



染色质三维结构重建及其生物学意义

彭城[†], 李国亮[†], 张红雨^{*}, 阮一骏^{*}

华中农业大学生命科学技术学院, 作物遗传改良国家重点实验室, 武汉 430070

[†] 同等贡献

^{*} 联系人, E-mail: zhy630@mail.hzau.edu.cn; yijun.ruan@gmail.com

收稿日期: 2014-05-10; 接受日期: 2014-06-15

国家自然科学基金青年科学基金(批准号: 31200951)、作物遗传改良国家重点实验室开放课题(批准号: ZK201406)和中央高校基本科研业务费专项资金(批准号: 2662014PY001)资助项目

doi: 10.1360/052014-108

摘要 真核生物的基因组在细胞核中以染色质的形式存在, 染色质的功能与它的三维结构紧密相关, 例如, 基因组的复制、转录、调控、DNA 突变、长链非编码 RNA 的传播和胚胎发育等生物功能都是在细胞核的三维空间中完成的. 随着染色体构象捕获及其衍生技术与高通量测序技术的结合, 产生了大量的染色质交互作用数据. 根据这些染色质交互作用数据, 研究人员已经提出很多种方法来重建染色质的三维结构. 这些方法有助于在不同分辨率下系统地研究染色质的三维结构, 为更好地了解染色质的调控功能提供了结构依据. 本文总结了近期染色质三维结构建模方法的进展, 并探讨了其在研究染色质生物学功能方面的应用.

关键词

染色质三维结构建模
3C
Hi-C
ChIA-PET
染色质结构
染色质功能

真核生物的基因组在细胞核内以染色质的形式存在, 其中染色质是由 DNA、组蛋白、非组蛋白和少量 RNA 组成的复合物, 其基本单位是核小体, 是遗传信息和表观遗传信息的载体. 随着人类基因组调控元件百科全书计划(encyclopedia of DNA elements, ENCODE)的进行, 研究人员根据转录因子的结合位点和染色质的状态, 发现了大量的调控元件^[1]. 其中很大一部分的调控元件处在远离基因的区域, 很难确定它们的目标基因. 一个假设是, 这些调控元件通过空间的交互作用, 控制距离它们很远的基因. 要正确理解这些调控元件的功能, 必须考虑它们在细胞核三维空间中的构象. 例如, 基因表达不仅受基因启动子的影响, 而且受基因组线性距离远的调控元件(如增

强子)的调控. 研究表明, 这些基因组线性距离远的增强子需要在三维空间中与目标基因进行物理接触, 来启动相应的生物功能^[2-4]. II 型 RNA 聚合酶(RNA polymerase II, RNAP II/Pol2)参与的染色质远程交互数据展示了增强子-启动子之间和启动子-启动子之间的交互作用在基因转录中的调控作用^[4-6]. 另外, 遗传事件, 如 DNA 突变^[7,8]和染色体易位^[9], 也受染色质高级结构的影响. 在细胞功能方面, 与封闭的染色质结构域相比, 开放染色质结构域通常会在细胞分裂间期的早期被复制^[10]. 研究表明, 长链非编码 RNA(long non-coding RNA, lncRNA)Xist 在细胞核内的扩散也受基因组三维结构的影响^[11]. 此外, 染色质结构域之间的开关在动物肢体发育中起着重要作用^[12].

引用格式: 彭城, 李国亮, 张红雨, 等. 染色质三维结构重建及其生物学意义. 中国科学: 生命科学, 2014, 44: 794-802

Peng C, Li G L, Zhang H Y, et al. Reconstruction of three-dimensional structures of chromatin and its biological implications. SCIENTIA SINICA Vitae, 2014, 44: 794-802, doi: 10.1360/052014-108

染色质结构和功能之间的密切关系促使人们研究染色质结构的形成机制和生物学功能. 真核生物的基因组都是在细胞核的三维空间中发挥功能, 如基因组的复制、DNA 突变、DNA 修复、基因的转录和调控、长链非编码 RNA 的传播和胚胎发育等. 人类 DNA 序列连接起来的线性长度大约为 3 m, 但细胞核半径仅约 $3.5 \mu\text{m}$ ^[13]. 如何将超长的染色质存放在非常小的细胞核中是一个非常有趣的问题. 越来越多的以染色体构象捕获(chromosome conformation capture, 3C)^[14]为基础的衍生技术与高通量测序技术相结合, 尤其是 Hi-C^[15]和基于配对末端标签测序的染色质交互作用分析(chromatin interaction analysis with paired-end tag sequencing, ChIA-PET)^[16]技术的应用, 不但揭示了染色质三维(three-dimensional, 3D)结构对染色质功能的重要性, 同时提供了大量的染色质远程交互作用数据, 可以用于染色质的三维结构重建. 重建染色质的三维空间结构因而成为当前的研究热点^[17]. 最近, 研究人员基于 3C 衍生技术产

生的数据, 应用物理或优化的理论, 已经提出了多种染色质三维结构建模方法.

