Modeling Epistatic Interactions Among Multiple Genetic Variants

Allan Jiang
The Jackson Laboratory
August 6, 2014

Mentor: Greg Carter

The Jackson Laboratory
Leading the search for tomorrow's cures

Understanding Complex Diseases

- Diabetes
- Alzheimer's
- Heart Disease
- Genetic background unclear
  - epistasis
  - pleiotropy
  - hairball

Untangling the hairball

- Detailed view of biological systems
  - Genotype and phenotype data
  - Across many individuals
  - Gene expression info → molecular level
- Can we model directed gene-phenotype relationships?
- What about directed gene-gene interactions?

Introducing cape

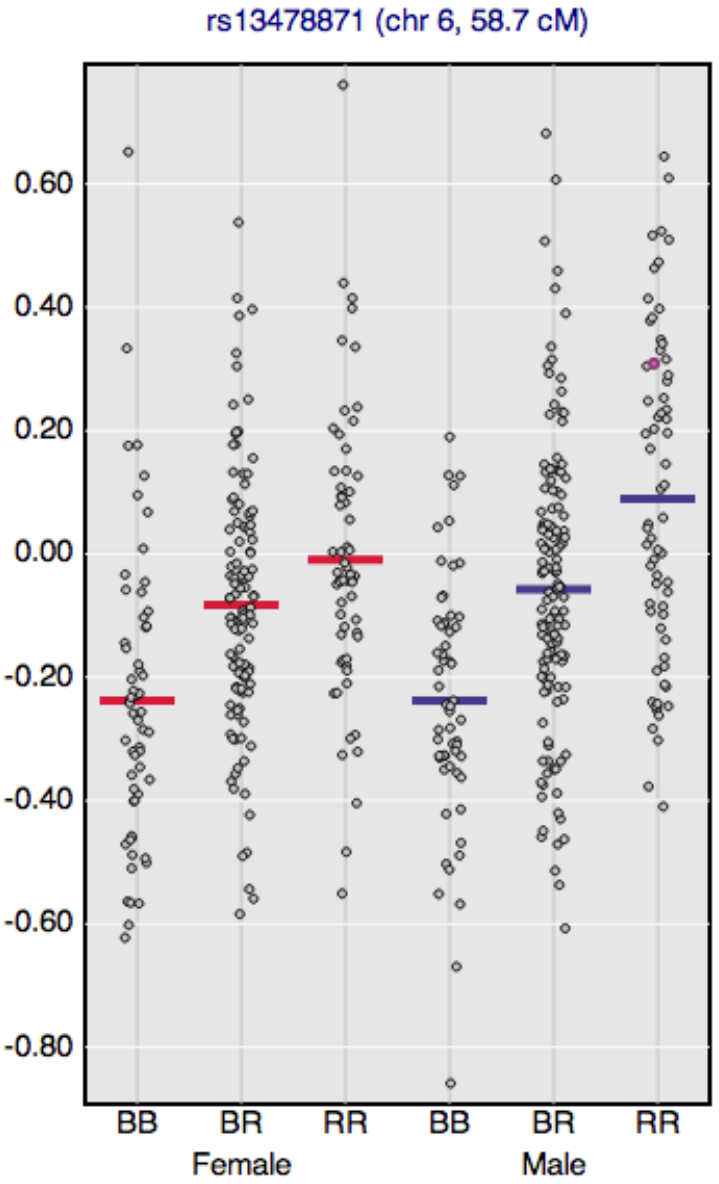
Math may be involved!

From Data to Network: The Data

- Genotype data
  - Data at various **markers** across the genome
  - Each marker is BB, BR, or RR
  - Arbitrarily define B as **reference** allele

- Phenotype data
  - Quantitative
  - Physiological or molecular

The Additive Assumption



- Lets us correlate genotype and phenotype
- Biological assumption brings us to causality
- Key assumption in the model

