代谢综合征风险、趋势预测和干预模型的构建

毕华瑞 耿明萌 李杨 龙永康 骆锦威
指导老师：李景治

南方科技大学

摘要

代谢综合征的致病机理复杂，其早期诊断与治疗是生物医学研究的重大挑战之一。本文通过分析基因型、基因表达水平和疾病之间的关系，结合生物信息学的知识和方法，通过对基因组数据和转录组数据进行分析与挖掘，找出了代谢综合征基因通路，构建代谢综合征风险预测模型。

本题所给的人类群体数据集可分为基因组数据和转录组数据：对于基因组数据，我们可以得到样本中的不同基因变异频率，对各基因与患病之间的关系进行分析，先后进行了基因简单筛选和趋势性基因筛选，构建并求解基于基因变异频率的模型，并对差异基因进行分析；对于转录组数据，利用GO功能富集分析、Pathway显著富集分析，利用变异系数和Mann-Kendall检验等方法处理数据，通过判断基因表达水平与患病程度之间的关系，剔除大部分不相关的基因，构建了基于基因表达量的预测模型。在求解以上模型的过程中，用到了支持向量机和决策树等数据挖掘的方法，提高了解决问题的速度和准确性，克服了样本数较少的缺点，达到了较高的交叉检验成功率。结合基因调控矩阵，证明了每个个体的基因交互状态向量中可能含有更多的信息。

现有的公开数据库中已经有大量关于生物分子相互作用和基因通路的信息，我们参考这些数据库并借助题目所给的数据，先研究了高频变异基因对其他基因表达量的影响、寻找上下游基因，实现了基于基因通路的预测。同时把之前模型的结果与实际数据库中的结论做对比，很多理论的结果与实际结论相吻合，在本文中以胰岛素信号通路中非常重要的PI3K基因为例进行说明。由于代谢综合征的成因较为复杂，本文还构造了“变异基因作用网络”的2-范数风险预测模型，可用于对造成代谢综合征的不同因素的预测。

本论文中所提到的方法能够依靠不同的数据，建立不同的模型，便于不同模型间的比较，同时也可以综合多种数据建模。本论文中的绝大多数模型都是基于数据而建立的，例如根据基因表达量和基因变异位点数据，其基本思想和方法不仅仅可以用于代谢综合征，可以应用于更多的疾病。虽然大部分模型是基于数据而建立的，但是我们同时还用实际科学研究发现和数据驱动的模型结果进行比较，来用于佐证模型实用性。

关键词：代谢综合征 生物信息学 数据挖掘

目录

插图

1 问题重述

当前的生命科学技术已经使得我们可以解析人类的遗传密码——基因序列。人与人之间基因序列的差异，即基因变异，影响着每个人罹患代谢综合疾病的风险高低，也影响着不同非遗传因素在每个人身上的具体作用效果。这也是为何代谢综合征在具有血缘关系的亲属之间有较高的发病关联的原因。此外，对人体生理运行动态特征，如基因的表达、各类小分子含量乃至于与人类紧密关联的微生物菌群的变化等，目前亦可以定量测量，从而实现了对人体当前运行状态动态监控。另一方面，各类自然和社会环境因素、生活方式等因素会对这些指征可能产生短期到长期的影响，最终形成多种外部表型特征，目前也可以通过各类移动医疗和健康设备加以记录，成为医疗及体检机构的临床检测数据。人体作为一个非线性复杂系统，其慢性疾病，特别是代谢综合征的发展，也是一个长期过程；而最终在临床上被诊断为代谢综合征，已经是这个慢性发展过程的结果。而我们平时所能够采取的健康和疾病预防手段，通常也只是针对于普适人群的平均推荐，不一定适合于每一个人。如果我们能够在早期对人体各类从内部到外部的因素进行测量、分析，可以构建一个早期的趋势预测模型，并且可以明晰这个复杂系统的具体问题所在，完成对每个人的个性化预防干预。

如果给你一个人类群体(~100人)，每个人具有较完整的下述生物医学数据：

1. 临床检测数据
2. 基因组数据
3. 表观基因组数据
4. 转录组数据
5. 蛋白质组数据
6. 代谢组数据

请你基于以上数据完成以下任务：

a) 请参考NCBI, EBI, DDBJ等公开数据库中的生物分子相互作用和基因通路信息，构建人类生命量化的动态模型；
b) 结合临床检测数据，哪些因素(数据特征或相互作用网络)是代谢综合征（一种临床诊断结论）的关键通路；
c) 若给定一个新的人类群体数据集（~10人），包含了每个个体的基因组、表观基因组、转录组、蛋白质组和（或）代谢组的部分测量，请问这些人有多大的代谢综合征风险？造成他们的代谢综合征风险的主要因素分别是什么？

2 问题背景与分析

2.1 代谢综合征的定义和诊断

自从工业革命以来，人们生活水平不断提高。但由于人体缺乏对营养过剩的抑制机制以及现代人缺乏良好的生活习惯等种种原因，代谢综合征(Metabolic Syndrome, MS)已经越来越成为人类健康问题的幕后推手。

代谢综合征是指在生理、生化、临床、代谢等层面能够直接增加心血管疾病、Ⅱ型糖尿病、和所有原因的死亡的风险的危险因子聚集现象。与代谢综合征相关的因素有很多，包括胰岛素抵抗、内脏肥胖、动脉粥样硬化、脂肪肝、血脂异常、内皮功能障碍、遗传易感性、血压升高、高凝状态、慢性应激等。

由于代谢综合征产生的因素较多，不同种族的特点也不同，目前无一致公认并适用于各种族的MS诊断标准，国际糖尿病联盟(International Diabetes Federation, IDF)、世界卫生组织(World Health Organization, WHO)、欧洲胰岛素抵抗研究小组(European Group

for the Study of Insulin Resistance, EGIR)、全美胆固醇教育计划(National Cholesterol Education Program, NCEP)、美国心脏协会(American Heart Association, AHA)等机构组织先后提出了不同的诊断标准。

2.2 代谢综合征相关的生物学背景

代谢综合征主要涉及人体物质能量代谢过程，人体细胞的能量摄取主要通过脂肪和糖代谢进行。物质代谢涉及到人体各个水平物质的动态平衡，血糖平衡是维持人体内稳态的关键，而胰岛素和胰高血糖素精确地控制着人体内血糖水平。自然地，胰岛素的信号通路也成为了调节内稳态的重要机制。

2.2.1 脂肪代谢(Lipid Metabolism)

脂肪代谢(Lipid Metabolism)涉及到人体的胆固醇(Cholesterol)、高密度脂蛋白(HDL)、血脂(Plasma lipid)等含量水平，其代谢过程如如图1所示。脂肪类物质从小肠吸收通过淋巴管形成乳糜微滴，在血管组织中通过甘油三酯水解形成甘油和脂肪酸运输到脂肪组织分解利用，部分多余乳糜微滴将流入肝脏在组装。从肝脏分泌的极低密度脂蛋白VLDL通过毛细血管水解形成IDL中密度脂蛋白，IDL经过脂蛋白脂肪酶作用形成低密度脂蛋白LDL，LDL由APOB基因表达合成。周围组织细胞将摄取LDL，作为能量来源，最终以HDL高密度脂蛋白的形式将多余脂肪排出，运送至肝脏回收利用。肝脏作为其中最重要的调控器官，调节脂肪摄入、运输、代谢等各项生理过程。脂肪代谢异常将引起血脂异常、高血压、脂肪肝等症状。

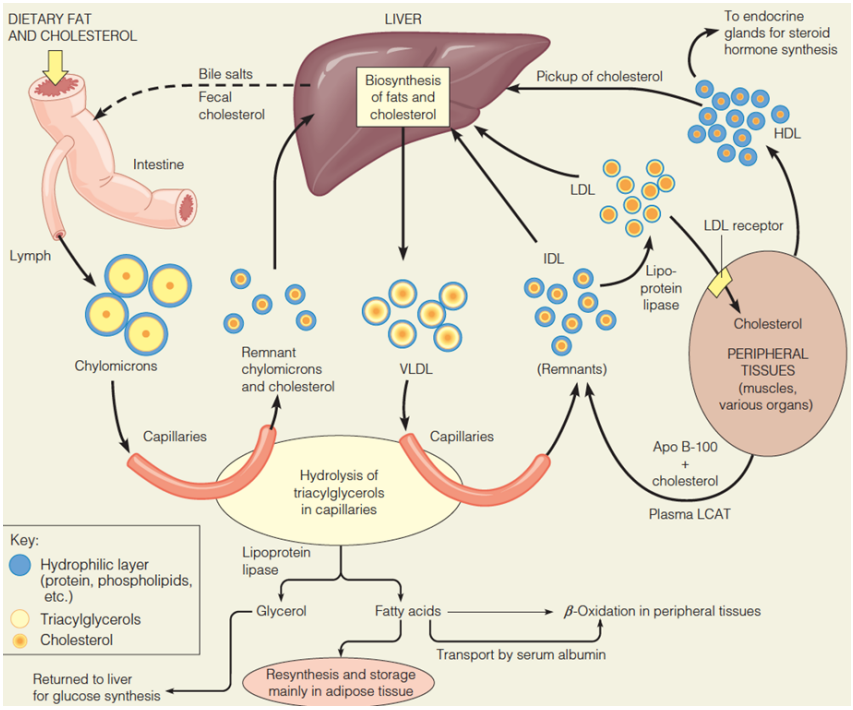


图 1: 脂肪代谢(Lipid Metabolism)

2.2.2 胆固醇代谢(Cholesterol Metabolism)

胆固醇的代谢直接涉及血液胆固醇的含量变化，高血液胆固醇将引起血管肥厚堵塞以至于破裂出血。胆固醇代谢(Cholesterol Metabolism)的过程如图2所示，细胞以胞吞方式摄取LDL，经过溶酶体酶解形成胆固醇脂滴，胆固醇分子经过内质网加工高尔基体组装重新形成LDL受体，LDL受体蛋白由LRP1B基因编码。同时胆固醇代谢异常也会引起心血管疾病(Cardiovascular disease)，例如冠心病、动脉粥样硬化。

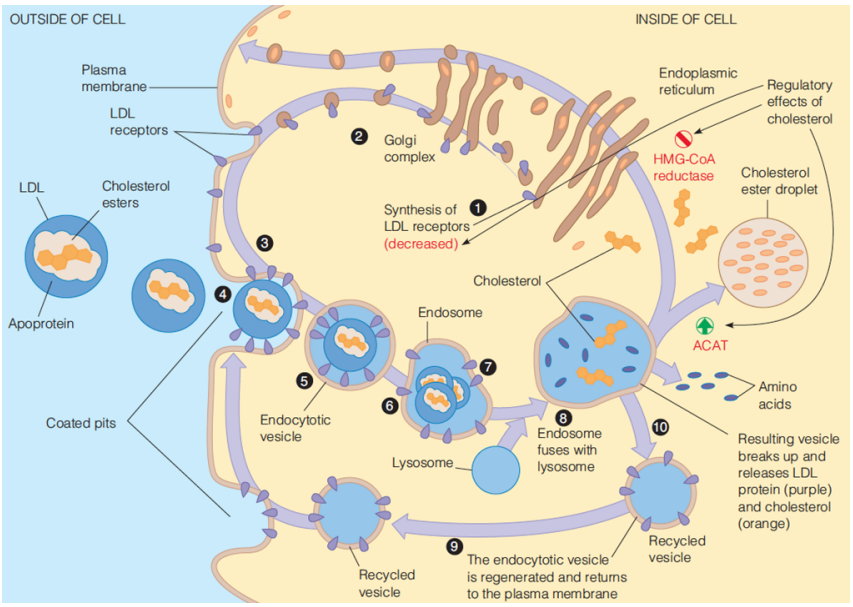


图 2: 胆固醇代谢(Cholesterol Metabolism)

2.2.3 胰岛素信号通路(Insulin Signaling Pathway)

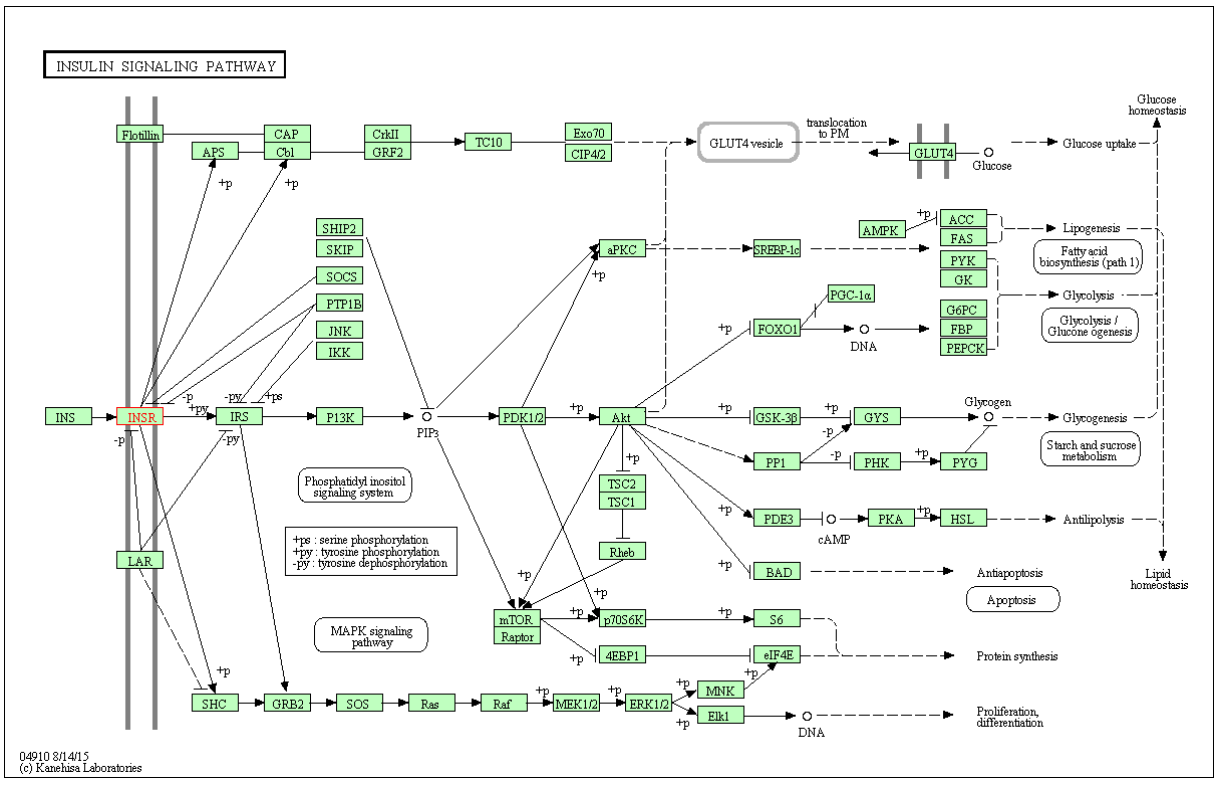


图 3: 胰岛素信号通路(Insulin Signaling Pathway)

