**CRC lncRNA分析流程**

**【研究意义】**

大肠直肠癌（Colo-rectal cancer）,又称为大肠癌、直肠癌、结肠直肠癌、结直肠癌、或肠癌，是源自结肠或直肠的癌症。CRC是第三大最常见的癌症类型，是全球癌症相关死亡的第三大主要原因。CRC的发生和发展是一个多步骤的过程。虽然科学家们一直都在尽最大努力理解CRC的复杂的发病机理，并改善其治疗效果，但是，CRC仍然是一种严重的疾病。因此，需要进一步理解CRC的精确机制，也需要开发新的诊断和预后生物标记。长链非编码核糖核酸（long non-coding RNAs,简称lncRNA），是指长于200核苷酸的不编码蛋白质的转录物。随着下一代测序的发展，越来越多的证据提示大多数lncRNA具有功能；比如LncRNAs在染色质组织、转录和转录后水平上调节基因表达，在不同的生物过程中扮演重要的角色，包括胚胎发育、细胞生长和肿瘤发生。目前已有研究表明，部分lncRNA与与肿瘤发生和癌症形成密切相关，可以作为肿瘤的标记物，因此找出CRC中lncRNA标记物，有助于CRC的检测，并为CRC治疗提供新途径，助力精准治疗研究。

本研究通过下一代测序方法获取到CRC病人（复发与未复发）基因测序数据，通过RNA-seq数据分析获取到基因表达数据，并得到lncRNA，通过差异表达分析得到正常组织与肿瘤组织之间差异基因和复发CRC患者和未复发CRC患者之间的差异基因和对应的功能；并找出其中的驱动lncRNA；并其中含有拷贝数拷贝数变异的lncRNA；通过实验验证lncRNA可能存在的功能；

**【研究内容与方法】**

我们拟先从以下几个方面开展研究：

一、测序与获取表达矩阵

收集获取20个人，总共40个测序标本：其中包含10个CRC复发患者病人的癌旁正常组织10例和癌组织10例，以及另外10个原发的CRC癌组织10例和癌旁正常组织10例；进行mRNA二代测序；并通过RNA-seq分析（STAR、cufflinks、RSEM等）获取到40样本的不同基因的表达信息。

二、基因分类与编码能力展示

通过注释文件，获取到样本中的蛋白质编码基因、lncRNA与未知的lncRNA,统计分类，并通过circos图展示出分类结果，从而得知患者基因中三者占比；为了展现不同基因编码能力的差异，通过coding potential 展示三种基因的编码图谱。

三、基因差异表达分析与功能富集分析

首先对正常样本与肿瘤样本做基因差异表达分析，找出两类样本间的差异表达基因，此部分基因或许与肿瘤转移相关。接着对复发患者与未复发患者样本做差异表达分析，找出两类样本之间差异表达基因，从而得到与复发相关基因，并取两个差异表达基因集的交集；并通过找出lncRNA附近的蛋白质编码基因，对蛋白质编码基因做功能富集分析，查看差异表达基因相关功能与同类；并找出差异基因中的驱动基因，从而筛选出与CRC相关的关键基因。

四、拷贝数变异分析

为了探究lncRNA 源头变异情况，对基因做拷贝数变异分析；从而得到有拷贝数变异的lncRNA，将获取到其与之前差异表达lncRNA的交集，从而得到有拷贝数变异的差异表达基因。

