

程序使用说明：

一、提取及筛选（GENOME_FILTER.py）

1. 程序作用：

将比对文件按照参数要求隔一段长度截取一段 bp 下来与目标区域被比对基因组使用 BLAST 进行比对，若符合要求则留下。比对完所有截取下的样本片段比对完后进行第二轮比对，即将所有合格样本与非目标区域被比对基因组使用 BLAST 进行比对，提取出不符合第二轮要求的样本。最后从第一轮的结果将不符合要求的样本剔除掉，输出含有探针位置信息、序列信息、在目标区域重复次数信息的文件。

2. 输入参数：

- ①read：将比对基因组文件名称（FASTA 格式）
- ②index：第一轮被比对目标基因组建的库名称
- ③INDEX：第二轮被比对目标基因组建的库名称
- ④fa：中间临时的比对者（-query）文件名称
- ⑤gap：隔多少个 bp 取一段样本（10000 左右时，程序耗时数十至数百秒；50 左右时，程序耗时十几小时）
- ⑥length：一段样本的长度
- ⑦start：目标区域的起始位置
- ⑧end：目标区域的结束位置

⑨output: 第一轮比对后的输出文件名称

⑩OUTPUT: 最终得到的探针文件名称

3. 输出参数:

①OUTPUT: 含有所有探针位置信息、序列信息、目标区域重复次数的.txt 文件

②OUTPUTbp1—4: 提取第二轮比对不合格样本, 含不合格样本位置信息、重复次数信息的中间产物文件

二、建库 (IndexCreator.py)

1. 程序作用:

将目标基因组 (可为多数) 及其目标片段截取下来, 以生成在 Linux 中用 BLAST 软件完成 makeblastdb 操作中的输入文件, 即-in 参数。

2. 输入参数:

①read: 基因组文件名称 (FASTA 格式)

①start: 基因组目标片段的起始位置

②end: 基因组目标片段的结束位置

③num: 基因组为几号染色体

3. 输出参数:

输出截取后的序列信息, 文件格式为 FASTA 格式。