Predicting Epistatic Interactions Using Information and Network Theory for Continuous Phenotypes

Krishna Bathina
bathina@umail.iu.edu
krishnacb.com

Indiana University
School of Informatics,
Computing, and Engineering

Predicting *Epistatic Interactions* Using *Information* and *Network Theory* for *Continuous Phenotypes*

Still working on a better title…

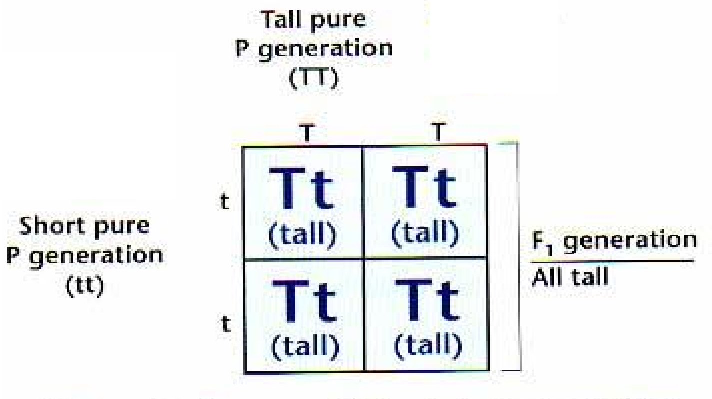
Genetics

Genetics

Motivation
Mutual Information
Information Gain
Finding Epistasis
Test Run

Genes & Alleles & Single Nucleotide Polymorphisms (SNPs)

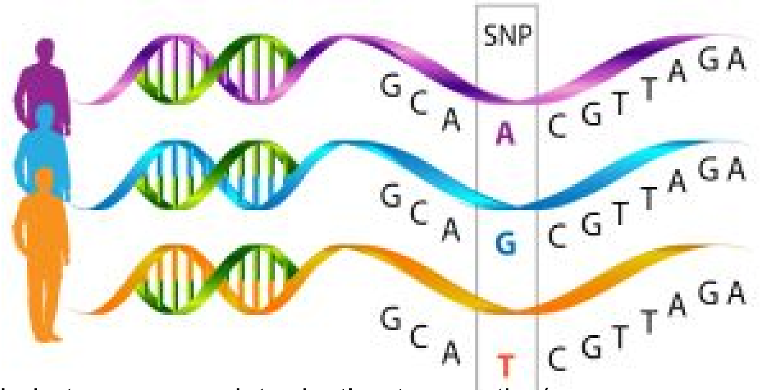
- Gene - basic unit of heredity - a region of nucleotides in DNA
- Allele - variant form of gene



	T	T
t	Tt (tall)	Tt (tall)
t	Tt (tall)	Tt (tall)

A Punnett Square of Mendel's Second Step

- Single Nucleotide Polymorphisms (SNPs) - variants at a single base that occur in at least 1% of the population
    - Mutation if less than 1%



https://neuroendoimmune.wordpress.com/2014/03/27/dna-rna-snp-alphabet-soup-or-an-introduction-to-genetics/

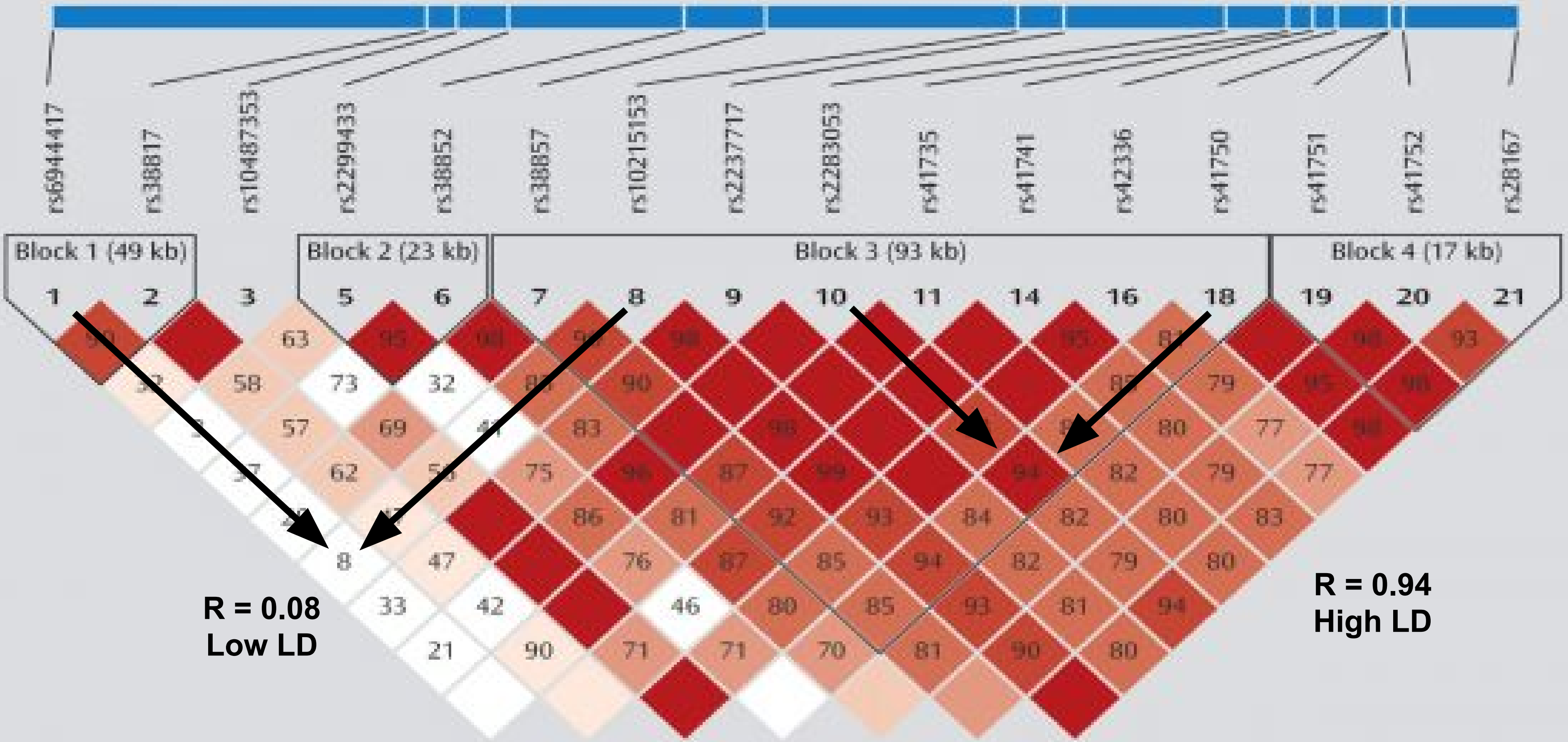
Linkage Disequilibrium (LD)

