Genetics 211 - 2018
Lecture 1

Genome Sequencing

Gavin Sherlock

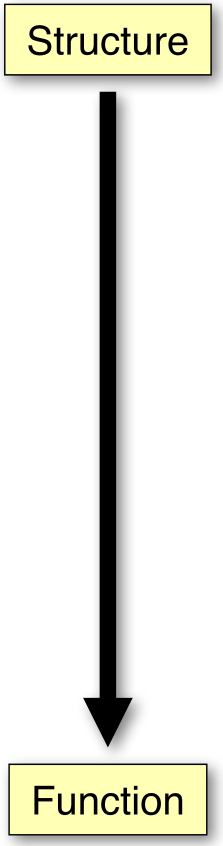
gsherloc@stanford.edu

January 9th 2018

Overview of My Lectures

- Genome Sequencing (Lecture 1)
  - Sanger Sequencing
    - Whole Genome Sequencing
    - Sequencing Theory
    - Genome Assembly
  - High Throughput Sequencing Technologies
    - Illumina
    - PacBio
    - Oxford Nanopore
- Short Read Genome (Re)sequencing (Lecture 2)
  - Making DNA sequence libraries
  - Data formats
  - Read alignment
  - Variant calling
  - *De novo* assembly from short reads
  - Gaining longer contiguity information
- Functional Genomics (Lecture 3)
  - Chromatin state
  - ChIP-Seq and Transcription factor binding sites
- Expression
  - RNA-Seq
  - Cluster Analysis

What to Sequence and Why?



- ***De novo* whole genome sequencing**
  - requires *de novo* whole genome assembly

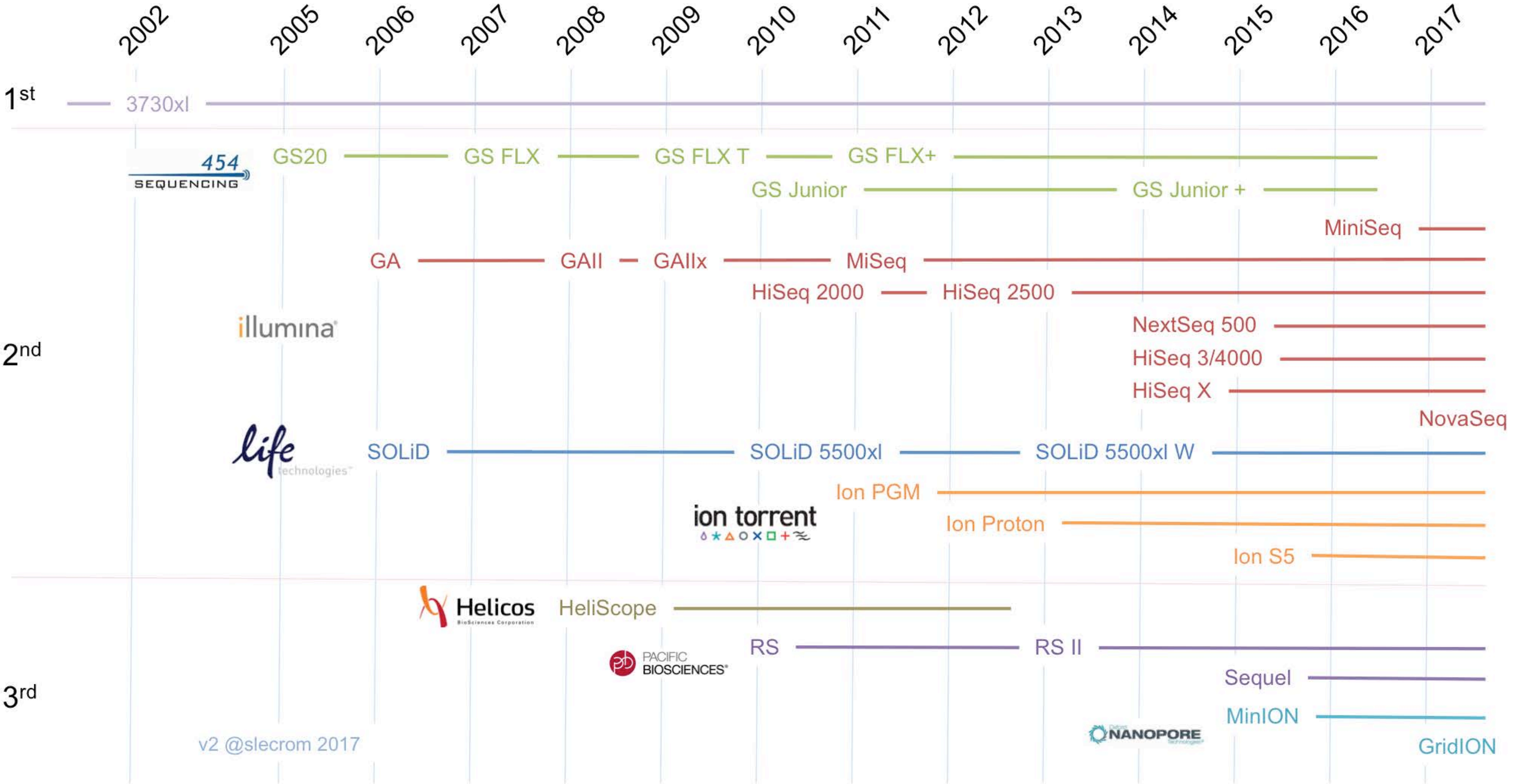
- **Polymorphism discovery** (distinct from genotyping)
  - Targeted approaches (exome)
  - Whole genome
  - SNPs, copy number variations, insertions, deletions, etc.

- **Expressed sequence discovery and functional genomics**
  - Expression profiling/RNA-Seq
  - ChIP
  - Nucleosome positioning
  - RNA editing
  - Hi-C
  - etc.

Four Fundamentally Different Approaches to DNA Sequencing

- Chemical degradation of DNA
  - Maxam-Gilbert
  - obsolete
- Sequencing by synthesis (“SBS”)
  - uses DNA polymerase in a primer extension reaction
  - most common approach
  - Sanger developed it (“Sanger sequencing”)
  - Illumina, Pacific Biosciences, Ion Torrent, 454
- Ligation-based
  - sequencing using short probes that hybridize to the template
  - SOLiD, Complete Genomics
- Nanopore
  - Inferring sequence by change in electrical current as ssDNA is pulled though a nanopore
  - Oxford Nanopore, NABsys, Genia, Illumina

