[特异调控元件鉴定软件]

[V1.0]

使用手册

（如文档为软件设计说明，请把“使用手册”更改为“设计说明”）

目录

[1. 引言 2](#_Toc146554841)

[1.1 编写目的 2](#_Toc146554842)

[1.2 背景 2](#_Toc146554843)

[1.3 参考文献 4](#_Toc146554844)

[2. 总体设计 5](#_Toc146554845)

[2.1 目标 5](#_Toc146554846)

[2.2 功能 5](#_Toc146554847)

[2.3 使用方法 5](#_Toc146554848)

[3. 运行环境 11](#_Toc146554849)

[3.1 硬件 11](#_Toc146554850)

[3.2 支持软件 12](#_Toc146554851)

[4. 使用说明 12](#_Toc146554852)

[4.1 软件安装和初始化 12](#_Toc146554853)

[4.2 输入文件 14](#_Toc146554854)

[4.3 配置程序 15](#_Toc146554855)

[4.4输出结果 17](#_Toc146554856)

[4.5求助查询 19](#_Toc146554857)

[5. 运行说明 21](#_Toc146554858)

[5.1程序存放位置 21](#_Toc146554859)

[5.2运行步骤 21](#_Toc146554860)

[5.3运行时间 23](#_Toc146554861)

[5.4运行错误说明 23](#_Toc146554862)

[5.5运行成功说明 24](#_Toc146554863)

# 1. 引言

## 1.1 编写目的

顺式调控元件是位于基因组上的一类能够调控细胞类型特异基因表达模式的非编码DNA序列，一般包括启动子、增强子和绝缘子。顺式调控元件通过转录调控在真核生物的生物发育、外界环境的响应适应等生物过程中发挥着重要作用。特异顺式调控元件是指能够调控基因在某一种或几种特定细胞类型中特异性高表达的顺式调控元件。通过挖掘和鉴定特异调控元件，不仅有利于理解细胞类型特异的表观调控转录机制，还可能应用到基因治疗的高效治疗载体研发中，避免因治疗基因组成型表达引起的潜在风险。然而，如何能够高效和高精度地筛选鉴定某一种或某几种细胞类型的特异调控元件是特异调控元件应用研究的必要前提和关键问题。

对此，我们开发了特异调控元件鉴定元件（CREfinder）。该软件以完成差异分析后得到的目标细胞类型的上调基因和上调开放区域作为输入，通过软件内部构建的机器学习算法对两者进行特征筛选，从而筛选出目标细胞类型特异的调控元件。通过使用该软件，研究者能够根据实际需求定制化地筛选和鉴定某一种或某几种目标细胞类型的特异调控元件，从而用于下游的特异调控元件应用研究。

## 1.2 背景

顺式调控元件(*Cis*-regulatory elements, CREs)是位于基因组上的一段非编码DNA序列。顺式调控元件包含用于激活和维持转录所需的转录因子和/或其他调控分子的结合位点。启动子和增强子是最容易理解的CREs类型。大多数真核生物基因含有一个位于转录起始位点附近的启动子，尽管有些基因也含有可选择的启动子，它们通常在特定条件下激活基因组中不同位置的转录。启动子是真核生物转录所必需的，但它们一般只能产生基础水平的mRNA。由于启动子序列与一组广泛使用且高度保守的转录调控因子结合，因此它们似乎不是顺式调控分歧的主要驱动因素。然而，启动子突变是人类疾病的常见原因。

相对于启动子，增强子在物种间的差异性更大，通常被认为是导致顺式调控分歧的主要CRE类型。在具有多个不同细胞类型的后生动物中，基因表达通常受到多个增强子的调控，每个增强子控制在一组受限的细胞或组织中的表达或在特定发育阶段中。通常认为每个增强子控制独特的基因表达子集。然而，最近已经识别出具有重叠功能的增强子对，它们有助于表型稳定性。

增强子的功能独立性意味着一个增强子的突变对其他增强子调控的基因表达产生有限的影响。增强子通常位于调控的基因的上游(5')、下游(3')或内含子中，但它们也可以位于一定距离之外。尽管功能同源的增强子通常位于相似的基因组位置，在物种之间也有保守性，但并非总是如此。

机器学习（Machine Learning，ML）是一门研究如何使用计算机模拟人类学习活动，研究计算机的自我改进方法，以获得新知识和新技能，识别现有知识，不断提高成绩和成就的学科。与人类学习相比，机器学习学习得更快，知识的积累更有利于学习的结果更容易传播。因此，人类在机器学习领域的任何进步，都将增强计算机的能力，从而对人类社会产生影响。

在机器学习的过程中，外部环境为系统提供的信息质量是主要因素。外部环境是以某种形式传递自己的外部信息集，它代表外部信息的来源；学习是处理外部信息到知识的过程，首先获取外部环境的信息，然后将信息处理为知识，并将这些知识放入库；库存储了许多指导实施行动的一般原则，由于环境为学习系统提供了各种信息，信息的质量直接影响学习实现。

机器学习解决了如何构建能够通过经验自动改进计算机的问题。它是当今发展最快的技术领域之一，处于计算机科学和统计学的交叉点，也是人工智能和数据科学的核心。机器学习的最新进展是由新的学习算法和理论的发展，以及在线数据可用性和低成本计算持续爆炸式的增长推动的。数据密集型机器学习方法的应用场景可以在整个科学、技术和商业中找到，从而在许多行业(包括医疗保健、制造业、教育、金融建模、警务和营销)中产生更多基于证据的决策。

当想要分析的数据集太大（许多单独的数据点），或者太复杂(包含大量的特征)，或是需要将数据分析过程自动化时建立一个可重复和时间效率高的流程时，机器学习特别有用。来自生物实验的数据经常拥有这些属性；生物数据不断扩大的规模和固有的复杂性鼓励了机器学习在生物学中的越来越多的应用，以建立潜在生物过程的信息和预测模型。

对此，我们开发了特异调控元件鉴定元件（CREfinder）。该软件以完成差异分析后得到的目标细胞类型的上调基因和上调开放区域作为输入，通过软件内部构建的机器学习算法对两者进行特征筛选，从而筛选出目标细胞类型特异的调控元件。通过使用该软件，研究者能够根据实际需求定制化地筛选和鉴定某一种或某几种目标细胞类型的特异调控元件，从而用于下游的特异调控元件应用研究。

## 1.3 参考文献

1. Wittkopp, P., Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 13, 59–69 (2012).
2. Kalay, G. & Wittkopp, P. J. Nomadic enhancers: tissue-specific cis-regulatory elements of yellow have divergent genomic positions among Drosophila species. PLoS Genet. 6, e1001222 (2010).
3. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134, 25–36 (2008).
4. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Rev. Genet. 12, 283–293 (2011).
5. Levine, M. Transcriptional enhancers in animal development and evolution. Curr. Biol. 20, R754–R763 (2010).
6. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers.Cell. 144, 327–339 (2011).
7. Cowles, C. R., Hirschhorn, J. N., Altshuler, D. & Lander, E. S. Detection of regulatory variation in mouse genes. Nature Genet. 32, 432–437 (2002).
8. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. Nature 430, 85–88 (2004).
9. Tirosh, I., Reikhav, S., Levy, A. A. & Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. Science 324, 659–662 (2009).
10. Roy, S. et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787–1797 (2010).
11. M. I. Jordan, T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science 349, 255-260(2015).
12. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2011).
13. Decatur S., Goldreich O., Ron D., Computational sample complexity. SIAM J. Comput. 29, 854–879 (2000).
14. Schmidhuber J., Deep learning in neural networks: An overview. Neural Netw. 61, 85–117 (2015).
15. Krizhevsky A., Sutskever I., Hinton G., Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105 (2015).
16. Blum A., Ligett K., Roth A., A learning theory approach to non-interactive database privacy. J. ACM 20, (2013).
17. Duchi J., Jordan M. I., Wainwright J., Privacy aware learning. J. ACM 61, 1–57 (2014).
18. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: toxicity prediction using deep learning. Front. Environ. Sci. 3, 80 (2016).
19. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl Acad. Sci. USA 117, 1496–1503 (2020).

