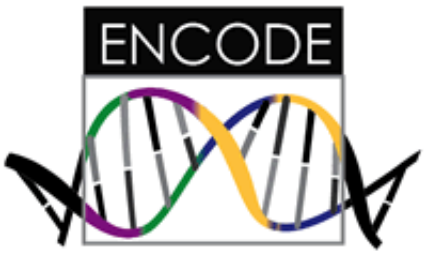
Statistical method will analyze important, poorly studied areas of human genome

October 3 2012



(Medical Xpress)—Each year, more and more pieces of the human genome puzzle fall into place, but large holes still remain. Researchers at the University of Wisconsin-Madison hope to fill in many more pieces with a new $1.1 million grant from the National Human Genome Research Institute (NHGRI).

The grant will support a School of Medicine and Public Health team of researchers who have created new computational tools to analyze important yet poorly studied areas of the human genome.

"We have developed new statistical methods that will help biologists look at the data more easily and effectively," says Dr. Sunduz Keles, associate professor of biostatistics and medical informatics who is one of three principal investigators on the project. "We hope that looking at the human genome data with our methods will contribute significantly to the

human genome puzzle."

Dr. Colin Dewey of the Department of Biostatistics and Medical Informatics and Dr. Emery Bresnick of the Department of Cell and Regenerative Biology are the co-investigators on the grant.

The Human Genome Project, completed in 2003, produced the identity of the entire human genetic code at the most fundamental level - the base. Three billion chemical bases from each parent pair together in a sequence along a twisting DNA ladder.

Only five percent of the material is actual genes; those 23,000 genes are the work horses that make molecules, usually proteins. The rest was initially thought to be useless "junk."

Wanting to understand how such waste could occur in nature, the NHGRI nearly a decade ago launched the Encyclopedia of DNA Elements, or ENCODE, to learn what that 95 percent was all about - particularly, where biological activity might be taking place in it. Last month, in a flurry of papers published in high-profile journals, ENCODE researchers concluded that, in fact, at least 80 percent of the human genome serves some biochemical purpose.

Now, building on the momentum, ENCODE has awarded another round of major grants to examine the data in new and even more rigorous ways to gain a deeper understanding of how the 80 percent affects genes. Keles' group will concentrate on areas of the genome that contain nearly identical repeating segments of base pairs. ENCODE did not include these repetitive areas in its earlier analysis.

Some DNA repeats are short, some are long. Many occur throughout the human genome, some repeating once, others thousands of times.

"We know these repetitive regions are important. They've been linked to inherited diseases, and may be the site of essential biological functions, such as telling genes what to do and when to do it," Keles says. "But repeats present a big challenge because when we observe data from them it is not immediately obvious which repetitive regions are generating the signal."

The UW School of Medicine and Public Health team has found a way around this problem.

"Our approach can statistically tease out where a real signal is coming from in a data sample that has noise from many sources," Keles says.

Researchers typically use a two-step technique called chromatin immunoprecipitation and sequencing (ChIP-seq) to identify sites in DNA samples where proteins such as transcription factors bind to DNA.

ENCODE researchers used ChIP-seq in 1,500 experiments, producing an initial map of hundreds of different proteins that control and regulate genomic activity.

But the standard methods for analyzing ChIP-seq data leave out a significant amount of the data that maps to repetitive DNA, an omission of up to 30 percent of the data that could be utilized. The probabilistic model developed by Keles, Dewey and Bresnick allows them to infer where the real signal is in the data.

The researchers will further enhance their model, recognized as a powerful new tool in the genomics community, and use it to reanalyze ENCODE datasets.

"We already have shown that even proteins that are not particularly known to interact with repetitive DNA can be doing just that," Keles

says.

The methods can also be applied to datasets generated on other National Institutes of Health projects. The end result will be software and other resources for biomedical researchers to leverage existing data and gain new insights into it.

"I think there's a good chance that we will help to significantly reshape the map of players that control and regulate genome activity," she says.

Provided by University of Wisconsin-Madison

Citation: Statistical method will analyze important, poorly studied areas of human genome (2012, October 3) retrieved 27 April 2024 from https://medicalxpress.com/news/2012-10-statistical-method-important-poorly-areas.html

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.