Treelet Covariance Smoothers

Estimation of Genetic Parameters

Benjamin Draves[1]

[1]Department of Mathematics
Lafayette College

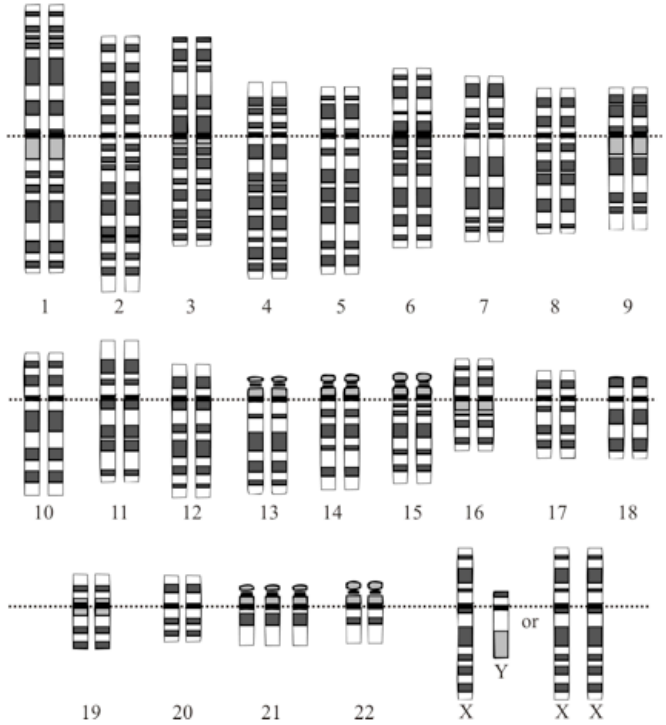
Advisor: T. Gaugler

Lafayette College, 2017

Overview

Molecular Biology Review

- Each person's genetic composition coded on *chromosomes*
- Most humans have 46 in total, all occurring in pairs
- The 23rd pair determines sex
- We can compare the genetic data coded by the first 22 pairs for all humans
- Find patterns between this genetic data and realized traits & diseases

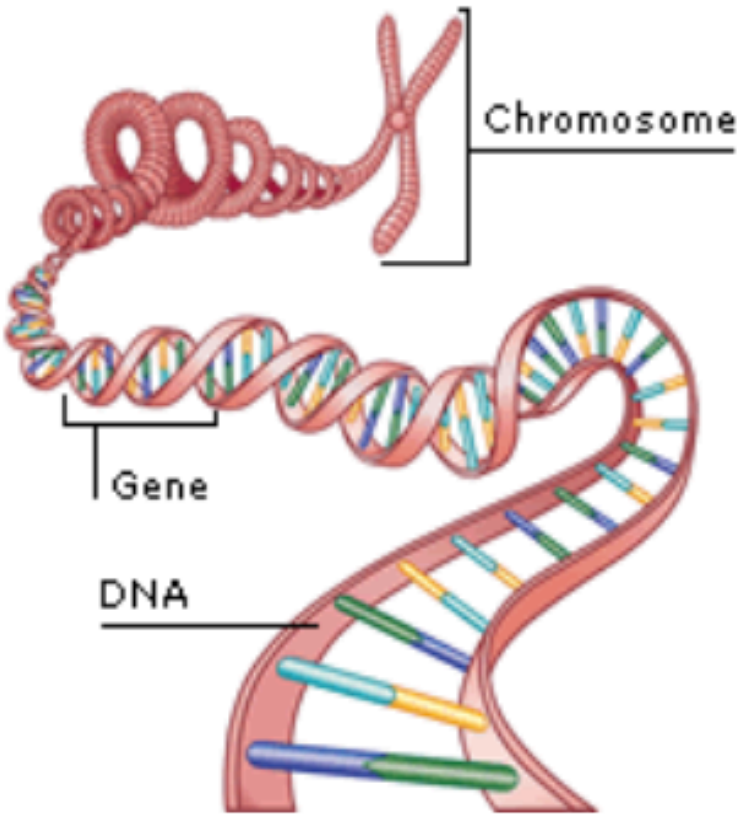


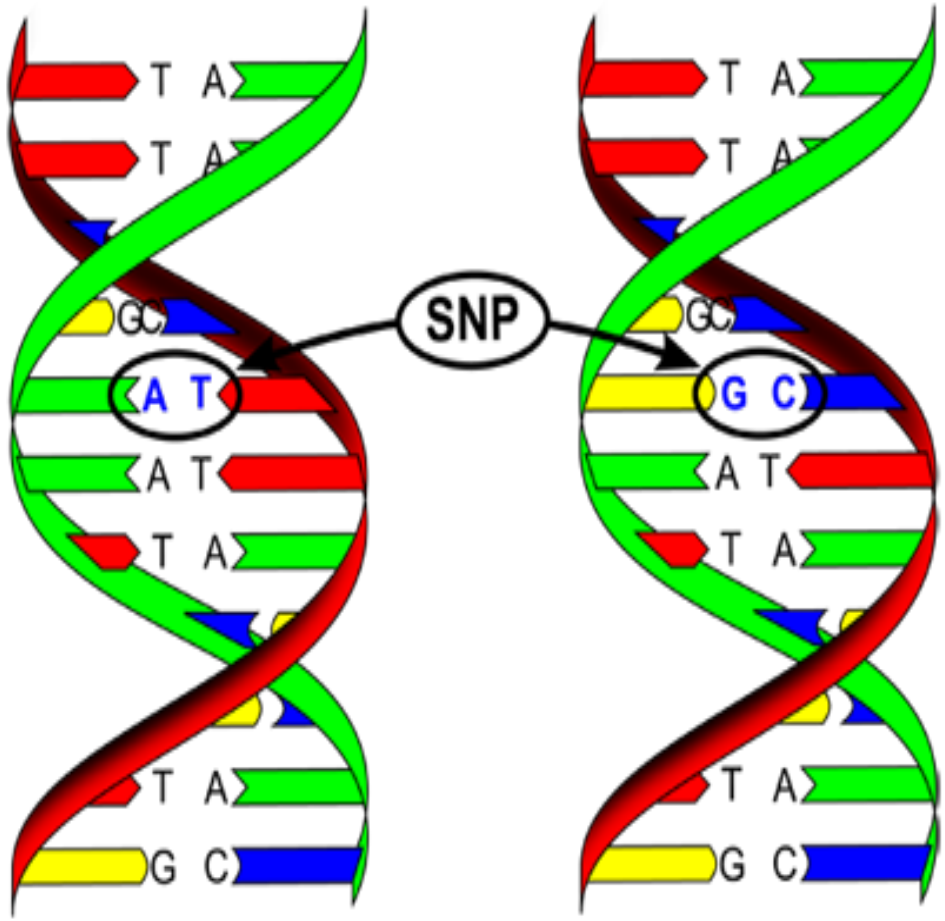
Traditional Genetic Studies

- We wish to estimate the *penetrance function*, $P(Y|\mathbf{G})$
  - Y is some phenotype of interest
  - $\mathbf{G}$ codes the underlying genotype
- Kinda hard to do without $\mathbf{G}$...
- Linkage Analysis studies have had considerable success understanding $\mathbf{G}$ indirectly by analyzing Y through numerous generations
- *Hard* to do with human genetics
- Next Generation Sequencing (NGS) technology allows us to sample from $\mathbf{G}$ directly

Single Nucleotide Polymorphisms (SNPs)

- So how do we encode this genetic information?
- Code the chromosome pairs
- Exploit the complimentary fashion of DNA





SNPs (cont.)

SNPs

(A,T)	(A,T)
(G,C)	(A,T)
⋮	⋮
(G,C)	(A,T)
(G,C)	(G,C)

$\Longrightarrow$

Recode

α	α
β	α
⋮	⋮
β	α
β	β

$\Longrightarrow$

Count Minor Alleles

2
1
⋮
1
0

- Each row in this diagram represents a *SNP*
- The pair, either (A, T) or (G, C), is called a *polymorphism* or an *allele*
- An allele is called a *minor allele* if appears less frequently in the population

Minor Allele Counts as Random Variables

- For each locus, k, we code can code individual i's minor allele count (MAC) by $c_k^{(i)} \in \{0, 1, 2\}$
- For m loci, we can describe the full genotype by

Minor Allele Count (MAC)

$$c_*^{(i)} = \{c_1^{(i)}, c_2^{(i)}, \ldots, c_m^{(i)}\} \in \{0, 1, 2\}^m$$

- If we assume random recombination of alleles, $c_k^{(i)} \sim \text{Binom}(2, p_k)$
  - Where p_k is the *minor allele frequency*
- This is a pretty strong assumption, but using this framework allows for simple model construction

Scaled Minor Allele Counts

- Under the assumption that alleles are independent, we can center our count vector
- Let $z_k^{(i)} := (c_k^{(i)} - 2p_k)/(2p_k(1-p_k))^{1/2}$ be the scaled minor allele count at locus k
- Then for each SNP, k, we define the scaled minor allele count by

Scaled Minor Allele Count (SMAC)

$$\mathbf{z}_k^* = (z_k^{(1)}, z_k^{(2)}, \ldots, z_k^{(n)})^t$$

- Where n is the number of individuals in the sample
- Then for a sample of m genetic markers, we organize this data as $\mathbf{Z} = (\mathbf{z}_1^*, \mathbf{z}_2^*, \ldots, \mathbf{z}_m^*) \in \mathbb{R}^{n \times m}$

