Using genetic markers to orient the edges in quantitative trait networks: the NEO software

Steve Horvath
dissertation work of Jason Aten

Aten JE, Fuller TF, Lusis AJ, Horvath S (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Systems Biology 2008, 2:34. April 15.

Using SNPs for learning directed networks

- Question: Can genetic markers help us to dissect causal relationships between gene expression- and clinical traits?
- Answer: yes, using the paradigm of Mendelian randomization
- Many authors have addressed this question both in genetics and in genetic epidemiology.

Motivating example

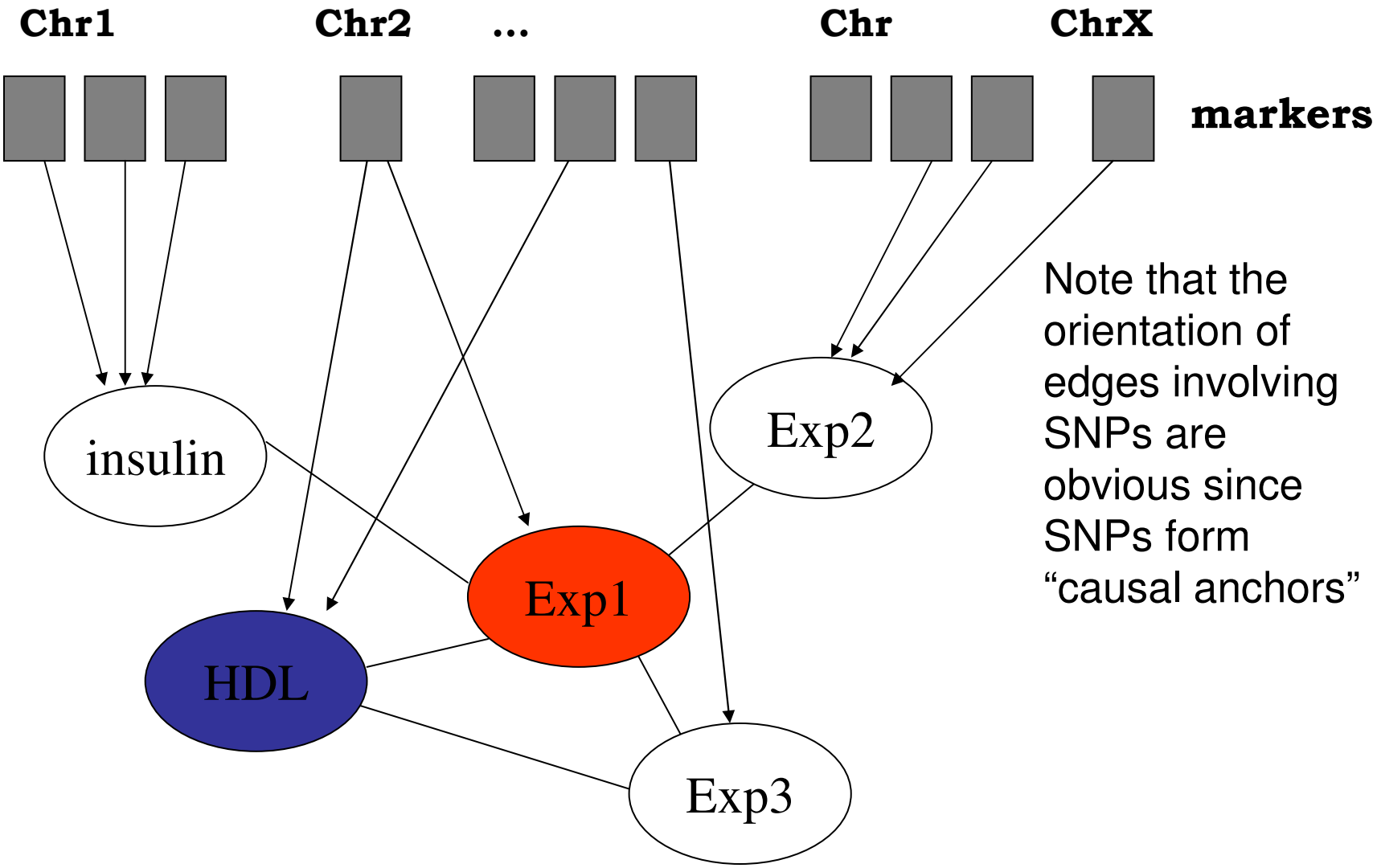
- Assume a high correlation between cholesterol levels *C* and the gene expression profile *Exp* of an unknown gene.
- Question: is the gene upstream (causal) or down stream (reactive) of cholesterol? Do high levels of the gene expression *Exp* cause high cholesterol levels *C* or the other way around?
- Answer: Genetic markers can be used to infer the directionality (orient the edge between *Exp* and *C*) if these markers are associated with either cholesterol or with the gene expression or both.

