University of Jordan
Faculty of Medicine

GENETICS & Molecular Biology

Number: 35 (7 for Dr.Mazin)

Done By: Hashim A. Mohammad

Corrected By: Abdullah S. Sulaiman

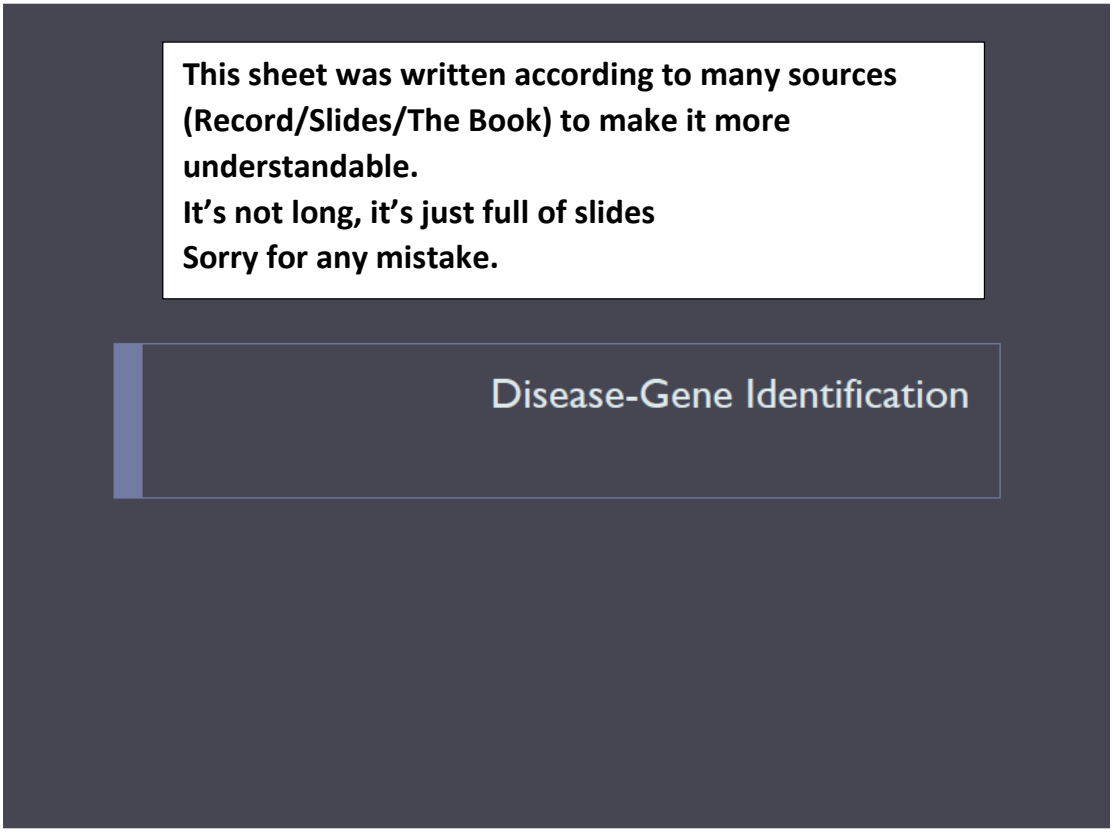
Subject: Linkage analysis

Doctor: Mazin Al-salihi

Price:

Date:

This sheet was written according to many sources (Record/Slides/The Book) to make it more understandable.
It's not long, it's just full of slides
Sorry for any mistake.



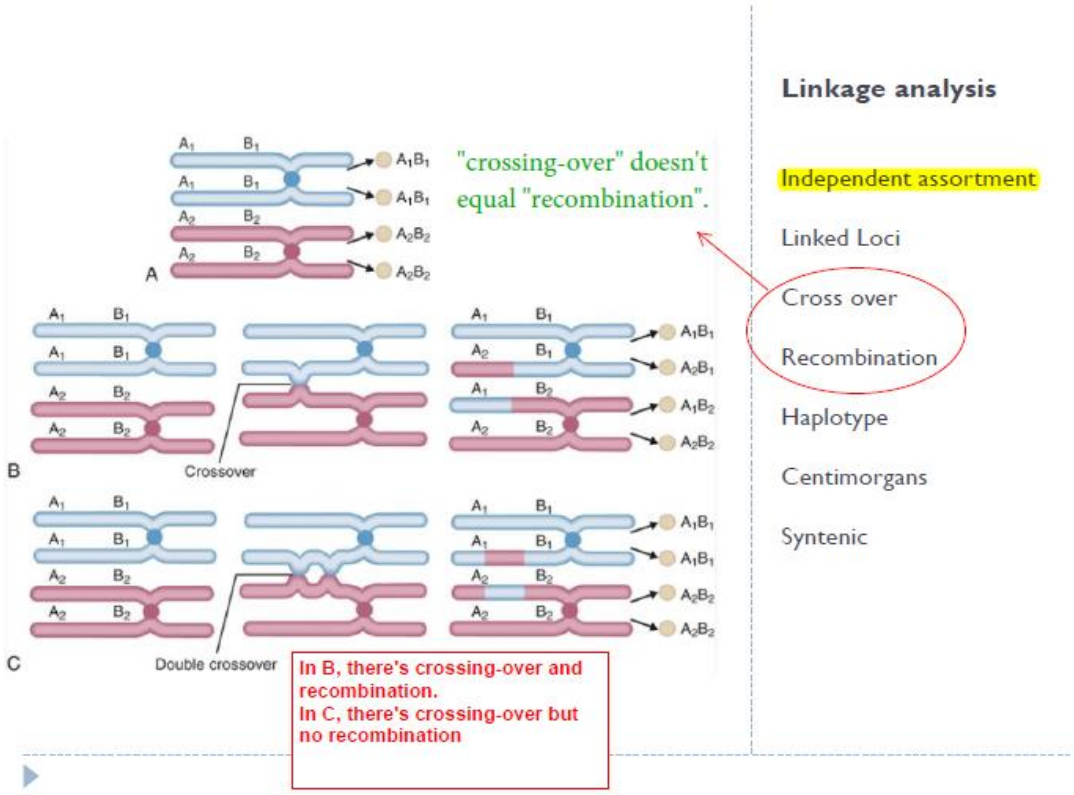
- Here, we are concerned with knowing the specific mutation responsible for a certain disease. This would be beneficial in many aspects of which the following are the most important:

  1- Understanding the disease and thus be more able to diagnose, and treat it.

  2- Getting a more accurate prognosis.

