Quality Control of GWAS Summary Statistics

Florian Privé

Aarhus Univ, Denmark

privefl

GWAS summary statistics

- $\hat{\gamma}_j$ — the GWAS effect size of variant j (marginal effect),
- $\text{se}(\hat{\gamma}_j)$ — its standard error,
- $z_j = \frac{\hat{\gamma}_j}{\text{se}(\hat{\gamma}_j)}$ — the Z-score of variant j,
- n_j — the GWAS sample size associated with variant j,
- f_j — the allele frequency of variant j,
- INFO_j — the imputation INFO score of variant j

The first quality control I already recommend

Compare standard deviations of genotypes estimated in 2 ways:

1. - When linear regression was used

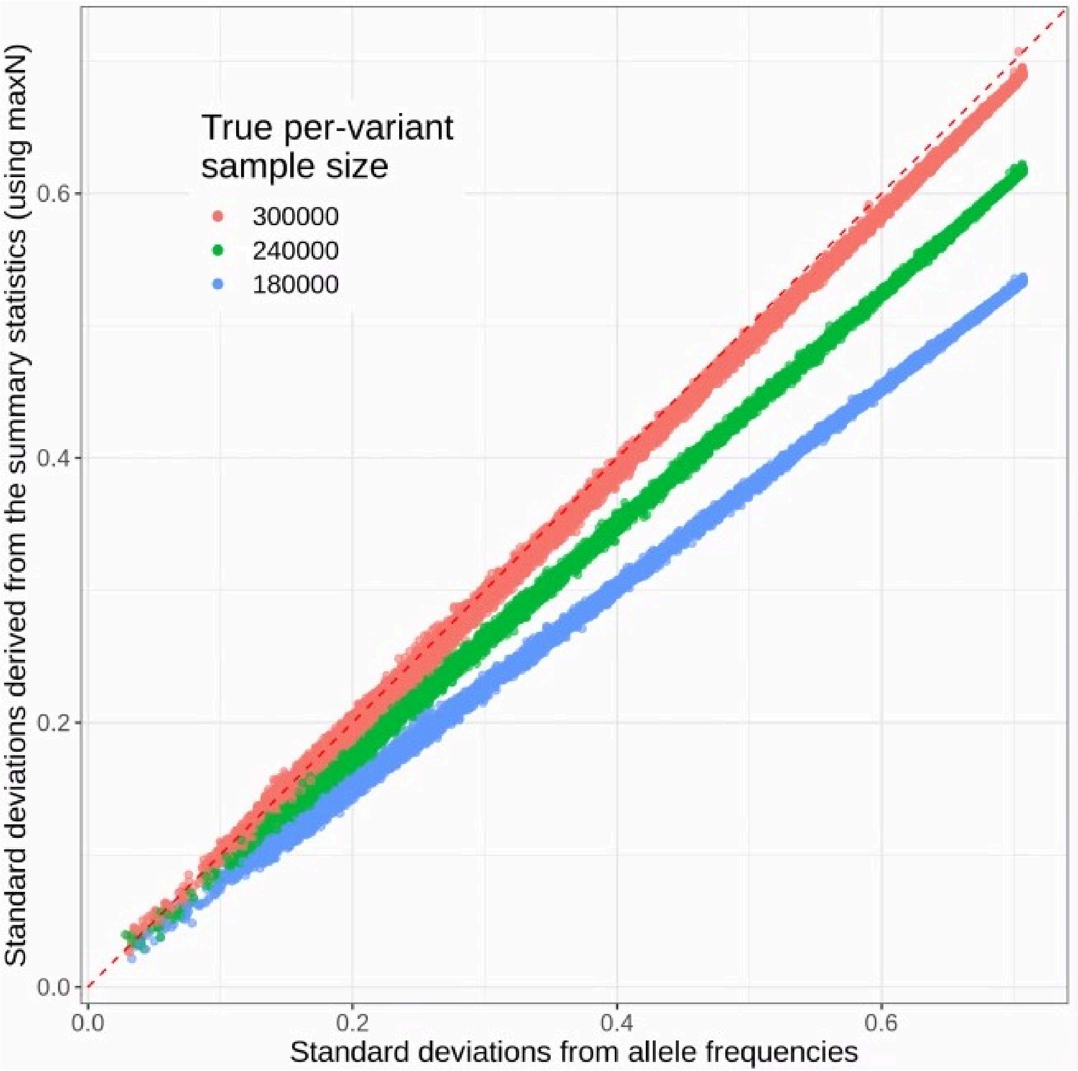
   $$\text{sd}(G_j) \approx \frac{\text{sd}(y)}{\sqrt{n_j \cdot \text{se}(\hat{\gamma}_j)^2 + \hat{\gamma}_j^2}}$$