1 染色质远程交互作用的实验数据与染色质三维结构建模

目前有两类 3C 衍生实验技术来产生基因组范围的染色质远程交互数据: Hi-C^[15]和 ChIA-PET^[16]. Hi-C 和 ChIA-PET 技术都是使用福尔马林将细胞核内 DNA 和蛋白质进行瞬间固定, 然后将 DNA 打断成片段, 再将同一个大分子中空间距离近的 DNA 片段进行交联, 最后进行双末端测序, 统计所得的交互频率反映了空间距离近、线性距离远的 DNA 片段间的染色质交互关系. 所不同的是, Hi-C 产生的数据是所有蛋白质参与的染色质交互作用, 而在 ChIA-PET 的实验流程中, 加了一个步骤, 对特定的蛋白因子进行富集, 所以 ChIA-PET 产生的数据是某个蛋白参与的染色质交互作用. 图 1 总结了 Hi-C 和 ChIA-PET 产生

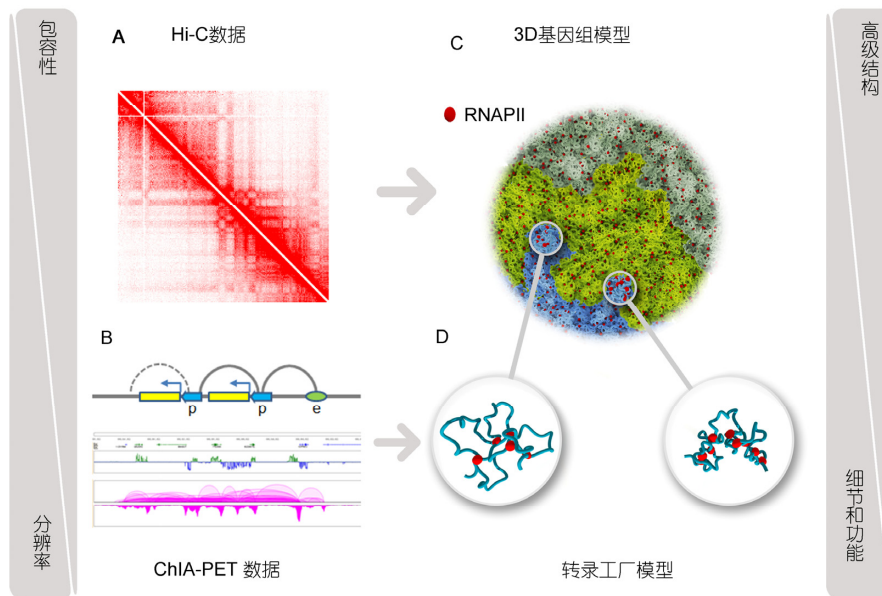


图 1 两类染色质远程交互作用数据产生的方法、数据和三维结构建模特征

A: 热图显示, Hi-C 数据能够反映染色质交互作用的层级结构, 基因组线性距离近的 DNA 片段总体上有较强的染色质交互作用; B: 示意图和浏览器的屏幕截图显示, ChIA-PET 数据能够反映精度更高的染色质交互作用, 包括增强子和启动子之间的交互作用, 以及启动子和启动子之间的交互作用, p 表示基因启动子, e 表示增强子; C: 根据分形球体模型, 线性较近的位点往往在三维空间中被组织在一起, 非特异性的力量, 如熵和结合分子也参与了染色质折叠; D: 根据 ChIA-PET 数据建立的局部染色质三维空间结构, 两个染色质结构域的高分辨率建模表明, 即使基于相同的组织原理, 这些结构域可以表现出不同的三维结构. 在本实验中, 其中的红点表示 II 型 RNA 聚合酶, 其他蛋白因子, 如 cohesin 和 CTCF (CCCTC-binding factor), 也具有类似的组织染色质结构的功能. 图两边的三角形总结了 Hi-C 和 ChIA-PET 数据的特征: Hi-C 产生的染色质交互作用数据更全面, 但是精度低, 因而可以用于低精度的三维结构建模; ChIA-PET 产生的染色质交互作用数据有特定蛋白因子参与功能, 精度更高, 可以用于高精度的三维建模

数据的特征. Hi-C 和 ChIA-PET 产生的数据都可以用热图来表示, 总的趋势是线性距离近的染色质之间有更强的交互作用, 线性距离远的染色质之间交互作用弱. 在热图中, 强的交互作用由深红色表示, 弱的交互作用由浅红色表示. 相对来说, Hi-C 产生的数据包容性更强, 涵盖了所有的染色质交互作用, 以及各种的随机噪声. ChIA-PET 产生的数据有更高的精度, 同时与特定的蛋白相关, 便于进行功能研究. 在 II 型 RNA 聚合酶参与的染色质交互作用中, 可以观察到增强子和启动子之间的交互作用, 也可以观察到启动子和启动子之间的交互作用.

