Supplementary material for

Multiple-testing correction in metabolome-wide association studies

Alina Peluso [1,*], Robert Glen[1,2,*], and Timothy M D Ebbels[1,*]

July 28, 2020

[1] Faculty of Medicine, Department of Surgery & Cancer, Imperial College London, South Kensington Campus, SW7 2AZ, London, UK.
[2] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, Cambridge, UK.
[*]To whom correspondence should be addressed: {a.peluso, r.glen, t.ebbels@imperial.ac.uk.}

List of Tables

List of Figures

1 Tables

Table 1: Descriptive statistics for the clinical outcome measures.

Outcome	mean	sd	median	min	max	skewness	kurtosis
Glucose (mg/dL)	98.28	31.10	90	38	507	4.17	28.89
Logarithm of Glucose	4.56	0.23	4.5	3.64	6.23	2.22	10.35
BMI (kg/m^2)	28.14	5.39	27.34	15.36	61.86	46.50	4.45
Logarithm of BMI	3.32	0.18	3.31	2.73	4.12	1.39	3.20

Table 2: Descriptive statistics for the fixed effects covariates.

Covariate	mean	sd
Age (years)	62.89	10.32
Gender	0.51	0.49
Height (cm)	166.43	10.23
Ethnicity: Caucasian	0.39	0.49
Ethnicity: Hispanic	0.23	0.42
Ethnicity: African-American	0.25	0.43
Ethnicity: Chinese-American	0.13	0.34
Smoking: Never	0.51	0.50
Smoking: Former	0.12	0.33
Smoking: Current	0.38	0.48
LDL cholesterol (mg/dL)	117.67	31.04
HDL cholesterol (mg/dL)	51.29	14.41
Systolic blood pressure (mmHg)	126.92	21.54
Blood pressure treatment	0.38	0.49
Diabetes	0.14	0.34
Lipids treatment	0.17	0.37

Table 3: Real data: Comparison of estimation of the number of non-redundant variates from the permutation method (ENT, obtained as the average of the ENT estimates for all the clinical outcomes measures considered via the multivariate Normal and the multivariate log-Normal methods) and via approximation procedures based on the eigenvalues of the correlation matrix of the metabolite concentrations (Meff). R=ENT/ANT(%) ratio in brackets. ANT=655 for the MESA BINNED data, and ANT=30590 for the NOESY and CPMG data. Meff estimates closest to ENT estimates in bold.

	MESA BINNED	MESA NOESY	MESA CPMG
ENT	352(53.8%)	2744(9.0%)	16014(52.3%)
Meff.MWSL	**345(52.7%)**	1931(6.3%)	**11570(37.8%)**
Meff.Galwey [1]	201(30.7%)	524(1.7%)	1815(5.9%)
Meff.Gao [2]	435(66.4%)	**2705(8.8%)**	3537(11.6%)
Meff.Liji [3]	226(34.5%)	2534(8.3%)	4972(16.3%)
Meff.Nyholt [4]	611(93.3%)	26823(87.7%)	29704(97.1%)

Table 4: Real data: MWSL estimation comparison between the permutation method and the approximation procedure generating the distribution of the minimum p-value as a Beta(1,$\mathrm{Meff}_{\mathrm{MWSL}}$).

MESA BINNED			MESA NOESY			MESA CPMG		
	Beta(1,345)	Permutation		Beta(1,1931)	Permutation		Beta(1,11570)	Permutation
ENT/ANT(%)	**51.4%**	**53.8%**	**ENT/ANT(%)**	**6.2%**	**9.0%**	**ENT/ANT(%)**	**36.9%**	**52.3%**
ENT	336	352	**ENT**	1883	2744	**ENT**	11279	16014
ENT_CI.low	335	338	ENT_CI.low	1874	2439	ENT_CI.low	11229	14509
ENT_CI.up	338	369	ENT_CI.up	1891	3080	ENT_CI.up	11330	17208
MWSL	**0.0001486**	**0.0001410**	**MWSL**	**0.0000266**	**0.0000179**	**MWSL**	**0.0000044**	**0.0000031**
MWSL_CI.up	0.0001493	0.0001476	MWSL_CI.up	0.0000267	0.0000195	MWSL_CI.up	0.0000045	0.0000036
MWSL_CI.low	0.0001480	0.0001340	MWSL_CI.low	0.0000264	0.0000160	MWSL_CI.low	0.0000044	0.0000028

