·论著·方法学研究·

R语言在分子流行病学中的应用

吴 娜，宋花玲

上海中医药大学公共健康学院（上海 201203）

【摘要】通过比较 R 语言和 SPSS 软件的特点，重点探讨 R 语言在分子流行病学领域的应用优势。R 语言作为一种自由开源的编程语言，具有强大的数据处理和分析能力，适用于处理大规模和复杂的分子数据。其丰富的统计函数和数据包使得医学院校研究生可以进行高级统计建模和生物信息学分析，满足分子流行病学研究的需求。本研究通过 R 语言在分子流行病学研究中的应用实例，展示了 R 语言处理相关数据的功能。医学类高等院校教师应根据时代要求和现实需要，培养研究生应用 R 语言处理分析大数据的能力。

【关键词】R 语言；分子流行病学；医学院校研究生

The application of R language in molecular epidemiology

Na WU, Hua-Ling SONG

School of Public Health, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China
Corresponding author: Hua-Ling SONG, Email: 99shL@163.com

【Abstract】By comparing the characteristics of R language and SPSS software, this paper focused on the application advantages of R language in molecular epidemiology. As a free and open source programming language, R language has powerful data processing and analysis capabilities, which is suitable for processing large-scale and complex molecular data. Its extensive statistical functions and packages enable graduate students in medical colleges to conduct advanced statistical modeling and bioinformatics analysis to meet the needs of molecular epidemiological research. Through the application examples of R language in molecular epidemiology research, the function of R language in analyzing relevant data was demonstrated. Teachers in medical colleges and universities should train graduate students to master the ability to process and analyze big data using R language according to the requirements of the times and practical needs.

【Keywords】R language; Molecular epidemiology; Graduate students in medical colleges

DOI: 10.12173/j.issn.1004-4337.202309020

基金项目：中国博士后科学基金项目（2022M722164）；2023 年度上海市青年科技英才扬帆计划（23YF1447700）；上海市卫生健康委员会中医药科研项目（2022QN014）；上海中医药大学课程建设项目（2023SHUTCMKCJS060、2023SHUTCMKCJS057）；上海中医药大学学科内涵建设专项（GJ202303）；上海中医药大学校级重点课程建设项目（SHUTCMKCJSZD201723）

通信作者：宋花玲，副教授，硕士研究生导师，Email：99shL@163.com

分子流行病学作为医学院校公共卫生与预防医学专业研究生的公共基础课，不仅可以帮助医学研究生探索疾病传播和控制的科学意义及环境对人类健康的影响，还能从分子标记的角度了解饮食和营养影响人类健康的内在生物学机制。目前，大数据在识别和干预人口健康决定因素方面具有革命性意义，被认为是未来科学的革命性发展。为积极应对大数据时代的挑战，公共卫生尤其流行病学相关专业的研究生不仅需要掌握传统流行病学的方法，还应该熟知分子流行病学相关知识，尤其是通过 R 语言编程处理大数据，通过大数据确定人口健康的干预目标。为培养相关大数据人才适应社会发展需求，医学院校教师有必要在传统流行病学的基础上，指导研究生掌握新兴技术和大数据分析方法，将 R 语言应用于分子流行病学研究，推动分子流行病学领域的发展。

1 分子流行病学概述

1.1 分子流行病学的定义和应用

分子流行病学是一种将先进的生物学实验方法纳入传统流行病学，以确定疾病病因并提出相应预防和干预措施的科学 [1-2]。它越来越多地被作为一种了解外部环境暴露与遗传及其他易感因素间相互作用的工具，从而确定易感人群，被广泛应用于遗传及代谢性疾病。

1973 年，Kilbourne 在“流感的分子流行病学”一文中首次引入了分子流行病学的概念 [3]。随着第一本关于分子流行病学的书籍《分子流行病学：原理与实践》的出版，这个术语变得更加正式 [1]。分子流行病学主要研究遗传和环境因素在分子或细胞阶段的作用及其相互作用。2003 年人类基因组计划完成 DNA 全测序工作是该领域的一个突破。随着基因分型和高通量测序技术的发展，研究人员可以全方位评估人类的 DNA、RNA、蛋白质或代谢组分，为更全面地检测与疾病风险因素和途径相关的潜在生物学变异奠定了基础。另外，高通量技术丰富了研究人员对疾病表型 – 基因型关联的理解，有助于寻找疾病的生物标志物，并利用其识别易感人群，从而帮助临床医生为患者制定个性化的治疗方案。分子流行病学通过各种工具研究疾病的生物标志物，如 DNA 甲基化谱、蛋白质谱、代谢物或新基因，有助于发现疾病的病因和决定因素，进而预防疾病以达到改善公众健康的目的。

