Improved Score Statistics for Meta-analysis in Single-variant and Gene-level Association Studies

Sai Chen, Illumina, Inc.
Jingjing Yang, Emory University
Gonçalo Abecasis, University of Michigan

Meta Analysis in GWAS

Mimicking joint GWAS using summary statistics from individual studies

- Test statistics, e.g., Z-scores, score statistics, effect-sizes with standard deviations (*Cochran's Method; Meta Score Test)*
- P-values (*Fisher's Method)*

Advantages

- Gaining power because of larger sample size
- Avoiding the hassle of combining individual-level data
- Without loss of efficacy under balanced setting (same case-control ratios)

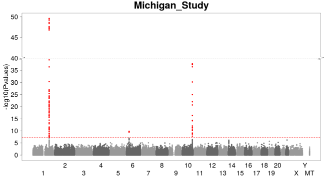
Power Loss Under Unbalanced Setting

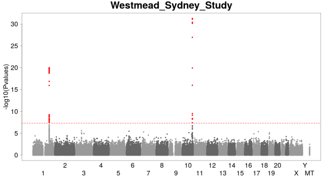
Current strategies

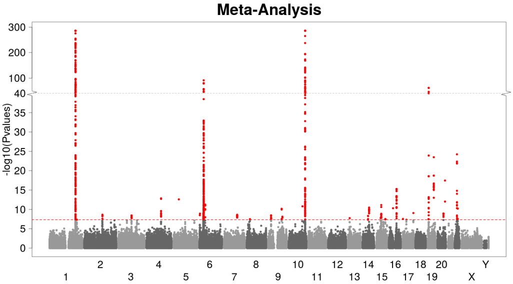
- Weight by effective sample sizes
- Weight by inverse standard errors of test statistics

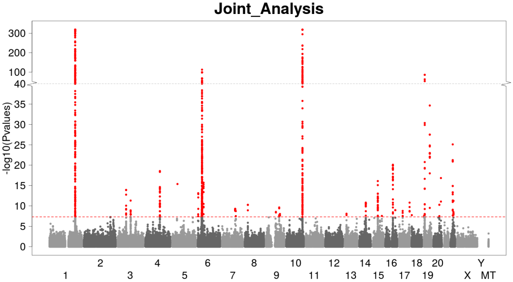
Fail for Gene-level tests based on Score Statistics

- Burden (Madsen & Browning, 2009; Liu et al., 2014)
- SKAT (Lee et al. 2013; Liu et al., 2014)
- Variable Threshold (Price et al., 2010; Liu et al., 2014)









Example two individual studies of AMD.

Score Statistics for Linear Regression Model

- Linear regression model for study k

$$y_k = C_k\alpha_k + X_k\beta_k + \varepsilon_k,\ \varepsilon_k \sim N(0, \sigma_k^2). \qquad (1)$$

- Score statistics

$$\begin{aligned} u_k &= (X_k - \overline{X_k})'(y_k - \widehat{\mu}_k), \\ V_k &= X_k'(\widehat{P}_k - \widehat{P}_k C_k (C_k' \widehat{P}_k C_k)^{-1} C_k' \widehat{P}_k) X_k, \end{aligned}$$

- where

$$\widehat{\mu}_k = C_k \widehat{\alpha_k},$$

$$\widehat{P}_k = \widehat{\sigma_k^2} I_k.$$

Estimates for Meta Score Statistics

- Joint analysis

$$u_{joint} = (X - \overline{X})'(y - \widetilde{\mu}), \; V_{joint} = X'(\widetilde{P} - \widetilde{P}C(C'\widetilde{P}C)^{-1}C'\widetilde{P})X.$$

- Current standard meta-analysis method

$$u_{std} = \sum_{k=1}^{K} u_k, \; V_{std} = \sum_{k=1}^{K} V_k.$$

- Our adjusted estimates

$$u_{adj} \quad = \quad \sum_{k=1}^{K} u_k - \sum_{k=1}^{K} 2n_k\delta_k(f - f_k), \; V_{adj} = \widetilde{\sigma^2}\left[\sum_{k=1}^{K}\left(\frac{V_k}{\widehat{\sigma_k^2}}\right) - \sum_{k=1}^{K} 4n_k(ff' - f_kf_k')\right],$$

where $\delta_k = \widetilde{\mu} - \widehat{\mu_k}$, $\widetilde{\sigma^2} = \frac{1}{n-1}\sum_{k=1}^{K}\left[(n_k - 1)\widehat{\sigma_k^2} + n_k\delta_k^2\right]$.

