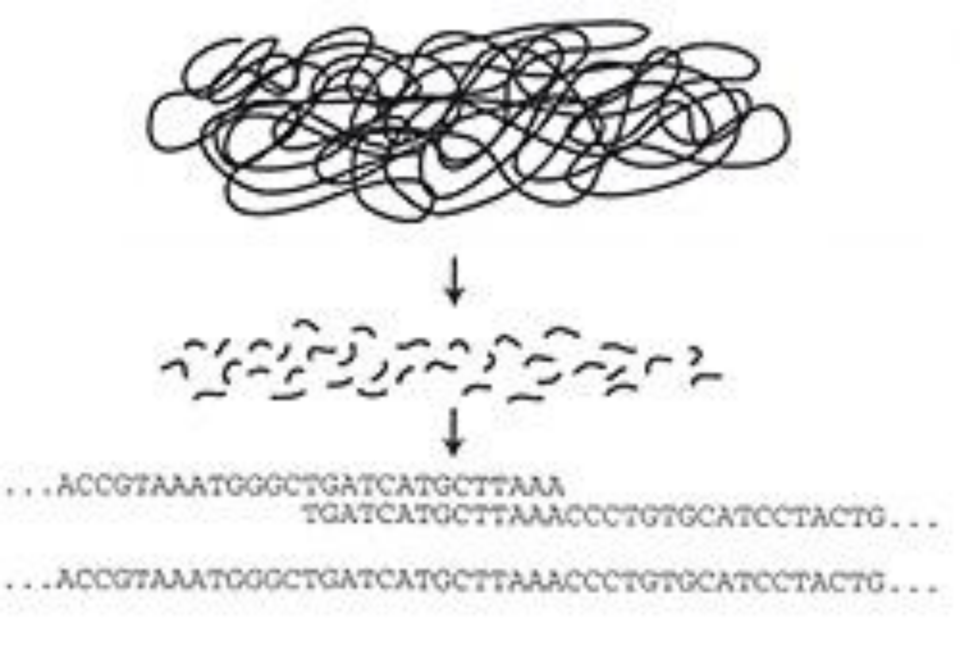
BENG 183

Trey Ideker

Genome Assembly and Physical Mapping


...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Reasons for sequencing

- Complete genome sequencing!!!
- Resequencing (Confirmatory)
  - E.g., short regions containing single nucleotide polymorphisms (SNPs) or other mutations
- Gene sequencing
  - Or associated upstream and downstream control regions (e.g., promoters, enhancers, intron splice sites)
  - cDNA sequencing and Expressed Sequence Tags (ESTs)

What sequencing methods (e.g. Sanger, pyro, SBH) are best suited for each of these scenarios?

Complete genome sequencing: ***Why bother?***

(For instance, why not just sequence expressed genes as cDNAs? Expressed genes constitute <5% of the entire human genome!)

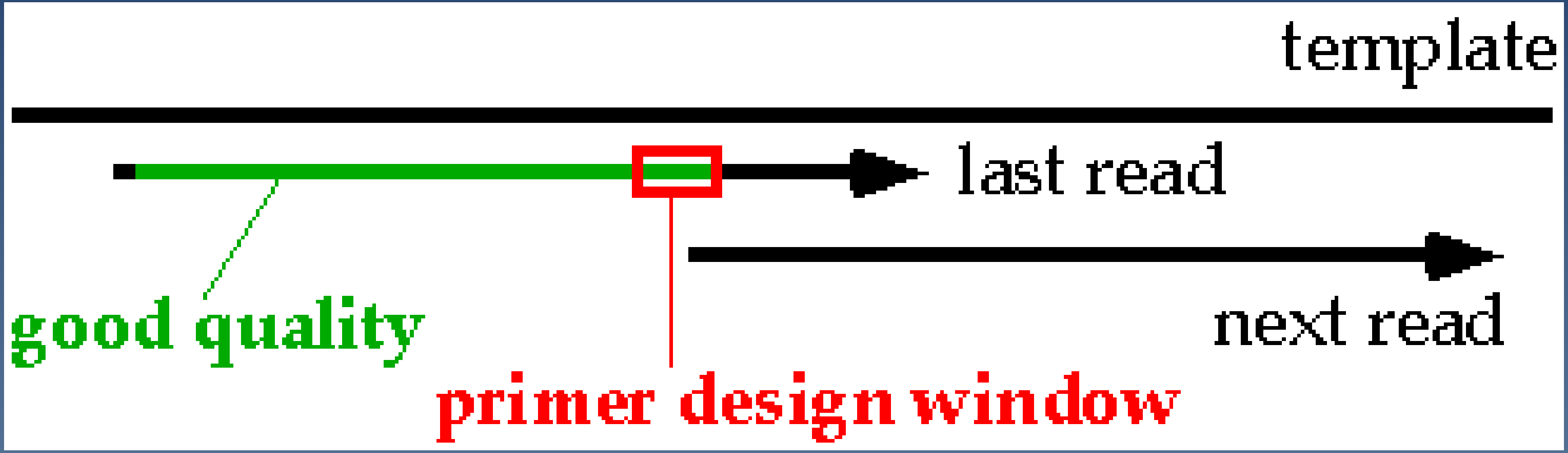
HOWEVER:

- Control elements not sequenced
- Many genes expressed at low levels
- Some genes difficult to recognize
- The remaining 95% of ‘junk’ DNA may have some as yet unknown but important function