1.2 分子流行病学在医学院校研究生教育中的作用

传统流行病学是研究人群中疾病与健康状况的分布及其影响因素，以及防治疾病及促进健康的策略和措施的科学 [4]。分子流行病学作为传统流行病学与分子生物学的交叉学科，强调通过先进的技术检测生物学标志的分布情况，借助传统流行病学的研究方法，从更深层次即分子或基因水平阐明疾病的病因及其致病过程 [5]。分子流行病学是由传统流行病学学科发展的需求，以及分子生物学理论和技术的巨大成就相结合的产物，是近十几年迅速发展的一门流行病学新分支 [6]。

分子流行病学课程在医学院校研究生教育中起着至关重要的作用 [7]：①疾病诊断和预防。研究生能够了解不同疾病的分子机制，如遗传变异、突变和表达模式，这些知识对于疾病诊断、预后评估和预防是必不可少的。通过学习如何利用分子工具和技术识别病原体，研究疾病病因，有助于制定个性化医疗和预防策略。②药物开发和治疗研究。分子流行病学是药物开发和治疗研究的工具。通过掌握分子技术评估药物的有效性和安全性，可以获得有关药物代谢、药物靶标相互作用和药物作用机制的知识，这些知识对于研究和开发新的治疗方法和个性化药物至关重要。③流行病学研究的设计和分析。通过学习设计和开展分子流行病学研究，掌握分析大规模分子数据的统计和生物信息学方法，对于研究疾病的遗传和环境风险因素、建立疾病预测模型和评估干预措施的有效性具有指导意义。④研究技能和科学素养的培养。分子流行病学教育不仅注重传授理论知识，而且注重培养研究技能和科学素养，包括学习文献综述，制定研究假设，设计实验方案，收集和分析数据等。

2 R语言与SPSS软件的比较

2.1 R语言的特点

R 是一种用于统计计算和图形绘制的编程语言，由统计学家 Ross Ihaka 和 Robert Gentleman 创建，核心 R 语言由大量包含可重复使用的代码和文档的扩展包组成 [8-9]。在过去的三十年里，R 语言在统计学和生物信息学领域发挥了重要作用，目前已产生了数以万计的扩展包，涉及范围

从机器学习（如支持向量机、随机森林、人工神经网络等）到单核苷酸多态性（single nucleotide polymorphisms, SNPs）数据、转录组数据和 DNA 甲基化数据分析等[10-12]。

R 语言具有以下特点[13]：①开源性，可以免费下载，并提供复杂的数据分析功能，同时还有一个活跃的在线用户社区，使用者们可以在其中寻求帮助。②跨平台的编程语言，其代码可以在多个操作系统上运行，程序员只需编写一次程序。③可以进行各种机器学习操作，如分类、回归以及开发人工神经网络的各种扩展包。④可以绘制高质量图片，通过 ggplot2 和 plotly 等 R 包制作精美图片。⑤在 CRAN 存储库中存有超过 10 000 个扩展包，可以执行各种数据分析功能。⑥既能使数据可视化，又能连接外部数据库如基因表达综合数据库（Gene Expression Omnibus, GEO）、京都基因与基因组百科全书数据库（Kyoto Encyclopedia of Genes and Genomes, KEGG）等以执行高级生物统计功能。⑦作为一种不断发展的编程语言，每当添加任何新功能时，R 都会提供更新服务，便于广大用户使用。

2.2 SPSS软件的特点

SPSS（statistical product and service solutions）是一种数据统计分析软件，由 SPSS 有限公司于 1968 年推出，2009 年被国际商业机器公司（International Business Machines Corporation, IBM）收购。由于 SPSS 简单易操作，常被用于数据处理、市场调查等。

SPSS 具有以下特点：①不需要编程，简单易上手；②不适用于大数据分析，如分子流行病学中有关 SNPs、转录组学及 DNA 甲基化等大数据；③作为一款商业软件包，正版软件需要付费才可以使用。

