Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations

Yi-Juan Hu\textsuperscript{a}, Yun Li\textsuperscript{b,c}, Paul L. Auer\textsuperscript{d}, and Dan-Yu Lin\textsuperscript{b,1}

\textsuperscript{a}Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322; \textsuperscript{b}Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420; \textsuperscript{c}Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7264; and \textsuperscript{d}Joseph J. Zilber School of Public Health, University of Wisconsin, Milwaukee, WI 53201-0413

Edited by Elizabeth A. Thompson, University of Washington, Seattle, WA, and approved December 9, 2014 (received for review April 3, 2014)
Outline

• Motivating example: Women’s Health Initiative (WHI) data

• Our approach: a robust variance estimator

• Simulation studies

• Application to the WHI data

• Conclusions
Outline

• Motivating example: Women’s Health Initiative (WHI) data
  • Our approach: a robust variance estimator
  • Simulation studies
  • Analyzing the WHI data
  • Conclusions
Motivating example: Women’s Health Initiative (WHI) data

- **Original WHI**, 1991
  - Enrolled $\geq 160,000$ postmenopausal women (aged 50–79)

- **WHI genome-wide association study (WHI-GWAS)**, 2007
  - Genotyped 12,008 women
  - Affymetrix 6.0 array: $\sim 550,000$ SNPs

- **WHI exome sequencing project (WHI-ESP)**, 2010
  - Due to the high cost of sequencing, only 2,150 women were sequenced
  - Whole-exome sequencing: all variants in the exome
Now we are interested in mapping SNPs for BMI in AA ...

- **Affymetrix 6.0 array**
  - 8,142 AA
  - ∼ 550,000 SNPs, most of which are common

- **Whole-exome sequencing**
  - 360 AA with BMI values >40 or <25 (*extreme-trait sampling*)
  - All variants, including all rare variants

![Histograms showing BMI frequency](image-url)
Mapping rare variants for BMI

- Rare variants have been hypothesized to have a large impact
- Assayed by sequencing, not arrays (missing by design)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>30.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>28.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>35.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>23.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>33.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>45.6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- Existing approach 1: use sequenced subjects only
  (Tennessen et al 2012, Science)

- Existing approach 2: genotype imputation
  (Auer et al 2012, AJHG)
  - Use a reference panel; fill in missing data by posterior means
  - MaCH, minimac (Li 2010, Gen Epid)
Observed Genotypes

Reference Haplotype
c g a g A t c t c c c c g A c c t c c A t g g

c g a a G c t c t t t t C t t t c A t g g

Index

1 2 3 4 5 6 7 8 9 10

Hidden
State S
Mapping rare variants using sequenced & imputed values

- Burden score: \( S = G_1 + \ldots + G_M \)
- \( Y = \gamma + \beta S + \epsilon \)
- Test \( H_0 : \beta = 0 \) with the *standard* score statistic

![Quantile-quantile (QQ) plot](image)

**Inflated type I error!**
Reasons for inflated type I error

- Imputation creates differential quality in genotype data
  - Rare variants cannot be imputed very accurately, so imputed values have a smaller variance than sequenced
- Extreme-trait sampling creates differential variation in BMI
  - Sequenced subjects have a greater variance

Thus, the variance of genotype values is related to the variance of phenotype values, causing the standard score statistic to fail
Existing solution to inflated type I error

- Use accurately imputed variants only
  - Quality control (QC): exclude poorly imputed variants
  - Type I error controlled
  - However, 82.9% variants were removed. Power loss?
Summarizing WHI data and generalizing the problem

Goal: to map rare variants for a disease/trait

- Sequence all in a large cohort? Economically infeasible
- A cost-effective sampling strategy: trait-dependent
- The past wave of GWAS have collected array genotype data
- Genotype imputation
- Inflated type I error when applying the standard score test
- Existing solutions lose power
Our goals in this work

Develop valid and efficient association tests for genotype data with differential qualities

- Show the score statistic is unbiased
- Show the standard variance estimator for the score statistic is invalid when the sequenced subjects are not a random subset
- Derive a robust variance estimator for the score statistic

Our tests have correct type I error (under any sampling scheme for sequencing) and improved power
Features of our methodology

