# Sequencing methods, BAC fingerprinting, physical maps and FISH

Biosciences 741: Genomics

Fall, 2013

Week 2

## Lecture outline

- Genetic mapping methods
- Review of DNA sequencing
- Shotgun ligation & shotgun sequencing
- Automated DNA sequence assembly
- Sequencing with custom oligonucleotides
- Methods for dealing with repetitive DNA
- YAC, BAC & PAC cloning and contigs

## Genetic mapping techniques

- Recombination mapping (conventional genetic mapping) assume that the distance between two genes is proportional to the number of crossovers between them. The next three are physical methods:
- Radiation hybrid mapping uses X-rays or gamma-rays to fragment chromosomes at random, followed by construction of mouse hybrid cell lines that have only one foreign chromosome. The assumption is that genes that are located closer together will have a greater probability of being inherited by the same subset of cell lines. This is useful for constructing low-resolution maps of ESTs and similar markers.
- Fluorescent *in situ* hybridization uses an entire BAC (or similar size clone) as a hybridization probe to cell chromosomes on a microscope slide. In most cases, a single band is labeled. The map order is unambigous, and the resolution is ~100 kb.
- Restriction mapping is the highest resolution and the most time consuming of the physical mapping techniques.

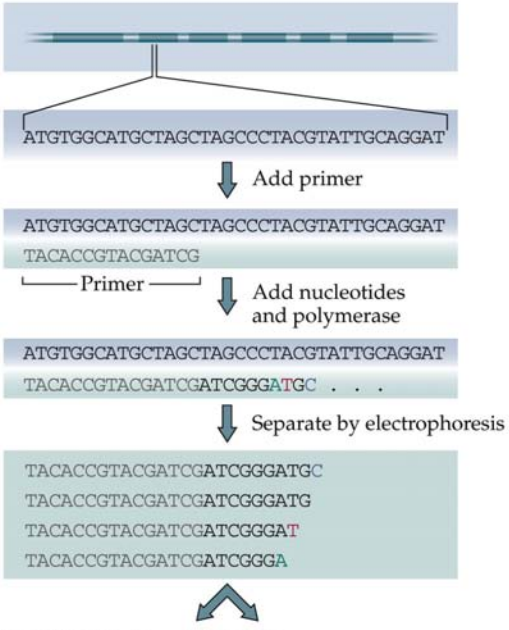
## Gel electrophoresis separates DNA fragments by size

- DNA is negatively charged, so it migrates in response to an electric field.
- The charge/mass ratio of DNA is constant (-1 per base), so all DNA fragments move at the same rate in a buffer solution (regardless of length or base composition).
- The gel matrix (agarose, acrylamide, etc.) slows DNA migration, and slows larger fragments more because they collide with the gel matrix more often.
- Denaturing reagents (formamide, urea) compete for hydrogen bonds, hence prevent base pairing of DNA, hence prevent sequence-specific effects on migration rates.
- If (and only if) we use relatively short DNA fragments (<600 bp), and high voltages (1000-5000 volts), and long migration distances (0.4 to 2.0 meters), then it is possible to clearly separate DNA fragments that differ in length by only one base!

## Dideoxy nucleotides are specific chain-terminators

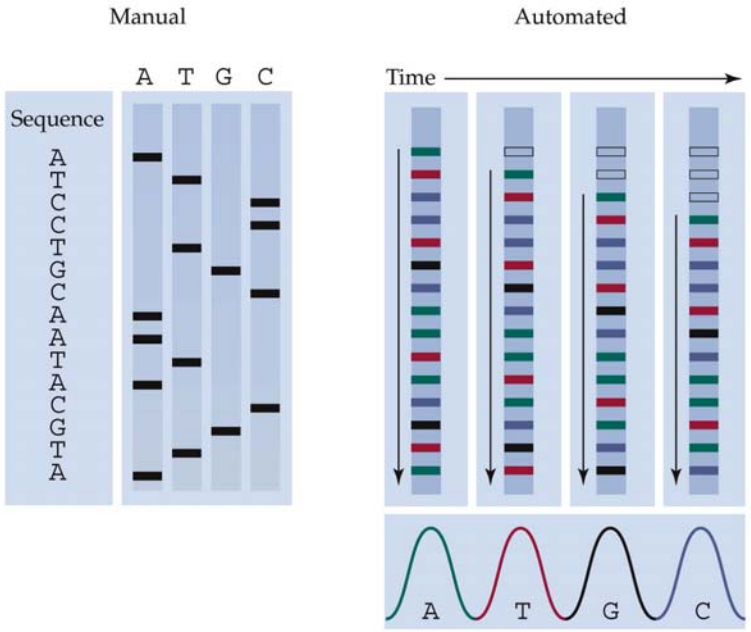
- 2',3' dideoxy nucleotides are specifically incorporated by DNA polymerase, but do not have a 3' OH group, and so the DNA strand can not be further extended beyond them.
- If each of the dideoxy nucleotides (ddATP, ddCTP, ddGTP, ddTTP) are covalently labeled with a fluorescent dye of a different color, then chains ending at A will all be red, chains ending at C will be blue, etc.
- If all of the newly-synthesized DNA chains start at the same point (a synthetic oligonucleotide primer), then the length of the chain will be different for each termination point.
- After separation of these DNA fragments by size, the base sequence can be read as a series of colors within the electrophoresis gel.