  3- If we identify the gene, we can know its DNA sequence and thus figure out the RNA product and the protein product, as well. Hence, we can prevent the occurrence of the disease. For example, we can prevent gene expression (i.e. we get rid of the protein causing the disease). Also, using recombinant DNA techniques, we can manufacture a normal gene product (i.e. a normal protein) for more effective treatment , [ex: Clotting factor VIII used in the treatment of hemophilia, and insulin used in diabetes type I] .

  4- Gene therapy



∞ Linkage analysis:

Independent assortment (2nd law of Gregor Mendel) states that genes are transmitted from one generation to the next independently of one another (i.e. they are unlinked and they don't travel together). This applies for most of genes but not all of them. Genes that occupy the same region on the chromosome (close to each other) tend to be linked and thus travel together from one generation to the next.

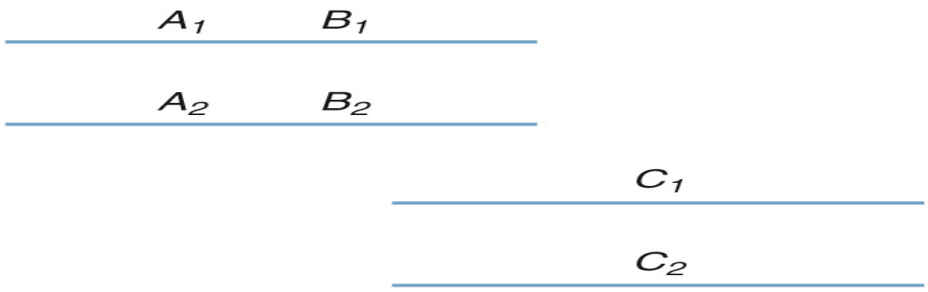
Independent assortment:
هو القانون الثاني لمندل ويسمى بالعربية **قانون التوزيع الحر**

Haplotypes:
A_1 and B_1 are present close to one another on the same chromosome, as do A_2 and B_2. The combination of alleles on one chromosome is called a "haplotype". So, A_1/B_1 and A_2/B_2 are haplotypes.

Crossing over doesn't necessarily result in recombination. In case of double crossing-over, there's no recombination.

Syntenic loci: are the loci present on the same chromosome.
Syntenic means “on the same thread”.

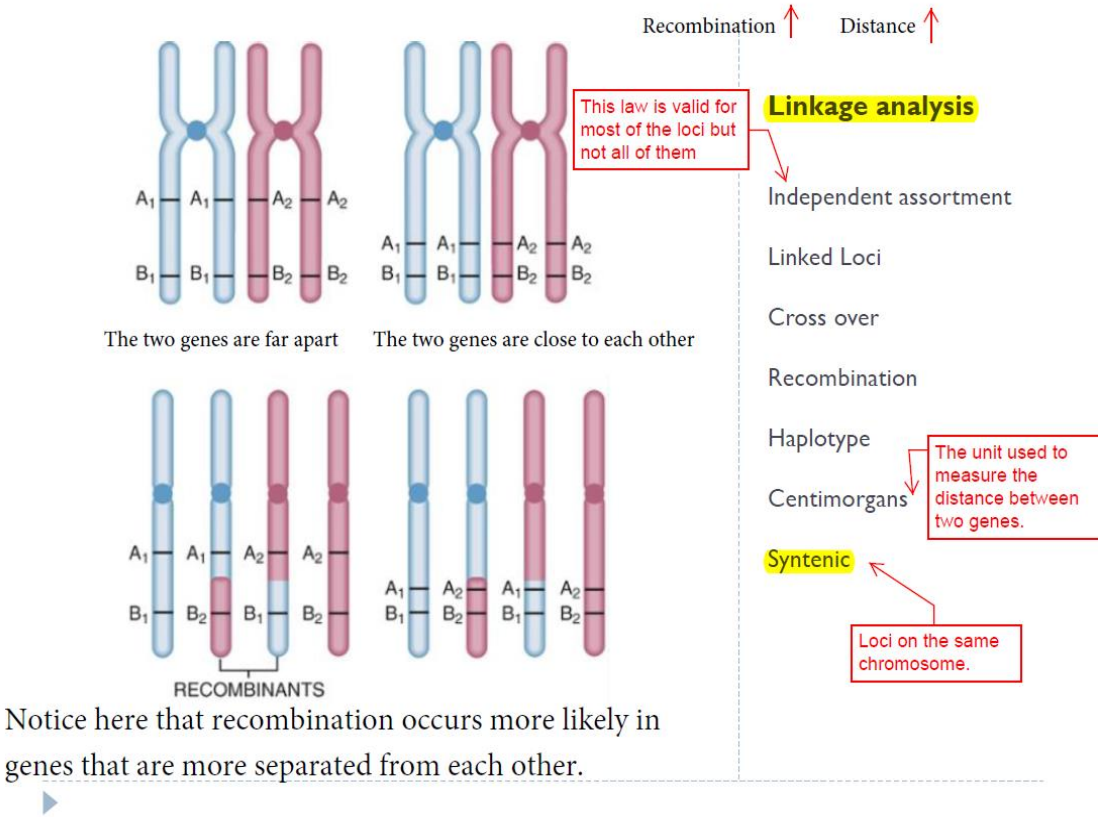
- To understand the concept of linkage and independent assortment, see this example (it wasn’t mentioned in the lecture, but I recommend reading it).



- A_1/B_1 and A_2/B_2 are haplotypes. A_1 and B_1 are present on the same chromosome. C_1 and C_2 are present on another chromosome. So, “A”s and “C”s will not travel together, because they are already on different chromosomes, hence following the independent assortment law (unlinked loci). On the other hand, A_1 and B_1 are on the same chromosome, so they may travel together (i.e. be linked loci) or not (i.e. be unlinked loci). What determines being linked or unlinked is recombination. If recombination occurs frequently, then A_1/B_2 and A_2/B_1 haplotypes are expected to be seen in gametes as a result of recombination, and thus the two loci that were present on the same chromosome don’t travel together (they are non-linked). If recombination is rare, then all gametes will have either A_1/B_1 or A_2/B_2 haplotypes, and thus we can say that the two loci are linked.

But now, what determines recombination frequency?
The distance between the two loci, as *recombination occurs more frequently when the two loci are far apart from each other*.

Memorize it in this way:
Genes that are closer to each other, love each other and travel together (linked genes). Genes that are far apart hate each other and tend not to travel together (unlinked genes).



- Cross-overs are more likely to occur between loci that are situated far apart on a chromosome than between loci that are situated close together. So, the more two genes are far apart, the more likely we may get a cross-over. Hence, we can make a relationship between distance and recombination frequency.