基于 3C 衍生技术产生的数据进行染色质三维建模的基本原理如下: 染色质是由线性的核苷酸缠绕在组蛋白上, 形成核小体, 再经过多次折叠而成; 所以, 假设染色质(或者至少部分)可由“线珠结构”(beads-on-a-string)所组成的物理模型来表示. 在这样的模型假设下, 可以通过拟合实验获得的染色质相互作用频率数据来重建染色质三维结构^[18]. 以前的实验已经表明, 真核生物的每条染色质位于各自的区域, 不同染色质之间的交互有限, 而染色质内部由大量的结构域组成. 从更高的分辨率层面来看, 染色质之间的交互由染色质环组成. 由于染色质的这种组织模式在某些特性上与多聚物类似, 因此人们常使用多聚物模型来表示染色质结构^[19]. 同时, 在染色质结构重建中可以集成其他信息, 包括实验数据和物理理论. 虽然基于 3C 的实验提供了高通量数据来拟合染色质结构, 但是整合其他实验方法获取的结构信息, 如通过荧光原位杂交(fluorescence *in situ* hybridization, FISH)实验检测到不同 DNA 片段间的空间距离, 对改进或评估染色质三维重建的可靠性及稳定性具有重要作用.

图 2 简单地总结了用 3C 衍生数据和其他信息来重建染色质三维结构的 2 种方法. 第一种方法利用物理/多聚物模型来解释实验的观察结果^[19]. 在这种方法中, 表示染色质的珠子通常被放在一个网格化的立方体中. 如果两个珠子之间的距离低于设定的网格距离, 则定义它们之间产生了物理交互, 反之则没有物理交互. 染色质相互作用的频率既可以转换成势能并随后定义 3D 空间中的物理交互, 也可以用于评估物理模拟的结果. 经典的约束条件, 如黏接波动和体积排斥, 可以很自然地结合在势能模型中. 如果

适用的话, 其他结构信息也可以嵌入到势能模型(图 2). 公式(1)给出了一种典型的势能函数^[20]:

$$H = H_{Bond} + H_{Ev} + H_{IF},$$

$$H_{Bond} = \begin{cases} -\frac{1}{2}kR_0^2 \log(1 - (r_{ij}/R_0)) & r_{ij} < R_0 \\ +\infty & r_{ij} \geq R_0 \end{cases}$$

$$H_{Ev} = \begin{cases} \beta \left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 + c, & r_{ij} < r_{cut} \\ 0, & \text{其他} \end{cases}, \quad (1)$$

$$H_{IF} = \sum_{|i-j|>1} \varepsilon_{ij} r_{ij}^2$$

$$\varepsilon_{ij} = \begin{cases} \varepsilon, & \text{以概率 } P \\ 0, & \text{以概率 } 1-P \end{cases}$$

其中, r_{ij} 表示两个节点之间的距离, 参数 $k, \beta, \varepsilon, \sigma, c, r_{cut}, R_0$ 为反映系统特征的经验参数, 概率 P 代表节点 i 和节点 j 之间的交互频率. 需要说明的是, 不同的处理方式可以给出不同的势能模型. 这种方法具有一定的优势: (i) 可以通过已有的统计力学和相应算法进行计算机模拟, 来解决上述问题; (ii) 这种方法可以产生不同染色质三维结构的集合, 从而可以用于研究染色质结构的动态, 如染色质折叠和结构群体的异构性. 然而, 通过计算机模拟产生这样的三维结构群体是非常耗时的, 这主要是因为计算机模拟很容易陷入能量势垒的局部极小区, 难以跳到其他的能量区, 不能向最优化的能量区转移.

第二种方法用受约束的优化模型来整合不同的数据^[18]并重建染色质结构. 为了拟合 3C 衍生的数据, 首先要将染色质相互作用的频率转换为空间距离, 并将其用于构建优化模型的目标函数. 其他模型信息和先验知识, 如黏接波动、体积排斥和核半径, 可以作为优化模型的约束条件(图 2). 公式(2)给出了一种典型的优化模型^[21]:

$$(P_1, \dots, P_n) = \operatorname{argmin}_{i < j} \frac{[\operatorname{dist}(P_i, P_j) - d_{ij}]^2}{d_{ij}^2},$$

$$\begin{cases} \operatorname{dist}(p_i, (0, 0, 0)) \leq r \\ b_l \leq \operatorname{dist}(p_i, p_{i+1}) \leq b_u \\ \operatorname{dist}(P_i, P_j) \geq m, |i - j| > 1 \end{cases}, \quad (2)$$

其中, d_{ij} 表示由节点 i 和节点 j 交互频率转换得到的空间距离, P_i 表示节点 i 在三维空间中的坐标, $\operatorname{dist}(P_i, P_j)$