2.3 R语言与SPSS软件的比较

分子流行病学是一门探究疾病病因相关分子生物标记的学科，SNPs 数据、转录组学数据及 DNA 甲基化数据等分子生物标记均属于大数据集，越来越多的研究人员选择使用 R 语言中的各种扩展包进行分析，而 SPSS 更适合分析样本量较少的人类测量学数据及血液学指标，见表 1。分子流行病学的教学目的之一是培养研究生掌握大数据处理与分析的能力以适应和满足社会需求。研究者可根据自身需要选择合适的统计软件，考虑到 R 语言在大数据处理上的优势，在分子流行病学研究中更推荐使用 R 语言。

3 R语言在分子流行病学中的应用

分子流行病学侧重研究生物标记物在疾病病因、风险评估和预防研究中的应用。通过选择和验证不同种类的生物标记物，采用不同的研究设计和 R 语言数据分析方法[14]。本研究通过案例介绍 R 语言在分子流行病学生物标记物 SNPs 和 DNA 甲基化修饰数据分析中的应用。

3.1 R语言应用于非酒精性脂肪肝的SNPs位点筛选

易感基因的 SNPs 位点是分子流行病学重点关注的一类生物标记物，也是分子流行病学课程教学的重要内容。利用 R 语言中的 SNPassoc 包的 association 函数分析非酒精性脂肪肝（non-alcoholic fatty liver disease, NAFLD）的易感基因 SNPs 在五种遗传模型下的基因型频率，操作简单，结果展示清晰明了。具体代码如下：

```
> setwd
> library(openxlsx)
> File<- read.xlsx("NAFLD.xlsx",5)
> File[File=="0 0"]<-NA
> File[File=="NA"]<-NA
> File[File==""]<-NA
> File<-as.data.frame(File)
> write.csv(File, file = "NAFLD_1.csv")
> library(SNPassoc)
> names(File)
> File.1<- setupSNP(File,colSNPs=2,sep="")
> zlassoc<- WGassociation(NAFLD~1,data=File.1)
> zlassoc
> dev.new()
> plot(zlassoc,ylim = c(-0,-2))
> write.csv(zlassoc,"NAFLD_5model.csv")
> asso<- association(NAFLD~rs1260326,data=File.1)
> asso
> write.csv(asso,"rs1260326_result.csv")
```

表 2 展示了 rs1260326 在五种遗传模型下基因型的频率，NAFLD 的葡萄糖调节蛋白基因（glucokinase regulator, GCKR）的 rs1260326 位点在显性模型（P=0.038）和超显性模型（P=0.040）下具有统计学意义。

表1 R语言与SPSS软件的比较

Table 1. Comparison between R language and SPSS software

特点	R语言	SPSS软件
开源免费	是	否
更新快	是	否
数据分析	功能强大，可分析大数据集	分析少量基础数据
图形可视化	根据不同软件包，呈现方式多样，可视化程度高	固定样式
交互性强	是，可以外连各种数据库进行高级生物统计学分析	是，可以外连部分数据库
丰富扩展包	是，各种功能的扩展包可以满足研究者各种需求	是，少量扩展包

表2 SNPs位点在五种遗传模型下的基因型频率分布

Table 2. Genotype frequency distribution of SNPs loci under five genetic models

	NAFLD（*n*，%）	Non-NAFLD（*n*，%）	OR值	95%CI	*P*值
共显性模型					0.113
C/C	46（70.8）	77（84.6）	1		
T/C	18（27.7）	13（14.3）	0.43	（0.19，0.96）	
T/T	1（1.5）	1（1.1）	0.60	（0.04，9.78）	
显性模型					0.038
C/C	46（70.8）	77（84.6）	1		
T/C-T/T	19（29.2）	14（15.4）	0.44	（0.20, 0.96）	
隐性模型					0.811
C/C-T/C	64（98.5）	90（98.9）	1		
T/T	1（1.5）	1（1.1）	0.71	（0.04, 11.58）	
超显性模型					0.040
C/C-T/T	47（72.3）	78（85.7）	1		
T/C	18（27.7）	13（14.3）	0.44	（0.20，0.97）	
加性模型					0.050
0,1,2	65（41.7）	91（58.3）	0.49	（0.24，1.01）	

