研究内容

小鼠精子发生过程中的三维构象变化

胚胎干细胞中的 TADs（拓扑结合域）类别

摘　要

细胞核中的 DNA 是编码所有基因，以表达维持生命的细胞功能所必需的蛋白质的蓝图。了解细胞核中的正常的信息组织、存储和展开方式有利于人类健康事业。已有研究发现染色质构象这种信息组织形式的变化可能引起癌症等疾病， 也可能改变人体对病毒等感染物的响应。本研究主要研究小鼠精子发生过程中的三维构象变化以及胚胎干细胞中的 TADs（拓扑结合域）类别。

论文第一部分是利用 Hi-C 和其他组学技术研究小鼠精子发生的五个时期中染色体的构效关系。我们发现染色质环存在于粗线期前后的细胞中。结合 Hi-C 和 RNA-seq 数据，我们发现在精子发生不同阶段间转换了 A/B 区室的区域与减数分裂特异的 mRNAs 和 piRNAs 的表达活性有关。此外，ATAC-seq 数据表明， 染色质可及性本身不能决定粗线期细胞中 TADs 和环的消失。而 ChIP-seq 数据表明，粗线期细胞中的 CTCF 和 cohesin 仍然结合在原始生殖细胞中的 TAD 边界位置上，说明TADs 和环的动态变化也不能仅由CTCF 和 cohesin 的结合决定。

论文第二部分是对拓扑结合域的分类。根据边界上基因的转录水平、边界序列的染色质状态、TADs 大小、TADs 内互作矩阵的特征值大小将TADs 分为了七类，并发现了其中 SA 类 TADs 表达最活跃且在细胞间最保守；A11 类和 S22 可能参与干细胞分化和增殖。

关键词：染色质构象，精子发生，拓扑结合域，小鼠胚胎干细胞

ABSTRACT

The DNA in the cell nucleus is the blueprint that encodes all the genes and can decode into the proteins necessary for the function of the cells to sustain life. Understanding the right way of information organization, storage and depackage in the cell nucleus is conducive to human health. Studies have found that changes in the information organization of chromatin conformation may lead to cancer and other diseases, and may also change the human body's response to infections such as viruses. This study mainly studies the three-dimensional conformational changes in mouse spermatogenesis and the TADs (topological binding domains) types in embryonic stem cells.

The first part of the paper is the study of the structure-function regulation of meiotic chromosomes by Hi-C and other omics techniques in mouse spermatogenesis across five stages. We demonstrated that chromatin loops are present prior to and after, but not at, the pachytene stage. By integrating Hi-C and RNA-seq data, we showed that the switching of A/B compartments between spermatogenic stages is tightly associated with meiosis-specific mRNAs and piRNAs expression. Moreover, our ATAC-seq data indicated that chromatin accessibility per se is not responsible for the TAD and loop diminishment at pachytene. Additionally, our ChIP-seq data demonstrated that CTCF and cohesin remain bound at TAD boundary regions throughout meiosis, suggesting that dynamic reorganization of TADs does not require CTCF and cohesin clearance.

The second part of the thesis is the classification of topological binding domains. According to the transcription level of genes on the borders, the chromatin state of the borders' sequences, the size of TADs, and the first eigenvalues of the interaction matrix within the TADs, TADs are divided into seven categories. Type-SA TADs are found to be the most active and most conserved among cells; Type-A11 and Type-S22 TADs may be involved in stem cell differentiation and proliferation.

Key Words: chromatin conformation; spermatogenesis; topologically associating domains; mouse embryonic stem cells

目　录

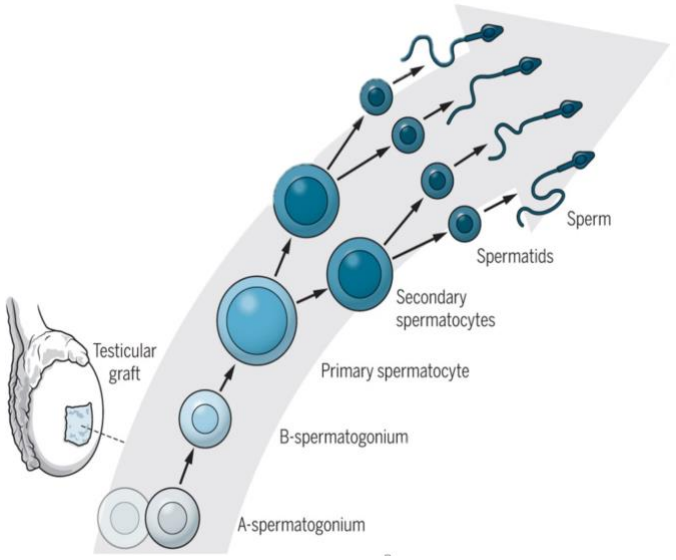
第 1 章　绪　　论

1.1　精子发生和染色质构象

1.1.1　男性不育与精子发生过程

许多男性幼儿时期接受了有性腺毒性的疗法，精原干细胞数量减少，导致生殖功能降低，甚至永久不育 [1]。这些性腺毒性疗法包括抗肿瘤疗法、针对良性镰刀细胞疾病或地中海贫血症的疗法。成年男性可以冷冻精子，但是青春期前的病人就无法如此 [2]。要保存青春期前的病人的生殖细胞，需要摘除或冷冻保存它们的睾丸组织。尽管冷冻保存组织已经可以实现了，但是目前还没有从保存的组织中获取精子的标准方法。早期的针对严重男性不育的疗法是对卵子进行ICSI（胞内精子注射）[3]。ICSI 与自然受精的最大区别在于其只需要少量的精细胞就可以实现受精。基于此，只要能得到少量精子，就可以治疗男性不育[4]。因此，研究精子发生过程十分重要。

相比其他动物，哺乳动物的生殖率低，后代数量少，并且生育的间隔时间长。因此，对哺乳动物需要保障稳定的配子发生过程，以产生高质量的配子。其中， 精子的



发育过程十分复杂（图1.1）。在精子发生过程中，SSCs（精原干细胞）首

图 1.1　精子发生过程

先分裂产生大量的未分化的精原细胞。在 RA（视黄酸）的刺激下，未分化的精

原细胞进行分化和几轮的分裂，产生精母细胞。精母细胞进行减数分裂，产生单倍体精子细胞。最后，精子细胞从圆形精子，经过变形形成高度分化，能使卵子受精的细长的成熟精子。在出生后个体的精子发生过程中，分裂的 SSCs 要么保

持自我更新的干性状态，要么进行分化，从而保证 SSCs 的数量，同时每天产生上百万数量的精子 [5-7]。

精子发生过程高度协调，需要受严格调控的阶段和细胞特异的基因转录。这种受严格调控的转录活动的实现依赖于特殊的染色质重塑、转录调控和睾丸特异的基因或亚型的表达。在发育过程中，三分之一的小鼠基因组在睾丸中差异表达。从出生到成年期间，大约一半的基因组在睾丸中表达。还有约 2%-4% 的小鼠/大鼠表达组的基因在睾丸特异表达，这些基因也参与调控精子发生过程。

在哺乳动物中，一般由染色体决定个体的性别。如果 Y 染色体缺失，早期性腺会发育为卵巢[8]。缺少 Y 染色体的小鼠在过表达Y 染色体相关的基因或生精的非 Y 染色体相关基因时，能产生男性生殖细胞（圆形精子）。在雌性（XX）小鼠中，尽管作为雄性生殖细胞的原始生殖细胞能促进睾丸环境，但这些细胞仅能维持到出生后两天 [9]。因此，基因因素对于传代和雄性生殖细胞的维持十分重要。

小鼠精子发生过程中的重要事件有：

第一次减数分裂前期中发生染色体重组时的 DNA 双链断裂能选择性地产生同源染色体间的 COs（交叉），从而产生遗传多样性，并保证染色体传递的准确性 [10]。在许多物种中，重组主要发生在被称为重组热点的基因组区域。在部分哺乳动物，如小鼠和人类中，热点的位置由特别的组蛋白甲基转移酶 PRDM9 决定[11]。PRDM9 是减数分裂特异的、DNA 结合锌指蛋白，其特异甲基化组蛋白 H3 的第 4 和第 36 位赖氨酸[12]。这种 H3K4me3/H3K36me3 双阳性特征能协

助这些位点的重组。在缺失 PRDM9 时，减数分裂 DSBs（双链断裂）的数量正常，但更多地位于 H3K4me3 信号富集的功能元件，如启动子。*B6.Prdm9*$^{-/-}$（缺失 *Prdm9* 基因的 C57BL/6 小鼠）的异位 DSBs 的修复异常，导致生殖细胞停滞在第一次减数分裂前期，从而引起不育 [13]。在男性人类中，*PRDM9* 的几种点突变与非阻塞性精子缺少有关 [14]。在人和小鼠中，也存在不依赖 Prdm9 的启动重组通路。在雄性小鼠中，常染色体上的重组热点是 PRDM9 依赖的，但在性染色体的假常染色体区域中的重组热点是由 PRDM9 非依赖的重组启动通路激活的 [12]。这表明在哺乳动物减数分裂过程中，PRDM9 依赖和 PRDM9 非依赖的重组通路先后被激活，并可能在 PRDM9 失活或缺失的情况下，满足生殖所需的功能。

重组的实现由 SPO11 参与，随后再次剪切形成 3’单链 DNA[15]。这些单链被

RAD51/DMC1 重组酶、单链 DNA 结合蛋白、RPA（复制蛋白 A）包覆 [16]。RPA 蛋白与另外两种减数分裂特异蛋白，MEIOB 和 SPATA22 互作，促进减数分裂重组 [17]。通过复杂且严密调控的链侵入、单端侵入、双端捕获和 dHJ（双 Holliday 连接体）的形成和解连，只有少部分（约 10%）的 DSBs 最终形成 COs[18]。但

一个同源染色体对至少需要一个COs。CO 形成或分布错误会导致生殖细胞丢失，非整倍型和胚胎发育异常，从而导致流产或不育 [19]。ZMM 蛋白也是促进 COs 和联会复合体形成的蛋白。其中 Spo16 是非同源染色体配对和 CO 形成所必须的。

在精子发生的最后，成熟精子中的 DNA 由带正电的小分子精蛋白紧密包装 [20]。参与精蛋白磷酸化修饰的分子有如热激蛋白 Hspa4l。Hspa4l 能保证精蛋白 2 特定丝氨酸位点的去磷酸化。缺失该蛋白的小鼠产生的精子的头部形状异常，具有不育表型 [21]。表达不可去磷酸化的组蛋白 2 变体能回复 Hspa4l 缺失小鼠的不育表型。总之，高度调控的精蛋白是成熟精子具有正常功能的必要条件。

对逆转座子的调控对于生殖细胞来说尤为重要。逆转座子占据了哺乳动物一半的基因组位置，其活性能损伤遗传材料，影响生育力以及后代的适应度 [22]。在哺乳动物中，生殖细胞会经历表观重组。其中，在出生时的精子发生过程中， 小 RNA 介导的 DNA 甲基化会形成持续终生的逆转座子表观抑制 [23]。piRNAs

（Piwi 蛋白互作 RNAs）是逆转座子转录本剪切的产物，能通过同源识别引导这些逆转座子启动子处的 DNA 甲基化 [24]。哺乳动物特异地进化出了无催化活性的 DNMT（DNA 甲基转移酶）辅酶，DNMT3L[23]。该辅酶作用于 piRNA 通路下游。DNMT3L 或 PIWI 通路蛋白的失活会导致低甲基化和逆转座子的重新激活，导致减数分裂失败，精子缺乏，以及以小的睾丸大小为特征的男性不育 [25]。

1.1.2　染色质构象及其与基因表达的关系

在高等真核生物中，间期染色体在细胞核内组织，它们的包装和折叠造成了基因组位点间的动态互作[26]。在细胞退出有丝分裂期后，核重新形成，每条染色体解凝缩，回到它们的领域组织 [27]（图1.2A）。基因组几乎没有缠结，在染色体间存在互作，有的在同一领域内，有的在域间混杂。在每个领域内，染色体空间区室化：早期的显微镜观察发现活性染色质域和非活性染色质域分别聚集形成真染色质A 区室和异染色质B 区室[28]（图1.2B）。染色质内的互作富集于同样类型的域之间的互作（A-A 和 B-B 互作）。利用全基因组染色质互作测序方法，比如 Hi-C，人们发现在百万碱基分辨率下，从反应互作频率的测序数据可以将染色体分为 A/B 区室。这种严格地讲染色体二分为区室的做法只是粗粒估计，但 A/B 区室可与活性的真染色体区域和非活性的异染色质区域分别对应。对 GM12878（人类淋巴母细胞系）的高深度的 Hi-C 数据分析，可进一步将 A/B 区室分为五个亚区室：A1,A2,B1,B2 和 B3[29]。这些亚区室内的基因表达和组蛋白修饰等功能特征各不相同。

在亚百万碱基尺度上，染色质组织形成倾向于在域内互作的 TADs（拓扑结合域）[30]（图1.2C）。TADs 在不同细胞类型中较为稳定，在物种间高度保守，常被

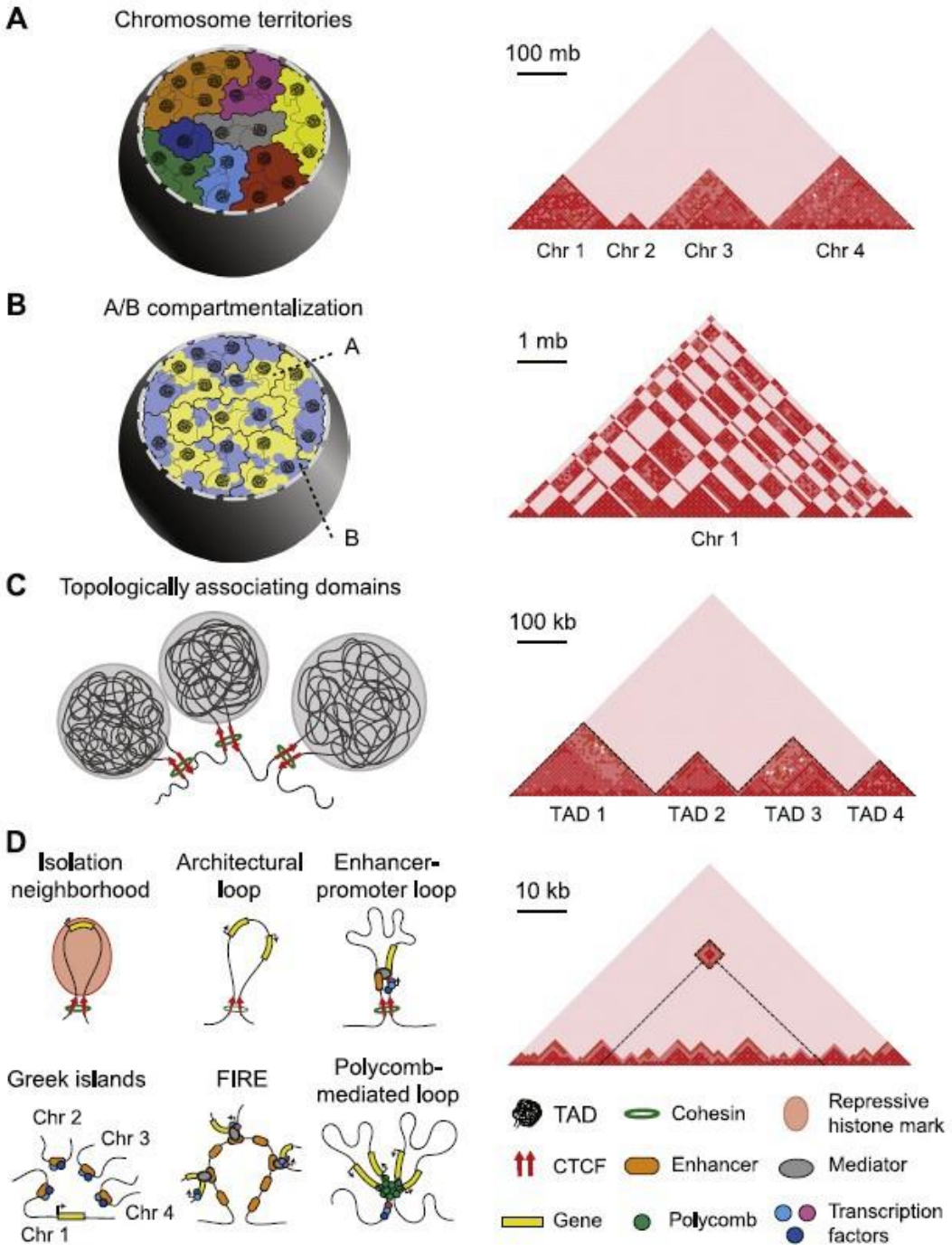
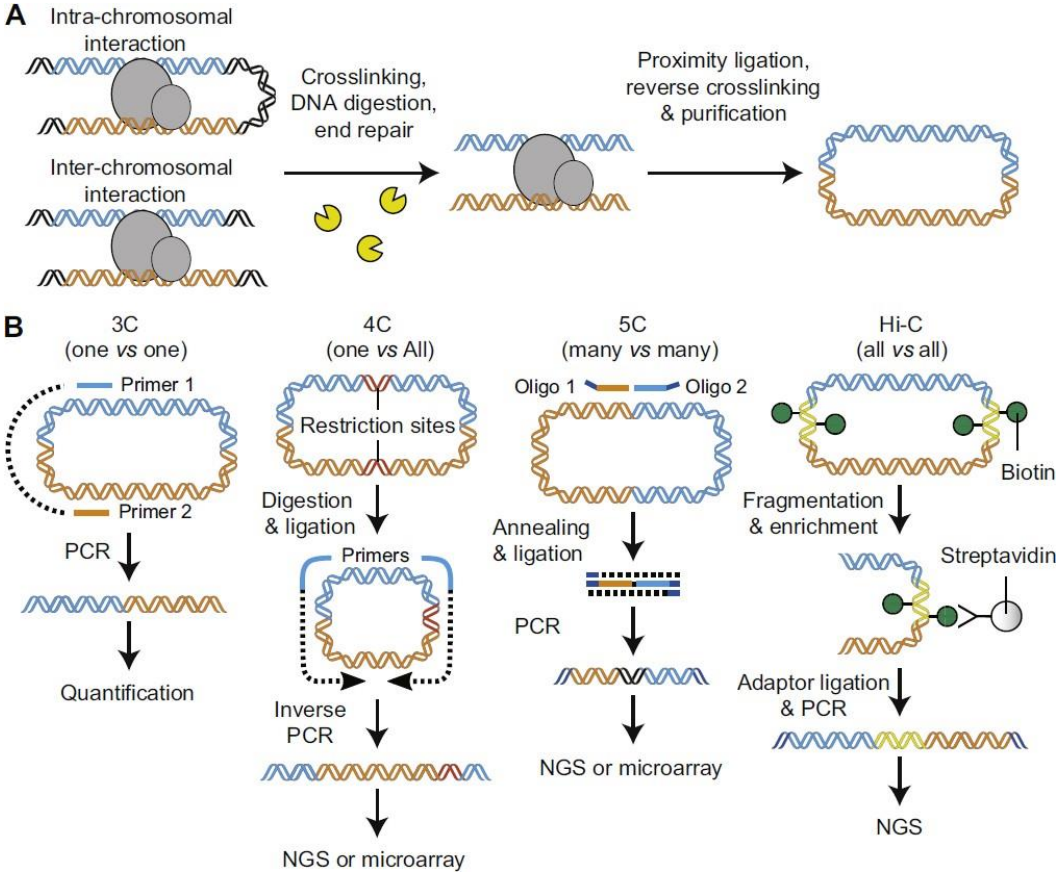


图 1.2　染色质构象

视为染色体折叠的基本单位。TADs 的形成主要依赖两种高度保守的因子：CTCF 和 cohesin 的相互作用[31]。Cohesin 是环状复合物。CTCF 能使用其 11 个锌指域的特定组合与其他蛋白互作并结合基因组的特定序列。Cohesin 复合物在环内拉染色质片段形成染色质环（图1.2D），造成染色质“挤压”，形成 TADs[32]。DNA 的挤压持续进行，直到 cohesin 碰到通常是 CTCF 结合位点的物理障碍（“TAD