胰岛素是人体调节血液葡萄糖含量的重要激素，对人体的物质能量代谢平衡有重要的作用。胰岛素的信号通路影响脂肪酸合成、糖酵解和糖异生、细胞分裂分化凋亡、蛋白质合成。胰岛素的异常代谢引起各种类型的糖尿病。胰岛素信号通路(Insulin Signaling Pathway)如图3所示，胰岛素分子与细胞膜表面的胰岛素受体蛋白INSR结合引

起IRS基因编码的胰岛素受体底物IRS磷酸化，IRS促进PI3KCA基因合成磷脂酰肌醇-3-激酶PI3K，进而促进IPCEF1基因表达使三磷酸肌醇PIP3的含量升高。同时INSR引起下游蛋白SHC磷酸化，通过促进GRB2、SC6、Ras、Raf基因表达翻译，引起MEK1/2蛋白的磷酸化，进而磷酸化MAPK4基因编码的ERK1/2蛋白。

2.3 代谢综合征相关代谢分子

自由脂肪酸分子(FFA) 高水平的FFA可能导致肝脂肪的积累从而导致脂肪肝，同时FFA可以通过抑制胰岛素介导的葡萄糖的摄取，而导致胰岛素抵抗，继而损害胰岛B细胞的功能。此外FFA还会增加纤维蛋白原和PAI-1的产生。在健康的个体中，胰岛素刺激胰岛素受体底物（IRS）-1允许其结合并激活PI3K活性，从而激活葡萄糖转运蛋白GLUT4，增加对葡萄糖的吸收，但是血浆中高的FFA水平能够激活蛋白激酶C(PKC)-*theta*，通过一个丝氨酸-苏氨酸激酶级联，减弱胰岛素刺激的IRS-1的酪氨酸磷酸化。

TNF-α ：诱导脂肪细胞的凋亡，通过抑制IRS-1通路从而促进胰岛素抵抗，但是TNF-α如何抑制IRS-1通路的机制目前还不太清楚。它能加剧FFA的释放，诱发动脉粥样硬化血脂异常。血浆中TNF-α浓度和体重，腰围，甘油三酯（TGs）呈正相关的关系，和高密度脂蛋白胆固醇（HDL-C）呈负相关的关系。

CRP 水平的升高和腰围的增加，胰岛素抗性，和高血糖相关联。此外，已经证明，CRP水平可以独立地预测CVD发生的概率，而不管个体是否患有或者患有何种程度的代谢综合征。

IL-6 是由脂肪组织和骨骼肌组织释放的，它同时具有炎性和抗炎作用。在大脑的多个区域，如下丘脑有IL-6的受体，用来控制食欲和能量的摄入。它不仅可以损害细胞对胰岛素的敏感性，也是肝产生CRP的主要决定因素。它被证明和二型糖尿病的程度呈正相关而和与HDL-C是呈负相关的。

PAI-1 通过抑制组织纤溶酶原激活剂（tPA）发挥其效果，因此被认为是一个障纤溶和动脉粥样硬化的一个标记，增加血管内血栓和有害心血管的风险。

脂联素(Adiponectin) 调节糖脂代谢，提高胰岛素敏感性，调节食物摄入量和体重，防止慢性炎征。它抑制肝脏酶和肝内源性葡葡萄糖异生酶和肝脏内源性葡萄糖产生的速率。它增加了肌肉的葡萄糖转运和脂肪酸的氧化，它能从多个方面的抗动脉粥样硬化，其中包括内皮细胞的活化，降低巨噬细胞转化成泡沫细胞的转化率，抑制平滑肌细胞的增殖和动脉重塑来抑制粥样硬化斑块的发展。脂联素的是与CVD危险因子的呈负相关的，比如血压，LDL-C和TGs。此外，Fumeron等人得出结论，低脂联素血征与胰岛素抵抗，高胰岛素血征，患二型糖尿病的概率相关，并且独立于脂肪含量。脂联素作为抗炎因子，它与HDL-C呈相关，和体重，腰围，TGs，空腹胰岛素，胰岛素抵抗（HOMA-稳态模型评估），BMI和血压呈负相关。它还与TNF-α形成拮抗作用。

2.4 思路分析

问题的第一问需要参考公开数据库中的生物分子相互作用和基因通路信息，构建人类生命量化的动态模型。不同的数据库有不同的特点，我们选择合适的数据库进行分析。

问题的第二问要求结合临床检测数据，判断哪些因素是代谢综合征的关键通路。原始数据较为繁杂，需要进行预处理，清除异常数据和重复数据，并纠正数据中的错误。

处理后的数据主要包含基因表达量信息和基因变异信息，可以用统计和机器学习的方法解决。但是由于本题中的样本数量不足，需要选用对数据要求较少的方法。

问题的第三问要求对新给定的人类群体数据集，给出这些人的代谢综合征的风险并给出主要的风险因素。基于第二问建立的模型，可以结合数据去预测。由于大多数与疾病相关的变异并不能改变一个基因的功能，只能改变基因的表达水平，将基因组数据和转录组数据结合，才能更好实现对全基因组数据转录水平的估算。由于本题中所给数据的种类较少，需要对用简化的模型实现对代谢综合征的影响因素的预测。

在MATLAB的工具箱中，包含了多种分类器，包括决策树(Decision Trees)、判别分析(Discriminant Analysis)、Logistic 回归(Logistic Regression)、支持向量机(Support Vector Machines)、最近邻分类器(Nearest Neighbor Classifiers)、集成分类器(Nearest Neighbor Classifiers)等。支持向量机在解决小样本、非线性和高维模式识别问题中表现出许多特有的优势。决策树也是一种常见的分类方法，在经过给定的数据集的训练之后，可以对新示例进行分类，易于理解和实现。我们可以利用这些方法对数据进行初步的分析处理，随后挖掘出更多的生物学含义。

3 模型假设

1. 假设基因测序的结果是正确的。
2. 假设每一个个体的高频基因患病程度中蕴含着其患病信息。
3. 假设同一个基因不同方式的变异具有相同的效果，即对于不同的变异位点坐标、变异碱基、变异类型，只要他们变异所在的基因相同就会产生同样的效果。
4. 假设被筛选出的异常数据不是正常数据。
5. 假设不同的个体生活环境相同。

4 模型建立

生物体是一个耗散结构，也是一个多层次的信息网络，其内部存在大量的相互作用和反馈循环。题目中已经给出了人类群体的生物医学数据，通过机器学习的相关方法，可以找出变异基因、基因表达量和患病程度之间的关系，从而完成对代谢综合征的风险和趋势的预测。

在题目所提供的RNA-seq基因表达文件中，每一个患者的数据都有超过2万个基因。基于生物学和统计学知识，可以过滤出无用的基因、减少数据量，进而根据基因的表达量信息对患者的患病程度分类。对于筛选后的基因，利用机器学习的方法对其进行分类，建立基于基因表达的模型的，得到代谢综合征的患病程度和基因表达量的关系。

在题目提供的变异位点信息中，由于变异分类和变异类型过于复杂，我们可以只考虑对基因发生变异的次数的统计，选出高频变异基因。同样运用机器学习的方法，建立基于基因变异次数的模型，寻找基因变异与代谢综合征的关系。

综合考虑基因表达量和基因变异次数，建立综合模型，通过临床数据来预测代谢综合症的患病程度和主要影响因素。现在的生物学已经有不少对于代谢综合征的研究，本题所得到的实验结果可以是其他科学研究的数据做对比。

流程图如下：

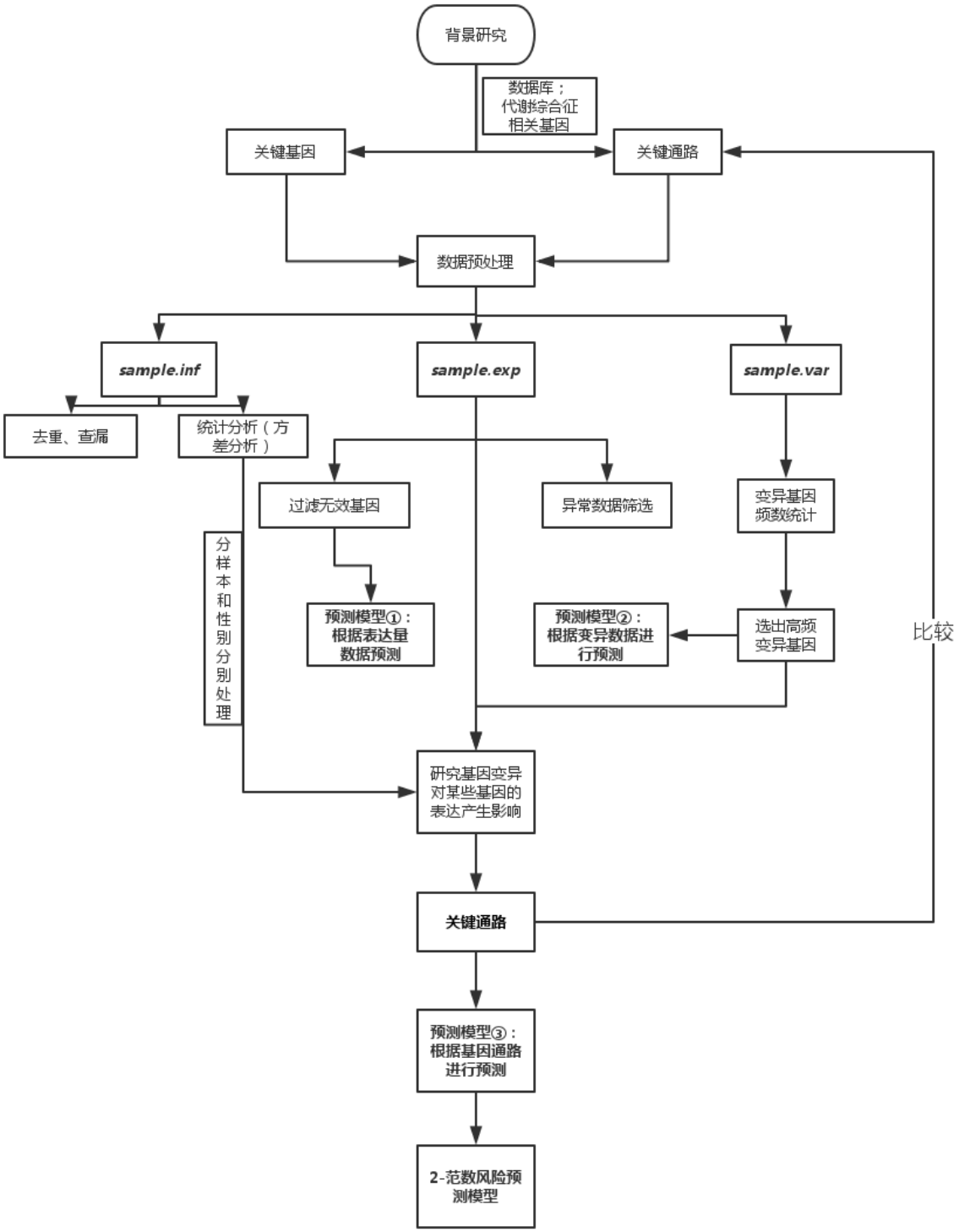


图 4: 流程图

5 数据处理

5.1 数据预处理

在题目的原始数据中，有些错误较为明显，需要进行简单的处理，如：

(1) 在sample.exp.0中，出现两个54号数据。处理方式：直接删除。

(2) 在sample.exp.0中，无样本205号表达量数据。处理方式：删除。

(3) 在sample.exp.1中，表达量数据异常。处理方式：删除。

(4) 在sample.var.0和sample.var.1中，数据大量重复。处理方式：删除。

对上述数据进行预处理之后，提高了数据处理的质量，使得结果更准确。

5.2 数据统计分析

本题所给的数据主要包含了患者的患病程度、变异位点信息和基因表达信息。但是也有性别信息和组别信息，在进行数据处理之前，为了考虑组别和性别两个因素是否会对患病情况产生影响，我们利用统计学检验方法——可重复双因素方差分析，进行检验，结果如表1、表2和表3所示：

表 1: 方差分析一

Source	SS	df	MS	F	Prob>F
Columns	0.612	1	0.612	0.06	0.8131
Rows	0.313	1	0.313	0.03	0.8658
Interaction	103.513	1	103.513	9.52	0.0028
Error	826.55	76	10.876		
Total	930.987	79			

