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

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

Supervisor: Prof. Paolo Provero

Candidate: Federico Marotta

TESI DI LAUREA, 18 LUGLIO 2018

Contents

List of Figures

List of Tables

Todo list

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

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

Abstract

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

Introduction

https://www.ebi.ac.uk/gwas/home

https://www.broadinstitute.org/news/after-decade-genome-wide-association-st

Genotype => Expression => Phenotype <= Environment

Actually, environment can influence gene expression and epigenome as well. Also the genome, for instance UV rays cause mutations. Most of the time the environmental effects are random, but not always (*e.g.* UV rays, smoking...).

somewhere: Height and most other quantitative traits are influenced by many variants of small effects.

gamazon2015:Gwas have found many associations, but a large sample size is needed.

other limit of gwas: they study single variants, but sometimes the disease manifest only when there is a certain *combination* of variants.

gamazon2015: Gwas on their own are not enough (cite https://www.nature.com/articles/nature08494). in particular, there is a missing link between the variant and the disease: how (not why, *how*) does the variant make one individual more susceptible to a disease? It is not true that the nearest gene is always involved.

fine mapping? (it may be necessary also for TWAS.)

gamazon2015: Many SNPs are found in regulatory regions, as evinced by the fact that they overlap with DNaseI sites (is this true? read Gusev, A. et al. Regulatory variants explain much more heritability than coding variants across 11 common diseases. bioRxiv 004309 (21 April 2014).), and that they often are found in eQTL (see Nicolae, D.L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet.

6, e1000888 (2010).)

gamazon2015: eQTL mapping has shown that intermediate phenotypes, especially gene expression, are important. New projects are generating huge amounts of epxression data: ENCODE, GTEx, GEUVADIS, to name a few.

gamazon2015: they provide a way to aggregate SNPs in a biologically meaningful way, since they combine those variants that influence the expression of a gene. In general, either the phenotype is influenced by a small number of variants (one in the extreme case of mendelian diseases) with a large effect, or by a large number of variants of small effect. PrediXscan should districate this entanglement.

marker1 marker2 } => gene expression => phenotype marker3 marker4

gamazon2015: multiple testing problems reduced.

Heritability

http://www.cureffi.org/2013/02/04/how-to-calculate-heritability/

Many traits vary among the individuals in a population: height and hair colour are obvious ones, but also the number of fingers can be different in some pathological cases (see the amish communities in the USA). The heritability of a trait is the proportion of variance which can be explained with the genetic variance among individuals.

Genetic variance where? only at the loci associated to the trait? It should be so, otherwise we underestimate heritability.

There are two definitions of heritability:

- 'narrow sense heritability', h^2 is the heritability due to additive genetic factors.
- 'broad sense heritability', H^2 is the heritability due to all genetic factors, taking into account dominance and gene-gene interactions.

According to the additive model, one individual's having m alleles (0, 1 or 2) influences the phenotype of a factor ma, where a is the effect of the allele. For instance, if you have 1 copy of allele

a then your height increases of 1 cm, and if you have 2 copies it increases of 2 cm. In the additive model, each allele and genotype is independent of the others.

There are many ways to estimate narrow-sense heritability. One is in selective breeding, where $R = h^2S$: we start from a population with mean λ, select a subpopulation with mean μ, then take the offspring of the subpopulation, which will have mean μ_1. R is the difference between the mean of the offspring and the mean of the original population (*i.e.*, $\mu_1 - \lambda$), that is a measure of the success of selection; S is the difference between the mean of the selected subpopulation and the mean of the original population (*i.e.* $\mu - \lambda$), that is a measure of the selective pressure applied, if you want.

Another way to interpret heritability in the narrow sense is the following. Make a plot of parents heights vs. offspring heights. If there is perfect heritability, the height of a son is equal to the average heights of the two parents, so the plot will be a straight line y=x. In general, h^2 is the slope of the regression line.

One way to get rid of environmental effects is to compare monozygotic twins with dizygotic ones. MZ twins share the same environment and the same genotype, whereas DZ twins share the same environment but have different genotypes (albeit pretty similar).

