Genomics for Energy and Environmental Science

What Is Genomics?

A **genome** is the complete set of genetic material encoding the biological instructions for developing and maintaining an organism. The genome for any free-living system—ranging from a single bacterial cell to a complex multicellular organism such as a human or a tree—is made up of DNA (deoxyribonucleic acid). **Genomics** is a scientific field focused on sequencing and analyzing genomes. Insights from genomic information are catalyzing an extraordinary transformation in our understanding of biological systems and how we can use their capabilities to address challenges in the environment, energy, industry, biomedicine, and other application areas.

DNA: Life's Molecular Archive of Genetic Information

Nucleotides are chemical subunits strung together to form DNA's double-stranded, twisted ladder structure. DNA is made up of four kinds of nucleotides, each containing a different nitrogenous base: adenine (A), cytosine (C), guanine (G), and thymine (T). Each base on one DNA strand pairs with a base on the other strand—A's pair with T's, and C's with G's. These base pairs form the "rungs" of the DNA ladder and hold the two DNA strands together. The long, parallel "side rails" of the ladder, forming the backbone of each DNA strand, consist of sugar and phosphate molecules. DNA molecules can be hundreds, thousands, or millions of base pairs in length.

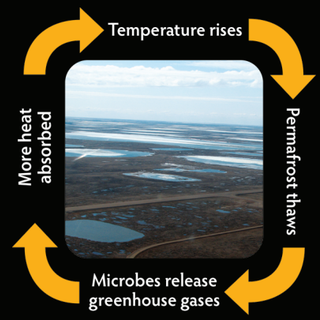
Metagenomics: Sequencing and Analyzing Microbial Communities in Diverse Environments

Microbes and their communities make up the foundation of the biosphere and sustain all life on Earth. Although largely unexplored, the dimensions of the microbial world are immense, comprising more than half the planet's biomass. The number of bacterial cells on Earth is estimated to be a billion times greater than the number of stars in the universe. The unique biochemistries that enable microbes to thrive in every niche on the planet represent a deep and virtually limitless wealth of capabilities that can meet diverse national needs.

Metagenomics is the sequencing and analysis of DNA extracted directly from microbial communities in the environment. Metagenomic studies are enabling the discovery of millions of previously unknown genes and proteins, thousands of species, and innumerable variations in critical biochemical capabilities. Pictured at right are some metagenome projects that are providing important insights to address challenges related to energy production and environmental science.

Termite gut. More than 200 species of microbes make up the community residing in the termite hindgut. Together they produce a bounty of wood-degrading enzymes that could be used by industry to make biofuels from woodchips and other forms of fibrous cellulosic biomass.

Thermal pools. To identify new microbes and enzymes resistant to the heat and stresses of industrial processing, researchers are investigating microbial communities living at near-boiling temperatures in pools at Yellowstone National Park.



Permafrost soils. By analyzing the DNA and proteins of microbial communities in Alaskan permafrost (permanently frozen soil), scientists can gain new insights into the biological mechanisms controlling some of the world's greatest reservoirs of terrestrial carbon.

