人类胚胎细胞分析

包亦爽 2015012037 蒋昊峻 2015012259 熊思儒 2016012092

**摘要：**在本次实验中，我们采用了KMeans，高斯混合模型以及层次聚类三种方法对E5和E7的数据进行了聚类，其中层次聚类得到的数据在有监督学习上性能最好。我们采用层次聚类的数据，尝试训练了多种分类器，包括线性判别器，支持向量机，随机森林，决策树，多层感知器以及k-邻近分类器，其中k-邻近分类器的性能最佳。利用训练所得的k-邻近分类器对E3,E4,E6,E7细胞进行谱系分类，得到的结果与生物学先验知识相吻合。

**关键词：**无监督学习，有监督学习，人类胚胎细胞，聚类，分类

# **介绍**

在人类发育的前7天，受精卵经历细胞分裂并分化成成熟胚泡前的三种不同细胞类型：滋养外胚层(trophectoderm-TE)，原始内胚层(primitive endoderm-PE)和外胚层(epiblast-EPI)，而未分化的细胞称为pre-lineage。现在有1529个人类第3天到第7天胚胎细胞的基因表达数据，我们需要利用无监督的聚类方法将第5天和第7天细胞进行聚类，分离出4个谱系的细胞，并用第5天无监督聚类所得数据进行有监督学习，对第3,4,6,7天细胞进行分类，并分析细胞的分化情况。

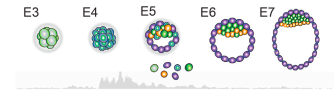


图 1 人类早期胚胎细胞

# **相关工作**

Petropoulos等人于2016年发表了文献**错误!未找到引用源。**，他们认为尽管对小鼠的研究能够帮助人类理解早期细胞谱有所帮助，但是人类细胞谱系的发展有着很大的不同，在小鼠上的研究结果无法很好地迁移到人类细胞中。因此他们收集了88个人类植入前胚胎的1529个单细胞基因测序转录组，在此基础上提出了了人类胚胎发育的综合转录图谱以及人类早期细胞谱系图。

# **实验方法与结果**

在实验任务中我们分别采用以下聚类和分类算法（见图1）：

下面分别介绍实验中方法。

图 2 聚类与分类算法

## **无监督学习**

在该任务中我们需要用无监督学习的方法对E5和

E7的细胞做谱系分类，并建立两者的映射。

一开始所有早期胚胎细胞都处于pre-lineage状态，即细胞未分化。随着细胞发育，在第一阶段，细胞先分化为inner cell mass (ICM)和TE，这一分离出现在E5细胞中。紧接着，ICM细胞分化为EPI和PE，这一分离也出现在E5细胞中。也就是说，ICM是一个中间态。在E7细胞中，几乎不应该存在pre-lineage细胞，所有细胞高度分化为TE, EPI和PE。结合以上先验知识以及文献[1]我们将无监督学习分为以下3个步骤：

1. 分离出E5中的pre-lineage细胞；
2. 用聚类方法将E5中的分化细胞分为ICM和TE；
3. 用聚类方法将第2步得到的ICM细胞分为EPI和PE。

由于我们通过机器学习方法得到的只是聚类，而不能分辨每一类的名字。所以我们根据文献所做的结果中各类细胞的数量关系（如下图），来给各类命名。(注: E5共有377个细胞，图中没列出pre-lineage细胞，通过差值可以得到E5有162个。)

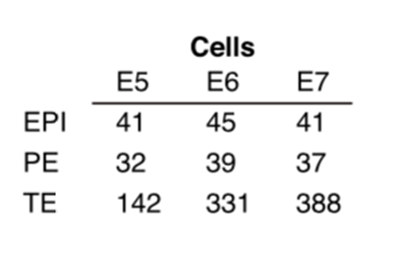


图 3 E5细胞文献[1]分类结果

### **分离出E5中的pre-lineage细胞**

在文献[1]中，作者采用的是一种叫做pseudo-time cutoff的方法。我们的理解是，通过t-SNE图像，可以明显看出E3到E7各天细胞之间的分散化（见图2）。E3, E4中的细胞几乎未分化，属于pre-lineage这一类。因此E5的pre-lineage应该和E3, E4中的细胞有很大的相似度，即在图中这些细胞位置会和E3,E4的细胞的位置比较接近。这样我们可以人为的设置第一个cutoff，分离出E5中的pre-lineage细胞。

