Garvan Institute of Medical Research

Estimating heritability of gene expression

Associate Professor Joseph Powell
Director, Garvan-Weizmann Centre for Cellular Genomics
Deputy Director, UNSW Cellular Genomics Futures Institute

SISG - 2019

Contents

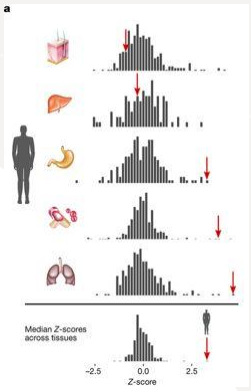
Garvan Institute of Medical Research

- What is heritability and why is it an important concept
- How can heritability be estimated
- What do we see in real data

Lets start with the following observation

Garvan Institute of Medical Research

- The expression levels of many (most) transcripts vary across individuals



RSP14

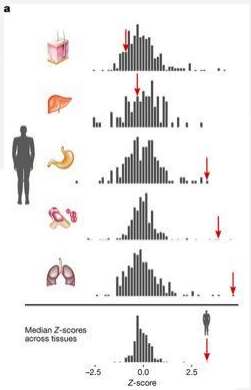
3

Lets start with the following observation

Garvan Institute of Medical Research

- The expression levels of many (most) transcripts vary across individuals

Why do they vary?



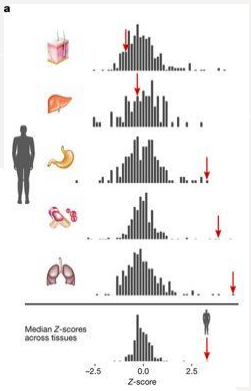
RSP14

4

Lets start with the following observation

Garvan Institute of Medical Research

- The expression levels of many (most) transcripts vary across individuals



RSP14

Why do they vary?

(1) Differences in the environment between individuals
(2) Technical variation in the sample collection, preparation and sequencing
(3) Stochastic variation
(4) Genetic variation between individuals

5

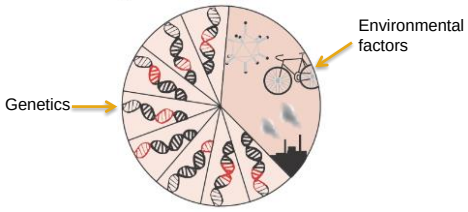
Variation

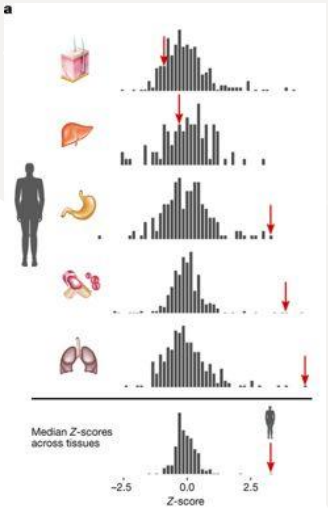
Garvan Institute of Medical Research

<u>Environmental</u>

Differences in the environment between individuals
Technical variation in the sample collection, preparation and sequencing
Stochastic variation

<u>Genetic</u>
Genetic variation between individuals





6

Definition

Garvan Institute of Medical Research

The proportion of variation in a phenotype (gene expression) that is explained/due to genetic variation

- Heritability: is a statistic that provides an estimate of how much variation in a trait in a population is due to genetic variation among individuals.
- Other causes of variation in a trait are characterized as environmental factors.

7

Definition

Garvan Institute of Medical Research

The proportion of variation in a phenotype (gene expression) that is explained/due to genetic variation

- Heritability: is a statistic that provides an estimate of how much variation in a trait in a population is due to genetic variation among individuals.
- Other causes of variation in a trait are characterized as environmental factors.

$$\text{Phenotype } (P) = \text{Genotypes } (G) + \text{Environment } (E)$$

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

8

Definition

Garvan Institute of Medical Research

The proportion of variation in a phenotype (gene expression) that is explained/due to genetic variation

- Heritability: is a statistic that provides an estimate of how much variation in a trait in a population is due to genetic variation among individuals.
- Other causes of variation in a trait are characterized as environmental factors.

