Course summary

- Goals
- Organization
- Labs
- Project
- Reading

Today

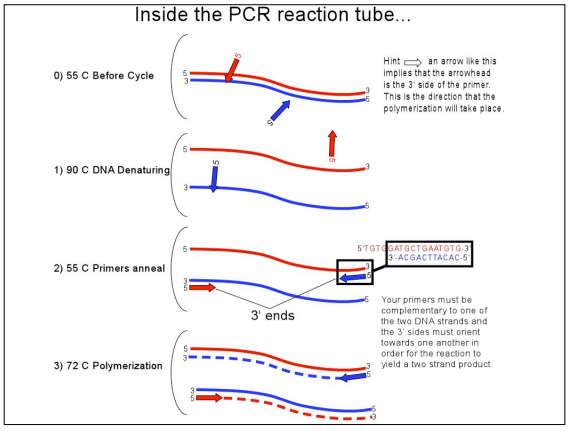
- DNA sequencing.
- Genome Projects.
- New DNA sequencing technologies.

Obtaining molecular data

- PCR
  - Typically used in empirical molecular evolution studies
- Sequencing
  - Sequence chromatograms and base calling
- Genome sequencing (data used by Mol. Evol.)
  - Shotgun versus Hierarchical
- Cloning vectors
  - Plasmids and BACs
- The human genome

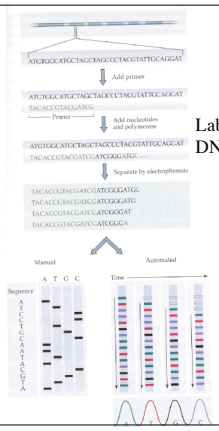
PCR
Polymerase chain reaction

- Template DNA
- Primers (must know sequence)
- dNTPs (dCTP, dGTP, dATP, dTTP)
- Polymerase
- Buffer
- $MgCl_2$



Sequencing

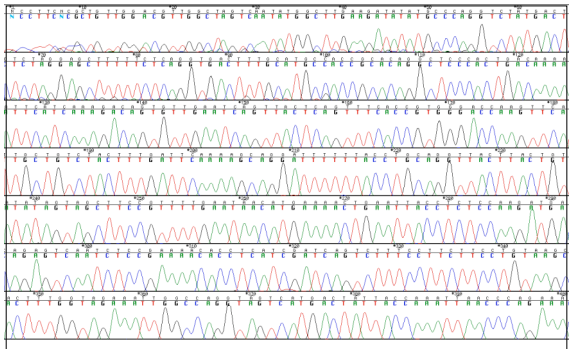
- Template DNA
- Primers
- dNTPs (dCTP, dGTP, dATP, dTTP)
- **Labeled ddNTPs**
- Polymerase
- Buffer
- $MgCl_2$



Labeled ddNTPs to terminate DNA synthesis.

Base calling

- Merge all four colors channels into one file
- Calculate spacing (mean distance between peaks)
- Detect local maxima for each four colors



How do you know which peaks are real?

Particularly important for high-throughput sequencing.

PHRED score

- Variance in peak spacing over 7 peaks
  - Should be low
- Ratio of largest to smallest called peak in window
  - Should be 1
- Same ratio over 3 peak window
- Number of bases between current base and nearest unresolved base.
  - Should be high

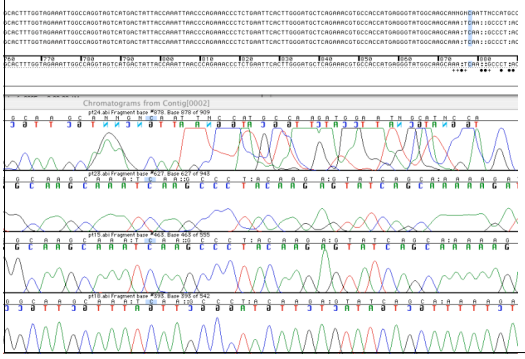
Phred score reported as:

$Q = -10 \log_{10}(Pe)$

where Pe is error probability of base call.

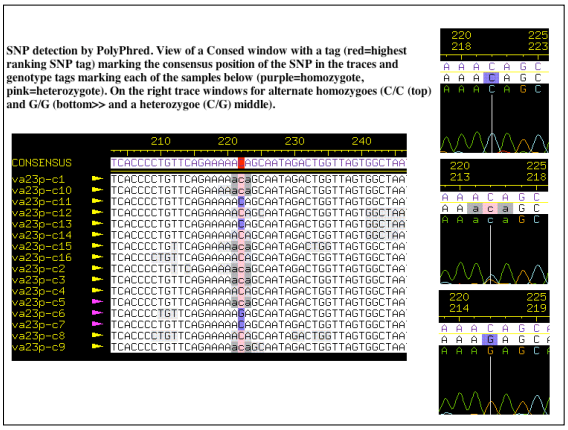
Score 30 means error probabilty is 0.001.

