

What You Need to Know

■ DNA provides the set of instructions, or genes, used by cells to carry out daily functions and interact with the environment. The complete set of DNA found in an organism is called its genome.

■ Since 1995, scientists have sequenced and analyzed the genomes of over 200 different organisms, with more being completed every day.

■ Comparing the genomes of multiple organisms identifies both common and divergent genetic regions.

■ At certain portions of the genome, the DNA sequence is highly similar or even identical across different organisms. These regions often contain genes that produce proteins critical to basic cellular functions common among all forms of life, but the functions of some of these conserved sequences serve as yet unknown functions.

■ Other genetic sequences may be shared among only a subset of organisms, suggesting the presence of genes related to characteristics unique to that group.

■ Identifying which mutations are present in the cancer offers physicians clues to both outcome and treatment options.

■ Comparative genomics means that studying the DNA of many different organisms can identify genes involved in human disease. Because they have similar genes, scientists can explore early forms of treatment and/or prevention in these organisms, as well.

If you want to know more:

Functional and Comparative Genomics Fact Sheet - http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml

The Genes We Share
<http://www.hhmi.org/geneshare/>

The Genome News Network
<http://www.genomenewsnetwork.org/>

BIOTECH 101

comparative genomics

Introduction

In 1995, scientists deciphered the complete genetic sequence of *Haemophilus influenzae* - a bacteria which causes the flu. This marked the first time an organism's genome had been sequenced and helped usher in the era of genomics - the study of an organism's entire collection of DNA. Since that time, the genomes of many additional organisms have been sequenced, ranging from single-celled bacteria to plants, fish, birds and mammals.



Although the human genome is perhaps the most famous sequencing project, over the past 14 years scientists have assembled a genomic library of nearly 200 different organisms. Knowing the genome of each species provides insight into the function of its DNA; however, there is additional information gained by comparing genomes across organisms. The field of comparative genomics helps discover previously un-

detected genes, identify the regulatory regions that control gene activity and determine gene function as it relates to health and disease.

Comparing Organisms

At first glance, humans may seem to have little in common with fruit flies, roundworms and mice. Upon closer inspection, each is composed of cells that must take in nutrients and remove waste, interact with neighboring cells and the outside environment, and grow and divide in response to specific signals. To varying degrees, each of these organisms contains digestive, circulatory, nervous and reproductive systems and is impacted by disorders that impair these systems. There are similarities in structures and the functioning of those structures across diverse members of the tree of life. Researchers value organisms like fruit flies, worms and mice because they can be used as simpler versions of more complex creatures, and are relatively easy to study in the laboratory as models for understanding health and illness.



With sequenced genomes in hand, scientists are able to directly compare the DNA sequences of these organisms. Essentially, comparative genomics involves the use of sophisticated computer programs that line up multiple genome sequences and look for regions of similarity in long strings of As, Cs, Gs and Ts. These similar segments suggest the corresponding DNA sequence has an important functional role - for example, a gene or a regulatory element that controls the activity of a gene.

Comparative studies have found that a number of similarities exist in the gene set of multicellular organisms. Although the organisms differ widely in the size and number of chromosomes containing their genomes (see table at right), a common ground is present with respect to an overlapping set of genes. In fact, one study suggests that 60 percent of the genes present in the fruit fly have a human counterpart - known as a homolog. For example, a well known set of genes organizes the fruit fly's body into head, chest, abdomen and limbs. These are called homeotic genes, with

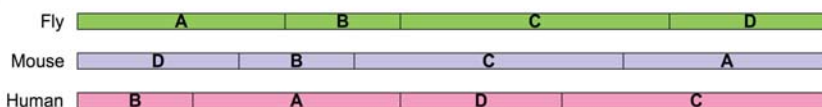


homeotic meaning that something has been changed into the likeness of something else: One mutation in a homeotic gene called *antp* can change one of the fly's antennae into a leg. Nearly identical copies of these genes are present in humans and are responsible for determining the structure of major body parts.

Genes affecting more advanced features, such as an immune system, are less likely to have direct counterparts in flies and worms. For these systems, organisms that have more structural and functional similarities to humans, such as other mammals, are examined. Among mammals, there is a striking degree of DNA sequence similarity, particularly with the sequence of genes. Nearly all mammals have similarly sized genomes and the likelihood of finding a homolog corresponding to a human gene approaches 100 percent. While the order and location of the genes on chromosomes may vary, large blocks of sequence identity are usually present (figure one). These similarities can be used in a variety of ways. For example, the genomic map of one species can be used to help find genes in a related organism that is in the process of being sequenced.

