申请代码	H2611
受理部门	
收件日期	
受理编号	8150120200

国家自然科学基金

申 请 书

（2 0 1 5 版）

资助类别： 青年科学基金项目

亚类说明：

附注说明：

项目名称： 基于家系测序数据的遗传风险预测模型的构建

申 请 人： 温雅璐 **电 话：** 0411-86118632

依托单位： 大连医科大学

通讯地址： 大连市旅顺南路西段9号蓝湾医院205

邮政编码： 116044 **单位电话：** 0411-86110146

电子邮箱： wenyalu@dlmedu.edu.cn

申报日期： 2015年02月22日

国家自然科学基金委员会

基本信息

申请人信息	姓　　名	温雅璐	性别	女	出生年月	1985年10月	民族	汉族
	学　　位	博士	职称	副教授		每年工作时间（月）		9
	电　　话	0411-86118632			电子邮箱	wenyalu@dlmedu.edu.cn		
	传　　真	0411-86118952			国别或地区	中国		
	个人通讯地址	大连市旅顺南路西段9号蓝湾医院205						
	工　作　单　位	大连医科大学/肿瘤干细胞研究院						
	主要研究领域	遗传统计学,生物信息学分析,疾病风险预测						
依托单位信息	名　　称	大连医科大学						
	联　系　人	潘艳			电子邮箱	panyan999@126.com		
	电　　话	0411-86110146			网站地址	www.dlmedu.edu.cn		
合作研究单位信息	单位名称							
项目基本信息	项目名称	基于家系测序数据的遗传风险预测模型的构建						
	英文名称	Genetic risk prediction using family data						
	资助类别	青年科学基金项目			亚类说明			
	附注说明							
	申请代码	H2611						
	基地类别							
	研究期限	2016年01月 -- 2018年12月						
	申请经费	24.9240万元						
中文关键词		全基因组关联分析；风险预测；空间统计学；遗传流行病学；阿尔茨海默病						
英文关键词		Genome-wide association studies; Risk prediction; Spatial statistics; Genetic epidemiology; Alzheimer's disease						

中文摘要	疾病的风险预测已成为国内外研究的热点，它对疾病的预防和早期诊断、新药开发以及个体化医疗方案的制订有着重要的意义。然而，现有的遗传风险预测模型并不完全适用于二代测序数据，无法深入挖掘高通量数据所蕴含的信息，其预测准确度普遍较低，无法满足临床的实际需求。本研究在综合考虑基因与基因及基因与环境的交互作用的基础上，针对家系测序数据构建基于随机场统计理论的遗传风险预测模型。该模型能够从二代测序数据中高效筛选出具有预测功能的SNP及罕见变异，能够考虑潜在的基因与基因及基因与环境的交互作用，还能够综合利用家族成员内部关联性所提供的信息，最终提高遗传风险预测的准确度。本研究将在此基础上以特定真实数据为例，利用新方法构建遗传风险预测模型，建立模型入选因子的重要性评价以及模型临床实用性评估的标准。本项目所建立的模型为进一步提高风险预测的准确度提供新的思路和理论根据，为临床医学提供数据上的支持。
英文摘要	Studies of genetic risk prediction represent high priority research projects worldwide. The advance in risk prediction will lead to new strategies in disease prevention and early detection. It will also improve drug development and benefit the research in personalized medicine. However, the risk prediction models formed to date are not suitable for next generation sequencing data and they can't make full use of the information from high-throughput data. These models lack sufficient accuracy for clinical use. In this study, with the consideration of gene-gene and gene-environment interactions we will build a risk prediction model within the random filed framework for family-based genome-wide studies. An efficient selection algorithm will be established to tease out noise factors from next generation sequencing data. The proposed method can incorporate the information provided by SNPs, rare variants and environmental factors. It can also make good use of the information provided from family members to further improve the accuracy of the risk prediction model. We will apply the new method to a real data example. We will first build a genetic risk prediction model, and then evaluate the importance of the selected predictors and the clinical utility of the prediction model. This study will shed light on the further improvement of risk prediction models, and provide data support for clinical use.

项目组主要参与者（注：项目组主要参与者不包括项目申请人）

编号	姓名	出生年月	性别	职 称	学 位	单位名称	电话	电子邮箱	证件号码	每年工作时间（月）
1	吕德康	1983-11-23	男	讲师	博士	大连医科大学	041186118636	dekanglv@126.com	372526198311230034	3
2	王佳	1984-03-03	女	讲师	博士	大连医科大学	18841138501	wangjia@dlmedu.edu.cn	150202198403030961	2
3	陈富顺	1990-04-20	男	硕士生	学士	大连医科大学	13998510849	970445123@qq.com	13050319900420063X	12
4	康志杰	1979-10-06	女	博士生	硕士	大连医科大学	15541192328	cathie1997@sina.com	211422197910063548	6
5	陈成军	1990-11-25	男	硕士生	学士	大连医科大学	15724571395	Jimmy_ccj@outlook.com	371327199011250016	6

总人数	高级	中级	初级	博士后	博士生	硕士生
6	1	2			1	2

国家自然科学基金项目资金预算表（定额补助）

项目名称： 基于家系测序数据的遗传风险预测模型的构建

项目负责人：温雅璐 金额单位：万元

序号	科目名称	金额	备注
	(1)	(2)	(3)
1	**一、 项目资金支出**	24.9240	/
2	（一）直接费用	20.9700	
3	1、 设备费	8.7000	
4	(1)设备购置费	1.2000	存储计算机购买，数据备份设备购买
5	(2)设备试制费	0.0000	
6	(3)设备改造与租赁费	7.5000	高性能计算机集群租赁使用费
7	2、 材料费	0.5000	墨盒，打印纸，学术海报等
8	3、 测试化验加工费	0.0000	
9	4、 燃料动力费	0.0000	
10	5、 差旅费	3.0000	2人每年1-2次国内交流探讨及国内会议
11	6、 会议费	0.0000	
12	7、 国际合作与交流费	3.1500	专家来华交流1次，参加国际会议1次
13	8、 出版/文献/信息传播/知识产权事务费	1.0000	文献检索，版面费等
14	9、 劳务费	4.3200	3位研究生3年劳务费（每月600元）
15	10、 专家咨询费	0.3000	阿尔兹海默病专家咨询费
16	11、 其他支出	0.0000	
17	（二）间接费用	3.9540	
18	其中：绩效支出	0.9885	
19	**二、 自筹资金**	0.0000	

预算说明书

（一）直接费用： 共20.97万元

1、设备费：8.7 万元
（1）设备购置费：1.2万元。
高性能存储计算机购买约1万元，数据备份设备（例如移动硬盘）购买约0.2万元。
（2）设备改造与租赁费：7.5万元
高性能计算机集群使用费每年约2.5万元(3年，合计7.5万元)。

2、材料费：0.5万元
墨盒约400元一个，每年约需要4个，3年合计消费0.48万元。打印纸和学术海报等共计约200元。

3、测试化验加工费： 0万元

4、 燃料动力费：0万元

5、 差旅费：3万元
2人每年1-2次到国内相关单位学术交流，探讨遗传流行病领域的最新进展，以及寻求新的合作（平均每次每人830元，包括火车票、住宿、当地交通等费用）。
2人每年参加一次国内统计学相关会议以及一次遗传流行病学相关会议。 统计学相关会议有利于交流算法设计等统计问题，遗传流行病学相关会议将促进本项目的临床实用性研究以及了解本领域内的最新进展。费用平均每次3200元，包括火车票（硬座或硬卧），住宿（标准间），会议注册费，当地交通等费用。

6、 会议费：0元

7、 国际合作与交流费：3.15万元
邀请国际知名专家（美国）来华交流1次。国际机票（往返）1.5万元，住宿每日300元，共5日，合计1500元。专家来华日常开销每日80元，共400元。市内交通费用（包括机场往返接送）共100元。
参加一次国际会议（美国）。签证费用1008元，国际机票1.2万元（往返），住宿（每日200元，共1000元），当地交通等费用每日100元（共500元）。

8、 出版/文献/信息传播/知识产权事务费：1万元
发表SCI论文3-5篇，文献检索，SAS软件租用等费用。

9、 劳务费：4.32万元
研究生每月600元劳务费。其中一位研究生每年工作12个月，3年合计2.16万元。这位研究生的主要工作是负责模拟实验。另外两位研究生每年各工作6个月，3年合计2.16万元。这两位研究生的主要职责是真实数据解读和模型临床实用性探究。

