Darwinian Fitness & Aster Models

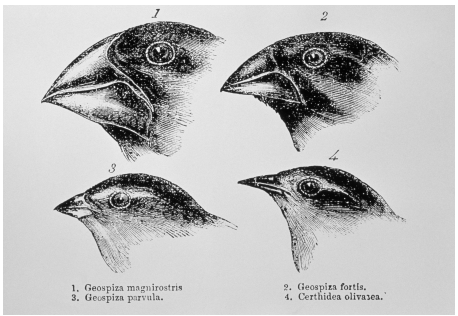
Allen J Clark

M.S. Statistics
University of Minnesota

December 20, 2022

3 Projects

Evolution
Estimate Darwinian fitness



Gene Expression
Find needles in a haystack

Movie Lens
Provide movie recommendations

Evolution

Darwinian Fitness

Fitness is the ability to pass down genetic information to future generations. How can we measure fitness?

$$\text{fitness} = \#\text{offspring}$$

Study Goals:

- Understand evolution better
- Plant and animal breeding
- Conservation

Statistical Goals:

- Estimate fitness
- Genes versus environment
- Rate of evolution

The Guinea Pig of Plants

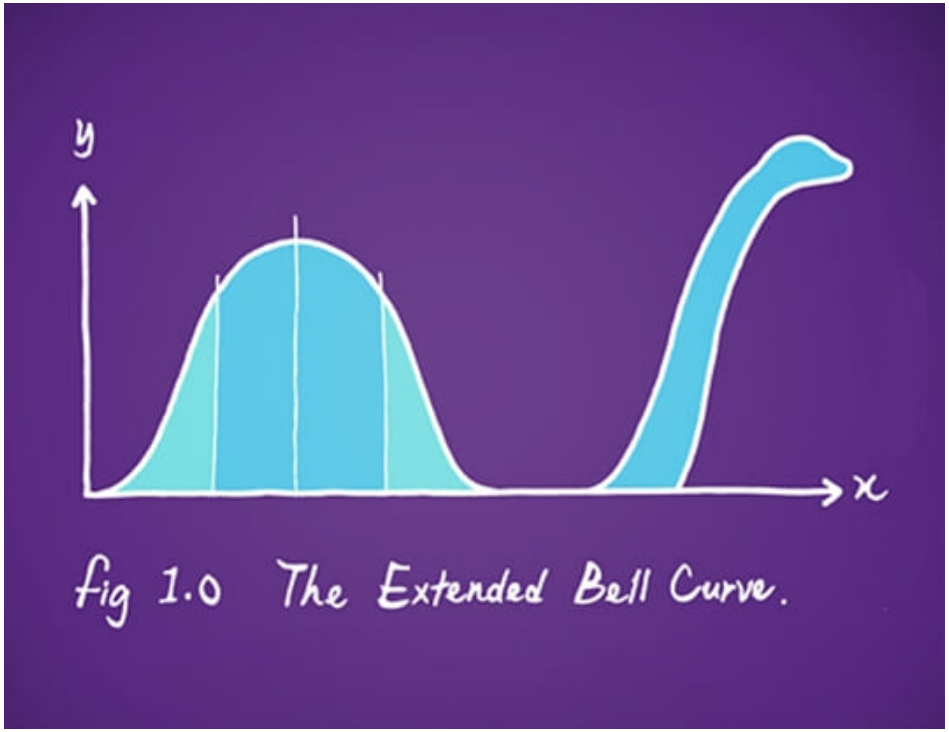
The partridge pea (Chamaecrista fasciculata):
Easy to study. Annual plant. Grows in Minnesota. Biologically convenient.

Key features

- Flowers
- Pea Pods
- Peas/Seeds

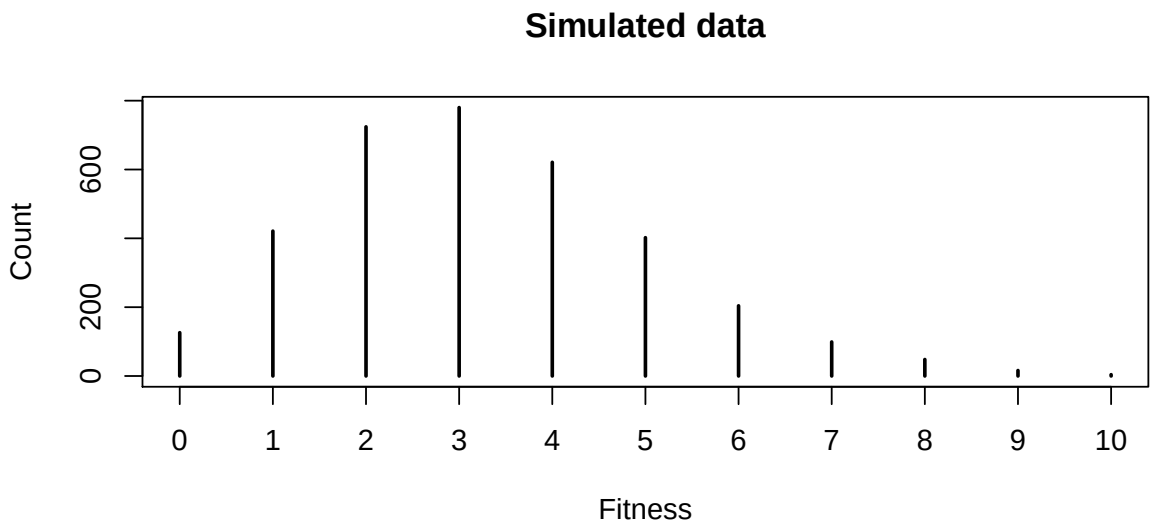
Fitness is proxied by seed count.