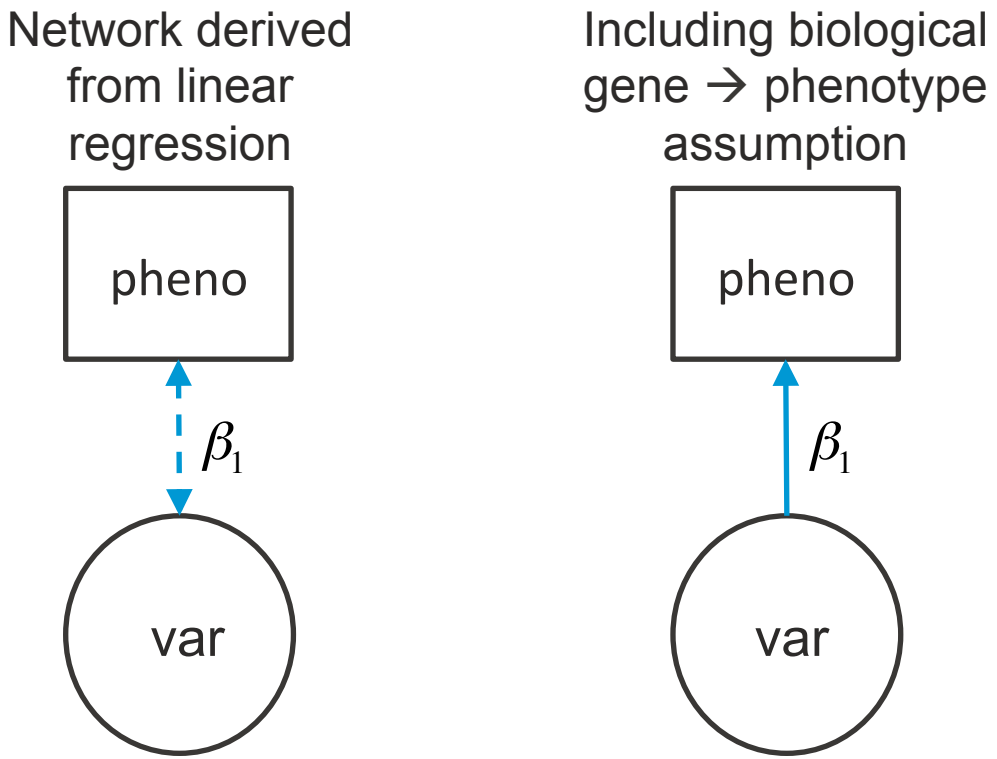
Single Marker Association

- Encode markers as 0, 1, 2
- Correlate genotype and phenotype values

$$y \sim \beta_0 + \beta_1 x_1 + \epsilon$$

- Significant β_1 value indicates linkage
- Genotypes explain phenotypic variance

Singlescan Network:



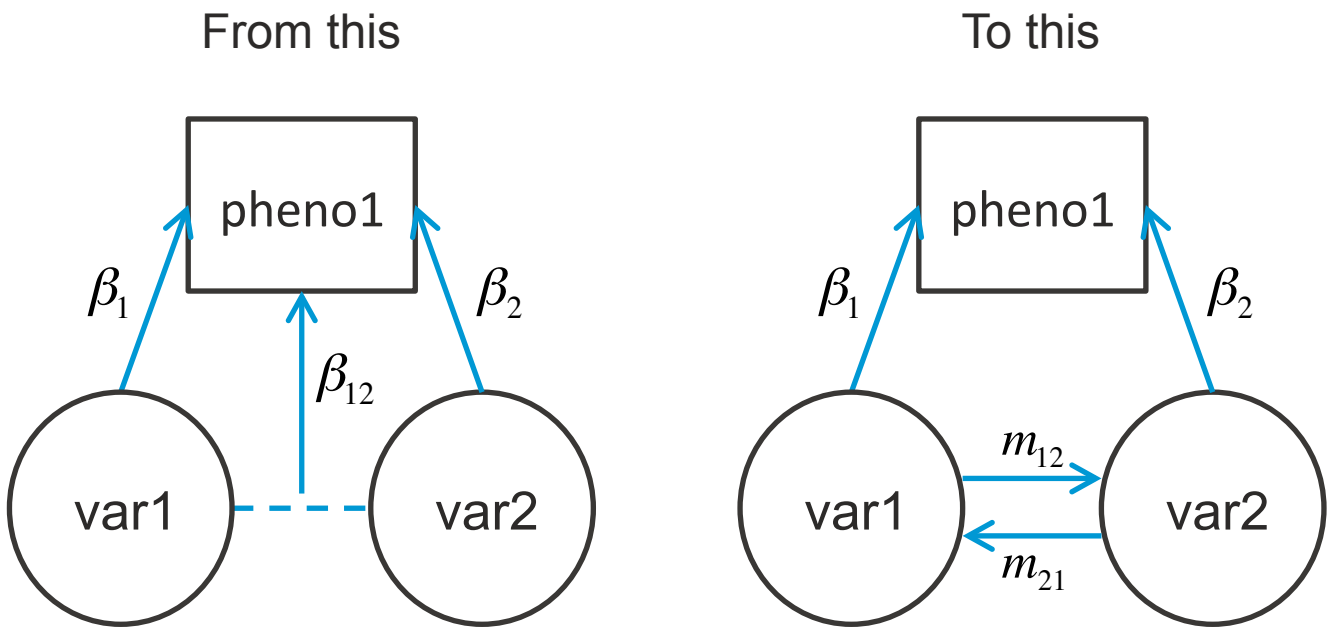
Two Marker Association (Pairscan)