   - When logistic regression was used (case-control phenotype)

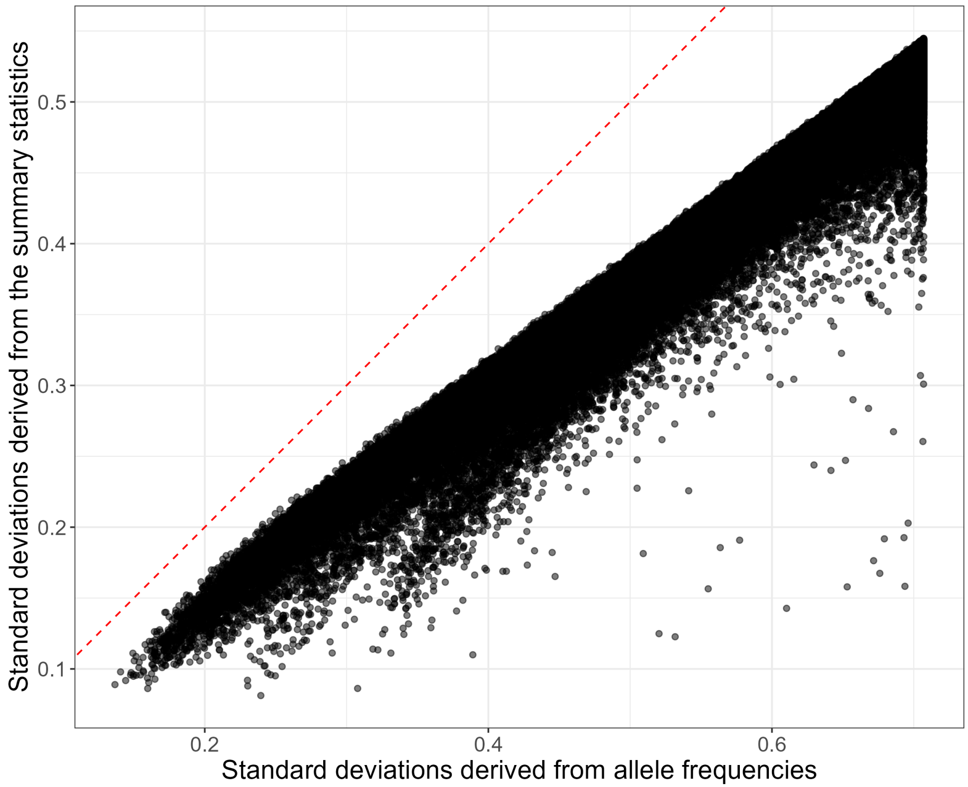
   $$\text{sd}(G_j) \approx \frac{2}{\sqrt{n_j^{\text{eff}} \cdot \text{se}(\hat{\gamma}_j)^2 + \hat{\gamma}_j^2}}$$

2. $$\text{sd}(G_j) \approx \sqrt{2 \cdot f_j \cdot (1 - f_j) \cdot \text{INFO}_j}$$