10、 专家咨询费：0.3万元
模型在阿尔兹海默病上的临床实用性咨询（6次左右咨询费用，平均每次约500元）。

11、 其他支出：0元

报告正文

1. 立项依据与研究内容

（1）项目的立项依据。（研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）

全基因组关联研究(Genome-wide Association Study, GWAS)目前已经成为研究复杂性状或疾病的主要手段之一。自 2005 年 *Science* 杂志上发表的第一篇 GWAS 成果以来，GWAS 已检测到了大量与复杂疾病相关的单核苷酸多态性(Single Nucleotide Polymorphism, SNP) [1, 2]。近年来，随着二代测序技术的发展与成熟，很多次等位基因频率(Minor Allele Frequency，MAF)较小的罕见变异(Rare Variant)也被证实与疾病发病机制相关[3, 4]。这些 GWAS 的成果为确定疾病发病易感位点和相关基因，揭示疾病的遗传机制提供了依据。目前，GWAS 的研究成果正在逐步向应用方面转化，这主要集中在预测疾病的遗传风险。疾病的风险预测有利于疾病在高危人群中的预防和早期诊断，有利于个体化医疗方案的制订，可以推动临床医学的发展。利用全基因组数据所蕴含的信息，建立准确的遗传风险预测模型已成为国内外十分关注的课题。

a) 罕见变异的预测作用

尽管关联研究已经成功检测到了大量与疾病相关联的易感区域，但是迄今为止，现有方法所构建的遗传风险预测模型的准确度偏低，能够解释的遗传性也较少，产生了遗传性丢失(Missing Heritability)的现象[5, 6]。随着二代测序技术的发展与成熟，高通量芯片技术涉及不到的人类基因组内的罕见变异得以测量。近年来，研究者相继发现罕见变异与许多复杂疾病相关，越来越多的学者也开始倾向于相信多因素控制的复杂疾病的遗传模式是由基因组上常见变异和罕见变异共同作用的[7-10]。Eichler、Gibson 和 Leal 等学者提出**忽视罕见变异对于疾病的影响是造成遗传性丢失的主要原因之一**[6]。因此，将罕见变异所提供的信息有效的结合到现有的风险预测模型中将会进一步提高模型的准确度[8]。然而，由于现有的预测模型多半是建立在 SNP 数据的基础之上，无法将罕见变异对于疾病的预测力有效发掘。现有数据表明人类基因内蕴含大量的罕见变异，而基于 SNP 数据设计的模型却无法利用这些罕见变异所提供的信息，造成了大量的信息浪费，也间接的导致了现有风险预测模型准确度较低。**因此，需要建立新的遗传风险预测模型，使其能够综合考虑罕见变异和 SNP 所提供的信息，从而进一步提升风险预测模型的准确度。**

b) 预测因子的筛选

目前**大多数的遗传风险预测模型是基于候选基因、已有的生物学知识、或者满足 GWAS 显著性条件的因子来构建的**[11-13]。例如，Miyake 运用 11 个和二型糖尿病显著相关的因子来构建风险预测模型的[9]。尽管这些方法促进了学界对于疾病风险的认知，但是由于它们的准确度普遍较低，目前还无法直接运用到临床[14]。近年来有学者研究表明，**合理利用全基因组数据有利于进一步解释复杂性状或疾病的遗传性**[15]。例如，满足 GWAS 显著性要求的约 180 个 SNP 只能解释人类高度 10%的变异[16]，而运用全基因组内大量 SNP 则可以解释约 45%的变异[17]。由于全基因组数据既包含预测疾病风险的有效信息又包含大量噪声因素，因此**利用全基因组数据构建风险预测模型时需采用适当的筛选机制，使其可以高效筛选出具有预测功效的因子**[11, 18]。

现阶段，基于全基因组数据的预测模型大多是以关联研究的显著性为标准来剔除噪声因素的[11, 12, 19]。然而，近年来有学者指出以关联研究的显著性作为前置筛选条件来筛选预测因子是有一定的局限性的[20-22]。这是因为关联研究的显著性受检测因子的比值比（Odds Ratio）和样本量大小的影响，而这二者都不是衡量模型预测力的统计量的一对一函数[21, 23]。在现有的样本量下，许多尽管微效但却具有预测功能的 SNP 无法被筛选检测出，这使得这些因子的预测功效被忽视[11]。这种情况即使适当放松显著性条件也无法完全避免。因此，**在构建预测模型时需将预测因子筛选和模型建立有机融合以减少预测因子的遗漏，从而提高风险预测模型的准确度**。近年来，有学者利用 LASSO，Elastic Net 等惩罚回归模型在建立模型的同时进行数据筛选[22, 24]。这类方法可有效避免部分有预测功能的因子因不满足前置筛选条件而被遗漏。诸如 SVM、神经网络等机器学习算法在应用到遗传风险预测领域时，多以优化预测准确率为目标函数，进而筛选具有预测功效的因子来构建模型[23, 25]。尽管这些算法可以提高模型的准确性，但受制于计算效率等因素，这类算法目前还无法使用到全基因组数据上。Schwarz 开发的 Random Jungle 和 Ye 开发的前向 ROC 都通过采取贪婪算法寻找局部最优解的方式使其开发的算法可以运用到全基因组的 SNP 中，但是其无法有效筛选和提取罕见变异所提供的信息[26, 27]。**因此需要建立基于全基因组数据的风险预测模型使其可以综合利用各种变异所提供的信息，并且能够从全基因组数据中高效的提取出对疾病具有预测功效的因子**。

c）基因与环境交互作用对于风险预测的影响

现代研究已经证实各种疾病的遗传率（Heritability）不尽相同[18]。例如，一型糖尿病的预测遗传率约为 90%，而二型糖尿病却只有 50%[18, 28, 29]。 **对于遗传率较低的疾病，环境因素及基因与环境的交互作用对于疾病的发生产生了巨大作用，因此仅仅依靠遗传因素去预测疾病的风险显然不能满足临床的需**

要[5, 6, 8, 18, 25]。由于数据的高维性，同时考虑多个 SNP 以及 SNP 和环境间的交互作用会使得计算量以指数级增长，目前参数预测模型大多数只考虑两个 SNP 或者单个 SNP 与环境间的潜在的交互作用，而无法考虑多维的交互作用[30, 31]。部分非参数的模型，例如 MDR、GMDR 等可以考虑多维的交互作用，但是这些模型不具备处理全基因组数据的能力[32, 33]。Ye 提出的模型可以在全基因组范围内考虑高维的交互作用，但随着因子的增多，该方法也会受到“维度的诅咒”（Curse of Dimensionality），实际该方法一般也只能考虑 10 个变量之间的交互作用，且其无法考虑相近 SNP 之间潜在的连锁不平衡性（Linkage Disequilibrium）以及基因内部不同 SNP 间的复杂交互关系等诸多因素对于预测的影响[27]。基因是人类基因组内的功能单元，以基因为基本研究单位有利于综合基因内部各个位点间的潜在的复杂关系，也有利于综合考虑基因内部多位点与环境的交互作用，且现有关联研究也显示以基因为研究单位的 GWAS 成果更具有可重复性和生物可解释性[34]。**因此，从基因角度构建风险预测模型，并考虑潜在的基因与基因以及基因与环境间的交互作用，有利于构建出具有生物可解释性的高准确度的风险预测模型。**

d）针对家系数据的风险预测模型

家系实验（Family-based Study）由于其对于人群分层等因素具有无偏性等优点目前在关联研究中也被广泛使用，然而家系实验数据却很少被用来构建遗传风险预测模型，造成了数据资源的浪费。**现有研究表明许多复杂疾病都具有家族遗传性，家族病史在许多复杂疾病的临床诊断中起到重要作用**[19, 35]。相比于以人群为基础的实验（Population-based study）运用家系实验数据构建风险预测模型不但可以利用所有已测量的基因以及环境变量所提供的信息，而且**还可以利用家族成员内部之间的相关性间接的捕捉到许多未被测量的与疾病相关联的因子对于疾病的预测作用**[19]。例如，Ruderfer 等通过模拟实验证实了运用家族中其他成员的基因型和表现型的信息有利于提高预测模型的准确度[19]。此外，**家系设计更容易发掘出对疾病具有预测功效的罕见变异**，因为罕见变异易在家族中聚集。类似于以人群为基础的风险预测，运用家系数据构建风险预测模型也同样存在着无法运用罕见变异所提供的信息、无法从全基因组数据中高效筛选具有预测功效的因子以及无法在全基因组范围内考虑各种潜在交互作用等问题。而与以人群为基础的实验数据所不同的是：家系数据中的每个样本并不满足统计学中独立性的假设，因此许多算法无法直接应用到家系数据中。目前广泛运用的一类分析方法是通过模型来调整家族成员间的相关性[36-38]。例如，Meigs 等人运用遗传计分方法（genetic scoring method）在 GEE 模型下考量基因检测（Genetic Test）的预测力[37]。Ye 等人提出的 Clustered-ROC 方法则通过在家族内部构建统计量之后加和的方法来解决家系数据中不独立性的问题[38]。尽