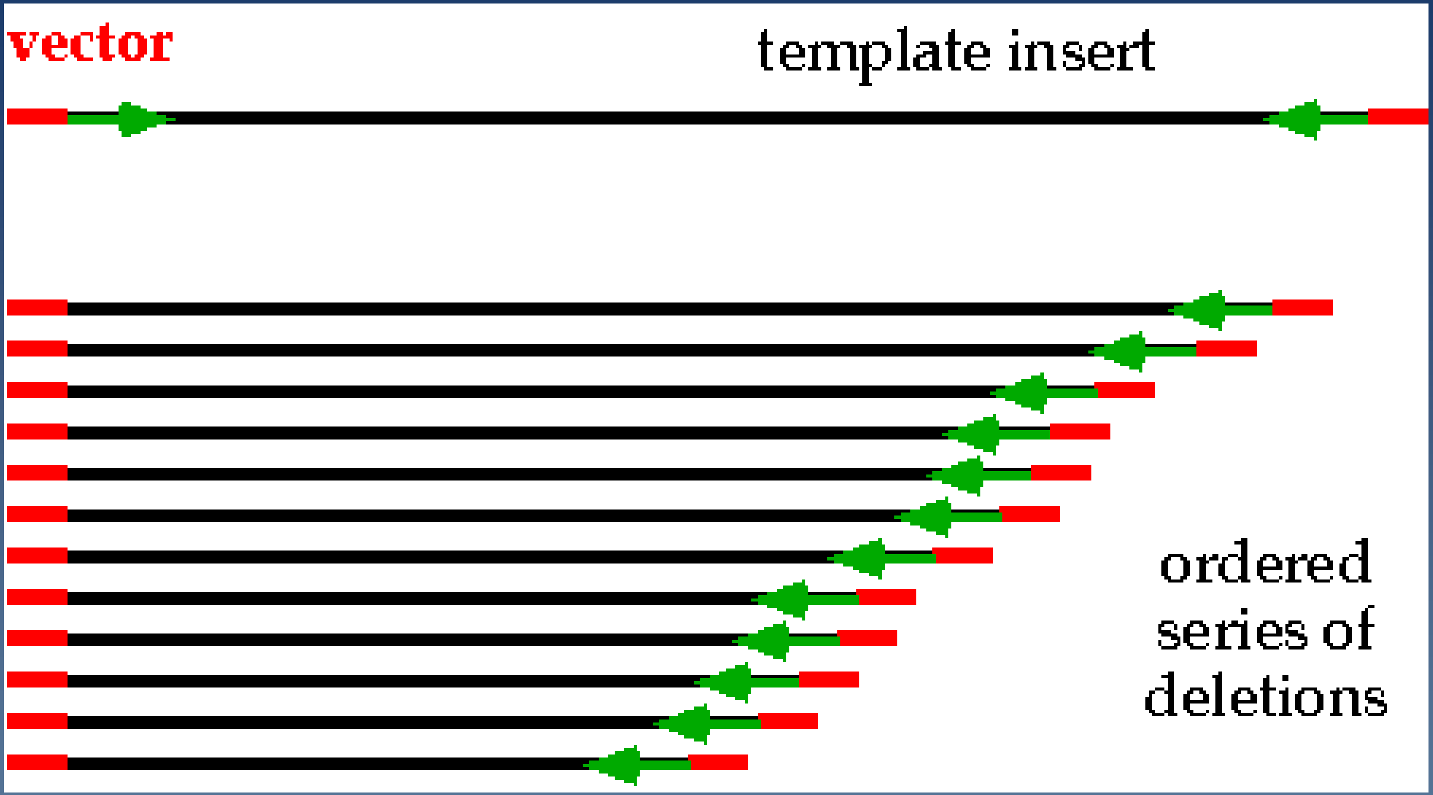
Sequencing DNA fragments > 1 kb

- A typical run produces a maximum of 600-800 bp of good DNA sequence.
- To sequence larger fragments:
  - Nested deletions
  - Primer walking
  - Subcloning and physical mapping
  - Shotgun cloning and assembly

Primer walking



Nested deletions



(1) Exonuclease mediated

(2) Transposon mediated

Strategies for Long-range and Genome Sequencing

The problem:

Sequencing reads are limited to 500 to 1000 bps. This is only partly handled by ***nested deletions*** and ***primer walking***

The 'shotgun' solution:

By oversampling, many reads can be assembled into a single target sequence. There are two competing strategies for this:

1) Clone-by-clone hierarchical approach

2) Whole-genome shotgun sequencing

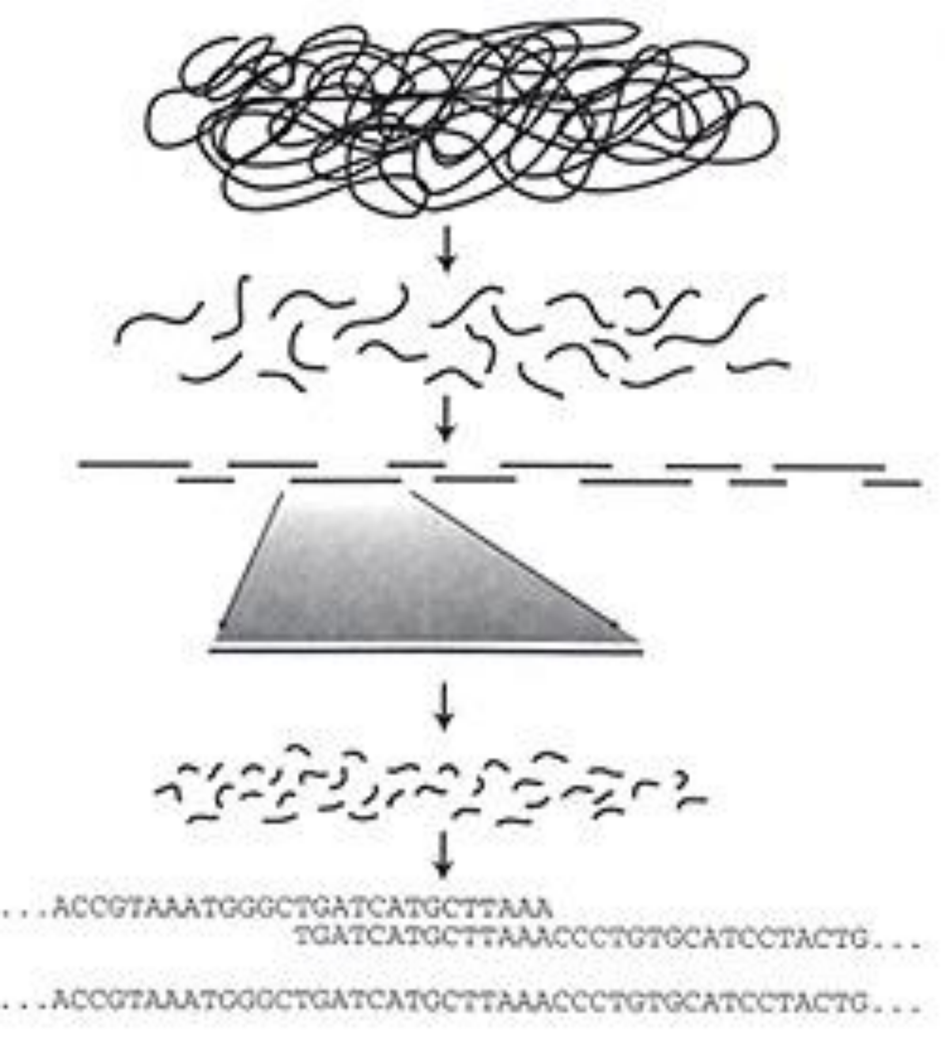
These two approaches sparked a huge debate…

Clone-by-clone (hierarchical) shotgun approach

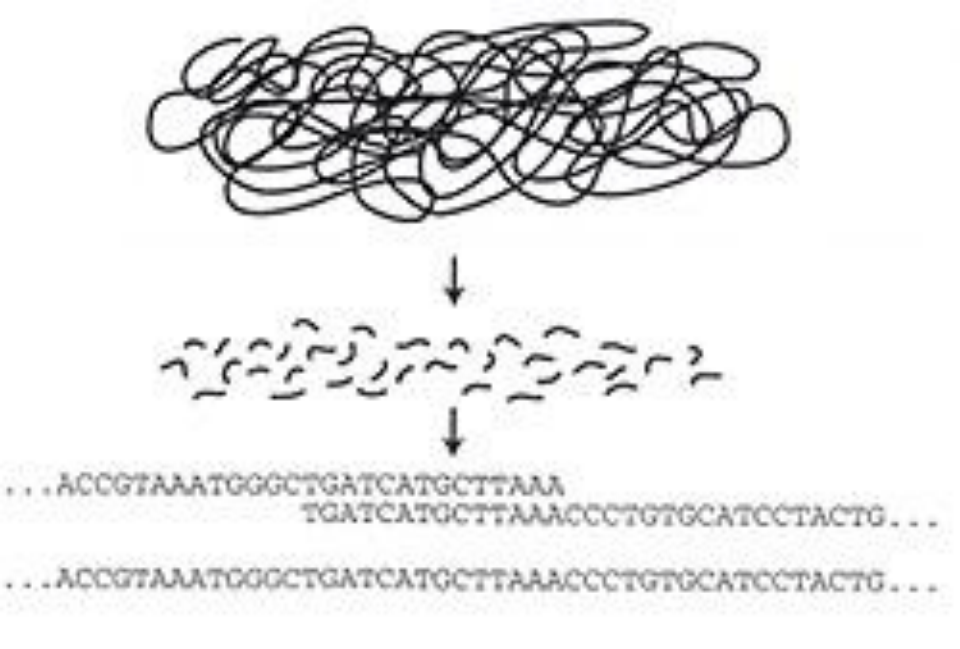
- Whole genomes are sequenced by first cloning large pieces into a set of overlapping cosmids, then ordering cosmids by *physical mapping*.
- Each ordered cosmid is sequenced by *shotgun sequencing,* i.e., random sub-cloning of fragments into vectors for sequencing.
- Sequences are pieced together for each cosmid using assembly software such as *PHRAP* (remember *PHRED*?).
- Remaining ‘gaps’ in a cosmid or between cosmids are closed using primer walking or probing of Southern blots to identify new fragments