Did everyone get that?

$$\mathbf{Z} = \begin{array}{c} \\ \mathbf{z}_*^{(1)} \\ \mathbf{z}_*^{(2)} \\ \vdots \\ \mathbf{z}_*^{(n)} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{z}_1^* & \mathbf{z}_2^* & \cdots & \mathbf{z}_m^* \end{array} \\ \begin{pmatrix} z_1^{(1)} & z_2^{(1)} & \cdots & z_m^{(1)} \\ z_1^{(2)} & z_2^{(2)} & \cdots & z_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^{(n)} & z_2^{(n)} & \cdots & z_m^{(n)} \end{pmatrix} \end{array}$$

Individual n

SNP 2

Genetic Parameters of Interest

- *Additive Genetic Relatedness* (A)
  - Denoted A_{ij} for relatedness between individuals i and j
  - Additive covariance between genetic markers
  - I'll refer to this as Relatedness
- *Narrow Sense Heritability* (h^2)
  - Incorporates a small contribution for the m genetic markers, independently
  - Doesn't try to understand the joint distribution of the alleles
  - Traditional studies implicitly use this joint distribution to infer *broad sense* heritability
  - I'll refer to this as Heritability

Estimating Relatedness

- We consider alleles *Identical By Descent* (IBD)
- Relatedness is the expected proportion of alleles IBD between individuals
- Under this interpretation of A, at SNP k, $A_{ij} = \mathrm{Cov}(z_k^{(i)}, z_k^{(j)})$
- Using this information, we can estimate A by

Method of Moments Estimate of A

$$\widehat{A} = \frac{1}{m}\sum_{k=1}^{m} \mathbf{z}_k^* (\mathbf{z}_k^*)^t = \frac{\mathbf{Z}\mathbf{Z}^t}{m}$$

- As m increases, we expect $\dfrac{\mathbf{Z}\mathbf{Z}^t}{m} \to A$

Estimating Heritability

Phenotype Model (1)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon} \qquad \text{with} \qquad \mathrm{Var}(\mathbf{y}) = \mathbf{Z}\mathbf{Z}^t\sigma_u^2 + \mathbf{I}\sigma_\epsilon^2$$

- $\mathbf{y}$ vector of phenotypes, $\mathbf{X}\beta$ fixed effects, $\mathbf{u}$ vector of random effects of the causal SNPs with $\mathrm{Var}(\mathbf{u}) = \mathbf{I}\sigma_u^2$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}\sigma_\epsilon^2)$ residual errors
- But remember, we want to understand the ratio of genetic variance to total variance
- Let $\mathbf{u} = (u_1, u_2, \ldots, u_J)^t \in \mathbb{R}^J$ be the vector of effects corresponding to the J casual SNPs
- Let $\sigma_g^2 = J\sigma_u^2$ be the variance explained by all the SNPs
- We can then write the genetic effect of individual i as $g_i = \sum_{j=1}^{J} z_j^{(i)} u_j$

  where $\mathrm{Var}(\mathbf{g}) = A\sigma_g^2$

Estimating Heritability (cont.)

Phenotype Model (2)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{g} + \boldsymbol{\epsilon} \qquad \text{with} \qquad \widehat{\mathrm{Var}}(\mathbf{y}) = A\sigma_g^2 + \mathbf{I}\sigma_\epsilon^2$$

- We can partition the variability of phenotypic expression into *genetic* (σ_g^2) and *environmental* (σ_ϵ^2) factors
- From here we define narrow sense heritability as

Narrow Sense Heritability

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\epsilon^2}$$

- We can estimate this value via restricted maximum likelihood (REML) algorithms

Possible Problems

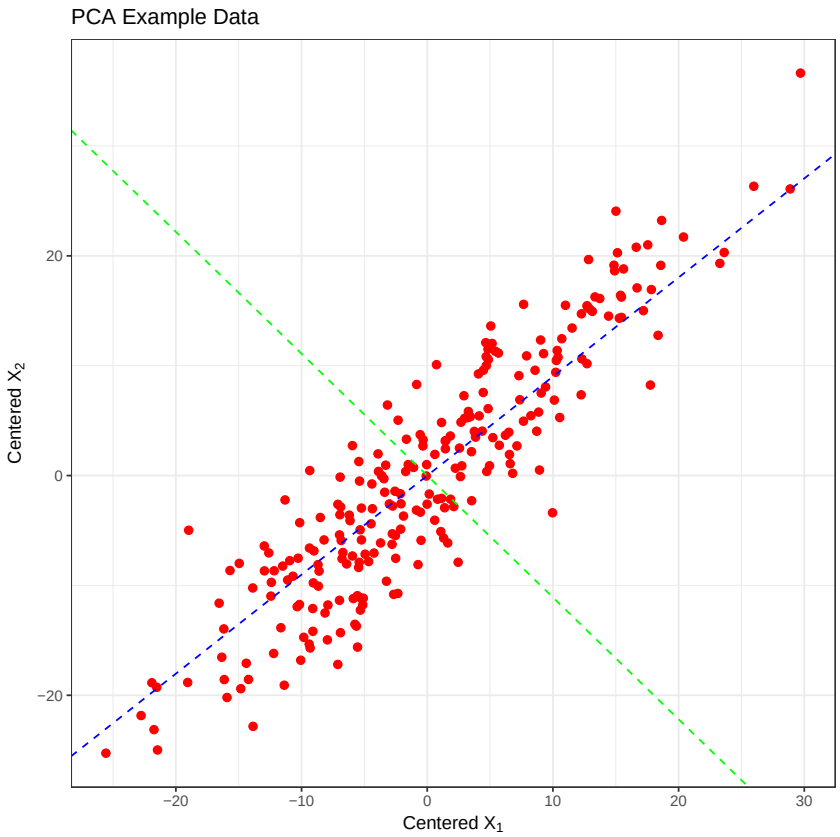
- Assume we have three random individuals, who happen to be named Ben, Josh, and Trent
- Trent and Ben, coming from small Midwest towns, are 7th degree relatives
- Josh, from the west coast, is unrelated to Trent and Ben

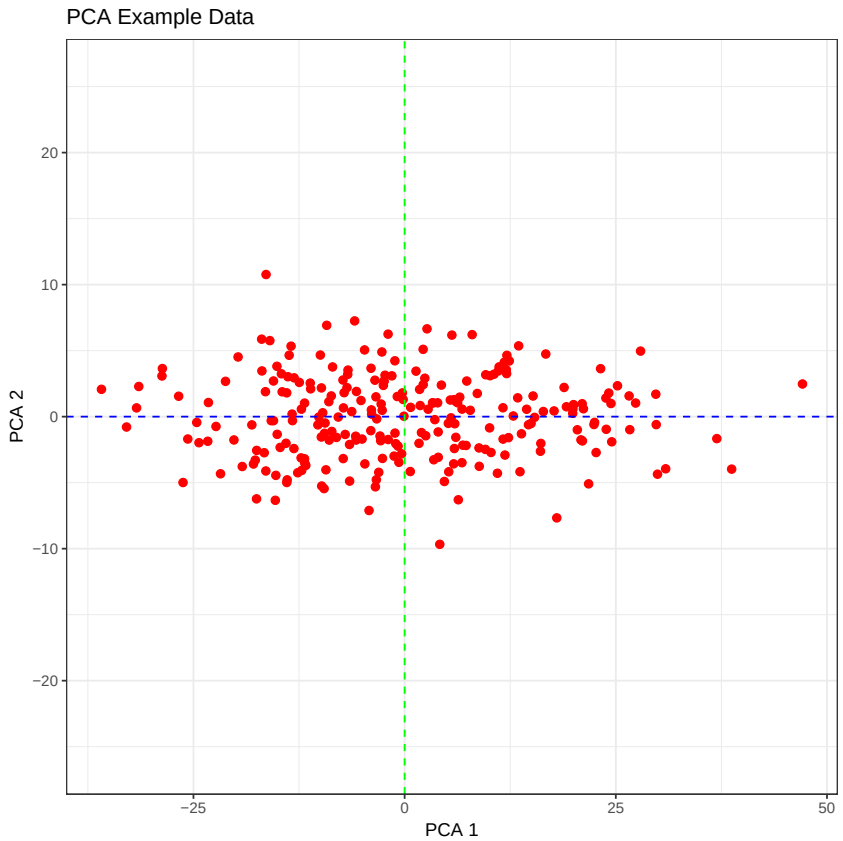
Ben :	0	2	1	·····	2
Trent :	1	2	0	·····	0
Josh :	1	2	1	·····	1

- $\widehat{A}_{(\text{Ben, Trent})} = \frac{1}{130}$, $\widehat{A}_{(\text{Ben, Josh})} = \frac{1}{130}$
- How do we differentiate between distantly and unrelated individuals?

