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

Yi-Juan Hu[a], Yun Li[b,c], Paul L. Auer[d], and Dan-Yu Lin[b,1]

[a]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322; [b]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420; [c]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7264; and [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

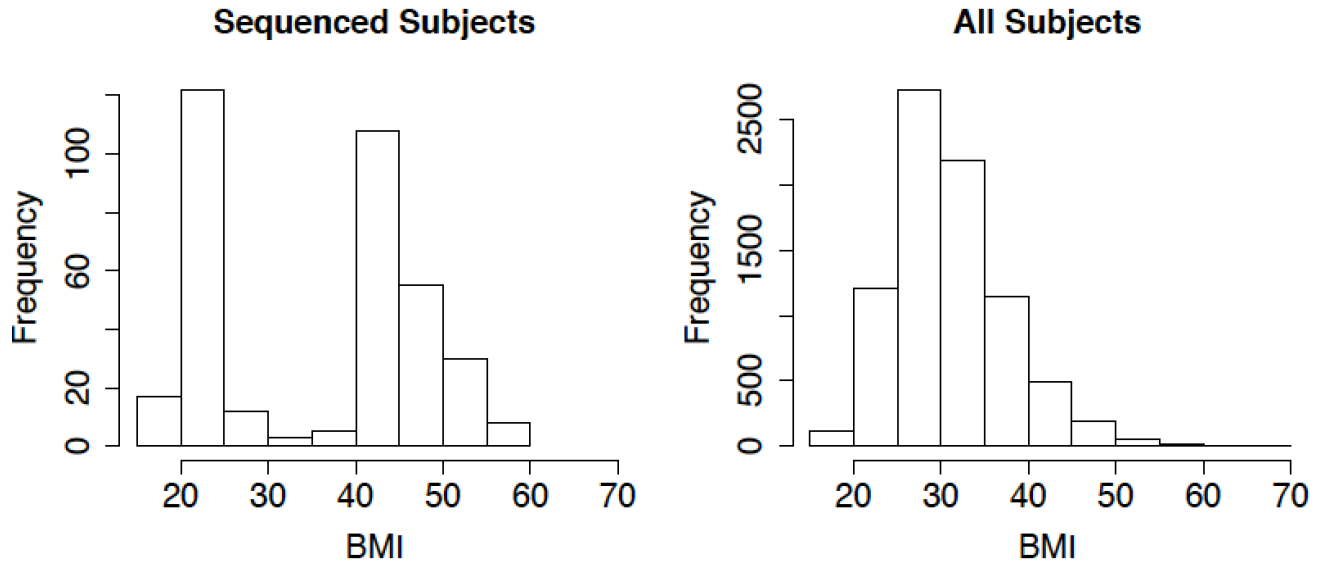
- Motivating example: Women's Health Initiative (WHI) data
- Our approach: a robust variance estimator
- Simulation studies
- Analyzing the WHI data
- Conclusions

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



Mapping rare variants for BMI

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

30.5	·	·	·	·	0	·	·	·	·	·	·	2	·	·	·
28.3	·	·	·	·	1	·	·	·	·	·	·	0	·	·	·
35.0	·	·	·	·	1	·	·	·	·	·	·	0	·	·	·
23.1	·	·	·	·	0	·	·	·	·	·	·	1	·	·	·
33.9	·	·	·	·	0	·	·	·	·	·	·	2	·	·	·
21.2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
45.6	0	1	0	0	0	0	1	0	0	0	0	2	0	0	0

- 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

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

Observed Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

Observed Genotypes

```
c g a g A t c t c c c g A c c t 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
```

