Cloning Disease Genes in the Human Genome
Arti Pandya, M.D., M.B.A.

REQUIRED READING


STUDENT OBJECTIVES

1. Define the term EST.
2. Understand the steps involved in cloning a disease gene of known function.
3. Understand the steps involved in cloning a disease gene of unknown function.
4. Interpret the results of association analysis for complex disorders.
5. Understand the utility of the Human Genome Project in identifying human disease genes.

I. Introduction

A. Many disorders exist which can be diagnosed clinically, but its gene is not identified or the pathogenesis of the disease is not understood. Cloning genes that cause specific disorders should lead to better diagnosis, the ability to understand the disease process and ultimately to better treatments.

B. Genes for disorders of known function, e.g. metabolic disorders, can be cloned by a biochemical approach. Genes for disorders where the function is not known can be cloned if their chromosomal position is known (positional cloning).

C. The latter has been greatly facilitated by the Human Genome Project.

II. Human Genome Project

A. The Human Genome Project represented an international effort to sequence the entire human genome by the year 2003. It was completed in April 2003, two years ahead of schedule.

B. The first step of this project was to construct a map of the genome using polymorphic markers (genetic mapping). Thousands of markers have been mapped and the project was finished in 1994 a year ahead of schedule. These markers help in identifying location of disease genes.

C. Smaller sequencing projects were also started as a test to determine the feasibility of such a large sequencing project. The genome sequence of many different species of bacteria, yeast, the worm C. elegans and the fly Drosophila melanogaster are finished.
D. One of the priorities of the **Human Genome Project** is to map as many genes as possible on to the human genome. They are mapped using a different type of mapping approach compared to the method used (**genetic mapping**) to map the **polymorphic markers**. Portions of cDNAs i.e. genes are sequenced. The sequence is used in mapping the gene. These mapped products are called **ESTs** (**expressed sequence tag sites**). ESTs are not polymorphic but become candidates of possible disease genes once the location of the disease has been identified.


III. **Cloning a Disease Gene of Known Function**

A. In general, the first step in cloning a gene of known function is to isolate the protein. Protein isolation requires an assay for the protein, such as an enzyme activity assay where a product can be measured. In essence one needs to know the function of the protein in order to purify it.

1. Once the protein is purified, small pieces of the protein are sequenced, i.e. the amino acid order is determined. Protein sequencing is a difficult technique and it will only be possible to get partial protein sequence. Based on the protein sequence, a synthetic oligonucleotide DNA probe is generated using the genetic code. This probe is used to search a cDNA library, i.e. isolate a molecular clone containing the gene for the protein.

2. Examples of genes of known function that have been cloned (associated disease): phenylalanine hydroxylase (phenylketonuria), hexosaminidase A (Tay-Sachs disease) and biotinidase (biotinidase deficiency).

IV. **Cloning a Disease Gene of Unknown Function, also called “Reverse Genetics” or “Positional Cloning”**

![Figure 1](image-url)
In Positional Cloning one identifies the disease gene based on DNA sequence differences between affected and unaffected individuals, and NOT based on information on its function (which is not known)

A. Overview

1. To find a disease causative gene the goal is to locate and identify sequence differences, which may be a single bp change between an affected versus an unaffected individual. Or 1 base pair in 3 billion base pairs (bp) of DNA (the size of the human genome).

2. There are 4 steps to cloning a gene based on its position.

   a. Step 1 – Narrow the location of the disease to a specific chromosomal region using linkage analysis, with polymorphic genetic markers of a. Additional markers in the region are typed at this time to refine the location of the disease gene. This is called identifying the gene critical region. This step can narrow the gene location to 1-3 Mb.

   b. Step 2 – Usually a gene critical region has several genes located in this region, all of which need to be identified, from which one needs to narrow down a single disease causing gene.

   c. Step 3 - Evaluate and prioritize the group of genes in a given region to determine which is the best candidate that might cause the disease.

   d. Step 4 - Search for disease causing mutations in candidate genes.

B. Identification of the Chromosomal Location of a Disease

1. Some diseases are associated with chromosomal aberrations (deletions, translocations, inversions) which give researchers a clue to the location of a gene. If that is the case then markers at a specific chromosomal region should first be used for linkage analysis.

2. If no clues of the location of a gene exist then a screen for markers that cover the whole genome should be initiated. For example you type the families using polymorphic markers that are 10 cM apart and use that data to perform linkage analysis (determine the LOD scores).

