Specific Aims

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

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

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

*Nonparametric Estimation of Gene-Environment Interactions*

Adina Crainiceanu, Kung-Yee Liang, Ciprian Crainiceanu
Nonparametric estimation of $G \times E$ interactions

- Estimated identity by descent (IBD) sharing
- 59 schizophrenia affected sib-pairs (ASP)
  - For each ASP
    - 28 micro-satellite markers on Chromosome 13
    - 64 cM have been genotyped
    - Onset age

**Observed and fitted means**

IBD sharing: observed (dotted lines), average (solid, red), estimated (solid, black).
\[
S(t_j|x_{i1}, x_{i2}) = C(x_{i1}, x_{i2}) \times \exp\{-0.04|t_j - \tau|\} + \epsilon_{ij}
\]

\[
\logit[\{1 + C(x_{i1}, x_{i2})\}/2] = m(x_{i1}, x_{i2})
\]

\[
\epsilon_{ij} \sim \text{Normal}(0, \sigma^2)
\]

- \((x_{i1}, x_{i2})\): schizophrenia onset age for \(i^{th}\) ASP.
- \(S(t_j|x_{i1}, x_{i2})\): estimated IBD sharing for \(i^{th}\) ASP at location \(t_j\).
- \([1 - \exp\{-0.02|t_j - \tau|\}]/2\): Haldane’s mapping function.
- \(C(x_{i1}, x_{i2})\): probability that an ASP share the same marker alleles at the disease locus.
- \(m(x_{i1}, x_{i2})\): unspecified function, modeled using penalized thin-plate splines.
Published results did not account for sampling variability. Pool (non-parametric) bootstrap results (green solid line): 140% longer CIs.
Conclusions

- Novel modeling approach + Bayesian posterior inference.
- Particularly useful for detecting G×E interactions.
- It is essential to incorporate sampling variability.
- Haldane’s function is not differentiable: no problem.
- Posterior distribution available.
- Exceedance probabilities can be calculated.
- The (incorrect) normal approximations are not used.
- Amount of smoothing is estimated using likelihood information.

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

*Detecting SNP-SNP Interactions in Case-Parent Trios*

Qing Li, Tom Louis, Dani Fallin, Ingo Ruczinski
Outline

- Extend the logic regression methodology (Ruczinski et al 2003) for trios with affected probands.

- Develop an R package to
  - impute missing genotypes, using a haplotype-based approach that employs mating tables;
  - simulate case-parent trios with disease risk dependent on SNP-SNP interactions.

- Carry out simulation studies to verify the validity of the methodology and the software.

- Analyze data from a study of Schizophrenia among Ashkenazi Jewish families.

Study details

- Goal: explore possible SNP-SNP interactions for Schizophrenia and Schizoaffective disorders.

- Study design:
  - Case-parent trios of Ashkenazi Jewish descents.
  - Diagnosis based on DSM IV.

- Target genetic markers:
  - Dense coverage of 64 candidate genes.
  - 375 SNPs on 11 chromosomes genotyped for 312 trios.
  - Original analysis: single marker TDT (Fallin et al 2005)
Logic regression

- Finds interactions in high dimensional model space:
  \[ g(\mathbb{E}[Y]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j \]
  The \( L_j \) are Boolean combinations of binary predictors.

- Example of a logic model:
  \[ \text{logit}(p) = \beta_0 + \beta_1 \times 1_{\{\text{SNP}_{17}^{D} \lor \text{SNP}_{33}^{R}\}}. \]

- Objective functions (deviance, likelihood, etc) assess model performance.

- Model search based on simulated annealing.

- Model selection via cross-validation / permutation tests.

---

Trio logic regression

- Pseudo controls are generated from the trio data, taking the LD block structure into account.

- Missing data are handled using haplotype-based imputation.

- The conditional logistic likelihood is used in logic regression to assess differences in cases and pseudo controls.
### Results

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>Z</th>
<th>“p-value”</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>$302^D$</td>
<td>0.51</td>
<td>-4.12</td>
</tr>
<tr>
<td></td>
<td>$166^D$</td>
<td>1.17</td>
<td>0.71</td>
</tr>
<tr>
<td>L</td>
<td>$302^D \land 166^D$</td>
<td>0.41</td>
<td>-4.89</td>
</tr>
<tr>
<td>A</td>
<td>$302^D$</td>
<td>0.51</td>
<td>-4.15</td>
</tr>
<tr>
<td></td>
<td>$166^D$</td>
<td>0.83</td>
<td>-0.86</td>
</tr>
<tr>
<td>A</td>
<td>$302^D$</td>
<td>1.21</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>$166^D$</td>
<td>1.18</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>$302^D : 166^D$</td>
<td>0.34</td>
<td>-2.87</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>OR</th>
<th>Z</th>
<th>“p-value”</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>302^D</td>
<td>0.51</td>
<td>-4.12 3.7 × 10^−5</td>
</tr>
<tr>
<td></td>
<td>166^D</td>
<td>1.17</td>
<td>0.71 0.480</td>
</tr>
<tr>
<td></td>
<td>148^D</td>
<td>1.54</td>
<td>1.70 0.088</td>
</tr>
<tr>
<td>L</td>
<td>302^D ∨ 166^D ∨ 148^D</td>
<td>3.14</td>
<td>5.67 1.5 × 10^−8</td>
</tr>
<tr>
<td>A</td>
<td>302^D</td>
<td>1.96</td>
<td>4.16 3.2 × 10^−5</td>
</tr>
<tr>
<td></td>
<td>166^D</td>
<td>1.20</td>
<td>0.83 0.410</td>
</tr>
<tr>
<td></td>
<td>148^D</td>
<td>1.54</td>
<td>1.70 0.088</td>
</tr>
</tbody>
</table>