Reference Haplotypes

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

Observed Genotypes

```
c g a g A t c t c c c g A c c t 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

Reference Haplotypes

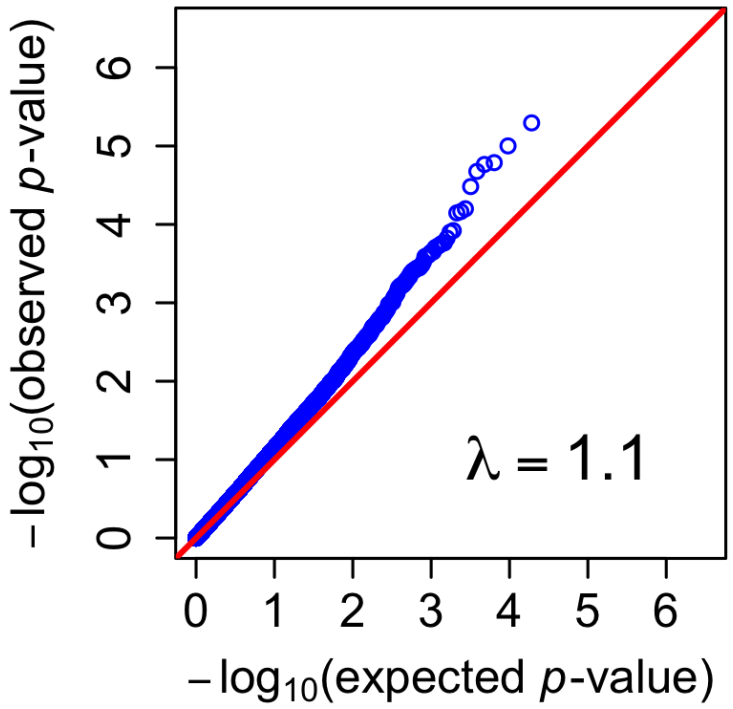
```
1  C G A G A T C T C C T T C T T C T G T G C
2  C G A G A T C T C C C G A C C T C A T G G
3  C C A A G C T C T T T T C T T C T G T G C
4  C G A A G C T C T T T T C T T C T G T G C
5  C G A G A C T C T C C G A C C T T A T G C
6  T G G G A T C T C C C G A C C T C A T G G
7  C G A G A T C T C C C G A C C T T G T G C
8  C G A G A C T C T T T T C T T T T G T A C
9  C G A G A C T C T C C G A C C T C G T G C
10 C G A A G C T C T T T T C T T C T G T G C
```

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



Inflated type I error!

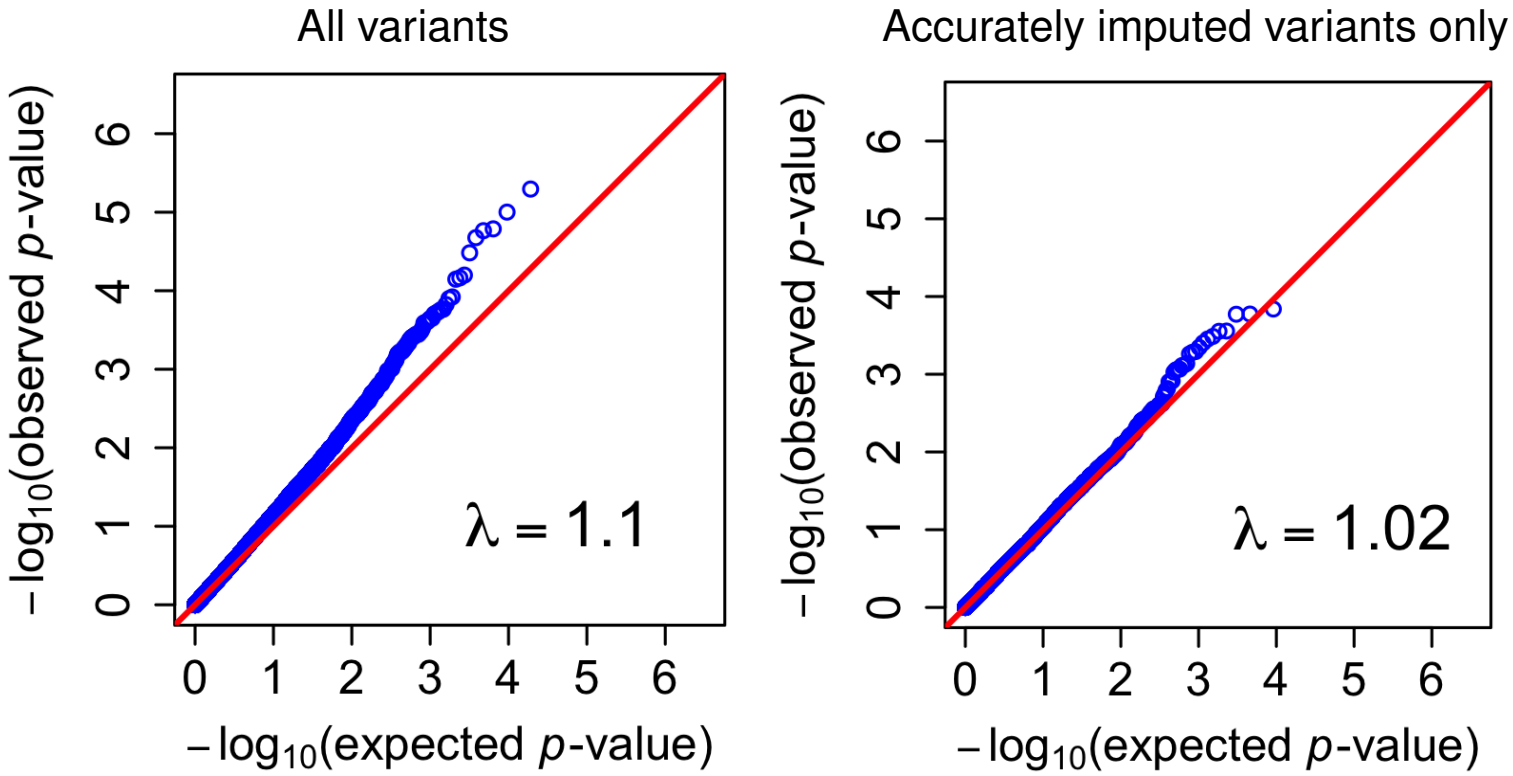
Reasons for inflated type I error

- 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

Notation

