UNIVERSITÀ DEGLI STUDI DI TORINO
SCUOLA DI MEDICINA
DIPARTIMENTO DI BIOTECNOLOGIE MOLECOLARI E SCIENZE PER LA SALUTE
CORSO DI LAUREA IN BIOTECNOLOGIE

# Transcriptome-Wide Association Studies: Bridging the Gap between Genome, Transcriptome and Disease

*Supervisor: Prof. Paolo Provero*

*Candidate: Federico Marotta*

TESI DI LAUREA, 18 LUGLIO 2018

# *Contents*

# *List of Figures*

# *List of Tables*

# *Todo list*

*'The harmony of the world is made manifest in Form and Number, and the heart and soul and all the poetry of Natural Philosophy are embodied in the concept of mathematical beauty.'*
—*D'Arcy Wentworth Thompson,* On Growth and Form

*I would like to thank my supervisor, Prof. Paolo Provero, and all the former and current members of the Computational Biology Unit whom I have met: Elisa Mariella, Davide Marnetto, Elena Grassi, Ugo Ala, Alessandro Lussana, Stefano Gilotto. Their help has often been invaluable and they have never failed to provide a stimulating environment for me to work in.*

# *Abstract*

Understanding how genetic variation among individuals can influence the manifestation of complex diseases, which stem from the interaction of many genes with each other and with the environment, is a relevant problem in medicine. Since the first genome-wide association study (or GWAS) was conducted in 2005, tens of thousands of SNP-trait associations have been reported, shedding light on the at least partly genetic roots of many diseases; most of such associations, however, do not provide much predictive value and are difficult to explain. Expression quantitative trait loci (eQTL) mapping, which identifies loci that influence gene expression, is a possible step towards a better understanding of the relationship between genetic variation and phenotypic trait, using gene expression as a proxy for the trait. Recently, a new approach has been devised which goes one step further and aims to directly find associations between the expression of each gene and a given trait by combining GWAS and eQTL data. In one of their versions, these 'transcriptome-wide association studies' (or TWAS) are performed in two phases, the first being the prediction of the genetic component of gene expression of the individuals in a GWAS cohort using reference transcriptome data, and the second being the evaluation of the association between predicted expression and trait in those individuals. On the whole, TWAS are powerful statistical methods to find associations between gene expression and complex phenotypic traits; while they can help in making sense of GWAS results, they also can find novel associations, pointing at potential candidate genes for further analysis: as such, their contribute to the characterisation of the relationship between genome and phenotype is substantial. After an introduction, the focus of the first part of this thesis will be a method to leverage individual-level data in order to detect genes associated with disease traits. The second part shall deal with how, conveniently, a TWAS can be performed starting only from the summary association statistics and the summary LD information of a GWAS. In the third part, we will discuss the advantages of integrating epigenetic markers in a TWA study and see an application to schizophrenia.

# *Integrative approaches for large-scale transcriptome-wide association studies, or: A test for significant cis genetic correlation between expression and traits*


## Abstract

Many genetic variants influence complex traits by modulating gene expression, thus altering the abundance of one or multiple proteins. Here we introduce a powerful strategy that integrates gene expression measurements with summary association statistics from large-scale genome-wide association studies (GWAS) to identify genes whose cis-regulated expression is associated with complex traits. We leverage expression imputation from genetic data to perform a transcriptome-wide association study (TWAS) to identify significant expression-trait associations. We applied our approaches to expression data from blood and adipose tissue measured in ~3,000 individuals overall. We imputed gene expression into GWAS data from over 900,000 phenotype measurements to identify 69 new genes significantly associated with obesity-related traits (BMI, lipids and height). Many of these genes are associated with relevant phenotypes in the Hybrid Mouse Diversity Panel. Our results showcase the power of integrating genotype, gene expression and phenotype to gain insights into the genetic basis of complex traits.


## *Introduction*

The *rationale* that lies behind the association of gene expression to phenotype is that many genetic variants influence traits by altering the regulation of the expression of some genes. Despite the strength of this argument, publications of studies in which both transcriptomic and phenotypic data are investigated simultaneously lag behind those of simple GWAS studies, for at least two reasons: first, although the cost of sequencing nucleic acids has been sharply decreasing for over a decade (Figure 1), it can become quite an expensive technology if applied to cohorts of tens of thousand samples, such as those of a typical modern GWAS; secondly, every tissue shows a different pattern of expressed genes, and to choose the right tissue to analyse for each phenotype is not always a trivial matter.

