Statistics and Its Interface Volume 0 (2012) 1

Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method

Jin Liu[*] Kai Wang Shuangge Ma and Jian Huang

APPENDIX A. APPENDIX SECTION

A.1 Accommodating Case-Control data with Logistic Regression

To accomandate the properties of case-control data, we use the marginal logistic regression with the proposed SMCP penalty.

$$\begin{aligned} L_n(\boldsymbol{\beta}) &= -\sum_{j=1}^{p}\frac{1}{n_j}\sum_{i=1}^{n_j}(y_{ij}\text{log}p_{ij}+(1-y_{ij})\text{log}p_{ij}) \\ &\quad+\sum_{j=1}^{p}\rho(|\beta_j|;\lambda,\gamma)|+\frac{1}{2}\lambda_2\sum_{j=1}^{p-1}\zeta_j(|\beta_{j+1}|-|\beta_j|)^2. \qquad (1)\end{aligned}$$

where $p_{ij}=\frac{e^{\beta_{0j}+x_{ij}\beta_j}}{1+e^{\beta_{0j}+x_{ij}\beta_j}}$, $\rho(t;\lambda,\gamma)$ is defined in Section 2.

Then, quadratic approximation can be applied piecewise to index j by using following equations.

$$\begin{aligned} z_{ij} &= \hat{\beta}_{0j}+x_{ij}\beta_j+\frac{y_{ij}-\tilde{p}_{ij}}{\tilde{p}_{ij}(1-\tilde{p}_{ij})} \\ w_{ij} &= \tilde{p}_{ij}(1-\tilde{p}_{ij})\end{aligned}$$

The new objective function after quadratic approximation is given as follows.

$$\begin{aligned} L_n(\boldsymbol{\beta}) &= \sum_{j=1}^{p}\frac{1}{2n_j}\sum_{i=1}^{n_j}w_{ij}(z_{ij}-\hat{\beta}_{0j}-x_{ij}\beta_j)^2 \\ &\quad+\sum_{j=1}^{p}\rho(|\beta_j|;\lambda,\gamma)|+\frac{1}{2}\lambda_2\sum_{j=1}^{p-1}\zeta_j(|\beta_{j+1}|-|\beta_j|)^2. \qquad (2)\end{aligned}$$

β_0 can be omitted for linear model by centering the response variable, but it must be included in the model for logistic regression .β_0s can be fitted marginal logistic regression and then fixed in objective function (2). ζ_js are defined the same as in Section 2. Then algorithm implemented in marginal linear regression with the SMCP penalty can be used to solve the marginal logistic regression with the SMCP penalty.

A.2 Application to Rheumatoid Arthritis Data

Due to the computational burden, we conduct the analysis for rheumatoid arthritis data only on chromosome 6 by marginal logistic regression. The plots of estimates by the SMCP, the MCP and the LASSO methods are presented in Fig. 5 and their significance estimates are large dots. By cross-sectional comparison with the results in Section 6.2, we found that there are 559 overlapping SNPs by the SMCP metho, in which 293 SNPs are significant. There are 535 overlapping SNPs by the MCP method, in which 293 SNPs are significant. while the LASSO method identifies the same set of SNPs. From simulation result and analysis results in Section 5, we see that despite that the logistic regression is a more natural choice for case-control studies, marginal linear regression can capture the pattern of SNPs' effect in GWAS. Furthermore, the computational burden prohibit us from conducting genome-wide scan by using marginal logistic regression, but it is possible to conduct it by marginal linear regression.

[Figure 1 about here.]

A.3 Application to dominant model with Heterogeneous Stock Mice Data

The proposed approach can be implemented to dominant and recessive models as well as additive model described in Section 2 to Section 6. We choose predetermined number to be 400 for the SMCP, the MCP and the LASSO methods. The multi-split method is used to evaluate the significance of the selected SNPs. The manhattan plots for all three methods are shown in Fig. 6. The large dots stand for SNPs with significant multi-split p-values while small dots for insignificant SNPs.