# 2. 总体设计

## 2.1 目标

从差异分析得到的目标细胞类型上调基因和上调开放区域中，通过机器学习进行特征筛选，高效和高精度地筛选和鉴定目标细胞类型的特异调控元件。

## 2.2 功能

1. **鉴定目标细胞类型的特异标记基因**

从差异分析得到的目标细胞类型上调基因列表中，鉴定目标细胞特异高表达的标记基因，这些特异标记基因可用于单细胞转录组分析中的细胞类型注释。

1. **鉴定目标细胞类型的特异调控元件**

通过软件内置的机器学习算法对差异分析得到的目标细胞类型上调基因和上调开放区域进行进一步的特征筛选，并结合开放区域注释信息进行数据整合，从而鉴定目标细胞类型的特异调控元件。

## 2.3 使用方法

使用该程序的时候，需要按要求准备好输入文件，并写入到程序运行配置文件中。其中上调基因的特征筛选步骤需要准备4个输入文件，上调开放区域的特征筛选步骤需要准备4个输入文件，数据整合环节需要准备2个输入文件。下面将对各步骤所用到的输入文件格式依次做介绍。

**2.3.1 上调基因的特征筛选**

1. **输入文件1**

输入文件1是上调基因特征筛选的目标细胞类型列表。该文件中的细胞类型列表需要和差异分析中上调基因来源的细胞类型保持一致。该文件中的每一行为目标细胞类型中的一种类型，细胞类型的名称需与输入文件3中的细胞类型名称保持一致。文件格式如下图所示。

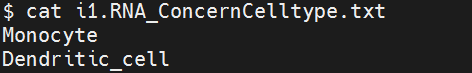


图 1 上调基因特征筛选的目标细胞类型列表

1. **输入文件2**

输入文件2是目标细胞类型的上调基因列表。该文件用于基因特征筛选来鉴定特异标记基因。该文件共包含2列，首行为列名，必须按照下图中分别命名为“id”和“celltype”。第一列为上调基因的基因id，第二列为上调基因来源的细胞类型，列之间以制表符进行分隔。文件格式如下图所示。

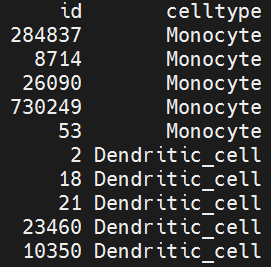


图 2 上调基因特征筛选的目标细胞类型上调基因列表

1. **输入文件3**

输入文件3是RNA-Seq样本的原始信息表。该文件用于生成样本再平衡的表达TPM矩阵和上调基因特征筛选。该文件共包含3列，首行为列名，必须按照下图中分别命名为“srr”、“celltype”和“celltype.uniq”。其中第一列为样本编号，与输入文件4的列名一致；第二列为样本的细胞类型名称；第三列是唯一标识的细胞类型名称，相同细胞类型以\*.1、\*.2等数字序号进行区分。文件格式如下图所示。

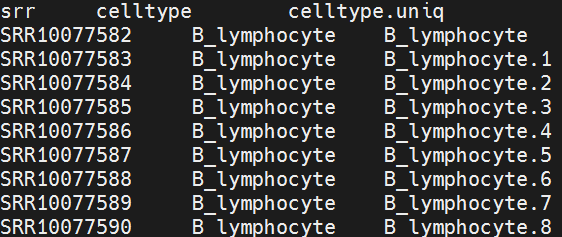


图 3 上调基因特征筛选的RNA-Seq样本原始信息表

1. **输入文件4**

输入文件4是基因表达的TPM矩阵。该文件格式必须包含行名和列名。行名为基因id和symbol以“|”进行连接。列名为样本的编号，与输入文件3中的第一列一致。列之间以制表符进行分隔。文件格式如下图所示。

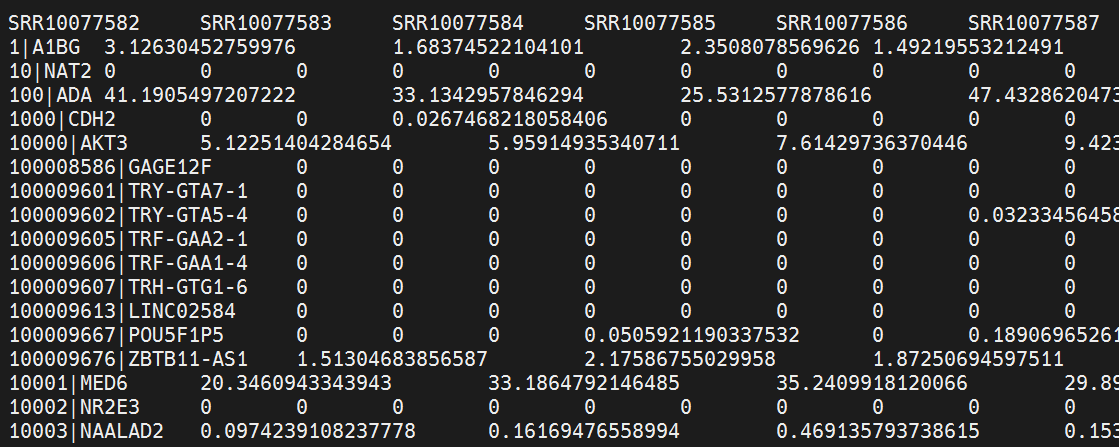


图 4 上调基因特征筛选的基因表达TPM矩阵

**2.3.2 上调开放区域的特征筛选**

1. **输入文件1**

输入文件1是上调开放区域特征筛选的目标细胞类型列表。该文件中的细胞类型列表需要和差异分析中上调开放区域来源的细胞类型保持一致。该文件中的每一行为目标细胞类型中的一种类型，细胞类型的名称需与2.3.1中的输入文件1保持一致。

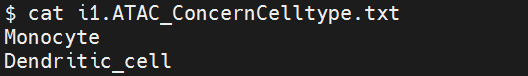


图 5 上调开放区域特征筛选的目标细胞类型列表

1. **输入文件2**

输入文件2是目标细胞类型的上调开放区域列表。该文件用于开放区域特征筛选来鉴定特异开放区域。该文件共包含2列，首行为列名，必须按照下图中分别命名为“id”和“celltype”。第一列为上调开放区域的id，命名方式可参考下图（开放区域所在染色体名称\_开放区域起始位置\_开放区域终止位置）。第二列为上调开放区域来源的细胞类型，列之间以制表符进行分隔。文件格式如下图所示。

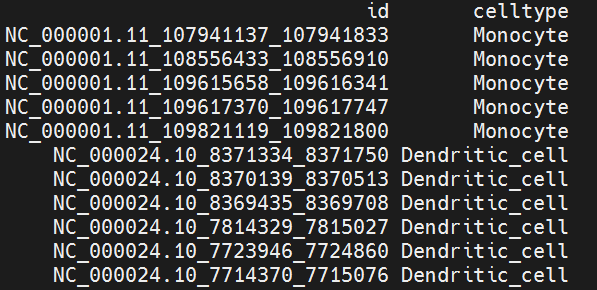


图 6 上调开放区域特征筛选的目标细胞类型上调开放区域列表

1. **输入文件3**

输入文件3是ATAC-Seq样本的原始信息表。该文件用于生成样本再平衡的CPM矩阵和上调开放区域特征筛选。该文件共包含3列，首行为列名，必须按照下图中分别命名为“srr”、“celltype”和“celltype.uniq”。其中第一列为样本编号，与下面的输入文件4的列名一致；第二列为样本的细胞类型名称；第三列是唯一标识的细胞类型名称，相同细胞类型以\*.1、\*.2等数字序号进行区分。文件格式如下图所示。