- LD - state of association between different alleles in a population
  - Low LD - random association
  - High LD - correlated association
- Coefficient of LD
  - Frequency of allele a: p_a
  - Frequency of allele b: p_b
  - Frequency of ab haplotype: p_{ab}

$$D = p_{ab} - p_a p_b$$

$$r = \frac{D}{p_a p_b (1 - p_a)(1 - p_b)}$$

https://estrip.org/articles/read/tinypliny/44920/Linkage_Disequilibrium_Blocks_Triangles.html



R = 0.08
Low LD

R = 0.94
High LD

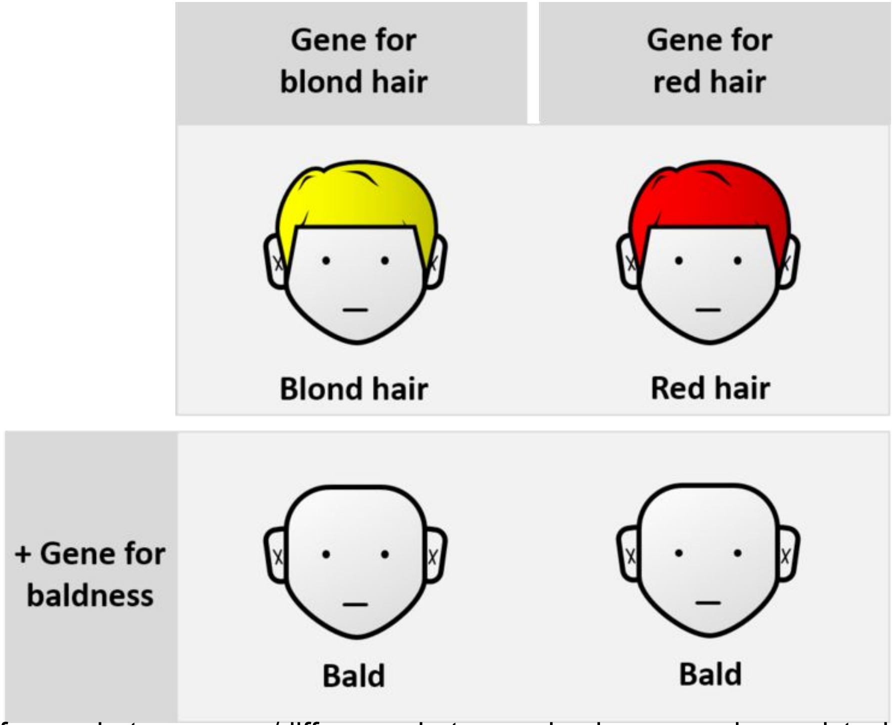
International HapMap Project

Epistasis

The effect of one gene is ***modified*** by the presence (or lack) of another gene.

- Synergistic effects
- Antagonistic effects

Dominant Epistasis - Baldness is dominant to blond and red hair



http://www.differencebetween.com/difference-between-dominance-and-vs-epistasis/

Motivation

Genetics

Motivation

Mutual Information

Information Gain

Finding Epistasis

Test Run

- Traditional **GWAS** only reports significant SNPS based on **single interactions**
- **GWAS** too slow to **discover joint interactions**
- Many complicated **proposed statistics**
- **Similar method proposed** by Hu et al, for binary phenotypes - Moore Lab
- **Continuous more common** than binary phenotypes

Hu, Ting, et al. "Genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. Vol. 20. NIH Public Access, 2015.

Mutual Information

Genetics
Motivation
Mutual Information
Information Gain
Finding Epistasis
Test Run

Definition

The amount of information learned about one variable from information about the other.

Given:

- Random variables: X,Y
- Joint probability function: p(x,y)
- Marginal probability distribution functions: p(x),p(y)

$$I(X,Y)_D = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$I(X,Y)_C = \int_x \int_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy$$

$$= D_{KL}(P(X,Y) || P(X)P(Y))$$

Example

X	Y
1	1
1	2
2	2
2	3
3	3

$$I(X,Y)_D = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$= p(1,1)*\log\frac{p(1,1)}{p(1)p(1)} + p(1,2)*\log\frac{p(1,2)}{p(1)p(2)} + \ldots$$

$$= 0.639$$

What about Mixed Data? (Ross et al 2014)

- Days of the week and traffic levels
- DNA bases and phenotype expression levels
- Population and City Size

$$I(X,Y) = \langle \log \frac{p(x_i, b_i)}{p(x_i)p(b_i)} \rangle$$

Binning data:

- each bin has N data points
- discrete variable X
- continuous variable Y
- probability of x_i $p(x_i)$
- fraction of data that falls in the same bin as y_i $p(b_i)$
- joint probability function $p(x_i,b_i)$.

