基于 Hi-C 的单细胞层次聚类与网络可视化及差异特征提取

——项目 07“染色体三维结构分析及其在疾病和基因组进化中的作用研究”

团队名称：阿尔斯丢丢实战队

指导老师：李立

团队成员：邵燕涵，周星宇，魏莫迟

贺芷涵，赵晓涵，宁晓彤

摘要

生物发育的过程受许多基因的调控，而基因和调控元件之间由于线性距离无法准确描述基因对个体生长发育的影响，同时部分调控元件在二维空间结构上的分布不能准确地反映基因之间真实存在的调控机制与结果，故我们通过分析 Hi-C 测序技术所得的单细胞数据，对其在三维空间结构中存在的染色质间交互以及染色质内交互的特征进行分析，来解释不同细胞中各基因间存在的相互作用与相互关联。本项目中，我们通过利用 PCA 及 t-SNE 等降维方法对小鼠胚胎干细胞的 Hi-C 数据进行降维处理，发现 PCA 在维度分析上面比破坏维度特征的 t-SNE 更加精确、有效。我们通过质检以及归一化处理进一步提高数据的可信度，并在染色质、Compartment 及 TADs 三个层面进行层次聚类。通过多种可视化以及网络图展示，我们实现了系统地展示小鼠胚胎干细胞以及其他细胞系在染色质、Compartment 及 TADs 三个层面所展现出来的差异，且在不同层面都能看到明显的差异存在。通过后续的二期实验，我们可以更好地从可视化结果中提取出基因在三维空间结构下存在的特征属性以及特征模式，进而去更好地探究在不同层面上，不同细胞内存在的调控机理。

关键字：单细胞 Hi-C，降维，层次聚类，差异特征提取

1 引言

研究发现，基因组序列、基因结构以及调控元件等所携带的遗传信息，在染色体上并不是以线性的方式表现的，且在二维空间结构上，部分调控元件的分布不能准确地反映基因之间真实存在的调控机制与结果。为了进一步探究基因组中特定调控元件之间的互作机制，对基因组三维结构的研究就成为了基因组学中一个新的发展方向。

三维基因组学（Three-Dimensional Genomics, 3D Genomics），研究基因组三维空间结构与功能。研究对象包括基因组序列、基因结构、调控元件及在生物发育过程中基因与基因之间的三维互作（基因转录、复制、调控功能等）变化。

提到三维基因组学，就必须要提到 Hi-C 技术与 CHIA-PET 技术，它们都是将空间结构上距离近的 DNA 片段交联，并将交联片段富集再高通量测序的技术，对这种测序数据分析就可以揭示染色质之间的远程互作，而不考虑线性距离远近。Hi-C 测序技术（High-through chromosome conformation capture，高通量染色质构象捕获技术）是源于染色质构象捕获技术，

以整个细胞核作为研究对象，研究全基因组范围内整个染色质 DNA 在空间位置上的关系的一种技术。而 CHIA-PET（Chromatin Interaction Analysis using Paired End Tag sequencing，双末端标签染色质免疫共沉淀测序技术）则是研究高通量测序基因组范围内染色质远程交互的技术。

单细胞测序技术（Single-cell sequencing），单细胞生物学中较为常见的一种测序技术，用来精细区分不同类型的细胞序列。由于普通测序结果会受到细胞异质性影响，而单细胞测序可以缩小此类误差，故而单细胞测序能够让我们更准确地认识细胞之间的差异，帮助我们了解细胞异质性对发育过程的影响，以及更加细化定义细胞类型等。此外，由于我们的研究对象是单细胞聚类，故而传统 Hi-C 数据不适用于本研究。

基因组的三维结构大致可以分为四种层面，从大到小分别是：染色质（Loose chromosome structure）、compartment（Chromatin compartments）、TADs（Topologically associating domains）和 loops（Chromatin loops）。大部分情况下，染色质是遗传物质的主要形态，结构松散，内部交联多于染色质间交联；compartment 是指染色质丝折叠形成的不同的亚区室，与转录激活（A 区室）或抑制（B 区室）相关，而 A 与 B 的分类由 Hi-C 的 contact 频率定义；TADs 是指折叠成具有高内部互作频率的自缔合染色质区域，边界处与邻区的局部互作明显下降，从而据此划定了界限；loops 是两个基因座之间互作频率较高的局部区域，这些区域构成了 DNA 环的基础，通常在不同结合因子(CTCF)位点的区域之间，或增强子与其靶启动子间形成。这四种层面的划分尺度有着明显的跨度，并揭示了不同的生物学意义。单细胞 Hi-C 技术的发展给予了我们从单细胞层面探索基因组三维结构的能力，与此同时，我们也可以试图从基因组三维结构上找出不同细胞系之间的差异，并将这种差异提取出来，而对于提取出来的差异，如果经过验证，其的确具有一定的模式特征或是统计学显著性，那么我们就可以将这种差异识别运用于对未知细胞的识别，或者对海量细胞系数据进行聚类，从而实现一种基于 Hi-C 的单细胞判别方法。

本项目中，我们使用 PCA 和 t-SNE 的方法对基因组交互数据进行了降维和聚类，希望能从染色质间和染色体内交互频率入手，实现层次聚类和差异特征提取。

2 方法与材料

2.1 数据来源

本次项目是针对单细胞层次聚类和差异特征提取所开展的，因此我们所选择的数据都来自于单细胞 Hi-C 的实验数据。

我们的研究数据一共有两个来源，第一套数据来源于对人类及小鼠卵母细胞受精过程进行研究的基因组单细胞 Hi-C 数据，共 227 个样本[1]。第二套数据来源于小鼠胚胎干细胞 E14 细胞系和 NMuMG 细胞系的单细胞 Hi-C 数据，共 447 个样本[2]。

在对数据进行 Compartment 和 TAD 的划分时，由于单细胞的数据数据量较小且差别波动较大，必须使用基准文件才能对其进行规范划分，我们需要一个较大的细胞系整体 Hi-C 测序数据进行 Compartment 和 TAD 的划分并得到划分的基准。本次使用的基准文件来自于网站 4D Nucleome Data Portal，该网站提供基因组数据的检索、下载和可视化等服务，其中就包括各种类型的实验流程处理得到的结果文件。我们的基准文件由于数据量非常庞大，无法在本校服务器上进行流程处理，因此使用了该网站提供的—使用小鼠胚胎干细胞细胞系的 in situ Hi-C 实验数据进行流程处理之后得到的.mcool 格式的多分辨率染色质交互信息文件，文件大小为 5.9G，作为基准文件。

2.2 前期准备

2.2.1 数据获取与整理

数据来源文献的作者在 NCBI 的 GEO 数据库中提供了单细胞样本经过 Hi-C 分析流程处理之后得到的染色质交互文件，而多细胞 Hi-C 数据的染色质交互文件能够直接从 4D Nucleome Data Portal 网站上免费下载，直接使用这些文件进行后续的分析。在得到交互文件之后，两套单细胞数据集都是染色质交互矩阵的形式，但是其矩阵列名和意义都并不统一，于是我们将交互矩阵统一成为染色体名称加交互位点的格式。对于基准文件，我们使用工具 cooltools 对多细胞 Hi-C 交互矩阵进行划分 Compartment 和 TAD 等操作。

2.2.2 数据的质控

数据的质量会对实验造成不可避免的系统误差，在获取数据后，第一步是对数据进行质控，在这次项目中我们的单细胞交互矩阵进行了三个方面的质控。

第一步和第二步较为常见，是对数据的深度和广度进行质控，这一步有点类似于测序数据的预处理。我们所处理的 Hi-C 数据中的深度和广度则主要是指的交互矩阵在真实矩阵上面的深度和广度。因为 Hi-C 数据的可视化需要借助热图矩阵，也就是我们需要将交互矩阵填充到热图矩阵，这样才能直观显示染色质交互，而热图矩阵上面的每一个点都有其存在的意义。

