参赛密码 

**（由组委会填写）**

****

**** 

**“华为杯”第十三届全国研究生**

**数学建模竞赛**

|  |  |
| --- | --- |
| **学 校** | **宁波大学** |
| **参赛队号** | **11646034** |
| **队员姓名** | **1.石翔夫** |
| **2.陈凯迪** |
| **3.董伟** |

参赛密码 

**（由组委会填写）**

题 目 具有遗传性疾病和性状的遗传位点分析

摘 要

全基因组关联分析（GWAS）是指在人类全基因组范围内找出存在的序列变异，即单核苷酸多态性（SNP），从而确定致病位点（SNPs）或者致病基因的位置。

对于问题一，要求对位点进行数值编码。位点信息都是由A、T、C、G四种碱基组成，不同位点间相同的碱基并不存在关联性。综合考虑多种编码方案后，我们采用二进制编码方式，分别统计位点内两种碱基的频率，按频率从高到低得到位点二进制编码方式分别为：00、01、11，并且进一步将两位的二进制编码数据转化为一位的十进制编码数据，最终得到单个位点的三种编码方式按碱基频率从高到低分别为：0、1、2。最后分别对数值编码前和数值编码后的位点使用F检验计算的P值和卡方计算的P值比较，得到两种方法计算得到的9445个位点P值排行相似度达到97.7%，并且分析了使用全部十种位点类型作为编码方式的不可行分析和过多加入人为经验对数据编码的影响。

对于问题二，分别从统计学和机器学习两种角度进行建立模型。模型一从统计学角度出发，逐个计算位点与性状的P值和OR值，综合考虑分析P值和OR值相关性，得到实验结果：当P-values<0.05时，共有430个位点；当P-values<0.01时，共有84个位点。并在P值的基础上，再对OR值进行分析。模型二选择随机森林（Random Forests）作为模型，使用第一题经数值编码后的样本进行监督训练，交叉验证后得到模型预测准确度为68%的模型，根据基尼指数，量化位点对随机森林算法的贡献度，得到位点重要性排行。最终，结合实际，考虑致病位点个数为一个或多个，综合比较模型一和模型二得到位点，发现相似度达80%。其中位点rs2273298各指标极其明显，其次是rs932372，rs12036216。

针对问题三，基因的引入需要进行多元自变量分析，并产生了冗余自变量和连锁不平衡的问题。采用统计学方法，首先使用主成分分析（PCA）完成基因内的降维，排除与性状无关的位点，并消除连锁不平衡的影响（通过排除线性相关的位点），再利用多元线性回归模型对主成分进行性状关联分析。本题使用工具软件MAGMA进行计算，MAGMA内置模型为多元主成分线性回归模型，完美契合我们的模型需求。通过计算获得基因P值与Z值（与P值正相关），比较得到指标最显著的是基因102。另外，按照统计学常用的0.05阈值还可以接受另外13个基因：217，128，169，235，121，76，113，55，280，85，17，30，260。

针对问题四，根据题意将10个性状看做整体，先对性状集合做处理。参考两篇文献内容，采用多角度综合分析。从主性状角度，找到性状10在集合中起代表作用；从性状组合角度，发现性状10与除5,6外的性状都有密切的相关性，即可以针对两个性状分类分别找出相关位点，从而得到的位点集合就是性状整体的相关位点；从位点交集角度，对每个性状单独进行位点相关分析，根据P值排序取TPO-50，再求主性状10与其他性状的TPO-50位点的交集，引入评分机制，最终得到优势显著的两个位点rs12746773与rs351617。

**关键词：GWAS，SNP，随机森林，主成分分析，多元线性回归**

1. 问题重述

**1.1问题背景**

人体的每条染色体携带一个DNA分子，DNA由分别带有A、T、C、G四种碱基的双螺旋长链分子组成。基因是DNA长链中的一些片段。SNP位点是组成DNA的大量碱基对中，那些特定位置上发生变异引起DNA多态性的单个核苷酸。因为SNP在DNA中频繁出现，并且多态性丰富，所以SNP是重要的遗传标志之一。

近些年，随着人类全基因组草图的完成和人类全基因组序列测定的完成，研究人员使用人类全基因组方法[[[1]](#endnote-0)]来确定致病位点或者致病基因，极大地促进了研究人员对于人类个体基因序列中包含的遗传信息的研究。具体做法：招募大量志愿者（样本），其中包含某种遗传病的人和健康的人，通常用1表示病人，0表示健康者。对每个样本，采用碱基(A,T,C,G)的编码方式来获取每个位点的信息(因为染色体具有双螺旋结构，所以用两个碱基的组合表示一个位点的信息），且每个位点的编码方式只有三种，两个纯合因子，一个杂合因子，形如TT、TC、CC。其他点位虽然碱基不同，但是都是由两种碱基组成的三种不同编码。

**1.2问题提出**

本题目给出某种遗传疾病（简称为疾病A）的1000个样本信息，每个样本含有9445个位点信息，以及包含这些位点的基因信息。根据题目给出的数据和文件，设计相关模型或者算法，设计最合适的碱基编码方式，计算寻找9445个SNP中的最有可能的致病SNP和基因，以及疾病相关性状的分析。其中包含如下文件：

phenotype.txt文件包含1000个样本的患病情况，0表示没有患病，1表示患病，由0和1组成1000行，500行0表示500个没有患病A的人，500行1表示500个患病A的人。

Genotype.dat文件由1001行和9445列组成，第一行为表头，为9445个位点的名称，接下来每行表示一个样本，每列表示一个位点信息，组成一个1000\*9445的矩阵。

文件夹gene\_info包含300个dat文件，表示300个基因信息，每个dat文件包含若干个位点名称，说明这个基因由这些位点组成。

Multi\_phenos.txt文件提供了上述1000个样本的10种相关性状的信息，为一个1000\*10的矩阵，每列表示一个性状，每一行表示一个样本，其中用0表示没有患病，用1表示患病。

根据题目给出的数据和文件，设计相关模型或者算法，计算寻找9445个SNP中的最有可能的致病SNP和基因，以及疾病相关性状的分析。

**1.3所需解决的问题**

问题一、用适当的方法，把genotype.dat中每个位点的碱基（A、T、C、G）编码方式转化成数值编码方法，便于进行数据分析。

问题二、根据genotype.dat和phenotype.txt文件，设计或采用一个方法，找出某种疾病最有可能的一个或几个致病位点。

问题三、根据phenotype.txt文件和gene\_info中的300个dat文件，找出最有可能致病的一个或者几个基因。

问题四、考虑疾病相关性状，根据multi\_phenos.txt文件，找出和multi\_phenos.txt中十个性状相关联的位点。

二、问题分析

本文研究的主要对象是位点，位点是人类基因组中最广泛的多态性现象，也是造成个体间差异的最主要因素。所以位点的研究在当前的人类基因组研究中备受关注。位点作为重要的遗传标志之一，位点的研究极大的促进了人类基因组序列中包含遗传信息的研究以及人类各个个体表现特征相关基因组序列片段识别的研究。越来越多的研究人员也开始在位点序列数据中挖掘基因信息。具体问题分析如下