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
- $\widetilde{G}$
  - 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 $(Y_i, \widetilde{G}_i)$ $(1 \le i \le N)$, $\overline{Y} = N^{-1} \sum_{i=1}^{N} Y_i$

$$U = \sum_{i=1}^{N} (Y_i - \overline{Y}) \widetilde{G}_i$$

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

$$U = \sum_{i=1}^{n} Y_i (\widetilde{G}_i - \overline{G}) + \sum_{i=n+1}^{N} Y_i (\widetilde{G}_i - \overline{G})$$

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

- Y is independent of G in both samples
- Imputation does not depend on Y

V_{std} tends to underestimate $\text{Var}(U)$

The standard variance estimator for U

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

- Consider balanced extreme-trait sampling for sequencing
- $\text{Var}(Y_u) < \text{Var}(Y_s)$, $\text{Var}(\widetilde{G}) < \text{Var}(G)$
- $\text{Var}(U) = n\text{Var}(Y_s)\text{Var}(G) + (N-n)\text{Var}(Y_u)\text{Var}(\widetilde{G})$
- $V_{\text{std}} \approx N^{-1} \{n\text{Var}(Y_s) + (N-n)\text{Var}(Y_u)\} \left\{n\text{Var}(G) + (N-n)\text{Var}(\widetilde{G})\right\}$
- By Chebyshev's sum inequality, $V_{\text{std}} < \text{Var}(U)$

We propose a robust variance estimator

$$V_{\text{rob}} = \sum_{i=1}^{n} \left\{ Y_i - \overline{Y} - (1 - r^2)(\overline{Y}_{\text{seq}} - \overline{Y}) \right\}^2 (\widetilde{G}_i - \overline{G})^2$$

$$+ \sum_{i=n+1}^{N} (Y_i - \overline{Y})^2 (\widetilde{G}_i - \overline{G})^2$$

$\overline{Y}_{\text{seq}} = n^{-1} \sum_{i=1}^{n} Y_i$

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

Deriving $V_{\mathbf{rob}}$ ...

- $\mathrm{E}(\widetilde{G}|G, \mathcal{G}_{\mathrm{seq}}) = (1 - r^2)\overline{G}_{\mathrm{seq}} + r^2 G$

  $\mathcal{G}_{\mathrm{seq}} = (G_1, \ldots, G_n)$, $\overline{G}_{\mathrm{seq}} = n^{-1} \sum_{i=1}^{n} \widetilde{G}_i$

- $\mathrm{Var}(U) = \mathrm{E}\{\mathrm{Var}(U|\mathcal{Y})\} + \mathrm{Var}\{\mathrm{E}(U|\mathcal{Y})\}$