Wikipedia

We assume that $P = G + E$, where P = phenotype, G = genetics, E = environment. Phenotypic variance can be expressed as follows:

$$Var(P) = Var(G) + Var(E) + 2Cov(G, E) \tag{1}$$

$$H^2 = Var(G)/Var(P) h^2 = Var(A)/Var(P) \tag{2}$$

Visscher 2006

Previous studies calculated genetic variance according to kinship (*i.e.* siblings share 1/2 of the genome, cousins 1/8, and so on). Visscher, instead, relies on the actual genotype of the samples, as assessed with markers. They had some 3000 pairs of siblings with genotype information. First, they calculated IBD sharing for each pair, then they calculated the heritability of height.

https://www.ncbi.nlm.nih.gov/books/NBK22001/

A trait is heritable if the variation of the trait in the individuals of a

population can be imputed to genes. Note that every gene plays a role in the development of a trait, but is the variation due to genes? For instance (by fmarotta), in a population of genetic clones there can be variability in a trait. In this case it is convenient to think about plants, for they are often propagated by vegetative methods, so each plant is genetically identical to the others; however, some plants may be better irrigated or manured, and hence grow taller. In this case, genes play a role in the "development" of height, but the variation in the trait is entirely due to environmental factors. The example made by the book I am following is this: 'there is no environment in which cows will speak. But, although the particular language that is spoken by humans varies from nation to nation, that variation is totally nongenetic'.

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 (or between siblings). In other words, if we have X = "phenotype of parent X" and Y(X) = "phenotype of offspring of parent X", then the plot of Y(X) should be a straight line with positive slope. This, however, is valid ONLY IF THE ENVIRONMENT IS NOT SHARED BETWEEN RELATIVES MORE THAN IT IS SHARED BETWEEN UNRELATED PEOPLE.

In order to estimate the heritability of a trait, we have to check whether individuals with different genetic markers have different phenotypes. If the phenotypes are different and the markers are different, then probably the markers are linked to genes that influence the phenotype (note that the markers are rarely involved directly in influencing the phenotype); on the other hand, if phenotypes differ but the markers are the same, then the trait is not heritable (otherwise it would have been inherited together with the markers).

There is also another sense in which heritability is not a measure of the role played by the genes during development: heritability is measured by taking into account all genetic variation, not variation in genes associated to the phenotype. This boggles me: we might see two individuals with different phenotypes and discover that they are genetically different, but maybe they differ at loci that have nothing to do with the phenotype! Perhaps, by using large cohorts, this phenomenon is less likely to appear, because different individuals will differ at different loci. It is also true that every locus influence every trait...

In experimental models, heritability is measured by artificial selection (see above).

https://www.nature.com/articles/ng0508-489

Combination of alleles may have specific effects both if they occur at the same locus (dominance) or at different loci (epistasis).

There is a nice picture showing the statistical power of association studies as a function of effect size and population size.

One could find genes near markers associated to a phenotype and look for overrepresented pathways.

'The main conclusion emerging from the current studies is that GWAS are able to robustly identify common variants that are associated with height but that the effect sizes of individual variants are small, so that very large sample sizes are needed to detect associations reliably. Single laboratories are unlikely to have sufficient sample sizes to do powerful studies on their own, and the trend in human complex trait mapping has been to create consortia of research groups and even consortia of consortia.'

At the same time, among the full sequences now available there are so many variants that trying to associate them with anything is very difficult. Statistics is not enough in this case.

https://www.nature.com/articles/nrg2322

Heritability deals with the old nature-nurture debate, in particular with how offspring resemble parents.

P = G + E (taking account of sex, age and other covariates while defining varP)

varP = varG + varE (assuming there is no genotype by environment covariance. there would be covariance if intelligent parents would provide an intelligence-stimulating environment for their offspring, or if cattle would be fed according to production. Also, the interaction between genotype and environment is neglected, i.e. when the effect of the genotype depends on the environment. they are ignored because they are difficult to evaluate.)

H2 = var(G) / var(P)

varG = varA + varD + varE (additive, dominant, epistatic effects. covariates are assumed 0)

h2 = varA / varP

I have not understood this part:

'in a non-inbred population, half of the additive genetic variance is between families and half is within families. This implies that

for a trait such as adult height in human populations, with a heritability of 0.8 and a standard deviation of approximately 7 cm in the population, the standard deviation of height in adult offspring around the mean value of the parents is 5.4 cm ($= sqrt[7^2(1^{\sim} * 0.8)]$), which is not much smaller than the standard deviation in the entire population. Hence, tall parents have on average tall children, but with a considerable variation around the parental mean.'

Wait a minute: h^2 because the variance is a squared thing! h would correspond to the standard deviation.

'Because individuals transmit only one copy of each gene to their offspring, most relatives share only single or no copies that are identical by descent (IBD) (the most important exceptions are identical twins and full siblings (sibs)), and dominance and other non-additive genetic effects that are based on sharing two copies do not contribute to their phenotypic resemblance. This is why the selection response and correlation of most relatives depend on h2 and not H2, and why h2 is the usual parameter.' That means that most people are heterozygous.