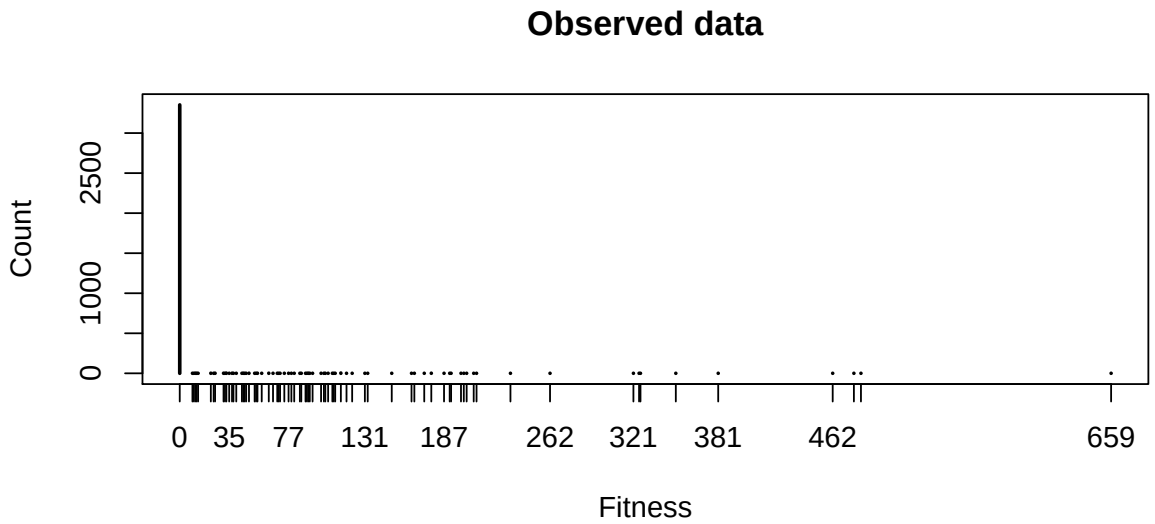
Fitness Doesn't Fit Any Distribution We Know



- Non-normal
- Heavy tailed (extreme values)
- Often zero
- Count data, but not Poisson

Not Poisson


Simulated data
Count
0
200
600
0
1
2
3
4
5
6
7
8
9
10
Fitness


Observed data
Count
0
1000
2500
0
35
77
131
187
262
321
381
462
659
Fitness

Components of Fitness

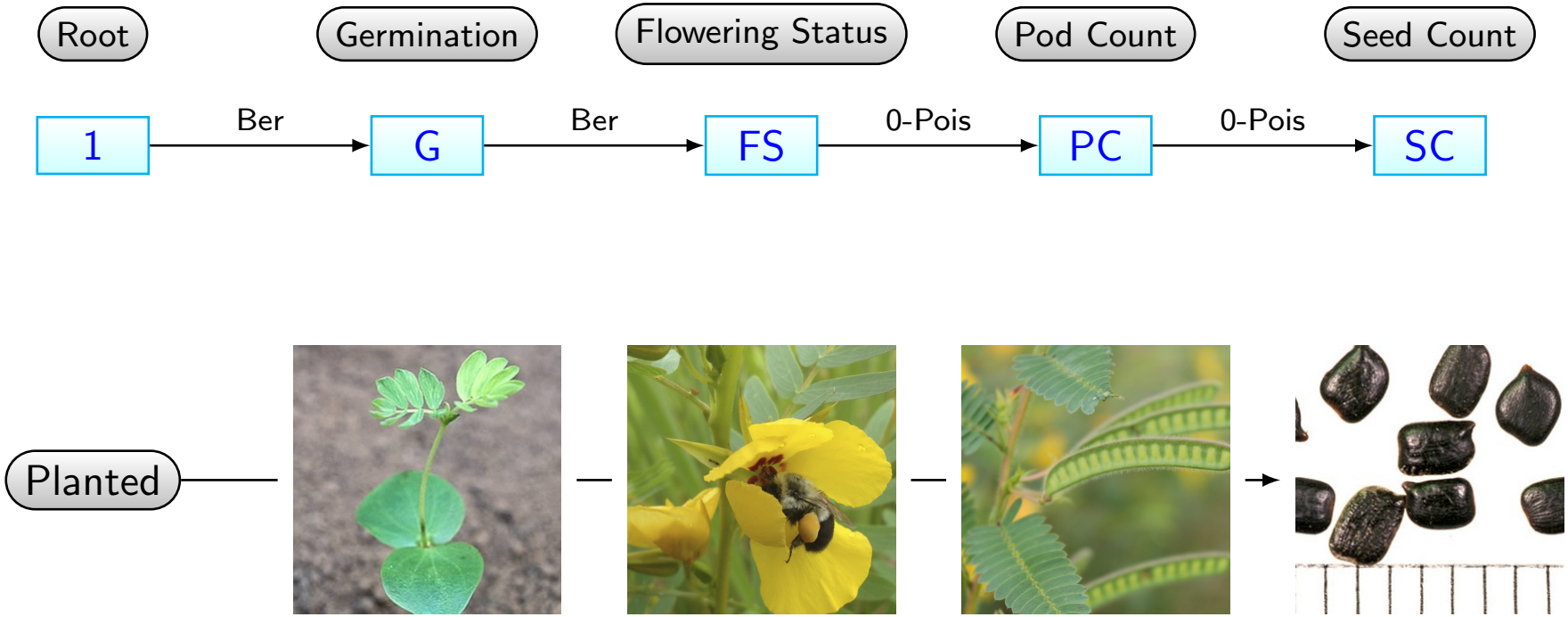


Figure: Components of fitness

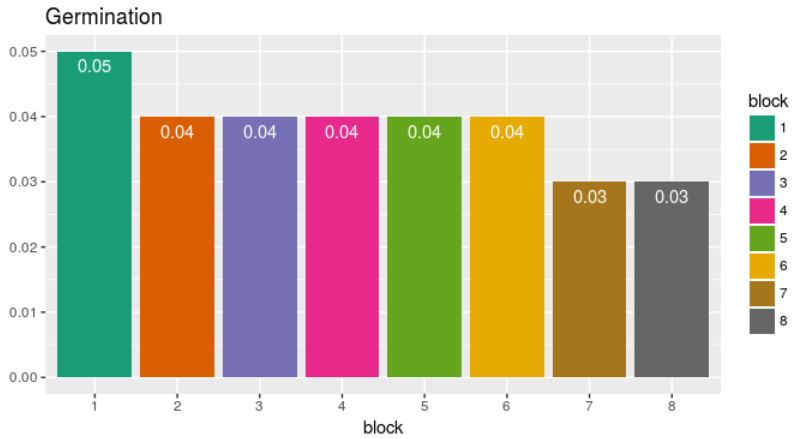
Aster Models

Linear models:	**GLM:**	**Aster models:**
$Y \sim \mathrm{N}(\mu, \sigma^2 I_n)$	$Y \sim \mathrm{Bin}(n \cdot p)$	$Y \sim \mathrm{ExpFam}(\varphi)$
$\mu = M\beta$	$\mathrm{logit}(p) = M\beta$	$\varphi = M\beta$
$Y_i \perp\!\!\!\perp Y_j \mid X$	$Y_i \perp\!\!\!\perp Y_j \mid X$	$Y_i \not\perp\!\!\!\perp Y_j \mid X$