于是，我们需要针对热图矩阵检测每一个点的深度以及该热图的覆盖度进行统计分析，并去除显著的离群点，保证剩下的交互矩阵其深度和广度都在可信范围之内。

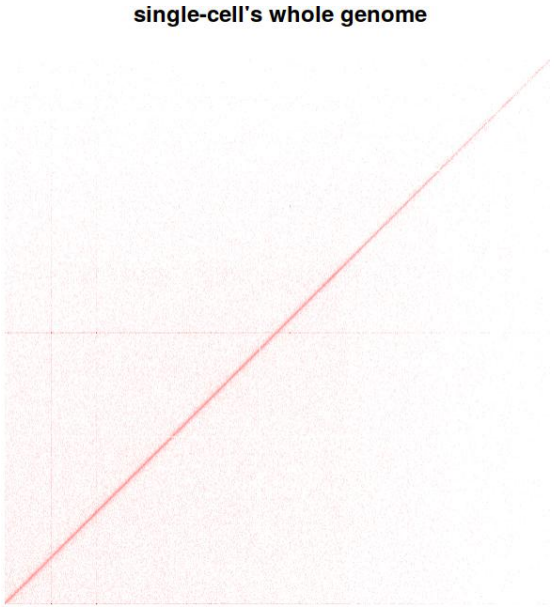

图 1 单细胞基因组热图。该图显示了单细胞数据染色质内交互明显多于染色质间交互

第三步是针对于交互矩阵文件的质控，第一步和第二步控制交互矩阵的深度和广度，但由于 Hi-C 数据分析的特殊性，具有同样的深度与广度，其代表的染色质间交互和染色质内交互却可能大不相同。所以，我们需要一个可信的“比例”，这个“比例”指的是每一个样本文件所记录的交互信息应该具有可信度。在对所有样本进行统计分析之后，我们发现有的文件其染色质间交互远远大于染色质内交互，但是在正常情况下染色质内交互应该是大于染色质间交互的（下面的 Hi-C 热图结果可以说明），这种现象可能就表明：在进行 Hi-C 测序的时候并不是纯的单细胞测序，其中可能混杂了其他细胞，因此这一类数据也应该剔除。

在进行上述三步质控之后，第一套数据原有 227 个样本，质控之后剔除 79 个样本，剩余 148 个样本，第二套数据原有 447 个样本，质控之后剔除 89 个样本，剩余 358 个样本。

2.3 层次聚类及差异特征提取实验流程

2.3.1 单细胞 Hi-C 数据不同层次的分 bin

前面我们讲到，Hi-C 的热图可以从四个层面来对基因组三维结构进行解释，包括染色质、Compartment、TAD 和 loop，由于 loop 的位点较小，且其主要对应于模体的识别，我们本次实验就只针对于染色质、Compartment 和 TAD 这三个较大跨度的层面。

染色质是最大的层面，我们对每一个交互位点所对应的两端序列进行了取中值的操作，并用这一对中值代表这一个交互位点的信息。在 Compartment 和 TAD 层面，我们按照 Compartment 和 TAD 的基准文件对每个样本进行分 bin 操作，而在 Compartment 层面，由于在基因组三维结构中，A/B Compartment 在整条染色质上是间隔分布的，所以我们在这里将 Compartment 再细分为合并（即每条染色质合并为 A 与 B 两个 bin）与非合并（即每条染色质合并为 A 与 B 间隔分布的 bin）两类以验证哪一种聚类效果更好，值得注意的是，由于 TAD 是 Compartment 的子结构，因此我们推断非合并的 Compartment 可能与 TAD 结果具有一定的相似性。（注：本次实验均在 100kb 分辨率下进行）

2.3.2 对不同层次分 bin 之后的染色质交互信息进行计数得到计数矩阵

在分 bin 之后，我们对每一个层次上的样本数据进行了染色质内交互信息的计数，这里忽略了染色质间交互，因为单细胞 Hi-C 的数据本身就非常稀疏，大部分数据都是集中在热图矩阵的主对角线上（以左上角为矩阵的零起始点），因此我们只对染色质间交互进行了计数。在计数之后所得到的计数矩阵，其每一行代表一个样本，每一列代表这个层次上的一个 bin，因此一套数据我们可以得到 4 个计数矩阵：染色质层面共 21 列，每一列为一条染色体，合并 Compartment 层面共 42 列，每两列为一条染色体的 A 与 B Compartment，非合并 Compartment 与 TAD 层面的每一列则代表一个单独的 Compartment 和 TAD。

2.3.3 样本归一化与 PCA 降维处理

在得到计数矩阵之后，我们就可以对矩阵进行降维并剔除那些不显著的维度，这样我们就能找到样本之间显著不相同的维度，这里指的维度就是计数矩阵里面的每一列，之前提到每一个层面的每一列都具有不同的意义，所以我们就能从四个层面去解释样本的生物学意义。

而在降维之前，我们还需要进行归一化的处理，这里指的归一化是针对于文件大小的归一化，之前质控之后，我们对样本的质量有了良好的控制，但是在降维的时候，每一行的样本值对降维都有一定的贡献，而 PCA（主成分分析）[3]的算法无法顾及样本本身文件大小所造成的影响，算法本身是基于维度上的纵向数据，并将其投影到不同维度，而不是基于横向的样本数据，因此我们就需要将计数矩阵里面的每一行都除以该样本文件的大小，从而达到样本的统一，接下来我们利用 PCA 对维度分析的强大处理能力，将计数矩阵进行 PCA 降维处理了。

2.3.4 t-SNE

t-SNE 是一种目前非常流行的降维聚类的算法，该算法由 SNE 算法改进而来，其降维原理与 PCA 完全不同[4]。PCA 是将数据投影到不同维度的平面上实现降维，而 t-SNE 则是根据分布概率来对样本进行归类。从聚类上来讲，t-SNE 比 PCA 更加有效且精确，但是由于 t-SNE 的算法会破坏数据的维度特征，因此我们无法从不同维度上有效地分析数据，也无

法得知不同维度上的贡献度大小。在该项目中，我们所要研究的每一个维度都有一定的意义，t-SNE 的算法只能得到固定的二维结果，而其他维度上均会被破坏，因此我们认为在此次实验当中 PCA 可能更加适合。

2.3.5 层次聚类

在经过 PCA 和 t-SNE 的降维之后，我们总能得到一个二维的可视化视图。对于 PCA 来讲，我们可以任意选择两个不同维度的主成分，将所有样本在所选二维平面上展现出来。但对于 t-SNE 来说，维度是不可选的，即使是我们所看到的二维分布，也是破坏原始数据的维度之后的结果，因此对于 t-SNE 我们只是进行了简单的绘图。

在使用 PCA 结果绘制二维平面分布图的时候，我们需要准确地确认哪些维度的贡献度更高，即哪些维度具有代表性，使得样本可以在该维度（或维度集）的投影面上显著地分开，这样该维度（或维度集）就更加适合用于可视化。本次实验的全部算法均在 R 语言中完成，其中，PCA 算法、分析贡献度、聚类可视化使用的是 prcome 函数、corrplot 函数和 factoextra 包的 fviz 系列函数，t-SNE 算法使用的是 tsne 函数。

2.3.6 差异特征提取

在得到 PCA 的结果之后，我们所需要了解的，即我们所迫切想知道的，是如果在某一维度（或维度集）上样本显示出显著的聚类效果，是何种因素导致了这种聚类，是哪些变量导致了这类分布的效果。进一步我们需要找到对该维度（或维度集）贡献较大的变量，这些变量就是每一个层次对应的 bin。通过提取出这些变量，我们能够实现在不同层次上，去验证那些“PCA 认为差异较大”的变量是否存在模式特征或是某种特殊的区分点，并通过可视化以及回溯原数据解释它存在的生物学意义。于是，我们使用了较为传统的热图方式来验证和探究。

3 结果展示

3.1 层次聚类及差异特征提取展示

3.1.1 基于 PCA 与 t-SNE 的层次聚类

我们共使用了两组数据，分别从染色质层面、合并 Compartment 层面以及非合并 compartment 和 TAD 层面进行对比展示。

第一组数据选择了来自小鼠受精卵与卵母细胞的 single-cell 数据。在染色质层面（图 2），我们通过水滴图，可以发现 PC1 对所有变量的代表性最强，即所有变量对 PC1 贡献程度最大。而通过聚类结果可以明显发现，PCA 和 t-SNE 算法都无法将两类细胞显著分开。在合并 Compartment 层面（图 3）和非合并 Compartment 和 TAD 层面（图 4），聚类结果类似。综合来看，第一套数据无法通过两种降维算法将不同种类的细胞完全分开。考虑到可能是因为从染色质交互计数上来讲，受精卵和卵母细胞的相似性较高，即使降维之后其数据也无法正常分开，所以我们没有继续对第一套数据进行后续分析，而是着重对第二套数据进行分析。