Phenotype (P) = Genotypes (G) + Environment (E)

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

In a planned experiment Cov(*G*,*E*) can be controlled and held at 0. Thus,

9

Definition

Garvan Institute of Medical Research

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

There are different forms of Var(G).

H^2 is the broad-sense heritability.

Var(G) includes all genetic contributions to phenotypic variance including additive (A), dominant, epistatic, as well as maternal and paternal effects.

Additive only heritability is calculated,

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

10

What do we need to estimate heritability?

Garvan Institute of Medical Research

- What data do we need?
  - Genotypes – either directly measures, or inferred indirectly from relationships
  - Normalized gene expression levels
  - Covariates – what factors do we think will be correlated with parameters in our equation?
- In an experimental setting, what else do we need to consider?
  - Are the sample matched? (Look up MixUpMapper)
  - Population stratification

11

Twin Models

Garvan Institute of Medical Research

Heritability estimates in humans are commonly made using the resemblance between monozygotic (MZ) and dizygotic (DZ) twins.

MZ twins are genetically identical whereas DZ twins, on average, have 50% of their alleles identical by descent (IBD).

12

Twin Models

Garvan Institute of Medical Research

Heritability estimates in humans are commonly made using the resemblance between monozygotic (MZ) and dizygotic (DZ) twins.

MZ twins are genetically identical whereas DZ twins, on average, have 50% of their alleles identical by descent (IBD).

The correlation of mRNA transcript levels in MZ (rMZ) and DZ (rDZ) can be used to estimate the additive genetic contribution (VA) to phenotypic variance by;

$VA = 2(rMZ - rDZ)$.

The contribution of environmental variance (VE) can be estimated by subtracting rMZ from1 as in $VE = 1 - rMZ$ and the contribution of common environmental effects (VC) by $VC = rMZ - VA$.

13

Parent-offspring

Garvan Institute of Medical Research

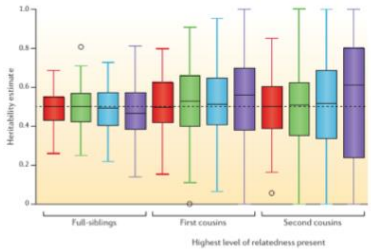
- Parents share 50% of their alleles IBD with their offspring. If we regress the transcript levels of an mRNA transcript measured in offspring again the levels measured in their parents then the slope of the regression (β) is equal to
- $\beta = cov(P,O)/Var(P) = \frac{1}{2} h^2$.
- In other words, the heritability can be estimated as $2 * \beta$. This method assumes no common environmental effects.

14

Shared genetics

Garvan Institute of Medical Research

By measuring actual genetic similiarity, rather than relying on expected similarity, we can obtains more precise estimate of h^2



Purple boxes are estimates using expected relatedness; red use actual relatedness (green, blue use less accurate measures of actual relatedness)

Even bigger benefit - we can use "unrelated" individuals

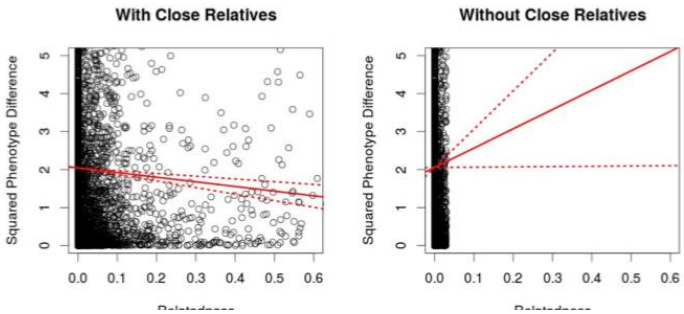
15

Estimating h^2 from unrelated individuals

Garvan Institute of Medical Research

In 2010, Jian Yang, Peter Visscher, et al. considered estimating heritability using only "unrelated individuals"

Why? Estimates of h^2 become less precise as number of close relatives in the sample decreases



16

Using unrelated individuals

Garvan Institute of Medical Research

However, using unrelated individuals has three key advantages:

Less of a problem that we ignore effects of common environment (and dominance / epistasis)