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

**Association Tests that Accommodate Genotyping Errors**

Ingo Ruczinski, Qing Li, Benilton Carvalho, Dani Fallin, Rafa Irizarry, Tom Louis
Genotype uncertainty

- SNP 11905
- SNP 18233
- SNP 8548
- SNP 7

---

Genotype uncertainty

- Number of Mistakes in Lab A:
  1. 000 AA
  2. 000 AB
  3. 000 BB

- Number of Mistakes in Lab B:
  1. 005 AA
  2. 001 AB
  3. 000 BB

---

Crainiceanu  Liang  Louis  Ruczinski  Statistical Methods to Assess Interactions
Genotype call comparison

BRLMM-P  CRLMM  BirdSeed

Concordance

Birdseed / CRLMM concordance

Confidence score
Incorporating genotype uncertainty

Let \( Y \) be the binary response, \( d \) be the score assigned to a genotype (reflecting the genetic model), \( w \) is the *working* genotype probability, and \( t \) is the *true* genotype probability.

The score test statistic is

\[
Z(w, d) = \frac{\sum_i (w_i d_1 + w_i d_2)(Y_i - \hat{\pi})}{\sqrt{n \times \text{Var}(w_1 d_1 + w_2 d_2) \times \hat{\pi} (1 - \hat{\pi})}}
\]

Here

\[
\hat{\pi} = \bar{Y} = \frac{1}{n} \sum Y_i
\]

and

\[
\text{Var}(w_1 d_1 + w_2 d_2) = d_1^2 \times \text{Var}(w_1) + d_2^2 \times \text{Var}(w_2) + 2 \times d_1 d_2 \times \text{Cov}(w_1, w_2)
\]

The variance term measures statistical information

\[
n \times \text{Var}(w_1 d_1 + w_2 d_2) =
\]

\[
\sum_{j=1}^{2} d_j^2 \bar{w}_j (1 - \bar{w}_j) - n \times 2 \times d_1 d_2 \bar{w}_1 \bar{w}_2 -
\]

\[
\sum_{j=1}^{2} d_j^2 \sum_i w_{ij} (1 - w_{ij}) + 2 \times d_1 d_2 \sum_i w_{i1} w_{i2}
\]

- **Blue** is the variance when \( w_{ij} = 0 \) or \( 1 \).
- **Red** is the loss of information.
- Note that \( w_{ij} = 0 \) or \( 1 \) does not imply validity!
Incorporating genotype uncertainty

- For $\theta = \frac{\theta^*}{\sqrt{n}}$ (a local alternative), $Z \sim N(m(\cdot), 1)$

- For a logistic, for example, we have

$$E(Z(w, d) | \mu, \theta^*, d, w, t) \approx \theta^* \{\hat{\pi}(1 - \hat{\pi})\}^{1/2} \times \sqrt{\text{Var}(t_1 d_1 + t_2 d_2)} \times \text{Corr}({\{w_1 d_1 + w_2 d_2\}; \{t_1 d_1 + t_2 d_2\}})$$

- The correlation is the relative efficiency of using $w$ rather than $t$ ($t$ can be $g$, the true genotype).

- For a fixed sample size and effect size, the power of the score test is a function of the correlation of the fuzzy genotype call with the true genotype.

\[ \text{Correlation} \]

\[ \text{OR} = 1.00 \]

\[ \text{OR} = 1.25 \]

\[ \text{OR} = 1.50 \]

\[ \text{OR} = 1.75 \]

\[ \text{OR} = 2.00 \]

Crainiceanu Liang Louis Ruczinski
Statistical Methods to Assess Interactions
Incorporating genotype uncertainty

![Graph showing the correlation between genotype uncertainty and ratio](image)

Crainiceanu Liang Louis Ruczinski  Statistical Methods to Assess Interactions
Incorporating genotype uncertainty

![Graph showing correlation values](http://biostat.jhsph.edu/~iruczins/)

http://biostat.jhsph.edu/~iruczins/