# 使用表观遗传学标记进行DNA序列分类

## 表观遗传学标记介绍

表观遗传标记主要包括DNA甲基化、RNA干扰、组蛋白修饰等。与经典遗传学以研究基因序列影响生物学功能为核心相比，表观遗传学主要研究这些“表观遗传现象”建立和维持的机制。

## 表观遗传学数据集的下载

在我的研究中主要从ENCODE项目下载数据，有多种方法可以下载数据，可以参考该文章（<https://www.plob.org/article/9514.html>）。

我通常使用文章中介绍的第一种方式。打开如下网址：<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/>

打开网址后可以看到其中包含了各种表观遗传学标记信息，我们通常更关心Dnase数据和Histone（组蛋白标记数据）



以组蛋白标记为例，进入下级目录后可以看到各个细胞系的组蛋白标记。

这里我们根据我们所需要的组蛋白标记类别和细胞系下载bigwig文件

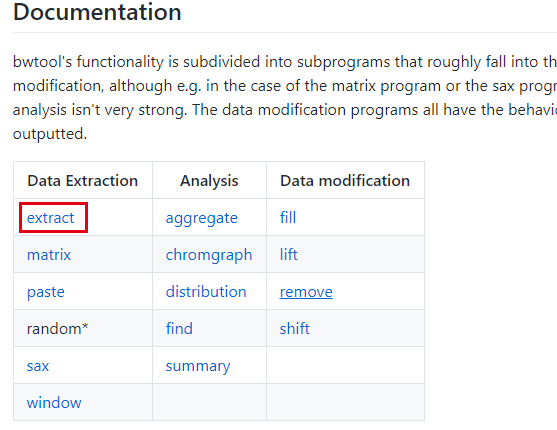
（这里的各种文件格式所包含意义可自行查阅）

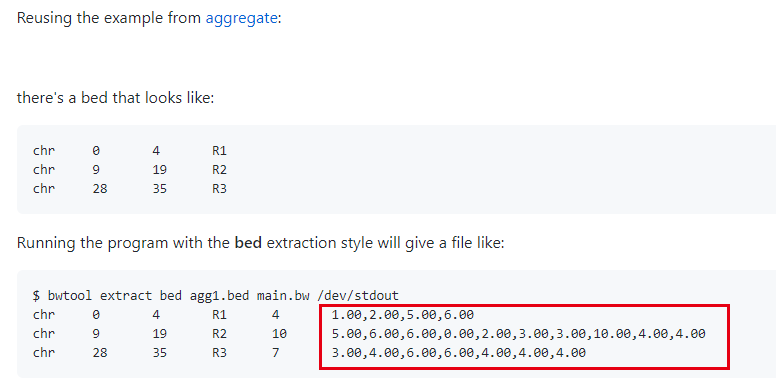
## Bigwig文件使用

我们可以使用bwtool（https://github.com/CRG-Barcelona/bwtool/wiki）软件来处理，BigWig文件是一种全基因组计算（例如GC含量）或实验（例如ChIP-seq/RNA-seq片段深度）信号数据的压缩的，索引的，二进制格式。bwtool是一个旨在快速且高效地读取bigWig文件的工具，提供了以几种方式，从全局或在特定区域提取数据并进行总结的功能。此外，该工具使得信号数据位置从一个基因组装配到另一个的转换，也称为“lifting”成为可能。我们相信bwtool对于经常处理bigWig数据的分析者是非常有用的，bigWig正在成为沿基因组展示功能信号的一种标准格式。

Bwtool的使用可以参考<https://www.jianshu.com/p/032ce018b84d>。

我们通常使用bwtool的extract功能将数据集的bed文件和bigwig文件进行映射。用法可以参考官网说明，写的很详细<https://github.com/CRG-Barcelona/bwtool/wiki/extract>。





我们通过该软件可以将我们数据集中的每一条数据就会一个一维的向量，每种表观遗传学标记都可以使一个DNA序列得到一个一维向量，我们可以通过对这些向量进行统计特征提取，例如求平均值，中位数，峰值，偏度，众数，最大值最小值，等，也可以将多个遗传学标记得到的向量进行堆叠，得到（N\*M）的矩阵，N为DNA序列的长度，M为使用的表观遗传学标记数量。得到的这个矩阵可以使用CNN进行处理。

## 具体操作步骤

这里我给出一个具体操作的示例，仅供参考

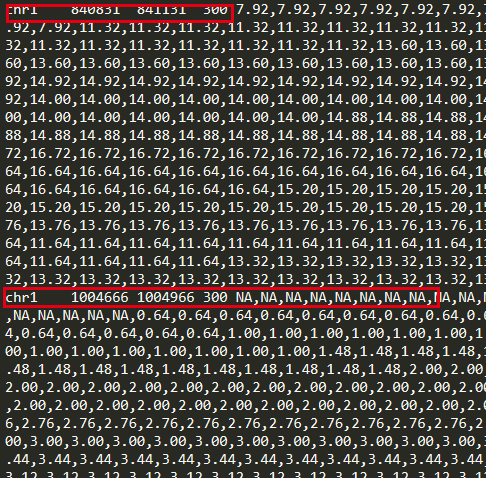
（ubuntu系统）

第一步：

Example文件夹中包含bigwig用于存放我们下载的bigwig文件（可以放多个进行批处理）

运行gfbw.py程序 ，gfbw.py程序需要提前安装好bwtool软件，实际上gfbw.py就是一个执行bwtool软件的一个脚本。（需要在代码里修改beb文件的名称）

得到的文件需要去掉前四列信息，以及对NA值进行0填充，这里可以手动处理，并且可以对这些数值进行归一化操作。



Scaler.py是一个归一化操作的程序，仅供参考

contact\_hms.py是将多个表观遗传学标记得到的向量进行合并成矩阵的程序，仅供参考使用

hms\_feature\_mean\_big.py是一个提取向量统计学特征的代码，主要提取了最大值最小值，仅供参考使用。