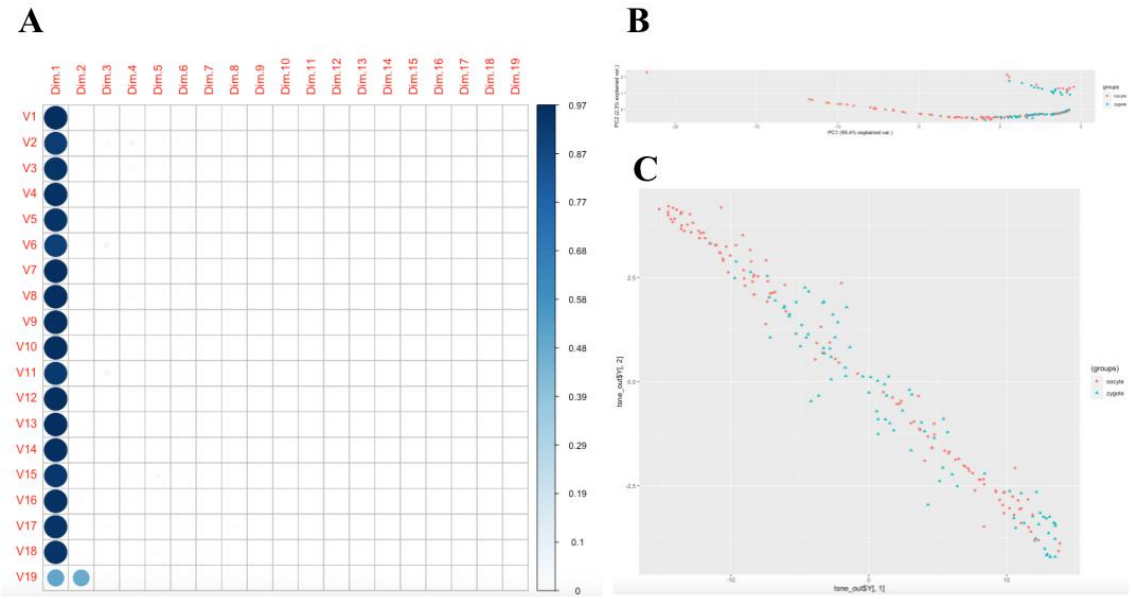

图 2 第一组数据染色质层面聚类结果。其中图 A 为主成分对变量的代表性的水滴图，行代表变量，列代表不同 PC。图 B 为二维样本图，取 PC1 与 PC2 作图，红色为卵母细胞，蓝色为受精卵细胞。图 C 为 t-SNE 降维聚类图，红色为卵母细胞，蓝色为受精卵细胞。

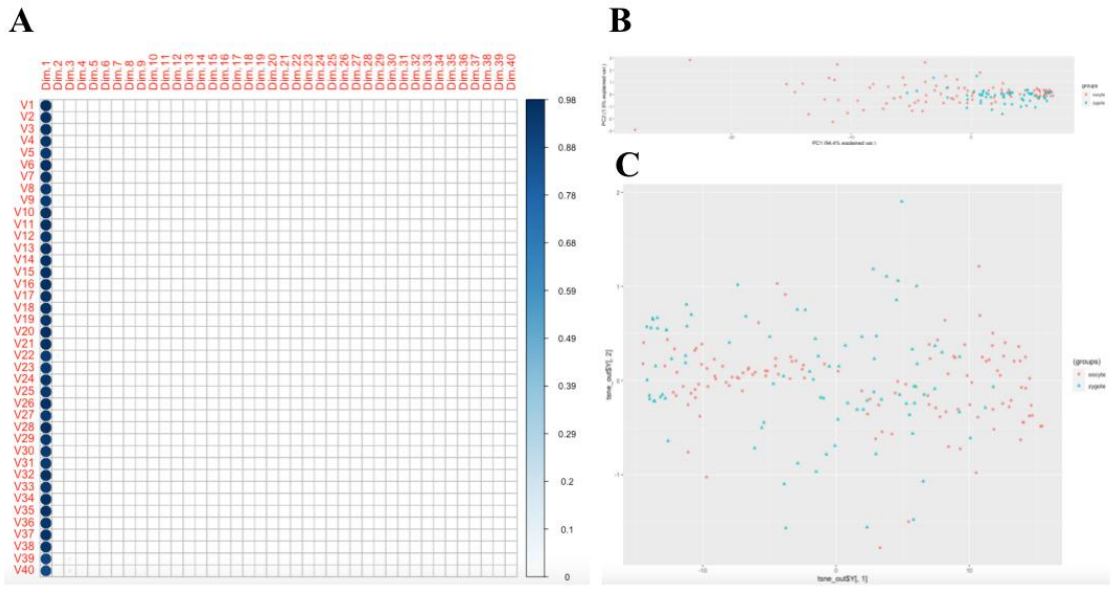

图 3 第一组数据合并 Compartment 层面聚类结果。其中图 A 为主成分对变量的代表性的水滴图。图 B 为取 PC1 与 PC2 的二维样本图，。图 C 为 t-SNE 降维聚类图。

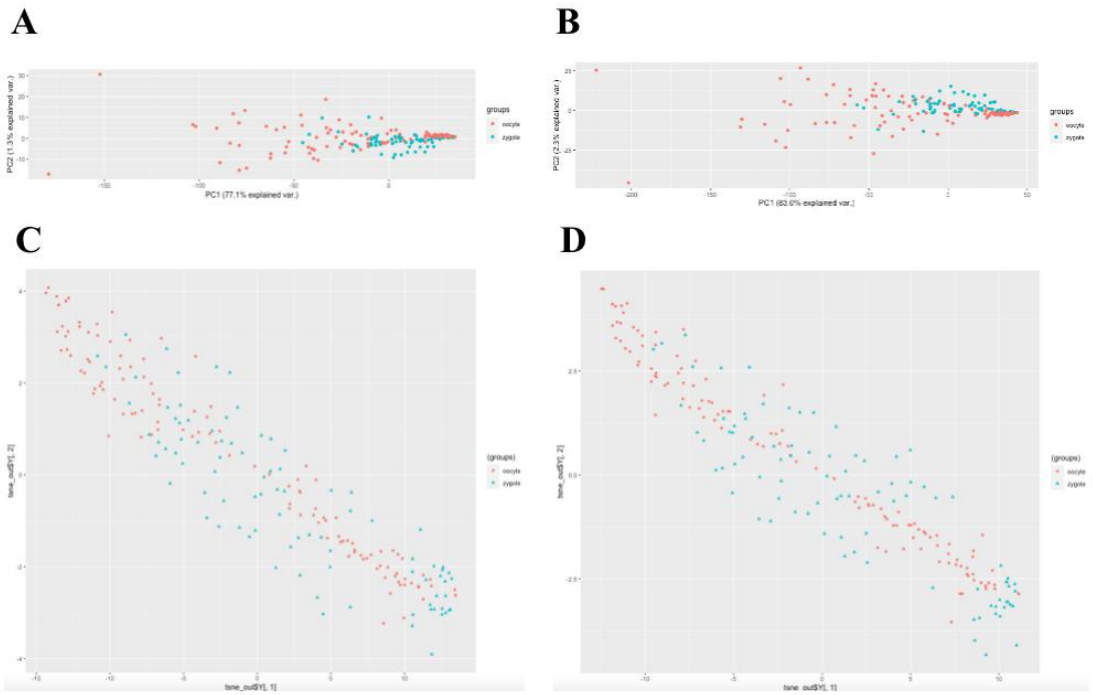

图 4 第一组数据非合并 Compartment 和 TAD 层面聚类结果。其中 A，C 分别为 Compartment 层面的二维样本图和 t-SNE 降维聚类图。B、D 为 TAD 层面。

第二组数据来自于小鼠胚胎干细胞 E14 细胞系和小鼠 NMuMG 细胞系。在染色质层面（图 5），水滴图可以发现 PC1 和 PC2 整体对所有变量的代表性最强，即所有变量对 PC1 和 PC2 整体贡献程度最大，从而证明 PC1 和 PC2 更适合用于作图聚类。随后，我们利用变量坐标（coord）与相关性（cor）进行可视化展示，发现变量 20（X 号染色体）相关性较大，变量 6（号染色体）相关性较小。除此之外，我们用点的大小表示当前的 PC 集[PC1 与 PC1]对该样本的解释程度（点越大，样本对 PC 集的贡献越大）绘制了基于 PCA 的样本聚类可视化图，直观展示了样本集的贡献程度。在该层面的 t-SNE 降维聚类图中，两类细胞可以完全分离。但图中可以发现，对于小鼠胚胎干细胞额外分开成为两类，我们考虑可能与论文中提到的小鼠胚胎干细胞的样本是两个批次有关。合并 Compartment 层面（图 6），我们根据水滴图，选择整体贡献程度最大的 PC1 和 PC3 进行后续分析。从该层面的 t-SNE 降维聚类图也能发现，两类细胞可以完整分离，但相对染色体层面更加散布。而在非合并 compartment 和 TAD 层面的 t-SNE 降维聚类结果也有类似特征。

在对第二套数据分析之后我们发现该套数据的质量较高，不管是从 PCA 还是 t-SNE 上都能得到不错的聚类效果，并且我们还找到了在每一层次上面贡献度较大的变量，这意味着这些变量是将细胞不同类群分开的主要因素，而接下来就需要对这些变量进行验证。