Expressing the Genome: DNA ▶ RNA ▶ Protein

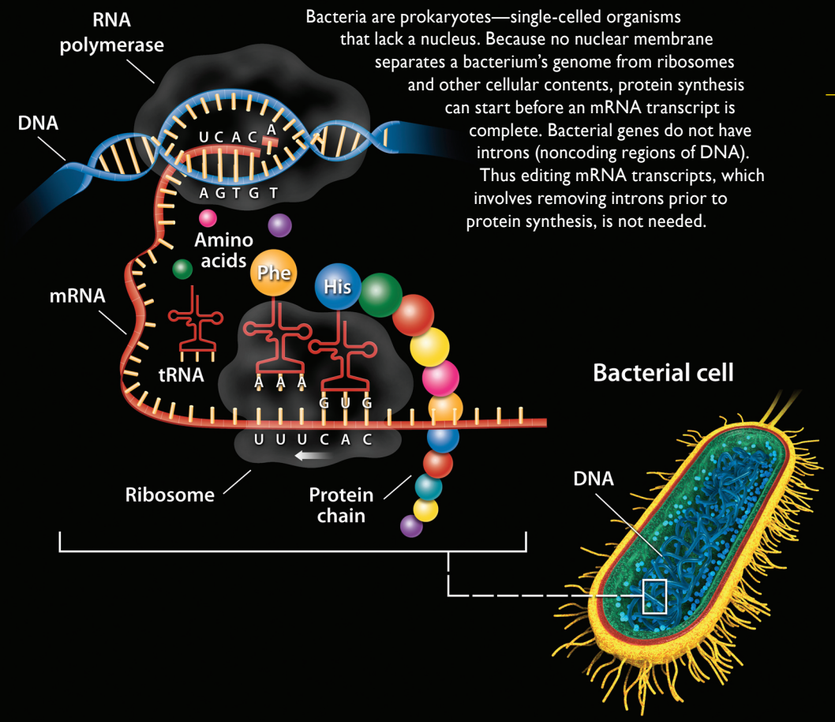
Each DNA molecule contains many **genes**—the basic physical and functional units of heredity. A gene is a segment of DNA that has a specific order of base pairs encoding the chemical subunits of a protein. Although genes get a lot of attention, proteins carry out most of the molecular-level functions of organisms. "Expressing" a gene involves using a gene's DNA sequence to direct the synthesis of a protein.

A gene is turned "on" when it is actively being expressed into protein, and turned "off" when expression is stopped. Producing proteins from genes occurs in two basic steps: (1) transcription and (2) translation.

Transcription is the process of making a temporary RNA copy of a gene's DNA sequence. This RNA copy of a gene is called messenger RNA (mRNA). Like DNA, RNA consists of a long chain of nucleotides, but RNA is typically single stranded and uses a nucleotide called uracil (U) in place of thymine (T). The enzyme RNA polymerase binds and separates the double strands of DNA and then uses one DNA strand as a template for assembling mRNA.

Translation is a process that uses the nucleotide base sequence of mRNA to direct the synthesis of a protein's amino acid sequence. A large molecular complex called a ribosome binds the mRNA to initiate translation. With the help of another RNA called transfer RNA (tRNA), the ribosome adds amino acids one at a time to build a linear protein chain. A set of three mRNA bases (called a codon) specifies a particular amino acid or signals the ribosome to start or stop protein synthesis. Once the linear chain of amino acids is complete, the chain folds into a specific three-dimensional protein structure that performs a particular biological function.

Expressing the Genome in Bacterial Cells



Bacteria are prokaryotes—single-celled organisms that lack a nucleus. Because no nuclear membrane separates a bacterium's genome from ribosomes and other cellular contents, protein synthesis can start before an mRNA transcript is complete. Bacterial genes do not have introns (noncoding regions of DNA). Thus editing mRNA transcripts, which involves removing introns prior to protein synthesis, is not needed.

