
National Center for Biotechnology Information



# Just the Facts: A Basic Introduction to the Science Underlying NCBI Resources

## ESTs: GENE DISCOVERY MADE EASIER

Investigators are working diligently to sequence and assemble the genomes of various organisms, including the mouse and human, for a number of important reasons. Although important goals of any sequencing project may be to obtain a genomic sequence and identify a complete set of genes, the ultimate goal is to gain an understanding of when, where, and how a gene is turned on, a process commonly referred to as **gene expression**. Once we begin to understand where and how a gene is expressed under normal circumstances, we can then study what happens in an altered state, such as in disease. To accomplish the latter goal, however, researchers must identify and study the protein, or proteins, coded for by a gene.

As one can imagine, finding a gene that codes for a protein, or proteins, is not easy. Traditionally, scientists would start their search by defining a biological problem and developing a strategy for researching the problem. Oftentimes, a search of the scientific literature provided various clues about how to proceed. For example, other laboratories may have published data that established a link between a particular protein and a disease of interest. Researchers would then work to isolate that protein, determine its function, and locate the gene that coded for the protein. Alternatively, scientists could conduct what is referred to as **linkage studies** to determine the chromosomal location of a particular gene. Once the chromosomal location was determined, scientists would use biochemical methods to isolate the gene and its corresponding protein. Either way, these methods took a great deal of time—years in some cases—and yielded the location and description of only a small percentage of the genes found in the human genome.

Now, however, the time required to locate and fully describe a gene is rapidly decreasing, thanks to the development of, and access to, a technology used to generate what are called **Expressed Sequence Tags**, or **ESTs**. ESTs provide researchers with a quick and inexpensive route for discovering new genes, for obtaining data on gene expression and regulation, and for constructing genome maps. Today, researchers using ESTs to study the human genome find themselves riding the crest of a wave of scientific discovery the likes of which has never been seen before.

An Expressed Sequence Tag is a tiny portion of an entire gene that can be used to help identify unknown genes and to map their positions within a genome.

## What Are ESTs and How Are They Made?

**ESTs** are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene. The idea is to sequence bits of DNA that represent genes expressed in certain cells, tissues, or organs from different organisms and use these "**tags**" to fish a gene out of a portion of chromosomal DNA by matching base pairs. The challenge associated with identifying genes from genomic sequences varies among organisms and is dependent upon genome size as well as the presence or absence of **introns**, the intervening DNA sequences interrupting the protein coding sequence of a gene.

### Separating the Wheat from the Chaff: Using mRNA to Generate cDNA

Gene identification is very difficult in humans, because most of our genome is composed of introns interspersed with a relative few DNA coding sequences, or genes. These genes are expressed as proteins, a complex process composed of two main two steps. Each gene (DNA) must be converted, or **transcribed**, into **messenger RNA** (mRNA), RNA that serves as a template for protein synthesis. The resulting mRNA then guides the synthesis of a protein through a process called **translation**. Interestingly, mRNAs in a cell do not contain sequences from the regions between genes, nor from the non-coding introns that are present within many genes. Therefore, isolating mRNA is key to finding expressed genes in the vast expanse of the human genome.

