

ENCODE project: Researchers catalogue functional elements of the genome

September 5 2012, by Haley Bridger



ENCODE project image. Credit: National Human Genome Research Institute

Most of the DNA alterations that are tied to disease do not alter protein-coding genes, but rather the "switches" that control them. Characterizing these switches is one of many goals of the ENCODE project – a sweeping, international effort to create a compendium of all of the working parts of the human genome that have not been well studied or well understood.

The function of the vast majority of the [human genome](#) has remained largely unknown, but the Encyclopedia of DNA Elements (ENCODE) project, launched in 2003, set out to change that. Comprised of more

than 30 participating institutions, including the Broad Institute, the ENCODE Project Consortium has helped to ascribe potential [biochemical function](#) to a large fraction of the non-coding genome. This work has revealed elements that act like dimmer switches, subtly turning up or down a gene's activity and influencing what parts of the genome are utilized in different kinds of cells. The team characterized and mapped out the locations of thousands of these switches and signals. More than 30 papers detailing these results appear online in *Nature*, *Science*, *Genome Research*, and *Genome Biology* this week.

"With these maps in hand, we can begin to understand why genetic variants that land in the annotated regions may predispose people to disease," said Brad Bernstein, a senior associate member at the Broad Institute and an associate professor of pathology at Massachusetts General Hospital (MGH) and Harvard Medical School. Bernstein is also a principal investigator in the ENCODE Consortium. "It turns out that many of the variants that genetic researchers have tied to various diseases – lupus, Crohn's disease, [metabolic diseases](#), [high cholesterol](#), and much more – sit in these regions that alter how genes are expressed in specific kinds of cells."

Researchers from the Broad, MIT, and MGH found that variants associated with [autoimmune diseases](#) such as lupus and [rheumatoid arthritis](#) sit in regions that are active only in immune cells, whereas variants tied to cholesterol and metabolic diseases sit in regions active in liver cells.

In order to generate detailed maps of the switches that lie between – and even nestled within – genes, data collection centers in the consortium amassed high-quality and comprehensive datasets detailing the function of elements in the genome. By looking across more than 140 cell types, they generated more than 1,500 datasets.

Several groups at the Broad Institute contributed to this work, including the Genome Sequencing and Analysis Program and the Epigenomics Program, which helped generate many of the datasets for the project. The term "epigenome" refers to a layer of chemical information on top of the genetic code, which helps determine when and where (and in what types of cells) genes will be active. This layer of information includes a suite of chemical changes that appear across the genetic landscape of every cell, and can differ dramatically between cell types. Researchers at the Broad and at the other ENCODE data collection centers developed ways to characterize these epigenetic "marks" across cell types.

"By bringing together computational groups from across the world and gathering all of the data generated, we can get at much more complex questions," said Manolis Kellis, an associate member of the Broad Institute, principal investigator at the MIT Computer Science and Artificial Intelligence Lab (CSAIL), and an associate professor of Computer Science at MIT. Kellis is also head of the MIT Computational Biology Group and a principal investigator in the ENCODE Consortium.

Maps and data generated through the ENCODE project have been publicly released as they have become available. With these maps, Bernstein, Kellis, and their Broad colleagues:

- Recognized regions upstream and downstream of genes that control when and where a gene is turned on or off
- Mapped small sequence patterns within these control regions that play important roles in region activity and disease association
- Studied the differences in the behavior of genetic variants inherited from the mother or father
- Grouped elements from across the genome into "neighborhoods" of similar activity that are typically associated with similar gene functions

- Proposed candidate biochemical functions for the majority of genetic variants associated with disease but residing outside of protein-coding genes

The ENCODE project challenged members of the Broad's Epigenomics Program to create new tools and methods and improve data generation. "As a data production center, we take very seriously our responsibility to only let high-quality data out the door," said Chuck Epstein, who manages the Epigenomics Program. "We're constantly trying to develop better ways to know that data are high quality, and that involves developing good systems to annotate and view data."

The project also led to challenging new problems for data integration across different labs and institutions using a variety of experimental protocols. "We developed novel statistical methods and robust automated pipelines for uniform processing, quality control, reproducibility analysis and integration of massive amounts of diverse data," said Anshul Kundaje who led the data coordination efforts of the consortium and is now a research scientist at the MIT Computational Biology group. "This resulted in very high standards of data quality across the consortium as well as freely available software pipelines that we believe will serve as a valuable resource to the larger scientific community."

With epigenomic maps in hand, researchers are now turning to the next phase of the project. They will look across more cell types – each of which contain different epigenomic instructions – and will begin looking at the wiring of these switches.

"We now have a map of the genomic locations of these switches, but we don't have a map showing which switch controls which gene," said Bernstein. "What turns on the switch? And when it turns on, what gene or genes get upregulated? Having a map of the way these elements are wired and connected is a critical goal."

More information: "An integrated encyclopedia of DNA elements in the human genome," *Nature*, [DOI: 10.1038/nature11247](https://doi.org/10.1038/nature11247)

"Landscape of transcription in human cells" is published online in *Nature* on September 5, 2012. The authors are: Sarah Djebali, Carrie A. Davis and 83 others. The paper can be obtained online at [doi:10.1038/nature11233](https://doi.org/10.1038/nature11233). Other new ENCODE results can be found in the following journals: *Nature* (6 papers); *Genome Research* (18 papers); and *Genome Biology* (6 papers).

www.nature.com/encode

Provided by Broad Institute of MIT and Harvard

Citation: ENCODE project: Researchers catalogue functional elements of the genome (2012, September 5) retrieved 20 March 2024 from <https://medicalxpress.com/news/2012-09-encode-catalogue-functional-elements-genome.html>

<p>This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.</p>
--