Fast and accurate 1000 Genomes imputation using summary statistics or low-coverage sequencing data

Bogdan Pasaniuc
GWAS study designs

Sequencing: >30x

Illumina 1M

Sequencing: 4x, <1x?

r² (True, Inferred genotypes) ~ 1

r² (True, Inferred genotypes) ~ 0.9

r² ~ 0.7 (0.24x)

High coverage & lots of samples ➔ too expensive
Imputation from reference panels improves power in GWAS

Howie et al. 2009 PLoS Genet; also see Marchini et al. 2007 Nat Genet
Genotype imputation: “two-thousand and late”?  

1. Key ingredient for increasing power in GWAS  
   [Marchini&Howie, Nat Rev Genet 2010,...]  

2. Enables powerful meta-analyses  

3. Accurate genotype calls from sequencing data  
   [1000 Genomes Project, Pasaniuc et al NatGen 2012,...]
Array-based imputation: existing methods are accurate but slow

Number of CPU days needed to impute 11.6 million SNPs using a 1000G reference panel of 292 European samples:

<table>
<thead>
<tr>
<th>Method</th>
<th>N=10,000 samples</th>
<th>N=50,000 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impute1(^1)</td>
<td>9,000 days</td>
<td>45,000 days</td>
</tr>
<tr>
<td>BEAGLE(^2)</td>
<td>2,500 days</td>
<td>12,500 days</td>
</tr>
<tr>
<td>Impute2(^3)</td>
<td>1,000 days</td>
<td>5,000 days</td>
</tr>
<tr>
<td>Impute2 with pre-phasing(^4)</td>
<td>200 days</td>
<td>1,100 days</td>
</tr>
</tbody>
</table>

\(^1\)Marchini et al. 2007 Nat Genet
\(^2\)Browning et al. 2009 Am J Hum Genet
\(^3\)Howie et al. 2009 PLoS Genet
\(^4\)Howie et al. 2012 Nat Genet
Imputation: limitations

• Imputation requires a lot of runtime

• Existing methods cannot be applied to summary statistics directly
  – Individual level genotype data is required
  – Challenge to obtain individual level data in meta-analysis

• Can we test for association untyped markers without access to individual level data?
Array-based imputation: why not use a Gaussian approach?

- Data at nearby SNPs is correlated (linkage disequilibrium)
  - Best performing methods use HMMs to model haplotype structure in the population

- Model correlations among SNPs with a Gaussian multi-variate
  - We assume $X \sim N(\mu,\Sigma)$
    - $\mu,\Sigma$ known from 1000G reference panel
    - LD blocks $\Rightarrow$ windows of fixed length (e.g. 0.5Mb).
    - $X$ – individual genotypes (standard imputation)
    - $X$ – association statistics (summary level imputation)

Gaussian imputation

- **Step 1.** Infer mean $\mu$ and covariance $\Sigma$ for summary data from reference panel
  - Allele frequencies:
    - $\Sigma(p_i, p_j) = 1/(2N-1) (p_{ij} - p_i p_j)$
    - $\mu = \text{(population allele frequencies)}$
  - Association z-scores:
    - $\Sigma(z_i, z_j) = r_{ij} \text{ (correlation coefficient)}$
    - $\mu = 0 \text{ (NULL)}$

- **Multivariate Central Limit**
  - $X \text{ summary statistics over sample of haplotypes}$
  - $X \sim N(\mu, \Sigma)$
Gaussian imputation

- Allele frequencies follow $N(\mu, \Sigma)$
  - $(\mu, \Sigma)$ inferred from reference panel

- **Step 2.** Infer conditional distribution of unobserved given typed

- $P_{\text{typed}}$ - Observed frequencies at subset of SNPs

- $X = \text{Frequencies at rest of SNPs}$

- **Conditional Distribution** $X_{i|t}$ is also Gaussian
  - $X_{\text{imputed}|\text{typed}} \sim N(\mu_{i|t}, \Sigma_{i|t})$
Conditional distribution is analytically derived

