Optimal screening for promising genes in two-stage designs


Beatrijs Moerkerke and Els Goetghebeur
Beatrijs.Moerkerke@UGent.be

Department of Applied Mathematics and Computer Science
Ghent University

http://www.stat.ucl.ac.be/IAP/PhaseVI/
http://www.cvstat.ugent.be/
Problem setting

Scope:

1. Detecting differentially expressed genes
   
   To identify genes whose patterns of expression differ according to phenotype or experimental condition

2. Screening SNP’s in whole genome scans in a case-control setting
   
   Human studies: to find SNP’s related to disease

3. Marker assisted selection in plant breeding
   
   - Observe marker genotype in a sample of plants
   - Find (a set of) genetic markers which predict phenotype of interest
   - Cross plants with good markers
Genetic markers: genes/SNP’s/markers

Aim: to develop cost-efficient designs for screening genetic markers

We define cost in terms of the number of marker evaluations that are needed

Outline

- Motivation for a new alternative-based procedure
- Balanced testing in one-stage designs
- Balanced testing in two-stage designs
Aim and outline

FIRST STAGE

Accept $H_{0j}$  Go to second stage  Accept $H_{A_j}$

$C_{opt,j} - \varepsilon$  $C_{opt,j}$  $C_{opt,j} + \varepsilon$

$T_{j,n_1}$

$T_{j,n_1}$ in grey zone

SECOND STAGE

Accept $H_{0j}$  Accept $H_{A_j}$

$C_{opt,j}$

$T_{j,n_m}$

Optimal screening for promising genes in two-stage designs – p. 4/23
Notation

- $\Delta_j$ denotes the population contrast of interest for marker $j$ ($j = 1, \ldots, m$)
- $\Delta^0$ is the value the contrast takes under the null hypothesis of no association with the phenotype, assume further $\Delta^0 = 0$
- $\Delta^1$ is the target value for contrast $\Delta_j$

Not mere non-null effects are of interest but we focus on markers with an effect of at least $\Delta^1$.

Interested in testing

$H_{0j} : \Delta_j = 0$ against $H_{A_j} : \Delta_j = \Delta^1 > 0$

for each marker $j$
Statistical challenges

To detect the ‘true’ associations among MANY observed associations

Avoid

- Further investigation of ‘untrue’ associations = wasted energy
- Ignoring ‘true’ associations = even worse

Multiple testing problem

$m$ genetic markers/genes = performing $m$ tests

Classical procedures: focus on type I error rate
The need for more powerful methods

Statistical significance $\leftrightarrow$ biological relevance

For detecting important alternatives: $p$-values not sufficient
Only measures of evidence against the null

Add summary of evidence against alternative $\Delta^1$ of interest

Take into account
- magnitude of observed effect
- magnitude of desired effect
- precision

What effect $\Delta^1$, if present, do we not want to miss?

Incorporate this directly in decision criterion

(Balanced test)
A measure for (im)potence: $p_1$

$\Delta^1$ Observed fold change

$p_0 \downarrow$: more evidence **against** the null

$p_1 \downarrow$: more evidence **against** the alternative $\Delta^1$
Balanced testing in one-stage designs

- \((p_0, p_1)\)-values: basis for a descriptive and a formal decision tool

- Balance \(p_{0j}\) and \(p_{1j}\) by optimizing a user-specific general expected gain:

\[
G_j = A_j \times P(\text{Accept } H_{0j} | H_{0j}) + B_j \times P(\text{Accept } H_{A_j} | H_{A_j})
\]

- Decision criterion determined by optimizing \(G_j\) (decision-theoretic)

- \(A_j, B_j\): user- (and marker-) specific weights

- Weight ratio \(A_j / B_j > (\leq) 1\): more weight on a correct decision under the null (alternative)
Balanced testing in one-stage designs

\[ G_j = A_j \times P(\text{Accept } H_{0j} | H_{0j}) + B_j \times P(\text{Accept } H_{A_j} | H_{A_j}) \]

1-sided test:

\( H_{0j} : \Delta_j = 0 \) and \( H_{A_j} : \Delta_j = \Delta^1 > 0 \)

maximize \( G_j \rightarrow c_{\text{opt},j} \): optimal cutoff for the test statistic \( T_j \)