表 2: 方差分析二

Source	SS	df	MS	F	Prob>F
Columns	1.8	1	1.8	0.18	0.6763
Rows	6.05	1	6.05	0.59	0.4447
Interaction	84.05	1	84.05	8.2	0.0054
Error	778.9	76	10.2487		
Total	870.8	79			

表 3: 方差分析三

Source	SS	df	MS	F	Prob>F
Columns	5.512	1	5.5125	0.55	0.4601
Rows	6.613	1	6.6125	0.66	0.4187
Interaction	49.613	1	49.6125	4.96	0.0289
Error	760.15	76	10.002		
Total	821.888	79			

从方差分析结果可以看出，我们不能得出性别因素对患病结果有显著影响；不能得出样本因素对患病结果有显著影响；但是，我们可以从P-value看出，样本因素与性别因素会产生交叉影响。因此，在处理数据时，我们首先分别分样本处理，再分别分性别处理，最后分析并比较结果。

5.3 数据库

生物分子相互作用和基因通路信息需要从公开的数据库中选择，需要对比不同数据库的特点，获取更多有用的信息。

GWAS 数据库 通过GWAS全基因组关联分析数据库可查找代谢综合征相关的单核苷酸多态性（SNPs），并将reported genes整理成表，用做关键基因的筛选匹配。

NCBI/DDBJ/EBI/GeneCards/Uniprot数据库 NCBI、DDBJ、EBI、GeneCards和Uniprot数据库共享基因和蛋白质信息。我们通过这三个数据库搜索查找从数据中得出的关键基因，找出这些基因的功能及相互作用。

Gene Ontology数据库 Gene Ontology 数据库可以通过基因富集分析得出相关的细胞组分、分子功能和生物过程。富集的基因由病患样本的差异表达基因筛选和变异基因筛选得出。

KEGG数据库 KEGG数据库可以查找跟代谢综合征相关的基因通路和代谢机理。

6 模型求解与结果分析

6.1 模型一：基于基因表达量的预测模型

6.1.1 基因简单筛选

在题目所给的两批样本的基因表达量信息分别储存在sample.exp.0文件和sample.exp.1文件中。对于不同批次不同类型的病人，对其基因分开进行处理。对于同一批的样本的第i个基因，记其全体为G_i，记该G_i的标准差为σ_i、均值为μ_i，则该类型样本的变异系数v_i可由下式求得：

$$v_i = \frac{\sigma_i}{\mu_i} \tag{1}$$

变异系数是概率分布离散程度的一个归一化量度，消除了测量尺度和量纲的影响，反应了数据离散程度的绝对值。在同一批次中表达量变化小的基因意味着此基因对于患病情况影响小将这些基因去除掉，因此筛选时作以0.1为阈值。

当基因的表达量过少时，RNA无法翻译足够多的多肽来合成蛋白质，且当平均值接近于0的时候，微小的扰动也会对变异系数产生巨大影响，因此造成精确度不足，因此取$\mu_i > 5$进行筛选。由于测量时存在误差，每一组G_i的样本数目较少，异常值可能会对该组数据的均值产生较大的影响，为了过滤掉存在异常值的基因，以$\mu_i + 3\sigma_i$作为条件再次进行筛选。经过这两次筛选后，对同一组别不同的患病类型的剩余的基因取交集得到初步筛选的结果：组别0（samlple.exp.0）中还剩下11684个基因，组别1(sample.exp.1)中还剩下12923个基因。

在之前的筛选中，剔除了有较大异常值影响该组基因表达量平均值的情况。对于同一批样本不同类型的第j个基因的表达量的平均值（每个基因共11个值，对应着11种不同的患病程度），记其平均值μ'_j,标准差为σ'_j，类似的，定义该类型样本的变异系数为v'_j

$$v'_j = \frac{\sigma'_j}{\mu'_j} \tag{2}$$

为了筛选出在不同患病程度下表达量有较大差异而在同一患病程度下基因表达量稳定的基因，以$v'_i > 0.5$为条件再次进行筛选。在组别0(sample.exp.0)中得到1723个基因，在组别1(sample.exp.1)中得到1682个基因，其中既在组别1中出现又在组别0中出现的基因为1214个。在此基础上，通过变异基因和数据跳跃分析对备选基因进行筛选可得到代谢综合症的关键通路。

6.1.2 趋势性基因筛选

对于初步筛选基因，即在组别0（samlple.exp.0）中的13379个基因和组别1(sample.exp.1)中的13253个基因进行趋势性基因筛选。我们采用mann-kendall趋势分析方法，此检验方法是非参数方法，不需要样本遵从一定的分布。

sample.info.txt文件中有两批样本中病人的患病程度的信息，在该文件中病人被划分为11种类型（I、IA、IB、II、IIA、IIB、 III、IIIA、IIIB、IV、IVA）。对于同一个样本病患（即sample.exp.0和sample.exp.1中的每一列）按照对应患病程度深浅按照由轻到重进行排序，并将其表达量记为$(X_1, X_2, \cdots, X_n)$，在本题中$n = 11$。

在 Mann-Kendall检验中，原假设是该序列数据是n个独立的随机变量同分布的样本。检验的统计量为S，可以由下面的式子计算：

$$S = \Sigma_{k=1}^{n-1}\Sigma_{j=k+1}^{n} Sgn(X_j - X_k) \tag{3}$$

即对于$m > n$：
若$(X_j - X_k) > 0$，则$Sgn(X_j - X_k) = 1$；
若$(X_j - X_k) > 0$，则$Sgn(X_j - X_k) = 0$；
若$(X_j - X_k) > 0$，则$Sgn(X_j - X_k) = -1$。
S为正态分布，均值为0，其方差可由下式计算：

$$Var(S) = \frac{n(n-1)(2n+5)}{18} \tag{4}$$

当n>10时，标准正态系统变量可分为下面的几种情况计算：
如果$S = 0$，则表示分类该基因的表达量在患病情况中没有趋势。
如果$S > 0$，则

$$Z = \frac{S+1}{\sqrt{Var(S)}} \tag{5}$$

如果$S > 0$，则

$$Z = \frac{S-1}{\sqrt{Var(S)}} \tag{6}$$

定义显著性水平α=0.05，由于S符合标准正态分布，将正态分布的0.025位点1.96作为筛选阈值。如果$|Z| > 1.96$，则将认定该基因具有趋势。进一步将最终筛选出的具有表达量趋势性的基因在生物数据库中进行查找验证，得到与代谢综合症具有重大关联的基因。

6.1.3 差异基因分析

在基因简单筛选中：经过第一步，组别0（samlple.exp.0）中还剩下13379个基因、组别1(sample.exp.1)中还剩下13253个基因；经过第二步，组别0(sample.exp.0)中还剩下3446个基因、在组别1(sample.exp.1)中还剩下3319个基因。

在第二步中剔除的基因中找出存在表达量之间差异大于1000倍的基因，即$\frac{|\mu_{i_1} - \mu_{i_2}|}{min(\mu_{i_1}, \mu_{i_2})} >$ 1000且表达量数值$\mu_{i_1} > 10$的基因，在组别1（sample.exp.1）中得到118个异常基因，组别0（sample.exp.0）中得到128个异常基因。这些表达量异常值可能由于该病患的对应基因产生突变或者与此基因调控关系的基因产生了变异，对于这些基因进行单独分析。

6.1.4 模型求解

针对筛选出的具有根据患病程度加深具有趋势性变化的基因，并以此基因为分类模型输入，患病程度为标签，利用决策树和支持向量机的相关算法进行分类，在MATLAB中进行交叉验证的结果如图5和图7所示，这几种方法的差异并不明显。对于组别0和组

别1各选取了一个混淆矩阵，如图6和8所示。从图中可以看出，本模型的预测结果一般。

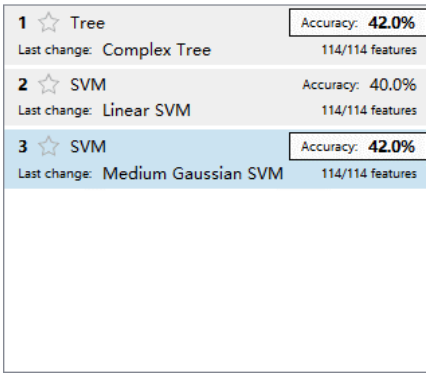


图 5: 组别0模型一交叉验证的结果

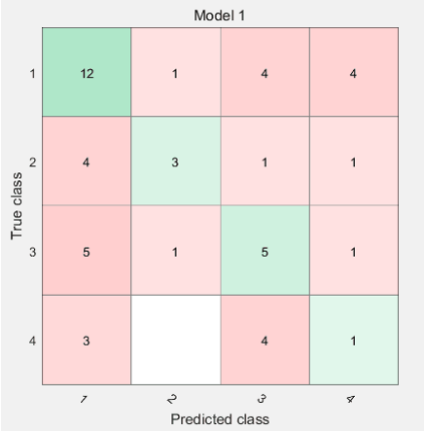


图 6: 组别0中方法1的混淆矩阵

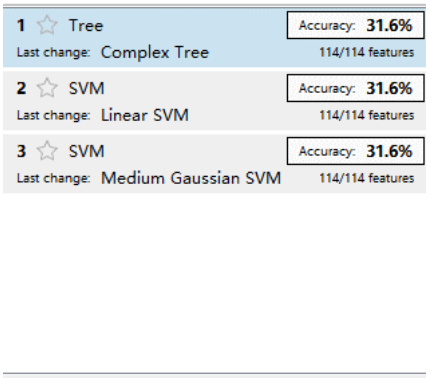


图 7: 组别1模型一交叉验证的结果

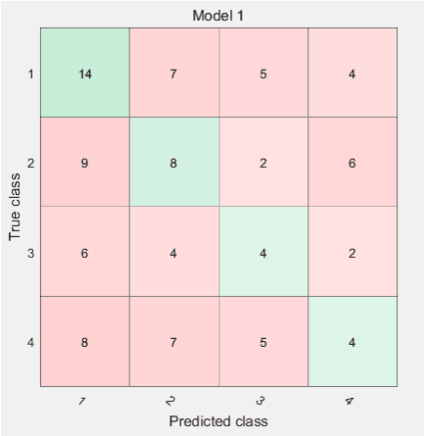


图 8: 组别1中方法1的混淆矩阵

6.2 模型二：基于基因变异频率的预测模型

6.2.1 基因变异频数统计

在本题所提供的变异位点文件中包含变异位点坐标、变异碱基、变异类型、变异所在基因和变异对基因的功能影响，对于同一个基因不同方式的变异，可能会产生不同的效果。在本题中，为了简化计算过程，我们假设同一个基因不同方式的变异具有相同的效果，即只统计该基因的变异的次数，对于组别0和组别1，统计结果如图9和图10所示。从图中可以看出，大部分基因变异的次数较低，与代谢综合征相关的基因发生变异的可能性要高于与代谢综合征无关的基因。

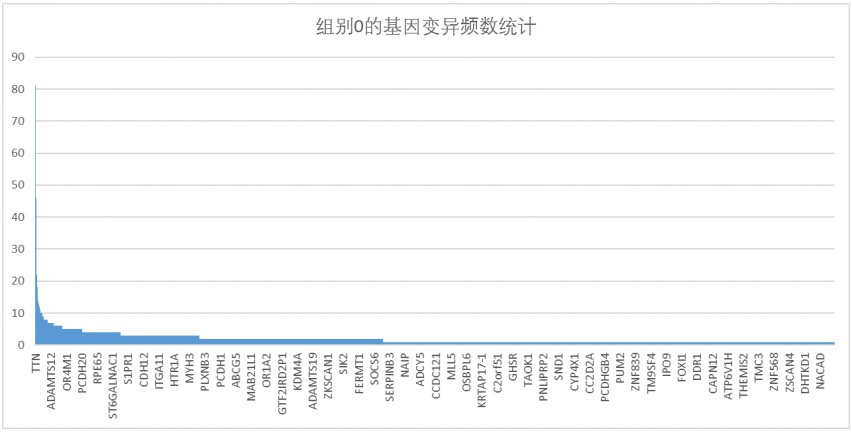


图 9: 组别0的基因变异频数统计

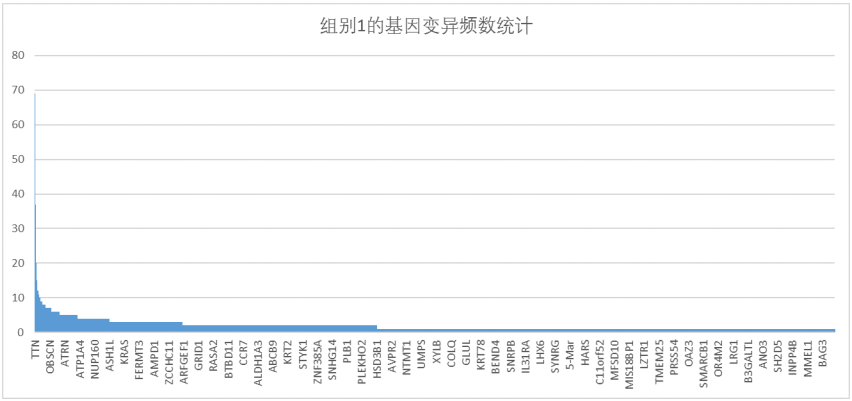


图 10: 组别1的基因变异频数统计

6.2.2 高频变异基因分析

对于不同的组别，我们分别对变异基因进行频数统计，并选出各个组别变异频率最高的前20个基因，结果见表4和表5。

表 4: 组别0的基因变异频数变异最高的20个基因

TTN	MUC16	TP53	RYR2	CSMD3	PIK3CA	LRP1B	USH2A	ZFHX4	FLG
SYNE1	PCLO	SPTA1	XIRP2	HMCN1	MUC17	CSMD1	OBSCN	ANK2	MLL3

表 5: 组别1的基因变异频数变异最高的20个基因

TTN	TP53	MUC16	VHL	CSMD3	FLG	LRP1B	RYR2	FAT1	PBRM1
ZFHX4	MALAT1	TRIP11	NEB	RYR3	AHNAK2	USH2A	LAMA4	PIK3CA	DMD

6.2.3 模型求解

在获得高频变异基因的基础上，我们通过合并样本信息文件（sample.inf.txt）与变异位点文件(sample.var.0, sample.var.1)，我们将每一个个体高频变异基因的变异个数与其患病程度联系起来。我们假设每一个个体的高频基因患病程度中蕴含着其患病信息，因此，我们以20个高频基因为其特征描述，以患病程度分为4个层次）为标签，利用决

