

ゲノムワイド関連解析における多重検定手法の比較評価

大谷 隆浩 リスク解析戦略研究センター 特任助教

【背景・目的】

ゲノムワイド関連解析 (GWAS) は、ゲノム全体をカバーする数十万～数百万の一塩基多型 (SNP) の遺伝子型を決定し、遺伝子型と疾患・量的形質との関連を統計的に調べる方法論である。遺伝疫学やバイオテクノロジーの発展に伴い、多くの疾患関連遺伝子がGWASによって同定されている。この方法では数百万程度のSNPが検定の対象となるが、現実的に用意できるサンプル数は数百～数万程度であるため、高次元小標本のデータを扱うことになる。このため、個別のSNPについて単回帰分析を行い、各回帰係数の尤度比検定によって得られた統計的有意性を多重検定補正する、という解析手順が採用されている。

近年の研究により、GWASによって同定された変異を集めても、家系分析から推定された遺伝率のごく一部しか説明できていないことがわかってきている。この原因の一つとして、GWASにおいては関連の有無を判定するために、 $P < 5 \times 10^{-8}$ を閾値とする「Genome-wide significance基準」が一般的に用いられていることが考えられる。この基準は、有意水準5%の検定を100万回行う際に、「少なくとも1つの偽陽性が起こる確率 (ファミリーワイズエラー率: FWER)」が5%未満になるように、Bonferroni補正法によって多重検定補正を行うことに対応する。この基準は過度に保守的で、多くの疾患関連SNPを見逃している可能性があると考えられる。

本研究では、実データを使用した解析例と、シミュレーション実験を通して、多重検定補正手法の性能評価を行う。Genome-wide significance基準の効率性について再評価し、また、マイクロアレイ技術による臨床研究などで広く用いられている False discovery rate (FDR: 棄却された全ての帰無仮説に含まれる偽陽性の割合) に基づいた補正手法との比較を行う。

【多重検定補正手法】

Bonferroni法: 有意水準を検定の数で割り、その値を補正された有意水準とすることでFWERを制御する。

Benjamini & Hochberg (BH) 法[2]、Storey (ST) 法[3]: 代表的なFDR制御手法。

Weighted Benjamini&Hochberg (wBH) 法 [4]: Minor allele頻度 (MAF) の逆数でSNPを重み付けした上で、BH法を適用する。

Wakefield's Bayesian framework (BF) [5]: ロジスティック回帰モデルに対する漸近近似ベイズ因子に基づいて偽陽性の確率を評価する。

Optimal discovery procedure (ODP) [6, 7]:

偽陽性の期待値を一定にした上で真陽性の期待値を最大化する、単一検定におけるNeyman-Pearsonの補題を一般化した検定方式。

【シミュレーション実験】

現実的なGWASの設定における各補正手法の性能を比較するために、文献[8]を参考に、統合失調症のケースコントロール研究を模した擬似的なデータを作成した。データの作成にはPLINK[9]を使用し、サンプル数 $N=20,000$ (ケース 10,000、コントロール10,000)、SNP数 1,000,000 (うち、83,000個が疾患関連SNP)、疾患関連SNPの相対リスク1.05、一般集団の有病率1%、というようにパラメータを設定した。この設定で1,000回の独立なシミュレーションを行い、指定した有意水準のもとで、実際のFWER/FDRがどの程度になるか、また、有意と判定されるSNPの数を調べた。

図1にシミュレーション実験より推定したFWER/FDRを示す。Bonferroni法、BH法、ST法、ODP法は指定した有意水準のもとで、適切に偽陽性を制御できている。しかしながら、BF法では疾患関連の事前確率 π_1 の設定によっては偽陽性が増加してしまっている。事前確率に対する感度は高く、実用の際には注意深い設定が必要であるといえる。

図2に指定した有意水準のもとで、有意と判定されるSNPの数を示す。FWERを制御するBonferroni法では過度に保守的で、有意となるSNPは少ない。これに対してFDRを制御する方法では、FDR=1%のような比較的厳しい水準のもとでも、より多くのSNPを検出できている。特に、一定の有意水準のもとで真陽性の期待値を最大化するODPでは、より多くのSNPを検出できている。BF法ではより多くのSNPを検出しているが、前述のとおり、事前確率の設定によっては擬陽性を指定した水準に適切に制御出来ないため注意しなければならない。