Breeder's equation: R = h2 S.

Methods

In this section I shall describe some well established methods which were applied in some of the papers described in the thesis.

Regression

Regression usually consists of three-steps:

1. Assumption making, where one chooses the parameters of the model and stuff.
2. Fitting of the model on a training dataset.
3. Prediction on a testing dataset.

Linear regression[1] is used to model 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

[1] **James2013**

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \,, \tag{3}$$

where the β'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 ϵ –a random error. The purpose of simple linear regression is first to fit the model, *i.e.* to find the values of the coefficients that better describe the relationship between X and Y in a training dataset, and then to apply the model to make predictions of Y on a testing dataset with known $X's$. The fitted model, where the parameters are estimated, is usually represented as follows:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j \,. \tag{4}$$

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 an x, we could estimate an $\hat{y}$; geometrically, $\hat{\beta}_0$ is the y-intercept of the regression line and $\hat{\beta}_1$ is its slope.

Ridge regression[2] is a regularisation method which allows to reduce the dependance 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

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

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 λ is properly chosen.

[2] **James2013**

Lasso regression[3] 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

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \,. \tag{6}$$

[3] **James2013**

Elastic net regression[4] combines the best features of ridge and lasso regression by introducing two penalty terms:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \alpha\sum_{j=1}^{p}\left(\beta_j\right)^2 + (1-\alpha)\sum_{j=1}^{p}\left|\beta_j\right| \,, \tag{7}$$

with α between 0 and 1. Its main advantage is that it works well when $p > n$, for it can potentially group correlated predictors and

[4] **Zou2005**

select only one representative variable for each group.

LOGISTIC REGRESSION[5] , 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.

[5] **James2013**

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

$$Y = P(Z = 1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} . \tag{8}$$

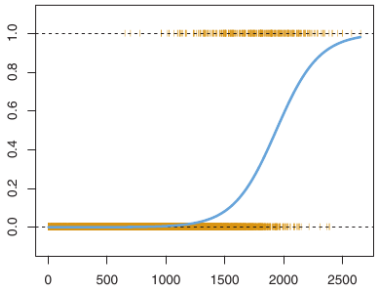


Figure 1: An example of logistic model. Adapted from James etal 2013, "An introduction to statistical learning".

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

$$\log\left(\frac{\hat{y}}{1-\hat{y}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x , \tag{9}$$

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.

Seminal work: TWAS from individual-level data

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 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; some of these consortia, whose datasets have been exploited by Gamazon *et al.*, are as follows.

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

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

GEUVADIS. This projects

DGN.

Braineac.

Method

first step: gene expression is decomposable in three components: genetically regulated expression (GReX), phenotype-influenced expression, and an environmental component. The phenotype can influence gene expression.

An additive model trained on reference transcriptome datasets finds for each SNP the coefficents of which gene expression is altered by that SNP, i.e. it says that SNP rs483920482905, when present in an individual, alters the expression of gene XXXX by a factor 1.5. Clearly, the training dataset must contain both genome and transcriptome data. Afterwards, the GReX is predicted in indivuduals for which only the genome sequence is available. They thus generated predictDB.

'(For specific results on the disease phenotypes analyzed here, we used logistic regression with disease status.)' by fmarotta: they used a binary variable, but what about using the liability to the disease, which is continuous? see visscher 2008.

$$T = \sum_{k=1}^{M} w_k X_k + \epsilon \tag{10}$$

T is the expression of a gene, w_k is the weigth of SNP k in influencing the expression of that gene, and X_k is the number of reference alleles of SNP k (I guess X_k is the sum of the alleles in all the individuals in the dataset).

consideration by fmarotta: there are other models available, for instance one could account for the penetrance, or use a dominant (recessive) model, and so on.

Important (by fmarotta): in linear regression, each regressor is considered independent of the others, but is that so? I think often a phenotype can depend on the *combination* of SNPs.

(also by fmarotta): is it possible to use PCA to find which SNPs are most relevant in influencing a phenotype?

In the second phase, the predicted GReX is correlated (with linear regression, logistic regression, Cox, or Spearman (the latter is non-parametric). They used logistic regression for the results discussed in this article.

limits: there is an attenuation bias because of the error in the estimation of the GReX.

features

The rationale for everything is that often SNPs influence a phenotype by altering gene expression (i.e. they have regulatory roles), as stated in this article: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000888

main advantages:

- directionality, possibility to get insights on mechanisms.
- small multiple-testing burden.
- possibility form functional units (e.g. basing on known pathways).
- possibility to reanalise gwas data (only the genome is needed).