Sequencher does not take into account quality of sequence: Gaps in "bad" sequence are included in consensus.



What about polymorphism in sequence data?

- Directly sequenced PCR products can be from diploid individuals with potential for heterzygotes.



Genome Projects

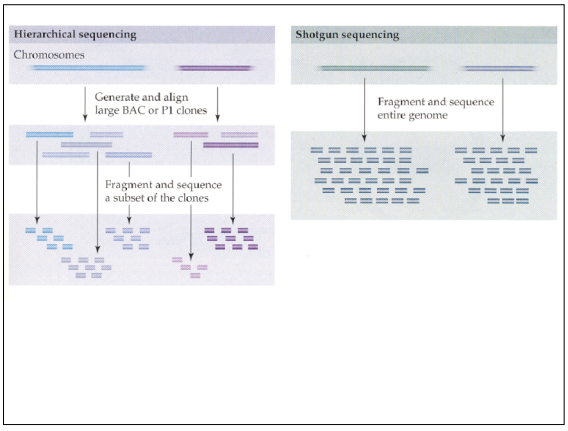
- Drosophila
- C. elegans
- Human, Chimpanzee
- Mouse, Rat
- Dog
- Plants
- Cow, Rhesus, …..

Goals of Human Genome Project

- Technology development
  - High throughput, inexpensive DNA sequencing
- Complete Sequence
  - Provide the genetic sequence
- Gene Identification
  - Identify the "parts' list. Also non-coding functional regions
- Polymorphism Data
  - Association studies, evolutionary studies
- Functional characterization
  - ENCODE

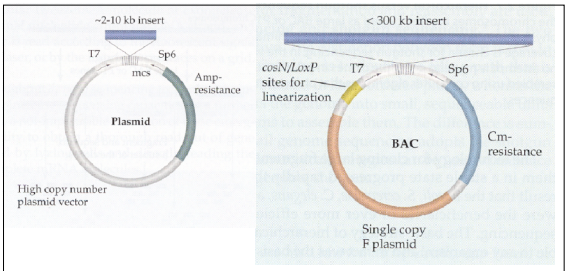
Two strategies for genome sequencing

- Whole Genome Shotgun
  - Random reads assembled by computer
- Hierarchical
  - Map bases, clone by clone



Genome sequencing Terminology

- Coverage
  - Estimated number of times a region of the genome is sequenced (typically 5 - 10 fold). Note on distribution.
- Consensus Sequence
  - Sequence based upon multiple reads
- Contig
  - Contiguous sequence aligned
- Scaffold
  - Ordered contigs based upon BAC-end sequencing or other knowledge



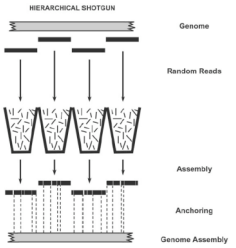
Plasmid multiple copy to allow easy purification large quantities
BACs single copy to prevent recombination

Sequencing clones

- Usually get better sequence than sequencing PCR products directly.
- Must remove vector sequence.
- Sometimes clones are chimeric

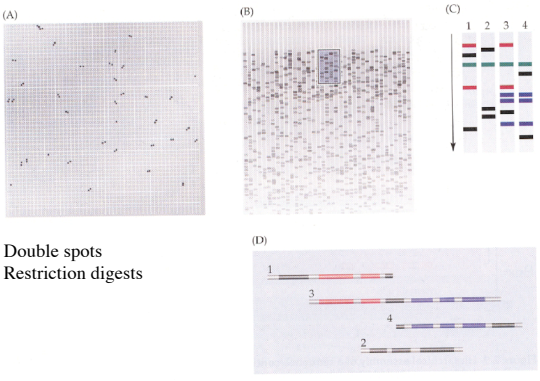
Hierarchical sequencing

- Break genome into smaller aligned chunks and sequence.

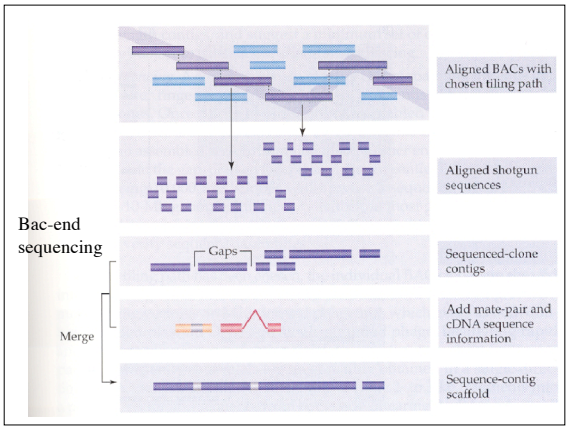


Aligning BACs

- Hybridize arrayed library with unique sequence.
- Isolate all hybridizing BACs.
- Perform restriction based mapping.
- Align BACs and make tilling path.

