Detecting Gene-Environment and Gene-Gene Interactions through Endophenotypes

Guan-Hua Huang and Shih-Kai Chu
National Chiao Tung University
TAIWAN

- Accumulating empirical evidences suggest that gene-environment and gene-gene interactions are major contributors to variation in complex diseases.

- Is there a rationale for modeling interactions in the absence of statistically significant marginal main effects?

Objective

- Identify SNPs that are weakly related to the disease by itself, but can have great impacts on the disease variability after combining with other SNPs and/or environmental effects.
- The endophenotype is closer to the underlying genotype than the phenotype in the course of disease's natural history.
- Select validate endophenotye to identify candidate SNPs with null marginal disease association for further interaction analysis.

Endophenotype

- Endophenotype provide a means for identifying the “downstream” traits of clinical phenotypes, as well as the “upstream” consequences of genes.

- Genotype → Endophenotype ⟶ Phenotype

Defining endophenotype

- Definition
  - $$f(E \mid G) = f(E) \Rightarrow f(P \mid G) = f(P)$$

    P: phenotype of interest

    E: candidate endophenotype

    G: underlying gene.

  - If the condition $f(P \mid E, G) = f(P \mid E)$ holds, then above definition holds.

Proportion of heritability explained by the endophenotype (PHE)

- Define

$$\mathrm{PHE} = 1 - \frac{h_{P|E}}{h_P}$$

$$(P_{ij} = \alpha + \gamma E_{ij} + \tau Z_{ij} + G_{ij} + \varepsilon_{ij})$$

  - $h_{P/E} =$ the heritability from the model using the candidate endophenotype (E) as one covariate
  - $h_P =$ the heritability from the model NOT using the candidate endophenotype as one covariate
  - the greater the PHE value, the more likely E is an endophenotype.
  - one-sided test

$$\begin{cases} H_0 : \mathrm{PHE} = 0 \\ H_1 : \mathrm{PHE} > 0 \end{cases}$$

Family dataset of GAW17 simulated data

- 697 individuals with 202 founders
- Genotypes contained 24487 SNPs, obtained from the 1000 Genomes Project.
- Genotypes were held fixed for all 200 replicates of the phenotype simulation.
- SEX, AGE, SMOKE, Q1, Q2, Q4, and AFFECTED were provided for each phenotype replicate.
  - AFFECTED – affected status of disease
  - Q1, Q2, and Q4 were quantitative traits related to the risk of disease
  - SMOKE – potential environmental causes of the disease

Family dataset of GAW17 simulated data

- AFFECTED was simulated using a liability threshold model and the top 30% of the distribution was declared affected.
- Q1, Q2, and Q4 were simulated as normally distributed phenotypes.
- All SNP effects are additive on liability scale or the quantitative trait.

Used data

- We used the data from the 1st replicate to develop the analytic procedure.
- Given the manner of the simulation, we assumed a lack of error in calling, and thus, did not perform initial quality assessment to exclude individuals and/or SNPs.

Endophenotype-based interaction detection

1. Select a validate endophenotype from Q1, Q2 and Q4
   - assessing the significance of PHE
2. Identify “endophenotypic SNPs”
   - SNPs that are significantly associated with the selected quantitative trait but only weakly related to the affected status
3. Form “candidate interactive SNPs” for interaction modeling
   - significant SNPs with the affected status, significant SNPs with the endophenotype and endophenotypic SNPs

Endophenotypic SNPs

- Perform FBAT to rank SNPs in their statistical significance to the affected status and the selected endophenotype, respectively.
- FBAT was done for one SNP at a time with the gene-environment interaction modeling:

$$\alpha(\mathrm{SNP}) + \beta(\mathrm{SMOKE}) + \gamma(\mathrm{SNP} \times \mathrm{SMOKE})$$

$$H_0 : \alpha = 0 \text{ and } \gamma = 0$$

- Identify SNPs that were both in the top 50 significant SNPs with the endophenotype and in the top 100 significant SNPs with the affected status

Interaction modeling

- MDR method was applied to candidate interactive SNPs and SMOKE for detecting possible gene-environment and gene-gene interactions.

