More about interactions

October 9, 2015

Toy data

Consider the data - a binary response Y, a binary environmental variable E and a binary gene G:

	$Y=1$	
	$G=1$	$G=0$
$E=1$	112	64
$E=0$	112	112

	$Y=0$	
	$G=1$	$G=0$
$E=1$	100	100
$E=0$	100	100

Three testing approaches

$$\text{logit}(Pr(Y = 1|G, E)) = \alpha_0 + \alpha_1 G + \alpha_2 E$$

P-value for $H_0 : \alpha_1 = 0$ is 0.070. Not significant!

$$\text{logit}(Pr(Y = 1|G, E)) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE$$

P-value for $H_0 : \beta_3 = 0$ is 0.051. Not significant!

But....

P-value for $H_0 : \beta_1 = \beta_3 = 0$ is 0.029. Significant!

More toy data

Consider the data - a binary response Y, a binary environmental variable E and a binary gene G:

	$Y=1$	
	$G=1$	$G=0$
$E=1$	115	85
$E=0$	85	115

	$Y=0$	
	$G=1$	$G=0$
$E=1$	100	100
$E=0$	100	100

Three testing approaches

$$\text{logit}(Pr(Y = 1|G, E)) = \alpha_0 + \alpha_1 G + \alpha_2 E$$

P-value for $H_0 : \alpha_1 = 0$ is 1. Not significant!

$$\text{logit}(Pr(Y = 1|G, E)) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE$$

P-value for $H_0 : \beta_3 = 0$ is 0.033. Significant!

But....

P-value for $H_0 : \beta_1 = \beta_3 = 0$ is 0.104. Not significant!

What does this teach us?

- Both the test $H_0 : \beta_3 = 0$ and the test $H_0 : \beta_1 = \beta_3 = 0$ involve the interaction parameter β_3, but one tests the interaction, one tests the genetic effect, in the situation of possible confounders.
- As always, it's important to be aware what the null hypothesis means -
  - Testing whether genes affect the outcome, where the gene effect may depend on the environment.
  - Testing for gene-environment interaction.

  Unfortunately not all genetic epi papers are that careful...
- Today focus on the interaction test $H_0 : \beta_3 = 0$.

Note: just as for the interaction test, testing for genetic effect in the situation of possible confounders can exploit G-E independence. See Dai et al. (2012), *Am J Epi* **176**:164–173 and the references therein.

There are some papers that have found genes testing for (G + GE), e.g.

- Gauderman and Siegmund (2001) *Hum Herid* **52**:34–46.
- Selinger-Leneman et al. (2003) *Gen Epi* **24**:200–7.
- Kraft et al. (2007) *Hum Herid* **63**:111-9.
- Huang et al. (2011) *Genome Med* **3**:42.

Power and sample size

Number of case-control pairs required for 80% power.

	GxE interaction			G main effect	
		$P(E=1)$			
$P(G=1)$	$\exp(\beta_3)$	0.1	0.5	$\exp(\beta_1)$	
0.05	1.5	19571	7520	1.5	860
	2.0	6263	2495	2.0	266
0.40	1.5	4107	1588	1.5	196
	2.0	1359	550	2.0	68

Power and sample size

Case-control	GxE interaction			G main effect	
		$P(E=1)$			
$P(G=1)$	$\exp(\beta_3)$	0.1	0.5	$\exp(\beta_1)$	
0.05	1.5	19571	7520	1.5	860
	2.0	6263	2495	2.0	266
0.40	1.5	4107	1588	1.5	196
	2.0	1359	550	2.0	68

Case-only	GxE interaction		
		$P(E=1)$	
$P(G=1)$	$\exp(\beta_3)$	0.1	0.5
0.05	1.5	9257	3916
	2.0	2640	1257
0.40	1.5	2114	885
	2.0	671	317

Genomewide... $\alpha = 5 \times 10^{-8}$

Case-control	GxE interaction			G main effect	
		$P(E=1)$			
$P(G=1)$	$\exp(\beta_3)$	0.1	0.5	$\exp(\beta_1)$	
0.05	1.5	98725	37946	1.5	4338
	2.0	31600	12591	2.0	1344
0.40	1.5	20707	7995	1.5	988
	2.0	6850	2776	2.0	342

Case-only	GxE interaction		
		$P(E=1)$	
$P(G=1)$	$\exp(\beta_3)$	0.1	0.5
0.05	1.5	46686	19739
	2.0	13304	6342
0.40	1.5	10654	4456
	2.0	3385	1579

Those sample sizes are large.....

- Interactions need larger sample sizes than main effects.
- Genome-wide searches need to correct for many comparisons.
- Third whammy: sometimes the E has serious measurement error, reducing power even further.

Idea: can we identify SNPs that are "more likely" to be involved in interactions, and only test those.

Issue: we need to select those SNPs in such a manner that we only need to multiple-comparisons correct for the number of SNPs we test for interactions, not the number we could have tested.

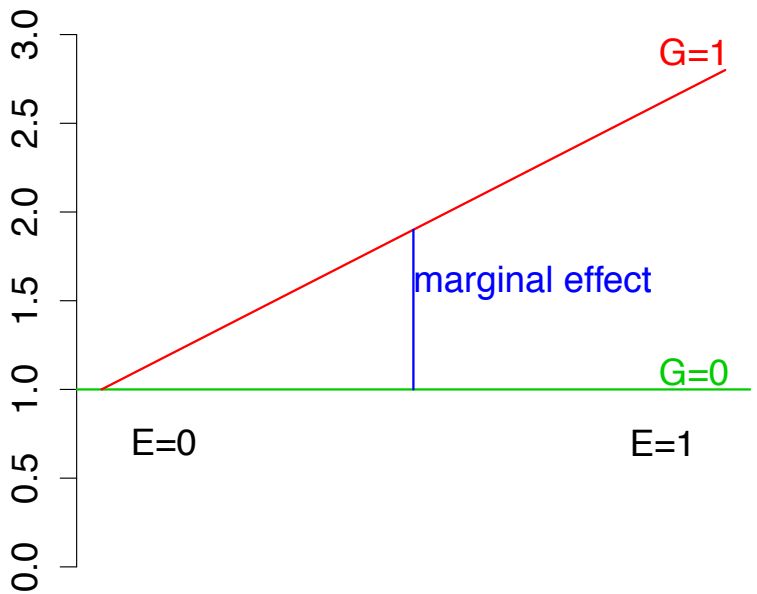
Simplest example

[Kooperberg & LeBlanc (2008) *Genet Epi* **32**:255–63.]

1. Genome-wide screen of the top M SNPs using "marginal-effect" test on all subjects:

$$\text{logit}(Pr(Y = 1|G)) = \gamma_0 + \gamma_1 G$$

Test $H_0 : \gamma_1 = 0$ for each SNP at α_M level.



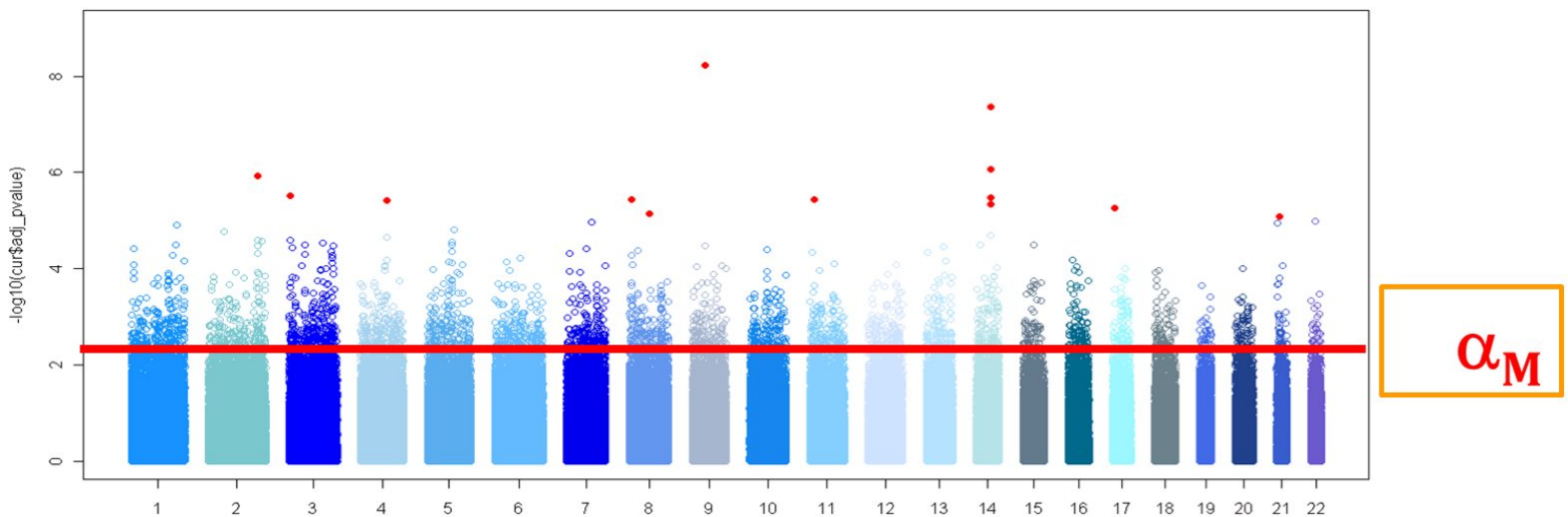
Simplest example

[Kooperberg & LeBlanc (2008) *Genet Epi* **32**:255–63.]

