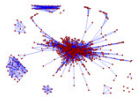


Systems Genetic Approaches for Studying Complex Traits

Steve Horvath
Human Genetics, Biostatistics
University of California, Los Angeles

Contents

- Using genetic markers to orient the edges in quantitative trait networks: the NEO software.
  - *Aten JE, Fuller TF, Lusis AJ, et al (2008) BMC Systems Biology 2008, 2:34. April 15.*
  - *Chapter 11 in Springer book "Weighted Network Analysis. Applications in Genomics and Systems Biology"*
- Application
  - *Plaisier et al (2009) A Systems Genetics Approach Implicates USF1, FADS3, and Other Causal Candidate Genes for Familial Combined Hyperlipidemia. PloS Genetics 2009;5(9)*

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

Aten JE, Fuller TF, Lusis AJ, et al (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

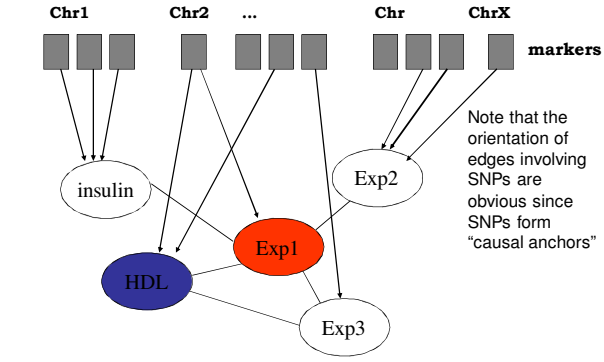
- Question: Can genetic markers help us to dissect causal relationships between gene expression- and clinical traits?
- Answer: yes, many authors have addressed this question both in genetics and in genetic epidemiology.
  - Vast literature->google search

Fundamental paradigm of biology can be used for inferring causal information

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

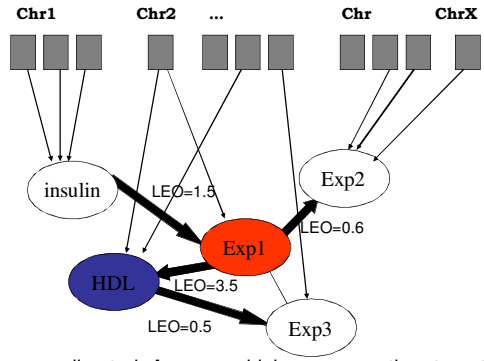
SNP -> gene expression

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)

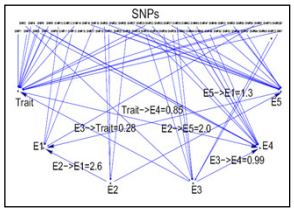
Output

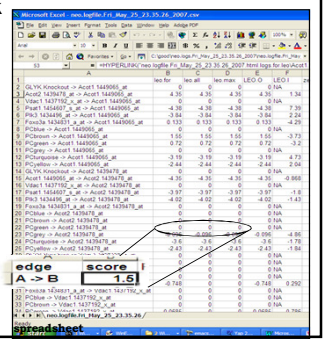
- Scores for assessing the causal relationship between correlated quantitative variables

Output of the NEO software

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

- graph of the directed network



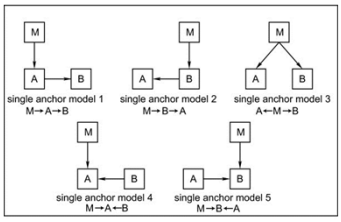


NEO Network Edge Orienting

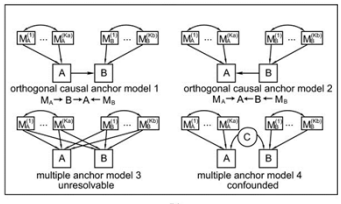
is a set of algorithms, implemented in R software functions, which compute scores for causal edge strength

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

Single marker causal models between traits A and B



Multi-marker causal models



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

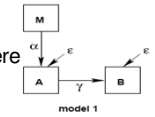
$$\chi^2 = (N-1)F(\hat{\theta}) \approx \chi^2(\frac{m(m-1)}{2} - t)$$

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

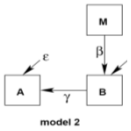
The higher the p-value, the better the causal model fits the data.

