Imputation and GWAS

Chien-Hsiun Chen
National Center of Genome Medicine
Institute of Biomedical Sciences
Academia Sinica, Taipei, Taiwan
2013-12-06
at
NCGM 使用者說明會

Topics
• Choose an imputation program
• Use proper references
• Clean genotype data
• Evaluate imputed data
• Perform GWAS with imputed data

Acknowledgements
• National Center of Genome Medicine
• Academia Sinica GMM study

Imputation

http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
Choose an imputation program

• IMPUTE2
• MACH (Minimac)
• BEAGLE
• Comparing BEAGLE, IMPUTE2, and Minimac Imputation Methods

IMPUTE version 2 (IMPUTE2)

• IMPUTE version 2 (also known as IMPUTE2) is a genotype imputation and haplotype phasing program based on ideas from Howie et al. 2009. (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
• SNPTEST is a program for the analysis of single SNP association in genome-wide studies. The program is designed to work seamlessly with the output of our genotype imputation software IMPUTE. (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html)

MACH

• MACH 1.0 is a Markov Chain based haplotype. It can resolve long haplotypes or infer missing genotypes in samples of unrelated individuals. (http://www.sph.umich.edu/csg/abecasis/MACH/index.html)
• MaCH was first used to imputed missing genotypes in the FUSION genome-wide association study (Scott et al, Science, 2007)
• With Mach2dat, users can directly use MACH output to assess association for quantitative and qualitative traits in unrelated. (http://www.unc.edu/~yunmli/software.html)

BEAGLE Genetic Analysis Software Package

• BEAGLE is a state of the art software package for analysis of large-scale genetic data sets with hundreds of thousands of markers genotyped on thousands of samples. BEAGLE can
  – phase genotype data (i.e. infer haplotypes) for unrelated individuals, parent-offspring pairs, and parent-offspring trios.
  – infer sporadic missing genotype data.
  – impute ungenotyped markers that have been genotyped in a reference panel. (http://faculty.washington.edu/browning/beagle/beagle.html)
• Beagle can also be used in conjunction with PRESTO, a program for fast and flexible permutation testing. PRESTO can compute empirical distributions of order statistics, analyze stratified data, and determine significance levels for one-stage and two-stage genetic association studies. (http://faculty.washington.edu/browning/presto/presto.html)
Comparing BEAGLE, IMPUTE2, and Minimac Imputation Methods ([http://blog.goldenhelix.com/?p=1911)]

- Imputation was performed both with and without pre-phasing the sample data with BEAGLE and IMPUTE2. Minimac is an implementation of the MaCH method that utilizes pre-phasing.

<table>
<thead>
<tr>
<th>Software</th>
<th>Total Compute Time</th>
<th>Mean SNP Concordance</th>
<th>Total # SNPs</th>
<th># High Quality SNPs</th>
<th>% High Quality Imputed</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMPUTE2</td>
<td>23 hours</td>
<td>99.98%</td>
<td>668,180</td>
<td>620,792</td>
<td>92.9%</td>
</tr>
<tr>
<td>BEAGLE</td>
<td>213 hours</td>
<td>98.43%</td>
<td>484,023</td>
<td>320,991</td>
<td>66.3%</td>
</tr>
<tr>
<td>IMPUTE2 with Pre-phasing</td>
<td>8 hours</td>
<td>99.92%</td>
<td>668,180</td>
<td>297,196</td>
<td>44.5%</td>
</tr>
<tr>
<td>BEAGLE with Pre-phasing</td>
<td>24 hours</td>
<td>98.05%</td>
<td>484,023</td>
<td>293,890</td>
<td>60.7%</td>
</tr>
<tr>
<td>Minimac</td>
<td>18 hours</td>
<td>96.25%</td>
<td>667,870</td>
<td>450,790</td>
<td>67.5%</td>
</tr>
</tbody>
</table>

Pre-phase

- The basic idea is to "pre-phase" your study genotypes to produce best-guess haplotypes, then impute into these estimated haplotypes in a separate program run.
- By contrast, the original IMPUTE2 method integrates over the unknown phase of your study data during the course of an imputation analysis. Pre-phasing leads to a small loss of accuracy since the estimation uncertainty in the study haplotypes is ignored, but this allows for very fast imputation.