Table 5: Simulated data: Comparison of estimation of the number of non-redundant variates from the permutation method (ENT, obtained as the average of the ENT estimates for all the simulated uncorrelated and correlated outcomes measures considered via the multivariate Normal and the multivariate log-Normal methods) and via the approximation procedure based on the eigenvalues of the correlation matrix of the metabolite concentrations (Meff). R=ENT/ANT(%) ratio in brackets. ANT=650. Meff estimates closest to ENT estimates in bold.

Correlation ∈	[0.95-1)	[0.85,0.95)	[0.75,0.85)	[0.65,0.75)	[0.55,0.65)	[0.45,0.55)	[0.35,0.45)	[0.25,0.35)	(0,0.25)
ENT	3(0.5%)	11(1.7%)	36(5.4%)	81(12.4%)	134(20.5%)	210(32%)	319(48.7%)	433(66.1%)	554(84.6%)
Meff.MWSL	**4(0.6%)**	**14(2.2%)**	**40(6.2%)**	**74(11.4%)**	**126(19.4%)**	195(30%)	**282(43.4%)**	**416(64%)**	416(64%)
Meff.Nyholt	34(5.2%)	121(18.6%)	239(36.8%)	331(50.9%)	419(64.5%)	491(75.5%)	543(83.5%)	588(90.5%)	638(98.2%)
Meff.Liji	18(2.8%)	64(9.8%)	133(20.5%)	195(30%)	263(40.5%)	320(49.2%)	365(56.2%)	413(63.5%)	382(58.8%)
Meff.Gao	65(10%)	131(20.2%)	213(32.8%)	275(42.3%)	327(50.3%)	372(57.2%)	412(63.4%)	449(69.1%)	317(48.8%)
Meff.Galwey	7(1.1%)	23(3.5%)	58(8.9%)	98(15.1%)	147(22.6%)	**200(30.8%)**	252(38.8%)	312(48%)	270(41.5%)

Table 6: Simulated data: MWSL estimation comparison between the permutation method and the approximation procedure generating the distribution of the minimum p-value as a Beta(1,$\mathrm{Meff}_{\mathrm{MWSL}}$).

Correlation ∈ **[0.95-1)**			Correlation ∈ **[0.65,0.75)**			Correlation ∈ **[0.35,0.45)**		
	Beta(1,4)	Permutation		Beta(1,74)	Permutation		Beta(1,282)	Permutation
ENT/ANT(%)	**0.60%**	**0.49%**	**ENT/ANT(%)**	**11.10%**	**12.48%**	**ENT/ANT(%)**	**42.30%**	**49.06%**
ENT	4	3	**ENT**	72	81	**ENT**	275	319
ENT_CI.low	4	3	ENT_CI.low	72	67	ENT_CI.low	274	293
ENT_CI.up	4	3	ENT_CI.up	72	93	ENT_CI.up	276	343
MWSL	**0.012740**	**0.015879**	**MWSL**	**0.000693**	**0.000620**	**MWSL**	**0.000182**	**0.000155**
MWSL_CI.up	0.012797	0.017769	MWSL_CI.up	0.000696	0.000710	MWSL_CI.up	0.000183	0.000170
MWSL_CI.low	0.012684	0.014545	MWSL_CI.low	0.000690	0.000536	MWSL_CI.low	0.000181	0.000138
Correlation ∈ **[0.85,0.95)**			Correlation ∈ **[0.55,0.65)**			Correlation ∈ **[0.25,0.35)**		
	Beta(1,14)	Permutation		Beta(1,126)	Permutation		Beta(1,416)	Permutation
ENT/ANT(%)	**2.10%**	**1.73%**	**ENT/ANT(%)**	**18.90%**	**20.68%**	**ENT/ANT(%)**	**62.37%**	**66.60%**
ENT	14	11	**ENT**	123	134	**ENT**	405	433
ENT_CI.low	14	10	ENT_CI.low	122	125	ENT_CI.low	404	440
ENT_CI.up	14	13	ENT_CI.up	123	143	ENT_CI.up	407	433
MWSL	**0.003657**	**0.004482**	**MWSL**	**0.000407**	**0.000373**	**MWSL**	**0.000123**	**0.000108**
MWSL_CI.up	0.003673	0.004964	MWSL_CI.up	0.000409	0.000410	MWSL_CI.up	0.000124	0.000117
MWSL_CI.low	0.003641	0.003900	MWSL_CI.low	0.000405	0.000328	MWSL_CI.low	0.000123	0.000097
Correlation ∈ **[0.75,0.85)**			Correlation ∈ **[0.45,0.55)**			Correlation ∈ **(0,0.25)**		
	Beta(1,40)	Permutation		Beta(1,195)	Permutation		Beta(1,535)	Permutation
ENT/ANT(%)	**6.00%**	**5.49%**	**ENT/ANT(%)**	**29.24%**	**32.29%**	**ENT/ANT(%)**	**80.23%**	**85.26%**
ENT	39	36	**ENT**	190	210	**ENT**	522	554
ENT_CI.low	39	32	ENT_CI.low	189	192	ENT_CI.low	519	526
ENT_CI.up	39	41	ENT_CI.up	191	228	ENT_CI.up	524	586
MWSL	**0.001282**	**0.001596**	**MWSL**	**0.000263**	**0.000221**	**MWSL**	**0.000096**	**0.000088**
MWSL_CI.up	0.001276	0.001796	MWSL_CI.up	0.000264	0.000241	MWSL_CI.up	0.000096	0.000096
MWSL_CI.low	0.001287	0.001386	MWSL_CI.low	0.000262	0.000201	MWSL_CI.low	0.000095	0.000078