Can use GWAS data, so sample sizes are much larger than using family data

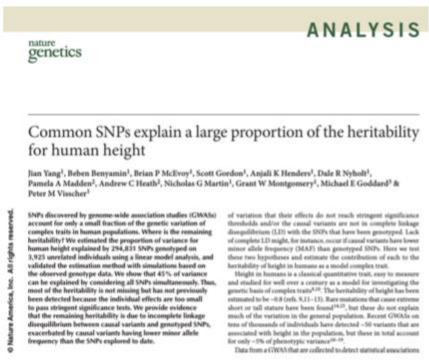
The resulting estimates, referred to as h^2_{SNP}, are estimates of "SNP heritability", the total variance explained by all SNPs

This area is referred to as **SNP-based heritability analysis**. The major software is GCTA; our software is LDAK

17

Unrelated individuals

Garvan Institute of Medical Research



ANALYSIS

nature genetics

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of complex traits in human populations. Where is the remaining heritability? We estimated the proportion of variance for human height explained by 294,831 SNPs genotyped on 3,925 unrelated individuals using a linear model analysis, and validated the estimation method with simulations based on the observed genotype data. We show that 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is not missing but has not previously been detected because the individual effects are too small to pass stringent significance tests. We provide evidence that the remaining heritability is due to incomplete linkage disequilibrium between causal variants and genotyped SNPs, exacerbated by causal variants having lower minor allele frequency than the SNPs explored to date.

of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped. Lack of complete LD might, for instance, occur if causal variants have lower minor allele frequency (MAF) than genotyped SNPs. Here we test these two hypotheses and estimate the contribution of each to the heritability of height in humans as a model complex trait.

Height in humans is a classical quantitative trait, easy to measure and studied for well over a century as a model for investigating the genetic basis of complex traits[6,7]. The heritability of height has been estimated to be ~0.8 (refs. 9,11–13). Rare mutations that cause extreme short or tall stature have been found[14,15], but these do not explain much of the variation in the general population. Recent GWASs on tens of thousands of individuals have detected ~50 variants that are associated with height in the population, but these in total account for only ~5% of phenotypic variance[16–19].

Data from a GWAS that are collected to detect statistical associations

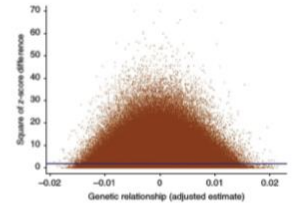


Figure 3 All pairwise comparisons contribute to the estimate of genetic variance. Shown are the squared z-score differences between individuals (Δy^2_{jk}) plotted against the adjusted estimates of genetic relationship (A'_{jk}). The blue line is the linear regression line of Δy^2_{jk} on A'_{jk}. The intercept and regression coefficient are estimates of twice the phenotypic variance and minus twice the genetic variances[23], respectively. The intercept is 1.98 (s.e. = 0.001), and the regression coefficient is −1.01 (s.e. = 0.27), consistent with estimates of the phenotypic and additive genetic variance of 0.990 and 0.505, respectively, and a proportion of variance explained by all SNPs of 0.51.

Estimating h^2_{SNP} using mixed model analysis with unrelateds found SNPs explain at least 45% of variation in height - over half the heritability

18

Unrelated individuals

Garvan Institute of Medical Research

For example, the GREML method (Yang *et al*. NG 2010, Powell *et al*. NRG, 2010) uses a linear mixed-effects model:

$y = g + e$, where y is a $nx1$ vector of normalized gene expression levels for a transcript;

g is n∗1 vector of random polygenic effects with $g \sim N(0, Var(G)\boldsymbol{A})$,

with $\boldsymbol{A}$ the genetic relationship matrix (GRM) estimated from common SNPs;

and is a $nx1$ vector of residuals with

e ~ N(0,Var(E)**I**), with **I** as the incidence matrix.

19

Resources

Garvan Institute of Medical Research

Related and unrelated individuals in the study design

PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses

Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham

plink - *(pngu.mgh.harvard.edu/ purcell/plink/)*

Unrelated individuals in the study design

GCTA

a tool for Genome-wide Complex Trait Analysis

GCTA - emphhttp://cnsgenomics.com/software/gcta/

20

What do we see in gene expression?