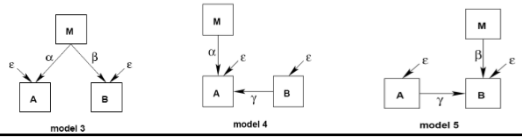
Causal models and corresponding model fitting p-values for a single marker M and the edge A-B.

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



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





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

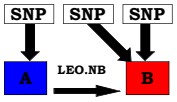
1) Merge genetic markers and traits

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



A Systems Genetics Approach Implicates USF1, FADS3, and Other Causal Candidate Genes for Familial Combined Hyperlipidemia
Chris Plaisier, Horvath S, Huertas-Vazquez A, Cruz-Bautista I, Herrera MF, Tusie-Luna T, Aguilar-Salinas C, Paivi Pajukante. PloS Genetics 2009;5(9)

Familial combined hyperlipidemia

- FCHL is a common atherogenic dyslipidemia conferring nearly two-fold greater risk for coronary heart disease.
- FCHL is characterized by familial segregation of elevated fasting plasma triglycerides (TGs), total cholesterol (TC), or both
- Another common characteristic of FCHL is elevated levels of fasting plasma apolipoprotein B (ApoB)

SNP rs3737787 in LD with USF1

- Linkage analysis and allelic association studies identified association within the region of chromosome 1q21-q23 consistently linked to FCHL with the associated linkage disequilibrium (LD) bin containing variants in upstream transcription factor 1 (USF1)
- A SNP (SNP rs3737787 residing in the 3′ UTR of USF1 captures the disease-associated signal
- Previous studies involving direct sequencing, extensive genotyping and gene expression analyses of the USF1 region have not identified any SNPs in the rs3737787 LD bin altering the coding sequence or the expression of USF1 itself in fat or lymphoblasts
- It has, however, been demonstrated that genes known to be regulated by USF1 were differentially expressed between rs3737787 genotype groups in Finnish fat biopsies.
- The direct targets of USF1 were previously identified using chromatin immunoprecipitation and high-resolution promoter microarrays (ChIP-Chip).

Mexican FCHL Families

- Originally, 872 individuals from 74 Mexican FCHL families were collected.
- 70 extremely discordant individuals
  - The 90th age-sex specific Mexican population percentiles for TGs and TC were used to determine the affection status
- Gene expression data: Affymetrix U133Plus2
- Our sample size of 70 extremely discordant individuals provided 80% power to detect a significant association (p-value≤0.05) with correlation coefficient = 0.33

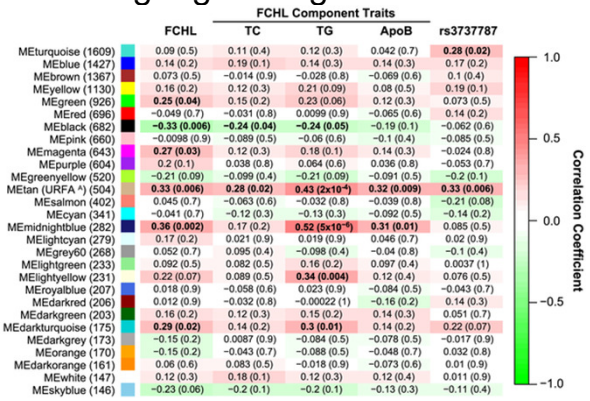
Effect of rs3737787 on FCHL is mediated through the transcription factor USF1