[http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#prephasing]

Use proper references

- Reference panels
  - HapMap II
  - HapMap III
  - 1000 Genome Projects
- Larger reference panels improve the power and resolution of imputation-based association mapping, but they also increase the computational burden of imputation.
- Ethnicity-specific references
  - HapMap II: CHB and JPT
  - HapMap III: CHB, CHD and JPT
  - 1000 Genome Projects: CHB, CHS, JPT

Download Reference Data (impute2)

<table>
<thead>
<tr>
<th>Link to download page</th>
<th>NCBI build</th>
<th>Allele set release date</th>
<th>Release status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Genomes Pilot + HapMap II</td>
<td>b37</td>
<td>Mar 2012</td>
<td>Includes chrX, updated 24 Aug 2012</td>
</tr>
<tr>
<td>1000 Genomes Pilot + HapMap II</td>
<td>b37</td>
<td>Jul 2011</td>
<td>Includes chrX, updated 28 Apr 2012</td>
</tr>
<tr>
<td>1000 Genomes Pilot</td>
<td>b37</td>
<td>Dec 2010</td>
<td>Includes chrX</td>
</tr>
<tr>
<td>1000 Genomes Pilot + HapMap II</td>
<td>b36</td>
<td>Apr 2010 / Feb 2009</td>
<td>Includes chrX</td>
</tr>
<tr>
<td>1000 Genomes Pilot</td>
<td>b36</td>
<td>Apr 2010</td>
<td>Includes chrX</td>
</tr>
<tr>
<td>HapMap II reference files</td>
<td>b36</td>
<td>Dec 2008</td>
<td>Includes chrX</td>
</tr>
<tr>
<td>HapMap II reference files</td>
<td>b35</td>
<td>Apr 2008</td>
<td>Includes chrX</td>
</tr>
<tr>
<td>HapMap II reference files</td>
<td>b35</td>
<td>Apr 2008</td>
<td>Includes chrX</td>
</tr>
<tr>
<td>HapMap II reference files</td>
<td>b35</td>
<td>Apr 2008</td>
<td>Includes chrX</td>
</tr>
</tbody>
</table>

[http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference]
Clean genotype data

- Genotype data
  - Affymetrix SNP6.0, CHB and TWB Axiom array,
  - Illumina 550K, 1M, 2.5M
- SNP quality control (call rate, HWE,...)
- Batch effect
  - Subjects were genotyped at different time points
  - Cases and controls were genotyped at different time points

Evaluate imputed data

- Imputation Concordance Table
  - For this analysis, IMPUTE2 masks the genotypes of one variant at a time in the study data (Panel 2), then imputes the masked genotypes with information from the reference data and nearby study variants.
  - The imputed genotypes are then compared with the original genotypes to evaluate the quality of the imputation.

Perform GWAS with imputed data

- Dosage data
- Make hard genotype calls by applying a threshold (default = 0.9) to the maximum value in each input probability triple.
  - a genotype with P(G=0,1,2) = (0.03, 0.95, 0.02) would be called as a '1' (heterozygous),
  - a genotype with P(G=0,1,2) = (0.1, 0.7, 0.2) would be left uncalled
Figure 1. Graphical summary of T2D GWAS in a Han Chinese population.


Imputation for genotype QC

<table>
<thead>
<tr>
<th>Genotype data</th>
<th>Imputed data</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>AB</td>
</tr>
<tr>
<td>41</td>
<td>281</td>
</tr>
<tr>
<td>89</td>
<td>247</td>
</tr>
<tr>
<td>54</td>
<td>250</td>
</tr>
<tr>
<td>89</td>
<td>247</td>
</tr>
<tr>
<td>54</td>
<td>250</td>
</tr>
<tr>
<td>60</td>
<td>53</td>
</tr>
<tr>
<td>69</td>
<td>235</td>
</tr>
<tr>
<td>53</td>
<td>272</td>
</tr>
<tr>
<td>78</td>
<td>206</td>
</tr>
</tbody>
</table>
Future works

- To provide bioinformatics services for GWA analysis (starting from Q2) and imputation services (starting from Q3 of 2014).
- To construct a Taiwanese Han-Chinese reference panel (to be released in Q3 of 2014).

Discussions

- Rare variants
- Different genotyping arrays
- Sample size