

**北 京 科 技 大 学**

本科生毕业设计(论文)选题报告

癌症组织监测与分析系统的设计与实现

题目： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

化学与生物工程学院

学院： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

生物技术专业

专业： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

黄凯辉

姓名： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

41467024

学号： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

指导教师签字： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2018 年 03 月 30 日

目 录

[1 课题背景及研究意义 1](#_Toc510445464)

[1.1 课题背景 1](#_Toc510445465)

[1.2 研究意义 2](#_Toc510445466)

[2 研究内容，预期目标 6](#_Toc510445467)

[2.1 研究内容 6](#_Toc510445468)

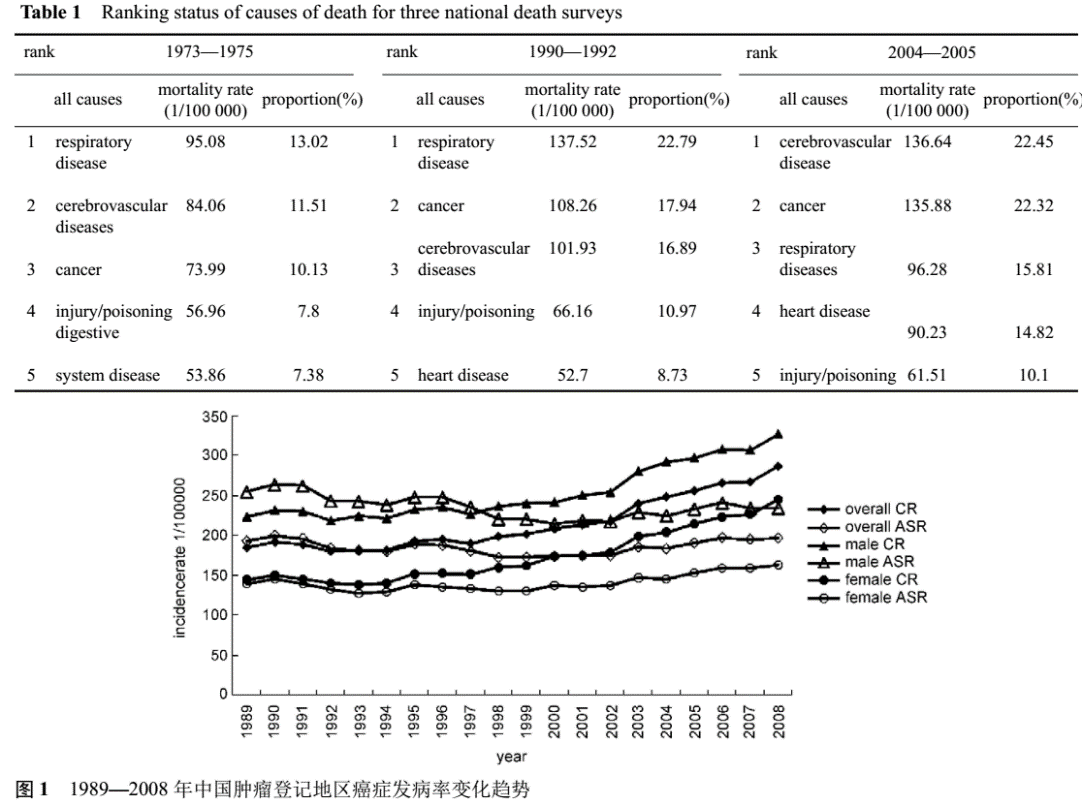
[2.2 预期目标 8](#_Toc510445469)

[3 研究进度安排 9](#_Toc510445470)

[参考文献 11](#_Toc510445471)

1. 课题背景及研究意义
   1. 课题背景

癌症是严重威胁人类生命和社会发展的重大疾病,近几十年来,随着疾病模式的转变和人口老龄化趋势,我国癌症负担日益增加,癌症防治面临严峻的形势。我国三次全国范围内的死因调查数据显示,近 30 年,中国癌症在死因中的构成比由 20 世纪 80 年代的 10.13% 上升至 22.32% (表 1),死亡率由 73.99/10 万上升至 135.88/10 万。在城市地区,癌症列居全死因的第一位,而在农村地区,列居全死因的第二位[1]。