Fundamental paradigm of biology can be used for inferring causal information

- Sequence variation->gene expression (messenger RNA)->protein->clinical traits
- SNPs are “causal anchors”

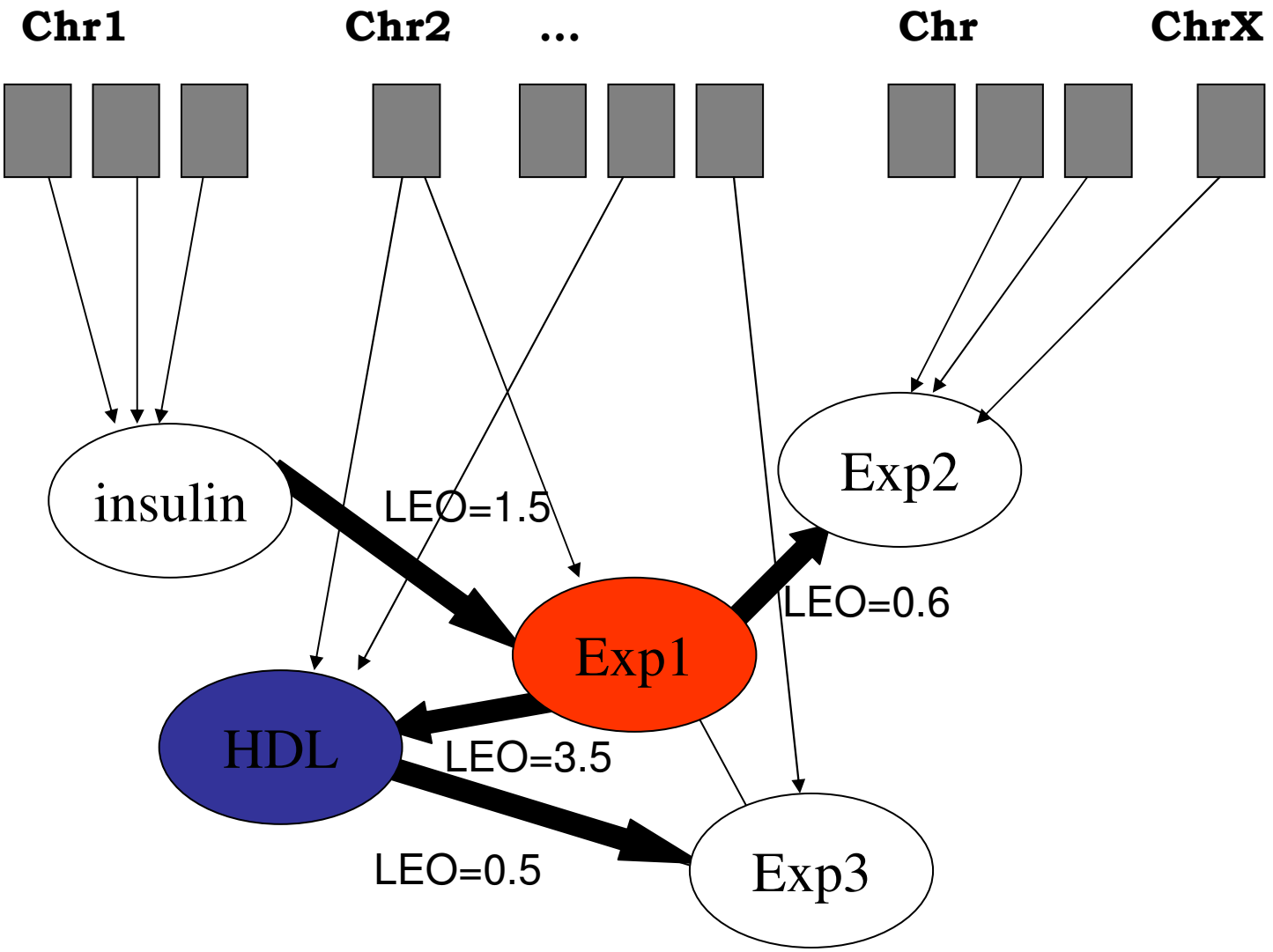
SNP -> gene expression

The edge orienting problem: unoriented edges between the gene expressions and physiologic traits



Edges between traits and gene expressions are not yet oriented

The solution to the edge orienting problem



Edges are directed. A score, which measures the strength of evidence for this direction, is assigned to each directed edge

NEO software

Input Data

- A set of quantitative variables (traits)
  - e.g. many physiological traits, blood measurements, gene expression data
- SNP marker data (or genotype data)