Preliminaries: Principal Component Analysis (PCA)

- Goal: Rotate underlying space so variability lies on few vectors
- We can rotate the space via a Jacobian matrix corresponding to the principal components
- Also used as a dimensionality reduction tool

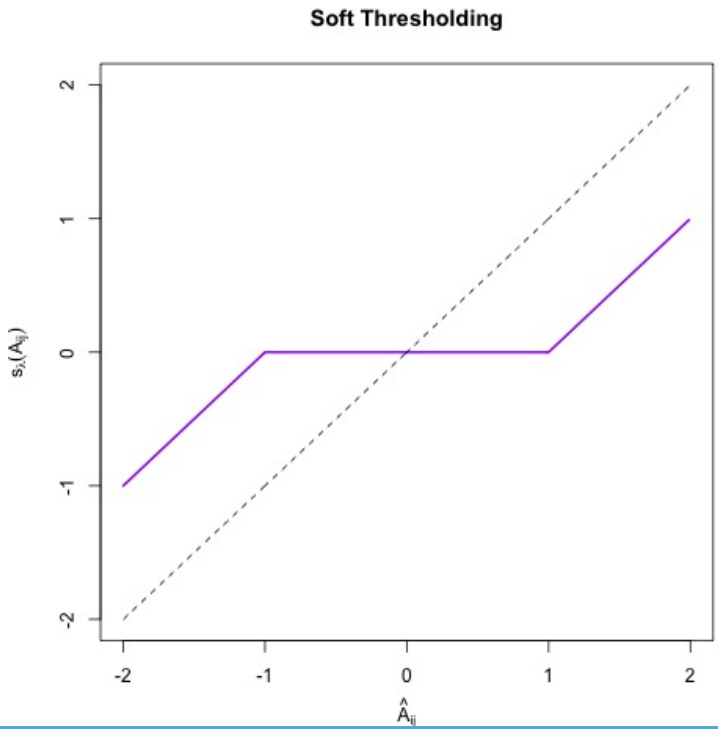


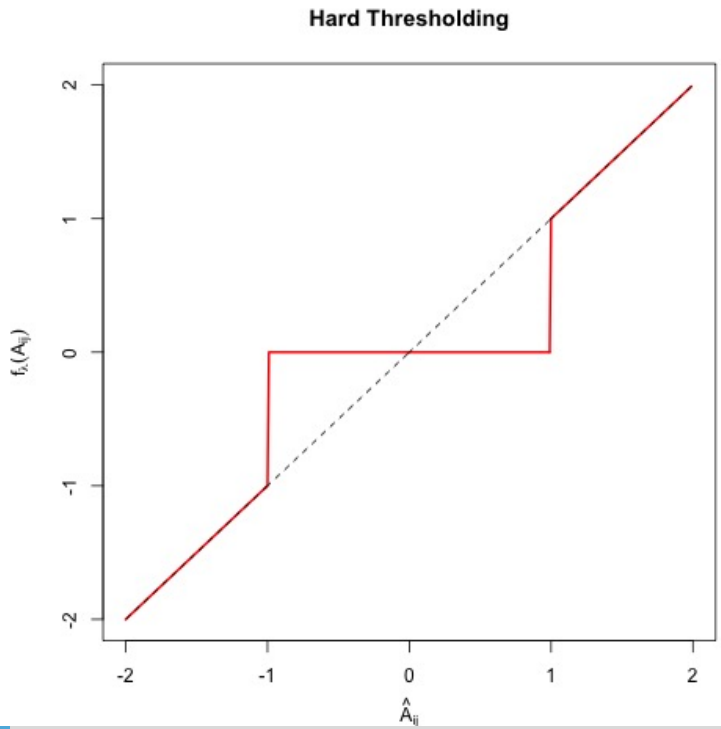


Preliminaries: Wavelet Thresholding

Soft and Hard Thresholding

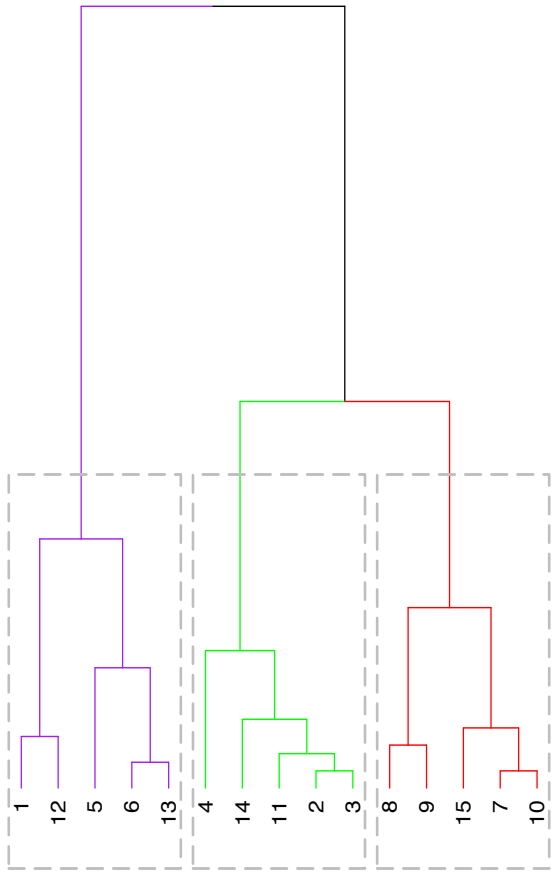
$$s_\lambda(\widehat{A}_{ij}) = \begin{cases} \widehat{A}_{ij} + \lambda & \text{if } \widehat{A}_{ij} < \lambda \\ 0 & \text{if } -\lambda \leq \widehat{A}_{ij} \leq \lambda, \\ \widehat{A}_{ij} - \lambda & \text{if } \widehat{A}_{ij} > \lambda \end{cases} \qquad f_\lambda(\widehat{A}_{ij}) = \begin{cases} \widehat{A}_{ij} & \text{if } |\widehat{A}_{ij}| \geq \lambda \\ 0 & \text{if } |\widehat{A}_{ij}| < \lambda \end{cases}$$





Treelet Algorithm: The Idea

- Focus on estimating close relatives well
- Preserve local familiar structures
- Try to extend that structure to distant relatives



Treelet Algorithm

1. Let $\mathbf{z}^*$ be a random vector representing the SMAC at any SNP with covariance $\Sigma = A$, which is the additive genetic relationship matrix, corresponding to $\ell = 0$
2. Let $\mathbf{V}^0$ be the basis corresponding to this vector
3. Compute the variance-covariance matrix $\widehat{\Sigma}^{(0)}$ with corresponding similarity matrix $\widehat{M}^{(0)}$ defined by

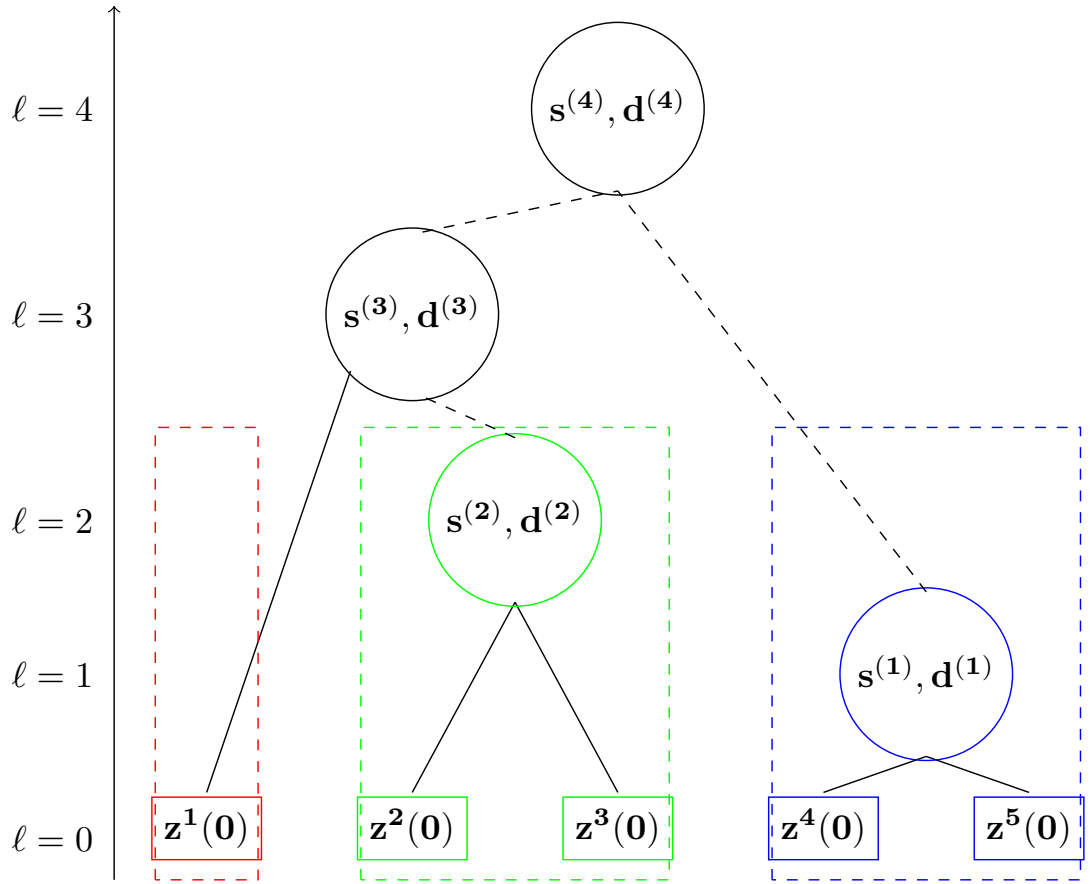
$$\widehat{M_{ij}}^{(0)} = \frac{\widehat{\Sigma_{ij}}^{(0)}}{\sqrt{\widehat{\Sigma_{ii}}^{(0)}\widehat{\Sigma_{jj}}^{(0)}}}$$

4. Initialize the *sum variable* indices to $S_0 = \{1, 2, \ldots, N\}$

Treelet Algorithm (cont.)