管这类方法可以调整数据不独立性所产生的统计学问题，但是其并没有利用家族成员所提供的信息去进一步提高模型的准确度。**因此针对家系数据构建风险预测模型，不但需要解决以人群为基础的实验数据所需要解决的问题，而且还需要综合利用家族成员内部关联性所提供的信息去进一步提升模型的准确度。**

近年来，测序技术已经逐步成为进行全基因组研究所常用的技术手段。二代测序技术可以在短时间内产生数以亿计的数据，其中包括大量以往未被测量的罕见变异，为进一步提高遗传风险预测模型的准确度提供了新的数据资源。然而，相比于迅猛发展的生物实验技术，针对高维数据的统计方法发展相对滞后，无法高效的运用现有数据所提供的全部信息，制约了医学等相关科学的发展。本次申请的基于家系测序数据的遗传风险预测模型的构建旨在深度挖掘二代测序数据所蕴含的信息从而建立准确的遗传风险预测模型。**本项目拟建立的方法将在综合考虑潜在的基因与基因及基因与环境的交互作用的情况下，从全基因组数据中筛选具有预测功能的因子，进而构建风险预测模型。新模型还会进一步利用家族成员内部关联性所提供的信息去进一步提升模型的准确度。**本研究将同时在**多种模拟实验的条件下和已获使用授权的真实实验数据**（阿尔茨海默氏病的家系全基因组测序数据）下，以新方法构建风险预测模型，分析评估新方法在不同遗传模型和真实数据应用中的预测精确度，比较与现有方法的优劣。**针对真实数据，本研究将建立模型入选因子的重要性评价以及模型的临床实用性评估的标准。**本项目所建立的遗传风险预测模型将有效的弥补现有方法的缺陷，为进一步提高风险预测模型的准确度奠定基础，也为遗传风险预测的临床应用、高危人群的筛选以及个体化医疗的可行性提供理论研究的前期基础。

参考文献

1. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.
2. Stranger, B.E., E.A. Stahl, and T. Raj, *Progress and promise of genome-wide association studies for human complex trait genetics.* Genetics, 2011. **187**(2): p. 367-83.
3. Kryukov, G.V., L.A. Pennacchio, and S.R. Sunyaev, *Most rare missense alleles are deleterious in humans: implications for complex disease and association studies.* Am J Hum Genet, 2007. **80**(4): p. 727-39.
4. Lee, S., M.C. Wu, and X. Lin, *Optimal tests for rare variant effects in sequencing association studies.* Biostatistics, 2012. **13**(4): p. 762-75.
5. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
6. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease.* Nat Rev Genet, 2010. **11**(6): p. 446-50.

7. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
8. Gibson, G., *Rare and common variants: twenty arguments.* Nat Rev Genet, 2011. **13**(2): p. 135-45.
9. Zuk, O., et al., *Searching for missing heritability: designing rare variant association studies.* Proc Natl Acad Sci U S A, 2014. **111**(4): p. E455-64.
10. Cheeseman, I.H., et al., *Pooled sequencing and rare variant association tests for identifying the determinants of emerging drug resistance in malaria parasites.* Mol Biol Evol, 2014.
11. Chatterjee, N., et al., *Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies.* Nat Genet, 2013. **45**(4): p. 400-5, 405e1-3.
12. Miyake, K., et al., *Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association.* J Hum Genet, 2009. **54**(4): p. 236-41.
13. van der Net, J.B., et al., *Value of genetic profiling for the prediction of coronary heart disease.* Am Heart J, 2009. **158**(1): p. 105-10.
14. Jostins, L. and J.C. Barrett, *Genetic risk prediction in complex disease.* Hum Mol Genet, 2011. **20**(R2): p. R182-8.
15. Zhang, Z., et al., *Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies.* PLoS One, 2014. **9**(3): p. e93017.
16. Lango Allen, H., et al., *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* Nature, 2010. **467**(7317): p. 832-8.
17. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.
18. Wei, Z., et al., *From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes.* PLoS Genet, 2009. **5**(10): p. e1000678.
19. Ruderfer, D.M., J. Korn, and S.M. Purcell, *Family-based genetic risk prediction of multifactorial disease.* Genome Med, 2010. **2**(1): p. 2.
20. Jakobsdottir, J., et al., *Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers.* PLoS Genet, 2009. **5**(2): p. e1000337.
21. Pepe, M.S., et al., *Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker.* Am J Epidemiol, 2004. **159**(9): p. 882-90.
22. Kooperberg, C., M. LeBlanc, and V. Obenchain, *Risk prediction using genome-wide association studies.* Genet Epidemiol, 2010. **34**(7): p. 643-52.
23. Kruppa, J., A. Ziegler, and I.R. Konig, *Risk estimation and risk prediction using machine-learning methods.* Hum Genet, 2012. **131**(10): p. 1639-54.
24. Austin, E., W. Pan, and X. Shen, *Penalized Regression and Risk Prediction in Genome-Wide Association Studies.* Stat Anal Data Min, 2013. **6**(4).
25. Okser, S., T. Pahikkala, and T. Aittokallio, *Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives.* BioData Min, 2013. **6**(1): p. 5.
26. Schwarz, D.F., I.R. Konig, and A. Ziegler, *On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data.* Bioinformatics, 2010. **26**(14): p. 1752-8.
27. Ye, C., et al., *A non-parametric method for building predictive genetic tests on high-dimensional data.* Hum Hered, 2011. **71**(3): p. 161-70.
28. Stumvoll, M., B.J. Goldstein, and T.W. van Haeften, *Type 2 diabetes: principles of pathogenesis and therapy.* Lancet, 2005. **365**(9467): p. 1333-46.

29. Hyttinen, V., et al., *Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study.* Diabetes, 2003. **52**(4): p. 1052-5.
30. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
31. Wan, X., et al., *BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies.* Am J Hum Genet, 2010. **87**(3): p. 325-40.
32. Ritchie, M.D., et al., *Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.* Am J Hum Genet, 2001. **69**(1): p. 138-47.
33. Lou, X.Y., et al., *A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence.* Am J Hum Genet, 2007. **80**(6): p. 1125-37.
34. Neale, B.M. and P.C. Sham, *The future of association studies: gene-based analysis and replication.* Am J Hum Genet, 2004. **75**(3): p. 353-62.
35. Do, C.B., et al., *Comparison of family history and SNPs for predicting risk of complex disease.* PLoS Genet, 2012. **8**(10): p. e1002973.
36. Choi, S., et al., *FARVAT: a family-based rare variant association test.* Bioinformatics, 2014. **30**(22): p. 3197-205.
37. Meigs, J.B., et al., *Genotype score in addition to common risk factors for prediction of type 2 diabetes.* N Engl J Med, 2008. **359**(21): p. 2208-19.
38. Ye, C., J. Zhu, and Q. Lu, *A clustered optimal ROC curve method for family-based genetic risk prediction.* Statistics and Its Interface, 2011. **4**: p. 373-380.

（2）项目的研究内容、研究目标，以及拟解决的关键科学问题。

（此部分为重点阐述内容）

研究目标

本项目旨在建立一套针对家系数据的高效准确的风险预测模型，满足医学和生物学研究的需要。具体的研究目标如下：1）建立有统计理论支撑的适用于家系实验数据的遗传风险预测模型，使其可以综合考虑SNP、罕见变异和环境因素对于疾病风险的影响；2）评价新方法在不同遗传模式下的精确度和稳定性；3）应用新的方法针对阿尔茨海默氏病的家系测序数据构建风险预测模型，探索并尝试完成“建立风险预测模型—预测因子的重要性评价—模型临床实用性评估—模型优化决策方案—高危人群筛查、个体化医疗方案制定”这一遗传风险预测的完整模式设计。