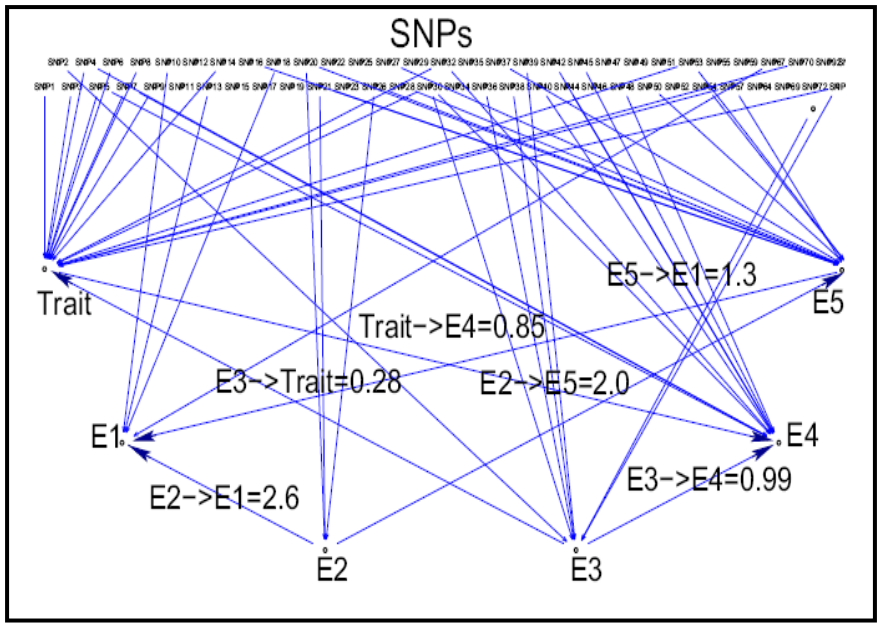
Output

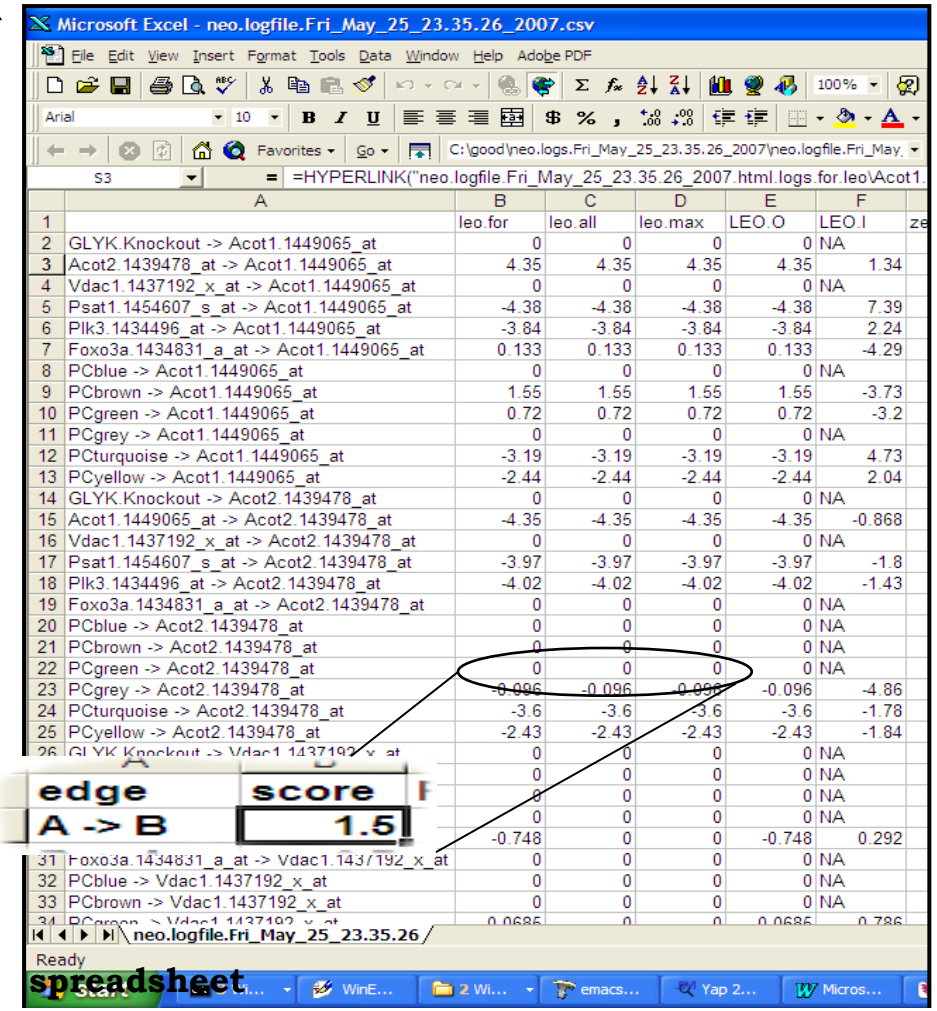
- Scores for assessing the causal relationship between correlated quantitative variables