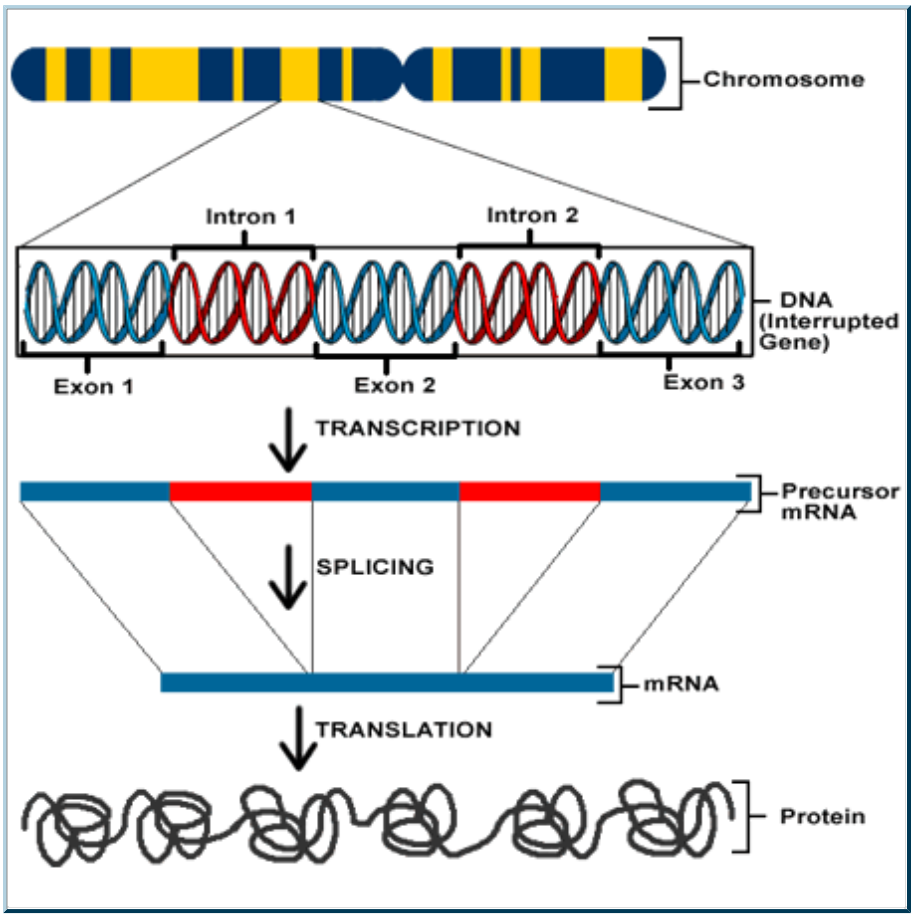


**Figure 1.** An overview of the process of protein synthesis.

*Protein synthesis is the process whereby DNA codes for the production of amino acids and proteins. The process is divided into two parts: transcription and translation. During transcription, one strand of a DNA double helix is used as a template by mRNA polymerase to synthesize a mRNA. During this step, mRNA passes through various phases, including one called splicing, where the non-coding sequences are eliminated. In the next step, translation, the mRNA guides the synthesis of the protein by adding amino acids, one by one, as dictated by the DNA and represented by the mRNA.*

The problem, however, is that mRNA is very unstable outside of a cell; therefore, scientists use special enzymes to convert it to **complementary DNA (cDNA)**. cDNA is a much more stable compound and, importantly, because it was generated from a mRNA in which the introns have been removed, cDNA represents only expressed DNA sequence.

cDNA is a form of DNA prepared in the laboratory using an enzyme called reverse transcriptase. cDNA production is the reverse of the usual process of transcription in cells

because the procedure uses mRNA as a template rather than DNA. Unlike genomic DNA, cDNA contains only expressed DNA sequences, or exons.

## From cDNAs to ESTs

A "**gene family**" is a group of closely related genes that produces similar protein products.

Once cDNA representing an expressed gene has been isolated, scientists can then sequence a few hundred nucleotides from either end of the molecule to create two different kinds of ESTs. Sequencing only the beginning portion of the cDNA produces what is called a **5' EST**. A 5' EST is obtained from the portion of a transcript that usually codes for a protein. These regions tend to be conserved across species and do not change much within a **gene family**. Sequencing the ending portion of the cDNA molecule produces what is called a **3' EST**. Because these ESTs are generated from the 3' end of a transcript, they are likely to fall within non-coding, or **untranslated regions (UTRs)**, and therefore tend to exhibit less cross-species conservation than do coding sequences.

A UTR is that part of a gene that is not translated into protein.

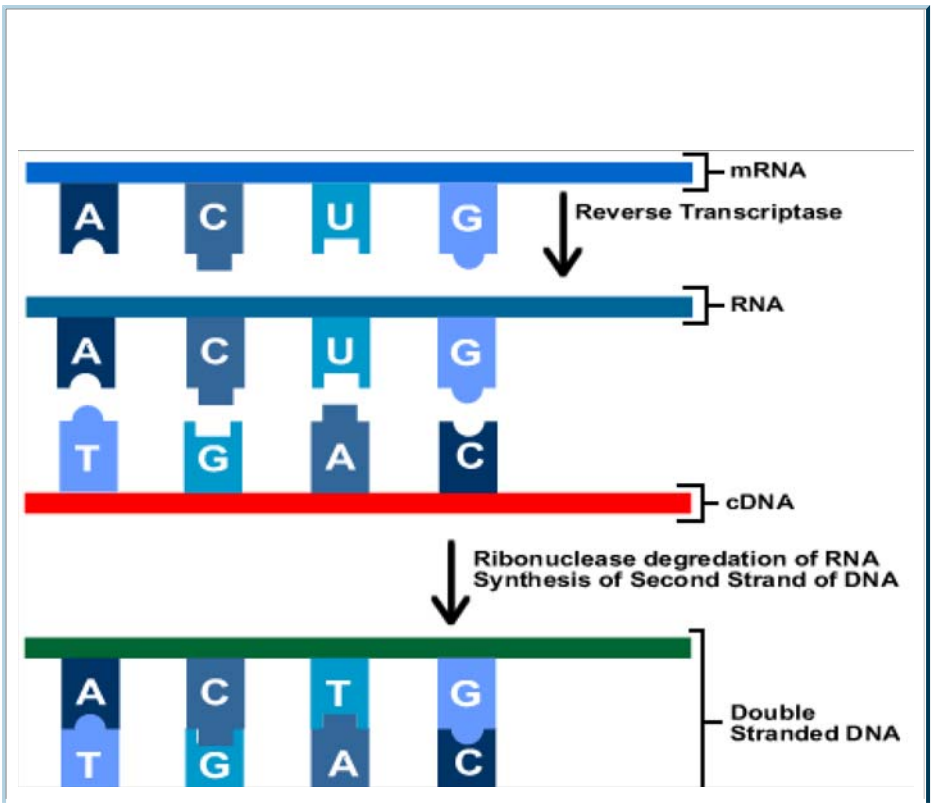


**Figure 2.** An overview of how ESTs are generated.

*ESTs are generated by sequencing cDNA, which itself is synthesized from the mRNA molecules in a cell. The mRNAs in a cell are copies of the genes that are being expressed. mRNA does not contain sequences from the regions between genes, nor from the non-coding introns that are present within many interesting parts of the genome.*