研究内容

针对上述研究目标，本项目首先将建立有统计理论支撑的无偏的风险预测模型，而后建立相应的高效筛选机制，再通过模拟数据来检验模型的预测能力，最后将针对真实实验数据完成遗传风险预测的完整模式设计。具体研究内容如下：

a) 遗传风险预测模型的构建及预测因子的筛选

数据降维是分析高维数据所要解决的首要问题。本项目拟以基因或者一定长度的DNA片段为基本研究单位，借鉴空间统计学 (Spatial Statistics) 常用的距离概念，运用核函数 (Kernel Function) 计算出的距离来衡量两个样本在各个基因片段以及环境等因素上的相似性。根据家系数据的特点,该模型将会进一步运用不同的核函数来考量家族成员间的相似性。本项目进而运用随机场 (Random Field) 理论，根据核函数测量的相似性建立风险预测模型。最后将通过随机场的相关理论证明模型的无偏性，并进一步通过理论计算出该模型的最大预测方差以证明该模型的稳定性。区别于基于SNP的预测模型，**本模型是建立在基因片段上的，因此可以有效的将高维的SNP和罕见变异数据降维**，并且使得模型的解释更直观和更具有生物意义。**本项目将在预测模型中考虑罕见变异对于风险预测的影响，综合利用家族成员的相关性所提供的信息，且将实现在家系测序数据中同时考虑多个基因以及各种潜在的交互作用。**

和其他预测模型相似，过多的包含噪声因子将影响模型预测的准确性以及增大模型预测结果的波动性。高通量数据不但蕴含着与疾病相关的预测因子而且还包含着大量的噪声因素。尽管拟建立的模型是基于基因水平的，但是**运用枚举法来进行基因筛选在现有的计算环境下依旧难**

以实现。因此，**本项目拟采用贪婪算法，将模型建立和预测因子筛选有机结合，使其可以在高维数据中快捷准确的发掘出预测因子**。

b) 通过模拟数据以及实际数据证实模型的实际预测能力

模型的预测能力和稳定性需要通过数据进一步检验。本项目将首先通过模拟数据考量**模型在不同的疾病模型、噪声 SNP 数、噪声罕见变异数、噪声环境变量数以及不同家系结构下的预测能力，包括其与理论最佳预测能力的差异和预测方差**。新方法将针对准确度及稳定性等指标与基于GEE模型的遗传计分方法 (GS-GEE) 和SVM等常用的预测方法进行全面的比较。

其次，在真实数据分析方面，本研究将分析来自于 Alzheimer's Disease Sequencing Project (ADSP) 项目中的阿尔茨海默氏病家系全基因组数据。目前，申请人已获得该数据的分析使用权限。本研究将按照2：1的比例随机的将整体数据划分成训练集和预测集，分别用于遗传风险预测模型的构建和新模型预测精确度的估算。新方法的准确度将与GS-GEE，SVM等传统方法进行比较。

c) 建立遗传风险预测的完整模式

以阿尔茨海默氏病家系全基因组数据为例，本研究将首先在全基因组水平上应用新的方法构建风险预测模型，而后从对模型精确度的贡献率角度，评价纳入模型的各个预测因子的重要性。随后，再从临床实用性的角度，评价模型的实用性和可行性，并基于上述指标实现模型的优化决策。最终，采用优化模型筛选高危人群和制定个体化医疗方案。本研究拟设计并实现 **“建立风险预测模型—预测因子的重要性评价—模型临床实用性评估—模型优化决策方案—高危人群筛查、个体化医疗方案制定”这一遗传风险预测的完整模式**，为复杂疾病的遗传风险预测提供前期工作基础。

拟解决的关键问题

本项目旨在建立适用于家系实验设计的、可以综合考虑环境和遗传等因素的、具有较高预测准确性的遗传风险预测模型。因此本项目拟解决的关键问题主要包括以下三点：

a) **家系数据建模**：目前适用于家系数据的风险预测模型无法有效的将家族成员所提供的信息融合到风险预测当中。本模型拟采用随机场模型的相应理论，通过建立核函数 (例如血缘系数) 来综合运用家族成员所提供的额外信息去进一步提升风险预测模型的准确度。

b) **罕见变异对于疾病预测的贡献**：大量关联研究已经证实罕见变异和疾病相关，因此本项目拟通过建立适合的核函数在家系数据中来综合考虑SNP、罕见变异以及环境因素对于疾病预测的影响，并采用随机场模型的相应理论构建可以考虑多种交互作用的遗传风险预测模型。

c) **遗传风险预测模式的建立**：以阿尔茨海默氏病家系全基因组数据为例，探讨从全基因组风险预测模型的构建到高危人群及个体化医疗方案的制定，这一整套适用于临床的风险预测模式的可行性。针对模型入选预测因子的重要性以及模型的临床实用性构建评价标准。

（3）拟采取的研究方案及可行性分析。（包括有关方法、技术路线、实验手段、关键技术等说明）

研究方案

本研究主要利用空间统计学中的随机场理论建立适用于家系大数据的遗传风险预测模型。本项目首先从统计理论上验证模型的理论特性，包括偏差和方差以及在特定准确性的要求下所需的样本量。而后，根据大数据的特点，在综合考虑运算速度及内存需求等因素下，设计一套能与模型有机结合且能从高维数据中快捷准确的筛选预测因子的算法。本项目将通过大量模拟及实际数据验证模型的预测能力，包括在不同条件下的稳定性和偏差性。最后，将以实际家系数据为例探讨建立遗传风险预测的完整模式。具体的研究线路如图1。

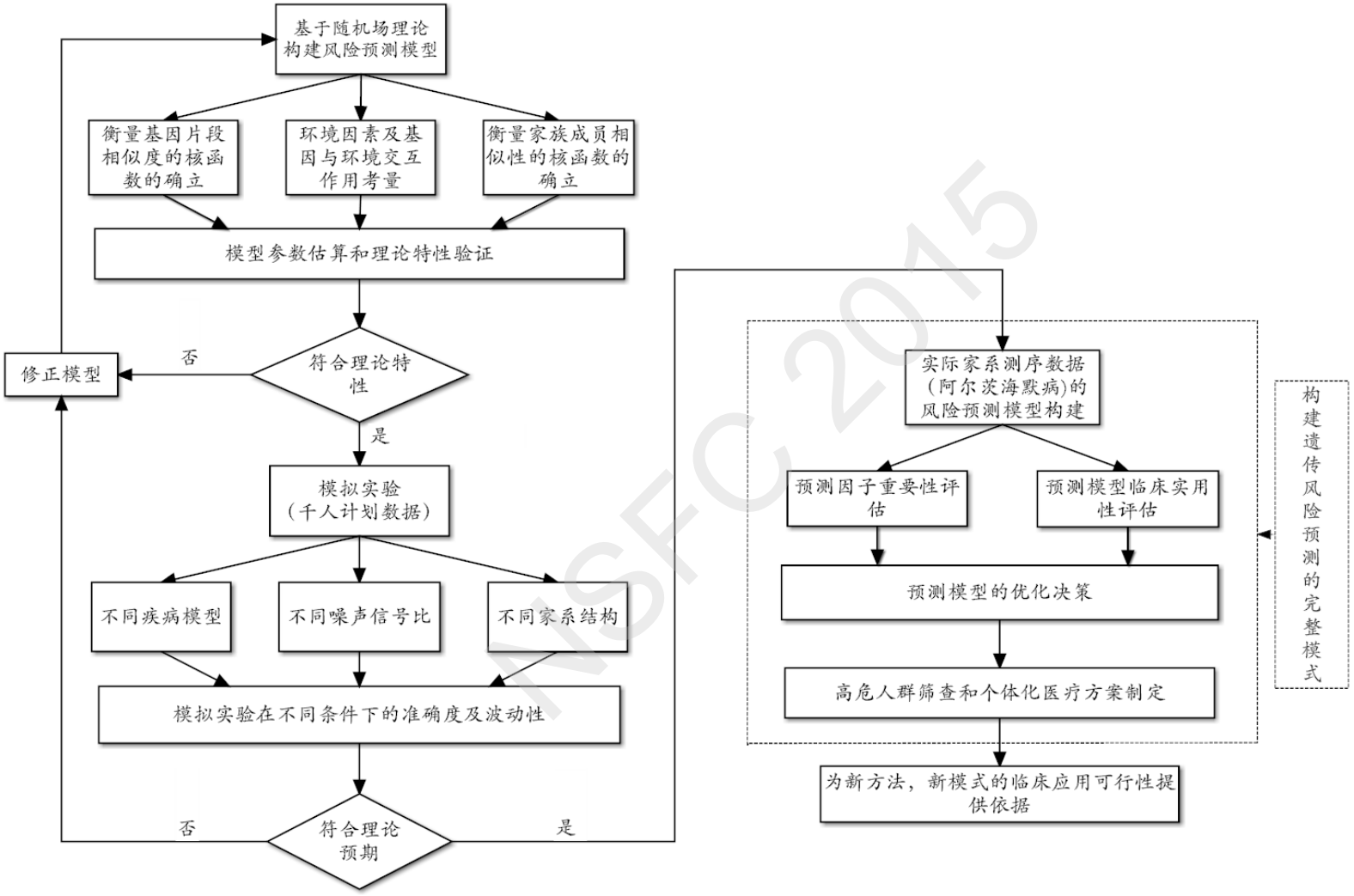

图 1. 技术路线图

a) **遗传风险预测模型的构建及预测因子的筛选**

借鉴关联研究中常用的假设，即相似的基因型将导致相似的表现型，**根据随机场理论本研究假设新样本表现型的预测可以根据已有样本中和其具有相似基因型及环境变量的其他样本的表现型来预测，即新样本的表现型是已有样本的一个加权平均**($Y_p = \sum_{i=1}^{n}\left[\omega_i\left(Y_i - X_i\beta\right)\right] + X_p\beta$，这里 X 是只考虑主作用的环境变量，β是这些主作用的大小）。权重的计算取决于相似性及各个基因作用大小。以考虑二维交互作用为例，

$$\omega_i = \sum_{e\in E}\sum_{k=1}^{K}\gamma_e^k s_{i,p}^k + \tau F_{i,p} + \sum_{(k,k')\in K} M_{k,k'}$$

