BIOL 525D - Bioinformatics for Evolutionary Biology

Location: MWF (4-6:30pm) - ANGU 437 (Henry Angus Building)
T/R (4-6:30pm) - LASR 107 (Frederic Lasserre Building)

Instructors: Sariel Hubner (sariel.hubner@botany.ubc.ca)
Evan Staton (statonse@biodiversity.ubc.ca)
Sam Yeaman (yeaman@zoology.ubc.ca)

Organizer: Loren Rieseberg (riesebe@mail.ubc.ca)

Website: https://github.com/UBCBio525/Bio525D

Course Objective:
Learn to manipulate and analyze large amounts of sequence data, with the goal of answering evolutionary or ecological questions.

Course Outline:

1 Broad introduction: Scope of course, goals, and VM setup [ALL]
2 Fastq files and quality checking/trimming [SAM]
3 Alignment: algorithms and tools [SARIEL]
4 Assembly: transcriptome and genome assembly [SARIEL]
5 RNAseq + differential expression analysis [SAM]
6 SNP and variant calling [SARIEL]
7 Population level statistics (sliding window analyses, nucleotide diversity, FST, SFS) [SAM]
8 Functional annotation and the use of ontologies [EVAN]
9 Phylogenetic inference [EVAN]
10 Among-species comparisons; conclusions [EVAN]

Overview of sequencing technology

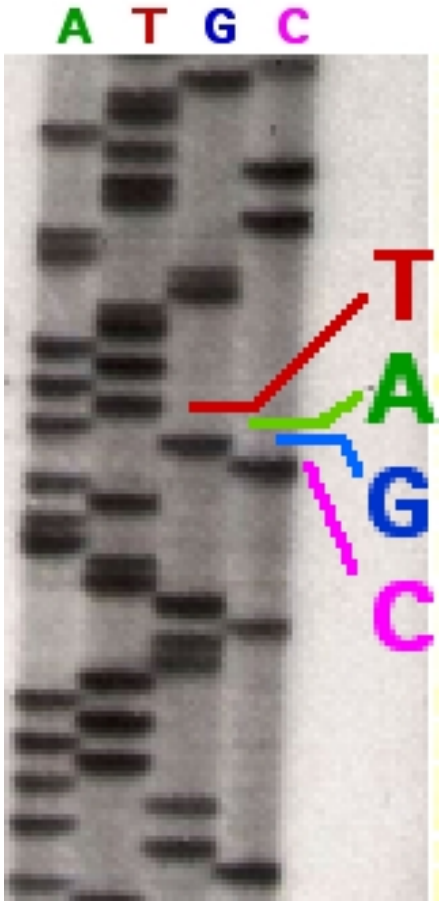
First Generation Sequencing Technology

Invented in the mid 1970s

Maxam-Gilbert Sequencing – chemical modification and cleavage paired with gel electrophoresis

Sanger Sequencing – dideoxy DNA sequencing paired with gel electrophoresis

- Became fully automated – dideoxy bases replaced with fluorescently labeled dideoxy bases
- And lasers and computers replace graduate students and postdocs
- Dominant sequencing method up until 2007 – 30 years!
- But only one fragment can be sequenced per Sanger reaction



