ENCODE project: Millions of DNA switches that power human genome's operating system discovered

September 5 2012

The locations of millions of DNA 'switches' that dictate how, when, and where in the body different genes turn on and off have been identified by a research team led by the University of Washington in Seattle. Genes make up only 2 percent of the human genome and were easy to spot, but the on/off switches controlling those genes were encrypted within the remaining 98 percent of the genome.

Without these switches, called regulatory DNA, genes are inert. Researchers around the world have been focused on identifying regulatory DNA to understand how the genome works. Using a new technology developed with funding from the National Human Genome Research Institute's ENCODE (ENCyclopedia Of DNA Elements) project, UW researchers created the first detailed maps of where regulatory DNA is located within hundreds of different kinds of living cells. They also compiled a dictionary of the instructions written within regulatory DNA—the genome's programming language.

The findings are reported in two papers appearing in the Sept. 5 online issue of *Nature*.

"These breakthrough studies provide the first extensive maps of the DNA switches that control human genes," said Dr. John A. Stamatoyannopoulos, associate professor of genome sciences and medicine at the University of Washington, and senior author on both

papers. "This information is vital to understanding how the body makes different kinds of cells, and how normal gene circuitry gets rewired in disease. We are now able to read the living human genome at an unprecedented level of detail, and to begin to make sense of the complex instruction set that ultimately influences a wide range of human biology."

Here are the key results:

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

The instructions within regulatory DNA are inscribed in small DNA 'words' that function as the docking sites for special proteins involved in gene control. In many cases, these switches are located far away from the genes that they control. To map the regulatory DNA regions, the researchers harnessed a special molecular probe—an enzyme called DNaseI—that snips the genome's DNA backbone. Under the right conditions, these snips occur precisely where proteins are docked at regulatory DNA. By treating cells with DNase I and analyzing the patterns of snipped DNA sequences using massively parallel sequencing technology and powerful computers, the researchers were able to create comprehensive maps of all the regulatory DNA in hundreds of different cell and tissue types. They found that of the 2.89 million regulatory DNA regions they mapped, only a small fraction—around 200,000—were active in any given cell type. This fraction is almost totally unique to each type of cell and becomes a sort of molecular bar code of the cell's identity. The researchers also developed a method for linking regulatory DNA to the genes it controls. The results of these analyses show that the regulatory 'program' of most genes is made up of more than a dozen switches. Together, these findings greatly expand the understanding of how genes are controlled and how that control may differ between normal and diseased cells.

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

The instructions for turning genes on and off are written in DNA switches called regulatory DNA. These switches are scattered throughout the non-gene regions of the human genome. Having mapped the locations of the regulatory DNA switches, UW researchers wanted to know what made them tick. These regions contain small chains of DNA 'words' that make up docking sites for special regulatory proteins involved in gene control. The human genome contains hundreds of genes that make such proteins.

However, current technologies only allow such proteins to be studied one at a time. They also lack the accuracy to resolve the DNA letters to which the proteins dock. As a result, most of the actual DNA words recognized by regulatory proteins in living cells were unknown. To find them, the researchers employed a simple, powerful trick that enabled them to study all the proteins at once.

Instead of trying to see proteins directly, they looked for their shadows or 'footprints' on the DNA. To accomplish this, they again turned to the DNaseI enzyme that snips the DNA backbone within regulatory DNA. Prior work had shown that DNaseI likes to snip DNA next to regulatory protein docking sites, but not within the docking site itself. By using next-generation DNA sequencing technology, the researchers analyzed hundreds of millions of DNA backbone breaks made when cells were treated with DNaseI. They then used a powerful computer to resolve millions of protein footprints. In total, they identified 8.4 million such footprints along the genome, some of which were detected in many cell types. Next, they compiled all of the short DNA sequences to which the proteins were docked. They analyzed them using a software algorithm that required hundreds of microprocessors working simultaneously. This

revealed that more than 90 percent of the protein docking sites were actually slight variants of 683 different DNA words—essentially a dictionary of the genome's programming language.

"These findings significantly advance the understanding of how the instructions for controlling genes are written and organized throughout the genome, and how combinations of different instruction sets function together to control genes, often at great distance along the genome," Stamatoyannopoulos said. "The broad spectrum of cell and tissue types included in these analyses provide an incredibly rich resource that can be mined immediately by researchers around the world to illuminate how the genes they are studying are controlled."

The scientists determined that genes are connected in a complex web. In this web, regulatory DNA regions typically control one or at most a few genes, but genes receive inputs from large numbers of regulatory regions. The researchers also found evidence for a combinatorial code that helps match regulatory DNA with the right genes. Another key finding was that the regulatory DNA controlling genes involved in cancer and other types of 'immortal' cells that can keep on growing indefinitely appears to acquire mutations at a different rate than other kinds of regulatory DNA. This result points to a previously unknown link between genome function and patterns of DNA variation in individual human genomes. The finding may have implications for understanding susceptibility to cancer.

The findings reported in these papers are expanded upon in two related papers to be published simultaneously in the journals *Science* and *Cell*. In the Science paper, UW researchers further expanded the regulatory DNA maps, and compared them with genetic maps of human disease. Their studies revealed that most DNA variants associated with specific human diseases or clinical traits are located in regulatory DNA rather than in gene sequences. In the *Cell* paper, the researchers describe using

the detailed information on regulatory protein docking sites to create a comprehensive map of how those proteins are wired.

Provided by University of Washington