- Just include more variants. And interaction term!

$$y \sim \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Main Effects}} + \underbrace{\beta_{12} x_1 x_2}_{\text{Interaction}} + \epsilon$$

- Variance in phenotype could be explained by interaction or main effects
- Significant β_{12} value indicates epistasis
  - No directionality?
  - **Deviation from additive model**

Pairscan Network:



Modeling Epistasis

- Modeled as an alteration from the additive assumption

Specifically,

$$
\begin{array}{cc|cc}
x_1 & x_2 & & x_1 & x_2 \\
0 & 0 & \longrightarrow & 0 & 0 \\
0 & 1 & \longrightarrow & 0 & 1 \\
1 & 0 & \longrightarrow & 1 & 0 \\
1 & 1 & \longrightarrow & 1+\delta_{21} & 1+\delta_{12}
\end{array}
$$

Reparameterization

- We now have 2 ways of looking at epistasis
  - Undirected, β_{12} interaction coefficients
  - Values of delta

- Lets us reparameterize beta values in terms of delta:

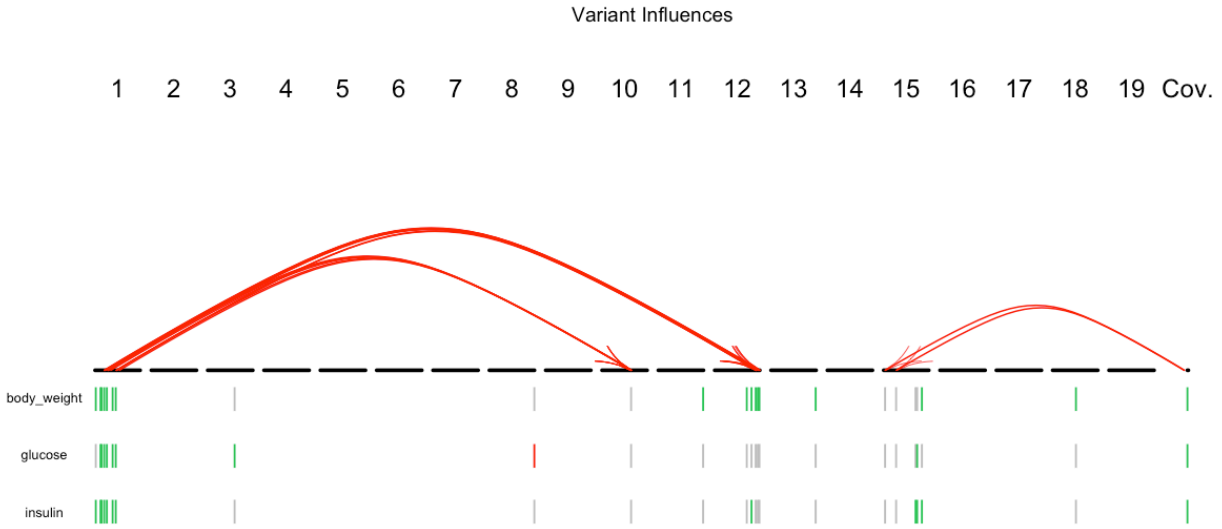
$$\begin{cases} \beta_{12}^1 = \beta_1^1 \delta_{21} + \beta_2^1 \delta_{12} \\ \beta_{12}^2 = \beta_1^2 \delta_{21} + \beta_2^2 \delta_{12} \end{cases}$$

- We need to have at least 2 phenotypes to constrain the model!