Comparison of two sequencing approaches

HIERARCHICAL (1990)


...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

GENOME-WIDE SHOTGUN (1998)


...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Sizes of sequencing vectors

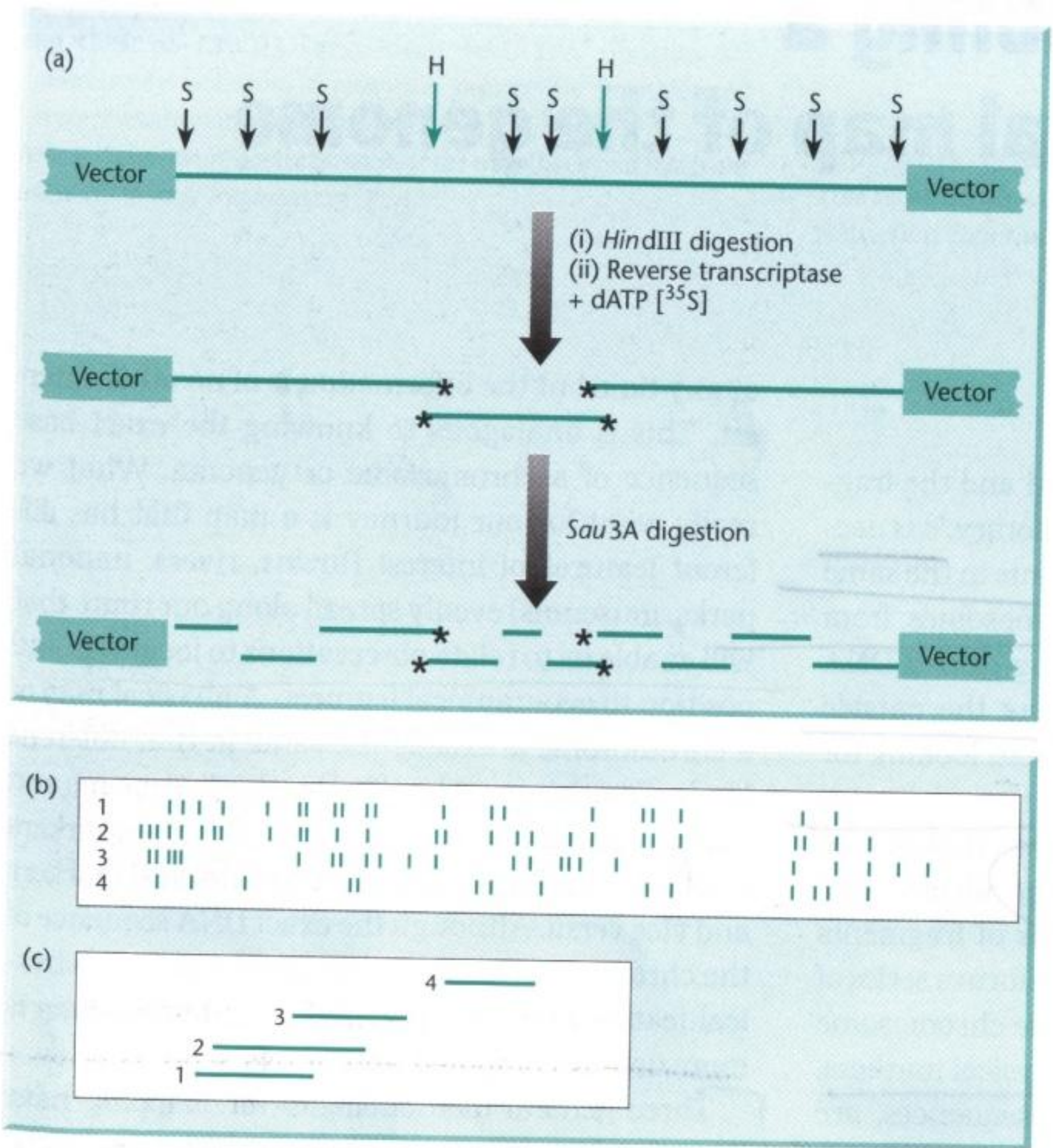
Vector	**Size (approx.)**
Whole chromosome	250 MB
YAC	1500 KB
BAC	150 KB
Cosmid	40 KB
Plasmid	5-10 KB
M13	1 KB

Methods for physical mapping

Mapping method	**Experimental resource**	**Breakpoints**	**Markers**
Fingerprinting	Library of clones	Endpoints of clones	Restriction sites or STSs
Hybridization mapping	Library of clones	Endpoints of clones	Whole clones or STSs
In situ hybridization (cytogenetic map)	Chromosomes	Cytological landmarks	DNA probes
Optical mapping	Chromosomes	Restriction fragments	Restriction sites or STSs
Radiation hybrid	Human/rodent fusion cells	Radiation-induced chromosome breaks	STSs
Genetic linkage (meiotic)	Pedigrees	Recombination sites	DNA polymorphisms

Table 4.2; Primrose and Twyman 3rd Edition 2003

Fingerprinting with restriction enzymes



FPC software (Fingerprinted contigs)

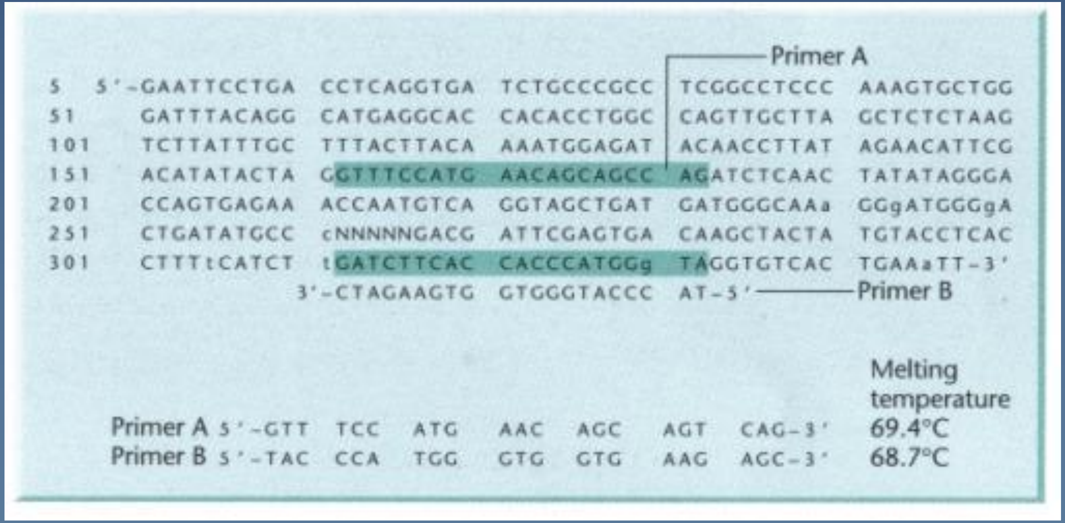
Fig. 4.1 The principle of restriction-fragment fingerprinting. (a) The generation of labelled restriction fragments (see text for details). (b) Pattern generated from four different clones. Note the considerable band sharing between clones 1, 2 and 3 indicating that they are contiguous whereas clone 4 is not contiguous and has few bands in common with the other three. (c) The contig map produced from data shown in (b). (Adapted and redrawn with permission from Coulson *et al*. 1986.)