4. For $\ell = 1, 2, \ldots, L$ for $L \leq N - 1$
   1. Find the two most closely related individuals according $\widehat{M}^{(\ell-1)}$. Let
   $$(\alpha_\ell, \beta_\ell) = \underset{i,j \in S_{\ell-1}}{\arg\max}\ \widehat{M}_{ij}^{(\ell-1)}$$
   2. Rotate the genetic space to decorrelate z_{α_ℓ} and z_{β_ℓ}
   3. Rotate $\widehat{\Sigma}^{(\ell-1)}$ and update $\widehat{M}^{(\ell-1)}$
   4. Assuming α_ℓ and β_ℓ represent the first and and second principal component, respectively
   5. Update the sum set $S_\ell = S_{\ell-1} \setminus \{\beta_\ell\}$

Treelet Algorithm Visualized



$$\left(\mathbf{s^{(4)}, d^{(3)}, d^{(2)}, d^{(4)}, d^{(1)}}\right)^t$$

$$\left(\mathbf{s^{(3)}, d^{(3)}, d^{(2)}, s^{(1)}, d^{(1)}}\right)^t$$

$$\left(\mathbf{v_1^{(0)}, s^{(2)}, d^{(2)}, s^{(1)}, d^{(1)}}\right)^t$$

$$\left(\mathbf{v_1^{(0)}, v_2^{(0)}, v_3^{(0)}, s^{(1)}, d^{(1)}}\right)^t$$

$$\left(\mathbf{v_1^{(0)}, v_2^{(0)}, v_3^{(0)}, v_4^{(0)}, v_5^{(0)}}\right)^t$$

Treelet Decomposition

- At each level ℓ we have an orthonormal basis $\mathbf{V}^\ell = \left[\mathbf{v}_1^{(\ell)} \ \ldots \ \mathbf{v}_N^{(\ell)}\right]$
- Using this basis, write $\mathbf{z}^*(0) = \sum_{i=1}^{N} \alpha_i^{(\ell)} \mathbf{v}_i^{(\ell)}$ where $\alpha_i^{(\ell)} = \langle \mathbf{z}^*(0), \mathbf{v}_i^{(\ell)} \rangle$ represent the projections onto that basis vector at level ℓ
- This gives rise to the decomposition of the variance of $\mathbf{z}^*(0)$

Treelet Decomposition

$$\Sigma = \mathrm{Var}[\mathbf{z}^*(0)] = \sum_{i=1}^{N} \gamma_{i,i}^{(\ell)} \mathbf{v}_i^{(\ell)} \left(\mathbf{v}_i^{(\ell)}\right)^t + \sum_{i \neq j}^{N} \gamma_{i,j}^{(\ell)} \mathbf{v}_i^{(\ell)} \left(\mathbf{v}_j^{(\ell)}\right)^t = \mathbf{V}^\ell \mathbf{\Gamma}^\ell \left(\mathbf{V}^\ell\right)^t$$

- Where $\gamma_{i,j}^{(\ell)} = \mathrm{Cov}[\alpha_i^{(\ell)}, \alpha_j^{(\ell)}]$ and $\mathbf{\Gamma}^\ell = \left[\gamma_{i,j}^{(\ell)}\right]$

Formalization of Problem

- For large samples, we expect A to be quite sparse
- We want to enforce this sparsity on our estimates of A
- We can do this directly, but run into the Trent, Ben, and Josh problem
- Idea: Use a Treelet representation of A and enforce sparsity of the projected covariances, $\mathbf{\Gamma}^{\ell}$, via wavelet hard thresholding

Treelet Covariance Smoothing (TCS)

- Crosset et al. (2013) first employed this method and called it *Treelet Covariance Smoothing* (TCS)
- Gaugler et al. (2014) used this method to show that the majority of risk of Autism resides in common variants
- It also partially got Trent a job at Lafayette

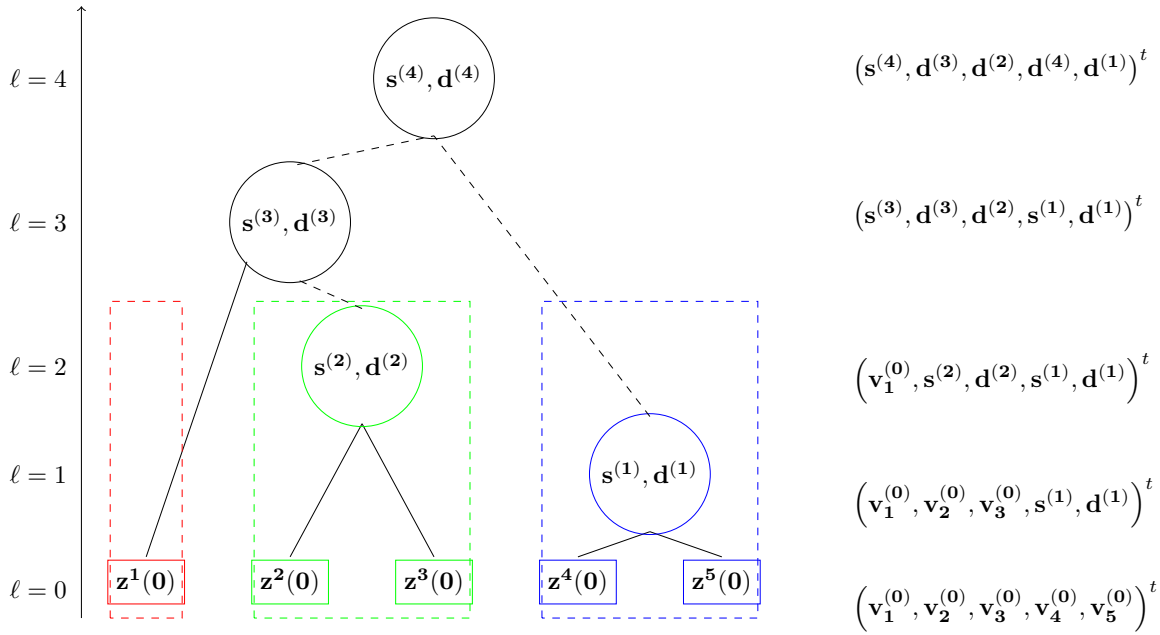
TCS Estimator

$$\widetilde{A}(\lambda) = \sum_{i=1}^{N} f_\lambda[\widehat{\gamma}_{i,i}]\widehat{\mathbf{v}}_i(\widehat{\mathbf{v}}_i)^t + \sum_{i \neq j}^{N} f_\lambda[\widehat{\gamma}_{i,j}]\widehat{\mathbf{v}}_i(\widehat{\mathbf{v}}_j)^t = \widehat{\mathbf{V}} f_\lambda\left[\widehat{\mathbf{\Gamma}}\right]\widehat{\mathbf{V}}^t$$

- Where f_λ is a hard-thresholding function with optimal smoothing parameter λ
- TCS utilizes the top level of the tree ($\ell = N - 1$)

Possible Improvements/Heuristic Strategies

- We anticipate clusters of closely related individuals in our samples
- Varying ℓ, we attain a more representative basis set for the underlying genetic space



- We can induce additional smoothing by projecting only onto the first principal component at each level

Treelet Covariance Blocking (TCB)

- We employ this idea in our proposed method *Treelet Covariance Blocking* (TCB)
- To utilize the first principal component only and write

$$\tilde{\mathbf{z}}^*(\ell) = \sum_{i \in S_l} \alpha_i^{(\ell)} \mathbf{v}_i^{(\ell)}$$

- Using this projection, we estimate $\mathrm{Var}\left(\tilde{\mathbf{z}}^*(\ell)\right)$ by

TCB Estimator

$$\widetilde{A}(\ell) = \sum_{i \in \widehat{S}_\ell} \widehat{\gamma}_{i,i}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_i^{(\ell)}\right)^t + \sum_{\substack{i,j \in \widehat{S}_\ell \\ i \neq j}} \widehat{\gamma}_{i,j}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_j^{(\ell)}\right)^t = \widehat{\mathbf{V}}^\ell \widehat{\mathbf{\Gamma}}^\ell \left(\widehat{\mathbf{V}}^\ell\right)^t$$

- Where ℓ is the optimal level of the tree
- We implicitly enforce sparsity by only projecting the data onto basis vectors that are supported by familial blockings in the data

Treelet Covariance Blocked Smoothing (TCBS)

- To further eliminate erroneous inter-familial relatedness, it may be advantageous to utilize a hard thresholding function
- We call this method *Treelet Covariance Blocked Smoothing* (TCBS)
- Using the same projection onto the first principal components we have

TCBS Estimator

$$\widetilde{A}(\theta) = \sum_{i \in \widehat{S}_l} f_\lambda[\gamma_{i,i}^{(\ell)}] \mathbf{v}_i^{(\ell)} \left(\mathbf{v}_i^{(\ell)}\right)^t + \sum_{\substack{i,j \in \widehat{S}_l \\ i \neq j}} f_\lambda[\gamma_{i,j}^{(\ell)}] \mathbf{v}_i^{(\ell)} \left(\mathbf{v}_j^{(\ell)}\right)^t = \widehat{\mathbf{V}}^\ell f_\lambda\left[\widehat{\mathbf{\Gamma}}^{\boldsymbol{\ell}}\right] \left(\widehat{\mathbf{V}}^\ell\right)^t$$

- Where $\theta = (\ell, \lambda)$ is the optimal level, smoothing parameter combination

Optimal Parameter Selection

- All of our methods rely on choosing optimal smoothing parameters

$$\tilde{A}(\ell) = \sum_{i \in \widehat{S}_\ell} \widehat{\gamma}_{i,i}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_i^{(\ell)}\right)^t + \sum_{i,j \in \widehat{S}_\ell, i \neq j} \widehat{\gamma}_{i,j}^{(\ell)} \widehat{\mathbf{v}}_i^{(\ell)} \left(\widehat{\mathbf{v}}_j^{(\ell)}\right)^t. \tag{15}$$

As with any smoothing technique, one will naturally ask how the optimal ℓ should be chosen. Who the hell knows? Maybe a cost function like TCS, or maybe based on heritability profiles, which will be covered in the next section.