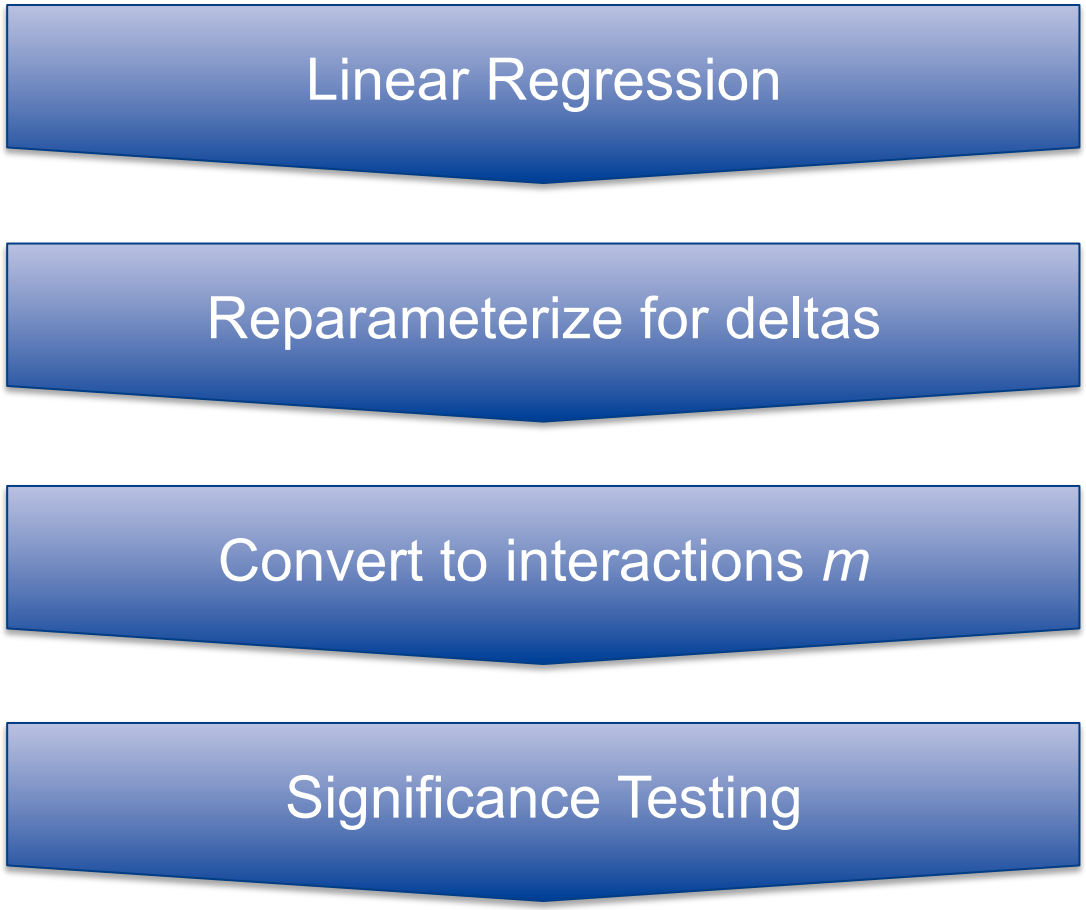
Creating the Network

- Delta values don't represent variant-to-variant influence!

$$\delta_{21} = m_{21}(1 + \delta_{12})$$
$$\delta_{12} = m_{12}(1 + \delta_{21})$$



Model Summary:



Extending the Model

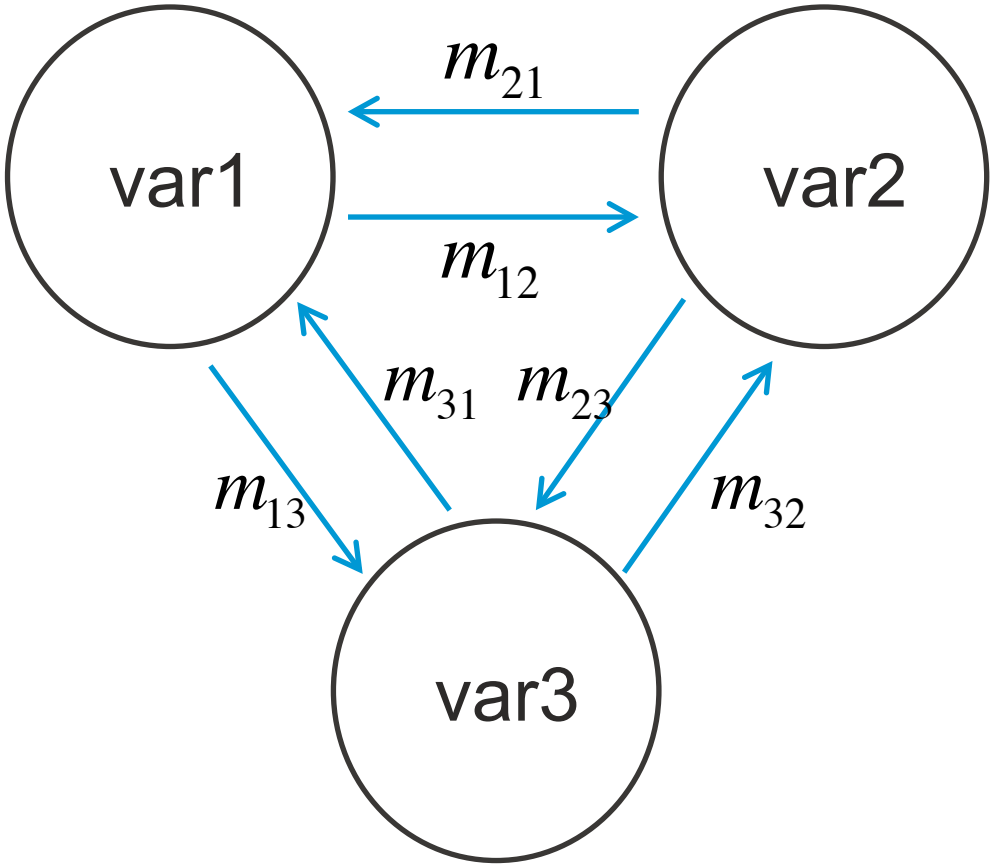
This will be shorter!

What's wrong?

- Currently, `cape` only considers two variants at a time without regard to other variants

- A more complete model includes multiple variants
  - find other genes that better explain interactions
  - detect "hidden" interactions

But how do you do that?



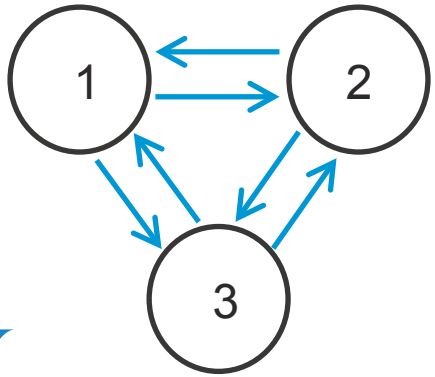
How about the delta approach?

- Model alterations to the additive assumption!

Genotype values

x_1	x_2	x_3		x_1	x_2	x_3
1	0	1	$\longrightarrow$	$1+\delta_{31}$	0	$1+\delta_{13}$

$$1 \quad 1 \quad 1 \longrightarrow \begin{matrix} 1+\delta_{21}+\delta_{31} & 1+\delta_{12}+\delta_{32} & 1+\delta_{13}+\delta_{23} \\ +\delta_{231}+\delta_{321} & +\delta_{132}+\delta_{312} & +\delta_{123}+\delta_{213} \end{matrix}$$



All “routes” of delta influence through **active** variants are included

Caveats

- This requires information about all combinations of variants under analysis!
  - That's 2^n
- Limits number of variants for simultaneous analysis
- Also power issues

Solving for Deltas

- Reparameterization gets harder
  - Example for 3 variants:

$$\begin{cases} \beta_{12} = \beta_1\delta_{21} + \beta_2\delta_{12} \\ \beta_{13} = \beta_1\delta_{31} + \beta_3\delta_{13} \\ \beta_{23} = \beta_2\delta_{32} + \beta_3\delta_{23} \\ \beta_{123} = \beta_1(\delta_{21} + \delta_{31} + \delta_{231} + \delta_{321}) + \\ \qquad\qquad \beta_2(\delta_{12} + \delta_{32} + \delta_{132} + \delta_{312}) + \\ \qquad\qquad \beta_3(\delta_{13} + \delta_{23} + \delta_{123} + \delta_{213}) \end{cases}$$

  - Loop over all phenotypes under analysis
  - Solve for deltas from betas using least squares optimization

Solving for Influences *m*

- Using delta information to infer values of *m*:

$$\begin{cases} \text{1,2 active} \begin{cases} A_1 = m_{21}A_2 + 1 \\ A_2 = m_{12}A_1 + 1 \end{cases} \\ \quad \vdots \\ \text{1,2,3 active} \begin{cases} A_1 = m_{21}A_2 + m_{31}A_3 + 1 \\ A_2 = m_{12}A_1 + m_{32}A_3 + 1 \\ A_3 = m_{13}A_1 + m_{23}A_2 + 1 \end{cases} \end{cases}$$

