Quantitative genetics of pathways

Key idea

- The machinery of quantitative genetics (such as QTL and association mapping, heritability, the fraction of trait variance due to genetics versus environment) apply to ANY trait we can measure
- Hence, molecular-level traits can also be handled via quantitative genetics
  - mRNA, protein, metabolite abundance
  - "epigenetic" traits: methylation patterns at a given gene
- The machinery for multivariate traits also applies to any set of (potentially interconnected) molecular traits

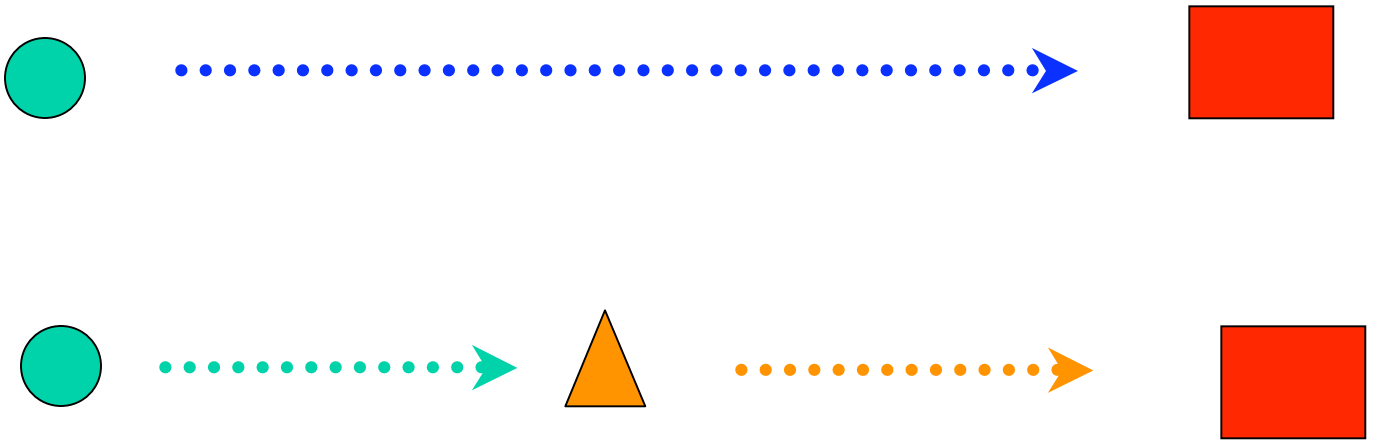
Overview

- Intermediate molecular traits between DNA and a phenotypic trait
- Directed acylic graphs
- Determining causality
- Path analysis, multiple regression, structural equation modeling

Intermediate molecular phenotypes

Genetic variation at a locus

Phenotypic variation in a trait



Locus controls variation in mRNA level at another gene

mRNA variation at controls trait variation

Perhaps we will have more resolution by focusing on these molecular intermediates

- Perhaps less noise (environmental variance) at these intermediate molecular traits
- We can search for these by looking for associations in transcript level and phenotype, protein level and phenotype, etc.
- Key point: don't need to know the pathway(s)

Terminology

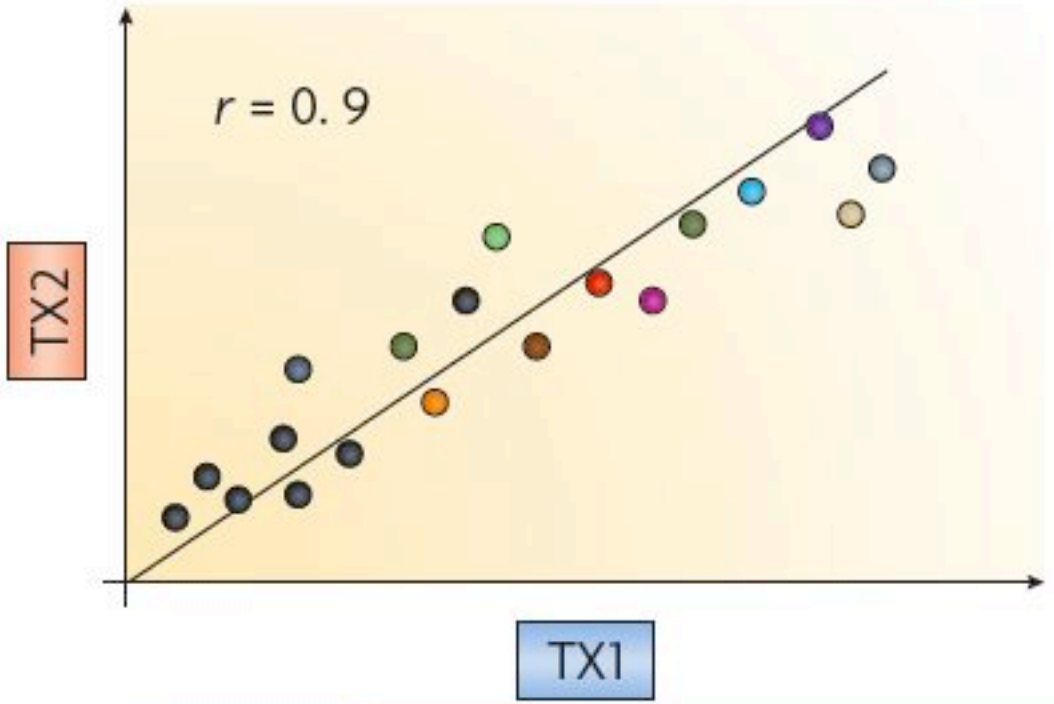
- QTL
  - Locus with effect on a trait
- Endophenotype
  - Any intermediate molecular phenotype associated with an organismal-level trait
- eQTL
  - The trait is expression level, typically of a mRNA, but could also be a protein or metabolote
- QTT -- **quantitative trait transcript**
  - Transcript levels at a gene influence trait values
- Genetical Genomics
  - The quantitative genetics of functional genomics traits (transcription levels, methylation, etc.)

Using correlations of transcript abundance to define modules

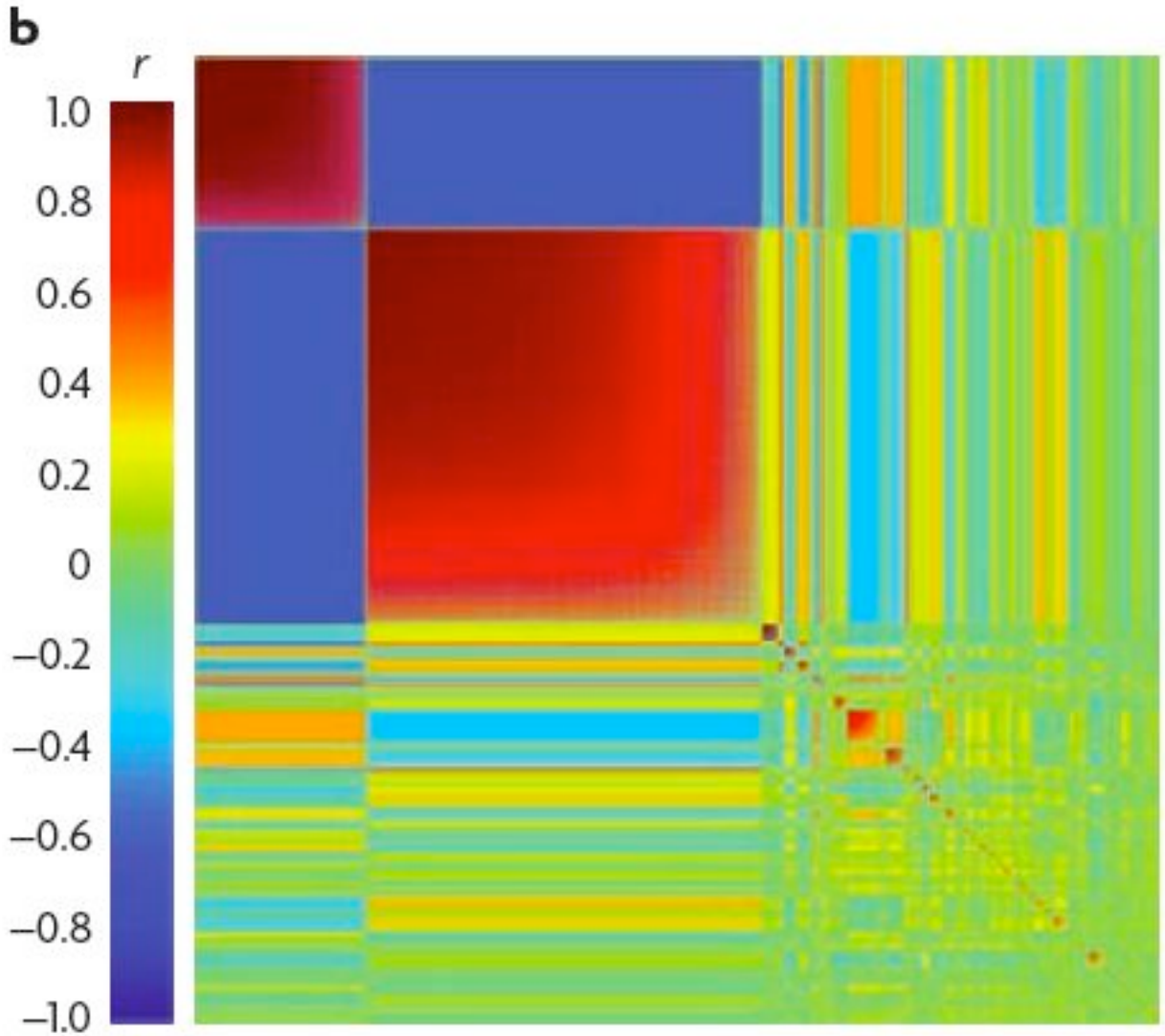
- By comparing either sets of individuals or sets of lines, one can compute the correlation among transcript abundance from different genes
- This can be used to assign genes into "modules" where there is high (absolute) correlation among all members within a module, low correlation with outside members
- The same exercise can also be done among QTT transcript, restricting attention to the covariance patterns at genes whose transcript abundance levels covary with out trait of interest.