这里γ_e^k代表第 k 段基因片段在 e 环境因素下作用力大小，$s_{i,p}^k$代表预测样本和第 i 个样本在第 k 段基因片段上的相似性，$F_{i,p}$代表预测样本和第 i 个样本由于家族所带来的相关性，τ代表家族相关性对于疾病预测的影响力，而最后一项是代表二维基因与基因的交互作用。尽管该模型首次在遗传风险预测领域中提出，但相似的思想在空间统计学中已广泛使用。例如气温的预测常是根据地理位置较近的区域的气温进行估算的，即预测值是已有样本根据欧氏距离的一个加权平均。

IBS (Identity-By-State) 核函数是目前分析二代测序数据比较广泛使用的函数之一，为合理考虑罕见变异对于疾病的影响，拟**采用以 MAF 为加权函数的 IBS 核函数来确定样本在各个基因片段上的相似性**，即

$$s_{i,p}^k = \sum_{l=1}^{m_k} f(MAF_l^k)\left(2 - \left|g_{i,l}^k - g_{p,l}^k\right|\right)$$

这里 m_k 是第 k 段片段上总变异数，$g_{i,l}^k$ 是样本 i 中第 k 段片段上第 l 位点上的基因型。针对家系数据，本项目**拟采用血缘系数（Kinship Coefficient）来考量家族成员之间的相似性。**模型以基因为基本研究单位，因此拟建立的模型在考虑交互作用时不是基于单点 SNP（罕见变异）的。由于本模型是在基因层面上考虑交互作用，因此拟建立的模型将会在每个基因片段上构建适当核函数，而后再考虑各个片段间的交互作用。以二维交互作用为例，公式中用来描述交互作用的 $M_{k,k'}$ 可以表述成 $M_{k,k'} = M_{k,k'}^1 \times M_{k,k'}^2$。这里 $M_{k,k'}^i$ 是在基因 i 上第 k 个样本和 k'个样本在去除主因素后的相似性。以基因为基本研究单位，将大大降低数据的维度，这使得在预测模型中考虑高维交互作用对于疾病预测的影响成为可能。拟建立的模型中的相关参数将采用最小化预测方差的方法来估算。

本项目拟建立的预测模型允许同时考虑多个 DNA 片段在不同环境因素下的作用，也允许不同的 DNA 片段对于疾病的发生具有不同方向的影响，即有些因素对于疾病有预防作用而有些因素将会加速疾病的发

生，这使得该预测模型可以适用于各种疾病发病机制。此外，由于基因片段相似性的引入，模型有效的降低了原始数据的维数，使得其可以同时考虑 SNP 和罕见变异等的影响。

尽管本模型是借鉴空间统计学中常用理论来建立的，但是鉴于生物大数据的独特特点以及新建模型的特异之处，本研究将会从统计理论上验证新建模型的各种统计学特点，包括对于不同分布的因变量预测的偏差性、方差、以及在指定样本量下的理论预测准确度。

高通量芯片技术和二代测序技术都已实现了全基因数据的覆盖，这些数据既包含了对疾病有预测功能的变量也包括大量的噪声因子，因此有效剔除数据中与疾病不相关的因素将有助于提高模型的准确性及稳定性。由于变量数目庞大，枚举式的基因片断筛选在现有的高性能计算机中很难实现。因此，**本项目拟采用向前选择（Forward Selection） 等贪婪算法进行快速高效的基因筛选，运用交叉验证（Cross-validation）以防止过度拟合（Over-fitting）等问题。**本项目将进一步探索贪婪算法对于基因筛选的影响。**区别于大多数适用于全基因组数据分析的方法，拟建立的模型将变量筛选和模型建立有机结合，不采用目前常用的两步法**（即第一步进行基因筛选，第二步建立预测模型）。

b) **通过模拟数据以及实际数据证实模型的实际预测能力**

对于**连续性的因变量，研究拟采用皮尔逊相关系数和误差平方和去衡量预测的效果，而对于二元变量，拟采用 ROC 曲线下面积（Area Under the Curve, AUC）和预测曲线（Predictiveness Curve)**来衡量模型预测效果。模拟实验拟在不同的疾病模型，不同的噪声 SNP 和罕见变异数与疾病相关 SNP 和罕见变异数的比值（噪声信号比），不同的环境变量数，以及不同的家系数据结构下，综合考量模型的预测能力以及稳定性。具体的家系模型包括得病子代-父母三元体（case-parents trios），同胞对（sib-pairs）以及复杂家系结构（complex pedigree）等。为保证模拟数据能够真实反应人类基因组中 SNP、罕见变异的分布以及相近位点间的连锁不平衡效应，本研究将从千人计划的数据中随机截取数段 DNA 片段作为基因型数据。千人计划中的样本将作为第一代，子代的基因型将根据父母基因型按照孟德尔遗传定律模拟产生。所有表现型数据将根据基因型信息以及环境变量的作用在不同的疾病模型下模拟产生。模拟实验将随机选取 2/3 的样本建立模型，采用剩余的样本检验模型的预测能力，并以此和理论值比较。而后，进一步将模型的准确度和波动性与其他预测模型相比较，以期能通过模拟研究全面的评价该预测模型。在本研究中，拟建立的方法将与目前应用最为广泛的 GS-GEE 方法（遗传

计分方法在家系数据中的拓展）以及 SVM 进行比较。最后，本项目将运用来自于 Alzheimer's Disease Sequencing Project (ADSP) 项目中的阿尔茨海默氏病家系全基因组测序数据来构建风险预测模型，并以此来考量模型在疾病发病机制未知的情况下的表现。

c) **建立遗传风险预测的完整模式**

Alzheimer's Disease Sequencing Project 项目拥有目前为止从全基因组角度研究阿尔茨海默病最为完整的全基因组家系数据。该项目旨在寻找预防及导致阿尔茨海默病的遗传因子，探求预防该疾病的有效机制、及解释携带治病因子却不发病的原因。该项目目前拥有来自于 111 个家庭的 578 份样本的全基因组测序数据，以及来自于 10939 个样本的全外显子测序数据。除了基因型的资料，该项目收集了性别、年龄、种族等基础信息以及阿尔茨海默病的患病情况等信息。

以此真实数据为例，本研究分别在候选基因（根据文献检索得到）水平以及全基因组水平上应用新的方法构建风险预测模型，并以 AUC 值作为准确度标准来评价新模型与现有模型的优劣。而后，借鉴随机森林等方法的相关理念来评价入选预测因子的重要性。通过重新分类比例（percentage of reclassification）和重新分类净提高（the net reclassification improvement)等指标来评价模型的临床实用性。根据上述指标，优化模型。最后，进一步咨询阿尔茨海默病的临床专家，对新建立的风险预测模型进行评估，探讨其应用推广的可能性。

可行性分析

a) 模型建立：尽管该项目首次运用家系数据中罕见变异所提供的信息建立风险预测模型，但罕见变异对于风险预测模型的意义以及罕见变异与许多疾病的关系已经在很多文献中有所报道。这使得本项目**所提出的观点有实际数据支持**。另外该项目所提出的风险预测模型是对空间统计学常用的模型的改进与创新，有**坚实的统计学理论做背景支撑**。这些基础可以保证本项目模型建立的可行性。

b) 数据来源：尽管现阶段国内的家系二代测序数据较少，但本项目负责人已经申请并且获得了许多国外相关机构的数据，这包括**千人计划的测序数据和阿尔茨海默氏病家系全基因组测序数据**。这些数据足以满足模拟实验以及实际数据分析的需求

c) 计算资源：二代测序技术产生的海量数据需要高性能的计算机来分析。本课题组所在实验室拥有**数台高性能计算机**，完全可以满足处理大数据的硬件要求。

d) 研究队伍：项目负责人有着多年大数据建模的经验，熟练掌握各种遗传统计学的算法。项目负责人拥有坚实的统计基础，可以从统计理论上验证模型的各种性能。课题组的主要成员也有着多年的程序撰写经验，这些都可以保证本项目的顺利实施。