- Conditional distribution is also Normal
  \[ \chi_{\text{imputed|typed}} \sim N(\mu_{i|t}, \Sigma_{i|t}) \]

\[ \mu_{i|t} = \mu_i + \Sigma_{i,t} \Sigma^{-1}_{t,t} (p_t - \mu_t) \]

[Lynch&Walsh, Genetics and Analysis of Quantitative Traits, 1998]
Conditional distribution is analytically derived

- Conditional distribution is also Normal
  \[ X_{\text{imputed|typed}} \sim N(\mu_{i|t}, \Sigma_{i|t}) \]

\[
\mu_{i|t} = \mu_i + \Sigma_{i,t} \Sigma_{t,t}^{-1} (p_t - \mu_t)
\]

Population Frequency

Correlation among typed & imputed SNPs

Deviation from the population frequency at typed SNPs

Correlation among typed SNPs

[Lynch&Walsh, Genetics and Analysis of Quantitative Traits, 1998]
Conditional distribution is analytically derived

- Conditional distribution is also Normal
  \[ X_{\text{imputed}|\text{typed}} \sim N(\mu_{i|t}, \Sigma_{i|t}) \]

\[ \mu_{i|t} = \mu_i + \Sigma_{i,t} \Sigma^{-1}_{t,t} (p_t - \mu_t) \]

- Linear transformation w/ weights pre-computed based on reference panel
- (Prediction world) \(\Rightarrow\) Best Linear Unbiased Predictor (BLUP) [Henderson 1975 Biometrics]
- For \(N=2\) \(\Rightarrow\) imputation of individual level data
- Similar approaches but with different var/cov: Best Linear IMPutation (BLIMP) [Wen&Stephens 2010]
Conditional distribution is analytically derived

• **Step 3.**
  – Derive $\mu_{i|t}$ and use as imputation

• **Step 4.**
  – Compute association statistics over imputed frequencies *(Imp-G-summary)*
  
  \[ \mu_{i|t} = \mu_i + \Sigma_{i,t} \Sigma^{-1} \Sigma_{t,t} (p_t - \mu_t) \]

  – Impute individual level data (n=2) *(Imp-G)*

  – Use conditional variance as measure for accuracy
Simulations

- 1000 Genomes data
- 292 Europeans used as reference
- The rest used to simulate case-control data sets
  - HAPGEN [Spencer et al PlosGen 2009]
- Randomly selected 0.5Mb loci from Chr 1
- Illumina 1M SNPs for array imputation
- Armitage Trend Test for case-control association (Armitage 1955 Biometrics)

- Beagle Imputation
  - Imputes genotypes
  - Requires individual level data

- ImpG
  - Imputes genotypes
  - Requires individual level data

- ImpG-summary
  - Imputes frequencies
  - z-scores over imputed freqs
  - Does not need individual level data
No inflation under null (odds ratio = 1)
Accurate 1000G imputation using Gaussian approach (ImpG)

Average ratio of $\chi^2$ statistics for imputed vs. true genotypes in simulations of 1K cases + 1K controls (odds ratio = 1.5):

<table>
<thead>
<tr>
<th></th>
<th>All SNPs (MAF&gt;1%)</th>
<th>Common SNPs (MAF&gt;5%)</th>
<th>Low-freq SNPs (1%&lt;MAF&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAGLE</td>
<td>0.87</td>
<td>0.89</td>
<td>0.65</td>
</tr>
<tr>
<td>ImpG</td>
<td>0.85</td>
<td>0.87</td>
<td>0.59</td>
</tr>
</tbody>
</table>
Accurate 1000G imputation using Gaussian approach (ImpG)

Average ratio of $\chi^2$ statistics for imputed vs. true genotypes in simulations of 1K cases + 1K controls (odds ratio = 1.5):