[Figure 2 about here.]

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

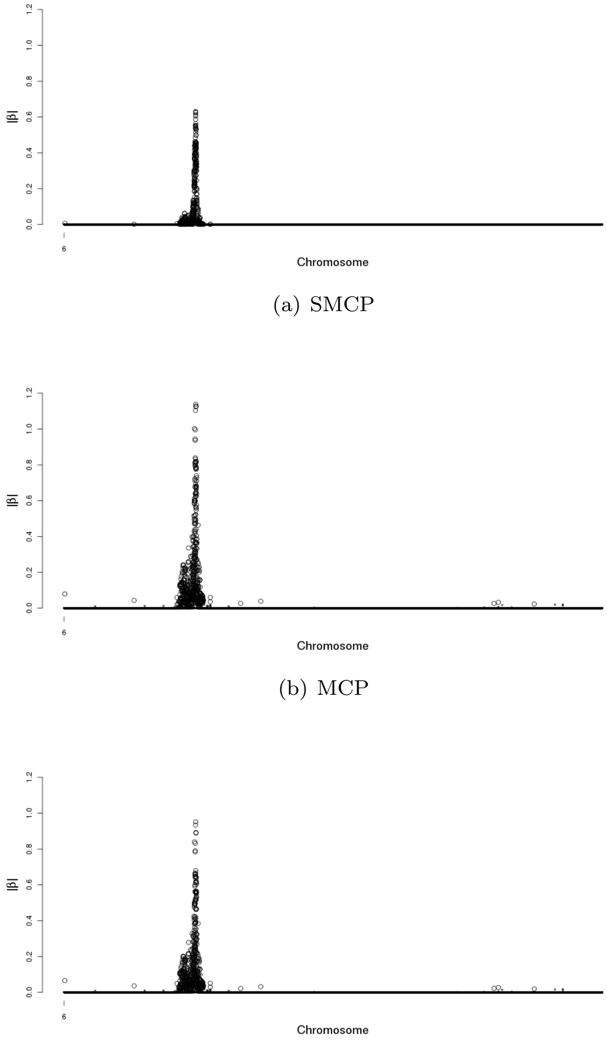
[*] Corresponding author

Jin Liu
School of Public Health, Yale University
New Haven, CT 06520, U.S.A.
E-mail address: jin.liu.jl2329@yale.edu

Kai Wang
Department of Biostatistics, University of Iowa
Iowa City, IA 52242, U.S.A.
E-mail address: jian-huang@uiowa.edu

Shuangge Ma
School of Public Health, Yale University
New Haven, CT 06520, U.S.A.
E-mail address: shuangge.ma@yale.edu

Jian Huang
Department of Statistics and Actuarial Science
Department of Biostatistics, University of Iowa
Iowa City, IA 52242, U.S.A.
E-mail address: jian-huang@uiowa.edu