Commercially Available Sequencers Timeline



2002 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017

1st

3730xl

2nd

454 SEQUENCING

GS20 — GS FLX — GS FLX T — GS FLX+

GS Junior — GS Junior +

illumina

MiniSeq

GA — GAII — GAIIx — MiSeq

HiSeq 2000 — HiSeq 2500

NextSeq 500

HiSeq 3/4000

HiSeq X

NovaSeq

life technologies

SOLiD — SOLiD 5500xl — SOLiD 5500xl W

ion torrent

Ion PGM

Ion Proton

Ion S5

3rd

Helicos BioSciences Corporation

HeliScope

PACIFIC BIOSCIENCES

RS — RS II

Sequel

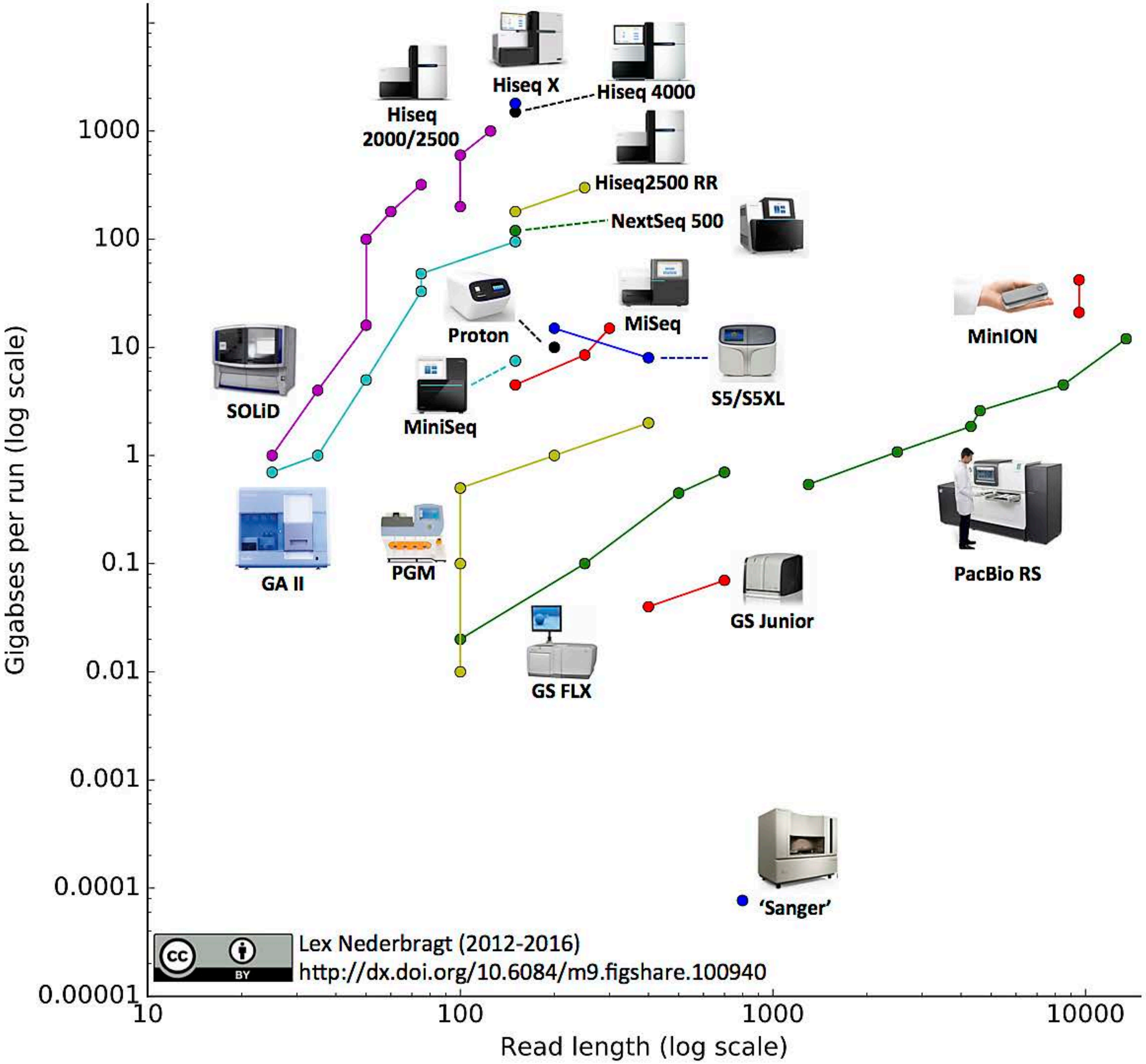
Oxford NANOPORE Technologies

MinION

GridION

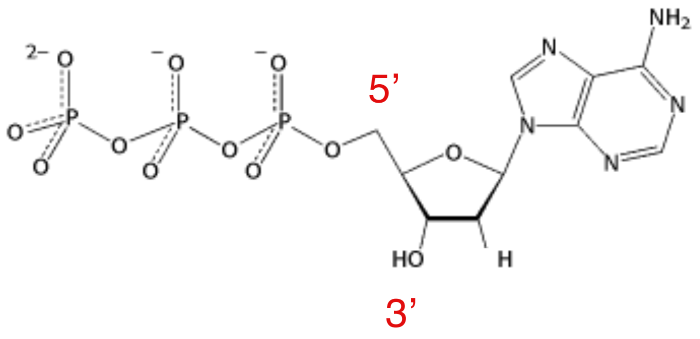
v2 @slecrom 2017

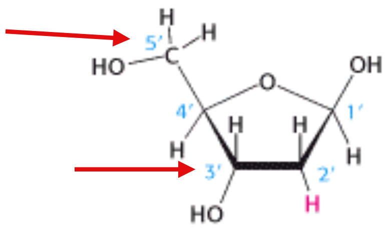
Developments in Sequencing



Lex Nederbragt (2012-2016)
http://dx.doi.org/10.6084/m9.figshare.100940

5’ and 3’



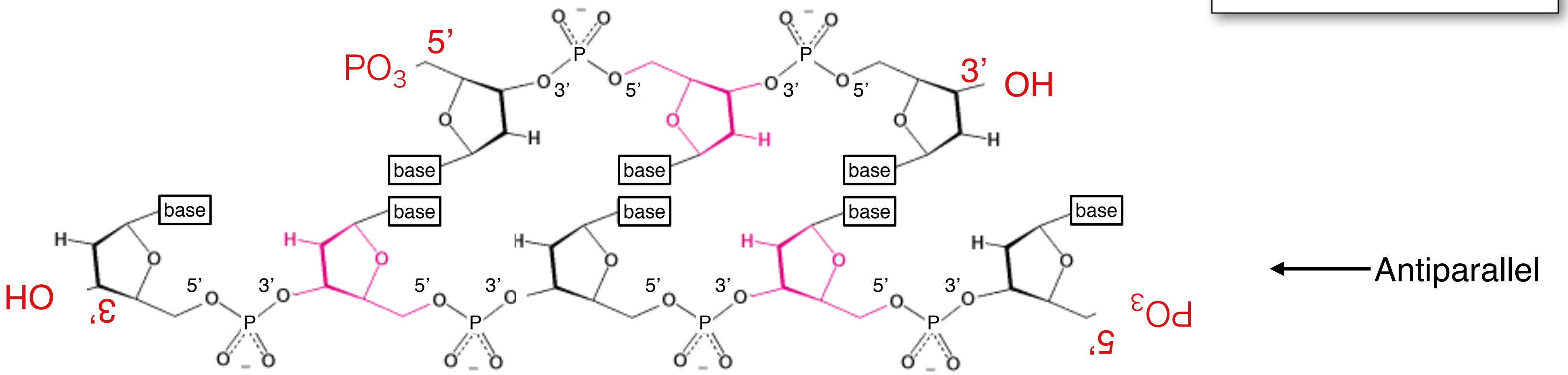


Deoxyribose

Base	plus sugar “nucleoside”
Adenine	Adenosine
Guanine	Guanosine
Cytosine	Cytidine
Thymine	Thymidine

in DNA: “deoxyadenosine”

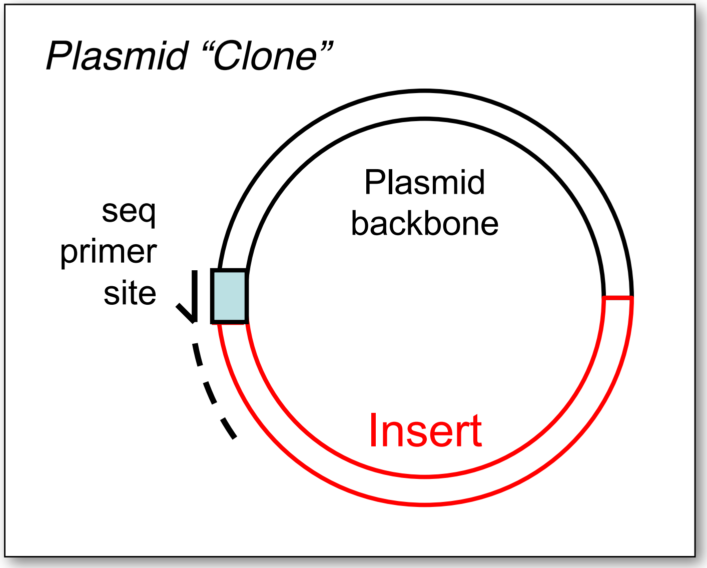
plus triphosphate
“deoxynucleotide”
“2’-deoxyadenosine 5’-triphosphate” = dATP

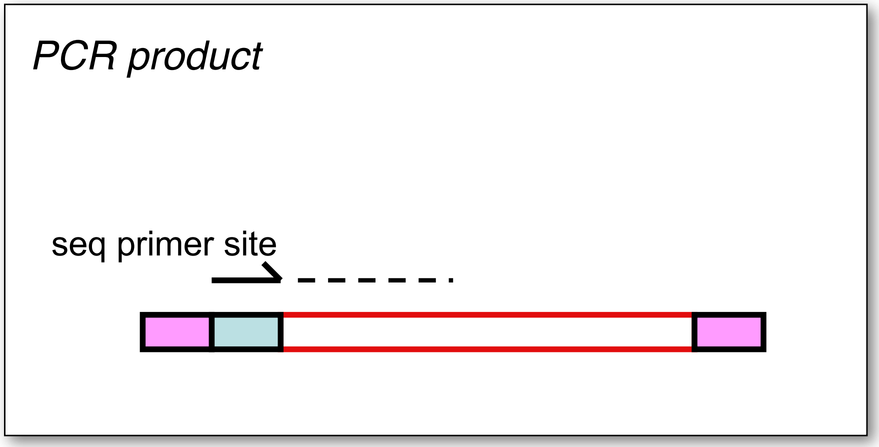


If I throw in DNA polymerase and free nucleotide, which end gets extended?

Adapted From Berg et al: Biochemistry 5th ed. Freeman+Co, 2002

Sanger Sequencing Templates





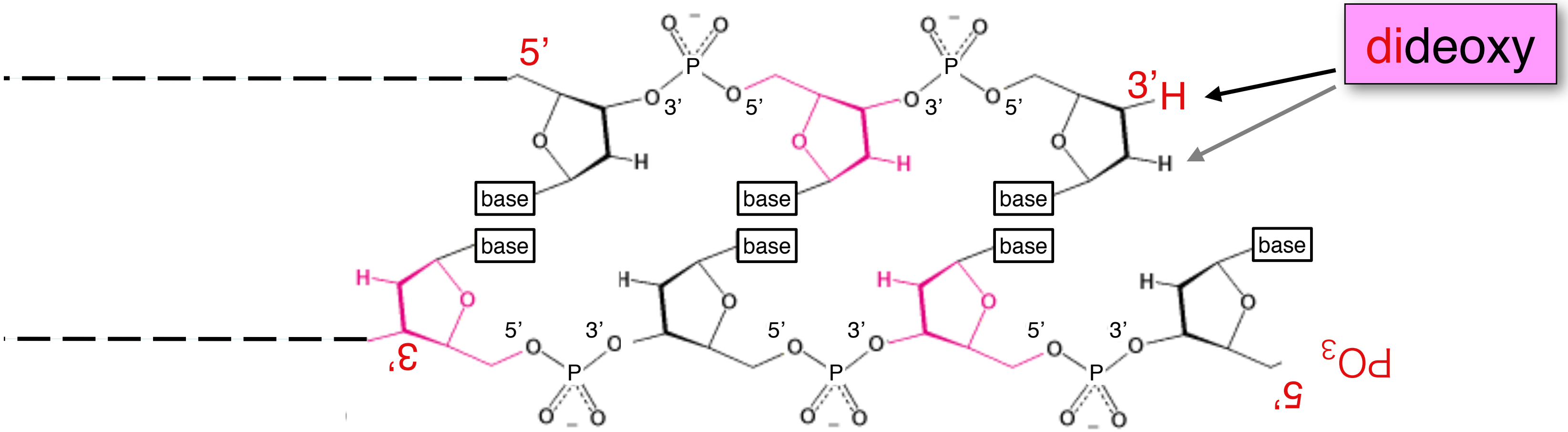
```
Watson  5′ .. T A G C G T C A G C T .. 3′
Crick   3′ .. A T C G C A G T C G A .. 5′

     5′Primer T A G C G 3′------------▸
      3′    .. A T C G C A G T C G A C .. 5′
```

In Sanger sequencing, Crick is the template and Watson’s synthesis starts at the primer’s 3’OH