策树和支持向量机的相关算法进行分类，利用MATLAB进行交叉验证的结果如图11。对于不同的方法，我们可以求出其混淆矩阵，例如图12为方法1复杂决策树的混淆矩阵：

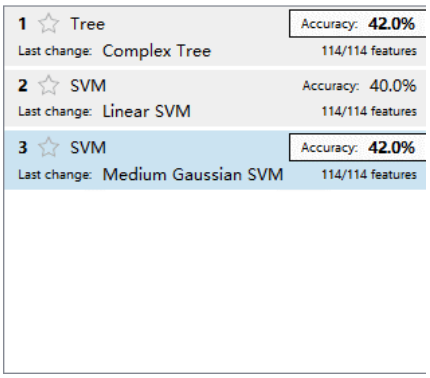


图 11: 组别0模型二交叉验证的结果

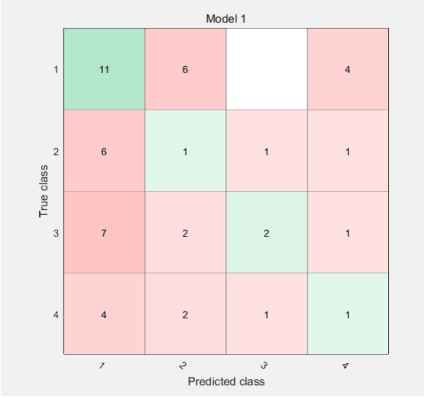


图 12: 组别0中方法1的混淆矩阵

类似的，我们可以对组别1进行分析，结果如图13和图14所示：

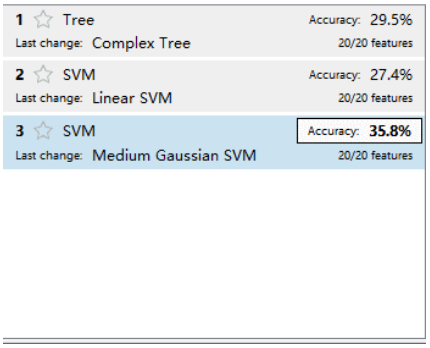


图 13: 组别1模型二交叉验证的结果

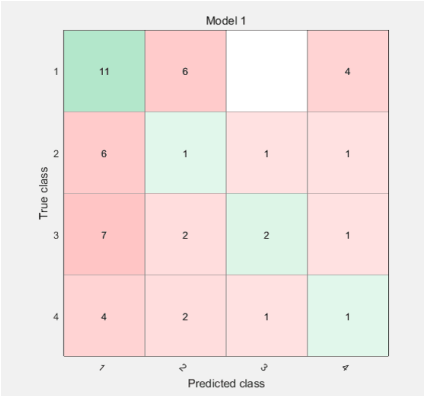


图 14: 组别1中方法1的混淆矩阵

6.2.4 结果分析

在数据量看，虽然分类标签量小（4个），但是由于样本个体实在太小，因此，机器学习算法较难准确获得每个标签的特征，这会导致准确率较低。

从预测准确率上看，三种算法预测准确率都较低，其中，高斯核的支持向量机算法准确率较高，结合三个算法的特点分析，我们认为这与基因和患病程度之间的有非线性关系有关；

从混淆矩阵看，错误比较集中在预测为stage1部分，我们可以因此推断，由于stage1的患病程度较轻，特征不如患病程度重的明显，因此更容易将样本误分到stage1。

比较组别0与组别1的结果，我们可以发现，从高频变异基因上看，组别0与组别1表现较为一致。

6.3 模型三：以混合状态进行分类

6.3.1 高频变异基因的查找

在图9和图10中，我们已经对两个基因变异文件（sample.var.0和sample.var.1）做频数统计的处理。为了探索高频变异基因对于其他基因表达量的影响，每个样本选取即变异频率在样本中超过12%的基因，即前24个基因。为了使每个样本的基因频率不要太

低，这24个基因在另外一个样本中超过5%的频率才被选取（如基因 A 在组别0中频率排在前24位，但是在组别1中频率低于5%不会被选取)。此外，如果这个基因在两个样本中的变异频率都超过10%也会被选取。我们得到了29个的基因列表（见表6）。

表 6: 29个高频变异基因

XIRP2	AHNAK2	ANK2	APOB	ARID1A	CSMD1	CSMD3	DMD
FAT1	FLG	HMCN1	LRP1B	MALAT1	MLL3	MUC16	MUC17
NEB	OBSCN	PCLO	PIK3CA	RYR2	RYR3	SPTA1	SYNE1
TP53	TTN	USH2A	VHL	ZFHX4			

6.3.2 高频变异基因对其他基因表达量的影响

我们通过NCBI和EBI数据库分析查找变异次数频率较高的29个基因，有17个基因跟代谢综合征相关。例如，PIK3CA基因是编码PI3K蛋白的基因，该基因调控IRS-1底物水平从而调控胰岛素信号通路；NEB基因翻译的蛋白与细胞骨架相关而且是2型糖尿病和缺血性心力衰竭中的关键基因；CSMD1基因维持血糖平衡的相关基因，同时也是潜在的肿瘤抑制因子；APOB和LRP1B基因是编码LDL和LDL受体蛋白的关键基因，在脂肪运输代谢过程中起到重要作用;VHL基因参与蛋白质泛素化过程，该基因编码的蛋白能够结合RNA聚合酶调控氧化代谢相关基因的转录过程；TP53基因是肿瘤抑制基因，调控许多细胞活动包括细胞周期、衰老、DNA修复、代谢水平调节等。

通过Gene Ontology数据库对29个基因进行基因富集，找到了相关的生物过程如下图所示：

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW! ⓘ)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
regulation of SA node cell action potential	2	2	.00	> 100	+	3.00E-02
protein localization to M-band	2	2	.00	> 100	+	3.00E-02
regulation of cardiac muscle contraction by regulation of the release of sequestered calcium ion	17	3	.02	> 100	+	1.56E-02
↳regulation of cardiac muscle contraction by calcium ion signaling	21	3	.03	> 100	+	2.93E-02
↳regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum	23	3	.03	94.33	+	3.85E-02
↳regulation of biological quality	3460	16	4.78	3.34	+	1.88E-02
response to muscle stretch	19	3	.03	> 100	+	2.18E-02
cardiac muscle contraction	67	5	.09	53.97	+	3.02E-04
↳striated muscle contraction	99	5	.14	36.52	+	2.06E-03
↳muscle contraction	232	6	.32	18.70	+	5.70E-03
↳muscle system process	281	6	.39	15.44	+	1.72E-02
↳heart contraction	78	5	.11	46.36	+	6.40E-04
↳heart process	85	5	.12	42.54	+	9.76E-04
actin-mediated cell contraction	74	5	.10	48.86	+	4.93E-04
↳actin filament-based movement	93	5	.13	38.88	+	1.52E-03
↳actin filament-based process	461	9	.64	14.12	+	6.57E-05
muscle cell differentiation	245	6	.34	17.71	+	7.81E-03
↳muscle structure development	451	9	.62	14.43	+	5.44E-05
cytoskeleton organization	904	9	1.25	7.20	+	1.91E-02
homeostatic process	1374	11	1.90	5.79	+	8.87E-03
Unclassified	4069	2	5.63	.36	-	0.00E00

图 15: 对高频变异基因进行基因富集的结果

由图中发现，突变频率较高的基因与心脏肌肉收缩、肌肉细胞分化以及心脏细胞收缩调控（钙离子通道）行为相关，可能与冠心病、动脉粥样硬化疾病相关。

得到变异频数较高的变异基因后，我们用perl对于每个基因，我们在基因变异文件中找到相应的病人编号，并且从筛选出的基因表达文件ample.var.0和sample.var.1中找到相应的病人编号，并且保存到同一个文件中，剩下的没有此基因变异的病人编号保存到

另一个文件中。每个变异频率较高的基因都产生了阳性和阴性基因两个文件，所有在两个样本中总共有29*2*2个文件。我们对这些文件中的每一行的基因表达量的均值并且取对数，然后用阴性的数值减去阳性的数值，在每个样本中得到29个变异频率较高的基因对我们筛选出的重要基因的表达量的影响。

图16和图17为exp.0.diff.FLG文件（FLG为29个高频变异基因之一）的前后两部分截图。第一列为基因名称，第二列为阴性病人（相对应基因没有变异，这里指FLG没有变异）基因表达量的均值取2为底的对数，第三列为对应的阳性病人的数据（处理方法与第二列相同），最后一列为第二列减去第三列的差，正值表明这个基因的变异之后，使得第一列的基因表达量下降，即若此基因正常，对应基因的表达量比基因变异时更强。负值表明这个基因变异之后第一列的基因表达量上升了，即若此基因正常，对应基因的表达量比不让基因变异时强。

```
 1 SLC3A1|6519     10.4014240969195        1.82763205847966        8.57379203843985
 2 SLC28A1|9154    8.12856529500338        1.75844814658012        6.37011714842326
 3 KCNJ16|3773     8.89674475201516        4.6748429040984 4.22190184791676
 4 USH1C|10083     8.55795655129566        4.8214415238428 3.73651502745286
 5 SALL1|6299      7.76397667597218        4.28915821342381        3.47481846254836
 6 ACMSD|130013    7.13228479014368        3.69995426979691        3.43233052034677
 7 SLC47A1|55244   9.42187824932903        7.55267755894455        1.86920069038448
 8 MYOM3|127294    7.20441572083358        5.80639697766709        1.39801874316649
 9 MAOB|4129       10.8097630315016        9.4454092418235 1.36435378967812
10 SLCO4C1|353189  8.56406608818022        7.23010675075261        1.33395933742761
11 PGAP3|93210     10.7211714444828        9.46530650463622        1.25586493984656
12 HOXD8|3234      7.59079091569021        6.37146058156552        1.21933033412469
13 PDZK1|5174      9.52406579672175        8.37955871027033        1.14450708645142
14 PLA2G16|11145   10.7419178828854        9.65232542856374        1.08959245432166
15 HPN|3249        10.0454605380072        9.03959728075318        1.005863257254
16 GIPC2|54810     6.87254467640632        5.86721481269768        1.00532986370864
17 HOXC10|3226     9.15960394769882        8.19023123748222        0.9693727102166
18 LOC100101266|100101266  2.73358944840166       1.80543004533013        0.928159403071529
19 ERBB3|2065      12.6391740587598        11.7210702053456        0.918103853414223
20 PECI|10455      10.1888582286305        9.30813374340829        0.880724485222256
```

图 16: FLG基因变异使部分基因表达量上升

```
1195 CBX8|57332      7.24968375570015        7.89828433876897        -0.648600583068825
1196 GPR115|221393   6.8206144463943 7.47854205282166        -0.657927606427361
1197 MERTK|10461     8.1305012305928 8.81120795083348        -0.680706720240677
1198 DCAF15|90379    8.88852968395152        9.57412690411602        -0.685597220164498
1199 NACA2|342538    6.34258960419827        7.04735478863917        -0.7047651844409
1200 FAM115C|285966  8.50114068271777        9.2244664667755 -0.723325784057735
1201 ITGB7|3695      7.95428046002111        8.68429510864843        -0.730014648627324
1202 PLAUR|5329      10.0251989403136        10.7650920878519        -0.73989314753832
1203 GNG7|2788       7.30289654711751        8.05556463634175        -0.752668089224248
1204 SHKBP1|92799    10.4679339216372        11.2305576429639        -0.762623721326745
1205 SH3BP1|23616    9.17899280755321        9.96972315857208        -0.790730351018867
1206 GPRIN2|9721     7.41400617515903        8.25154395940059        -0.837537784241561
1207 PVRL1|5818      11.0840949124511        11.9543691030275        -0.870274190576426
1208 HMGA1|3159      11.631269425357 12.5054983903238        -0.87422896496679
1209 ZNF215|7762     4.71437469656065        5.60845190211831        -0.894077205557651
1210 GUSBP1|728411   6.82757943923652        7.73586659866593        -0.908287159429407
1211 MFSD2B|388931   5.13738969726978        6.06458442265595        -0.927194725386162
1212 LRIG3|121227    9.18413762184164        10.2678604276406        -1.083722805799
1213 GJB5|2709       7.16315960060121        8.45318904136803        -1.29002944076682
1214 ITGAM|3684      9.0230015463807 10.4202902494282        -1.39728870304747
```

图 17: FLG基因变异使部分基因表达量下降

为了消除随机性的影响，我们用哈希表随机给阴性和阳性分配病人编号，并且对对称(阴性和阳性病人一样多）和不对称（阴性和阳性最多差别 3 倍）都做了多次随机模拟，发现每次排在两端的基因都不一样，而且最大值不会超过0.9。图18为某次随机模拟截图。

