# Work flow

# 数据收集和处理

1. 打开ENCODE 网站(https://www.encodeproject.org/), 选择DNA binding, 选择Filtered Data Matrix
2. 在Assay title选择TF CHIP-seq，在Organism选择Homo sapiens，Genome assembly选择GRCh38，Available file types 选择bed narrowPeak, 选择Download下载数据下载地址
3. 对于下载好的文件，使用wget软件下载所有文件，其中第一个文件为meta data，包含了peak文件的实验信息，其他文件为peak（bed narrowPeak）文件
4. 为了保证一个同一个实验只取一个peak文件作为结果，使用R语言过滤得到的meta data文件，具体来说：
   1. 过滤得到所有out type为IDR thresholded peaks的文件
   2. 对于剩下的文件，挑选含有Biological replicate最多的文件
   3. 对于剩下的文件，挑选含有peak数量最多的文件
5. 最后得到2762个实验的唯一peak文件
6. 对于所有的peak文件，使用awk在最后一列添加对应的文件ID，并将所有单个实验的peak文件整合为一个包含所有实验的peak文件

# 数据库后端搭建

根据数据设计ER表格，因为数据简单，所有数据都是依赖于“File\_accession”实验编号的，所以数据表格符合第三范式。

表格设计：

experiment\_meta\_final

File accession VARCHAR 50

File assembly VARCHAR 10

Experiment accession VARCHAR 50

Biosample term name VARCHAR 50

Biosample type VARCHAR 50

Biosample organism VARCHAR 20

Experiment target VARCHAR 20

Experiment date released VARCHAR 15

Project VARCHAR 10

File download URL TEXT

1. 创建表格命令：（本地）

CREATE TABLE `homework1\_test`.`experiment\_meta\_final` (
`File\_accession` VARCHAR(50) ,
`File\_assembly` VARCHAR(10) NOT NULL ,
`Experiment\_accession` VARCHAR(50) NOT NULL ,
`Biosample\_term\_name` VARCHAR(50) NOT NULL ,
`Biosample\_type` VARCHAR(50) NOT NULL ,
`Biosample\_organism` VARCHAR(20) NOT NULL ,
`Experiment\_target` VARCHAR(20) NOT NULL ,
`Experiment\_date\_released` VARCHAR(15) NOT NULL ,
`Project` VARCHAR(10) NOT NULL ,
`File\_download\_URL` TEXT NOT NULL ,
PRIMARY KEY (`File\_accession`)) ENGINE = InnoDB;

1. 通过phpmyadmin批量导入数据。

1. 制作后端：

通过php中PDO类，对数据库进行访问，这种方式可以连接多种数据库，同时使用了“prepare”语句，可以提升应用的性能，让客户端/服务器缓存查询和元信息，还能阻止 SQL 注入攻击。

前端通过Ajax工具，连接到后端服务器，能够在不完全刷新页面的情况下，返回数据，并填入表格中。

1. 前后端连接。前端使用html展示文档信息，使用css添加样式，使用JavaScript对网页响应进行操作，我们通过Boostrap和jQuery框架来进行简化操作。在后端使用php接受用户前端输入的数据，同时使用Ajax异步处理数据，来做到用户友好。

# 前端网页搭建

1. 导入bootstrap、datatable的css以及js文件，并导入jquery3.6.js
2. 创建自己的style.css
3. 根据需求 建立了 Home.html、TF-search.html、Browse.html、Management.html、statistic。其中Home为index页面，TF-search用于搜索，management 用于更新数据。主要就是用了form、nav、ul、li、img、input等标签
4. 主要页面就是Home和TF-search Home是主页 ，TF-search提供搜索内容，输入symbol即可获得信息，management.html实现的内容为提交信息。

# 使用方法

1. **Home界面:**

目前已经完成Home以及TF-search页面，点击上方的导航栏进行跳转；点击未完成的页面后，会在3秒后自动跳转回Home页面； Information模块介绍本数据库的数据来源个基本信息

**已完成部分：**建立项目网站；建立后台数据库，可以利用数据库的phpmyadmin工具进行数据库内容的动态更新； 可以通过html页面对数据库数据进行添加条目的操作；每个页面中都有对本数据库dEMO信息进行了介绍，在页面下方是对当前页面功能介绍。  
对于未完成的前端页面，会自动跳转回home页面。

1. **TF-search页面**：

使用:TF-search页面用于搜索并展示相应的结果。输入需要查询的转录因子名称（TF name），选择想要查询的 物种名(Biosample organism),点击Submit，即可返回所需要的结果;如果不进行输入，默认返回数据库中前10条数据；如果输入的名称在数据库中不存在（例如‘a’），会有提示框。表格大小会根据返回数据量的多少进行变化。

**返回的结果**:返回的数据一共9列，包括 FIle acc: 文件ID、 File assembly：基因组版本、 Experiment acc：实验ID、 Biosample term：生物样本名称、 Biosample type：生物样本类型、 Experiment target：转录因子名称、 Experiment data released：数据发布时间、 Project：项目、 Download URL：数据下载地址。

可以通过点击downlowad进行下载。

1. **Management Data页面：**

**使用**:此页面用于向数据库中插入数据，现支持每次插入一条数据，填写所有表单后，点击“提交插入数据申请”。点击后进入跳转新页面展示php返回信息，默认3秒中后跳转会原始页面。

用户输入数据后，前后端会进行输入检验。前端页面会判断是“text”类型，后端会判断是否输入全部条目。如果条目不全，不会进行插入操作。