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*

*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.

# *Introduction*

## *The old nature-nurture debate*

Both genes and environment play a crucial role during the entire life of an organism, from development to senility, and in particular in the manifestation of complex diseases. Every phenotype arises as a consequence of the interactions of genes both with each other and with the environment. Other things being equal, genetic variation among individuals results in phenotypic variation; on the other hand, environment can influence a trait as much as any gene, especially if the trait is *complex*.

A complex trait, as opposed to a mendelian trait, is a phenotype whose variation in the population cannot be explained by the variation in a single gene. For instance, height is a complex trait as there have been found many loci contributing to it: each of them gives a small contribution, and the final height depends on the combination of all the alleles in the genome of the individual. Clearly, not *all* phenotypic variation in a population can be explained by genetic variation, for the environment has an effect as well. The proportion of phenotipic variance that can be explained by genetic variance is called heritability (see Section *Heritability of gene expression* on page 25 for an in-depth discussion), which for height is about 80%.[1] It cannot be said that an allele *determines* a phenotype, but rather variation at that locus can result in phenotypic variation, under the influence of an appropriate environment.

[1] Visscher, 'Sizing up human height variation'.

Initially, research concerning the genetic basis of diseases relied on two approaches: linkage studies, which searched for loci that correlated with a disease in a family; and candidate-gene association studies, which explored genetic variation in a limited region of a chromosome. Gradually, it happened that the focus moved from candidate genes to whole genomes, and from families to populations. As a consequence of the large amount of genetic variation among individuals, larger sample sizes were needed to study variation at a population level, thus many laboratories decided to join their forces and create consortia.

One of the driving ideas of the Human Genome Project,[2] indeed, was that the knowledge of the human genome sequence and its annotation would have helped to explain and cure diseases. After all, the fact that DNA is an inherited molecule strongly suggests that heritable traits must be related to it. In 2005, a powerful method was developed in order to harness the huge amount of data collected from the sequencing of genomes: genome-wide association studies (see Section *GWAS* on page 18 for a technical description of the method), which find associations between genetic variants and phenotypic traits, in the sense that people harbouring a particular allele might be more liable to develop a particular phenotype. Such liability can be quantified by an odds ratio or effect size.

[2] Lander et al., 'Initial sequencing and analysis of the human genome'; Venter and al., 'The sequence of the human genome'.

Another method to investigate complex traits is the mapping of loci that influence gene expression (see Section *eQTL mapping* on page 18). Gene expression is an intermediate phenotype in the sense that it is one of the mechanistic steps that bring from a gene to an accomplished phenotype, therefore expression can be used as a proxy for the phenotype. In eQTL mapping, each SNP is assigned a coefficient according to how much the expression of a gene is altered by the presence of that SNP.

Environment can influence gene expression, and in fact also gene sequence —for instance, exposure to UV rays can cause mutations. However, since the effects of environment are difficult to quantify, much of the reaserch on complex traits and diseases was concerned on genetic factors. Here, our focus will be primarily on SNPs as genetic variants, and on humans as organisms of interest, with particular reference to their diseases.

## *Genome-wide association studies and expression quantitative trait loci mapping*

The classical method to find associations between SNP and disease is the GWAS: several individuals in a cohort of cases and controls are genotyped, and the variants that occur more frequently in cases than in controls are said to be associated with the disease. Unless the full genomes of the samples are sequenced, however, not every single variant in the population will be known, and in fact those that result associated to the disease are almost never the causal ones, but are only in linkage disequilibrium with the unknown causal SNPs.[3] Moreover, even if the causal variant were known, the underlying biological mechanism of the disease would still be obscure, although it is possible to leverage a functional annotation of the genome to draw conclusions[4].

As of 2018-06-25, the GWAS Catalog contains 3420 publications and 62652 unique SNP-trait associations. https://www.ebi.ac.uk/gwas/home

[3] Visscher, Brown, et al., 'Five years of GWAS discovery'.

[4] For instance, most disease-associated SNPs fall in enhancers, as reported in a paper by Ernst published in 2011, using data from the ENCODE project.

Despite these limitations, GWAS have revealed some interesting facts.

First and foremost, complex traits are highly polygenic:[5] there are many loci which, together, can carry a number of combination of alleles, each of which increases or decreases the probability of disease by a small amount. One of the hypothesis that led scientists to investigate the genome searching for relationship with diseases, was that of 'common disease, common variant'. Actually, however, there can be few rare variants that contribute substantially to the disease risk, as well as many common variants with a small effect size. The first GWAS, published in 2005, found an association between a SNP falling inside an intron of the complement factor H (*CFH*) gene and age-related macular degeneration;[6] such polimorfic allele was in linkage disequilibrium with another allele causing a T402H mutation in the resulting protein. Since then, many disease-associated variants have been found, but the majority of them fall in non-coding regions, making their interpretation very difficult. Figure 1 reports all the SNPs associated to a disease to date, coloured by disease.

[5] Visscher, Wray, et al., '10 Years of GWAS Discovery: Biology, Function, and Translation'.

[6] Klein et al., 'Complement factor H polymorphism in age-related macular degeneration.'

Figure 1: Some of the most notable successes of GWAS have been reported for Crohn's disease, schizophrenia, type 2 diabetes, and cardiovascular traits and diseases. The figure reports all the association found as of 2018-06-25, coloured by the disease to which the SNP is associated.

The genomes of closely related species do not exhibit extreme differences, even though the species are far apart at the phenotypic level; the majority of differences are found in non coding regions[7], suggesting that much of the phenotypic differences are due to differences in how genes are regulated. It seems to be even more so for variations among individuals of the same species. The mapping of eQTL contributed to explain how expression levels differ in different individuals, and revealed that gene expression, as a phenotype, is predominantly regulated in *cis*, is quite heritable and differs in different populations.[8]

[7] For instance, 80% of human genes are shared with the mouse, and 40% of nucleotides are identical. Emes et al. ('Comparison of the genomes of human and mouse lays the foundation of genome zoology')

[8] Gilad, Rifkin, and Pritchard, 'Revealing the architecture of gene regulation: the promise of eQTL studies'.

In general, genetic variants can influence a phenotype in at least three ways (Figure 2 on the following page): chromatin modification, splicing sites alteration, and change of expression levels through direct mechanisms.[9] Most of the times, this three proposed mechanism lead

[9] Li et al., 'RNA splicing is a primary link between genetic variation and disease'.

to an alteration of gene expression, and it has been shown that disease-associated variants often are linked to eQTL,[10] suggesting that gene expression is one of the most important intermediate links in the chain of events leading from genome to phenotype.

[10] Nicolae et al., 'Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS'.

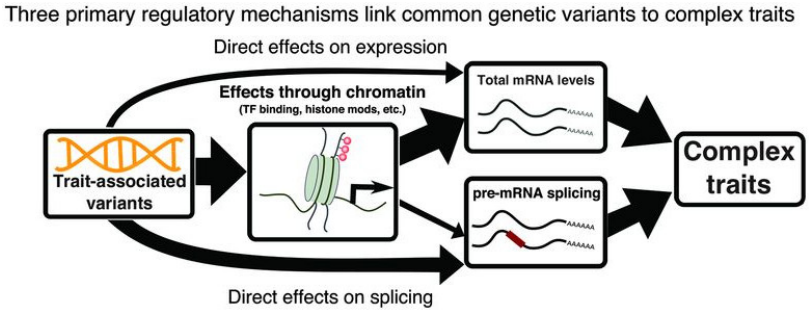


Figure 2: The main effects of regulatory eQTL. Many GWAS hits influence disease in one of these ways.

## *Limitations of GWAS and eQTL mapping*

The GWAS has been an important method in finding genetic variants associated to disease. Its integration with eQTL and functional annotation (mostly suggested by chromatin activity) of DNA, can provide and has provided valuable knowledge both for science and for clinical applications. Nevertheless, most variants are characterised by a small effect size and are not able to explain fully the heritability.[11] For example, height has an heritability of 80%, but the about 40 loci associated with this trait in 2009, after a series of GWA studies involving tens of thousands of people, only explained about 5% of the total phenotypic variance.

[11] Manolio et al., 'Finding the missing heritability of complex diseases'.

Some of the causes proposed to explain this 'missing heritability' are the following.[12]

[12] Ibid.

- Natural selection tends reduce the number of alleles conferring a strong disease risk, and the remaining alleles of small effect size are difficult to detect without very large sample sizes.
- Rare variants of large effect may be present, but genotyping arrays are not fully able to detect rare variants, and sequencing the genomes of tens of thousand of people is infeasible.
- Phenotypic variation might be explained by types of variation different from single nucleotide polimorfisms, like CNV, but genotyping arrays focus on SNPs.
- The effects may be due to combinations of alleles which behave in a non-additive fashion; this can happen if they occur at the same locus (dominance) or at different loci (epistasis).

Another limitation of GWAS and eQTL is that, although most hits fall in regulatory regions and are known to have an effect on gene expression, these studies do not explain how the variants make one individual more liable to a disease.

By exploiting the fact that most trait-associated variants influence gene expression, transcriptome-wide association studies aim to find correlations between expression and traits. Gene expression lies at an higher level than DNA, therefore variation in expression levels are the result of the combination of many genetic variants; TWAS naturally take account of this and aggregate the effects of many SNPs into an intermediate phenotype, which then is correlated to a yet higher level phenotype, such as a disease.

Thus, while GWAS and eQTL mapping find associations between genetic variants and phenotype or gene expression, respectively, TWAS find associations between gene expression and phenotypes. On the one hand, this reduces the statistical tests to perform; on the other hand, it increases the interpretability of results. Indeed, it is easier to understand the function of a gene rather than of a nucleotide.

Despite the strength of these arguments, 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 3), 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.

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 (Chapter *A gene-based association method for mapping traits* on page 21 ), which solves the problem of the missing expression in a GWAS by imputing the expression of an individual on the basis of its genotype. Since the genotype of the individuals who took part in an association study are often unknown due to privacy or logistic issues, some authors focused on an approach based on summary-level statistics, which are widely available (Chapter *Integrative approaches for large-scale transcriptome-wide association studies* on page 31 ). Finally, because an association does not imply a causal mechanism, methods to understand the underlying biology between the correlation were developed, such as that discussed in Chapter *Transcriptome-wide association study of schizophrenia and chromatin activity* on page 39 .

Besides, a sort of dilemma arises: to detect rare variants, whole genome sequencing of a huge number of people is necessary. At the same time, the differences among individuals revealed by WGS are so many that trying to associate them with anything can become statistically very dificult due to allelic heterogeneity. At each generation, some 40 *de novo* mutations per genome are introduced in the population, since the mutation rate for humans is about $1.2 \cdot 10^{-8}$ nucleotides per genome per generation. Kong et al. ('Rate of de novo mutations and the importance of father-s age to disease risk')

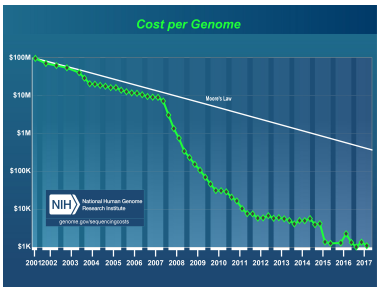


Figure 3: The decrease in the cost of genome sequencing; the same technology is used to sequence RNA. https://www.genome.gov/sequencingcosts/

# *Methods*

In this section I shall describe some general methods which are widely used in quantitative genetics and bioinformatics, and were also employed in the articles described in the thesis.

## *Regression*

Regression is used to find relationship between data. It usually consists of three-steps:

1. Assumption-making, where one chooses the type of relationship (*e.g.* linear, logistic, polynomial...).
2. Fitting of the model, where the parameters of the model are evaluated on a training data set.
3. Prediction of the response, where known data from a testing data set are fed to the model, which returns an estimation of the outcome.

Linear regression[13] models a linear relationship between a continuous variable, $Y$, and one or more other variables, the $X$'s, which may be continuous or categorical. In other words, $Y$ can be expressed as

[13] James et al., *An Introduction to Statistical Learning*.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \,, \tag{1}$$

where the $\beta$'s are called the model coefficients; the approximation sign is due to random errors, which causes the $Y$ to differ from the right-hand side by a term $\epsilon$ —a random residual error. In simple linear regression, at first the model is fitted, *i.e.* the values of the coefficients that better describe the relationship between $X$ and $Y$ are found relying on a training dataset, and the model is applied to a testing dataset with known $X$'s to make predictions of $Y$. The fitted

model, where the parameters are estimated, is usually represented as follows, with the parameters denoted by an hat:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j \,. \tag{2}$$

The estimation of the coefficients is often made by the minimisation of the residual sum of squares, which is defined as $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, *i.e.* $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j x_j))^2$.