The Chain Terminator

- Dideoxy nucleotides cannot be further extended, and so terminate the sequence chain



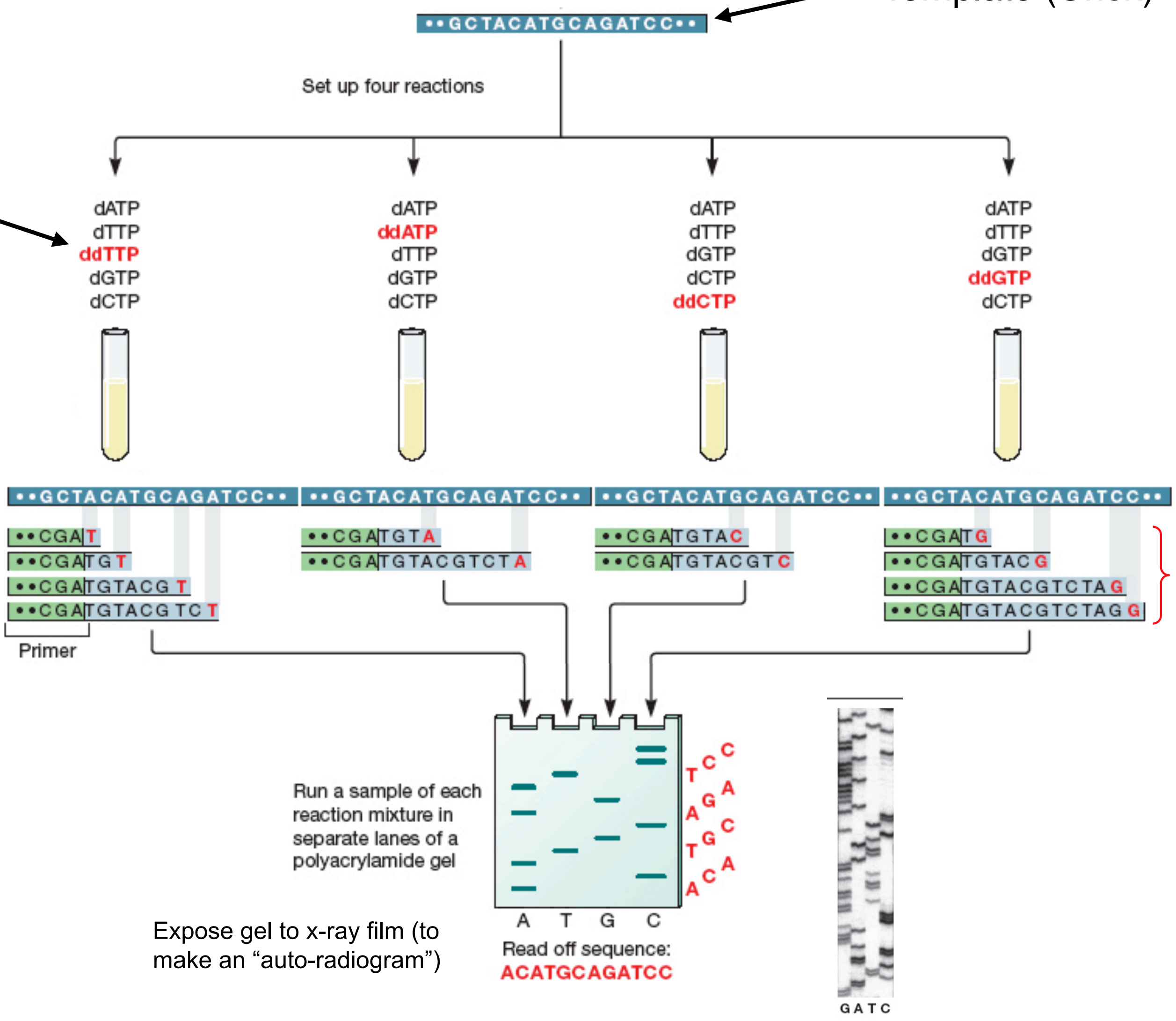
Adapted From Berg et al: Biochemistry 5th ed. Freeman+Co, 2002

Original Sanger Sequencing with Radioactive Signal

Template (Crick)

very low concentration of ddNTPs compared to dNTPs

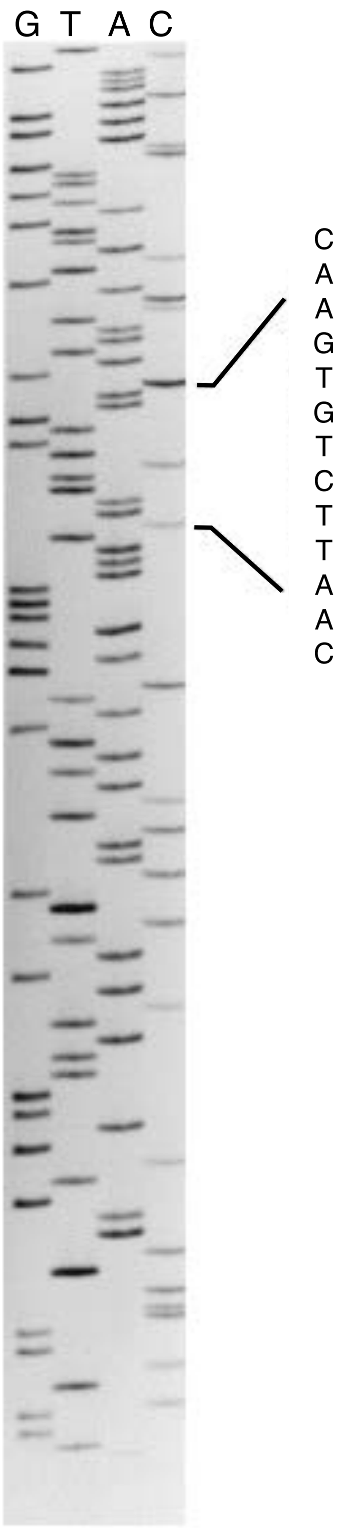
Watsons



Expose gel to x-ray film (to make an “auto-radiogram”)

A nested series of DNA fragments ending in the base specified by the terminator-ddNTP

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



This is great, but…

Wouldn’t it be great to run everything in one lane?

- Save space and time, more efficient

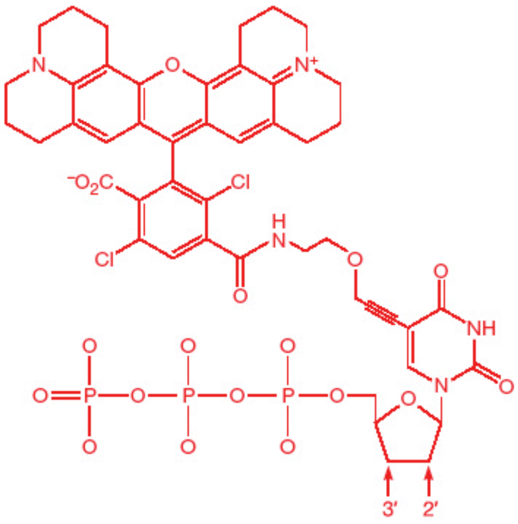
Also, would be nice to read everything at the same point in the gel

- Unable to read sequence near the top, as the bands get closer and closer together.

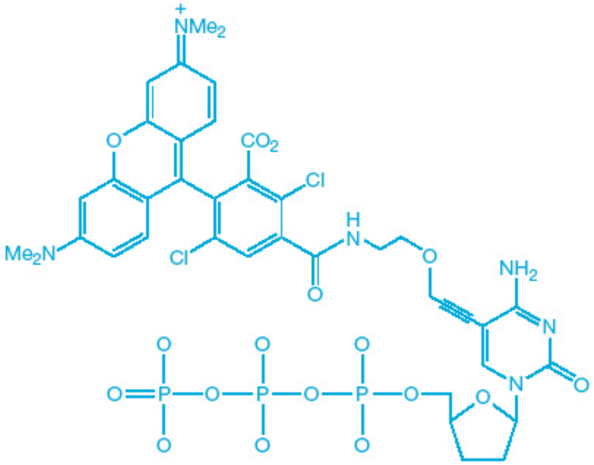
Fluorescently label the ddNTPs so that they each appear a different color, and can be read by a laser at a fixed point

Fluorescent Sanger Sequencing: “Dye-terminators”

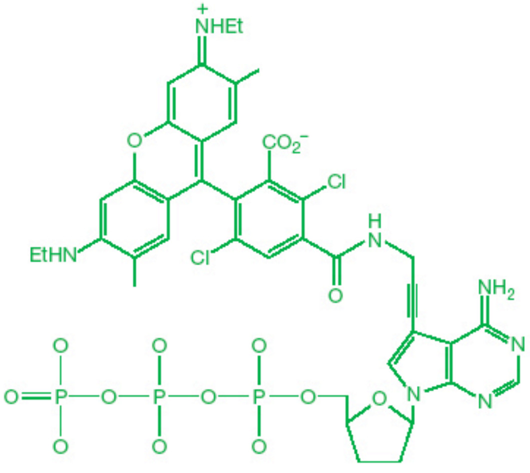
Each of the 4 ddNTPs is labeled with a different fluorescent dye (instead of radioactivity)