1. Genome-wide screen of the top M SNPs using "marginal-effect" test on all subjects:

$$\text{logit}(Pr(Y = 1|G)) = \gamma_0 + \gamma_1 G$$

Test $H_0 : \gamma_1 = 0$ for each SNP at α_M level.



Simplest example

[Kooperberg & LeBlanc (2008) *Genet Epi* **32**:255–63.]

1. Genome-wide screen of the top M SNPs using "marginal-effect" test on all subjects:

$$\text{logit}(Pr(Y = 1|G)) = \gamma_0 + \gamma_1 G$$

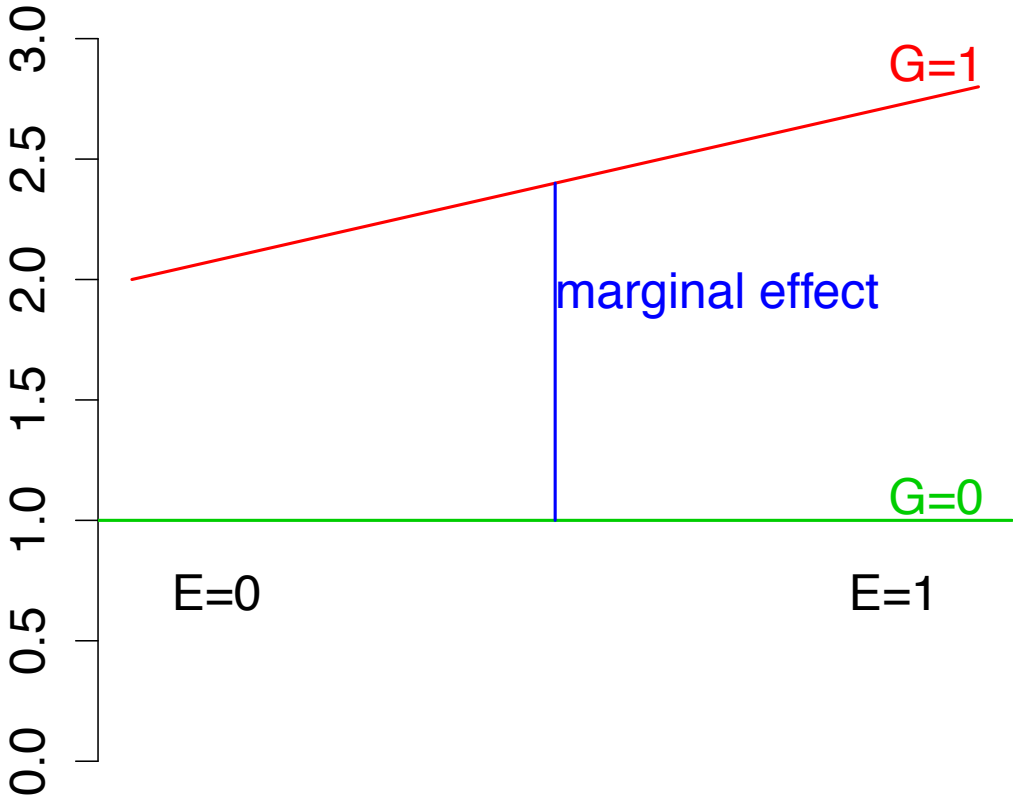
   Test $H_0 : \gamma_1 = 0$ for each SNP at α_M level.

2. Test m SNPs that pass screen with standard logistic model

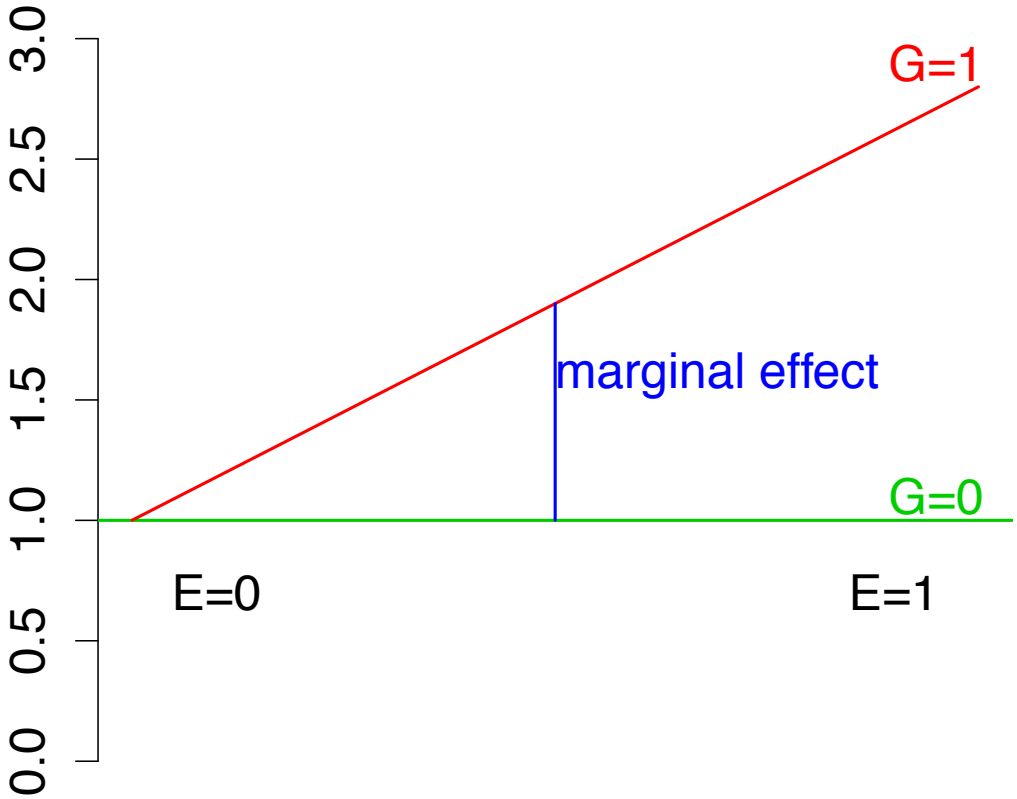
$$\text{logit}(Pr(Y = 1|G, E)) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE$$

   Significance threshold α/m.