Table 7: BINNED data: ENT estimates with 95% confidence intervals in brackets, and type I error estimation from the permutation procedure for various simulated outcome measures: continuous, discrete-binary, discrete-count, time-to-event survival. K=5,000 permutations. ANT=655.

Outcome type	continuous			binary			count			survival		
MESA Binned data	ENT	R (%)	Type I error (%)	ENT	R (%)	Type I error (%)	ENT	R (%)	Type I error (%)	ENT	R (%)	Type I error (%)
identity	482(442;506)	74%	5.16%	409(379;443)	62%	5.11%	221(198;256)	34%	4.71%	466(431;505)	71%	4.97%
multivariate log-Normal	351(316;388)	54%	5.02%	338(310;380)	52%	5.01%	344(305;377)	53%	4.71%	344(309;366)	53%	4.92%
multivariate Normal	376(345;423)	57%	5.04%	355(318;395)	54%	5.21%	366(326;404)	56%	4.69%	361(331;397)	55%	4.81%

Table 8: PCA simulated data (ANT=655, n_t=1,500, PCs=350): ENT estimates with 95% confidence intervals in brackets, and type I error estimation from the permutation procedure for various simulated outcome measures: continuous, discrete-binary, discrete-count, time-to-event survival. K=5,000 permutations.

Outcome type	continuous			binary			count			survival		
PCA simulated data	ENT	R (%)	Type I error (%)	ENT	R (%)	Type I error (%)	ENT	R (%)	Type I error (%)	ENT	R (%)	Type I error (%)
identity	292(271;314)	45%	5.10%	333(307;360)	51%	4.93%	500(469;542)	76%	5.05%	394(371;427)	60%	4.95%
multivariate log-Normal	391(365;417)	60%	4.99%	377(347;401)	58%	5.09%	389(360;414)	59%	4.98%	379(356;410)	58%	4.93%
multivariate Normal	368(340;401)	56%	4.98%	356(332;384)	54%	5.01%	370(345;398)	56%	5.11%	343(320;368)	52%	5.19%