The coefficients are obtained by using some calculus to find the partial derivatives of the RSS function with respect to all the $\hat{\beta}$, and then some algebra to solve a $p$th-order linear system where we impose such derivatives equal to 0.

When $p = 2$, the partial derivatives are

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)$$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} -2x_i(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0)$$

and the system yields

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Once we have the coefficients, given the $x$'s, we could estimate an $\hat{y}$.

RIDGE REGRESSION[14], which is similar to linear regression, is a regularisation method allowing to reduce the dependence of the fitting on the training set of values by shrinking the coefficients towards zero. This is achieved with a slight modification of the least squares, that is, the function to minimise is

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\beta_j^2 \,. \tag{3}$$

This function is the sum of the RSS and a penalty term, the minimisation of which can improve the fitting of the model, provided that the tuning parameter $\lambda$ is properly chosen.

[14] James et al., *An Introduction to Statistical Learning*.

THE LASSO[15] is another regularisation method which allows to effectively select a subset of relevant predictors by setting the coefficients of the others to zero. In this case, the minimisation function is

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right| \,. \tag{4}$$

[15] Tibshirani, 'Regression Selection and Shrinkage via the Lasso'.

ELASTIC NET[16] combines the best features of ridge regression and

[16] Zou and Hastie, 'Regularization and variable selection via the elastic net'.

lasso by introducing two penalty terms:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\alpha\sum_{j=1}^{p}(\beta_j)^2+(1-\alpha)\sum_{j=1}^{p}|\beta_j|\,, \qquad (5)$$

with $\alpha$ between 0 and 1. Its main advantage is that it works well when $p > n$, for it can potentially group correlated predictors and select only one representative variable for each group.

Elastic net is used in the paper by Gamazon *et al.* (Chapter *A gene-based association method for mapping traits* on page 21) to predict the genetically regulated component of gene expression.

Logistic regression[17], which applies when the predicted outcome is a binary variable indicating whether the response falls into one of two categories, models the probability that the variable belongs to a particular category.

[17] James et al., *An Introduction to Statistical Learning*.

While for linear regression we assumed an equation of the form of a straight line (or a multi-dimensional equivalent), for logistic regression we need a function that returns values between 0 and 1, thus we rely on the logistic function (Figure 4):

$$Y = P(Z = 1|X) = \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\,. \qquad (6)$$

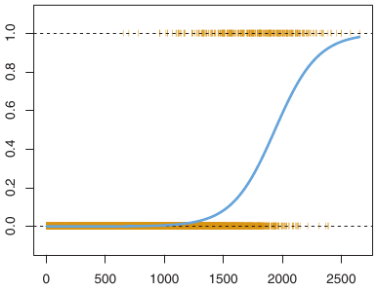


Figure 4: An example of logistic model. Adapted from James et al. (*An Introduction to Statistical Learning*)

In order to fit this model we switch to the 'estimated' values, denoted by an hat, and manipulate the logistic function until we have

$$\log\left(\frac{\hat{y}}{1-\hat{y}}\right) = \hat{\beta}_0+\hat{\beta}_1 x\,, \qquad (7)$$

where the left-hand member is the logarithm of the odds, or logit. At this point the coefficients can be found with the maximum likelihood method, and the predictions be made. As with linear regression, this model can be easily extended to include more than one predictor.

Logistic regression is used in Gamazon *et al.* to find the probability of disease given the expression level of a gene.

Linear mixed models are useful when data have an hierarchical structure, *i.e.* there can be identified many groups and there is variability both inside and among groups. Such models cobine fixed and random effects as predictors of the outcome. The relationship

between the predictors and the outcome is as follows:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + u_0 + \sum_{j=1}^{q} u_j Z_j + \epsilon \tag{8}$$

The $X$'s are the predictor variables, with the $\beta$'s as their coefficients; the $uZ$'s are the random effects, random because the $u$'s are assumed to be normally distributed; and $\epsilon$ is the residual random error.

A BAYESIAN SPARSE LINEAR MIXED MODEL[18] is a linear mixed model because the outcome depends on both fixed and random effects, is sparse because it selects only a subset of all the predictors (like the lasso or elastic net), and is Bayesian because the predictors are selected using a Bayesian approach.

[18] Zhou, Carbonetto, and Stephens, 'Polygenic Modeling with Bayesian Sparse Linear Mixed Models'.

One of these models is used in both papers by Gusev *et al.* (Chapter *Integrative approaches for large-scale transcriptome-wide association studies* on page 31 , Chapter *Transcriptome-wide association study of schizophrenia and chromatin activity* on page 39 ) to predict the expression levels of genes in a reference transcriptome data set.

## *eQTL mapping*

There are many ways to perfom an eQTL analysis,[19] but the underlying idea is that of considering gene expression as a quantitative trait and finding associations between a genetic variant and this phenotype. The association, which can exist both for *cis* and for *trans* variants, can be found either through correlating genotype and expression, or by using a linear regression model for each gene-marker pair between the number of alleles that individuals have at the marker locus and the expression of the gene.

[19] Gilad, Rifkin, and Pritchard, 'Revealing the architecture of gene regulation: the promise of eQTL studies'.

## *GWAS*

To perform a genome-wide association study,[20] two types of data about the individuals are needed: genetic markers and phenotypes. For each individual, its two alleles at each of the loci under study are reported; such loci are called markers, because they mark a specific position on a chromosome. Phenotypic data are needed to split the population in cases and controls, and possibly to detect and correct false associations: indeed, differences in the allele frequencies of cases and controls could be due to differences in sex, population stratification or other conditions that differ from cases to controls. Since

[20] Clarke et al., 'Basic statistical analysis in genetic case-control studies'.

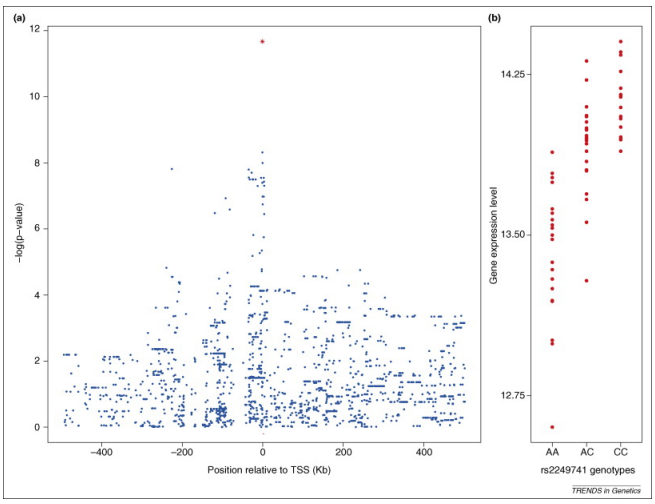


Figure 5: Typical eQTL results.

only a few loci are typed in a normal GWAS, it is often performed an imputation step, where SNPs at other loci are estimated on the basis of their LD structure.

After making quality controls and corrections for causes of variance not due to the phenotype of interest (*i.e.*, covariates, which are a possible source of confounding), the frequency of each allele at each marker in cases and controls is computed. At this point, there are a few choices about how to perform the test of association. The simplest model is based on allele counts.

First of all, a contingency table like that in Table 1 is built.

| Allele | a | A | Total |
|---|---|---|---|
| Cases | $m_{11}$ | $m_{12}$ | $m_{1\cdot}$ |
| Controls | $m_{21}$ | $m_{22}$ | $m_{2\cdot}$ |
| Total | $m_{\cdot 1}$ | $m_{\cdot 2}$ | $2n$ |

Table 1: Contingency table for an allelic model of association.

The odds ratio for allele A is estimated by

$$OR_A = \frac{\frac{m_{12}}{m_{11}}}{\frac{m_{22}}{m_{21}}} = \frac{m_{12}m_{21}}{m_{22}m_{11}} \tag{9}$$

And the association test is actually a $\chi^2$ test of independence of rows and columns, the significance threshold of which should be corrected for the multiple testing performed at each marker.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(m_{ij} - E[m_{ij}])^2}{E[m_{ij}]} \tag{10}$$

# *A gene-based association method for mapping traits using reference transcriptome data*

*Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox & Hae Kyung Im; Nature Genetics 2015*

*Abstract*

Genome-wide association studies (GWAS) have identified thousands of variants robustly associated with complex traits. However, the biological mechanisms underlying these associations are, in general, not well understood. We propose a gene-based association method called PrediXcan that directly tests the molecular mechanisms through which genetic variation affects phenotype. The approach estimates the component of gene expression determined by an individual's genetic profile and correlates 'imputed' gene expression with the phenotype under investigation to identify genes involved in the etiology of the phenotype. Genetically regulated gene expression is estimated using whole-genome tissue-dependent prediction models trained with reference transcriptome data sets. PrediXcan enjoys the benefits of gene-based approaches such as reduced multiple-testing burden and a principled approach to the design of follow-up experiments. Our results demonstrate that PrediXcan can detect known and new genes associated with disease traits and provide insights into the mechanism of these associations.

## *Introduction*

Albeit it is accepted that in the majority of cases the biological role of variants associated to diseases is regulatory, as confirmed by the fact that many such variants are linked to eQTL and fall in regions that are epigenetically marked as regulatory, GWAS results remain mainly uncharacterised from a functional point of view, and are only able to explain a little proportion of phenotypic variance. The wealth of biological data that is now being released by large-scale consortia provides an unprecedented opportunity to integrate information and obtain insight into the genetic and biological processes underlying disease susceptibility[21].

This seminal article is based on two key ideas: first, genetic variants most often impact gene expression, as shown by the many eQTL studies; second, SNP aggregation methods that combine many variants in a biologically meaningful way have the potential to improve GWAS results. In particular, the authors propose to group together all the SNPs that regulate the expression of a given gene. One advantage of

[21] Some of these consortia, whose data sets have been used by Gamazon *et al.*, are the following.

*ENCODE.* The focus is on the systematic functional annotation of each element of the human genome.

*GEUVADIS.* This project endeavours to uncover functional variation in humans through the study of how genetic variants affect gene expression.

*DGN.* Variants regulating gene expression, splicing and allelic expression are detected.

*Braineac.* The authors find eQTL in ten human brain regions.

*GTEx Project.* Its aim is to collect data on genotype and gene expression levels of a number of tissues from post-mortem samples.

this approach, which they called PrediXcan, is that statistical tests performed on group of SNPs are more powerful than those performed on each and every SNP, due to less multiple testing; besides, by choosing the gene as a grouping unit, information about the directionality of the effect is intrinsically provided, *i.e.* it is possible to say whether the disease is associated to an increase or a decrease in the gene's expression. Additionally, from a functional point of view a gene is much more interpretable than a simple genetic polymorphism.

### *Imputation of gene expression*

The main goal is to provide a framework to better interpret GWAS results. However, in a typical GWA study the individuals are simply genotyped with a SNP microarray, and expression data are completely missing. Therefore, gene expression has to be predicted by exploiting the knowledge of expression quantitative trait loci. Here, a linear regression model is fitted on a reference transcriptome data set, where both genotype and expression data are available; then it is applied on the individuals of a GWAS, for which only genotype data is collected, in order to predict expression levels.

The first assumption here is that gene expression can be decomposed into three components: a genetically regulated expression (GReX), a phenotype-influenced expression, and an environment-determined component (Figure 6). Some phenotype, including many diseases, can indeed influence gene expression, but, as we shall see in a moment, since the models that predict gene expression are trained on healthy individuals from reference transcriptome experiments, they already exclude that component from what they predict.