**2.1 问题一的分析**

本问中要求选择合适编码方式将每个位点的碱基（A,T,C,G）编码方式转换数值编码方式，以便于进行数据分析和处理。从替换碱基原始数据的数据属性的角度着手，显然需要让数值化后的编码尽可能不丢失原本的信息且便于数据处理。换而言之，数值化的SNP编码要与碱基编码获得相同的结果。可以考虑以统计指标——P值，作为比较指标，通过计算各种编码方式获得的所有SNP的P值序列与碱基编码结果的差异，来衡量编码方式的效果。

**2.2 问题二的分析**

问题二要求用genotype.dat中的SNP位点样本和phenotype.txt中的患病情况进行分析，找出最有可能的一个或者几个致病位点。这是背景专业中极其常见的问题，所以可以采用统计学方法分析位点与患病情况的相关性，找出最有可能致病位点。同时结合第一问中的编码方式，将genotype.dat转化为数值矩阵，使用机器学习的方法，利用样本训练模型，通过达到一定预测能力后，提取模型中的参数分布，从而获得位点的影响因素排序，检验统计学方法[[[2]](#endnote-1)]的筛选结果。

**2.3 问题三的分析**

问题三要求结合gene\_info中的300个基因dat文件，从基因层面分析其与疾病的相关性。这里题目给出的暗示是可以把基因理解为若干个位点组成的集合，遗传疾病与基因的关联性可以由基因中包含的位点的全集或其子集合表现出来。这表示问题二得到的高相关度SNP位点会与最终的基因选择结果存在较高的一致性。根据对数据文件的分析，我们发现300个基因囊括了全部9445个SNP位点，且基因及其中的位点都是连续。因此本题所需建立的模型，首先需要考虑剪枝消除基因中相关性低或者无关的SNP的影响。其次模型要能够建立多元自变量与单一因变量之间的关系。

**2.4 问题四的分析**

问题四是在问题二的基础上，给出因变量一方的干扰。10个性状使问题变得复杂，前面问题适用的一些模型会因为，因变量变成了10维数组而无法直接使用。本题给出的信息是将相关疾病或性状看作一个整体。由于目标是找出一个或几个关联位点。可以从两个方向入手，一是对10维的输出进行适当得降维，再用原来的模型找出与低维度的输出呈高相关性的位点。二是对性状进行聚类，分别找出不同性状类的最相关位点，则这些位点的组合就是10个性状的最相关位点。

三、基本假设

1、不考虑样本中人与人之间的差异。

2、不考虑样本中性别差异。

3、不考虑基因与SNP位点在染色体上的位置信息。

4、不考虑样本收集过程中产生的数据噪声。

5、假设样本选择随机，不存在样本特异性。

四、符号说明

|  |  |  |
| --- | --- | --- |
| 序号 | 符号 | 符号说明 |
| 1 |  | 等位基因中低频碱基 |
| 2 |  | 等位基因中高频碱基 |
| 3 |  | i类样本第j种位点的数量和 |
| 4 |  | i类碱基在j类样本中的数量和 |
| 5 |  | 样本中某类碱基的频率 |
| 6 |  | i类碱基在j类样本中的期望频率 |
| 7 |  | 样本中某类碱基的期望频数 |
| 4 | O | 即Observe值，表示样本的观察值 |
| 7 |  | 熵 |
| 8 |  | 基尼指数（Gini Index） |
| 9 | t | 随机森林中的节点 |
| 10 |  | 第i个类别在整个类别种办法的比例 |
| 11 | s | 节点分裂点 |
| 12 |  | 孩子节点中样本数量各自的比例 |
| 13 | S | 训练集 |
| 14 | F | 特征维度 |
| 15 | n\_estimators | 树的数量 |
| 16 | criterion | 分割标准 |
| 17 | max\_features | 分割时选择的特征数 |
| 18 | rank | 子模型排行函数 |
| 19 |  | log求和函数 |

五、模型建立与求解

**5.1问题一的模型与求解**

为了便于数据分析，直接使用位点碱基（A，T，C，G）字母编码方式显然不太合适，所以需要寻找一种合适的数值编码方式，使得在对位点分析能更加便利，且不丢失位点原有的信息。我们从多个角度分析，经过检验选择了使用基于单碱基频率的0，1，2编码。

首先考虑不同位点的碱基对之间的关系，即假如位点x是由A和T碱基组成，而位点y也是由A和T碱基组成，那么由基因学知识可以知道，虽然同一位点的AT，AA，TT碱基对三者之间会存在相关性，但是不同位点间的相同碱基编码不存在直接相关性，所以直接不考虑全部位点可能的出现形式，即AA，GG，TT，CC，AG，GT，AC，AT，GC，CT这十种碱基对情况。而后，所以我们只考虑单个位点内的三种不同编码方式，一开始，我们将患病和未患病分开进行统计，将患病样本中出现频率最高的编码作为2，再在未患病中寻找最高频出现的编码作为1，若该编码和患病样本一样，则退而求其次，选择次高频率的编码作为1，剩余的为0，但是经过检验发现该种编码方式由于过度的加入人为先验知识，使得检验结果并不明显。

最终，我们选择出了一种能最高还原原始数据并且还能方便数据分析的编码方式，即将单个碱基位点，拆成两个碱基考虑，考虑到低频碱基为致病主要因素的概率很大，为检测时或者数据处理时，能突出低频信息，提高区分度，给予低频较大的权重，将单个碱基对拆分成两个碱基，对单个碱基对内的碱基出现频率进行统计，再以这个频率大小进行区分，使用二进制编码对碱基进行数值编码。低频碱基定义为1，高频碱基定义为0



其中表示频率较低的碱基，定义为1；表示同一碱基对中，频率较高的碱基，定义为0。然后对同一碱基对中的，，分别编码为二进制01，11，00，其中杂合子碱基编码固定，定义为01，考虑到为数值计算，进一步将二进制编码转化为十进制编码，最终确定编码如下：

=0，=1，=2

先使用基因分析专业软件PLINK计算碱基对编码的P-values值，再对编码进行数字化，然后使用python语言的机器学习工具SKLEARN中的features库的P-values函数计算数值型编码的P-values值。按照筛选条件P-values值<0.05，从PLINK软件的结果中可以筛选出440个位点，而在机器学习工具SKLEARN结果中则能筛选出的430个位点，两者获得的位点集合的相似度为97.72%，而且这些差异是由编码的方式之间的微小差异造成。成功说明了所采用的的数值编码方式的有效性，具体P-values介绍与应用将在问题二中详细阐述。

下表是PLINK在碱基（A,T，C，G）编码方式(简称字母编码)上计算的位点P-values值和数值编码方式（简称数值编码）计算的位点P-values值的排序比较：

表 1 用数值编码和字母编码分别计算位点的P-values值排行



从表1中可以发现，用数值编码计算的P-values值和字母编码计算的P-values值之间存在排行虽然存在差异，但这是由于P-values值差异极小造成的，对采用数值编码得到的结果没有影响，可以发现，选择出的位点全部相同，证明采用数值编码方法的可靠性。

