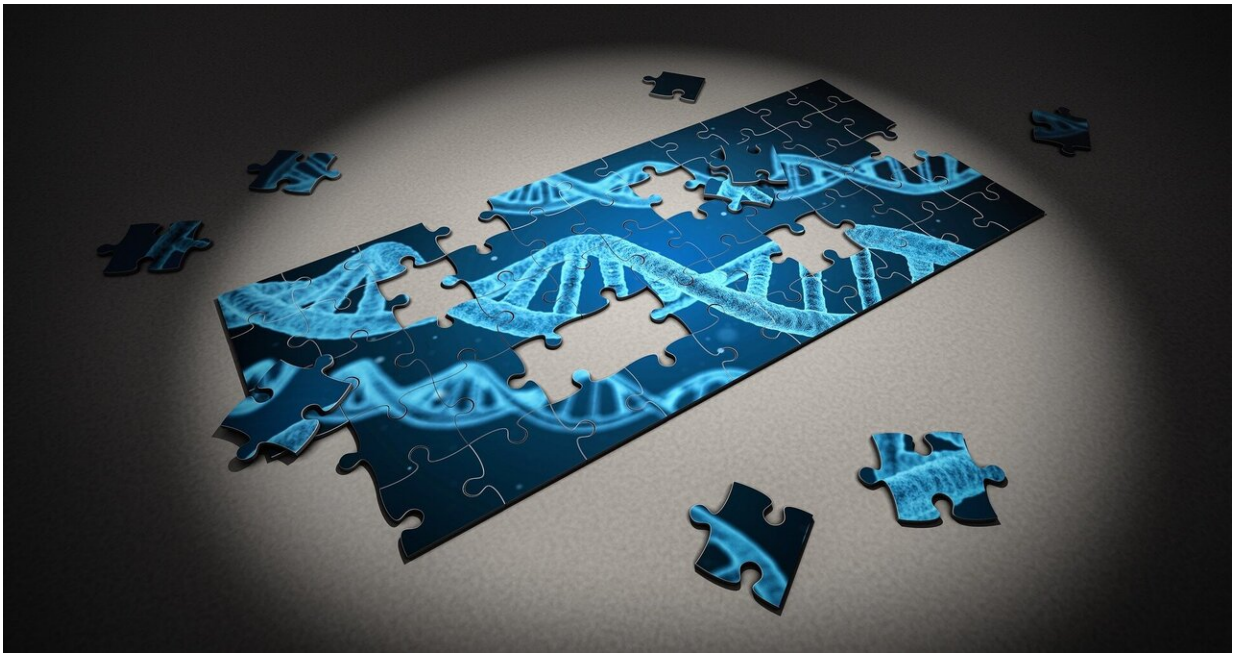
Epigenomic map reveals circuitry of 30,000 human disease regions

February 3 2021



Credit: CC0 Public Domain

Twenty years ago this month, the first draft of the human genome was publicly released. One of the major surprises that came from that project was the revelation that only 1.5 percent of the human genome consists of protein-coding genes.

Over the past two decades, it has become apparent that those noncoding

stretches of DNA, originally thought to be "junk DNA," play critical roles in development and gene regulation. In a new study published today, a team of researchers from MIT has published the most comprehensive map yet of this noncoding DNA.

This map provides in-depth annotation of epigenomic marks—modifications indicating which genes are turned on or off in different types of cells—across 833 tissues and cell types, a significant increase over what has been covered before. The researchers also identified groups of regulatory elements that control specific biological programs, and they uncovered candidate mechanisms of action for about 30,000 genetic variants linked to 540 specific traits.

"What we're delivering is really the circuitry of the human genome. Twenty years later, we not only have the genes, we not only have the noncoding annotations, but we have the modules, the upstream regulators, the downstream targets, the disease variants, and the interpretation of these disease variants," says Manolis Kellis, a professor of computer science, a member of MIT's Computer Science and Artificial Intelligence Laboratory and of the Broad Institute of MIT and Harvard, and the senior author of the new study.

MIT graduate student Carles Boix is the lead author of the paper, which appears today in *Nature*. Other authors of the paper are MIT graduate students Benjamin James and former MIT postdocs Yongjin Park and Wouter Meuleman, who are now principal investigators at the University of British Columbia and the Altius Institute for Biomedical Sciences, respectively. The researchers have made all of their data publicly available for the broader scientific community to use.