3. Example of a linkage analysis of a disease (D) and the marker S300

<table>
<thead>
<tr>
<th></th>
<th>0.0</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D with S300</td>
<td>-∞ (minus infinity)</td>
<td>4.95</td>
<td>2.15</td>
<td>1.54</td>
<td>0.91</td>
<td>0.35</td>
<td>-0.27</td>
</tr>
</tbody>
</table>

Table 1: LOD scores

   a. The disease (D) is linked to the marker S300 at a recombination distance of \( \theta \) (theta) of 0.01 or 1% recombination.

   b. There has been a recombination event between the marker and the disease because at 0% recombination the LOD score is - \( \infty \) (minus infinity).
c. The estimated physical distance of the marker and the disease is about 1,000 kb or 1 Mb (See syllabus for Genetic Mapping).

4. Once linkage is established for a disorder, prenatal diagnosis and/or presymptomatic diagnosis can be offered to affected families using linkage analysis. This involves the participation of many family members.

5. Once the initial chromosomal location of the disease is identified, other markers very close to the first linked marker are typed to establish the smallest region where the disease gene may reside. This is defined as the gene critical region
   a. The gene critical region is typically 1-3 Mb in size
   b. The gene critical region must have defined borders.

6. To define the gene critical region the two markers that are closest to the disease gene yet recombine with it are identified. The DNA between these two markers is the gene critical region.
   a. Using the data from the typed markers you can determine if any of the markers have recombined with the disease
   b. Some markers do not recombine with the disease gene. This shows that the marker is either very close to or within the disease gene.
   c. The polymorphic markers that recombine with the disease suggest that your marker can not be within the disease gene.

7. Example of identifying the gene critical region. As part as the human genome project, the chromosomal location of many polymorphic markers have been identified. The genetic map near the marker S300 is known and shown below. S300 is the marker originally linked to the disease you are studying. You type your families with the other four makers (S90, S27, S32 and S15) the positions of which in relation to S300 are indicated in the Figure 2.

<table>
<thead>
<tr>
<th></th>
<th>0.4cM</th>
<th>0.5cM</th>
<th>1cM</th>
<th>5cM</th>
</tr>
</thead>
<tbody>
<tr>
<td>S90</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2**
Table 2: LOD scores at several recombination distances

<table>
<thead>
<tr>
<th></th>
<th>0.0</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D with S90</td>
<td>$-\infty$</td>
<td>4.95</td>
<td>3.18</td>
<td>2.35</td>
<td>1.19</td>
<td>0.35</td>
<td>-0.27</td>
</tr>
<tr>
<td>D with S27</td>
<td>5.41</td>
<td>3.22</td>
<td>2.15</td>
<td>1.11</td>
<td>0.82</td>
<td>-0.10</td>
<td>-0.35</td>
</tr>
<tr>
<td>D with S32</td>
<td>$-\infty$</td>
<td>3.12</td>
<td>2.78</td>
<td>1.32</td>
<td>0.97</td>
<td>0.25</td>
<td>-1.72</td>
</tr>
<tr>
<td>D with S15</td>
<td>$-\infty$</td>
<td>1.15</td>
<td>2.95</td>
<td>1.89</td>
<td>1.34</td>
<td>1.16</td>
<td>0.91</td>
</tr>
</tbody>
</table>

a. All the markers are linked to the disease but S90, S32 and S15 also recombine with the disease but the marker S27 does not. You already know that the marker S300 recombines with your disease though it is linked.
b. You see that there has not been any recombination between S27 and the disease gene so the disease gene must be near S27.
c. The two markers that flank S27 are S90 and S32. You know that the gene causing the disease cannot be at the same location as S90 and S32 but you do not know where the recombination events occurred between the disease and the markers. These markers become the flanking markers of your gene critical region.
d. In this example, the gene critical region is between the markers S90 and S32.

e. The genetic distance between S90 and S32 is 0.9cM which corresponds to a physical distance of 0.9 Mb or 900 kb.

C. Identify the candidate genes in the gene critical region

1. Due to the Human Genome Project the DNA sequence between the flanking markers is now known.

   a. Genes can be identified by using gene prediction computer programs. These programs look for splice sites and open reading frames to distinguish possible genes from introns and repetitive sequence.
b. Known genes and ESTS can be mapped to the region by matching the cDNA sequence to the genomic sequence in the region

2. In a 900 kb stretch of DNA there may be as many as 24 genes in the region and only one of them when mutated causes the disease you are studying.
3. All genes found in the gene critical region are considered candidates to cause the disease just due to their chromosomal location.