3.2 R语言应用于非酒精性脂肪肝的DNA甲基化修饰标记筛选

在后基因组时代，随着高通量技术成本的降低，海量组学数据与研究结果展现了生命现象的复杂性。目前，分子流行病学研究越来越倾向于从多组学的角度出发，从遗传和表观遗传到转录和代谢，从机制到表型，进行整合研究以得到全局结果。DNA 甲基化是表观遗传学中研究最多的一种修饰，是将甲基基团（CH3）转移至 DNA，从而使基因活性发生改变的修饰方式。

在当前的科研需求下，Illumina 的甲基化芯片 Infinium Methylation EPIC BeadChip（简称 850k 芯片）可以检测超过 853 000 个 CpG 位点，全面覆盖 CpG 岛、启动子、编码区、开放染色质和增强子，提供了性能优越且经济可靠的解决方案。R 语言中 CHAMP 包的 CpG.GUI 函数可以分析 CpG 位点在染色体、CpG 岛、转录起始区域（transcription star site, TSS）的分布情况，见图 1。

差异甲基化位点的筛选是数据分析过程的主要环节，R 语言中 CHAMP 包的 champ.DMP() 函数可以计算差异甲基化，使用 ggplot2 包可以绘制火山图，以展示 NAFLD 患者相比于健康人群的差异甲基化位点，见图 2。

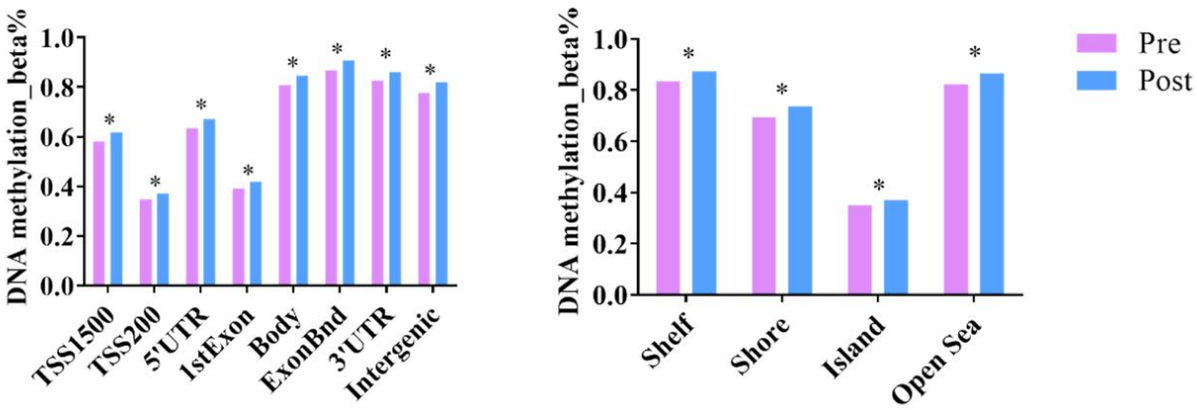

图1 NAFLD患者与健康人群DNA甲基化的分布情况

Figure 1. Distribution of DNA methylation between NAFLD patients and healthy people