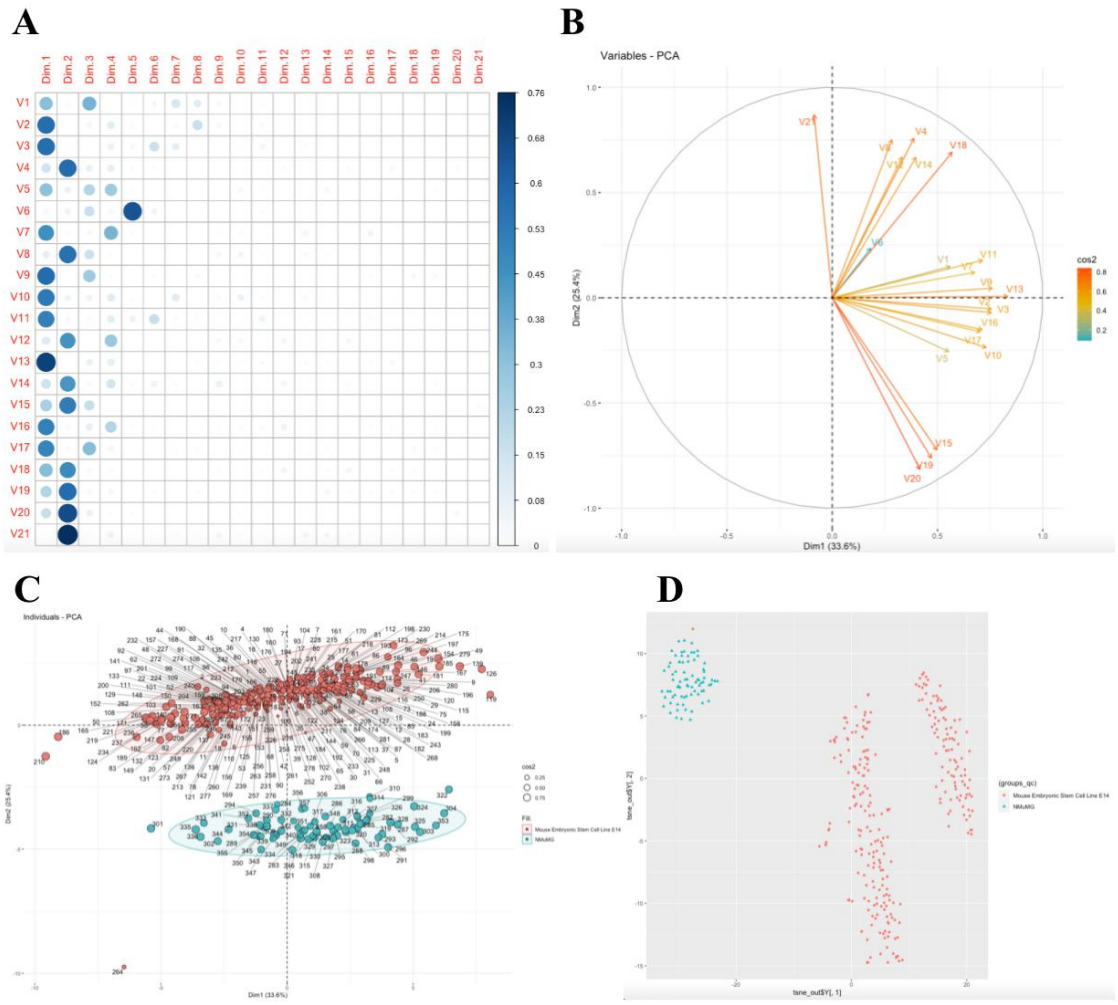

图 5 第二组数据染色质层面聚类结果。图 A 为主成分对变量的代表性的水滴图。图 B 为变量坐标(coord)与相关性(cor)可视化，其中 coord 是坐标，与 cor 数值相同 coord=loading*stdev=loadings*sqrt(eig)。相关图中，靠近的变量表示正相关；对向的是负相关，直角为正交不相关。箭头越远离原点、越靠经圆周则表明 PC 对其的代表性高，即相关性越强。图 C 是基于 PCA 的样本聚类可视化，图中标示出了两类样本的序号，红色代表小鼠胚胎干细胞，蓝色代表小鼠 NMuMG 细胞，点的大小代表当前的 PC 集[PC1 与 PC1]对该样本的解释程度。图 D 为 t-SNE 降维聚类结果，红色为小鼠胚胎干细胞，蓝色为小鼠 NMuMG 细胞。

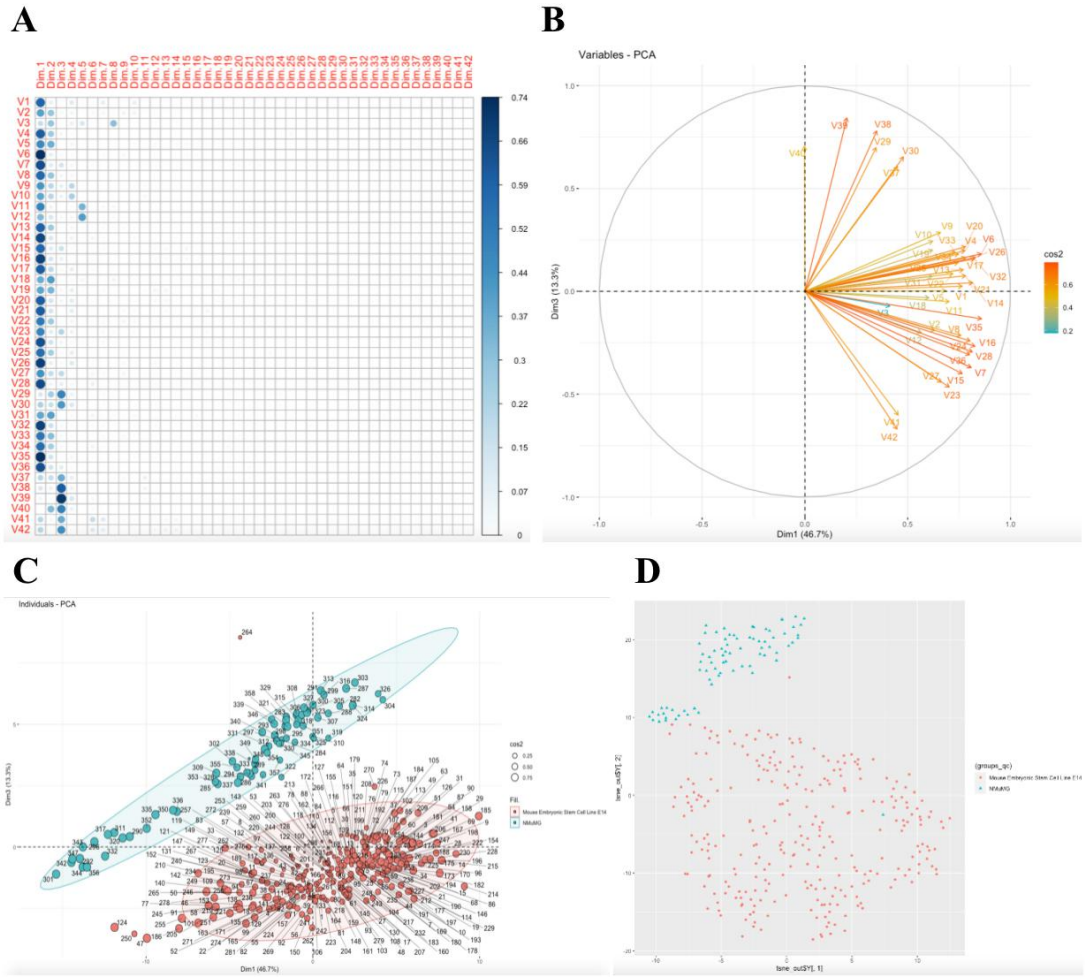

图 6 第二组数据合并 Compartment 层面聚类结果。图 A 为主成分对变量的代表性的水滴图。图 B 为变量坐标(coord)与相关性(cor)可视化，可以看出变量 7（第 4 号染色体的 A compartment）相关性较大，变量 3（第二号染色体的 B compartment）相关性较小。图 C 是基于 PCA 的样本聚类可视化。图 D 为 t-SNE 降维聚类结果，两类细胞可以完全分开，但较为散布。