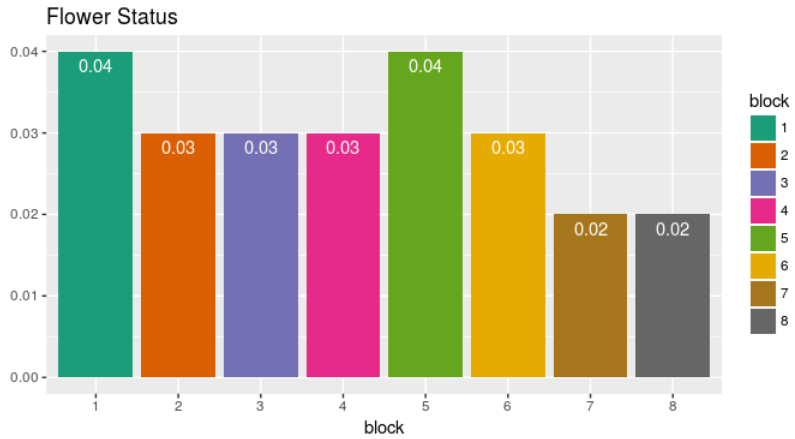
Aster models are . . .

- Generalized generalized linear models
- Exponential family models
- Graphical models
- Peculiar in that data is both a response and predictor
- Excellent for dealing with data of mixed types (continuous, categorical, binary, skewed, etc)

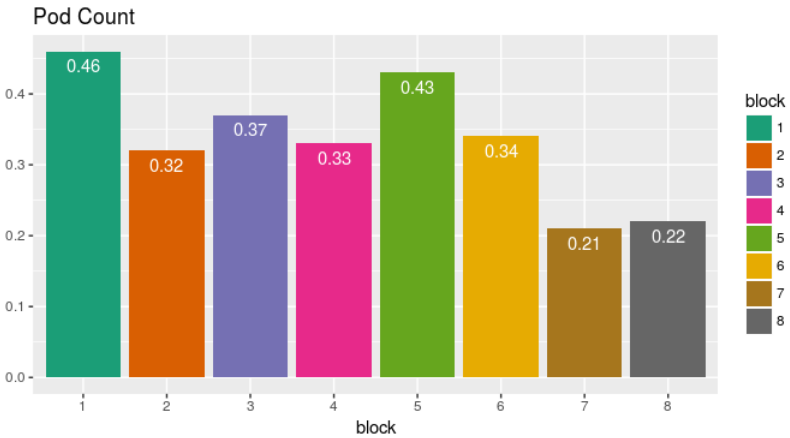
Some Results



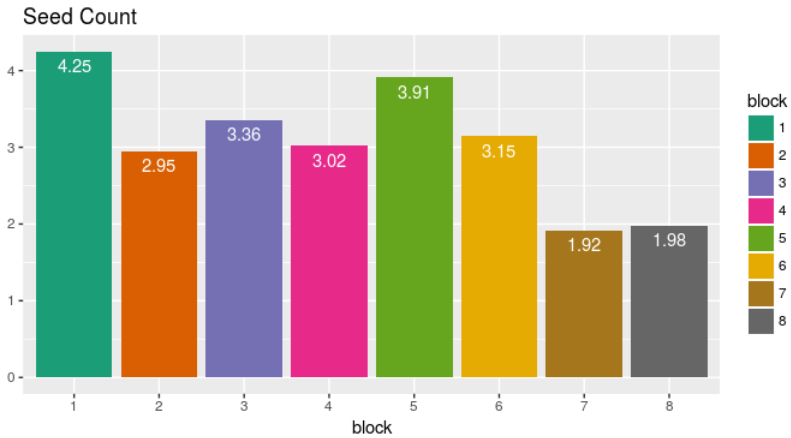
(a) Germination Probability ≈ 0.04



(b) Flowering Probability ≈ 0.03



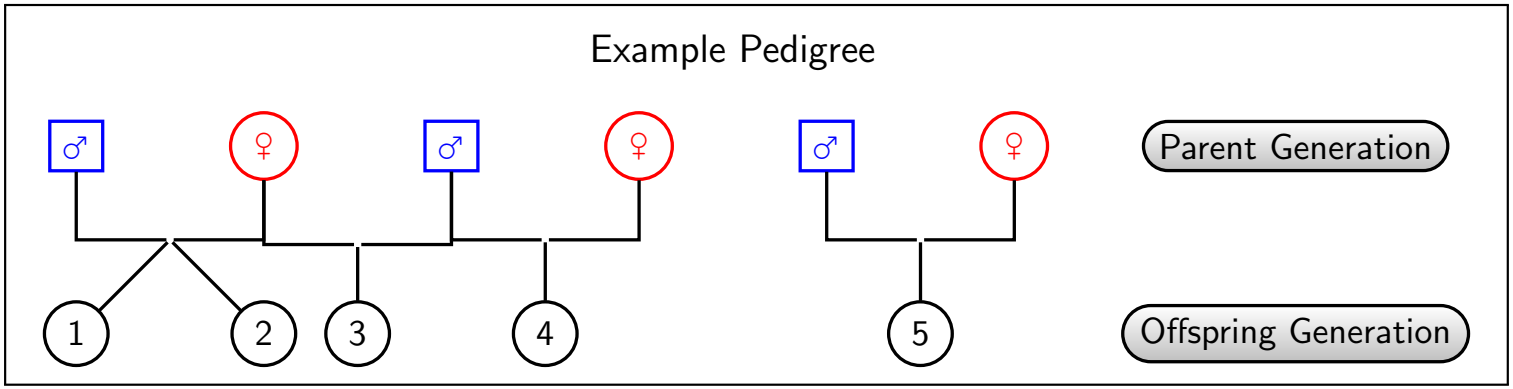
(c) Pod Count ≈ 0.36



(d) Seed Count ≈ 3.07

Adding Genetic Information

Genetic information is included through a pedigree. The parentage of each plant is recorded in the data set.



