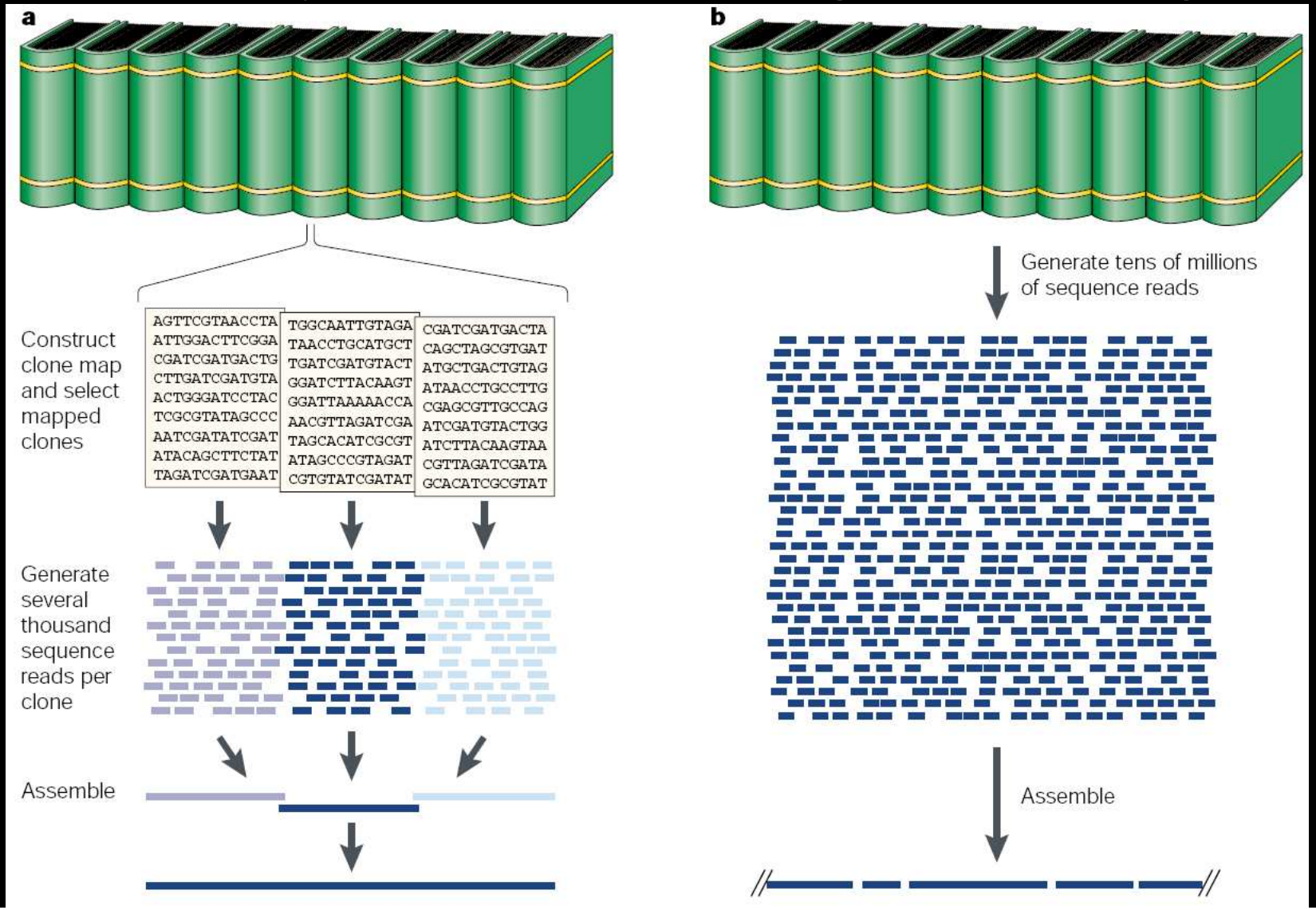
Genome Sequencing and Assembly
High throughput Sequencing

