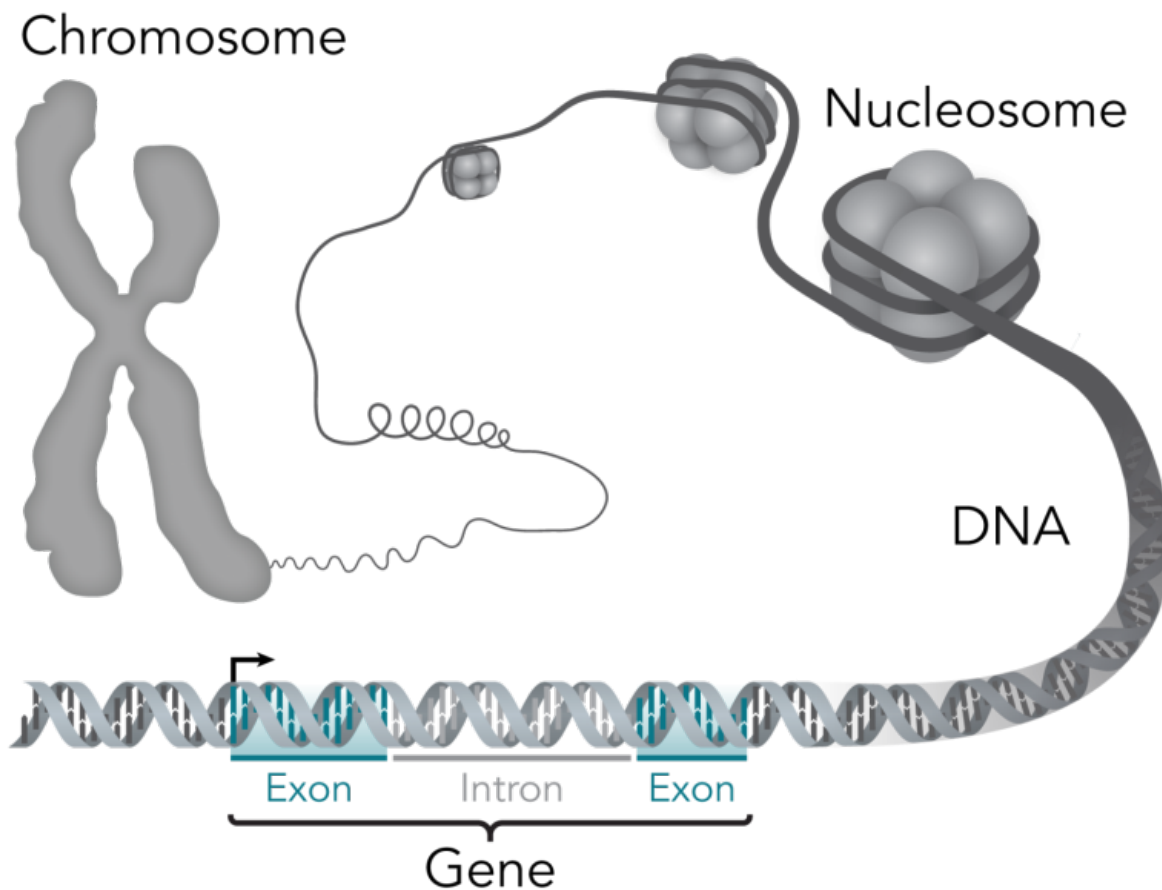# New computational method predicts genes likely to be causal in disease

August 10 2015

This stylistic diagram shows a gene in relation to the double helix structure of DNA and to a chromosome (right). The chromosome is X-shaped because it is dividing. Introns are regions often found in eukaryote genes that are removed in the splicing process (after the DNA is transcribed into RNA): Only the exons encode the protein. The diagram labels a region of only 55 or so bases as a gene. In reality, most genes are hundreds of times longer. Credit: Thomas

A new computational method developed by scientists from the University of Chicago improves the detection of genes that are likely to be causal for complex diseases and biological traits. The method, PrediXcan, estimates gene expression levels across the whole genome - a better measure of biological action than single mutations - and integrates it with genome-wide association study (GWAS) data. PrediXcan has the potential to identify gene targets for therapeutic applications faster and with greater accuracy than traditional methods. It is described online in *Nature Genetics* on Aug 10, 2015.