Assume normally distributed test statistic: \( T_j \sim N(\sqrt{n}\Delta_j / \sigma_{D_j}, 1) \)

\[
c_{\text{opt},j} = \ln\left(\frac{A_j}{B_j}\right) + 0.5 \times \left(\frac{\sqrt{n}\Delta^1 / \sigma_{D_j}}{\sqrt{n}\Delta^1 / \sigma_{D_j}}\right)^2
\]

(Even if \( A_j = A, B_j = B \rightarrow \) cutoff marker-specific!)
Balanced testing in two-stage designs

Measurements in genetic research: highly expensive
→ need for cost-reducing designs

Basic idea of two-stage designs:

\[ m \] genetic markers need to be tested with a fixed budget
→ \( n_E \) observations/individuals - cost: \( m \times n_E \) measurements

**Step 1:** genotype \( m \) markers on \( n_1 \) individuals

**Step 2:** genotype \( m_2 \) markers for which results are inconclusive
on \( n_2 \) extra individuals

with maximum sample size \( n_{\text{max}} = n_1 + n_2 \) such that expected cost per marker equals \( n_E \)
Balanced testing in two-stage designs

\[ G_j = A_j \times P(\text{Accept } H_{0j} | H_{0j}) + B_j \times P(\text{Accept } H_{A_j} | H_{A_j}) \]

\( T_{j,n_k} \): test statistic for marker \( j \) on data gathered in stage \( k \) only

Assume \( T_{j,n_k} \sim N \left( \sqrt{n_k} \Delta_j / \sigma_{Dj}^{(k)}, 1 \right) \)

Test statistic \( T_{j,n_{\text{max}}} \) for combining data from both stages:

\[ T_{j,n_{\text{max}}} = \frac{\sqrt{n_1}}{\sqrt{n_{\text{max}}}} T_{j,n_1} + \frac{\sqrt{n_2}}{\sqrt{n_{\text{max}}}} T_{j,n_2} \]
Balanced testing in two-stage designs

**FIRST STAGE**

- Accept $H_{0j}$
- Go to second stage
- Accept $H_{1j}$

$C_{opt,j} - \varepsilon$  $C_{opt,j}$  $C_{opt,j} + \varepsilon$  $T_{j,n_1}$

$T_{j,n_1}$ in grey zone

**SECOND STAGE**

- Accept $H_{0j}$
- Accept $H_{1j}$

$C^{(2)}_{opt,j}$  $T_{j,n_{max}}$
Balanced testing in two-stage designs

Note:
Marker-specific decision criterion and fixed budget per marker → different $n_{\text{max}}$ per marker: $n_{\text{max},j}$

Difficult to specify marker specific $\sigma_{D,j}$-values at design stage

When possible: added degree of freedom allows to gain efficiency

When marker-specific designs not attainable, one can choose to:

1. take the minimum over all $n_{\text{max},j}$ ($j = 1, \ldots, m$)
2. prioritize some markers
Balanced testing in two-stage designs

\[ G_j = A_j \times P(\text{Accept } H_{0j}|H_{0j}) + B_j \times P(\text{Accept } H_{A_j}|H_{A_j}) \]

Optimal cutoff second stage when \( n_{1j} = n_{2j} = n_{\text{max},j}/2 \):

\[ c^{(2)}_{\text{opt},j} = \ln\left(\frac{A_j}{B_j}\right) + \left(\frac{\sqrt{\frac{n_{\text{max},j}}{2}} \Delta^1/\sigma_{D_j}}{\sqrt{2}}\right)^2 \]

depends only on length of grey zone \( \varepsilon_j \) through \( n_{\text{max},j} \)

Length of grey zone \( \varepsilon_j \)?

\[ \varepsilon_j \to 0 : \quad \text{Never go to second stage: } n_{\text{max},j}/2 = n_E \]
\[ \varepsilon_j \to +\infty : \quad \text{Always go to second stage: } n_{\text{max},j} = n_E \]
 Balanced testing in two-stage designs

I. What is optimal grey zone ($\varepsilon_j$)?

II. 2-stage versus 1-stage comparison?

An example:

$H_{0j} : \Delta_j = 0 \leftrightarrow H_{Aj} : \Delta_j = 0.3$

with $P(H_{0j}) = 0.9$ and $P(H_{Aj}) = 0.1$

$A_j/B_j = 4$

(null 4 times more important than alternative in optimization)

Budget: $n_E = 80$

$(\varepsilon_j, n_{\text{max},j}) \rightarrow P(\text{Binary decision in stage I}) \rightarrow n_{\text{max},j}$

$\Rightarrow$ Numerical optimization of two-stage designs
For $n_E = 80$: optimal $\varepsilon_j = 0.904$ with $n_{\text{max},j} = 132$
What is expected cost of a two-stage design with the same expected gain as one-stage design?
\textbf{Balanced Testing in Two-stage designs}

Algorithm to account for unknown \( P(H_{0j}) \) and \( P(H_{Aj}) \)

Fix \( n_E \), let \( P(H_{0j}) \) vary from 0 to 1, \( P(H_{Aj}) = 1 - P(H_{0j}) \)

Result: range of optimal prevalence-specific two-stage designs

\rightarrow \text{choose minimum/average/maximum } n_{\text{max},j}

Simulations

More complex situations analyzed with algorithm built on two-stage principle

- Distribution of true underlying effects to depict realistic scenarios
- Possibility of different variance structure among markers
- Correlation among markers
Balanced Testing in Two-stage designs

\[ m = 3000 \text{ markers with } A/B = 4, \text{ 1000 simulations} \]

\[ P(\Delta_j = 0) = \frac{989}{1000}, \quad P(\Delta_j = 0.25) = \frac{10}{1000}, \quad P(\Delta_j = 0.5) = \frac{1}{1000} \]

Pairwise correlation coefficient \( \rho \)

\[ \Delta^1 = 0.4 \]

All markers with \( \Delta_j \geq \Delta^1 \) are of interest

→ achieved gain in simulations is evaluated accordingly
<table>
<thead>
<tr>
<th>$n_E$</th>
<th>$\Delta^1$</th>
<th>Average gain</th>
<th>Relative increase in gain (%)</th>
<th>Relative cost reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0.4</td>
<td>0.964 (50)</td>
<td>5.48</td>
<td>37.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.975 (65)</td>
<td>8.97</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.979 (73)</td>
<td>7.48</td>
<td>37.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.963 (50)</td>
<td>5.48</td>
<td>31.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.974 (59)</td>
<td>9.47</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.979 (71)</td>
<td>7.98</td>
<td>35.3</td>
</tr>
<tr>
<td>(ρ = 0)</td>
<td>0.4</td>
<td>0.964 (50)</td>
<td>5.48</td>
<td>37.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.975 (65)</td>
<td>8.97</td>
<td>37.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.979 (73)</td>
<td>7.48</td>
<td>37.2</td>
</tr>
<tr>
<td>(ρ = 0.4)</td>
<td>0.4</td>
<td>0.963 (50)</td>
<td>5.48</td>
<td>31.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.974 (59)</td>
<td>9.47</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.979 (71)</td>
<td>7.98</td>
<td>35.3</td>
</tr>
</tbody>
</table>
Balanced testing in two-stage designs

Main conclusions:

- Two-stage designs: better balance between true negatives and true positives for same cost

- Correlations do not change results on average, higher instability

- Comparison with an FDR-controlling two-stage design (Zehetmayer et al., 2005):
  More markers need to be selected with FDR procedure to select same amount of biologically relevant markers
Discussion

- Problem of screening genetic markers from perspective of important alternatives
  - Clinical trials: focus on classical $p$-values and type I error
  - Many other applications: null and alternative deserve better balance
  - Careful reflection of analysis goals
- Balanced testing in 2-stage designs: expected cost for 2-stage designs is lower
- Two-stage designs with equal sample sizes in both stages
  - Proportion of data used in first stage is also a design parameter
  - Extra level of complexity provides only small increase in gain.

Paper of this talk to appear in Biostatistics
http://biostatistics.oxfordjournals.org/cgi/content/abstract/9/4/700
References