Use microarrays (or other) to compute pair-wise correlation among transcripts

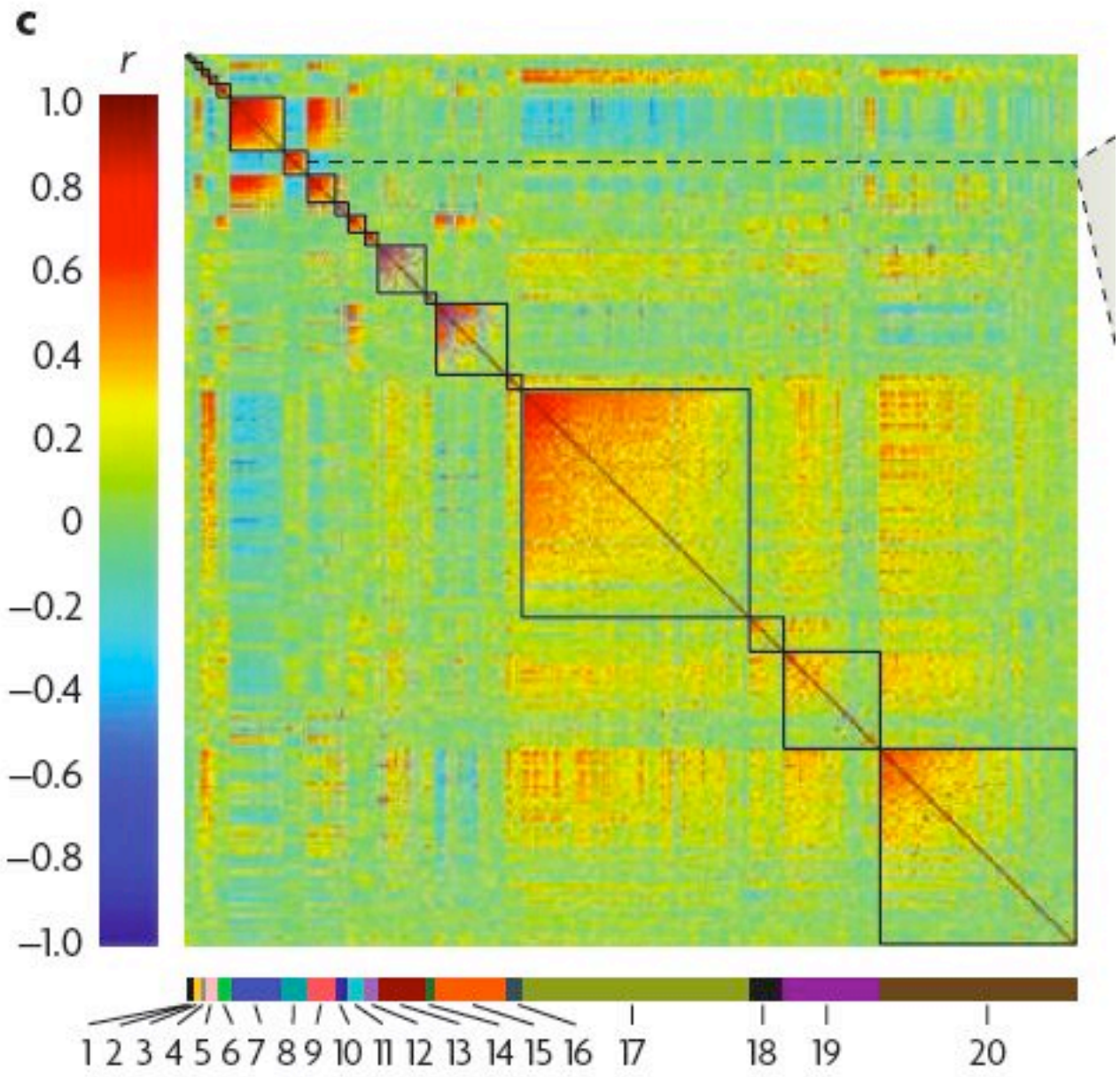
a



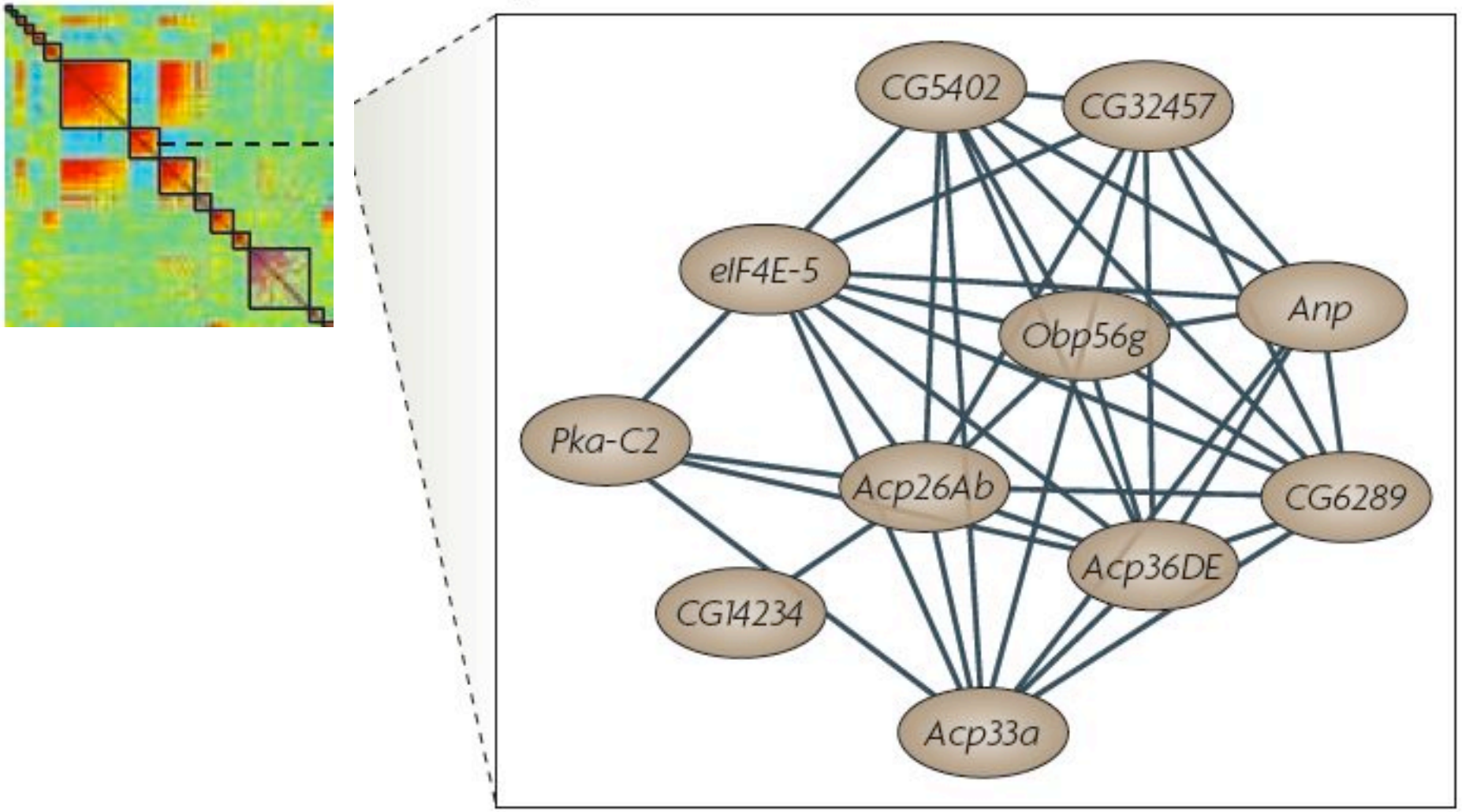
	TX1	TX2	TX3	...	TXn
TX1	1.00	0.90	–0.02	...	0.17
TX2	0.90	1.00	–0.05	...	0.19
TX3	–0.02	–0.05	1.00	...	–0.96
...	...	...	...	...	0.25
TXn	0.17	0.19	–0.96	0.25	1.00



In *Drosophila melanogaster*, the coexpression patterns of 10,096 variable transcripts (based on 40 inbred lines) clustered into 241 modules (red blocks) with high correlation within each and low between



The 414 QTT for competitive fitness formed 20 modules, using an absolute correlation threshold of 0.6 to define modules



Close-up of one of these modules (here involving 11 genes)

Tests of functional enrichment

- Are modules enriched for certain types of genes?
  - Enrichment analysis
- One standard approach is to look at GO (gene ontology) terms associated with the elements within each cluster. GO terms attempt to classify all known genes into categories
- Suppose there are 600 genes in a particular GO category (out of, say, 10,000 genes scored). Of the 150 genes in your cluster, you find 50 in this category . Is there an excess in your module?
