Genetic control underlying transcriptional variation

Joseph Powell

Computational and Single Cell Genomics, Institute for Molecular Bioscience

---

Summer Institute in Statistical Genetics

Brisbane, 14th Feb 2016

Genetic control of gene expression

Observation

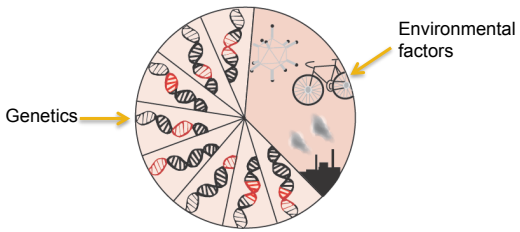
The expression levels of many (most) transcripts vary across individuals.
Why does it vary?

Genetic control of gene expression

Observation

The expression levels of many (most) transcripts vary across individuals.
Why does it vary?

- Differences in the environment between individuals
- Technical variation in the sample collection, prep and sequencing
- Stocatic variation
- Genetic variation between individuals



Heritability

Heritability

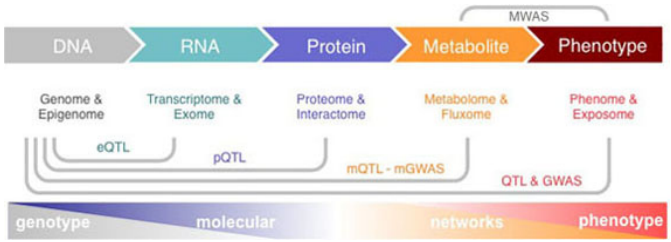
Heritability is a statistic that provides an estimate how much variation in a phenotypic trait in a population is due to genetic variation among individuals in that population. Other causes of measured variation in a trait are characterized as environmental factors, including measurement error.

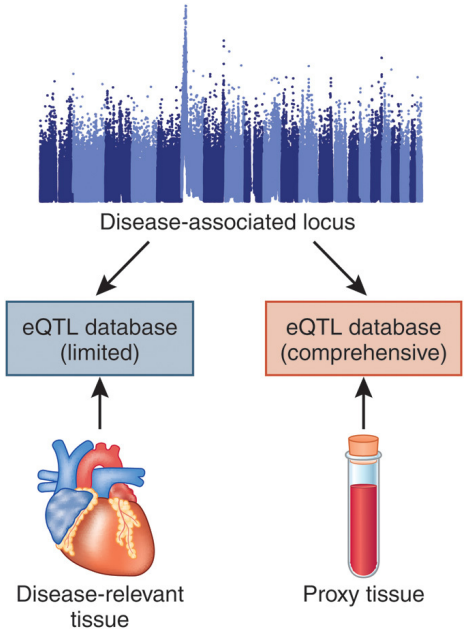
$$h^2 = \sigma_A^2 / \sigma_P^2$$

Why is heritability important?

Expression Quantitative Trait Loci

What is an eQTL?
Why are eQTL important (in the context of both biology and disease genomics)





Data

What data do we need?

- Genotype
- Normalised expression data
- Covariates

What do we need to consider with the data QC?

- Are the samples matched - look at MixUpMapper?
- Population stratification

Methods

Twin model

Twins: Heritability estimates in humans are most 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 (r_{MZ}) and DZ (r_{DZ}) can be used to estimate the additive genetic contribution (V_A) to phenotypic variance by; $V_A = 2(r_{MZ} - r_{DZ})$. The contribution of environmental variance (V_E) can be estimated by subtracting r_{MZ} from 1 as in $V_E = 1 - r_{MZ}$ and the contribution of common environmental effects (V_C) by $V_C = r_{MZ} - V_A$.

