Molecular and General Genetics – BIOLOGY 261, Lecture Notes

Lectures 1 and 2.

Mendelian Genetic Analysis – Chapter 2

1. MENDEL"s FIRST LAW – Independent assortment.

Mendel's first law applies to a cross to monitor the segregation of a single gene with two alleles.

Diploid organisms carry two copies of each gene. The genes can have variants and each variant is called an allele – (such as the blue and brown alleles of the human eye color gene). When organisms produce gametes (egg, sperm or pollen), each gamete receives one of the two gene copies, with equal chance (50/50). During fertilization two gamete at random and the fertilized egg has two copies. The pairing of the two gametes is random relative to the alleles that they carry. All possible combination of pairs the alleles from the intercross of two heterozygotes have an equal chance of occurring.

In practice leads to offspring with different phenotypes at predictable frequencies or ratios.

1.1. A cross between two heterozygotes:

Aa x Aa → AA, Aa, aa 1:2:1 genotypic ratio

3:1 phenotypic ratio.

Male Gametes \ Female Gametes	A	a
A	AA	Aa
a	Aa	aa

1.2. A cross between a heterozygote and a heterozygote.

Aa x aa → Aa, aa 1:1 genotypic ratio

1:1 phenotypic ratio.

If we know the genotypes of the parents we can predict the genotypic and phenotypic ratios of the offspring.

If we know the phenotypic ratio of the offspring, we can often decipher the genotypes of the parents.

2. TEST CROSS is a cross of a homozygous recessive individual (aa) to any other individual. It is very informative for deciphering the genotype of the other individual. Any allele present in the "non-tester" parent will be expressed in the offspring.

Aa x aa → 1:1 Aa : aa

3. PROBABILITY RULES

The product rule. The probability of two independent events occurring is the product of the two individual probabilities. For example, in the cross described in (1) above, Aa x Aa, the probability of having an individual receiving a recessive allele from its maternal parent is ½ and the probability of receiving a recessive allele from its paternal parent is ½ the probability of both these events happening is 1/2 x ½ = ¼ .

The sum rule. The probability of having of having one or the other of two mutually exclusive events is the sum of their individual probabilities. Consider again the cross described in (1) above, Aa x AA,

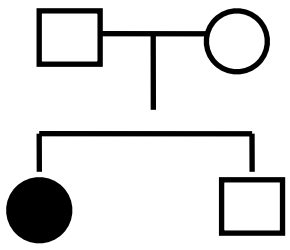
The probability an individual receiving a dominant allele from his maternal parent and recessive allele from its paternal parent is ½ x ½ = ¼ .
The probability of the opposite happening, a recessive allele from its mother and a dominant allele from its father is ½ x ½ = ¼.
The probability of the first case **or** the second case occurring is ¼ + ¼ = ½.

PROBABILITY AND RATIOS: "Expected segregation ratios", like 3:1, can be recognized in large populations. However, even in small families, you can think of expected ratios in terms as probabilities.
Example: If two individuals of "Bb" genotype mate and have one offspring there is a ¼, (i.e 25%) chance that the offspring will be "bb" genotype. The the frequency of a bb genotype will be 25%

3. THE PEDIGREE

For species that don't produce very large numbers of offspring from a single mating pair (such as human), familial relationships and inheritance of traits are often represented by pedigree diagram.

By analyzing pedigree, we can often decipher the genetic control of a trait. Squares represent males, circles represent females. The rows represent generations, in this example parents with a male and female offspring are represented.



The key elements of the pedigree to observe are:
1) An affected offspring resulting from the mating of two unaffected individuals indicates that this

trait is caused by a recessive allele. This also indicates that the parents were both heterozygotes.

2) It is a strong indication, though not proof, that the trait is controlled by a dominant allele if all affected individuals in a pedigree have at least one affected parent.

3) A high frequency of individuals being affected by a rare trait within a single family also suggests the effect of a dominant allele.

In pedigree analysis we often ask two questions: Is a trait controlled by a dominant or recessive allele? What is the probability of a certain individual being affected by a certain genetic trait?

4. THE SIMPLE LOGIC OF GENETICS

1) If we know that a trait is controlled by different of alleles of a single gene or two genes we can predict the inheritance of the trait based on probability.

2) The inverse is also true. If we can recognize a pattern of inheritance for a trait that is like a Mendelian ratio 3:1 or 1:1, we can conclude that the trait is controlled by a single gene or two genes.

For example we know that cystic fibrosis and Huntington's disease are controlled by single genes with at least one allele that causes the disease. This is known because in families that have one affected child, 1 out of the 4 of the other children will also be affected – when averaged over a reasonably large number of families eg.50 children total.

5. MINI SUMMARY: WHAT ARE GENES AT THE MOLECULAR LEVEL –. This summary material is to help you understand genes, but will be covered in detail in the second half of the course, after the midterm.

The molecular basis of inheritance.
Body is made up of cells, each cell has a nucleus which contains chromosomes. The genes are on the chromosomes. They are composed of DNA.

DNA is the template for RNA synthesis. RNA is translated into proteins. Proteins run the mechanics of the cell.

DNA $\longrightarrow$ RNA $\longrightarrow$ Proteins $\longrightarrow$ Development and Metabolism

Mutations are changes in DNA sequence that create different alleles of a gene. These often leads to non-functional proteins. The altered metabolism that results from mutations gives the varied phenotypes that we see. Mutations in genes that encode the enzymes that synthesize pigments lead to different colors found among individuals of a species. Altered proteins are the causes of many human genetic diseases.

6. CHROMOSOMAL BASIS OF INHERITANCE

Mendel's the laws of genetic inheritance were developed before we knew that genes are located on chromosomes. Now we understand that the nature of the segregation of alleles and random nature of fertilization is based on the behavior of chromosomes during meiosis, the cell division that gives rises to gametes.

In diploid organisms each cell has two sets of **homologous** chromosomes. Humans have 46 chromosomes, including 22 pairs of autosomal chromosomes and two sex chromosomes. The fruit fly, *Drosophila meanogaster* has 8 chromosomes, 3 pairs of autosomes and 2 sex chromosomes. The chromosomes are in pairs, and the pairs are nearly identical copies of each other. The differences are primarily small changes in DNA sequence that lead to the different alleles for genes.

Today we know that the segregation pattern of chromosomes during meiosis is the basis of gene segregation and assortment observed by Mendel. This was first hypothesized when scientists observed chromosomes under a microscope and recognized that chromosome segregation resembled gene segregation described by Mendel.

Important chromosomal vocabulary:

Chromosome	Centromere
Chromatid	Sister chromatids
Chromosome arm	Homologous chromosomes, also called homologous pairs of chromosomes.

There are two types of cell division:

6.1 Mitosis - cell division that occurs in somatic cells (all cells except those producing gametes). In mitosis, chromosomes replicate then divide in conjunction with cell division. This conserves the number of chromosomes during cell division.

6.2. Meiosis - a particular kind of cell division in reproductive cells which give rise to gametes, i.e. egg cells, sperm and pollen.
In meiosis the number of chromosomes becomes half of the number of a somatic cell; chromosome replication is followed by two cell divisions. The first cell division separates the members of each chromosome pair from each other; the individual chromosomes maintain their duplicated structures during this first division (meiosis I). The second division (meiosis II) splits each of the duplicated chromosomes into two daughter chromosomes.
The pairing of homologous chromosomes and their segregation into separate gametes accounts for the segregation of alleles of a gene during gametogenesis. The random assortment of genes on different chromosomes accounts for random assortment of genes.

6.3. How did early geneticists demonstrate that chromosomes carried the genetic material?

a. Chromosomes were found to determine sex type.
b. Gene inheritance correlated with sex chromosomal inheritance for some traits. These traits are called sex linked traits.
c. Abnormal sex linked gene inheritance occurred with abnormal sex chromosome inheritance.

7. SEX CHROMOSOMES

XX, XY - in humans, and flies. Individuals with two X chromosomes are female, males have an X and a Y chromosome. The X and Y chromosomes pair during meiosis and males produce X and Y gametes in equal numbers. Females produce gametes only with X chromosomes. Fertilization gives rise to XX and XY offspring.

ZZ, ZW - In birds the females have heteromorphic sex chromosomes, called Z and W. Males have homomorphic sex chromosomes called ZZ

7.1. Sex linkage.
Disproportionate inheritance of a trait between male and female offspring usually indicates that a trait is sex linked i.e. is a trait controlled by a gene on either the X or Y chromosome. (Most often it

is on the X chromosome).

7.2. Genes on the X chromosome are passed from fathers to daughters and from mothers to either daughters or sons.

Because males have only one X chromosome, there are some striking features of sex linkage in cases of deleterious X linked mutations such as hemophilia. Mothers who are heterozygous carriers of deleterious recessive traits pass it to half her sons, and those sons all manifest the trait. For rare recessive traits, daughters usually don't show the trait but half will be heterozygous carriers of the deleterious.

Notes on gene symbols for recessive and dominant alleles, mutant and wild type vary, by tradition and standardization for different species.

1. A, a In generic discussions when we may not know the system used for a particular species the dominant allele is designated by capital letter and the recessive allele by a lower case letter.

2. Drosophila : w, w^+ B, B^+ $w^{,+}$ B,+

In Drosophila a different system is used: The name of gene is based on the mutant type as compared to the wild type fly. Lowercase letters are used if the mutant allele is recessive; capital letters are used for dominant mutant alleles (w, - white eye, B - black body) a plus superscript indicates the wild type allele. Eg., w and B^+ are recessive alleles, w^+ and B are dominant alleles.

3. cf9, Cf9 . For many organisms, three letter symbols are preferred.

4. *E. coli*, bacteria. In this course we will use superscripts +/– to designate the normal and mutant alleles e.g.: *dcm*$^+$ *or dcm*$^-$. However in the real world, when bacterial geneticists write a strain description they only write the mutations and leave out superscripts. E.g., a strain that was written: *dcm, met,* would be *dcm*$^-$ *met*$^-$ *,* that is mutant or inactive *dcm or met* genes.

Mendel's Second Law- Independent assortment.
Lectures 3

1. The TWO GENE CROSS. Mendel's second, Independent Assortment. The alleles of two genes are partitioned independently and randomly into gametes, and fertilization between two gametes is also random. This means that organisms produce gametes with all possible gene assort independently of the alleles combination of alleles of two genes in equal frequency. In a dihybrid cross (Aa Bb x Aa Bb) each of the 4 possible permutations AB, Ab, aB, ab, each occur ¼ of the time. [This law applies to genes that are on different chromosomes]

1) 9:3:3:1

AaBb x AaBb →	A_B_	9	
	aaB_	3	
	A_bb	3	
	aabb	1	

(the underscore, "_ " indicates that the allele may be either the dominant or recessive allele)

Below are the results of a cross between two plants that differed in the alleles they carried for two genes:

Parents: RRyy x rrYY (Round, green seed x wrinkled , yellow seeded)

F1 → RrYy (Round, Yellow seeded)

F2 →	R_Y_*	9	(Round, Yellow (2 dominant phenotypes)
	R_yy	3	(Round, green (1 dominant phenotype)
	rrY_	3	(wrinkled, Yellow (1 dominant phenotype)
	rryy	1	(wrinkled, green (2 recessive phenotypes)

* R_Y_ - "_" means this allele could be either a dominant or recessive allele "R or "r", Y or y. The same, phenotype would be result in either case.

