## 

## 快速比对分析软件技术文档

版本号：V1.0

发布时间：2016.04.28

目录

[快速比对分析软件技术文档 1](#_Toc776440266)

[1. 运行说明 4](#_Toc1286711254)

[1.1运行环境 4](#_Toc1426635631)

[1.2硬件环境 4](#_Toc1072512606)

[1.３运行 4](#_Toc271177644)

[1.4输入 4](#_Toc570202357)

[1.2输出 5](#_Toc1964304311)

[2. 技术路线 7](#_Toc877668239)

[2.1 8](#_Toc672843363)

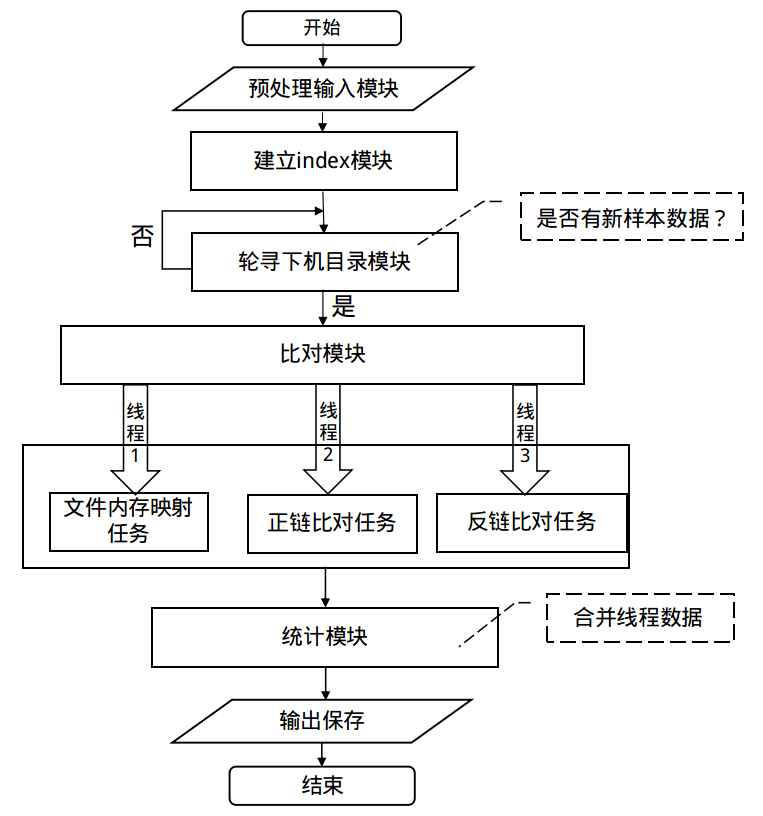
[2.2 构建index库 8](#_Toc987508224)

[2.3 比对 9](#_Toc1356384701)

[2.4生成结果文件 9](#_Toc508009410)

[3.声明 10](#_Toc600641306)