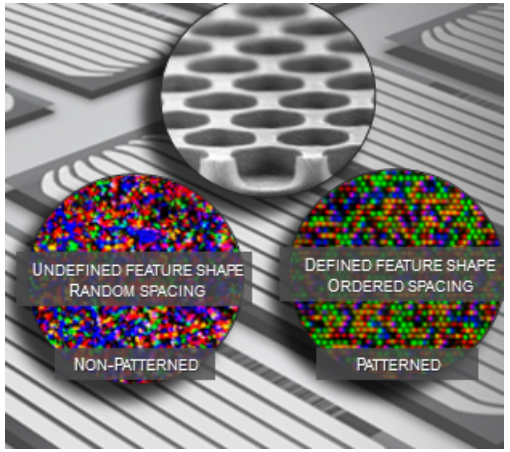
Sanger

Second generation sequencing technology

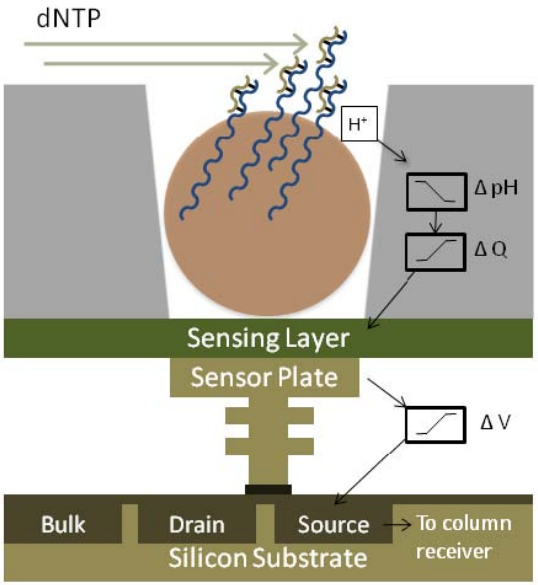
Developed to increase throughput

Can sequence many molecules in parallel

- DNA sequenced as clusters or in nanowells
- Single machine can sequence 3-10 Billion independent DNA fragments at same time
- Single Sanger Sequencer maxes out at 1152 reactions per machine