## Figure 2.1 The principle of dideoxy (Sanger) sequencing




*A PRIMER OF GENOME SCIENCE 3e*, Figure 2.1 (Part 1)

Figure 2.1 The principle of dideoxy (Sanger) sequencing



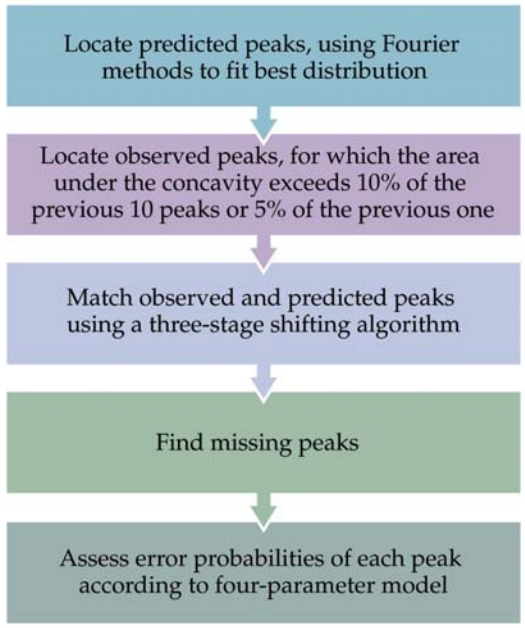

***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.1 (Part 2)**



## Reading sequences (aka "base calling")

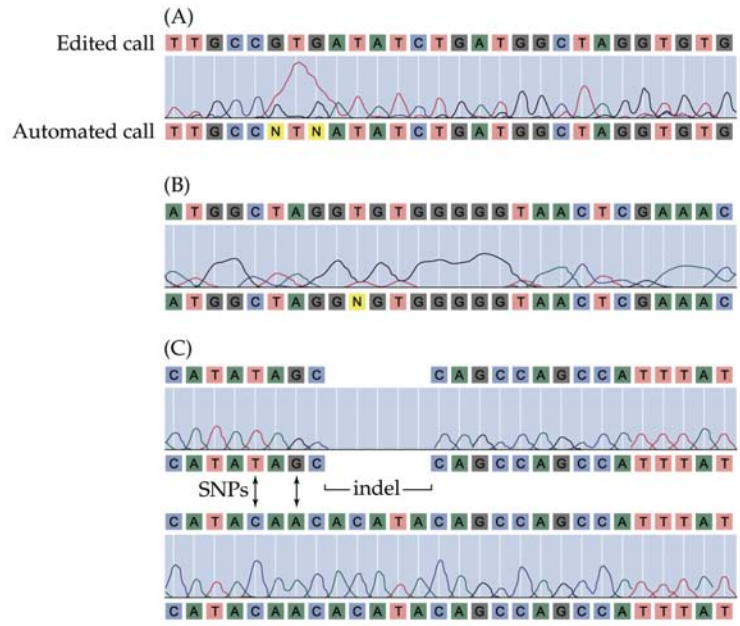
- The confidence in reading each base is assessed by the peak height (signal-to-noise ratio) and the regularity of peak spacing (phred).
- The signal-to-noise ratio is based on the local fluctuations in all four channels. The regularity of peak spacing is based on the spacing of nearby sequence peaks.
- The result is a "phred score", which is 10 times the negative logarithm of the error probability (if error prob $\leq 0.01$, then phred $\geq 20$).
- Bases 20-50 (after the primer) usually have low confidence because of both low peak height and irregular peak spacing.
- Bases 50-450 usually have high confidence. It is also possible in this range to read heterozygous sequences with high confidence (Marcus).
- Bases 450-700 generally have gradually declining confidence, due to gradually declining peak height and gradually declining peak spacing.

Figure 2.2 - the phred base-calling algorithm.




*A PRIMER OF GENOME SCIENCE 3e*, Figure 2.2






*A PRIMER OF GENOME SCIENCE 3e*, Figure 2.3

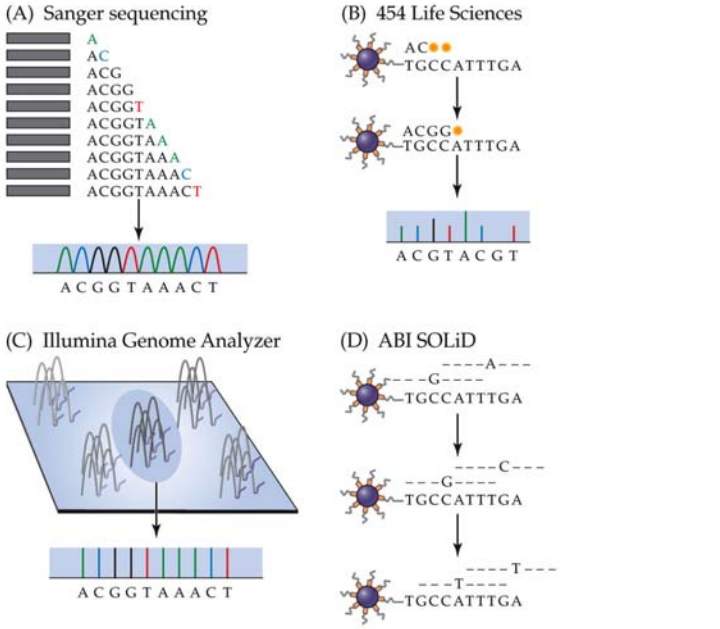
## DNA sequencing has become much faster and cheaper

- In the original Sanger method (using slab gels and $^{35}S$), a single person could process about 10 sequencing reactions per week, and read about 500-600 bases per reaction.
- A single capillary machine (ABI 310) can process about 10 sequencing reactions per day. The increase is due to automated gel pouring, sample loading and base calling. We still read about 500-600 bases per reaction.
- Capillary array machines (ABI 3700) can process about 1,000 sequencing reactions per day. The increase is due to parallel automation of many capillaries inside one machine.
- Laboratory robots (Beckman Biomek) and quad microtiter plate thermocyclers (MJ Research) make it possible for one technician to supervise ~10 capillary array machines, or 10,000 reactions per day.