1	DQX1|165545	6.25450202366513	5.40524614662295	0.849255877042173
2	GPR115|221393	8.38225842037636	7.53974200889063	0.842516411485728
3	GJB5|2709	9.69624822821534	8.95571171512544	0.740536513089896
4	PLEKHN1|84069	7.44931795002295	6.75462281160731	0.69469513841564
5	BOK|666	10.1891159127561	9.61520701438827	0.573908898367817
6	TUBB2C|10383	13.1716286295095	12.6030035319725	0.568625097537014
7	PVRL1|5818	12.7487702835702	12.2255654380146	0.523204845555535
8	ESRRA|2101	10.4679546238098	9.95559921500616	0.512355408803677
9	ATAD3B|83858	8.22305075493825	7.74534693878453	0.477703816153716
10	KREMEN1|83999	10.8522315107196	10.3878204320912	0.464411078628403
11	VPS37B|79720	10.2901895869866	9.82972124163571	0.460468345350858
12	MFSD2B|388931	5.7522313319634	5.30840182883888	0.443829503124525
13	PUSL1|126789	7.91536957590957	7.47576690802035	0.439602667889221
14	MAP7D1|55700	11.749780936882	11.3149841322061	0.434796804675967
15	CNKSR1|10256	8.61191456343028	8.18115177158094	0.430762791849332
16	SH3BP1|23616	9.98457764224836	9.5694352609873	0.415142381261058
17	DVL1|1855	10.5635262799935	10.1505565987533	0.412969681240231
18	DGKA|1606	10.2651223208175	9.8536105851595	0.411511735657962
19	TSPO|706	11.3669293154581	10.9668194870378	0.40010982842033
20	PRICKLE3|4007	8.18745415453282	7.79085260754262	0.396601546990195

图 18: 高频变异基因对其他基因表达影响的随机模拟

6.3.3 寻找上下游基因

我们选取阈值为-1和1（即表达量受基因变异的影响大于两倍），对于在-1和1范围内的基因，我们认为高频变异的基因对于这些基因的表达量没有显著影响，赋予他们数值0，对于超过1的基因，赋值1，低于-1的基因，赋值-1。用perl把这29个变异基因联系起来，每一行为一个变异基因，每一列为一个我们筛选出来的重要基因，生成一个只有0,1和-1的矩阵。如果有一列基因全部为0，那么表明，这个基因的表达量不会受到这29个变异基因的影响，我们将会删除这些基因。这样，对于组别0，我们得到254个相关基因，对于样本1，我们得到312个相关基因。因而达到下图中组别0的（matrix.0）调控矩阵：

1		A2M|2	ABI3|51225	ACACA|31	ACADS|35	ACAT1|38	ACAT2|39	
2	AHNAK2	0	0	0	0	1	0	0	1	0
3	ANK2	0	0	0	0	0	0	0	0	0
4	APOB	0	0	0	0	1	0	0	1	0
5	ARID1A	0	0	0	0	1	0	0	0	-1
6	CSMD1	0	0	0	0	0	0	0	0	0
7	CSMD3	0	0	0	0	1	0	0	1	0
8	DMD	0	0	0	0	1	0	0	1	0
9	FAT1	0	0	0	0	1	0	0	1	0
10	FLG	0	0	0	0	1	0	0	0	0
11	HMCN1	0	0	0	0	0	0	0	0	0
12	LRP1B	0	0	0	0	1	0	0	1	0
13	MALAT1	0	0	0	-1	1	0	1	0	0
14	MLL3	0	0	0	0	1	0	-1	0	0
15	MUC16	0	0	0	0	1	0	0	1	0
16	MUC17	0	0	0	0	1	0	0	0	0
17	NEB	0	0	0	0	1	0	0	0	0
18	OBSCN	0	0	0	0	1	0	0	0	0
19	PCLO	0	0	0	0	1	0	0	0	0
20	PIK3CA	0	0	0	0	1	0	0	0	0
21	RYR2	0	0	0	0	1	0	0	0	0
22	RYR3	0	0	0	0	1	0	0	0	0
23	SPTA1	0	0	0	0	0	0	0	0	0
24	SYNE1	0	0	0	0	1	0	0	0	0
25	TP53	0	0	0	0	1	0	0	0	0
26	TTN	0	0	0	0	1	0	0	0	0
27	USH2A	0	0	0	0	1	0	0	0	0
28	VHL	-1	-1	-1	-1	-1	-1	0	0	0
29	XIRP2	0	0	0	0	0	0	0	1	0
30	ZFHX4	0	0	0	0	1	0	0	0	0

图 19: 调控矩阵

6.3.4 基因调控矩阵

在前文中，我们已经找到了20个高频基因，通过算法分析最终得出突变高频基因与差异表达基因的调控关系，其中突变高频基因对差异表达基因分别有促进（+）和抑制（-）作用。他们总共促进或抑制了114个基因，调控矩阵的拓扑结构如下图：

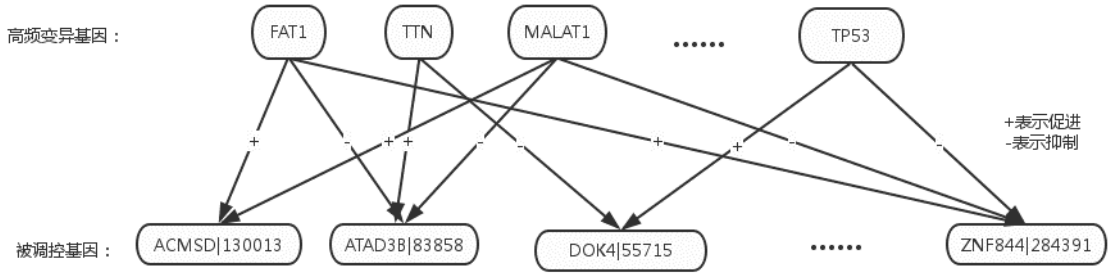


图 20: 调控矩阵的拓扑结构

将这114个基因作为一个1×114的列表T。对于每个高频基因，如该基因对于T中的某个基因为促进则记该位点为1，抑制则记该位点为-1，无影响则记为0，得到一个1×114的行向量记为$T1$。总共20个行向量按照高频基因排序合并为20×145的矩阵M，此矩阵代表了高频基因对于其它基因的调控情况。将每个人的高频基因变异情况作为一个行向量，记为P，其中，此人的20个高频基因中对应的基因变异次数，没变异则记为0. 将这个行向量与矩阵相乘（$P \times M$）得到此人的基因交互状态向量记为$M1$。基因调控矩阵的示意图如图21：

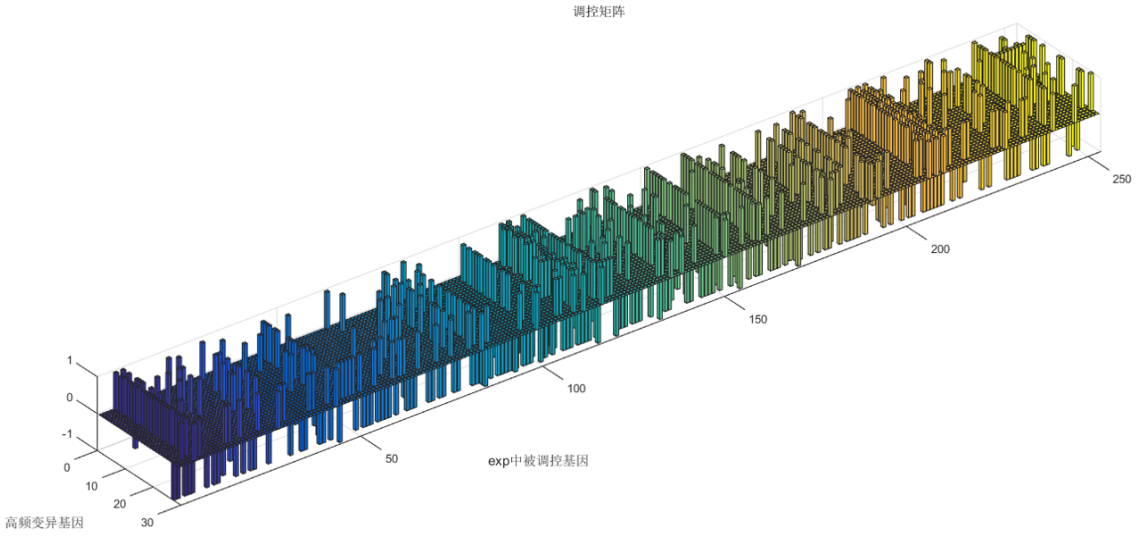


图 21: 调控矩阵3D示意图

6.3.5 结果分析

在基于调控矩阵的基础上，我们将变异基因数据与调控矩阵相结合，得到每个个体的基因交互状态向量M1。在本论文中，我们认为M1状态中每个stage的信息可能增强，同时也可能会使其模式更加紊乱。

我们以每个个体M1状态为输入，以样本个体患病程度为标签，仍然使用复杂树、线性核支持向量机、高斯核支持向量机进行分类，其结果要稍好于模型一和模型二的结果。

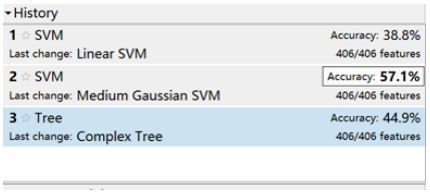


图 22: 组别0模型三交叉验证的结果

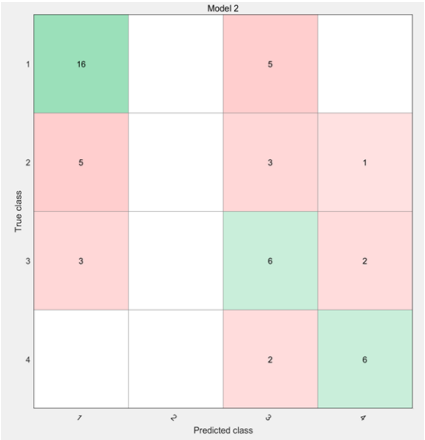


图 23: 组别0中方法1的混淆矩阵

图 24: 组别0模型三交叉验证的结果

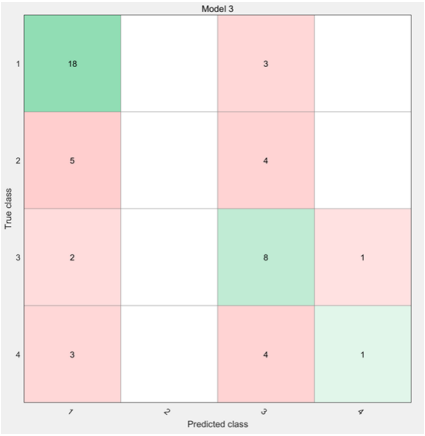
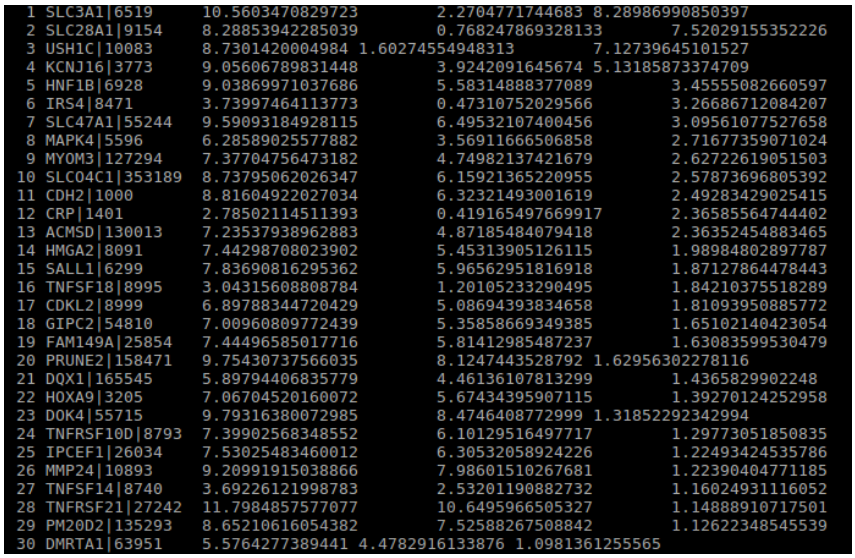


图 25: 组别0中方法1的混淆矩阵

6.3.6 实例说明

从29个高频基因中，我们选择了胰岛素信号通路中非常关键的蛋白激酶的PI3K来举例说明，在文件中的基因名为PI3K，下图是PIK3CA对部分基因表达量有重要影响的截图。



1	SLC3A1|6519	10.5603470829723	2.2704771744683	8.28986990850397
2	SLC28A1|9154	8.28853942285039	0.768247869328133	7.52029155352226
3	USH1C|10083	8.7301420004984	1.60274554948313	7.12739645101527
4	KCNJ16|3773	9.05606789831448	3.9242091645674	5.13185873374709
5	HNF1B|6928	9.03869971037686	5.58314888377089	3.45555082660597
6	IRS4|8471	3.73997464113773	0.47310752029566	3.26686712084207
7	SLC47A1|55244	9.59093184928115	6.49532107400456	3.09561077527658
8	MAPK4|5596	6.28589025577882	3.56911666506858	2.71677359071024
9	MYOM3|127294	7.37704756473182	4.74982137421679	2.62722619051503
10	SLCO4C1|353189	8.73795062026347	6.15921365220955	2.57873696805392
11	CDH2|1000	8.81604922027034	6.32321493001619	2.49283429025415
12	CRP|1401	2.78502114511393	0.419165497669917	2.36585564744402
13	ACMSD|130013	7.23537938962883	4.87185484079418	2.36352454883465
14	HMGA2|8091	7.44298708023902	5.45313905126115	1.98984802897787
15	SALL1|6299	7.83690816295362	5.96562951816918	1.87127864478443
16	TNFSF18|8995	3.04315608808784	1.20105233290495	1.84210375518289
17	CDKL2|8999	6.89788344720429	5.08694393834658	1.81093950885772
18	GIPC2|54810	7.00960809772439	5.35858669349385	1.65102140423054
19	FAM149A|25854	7.44496585017716	5.81412985487237	1.63083599530479
20	PRUNE2|158471	9.75430737566035	8.1247443528792	1.62956302278116
21	DQX1|165545	5.89794406835779	4.46136107813299	1.4365829902248
22	HOXA9|3205	7.06704520160072	5.67434395907115	1.39270124252958
23	DOK4|55715	9.79316380072985	8.4746408772999	1.31852292342994
24	TNFRSF10D|8793	7.39902568348552	6.10129516497717	1.29773051850835
25	IPCEF1|26034	7.53025483460012	6.30532058924226	1.22493424535786
26	MMP24|10893	9.20991915038866	7.98601510267681	1.22390404771185
27	TNFSF14|8740	3.69226121998783	2.53201190882732	1.16024931116052
28	TNFRSF21|27242	11.7984857577077	10.6495966505327	1.14888910717501
29	PM20D2|135293	8.65210616054382	7.52588267508842	1.12622348545539
30	DMRTA1|63951	5.5764277389441	4.4782916133876	1.0981361255565