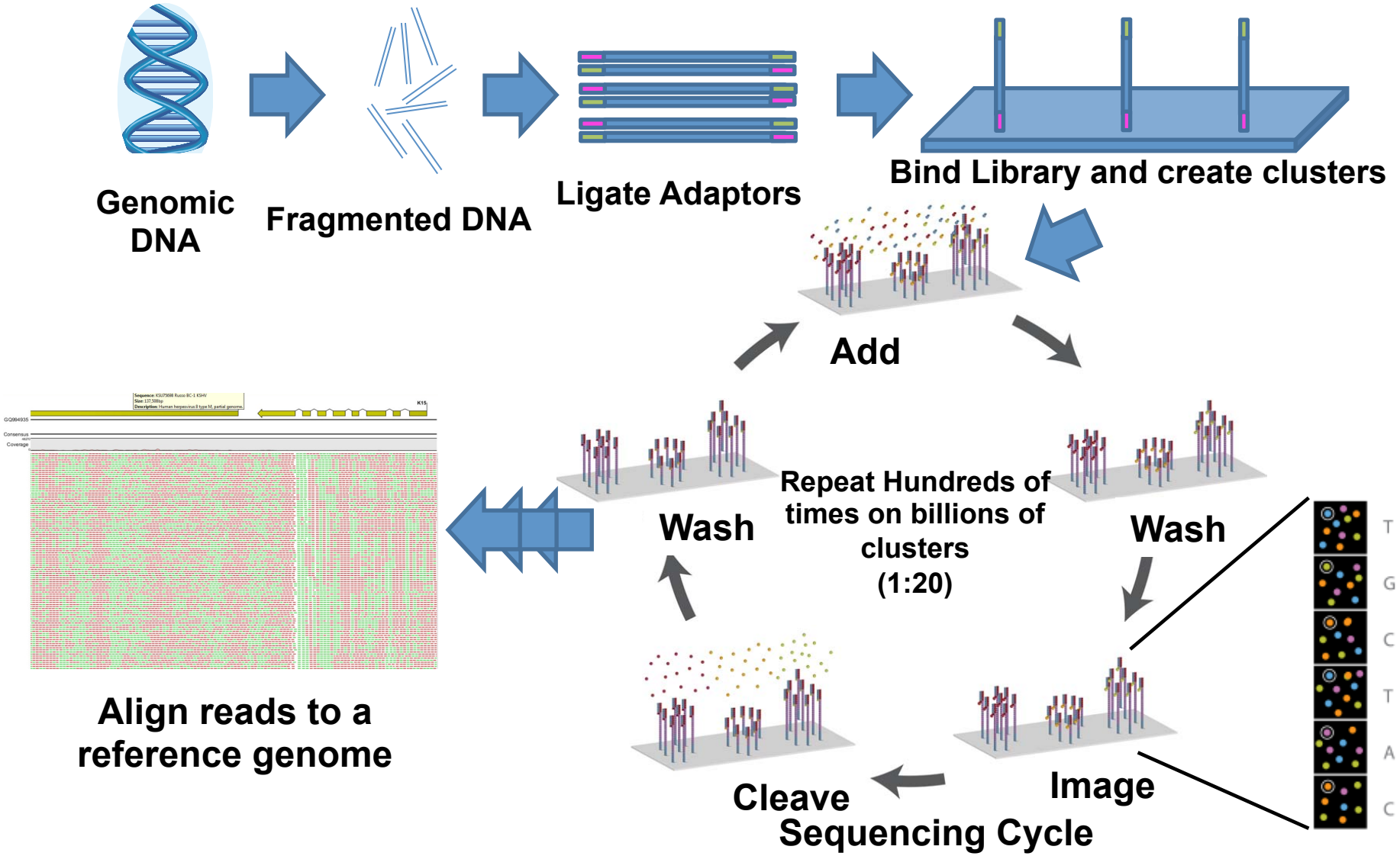


Illumina HiSeq (3-9 billion clusters – 600GB-1.8TB)



Ion Torrent Proton
(100 - 300 million nanowells - 20 - 60GB)

2nd Gen: Sequencing by Synthesis Overview



Flavors of Sequencing

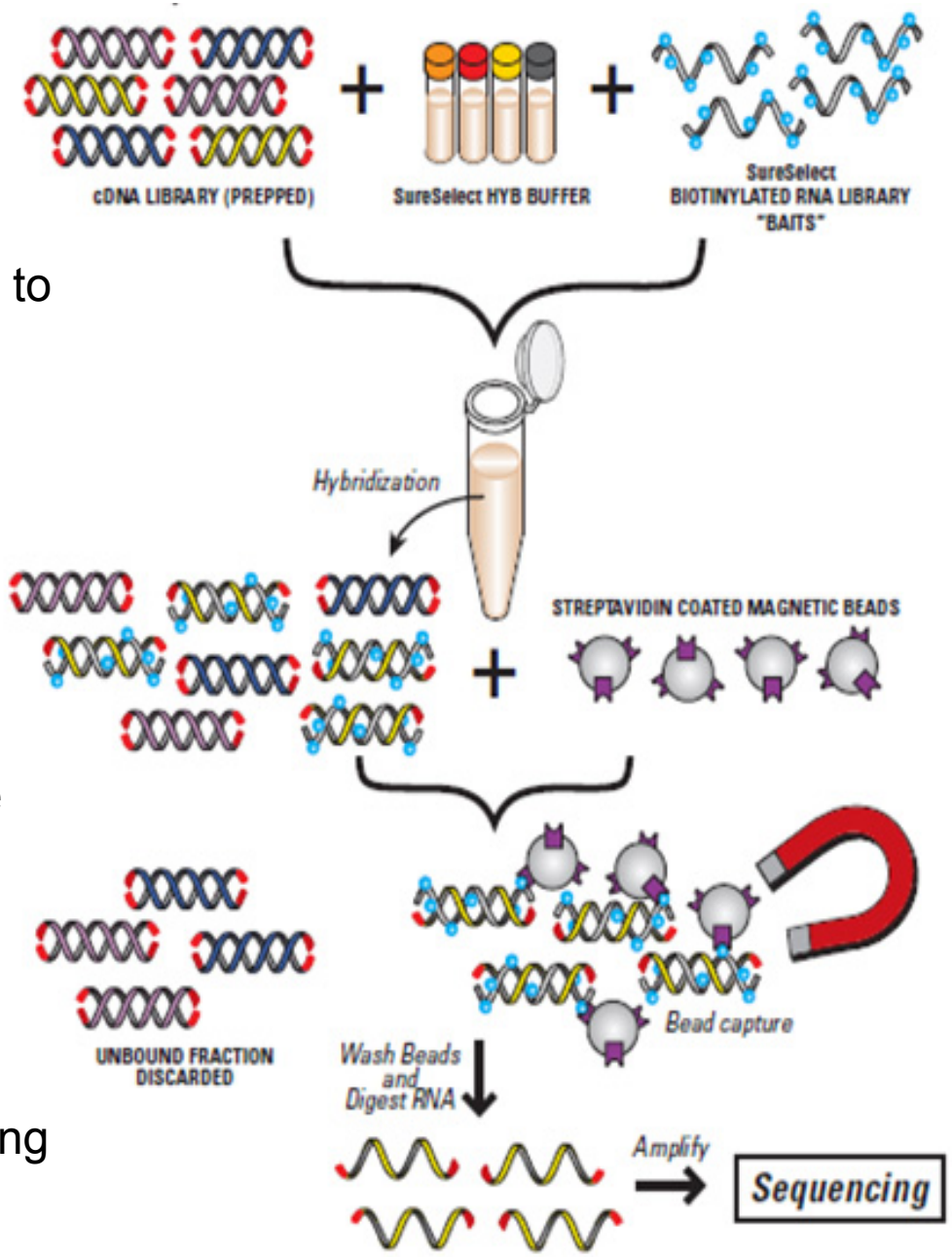
Whole Genome Sequencing
- Create sequencing libraries of all DNA fragments