二代测序(next generation sequencing)为快速检测病人遗传信息提供了可能。第二代测序技术也叫做高通量测序，有通量大， 可以同时进行平行测序，可定量的优点。其应用平台包括罗氏公司 454 测序平台、ABI 公司的 Solid 测序平台，以及 Illumina 公司的 Solexa 测序平台[2]。有学者利用二代测序鉴定出了食管鳞癌中发生频繁突变的驱动基因[3]，也有学者通过该技术对溃疡性结肠炎患者癌变组织进行全基因组和外显子组测序，进而研究不同体细胞突变与DNA复制修复功能之间的关系[4]。

然而测序需要从病人组织中提取样本,但通常癌症组织并不是只单纯地包含癌症细胞,它还有非常丰富的微环境。肿瘤微环境是存在于肿瘤内部和周围的非癌性细胞，主要包括免疫细胞，还包括成纤维细胞和包含支持血管的细胞。它们对肿瘤样本的基因组分析有很强的影响，并可能改变结果的生物学解释[5]。癌细胞样本提取时,这些微环境会和癌细胞一起被提取,并会伴随癌细胞一起被测序[6]。癌症细胞在癌症样本中的比例被定义为癌症样本的纯度[7]。

癌症基因组通常包含着大量体细胞序列拷贝数变异,这些变异主要由基因组片段扩增或删除造成。识别特定肿瘤基因组的基因组片段拷贝数变化,是癌症基因组研究的一个重要课题。要准确鉴定基因组片段拷贝数具有一定挑战,因为癌症片段拷贝数主要由两个因素混合决定,一是癌症样本纯度,即癌症细胞在癌症样本中所占比例,二是染色体倍性[8，9]。

因此，计算癌症样本纯度和癌症细胞基因组倍性是分析肿瘤样本的二代测序数据时的关键问题。

* 1. 研究意义

随着测序技术地发展，测序数据地快速增长，以及测序数据分析技术的积累，各种各样的癌症样本纯度算法被提出，并开发出了相对应的软件。基于基因组片段拷贝数变异和基于等位基因频率(突变位点的 B-等位基因(B-allele))的计算方法被相继提出。基于等位基因频率的方法有 PurityEst[10]和 PurBayes[11],主要是依赖于随着肿瘤样本纯度和肿瘤基因组倍性的不同,等位基因的频率会有所不同。基于拷贝数变异的方法有 CNAnorm[12]、THetA[13]、和 ABSOLUTE[14]等。然而这两种方法都有不同程度的问题,使用等位基因频率的方法由于数据量的问题会有较大的误差,而运用拷贝数变异的方法虽然较稳定,但却无法区分样本纯度和染色体倍性的补偿效应,即存在识别问题。以上基于片段拷贝数的软件都没有解决这一问题,CNAnorm 倾向于选择染色体倍性离二倍体最近的解决方案,ABSOLUTE 结合了其他的经验数据,THetA 则直接将所有可能结果都列出来了。

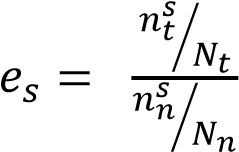
更优的方案应该是结合等位基因频率信息和片段拷贝数信息共同计算肿瘤样本纯度PyLOH[15], patchwork[16] 使用了基因组上杂合 SNV(单核苷酸变异)位点的频率信息,和基因组片段的拷贝数。PyLOH 一定程度上解决了“识别困难”的问题,可以更合理给出唯一解决方案。但是其准确性较差,特别是遇到基因组中存在亚克隆(subclone)的情况下。Patchwork 同时使用了两种信息,但是在计算基因型的中间步骤中,需要人工识别,人工判断的结果缺乏准确性,并且这种半自动化软件给应用带来很多不便。