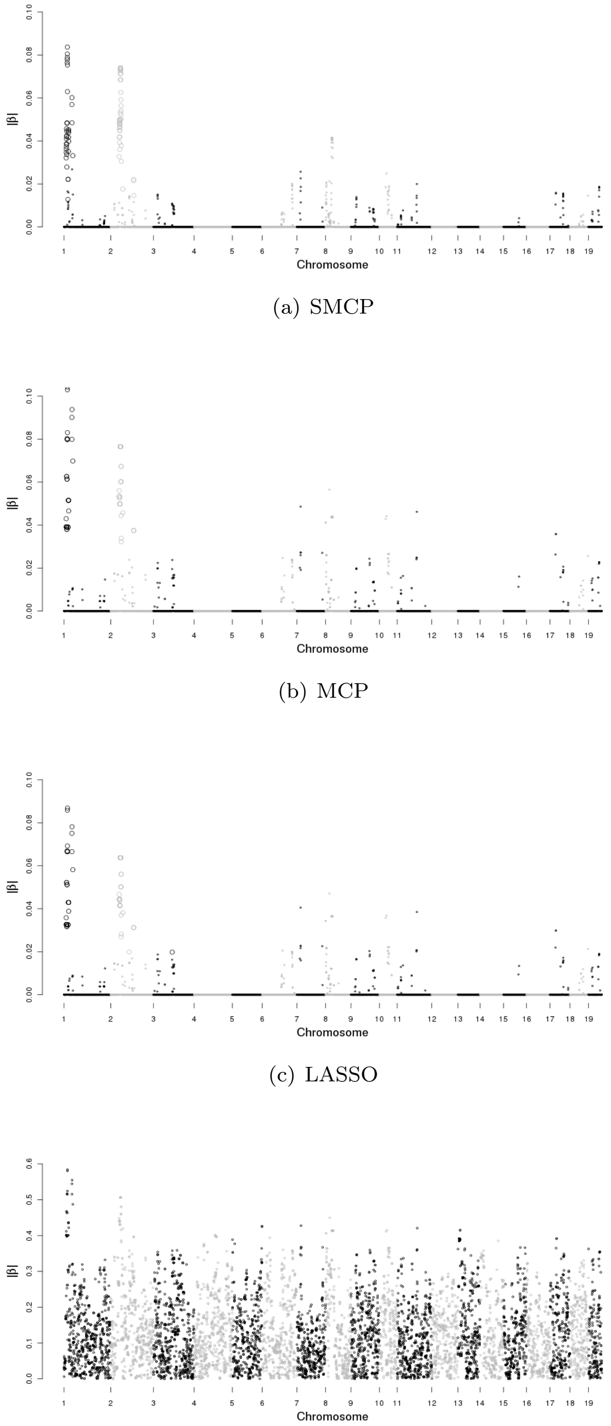


(a) SMCP

(b) MCP

(c) LASSO

Figure 5. Genome-wide plot of $|\beta|$ estimates for RA data on chromosome 6 by marginal logistic loss function.



(a) SMCP

(b) MCP

(c) LASSO

(d) Regular Single-SNP Linear Regression

Figure 6. Genome-wide plot of $|\beta|$ *estimates for heterogeneous stock mice data by dominant genetic model.*

Table 4. List of SNPs selected by the SMCP, the MCP and the LASSO method for a simulated data set with quantitative trait. Recall that the 31 disease-associated SNPs are 2287 – 2298 and 2300 – 2318.