Garvan Institute of Medical Research

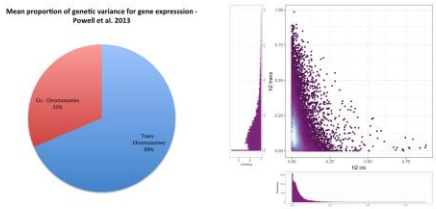
- There are >20,000 protein coding genes….

21

What do we see in gene expression?

Garvan Institute of Medical Research

- There are >20,000 protein coding genes….



The majority of heritability is located on *trans* chromosomes

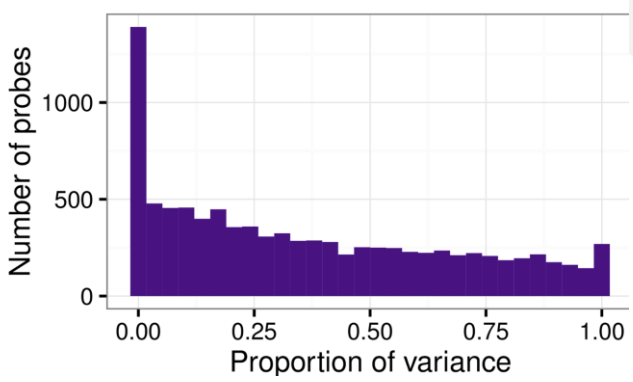
What about other forms of genomic architecture?

22

Not all genes are the same

Garvan Institute of Medical Research

- There is huge variance in the heritability of a given gene



What does this tell us?

23 Powell et al. Genome Research, 2014

What about other tissues?

Garvan Institute of Medical Research

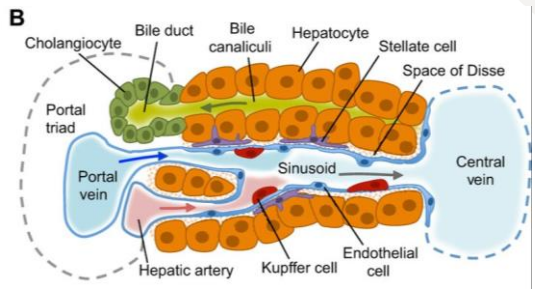
Table: Summary of the estimates of heritability for gene expression levels from large-scale studies.

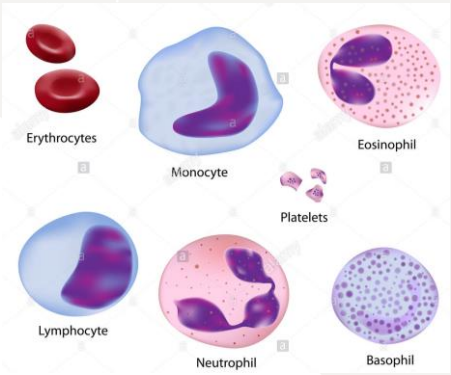
Study	Tissue	N Transcripts or Genes	Mean h^2	Method	Sample Size
Dixon *et al.*	LCLs	20,599	0.23	Sib pairs	400
Price *et al.*	Peripheral blood	18,735	0.16	Population IBD	687
Price *et al.*	Adipose tissue	19,099	0.24	Population IBD	496
Wright *et al.*	Peripheral blood	18,392	0.14	Twin model	2,752
Powell *et al.*	LCLs	9,555	0.38	Twin model	100
Powell *et al.*	Peripheral blood	9,555	0.32	Twin model	100
Powell *et al.*	Peripheral blood	17,994	0.24	Complex family	862
Grundberg *et al.*	Adipose tissue	23,596	0.26	Twin model	714
Grundberg *et al.*	Skin	23,596	0.16	Twin model	540
Grundberg *et al.*	LCLs	23,596	0.21	Twin model	718
Lloyd-Jones *et al.*	Peripheral blood	36,778	0.192	Population	2813

24

What about different cells in a tissue?

Garvan Institute of Medical Research





Do we expect heritability of gene expression to be the same in each cell type?

25

Garvan Institute of Medical Research

Thank You!

26