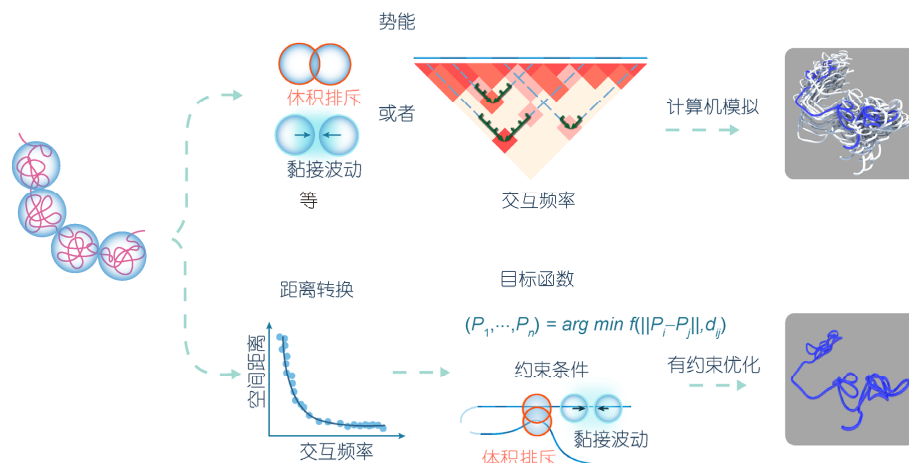


图2 基于3C衍生技术(如Hi-C或ChIA-PET)产生的数据来重建染色质三维结构的两种方法

在物理模型中, 势能是由经典的成分(黏接波动、体积排斥)或染色质相互作用的频率构成。在优化模型中, 染色质相互作用频率转换产生的空间距离用于建立目标函数, 而黏接波动和体积排斥作为约束条件

表示节点 i 和节点 j 在三维空间中的空间距离, 参数 r, b_l, b_u, m 为经验参数。因为3C衍生数据具有很高的噪音^[22], 因而如何将染色体交互频率转化成三维空间距离对于定义一个好的解决方案是至关重要的。常见的转换方式包括反比例转换和线性转换。反比例转换是幂函数的一种, 指数为-1。Zhang 等人^[23]证明, 幂函数的绝对值将随着节点在基因组上分辨率的减小(对应于节点代表的DNA片段增长)而减小, 所以, 在应用不同的分辨率进行基因组三维结构重建时, 应选用不同的转换函数。最近 Trieu 和 Cheng^[24]提出了一种新的目标函数, 建立了一个不需要将交互频率转换为节点距离的方法。在该方法中, 重建的三维结构是否满足节点的交互频率成为重要的条件; 然后通过梯度下降法, 找到三维结构的最优解。结果表明, Trieu 和 Cheng^[24]的方法产生的染色质三维结构与预期的染色质疆域(chromosome territory)相吻合。

由于在优化过程中约束条件是必须满足的, 所以约束条件的设置必需谨慎, 只有必须满足的前提条件或信息才能够被设置为优化模型的约束条件, 如上面提到的黏接波动、体积排斥和核半径。此外, 有些物种的染色质空间结构具有特殊性, 如在裂殖酵母(*Schizosaccharomyces pombe*)中着丝粒占据着空间特殊位置^[25]。这种系统特殊性也可以设置成约束条件来限制染色质三维结构重建。与第一种方法相比, 这种方法的优点是可以使用众所周知的优化理论有效地解决问题。同时, 在优化模型框架下, 研究

人员更容易理解三维结构重建问题, 而且不需要太多的统计力学背景知识。然而, 这种方法每次运行只返回单个的染色质结构, 使得群体结构分析变得困难。此外, 该解决方案存在一个优化中的普遍问题, 即容易陷入局部最优解。

2 染色质三维建模进展及其所揭示的生物学意义

虽然染色质三维建模的两种方法在实现过程上有所不同, 但是两者的目标相似: 获得准确的染色质三维结构, 从而用于理解染色质高级结构的形成机制或生物学功能。以下简要介绍如何从染色质三维建模来研究染色质结构的形成机制和生物功能, 包括从染色质折叠到染色质区域的变化。

在第一篇 Hi-C 文献^[15]中, 作者根据不同的物理模型, 在 10^6 bp 的分辨率下, 模拟了人类细胞系 K562 和 GM06990 中染色质的折叠情况, 产生了不同染色质结构的群体, 发现分形球体模型所产生的数据集和 Hi-C 染色质相互作用数据之间具有最好的拟合。这一结果表明, 染色质是按照一种良好的组织方式来进行全局折叠的, 其中线性临近的位点通常在 3D 空间中比较接近。在芽殖酵母(*Schizosaccharomyces cerevisiae*)和裂殖酵母中通过优化理论所预测的染色质三维结构也具有类似的全局性结构属性^[25,26]。

随之而来的问题是, 染色质在更高的分辨率下是否具有同样的组织方式. 一个对 5C 实验数据进行染色体三维建模的工作对解决这个问题提供了一些线索^[27]. 通过以人类 α 血球蛋白 (CHR16: 60002-55999) 作为模板进行结构分析, 发现该染色质区域也采用了分形球体模型方式来组织该染色质区域进行折叠. 这些在不同分辨率下的染色质三维建模研究表明, 染色质结构存在着层次化组织机制^[28]. 以前的模拟研究^[15]已经表明, 分形球体模型一般是无结的结构, 具有易扩展和重新聚合的结构优势, 促进了染色质对于不同生物学功能所需的结构重排.