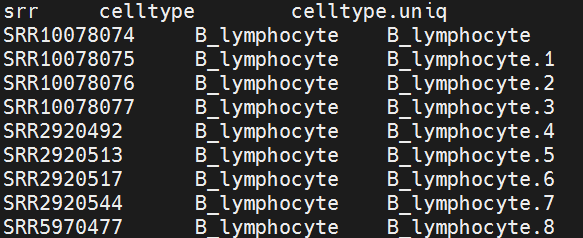


图 7 上调开放区域特征筛选的ATAC-Seq样本原始信息表

1. **输入文件4**

输入文件4是开放区域的CPM矩阵。该文件格式必须包含行名和列名。行名为开放区域id。列名为样本的编号，与输入文件3中的第一列一致。列之间以制表符进行分隔。文件格式如下图所示。

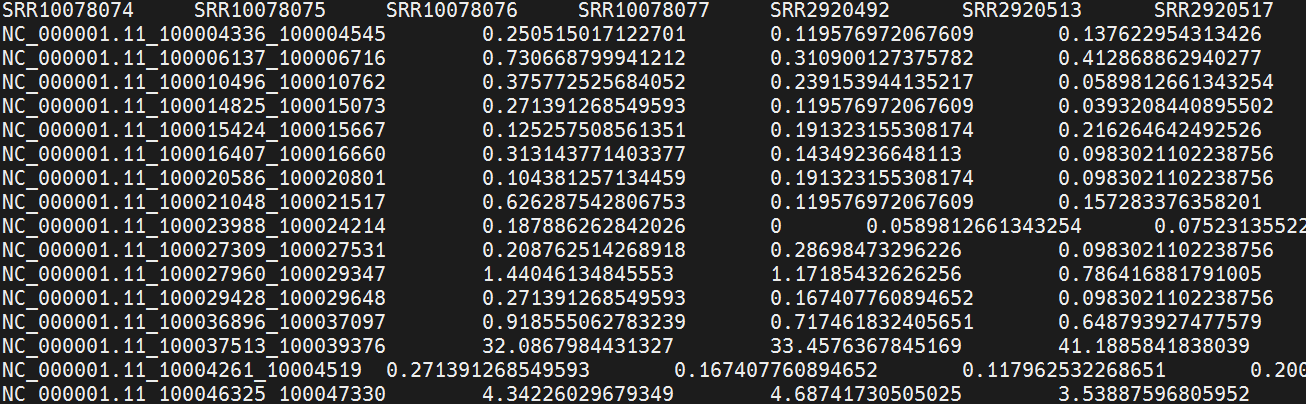


图 8 上调开放区域特征筛选的CPM矩阵

**2.3.3 数据整合**

1. **输入文件1**

输入文件1是上调开放区域的基因注释信息。该文件用于整合特征筛选后的特异开放区域和特异标记基因。该文件格式共包含3列，首行为列名，列名必须按下图分别命名为“peakId”、“geneSymbol”和“geneId”。其中第一列为上调开放区域的id，第二列为上调开放区域调控的基因名称，第三列为上调开放区域调控的基因id。文件格式如下图所示。

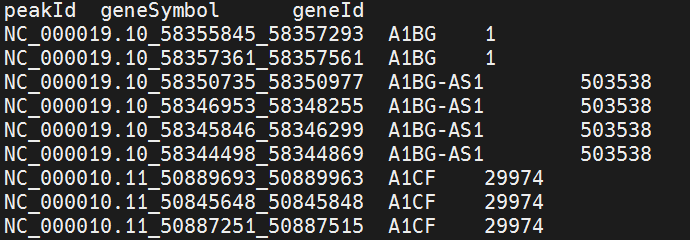


图 9 上调开放区域的基因注释信息

1. **输入文件2**

输入文件2是目标细胞类型的上调基因和上调开放区域特征筛选结果文件路径。该文件格式共包含3列，首行为列名。列名必须按下图分别命名为“celltype”、“RNA\_AUC\_Path”和“ATAC\_AUC\_Path”。从第二行开始，每一行的3列分别代表目标细胞类型、上调基因特征筛选的结果文件路径和上调开放区域特征筛选的结果文件路径。文件格式如下图所示。