- We observed 972 genes (gene expression profiles) significantly correlated with rs3737787 genotypes using an additive model.
- The rs3737787 correlated genes had significant overlap both with
  - i) the set of USF1 regulated genes identified in our USF1 over-expression experiment (n = 277; p-value = 3.0×10^{-5}; fold-enrichment = 1.22)
  - and ii) the previously published genes identified by ChIP-Chip which are directly regulated by USF1 (n = 117; p-value = 0.0051; fold-enrichment = 1.23).
  - Furthermore, we also observed significant overlap between the rs3737787 correlated genes and the 2,189 genes differentially expressed between FCHL cases and normolipidemic controls (n = 245; p-value = 0.0030; fold-enrichment = 1.16) supporting a link from rs3737787 to FCHL etiology.
- Taken together, the overlap between rs3737787 correlated genes and genes regulated by USF1 suggest that the effect of rs3737787 on FCHL is mediated through the transcription factor USF1.

WGCNA analysis leads to 28 modules and 2*28 multiple tests

- Using blockwise module detection, WGCNA method clustered the 14,942 gene expression probes on the Mexican FCHL case/control microarrays into 28 gene co-expression modules.
- The module eigengene (ME) of each module was correlated with the quantitative FCHL component traits: TC, TG and ApoB.
- Regarding Bonferroni correction: Given the high level of correlation between the FCHL component traits (correlation TC & TG = 0.57; correlation TC & ApoB = 0.90; correlation TG & ApoB = 0.59), we approximate the total number of independent tests to be 2. Therefore, a Bonferroni correction would have to account for a total of 56 multiple comparisons (28 modules×2 independent tests).
- This highlights a major statistical advantage of our module based analyses over conventional differential expression analyses which have to account for tens of thousands of multiple comparisons.

Eigengene Significances



USF1-Regulated FCHL-Associated (URFA) Module Characterization

- Tan module (504 genes) was renamed URFA module
- The fact that the URFA ME was associated with rs3737787 reflects the fact that most module genes are at least partially regulated by this SNP
- Gene ontology enrichment
  - Cellular Lipid Metabolic Processes (p-values = 1.0×10–5) and Lipid Metabolic Processes (9.3×10–6)
- The URFA ME accounts for 10% of the variation of FCHL, 6% of TC, 17% of TG, and 9% of ApoB

NEO analysis

- To evaluate whether the module causally affects FCHL component traits, we utilized the Network Edge Orienting (NEO) R software package
- Since we are only considering a single SNP (rs3737787) we computed LEO.NB.SingleMarker scores for the causal orientation of a ME→trait.
- The LEO.NB.SingleMarker score is a relative model fitting index for the causal model rs3737787→ME→trait relative to alternative causal models
- We required that the our causal model fit at least two times better than the next best alternative model, which equates to a LEO.NB.SingleMarker score of 0.3

NEO analysis indicates that the URFA eigengene is upstream of TG levels

- We found sufficient evidence to infer a causal relationship between the URFA ME and fasting plasma TG levels (LEO.NB.SingleMarker score = 0.31), the key component trait of FCHL.
- The LEO.NB.SingleMarker score for the URFA ME to FCHL was 0.25.

Using NEO to screen for causal candidate genes

- We then used the NEO software to prioritize genes inside the URFA module by calculating the LEO.NB.SingleMarker scores evaluating the causal model (rs3737787→gene expression→trait).
- We identified
  - 18 causal candidate genes for FCHL,
  - 171 causal candidate genes for fasting plasma TGs levels
  - 13 causal candidate genes for both FCHL and TG
- Since our interest was in FCHL disease status, we characterized the 18 causal candidate genes for FCHL disease status as potential candidate genes for genetic association studies in Mexican FCHL families