Sequence Capture
- Attach complimentary RNA or DNA strands to beads
- Fish out desired DNA sequences
- Create sequencing libraries from enriched DNA
- Reduces cost and analysis time

RNASeq
- Extract RNA and convert to cDNA
- Create sequencing library from cDNA
- Permits analysis of expression & sequence variation in expressed genes

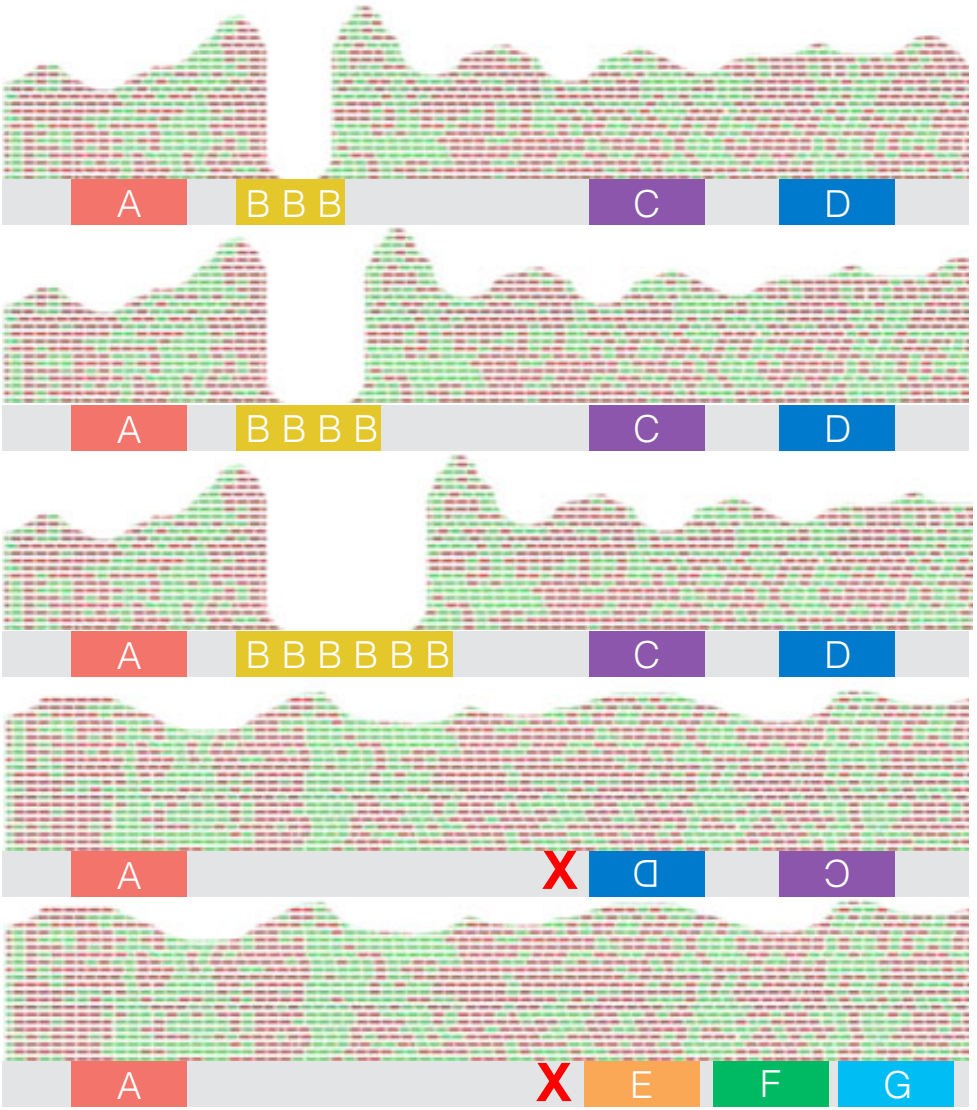
RadSeq or Genotyping by Sequencing
- Reduce genome complexity with REs
- Sequence highly multiplexed libraries of restriction fragments
- Simultaneous marker discovery & genotyping