图 26: PIK3对部分基因表达量的影响

从前面30个基因我们看到，在胰岛素信号通路中的基因如IRS4，MAPK4(ERK)，IPCEF1(PIP3),还有肿瘤坏死因子家族等与代谢综合征相关的基因都受到了PIK3CA变异的影响，PIK3CA的变异使得他们的表达量都上升。肿瘤坏死因子诱导脂肪细胞的凋亡，通过抑制IRS1通路从而促进胰岛素抵抗。第六位为IRS4基因，IRS4的主要作用是抑制IRS1和IRS2的功能，和肿瘤坏死因子一样是抑制IRS1通路，而IRS1/2的酪氨酸磷酸化激活磷脂酰肌醇3激酶（PI3K）和Ras-MAPK信号通路，二者分别介导胰岛素的能量代谢和生长调节反应。如图27。

对于第8位的MAPK4，其编码的蛋白ERK和之前的肿瘤坏死因子，IRS4表现为协同作用，抑制IRS1通路。第12位为CRP，CRP为C反应蛋白，炎症刺激肝脏合成的急性蛋白，其水平的升高和腰围的增加，胰岛素抗性，和高血糖相关联。第14位的HGMA2在脂肪形成和间质分化中起作用，在大量的癌症病人中发现高水平的HGMA2，特别是和粘液样脂肪肉瘤病人相关。虽然通过此方法，找到的基因并不一定都是与高频变异基因相关的基因。但是生物体是一个复杂的回路，不排除可能通过DNA与DNA，DNA与蛋白质，蛋白质与蛋白质相互作用间接影响。我们可以通过此方法可以知道，一个基因的变异会使得很多的基因受直接或间接受到影响。此外，还可以通过此方法，寻找一些目前还没有研究透彻的基因通路。

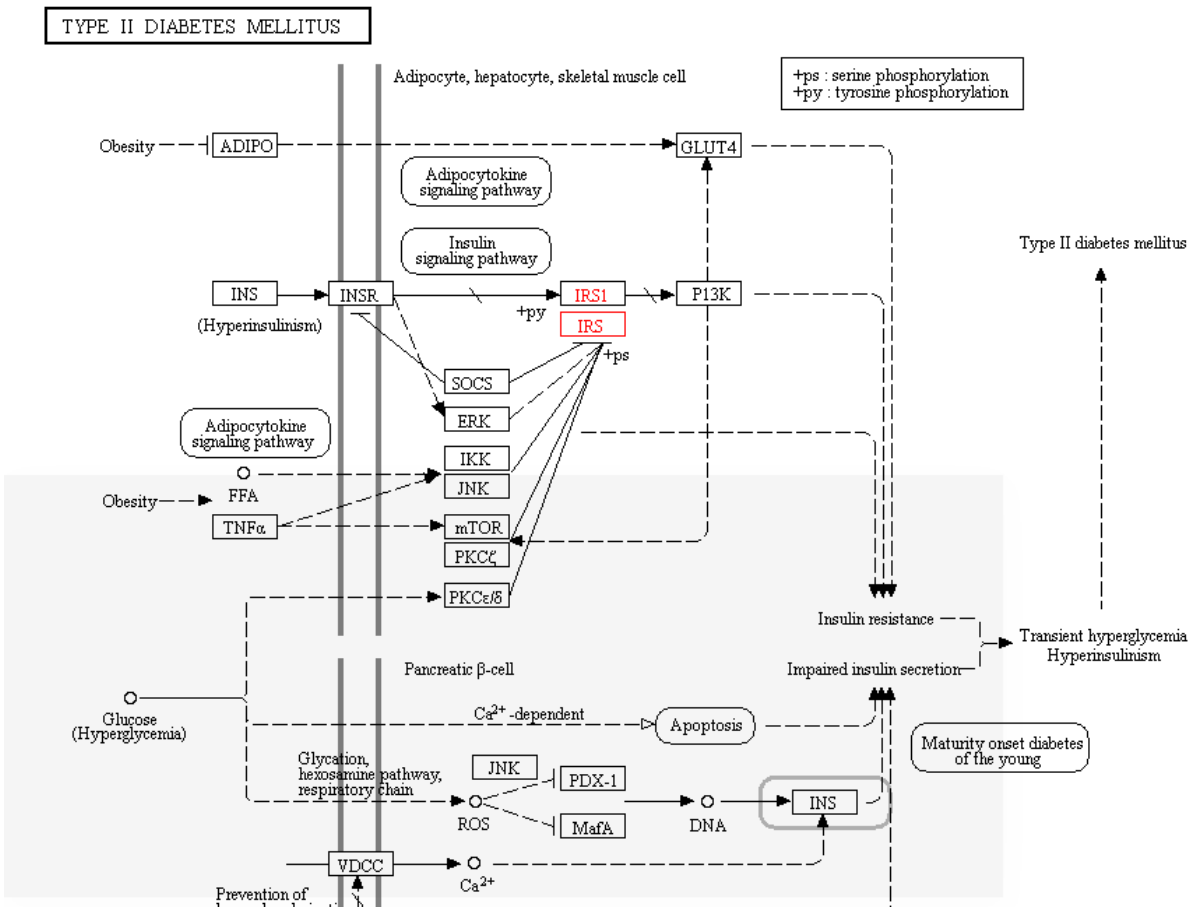


图 27: 二型糖尿病基因通路

6.3.7 寻找相关基因

在生物过程中，有些基因的有具有相似的功能，而且共同起作用，我们对调控矩阵中第一列中的29个基因做两两分析，如果两个基因在254或312的个基因列表中有超过10%的作用相同，即对超过10%的基因有相同的调控作用，同时为1或者-1。我们认为这两个基因可能有协同作用，赋予这两个基因值为1，意为有可能协同,若没有，则为0。这样我们就得到29*29的只有0和1的协同矩阵。如下图：

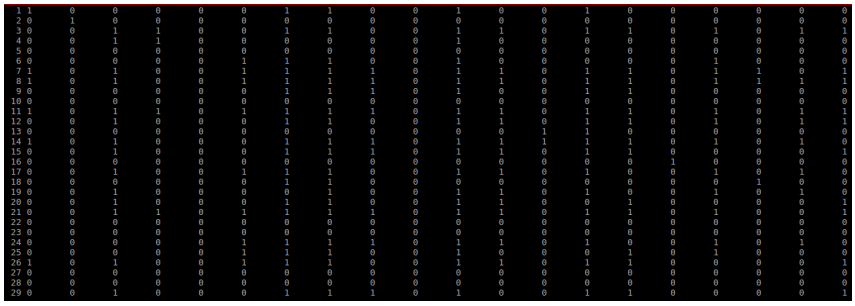


图 28: 协同矩阵

通过对比NCBI和EBI数据库，寻找具有协同作用的基因是否在生物意义上具有共同作用。经发现APOB基因和LRP1B基因，VHL基因和CSMD1基因有共同的作用。因为APOB基因编码LDL低密度脂蛋白表达，而LRP1B基因编码LDL受体蛋白表达，当LDL含量增加，受体蛋白含量也相应增加，两者共同完成吸收过多的LDL的作用。而VHL基因

和CSMD1基因共同抑制基因的表达，因为VHL基因和CSMD1均为肿瘤抑制因子，对于癌症增殖有相同的抑制作用。

6.4 模型四：基于“变异基因作用网络”的2-范数风险预测模型

6.4.1 模型说明

在之前的三个预测模型中，在已知个体患病的情况下，我们利用分类模型对其患病状态进行预测，但以上模型均不能反映未知个体处于各个阶段的的概率。基于此，我们提出一个基于“变异基因作用网络”的2-范数风险预测模型。由于我们未获得健康个体样本，我们依旧在该个体患病前提下，预测其处于某阶段的风险，进一步根据所有变量中对2-范数贡献最大的属性找出主要影响因素。

理想情况下，每个Stage（患病程度）所对应的基因交互状态向量M1在高维空间中应该组内相聚，组间相离（如图29）；但根据预测模型三，由于预测精度较低，说明Stage组间分离效果不是很明显，呈现出一定的交叉区域（如图30）。

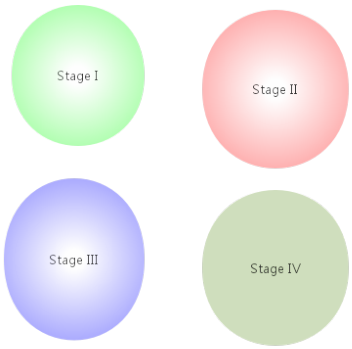


图 29: 高维空间中的组间分离

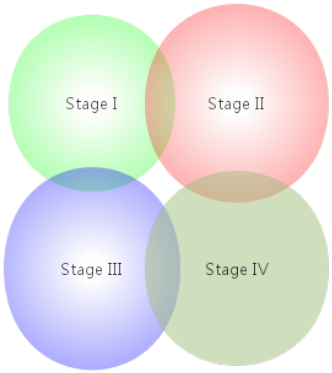


图 30: 高维空间中的组间交叉

为了简化计算，我们暂时忽略不同Stage之间的交叉区域，从而我们可以通过样本计算每个Stage的中心。当获得新未知个体数据时，我们通过其与四个中心点的距离去量化患病程度风险。

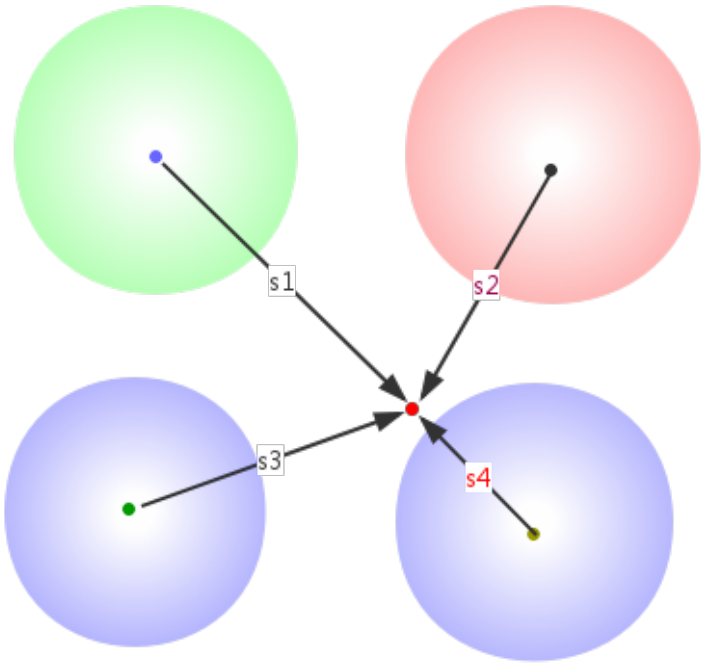


图 31: 患病程度风险量化示意图

例如，处于Stage I的概率P_1可由下述式子计算：

$$P_1 = \frac{\frac{1}{s1}}{\frac{1}{s1}+\frac{1}{s2}+\frac{1}{s3}+\frac{1}{s4}}c \tag{7}$$

6.4.2 模型应用

根据前文中所得基因变异状态向量M1以及M1临近的中心点，比较其在不同维度(基因)的范数成分，并由此排序，取成分最小的几个维度，因此获得影响导致此状态M1患病的主要基因；基于此，再根据前文所得基因通路与数据库信息找出主要基因对应的代谢综合症症状，从而确定导致此病患表现型因素。

表7是从组别0的数据中提取的病患（患病程度为unknown）的预测分析

表 7: 对病患的预测分析

	某M1	StageI 中心	范数成分
HPN|3249	1	1.066667	0.004444
MGAT4A|11320	1	0.9	0.01
SLC47A1|55244	5	5.133333	0.017778
ECHDC1|55862	0	-0.16667	0.027778
SEMA3F|6405	0	0.166667	0.027778
LRIG3|121227	0	-0.23333	0.054444
AMOT|154796	2	1.766667	0.054444
CDH2|1000	2	2.266667	0.071111
PVRL1|5818	-1	-1.26667	0.071111
ECM2|1842	0	0.266667	0.071111
ERGIC1|57222	0	0.266667	0.071111
KIF4B|285643	-1	-0.73333	0.071111
HMGA1|3159	-2	-1.66667	0.111111
MFSD2B|388931	-2	-1.66667	0.111111
CDCA5|113130	0	-0.33333	0.111111
CRYL1|51084	0	0.333333	0.111111
CYFIP2|26999	0	0.333333	0.111111
MARK1|4139	0	-0.33333	0.111111
ORC1L|4998	0	-0.33333	0.111111
DFNB31|25861	0	0.366667	0.134444