IV MSc Botany

Dr Giby Kuriakose

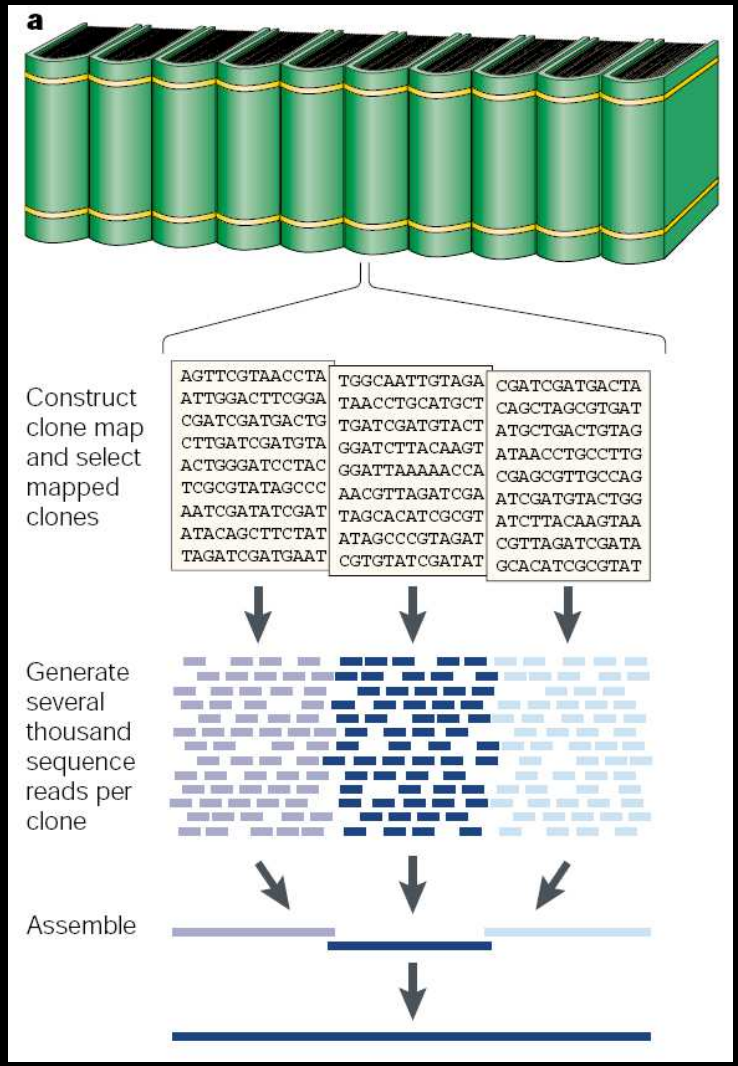
Competing Sequencing Strategies

- Clone-by-clone and whole-genome shotgun



Clone-by-Clone Shotgun Sequencing

- E.g. Human genome project
- Map construction
- Clone selection
- Subclone library construction
- Random shotgun phase
- Directed finishing phase and sequence authentication

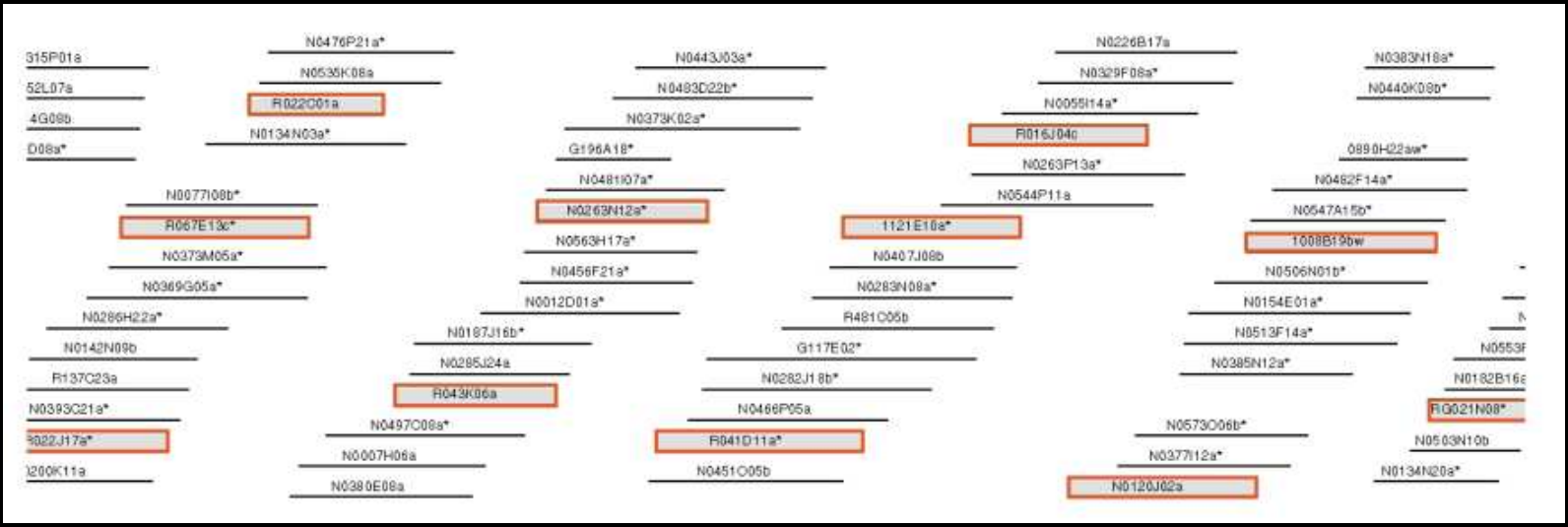


Map Construction

- Clone genomic DNA in YACs (~1MB) or BACs (~200KB)
- Map the relative location of clones
  - Sequenced-tagged sites (STS, e.g. EST) mapping
  - PCR or probe hybridization to screen STS
  - Restriction site fingerprint
- Most time consuming
  - 1990-98 to generate physical maps for human http://www.ncbi.nlm.nih.gov/genemap99/