  - Use Fisher's exact test (exact 2 x 2 contingency table tests)

	In module	Not in module
Total genes	150	10000-150 = 9850
Particular *GO* category	50	600-50 = 550

```
> d<-matrix(c(150,50,9850,550),nrow=2)
> d
     [,1] [,2]
[1,]  150 9850
[2,]   50  550
> fisher.test(d)

 Fisher's Exact Test for Count Data

data:  d
p-value = < 2.2e-16
```

This *GO* group is significantly over-represented in the module. Multiple comparison correction required since multiple *GO* groups tested.

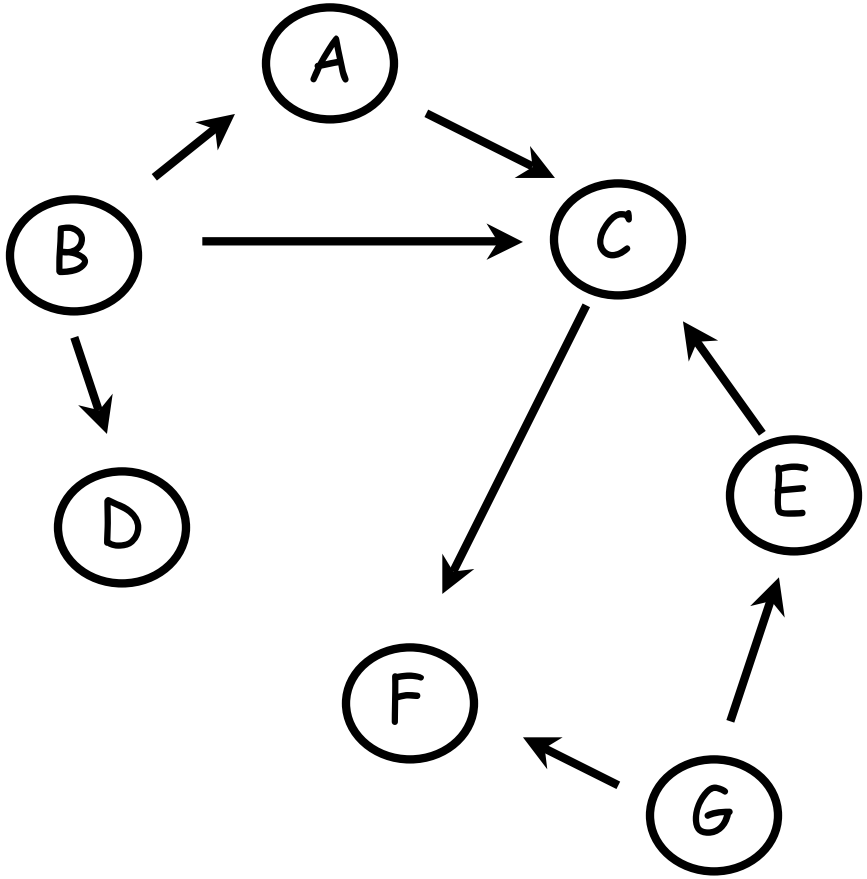
Other tools

- KEGG: Kyoto Encyclopedia of genes and genomes
- Essentially a list of all known pathways
- Similar to a GO analysis, can look for enrichment of elements of a particular KEGG within a module

More graph theory

- Edges (lines) connect nodes (also called vertices)
- A **directed graph** uses arrows in place of edges to indicate influence (causality)
- In an **acyclic graph**, no pathways (feedback loops) that connect a node back to itself through a series of intermediate nodes
- Simple models of basic pathways are **directed acyclic graphs**

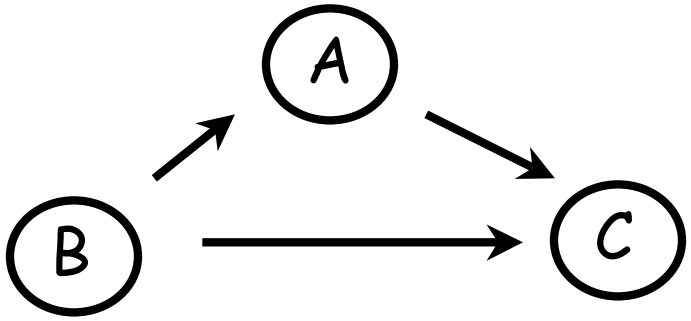
A directed, acyclic graph



Acyclic, as no directed pathways ever return to a node

Matrices (again)!

