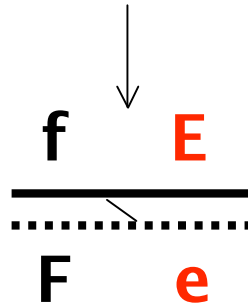
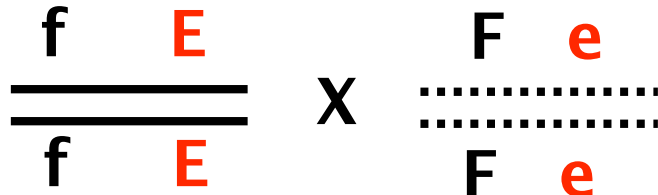


Mapping with a CAPS marker



E/e: CAPS marker
F/f: a mutation causing abnormal flowers
Ler: solid line
Col: dotted line

self
 ↓

No. of plants

No. of
recombinant Chr.

11 $\frac{f}{f} \frac{E}{E}$ 0

6 $\frac{f}{F} \frac{E}{e}$ 1

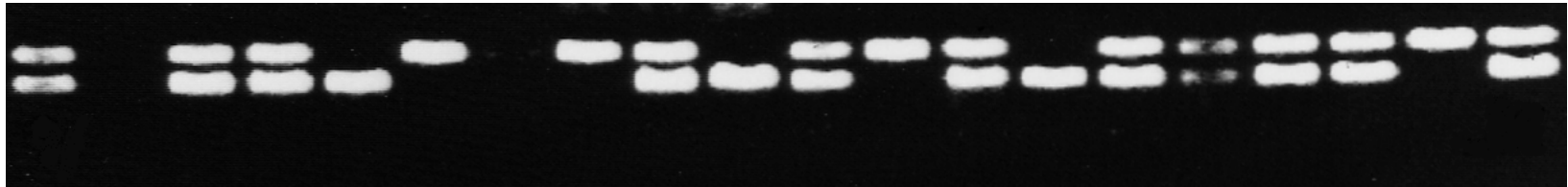
0 $\frac{F}{F} \frac{e}{e}$ 2

Recombination Frequency

$$\text{Recombination Frequency} = \frac{\text{Total \# of recombinants}}{\text{Total Progeny}}$$

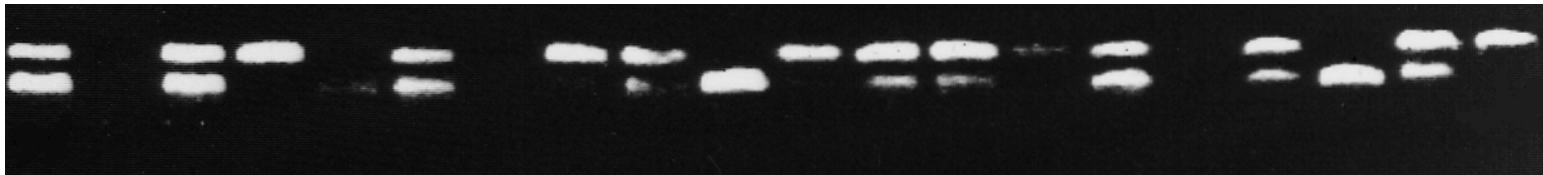
C

E
e



$$\text{C1: } \frac{10 + (3 \times 2)}{17 \times 2} = \frac{16}{34} = 47.0\% = \text{UNLINKED}$$

C

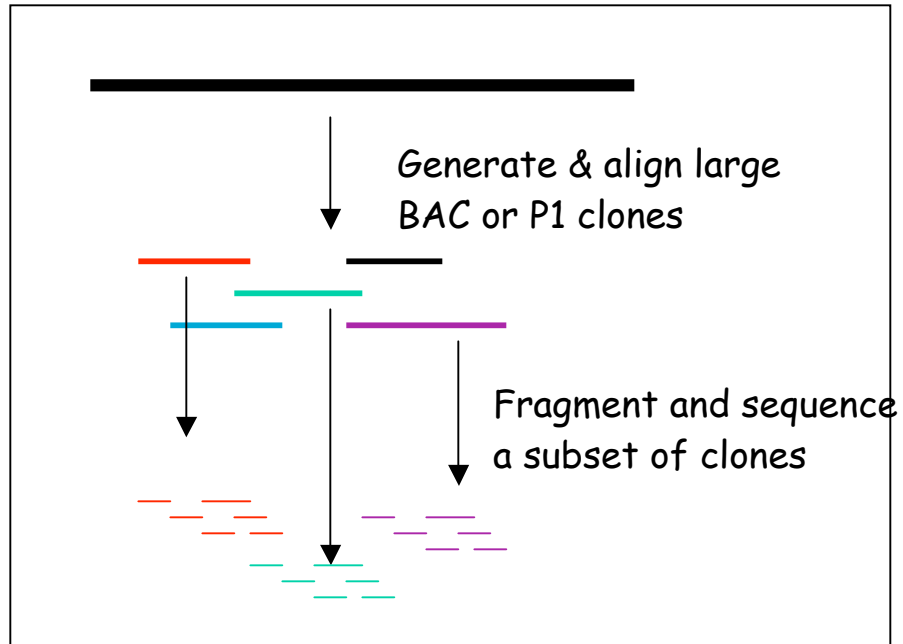
M
m

C2: $\frac{8 + (3 \times 2)}{17 \times 2} = \frac{14}{34} = 41.1\% = \text{Possibly UNLINKED}$