【実データの解析】

実データの解析における各補正手法の効率性を評価するために、統合失調症に関するケースコントロール研究から得られた大規模なGWASのデータセット[1]を対象に解析を行った。このデータセットは17の研究から得られたデータを統合したもので、サンプル数は全体で21,856 (ケース 9,394、コントロール12,462) である。データセットには1,252,901個のSNPに対するロジスティック回帰を用いた関連解析の結果が記録されている。実際のデータファイルはPGCのWebサイト (<http://www.med.unc.edu/pgc/>) からダウンロードできる。このデータに対して各補正手法を適用した。

図3に有意と判定されたSNPの数を示す。シミュレーション実験の結果と同様に、FDRを制御する補正手法ではBonferroni法よりも多くのSNPを有意と判定している。表1に、文献[1]で報告された疾患関連SNPに対する、有意性のランキングを示す。BH法とST法によるランキングはP値によるランキングと一致する。BF法によるランキングはP値によるランキングと似ているに関わらず、FDRの推定値は、妥当な事前確率の設定 (8%) においても、より小さくなっている (図3)。この結果は、推定値に強いバイアスがあり、また事前確率の設定に対して非常に感度が高いことを示唆している。したがって、少なくとも偽陽性の割合を評価することにおいては、これらの結果を単純に用いることは問題があると考えられる。

以上の結果に対して、偽陽性の期待値を一定にした上で真陽性の期待値を最大化するODP法でのランキングは、P値のランキングでは上位に来ないSNPを検出する傾向がある。表2では、ODP法でのランキング上位100個のSNPのうち、P値では上位にならないSNPを示した。特に、rs1107592とrs10226475の2つのSNPはGenome-wide significance基準を満たさないにもかかわらず、ODPでのランキングでは比較的上位となった。このように、ODPでは偽陽性の発生を適度に抑えた上で、GWASにおいて一般的に用いられているP値に基づく方法よりも効率的に疾患関連の候補となるSNPをスクリーニングできる可能性があることを示唆している。

【参考文献】

- [1] The Schizophrenia Psychiatric GWAS Consortium, 2011; nature genetics 43: 969-978
- [2] Benjamini and Hochberg, 1995; J R Statist Soc B 57: 289-300
- [3] Storey, 2002; J R Statist Soc B 64: 479-498
- [4] Xing et al., 2010; Am J Hum Genet 86: 440-446
- [5] Wakefield, 2007; Am J Hum Genet 81: 208-227
- [6] Storey 2007; J Royal Statist Soc B 69: 347-368
- [7] Noma and Matsui, 2012; Statist Med 31, 165-176
- [8] Ripke et al., 2013; nature genetics 45: 1150-1159
- [9] Purcell et al., 2007; Am J Hum Genet 81: 559-575
<http://pngu.mgh.harvard.edu/~purcell/plink/>

