Leveraging 1000 Genomes to Improve Disease Gene Localization

William C. L. Stewart, Ph.D.
Assistant Professor of Statistics and Pediatrics
Principal Investigator at The Research Institute

Nationwide Children's Hospital
William.Stewart@nationwidechildrens.org

June 9, 2017

Nationwide Children's
When your child needs a hospital, everything matters.℠

The Ohio State University
College of Medicine

Talk Outline

I Multipoint Likelihoods & Large Families
  - A Statistical problems
  - B Computational problems

II CIs for trait location (denoted θ)
  - A The standard estimator: $\hat{\theta}$
  - B The Stewart & Peljto estimator: $\tilde{\theta}$
  - C Variance estimators

III A Simulation Study

IV Approximate Importance Sampling

Likelihoods on Large Families w/ Dense SNPs

Statistical Problems:

- $\mathcal{H}_o : \theta = \pm\infty$
- $L(\delta) \neq L(\theta)$
- *iid* no longer applies.
- Confidence Intervals (CIs)? Validity? Efficient?
- How do we know that the *chosen* SNPs are the right SNPs?

Likelihoods on Large Families w/ Dense SNPs

Computational Problems:

- A large latent space
- Sparse SNP panels *increase* error!
- Likelihoods involving dense SNPs are intractable

CIs for θ with Dense SNPs: Notation

Let **G** denote the obs'd dense SNP genotype data.

Ideally, we want to base inference on $\hat{\theta}(\mathbf{G})$, but this likelihood is intractable. Why? Because for certain pairs of SNPs there is linkage disequilibrium (LD), which means that

$$Pr(A_i - B_j) \neq Pr(A_i)Pr(B_j) \tag{1}$$

for any allele A and B of loci i and j, resp.

Note that LE (linkage equilibrium) implies equality in (1), and that likelihoods for sparse subsamples (denoted **M**) are tractable provided that the SNPs are all in LE.

Thus, $\mathbf{M} \equiv \mathbf{M}(\mathbf{G}, \mathbf{S})$, where **S** denotes a sparse SNP panel.

CIs for θ with Dense SNPs: The Standard Estimator

Most researchers choose an $\mathbf{S}$ that they think is *best* and compute $\hat{\theta}(\mathbf{G} \mid \mathbf{S})$. But no one knows what $\mathbf{S}$ is best, and so everyone uses a different $\mathbf{S}$ (either by choice or by force), which means that the standard estimator $\hat{\theta}(\mathbf{G}, \mathbf{S})$ is, in fact, random in *both* $\mathbf{G}$ and $\mathbf{S}$.

Furthermore, although dense SNP panels change every year, and from platform to platform, *etc.*, efficient estimators of θ are invariant to dense SNP panels because two different dense SNP panels will provide nearly identically information about θ.

CIs for θ with Dense SNPs: The Stewart-Peljto Estimator

In '10, we proposed $\tilde{\theta} \equiv \mathbf{E}\hat{\theta} \mid \mathbf{G}$, where $\mathbf{E}(\cdot)$ is taken *wrt* $Pr(\mathbf{S} \mid \mathbf{G})$.

In practice however, we estimate $\tilde{\theta}$ by $\frac{1}{k}\sum \hat{\theta}(\mathbf{G}, \mathbf{S}_j)$, from $\mathbf{S}_1, \ldots, \mathbf{S}_k$ realizations of $\mathbf{S} \sim Pr(\mathbf{S} \mid \mathbf{G})$.

Note that (if you can compute it),

$$\begin{aligned} Var(\tilde{\theta}) &= Var[\,\mathbf{E}\,\hat{\theta} \mid \mathbf{G}\,] \\ &\leq Var[\,\mathbf{E}\,\hat{\theta} \mid \mathbf{G}\,] + \mathbf{E}Var(\hat{\theta} \mid \mathbf{G}) \\ &= Var[\hat{\theta}(\mathbf{G}, \mathbf{S})] \end{aligned}$$

CIs for θ with Dense SNPs: Variance Estimators

For a large number of small families a nonparametric bootstrap approach is quite effective.

For a small number of large families

(1) nonparametric bootstrap is no longer applicable,
(2) a minus 1-LOD unit approach is approximate, at best
(3) simulation of **G** conditional on obs'd trait data is computationally infeasible due to LD.