2 Figures

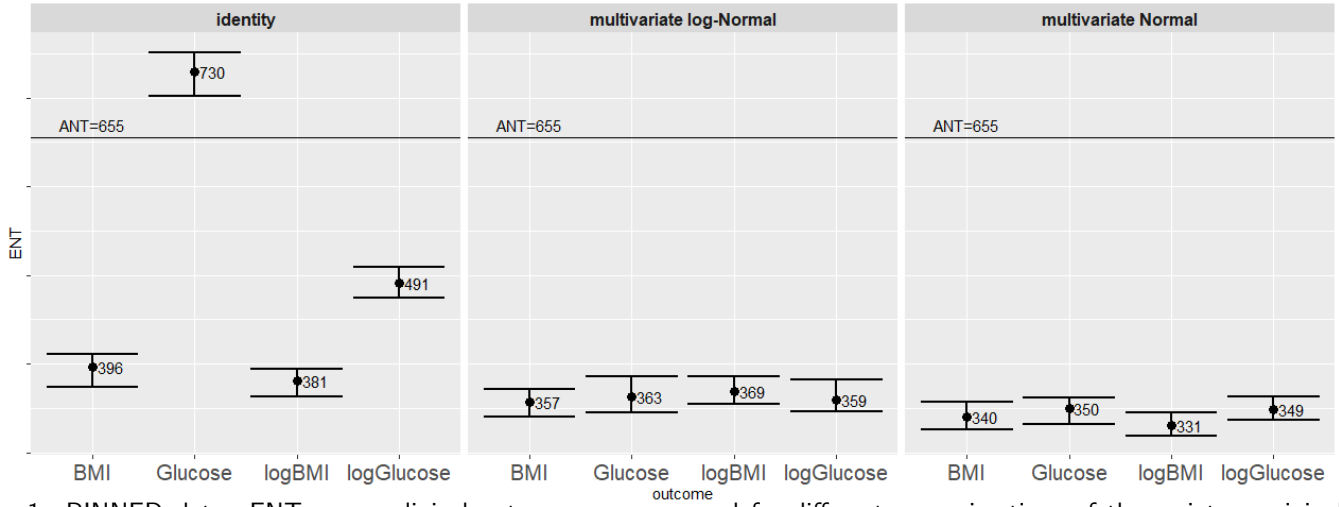

Figure 1: BINNED data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.

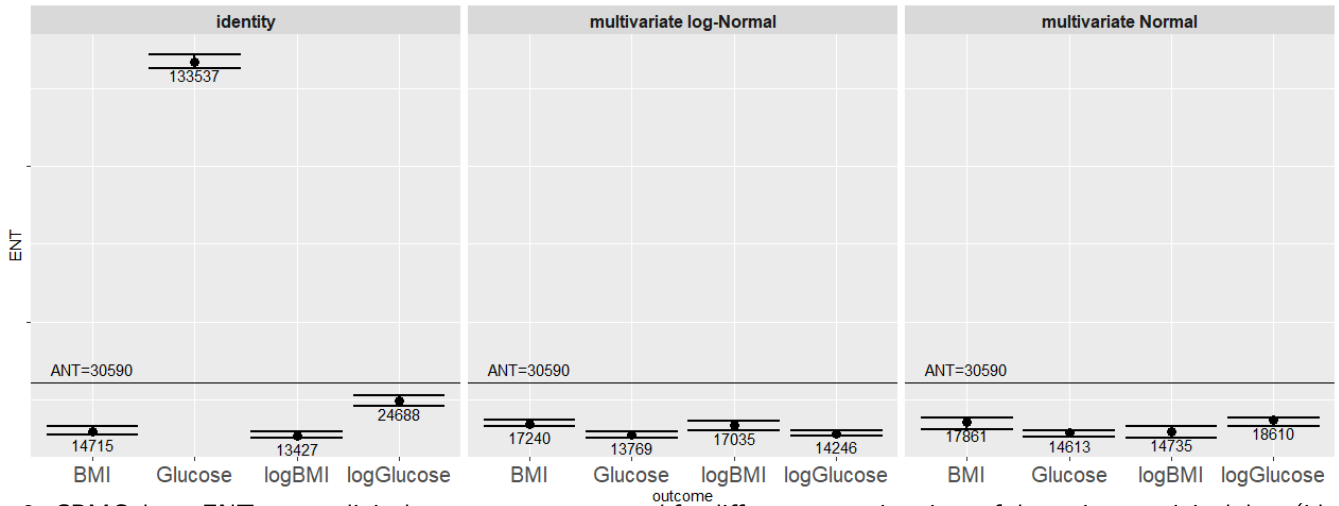

Figure 2: CPMG data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.