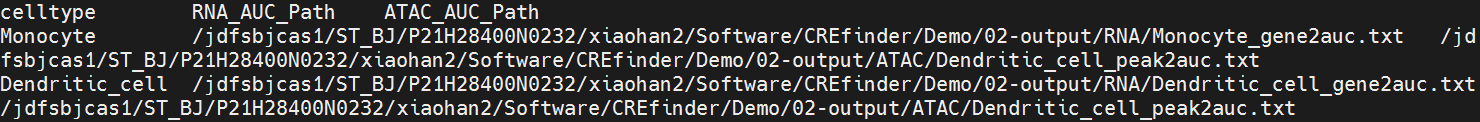


图 10 上调基因和上调开放区域特征筛选结果文件路径

1. **程序运行配置文件**

上述输入文件需配置到程序运行配置文件中。配置文件包括各程序步骤用到的软件、输入参数以及输入文件。具体示例见4.2节。

该程序的总体**运行流程**如下图所示：

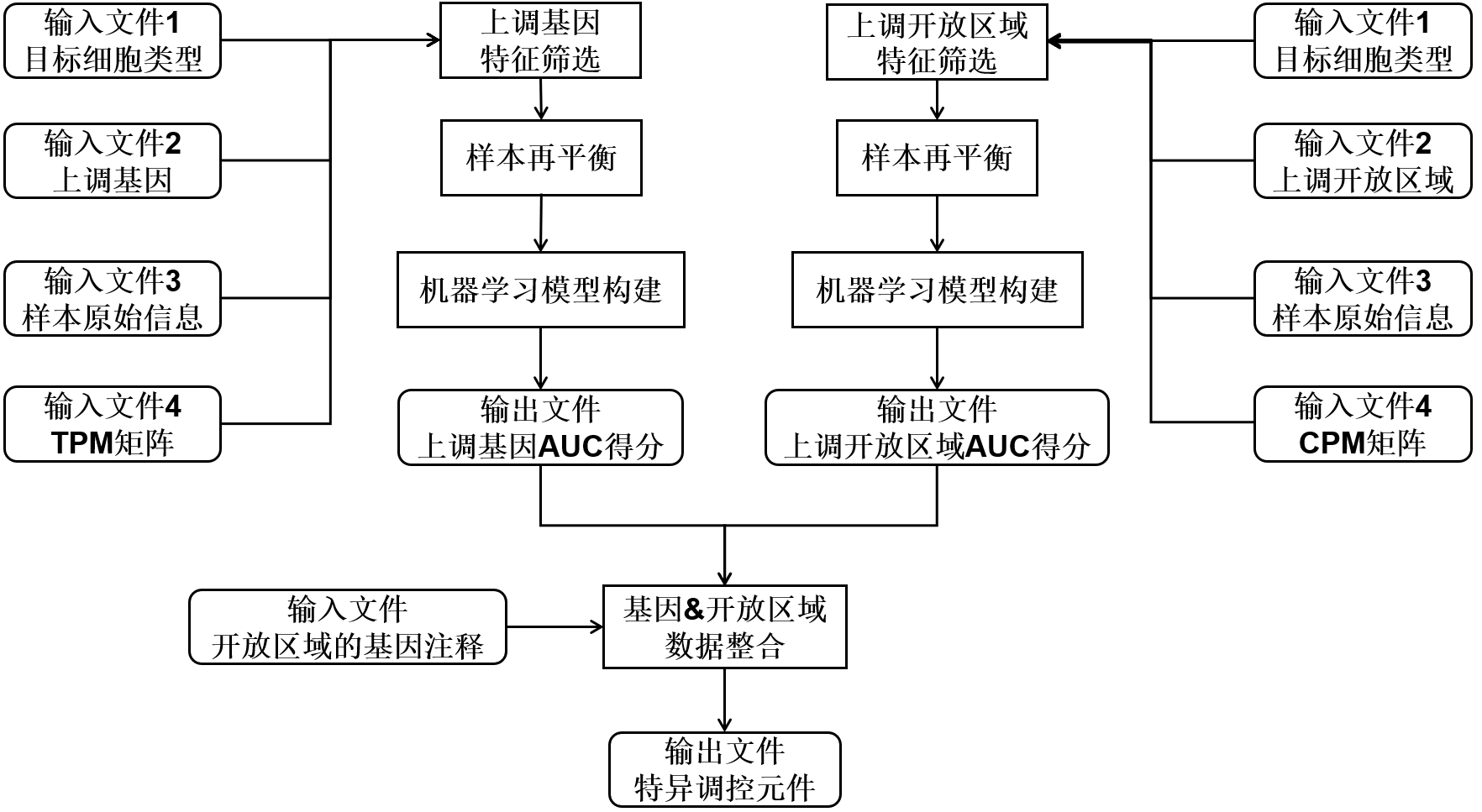


图 11 程序总体运行流程图

# 3. 运行环境

## 3.1 硬件

开发环境：

CPU：Intel（R）Core（TM）i5-7300HQ CPU@2.50GHz 2.50 GHz

RAM：8.00GB

内存：128G固态硬盘

## 3.2 支持软件

Miniconda3 Linux 64-bit

Hisat2 v2.2.1

BWA v0.7.17

BLAST v2.5.0

Perl v5.32.0

# 4. 使用说明

## 4.1 软件安装和初始化

**4.1.1 安装Miniconda3**



图 12 Miniconda3安装步骤

**4.1.2 通过Miniconda3安装相关软件**

1. **激活conda**。激活后命令行左侧会出现base字样，代表已激活conda环境



图 13 conda环境激活

1. **安装分析软件**。conda后接：软件名=版本号，通过这种方式能便捷安装3.2节的所有软件。下图示例的是conda安装R v4.2.2。



图 14 conda安装软件

1. **检查软件安装。**安装完毕后，可通过which加软件名找到和调取软件安装的绝对路径

## 4.2 输入文件

程序运行第一步，上调基因特征筛选和上调开放区域特征筛选的输入文件分别为4个。运行第二步，数据整合的输入文件为2个。每个输入文件具体格式示例如下，格式要求在2.3节已做说明。同时，由于上调基因特征筛选和上调开放区域特征筛选的输入文件格式较为一致，在此作合并介绍，只列出要点：

**4.1.1 上调基因和上调开放区域特征筛选**

1. **输入文件1：**目标细胞类型，列名必须一致，一行一个细胞类型；
2. **输入文件2：**上调基因或上调开放区域列表，列名必须一致，每行一个基因或上调开放区域；
3. **输入文件3：**样本原始信息，列名必须一致，每行一个样本；
4. **输入文件4：**TPM或CPM矩阵，需包含行名和列名。

**4.1.2 数据整合**

1. **输入文件1：**开放区域的基因注释信息，包含开放区域id，基因名称和id；
2. **输入文件2：**目标细胞类型以及4.1.1步骤的两个结果文件路径。

Demo示例程序的输入文件在程序主目录下的/Demo/01-input目录中，其文件目录结果如下图所示。

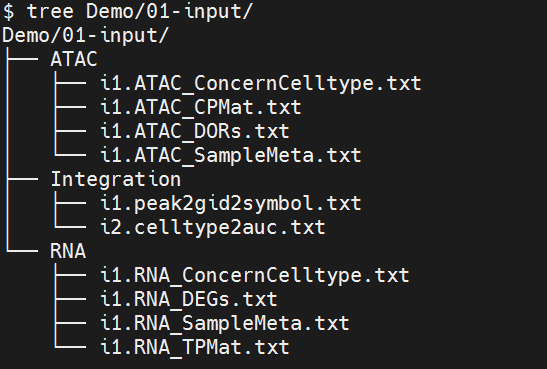


图 15 输入文件目录结构

## 4.3 配置程序

软件安装完成并准备好输入文件后，需要配置程序运行参数文件。程序运行第一步（上调基因特征筛选）的参数配置文件写入demo\_run\_rna.sh，程序运行第二步（上调开放区域特征筛选）的参数配置文件写入demo\_run\_atac.sh，程序运行第三步（数据整合）的参数配置文件写入demo\_run\_integration.sh。程序第一、二步相互独立，因此可并行执行demo\_run\_rna.sh和demo\_run\_atac.sh，成功运行后再执行demo\_run\_integration.sh获取最终结果。由于demo\_run\_rna.sh和demo\_run\_atac.sh参数设置方法较为一致，在此合并作介绍。

1. **demo\_run\_rna.sh和demo\_run\_atac.sh：**

demo\_run\_rna.sh和，demo\_run\_atac.sh文件内容分别如下图所示。其中以“#”开头或包含在“COMMENTS”内部的代码为注释行。其他以“--”开头的行中，每一行对应一个参数变量，参数名称和值之间以空格符进行分隔，“--”开头的变量为程序识别的参数名称，可根据需要设置对应参数；空格后面为变量所对应的值（运行模式、参数、文件路径等）。

“--Mode”指定程序运行模式，后接“RNA”为上调基因特征筛选模式，后接“ATAC”为上调开放区域特征筛选模式，后接“Integration”为数据整合模式；

