人类基因组中心粒测序、组装及评价的关键技术


艾云灿 教授

孟繁梅，艾汉南，艾云灿

中山大学生命科学学院，有害生物控制与资源利用国家重点实验室 广州 510275

人类基因组是测序和组装的质量标杆，但是迄今未能完成(以 Ns 占位)组装泛中心粒(中心粒及其邻近异染色质区域，centromeric & pericentromeric heterochromatin regions)。泛中心粒由大尺度重复序列构成。长期难以组装大尺度重复序列，是基因组学及生物信息学领域的关键技术挑战之一。

最新版人类参考基因组 GRCh38 中心粒序列并不是真实的线性序列，而是采用图论模拟方法模拟的，参见国际人类参考基因组(GRCh)小组于 2017 年 4 月 10 日在线发表于 *Genome Research* (doi:10.1101/gr.213611.116.)。美国加州大学 UCSC 基因组学研究所曾建立了该图论模拟方法，并模拟人类 Y 染色体中心粒的局部序列(~227 kb, GJ212193.1), 参见 2014 年 2 月 5 日在线发表于 *Genome Research* (doi:10.1101/gr.159624.113)。新近，该研究所又利用纳米孔测序技术(MinION)测定以往积累的 BAC，获得长读数文库，结合物理和遗传图谱，完成了首例人类 Y 染色体中心粒的真实序列(~301 kb, MF741337.1)的测序和组装。该项研究成果于 2018 年 3 月 19 日在线发表于 *Nature Biotechnology* (doi:10.1038/nbt.4109)。

本实验室重点关注如何评价人类参考基因组泛中心粒大尺度重复序列的质量。我们建立了全基因组指纹图谱几何学分析方法，反向追踪溯源和分析大尺度重复序列的组成(单体和复体)及其来源和质量。该方法克服了依赖先验性知识(包括物理和遗传图谱)的局限性，能够开展大数据挖掘驱动的自动分析，客观表征大尺度重复序列的组成及其来源和质量。例如，采用本方法分析最新版人类参考基因组 GRCh38 之 Y 染色体，通过可视化技术提取泛中心粒区域的大尺度重复序列(~1.30 Mb)(图 1)，并预测和提取其中的单体和复体序列，追踪溯源分析各个组成序列的同源物，进而开展聚类分析(图 2)，结果

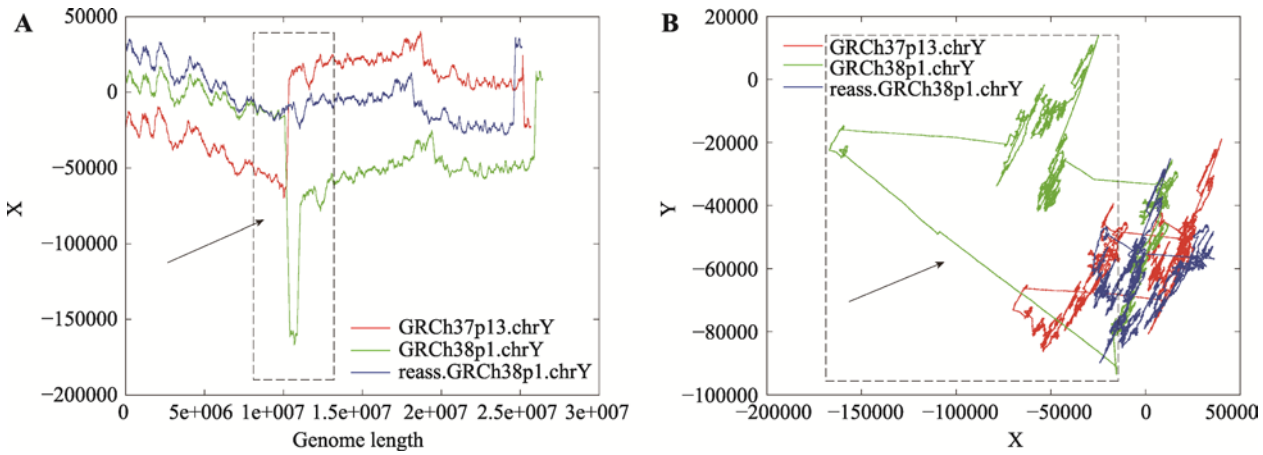

图 1 基于可视化展示提取人类 Y 染色体泛中心粒大尺度重复序列

Fig. 1 Visualization-guided extraction of large-scale centromeric and pericentromeric repeats of the human chromosome Y

A：X 坐标值随着基因组长度变化而出现的大直线；B：在 X 和 Y 二维平面上出现的大直线。GRCh38p1.chrY 和 GRCh37p13.chrY 分别是最新版及前一版的人类染色体 Y 基因组序列。reass.GRCh38p1.chrY 是定位删除大直线所对应的序列之后的版本。