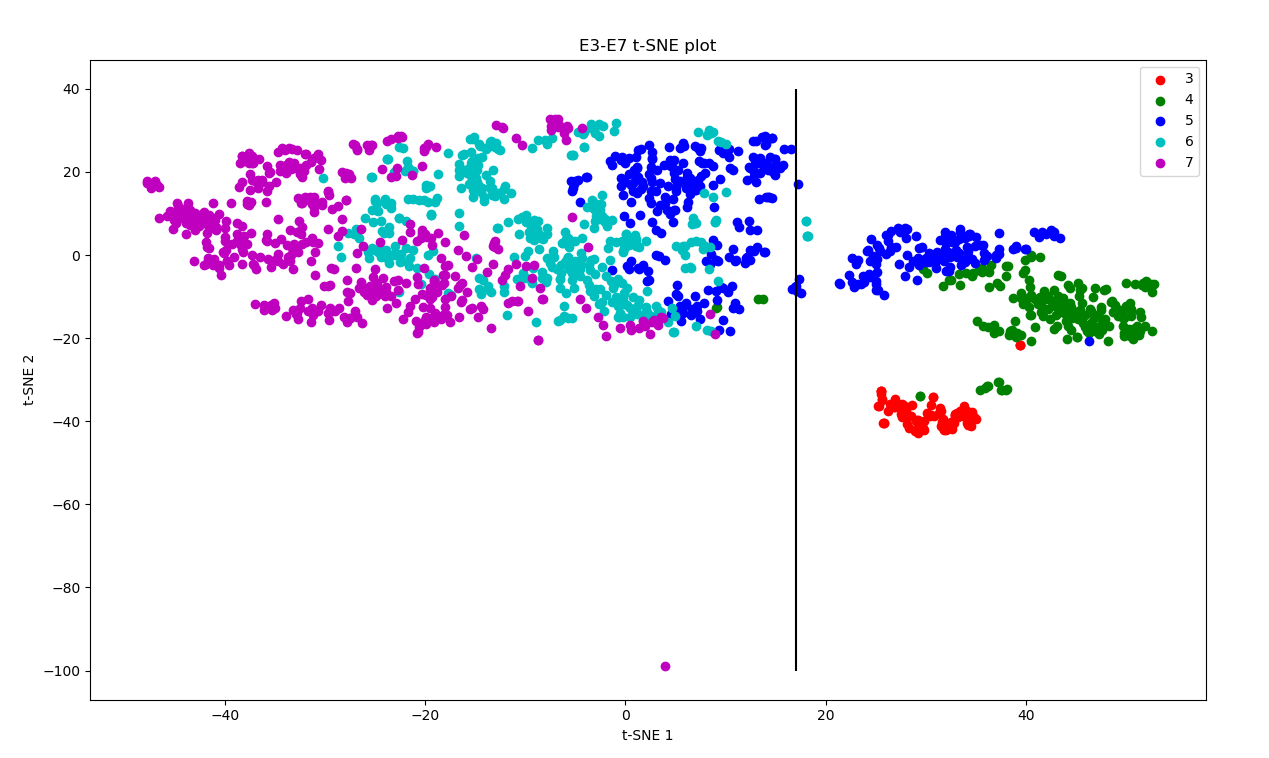


图 4 E3至E7细胞t-SNE图

从上图可以看到，在黑线右边的E5细胞 (蓝点)和E3，E4比较靠近，而和黑线左边的较分化细胞有明显的分离。因此我们选取的所有t-SNE 1（横坐标）大于cut-off的蓝点，将它们认定为E5的pre-lineage细胞。cut-off可作为一个参数调整，实验中我们测试发现当cut-off设置为t-SNE 1=15（图中黑线位置），可以得到140个E5 pre-lineage细胞，和文献结果较为接近，并且从t-SNE图看该分界线也比较合理。

