Next gen sequencing technology pinpoint 'on-off switches' in genomes

February 12 2009

Activity pattern of one of nearly 5000 potential genetic switches (enhancers) identified by Visel et al. This particular switch is located on human chromosome 5 and turns on genes in developing mammalian limbs, as shown here by reporter gene staining (dark blue) in a transgenic mouse embryo. Credit: Axel Visel, Lawrence Berkeley National Laboratory

Scientists from the U.S. Department of Energy (DOE) Joint Genome Institute (JGI), Lawrence Berkeley National Laboratory, and the University of California, San Diego have developed a set of molecular tools that provide important insight into the complex genomes of multicellular organisms. The strategy promises to clarify the longstanding mystery of the role played by vast stretches of DNA

sequence that do not code for the functional units—genes—that nevertheless may have a powerful regulatory influence. The research is described in the 12 February edition of the journal *Nature*.

DOE bioenergy researchers have an interest in identifying these regulatory regions in plants, where proteins interact with DNA and exert control over gene expression and development, so that plants used as biomass "feedstocks" can be optimized for biofuels production.

"From the Human Genome Project we have a good idea where in the genome the protein-coding genes are located, but these constitute only about two percent of the human genome, the remaining 98 percent are non-coding sequence whose function is largely unknown," said Len Pennacchio, the paper's senior author and DOE JGI Genomic Technologies Department Head.

"Our approach employs next generation sequencing technology to find regulatory regions, the 'switches' on a genome-wide scale and much more cost effectively," said Pennacchio. "It's the next layer of knowledge that's been missing."

The DOE JGI was founded in 1997 to accelerate the completion of the HGP and completed the DOE's commitment to sequence three (5, 16, 19) of the 23 chromosomes, totaling 11 percent of the human genome, and published the analysis in Nature back in 2004.

In this newly published study, Pennacchio, lead authors DOE scientist Axel Visel and postdoctoral fellow Matthew Blow, and their colleagues, describe a shortcut for identifying gene regulatory regions or the molecular switches that turn on or off gene expression.

Using what's called ChIP-Sequencing or ChIP-Seq, chromatin immunoprecipitation (ChIP) is combined with massively parallel DNA

sequencing to identify binding sites of DNA-associated proteins.

Traditionally researchers have relied on evolution to guide them to non-coding sequences that are likely to have a function—such as enhancing the expression of genes. Via the public genome databases, they would align the entire human genome code with that of other vertebrate species (e.g. other mammals, birds, frogs, fish) and then look for sequences that are conserved in evolution.

"Most protein-coding sequences show signs of conservation between species, but there is also a large number of non-coding sequences that have been surprisingly well conserved for tens or even hundreds of millions of years," said Visel. "This suggests that these regions, formerly thought to be "junk" DNA, actually have some functional relevance and are under selection because sequence changes reduce fitness of affected individuals. Using such sequence conservation, we have in previous studies identified enhancer candidate regions and shown in transgenic mouse experiments that these conserved non-coding regions are in fact often enhancers that are active during embryonic development. Conservation-based methods are relatively good at finding enhancers in the genome, but an important limitation is that they don't tell us where and when that particular enhancer would be active and thereby drive the expression of its neighboring target gene(s).

The older methods lacked specificity, Blow said. "For example, if we have a gene that is important both for brain and for limb development, we would not have been able to specifically identify the enhancer sequences near that gene that would drive the expression in the brain or limb, the only way to find out was to test these activities in experiments one-by-one, which is slow and can't be done on a genomic scale.

"Using this new method, we can directly identify a genome-wide set of enhancers that are active in a particular anatomical region or tissue at a

particular time-point, which is an important advantage over conservation-based methods because in addition to telling us where an enhancer is located in the genome, it also provides an initial experimental characterization where we should expect this enhancer to be active."

The team used ChIP linked with a particular enhancer-associated protein, p300, then directed DOE JGI's massively parallel next generation sequencing capacity to map several thousand sites in mouse embryonic forebrain, midbrain and limb tissue. Over 80 of these fragments were tested in transgenic mouse experiments indicating an almost perfect success rate of p300-ChIP-Seq for identifying enhancers active in vivo.

"Enhancers are especially important for regulating genes during embryonic development," said Pennacchio. "They can regulate genes over long distances and switch on their target genes during very specific time-points and in very specific anatomical structures during development. There are several examples of mutations in such enhancers that cause disease in humans because genes are not expressed at the right time or in the right place anymore. A fundamental problem in studying such enhancers is that until recently we did not have effective tools to even find them in the genome on a large scale.

Pennacchio said that this new method will prove useful to the greater genomics and biomedical community for characterizing the role of the vast non-coding regions—dubbed genome "dark matter"—about which little is known.

"These datasets will also help to identify mutations in enhancers that play a role in human disease," Pennacchio said. "Human genetic studies indicate that in many cases disease is caused by mutations in non-coding sequences, but it has been difficult to study this in detail because the function of most non-coding sequences is poorly understood. Eventually,

this will be useful for purposes including disease detection and personalized medicine."

With the rapidly increasing efficiency and cost-savings of the next generation sequencing technologies, a deluge of data from individual human genomes are being to come to light, to the point where whole-genome sequencing of patients may soon become a standard diagnostic tool.

"While progress is being made towards this goal, it is important to keep in mind that our current understanding of the genome has focused on protein-coding sequences," said Pennacchio. "Datasets like the one provided through this study will be important to understand the remaining 98 percent of the genome and what its role in health and disease is."

The published study provides an important proof of principle to establish and validate a new method in three different mouse tissues at a single embryonic time-point, Pennacchio said. "We can now generate genome-wide enhancer datasets directly from human tissues and compare genome-wide sets of enhancer activities between healthy people and people suffering from disease, which may reveal how enhancer activities change on a global scale in these disease states."

Source: DOE/Joint Genome Institute

study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.