昆明学院学报　2019，41（6）：106～113　ISSN 1674－5639　CN 53－1211/G4
Journal of Kunming University

阿尔茨海默病患病风险评估模型研究

余彩裙，裴晶晶
（云南民族大学 数学与计算机科学学院，云南 昆明 650500）

摘要：随着人口老龄化日趋加重，我国患阿尔兹海默病的人数越来越多．为预测个体患阿尔兹海默病的风险，采用基于高通量测序技术下的样本数据，并根据 GWAS 和 GRASP 数据库中已知的 SNP 位点的 *OR* 值，对未报道的 SNP 位点的 *OR* 值利用多元线性回归进行预测，同时融合环境和遗传风险因素的 *OR* 值构建了阿尔兹海默病风险评估模型，计算出多个 SNP 位点导致患病的概率，从而评估患病风险．所得结果与临床结果比较表明，该模型具有一定的可靠性．

关键词：阿尔茨海默病；高通量测序；SNP 位点；风险因素；*OR* 值
中图分类号：R749.16；TP391.7　**文献标识码**：A　**文章编号**：1674－5639（2019）06－0106－08
DOI：10.14091/j.cnki.kmxyxb.2019.06.020

Study on Risk Prediction Model of Alzheimer's Disease

YU Caiqun，PEI Jingjing
（College of Mathematics and Computer Science，Yunnan Nationalities University，Kunming，Yunnan，China　650500）

Abstract：As the population ageing becomes more serious，the number of people suffering from Alzheimer's disease is increasing in our country．Therefore，in order to predict the risk of Alzheimer's disease，the sample data from high-throughput sequencing are used．The *OR* values of unreported SNPs are predicted using multiple linear regression，based on the *OR* values of known SNPs in the GWAS and GRASP databases．And a risk assessment model is constructed by integrating the *OR* values of environmental and genetic risk factors to calculate the probability level of multiple SNPs leading to disease assessing the risk of illness．Compared with clinical results，the model has certain reliability．

Key words：Alzheimer's disease；high-throughput sequencing；SNPs；risk factor；*OR* values

阿尔茨海默病（Alzheimer's disease，AD）又称老年痴呆病，是一种神经退行性疾病．阿尔茨海默病这一概念是由德国精神病与神经病理学家 Alois Alzheimer 在 1906 年最先提出．AD 致病机理非常复杂，发病原因尚未明确阐明，且每个人的发病机制大多都存在较大的差异．临床统计结果[1]显示，男性患阿尔茨海默病的频率（9%）低于女性的患病频率（17%）．目前，认为患该病的主要影响因素分别是环境因素和遗传因素．而在环境因素中，年龄是一个很重要的影响因素[1-2]．据统计[1]，在超过 65 岁（包括 65 岁）的大部分人群中，患有阿尔茨海默病的人的频率大约在 1% 左右，而在超过 95 岁（包括 95 岁）的大多数人群中患此病的大约有 40%～50%．另外，阿尔茨海默病具有较高的遗传性，全基因组相关性（GWAS）研究[3]显示，近 2/3 的遗传学变异定位在与阿尔茨海默病发病相关的候选基因上．

目前，阿尔茨海默病按年龄可分为：早发性阿尔茨海默病（年龄 <65 岁）和晚发性阿尔茨海默病（年龄≥65 岁）．其中，早发性阿尔茨海默病发病原因主要有 β-淀粉样蛋白前体（*app*）、早老素-1（*ps*-1）和早老素-2（*ps*-2）基因突变[4]，而晚发性阿尔茨海默病则往往是由多个基因的相互作用所导致的，且提高了个体对阿尔茨海默病的易感性．因此，由基因突变引发的阿尔兹海默病患病率较小，而大多数患者发病原因都是散发性的[2]．有数据显示，有 60%～80% 的阿尔茨海默病的发病风险源自于遗传．而载脂蛋白 E4（*ApoE*4）是

收稿日期：2019－06－22
作者简介：余彩裙（1996—），女，云南丽江人，硕士研究生，主要从事智能计算研究．

引起散发性阿尔兹海默病的最有显著意义的影响因素，携带有 *ApoE-ε*4 等位基因的人群比没有携带的人群可能会患有晚发性阿尔茨海默病的概率大约增加了 2 ~ 4 倍[5]．此外，在阿尔兹海默病家族中，早发性阿尔兹海默病的 *ps* 基因突变会导致 Aβ 的过度生成，Aβ 的代谢可能还与 *ApoE* 和其他新发现的晚发性阿尔茨海默病易感基因有关[2]．对晚发性阿尔茨海默病遗传相关性的研究主要有两种方式，一种是候选基因模式，而另一种是高通路全基因组相关性研究模式[6]．目前，后者已经取代前者，因该方法能够同时检测百万以上的单核苷酸多态性（SNP），使研究者能够更全面地对全基因组相关性分析进行研究[1]．