Methods

Parent-Offspring

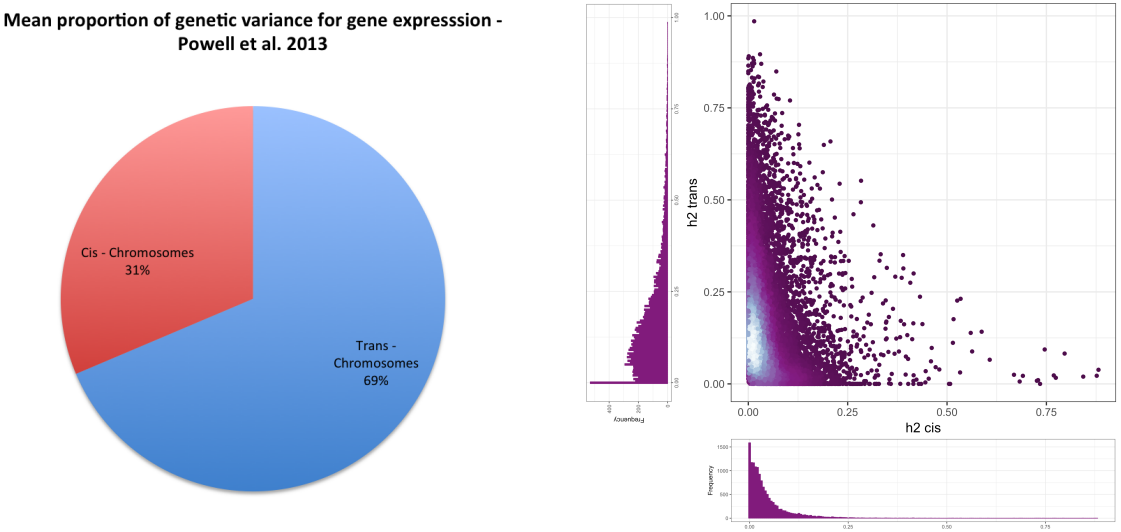
Parent-offspring: Parents share 50% of their alleles IBD with their offspring. If we regress the transcript levels of an mRNA transcript measured in offspring again the levels measured in their parents then the slope of the regression (β) is equal to $\beta = \frac{cov(P,O)}{Var(P)} = \frac{1}{2}\frac{V_A}{V_A}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

Methods

Population-based

For example, the GREML method presented in Yang et al. NG 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 $g \sim N(0, \sigma_g^2 \mathbf{A})$, with $\mathbf{A}$ the genetic relationship matrix (GRM) estimated from common SNPs; and $e$is a $nx1$ vector of residuals with $e \sim N(0, \sigma_{e_i}^2 \mathbf{I})$, with $\mathbf{I}$ as the incidence matrix.

What about partitioning across the genome?



The majority of heritability is located on *trans* chromosomes

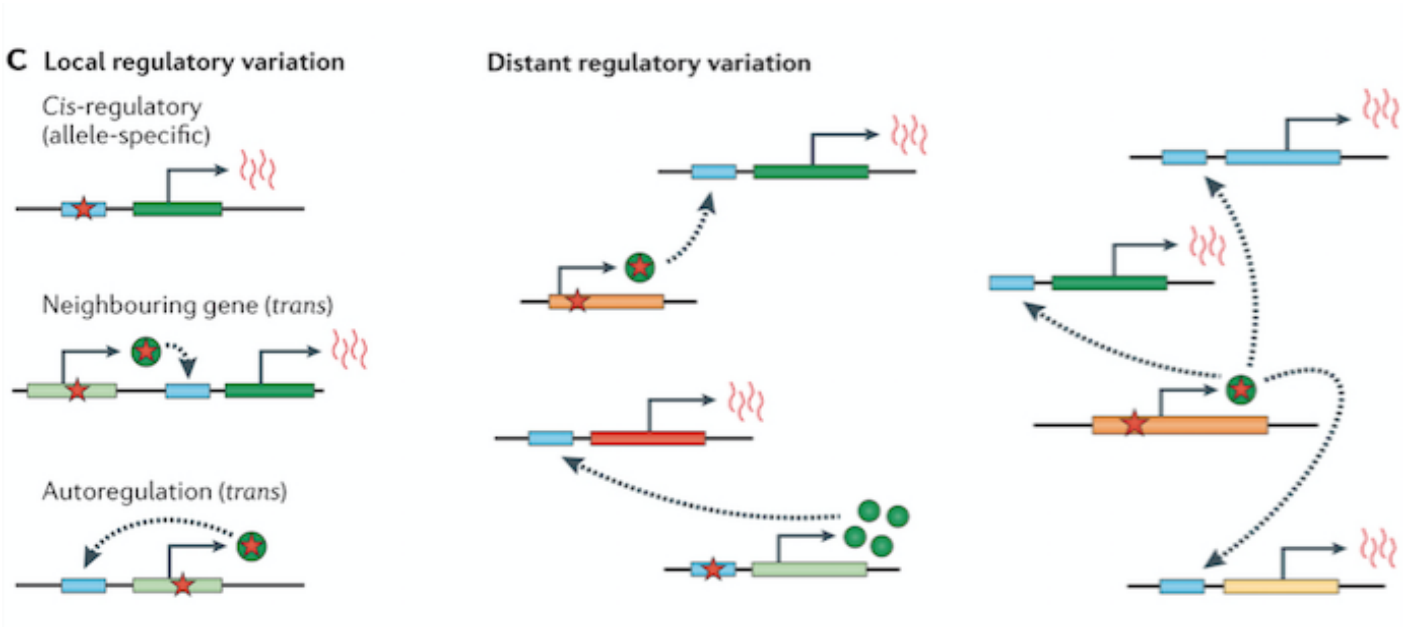
What about other forms of genomic architecture?