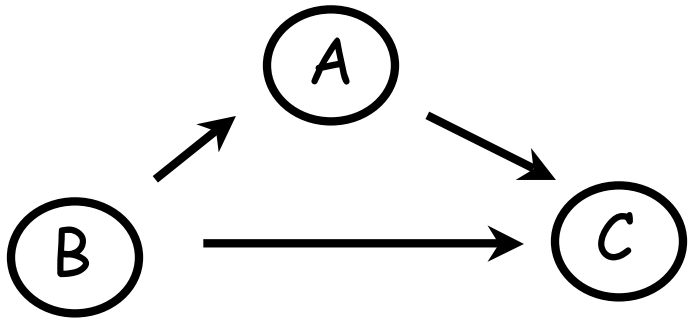
The topology (patterns of connection) of a graph can be represented as a matrix



A direct graph has a nonzero value when node i influences node j. If we simply want to represent the topology, use 0/1

From \ To	A	B	C
A	0	0	1
B	1	0	1
C	0	0	0

We can also indicate the nature of the directed interactions by +1 (increases value) or -1 (decreases value)

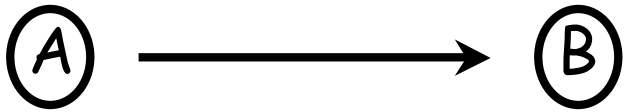


More generally, the dynamic behavior of the system can be indicated by a strength coefficient in the matrix.This value may change with the value of the influencing node

		To		
		A	B	C
From	A	0	0	-0.5
	B	1.2	0	0.1
	C	0	0	0

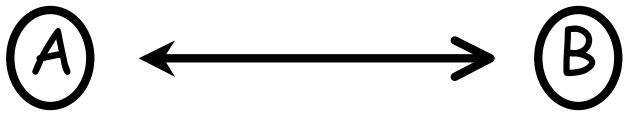
Type of directed interactions

A & B are correlated



A influences B

B influences A

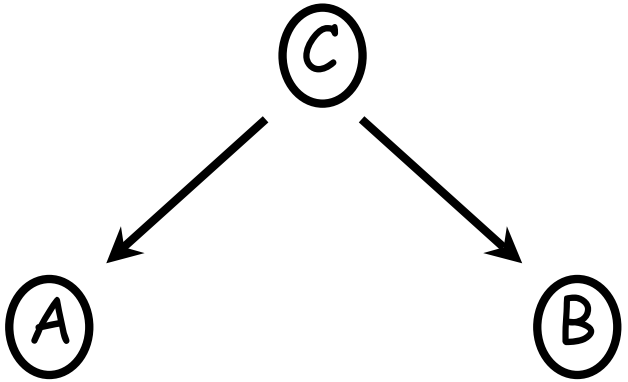


A and B influence each other

Neither A or B influence each other

Correlation and Causation

This indicates A, B correlated, but we know nothing about the causation of this association

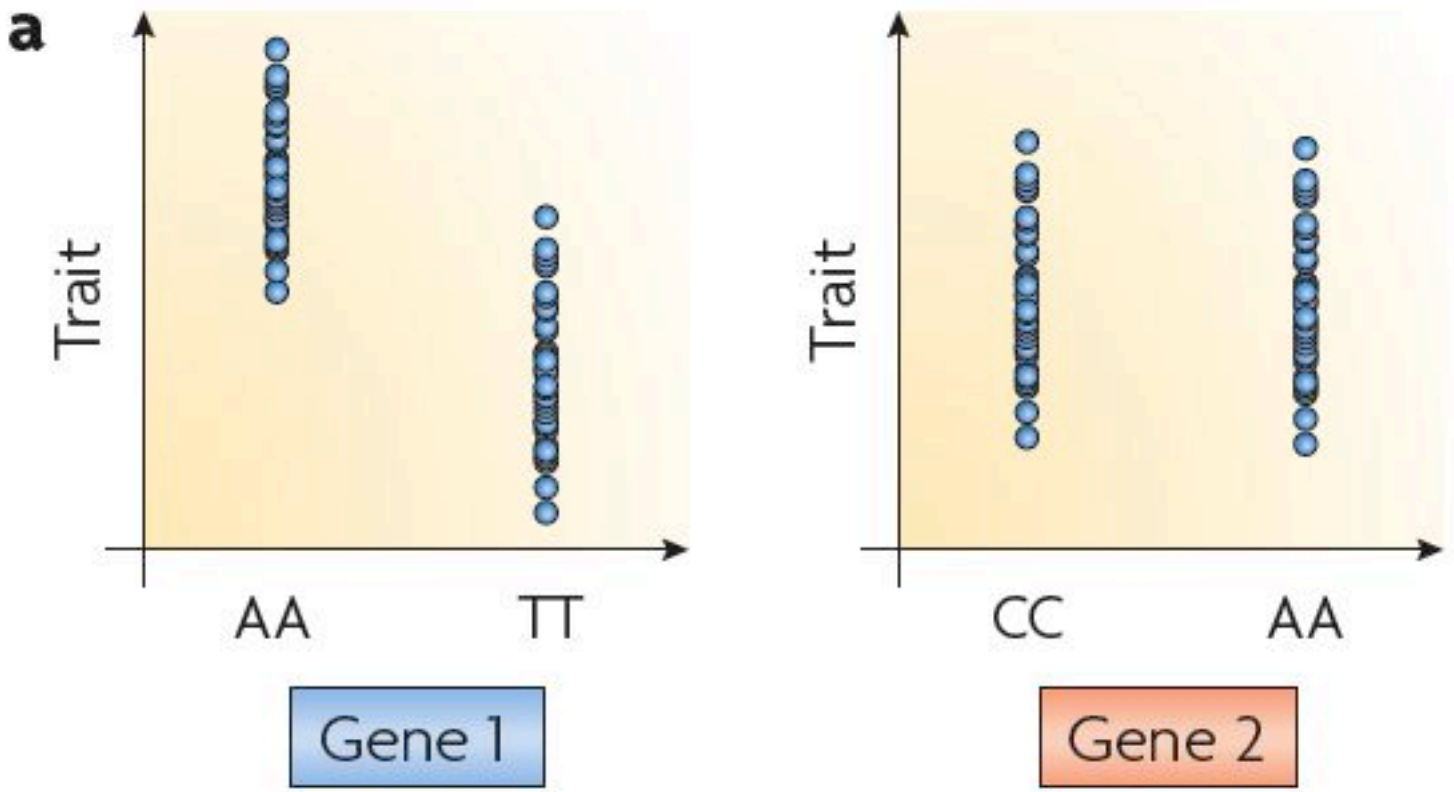


Suppose the node C influences both A & B. This generates their correlation. When C is included in the graph, the edge (connection) between A & B disappears.