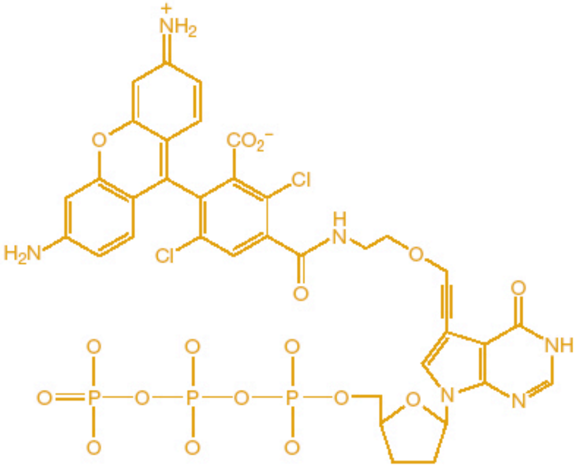
ddTTP



ddCTP



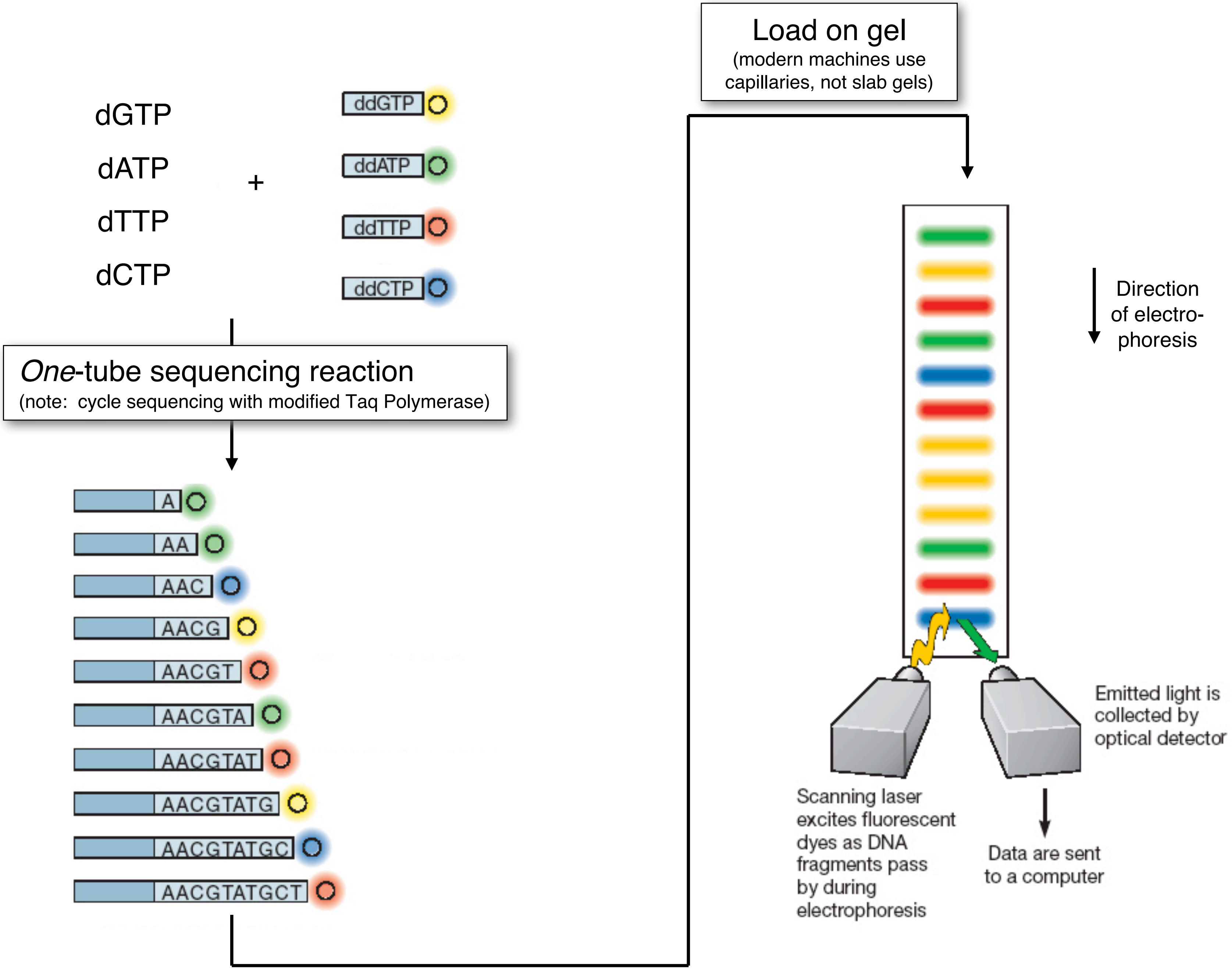
ddATP



ddGTP

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

Fluorescent Sanger Sequencing



One-tube sequencing reaction
(note: cycle sequencing with modified Taq Polymerase)

Load on gel
(modern machines use capillaries, not slab gels)

Direction of electro-phoresis

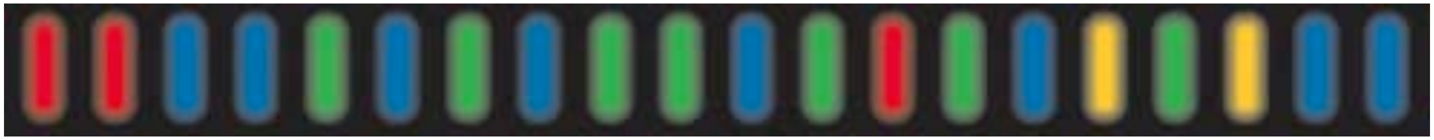
Emitted light is collected by optical detector

Scanning laser excites fluorescent dyes as DNA fragments pass by during electrophoresis

Data are sent to a computer

Fluorescent Sanger Sequencing Trace

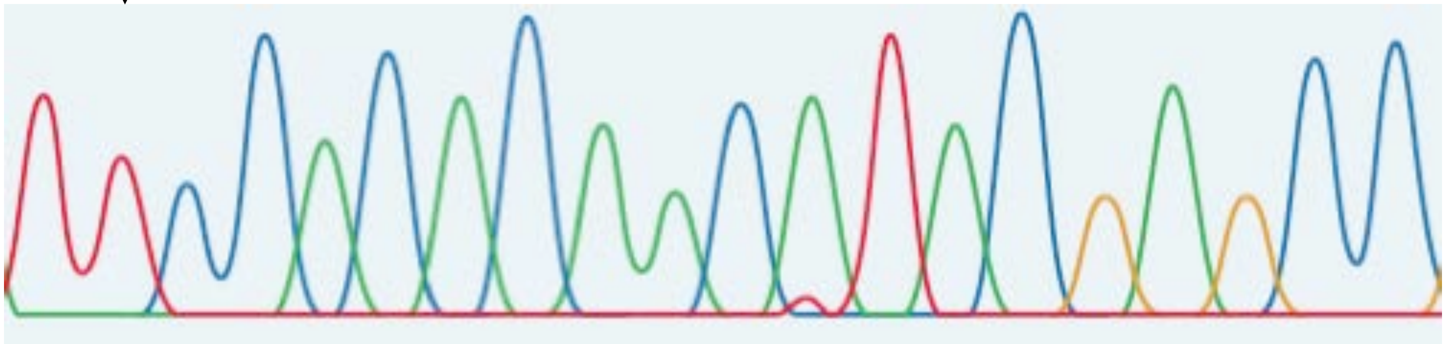
Lane signal



(Real fluorescent signals from a lane/capillary are much uglier than this).

Various algorithms to boost signal/noise, correct for dye-effects, mobility differences, etc., generates the ‘final’ trace (for each capillary of the run)

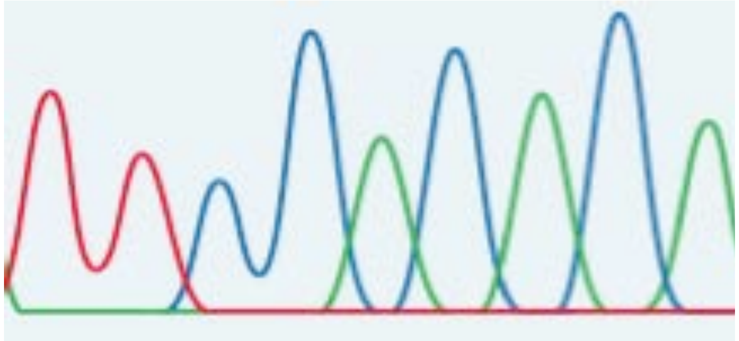
Trace



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

Sanger Base Calling

Low quality over here…



…low quality over there

Base Caller (Phred)

```
... 44 45 46 47 48 49 50 51 52 53 54 55 ... 718 719 720 ...
...  N  A  G  C  G  T  T  C  C  G  C  G ...   A   N   N ...

...  0  3 20 25 40 88 95 99 99 99 99 99 ...  10   0   0 ...
```

Quality score = -10 * log(probability of error) or P=10^-Q/10
For Q20, probability of error = 1/100
For Q99, probability of error ~10^{-10}

Phred: *The* base-calling program

- Algorithm based on ideas about what might go wrong in a sequencing reaction and in electrophoresis

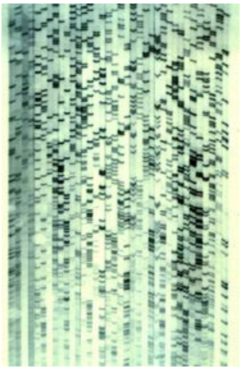
- Tested the algorithm on a huge dataset of “gold standard” sequences (finished human and *C. elegans* sequences generated by highly-redundant sequencing)

- Compared the results of phred with the ABI Basecaller

- Phred was considerably more accurate (40-50% fewer errors), particularly for indels and particularly for the higher quality sequences

(Ewing et al., 1998, *Genome Research* **8**: 175-185; Ewing and Green 1998, *Genome Research* **8**: 186-194)

Progress of Sanger Sequencing Technology



Radioactive polyacrylamide slab gel

Low throughput,
labor intensive

AB slab gel sequencers (370, 373, 377)

Fluorescent sequencing
1990-1999
6 runs/day
96 reads/run
500 bp/read
288,000 bp/day

AB capillary sequencers (3700, 3730)

1998-now
24 runs/day
96 reads/run
550 – 1,000 bp/read
1-2 million bp/day

~1,000-fold increase in throughput since 1985 accomplished by incremental improvements of the same underlying technology

2nd Generation Sequencing Technologies have up to 1e6x more throughput than 3730

Whole Genome Sequencing

- Two main challenges:
  - Getting sufficient “coverage” of the genome
    - A function of read length, number of reads, complexity of library, and size of genome
  - Assembling the sequence reads into a complete genome
    - A function of coverage, and repeat size (relative to read lengths) and repeat frequency

How much sequence do you need?