**5.2问题二的模型与求解**

问题二的目标是利用1000个样本，从9445个位点中找出与某种疾病最有可能的相关的一个或几个位点。由于自变量远远大于样本数，因此决定采取两种不同的方法来寻找目标位点。方法一是采用统计学方法，利用统计指标，P-values值与OR值，其为生物学界最常用的分析方法。方法二是采用机器学习方法，通过训练随机森林，给位点进行逐个评分，由于在随机森林中，评分表示位点对最终决策的影响重要程度，所以可以通过评分排序找出目标位点。

**5.2.1统计计算方法**

P-values(以下简称P)值是统计学中的常用指标，它的含义是当原假设为真时所得到的样本观察结果或更极端结果出现的概率。如果P值很小，说明原假设情况的发生的概率很小，而如果出现了，根据小概率原理，我们就有理由拒绝原假设，P值越小，我们拒绝原假设的理由越充分。总之，P值越小，表明结果越显著。但是检验的结果究竟是“显著的”、“中度显著的”还是“高度显著的”需要根据P值的大小和实际问题来解决。

在全基因组关联分析（GWAS）中，P值的计算需要将碱基对等位基因分开，首先进行卡方检验。卡方（）检验，就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，卡方值越大，越不符合；卡方值越小，偏差越小，越趋于符合；若两个值完全相等时，卡方值就为0，表明理论值完全符合。其具体计算过程如下所示：

表 2 统计表现型与位点分型的频数情况

|  |  |  |  |
| --- | --- | --- | --- |
| 表现型 | AA | AB | BB |
| 患病 |  |  |  |
| 未患病 |  |  |  |

表 3 统计观察样本的频数分布

|  |  |  |  |
| --- | --- | --- | --- |
|  | 患病 | 未患病 | 频数 |
| A |  |  |  |
| B |  |  |  |
| 频数 |  |  |  |

表 4 统计观察样本的频率分布

|  |  |  |  |
| --- | --- | --- | --- |
|  | 患病 | 未患病 | 概率总和 |
| A |  |  |  |
| B |  |  |  |
| 概率总和 |  |  | 100% |

表5 计算样本的期望概率

|  |  |  |  |
| --- | --- | --- | --- |
|  | 患病 | 未患病 | 概率总和 |
| A |  |  |  |
| B |  |  |  |
| 概率总和 |  |  | 100% |

表 6 计算样本的期望频数

|  |  |  |  |
| --- | --- | --- | --- |
|  | 患病 | 未患病 | 频数和 |
| A |  |  |  |
| B |  |  |  |
| 频数和 |  |  |  |

利用上述四表运算得到样本的期望频数E，再利用观察到的样本频数O，由此可以计算卡方检验值。



利用卡方值我们可以得到极具统计意义的P值，P值代表假设的可信程度。通常认为P值小于0.05时，拒绝极端假设，即可接受变量间的相关性。

OR值（Odds Ratio）又称比值比、优势比，是在对照实验中常用的统计值。在本题中，由于位点仅存在两种碱基和三种表达方式，故OR值表示的是碱基因素对疾病作用程度。



通常认为，OR值等于1为因素不起作用，OR值大于1为危险因素，OR值小于1为保护因素。在本题中，由于SNP位点同时具有两种碱基，故当一种碱基为保护因素时，另一种必为危险因素，所以OR值对1的偏离越大，表明作用越明显。方便起见，我们在对OR值进行了处理，对小于1的OR值取其倒数以方便排序。