- The unit used to measure the distance between two loci is cM (centimorgans).
- If the recombination frequency is 1%, approximately, the distance between the two loci is one cM. One cM corresponds approximately to one million base pairs. This number isn't accurate because the frequency of recombination isn't the same in all regions on the chromosome.

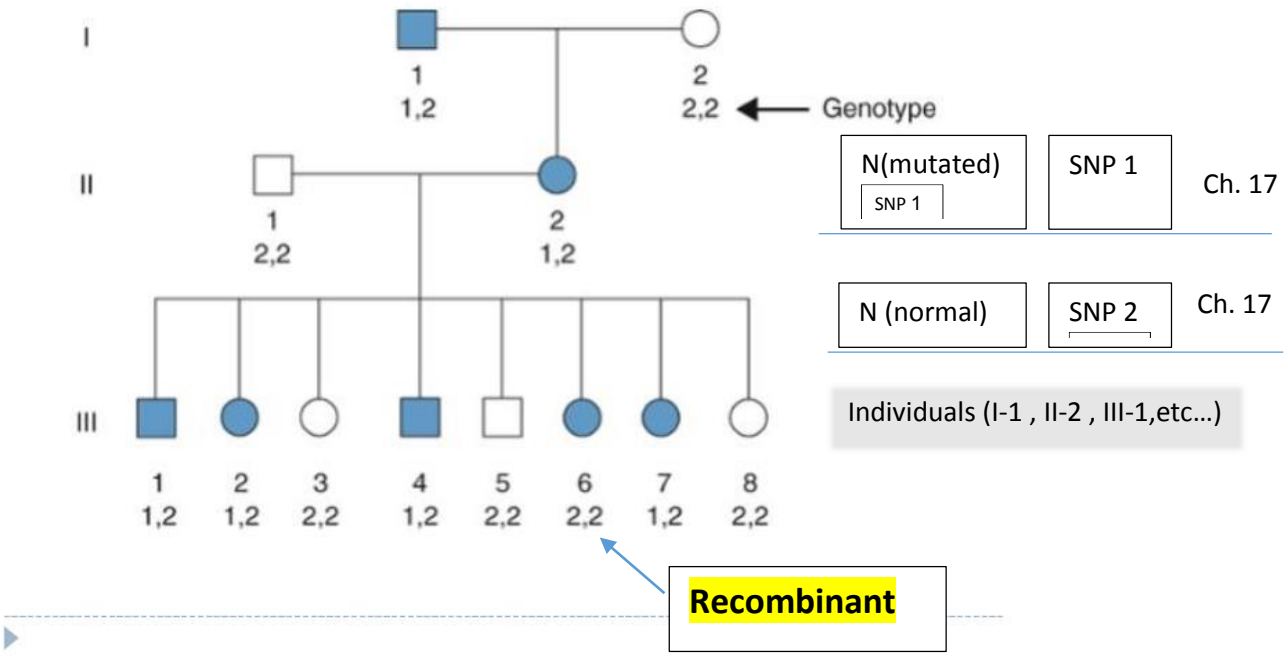
- **Factors affecting the rate of crossover**:

I. Crossovers in female meiosis = 1.5 times that in male meiosis.

II. Higher recombination frequency near telomeres.

III. Recombination hot spots: small regions on the chromosome that exhibit 10-fold increase in recombination rate.

☞ How can we know recombination frequency, and then figure out an approximation of the distance between two loci?
We have to look at large <u>family pedigrees</u>, find out the recombination frequency. For example, if it was 1% we can say that the distance between the two loci is 1 cM.



Recombination frequencies can be estimated by observing the transmission of genes in pedigrees. The figure shown above is an example of a pedigree in which neurofibromatosis type 1 (NF1) is being transmitted. The members of this pedigree have also been typed for a two-allele single nucleotide polymorphism (SNP) which, like the *NF1* gene, is located on chromosome 17. The SNP genotypes are shown below each individual's number in the pedigree. Examination of generations I and II allows us to determine that ,*under the hypothesis of linkage between NF1 and the SNP,* the disease-causing mutation in the *NF1* gene must be on the same copy of chromosome 17 as allele *1* of the SNP in this family, because individual I-2, who is homozygous for allele *2,* is unaffected with the disease. Only the affected father (I-1), who is a heterozygote for the SNP, could have transmitted a copy of chromosome 17 that contains both the *NF1* disease allele and SNP allele *1* to the daughter (II-2).

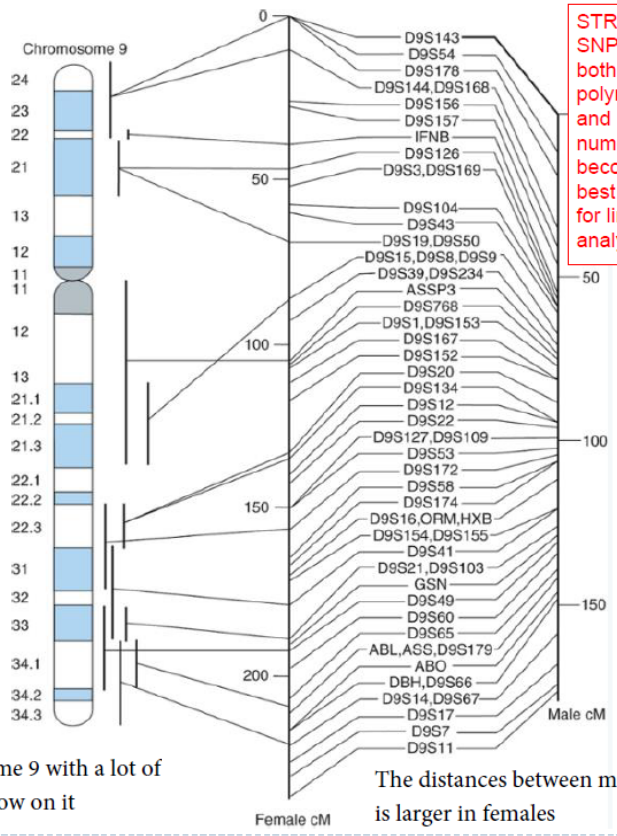
The arrangement of these alleles on each chromosome is referred to as **linkage phase.** With the linkage phase known, individual II-2's haplotypes would then be *N1/n2,* where *N* indicates the mutated allele causing *NF1, n* indicates the normal allele, and *1* and *2* are the two SNP alleles (in other words, individual II-2 has one copy of chromosome 17 that contains both the disease-causing mutation *N* and SNP allele *1,* and her other copy of chromosome 17 contains the normal allele *n* and SNP allele *2*). This woman's husband (individual II-1) is not affected with the disease and is a homozygote for SNP allele *2.* He must have the haplotypes *n2/n2.* If the *NF1* locus and the SNP are linked, the children of this union who are affected with NF1 should usually have SNP allele *1,* and those who are unaffected should have allele *2.*In seven of eight children in generation III, we find this to be true. In one case, a recombination occurred (individual III-6). This gives a recombination frequency of 1/8, or 12.5%, supporting the hypothesis of linkage between *NF1*and the SNP. A recombination frequency of 50% would support the hypothesis that the two loci are not linked. Note that the pedigree allows us to determine linkage phase in individual II-2, but we cannot determine whether a recombination took place in the gamete transmitted to II-2 by her father. Thus, the recombination frequency is estimated only in the descendants of II-2. In actual practice, a much larger sample of families would be studied to ensure the statistical accuracy of this result.