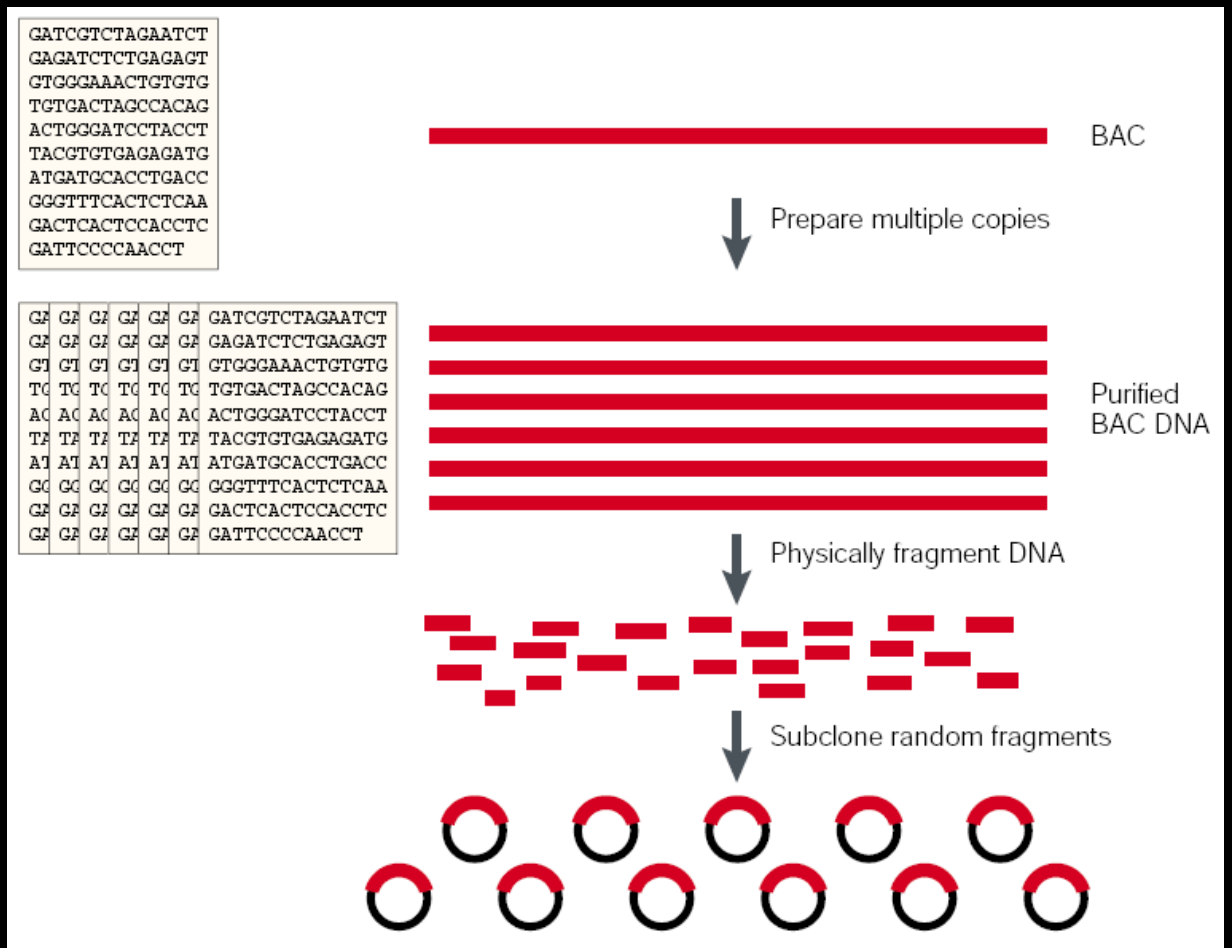
Clone Selection

- Based on clone map, select authentic clones to generate a minimum tiling path
  - Most important criteria: authentic