Detect differences in per-variant GWAS sample sizes



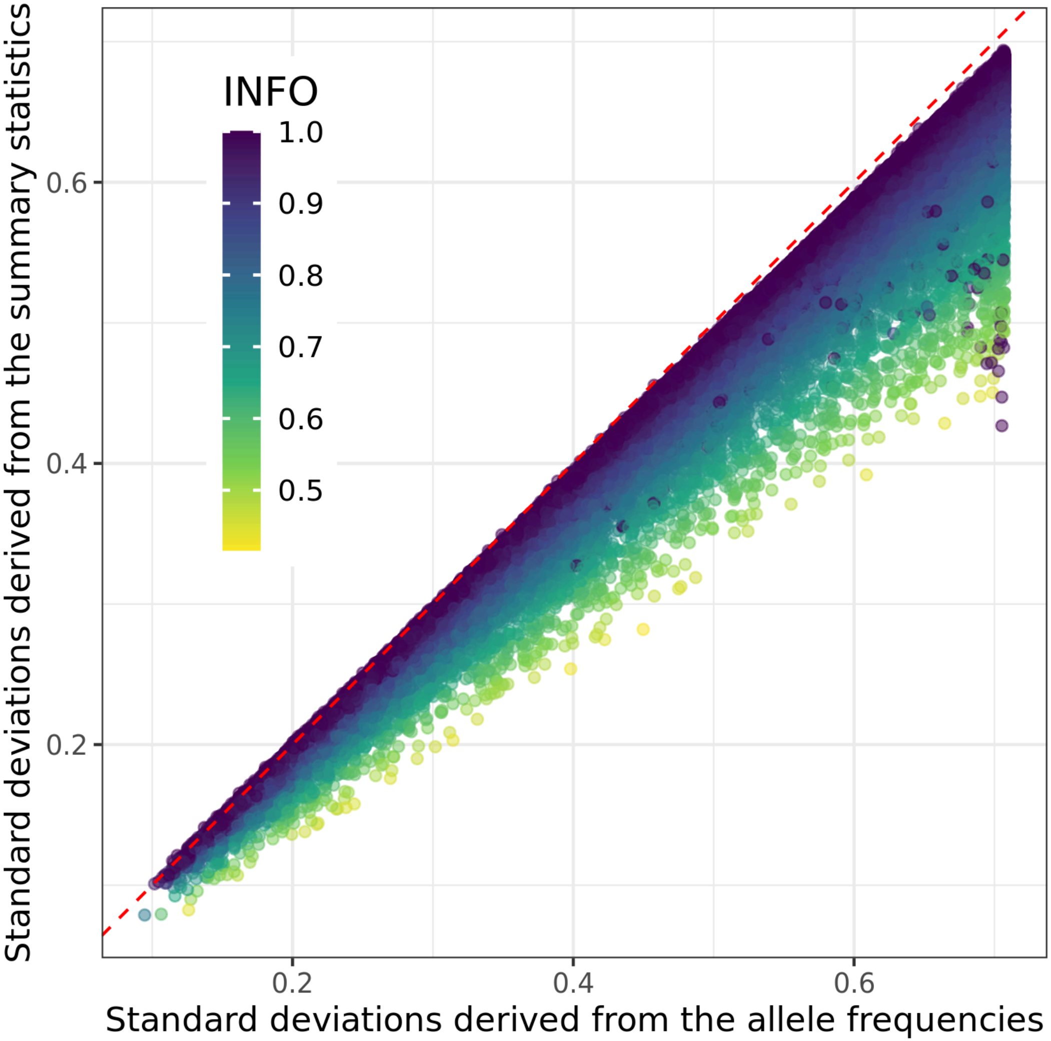
Detect bias in total effective GWAS sample size