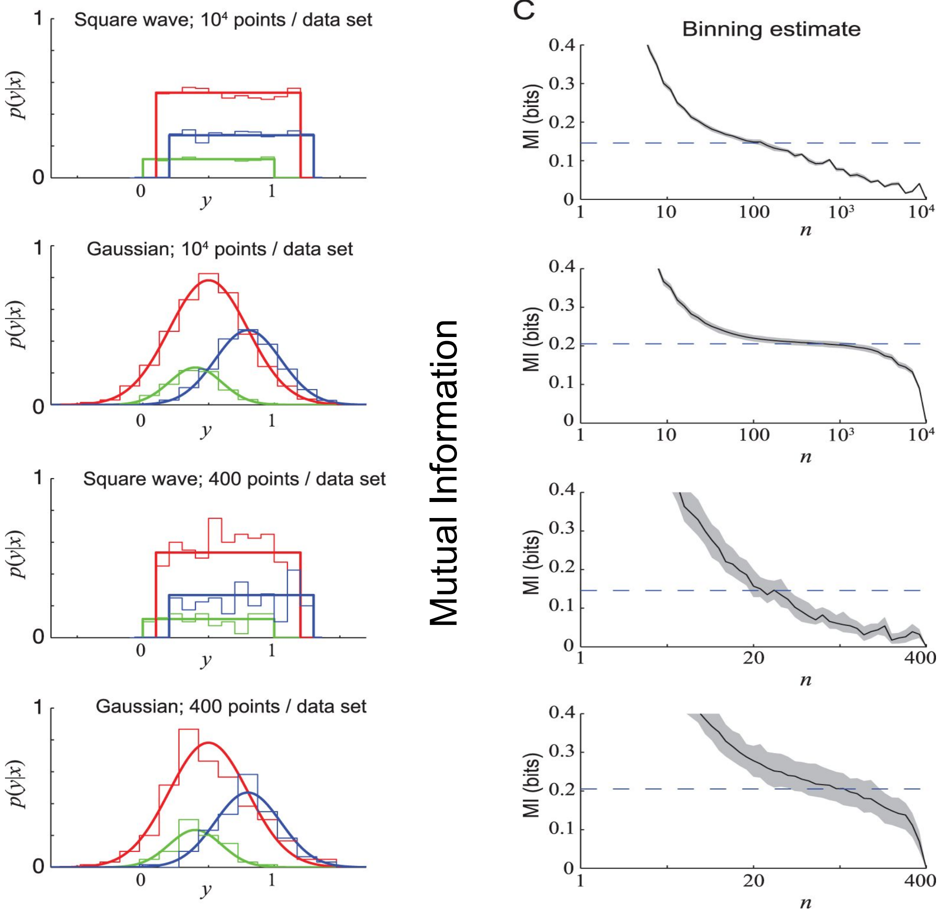
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0087357#pone.0087357-Kraskov1

$$\{(x, y)|$$

$$x \in [R, B, G]$$

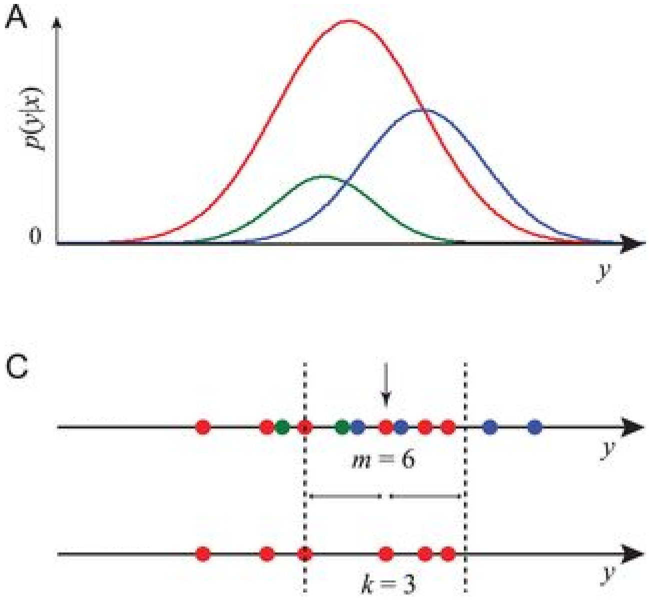
$$\wedge$$

$$y \in R\}$$



Estimation using binning relies on bin size - *not reliable*

K-Nearest Neighbors Method (Ross et al 2014)



