基因型与表型关联分析

基因型数据可以是 SNP 数据（0=AA 1=AB 2=BB）也可以是连续性的基因表达数据。表型数据可以是连续性变量，二分类变量，时间依赖的生存状态变量等。

检验每个位点或每个基因表达和表现型之间的关系，用于对大量基因位点进行初筛，帮助快速发现有意义的基因位点。

对于每个基因位点 SNP 数据，将分别采用显性、隐性、相加三种模型进行检验。如果表现型是连续变量，系统将对每种基因型（AA，AB 或 BB）分别计算出表现型的调整均值。

对于连续性的基因表达数据，将分别采用直线回归（输出回归系数、SE、P 值）与曲线拟合（输出自由度与 P 值），如果表型数据是时间依赖的生存状态，将采用 COX 回归分析。研究设计也可以是配对的病例对照研究，选用 cox 回归并给定配对组编号，软件将采用条件 logistic 回归分析。

基因型数据可以与表现型数据存在一个数据文件内，也可以直接读取.ped 文件（存放 SNP 数据的 pedgree 文件）或制表符分隔（或空格或逗号）的.txt 文本文件。

.ped 文件格式要求：

1. 第一行列出 SNP 标记名，第二行开始为每个测量对象的数据；
2. 每行数据按如下顺序：家系编号、个人编号、父亲编号、母亲编号、性别、是否先证者，然后是每个 SNP 的两个等位基因编码（1=A、2=C、3=G、4=T）；
3. 所有数据必须是数字型，字段间用空格分隔。

.txt 文件格式要求：

1. 第一行列出变量名（GENE ID），第二行开始为每个测量对象的数据；
2. 第一列为研究对象编号，第二列开始为每个基因的数据，如果是 SNP 数据必须编码为：0=AA 1=AB 2=BB；
3. 所有数据必须是数字型，字段间用制表符（或空格或逗号）分隔。

如果基因型数据来自.ped 文件或.txt 文件，要求给出表型数据里的研究对象编码（ID）变量，以便与之链接。

SNP 与表型关联分析输出表格列中包括：

- 基因位点名称
- 表现型名称
- 组别（如果指定了分层变量）
- N0, N1, N2：不同基因型（AA、AB、BB）的观察对象个数
- b（add），se（add），p（add）：相加模型的回归系数、标准误和 P 值
- b（dom），se（dom），p（dom）：显性模型的回归系数、标准误和 P 值
- b（rec），se（rec），p（rec）：隐性模型的回归系数、标准误和 P 值
- lsmean0, lsmean1, lsmean2：不同基因型（AA、AB、BB）人群调整的表型变量的均值（仅适用于连续性的表现型变量）

- FDRCUT：根据相加模型 p 值计算的 FDR（false discovery rate）切点，p<=0.05*k/m，k 为按 p 值排序的序数，m为总的位点数（Benjamini and Hochberg 1995）。如果观察的 P 值小于该值，表示 FDR 校正的 P<0.05。

基因表达数据与表型关联分析输出表格列中包括：

- 各基因表达的统计量：N mean sd min 5% 10% 25% 50% 75% 90% 95% max p-normal(pearson test for 正态性检验的 p 值)
- 各基因表达与表型的关联分析结果

例 1：对练习数据 demo.xls 中的 SNP1、SNP2 与 FEV1、FVC 的关联关系进行分析。因为数据来自家系，需要用 GEE 调整同一家系内成员的内部相关性。同时要调整 AGE、HEIGHT、WEIGHT、SEX 对表现型的影响。

输入界面如下：

输出结果如下：

SNPs 关联分析

SNP	结局变量	N0	N1	N2	MEAN0	MEAN1	MEAN2	beta-add	se-add	P-add	beta-dom	se-dom	P-dom	beta-rec	se-rec	P-rec	FDRCUT
SNP1	FEV1	438	237	36	3.3431	3.6125	3.6597	0.0818	0.0525	0.11942715	0.1201	0.0624	0.05422143	-0.0024	0.1360	0.98616133	0.025
SNP2	FEV1	384	266	54	3.5879	3.3418	3.0194	-0.1890	0.0455	3.31e-05	-0.2308	0.0574	5.73e-05	-0.2643	0.0914	0.00382108	0.05
SNP1	FVC	438	237	36	4.3971	4.7162	4.8903	0.1398	0.0497	0.00487977	0.1899	0.0619	0.00214977	0.0728	0.1255	0.56223487	0.025
SNP2	FVC	384	266	54	4.6738	4.4359	3.9081	-0.2232	0.0459	1.19e-06	-0.2461	0.0603	4.52e-05	-0.4051	0.0886	4.79e-06	0.05