Amplicon Sequencing
- Use PCR to amplify target DNA
- Sequence amplified DNA (Amplicon)



Disadvantages of 2nd Generation Tech

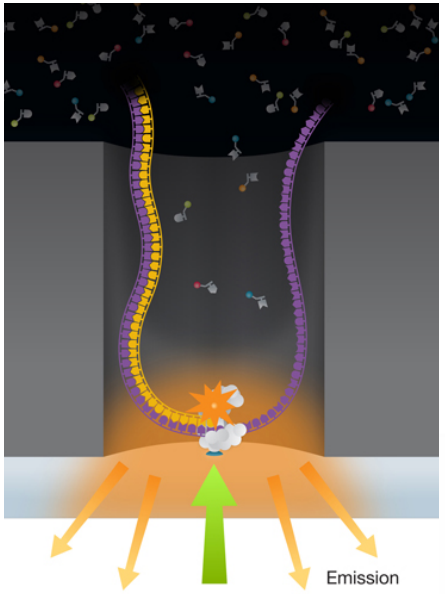
- Rely on amplification to create libraries and clusters
  - All polymerases have an inherent error rate (10^{-6}-10^{-7})
  - Errors introduced every 10 million to 100 million bases
  - Secondary validation of variants is key
- *De novo* genome assembly is challenging with short reads
  - Most 2nd Generation sequencers have a maximum read length of < 400bp
  - Too short to span long repeat regions
- Short reads can miss large structural variations
  - Translocations & inversions will be missed
  - Significant read depth at break points needed to detect variants
- Trouble detecting small insertions & deletions
- Short reads challenging to align



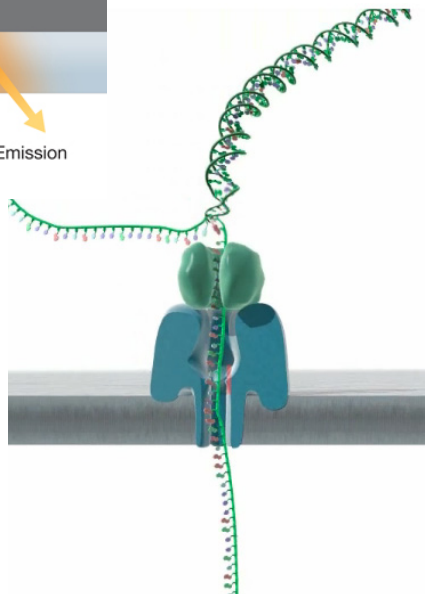
Very high quality single molecule long reads would fix many of these problems!