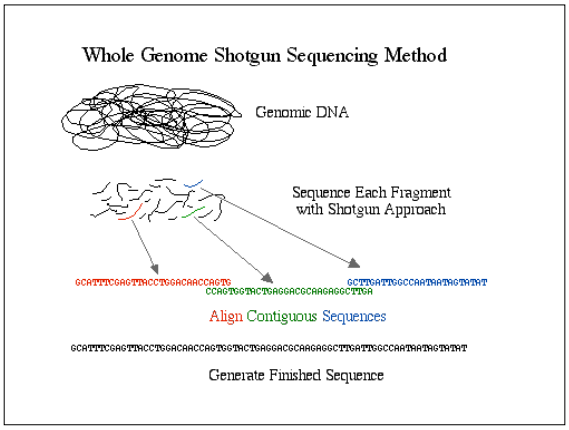


Double spots
Restriction digests



Shotgun sequencing

- Screen sequence for known repeats
  - Note there will still be duplications
- Finding overlapping sequence
- Link contigs into scaffolds
- Fill in gaps by BAC sequencing



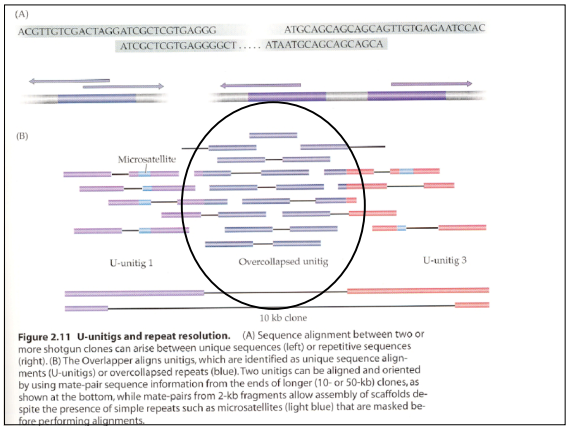


Figure 2.11 U-unitigs and repeat resolution. (A) Sequence alignment between two or more shotgun clones can arise between unique sequences (left) or repetitive sequences (right). (B) The Overlapper aligns unitigs, which are identified as unique sequence alignments (U-unitigs) or overcollapsed repeats (blue). Two unitigs can be aligned and oriented by using mate-pair sequence information from the ends of longer (10- or 50-kb) clones, as shown at the bottom, while mate-pairs from 2-kb fragments allow assembly of scaffolds despite the presence of simple repeats such as microsatellites (light blue) that are masked before performing alignments.

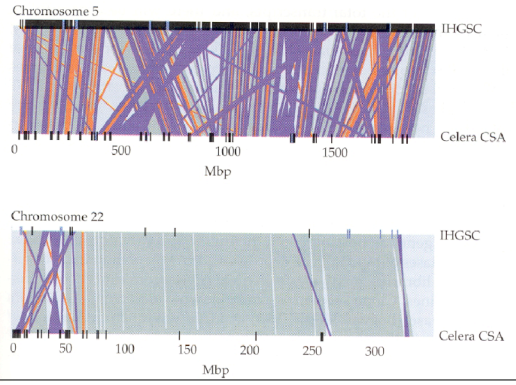
Sequence verification

- Completeness
  - Repetitive regions, heterochromatin
- Accuracy
  - Based upon probability scores for base calls (Phred scores)
- Validity of assembly
  - Difficult to assess. Predict restriction profiles and compare to observed. Correct spacing of pair-end sequences (correct for repeats), consistent read depth.

Is there any difference between two draft human genomes?

- International public consortium
  - Used hierarchical sequencing
    - Published in Nature 2001
- Private company Celera
  - Used Whole Genome Shotgun
    - Published in Science 2001

DRAFT GENOMES ARE FAR FROM COMPLETE



How effective was the whole genome shotgun approach?

Celera spiked their data with "shredded" reads from the human genome project

Celera used "shredded" reads from public data.
This was not done randomly was would be obtained from WGS

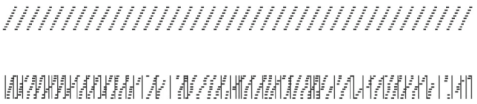


Fig. 2. Random vs. perfect spacing in 100 kb. (*Upper*) Two-fold coverage in perfectly spaced reads. (*Lower*) Two-fold coverage in randomly selected reads. There are frequent regions in which the adjacent reads either fail to overlap or the overlap is too small to allow reliable detection (< 40 bp). These breaks in continuity are indicated by vertical lines.