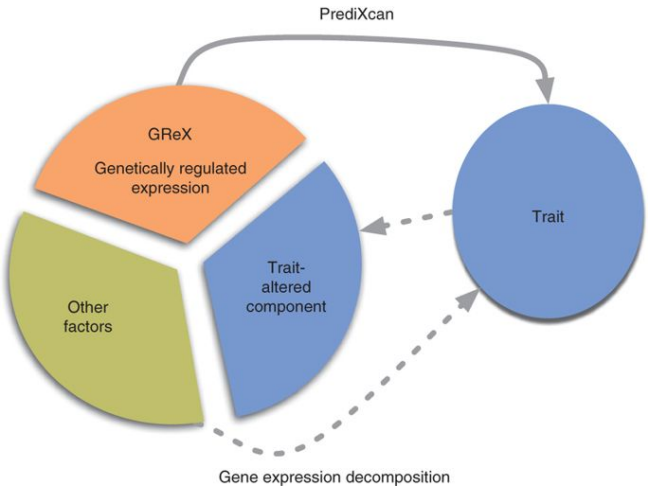


Figure 6: Gene expression can be decomposed into three components

The prediction relies on data sets where both genotype and expression are present, such as the aforementioned GEUVADIS and GTEx projects, and the model is additive (Figure 7 on the next page), mean-

ing that a given variant in a homozygous individual is supposed to have twice the effect of that same variant in an heterozygous individual. This is surely an oversimplification, for it does not take into account three biologically important effects —epistasis, dominance and penetrance—, but an additive model is much simpler to implement. Moreover, the model performs multiple linear regression, which is a first attempt to quantitatively model the interactions among *many* genetic variants: indeed, it may well be that a given phenotype is influenced by a *combination* of SNPs rather than a single SNP. The purpose of this regression model is to find for each SNP the coefficient of which gene expression is altered by a copy of that SNP. Once the coefficients have been estimated, the genetically regulated component of gene expression can be predicted starting only from the genotype of an individual; the predicted GReX is denoted as $\widehat{GReX}$.

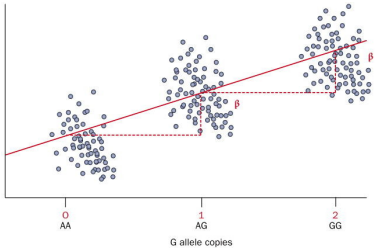


Figure 7: An example of additive model for one SNP; Gamazon *et al.* extended it for several SNPs. Image adapted from Conall M. O'Seaghdha and Caroline S. Fox, 'Genome-wide association studies of chronic kidney disease: what have we learned?'

They thus generated predictDB, which stores the coefficients of which each SNP influence the GReX. By using healthy individuals from reference transcriptome and genome data sets, they disregard the disease-determined component of gene expression, and by using a regression model, they disregard the random environmental component. It is now possible to 'impute' the transcriptome of an individual from its genotype, just like it is possible to impute unknown genetic variants in an individual from its known genotyped variants.

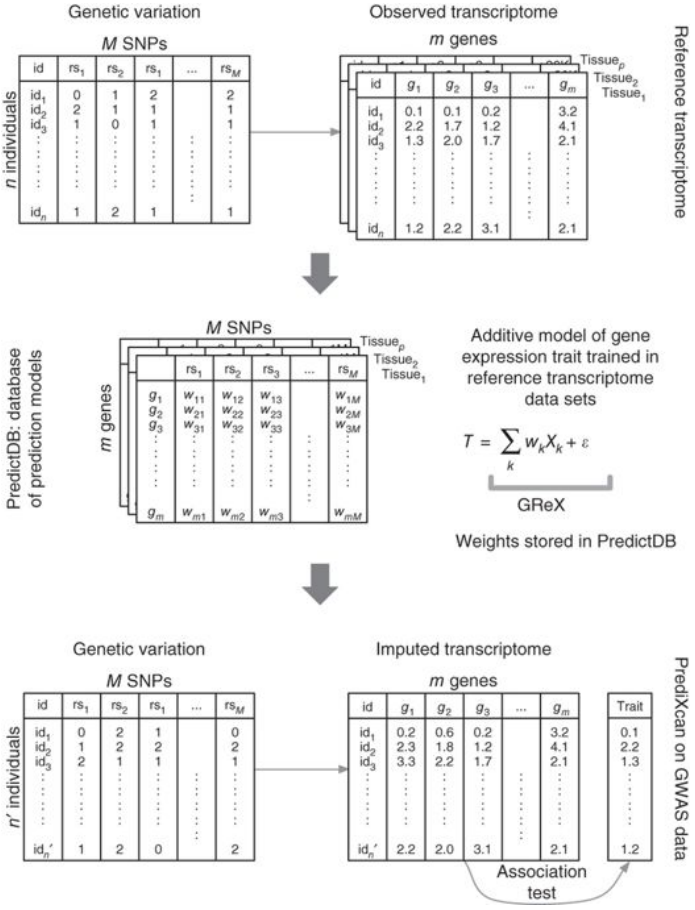


Figure 8: The framework to estimate the coefficient by which each SNP alters the expression of a gene.

The regression model employed can be summarised with the following equation (referring to Figure 8 on the preceding page):

$$T = \sum_{k=1}^{M} w_k X_k + \epsilon \tag{11}$$

where T is the expression of a given gene in a given tissue, $w_k$ is the weigth (or the effect size) of SNP $k$ in influencing the expression of that gene, $X_k$ is the number of reference alleles of SNP $k$ in all the data set, and $\epsilon$ is a residual error. Only SNPs falling 1Mb within the gene's start or end were considered. To fit this model, the authors tried various types of regression methods —lasso, elastic net and polygenic scrores—, but in the end settled to elastic net[22] (with the parameter $\alpha = 0.5$), whose main advantage is the 'automatic' selection of the most important regressors. In this case, the SNPs that have a small influence on gene expression, or that are correlated to another SNP that can explain more variation in gene expression, are neglected. 10-fold cross-validation[23] was used to asses the predictive performance.

[22] In general, elastic net is used for two reasons: first, when the number of predictors is large, especially if compared to the number of samples; and second, to avoid overfitting.

[23] In $k$-fold cross-validation, the dataset is split in $k$ portions, and for each part, the model is trained on the remaining $k-1$ parts, then the $R^2$ of the predicted and real values is calculated on the selected part. The average of the $R^2$ is finally reported.

Having seen the general method, we now turn to the actual protocol followed by the authors.

From DGN, they obtained whole-blood RNA-seq and genome-wide genotype data for 922 European individuals, normalised and filtered to retain only SNPs with MAF > 0.05, in Hardy-Weinberg equilibrium and univoquely mapped onto a strand. The SNPs that were not genotyped were imputed[24]. This data set was used to train the predictive models.

[24] Genome imputation is a routinely-used method to estimate the genotype of an individual at loci that were not analysed, basing on known linkage disequilibrium information in a reference population

From GEUVADIS, normalised RNA-seq data from lymphoblastoid cell lines established from 421 individuals was downloaded; genotype data was also available for the same individuals, since they were part of the 1000 genomes project as well. This data set was used for the validation of the models (Figure 9).

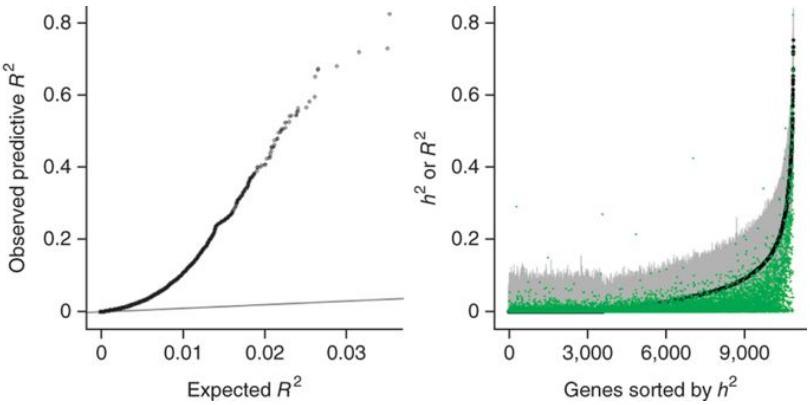


Figure 9: The performance in the GEUVADIS data set of the elastic net model trained on the DGN data set, showed by a quantile-quantile plot (left) and a distribution of $R^2$ (right).

In the quantile-quantile plot of Figure 9 on the facing page each point represents an expression value; on the $x$-axis it is reported the $R^2$ that would be expected if the null hypothesis were true (*i.e.*, if there were no correlation between predicted and observed expression values), while on the $y$-axis there is the observed $R^2$. A 45° line is displayed in gray: the farther the points are from that line, the more different the two distributions of $R^2$ are. Since in this plot the distributions are fairly discordant, we can reject the null hypothesis and state that the correlation between predicted and real expression is quite good. In the scatter plot to the right, the distribution of $R^2$ is plotted; the black line represents heritability, a threshold which theoretically the $R^2$ cannot pass, as explained in the next section.

RNA-seq data from GTeX, normalised and adjusted for the most common covariates such as sex, was used to test the predictive performance of the model in different tissues. Surprisingly, the model was able to predict gene expression quite accurately in tissues other than the blood, even though it was trained on blood expression data.

## *Heritability of gene expression*

Heritability is an important idea in genetics and is especially relevant in the scope of association studies, therefore we dedicate some space to its analysis.

Many traits vary among the individuals of a population: height and hair colour are obvious ones, but for instance also disease status (or the liability to it) can be considered a phenotypic trait. The heritability of a trait is the proportion of trait variance which can be explained by the genetic variance among the individuals of the population. In order not to underestimate heritability, only genetic variance at loci associated to the phenotype must be taken into account. Heritability does not deal with the influence of genes in the development of the trait, but rather is concerned with the role of genetic *variation* in determining phenotypic variation. There are two definitions of heritability:

*Narrow sense heritability,* or $h^2$, is the heritability due to additive genetic factors (see Figure 7 on page 23 and related discussion).

*Broad sense heritability,* or $H^2$, is the heritability due to all genetic factors, taking into account dominance and gene-gene interactions.

Usually, the first definition is used, and it is a reasonable approximation, for in the majority of case, alternate alleles are homozygous only in a minority of individuals due to their low frequency, hence the effects of dominance or epistasis manifest only rarely.[25]

If we assume that $P = G + E$, where $P$ is the phenotype, $G$ the genetics, $E$ the environment, then phenotypic variance can be expressed as follows:

$$Var(P) = Var(G) + Var(E) + 2Cov(G, E)$$

and, assuming the independence of genetics and environment,

$$H^2 = Var(G)/Var(P)$$

$$h^2 = Var(A)/Var(P)$$

[25] Visscher, Hill, and Wray, 'Heritability in the genomics era–concepts and misconceptions.'

There are many ways to estimate narrow-sense heritability. One is in selective breeding, where the heritability is the proportionality coefficient between the intensity of the applied selective pressure, $S$, and the response to selection, $R$ (Figure 10). In other words, $R = h^2 S$, wihch is the famous breeder's equation. The larger the heritability, the greater the response to selection.

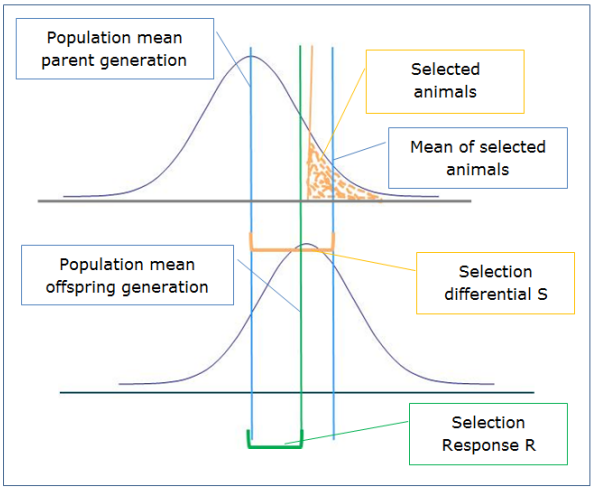


Figure 10: The breeder's work. Source: https://wiki.groenkennisnet.nl/

Another way to interpret heritability in the narrow sense is the plotting of the averaged trait in the two parents versus the trait in their offspring. In principle, if genes determine variations in phenotypes, then offspring should be more similar to their parents than to unrelated individuals. This can be expressed as a correlation between parents and offspring (Figure 11). In general, $h^2$ is the slope of the regression line of the phenotypes of offspring and parents. This, however, is valid only if the environment is not shared between relatives more than it is shared between unrelated people.

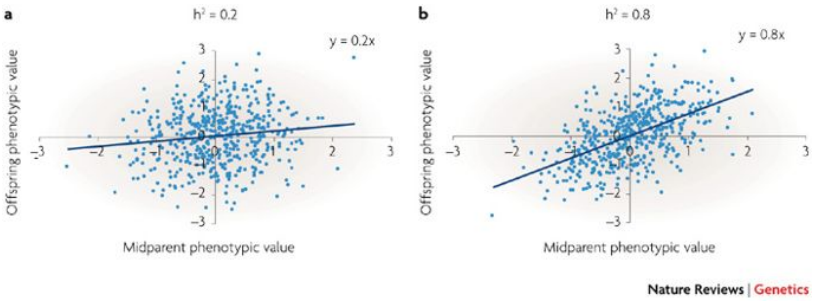


Figure 11: Parents-offspring regression. Source: Visscher, Hill, and Wray ('Heritability in the genomics era–concepts and misconceptions.')