Summary

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

Which loci to test?

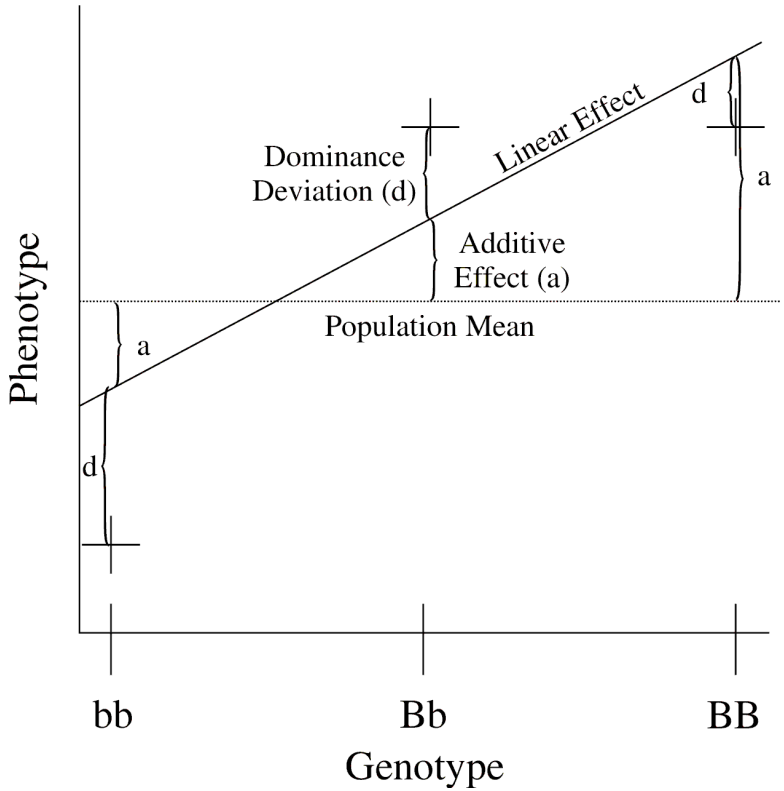


Albert and Kruglyak, NRG 2015

What issues may arise when performing either a *cis*- or *cis and trans*-eQTL analysis?

It is important to think about covariates in the model.

Common model - additive effect



Software

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

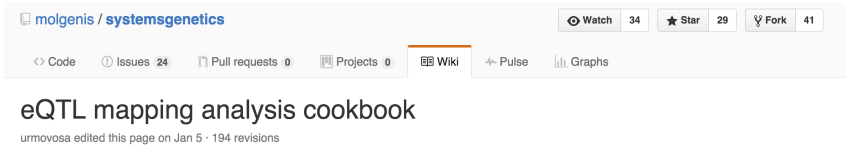
plink - *(pngu.mgh.harvard.edu/ purcell/plink/)*

Package 'MatrixEQTL'

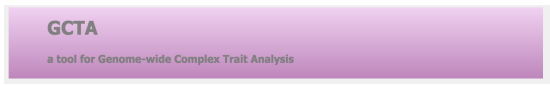
February 19, 2015

Type Package
Title Matrix eQTL: Ultra fast eQTL analysis via large matrix operations
Version 2.1.1
Date 2014-02-24
Author Andrey Shabalin