### **. 将分化细胞分为ICM和TE**

第二步，我们先用主成分分析（PCA）对E5进行了特征提取，采用了前3个主成分。我们采用三种聚类方法对E5的所有分化细胞（E5细胞出去上面分离掉的pre-lineage细胞）进行二分类。我们用的主要聚类方法有KMeans聚类, 高斯混合模型和层次聚类。

1. KMeans聚类

经过参数调整，我们得到较好的分类效果如下图所示（见图3），可以发现KMeans的分离效果是比较好的，红点（）和绿点有比较明显的分界。但是其聚类得到ICM细胞数量与文献[1]中报道的数量相差甚远。

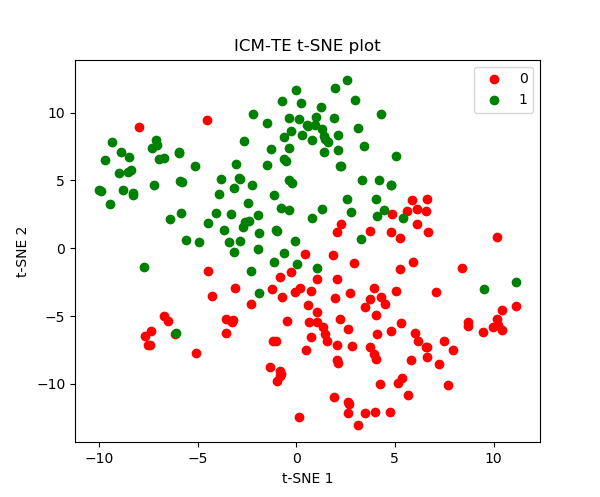


图 5 E5 ICM-TE KMeans聚类t-SNE图

1. 高斯混合模型聚类

从下图可以看出，分离效果也比较好，但是同样两类的数量差异较大，关于这一点，我们在之后会统计数量然后和文献中的各类数量比较。

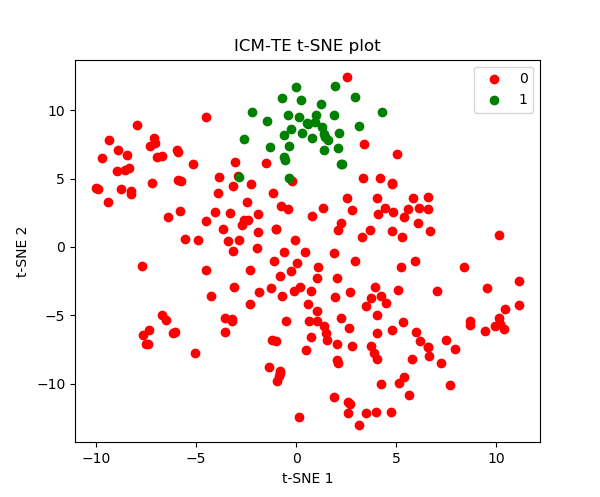


图 6 E5 ICM-TE GMM聚类t-SNE图

1. 层次聚类

实验结果如下图所示（见图5），分离效果比较好，并且两类的数量相比前两种聚类方法更加符合实际情况。

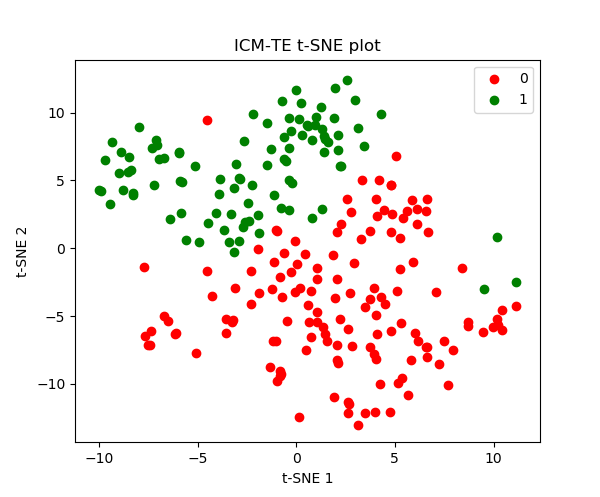


图 7 E5 ICM-TE 层次聚类t-SNE图

统计出上面三种方法的分类结果，注意，在这里ICM和TE的区分主要是基于文献中的结果，TE的数量比ICM要多很多，因此我们将数量较多的一类定义为TE细胞。由下表可知，层次聚类的结果与实际情况接近。



表 1 E5 ICM-TE聚类数量

### **将ICM分成EPI和PE**

根据文献[1]，ICM在E5的后期会继续分裂成EPI和PE两种谱系，ICM是一种过渡状态。于是我们基于上面挑出的ICM细胞，继续做聚类分析。方法与3.1.1中一致，不再赘述，直接给出聚类结果。