Returning to the paper by Gamazon *et al.*, they rightly claim that heritability is an upper bound to how well the trait can be associated to the genotype. A high heritability means that the parents' trait can predict the offspring trait, or, equivalently, since this predictability is due to genetic factors, that people with a similar genotype will have a similar phenotype (indeed, $h^2$ is precisely the correlation between the phenotypes in parents and offspring).

The heritability of gene expression in DGN cells was computed, resulting in an average value of 0.153, whereas the average 10-fold cross-validation $R^2$ between the $\widehat{GReX}$ and the real expression was 0.137. Figure 12 reports the results of the cross-validation. Expression, contrary to other phenotypes, can be predicted very accurately from genotypes.

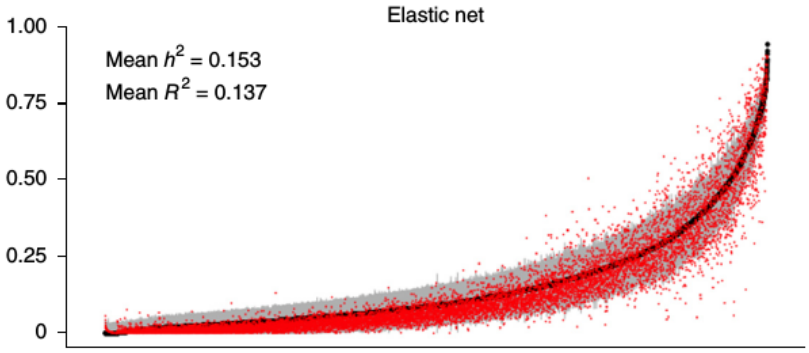


Figure 12: $R^2$ (red) of $\widehat{GReX}$ versus observed expression; heritability (black) of gene expression. The fact that this two measures are so close indicates that the prediction of the expression works.

## *Correlation between expression and phenotype*

In the second phase, the predicted $\widehat{GReX}$ is associated with the phenotypic status. To this aim, linear regression, logistic regression, Cox regression, or Spearman correlation (the latter is non-parametric) can be employed. For the results discussed in this article, logistic regression was chosen. Much like in a GWAS some variants are more frequent in cases than in controls, in a TWAS some genes will be more or less expressed in patients than healthy people.

This is the core part of a TWAS. It goes beyond genetic variants, and allows to explore the genetic basis of disease through the proxy of expression.

## *Application of PrediXcan to WTCCC GWAS*

At last, the method was applied to seven autoimmune diseases which had previously been the object of as many GWAS by the Wellcome Trust Case Control Consortium. They used DGN whole-blod elastic net prediction models to predict the expression in each WTCCC cohort, then correlated the predicted GReX with the disease status. After applying Bonferroni correction, 41 significant associations (P-value < 0.05) with 5 diseases were reported. Most of the significant associations were for autoimmune disease and were located in the extended MHC region[26]. Moreover, some genes were associated to multiple diseases[27] The majority of these associations were supported

[26] This region, located on chromosome 6, harbours 421 loci, including 252 expressed genes, 139 pseudogenes and 30 transcripts. Many of these loci are associated to diseases.

[27] In these cases, what determines which disease shows up if the expression of that gene is altered in an individual? Perhaps the environment, or gene expression level. This is an example of the complexity of the situation: the relationship between genotype and phenotype is not biunivocal at all.

by previous evidence, and often they were enriched in known GWAS; but two completely novel disease-associated genes were also found: low expression of *KCNN4* was associated with hypertension, and high expression of *PTPRE* with bipolar disorder.

One of the most interesting features of this new approach is that not only can it provide association results, but also a directionality. As an example, *PTPN22* expression was positively associated with rheumatoid arthritis and type 1 diabetes, and negatively associated with Crohn's disease.

The results of the associations with type 1 diabetes are reported in Figure 13.

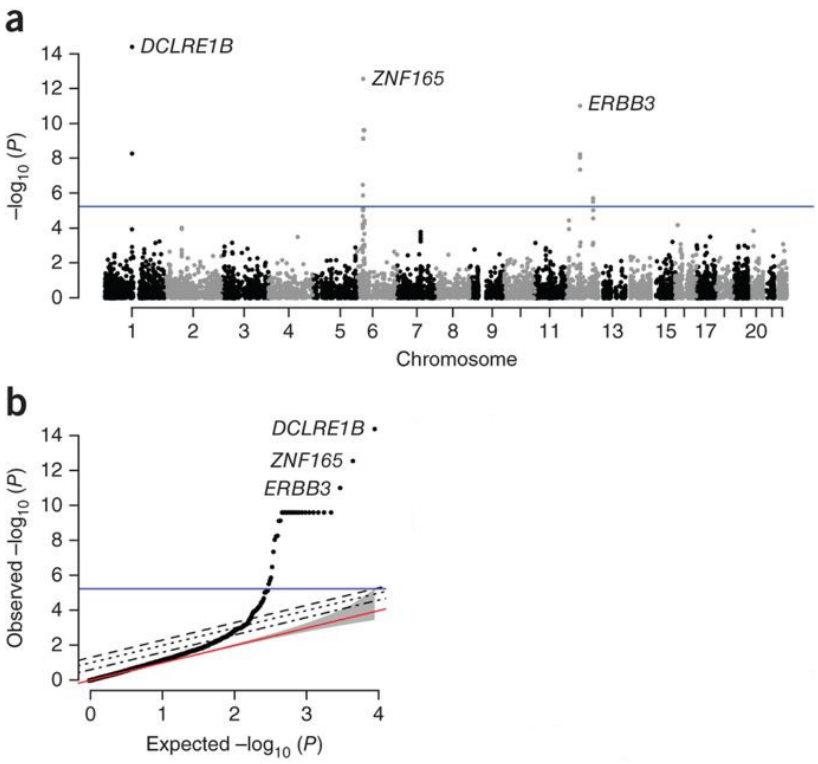


Figure 13: (a) Manhattan plot of disease-gene associations P-values. (b) Q-Q plot of the same P-values. The top three genes are emphasised.

## *Discussion*

General considerations on transcriptome-wide association studies are postponed until all the articles are discussed; in the discussion section of each paper, ideas about the particular article will be considered.

In this work, which can be regarded as the founder of the field of TWAS, the authors developed a framework to group together genetic variants, predict how such variants influence gene expression, and finally correlate gene expression to a disease. The last step was the

real association study, whereas the imputation of gene expression was needed only because expression data are missing in GWAS samples.

This imputation step makes PrediXcan economic, in the sense that one only needs reference transcriptome and GWAS data, which are already available, to perform a TWAS; therefore, many existing GWAS dataset can be reanalysed 'for free'. Nevertheless, due to privacy or logitic reasons, individual-level data[28] for published genome-wide association studies are often unavailable. The article analysed in the next chapter, by Gusev *et al.*, shall eliminate this limitation.

[28] That is, the genotype of each individual in the cohort.

The authors chose to use elastic net to predict gene expression, but argue that this expression might be biased, and that more sophisticated methods, like a combination of K nearest neighbours (KNN), elastic net and the use of genomic annotation, may perform better. Besides, we add that elastic net does not take into account biologically important concepts such as epistasis, dominance and penetrance.

The use of logistic regression to perform the association between expression and disease is arguable, too. For instance, another possibility could have been that not to model the disease status, which is a binary variable, but rather the liability to the disease, following what Vissher[29] reported showing that this approach is able to explain a larger proportion of genetic variance.

[29] Visscher, Hill, and Wray, 'Heritability in the genomics era–concepts and misconceptions.'

When applied to real world GWAS, PrediXcan performed well. Most genes had already been found, but two novel ones were reported as well. The features of these two genes were not investigated, but in the next articles we shall see that outliers like these were found to be regulated by multiple causal SNPs, rather than one (*e.g.*, see Section *Allelic heterogeneity* on page 37).

# *Integrative approaches for large-scale transcriptome-wide association studies*

*Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis, Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta & Bogdan Pasaniuc; Nature Genetics 2016*

*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*

As we have said, 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. PrediXcan, with which we dealt in the previous section, is not the only method to perform a TWAS. In particular, in 2016 it has been proposed a new approach where individual-level data are superfluous: only summary association statistics[30] from a GWAS is needed. This is an important advantage since, normally, only the summary-level data of a study are publicly available due to privacy concerns.

The description of this new approach is as follows. First, a Bayesian sparse mixed regression model finds the correlation between each SNP and gene expression from a reference transcriptome data set, and accordingly assigns a weight to each SNP; next, starting from the

[30] By summary association statistics we mean the effect size of all the SNPs and, optionally, the summary linkage disequilibrium information for the samples (*i. e.* the pairwise LD among typed SNPs). The genotype of single individuals is unknown.

summary association score of a genotyped SNP with the disease status, and from the weight by which the SNP alters gene expression, the association between expression and disease status can be estimated. Furthermore, by considering the linkage disequilibrium between the genotyped SNP and a non-genotyped SNP, and the weight by which the non-genotyped SNP alters expression, the association between expression of non-genotyped SNPs and disease status can be 'imputed'. The approach is quite different from PrediXcan, especially in the second part (Figure 14).

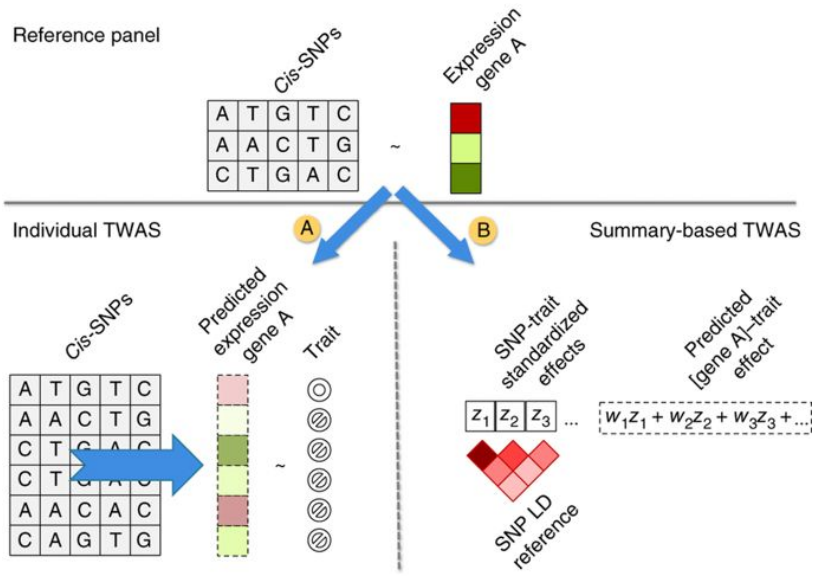


Figure 14: Schematics and comparison of individual-level and summary-level TWAS.

There are some relevant points in this new method: its being based on summary association statistics greatly increases the effective sample size, for the method can in principle be applied to any GWA study; moreover, the authors emphasise the specificity of their approach, for its focus is on the genetic component of expression only, computed from a reference transcriptome data set of healthy individuals; therefore, it is guaranteed that, save for pleiotropic effects[31], if an association between expression and trait is detected, it is ultimately due to genetic factors. In general, the TWAS approach is not able to detect associations between gene and disease if the mechanism does not involve a change in gene expression. Several are the ways in which genomic variation can be related to gene expression and phenotypic variation (Figure 15 on the facing page). Transcriptome-wide association studies work only under the hypothesis that a genetic variant influences, directly or indirectly, the phenotype through gene expression.

[31] Pleiotropy is a phenomenon where a single genetic locus *independently* influences more than one phenotype; for instance, an allele could alter gene expression on the one hand, and lead to a disease through a different mechanism. Pleiotropic effects cannot be modeled by TWAS.