n_{ij} = genetics shared by individuals i and j. Siblings share half, half-siblings share a quarter, etc

$$N = \begin{bmatrix} 1 & 1/2 & 1/4 & 0 & 0 \\ 1/2 & 1 & 1/4 & 0 & 0 \\ 1/4 & 1/4 & 1 & 1/4 & 0 \\ 0 & 0 & 1/4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The Random Effect Aster Model

$$Y \sim \text{ExpFam}(\varphi)$$
$$\varphi = M\beta + Zb$$

- φ - Exponential family parameter
- M - Fixed effects model matrix
- β - Fixed effect coefficients
- Z - Random effect model matrix
- b - Breeding values

Assume $b \sim \text{N}\left(0, \sigma_b^2 N\right)$
φ has aster model likelihood

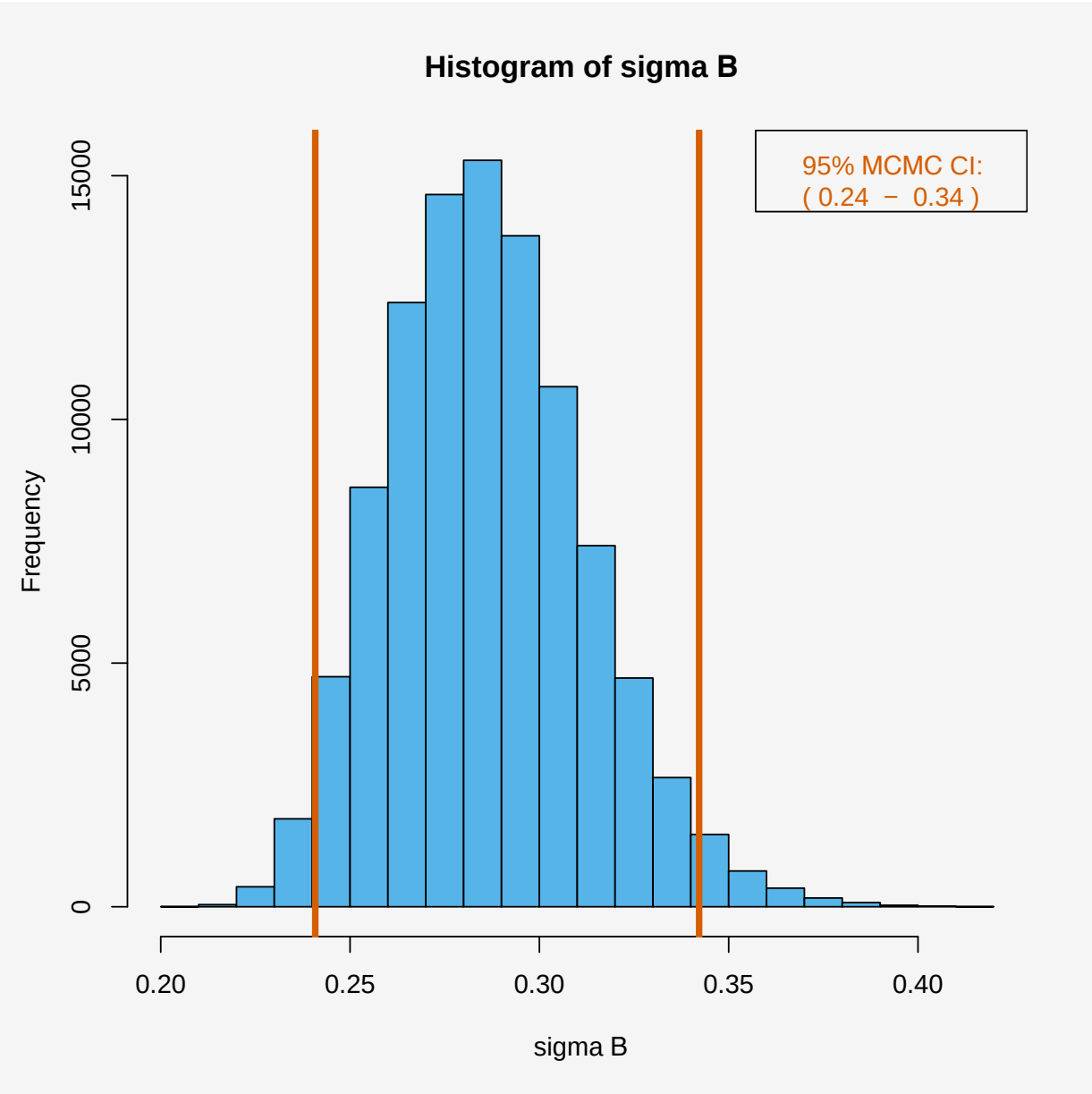
Fitting the Random Effects

- High dimensional data - 3300 random effects
- No known exact methods, previous approximate methods failed
- Bayesian Markov chain Monte Carlo (MCMC) approach
- Bayes theorem:

$$\underbrace{P(\theta|X)}_{\text{posterior}} = \frac{\overbrace{L(\theta;X)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{\int L(\theta;X) \cdot P(\theta) d\theta}_{\text{normalizing constant}}}$$

- MCMC: Metropolis random-walk algorithm

Breeding Value Standard Deviation



The Frontier

Figure: Fisher the ~~statistician~~ geneticist

Never been done with aster models before.
Fisher's fundamental theorem of natural selection:

> *'Change in fitness due to natural selection in one generation is mean fitness over additive genetic variance.'*

$$\Delta^{\text{ns}}\left(\text{Fitness}\right) = \frac{\mu(\text{Fitness})}{\sigma_a^2}$$

To Do

Additive genetic variance σ_a^2 related to breeding value variance σ_b^2.