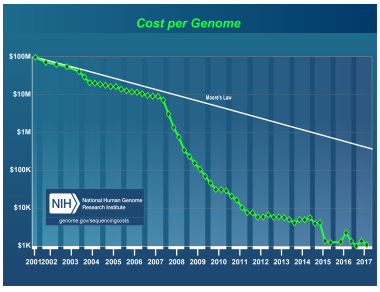



Figure 1: The decrease in the cost of genome sequencing; the same technology is used to sequence RNA. https://www.genome.gov/sequencingcosts/

In order to harness the plethora of data available from existing large-cohort GWAS studies, which, due to their great sample size, have the statistical power to find association even for rare and

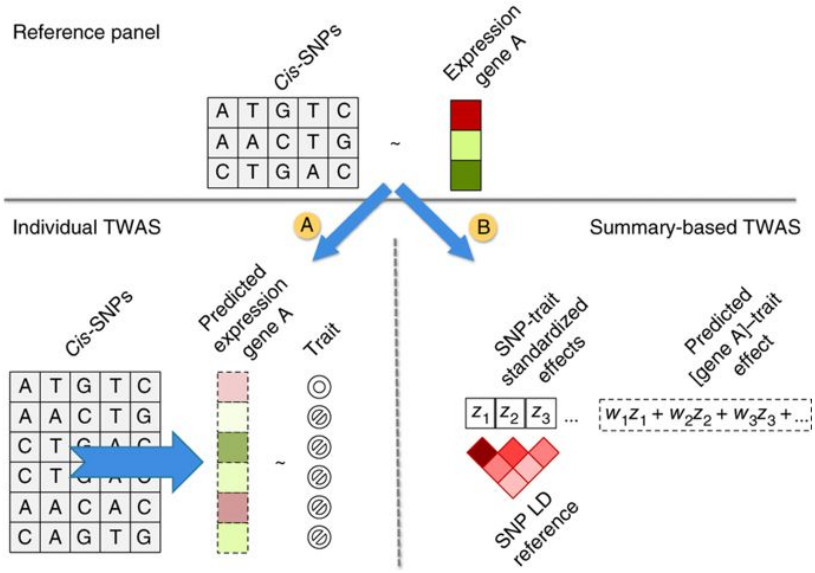
small-effect variants, many new methods are being developed. One of such methods is PrediXcan, with which we dealt in the previous section, but it is by no means the only one. In particular, in 2016 a new approach has been proposed which does not need individual-level data, but only summary association statistics[1] from a GWAS, which is an important advantage since, normally, only summary-level data is publicly available due to privacy concerns.

[1] By summary association statistics we mean, for instance, the effect size of all the SNPs

In essence, this approach is not different from PrediXcan: first, a linear regression model finds the correlation between each SNP and gene expression and accordingly assigns a weight to the SNP ; next, the SNPs weights are used to impute the *cis* genetic component of expression; finally, the imputed gene expression is tested for an association with a complex trait.

taking into account LD: gamazon used elastic net to prune correlated predictors

through correlation




Nevertheless, there are some relevant points in this new method, relative to PrediXcan: its being based on summary association statistics greatly increases the effective sample size, because the method can in principle be applied to any GWA study; moreover, the authors emphasise the robustness of their approach, for its focus is on the genetic component of expression only, therefore it is guaranteed that the association between expression and trait is ultimately due to genetic factors.

it doesn't mean it is robust

Indeed, there are several ways in which genomic variation can be related to gene expression and phenotypic variation.

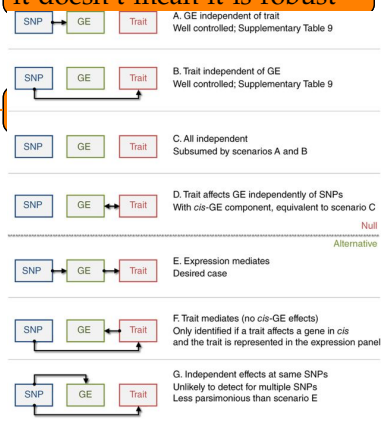



Figure 2:

'We applied our approaches to expression data from blood and adipose tissue measured in ~3,000 individuals overall. Through extensive simulations and analyses of real data, we show that our proposed approach increases performance over standard GWAS and eQTL-guided GWAS. Furthermore, we reanalyzed a 2010

lipids GWAS17 to find 25 new expression-trait associations in those data. Among these associations, 19 of 25 contained genome-wide significant SNPs in the more recent and expanded lipids study5, thus showcasing the power of our approach to find robust associations. We imputed gene expression into GWAS data from over 900,000 phenotype measurements5,6,7 to identify 69 new genes significantly associated with obesity-related traits (body mass index (BMI), lipids and height). Many of these genes were associated with relevant phenotypes in the Hybrid Mouse Diversity Panel (HMDP). Overall, our results showcase the power of integrating genotype, gene expression and phenotype to gain insights into the genetic basis of complex traits.'

# *Complete Bibliography*


[1] R. C. Lewontin. 'THE INTERACTION OF SELECTION AND LINKAGE. I. GENERAL CONSIDERATIONS; HETEROTIC MODELS'. In: *Genetics* 49.1 (Jan. 1964), pp. 49–67. ISSN: 0016-6731. URL: http://www.genetics.org/content/49/1/49.article-info.

[2] W. G. Hill and Alan Robertson. 'Linkage disequilibrium in finite populations'. In: *Theor. Appl. Genet.* 38.6 (June 1968), pp. 226–231. ISSN: 1432-2242. DOI: 10.1007/BF01245622.

[3] *LinkageDisequilibrium.pdf*. [Online; accessed 6. May 2018]. Oct. 2017. URL: http://www.handsongenetics.com/PIFFLE/LinkageDisequilibrium.pdf.

[4] Peter JP Croucher. *Linkage Disequilibrium*. [Online; accessed 6. May 2018]. Apr. 2013. DOI: 10.1002/9780470015902.a0005427.pub3.