The models to find the coefficients by which each SNP alters gene expression were trained on about 3,000 individuals whose expression data from blood and adipose tissues, as well as genotype data, were available. With the help of a simulated dataset, they compared their approch with others previously proposed, showing that theirs is

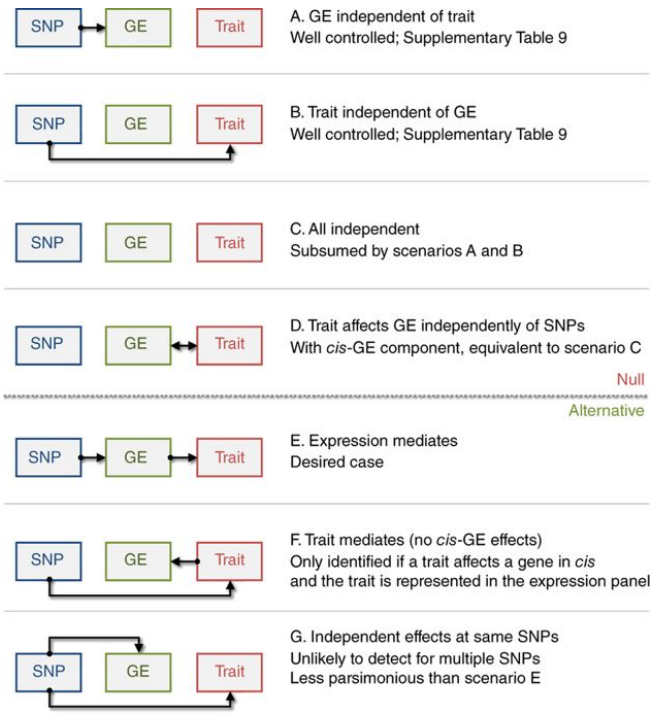


Figure 15: The possible models of causality. TWAS are useless in cases A-D, but work for E-G.

a significant improvement. Moreover, they reanalysed an existing dataset of a small-cohort lipid GWAS, finding that most of the novel associations they obtained with the TWAS had been reported in a larger-cohort GWAS, and implying that their method is statistically more powerful than SNP-based approaches, especially when the sample size is small. Finally, they applied their method to GWAS data for over 900,000 phenotype measurements, identifying many new disease-associated genes.

Table 2 illustrates the differences between the Gamazon and the Gusev methods.

| | PrediXcan | Integrative |
|---|---|---|
| Training data sets | DGR | METSIM, YFS, NTR |
| Prediction of expression | elastic net | BSLMM |
| Input data | individual level | summary level |
| Gene-disease association | logistic regression | correlation |

Table 2: Comparison between PrediXcan and the integrative approach.

## *Estimation of SNP-expression weights*

The accuracy of the prediction of a gene's expression cannot be greater than the heritability of the expression of that gene itself (see the discussion in Section *Heritability of gene expression* on page 25 ). For example, if a quantitative trait is normally-distributed in a population, but every individual has the same alleles at the same trait-associated loci, the

genetic variance in that population will be 0, and the heritability for such a trait would consequently be 0 as well. In such circumstances, it is not possible to predict the trait using the *cis*-genetic component of gene expression, for there is no such component: the differences in the individuals' traits depend only upon the environment, and it is notoriously difficult to quantitatively measure the effect of environment, especially outside of the laboratory. On the other hand, if the trait has an $h^2$ of 1, its manifestation can be predicted from the genotype with arbitrary accuracy, save for random variation due to chance[32].

[32] Indeed, environment and chance have different effects: the former generates a systematic bias in the trait, but is difficult to quantitfy, while the latter alters the trait because of the stochastic nature of life, and its average effect is zero in a large enough population.

In order to predict a quantitative trait from the genotype of the individuals, samples for which both gene expression and genotype data are present are necessary. The authors collected about 3,000 samples from three data sets: METSIM, YFS and NTR. Both the quantitative measures of the phenotypes and the gene expression levels were normalised and standardised before the analysis.

*YFS* is a long-term study of cardiovascular diseases in young finns.

*METSIM* studied the metabolic syndrome in men, collecting adipose tissue data in follow-ups of the young finns study.

*NTR* measured gene expression in peripheral blood in more than 2000 twins, computing the heritability of genes and finding eQTL.

From the data obtained from the about 3,000 individuals, the heritability of the expression of each gene was computed (Figure 16) using the tool GCTA.[33] For each gene, two heritability measures were estimated: *cis*- and *trans*- heritability, labelled $h^2_{g,cis}$ and $h^2_{g,trans}$; *cis*-heritability refers to the proportion of variance in gene expression that is imputable to variance in loci up to 1Mb from the gene, whereas *trans*-heritability is the proportion of variance in gene expression explained by the rest of the loci. Since on average any two non-related individuals differ at 0.1% of loci,[34] in order to estimate *trans* variance a very large sample size is needed, far larger than the 3,000 individuals used in this study, and this is the reason why estimates of *trans*-heritability are close to 0. All subsequent analysis were based on the 6,924 *cis*-heritable genes (Figure 17 on the facing page). Restricting the analysis to *cis*-SNPs greatly increases the statistical power of the study, for the number of predictors of gene expression becomes quite small; as previously explained, the multiple testing burden is also decreased.

[33] Yang et al., 'GCTA: A tool for genome-wide complex trait analysis'.

[34] Auton et al., 'A global reference for human genetic variation'.

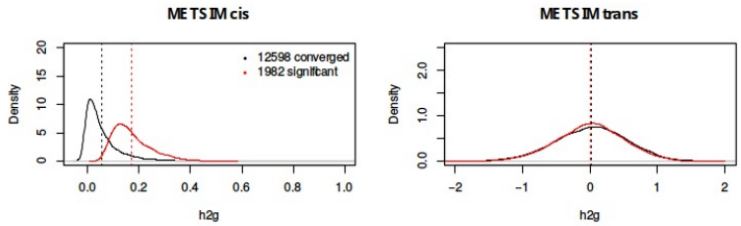


Figure 16: Heritability distribution in the METSIM data set. Distributions in the other data sets are not reported.

Having computed heritability, a statistical model could be trained to predict gene expression from genotype data; this is only needed to estimate the weights of which each SNP alters gene expression. Two different models, all based on the *cis*-SNPs, were employed: the first was a best linear unbiased model (BLUP) and the second a Bayesian

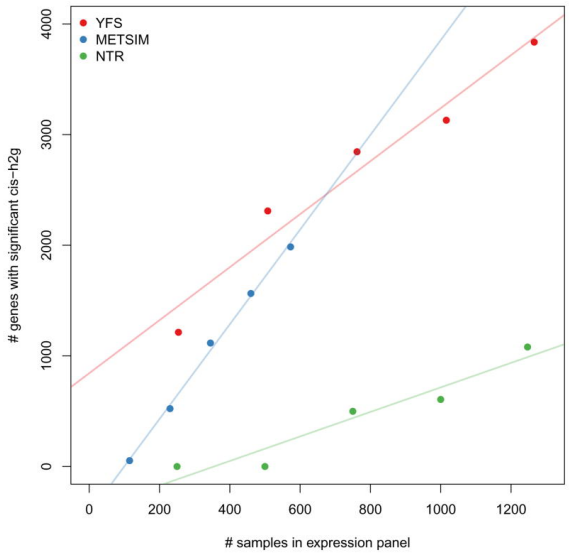


Figure 17: The 6,924 heritable genes, distributed according to their origin

sparse linear mixed model (BSLMM, see Section *Regression* on page 15 for technical details). The performance of each model was evaluated by cross-validation. Moreover, these two models were compared to the predictions of gene expression made from the best *cis*-eQTL. The Bayesian model was the best one (Figure 18), therefore it was used for subsequent analysis.

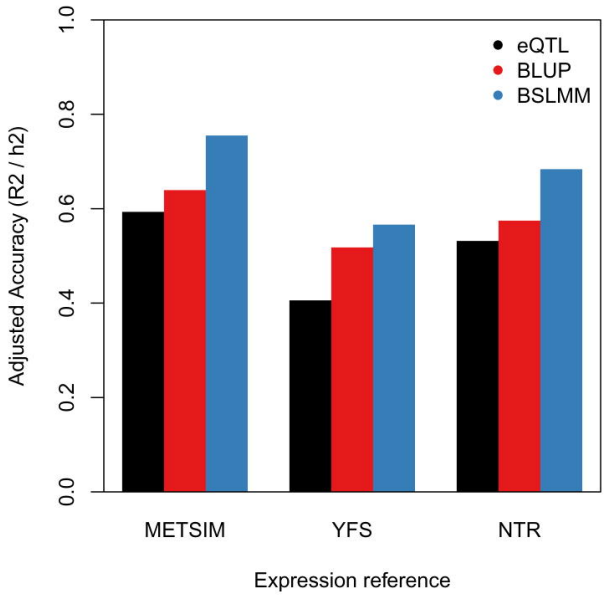


Figure 18: BSLMM performs better, as can be seen by the cross-validation results.

## *Application to a small-cohort GWAS*

The BSLMM was trained on the three data sets (METSIM, YFS and NTR) as described previously, and the weight of each SNP was estimated; then, the correlations between the *cis*-regulated expression of each gene and the lipid phenotypes from the GWAS were calculated

using summary-level statistics. 25 correlations were found between phenotype and genes that were more than 500 kb far from any significant SNP in that study, and 19 of these genes contained at least a significant SNP in a larger blood lipid GWAS. This result is an additional confirmation that the proposed method is valid.

The method they used to find associations between gene expression and phenotype from the summary statistics of the GWAS is a generalisation of a method previously proposed by the same authors to impute SNP-phenotype associations in a GWAS, knowing only the association score of genotyped SNPs[35] and optionally the summary LD statistics. In essence, the original method is based on the assumption that the z-scores[36] of the SNPs in a locus are normally-distributed, with mean 0 if there is no association, and with a standard deviation depending on the correlation among the SNPs, that is, on the LD structure of the locus. The LD structure can be taken from the reference genome or, if available, directly from the population of the GWAS. Thus, with the previously developed method, by knowing the z-score of a genotyped SNP and the LD between a non-genotyped SNP and the genotyped one, it is possible to estimate the z-score of the non-genotyped SNP. The method was adapted in this paper so as to wheight the imputed z-score of a SNP by the coefficient of which that SNP alters gene expression. This coefficient was previously estimated in a reference transcriptome data set using the BSLMM.

[35] Pasaniuc et al., 'Fast and accurate imputation of summary statistics enhances evidence of functional enrichment'.

[36] The z-score is the standardised association score and measures of how many standard deviations the score differs from the mean.

In other words, the association score for a gene can be expressed as

$$\mathbf{Z}_{TWAS} = \frac{\mathbf{W}\mathbf{Z}_{GWAS}}{\sqrt{\mathbf{W}\mathbf{D}\mathbf{W}^t}}, \tag{12}$$

where $\mathbf{Z}$ is a vector of the standardised effect sizes of the *cis* SNPs of one gene; $\mathbf{W}$ is a vector of the weights of those SNPs; $\mathbf{D}$ is the linkage disequilibrium matrix, representing the correlation between SNPs. Therefore, the numerator is a linear combination of effect sizes and weights: $w_1 z_1 + w_2 z_2 + \ldots$, while the denominator is the standard deviation of the numerator.

## *Application to 900,000 phenotypes*

One of the most innovative features of this approach is its broad applicability. Indeed, its potential was unleashed on three GWAS which accounted for over 900,000 phenotype measurements of obesity-related traits[37]. 665 significant gene-trait associations were found, 69 of which genes did not overlap any SNP which was reported by the original GWA studies.

[37] Lipid measures (high-density lipoproteins [HDL] cholesterol, low-density lipoprotein [LDL] cholesterol, total cholesterol [TC], and triglycerides [TG]); height; and BMI

Those 69 novel associations are the most interesting ones, therefore they were the focus of a functional analysis: on the one hand, their presence was sought in the Hybrid Mouse Diversity Panel (HMDP), which collects obesity-related phenotypes; on the other hand, tissue-specific enrichments of these genes was evaluated. Many of the 69 genes were indeed present and they were associated with an obesity-related trait. Moreover, the tissue enrichment analysis, performed with DEPICT, showed that the novel genes were specific of hypothalamus and neurosecretory systems, which is consistent with recent discoveries on obesity. No pathway enrichment analysis was performed.

## *Allelic heterogeneity*

For comparison purposes, the authors built an array of simulated data sets, each modelling a possible scenario (1 causal variant, 5% causal or 10% causal), and performed a TWAS, a GWAS and an eGWAS[38] on them. On the whole, the performance of the TWAS was comparable to the others when the number of causal variants was small, but it was better at associating multiple causal variants to the trait (Figure 19).

[38] In the GWAS, trait-associated variants falling in a gene are retained; in the eGWAS, only the best eQTL for each gene is tested for associations with the trait.