边界”）。CTCF 结合模序不是回文序列，它们的方向对控制挤出的环十分重要。对向的 CTCF 结合模序通常位于小环的端点上，而同向的 CTCF 模序通常出现

在 TAD 边界上。但并非所有的 TAD 边界都是CTCF 依赖的，在一些高转录的基因区域，cohesin 依赖的环挤出过程会被 RNA 聚合酶停止。TADs 内部的基因和调控区域的组织在一定程度上保证了增强子和启动子间的交流，参与基因表达模式的保存，协助基因的共表达。并且，肿瘤细胞中 CTCF/cohesin 结合位点常发生突变，可能引起 TAD 的改变。在另一些研究中，一些位点相对 TAD 结构的扰动是稳定的，对 cohesin 依赖的TADs 的剧烈干扰对整体基因表达的影响很小。因此，



TADs 确切的功能目前尚不清楚。

图 1.3　研究染色质构象的测序方法

在 TADs 内，约几百 kb 距离的染色质组织形成染色质环。环通过基因组组织的最基本单位的三维折叠形成 [29]。染色质纤维的最基本单位是由连接 DNA 连接的 DNA 包绕的组蛋白八聚体核小体 [33]。一些环由 CTCF/cohesin 依赖的挤出机制形成。这些 CTCF 介导的环非常保守，参与到染色质构象的形成中。另一些环是CTCF 非依赖的，它们大多更动态，主要参与到对转录的直接控制，如通过介导增强子-启动子互作促进多梳复合物的形成。TADs 内部的环的组织可以是细胞特异的。TAD-嵌套结构包括亚-TADs、绝缘邻域和 FIREs（频繁互作区

域）。亚 TADs 和绝缘邻域通过参与局部染色质微环境对基因表达有显著影响。亚 TADs 通常聚集多个环 [34]. 而绝缘邻域通常由 CTCF 介导的环形成，以保证

其中基因的隔离 [35]。FIREs 是在 TADs 中间具有明显顺式连接度的区域，富集有活性增强子和超级增强子。FIREs 高度组织特异，它们的活性与决定细胞身份和组织功能的基因表达调控有关 [36]。这些局部结构的组织多样性与细胞类型间 TAD 水平的一致性相反，提示在染色质组织的这一精细水平上，特别的拓扑特征能调控种系特异的基因组。

1.1.3 染色质构象与精子发生关键事件

精子发生伴随着显著的染色质组织的重构。在哺乳动物中，精子基因组通过精蛋白包装形成，只有 1-15% 的 DNA 通过组蛋白包装 [37]。精子 DNA 螺旋形成大而紧密的精蛋白环形线圈。Hi-C 分析的结果也印证精子染色质的凝缩程度， 精子相比ESCs 和成纤维细胞，有更多的长程染色质互作[38]。尽管小鼠精子和 ESCs 以及体细胞由不同蛋白包装，它们中的TADs 的位置和区室A 区室B 的区分都是类似的[39]。CTCF 和 cohesin 在精子、圆形精子和 ESCs 中结合相似的位点 [40]。

精子发生过程伴随着显著的染色质构象重编程 [41]。在男性减数分裂细胞周期中，染色质组织经历几次事件，包括同源染色体配对、联会复合物的形成、减数分裂重组、解联会等等 [42]。在小鼠和恒河猴中，第一次减数分裂前期的粗线期细胞中同源染色体配对形成联会复合体，此时 TADs 缺失 [43-45]。粗线期染色体转录活跃，因此这一结果说明在发育的某些阶段，转录可以几乎与 TADs 无关 [42]。这一猜想也与 CTCF 对维持联会复合物和同源重组是非必需的一致[46]。传统的区室化在粗线期也减弱了。A 区室富集于减数分裂双链断裂热点和交叉位点。在不同物种的粗线期细胞中，有明显程度不同的精细的 A/B 区室。在小鼠粗线期细胞中，一些活跃转录的区域在空间上聚集形成“节点”或“点互作”。精子中的 X 染色体在减数分裂性染色体失活过程中被失活，而精细的 A/B 区室和与转录相关的“节点”在精子中的 X 染色体上都不存在。在缺乏 Sycp2（联会复合物蛋白 2）或 Top6bl（类 II 型 DNA 拓扑异构酶 6 亚基B）的小鼠精细胞中，精细 A/B 结构、传统的区室和 TADs 都部分恢复了[43]。这两种突变都能引起联会复合物的形成失败，细胞周期停滞在进入粗线期前。说明完整的联会复合物对观测到粗线期特异的 Hi-C 特征是重要的。

在机制方面，目前还不知道粗线期细胞为什么缺失 TADs，又为何形成精细的 A/B。联会复合物可能限制染色体的 TAD 相关的自由挤出，从而消除 TADs。或者，cohesin 的功能可能用于形成稳定的染色质环阵列。在这两种可能中，TADs 的缺失

都可能促进精细的染色质区室化如精细的 A/B、节点和点互作的形成 [47]。但在 cohesin 缺失的细胞中，粗线期细胞中这种转录相关的精细区室化甚至更为明显。

1.2　干细胞分化和拓扑结合域

1.2.1　胚胎干细胞

ESCs（胚胎干细胞）具有自我更新和分化形成胚胎的任一细胞类型的能力。自我更新和多能性的实现由转录因子和多种染色质修饰蛋白协作的自调节网络维持 [48]。内源和外源共同决定 ESC 命运。内源因此 Oct4、Nanog 和 SOX2 是 ESC 实现自我更新和多能性的核心转录因子。

基因表达的随机性，也即基因表达噪声，能促发细胞间的基因表达异质性，给单细胞生物带来表型多样性，通过对冲环境突变来提高种群适应度。在多种多细胞生物中也观测到了基因表达噪声。谱系特异基因参与胚胎干细胞自发分化和多能性。对这些基因的沉默涉及 PcG（多梳家族）蛋白带来的染色质表观修饰 [49-50]。PRC1 和 PRC2 分别有四个核心亚基。PRC2 甲基化 H3K27me3（组蛋白 H3 的 27 位赖氨酸），这种甲基化修饰是抑制 *Hox* 基因所必需的。典型的PRC1 有包含结合 H3K27me3 的亚基, 其既可能通过形成染色质环抑制基因表达[51]，也能通过形成增强子启动子环促进基因表达[52]，这使得它能参与对胚胎发育过程中的基因表达的精密调控。PRC1 的 Phc2 亚基的自多聚化是模式发育所必需的。PRC1 的第三种功能是将附近的核小体压缩形成球状结构。

PRC1 和 PRC2 对胚胎干细胞的多能性和自我更新能力的影响还存在争议。比如对体外分化来说，PRC2 亚基是必需的，但 Suz12 或 Eed 亚基缺失对mESCs 的基因表达和自我更新能力没有显著影响 [53]。尽管 EZH1/2 对活化的 hESCs（人类胚胎干细胞）的自我更新是必需的，但对于加入 MEK 和 GSK3 抑制剂的培养环境中的保持类 mESCs 的初始状态的 hESCs 的增值是非必需的 [54]。另外，缺失 PRC1 催化核心亚基 Ring1 和 Rnf2 会导致对谱系特异性基因抑制强度降低和自发分化，说明 PRC1 对 mESC 身份的维持是必需的[55]。研究结果显示，PRC2/cPRC1 和 vPRC1 都可以抑制谱系特异基因，这种冗余性可以保证 mESC 稳定地自我更新 [56]。

1.2.2　拓扑结合域的形成和功能

在几十到几百kb 尺度上，染色体折叠形成优先域内互作的 TADs[30]。TADs 是的定义转录谱的染色质构象单位，它们对塑造功能性染色体构象有基础作用。TADs 的尺度刚好利于其发挥功能，将功能性互作限制在 TADs 内可能对保证正常的基因调控是重要的[57]。例如：TAD 边界与细胞分裂过程的复制域对应。在

细胞分化过程中，位于同一个 TAD 的基因倾向于被共调节 [58]。增强子和基因启动子之间的互作主要被限制在TADs 内 [59]。在基因组中插入的报告基因受到大的、与 TADs 强相关的调节域中的增强子的影响 [60]。通过改变 TAD 边界破

坏 TAD 结构会导致顺式调节元件和基因启动子间的异位互作，从而造成基因表达异常，导致发育缺陷或癌症[61]。因此，研究TADs 的结构与功能对阐明基因组构象及其调节机制十分重要，也对医学研究有重要意义。但是，即便基因组折叠为自互作域是演化上被广泛采用的策略，TADs 仍然有不同的大小、染色质特征和不同的形成机制。这提示 TADs 能被细分为具有不同结构和功能特征的亚类。另外，TADs 的鉴定依赖于 Hi-C 数据的分辨率和 TAD 注释的方法。在更高的测序深度和分辨率下能定义更精细的染色质互作模式和内部绝缘区域。因此， TAD 边界的确认是困难的。在不同物种中，这些染色质域是否代表同一层面的染色质构象也尚不清楚。

在 TAD 的形成机制方面，主流的理论是环挤出决定着域的形成（图1.4）。哺乳动物基因组包含许多在 Hi-C 上有“角点”（相较周围的域结构有显著高的互作频率的点状的邻近像素集）特征的域。“角点”结构代表长程环互作。顶点上有“角点”的染色质域代表了所谓的环域。在人类细胞中，大约有一万到六万个角-点结构。大多数交点的端点都 CTCF 结合的模序 [62]。60%-90% 的角点在两个端点上都有对向的 CTCF 模序。这种对象的 CTCF 模序对环域的形成是必要的。现有的“环挤出”模型认为基因组上的分子马达会沿着 DNA 序列运动，随之“挤出”其中的 DNA。计算方法能模拟环挤出，重现环域，也能预测 DNA 挤出因子的存在 [63]。

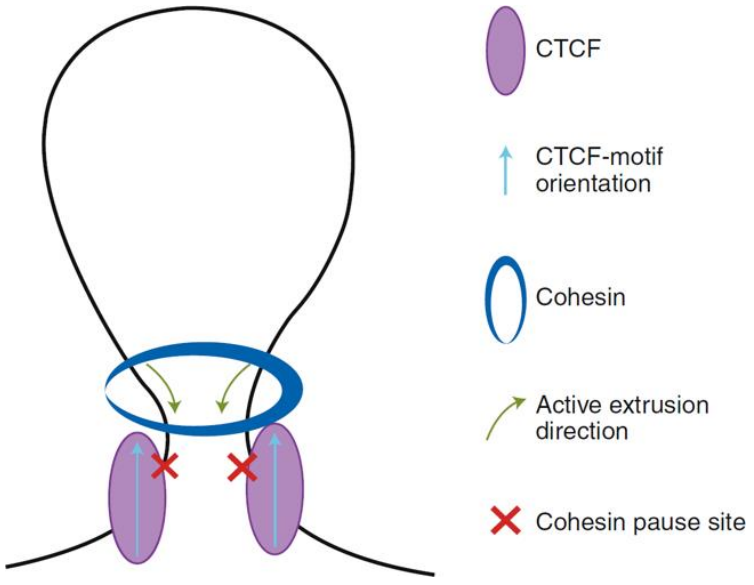


图 1.4　环挤出模型

SMC（染色体结构稳定）复合物，例如 cohesin 和 condensin，一直被认为是环锚定因子，它们可通过稳定已形成的环或通过主动的挤出机制发挥功能。通过染色

质免疫沉淀测序测定的富集 cohesin 结合的峰位于 CTCF 结合位点附近，稍微向对向的CTCF 模序的 3’端偏移的位置[32]。这一结果支持招募cohesin-CTCF

的滑动机制。单细胞成像研究给出了支持环挤出的更直接的证据：condensin 和 cohesin 能以 ATP 依赖的形式在裸 DNA 上移位 [64]。因此，SMC 复合物停留在对向的 CTCF 模序后的环挤出机制是目前主流的环域形成机制。

除了位于单一区室中的TADs 和有角点的TADs，还有别的 TADs。这些 TADs 可能也是通过挤出形成的。因为非顶点处的互作也可能是活跃的挤出事件的附带信号。有假说认为具有不同分子特征的边界可以得到不同的挤出-阻碍强度，从而导致角点互作频率不同的假说，这一假说可能通过对影响挤出速率的不同蛋白的注释得到验证。其他的解释既非区室也没有角点的域的假说包括不能形成强的瞬时边界的环挤出、不能在多数细胞中形成弱的边界的环挤出等。

第二种域形成机制是区室化。通过高分辨率分析，哺乳动物细胞中存在没有角点的类域结构并且和区室坐标一致的“区室域”[65]。问题在于在区室化是染色质域形成的主要原因的物种中，环挤出是否发生，如果发生，又如何发生。

1.2.3　对转录调控的综合计算分析

解析基因型如何解码形成多细胞表型对于理解个体的发育、结构和功能是重要的。解码过程的第一步是受 TF 调控的基因转录活动。TF 对细胞的总体作用是通过其对靶基因的调控效应在转录网络中的传播来实现的。TF 的调节活性受自身表达水平、翻译后修饰和辅因子的存在与否的影响。鉴于这些生物学事实，人们已开发出许多推断 TFs 的活性的统计或机器学习方法。一些方法只关注 TF 的局部活性[66-67]; 而另一些方法结合空间长程互作信息，利用包括染色质状态（从 ChIP-seq（染色质免疫沉淀后高通量测序）和 ATAC-seq（转座酶可及的染色质高通量测序检测得到））、基因表达情况（从 RNA-seq（RNA 测序）或芯片阵列得到）、染色质长程互作（从 Hi-C 或计算预测得到）的信息整合分析来预测有驱动作用的 TFs[68]。

在分析的数据源方面，ENCODE（DNA 元件百科全书）项目系统性地测量了小鼠胚胎发育的 12 种组织和 8 个发育阶段的表观动态信息。这些数据集提供了分析发育过程中的复杂转录调节回路的前所未有的机会，也为整合计算分析带来了新的挑战。

第 2 章　小鼠精子发生过程中的染色质构象研究

2.1　背景

应用成像技术，人们早已发现哺乳动物的精子发生过程中涉及显著的染色质构象重新组织的现象。精子发生的起始是少量的 A 型原始生殖细胞通过有丝分裂分化形成A 型和B 型精原细胞。随后，精原细胞进行第一次减数分裂，其中的标志性事件是双链断裂以及粗线期细胞中同源染色单体的联会。此后，次级精母细胞很快进行第二次减数分裂，形成单倍体的圆形精子并最终形成成熟精子。对这一过程的多组学研究揭示了其中的转录组、DNA 甲基化、染色质可及性的变化。这些研究促成了对精子发生过程中更深入的分子事件（如染色质构象和转录调控）及其之间的关系的探索。

精子发生过程涉及高度协调的染色体重组事件，因此对精子发生过程，尤其是减数分裂过程的染色质构象研究的表征一直是热点。人们已经考察了小鼠和恒河猴精子发生过程的一些阶段中的染色质构象。这些研究都发现了在粗线期，TADs 几乎消失，同时区室化的强度很低。而在恒河猴的数据中，研究人员在更小的分辨率上考察区室，发现了在粗线期细胞中与转录活动相对应的更小的区室结构。黏粘蛋白在 pacSC 和 rST 中的结合也会影响基因表达。但在减数分裂过程中，染色质构象、染色质可及性是如何参与减数分裂特异的基因表达活动， 仍然是未解决的问题。

在本研究中，我们结合了 Hi-C、ATAC-seq、ChIP-seq 和 RNA-seq 技术，检测小鼠精子发生过程中各个阶段的基因组、表观组和表达组信息，来探索染色质构象、染色质可及性、蛋白质结合、转录活动等在这个过程中的协调互作关系[69]。

2.2　研究成果

2.2.1　小鼠精子发生过程中染色质构象的总体变化

在主要由徐倩岚分选出小鼠精子发生的四个阶段的细胞（priSG-A：A 型生殖细胞、SG-A：A 型精原细胞、pacSC：粗线期细胞、rST：圆形精子），罗正誉进行原位 Hi-C 测序得到测序数据并收集 SZ(成熟精子) 时期细胞的 Hi-C 数据后，我对这些测序数据进行了分析。初步的序列比对后，我们考察了生物学重复数据间

的相关性，高的相关性验证了实验数据的可靠性（附图A.1）。常染色体的 A/B区室化在整个过程中没有明显的消失（图2.1（B）。鞍形图和计算得到的区室化强度的值显示 A/B 区室化的强度有变化 (图2.1（B,C）。具体而言，区室化强度

从 priSG-A 到 pacSC 中递减，随后逐渐恢复直到达到 SZ 中的最高强度。

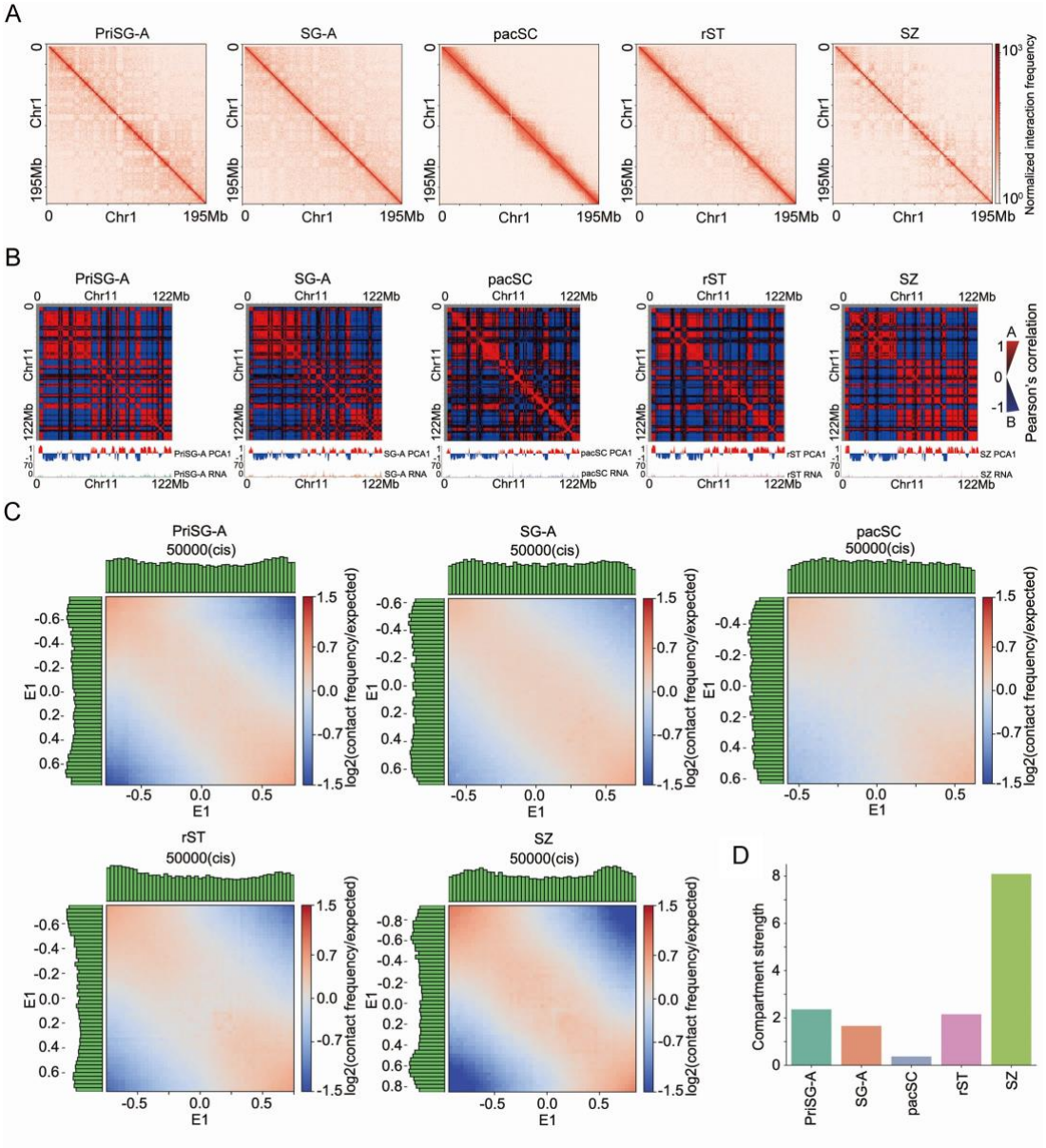


图 2.1　小鼠精子发生过程中区室和 TADs（拓扑结合域）的重组

从互作的热图以及各个时期细胞中所有拓扑结合域边界附近的 o/e（观测值/预期值）互作频率的均值可以看出，互作 TAD 结构在 priSG-A 和 SG-A 中还很显著，但在pacSC 和 rST 中几乎消失，并在 SZ 中恢复（图2.2[①]（A, B）。