2. Probability Rules, "and" – multiply; "or" - add

The product rule. The probability of two independent events occurring is the product of the two individual probabilities. For example, in the cross described in (1) above, the probability of having an individual in the F2 generation that is homozygous for rr **and** for yy is ¼ x ¼ = 1/16.

The sum rule. The probability of having of having one or the other of two mutually exclusive events is the sum of their individual probabilities. For example, in the cross described in (1) above, the probability of having picking an individual in the F2 generation that is homozygous for rr **and** for yy or is round and green (R_yy) 1/16 + 3/16 = 1/16 = ¼

Memory aid : “and” – multiply; “or” - add

3 Other two gene crosses,	**AaBb X Aabb →**	3	A_Bb
		1	aaBb
		3	A_bb
		1	aabb

This outcome can be seen using the **Punnett square** shown in the book**.** The **tree diagram** is also useful and versatile. We will take another example of a cross AaBb X Aabb, to see the use of a **tree diagram** to get similar information for a two gene cross. In this example one parent contributes the four possible gametes: AB, Ab, aB, ab; the other parent contributes Ab and ab only.

The results can be seen with a **tree diagram**, results for the first gene and second gene are seen in the first two columns, the combinations are seen in the third column in this example:

gene 1	gene 2		genotypes
1/4 AA	1/2 Bb	→	1/8 AABb
	1/2 bb	→	1/8 AAbb
1/2 Aa	1/2 Bb	→	1/4 AaBb
	1/2 bb	→	1/4 Aabb
1/4 aa	1/2 Bb	→	1/8 aaBb
	1/2 bb	→	1/8 aabb

phenotypes	
3/8	A_Bb
3/8	A_bb
1/8	aaBb
1/8	aabb

3) Tree diagrams are very practical for analyzing crosses that involve more than two genes - the Punnett square is very impractical for such crosses.
Example. Add Cc to cross to the cross shown above:

AaBbCc x AabbCc

Each of the 6 genotypes shown above in example 2, would be in turn subdivided into three genotype for "C": CC, Cc, cc. at the frequency of 1/4, 1/2, 1/4.

Example problem.

a) What is the expected frequency of the aabbcc genotype from the cross listed above?
Answer: 1/4 aa x 1/2 bb x 1/4 cc = 1/32 aabbcc

b) What is the expected frequency of individuals with the dominant phenotypes for genes A and C

and the recessive phenotype for the gene B?
Answer: 3/4 A_ x 3/4 C_ x 1/2 bb = 9/32 A_bb C_

Chi square (χ^2) Test

How do we know if real data set is close enough to an expected ideal ratio such as 1:1:1:1 (or 9:3:3:1 etc. etc.) to be confident that each class can actually be explained by our genetic hypothesis - such as, "Genes A and B follow Mendellian inheritance. They segregate independently; they are not linked on the same chromosome"? There is a statistical test to decide if a real data set matches theoretically expected results. Real data will rarely match a Mendellian ration perfectly, this test helps us decide if deviations from the ideal could be explained by mere chance or if they are likely due to some other cause, such as linkage. The statistical test is the **Chi squared test**. [Chi is written with the Greek letter: $\chi^2$and pronounced "Kaie" or KI with a hard English i sound like the word "eye".]

Consider the data from a test cross CcDd x ccdd

progeny:		
	C D	300
	C d	272
	c D	278
	c d	350
		1200

We can test whether this deviation from a 1:1:1:1 could be explained by variation due to chance, or not.

1. **Make null hypothesis**. Usually : "There is no linkage." With this hypothesis we expect a 1:1:1:1 segregation ratio. The contrary hypothesis, "The genes are linked." is not good because we would not be able to know what segregation values to expect and to calculate expected values.

2. Calculate Chi squared (with numbers not ratios!)
O are observed values, E are expected values.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

In a population of 1200, a 1:1:1:1 ratio would give expected values of 300 for each class.

O	**E**	**(O-E)**		**$(O-E)^2$**		**$\frac{(O-E)^2}{E}$**
300	300	= 0	squared --->	0	divide by E -->	0
272	300	= 28		784	(300)	2.61

278 - 300 = 22	484	1.61
350 - 300 = 50	2500	8.33
		sum 12.51

$\chi^2 = 12.51$

3. Degrees of freedom = (number of classes) -1

df = 4 - 1 = 3

4. Compare values to χ^2 distribution in Table 3-1, page 97 of text book.

With a χ^2 value of 12.10 and 3 degrees of freedom, the P value (or probability) of having such a deviation from expected values by mere chance is between .01 and .005, i.e. between 1% and 0.5%. This means it is very unlikely.

We reject the null hypothesis, which was "the genes are unlinked".
Therefore, the genes are likely to be linked.

The language of hypothesis testing is subtle and precise. If we reject the null hypothesis, we say that it is very unlikely that we would obtain such a data set if the null hypothesis were true. **The cut-off for "unlikely" is 5%**. That is, we would expect such a poor match between expected and observed data less than 5% of the time if the genes were unlinked and the deviation from the ideal was occurring by chance alone. In cases where we accept the null hypothesis, we are only saying that the observed data is consistent with the hypothesis, this is not proof of that the null hypothesis is true. Accepting the null hypothesis does not mean that there are no other possible explanations for the data.

SOME TEXT BOOKS insist that one should use the terminology "fail to reject" the null hypothesis, instead of "accept" the null hypothesis. "Fail to reject" is not universally used, and it is not the terminology used by our text book. "Accept" is acceptable in this course, but you should be aware that there is some controversy over the choice of words and that other terminology maybe used in other courses. However, there is universal agreement over the meaning of these terms in relation to hypothesis testing.

Linkage - Eukaryotic Chromosome Mapping Chapter 4

Mendel's second law, the law of independent assortment, applies to genes which are on different chromosomes. In contrast, alleles of genes that are close together on the same chromosome tend to be inherited together, i.e. they do not assort independently.

The tendency for genes to be inherited together due to their location on the same chromosome is called **linkage.** Alleles of genes that are very close together are said to be closely linked and have a high probability to be inherited together. Those with lower probability to be inherited together are loosely linked. The degree of linkage is proportional to the distance between genes on chromosomes and is measured in units called **map units**. Genes are arranged in a linear order on chromosomes that is virtually identical for every individual within a species. (Exceptions to this are discussed later in the course)

I. Detecting linkage. - two genes, two alleles each.

With two independently assorting genes (i.e. unlinked genes) we expect a 9:3:3:1 segregation from a dihybrid cross, and a 1:1:1:1 segregation from a test cross. If a deviation from these ratios is observed, we suspect that linkage is involved.

Example: In tomato hairlessness on the stem is controlled by a recessive allele of the hairlessness gene. The presence or absence of a joint on the stem of the fruit is controlled by a single gene, jointlessness is controlled by the recessive allele, j.

	P1		P2		F1
The cross:	**Hl Hl J J**	x	**hl hl j j**	----->	**Hl hl J j**

followed by test cross: **Hl hl J j x hl hl j j**

resulted in the following offspring

Hl __, J __	162
Hl __, j j	39
hl hl, J __	42
hl hl, j j	157
	400

1) These results deviate significantly from the 1:1:1:1 ratio expected from a test cross involving independently assorting genes. These genes are linked.

2) The most abundant classes of offspring have the same phenotype as the parents of the cross, but most importantly in the linkage conformation of their alleles is like the parents. The parents had coupling conformation, i.e. both genes had dominant alleles on the same chromosome for the P1 parent and both genes had recessive alleles on the same chromosome for the P2 parent.

3) The two recombinant classes of offspring account for 20% of the offspring, calculated by (39 + 42)/400. This means that the two genes are linked on the same chromosome,

and are 20 **map units** apart, also referred to as 20 **centimorgans.**

Note: Linkage can be determined whether the alleles of the genes are in coupling or in repulsion. Recombination is always determined by comparing offspring to the parental genotype and linkage conformation.

II. Symbols for linkage: If you know that genes are linked it is useful to designate this and to indicate coupling or repulsion conformation. The following are four designation systems for the same thing.

hl j / Hl J	hl j / Hl J	hl j/Hl J	hl j/ + +

II. Three-point Test cross. Mapping three genes relative to each other.

1. Consider three genes in tomato. A single gene in which the recessive allele f causes fasciation controls fasciated fruit, a normal F__ plants has round fruit. Hairlessness found in the hl hl genotype was described above. Green stem is controlled by the recessive allele, a, of a gene that controls anthocyanin synthesis, the seedlings of A__ plants have purple stems.

P1 P2

Consider crosses: (1) **F F A A Hl Hl x f f a a hl hl** ----->

F1: **F f A a Hl hl**

(2) **F f A a Hl hl x f f a a hl hl**

The following offspring were observed (note: since this is a test cross, all individuals have at least one recessive allele for gene. We list the variable, allele in the table; this allele will correspond to the phenotype, since the unlisted allele is recessive). F A Hl = Ff Aa Hl hl

F A Hl	302
f a hl	308
F a hl	101
f A Hl	109
F A hl	82
f a Hl	77
F a Hl	10
f A hl	11
	1000

a) There is not an equal distribution of offspring classes, therefore there is linkage.

b) The most abundant classes are the parental types (like P1 and P2) in gene linkage arrangement.

c) The recombination rate between F and A is measured by counting all individuals in which F and A or f and a are not inherited together and dividing this number by the total. This frequency of recombination is:

(101 + 109 + 10 + 11)/ 1000 = 230/1000 = .23

c2) F and A are 23 map units apart on the same chromosome.

d) The recombination frequency between A and Hl is:
(82 + 77 + 10 +11)/1000 = 180/1000 = .18

d2) A and Hl are 18 map units apart on the same chromosome.

e) The recombination frequency between F and Hl is:
(101 + 109 + 82 + 77 + 2(10 + 11))/1000 = 411/1000 = .41
["10 + 11" is added twice because these two classes, in reality, represent two crossovers between the two genes]

f) If all 8 classes of offspring are present, the most rare class is the double crossover class - if you compare these to the parental class you can determine which gene is in the middle.

Interference
In most cases, the occurrence of one crossover between a chromosome pair has an inhibitory effect on additional crossovers in the adjacent regions. This tendency is called interference. It is observed as a lower frequency of double crossovers than expected when our expected values are based simply on map distance.