  $= \mathrm{E}\Big[\mathrm{E}\{\mathrm{Var}(U|\mathcal{Y}, \mathcal{G}_{\mathrm{seq}})|\mathcal{Y}\} + \mathrm{Var}\{\mathrm{E}(U|\mathcal{Y}, \mathcal{G}_{\mathrm{seq}})|\mathcal{Y}\}\Big] + \mathrm{Var}\{\mathrm{E}(U|\mathcal{Y})\}$

Connection between $V_{\mathbf{rob}}$ and $V_{\mathbf{std}}$

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

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

$$\sum_{i=1}^{N} (Y_i - \overline{Y})^2 (\widetilde{G}_i - \overline{G})^2 \tag{1}$$

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

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

... V_{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, *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: $\text{logit}\{\Pr(Y = 1)\} = \beta S + \gamma_1 X + \gamma_0$
  - $X \sim N(0, 1)$
- $N = 5,000$

Sampling schemes for selecting subjects for sequencing

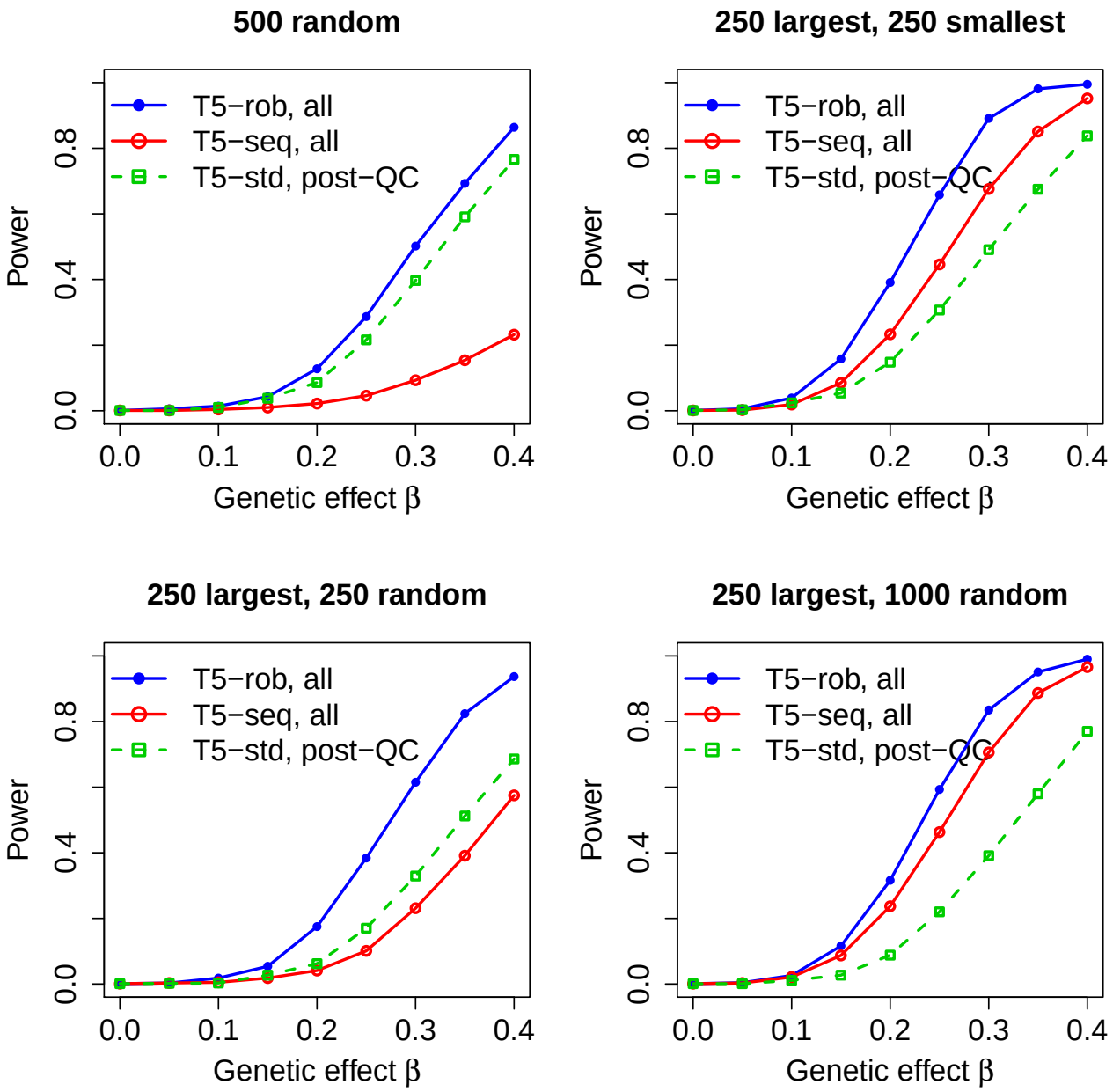
- Quantitative trait
  - 500 random