我

[①]（A）减数分裂不同时期样本中在priSG-A 中找到的TAD 边界及其附近区域的平均观测/预期互作频率热图。

(B)　减数分裂不同阶段样本的染色质观测/预期互作频率热图（25 kb bin）。黑色的三角形线条标注各个

时期的 TADs，纵线标注 priSG-A 中找到的 TAD 边界位置。pacSC 和 rST 中的 CTCF ChIP-seq 覆盖度以及 gencode 基因展示在下方。

(C)　减数分裂各时期样本中的 TADs 及其附近区域（±0.5 TADs 大小）的平均 IS（绝缘值）。

(D)　根据样本的 IS 相关性作出的树状图。

们进一步计算了以染色体每隔xxbp 的坐标位置为中心，周围 xxbp 内的互作均值为该点的一维 IS 值（绝缘值），考察了 IS 值在拓扑结合域附近的分布，发现 IS 值也从 priSG-A 到 pacSC 逐渐下降，随后增大，印证了以上结果（图2.2（C）。对样本来源的时期按照其中的TAD 及其附近区域的 IS 值进行聚类分析，我们发现 priSG-A 和 SZ 的 TAD 更为类似（图2.2（D），说明TADs 重组后大体恢复了在 priSG-A 中的构象。

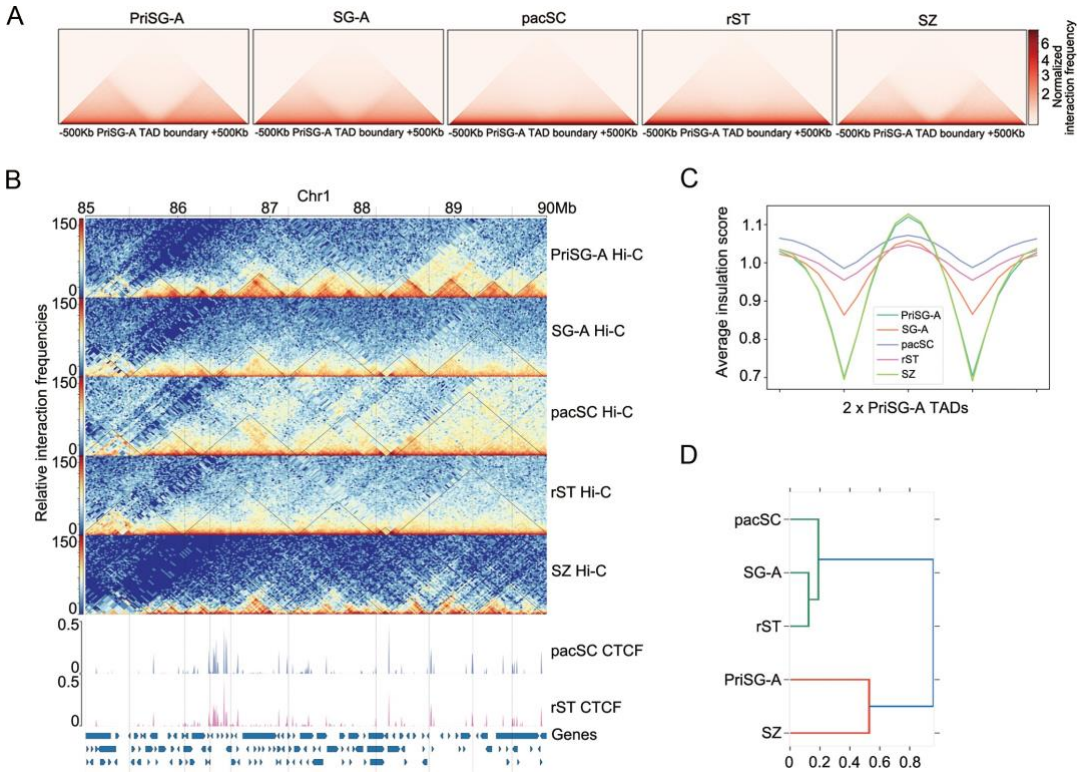


图 2.2　小鼠精子发生过程中 TADs（拓扑结合域）的重组

2.2.2　priSG-A 和 SZ 的染色质构象类似

priSG-A 和 SZ 的 TAD 更为类似，这一数据引导我们进一步研究各时期的染色质互作模式。不同时期的互作热图的比较和不同距离间染色质互作频率大小的比较反应了整体上，priSG-A，SG-A 和 SZ 的染色质折叠模式类似，而 pacSC 与他们有所不同（图2.3[1] (A-C)），这与根据 IS 值的聚类结果一致(图2.2 (D))。整体来看，priSG-A、SG-A 和 SZ 中的染色质互作比较类似。相比 SG-A，

pacSC 中

[1]（A）精子发生各个阶段前后一号染色体观测/预期互作频率倍数的热图（100 kb bin）。

（B） 精子发生pacSC 相比priSG-A、SZ 相比pacSC 和 SZ 相比priSG-A 的一号染色体观测/预期互作频率倍数的热图（100 kb bin）。

（C） 精子发生各阶段样本的 P(s) 图（纵坐标为互作频率，横坐标为基因组距离）。

（D） pacSC 和有丝分裂期样本的 P(s) 图以及分别展示斜率为-1（蓝色）和-0.5（橘色）的虚线。

的短程染色体互作（距离＜ 5Mb）更强，而长程互作更弱（图2.3（C）。

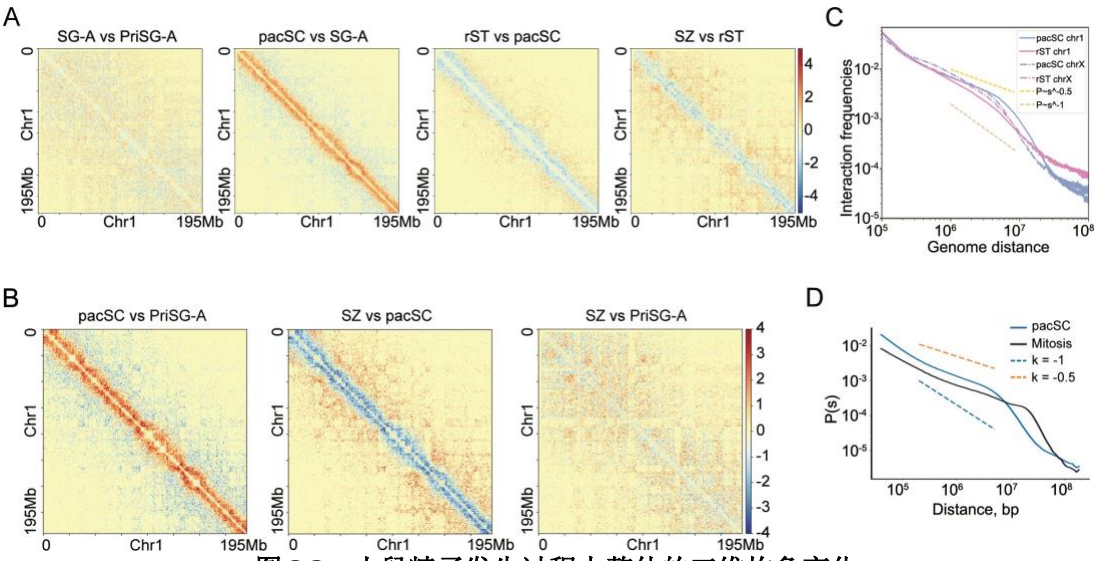


图 2.3　小鼠精子发生过程中整体的三维构象变化

2.2.3　pacSC 中，位于活性区室的基因和 piRNA 簇具有减数分裂相关的功能

我们进一步考察了减数分裂过程中的染色质 A/B 区室化变化。A 区室由对染色质内的 o/e 互作矩阵（在本研究中为 50kb 分辨率）进行特征向量分解后得到的最大特征值为正的区域，而 B 区室定义为最大特征值为负的区域。这些计算得到的 A 区室与生物学上的活性的真染色质区域对应。相反，B 区室更多地与无活性的异染色质区域对应。在 priSG-A 中，由 43.2% 的区域属于A 区室，其中由 13.9% 在 pacSC 中已处于 B 区室中；而 priSG-A 中的 B 区室，有 15.3% 在 pacSC 中转变为了A 区室（图2.4[①]（A）。也即，在pacSC 中，更多的区域属于 A 区室，提示 pacSC 中的染色体的转录状态更活跃。

PIWI 蛋白和 piRNAs（PIWI 互作 RNAs）在动物细胞中介导对靶序列的抑制。在粗线期早期，piRNA 从大的被称作 piRNA 簇的基因片段转录得到。这些

[①]（A）展示相比 priSG-A，在 pacSC 中转换了区室化状态（A 到 B 或 B 到 A）的基因组区域的比例。

Ⓑ　展示在 priSG-A 中分别位于 A 或 B 区室的 piRNA 簇的数量以及在 pacSC 中转换了区室化状态的 piRNA 簇数量的条形图。

Ⓒ　在 priSG-A 和 pacSC 中都位于A 区室的piRNA 簇的平均ucsc phastCons 值以及在priSG-A 中位于 B 区室而在 pacSC 中位于 A 区室的 piRNA 簇的平均 ucsc phastCons 值。

Ⓓ　展示转换了区室化状态的区域中的差异表达基因数量的条形图。

Ⓔ　特定基因Boll 位于的一号染色体，52-58 Mb 区域上的 PCA1（最大特征值）和 RNA-seq 覆盖度曲线。Gencode 基因展示在下方。

Ⓕ　展示在pacSC 中相比priSG-A 中区室化状态从B 变为A 且表达上调的基因所富集的生物学过程的条形图。

Ⓖ　展示（F）中所用的基因富集到的哺乳动物表型的条形图。

RNA 能控制转座子，也参与到亲代印记中，是后续减数分裂过程必须的因素。我们也研究了在 pacSC 中相比 priSG 中染色质区室状态发生转化的基因区域中的 piRNA 簇（piRNA 簇的信息从公共数据库中得到）。我们发现 79.9% 的 piRNA 簇在 priSG-A 中位于 A 区室中，其中只有 12% 在 pacSC 中转变了区室化状态；而 80% 的在 priSG-A 中位于 B 区室的 piRNA 簇在 pacSC 中是位于 A 区室的(图2.4 (B))。既然 A 区室区域更多的是染色质活性区域，这一结果也与 piRNA 簇在 pacSC 中的表达活动一致。

为了进一步研究未转变和转变了区室化状态的 piRNA 簇的区别，我们利用衡量每个碱基属于保守原件的概率的 phastCons 值，考察了这两种 piRNA 簇的保守性。具体而言，我们比较了转变区室化状态和未转变区室化状态的 piRNA 簇中的碱基的平均 phastCons 值，发现没有转变区室化状态的 piRNA 簇在考察的 60 种脊椎动物的系统发育中更为保守（图2.4（C）。这一结果提示在 priSG-A 和 pacSC 中都处于A 区室中的piRNA 簇受到更强的选择压力而更为保守，对这些物种可能具有有利功能。总之，piRNAs 的功能活动可能受到其在染色体构象中位于的区室的影响。

结合相应的 priSG-A 和 pacSC 样本测得的 RNA-seq 数据，我们考察了位于转换了染色体区室状态的区域中的基因。我们发现位于从 priSG-A 的 B 区室变换到 pacSC 中的 A 区室区域的基因大多（1037 个）转录水平有所上升（图2.4（D）。这一结果提示染色质所处的状态由 B 区室转变为 A 区室这一变化可能基因转录的上调从而发挥它们的生物学功能。这些基因中有的参与减数分裂的正常进行，其中一个例子是减数分裂 G2/M 转换和精细胞发育必须的 DAZ 家族蛋白 Boll（图2.4（E）。通过 GO（基因功能）富集分析，其中确实有许多基因负责与正常精子生理功能关系密切的纤毛形成和 DNA DSBs 修复（图2.4（F）。

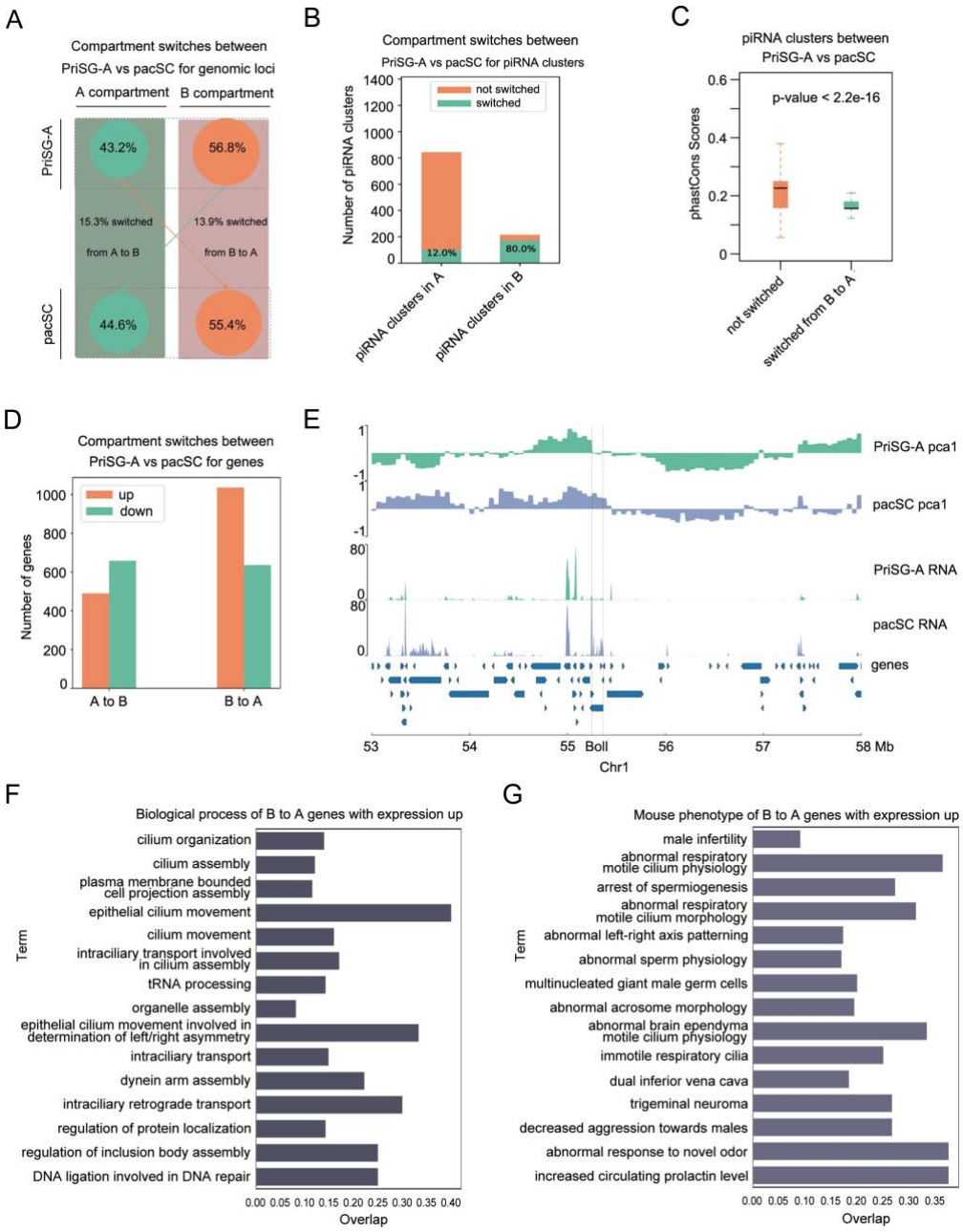


图 2.4　在 pacSC 中位于活性区室的基因和 piRNA 簇

2.2.4　小鼠精子发生过程中染色质可及性和 CTCF/cohesin 的结合依然保留

我们已观察到TADs 在精子发生过程中经历的重组，特别的，在 pacSC 中其结构的消失。在此基础上，我们进一步探究这种染色质三维构象的变化是否与其上的基因组或表观基因组特征有关。主要由徐倩岚和罗正誉在四种精子发生过程

中的细胞（priSG-A、SG-A、pacSC 和 rST）样本进行了 ATAC-seq 测序，以获得染色质可及性信息。在检验了测序数据的重复性后（图2.5[①](A)），我们分析

[①]（A）分别展示 priSG-A、SG-A、pacSC 和 rST 的 ATAC-seq 信号的重复性的散点图。

发现染色质开放的位置大多位于基因启动子或增强子附近；染色质开放的程度随着精子发生过程的进展逐渐降低（图2.5（A，B）。在priSG-A 和 pacSC 中的染色质可及性水平没有明显的不同。考虑到在这两种细胞中的染色体三维构象有较大区别，这一结果提示我们，在精子发生过程中 TADs 水平的重组可能与染色体的开放性无关。

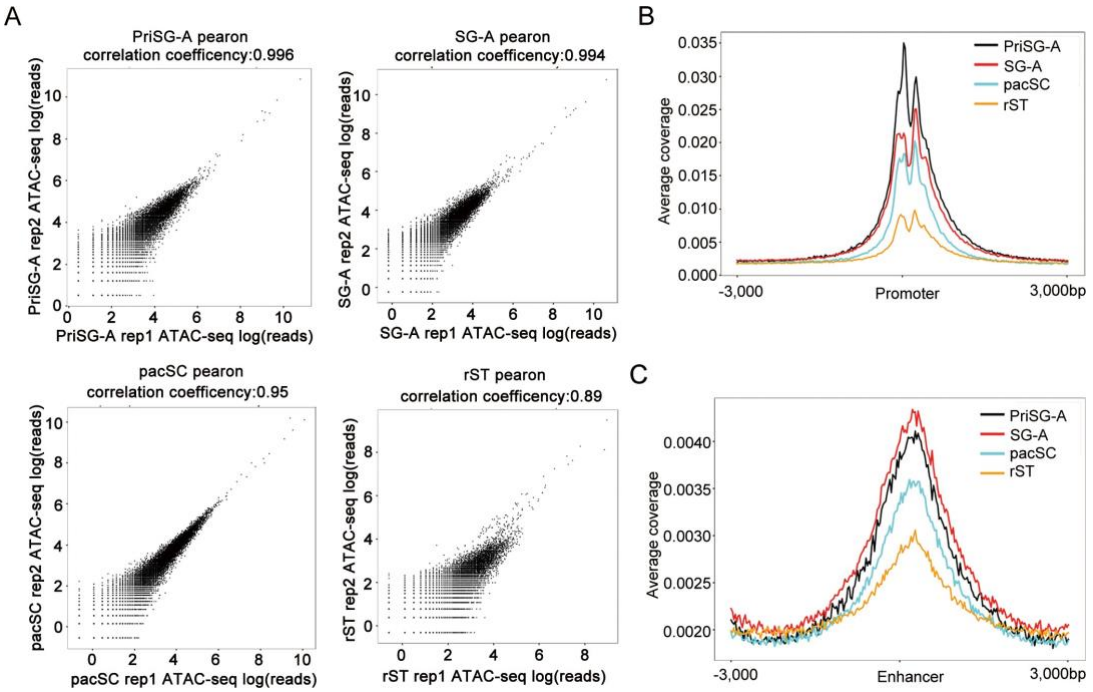


图 2.5　小鼠精子发生过程中由 ATAC-seq 测得的染色质可及性

另一方面，染色质可及性与转录活性有关。因此，徐倩岚和罗正誉应用抗Pol II S2P (延长形式的 RNA 聚合酶 II) 进行了 ChIP-seq 测序。在 pacSC 和 rST 中，Pol II S2P 富集于 TSSs（基因转录起始位点），这些富集位点的基因可能在这两种细胞中有表达（图2.6[①]（A）。但这两种细胞中的染色体可及性并不类同， 在 priSG-A 和 SG-A 间也不类同。具体地，在priSG-A 和 pacSC 之间，我们找到了 6580（总共 15815）个差异可及性区域，而在 pacSC 和 rST 之间，有 1288（总共 15815）个差异可及性区域（图2.6（B）。对位于priSG-A 和 pacSC 差异可及