患病程度风险预测：通过“变异基因作用网络”的2-范数风险预测模型我们得到一个基因列表，我们通过预测得知，该病人患Stage1的风险较大。

基因作用网络：从本预测模型得出的与Stage1密切相关的基因当中，有至少一半以上的基因都与代谢综合征相关，在NCBI、EBI数据库中搜寻范数成分小于0.08的相关基因（共12个基因），最终找出有8个与Stage相关的基因，如图32。

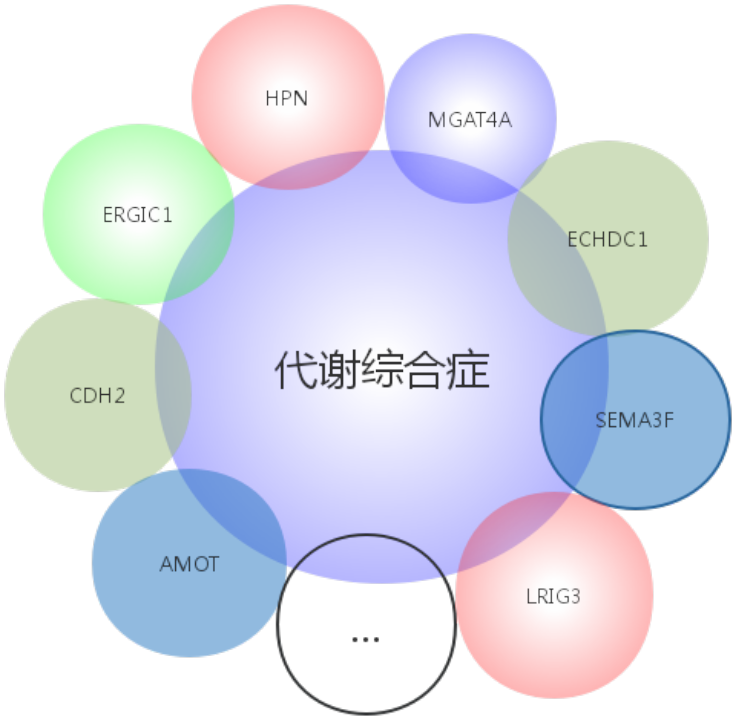


图 32: 与代谢综合征相关的基因

例如，范数成分最低的HPN基因与胰腺癌密切相关，该基因表达翻译丝氨酸蛋白酶，该蛋白对细胞增殖和细胞形态维持作用有重要作用；MGAT4T基因表达翻译糖基转移酶，该酶参与葡糖糖转运蛋白GLUT2的糖基化过程，对于葡萄糖运输有重要作用；此外HPN、AMOT、CDH2等基因均与前列腺癌变相关。

综上，通过图中基因信息得知，该病患的代谢综合征风险与葡萄糖代谢紊乱和胰腺功能损伤前列腺细胞异常等因素密切相关。

6.4.3 结果分析

- 由于我们未获得表观基因组数据、蛋白质组数据、代谢组数据，并且在转录组数据、基因组数据中并无健康个体数据，我们无法得知健康个体的基因变异以及表达情况，也就无法获得基因交互状态向量M1，从而无法预测未知个体患病与不患病风险。但在本风险模型中，添加健康群体信息，即可预测个体患病与不患病风险；

- 由于样本个体太少，每个Stage群体聚类效果交叉，无法准确地得到Stage中心数据，因此可能导致潜在的误差；若提高样本数量，获得更准确的中心数据，可以提高预测的准确率；

- 在进一步完善模型时，我们可以进一步将交叉区域与非交叉区域单独分析，更加准确地描述患病状态；

- 在划分群组时，忽略了每个Stage中A型、B型的差别，在进一步研究中，可以将群组细分为12个群组，从而获得更加精确的风险预测。

7 模型的评价、改进及推广

7.1 模型的评价

7.1.1 模型优点

- 对于不同的数据，可以建立不同的模型，进行对不同模型的进行比较。
- 能够用实际科学研究发现来用于佐证模型有效。

7.1.2 模型缺点

- 没有充分利用给定的数据，对于疾病的划分不够细致。
- 样本数量不够多，没有直接用生命科学的研究成果来建模。
- 所用方法普适性高，便于推广。

7.2 模型的改进

- 在变异文件中，本文中的模型只用到了变异所在基因的信息。但是文件中还包含位点坐标，变异碱基，变异类型，可以根据这些信息进行更为细致的讨论、构建更为复杂的模型。

7.3 模型的推广

- 本文最后的模型是2-范数风险预测模型，使用不同的范数可能会获得更多样的结果。
- 本论文中的绝大多数模型都是基于数据而建立的，例如根据基因表达量和基因变异，其基本思想和方法不仅仅可以用于代谢综合征，可以应用于更多的疾病。

参考文献

[1] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(01):32-42.

[2] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(01):2-10.

[3] Matej Oresic, Antonio Vidal-Puig. A Systems Biology Approach to Study Metabolic Syndrome[M]. 2014.

[4] Kaur J. A comprehensive review on metabolic syndrome [J]. Cardiology research and practice, 2014, 2014.

[5] ． Kahn R, Buse J, Ferrannini E, et al. The metabolic syndrome: time for a critical appraisal Joint statement from the American Diabetes Association and the European Association for the Study of Diabetes[J]. Diabetes care, 2005, 28(9): 2289-2304.

[6] Alberti K G M M, Zimmet P Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation[J]. Diabetic medicine, 1998, 15(7): 539-553.

[7] Wang G. Raison d'etre of insulin resistance: the adjustable threshold hypothesis[J]. Journal of The Royal Society Interface, 2014, 11(101): 20140892.

[8] Wang G. Singularity analysis of the AKT signaling pathway reveals connections between cancer and metabolic diseases[J]. Physical biology, 2010, 7(4): 046015.

[9] Neel J V. Diabetes mellitus: a “thrifty” genotype rendered detrimental by “progress” ? [J]. American journal of human genetics, 1962, 14(4): 353

[10] Hales C N, Barker D J P. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis[J]. Diabetologia, 1992, 35(7): 595-601.

附 录

A 代谢综合征相关症状

A.1 肥胖症(Obesity)

肥胖症是由于人体摄入过多脂肪类物质，经过长期积累无法代谢造成的。肥胖症会增加胃、肝脏和脂肪组织的氧化代谢压力，最终引起胰岛素抵抗的血脂障碍和高血压等代谢综合征症状。低高密度脂蛋白（HDL）水平是肥胖症的重要判断依据（临床标准：男性<40mg/dL，女性<50mg/dL），中心性肥胖是判断代谢综合征重要的标准之一（临床标准：男性腰围≥102厘米，女性腰围≥88厘米），由肥胖引起代谢综合征的过程如图33所示。

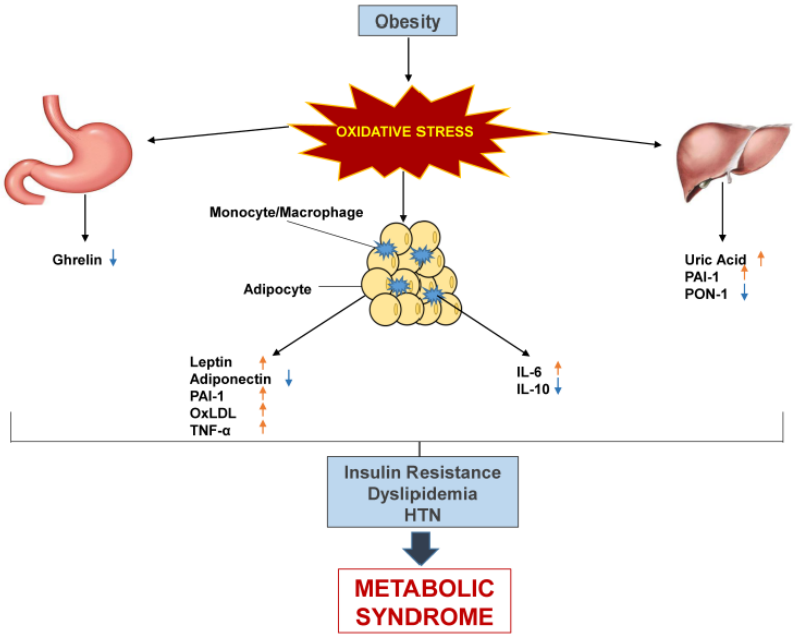


图 33: 肥胖症(Obesity)

A.2 II 型糖尿病(Type 2 diabetes)

II 型糖尿病大多由肥胖、衰老以及不良的生活习惯造成的各个组织细胞对胰岛素的抵抗作用。组织细胞对织细胞胰岛素抵抗将引起肝脏产生过多葡萄糖代谢产物、肌肉组织减少葡萄糖摄入以及脂肪组织向血液分泌过多的自由脂肪酸分子(Free Fatty Acid)。在早期阶段，beta细胞将分泌大量胰岛素补偿胰岛素抵抗作用，即到时血液胰岛素水平极高；当补偿作用失效，糖分无法被组织细胞摄入，从而造成高血糖，同时高血糖和高脂肪酸水平将引起严重的细胞损伤及病变。90%的糖尿病患者均为 II 型糖尿病患者。

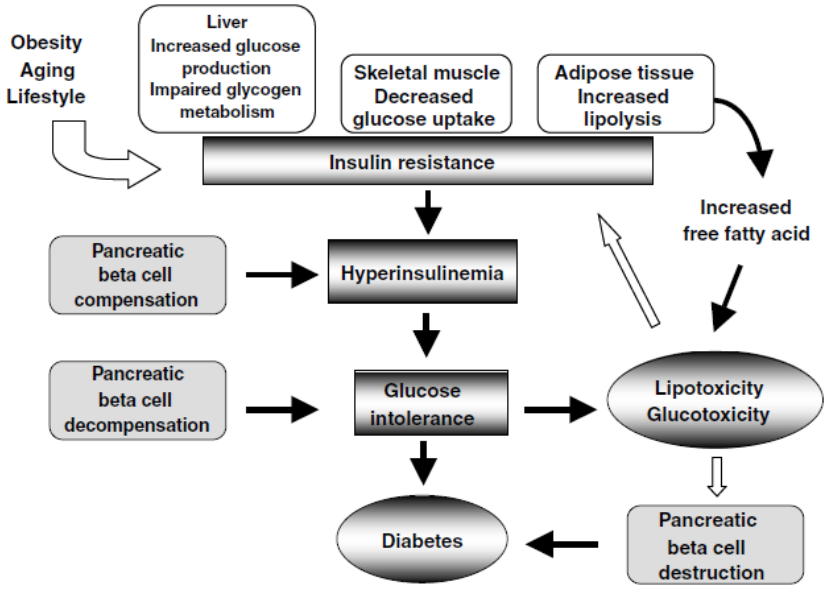


图 34: II 型糖尿病(Type 2 diabetes)

A.3 高血压(Hypertension)、高血糖(Hyperglucomia)、高血脂(Hyperlipidemia)

血液中的血压、血糖、血脂值是临床判断代谢综合征的主要依据。高血压主要由于血管壁的胆固醇沉积导致的血管堵塞血压增高，高血糖主要由于糖尿病引起的血糖含量增高，高血脂是由于脂肪代谢异常导致血液脂肪类含量过高。

A.4 动脉粥样硬化(Atherosclerosis)

动脉粥样硬化是由于脂肪代谢异常导致大量白细胞在血管壁中聚集沉积血液中过量的胆固醇和甘油三酯造成血管堵塞的一种症状，其原理如图所示。

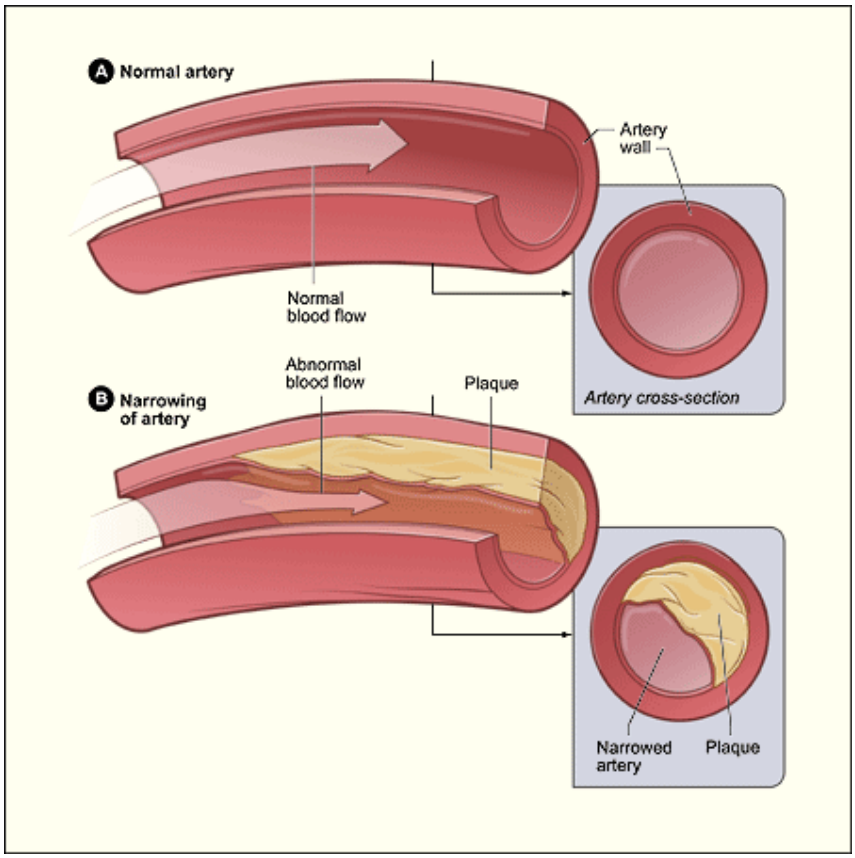


图 35: Ⅱ型糖尿病(Type 2 diabetes)

