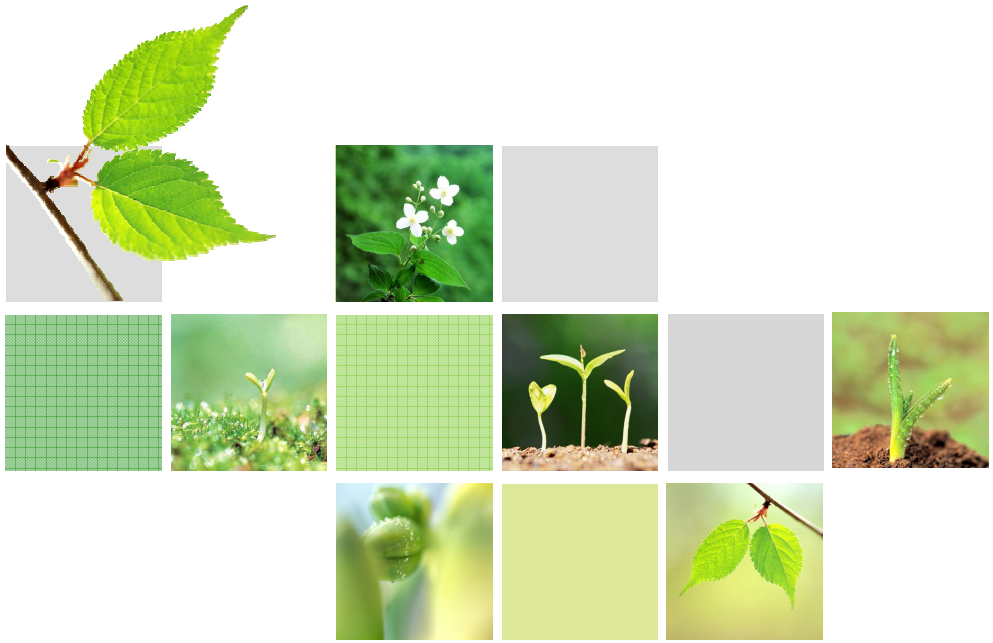
R在肿瘤生存分析中的应用



钟春燕

zhngchy@163.com

内容提要

1 肿瘤生存分析简介

2 **SS-RPMM**

3 应用实例

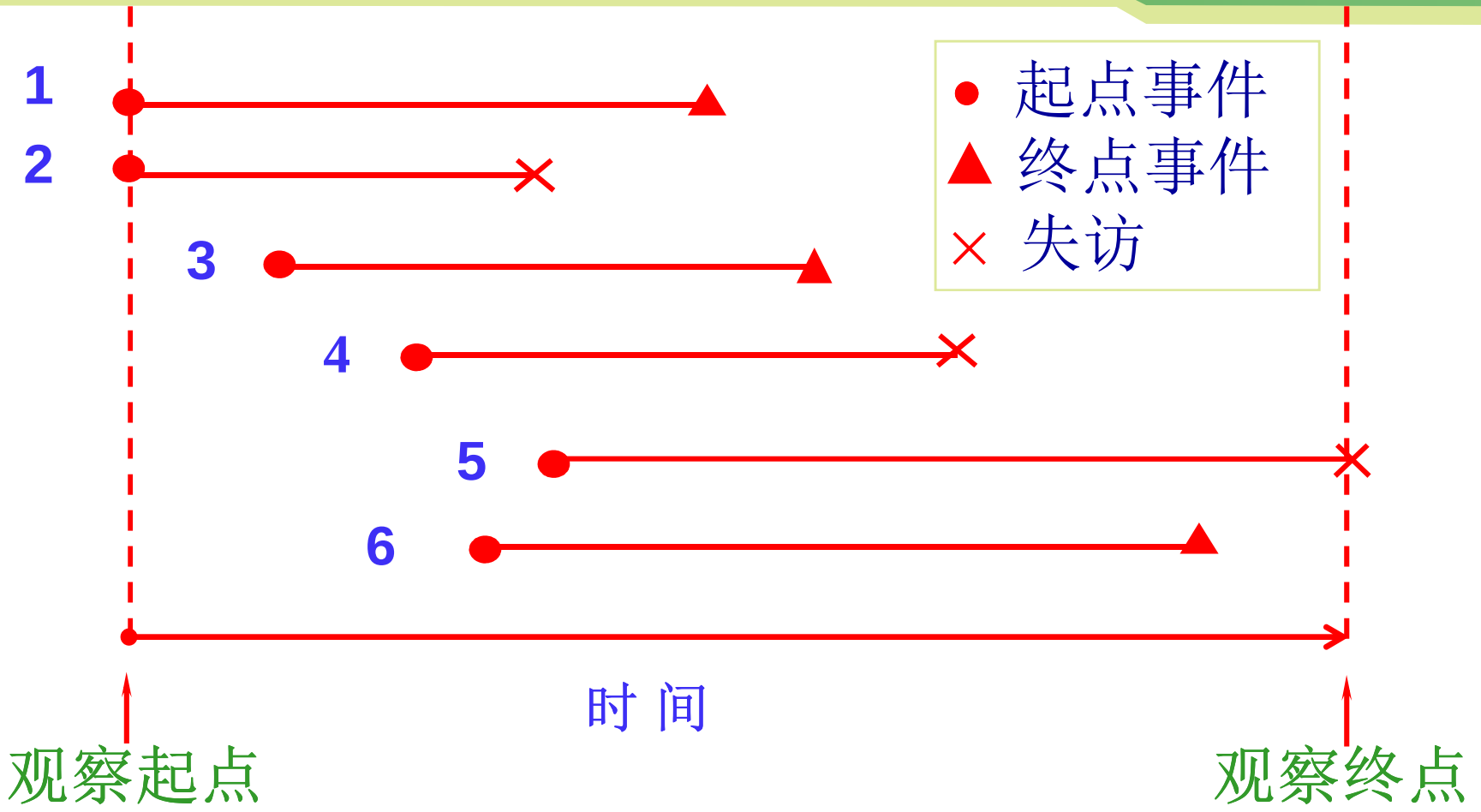
肿瘤生存分析简介（一）

- 生存时间

（狭义）：患某种疾病的病人从发病到死亡所经历的时间跨度。

（广义）：从某种起始事件到终点事件所经历的时间跨度。

- 生存时间是一个随机变量，服从偏态分布
- 含有截尾值 **(censored value)**



临床随访研究中的完全数据与截尾数据

产生截尾值的原因

病例死于与所研究的疾病无关的其他原因

病人因搬迁等失去联系

病人的生存期超过了研究的终止期

截尾值

不知道病人真正的生存时间，但其真实的生存时间肯定长于观察到的时间。

肿瘤生存分析简介（二）

- 生存函数/累积生存率：

$$S(t)=P(T>t)=\frac{\text{生存时间}T>t\text{的病人数}}{\text{观察病人总数}}$$

- 风险函数：生存时间已达到t的一群观察对象在时刻t的瞬时死亡率。

$$h(t)=\frac{\text{死于区间}(t,t+\Delta t)\text{的病人数}}{\text{在}t\text{时刻尚存的病人数}\times\Delta t}$$