Results

- Q1, Q2 and Q4 were significantly associated with AFFECTED after adjusting for SEX and AGE.
- PHE analysis

	PHE	S.E.	P-value
Q1	0.49	0.14	0.00022
Q2	0.06	0.12	0.29
Q4	-0.15	0.18	0.80

Endophenotypic SNPs

- Analyze 5753 SNPs with 10 or more informative families
- AFFECTED
  - None of the SNPs was significant after multiple testing adjustment (pFDR≤0.05)
- Q1
  - C6S2981 was significant under pFDR≤0.05
- Endophenotypic SNPs:
  - C22S1222, C6S2367, C11S164, C12S4103, C12S4082, C19S4377, C6S2366, C11S3810, C17S1350 and C4S1220

Interaction detection

Model size	Top model	Testing accuracy	*p*-value with AFFECTED
1	SMOKE	0.5681	0.98
2	C11S164, SMOKE	0.5811	0.275
3	C6S2981, C11S164, SMOKE	0.5951	0.0365
4	C4S1220, C6S2366, C12S4082, C22S1222	0.5418	0.14
5	C4S1220, C6S2367, C12S4103, C22S1222, SMOKE	0.522	0.321
6	C4S1220, C6S2367, C12S4103, C19S4377, C22S1222, SMOKE	0.5654	0.0175
7	C4S1220, C6S2366, C6S2367, C12S4103, C19S4377, C22S1222, SMOKE	0.5678	0.0305
8	C4S1220, C6S2366, C6S2367, C11S164, C12S4103, C19S4377, C22S1222, SMOKE	0.5596	0.0735
9	C4S1220, C6S2366, C6S2367, C11S164, C12S4082, C12S4103, C19S4377, C22S1222, SMOKE	0.5538	0.1465
10	C4S1220, C6S2366, C6S2367, C6S2981, C11S164, C12S4103, C17S1350, C19S4377, C22S1222, SMOKE	0.5586	0.0895
11	C4S1220, C6S2366, C6S2367, C6S2981, C11S164, C12S4082, C12S4103, C17S1350, C19S4377, C22S1222, SMOKE	0.5686	0.0295
12	C4S1220, C6S2366, C6S2367, C6S2981, C11S164, C11S3810, C12S4082, C12S4103, C17S1350, C19S4377, C22S1222, SMOKE	0.5679	0.017

Apply to Q2 and Q4

- Q2
  - None of the SNPs was significant after multiple testing adjustment (pFDR$\leq$0.05)
- Q4
  - None of the SNPs was significant after multiple testing adjustment (pFDR$\leq$0.05)
- Endophenotype-based interaction detection
  - Both Q2 and Q4 did not result in any significant SNP-SMOKE and SNP-SNP interactions

Problems with rare variants

- GAW17 simulated data includes many rare SNPs with a minor allele frequency (MAF) smaller than 0.05.
- Current statistical strategies for detecting disease associated variants may lose power when applied to rare variants.
- In fact, C6S2981 in gene VEGFA was the only causal SNP (provided in the “Answers”) detected by FBAT.

Analyze rare variants

- Collapse multiple rare variants within a gene to form a combined variant
  - can enrich the signal of association
- $$R_{ij} = \begin{cases} 1 & \text{the miner allele was obssrved for any of the rare SNPs} \\ 0 & \text{otherwise} \end{cases}$$
- The variance component model was used to obtain its association with AFFECTED, Q1, Q2, and Q4

Analyze rare variants

- Excluded SNPs (MAF=0): 10703
- Common SNPs (MAF≥0.05): 3074
- Rare SNPs (MAF<0.05): 10710
  - rare SNPs were then collapsed to form 2575 combined variants.
- AFFECTED
  - None of the combined variants was significant after multiple testing adjustment (pFDR≤0.05)
- Q1
  - VEGFC, VEGFA, PSG1, KIT, LOC728326, SMYD2, and NR2C2AP were significant under pFDR≤0.05.

Analyze rare variants

- Two causal genes for Q1 (VEGFC and VEGFA) were identified, but none were identified for AFFECTED.
- It appears that the collapsing approach does not work well in family-based association tests.
- Apply MDR to candidate interactive variants formed from common SNPs and combine variants
  - no significant interaction was identified.