Matrix eQTL - *(https://cran.r-project.org/web/packages/MatrixEQTL/MatrixEQTL.pdf)*



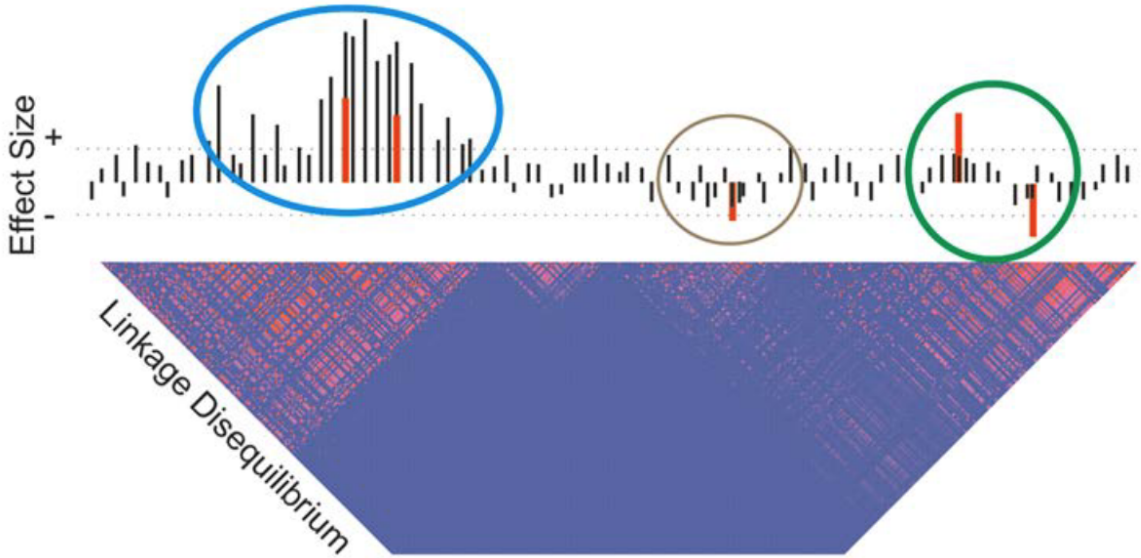
eQTL CookBook - *(https://github.com/molgenis/systemsgenetics/wiki/eQTL-mapping-analysis-cookbook)*



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

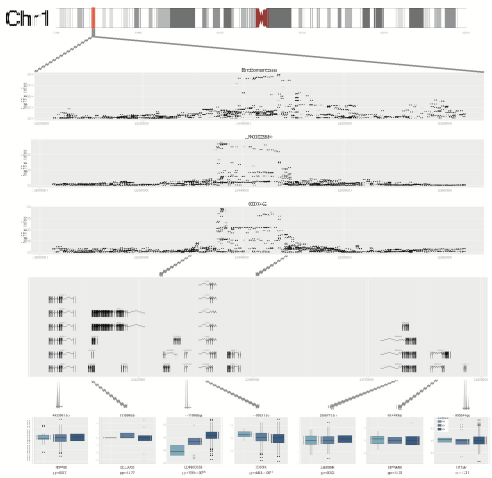
Conditional analysis

Multiple independent eQTL within a locus



We can use methods such as multiple regression, or sequential rounds of conditional analyses to identify secondary (and so on) effects.