**软件设计流程图附图**



# 运行说明

## 1.1运行环境

* Linux 集群
* 3核以上
* 35G内存以上

## 1.2硬件环境

* Linux 平台下C库
* Gcc编译器
* Gdb调试软件

## 1.３运行

在命令行运行主程序hashtb\_pth即可，它可以实时监控下机数据目录，一旦有数据出现，能够实时处理，并输出结果文件。Shell命令行下输入：

./hashtb\_pth

## 1.4输入

输入文件格式要求：

fastq下机样本文件，如：

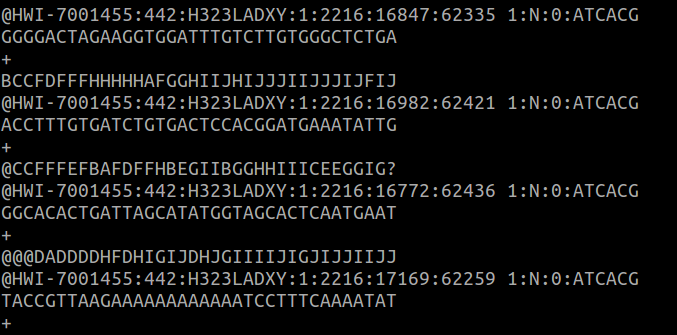


图１

说明：

每4行记录一条reads信息：

第一行：以@开头，标示一条序列，包括测序仪名字，标记位置，单/双端测序以及过滤情况和引物接头等相关信息描述；

第二行：reads序列；

第三行：以‘+’开头，后面是序列标示符、描述信息，或者什么也不加；

第四行：是reads质量信息，和第二行的序列相对应。

## 1.**2输出**

比对结果输出文件内容列表如下：

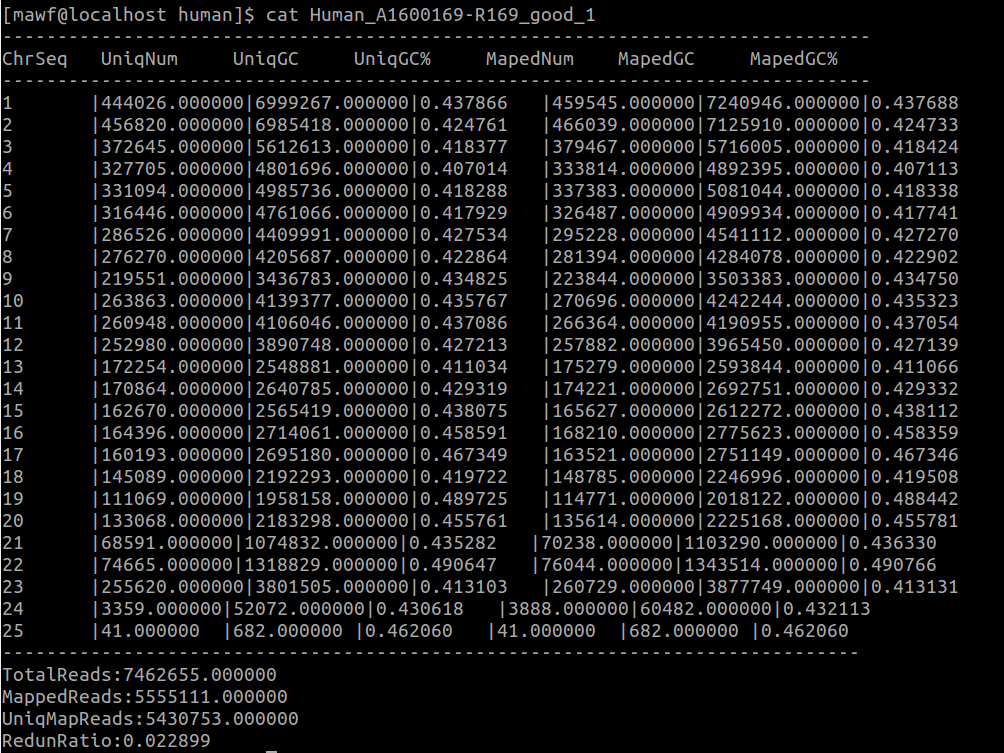


图２

**1.3参数说明**

TotalReads：样本总reads量；

MappedReads：比对上的总reads量（所有染色体比对的总和）；

UniqMapReads：唯一比对上的总reads量（所有染色体比对的总和）；

RedunRatio：比对冗余率；

chrSeq:比对上的reads所在染色体号；

MapedNum: 比对上的reads数量；

MapedGC: 比对上reads的GC含量；

MapedGC%: 比对上reads的GC含量占reads总碱基的百分率；

UniqNum：唯一比对上reads数量；

UniqGC：唯一比对上的reads的GC含量；

UniqGC%: 唯一比对上的reads的GC占reads总碱基的百分率；