在蛋白折叠中, 非特异性作用力, 如氢键、范德华力、疏水相互作用和电荷相互作用, 在塑造蛋白折叠结构中具有重要作用. 然而, 人的染色质在长度上比典型的蛋白高出了 7 个数量级, 并且染色质是一个高度复杂的复合体, 包含了 DNA、各种蛋白和小分子等, 那么什么是驱使染色质折叠和分隔的动力呢? 一个称为动态成环模型的计算机模拟程序^[29]对人类染色质的模拟表明, 基于扩散驱动的动力学模拟能够准确地抓住 Hi-C 染色质相互作用的模式, 表明这个动态的扩散可能是导致染色质分形结构的重要原因. 一个针对细菌的蒙特卡罗模拟也表明, 染色体不需要其他的机制, 可以仅通过随机扩散迅速地偏向细胞两极^[30]. 除了非特异性作用力, 结合在 DNA 上的分子也可能显著地作用于染色质折叠. 通过模拟染色质的组织蛋白, 如 CTCF^[31] 和转录工厂 (transcription factories)^[32], 另一种称为细绳和黏合剂转换的聚合物模型^[33]被用来研究结合分子在复杂染色质空间组织形成中的功能. 除了使用传统的高分子物理, 这个模型将结合分子定义为具有布朗运动和一定浓度的扩散粒子. 通过拟合 FISH 和 Hi-C 数据集, 这个模拟研究表明, 结合分子的浓度和结合能力在染色质折叠中也起到了关键作用. 这一结果与最近的转录因子 cohesin 实验结果高度一致^[34], cohesin 的缺失将显著地重塑染色质结构. 总之, 染色质三维建模表明, 非特异性力量和结合分子在形成染色质结构中共同起着决定性作用(图 1).

虽然染色质折叠研究表明, 临近的 DNA 位点可以由不同的驱动力被包装在一起而形成染色质结构域, 但目前还不清楚这些染色质区域是否在三维结构上具有相同的特性, 以及这些局部结构与染色质功能之间具有何种联系. 高分辨率的染色质三维建

模有助于研究这些悬而未决的问题. 然而, Hi-C 染色质相互作用数据有着显著的噪音, 而且在高分辨率下, 数据的内在随机性^[22]使得问题变得更加复杂. 研究人员试图在不同的分辨率下减少 Hi-C 数据的噪音, 从而更加准确地重建染色质三维结构. 在最近的一项研究^[35]中, 研究人员利用泊松回归, 减少了原始 Hi-C 数据的系统偏差, 随后这些归一化的数据作为贝叶斯方法的输入, 重建染色质的三维结构.

最近, 华中农业大学系统生物学团队提出了一种自动化的染色质三维建模流程, 称为 AutoChrom3D^[21]. 该模型分析了不同实验和染色质区域的 Hi-C 数据在高分辨率下的内在特征, 并利用这种内在特性对不同染色质区域和实验来源的 Hi-C 交互数据进行归一化处理, 从而实现了以前所未有的高分辨率来重建染色质的三维结构. 这使得该方法适合在局部区域内系统地研究染色质三维结构的特性和生物学功能. 该研究表明, 不同的染色质结构域表现出不同的三维结构, 这些差异通常与基因组/表观遗传特性和染色质的功能相关. 现在该团队正在将染色质三维建模流程和其他程序应用于 ChIA-PET 数据, 希望建立更详细的转录工厂三维模型(图 1D), 来说明转录调控的拓扑结构机理.

Zhang 等人^[23]提出了用半正定规划方法, 通过归一化的 Hi-C 数据来重建细胞核的三维结构. 图 3 是一个用 Hi-C 数据重建基因组三维空间结构的例子. 基因组序列用不断折叠的实线表示, 不同片段之间的距离表示它们在空间相互作用的可能性——距离越近, 不同片段之间相互作用的可能性越大. 线段上不同的颜色表示不同的染色质状态(chromatin states): 绿色表示“开放的染色质区”(open chromatin), 黄色表示基因高表达的区域, 蓝色和紫色表示“凝聚染色质区”(condensed region). 图 3 显示, “凝聚染色质区”处在一个紧密折叠的区域, 高表达的区域处在一个开放的染色质区. ChIA-PET 数据证实, 线性距离远但空间距离近的 DNA 片段间存在着 II 型 RNA 聚合酶参与的染色质远程交互作用.

图 3 中重建的染色质三维结构表明, 开放的染色质区域具有松散的三维结构, 而封闭的染色质区域具有较紧凑的三维结构. 研究表明, 和封闭的染色质结构(closed chromatin)比较, 开放的染色质结构和该区域的 GC 含量、基因密度和表观遗传状态是高度相关的^[36]. 空间距离近的染色质区域在细胞分裂周期