“--ExpMatrix”指定CPM或TPM矩阵的文件路径；

“--SampleMeta”指定样本原始信息的文件路径；

“--Features”指定上调基因或上调开放区域列表；

“--TargetCelltype”指定目标细胞类型列表；

“--Outdir”指定程序输出目录。

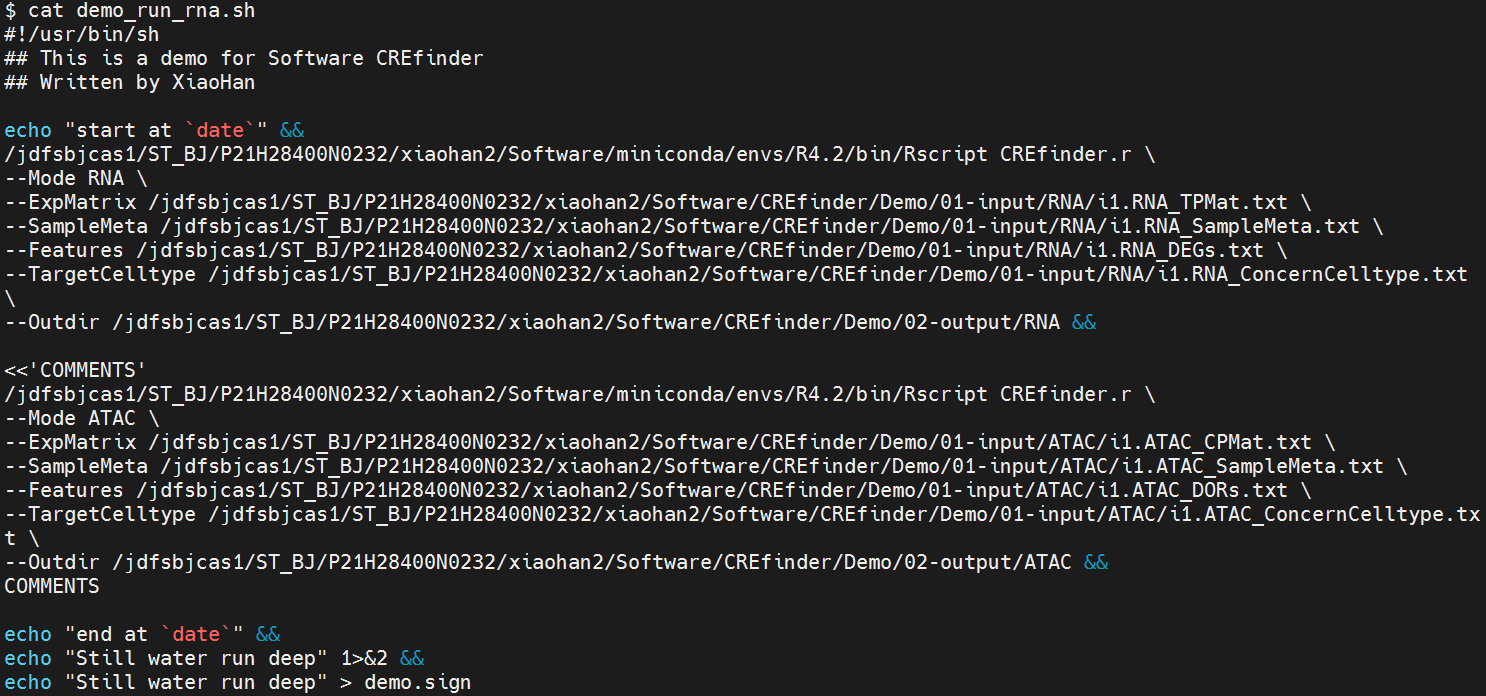


图 16 demo\_run\_rna.sh输入参数

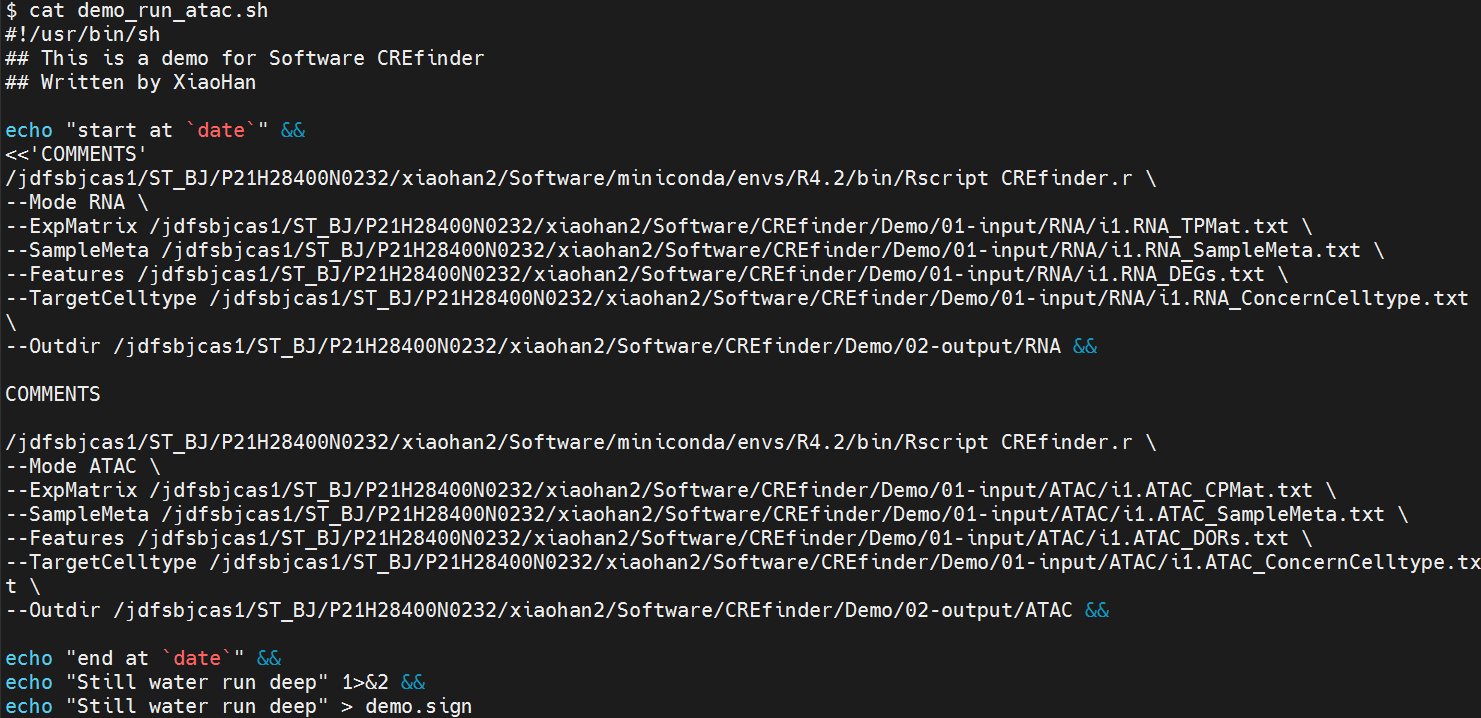


图 17 demo\_run\_atac.sh输入参数

1. **demo\_run\_integration.sh:**

文件内容示例下图所示。通过运行该脚本整合程序第一、二步运行产生的结果文件，获取最终的特异调控元件列表。

“--Mode”指定程序运行模式为Integration模式；

“--PeakGene”指定开放区域的基因注释文件路径；

“--AUCtbl”指定程序第一、二步运行产生的结果文件路径；

“--Outdir”指定程序输出目录。

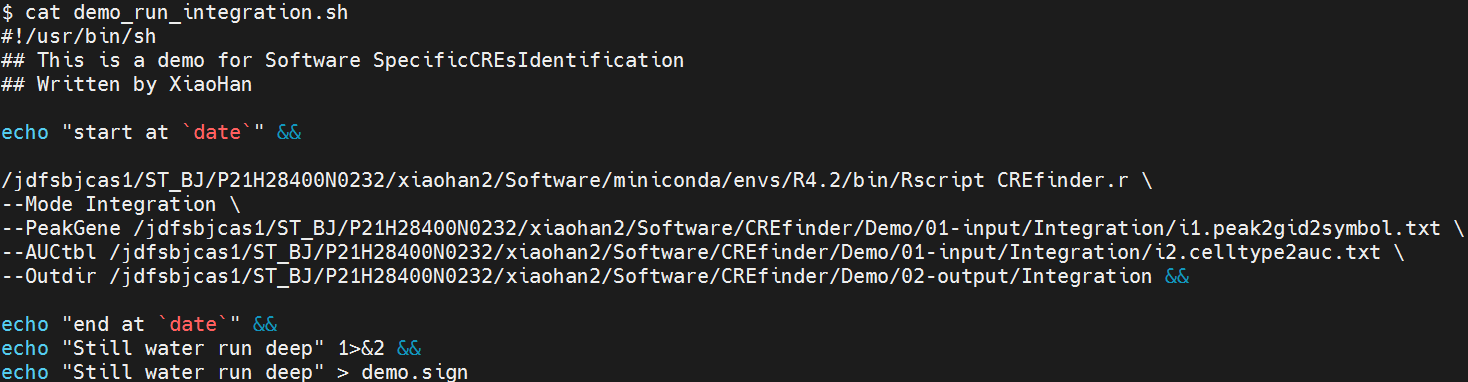


图 18 demo\_run\_integration.sh输入参数

**注意：在整个过程中最好均使用绝对路径！！！**

## 4.4输出结果

在成功运行上述三个配置程序后，在输出主目录下会产生3个子目录。“RNA”目录用于存放上调基因特征筛选的结果文件，“ATAC”目录用于存放上调开放区域特征筛选的结果文件，“Integration”目录用于存放数据整合的结果文件。结果文件目录结构如下图所示。**下面以Demo示例数据为例，解释结果文件。**其程序运行脚本存放在程序主目录下demo\_run\*.sh格式的脚本文件中。Demo示例数据的输入文件存放在程序主目录下的Demo/01-input子目录中，输出文件存放在程序主目录下的Demo/02-output子目录中。

