

Generalized Regional Admixture Mapping (RAM) and Structured Association Testing (SAT)

David T. Redden, Associate
Professor, Department of
Biostatistics, University of
Alabama at Birmingham

Acknowledgements

Jose Fernandez
Jasmin Divers
Kelly Vaughan
Solomon Musani
Hemant Tiwari
Miguel Padilla
Michael B. Miller

Rui Feng
Nianjun Liu
Guimin Gao
T. Mark Beasley
Robert P. Kimberly
David B. Allison

The problem and the promise

- Admixture, the event of two or more populations with different allele frequencies intermating, creates offspring with linkage disequilibrium that spans a greater distance than in a panmictic population.

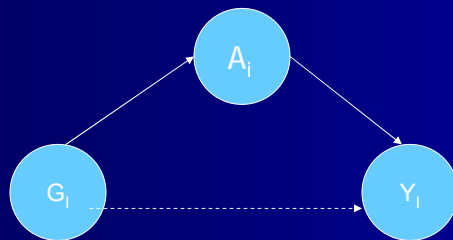
The problem and the promise

- This admixture process can, under some circumstances, create disequilibrium between pairs of unlinked loci and thus create confounding (spurious associations, inflated false positive results) in genetic association studies between trait and marker.

The problem and the promise

- The classic example is found in Knowler et al (1988). They reported an association between an HLA haplotype and diabetes for Pima Indians. When the analysis was repeated stratifying subjects by amount of European ancestry, the observed association between HLA haplotype and diabetes was not present.

The Problem



Structured Association Tests

- In response to this problem, many authors have proposed a collection of methods we will collectively call SAT.

The promise

- Regional admixture mapping (RAM) methods use genome wide ancestry and region specific admixture estimates to identify specific regions of the genome potentially harboring loci influencing the trait.

The promise (Illustrated)

- Hypothetical Segment of an Admixed Individual.



Objective of the Paper

- To extend both Regional Admixture Mapping (RAM) and Structured Association Tests (SAT) into a regression modeling framework.
- This would allow for tests of dominance, allow for either continuous or dichotomous outcomes, and allow for inclusion of covariates.

Regional Admixture Mapping

- Using a sample of admixed individuals, estimate each individual's ancestry as well as estimate the ancestry of an individual's alleles within specific genomic regions.
- Given the increased linkage disequilibrium in admixed populations and assuming a disease/phenotype which is more prevalent within a parental population (P_d) genomic regions exhibiting a high number of alleles from P_d may harbor/be linked to causative alleles.

Regional Admixture Mapping

- By comparing regional admixture estimates to an individual's total admixture estimate, some authors (Zhu et al 2004, Patterson et al 2004, Montana and Pritchard 2004) have recommended case only designs.
- Other authors have recommended comparing regional admixture estimates between cases and controls.

Structured Association Tests

- The SAT approach seeks to test for the association of a marker and phenotype after making adjustments for population stratification.
- Examples include Devlin and Roeder (1999), Pritchard et al (2001), Satten et al (2001).

Structured Association Tests

- All examine for association between markers and case/control status. The methods have not been generalized to continuous outcomes.

Our Proposed Model (RAM)

- $Y_i = \beta_0 + \beta_1 A_i + \beta_2 (P_{1i} * P_{2i}) + \beta_3 A_{i,j,1} + \beta_4 A_{i,j,2} + \varepsilon_i$
- A_i is the ancestry for the i^{th} individual (to be estimated) which is $(P_{1i} + P_{2i})/2$.
- P_{1i} is the ancestry of Parent 1 and P_{2i} is the ancestry of parent 2.

Controlling for Ancestry

- Let P_1 = ancestry of parent 1 from population D
- Let P_2 = ancestry of parent 2 from population D
- Let V_i be the number of alleles at a random loci that their child has from population D

Controlling for Ancestry

- $P(V_i = 0 | P_1 P_2) = (1-P_1)*(1-P_2)$
 $= 1 - P_1 - P_2 + P_1 * P_2$
- $P(V_i = 1 | P_1 P_2) =$
 $(1-P_1)*P_2 + (1-P_2)*P_1 = P_1 + P_2 - 2*P_1 * P_2$
- $P(V_i = 2 | P_1 P_2) = P_1 * P_2$

One Possible Model (RAM)

- $Y_i = \beta_0 + \beta_1 A_i + \beta_2 (P_{1i} * P_{2i})$
 $+ \beta_3 A_{i,j,1} + \beta_4 A_{i,j,2} + \epsilon_i$
- $A_{i,j,k}$ is a (0,1) indicator variable indicating whether the i^{th} individual inherited k allele at the j^{th} locus from population D .

Proposed Model (SAT)

- $Y_i = \beta_0 + \beta_1 A_i + \beta_2 (P_{1i} * P_{2i}) + \beta_3 G_{i,j,1} + \beta_4 G_{i,j,2} + \varepsilon_i$
- $G_{i,j,k}$ is a (0,1) indicator variable indicating whether the i^{th} individual inherited k allele at the j^{th} locus of type M .

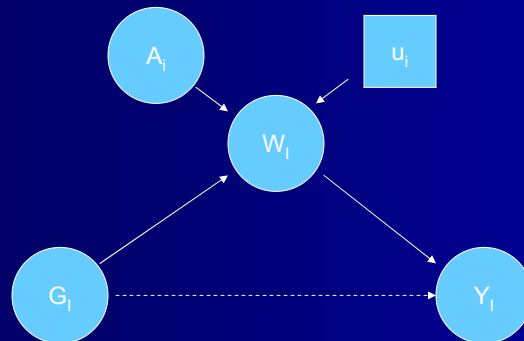
Further Issues

- Literature confuses the terms ancestry and admixture.
- Individual admixture (W_i) is the proportion of alleles in an individual's genome that an individual has from population D .
- Ancestry (A_i) is simply the midpoint of the parental ancestries.

