# On Genotyping Errors and Missing Data in Genome-Wide Association Studies

Kung-Yee Liang  Tom Louis  Ingo Ruczinski

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

May 5, 2008

# Specific Aims of our R01

1. Develop and evaluate new statistical methods to prioritize genes through proper ranking in genome-wide association (GWA) studies that address GxE interactions.

2. Develop and evaluate new statistical methods to localize causal genes as part of linkage and fine mapping studies while considering GxE interactions.

3. Develop and evaluate new statistical methods to identify higher order interactions between environmental variables and SNPs in candidate genes studies.

# Specific Aims (cont.)

4. Adapt existing and develop new statistical methods to address imprecise and missing environmental and genetic measurements.

5. Develop and disseminate efficient algorithms for GxE analyses, and apply these methods in several ongoing genetic studies of complex diseases.

# Addressing Imprecise and Missing Data

In genetic studies, genomic data can be missing, and genotyping errors can occur. The same holds true for environmental variables. In a stochastic sense, the structure of missing data and errors in variables are very similar. Missing data likelihood methods and multiple imputation are often practical and effective.

$\longrightarrow$ As part of this R01, we develop methods that incorporate knowledge about genotype uncertainty, for example scoring functions for tests of association that take the genotype likelihoods derived from genotype calling methods such as CRLMM into account.

# Missing Environmental Data

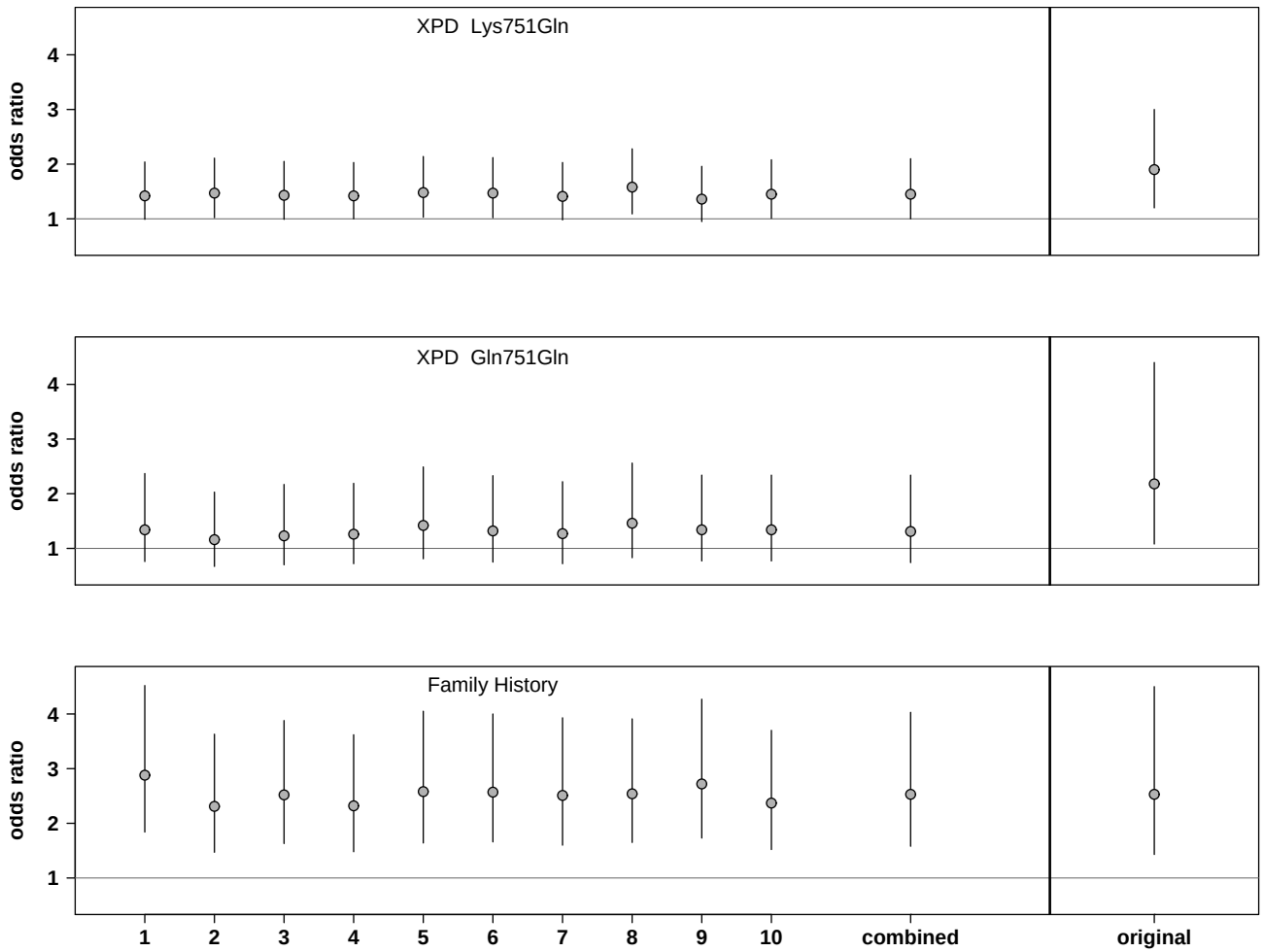
| | Number of Pairs | Odds Ratio | Confidence Interval |
|---|---|---|---|
| | | XPD Lys751Gln | |
| original data set | 202 | 1.90 | ( 1.20 – 3.00 ) |
| multiple imputations | 321 | 1.45 | ( 1.00 – 2.10 ) |
| | | XPD Gln751Gln | |
| original data set | 202 | 2.18 | ( 1.08 – 4.40 ) |
| multiple imputations | 321 | 1.31 | ( 0.74 – 2.34 ) |
| | | Positive Family History | |
| original data set | 202 | 2.53 | ( 1.43 – 4.50 ) |
| multiple imputations | 321 | 2.53 | ( 1.58 – 4.03 ) |