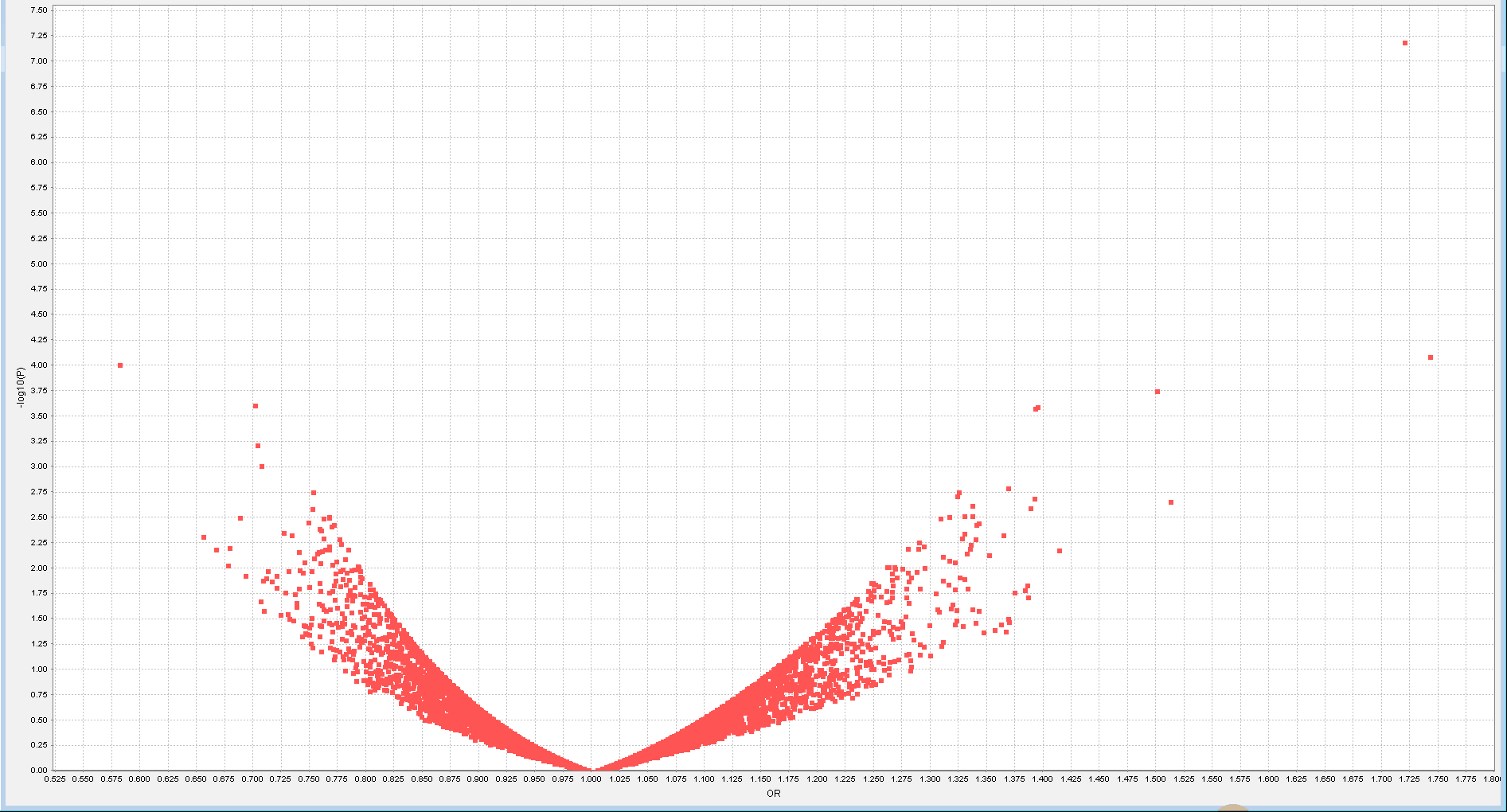


图 1 PLINK的OR值散点图

表 7 利用EXCEl计算的P值

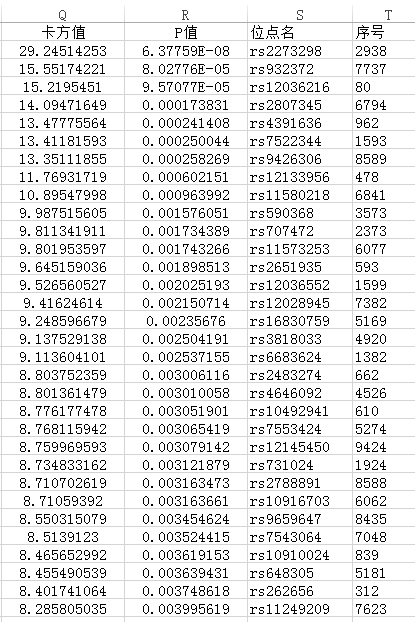
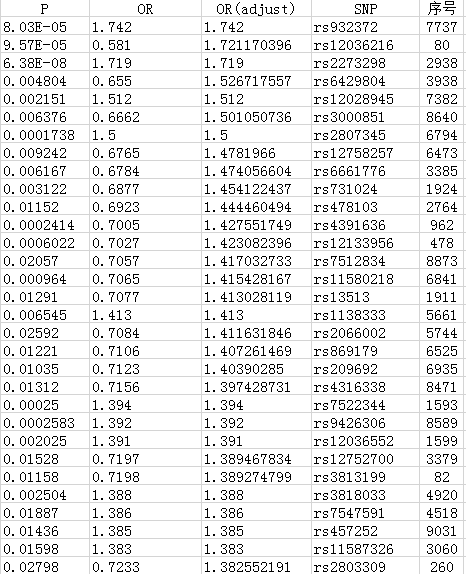


表 8 利用基因分析软件PLINK检验P值并计算OR值



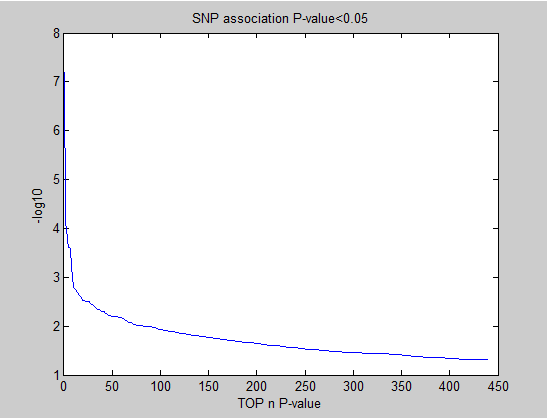


图 2 P值<0.05的排序（440个位点）

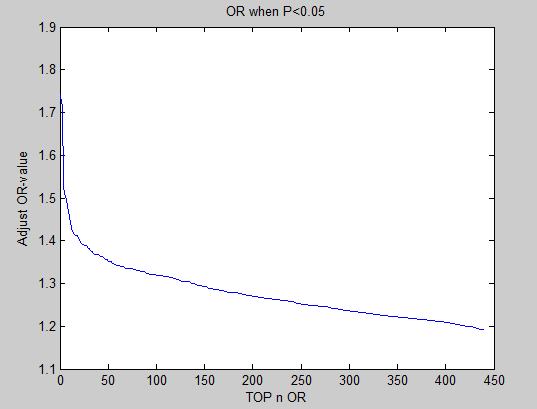


图 3 OR值排序（当P值<0.05）

**5.2.2随机森林（RF）模型**

RF[[[3]](#endnote-2)]是一个由多棵决策树对样本进行训练和预测的一种组合分类器。简单的说，RF就是由多个CART（Classification And Regression Tree）构成的。对于每棵树，总是从有放回的全部样本中采样，所以样本可能会多次出现在一颗树的训练集中。而RF用于分析位点的优势有如下优势：

（1）可以计算出每个位点的重要性评分；

（2）对于以位点作为训练特征时，其对相加、显性、隐性或者共线性遗传模型没有要求；

（3）和传统的机器学习方法，如回归分析方法相比，随机森林由于利用了bootstrap抽样，考虑袋外数据的预测，不容易产生过拟合（over-fitting）问题。

决策树形成的关键是分类每个节点的选择都是最佳分裂属性，根据这个属性，样本确定最佳的划分结果。两种常用的不纯度的度量包括熵和基尼指数（Gini Index）,计算公式分别为：





其中，t表示当前节点，l为样本类别个数，表示每个类别在整个类别中的比例。在每一颗决策树中，每个分裂点都是以不纯度下降来进行选择，计算方法如下：



其中t依旧表示当前节点，s表示当前节点的一个分裂值，和表示左右两个孩子节点中样本数量各自的比例。

综合考虑后我们使用Gini指数来评价训练完成后的特征最训练结果的贡献程度。训练过程：

(1)从训练集有放回采样（Bootstrap方法）。

(2)利用获取的采样样本，进行单棵决策树训练，利用Gini指数选择最佳属性进行划分节点。

(3)重复采样，建树。

最终代码实现时，考虑到位点信息量比较大，合理选择语言，使用C/C++这种较为底层的语言做位点数据处理，读取时间快，缩短程序运行的时间，而python运行环境是搭建在C/C++上面，但是其有着丰富的开源开发包和开源库，加快了的建立随机森林模型的速度，节省了大量的时间。

为验证我们训练的模型是否可行，采用了交叉验证的方法，将1000个样本进行划分，700个用于训练，300个用于验证模型的准确程度，经过一定的参数调整，模型的预测精度可以达到68%以上。最终根据训练好的随机森林模型，根据Gini指数对特征进行重要性评估，并且计算出全部特征的特征重要程度。

随机森林参数选择：

训练集S：1000（全部样本）

特征维度F：9445(全部位点)

树的数量n\_estimators：1000

分割标准criterion：gini

分割时选择的特征数max\_features：97

最后，我们将通过使用P值、OR值选出的位点和由随机森林选出的位点比较分析，发现两者相似度较高,考虑到实际情况中，致病位点往往只有一个或者几个，所以直接选择了各个方法重要性排列后的前20个，进行比较分析。其中位点rs2273298特别明显。使用各个方法均有明显的排名优势，其中P值达到了，RF重要性评价为0.00148767475755，排名第二的位点rs932372，RF重要性评价为0.000654733645671，两者相差两倍。为确定我们数据的正确性，我们还使用了PLINK软件进行验证，这更加确定了我们的结果。