Output of the NEO software

NEO spreadsheet summarizes LEO scores and provides hyperlinks to model fit logs

- graph of the directed network





	A	B	C	D	E	F
1		leo.for	leo.all	leo.max	LEO.O	LEO.I
2	GLYK.Knockout -> Acot1.1449065_at	0	0	0	0	NA
3	Acot2.1439478_at -> Acot1.1449065_at	4.35	4.35	4.35	4.35	1.34
4	Vdac1.1437192_x_at -> Acot1.1449065_at	0	0	0	0	NA
5	Psat1.1454607_s_at -> Acot1.1449065_at	-4.38	-4.38	-4.38	-4.38	7.39
6	Plk3.1434496_at -> Acot1.1449065_at	-3.84	-3.84	-3.84	-3.84	2.24
7	Foxo3a.1434831_a_at -> Acot1.1449065_at	0.133	0.133	0.133	0.133	-4.29
8	PCblue -> Acot1.1449065_at	0	0	0	0	NA
9	PCbrown -> Acot1.1449065_at	1.55	1.55	1.55	1.55	-3.73
10	PCgreen -> Acot1.1449065_at	0.72	0.72	0.72	0.72	-3.2
11	PCgrey -> Acot1.1449065_at	0	0	0	0	NA
12	PCturquoise -> Acot1.1449065_at	-3.19	-3.19	-3.19	-3.19	4.73
13	PCyellow -> Acot1.1449065_at	-2.44	-2.44	-2.44	-2.44	2.04
14	GLYK.Knockout -> Acot2.1439478_at	0	0	0	0	NA
15	Acot1.1449065_at -> Acot2.1439478_at	-4.35	-4.35	-4.35	-4.35	-0.868
16	Vdac1.1437192_x_at -> Acot2.1439478_at	0	0	0	0	NA
17	Psat1.1454607_s_at -> Acot2.1439478_at	-3.97	-3.97	-3.97	-3.97	-1.8
18	Plk3.1434496_at -> Acot2.1439478_at	-4.02	-4.02	-4.02	-4.02	-1.43
19	Foxo3a.1434831_a_at -> Acot2.1439478_at	0	0	0	0	NA
20	PCblue -> Acot2.1439478_at	0	0	0	0	NA
21	PCbrown -> Acot2.1439478_at	0	0	0	0	NA
22	PCgreen -> Acot2.1439478_at	0	0	0	0	NA
23	PCgrey -> Acot2.1439478_at	-0.096	-0.096	-0.096	-0.096	-4.86
24	PCturquoise -> Acot2.1439478_at	-3.6	-3.6	-3.6	-3.6	-1.78
25	PCyellow -> Acot2.1439478_at	-2.43	-2.43	-2.43	-2.43	-1.84
26	GLYK.Knockout -> Vdac1.1437192_x_at	0	0	0	0	NA
27		0	0	0	0	NA
28		0	0	0	0	NA
29		0	0	0	0	NA
30		-0.748	0	0	-0.748	0.292
31	Foxo3a.1434831_a_at -> Vdac1.1437192_x_at	0	0	0	0	NA
32	PCblue -> Vdac1.1437192_x_at	0	0	0	0	NA
33	PCbrown -> Vdac1.1437192_x_at	0	0	0	0	NA
34	PCgreen -> Vdac1.1437192_x_at	0.0686	0	0	0.0686	0.786

Correlation and causation

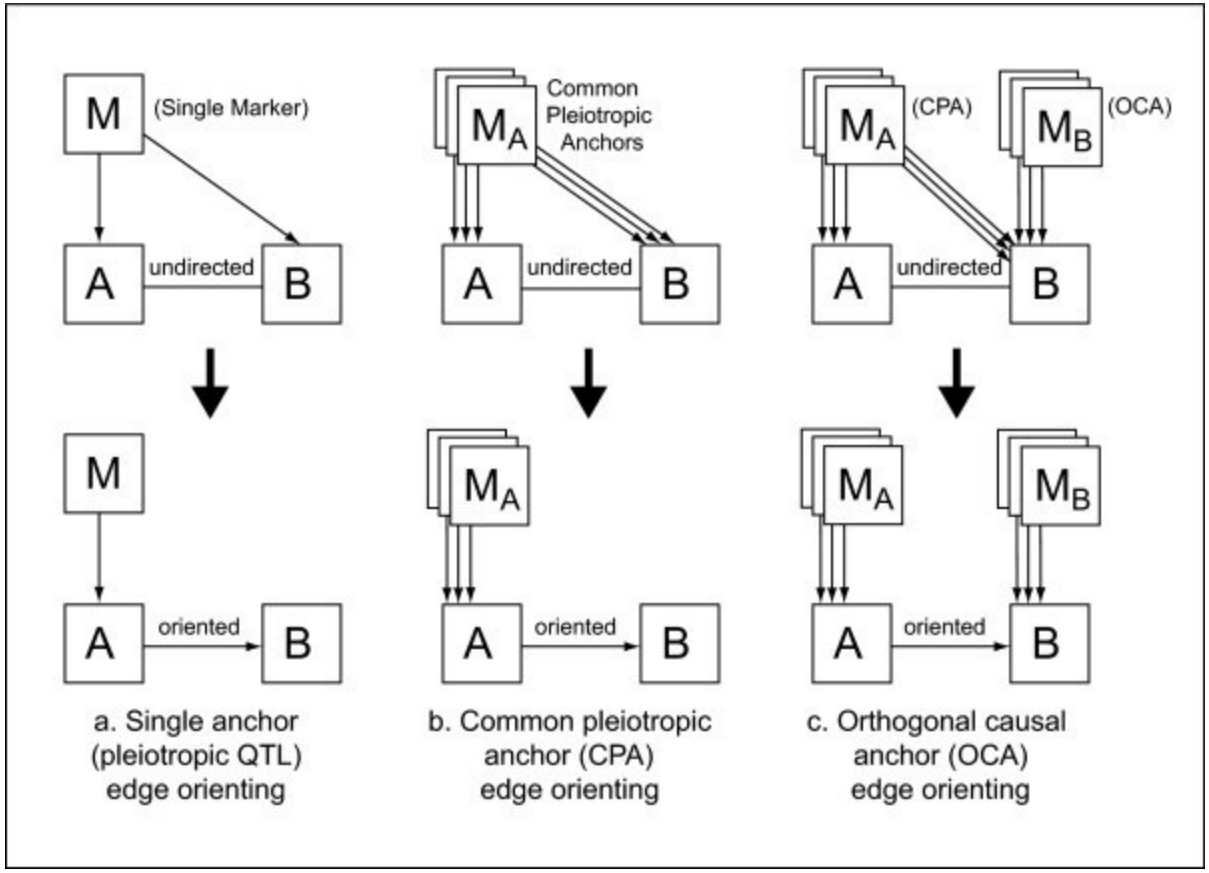
- Background: by comparing correlation coefficients one can sometimes infer causal information.
  - The saying that “correlation does not imply causation” should be changed to “correlation does not *always* imply causation”
- A causal graph implies statements about the relationship of the pairwise correlations.
- More generally it implies statements about the likelihood of a corresponding structural equations model
- Several good introductory books, e.g. Shipley