	SMCP		MCP		LASSO		Regression	
SNP	$\|\hat{\beta}\|$	p-value*	$\|\hat{\beta}\|$	p-value*	$\|\hat{\beta}\|$	p-value*	$\|\hat{\beta}\|$	p-value**
1866			-0.011	1.000	-0.005	1.000	-0.211	1.2E-04
2144	-3.6E-04	1.000	-0.044	0.031	-0.019	0.056	-0.227	3.3E-05
2167			-0.038	0.034	-0.017	0.090	-0.225	4.1E-05
2171	-0.029	0.168	-0.096	1.000	-0.043	1.000	-0.253	3.6E-06
2173	-0.026	0.112	-0.115	1.000	-0.051	1.000	-0.262	1.6E-06
2178	0.001	1.000	0.043	0.015	0.019	0.078	0.227	3.3E-05
2209			0.005	0.080	0.002	0.097	0.208	1.5E-04
2212			0.002	0.080	0.001	0.098	0.207	1.6E-04
2235			0.027	0.050	0.012	0.040	0.219	6.4E-05
2240	0.042	0.013	0.162	0.008	0.072	0.004	0.286	1.5E-07
2241	0.037	0.241	0.038	0.031	0.017	0.030	0.225	4.1E-05
2242	0.052	0.019	0.139	0.012	0.062	0.005	0.274	5.0E-07
2243	0.035	0.193	0.069	1.000	0.031	1.000	0.240	1.2E-05
2269	-0.065	0.015	-0.167	0.014	-0.074	0.005	-0.288	1.2E-07
2270	0.059	0.034	0.097	0.032	0.043	0.021	0.254	3.5E-06
2271	-0.038	0.059	-0.121	0.024	-0.054	0.024	-0.265	1.1E-06
2272	-0.009	0.950	-0.007	0.057	-0.003	0.095	-0.210	1.3E-04
2275			-0.029	1.000	-0.013	1.000	-0.220	6.0E-05
2279	-0.081	0.002	-0.237	1.000	-0.105	1.000	-0.322	2.7E-09
2281	-0.016	0.413	-0.080	1.000	-0.036	1.000	-0.245	7.2E-06
2284	-0.048	0.011	-0.159	0.007	-0.071	0.006	-0.284	1.8E-07
2285	0.039	0.470					0.205	1.9E-04
2286	-0.183	3.0E-04	-0.265	1.000	-0.118	1.000	-0.336	5.1E-10
2287	0.274	3.3E-04	0.271	3.1E-04	0.120	0.001	0.339	3.5E-10
2288	0.287	3.3E-04	0.277	2.7E-04	0.123	0.001	0.342	2.4E-10
2289	-0.352	6.0E-05	-0.412	3.2E-05	-0.183	8.1E-05	-0.409	2.0E-14
2290	0.428	3.1E-11	0.841	1.000	0.374	1.000	0.619	1.6E-34
2291	-0.037	0.187					-0.159	0.004