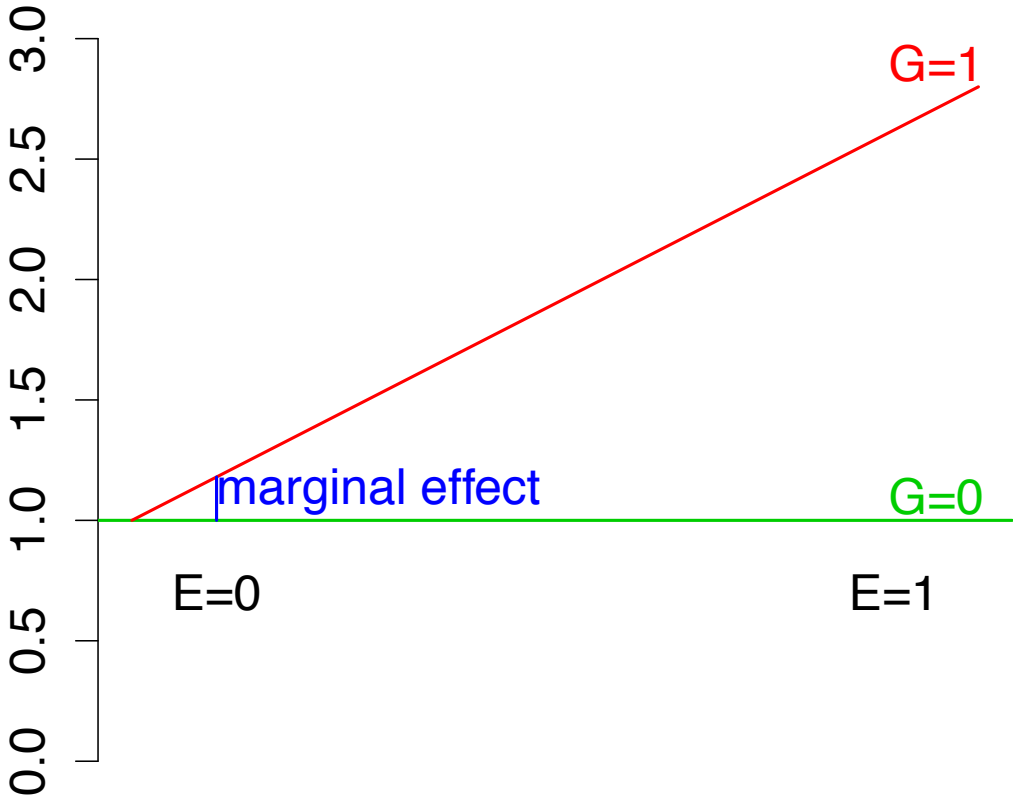
Here this works well



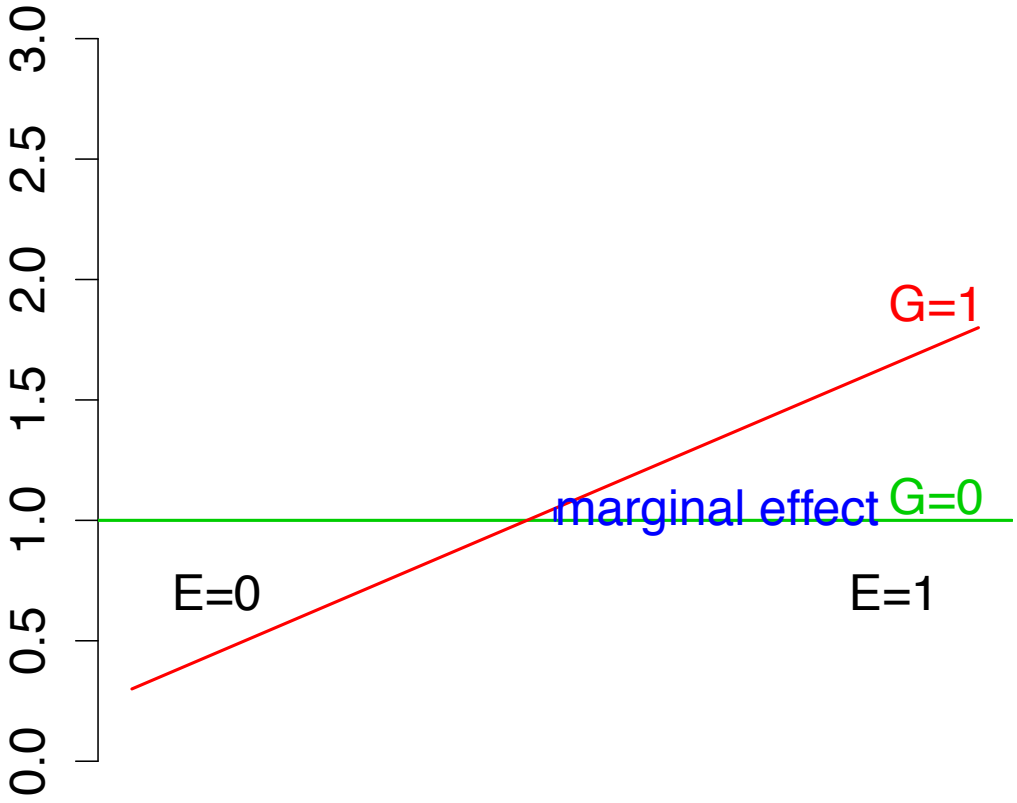
Here as well



Here it is more problematic



Here it won't work well



Two stage procedures

1. Screen all SNPs
   - Using marginal association
   - Using the correlation of G and E in cases and controls combined
   - A combination of the above
2. Test only those SNPs that pass the screen.
   - Case-control
   - Case-only
   - Data-adaptive (e.g. Empirical Bayes)

Testing approaches

Case-control	▶ Robust ▶ Does not assume G-E independence
Case-only	▶ Substantial gain in power when G-E independence holds ▶ Type 1 error increases when G-E independent incorrectly assumed ▶ Also assumes "rare disease" (this can be relaxed)
Data-adaptive: Empirical Bayes, Bayesian Model Averaging	▶ Increased power versus case-control ▶ Improved control of type 1 error versus case-only

Key result

[Dai et al. (2012) *Biometrika*, **99**:929–44.]

For a screening procedure to maintain the correct type 1 error, the test statistics for screening and testing are (pairwise) independent.

We can then work out that

- For generalized linear models and Cox proportional hazards models, the marginal association screening is independent of the case-control, the case-only, and the empirical Bayes estimators.
- The correlation between G and E is independent of the case-control estimator, but not independent of the case-only or the empirical Bayes estimators.

Thus, when using a two-stage procedure we need to consider the pair of tests: not every screening statistic can be matched up with every GxE test.

Formally...

- Subject $i = 1, \ldots, n$ iid: outcome Y_i, genes $G_{i1}, \ldots, G_{im}$, environmental variable E_i, confounders W_i.
- θ_j is the G-E interaction between G_j and E. The Wald statistic for $H_{0j} : \theta_j = 0$ is $T_j = \widehat{\theta}_j / \widehat{\text{var}}(\widehat{\theta}_j)^{1/2}$.
- Let $K_{0j} : \xi_j = 0$ be another hypothesis with an assymptotically linear estimator; let $T_j^0 = \widehat{\xi}_j / \widehat{\text{var}}(\widehat{\xi}_j)^{1/2}$ be its Wald statistic.
- Specify $0 \leq \alpha_0 \leq 1$. Set $\Gamma_j^0 = \{T_j^0 : |T_j^0| > \Phi^{-1}(1 - \alpha_0/2)\}$. Suppose m_0 genetic variants pass the filter.
- Define $\Gamma_j = \{T_j : |T_j| > \Phi^{-1}(1 - \alpha/(2m_0))\}$.
  We declare the jth test significant if $T_j^0 \in \Gamma_j^0$ and $T_j \in \Gamma_j$.
- **Theorem:** if $\text{cov}\{n^{1/2}(\widehat{\xi}_j - \xi_j), n^{1/2}(\widehat{\theta}_j - \theta_j)\} \to 0$. for all j, and m_0/m converges to a constant α_0' in probability, then the two-step procedure preserves the family wise error rate.

Using this result

For most of the filtering/testing approaches we can establish this independence on a case-by-case basis. But the following result is more general useful.

- Let $(Y_i, V_{i1}, \ldots, V_{ip})$, i=1,. . . , n be iid random variables, with Y the outcome variable in a GLM with canonical link g.
- Let $q < p$. Consider the nested GLMs

$$g\{E(Y|V_1, \ldots, V_q)\} = \beta_0 + \sum_{j=1}^{q} \beta_j V_j,$$

$$g\{E(Y|V_1, \ldots, V_p)\} = \beta_0 + \sum_{j=1}^{p} \gamma_j V_j.$$

- **Theorem:** the MLEs $(\widehat{\beta}_0, \ldots, \widehat{\beta}_q)$ and $(\widehat{\gamma}_{q+1}, \ldots, \widehat{\gamma}_p)$ are assymptotically independent.

Thus, the independence of $\widehat{\beta}_1$ and $\widehat{\gamma}_3$ is immediate in these models:

$$g\{E(Y|G, W)\} = \beta_0 + \beta_1 G + \beta_2 W,$$

$$g\{E(Y|G, E, W)\} = \gamma_0 + \gamma_1 G + \gamma_2 E + \gamma_3 GE + \gamma_4 W.$$

Why does screening on G-E correlation make sense?

$2 \times 2 \times 2$ data:

	$Y=0$		$Y=1$	
	$G=0$	$G=1$	$G=0$	$G=1$
$E=0$	n_a	n_b	n_e	n_f
$E=1$	n_c	n_d	n_g	n_h

$$\text{Interaction} = \frac{\text{OR}(Y, G|E=1)}{\text{OR}(Y, G|E=0)} = \frac{(n_h n_c)/(n_g n_d)}{(n_f n_a)/(n_e n_b)}$$

$$= \frac{(n_h n_e)/(n_g n_f)}{(n_d n_a)/(n_c n_b)} = \frac{\text{OR}(G, E|Y=1)}{\text{OR}(G, E|Y=0)}$$

- So the G-E correlation is different between cases and controls, therefore, at least one of these strata has a correlation that is not 0.

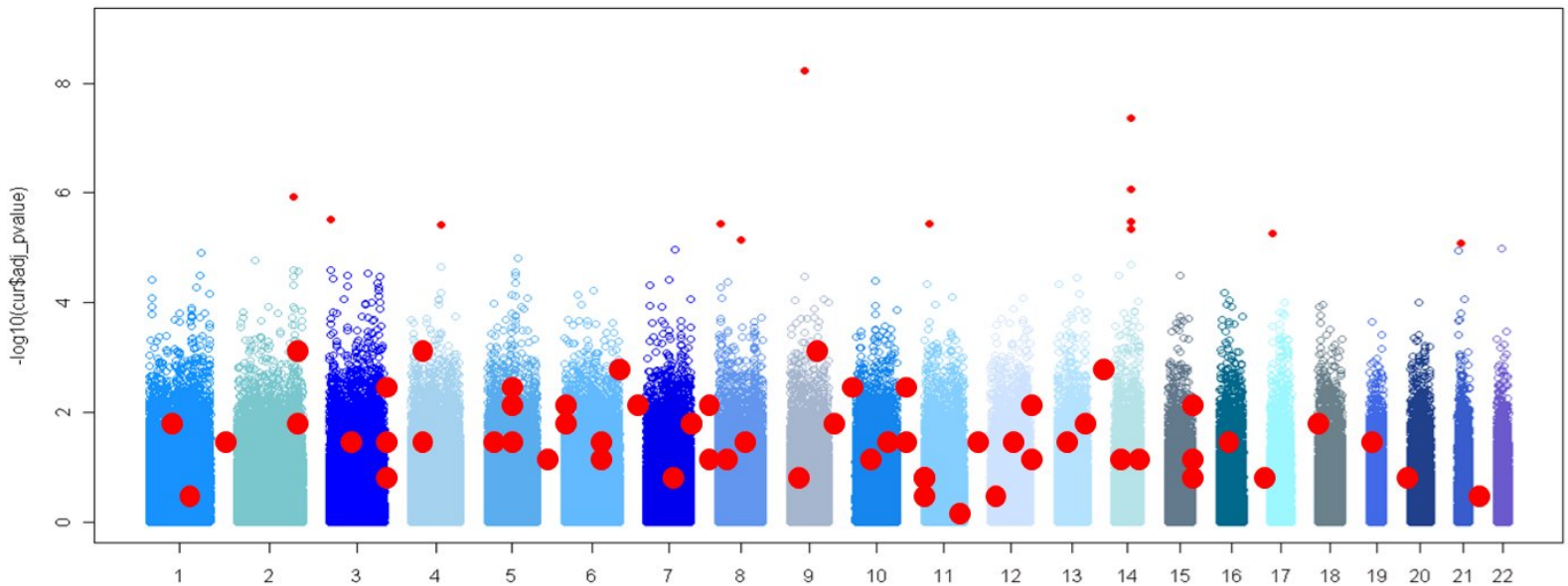
G-E interaction screening

[Murcray et al. (2009) *Am J Epi* **169**:219–26.]