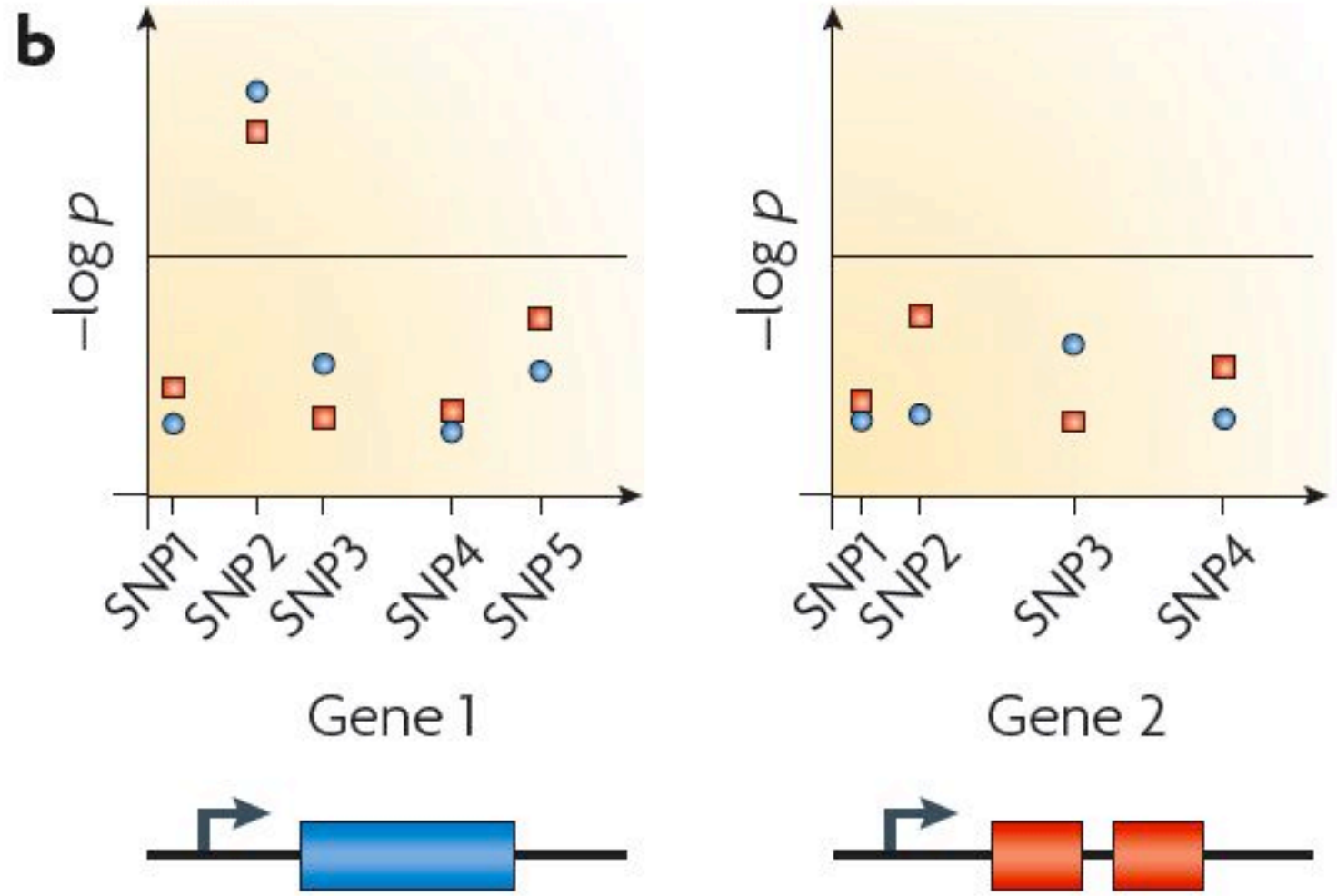
Putting these together

The following toy example is from the excellent review by Mackay et al. (2009 Nature Genetics)

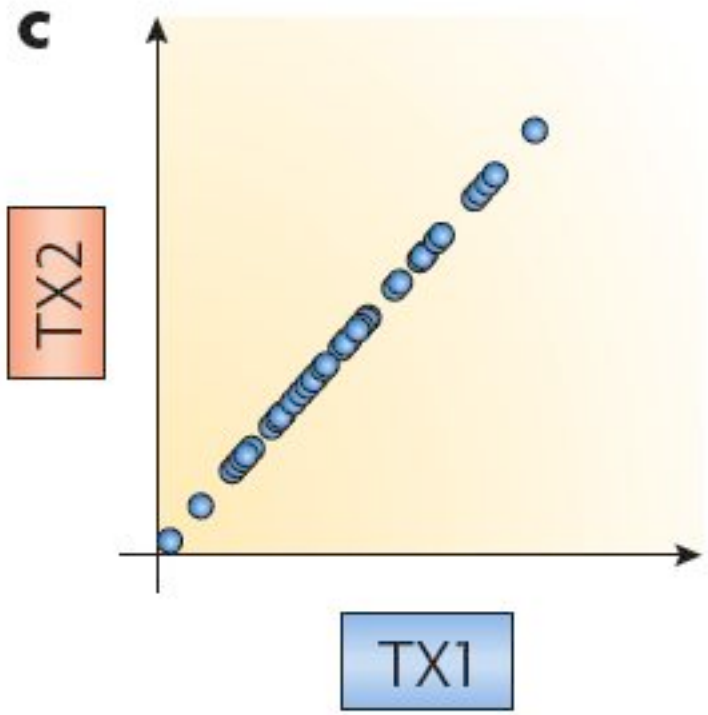
You have a series of SNPs in two genes, and use association mapping to both look for marker-trait associations and eQTLs. Further, you also look at the association between levels of expression and the (organismal-level) trait of interest



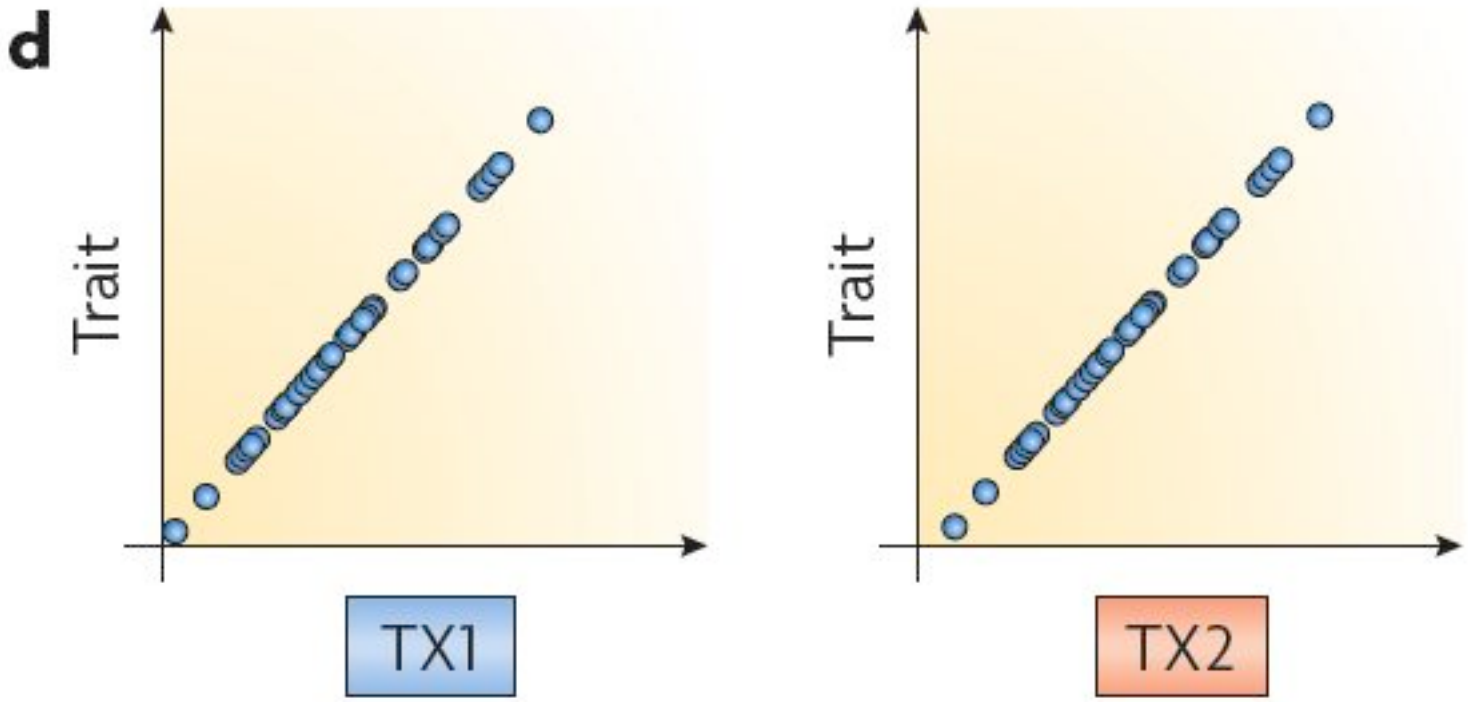
Gene one is a QTL for the organismal-level trait, while gene 2 is not