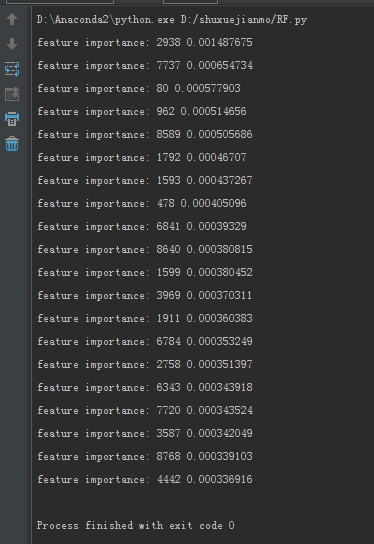


图4 RF特征重要性表

综合考虑三种方法的结果，我们选出各个方法排名均靠前的三个位点，其中显著性极高的rs2273298，还有显著性也相对明显的rs932372和rs12036216。

**5.3问题三的模型与求解**

问题三与问题二的区别在于引入了基因的概念。基因可以认为是SNP的合集，这表示问题从简单的一元回归变成了多元回归问题，同时，基因分析在SNP自变量的增加之外，它还引入了基因内部的关系考虑。

在全基因组关联分析中，基因和基因组分析被认为是典型的单SNP分析的强力替代。在基因分析中，基因标记数据在全基因层面上整合，考察了基因上所有标记的联合关联。同样的，在基因组分析中，个体基因通过相近的生物、功能或其他特征聚集在基因组中。这样做的好处就是既减少了所需的测试次数，又能检测到多个弱关联引起效应，而这些弱关联往往在单标记检测中容易被忽略。

针对上述特征，我们选择主成分多元线性回归模型来解决问题。

多元线性主成分回归模型使用F检验来计算基因的P值，这个模型首先把基因的SNP位点矩阵进行主成分投影（Principal Components），即裁剪掉那些特征值极小的成分，然后则利用这些主成分作为表现型的线性回归模型中的自变量。通过移除冗余参数保证了模型对于高共线性的SNP位点具有很强的识别能力。

用表示主成分，Y表示表现型，W表示协变量的可选矩阵，则该模型可以写作：



其中E是单位向量，表示基因效果的参数向量，是可选协变量的效果，是截距，是残差向量。F检验采用零假设，即是零向量，表示基因g对表现型Y不存在影响。

F检验[[[4]](#endnote-3)]，又称方差齐性检验。从研究总体中随机抽取两个样本，要对这两个样本进行比较的时候，首先要判断两总体方差是否相同，即方差齐性。F检验法是英国统计学家Fisher提出的，主要通过比较两组数据的方差 S^2，以确定他们的精密度是否有显著性差异。至于两组数据之间是否存在系统误差，则在进行F检验并确定它们的精密度没有显著性差异之后，再进行t检验。



计算出F值之后，再根据两个样本各自的自由度查阅F表，如果F 值小于表中值，表明两组数据没有显著差异，相反则两组数据存在显著差异。

上述模型其实可以分为两个步骤，一是主成分分析，二是多元线性回归。这种设定是充分考虑了统计和实际相关的平衡。该多元回归模型确保了SNP位点之间的连锁不平衡被充分考虑到，同时也为调节额外协变量和交互作用项。

**5.3.1主成分分析**

主成分分析PCA，是数据处理常用的工具，通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维。降维不仅仅是减少计算量的问题，对于高维数据而言，以第三题为例，基因中存在从几个到几十个不等的SNP位点，根据第二题的结果我们可以很清楚地得知，SNP位点的致病影响力大小不一，差别显著，因此在基因层面的分析中，如若采用多元线性回归的模型，则冗余的无关位点会对分析结果造成极大影响，即妨碍有效模型的建立，干扰我们找到基因与致病之间的规律。另外，降维还能够消除多重共线性，即自变量之间的相互关联，从数学角度讲，多重共线性会导致解空间的不稳定，从而可能导致结果的不连贯。

多重共线性，在基因的场景中是着实存在，且需要考虑的。在生物学上，这个现象为连锁不平衡(Linkage Disequilibrium)[[[5]](#endnote-4)]，又称为等位基因关联，是指同一条染色体上，两个等位基因间的非随机相关，即，当位于同一条染色体的两个等位基因（A,B）同时存在的概率要大于人群中随机分布而同时出现的概率。如下图；

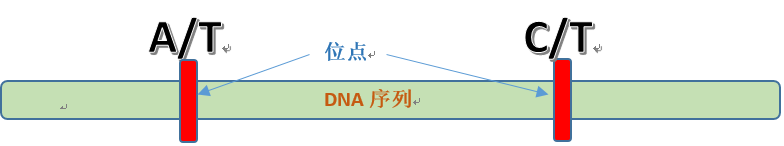


图5. 连锁不平衡图示

**A/T**

**C/T**

**DNA序列**

**位点**

**A/T**

**C/T**

**DNA序列**

**位点**

左边的位点可能出现AA/AT/TT，右边的位点可能出现CC/CT/TT，假设出现概率都是40%/20%/40%，则出现组合为AACC的概率P(AACC)的概率为16%，若实际观测到人群中的P(AACC)=40%，即显然高于两者独立的联合概率，那么就可以认为左右两个位点存在连锁关系，我们用LD表示两者的连锁不平衡关系：



研究发现，连锁不平衡由突变或重组形成，在染色体某一SNP位点附近有新的突变产生时，则连锁不平衡出现，若重组发生，则两位点间的连锁不平衡程度下降。理论上，连锁不平衡的强度与两个SNP位点间的距离有关，距离越小，发生重组机会越小，即连锁不平衡越强；而距离越大，发生重组的机会越大，即连锁不平衡越弱。不过，实际上也有距离很近而不存在连锁不平衡，但是距离相当远（超过100个碱基对距离），却存在连锁不平衡的情况。

连锁不平衡的度量一般不直接使用LD的定义式，而是对LD的值进行归一化，然后用连锁不平衡系数D­’­和r­­2进行检验。



D’是与频率无关的量。

D’=1，称为完全连锁不平衡，说明两个位点间没有发生重组；D’=0，称为连锁平衡。

r2则是与频率有关的量，这意味着在两位点间无重组时，r2也不一定达到最大值1。

r2=1，说明两位点无重组，亦称为完美LD，即观察一个标记就可以得到另一标记的全部信息。

r2=0，表示连锁平衡。

连锁不平衡的一种应用方式是基于SNP分析易患病性，其原理是利用已知的标记SNP，即该SNP已知与疾病易患性有关，则可以利用连锁不平衡找出连锁的其他位点，若这些连锁位点的观察样本在患病组明显高于对照群体，则可以认为这组SNP位点与疾病易患性有关。