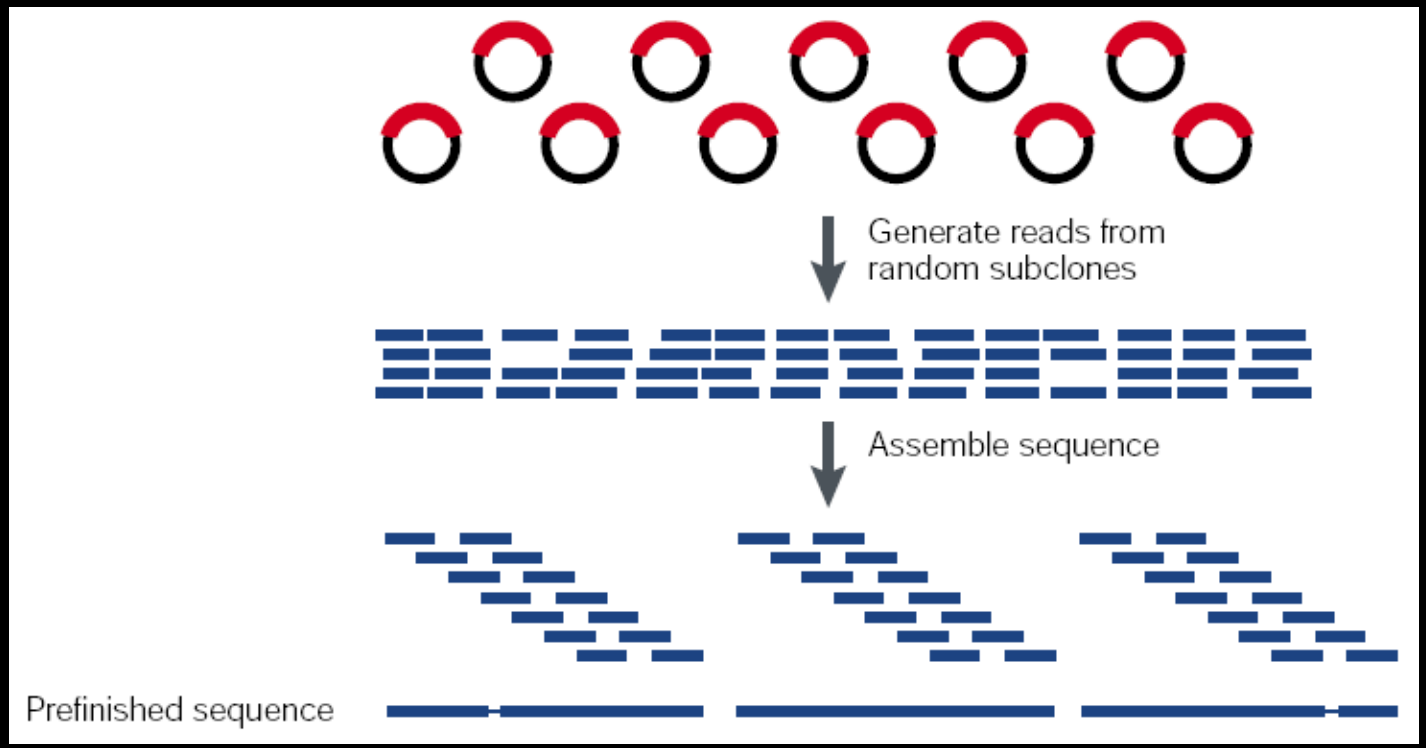
Subclone Library Construction

- DNA fragmented by sonication or RE cut
- Fragment size ~ 2-5 KB



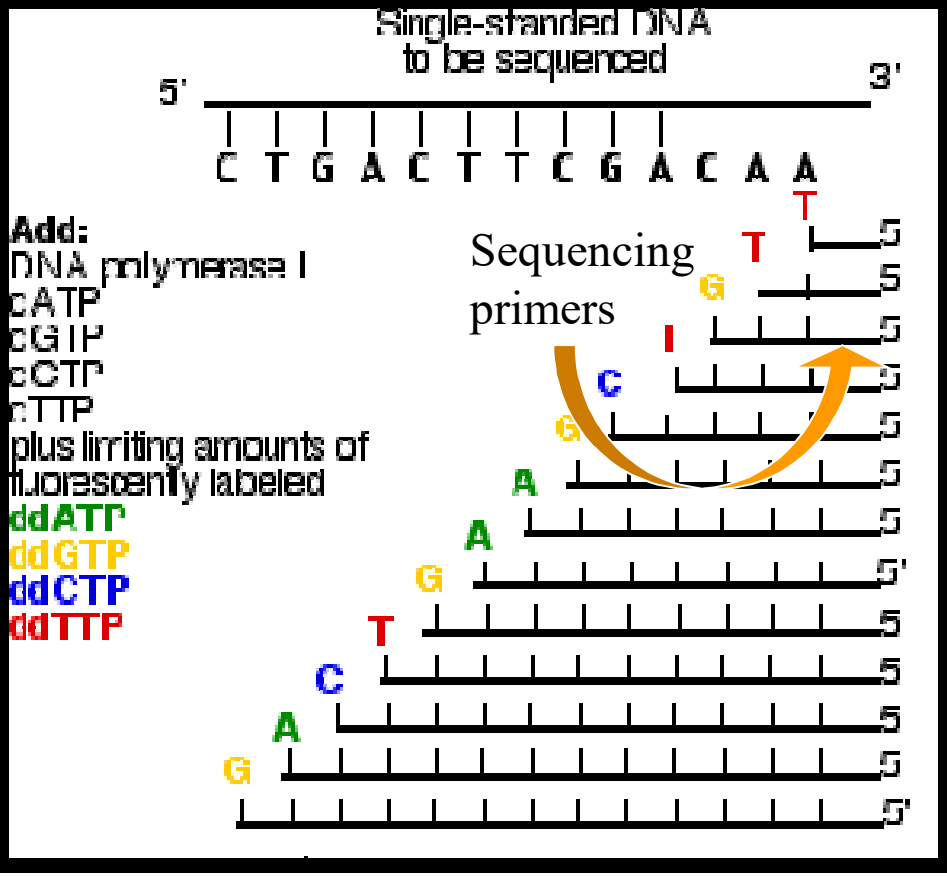
Random Shotgun Phase

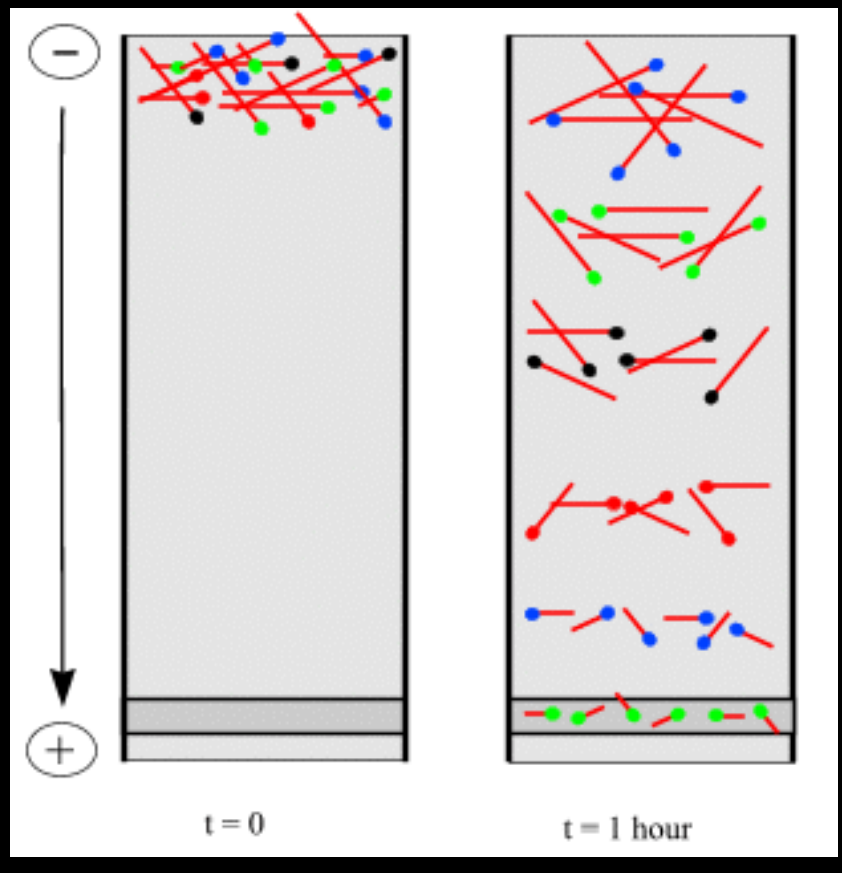
- Dideoxy termination reaction
- Informatics programs
- Coverage and contigs