- We considered clustering techniques, cross validation, and likelihood based methods to search over the parameter space Θ

Cross Validation

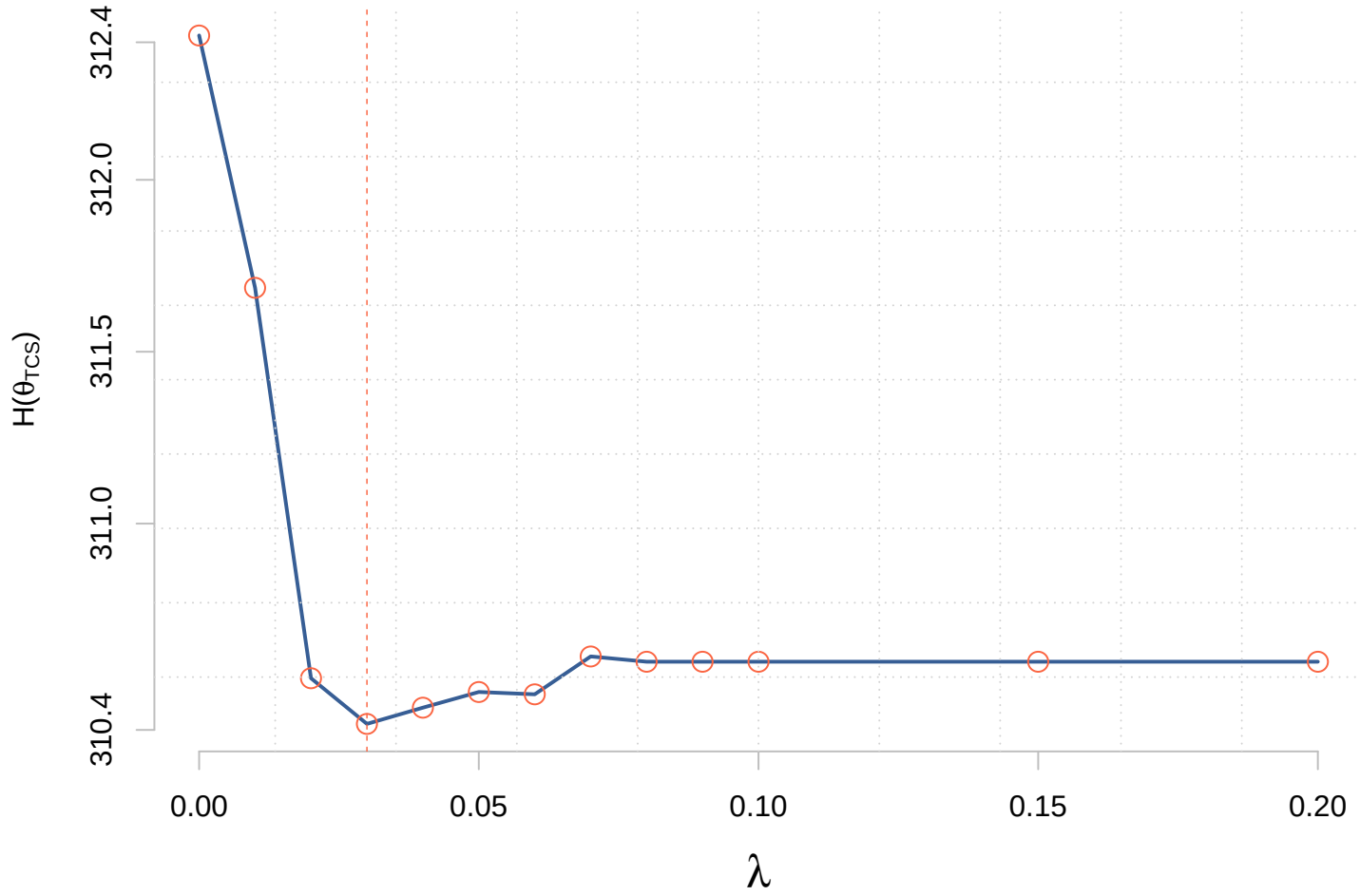
- Partition chromosomes into two sets: $\mathcal{A}$ and $\mathcal{B}$
- Find a robust, no smoothing estimate, $\widehat{A}$ using the SNPs from $\mathcal{A}$
- Train our algorithms on $\widehat{A}$ to attain $\widetilde{A}(\theta)$ for each $\theta \in \Theta$
- Partition $\mathcal{B}$ into K groups
- For each $k = 1, 2, \ldots, K$, attain $\widehat{A}_k$ and compare to smoothing estimates $\widetilde{A}(\theta)$ via

Cost Function

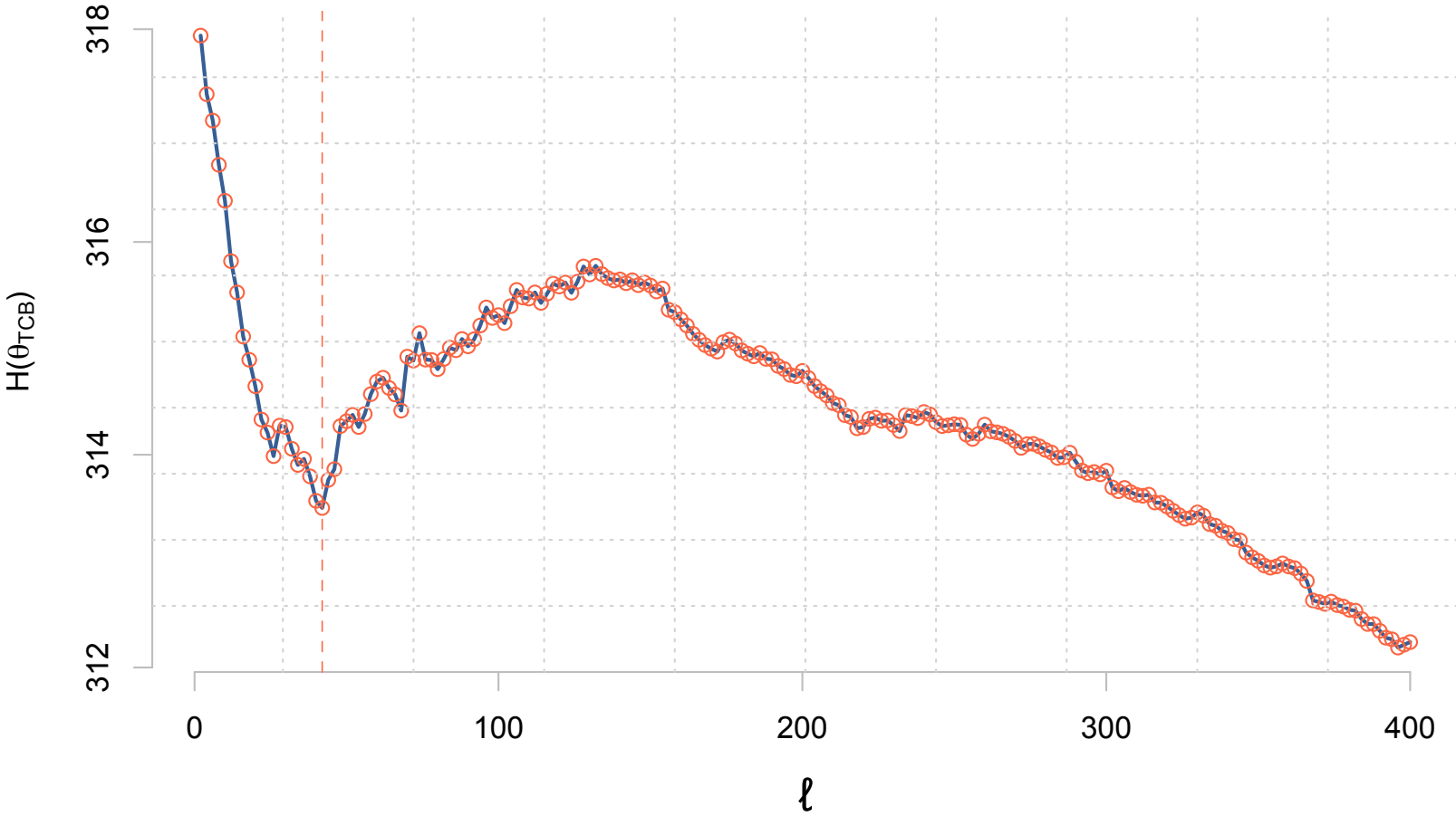
$$H(\theta) = \frac{1}{(N-1)NK} \sum_{k=1}^{K} \sum_{i<j}^{N} w_{ij} (\widehat{A}_{ij,k} - \widetilde{A}_{ij}(\theta))^2$$

- $w_{ij} = |\mathbf{\Gamma}_{ij}^{(\ell)}|$ corresponds to the $\mathbf{\Gamma}$ matrix at level ℓ determined by θ
- The optimal parameter is given by $\widehat{\theta} = \arg\min_{\theta \in \Theta} H(\theta)$