随着人口老龄化的日趋加重，我国阿尔茨海默病的发病率已呈逐年递增趋势．目前，全球痴呆患者人数有 4 680 万人，其中阿尔兹海默病患者占 50% ~75%．据统计，我国老年人群中阿尔兹海默病患病人数已经超过 600 万，预计到 2050 年患病人数将超过 2 000 万．而高通量测序技术为阿尔茨海默病的诊断和预测提供了可能．因此，本文采用基于高通量测序技术下的 Illumina 测序平台得到的 FASTQ 文件作为样本数据，筛选出与阿尔兹海默病强相关的 20 个基因，找到这 20 个基因对应的 SNP 位点，并根据 GWAS 和 GRASP 数据库找到 SNP 位点对应的 *OR* 值，对部分未报道 *OR* 值的 SNP 位点采用多元线性回归模型进行预测．最后将遗传因素和环境因素的 *OR* 值信息融入至阿尔茨海默病风险评估模型，计算出多个 SNP 位点导致个体患阿尔茨海默病的概率，于是根据概率水平评估患病风险．

1　样本数据准备及数据分析流程

1.1　高通量测序平台

高通量测序技术的出现不仅极大地降低了测序所需费用及时间，更有助于研究者全面、深入地分析基因组、转录组、蛋白质组等不同组学产生的数据，促使该技术成为实验室研究及临床检测最常规的手段之一[7]．目前使用最普遍的是以 HiSeq 系列为主的 Illumina 测序技术，该技术主要利用了一边合成一边测序的方法（Sequencing by synthesis，SBS）[8]．Illumina 测序基本流程总体可概括为 4 个方面：构建 DNA 测序文库，测序流动槽（Flowcell）制备，桥式 PCR 扩增与变性和上机测序[9]，如图 1 所示．

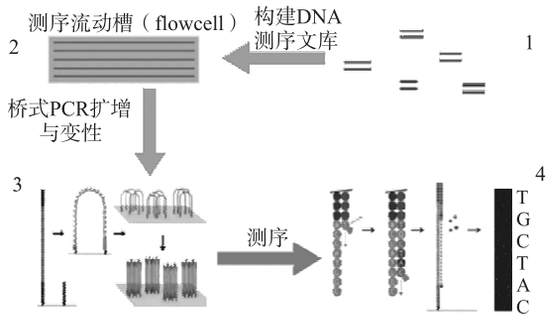


图1　Illumina测序流程

Illumina 平台测序结果以 FASTQ 格式[10]进行存储测序，是后续数据分析的基础．FASTQ 文件是一种包含生物序列和序列质量信息标准格式的文本文件，如图 2 所示．每 4 行组成一个 read，其中以@ 开头的@ SRR016186. 230160BI 为序列的名称，是每一条 read 的唯一标识符，包含序列标识及相关的信息描述；第 2 行是序列，由 A，T，C，G 组成；第 3 行以“ + ”开头，一般后面什么也不加；第 4 行是测得序列的质量分数[8]，与第 2 行的序列相对应，描述每个碱基的可靠程度．

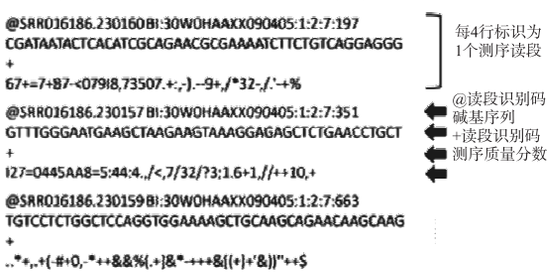


图2　FASTQ基本格式信息

本研究从临床获取了 7 份待检测者的干血片样本，样本信息如表 1 所示．样本信息包括待检测者的年龄、性别、运动频率、是否饮酒、血压、教育程度．该样本用 Illumina 高通量测序平台进行全外显子测序，得到原始的 FASTQ 文件的测序数据，用于后续的数据处理．

1.2　数据分析流程

目前，针对高通量测序技术出现了很多处理高通量数据的 python 软件包，比如 HTSeq，HISAT2，STAR，bcbio-nextgen[11] 等，其中 bcbio-nextgen 为高通量测序数据分析提供了更好的管道．考虑到效率和实用性，本研究使用 bcbio-nextgen[11] 框架对 7 份临床样本测序得到的 FASTQ 全外显子数据进行标准化分析．