Improved Estimates for Meta Score Statistics

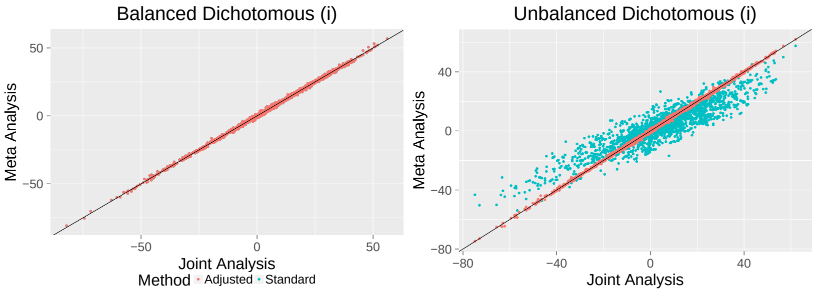


Figure 2: Simulations without population stratification.

$-log10$(P-values) of Single-Variant Meta Score Tests

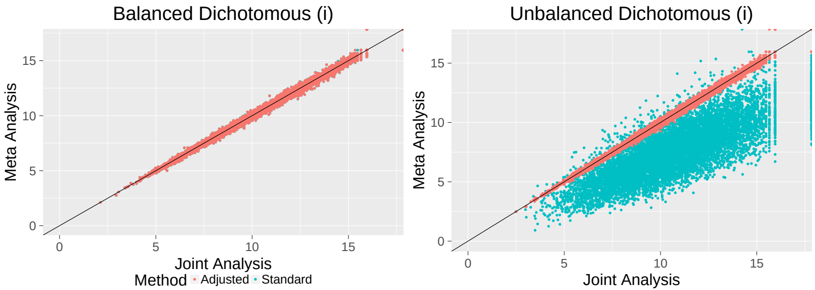
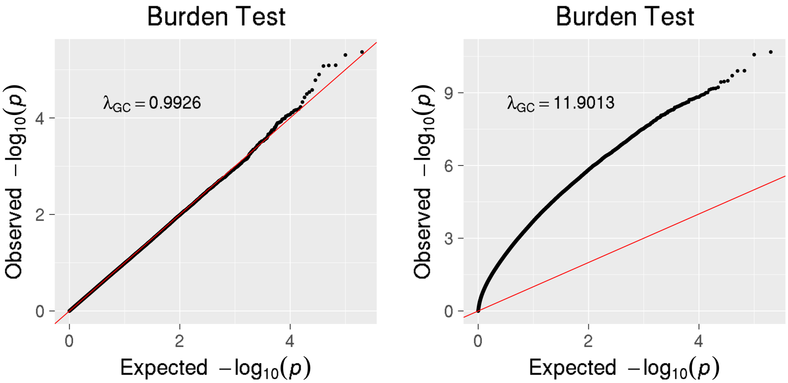


Figure 3: Simulations without population stratification.

Side Effect with Population Stratification



(a) Standard Method (b) Adjusted Method

Figure 4: Quantile-Quantile (QQ) plots of 20,000 null simulations.