## Alternative sequencing technologies

- Pyrosequencing - DNA is fragmented and coupled to beads. Each nucleotide incorporation is associated with a flash of light. Nucleotides are added sequentially. Sequence reads rather short (100 bp). Can be adapted to massively parallel processing.
- Microislands (Illumina) - cell free PCR "colonies" are labeled with fluorescent ddNTPs, then deprotected and labeled again. Short reads (2 x 50 bp), massively parallel.
- Polony sequencing - similar to above, but fluorescent dNTPs are added one at a time. Very short reads (20-25 bp).
- Oligonucleotide hybridization (and/or ligation)…
- Mass spectrophotometric - identify oligonucleotide fragments by time of flight. High signal/noise ratio, high resolution, short reads.

Figure 2.5 Principles of massively parallel genome resequencing



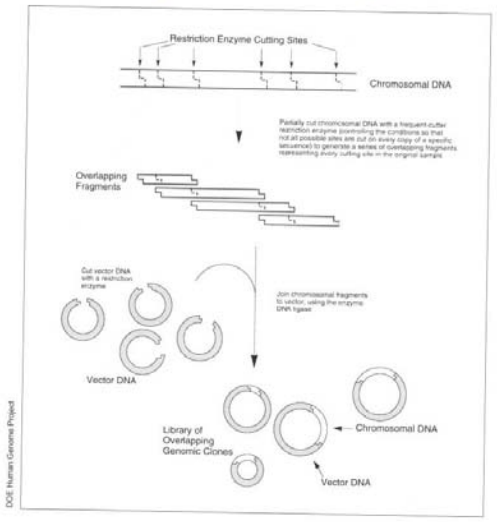

***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.5**



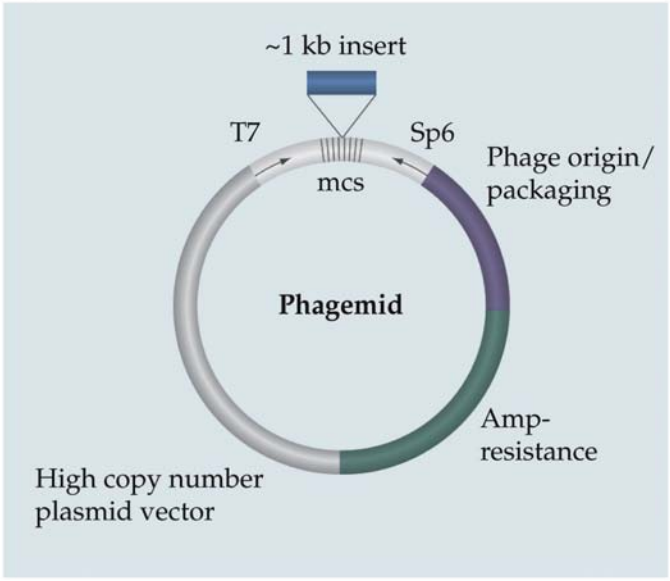
## *Sau*3AI restruction fragments are a convenient size for DNA sequencing

- The restriction enzyme *Sau*3AI has a 4-base recognition sequence (GATC).
- In DNA sequences with a GC content of 50% GC, *Sau*3AI fragments average 256 bp in length (why??).
- Even in DNA sequences with an unusual base composition, the fragment size does not differ much. For example, in DNA sequences with a GC content of 70% GC, *Sau*3AI fragments average 362 bp in length (why??).
- What would the fragment length be for 30% GC?
- As mentioned previously, it is possible to read an average of ~500-600 base pairs from a single sequencing reaction.
- Thus, a single sequencing reaction can read all the way across most *Sau*3AI fragments.

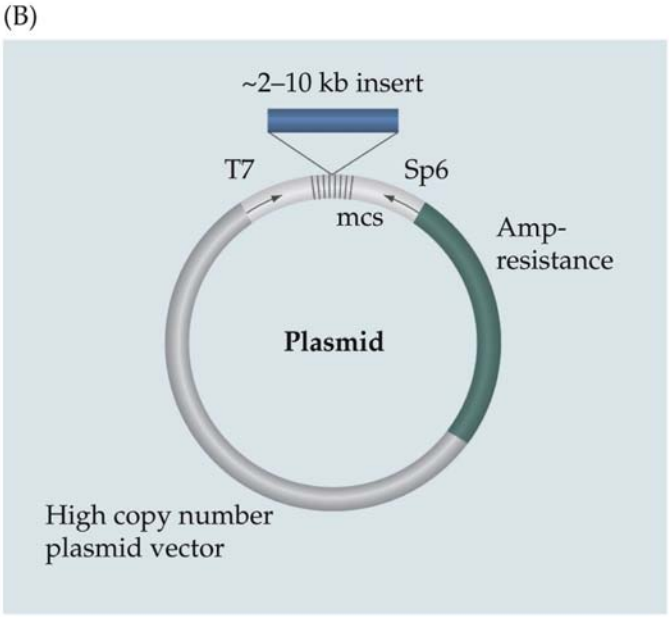
## Shotgun ligation is used to clone millions of recombinant DNA fragments in one experiment




(A)