表 1 样本信息

样本编号	年龄/岁	性别	运动频率	是否饮酒	血压/mmHg	教育程度
SN0001	62	男	偶尔	是	118	小学
SN0002	58	女	偶尔	否	141	高中
SN0003	34	女	经常	是	128	硕士
SN0004	55	男	经常	否	118	本科
SN0005	72	女	偶尔	否	156	小学
SN0006	45	女	偶尔	否	118	高中
SN0007	53	男	经常	否	102	初中

bcbio-nextgen 全外显子分析流程可概括为如下几个方面，比对 - 转换 - 去噪 - 找变异位点 - 注释，具体如下所述[9]：

1）使用 BWA 短序列软件将 reads 比对至参考基因组，以此确定 reads 在基因组上的位置；

2）使用 SAMTOOLS 软件将已比对定位好的短序列进行排序（1 ~ 22，X，Y，其他），并进行数据格式的转换；

3）使用 PICARD 软件过滤掉测序过程中产生的冗余信息和噪声；

4）使用 GATK 检测测序数据与参考基因组的差异位点；

5）使用 ANNOVAR[12] 对上述变异位点进行功能注释，得到一个易于理解的变异位点列表，进入后续疾病风险评估及预测分析.

2 构建风险评估模型

2.1 筛选与阿尔茨海默病强相关的基因

单个样本经数据预处理后与阿尔兹海默病相关的 17 444 个基因的权重分布如图 3 所示，图中显示了部分基因的名字，颜色越深，图中所占面积越大，表示该基因与阿尔兹海默病的相关分值越高. 利用 Phenolyzer[13]（http://phenolyzer.wglab.org/）工具预测并获得与阿尔兹海默病相关的基因有 17 444 个，对这 17 444 个基因按照从大到小的顺序进行排序，选取与该病强相关的前 20 个基因（见表 2）. 由于直接从 bcbio-nextgen 全外显子分析流程中获得的单个样本与阿尔兹海默病相关的 SNP 位点大约有 6 000 个左右，但是大多数 SNP 位点与阿尔兹海默病的相关性很弱，因此需要合理过滤掉那些意义不大的位点用于后续的数据分析. 首先，筛选出这 20 个基因对应的 SNP 位点. 其次，对筛选的得到的位点进行 ANNOVAR[12] 注释，把在 CLINVAR 数据库中显示为良性的 SNP 位点进行过滤. 最后，将上述经过筛选过滤后得到的 SNP 位点用于后续的风险评估分析模型.

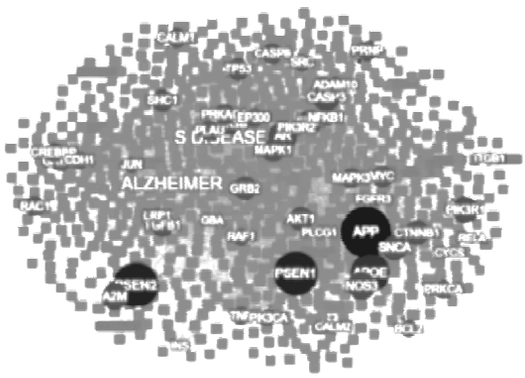


图3 阿尔茨海默病关联基因权重分布

2.2 阿尔茨海默病的风险因素

阿尔茨海默病的风险因素众多，包括环境因素、遗传因素、不健康的生活方式等，其中影响最大的是环境和遗传风险因素.

2.2.1 环境风险因素

1）年龄. 年龄越大，患病风险越高. 超过 65 岁时，患病的风险会每 5 年增加 1 倍.

2）教育程度. 教育程度低是国际上普遍认可的风险因素之一.

3）体育锻炼和社交活动. 低频率的体育锻炼和社交活动也会增加阿尔茨海默病的发病风险.

4）高血压. 有研究表明，高血压与阿尔兹海默病有一定的关联性，高血压可能会诱发阿尔兹海默病的发生.

表 3[14] 给出了不同环境风险因素的 odds ratio（*OR*）值，*OR* 值是一个比值，在统计患病风险因素时，如果某一因素的 *OR* 值等于 1，说明该因素对疾病不起任何作用；如果 *OR* 值大于 1，说明该因素是疾病的危险因素，含有该因素的患者患病风

险较高；反之，如果 *OR* 值小于 1，则说明该因素是保护因素，含有该因素的患者的患病风险较低. 在构建风险评估模型时，将不同因素的 *OR* 值融入模型中，得到患病概率，从而评估患病风险.