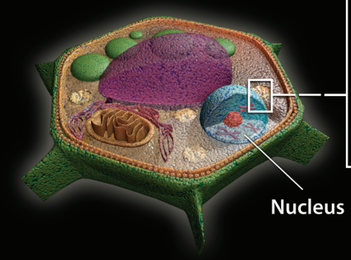
Second Base

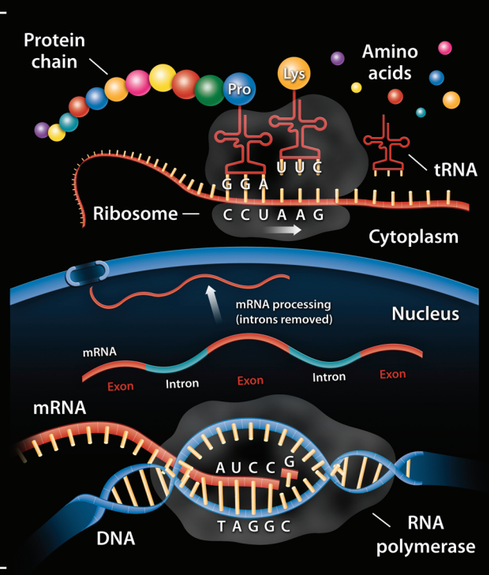
First Base	U	C	A	G	Third Base
U	UUU Phenylalanine (Phe)	UCU Serine (Ser)	UAU Tyrosine (Tyr)	UGU Cysteine (Cys)	U
	UUC Phenylalanine (Phe)	UCC Serine (Ser)	UAC Tyrosine (Tyr)	UGC Cysteine (Cys)	C
	UUA Leucine (Leu)	UCA Serine (Ser)	UAA Stop	UGA Stop	A
	UUG Leucine (Leu)	UCG Serine (Ser)	UAG Stop	UGG Tryptophan (Trp)	G
C	CUU Leucine (Leu)	CCU Proline (Pro)	CAU Histidine (His)	CGU Arginine (Arg)	U
	CUC Leucine (Leu)	CCC Proline (Pro)	CAC Histidine (His)	CGC Arginine (Arg)	C
	CUA Leucine (Leu)	CCA Proline (Pro)	CAA Glutamine (Gln)	CGA Arginine (Arg)	A
	CUG Leucine (Leu)	CCG Proline (Pro)	CAG Glutamine (Gln)	CGG Arginine (Arg)	G
A	AUU Isoleucine (Ile)	ACU Threonine (Thr)	AAU Asparagine (Asn)	AGU Serine (Ser)	U
	AUC Isoleucine (Ile)	ACC Threonine (Thr)	AAC Asparagine (Asn)	AGC Serine (Ser)	C
	AUA Isoleucine (Ile)	ACA Threonine (Thr)	AAA Lysine (Lys)	AGA Arginine (Arg)	A
	AUG Methionine (Met) or Start	ACG Threonine (Thr)	AAG Lysine (Lys)	AGG Arginine (Arg)	G
G	GUU Valine (Val)	GCU Alanine (Ala)	GAU Aspartic acid (Asp)	GGU Glycine (Gly)	U
	GUC Valine (Val)	GCC Alanine (Ala)	GAC Aspartic acid (Asp)	GGC Glycine (Gly)	C
	GUA Valine (Val)	GCA Alanine (Ala)	GAA Glutamic acid (Glu)	GGA Glycine (Gly)	A
	GUG Valine (Val)	GCG Alanine (Ala)	GAG Glutamic acid (Glu)	GGG Glycine (Gly)	G

The Genetic Code. Universal among all life forms, the genetic code is the language used to write the mRNA instructions for building proteins. Each three-letter codon in the mRNA specifies a particular amino acid. Since there are 20 different amino acids and 64 different codons, an amino acid can be represented by more than one codon. The methionine codon (AUG) is the start codon initiating protein synthesis. Three stop codons signify the end of a protein sequence.

Expressing the Genome in Plant Cells

Plants are eukaryotes—organisms with cells that contain a membrane-bound nucleus. A eukaryote's DNA is in the nucleus where mRNA is transcribed. The genes in plants and other eukaryotic organisms, such as humans and animals, contain noncoding regions called introns. In the nucleus, introns are removed from mRNA transcripts, and the remaining coding regions (called exons) are spliced back together. Once edited, the mRNA is transported outside the nucleus for translation into proteins by ribosomes.