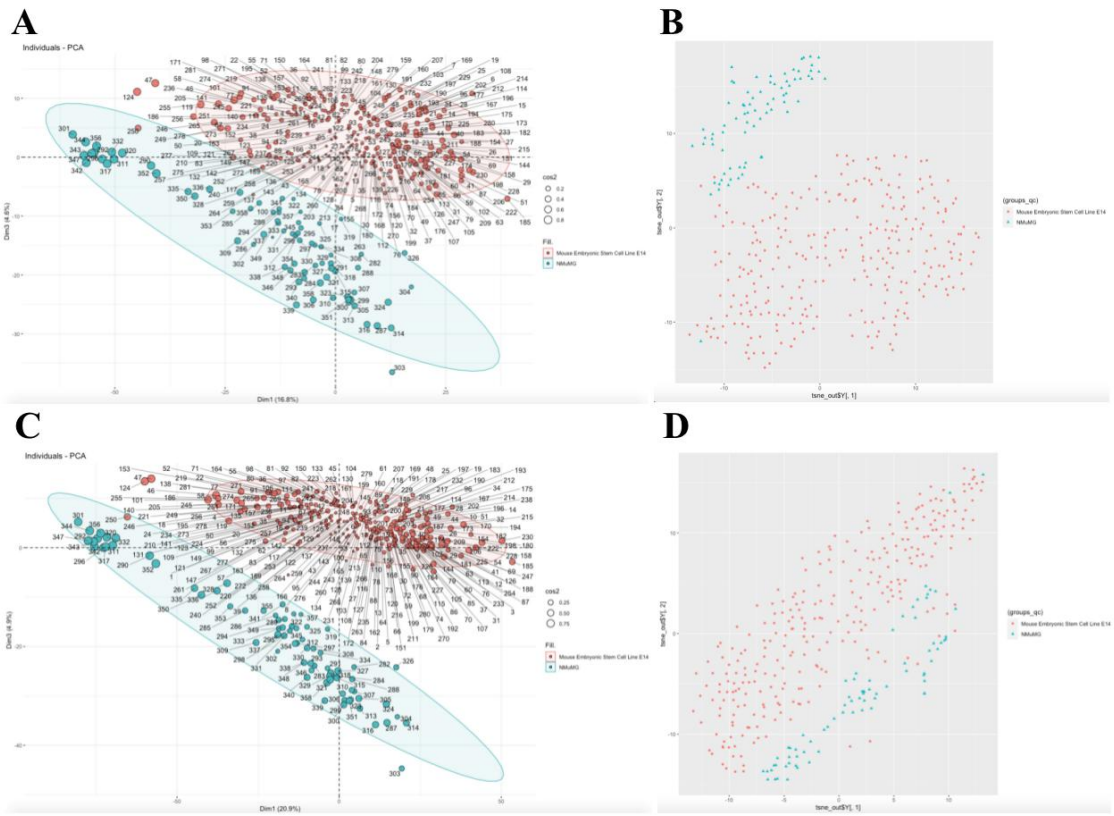

图 7 第二组数据非合并 Compartment 和 TAD 层面聚类结果。其中 A，B 分别为 Compartment 层面的 PCA 的样本聚类图和 t-SNE 降维聚类图。C、D 为 TAD 层面。

3.1.2 差异特征提取

前文讲到，我们可以将每一层面贡献度较大的变量提取出来进行差异验证，看其是否存在某种结构差异或是模式特征，在这里我们选取了第二套数据的 PCA 结果，并使用了热图的方式来展现不同层面的差异性。

- 染色体层面

我们首先通过对各变量的贡献度在 1、2 主成分上的加和结果绘制柱状图（图 8），发现 X 号染色体的加和贡献度最高，即在染色体层面上，各样本间在 X 号染色体的差异最显著。因此我们选择 X 号染色体作为主要分析对象。

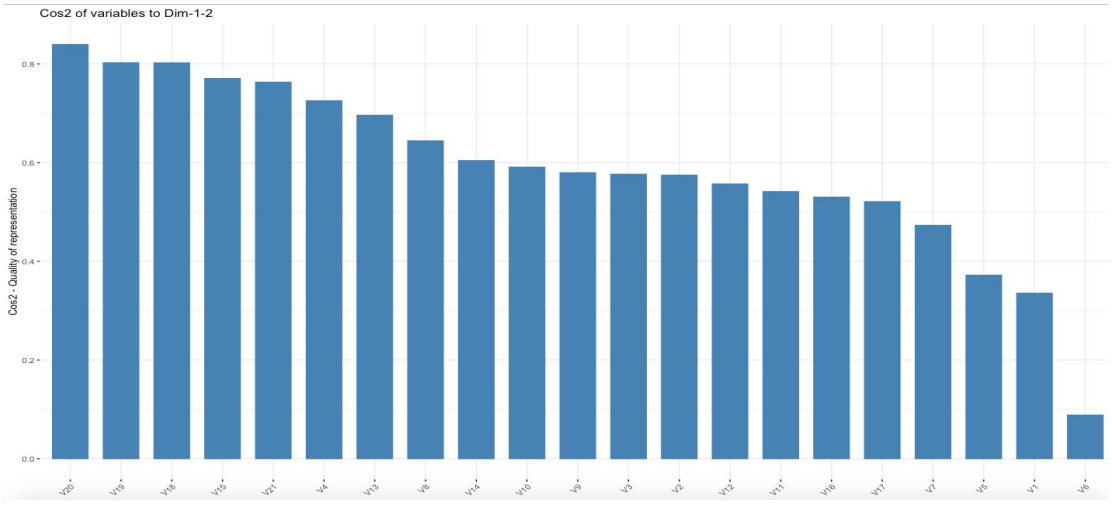

图 8 染色体层面，各变量的贡献度在 1、2 主成分上的加和

图 9 展示了 X 号染色体的热图情况。染色体层面的 PCA 结果表示各样本间 X 号染色体差异较为明显，该图一共进行了两组对照实验，分别为样本 38-321，样本 90-332，每一组里面的两个样本来自于之前展示的 PCA 样本聚类可视化图中点较大的样本，在聚得的两类样本中各选取一个较大的点进行对照实验。由于染色体的尺度较大，热图结果展现了差异较为集中在对角线的空位上。