表 2　与阿尔茨海默病强相关的前 20 个基因

排序	基因	基因识别号	得分	状态	单倍剂量不足得分	基因不耐受得分
1	*APP*	351	1.000	种子基因	0.971	0.912
2	*PSEN*1	5 663	0.959	种子基因	0.955	0.696
3	*PSEN*2	5 664	0.941	种子基因	0.648	0.920
4	*APOE*	348	0.544	种子基因	0.242	0.000
5	*CASP*3	836	0.425	种子基因	0.933	0.726
6	*GSK*3*B*	2 932	0.348	种子基因	0.980	0.671
7	*CTNNB*1	1 499	0.315	预测基因	1.000	0.816
8	*TGFB*1	7 040	0.297	预测基因	0.752	0.551
9	*SHC*1	6 464	0.270	预测基因	0.710	0.596
10	*GRB*2	2 885	0.273	预测基因	1.000	0.555
11	*PRKACA*	5 566	0.268	预测基因	0.371	0.770
12	*MAPK*1	5 594	0.262	预测基因	0.999	0.738
13	*SNCA*	6 622	0.258	种子基因	0.698	0.571
14	*AKT*1	207	0.254	预测基因	0.997	0.917
15	*PIK*3*R*1	5 295	0.243	预测基因	0.894	0.866
16	*TP*53	7 157	0.237	预测基因	1.000	0.640
17	*A*2*M*	2	0.205	种子基因	0.264	0.392
18	*NOS*3	4 846	0.185	种子基因	0.209	0.930
19	*AR*	367	0.135	预测基因	0.990	0.827
20	*ETS*1	2 113	0.085	预测基因	0.951	0.612

表 3　阿尔茨海默病受不同环境因素影响的 *OR* 值

环境因素	条件	*OR* 值
年龄/岁	<47	1.000
	47 ~ 53	2.958
	>53	5.825
教育程度/年	<6	3.599
	6 ~ 9	2.485
	>9	1.000
血压/mmHg	<140	1.000
	140 ~ 160	2.206
	>160	2.300
运动	经常	1.000
	偶尔	1.000
	几乎不	1.693
饮酒	是	0.490
	否	1.000

2.2.2　遗传风险因素

影响阿尔茨海默病的遗传基因有很多，如表 2 所示，其中晚发性阿尔茨海默病中最常见的风险因素是称为载脂蛋白 E 或 *APOE* 的基因. 该基因具有许多不同的等位基因，常见的有 3 种，即 *APOEε*2，*APOEε*3 和 *APOEε*4，其中，等位基因 *APOEε*2 和 *APOEε*3 可防止阿尔茨海默病，而 *APOEε*4 会增加该病的发病风险. 每个人都有这个等位基因的两个拷贝，不同的组合决定了该疾病发生的可能性，其中，*APOEε*2/*APOEε*2 是最低风险，*APOEε*4/*APOEε*4 为最高风险. 各不同基因亚型对应的 *OR* 值见表 4.

表 4　*APOE* 各亚型对应的 *OR* 值

APOE Allele	*rs*429358	*rs*7412	*OR*
*APOEε*2/*APOEε*2	T T	T T	0.26
*APOEε*2/*APOEε*3	T T	T C	0.36
*APOEε*2/*APOEε*4	T C	T C	1.38
*APOEε*3/*APOEε*3	T T	C C	0.56
*APOEε*3/*APOEε*4	T C	C C	2.00
*APOEε*4/*APOEε*4	C C	C C	7.71

由于很多阿尔兹海默病患者并没有携带 *APOE*

基因或者 *APP* 基因，所以这些基因只是风险因素，而不是致病原因. 直接采用 *APOE* 单个基因型衡量患阿尔茨海默病风险会产生一定的偏差. 为保证预测评估模型更加可靠，将其他与阿尔茨海默病相关的风险基因也融入至风险评估模型.

2.3 SNP 位点 *OR* 值的预测

因大部分的 SNP 位点在 GWAS 数据库中不存在相应的 *OR* 值，所以采用多元线性回归模型来预测 SNP 位点的 *OR* 值，这里运用最小二乘法对回归参数进行估计. 具体的计算过程如下：

1）训练集. 在 GWAS 以及 GRASP（https：//grasp. nhlbi. nih. gov/Overview. aspx）数据库中找到含有纯合突变和杂合突变的 *OR* 值信息的位点，对这些位点进行 ANNOVAR[12] 注释，并以此作为训练集.