2293	0.201	4.9E-07	0.524	6.3E-07	0.233	5.1E-06	0.463	1.7E-18
2294	0.190	0.001	0.294	1.1E-04	0.131	0.001	0.351	8.2E-11
2295	-0.121	4.6E-04	-0.252	1.6E-04	-0.112	0.001	-0.330	1.1E-09
2296	0.035	1.000					0.159	0.004
2297	-0.015	0.211	-0.077	0.064	-0.034	0.031	-0.244	8.4E-06
2299	0.054	1.000					0.033	5.5E-01
2300	0.716	1.8E-15	0.643	2.3E-16	0.456	7.2E-15	0.711	4.0E-48
2301	-0.789	2.2E-19	-0.706	8.2E-19	-0.520	1.6E-17	-0.781	7.4E-62
2302	0.718	2.7E-12	0.913	1.4E-13	0.406	1.3E-12	0.655	2.6E-39
2303	-0.401	0.089					-0.191	5.1E-04
2304	-0.615	4.4E-17	-0.681	5.9E-18	-0.494	3.3E-18	-0.753	6.3E-56
2305	-0.531	8.5E-10	-0.762	1.7E-10	-0.339	1.2E-09	-0.580	9.0E-30
2306	0.384	0.290					0.175	0.002
2307	-0.406	1.5E-06	-0.559	1.000	-0.249	1.000	-0.481	6.1E-20
2308	0.237	0.114					0.195	3.8E-04
2309	0.359	6.9E-09	0.695	1.8E-10	0.309	7.3E-10	0.547	3.5E-26
2310	-0.291	3.5E-05	-0.452	1.000	-0.201	1.000	-0.428	8.4E-16
2312	0.153	4.7E-04	0.331	1.000	0.147	1.000	0.369	7.2E-12
2313	0.146	0.092	0.047	1.000	0.021	1.000	0.229	2.9E-05
2314	-0.276	6.6E-05	-0.368	8.8E-05	-0.164	4.1E-05	-0.387	5.4E-13
2315	0.296	6.6E-05	0.368	8.8E-05	0.164	4.1E-05	0.387	5.4E-13
2316	-0.322	3.4E-07	-0.597	1.88E-07	-0.265	1.21E-07	-0.499	1.5E-21
2317	-0.260	0.005	-0.181	0.003	-0.081	0.002	-0.295	5.8E-08
2318	0.228	0.003	0.236	1.000	0.105	1.000	0.322	2.8E-09
2320	0.014	0.735	0.065	0.009	0.029	0.015	0.238	1.4E-05
2321	-0.012	0.992	-0.055	0.009	-0.024	0.018	-0.233	2.1E-05
2337	-0.087	0.002	-0.317	1.000	-0.141	1.000	-0.362	1.8E-11
2363			-0.024	0.047	-0.011	0.054	-0.218	7.1E-05
2371	-0.023	0.035	-0.124	0.023	-0.055	0.005	-0.267	1.0E-06

* Computed using the multi-split method.