Dideoxy Termination

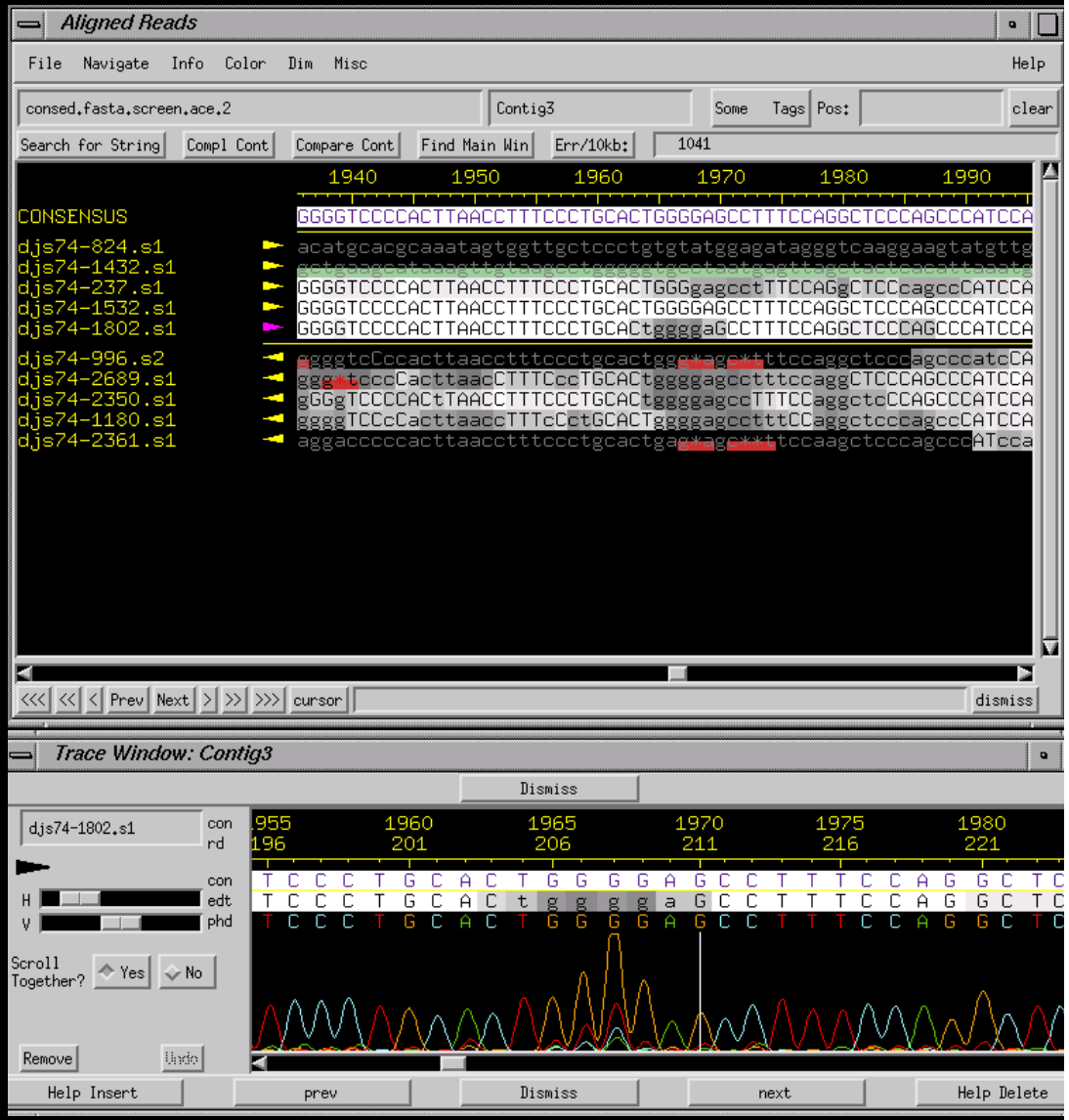
- Method invented by Fred Sanger
- Automated sequencing developed by Leroy Hood (Caltech) and Michael Hunkapiller (ABI)





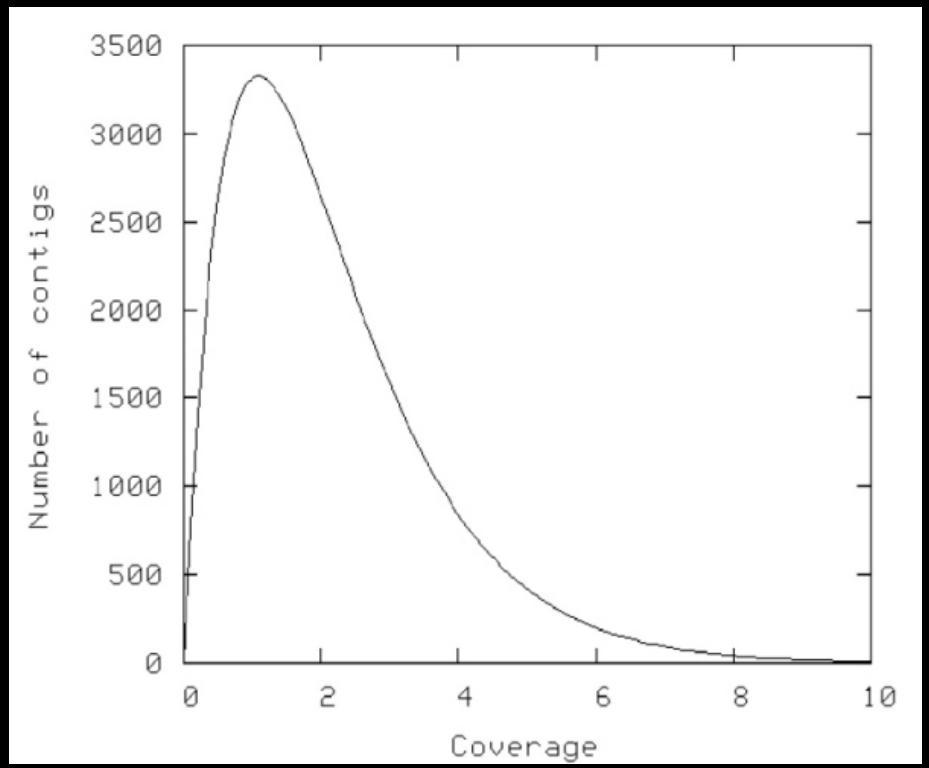
Bioinformatics Programs

- Developed at Univ. Wash
- Phred
  - Base calling
  - Phil Green
- Phrap
  - Assembly
  - Brent Ewing
- Consed
  - Viewing and editing
  - David Gordon