1. KMeans聚类

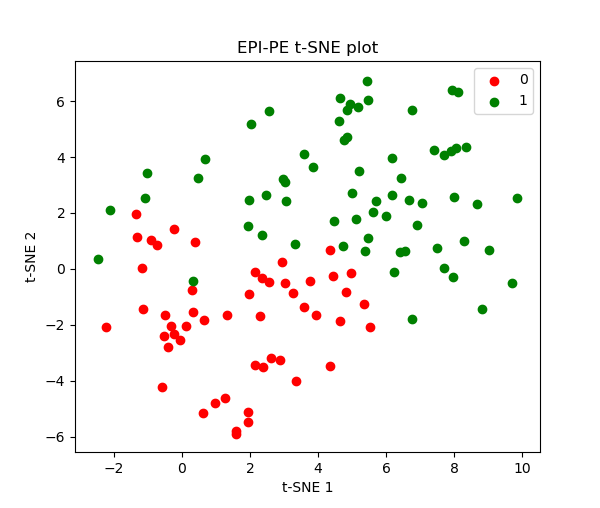


图 8 E5 EPI-PE KMeans聚类t-SNE图

1. 高斯混合模型聚类

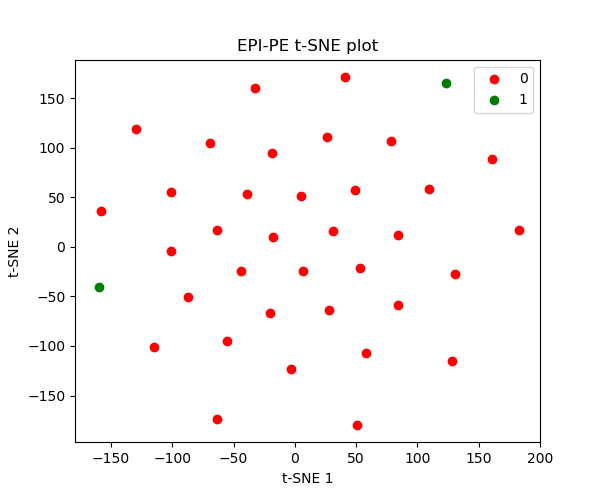


图 9 E5 EPI-PE GMM聚类t-SNE图

1. 层次聚类

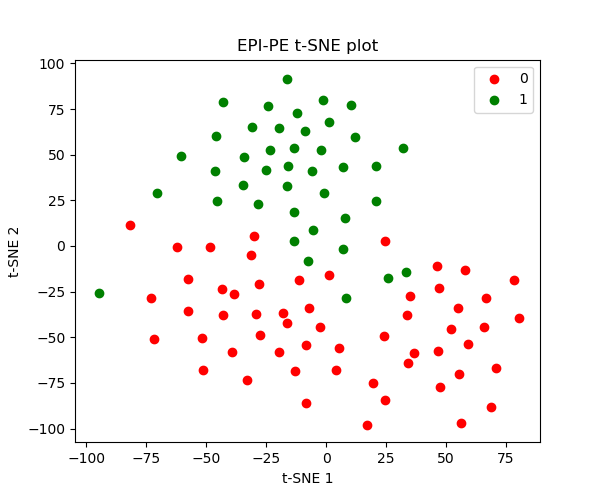


图 10 E5 EPI-PE 层次聚类t-SNE图

综上，我们有以下统计结果（见表2），可以看到，KMeans和层次聚类的结果与文献结果较为吻合。为了找到最好的聚类，必须从第二个任务，即有监督学习的测试结果中观察，并根据结果调整这部分的参数。从表中也能看出，结果与文章有一定差异，这可能是由于信息量有限，文章中有15000左右个基因，而我们只用了490个。



表 2 E5聚类结果

## **监督学习**

在任务一中，我们使用3.1中无监督学习方法，对E5细胞和E7细胞进行了分类。在监督学习中，我们用聚类的结果训练出一个分类器，然后用交叉验证评价分类器的性能，并将分类器运用于E3,E4,E6,E7类细胞，分析结果。特别的，对于E7类细胞（作为测试集），我们将该分类器的分类结果与之前无监督聚类的结果进行了比较，从侧面验证E5与E7无监督聚类的合理性和一致性。在3.1中我们已经知道KMeans聚类和层次聚类的结果与实际情况比较吻合，因此在监督学习中我们测试两种聚类的数据的性能，实验中发现层次聚类的方法要比KMeans好很多，因此最终我们采用了层次聚类，下面的实验结果都是在层次聚类的数据上得到的。