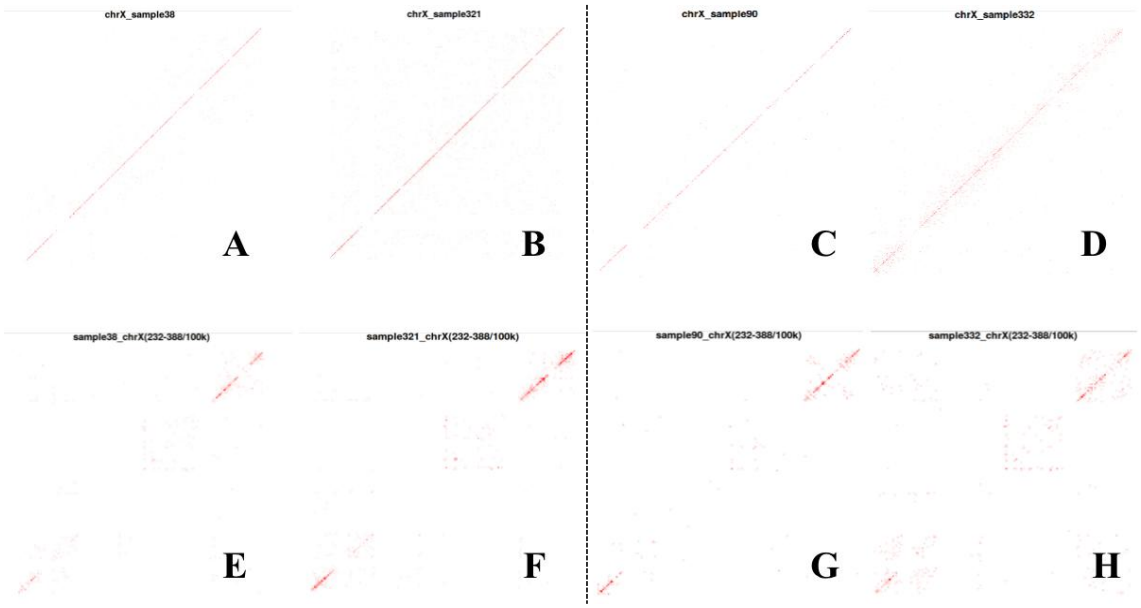

图 9 样本 38-321，样本 90-332 的 X 号染色体的热图。A-D 为全局热图，E-H 为局部热图

- 合并 Compartment 层面

根据贡献度在 1、3 主成分上的加和情况，我们选择 4 号染色体作为主要分析对象。

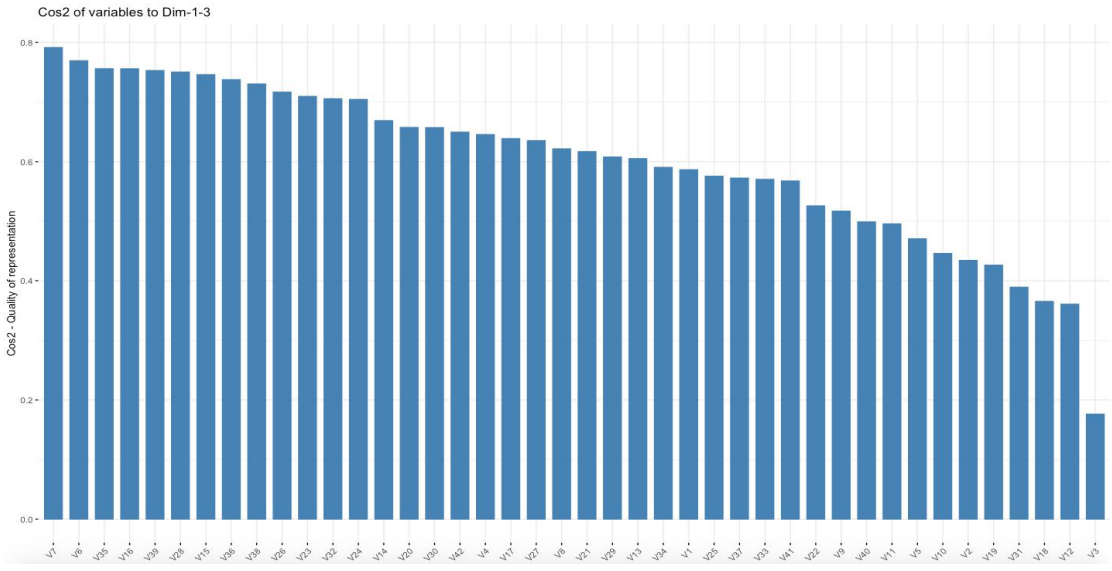

图 10 合并 Compartment 层面，各变量的贡献度在 1、3 主成分上的加和

图 11 展示了 4 号染色体合并 Compartment 层面的热图，其中合并层面的 PCA 结果表明各样本间 4 号染色体的 A Compartment 存在显著差异。此处选取样本 90-332 进行对照实验，两个样本来自于之前展示的 PCA 样本聚类可视化图中点较大的样本，在聚得的两类样本中

各选取一个较大的点进行对照实验。在该热图中，我们分别将 A Compartment 与 B Compartment 划分为正值和负值，在热图中使用暖色调和冷色调将其进行区分。热图的背景底色表示零值的偏向，如果热图底色为暖色调，证明正值即 A Compartment 的矩阵数值较大，而如果底色为冷色调，证明负值即 B Compartment 的矩阵数值较大。在所选两个样本中，90 号样本的 A Compartment 数值偏大，底色呈现暖色调，但仍然可以看到冷色调的 B Compartment，332 号样本的 B Compartment 数值较大，但几乎看不到暖色调的 A Compartment，因此可以证明 PCA 的结果的可信性。

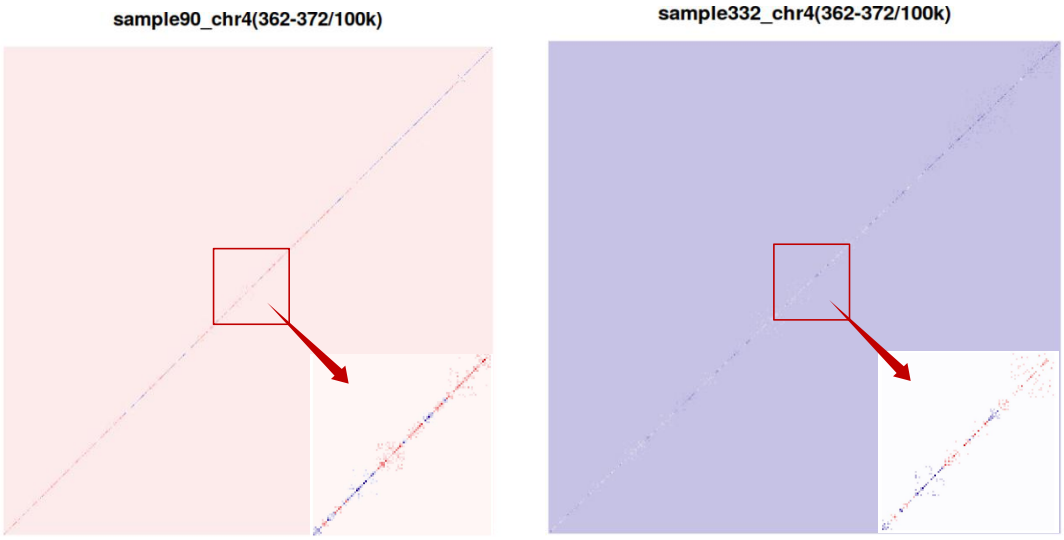

图 11 样本 90 和样本 332 的 4 号染色体热图

- 非合并 Compartment 层面

根据贡献度在 1、3 主成分上的 top10 加和情况，我们选择 19 号染色体作为主要分析对象。

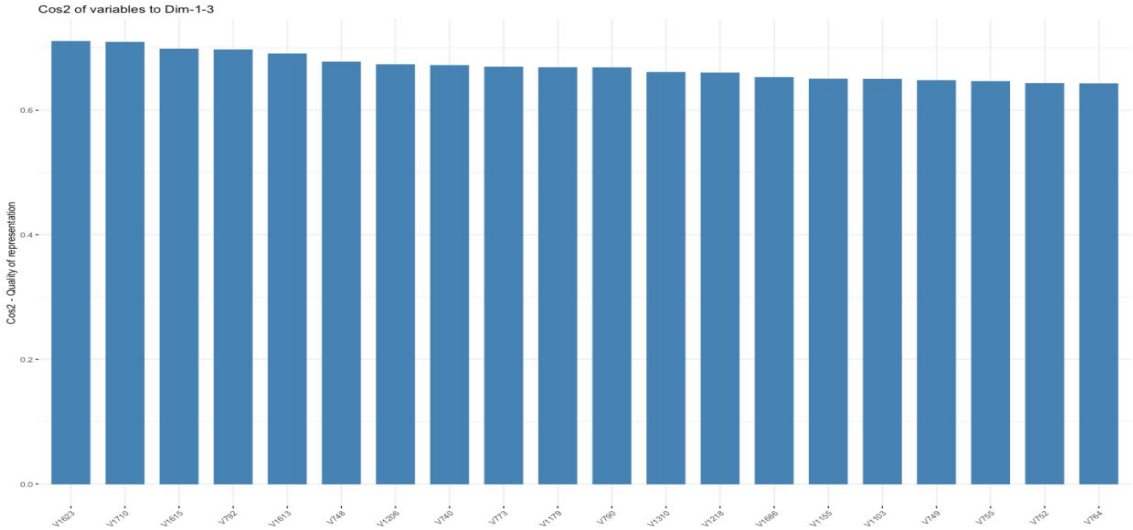

图 12 非合并 Compartment 层面，各变量的贡献度在 1、3 主成分上的 top10 加和

图 13 展示的是 19 号染色体 56M-58M 区间的 B Compartment 热图，其中非合并 Compartment 层面的 PCA 结果显示各样本间 19 号染色体 56M-58M 区间的 B Compartment 存在明显差异，这两张图片展示了该 B Compartment 以及其相邻的两个 A Compartment 的热图。此处选取样本 46-332 进行对照实验，两个样本来自于之前展示的 PCA 样本聚类可视化图中

点较大的样本，在聚得的两类样本中各选取一个较大的点进行对照实验。在 46 号样本的热图中，底色偏冷色调，证明 B Compartment 的数值较大，在 332 号样本中，底色偏暖色调，证明 A Compartment 数值较大，并且可以看到 B Compartment 的数值极少，在原本对应的 B Compartment 区域出现了从暖色调过渡到冷色调的白点，因此可以证明 PCA 结果的可信性。