Adjusting for Population Stratification

Recall our adjusted formulas for score statistics:

$$u_{adj} = \sum_{k=1}^{K} u_k - \sum_{k=1}^{K} 2n_k \delta_k (f - f_k), \; V_{adj} = \widetilde{\sigma^2} \left[\sum_{k=1}^{K} \left(\frac{V_k}{\widehat{\sigma_k^2}} \right) - \sum_{k=1}^{K} 4n_k (ff' - f_k f_k') \right].$$

First, regress $f_k \sim$ known population MAFs

$$f_k = \sum_{pop} \gamma_{pop} f_{pop} + \varepsilon.$$

Requirements:

- Phenotypes are of the same metrics, or distributions (i.e., δ_k dose not contain population differences)
- Good reference panel with accurate population MAFs f_{pop}

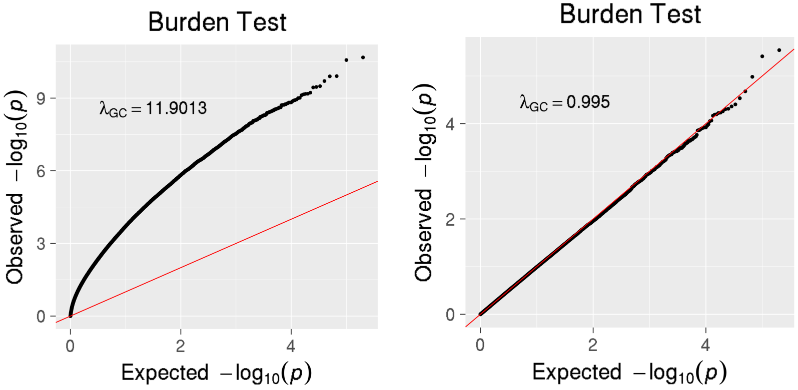
Adjusting for Population Stratification

- Replace f_k by

$$\zeta_k = f_k - \widehat{f_k},\ \widehat{f_k} = \sum_{pop} \widehat{\gamma_{pop}} f_{pop}$$

  and replace f by $\overline{\zeta} = \frac{\sum_{k=1}^{K} n_k \zeta_k}{\sum_{k=1}^{K} n_k}$ in our adjusted formulas.
- Set ζ_{ki} at 0 for variants without corresponding population MAFs, or with $\widehat{f_{ki}}$ falling outside of the 95% prediction confidence interval

Successfully Adjust for Population Stratification



(a) Without adjustment (b) With Adjustment

Figure 5: Quantile-Quantile (QQ) plots of 20,000 null simulations.

Simulation Studies

Considered 5 individual studies, each with sample size 600 (cases, controls)

	Study 1	Study 2	Study 3	Study 4	Study 5
Balanced	(300, 300)	(300, 300)	(300, 300)	(300, 300)	(300, 300)
Unbalanced	(60, 540)	(180, 420)	(300, 300)	(420, 180)	(540, 60)

- Considered **without and with population stratification**
- Simulated genotypes in a 5KB region, 80% MAFs < 5%
- Repeated null simulations for empirical **Type I Errors**
- Compared **power** for gene-level **Burden** and **SKAT** tests

Empirical Type I Errors with $\alpha = 2.5 \times 10^{-6}$

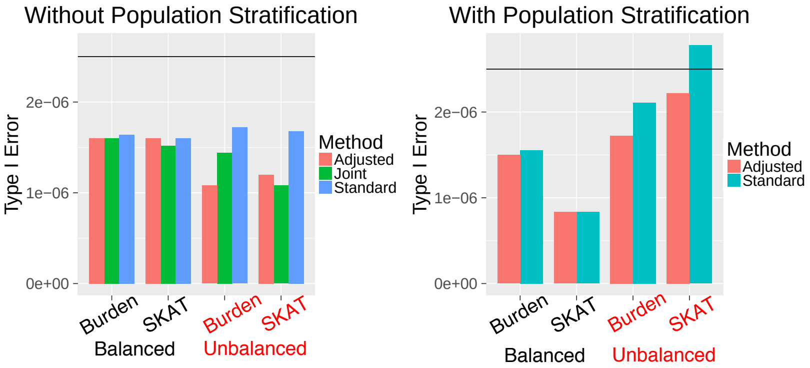


Figure 6: Type I errors are well controlled by our meta-analysis methods under all scenarios.