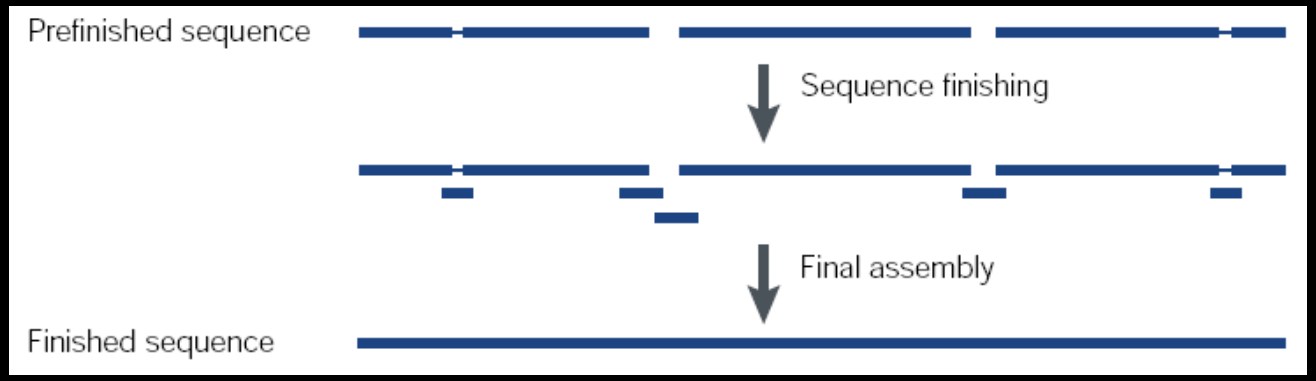
Coverage and contigs

- Coverage: sequenced bp / fragment size
  - E.g. 200KB BAC, sequenced 1000 x 500bp subclones, coverage = 1000 x 500bp / 200KB = 2.5X
- Lander-Waterman curve



Directed Finishing Phase

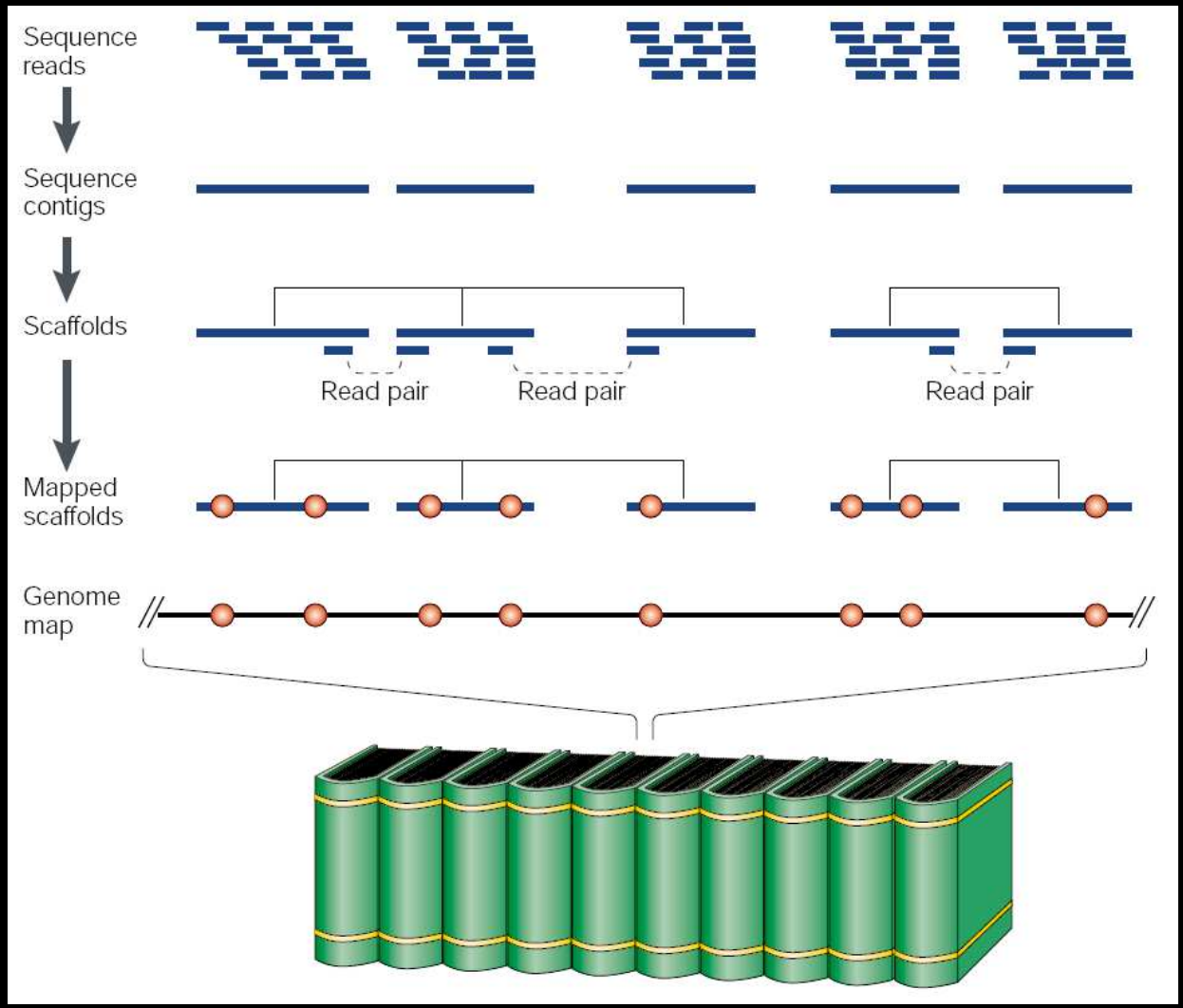
- David Gordon: auto-finish
  - Deign primers at gap 2 ends, PCR amplify, and sequence the two ends until they meet



- Sequence authentication: verify STS and RE sites
  - Finished: < 1 error (or ambiguity) in 10,000bp, in the right order and orientation along a chromosome, almost no gaps.

