# 14B0305A-2关联分析报告

## 一、项目简介

复杂疾病是由遗传因素与环境因素共同作用的结果，探索影响复杂疾病发生、发展的遗传因素，是遗传学的重要任务。研究人员期望从疾病个体和正常个体的比较中来发现基因组上的差别，进而寻找与疾病发生关联的基因或位点。

为寻找与疾病相关的位点，我们通常使用关联分析方法，根据已有样本的突变位点的基因型数据以及表型数据（如case/control数据、性别、年龄等），使用数学工具如逻辑回归、卡方、T检验等进行统计分析。

逻辑回归以单个位点为单位，按5个假设的遗传模型，Dominant (频率低的allele为显性)，Recessive(频率低的allele为隐性)，Log-additive (叠加作用)，HOM/HET（共显性模型，以纯合正常为参照）进行关联分析。

卡方分析同样对每个位点进行分析，并设4个分析模型，Codominant共显性、Dominant显性、Recessive隐性、Allele等位基因。

T检验主要用于检测表型在当前case/control分层下是否显著，并把显著的表型用于逻辑回归的校正中。

本项目的基本数据信息如下：

|  |  |
| --- | --- |
| SNP位点 | 9个 |
| Case样本 | 242个 |
| Control样本 | 310个 |
| 可用表型数据 | 2个 |

## 二、质控

（1）Hardy-Weinberg 平衡控制

在理想状态下，各等位基因的频率和等位基因的基因型频率在遗传中是稳定不变的，即保持着基因平衡。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNP | rs3825942 | rs4886467 | rs4558370 | rs2165241 |
| CHR | 15 | 15 | 15 | 15 |
| Ref allele | G | G | G | C |
| Alt allele | A | T | T | T |
| Total(11/01/00) | 6/117/429 | 70/239/243 | 8/123/421 | 112/241/199 |
| HWE | 0.6804 | 0.3463 | 1 | 0.01453 |
| Case(11/01/00) | 2/18/222 | 47/113/82 | 4/33/205 | 75/107/60 |
| Case\_refallele\_number | 462 | 277 | 443 | 227 |
| Case\_altallele\_number | 22 | 207 | 41 | 257 |
| HWE\_Case | 0.07615 | 0.5114 | 0.0753 | 0.09296 |
| Control(11/01/00) | 4/99/207 | 23/126/161 | 4/90/216 | 37/134/139 |
| Control\_refallele\_number | 513 | 448 | 522 | 412 |
| Control\_altallele\_number | 107 | 172 | 98 | 208 |
| HWE\_Control | 0.04508 | 0.8877 | 0.1363 | 0.6107 |

在结果HWE.xlsx文件里，我们对每一个位点划分所有样本、case样本以及control样本进行了HWE检验，从示例结果看出，rs2165241位点存在Pvalue<0.05，差异极显著，在该群体内并不是H-W平衡体。（一般case里面hwe显著，可能恰恰就是因为它和疾病发生相关，所以仍会继续进行后续分析。但如在control样本中HWE是显著的，则往往认为该SNP在该群体内并不是H-W平衡，进而会被删除，不再继续后续分析）。

（2）表型的质控

能够与疾病产生关联的除了样本的基因型外，表型同样与样本的患病与否存在关联性（如是否抽烟可能影响患肺癌的风险）。根据表型的数据类型，可划分为连续型变量和分类变量。对连续型变量我们使用T检验进行分析，如年龄；而对分类变量，可使用卡方检验进行分析，如性别。

T检验结果示例：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 年龄 |  |  |  |  |  |  |
| 项目 | 例数 | 均值 | 中位数 | 标准差 | 最小值 | 最大值 |
| case | 242 | 69.90277778 | 70 | 8.470731643 | 48 | 90 |
| control | 310 | 62.45791246 | 64 | 9.940315793 | 42 | 89 |
| t | 9.130273424 |  |  |  |  |  |
| p | 1.11022E-16 |  |  |  |  |  |