- Transform σ_b^2: ExpFam scale $\rightarrow$ mean value scale
- Regress fitness on σ_b^2
- σ_a^2 are residuals from regression
- Repeat for every iteration of the Markov chain
- Repeat for every block

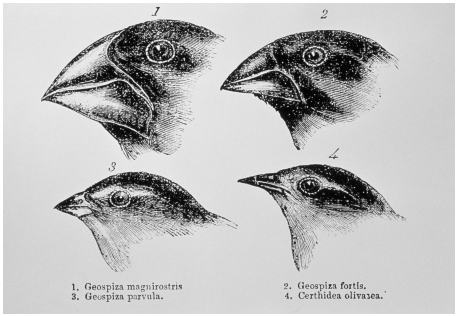
Pea-Pods and Beyond

Other applications of aster models:

- Insurance loss (claims y/n, claim count, claim cost)
- Manufacturing yield
- Complicated systems that can be modeled by simple parts

Questions?

Finished:
Aster Models



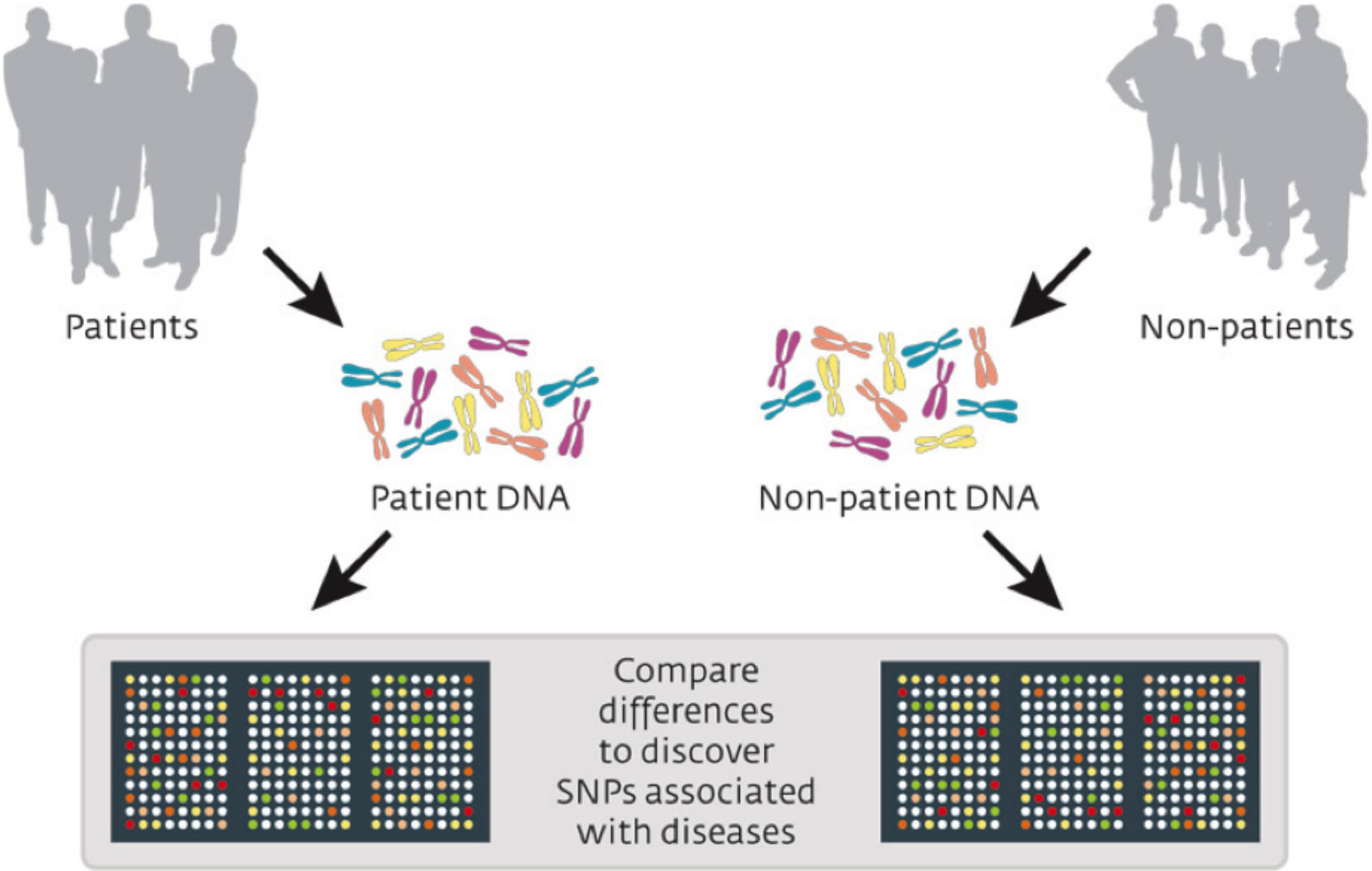
Next:
Gene Expression

Need for Precision Medicine

'An emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment and lifestyle for each person.' National Institute of Health (NIH)