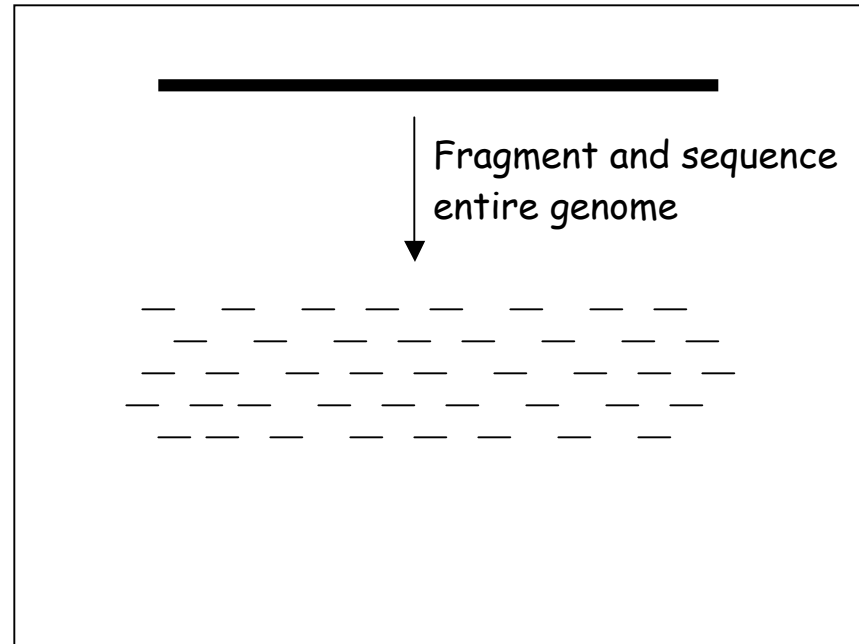
Lecture 10: Genome Analyses

1. Read our textbook: page 321-324
Table. 10-1
2. Exploring genomes
Web-based Bioinformatics Tutorial
by Paul Young

Hierarchical sequencing



Shotgun sequencing



Adapted from Fig. 2.7 Gibson and Muse

TABLE 10.1 A Comparison of the Developmental Complexity and Genome Features* of Model Organisms

Organism		Developmental Complexity	Genome Size* (Megabases)	Predicted Genes	# of Genes per Million bp Sequenced	Date Genome Finished
Type	Species					
Bacterium	<i>Escherichia coli</i>	1-cell prokaryote	4.64	4200	905	1997
Archaeobacterium	<i>Halobacterium</i> sp. NRC-1	1-cell	2.57	2630 (2411)	1023	2000
Yeast	<i>Saccharomyces cerevisiae</i>	1-cell eukaryote	12.07	5800	483	1996
Worm	<i>Caenorhabditis elegans</i>	~1000 cells	100	19,099	197	1998
Fly	<i>Drosophila melanogaster</i>	~50,000 cells	180	13,601	117	2000
Mustard weed	<i>Arabidopsis thaliana</i>	10 ¹⁰ cells	125	25,000	221	2000
Pufferfish (draft)	<i>Fugu rubripes</i>	10 ¹² cells	380	38,000	118	2002
Rice (draft)	<i>Oryza sativa</i>	5 × 10 ¹⁰ cells	Ind/jap 466/420	46,000–56,000/ 32,000–50,000	127–155/ 82–128	2002
Mouse (unfinished)	<i>Mus musculus</i>	10 ¹¹ cells	3200	40,000	8–11	2002
Human (draft)	<i>Homo sapiens</i>	10 ¹⁴ cells	3200	40,000–60,000	9–12	2002

*Haploid genome size, including heterochromatin.

Note: For the sequenced genomes of model organisms: Gene numbers are taken from the original sequence publications; most numbers have since changed slightly and different sources give different estimates, depending on protocols. For rice, two different strains have been sequenced: *Oryza sativa* L. ssp. *Japonica* and *Oryza sativa* L. ssp. *Indica*.