***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.8 (Part 1)**




***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.8 (Part 2)**

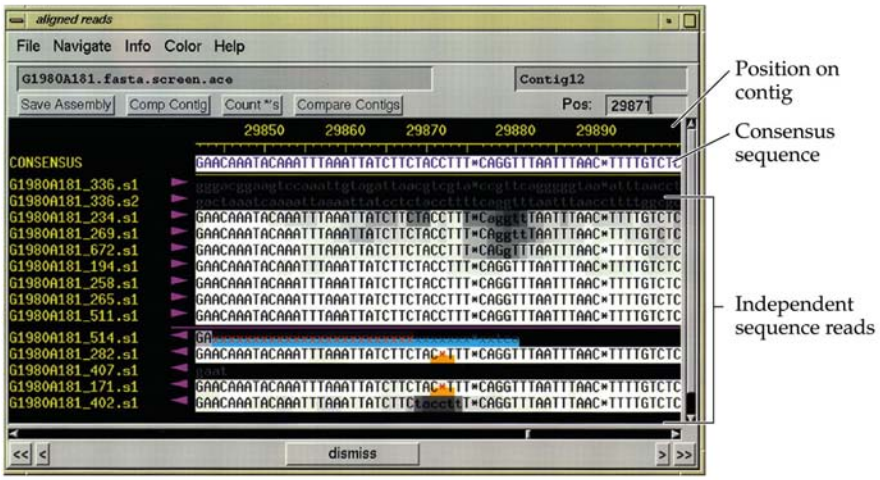


## Sequencing shotgun clones: a few practical considerations

- Identifying specific clones is time-consuming and expensive. But simply cloning random DNA fragments is very fast and can be done on a large scale (thousands per person per day).
- All of these shotgun clones can be sequenced with the same protocol, and the same primers, so there is no real need to identify which clone you are working with.
- In order to join the shotgun sequences to each other, additional clone libraries (i.e. cut with a second restriction enzyme, or cut with a partial restriction enzyme digest, or cut by random mechanical shear) must be used.
- You are going to have to sequence all the DNA, on both strands, anyway, so it really doesn't matter where you start ("We'll start the war right here.").
- Coverage of 5-10x is needed for a shotgun strategy to approach completion (think about a dart board…). But this effort is not wasted, because this level of redundancy is needed for error correction purposes.

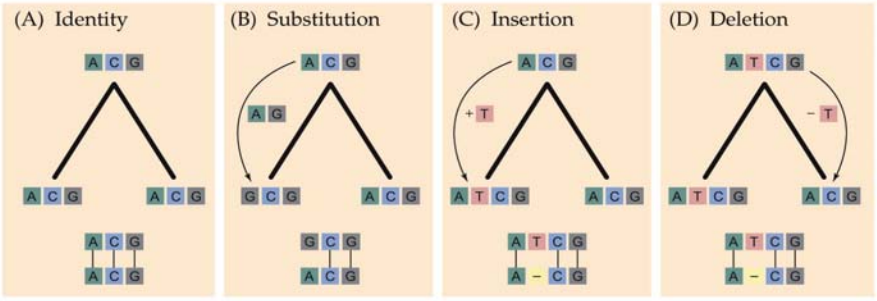
## Automated sequence alignment to build contigs

- Rigorous sequence alignment programs search both strands (why??), in all possible relative alignments (plus all possible gaps in both sequences for each relative alignment). They are searching for the alignment with the highest identity score (minus penalties for gaps and base substitutions). Thus a partial match is always found, even in unrelated sequences.
- Local alignment programs such as BLAST speed up this search by restricting their attention to an empirically-determined subset of the alignments and gaps. They also return an E-value, which is an empirically-derived estimate of the number of such matches that would be found in the sequence database by chance.
- Sequencing projects typically enforce an E-value cutoff, such that matches below a certain E-value are interpreted as overlapping sequences and joined into larger sequence "contigs".
- After shotgun sequencing to 5-20x, the data typically consists of several large contigs, separated by gaps of unknown length and orientation (why??).
- These gaps can be closed with custom oligonucleotide sequencing strategies.




**A PRIMER OF GENOME SCIENCE 3e, Figure 2.4**

Box 2.1, Figure A Common evolutionary events and their effects on alignment




***A PRIMER OF GENOME SCIENCE 3e*, Box 2.1, Figure A**



**EXERCISE 2.2** *Computing an optimal sequence alignment*

Compute the best possible alignment for the following two sequences, assuming a gap penalty of –5, a mismatch penalty of –1, and a match score of +3. Would your answer be any different if the gap penalty was also –1 (rather than –5)?

AGCGTAT and ACGGTAT

**ANSWER:** Three possible high-quality alignments are:

```
AGCGTAT        AGC-GTAT        AGCG-TAT
|··||||  and   |-|-||||  and   |-||-|||
ACGGTAT        A-CGGTAT        A-CGGTAT
```

The second and third alignments will produce the same score with these penalties, namely (6 × 3) – (2 × 5) = 8 with a gap penalty of –5; or (6 × 3) – (2 × 1) = 16 with a gap penalty of –1. By contrast, the first alignment gives a score of (5 × 3) – (2 × 1) = 13. Consequently, the first alignment is best with a large gap penalty, but either of the other two alignments would be better with a small gap penalty.

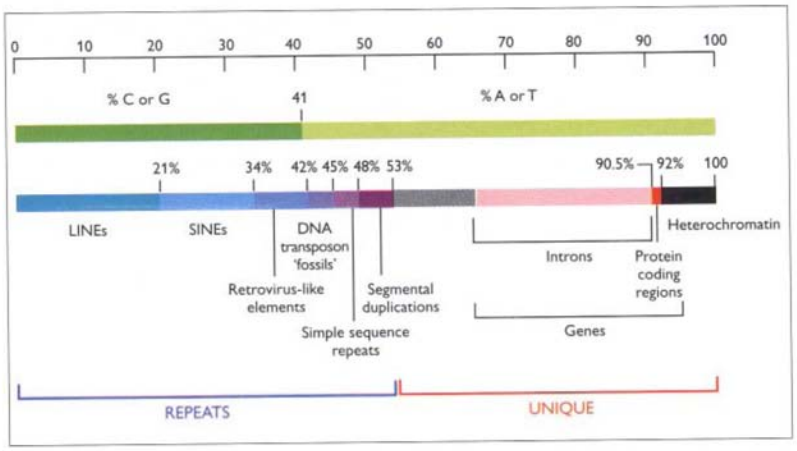
## Sequencing with custom oligonucleotide primers