NEO Network Edge Orienting

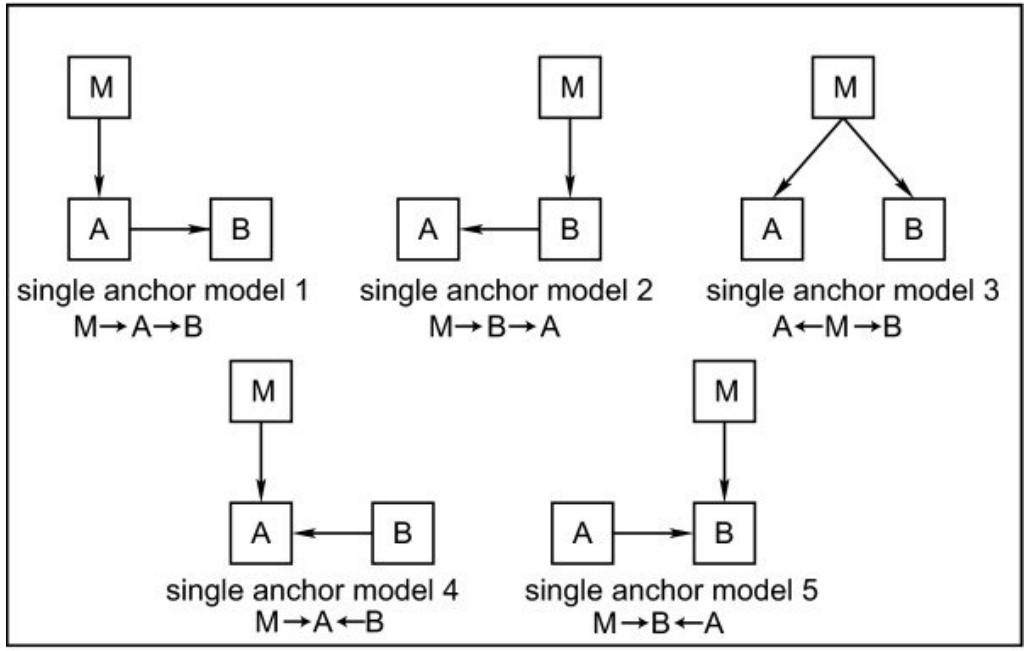
is a set of algorithms, implemented in R software functions, which compute scores for causal edge strength

- **LEO** – compares local structural equation models; the more positive the score, the stronger the evidence

Candidate common pleiotropic anchors (CPA) versus candidate orthogonal candidate anchors (OCA) for the edge A-B

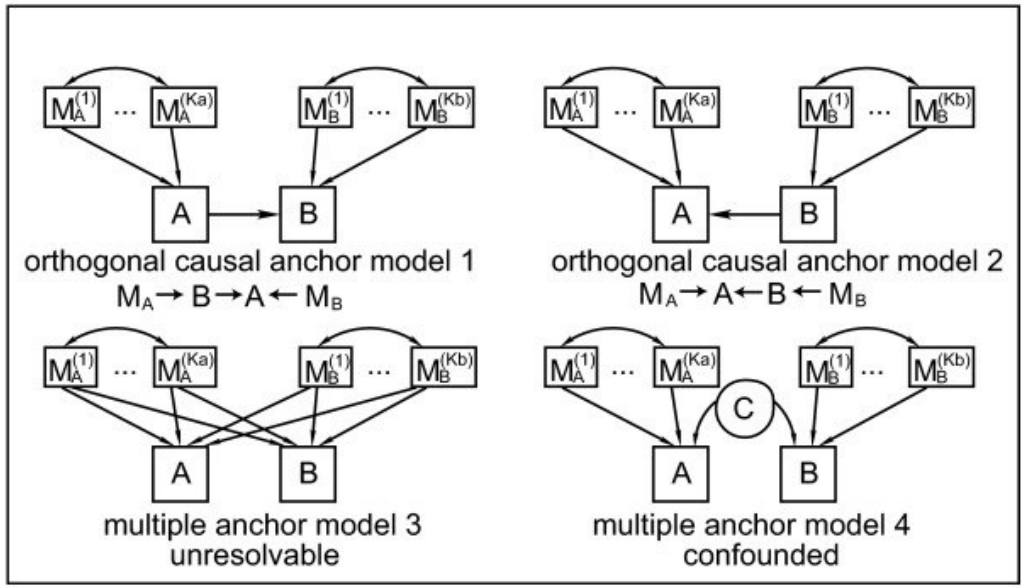


Single marker causal models between traits A and B



(a)