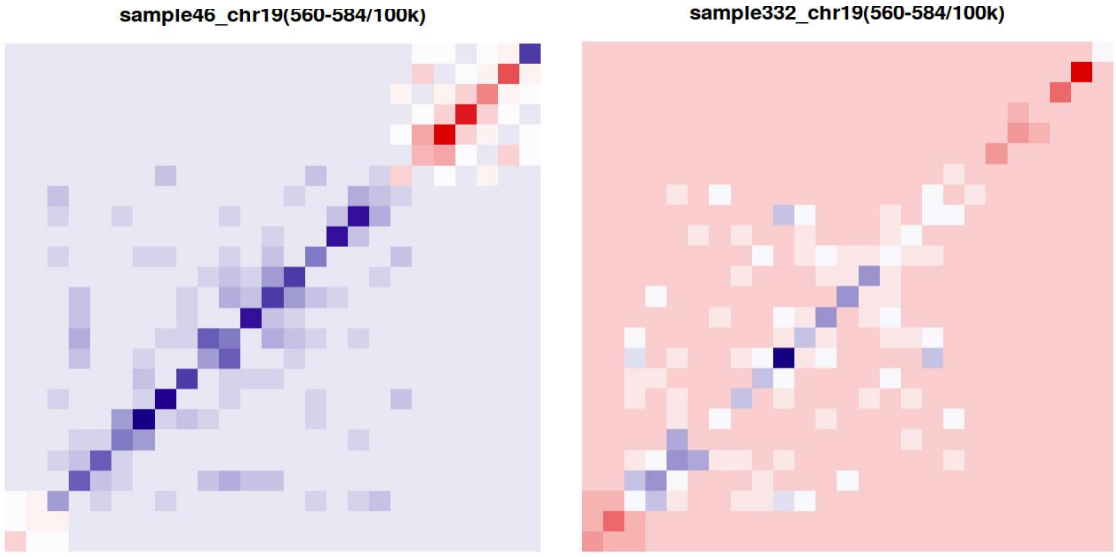

图 13 样本 46 和样本 332 的 19 号染色体局部热图

- TAD 层面

根据贡献度在 1、3 主成分上的 top10 加和情况，我们选择 8 号染色体作为主要分析对象。

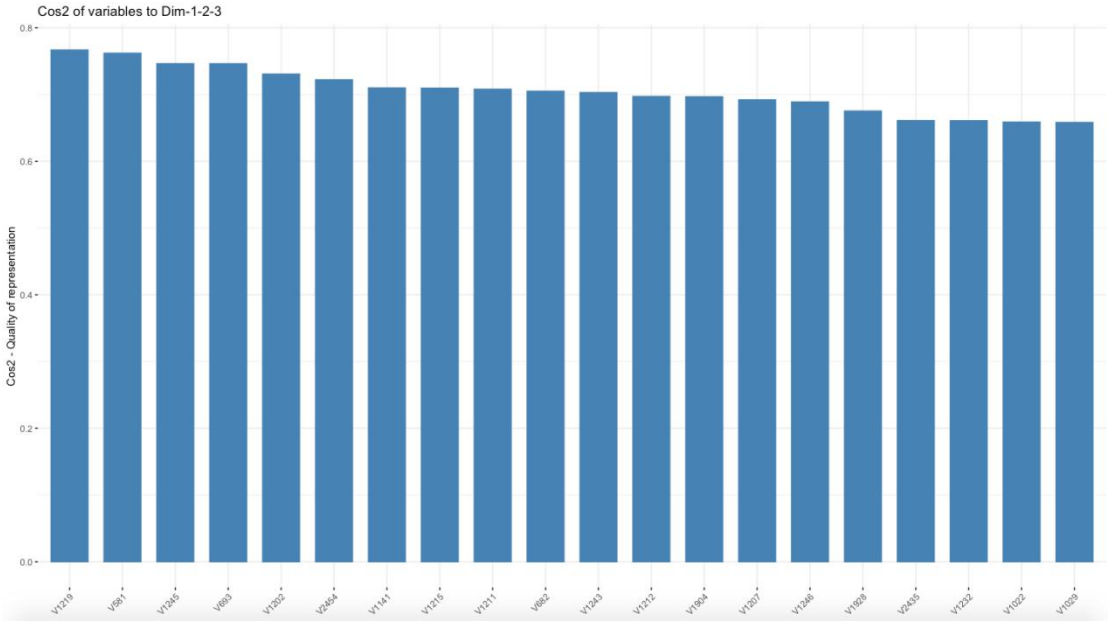

图 14 TAD 层面，各变量的贡献度在 1、3 主成分上的 top10 加和

在该层面，我们分别进行了 3 组对照：

对照一：类间显著样本对照。选取样本 46-343，样本 90-332 进行两两对比，每个样本均来自于图 7C 中节点较大的样本，在聚得的两类样本中各选取一个较大的点进行对照实验。图 15 展示了四个样本中 8 号染色体 36M-37M 区域的 TAD 热图，可以看出，TAD 内部存在明显差异，左侧两个样本（来自小鼠胚胎干细胞 E14 细胞系）比右侧两个样本（来自小鼠

NMuMG 细胞系）分散且数值更大。

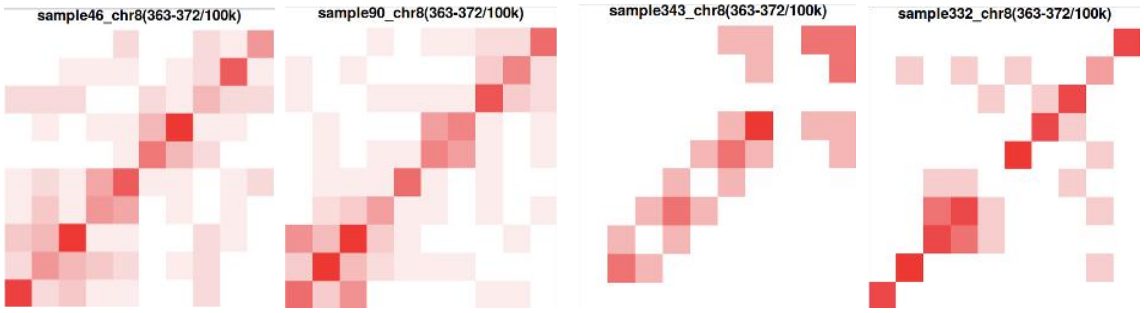

图 15 样本 46 与 90 和样本 343 与 332 的 8 号染色体局部热图

对照二：类间不显著样本对照。选取样本 128-304，样本 243-302 进行两两对比，每个样本均来自于图 7C 中节点较小的样本，在聚得的两类样本中各选取一个较小的点进行对照实验。图 16 展示了四个样本中 8 号染色体 36M-37M 区域的 TAD 热图，同样存在显著差异，相对对照一，TAD 中交互强度普遍偏低，但左侧两样本（来自小鼠胚胎干细胞 E14 细胞系）依旧比右侧两样本（来自小鼠 NMuMG 细胞系）分散且数值更大。

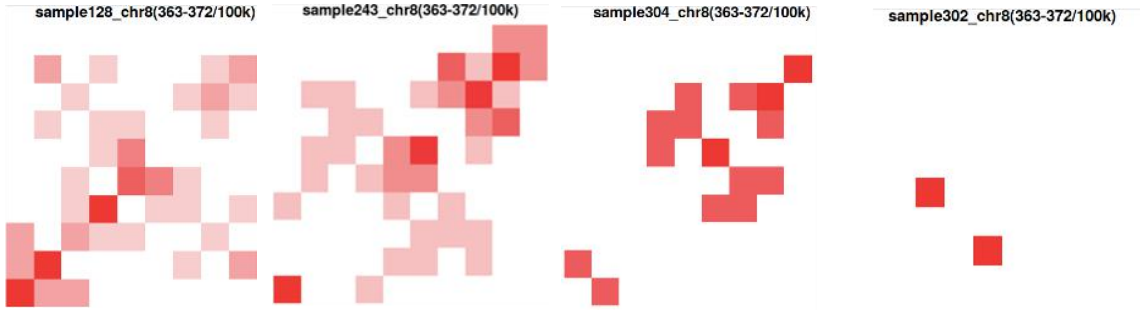

图 16 样本 128 与 243 和样本 304 与 302 的 8 号染色体局部热图

对照三：类内样本对照。选取样本 21-87，样本 328-358 进行两两对比，每组样本来自于同一类且贡献程度相似（节点大小相似）。结果显示小鼠胚胎干细胞这一类样本相对小鼠 NMuMG 细胞样本，其交互关系更为分散且普遍数值更大。