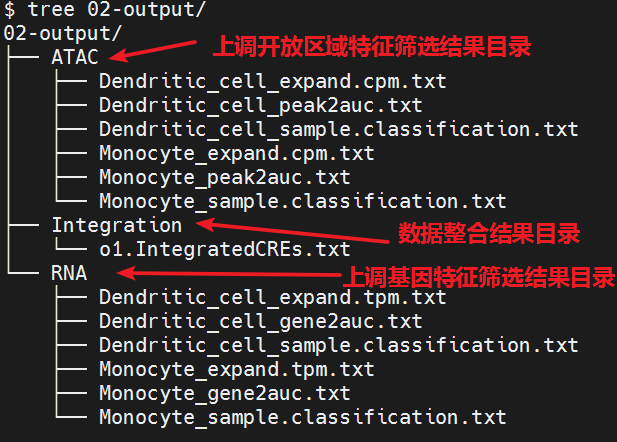


图 19 输出结果目录

**4.4.1 上调基因特征筛选的结果说明**

结果文件存放在02-output/RNA目录下，结果文件目录结构如下图所示。箭头所指的\*\_gene2auc.txt文件是两个目标细胞类型的上调基因特征筛选的结果文件，存储了上调基因的AUC得分，也是程序下一步进行数据整合的输入文件。\*\_expand.tpm.txt是两个目标细胞类型样本再平衡后的TPM矩阵。\*\_sample.classification.txt是两个目标细胞类型样本再平衡后的样本信息。

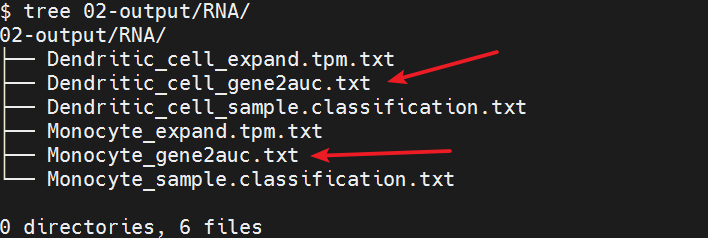


图 20 上调基因特征筛选的结果文件目录结构

**4.3.2 上调开放区域特征筛选的结果说明**

结果文件存放在02-output/ATAC目录下，结果文件目录结构如下图所示。箭头所指的\*\_peak2auc.txt文件是两个目标细胞类型的上调开放区域特征筛选的结果文件，存储了上调开放区域的AUC得分，也是程序下一步进行数据整合的输入文件。\*\_expand.cpm.txt是两个目标细胞类型样本再平衡后的CPM矩阵。\*\_sample.classification.txt是两个目标细胞类型样本再平衡后的样本信息。

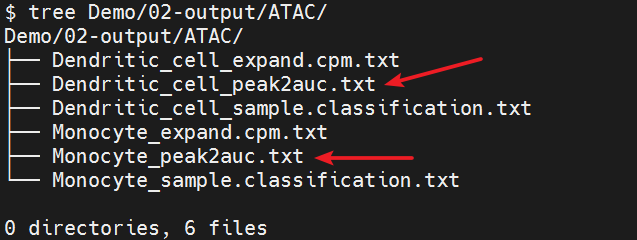


图 21 上调开放区域特征筛选的结果文件目录结构

**4.3.3 数据整合的结果说明**

结果文件存放在02-output/ATAC目录下，结果文件o1.IntegratedCREs.txt的内容如下图所示。该文件包含以制表符分隔的7列信息，记录了经过特征筛选被保留下来的目标细胞类型特异调控元件，其中每一行代表一个调控元件。在每一行中，第一列是调控元件对应上调开放区域的id，第二、三列分别是调控元件调控的基因id和基因名称，从第四列开始，每两列记录该调控元件在目标细胞类型的上调基因和上调开放区域AUC得分情况。由于经过特征筛选，因此该文件中AUC得分均大于或等于0.85。该值越大表明该调控元件的特异性越强。

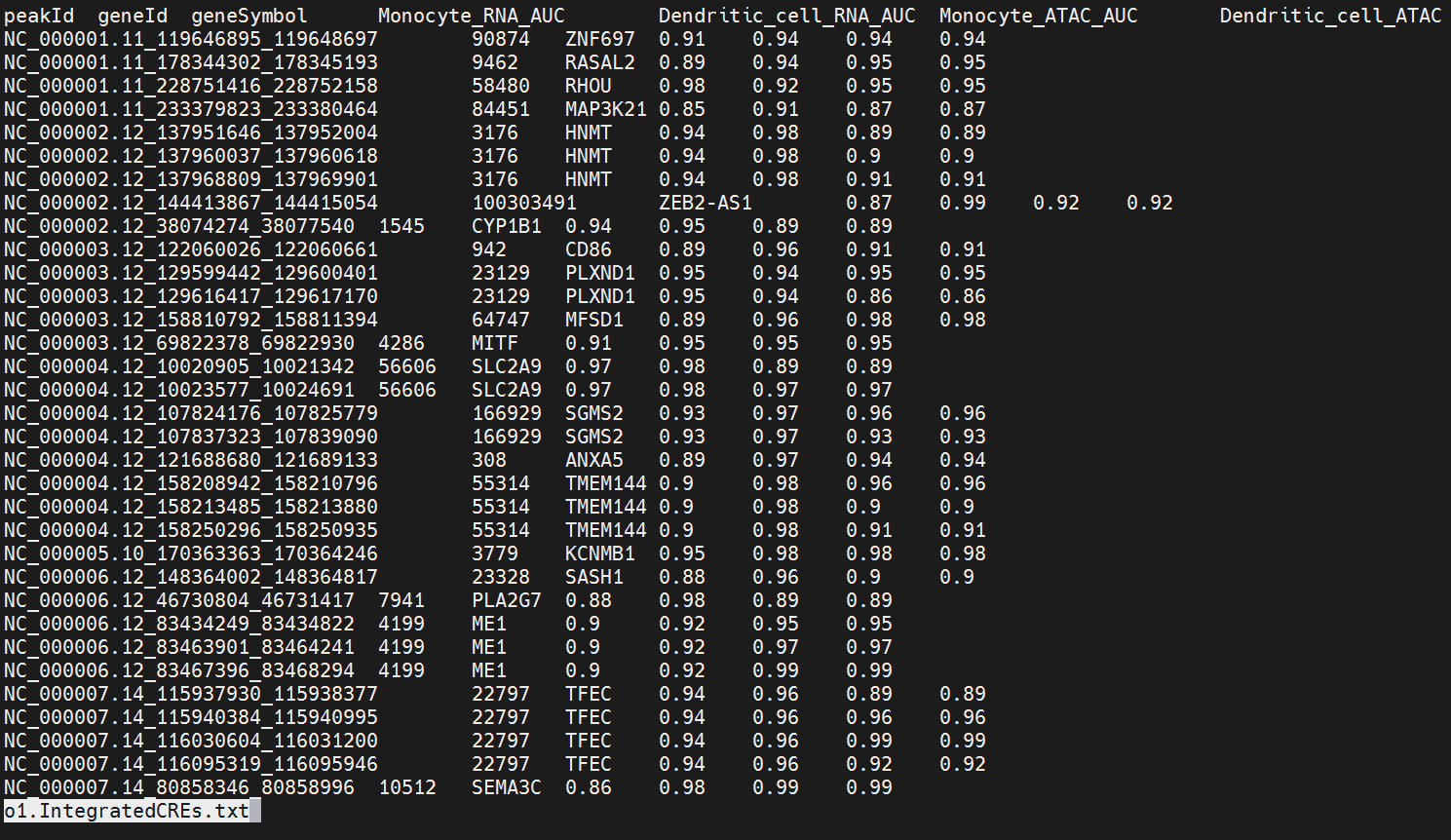


图 22特异调控元件的结果文件格式

## 4.5求助查询

1. 当不清楚主程序输入参数的时候，可以通过输入以下命令查看帮助文档（在脚本后面添加-h参数）：

Rscript /jdfsbjcas1/ST\_BJ/P21H28400N0232/xiaohan2/Software/CREfinder/CREfinder.r --help