这种情况下，我们希望能设计一种算法，能快速，准确地计算癌症样本纯度和癌症细胞基因组倍性。

1. 课题背景及研究意义

我们的研究计划采用癌症样本和匹配的正常样本的全基因组测序数据，对不同拷贝数片段的TRE和HGSNV的MAF分布构建混合高斯模型。计算癌症样本纯度和染色体倍性。其中将运用全基因组测序数据的 TRE 信息和 HGSNV 的 MAF 信息。TRE 基本反映了癌症样本拷贝数变异情况，HGSNV 的 MAF 信息基本反映了癌症样本的基因型。

Tumor Read Enrichment（TRE）， 癌症片段读长富集程度𝑒𝑠，指癌症样本中某一片段 S 内 read 数量与相应正常样本中对应片段 read 数量的比值，定义公式如下：

 （1）

上式中，𝑛𝑡𝑠和𝑛𝑛𝑠 分别表示在癌症样本中覆盖片段 s 的 read 数量和相匹配的正常样本中覆盖片段 s 的 read 数量，𝑁𝑡表示癌症样本全基因组测序获得 read 总数量，𝑁𝑛表示相应正常样本全基因组测序获得 read 总数量。

TRE 的差别主要来源于基因组片段的拷贝数差异，高拷贝数基因组片段内测序获得的 read 数量一定大于低拷贝数基因组片段测序获得的 read 数量，通过片段内 read 数量差异计算片段拷贝数差异是基因组拷贝数检测中的常用方法。但是大多数研究中，直接使用癌症样本片段内 read 数量除以正常样本该片段内 read 数量的比值（ratio）来进行 read 数量差异评估。我们预计使用公式（1）所示的 TRE 来评估片不同片段的 read 数量差异。传统方法计算所得 ratio 不仅受癌症样本纯度与染色体倍性的影响，还受到癌症样本和正常样本测序深度的影响，而 TRE 不会受到样本测序深度的影响。 单独依赖 read 数量差异，无法确定各拷贝数片段的基因型，更重要的是无法区分样本纯度与样本倍性的补偿效应。

Heterozygous Germline Single Nucleotide Variants（HGSNV）：杂合生殖系细胞单碱基变异，由于人类染色体属于二倍体，体细胞均由胚胎细胞发育而来，而生殖细胞中HGSNV 位点只有两种碱基类型 A 和 B，其中一种来源于父本，另一种来源于母本。

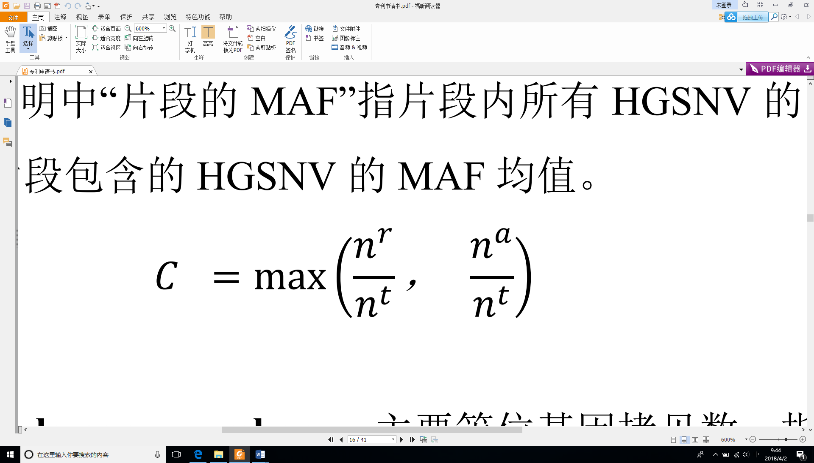
结合拷贝数差异片段内的 HGSNV 可以提供基因型信息，并帮助解决纯度与倍性的补偿效应，然而在前人的研究中，并没有高效利用 HGSNV 的方法，大部分方法采用枚举的方式将不同拷贝数片段可能对应的基因型一一列出，然后对排列组合的结果进行计算，挑选最可信的结果。这些方法共同的特点是，方法计算时间长，准确性差，对拷贝数很高或基因组变化较大的样本效果很差。本研究将依据 HGSNV 的 MAF 与 TRE 的混合高斯模型计算癌症样本纯度和染色体倍性，能显著减少计算时间，并提高计算结果准确率。