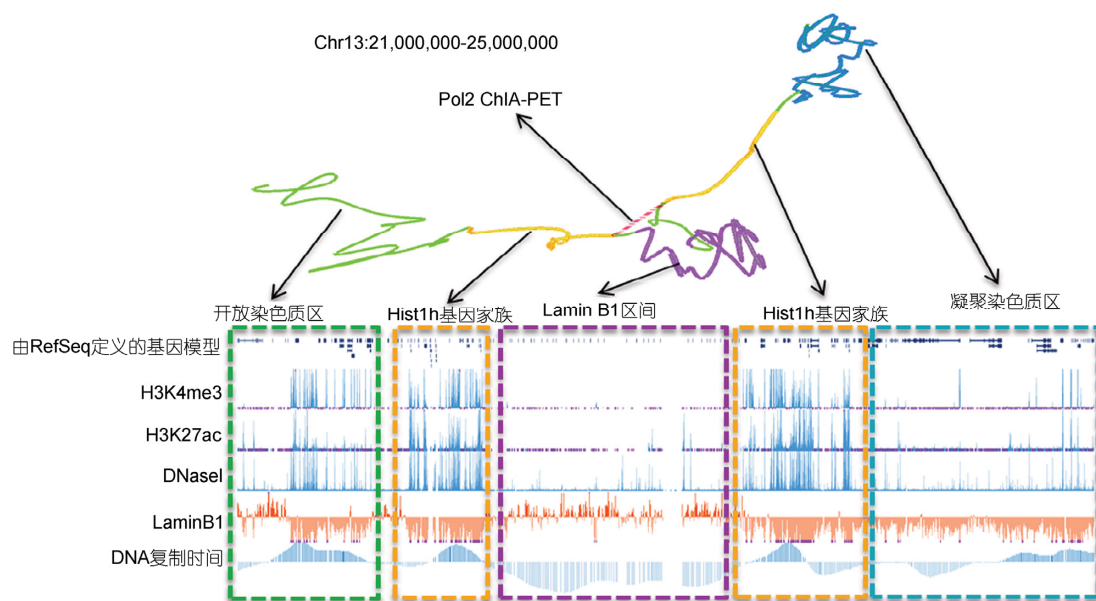


图3 由小鼠胚胎干细胞 Hi-C 数据重建的三维结构(mm10, chr13:21~25 Mb, 分辨率 40 kb), 及相关的基因、组蛋白、DNase I、LaminB1 和 DNA 复制时间信号^[23]

Pol2 ChIA-PET 数据说明, 从 Hi-C 重建的近距离 DNA 片段间存在 II 型 RNA 聚合酶介导的染色质远程交互作用

中具有类似的 DNA 复制时间^[10]. 组蛋白所标识的非活跃区具有富集的 LaminB1 信号, 在细胞分裂周期中具有较晚的 DNA 复制时间(图 3).

此外, 染色质的三维结构重排与疾病和发育不同时期的基因调控高度相关^[12,37]. 在老鼠 (*Mus musculus*) 的肢体发育中, 调控 *HoxD* 基因表达的基因拓扑结构域发生了变化, 对应于老鼠肢体发育的不同阶段^[12]. 在早年衰老综合征 (Hutchinson-Gilford progeria syndrome, HGPS) 疾病中, 一个位点的变异导致了染色质三维结构的变化^[37], 这很可能与该疾病的发生发展有紧密的联系.

3 讨论与展望

真核生物的染色质是按层次结构进行组织的, 这样便于局部结构的扩展和重新聚合. 虽然染色质结构域具有类似的组织形式, 但它们的具体三维结构有所不同. 染色质的三维结构差异和重塑都与染色质的功能有关, 如表观遗传状态与基因转录调控. 这些问题激发了研究人员对染色质三维结构的探索.

最近, 染色质三维结构建模的研究取得了快速发展, 为染色质折叠和区域形成的基本机制及其与染色质功能之间的关系提供了新的见解. 但是, 目前

的染色质三维结构建模研究距离普遍揭示生物学功能的目标还很遥远. 为了让染色质三维结构在生物学研究中获得更加广泛的应用, 需要提高染色质三维结构建模的精度、速度和易用性. 具体地讲, 染色质三维建模今后需要解决以下几个问题:

(1) 减少染色质交互数据的噪声, 提高数据的可靠性. Hi-C 交互数据包含相当高的随机噪声. 目前主要是通过计算染色体内部交互数据和染色体间交互数据的比例来估计 Hi-C 数据的信噪比. 为了减少 Hi-C 数据中随机交联的噪声, 研究人员已经开发一种 Hi-C 衍生技术, 称为栓系染色体构象捕获 (tethered conformation capture, TCC), 能够显著地提高染色体交互数据的信噪比^[38]. 但是 TCC 所产生数据和 Hi-C 数据一样, 只能近似估计而不能直接测量交互数据的信噪比, 并且该方法用于染色质三维结构建模的精度依然较低, 还达不到单个调控元件的级别. 在 ChIA-PET 实验中, 为了通过实验方法直接测量交互数据的信噪比, 设计了两个连接子 (linker), 然后通过嵌合连接子 (chimeric linkers) 的比例来估计 ChIA-PET 中随机交联的噪声比例. 此外, ChIA-PET 产生的数据与特定蛋白相关, 能够高精度地定位到调控元件, 达到单个调控元件的级别, 从而便于研究染色质交互和三维空间结构的生物学功能. 但是, 特定蛋白参

与的 ChIA-PET 交互数据不是染色质三维结构的完整反映, 在染色质三维结构重建中有偏差. 如何将高噪声的 Hi-C(或 TCC)数据和高精度的 ChIA-PET 数据结合, 将是染色质三维结构重建中的一个挑战.