结局变量：FEV1 和 FVC
应用广义估计方程(GEE) 研究对象=FMYID,相关类型=independence
N0，N1，N2：分别表示基因型为 0，1，2 的人数
MEAN0，MEAN1，MEAN2：分别表示基因型为 0，1，2 的均数
beta-add，se-add，P-add：分别表示相加模型中的回归系数，标准误，P 值
beta-dom，se-dom，P-dom：分别表示显性模型中的回归系数，标准误，P 值
beta-rec，se-rec，P-rec：分别表示隐性模型中的回归系数，标准误，P 值
fdrcut：根据相加模型 p 值计算的 FDR 切点，p<=0.05*k/m (Benjamini and Hochberg 1995)

例 2. 下载练习数据 pheno.xls 与 geno.xls，分析 35 个（G1-G35）基因表达与表型 SURV 的关系，DAYS 为生存时间。
http://r.empowerstats.cn/empowerStats/exdata/pheno.xls
http://r.empowerstats.cn/empowerStats/exdata/geno.xls

输入界面如下图所示：

点击“浏览”找到 geno.xls 文件，注意：如果手动输入文件路径，必须将 windows 路径里的“\”改为 “/”。

点击查看结果，结果输出如下：

基因型与表型关联分析

Geno expression

GENE.ID	unique	N	mean	sd	min	5%	10%	25%	50%	75%	90%	95%	max	p-normal
G1	257	260	12.09	7.82	0.00	2.95	4.01	6.27	10.52	16.32	23.64	26.25	41.64	<0.0001
G2	259	260	9.2	5.99	0.00	2.06	3.12	4.90	8.21	11.86	15.80	20.4	40.78	<0.0001
G3	260	260	234.55	86.04	87.09	118.76	133.3	169.25	226.08	277.38	348.84	411.58	576.38	<0.0001
G4	260	260	1004.63	336.41	264.52	598.03	637.14	781.71	962.3	1180.25	1347.96	1530.42	3043.84	0.0061
G5	260	260	185.85	110.18	22.66	55.44	62.83	105.92	170.89	241.62	322.53	425.94	607.19	<0.0001
G6	154	260	0.56	0.67	0.00	0.00	0.00	0.00	0.43	0.87	1.41	1.80	4.66	<0.0001
G7	201	260	1.79	4.99	0.00	0.00	0.00	0.41	0.82	1.88	3.43	5.16	69.74	<0.0001
G8	260	260	4.98	2.92	0.56	1.35	2.09	3.18	4.27	6.23	8.95	10.67	23.4	<0.0001
G9	43	260	0.10	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.59	1.86	<0.0001
G10	77	260	0.42	1.64	0.00	0.00	0.00	0.00	0.00	0.41	0.92	1.68	19.70	<0.0001
G11	260	260	491.59	393.19	42.55	93.89	141.93	206.42	397.38	655.78	968.46	1178.21	3131.6	<0.0001
G12	260	260	8.41	6.15	0.41	1.32	2.37	3.82	6.86	11.09	16.1	21.34	35.25	<0.0001
G13	141	260	10.62	19.13	0.00	0.00	0.00	0.00	2.35	12.61	32.82	43.89	143.51	<0.0001
G14	260	260	256.15	184.88	34.05	78.14	97.88	138.45	199.08	324.8	428.8	598.63	1473.92	<0.0001
G15	98	260	0.35	0.65	0.00	0.00	0.00	0.00	0.00	0.47	1.20	1.55	5.52	<0.0001
G16	129	260	21.12	201.31	0.00	0.00	0.00	0.00	0.00	1.09	6.52	21.02	3056.96	<0.0001
G17	260	260	903.7	356.32	223.75	436.37	519.3	688.34	822.01	1060.6	1310.08	1605.37	2522.12	<0.0001
G18	245	260	2.77	4.42	0.00	0.00	0.43	0.81	1.60	2.71	5.20	8.65	40.84	<0.0001
G19	260	260	23.79	20.46	3.71	6.76	8.24	12.09	19.42	28.44	39.12	53.68	186.92	<0.0001
G20	258	260	171.65	114.61	0.00	10.07	32.45	94.96	162.86	223.67	291.18	374.85	709.53	0.0006
G21	173	260	7.92	21.88	0.00	0.00	0.00	0.00	1.03	4.15	16.96	33.69	174.13	<0.0001
G22	260	260	167.62	72.81	28.62	73.55	89.45	122.32	154.55	198.78	259.00	297.68	472.12	0.0018
G23	121	260	27.51	209.21	0.00	0.00	0.00	0.00	0.00	0.90	2.83	16.64	2464.29	<0.0001
G24	260	260	6348.66	5613.28	430.28	1197.82	1597.91	2663.92	4672.57	7709.9	13198.25	16328.06	36199.69	<0.0001
G25	260	260	213.2	248.21	16.08	32.77	45.21	75.9	144.31	248.95	460.00	654.82	2825.05	<0.0001
G26	105	260	0.53	2.59	0.00	0.00	0.00	0.00	0.00	0.50	1.05	1.84	40.91	<0.0001
G27	143	260	1.95	3.85	0.00	0.00	0.00	0.00	0.42	2.15	5.73	8.97	25.39	<0.0001
G28	260	260	905.66	235.6	384.93	579.7	661.26	756.22	860.72	1017.97	1180.18	1375.59	1860.34	0.0019
G29	260	260	1054.67	351.24	314.94	560.88	662.47	813.74	1002.35	1249.65	1503.88	1738.34	2411.81	0.0004
G30	244	260	19.9	67.68	0.00	0.00	0.49	1.40	3.86	12.72	40.9	65.97	658.50	<0.0001
G31	260	260	156.45	95.63	3.41	36.51	49.64	81.64	146.84	206.37	276.55	358.75	452.48	<0.0001
G32	260	260	1216.45	339.41	76.39	754.57	849.32	1005.33	1195.96	1393.21	1622.68	1803.98	2921.8	0.1454
G33	259	260	639.92	421.03	102.03	221.07	296.76	399.35	532.56	731.72	1123.12	1517.06	2931.42	<0.0001
G34	260	260	3215.02	671.89	1605.94	2277.97	2500.87	2747.50	3103.35	3598.16	4174.96	4343.04	5651.53	0.0243
G35	151	260	0.59	0.86	0.00	0.00	0.00	0.00	0.41	0.79	1.44	2.19	5.92	<0.0001