- Let L = read Length; G = Genome size.
- Assume $L \ll G$.
- $P_{\text{obs\_with\_a\_given\_read}} = L/G$
- $P_{\text{not\_obs\_with\_a\_given\_read}} = 1\text{-}L/G$
- $P_{\text{not\_obs\_with\_N\_reads}} = (1\text{-}L/G)^N$
- $P_{\text{covered\_by\_at\_least\_one\_read}} = 1 - (1\text{-}L/G)^N$
- Rearranging gives: $N = \ln(1\text{-}P)/\ln(1\text{-}L/G)$

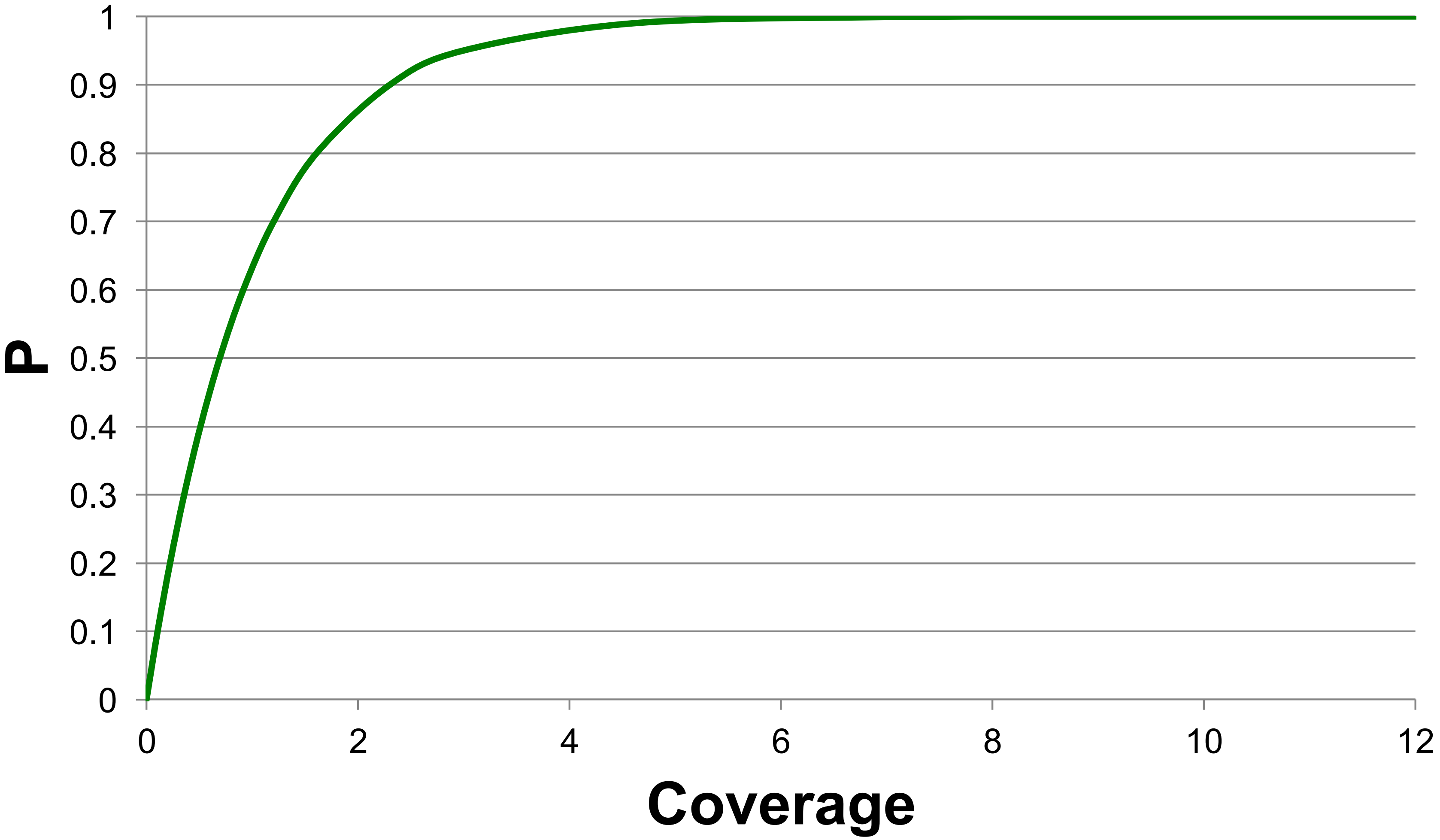
Example Calculation, Sanger Sequencing

- *E. coli* genome G = 4.6Mb, read length L = 800bp
- How many reads do I need to have a certain probability of observing any particular piece of my genome?
- Remember $N = \ln(1-P)/\ln(1-L/G)$
- P = 0.9 => ~13,000 ~2.3x coverage
- P = 0.95 => ~17,000 ~3x coverage
- P = 0.99 => ~26,500 ~4.6x coverage

Back of the Envelope

- Remember, $P = 1 - (1-L/G)^N$
- Given $(1-L/G)^N \approx \mathrm{e}^{-NL/G}$
- And, coverage, $R = NL/G$
- Then, $P \approx 1-\mathrm{e}^{-R}$
- This is a widespread back of the envelope calculation for any project involving redundancy.

Probability as a Function of Coverage

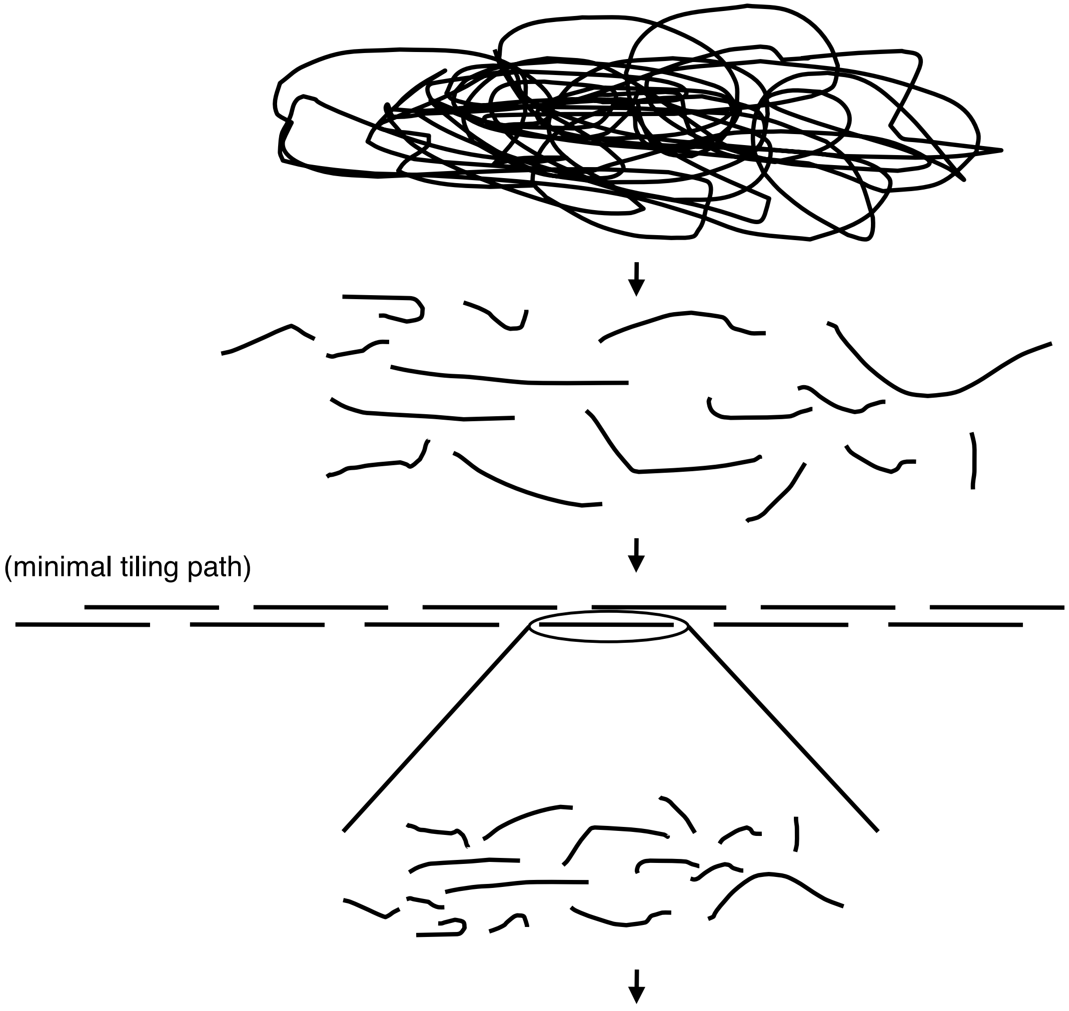


Overcoming repeats

- Most problematic when:
  - Repeats are longer than read lengths
  - Repeats are present in many copies
- Recognize based on coverage
- Resolve with longer range continuity information:
  - Paired-end reads
  - Multiple insert size libraries
    - Plasmids
    - Fosmids
    - BAC ends
    - Other tricks (which I'll come to later)

Whole Genome Sequencing Approaches

Hierarchical Shotgun Approach



Genomic DNA

BAC library

Organized, Mapped Large
Clone Contigs

Shotgun Clones

GCAATGAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGG
AGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