Multi-marker causal models



(b)

Computing the model chi-square test p-value for assessing the fit

The following function is minimized to estimate the model based covariance matrix $\Sigma(\theta)$

$$F(\theta) = \ln|\Sigma(\theta)| - \ln|S| + trace(S\Sigma(\theta)^{-1}) - m$$

where m denote the number of variables.

Denote the minimizing value by $\hat{\theta}$.

Then following follows a chi-square distribution

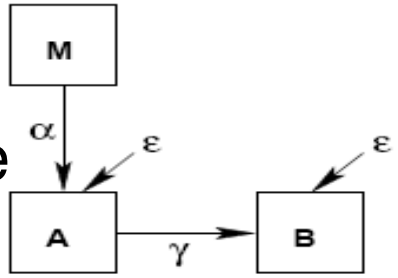
$$\chi^2 = (N-1)F(\hat{\theta}) \approx \chi^2\left(\frac{m(m-1)}{2} - t\right)$$

which can be used to compute a p-value for the causal model.

The higher the p-value, the better the causal model fits the data.

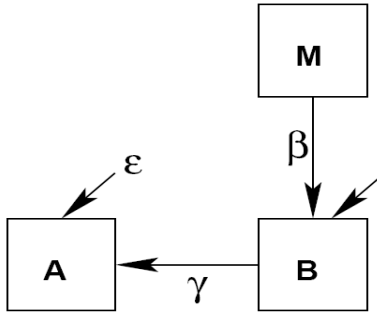
Causal models and corresponding model fitting p-values for a single marker M and the edge A-B.

P(M->A->B)= P(model 1) where

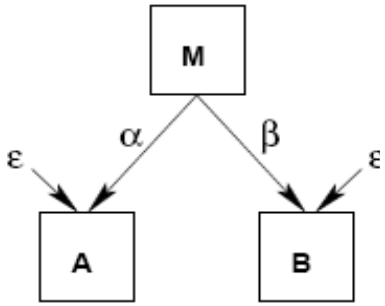


model 1

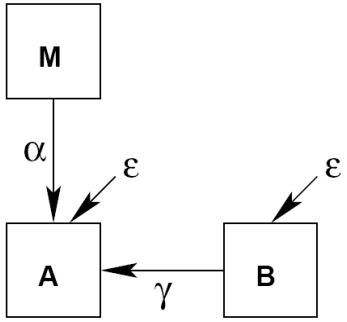
P(M->B->A)= P(model 2) where



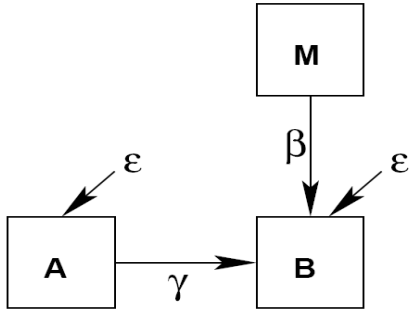
model 2



model 3



model 4



model 5

LEO.NB.SingleMarker(A->B) = $\log_{10}$(RelativeFit)

compares the model fitting p-value of A->B with that of the **N**ext **B**est model

$$\text{LEO.NB.SingleMarker}(A \rightarrow B)$$

$$= \log_{10}\left(\frac{P(M->A->B)}{\text{Model fitting p-value of the next best model}}\right)$$

where the model fitting p-value of the next best model is given by

$$\max(P(M->B->A), P(A \leftarrow M->B),$$
$$P(M->A \leftarrow B), P(A->B \leftarrow M))$$

Overview Network Edge Orienting

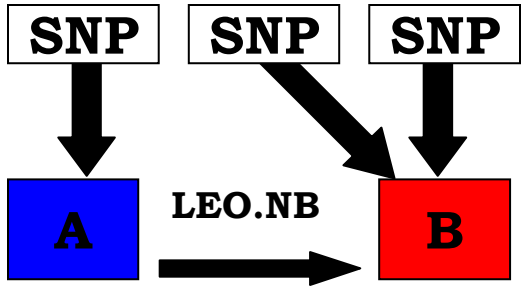
1) Merge genetic markers and traits

2) Specify manually genetic markers of interest, or invoke automated marker selection & assignment to trait nodes
Automated tools:
- greedy & forward-stepwise SNP selection;

3) Compute Local-structure edge orienting (LEO) scores to assess the causal strength of each A-B edge
- based on likelihoods of local Structural Equation Models
- integrates the evidence of multiple SNPs