** Single SNP analysis, not corrected for multiple testing.

*** Empty cells stand for SNPs that are not identified from the model

Table 5. List of SNPs selected by the SMCP and the LASSO method for a simulated data set with binary trait. The analysis is based on marginal negative log-likelihood loss. Recall that the 31 disease-associated SNPs are 2287 – 2298 and 2300 – 2318.

	SMCP		MCP		LASSO		Regression	
SNP	$\lvert\hat{\beta}\rvert$	p-value*	$\lvert\hat{\beta}\rvert$	p-value*	$\lvert\hat{\beta}\rvert$	p-value*	$\lvert\hat{\beta}\rvert$	p-value**
366			-0.009	1.000	-0.004	1.000	-0.071	0.004
368	-0.001	1.000	-0.045	1.000	-0.020	1.000	-0.075	0.002
506	-0.002	1.000	-0.103	1.000	-0.043	1.000	-0.081	0.001
656	0.001	1.000	0.056	1.000	0.025	1.000	0.077	0.002
932			0.001	1.000	0.001	1.000	0.071	0.005
948			0.020	1.000	0.009	1.000	0.073	0.004
1047			0.009	1.000	0.004	1.000	0.071	0.004
1476			-0.003	1.000	-0.001	1.000	-0.071	0.005
1477			0.025	1.000	0.011	1.000	0.073	0.003
1478			-0.011	1.000	-0.005	1.000	-0.072	0.004
1678	-0.001	1.000	-0.033	1.000	-0.015	1.000	-0.074	0.003
1978	-0.008	1.000	-0.195	0.788	-0.083	0.788	-0.091	2.6E-04
1980	-3.8E-05	1.000	-0.028	1.000	-0.012	1.000	-0.073	0.003
1990	0.005	1.000	0.068	1.000	0.030	1.000	0.078	0.002
2048	0.001	1.000	0.039	1.000	0.016	1.000	0.074	0.003
2283	0.002	1.000					0.060	0.017
2284	-0.030	1.000	-0.108	1.000	-0.047	1.000	-0.082	0.001
2285	0.034	1.000					0.060	0.016
2286	-0.144	0.015	-0.436	0.026	-0.180	0.026	-0.113	4.9E-06
2287	0.150	0.425	0.168	0.720	0.072	0.720	0.088	3.9E-04
2288	0.151	0.354	0.187	0.615	0.080	0.615	0.090	2.9E-04
2289	-0.152	0.218	-0.192	1.000	-0.077	1.000	-0.089	3.6E-04
2290	0.152	1.0E-04	0.751	8.1E-05	0.313	8.1E-05	0.144	4.2E-09
2291	-0.034	1.000					-0.054	0.031
2292	-0.018	1.000					-0.006	0.820
2293	0.065	0.014	0.444	0.013	0.187	0.013	0.116	2.8E-06
2294	0.067	0.126	0.268	0.191	0.117	0.191	0.099	6.2E-05
2295	-0.048	0.629	-0.167	1.000	-0.072	1.000	-0.088	3.9E-04
2296	0.030	1.000					0.061	0.014
2299	-0.097	1.000					-0.021	0.399
2300	0.275	2.0E-04	0.553	0.002	0.238	0.002	0.128	0.000
2301	-0.307	2.3E-06	-0.887	2.3E-06	-0.438	2.3E-06	-0.170	2.4E-12
2302	0.294	1.9E-04	0.684	3.1E-04	0.278	3.1E-04	0.136	3.0E-08
2303	-0.211	1.000					-0.048	0.053
2304	-0.206	1.1E-05	-0.876	1.1E-05	-0.371	1.1E-05	-0.157	1.4E-10
2305	-0.176	0.003	-0.490	0.008	-0.196	0.008	-0.118	1.9E-06
2306	0.131	1.000					0.020	0.421
2307	-0.076	1.000	-0.003	1.000	-0.001	1.000	-0.071	0.005
2308	0.041	1.000					0.053	0.034
2309	0.053	0.117	0.313	0.134	0.130	0.134	0.102	3.7E-05
2310	-0.040	1.000	-0.148	1.000	-0.061	1.000	-0.085	0.001
2316	-0.005	0.753	-0.216	0.591	-0.086	0.591	-0.091	2.4E-04
2329			0.003	1.000	0.001	1.000	0.071	0.005
2337	-0.016	0.328	-0.299	0.253	-0.113	0.253	-0.097	9.8E-05
2360	-0.002	1.000	-0.055	1.000	-0.024	1.000	-0.076	0.002
2362			-0.028	1.000	-0.012	1.000	-0.073	0.003
2461	0.001	1.000	0.049	1.000	0.020	1.000	0.075	0.003
2550	0.009	1.000					0.068	0.007
2551	0.038	0.460	0.269	0.514	0.100	0.514	0.093	1.7E-04
2552	-0.033	1.000	-0.134	1.000	-0.057	1.000	-0.085	0.001
2553	0.029	1.000	0.146	1.000	0.062	1.000	0.086	0.001
2554	-0.015	1.000					-0.056	0.024
2912	0.001	1.000	0.031	1.000	0.014	1.000	0.074	0.003
3140	0.002	1.000	0.066	1.000	0.028	1.000	0.077	0.002
3329	0.015	1.000	0.117	1.000	0.050	1.000	0.083	0.001
3388	0.001	1.000	0.045	1.000	0.020	1.000	0.075	0.002
3620	0.001	1.000	0.053	1.000	0.023	1.000	0.076	0.002
4018	0.006	0.576	0.243	0.598	0.096	0.598	0.094	1.5E-04
4078	0.002	1.000	0.059	1.000	0.026	1.000	0.077	0.002
4745			-0.007	1.000	-0.003	1.000	-0.071	0.004
4877			0.007	1.000	0.003	1.000	0.071	0.004