1. Genome-wide screen of the top M SNPs testing for G-E correlation (i.e., a "case-only" test on all subjects

$$\text{logit}(Pr(E = 1|G)) = \delta_0 + \delta_1 G$$

Test $H_0 : \delta_1 = 0$ for each SNP at α_M level.



G-E interaction screening

[Murcray et al. (2009) *Am J Epi* **169**:219–26.]

1. Genome-wide screen of the top M SNPs testing for G-E correlation (i.e., a "case-only" test on all subjects

$$\text{logit}(Pr(E = 1|G)) = \delta_0 + \delta_1 G$$

   Test $H_0 : \delta_1 = 0$ for each SNP at α_M level.

2. Test m SNPs that pass screen with standard logistic model

$$\text{logit}(Pr(Y = 1|G, E)) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE$$

   Significance threshold α/m.

Multiple possibilities

- Marginal G screening $\Longrightarrow$ case-control testing.
- G-E correlation screening $\Longrightarrow$ case-control testing.

But also

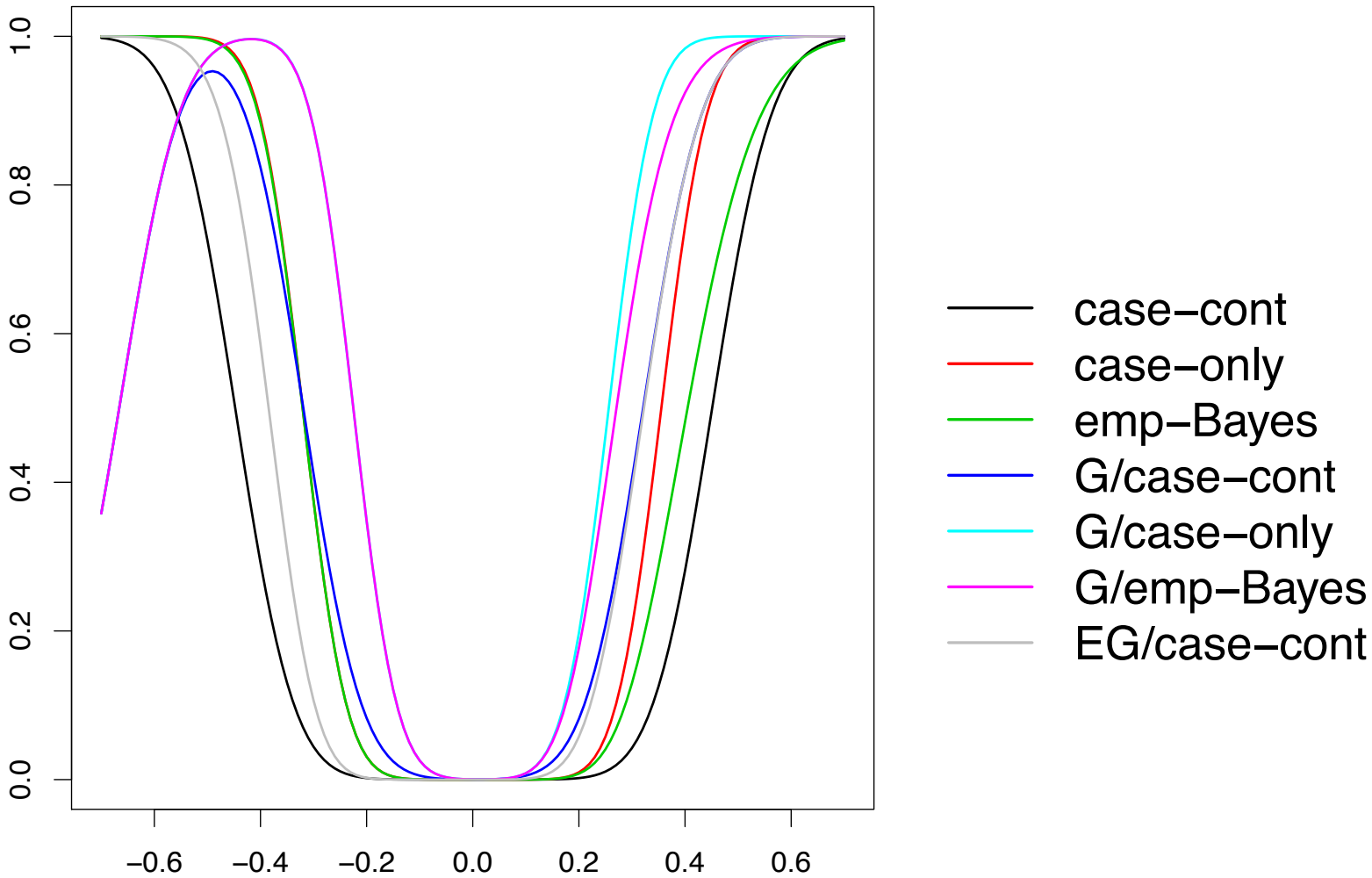
- Marginal G screening $\Longrightarrow$ empirical Bayes testing.
- Marginal G screening $\Longrightarrow$ case-only testing, if G-E independence holds.

Which one works best?

Example 1

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1$

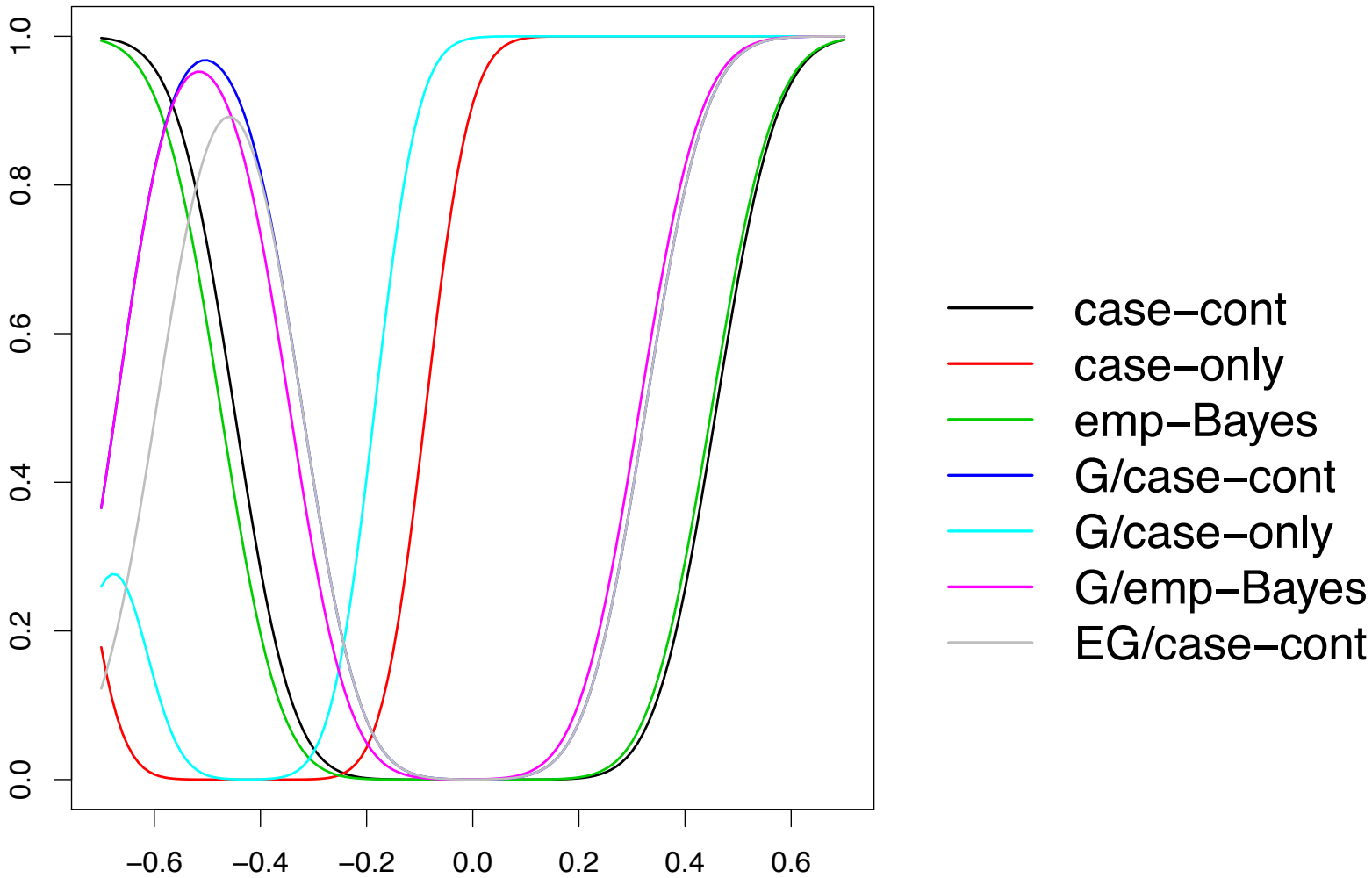
$$\text{logit(Y=1—G,E)} = \beta_0 + 0.5G + 0.5E + \beta_3 GE$$