<table>
<thead>
<tr>
<th></th>
<th>All SNPs (MAF&gt;1%)</th>
<th>Common SNPs (MAF&gt;5%)</th>
<th>Low-freq SNPs (1%&lt;MAF&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAGLE</td>
<td>0.87</td>
<td>0.89</td>
<td>0.65</td>
</tr>
<tr>
<td>ImpG</td>
<td>0.85</td>
<td>0.87</td>
<td>0.59</td>
</tr>
</tbody>
</table>

- Or, run ImpG genome-wide, then run BEAGLE only on regions of significant or suggestive association.
Accurate 1000G imputation using summary statistics (ImpG-summary)

Average ratio of $\chi^2$ statistics for imputed vs. true genotypes in simulations of 1K cases + 1K controls (odds ratio = 1.5):

<table>
<thead>
<tr>
<th>Method</th>
<th>All SNPs (MAF&gt;1%)</th>
<th>Common SNPs (MAF&gt;5%)</th>
<th>Low-freq SNPs (1%&lt;MAF&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAGLE</td>
<td>0.87</td>
<td>0.89</td>
<td>0.65</td>
</tr>
<tr>
<td>ImpG</td>
<td>0.85</td>
<td>0.87</td>
<td>0.59</td>
</tr>
<tr>
<td>ImpG-summary</td>
<td>0.82</td>
<td>0.84</td>
<td>0.52</td>
</tr>
</tbody>
</table>
Accurate 1000G imputation using summary statistics (ImpG-summary)

Average ratio of $\chi^2$ statistics for imputed vs. true genotypes in simulations of 1K cases + 1K controls (odds ratio = 1.5):

<table>
<thead>
<tr>
<th></th>
<th>All SNPs (MAF&gt;1%)</th>
<th>Common SNPs (MAF&gt;5%)</th>
<th>Low-freq SNPs (1%&lt;MAF&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAGLE</td>
<td>0.87</td>
<td>0.89</td>
<td>0.65</td>
</tr>
<tr>
<td>ImpG</td>
<td>0.85</td>
<td>0.87</td>
<td>0.59</td>
</tr>
<tr>
<td>ImpG-summary</td>
<td>0.82</td>
<td>0.84</td>
<td>0.52</td>
</tr>
</tbody>
</table>

• Consortia can impute meta-analysis summary statistics into new reference panels without having to repeat imputation separately in each individual cohort.
Accurate 1000G imputation when imputing GBR from rest of EUR

• Used all Great Britain data from 1000G for simulations and the rest as reference panel
• Average ratio of $\chi^2$ statistics for imputed vs. true genotypes in simulations of 1K cases + 1K controls (odds ratio = 1.5):

<table>
<thead>
<tr>
<th></th>
<th>All SNPs (MAF&gt;1%)</th>
<th>Common SNPs (MAF&gt;5%)</th>
<th>Low-freq SNPs (1%&lt;MAF&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAGLE</td>
<td>0.867</td>
<td>0.888</td>
<td>0.630</td>
</tr>
<tr>
<td>ImpG</td>
<td>0.842</td>
<td>0.867</td>
<td>0.570</td>
</tr>
<tr>
<td>ImpG-summary</td>
<td>0.816</td>
<td>0.843</td>
<td>0.516</td>
</tr>
</tbody>
</table>
Gaussian imputation is extremely fast

Number of CPU days needed to impute 11.6 million SNPs using a 1000G reference panel of 292 European samples:

<table>
<thead>
<tr>
<th>Method</th>
<th>N=10,000 samples</th>
<th>N=50,000 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impute1</td>
<td>9,000 days</td>
<td>45,000 days</td>
</tr>
<tr>
<td>BEAGLE</td>
<td>2,500 days</td>
<td>12,500 days</td>
</tr>
<tr>
<td>Impute2</td>
<td>1,000 days</td>
<td>5,000 days</td>
</tr>
<tr>
<td>Impute2 with pre-phasing</td>
<td>200 days</td>
<td>1,100 days</td>
</tr>
<tr>
<td>ImpG</td>
<td>4 days</td>
<td>20 days</td>
</tr>
<tr>
<td>ImpG-summary</td>
<td>0.4 days</td>
<td>0.4 days</td>
</tr>
</tbody>
</table>