# ESTs: Tools for Gene Mapping and Discovery

## ESTs as Genome Landmarks

Just as a person driving a car may need a map to find a destination, scientists searching for genes also need **genome maps** to help them to navigate through the billions of nucleotides that make up the human genome. For a map to make navigational sense, it must include reliable landmarks or "markers". Currently, the most powerful mapping technique, and one that has been used to generate many genome maps, relies on Sequence Tagged Site (STS) mapping. An STS is a short DNA sequence that is easily recognizable and occurs only once in a genome (or chromosome). The 3' ESTs serve as a common source of STSs because of their likelihood of being unique to a particular species and provide the additional feature of pointing directly to an expressed gene.

## ESTs as Gene Discovery Resources

ESTs are powerful tools in the hunt for known genes because they greatly reduce the time required to locate a gene.

Because ESTs represent a copy of just the interesting part of a genome, that which is expressed, they have proven themselves again and again as powerful tools in the hunt for genes involved in hereditary diseases. ESTs also have a number of practical advantages in that their sequences can be generated rapidly and inexpensively, only one sequencing experiment is needed per each cDNA generated, and they do not have to be checked for sequencing errors because mistakes do not prevent identification of the gene from which the EST was derived.

Using ESTs, scientists have rapidly isolated some of the genes involved in Alzheimer's disease and colon cancer.

To find a disease gene using this approach, scientists first use observable biological clues to identify ESTs that may correspond to disease gene candidates. Scientists then examine the DNA of disease patients for mutations in one or more of these candidate genes to confirm gene identity. Using this method, scientists have already isolated genes involved in Alzheimer's disease, colon cancer, and many other diseases. It is easy to see why ESTs will pave the way to new horizons in genetic research.

## ESTs and NCBI

For ESTs to be easily accessed and useful as gene discovery tools, they must be organized in a searchable database that also provides access to genome data.

Because of their utility, speed with which they may be generated, and the low cost associated with this technology, many individual scientists as well as large genome sequencing centers have been generating hundreds of thousands of ESTs for public use. Once an EST was generated, scientists were submitting their tags to **GenBank**, the NIH sequence database operated by NCBI. With the rapid submission of so many ESTs, it became difficult to identify a sequence that had already been deposited in the database. It was becoming increasingly apparent to NCBI investigators that if ESTs were to be easily accessed and useful as gene discovery tools, they needed to be organized in a searchable database that also provided access to other genome data. Therefore, in 1992, scientists at NCBI developed a new database designed to serve as a collection point for ESTs. Once an EST that was submitted to GenBank had been screened and annotated, it was then deposited in this new database, called **dbEST**.

## dbEST: A Descriptive Catalog of ESTs

Scientists at NCBI annotate EST records with text information regarding DNA and mRNA homologies.

Scientists at NCBI created dbEST to organize, store, and provide access to the great mass of public EST data that has already accumulated and that continues to grow daily. Using dbEST, a scientist can access not only data on human ESTs but information on ESTs from over 300 other organisms as well. Whenever possible, NCBI scientists annotate the EST record with any known information. For example, if an EST matches a DNA sequence that codes for a known gene with a known function, that gene's name and function are placed on the EST record. Annotating EST records allows public scientists to use dbEST as an avenue for gene discovery. By using a database search tool, such as NCBI's BLAST, any interested party can conduct sequence similarity searches against dbEST.

## UniGene: A Non-Redundant Set of Gene-oriented Clusters

Because a gene can be expressed as mRNA many, many times, ESTs ultimately derived from this mRNA may be **redundant**. That is, there may be many identical, or similar, copies of the same EST. Such redundancy and overlap means that when someone searches dbEST for a particular EST, they may retrieve a long list of tags, many of which may represent the same gene. Searching through all of these identical ESTs can be very time consuming. To resolve the redundancy and overlap problem, NCBI investigators developed the **UniGene database** UniGene automatically partitions GenBank sequences into a non-redundant set of gene-oriented clusters.

Although it is widely recognized that the generation of ESTs constitutes an efficient strategy to identify genes, it is important to acknowledge that despite its advantages, there are several limitations associated with the EST approach. One is that it is very difficult to isolate mRNA from some tissues and cell types. This results in a paucity of data on certain genes that may only be found in these tissues or cell types.

Second is that important gene regulatory sequences may be found within an intron. Because ESTs are small segments of cDNA, generated from a mRNA in which the introns have been removed, much valuable information may be lost by focusing only on cDNA sequencing. Despite these limitations, ESTs continue to be invaluable in characterizing the human genome, as well as the genomes of other organisms. They have enabled the mapping of

many genes to chromosomal sites and have also assisted in the discovery of many new genes.

Back to top

Revised: March 29, 2004