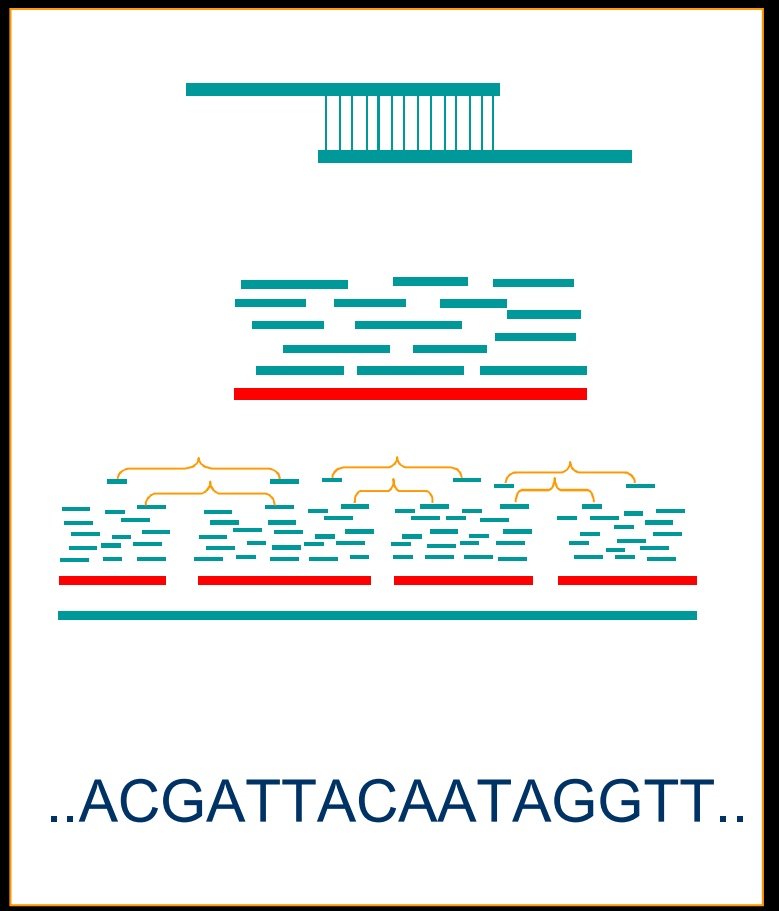
Genome-Shotgun Sequencing

- Celera human and drosophila genomes
- No physical map
- Jigsaw puzzle assembly
- Coverage ~7-10X

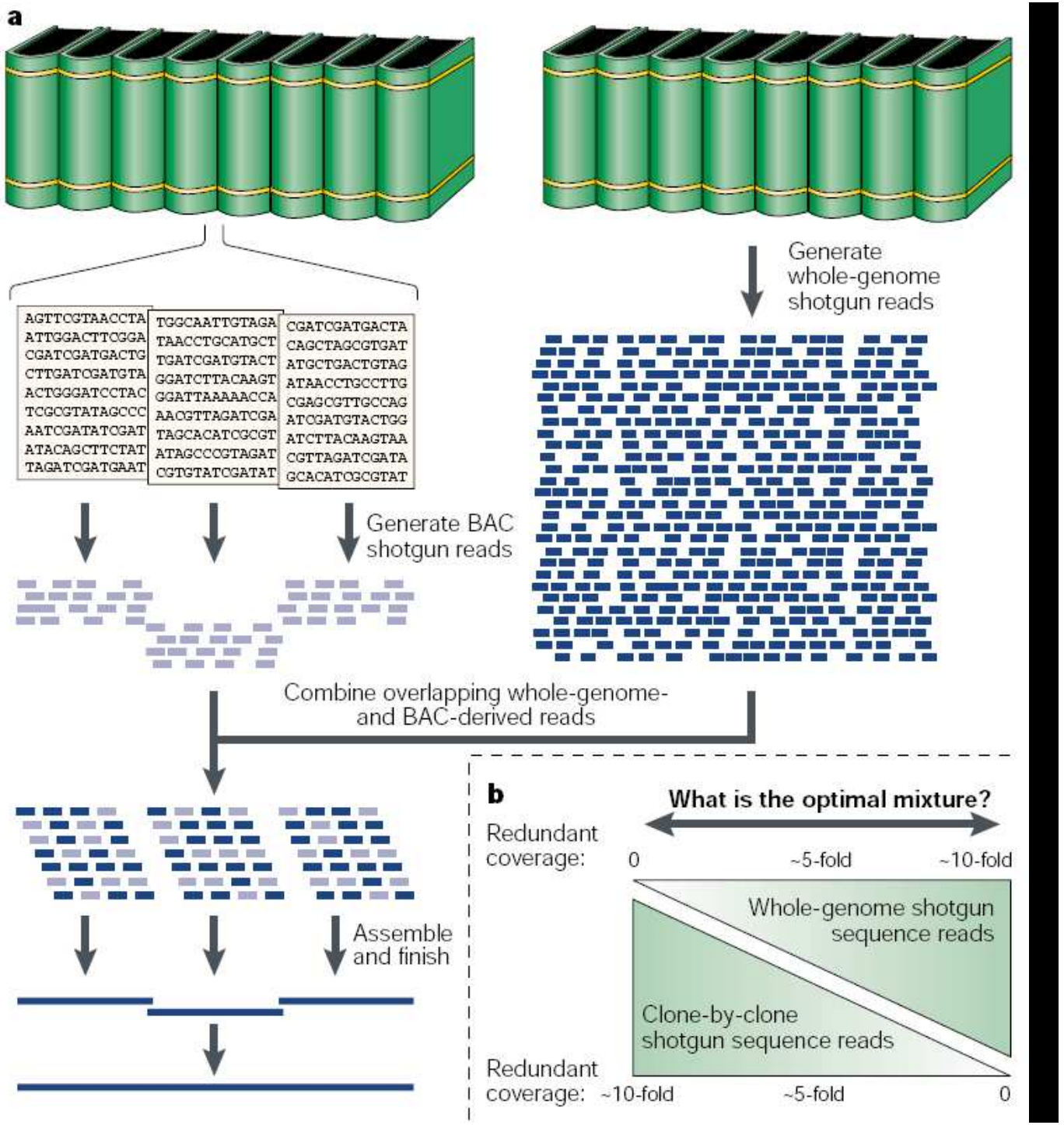


Shotgun Assembly

- Screener
  - Identify low quality reads, contamination, and repeats
- Overlapper
  - >= 40bp overlap with <= 6% mismatches
- Unitigger
  - Combine the easy (unique assembly) subset first
- Scaffolder & repeat resolution
  - Generate different sized-clone libraries, and just sequence the clone ends (read pairs)
  - Use physical map information if available
- Consensus