- Again, least squares optimization

Little Cross Analysis

Testing my model

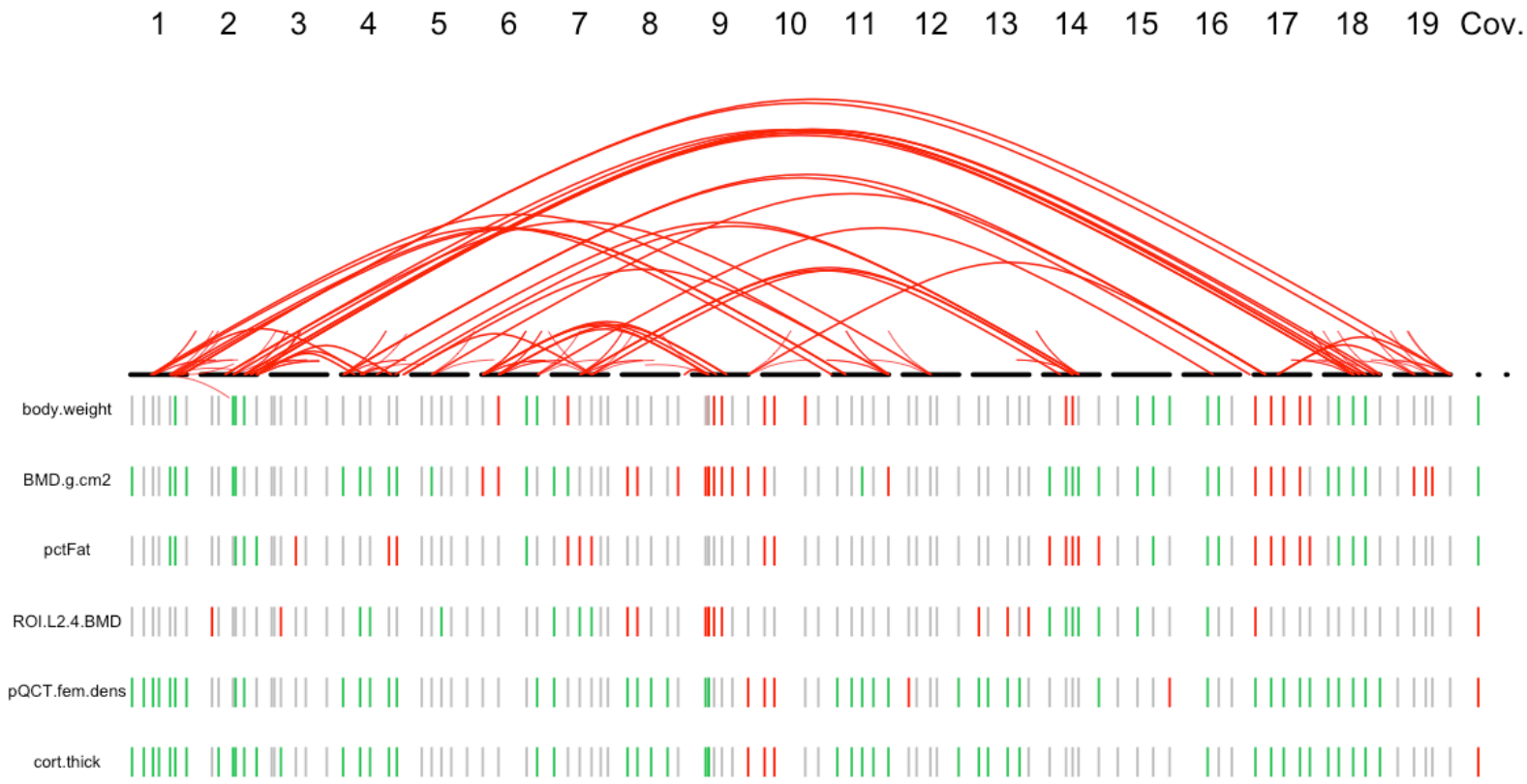
“Little cross” data

- 2133 mice
- F2 cross of B6 and C3H mice, both with *Ghrhr*^{lit} mutation
  - Growth hormone deficiency
  - low IGF1
- Marker data on 100 genotypes
- Physiological phenotypes targeted towards studying bone morphology/body composition

Strategy

1. Use cape to find markers of interest
2. Plot cape's network
3. Run my analysis on the same markers
4. Create a network
5. Compare