# 2. 技术路线

2.1 参考基因组的预处理，用perl脚本把参考基因组预处理为如下格式。

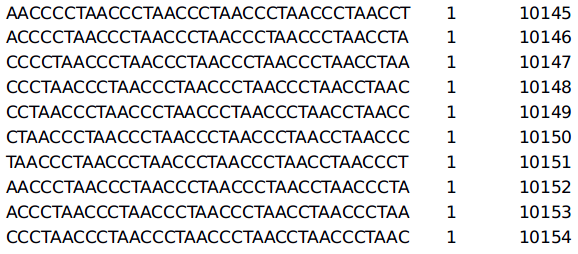


图３

说明：

用于index建库。

第一列：36kmer DNA序列片段；

第二列：片段所在染色体；

第三列：片段所在染色体上的位点；

预处理脚本说明：

主要由perl语言和shell命令完成：

segmentation.pl 分割提取hg19.fa单染色体信息

clc\_break.pl 滑窗式分割为36kmer序列片段

clc\_N.pl 去除含N的序列判断

clc\_cut.pl 分割文件以待处理

clc\_fq.pl 处理成fq格式文件

clc\_bwa.pl 提取bwa唯一比对上的reads

Prepro.sh 辅助处理的各种shell命令

2.2 构建index库

1. 利用XUDDHash算法和MWFHash算法分别对图3文件reads序列进行一次hash和二次hash运算求得index1和index2，分别作为构建index库的索引值和索引冲突分离值。
2. 利用求整求余的算法压缩reads序列所在位置为整数值value。
3. 构建结构体数据结构存储冲突分离索引index2和reads位置压缩值value。
4. 构建结构体数组用来存储结构体，索引为index1。

## 2.3 比对

1. 下机数据reads序列利用XUDDHash算法和MWFHash映射为index库索引值
2. 根据索引值index1查询reads序列位置压缩值value，并提取出参考基因组的位置。
3. 若有冲突，利用index2信息分离冲突值，然后查询reads序列位置压缩值value，提取reads序列所在的位置。

## 2.4生成结果文件

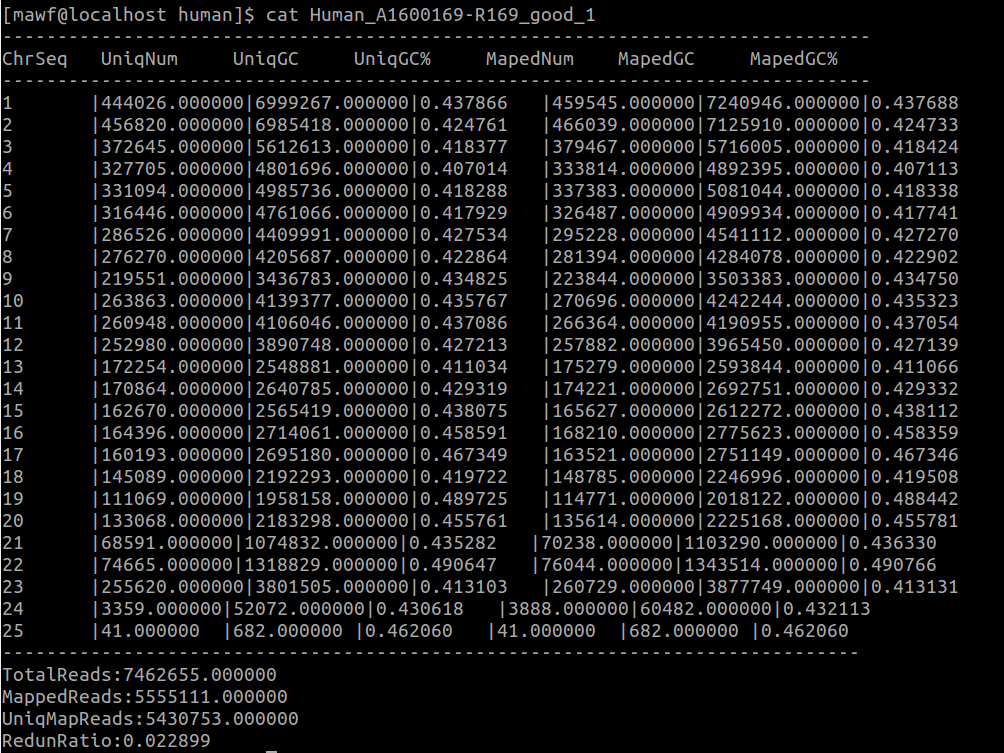


图４

# 3.声明

本软件可以实时监控下机数据目录，可以利用nohup ....＆，将其转化为守护进程，实时监控。