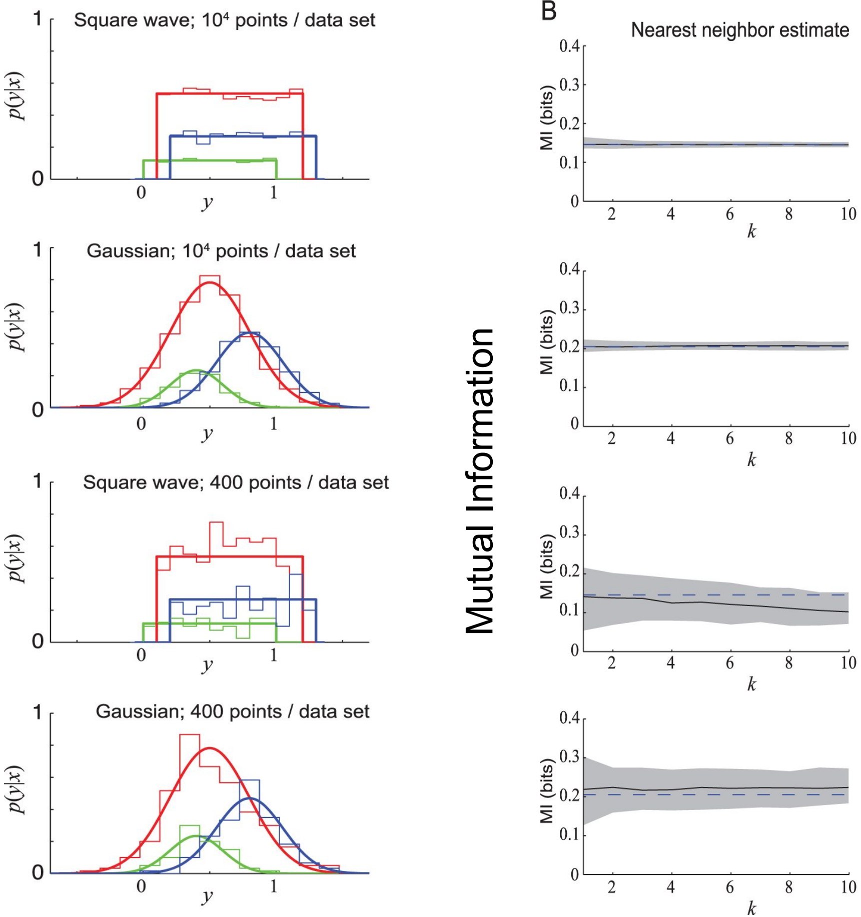
- **N** = number of data points: **12**
- **x$_i$** = category of data point i: **Red**
- **N$_x$** = number of data points in the same category as x: **6**
- **K** = nearest neighbors: **3**
- **M** = *total* number of data points within the radius of the farthest k-neighbor datum of category x: **6**

$$\psi(x) = \frac{d}{dx}\ln(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)}$$

Information Gain

$$I_i = \psi(N) - \psi(N_{x_i}) + \psi(k) - \psi(m_i)$$

$$\begin{aligned} I(X,Y) &= \langle I_i \rangle \\ &= \psi(N) - \langle \psi(N_x) \rangle + \psi(k) - \langle \psi(m) \rangle \end{aligned}$$



Estimation using K-nearest neighbor: more accurate and more precise

Information Gain

Genetics
Motivation
Mutual Information
Information Gain
Finding Epistasis
Test Run

Information Gain (McGill 1954)

Information Gain(X,Y;Z): a measure of the **combined interaction** between joint variables X and Y with Z

- Amount of **synergy in the set (X,Y,Z) beyond** the synergy from **the subsets of (X,Y,Z)**
- The difference between the mutual information of the **joint variables X and Y with Z** from the **individual mutual information**

$$IG(X,Y;Z) = I(X,Y,Z) - I(X,Z) - I(Y,Z)$$

McGill, W J (1954). "Multivariate information transmission". *Psychometrika*. **19**: 97–116. doi:10.1007/bf02289159

Example

X	Y	**Z**
1	1	**0**
2	2	**0**
2	2	**1**
2	3	**1**
1	1	**0**

$$I(X,Y;Z) - I(X;Z) - I(Y;Z)$$

$$= 0.395753 - 0.0138443 - 0.395753$$

$$= -0.0138443$$

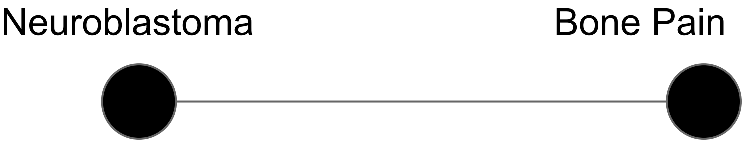
Joint interaction does not give any extra information

Finding Epistasis

Genetics
Motivation
Mutual Information
Information Gain
Finding Epistasis
Test Run