From Primrose and Twyman 3rd Edition 2003

Sequence Tagged Sites (STS’s)

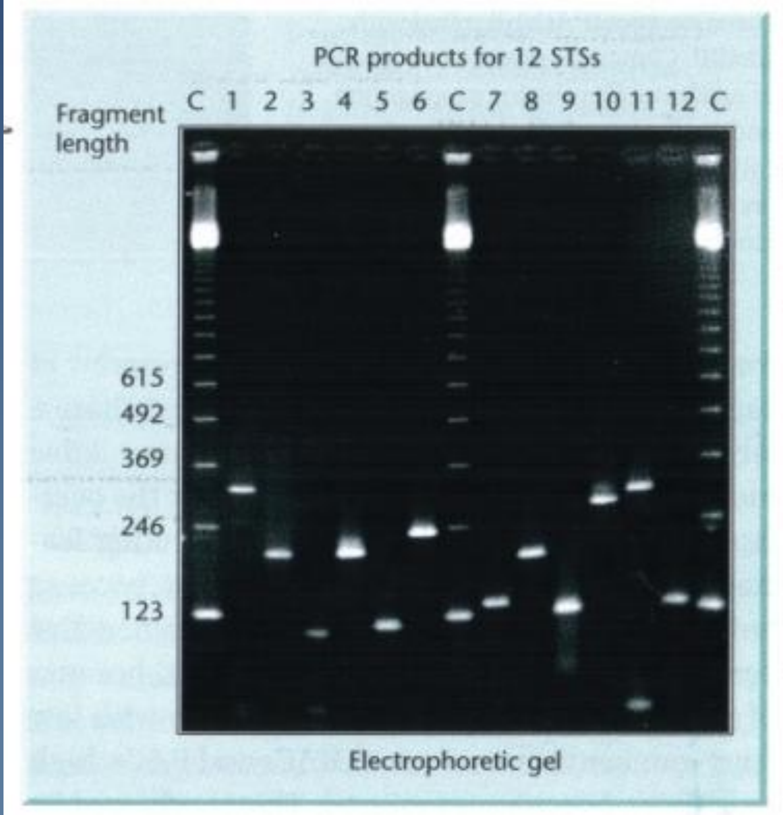
An STS is a primer pair that amplifies a unique region of the genome (i.e., produces a single PCR band– see below right).

These can also be used for identifying overlapping cosmids, i.e. fingerprinting…



```
                                                  ┌──────Primer A
5    5'-GAATTCCTGA  CCTCAGGTGA  TCTGCCCGCC │ TCGGCCTCCC  AAAGTGCTGG
51      GATTTACAGG  CATGAGGCAC  CACACCTGGC │ CAGTTGCTTA  GCTCTCTAAG
101     TCTTATTTGC  TTTACTTACA  AAATGGAGAT │ ACAACCTTAT  AGAACATTCG
151     ACATATACTA  GGTTTCCATG  AACAGCAGCC   AGATCTCAAC  TATATAGGGA
201     CCAGTGAGAA  ACCAATGTCA  GGTAGCTGAT   GATGGGCAAa  GGgATGGGgA
251     CTGATATGCC  cNNNNNGACG  ATTCGAGTGA   CAAGCTACTA  TGTACCTCAC
301     CTTTtCATCT  tGATCTTCAC  CACCCATGGg   TAGGTGTCAC  TGAAaTT-3'
              3'-CTAGAAGTG  GTGGGTACCC   AT-5'───────Primer B

                                                         Melting
                                                         temperature
Primer A 5'-GTT  TCC  ATG  AAC  AGC  AGT  CAG-3'         69.4°C
Primer B 5'-TAC  CCA  TGG  GTG  GTG  AAG  AGC-3'         68.7°C
```