D. Evaluation of candidate genes

1. Once you identify all the genes in the gene critical region, you need to evaluate and prioritize them to determine which ones are the most promising candidate genes.
2. One method is to identify sequence homology to other known genes using a homology search program called BLAST. This program searches GenBank which is a depository of all nucleic acid and protein sequence data compiled world-wide.
   a. If a gene in your gene critical region has homology to a gene whose function is known then you may be able to infer the function of the gene that you found.
   b. If the function can explain the disease phenotype this gene becomes a good candidate gene.
   c. For example, if the disease is known to affect growth and a gene in the gene critical region encodes a protein that is homologous to a group of proteins known as growth factors then this gene becomes a good candidate gene.
3. Identify expression pattern of a gene through northern hybridization.
   a. The disease gene should be expressed in the tissue that the disease affects. For a neurological disease the gene should be expressed in the brain. Of 20 genes you identified in your gene critical region only four are expressed in the brain. Those genes become better candidates then the remaining 16 and will move up the priority list.
E. Identification of mutations

1. There are many methods to identify mutations in a gene and one method is to sequence the DNA.

   a. In this method you need to sequence the DNA of your candidate gene from an affected person and compare it to the sequence from an unaffected individual.

      i. The easiest method is to use PCR to amplify the gene of interest from the DNA of affected and unaffected individuals.

      ii. The PCR products can be directly sequenced.

   b. If there is a difference in the sequence between affected and unaffected individuals this change may be the causative mutation. If the change creates a nonsense or missense change in the protein this provides stronger but not conclusive evidence that this is the mutation which causes the disease.

   c. To be sure the change you identified is the causative mutation, a large number (at least 100) of unaffected, unrelated individuals must be screened for that change. If the change is found in unaffected individuals it is likely to be a polymorphism and not a mutation.
2. Once a mutation in a gene is identified, direct diagnosis, pre-implantation, prenatal and presymptomatic testing can be offered without the need to involve other family members.

F. Progress of cloning genes

1. Hundreds of disease genes have been identified by positional cloning.

V. Identification of genes from Genome Wide Association Studies

1. It is often possible to determine sequence variants in a gene likely to increase the risk for a complex disorder.
2. More often, these variants are neither necessary, not sufficient to cause the particular disorder, and other genetic or environmental factors are necessary to develop the disorder.
3. The steps in gene identification for complex disorders is similar to that for genes for single gene disorders, but the gene critical region is identified based on association studies and the ODDS ratios or Linkage Disequilibrium to a SNP or set of SNP’s.
4. Implication of a gene variant as a susceptibility locus is dependent on the ability to replicate the results in independent population studies AND demonstrate a functional role for the polymorphism identified.
5. Examples of Complex disease genes which have been identified include the NOD2 gene for Crohn’s disease and the Complement Factor H (CFH) for age related macular degeneration (refer to text, page 227-229)

VI. Things to Think About

A. Cloning a gene is only the first step in finding therapies for disorders. Once a gene has been identified and cloned, if the gene does not have a known function then its function has to be determined before any clinical benefits besides diagnostic testing can be seen. This will likely take longer than the actual cloning of the gene.

1. The first disease gene to be cloned was the Duchenne muscular dystrophy gene in 1987. A fairly good idea about what the gene does is known but still no therapy for this disorder is available. No therapies from positionally cloned genes that are not biochemical in nature have been forthcoming.

B. What are the ethical ramifications of genetic testing of disorders for which no therapies are available?
C. Scientist are now using similar approaches to try to identify genes important in common disorders such as heart disease, arthritis and asthma. These are more complex problems but potentially can benefit a much greater number of people.

D. In the future, a physician may be able to order a series of tests to identify all changes in an individual’s genome that will affect their health. This will give physicians the ability to predict which diseases an individual may acquire during their lifetime.

1. Possible positive consequences
   a. Early and more stringent screening can be performed in individuals with higher risk for developing a specific disorder.
   b. The physician may be able to recommend preventative measures to avert disease occurrence.
   c. Pharmacogenetics or personalized medication.

2. Possible negative consequences.
   a. In complex disorders, it is difficult to determine which individuals may actually become symptomatic.
   b. An individual may react negatively if they know they have a high chance of developing a disorder.
   c. Privacy of such genetic information must be guarded.
   d. An individual may lose their health insurance if they are deemed to have a “pre-existing” condition.
   e. Such information could be used for eugenic practices.

VI. Resources

Gene Tests/Gene Clinics at www.geneclinics.org