Hybrid Method



Hybrid Method

- Optimal mixture of clone-by-clone vs whole-genome shotgun not established
  - Still need 8-10X overall coverage
  - Bacteria genomes can be sequenced WGS alone
  - Higher eukaryotes need more clone-by-clone
  - Comparative genomics can reduce the physical mapping (clone-by-clone) need

First Generation Sequencing

PRODUCTION

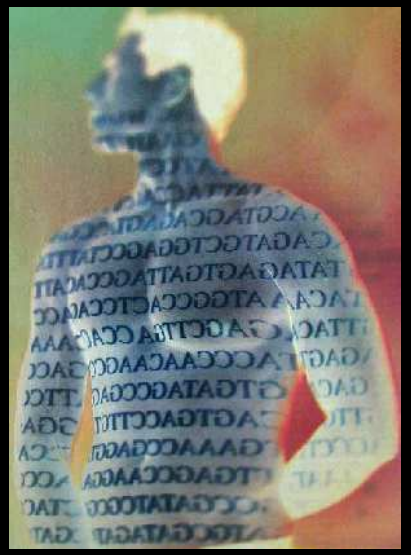
Rooms of equipment
Sample preparations
35 people
3-4 weeks

SEQUENCING

74x Capillary Sequencers
10 people
15-40 runs per day
1-2Mb per instrument per day
120Mb total capacity per day

Human Genome Project

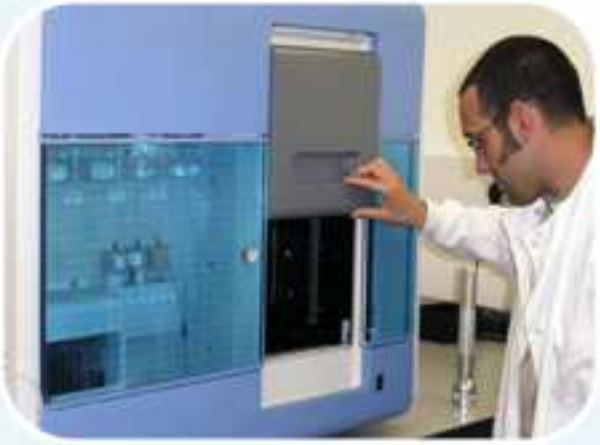
- 1990-2003
- Cost $3,000,000,000
- Thousands of scientists in six countries
- Triumph of automation and bioinformatics
- More significant than the Manhattan Project and moon landing



Second Generation Sequencing

PRODUCTION

1x Cluster Station
1 person
1 day



SEQUENCING

1x Genome Analyzer
Same person as above
1 run per 3-5 days
0.5Gb per day per instrument

2nd Gen Sequencing Tech

- Traditional sequencing: 384 reads ~1kb / 3 hours
- 454 (Roche):
  - 1M reads 450-1000bp / 10-24 hours
- HiSeq (Illumina):

  - http://www.youtube.com/watch?v=HtuUFUnYB9Y
  - 100-200M reads of 50-100bp / 3-8 days * 16 samples
- SOLiD (Applied Biosystems)
  - >100M reads of 50-60bp / 2-8 days * 12 samples
- Ion Torrent (Roche):
  - http://www.youtube.com/watch?v=yVf2295JqUg
  - 5-10M reads of 200-400bp / < 2 hours

Illumina HiSeq2000

- Throughput:
  - $1000-2500 / lane (depends on read length, SE / PE)
  - 50-100 bp / read
  - 16 lanes (2 flow cells) / run
  - 150-200 million reads / lane
  - Sequencing a human genome: $3000, 1 week
- Bioinfo challenges
  - Very large files
  - CPU and RAM hungry
  - Sequence quality filtering
  - Mapping and downstream analysis

(Potential) Applications

- Metagenomics and infectious disease
- Ancient DNA, recreate extinct species
- Comparative genomics (between species) and personal genomes (within species)
- Genetic tests and forensics
- Circulating nucleic acids
- Risk, diagnosis, and prognosis prediction
- Transcriptome and transcriptional regulation
  - More later in the semester…

Third Generation Sequencing

- Single molecule sequencing (no amplification needed)
- Oxford Nanopore: Read fewer but longer sequences

  http://www.youtube.com/watch?v=_rRrOT9gfpo
- In 1-2 years, the cost of sequencing a human genome will drop below $1000, storage will cost more than sequencing
- Personal genome sequencing might become a key component of public health in every developed country
- Bioinformatics will be key to convert data into knowledge

Summary

- Genome sequencing and assembly
  - Clone-by-clone: HGP
    - Map big clones, find path, shotgun sequence subclones, assemble and finish
    - Sequencing: dideoxy termination
  - Whole genome shotgun: Celera
- Massively Parallel Sequencing
  - 454, Solexa, SOLiD, Ion Torrent, Oxford Nanopore…
  - Many opportunities and many challenges

Acknowledgement

- Fritz Roth
- Dannie Durand
- Larry Hunter
- Richard Davis
- Wei Li
- Jarek Meller
- Stefan Bekiranov
- Stuart M. Brown
- Rob Mitra