Reads

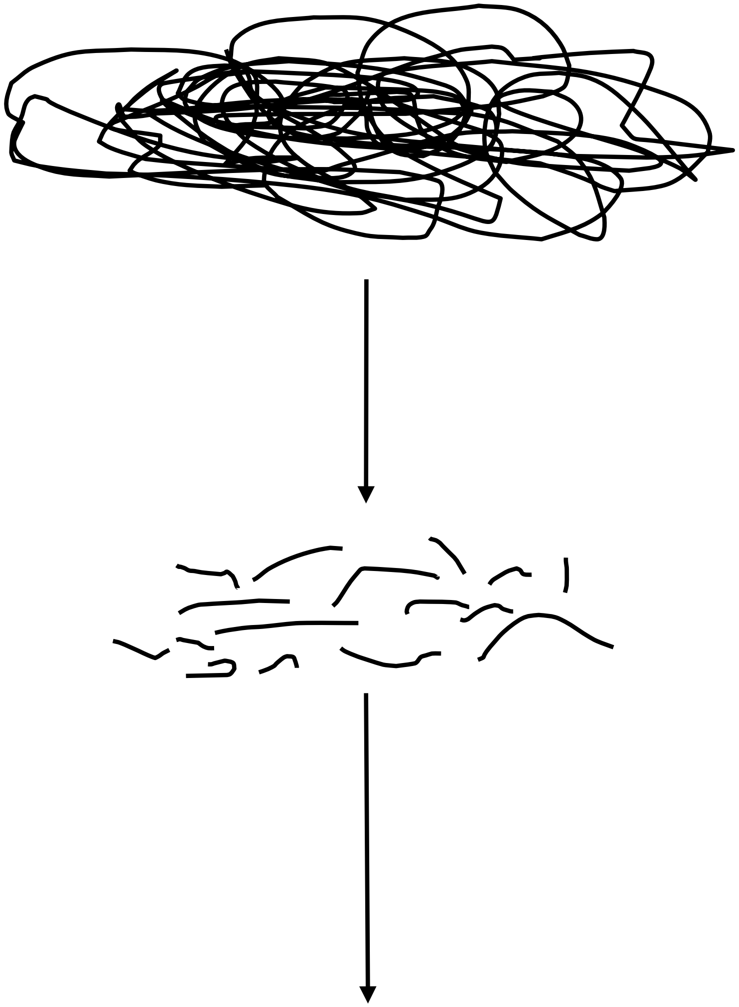
↓

TATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

Assembly

Whole Genome Sequencing Approaches

Shotgun Approach



Genomic DNA

Shotgun Clones

Reads

GCAATGAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGG

AGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

TATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATGGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAG

Assembly

Rationale for Hierarchical Strategy

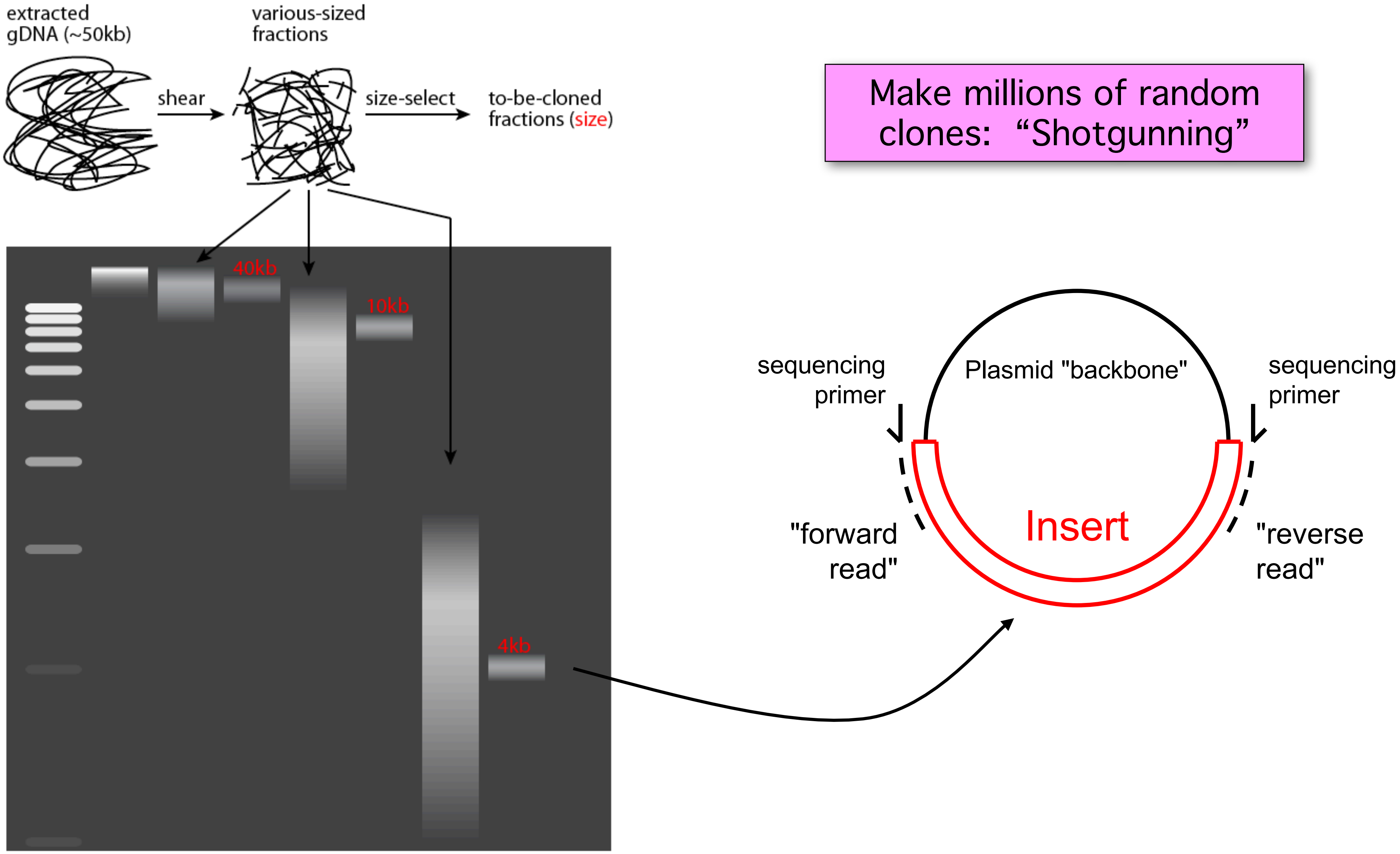
- Better for a repeat-rich genome
  - *less misassembly of finished genome*
    - *long-range misassembly largely eliminated and short-range reduced*
- Better for an outbred organism
  - *each clone from an individual and no polymorphisms in the final sequence.*
  - *(Added bonus: get SNPs from regions of overlapping clones)*
  - *Can also get some haplotype information, if individual BACs shotgun sequenced.*
- Better if there are cloning biases
  - *use minimum tiling path,so the same coverage for each region*
- Easier to identify and fill gaps (from unclonable regions) sooner

BUT

- Time consuming and expensive to make minimum tiling path

De Novo Whole Genome Sequencing

Make millions of random clones: “Shotgunning”