五、lncRNA实验验证

筛选出感兴趣的lncRNA，并通过敲除实验，查看lncRNA的缺失是否与肿瘤细胞迁移和凋亡相关。

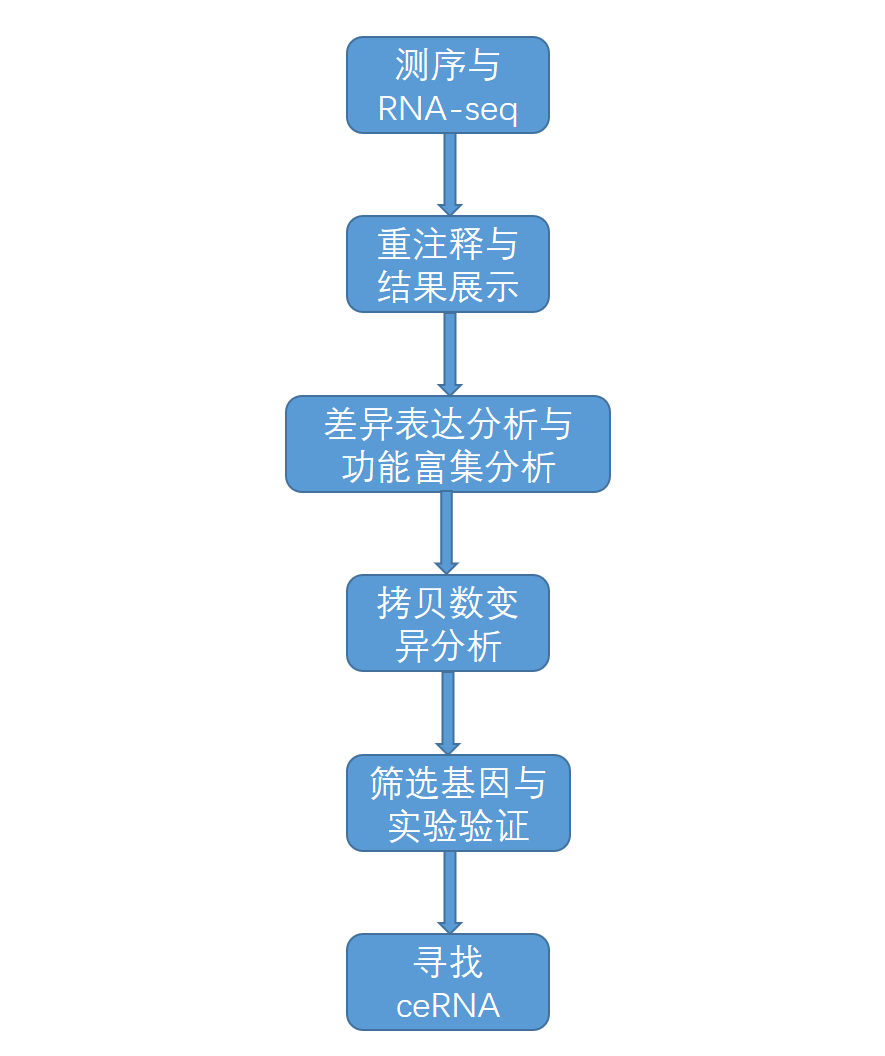
六、ceRNA

ceRNA(competing endogenous RNAs，内源性竞争RNA),其可以竞争性地结合microRNA调节基因的表达；为了查看lncRNA与哪些基因的变化，以及是如何影响基因的变化；提取与mRNA表达正相关并且共享microRNA的lncRNA,如果存在通过竞争性结合microRNA而影响基因表达，此部分lncRNA可能是ceRNA，从而有助于解释lncRNA的功能。

**【研究目标】**

1. 找出与CRC细胞发生发展和转移相关的lncRNA
2. 了解lncRNA是如何影响CRC细胞转移发展

**【技术路线】**



先获取到gencode、lncpedia、refgene三个数据库的gtf注释文件,并整合在一起；

使用Cufflinks中STAR和Tophat工具对质控后Fastq原始reads文件进行比对与组装；

选择长度大于200bp,Exon个数大于2的转录本，

使用Cuffcompare软件去除已知的基因，剩下未知的novel，其中包含编码和非编码的RNA。

使用CPAT、CNCI软件编码能力筛选出需要的novel lncRNA；

RSEM对reads做定量，获取RNA表达矩阵，使用DESeq、EdgeR差异表达分析。