R/cape's output



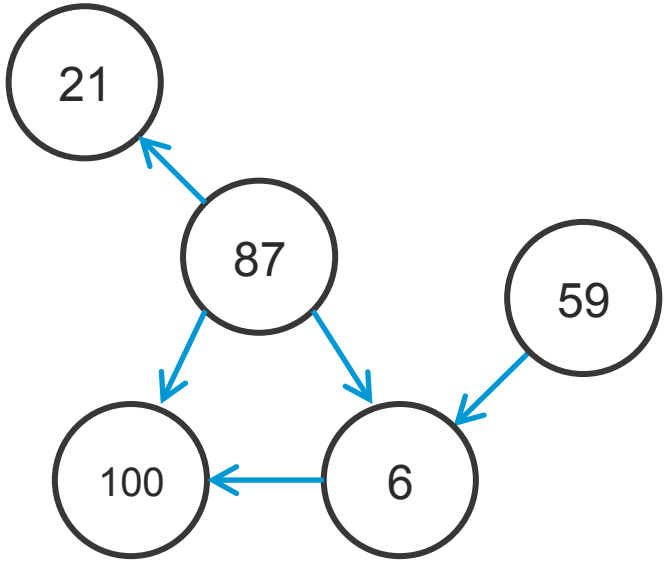
Teasing apart that hairball

- Network analysis finds connected components
  - `2, 3, 38, 39, 74, 76, 13, 22, 23, 24, 25, 58, 62`
  - `97, 99, 100, 5, 6, 7, 53, 86, 87, 59, 28, 21`
  - `65, 9, 10, 12, 93, 94, 95`
  - `32, 33, 50, 51, 52`
  - `34, 83`
  - `48, 46`

- Choose carefully!

Variants chosen for analysis

- I chose 6, 21, 59, 87, 100
  - Not necessarily best choice



Ran the analysis for *m*

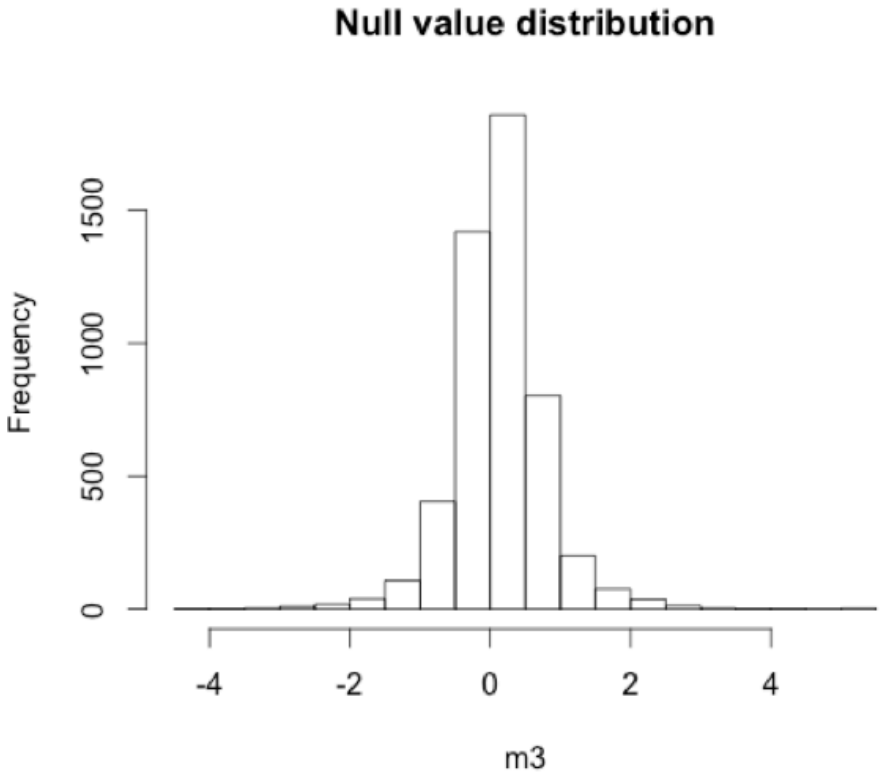
- Values inferred:

```
> solve.m.vals
Nonlinear regression via the Levenberg-Marquardt algorithm
parameter estimates: -0.0207505276495917, 0.641153346815994, -1.99878339291354, -0.488945002382586, 0.43075
0655463081, -0.438145950850714, 0.844890119224868, -0.10304610629287, 0.110499097774666, -1.87289630528937,
 0.190612721232149, -0.574979521386688, -0.252636912146147, -0.165109777602683, 0.268486466920171, -0.08828
70367685985, -0.0633981253006718, -0.633459302518879, 0.562567533439969, -1.05205410910421
residual sum-of-squares: 62.01
reason terminated: Relative error in the sum of squares is at most `ftol'.
```

```
> signif(m, 2)
      [,1]  [,2]  [,3]   [,4]   [,5]
[1,]  0.000  0.43  0.11 -0.250 -0.063
[2,] -0.021  0.00 -1.90 -0.170 -0.630
[3,]  0.640 -0.44  0.00  0.270  0.560
[4,] -2.000  0.84  0.19  0.000 -1.100
[5,] -0.490 -0.10 -0.57 -0.088  0.000
```

