北京基因组所实现对大样本全基因组数据的群体遗传学分析

文章来源：北京基因组研究所 发布时间：2015-11-26 【字号： 小 中 大 】

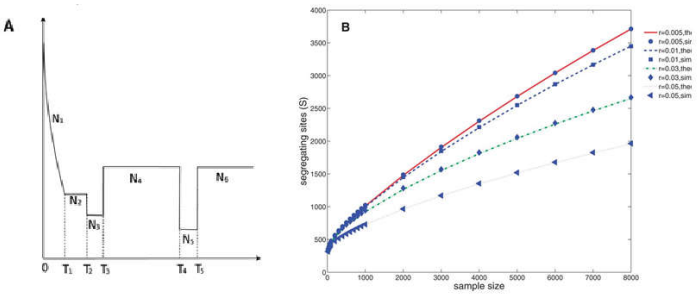
计算基因组学旨在发展理论和方法学，对基因组数据进行数据挖掘、提取信息。其中对遗传多态性，如基因频谱、单倍型结构等进行建模，利用群体基因组水平的遗传变异数据来推断群体的历史和变迁是群体遗传学的核心内容。传统的群体遗传学分析方法通常基于小样本数据，所推断的多为相对古老的历史事件，例如，被广泛应用的Li and Durbin (2011)的PSMC方法（递次式对偶溯祖方法）适用于2万年到300万年之间的群体大小变化，对一万年以内群体历史的推断精度很有限。该时间区间是这些方法进行参数推断的“盲区”。而人类在过去一万年左右从漫长的狩猎-采集文明逐渐过渡到农业、畜牧业和工业文明，深入了解人类两万年内的群体进化和变迁，对解析环境适应性、遗传性疾病的易感性和发病机制等都有重要意义。

日新月异的测序新技术正在产生海量的基因组序列数据。这些大样本或群体水平的测序数据为基因组时代的群体遗传学研究提供了前所未有的机遇。大数据蕴含的丰富信息使得更精细推断群体历史，包括1万年乃至几千年以内的群体变化成为可能。但另一方面，也给现有的理论和方法带来了新的挑战。多数现有的分析方法并不适用分析大数据：一方面是由于这些传统方法大多是基于随机取样方法，计算量太大；另一方面则源于一些公式在大样本条件下存在数值不稳定性。

中国科学院北京基因组研究所计算基因组中心陈华课题组针对以上问题提出了一个群体遗传学新算法（TNSFS）。该方法克服了大样本时的数值计算问题，首次实现了对大样本全基因组数据进行计算高效的群体遗传学分析，可用于检测群体的增长模式，有效推断一万年以内的群体大小变化的相关参数。新算法拥有若干计算上的优势：该方法给出解析形式的公式，不依赖于仿真，计算便捷高效，而且在大样本时无数值问题；具有很好的灵活性，能涵盖复杂的群体模型；此外，现代群体遗传学模型以Kingman溯祖理论为基本构架，理论上只适用于样本量远小于群体大小的前提下。当样本数目很大，甚至于接近群体水平时，Kingman溯祖理论会有严重偏差。新提出的方法即使在这种情况下，也具有很好的鲁棒性（robustness）。

该工作的研究成果于2015年11月发表于进化生物学期刊Molecular Biology and Evolution。该项研究得到中国科学院百人计划的资助。

论文链接



TNSFS方法能有效模拟复杂的群体模型（如图A所示，用于描述欧洲群体变化的Gazave模型），并且在不同参数值下的理论预测值与电脑仿真产生的结果吻合（图B）

（责任编辑：叶瑞优）