（4）本项目的特色与创新之处。

本项目针对家系全基因组测序数据来建立遗传风险预测模型。该模型突破了许多制约风险预测模型准确性的瓶颈，具体归纳如下：

a) 运用罕见变异及家族成员的相关性所提供的额外信息来提升遗传风险预测模型的准确性。

b) 本项目实现了在基因层面上对疾病风险进行预测。本项目拟建立的模型是多基因且可以考虑高维交互作用的风险预测模型。相比于基于SNP的风险预测模型，该模型可以综合考虑基因内部各个位点间的复杂关系。此外，该模型所采用的筛选机制可以高效剔除噪声因子对于风险预测模型准确度的影响。

c) 本研究尝试建立一套遗传风险预测的完整模式。即通过新方法构建基于家系数据的全基因组风险预测模型，而后评价入选预测因子的重要性以及风险预测模型的临床实用性，优化模型进而为筛选高危人群及制定个体化医疗方案提供数据上的支持。该套完整的预测模式有望在阿尔茨海默氏病家系全基因组测序数据中首次实现。

（5）年度研究计划及预期研究结果。（包括拟组织的重要学术交流活动、国际合作与交流计划等）

年度计划：

2016.1-2016.12：

a) 遗传风险预测模型的建立，包括衡量基因片段及衡量家族内部成员间的相似度的核函数确立及相应的参数估计。

b) 通过统计理论证明该模型的特性，例如偏差性，波动性等。

2017.1-2017.12

a) 通过模拟实验验证模型在不同实验条件下的表现。

b) 进行必要的模型修改，使模型在不同情况下都具有较高的预测准确性。

c) 通过实际实验数据验证模型的预测效果。

d) 完成算法代码编写，在公共平台上发布该算法。

2018.1-2018.12

a) 以阿尔茨海默氏病的风险预测模型为例，探讨建立遗传风险预测的完整

模式。设计评价预测因子的重要性的指标。评价该模型的临床实用性，并根据临床实用性等指标进行适当的模型优化。

b) 撰写算法流程，发表论文，提交研究报告。

预期成果：

a) 建立可以综合考虑 SNP、罕见变异和环境等因素的，适用于家系二代测序数据的高效准确的遗传风险预测模型。

b) 建立面向大众的、易于使用的程序包来执行此算法。

c) 应用新算法建立阿尔茨海默氏病的遗传风险预测模型，并完成阿尔茨海默氏病的遗传风险预测完整模式的构建。

d) 培养 1-2 名硕士研究生。

e) 发表 5-8 论文，其中 3-5 篇被 SCI 收录。

学术交流：

a) 邀请 1-2 位统计遗传学领域著名的专家教授交流 1-2 次。

b) 参加统计及统计遗传学领域的国际会议 1-2 次。

2. 研究基础与工作条件

（1）工作基础（与本项目相关的研究工作积累和已取得的研究工作成绩）

项目申请人多年来一直从事针对大数据的建模工作，积累了大量的处理 GWAS 数据的经验。自 2009 年以来，申请人在国际 SCI 期刊发表论文 6 篇（其中 4 篇为第一作者），同时为每个自主开发的算法用 R 语言或者 C++语言编写相应的易于使用的程序包。申请人攻读博士期间建立的分析高通量芯片数据的模型已在 2013 年获得美国专利（专利号 8412465B2）。申请人作为项目的主要执行者参与了 NIH 资助的“Gene-Gene/Gene-Environment Interactions Associated with Nicotine Dependence”以及“A High-dimensional Statistical Genetic Approach for Family-based Orofacial Clefts Risk Prediction”两个项目。近年来，申请人研究了如何在 GWAS 数据中探求导致合病（Comorbidity）的遗传因素（已发表 SCI 论文 1 篇，IF=4.015）以及在考虑表观异质性（Phenotypic Heterogeneity）的情况下在高维数据中建立遗传风险预测模型（已发表 SCI 论文 1 篇，IF=4.015）。

随着二代测序技术的发展，大量的与疾病相关的罕见变异被发现，而针对 SNP 数据建立的多点风险预测模型无法全面的分析二代测序技术产生的数据。现阶段申请人着重研究二代测序技术这一新兴技术在遗传流行病学中的运用，以及如何利用 SNP、罕见变异、环境因素、各种交互作用以及家族信息对于疾

病的影响来进一步提升风险预测模型的准确性。目前申请人已熟练掌握现有的检测罕见变异是否与疾病相关的统计学方法，如基于负荷检验(Burden Test)和SKAT检测的系列统计方法。利用随机场理论，申请人针对以人群为基础的实验数据（population based genetic studies）已建立了可以综合考虑SNP、罕见变异以及二维交互作用的遗传风险预测模型（称之为FSRF）。通过模拟实验，在以人群为基础的实验数据中证实了，在测序数据中合理的考虑罕见变异有利于提高模型预测的准确性。如图2所示，基于随机场理论所建立的适用于以人群为基础的二代测序数据的遗传风险预测模型（FSRF）较SVM等常用方法更为准确。此外，在实际数据的分析中，以随机场理论所建立的模型也具有较高的AUC (FSRF的AUC为0.76，SVM的AUC为0.69)。目前，项目申请人开发的适用于以人群为基础的二代测序数据的遗传风险预测模型（FSRF）正在投稿阶段。通过大量的前期工作积累，项目申请人对于随机场理论、多因素控制疾病的遗传模式以及遗传风险研究中的难点和热点有着深入的认识，对于如何利用ROC曲线以及预测曲线构建风险预测模型有着丰富的经验。本项目拟建立的针对家系数据的遗传风险预测模型是基于随机场理论构建的，是对于FSRF的进一步延伸。FSRF的成功建立为探索SNPs、罕见变异以及环境变量在家系测序数据中对于遗传风险预测模型的作用奠定了坚实的基础。

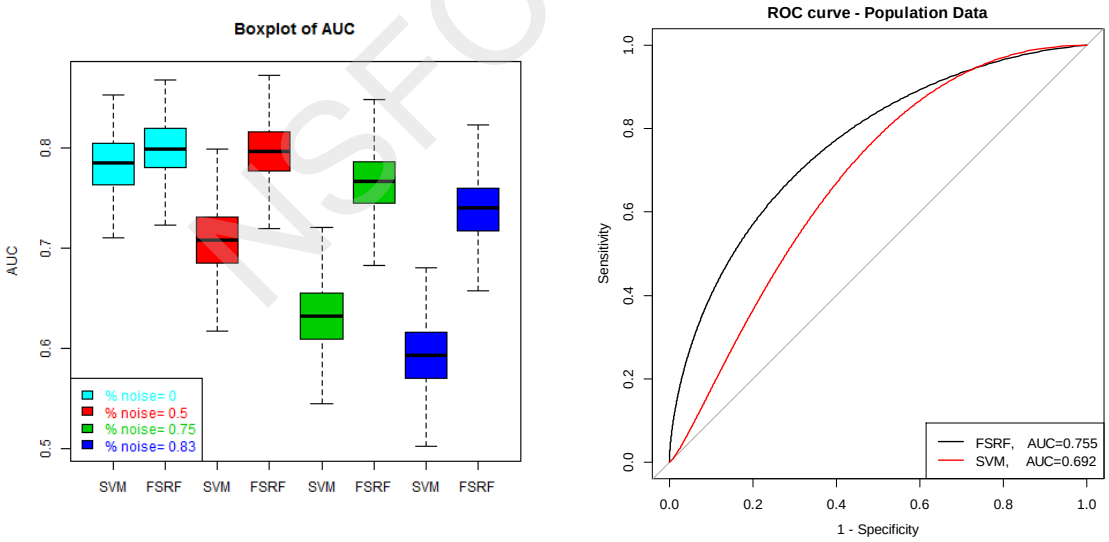

图2. 以人群为基础的数据，基于随机场理论所建立的遗传风险预测模型（FSRF）与SVM在不同噪声信号比的条件下的预测准确度比较

利用家系测序数据构建风险预测模型除了可以利用以人群为基础的测序数据所提供的信息外，还可以进一步利用家族成员间的相关性来提升风险预测模型的准确度。目前，初步模拟结果显示合理考虑家族成员间的相关性所提供的额外信息有利于进一步提高模型的准确度（图3）。在初步模拟实验中，我们仅考虑了三人家系数据，即以千人计划中的样本作为第一代，子代的基因型根据