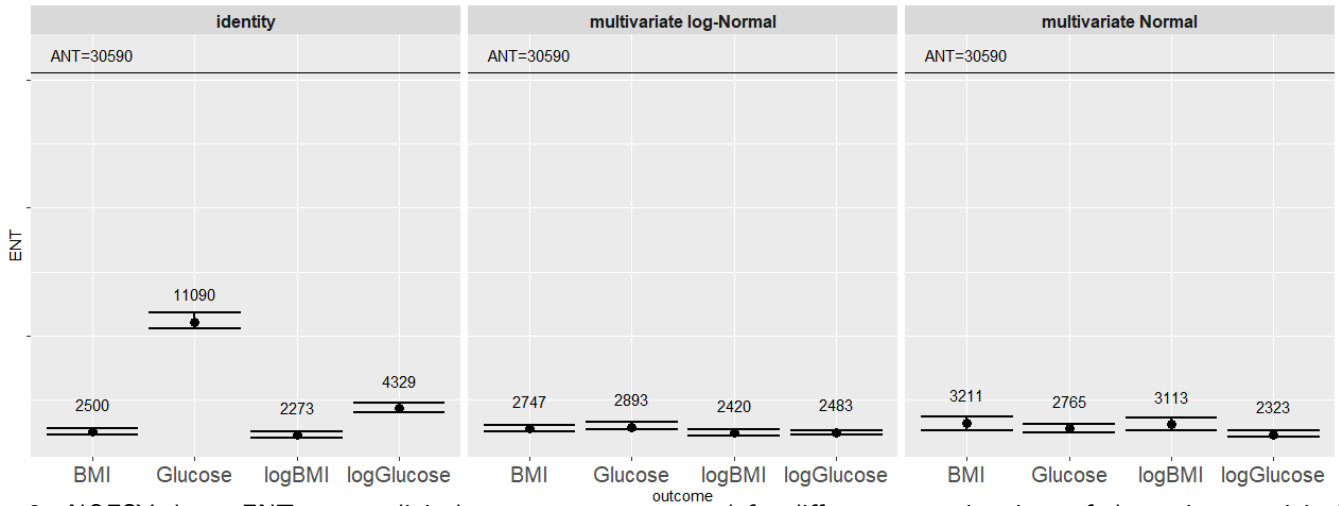

Figure 3: NOESY data: ENT across clinical outcome measures and for different approximations of the variates: original data (identity), multivariate Normal, multivariate log-Normal. Error bars represent 95% confidence limits. K=10,000 permutations.