Example: In the example given above genes F and A are 23 map units apart, the genes A and Hl are 18 map units apart. The frequency of crossover between each gene pair is .23 and .18 respectively. In the absence of interference you would expect a frequency of double crossover to be 0.23 x 0.18 = .0414. Among 1000 progeny you would expect find 1000 x .0414 = 41 double crossover individuals. In fact 21 double crossovers were observed (10 + 11).

$$\text{Interference} = 1 - \frac{\text{Observed double crossovers}}{\text{Expected double crossovers}}$$

$$1 - \frac{21}{41} = 1 - .51 = .49$$

Note the value: observed/ expected is called the coefficient of coincidence

GENETICS OF BACTERIA AND THEIR VIRUSES. – Chapter 5

Indicators that a physical character is controlled by genes include:

1. Inheritance - trait is passed though successive generations in families.
2. Mendelian segregation ratios - 1:3, 1:1 can indicate that the trait is controlled by one gene.
3. Linkage to other genes. Co-segregation of traits.

In order to observe these things it is advantageous to study organisms that have short generation time and large numbers of descendants. Bacteria are good for criteria 1 and 3, inheritance and linkage. They are not good for criteria 2, Mendelian ratios.

Under optimal conditions, the bacteria, Escherichia coli, divides every 20 minutes. A single cell produces a colony of 10^7 cells, which is visible to the eye, after 16 hours of growth.

A. *Escherichia coli* growth and genetics

a) *E. coli* are grown in liquid media or on agar, a semisolid media.
b) Serial dilution of cultures, plating and counting of colonies are done to determine the number of cells in a culture per a given volume, e.g. per ml.
c) The use of selective media is used to determine the phenotype and genotype of bacterial cells.

Three classes of mutants are commonly characterized in *E coli*.

I. Resistance to antibiotics

Some examples: resistance alleles are designated with a "superscript r", sensitive are designated with a "superscript s".

ampicillin	(amp^r)	kanamycin	(kan^r)
streptomycin	(str^r)	tetracycline	(tet^r)
neomycin	(neo^r)		

II. Requirement for a particular nutrient in order to grow.

Many of these gene mutations are related to a requirement for particular amino acids; eg.

argenine	arg-	cysteine	cys-
glycine	gly-	histidine	his-
leucine	leu-	lysine	lys-
methione	met-	proline	pro-

III. Mutants that cannot grow on a particular compound as a sole carbon source. Many of these gene mutations prevent *E. coli* from growing with a particular sugar as a sole carbon source.

lactose	lac-
galactose	gal-
mannose	man-

B. Conjugation in *E. coli* - i.e. mating.

1) The F factor, for fertility. F+ strains can donate genes and F- can accept genes. F+ factor is an episome, a small circular chromosome.

2) Hfr strains, high frequency recombination strains, have the F factor integrated into their large principal chromosome.

C. Methods of gene mapping in bacteria.

1) Gene mapping by timed matings. - ie interrupted matings.
- note the genome of *E. coli* is circular.

2) Mapping by recombination frequency (for higher resolution).

3) Mapping by transformation. Transformation is the introduction of chromosomic material (DNA) by physical means, by an experimenter, rather than by mating.

4) Mapping by phage transduction, (see below) .

Example:

1. Mapping in *E. coli* by interrupted matings; this allows partial transfer of the bacterial genome (unlike higher eukaryotes). The chromosome passes into the recipient cell in a linear order and the timing of passage becomes a measure of distance between genes on the chromosome. The HFR factor is the last gene to transfer.

Genes: Streptomycin sensitive, argenine+, leucine+, histidine+

The matting: Str^s, arg^+,leu^+,his^+,lac^+ (Hfr) X Str^r,arg^-,leu^-,his-,lac- (recipient)

The cells are plated (grown) on media with streptomycin (The two strains cannot be physically separated after they are mixed, but after the mating is allowed to proceed for a controlled set of times, the streptomycin antibiotic is used to kills or inhibit the growth of the Hfr, donor strain, since it is Str^s, but allows the recipient to grow.

Selective media to determine the genotype of the cells after mating, minimal media plus various supplements:

Media	Strains that grow
Minima media (MM)	his^+, leu^+, arg^+
MM plus arg and leu	his^+
MM plus leu and his	arg^+
MM plus his and arg	leu^+
MM plus glucose	glu^+
MM plus lactose	lac^+

phenotype	10 min.	20 min.	30 min.	40 min.
arg^+	0	0	3	20
leu^+	6	30	60	80
his^+	0	8	41	56
lac^+	0	0	0	10

The leu gene is the first to be passed, at 10 min, followed by his, arg and lac. Therefore the gene order is: leu, his, arg, lac

---

Bacterial gene maps are circular!!!! The bacterial chromosome is circular.

D. Bacteriophages (synonym: phages, bacterial viruses)

1) infection
2) lysis
3) phages have genes too!, which can be mapped
4) lysogenic phages
5) transducing phages (P1) - generalized transduction. The virus as a transporter of E. coli genes.

*6) mapping by co-transduction of genes.

Genes in *E. coli* can be mapped by measuring the rate of co-transduction via the P1 phage form a donor bacterial strain to a recipient strain. The more closely linked two genes are, the higher the rate of cotransduction. The maximum distance that can be cotransduced by a p1 phage is about 2 minutes, about 2% of the *E.coli* genome. It is most useful for mapping relatively short distances.

For interest only: The *E. coli* genome is 4,000,000 base pairs (4 Mega bases or 4 Mb) of DNA long. The genetic map of E. coli is 90 minutes long. P1 phage can transduce about 2 minutes of DNA, or about 80,000 (80 kb) of DNA.
The sequencing of the complete genome of *E. coli* was completed in 1997, the relative position of every gene is now known. The actual physical map accurately corresponds to the genetic map that was developed by the mapping methods described above.

GENETICS OF DNA FUNCTION Chapter 6

A. GENES ENCODE PROTEINS

Proteins are the molecular machines of organisms and are central to the functioning of all organisms. Enzymes of metabolism are proteins. The transporters that carry substances across membranes into and out of the cell are proteins. The receptors of signals and transmitters of signals are proteins. Antibodies of the immune system are proteins. They play many, many roles.

The one-gene-one-enzyme hypothesis proposed that genes encode proteins – genes are the templates from which proteins are produced. This hypothesis was first proposed by George Beadle and Edward Tatum in the 1940's – even before the structure of DNA was known.

1. The work of Beadle and Tatum:
The basis of the hypothesis came from the analysis of mutants of *Neurospora* which required arginine to grow.

a) The biosynthetic pathway for arginine was proposed (and subsequently demonstrated) to be:

precursor → ornithine → citrulline → arginine

Three different enzymes catalyze these three steps in synthesis.

b) Three different mutations were shown to be mutations in different genes. They **mapped** to three different chromosomes, therefore they must be different genes.

c) All three mutants could grow if they were given arginine, however two of the mutants could also grow in they were given the intermediate compounds from the biosynthetic pathway. See the table below for details.

Table 12-2 in the text:
Growth of arg mutants in response to supplements. "+" is for growth; "-", no growth

	Supplement		
Mutant	Ornithine	Citrulline	Arginine
agr-1	+	+	+
arg-2	-	+	+
arg-3	-	-	+

d) The hypothesis "one-gene-one-enzyme" proposed that each of the different mutations was responsible for inactivating different enzymes in the biosynthetic pathway for arginine. These

could be recognized because intermediates "downstream" of the inactivated enzyme could satisfy the nutritional requirements of the mutant.

2. **The structure and function of proteins.** Proteins are long polypeptide chains composed amino acids. There are 20 common amino acids found in all organisms. Proteins function as enzymes and play central roles in the functioning of organisms. Each protein class has a unique amino acid sequence.

3. Further demonstration that genes encode proteins came from the analysis of mutations of causing sickle cell anemia. By studying people that had the genetic disease it was found that they had changes in some of the amino acids in their globin proteins.

4. There was **co-linearity** of mutations in the gene for tryptophan synthase and the resulting changes in amino acid changes in tryptophan synthase enzyme was shown by Charles Yanofsky.

B. The molecular basis of dominance and recessiveness.

Mutations will often cause gene product, e.g. an enzyme, to be dysfunctional. However if the other gene copy on the homologous chromosome is not mutated, and if the non-mutant allele produces enough normal, functional enzyme for the cell to have normal metabolism, then the normal allele will be dominant and the mutant will be recessive. The heterozygote will have the same phenotype as the homozygous dominant genotype.

C. COMPLEMENTATION tests are done to determine whether two mutations are mutations in the same gene or in two different genes. Wild type or normal phenotype will be restored in a cross between two mutants with mutations that effect the same trait if the mutations are in different genes. An example would be mutations in two different enzymes in the same biosynthetic pathway.

Mutations in different genes can cause the same phenotype (eg arg-). We can make a diploid (or merodiploid) with the two mutant genes in question by crossing the two mutants. If the mutations are in different genes, wild type phenotype will result in the presence of two mutations. An example would be mutations in two different enzymes in the same biosynthetic pathway. "Cistron" = gene.

Experimental methods of complementation testing vary from organism to organism - depending on their mating system.

a. Neurospora – fusion of fungal hyphae for the production of heterokaryon.

b. E. coli - merodiploid ("partial diploid"), is produced by introduction of episome or plasmid carrying the gene to be tested.

c. Phage - co-infection with two strains.

d. Higher eukaryotes, plants and animals – complementation is tested by mating, and more recently by introducing genes by recombinant DNA methods.

Recombination can also restore wild type phenotype when two different mutations are in the same gene but at different locations within the gene, but this occurs at a very low frequency because it depends on very rare recombination between two points within a gene that are very close together and can readily be distinguished from COMPLEMENTATION, on the basis of the frequency of its occurrence.

Gene Interaction and Variation of Mendelian Inheritance ratios Chapter 6 continued

The variations of inheritance patterns discussed here are involved with a fascinating and often beautiful array of phenotypes, which often appear very complex. Usually the inheritance can be understood them in terms of simple Mendelian ratios.

I. Variations of single gene traits.

1. Incomplete dominance (and no dominance).

With incomplete dominance the heterozygote, Aa, has an intermediate phenotype between the two homozygotes. For example: Genotypes AA, Aa, aa are respectively: red, pink, white. The phenotypic ratio in Aa x Aa cross is 1:2:1.

2. Codominance

Both alleles are fully expressed in the heterozygote.
For example: in the ABO blood groups. Heterozygotes with the A and B alleles have both the A and B blood antigens and are blood type AB.

3. Multiple alleles of a single gene.

A single gene can have more than two alleles, within a population, though any one diploid individual can carry only two of these alleles. For example: in the ABO blood groups. This trait is controlled by a single gene. There exist A, B and O alleles. A and B are controlled by codominant alleles I^A, I^B, respectively and the third allele is "i" which is a recessive allele that accounts for the O blood type. Individuals can have the following genotypes and phenotypes.

Genotype	Blood type
ii	O
I^AI^A,I^Ai,	A
I^BI^B, I^Bi	B
I^AI^B	AB

Though a single gene may control many phenotypes among different individuals, a <u>single gene trait</u> can be recognized by the typical single gene segregation rations including 3:1, 1:2:1, 1:1 in the progeny from individual crosses.