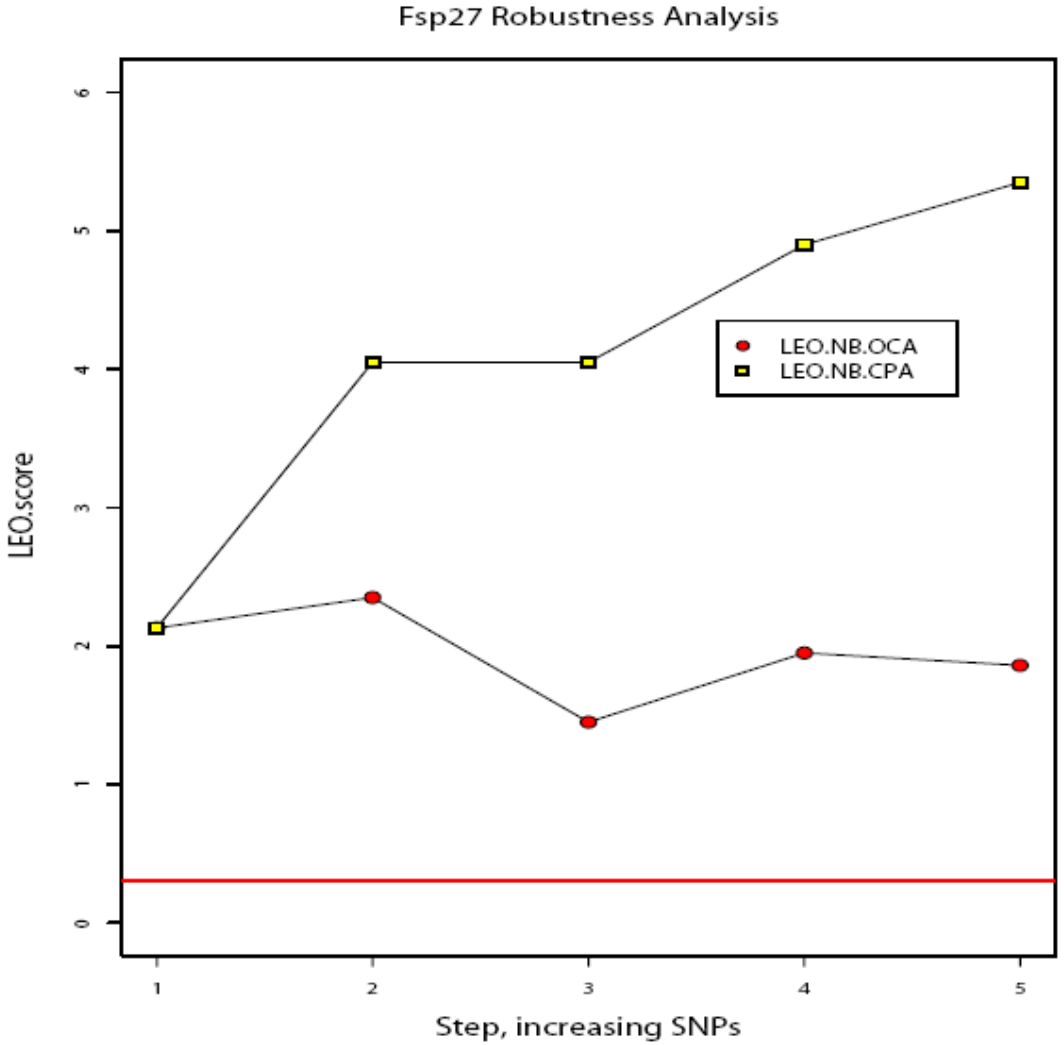
4) For each edge with high LEO score, evaluate the fit of the underlying local SEM models
- fitting indices of local SEMs: RMSEA, chi-square statistics

5) Robustness analysis
with regard to automatic marker selection;
6) Repeat analysis for next A-B edge



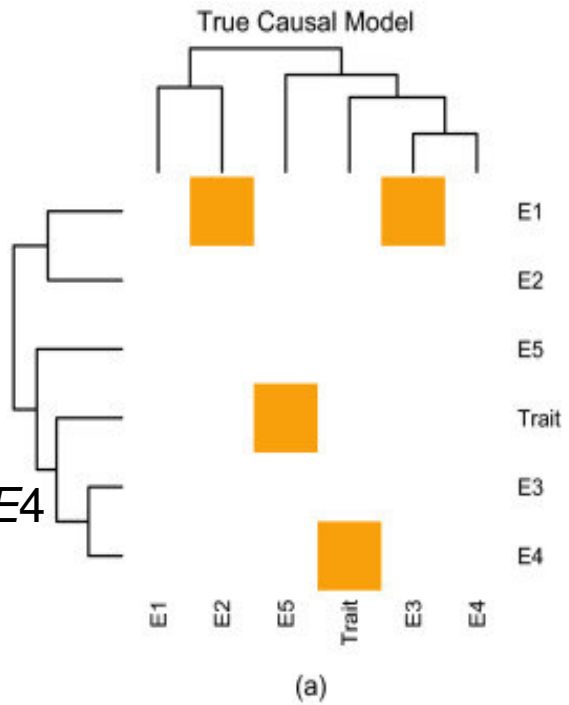
Robustness analysis
Fsp27 is a causal driver of a biologically important co-expression module