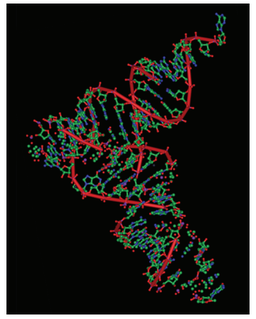




Omics: Exploring the Molecular Universe within Biological Systems

Instead of studying one or a few genes or proteins at a time, "**omics**" collectively describes the comprehensive analysis of genes, RNA transcripts, proteins, metabolites, and other molecules present in a biological system. Ongoing advances in computing power and automated technologies for DNA sequencing and experiments continue to improve our ability to analyze increasing numbers of molecules and how they function as a system. **Systems biology** integrates the data from various omic analyses using computational tools to build predictive models of biological systems.

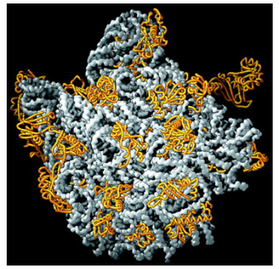
Transcriptomics is the analysis of the transcriptome—the complete set of RNA molecules present in a cell or population of cells. RNA, which is much less stable than DNA, is constantly being synthesized and then broken down to facilitate rapid changes in patterns of protein expression that occur as an organism dynamically responds to its environment. In addition to the three major classes of RNA (mRNA, tRNA, and ribosomal RNA), single-stranded RNA is very flexible and can fold into complex shapes that carry out specific functions. RNA also can be broken into small segments that bind molecules and influence gene expression.



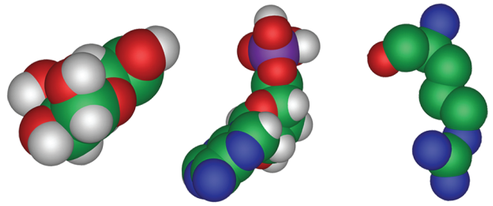
Single-stranded RNA can fold up and twist to form 3-D structures like this tRNA.

Metabolomics is the analysis of the metabolome—the complete set of an organism's metabolites (small molecular products of cellular processes). Characterizing the metabolome provides a snapshot of the metabolic status of an organism and helps determine which of its enzyme-mediated reactions and biochemical pathways are active.

Proteomics is the analysis of the proteome—the complete set of proteins expressed by a cell or population of cells. Proteins, the workhorse molecules of life, catalyze biochemical reactions; provide structural support; and recognize, bind, or transport other molecules throughout the cell. Hundreds of different types of proteins can be expressed at a time, and most are part of large complexes made up of many proteins and other molecules.



Hundreds to thousands of different protein types exist within a cell. This large subunit of a ribosome contains about 3,000 RNA nucleotides (gray) and 30 protein chains (gold).



Small molecular metabolites (e.g., sugars, amino acids, and fatty acids) are the substrates and products of biochemical processes.

Why Is the Department of Energy Involved in Genomics and Systems Biology?

For decades, the U.S. Department of Energy (DOE) Office of Science and its predecessor agencies have supported basic research on the biological effects of energy production and use. Recognizing the need for a complete reference sequence of human DNA, the Office of Biological and Environmental Research (BER) initiated its Human Genome Program in 1986, which led to the international Human Genome Project and launched a new era of genome-enabled biology.

Operating within the Office of Science, BER is focused today on DOE missions to produce sustainable bioenergy alternatives and understand energy-related impacts on the environment, carbon cycle, and climate. BER's Genomic Science program is advancing and using the latest genome-based biotechnologies to explore and harness the diverse biochemical capabilities of microbial and plant systems. An ultimate goal for the Genomic Science program is to develop computational models that provide a predictive understanding of DOE-relevant organisms.

Websites for More Information

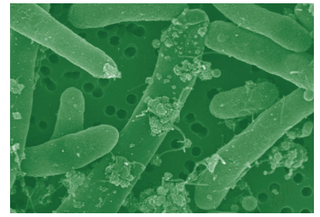
U.S. Department of Energy Office of Science Office of Biological and Environmental Research
science.energy.gov/ber/

DOE Genomic Science Program Websites
science.energy.gov/ber/research/bssd/genomic-science/
genomicscience.energy.gov

Download or request placemat copies and provide comments
genomicscience.energy.gov/posters/

Genomics Enables Systems-Level Studies of Microbes and Plants for Energy and Environmental Solutions

In addition to providing the research community with state-of-the-science capabilities for sequencing and analyzing genomes at the DOE Joint Genome Institute (JGI), the DOE Office of Biological and Environmental Research (BER) supports a broad portfolio of projects and Bioenergy Research Centers that provide the science and technology development needed for a systems-level understanding of DOE-relevant organisms. Some of these projects are described below.