Note: ImpG/ImpG-summary running time ~ (#reference samples)
BEAGLE and Impute2 running time ~ (#reference samples)^2
Sequencing-based imputation is different from array-based imputation

Li et al. 2011 Genome Res
Low-coverage sequencing + imputation increases power vs. genotyping arrays

Effective sample size of a GWAS with a $300,000 budget:

<table>
<thead>
<tr>
<th></th>
<th>Cost per sample</th>
<th>Actual #samples</th>
<th>Average imputation $r^2$</th>
<th>Effective #samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina 1M array</td>
<td>$400</td>
<td>750</td>
<td>1.00</td>
<td>750</td>
</tr>
<tr>
<td>0.4x sequencing</td>
<td>$83*</td>
<td>3,600</td>
<td>0.81**</td>
<td>2,900</td>
</tr>
<tr>
<td>0.1x sequencing</td>
<td>$43*</td>
<td>7,000</td>
<td>0.64**</td>
<td>4,500</td>
</tr>
</tbody>
</table>

*Based on sample preparation cost of $30/sample, which is conservatively double the $15/sample reported by Rohland & Reich 2012 Genome Res, and on $133 per 1x sequencing (Illumina Network cost).

**Imputation $r^2$ attained at Illumina 1M SNPs by downsampling reads from real off-target exome sequencing data. Relative performance of low-coverage sequencing will be even higher at non-Illumina 1M SNPs.

Pasaniuc et al. 2012 Nat Genet
How much more powerful is low-coverage sequencing than genome-wide arrays?

Pasaniuc et al. 2012 Nat Genet
Sequencing-based imputation: existing methods are accurate but slow

Number of CPU days needed to impute 11.6 million SNPs from 1x low-coverage sequencing data using a 1000G reference panel of 292 European samples:

<table>
<thead>
<tr>
<th>Method</th>
<th>N=10,000 samples</th>
<th>N=50,000 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAGLE</td>
<td>2,500 days</td>
<td>12,500 days</td>
</tr>
<tr>
<td>Impute2</td>
<td>3,700 days</td>
<td>18,500 days</td>
</tr>
<tr>
<td>Impute2 with pre-phasing</td>
<td>not applicable</td>
<td>not applicable</td>
</tr>
</tbody>
</table>
Sequencing-based imputation: Gaussian approach

Let \( g \) denote genotypes.
Let \( \Sigma \) denote covariance between SNPs.
We assume \( g \sim N(\mu, \Sigma) \) with \( \mu, \Sigma \) known from 1000G reference panel, restricting to windows of fixed length.

Max-Likelihood framework: find \( g \) that maximizes likelihood

\[
P\left( \text{Read data} \mid \text{genotype } g \right) \times P\left( \text{genotype} \mid \mu, \Sigma \right)
\]

Error model: product of binomials (error rate \( \varepsilon \))

\[
P(R_i \mid g_i^j = 0; N_i^j) = \binom{N_i^j}{R_i} \times \varepsilon^{R_i} \times (1 - \varepsilon)^{N_i^j - R_i}
\]
\[
P(R_i \mid g_i^j = 1; N_i^j) = \binom{N_i^j}{R_i} \times \frac{1}{2}^{N_i^j}
\]
\[
P(R_i \mid g_i^j = 2; N_i^j) = \binom{N_i^j}{R_i} \times \varepsilon^{N_i^j - R_i} \times (1 - \varepsilon)^{R_i}
\]

\[
\exp(-[g-\mu]^T\Sigma^{-1}[g-\mu]/2)
\]
Sequencing-based imputation: extremely-fast algorithm

- Use $\Sigma$ to augment read counts using linked SNPs, then infer posterior $P(g_i)$ at each SNP $i$ independently.
- Borrow reads from nearby SNPs in LD
  - New counts are linear combination of reads from nearby SNPs