The Future: Third Generation Sequencing

- Single molecule sequencing
- Less complex sample prep and much longer read length (1-100kb)
- Two categories:
  1) Sequencing by synthesis
     - Pioneered by Pacific Biosciences
     - Microscopes and polymerase bound nanowells are used to WATCH DNA as it is sequenced in real time
     - Fluorescence of base detected at polymerase
  2) Direct sequencing by passing DNA through a nanopore
     - Bases fed through a membrane bound nanopore
     - Detect how ion flow changes at the pore as each base passes through
- Technical hurdles
  - High error rates (10-25%)
  - Expensive (3-10x more than Illumina)

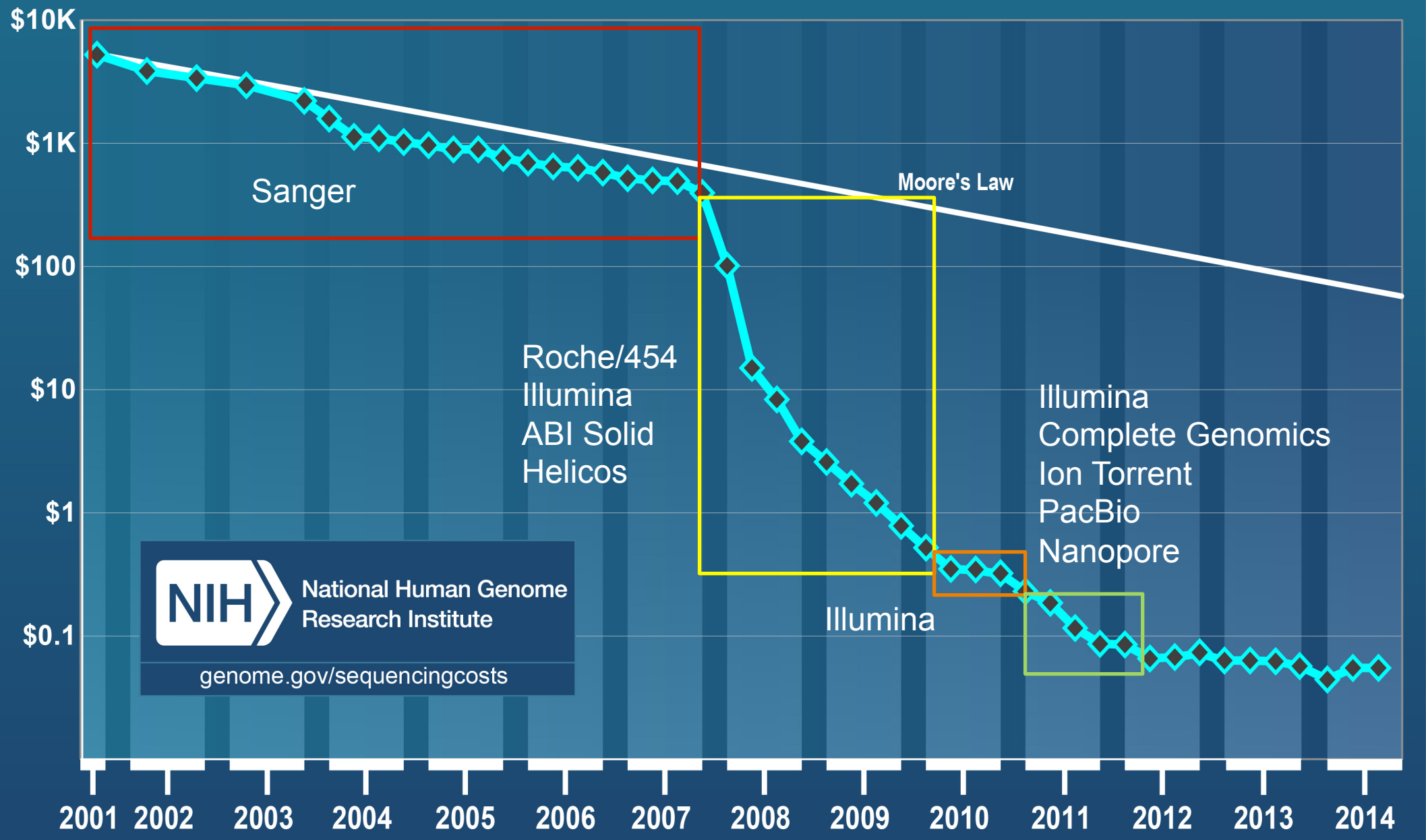


PacBio

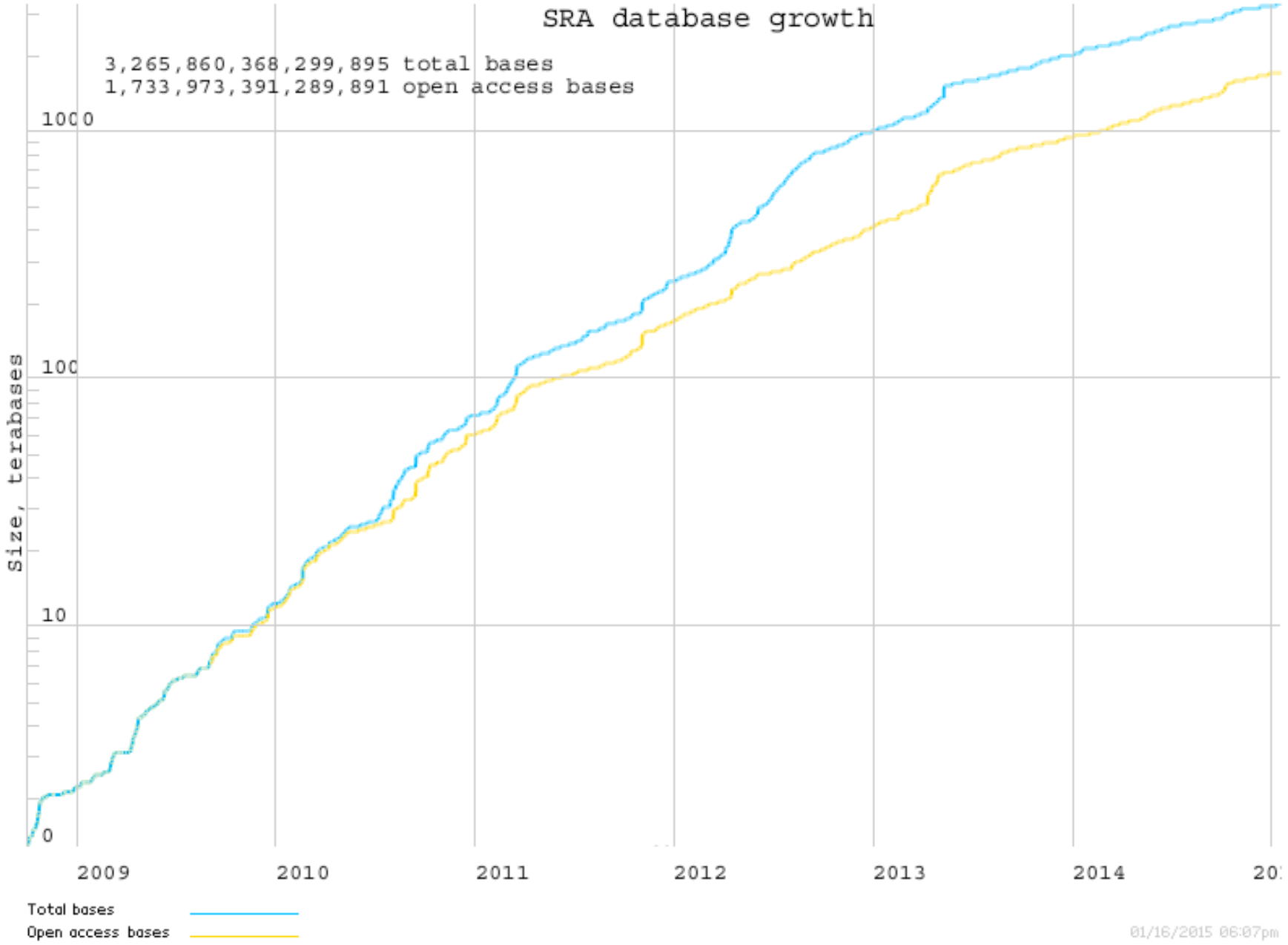


Oxford Nanopore

Cost per Raw Megabase of DNA Sequence



Sequence read archive (SRA)



Drowned in next generation sequencing data