EMBARGOED UNTIL 1 P.M. EDT (17:00 GMT) WEDNESDAY, SEPT. 5:
Major advances made in compiling an encyclopedia of DNA elements; UW a key force in Project ENCODE

An international team of researchers has made significant progress toward compiling a comprehensive listing of all the working parts of the human genome. Their results will be reported in more than 30 papers available to the public, Wednesday, Sept. 5. The University of Washington in Seattle is a major contributor to this effort, which is being conducted largely under the auspices of a multi-nation consortium called ENCODE (ENcyclopedia Of DNA Elements). The National Human Genome Research Institute of the National Institutes of Health is a chief source of ENCODE funding.

Dr. John A. Stamatoyannopoulos, associate professor of genome sciences and medicine at the UW, director of the UW ENCODE center, and a senior author on seven ENCODE-related papers, explains why understanding how the human genome functions is important to progress in genomic medicine:

"The first phase of the human genome project provided the primary genome sequence, and a basic catalog of genes, which occupy only 2 percent of the genome. Every cell in the body has the same genes, but different kinds of cells, such as liver or heart, switch on different combinations of genes. When cells become unhealthy, these combinations change. Understanding how genes turn on and off is therefore vital to deciphering their role in both normal health and disease. The instructions for how genes are controlled are contained in small DNA 'switches' that are scattered around the 98 percent of the genome that does not contain genes. Mapping and decoding these instructions is a central mission of the ENCODE project, and the focus of work at the UW ENCODE center. Data generated in this project so far have already shown, for example, that common DNA variations in the gene-controlling switches can affect the risk of developing different common diseases. This finding, together with the emerging wealth of information about the basic mechanisms of gene control, is opening new vistas on preventing, diagnosing, and treating disease."

Stamatoyannopoulos is also author of a headlining article for the ENCODE special issue of *Genome Research* that gives an overview and perspective on the project, its accomplishments, their significance, and prospects for the future.

Many UW researchers contributed to ENCODE-related research. Major discoveries include:

1) The first detailed maps of regulatory DNA switches that make up the genome's 'operating system.

Researchers located millions of DNA 'switches' that dictate how, when, and where in the body different genes turn on and off. These switches, or regulatory DNA, contain small chains of DNA 'words' that make up docking sites for proteins involved in gene control. Often these switches are far away from the genes that they control. Of the millions of regulatory DNA regions, only a small fraction, around 200,000, are active in any given cell type. This fraction is almost unique to each type of cell, a sort of molecular bar code of its identity. The regulatory 'program' of most genes has more than a dozen switches. _Nature_ paper: The accessible chromatin landscape of the human genome.

2) The first extensive map of regulatory protein docking sites on the human genome reveals the dictionary of DNA words that comprise the genome's programming language.

To find the DNA words recognized by regulatory proteins, researchers employed a simple, powerful trick to study all the proteins at once. Instead of trying to see proteins directly, they looked for their footprints on the DNA. They discovered that over 90 percent of the protein docking sites were actually slight variants of about 680 different DNA words -- a dictionary of the genome's programming language. _Nature_ paper: An expansive human regulatory lexicon encoded in transcription factor footprints.

3) A comprehensive wiring diagram provides insights into how cells 'think'

 The genome senses and responds to signals received from other parts of the cell and from the environment by changing the activity of regulatory proteins. Scientists mapped all of the connections between regulatory protein genes to create a central wiring diagram for the cell. Using powerful computers, they created, in a matter of weeks, wiring diagrams o how 475 regulatory protein genes were connected to each other, and how those connections changed across 41 different types of human cells. Conventional methods would have required nearly 20,000 different experiments, taking several years to complete and costing over one hundred million dollars. Even though individual connections between regulatory proteins differed among cell types, the overall connection was nearly the same in all cell types. When compared to the best-studied biological network -- the map of all connections between neurons in the worm brain, created by Nobel Prize winner Sydney Brenner – the layout is almost identical. Nature seems to have settled on an ideal 'brain-like' architecture to process complex biological information; this plan can be found in the genomic wiring of every living cell. _Cell_ paper: Circuitry and dynamics of human transcription factor regulatory networks.

4) Unlocking disease information hidden in the genome's control circuitry.

Hundreds of studies have attempted to map the genes causing common diseases and physical traits. Frustratingly, most of these studies have pointed to regions of the genome that do not contain gene sequences that make protein. Researchers set out to chart a global map of the relationship between disease-associated genetic changes and the gene-controlling switches scattered around the genome, With support from National

Institutes of Health's Common Fund, researchers collected regulatory DNA maps from 349 tissue samples covering all major organ systems in adults and stages of human development. Using powerful computers, the researchers crossed these maps with data from genetic studies of over 400 common diseases and clinical traits. Instead of isolated instances, they found that most disease-associated genetic changes occurred within gene-regulating switches, often located far away from the genes they control. Most changes affected circuits active during early human development, when body tissues are most vulnerable. Extensive blueprints of control circuitry revealed previously hidden connections between diverse diseases, may explain common clinical features, and will open new avenues for developing diagnostics and treatments. _Science_ (cover feature): Systematic localization of common disease-associated variation in regulatory DNA.

5) Differences in regulatory DNA between people and human populations, and evolutionary changes from natural selection

In comparing genomes from individuals from several parts of the world, a team led by Benjamin Vernot, Joshua Akey, and Stamatoyannopolous found that changes affecting gene control regions are very frequent. In the average individual these changes dwarf those found in DNA that encodes proteins. By performing genome-wide scans for areas of recent evolutionary change, scientist discovered evidence that hundreds regulatory DNA regions had been targeted by natural selection, presumably because of their roles in biological pathways important for human survival, such as skin pigmentation and fat storage. _Genome Research_: Personal and population genomics of human regulatory variation.

6) New insights into the genome's 'master weaver'

 Not all regulatory proteins are created equal: CTCF has earned the title of the genome's 'master weaver' because it not only controls genes but determines how DNA is wound up within the cell nucleus. Prior research had suggested that the sites along the genome to which CTCF liked to dock were nearly the same in every cell. By CTCF binding across many cell types, UW researchers have found patterns varied among different cells, and between normal cells and cells that keep growing indefinitely, such as cancer cells. Many changes between cell types were accompanied by chemical changes in DNA known as methylation, which has been linked with aging. _Genome Research_: Widespread plasticity in CTCF occupancy linked to DNA methylation.

7) Teaching computers to find patterns in ENCODE "big data"

UW researchers in Genome Sciences and Electrical Engineering, led by William Noble and Michael M. Hoffman and have applied machine learning techniques to find patterns in the structure of human DNA and associated biomolecules. Machine learning involves the use of computer programs that can learn to classify big data sets into human-interpretable categories. The team designed a computer program and trained it to examine and characterize data on the location of chromatin modifications. Chromatin is the three-dimensional structure that DNA forms as it wraps around bead-like protein structures known as nucleosomes. The team used the software to identify patterns associated with genes and other DNA elements important in the regulation of gene activity

###