### **特征选择**

在监督学习中，我们采用了有recursive feature elimination的特征选择方法，我们用E5的数据进行了特征的选择，并对E3/E4/E6/E7数据进行同样的特征选择，用交叉验证分析分类器在E5数据上的性能以及比较聚类和分类结果，分析分类器在E7数据上的性能。由于该方法要求选用的分类器必须有“coef\_”和“feature\_importances\_”的性质，我们仅能在线性判别分析和线性支持向量机上运用特征选择。实验的结果如下表所示，可以看到运用特征选择后，在分类器在E5数据上的表现都有所提高。但是在LDA模型上，采用特征选择后，分类的泛化能力大大减弱，我们分析这是由于特征选择是基于E5的分类效果，在选择过程中可能筛去了E7关键的基因表达特征，



表 3 特征选择结果

### **分类方法**

在监督学习中，我们还尝试了随机森林，决策树，k-近邻，MLP，支持向量机。以上几种方法的实验结果如下两表所示，其中包括分类器在E5和E7上的准确率（实验时准确率有一定的浮动）:



表 4 不同分类器分类准确率

从上表中，可以看出k-近邻法的效果最好。实验中，我们测试了多种k值，最好的是k=6。验证集正确率在91.81%，测试集正确率则达到了81.12%。这是我们做出的最理想的结果。

下面分析k=6的分类器对E3,E4,E6,E7细胞谱系分类的情况（其他分类器对E3,E4,E6,E7细胞谱系分类的情况见supplementary的figure）：

1. E3：能够将E3的81个细胞全部识别为Pre-Lineage。E3是年龄较小的细胞，可能还未分化。因此这个分类结果是符合先验知识的，比较理想。

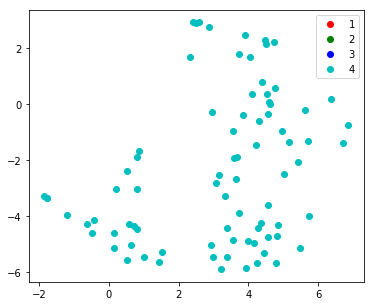


图 11 E3谱系分类t-SNE图(1-TE,2-EPI,3-PE,4-PRE)

1. E4：186个细胞被分类为Pre-Lineage，其余4个细胞被分类为PE。我们预期会有很多细胞被分类为Pre-Lineage。但是这个结果比较极端，让我们比较意外，于是我们分析了其他方法的分类情况，发现如果使用MLP ，SVM或随机森林，结果类似。这说明，E4细胞的分化程度应该是比较低的，仅有个别细胞开始分化。