"PrediXcan tells us which genes are more likely to affect a disease or trait by learning the relationship between genotype, gene expression levels from large-scale transcriptome studies, and disease associations from GWAS studies," said study leader Hae Kyung Im, PhD, research associate (assistant professor) of genetic medicine at the University of Chicago. "This is the first method that accounts for the mechanisms of gene regulation, and can be applied to any heritable disease or phenotype."

Genome-wide association studies are a critical tool in the detection of genes involved in complex diseases such as diabetes and cancer or traits such as height and obesity. GWASs determine these links by identifying single letter DNA variants that appear more frequently in individuals with a disease or trait of interest. However, significant follow-up work is needed to understand the mechanism of action of these variants. Most disease-associated variants are do not alter the function of a gene but instead change the amount of the gene copied in the cells. These studies are unable to determine a causal relationship due to factors such as gene regulation - a genetic variant may instead contribute to altered

expression levels of true causal genes, which remain undetected by a GWAS.

Transcriptome studies such as the National Institute of Health's Genotype-Tissue Expression (GTEx) program aim to overcome this limitation by studying gene expression levels and regulation mechanisms and their relationship with diseases, instead of only DNA sequence. But transcriptome studies also have significant limitations, such an inability to determine reverse causality - whether gene expression levels are altered by disease, or whether disease arises due to altered gene expression.

To develop a method of detecting associations between genes and traits that avoids these issues, Im and her colleagues integrated both transcriptome and GWAS data into a single computation framework, which they named PrediXcan. The method uses computational algorithms to learn how genome sequence influences gene expression, based on large-scale transcriptome datasets such as GTEx. This can then be used to create computational estimates of gene expression levels from any whole genome sequence or chip dataset.

Genomes that have been sequenced as part of a GWAS can be run through PrediXcan to generate a gene expression level profile, which is then analyzed to determine the association between gene expression levels and the disease states or the trait of interest being studied.

The method not only can identify potentially causal genes, it can determine directionality - whether high or low levels of expression might cause the disease or trait. As calculations are based on DNA sequence data and not physical measurements, it can tease apart the genetically determined component of gene expression from the effects of the trait itself (avoiding reverse causality) and other factors such as environment. With PrediXcan, validation studies only need to test a few thousand

genes at most, instead of millions of potential single mutations. In addition, the method can be used to reanalyze existing genomic datasets with a focus on mechanism in a high-throughput manner, addressing a major gap in GWAS studies.

"This integrates what we know about consequences of genetic variation in the transcriptome in order to discover genes, instead of just looking at mutations," Im said. "In a way, we're modeling one mechanism through which genes affect disease or traits, which is the regulation of gene expression level."

While PrediXcan can discover links between genes and traits in a high-throughput manner, Im notes that because it creates estimates based on genome sequence data, it is most accurate for strongly heritable traits. However, almost every complex trait or disease has a genetic component. The method can be used to predict the influence of that component, reducing the complexity of follow-up studies.

Im is now working to improve the prediction of PrediXcan and applying it to mental health disorders. In addition, she is working to expand it beyond gene expression levels, to predict the links between diseases or traits and protein levels, epigenetics and other measurements that can be estimated based on genomic data.

"GWAS studies have been incredibly successful at finding genetic links to disease, but they have been unable to account for mechanism," Im said. "We now have a computational method that allows us to understand the consequences of GWAS studies."

  More information: A gene-based association method for mapping traits using reference transcriptome data, DOI: 10.1038/ng.3367