高斯混合模型（Gaussian Mixed Model）指的是多个高斯分布函数的线性组合，理论上GMM可以拟合出任意类型的分布，通常用于解决同一集合下的数据包含多个不同的分布的情况，或者是同一类分布但参数不一样，或者是不同类型的分布，比如正态分布和伯努利分布。

1. 研究内容，预期目标
   1. 研究内容

我们的研究目标是通过癌症样本和匹配的正常样本的全基因组测序数据,经过一系列计算获得不同拷贝数片段的TRE和 HGSNV ，并对TRE和HGSNV的 Major Allele Fraction(MAF)分布构建混合高斯模型,计算癌症样本纯度和染色体倍性。最终利用这一算法，使用rust语言编写出对应的软件。

MAF指主要等位基因(allele)分数,研究中使用的 HGSNV只有两种等位基因,一种等位基因与参考基因组相同,另一种与参考基因组不同。这两种等位基因型分数的计算方法为覆盖某一等位基因的 read 数量除以覆盖该位点总 read 数量的比值,MAF 就是两种等位基因分数中的较大值。计算公式如下所示,nr为包含与参考基因组相同等位基因的 read 数量,na为包含另一种等位基因的 read 的数量,nt表示覆盖该 HGSNV 位点的总 read 数量, C为该HGSNV的MAF值。MAF 是相对于 HGSNV的概念,本发明中“片段的MAF”指片段内所有 HGSNV 的 MAF 均值,“peak的MAF”指peak中所有片段包含 HGSNV的MAF均值。



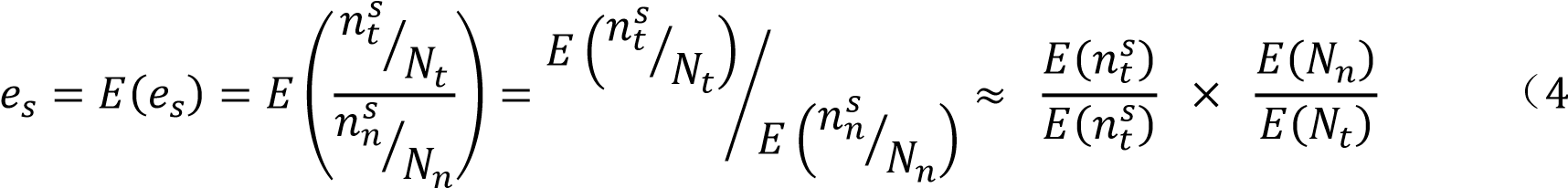
假设某癌症样本的纯度为𝛾，那么癌症样本中的正常细胞比例为1–𝛾。癌症样本中正常细胞的染色体倍性为 2，癌症细胞的染色体倍性为𝜅。那么癌症样本的染色体倍性𝜔如下公式（2）所示。

𝜔 = (1 − 𝛾) × 2 + 𝛾 × 𝜅 (2)

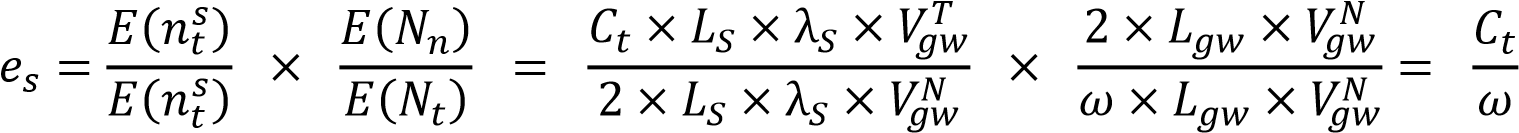
假设在癌症细胞中某一片段 S 的拷贝数为𝐶𝑆。那么癌症样本的片段 S 的拷贝数𝐶𝑡应该为如下公式（3）所示。