# Missing Environmental Data

| | Family History not complete | | | | Family History complete | | | |
|---|---|---|---|---|---|---|---|---|
| | AA | AC | CC | na | AA | AC | CC | na |
| | raw numbers | | | | | | | |
| case | 43 | 54 | 5 | 5 | 61 | 121 | 25 | 7 |
| control | 35 | 57 | 12 | 3 | 90 | 102 | 22 | 0 |
| | percentages | | | | | | | |
| case | 40.2 | 50.5 | 4.7 | 4.7 | 28.5 | 56.5 | 11.7 | 3.3 |
| control | 32.7 | 53.3 | 11.2 | 2.8 | 42.1 | 47.7 | 10.3 | 0.0 |

Reference: Brewster AM et al (2006). *Polymorphisms of the DNA Repair Genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): Relationship to Breast Cancer Risk and Familial Predisposition to Breast Cancer.* Breast Cancer Res Treat, 95(1): 73-80.

# Missing Environmental Data




The missing data were imputed using decision trees.

Reference: Dai J et al (2006). *Imputation Methods to Improve Inference in SNP Association Studies.* Genetic Epidemiology, 30(8): 690-702.

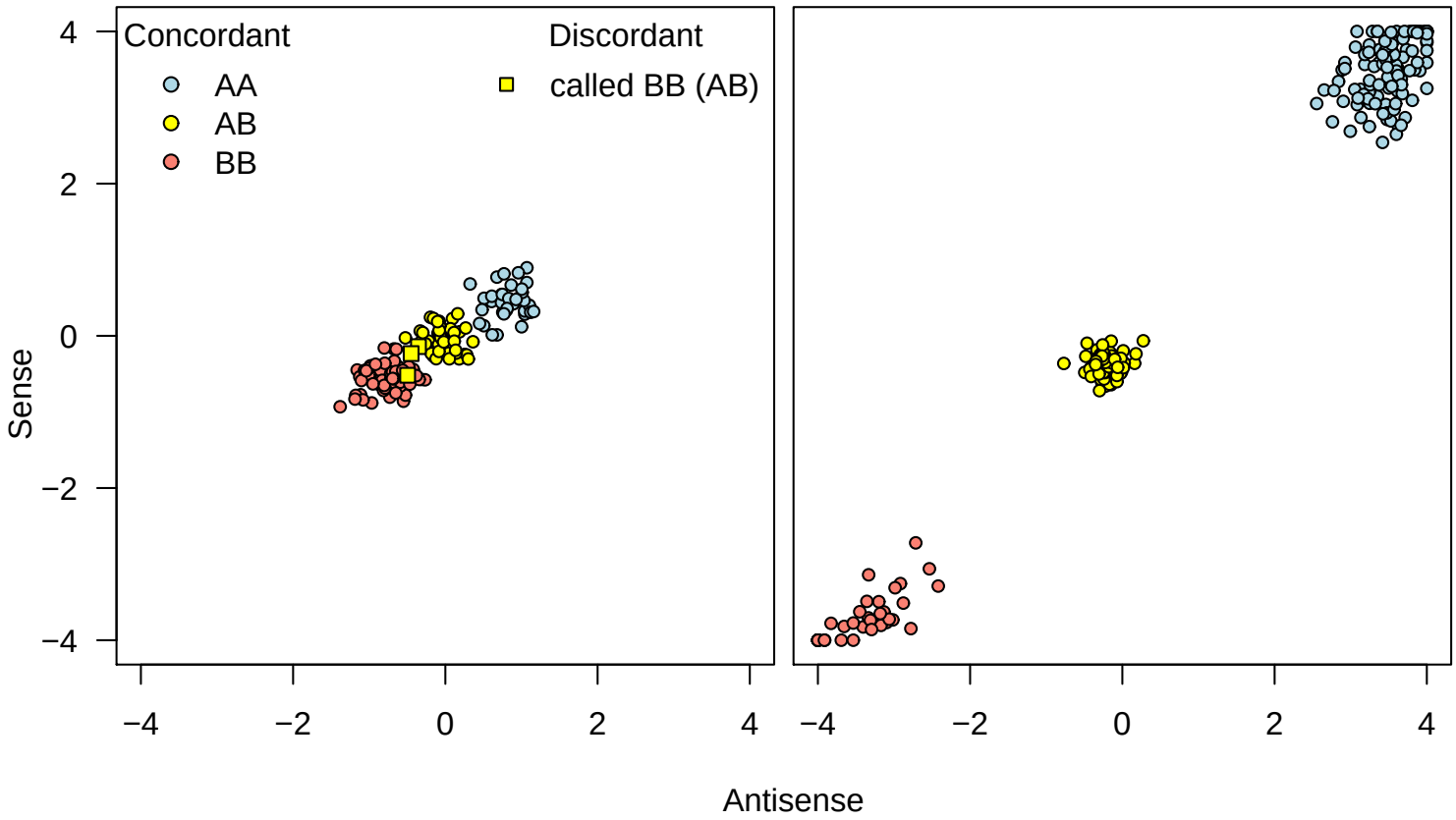
# Missingness at Random?

| Method | Confidence Threshold | Overall Call Rate | Hom Call Rate | Het Call Rate |
|---|---|---|---|---|
| **DM** | 0.26 | **94.16%** | 97.24% | 86.32% |
| **DM** | 0.33 | **95.96%** | 98.24% | 90.16% |
| **BRLMM** | 0.3 | **97.40%** | 97.40% | 97.75% |
| **BRLMM** | 0.4 | **98.27%** | 98.30% | 98.48% |
| **BRLMM** | 0.5 | **98.79%** | 98.82% | 98.93% |
| **BRLMM** | 0.6 | **99.15%** | 99.18% | 99.25% |