- Handle many types of trait and any sub-sampling scheme
- Encompass all commonly used rare variant tests (Burden, SKAT, etc); include single-variant tests as special cases
- Allow for covariates
- Simple implementation: replacing the standard variance estimator with the robust one
Outline

• Motivating example: Women’s Health Initiative (WHI) data

• Our approach: a robust variance estimator

• Simulation studies

• Application to the WHI data

• Conclusions
Consider the simplest case: a single variant, no covariate

- \( G \): genotype of the variant
- \( Y \): trait (here, quantitative)
- \( N \): number of all cohort members, having array data
- \( n \): number of subjects selected for sequencing, having array and sequencing data
  - The first \( n \) subjects are the sequenced ones
- \( \tilde{G} \)
  - imputed \( G \) (by posterior mean) for a non-sequenced subject
  - observed \( G \) for a sequenced subject
The score statistic is unbiased

To test $H_0 : \beta = 0$ in the linear regression

\[ Y = \gamma + \beta G + \epsilon \]

The score statistic based on $(Y_i, \tilde{G}_i)$ ($1 \leq i \leq N$), $\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i$

\[ U = \sum_{i=1}^{N} (Y_i - \bar{Y})\tilde{G}_i \]

By some simple algebra, denoting $\bar{G} = N^{-1} \sum_{i=1}^{N} \tilde{G}_i$

\[ U = \sum_{i=1}^{n} Y_i(\tilde{G}_i - \bar{G}) + \sum_{i=n+1}^{N} Y_i(\tilde{G}_i - \bar{G}) \]

$E(U) = 0$, because $Y$ is independent of $\tilde{G}$ in both samples

- $Y$ is independent of $G$ in both samples
- Imputation does not depend on $Y$
**$V_{\text{std}}$ tends to underestimate $\text{Var}(U)$**

The standard variance estimator for $U$

$$V_{\text{std}} = N^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \sum_{i=1}^{N} (\tilde{G}_i - \bar{G})^2$$

- Consider balanced extreme-trait sampling for sequencing
- $\text{Var}(Y_u) < \text{Var}(Y_s)$, $\text{Var}(\tilde{G}) < \text{Var}(G)$
- $\text{Var}(U) = n \text{Var}(Y_s) \text{Var}(G) + (N - n) \text{Var}(Y_u) \text{Var}(\tilde{G})$
- $V_{\text{std}} \approx N^{-1} \{n \text{Var}(Y_s) + (N - n) \text{Var}(Y_u)\} \{n \text{Var}(G) + (N - n) \text{Var}(\tilde{G})\}$
- By Chebyshev’s sum inequality, $V_{\text{std}} < \text{Var}(U)$
We propose a robust variance estimator

\[
V_{\text{rob}} = \sum_{i=1}^{n} \left\{ Y_i - \bar{Y} - (1 - r^2)(\bar{Y}_{\text{seq}} - \bar{Y}) \right\}^2 (\tilde{G}_i - \bar{G})^2
\]

\[+ \sum_{i=n+1}^{N} (Y_i - \bar{Y})^2 (\tilde{G}_i - \bar{G})^2 \]

\[
\bar{Y}_{\text{seq}} = n^{-1} \sum_{i=1}^{n} Y_i
\]

- \( r \): correlation coefficient between true and imputed genotypes
- \( r^2 \): estimated by \( \text{Rsq} = \text{Var}(\tilde{G})/[2\hat{p}(1 - \hat{p})] \), where \( \hat{p} \) is MAF
- \( \text{Rsq} \): imputation accuracy
Deriving $V_{\text{rob}}$ ...

- $E(\tilde{G}|G, G_{\text{seq}}) = (1 - r^2)\overline{G}_{\text{seq}} + r^2 G$
  
  $G_{\text{seq}} = (G_1, \ldots, G_n)$, $\overline{G}_{\text{seq}} = n^{-1} \sum_{i=1}^{n} \tilde{G}_i$

- $\text{Var}(U) = E\{ \text{Var}(U|Y) \} + \text{Var}\{E(U|Y)\}$
  
  $= E\left[ E\{ \text{Var}(U|Y, G_{\text{seq}})|Y \} + \text{Var}\{E(U|Y, G_{\text{seq}})|Y \} \right] + \text{Var}\{E(U|Y)\}$
Connection between $V_{\text{rob}}$ and $V_{\text{std}}$

If the imputation is perfect or the selection for sequencing is random ...