SNP 2 in gene one is an eQTL for expression levels in both gene 1 and gene 2

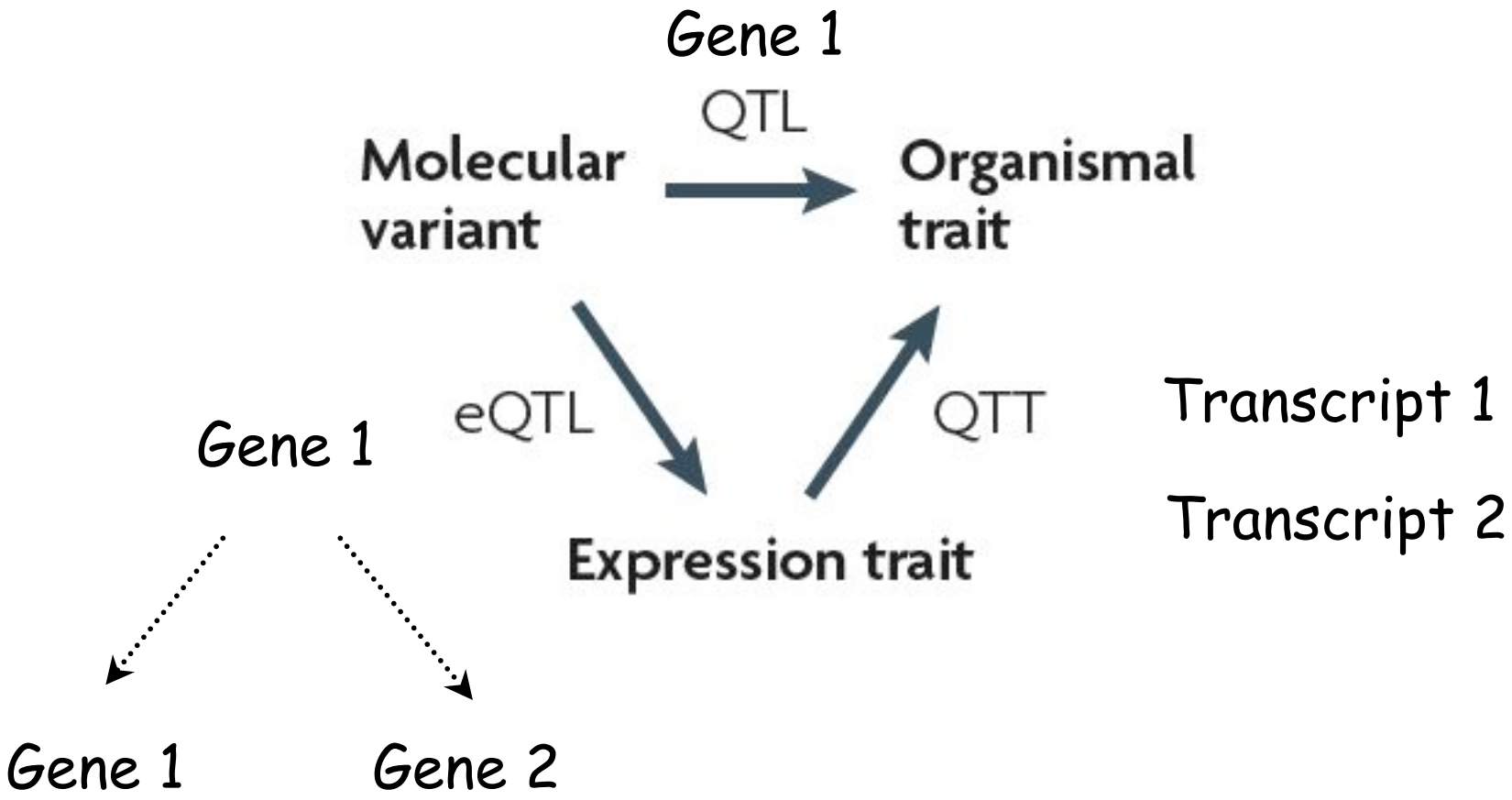


Expression levels of both genes are correlated

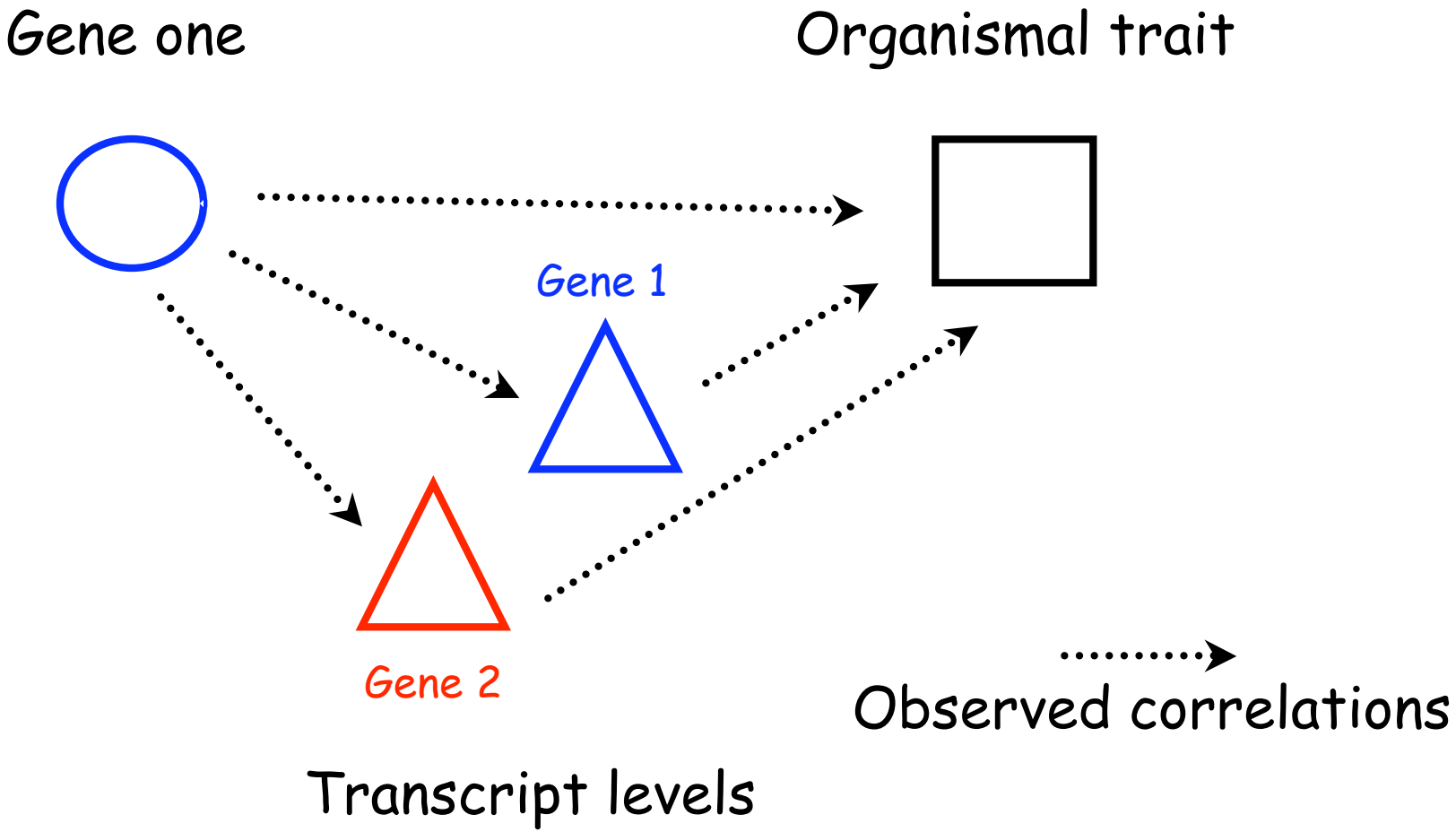


Observed trait value is a function of the level of transcripts for both gene 1 and gene 2

