BIOTECH Basics

Exploring and Understanding the ENCODE Project

What you need to know:

- The Human Genome Project identified the sequence of the ~ 3.2 billion letters in a human genome and the location of ~ 21,000 genes. The ENCODE Project seeks to understand the ways those genes are regulated in cells.
- The location of millions of functional elements has been identified, although the biological significance of many of these locations is uncertain.
- The ENCODE data is useful for biological and disease-related research, helping correlate DNA regions associated with disease and the functional significance of those regions.

ENCODE stands for the ENCyclopedia Of DNA Elements. Launched in 2003, the decade-long project has sought to catalog all of the regions in human genomes that control how genes function and facilitate understanding of when and in which cells they are used. Following up a series of major publications in 2007, findings from the most recent phase of ENCODE were published in September 2012, providing new insights into how our genome is organized and regulated.

As an analogy, think of a passage from a story or poem. Imagine the text is missing punctuation, spacing and those visual cues that make sense of the language. All that is present are the individual letters that make up the words. As an example, we'll consider the first sentence from Edgar Allan Poe's famous poem *The Raven*.

onceuponamidnightdrearywhileipondered-weakandwearyovermanyaquaintandcurious-volumeofforgottenlore

Adding in other parts of language provides meaning to the string of text – things like capital letters, commas and spaces. From a string of seeming nonsense, familiar words appear.

Once upon a midnight dreary,
while I pondered, weak and weary,
Over many a quaint and curious
volume of forgotten lore –

In a similar way, the ENCODE Project seeks to help scientists make sense of human genomes – by understanding the biological language contained in the sequences of letters in our DNA. The Human Genome Project determined the locations of more than ~21,000 genes among ~3.2 billion DNA nucleotides. However, these genes only account for 1-2 percent of an entire genome. This means the functional significance of much of the remainder has been a mystery. The ENCODE Project is the genetic equivalent of providing spacing and punctuation – a set of experiments to determine which pieces of DNA regulate the action and storyline of a human genome.

Some analyses identified regions where proteins known as transcription factors bind DNA and control gene transcription – the process that copies the genetic message into RNA. Other experiments searched for DNA methylation – small molecules that attach to the DNA sequence. Transcription factor binding and DNA methylation can greatly increase or completely silence the activity of a corresponding gene - dramatically altering the level of RNA it produces. Varying levels of RNA influence how much protein can be produced. This may have important consequences for how cells function or interact with neighboring cells.

The ENCODE Project involved 442 researchers from 32 institutions around the world. All told, 1,649 experiments that looked across the genome were performed. From these, scientists identified the location of millions of functional elements in our genome. The data show the genome as a very active place – a beehive of molecules docking at specific DNA sequences. These molecules initiate transcription, regulate genes thousands of bases away or control how the DNA is folded. Scientists have known these regulatory processes exist, but the scope and scale took many by surprise. Each cell type uses only a small subset of these regulators, which in part is what differentiates those cells – e.g. skin cells rely on a different combination of genes and regulators than muscle or liver cells. Choosing from a vast array of biological controls lets the timing and level of gene activation be finely tuned to the needs of a cell.

However, the biological significance of many of these active elements is hotly debated in the genomics community. For example, roughly half of our genome is composed of repetitive sequences of DNA and fragments of inactive genes - leftovers from our evolutionary history. Although sometimes described in the popular press as "junk" DNA, scientists avoid using this word as it is both

If you want to know more:

www.encodeproject.org

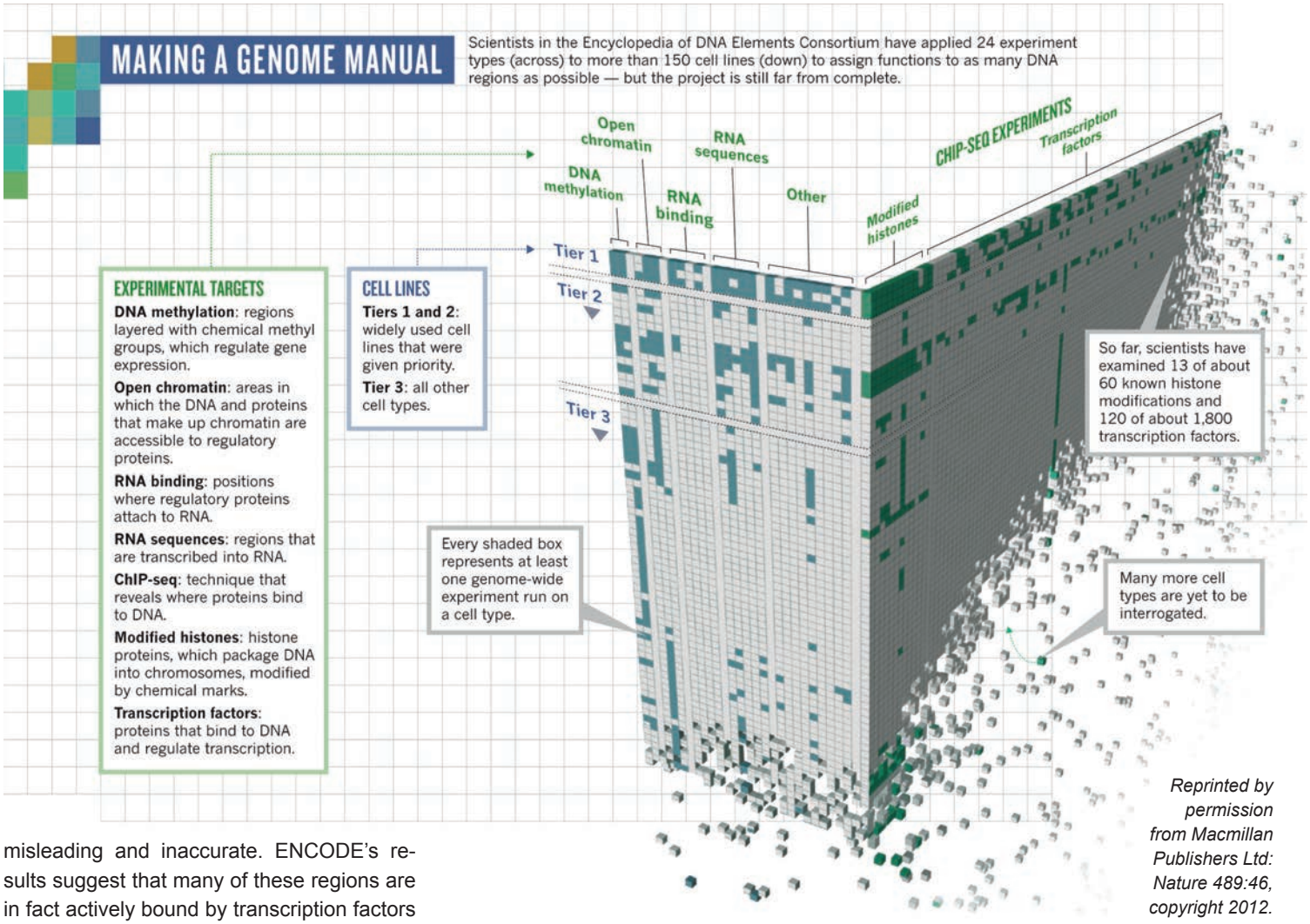
As with many large-scale genomic projects, the data are publicly available, allowing scientists in all fields to better understand how human genomes are organized and regulated.

www.nature.com/encode

The results were simultaneously published in 30 different papers. In order to navigate through the findings, the scientific journals created an online portal site that allows individuals to explore one of 13 topics of interest, following "threads" that gather the relevant information from each paper. This data are also available in a free app for iPhone and iPad users.

http://www.hudsonalpha.org/more-encode

Still not sure you really understand what ENCODE is all about? Take a look at HudsonAlpha's "Understanding ENCODE" video on our YouTube channel.



Reprinted by permission from Macmillan Publishers Ltd: Nature 489:46, copyright 2012.

misleading and inaccurate. ENCODE's results suggest that many of these regions are in fact actively bound by transcription factors and exhibit other types of molecular activity. If we incorporate this concept into the literary example mentioned earlier, the opening line of the poem would looks more like this:

Once up, up, up, upon a midnight very, very, very, very, dreary, while I, I, I, I, I pondered, weak and w-, w-, w-, weary,

Genomic scientists don't yet know what fraction of this binding and transcriptional activity is important for the daily activities of a cell. For example, because DNA binding proteins use short DNA sequences (just a handful of letters) as a binding target, the chance occurrence of these docking sites is high and will likely occur many times in genomes that are billions of letters in length, especially if those target sequences are repeated over and over again. These segments may in turn produce RNA and mimic the features of being functional, even if they serve no purpose. Although further studies are needed, it is likely that many of the ENCODE-detected activities are not biologically important or needed by our cells. They are essentially molecular noise.

Setting aside the debate over what fraction of the genome is functionally important, the ENCODE Project represents a landmark scientific achievement. It picks up where the Human Genome Project left off, in terms of identifying context and meaning for the genome. ENCODE has provided a deep and high-quality description of genomic activity that will be useful throughout many biological and disease research areas. The next step is to figure out how the various players in this regulatory symphony interact. For example, if a binding site is altered or deleted through mutation, is there an effect on the body? Over the last several years, researchers have used a technique called genome-wide association study (see the Summer 2009 *Biotech Basics* for more details) to identify thousands of DNA variants associated with disease. The vast majority of these associations occur outside of known gene regions. The assumption has been that they occur in regulatory regions that control the activity of disease-related genes. In support of this hypothesis, all known associated regions were cross-referenced to the ENCODE data. At least 88 percent lie outside of protein-producing sections of the genome, many of which land in the ENCODE-identified regulatory sites. If these findings are validated, they offer insight into the process of disease progression and may lead to new candidates for therapy. The ability to correlate association with function is a major benefit of the ENCODE data.

While the ENCODE Project represents an important milestone, a great deal of work still remains to fully decipher the functional components of our genome. Less than 150 of the thousands of human cell types were examined and only a fraction of known transcription factors were examined. A new phase of the project is just beginning to dig deeper into the regulatory process.

- Neil Lamb, Ph.D.
director of educational outreach
HudsonAlpha Institute
for Biotechnology