Predicting the transcriptome

They tried lasso, elastic net and another thing which turned out to be pretty bad and was abandoned. Lasso is a regression analysis whereby the number of regressors is diminished: some covariates are disregarded (their coefficient becomes 0), and in general only one covariate is selected from a set of highly correlated covariates. This is achieved by imposing the constraint that the sum of all the coefficients (in absolute value) must be less than a fixed amount t (see wikipedia). Lasso explores a different path from ridge regression, where the sum of the squares of the coefficients must be less than an amount. Elastic net is an improvement of lasso, where an additional constraint is imposed: the linear combination of lasso and ridge constraints. The relative weight of the two constraints is moduled by the alpha coefficient: the linear combination is alpha*L1 + (1-alpha)*L2.

In general lasso (elastic net) is used 1) because the number of predictors is HUGE, especially if compared to the number of samples; 2) to avoid overfitting.

They chose to use elastic net and used tenfold cross-validation (*i.e.* they looked at the R square of estimated GReX vs observed expression).

They also computed the heritability of gene expression in DGN and claim that heritability is an upper bound to how well the trait can be associated to the genotype. This makes sense, because if the variance in a trait is entirely due to the environment, the association study will not find anything.

Heritability is defined as the proportion of phenotypic variance that is due to genetic variance

The average heritability calculated in DGN was 0.153, while the average tenfold cross-validated prediction R^2 value for elastic net was 0.137, so: not bad. Does this make sense? I think so, because heritability can be interpreted as the slope of the line of the predicted variable as a function of the predictor (*e.g.* offspring height as a function of parent height). In this case the "predictor" is the phenotype of the training sample, and the "predicted" is the phenotype in the testing sample. A high heritability says that parent height can predict offspring height, and that this predictability is due to genetic factors; here, the authors are trying to predict a phenotype (gene expression, but it is hardly relevant) from a genotype, which in turn was attributed scores according to the phenotype. The linear model (actually elastic net) they made, mimics the actual process of development: they go from one genotype to a phenotype, attributing to each element of the genotype (*i.e.* to each SNP) a coefficient that measures how much that element is involved in the manifestation of that phenotype. Then, let us imagine that the training people have offspring, and that the offspring is the "testing people": many of the SNPs will be transmitted from parents to offspring. offspring have a similar genotype to their parents (abstracting things, offspring are merely people which have a genome similar to their parents), but do they have also a similar phenotype? If the trait is heritable, yes. In yet other words, if the trait is heritable, people with a simila genotype (*i.e.* relatives) will have a similar phenotype (h2 is precisely the correlation between the phenotypes in parents-offspring); in a sense, in this paper the testing people (those in the gwas) are relatives of the training people (those in the reference transcriptome study like GEUVADIS). All this is expressed in figure 3.

An important thing that perhaps I did not say before is that previously they had imputed the SNPs of the DGN people. They used both 1000 Genomes and HapMap Phase 2 to impute, and achieved similar results, therefore they chose to restrict the imputation to hapmap to save computation time.

They then tested their models to see wether, given a genotype, they could predict the expression. That is to say, they predicted expression from a set of genotypes and then confronted it with the real expression. They used GEUVADIS and GTEx as tests. In figure 4 they show that the correlation between predicted and actual expression in this separate cohort is very different from the expected correlation under the hypothesis that the two vectors (predicted and actual expression) are independent. In gray there is a 45-degree line: if the point lay on this line, then on the two axes there are identical variables. on the y axis there are the quantiles of observed R2 (a quantile is the percentage of points below a given value), while on the x axis there are the quantiles of expected R2. A point is produced when the two quantiles are equal. For instance,

if the 10th quantile in the observed R squareds have an R2 of 0.2, what is the R2 of the expected R2 at the 10th quantile? In other words, the curve is parametric. This Q-Q plot is used to check whether two populations are similar.

They also note that the same situation arises for different tissues.

In fig 5 they present some example genes.

they also made a linear model for trans eQTL, with linear regression, but it had a poorer predictive power, so they resolved to use local SNPs only.

Application of PrediXcan to WTCCC

At last, they apply their method. they used DGN whole-blod elastic net prediction models to predict the expression in the WTCCC cohort, then correlated the predicted GReX with the disease status.