$$N_{\mathrm{eff}} = \frac{4}{1/N_{\mathrm{ca}} + 1/N_{\mathrm{co}}}$$

CAD study	Nca	Nco	Neff
1	5719	6545	12208
2	206	259	459
3	278	312	588
4	505	1021	1352
5	392	410	802
6	1010	3998	3225
7	1628	368	1201
8	2083	2048	4131
9	1216	653	1699
10	658	5841	2366
11	1802	466	1481
12	2099	2690	4716
13	634	1608	1819
14	1207	1288	2492
15	1061	1467	2463
16	1089	1147	2234
17	877	2187	2504
18	2700	2758	5457
19	361	2778	1278
20	487	1381	1440
21	758	3337	2471
22	2791	3757	6405
23	2095	503	1622
24	933	468	1247
25	2905	2998	5902
26	947	1008	1953
27	1294	1529	2803
28	843	318	924
29	933	468	1247
30	119	830	416
31	631	334	874
32	836	761	1593
33	426	594	992
34	814	5999	2867
35	322	857	936
36	1926	2938	4653
37	4651	4452	9099
38	4380	3929	8285
39	1535	772	2055
40	1007	22286	3854
41	402	448	848
42	745	1389	1940
43	397	2474	1368
44	506	5335	1849
45	259	4202	976
46	334	3446	1218
47	2034	3210	4980
48	454	8443	1723
Total	**61289**	**126310**	
Total Neff	**165063**		**129015**

Detect low imputation INFO scores



Read more about this

- Privé, F., et al. (2022) "Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores." *Human Genetics and Genomics Advances* 3.4.

- Grotzinger, A.D., et al. (2023) "Pervasive downward bias in estimates of liability-scale heritability in genome-wide association study meta-analysis: a simple solution." *Biological Psychiatry* 93.1.

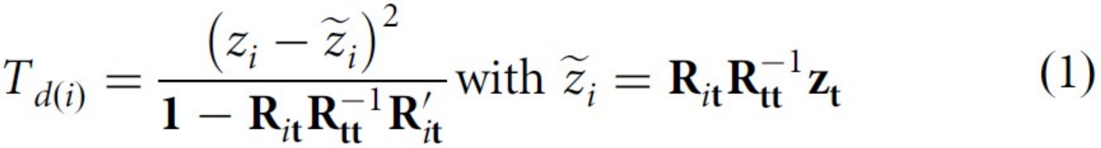
- Gazal, S., et al. (2018) "Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations." *Nature Genetics* 50.11.

- Privé, F. (2022) "Using the UK Biobank as a global reference of worldwide populations: application to measuring ancestry diversity from GWAS summary statistics." *Bioinformatics* 38.13.

Additional (complementary) QC — DENTIST methodology

GCTA method which compares reported Z-scores with imputed Z-scores.

$\chi^2(1)$ test statistic:



$$T_{d(i)} = \frac{\left(z_i - \widetilde{z}_i\right)^2}{\mathbf{1} - \mathbf{R}_{it}\mathbf{R}_{\mathbf{tt}}^{-1}\mathbf{R}_{it}'} \text{ with } \widetilde{z}_i = \mathbf{R}_{it}\mathbf{R}_{\mathbf{tt}}^{-1}\mathbf{z}_{\mathbf{t}} \tag{1}$$

where i is the variant of interest, and t the variants used for imputing.

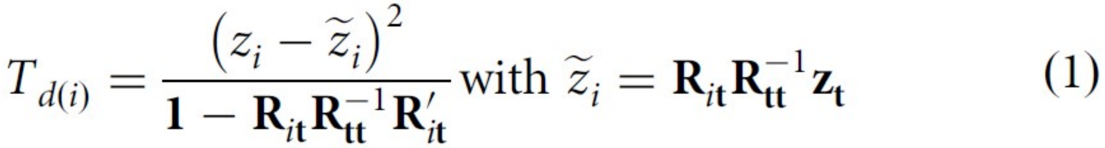
It is particularly good at detecting allelic errors (opposite effect).

DENTIST citation: Chen, W., et al. (2021) "Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors." *Nature Communications* 12.1.