具体参数说明如下图16所示：

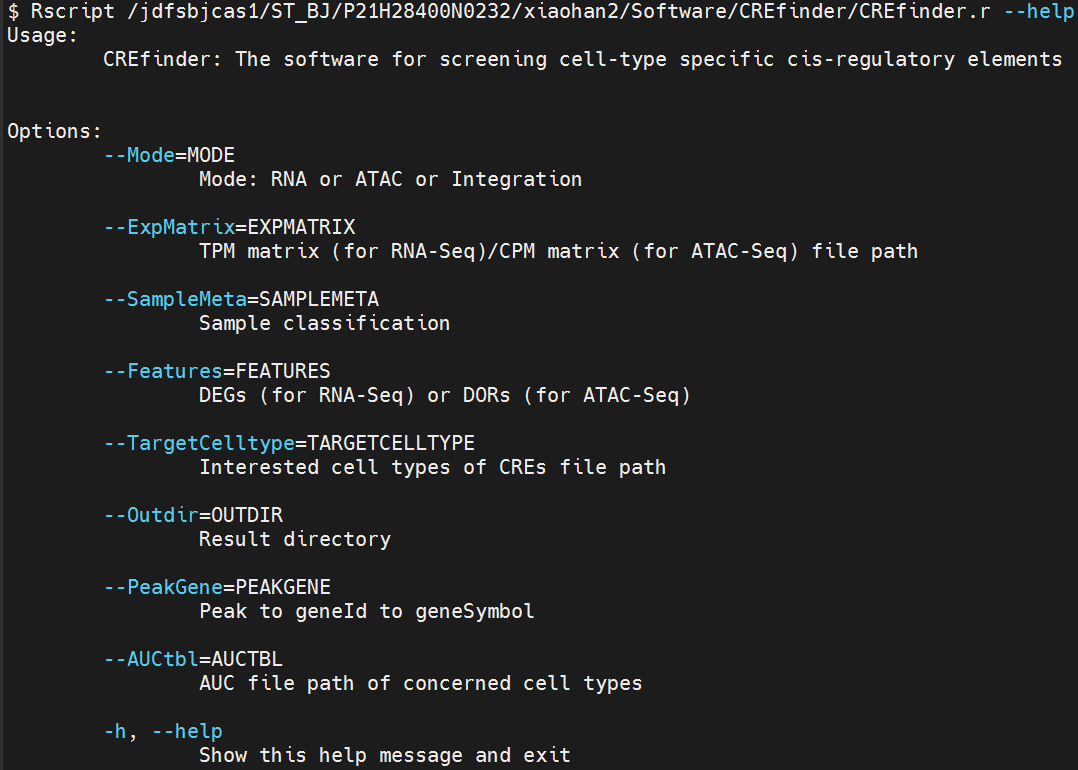


图 23 主程序参数说明

1. 当不清楚程序使用到的软件参数含义的时候，可以通过输入软件的绝对路径加—help参数查看软件说明文档：

例如，查看Miniconda3软件的参数说明：

/jdfsbjcas1/ST\_BJ/P21H28400N0232/xiaohan2/Software/miniconda/bin/conda --help

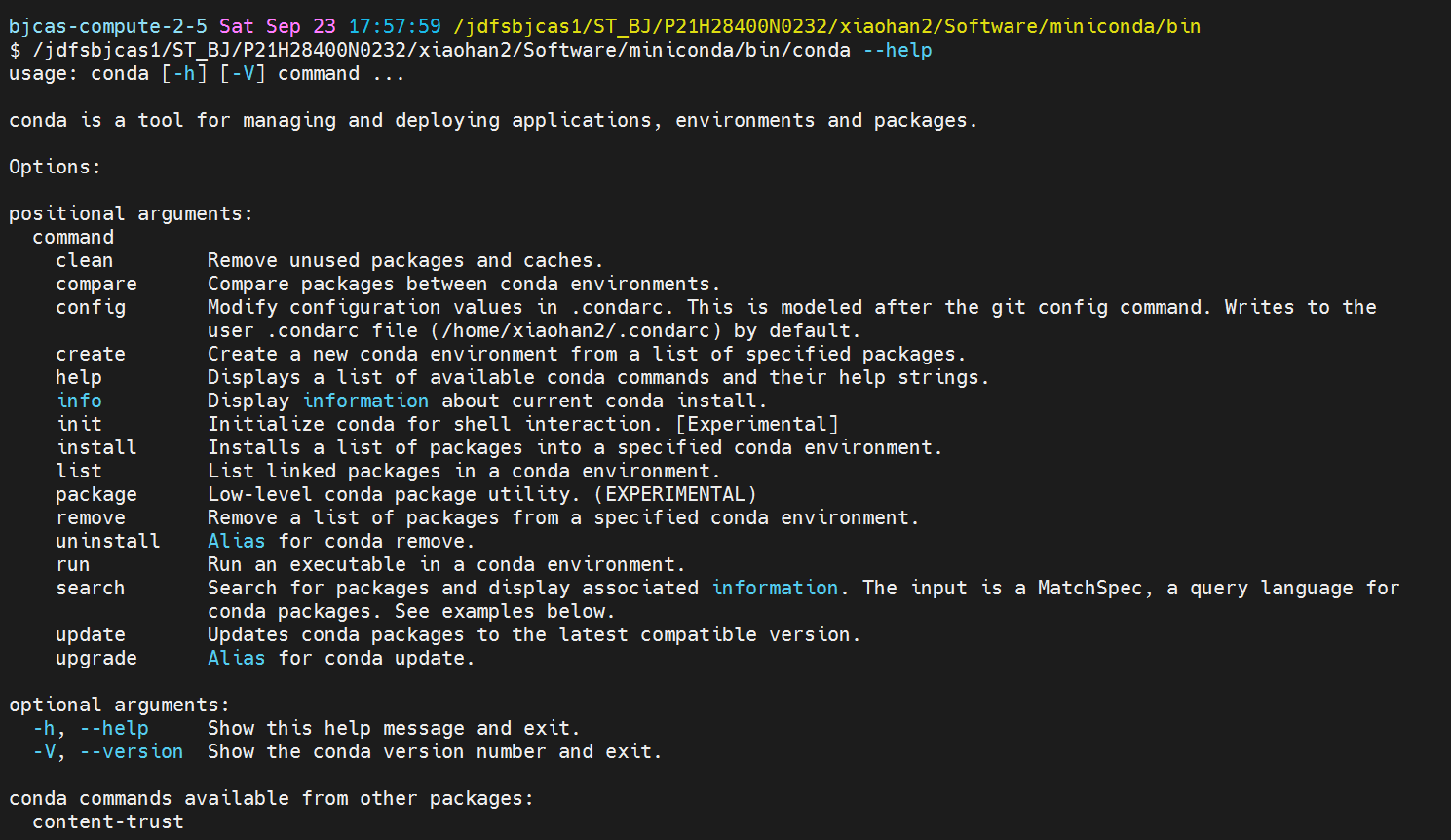


图 24 调取Miniconda3参数说明

# 5. 运行说明

## 5.1程序存放位置

**脚本存放路径：**

/jdfsbjcas1/ST\_BJ/P21H28400N0232/xiaohan2/Software/CREfinder目录下的可执行脚本文件。

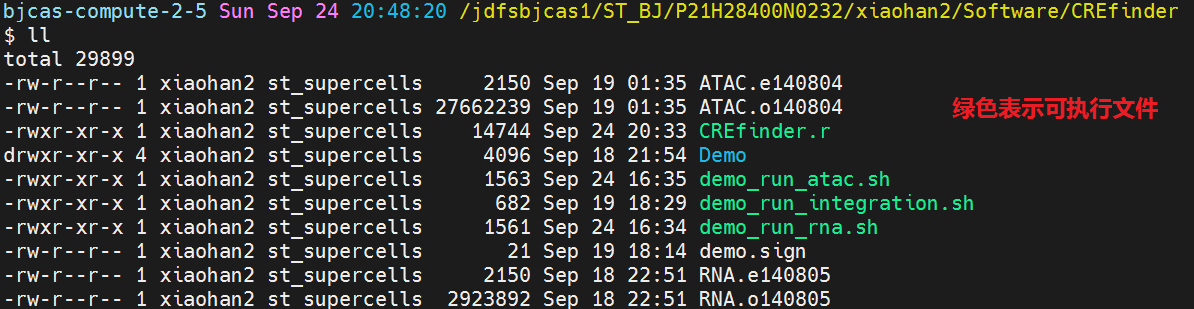


图 25 程序脚本

## 5.2运行步骤

在此以一个**Demo数据作为示例**，展示使用程序筛选单核细胞（Monocyte）和树突状细胞（Dentritic\_cell）共有的特异调控元件。

1. **准备好各步骤的输入文件，放在Demo/01-input目录下**

**示例输入文件主目录路径：**

/jdfsbjcas1/ST\_BJ/P21H28400N0232/xiaohan2/Software/CREfinder/Demo/01-input



图 26 Demo示例数据的输入文件目录结构

1. **修改配置demo\_run\_rna.sh、demo\_run\_atac.sh和demo\_run\_integration.sh文件内容，放在程序主目录下**

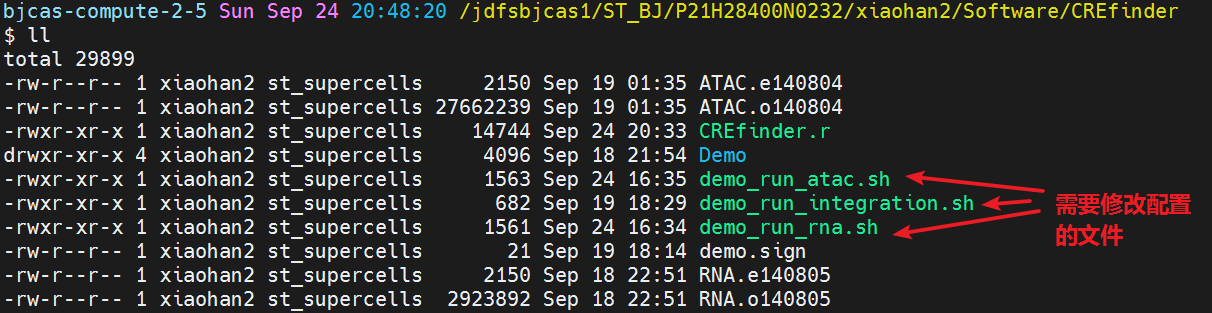


图 27 需要修改的配置文件

1. **首先分别运行demo\_run\_rna.sh和demo\_run\_atac.sh生成对应的输出文件主目录**

**上调基因特征筛选的示例输出文件主目录路径**：/jdfsbjcas1/ST\_BJ/P21H28400N0232/xiaohan2/Software/CREfinder/Demo/02-output/RNA

**上调开放区域特征筛选的示例输出文件主目录路径**：/jdfsbjcas1/ST\_BJ/P21H28400N0232/xiaohan2/Software/CREfinder/Demo/02-output/ATAC/

1. **第三步程序成功运行结束后，运行demo\_run\_integration.sh，生成对应的输出文件主目录**

**数据整合后的目标细胞类型特异调控元件的示例输出文件路径：**

/jdfsbjcas1/ST\_BJ/P21H28400N0232/xiaohan2/Software/CREfinder/Demo/02-output/Integration/o1.IntegratedCREs.txt