An interesting consideration (by fmarotta) is that many of the genes that were associated to the disease status were in the HLA or MHC region; also, the most significant results were for autoimmune diseases (this can be due to the fact that also in the WTCCC work the most significant results were for autoimmune diseases. Nay, I think that the fact that the two studies (WTCCC and gamazon) have found the same result is due to a common underlying cause, which is the same for which most GWAS hits are in the MHC region.) Think more about this.

They made a manhattan plot and another quantile-quantile plot. Also a gwas enrichment.

Some genes were associated with multiple autoimmune disease. Question by fmarotta: what determines which disease you have if the expression of that gene is altered in you? environment? gene expression level? Anyway, this is an example of the complexity of the situation: the relationship between genotype and phenotype is not biunivocal at all. the authors say that lower expression of dclre1b was associated with rheumatoid arthritis and T1D, whereas higher expression with crohn's disease.

An advantage of predicscan is that it provides directionality: we know if higher or lower expression is associated with the disease. See the example of ERBB3.

Globally, many genes were previously reported or fell near reported genes. Or: they were in the MHC.

Using less stringent significance thresholds, they found the same high enrichments of reported genes among the results of predicscan. This suggests that there are many false-negatives at the higher thresholds. The method is not so powerful afterall. They found also two completely novel genes.

Finally, they compared their method with two other gene-based tests: vegas and skat. In the quantile-quantile plot, predixcan was the best in the tail-end.

Discussion

Why gene expression? It is the most direct phenotype (indeed, sometimes we speak about 'extended phenotypes'; gene expression can also be viewed as an intermediate phenotype), it is heritable, and virtually all the other phenotypes depend on it. Moreover, it is easy to measure.

fmarotta:Genes can do few things: either they bind proteins whith a structural or regulatory function (and when it is structural, it can be regulating: TAD are coregulated), or they are transcribed, starting a series of biochemical reactions that ultimately lead to functional molecules, be they RNA or proteins. The complexity stems from the interactions of many genes together and with the environment

Limits:

- the prediction of gene expression can be biased, and some models, namely a combination of K nearest neighbour (KNN), elastic net, and the use of genomic annotation may perform better.
- (https://www.cell.com/ajhg/fulltext/S0002-9297(18)30108-3?code=cell-site) Genetic variation does not alter only gene expression. There can be *trans*-acting effects, where a SNP alters how a gene (be it a TF or a miRNA) modulates the expression of others, without altering the expression of the modulator gene itself. Moreover, a SNP can have effects on splicing, transcription start or end site or other RNA editing processes, without altering the expression of the gene.

One of the main advantages is that it is economic: one only needs existing data, therefore many existing GWAS dataset can be reanalysed 'for free'.

Another virtue is that predicscan provides directionality, hinting at potential strategies to cure disease, *e.g.* if a gene's upregulation is

linked to a disease, then a drug may be developed to downregulate it. (by fmarotta: It probably woul have no effect whatsoever because of compensatory effects.)

Multiple testing: here they have used bonferroni, which is pretty conservative. They only corrected gene-based tests of association, and not SNP-based ones, because the gene-based association is the last step, and because bonferroni is conservative. (by SNP-based, I do not know if they mean SNP-geneexpression association or SNP-trait association performed in a classical GWAS.)

They do not claim causality, for SNPs may contribute both to expression and to other things, and it may be that the other things are the cause of the disease, not gene expression.

They state that their method provides insights into gene regulation and directionality.

by fmarotta: Actually, they do not prove that a SNP associated to gene expression regulates the gene: they only say that variation between individuals at that locus results in variation in gene expression. Or do they?

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

Abstract

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

Introduction

The *rationale* that lies behind the association of gene expression to phenotype is that many genetic variants influence traits by altering the regulation of the expression of some genes. Despite the strength of this argument, publications of studies in which both transcriptomic and phenotypic data are investigated simultaneously lag behind those of simple GWAS studies, for at least two reasons: first, although the cost of sequencing nucleic acids has been sharply decreasing for over a decade (Figure 2), it can become quite an expensive technology if applied to cohorts of tens of thousand samples, such as those of a typical modern GWAS; secondly, every tissue shows a different pattern of expressed genes, and to choose the right tissue to analyse for each phenotype is not always a trivial matter.