Pretty pictures for those who are lost or bored



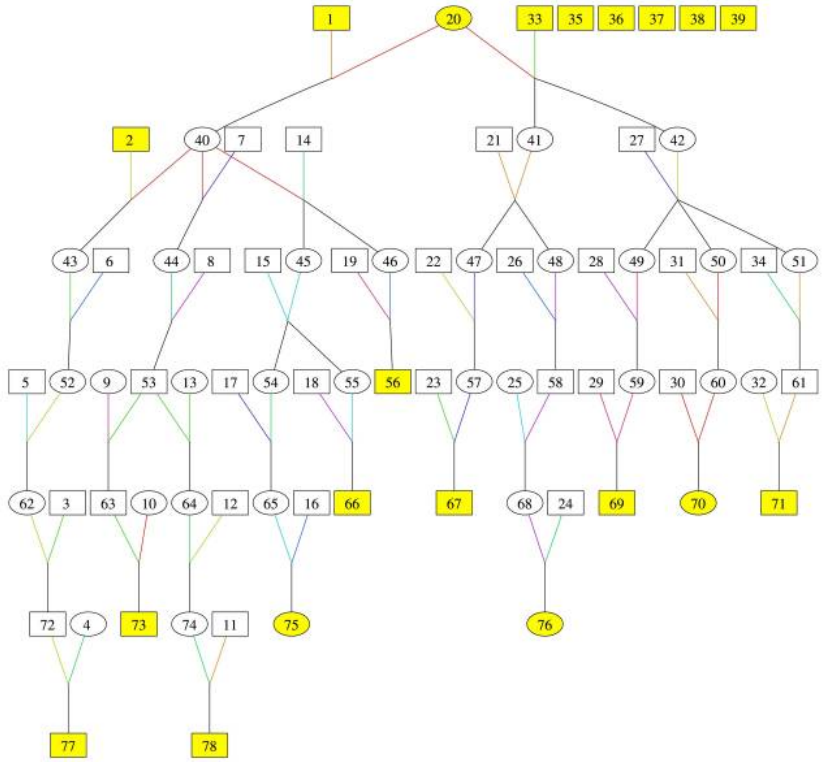
Pretty pictures for those who are lost or bored



Simulation Data - HapMap3 Data

- We utilize a pedigree structure used in other simulation studies
- Seven generation family - only consider 20 individuals
- Most closely related was degree three ($\frac{1}{8}$ genetic information)
- Most distantly related was degree eleven ($\frac{1}{2048}$ genetic information)
- Still unrelated individuals in this sample

Liebners



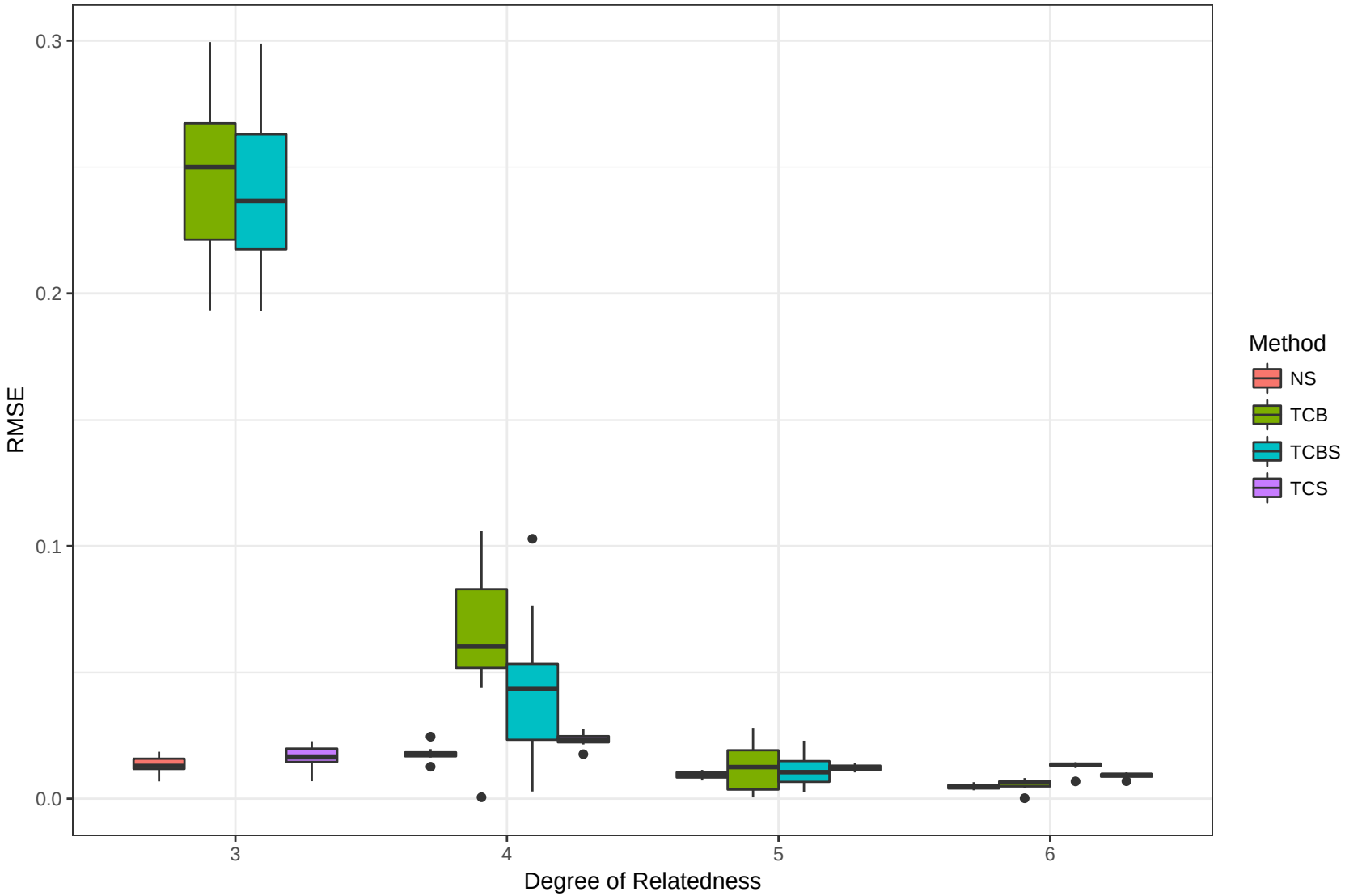
Simulation Design - Relatedness

1. Create a sample of 500 individuals by iteratively sampling 10 person blocks from the Liebner pedigree
2. Record the relatedness of individuals within the blocks
3. Set relatedness for individuals *not* in the same block to 0

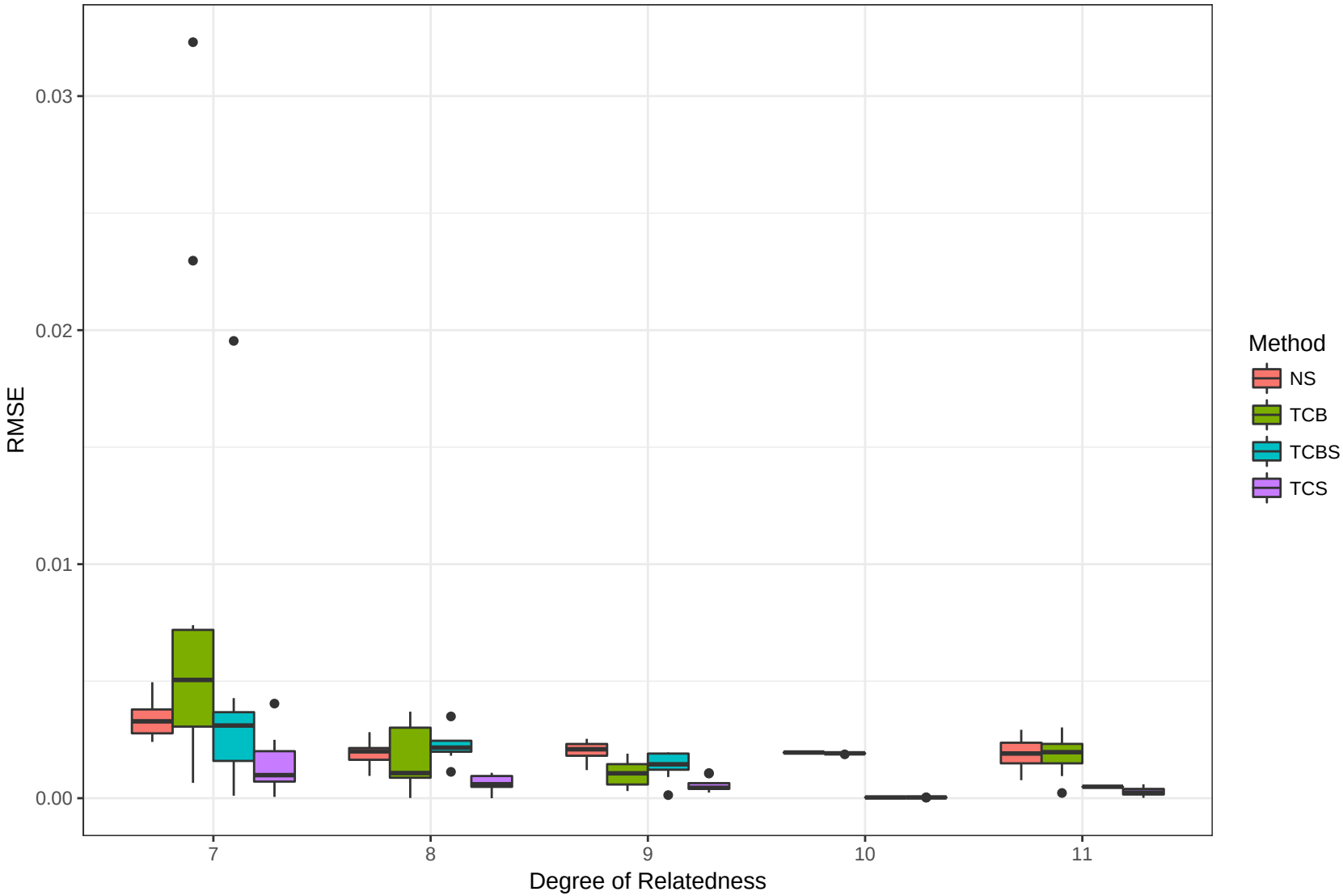
$$A_1 = \begin{bmatrix} \mathbf{A}^1 & 0 & \dots & 0 \\ 0 & \mathbf{A}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{A}^{50} \end{bmatrix}$$