父母基因型按照孟德尔遗传定律模拟产生。模拟产生的环境变量具有家族聚集性，即家庭成员间具有相似的环境变量。表现型的信息根据基因型及环境变量产生。如图 3 所示，利用家族成员所提供的信息所建立的方法（FSRF_Fam）的 AUC 高于 FSRF，SVM 以及遗传计分法（GS-GEE）。

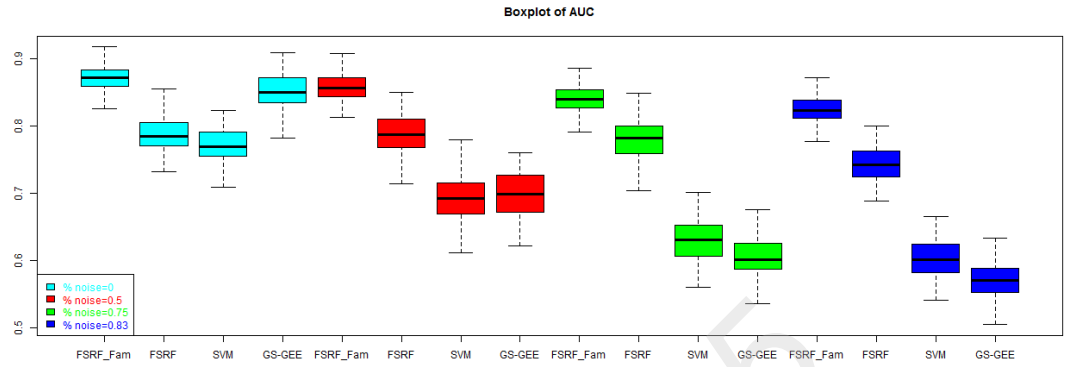

图 3. 针对家系数据，利用家族成员所提供的信息所建立的模型（FSRF_Fam）与忽视家族成员所提供的信息所建立的模型（FSRF）、SVM 及 GS-GEE 的准确度比较

尽管拟建立的模型在高维交互作用、复杂家系数据以及复杂疾病模型下的综合预测能力还有待于通过理论证明、大量模拟实验以及实际数据去检验，但这些初步研究工作为本项目的展开奠定了良好的基础。

为了调查基于阿尔茨海默病家系数据的遗传风险预测模型可能达到的精确度，申请人分别采用了 GS-GEE 和 FSRF 两种方法构建了遗传风险预测模型，并采用 AUC 来评估了上述两个模型的准确度（GS-GEE AUC=0.57，FSRF AUC=0.71）。由于基于随机场理论建立的风险预测模型（FSRF）并未利用家族成员间的相似性所提供的信息，且前期模拟实验也证实合理考虑家族相关性所提供的信息将有利于提升风险预测模型的准确度（图 3），因此申请人相信本项目所建立的适用于家系测序数据的预测模型（FSRF_Fam）具有达到更高精确度的条件。

交流合作方面，申请人目前正积极寻求和国内外相关领域的专家学者合作。目前，与美国密西根州立大学 Qing Lu 副教授和美国阿肯色大学的 Ming Li 助理教授有着密切的合作关系，对有关研究工作将进行广泛的学术交流。在项目的实施过程中，美国密西根州立大学 Qing Lu 副教授等专家可以提供技术支持。此外，神经系统疾病的临床医生 Gretchen Birbeck 博士将对风险预测模型的临床应用等问题提供意见和建议。

本项目组主要成员具有生物学、流行病学、统计学和计算机学背景，具有从事多学科交叉的研究基础。项目组主要成员有着多年的科研经验，严谨治学态度和对于遗传风险预测模型建立的旺盛的求知欲。这些都为本项目的开展提供了保障。

（2）工作条件（包括已具备的实验条件，尚缺少的实验条件和拟解决的途径，包括利用国家实验室、国家重点实验室和部门开放实验室等研究基地的计划与落实情况）

申请人所在的实验室隶属于大连医科大学肿瘤干细胞研究院下的基因组中心。针对国家对于癌症诊断和治疗的重大需求，基因组中心目前正在建设集肿瘤预防、早期诊断和个体化治疗为一体的肿瘤防治中心。基因组中心现拥有数台高性能计算机（CPU：Intel 酷睿 i7 3930K，内存：64G,硬盘：18T），其配置完全可以满足分析二代测序数据的需要。同时，为适应学院发展的需要，肿瘤干细胞研究院已初步建成了高性能计算机集群。除了高性能计算机，基因组中心拥有高技术实验室，基因组测序平台，以及各种深度测序技术所需的仪器。人员组成方面，本实验室及基因组中心有多位生物学、医学、生物信息学、统计学以及计算机学的科研人员。这些人员有多年进行数据处理与解读的经验。数据方面，项目申请人已申请并且获得了本项目所需要的二代测序数据，包括千人计划数据以及阿尔茨海默氏病家系全基因组测序数据。这些数据完全可以满足模型检验的要求。因此，现有实验条件以及人员组成可以满足本申请项目的实施。

（3）承担科研项目情况（申请人正在承担或参加科研项目的情况，包括国家自然科学基金的项目。要注明项目的名称和编号、经费来源、起止年月、与本项目的关系及负责的内容等）

大连医科大学人才引进启动基金，15 万元，2014-2019，该基金主要负责项目的启动。

（4）完成国家自然科学基金项目情况（对申请人负责的前一个已结题科学基金项目（项目名称及批准号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要（限 500 字）和相关成果的详细目录）

无。

3. 资金预算说明

购置单项经费 5 万元以上固定资产及设备等，须逐项说明与项目研究的直接相关性及必要性。

直接费用： 共 20.97 万元

a) 设备费：8.7 万元

i) 设备购置费（1.2 万元）：高性能存储计算机购买约 1 万元，数据备份设备（例如移动硬盘）购买约 0.2 万元。

ii) 设备改造与租赁费（7.5 万元）：高性能计算机集群使用费每年约 2.5 万元(3 年，合计 7.5 万元)。

b) 材料费：0.5 万元

墨盒约 400 元一个，每年约需要 4 个，3 年合计消费 0.48 万元。打印纸和学术海报等共计约 200 元。

c) 差旅费：3 万元

2 人每年 1-2 次到国内相关单位学术交流，探讨遗传流行病领域的最新进展，以及寻求新的合作（平均每次每人 830 元，包括火车票、住宿、当地交通等费用）。

2 人每年参加一次国内统计学相关会议以及一次遗传流行病学相关会议。统计学相关会议有利于交流算法设计等统计问题，遗传流行病学相关会议将促进本项目的临床实用性研究以及了解本领域内的最新进展。费用平均每次 3200 元，包括火车票（硬座或硬卧），住宿（标准间），会议注册费，当地交通等费用。

d) 国际合作与交流费：3.15 万元

邀请国际知名专家（美国）来华交流 1 次。国际机票（往返）1.5 万元，住宿每日 300 元，共 5 日，合计 1500 元。专家来华日常开销每日 80 元，共 400 元。市内交通费用（包括机场往返接送）共 100 元。

参加一次国际会议（美国）。签证费用 1008 元，国际机票 1.2 万元（往返），住宿 (每日 200 元，共 1000 元) ，当地交通等费用每日 100 元（共 500 元）。

e) 出版/文献/信息传播/知识产权事务费：1 万元

发表 SCI 论文 3-5 篇，文献检索，SAS 软件租用等费用。

f) 劳务费：4.32 万元

研究生每月 600 元劳务费。其中一位研究生每年工作 12 个月，3 年合计 2.16 万元。这位研究生的主要工作是负责模拟实验。另外两位研究生每年各工作 6 个月，3 年合计 2.16 万元。这两位研究生的主要职责是真实数据解读和模型临床实用性探究。

g) 专家咨询费：0.3 万元

模型在阿尔兹海默病上的临床实用性咨询（6 次左右咨询费用，平均每次约 500 元）。

4. 其他需要说明的问题

无。

温雅璐 简历

大连医科大学，肿瘤干细胞研究院，副教授

教育经历（按时间倒排序）：

2008/8 - 2012/8, 密西根州立大学, 流行病与生物统计, 博士, 导师: Fu Wenjiang

2004/9 - 2008/7，浙江大学，制药工程，学士，导师：任其龙

工作经历（科研与学术工作经历，按时间倒序排序）：

2013/12 - 至今，大连医科大学，肿瘤干细胞研究院，副教授

2012/8 - 2013/11，密西根州立大学，流行病与生物统计，博士后

曾使用证件信息（限3个）

护照，G28423769

主持或参加科研项目及人才计划项目情况（按时间倒序排序）：

项目类别、批准号、名称、研究起止年月、获资助金额、项目状态（已结题或在研等）、主持或参加。

大连医科大学人才引进启动基金，15万元，2014-2019，该基金主要负责项目的启动。

吕德康简历(参加者)