(2) 数据的测序深度. 染色质交互数据的测序深度对三维建模影响很大. 以人类基因组为例, 大约有 6×10^9 bp (3×10^9 bp 的 2 倍). 如果每个调控元件的大小在 1000 bp, 人类基因组可以分成 6×10^6 个单元. 如果要保证任意两个单元之间有一个测序序列, 至少需要 36×10^{12} 条配对序列. 这是一个非常大的数字, 远远超出了目前的测序量(通常小于 10^9). ChIA-PET 技术只针对一个特定蛋白所参与的染色质交互, 所需要的数据量大大减少. 受到测序深度的影响, 为了基因组单元之间有足够的交互数据用于分析, Hi-C 数据分析精度通常在 1×10^6 bp. 如果分辨率要从 1 Mb 增加到 40 Kb, 测序深度要增加 625 倍. 如果测序深度不变而分辨率提高, 如将分辨率提高到 40 或 8 Kb, 那么目前的数据只能用于局部染色质三维结构重建, 而不能用于整体染色质三维结构重建.

(3) 染色质三维群体结构和平均结构之间的关系. 目前用于染色质三维结构重建的实验数据是从大量细胞中产生的, 并且这些细胞可能处在细胞周期中的不同时期. 通过这样的数据所重建的单个三维结构是所有用于测序细胞的染色质平均三维结构, 对该细胞群体的三维结构具有统计意义. 但是在真实的细胞中, 可能没有任何一个细胞具有这样的染色质三维结构. 目前已经有单细胞的 Hi-C 数据^[39], 但是这些 Hi-C 数据不足以重建染色质的三维结构. 此外, 有限的单细胞三维结构不足以反映染色质三维结构的群体特征, 不具有足够的生物学意义. 因此, 结合更多的传统 Hi-C 数据和高质量单细胞 Hi-C 数据对于研究群体和个体三维结构之间的关系及其生物学意义具有重要作用.

(4) 染色质三维结构的验证. 目前重建的染色

质三维结构难以从全局上进行直接验证. 最理想的情况是从一个细胞产生足够多的染色质交互作用数据, 用于染色质三维结构重建, 并且同时将这一个细胞用于 FISH 实验, 使用超分辨率显微镜观察细胞核内染色质的三维结构, 最后再将重建的染色质三维结构和显微镜观察到的三维结构进行比较, 从而验证通过染色质交互重建的染色质三维结构的可靠性和稳定性. 但是, 目前的技术不支持这样的实验, 这是因为同一个细胞不可能用于不同的实验, 并且 FISH 实验目前也不能高通量观察所有位点之间的交互情况. 在今后的技术发展中, 可以期望的结果是, 通过超分辨率的显微镜, 观察到更多的 FISH 结果, 然后和染色质远程交互作用数据重建的三维结构进行相关性分析, 从而在统计意义上评价和验证重建的染色质三维结构.

(5) 同源染色体三维结构重建. 真核生物通常都是多倍体, 如何在染色质三维结构重建中区分同源染色体是一个还没有解决的问题. 对于这个问题, 目前的解决思路是通过染色体中的杂合单核苷酸多态性(single nucleotide polymorphism, SNP)来区分不同的同源染色体. 但是, 染色体中的杂合 SNP 密度通常不足以区分不同的同源染色体, 为同源染色体三维结构建模引入了复杂性. 这个问题目前没有其他好的解决方案, 还需等待技术上的突破.

随着生物技术的进步, 希望在未来的研究中, 实验技术的改进和发展能够产生更高质量的染色质交互数据. 其次, 随着测序技术的发展和测序价格的下降, 可以对染色质交互作用进行更高深度的测序, 提供足够的数据用于染色质三维结构重建. 此外, 需要开发更快、更稳定和更合理的多尺度染色质三维模型来产生精确的结构, 包括染色质三维群体结构和平均结构. 这些进步将可以提供更深刻的关于染色质结构和功能的认识. 例如, 可以阐释染色质从一个状态到另一个状态的结构转变以及这种结构转变所导致的生物学功能变化.

致谢 感谢审稿人给出的宝贵意见和建议.