不过，在本题背景下，连锁不平衡的应用方式截然不同。由于本题是要分析基因与疾病间的关系，即多个SNP位点与疾病之间的关系。回顾第二题，基于逐个SNP的分析，我们可以获得单个SNP位点与疾病之间的相关性，必然的，这些相关的SNP又会显著影响基因整体与疾病的相关性。考虑到连锁不平衡，需要注意连锁不平衡与距离有关，因此存在一对或多对疾病相关SNP位于同一基因，且碰巧它们呈现连锁不平衡性，假设不考虑冗余无关SNP位点的影响，这就意味着这一对或几对连锁SNP位点在回归分析或采用其他模型时，会提高该基因与疾病的相关性指标。由于我们采用多元线性回归模型去分析基因与疾病的关联，因此相关性指标对最终选择与评价起至关重要作用，即错误的指标必然导致错误的结论。

因此通过PCA主成分分析，消除自变量SNP位点之间的共线性，即消除了它们之间的连锁不平衡对后续模型分析的影响。主成分分析的做法如下：

1，特征中心化，即每一维的数据都减去该维的均值，得到矩阵B。

2，计算B的协方差矩阵C。

3，计算协方差矩阵C的特征值和特征向量。

4，选取大的特征值对应的特征向量，得到新的数据集。

**5.3.2多元线性回归模型**

多元线性回归模型是现实研究中常用的模型，其特点在于解决因变量的变化受到几个或多个重要因素的影响。当多个自变量与因变量之间是线性关系时，所进行的回归分析就是多元线性回归。本文使用的是多元线性主成分回归模型，其与前者不同之处在于多了主成分分析的步骤，该步骤主要用于解决自变量过多，特别是影响因子小的冗余自变量过多的问题。

设y是因变量，x1,，x2...是自变量，b是常数系数，则y可以表达为：



由于多元线性回归模型往往没有解析解，故多采用近似解法。通常采用最小二乘法来估计参数，采用EM算法来逐渐逼近数值解。EM算法，又称最大期望算法，是一种迭代算法，用于含有隐变量的概率参数模型的最大似然估计或极大后延概率估计。EM算法经过两个步骤交替进行计算[[[6]](#endnote-5)]：

第一步是计算期望（E），利用对隐藏变量的现有估计值，计算其最大[似然](http://baike.baidu.com/view/1055533.htm" \t "http://baike.baidu.com/_blank)估计值；

第二步是最大化（M），最大化在 E 步上求得的最大似然值来[计算参数](http://baike.baidu.com/view/862797.htm" \t "http://baike.baidu.com/_blank)的值。

M 步上找到的[参数估计](http://baike.baidu.com/view/123231.htm" \t "http://baike.baidu.com/_blank)值被用于下一个 E 步计算中，这个过程不断交替进行。

总体来说，EM的算法流程如下：

1.初始化分布参数

2.重复直到收敛：

E步骤：估计未知参数的期望值，给出当前的参数估计。

M步骤：重新估计分布参数，以使得数据的[似然](http://baike.baidu.com/view/1055533.htm" \t "http://baike.baidu.com/_blank)性最大，给出未知变量的期望估计。

**5.3.3工具介绍与结果分析：MAGMA**

MAGMA[[[7]](#endnote-6)]是一款专门用于基因和基因组的GWAS数据分析工具。它采用主成分多元线性回归模型，恰好符合我们的建模需要。MAGMA的基础分析包含二到三个步骤：首先生成一个注释，记录SNP与基因的映射。然后基于基因层面，计算基因的P值。最后，根据需求，进行基因组或者基因性质的分析。

MAGMA运算结果词汇解释：

START、STOP：基因所包含的SNP位点的起始标号和结束编号。

NSNPS：基因包含的SNP位点数量。

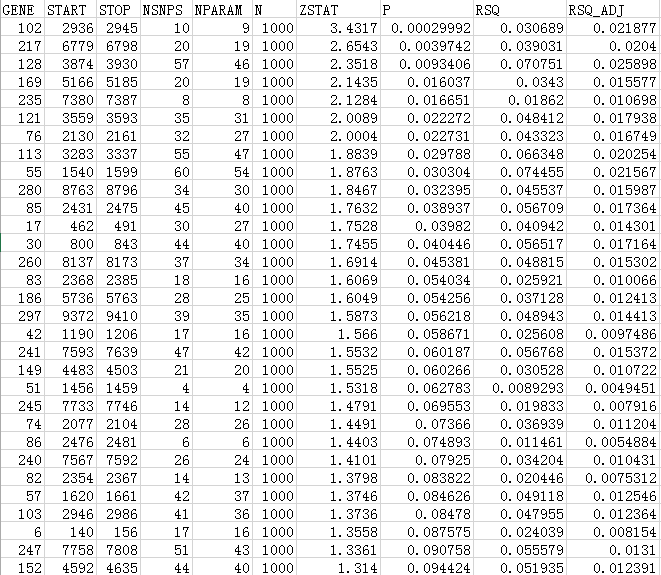
NPARAM：模型使用的相关参数数量。

N：样本数量。

ZSTAT：Z值，与P值相关。用于衡量基因的相关性。在数学上，表示当已知标准差时，验证一组数的均值是否与某一期望值相等时。

RSQ/RSQ\_ADJ:R2值与调整后的R2值，与主成分分析中的协变量有关。

表9，利用MAGMA进行基因分析的结果



**5.4问题四的模型与求解**

根据题设，“科研人员往往把相关的性状或疾病看成一个整体”，“找出与multi\_phenos.txt中10个性状有关联的位点”,可以将10个性状作为一个整体输出Y，去求9445个位点中与Y相关性最强的一个或几个位点。

根据文献查找结果，我们发现多相关性状与基因的关联分析在学术界也属于比较前沿，没有固定解决方式的问题。因此，我们决定参考两篇文章用三种方法综合分析点位与10个性状整体的相关性。

**5.4.1主基因方法**

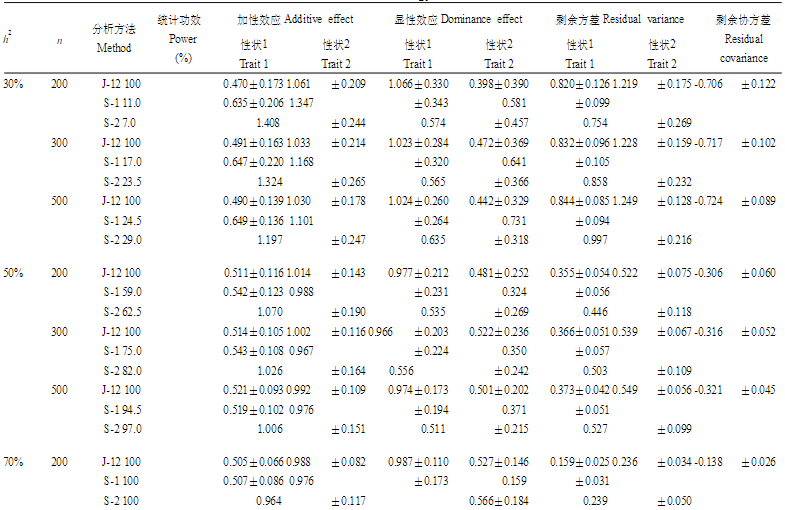
《多个相关数量性状主基因的联合分析方法》[[[8]](#endnote-7)]中提出了主基因的概念。该文献认为由于单性状的相关基因候选数太多，且有些基因的影响力不强，所以单性状发现基因的能力有限，所以提出了利用多个相关性状来发现主基因的方法。其隐含假设就是，主基因对主性状产生超显性作用，而对其他性状产生弱显性作用。文献设计了两种方案：