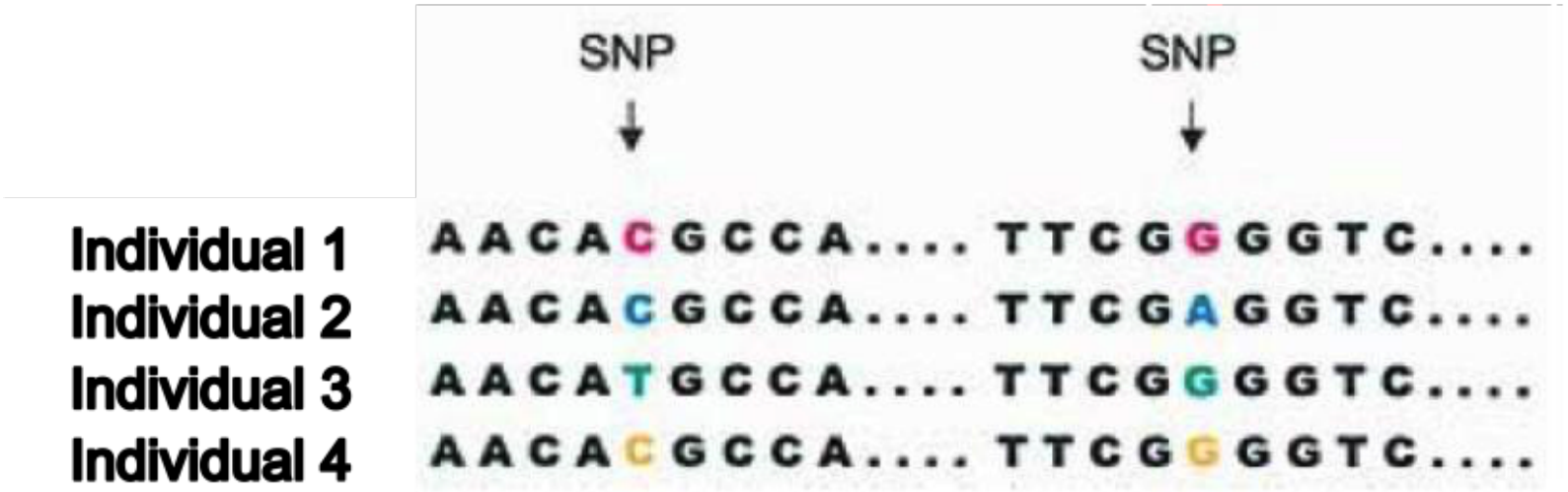
Genome-Wide Association Study (GWAS)

Examining a genome-wide set of genetic variants (SNPs) in different individuals to see if any SNP is associated with a trait (complex disease)



Single-Nucleotide Polymorphism (SNPs)

- Variation in a single-nucleotide that occurs at a specific position in the genome (e.g.$>$ 1% of the population)



Challenges in GWAS

- Limited statistical power to identify disease-associated loci ($p >> n$)
- Small effect size of each SNP (needle in a haystack)
- Difficulty to interpret the biological relevance of susceptible loci and link with disease etiology.

Normal Prostate Tissue Data

Goal: predict phenotype from SNPs

- $n = 471$ subjects
- 88 data sets corresponding to 88 genes
- Response: phenotype in normal prostate tissue after regressing out non-SNP covariates
- Predictors: each gene has 7000 to 32000 SNPs

index	phenotype	rs111417370_T	rs2003816_T	rs116897926_A	$\cdots$	rs114857962_A
1	0.004	0.0	1	0.0	$\cdots$	0.002
2	0.328	0.0	1	0.0	$\cdots$	0.004
3	0.164	0.4	0	0.0	$\cdots$	0.000
4	-0.107	0.0	1	0.0	$\cdots$	0.989
5	-0.109	0.0	0	0.9	$\cdots$	1.000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Statistical Methods

Method Criteria:

- Predictive ability
- Simpler is better
- Penalization useful because $p >> n$

Methods used:

- Lasso
- Elastic Net
- Truncated Lasso Penalty (TLP)
- Smoothly Clipped Absolute Deviation (SCAD)
- Random Forest

Model Evaluation

- Divided data into 80% training and 20% test data sets.
- 5-Fold Cross Validation (CV) to evaluate the performance of the models.
- Performance measures:
  - Mean Squared Error (MSE)
  - Relative Error (MSE/variance(y))

Methods	Parameters tuned
LASSO	λ
SCAD	λ
TLP	λ
Elastic Net	α, λ
Random Forest	*mtry*, *ntree*