1a. Phenotype–Phenotype Network

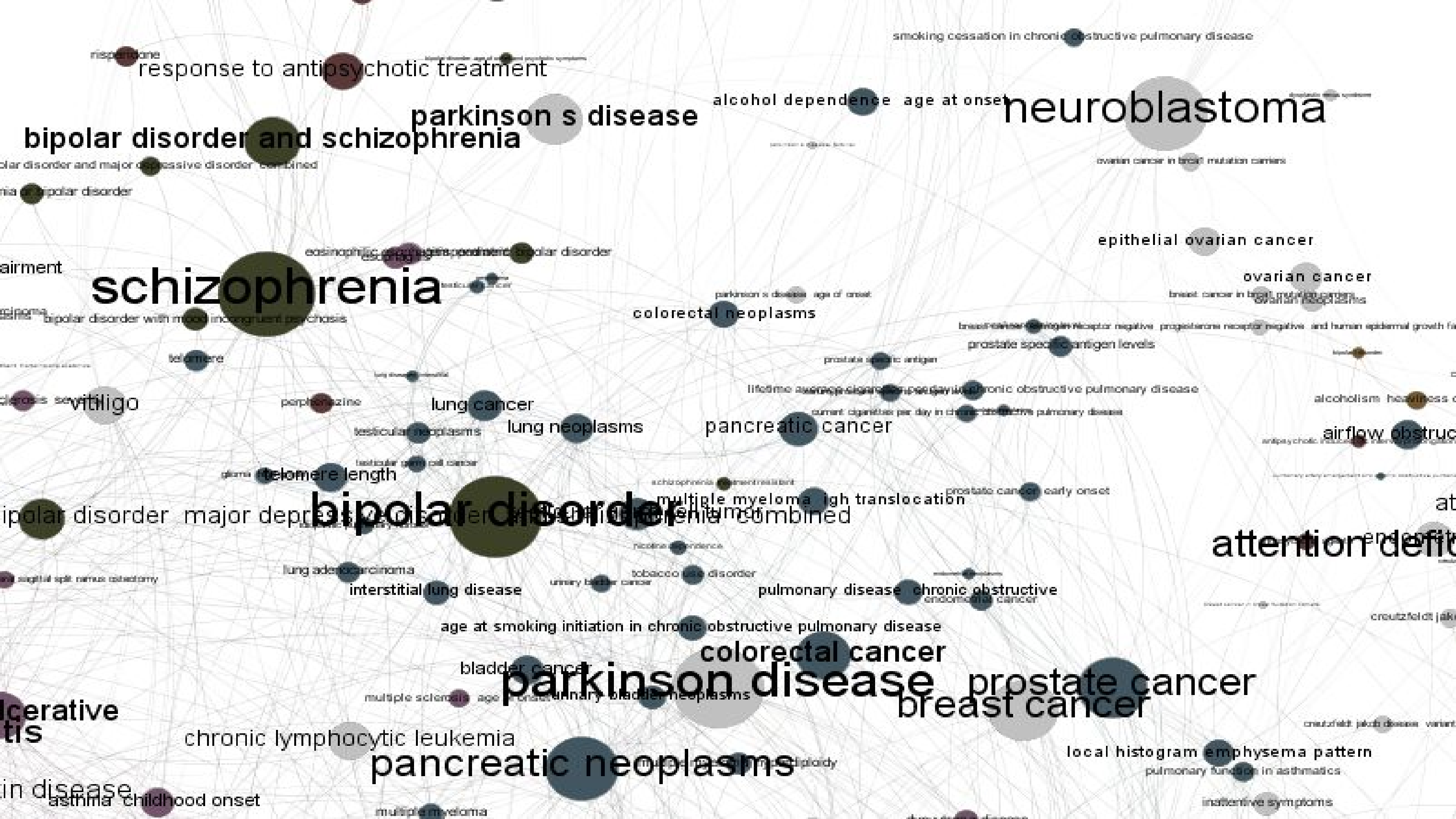
1. Dataset of Phenotypes and their **statistically significant associated SNPs** - federally funded studies
    a. dbGaP - Database of Genotypes and Phenotypes
    b. GWAS Catalog EMBL-EBI
2. **Phenotypes** = Nodes
3. **Jaccard Index of SNP overlap** = edge weights