Ⓑ　在基因的启动子区域（±3 kb）上的 ATAC-seq 信号基因荟萃图。

Ⓒ　在增强子区域（±3 kb）上的 ATAC-seq 信号基因荟萃图。

[①]（A）在 pacSC 和 rST 中所有基因的转录起始位点附近（±3kb）上的平均 Poll II 信号。

（B） 在四个精子发生时期样本中找到的染色质可及性热图。

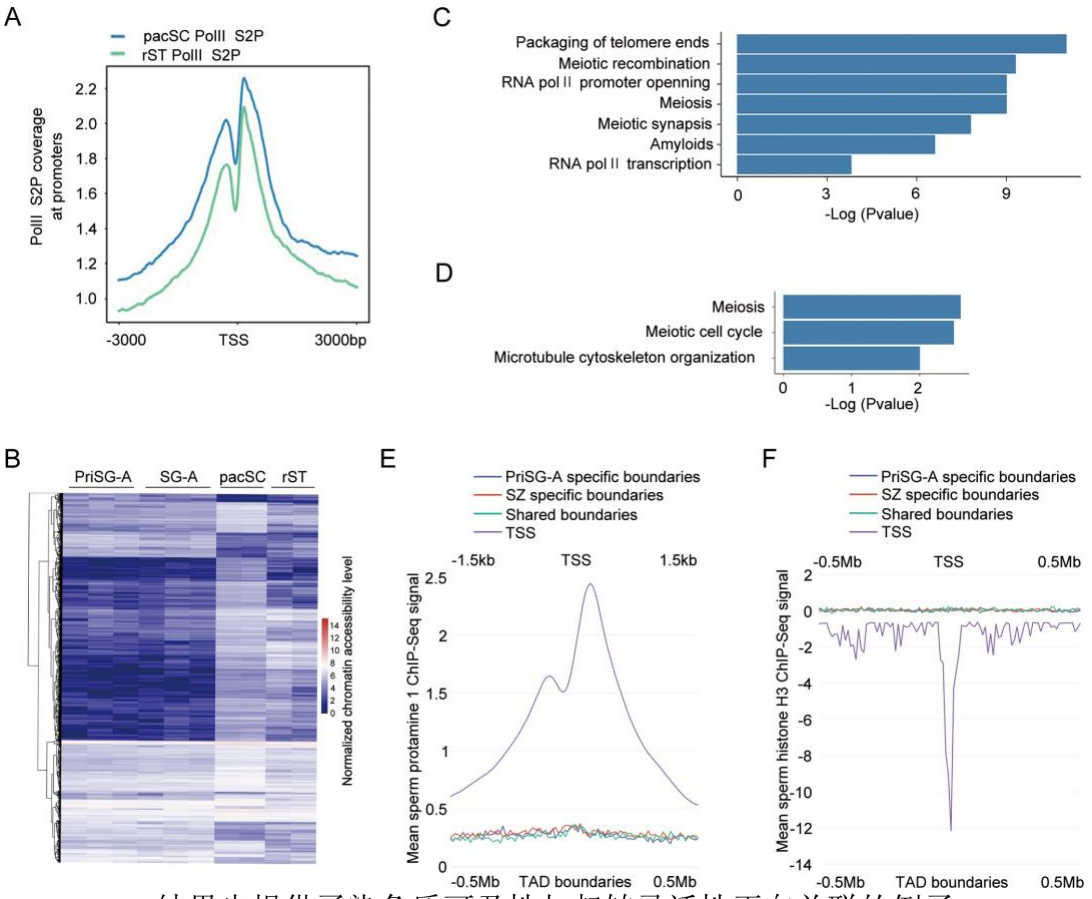
（C） priSG-A 和 pacSC 间的差异可及性位点的功能富集。

（D） pacSC 和 rST 间的差异可及性位点的功能富集。

（E） 精子中平均精蛋白 1 ChIP-seq 信号在精子发生各时期中的TAD 边界及其附近（±0.5Mb）的。橘色的线展示 SZ 中 TSS 附近的平均精蛋白 1 ChIP-seq 信号。

（F） 精子中平均组蛋白 H3 ChIP-seq 信号在精子发生各时期中的TAD 边界及其附近（±0.5Mb）的。橘色的线展示 SZ 中 TSS 附近的平均精蛋白 1 ChIP-seq 信号。

性区域的基因进行 GO 分析的结果显示，这些基因的功能集中于参与减数分裂的关键事件，如减数分裂重组或联会（图2.6（C）。位于 pacSC 和 rST 差异可及性区域的基因也多是减数分裂相关基因，但无法富集到减数分裂重组或联会事件中（图2.6（D）。这一结果可以用 pacSC 和 rST 各自处于的减数分裂阶段来解释：pacSC 是第一次减数分裂前期的细胞，而 rST 形成于减数分裂完成后。这



一结果也提供了染色质可及性与起转录活性正向关联的例子。

图 2.6　小鼠精子发生过程中的染色质可及性和精子中保留的组蛋白

在哺乳动物精子发生过程中，组蛋白和精蛋白互换以形成高度聚缩的染色体结构。其中在小鼠精子中的染色质上，仅保留了 1%-8% 的组蛋白。在先前的核纤溶酶处理的精子中进行的 ChIP-seq 结果表明，这些保留的组蛋白主要位于远端基因间区域。此前尚不清楚 TAD 边界是否与组蛋白的保留有关。我分析了公共的精子样本中的精蛋白 1 和组蛋白 H3 ChIP-seq 数据，并考察了精子中 TAD 边界附近的 ChIP-seq 平均信号，发现精蛋白 1 在基因转录起始位点有富集，但这两种蛋白在 TAD 边界都没有富集（图2.6（E，F））。这些结果表明精子中的

TAD 结构与精蛋白或组蛋白的组成无关。

我们进一步研究了精子中保留的组蛋白是否在其可能形成的后代二倍体细胞中发挥功能。在 171 个其外显子上有组蛋白的基因中，有 16 个基因有报道在

早期胚胎发育中表达（Theiler stage 1-5）（Table 2）。将来对这些保留的不同修饰的组蛋白的分布和其可能在早期胚胎发育中的功能的研究可能提供更细致的结果和假说。

表 2.1　MGI-小鼠基因表达数据库（Smith et al., 2019）中检测到在 Theiler Stage 1-5 有表达的基因。

MGI 基因 ID	基因符号	基因名	类型	染色体	基因位置 - GRCm38	cM	链
MGI:87911	Acvr1	activin A receptor, type 1	PDG[1]	2	58446438-58566828	33.05	-
MGI:2448562	Adnp2	ADNP homeobox 2	PDG	18	80126311-80151482	53.28	-
MGI:2442590	Ankrd35	ankyrin repeat domain 35	PDG	3	96670131-96691032	41.94	+
MGI:108405	Apbb2	amyloid beta (A4) precursor protein-binding, family B, member 2	PDG	5	66298703-66618784	34.43	-
MGI:108028	Atr	ataxia telangiectasia and Rad3 related	PDG	9	95857597-95951781	50.27	+
MGI:1298392	Bscl2	Berardinelli-Seip congenital lipodystrophy 2 (seipin)	PDG	19	8837467-8848683	5.76	+
MGI:88455	Col4a2	collagen, type IV, alpha 2	PDG	8	11312805-11449287	5.62	+
MGI:94865	Dbi	diazepam binding inhibitor	PDG	1	120113280-120121078	52.65	-
MGI:99892	Lama1	laminin, alpha 1	PDG	17	67697259-67822647	38.8	+
MGI:97275	Myod1	myogenic differentiation 1	PDG	7	46376474-46379092	30.03	+
MGI:2441856	Sf3b2	splicing factor 3b, subunit 2	PDG	19	5273932-5295455	4.29	-
MGI:108078	Sfrp2	secreted frizzled-related protein 2	PDG	3	83766321-83774316	37.37	+
MGI:103063	Stat1	signal transducer and activator of	PDG	1	52119440-	26.81	+

		transcription 1			52161865		
MGI:98729	Tgfbr2	transforming growth factor, beta receptor II	PDG	9	116087695-116175363	68.39	-
MGI:1341872	Tjp2	tight junction protein 2	PDG	19	24094505-24225030	19.17	-
MGI:1926031	Zc3hav1	zinc finger CCCH type, antiviral 1	PDG	6	38305286-38354603	17.72	-

〇1 蛋白编码基因

我们进一步验证了染色质可及性的差异与 pacSC 中 TADs 的消失之间可能的分子水平的关联。在哺乳动物细胞中，CTCF 富集于 TAD 边界。一些研究中，缺失 CTCF 会导致 TADs 的丢失。因此我们考察了 CTCF 模序的可及性是否与 TAD 边界的形成有关。我们发现在 priSG-A 和 pacSC 中的 CTCF 模序上的可及性是类似的，也即在没有显著的 TADs 的细胞中，CTCF 仍然可能结合其结合域。为了验证此可能，徐倩岚和罗正誉对 pacSC 和 rST 样本做了 CTCF ChIP-seq 和 Rad21 ChIP-seq。结果显示在 priSG-A 的 TADs 边界位置上仍然结合了 CTCF 和 cohesin。这些结果表明，pacSC 中的 TADs 的消失不是由于 CTCF 或 cohesin 从染色体上脱落，也就是说，CTCF 或 cohesin 的结合不足以维持 pacSC 中的 TADs。

2.2.5　减数分裂 DSB 位点在 priSG-A 中提前开放，piRNA 簇只在 pacSC 中开放

DSB（DNA 双链断裂）对于减数分裂中重组的发生是必须的。DSB 开始于细线期，结束于双线期。我们利用 ATAC-seq 数据考察了已经报道的 DSB 热点区域中点附近的染色体可及性。在 Hop2 敲除鼠的 DSB 阳性细胞中，减数分裂 DSB 不被修复，因而减数分裂进程停滞在类粗线期。我首先利用这种细胞样本中的 Dmc1 ChIP-seq 数据得到减数分裂 DSB 热点区域中点。随后，我比较了priSG-A 和 pacSC 中 DSB 热点区域中点附近的染色质可及性，发现其在这两种细胞中没有显著差异（图2.7[①]（D））。这一结果提示，DSB 热点区域的开放可能早于 DSB 的发生，并且持续到 DSB 在 pacSC 中的修复完成。

我们已经发现在 pacSC 中更多的 piRNA 簇位于活性的区室中，便进一步比较了 priSG-A 和 pacSC 中可及的 piRNA 簇的数量。其结果是 pacSC 中有更多的可及的 piRNA 簇（图2.7（E））。这一结果与 piRNA 簇位于的区室类别以及这类 piRNA 在 pacSC 中表达的事实一致。

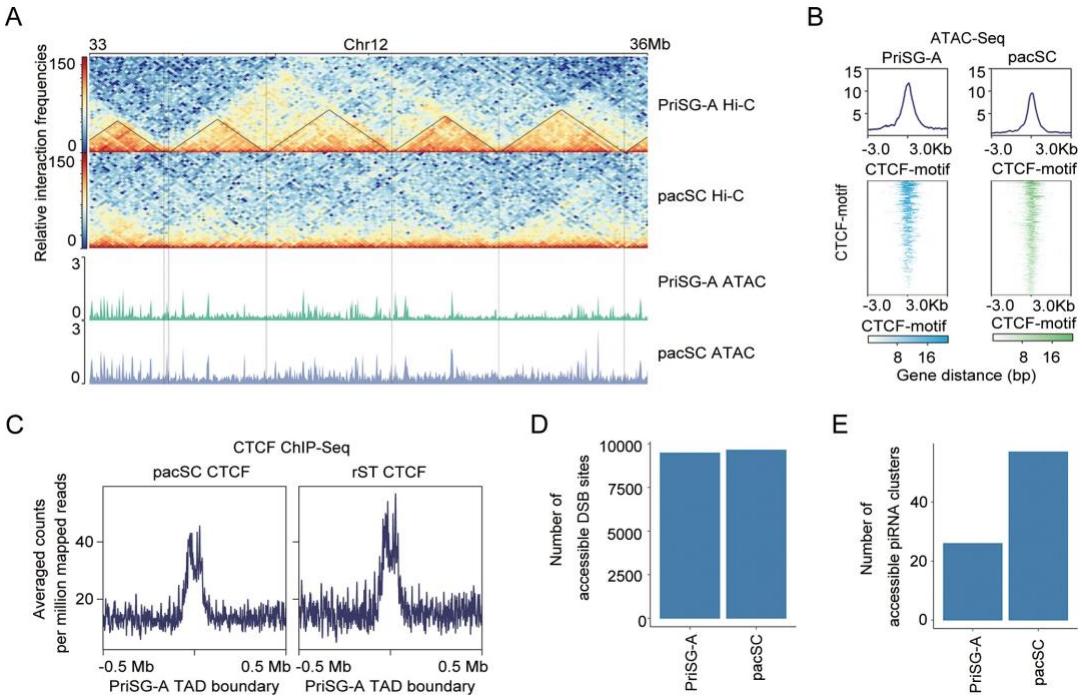


图 2.7　pacSC 中的染色体可及性区域的特征

[①]（A）priSG-A 和 pacSC 中的染色体观测/预期的互作频率图（25kb bin）。纵向虚线标注 priSG-A 中的 TAD 边界位置。pacSC 和 rST 中的标准化 ATAC-seq 覆盖度展示在下方。

（B）ATAC-seq 信号在 CTCF 模序区域（±3kb）上的热图。上方是 CTCF 模序区域上的平均 ATAC-seq 信号图。

（C） 在 priSG-A 中发现的TAD 边界附近（±0.5Mb）上的pacSC 和 rST 中的CTCF 和 Rad21 平均信号。

（D） priSG-A 和 pacSC 中可及的 DSB 位点数量。

（E） priSG-A 和 pacSC 中可及的 piRNA 簇数量。

本节结果指出了 DSB 和 piRNA 簇在哺乳动物精子发生过程中的功能受到精密调控的一个方面。

2.2.6　在精子中重组形成的染色质环参与早期胚胎发育

我利用 cLoops 确定了精子发生各个时期的细胞中的染色质环以研究更高分辨率的构象。我考察了染色质环的数量以及 priSG-A 的环端点附近各个时期细胞中染色体互作的总强度，发现二者都与区室化强度和 TADs 的动态变化类似，在过程中也经历了重组：在 pacSC 和 rST 中环的数量和强度都很低，随后逐渐恢复（图2.8[①] (A-C)）。为了研究 SZ 中数量众多的染色质环的功能，我分析了已发表的小鼠睾丸中的组蛋白修饰和 RNA Pol II 结合的 ChIP-seq 数据，以此找到了可能的增强子在基因组上的位置。与 SZ 中的环端点取交集后，发现共 1572 个基因与 SZ 中的增强子通过染色质环互作（图2.8（D）。在这些基因中，只有少部分（317，即 20%）在 priSG-A 中也被染色体环连接（图2.8 (E)）。这一结果提示，尽管在 SZ 中恢复了较大尺度上的染色质三维构象（如区室和TADs），在更小尺度上，重新建立的染色质环与 priSG-A 中的是不同的。

[①]（A）精子发生过程各时期细胞中的观测/预期的 Hi-C 互作频率热图（5kb bin,13 号染色体，50-54Mb）。黑色方框显示的是 priSG-A 中找到的染色质环的位置。

Ⓑ　在 priSG-A 中找到的环的端点附近（±50 kb）的观测/预期的 Hi-C 互作频率热图。

Ⓒ　精子发生过程中各时期细胞中的 APA 峰对左下的比值。

Ⓓ　精子发生过程中各时期细胞中的活性启动子-增强子环的数量。青色柱状图展示与这些环相连的基因数量

Ⓔ　在 priSG-A 中由 1096 个基因的启动子位于环的端点。在 SZ 中有 1572 个这样的基因。这两组基因的交集内有 317 个基因。

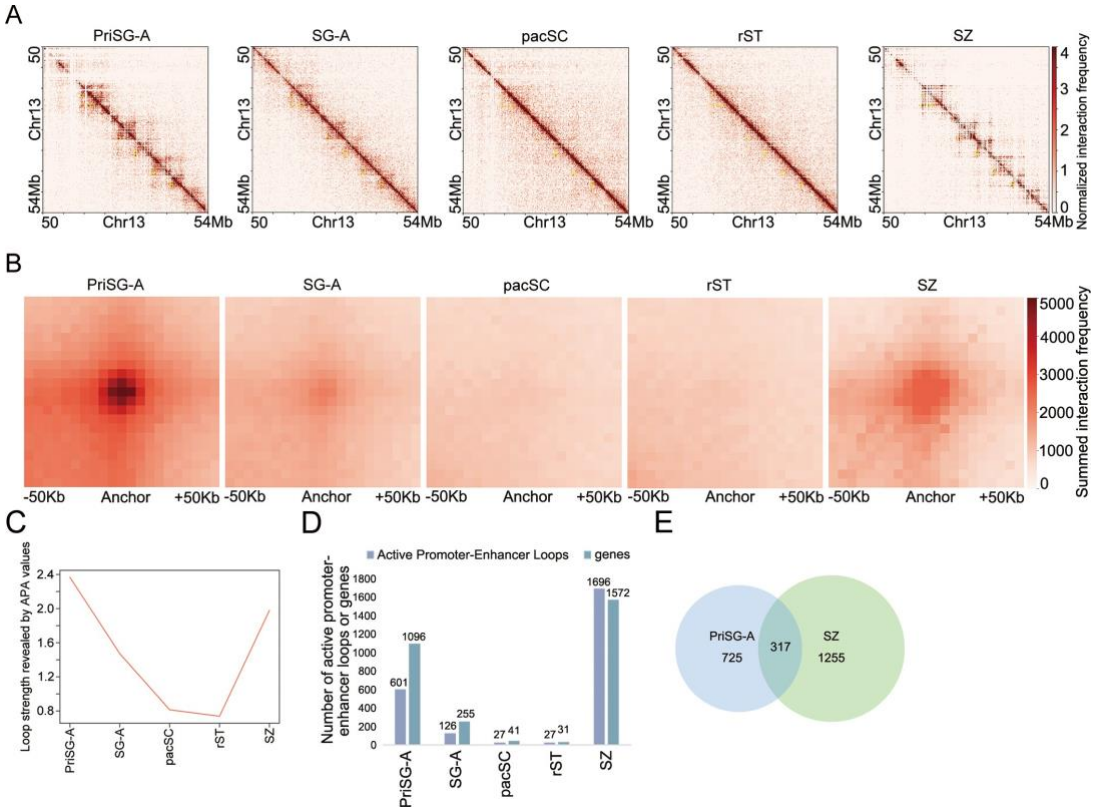


图 2.8　小鼠精子发生过程中的染色质环的重组

为验证在 SZ 中形成了染色质环的基因可能在后续的早期胚胎发育过程中具有转录活性的假说，我考察了这 1572 个基因被报道的表达情况，发现它们大部分在胚胎中有表达（1174，即 75%），其中，有 26 个基因从 TS1 开始表达并持续到 TS4，63 个基因从 TS2 到 TS4 持续表达（图2.9[①]（A））。217 个基因在早期的 TS（Theiler stages）1-5 中的胚胎细胞可检测到表达（图2.9（B））。

