作品介绍-许怡冉

在父系遗传研究中，多个短串联重复序列（short tandem repeat，STR）位点的数值组合可以代表个体的 STR 单倍型，而单核苷酸多态性（single nucleotide polymorphism，SNP）的集合可以确定个体所属的 Y 染色体单倍群。由于二者的突变速率不同，STR 更能代表晚近的遗传突变，而 SNP 与早期的遗传突变关联较大。目前的数据库大多基于 Y 染色体 STR 建立，但高度降解的生物样本难以获取序列较长的 STR 信息，因此，可以使用 AI 从 Y 染色体 STR 推断单倍群。该模型基于随机森林算法，在训练集中通过有放回抽样的方式，形成子训练集，从中随机选择特征以构建决策树，多个决策树形成随机森林。将从文献中整理的 Y-STR 与单倍群数据库作为训练集和测试集，建立词典映射关系，训练模型并对其各项参数进行评估。此后，输入新样本的 Y-STR 信息，利用模型推断其所属单倍群。进一步扩充 Y 染色体数据库，可以提高模型的准确性。