Also note that the $Var(\hat{\theta}) \approx \mathbf{E}[Var(\hat{\theta} \mid \mathbf{S}]$. Eq. (1)

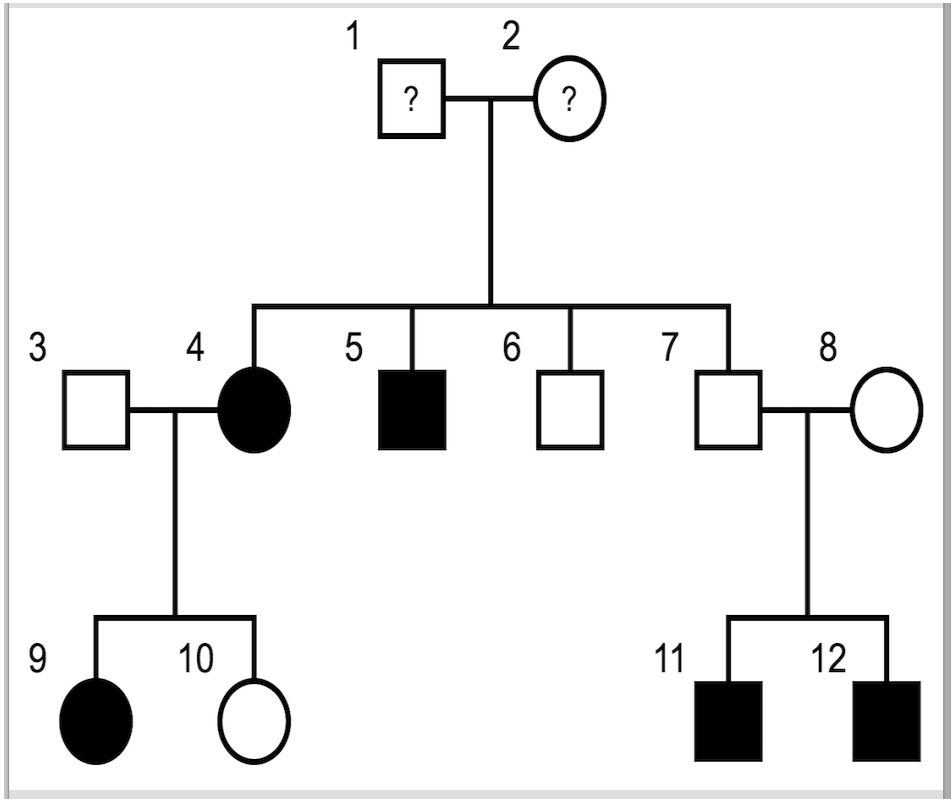
A Simulation Study: Design

To quantify the gains in precision of $\tilde{\theta}$ for large families,

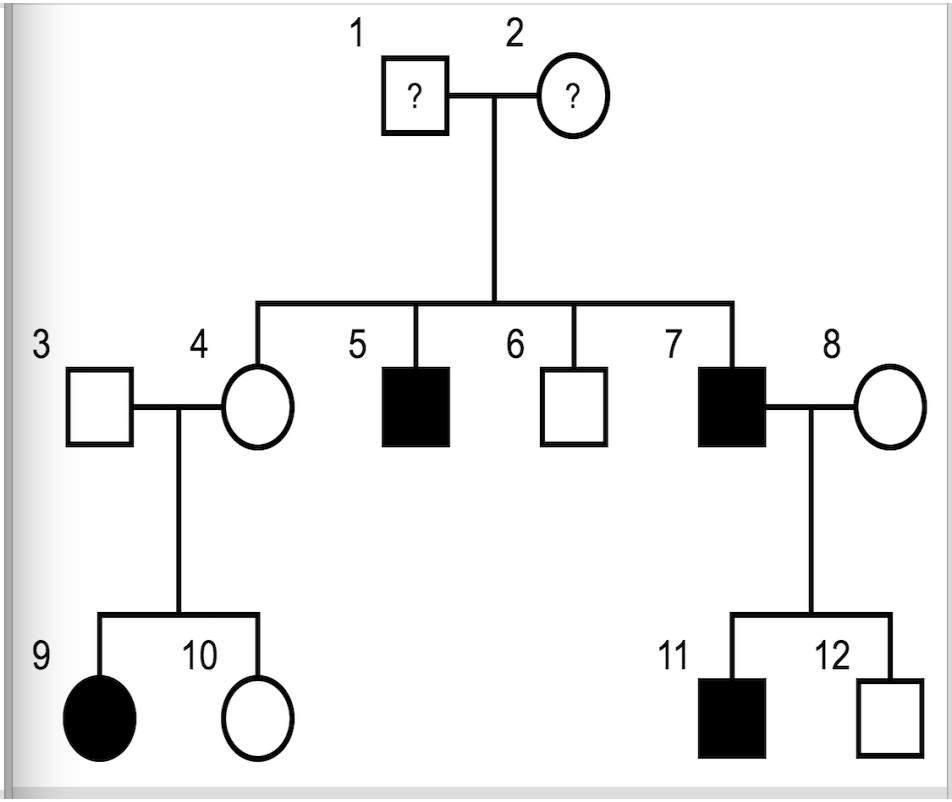
- We generated dense SNP cosegregation data on 3-generation families for dominant (DOM) and recessive (REC) traits with incomplete penetrance, positioned in the middle of 132 haplotype-blocks (avg spacing between blocks is 0.5 cMs; avg spacing between SNPs is 0.25 cMs; each block has 3 SNPs).
- For DOM: disease allele frequency of 1%, a phenocopy rate of 1%, and a penetrance of 20%.
- For REC: the corresponding parameters were 10%, 1%, and 50%, respectively; each replicate contained seven families.