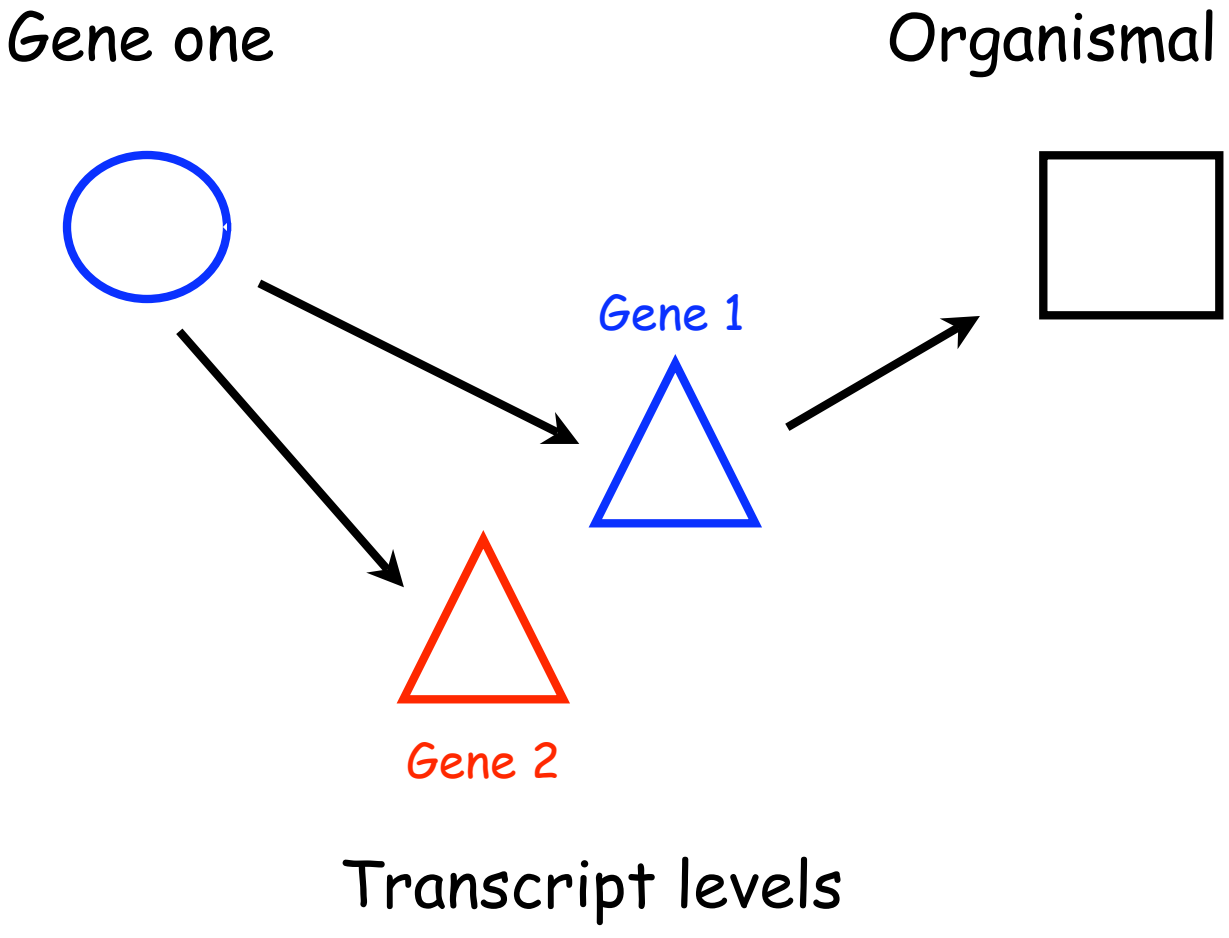
Possible pathways



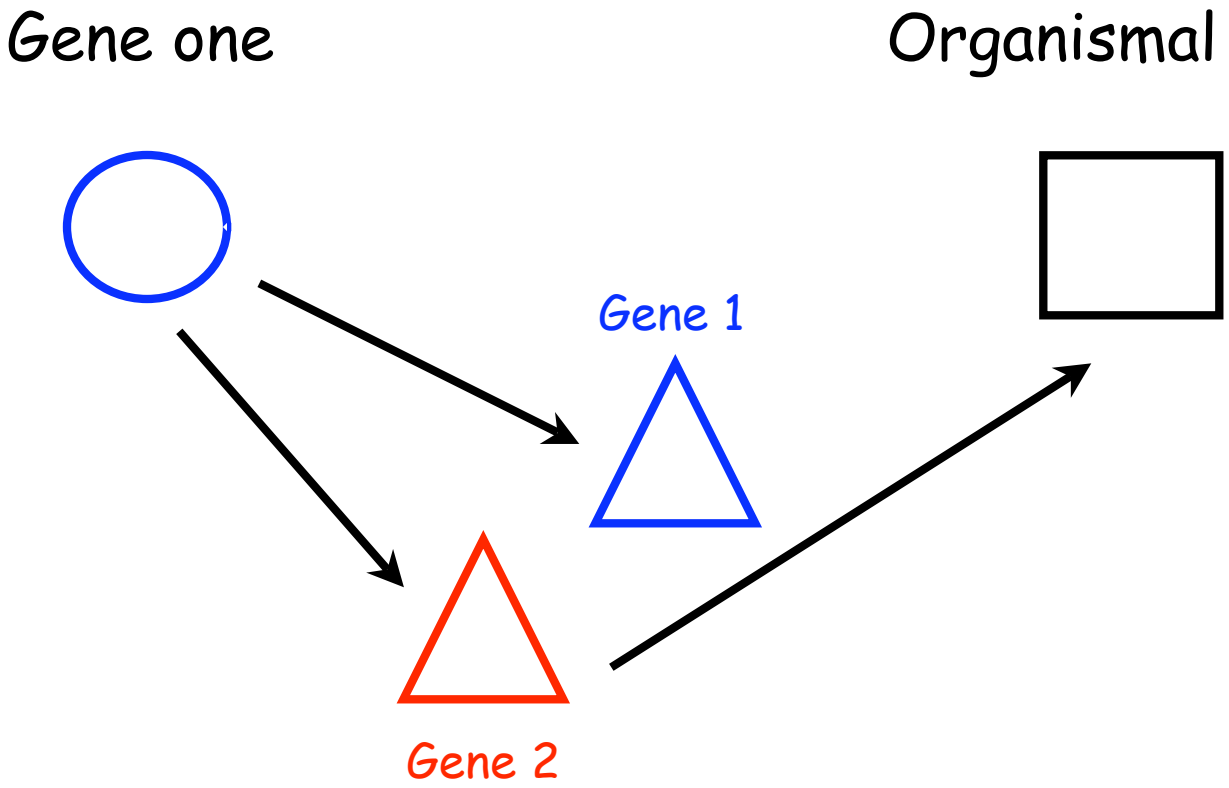
What is the model?



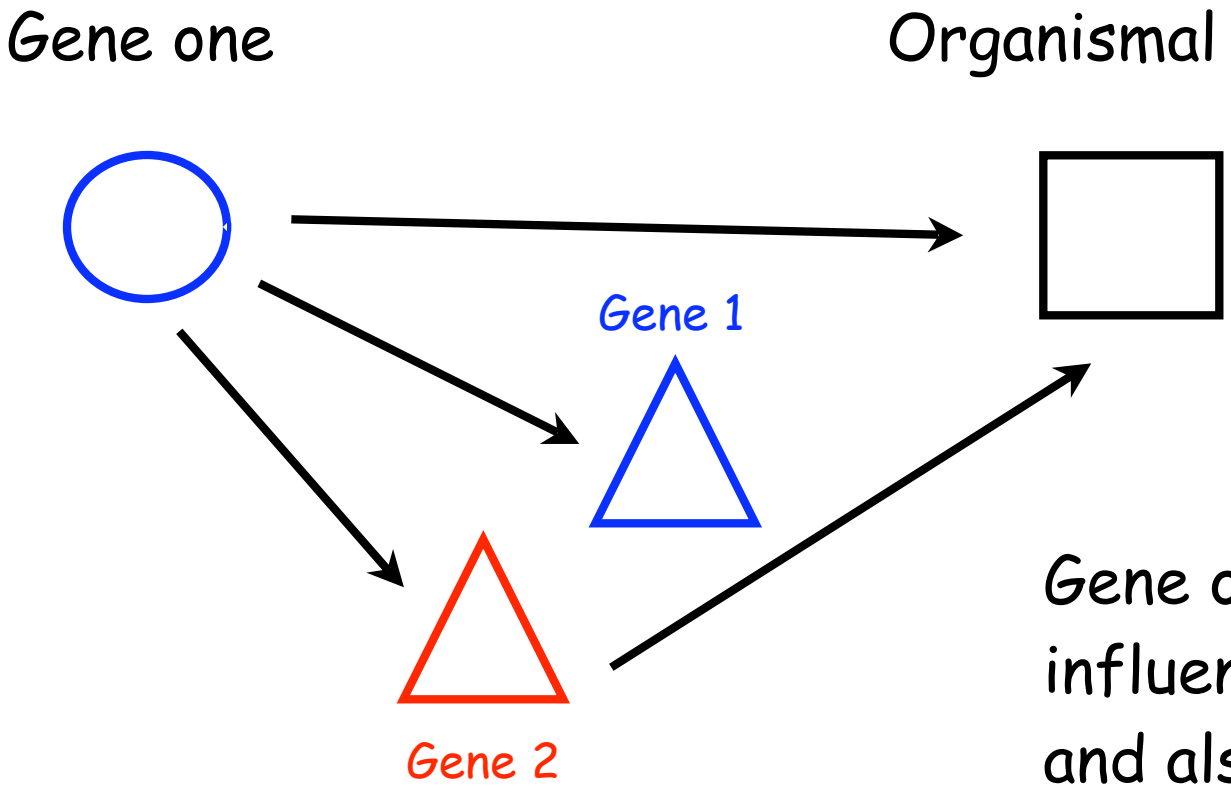
Model 1: Transcript levels of Gene 1 determine trait