Table of 18 causal candidate genes

Affymetrix ProbeSet	Gene Name	Associations	Network Connectivity	Kathiresan GWAS Minimum Regional P-values[B]	
				LDL	TG
214033_at	*ABCC6*	CAD, HDL, PSE, TG	34.1	7.5×10^{-3} (rs212077)	1.2×10^{-4} (rs9924674)
203925_at	*GCLM*	AvgIMT, CHD, MI, T2D, VCI	33.1	8.4×10^{-4} (rs2281525)	4.1×10^{-3} (rs12070273)
209740_s_at	*PNPLA4*	NA	28.3	NA[C]	NA[C]
204257_at	*FADS3*	HDL, PUFA, TG	27	4.2×10^{-4} (rs174549)	3.3×10^{-7} (rs102275)
227117_at	*XPOT*	NA	20.3	1.2×10^{-3} (rs10878151)	1.1×10^{-3} (rs11504159)
212799_at	*STX6*	NA[A]	17.2	1.5×10^{-3} (rs17299701)	9.1×10^{-3} (rs6658713)
204057_at	*IRF8*	HEPC	15.3	2.6×10^{-4} (rs903194)	5.7×10^{-4} (rs4843966)
214152_at	*CPR8*	NA	12.9	7.9×10^{-3} (rs12902248)	1.6×10^{-2} (rs4774780)
214696_at	*C17orf91*	NA	12.1	7.9×10^{-5} (rs17761734)	6.6×10^{-4} (rs2955626)
205404_at	*HSD11B1*	BC, HT, T2D	11.1	1.6×10^{-4} (rs6659502)	2.1×10^{-3} (rs7536585)
236664_at	*AKT2*	MetS, T2D, TC/HDL	9.8	7.0×10^{-3} (rs10412191)	4.6×10^{-4} (rs4803342)
203625_x_at	*SKP2*	NA	8.9	9.7×10^{-3} (rs1610218)	4.8×10^{-4} (rs6895261)
50374_at	*C17orf90*	NA	8.5	2.5×10^{-3} (rs11150780)	1.8×10^{-2} (rs7210742)
229144_at	*KIAA1026*	NA	5.3	9.3×10^{-4} (rs9442193)	3.5×10^{-3} (rs12141589)
205452_at	*PIGB*	NA	3.6	7.9×10^{-3} (rs12902248)	1.6×10^{-2} (rs4774780)
209268_at	*VPS45A*	NA[A]	2.7	5.1×10^{-3} (rs7537292)	1.3×10^{-2} (rs6587552)
214252_s_at	*CLN5*	NCL	2.2	3.5×10^{-3} (rs851251)	9.5×10^{-3} (rs1537063)
228641_at	*CARD8*	AZ, RA	0.8	4.9×10^{-3} (rs3760802)	1.6×10^{-4} (rs11669775)

AvgIMT indicates average intima-media thickness; AZ, Alzheimer's disease; BC, body composition; CAD, coronary artery disease; CHD, coronary heart disease; HDL, low levels of high-density lipoprotein cholesterol; HEPC, hepatitis-C; HT, hypertension; MI, myocardial infarction; NCL, neuronal ceroid-lipofuscinosis; PSE, pseudoxanthoma elasticum; PUFA, polyunsaturated fatty acids; T2D, type II diabetes; TC/HDL, total cholesterol/high-density lipoprotein cholesterol ratio; TG, hypertriglyceridemia; and VCI, vascular cognitive impairment.

[A]*STX6* and *VSP45A* proteins are known to physically interact.

[B]Minimum P-value from a region spanning 1 Mbp on either side of the gene from Kathiresan et al., 2009 [38].

