· U C · FCTUC FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE DE COIMBRA

Computational Discovery of Genetic Markers for Type 2 Diabetes

João Roque[1], Conceição Egas[2], Joel P. Arrais[3]

[1]Physics Department and [3]Informatics Engineering Department of the Faculty of Science and Technology of the University of Coimbra, Portugal
[2]UC-BIOTECH

Introduction

This project is being performed in the Laboratory of Neural Networks in the Department of Informatics Engineering, with the guidance of Joel P. Arrais and support of Conceição Egas from UC-BIOTECH.
Type 2 Diabetes is one of the most common diseases encountered in the world, and the fifth leading cause of death worldwide. There is evidence supporting that T2D is strongly influenced by genetic and epigenetic factors, as well as environmental ones[1]. Therefore, with careful examination of genetic variants it is possible to find associations between genotypes and phenotypes. The most commonly used variants are Single Nucleotide Polymorphisms (SNP), which are single nucleotide variations that occur in a specific position in the genome[2]. Considering SNPs, allele frequencies, Linkage Disequilibrium and even other types of variants such as Copy Number Variations as features, machine learning methods can be used to discover this association.

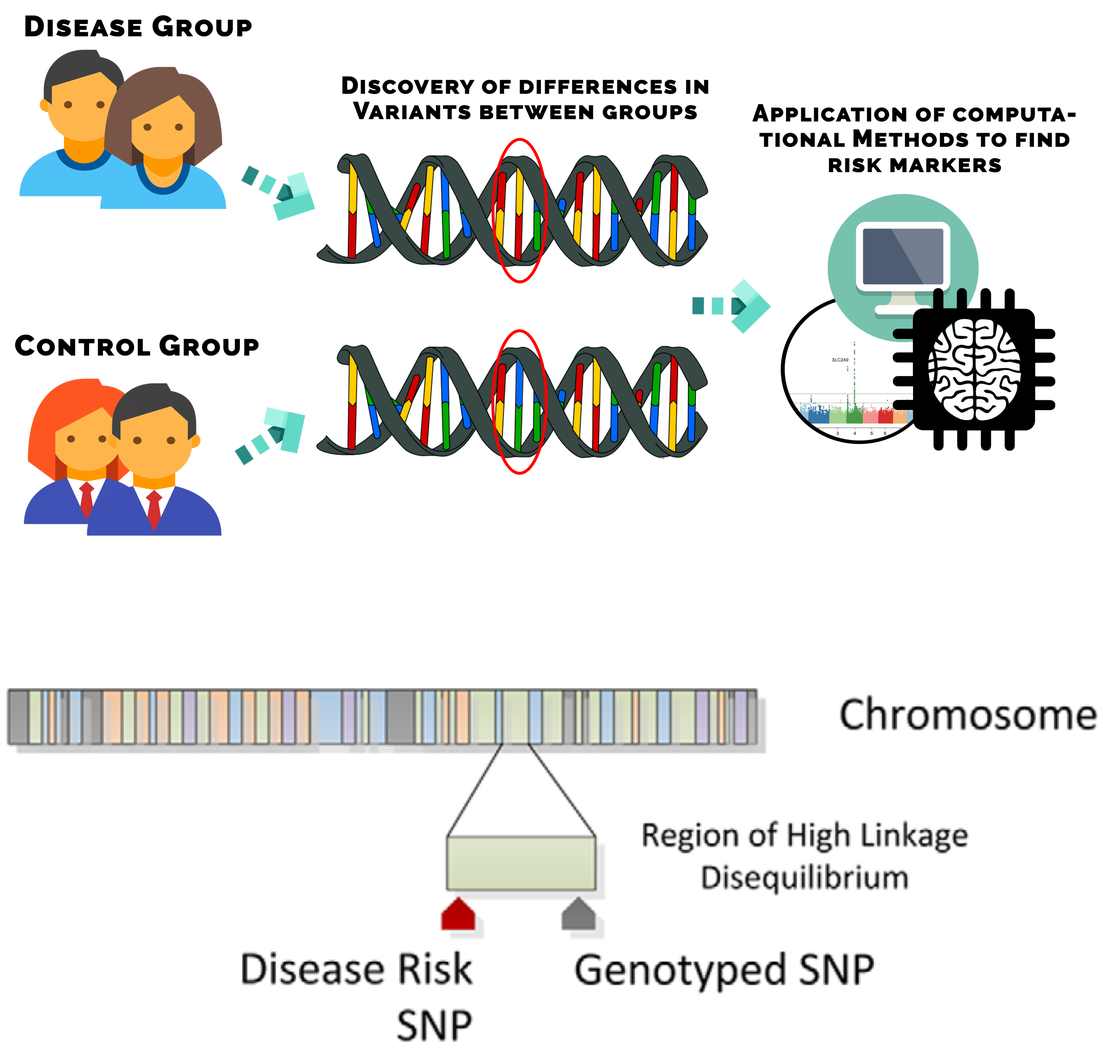

Figure 1: **Top:** Process of a Genome Wide Association Study: from a population, the genome of each person is collected, for control and study groups. Then, computational and statistical methods are applied to find the proper markers with association to the target phenotype.
Bottom: Association between allele frequencies and their effect on phenotype. Mendelian disorders are associated with extremely rare SNPs with large effects, but T2D is mostly associated with common SNPs with small additive effects[3].