- The mating in the first generation is uninformative.

- If there's a disease caused by one-gene mutation. How can we know where is this gene on the chromosome?
  We have a lot of markers that we know their locations on the chromosome. If the disease-causing gene is linked to one of these markers, this means that they are close to each other, and thus we will know where the gene is.
- In the past, a few markers where known, that's why linkage analysis wasn't that much useful and successful. Now, we know large numbers of markers, so it's becoming easier to map a disease-causing gene very quickly. Here, it's noteworthy to remind ourselves with great effort made in the Human Genome Project, which enables us to know all of these markers.

- **Two characteristics of a marker are important**:
  - To be highly polymorphic.
  - Numerous loci.

To understand the importance of these characteristics, read page 159.



Chromosome 9 with a lot of markers show on it

The distances between markers is larger in females

STRs and SNPs are both highly polymorphic and numerous, becoming the best markers for linkage analysis

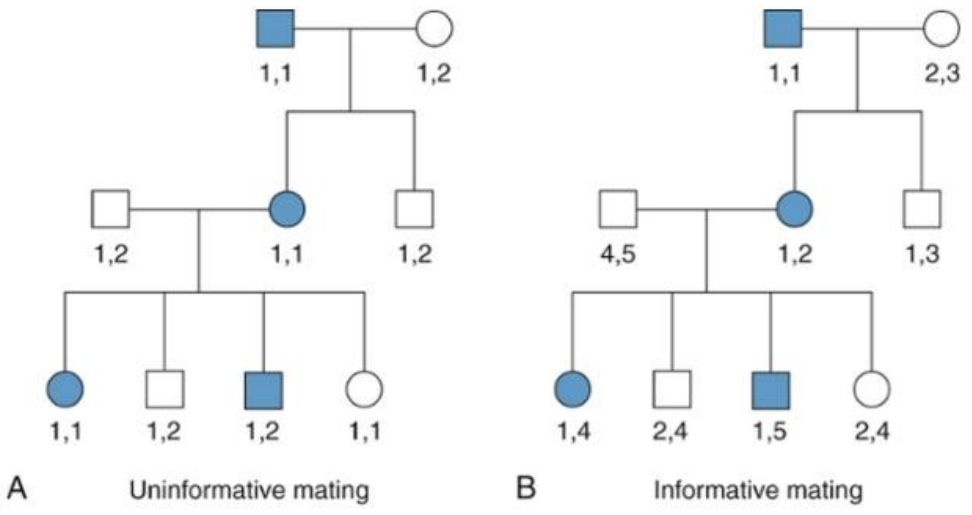
Linkage analysis

Useful markers:
- Highly polymorphic
- Numerous loci

In a 1000-km-long highway, how can we help drivers know where they are?

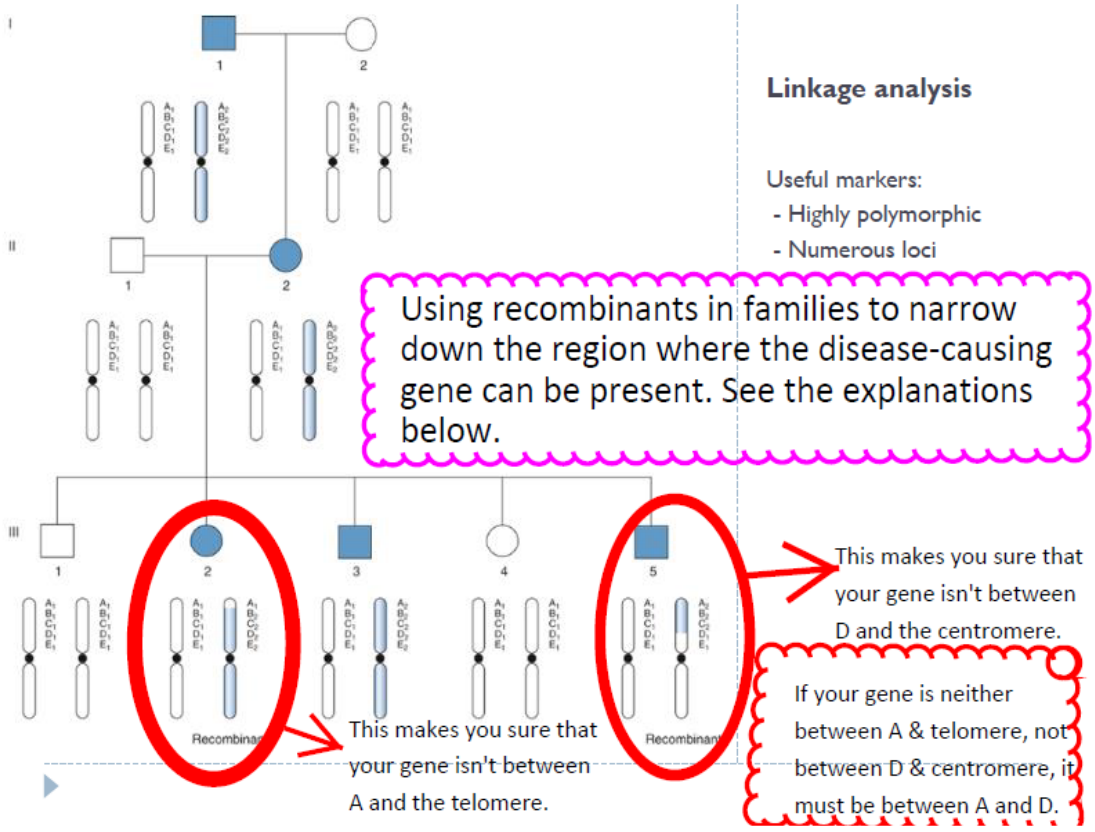
In the US, they use mile markers.