- 风险比（**HR**）：同一时间两组观察对象的风险函数之比，相当于相对危险度（**RR**）

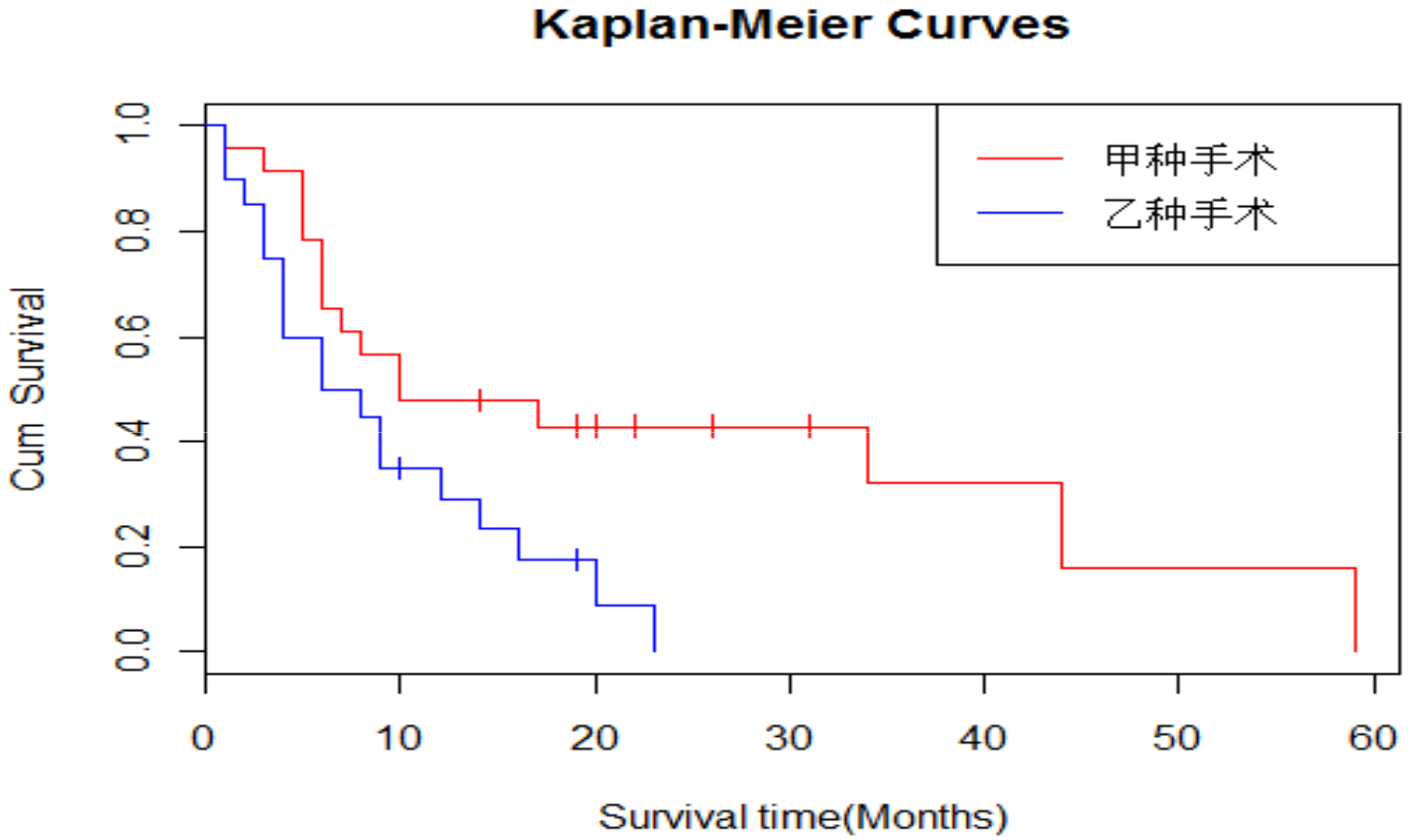
$$\text{风险比}=\frac{\text{第一组的}h_1(t)}{\text{第二组的}h_2(t)}$$