卡方检验结果示例：

|  |  |  |
| --- | --- | --- |
| 性别 |  |  |
| 项目 | 1 | 2 |
| case | 164 | 78 |
| control | 167 | 143 |
| t | 10.25095796 |  |
| p | 0.0013661 |  |

从结果PhenoTest.xlsx文件里给出的两个检测结果来看，年龄和性别与样本的患病与否存在显著的关联性。

## 三、SNP位点关联分析

#### 3.1 SNP位点卡方检验

对所有SNP位点进行卡方检验分析，分析模型有Codominant共显性、Dominant显性、Recessive隐性、Allele等位基因。

卡方检验结果示例：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SNP | rs4886761 | rs4886761 | rs4886761 | rs4886761 |
| CHR | 15 | 15 | 15 | 15 |
| Ref allele | C | C | C | C |
| Alt allele | T | T | T | T |
| Model | Codominant | Dominant | Recessive | Allele |
| AFF(11|10|00) | 48/118/76 | 166/76 | 48/194 | 214/270 |
| UNAFF(11|10|00) | 23/118/169 | 141/169 | 23/287 | 164/456 |
| ChiScore | 36.28 | 29.41 | 18.69 | 38.09 |
| OR(95%CI) |  |  |  | 2.204(1.711-2.838) |
| P-value | 0.00000001325 | 0.00000005866 | 0.00001537 | 0.0000000006752 |
| FDR\_BH adjusted |  |  |  | 0.000000003376 |

Model：分析模型

AFF(11|10|00)：case样本中纯合突变、杂合突变、纯合正常数量统计

UNAFF(11|10|00)：control样本中纯合突变、杂合突变、纯合正常数量统计

ChiScore：卡方值

P-value：卡方分析P值

详细结果见ChiScore.xlsx

#### 3.2 SNP位点逻辑回归分析

对所有SNP位点进行逻辑回归分析，分析模型有Dominant (频率低的allele为显性)，Recessive(频率低的allele为隐形)，Log-additive (叠加作用)，HOM/HET（共显性模型，以纯合正常为参照）进行关联分析。

逻辑回归结果示例：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SNP | rs4886761 | rs4886761 | rs4886761 | rs4886761 | rs4886761 |
| CHR | 15 | 15 | 15 | 15 | 15 |
| BP | 74215551 | 74215551 | 74215551 | 74215551 | 74215551 |
| Alt Allele | T | T | T | T | T |
| Model | Dominant | Recessive | Additive | HOM | HET |
| NMISS | 552 | 552 | 552 | 552 | 552 |
| OR | 2.618 | 3.087 | 2.175 | 4.641 | 2.224 |
| SE | 0.1794 | 0.2701 | 0.1323 | 0.2888 | 0.1898 |
| L95 | 1.842 | 1.818 | 1.678 | 2.635 | 1.533 |
| U95 | 3.721 | 5.242 | 2.819 | 8.173 | 3.226 |
| STAT | 5.364 | 4.174 | 5.871 | 5.315 | 4.211 |
| P-value | 0.00000008143 | 0.00002995 | 0.000000004334 | 0.0000001065 | 0.00002547 |
| FDR\_BH adjusted | 0.0000002443 | 0.00008985 | 0.000000007801 | 0.0000003195 | 0.0000003195 |