Standard approach:
Reads at $g_i$: $(R_i, A_i)$

Proposed approach (ImpG-seq):
Reads at $g_i$: $(R_i, A_i)$

- If $R_i + A_i$ small
  - $R_i', = \Sigma \rho R_j$
  - $A_i', = \Sigma \rho A_j$

$P(g_i \mid R_i, A_i, f_i)$

$P(g_i \mid R_i', A_i', f_i)$
Accuracy of sequence-based imputation using Gaussian approach (ImpG-seq)

Average imputation $r^2$ for 0.5x low-coverage sequencing data using a 1000G reference panel of 292 European samples:

<table>
<thead>
<tr>
<th></th>
<th>All SNPs (MAF&gt;1%)</th>
<th>Common SNPs (MAF&gt;5%)</th>
<th>Low-freq SNPs (1%&lt;MAF&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-SNP*</td>
<td>0.18</td>
<td>0.19</td>
<td>0.16</td>
</tr>
<tr>
<td>BEAGLE</td>
<td>0.78</td>
<td>0.86</td>
<td>0.57</td>
</tr>
<tr>
<td>ImpG-seq</td>
<td>0.57</td>
<td>0.65</td>
<td>0.37</td>
</tr>
</tbody>
</table>

*Simple genotype calling strategy that analyzes each SNP independently using allele frequencies $\mu$ (but not covariance $\Sigma$) from reference panel
Accuracy of sequence-based imputation using Gaussian approach (ImpG-seq)

Average imputation $r^2$ for 4x low-coverage sequencing data using a 1000G reference panel of 292 European samples:

<table>
<thead>
<tr>
<th></th>
<th>All SNPs (MAF&gt;1%)</th>
<th>Common SNPs (MAF&gt;5%)</th>
<th>Low-freq SNPs (1%&lt;MAF&lt;5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-SNP*</td>
<td>0.65</td>
<td>0.68</td>
<td>0.58</td>
</tr>
<tr>
<td>BEAGLE</td>
<td>0.93</td>
<td>0.96</td>
<td>0.85</td>
</tr>
<tr>
<td>ImpG-seq</td>
<td>0.77</td>
<td>0.80</td>
<td>0.69</td>
</tr>
</tbody>
</table>

*Simple genotype calling strategy that analyzes each SNP independently using allele frequencies $\mu$ (but not covariance $\Sigma$) from reference panel
Sequencing-based imputation using Gaussian approach is extremely fast

Number of CPU days needed to impute 11.6 million SNPs from 1x low-coverage sequencing data using a 1000G reference panel of 292 European samples:

<table>
<thead>
<tr>
<th>Method</th>
<th>N=10,000 samples</th>
<th>N=50,000 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAGLE</td>
<td>2,500 days</td>
<td>12,500 days</td>
</tr>
<tr>
<td>Impute2</td>
<td>3,700 days</td>
<td>18,500 days</td>
</tr>
<tr>
<td>Impute2 with pre-phasing</td>
<td>not applicable</td>
<td>not applicable</td>
</tr>
<tr>
<td>ImpG-seq</td>
<td>4 days</td>
<td>20 days</td>
</tr>
</tbody>
</table>

Note: ImpG-seq running time $\sim (#\text{reference samples})$
BEAGLE and Impute2 running time $\sim (#\text{reference samples})^2$
Conclusions

• Gaussian models ➞ fast linear predictors

• Linear models recover most of the association signal for 1000 Genomes imputation!

• Array-based imputation: Gaussian imputation is very fast and accurate, and can be applied to summary statistics.

• Sequencing-based imputation: Low-coverage sequencing is far superior to genotyping using genome-wide arrays. Gaussian imputation is very fast and moderately accurate.
Acknowledgements

Harvard School of Public Health:
Alkes Price
Sasha Gusev
Gaurav Bhatia

UCSF
Noah Zaitlen

Harvard Medical School/Broad
Nick Patterson
David Reich

Postdoctoral positions available at UCLA!