2）设所要预测的 *OR* 值为 *Y*，经过 ANNOVAR[12] 注释后得到的 SNP 位点可靠性的 6 个指标作为协变量，这 6 个指标分别为：Func. refGene[15]，ExonicFunc. refGene[16]，1000g2015aug_all[17]，SIFT_score[18]，MCAP[19]，CLINSIG[20].

3）利用训练集和协变量拟合回归方程，设拟合的多元线性回归模型为：

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\cdots+\hat{\beta}_nX_n, \tag{1}$$

其中，$\hat{\beta}_0$ 为回归常数，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_n$ 为待估参数，X_1，X_2，…，X_n 为所选的特征值.

运用普通最小二乘法来估计回归参数，得

$$\hat{\beta}=(X'X)^{-1}X'y. \tag{2}$$

对训练集进行训练，求出 $\hat{\beta}$ 并代入上式即可得到估计的回归方程：

$$\hat{Y}=0.4536+3.7602X_1+0.8097X_2-4.3605X_3+1.7676X_4+1.0211X_5+2.0453X_6.$$

经 *F* 检验得到的 *P* 值为 8.91×10^{-8}，所以拒绝原假设，认为 *Y* 与 X_1，X_2，X_3，X_4，X_5，X_6 有显著的线性关系，回归方程显著.

4）将过滤后所得的位点作为测试集，根据上面得到的回归方程对测试集进行预测，获取测试集中基于纯合突变和杂合突变各位点对应的 *OR* 值信息，用于后续的数据分析.

2.4 建立风险回归模型

2.4.1 单个 SNP 位点决定的疾病

假设阿尔兹海默病的发生只与单个的 SNP 位点有关. 根据 GWAS 数据库中与 SNP 相关的数据信息，设有致病风险的单个碱基为 A，没有致病风险的单个碱基为 B，有致病风险的杂合子基因型（AB）的 *OR* 值用 OR_1 表示，有致病风险的纯合子基因型（AA）的 *OR* 值用 OR_2 表示. 3 种基因型 AA，AB，BB 的基因型分布频率分别为 $\Pr(G_1)$，$\Pr(G_2)$ 和 $\Pr(G_3)$，患病率分别为 $\Pr(D/G_1)$，$\Pr(D/G_2)$，$\Pr(D/G_3)$，而 $\Pr(D)$ 表示该病与单个的 SNP 位点有关的患病率，则

$$\Pr(D)=\Pr(D/G_1)\Pr(G_1)+\Pr(D/G_2)\Pr(G_2)+\Pr(D/G_3)\Pr(G_3). \tag{3}$$

同样，也可以把 OR_1 和 OR_2 分别表示为如下：

$$OR_1=\frac{\Pr(D/G_2)/(1-\Pr(D/G_2))}{\Pr(D/G_1)/(1-\Pr(D/G_1))}; \tag{4}$$

$$OR_2=\frac{\Pr(D/G_3)(1-\Pr(D/G_3))}{\Pr(D/G_1)/(1-\Pr(D/G_1))}, \tag{5}$$

其中，OR_1 和 OR_2，$\Pr(G_1)$，$\Pr(G_2)$ 和 $\Pr(G_3)$ 通过查找文献［21］可知，根据上面的式子，即可求得单个 SNP 位点的各项特定基因型疾病下的患病率.

2.4.2 多个 SNP 位点决定的疾病

在实际研究中，疾病往往是和多个 SNP 位点相关的. 因此对于这种情况，计算疾病的患病风险. 假设多个 SNP 位点基因型组合的最终患病率为：

$$\Pr(D/\cap_{k=1}^{k}G_{m_k,k})=odds^{-1}[OR_c^*\times(odds(D)+OR_i^*)],i=1,2, \tag{6}$$

其中，OR_1^* 指遗传风险因素对应的 *OR* 值，OR_2^* 指环境风险因素对应的 *OR* 值，$odds(D)$ 表示单个样本患该疾病的 *OR* 值，OR_c^* 表示带有某些特定基因型组合的个体患病风险与总体人群患病风险的比值，单个 SNP 位点的特定基因型的 *OR* 值用 OR_x^* 表示，根据公式（4）和（5）即可求出，则

$$OR_x^*=\Pr(D/G_x)/(1-\Pr(D/G_x)). \tag{7}$$

单个 SNP 的患病风险值为 OR_m^*，可由下式得到：

$$OR_m^*=OR_x^*/OR_D. \tag{8}$$

假设与疾病相关的 SNP 位点有 *k* 个具体的基因型，得

$$OR_c^*=\prod_{k=1}^{k}OR_{m_k,k}^*. \tag{9}$$

同时 OR_c^* 也可用下面公式表示：

$$OR_c^* = odds(D/\cap_{k=1}^{k} G_{m_k,k})/odds(D). \quad (10)$$

通过上述式子，求出多个 SNP 位点基因型组合的患病率.(6)式即为最终的患病风险率的模型.