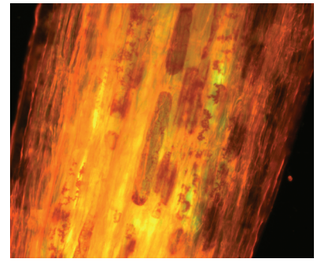


Metabolically engineered bacteria produce biodiesel.

Establishing new biofuel production pathways in microbes. To accelerate development of next-generation biofuels, researchers supported by DOE BER are studying the biochemical pathways encoded in microbial genomes. Using this genomic information, researchers are developing new microbial systems that can produce biodiesel and other chemical replacements for petroleum products.

Genome-wide analysis of small RNAs in grasses. Researchers are just beginning to explore the roles that microRNAs (small RNA molecules consisting of a few dozen nucleotides) play in regulating development, stress responses, and other plant processes. Understanding microRNAs in grasses such as switchgrass and *Miscanthus* is of particular interest to DOE because these plants grow fast, creating large amounts of biomass for biofuel production, and they help increase carbon storage in soils.

Discovery of growth-promoting bacteria in tree roots. The fast-growing poplar tree *Populus trichocarpa* is a model species for bioenergy production because it can grow on marginal lands unsuitable for food crops. Researchers working with DOE JGI have sequenced the genome of a bacterium that can boost tree growth up to 40% when it invades poplar roots. This work has revealed a wide range of genes associated with this symbiotic relationship between plants and bacteria.



Symbiotic relationships between bacteria and root cells promote tree growth.

Predicting microbial response to global change. To examine the impacts of altered water and nitrogen availability on microbial carbon degradation in a grassland ecosystem, investigators are using DNA sequencing to monitor changes in microbial community composition and the expression of genes mediating the processing of plant litter. This information is being used to develop a predictive mathematical model of microbial community response to climate change variables.

Select Milestones in Genomics

1977
- First complete genome sequence, bacteriophage ϕX174 (a virus that infects bacteria).

1986
- Announcement of DOE's initiative to map and sequence the human genome.

1990
- Start of international, publicly funded Human Genome Project, which was led in the United States by the National Institutes of Health and DOE.

1991
- Publication of first analysis of bulk DNA isolated from the environment (northern Pacific Ocean).

1995
- First complete genome of a free-living organism, the bacterium *Haemophilus influenzae.*

1996
- DOE-supported, first complete genome of an archaeal microbe (methane-producing *Methanocaldococcus jannaschii*), confirming Archaea as a biologically distinct third domain of life.
- First complete genome of a eukaryote, the yeast *Saccharomyces cerevisiae.*

1999
- DOE Joint Genome Institute (JGI) facility established in Walnut Creek, California, to complete chromosomes 5, 16, and 19 in the Human Genome Project. JGI's capabilities are later used to sequence microbial genomes, metagenomes, and plant genomes important to DOE missions.

Methane producer, *M. jannaschii.*

- DOE BER–supported completion of genome for bacterium *Deinococcus radiodurans* R1, one of the most radiation-resistant organisms known.

2000
- Completion of a "working draft" sequence covering 85% of the 3 billion base pair human genome. Public versus private efforts to finish the human genome declared a tie.
- First complete plant genome, the small flowering plant *Arabidopsis thaliana.*

2001
- Start of DOE BER Genomes to Life program (now called Genomic Science program) to build a predictive understanding of DOE-relevant biological systems.