Power Comparison

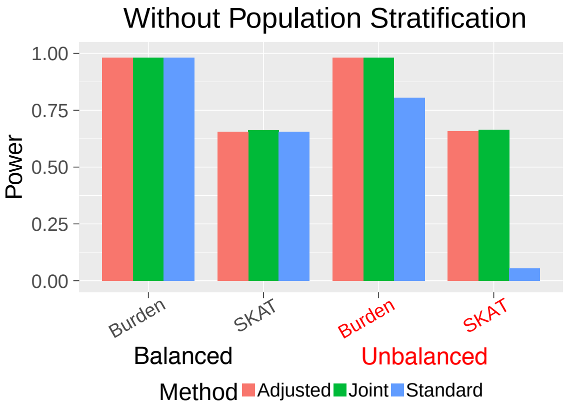


Figure 7: Our method has equivalent power as joint analysis under unbalanced designs.

Power Comparison

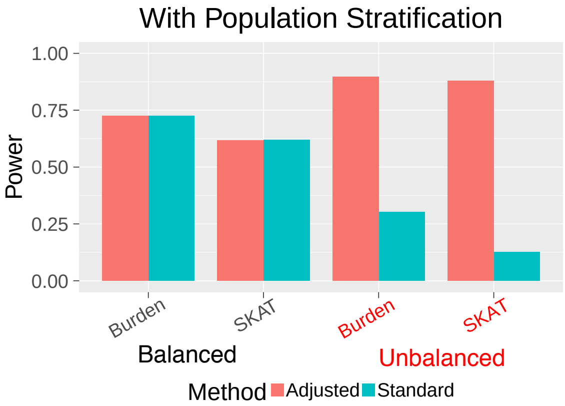


Figure 8: Our method has higher power than standard meta-analysis method under unbalanced designs.

AMD Study

- Consisted with 26 individual studies (IAMDGC) with various case-control ratios (Fritsche et al., 2016)
- European ancestry samples (33,976) without population stratification
- Analyzed rare coding variants only, with optimal MAF threshold given by Variable Threshold (VT) test
- Adjusted for independent common signals and covariates

Gene-level Association Studies

Burden tests on 3 known AMD risk loci

Gene	Joint VT	Std Meta Burden	Adj Meta Burden	Joint Burden
CFH	1.2×10^{-6}	3.2×10^{-5}	$\mathbf{2.1 \times 10^{-6}}$	$\mathbf{2.4 \times 10^{-7}}$
CFI	1.0×10^{-8}	9.6×10^{-10}	$\mathbf{3.3 \times 10^{-14}}$	$\mathbf{8.9 \times 10^{-15}}$
TIMP3	9.0×10^{-8}	9.8×10^{-4}	$\mathbf{1.0 \times 10^{-5}}$	$\mathbf{1.8 \times 10^{-5}}$

Table 1: P-values of Joint VT (Fritsche et al., 2016), Standard (Std) Meta Burden, our Adjusted (Adj) Meta Burden, and Joint Burden tests (Madsen & Browning, 2009).

Single Variant Association Studies of T2D

- Three individual studies of type 2 diabetes (T2D):

Study	FUSION	METSIM	MGI
Population	Finnish	Finnish	American European
Cases	1142	673	1942
Controls	155	2667	14553

- Consider genotyped variants in METSIM
- Jointly correct phenotype for Age, Gender, BMI, PC1-4
- Use 1000 Genome as reference panel for adjusting population stratification

Figure 9: Top two PCs show population stratification with these three studies.