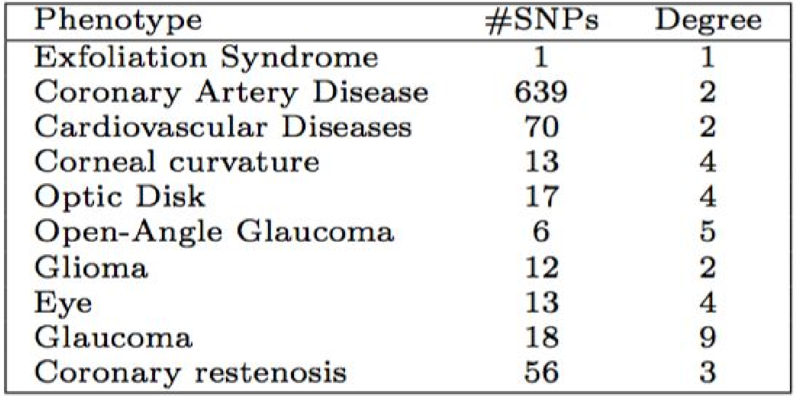
$$J = \frac{3}{8} = 0.375$$

SNP1
SNP2
SNP3
SNP4
SNP5
SNP6

SNP1
SNP2
SNP3
SNP7
SNP8

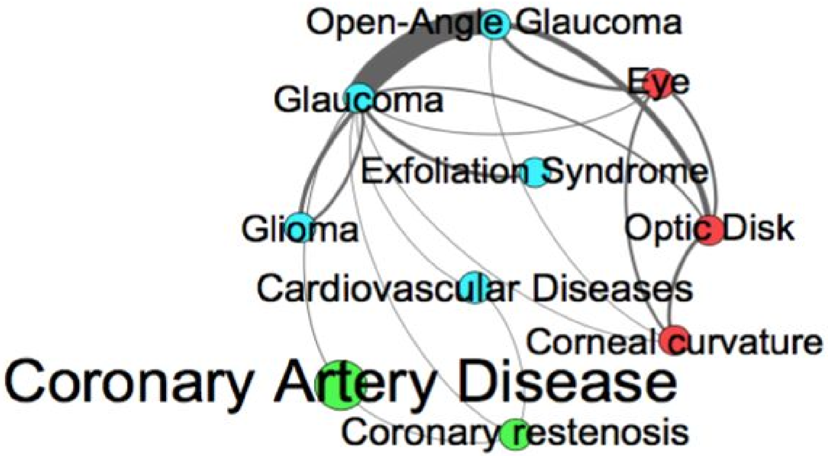

smoking cessation in chronic obstructive pulmonary disease
risperidone
response to antipsychotic treatment
parkinson s disease
alcohol dependence age at onset
neuroblastoma
bipolar disorder and schizophrenia
olar disorder and major depressive disorder combined
ovarian cancer in brca1 mutation carriers
nia or bipolar disorder
epithelial ovarian cancer
airment
ovarian cancer
schizophrenia
bipolar disorder
parkinson s disease age of onset
breast cancer in brca1 mutation carriers
ovarian neoplasms
bipolar disorder with mood incongruent psychosis
colorectal neoplasms
telomere
prostate specific antigen levels
prostate specific antigen
vitiligo
perphenazine
lung cancer
lifetime average cigarettes per day in chronic obstructive pulmonary disease
current cigarettes per day in chronic obstructive pulmonary disease
alcoholism heaviness
testicular neoplasms
lung neoplasms
pancreatic cancer
airflow obstruc
testicular germ cell cancer
glioma
telomere length
schizophrenia treatment resistant
prostate cancer early onset
multiple myeloma igh translocation
bipolar disorder major depressive disorder and schizophrenia combined
bipolar disorder
attention defic
nicotine dependence
lung adenocarcinoma
sagittal split ramus osteotomy
tobacco use disorder
urinary bladder cancer
interstitial lung disease
pulmonary disease chronic obstructive
endometrial cancer
creutzfeldt jakob
age at smoking initiation in chronic obstructive pulmonary disease
colorectal cancer
bladder cancer
parkinson disease
prostate cancer
breast cancer
multiple sclerosis age of onset
urinary bladder neoplasms
lcerative
tis
creutzfeldt jakob disease variant
chronic lymphocytic leukemia
local histogram emphysema pattern
pancreatic neoplasms
pulmonary function in asthmatics
in disease
asthma childhood onset
inattentive symptoms
multiple myeloma

1b. Choose Subset of Phenotypes



Phenotype	#SNPs	Degree
Exfoliation Syndrome	1	1
Coronary Artery Disease	639	2
Cardiovascular Diseases	70	2
Corneal curvature	13	4
Optic Disk	17	4
Open-Angle Glaucoma	6	5
Glioma	12	2
Eye	13	4
Glaucoma	18	9
Coronary restenosis	56	3

(a)

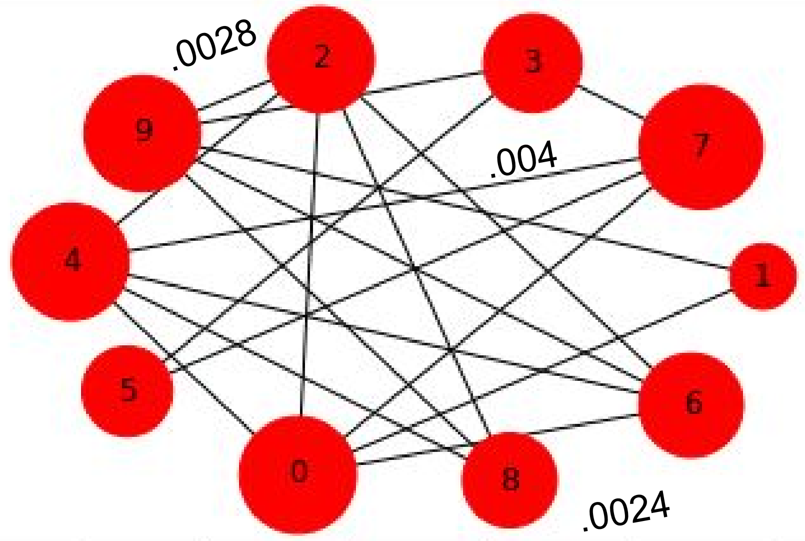


(b)

Hu, Ting, et al. "Genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. Vol. 20. NIH Public Access, 2015.

2. SNP-SNP Network

1. Build new network with relevant SNPs - Include SNPs in high LD
2. **SNPs** = Nodes
3. **Information Gain** = Edge weights
   a. The difference between the epistatic effect on the phenotype from the individual effects



$$IG(A,B;\mathcal{P}) = I(A,B;\mathcal{P}) - IG(A;\mathcal{P}) - IG(B;\mathcal{P})$$

3. Network Analysis

1. **Threshold network edges** from [0,max(IG)] in increments of 0.0001
   a. Only include edges with IG ≥ threshold
   b. Find size of largest connected component
2. **Create 100 new graphs** - shuffle phenotypes across subjects
   a. Repeat thresholding process
4. **Permutation Test** - find threshold for which the connected component is statistically larger in the original graph than the permutation graphs
5. Find most **central nodes**

4. SNP Annotation

Annotate discovered SNPs for current pathway information

SNPnexus

Test Run

Genetics
Motivation
Mutual Information
Information Gain
Finding Epistasis
Test Run

Data

‘The investigator must be a tenure-track professor, senior scientist, or equivalent’

-dbGaP

Mixed Linear Model:

- 4000 subjects
- 200 total SNPs
- MAF < 0.5 - Frequency of second most common allele
  - Uniform, Inversely proportional to frequency, etc.
- Risk variants assigned by HW equilibrium