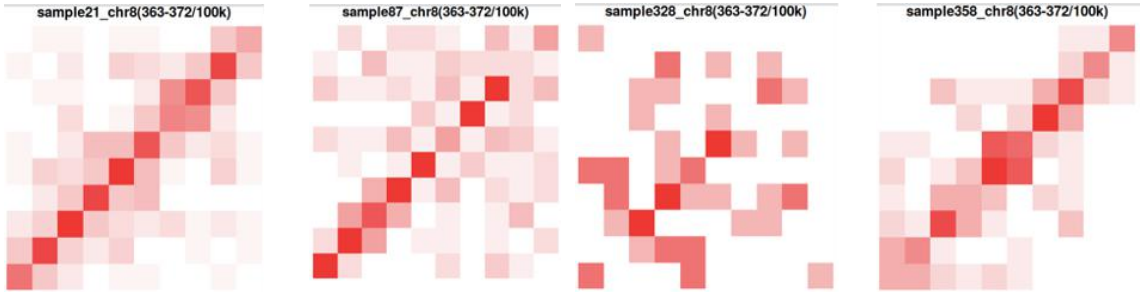

图 17 样本 21 与 87 和样本 328 与 358 的 8 号染色体局部热图

针对对照三所展示的热图，我们利用小组开发的 hicGraph 工具（GitHub 链接了解该工具的更多详细功能：http://www.github.com/LittleHan/HicGraph），将热图以网络图的形式展现。从网络图也能明显看出，在 PCA 分析结果显示差异较为显著的这个 TAD 上，胚胎干细胞的交互较多且均匀，NMuMG 细胞的交互较少且差异较大。

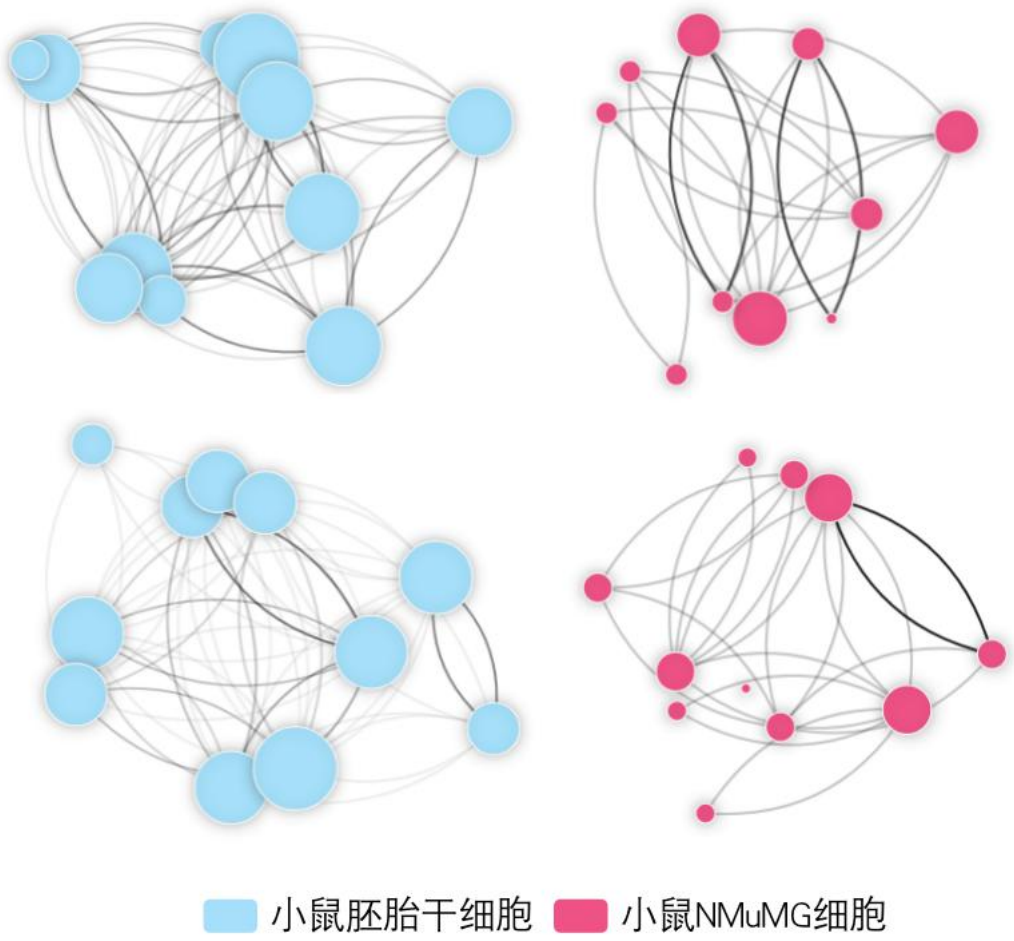

图 18 样本 21 与 87 和样本 328 与 358 的 8 号染色体局部热图网络化

在上述的展示中，我们对染色质、合并/非合并 Compartment 与 TAD 层面进行了差异特征提取，发现 PCA 的结果具有一定的可信度，且在不同层面都能看到明显的差异存在，但是由于我们目前的算法只能精确到一块 TAD 区域内的计数，并没有精确到区域内的所有交互，因此该方法在模式特征的挖掘上还缺乏一定的严谨性。

4 总结与讨论

本项目中，为了挖掘出基因组序列、基因结构以及调控元件等所携带的遗传信息在三维空间结构上存在的调控关系以及相互作用，一方面，我们通过对 Hi-c 数据进行处理，得到了染色质、合并/非合并 Compartment 与 TAD 三个大层面上的单细胞聚类结果。随后基于交互区域计数方法对聚类结果进行了差异特征提取。另一方面，通过多种软件，对不同细胞的在 Compartment 和 TAD 两个层面进行了网络图以及可视化处理，获取了具有特征的聚集以及相互联系的“模式”,并将整套可视化流程进行了打包处理，即：实现了对大部分的 Hi-c 单细胞数据，可以通过修改命令行参数的方式进行统一处理，大大提高了数据处理及可视化的效率与可信度，并将有助于后续项目的进行。

在层次聚类与差异特征提取方面，我们选取了两套单细胞测序数据，第一套为小鼠受精卵与卵母细胞数据，第二套为小鼠胚胎干细胞和 NMuMG 细胞系数据，对数据进行三步严密的质控，从深度、广度和交互比例控制染色质交互数据的质量。在对质控之后的数据进行

分 bin 之后，我们得到了染色质、合并/非合并 Compartment 与 TAD 三个大层面的计数矩阵，并使用目前主流的 PCA 与 t-SNE 降维方式对其进行降维处理，最终发现 PCA 在维度分析上面比破坏维度特征的 t-SNE 更加精确、有效，于是我们使用了 PCA 的分析结果对样本进行了聚类，在聚类之后我们舍弃了效果较差的第一套数据，我们认为其聚类差的原因可能是受精卵与卵母细胞在染色质交互计数上存在较大的相似性，因此在之后的差异特征提取中我们只针对于第二套数据开展了实验。

通过利用每一层面对维度集贡献度较大的变量，我们实现了差异特征提取。由于贡献度较大则意味着各样本间在该变量上的差异性较为显著，在对染色质、合并/非合并 Compartment 与 TAD 层面进行了差异特征提取之后，我们发现 PCA 的结果具有一定的可信度，且在不同层面都能看到明显的差异存在，但是由于我们目前的算法只能精确到一块 TAD 区域内的计数，并没有精确到区域内的所有交互，因此该方法在模式特征的挖掘上还缺乏一定的严谨性。目前我们的层次聚类和差异特征提取是基于交互区域计数的阶段性成果，其整体流程也较为成熟，在之后的研究里，我们也将继续从不同层面的单一交互作用来进行模式挖掘，希望能够将固定的模式差异提取出来，以对基于 Hi-C 的单细胞判别方法进行更加全面的完善。

参考文献

[1] Ilya M Flyamer, Johanna Gassler, Kikuë Tachibana-Konwalski.Single-cell Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature,2017;544(7648):110-114

[2] Dong-Sung Lee, Chongyuan Luo, Jingtian Zhou. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nature Methods,2019;16(10):999-1006

[3] 曹靖城,张继东,王培才.基于 PCA 降维的海量数据特征抽取技术研究[J].通讯世界,2020,27(07):83-84

[4] 于慧伶,霍镜宇,张怡卓,蒋毅.基于 PCA 与 t-SNE 特征降维的城市植被 SVM 识别方法[J].实验室研究与探索,2019,38(12):135-140

致谢

在这长达一年的项目研究过程中，我们遇到了不少困难与阻碍，如今看到文稿最终写完，我们每一个人内心都是满满的成就感。

在这篇文章中还有许多老师及师兄师姐的帮助，在这里，我们特别感谢不厌其烦为我们调试环境的钟权师兄、参与讨论的王子林师兄、维护服务器的张巍瀚师兄。我们还要特别感谢指导老师李立老师，感谢李立老师在整个项目推动的过程中提供的及时指导、前瞻性建议及思考，及时给予的反馈、解答疑惑，每次参与组会讨论总是收获颇多。此外，还十分感谢信息学院对“项目育人”的支持，是“项目育人”让我们有了这次难忘的经历，充分锻炼了自己的能力。

最后，还要感谢这篇论文引用的参考文献、相关论坛资料的作者及拥有者，除却这些资料，我们将不能顺利完成项目。在此，我们献上真挚的感谢！此致！