Presentation Structure

João Roque
Genetic Markers for T2D

January 9, 2018

T2D mechanism

In type 2 diabetes, a hormone called insulin cannot adequately control the use of sugar from food which leads to a build up of sugar in the blood. The incidence of diabetes is increasing because of the aging and changing ethnic mix of the population and because of worsening obesity. On the basis of current trends, the prevalence of diabetes is expected to nearly double by 2030 [1]. It is estimated that by 2013, 382 million people in the world lived with diabetes [2]. The causes of type 2 diabetes are complex. This condition results from a combination of genetic and lifestyle factors, some of which have not been identified. This makes it difficult to find specific variations on the genome that implicate risk of diabetes, since it's such a complex disease not yet fully understood.

There are several different types of diabetes, but of those, T2D is the most prevalent. It is also the one which has proven itself to be the hardest to explain genetically. Our definition of diabetes and subgroups is imprecise and thereby makes the identification of genetics causes difficult.

Underlying Genetics

It is fairly well known that human phenotype and traits are mostly defined by single nucleotide polymorphisms. These single nucleotide variations can underlie differences in our susceptibility to disease.

There is compelling evidence that the individual risk for T2D is strongly influenced by genetic and environmental factors [3]. First degree relatives from one carrier of the disease, have up to 40% chance to develop T2D sometime throughout their lifetimes, and 70%, if both of the parents are carriers. This shows a certain degree of heritability, even though we don't entirely understand how this mechanism works for T2D . There seems to be no unique variants that are able to explain it, which only raises more questions [2]. Efforts to identify genetic risk markers have only seen limited success, both in common and rare variants of SNP's [4]. So far, around 153 variants mapping up to 120 loci have shown, only explaining a small portion of the total heritability [2] .

So, with evidence that T2D has at least a genetic component, what should we look for in the genome to identify it's risk? T2D is not our common genetic disease in which a find of a few markers can dictate high penetrance. With rare variants proving themselves not to be the key [4], it is required of recent searches to acknowledge the complexity of DNA, and include some aspects beyond linear correlation of SNP's. We must try and build models that account for epigenetics (DNA methylation), epistasis (gene-gene interaction), both rare, common and protective variants, CNV's and so on [2].

Genome Wide Association Studies & Methods

When scientists started to get their hands on whole human genomes, to try and verify if it was possible to find traits of heritability and correlation for some diseases, genome wide association studies were made. GWAS are typically done by looking at the allele frequency for each SNP, and looking for alleles with higher frequency in the case group than the control. A Manhattan plot can be made to visualize these associations.

Since in more complex diseases, such as T2D, this only seems to explain 5-10% of heritability, models need to take into account the factors explained before (epigenetics (DNA methylation), epistasis (gene-gene interaction), both rare, common and protective variants, CNV's). To try and do this, we can use non-linear machine learning methods such has Random Forests or even deep learning [5]. Despite my best efforts, I couldn't find meaningful literature on deep learning applied to T2D. Nevertheless, the ones who have tried to do it in GWAS have seen drastic improvement in results, which might make it worth it to try out [6]. Deep learning has the advantage of trying to find the best features in SNP's by itself, and being very widely used with good results. However, it is very computationally heavy and makes a complex network that is impossible to explain and interpret in a biological sense.

Before approaching such complex algorithms, it makes sense to approach the problem with a fresh look and find new features to improve simpler models (even though everything points towards non-linear models at least). Trying to include the factors referred before, can also prove itself to be useful.

Challenges

The first challenge that arises is the sheer difficulty of processing and building models that can process such vast quantities of data (there are around 10 million SNP's in the human genome). This vastly increases computing time, so it needs to be very well thought out and considered during the whole project. Even thinking of considering phenomenons such as epistatis need to take this issue into consideration. Secondly, comes trying to integrate gene-environment interaction on the model (if we even want to do it at all), since it's hard to gather data from each patient. Even though there are loads of data for each sample, the datasets needed to study these hypothesis should preferably include people's information (and genome) from different ethnic backgrounds, since it makes it simpler to identify common risk markers. Lastly, there is the problem of phantom heritability. Although studies have discovered <1,200 variants associated with common diseases and traits, these variants typically appear to explain only a minority of the heritability [7].

References

[1] Sandeep Vijan. Type 2 diabetes. *Annals of internal medicine*, 152(5):ITC3–1, 2010.

[2] Rashmi B Prasad and Leif Groop. Genetics of type 2 diabetes—pitfalls and possibilities. *Genes*, 6(1):87–123, 2015.

[3] Mark I McCarthy and Eleftheria Zeggini. Genome-wide association studies in type 2 diabetes. *Current diabetes reports*, 9(2):164–171, 2009.

[4] Christian Fuchsberger, Jason Flannick, Tanya M Teslovich, Anubha Mahajan, Vineeta Agarwala, Kyle J Gaulton, Clement Ma, Pierre Fontanillas, Loukas Moutsianas, Davis J McCarthy, et al. The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47, 2016.

[5] Jason H Moore, Folkert W Asselbergs, and Scott M Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

[6] Suneetha Uppu, Aneesh Krishna, and Raj P Gopalan. Towards deep learning in genome-wide association interaction studies. In *PACIS*, page 20, 2016.

[7] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.