Figures 4.3 and 4.4; Primrose and Twyman 3rd Edition 2003

Hybridization mapping

Clones hybridizing to repeats are pre-screened and removed

End-sequences of clones can be used as STS's

From Primrose and Twyman 3rd Edition 2003

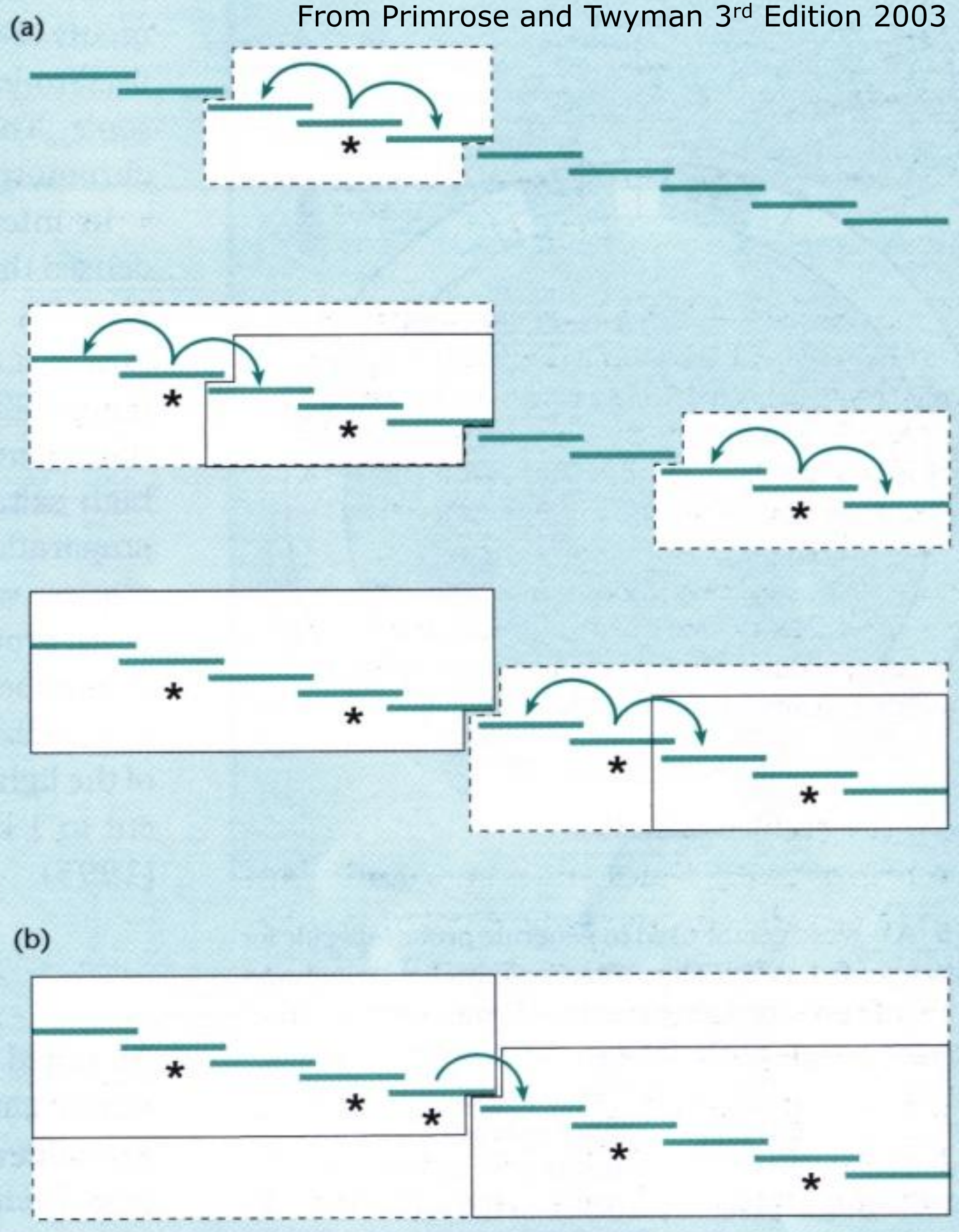


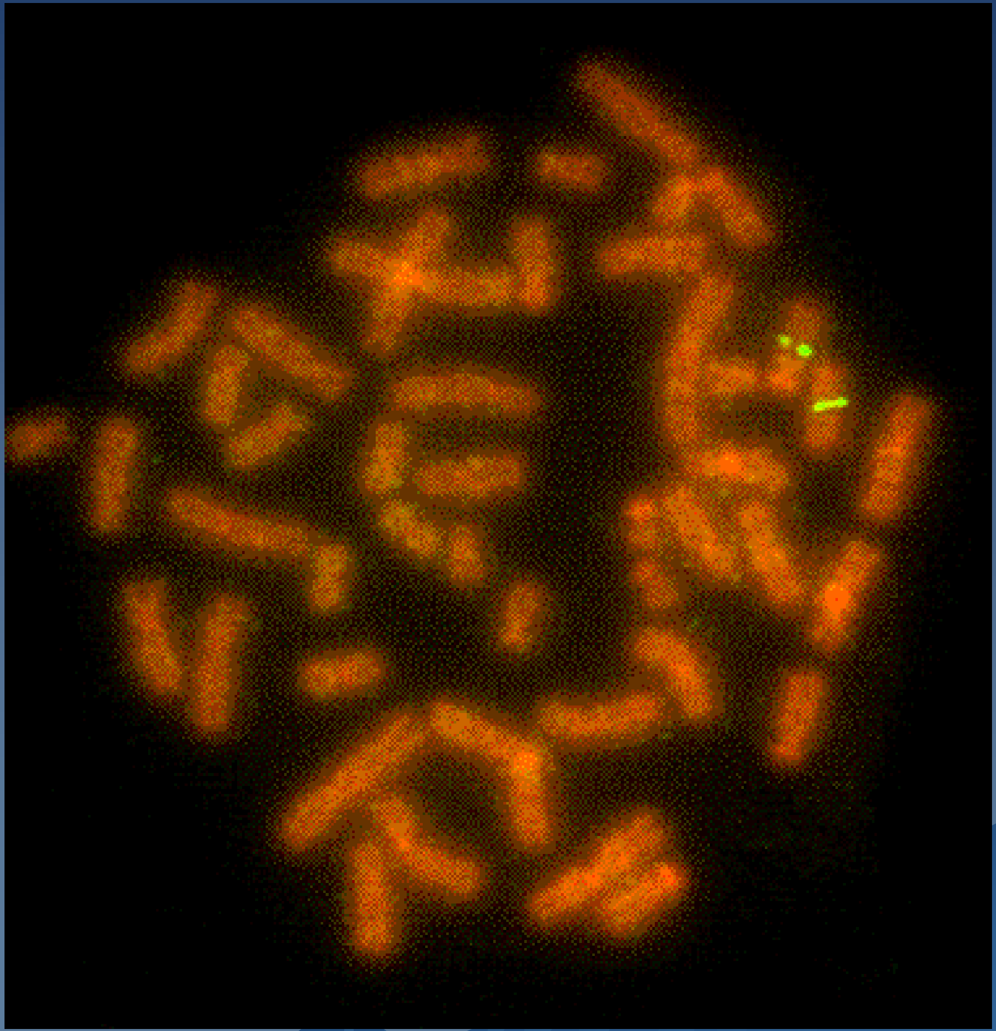
Fig. 4.14 The principle of hybridization mapping. (a) Clones for use as probes are randomly picked (*) from a given set of cosmids whose map order is not known. Hybridization identifies overlapping clones (arrows). From clones that do not give a positive signal in any earlier hybridization assay (unboxed areas), probes for the next round of experiments are chosen until all the clones show positive hybridization at least once. (b) Gaps in the map caused by the lack of probes for certain overlap regions are closed by using terminal contig clones. (Reprinted from Hoheisel 1994 by permission of Elsevier Science.)

Cytogenetic mapping

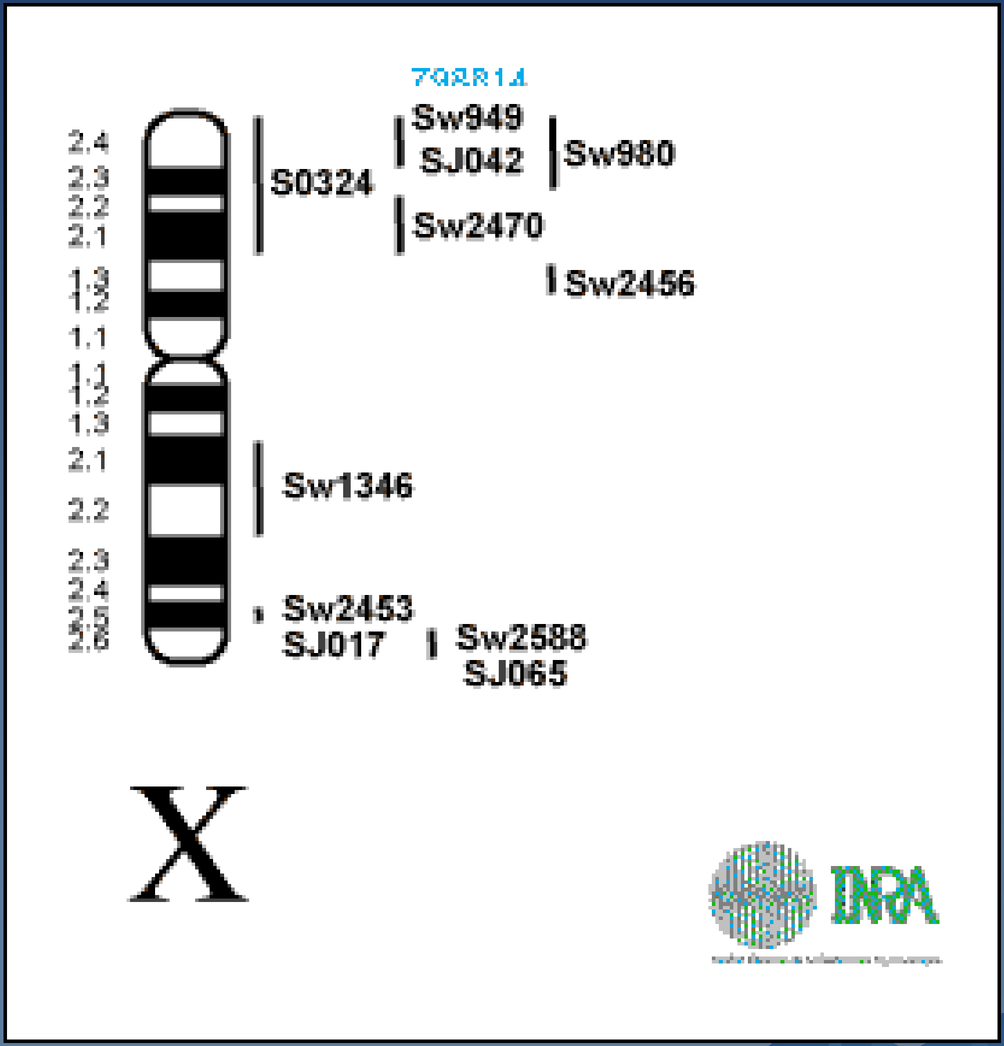
Uses FISH: Fluorescent *In-Situ* Hybridization