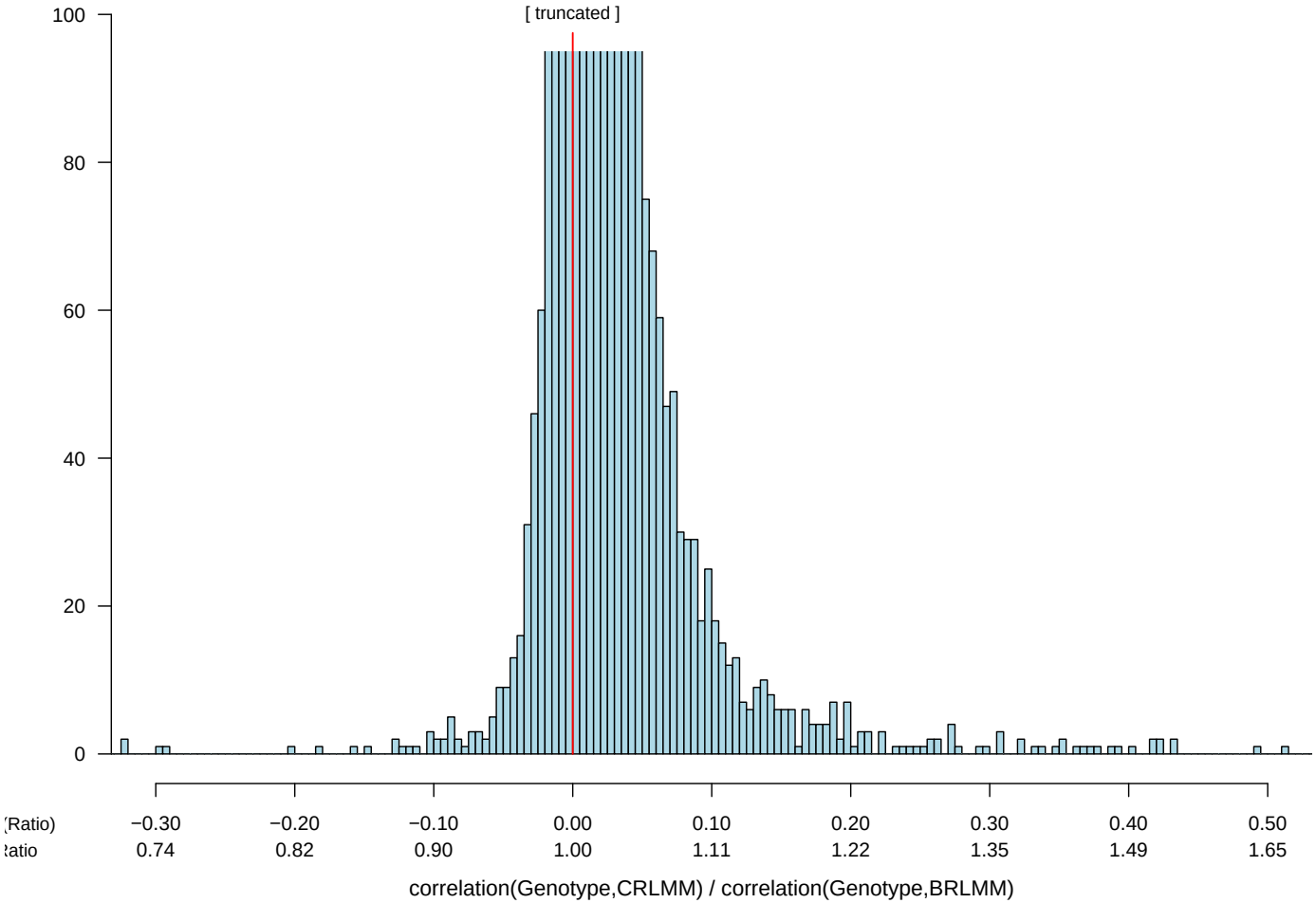
From the “white paper”,
`http://www.affymetrix.com/support/technical/product_updates/brlmm_algorithm.affx`