Link to GWAS hits



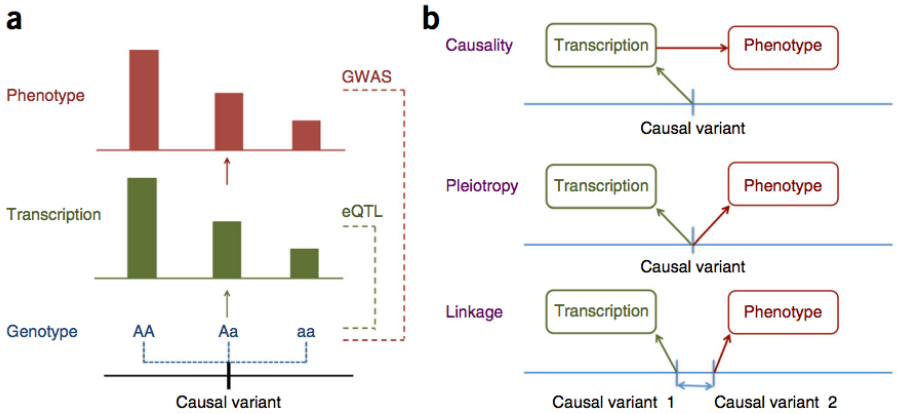
Methods such as SMR can test (a) gene(s) in a region are the functional gene of a GWAS hit (b) is the eQTL and GWAS loci is the same variant

Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets

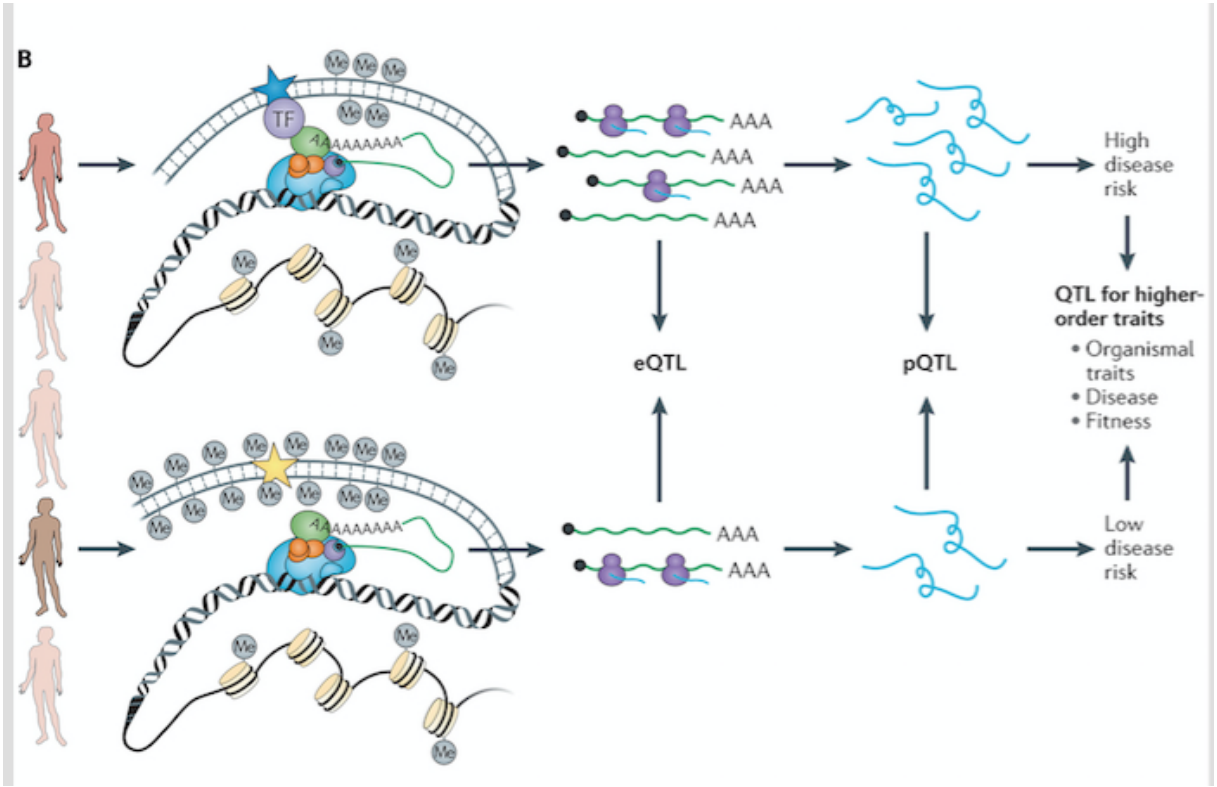
Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher & Jian Yang

Affiliations | Contributions | Corresponding author

Nature Genetics **48**, 481–487 (2016) | doi:10.1038/ng.3538



Mechanisms



Albert and Kruglyak, NRG 2015