方案一是某一主基因同时控制两个性状的表达。设该主基因对性状1的作用表现为超显性，其加性效应和显性效应分别为0.5和1；该主基因对性状2的作用表现为部分显性，其加性效应和显性效应分别为1和0.5。

方案二是某一主基因仅控制性状2的表达。方案参数设置同方案一，其区别为该主基因对性状1无作用，即此时性状1的表现型变异完全由剩余变异造成。

该实验表明，采用主性状联合分析方法发现相关基因的能力要远大于采用单性状分析。部分结果如下表10所示，其中分析方法一栏，J表示性状1和形状2联合分析，S-1,S-2分别表示单性状分析。

表10，《多个相关数量性状主基因的联合分析方法》文献实验结果



根据上文的发现，我们结合本题题设中将10个性状整体考虑的暗示，提出10个性状中存在某一个性状为主性状，主性状受主SNP位点的超显性作用，而其他9个性状亦受到部分显性作用。即，可以将10个性状的整体与SNP位点的相关性用某一主性状与SNP位点的相关性来替代。模型方法为用PCA的10个性状的样本进行主成分分析，得到如下特征值：

表11. 10个性状的PCA特征值结果。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 特征值 | 0.0529 | 0.0583 | 0.0606 | 0.0621 | 0.0689 | 0.0753 | 0.0780 | 0.0803 | 0.1083 | 1.8577 |

由上表，可得知第十个性状的特征值显著大于其他，故10个性状存在且仅存在一个主性状。之后，我们将利用PLINK分析SNP位点与性状10的相关性。

**5.4.2性状组合评价法**

固然题目中指出将10个性状整体考虑，我们不能排除是多个SNP位点组合作用导致，即10个性状分为几个聚类，分别对应某几个SNP位点。故我们先采用性状间两两组合进行相关性分析，然而结果显示，45种组合的相关性差异极不明显，所以难以发现协同性极强的组合，所以考虑结合其他角度的数据分析。

表12，10个性状两两间相关性

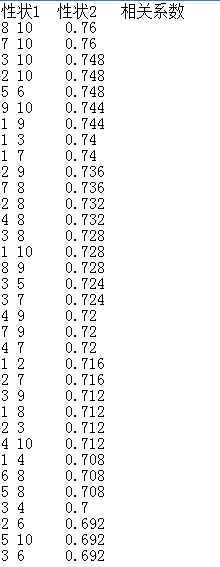
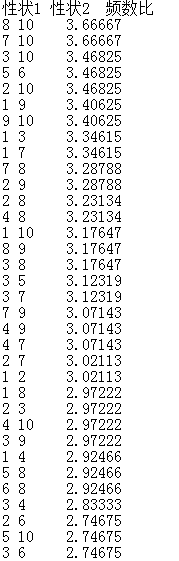
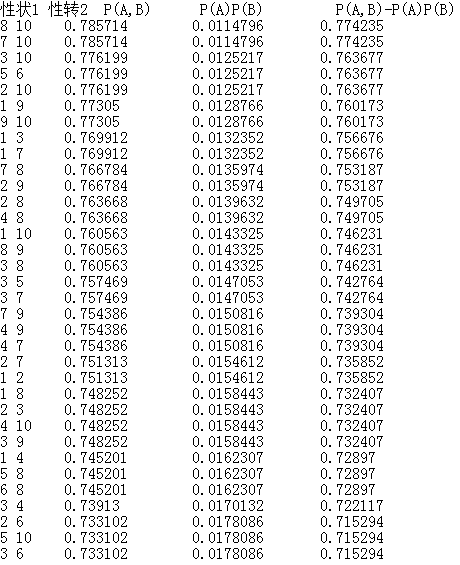


表13，性状组合为11的频数比与联合概率

受到第二题中，OR值计算的启发。我们选择对10个性状进行两两组合后，排除00的样本，计算11的频数与01,10的频数比，并计算了联合概率与概率积的差。如上表13所示。概率差表明10个基因之间两两的相关性都很显著，即两两之间都存在通常意义上的相关，毕竟由样本中大量的全1样本即可预料到这种情况。因此通过对排序结果的分析，可得知性状10与1,2,3,7,8,9的组合，频数比都在3以上，而4与10的频数比也极其接近3。而5,6的频数比极高。所以根据上述结果，可以将10个性状聚为两类，其中 5、6为一类，其他为一类。即至少存在两个SNP位点，它们的组合对10个性状的整体产生显著作用。

**5.4.3单性状相关位点交集法**

根据参考文献[[[9]](#endnote-8)]《绵羊全基因组CNV与部分性状的关联分析》，作者采用混合模型对采用混合线性模型对绵羊11 个性状表型的 CNP 关联结果进行分析，

“共 32 个 CNP 区域与各表型性状显著关联（P<0.05），发现其中 5 个 CNP 与初生重显著相关，2 个 CNP 与 125 天校正日龄重显著相关，9 个 CNP 与 190 天校正日龄重显著相关，6 个 CNP 与背膘厚度显著相关，3 个 CNP 与眼肌面积显著相关，3 个 CNP 与断奶前日增重显著相关，8 个 CNP 与断奶后日增重显著相关，10个 CNP 与平均日增重显著相关，10 个 CNP 区域与体高显著相关，5 个 CNP 与胸围显著相关，5个 CNP 与管围显著相关。”

文献最终得到结论：多数 CNP 区域影响多个性状，同时一些性状也受到多个 CNP 区域的影响，说明上述性状表型受到多微效基因的影响，并且各性状之间存在一定的遗传相关。

受此启发，我们想到大多数复杂疾病都会包含一系列高度关联的临床或者分子表现型，考虑到十种性状相关，从两个角度对本问题进行分析建模。模型一，各个性状虽然相关性强，但是也可以单独作为一种疾病进行分析。那么就可以将一个问题分解成十个子问题进行建模求解，使用PLINK计算所有SNP与个性状子集的P值：  
 

考虑到数据整体的相关度比较高和致病位点的个数一般在一个或者几个，我们采用TOP的方式进行对每个子集进行选择的数据，然后对获取的十个子模型再次建立Rank模型：



其中表示全部十个子模型中存在的位点，I表示总共的位点个数，表示位点在第k个子模型中的排行，然后对各个排行进行求和，考虑到我们要选择的是排名越靠前越好的位点，因此在rank前加入了log10()函数，利用log()函数随着值的增加梯度下降速度逐渐减慢，更加突出排名靠前的位点的重要性。另外，若某子模型TOP排序中不存在该SNP，则获得惩罚，惩罚值等于TOP的数量。最后再对处理过的排行值求和，并对求和后的CRank重新按数值从大到小进行排行。