Chromosomes contain billions of base pairs, so they are very similar to such long highways. How can we know where we are on the chromosome? Similarly, by markers. But our markers here are polymorphic alleles rather than mile markers. If the disease-causing gene is linked to a certain polymorphic allele, we can say that "Oh! the gene and the marker are linked, so they are definitely very close to each other."



A- **<u>Uninformative mating</u>**:
Uninformative means that mating doesn't allow you to know the disease-causing gene is linked to which marker.
Here, mating in generation I is not informative (uninformative mating).

B- **<u>Informative mating</u>**: Here, we can obviously notice that the disease-causing gene is linked to marker 1 .



The figure is not completely clear here. See it on page 162

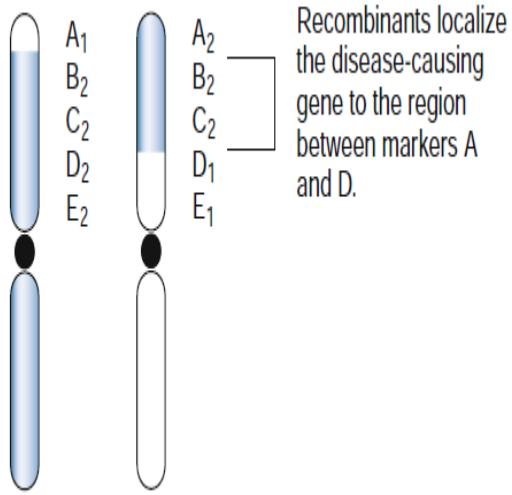
- Using one marker, we can know approximately where the disease-causing gene is. Nevertheless, narrowing down the location of the gene is much better; hence we use more than one marker.

- Observation of recombinations within families enables a researcher to narrow down the location of the disease-causing gene.
(I know that the idea still unclear, be patient, the following text will explain everything to you, just read). Keep looking at the figure while reading.

- Suppose that a series of marker polymorphisms, labelled *A, B, C, D,* and *E,* are all known to be closely linked to a disease-causing gene. The family shown in the figure above has been typed for each marker, and we observe that individual II-2 carries marker alleles A_2, B_2, C_2, D_2, and E_2 on the same copy of the chromosome that contains the disease-causing mutation (linkage phase). The other (normal) copy of this chromosome in individual II-2 carries marker alleles A_1, B_1, C_1, D_1, and E_1. Among the affected offspring in generation III, we see evidence of two recombinations. Individual III-2 clearly inherited marker allele A_1 from her affected mother (II-2), but she also inherited the disease-causing mutation from her mother. This tells us that there has been a recombination (crossover) between marker *A* and the disease-causing gene. Thus, we now know that the region of the chromosome between marker *A* and the telomere cannot contain the disease-causing gene (because she lost the region between A and the telomere and she's still diseased).

We observe another recombination in the gamete transmitted to individual III-5. In this case, the individual inherited markers D_1 and E_1 but also inherited the disease-causing mutation from the mother. This indicates that a crossover occurred between marker locus *D* and the disease-causing locus. We now know that the region between marker *D* and the centromere (including marker *E*) cannot contain the disease-causing locus.

- These two key recombinations have thus allowed us to substantially narrow the region that contains the disease-causing locus.



Two recombinants:
1- Individual III-2: from this, we can know that the disease-causing gene isn't present in the region between A and the end of the chromosome (the telomere).

2- Individual III-5: from this, we can know that the disease-causing gene isn't present in the region between D and the centromere.

Excluding these two regions, we can localize the disease-causing gene to the region between markers A and D.

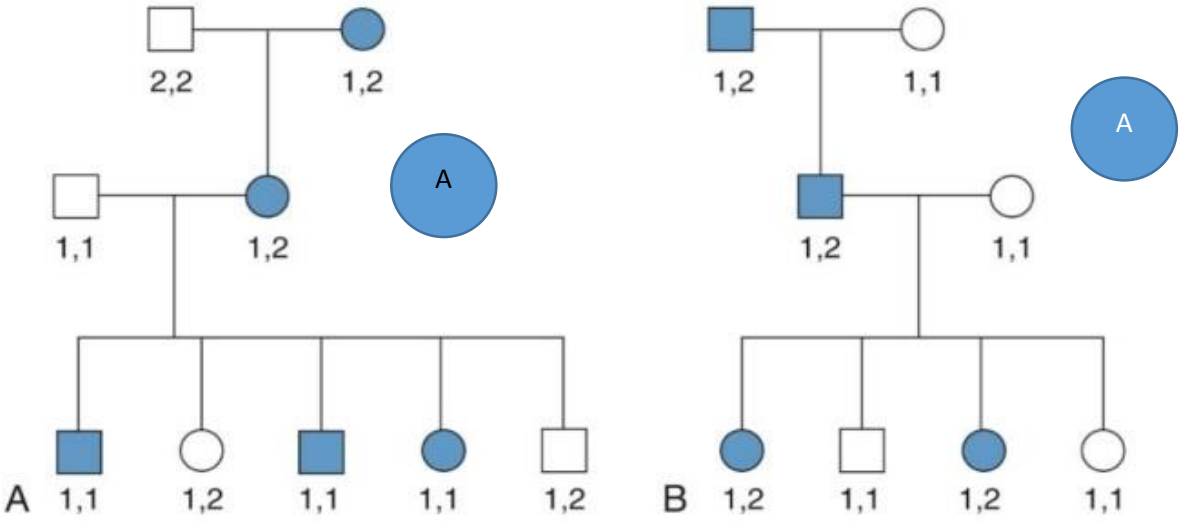
This is what I meant by “Using recombinants in families to narrow down the region where the disease-causing gene can be present”.

Now, revise the concept from the edited slide that’s present at the beginning. I hope everything is clear now.

Linkage Disequilibrium:

Non-random association of alleles at linked loci within families, one allele of a marker locus will usually be transmitted along with the disease-causing allele if the marker and disease loci are linked.

For example, allele I of a linked two-allele marker could co-occur with the Huntington disease (HD) allele, located on chromosome 4, in a family. This association is part of the definition of linkage. However, if one examines a series of families for linkage between HD and the marker locus, allele I will co-occur with the disease in some families (A), and allele 2 of the marker will co-occur with the disease in others(B) .



This reflects two things :

- First, disease-causing mutations might have occurred numerous times in the past, sometimes on a copy of chromosome 4 carrying marker allele 1 and other times on a copy of chromosome 4 carrying marker allele 2.
- Second, even if the disease is the result of only one original mutation, crossovers occur-ring through time will eventually result in recombination of the marker and disease alleles.