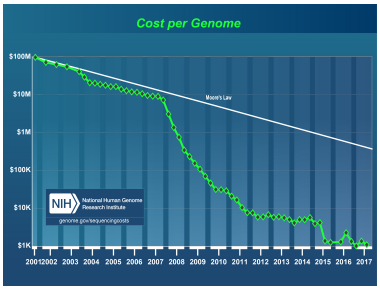


Figure 2: The decrease in the cost of genome sequencing; the same technology is used to sequence RNA. https://www.genome.gov/sequencingcosts/

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

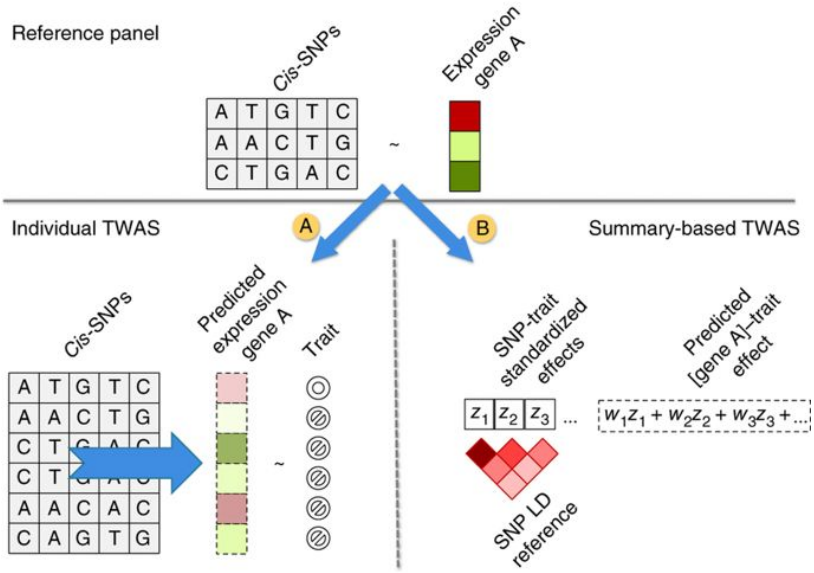
small-effect variants, many new methods are being developed. One of such methods is PrediXcan, with which we dealt in the previous section, but it is by no means the only one. In particular, in 2016 a new approach has been proposed which does not need individual-level data, but only summary association statistics[6] from a GWAS, which is an important advantage since, normally, only the summary-level data of a study is publicly available due to privacy concerns.

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

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

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

through correlation



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

it doesn't mean it is robust

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

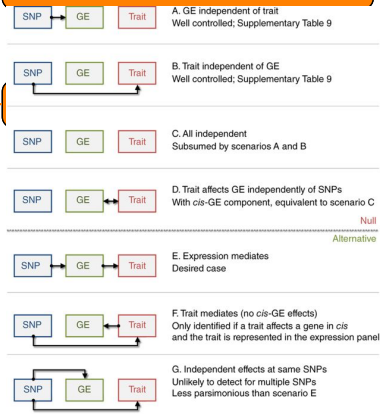


Figure 3:

Technically, here the weight of a SNP is its effect size.

The models 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 a significant improvement. Moreover, they reanalysed an existing dataset of a small-cohort lipid GWAS, finding that most of the novel associations they obtained had been previously reported in a large-cohort GWAS, and implying that their method is statistically more powerful than SNP-based approaches. Finally, they applied their method

additional sections: heritability and genetic heterogeneity.

Training of the expression model

The accuracy of the prediction of a gene's expression cannot be greater than the heritability of the expression of that gene itself. For instance, let us consider height[7] and suppose that this trait is normally-distributed in the population. If every individual has the same alleles at the same height-associated loci, the genetic variance in that population will be 0, and the heritability for height would consequently be 0 as well. In such circumstances, it is not possible to predict height using the *cis*-genetic component of gene expression, because there is no such component: the differences in the individuals' heights depend only upon the environment. Theoretically, the effect of the environment could be decomposed in a deterministic one and a random component, and the deterministic one could be associated with the trait; however, it is notoriously difficult to quantitatively measure the effect of the 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.

heritability: where to put this discussion?

[7] Height is a typical, quantitative trait, and we choose to base our discussion on it because it is also quite easy to visualise; nevertheless, everything is still valid for gene expression, which is another quantitative trait.

there are still effects of chance

In order to predict a quantitative trait from the genotype of the individuals, a sample for which both gene expression and genotype data are present is necessary. The authors collected these data from three data sets: METISM, YFS and NTR.

describe data sets

From such individuals' data, the heritability of the expression of each gene was computed. 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 , 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

citation 1000 genomes

were based on the 6,924 *cis*-heritable genes (Figure 4). Restricting the analysis to *cis*-SNPs greatly increases the statistical power of the study, for the number of predictors of gene expression is quite small; as previously explained, the multiple testing burden is also decreased.