---

4. Lethal alleles. Recessive lethal alleles cause the death of homozygous recessive individuals (sometimes *in utero*). This causes a 3:1 ratio to become a 2:1 among surviving offspring. The allele has two effects in the example of the tailless Manx cat: the M^L allele is dominant for taillessness but is recessive for lethality (death).

II. Multiple genes affecting the same trait.

1. Two genes that affect the same trait give rise to segregation ratios based on the 9:3:3:1 ratio.

For example, there are two genes which affect coat color in mice. This example involves two genes, two alleles each, each with one dominant allele and one recessive allele.

BB and Bb	- black
bb	- brown
AA, Aa	- agouti (yellow band near the tip of the hair)
aa	- solid color.

AaBb x AaBb

9	A_B_	agouti
3	A_bb	cinnamon
3	aaB_	black
1	aabb	brown

Note: The examples presented above and below are primarily shown as results of a dihybrid cross like AaBb x AaBb, to facilitate comparison between different types gene interactions. However in each case, crosses can be made with other genotypes for example AaBb x aaBB would give other interesting segregation ratios.

2. Epistasis and epistatic genes. In epistasis the expression of one gene affects the expression of another gene - as if overriding the effects of the other gene.

Example. C is a color gene, cc homozygous recessive is albino.
Albino individuals are without pigment even though other genes for pigment are present in the individual. The albino genotypes overrides (i.e. is epistatic to) the effects of other genes.

BBCC -	are black
bbCC -	are brown
BBcc -	are albino
bbcc -	are albino

In a dihybrid cross as listed for 1. (above) the genotypic segregation ratios would lead to a phenotypic segregation ratio of 9:3:4. Because two classes of the 9:3:3:1 ratio have the same phenotype.

3. Duplicate genes.

Genes can be duplicated in some organisms, so instead of the normal ratio of 3:1 for a single gene one can observe the segregation from a cross like AaAa x AaAa:

A_A_	9		
A_aa	3	= 15	This leads to a 15:1 ratio.
aaA_	3		
aaaa	1	--- 1	

THE STRUCTURE OF DNA – Chapter 7

How do we know what we know about DNA?

A. Landmark experiments in the understanding DNA as genetic material:

1. Frederick Griffith, 1928. Transformation: - A substance, "the transforming principle", from *Streptococcus pneumoniae* could change the genotype of a non-virulent, rough coated strain, causing it to a virulent and smooth coated.

2. Oswald Avery, CM MacLeod and M McCarty 1944, identified the "transforming principle" as DNA – deoxyribonucleic acid.

Initially this was met with skepticism because DNA was considered to be a simple molecule, a polymer made of four basic constituents.

3. Alfred Hershey and Martha Chase in 1952, confirmed DNA as the genetic material with experiments with T2 phage. Components of DNA were known: deoxyribose nucleotides giving the name: deoxyribonucleic acid.

4. Chargaff's rules (Erwin Chargaff) described the relative frequencies of the four specific nucleotides in any sample of DNA. The amount or Thymadine in a DNA sample was equal to the amount of adenine, and the amount of guanine was the same as the amount of cytosine.

T + C = A + G
A=T C=G, A+T does not necessarily = G+C
G+C = 1 - A+T and is different for different species.

5. James Watson and Francis Crick first proposed the structure of DNA as an antiparallel double helix, in 1953. This was based on the X-Ray diffraction data of Rosiland Franklin and Maurice Wilkins.

B. Implications of the structure. Coding potential and replication of information!

C. Replication

Mathew Meleson and Franklin Stahl (1958) showed that DNA replicated in a semi-conservative fashion. This was demonstrated by labeling *E. coli* DNA with ^{15}N and characterizing it by density gradient centrifugation, after one and two cycles of replication.

A chromosome contains a single double strand molecule of DNA.
In replicated chromosomes, before cell division, each chromatid is a single double stranded molecule of DNA.

D. The enzymes of DNA replication.

Arthur Kornberg isolated the first DNA polymerase,

DNA polymerase I, called Pol I.

Another polymerase, Pol III, was latter discovered and shown to be the enzyme that carries out most of the synthesis of DNA. It synthesizes DNA on both leading and lagging strand and pol I fills in the gaps in the lagging strand.

DNA REPLICATION is done by a number of enzymes (which are proteins):

Primase or primasome complex includes dnaB, dnaT, priA, priB priC
Pol III does most of the synthesis, pol I fills in the gaps
DNA ligase joins adjacent nucleotides by a covalent bond.
Helicases disrupt the H bonds
Topoisomerases produce supercoiling or relax supercoiling of the DNA
Pol I and III have exonuclease activity and can proof read or edit, in regions where base matching has errors, the enzymes go in reverse and remove the errors and then go forward with the correct replication.

DNA polymerases move 5' to 3'
DNA polymerases require a primer, i.e. a pre-existing strand of DNA or RNA to which they add additional nucleotides.

Lagging strand of DNA replication is primed by Okazaki fragments, which are short RNA fragments

E. Origins of replication

The *E. coli* chromosome has a single origin of replication. (Note, the F factor and other episomes also have their own origins of replication, this makes them autonomously replicating.)

Eukaryotic cells have many origins of replication along the chromosomes. For example, in yeast, *Saccharomyces cervisiae*, which has 17 chromosomes, there are 400 origins of replication. In eukaryotic cells, replication occurs at S phase of the cell cycle. Cell cycle: G1, S, G2, M.

Pulse-chase labeling can be used to see multiple origins of replication. This involves a pulse feeding of radioactive nucleotides followed by feeding with non-radioactive nucleotides. The label is incorporated into the DNA causes sections of a chromosome to be labeled - this can give images on film sensitive to radioactivity, and show the pattern of replication.

F. Teleomeres

Teleomeres are the ends of linear chromosomes in Eukaryotes. Telomerase is the enzyme that replicates the DNA at the end of the chromosome; it contains RNA that serves as a template for

DNA synthesis

RNA TRANSCRIPTION AND PROCESSING Chapter 8

DNA ⟶ **RNA** ⟶ **proteins** ⟶ metabolism ⟶ phenotypes

In eukaryotes DNA is in the nucleus, and protein synthesis takes place in the cytoplasm. The link (messenger) between the genes and protein synthesis is mRNA. RNA was shown by "pulse-chase" experiments to be synthesized in the nucleus and to move to the cytoplasm. Pulse-chase also showed that RNA has a short half-life.

1. RNA structure: RNA is similar to DNA except:

a) RNA is single stranded,
b) RNA has ribose sugar instead of a deoxyribose,
c) RNA has uracil instead of thymine

RNA classes:

ribosomal RNA (rRNA) - part of the translational complex.
message RNA (mRNA) - actual encoded message for protein synthesis.
transfer RNA (tRNA) - part of the translational complex. It is the link between the message and the amino acids.

2. TRANSCRIPTION - RNA synthesis. RNA polymerase, a DNA dependent enzyme, synthesizes RNA from the DNA template. Correct transcription of the genetic material is dependent on base complementarity of RNA to DNA (like the complementarity between the two DNA strands).

For any gene, only one DNA strand is the template for RNA synthesis. The DNA strand that is the template for RNA synthesis is the "sense strand", the complementary strand is the "antisense strand."

In *E coli*, the RNA polymerase has four different protein subunits
ß, ß', a, a, s (beta, beta', alpha, alpha, sigma).

BEGINNING AND END of the transcription of a gene.
Initiation: binding to promoter region -35 and -10 to a "consensus sequence" i.e. a similar, though not identical sequence found in many genes.
Elongation: synthesis proceeds in a 5' to 3' direction .
Termination: a) GC rich sequence, self complementarity, forming a hairpin loop, followed by A's

b) Rho factor interacting with Rut site on RNA

3. TRANSCRIPTION IN EUKARYOTS

Eukaryots are organisms whose cells have nuclei, including yeast, fungi, animals and plants. The mRNA in eukaryots is more processed. A eukaryotic gene usually has non-coding regions called **introns** interspersed between coding regions called **exons**. Initially the RNA from a gene

is transcribed with both introns and exons. Eukaryotic mRNAs are have 3 processing steps: (1) the 5' end is capped with a 7-methylguanasine (2) a poly AAAA string of about 300 A's are added to the 3' end (3) the introns are spliced out leaving only the exons.

TRANSLATION - PROTEIN SYNTHESIS CHAPTER 9

mRNA, encodes the message from the gene to make a protein.
rRNA, tRNA essential parts of the translation machinery, act in concert with the proteins of the translation complex for the "de novo" synthesis of proteins.

4. DNA and THE CODE

THE CODE see Figure 9-8 , pg 282 in the text book.
Three letter non overlapping code, called a CODON. One codon codes for one amino acid. The mRNA is read by consecutive codons of three nucleotides each with no gaps. The possible number of triplets that can be composed from the four bases (A,U,C,G in RNA, or A,T,C,G in DNA) is 4x4x4=64, The code must encode 20 different amino acids, most amino acid have more than one codon (the number of codons for a single amino acid ranges from 1 to 6 codons, depending on the amino acid).

Protein synthesis proceeds in a stepwise fashion, by a translation complex composed of the large and small ribosomal subunits with protein factors. Translation proceeds in a 5' to 3' direction on the mRNA, the protein is synthesized from the "amino" terminus to the "carboxyl" terminus.
Translation starts at an AUG codon (it codes for methionine), STOP – is signaled by UAG, UAA, or UGA. In DNA these are ATG, for start, TAG, TAA, TGA for stops.)

5. MUTATIONS

- point mutations - single base pair (bp) substitutions, may change the amino acid.
- bp insertions and deletions - frame shift, drastically disrupts the coding sequence.
-3 bp deletions or insertions, cause one amino acid to be deleted or inserted respectively
-large deletions, cause parts of the coding sequence to be lost

6. Eukaryotes vs. prokaryotes (bacteria): transcription and translation.

a) the genetic code is universal - the same code is used by *E. coli* and humans.
b) monocistronic (eukaryotes), polycistronic (prokaryotes)
c) eukaryotic genes have introns which are spliced out of the mRNA
d) other processing in eukaryotes include: 5' end cap of mRNA with 7-methylguanosine, poly (A) tails at the 3' end.

Structure of a Eukaryotic mRNA

Sequence in
DNA: **ATG** **TAG, TAA or TGA**

mRNA: Start Stop
AUG UAG or UAA or UGA
_____|_____|_____AAAAAAAAAAA

5' UTR Coding region 3' UTR T tail

UTR = Untranslated region

Control of gene expression. Chapter 11, pg 397-413

Gene expression: the transcription of a mRNA and its translation to produce a protein.
A gene is being expressed, or is "on", when it is producing its mRNA, it is repressed, or "off", when it is not.

It is estimated that there are between 25,000 and 35,000 genes in multicellular organisms such as humans, flies, and plants. The genes are sometimes "on", sometimes "off", and the regulation of their expression is precisely controlled. Different genes are expressed in different tissues, at different times of development and in response to different stimuli, such as nutrients, light, stress, infection.