Mixed Linear Model

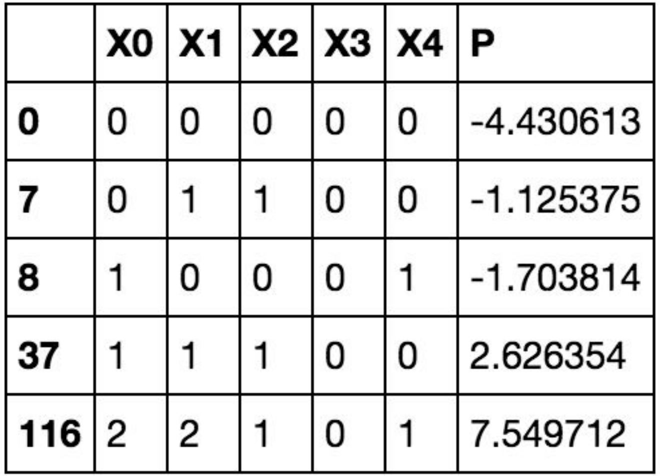
$$\underset{\text{Phenotype}}{P} = \overset{\text{Intercept}}{\beta_i} + \underset{\text{Effect size of epistatic interaction between SNP0 and SNP1}}{\beta_{0,1}} \overset{\text{Number of Risk Variants for SNP0 and SNP1}}{X_0 X_1} + \sum_{n=0}^{N} \underset{\text{Effect Size}}{\beta_n} \overset{\text{\# Risk Variants}}{X_n} + \overset{\text{Random Variation}}{\mathcal{N}(0, 1)}$$

Given **A** is the risk allele and **a** is the common allele

AA = 2 Risk Variants
Aa = 1
aa = 0

Result – 1 sample run

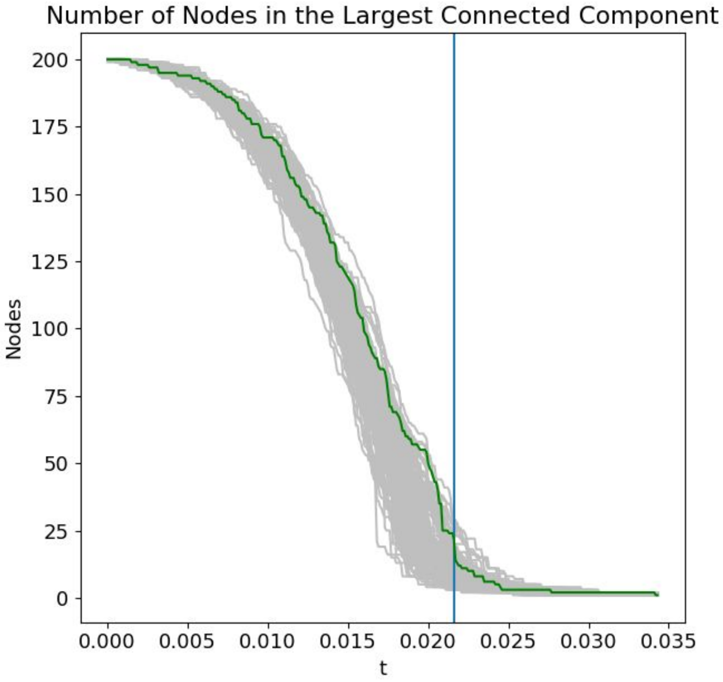
$$P = 1 + 2.2X_0X_1 + 1.5(X_0 + X_1) + \sum_{n=2}^{N} \mathcal{N}(0, 0.5)X_n + \mathcal{N}(0, 1)$$

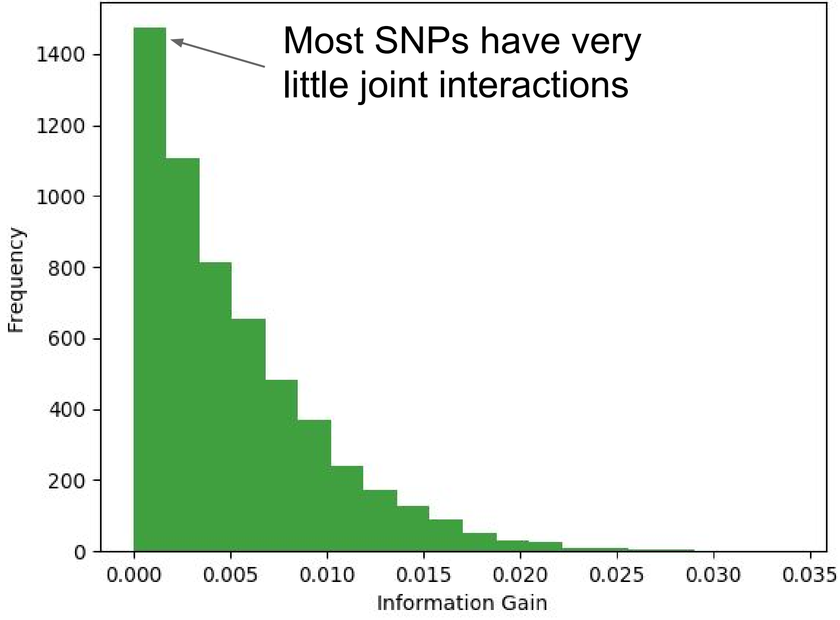


	X0	X1	X2	X3	X4	P
0	0	0	0	0	0	-4.430613
7	0	1	1	0	0	-1.125375
8	1	0	0	0	1	-1.703814
37	1	1	1	0	0	2.626354
116	2	2	1	0	1	7.549712

Interactions with **negative** IG: 53.8%
Interactions with **IG = 0**: 17.7%
Statistically Significant **cutoff** = 0.0216 (p = 0.05)