From Waterston et al. 2002 *PNAS* **99**, 3712 - 3716

Table 1. Reported statistics for genome assemblies in the HGP and Celera papers

Category	Celera WGS	Celera Faux WGS	Celera Faux CSA	HGP
Sequence coverage	5.1 × Celera	5.1 × Celera + 7.5 × HGP 12.6 × total	5.1 × Celera + 7.5 × HGP 12.6 × total	7.5 × HGP
Length (in Gb) of draft genome assembly, counting only bases with known sequence*	NR	2.587	2.654	2.693
Length (in Gb) of draft genome assembly, including unknown nucleotides in gaps†	NR	2.848	2.906	2.916
Proportion of sequence in euchromatic genome present in draft genome assembly, %‡	NR	89	91	92
Number of contigs§	NR	221,036	170,033	149,821
Number of scaffolds§	NR	118,968	53,591	87,757
Number of components, to be anchored in genome¶	NR	118,968	3,845	942

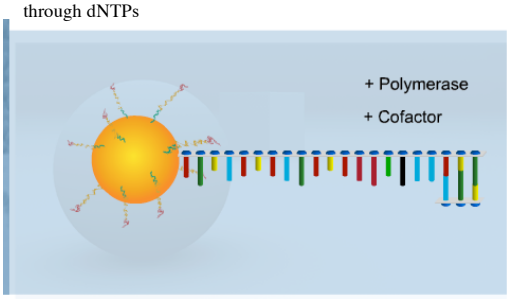
WGS: Whole genome shotgun
CSA: Compartmentalized sequence assembly
HGP: Human genome Project (hierarchical sequencing)

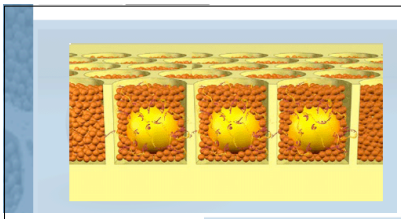
New sequencing technology Pyrosequencing (454)

- 4 enzymes
  - DNA polymerase
    - DNA synthesis
  - ATP sulfurylase
    - Converts PPi to ATP
  - Luciferase
    - Uses ATP to generate light
  - Apyrase
    - Degrades dNTPs
- Cheaper, higher throughput
  - 20 million bases per run, versus 67K
  - Moving towards the $1,000 genome

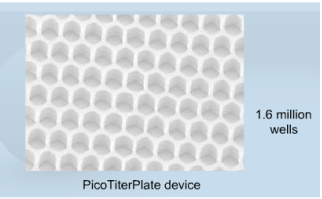
Show chemistry

Reaction carried out on beads so that can wash through dNTPs

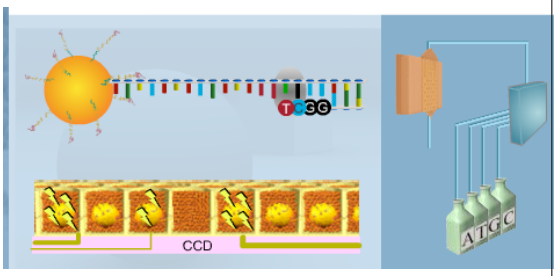




Beads placed individually into wells that can be recorded by CCD camera



Each nucleotide washed though and incorporation recorded by light production.



Some examples of pyrosequencing

- Marine Microbes
  - Large amount of sequencing to get "unculturable" microbes
- Extinct cave bears
  - Test case for Neanderthals
- Neanderthals
  - Small amount of highly degraded DNA, lots of microbe contamination, need to sequence a lot.

Disadvantages

- Short reads
  - Only 100 bps compared to 800+
- Homo-polymer runs difficult
  - Beyond 2 or 3, can not quantify
- Needs for reference genome
  - Reads potentially too short for shotgun assembly. No possibility of clone based sequencing
- Initial investment high
  - Instrument cost.

1. Describe DNA sequencing.
2. What is the difference between a BAC and Plasmid?
3. What are two methods for sequencing genomes?
4. What is hierarchical sequencing?
5. What is whole genome shotgun sequencing?
6. Describing a tilling path.
7. What is a phred score?
8. What is the difference between a scaffold and contig sequence?
9. What is the difference between dNTP and ddNTP?
10. Describe pyrosequencing? What is advantage/disadvantage?