Significance Testing

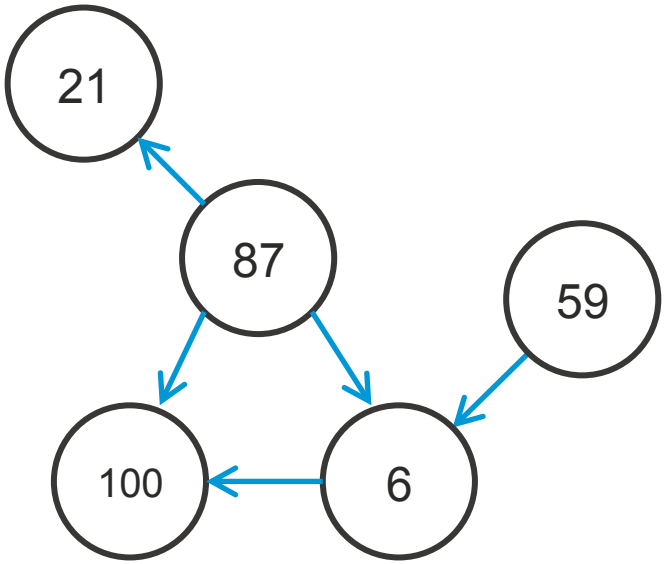
- 5000 permutations!
  - Break relationship between genotype and phenotype data
  - Generate null value distribution
  - Didn't have time to implement correction for multiple testing
- Empirically determine *p* values



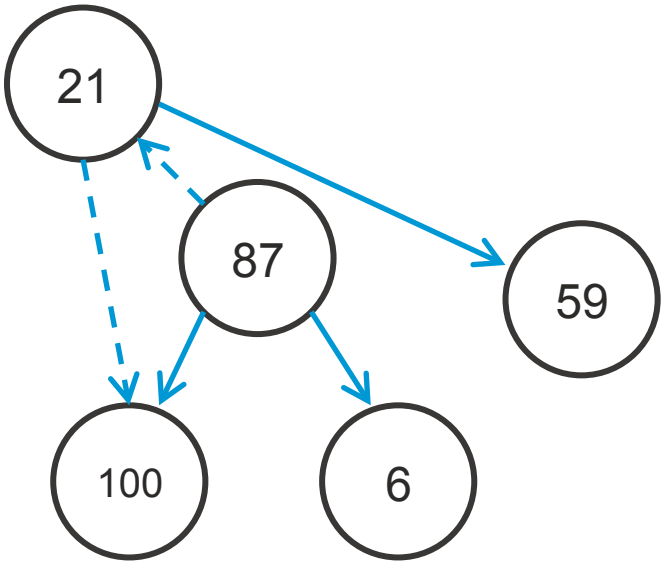
Network Output Comparison

- Only plotting significant values of *m*

Original cape network



Extended Model network



Conclusions

- Extension and implementation of existing model
- We verify some interactions found by `cape`'s analysis
- We find novel interactions for further analysis

Future work:

- Choose a better set of variants for analysis
- Correction for multiple testing
- Further testing for model sensitivity in real data

Acknowledgements

- **NSF Federal grant** supported by the Research Experience for Undergraduates Site at the Jackson Laboratory (DBI-1262049, Robert Braun, PI).
- The Jackson Laboratory
  - Greg Carter
  - Xulong Wang
  - Robyn Ball
- The Summer Student Program

Questions?

More math behind the model:
http://jiangts.github.io/epistasis-simulation/presentation/

Implementation:
https://github.com/jiangts/epistasis-simulation

The Jackson Laboratory

Leading the search for tomorrow's cures

Appendix

- For more (math) online: http://jiangts.github.io/epistasis-simulation/presentation/
- Implementation: https://github.com/jiangts/epistasis-simulation