本课题组通过分析转座子（TEs）的重复特性、独特的结构特点以及motif特征，使用低频k-mer覆盖以及具有容错的比对扩展算法，实现了高精度的TE识别和注释，解决了De novo TE识别困难的问题。

2.1.1 高精度的TE识别方法研究

通过对TEs的重复特性、保守的motif特征以及结构特点进行分析，提出了一种具有容错特点的高精度TE识别和注释方法，HiTE。转座子（TEs）在基因组中广泛分布，通过产生多样性的等位基因、诱导结构变异和调节基因表达等多种方式对基因组的进化产生重要的影响。随着时间的推移，转座子在复制和拷贝自身时常常伴随着大量的删除和插入变异，同时其在基因组上的插入位点具有随机性，导致基因组中的TE呈现出复杂的序列模式，包括大量零散且高度分化的衍生TE序列以及嵌套的TE结构，使得准确的TE识别变得异常困难；同时TE的识别需要进行大量的比对运算，是一件非常具有挑战性的计算密集型任务。为了克服高度分化的TE造成的识别难题以及减少计算密集的比对运算，HiTE首先使用低频k-mer覆盖的方法初步定位基因组中的候选重复区，避免了整个基因组直接参与比对，减少了大量的计算资源。然后针对高度分化的TE，HiTE使用具有容错的比对扩展算法，能够跨过由于大量插入和删除导致的TE变异问题，并识别出具有粗边界的TE候选序列。最后根据不同的TE中所包含的特有的motif和结构特征对候选TE序列进行高精度的边界识别，并设计了一套高可靠的假阳性TE的过滤方法，极大减少了由于结构特征识别所带来的假阳性问题。HiTE在水稻、线虫、斑马鱼和果蝇四种模型物种上进行了全面的测试，与其他TE识别工具相比，取得了最高的识别精度以及复原了最多的“完美”金标准序列。