- Custom oligonucleotide primers are short (~20 bases), synthetic, single-stranded DNA molecules.
- Custom oligonucleotides are synthesized in the 3' to 5' direction, on a solid-phase support, in an anhydrous environment.
- The cost and time required to synthesize a custom oligonucleotide has come down steadily. A typical commercial cost is now roughly $0.50 per base, with a turn-around time of about 48 hr.
- The sequence data from the end of a sequencing run can be used to design a custom oligonucleotide, which can then be used in a second sequencing reaction to generate an overlapping sequence (Hood).
- This is the exact opposite of shotgun sequencing: one large clone is sequenced with many primers (vs. many small clones sequenced with one primer).

## Some practical considerations re: sequencing with custom oligonucleotides

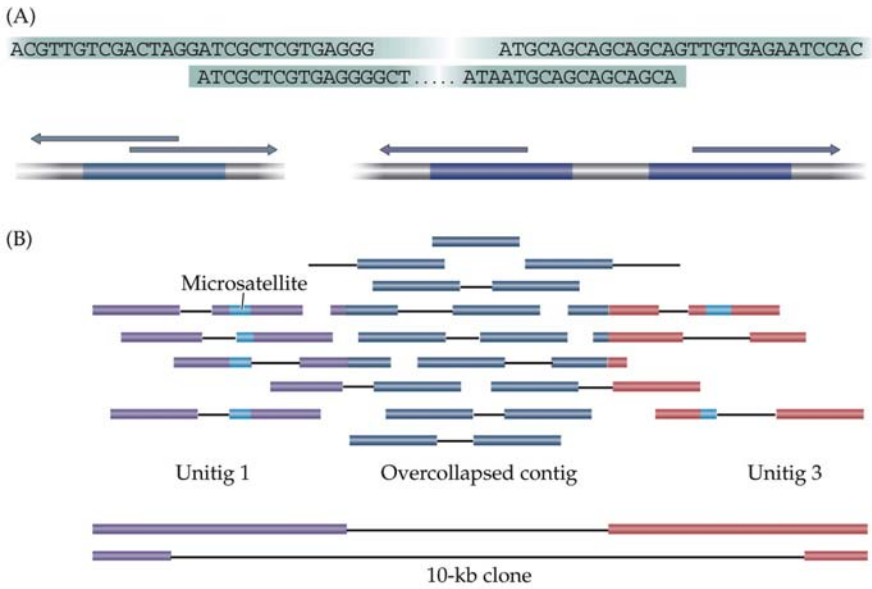
- It does use machine time efficiently, because overlap between sequencing reactions is optimal.
- It does not use calendar time efficiently, because each sequencing reaction has to wait on the results of the previous sequencing reaction, plus time for designing oligos (which is non-trivial) and ordering them.
- Additional reads are often necessary to generate sufficient overlaps, sequence the second strand, resolve discrepancies and low-confidence portions of the sequence data, read through repetitive DNAs, etc., giving about 2-4 fold coverage. This reduces the advantage of oligos.
- On a large scale, sequencing with custom oligos is expensive. On a large scale, costs matter.
- Most genome projects therefore use a combination of sequencing strategies: shotgun sequencing in the first ~2/3 of the project (because its cheaper and faster), followed by sequencing with custom oligos (to specifically target the gaps in the shotgun sequence data).

## The DNA sequence content of the human genome




## Sequencing through repetitive DNAs

- Shotgun contig assembly is facilitated by computer software that screens out known repeat sequences (RepeatMasker).
- Short, imperfect repeats can be handled with conventional shotgun sequencing methods, provided that the sequence reads are long enough (long capillaries), and that the thresholds for contig assembly are set high enough.
- Medium-length perfect repeats can sometimes be sequenced across with custom oligonucleotides. Longer perfect repeats require end-sequencing of specific clones (directed deletions, mate-pair clones).
- In many cases, chromosome regions that are particularly rich in repeats (rRNA arrays, centromeric heterochromatin, etc.) have simply been skipped by genome projects. They are more expensive to sequence and of less biomedical interest.
- Some sequencing errors related to repeats have probably slipped through all of the genome projects.




***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.11**



## Prokaryotes vs. Eukaryotes

- Prokaryotic genomes are relatively small (0.5 - 5.0 Mb) and contain almost no repetitive DNA. Thus the most cost-effective way to sequence bacterial genomes is by random shotgun sequencing across the entire genome.
- Random whole-genome sequencing of eukaryotic genomes was difficult at first, but is now cost-effective due to increases in computing power and the availability of related genomes to aid contig assembly.
- Novel and/or difficult genomes are assembled by "hierarchical" shotgun sequencing of individual BACs (bacterial artificial chromosomes) and PACs (P1 phage clones). Higher-level sequence assembly into "contigs" relies heavily on physical mapping methods to arrange the BACs in the correct order.
- BAC inserts are typically 100-300 kb in length. BACs can be ordered by fingerprinting, by FISH, by radiation hybrid mapping, or by classical genetic mapping. Usually a combination of several of these methods are used.

## Genomic DNA: purification for making libraries

- Purity of genomic DNA is important because it affects the frequency of various artifacts
- But, length of DNA is critical if you are to succeed in making a large-insert library. Simple manipulations (such as pipetting) can shear DNA to ~20 kb in length, so elaborate, high-purity protocols can not be used.
- Compromise protocols typically involve embedding mammalian cells in agarose plugs.
- Cells can be lysed within the agarose matrix, followed by a partial restriction enzyme digest, ligation to the vector, and electroporation into E. coli cells.