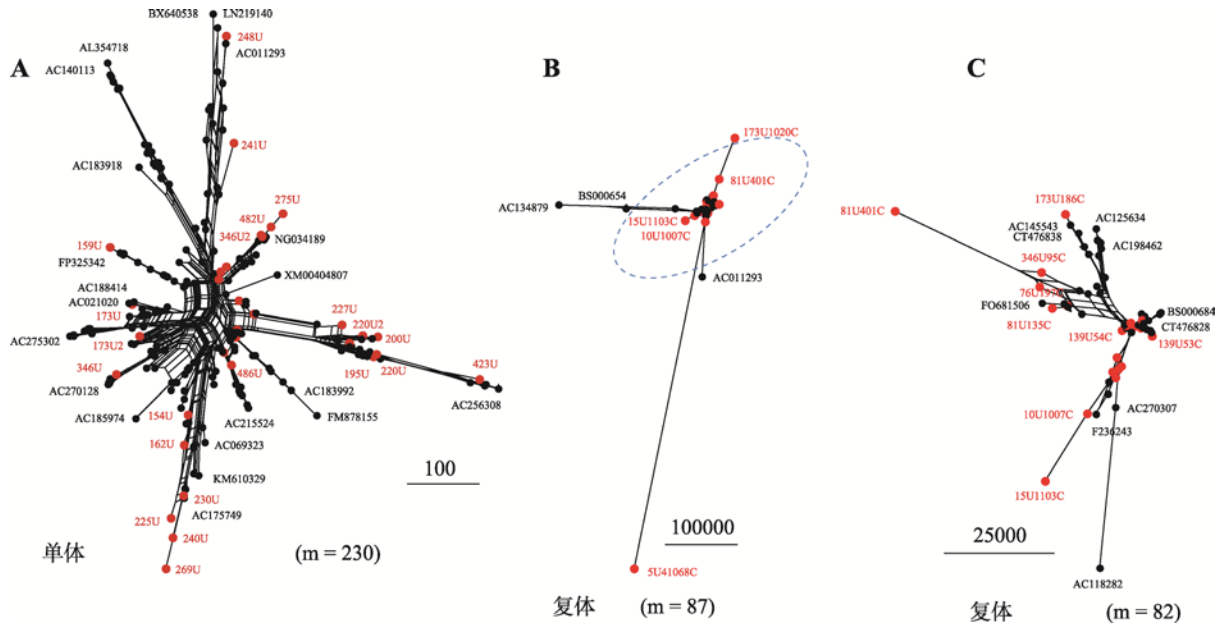

图 2 追踪溯源分析人类 Y 染色体泛中心粒模拟大尺度重复序列

Fig. 2 Tracking-back analysis of the modeled large-scale centromeric and pericentromeric repeats of the human chromosome Y

A：单体(红色)及同源物(黑色)序列的系统发育树；B：复体(红色)及同源物(黑色)序列的系统发育树；C：将 B 虚线圈中心放大。m 为序列个数。

显示所分析序列中的全部单体(图 2A)和绝大多数复体(图 2：B，C)的序列质量都是可靠的，但是存在个别孤儿复体序列(例如复体 5U41068，包含 41068 个拷贝的 5 bp 单体)(图 2B)，因为找不到同源序列，暗示其有可能存在“过度模拟”。该项研究成果于 2018 年 1 月 18 日发表于 *Scientific Reports* (doi:10.1038/s41598-018-19366-2)。

最近，本实验室又比较了人类 Y 染色体中心粒的真实序列(~301 kb, MF741337.1)与模型序列(~227 kb, GJ212193.1)，结果显示两者之间存在明显差异，大量单体的局部序列的方向相反(图 3，未发表)。图 3 仅展示了比对整齐的核心部分(~227 kb)。上述的研究工作受到天河二号超级计算机及国家超级计算专项项目(No. U1501501-201603534)资助。

综上所述，从图论模拟法建立模型序列，到利用无偏见数据挖掘分析法追踪溯源和反向评价大尺度重复序列的组成及其来源和质量，再到纳米孔测序法跨越长读数测定完整的真实序列，标志着中心粒大尺度重复序列的测序和组装在技术层面上取得了关键性突破。这必将有力推动完成更多的大型哺乳动物基因组(通常缺乏物理和遗传图谱)泛中心粒区域的测序、组装及下游分析。

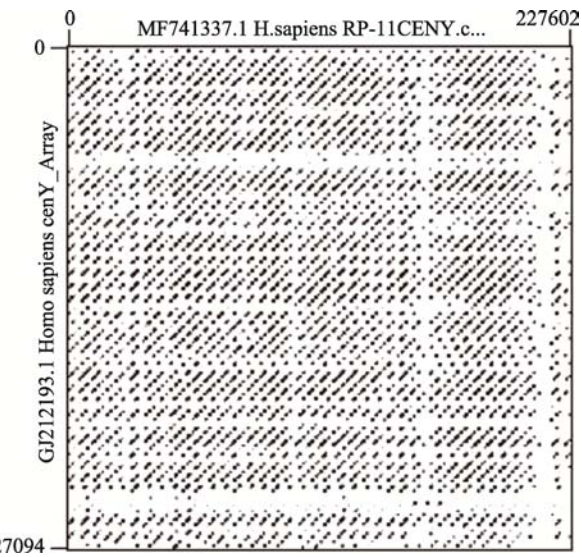

图 3 比较人类 Y 染色体中心粒的真实序列与模型序列

Fig. 3 Comparison between the native and modeled centromere sequences of the human chromosome Y