肿瘤生存分析简介（三）

- 中位生存时间：刚好有50%的个体其存活期大于该时间，是评价疾病预后的指标。

- 生存曲线：以生存时间为横轴，生存率为纵轴（阶梯状曲线）

生存曲线



两种手术治疗方式术后病人生存曲线的比较

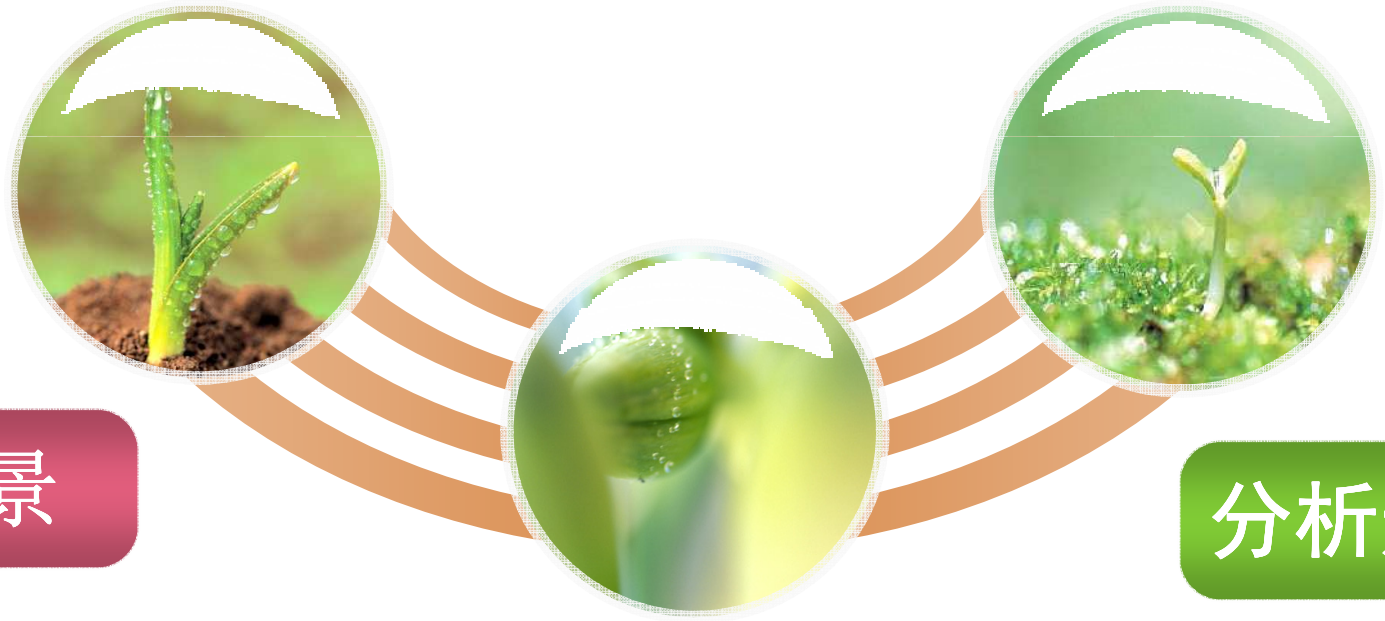
肿瘤生存分析研究的主要内容

- 描述生存过程：估计生存率（乘积极限法/**Kaplan-Meier**法）、平均存活时间、中位生存时间，绘制生存曲线
- 比较生存过程：生存曲线的**log-rank**检验
- 影响生存时间的因素分析：拟合生存分析模型，筛选影响生存时间的保护因素和危险因素。如：**cox**比例风险回归模型

- http://www.r-project.org/
- CRAN
- 选择其中一个镜像
- Contributed extension packages
- CRAN Task Views
- Survival

SS-RPMM

半监督回归分割混合模型 （**SS-RPMM，Semi-Supervised Recursively Partitioned Mixture Models**



背 景

优 点

分析过程

SS-RPMM背景（一）

- 非监督的聚类（常用分层聚类）用于肿瘤分型时，只利用了微阵列为基础的遗传学信息。因为没有利用临床数据，所以不能保证肿瘤的分型能够很好的预测结局。
- 半监督的聚类（**Semi-Supervised Clustering**）用于肿瘤分型时虽然结合了基因表达数据和临床数据，但当聚类的数目不明确时，特别是亚类之间存在较大的重叠时不适用。

SS-RPMM背景（二）

- **RPMM**类似于半监督的聚类的思想，用基于分层聚类的方法，可以在短时间内准确地估计出分类的数目。

- 因此，提出半监督回归分割混合模型（**Semi-Supervised Recursively Partitioned Mixture Models，SS-RPMM**）