参考文献

- 1 The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012, 489: 57-74

- 2 Dekker J, Marti-Renom M A, Mirny L A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, 2013, 14: 390–403
- 3 Handoko L, Xu H, Li G, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*, 2011, 43: 630–638
- 4 Li G, Ruan X, Auerbach R K, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 2012, 148: 84–98
- 5 Kieffer-Kwon K R, Tang Z, Mathe E, et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. *Cell*, 2013, 155: 1507–1520
- 6 Zhang Y, Wong C H, Birnbaum R Y, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 2013, 504: 306–310
- 7 Lawrence M S, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 2013, 499: 214–218
- 8 Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun*, 2013, 4: 1502
- 9 Fudenberg G, Getz G, Meyerson M, et al. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat Biotechnol*, 2011, 29: 1109–1113
- 10 Ryba T, Hiratani I, Lu J, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*, 2010, 20: 761–770
- 11 Engreitz J M, Pandya-Jones A, McDonel P, et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science*, 2013, 341: 1237973
- 12 Andrey G, Montavon T, Mascrez B, et al. A switch between topological domains underlies *HoxD* genes collinearity in mouse limbs. *Science*, 2013, 340: 1234167
- 13 Kuse R, Schuster S, Schübbe H, et al. Blood lymphocyte volumes and diameters in patients with chronic lymphocytic leukemia and normal controls. *Blut*, 1985, 50: 243–248
- 14 Dekker J, Rippe K, Dekker M, et al. Capturing chromosome conformation. *Science*, 2002, 295: 1306–1311
- 15 Lieberman-Aiden E, van Berkum N L, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 2009, 326: 289–293
- 16 Fullwood M J, Liu M H, Pan Y F, et al. An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 2009, 462: 58–64
- 17 李国亮, 阮一骏, 谷瑞升, 等. 起航三维基因组学研究. *科学通报*, 2014, 59: 1165–1172
- 18 Marti-Renom M A, Mirny L A. Bridging the resolution gap in structural modeling of 3D genome organization. *PLoS Comput Biol*, 2011, 7: e1002125
- 19 Tark-Dame M, van Driel R, Heermann D W. Chromatin folding—from biology to polymer models and back. *J Cell Sci*, 2011, 124: 839–845
- 20 Mateos-Langerak J, Bohn M, de Leeuw W, et al. Spatially confined folding of chromatin in the interphase nucleus. *Proc Natl Acad Sci USA*, 2009, 106: 3812–3817
- 21 Peng C, Fu L Y, Dong P F, et al. The sequencing bias relaxed characteristics of Hi-C derived data and implications for chromatin 3D modeling. *Nucleic Acids Res*, 2013, 41: e183
- 22 Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 2011, 43: 1059–1065
- 23 Zhang Z, Li G, Toh K C, et al. 3D chromosome modeling with semi-definite programming and Hi-C data. *J Comput Biol*, 2013, 20: 831–846
- 24 Trieu T, Cheng J. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res*, 2014, 42: e52
- 25 Tanizawa H, Iwasaki O, Tanaka A, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res*, 2010, 38: 8164–8177
- 26 Duan Z, Andronescu M, Schutz K, et al. A three-dimensional model of the yeast genome. *Nature*, 2010, 465: 363–367
- 27 Baù D, Sanyal A, Lajoie B R, et al. The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, 2011, 18: 107–114
- 28 Sanyal A, Baù D, Marti-Renom M A, et al. Chromatin globules: a common motif of higher order chromosome structure? *Curr Opin Cell Biol*, 2011, 23: 325–331
- 29 Bohn M, Heermann D W. Repulsive forces between looping chromosomes induce entropy-driven segregation. *PLoS One*, 2011, 6: e14428

- 30 Jun S, Mulder B. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. *Proc Natl Acad Sci USA*, 2006, 103: 12388–12393
- 31 Phillips J E, Corces V G. CTCF: master weaver of the genome. *Cell*, 2009, 137: 1194–1211
- 32 Cook P R. A model for all genomes: the role of transcription factories. *J Mol Biol*, 2010, 395: 1–10
- 33 Barbieri M, Chotalia M, Fraser J, et al. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci USA*, 2012, 109: 16173–16178
- 34 Tedeschi A, Wutz G, Huet S, et al. Wapl is an essential regulator of chromatin structure and chromosome segregation. *Nature*, 2013, 501: 564–568
- 35 Hu M, Deng K, Qin Z, et al. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol*, 2013, 9: e1002893
- 36 Chambers E V, Bickmore W A, Sempé C A. Divergence of mammalian higher order chromatin structure is associated with developmental loci. *PLoS Comput Biol*, 2013, 9: e1003017
- 37 McCord R P, Nazario-Toole A, Zhang H, et al. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Res*, 2013, 23: 260–269
- 38 Kalhor R, Tjong H, Jayathilaka N, et al. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, 2012, 30: 90–98
- 39 Nagano T, Lubling Y, Stevens T, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 2013, 502: 59–64

Reconstruction of Three-Dimensional Structures of Chromatin and Its Biological Implications

PENG Cheng, LI GuoLiang, ZHANG HongYu & RUAN YiJun

National Key Laboratory of Crop Genetic Improvement, College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, China

Chromatin is the basic element of eukaryotic chromosomes that package the genome in the nucleus, and its function is closely related to its three-dimensional (3D) structure. For example, many biological functions such as DNA replication, DNA damage, DNA repair, gene transcription and regulation, spread of long non-coding RNAs, and embryonic development, are processed in and affected by 3D chromatin architecture in nuclear space. In recent years, chromosome conformation capture (3C) and its derivative technologies combined with high-throughput sequencing such as ChIA-PET and Hi-C, have generated extensive chromatin interaction data. With these data, researchers have proposed a number of ways to reconstruct the 3D structures of chromosomal folding, which help to systematically study the chromatin 3D structures at different resolutions and to provide better understanding of chromatin functions with structural basis. In this review, we summarize recent advances in the methods for chromatin 3D structure modeling, and discuss its implications in chromatin function.

chromatin 3D structure modeling, 3C, Hi-C, ChIA-PET, chromatin structure, chromatin function

doi: 10.1360/052014-108