Comparative genomics can also be used to identify regulatory regions that control gene activity. While DNA sequence is highly similar within genes, lower levels of sequence similarity usually are observed in the segments between genes. If a comparison of multiple organisms

a) Sequenced DNA fragments



b) Comparing similar regions



c) Searching for high similarity

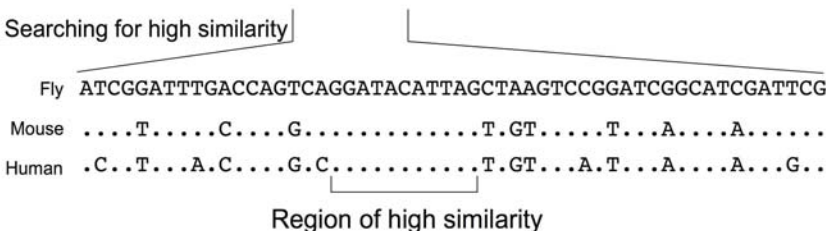


Figure: sequence homology among organisms

a) DNA fragments are sequenced across a number of different organism and b) regions of sequence similarity are identified and compared. c) Sequence alignments – mouse and human DNA sequence identical to the fly is noted with “.” and different sequence is shown as the corresponding DNA nucleotide (A, T, C or G). Within these alignments, areas of highest similarity are flagged as identifying possible genes or regulatory regions that require additional analysis. Figure modified from Marguiles and Birney Nature Reviews Genetics 9:304 (2008)

identifies high similarity within a non-gene sequence, it may signal the presence of an important regulatory element used to activate or silence nearby genes. A hot question in recent genomic studies has centered on the presence of numerous conserved regions in many species that don't fit any definition we currently have for a gene or a regulatory sequence.

What are these sequences doing, and why have they stayed the same over such long periods of time?

Linking to Disease

Genomic comparison also extends to genes involved in disease. If we examine the current list of human disease genes, approximately 20 percent have a homolog in yeast and nearly two-thirds have one in flies and worms. Initial studies suggest these counterparts may function in nearly identical ways, meaning these organisms can serve as models for understanding human disease and potential treatment. Recently, researchers inserted into fruit flies an altered version of a human gene associated with early-onset Parkinson disease. The flies containing the altered human gene showed symptoms similar to patients affected with Parkinson.

Although, a nearly one-to-one gene correlation exists between most mammals, species-specific genetic differences do exist. Comparative genomics can help

Table: Comparative Genome Sizes and Gene Number for Various Organisms

Modified from the US Department of Energy Office of Science website: http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml

Organism	Estimated size of genome	Estimated gene number	Average gene density	Chromosome number
<i>Homo sapiens</i> (human)	3.2 billion	approx. 25,000	1 gene per 100,000 bases	46
<i>Mus musculus</i> (mouse)	2.6 billion	approx. 25,000	1 gene per 100,000 bases	40
<i>Glycine max</i> (soybean)	1 billion	approx. 66,000	1 gene per 15,000 bases	40
<i>Drosophila melanogaster</i> (fruit fly)	137 million	13,000	1 gene per 9,000 bases	8
<i>Caenorhabditis elegans</i> (roundworm)	97 million	19,000	1 gene per 5,000 bases	12
<i>Saccharomyces cerevisiae</i> (yeast)	12.1 million	6,000	1 gene per 2,000 bases	32
<i>Escherichia coli</i> (bacteria)	4.6 million	3,200	1 gene per 1,400 bases	1
<i>Haemophilus influenzae</i> (bacteria)	1.8 million	1,700	1 gene per 1,000 bases	1

: CONTINUED ON PAGE 6

BIOTECH 101 CONTINUED

identify the genetic spellings that make each species distinct. This is a topic with broad general interest (“What is it that makes us uniquely human?”), but one that also carries important medical applications. For example, certain mammals do not frequently develop cancer. Chimpanzees are not affected by malaria and AIDS. A comparison of the genes involved in these diseases may identify the genetic changes that prevent the disorder from occurring, leading to new pathways for prevention in humans.

Comparative genomics is still a relatively young field of research. As scientists sequence the genomes of other organisms, comparative studies will continue to uncover common threads and unique genetic sequences. By identifying the distinctive genetic changes that alter disease frequency and susceptibility, we can better understand gene function and the connection to disease. This identification will also help us increase the speed of potential treatment and pharmaceutical interventions. ■