DOE CONTRIBUTION TO SEQUENCING THE HUMAN GENOME

The U.S. Department of Energy Joint Genome Institute (JGI) was launched in 1997 based on the same premise that motivated the Nobel Prize-winning University of California (UC) physicist Ernest Orlando Lawrence to found the first national labbratory more than six decades earlier: large-scale, groundbreaking science can best be achieved by combining the intellectual assets of physicists, biologists, mathematicians, and engineers. At the JGI, the DNA sequencing resources of the DOE national laboratories managed by UC (Lawrence Berkeley, Lawrence Livermore, and Los Alamos) were united to speed the sequencing and characterization of human chromosomes 5, 16, and 19, which total 11% of the human genome.

The broader JGI partnership entails productive relationships with other academic institutions and national laboratories. The Stanford Human Genome Center played a vital role in the capacity of "finishing" the human DNA sequence. The finishing process ensured that the informa-tion made available through the public databases was completely contiguous, with all ambiguities resolved. Other JGI partners include Oak Ridge National Laboratory (for genome annotation), Brookhaven National Laboratory (for genome sequencing-related molecular biology), and Pacific Northwest National Laboratory (proteomics— elucidating protein structure and function). The JGI's budget of approximately $60 million per year comes from the Office of Biological and Environmental Research in the DOE Office of Science to support primarily 180 researchers, administrators, and support staff at the JGI's Production Genomics Facility in Walnut Creek, California. Since the Human Genome Project completion, JGI has extended its charge to whole-genome shotgun sequencing projects devoted to comparative studies of model system vertebrates, microbes and microbial communities, aquatic organisms, and plants. In 2004, JGI initiated the Community Sequencing Program (CSP) to provide the greater scientific community with access to JGI's high-throughput sequencing for large-scale (typically over 50 million base pairs) projects geared toward advancing the frontiers of genome sciences.

Of particular emphasis will be the emerging field of ecogenomics where microbial communities in the environment are revealed through sequencing rather than through laboratory cultivation. An example is the JGI collaboration with UC Berkeley characterizing the microbial communities from Iron Mountain, California. (*Nature,* **248,** 37-43 [04 March 2004]).

CHROMOSOME 5

Expansive Gene "Desert"

Chromosome 5, at 180.8 million base pairs containing 923 protein-encoding genes, is one of the largest human chromosomes yet has one of the lowest gene densities. Vast regions known as gene deserts feature extensive stretches of non-coding DNA that are conserved across numerous vertebrates. The ancient evolutionary roots of this genetic motif suggest a vital functional role.

Pilot studies on chromo-some 5 at Lawrence Berkeley National Laboratory focused on a cluster of interleukin genes which enhance the immune system against disease.

This megabase (million base) region illustrates how multi-mammalian sequence comparisons have lead to the identifi-cation of non-coding elements possess-ing gene regulatory activities. These gene deserts appear to influence the regulation of genes separated by distan-ces of as much as a megabase or more. Genes of interest include ADHD (attention-deficit/hyper-activity disorder), obesity, asthma, and colorectal cancer.

Nature **431,** 268 - 274 (16 Sep 2004)

Chromosomes are packages of DNA and protein located in the nuclei of human and other eukaryotic cells.

CHROMOSOME 16

Highly Repetitive Terrain

A focus of early studies at Los Alamos National Laboratory, Chromosome 16 is highly enriched for segmental sequence duplications—regions that have been copied to other places within the chromosome or copied to other chromosomes. Excluding the Y sex chromosome, chromosome 16 has the most segmental duplications in the human genome.

At 88.7 million bases, it has 880 genes. It features genes implicated in the development of breast and prostate cancer, Crohn's disease, and adult polycystic kidney disease, which affects an estimated five million people worldwide. Half the affected people require dialysis or kidney transplant.

The human genome has undergone extensive periods of segmental duplication, which played a funda-mental role in disease and chrom-osome evolution. These duplications, depicted in purple, vary across the genome, with the level of conserv-ation indicative of when they took place along the primate evolu-tionary tree. Duplications with higher levels of similarity are presumed to have occurred more recently than duplications with lower percent identity. Deletions of these duplicated regions can lead to instability and disease.

CHROMOSOME 19

Gene Mother Lode

Chromosome 19, at 63.8 million bases, although representing only about two percent of the human genome, features nearly 1461 genes including those that code for cardiovascular disease, insulin-dependent diabetes, and migraines.

For nearly 20 years, investigators at Lawrence Livermore National Laboratory sought to uncover the critical regulatory networks embedded in the sequence of Chromosome 19 tasked with repairing DNA damage caused by exposure to radiation and to other environmental pollutants. DNA-repair mechanisms have a role in detoxifying and excreting chemicals foreign to the body. Defects in these pathways are implicated in the development of certain cancers.