A Simulation Study: Design

5–DOM pedigrees per replicate

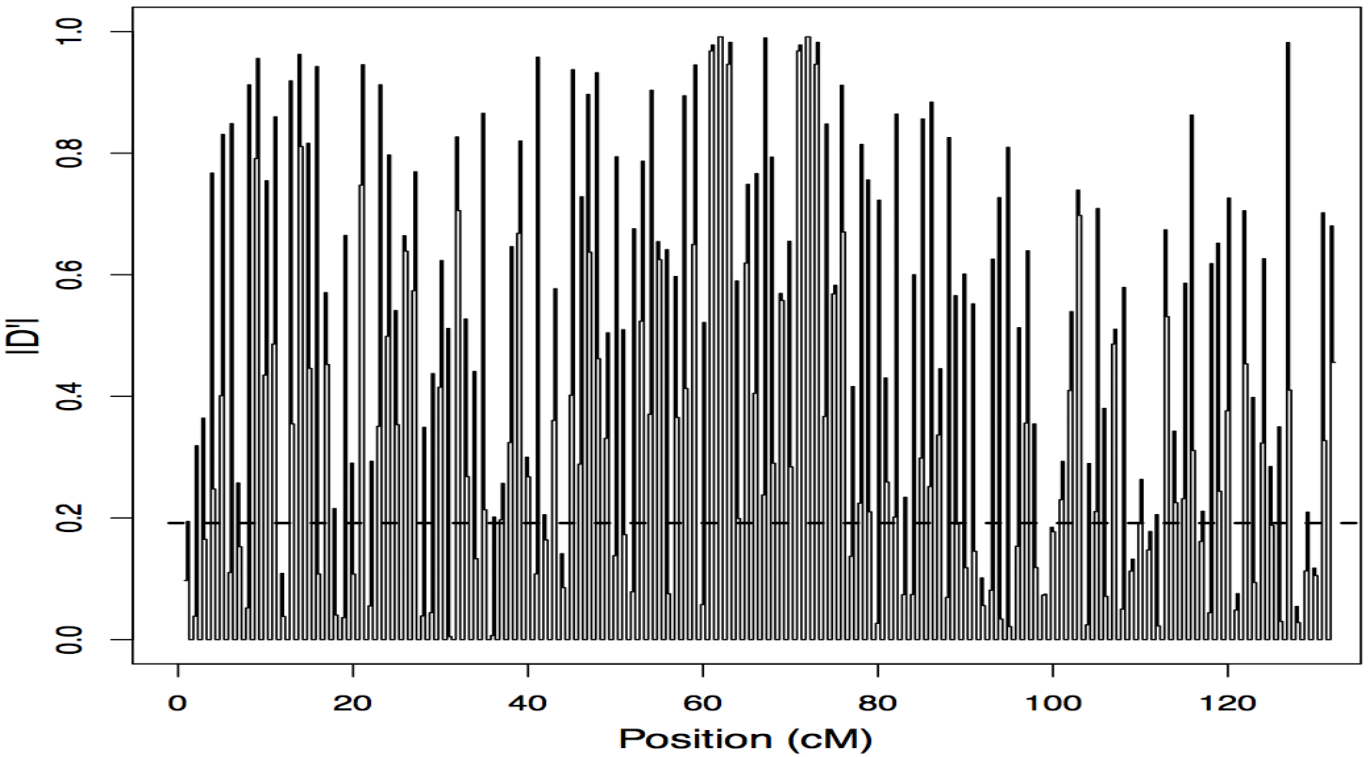


7–REC pedigrees per replicate



A Simulation Study: Design

LD structure of the 132 haplotype blocks:



A Simulation Study: Results

The conditional variance formula holds: $V\hat{\theta} = V\tilde{\theta} + \mathbf{E}V\hat{\theta} \mid \mathbf{G}$

Trait	$\mathbf{V}\hat{\theta}$	$\mathbf{V}\tilde{\theta}$	**AMCE**
DOM	47.27 (46.39)	38.43	8.84
REC	64.25 (63.13)	53.01	11.24

- CI lengths are reduced by 10%.
- $V\hat{\theta}$ in parentheses is computed by $\mathbf{E}V\hat{\theta} \mid \mathbf{S}$.

Approximate Importance Sampling (AIS)

Recall that the average Monte Carlo error (**AMCE**) is:

$$\begin{aligned} \mathbf{E}V\hat{\theta} \mid \mathbf{G} &= \sum [V\hat{\theta} \mid \mathbf{G}]\, Pr(\mathbf{G}) \\ &= \sum [V\hat{\theta} \mid \mathbf{G}]\, \frac{Pr^*(\mathbf{G})}{Pr(\mathbf{G})} Pr(\mathbf{G}) \\ &\approx \sum [V\hat{\theta} \mid \mathbf{G}]\, \frac{Pr^*(\mathbf{M}')}{Pr(\mathbf{M}')} Pr(\mathbf{G}), \end{aligned}$$

where $\mathbf{M}' \equiv \mathbf{M}(\mathbf{G}, \mathbf{S}')$ and $\mathbf{S}'$ minimizes $V\hat{\theta} \mid \mathbf{S}$.

For REC: **AMCE** $= 11.24$ (cond. sim), and **AMCE** $= 11.78$ (AIS).

Incorporating 1000 Genomes

Currently, we use the EM algorithm to estimate the dependence structure of **G**, which facilitates the rapid and efficient subsampling of **M**. But the accuracy of this estimation could be greatly improved by the inclusion of ethnically matched data from 1000 Genomes data.

Then, the **G** simulation with LD but without cosegregation could be improve by coalescent-based programs like FastPhase, or Markov chain-type programs like Haplodrop.

Futhermore, with AIS and Eq. (1), all condition simulations can happen under LE, and all LD simulations can be unconditional.

Conclusions

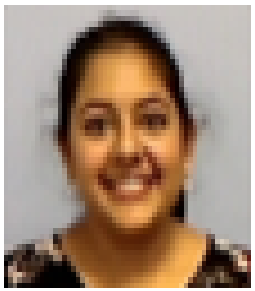
1. In principle, accurate, narrow CIs for a small number of large families is, now a reality.
2. By shrinking the candidate gene region, sequencing costs will decrease, and the power to detect associations underneath cosegregation peaks will increase.

Acknowledgements

I would like to thank the Stewart Lab...

Post-docs: Meng Wang, Jane Cerise
Students: Jessica Fladen, Akanksha Malhotra, Komla Gnona, Valerie Hager, Esther Drill, and Anna Peljto.







Nationwide Children's
When your child needs a hospital, everything matters.℠

The Ohio State University
College of Medicine

For those who are interested in our software:

William.Stewart@nationwidechildrens.org

Thank You!

http://u.osu.edu/stewart.1212/

The Kong & Cox Likelihood ('97) at location j

$$\log L(\delta) = c(\mathbf{D}) + \sum_i \log[1 + \delta\,(f(\mathbf{D}_i) - \mu_i)/\sigma_i],$$

where $f(\mathbf{D}_i) = \mathbf{E}(S_i \mid \mathbf{D}_i, H_o : \delta = 0)$ and $i = 1, 2, ..., n$ for n independent families in the data set.

Now define $Z^j \equiv \mathrm{sgn}(\hat{\delta})\sqrt{2\,[\log L(\hat{\delta}) - \log L(0)]}$ for $j = 1, 2, ..., m$ for m markers in a genome-wide cosegregation scan.

$$Z^j \to N(0, 1) \ \text{ as } \ n \to \infty$$