# Incorporating Genotype Uncertainty

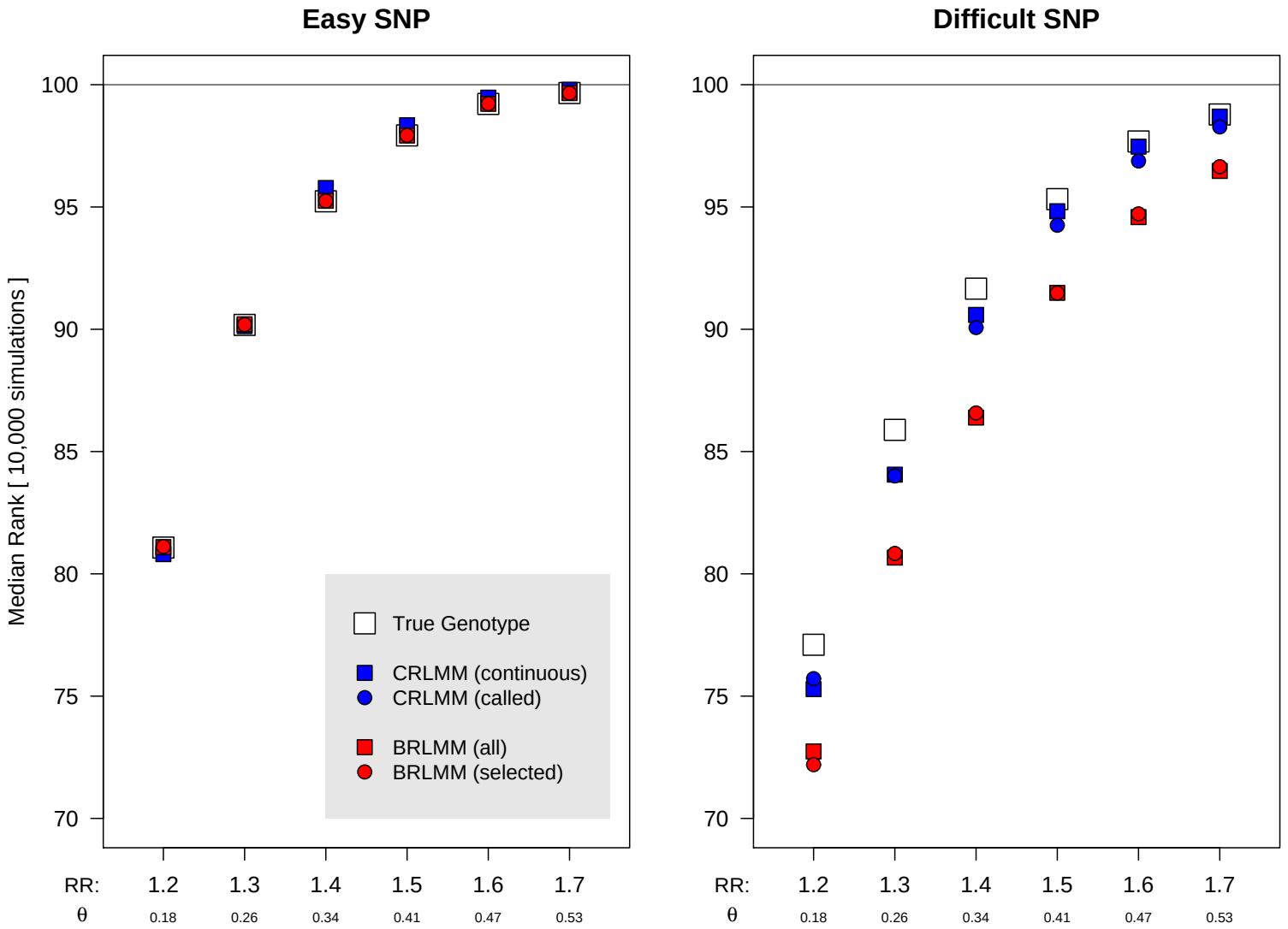
→ The confidence in genotype calls can differ substantially between SNPs!