Chromosome 19 not only has more than twice the gene density of the genome-wide average but also highlights large blocks of rodent gene conservation and segments of coding and non-coding conser-vation with *Fugu.*

Nature, **428,** 529-535 (01 April 2004)

P ARM

Q ARM

Gene density, depicted by the violet shading on the outer left quadrant of the chromosome figures, varies throughout the genome. Approx-imately one-quarter of the human genome consists of gene-poor regions, or deserts, greater than 500,000 bases. Comparisons of gene deserts in humans, mice, and fish have revealed regulatory elements residing in these deserts that have the ability to modulate gene expression.

Each chromosome is divided into two segments or "arms"—the short, or "p" arm (from the French "petit," meaning small) and the "q," or long arm. The symbol "q" was chosen simply because it followed "p" in the alphabet and is placed below the "p" arm. These sections are linked at the centro-mere, the juncture where the chromosome attaches during cell division.

Only 2% of the human genome encodes protein, leaving the vast proportion of non-coding DNA sequences as one of the mysteries still being deciphered. Comparative genomics between human and other vertebrates— rodents and the puffer fish, *Fugu rubripes*— serves to distinguish functional regions based on conservation between species. Mouse-human non-coding conservation is depicted here in red. With genomes of comparable size (three billion bases), it has been 80 million years since the last human-mouse common ancestor. Human-mouse comparisons have provided clues to such functional non-coding sequences as those regulating immune response.

More than 400 million years of evolution has transpired since primates and fish last shared a common ancestor. Although the *Fugu* genome contains a significant proportion of similar genes and regulatory sequences as a human, they are condensed in approximately 400 million bases or nearly eight-fold less DNA than the human genome. With far less non-coding DNA to sift through, conserved regions between these species have proven successful in revealing functionality. *Fugu*-human non-coding sequence conservation is depicted in turquoise.

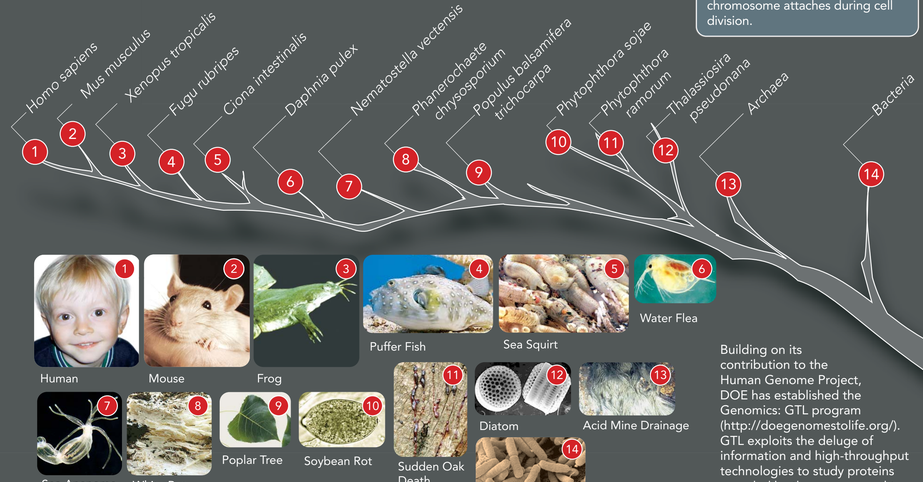
1 Human gene density
2 Segmental duplication
3 Human/mouse non-coding conservation
4 Human/*fugu* non-coding conservation

The chromosome figures depicted here are sister chromatids, that is, two identical copies of a single chromosome that are connected by a centromere. Sister chromatids are created during the interphase period of the cell cycle.

The number assigned to chromo-somes is inversely proportional to their size, so that, excluding the sex chromosomes X and Y, Chromosome 1 is the largest and the smallest is Chromosome 22.

Each chromosome contains a single molecule of the DNA double helix arranged in long stretches of nucleotides, one of four different nitrogenous bases: adenine (A), thymine (T), cytosine (C), and guanine (G). The order of these bases along a strand of DNA is the genome sequence.

Histones are the proteins around which DNA coils to form chromatin, which organize into chromosomes. The human genome is comprised of 23 pairs of chromosomes, a set donated by each parent.



The figure above, a phylogenetic tree or tree of life, is one way to depict evolutionary relationships among organisms that the JGI has sequenced. The branch lengths are not proportional to the esti-mated evolutionary time between organisms or their sequences.

Building on its contribution to the Human Genome Project, DOE has established the Genomics: GTL program (http://doegenomestolife.org/). GTL exploits the deluge of information and high-throughput technologies to study proteins encoded by the genome and to harness the diverse natural capabilities in microbes.

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY

Office of Science U.S. DEPARTMENT OF ENERGY

TEID JO9313

Lawrence Livermore National Laboratory · Los Alamos NATIONAL LABORATORY · shgc

http://www.jgi.doe.gov/