Example 2

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1.5$

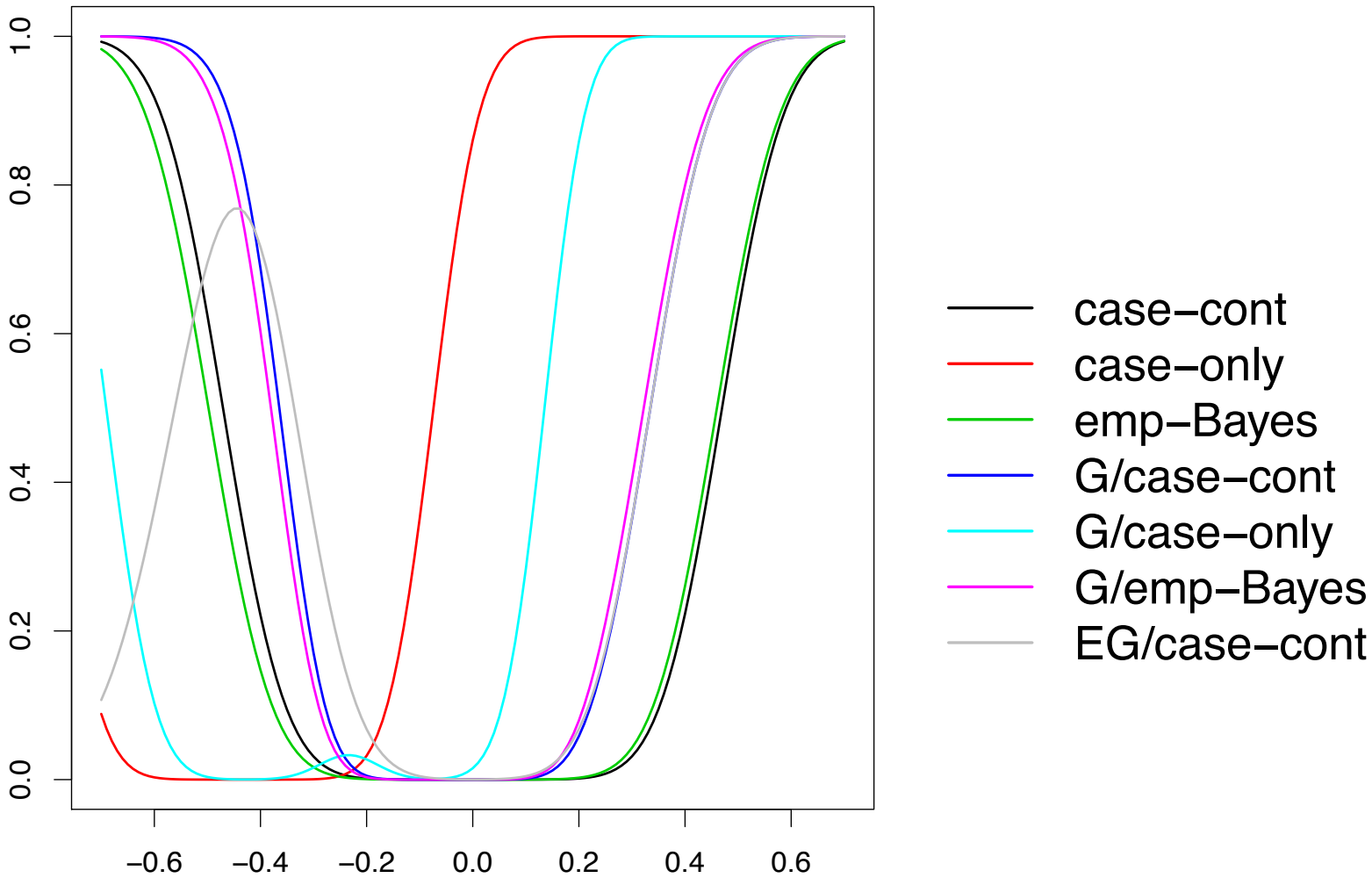
$$\text{logit(Y=1—G,E)} = \beta_0 + 0.5G + 0.5E + \beta_3 GE$$



Example 3

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1.5$

$$\text{logit(Y=1—G,E)} = \beta_0 + 0G + 0.5E + \beta_3 GE$$



A general framework

[Hsu et al. (2012) *Gen Epi* **36**:183–94.]

Module A: Screening

- No Screening
- Marginal G-D association
- Correlation G-E
- Hybrid approaches

Module B: Multiple Comparisons

- Bonferroni testing
- Permutations
- Weighted hypothesis testing

Module C: Testing

- Case-control
- Case-only
- Empirical Bayes (EB)
- Bayesian Model Averaging

Cocktail approach

[Hsu et al. (2012) *Gen Epi* **36**:183–94.]

- Let p^{marg} be the P-value for $H_0 : \gamma_1 = 0$ in

$$\text{logit}(Pr(Y = 1|G)) = \gamma_0 + \gamma_1 G$$

- Let p^{corr} be the P-value for $H_0 : \delta_1 = 0$ in

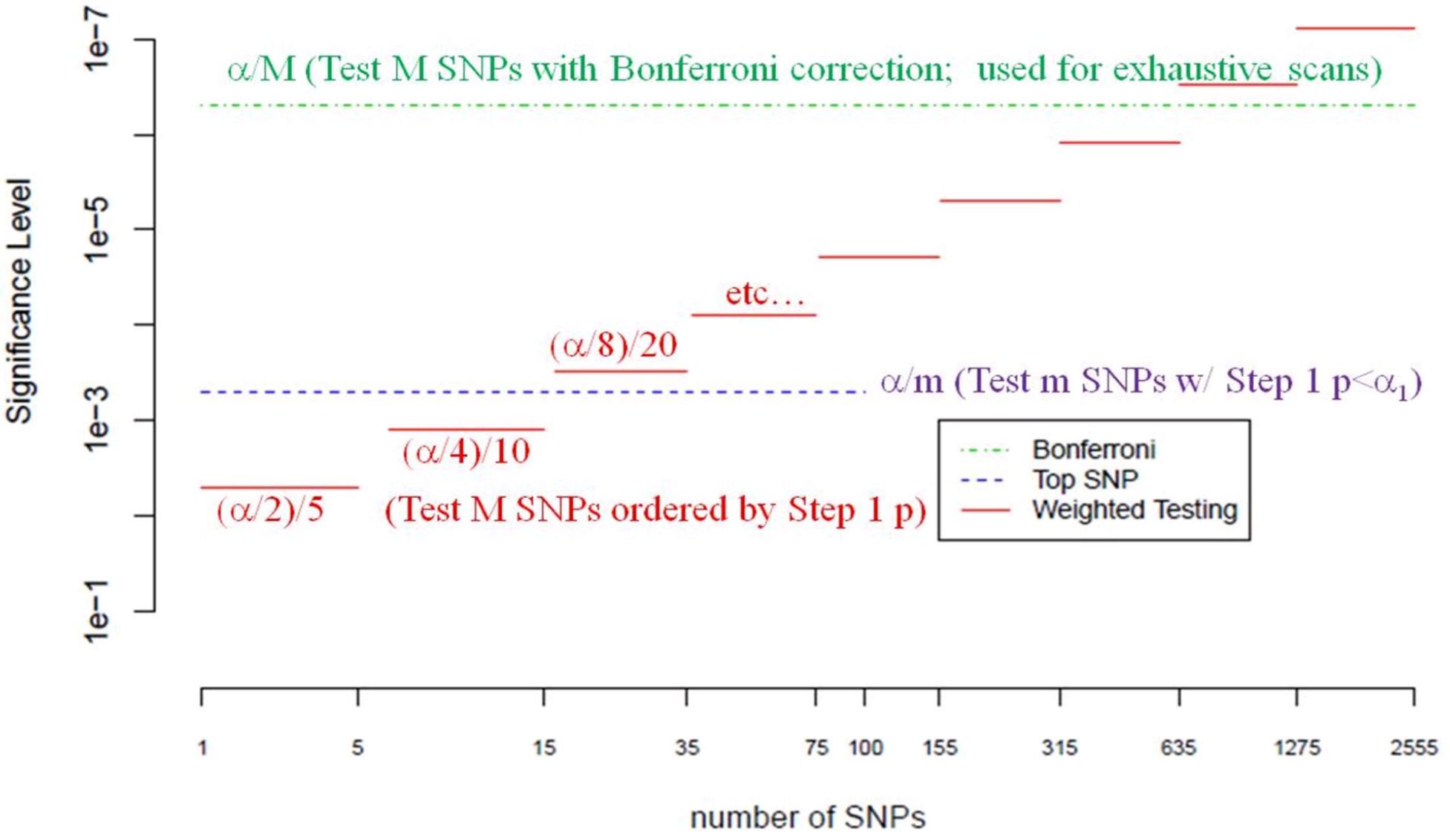
$$\text{logit}(Pr(E = 1|G)) = \delta_0 + \delta_1 G$$

- Use $p^{screen} = \min(p^{marg}, p^{corr})$ for screening.
- Screening can be done using a fixed α or using weighted testing (next slide).
- If $p^{marg} \leq p^{corr}$ test using empirical Bayes, if $p^{marg} > p^{corr}$ test using case-control.

Somewhat similar method: H2 [Murcray et al. (2011) *Gen Epi* **35**:201–10.]

Weighted hypothesis testing

[Ionita-Laza et al. (2007) *AJHG* **81**:601–14.]