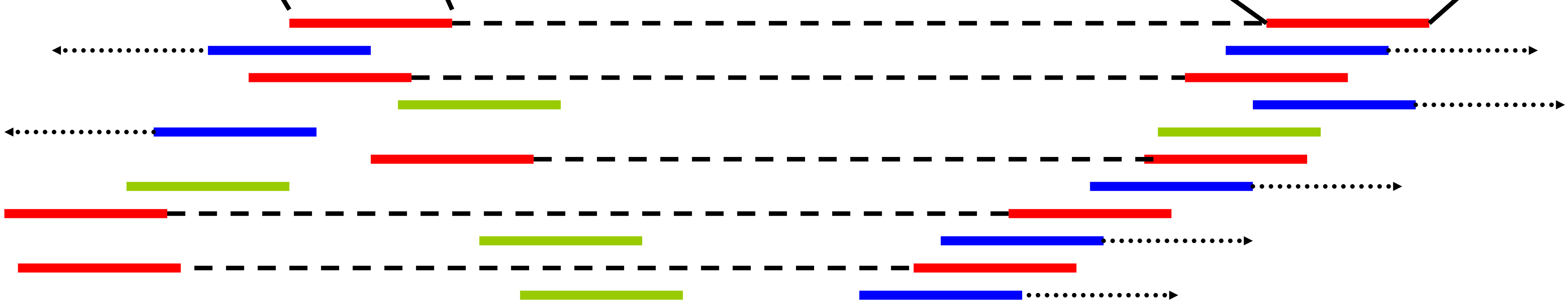


Sequencing Read

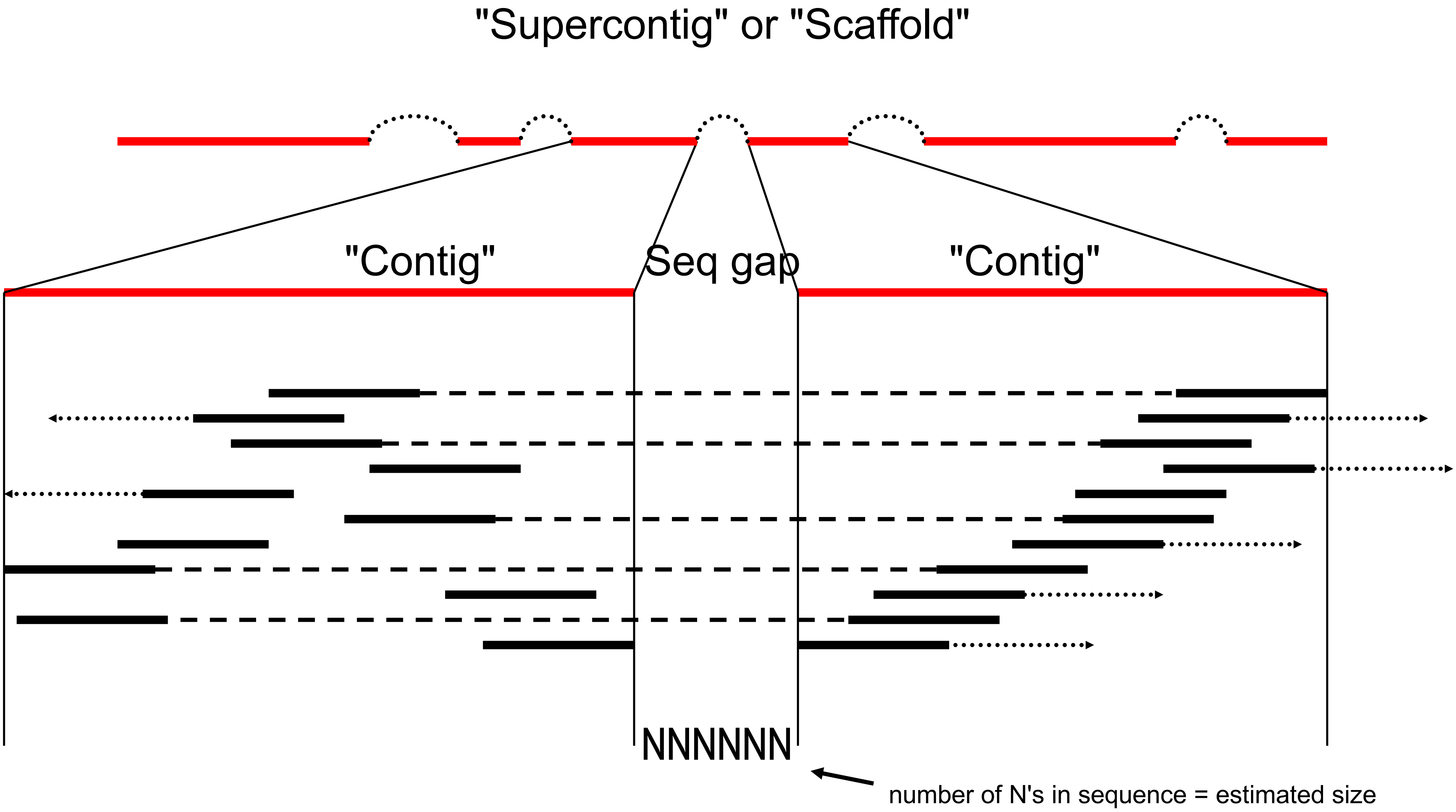
```
GCAATGAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTATG
GATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAGAAACCCCCAAGCTAGGATTTCTGCA
GCTCATGAAGCCTTGGAGATAAATGAGTAAGTGGGGGAAAATCTTGCTGTTAAAAAGGAAATCTCATCCTTTGCTGAATATATT
CAGTTGCCATTGATAGGATACTTAAATTAAACTGCATTTGAACTGGAGGATTATTTGGGGAGTTATTACTCTATTTAAAAAAGT
TTTTTTTTTAAATGAAGGACAGCCACCATGTGGAGGTGGTTTTAGTCATTTTATGAATTCAATGGCTTTGCTGTGATCCTAAAT
TAATTTCTTGAAGGGCTATCCCTAGGATATTGTGAGGATATAAAATAAATACAATTCTTTACATATCTAAAACATTCTGACAGG
GAAAATTTTCCAGATGTAGAATGCTCATCTGCACTAGAACATTTTCTAGTAGAACTTCTGCTAGTGGGGAAAACATGATAACAA
CATAAGGTTTAAAAAAAAAATTTTAGAAAATACTTCAAGATTAAGACAAAGATAAGAGGAAATGCTGTCTTGAGTGTTGTTAAA
CATTCTGTGGGTTACCAAGGAAGGCTGGGAAATCTCTTCTGGAGATCTCAGAAAATGAGAAAGATTCTTAAAGTTGGAGTCATA
AAAACTCAGGGTTGGCAGAGACCTTAAAGGTCACTTAGCTGAACCACCCATCTGGTGCTTGAATCACCTCAACACTATCCTTGC
CAAGTGGTCATTGTTAAACTATTTTATGATTTTTCTGAAGAAGGTTACAGAATCTTCTTCAGAGATCTTAGGGAAAAAAAAAAA
AGATTGTCGTGAGAGTTGAAAATCCTGCCATTGTAACCAGTTGATCTACGGTTTCTGATTCTGTCATGCAACATATTTATTTTC
CAGTTTCTTGTCATCTACAAATTCGATATGCCTGCCTTCTGTGTGTCATCCATATTTCTGAGAAAAATATGAAGGCCAGGAATA
GAGCCCTGTGACATGACATAGAAACTACCCTCCAGGTTCATGTCTTCATGAATCACCATCTTTTGTATTGTTCACTCAATTACT
AAGCCACCCAGTTACACTGTGACTCAGCTCATATTTCTCCATTTGGATCTTAAGAATGCCAATCGTAGCTGCGGATCTTAAATT
TATAGTAAATCTATTACAGTAAATTAAGCTAGCACAATCTGATTTATTTATTCTTAGTGAATATAAGCTGGCTTCTAGTCGTCA
CTACTTTCTTTTTAAAGTGCTTGGAGACCATTCCTTTAATAATCCATTAGAATATCTTTCCAAATCACTGTGTTCTGTAGTTTG
GGAAGTCTGCCTTCTTCCCCTTTTTGAAAATTTATGCTACATTTATCATCTCATCTTCTAGCACCTCTCCATTCTTTGTGATTC
CTCAACTATCCACAGAGAGCAATTCCATGGCCTGCCTACAAGGTCTTTCGGTTTCCTGGGATTTGCCCATCCAGTCCAGTAATT
CATTTAGAATGGATCAATTATTTGCTATCTTACATCTTTTTACCCATTTTAGAGTTTAATTTCTTCTCCCTTTTTCAGTCTGAC
AGTCATTCTCCTTGATAGAGAAGCCAGGAACAAAATAGGAGGGAGAGAGTTTTGCTTTTTCTTTATTATCTACTGCTTTTAACA
ATAAACCTTCCTTGTTTTGATGTTATTATGTTGTTTGTCTTTTTTTTTTACTTATTTGCCTTTGTGACATGGGGACGGTGATAG
GGCCTTAAATATAATTTTAAAATAGGGAATAAATGGTTGTCTTTAGTATTTTATTTTGTTTTATTATTATTATTATTATTGTTA
TTTTTGCAAGCTTCAGCTAATTTGGAATTGTAGCTCTCCTGACATTATTCTTATAAGCTCATTCCACTCTCTTATAGACCATCA
TTACATGCCCTCTTTCCATCTTTTAAAATATGTCCTTTAAAAATCTGACCTGGGAGAAATCTCTGTGAAGCCGTGTTGGTTACT
TAAGTGCCACCCCTCTTTTCTTCCTGAGAGGATCATTTGTGATTGCAGTTACAGTTGA
```

Paired End Sequencing Reads

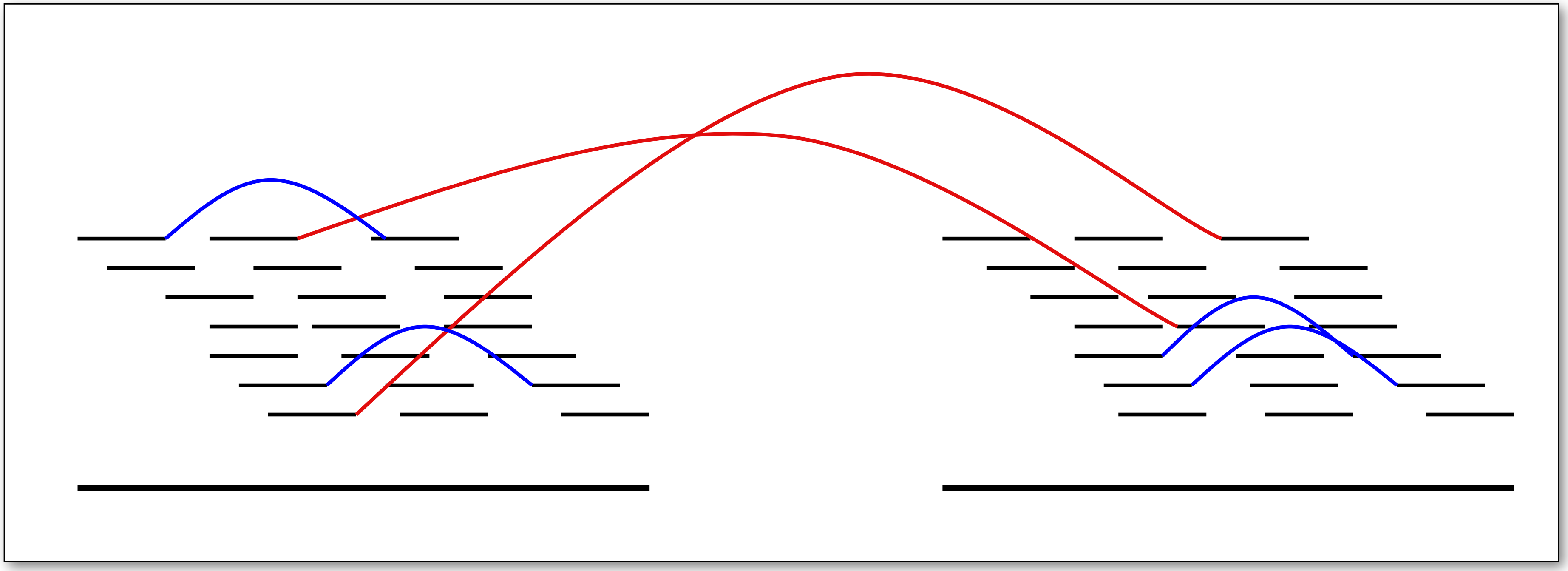
```
GCAATGAAATATGTTCTTGTAATTTAAGCTGACACTCCTAATTTAGCTCTTGTCCTCTACTGAGTCTACCTAATTATATGTAT
GGATTGACTTGGTGTTTTCTCTTTTTCTTAAATAGTAATGCAGAAAGCCTGGAGAGAGAGAAACCCCCAAGCTAGGATTTCTG
CAGCTCATGAAGCCTTGGAGATAAATGAGTAAGTGGGGGAAAATCTTGCTGTTAAAAAGGAAATCTCATCCTTTGCTGAATAT
ATTCAGTTGCCATTGATAGGATACTTAAATTAAACTGCATTTGAACTGGAGGATTATTTGGGGAGTTATTACTCTATTTAAAA
AAGTTTTTTTTTTAAATGAAGGACAGCCACCATGTGGAGGTGGTTTTAGTCATTTTATGAATTCAATGGCTTTGCTGTGATCC
...................................................................................
TAAATTAATTTCTTGAAGGGCTATCCCTAGGATATTGTGAGGATATAAAATAAATACAATTCTTTACATATCTAAAACATTCT
GACAGGGAAAATTTTCCAGATGTAGAATGCTCATCTGCACTAGAACATTTTCTAGTAGAACTTCTGCTAGTGGGGAAAACATG
ATAACAACATAAGGTTTAAAAAAAAAATTTTAGAAAATACTTCAAGATTAAGACAAAGATAAGAGGAAATGCTGTCTTGAGTG
TTGTTAAACATTCTGTGGGTTACCAAGGAAGGCTGGGAAATCTCTTCTGGAGATCTCAGAAAATGAGAAAGATTCTTAAAGTT
GGAGTCATAAAAACTCAGGGTTGGCAGAGACCTTAAAGGTCACTTAGCTGAACCACCCATCTGGTGCTTGAATCACCTCAACA
CTATCCTTGCCAAGTGGTCATTGTTAAACTATTTTATGATTTTTCTGAAGAAGGTTACAGAATCTTCTTCAGAGATCTTAGGG
```