A disease-causing allele and a linked marker allele will thus be **associated within families but not necessarily between families**. In other words, if we examine a marker locus and a disease-causing locus in a large series of families in a population, we do not necessarily expect that one specific marker allele will be associated with the disease-causing mutation in most or all families.

Sometimes, however, we do observe preferential association of a specific marker allele and the disease-causing allele in a population. The chromosome haplotype consisting of one marker allele and the disease-causing allele is found **more often** than we would expect based on the frequencies of the two alleles in the population.

NOTE (JUST TO GET THE IDEA):

Suppose, for example, that the disease-causing allele has a frequency of 0.1 in the population (disease associated frequency) and the frequencies of the two alleles (labelled 1 and 2) of the marker locus are 0.4 and 0.6, respectively (marker population frequency). Assuming statistical independence between the two loci {i.e., linkage equilibrium), the multiplication rule predicts that the population frequency of the haplotype (haplotype frequencies) containing both the disease-causing allele and marker allele 1 would be 0.1 x 0.4 = 0.04. By collecting family information, we can directly count the haplotypes in the population. If we find that the actual frequency of this haplotype is 0.09 instead of the predicted 0.04, then the assumption of independence has been violated, indicating preferential association of marker allele 1 with the disease allele. This association of alleles at linked loci is termed linkage disequilibrium.

Figure (8-10) : **The doctor did not mention the disease , but discussed the topic.**

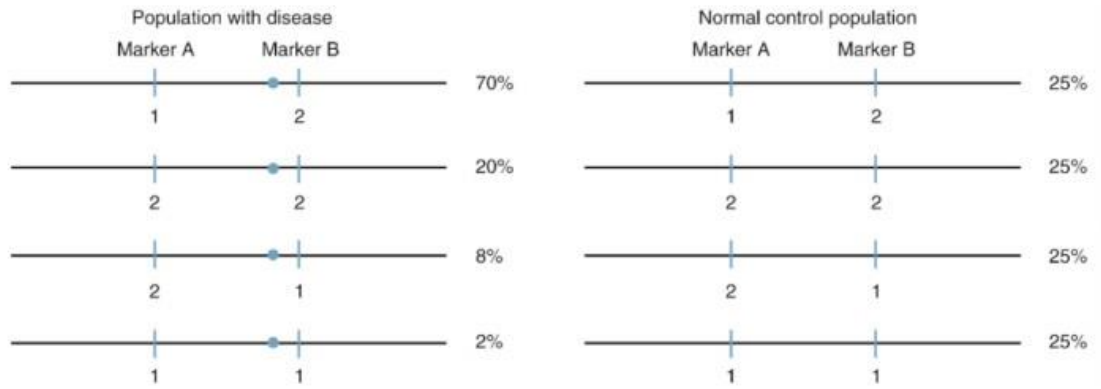


Figure 8-10 illustrates how linkage disequilibrium can come about. Imagine two marker loci that are both linked to the myotonic dystrophy locus on chromosome 19. Marker B is closely linked. Marker A is less closely linked. Because each of these marker loci has two alleles (denoted 1 and 2), there are four possible combinations of marker alleles at the two loci, as shown in Figure 8-10. When a new myotonic dystrophy mutation first occurs in a population, it can be found on only one copy of a chromosome, in this case the one with the A1B2 marker combination. As the disease-causing mutation (allele) is passed through multiple generations, crossovers will occur between it and the two markers. Because the disease locus is more closely linked to marker B than marker A, fewer cross-overs will occur between the disease-causing allele and marker B. As a result, the disease-causing allele is found on a B2 containing chromosome 90% (70%+20%) of the time, and it is found on an A2 containing chromosome 72% (70%+2%) of the time. The degree of linkage disequilibrium is stronger between marker B and the disease-causing allele than between marker A and the disease-causing allele. Notice also that both the A1 and the B2 alleles are still positively associated with the disease-causing allele, because each marker allele has a much lower frequency (50%) in the population of individuals who lack the disease-causing allele (see Fig. 8-10). If enough generations elapse, recombination would eventually eliminate the allelic associations completely, and the loci would be in linkage equilibrium. Because linkage disequilibrium is a function of the distance between loci, it can be used to help infer the order of genes on chromosomes. Linkage disequilibrium provides an advantage over linkage analysis in that it reflects the action of recombinations that have occurred during dozens or hundreds of past generations (i.e., the number of generations that have elapsed since the disease-causing mutation first occurred in a population). Linkage analysis, in contrast, is limited to the recombinations that can be directly observed in only the past several generations. Consequently, there are seldom enough recombinants in a series of families to map a gene to a region smaller than several centimorgans using linkage analysis, whereas linkage disequilibrium analysis can some-times map a gene to an interval of 0.1 cM or less.

However, linkage disequilibrium can be influenced by evolutionary forces, such as natural selection or genetic drift that have acted during the history of a population.

- ✓ *Linkage disequilibrium between loci diminishes through time as a result of recombination. It can be used to infer the order of genes on chromosomes.*

Physical mapping

If you don't understand Arabic, skip this page. It's not that much important.

إذا كنت تبحث عن إبرة في كومة قش، فكم سيكون من الصعب عليك إيجادها؟
صعب جدًّا ... أليس كذلك؟
إذن ماذا ستفعل حتى تجدها؟
أمامك طريقان، إما أن تبحث وتبحث حتى تجدها، أو أن تستخدم مؤشرات أو أدلة تساعدك على إيجادها. وهذا حالنا في إيجاد الجين المسؤول عن إحداث المرض، فكم من الصعب أن نجد الجين الذي نريد بين آلاف الجينات الممتدة على طول الكروموسوم؟؟
إذن يجب أن نبحث عن طرق ملتفة وذكية لتسهيل مهمتنا ...

الأولى، أن نربط مكان وجود الجين المسؤول عن إحداث المرض ببعض الاختلالات الكروموسومية كالحذف أو المضاعفة (Deletion and Duplication Mapping) أو نربط المرض بكمية البروتين/الإنزيم الموجودة ونقارنها بعمليات الحذف أو المضاعفة ونربط ذلك بالجينات لكشف جين المرض (Dosage Mapping) .
الثانية، أن نركز بحثنا في منطقة معينة محدودة من الكروموسوم ونهمل المناطق الأخرى وبالتالي يصبح مكان البحث أصغر وتصبح المهمة أسهل، فإذا أهملنا المناطق التي تحتوي على جينات مصمتة (Silent genes) أو المناطق التي تبدو لنا غير مهمة فسيصبح مكان البحث أصغر وأصغر. وهذا ما نسميه (The Use of Functional v.s Nonfunctional DNA for disease-gene identification).
الثالثة، أن نرشح جينات محددة لتكون هي الجينات المسؤولة عن إحداث المرض بناء على وظائف البروتينات الناتجة عنها وطبيعة المرض وبالتالي نستطيع تضييق المكان الذي نبحث فيه، وهذا ما نسميه (candidate genes).