How are genes turned off or on??

1. THE LAC OPERON is one of the first models of gene regulation.
The regulation of this gene was studied by Francois Jacob and Jaques Monod.

The lactose operon is a cluster of three genes (**lac Z, Y, A**) a) lac Z: encodes Beta-galactosidase, an enzyme which breaks the ß linkage in lactose sugar, a disaccharide to produce two monosaccharide sugars galactose and glucose.
b) lac Y, a permease - which facilitates the transport of lactose across the plasma membrane into the cell. c) lac A, a transacetylase (not important for lactose metabolism). A single mRNA is expressed which encodes all three genes. This is called a **polycystronic message**, because it encodes multiple proteins

1) The operon is not expressed in the absence of lactose.
2) The genes are expressed when lactose is added to the media.

This regulation has three elements:

1) The lac repressor - a protein which physically binds to the DNA upstream of the lac and inhibits its expression. The lac repressor is encoded by another gene, the **lac I** gene.
2) The **O, operator site** is the site in the DNA where the lac I repressor binds.
3) The **P, promoter** region of the DNA is upstream of the O site, is the binding site of the RNA polymerase protein - it is necessary for transcription.

The lac repressor - has two binding abilities (and two binding sites) one for DNA the other for lactose. When lactose is present, it binds to the repressor protein rendering it unable to bind the DNA, thus the gene becomes derepressed and expressed, RNA begins to be synthesized.

Operon: I PO ZYA

The function of the lac operon was analyzed by the use of mutations in the various elements of the operon. Mapping was used to determine the linear order of the mutations on the chromosome. Merodiploids were used to determine dominance and recessiveness between the alleles, and epistatic relations between the elements.

Lac Operon - mutations

Z ß-galactosidase , cleaves lactose into galactose and glucose
Z^- mutation leads to an inactive enzyme, recessive to Z+

Y permease, transports lactose into the cell
Y^- inactive enzyme, recessive to Y+

A transacetylase (not very relevant to the model)

I Repressor. It represses the expression of the lac operon by binding to the operator region. In the presence of lactose, it binds the lactose and does not bind the operator and the lac operon becomes de-repressed or actively transcribed.

I^{--} inactive repressor. It cannot bind the operon, cannot repress. It is recessive to I+, causes **constitutive** expression of the lac operon. Constitutive means constantly "on" with no regulation.

I^s super repressor mutant form of the repressor. It binds the operator site and represses the expression of the operon. It cannot bind lactose and therefore the operon cannot be derepressed. Gene is OFF. I^s is dominant to I^+ and I^-, its repressive capacity can override the effect of any normal repressor product produced by a I^+ copy of the gene.

O Operator. The binding site of the repressor. It is a DNA sequence, upstream of the lac Z gene.

O^c Mutations in the Operator region can inhibit binding of the repressor. This leaves the gene ON **constitutively**. O^c mutations also do not bind the I^s type repressors. I^s therefore cannot repress a O^c type mutant (O^c mutants are **epistatic** to I^s mutants).

P Promoter. The region of DNA where the RNA polymerase binds to the lac operon. It is upstream of the O site

P^- Promoter mutations inhibit the binding of RNA polymerase and inhibit transcription, and repress the operon (such mutations are **epistatic** to O^c mutations)

Note: Dominance and recessiveness describe the relationship between alleles of the same gene. The relationship between alleles of different genes is described by epistasis.

CIS elements and TRANSACTING FACTORS

O and P are CIS elements – they are physically adjacent to the gene they are controlling.
The Repressor, I, is a transacting factor. It acts "in trans", " on a gene "at a distance".

The action of the lac repressor is an example of negative regulation.

2. Positive control - another level of control of the lac Operon is **catabolite repression**. **Catabolite Activator Protein** (CAP) and the lac operon.

E. coli will metabolize lactose if it doesn't have glucose available in the media. The switch between glucose and lactose based energy metabolism is regulated by the interaction of the positive regulator, CAP, and the lac operon.

With high glucose in media: no ß galactosidase produced.

high glucose → low cAMP (cyclic AMP)
low glucose → high cAMP

At high cAMP concentrations, cAMP binds to CAP (Catabolite Activator Protein). The complex binds to the promoter and activates lac operon. The binding occurs at the CAP binding site in the DNA. This is an example of positive regulatory element.

---

3. **Attenuation** is another mechanism of gene regulation that has been described in *E coli* on the tryptophan operon. When tryptophan levels are high, trp-tRNA levels are also high, translation proceeds quickly but secondary structures formed in the trp operon mRNAs and this inhibits further **transcription** of the message**.** When tryptophan levels are low, trp-tRNAs levels are low, translation stalls at the trp codons and no inhibitory secondary structures are formed in the mRNA. **Transcription** is not inhibited.

This interaction between translation and transcription is possible in bacteria because a mRNA can be simultaneously transcribed and translated. This is not the case in eukaryotes. However gene expression in eukaryotes can be regulated at the level of translation by protein-mRNA interactions.

4. **Gene regulation in Eukaryotes.**

Eukaryotic gene promoters usually have many *cis* acting elements, both positive and negative elements. They contribute to the complex way in which genes are expressed. The same gene may be expressed at very different levels in different tissues, at different times during development, or cell cycle, under different environmental conditions, etc.

Regulatory elements –RE ,

e.g. DRE is a "drought response element" found in certain plant genes

Upstream activating sequences –UAS
Enhancers - increase gene expression and act additively to other tissue specific regulation
Silencers – act to lower or stop gene expression

Recombinant DNA Techniques - Chapter 10

The remarkable advances in our understanding of the molecular basis of genetics in the last three decades have increased the understanding of genes, gene regulation and genome of many organisms. Individual genes can be isolated and added to, or inactivated in the genomes of plants and animals and micro-organisms. Recombinant DNA methods are used for diagnostics, medical treatment, and forensics.

Recombinant DNA technology can be used to cut DNA precisely into discrete pieces and reconnect or ligate them to other DNA. DNA is cloned into vectors that allow its replication in cells. DNA can be sequenced; the coding portion of a gene can be recombined with the promoters of other genes. These methods are constantly changing and advancing. In our class we will learn about the foundation methodologies of recombinant DNA.

1. Restriction endonucleases. Restriction enzymes are proteins, isolated from different species of bacteria, which can cut double stranded DNA specifically. The enzyme recognizes a specific sequence in the DNA and cuts it in a precise manner.

The recognition sequence for restriction enzymes, restriction site, is specific for each restriction enzyme. Restriction sites are normally 4,5,6,or 8 base pair (bp) long, the most commonly used enzymes recognize sequences that are 6 bp long. Restriction sites are usually palindromes, they read the same if you rotate the site 180 degrees.

EcoRI 5' GAATTC
CTTAAG 5'

Many enzymes cut in a staggered pattern. Each strand of the double stranded DNA is cut off-centered, this generates short single stranded overhanging ends referred to as "sticky ends" because they can rehybridize to each other by hydrogen bonding.

5'AAATACCTTTGAATTCTCAATA → AAATACCTTTGAATT CTCAATA
TTTATGGAAACTTAAGAGTTAT TTTATGGAAAC TTAAGAGTTAT

The sticky ends can be rejoined by covalent bonds rejoining deoxyribose-phosphate backbone of the DNA by another enzyme called **ligase.**

The expected frequency of any given restriction site in the DNA is proportional to the size of the sequence of the recognition site. $(1/4)^n$ where n is the number of bp in the site.

eg.	HaeIII	GGCC	4 bp : 1/4 x 1/4 x 1/4 x 1/4 = 1/254
	TfiI	GAATC	5 bp : 1/1024
	BamHI	GGATCC	6 bp : 1/4096
	NotI	GCGGCCGC	8 bp : 1/66,000

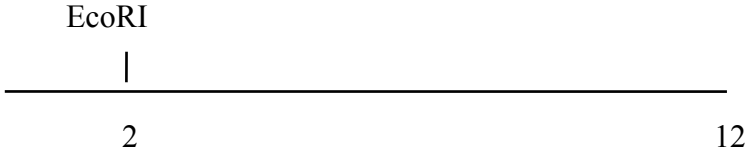
i.e. the site recognized by BamHI occurs about once in 4000 bases in a piece of DNA, so DNA digested by BamHI will produce fragments with the average size of about 4000 bp long.

2. **Separation of DNA by size.** The size of fragments of DNA that result from digestion by restriction enzymes can be determined by **gel electrophoresis**. In electrophoresis the DNA fragments are moved through a gel slab by an electric current and are separated by size. The speed of migration of a fragment is correlated with the size of the fragment. The smaller fragments of DNA migrate through the gel faster.

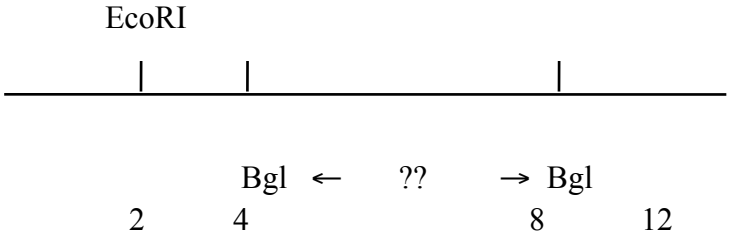
The DNA can be visualized with by staining with ethidium bromide. The bands of DNA appear orange when viewed with ultra violet light. Bands can also be seen by autoradiography if the DNA is labeled with radioactive nucleotides.

The location of restriction sites in a piece of DNA can be determined by knowing the size of the starting piece of DNA and determining the size of the fragments that result from the digestion.

If a 12,000 base pair (usually called **12 Kb**) long piece of DNA is digested with EcoRI restriction enzyme and the resulting fragments are 2 and 10 kb we would know that a EcoRI site is 2 kb from the end:

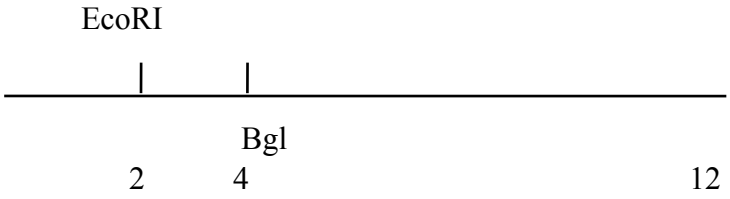


cut the same 12 kb piece with BglI, this results in 4 and 8 kb fragments. Bgl is 4Kb from the end, but it would be unclear which end - relative to the EcoRI site.



A digestion with both enzymes would resolve this question.

If digestion with both enzymes resulted in fragments of 2 and 8 kb, this would indicate that the Bgl site was near the EcoRI site, as in the map below.



(Note that the digestion would result in two fragments of 2 kb. These would be distinct, having different DNA sequence, but would be indistinguishable in an electrophoresis gel since they are the same size. The gel would have two recognizable bands one 8 kb and one 2 kb.)