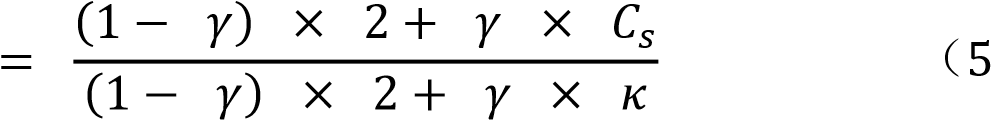
𝐶𝑡 = (1 − 𝛾) × 2 + 𝛾 × 𝐶𝑠 (3)

对于基因组片段 S，TRE 的计算方式如公式（1）所示。而 TRE 的期望值（expectation）𝐸(𝑒𝑠)的推导公式如下公式（4）所示，式中的𝑛𝑡𝑠、𝑛𝑛𝑠 、𝑁𝑛和𝑁𝑡与公式（1）中含义相同。

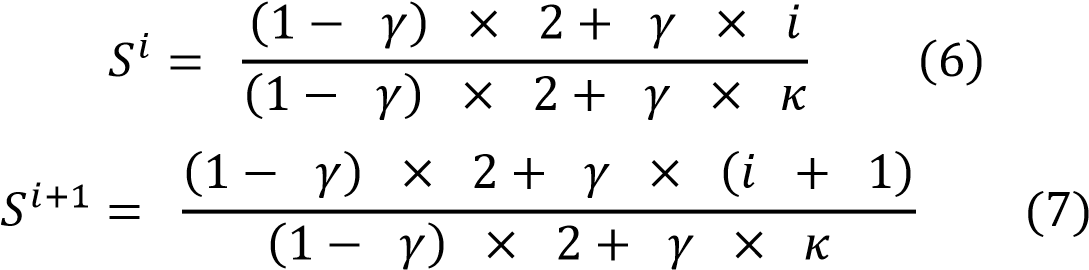


为了更进一步引出𝑒𝑠，本方法定义了一些帮助理解的参数。片段 S 的长度𝐿𝑆，人类参考基因组的长度𝐿𝑔𝑤，癌症样本的测序深度𝑉𝑔𝑤𝑇 ， 正常样本的测序深度𝑉𝑔𝑤𝑁 。那么片段 S 在癌症样本中测序深度为 ，片段 S 在正常样本中测序深度为 是指与片段 S 特性（如 GC 含量等引起测序深度偏好的特性）有关的参数，所以在癌症和正常样本中是一样的。进一步通过𝛾， 𝜅， 𝐶𝑠 来表示𝑒𝑠，如公式（5）所示。

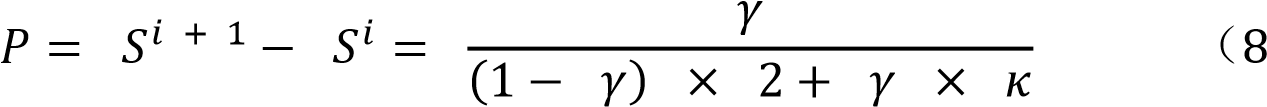




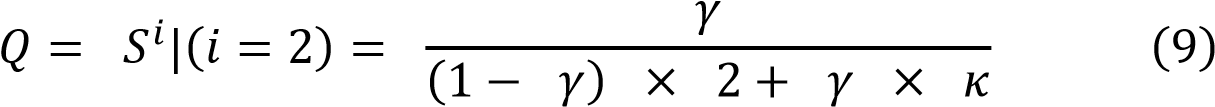
公式（5）中𝐶𝑠表示在癌症细胞中片段 S 的拷贝数，那么当片段 S 的拷贝数为 i 时和 i+1 时对应的 TRE 均值𝑆𝑖和𝑆𝑖+1分别如公式（6）和公式（7）所示：