Additional (complementary) QC — DENTIST methodology

GCTA method which compares reported Z-scores with imputed Z-scores.

$\chi^2(1)$ test statistic:



$$T_{d(i)} = \frac{\left(z_i - \widetilde{z}_i\right)^2}{\mathbf{1} - \mathbf{R}_{it}\mathbf{R}_{\mathbf{tt}}^{-1}\mathbf{R}_{it}'} \text{ with } \widetilde{z}_i = \mathbf{R}_{it}\mathbf{R}_{\mathbf{tt}}^{-1}\mathbf{z}_{\mathbf{t}} \qquad (1)$$

where i is the variant of interest, and t the variants used for imputing.

It is particularly good at detecting allelic errors (opposite effect).

DENTIST citation: Chen, W., et al. (2021) "Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors." *Nature Communications* 12.1.

Quick simulation results

Design:

- Use 145K variants on chromosome 22 with MAF > 0.005 and INFO > 0.8
- Simulate some phenotype with heritability of 0.1 and polygenicity of 0.01
- Compute the GWAS summary statistics using N=50K
  (Z-scores in [-20; 20], mostly in [-10; 10])
- For 1000 variants at random, assign them an opposite effect (allelic error)

Quick simulation results

Design:

- Use 145K variants on chromosome 22 with MAF > 0.005 and INFO > 0.8
- Simulate some phenotype with heritability of 0.1 and polygenicity of 0.01
- Compute the GWAS summary statistics using N=50K
  (Z-scores in [-20; 20], mostly in [-10; 10])
- For 1000 variants at random, assign them an opposite effect (allelic error)

Results:

- 802 true positives (TP, real errors) and 3209 false positives (FP)
  with DENTIST

Quick simulation results

Design:

- Use 145K variants on chromosome 22 with MAF > 0.005 and INFO > 0.8
- Simulate some phenotype with heritability of 0.1 and polygenicity of 0.01
- Compute the GWAS summary statistics using N=50K
  (Z-scores in [-20; 20], mostly in [-10; 10])
- For 1000 variants at random, assign them an opposite effect (allelic error)

Results:

- 802 true positives (TP, real errors) and 3209 false positives (FP)
  with DENTIST
- vs 686 TP and 9 FP
  with my alternative methodology

Current project

- Check and improve the DENTIST methodology,
  to ideally get more power and less false positive

- As an R implementation

- [I NEED YOUR HELP]
  Do you have/know GWAS summary statistics with allelic errors?

Current project

- Check and improve the DENTIST methodology,
  to ideally get more power and less false positive
- As an R implementation
- [I NEED YOUR HELP]
  Do you have/know GWAS summary statistics with allelic errors?

Part of a larger project

- Provide some very well quality-controlled GWAS summary statistics
- In a standardized format
- Probably as a GitHub repo of R scripts,
  where each script processes a specific GWAS summary statistics file

Take-home messages

- There can be many issues in GWAS summary statistics

Take-home messages

- There can be many issues in GWAS summary statistics
- You can detect many of them by comparing SDs estimated in two ways

Take-home messages

- There can be many issues in GWAS summary statistics
- You can detect many of them by comparing SDs estimated in two ways
- You can detect other (complementary) issues with DENTIST

Take-home messages

- There can be many issues in GWAS summary statistics
- You can detect many of them by comparing SDs estimated in two ways
- You can detect other (complementary) issues with DENTIST
- DENTIST is currently prone to false positives; it needs to be improved

Take-home messages

- There can be many issues in GWAS summary statistics
- You can detect many of them by comparing SDs estimated in two ways
- You can detect other (complementary) issues with DENTIST
- DENTIST is currently prone to false positives; it needs to be improved
- I hope to provide QCed GWAS summary statistics for everyone to use

Thank you for your attention

Presentation available at bit.ly/qc_sumstats_EMGM

privefl

Slides created via the R package **xaringan**