Assembly: Contigs and Supercontigs


"Supercontig" or "Scaffold"
"Contig"
Seq gap
"Contig"
NNNNNN
number of N's in sequence = estimated size

Why Different Insert Sizes are Useful



Longer (fosmid) mate pairs connect assembly pieces that are not connected by shorter (plasmid) paired ends

Key Concepts in Assembly

- **Contig N50**
  - 50% of the genome assembly is in contigs larger than this size
- **Supercontig** (scaffold) **N50**
  - same, but for scaffolds

- **k-mer**
  - string of bases of length k
  - for computational efficiency, long sequences such as sanger reads are often chopped up into their constituent k-mers; usually *overlapping* k-mers are used because converting a sequence into nonoverlapping k-mers loses information

The first three overlapping 22-mers and their positions in a Sanger read

```
Read   tagcgactacctgaactggacctttgaacgag...
0      tagcgactacctgaactggacc
1       agcgactacctgaactggacct
2        gcgactacctgaactggacctt
```

- **High-quality mismatch**
  - A position in two well-aligning reads in which the base calls are *high quality* but *disagree*
  - Indicative of **allelism** or **paralogy**

A high-quality mismatch: High Phred scores (like Q99) on both mismatched bases

```
Read 1 ..actacctgaactggacctttgaacg...
Read 2 ..actacctgaactagacctttgaacg...
```

Assemblies are not Perfect

- Sequence coverage may vary
  - missing regions; strong fragmentation
- Some regions don't clone well
  - results in low sequence coverage
  - which causes gaps in assembly
- Some regions don't sequence well
  - extreme GC content
  - homopolymeric or otherwise low-complexity runs
- Some regions don't assemble well
  - mobile elements
    - high identity, large copy number
  - segmental duplications
    - Repeats are the single biggest impediment to assembly
- Polymorphism
- *Best way to improve assemblies is longer reads and better long range continuity*

High Throughput Sequencing

"The cost of DNA sequencing has plunged orders of magnitude in the last 25 years. Back in 1990, sequencing 1 million nucleotides cost the equivalent of 15 tons of gold (adjusted to 1990 price). At that time, this amount of material was equivalent to the output of all United States gold mines combined over two weeks. Fastforwarding to the present, sequencing 1 million nucleotides is equivalent to the value of ~30 g of aluminum. This is approximately the amount of material needed to wrap five breakfast sandwiches at a New York City food car."

Erlich Y. (2015). A vision for ubiquitous sequencing. *Genome Res.* **25(10)**:1411-6.

The Players

- Commercially available now:
  - Illumina – most prevalent technology
  - SOLiD (Life Technologies)
  - Ion Torrent (Life Technologies)
  - Pacific Biosciences
  - Complete Genomics – aquired by BGI, possibly dead
  - 454, Helicos – both commercially dead
- Next generation approaches
  - Oxford Nanopore
  - Illumina Nanopore (nothing released yet)
    - Recently licensed an alternative nanopore technology
  - NABsys, Genia, Noblegen – might all be dead

Sequencing Template Approaches

- Clonal Amplification of Single Molecules
  - Single molecule only briefly needed as a template
  - Thousands of identical molecules boost signal
  - Two different methods
    - Bridge amplification of molecules immobilized on surface
      - Illumina
    - Emulsion PCR
      - SOLiD and Ion Torrent, 454
- Single DNA molecule as a sequencing template.
  - Challenges include:
    - Keeping single molecules stable during insults of sequencing
    - Signal to noise ratio in base detection

    BUT

    - Avoid amplification biases
  - Pacific Biosciences, Oxford Nanopore, Helicos

Recommended Reading

Early Sequencing Technology:

- Maxam, A.M., Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci USA* **74(2)**:560-4.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *PNAS* **74**, 5463-7.
- Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B. and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature* **321(6071)**:674-9.
- Sanders, J.Z., Petterson, A.A., Hughes, P.J., Connell, C.R., Raff, M., Menchen, S., Hood, L.E. and Teplow, D.B. (1991). Imaging as a tool for improving length and accuracy of sequence analysis in automated fluorescence-based DNA sequencing. *Electrophoresis* **12(1)**:3-11.
- McCombie WR, Heiner C, Kelley JM, Fitzgerald MG, Gocayne JD. (1992). Rapid and reliable fluorescent cycle sequencing of double-stranded templates. *DNA Seq.* **2(5)**:289-96.
- Kasianowicz JJ, Brandin E, Branton D, Deamer DW. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *PNAS* **93(24)**:13770-3. **Initial nanopore paper**

New Sequencing Technologies:

- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456(7218)**:53-9. **Illumina**
- Eid, J. *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science*. **323**, 133-8. **PacBio**
- Flusberg, B.A. *et al.* (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7(6)**:461-5. **PacBio**
- Rothberg J.M., Hinz, W. et al (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475(7356)**:348-52. **IonTorrent**
- Ayub, M. and Bayley, H. (2012). Single Molecule RNA Base Identification with a Biological Nanopore. *Biophysical Journal* **102**:429. **Oxford Nanopore**
- Quick J, Quinlan AR, Loman NJ. (2014). A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience* **3**:22. **Oxford Nanopore – has data.**
- Ashton, P.M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., O'Grady, J. (2015). MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol*. **33(3)**:296-300. **Oxford Nanopore - has data.**
- Manrao, E.A., Derrington, I.M., Laszlo, A.H., Langford, K.W., Hopper, M.K., Gillgren, N., Pavlenok, M., Niederweis, M., Gundlach, J.H. (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat Biotechnol*. **30(4)**:349-53. **Nanopore technology licensed by Illumina**
- Derrington, I.M., Craig, J.M., Stava, E., Laszlo, A.H., Ross, B.C., Brinkerhoff, H., Nova, I.C., Doering, K., Tickman, B.I., Ronaghi, M., Mandell, J.G., Gunderson, K.L., Gundlach, J.H. (2015). Subangstrom single-molecule measurements of motor proteins using a nanopore. *Nat Biotechnol*. **33(10)**:1073-5. **First nanopore paper with Illumina authors**

Recommended Reading

Landmark Genome Sequencing Papers:

- Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. and Ysebaert, M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260(5551)**:500-7. First viral RNA genome
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265(5596)**:687-95. First DNA genome
- Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996). Life with 6000 genes. *Science* **274(5287)**:546, 563-7. Yeast Genome Paper – 1st sequenced eukaryote
- *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282(5396)**:2012-8. 1st sequenced multicellular eukaryote
- Adams, M.D., *et al.* (2000). The genome sequence of *Drosophila melanogaster*. Science 287(5461):2185-95.
- Lander, E.S., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409(6822)**:860-921.
- Venter, J.C. *et al.* (2001). The sequence of the human genome. *Science* **291(5507)**:1304-51.
- Mouse Genome Sequencing Consortium, *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420(6915)**:520-62.

Assembly Algorithms:

- Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P. and Lander, E.S. (2002). ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177-89.
- Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C. and Lander, E.S. (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-6.
- Phillippy, A., Schatz, M. and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008).
- Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D. and Phillippy, A.M. (2012). Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat Biotechnol*. **30(7)**, 693-700.
- Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M.C., McCombie, W.R. (2015). Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res*. **25(11)**:1750-6.

Recommended Reading

Recent Reviews:

- Erlich, Y. (2015). A vision for ubiquitous sequencing. *Genome Res.* **25(10)**:1411-6.
- Feng, Y., Zhang, Y., Ying, C., Wang, D., Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics* **13(1)**:4-16.
- Heather, J.M., Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* **107(1)**:1-8.
- Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., McCombie, W.R., Schatz, M (2016). Third-generation sequencing and the future of genomics. bioRxiv https://doi.org/10.1101/048603.
- Jiao, W.B., Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant. Biol.* **36**:64-70.
- Mardis, E.R. (2017). DNA sequencing technologies: 2006-2016. *Nat. Protoc.* **12(2)**:213-218.
- Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A., Waterston, R.H. (2017). DNA sequencing at 40: past, present and future. *Nature* **550(7676)**:345-353.
- Green, E.D., Rubin, E.M., Olson, M.V. (2017). The future of DNA sequencing. *Nature* **550(7675)**:179-181.

Sequencing Theory:

- Clarke, L. and Carbon, J. (1976). A colony bank containing synthetic Col El hybrid plasmids representative of the entire *E. coli* genome. *Cell* **9(1)**:91-9.
- Lander, E.S. and Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2(3)**:231-9.
- Roach, J.C., Boysen, C., Wang, K. and Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26(2)**:345-53.
- Roach J.C. (1995). Random subcloning. *Genome Res.* **5(5)**:464-73.