1. **随后可通过查看输出目录中的如下三个文件查看运行结果：**

其中

\*.sh.o 该文件中保存了一些输出到屏幕的信息，例如运行的起始和终止时间；

\*.sh.e 该文件中保存报错信息，如果该程序报错，则会将报错信息输入到该文件中，最后可以通过查询该文件进行问题溯源，从而解决问题。

demo.sign 当程序正确且成功运行后，会生成demo.sign文件，以提示用户程序已经成功运行完成，用户不需要通过其他方式对程序进行监控。

## 5.3运行时间

可以通过cat \*.sh.o可以通过键入cat \*.sh.o查看文件，其中第一行是起始时间，最后一行是程序运行完后的时间。结果如下图所示：

以Demo示例数据为例，该数据的RNA-Seq样本数为368，ATAC-Seq样本数为298，上调基因数为1783，上调开放区域数为9089。在上调基因特征筛选中，程序运行时间约为40分钟；在上调开放区域特征筛选中，程序运行时间约为3个小时；数据整合的程序运行时间只需19s由于上调基因和上调开放区域的特征筛选可并行处理，因此程序运行耗时主要在上调开放区域特征筛选中，即整个程序总运行时长大约为3个小时。若用户输入的RNA-Seq和ATAC-Seq样本量进一步增大，或上调基因和上调开放区域数目的进一步增多，则总时长将进一步增大，总体上取决于服务器运算速度和输入数据总量的大小。



图 28 \*.sh.o示例图

## 5.4运行错误说明

当运行程序时，会生成两个文件\*.sh.o和\*.sh.e，其中错误输出会保存在文件\*.sh.e中。也就是说，当程序出错的时候会将报错的信息保存到\*.sh.e中，我们可以通过查看\*.sh.e文件来检查程序是否出错，如果出错也可以从该文件中得到报错信息，进行问题溯源，从而修改输入文件或者程序来实现成功运行。

可以通过键入cat \*.sh.e查看文件，结果如下图所示：

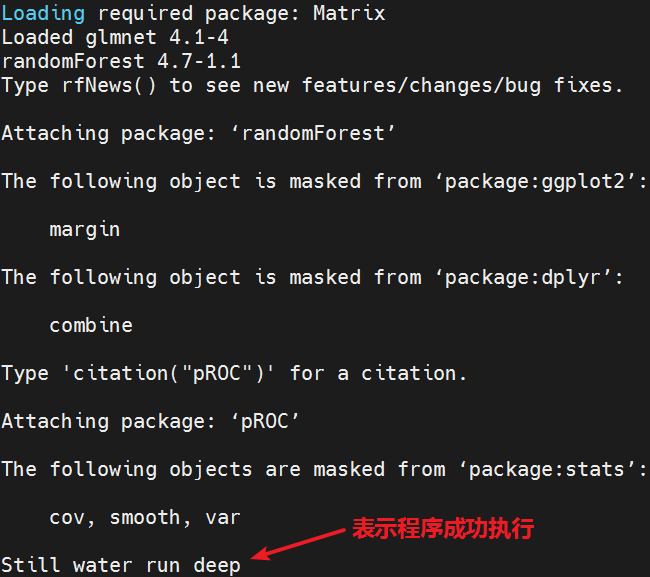


图 29 \*.sh.e示例图

其中如果出现“Attaching package”是将R脚本调用包产生的，可忽略。

当该文件中出现“Still\_waters\_run\_deep”则表示该程序运行成功且运行完成，反之如果出现了其他字样可能是报错信息，需要仔细阅读，找到错误原因并修正。修正后可以重新运行。

## 5.5运行成功说明

当程序成功完成后会生成另一个新的文件demo.sign。该文件的生成表明程序的顺利执行，并且已经完成所有运行。用户可直接查看结果内容。无需通过其他操作对程序进行监控。

可以通过键入cat demo.sign查看文件，结果如下图所示：



图 30 demo.sign示例图

该文件中只有一行内容“Still\_waters\_run\_deep”，表示程序运行完成并且成功运行。