[C]*PNPLA4* resides on the X chromosome, and there weren't any SNPs tested within the specified region in the Kathiresan et al., 2009 [38].

doi:10.1371/journal.pgen.1000642.t004

Prior literature on 18 causal candidates

- Three of the FCHL causal candidate genes were directly related to lipid phenotypes (ABCC6, AKT2, HSD11B1), and two others were likely to be related to lipid phenotypes (FADS3, PNPLA4) via homology.
- We also identified genes which were related to atherogenic processes such as inflammation (CARD8, ICSBP1, STX6) and reactive oxygen species (GCLM).
- Importantly, some of the genes causally linked to FCHL (ABCC6, AKT2, FADS3, GCLM, HSD11B1) have already been associated with FCHL related traits in humans. Studies in mice have also demonstrated that genetic manipulation of Abcc6, Akt2 and Hsd11b1 affect FCHL component traits or related phenotypes.
- **Together these data support a causal association between the 18 causal candidate genes from the URFA module and FCHL**.
- Among the 18 genes there are also putative genes and genes with little known function.
- Additional biological validation studies are warranted

Fatty acid desaturase 3 (FADS3)

- Interestingly, variation from the FADS1-2-3 genomic region was previously associated with TGs in a recent meta-analysis of GWAS in Caucasians.
- We chose to follow-up the SNP rs174547 residing in the FADS1-2-3 locus which was significantly associated with TGs at the genome-wide level in this previous meta-analysis.
- The same study demonstrated that the SNP rs102275, in complete LD with rs174547 in Caucasians, predicted the expression of FADS1 and to a lesser extent FADS3 in human liver (Kathiresan 2009).
- We hypothesized that because FADS3 expression was associated with FCHL, any variation affecting the expression of FADS3 could be associated with FCHL or an FCHL component trait, especially TGs.
- Therefore we genotyped both rs174547 and rs102275 in the Mexican FCHL case/control fat biopsies (n = 70). Our results replicated the findings for the FADS1-2-3 locus in the Mexican population.

Why did we use WGCNA?

- First, co-expression modules may be comprised of sets of genes that are likely to be co-regulated by similar factors (e.g. shared transcription factors, genetic variants or environmental effects).
- Second, modules (and corresponding module eigengenes) represent a biologically motivated data reduction method which greatly alleviates the multiple comparison problem inherent in genomic data analysis.
- Third, kME (intramodular connectivity) can be used to provide annotation tables for module membership e.g. to the URFA module

Conclusion

- By integrating a genetic polymorphism with genome-wide gene expression levels, we were able to attribute function to a genetic polymorphism in the USF1 gene.
- We demonstrate that this genetic polymorphism in USF1 contributes to FCHL disease risk by modulating the expression of a group of genes functionally related to lipid metabolism, and that this modulation is mediated by USF1.
- Our unbiased module detection analysis identified a module (the URFA module) that was associated with rs3737787 genotypes, fasting plasma TG levels, FCHL disease status, and contained genes that are causal drivers of TG levels.
- Our approach provides insight to how the SNP rs3737787 confers increased risk for FCHL, by demonstrating that it regulates the URFA module eigengene which in turn contributes to increased TG levels, a key component trait of FCHL.
- One of the genes whose expression is modulated by USF1 is FADS3, which was also implicated in a recent genome-wide association study for lipid traits.
- We demonstrated that a genetic polymorphism from the FADS3 region, which was associated with triglycerides in a GWAS study of Caucasians, was also associated with triglycerides in Mexican FCHL families.
- Our analysis provides novel insight into the gene expression profile contributing to FCHL disease risk, and identifies FADS3 as a new gene for FCHL in Mexicans.

Software and Data Availability

- For R code see “Corrected Tutorial for Chapter 12” at the following webpage:
  - http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Book/
- Or the original NEO webpage:
- www.genetics.ucla.edu/labs/horvath/aten/NEO/

Acknowledgement

- (Former) students and Postdocs:
- Peter Langfelder, Jun Dong, Tova Fuller, Mike Oldham, Ai Li, Wen Lin, Jeremy Miller, Chris Plaisier, Anja Presson, Bin Zhang, Wei Zhao, Jason Aten, Lin Song
- Colleagues/Collaborators
- Jake Lusis, Paivi Pajukante, Dan Geschwind, Giovanni Coppola