حقيقًة، هذه الصفحة هي آخر ما كتبت، وما جاءت في هذا المكان إلا لأنني شعرت بعدم ترابط المواضيع المشروحة بعد ذلك، فقد يتوه القارئ عن الهدف الأسمى مما سيأتي، ألا وهو تسهيل البحث عن الجين المسؤول عن إحداث المرض، فلعل هذه المقدمة تكون مفيدًة موجِّهَة مُعِيْنَة. وفقكم الله وأعانكم ...

Physical Mapping and Cloning

Linkage analysis allows us to determine the relative distances between loci, but it does not assign specific chromosome locations to markers or disease-causing genes. Physical mapping, which has involved a variety of methods, accomplishes this goal, and considerable progress has been made in developing high-resolution physical mapping approaches.

Chromosome Morphology

A simple and direct way of mapping disease-causing genes is to show that the disease is consistently associated with a cytogenetic abnormality, such as a duplication or deletion. Such abnormalities might have no clinical consequences themselves (thus serving as a marker), or they might cause the disease. Because these approaches are historically the oldest of the physical mapping approaches, they are discussed first.

Deletions

Karyotypes or array CGH of patients with a genetic disease occasionally reveal deletions of a specific region of a chromosome. This provides a strong hint that the locus causing the disease might lie within the deleted region. The extent of a deletion can vary in several patients with the same disease. Deletions are compared in many patients to define the region that is deleted in all patients, thereby narrowing the location of the gene (See the figure below).

Examples showing the importance of deletion detection in physical mapping: Deletion mapping has been used, for example, in locating the genes responsible for:

1- retinoblastoma
2- Prader–Willi and Angelman syndrome.
3-Wilms tumor

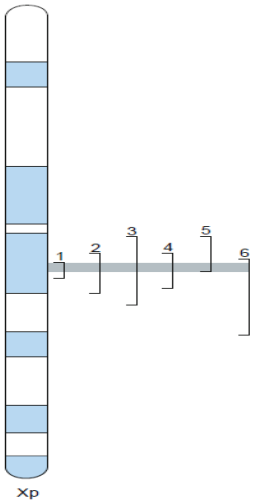


FIGURE 8-12
Localization of a disease-causing gene through deletion mapping. A series of overlapping deletions is studied in which each deletion produces the disease phenotype. The region of overlap of all deletions defines the approximate location of the gene.

Translocations

Balanced chromosome translocations often have no effect on a translocation carrier because the individual still has a complete copy of his or her genetic material. However, when a translocation happens to interrupt a gene, it can produce genetic disease. For example, after linkage analysis had mapped the *NF1* gene approximately to the long arm of chromosome 17, a more refined location was obtained when two patients were identified—one with a balanced translocation between chromosomes 17 and 22, and the other with a balanced translocation between chromosomes 17 and 1. The breakpoints of these translocations on chromosome 17 were located very close to each other, in the same region implicated by linkage analysis. They provided a physical starting point for experiments that subsequently led to identification of the *NF1* gene.

A similar example is provided by translocations observed between the X chromosome and autosomes in females with Duchenne muscular dystrophy (DMD). Because this is a lethal X-linked recessive disorder, affected homozygous females are rare. The translocation breakpoint on the X chromosome was found to be in the same location (Xp21) in several affected females, suggesting that the translocation interrupted the *DMD* gene. This proved to be the case, and these translocations aided considerably in mapping and identifying the *DMD* gene. (Although these females also carried a normal X chromosome, the normal X was preferentially inactivated, leaving only the interrupted X as the active chromosome.)

Dosage Mapping Using Deletions and Duplications

Examples are not for memorization

Before going into the whole area of genetics, we were depending on biochemical assays to detect genetic diseases (i.e. we were looking for a substance that does accumulate over time or a product that's not produced efficiently, and thus know

that the enzymatic activity is reduced, or we have more of the product and thus increased enzymatic activity). Now, we can combine these biochemical assays and karyotypes for physical mapping of genes. Here, any deletion indicates a decrease in enzymatic activity, whereas a duplication indicates an increase. See the following text and figures to understand this concept better.

When a deletion occurs on a chromosome, it stands to reason that the protein products encoded by genes in the deleted region will be present in only half the normal quantity. This is the basis of a simple approach known as **dosage mapping.** For example, it was observed that a 50% reduction in the level of the enzyme adenylate kinase was consistently associated with a deletion on chromosome 9, mapping the adenylate kinase gene to this chromosome region.

We used a deletion/the amount of the enzyme to know the location of a gene (Dosage Mapping).

Similarly, a duplication of chromosome material should be associated with an increase in gene product levels. Because three genes are present instead of two, the increase should be approximately 50% above normal. This form of dosage mapping was used to assign the gene encoding superoxide dismutase-1 (SOD-1) to the long arm of chromosome 21.

As a conclusion, what we mean by "Dosage Mapping" is that we establish a relationship between the amount of the protein present with certain chromosomal abnormalities to know where the gene is (i.e. on which chromosome).

Examples:

Deletion on chromosome 9 → 50 % reduction in the amount of the enzyme → so the gene coding for this enzyme must have been on chromosome 9.

Duplication on chromosome 21 → 50 % increase in the amount of the enzyme → so the gene coding for this enzyme must have been on chromosome 21.

See slide 10.

Positional Cloning

Sometimes the gene product responsible for a genetic disease is known before the gene itself is identified. This was the case, for example, with the β-globin polypeptide and sickle cell disease. In such cases, one can deduce the DNA sequence from the amino acid sequence of the polypeptide; this DNA sequence can be used to make a probe in order to locate the disease-causing gene. This type of approach, in which the gene product and its function are used to pinpoint the gene, is an example of **functional cloning** (i.e., identifying a specific gene or gene segment and making copies of it for study.

More often, however, we have only a linkage result that has localized the disease-causing gene to a region near the linked marker polymorphism (the locations of these markers have been established previously). Because of the limited resolution of linkage analysis, the region that contains the disease-causing locus may be several mega-bases or larger and can easily contain dozens of genes interspersed with noncoding DNA. A common approach has been to begin with a linked marker and then canvas the region around the marker to locate and identify the disease-causing gene itself. Because this process begins with an approximate knowledge of the gene's position on a chromosome, it has traditionally been termed positional cloning.