（C）例如 $Actr3Gt_{tm2Knk}/Actr3Gt$ $(A009F03)VauA$ $(A009F03)VauA$，正常的胚胎发育、停滞和致死（图2.9 217 个早期表达的基因的敲除表型集中于不

，小鼠表现出胚胎发育停滞。$Adar$ $/Adar$ 小鼠表现出胚胎红细胞生成异常和肝发育异常。这 217 个基因多参与到调节基因表达和细胞分裂的生物学过程中（图2.9（D））。

[①]（A）在 SZ 中启动子位于环端点的 1572 个基因在早期胚胎发育各时期（Theiler Stage 1-5）中的表达数量。

（B） 在 SZ 中启动子位于环端点的 1572 个基因在早期胚胎发育各时期（Theiler Stage 1-5）中的表达情

况。

（C） 在 SZ 中其启动子位于环端点且在胚胎发育早期表达的 217 个基因的 GO 哺乳动物表型富集分析的结果。

（D） 在 SZ 中其启动子位于环端点且在胚胎发育早期表达的 217 个基因的 GO 生物学过程富集分析的结果。

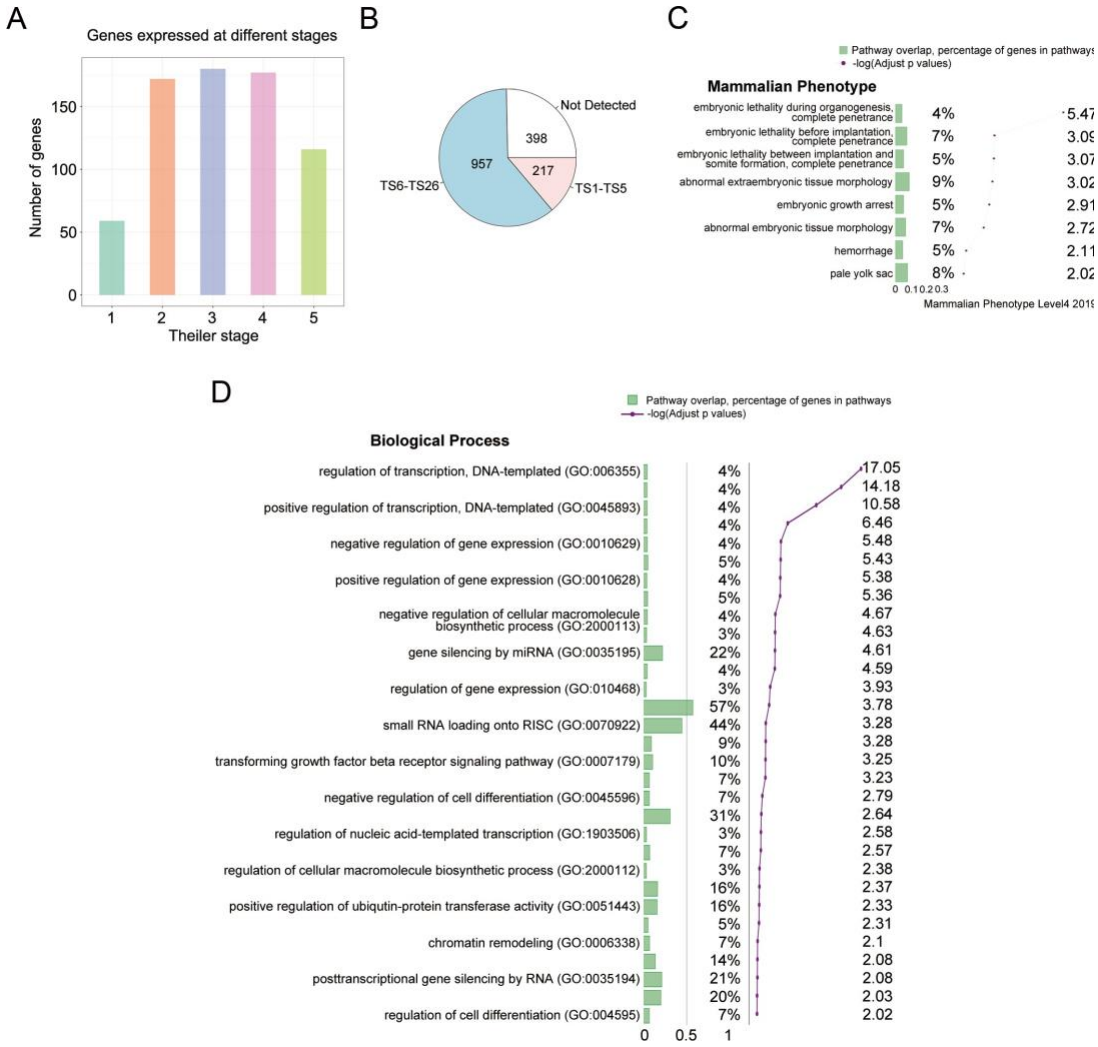


图 2.9　SZ 中建立的染色质环在早期胚胎发育中的潜在作用

本节的结果提示 SZ 中重新组织的染色质环协助了胚胎发育过程中细胞分化需要的基因的表达。

2.3　实验结果与讨论

已有研究表征了哺乳动物精子发生过程中 TADs 和染色质区室的动态变化。在本研究中，我们结合 Hi-C，ATAC-seq，ChIP-seq，和 RNA-seq 数据，探索了染色质构象的动态变化以及它们对转录调控的作用。本研究为已有的科学认知增加了染色质环在此过程中与TADs 类似的动态变化、A/B 区室的转化与减数分裂特异的 mRNAs 和 piRNAs 的表达有关。

为研究染色质构象参与转录调控的机制，本研究考察了染色质可及性、CTCF 和黏粘蛋白在 TAD 边界上的结合，发现它们与粗线期时期 TAD 和 loop 的消失无关。在单细胞高分辨率成像的染色质示踪下可观察到，TADs 边界通常形成于

CTCF 和黏粘蛋白结合位点附近。将黏粘蛋白从染色质上去除后，在单细胞内，仍然有类TAD 结构，但在群体细胞水平，观察不到 TADs 信号了。化学降解黏粘蛋白后，单细胞中的 TADs 倾向于随机分布。这些结果提示黏粘蛋白虽然可能并非形成 TADs 所必须的，但却是形成适合的 TAD 边界位置所必须的。而在原始精细胞中，黏粘蛋白介导的转录可能有助于基因表达和 DSBs 的形成。而在本研究中，在精子发生过程中的 pacSC 中，TADs 消失的同时，CTCF 和黏粘蛋白仍然结合在 priSG-A 中的 TAD 边界位置，这一发现可能有几种原因。一方面，在减数分裂过程中可能由别的分子调控TAD 结构。将来结合三维构象、免疫沉淀、染色质分离和质谱等方法，设计更好的控制变量实验，可能找到这些未知分子。另一方面，在更高的分辨率上，更小尺度的 microTADs 的形成是基因依赖的，并于染色质可及性和转录活性相关。因此可能在 pacSC 中，利用更高分辨率的测序技术 (如 micro-C)，可以检测到 microTADs 信号。

已有研究发现较小尺度上的 A/B 区室的转换与精子发生过程中特定基因的表达有关。本研究进一步发现在从 priSG-A 到 pacSC 发生的 A/B 区室转换与其中基因表达变化的关系。我们还发现在 priSG-A 中位于 A 区室的 piRNA 簇的区室状态更为稳定、在种群间也更为保守，而大量的在 priSG-A 中位于 B 区室的 piRNA 簇到了 pacSC 中的状态就变为了 A 区室。这与这些 piRNAs 在 pacSC 中活跃表达的情况一致。也即，本研究指出染色质区室化状态的转换可能影响 piRNAs 的活性。

在染色质三维构象与其可及性的关系方面，本研究发现在 pacSC 中，染色质可及性没有发生显著的变化，则染色质可及性可能与 pacSC 中 TADs 和环的消失无关。另一方面，在pacSC 中的A 区室中富集有减数分裂 DSB 位点，而一些研究提示染色质可及性可能与减数分裂 DSBs 相关，但在 pacSC 中，可及性区域与减数分裂 DSBs 的重叠较少。本研究发现许多减数分裂 DSB 位点都位于开放区域内，并且在 priSG-A 中就已提前开放。

根据已有的结果，减数分裂和有丝分裂过程中的染色质构象变化存在的异同大致有如下几点: 从 P(s) 曲线图来看，总体上，它们分裂间期的构象是类似的。减数分裂过程中染色体能观察到 A/B 区室的存在，但在有丝分裂中已没有区室的存在。比较减数分裂和有丝分裂过程中的因此可能发现区室化调控因子。在 TADs 层面，有丝分裂和减数分裂中 TADs 的维持机制可能有所不同，因为有丝分裂染色质中TADs 和 CTCF 结合的消失是同步的。另外，减数分裂中转录活动仍在进行，而有丝分裂中转录活动停止。转录活动的差异与三维构象的差异可能有某个方向的调控关系。

总之，本研究发现了小鼠精子发生过程中区室化强度、TADs 的动态变化；揭示了染色质三维构象、CTCF 或黏粘蛋白结合的功能关系；显示了 A/B 区室转化

在减数分裂基因表达调控中的功能以及 SZ 中染色质环对随后胚胎发育的潜在功能。

2.4　实验材料与方法

2.4.1　实验数据

本研究使用的数据存放于 NCBI 基因表达文库中，序列号为 GSE147536。

2.4.2　实验方法

1. Hi-C: 序列比对

将 Hi-C 文库测序得到的双端`.fastq`文件用 `HiC-Pro (v2.8.1)` 进行比对。用 `cooler (v0.8.6.post0)` 生成各分辨率矩阵并进行矩阵平衡。具体步骤有：

1　利用 `bowtie2 (v2.2.5)` 将双端序列比对到 Mus musculus mm10 参考基因组上。

2　去掉含有 Mboi 连接位点但未比对上的序列的末端部分序列，再次比对到 mm10 基因组上。

3　合并两次比对结果，去掉其中的 PCR 重复和光学重复序列并将剩下的序列与 Mboi 限制酶切片段对应起来。

4　将来自同一酶切片段的双端序列筛去，得到有效互作对。

经过这四个步骤，能得到表 (Table S1) 中描述的测序数据。

将生物学重复合并起来，并从每组样本的可靠互作对中随机抽取了等量的序列（172252595），再用这些序列对在 100kb, 50kb, 25kb 和 5kb 尺度上分别建立 Hi-C 互作矩阵。得到的各分辨率矩阵进一步进行迭代校正。随后将矩阵转换为`.hic`格式，以应用 `juicebox` 进行可视化。

2. Hi-C: 互作频率曲线

在 100kb 分辨率下对校正后的互作矩阵根据 () 方法计算互作频率曲线 $P(s)$。首先，互作的距离被分成对数区间，并计算在相应距离区间内的互作次数。用互作次数该距离上所有可能的互作数即得到 $P(s)$。

3. Hi-C: A/B 区室、TADs 和染色质环

（1） A/B 区室

在 50kb 分辨率下的 Hi-C 矩阵的区室数据由 cooltools [①]的脚本 eigdecomp.py 计算得到。步骤如下：

1. 计算染色体内的 o/e 互作矩阵。在 500kb 分辨率下计算期望矩阵，也即平

[①]https://github.com/mirnylab/cooltools

均在此分辨率下各个互作距离间的，随着距离线性增长的区域范围内的互作频率，则为该距离 d 间的期望互作频率。

2 计算 o/e 矩阵的 Pearson 相关矩阵，再对相关矩阵进行 PCA(主成分分析)。

3 利用第一主成分的特征值来区分 A/B 区室。

按照惯例，根据染色质活跃状态来确定 A(活性、真染色质区室) 和 B(一直、异染色质区室)。利用 cooltools 进行区室的鞍形分析。具体步骤有：

1. 对每个 o/e 互作矩阵，重排矩阵的行和列以得到使各列的特征值按递增的方式排列。

2 将得到的每个矩阵粗粒化为 30 x 30 的矩阵，并合并所有矩阵。

3 对得到的矩阵以热图形式可视化便得到区室图（鞍形图）。

区室化强度的计算过程如下：[①]

$$I_{AA} + I_{BB} - 2I_{AB} \tag{2.1}$$

其中 strong A 区域定义为有最高 25% 的 PCA1(第一特征值) 的区域，strong B 区域定义为具有最低 25% PCA1 值的区域。

（2） TADs 位置的确定

采用 IS（绝缘值）方形分析确定 TADs 的位置。具体步骤如下：

1. 沿着 25kb 分辨率的矩阵的每条染色体对角线分别计算每隔 100kb, 1000kb 和 50kb 步长位置附近 125kb 内的互作平均值。

2 对位置求导得到 delta 向量，保留边界强度大于 0.25 的边界。

3 将所有 priSG-A 中的 TADs 及其附近区域（±0.5 TAD 长度）上的 IS 平均起来对位置作图。

4 将各个时期细胞在 priSG-A 中的 TAD 边界位置附近的互作频率平均起来，以考察各个时期细胞相对 priSG-A 中的 TAD 位置和强度的变化。

（3） 染色质环的确定

应用 cLoops，以参数-w -j -s -m 3 来确定染色质环。参考（）进行 APA（峰集合分析），具体过程如下：对所有 loop 及其附近 50kb x 50kb 区域的 KR 标准化后的互作频率求和并以热图形式呈现。通过将区域中心的值除以左下角位置（25kb x 25kb）的平均值，得到 APA 值。

下载活性启动子（H3K4me3 或 Pol II 标记区域）和增强子（H3K4me1 或 H3K27ac 标记区域）的数据，并用它们的位置来注释每个时期细胞中的染色质环端点。随后将有增强子-启动子环连接的基因在 MGI-小鼠表达数据库中查询，并筛选其中在 Theiler Stage 1-5 期间可检测到表达的基因。

① I_{AA}：strong A-strong A interactions

BB：strong B-strong B interactions

I_{AB}：strong A-strong B interactions

4. ATAC-seq 数据分析

用 snap-aligner (v1.0) 将 ATAC-seq 序列比对到mm10 参考基因组上。用 samtools 对比对得到的文件进行排序、生成索引以及去除重复序列。以默认参数对这些比对文件允许 MACS2 (v2.1.0) 找到 ATAC-seq 峰的位置。对各个时期样本中在峰位置测到的序列数（RPKM 标准化）计算 Pearson 相关系数。利用 bedtools 对找到的峰计数，利用 R 中的 quantile norm 函数对其标准化。然后利用 DESeq2 找到差异可及性区域。利用 GREAT 对这些差异可及性区域做功能富集分析。利用 HOMER 中的 findMotif 函数对这些差异可及性区域做模序富集分析。通过已公开的 DMC1 ChIP-seq 数据得到减数分裂 DNA 双链断裂位点数据。在 piRNA 簇数据库中获取 14.5dpp 小鼠中的piRNA 簇位点。在 ucsc genome browser[①]中下载 phastCons 值。

5. ChIP-seq 数据分析

对 Poll II S2P , CTCF, Rad21, protamine 1 (从 GSM2088400 和 GSM2401441 中获取)，histone H3 (从 DRA006537 中获取) ChIP-seq 测序数据，首先用 NGmerge (v0.3) 去除它们的接头序列，然后用 BBMap (v37.62) 中的 BBDuk 去除右端质量低的序列。利用 bowtie2 (v2.2.6) 将序列比对到 mm10 参考基因组上，然后用 samblaster (v0.1.24) 去除重复序列。

用 sambamba (v0.7.0) 在 Pol II S2P ChIP-seq 和 Rad21 ChIP-seq 的比对序列中分别抽取 21445601 条， 并采用参数 -normalizeUsing RPKM --operation ratio --binSize 30 --smoothLength 300 --scaleFactorsMethod None -- extendReads 200， 运行 Compare (v3.3.0)，对这些同样数量的序列相比各自的 Input 比对数据的多少进行比较。

对 protamine 1 和 histone H3 ChIP-seq 数据，利用 sambamba (v0.7.0) 比对，并采用参数 -normalizeUsing RPKM --operation subtract --binSize 30 --smoothLength 300 --scaleFactorsMethod None -- extendReads 200，运行 Compare (v3.3.0)，对这些同样数量的序列相比各自的 Input 比对数据的多少进行比较。

mESC 样本的 CTCF 和 Rad21 ChIP-seq 数据来自于 GSE102997。

用默认参数运行 MACS2 (v2.1.0) 以找到 CTCF ChIP-seq 的峰。用 deepTools (v3.3.0) 中的 computeMatrix 和 plotProfile 函数

计算 TAD 边界附近的 CTCF 信号的平均值。用 `pyGenomeTracks` 完成这部分的所有作图。

[①]http://hgdownload.cse.ucsc.edu/goldenPath/mm10/phastCons60way/mm10.60way.phastCons.bw

6. RNA-seq 数据分析

本研究用到的 RNA-seq 数据来自于（）。用 fastQC 对双端测序文件进行质量控制。用 hisat2 将质控后的序列比对到mm10 参考基因组上。用 samtools 去除重复或比对到多个位点的序列。用 stringtie 注释并对比对的序列计数。用 R 中的 DESeq2 包找到差异表达基因。用 enrichR 对基因集合作基因本体分析。

第 3 章　小鼠胚胎干细胞中的拓扑结合域的分类研究

3.1　背景

在真核生物中，基因组构象对以 DNA 为模板的过程（例如转录和 DNA 修复）具有广泛的影响。但是，当前基因组构象分析较为粗略，无法与精细的转录调控对应。在染色质不同尺度的构象中，TAD 的基因组大小与增强子-启动子互作之间的典型基因组距离是可比的，因此许多研究关注于 TAD 的结构与功能。关于TAD 的热点问题有各种细胞类型的动态折叠特性，以及关于 TAD 作为

功能增强剂调节基因的功能性中介物的必要性。具体而言，一些遗传扰动研究结果支持了一种TAD 参与转录调控的模型。在该模型中，TAD 创建了隔离的邻域，为增强子划分出了搜索靶基因的空间。另一方面，一些研究不支持该功能模型。例如，快速降解 cohesin 的亚基或 CTCF 耗竭对基因表达的影响不大。在发育研究中，一些研究者在 *Sonic Hedgehog*（*Shh*）基因座处发现了与小鼠肢体发育相关的 TAD 结构。但是，已经发现，缺失该结构附近的 CTCF 位点或围绕 Shh 调节性 ZRS 的边界的 35kb 区域对 *Shh* 表达的影响较小，并且没有导致明显的发育缺陷。

TAD 的功能差异可能来自其不同的内部结构。尽管TAD 边界在不同细胞类型之间变化不大，但利用新的技术，可获得超高分辨率染色质互作图，从这些图上能观察到在不同细胞类型间有动态变换的 sub-TAD（亚兆碱基级拓扑结合域）。TAD 的差异也可能源于它们的形成机制。染色质首先以位于同一条染色体上的基因组序列在物理上彼此靠近的方式组织成染色质环。在计算建模和蛋白缺失实验的支持下，TAD 可能由多个动态环的挤出过程建立。目前认为这样的环挤出过程可能地由 CTCF 结合位点限制也可能由转录机器的活动限制。

利用导致TAD 功能差异的这些潜在原因构造输入变量，我们将TAD 分为 7 种类型，并评估它们在细胞类型间的动态性，表达活性和参与干细胞分化的情况。

3.2　研究成果

321　　对 TADs 和 TAD 边界的初步分类

收集了对E14TG.2a 细胞系测得的 Hi-C、ChIP-seq 和 RNA-seq 数据并进行初

步处理后，我检验了数据集比对到基因组上的结果间的重复性（图3.1[①]（A-C）。

[①]（A）研究中所用的各组 Hi-C 数据在基因组上的覆盖度的 Pearson 相关系数热图。

(B)　研究中所用的各组 ChIP-seq 数据在基因组上的覆盖度的 Pearson 相关系数热图。

为得到 mESC 中 TAD 边界的可靠位置，我从高重复性的三组 Hi-C 的原始数据集中得到 TAD 边界，并保留其中在至少两组数据集中都出现，且在 blacklist 区域或 Y 染色体外的 TAD 边界（图3.1（D，E）。