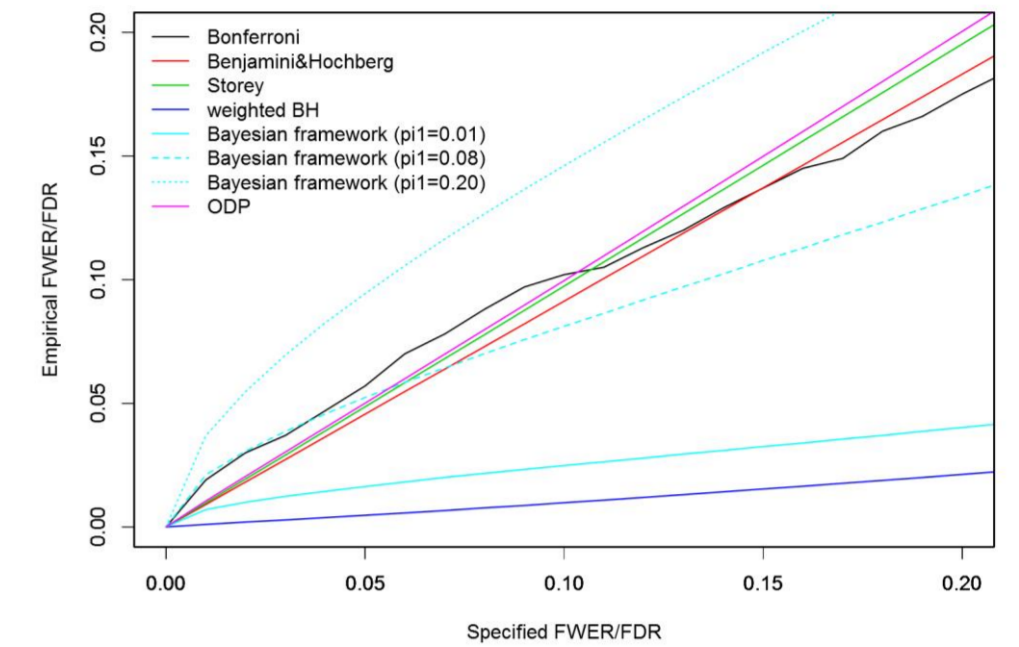


図1. 指定した有意水準に対する、実際のFWER/FDRの比較

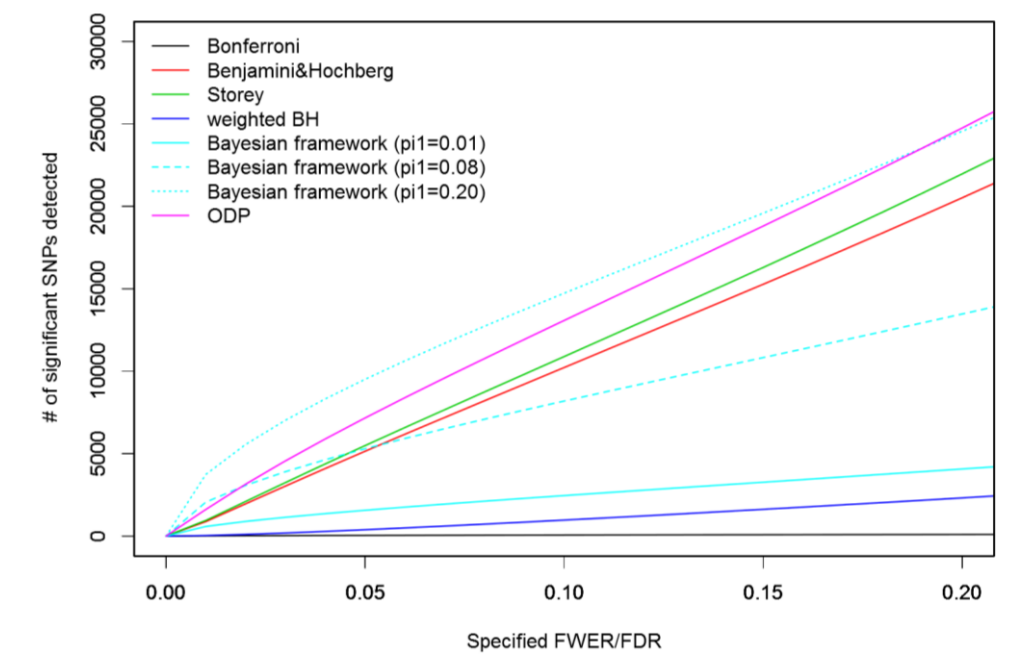


図2. 指定した有意水準のもとで、有意と判定されるSNPの数

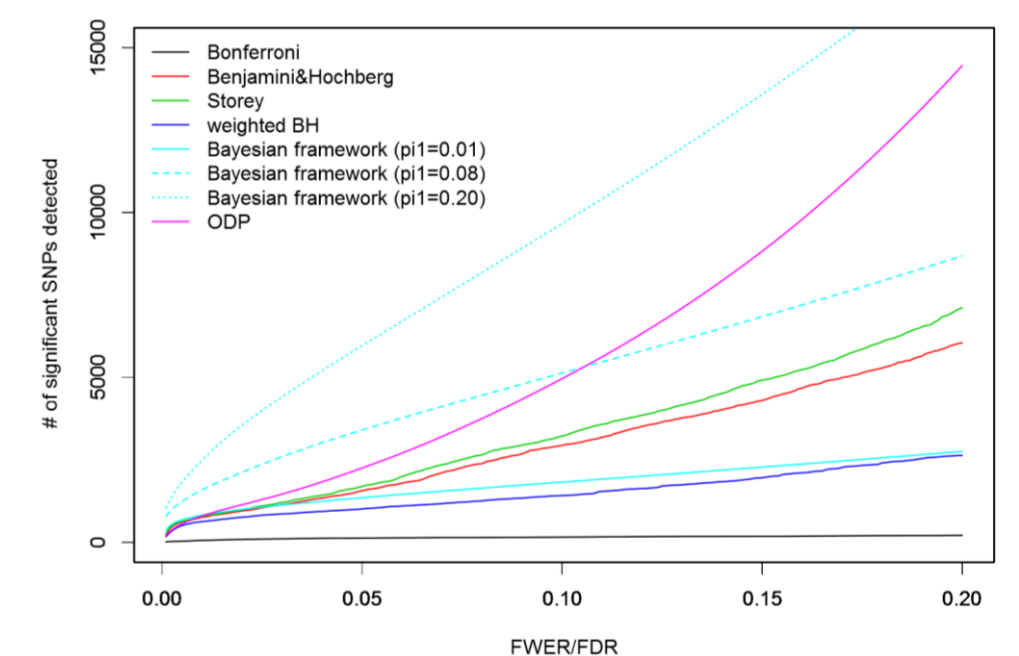


図3. 実データ[1]における、有意と判定されたSNPの数の比較 (Genome-wide significance基準によって検出されるのは136個)

表1. 文献[1]で報告された疾患関連SNPに対する、各補正手法での有意性のランキング

SNP rs ID	Chr.	BP	MAF	OR	SE	P	Ranking of significance							
							P&BH&ST	wBH	BF	BFDP (p1=0.01, 0.08, 0.20)		ODP	PP	
rs1625579	1	98275522	0.17	1.142	0.027	5.72E-07	395	395	385	3.53E-03	4.11E-04	1.43E-04	415	7.56E-03
rs17662626	2	193692866	0.12	1.222	0.043	3.09E-06	645	644	618	1.46E-02	1.71E-03	5.96E-04	1601	7.73E-02
rs2021722	6	30282110	0.20	0.843	0.026	4.30E-11	1	5	1	4.09E-07	4.75E-08	1.65E-08	1	1.61E-06
rs10503253	8	4168252	0.19	1.141	0.026	3.84E-07	333	333	331	2.33E-03	2.71E-04	9.44E-05	341	5.18E-03
rs7004633	8	89829427	0.17	0.858	0.027	1.45E-08	83	86	86	1.03E-04	1.20E-05	4.17E-06	55	3.37E-04
rs7914558	10	104765898	0.38	0.896	0.021	1.58E-07	210	210	235	1.30E-03	1.51E-04	5.25E-05	141	1.53E-03
rs11191580	10	104896201	0.08	1.227	0.037	2.23E-08	104	104	101	1.47E-04	1.70E-05	5.92E-06	129	1.32E-03
rs548181	11	124966919	0.11	0.832	0.033	2.91E-08	117	117	111	1.85E-04	2.15E-05	7.49E-06	109	9.30E-04