Weighted hypothesis testing

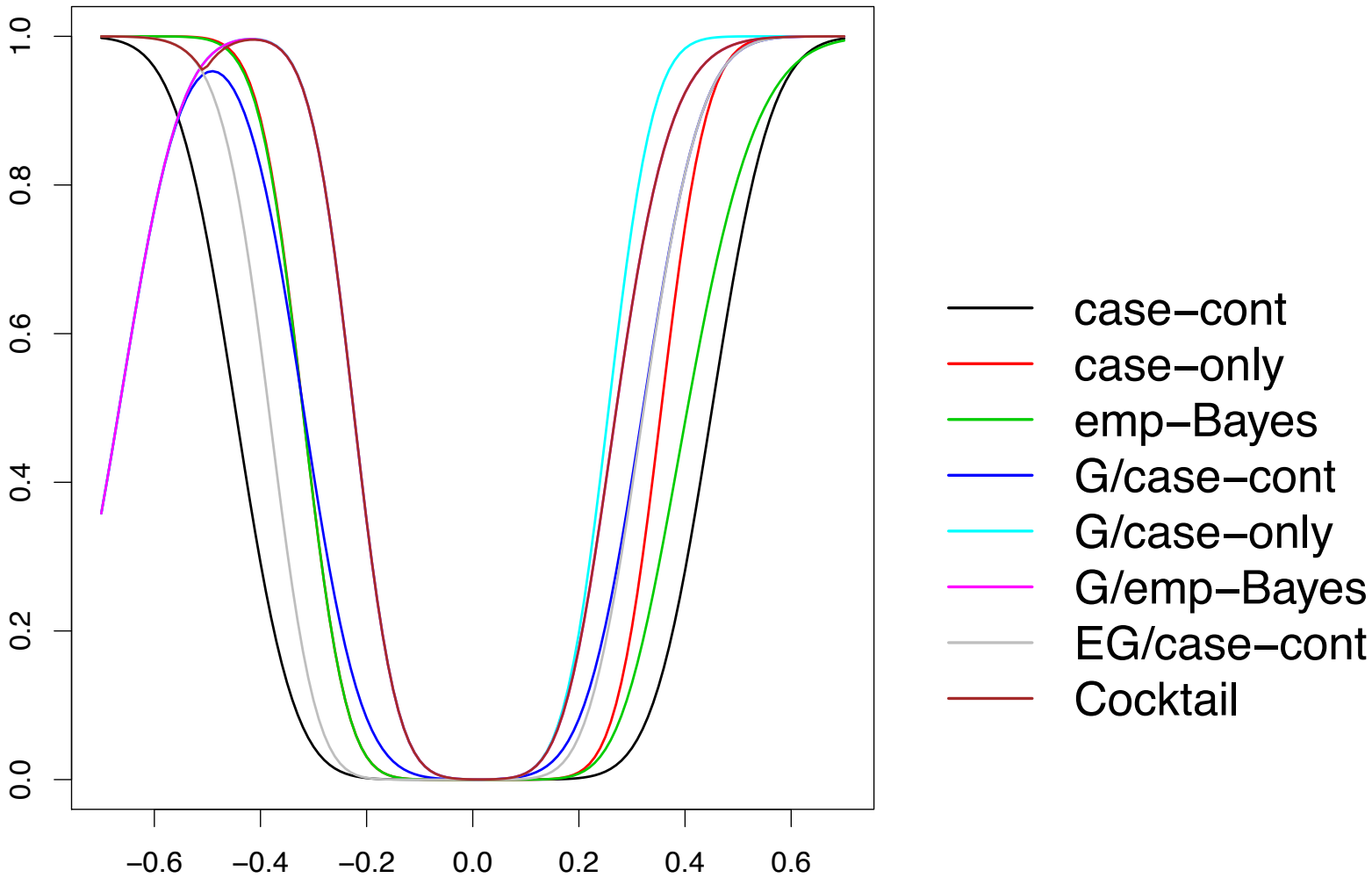
- All SNPs are tested, but with different significant thresholds.
- Rank SNPs by screening P-value, e.g.
  1. 5 SNPs with smallest screening P-value.
  2. next 10 SNPs
  3. next 20 SNPs

Group	# SNPs	Alpha
1	5	5.00E-3
2	10	1.25E-3
3	20	3.13E-4
4	40	7.81E-5
5	80	1.95E-5
6	160	4.88E-6
7	320	1.22E-6
8	640	3.05E-7
9	1280	7.63E-8
10	2560	1.91E-8
11	5120	4.77E-9
12	10240	1.19E-9
...	...	...

Example 1

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1$

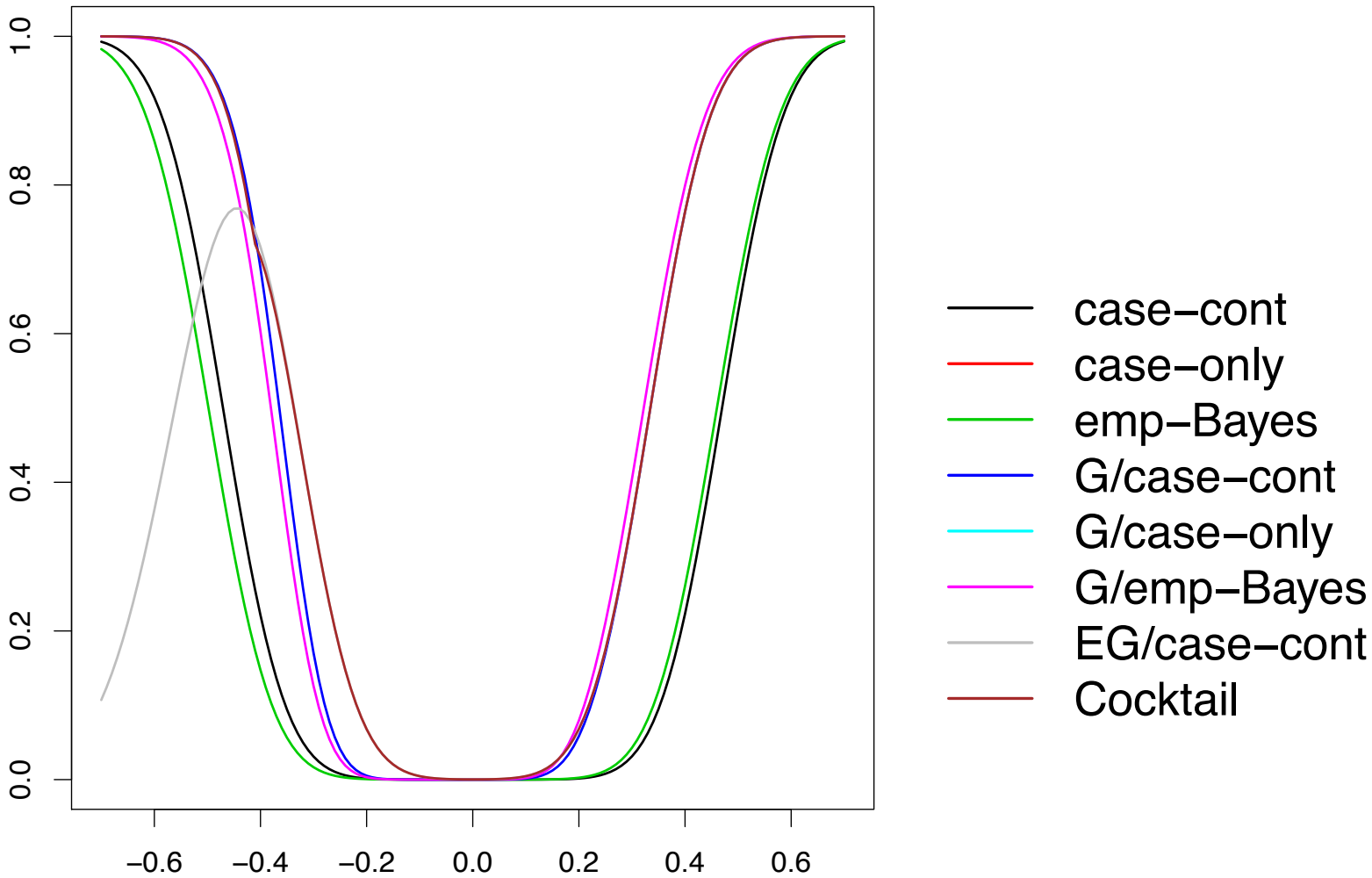
$$\text{logit(Y=1—G,E)} = \beta_0 + 0.5G + 0.5E + \beta_3 GE$$



Example 3

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1.5$

$$\text{logit(Y=1—G,E)} = \beta_0 + 0G + 0.5E + \beta_3 GE$$



One more modification.... EDGxE

[Gauderman et al. (2013) *Gen Epi* **37**:603–13.]

Instead of either using marginal or G-E ranking, can we use both simultaneously?