3　结果

为了更好地描述患病风险评估结果，将患病风险等级分为 4 个等级：

1）正常风险．截至目前为止，该疾病未发现或未收录任何有害突变．男性的相对患病风险率为 9%，女性的相对患病风险率为 17%.

2）低风险．该疾病的相对患病风险率高于正常人的相对患病风险率，低于相对患病风险率的 2 倍以内.

3）中等风险．该疾病的相对患病风险率高于正常人的相对患病风险率 2 倍，低于相对患病风险率 3 倍以内.

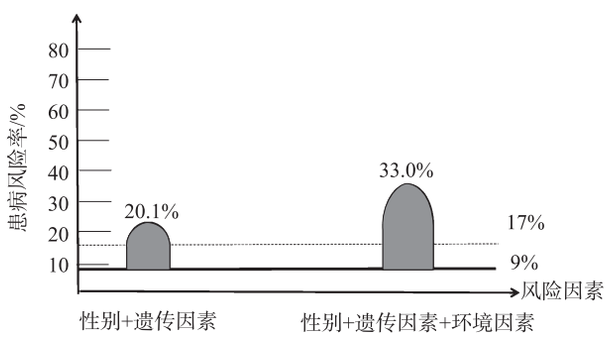
4）高风险．该疾病的相对患病风险接近正常人的相对患病风险率 4 倍及以上.

需要说明的是，由于疾病的发生是内因（基因、转录本、表观遗传性）和外因（生活、环境）等多种因素决定的，因此该疾病风险预测只是根据基因数据进行评估，不是临床诊断疾病的依据.

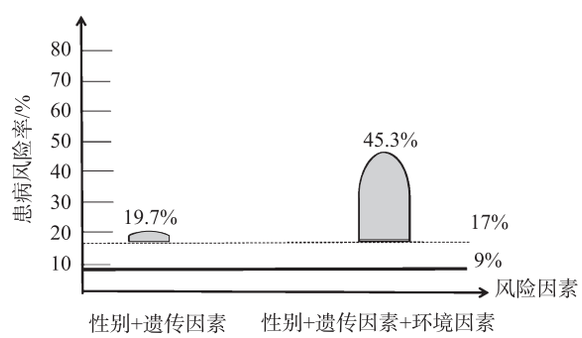
由于本文主要围绕环境风险因素和遗传风险因素对阿尔兹海默病患病风险进行研究，而男性和女性患阿尔兹海默病的患病率不同，因此在作评估时加入性别变量．将 7 个样本应用于上述的风险模型进行评估，得到阿尔兹海默病的风险评估结果，如图 4 所示.

由图 4（a）可知，样本 SN0001 为男性，风险因素为遗传因素时，患病风险率为 20.1%，高于正常男性患病率（9%）近 2 倍，为中等风险；风险因素为遗传和环境因素时，患病风险率为 33%，接近正常男性患病率（9%）近 4 倍，为高风险．此外，由表 1 可知，该男性年龄偏大，运动量少，饮酒，教育程度较低，因此，认为他患阿尔兹海默病的风险较高．由图 4（e）可见，样本 SN0005 为女性，风险因素为遗传因素时，患病风险率为 22.9%，略微高于正常女性患病率（17%），为低风险；风险因素为遗传和环境因素时，患病风险率为 35.2%，高于正常女性患病率（17%）的 2 倍以上，为中等风险．此外，由表 1 可知，该女性年龄偏大，运动量少，教育程度较低，因此，认为她有可能会患阿尔兹海默病.

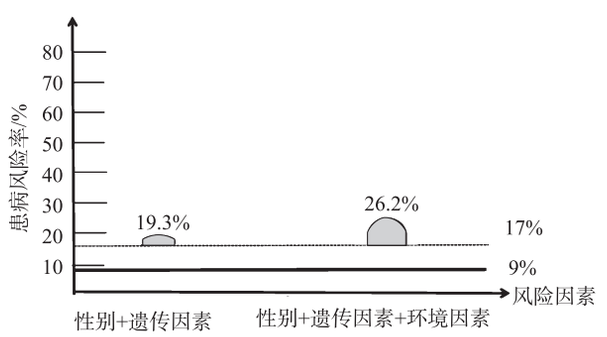
根据上述的分析过程，从而给出了图 4 中 7 个样本的患病风险率的患病风险等级评估表，如下表 5 所示.



