending on other steps
- More uniform coverage by GC
- Fewer regions of 0 coverage
- Still some bias as cluster formation is PCR-like
- Requires more DNA (1-2 µg *vs*. 100-200 ng)

GC Bias on Modern Illumina



Sequencing of primers, adapters & tags

- Not every base you sequence is useful
- Primers will be present if you used PCR to target your input DNA
  - Sequence from primers does not represent target
  - Variation seen (or not) under primers is not real
  - Overlapping products will allow analysis of the primer-covered regions
- Short fragments may read through to adapter
- Custom barcodes or other tags may get sequenced too, though most vendor tags will be removed automatically