SS-RPMM优点

- **SS-RPMM**可以利用微阵列为基础的遗传学信息和病人临床数据发现与病人生存相关的肿瘤分型
- **SS-RPMM**不需要知道肿瘤分类的数目

SS-RPMM分析过程

假如我们有**n**个样本**J**个基因位点以及临床生存结局的变量，我们要选出与生存最相关的**M**个（**M≦J**）基因位点。我们采用**Cox**比例风险模型计算**Cox**得分，用**Cox**得分测量基因表达水平和患者生存的相关性。 **Package**确定**M**个**Cox**得分绝对值最大的基因位点进行**RPMM**分析，画出相应的**heatmap**聚类图，并计算出风险比的估计值及**95%CI**。

应用实例

- 收集病例的基本资料和血样
- 提取DNA，测定基因表达情况
- 随访观察，获得生存时间
- 将全部数据分为：training和testing两个子集
- Training数据用于识别与生存时间有关的CpG位点，应用RPMM Package确定M个Cox得分绝对值最大的CpG位点进行RPMM分析
- 用模型去预测testing子集中的每一个观察对象，选出后验概率最大的类。

例1

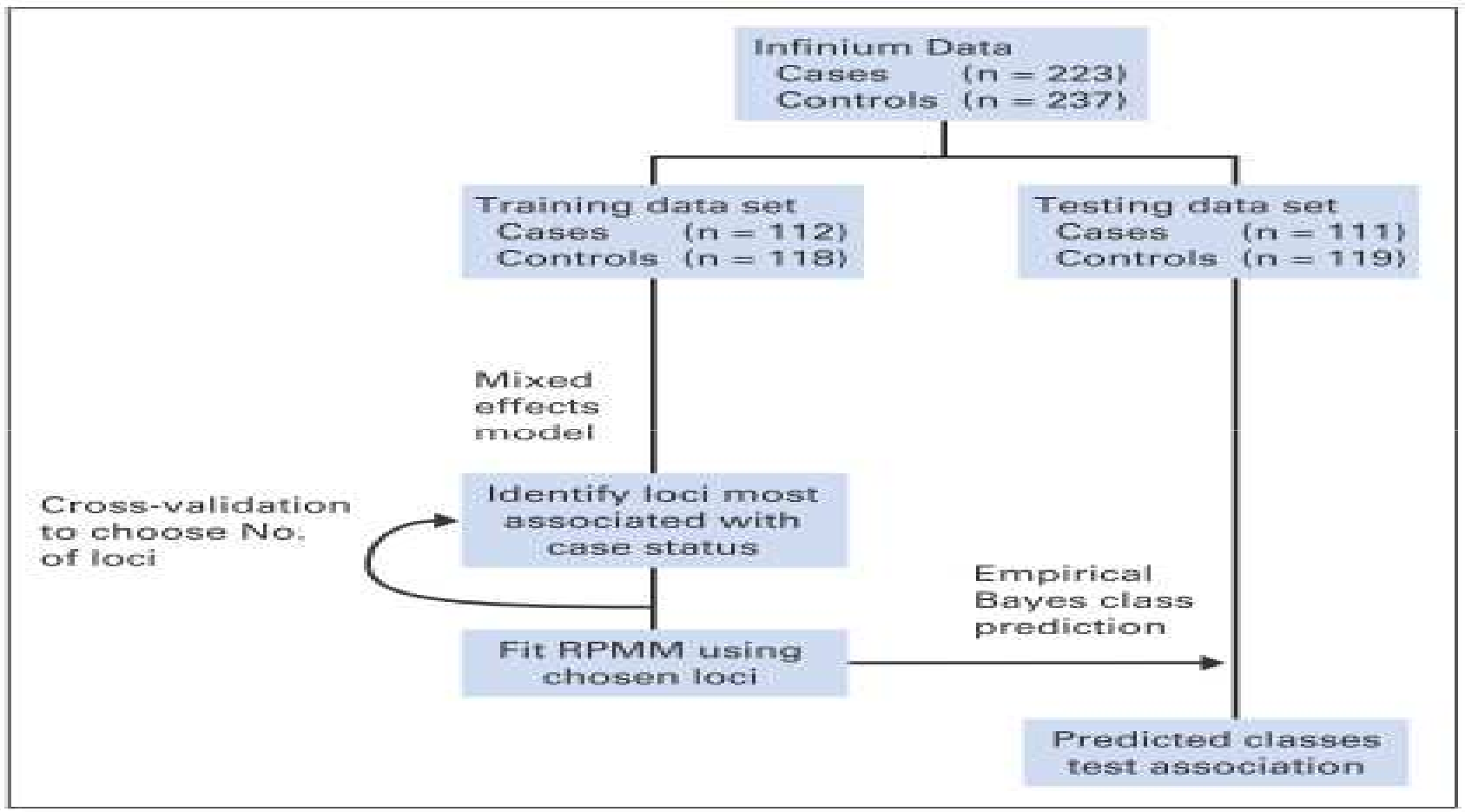


Fig.2. RPMM分析的示意图

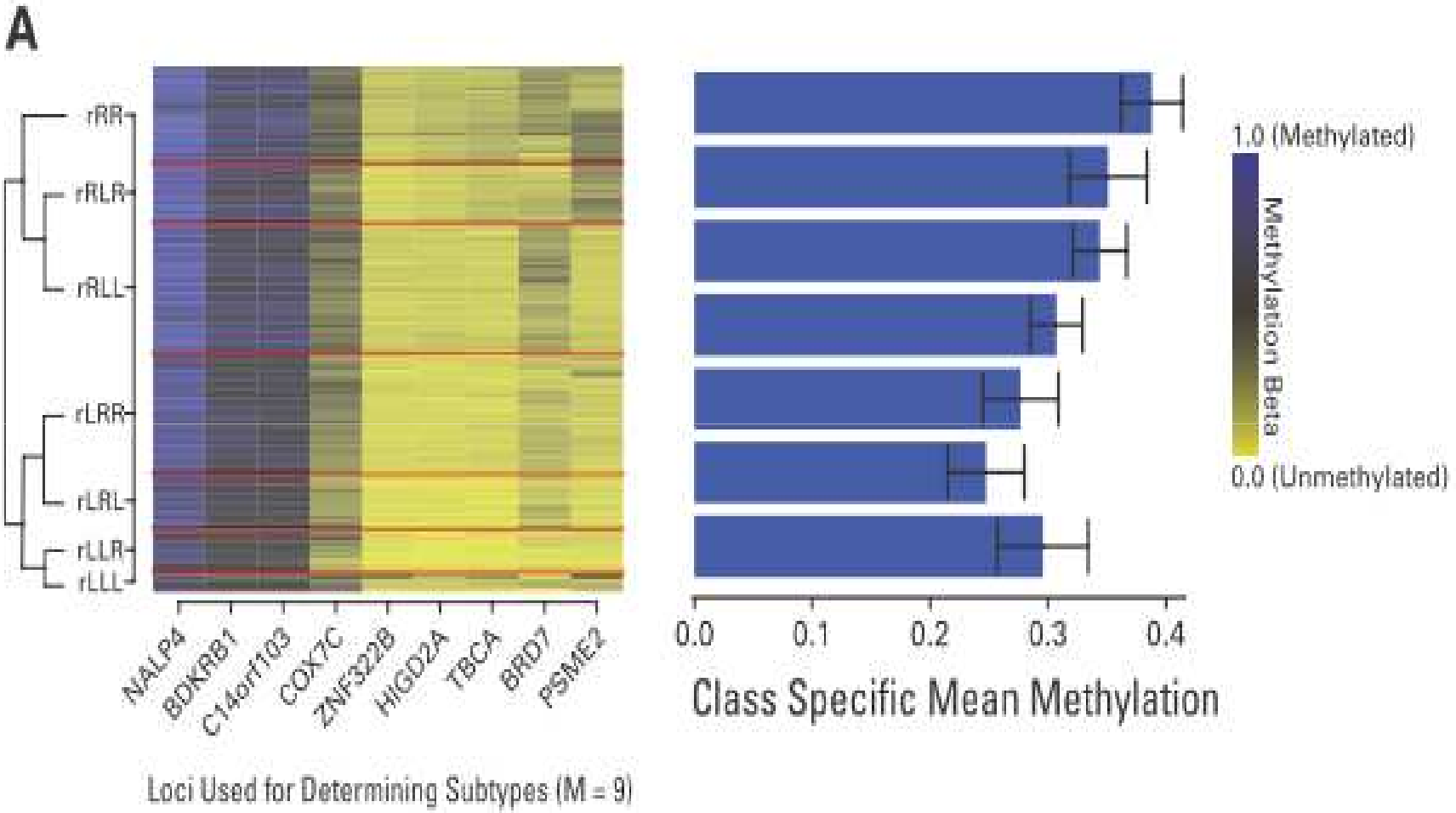


Fig.3. 9个与膀胱癌相关的甲基化位点的Heatmap聚类图

例2

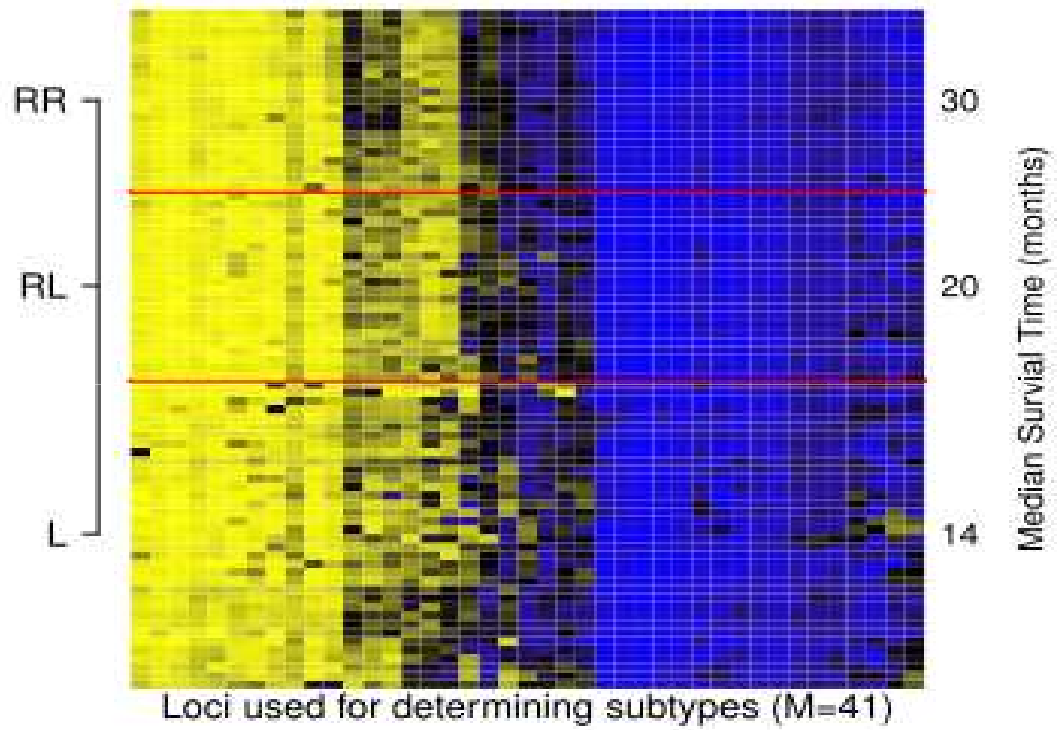


Fig.1. Heatmap of predicted class memberships for the observations in the testing set using the average beta values for the 41 loci with largest absolute Cox-Scores.

Covariate	HR Estimate	95 % CI for HR
RR vs L	0.35	[0.17, 0.73]
RL vs RR	1.81	[0.78, 4.30]
RL vs L	0.64	[0.31, 1.30]
Gender	0.56	[0.28, 1.13]
Age	1.03	[1.00, 1.06]

Table 1. The hazard ratio (HR) estimates

参考文献

- **Semi-Supervised Recursively Partitioned Mixture Models for Identifying Cancer Subtypes.（Koestler DC, Marsit CJ and Christensen BC, et al.2010)**

- **DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer.（Marsit CJ, Koestler DC and Christensen BC, et al.2011)**

谢谢大家！