1. Let T_{marg} be the χ^2 statistic for for $H_0 : \gamma_1 = 0$ in

$$\text{logit}(Pr(Y = 1|G)) = \gamma_0 + \gamma_1 G$$

and let T_{corr} be the χ^2 statistic for for $H_0 : \delta_1 = 0$ in

$$\text{logit}(Pr(E = 1|G)) = \delta_0 + \delta_1 G$$

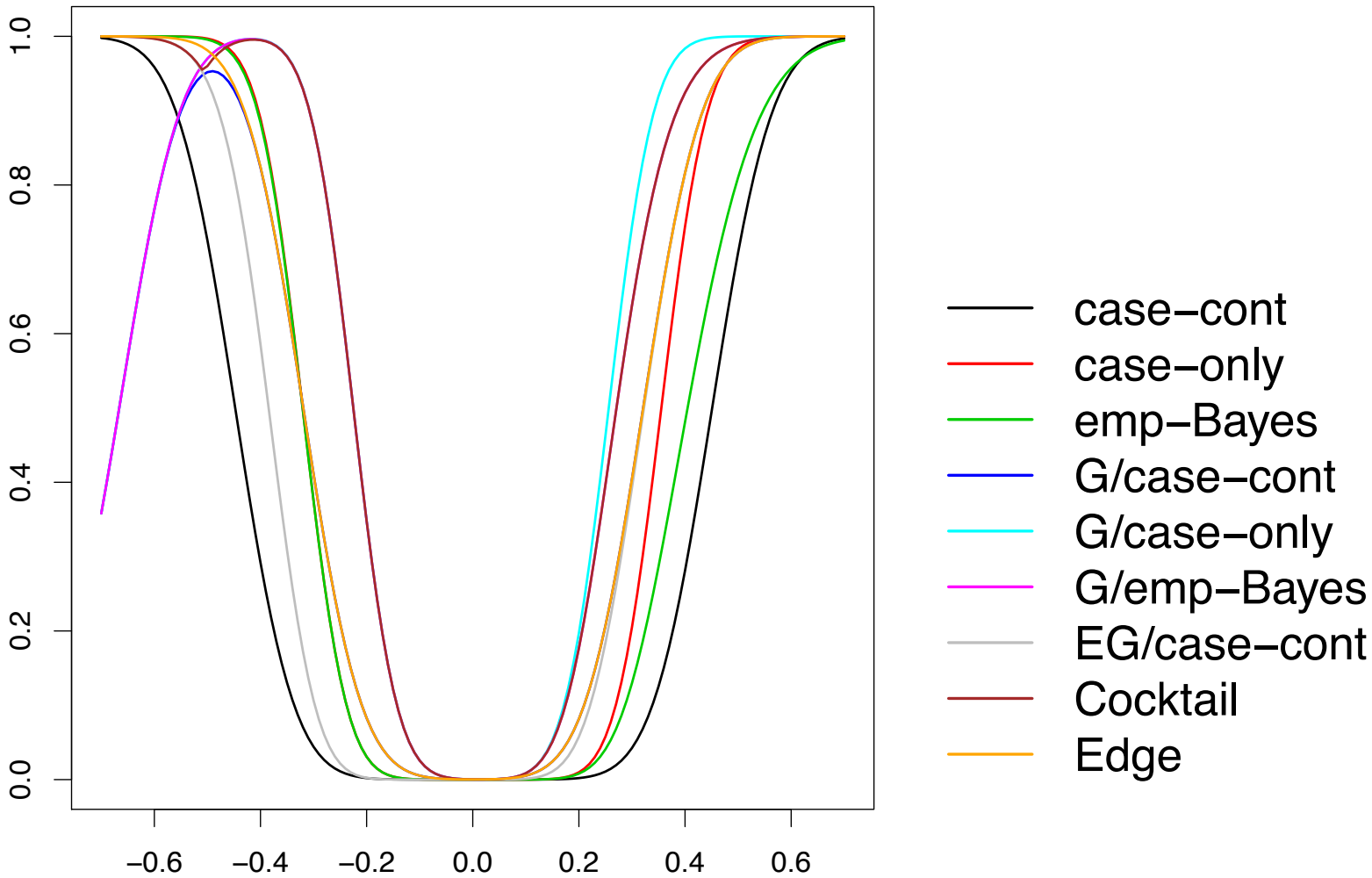
Then rank using $T_{EDGxE} = T_{marg} + T_{corr}$.

2. Test only those wit $p_{EDGxE} < \alpha_M$ or use weighted hypothesis testing.

Example 1

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1$

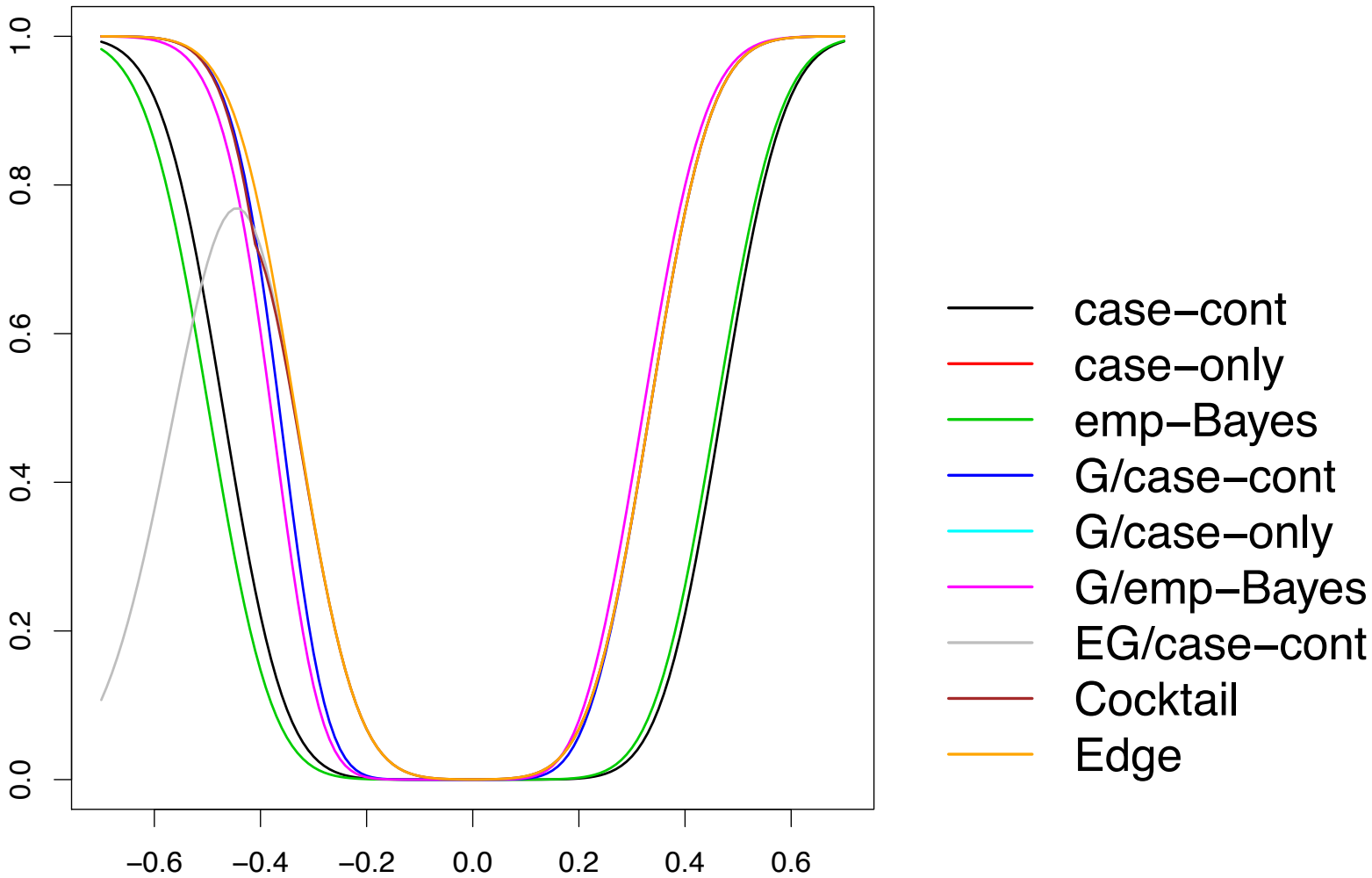
$$\text{logit(Y=1—G,E)} = \beta_0 + 0.5G + 0.5E + \beta_3 GE$$