3. Cloning vectors. Plasmids.

Plasmids are small circular autonomously replicating circles of DNA. About 4000 bp, they can replicate in *E. coli.* Plasmids can be cut with restriction enzymes, and a piece of DNA can be ligated into them and the circle is reformed. Plasmids will replicate in *E coli* thus forming clones with many copies.

Most plasmids have:

Origin of replication for replication in *E coli*.
Gene for antibiotic resistance - a selectable marker gene.
Multiple cloning site (MCS) with many different restriction sites - also called polylinker.
b- galactosidase gene - to detect the presence of the inserted piece of DNA.
lac promoter - to express any foreign gene you insert into the plasmid next to the promoter.

4. Southern blot - electrophoresis of DNA, followed by transfer to a membrane. The DNA will be transferred in the same pattern that it had in the gel. The DNA can be denatured to single stranded forms and fixed on the membrane.

Hybridization: A "probe" can be prepared from an isolated gene which has been cloned into a plasmid. The fragment of DNA encoding the gene is removed from the plasmid by restriction digestion; the probe is made by duplicating the DNA in vitro with radioactive nucleotides. The probe can be denatured by brief boiling, to produce single stranded DNA. The probe can then be hybridized to the DNA on the membrane of a Southern blot. The probe forms double stranded DNA with DNA on the membrane which has complimentary sequence - thus "finding' the gene which matches the probe.

Southern blots have many uses. It is the basis of the DNA fingerprinting methods used to identify individuals in criminal investigations, for verifying parentage, identifying carriers of disease related alleles, etc.

5. DNA sequencing.

DNA to be sequenced is often cloned into a plasmid vector. It is purified melted to single a stranded state and a new second strand is synthesized from a fixed position on one of the original template strands. The fixed position is determined by a primer, short synthetic piece of DNA, an oligonucleotide usually about 20 nucleotides (nt) long, that anneals to a fixed starting point. The DNA is synthesized with DNA polymerase and the 4 normal deoxyribonucloetides and 4 dideoxyribonucleotides that are linked to colored dyes, with each dideoxy nucleotide, ddA ddC ddG ddT, having a different color. When the dideoxy nucleotides are incorporated into the newly synthesized DNA they stop the DNA chain elongation. Thus a strand is extended until a dideoxy nucleotide is incorporated which blocks the end of the strand and attaches a colored dye to the strand indicating the last base added to the strand. The mixture of normal and dideoxynucleotides is such that a population of fragments of DNA ranging from the size of the primer (20 nt) plus 1 nt up to several hundred nt long, each ending with a colored dye which indicates the last base added. Every possible length is represented many times over , ie 21, 22, 23, 24 …… 300, 301, 302, etc. …. nucleotides long. These fragments are separated by size by

electrophoresis and read in the order of their size as they leave the gel matrix. Thus if the fragment 21 nt. long ends in an A (green), 22 in G (black), 23 in G, 24 in T (red), 25 in C (blue) … the DNA sequenced would begin AGGTC, etc. Sequences of over 600 nt. can be achieved from a single reaction sample and electrophoretic separation. Overlapping sequences from many such reactions can be combined to determine the sequence of thousands or millions of nt. long.

POPULATION GENETICS - Chapter 18

GENE POOL = the sum of ALLELES for all genes IN A POPULATION
The term can encompass a species or refer to populations or sub populations.

In the early part of the course we looked at segregation from the point of view of controlled crosses and observed Mendelian ratios (3:1, 9:3:3:1, etc.) for genotypes and phenotypes among offspring from these crosses.

In natural populations the frequency of occurrence of specific alleles varies from vary rare to very frequent. In our study of population genetics we will study the relationship between allelic frequency and genotype frequency, as well as the main forces that effect allelic and genotype frequencies: mating scheme (random mating vs. inbreeding), selection, mutation, migration, and population size.

1. Allele frequency - at a given locus (gene). We can calculate the frequency of alleles if we know the genotypes of the individuals in a population. p is the frequency of allele A, q is the frequency of a.

p= fAA +1/2 fAa q= 1/2 fAa + faa

where fAA, fAa, faa are the frequencies of the three genotypes in the population.

2. The Hardy-Weinberg Equilibrium

If p and q are known, the frequencies of the three genotypes in a population can be predicted. The conditions which are assumed for Hardy-Weinberg equilibrium are: a large population, random mating, and insignificant effects of migration, mutation, selection.

If p and q are the frequencies of the alleles "A" and "a", respectively, the frequencies of the genotypes: AA, Aa and aa will be:

p^2 2pq q^2

Demonstration of this is shown on the back of this page.

2b. This same equilibrium applies to more than one allele at a locus. If p,q,r are the frequencies of alleles a1, a2, a3 then the frequencies in a population in H-W equilibrium of the genotypes a1a1, a1a2, a2a2, a3a3, a1a3, a2a3 would be:

p^2 2pq q^2 r^2 2pr 2qr

HARDY-WEINBERG EQUILIBRIUM

The following is a numerical example that shows that random mating will lead to Hardy Weinberg equilibrium.

We start with an arbitrary distribution of genotypes, not necessarily in any equilibrium. You can try this with other initial numbers.

Arbitrary choice:	We can calculate allele frequency:
.2 AA	
.2 Aa	fA = .2 + (1/2).2 = .3 = p
.6 aa	fa = .6 + (1/2).2 = .7 = q

Assume random mating and determine the frequency of mating between all genotypes.

	.2 AA	.2 Aa	.6 aa
.2 AA	.04	.04	.12
.2 Aa	.04	.04	.12
.6 aa	.12	.12	.36

For example: a mating between AA and AA will occur 4% of the time in this population.

These nine matings are six different classes (AA x Aa is the same as Aa x AA etc.) These are listed below with their frequency and the frequency of offspring genotypes from each mating. Note these are from Mendelian ratios 1:1 or 1:2:1.

MATINGS:

	AA x AA	AA x Aa	AA x aa	Aa x Aa	Aa x aa	aa x aa
	.04	.04	.12	.04	.12	.36
		.04	.12		.12	
Total	.04	.08	.24	.04	.24	.36

OFFSPRING:

	AA .04	AA .04	Aa .24	AA .01	Aa .12	aa .36
		Aa .04		Aa .02	aa .12	
				aa .01		

Total

AA	= .04 + .04 + .01	= .09	= p^2
Aa	= .04 + .24 + .02 + .12	= .42	= 2pq
aa	=.01 + .12 + .36	= .49	= q^2

Note: From the arbitrary choice of population composition, we started with allelic frequency of p

= .3, q = .7, from the result of random mating the frequency of genotypes AA, Aa, and aa are p^2, 2pq and q^2.

2c. The Hardy Weinberg equilibrium can be used in the reverse, namely to calculate the frequencies of alleles in a population when only the frequency of certain genotypes are known.

a. Example: Cystic Fibrosis occurs at the frequency of 1/2500 births among Caucasians of northern European descent. Individuals are homozygous for the disease allele; therefore they represent "aa" genotype and q^2.

Since q^2 is 1/2500,
q is the square root of 1/2500, which is 1/50.
2pq = 2 x 1/50 x 49/50 = 1/25. This the frequency of
heterozygote carriers.
The probability of any marriage within this population group being between two heterozygote carriers is 1/25 x 1/25 = 1/625. In such a marriage, the risk is 1/4 for each child having cystic fibrosis. (Note: 1/625 x 1/4 = 1/2500)

b. Example. Sickle cell anemia affects 1/400 blacks in USA.
Since q^2 is 1/400 q=1/20,
2pq =.095 =.1 is the frequency of heterozygote carriers.

In some West African populations heterozygote carriers of sickle cell anemia (*sca*) comprise 40% of the population. (This is because heterozygotes are more resistant to malaria than homozygotes with normal alleles and the *sca* allele has been selected for.)

Inbreeding

Inbreeding and assortative mating (mating between like partners) increases the frequency of homozygotes and decreases the frequency of heterozygotes relative to Hardy-Weinberg equilibrium numbers.

Humans are outcrossers. Plants include both self pollinating species, obligate outcrossers, and partially inbreeding species (due to pollination to nearby plants which are often sibs).

Self-pollination in plants through several generations of selfing: Half heterozygosity is lost in each generation. (1/2)n

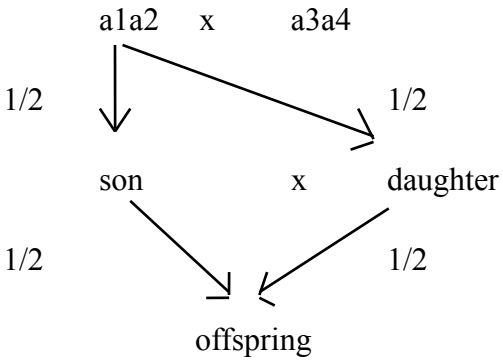
.5 → .25 → .125 → .0625 → .03125 →

.0156 → .0078 → .0039 → .0195 → .00975

Note this is change for one locus, same number applies to proportion of loci that will remain homozygous in an individual.
This is the way that tomato and wheat is bred, by a single outcross followed by self-pollination in conjunction with selection. These plants breed true because after 10 generations of self-pollination, they are 1 - .00975 = .99125% homozygous.

Inbreeding leads to higher frequencies of <u>homozygosity by descent</u>
This probability is called **inbreeding coefficient**.
Inbreeding coefficient is the probability of being homozygous BY decent, i.e. from common parent which appears in two lineages in the parents pedigree.

The probability of a single allele being passed though two lineages from one parent times the number of alleles that are in common parents. Full sib matting (brother-sister mating), means that grand parents are in both the father and mother's lineage for the offspring in question.



Probability of a1a1 homozygote is 1/2 x 1/2 x 1/2 x 1/2 = 1/16
Probability of homozygote for any allele is 4 x 1/16 = 1/4

Self	1/2
Full sib	1/4
parent-offspring	1/4
half sibs	1/8
cousins	1/16

Another question is:
What is the probability of an offspring of a consanguineous mating being affected by a rare genetic disease? For full sib mating:
Probability: that either grand parent is carrier = 2pq + 2pq
Probability: disease is passed to sib (1) = 1/2
Probability: disease is passed to sib (2) = 1/2
Probability: child being homozygous for disease = 1/4

(2pq +2pq) x 1/2 x 1/2 x 1/4 = pq/4, for full sib mating

For Cystic Fibrosis this would be 1/50 x 1/4 = 1/200
Contrast this to the rate from random mating: 1/2500 for q^2.

4. Other forces: Selection, Mutation, Migration, Drift (effect of small population size)

5. Models of equilibrium for balance between these forces that effect allele frequency.

LARGE SCALE CHROMOSOME CHANGES - Chapter 17

CHANGES IN CHROMOSOME NUMBERS

Euploidy is the unusual numbers of chromosome sets, i.e. those that differ from the diploid conformation. These include monoploid, and polyploid types. Polyploids include triploid, tetraploid, pentaploid, hexaploid etc. (3x, 4x, 5x and 6x respectively). Euploidy is rare in animals, but is common among plants species.