由于本题的数据不同于真实世界，SNP位点与性状间的相关性普遍太强，使得P值小于0.05的统计学常规方法不能体现效果，因为在子模型中通常有400个左右位点其P值小于0.05。故我们采用TOP-50的方式进行筛选目标SNP位点。

结合前两种方法，即PCA与性状组合评比的结果，我们以性状10为主性状，对其P值TOP的SNP位点逐个计算CRank值，得到两个具有显著优势的位点rs12746773与rs351617。其中rs12746773在除了5,6,8之外的性状上都具有极高的P值排名，而在性状5的实际排名为708，性状6的实际排名为487。这与性状组合评比的结果完全相同。而rs351617位点则是TOP榜覆盖了最多性状的位点，其在唯一不受覆盖的形状2上的实际排名为67，也相当接近TOP50。

由上，我们认为SNP位点rs12746773与rs351617的组合，与10个性状的整体具有最强的相关性。

表14，性状的TOP-50个SNP位点的CRank值排序

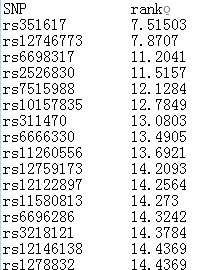


表15，单性状相关位点交集法结果1

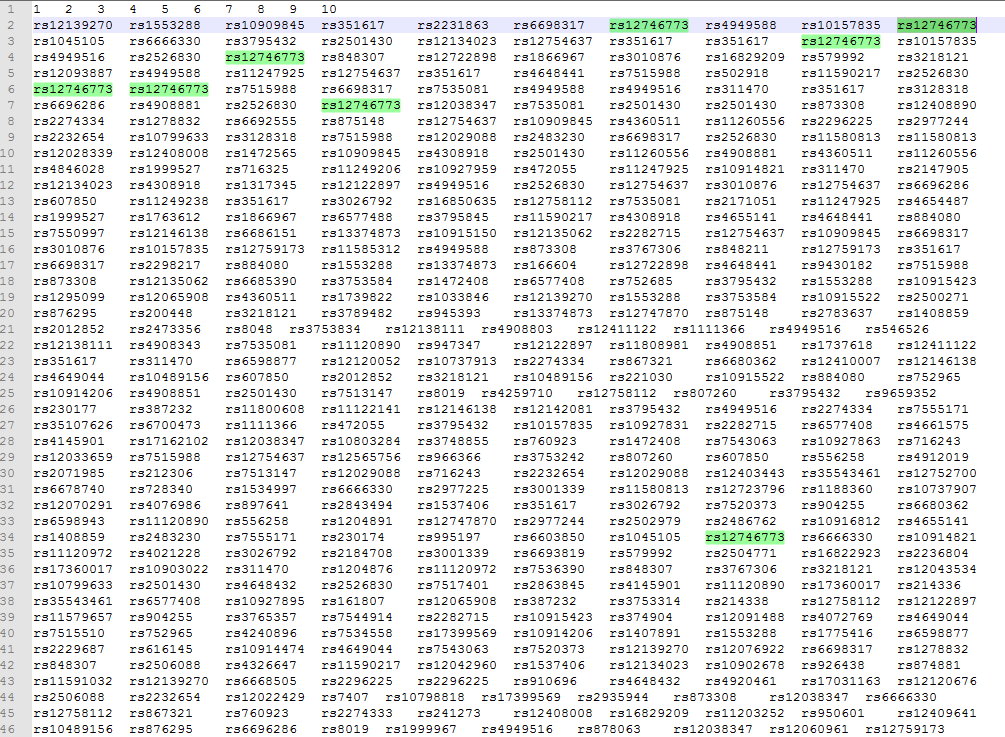


表16，单性状相关位点交集法结果2



六、模型的评价和改进

本次建模针对问题类型，我们在模型的选择上充分倾向于背景学科常用的方法与模型，即统计学与线性回归模型。从模型本身来讲，由于问题本身给出的数据并不具有真实性，因此在现实生物研究，或者说基因研究中需要被考量的众多因素都被忽略了。所以光从统计学模型来讲，我们的模型由于紧跟相关学科的研究方法，很难讲还有多少提升的余地，当然这是基于问题的抽象性，一旦应用于真实的数据与背景，则我们的模型并不具备足够的复杂性。

因此，由于问题与数据并非真实，或许仅仅从统计学角度进行分析，解释力度会不足。从数据结果中，很容易看出这个问题。在第二问，单SNP与病症的相关性分析中，有高达440个SNP位点的P值小于0.05，即通常情况下，我们会接受这全部440个SNP位点。然而，从数学建模竞赛的角度来看，显然440个解是很荒谬的。因此，单从统计学角度来分析，则不得不设置阈值来截取解，但是阈值的设置是非常主观的，解释力度欠缺。

所以，如果能用机器学习方法通过样本学习获得模型，并得到较理想的模型预测水平，则可以逆向增加解释力度。然而本题的数据模式使得模型的选择很有限，因为自变量有9445个，而样本只有1000个。从问题一到问题四，我们提出的模型都是基于统计理论和一些简单机器学习算法，理论上说这些算法对数据高维特性的反映并不是很强甚至没有，从数据中提取的也都是低维信息。很多显隐性基因表达方式的多样性无法被简单算法所刻画。使用深度学习（Deep Learning）恰恰能解决这些问题，深度学习能很好的刻画数据的高维特征，通过逐层的特征提取可以获取到很多一般算法无法分析出来的高维特征。而且位点基因分析就是从大量的数据中去寻找数据之间的相关性，这恰恰是深度学习最强的地方。考虑到深度网络搭建、训练和参数优化调整的时间较长，我们在本文中并未使用。但是使用深度神经网络分析位点或者基因之间的相关信息是一个很不错思路。

七、参考文献

1. [] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007, 447 (7145): 661-678. 10.1038/nature05911. [↑](#endnote-ref-0)
2. [] Balding DJ(2006) A tutorial on statistical methods for population association studies.Nat Rev Genet 7:781-91. [↑](#endnote-ref-1)
3. [] Winham SJ, Colby CL, Freimuth RR, Wang X, de Andrade M, Huebner M, and Biernacka JM, “SNP interaction detection with Random Forests in high-dimensional genetic data” ,BMC Bioinformatics , 2012 13:164. [↑](#endnote-ref-2)
4. [] F检验，百度百科 [↑](#endnote-ref-3)
5. [] 刘广智，基于SNP的连锁不平衡分析，百度文库 [↑](#endnote-ref-4)
6. [] EM算法，百度百科 [↑](#endnote-ref-5)
7. [] de Leeuw C, Mooij J, Heskes T, Posthuma D: MAGMA: Generalized gene-set analysis of GWAS data. PLoS Comput Biol 11(4): e1004219. doi:10.1371/journal.pcbi.1004219  [↑](#endnote-ref-6)
8. [] 肖静，胡治球。多个相关数量性状主基因的联合分析方法。中国农业科学 2005,38(9):1717-1724。 [↑](#endnote-ref-7)
9. [] 刘佳森，绵羊全基因组CNV与部分性状的关联分析，博士学位论文。 [↑](#endnote-ref-8)