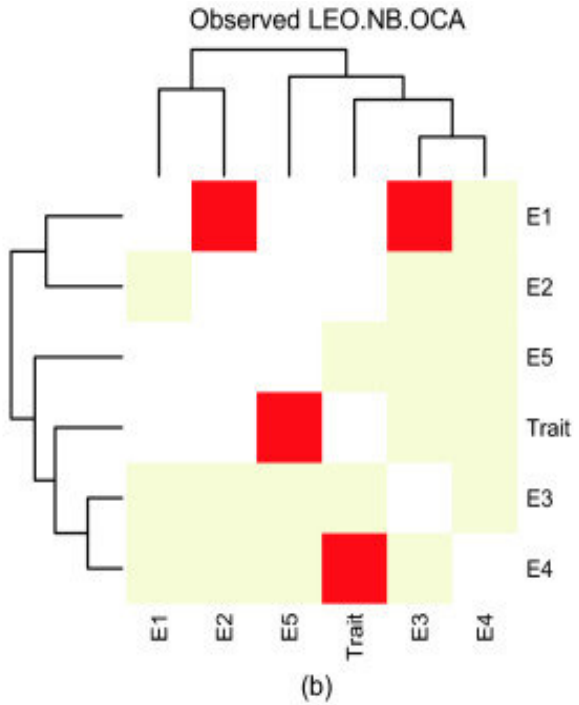
- LEO.NB(Fsp27-> MEblue) with respect to different choices of genetic markers sets (x-axis)
- Here we used automatic SNP selection to determine whether Fsp27 is causal of the blue module gene expression profiles.
- Both LEO.NB.CPA and LEO.NB.OCA scores show that the relationship is causal.

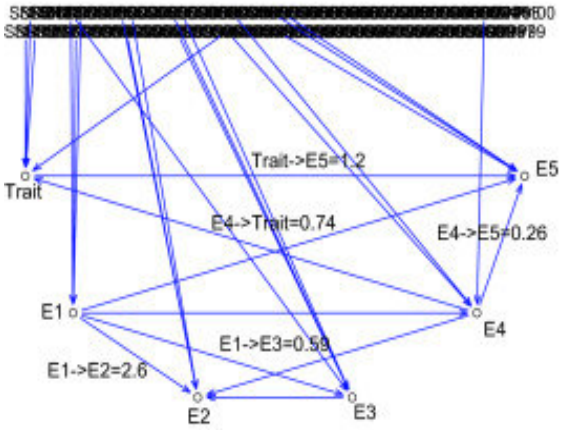


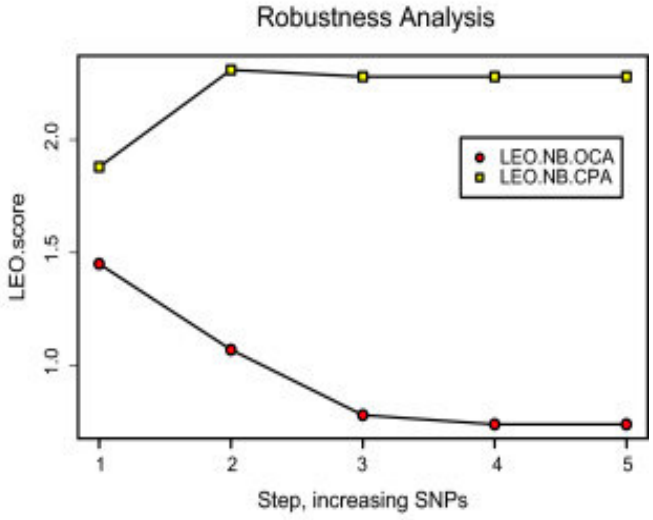
Multi edge simulations

$E1 \rightarrow E2$
$E1 \rightarrow E3$
$E3 \leftarrow HiddenConfounder \rightarrow E4$
$E4 \rightarrow Trait$
$Trait \rightarrow E5.$









Conclusion

- Genetic markers allow one to derive causality tests that can be used to assess the causal relationships between different traits.
- Systems genetic approaches that combine network methodology with traditional gene mapping approaches promise to bridge the chasm between sequence and trait information.
- An integrated gene screening approach can be used to find highly connected intramodular hub genes that are upstream of clinically interesting modules.

Software and Data Availability

- R software tutorials etc can be found online
- www.genetics.ucla.edu/labs/horvath/aten/NEO/
- Google search
  - weighted co-expression network
  - “WGCNA”
  - “co-expression network”
- http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork

Acknowledgement

- Doctoral dissertation work of Jason Aten
- (Former) lab members: Peter Langfelder, Jun Dong, Tova Fuller, Ai Li, Wen Lin, Anja Presson, Bin Zhang, Wei Zhao
- Collaborators
- Mice: Jake Lusis, Tom Drake, Anatole Ghazalpour