So far, for T2D, more than 80 robust markers were found, even though they only account for 20% of the heritability for this disease, which is also known as the missing heritability problem[2].

Objectives

The objective of this work is to apply computational methods to perform a Genome Wide Association Study. The goals are to discover new and relevant genetic markers from Single Nucleotide Polymorphisms datasets for T2D, and to uncover the missing heritability problem that haunts complex diseases since fast sequencing techniques emerged.

Challenges

There are around 10 million SNPs in the human genome that can be associated with numerous traits and diseases. For T2D, the markers discovered are predominantly common (present in great part of the population) with additive effects, which makes their detection among millions of other variants an extremely hard problem. Most datasets contain 200,000 to 2 million SNPs because only high-end sequencers target the whole genome. Missing data needs to be imputed with the 1000 Genome Project[4], and association techniques using Linkage Disequilibrium are required to overcome the missing data problem. Furthermore, these are issues that exist already when considering only a Single-Locus Analysis, which focus on each SNP statistical association with the phenotype. When trying to account for Epistasis or gene-gene interactions, the problem becomes even more so statistically and computationally complex, requiring extensive feature selection and attention to not lose understanding of the biological context of each variant.

Expected Results

It is expected to find a statistical association between the known markers and T2D in the dataset provided, as well as finding novel methodologies to perform Multi-Locus analysis in a feasible time set.

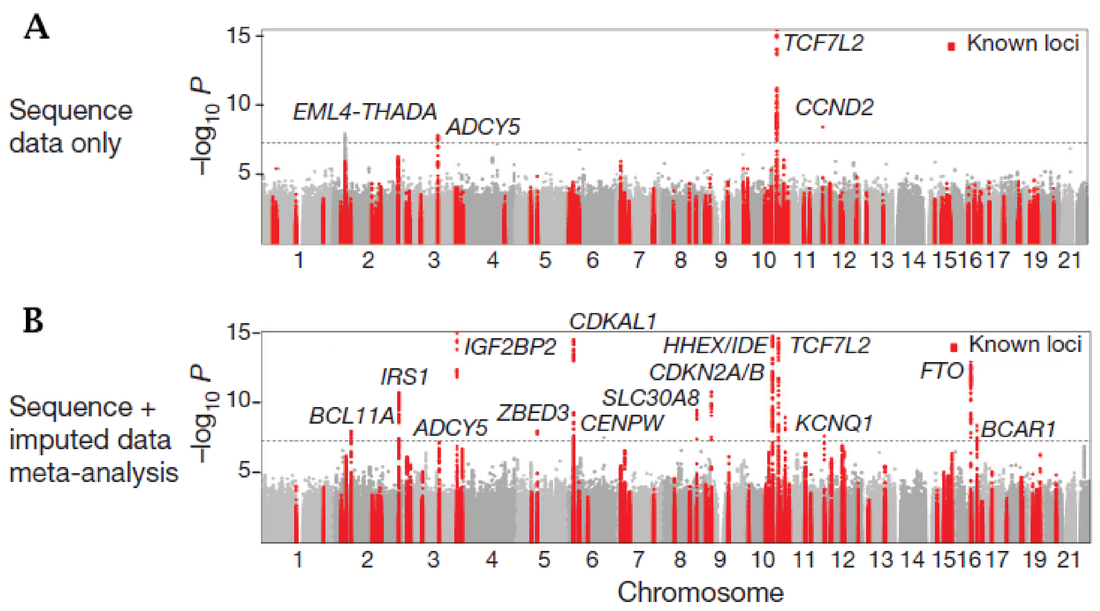

Figure 2: a, b, Manhattan plots of single-variant association analyses for: sequence data alone (a, 1,326 cases and 1,331 controls) and meta-analysis of sequence and imputed data (b, total of 14,297 cases and 32,774 controls). 1000G, the 1000 Genomes Project data[2, 4].

Validation of new methods that predict T2D risk are performed with different public datasets. These will be of European populations, such as the original one. Besides these, any new markers discovered will be extensively studied and validated, to add to the explanation of missing heritability.

Acknowledgements

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

financed by national funding via the Foundation for Science and Technology and by the European Regional Development Fund (FEDER), through the COMPETE 2020 – Operational Program for Competitiveness and Internationalization (POCI)

Cofinanciado por:
COMPETE 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional

References

[1] R. B. Prasad and L. Groop, "Genetics of type 2 diabetes—pitfalls and possibilities," *Genes*, vol. 6, no. 1, pp. 87–123, 2015.

[2] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, *et al.*, "The genetic architecture of type 2 diabetes," *Nature*, vol. 536, no. 7614, pp. 41–47, 2016.

[3] W. S. Bush and J. H. Moore, "Genome-wide association studies," *PLoS computational biology*, vol. 8, no. 12, p. e1002822, 2012.

[4] . G. P. Consortium *et al.*, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, p. 56, 2012.