* Computed using the multi-split method.

** Single SNP analysis, not corrected for multiple testing.

*** Empty cells stand for SNPs that are not identified from the model

Table 6. List of SNPs selected by the SMCP and the LASSO method for a simulated data set with binary trait. The analysis is based on marginal least-square loss. Recall that the 31 disease-associated SNPs are 2287 – 2298 and 2300 – 2318.

	SMCP		MCP		LASSO		Regression	
SNP	$\lvert\hat{\beta}\rvert$	p-value*	$\lvert\hat{\beta}\rvert$	p-value*	$\lvert\hat{\beta}\rvert$	p-value*	$\lvert\hat{\beta}\rvert$	p-value**
366			-0.002	1.000	-0.002	1.000	-0.071	0.004
368	-0.002	1.000	-0.012	1.000	-0.010	1.000	-0.075	0.002
506	-0.005	1.000	-0.025	1.000	-0.021	1.000	-0.081	0.001
656	0.002	1.000	0.015	1.000	0.012	1.000	0.077	0.002
932			3.4E-04	1.000	2.9E-04	1.000	0.071	0.005
948	0.001	1.000	0.005	1.000	0.004	1.000	0.073	0.004
1047			0.002	1.000	0.002	1.000	0.071	0.004
1476	0.000	1.000	-0.001	1.000	-0.001	1.000	-0.071	0.005
1477	0.001	1.000	0.006	1.000	0.005	1.000	0.073	0.003
1478	-0.001	1.000	-0.003	1.000	-0.002	1.000	-0.072	0.004
1678	-0.002	1.000	-0.009	1.000	-0.007	1.000	-0.074	0.003
1978	-0.013	0.240	-0.049	0.230	-0.041	0.230	-0.091	2.6E-04
1980	-0.001	1.000	-0.007	1.000	-0.006	1.000	-0.073	0.003
1990	0.008	1.000	0.018	1.000	0.015	1.000	0.078	0.002
2048	0.003	1.000	0.009	1.000	0.008	1.000	0.074	0.003
2284	-0.009	1.000	-0.028	1.000	-0.023	1.000	-0.082	0.001
2285	0.005	1.000					0.060	0.016
2286	-0.076	0.006	-0.102	0.006	-0.085	0.006	-0.113	4.9E-06
2287	0.049	0.250	0.043	0.442	0.036	0.442	0.088	3.9E-04
2288	0.051	0.222	0.047	0.282	0.039	0.282	0.090	2.9E-04
2289	-0.060	0.206	-0.044	0.328	-0.037	0.328	-0.089	3.6E-04
2290	0.093	0.001	0.177	0.001	0.147	0.001	0.144	4.2E-09
2291	-0.003	1.000					-0.054	0.031
2293	0.051	0.028	0.109	0.028	0.091	0.028	0.116	2.8E-06
2294	0.049	0.153	0.069	0.259	0.058	0.259	0.099	6.2E-05
2295	-0.028	0.187	-0.042	0.500	-0.035	0.500	-0.088	3.9E-04
2296	0.007	1.000					0.061	0.014
2300	0.122	0.009	0.138	0.009	0.115	0.009	0.128	2.1E-07
2301	-0.148	4.2E-05	-0.240	4.2E-05	-0.200	4.2E-05	-0.170	2.4E-12
2302	0.126	0.003	0.158	0.003	0.131	0.003	0.136	3.0E-08
2303	-0.040	0.707					-0.048	0.053
2304	-0.090	0.001	-0.207	0.001	-0.172	0.001	-0.157	1.4E-10
2305	-0.060	0.027	-0.113	0.027	-0.095	0.027	-0.118	1.9E-06
2306	0.007	1.000					0.020	0.421
2307	-0.001	1.000	-0.001	1.000	-0.001	1.000	-0.071	0.005
2309	0.030	0.081	0.076	0.081	0.064	0.081	0.102	3.7E-05
2310	-0.024	0.313	-0.035	0.689	-0.029	0.689	-0.085	0.001
2316	-0.010	0.299	-0.050	0.214	-0.041	0.214	-0.091	2.4E-04
2329			0.001	1.000	0.001	1.000	0.071	0.005
2337	-0.022	0.238	-0.063	0.238	-0.052	0.238	-0.097	9.8E-05
2360	-0.005	1.000	-0.014	1.000	-0.012	1.000	-0.076	0.002
2362	-0.002	1.000	-0.007	1.000	-0.006	1.000	-0.073	0.003
2461	0.002	1.000	0.011	1.000	0.010	1.000	0.075	0.003
2550	0.003	1.000					0.068	0.007
2551	0.031	0.172	0.055	0.172	0.046	0.172	0.093	1.7E-04
2552	-0.022	0.639	-0.034	1.000	-0.028	1.000	-0.085	0.001
2553	0.018	0.768	0.037	0.902	0.031	0.902	0.086	0.001
2912	0.003	1.000	0.008	1.000	0.007	1.000	0.074	0.003
3140	0.004	1.000	0.016	1.000	0.014	1.000	0.077	0.002
3329	0.016	1.000	0.029	1.000	0.024	1.000	0.083	0.001
3388	0.003	1.000	0.012	1.000	0.010	1.000	0.075	0.002
3620	0.002	1.000	0.013	1.000	0.011	1.000	0.076	0.002
4018	0.011	0.124	0.057	0.124	0.047	0.124	0.094	1.5E-04
4078	0.004	1.000	0.015	1.000	0.013	1.000	0.077	0.002
4745			-0.002	1.000	-0.001	1.000	-0.071	0.004
4877			0.002	1.000	0.002	1.000	0.071	0.004

* Computed using the multi-split method.

** Single SNP analysis, not corrected for multiple testing.

*** Empty cells stand for SNPs that are not identified from the model