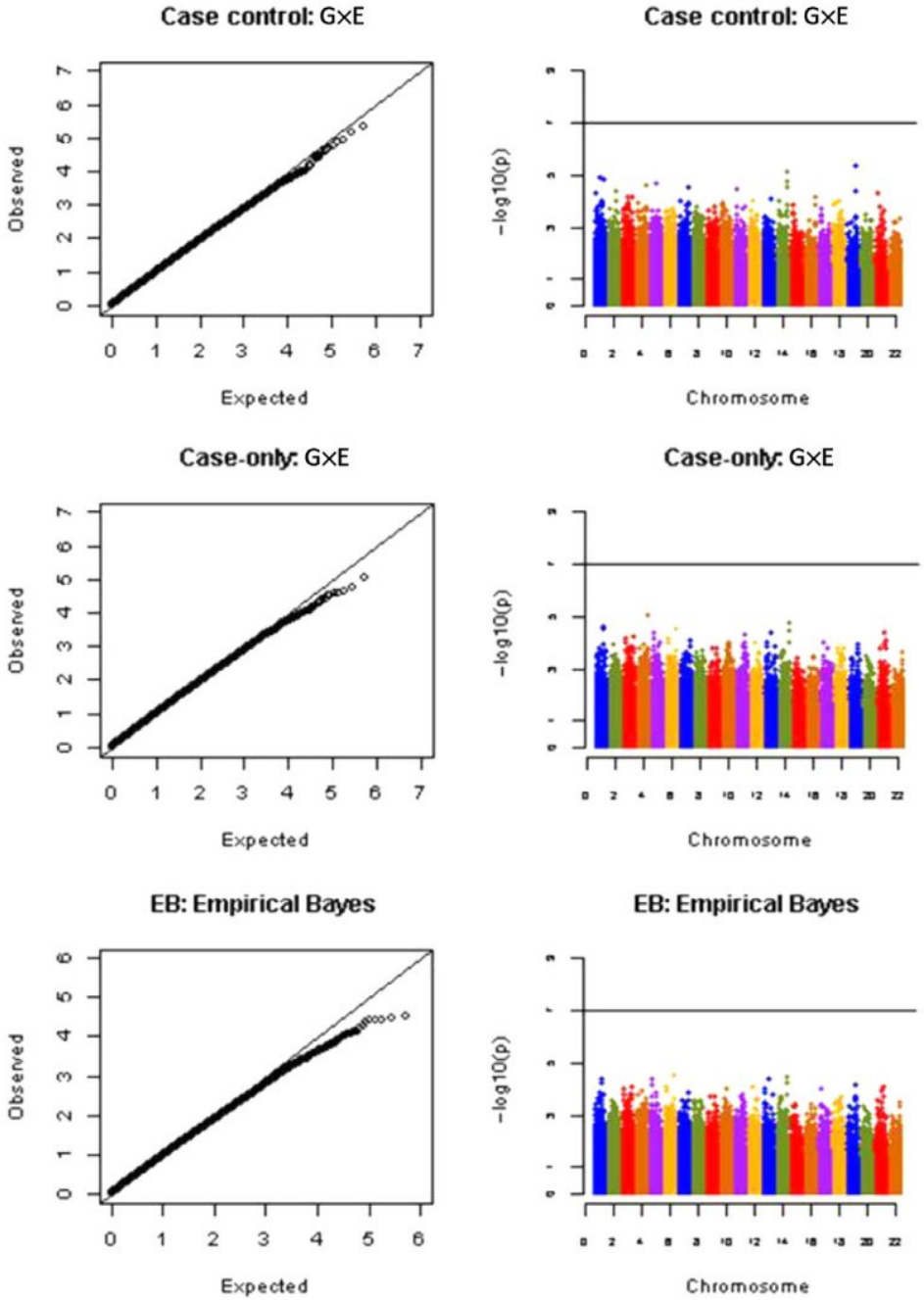
Example 3

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1.5$

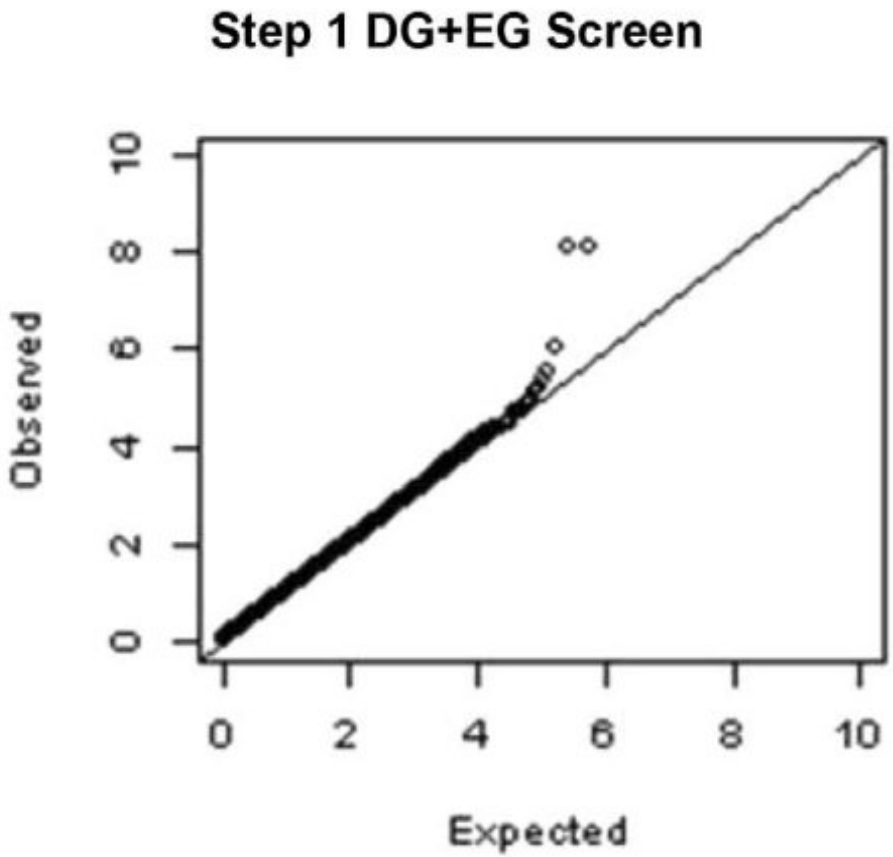
$$\text{logit(Y=1—G,E)} = \beta_0 + 0G + 0.5E + \beta_3 GE$$

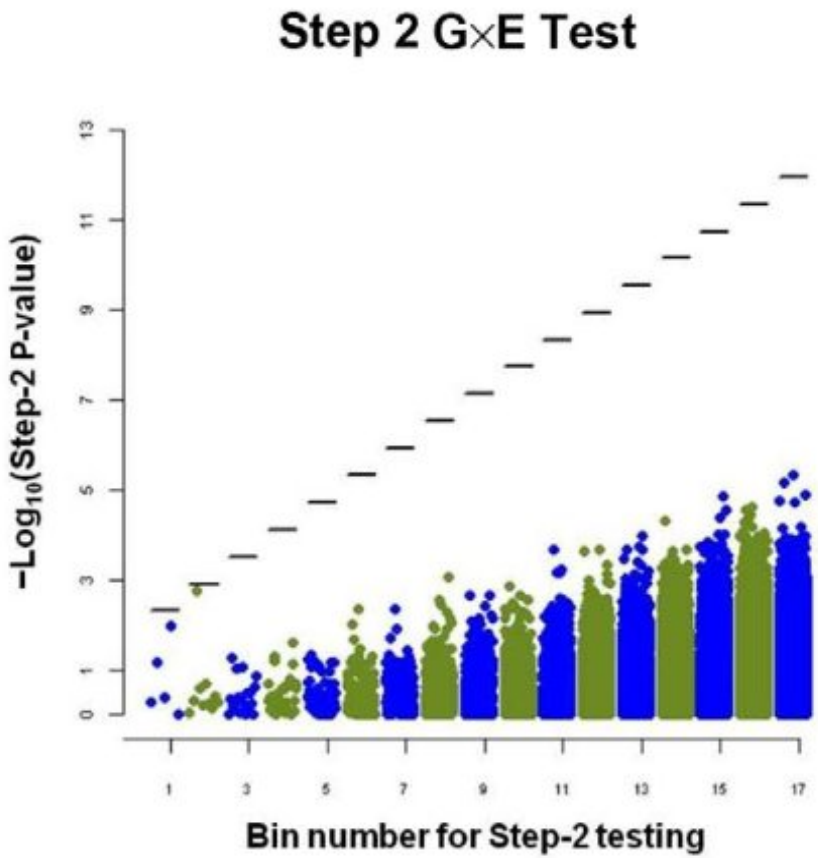


Asthma GWAS from Gauderman et al.



Asthma GWAS from Gauderman et al.





Asthma GWAS from Gauderman et al.

Table 6. Top 15 SNPs from EDG×E analysis of 536,857 SNPs for G × Sex interaction with young-onset childhood asthma

Chr	SNP	Location	Reference Allele	Step 1		Step 2				Significance Threshold
				Chi-square	*P*-value	Bin	$OR_{G\times E}$	*t*-Test	*P*-value	
1	rs6697552	241192975	T	37.48	7.3×10^{-9}	1	1.13	0.64	0.52	0.005
1	rs1832719	200713649	C	37.23	8.2×10^{-9}	1	0.42	−1.81	0.07	0.005
7	rs1229492	81402058	T	27.88	8.8×10^{-7}	1	0.88	−0.82	0.41	0.005
4	**rs6842542**	**158594467**	**C**	**25.50**	$\mathbf{2.9 \times 10^{-6}}$	**1**	**1.60**	**2.55**	**0.011**	**0.005**
9	rs520613	109925873	G	24.79	4.1×10^{-6}	1	1.01	0.04	0.97	0.005
4	rs719525	76578274	C	24.18	5.6×10^{-6}	2	0.98	−0.11	0.91	0.0012
5	rs10069175	21923777	C	23.57	7.6×10^{-6}	2	1.15	0.70	0.48	0.0012
8	**rs7000310**	**119837792**	**C**	**22.91**	$\mathbf{1.1 \times 10^{-5}}$	**2**	**0.57**	**−3.13**	**0.0017**	**0.0012**
8	rs10505105	108376513	T	22.61	1.2×10^{-5}	2	0.82	−1.11	0.27	0.0012
9	rs630965	109925300	G	22.14	1.6×10^{-5}	2	1.07	0.53	0.59	0.0012
13	rs1988388	52944609	T	21.75	1.9×10^{-5}	2	1.19	1.26	0.21	0.0012
9	rs2767777	125998894	C	21.69	2.0×10^{-5}	2	0.91	−0.59	0.56	0.0012
9	rs865686	109928299	C	21.67	2.0×10^{-5}	2	1.05	0.34	0.74	0.0012
12	rs4765748	3724788	A	21.66	2.0×10^{-5}	2	0.84	−0.87	0.38	0.0012
15	rs1523526	59051579	C	21.31	2.4×10^{-5}	2	0.92	−0.62	0.53	0.0012

What about $G \times G$? Continuous Y? Continuous E?

- Most things go through the same way.
- Except, case-only and empirical Bayes estimators need a binary Y.
- With GxG computational efficiency becomes more of an issue.
- Other complications arise when the G were imputed, and are not exactly 0/1/2.

A sobering note

There likely have been more papers written about methods to identify *GxE* and *GxG* interactions, than the number of interactions that have successfully been identified.