Many cultivated plants are natural polyploids. Wheat used for pasta (*Triticum turgidum*) is a tetraploid. It has four sets of seven chromosomes, 28 in all. These are classified as two diploid sets, of 14 chromosomes each in every cell nucleus. The sets are called A and B sets. The chromosome pairs within a set are called homologous chromosomes. Bread wheat *(Triticum aestivum*) is a hexaploid. It has six sets of 7 chromosomes or 42 in all. These sets are called the A, B and D sets (actually they are called genomes, though in wheat they behave as a single genome).

Other common tetraploid species include potato, tobacco, petunia, broccoli, and canola. Individuals with even number sets of chromosomes, i.e. tetraploid, hexaploid, octaploid, are often fully viable. Individuals with odd sets of chromosomes are usually sterile. For example, the cultivated banana is triploid; it is sterile and for this reason you don't find viable seeds in a banana.

Symbols:
"x" indicates the basic number of chromosomes in a set. "n" indicated the number of chromosomes in a gamete. For diploid organisms n = x. In polyploids x does not equal n; for example, in tetraploid pasta wheat (Triticum durum) 2n = 4x = 28; x = 7;
n = 14.

Triploidy leads to sterility because at meiosis the third homologous chromosome does not have a pairing partner and it segregates at random to one of the two daughter cells. Since this occurs for each chromosome threesome, gametes end up with some variable number of chromosomes between n and 2n. .

Monoploids are sterile because they do not have normal meiosis. They may occur spontaneously or can be induced to form under controlled culture conditions. Monoploids (or an individual of any other ploidy level) can be induced to double its chromosome level by treatment with **colchicine**. Doubled monoploids are diploid and are useful because they are fully homozygous for all their genes.

Tetraploids.
Hexaploids.

Autopolyploids and Allopolyploids.

Autopolyploids are polyploids for which both sets of chromosomes were derived from

the same parental species.

Allopolyploid are polyploids for which each set of chromosomes is derived from a different parental species.

Aneuploidy

Aneuploids differ in individual chromosome numbers. They have complete sets of chromosomes that are either missing or have extra individual chromosomes. Usually aneuploids differ by a single chromosome or chromosome pair from the normal individuals.

Aneuploids usually result from gametes that have had non-disjunction of chromosomes during meiosis.

Nullisomics (2n - 2)

Monosomics (2n - 1)

Turner females XO.

Trisomics (2n + 1).

Kleinfelters (XXY)

Downs syndrome (trisomic 21)

Aneuploidy for autosomes in humans is very deleterious. However aneuploids in polyploid species of plants are often viable and are used for research including gene mapping.

Genes are arranged in a fixed linear order along the chromosomes. This gene order is relatively constant within a species. However, chromosomal rearrangements sometimes do occur. They occur in natural populations and can be induced by radiation.

The typical full set of chromosomes in a cell of any particular organism is called the **karyotype.** Individual chromosomes are classified by their relative length and the location of the centromere.

CHANGES IN CHROMOSOME STRUCTURE

Metacentric chromosomes have centromeres in the center. **Acrocentric** chromosomes have centromeres off-center. **Telomeric** chromosomes have the centromere at the end of the chromosome.

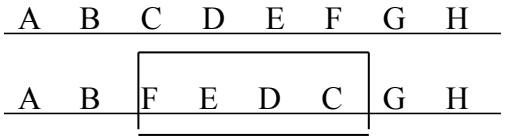
Chromosomal mutations (i.e. rearrangements) include duplications, deletions, inversions and translocations of sections of a chromosome.

Polytene chromosomes are found in the salivary gland of Drosophila. These large chromosomes facilitate the visualization of chromosomes and chromosome structure.

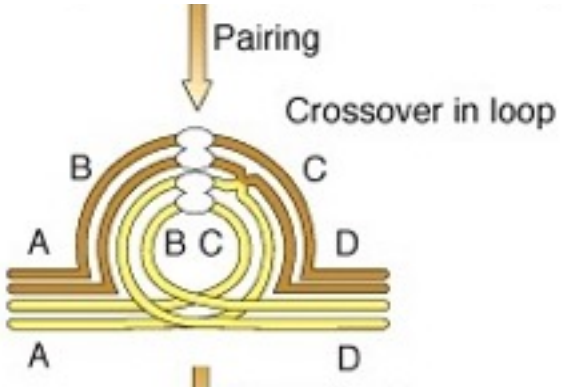
Deletions. Small parts of chromosomes are sometimes lost. Even "small" parts of chromosomes can contain many genes. Deletions are deleterious, especially in diploid organisms and large deletions in diploid organisms are not observed, since individuals with such deletions do not survive. Deletions can cause the loss of important genes, and can "uncover" deleterious recessive alleles of genes that are on the corresponding region of the homologous chromosome. Deletions can be recognized cytologically as looped regions in paired chromosomes. Cri du Chat is caused by a small deletion on chromosome 5 in humans.

Duplications. These are duplicated regions of a chromosome and carry extra copies of genes on from the duplicated region. Tandem duplications are adjacent to each other. Unequal crossing over can result in gene duplication. Duplicated genes are considered to an important source of new genes in evolution of a species.

Inversions. Inversions are regions of a chromosome that have flipped over:



Paracentric inversions do not include the centromere. **Paricentric inversions** include the centromere. Inversions alone are not necessarily deleterious. However in an individual who is heterozygous for an inversion, crossing over within the inversion prior to meiosis leads to bridge structures and chromosome breakage and loss of chromosome parts during meiosis in some of the gametes - causing partial sterility. Inversions inhibit crossing over within the inversion region in inversion heterozygotes.



Translocations.

Reciprocal translocations are the exchange of chromosome parts between two non-homologous chromosomes. Alone they are not deleterious, however in translocation heterozygotes, segregation during meiosis can lead to inviable gametes by gain or loss of particular chromosome regions. The loss and duplication of genetic material in some gametes leads to semisterility. In the segregation of translocation chromosomes at meiosis, adjacent-1 type segregation leads to inviable gametes, alternate segregation gives rise to viable gametes. This causes genes to appear to be linked between the two chromosomes - i.e. there is not independent assortment.

Genomics. Chapter 14

Genomics is the study of genomes in their entirety.
The genomes of several organisms have been completely sequenced. For many other organisms, large sets of sequence of cDNA clones constitute a large but partial insight into their genome composition. In addition, the similarity between model species and related species give additional insight into the uncharacterized portions of the related species. For example the complete genome sequence of rice and the experimental plant Arabidopsis has given important insight into the genome structure and function of many crop species.

Once the sequence of the human genome was known, it opened the the possibility of identifying mutations or alleles of specific genes that may cause genetic diseases for increase susceptibility to diseases. The application of this knowledge includes approaches such as current programs to identify all genes in tumors that are expressed differently than they are in normal tissue in order to discover key genes that may control cancer. These can subsequently be the targets of new drugs or chemotherapies that are more specific to different tumor types.

The field of genomics has been developed from molecular genetics. It has become possible through large-scale DNA sequencing, large scale measurement of gene expression by measuring changes in mRNA levels or the protein products for many genes, large-scale creation of experimental transgenic organisms and gene knockouts organisms to study the role of genes. In the gross sense, "large-scale" analysis means that thousands to tens of thousands of genes are characterized.

Structural genomics involves the analysis of gene sequence, gene number, order and physical nature of chromosomes.

Functional genomics studies the function of genes by studying gene expression, and the effect of modified gene expression through mutation, trans-gene expression or through comparative studies between different genotypes or by comparison between species. It asks, under what conditions, or in what tissue is a gene expressed. For example: What genes are expressed in tumors that are not expressed in normal tissue? Or: If we mutate a specific gene, how will the phenotype change? Will the mutation of a gene increase or decrease the development of cancer.

Bioinformatics is the use of computer analysis for structural or functional genomics.

Proteomics – the study of all the proteins of an organism.

Analyzing DNA sequence

DNA sequence. There are several databases of DNA sequences. Whole genome sequence is emphasized in the popular media, but there are many other important sets of DNA. There are thousands of gene sequences available for organisms that have not been completely sequenced. In addition, cDNA sequences are very important for analyzing genomes, but they are not usually

mentioned in media stories about genomes – probably because it is not yet possible to have the complete set of cDNAs that represent all the genes of an organism.

Types of DNA sequence:

1. Single gene sequence based on full-length cDNA clones that include the entire coding region and usually the 5' and 3' untranslated regions.

2. Single gene sequence based on genomic DNA sequence. These include coding region and all the introns and usually about 1000 base pairs (bp) of the DNA sequence upstream of the start of transcription – which normally contains the gene promoter, and some region downs stream of the coding region, usually a few 100 bp down stream of the 3' UTR.

3. Sequence of large genomic clones – often derived from Bacterial Artificial Chromosomes (BAC) of 50,000 bp (50 kilo bases (kb)) to 500 kb. These contain several adjacent genes and all the intervening sequence.

4. Whole genome sequence, normally based on sequencing several thousand BACs. This has been done for several species including many micro-organisms and a few plants and animals.

5. Expressed Sequence Tags (ESTs), are partial sequences of cDNA clones, usually from 300 to 900 bp long. The sequences are often long enough to identify the gene they encode based on similarity to sequences from other species. Projects typically sequence from 1000 to 100,000 ESTs from a single organism.

6. Tentative contigs (TCs). These are overlapping EST sequences from single species that can be combined to deduce longer and sometimes full-length cDNA sequences for single genes. They are called "tentative" because the sequences may have minor errors due to imperfections in the DNA sequence and because the assembly of individual overlapping sequences into longer contiguous sequence is done by computer algorithms that can sometimes make errors, such as combining sequences that are highly similar but may not be from the same gene.

Whole genome sequencing has been completed for several model organisms. The size of the genome of on organism is proportional to the complexity of the organism over large range of species. However there are very large differences in genome size with in species groups that are not related to complexity, for example the plant Arabidopsis has a genome of 125 million (Mega bases or MB) but another plant species, barley has 6000 Mega bases, even though the number of genes is similar.

Identification of genes in the DNA sequence is a major challenge. Locating the majority of genes by computer based sequence analysis is straight forward for most genes, but it is difficult to know if 100% of the genes have been identified. It may also be difficult to predict the precise intron and exon junctions.

The principal methods of identifying the sequence of genes in the genomic sequence are:

1. Finding similarity to cDNA sequences
2. Finding Long Open Reading Frames (ORFs) stretches of more than 100 bp with no stop codes in one reading frame.
3. Finding similarity to known genes in other species
4. Finding similarity to DNA sequence in other species

Genome structure

Much of the genome of eukaryots with large genomes, such as animals and plants, is composed of repeated sequences interspersed among the gene sequence. Much of repeated sequences are primarily derived from transposable elements that have replicated in the genome over long evolutionary time.

In humans, 3% of the genome is coding, 45% or more is derived from transposable elements, called transposons and retroposons. These are basically inactive hitchhiking DNA.