Association tests

GENE.ID	Outcome	N	beta	se	p-value	FDRCUT
G1	SURV	260	-0.0218	0.0183	0.23247715	0.01428571
G2	SURV	260	-0.0249	0.0247	0.31446023	0.02000000
G3	SURV	260	0.0029	0.0015	0.05293515	0.00428571
G4	SURV	260	5e-040	5e-040	0.23614234	0.01571429
G5	SURV	260	2e-040	0.0012	0.85377152	0.04285714
G6	SURV	260	-0.1593	0.2246	0.47815569	0.02714286

G7	SURV	260	-0.0426	0.0638	0.50386185	0.02857143
G8	SURV	260	-0.0010	0.0466	0.98291383	0.05000000
G9	SURV	260	0.8193	0.5230	0.11722583	0.01142857
G10	SURV	260	0.0659	0.0584	0.25908500	0.01714286
G11	SURV	260	-4e-04	4e-040	0.35504632	0.02285714
G12	SURV	260	-0.0068	0.0240	0.77765090	0.03714286
G13	SURV	260	-0.0128	0.0094	0.17021306	0.01285714
G14	SURV	260	2e-040	7e-040	0.80708517	0.04000000
G15	SURV	260	0.1426	0.2318	0.53858138	0.03000000
G16	SURV	260	2e-040	9e-040	0.79356309	0.03857143
G17	SURV	260	-8e-04	4e-040	0.05850691	0.00571429
G18	SURV	260	0.0267	0.0284	0.34655845	0.02142857
G19	SURV	260	-0.0059	0.0114	0.60370594	0.03285714
G20	SURV	260	0.0019	0.0011	0.08891485	0.00857143
G21	SURV	260	0.0032	0.0061	0.59887522	0.03142857
G22	SURV	260	-6e-04	0.0018	0.74410406	0.03571429
G23	SURV	260	4e-040	4e-040	0.30185925	0.01857143
G24	SURV	260	0.0000	0.0000	0.86087523	0.04428571
G25	SURV	260	3e-040	4e-040	0.43259139	0.02571429
G26	SURV	260	0.0676	0.0303	0.02564796	0.00285714
G27	SURV	260	-0.0737	0.0432	0.08793972	0.00714286
G28	SURV	260	-0.0011	7e-040	0.10956734	0.01000000
G29	SURV	260	-1e-04	4e-040	0.73754756	0.03428571
G30	SURV	260	-8e-04	0.0038	0.82724115	0.04142857
G31	SURV	260	-2e-04	0.0015	0.90168258	0.04714286
G32	SURV	260	4e-040	4e-040	0.38942142	0.02428571
G33	SURV	260	0.0000	3e-040	0.93383657	0.04857143
G34	SURV	260	-7e-04	2e-040	0.00478826	0.00142857
G35	SURV	260	0.0217	0.1518	0.88648171	0.04571429

N: number of subjects with non-missing data
beta, se, p-value: regression coefficient, standard err, pvalue
edf, p-smooth: degree of freedom, pvalue from splint smoothing
FURCUT: FDR cut point for P-value from regression model based on p<=0.05*k/m
(Benjamini and Hochberg 1995)
Cox model time variable: DAYS
调整变量：SEX; AGE; STAGE
此表用易侕统计软件 (www.empowerstats.com) 和 R 软件生成，生成日期： 2017-05-18