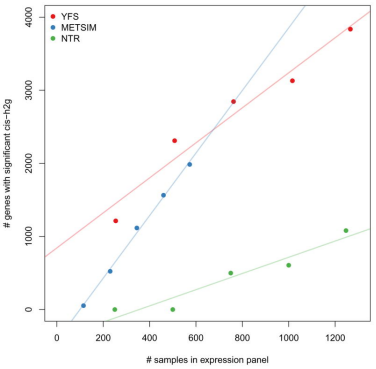
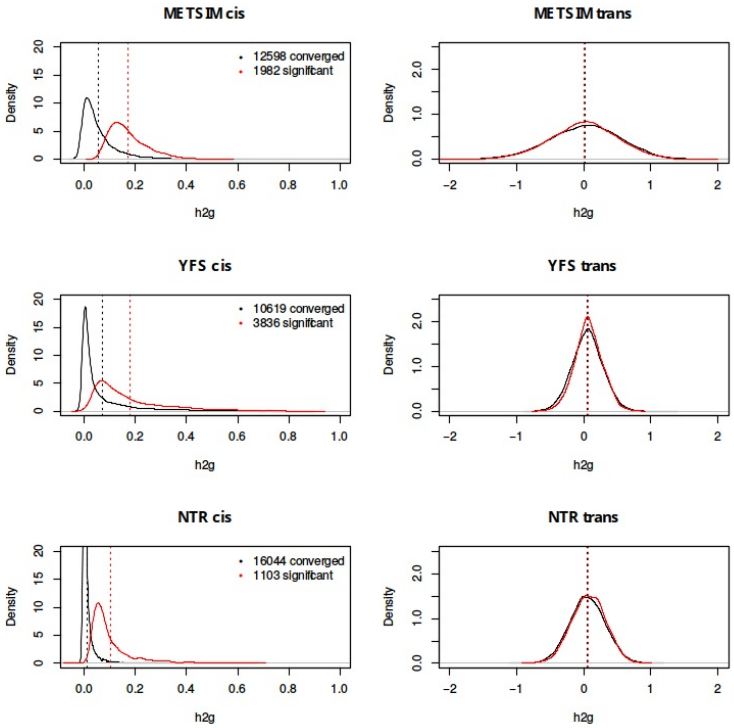


Figure 4: The 6,924 heritable genes, distributed according to their origin

Figure 5: Heritability distribution.



Having computed heritability, a statistical model could be trained to predict gene expression from genotype data. Two different models, starting with the *cis*-SNPs, were employed: the first was a best linear unbiased model (BLUP) and the second a Bayesian model (BSLMM); 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.

What is the trick? every gene is averaged from all the people? I think the trick is that the weigh of a snp is its effect size.