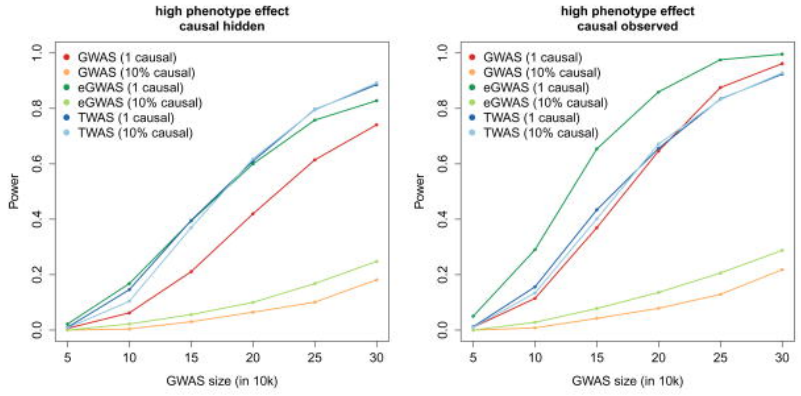


Figure 19: The TWAS approach compared to GWAS and eGWAS.

This can be explained by allelic heterogeneity, *i.e.* the presence of multiple different variants affecting the same gene in the same way. For instance, $\beta$-thalassemia can be due to several mutations. In these cases, GWAS are disadvantages because they focus on single variants, each of which can be very rare. On the other hand, whatever genetic variant is present, the expression of the involved gene will be altered in nearly the same way.

## *Discussion*

The main idea introduced with this work, with respect to the previous one, is that individual-level data are no longer necessary. In the words of the authors, this method can be viewed as 'a test for the correlation between the genetic component of expression and the genetic component of a trait'. Indeed, the z-score of the association between gene expression and trait is a function of the z-scores of the *cis* SNPs of that gene, the weight of which those SNPs alter the expression of the gene, and the linkage disequilibrium between them (see equation (12)).

This complicates the statistics, as more assumptions are needed (in particular, that z-scores are normally distributed with mean 0 in the case of no association), and another downside is that rare variants are less likely to be identified. Nevertheless, the benefits outweight the costs, for the method can be used on virtually every GWAS performed to date, and just in this article many novel genes were found associated to phenotypes.

Here, a Bayesian sparse linear mixed model is used instead of elastic net; both perform variable selection, meaning that only the best predictors are employed. The main difference between the two regression models is that the former takes account of random effects, modeled with a normal distribution.

Another possible point of debate about this work is that genetic variation does not only alter gene expression. A SNP can have effects on splicing, transcription start or end site or other RNA editing processes, without altering the expression of the gene. The next article shall deal with such problems.

# *Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights*

*Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K. Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Steven McCarroll, Benjamin M. Neale, Roel A. Ophoff, Michael C. O'Donovan, Gregory E. Crawford, Daniel H. Geschwind, Nicholas Katsanis, Patrick F. Sullivan, Bogdan Pasaniuc & Alkes L. Price; Nature Genetics 2018*

Abstract

Genome-wide association studies (GWAS) have identified over 100 risk loci for schizophrenia, but the causal mechanisms remain largely unknown. We performed a transcriptome-wide association study (TWAS) integrating a schizophrenia GWAS of 79,845 individuals from the Psychiatric Genomics Consortium with expression data from brain, blood, and adipose tissues across 3,693 primarily control individuals. We identified 157 TWAS-significant genes, of which 35 did not overlap a known GWAS locus. Of these 157 genes, 42 were associated with specific chromatin features measured in independent samples, thus highlighting potential regulatory targets for follow-up. Suppression of one identified susceptibility gene, mapk3, in zebrafish showed a significant effect on neurodevelopmental phenotypes. Expression and splicing from the brain captured most of the TWAS effect across all genes. This large-scale connection of associations to target genes, tissues, and regulatory features is an essential step in moving toward a mechanistic understanding of GWAS.

## *Introduction*

GWAS hits are difficult to explain from a mechanistical point of view, for the association with the disease can arise in many different circumstances. In the great majority of cases, the GWAS hit is not even the real causal variant, but is merely in linkage disequilibrium with it; and even if we knew which is the actual causal variant, we still could not infer much about its functional role without a deeper knowledge of the biology at the locus where the variant lies. Integrating GWAS signals with a functional annotation of the genome can give insight into the mechanisms through which the variant affects the phenotype; in particular, it has been shown that schizophrenia GWAS hits were enriched in regulatory elements. In this paper, this concept was extended to transcriptome-wide association studies to identify putative regulatory mechanisms, through which the associations could make biological sense.

The regulatory role of a genetic region is mainly determined by its chromatinic state, *i.e.* by how histones are modified, by which proteins bind in that region, and by whether the DNA is methylated or not; and its chromatinic state is in turn influenced by DNA elements either in *cis*, for an alteration in a sequence that binds a protein can impair the protein's regulatory activity, or in *trans*, for an altered protein that does not recognise a DNA motif any more cannot work properly. Ultimately, however, the chromatinic state of a region is under genetic control as much as any other phenotypic trait, therefore we have two effects that can spring from a genetic variant: the association with the disease or the association with the chromatinic state of a region. Such effects may or may not be independent. In this study the focus is on those genetic variants which both affect the chromatin structure of a locus, and alter gene expression, leading to a disease. The purpose, then, is to find a causal mechanism of action for variants associated with a disease.

The authors, by exploiting the method they had previously developed (see Chapter *Integrative approaches for large-scale transcriptome-wide association studies* on page 31 ), performed a schizophrenia TWAS relying on summary-level data from a published large-scale GWAS, and subsequently performed a chromatin TWAS in order to find genes whose expression was associated with a chromatin phenotype. They then compared the two sets of genes aiming to gain insight into the biological function of the genes associated with schizophrenia. Their approach is summarised in Figure 20.

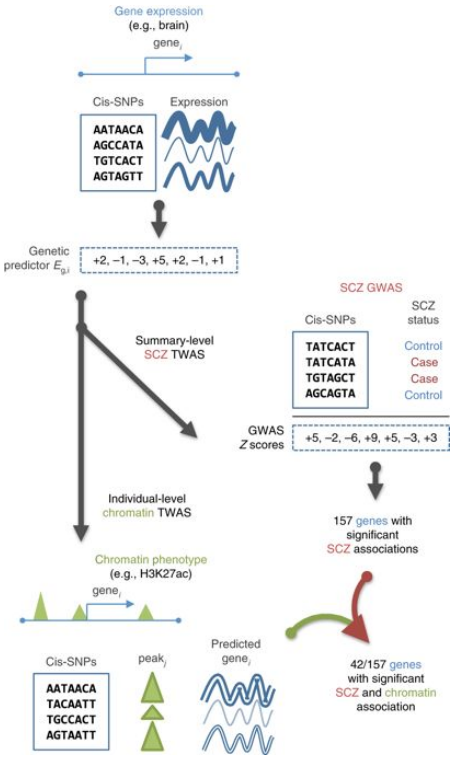


Figure 20: The schematic of the TWAS approach used in the schizophrenia work.

## *Schizophrenia*

Schizophrenia (MIM: 181500), which affects more than 20 million people worldwide with symptoms such as hallucinations, delusions, depression, and poor cognitive functionality, is a psychiatric disorder characterised by a loss of contact with reality. As with any complex disease, the manifestation of schizophrenia is influenced both by genetic and by environmental factors; to date, more than 100 loci have been associated to this disease, and exposure to viruses, malnutrition or psychosocial factors are thought to have a contribution to it as well. Schizophrenia is treated with drugs and psychosocial support.

## *Different levels of phenotype*

In this paper, the authors treated chromatin peaks as quantitative traits, much like they did with gene expression. This warrants further considerations on the nature of phenotypes.

The process of biological development of a multicellular organism is very complex: it involves signaling, differentiation, growth, cell division, and everything must happen in a coordinated fashion. The fascinating thing is that this process is self-regulated: the first cell contains all the information necessary to produce the final organism, and this information flows from moment to moment, from cell generation to cell generation, like a cascade carrying each cell towards its destiny. During development, all the genes interact with each other and contribute to the expression of phenotypes. No single gene can be held *responsible* for a phenotype, but we can say that genetic differences at that gene in a population result in different phenotypes.[39] For example, a mutation in the gene *white* of *Drosohpila melanogaster* results in insects with white eyes instead of red; however, that gene just encodes for an ABC transporter,[40] thus it cannot be responsible for the eye colour unless we insert it in the context of an already developed eye, which consists of many cells which assumed the structure of the eye during development. If we traced back the history of an eye cell, we would see that it changed dramatically, thanks to the collaboration of many genes; the *white* gene is only the last one, and would not make sense without the previous history of the whole eye, and indeed of the whole organism.

Furthermore, DNA should not be seen as the only lowest-level determinant of every phenotype. Indeed, it is becoming clearer and clearer that epigenetic mechanisms can change gene expression (and other phenotypes as well) without altering DNA, and at the same time being inherited. This means that epigenetic mechanisms are under the effects of natural selection.[41] But epigenetics is not the only

[39] Griffiths et al., *Introduction to genetic analysis*.

[40] Mackenzie et al., 'Mutations in the white gene of Drosophila melanogaster affecting ABC transporters that determine eye colouration'.

[41] Hunter, 'Extended phenotype redux. How far can the reach of genes extend in manipulating the environment of an organism?'

high-level mechanism that can influence gene expression. When a cell undergoes mitosis or meiosis, each of the daughter cells stochastically inherits about half of the cytoplasm of the mother cell, and since the cytoplasm contains many molecules capable of altering gene expression —such as transcription factors—, daughter cells are already 'primed' towards a particular destiny, without any change to nucleotides. The fact that the cytoplasm can influence DNA can be clearly observed in cloning experiments, where embryogenesis is not triggered by the DNA itself, but by the cytoplasmic environment in which DNA lies.

Phenotypes can be seen at many levels: from chromatin states to gene expression, from the proteins present in the cell to eye colour, and even beyond the physical aggregate of cells which is the organism (for instance, a bird's nest is a phenotype, for a better nest increases the fitness of the bird, according to the view of the extended phenotype[42]). Any phenotype can, in principle, affect any other phenotype, as well as DNA itself (*e.g.* some phenotypes increase mutation rate of DNA). Therefore, there is no linear relationship between genotype and phenotype, but rather a tight integration of all the levels. Evolution can act at different levels, too.

[42] Dawkins, *The extended phenotype: the long reach of the gene*.

Genetic variation, however, can be argued to be the ultimate source of variation. Epigenetic modifications, for instance, need two pieces of DNA: a gene encoding for a DNA binding protein (or a methyltransferase), and a sequence recognised and bound by the protein. According to what regions of DNA, be they coding or non-coding, are expressed, different enzymes and specific membrane transporters can be produced, thereby determining which reactions can happen and what molecules can enter the cell.[43] Moreover, genetic variation is very easy to measure, so most efforts have concerned it.

[43] Alberts et al., *Molecular Biology of the Cell 6e*.

## *Traning of the expression models*

Coming back to the work of Gusev *et al.*, four datasets of either RNA-seq or genome-wide SNP-array expression measurements, for a total of nearly 4,000 individuals, were used to train the expression models[44]. In particular, some of the data came from the YFS and METSIM studies, and the weights for the SNPs in those cases were pre-computed by Gusev *et al.* for the 2016 paper. The other dataset, the CMC, was added in order to train expression models in the brain, which is in all probability a relevant tissue in schizophrenia. The fact that both healthy individuals and patients were present in this last cohort did not impair the prediction of gene expression.

For each reference panel, *cis*- and *trans*- heritability of gene expression were computed and found significant for 18,084 genes (10,819 unique).

[44] RNA-seq from the dorsolateral prefrontal cortex of 621 individuals (including 283 schizophrenia cases, 47 bipolar cases, and 291 controls) collected by the CommonMind Consortium (CMC); expression array data measured in peripheral blood from 1,245 unrelated control individuals from the Netherlands Twin Registry (NTR); expression array data measured in blood from 1,264 control individuals from the Young Finns Study (YFS); and RNA-seq data measured in adipose tissue from 563 control individuals from the Metabolic Syndrome in Men study (METSIM).

In addition, 9,009 splicing events in the brain were characterised, since an alteration of this kind of regulation is implied in disease. From such datasets, a sparse midex linear model (the BSLMM also used in the previous paper) was trained to find the weight of each *cis*-SNP for each gene. Furthermore, since a population LD structure is necessary to estimate the correlation between the genetic component of expression and a phenotypic trait using summary GWAS information, LD informaton was also extracted from these data sets.