Clones are fluorescently labeled and hybridized directly to metaphase plates

Repeats and duplicated genes cause problems

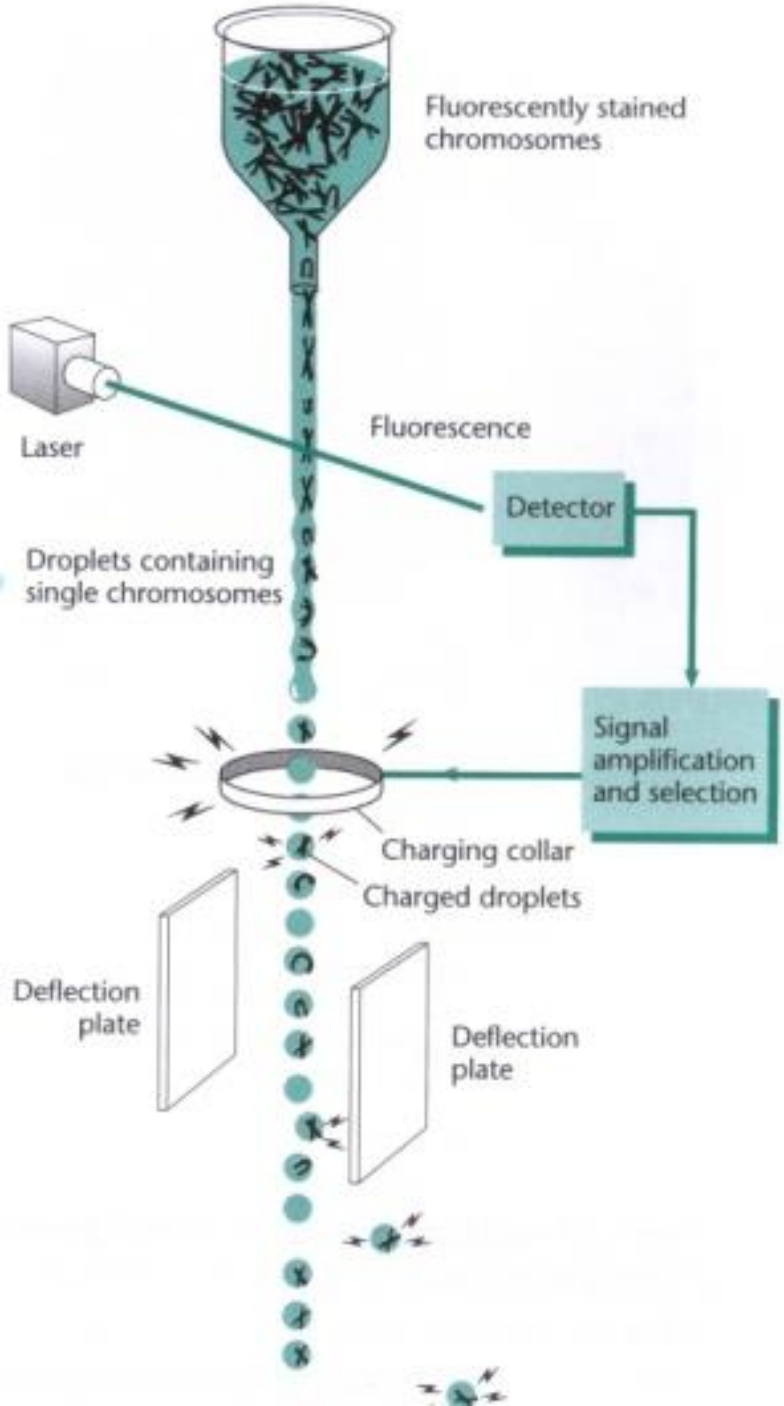


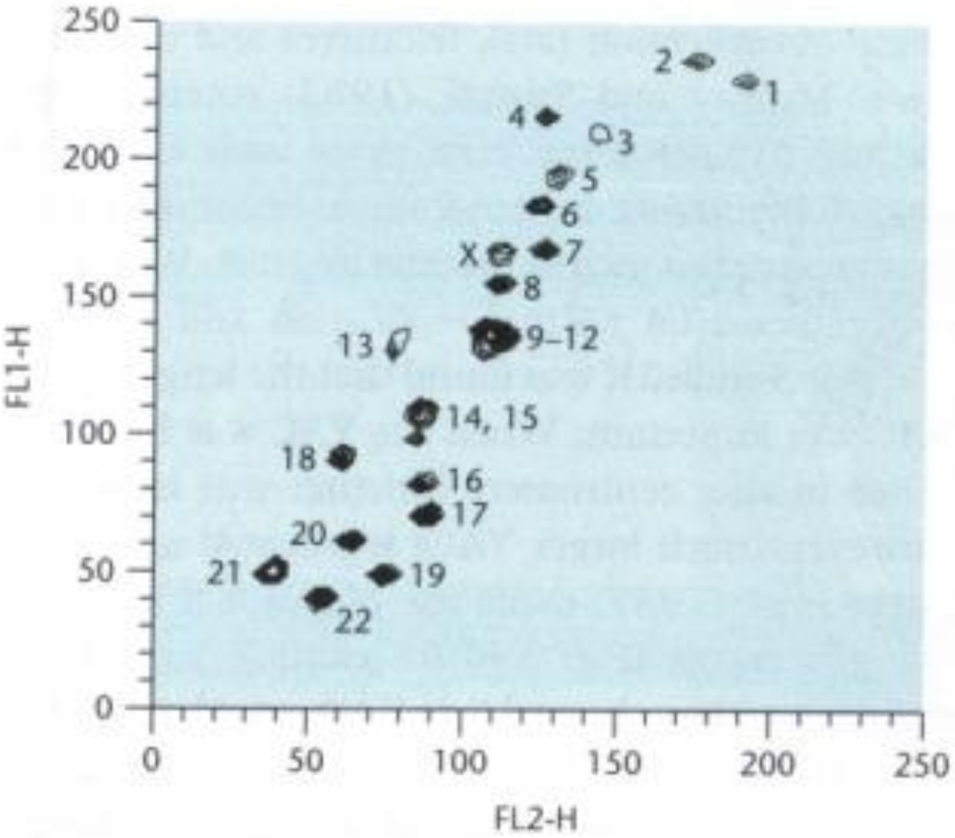
Example pig cytogenetic map



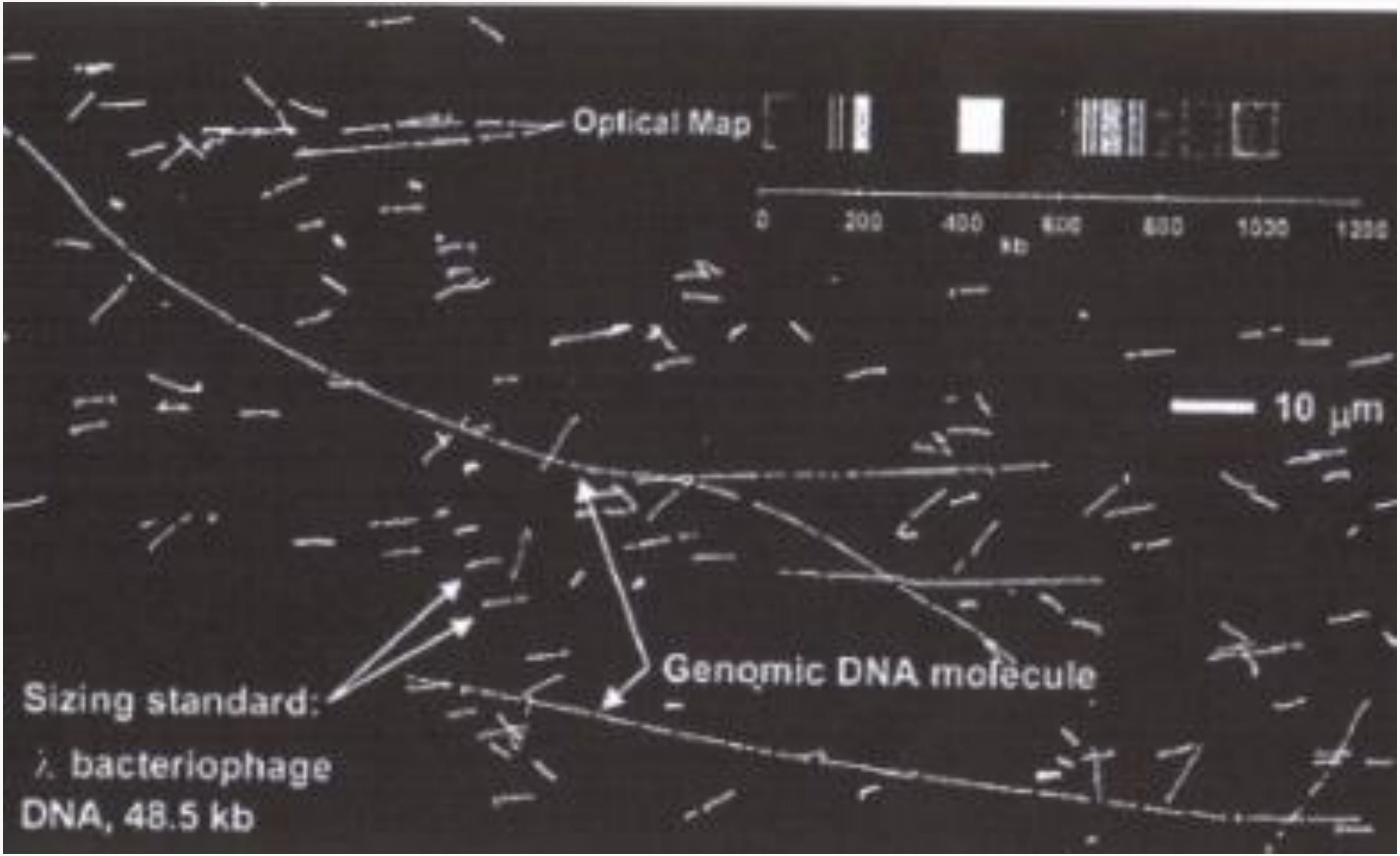
Separating chromosomes with Fluorescence-Activated Cell Sorting (FACS)

(Figs. 3.4 and 3.5 from Primrose and Twyman 3rd Edition 2003)


Fluorescently stained chromosomes
Laser
Fluorescence
Detector
Droplets containing single chromosomes