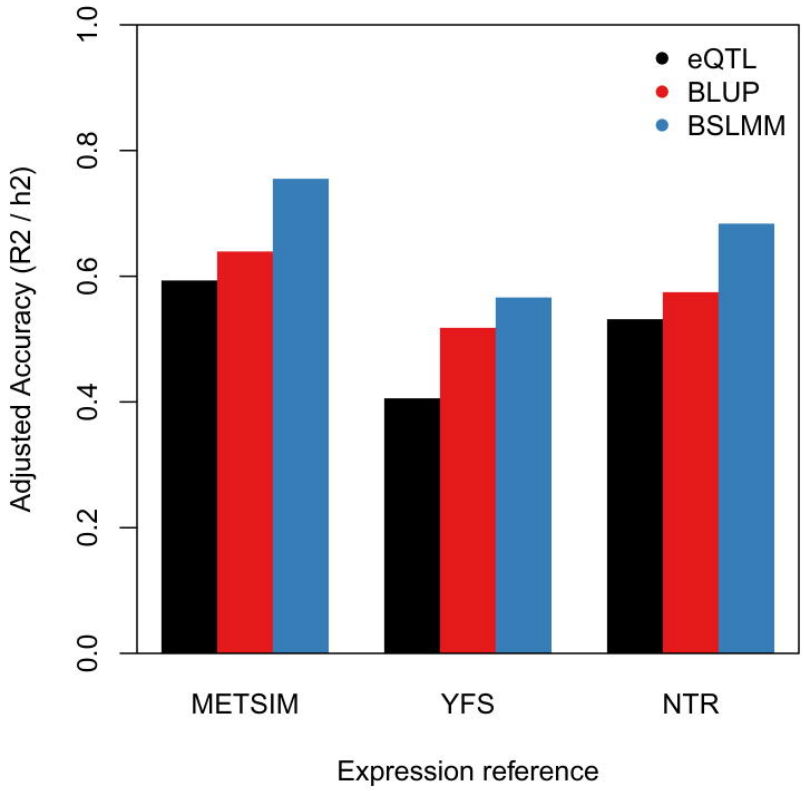


Figure 6: BSLMM performs better

Simulations

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

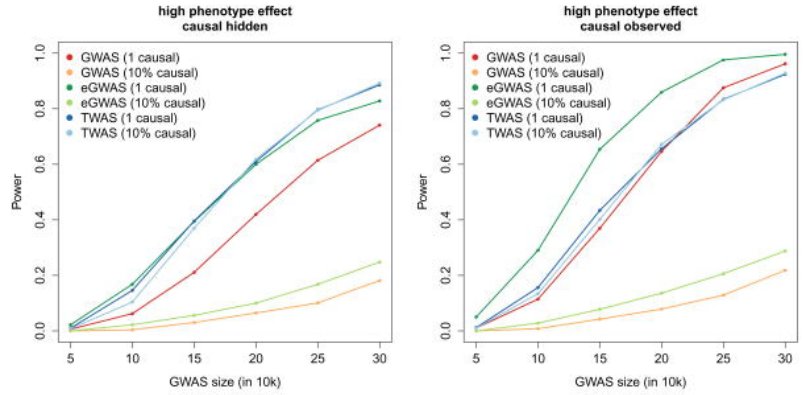


Figure 7: Powerful!

I am skipping three paragraphs where they 1) compare TWAS with coloc 2) and another thing; 3) investigate on the effect of a larger

sample size for expression.

small-cohort GWAS

In the previous section, the authors showed that predicting gene expression from the summary-level statistics of a GWAS is feasible; here they show that associating gene expression to the trait is useful. In other words, they split their approach in two parts and validated them separately.

A previous study had reported all the 697 known loci associated to height. For each locus, Gusev etal selected a single causal gene according to three different strategies:

The third strategy seems a bit circular: they select the best twas gene, and then associate its expression with the trait; but is the best gene not found already by its association with the trait?

I give up this section. I think it tries to say that performing twas on summary level or directly on real expression is identical.

900000 phenotypes

One of the most innovative features of this approach is its broad applicability. Indeed, its potential was unleashed on three GWAS which account for over 900,000 phenotype measurements of obesity-related traits[8]. They first imputed gene expression for the 6,924 genes whose expression is heritable, then associated such imputed expression to the trait, correcting for the multiple testing, and finding 665 significant gene-trait associations, 69 of which genes did not overlap any SNP which was reported by the original GWA studies.

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

'Averaging over the novel genes, the Z2 statistics from TWAS were 1.5x higher than the strongest eQTL SNP for the same gene(though this may be slightly inflated due to winner's curse).'

They used a permutation test. Allelic heterogeneity strikes back.

come back here when the methods are understood

Paragraph on the contribution to heritability of the associations. I think they say that if a gene is associated, it contributes to the heritability.

put a subsection on heritability in gamazon and a subsection on heretogeneity here

Paragraph where they used muther and another thing to train the expression models. They still found many of the associations. (only

the training changes, the three gwas summary are the same.)

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, 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 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.

cite obesity papers

Methods

Heritability computation

Application: schizophrenia

Application: schizophrenia

Bonus: medical description of schizophrenia

Putative regulatory mechanisms

Of the 157 genes associated to a schizophrenia, 42 were also associated to a chromatin state, and, in particular, 8 of the 42 genes were associated to a chromatin peak located in the promoter of the gene itself. First and foremost, (by fmarotta) the fact that a gene's expression level is associated both to a chromatin state and to a disease does not imply that the chromatin state is associated to the disease as well; however, it does suggest that the gene is regulated by a region where the chromatin peak was found (note that this is the reverse of what happens in the other TWAS, where the suggestion is that the gene is causal to the phenotype; the reason of this is that chromatin state comes before gene expression, and gene expression before phenotype manifestation.), and the fact that many chromatin peaks harboured GWAS hits for schizophrenia supports the hypothesis that GWAS hits are regulatory. Moreover, (by gusev) the fact that the majority of genes were associated to distal chromatin peaks suggests that they be regulated by enhancers.

They applied coloc again... skipping.

Regarding what I said previously, they did not perform a chromatin-wide association study only because the sample size was not large enough.

Examples

The authors used a scatter plot to represent loci associated to schizophrenia and chromatin state, reporting the Z score on the one hand (y axis), and the correlation between predicted expression and GWAS-QTL association.

The Z score is a measure of the association between GWAS and QTL.

References