## *Schizophrenia TWAS*

A separate TWAS was performed for each of the four reference gene expression training datasets, using the GWAS summary information from a large study of about 80,000 samples. 157 unique genes, of which 35 completely novel (*i.e.* being farther than 500kb from any previously associated SNP), were found significantly associated to the disease, as shown in Figure 21. Interestingly, 33 loci were found to harbour hotspots of multiple TWAS hits (*i.e.* many genes less than 500kb apart), but, with a statistical test, it was found that in 27 of these cases the genes were correlated, suggesting a single underlying effect. For instance, this could be due to the alteration of the structure of an entire topologically associating domain (TAD).

For these analysis, the MHC region was not considered due to its complexity.

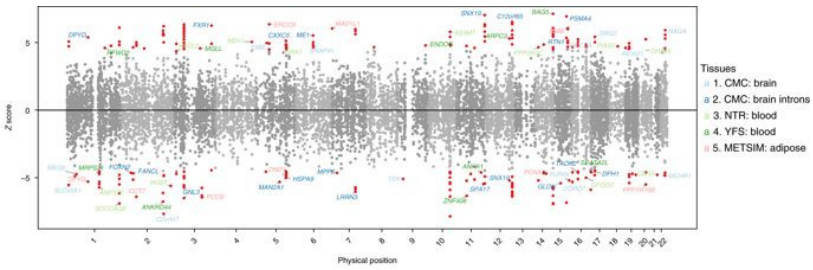


Figure 21: Results of the schizophrenia TWAS. Red dots in the manhattan plot represent significant genes.

Again, some evidence emerged that TWAS are more powerful than GWAS at detecting multiple variants whose combined effects explain the phenotype, rather than events where a single variant is involved (see Section *Allelic heterogeneity* on page 37): indeed, 27% of the novel genes were associated to the disease more strongly than the top GWAS hit at the locus of the gene, whereas only 3% of the genes overlapping a reported GWAS hit were more strongly associated than the GWAS hit (Figure 22). This indicates that when there is a single causal variant, the GWAS is best at identifying it, but when there are multiple variants affecting the trait, the TWAS performs better.

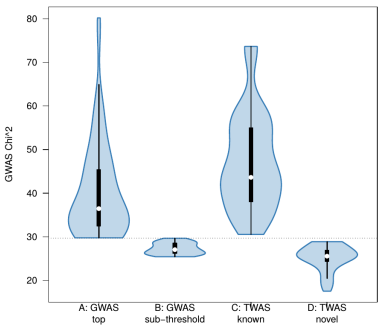


Figure 22: Violin plot of GWAS $\chi^2$ for different sets of variants.

After finding those genes whose expression is associated to schizophrenia, the authors evaluated on the one hand whether there was an association between splicing events and disease, and on the other hand whether the significant genes had chromatin interactions with

one of the causal SNPs associated to schizophrenia.

46 splicing events in the brain were also found associated to the disease. Splice variants were detected using LeafCutter, a previously published algorithm which is based on the clustering of reads that span intron junctions; each cluster identify an isoform. The abundance of each isoform, after normalisation and PEER-correction, was treated as a quantitative trait in the same way as gene expression.

The chromatin interactions were derived from a Hi-C study of the human brain during development, while the causal SNPs were taken from the set of the fine-mapped ones[45]. From these data it was possible to trace back all the genes which physically interact with one of the causal SNPs. Among such 'risk genes' there were 105 of the 157 TWAS-associated genes, pointing at a mechanistic reason for the associations and relating gene expression and regulation[46].

[45] Fine mapping aims to find the causal variants associated to a trait. First, all the SNPs in linkage disequilibrium with a GWAS hit are selected, then the probability of each variant to be causal is estimated with a difficult statistical method.

[46] Actually, this mechanism is only valid if it is assumed to manifest during development, for the chromatin interaction experiment was performed in a zebrafish embryo.

## *Chromatin TWAS*

Chromatin data for nine markers was obtained from ChIP-seq[47], and the intensity of each peak was treated as a quantitative trait. Inside the cohorts with chromatin information, a TWAS was performed through the usual two steps: first, gene expression was imputed for the 10,819 heritable genes and the spliced introns; then, gene expression and spliced introns expressionon was correlated to the chromatin state. Overall, 806 genes and 224 splicing events associated to at least one chromatin state were found (Figure 23 on the next page).

[47] Chromatin immunoprecipitation-DNA sequencing consists in the sequencing of DNA fragments bound to specific proteins, captured by an antibody. In this case, the data came from lymphoblastoid cell lines from Yoruban and European populations:

*LCL of YRI:* acetylated histone H3 Lys27 (H3K27ac; marking active enhancers), methylated H3 Lys4 (H3K4me1; enhancers), trimethylated H3 Lys4 (H3K4me3; promoters), and DNase I–hypersensitive sites (DHS; open chromatin).

*LCL of CEU:* H3K27ac, H3K4me1, H3K4me3, the regulatory transcription factor PU1, and RNA polymerase II (RPB2, associated with active transcription).

## *Putative regulatory mechanisms*

The integration of the two types of TWAS, that on chromatin and that on schizophrenia, can provide biological insight into the mechanism through which the associated genes influence the disease. Of the 157 genes associated to schizophrenia, 42 were also associated to a chromatin state, and, in particular, only 8 of the 42 genes were associated to a chromatin peak located in the promoter of the gene itself, suggesting that the majority of disease-associated genes are regulated by enhancers.

There are two possible models to explain the association between SNP and trait: one in which the SNP mediates the change in chromatin structure, which in turn alters gene expression leading to the disease; and one where the SNP directly mediates a change in the expression of some genes, and this in turn affects chromatin activity. In either case, the association of a gene both to a chromatin state and to a

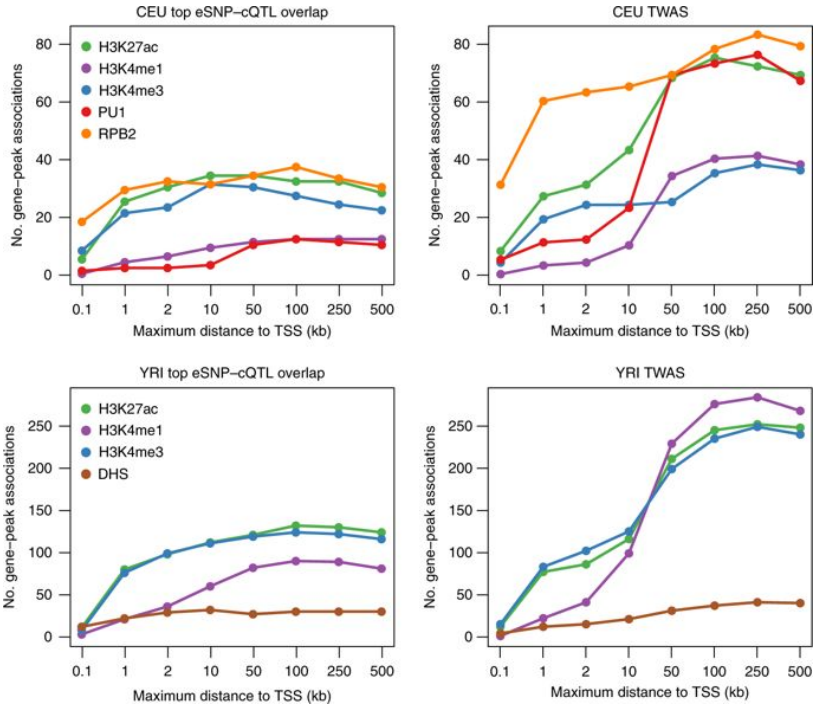


Figure 23: Chromatin TWAS. eSNP-cQTL are those variants that influence both expression and chromatin, find by traditional analysis. The TWAS was able to find more associations, especially if the chromatin peak was far from the gene.

disease suggests the existence of particular regulatory mechanisms, which deserve experimental validation.

## *Example*

As an example, the locus around the gene *KLC1* was chosen. This gene encodes for the light chain of kinesin, which, as a tetramer composed of two heavy and two light chains, exploits microtubules to carry various molecules and other cargos along the cell (Figure 24). Figure 25 on the next page shows in (a) an overview of the locus: the gene is associated to schizophrenia and to two chromatin peaks —H3K4me1 and H3K4me3—, and an Hi-C signal confirms the interaction between the promoter of *KLC1* and the regions where the chromatin peaks lie. The other portion of the figure, (b), display on the left a manhattan plot of the P-values of association between SNP and disease status, with the association either being weighted for the expression (coloured dots, like in a TWAS) or not (black dots, like in a GWAS); on the right, there are scatter plots reporting on the $y$ axis the z-score of the association between GWAS and eQTL, and on the $x$ axis the correlation between the z-score and the predicted expression.

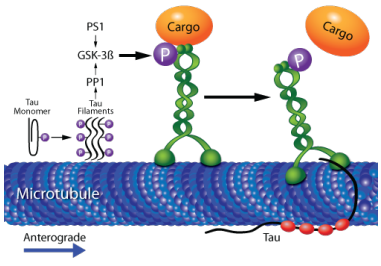


Figure 24: A kinase phosphorilates the kinesin triggering a conformation change in the protein which results in its movement along the microtubule filament. https://www.cytoskeleton.com

It can be seen from (b) that no SNP was genome-wide significant at that locus, but after accounting for their contribute to gene expression, many of the variants passed the significancy threshold and were therefore transcriptome-wide significant. Finally, the relationship

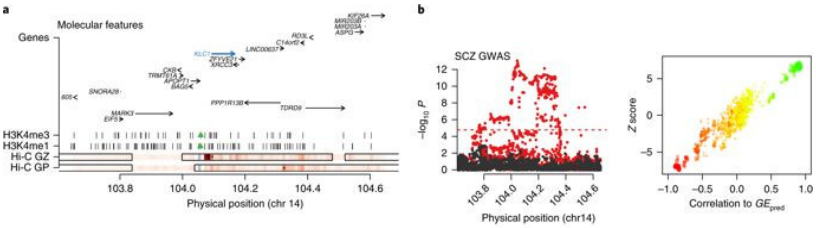


Figure 25: the *KLC1* locus.

between the z-score and the predicted expression is linear and has a positive slope, meaning that the more the gene is expressed, the higher the risk to develop schizophrenia.

## *Functional validation*

Another interesting gene whose expression correlated with schizophrenia and with two chromatin peaks, is *MAPK3*, encoding a mitogen-activated protein kinase which regulates many aspects of cell growth and proliferation.

Previous studies found that *MAPK3* and *KCTD13* are coregulated, and that *KCTD13* over-expression causes microcephaly because of its impairment of neuronal proliferation. The TWAS results show that if *MAPK3* is over-expressed, then the risk of disease increases; the authors hypothesised that, if *MAPK3* acts through *KCTD13* on schizophrenia, then down-regulating the gene should rescue the disease.

A zebrafish model over-expressing *KCTD13* was built to test this hypothesis (Figure 26). As expected, the heads of the fish embryos were small and the cells therein were less proliferative, but after the suppression of *MAPK3*, the embryos developed normally.

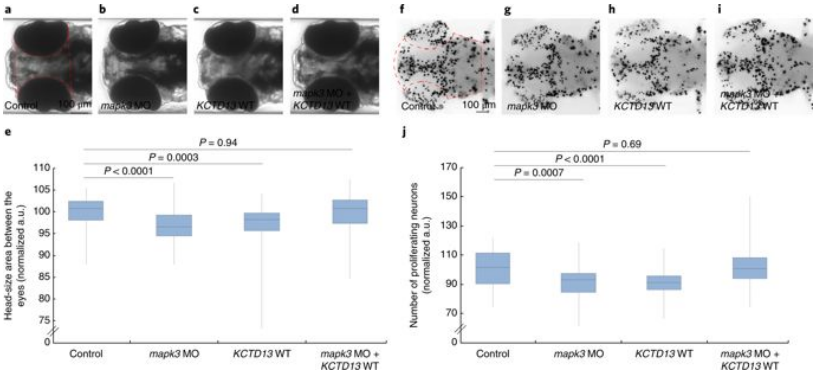


Figure 26: A zebrafish model validates the discovery of a gene associated to schizophrenia.

It could be argued that this is not a real validation, for fish do not have schizophrenia. However, modeling human psychiatric disorders

is not easy either. At least, this experiment validates that the gene and the phenotype are indeed associated, even in the real world.

## *Discussion*

The motivation behind this work was double: first, an association does not imply a causal mechanism; secondly, gene expression is not the only way through which genetic variants can alter a phenotype. For instance, a gene can be differentially spliced without its expression level changing.