  - 250 largest, 250 smallest
  - 500 largest, 250 smallest
  - 250 largest, 250 random
  - 250 largest, 1000 random
- Binary trait
  - Five disease rates: 50%, 30%, 20%, 10%, 5%
  - Always sample 250 cases and 250 controls
- Imputation by minimac: the largest Rsq is 0.22

Simulation results: type I error

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

Nominal significance level: $\alpha = 0.001$; Replicates: $100,000$

Simulation results: power for quantitative traits



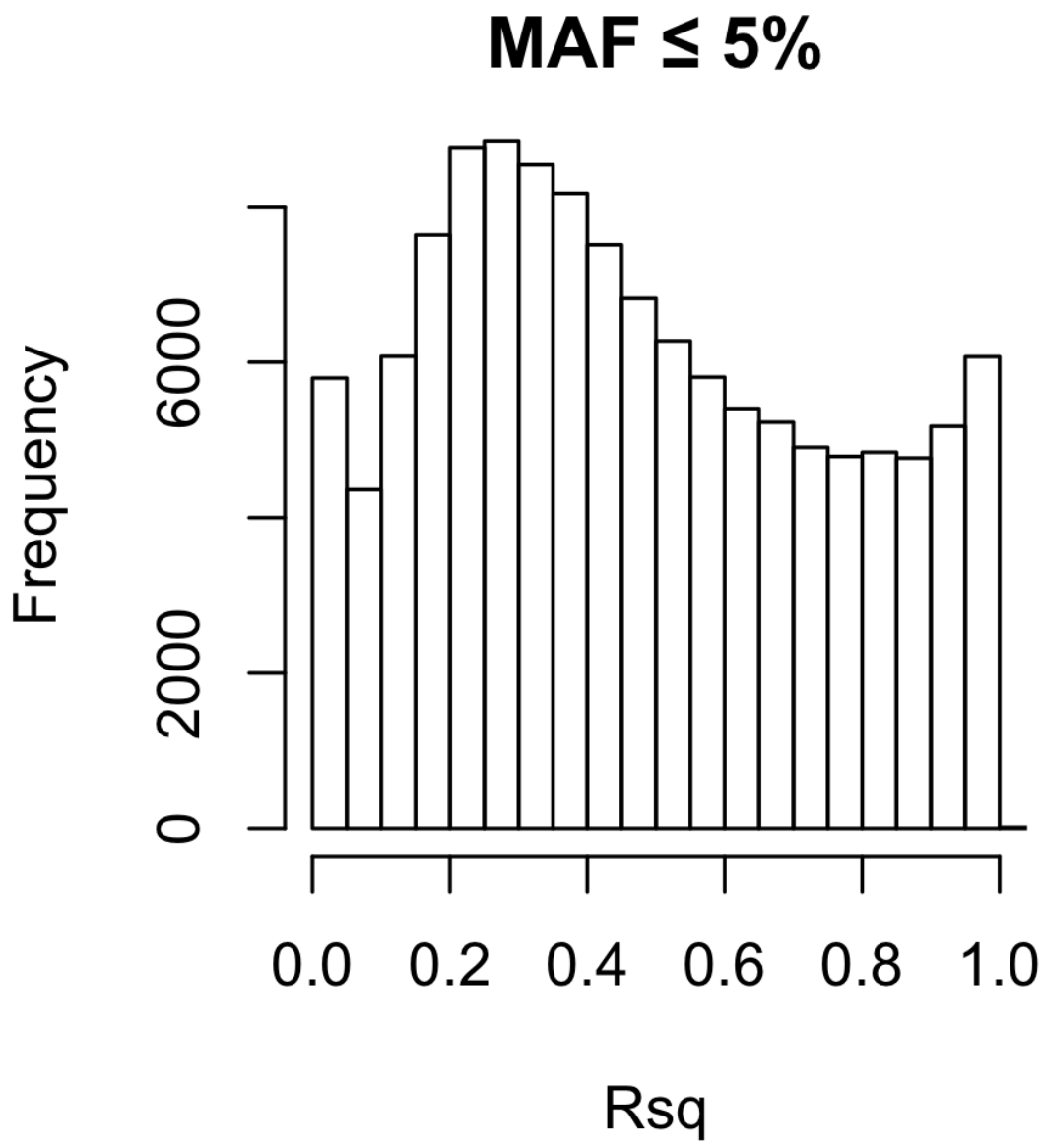
Outline

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

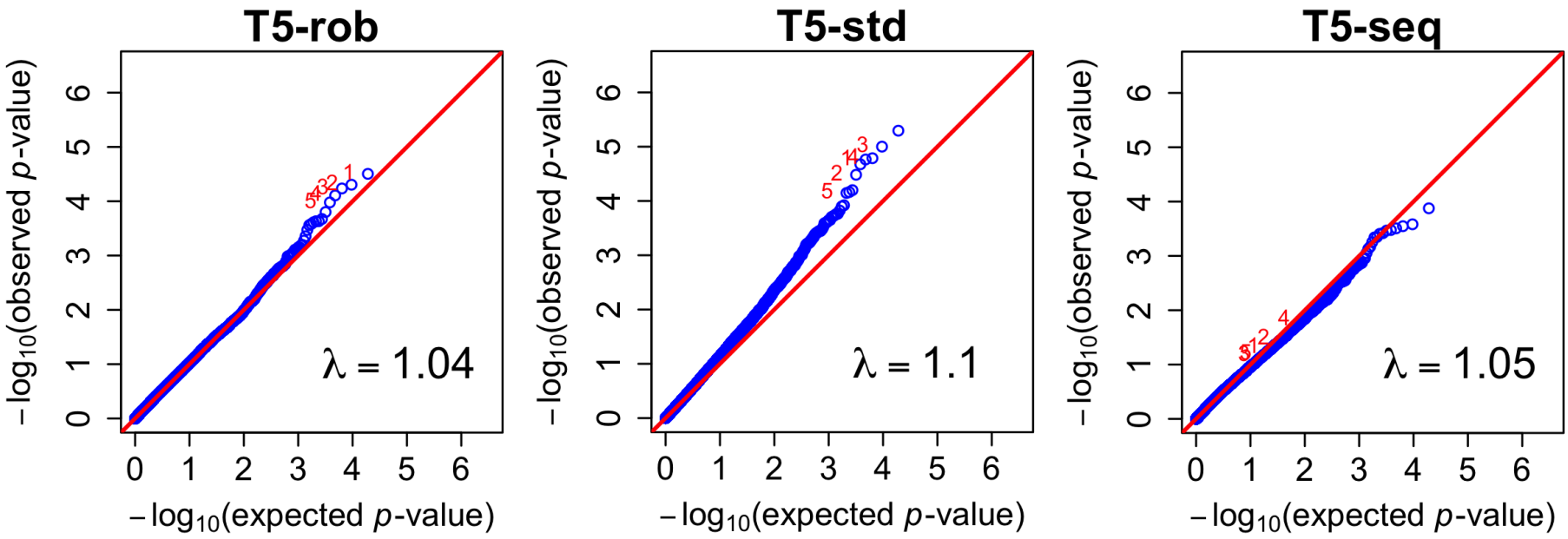
WHI data: recall...

- **Affymetrix 6.0 array**
  - 8,142 AA
  - Assay $\sim$ 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



WHI data: QQ-plots



WHI data: top ten genes for BMI identified by T5-rob

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

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