The Human Genome

The analysis of the sequence of the human genome was first published on Feb 16, 2001) in the journal Science 291: 1304-1350

The size of the human genome sequence is 2.91 Giga base pairs (Gbp)-
There are 26,383 annotated genes in the human genome.
There are 39,000 annotated and hypothetical genes. The latter are sequences that resemble genes but that have not been verified by the detection of a corresponding mRNA
42% of the annotated genes are of unknown function.
The gene with the most exons is Titin, which has 234 exons
The average human gene size is
27 kb

% of base pairs of spanned by genes 25.5 to 36.4%
% of base pairs of spanned by exons 1.1 to 1.4 %
% of base pairs of spanned by introns 24.4 to 36.4%
% of base pairs inn intergenic regions 75 to 64%

The largest intergenic region, or stretch of DNA with no genes, is on Chromosome 13. It is 3,038,416 bp long..

Exons in human genes average 150 bp and introns are larger. On average there are 10 exons per gene and 3 types of exon splicing variants per gene.
The mouse genome is 86% the size of the human genome 2.5 Gbp. Mouse has a similar number

of genes.

GENOME COMPLEXITY

How large are genomes in terms of amount of DNA and number of genes?

Whole genome sequencing has been completed for several model organisms. The size of the genome of an organism is proportional to the complexity of the organism over large range of species. However there are very large differences in genome size between species groups that are not related to biological complexity. For example the plant Arabidopsis has a genome of 125 Mega bases (MB) but barley has a genome of 6000 Mega bases, though the number of genes is similar. I.e. there is more space between genes.

Genome complexity for several model species:

Species	Size	Genes	Ave. gene	intron gene	Year sequenced
Escherichia coli bacteria	4.6 Mb	4000	1kb	none	1997
Saccharomyces cerevisiae yeast	12 Mb	6000	1.5 kb	0.03	1996
Neurospora crassa fungi	43 Mb	10,000	1.7	1.7	2003
Arabidopsis thaliania plant	125 Mb	25,000	2 kb	4	2000
Caenorhabditis elegans 1998 nematode	97 Mb	19,000	5 kb	5	
Drosophila melanogaster fruit fly	180 Mb	13,000	3 kb	4	2000
Mus musculus 2002 mouse	2500 Mb	28,000	40 kb	8.3	

JOINING DNA SEQEUNCE INTO CONTIGS

The genomes of organisms are millions to billions of base pair long. Each DNA sequence reaction will give about 300 to 800 bp of sequence. Thus long continuous sequence has to be pieced together from these individual sequences by identifying overlapping sequence. Contiguous sequences assembled from smaller sequences are called **Contigs.**

To join DNA sequences together we use a computer program to find the correct alignment between two or more overlapping sequences based on identical or near identical sequence:

```
Seq 1           GCCCGTGAGGATCTCGCCGCGTTGGAGAAGGACTACGAGGAGGTTGGCTCTGAGTCCGAC   60
Seq 2           ----------------------------------------------TCTGAGTCCGAC   12


Seq 1           GAGAATGAGGATGGCGATGATGGTGACGAGTACTAG-------------------------   95
Seq 2           GAGAATGAGGATGGCGATGATGGTGACGAGTACTAGAGGAGTCGTCGTCGTCTGGGGGCT   72


Seq 1           ------------------------------------------------------------
Seq 2           TGATGTTCTGTGTGTCAAGGCCTGATTGATAACTGCTGCTATCCCATGATCTGCCAGTGT    130
```

These sequences have a region of overlap with a perfect match; they can be combined to make a single sequence of about 180 bp. The search for overlap can be done on a large scale to combine sequences of 100,000's of bp.

The overlapping of many independent sequences for a single region gives high confidence in the accuracy of the sequence. To complete the human genome, DNA sequence equivalent of 10

times the size of the genome was sequenced, in order to have high confidence in the accuracy of the sequence and to have sufficient sequence to overlap all the individual sequences to build a proper contiguous sequence (or "contig") for each chromosome. Each region was sequenced an average of 10 times. The genomes of some other species have been sequenced with less redundancy to reduce the cost of sequencing.

Finding genes by searching the DNA sequence with the amino acid sequence of a previously known gene – deduced from its cDNA clone.

Gene sequences can be located by finding similarity to known gene sequence or the translation of the sequence into protein sequence. This can be done in two steps

(1) Translate the DNA sequence into amino acid sequence in the six possible reading frames and (2) search for similarity to know protein sequence.

Remember that codons are 3 nucleotides long and DNA has two strands. Thus, when analyzing DNA sequence for the first time there are 3 possible reading frames on each strand that may encode proteins. The reading frame starting with the 4th base is the same as that beginning with the 1st base.

The beginning amino acid sequence of rice actin gene is:
MRECISIHIGQAGIQVGNACWELYCLEHGIQADGQMPSDRTVGG

In the DNA sequence below, the beginning of the actin amino acid sequence is found in reading frame +3. This marks the beginning of the rice actin gene, the end of the first exon is also shown by the end of the region with matching to the translation of the nucleic acid sequence. The match begins again farther down stream, which is not shown.

5'3' Frame 1 –starts with the first base

```
tctccgcgcctccgcgcttttcctcctcctctcccctctctcccttctccgccgccgtcg
 S  P  R  L  R  A  F  P  P  P  L  P  S  L  P  S  P  P  P  S
cagcatcaacccaatccgccgccatgagggagtgcatctcgatccacatcggccaggccg
 Q  H  Q  P  N  P  P  P  -  G  S  A  S  R  S  T  S  A  R  P
gtatccaggtcgggaacgcgtgctgggagctctactgcctcgagcatggcatccaggtac
 V  S  R  S  G  T  R  A  G  S  S  T  A  S  S  M  A  S  R  Y
ggatccgcgtcccatctccctcacccccgtgttcttcgtgcctgcttctgggtcagatc
 G  S  A  S  H  L  P  H  P  P  C  S  S  C  L  L  L  G  Q  I
```

5'3' Frame 2 – starts with the second base

```
tctccgcgcctccgcgcttttcctcctcctctcccctctctcccttctccgccgccgtcgc
 L  R  A  S  A  L  F  L  L  L  S  P  L  S  L  L  R  R  R  R
agcatcaacccaatccgccgccatgagggagtgcatctcgatccacatcggccaggccgg
 S  I  N  P  I  R  R  H  E  G  V  H  L  D  P  H  R  P  G  R
tatccaggtcgggaacgcgtgctgggagctctactgcctcgagcatggcatccaggtacg
 Y  P  G  R  E  R  V  L  G  A  L  L  P  R  A  W  H  P  G  T
gatccgcgtcccatctccctcacccccgtgttcttcgtgcctgcttctgggtcagatct
 D  P  R  P  I  S  L  T  P  R  V  L  R  A  C  F  W  V  R  S
```

5'3' Frame 3 – starts with the third base.

The perfect match with the amino acid sequence of actin shows that the third reading frame in the coding frame. In comparison to an actin sequence of another species the match would not be

perfect. A match of even 75% similarity would be a reliable method to detect genes in the DNA of another species.

```
tctccgcgcctccgcgcttttcctcctcctctcccctctctcccttctccgccgccgtcgca
 S  A  P  P  R  F  S  S  S  S  P  L  S  P  F  S  A  A  V  A
gcatcaacccaatccgccgccatgagggagtgcatctcgatccacatcggccaggccggt
 A  S  T  Q  S  A  A  M  R  E  C  I  S  I  H  I  G  Q  A  G
                      M  R  E  C  I  S  I  H  I  G  Q  A  G
atccaggtcgggaacgcgtgctgggagctctactgcctcgagcatggcatccaggtacgg
 I  Q  V  G  N  A  C  W  E  L  Y  C  L  E  H  G  I  Q  V  R
 I  Q  A  D  G  Q  M  P  S  D  R  T  V  G  G
atccgcgtcccatctccctcacccccgtgttcttcgtgcctgcttctgggtcagatctg
 I  R  V  P  S  P  S  P  P  V  F  F  V  P  A  S  G  S  D  L
```

A computer program called **tBlastx** can do this search**.** It can check the translation of DNA against a database of known protein sequences. There are five versions of Blast, to compare nucleic acid (nt) sequences (Blastn), protein sequences (Blastp), protein to nt (tBlastn), nt to protein (blastx) and translated nt to translated nt (tBlastx).

LOCATING EXONS – THE CODING REGION

By comparing cDNA sequence to genomic DNA sequence you can locate genes in the genomic DNA sequence and identify introns and exons. Below you can see the alignment of a cDNA sequence starting with the initial ATG aligned with genomic sequence. This allowed us to locate the gene and to find the first exon and first intron. The dashed lines represent the intron in the cDNA sequence – it is missing since the cDNA does not have the introns. The intron extends from nt 95 to nt 380 in the genomic sequence.

```
                          THIS REGION IS AN EXON

cDNA-w3'UTR      ATGAGGGAGTGCATCTCGATCCACATCGGCCAGGCCGGTATCCAGGTCGGGAACGCGTGC 60
genomic          ATGAGGGAGTGCATCTCGATCCACATCGGCCAGGCCGGTATCCAGGTCGGGAACGCGTGC 60
                 ************************************************************

cDNA-w3'UTR      TGGGAGCTCTACTGCCTCGAGCATGGCATCCAGG-- THIS REGION IS INTRON -- 94
genomic          TGGGAGCTCTACTGCCTCGAGCATGGCATCCAGGTACATCTGTGGAACAAATACTCCACG 120
                 *********************************

cDNA-w3'UTR      ------------------------------------------------------------
genomic          CATGTATGGTAGTTTTGAAACGATCTTGTATCTTCCATTGTTGTAGTAACAACTAAATAA 180


cDNA-w3'UTR      ---------------THIS REGION IS INTRON ------------------------
genomic          AGTACAATTGTTCAATTATTGGGAATCGTATTTTCTGTAGTGCCGATGTACAGCATATTC 240


cDNA-w3'UTR      ------------------------------------------------------------
genomic          ATAGATGTCTATTTAGGAACTCAAATTTTAAATTGAGGACTAGTTATTTATTGTGGGTCA 300


cDNA-w3'UTR      ----------- THIS REGION IS INTRON --------------------------
genomic          GTCTTTTGAATTGTGTTATCTTGCTGTACTGAAATAATAATGTACCACTAAGGCGCTAAC 360

                                        THIS REGION IS AN EXON
cDNA-w3'UTR      --------------------CTGATGGTCAGATGCCCAGTGACAGGACTGTTGGTGGAGG 134
genomic          ATGTATTTTGTCCTTTCAGGCTGATGGTCAGATGCCCAGTGACAGGACTGTTGGTGGAGG 420
                                    ****************************************
```

```
cDNA-w3'UTR      TGATGATGCTTTCAACACCTCCTTCAGTGAGACTGGTGCTGGGAAGCATGTTCCCCGTGC 194
genomic          TGATGATGCTTTCAACACCTTCTTCAGTGAGACTGGTGCTGGGAAGCATGTTCCCCGTGC 480
```