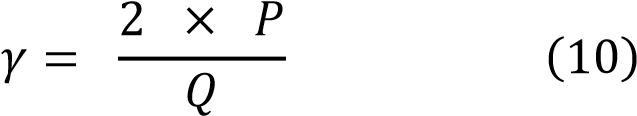
通过公式（6）和公式（7），对于相邻的拷贝数对应的片段，它们的 TRE 的差值 P 如公式（8）所示，可见 P 值的大小与片段具体拷贝数没有关系，它只决定于癌症样本纯度和染色体倍性。

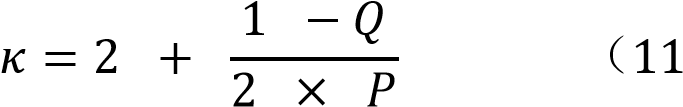


此外，对于 i = 2 即拷贝数为 2 的片段，它们的 TRE 值 Q 如公式 9 所示。



通过上述公式（8）和（9），可以解得癌症样本的纯度（𝛾）和染色体倍性(𝜅) 分别为：





通过以上分析可以得知，通过确定 P 和 Q 可以计算出癌症样本纯度𝛾和染色体倍性𝜅。

* 1. 预期目标

本研究的目的在于克服现有技术的缺陷,提供一种全自动、高效率、高准确性的计算癌症样本纯度和染色体倍性的方法和装置。本研究在癌症样本纯度和染色体倍性计算上具有广阔的运用前景。

1. 研究进度安排

第1周到第3周：确定题目，查阅文献，初步设计算法；   
第4周：撰写开题报告；   
第5周到第9周：编写癌症组织监测功能代码，翻译文献；   
第10周：中期检查，完善癌症组织监测功能；   
第11周到第13周：设计实现数据可视化系统，开始撰写论文；  
第14-15周：完成整个系统的设计与实现，完成论文撰写，准备答辩。

学生本人签字：

2018年 4月 2 日

参考文献

1. 曾磊, 王国平. 中国癌症流行病学与防治研究现状. 世界最新医学信息文摘, 2016(87): p. 36-37.
2. 张丁予, 章婷曦,王国祥, 第二代测序技术的发展及应用. 环境科学与技术, 2016(09): p. 96-102.．
3. 程彩霞, 基于全基因组测序的食管鳞癌基因组结构突变及靶基因功能验证研究.山西医科大学. 2016,．
4. 曹丹丹, 基于二代测序技术的溃疡性结肠炎相关性大肠癌变异谱分析. 中国科学院北京基因组研究所，2015,.
5. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications. 2013;4:2612. doi:10.1038/ncomms3612.
6. O’BRIEN J A. Introduction to information systems [M].7th ed.Burr Ridge, Ⅲ.:Irwin,1994.
7. Junttila M R, de Sauvage F J. Influence of tumour micro-environment heterogeneity on therapeutic response [J].Nature, 2013, 501(7467):346-354.
8. Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations 10 in human cancer [J]. Nature Biotechnology, 2012, 30(5):413-21．
9. Oesper L, Mahmoody A, Raphael B J. Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data [J]. Genome Biology, 2013, 14(7):R80．
10. Su X, Zhang L, Zhang J, et al. PurityEst: estimating purity of human tumor samples usingnext-generation sequencing data.[J]. Bioinformatics, 2012, 28(17):2265-2266．
11. Larson N B. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data [J]. Bioinformatics, 2013, 29(15):1888-9．
12. Gusnanto A, Wood H M, Pawitan Y, et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data [J]. Bioinformatics, 2012, 28(1):40-47．
13. Oesper L, Mahmoody A, Raphael B J. Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data [J]. Genome Biology, 2013, 14(7):R80．
14. Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer [J]. Nature Biotechnology, 2012, 30(5):413-21．
15. Li Y, Xie X. Deconvolving tumor purity and ploidy by integrating copy number alterations 30 and loss of heterozygosity [J]. Bioinformatics, 2015, 30(4):2121.
16. Mayrhofer M, Dilorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue [J]. Genome Biology, 2013, 14(3):R24.

指导教师意见

指导教师签字：

年 月 日