## Agarose plugs are used for preparing high m.w. DNA

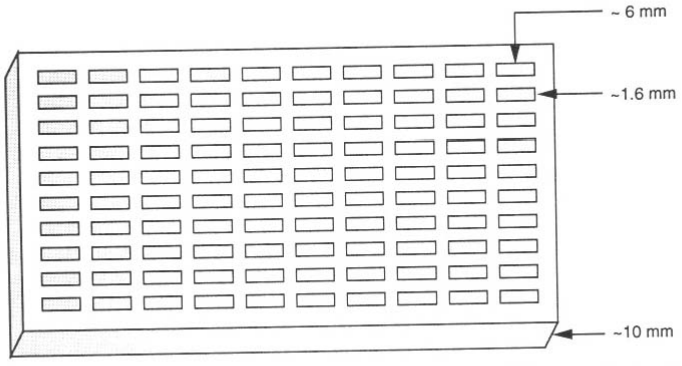
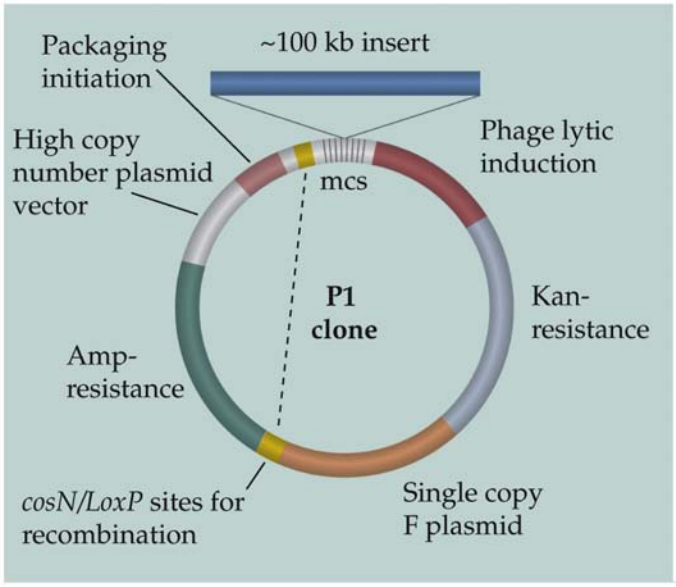



**Figure 2.** A 100-slot Perspex plug mould used for preparing DNA embedded in LGT agarose. The approximate dimensions of the slots are 6 mm × 1.6 mm × 10 mm resulting in ~ 100 μl agarose plugs.

## P1 artificial chromosome vectors (PACs)

- P1 is a bacteriophage that infects *E. coli* and has a genome size of ~110 kb.
- Only ~20 kb of the P1 genome consists of essential genes, and so P1 phage vectors can hold ~90 kb.
- P1 phage replicate as a low copy-number plasmid, then replicates as a high copy-number concatemer that is packaged into phage heads by a "head-full" mechanism similar to that of lambda.
- P1 artificial chromosome (PAC) vectors dispense with most of the "essential" genes. Special features of PAC clones allow the initiation of phage packaging during the cloning step, excision of high-copy number plasmid sequences after bacterial infection, and induction of replication immediately prior to harvesting of DNA.
- PAC vectors use the cre-lox system to force circularization (by site-specific recombination between lox sites). Inserts are typically 100-200 kb.
- PAC DNA preps provide a better yield and purity than BACs, because of the ability of P1 phage vectors to induce as a lytic (high copy-number) replicon.

(C)

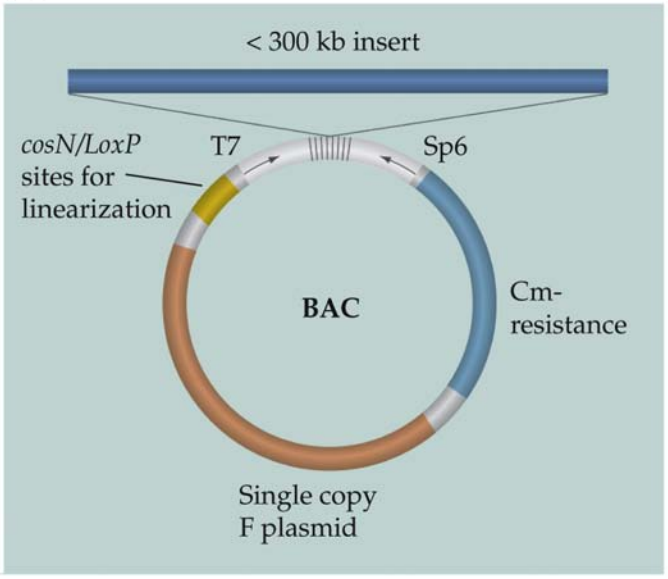



***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.8 (Part 3)**

## Bacterial artificial chromosomes (BACs)

- Bacterial artificial chromosome (BAC) vectors have the following components:
  - *E. coli* origin of replication (single copy)
  - *E. coli* selectable marker
  - Site-specific recombination system (cre-lox), to force circularization of large inserts
- Some BAC libraries have inserts as large as 2.5 Mb, but typical insert sizes are in the 200 - 300 kb range.
- Most clones are genetically stable.
- DNA preps work better than in YACs.
- Most of human genome project was based on BACs, although many researchers now prefer PACs.