- $(1 - r^2)(\bar{Y}_\text{seq} - \bar{Y}) = 0$, and $V_{\text{rob}}$ becomes

$$
\sum_{i=1}^{N} (Y_i - \bar{Y})^2 (\tilde{G}_i - \bar{G})^2
$$

- $(Y - \bar{Y})^2$ and $(\tilde{G} - \bar{G})^2$ are uncorrelated, and (1) is equivalent to

$$
V_{\text{std}} = N^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \sum_{i=1}^{N} (\tilde{G}_i - \bar{G})^2
$$

... $V_{\text{std}}$ will be valid
Outline

• Motivating problem: Women’s Health Initiative (WHI) data
• Our approach: a robust variance estimator
• Simulation studies
• Application to the WHI data
• Conclusions
Simulation setup

- Choose one gene, \textit{NPHS2}; restrict analysis to 5 rare variants
- Generate genotype data of all variants by GWAsimulator (Li and Li, 2008)
- Generate the trait
  - $Y$ quantitative: $Y = \beta S + \gamma_1 X + \epsilon$
  - $Y$ binary: $\text{logit}\{\text{Pr}(Y = 1)\} = \beta S + \gamma_1 X + \gamma_0$
  - $X \sim N(0, 1)$
- $N = 5,000$
Sampling schemes for selecting subjects for sequencing

- Quantitative trait
  - 500 random
  - 250 largest, 250 smallest
  - 500 largest, 250 smallest
  - 250 largest, 250 random
  - 250 largest, 1000 random

- Binary trait
  - Five disease rates: 50%, 30%, 20%, 10%, 5%
  - Always sample 250 cases and 250 controls

- Imputation by minimac: the largest Rsq is 0.22
## Simulation results: type I error

<table>
<thead>
<tr>
<th>Sampling scheme</th>
<th>Bias</th>
<th>SE</th>
<th>$V_{rob}$</th>
<th>SEE</th>
<th>Size</th>
<th>$V_{std}$</th>
<th>SEE</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 random</td>
<td>0.000</td>
<td>0.120</td>
<td>0.118</td>
<td>0.87</td>
<td>1.02</td>
<td>0.118</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>250 largest, 250 smallest</td>
<td>−0.001</td>
<td>0.181</td>
<td>0.180</td>
<td>0.68</td>
<td>36.22</td>
<td>0.118</td>
<td>36.22</td>
<td></td>
</tr>
<tr>
<td>500 largest, 250 smallest</td>
<td>−0.006</td>
<td>0.192</td>
<td>0.191</td>
<td>0.96</td>
<td>27.95</td>
<td>0.131</td>
<td>27.95</td>
<td></td>
</tr>
<tr>
<td>250 largest, 250 random</td>
<td>−0.006</td>
<td>0.134</td>
<td>0.130</td>
<td>0.89</td>
<td>3.61</td>
<td>0.118</td>
<td>3.61</td>
<td></td>
</tr>
<tr>
<td>250 largest, 1000 random</td>
<td>−0.005</td>
<td>0.171</td>
<td>0.169</td>
<td>0.97</td>
<td>3.39</td>
<td>0.152</td>
<td>3.39</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>0.000</td>
<td>0.060</td>
<td>0.059</td>
<td>0.78</td>
<td>0.93</td>
<td>0.059</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>30%</td>
<td>−0.001</td>
<td>0.057</td>
<td>0.056</td>
<td>0.94</td>
<td>1.63</td>
<td>0.054</td>
<td>1.63</td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>−0.002</td>
<td>0.053</td>
<td>0.052</td>
<td>0.91</td>
<td>3.16</td>
<td>0.047</td>
<td>3.16</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>−0.003</td>
<td>0.047</td>
<td>0.046</td>
<td>1.00</td>
<td>13.47</td>
<td>0.036</td>
<td>13.47</td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>−0.003</td>
<td>0.043</td>
<td>0.041</td>
<td>1.09</td>
<td>50.95</td>
<td>0.026</td>
<td>50.95</td>
<td></td>
</tr>
</tbody>
</table>

Nominal significance level: $\alpha = 0.001$; Replicates: 100,000
Simulation results: power for quantitative traits

**500 random**

- T5-rob, all
- T5-seq, all
- T5-std, post-QC

**250 largest, 250 smallest**

- T5-rob, all
- T5-seq, all
- T5-std, post-QC

**250 largest, 250 random**

- T5-rob, all
- T5-seq, all
- T5-std, post-QC

**250 largest, 1000 random**

- T5-rob, all
- T5-seq, all
- T5-std, post-QC
Outline

• Motivating problem: Women’s Health Initiative (WHI) data

• Our approach: a robust variance estimator

• Simulation studies

• Application to the WHI data

• Conclusions
WHI data: recall...