4. Use the genetic information for this pedigree to attain $\widetilde{A}_1(\theta)$
5. Compare these estimates to the true A_1
6. Repeat this process ten times (e.g. $A_1, A_2, \dots, A_{10}$)

Relatedness Results



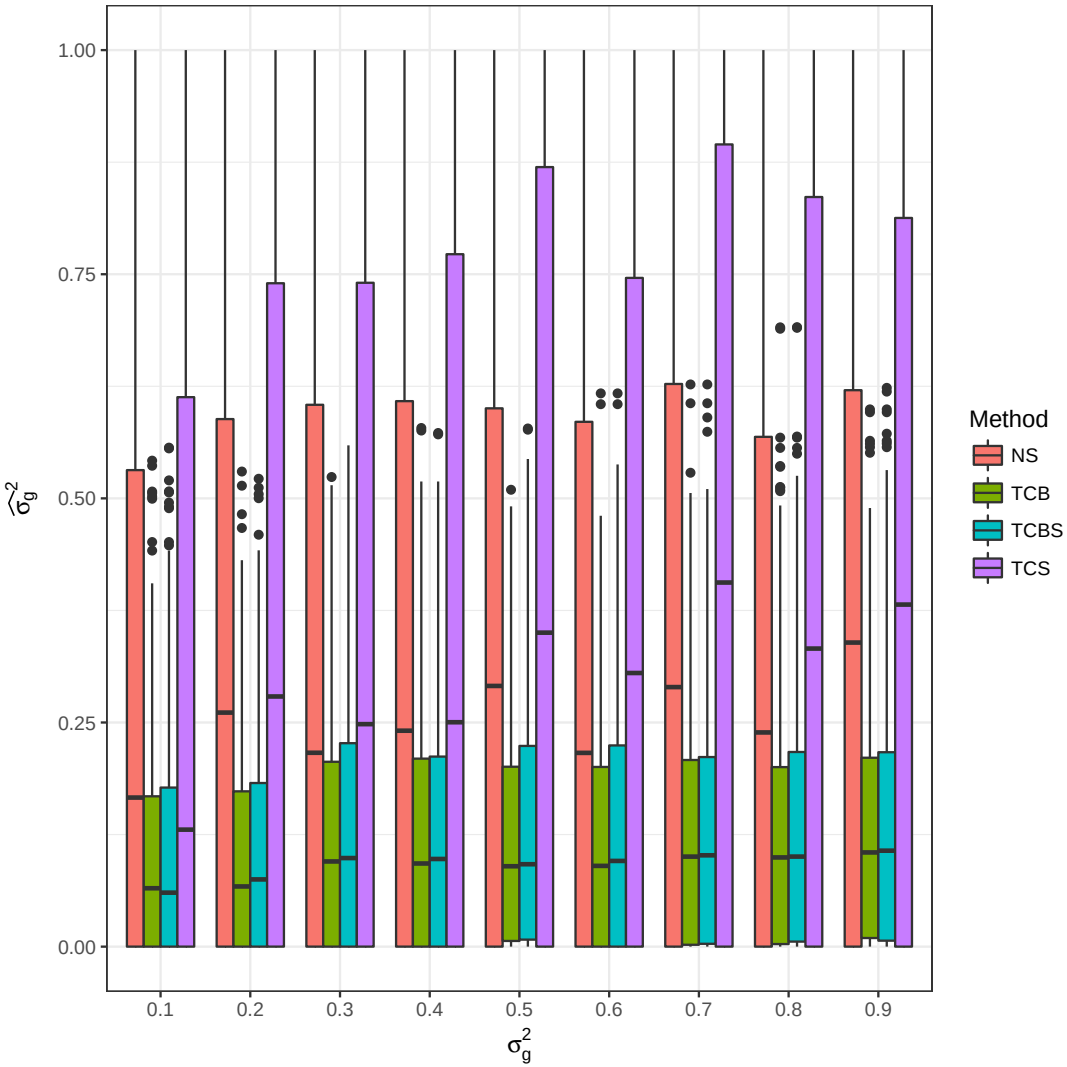
Relatedness Results (cont.)



Simulation Design - Heritability

1. Use Phenotype Model (2), $\mathbf{y} = \boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\epsilon}$, to generate ten phenotype vectors with heritability σ_g^2
2. Do this for $\sigma_g^2 \in \{.1, .2, \ldots, .9\}$
3. Do this for all ten population structures represented by $A_1, A_2, \ldots, A_{10}$
4. In aggregate, each population will have 10 phenotype vectors for each σ_g^2 considered
5. Use $\widetilde{A}(\theta)$ in the REML algorithm to estimate heritability, $\widehat{\sigma}_g^2$, and compare to the known σ_g^2

Heritability Results



Huh?

Is this *really* how academics fight?

Kumar et al. (2016) - January 5, 2016

Here, we show that GCTA applied to current SNP data cannot produce reliable or stable estimates of heritability.

Yang et al. (2016) - July 25, 2016

We show below that those claims are false due to their misunderstanding of the theory and practice of random-effect models underlying genome-wide complex trait analysis.

Kumar et al. (2016) - July 25, 2016

We do not understand the basis for the claim that "the GREML fits all of the SNPs jointly in a random-effect model so that each SNP effect is fitted conditioning on the joint effects of all of the SNPs." Although Yang and colleagues insist on this fact, they do not provide any mathematical justification for this conclusion.

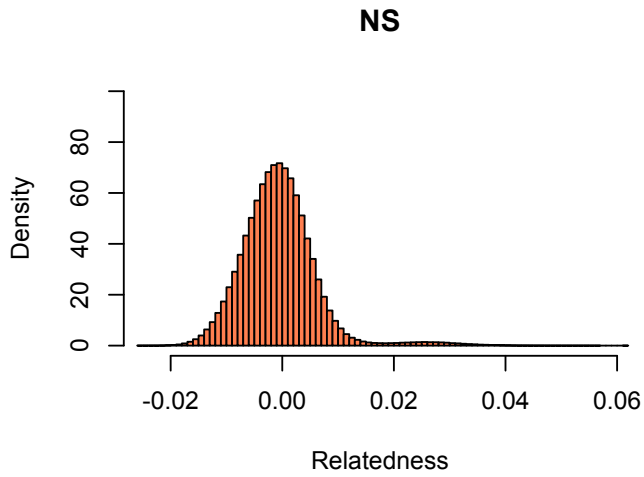
Simulation Take-Aways

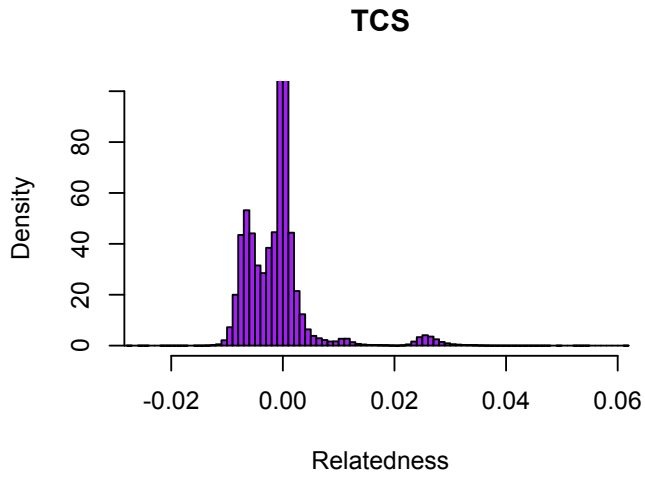
- Relatedness
  - Our newly proposed methods, like most shrinkage estimators, fail to estimate close relatives accurately
  - Refine the estimate of distant relatedness
  - Offer comparable, if not better, estimates for relatedness above 5th degree
- Heritability
  - Uhhh...
  - Quite difficult to attain any reasonable interpretation of these results
  - Need a better REML algorithm to utilize this model
  - Doctoral thesis? Draves et al. (2021)