大连医科大学，肿瘤干细胞研究院，讲师

教育经历（从大学本科开始，按时间倒排序）：

2006/9 - 2011/12，东北农业大学，植物学，博士，导师：朱延明

2002/9 - 2006/6，山东农业大学，生物技术，学士；

2002/9 - 2006/6，山东农业大学，计算机科学与技术，学士；

工作经历（科研与学术工作经历，按时间倒排序）：

2013/11-至今，大连医科大学，肿瘤干细胞研究院，讲师

2012/05-2013/09，中科院生化细胞所，生物学博士后流动站，博士后

曾使用证件信息（限3个）

无

主持或参加科研项目及人才计划项目情况（按时间倒排序）：

项目类别、批准号、名称、研究起止年月、获资助金额、项目状态（已结题或在研等）、主持或参加。

1、国家自然科学基金面上项目、81472637、DNMT1 通过与 RNA 聚合酶 polII 相互作用而导致抑癌基因启动子区域异常甲基化的机制研究、2015/01-2018/12、72 万元、在研、参加。

王佳 简历(参加者)

大连医科大学，基础医学院，讲师

教育经历（从大学本科开始，按时间倒排序）：

2009/09 - 2013/12，大连理工大学，控制科学与工程，博士，导师：顾宏

2011/10 – 2012/10, University of California，San Diego, Mechanical and Aerospace Engineering，Mentor：Raymond De Callafon

2006/09 - 2009/01，大连理工大学，控制科学与工程，硕士，导师：王宏伟

2002/09 - 2006/07，河北科技大学，本科，

工作经历（科研与学术工作经历，按时间倒排序）：

2013/12 - 至今，大连医科大学，基础医学院生物技术系，讲师

曾使用证件信息（限 3 个）

无

主持或参加科研项目及人才计划项目情况（按时间倒排序）：

项目类别、批准号、名称、研究起止年月、获资助金额、项目状态（已结题或在研等）、主持或参加。

无

近5年（2010年1月1日以后）代表性研究成果列表

1、 Wen, Yalu(#), Lu, Qing(*), A Multiclass Likelihood Ratio Approach for Genetic Risk Prediction Allowing for Phenotypic Heterogeneity, Genetic Epidemiology, 2013, 37 (7) : 715-725。0, SCI 期刊论文

2、 Wen, Yalu(#), Li, Ming, Fu, Wenjiang J.(*), Catching the Genomic Wave in Oligonucleotide Single-Nucleotide Polymorphism Arrays by Modeling Sequence Binding, Journal of Computational Biology, 2013, 20 (7) : 514-523。0, SCI 期刊论文

3、 Wen, Yalu(#), Schaid, Daniel J., Lu, Qing(*), A Bivariate Mann-Whitney Approach for Unraveling Genetic Variants and Interactions Contributing to Comorbidity, Genetic Epidemiology, 2013, 37 (3) : 248-255。0, SCI 期刊论文

4、 Li, Ming(#), Wen, Yalu, Lu, Qing, Fu, Wenjiang J.(*), An Imputation Approach for Oligonucleotide Microarrays, PLos One, 2013, 8 (3) 。1, SCI 期刊论文

5、 Alaimo, Katherine(#)(*), Oleksyk, Shannon C., Drzal, Nick B., Golzynski, Diane L., Lucarelli, Jennifer F., Wen, Yalu, Velie, Ellen M., Effects of Changes in Lunch-Time Competitive Foods, Nutrition Practices, and Nutrition Policies on Low-Income Middle-School Children's Diets, Childhood Obesity, 2013, 9 (6) : 509-523。1, SCI 期刊论文

6、 Wen, Yalu(#), Li, Ming, Fu, Wenjiang J.(*), MA-SNP - A New Genotype Calling Method for Oligonucleotide SNP Arrays Modeling the Batch Effect with a Normal Mixture Model, Statistical Applications in Genetics and Molecular Biology, 2011, 10 (1) 。0, SCI 期刊论文

7、 Fu, Wenjiang, Wen, Yalu, Li, Ming, Microarray-based gene copy number analyses, USA, 8412465B2。 专利

8、 Fu, Wenjiang, Li, Ming, Wen,Yalu, Preeyanon, Likit, Some critical data quality control issues of oligoarrays, Frontiers in Computational and System Biology, Feng, Jianfeng;Fu, Wenjiang; Sun, Fengzhu, Springer, pp 39-59, USA, 2012 书籍章节

9、 Jia Wang(#), Hongwei Wang, Hong Gu(*), A novel recursive subspace identification approach of closed-loop systems, Mathematical and Computer Modelling of Dynamical System, 2013, 19 (6) : 526-539。 期刊论文

10、 Hongwei Wang(#), Jia Wang, Hong Gu(*), Fuzzy identification of nonlinear systems via orthogonal transform, Asian Journal of Control, 2011, 13 (6) : 1-6。 期刊论文

11、 Dekang Lv(#), Ying Ge, Bei Jia, Xi Bai, Peihua Bao, Hua Cai, Wei Ji, Yanming Zhu(*), miR167c is induced by high alkaline stress and inhibits two auxin response factors in Glycine soja, Journal of Plant Biology, 2012, 55 (5) : 373-380。SCI 期刊论文

12、 Ge Ying(#), Li Yong, Lv DeKang, Bai Xi, Ji Wei, Cai Hua, Wang AoXue, Zhu YanMing(*), Alkaline-stress response in Glycine soja leaf identifies specific transcription factors and ABA-mediated signaling factors, Funct Integr Genomics, 2011, 11 (2) : 369-379。SCI 期刊论文

13、 Ge Ying(#), Li Yong, Zhu YanMing, Bai Xi, Lv DeKang, Guo Dianjing, Ji Wei, Cai Hua(*), Global transcriptome profiling of wild soybean (Glycine soja) roots under NaHCO3 treatment, BMC Plant Biology, 2010, 10: 153-153。SCI 期刊论文

14、 Shu Yongjun(#), Li Yong, Zhu Yanming, Zhu Zhenlei, Lv Dekang, Bai Xi, Cai Hua, Ji Wei, Guo Dianjing(*), Genome-wide identification of intron fragment insertion mutations and their potential use as SCAR molecular markers in the soybean, Theoretical and Applied Genetics, 2010, 121 (1) : 1-8。SCI 期刊论文

15、 Gao Peng(#), Bai Xi, Yang Liang, Lv Dekang, Li Yong, Cai Hua, Ji Wei, Guo Dianjing, Zhu Yanming(*), Over-expression of osa-MIR396c decreases salt and alkali stress tolerance, Planta, 2010, 231 (5) : 991-1001。SCI 期刊论文

附件信息

序号	附件名称	备注	附件类型
1	专利		专利
2	4篇代表论文		代表性论著
3	伦理证明		其他

签字和盖章页

申 请 人： 温雅璐 依托单位： 大连医科大学

项目名称： 基于家系测序数据的遗传风险预测模型的构建

资助类别： 青年科学基金项目 亚类说明：

附注说明：

申请人承诺：

我保证申请书内容的真实性。如果获得资助，我将履行项目负责人职责，严格遵守国家自然科学基金委员会的有关规定，切实保证研究工作时间，认真开展工作，按时报送有关材料。若填报失实和违反规定，本人将承担全部责任。

签字：

项目组主要成员承诺：

我保证有关申报内容的真实性。如果获得资助，我将严格遵守国家自然科学基金委员会的有关规定，切实保证研究工作时间，加强合作、信息资源共享，认真开展工作，及时向项目负责人报送有关材料。若个人信息失实、执行项目中违反规定，本人将承担相关责任。

编号	姓名	工作单位名称	证件号码	每年工作时间（月）	签字
1	吕德康	大连医科大学	372526198311230034	3	
2	王佳	大连医科大学	150202198403030961	2	
3	陈富顺	大连医科大学	13050319900420063X	12	
4	康志杰	大连医科大学	211422197910063548	6	
5	陈成军	大连医科大学	371327199011250016	6	
6					
7					
8					
9					

依托单位及合作研究单位承诺：

已按填报说明对申请人的资格和申请书内容进行了审核。申请项目如获资助，我单位保证对研究计划实施所需要的人力、物力和工作时间等条件给予保障，严格遵守国家自然科学基金委员会有关规定，督促项目负责人和项目组成员以及本单位项目管理部门按照国家自然科学基金委员会的规定及时报送有关材料。

依托单位公章 合作研究单位公章1 合作研究单位公章2

日期： 日期： 日期：