Epistasis Testing on the Cloud

Limsoon Wong
5 October 2011

Plan

• Empirical Comparison of Recent Epistasis Analysis Methods

• Epistasis Analysis on the Cloud
Acknowledgements

- Wang Yue
- Feng Mengling
- Liu Guimei

- Wang Zhengkui
- Tan Kian Lee
- Divyakant Agrawal

- A*STAR SERC PSF grant

Empirical Comparison of Recent Epistasis Analysis Methods
Epistasis Analysis Methods

- BOOST (B), TEAM (T), SNPRuler (SR), SNPHarvester (SH), Screen and Clean (SC)

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>T</th>
<th>SR</th>
<th>SH</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhaustive Search</td>
<td>×</td>
<td></td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Logit Model Assumed</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multi-Stage</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>Permutation Test</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Bonferroni correction</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Programming language</td>
<td>C</td>
<td>C++</td>
<td>Java</td>
<td>Java</td>
<td>R</td>
</tr>
</tbody>
</table>

- BOOST detected most of the ground-truth epistatic interactions
- TEAM detected most of the ground-truth epistatic interactions
Type-1 Error Rate

- **Type-1 error rate**
  - Proportion of datasets that a method reports existence of significant epistatic interactions, out of the 1,000 datasets in which no epistatic interactions are embedded

<table>
<thead>
<tr>
<th>Method</th>
<th>Type-1 Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEAM</td>
<td>0.018</td>
</tr>
<tr>
<td>BOOST</td>
<td>0.065</td>
</tr>
<tr>
<td>SNPRuler</td>
<td>0.003</td>
</tr>
<tr>
<td>SNPHarvester</td>
<td>0.003</td>
</tr>
<tr>
<td>Screen and Clean</td>
<td>0.860</td>
</tr>
</tbody>
</table>

Completeness

- **Non-exhaustive methods wrongly pruned many top significant epistatic interactions**
Performance

Table 3. Running time comparison of the five methods. SR is short for SNPRuler, SH is SNPHarvester, SC is Screen and Clean.

<table>
<thead>
<tr>
<th># SNPs</th>
<th>TEAM</th>
<th>BOOST</th>
<th>SR</th>
<th>SH</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>58.23s</td>
<td>0.162s</td>
<td>2.43s</td>
<td>2.29s</td>
<td>7.389s</td>
</tr>
<tr>
<td>1000</td>
<td>353s.20s</td>
<td>2.47s</td>
<td>21.73s</td>
<td>22.33s</td>
<td>55.475s</td>
</tr>
<tr>
<td>10000</td>
<td>7406.29s</td>
<td>156.16s</td>
<td>1097.65s</td>
<td>224.24s</td>
<td>626.96s</td>
</tr>
<tr>
<td>100000</td>
<td>~36 days</td>
<td>15010.42s</td>
<td>NA</td>
<td>6616.65s</td>
<td>5858.34s</td>
</tr>
</tbody>
</table>

Remarks

- Methods that are more “exhaustive” in testing of SNP pairs (i.e., BOOST, TEAM)
  - Are more sensitive
  - Have good type-1 error rate
  - But take long time to compute

- Stick to exhaustive methods
- Need large-scale parallelism
  ⇒ Cloud computing!
Cloud Computing: What

- Cloud computing is the delivery of computing as a service, whereby shared resources, software and information are provided to computers and other devices as a utility over the Internet.
Cloud Computing: Why

- **Reduced Cost**
  - Cloud computing is paid incrementally, saving money

- **Increased Storage**
  - Can store more data than on private computer systems

- **Highly Automated**
  - No need to worry about keeping software up to date

- **Flexibility**
  - More flexible than past computing methods

- **More Mobility**
  - Access info from anywhere

- **Allows IT to Shift Focus**
  - No worry about server updates and other computing issues

  ⇒ Free to concentrate on innovation

Cloud Computing: How

• eCEO uses bit string data representation
Cloud Computing: How

- eCEO uses
  - Map-reduce processing framework
  - Load-balanced distribution of SNP pairs to 2M reducers


eCEO Performance

It takes 1.2 years to do pairwise epistasis testing of 500 000 SNPs using a serial program on a 2.66 GHz single processor w/o parallel processing. Our eCEO can do this in < 9h using only a 43-node cluster
Any Question?