(D)




***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.8 (Part 4)**

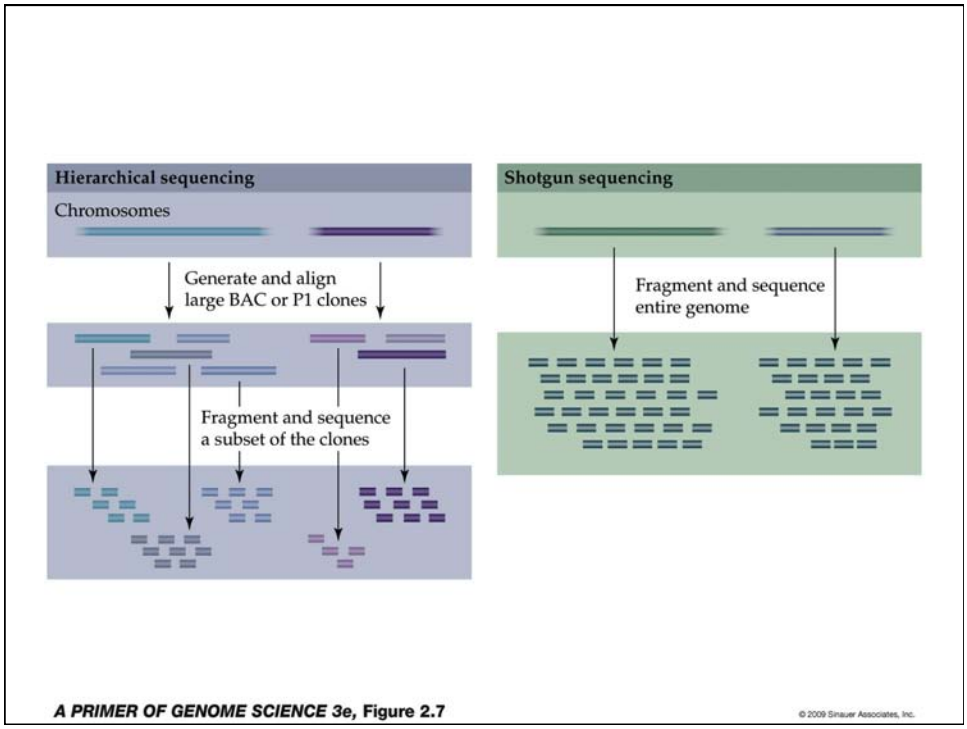
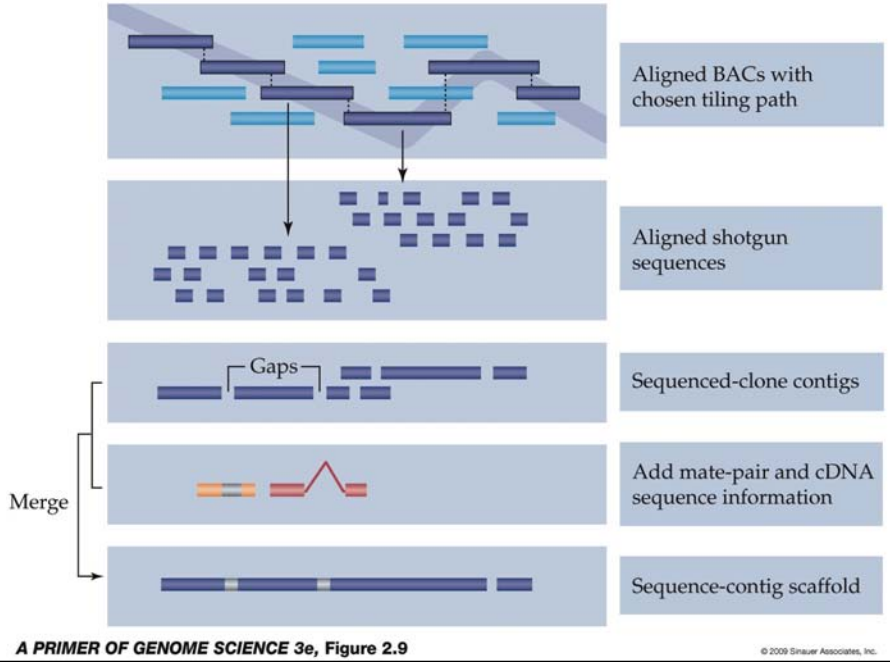
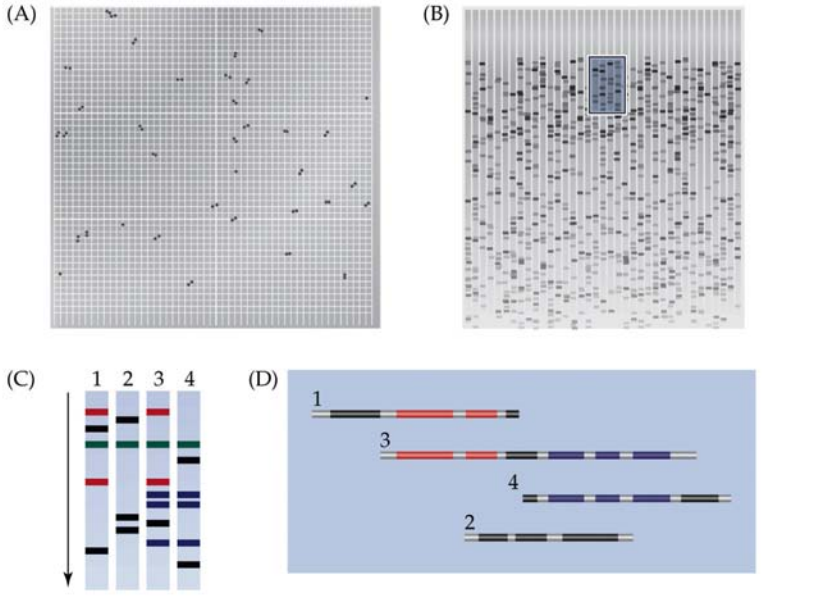