Signal amplification and selection
Charging collar
Charged droplets
Deflection plate
Deflection plate



Optical Mapping



From Primrose and Twyman 3rd Edition 2003

Radiation Hybrid (RH) Mapping

(1) Human cells are X-ray irradiated to fragment chromosomes

(2) These fragments are introduced into rodent cells

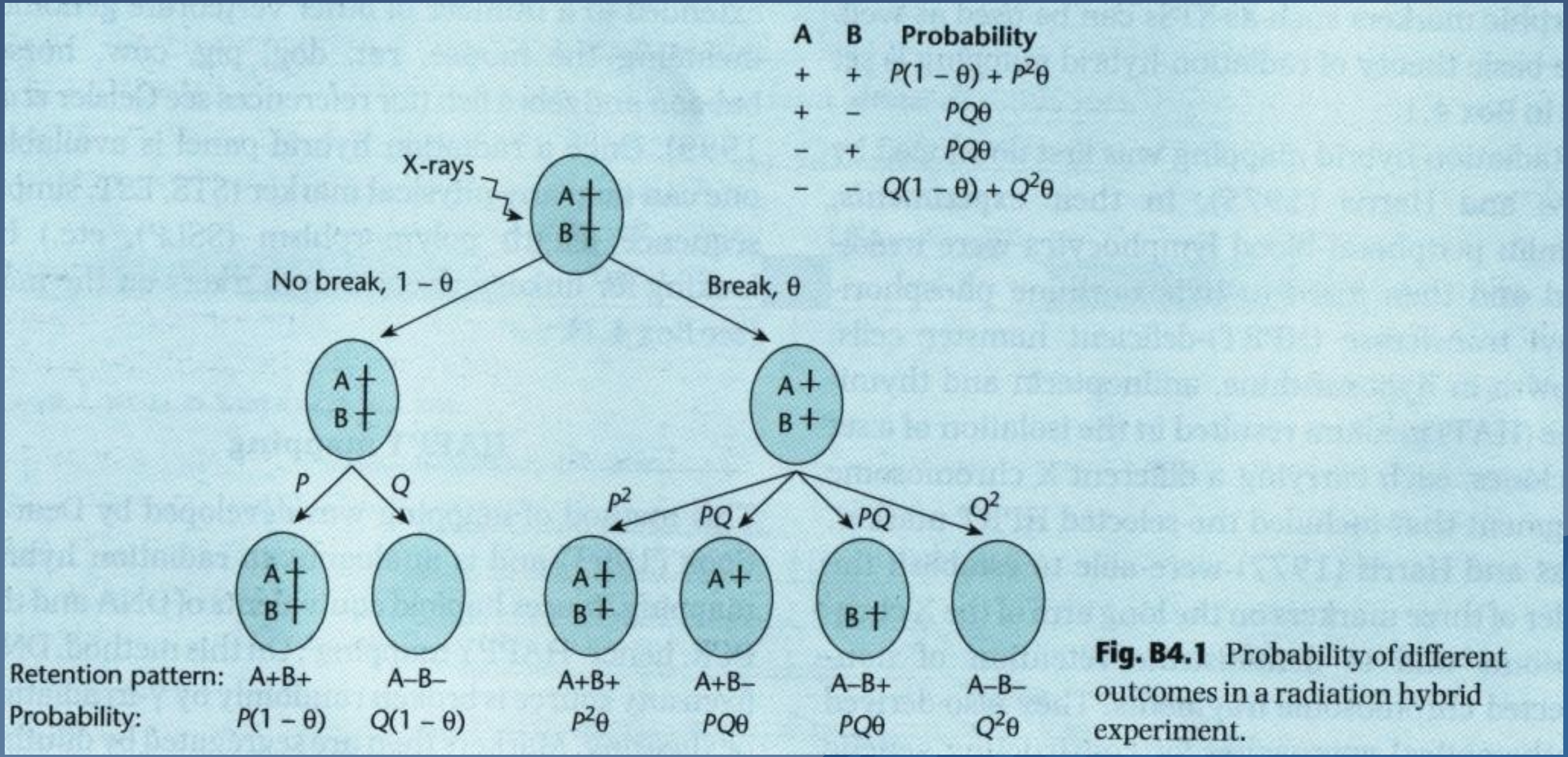
(3) Human/rodent hybrids lose human chr. fragments randomly

(4) Screen hybrids for markers A and B

θ = prob. of breakage between A and B (i.e. separate fragments)