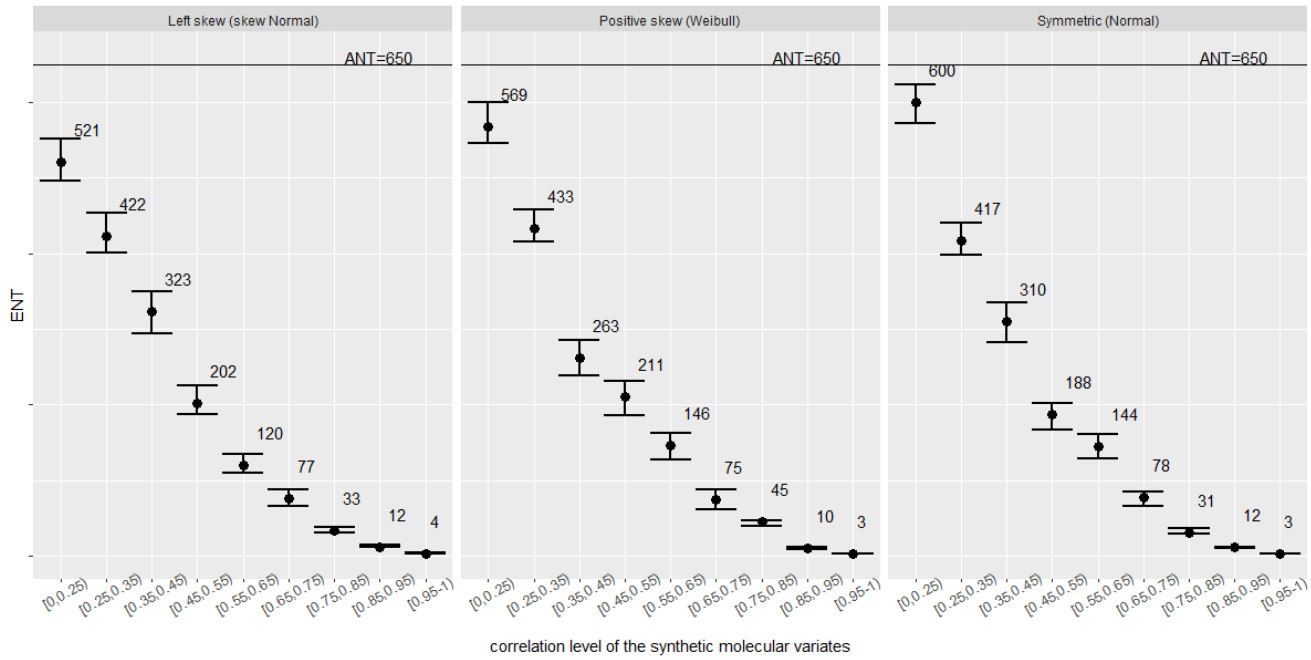

Figure 4: ENT for uncorrelated outcomes across correlated variates. Error bars represent 95% confidence limits. K=5,000 permutations.

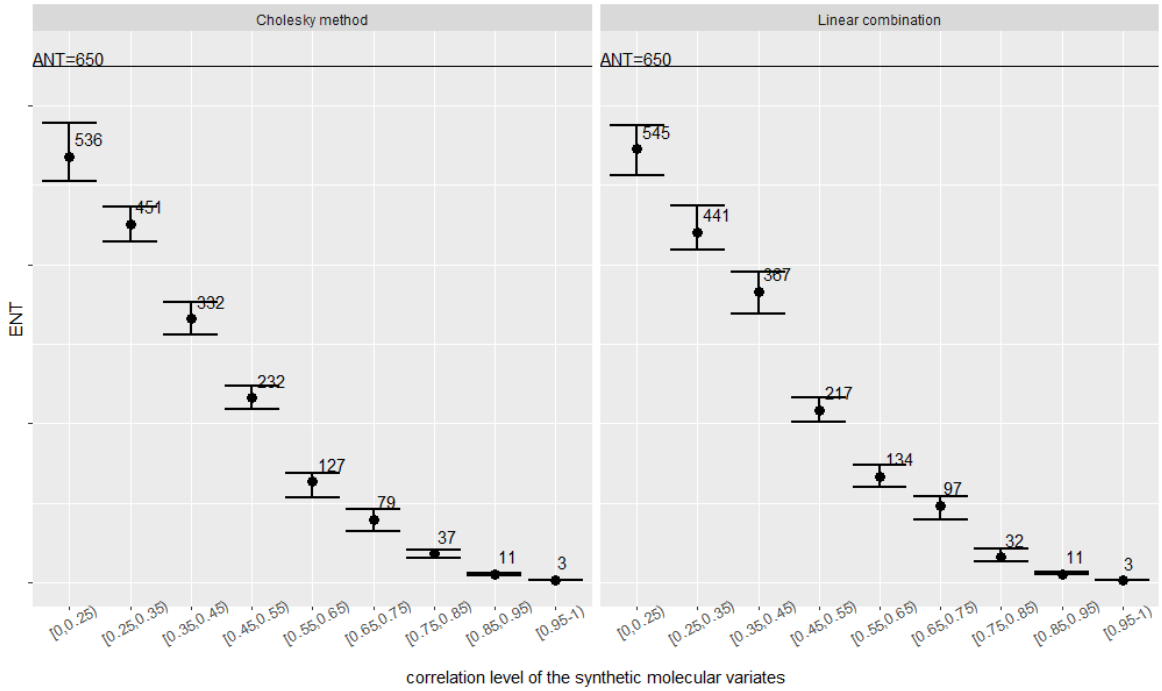

Figure 5: ENT for correlated outcome across correlated variates. Error bars represent 95% confidence limits. K=5,000 permutations.

References

[1] Nicholas W Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(7):559–568, 2009.

[2] Xiaoyi Gao, Joshua Starmer, and Eden R Martin. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic epidemiology*, 32(4):361–369, 2008.

[3] J Li and L Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221, 2005.

[4] Dale R Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.