Further Issues

- In fact $W_i = A_i + u_i$ where u_i is a combination of measurement error and biological effect.
- $E[W_i] = A_i$.
- All software (Structure, Admixmap) we are aware of provides estimates of individual admixture, which are error contaminated estimates of ancestry.

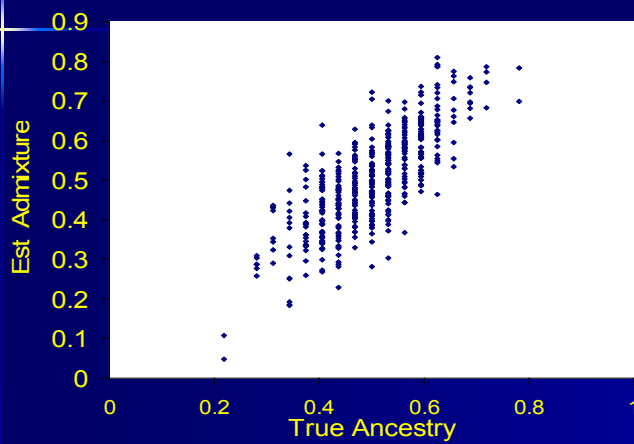
The Big Problem



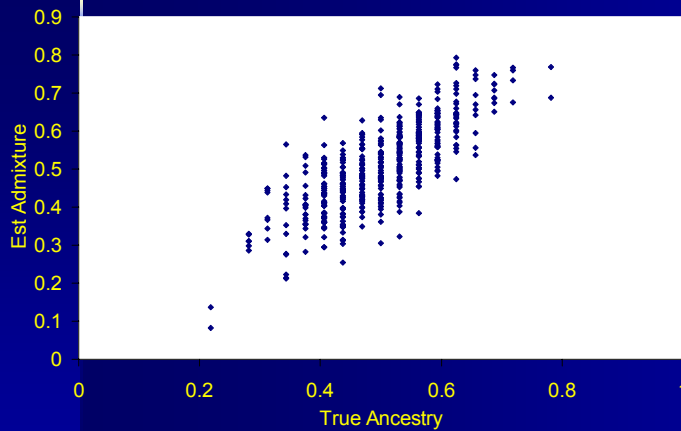
Relationship between True Admixture and True Ancestry, $r = .97$



Relationship between Maximum Likelihood Estimate of Admixture versus True Ancestry, $r = .78$



Relationship between Structure Estimate of Admixture
and True Ancestry $r = .78$

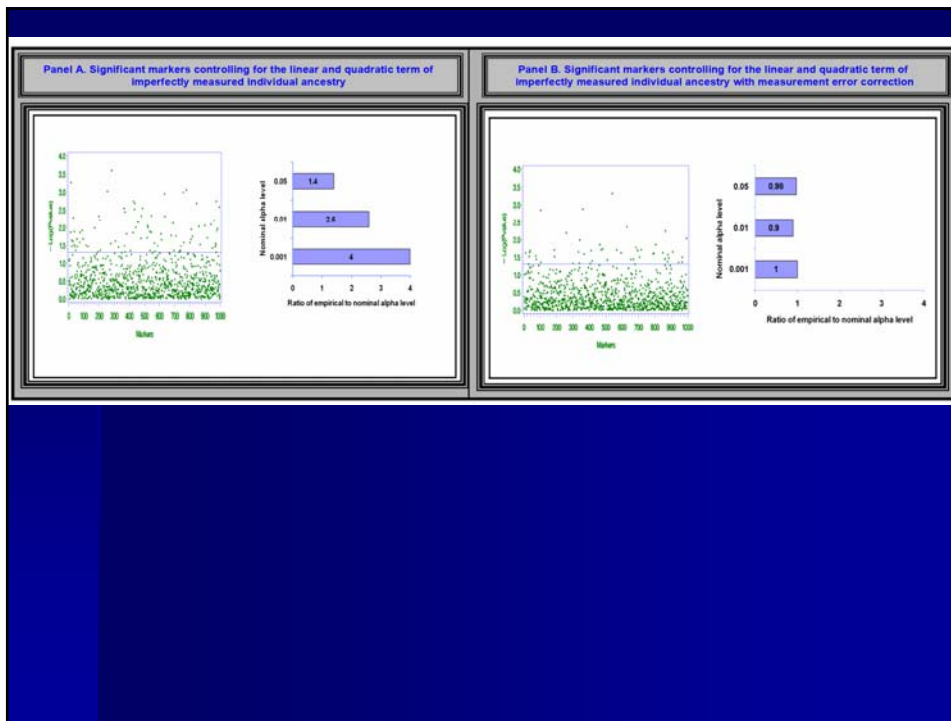


If we ignore the error in estimates...

- In regression, we assume all covariates are measured without error.
- The admixture estimates violate this assumption and create the possibility of residual confounding and biased estimation.
- We are currently investigating multiple models to address these issues.

Measurement errors

- Some knowledge about the distribution of the measurement error is required.
- There are several approaches to the measurement error problem:
 - 1) SIMEX
 - 2) Multiple Imputation
 - 3) Regression calibration
- Regression Calibration methods
 - 1) Regression calibration
 - 2) Moment reconstruction method
 - 3) Extended Regression calibration



What controls the measurement error for admixture?

- How many of markers will be used to estimate admixture?
- How informative the markers are with regard to ancestry?
- How many individuals within your study have pure ancestry from a founding population?

What controls the measurement error for admixture?

- How variable is admixture/ancestry within your sample?

Conclusion

- The research continues...