P = probability that a DNA fragment is retained; $Q = 1-P$

Rel. distance between A and B = $-\log(1-\theta)$ Rays



A	B	Probability
+	+	$P(1-\theta) + P^2\theta$
+	–	$PQ\theta$
–	+	$PQ\theta$
–	–	$Q(1-\theta) + Q^2\theta$

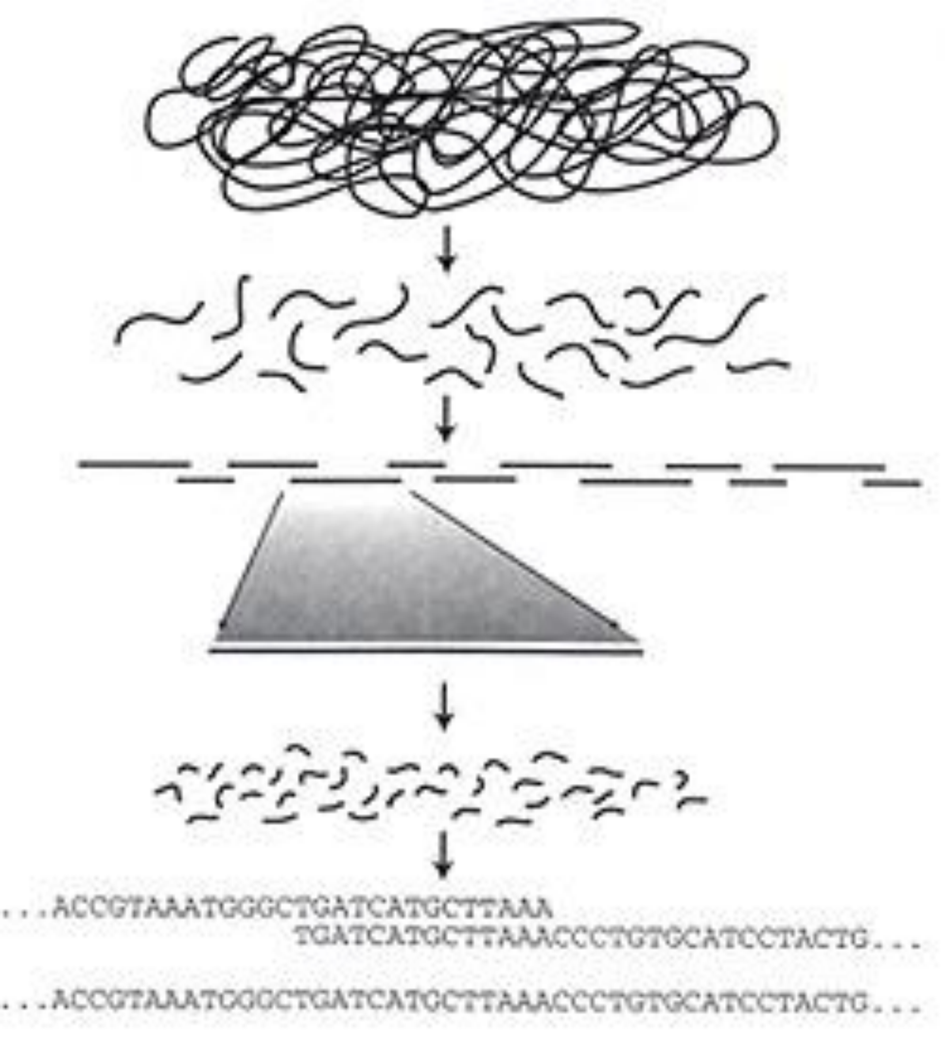
Fig. B4.1 Probability of different outcomes in a radiation hybrid experiment.

Moving from hierarchical assembly to the whole-genome shotgun approach

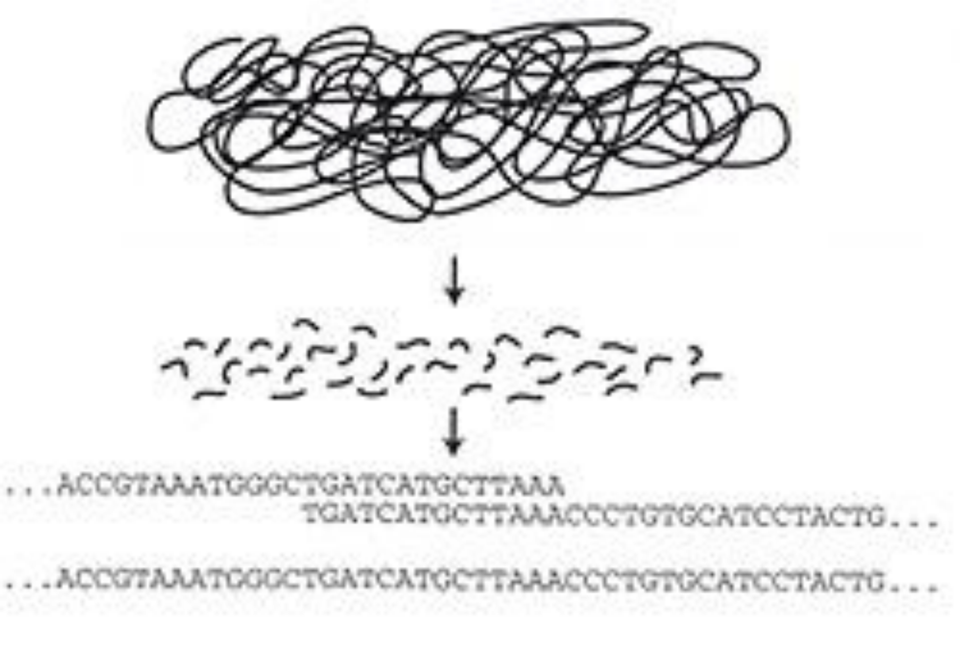
- Little or no physical mapping required to hierarchically order clones
- Previously thought that cosmids were the upper size limit for shotgun sequencing
- This idea was destroyed when Fleischmann et al. (1995) determined the sequence of *Haemophilus influenzae* through a pure shotgun approach (no cosmid intermediary)

Comparison of two sequencing approaches

HIERARCHICAL SHOTGUN (1990)


...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

GENOME-WIDE SHOTGUN (1998)


...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

Sequence Assembly Algorithms

- Start from an initial sequence fragment (<1kb) chosen at random
- Chose the second fragment as having the best overlap with the first based on DNA sequence
- The overlap is based on strict match criteria specifying minimum length of match, max. length of unmatched segment, and the min. percentage of matching nucleotides
- A set of overlapping sequence reads is called a *contig*.
- Examples are PHRAP, Arachne, TIGRassembler

Even with all this sophistication, sequencing is still work

Statistics for the genome sequence of *Haemophilus influenzae*:

- 1,830,137 bp of DNA in total
- Full shotgun approach producing DNA fragments 1.6-2.0 kb in length.
- 28,643 sequencing reactions
- 24,304 give useful & high-quality sequence
- Assembled directly into 140 contigs
- 8 technicians, 14 automated sequencers
- Total amount of time: 3 months

Genome sizes and coverage

Organism	Year	MB sequenced	% coverage (total)	% coverage (euchrom.)
S. cerevisiae (yeast)	1996	12	93	100
C. elegans (nematode worm)	1998	97	99	100
D. melanogaster (fruit fly)	2000	116	64	97
A. thaliana (flowering plant)	2000	115	92	100
Human chr. 22	1999	34	70	97
Human genome (consortium)	2001	2693	84	90
Human genome (Celera)	2001	2654	83	88-93

Reprinted from Table 5.2 of Primrose and Twyman