2002
- First 100 genomes completed and published.

2003
- International Human Genome Project officially declared complete.

2004
- Extracting DNA from a microbial biofilm thriving in highly acidic mine drainage, a DOE metagenome project is first to reconstruct complete genomes from environmental DNA.

2006
- DOE BER–supported completion of the first tree genome, *Populus trichocarpa,* a type of poplar.

2007
- Researchers working with DOE JGI publish results from termite hindgut metagenome.

2009
- First publication reporting progress on GEBA (Genomic Encyclopedia of Bacteria and Archaea), a DOE BER–supported project to sequence thousands of prokaryotic genomes from diverse branches of the Tree of Life.

2010
- Individual machines can sequence billions of base pairs per week with the potential to sequence a bacterial genome in hours.
- Soybean genome completed by multi-institutional partnership including DOE JGI.

Microbial community thrives in acidic runoff from a mine.

Poplar, the first sequenced tree genome.

The DOE Joint Genome Institute (jgi.doe.gov) has sequenced hundreds of microbial genomes as well as several plant genomes and metagenomes from diverse environments. The table lists some examples of the range of biological systems relevant to DOE missions.

Some DOE-Supported Genome Projects

Organism	Description	Domain	Base pairs	Genes
Methane-producing microbe *(Methanocaldococcus jannaschii)*	Isolated from a hot vent in the Pacific floor, the first archaeal microbe sequenced.	Archaea	~1.7 million	~1,700
Bioenergy microbe *(Caldicellulosiruptor saccharolyticus)*	Thermal spring bacterium that converts plant materials into diverse energy products.	Bacteria	~3 million	~2,700
Metal-reducing microbe *(Desulfovibrio vulgaris Hildenborough)*	Subsurface microbe that can immobilize contaminants in groundwater.	Bacteria	~3.8 million	~3,500
Soil bacteria *(Solibacter usitatus Ellin6076)*	Abundant soil microbe involved in terrestrial carbon cycling.	Bacteria	~10 million	~8,500
Biomass-degrading fungus *(Trichoderma reesei)*	Industrially important fungus that secretes biomass-degrading enzymes for biofuel production.	Eukarya	~34 million	~9,000
Single-celled green alga *(Chlamydomonas reinhardtii)*	Model system for studying the genes for converting carbon dioxide into biomass.	Eukarya	~120 million	~15,000
Black cottonwood tree *(Populus trichocarpa)*	First complete tree genome. Potential bioenergy crop and model for studying carbon accumulation in plants.	Eukarya	~500 million	~45,000
Human *(Homo sapiens)*	Chromosomes 5, 16, and 19 sequenced by DOE for the Human Genome Project.	Eukarya	~3 billion	~25,000

U.S. DEPARTMENT OF ENERGY | Office of Science

February 2011

Image credits. Termite gut microbes: R. Webb, University of Queensland. Termite: ©Dennis Kunkel Microscopy/Visuals Unlimited. Thermal pool microbes: D. Mead, Lucigen. Thermal pool: iStockphoto. Permafrost: DOE ARM Climate Research Facility. 3-D tRNA: Protein Data Bank (www.rcsb.org) PDB ID 4TNA visualized using iMol. Ribosome: Ban, N., et al. 2000. *Science* **289**, 905–20, reprinted with permission from American Association for the Advancement of Science. Metabolites: ChemDB (cdb.ics.uci.edu) chemical ID 3965499, 3966782, and 3965278 visualized using iMol. Methane-producing microbe: B. Boonyaratanakornkit, D. Clark, and G. Vrdoljak, University of California, Berkeley. Acid mine microbes: L. R. Comolli, Lawrence Berkeley National Laboratory. Poplar trees: DOE National Renewable Energy Laboratory. Biodiesel microbe: J. Remis, Lawrence Berkeley National Laboratory. Symbiotic relationship in tree roots: S. Taghavi and D. Nykypanchuk, Brookhaven National Laboratory.