Model 2: Transcript levels of gene 2 determine trait



Other models



Gene one directly influences trait, and also through expression levels of gene 2

Determining which model is correct

- This is a standard multiple regression problem, which accounts for any correlation structure among MEASURED variables (factors left out of the model can cause false associations)
- Analysis of these sort of graphs (is there any non-zero value on the observed connections) also goes under the name of Path Analysis, or Structural Equation Modeling (SEM)
- Path analysis dates to Sewall Wright in the 1920's, and is reviewed in one of the Appendices of Lynch and Walsh.

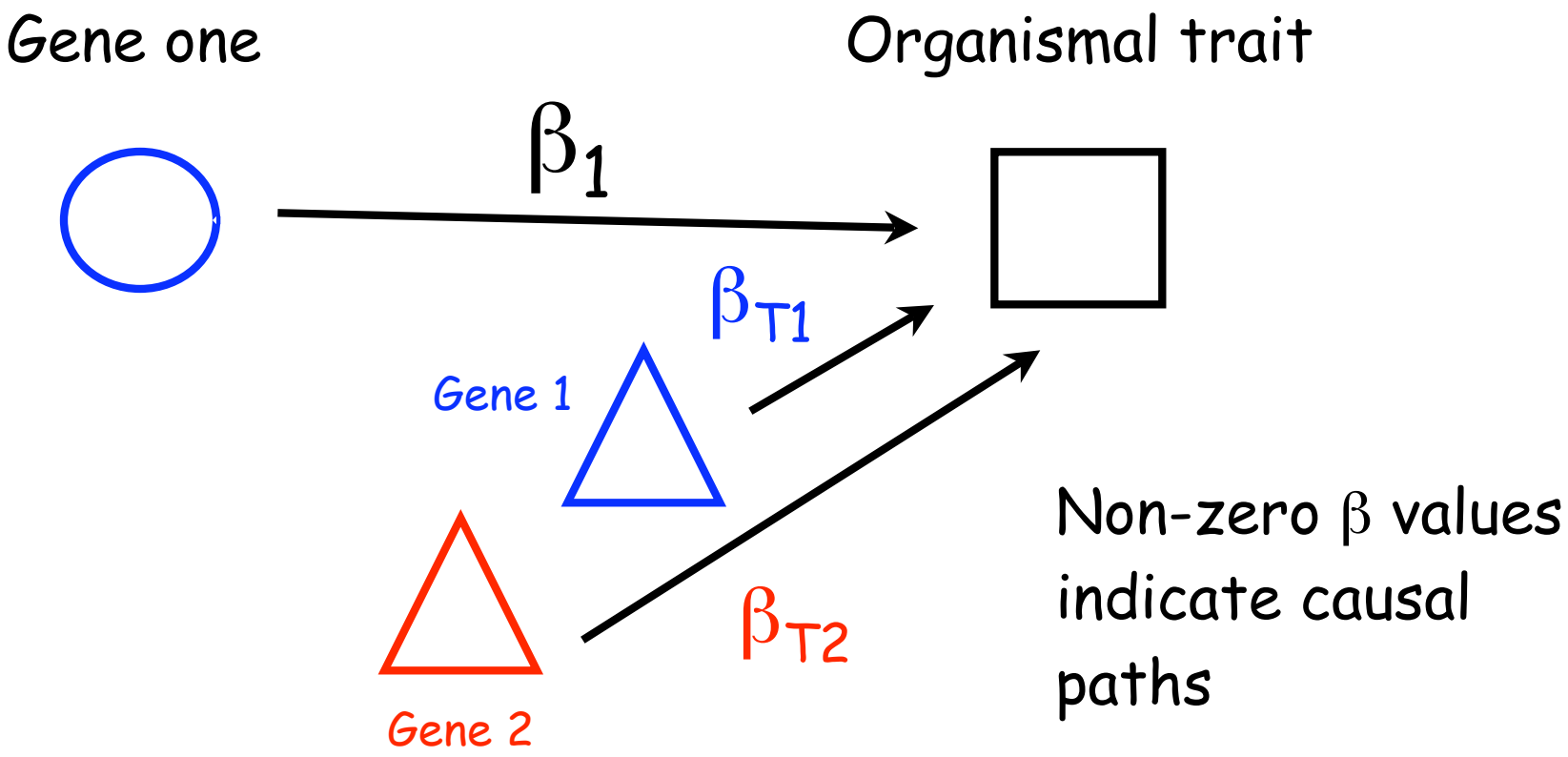
Using a multiple (aka partial) regression approach

$$z = \mu + \beta_1 G + \beta_{t1} T_1 + \beta_{t2} T_2 + e$$

z is the observed trait value, G is the marker state at gene one, T_1 and T_2 are the levels of transcript at genes one and tw.

The partial regression coefficients β determine the effect of a variable on z when all other values are held constant.

$$z = \mu + \beta_1 G + \beta_{T1} T_1 + \beta_{T2} T_2 + e$$

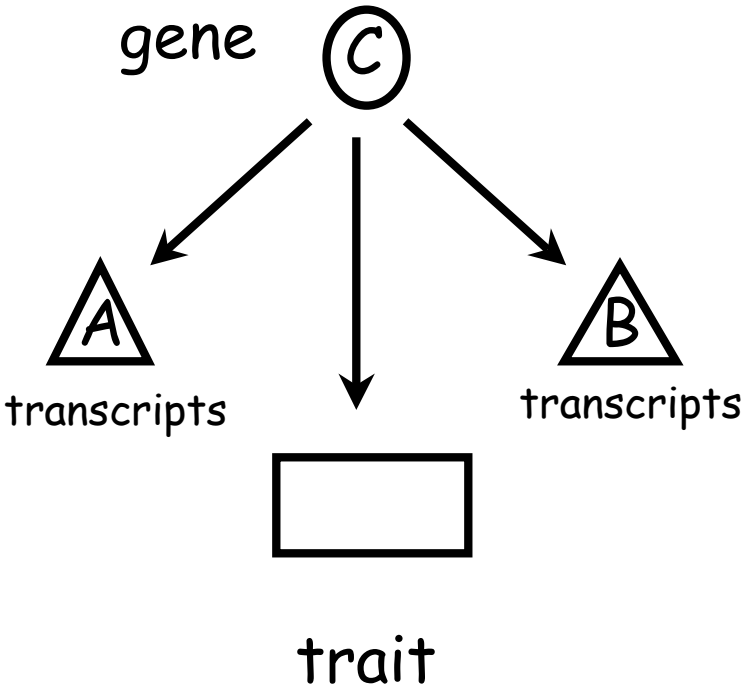


Conditional dependence

- A partial regression coefficient is an example of conditional dependence: conditioning on knowledge of the other variables, how does the target on affect the trait of intrest
- For example, if expression levels of gene two entirely determine the trait, then for fixed genotype at gene 1 and a fixed level of gene one transcript, variation in transcript two abundance should covary with the trait
- Conversely, under these assumptions if we condition on the level of transcript 2, the trait should not vary with abundance levels at transcript one.

Limitations

Correlations generated by causal effects of unmeasured parameters can be fatal.



Suppose the genotype at locus C determines both the transcript levels of A and B as well as directly influencing the trait itself.

If just A and B included in the regression model, correlations btw their transcript levels and the trait would generate nonzero β values for each. However, inclusion of the genotype of C results in both being zero