本研究中的分类分析用到了从高重复的数据集中计算得到的 12 个变量。其中，TAD 边界上的表达水平、TAD 的大小都是单峰分布的，而 TAD 内部的平均第一特征值的分布是双峰的（图3.2)。为避免对样本间距离的计算以及随后的分类结果过于倾向于依赖TAD 内部的平均第一特征值的大小，我将每个 TAD 的两个边界上基因的平均表达水平、TAD 的大小和TAD 内互作矩阵的第一特征值标准化到同样的总值。这三个变量之间线性相关性很低（图3.1（F）。对这三个变量进行聚类分析，可将 2117 个 TADs 分为三类（SA 类、A 类和 S 类）（图3.3[①]（A,C）。根据输入的分类信号，S 类 TADs 大多位于 B 区室；S 类 TADs 主要位于 A 区室，并且其边界上的表达量很高；而A 区室虽然也主要位于A 区室，但其边界上的表达量没有那么高。

ChIP-seq 的信号有较高的共线性、不能被认为是完全的独立变量。从这些 ChIP-seq 信号中利用隐马尔可夫模型可以推断出染色质状态（图3.3（B）。对边界上的染色质状态进行聚类，可将 TAD 边界分为两类（图3.3（D）。其中 I 类边界（其染色质状态多为活性状态）上有更多的 SINE 序列（图3.3（E）。I 类边界上也有更多的顺式调控元件（图3.3（F-I）。

Ⓒ　研究中所用的各组 RNA-seq 数据在基因组上的覆盖度的 Pearson 相关系数热图。

Ⓓ　利用三组数据分别找出的 TAD 边界及其与不能被良好测量的区域及 Y 染色体的交集内元素的数量。

Ⓔ　利用三组数据分别找出的常染色体上位于可被良好测量的 TAD 边界的交集。

Ⓕ　对 TAD 计算的 12 种变量的相关性热图。

[①]（A）将TADs 根据边界上基因的的平均表达水平、TAD 大小和TAD 内的平均第一特征值聚集聚类为三类。

Ⓑ　根据 ChIP-seq 信号推断染色质的 12 种状态。

Ⓒ　对 1 号染色体 156M-166M 的可视化展示三类 TADs。

Ⓓ　将 TAD 边界按其染色质状态的分布聚集聚类为两类。

Ⓔ　两类边界上分别的重复序列的数量。

Ⓕ　两类边界上 CTCF 结合位点数量的分布。

（G-I）两类边界上远程类增强子特征、近程类增强子特征和类启动子特征数量的分布。

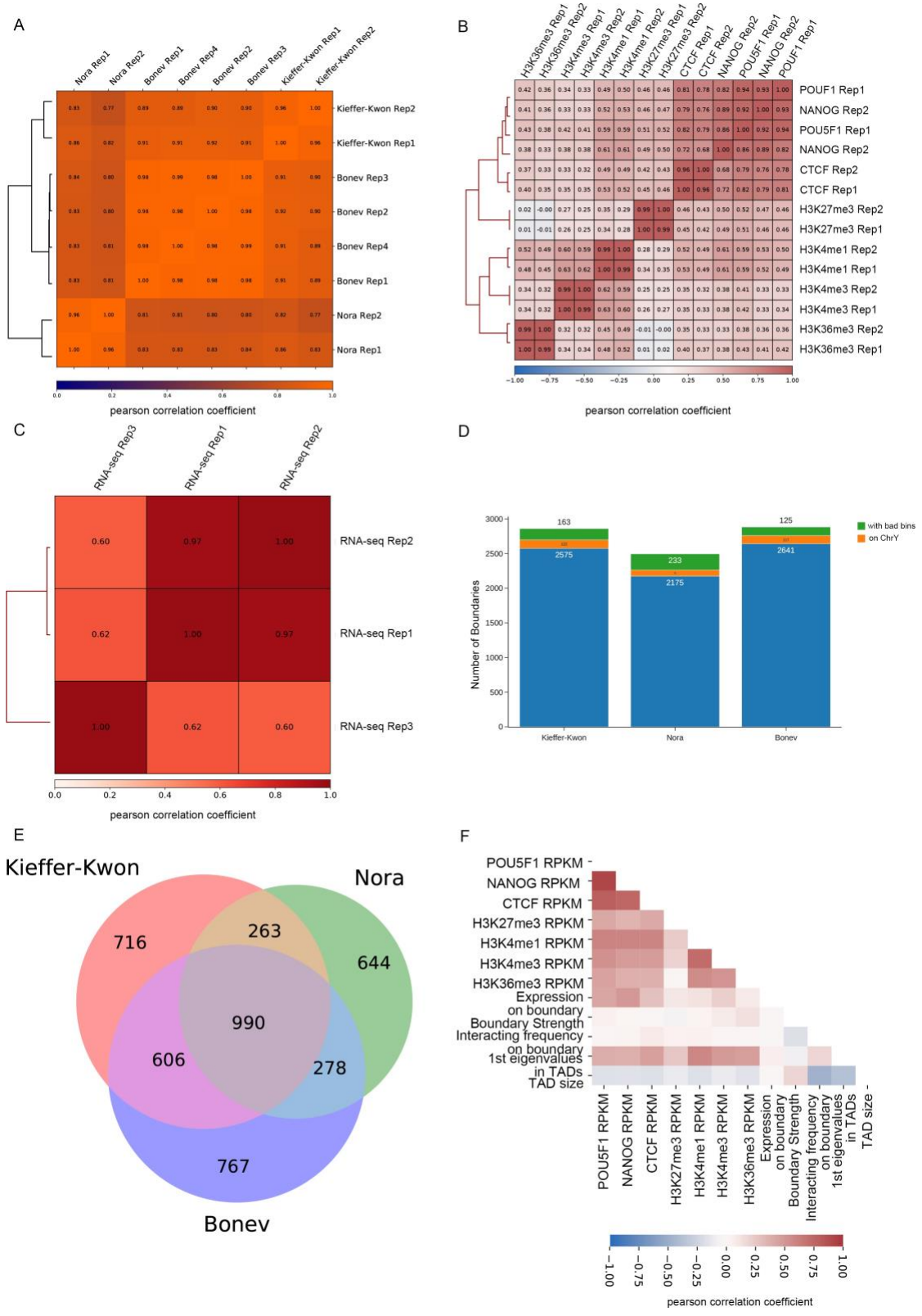

A
Nora Rep1
Nora Rep2
Bonev Rep1
Bonev Rep4
Bonev Rep2
Bonev Rep3
Kieffer-Kwon Rep1
Kieffer-Kwon Rep2
Kieffer-Kwon Rep2
Kieffer-Kwon Rep1
Bonev Rep3
Bonev Rep2
Bonev Rep4
Bonev Rep1
Nora Rep2
Nora Rep1
pearson correlation coefficient
B
H3K36me3 Rep1
H3K36me3 Rep2
H3K4me3 Rep1
H3K4me3 Rep2
H3K4me1 Rep1
H3K4me1 Rep2
H3K27me3 Rep1
H3K27me3 Rep2
CTCF Rep1
CTCF Rep2
NANOG Rep2
POU5F1 Rep1
NANOG Rep2
POUF1 Rep1
POUF1 Rep1
NANOG Rep2
POU5F1 Rep1
CTCF Rep2
CTCF Rep1
H3K27me3 Rep2
H3K27me3 Rep1
H3K4me1 Rep2
H3K4me1 Rep1
H3K4me3 Rep2
H3K4me3 Rep1
H3K36me3 Rep2
H3K36me3 Rep1
pearson correlation coefficient
C
RNA-seq Rep3
RNA-seq Rep1
RNA-seq Rep2
RNA-seq Rep2
RNA-seq Rep1
RNA-seq Rep3
pearson correlation coefficient
D
Number of Boundaries
with bad bins
on ChrY
163
2575
233
2175
125
2641
Kieffer-Kwon
Nora
Bonev
E
Kieffer-Kwon
Nora
Bonev
716
263
644
990
606
278
767
F
POU5F1 RPKM
NANOG RPKM
CTCF RPKM
H3K27me3 RPKM
H3K4me1 RPKM
H3K4me3 RPKM
H3K36me3 RPKM
Expression on boundary
Boundary Strength
Interacting frequency on boundary
1st eigenvalues in TADs
TAD size
pearson correlation coefficient

图 3.1　TADs 位置的确定以及与 TADs 有关的变量

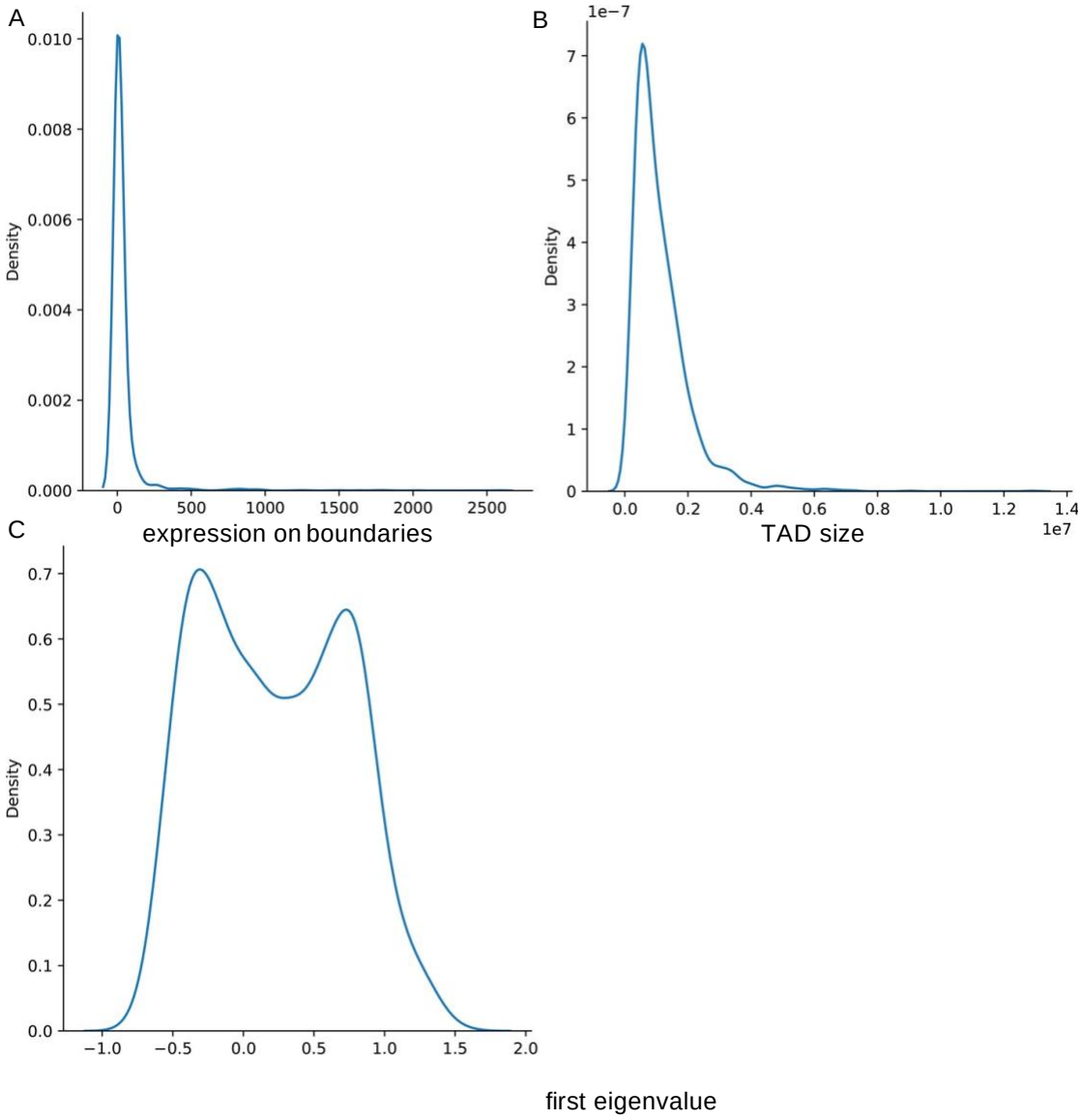


图 3.2　TAD 边界上的表达水平、TAD 的大小和 TAD 内部的互作矩阵的平均第一特征值的分布。

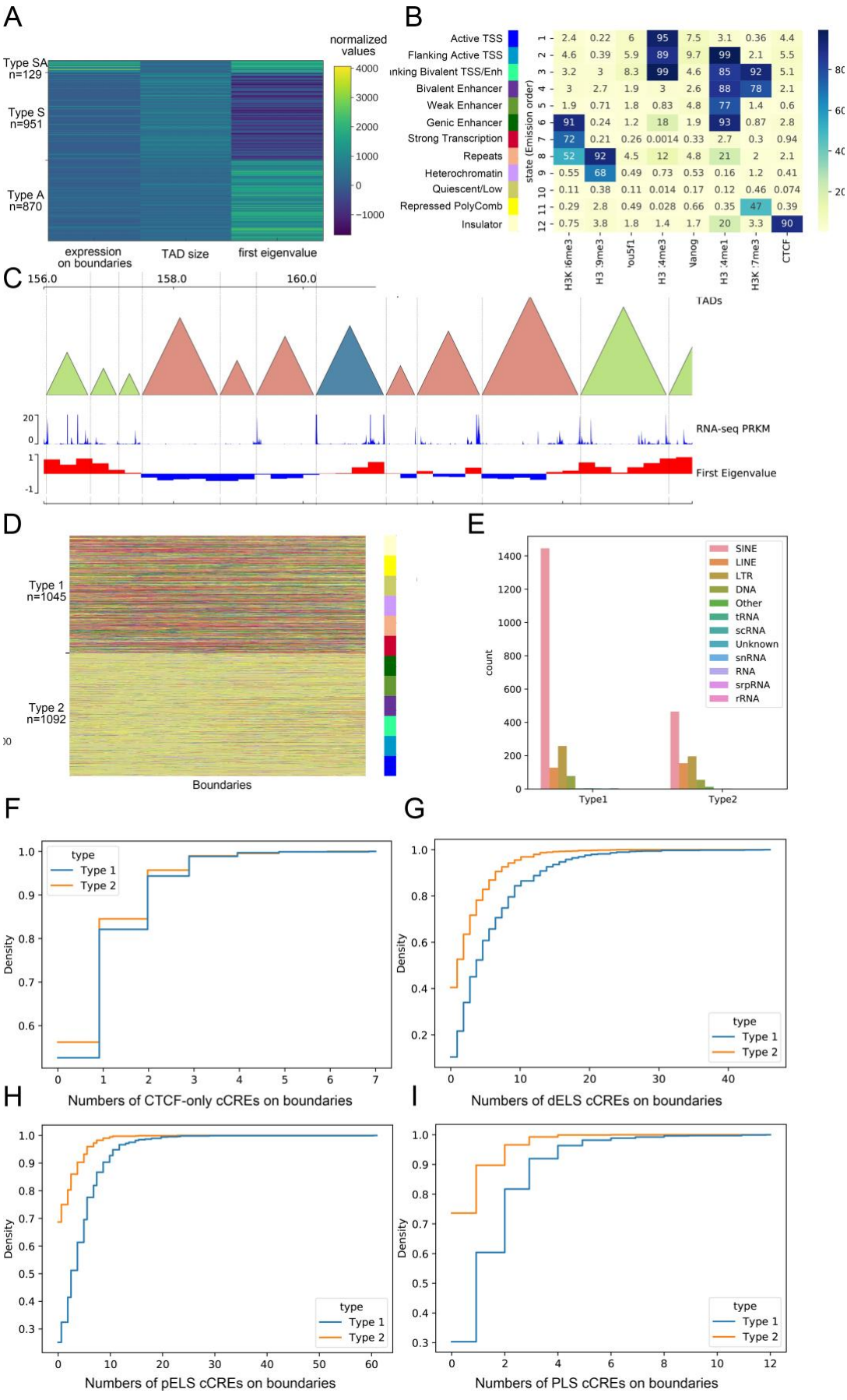

A
Type SA
n=129
Type S
n=951
Type A
n=870
normalized values
4000
3000
2000
1000
0
−1000
expression on boundaries
TAD size
first eigenvalue
B
Active TSS
Flanking Active TSS
Bivalent Enhancer
Weak Enhancer
Genic Enhancer
Strong Transcription
Repeats
Heterochromatin
Quiescent/Low
Repressed PolyComb
Insulator
state (Emission order)
H3K 6me3
H3 9me3
H3 4me3
Nanog
H3 4me1
H3K 7me3
CTCF
C
156.0
158.0
160.0
TADs
RNA-seq PRKM
First Eigenvalue
D
Type 1
n=1045
Type 2
n=1092
Boundaries
E
SINE
LINE
LTR
DNA
Other
tRNA
scRNA
Unknown
snRNA
RNA
srpRNA
rRNA
count
Type1
Type2
F
type
Type 1
Type 2
Density
Numbers of CTCF-only cCREs on boundaries
G
Numbers of dELS cCREs on boundaries
H
Numbers of pELS cCREs on boundaries
I
Numbers of PLS cCREs on boundaries

图 3.3　对 TADs 和 TAD 边界的分别聚类分析

3.2.2　结合对 TADs 和 TAD 边界的初步分类将 TADs 分为七类

A 类和 S 类 TADs 可根据它们边界的类型分别分为三类。这样，TADs 就被分为了七类。为验证这样的分类是否有生物学意义，我考察了这七类 TADs 的生物学特征的分布并检验不同类别间是否存在显著性差异。在表达活性方面，SA 类 TADs 中的基因的表达水平最高。S 类 TADs 中的基因比 A 类 TADs 中的基因的表达水平低。在A 类和S 类 TADs 中，两个边界都是I 类边界的TADs 中的基因表达水平更高（图3.4[①]（A）, 图3.5[②]（A））。SA 类和 A 类 TADs 各自的两个边界互作更频繁。A 类 TADs 中两个属于不同类别边界相互之间的互作最强，但S 类 TADs 没有这种趋势(图3.4（B）, 图3.5（B）)。平均来看，两个边界的类型不同的TADs 以及 SA 类 TADs 的边界强度较低（图3.4（C）, 图3.5（C））。类别间 TADs 之间的互作方面，SA 类 TADs 与 SA 类 TADs 之间的互作更频繁，而互作最不频繁的是 S 型 TADs 之间（图3.4（D），图3.5（D））。这些结果显示 SA、 A 和 S 类 TADs 在它们的结构特征上可以区分，但更细节的亚类在 TAD 内的表达水平上具有更高的区分度。