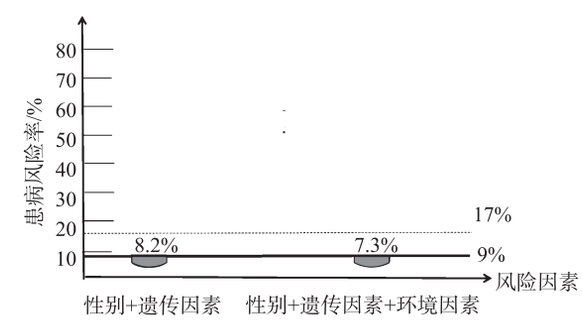
（a）SN0001



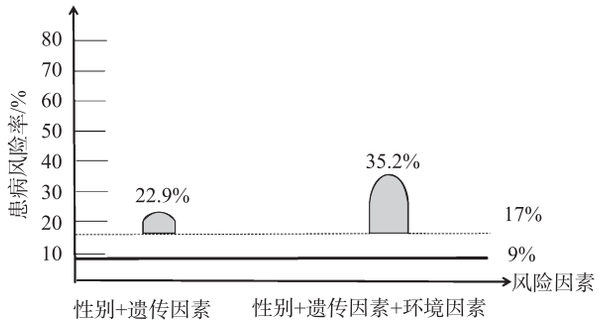
（b）SN0002



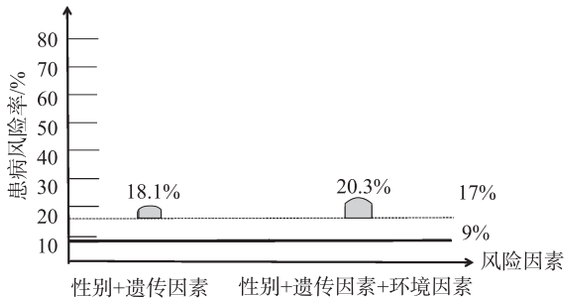
（c）SN0003



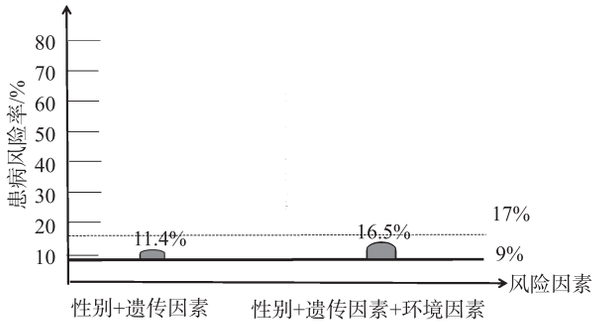
（d）SN0004



（e）SN0005



（f）SN0006



（g）SN0007

注：1)图中9%为男性患病风险率，17%为女性患病风险率；2)深色图形表示男性，浅色图形表示女性.

图4 性别、遗传和环境因素的影响评估

表 5 临床样本风险等级评估表

风险因素	SN0001	SN0002	SN0003	SN0004	SN0005	SN0006	SN0007
性别 + 基因	中等风险	低风险	低风险	正常	低风险	低风险	低风险
性别 + 基因 + 环境	高风险	中等风险	低风险	正常	中等风险	低风险	低风险

如表 5 所示，样本 SN0001 的患病风险偏高，样本 SN0002 和 SN0005 有可能患病，样本 SN0003，SN0006 和 SN0007 的患病风险较低，样本 SN0004 为正常. 由表 1 可知，这 7 个样本的年龄相对于患阿尔兹海默病的年龄偏低，因此大多数样本有一定的患病风险或风险偏低. 而临床上的结果显示，样本 SN0001 在临床上已被确诊为阿尔兹海默病，样本 SN0004 在临床上显示为正常. 因此，根据 7 个样本的分析结果，说明本文基于 SNP 位点预测患病风险具有一定的可靠性. 需要指出的是，存在阿尔茨海默病的风险因素仅仅是提示患病风险. 而患病风险提高，并不一定会患病. 对于存在阿尔茨海默病患病风险因素的个体，应在医生的指导下，积极调控风险因素，则可有效降低患病风险.

