Untangling our genes

1 message

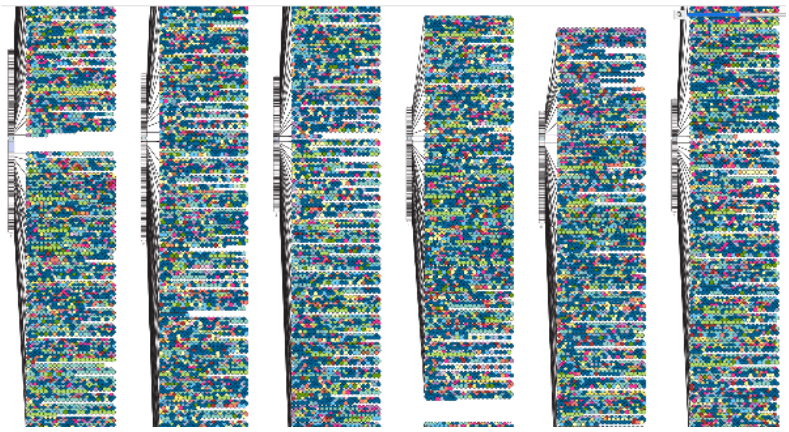
Shelley Edmunds <newsletters@findingpheno.eu>
Reply-To: Shelley Edmunds <newsletters@findingpheno.eu>
To: findingpheno.eu@gmail.com

Fri, May 6, 2022 at 2:59 PM

FindingPheno

This month's email about GWAS analysis, or how we can find which genes are causing a disease or other phenotype. This post is a companion piece to our ML explainer video giving an example of how this technology can be applied. GWAS have been getting larger and larger, going from a few thousand participants to now over a million, and part of what's driving this is the new analytical methods unlocked by machine learning.

This was a lot of fun to write since it calls back to work I was doing in my own PhD back when GWAS were still new, and it really shows how the field has progressed.



Untangling our genes

Genome Wide Association Studies, or GWAS, aim to identify the genetic causes of a specific disease or other outcome. These studies are important because understanding how a disease develops at a molecular level allows us to design better drugs or other treatments for that disease. GWAS are done by sequencing the DNA of individual people, some with the disease and some without, and analysing this sequencing data to see which parts are found more often in the people with the disease – known as a positive association. The genes that occur within these DNA sequences are then identified and undergo functional analysis to figure out how they may be causing the disease that we see.

It is important to look across the entire genome for these studies because most diseases are caused by many interacting genes instead of just one, with each gene coming in multiple versions and each version causing different amounts of risk or protection. On top of this, most diseases are not caused by

genetics alone but also include environment or life style factors in their development. Together this makes for a complicated, messy system with many interconnecting parts and a lot of noise. To deal with this, GWAS have been getting larger, with studies including over 1 million individuals now being published. These large studies provide statistical power to give robust and reproducible associations while also increasing the dynamic range to include more of the rare variants found only in a few individuals. However, more individuals also means more data, increasing the difficulty and computational burden of analysis. Researchers are now turning to Machine Learning (ML) to come up with better strategies for dealing with the very large data sets now being generated.

One key problem with GWAS is figuring out which associations are causative and which ones are not. The way that the DNA sequence varies between people is not random, with some parts more likely to stick together when the DNA is reshuffled during meiosis (linkage disequilibrium). So it may be that a gene with a positive association isn't truly involved in disease risk, but is instead just stuck next to another part of DNA that is really to blame. Alternatively, an associated gene may be involved in disease risk, but its contribution is so small that it has no real effect. In both of these situations a lot of effort is spent trying to understand the function of a gene which is not really relevant to the patient.

This problem can be solved with supervised ML methods, where statistical models are applied to the list of associations to accurately classify them based on if they are causal or not. By using ML, many different models can be developed, combined and tested against the data very, very quickly. These models can also incorporate complex phenotype data, i.e. information about the patient beyond just diseased or healthy, along with known biological information about what genes or pathways are likely to be involved in the disease process. This combination of speed, complexity and processing power allows us to develop much better predictions about the true relevance of each association than is possible from simple statistical modelling without ML.

There are several other ways which ML can be used to improve GWAS data analysis, such as multi-loci feature selection to find novel groups of genes working together, neural networks or deep learning to account for incomplete data, or permutation testing to simplify analysis by removing non-significant data. In addition, genes found by existing ML methods are starting to be used in the clinic. For example, a drug targeting Uromodulin is now being tested in patients for treatment of hypertension after this gene was identified as relevant by the OPEN unbiased ML GWAS approach. ML is fast becoming a standard part of the GWAS analysis pipeline, a trend which is expected to continue.

Read and comment on the full post here.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952914.