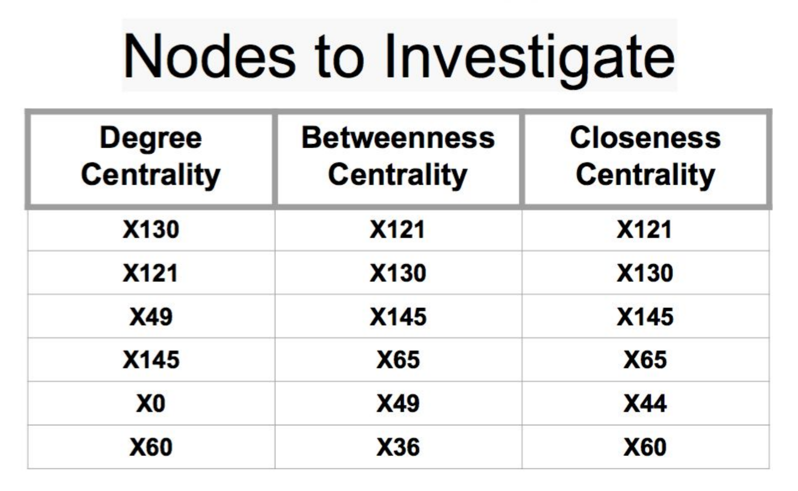
Result





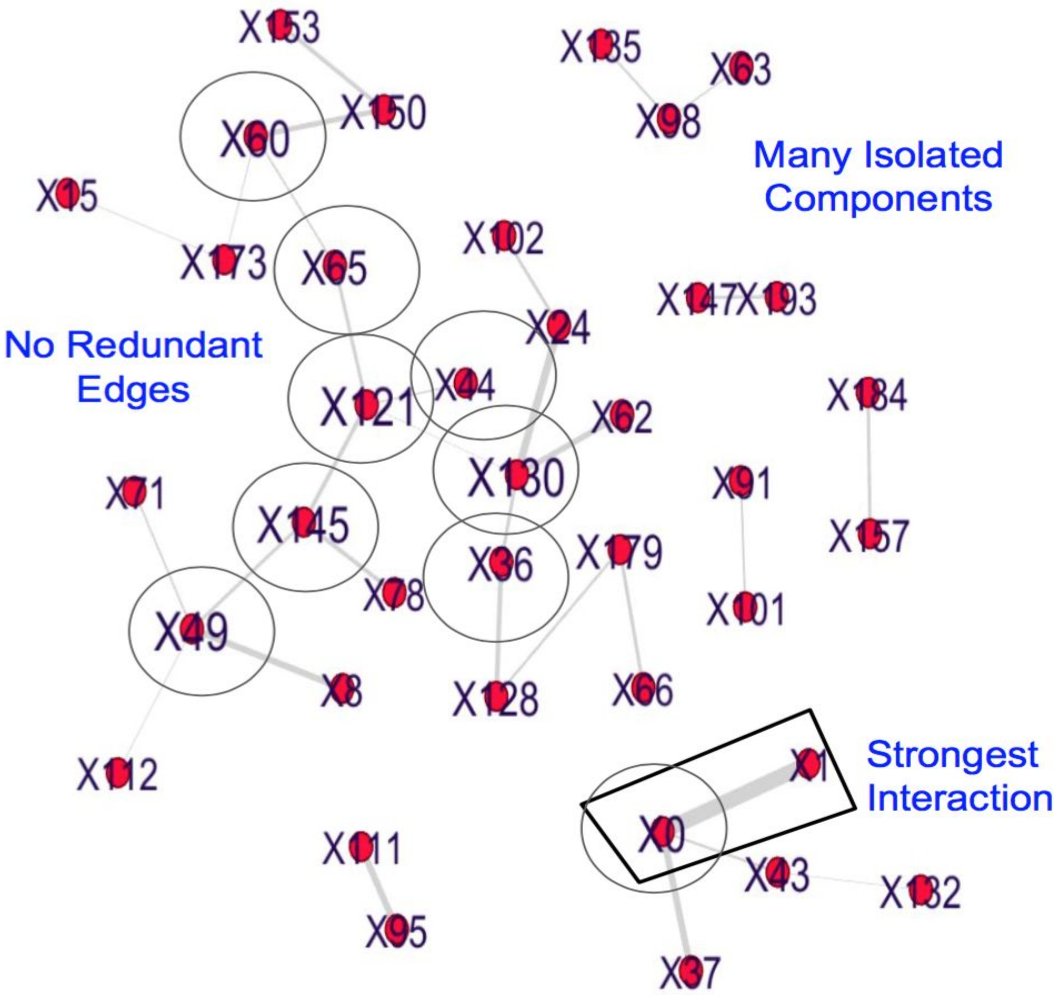
Most SNPs have very little joint interactions

Result



Nodes to Investigate

Degree Centrality	Betweenness Centrality	Closeness Centrality
X130	X121	X121
X121	X130	X130
X49	X145	X145
X145	X65	X65
X0	X49	X44
X60	X36	X60



Many Isolated Components

No Redundant Edges

Strongest Interaction

Future Work

Standard GWAS Method Evaluation

1. **Make series of toy datasets** over reasonable **parameter** ranges
    a. Need to check literature for possible values because parameters vary greatly by phenotype
2. **Compare method** with current, well established methods - find ranges in which new method does well
3. Compare **computational complexity and speed**

Intercept	**Distribution of Effect Sizes**	**Distribution of Risk variants**
Effect Size of Epistasis	**Number of Epistatic Interactions**	**Population Size**

Future Work cont.

1. Investigate **new ways to choose relevant phenotypes**
   a. 1° neighbors might be too restrictive.
   b. Looking at **communities** will be more informative for non-obvious phenotype relatedness
2. **Important Nodes** should not be found from trying every possible measure
   a. Each measure represents a specific kind of important node
3. **Extend Information Gain** to 3,4,5,...n variables - many different extensions
4. **Different measures** of co-interaction
   a. Not all measures can find triadic interactions in all distributions (Ryan James)
5. **Apply method on individual genomic data** from dbGaP.

Questions?