NMISS:本分析模型下，该位点未缺失数据量

OR：逻辑回归OR值

L95,U95：逻辑回归95%置信区间

SE：标准差

STAT：系数的t统计量

FDR\_BH\_adjusted:FDR校正结果

详细结果见Logistic.xlsx

#### 3.3 环境因子校正的逻辑回归分析

使用性别、年龄用于逻辑回归校正。

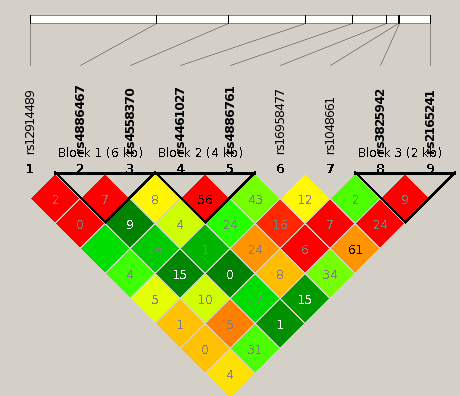
逻辑回归校正结果示例：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SNP | rs4886761 | rs4886761 | rs4886761 | rs4886761 | rs4886761 |
| CHR | 15 | 15 | 15 | 15 | 15 |
| BP | 74215551 | 74215551 | 74215551 | 74215551 | 74215551 |
| Alt Allele | T | T | T | T | T |
| Model | Dominant | Recessive | Additive | HOM | HET |
| NMISS | 513 | 513 | 513 | 513 | 513 |
| OR | 2.764 | 3.756 | 2.337 | 5.655 | 2.241 |
| SE | 0.2017 | 0.3022 | 0.1482 | 0.3235 | 0.2139 |
| L95 | 1.861 | 2.077 | 1.748 | 3 | 1.473 |
| U95 | 4.104 | 6.791 | 3.125 | 10.66 | 3.408 |
| STAT | 5.04 | 4.379 | 5.729 | 5.355 | 3.772 |
| P-value | 0.0000004648 | 0.0000119 | 0.00000001011 | 0.00000008549 | 0.000162 |
| FDR\_BH adjusted | 0.000001394 | 0.00003569 | 0.00000002274 | 0.0000002565 | 0.0000002565 |

详细结果见Logistic\_CORRECT.xlsx

## 四、连锁不平衡分析

对每个基因内的位点使用haploviev软件进行连锁不平衡分析，从而获得存在较强关联性的单倍型。分析结果如下图：



LOXL1基因的Haploview分析Block图

其他基因分析结果与原始数据见文件夹haploview。

## 五、单倍型分析

通过章节四中的连锁不平衡分析，找出存在较强关联性的Block，对这些Block进行单倍型分析并使用case/control信息进行逻辑回归的关联分析。

LOXL1基因单倍型关联分析分析结果（在我们的分析里，以需要分析的单倍型之外的所有单倍型作为参照）

|  |  |  |
| --- | --- | --- |
| Hap | LOXL1-1 | LOXL1-1 |
| CHR | 15 | 15 |
| SNPS | rs12914489;rs4886467;rs4558370 | rs12914489;rs4886467;rs4558370 |
| HAPLOTYPE | GGG | AGG |
| case\_F | 212(0.438) | 24(0.050) |
| control\_F | 322(0.519) | 28(0.045) |
| OR | 0.7213 | 1.1031 |
| 95%CI | 0.5680-0.9160 | 0.6309-1.9286 |
| P-value | 0.0074 | 0.7306 |

Hap：单倍型名称

SNPS：构造单倍型的SNP位点

HAPLOTYPE：单倍型

case\_F：case样本单倍型的数量与频率

control\_F：control样本单倍型的数量与频率

其他单倍型与详细结果见haplotype.xlsx

#### 5.2 环境因子校正的单倍型分析

使用性别、年龄用于逻辑回归校正。

|  |  |  |
| --- | --- | --- |
| Hap | LOXL1-1 | LOXL1-1 |
| CHR | 15 | 15 |
| SNPS | rs12914489;rs4886467;rs4558370 | rs12914489;rs4886467;rs4558370 |
| HAPLOTYPE | GGG | AGG |
| case\_F | 186(0.431) | 22(0.051) |
| control\_F | 307(0.517) | 27(0.045) |
| OR | 0.7612 | 0.9952 |
| 95%CI | 0.5811-0.9971 | 0.5320-1.8617 |
| P-value | 0.0476 | 0.9879 |

Hap：单倍型名称

SNPS：构造单倍型的SNP位点

HAPLOTYPE：单倍型

case\_F：case样本单倍型的数量与频率

control\_F：control样本单倍型的数量与频率

其他单倍型与详细结果见haplotype\_correction.xlsx

## 六、主要软件方法以及参考文献

(1)关联分析：本分析流程主要基于plink分析软件实现。http://pngu.mgh.harvard.edu/~purcell/plink/index.shtml

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007).PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

（2）连锁不平衡分析：本流程使用软件Haploview实现