[①]（A）七类 TADs 中基因的平均表达水平。

（B）TADs 的两个边界间的互作频率。

（C）TADs 的边界强度。

（D）同类别的 TADs 的平均互作频率。

[②]（A）七类 TADs 中基因的平均表达水平之间的 Mann-Whitney rank test 的 p 值。

（B）TADs 的两个边界间的互作频率之间的 Mann-Whitney rank test 的 p 值。

（C）TADs 的边界强度之间的 Mann-Whitney rank test 的 p 值。

（D）同类别的 TADs 的平均互作频率之间的 Mann-Whitney rank test 的 p 值。

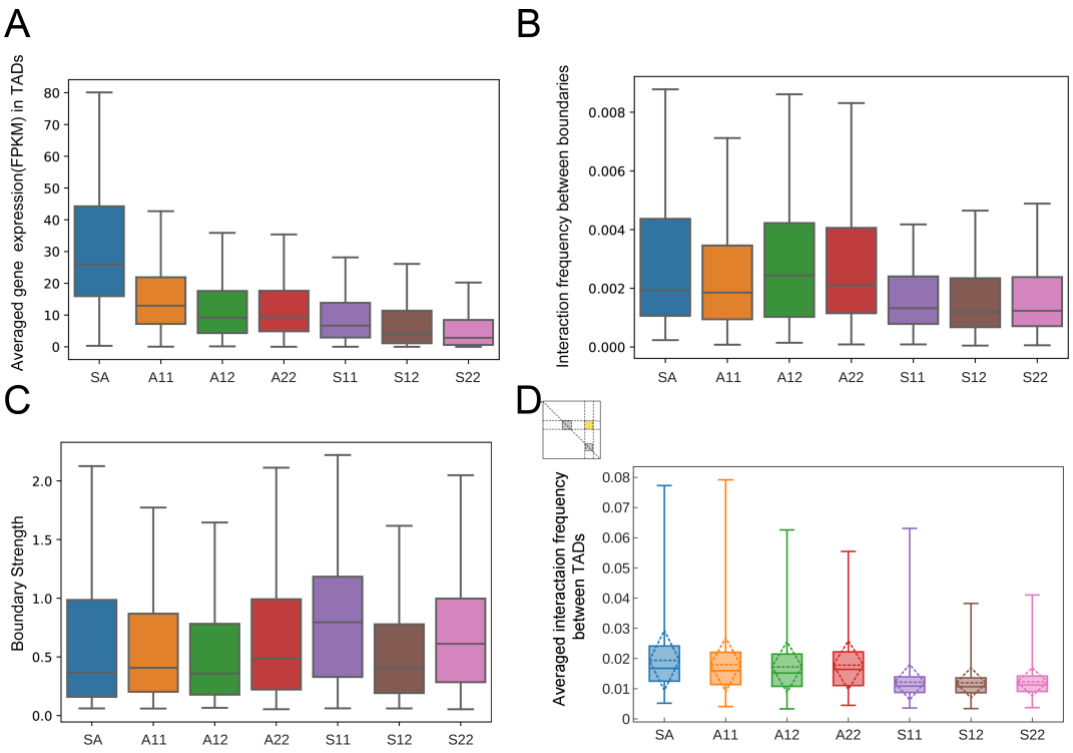


图 3.4　七类 TADs 的特点


-log10(p-value)

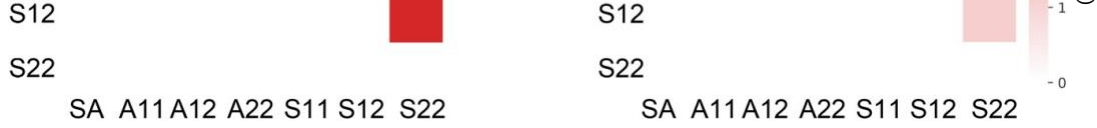


图 3.5　七类 TADs 的特点的同分布假设检验

3.2.3　七类 TADs 各自在胚胎干细胞分化过程中的表现

为进一步研究这一分类对基因组功能的提示，在基因本体 terms 中选取了“干细胞分化”和“干细胞增值”中的基因集并研究它们的分布。参与干细胞分化的基因（n=173）更多地位于 A11（具有两个 1 类边界的 A 类 TADs）和 S22 类 TADs（具有两个 2 类边界的S 类 TADs）中（图3.6[①]（A），而参与干细胞增值的基因 (n=29) 更多地分布于 A22 类 TADs（具有两个 2 类边界的 A 类 TADs） 中（图3.6（B）。

为研究 mESC 中各类别 TADs 在 mESC 分化为 NPC（神经前体细胞）后的稳定性以及细胞类型变化性，我构造了作为基因组不同划分的 TADs 集合之间的 Jaccard 指数。mESC 种某 TAD 的 Jaccard 指数越高，该 TAD 在另一细胞类型中越能找到位置与其一致的 TAD。该指数的分布提示，SA 类的 TADs 在不同细胞

[①]（A）干细胞分化相关的基因在不同 TADs 上的分布。

（B）干细胞增殖相关的基因在不同 TADs 上的分布。

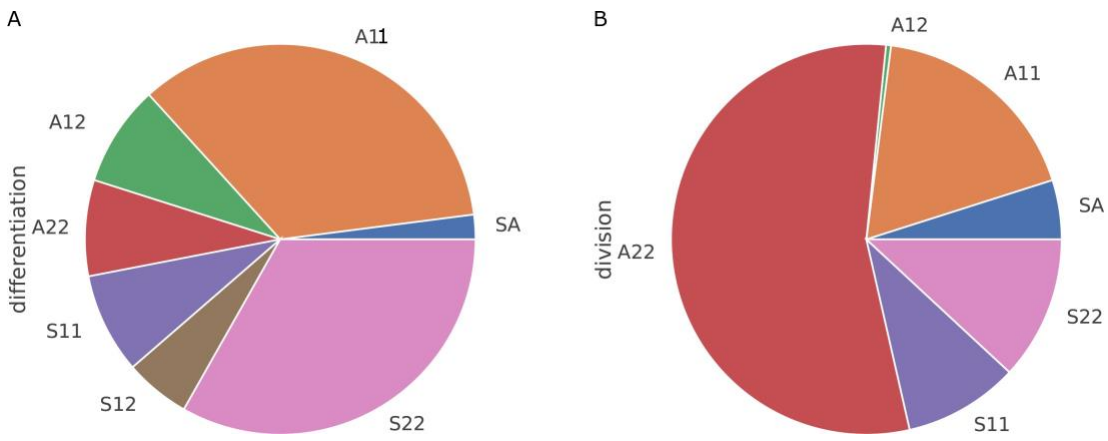
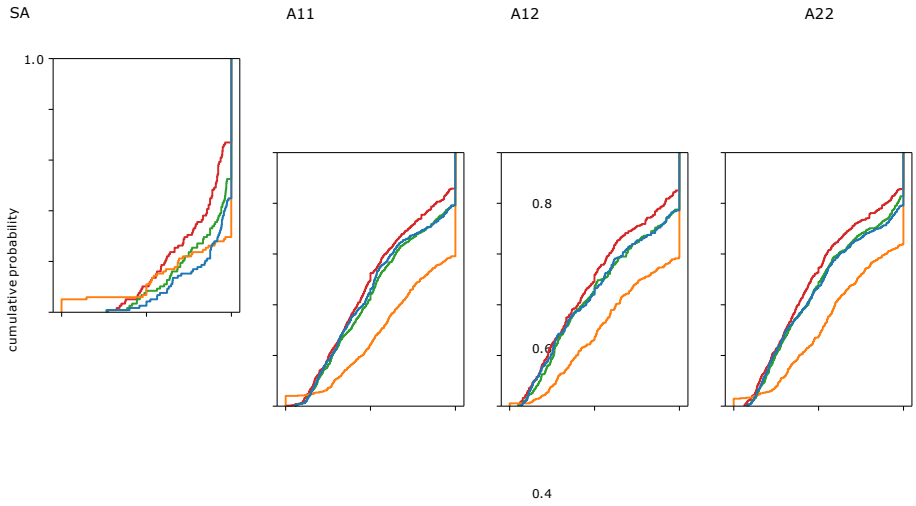


图 3.6　干细胞增殖与分化相关的基因在不同 TADs 上的分布

类型间都更加稳定，具有两个不一样类别的边界的 S 类 TAD 也比较稳定; 在细胞类型间不稳定的 TADs 类别，相比配子及 PN3（原核合子），与 mESC 分化的 NPC 中的TADs 更为相似（图3.7）。与 NPC 中的TADs 最不相似的是 A11 和 S22 类，除了最稳定的 SA 类 TADs，与 NPC 最相似的是 A22 类 TADs。将来对这些类别的 TADs 的深入研究可能提示我们关于干细胞定向分化为神经前体细胞的分子机制。



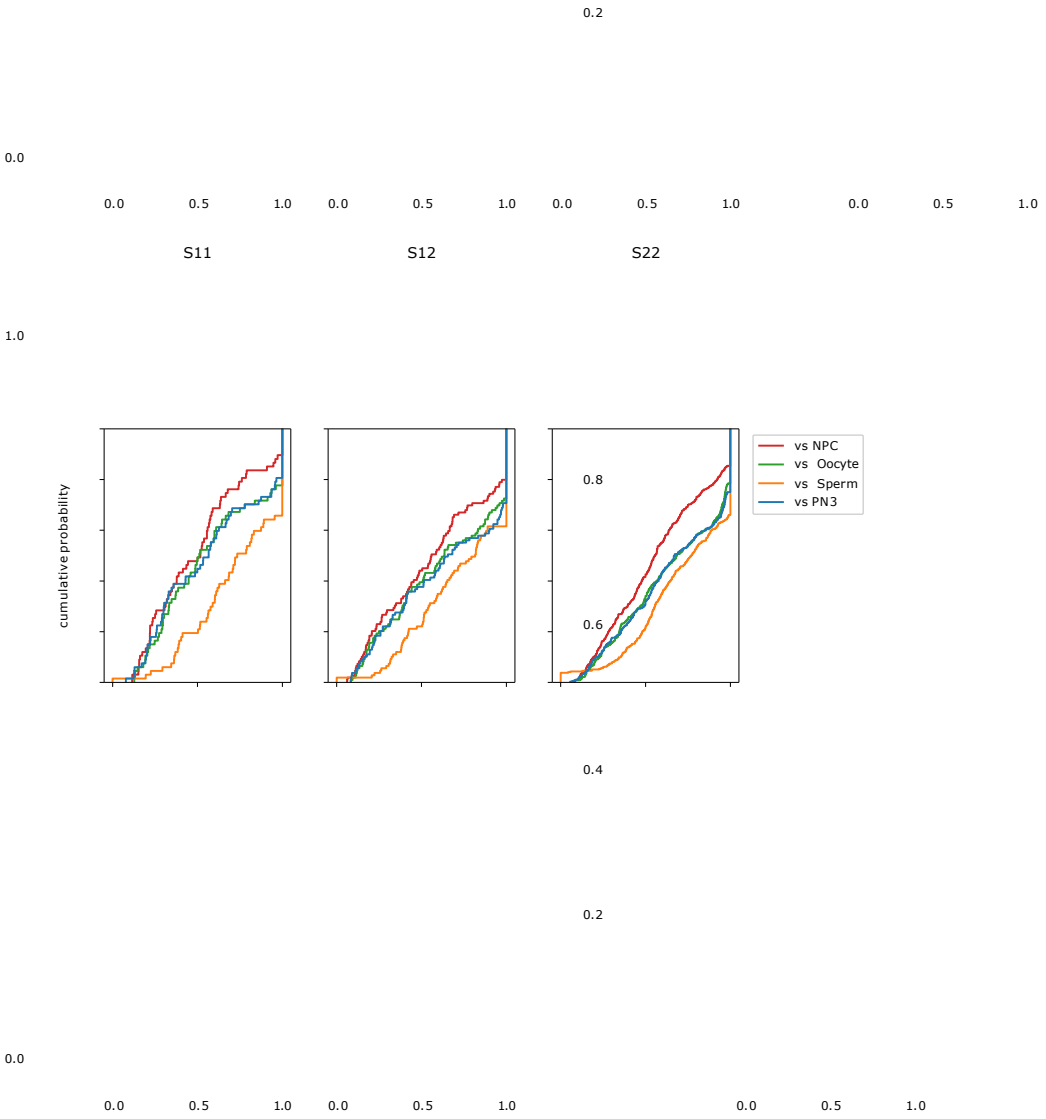


图 3.7　SA 类 TADs 在细胞间更稳定

比较不同类别的 TADs 在五种细胞类型间的成对相似性。

3.3　实验结果与讨论

在真核生物的染色质高级结构中，TADs 的尺度与增强子作用的范围比较一致，被认为可能帮助增强子发挥界限明确的功能。但是，TADs 的数量众多，例如在小鼠细胞内就有超过两千个 TADs。已有的研究表明，有的 TADs 对特定的生物学过程是必需的，有的却显得无关紧要。因此，对TADs 的分类和注释可能补充我们对 TADs 在结构与基因组信息传递中作用的认识。

在本研究中，我计算了 mESC 样本的 Hi-C 互作频率、组蛋白修饰和转录因子结合信号以及RNA-seq 信号，并以此分别对 TADs 和 TAD 边界进行聚类分析。这一工作成功将 TADs 分为了七类，并且发现了 SA 类的 TADs 在细胞间最为稳定，mESC 中的 TADs 在与其分化的 NPC 中的 TADs 更为类似的同时，S22 类 TADs 与 NPC 中的最为不同，可能与干细胞定向分化有关。这七类 TADs 内的基因有较明显的表达趋势，在 SA 类中最为活性，位于S 类 TADs 中的基因最为沉默。在功能方面，发现 A11 和 S22 类 TADs 可能与干细胞分化有关，而 A22 中的基因可能参与干细胞增殖。

将来还可以在本研究的基础上，选取方便获取的变量，继续建立可解释的或可用于预测的有监督模型，用以考察其他细胞系中的 TADs 类别。根据预测的 TADs 类别，缩小与想要研究的生物学过程相关的 TADs 数量，对相应的 TADs 进行实验研究，以验证并准确定位具有表达调控功能的 TADs。

3.4　实验材料与方法

3.4.1　实验材料

本研究所用数据均来源与公共数据库。数据来源如表 3.1-3.3。

表 3.1　Hi-C 数据集

ExperimentSet	Organism	Biosource Type	Experiment Type	Assay Details
4DNES87HWQAX	*M.musculus*	129 E14TG2a.4	Dilution Hi-C	Enzyme: *Hin*dIII
GSE96107	*M.musculus*	129 E14TG2a.4	*in-situ* Hi-C	Enzyme: *Dpn*II
GSE82144	*M.musculus*	129 E14TG2a.4	*in-situ* Hi-C	Enzyme: *Mbo*I
GSE116856	*M.musculus*	sperm	*in-situ* Hi-C	Enzyme: *Dpn*II

GSE82185	*M.musculus*	MII oocyte and embryo	*in-situ* Hi-C	Enzyme: *Mbo*I
4DNESJ9SIAV5	*M.musculus*	mouse neural progenitors cells	*in-situ* Hi-C	Enzyme: *Dpn*II

表 3.2　ChIP-seq 数据集

ExperimentSet	Organism	Biosource Type	Experiment Type	Target
GSE136488	*M.musculus*	129 E14TG2a.4	TF ChIP-seq	CTCF
GSE136489	*M.musculus*	129 E14TG2a.4	TF ChIP-seq	POU5F1
GSE136517	*M.musculus*	129 E14TG2a.4	TF ChIP-seq	NANOG
GSE136479	*M.musculus*	129 E14TG2a.4	Histone ChIP-seq	H3K4me3
ENCSR059MBO	*M.musculus*	129 E14TG2a.4	Histone ChIP-seq	H3K27me3
ENCSR253QPK	*M.musculus*	129 E14TG2a.4	Histone ChIP-seq	H3K36me3
GSE136526	*M.musculus*	129 E14TG2a.4	Histone ChIP-seq	H3K9me3
GSE136454	*M.musculus*	129 E14TG2a.4	Histone ChIP-seq	H3K4me1
GSE136485	*M.musculus*	129 E14TG2a.4	Control ChIP-seq	input

表 3.3　RNA-seq 数据集

ExperimentSet	Organism	Biosource Type	Experiment Type	Assay Details
4DNESX6IF3FC	*M.musculus*	129 E14TG2a.4	RNA-seq	None

3.4.2　Hi-C 数据分析

将 Hi-C 文库测序得到的双端`.fastq`文件用 `HiC-Pro (v 2.8.1)` 进行比对。用 `cooler(v 0.8.6.post0)` 生成各分辨率矩阵并进行矩阵平衡。具体步骤有：

1　利用 `bowtie2 (v 2.2.5)` 将双端序列比对到 *Mus musculus*(mm10) 参考基因组上。

2　去掉含有 *Mbo*I 连接位点但未比对上的序列的末端部分序列，再次比对到 mm10 基因组上。

3　合并两次比对结果，去掉其中的 PCR 重复和光学重复序列并将剩下的序列

与 *Mbo*I 限制酶切片段对应起来。

4 将来自同一酶切片段的双端序列筛去，得到有效互作对。经过这四个步骤，能能到本研究所用的染色质互作数据。

合并各批生物学重复，再用这些序列对在 100kb, 50kb, 25kb 和 5kb 尺度上分别建立 Hi-C 互作矩阵。得到的各分辨率矩阵进一步进行迭代校正。随后将矩阵转换为`.hic`格式，以应用 `juicebox` 进行可视化。

采用 IS（绝缘值）方形分析确定 TADs 的位置。具体步骤如下：沿着 25kb 分辨率的矩阵的每条染色体对角线分别计算每隔 100kb, 1000kb 和 50kb 步长位置附近 125kb 内的互作平均值，并对位置求导得到 delta 向量。delta 向量为 0 且该处的差分为正的位置为局部最小值，即为可能的边界位置。其中边界强度大于 0.25 的边界，同时至少在 2/3 组数据集中出现边界被认为是可靠边界。

3.4.3　ChIP-seq 数据分析

用 bbduk 对去除下机.fastq 文件中的低质量数据（trimq=10 maq=10 minlen=20）。用 bowtie2(v 2.3.5.1) 将 ChIP-seq 序列比对到 mm10 参考基因组上。用 sambamba,samtools(v 1.7) 和 bedtools 对比对得到的文件进行排序、生成索引以及去除重复序列。

用 ChromHMM 的 BinarizeBam 将处理过后的.bam 以每 200bp 为一个 bin 进行二元化。对得到的二元信号通过 LearnModel 学习马尔可夫模型，推断出染色质的 12 种隐状态，并根据隐状态产生可观测的表观信号对状态进行生物学注释。

3.4.4　Hi-C, ChIP-seq 和 RNA-seq 数据的重复性

对 Hi-C 和 RNA-seq 的比对文件，用 bamCoverage 计算覆盖度并采用 RPKM 标准化。对 ChIP-seq 数据， 用 macs2 predictd 预测片段长度， 并在计算覆盖度时延长所有序列到该长度。对每种测序方法得到的覆盖度文件用 multiBigwigSummary bins 总结到一个文件中， 并用 plotCorrelation 可视化为相关性热图。

3.4.5　计算与 TAD 相关的变量

对于TAD 边界上的转录因子和组蛋白修饰信号，平均 TAD 的两个边界上的 ChIP-seq RPKM 值。对于边界上的转录水平，平均 TAD 两个边界上的 RNA-seq FPKM 值。边界强度由 TAD 内最大的 IS 值与边界的 IS 值的差得到。边界间的互作频率直接从互作矩阵中提取得到。计算互作矩阵的最大特征值，过程如下：

1　计算染色体内的 o/e 互作矩阵。在 500kb 分辨率下计算期望矩阵，也即平均在此分辨率下各个互作距离间的，随着距离线性增长的区域范围内的互作频率，则为该距离 d 间的期望互作频率。

2　计算 o/e 矩阵的 Pearson 相关矩阵，再对相关矩阵进行 PCA（主成分分析）得到最大特征值向量。

TAD 的大小由坐标之差得到。同一类别内的TAD 之间的平均互作频率在互作矩阵中提取相应数据得到。

下载ENCODE 项目中注册的cCREs（候选顺式作用元件）并用以注释TAD

边界。从 EBI 数据库中获取重复序列在基因组上的位置，计算它们与不同边界的交集数量分析边界上的重复序列情况。下载 AmiGO 2 中的 stem cell differentiation 的 GO 生物学过程项中的基因和 stem cell division 的 GO 生物学过程项中的基因并注释其在不同类别 TAD 中的分布。

3.4.6　聚类分析

将输入矩阵正则化后采用聚集聚类，递归地合并增加类间联系程度距离最少的类别，最终选择最大化 silhouette 系数[①]的类别数量得到聚类结果。其中 a 为类别内的平均距离，b 为样本和不包含该样本的最近类别中样本的平均距离。对于边界上的染色质状态矩阵，采用 Otsuka-Ochiai 系数[②]作为相似度度量进行聚集聚类。

3.4.7　评估两组 TAD 集合的一致性

应用 Jaccard 指数比较两组 TAD 集合

$$\text{Jaccard}(A_i, B_i) = \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \tag{3.1}$$

。且 A_i 对 B 划分的 Jaccard 指数定义为 B 中与最优匹配的 Jaccard 指数：

$$\text{Jaccard}(A_i, B) = \max_{j \leq l} \text{Jaccard}(A_i, B_j) \tag{3.2}$$

①样本的 Silhouette 系数定义为 $S=\frac{b-a}{\max\{a,b\}}$

②样本的 Otsuka-Ochiai 系数定义为 $K=\frac{|A\cap B|}{\sqrt{|A|\times|B|}}$

参 考 文 献

[1] GREEN D M, KAWASHIMA T, STOVALL M, et al. Fertility of male survivors of childhood cancer: a report from the childhood cancer survivor study[J]. Journal of Clinical Oncology, 2010, 28(2): 332.