- Repeat the procedure for each method for 88 data sets

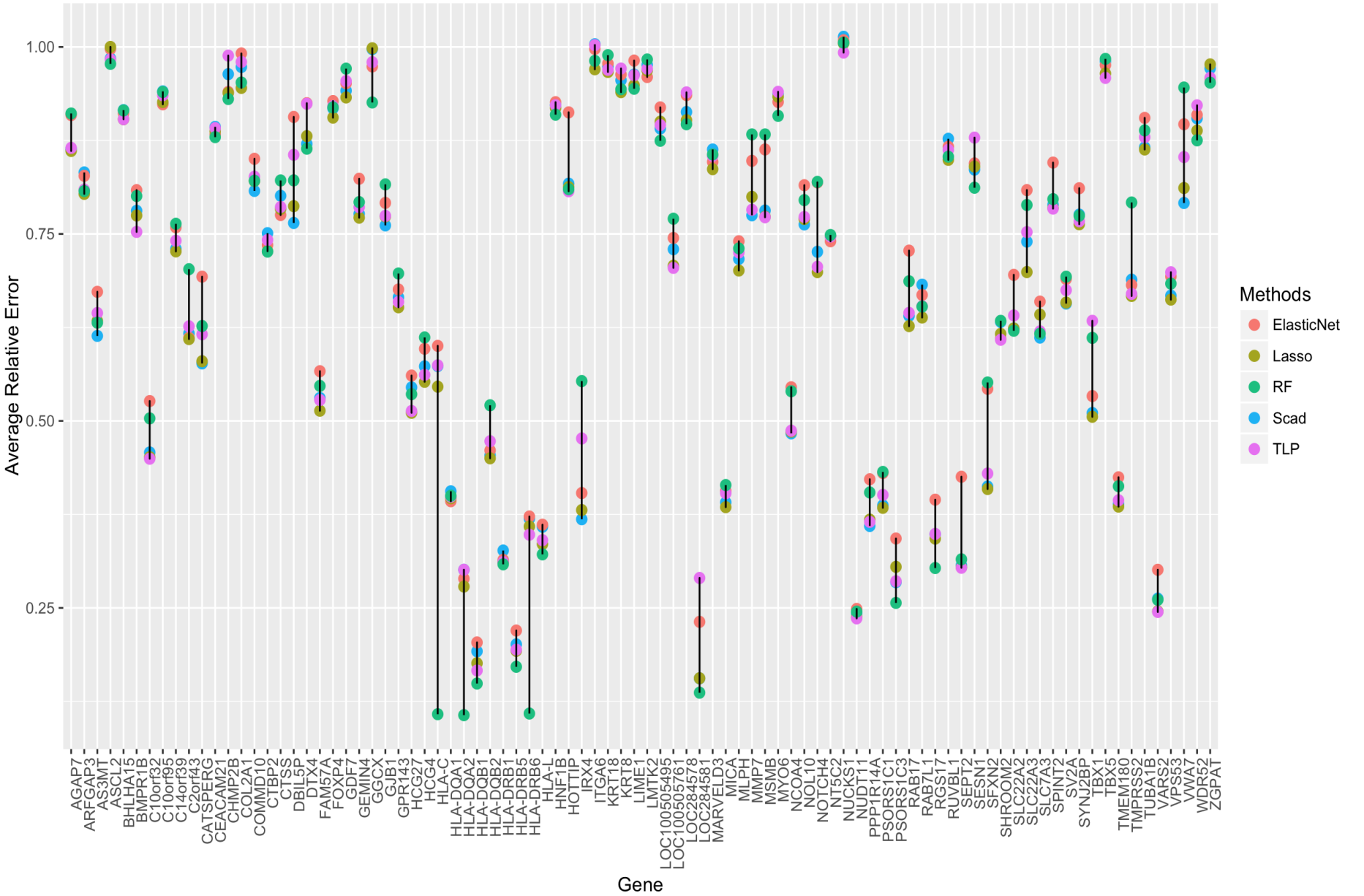
Parallel Processing

Saving computation time on multi-core linux platforms

- Parallel processing speeds up the code.
- Code run in parallel with one core for each CV fold
- Reduced running time by 71% (450 sec $\rightarrow$ 130 sec)

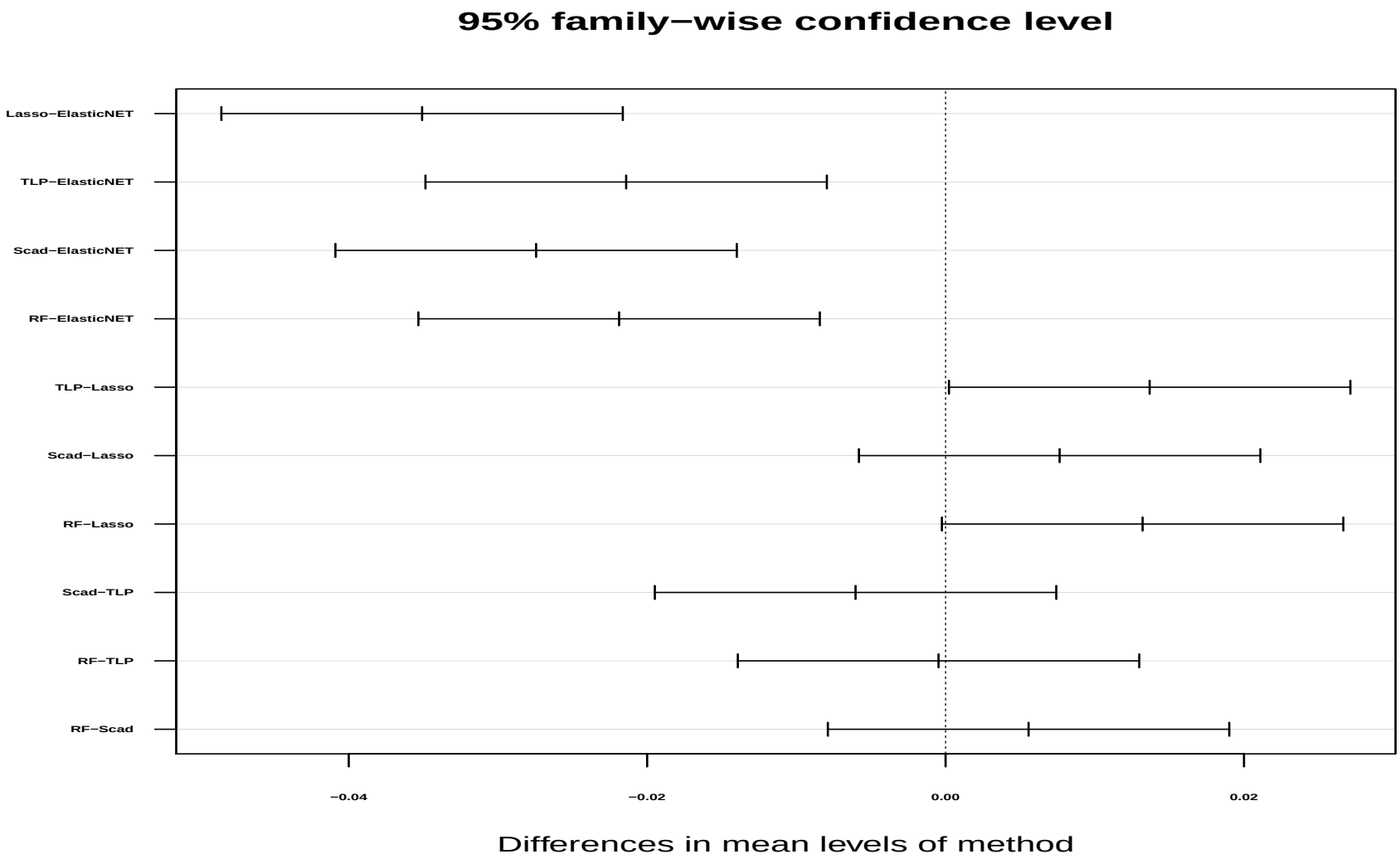
```
library ( parallel )

no_cores <- 5 # Number of Cores

cl <- makeCluster(no_cores) # Initiate Cluster

# Export Necessary Functions
toexport = c("tlp_fold", ... )
clusterExport(cl, toexport)

parApply(cl = cl, X = CVFolds ,MARGIN = 2, FUN = tlp_fold, data = data)
```