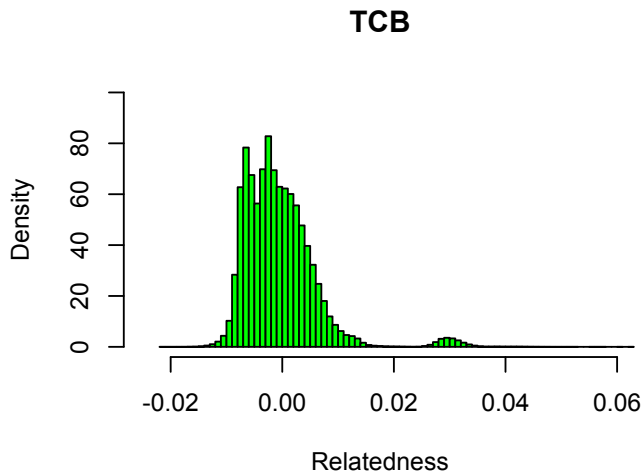
Health ABC - Study Description

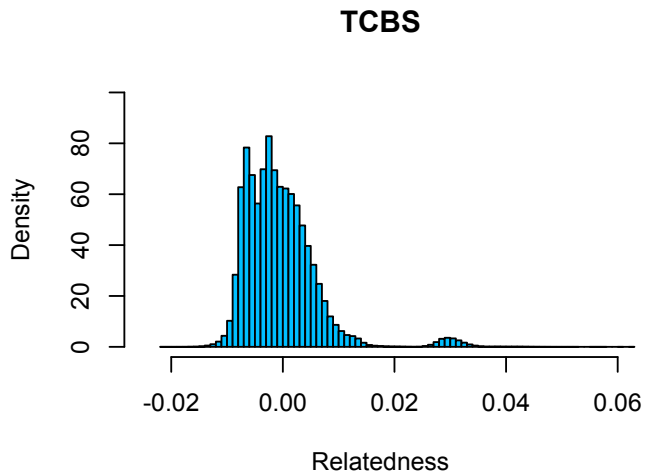
- 3,075 men and women from Memphis and Pittsburgh areas between the ages of 70 and 79
- 45% of women and 33% of men self reported African-American race
- We only consider the 1663 individuals who self reported White race seeing they made up the majority of the sample
- The study records and maintains SNP level information as well as several phenotypes
  - Body Mass Index (BMI)
  - Abdominal Visceral Fat Density (AVFD)

Relatedness Estimates









Heritability of BMI and AVFD

- BMI is 30-40% heritable [Zhang and Lupski (2015)]
- AVFD has maximal heritability, including non-genetic factors, of 48% [Rice et al. (1997)]

Method	BMI	AVFD
NS	44.5%	14.6%
TCS	99.9%	54.0%
TCB	22.8%	17.0%
TCBS	15.4%	18.0%

- TCS over estimates the heritability for both traits
- TCB and TCBS have more stable behavior and appropriately estimate these parameters

Conclusions

- This thesis develops two new methods that better utilize genome-level genetic data
- These methods better represent the inherent familial blockings within large samples to better estimate distant relatedness
- Our methods offer comparable estimates of relatedness for degree 5 relatives and higher
- We refine the estimate of relatedness for degree 7 and higher
- These better estimates *should* lead to better estimates of heritability
- Applying these methods to the Health ABC study, we show our methods stabilize the estimate of heritability in this setting

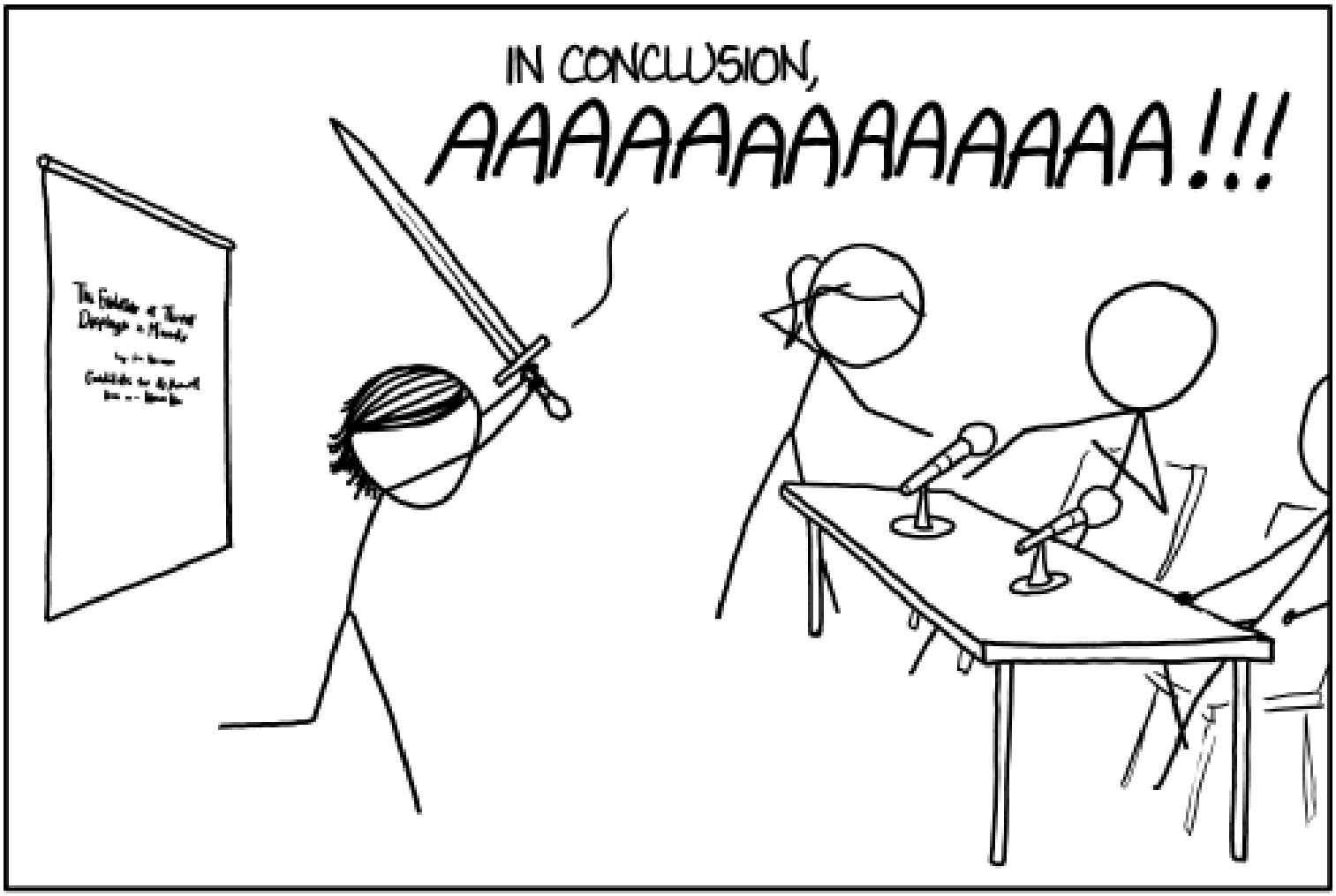
Future Work

- Better parameter selection - hierarchical clustering methods
- Account for SNP - SNP correlation via genetic distant & other correlation metrics
- Implement decompositions into software package
- Implement alternative methodologies for estimating heritability (e.g. regression techniques, mixture modeling)

Thank You

- Advisor: Trent Gaugler
- Committee: Eric Ho & Joy Zhou
- Jayne Trent
- Josh Arfin
- Math Lounge Rabble

Questions?


IN CONCLUSION,
AAAAAAAAAAAAAAA!!!
THE BEST THESIS DEFENSE IS A GOOD THESIS OFFENSE.

References I

- Crossett, A., A. B. Lee, L. Klei, B. Devlin and K. Roeder (2013): "Refining genetically inferred relationships using treelet covariance smoothing," The Annals of Applied Statistics, 7, 669-690.
- Gaugler, T., L. Klei, S. J. Sanders, C. A. Bodea, A. P. Goldberg, A. B. Lee, M. Mahajan, D. Manaa, Y. Pawitan, J. Reichert, S. Ripke, S. Sandin, P. Sklar, O. Svantesson, A. Reichenberg, C. M. Hultman, B. Devlin, K. Roeder and J. D. Buxbaum (2014): "Most genetic risk for autism resides with common variation," Nature Genetics, 46, 881-885.
- Kumar S.K.,Feldman M.W., Rehkopf D.H., and Tuljapurkar S. (2015). Limitations of GCTA as a solution to the missing heritability problem. PNAS 2016 113 (1) E61-E70; published ahead of print December 22, 2015.

References II

- Lee, A. B., Nadler, B. and Wasserman, L. (2008). Treeletsan adaptive multi-scale basis for sparse unordered data. Ann. Appl. Stat. 2 435471.
- Rice, T., Desprs, J. P., Daw, E. W., Gagnon, J., Borecki, I. B., Prusse, L., Leon, A. S., Skinner, J. S., Wilmore, J. H., Rao, D. C., and Bouchard, C. (1997). Familial Resemeblance for Abdominal Viseral Fat: The HERITAGE Family Study. Int J Obes Relat Metab Disord 21 (11), 1024-1031.
- Yang, J., Benyamin, B., McEvoy, B. P.,Gordon, S.,Henders, A. K.,Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010a). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42 565569.

References III


Yang, J., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2010b). GCTA: A tool for genome-wide complex trait analysis. The American Journal of Human Genetics 88 7682.


ang, J., Lee, S.H., Wray, N.R., Goddard, M.E., and Cisscher, P.M. (2016). GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. PNAS 2016 113 (32) E4579-E4580; published ahead of print July 25, 2016.


Zhang F, Lupski JR (2015). Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24:R10210. doi: 10.1093/hmg/ddv259