Epigenomic control

Layered atop the human genome—the sequence of nucleotides that

makes up the genetic code—is the epigenome. The epigenome consists of chemical marks that help determine which genes are expressed at different times, and in different cells. These marks include histone modifications, DNA methylation, and how accessible a given stretch of DNA is.

"Epigenomics directly reads the marks used by our cells to remember what to turn on and what to turn off in every cell type, and in every tissue of our body. They act as post-it notes, highlighters, and underlining," Kellis says. "Epigenomics allows us to peek at what each cell marked as important in every cell type, and thus understand how the genome actually functions."

Mapping these epigenomic annotations can reveal genetic control elements, and the cell types in which different elements are active. These control elements can be grouped into clusters or modules that function together to control specific biological functions. Some of these elements are enhancers, which are bound by proteins that activate gene expression, while others are repressors that turn genes off.

The new map, EpiMap (Epigenome Integration across Multiple Annotation Projects), builds on and combines data from several large-scale mapping consortia, including ENCODE, Roadmap Epigenomics, and Genomics of Gene Regulation.

The researchers assembled a total of 833 biosamples, representing diverse tissues and cell types, each of which was mapped with a slightly different subset of epigenomic marks, making it difficult to fully integrate data across the multiple consortia. They then filled in the missing datasets, by combining available data for similar marks and biosamples, and used the resulting compendium of 10,000 marks across 833 biosamples to study gene regulation and human disease.

The researchers annotated more than 2 million enhancer sites, covering only 0.8 percent of each biosample, and collectively 13 percent of the genome. They grouped them into 300 modules based on their activity patterns, and linked them to the biological processes they control, the regulators that control them, and the short sequence motifs that mediate this control. The researchers also predicted 3.3 million links between control elements and the genes that they target based on their coordinated activity patterns, representing the most complete circuitry of the human genome to date.

Disease links

Since the final draft of the human genome was completed in 2003, researchers have performed thousands of genome-wide association studies (GWAS), revealing common genetic variants that predispose their carriers to a particular trait or disease.

These studies have yielded about 120,000 variants, but only 7 percent of these are located within protein-coding genes, leaving 93 percent that lie in regions of noncoding DNA.

How noncoding variants act is extremely difficult to resolve, however, for many reasons. First, genetic variants are inherited in blocks, making it difficult to pinpoint causal variants among dozens of variants in each disease-associated region. Moreover, noncoding variants can act at large distances, sometimes millions of nucleotides away, making it difficult to find their target gene of action. They are also extremely dynamic, making it difficult to know which tissue they act in. Lastly, understanding their upstream regulators remains an unsolved problem.

In this study, the researchers were able to address these questions and provide candidate mechanistic insights for more than 30,000 of these noncoding GWAS variants. The researchers found that variants

associated with the same trait tended to be enriched in specific tissues that are biologically relevant to the trait. For example, genetic variants linked to intelligence were found to be in noncoding regions active in the brain, while variants associated with cholesterol level are in regions active in the liver.

The researchers also showed that some traits or diseases are affected by enhancers active in many different tissue types. For example, they found that genetic variants associated with coronary heart disease (CAD) were active in adipose tissue, coronary arteries, and the liver, among many other tissues.

Kellis' lab is now working with diverse collaborators to pursue their leads in specific diseases, guided by these genome-wide predictions. They are profiling heart tissue from patients with coronary artery disease, microglia from Alzheimer's patients, and muscle, adipose, and blood from obesity patients, which are predicted mediators of these disease based on the current paper, and his lab's previous work.

Many other labs are already using the EpiMap data to pursue studies of diverse diseases. "We hope that our predictions will be used broadly in industry and in academia to help elucidate genetic variants and their mechanisms of action, help target therapies to the most promising targets, and help accelerate drug development for many disorders," Kellis says.

More information: Regulatory genomic circuitry of human disease loci by integrative epigenomics, *Nature* (2021). DOI: 10.1038/s41586-020-03145-z , www.nature.com/articles/s41586-020-03145-z

Provided by Massachusetts Institute of Technology