Genome Annotation

Ab initio Gene discovery

GeneFinder

Grail

Genie

Genscan

HMM gene

FGENES

Blast Search

BLAST: Basic Local Alignment Search Tool

Performs pairwise comparisons of sequences, seeking regions of local similarity rather than optimal global alignment between two sequences

```
ATGTTCTT CAACAAAGCC GCGGAGACAA ACTTCGTATT CCATCTCTCG ATTCCCACTT CCACTTTTAC CCTCCTCCTC
CTCCTTCCTC CGGCGGCGGA GGTGGCGTCT TTCCTCTCGC TGATTCCGAT TTCCTCGCAG CCGGTGGCTT TCACTCCAAC
AACACAACA ACCACATATC TAACCCTAGC TACAGTAATT TCATGGGATT TCTCGGTGGC CCTTCTTCTT CTTTCATCCAC
CGCAGTCGCC GTCGCCGGAG ATCATTCCTT TAACGCCGGA CTTTCTTCCG GAGACGTTCT TGTCTTCAA CCCGAGCCTC
TATCTCTATC TTTGTCCTCT CACCCTAGAC TCGCTTACGA TCTAGTCGTT CCCGGTGTTG TTAAGTCCGG ATTCTGTAGA
TCTGCCGGTG AAGCCAACGC CGCCGCCGTC ACCATCGCGT CTAGAAGCTC TGGTCCTCTC GGACCTTTCA CGGGCTACGC
GTCGATTCTT AAAGGATCAA GGTTCCTGAA ACCAGCACAG ATGCTTCTTG ATGAGTTTTG TAATGTGGGT CGTGGGATTT
ACACCGACAA AGTCATCGAC GACGATGATT CTTCTCTGCT TTTTGATCCG ACGGTTGAGA ATCTCTGCGG TGTTTCTGAT
GGCGGCGGAG GAGATAATGG AAAGAAAAAA TCAAACTCA TCTCCATGCT CGACGAGGT TACAAGAGGT ATAAGCAATA
CTATGAGCAG CTACAAGCTG TGATGGGATC ATTCGAATGC GTTGCGAGTC TCGGGCACGC TGCTCCGTAC GCTAACTTAG
CCTTGAAAGC ATTGTCTAAG CATTTCAAGT GTTTGAAGAA TGCTATAACG GACCAGCTTC AATTCAGCCA CAACAACAAG
ATCCAACAAC AACAACAATG TGGTCATCCG ATGAACTCTG AGAATAAGAC TGATTCTTTA AGATTTGGAG GAAGTGATAG
TTCTAGAGGC TTATGTTCTG CTGGTCAAAG ACATGGATTT CCTGATCATC ATGCTCCTGT TTGGAGACCG CACCGTGGCC
TACCCGAACG TGCTGTTACT GTTCTAAGGG CTTGGCTCTT CGATCATTTT TTGCATCCTT ATCCAACAGA TACAGACAAA
CTCATGCTGG CTAAGCAGAC AGGTCTCTCC AGAAATCAGG TATCGAATTG GTTCATAAAC GCAAGAGTTA GGGTTTGGAA
GCCGATGGTG GAAGAGATTC ACATGCTGGA GACTCGACAA TCTCAGAGAT CTTCTTCTTC CTCTTGGAGA GACGAACGTA
CTAGCACCAC CGTCTTCCCT GACAACAACA ACAACAACCC ATCTTCGTCC TCGGCACAGC AAAGACCTAA CAACTCATCT
CCGCCTAGAC GGGCACGAAA CGACGACGTT CATGGCACAA ACAACAACAA CAGCTATGTA AACAGTGGGA GCGGCTGCGG
TAGTGCGGTT GGTTCCTCGT ATGGAATTGG GTCGTCGAAT GTGCCGGTGA TGAATAGCAG CACAAACGGA GGAGTGTCTT
TGACGTTAGG GCTTCATCAT CAGATTGGGT TACCGGAGCC TTTTCCGATG ACAACTGCTC AGAGGTTTGG TCTTGATGGT
GGTAGTGGCG ATGGTGGTGG TGGGTATGAA GGGCAAAATC GTCAGTTTGG GAGAGATTTT ATTGGTGGTA GTAATCATCA
GTTTCTACAT GATTTTGTAG GTTGAGATTA TTTGTGTGGA AAGGAAAAAA TATGTTTGAC GTTTGGGTAT GTATAAGAAG
ATATGGGGGA ATTGAAATGC ATATGATGTG TATATTAGAA TGTTTCTTC
```

NCBI (<http://www.ncbi.nlm.nih.gov/>)

Searching sequence databases

Query: the submitted sequence

Genbank accession number: every sequence submitted to Genbank has an assigned number

E-value: probability of, by chance, obtaining a seq similarity as similar as the blast result.

Scores: based on scoring matrix, penalizes mismatches according to certain rules or seq alignment.

Blast result

Exploring genomes

Web-based Bioinformatics Tutorial

by Paul Young

www.whfreeman.com/young