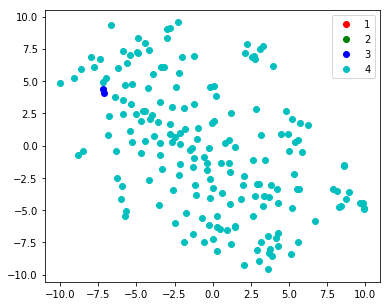


图 12 E4谱系分类t-SNE图(1-TE,2-EPI,3-PE,4-PRE)

1. E6：4个细胞被分类为Pre-Lineage，79个细胞被分类为PE，99个细胞被分类为EPI，233个细胞被分类为TE。Pre-Lineage细胞很少，我们预期没有Pre-Lineage细胞，这里可能存在一点误差，但整体效果还是比较符合先验知识的。分类情况见下图：

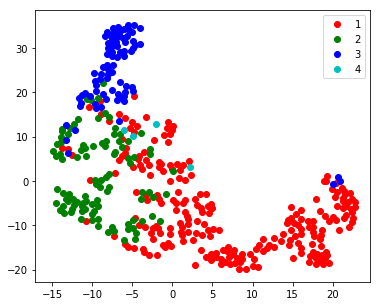


图 13 E6 谱系分类t-SNE图(1-TE,2-EPI,3-PE,4-PRE)

1. E7：59个细胞被分类为PE，93个细胞被分类为EPI，314个细胞被分类为TE。注意，没有细胞被分类为Pre-Lineage，这个结果与先验也是比较一致的，因为E7细胞的年龄较大，我们期望其分化程度相对之前要高一些，故没有Pre-Lineage细胞。我们先做个PCA，然后用t-SNE可视化数据。1类细胞（TE 滋养层）占比变大，与先验知识相符合。分类情况如下如所示：

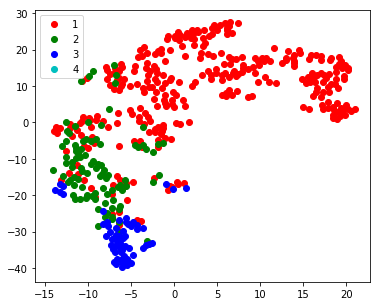


图 14 E7 谱系分类t-SNE图(1-TE,2-EPI,3-PE,4-PRE)

# 实验结论

在本次实验中，我们采用了KMeans，高斯混合模型以及层次聚类三种方法对E5和E7的数据进行了聚类。 我们在有监督学习中比较了三种聚类方法的效果，其中层次聚类得到的数据在有监督学习上性能最好。

我们采用层次聚类的数据，尝试训练了多种分类器，包括线性判别器，支持向量机，随机森林，决策树，多层感知器以及k-邻近分类器，其中k-邻近分类器的性能最佳。根据k-邻近分类器对E3,E4,E6,E7细胞的谱系分类结果可知：E3细胞是处于未分化阶段，所有细胞是pre-lineage；E4细胞处于刚开始分化阶段，绝大部分细胞是pre-lineage，仅有个别细胞是PE；E6，E7细胞都处于高度分化的阶段，几乎全部属于TE,EPI和PE，其中TE占大部分。

# 组员分工

全组成员参与了前期的文献调研，方案设计与探讨，在定下正式方案后：

1. 包亦爽同学负责了无监督学习部分的代码以及报告的撰写；
2. 熊思儒同学负责了监督学习部分的代码以及报告的撰写；
3. 蒋昊峻同学负责了代码的整合以及调试修改，调整参数进行实验以及报告的整合排版。

# 感谢

最后感谢两位助教在实验过程中的大力帮助以及胡彦同学帮忙解答了相关生物学问题。

# 参考文献

1. Petropoulos S, Edsgärd D, Reinius B, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos[J]. Cell, 2016, 165(4):1012-1026.