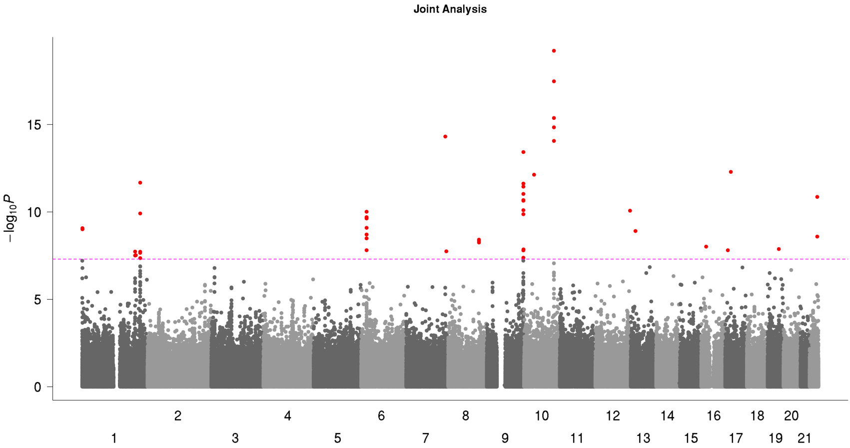


Figure 10: Joint analysis results with inflated false positives.

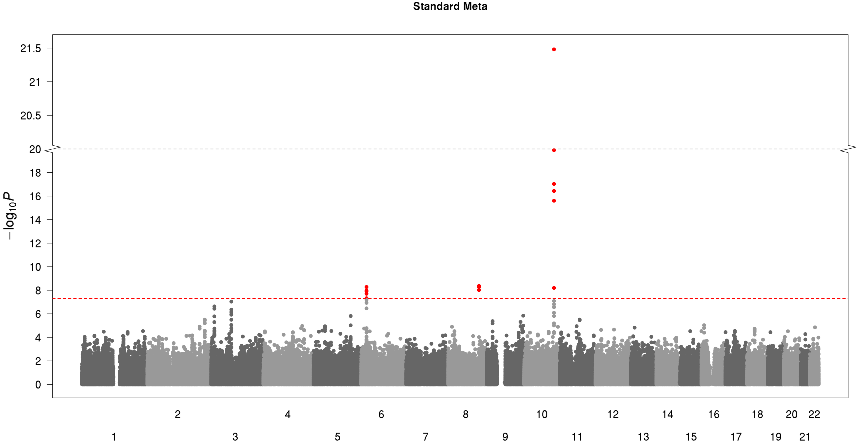


Figure 11: Standard meta-analysis results with power loss.

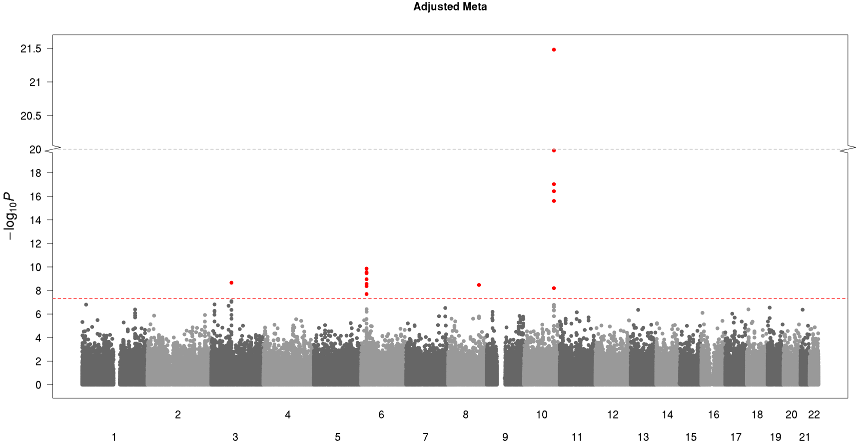


Figure 12: Meta-analysis results by our method with adjustment for population stratification.

Summary

- Improved estimates for meta score statistics
- Novel strategy adjusting for population stratification
- Suitable for both single-variant and gene-level association studies
- Ensure the efficiency of meta-analysis under general settings
- Require phenotypes of the same distribution and good reference panel

Acknowledgements

Michigan Genomics Initiative (MGI), FUSION, METSIM

illumina®

EMORY UNIVERSITY SCHOOL OF MEDICINE

UNIVERSITY OF MICHIGAN

International AMD Genomics Consortium