TOOLBOX

Massive encyclopedia catalogs genome's regulatory regions

BY **CHLOE WILLIAMS**

26 AUGUST 2020

Researchers have created the most comprehensive catalog to date of DNA regions that modify gene expression[1,2]. The encyclopedia — first released in 2007 and updated last month — is helping researchers uncover the functions of DNA segments between genes and their role in conditions such as autism.

Genes make up less than 2 percent of the human genome. Many genetic variants linked to autism are found **outside of genes**, in 'noncoding' sections of DNA. A major challenge for scientists has been deciphering the functions of these noncoding regions.

In 2003, researchers began working on the **Encyclopedia of DNA Elements** (ENCODE), a project that aims to interpret the function of DNA beyond genes.

Initially, scientists combed through 1 percent of the genome to identify stretches of DNA that encode molecules such as proteins, and so-called 'functional elements' that regulate gene expression. In 2012, the team expanded the encyclopedia to include the entire genome, and two years later, they mapped functional elements in the mouse genome.

The project's latest release, ENCODE 3, described in July in *Nature*, includes data from nearly 6,000 experiments, more than twice as many as in the previous version. It also details functional elements in more than 1,300 cell samples from mice and people, representing 503 cell and tissue types — about four times more than in previous releases[3]. A future release is slated to include even more experiments and cell samples, the researchers say.

Researchers can explore the data from all ENCODE experiments **online**.

Regulation register:

To create the encyclopedia, hundreds of researchers ran a slew of experiments to collect molecular clues about the location of functional elements in the genome and their likely roles. For the new release, the researchers studied cells taken directly from people and mice, which may recapitulate healthy tissue more accurately than the lab-grown cells used for previous ENCODE versions.

In one test, for instance, researchers identified sections of **chromatin** — the coiled complex of DNA and proteins called histones — that are accessible to enzymes. Regions that are accessible to enzymes and other proteins are thought to play a role in gene expression.

In another experiment, they mapped the binding sites within the genome of 282 'transcription factors,' proteins involved in modulating gene expression. The team also sequenced stretches of DNA near histones that have chemical tags that alter how tightly DNA is wound. A tighter wind blocks access to proteins that initiate gene expression.

The researchers tapped all of these molecular clues to identify likely regulatory elements, such as promoters, which modify the expression of neighboring genes, and enhancers, which modulate the expression of distant genes. The team classified these regulatory elements based on their associations with distinct histone tags or transcription factors, and on their proximity to the start of a gene sequence. The new catalog documents 926,535 likely regulatory elements in people and 339,815 in mice.

"In any given tissue, there are a few hundred thousand that are active," says **Mark Gerstein**, professor of biomedical informatics at Yale University and one of the project's lead investigators.

The catalog, which can be accessed through a **web-based server**, could help researchers gain a better understanding of noncoding regions of the genome, Gerstein says. Paired with other datasets, it could help scientists identify regulatory regions that are particularly active in the brain.

RNA rank:

In another study, published in July in *Genome Biology*, researchers used ENCODE to develop a way to identify genetic variants that are most likely to compromise proteins that bind to RNA, a cell's protein-building instructions[4].

These 'RNA-binding proteins' help stabilize, change or destroy RNA, altering how genes are expressed in a cell. Studies have linked mutations in **some RNA-binding proteins to autism.**

ENCODE's latest release details the binding sites of 112 of these proteins. Using these data, the researchers cataloged 52.6 million letters of genetic code that make up these sites.

They then created software that scores variants within these regions based on several lines of evidence, such as whether the variant is found in a highly conserved sequence, suggesting it has

an important function, or the extent to which the variant's sequence alters the protein's binding site. A high score suggests the variant disrupts protein binding.

The software can also be programmed to factor into the score tissue-specific criteria, such as whether a variant lies within a known regulatory element active in the brain.

To test the software, the team compared scores for variants documented in two other datasets, the **Human Gene Mutation Database** and the **1000 Genomes Project**.

As expected, variants with ties to medical conditions had significantly higher scores than those without, the researchers reported. The software also highlighted 720 variants related to medical conditions that other variant-ranking methods did not discover.

The software could help researchers uncover variants linked to RNA-binding proteins with important functions that other approaches have missed, including those involved in autism, the researchers say. It is freely **available online**.

REFERENCES:

1. ENCODE Project Consortium *et al. Nature* **583**, 699-710 (2020) **PubMed**
2. Hon C. and P. Carninci *Nature* **583**, 685-686 (2020) **PubMed**
3. Yue F. *et al. Nature* **515**, 355-364 (2014) **PubMed**
4. Zhang J. *et al. Genome Biol.* **21**, 151 (2020) **PubMed**