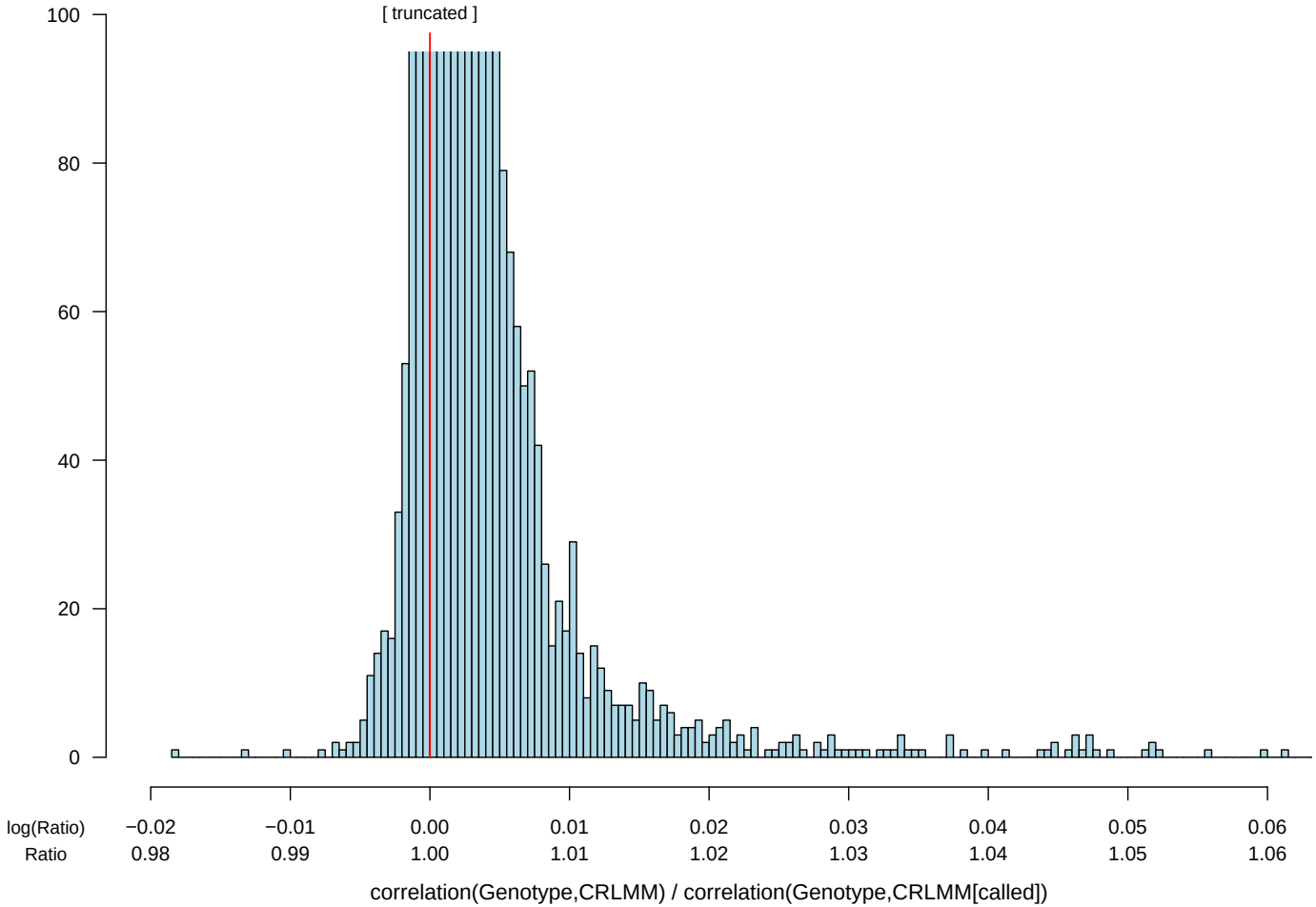
# Incorporating Genotype Uncertainty



# Incorporating Genotype Uncertainty



# Incorporating Genotype Uncertainty



# Prioritizing Genes in GWAs

- Based on ranking and selection approaches rather than on testing the null hypothesis of no GxE interaction.
- Requiring the specification and estimation of biologically relevant, SNP specific parameter (effect size) and specification of a joint prior distribution of these parameters.
- Ranking and selecting SNPs by combining estimates of effect sizes and their uncertainties.
- Incorporating knowledge obtained through linkage, fine mapping and candidate gene studies of GxE interaction.

# Acknowledgments


- Qing Li, Dani Fallin
- Rafael Irizarry, Benilton Carvalho
- Marvin Newhouse, Jiong Yang