注：*$P<0.05$；pre：健康人群；post：NAFLD患者。

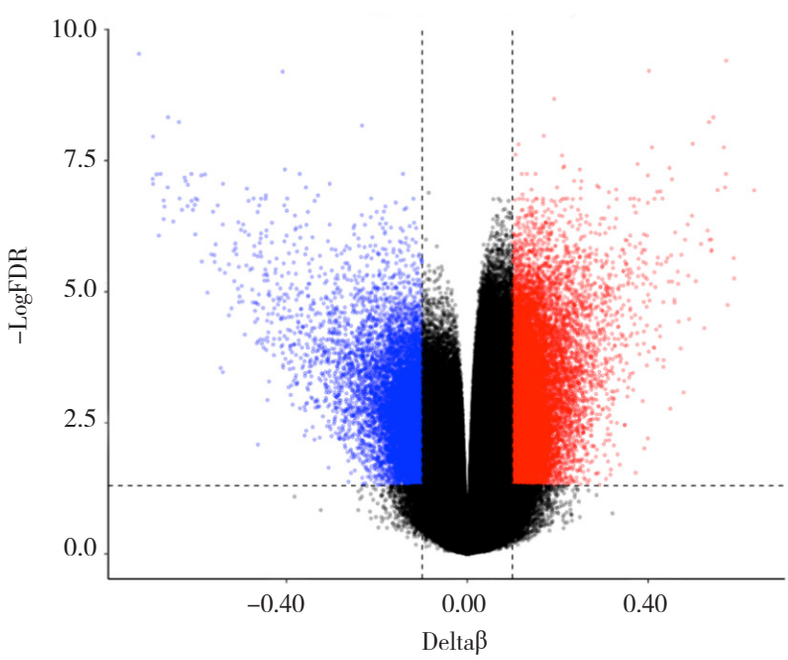

图2 NAFLD的差异DNA甲基化位点

Figure 2. Differential DNA methylation sites in NAFLD

注：红色代表试验组高甲基化位点，蓝色代表试验组低甲基化位点，黑色表示甲基化位点在组间无变化。

4 结语

本文通过比较 R 语言和 SPSS 软件的特点，重点探讨了 R 语言在分子流行病学中的应用，R 语言具有强大的绘图及数据分析能力，在大数据处理与分析上更具优势。医学类高等院校教师应根据时代要求和现实需要，培养研究生掌握应用 R 语言处理和分析大数据的能力，进一步满足分子流行病学领域的研究需求。

参考文献

1 Schulte Paul A, Perera Frederica P. Molecular epidemiology: principles and practice[M]. Cambridge, Massachusetts: Academic Press, 1993.

2 Rebbeck TR, Ambrosone CB, Shields PG. Molecular epidemiology[M]. Florida: Chemical Rubber Company, 2013.

3 Kilbourne ED. The molecular epidemiology of influenza[J]. J Infect Dis, 1973, 127(4): 478–487. DOI: 10.1093/infdis/127.4.478.

4 金倩莹, 李星明. 流行病学方法在医学研究中的应用概述[J]. 北京医学, 2020, 42(5): 444–451. [Jin QY, Li XM. Overview of the application of epidemiological methods in medical research[J]. Beijing Medical Journal, 2020, 42(5): 444–451.] DOI: 10.15932/j.0253–9713.2020.05.023.

5 Honardoost M, Rajabpour A, Vakili L. Molecular epidemiology; New but impressive[J]. Med J Islam Repub Iran, 2018, 32: 53. DOI: 10.14196/mjiri.32.53.

6 Slattery ML. The science and art of molecular epidemiology[J]. J Epidemiol Community Health, 2002, 56(10): 728–729. DOI: 10.1136/jech.56.10.728.

7 Tümmler B. Molecular epidemiology in current times[J]. Environ Microbiol, 2020, 22(12): 4909–4918. DOI: 10.1111/1462–2920.15238.

8 Chan BKC. Data analysis using R programming[J]. Adv Exp Med Biol, 2018, 1082: 47–122. DOI: 10.1007/978–3–319–93791–5_2.

9 Ihaka R, Gentleman R. R: a language for data analysis and graphics[J]. J Comput Graph Stat, 1996, 5(3): 299–314. DOI: 10.1080/10618600.1996.10474713.

10 Rhys H. Machine learning with R, the tidyverse, and mlr[M]. New York: Simon and Schuster, 2020.

11 Giorgi FM, Ceraolo C, Mercatelli D. The R language: an engine for bioinformatics and data science[J]. Life (Basel), 2022, 12(5): 648. DOI: 10.3390/life12050648.

12 Boehmke B, Greenwell BM. Hands–on machine learning with R[M]. Florida: Chemical Rubber Company, 2019.

13 Wickham H, Bryan J. R packages[M]. California: O'Reilly Media, 2023.

14 Bryan K. Epidemiology and biostatistics[M]. Berlin, Heidelberg: Springer, Cham, 2018.

收稿日期：2023 年 09 月 04 日 修回日期：2023 年 10 月 23 日
本文编辑：张 苗 黄 笛

引用本文：吴娜, 宋花玲. R语言在分子流行病学中的应用[J]. 数理医药学杂志, 2023, 36(11): 797–802. DOI: 10.12173/j.issn.1004–4337.202309020
Wu N, Song HL. The application of R language in molecular epidemiology[J]. Journal of Mathematical Medicine, 2023, 36(11): 797–802. DOI: 10.12173/j.issn.1004–4337.202309020