A.5 脂肪肝(Fatty liver)

糖尿病、高血压、肥胖、脂肪代谢紊乱和酒精过量摄入均是导致脂肪肝的重要因素。脂肪组织在肝脏大量富集将严重影响肝脏的正常功能，从而引起一系列的疾病。

B 数据处理代码

B.1

```
#!/usr/bin/perl –w

open LIST, '/media/jayden/My␣Files/format/gene.list';

while(<LIST>){
chomp;
my @data = split /\t/;
open VAR, '/media/jayden/My␣Files/format/sample.var.0';
my @positive;
while(<VAR>){
chomp;
my @var = split /\t/;
if ($var[0] eq $data[0] && $var[11] ne 54){
push @positive, $var[11];
}
}
```

```
close VAR;

open EXP, '/media/jayden/My␣Files/format/exp.important.0';
my $probe = <EXP>;
chomp $probe;
my @Probe = split /\t/,$probe;
my @Po;
my @Ne;

for( my $i=0;$i<= $#Probe;$i++){
if (grep{$Probe[$i] eq $_} @positive){
push @Po, $i;
}
else{
push @Ne, $i;
}
}

open Pos, '>/media/jayden/My␣Files/format/exp.0.pos.'.$data[0];
open Neg, '>/media/jayden/My␣Files/format/exp.0.neg.'.$data[0];
print Pos "$Probe[0]\t";
foreach my $item (@Po){
print Pos "$Probe[$item]\t";
}
foreach my $item (@Ne){
print Neg "$Probe[$item]\t";
}

while(<EXP>){
chomp;
my @exp = split /\t/;
print Pos "\n$exp[0]\t";
print Neg "\n";
foreach my $item (@Po){
print Pos "$exp[$item]\t";
}
foreach my $item(@Ne){
print Neg "$exp[$item]\t";
}
}
close EXP;
}
```

B.2

```
#!/usr/bin/perl –w

open EXP, '/media/jayden/My␣Files/format/sample.exp.1';
open INFO, '/media/jayden/My␣Files/format/sample.inf.txt';
```

```
my @FSI;
my @FSIA;
my @FSIB;
my @FSII;
my @FSIIA;
my @FSIIB;
my @FSIII;
my @FSIIIA;
my @FSIIIB;
my @FSIV;
my @FSIVA;
my @unknow;

while(<INFO>){
chomp;
my @data = split /\t/;
if ($data[3] eq '1'){
if ($data[2] eq 'Stage I'){
push @FSI, $data[0];
}
elsif($data[2] eq 'Stage IA'){
push @FSIA, $data[0];
}
elsif($data[2] eq 'Stage IB'){
push @FSIB, $data[0];
}
elsif($data[2] eq 'Stage II'){
push @FSII, $data[0];
}
elsif($data[2] eq 'Stage IIA'){
push @FSIIA, $data[0];
}
elsif($data[2] eq 'Stage IIB'){
push @FSIIB, $data[0];
}
elsif($data[2] eq 'Stage III'){
push @FSIII, $data[0];
}
elsif($data[2] eq 'Stage IIIA'){
push @FSIIIA, $data[0];
}
elsif($data[2] eq 'Stage IIIB'){
push @FSIIIB, $data[0];
}
elsif($data[2] eq 'Stage IV'){
push @FSIV, $data[0];
```

```
}
elsif($data[2] eq 'Stage IVA'){
push @FSIVA, $data[0];
}
elsif($data[2] eq '.'){
push @unknow, $data[0];
}

}

}

my $probe = <EXP>;
chomp $probe;
close EXP;

my @PI;
my @PIA;
my @PIB;
my @PII;
my @PIIA;
my @PIIB;
my @PIII;
my @PIIIA;
my @PIIIB;
my @PIV;
my @PIVA;
my @Punknow;

my @Probe = split /\t/, $probe;


for( my $i=0;$i<= $#Probe;$i++){
if (grep{$Probe[$i] eq $_} @FSI){
push @PI, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIA){
push @PIA, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIB){
push @PIB, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSII){
push @PII, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIIA){
push @PIIA, $i;
}
```

```
elsif(grep{$Probe[$i] eq $_} @FSIIB){
push @PIIB, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIII){
push @PIII, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIIIA){
push @PIIIA, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIIIB){
push @PIIIB, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIV){
push @PIV, $i;
}
elsif(grep{$Probe[$i] eq $_} @FSIVA){
push @PIVA, $i;
}
elsif(grep{$Probe[$i] eq $_} @unknow){
push @Punknow, $i;
}
}
```

```
open EXP, '/media/jayden/My␣Files/format/sample.exp.1';
open SI, '>/media/jayden/My␣Files/format/exp.I.1';
open SIA, '>/media/jayden/My␣Files/format/exp.IA.1';
open SIB, '>/media/jayden/My␣Files/format/exp.IB.1';
open SII, '>/media/jayden/My␣Files/format/exp.II.1';
open SIIA, '>/media/jayden/My␣Files/format/exp.IIA.1';
open SIIB, '>/media/jayden/My␣Files/format/exp.IIB.1';
open SIII, '>/media/jayden/My␣Files/format/exp.III.1';
open SIIIA, '>/media/jayden/My␣Files/format/exp.IIIA.1';
open SIIIB, '>/media/jayden/My␣Files/format/exp.IIIB.1';
open SIV, '>/media/jayden/My␣Files/format/exp.IV.1';
open SIVA, '>/media/jayden/My␣Files/format/exp.IVA.1';
open UNKNOWN, '>/media/jayden/My␣Files/format/exp.unknown.1';


while(<EXP>){
my @exp  = split /\t/;

print SI "$exp[0]\t";
print SIA "$exp[0]\t";
print SIB "$exp[0]\t";
print SII "$exp[0]\t";
print SIIA "$exp[0]\t";
```

```
print SIIB "$exp[0]\t";
print SIII "$exp[0]\t";
print SIIIA "$exp[0]\t";
print SIIIB "$exp[0]\t";
print SIV "$exp[0]\t";
print SIVA "$exp[0]\t";
print UNKNOWN "$exp[0]\t";

foreach my $item (@PI){
print SI "$exp[$item]\t";
}

foreach my $item (@PIA){
print SIA "$exp[$item]\t";
}

foreach my $item (@PIB){
print SIB "$exp[$item]\t";
}

foreach my $item (@PII){
print SII "$exp[$item]\t";
}

foreach my $item (@PIIA){
print SIIA "$exp[$item]\t";
}

foreach my $item (@PIIB){
print SIIB "$exp[$item]\t";
}

foreach my $item (@PIII){
print SIII "$exp[$item]\t";
}

foreach my $item (@PIIIA){
print SIIIA "$exp[$item]\t";
}

foreach my $item (@PIIIB){
print SIIIB "$exp[$item]\t";
}

foreach my $item (@PIV){
print SIV "$exp[$item]\t";
}
```

```
foreach my $item (@PIVA){
print SIVA "$exp[$item]\t";
}
foreach my $item(@Punknow){
print UNKNOWN "$exp[$item]\t";
}



print SI "\n";
print SIA "\n";
print SIB "\n";
print SII "\n";
print SIIA "\n";
print SIIB "\n";
print SIII "\n";
print SIIIA "\n";
print SIIIB "\n";
print SIV "\n";
print SIVA "\n";
print UNKNOWN "\n"
}

print "@FSIIIA\n@PIIIA\n";
```

B.3

```
#!/usr/bin/perl -w

open FILE, '/media/jayden/My_Files/format/sample.var.1';
open FRE, '>/media/jayden/My_Files/format/var.fre.1';

<FILE>;
my %hash;

while(<FILE>){
chomp;
my @data = split /\t/;
$hash{$data[0]}++;
}

foreach my $keys (sort{$hash{$b}<=>$hash{$a}}keys %hash){
print FRE "$keys\t$hash{$keys}\n";
}
```

B.4

```
open LIST, '/media/jayden/My_Files/format/gene.list';

while(<LIST>){
```

```
chomp;
my @data = split /\t/;
open POS, '/media/jayden/My␣Files/format/exp.0.pos.'.$data[0];
open NEG, '/media/jayden/My␣Files/format/exp.0.neg.'.$data[0];
open LOG, '>/media/jayden/My␣Files/format/exp.0.diff.'.$data
    [0];

<NEG>;
<POS>;

my %hash;
while(<NEG>){
chomp;
my @neg = split /\t/;
my $gene = shift @neg;

my $line = <POS>;
chomp $line;
my @pos = split /\t/, $line;
shift @pos;

my $pos = &Average(@pos);
my $neg = &Average(@neg);

my $diff = $neg - $pos;
my @data = ($neg,$pos,$diff);
my $data_ref = \@data;
$hash{$gene} = $data_ref;

}



foreach my $keys (sort{$hash{$b}->[2] <=> $hash{$a}->[2]} keys
    %hash){
my $data = $hash{$keys};
print LOG "$keys\t$data->[0]\t$data->[1]\t$data->[2]\n";
}

}




sub Average{
my $sum = 0;
foreach my $item (@_){
$sum += $item;
```

```
}
my $average = $sum/($#_ +1);
if ($average == 0){
return -1;
}
my $log = log($average)/log(2);
}
```

B.5

```
#!/usr/bin/perl -w

open LIST, '/media/jayden/My_Files/format/gene.list';

open MAT, '>/media/jayden/My_Files/format/matrix.0';
my %matrix0;
my %matrix1;

while(<LIST>){
chomp;
my @data = split /\t/;
open LOG0, '/media/jayden/My_Files/format/exp.0.diff.'.$data
    [0];
open LOG1, '/media/jayden/My_Files/format/exp.1.diff.'.$data
    [0];
my %hash0;
while(<LOG0>){
chomp;
my @gene0 = split /\t/;
if ($gene0[3]>1){
$hash0{$gene0[0]} = 1;
}
elsif($gene0[3]<-1){
$hash0{$gene0[0]} = -1;
}
else{
$hash0{$gene0[0]} = 0;
}

}

my %hash1;
while(<LOG1>){
chomp;
my @gene1 = split /\t/;
if ($gene1[3]>1){
$hash1{$gene1[0]} = 1;
}
elsif($gene1[3]<-1){
```

```
$hash1{$gene1[0]} = -1;
}
else{
$hash1{$gene1[0]} = 0;

}

}



$matrix0{$data[0]} = \%hash0;
$matrix1{$data[0]} = \%hash1;
}


my $hash0_ref = $matrix0{"TTN"};
foreach my $key0 (keys %$hash0_ref){
my $i =0 ;
foreach my $key (keys %matrix0){
my $ref = $matrix0{$key};
$i += abs($ref->{$key0});
}

foreach my $key (keys %matrix0){
if ($i==0){
my $ref = $matrix0{$key};
delete $ref->{$key0};

}
}

}

print MAT "\t";
foreach my $key0 (sort keys %$hash0_ref){
print MAT "$key0\t";
}
print MAT "\n";

foreach my $key (sort keys %matrix0){
print MAT "$key\t";
my $hash0_ref = $matrix0{$key};
foreach my $key0 (sort keys %$hash0_ref){
print MAT "$hash0_ref->{$key0}\t";

}
print MAT "\n";
```

```
}
```

B.6

```
#!/usr/bin/perl –w

open FRE1, '/media/jayden/My␣Files/format/var.per.1';
open FRE0, '/media/jayden/My␣Files/format/var.per.0';
open LIST, '>>/media/jayden/My␣Files/format/gene.list2';

my %hash;

while(<FRE0>){
chomp;
my @data = split /\t/;
$hash{$data[0]} = $data[2];
}

my $i = 0;
my %hash2;
while(<FRE1>){
chomp;
my @data = split /\t/;
$hash2{$data[0]} = $data[2];
$i++;
last if ($i>10);
}


my %hash3;
foreach my $key (keys %hash2){
if ($hash2{$key} >0.123 && $hash{$key}>0.05){
$hash3{$key} = $hash{$key} + $hash2{$key};
}

}

foreach my $key (sort{$hash3{$b}<=> $hash3{$a}}keys %hash3){
print LIST "$key\t$hash{$key}\t$hash2{$key}\t$hash3{$key}\n";
}
```

B.7

```
#!/usr/bin/perl –w

open FILE, '/media/jayden/My␣Files/format/sample.var.1';
open FILE2, '/media/jayden/My␣Files/format/sample.var.1';
open INFO, '/media/jayden/My␣Files/format/sample.inf.txt';

open TAG, '>/media/jayden/My␣Files/format/var.tag.1';
```

```
my %info;
while(<INFO>){
chomp;
my @data = split /\t/;
$info{$data[0]} = $data[2];
}

open MERGE, '/media/jayden/My␣Files/format/matrix.0';
<MERGE>;
my @merge;
while(<MERGE>){
chomp;
my @data = split /\t/;
push @merge, $data[0];
}

<FILE>;
<FILE>;
<FILE2>;

my %hash;
foreach my $item (@merge){
$hash{$item} = 0;
}

while(<FILE>){
chomp;
my @data = split /\t/;
my $line = <FILE2>;
chomp $line;
my @data2 = split /\t/, $line;
if ($data[11] eq $data2[11]){
foreach my $item (@merge){
my @head = split /\_/, $item;
if ($data[0] =~ /^$head[0]/){
$hash{$item}++;
}
}
}
else{
print TAG "$data2[11]";
foreach my $keys (sort (keys %hash)){
print TAG "\t$keys";
}
print TAG "\n";
print TAG "$info{$data2[11]}";
foreach my $keys (sort (keys %hash)){
```

```
print TAG "\t$hash{$keys}";
}
print TAG "\n";
undef %hash;
foreach my $item (@merge){
$hash{$item} = 0;
}
}
}
```