rs12966547	18	50903015	NA	0.902	0.021	1.00E-06	492	492	510	7.05E-03	8.24E-04	2.87E-04	362	5.91E-03
rs17512836	18	51345959	0.03	0.711	0.061	2.35E-08	106	106	118	2.09E-04	2.42E-05	8.43E-06	943	4.12E-02

表2. ODPランキングでの上位100SNPのうち、P値では上位にならないSNP

SNP rs ID	Chr.	BP	MAF	OR	SE	P	Ranking of significance							
							P&BH&ST	wBH	BF	BFDP (p1=0.01, 0.08, 0.20)		ODP	PP	
rs1107592	7	2007958	0.34	1.124	0.021	5.28E-08	141	141	142	3.91E-04	4.54E-05	1.58E-05	75	5.79E-04
rs10226475	7	2192688	0.32	1.124	0.021	5.06E-08	139	139	147	4.18E-04	4.86E-05	1.69E-05	80	6.12E-04
rs13196606	6	28478057	0.22	0.876	0.024	3.17E-08	122	122	124	2.34E-04	2.72E-05	9.45E-06	77	5.82E-04
rs615672	6	32682149	0.21	0.886	0.022	3.07E-08	120	121	125	2.40E-04	2.78E-05	9.68E-06	62	4.29E-04
rs735765	6	28278276	0.12	0.844	0.031	2.88E-08	116	116	112	1.92E-04	2.23E-05	7.77E-06	93	7.51E-04
rs1339898	6	28503485	0.29	0.879	0.023	2.71E-08	112	112	122	2.20E-04	2.55E-05	8.88E-06	69	5.07E-04
rs2858867	6	32683303	0.22	1.130	0.022	2.42E-08	107	107	116	2.06E-04	2.40E-05	8.33E-06	61	3.74E-04
rs1264350	6	30904524	0.13	1.187	0.031	2.22E-08	103	103	103	1.53E-04	1.78E-05	6.18E-06	95	7.78E-04
rs3130377	6	30431372	0.12	0.829	0.034	2.17E-08	101	102	99	1.44E-04	1.67E-05	5.81E-06	100	7.93E-04