[2] VALLI H, PHILLIPS B T, SHETTY G, et al. Germline stem cells: toward the regeneration of spermatogenesis[J]. Fertility and sterility, 2014, 101(1): 3-13.

[3] RUBINO P, VIGANÒ P, LUDDI A, et al. The icsi procedure from past to future: a systematic review of the more controversial aspects[J]. Human reproduction update, 2016, 22(2): 194-227.

[4] FAYOMI A P, PETERS K, SUKHWANI M, et al. Autologous grafting of cryopreserved prepubertal rhesus testis produces sperm and offspring[J]. Science, 2019, 363(6433): 1314-1319.

[5] DE ROOIJ D G. The nature and dynamics of spermatogonial stem cells[J]. Development, 2017, 144(17): 3022-3030.

[6] GRISWOLD M D. Spermatogenesis: the commitment to meiosis[J]. Physiological reviews, 2016, 96(1): 1-17.

[7] OATLEY J M, BRINSTER R L. Regulation of spermatogonial stem cell self-renewal in mammals[J]. Annual review of cell and developmental biology, 2008, 24: 263-286.

[8] YAMAUCHI Y, RIEL J M, RUTHIG V A, et al. Two genes substitute for the mouse y chromosome for spermatogenesis and reproduction[J]. Science, 2016, 351(6272): 514-516.

[9] ISOTANI A, NAKANISHI T, KOBAYASHI S, et al. Genomic imprinting of xx spermatogonia and xx oocytes recovered from xx↔ xy chimeric testes[J]. Proceedings of the National Academy of Sciences, 2005, 102(11): 4039-4044.

[10] ZICKLER D, KLECKNER N. Recombination, pairing, and synapsis of homologs during meiosis[J]. Cold Spring Harbor perspectives in biology, 2015, 7(6): a016626.

[11] POWERS N R, PARVANOV E D, BAKER C L, et al. The meiotic recombination activator prdm9 trimethylates both h3k36 and h3k4 at recombination hotspots in vivo[J]. PLoS genetics, 2016, 12(6): e1006146.

[12] BRICK K, SMAGULOVA F, KHIL P, et al. Genetic recombination is directed away from functional genomic elements in mice[J]. Nature, 2012, 485(7400): 642-645.

[13] HAYASHI K, YOSHIDA K, MATSUI Y. A histone h3 methyltransferase controls epigenetic events required for meiotic prophase[J]. Nature, 2005, 438(7066): 374-378.

[14] IRIE S, TSUJIMURA A, MIYAGAWA Y, et al. Single-nucleotide polymorphisms of the prdm9 (meisetz) gene in patients with nonobstructive azoospermia[J]. Journal of andrology, 2009, 30

(4): 426-431.

[15] BAUDAT F, MANOVA K, YUEN J P, et al. Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking spo11[J]. Molecular cell, 2000, 6(5): 989-998.

[16] PITTMAN D L, COBB J, SCHIMENTI K J, et al. Meiotic prophase arrest with failure of chromosome synapsis in mice deficient for dmc1, a germline-specific reca homolog[J]. Molecular cell, 1998, 1(5): 697-705.

[17] SOUQUET B, ABBY E, HERVÉ R, et al. Meiob targets single-strand dna and is necessary for meiotic recombination[J]. PLoS Genet, 2013, 9(9): e1003784.

[18] JONES G. The control of chiasma distribution.[C]//Symposia of the Society for Experimental Biology: volume 38. 1984: 293-320.

[19] HASSOLD T, HALL H, HUNT P. The origin of human aneuploidy: where we have been, where we are going[J]. Human molecular genetics, 2007, 16(R2): R203-R208.

[20] BAO J, BEDFORD M T. Epigenetic regulation of the histone-to-protamine transition during spermiogenesis[J]. Reproduction (Cambridge, England), 2016, 151(5): R55.

[21] ITOH K, KONDOH G, MIYACHI H, et al. Dephosphorylation of protamine 2 at serine 56 is crucial for murine sperm maturation in vivo[J/OL]. Science Signaling, 2019, 12(574). https://stke.sciencemag.org/content/12/574/eaao7232. DOI: 10.1126/scisignal.aao7232.

[22] GOODIER J L, KAZAZIAN JR H H. Retrotransposons revisited: the restraint and rehabilitation of parasites[J]. Cell, 2008, 135(1): 23-35.

[23] ARAVIN A A, SACHIDANANDAM R, BOURC'HIS D, et al. A pirna pathway primed by individual transposons is linked to de novo dna methylation in mice[J]. Molecular cell, 2008, 31(6): 785-799.

[24] MANAKOV S A, PEZIC D, MARINOV G K, et al. Miwi2 and mili have differential effects on pirna biogenesis and dna methylation[J]. Cell reports, 2015, 12(8): 1234-1243.

[25] ZAMUDIO N, BOURC'HIS D. Transposable elements in the mammalian germline: a comfortable niche or a deadly trap?[J]. Heredity, 2010, 105(1): 92-104.

[26] HUG C B, VAQUERIZAS J M. The birth of the 3d genome during early embryonic development[J]. Trends in Genetics, 2018, 34(12): 903-914.

[27] CREMER T, CREMER M. Chromosome territories[J]. Cold Spring Harbor perspectives in biology, 2010, 2(3): a003889.

[28] HUGHES J R, ROBERTS N, MCGOWAN S, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment[J]. Nature genetics, 2014, 46(2): 205.

[29] RAO S S, HUNTLEY M H, DURAND N C, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping[J]. Cell, 2014, 159(7): 1665-1680.

[30] DIXON J R, SELVARAJ S, YUE F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions[J]. Nature, 2012, 485(7398): 376-380.

[31] VIETRI RUDAN M, BARRINGTON C, HENDERSON S, et al. Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. cell rep. 10: 1297–1309 [Z]. 2015.

[32] FUDENBERG G, IMAKAEV M, LU C, et al. Formation of chromosomal domains by loop extrusion[J]. Cell reports, 2016, 15(9): 2038-2049.

[33] LUGER K, MÄDER A W, RICHMOND R K, et al. Crystal structure of the nucleosome core particle at 2.8 å resolution[J]. Nature, 1997, 389(6648): 251-260.

[34] PHILLIPS-CREMINS J E, SAURIA M E, SANYAL A, et al. Architectural protein subclasses shape 3d organization of genomes during lineage commitment[J]. Cell, 2013, 153(6): 1281-1295.

[35] JI X, DADON D B, POWELL B E, et al. 3d chromosome regulatory landscape of human pluripotent cells[J]. Cell stem cell, 2016, 18(2): 262-275.

[36] SCHMITT A D, HU M, JUNG I, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome[J]. Cell reports, 2016, 17(8): 2042-2059.

[37] MEISTRICH M L, MOHAPATRA B, SHIRLEY C R, et al. Roles of transition nuclear proteins in spermiogenesis[J]. Chromosoma, 2003, 111(8): 483-488.

[38] DU Z, ZHENG H, HUANG B, et al. Allelic reprogramming of 3d chromatin architecture during early mammalian development[J]. Nature, 2017, 547(7662): 232-235.

[39] KE Y, XU Y, CHEN X, et al. 3d chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis[J]. Cell, 2017, 170(2): 367-381.

[40] JUNG Y H, SAURIA M E, LYU X, et al. Chromatin states in mouse sperm correlate with embryonic and adult regulatory landscapes[J]. Cell reports, 2017, 18(6): 1366-1382.

[41] TANG W W, KOBAYASHI T, IRIE N, et al. Specification and epigenetic programming of the human germ line[J]. Nature Reviews Genetics, 2016, 17(10): 585.

[42] TURNER J M. Meiotic sex chromosome inactivation[J]. Development, 2007, 134(10): 1823-1831.

[43] WANG Y, WANG H, ZHANG Y, et al. Reprogramming of meiotic chromatin architecture during spermatogenesis[J]. Molecular cell, 2019, 73(3): 547-561.

[44] ALAVATTAM K G, MAEZAWA S, SAKASHITA A, et al. Attenuated chromatin compartmentalization in meiosis and its maturation in sperm development[J]. Nature structural & molecular biology, 2019, 26(3): 175-184.

[45] PATEL L, KANG R, ROSENBERG S C, et al. Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase[J]. Nature structural & molecular biology,

2019, 26(3): 164-174.

[46] HERNÁNDEZ-HERNÁNDEZ A, LILIENTHAL I, FUKUDA N, et al. Ctcf contributes in a critical way to spermatogenesis and male fertility[J]. Scientific reports, 2016, 6(1): 1-13.

[47] FUDENBERG G, ABDENNUR N, IMAKAEV M, et al. Emerging evidence of chromosome folding by loop extrusion[C]//Cold Spring Harbor symposia on quantitative biology: volume 82. Cold Spring Harbor Laboratory Press, 2017: 45-55.

[48] ORKIN S H, HOCHEDLINGER K. Chromatin connections to pluripotency and cellular reprogramming[J]. cell, 2011, 145(6): 835-850.

[49] LOUBIERE V, MARTINEZ A M, CAVALLI G. Cell fate and developmental regulation dynamics by polycomb proteins and 3d genome architecture[J]. BioEssays, 2019, 41(3): 1800222.

[50] ALOIA L, DI STEFANO B, DI CROCE L. Polycomb complexes in stem cells and embryonic development[J]. Development, 2013, 140(12): 2525-2534.

[51] OGIYAMA Y, SCHUETTENGRUBER B, PAPADOPOULOS G L, et al. Polycomb-dependent chromatin looping contributes to gene silencing during drosophila development[J]. Molecular cell, 2018, 71(1): 73-88.

[52] LOUBIERE V, PAPADOPOULOS G, SZABO Q, et al. Widespread activation of developmental gene expression characterized by prc1-dependent chromatin looping[J]. Science advances, 2020, 6(2): eaax4001.

[53] RIISING E M, COMET I, LEBLANC B, et al. Gene silencing triggers polycomb repressive complex 2 recruitment to cpg islands genome wide[J]. Molecular cell, 2014, 55(3): 347-360.

[54] SHAN Y, LIANG Z, XING Q, et al. Prc2 specifies ectoderm lineages and maintains pluripotency in primed but not naïve escs[J]. Nature communications, 2017, 8(1): 1-14.

[55] ENDOH M, ENDO T A, ENDOH T, et al. Polycomb group proteins ring1a/b are functionally linked to the core transcriptional regulatory circuitry to maintain es cell identity[J]. Development, 2008, 135(8): 1513-1524.

[56] ZEPEDA-MARTINEZ J A, PRIBITZER C, WANG J, et al. Parallel prc2/cprc1 and vprc1 pathways silence lineage-specific genes and maintain self-renewal in mouse embryonic stem cells[J/OL]. Science Advances, 2020, 6(14). https://advances.sciencemag.org/content/6/14/e

aax5692. DOI: 10.1126/sciadv.aax5692.

[57] ZHAN Y, MARIANI L, BAROZZI I, et al. Reciprocal insulation analysis of hi-c data shows that tads represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes[J]. Genome research, 2017, 27(3): 479-490.

[58] MARCHAL C, SIMA J, GILBERT D M. Control of dna replication timing in the 3d genome [J]. Nature Reviews Molecular Cell Biology, 2019, 20(12): 721-737.

[59] BONEV B, COHEN N M, SZABO Q, et al. Multiscale 3d genome rewiring during mouse neural development[J]. Cell, 2017, 171(3): 557-572.

[60] SYMMONS O, USLU V V, TSUJIMURA T, et al. Functional and topological characteristics of mammalian regulatory domains[J]. Genome research, 2014, 24(3): 390-400.

[61] WEISCHENFELDT J, DUBASH T, DRAINAS A P, et al. Pan-cancer analysis of somatic copy-number alterations implicates irs4 and igf2 in enhancer hijacking[J]. Nature genetics, 2017, 49(1): 65-74.

[62] TANG Z, LUO O J, LI X, et al. Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription[J]. Cell, 2015, 163(7): 1611-1627.

[63] GOLOBORODKO A, MARKO J F, MIRNY L A. Chromosome compaction by active loop extrusion[J]. Biophysical journal, 2016, 110(10): 2162-2168.

[64] GANJI M, SHALTIEL I A, BISHT S, et al. Real-time imaging of dna loop extrusion by condensin[J]. Science, 2018, 360(6384): 102-105.

[65] ROWLEY M J, NICHOLS M H, LYU X, et al. Evolutionarily conserved principles predict 3d chromatin organization[J]. Molecular cell, 2017, 67(5): 837-852.

[66] SCHACHT T, OSWALD M, EILS R, et al. Estimating the activity of transcription factors by the effect on their target genes[J]. Bioinformatics, 2014, 30(17): i401-i407.

[67] ARRIETA-ORTIZ M L, HAFEMEISTER C, BATE A R, et al. An experimentally supported model of the bacillus subtilis global transcriptional regulatory network[J]. Molecular systems biology, 2015, 11(11): 839.

[68] ZHANG K, WANG M, ZHAO Y, et al. Taiji: System-level identification of key transcription factors reveals transcriptional waves in mouse embryonic development[J/OL]. Science Advances, 2019, 5(3). https://advances.sciencemag.org/content/5/3/eaav3262. DOI: 10.1126/sciadv.aav3262.

[69] LUO Z, WANG X, JIANG H, et al. Reorganized 3d genome structures support transcriptional regulation in mouse spermatogenesis[J]. Iscience, 2020, 23(4): 101034.

附录 A　补 充 材 料

A.1　补充图片

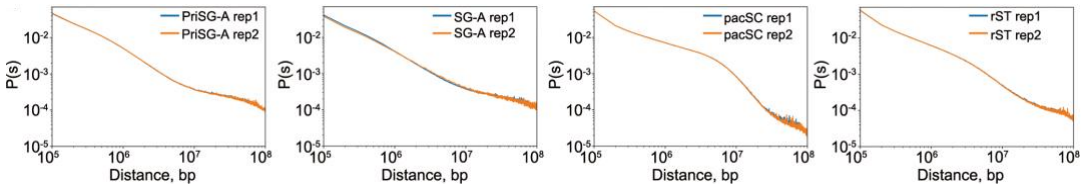


图 A.1　Hi-C 数据的重复度

对小鼠精子发生各阶段的技术重复样本的 P(s) 分析。

致　谢

我在 18 年 2 月定下读研的事离家的时候，写了一首诗——鹏鸟翱翔兮喜忧，晚舟不系兮浮休，问道泰山以北海，趋意折枝兮沙洲。那时候的我出于缺乏信息和多余的道德感和自卑，对于读研其实是悲观的，但是为了提振精神，写了这样四不像的诗——既立志、又退缩。在这之后的三年间，我并未向预期地那样全情投入科研和学习。在这三年间我所取得的大部分成果，都来自于整个课题组给予的机会与信任。在此，我必须向导师和所有同学表示最诚挚的感谢。

首先要感谢宋晓元老师。宋老师自身一路走来的每个角色都做得很好，是我的榜样。宋老师尽己所能地自我学习、利用自己所长提出并争取重要的生物学课题，建设了我们的实验室。宋老师为人温和，温柔宽宏的老师能为学生提供舒适的成长和进步的环境。在具体的课题和研究工作产出方面，宋老师积极主动为学生着想，为学生提供各方面的帮助。能做宋老师的学生我真的感到非常地幸运。

接下来感谢的 813 办公室的全体成员。813 办公室是我这三年办公的场所，在其中我有时过于不分彼此，可能有时又过于冷漠，在我对自己没有严格要求的时候，可能也给一个环境内工作的同事带来了坏心情。但大家对我仍然给予着关心和帮助。在来到科大的前两年里，我和 813 小课题组在工作上关系也较为密切。通过参与团队的课题和小组会，我学到了很多实现高效团队协作的小技巧。特别的，我要感谢张远伟师兄给予的机会和指点。张师兄虽然自己能力很强，科研成果丰硕，同时又平易近人，对待我这个异门师妹也很耐心。在 813 期间，我和宋老师课题组的设备等出现问题，也总是张师兄帮忙解决的。我要感谢高佳宁师兄。我能进入宋老师课题组，也是靠高师兄的推荐。一开始，是高师兄带我入门了基因组学，并教给我一些生物学信息学分析的脚本和工具。高师兄非常聪明，在科研上很活跃，遗憾我不总是能跟上他的工作节奏。我还要感谢周建腾师兄和赵达仁师弟在生活上给予的帮助。我有任何困扰和烦恼，与他们沟通，总会豁然开朗。

接下感谢们实验室的博后王斐师姐。在我的每次工作汇报之后，师姐总是提出许多建设性的意见，使我受益匪浅。和我同一届的朱志强、陈佩泽、张可和简仕坤同学，都非常优秀，在我面对各种困难的时候也愿意伸出援手，在我需要的时候也总是乐于支持我，包括作为我的入党介绍人等等。

感谢吴玥明师弟、何梦真师妹。他们对待学习和科研非常主动积极和认真，也热心地对我提供帮助，他们看到有利于我的信息时总是想起我，并及时提醒我的

工作进度。我们曾经一起组织过生信小组会，也在一直以来的分小组的科研活动中合作和配合得很好。

同时要特别感谢中国科学技术大学。中科大不愧是名校，在学习环境上足够先进，教室、图书馆、实验室的资源都特别好，使得学生在其中能自觉投入严肃的学习和科研工作。中科大的校园环境也美丽宜人又亲切。熟悉了中科大之后， 我觉得自己很适合这里的校园环境，也有些不舍。在这个学校里，我遇到了许多亲切的课任老师，选修了各个学院的、课程安排很好、内容和形式都非常详实的课程。我也结识了不同学院的、性格各异但都很美好的同学。能在中科大就读， 我感到十分幸运。

我还要感谢理科讨论群内的众位群友。作为经历过或正在经历读研读博的年轻学生，大家在读研读博的历程、具体每一步的工作上，几乎做到了我有问必有达。在整个学术道路上能有志同道合的人同行，也让我感到十分幸福。

最后，要特别感谢的父母和哥哥在生活上和学习上的帮助和支持。我的父母全力支持我的人生，我的哥哥一直关心我所走的每一步。我会继续认真面对我的生活，更好地回报家人和朋友的支持。

在读期间发表的学术论文与取得的研究成果

已发表论文

1. Luo, Z., Wang, X., Jiang, H., Wang, R., Chen, J., Chen, Y., ... & Yang, Y.(2020). Reorganized 3D genome structures support transcriptional regulation in mouse spermatogenesis. iScience, 101034. **(Co-first author)**

2. Luo, Z., Hu, T., Jiang, H., Wang, R., Xu, Q., Zhang, S., ... & Song, X. (2020). Rearrangement of macronucleus chromosomes correspond to TAD-like structures of micronucleus chromosomes in Tetrahymena thermophila. Genome Research, 30(3), 406-414. **(Second author)**

3. Cao, J., Jiang, H., Zhang, R., Ghanam, A. R., Xia, H., Zhu, Y., ... & Song, X. (2019). Epigenomic profiling identifies the role of Nr5a2 and CYP1B1 in hepatocellular carcinoma stemness maintenance. Science Bulletin, 64(16), 1132-1135. **(Second author)**

4. Azhar Muhammad, Ramay Waheed, Nauman Ali Khan, Hong Jiang, Xiaoyuan Song, piRDisease v1.0: a manually curated database for piRNA associated diseases, Database, Volume 2019, 2019, baz052.

待发表论文

1. The Genome is Partitioned into Contact Domains that Segregate into Seven Types with Distinct Expression and Interaction Patterns. **(First author)**