Figure 2.9 Hierarchical assembly of a sequence-contig scaffold



## Aligning BAC clones by hybridization and fingerprinting




*A PRIMER OF GENOME SCIENCE 3e*, Figure 2.10



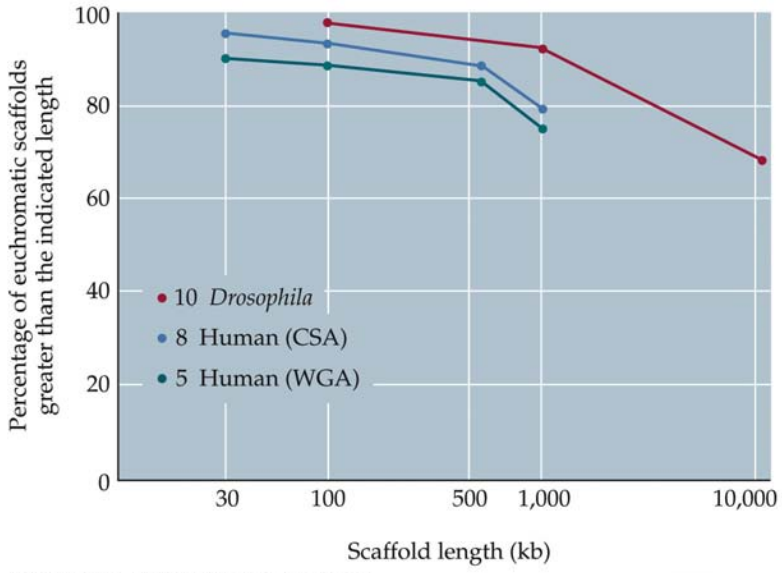
## The problem of chimeric clones (two inserts joined together)

- The frequency of chimeric clones can be limited by limiting the concentration of insert DNA during the ligation.
- Most chimeric clones can be prevented by partially end-filling the sticky ends of insert and vector. This has the advantage of increasing the ligation efficiency.
- Another strategy for eliminating chimeric clones involves phosphatasing the insert DNA. This has the disadvantage of decreasing the ligation efficiency.
- Sequence assembly is another line of defense against chimeric clones (any given joint should be observed in more than one ligation to be considered confirmed). Note that a single clone can be recovered more than once per ligation.
- The bottom line is that chimeric clones should be rare in assembled sequences, but can not be taken for granted.

## "Unclonable" DNA sequences

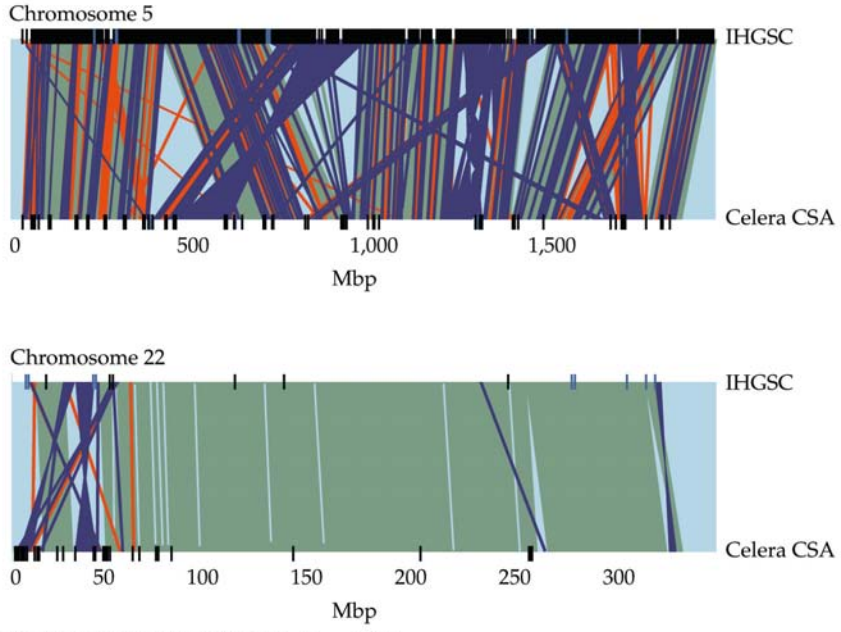
- Virtually every cloning system is incapable of cloning certain DNA sequences. Some of the problems include:
  - Sequences that cause the expression of toxic genes
  - Sequences that delete themselves, due to a high frequency of recombination
  - Sequences that cause (or experience) a high frequency of mutation
  - Sequences that provoke host defense mechanisms
- Many of these problems were historically fixed by straightforward *E. coli* genetics, such as using *r-m-* hosts, *recA-* hosts, and/or *recBC-* hosts.
- Nevertheless, long inverted repeats, AT-rich sequences, and sequences that form unusual secondary structures (triple helixes, Z-DNA, etc) are still highly unstable in modern *E. coli* host strains. These sequences did contribute to errors in the draft human sequence.
- Transformation-associated recombination (TAR) allows such sequences to be cloned in yeast, where they are more stable (Larionov).

## The initial assembly of human and fruit fly genomes at 5x to 10x coverage.




*A PRIMER OF GENOME SCIENCE 3e*, Figure 2.12

## Comparison of two draft human genome assemblies




***A PRIMER OF GENOME SCIENCE 3e*, Figure 2.13**



## Discussion Questions

- Briefly describe the steps involved in conventional shotgun sequencing projects (shotgun subcloning, cycle sequencing with fluorescent dideoxy nucleotides, and automated sequence assembly). Why is this approach so efficient?
- Briefly describe at least three methods of dealing with repetitive DNA in genome sequencing. Why is this a significant problem for most eukaryotic (but not prokaryotic) genome projects?
- How do physical maps differ from cytological or recombinational maps? Which is most relevant to genome projects?
- Briefly describe how BAC (and P1) fingerprinting works. How can fingerprinting be used together with FISH to build physical genome maps?