- **Affymetrix 6.0 array**
  - 8,142 AA
  - Assay ~ 550,000 SNPs, most of which are common

- **Whole-exome sequencing**
  - 360 AA with BMI values >40 or <25
  - Assay all variants, including all rare variants

- **Goal:** to map rare variants for BMI

- **Imputation** has been done using minimac
WHI data: imputation accuracy

MAF ≤ 5%

Frequency

0 2000 6000

Rsq

0.0 0.2 0.4 0.6 0.8 1.0
WHI data: QQ-plots

- **T5-rob**
  - $\lambda = 1.04$

- **T5-std**
  - $\lambda = 1.1$

- **T5-seq**
  - $\lambda = 1.05$
### WHI data: top ten genes for BMI identified by T5-rob

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Chr</th>
<th>m</th>
<th>Rsq</th>
<th>P value</th>
<th>T5-rob</th>
<th>T5-std</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODF2L</td>
<td>outer dense fiber of sperm tails 2-like</td>
<td>1</td>
<td>11</td>
<td>0.685</td>
<td>$3.1 \times 10^{-5}$</td>
<td>$1.7 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>ITSN1</td>
<td>intersectin 1 (SH3 domain protein)</td>
<td>21</td>
<td>7</td>
<td>0.609</td>
<td>$5.0 \times 10^{-5}$</td>
<td>$3.3 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>KDM6B</td>
<td>lysine (K)-specific demethylase 6B</td>
<td>17</td>
<td>30</td>
<td>0.266</td>
<td>$5.8 \times 10^{-5}$</td>
<td>$1.0 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>SOCS1</td>
<td>suppressor of cytokine signaling 1</td>
<td>16</td>
<td>2</td>
<td>0.348</td>
<td>$7.8 \times 10^{-5}$</td>
<td>$1.6 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>ODF2L</td>
<td>[with a different accession number]</td>
<td>1</td>
<td>9</td>
<td>0.689</td>
<td>$1.1 \times 10^{-4}$</td>
<td>$7.1 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>ACADVL</td>
<td>acyl-CoA dehydrogenase, very long chain</td>
<td>17</td>
<td>15</td>
<td>0.189</td>
<td>$1.6 \times 10^{-4}$</td>
<td>$6.8 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>BDNF</td>
<td>brain-derived neurotrophic factor</td>
<td>11</td>
<td>2</td>
<td>0.628</td>
<td>$2.1 \times 10^{-4}$</td>
<td>$2.5 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>TRDMT1</td>
<td>tRNA aspartic acid methyltransferase 1</td>
<td>10</td>
<td>3</td>
<td>0.718</td>
<td>$2.3 \times 10^{-4}$</td>
<td>$1.8 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>FAM60A</td>
<td>family with sequence similarity 60, member A</td>
<td>12</td>
<td>1</td>
<td>0.768</td>
<td>$2.3 \times 10^{-4}$</td>
<td>$4.0 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>PDGFRA</td>
<td>platelet-derived growth factor receptor, alpha polypeptide</td>
<td>4</td>
<td>12</td>
<td>0.563</td>
<td>$2.4 \times 10^{-4}$</td>
<td>$2.2 \times 10^{-4}$</td>
<td></td>
</tr>
</tbody>
</table>
Outline

• Motivating problem: Women’s Health Initiative (WHI) data
• Our approach: a robust variance estimator
• Simulation studies
• Application to the WHI data
• Conclusions
Conclusions

We developed an approach to integrative analysis of sequencing and GWAS array data

- Simple and versatile (handle any trait, sampling scheme, test)
- Have correct type I error
- More powerful than
  - use of sequencing data alone
  - use of accurately imputed variants only
- Software: **SEQGWAS**, ~ 2 hrs to analyze the WHI data