Results by Gene



Overall Method Comparison

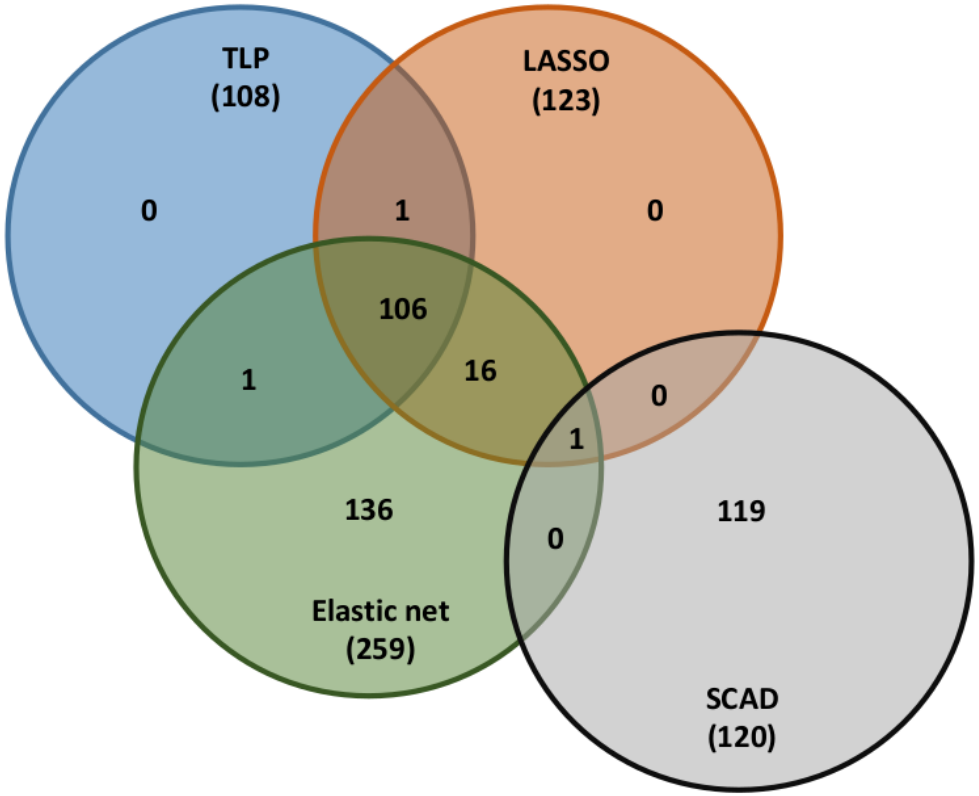
- Elastic Net most predictive. Other methods similar.

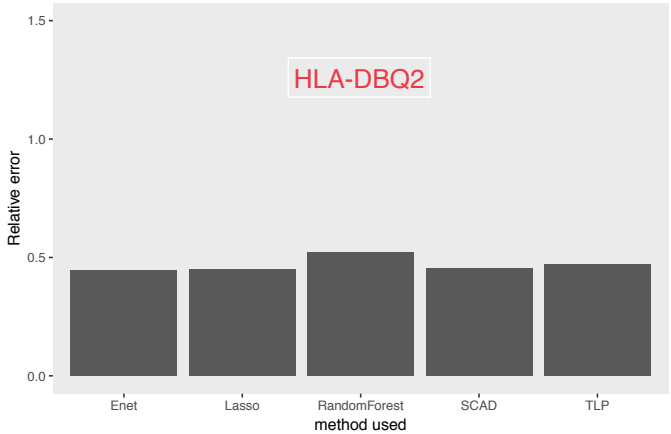


Variable Selection

Gene: HLA-DQB2.chr6

- 24743 SNPs
- Shrunk to $\sim 100 - 300$





- TLP - simple and interpretable
- SCAD - strange
- To do - check for correlations

Questions?

Finished:
Gene Expression

Next:
Movie Predictions

Netflix Prize

Goal: Improve movie recommendations

- Beat Netflix's Cinematch algorithm by 10%.
- RMSE Metric: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- \$1M prize.
- Won in 2009 by Pragmatic Chaos team

Figure: Belkor's Pragmatic Chaos

Movie Lens 100K

Netflix data taken down after 2009. The UMN research group "Grouplens" provides similar data.

- 100K ratings
- 943 users
- 1682 movies
- $R(u, i) =$ user u rating on item i
- Movie info: Genre and release year

user	Star Wars	LOTR	The Godfather	Sprited Away	$\cdots$
1	5	4	1	NA	$\cdots$
2	4	3	5	NA	$\cdots$
3	NA	NA	3	2	$\cdots$
4	3	5	NA	3	$\cdots$
5	5	NA	NA	5	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Recommender Systems

Goal: Estimate $\hat{R}(u, i)$

Content Based RS

- Relies on similarity of items
- User independence
- Over specialization
- New user issue: Cannot make predictions for new users

Collaborative Filtering RS

- Relies on similarity between users
- User dependence
- More diverse recommendations than CB
- New item issue: Cannot make predictions on new items

Content Based RS

- How it works: item similarity matrix S
- Cosine similarity

$$S_{ij} = \cos\theta = \vec{m}_1 \cdot \vec{m}_2$$

- Predicted ratings:

$$\hat{R}(u,i) = \frac{\sum_j R(u,j) S_{ij}}{\sum_j R(u,j)}$$

- Code: implemented in `R`
- 5-fold CV: $RMSE \approx 1.43$

Collaborative Filtering RS

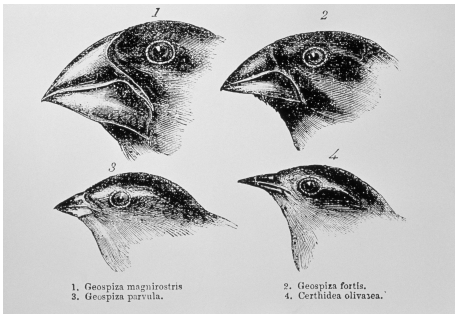
- How it works: user similarity matrix S
- Pearson similarity on items rated by both user u and user v

$$S_{uv} = \frac{\sum_i \left(R(u,i) - \overline{R}_i\right)\left(R(v,i) - \overline{R}_i\right)}{\sqrt{\sum_i \left(R(u,i) - \overline{R}_i\right)^2 \sum_i \left(R(v,i) - \overline{R}_i\right)^2}}$$

- Predicted ratings:

$$\hat{R}(u,i) = \overline{R}_i + \frac{\sum_v S_{uv}\left(R(v,i) - \overline{R_v}\right)}{\sum_v S_{uv} S_{uv}}$$

- Code: `graphlab` module in `python`
- 5-fold CV: $RMSE \approx 1.02$



Applying numerical optimization techniques to ongoing scientific research

Finding small signals in high noise situations

NETFLIX

Predicting outcomes in complex data environments

Thank you