4 结语

本文围绕环境和遗传风险因素对阿尔兹海默病的患病风险进行研究. 首先，通过查找文献，获取与阿尔兹海默病相关的环境风险因素所对应的 *OR* 值. 其次，为提升评估模型的可靠性，对部分未报道的 SNP 位点的 *OR* 值采用 OLS 拟合预测. 最后，将预测到的 SNP 位点的 *OR* 值与环境风险因素和遗传因素融入到后续的疾病风险评估模型中，得到多个 SNP 位点导致患阿尔茨海默病的风险率. 本文主要根据 GWAS 和 GRASP 数据库中 SNP 位点的 *OR* 值，对未知的 *OR* 值建立回归模型进行预测，SNP 位点信息的广度和深度得到了很好的提升，具有一定的意义. 此外，将本文的评估模型所得结果和临床结果进行比较，两者之间具有一定的一致性，说明本文建立的风险评估模型具有一定的可靠性，对个体疾病诊断和预防具有积极的意义.

[参考文献]

[1] 黄越，马建芳，黄山. 阿尔茨海默病基因学研究进展[J]. 中国现代神经疾病杂志，2010，10（2）：171－176.

[2] BERTRAM L. Alzheimer′s disease genetics：current

status and future perspectives [J]. International Review of Neurobiology, 2009, 84: 167 – 184.

[3] GLENNER G G. Alzheimer's disease, in biomedical advances in aging [M]. Berlin: Springer, 1990: 51 – 62.

[4] BROOKMEYER R, JOHNSON E, ZIEGLER-GRAHAM K, et al. Forecasting the global burden of Alzheimer's disease [J]. Alzheimer's and Dementia, 2007, 3 (3): 186 – 91.

[5] CHAN K Y, WANG W, WU J J, et al. Epidemiology of Alzheimer's disease and other forms of dementia in China, 1990—2010: a systematic review and analysis [J]. Lancet, 2013, 381: 2016 – 2023.

[6] BERTRAM L, ANZI R E. Genome-wide association studies in Alzheimer's disease [J]. Human Molecular Genetics, 2009, 18 (2): 137 – 45.

[7] 张鑫，杨森，张学军. 全基因组外显子测序及其应用 [J]. 遗传，2011，33 (8)：847 – 856.

[8] MEYER M M, KIRCHER M M, KIRCHER M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing [J]. Cold Spring Harbor Protocols,2010 (6): 5448 – 5459.

[9] 李文轲，李丰余，张思瑶，等. 基因组二代测序数据的自动化分析流程 [J]. 遗传，2014，36 (6)：618 – 624.

[10] COCK P J A, FIELDS C J, GOTO N, et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants [J]. Nucleic Acids Research, 2009, 38 (6): 1767 – 1771.

[11] GUIMERA R V. Bcbio-nextgen: Automated, distributed next-gen sequencing pipeline [J]. Embnet Journal, 2012, 17: 30 – 39.

[12] WANG K, LI M, HAKONARSON H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data [J]. Nucleic Acids Research, 2010, 38 (16): 164.

[13] YANG H, ROBINSON P N, WANG K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases [J]. Nature Methods, 2015, 12 (9): 841 – 843.

[14] SO H C, GUI A H, CHEMY S S, et al. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases [J]. Genetic Epidemiology, 2011, 35 (5): 310 – 317.

[15] YANG H, WANG K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR [J]. Nature Protocols, 2015, 10 (10): 1556 – 1574.

[16] SHAFINA-NADIAWATI A, NURUL-SYAKIMA A B M, KHOR S S, et al. Molecular characterization of somatic alterations in dukes' B and C colorectal cancers by targeted sequencing [J]. Frontiers in Pharmacology, 2017, 8: 465 – 468.

[17] WANG N, ZHANG Y, GEDVILAITE E, et al. Using whole exome sequencing to investigate the genetic bases of lysosomal storage diseases of unknown etiology [J]. Human Mutation, 2017, 38 (11): 1491 – 1499.

[18] KUMAR P, HENIKOFFS, NG P C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm [J]. Nature Protocols, 2009, 7 (4): 1073 – 1081.

[19] DEY A, ELLENBERG J, FANINAA, et al. A bromodomain protein, MCAP, associates with mitotic chromosomes and affects G2-to-M transition [J]. Molecular & Cellular Biology, 2000, 20 (17): 6537 – 6549.

[20] JOSE J, SUBRAMONII H, PANDA D K, et al. Memcached design on high performance RDMA capable Interconnects [C] //IEEE Computer Society. 2011 International Conference on Parallel Processing. 2011: 743 – 752.

[21] ZUO L, DYCK C H V, LUO X, et al. Variation at APOE and STH loci and Alzheimer's disease [J]. Behavioral and Brain Functions, 2006, 2 (1): 1 – 26.