Functional versus Nonfunctional DNA

- Conserved DNA sequences are always important.
- Some of our genetic material is conserved back to yeast.
- Conservation isn't only seen in coding regions. 5 % of our non-coding regions is highly conserved and therefore likely to perform regulatory functions (Remember: Conservation is an indication of importance).
- See slide 14. There's a lot of consensus between the genetic material of the fly protein and that of respiratory cilia.

Most of our DNA sequence has no known function and is unlikely to contribute to disease. Thus, in searching for disease-causing alterations, we typically focus on DNA that encodes proteins or

performs important regulatory functions (e.g., enhancer or promoter sequences). Because of their functional significance, coding DNA or regulatory DNA sequences generally cannot change very much through the course of evolution. This means that such DNA sequences will be **conserved,** or similar in base-pair sequence, in many different species. In contrast, nonfunctional DNA sequences are likely to change more rapidly and to differ substantially among species. DNA sequences can be compared among species using computer algorithms to distinguish functional (conserved) DNA from nonfunctional (nonconserved) DNA. As the genomes of more and more species are sequenced, these comparisons are done increasingly by comparing whole-genome sequences among species. Such comparisons have shown that about 5% of the noncoding human genome is highly conserved and is therefore likely to perform regulatory functions.

Summary:
If we find the conserved sequences, we will just focus on them and narrow down the region in which the disease-causing gene is present. So, instead of looking at the whole genome, we will just look at 5-6 % of it, therefore making our task much easier.

Most CG dinucleotides are methylated. However, approximately 60% of human genes have unmethylated CG dinucleotides **(CG islands)** in their 5′ region. (A lack of methylation in the 5′ region of the gene probably makes it more accessible to transcription factors required for active expression.) The identification of a series of CG islands has often been used to pinpoint the locations of coding genes.

Conclusion:

Conserved sequences→ important sequences that code for important proteins → it's a region where our gene can be present.
Unmethylated region → expressed (not silent) → our gene can be present her, as well.
By looking at conservation and methylation status, we can know where the disease-causing gene is more likely to be present, hence minimizing the area

where we have to search for the gene, and making our task easier.
Don't forget our goal, we want to pick the important regions of the genome to find our gene more quickly.

How can we find new mutations?
1- FISH
2- aCGH: can detect small deletions.
3- sequencing of affected and unaffected individuals:
the disease-causing gene should be only mutated in affected individuals.

Read page 170 for further details.

Test for gene expression:
1- Northern Blotting
2- RNA sequencing
3- Microarrays
4- Transfection

Candidate genes:
Just understand the concept. Don't memorize anything.

> "It's all about a little bit of knowledge, and a lot of common sense."
> Dr.Mazin Salihi

If we have a disease and we want to search for the disease-causing gene, we will look at the protein product of a few genes that may be related to the disease of interest. For example, in Marfan syndrome, all symptoms are caused by weakness of the connective tissue. So, what are the proteins you may think of?

You will definitely think of the genes, whose protein products are components of the connective tissue (collagen, elastin, fibrillin … etc).

- All these genes that we may think of as the disease-causing genes are called "Candidate Genes".

- But how can we know what's the real disease-causing gene?
We have to examine linkage analysis. In our example, linkage analyses showed that there's no linkage between collagen and collagen markers, while showing that the disease-causing gene is present on chromosome 15. Knowing that FBN-1 gene is also present on chromosome 15 makes it a stronger candidate than collagen.

Shortly, linkage analysis showed that FBN-1 gene is a stronger candidate than collagen gene (i.e. it's more likely to be the disease-causing gene).

But, is that enough to say that "FBN-1 gene is the disease-causing mutation?" Definitely, No.

We have to study the FBN-1 gene mutations to confirm that it's the disease-causing gene.

> Candidate genes are those whose characteristics (e.g., protein product) suggest that they may be responsible for a genetic disease. The analysis of candidate genes in a region known to contain the disease-causing gene is termed the *positional candidate approach.*

Now what if common sense failed , you don't have candidate genes and there is nothing that really is pointing you to an area or another ?!?!

GENE MAPPING BY ASSOCIATION: GENOME-WIDE ASSOCIATION STUDIES

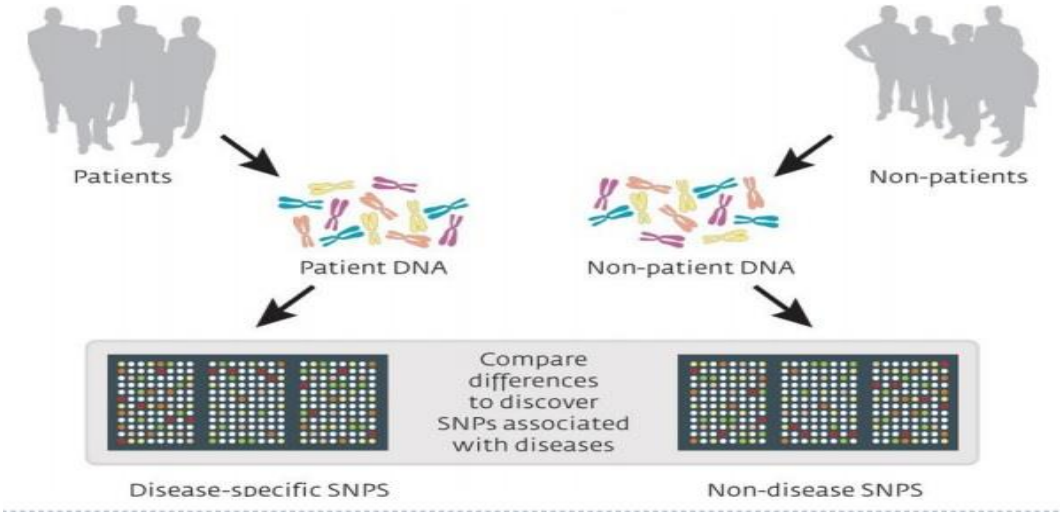
- If all of the previously explained ways of disease-genome identification failed, we will try GWAS.

  - It's typically done by examining certain SNPs in affected cases and a large sample of unaffected cases. Then, we compare the frequency of SNPs in both samples.

- SNPs that are present in a much higher frequency in the diseased sample indicate that the disease-causing genes may be present close to these SNPs.

Manhattan plot : The plot represents the significance of the association between a SNP or haplotype and the trait being measured. [the tool to read a GWAS].

The plot mentioned in the slide illustrates genetic heterogeneity (multiple mutations in different genes).



Please, read page 172.