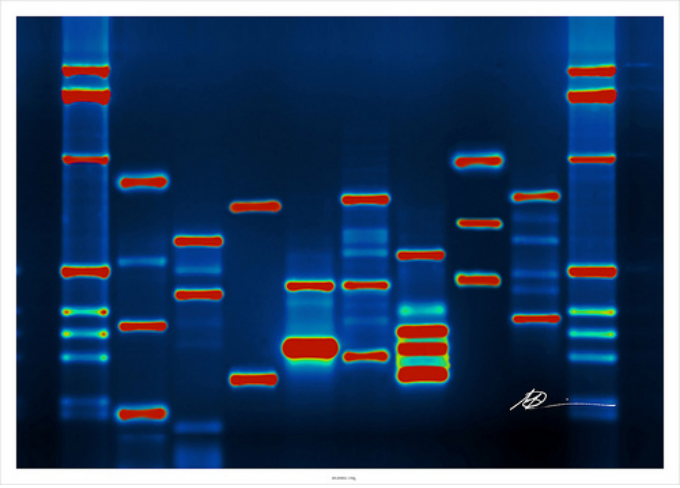
Human Genome Project 2.0

Meet ENCODE, the ultimate guide to the human genome. But the project isn't just changing the way we look at our DNA, it's also revolutionizing the way in which scientist publish their research findings.



The ENCODE project aims to revolutionize the way large
scientific consortia present their data. Image courtesy Micah
Baldwin, Flickr.

Posted on SEP 19 2012 5:01AM

Andrew Purcell
European editor

Share this story

On 5 September, the ENCODE project simultaneously published 30 research papers in three top journals: *Nature*, *Genome Research* and *Genome Biology*. Review articles have also been published in the journals *Science*, *Cell* and *The Journal of Biological Chemistry*. In other words, ENCODE is a big deal; a very, very big deal in fact.

But what exactly is ENCODE? Why are scientists getting so excited about it? And what on Earth does it have to do with computing anyway?

ENCODE stands for 'The Encyclopedia of DNA Elements'; it's basically the Human Genome Project 2.0. It seeks to move science beyond simply telling us what the human genome looks like to telling us how it works and this is what each part does. Achieving this has involved a team of 442 researchers from various scientific fields working for over a decade in labs across the globe. Collectively, they have produced 1,640 genome-wide data sets taken from experiments conducted on 147 different human cell types.

These data sets have, for the first time, comprehensively shown that over 80% of the human genome has a biochemical function*. And, Ewan Birney, the project's Lead Analysis Coordinator, claims that this figure could end up being as high as 100% once further cell types have been studied. To put this into perspective, we know that only 1.5 % of the human genome actually codes for the production of proteins, so to find out what the rest does is hugely important. ENCODE not only shows us that the rest of the genome does

Tags

code DNA ENCODE genetics genome human

actually do something, thus dealing a major blow to the theory that a large portion of our genome is 'junk DNA', but it also shows us what it is exactly that each part does, for example: acting as switches to turn genes on or off, influencing the activity of genes (often over great distances), or altering the way in which DNA is folded and packaged.

Of course, finding all of this out required a staggering amount of data. While DNA may only be made up of a simple four-letter code, the sheer amount of genetic material studied meant that the ENCODE project nevertheless needed some serious computing muscle. In total, the project generated over 15 trillion bytes of raw data, requiring the equivalent of more than 300 years of computer time to analyze. If one were to attempt to print this to paper, even at a resolution as high as 1,000 base pairs per square centimeter, the resultant printout would stretch 16 meters high and at least 30 kilometers long.

To overcome the problems associated with analysis of such a large dataset, William Noble, an expert on machine learning from the University of Washington, led a team which designed artificial intelligence programs to analyze the ENCODE data. These computer programs are able learn, recognize patterns, and organize information into categories understandable to scientists. The computer center at the University of Washington is a major contributor to the ENCODE project, analyzing over four petabytes of genomic data a year.

However, the most exciting data innovation of the ENCODE project is undoubtedly its 'virtual machine'. Not content with merely providing an online portal where raw data from the project is available, the ENCODE team created a freely available virtual machine. By running this virtual machine on a cloud computing service, anyone can - at least in theory - verify the project's findings by exactly repeating the analysis steps carried out in the original research. "You can absolutely replay step by step what we did to get to the figure," says Birney. Writing on his blog, he adds: "I believe this virtual machine substantially increases the transparency of this data-intensive science, and that we should produce virtual machines in the future for all data-intensive papers... think of this a bit like the ultimate materials and methods section of the paper."

** Please note: The researchers' definition of what exactly it means to be functional has come in for heavy criticism from a variety of quarters. A thorough summary of this controversy can be found over on Brendan Maher's Nature blog.*

Join the conversation

Contribute

Do you have story ideas or something to contribute? **Let us know!**

OUR UNDERWRITERS

Thank to you our underwriters, who have supported us since the transition from International Science Grid This Week (iSGTW) into Science Node in 2015. We are incredibly grateful.

View all underwriters

CATEGORIES

Advanced computing
Research networks
Big data
Tech trends
Community building

CONTACT

Science Node
Email:
editors@sciencenode.o
Website:
sciencenode.org