This was also an attempt to go beyond genetic variants alone and relate chromatin structure to gene expression. The integration of a disease TWAS with a chromatin activity TWAS allowed to make hypothesis on the mechanism through which the expression is regulated. A direct association between chromatin activity and disease, however, was never proven[48], therefore these result provide possible mechanistic explainations for how genetic variants regulate gene expression through the modification of chromatin (or *vice versa*), and not for how the altered expression of a gene lead to the disease.

[48] Indeed, the sample size was too small to perform a chromatin-wide association study

The method and most of the data set where the expression models were trained, were the same as those used in the previous article. A new entry here was the integration of experimental biology to suggest or validate hypothesis. First, the chromatin interactions were used to confirm that, at least once in the life of a living organism, disease-associated genes interacted with disease-associated SNPs. Then a zebrafish model validated an hypothesis on a particular gene, *MAPK3*.

# *Conclusions and future perspectives*

Transcriptome-wide association studies were born to address some of the limitations of classical association studies by combining many SNPs in a biologically meaningful way, *i.e.* through their effect on the expression of a gene. However, one first problem was the paucity of large-cohort studies where genomic, transcriptomic and phenotypic data are collected simultaneously; on the other hand, GWAS, where only genotype and phenotype are characterised, are very common. This problem was solved by imputing gene expression with a regression model able to predict the levels of expression starting from genotypic information. But another problem arised: in published GWA studies the full genotype of each individual is not available; in order to solve this problem, the association between gene expression and trait was performed at the level of the whole GWAS cohort, still relying on reference transcriptome data sets to grasp the relationship between SNPs and expression.

The aggregation of SNPs in functional units ensure that TWAS perform better than GWAS when there are multiple causal SNPs (see Section *Allelic heterogeneity* on page 37). Statistically, TWAS are more powerful than single-variant-based approaches, for the multiple testing buden is reduced. Moreover, the interpretability of results increases and directionality information is provided. From a human point of view, knowing that having a 'G' rather than a 'T' at a locus makes one more liable to a disease does not make much sense. On the contrary, knowing that one is at risk if a gene is more (or less) expressed than normal, is somewhat reassuring.

The other side of the coin is that TWAS miserably fail when genetic variants influence the phenotype independently of gene expression. Even worse, if a variant pleiotropically affects both gene expression and disease independently, TWAS are confounded, in the sense that they report an association between expression and disease, where in fact there is none.

A genetic variant can affect phenotypes in many ways, which fall in two main broad categories: a regulatory or a structural one. In the paper by Gusev *et al.* on schizophrenia, the possibility of structural

alterations was explored by associating the expression of particular spliced isoforms to the disease, to perform a 'spliceome-wide association study'. In the same paper, gene expression was associated to chromatin activity. These examples highlight the fact that association studies can go beyond genetic variants alone. In principle, every intermediate phenotype in the path from genome to disease could be associated with the disease status. A possible future perspective is to integrate associations from more than one phenotypic levels (see Section *Different levels of phenotype* on page 41 ). Indeed, genetic variants can have different effects in different contexts.

An advantage of performing associations using gene expression is that this intermediate phenotype is fairly heritable, and above all its prediction from genetic variants explains a good proportion of such heritability. On the contrary, higher-level phenotypes are difficult to predict from the genotype. This could mean that there is a more direct link between gene sequence and gene expression than between gene sequence and height. The 'missing heritability', not explained by GWAS, could be lost in processes that do not depend directly on gene sequence: each time a phenotypic level is crossed, the possible combinations expand.

Is there a way to understand the mechanisms behind association and to find causal variants? This should remain an open question for now. Although transcriptome-wide associations can lead to putative causal genes, they are cannot replace experimental validation. Fine mapping of causal genes is still necessary.

But the fact that we do not know the mechanism does not prevent us from exploiting existing associations. For instance, if the expression of a gene is positively correlated with the risk of a disease, we could develop drugs that down-regulate that gene. Not only are genes more interpretable: they are also more druggable than SNPs.

In principle, any population whose genotypic and phenotypic data are available can be used to perform a TWAS. For instance, one could start from a population of tumours, such as those collected by TCGA, and leverage this data to find driver genes. Another possibility is to start from a population of single cells and ask which are the genes that make each of them unique.

A possible limitation of TWAS is as follows. If a disease-causing variant alters the activity of a transcription factor, and this altered transcription factor affects the expression of hundreds of genes, then every one of these genes would be associated to the disease, albeit the 'true' causal gene was that encoding for the transcription factor. Moreover, multiple variants in different individuals can lead to the same disease, because they alter either the same gene or different genes partaking in the same functional pathway or coregulated. All

this heterogeneity decreases the statistical power to find associations. Perhaps one of the possible solutions is to aggregate genes in even higher level biological units, like functional pathways. Such 'network approach' seems very promising, for it naturally accounts for the fact that the behavious of living systems emerge from the interactions of its elements.

On the whole, transcriptome-wide association studies are one of the methods through which the relationship between genome, transcriptome and disease can be investigated. They are already applied quite widely and hopefully in the future will enable us to understand and cure many diseases.

# *References*

## *Main Articles*


Gamazon, Eric R. et al. (Sept. 2015). 'A gene-based association method for mapping traits using reference transcriptome data'. In: *Nat. Genet.* 47.9, pp. 1091–1098. DOI: 10.1038/ng.3367.

Gusev, Alexander, Arthur Ko, et al. (Mar. 2016). 'Integrative approaches for large-scale transcriptome-wide association studies'. In: *Nat. Genet.* 48.3, pp. 245–252. DOI: 10.1038/ng.3506.

Gusev, Alexander, Nicholas Mancuso, et al. (Apr. 2018). 'Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights'. In: *Nat. Genet.* 50.4, pp. 538–548. DOI: 10.1038/s41588-018-0092-1.


## *Further Reading*


Visscher, Peter M. (2008). 'Sizing up human height variation'. In: *Nat. Genet.* 40.5, pp. 489–490. DOI: 10.1038/ng0508-489.

Lander, Eric S. et al. (2001). 'Initial sequencing and analysis of the human genome'. In: *Nature* 409.6822, pp. 860–921. DOI: 10.1038/35057062.

Venter, J C and Et al. (2001). 'The sequence of the human genome'. In: *Science (80-. ).* 291.5507, pp. 1304–1351. DOI: 10.1126/science.1058040.

Visscher, Peter M., Matthew A. Brown, et al. (2012). 'Five years of GWAS discovery'. In: *Am. J. Hum. Genet.* 90.1, pp. 7–24. DOI: 10.1016/j.ajhg.2011.11.029.

Visscher, Peter M., Naomi R. Wray, et al. (2017). '10 Years of GWAS Discovery: Biology, Function, and Translation'. In: *Am. J. Hum. Genet.* 101.1, pp. 5–22. DOI: 10.1016/j.ajhg.2017.06.005.

Klein, Robert J et al. (2005). 'Complement factor H polymorphism in age-related macular degeneration.' In: *Science (80-. ).* 308.5720, pp. 385–389. DOI: 10.1126/science.1109557.

Emes, Richard D. et al. (2003). 'Comparison of the genomes of human and mouse lays the foundation of genome zoology'. In: *Hum. Mol. Genet.* 12.7, pp. 701–709. DOI: 10.1093/hmg/ddg078.

Gilad, Yoav, Scott A. Rifkin, and Jonathan K. Pritchard (2008). 'Revealing the architecture of gene regulation: the promise of eQTL studies'. In: *Trends Genet.* 24.8, pp. 408–415. DOI: 10.1016/j.tig.2008.06.001.

Li, Yang I. et al. (2016). 'RNA splicing is a primary link between genetic variation and disease'. In: *Science (80-. ).* 352.6285, pp. 600–604. DOI: 10.1126/science.aad9417.

Nicolae, Dan L. et al. (2010). 'Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS'. In: *PLoS Genet.* 6.4. DOI: 10.1371/journal.pgen.1000888.

Manolio, Teri A. et al. (2009). 'Finding the missing heritability of complex diseases'. In: *Nature* 461.7265, pp. 747–753. DOI: 10.1038/nature08494.

Kong, Augustine et al. (2012). 'Rate of de novo mutations and the importance of father-s age to disease risk'. In: *Nature* 488.7412, pp. 471–475. DOI: 10.1038/nature11396.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103.

Tibshirani, Robert (1996). 'Regression Selection and Shrinkage via the Lasso'. In: *J. R. Stat. Soc. B* 58.1, pp. 267–288. DOI: 10.2307/2346178.

Zou, Hui and Trevor Hastie (2005). 'Regularization and variable selection via the elastic net'. In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.2, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

Zhou, Xiang, Peter Carbonetto, and Matthew Stephens (2013). 'Polygenic Modeling with Bayesian Sparse Linear Mixed Models'. In: *PLoS Genet.* 9.2. DOI: 10.1371/journal.pgen.1003264.

Clarke, Geraldine M. et al. (Feb. 2011). 'Basic statistical analysis in genetic case-control studies'. In: *Nat. Protoc.* 6.2, pp. 121–133. DOI: 10.1038/nprot.2010.182.

Visscher, Peter M., William G. Hill, and Naomi R. Wray (Sept. 2008). 'Heritability in the genomics era–concepts and misconceptions.' In: *Nat. Rev. Genet.* 9.4, pp. 255–266. DOI: 10.1038/nrg2322.

Yang, Jian et al. (2011). 'GCTA: A tool for genome-wide complex trait analysis'. In: *Am. J. Hum. Genet.* 88.1, pp. 76–82. DOI: 10.1016/j.ajhg.2010.11.011.

Auton, Adam et al. (2015). 'A global reference for human genetic variation'. In: *Nature* 526.7571, pp. 68–74. DOI: 10.1038/nature15393.

Pasaniuc, Bogdan et al. (2014). 'Fast and accurate imputation of summary statistics enhances evidence of functional enrichment'. In: *Bioinformatics* 30.20, pp. 2906–2914. DOI: 10.1093/bioinformatics/btu416.

Griffiths, Anthony et al. (2007). *Introduction to genetic analysis*.

Mackenzie, Susan M. et al. (1999). 'Mutations in the white gene of Drosophila melanogaster affecting ABC transporters that determine eye colouration'. In: *Biochim. Biophys. Acta - Biomembr.* 1419.2, pp. 173–185. DOI: 10.1016/S0005-2736(99)00064-4.

Hunter, Philip (2009). 'Extended phenotype redux. How far can the reach of genes extend in manipulating the environment of an organism?' In: *EMBO Rep.* 10.3, pp. 212–215. DOI: 10.1038/embor.2009.18.

Dawkins, Richard (1982). *The extended phenotype: the long reach of the gene*.

Alberts, Bruce et al. (2014). *Molecular Biology of the Cell 6e*.

Ardlie, Kristin G. et al. (2015). 'The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans'. In: *Science (80-. ).* 348.6235, pp. 648–660. DOI: 10.1126/science.1262110.

Battle, Alexis et al. (2014). 'Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals'. In: *Genome Res.* 24.1, pp. 14–24. DOI: 10.1101/gr.155192.113.

Dunham, Ian et al. (2012). 'An integrated encyclopedia of DNA elements in the human genome'. In: *Nature* 489.7414, pp. 57–74. DOI: 10.1038/nature11247.

Ernst, Jason et al. (2011). 'Mapping and analysis of chromatin state dynamics in nine human cell types'. In: *Nature* 473.7345, pp. 43–49. DOI: 10.1038/nature09906.

Iacobuzio-Donahue, Christine A. et al. (2003). 'Highly Expressed Genes in Pancreatic Ductal Adenocarcinomas: A Comprehensive Characterization and Comparison of the Transcription Profiles Obtained from Three Major Technologies'. In: *Cancer Res.* 63.24, pp. 8614–8622. DOI: 10.1126/science.1058040.

Lappalainen, Tuuli et al. (2013). 'Transcriptome and genome sequencing uncovers functional variation in humans'. In: *Nature* 501.7468, pp. 506–511. DOI: 10.1038/nature12531.

Lonsdale, John et al. (2013). 'The Genotype-Tissue Expression (GTEx) project'. In: *Nat. Genet.* 45.6, pp. 580–585. DOI: 10.1038/ng.2653.

Ramasamy, Adaikalavan et al. (2014). 'Genetic variability in the regulation of gene expression in ten regions of the human brain'. In: *Nat. Neurosci.* 17.10, pp. 1418–1428. DOI: 10.1038/nn.3801.