VCF文件格式

1. 什么是vcf文件

- VCF是用于描述SNP（单个碱基上的变异），INDEL（插入缺失标记）和SV（结构变异位点）结果的文本文件
- 由sam/bam文件而来

2. VCF的主体结构

- 以“#”开头的注释部分：一些对VCF的介绍信息，和每一列所代表的意义
- 主体部分中每一行代表一个Variant的信息

列名	意义
CHROM	表示变异位点是在哪个contig
POS	变异位点相对于参考基因组所在的位置
ID	variant的ID，若在dbSNP中则写出，若无则写'.'
REF	参考基因组中所对应的碱基
ALT	研究对象基因组的碱基
QUAL	质量值，值越高，variant可能性越高
FILTER	表示是否通过过滤
INFO	变异的描述信息
FORMAT	表示基因型信息的多个标签,具体信息在第十列
SAMPLES	样本信息，以冒号分割

> INFO
> AC:表示基因型与变异一致的等位基因
> AN:表示等位基因的总数目
> AF:Allele的频率，AF=AC/AN
> DP:reads被过滤后的覆盖度
> FS：正负链偏移，该值越小越好

> FORMAT
> GT:样品的基因型。0/0，纯合位点；0/1，杂合突变；1/1纯合突变；1/2，杂合突变
> AD:每一种等位碱基的reads覆盖度，逗号分开，前者对应REF，后者对应ALT基因型
> DP:为该位点的覆盖度，是所支持的两个AD的值的和
> GQ:基因性的质量值，表示该位点基因型存在的可能性
> PL：三种基因性的质量值

GATK大致流程

1. GATK

- 用于分析高通量测序数据的命令行工具集
- 是一款从高通量测序数据中分析变异信息的软件，是目前最主流的snp calling 软件之一。GATK 设计之初是用于分析人类的全外显子和全基因组数据，随着不断发展，现在也可以用于其他的物种，还支持CNV和SV变异信息的检测

2. 判断变异检测结果的好坏

- 变异数目
- indel ratio（插入和缺失比率）
- titv ratio（转换和颠换比率）

spark基础

1. 什么是spark

- spark是一个实现快速通用的集群计算平台。它是由加州大学伯克利分校AMP实验室开发的通用内存并行计算框架，用来构建大型的、低延迟的数据分析应用程序。
- 它扩展了广泛使用的MapReduce计算模型。高效的支撑更多计算模式，包括交互式查询和流处理。spark的一个主要特点是能够在内存中进行计算，及时依赖磁盘进行复杂的运算，Spark依然比MapReduce更加高效。

2. RDD

- 弹性分布式数据集，是Spark中最基本的数据抽象，它代表一个不可变、可分区、里面的元素可并行计算的集合
- 可以分布在集群的节点上，以函数式操作集合的方式，进行各种并行操作

3. RDD编程

- Transformation：

转换	
map	返回一个新的RDD，该RDD由每一个输入元素经过func函数转换后组成
filter	返回一个新的RDD
flatMap	类似于map，但是每一个输入元素可以被映射为0或多个输出元素。 flatMap会将其返回的数组全部拆散，然后合成到一个数组中。
reduceByKey	使用指定的reduce函数，将相同key的值聚合到一起

- Action

动作	
reduce	通过函数聚集RDD中的所有元素，这个功能必须是可交换且可并联的
collect	在驱动程序中，以数组的形式返回数据集的所有元素
count	返回RDD的元素个数

4. Hadoop

1. hadoop

- Hadoop是Apache开源组织的一个分布式计算开源框架，用java语言实现开源软件框架，实现在大量计算机组成的集群中对海量数据进行分布式计算
- Hadoop框架中最核心设计就是：HDFS和MapReduce，HDFS实现存储，而MapReduce实现原理分析处理，这两部分是hadoop的核心。是一个高性能处理海量数据集的工具。

2. HDFS

- 分布式文件系统，它是一个高度容错性的系统，适合部署在廉价的机器上。HDFS能提供高吞吐量的数据访问，适合那些有着超大数据集的应用程序。
- 特点：大数据文件，文件分块储存，流式数据访问（一次写入多次读写），廉价硬件，硬件故障
- 构架：一个**Namenode**和一定数目的**Datanode**组成。Namenode是一个中心服务器，负责管理文件系统的namespace和客户端对文件的访问。Datanode在集群中一般是一个节点一个，负责管理节点上它们附带的存储。在内部，一个文件其实分成一个或多个block，这些block存储在Datanode集合里。Namenode执行文件系统的namespace操作，例如打开、关闭、重命名文件和目录，同时决定block到具体Datanode节点的映射。Datanode在Namenode的指挥下进行block的创建、删除和

复制。Namenode和Datanode都是设计成可以跑在普通的廉价的运行linux的机器上。

3. MapReduce

- 是一种编程模型，用于大规模数据集（大于1TB）的并行运算。MapReduce将分成两个部分"Map（映射）"和"Reduce（归约）"。
- 当你向MapReduce框架提交一个计算作业时，它会首先把计算作业拆分成若干个Map任务，然后分配到不同的节点上去执行，每一个Map任务处理输入数据中的一部分，当Map任务完成后，它会生成一些中间文件，这些中间文件将会作为Reduce任务的输入数据。Reduce任务的主要目标就是把前面若干个Map的输出汇总到一起并输出。

4. HBase

- HBase 是 BigTable 的开源 java 版本。是建立在 HDFS 之上，提供高可靠性、高性能、列存储、 可伸缩、实时读写 NoSQL 的数据库系统。
- 非结构化数据储存的数据库，分布式储存系统，面向列的开源的数据库
- HBase数据表”schema-less“的特点 : 每一行中，列的组成都是灵活的，行与行之间并不需要遵循相同的列定义，
- **RowKey**:用来表示唯一一行记录的主键，HBase的数据是按照RowKey的字典顺序进行全局排序的，所有的查询都只能依赖于这一个排序维度。
- **Region**:将HBase中拥有数亿行的一个大表，横向切割成一个个”子表“，这一个个”子表“就是Region
- **Column Family**：每一个列，都必须归属于一个Column Family，这个归属关系是在写数据时指定的，而不是建表时预先定义。
- HBase的操作：

1. 可以直接通过bbase shell的方式进行导入：create 表名，列族
   put 表名 行名 列名 值
2. 可以通过编程额方式进行导入（python api：thrift等）

Hail

1. Hail

是一个开源的，通用的，基于Python的数据分析工具，具有用于处理基因组数据的其他数据类型和方法。Hail是按比例建立的，并且具有对多维结构化数据的一流支持，如全基因组关联研究（GWAS）中的基因组数据。

2. 具体操作

其他

- 虚拟机的安装
- zsh
- SwitchyOmega代理扩展插件

总结

1. 不足

- 在安装虚拟机的时候浪费了很多时间
- 由于对linux操作系统运用不熟练，在一些程序的环境配置上浪费了时间
- 由于英文水平不足，在看一些包和软件的官方文档时比较困难
- 对集群，节点，环境等概念理解不是很深

2. 收获

- 进一步了解了VCF文件的格式
- 了解了spark的基本概念和基本操作
- 了解了大数据计算，并行计算的基本工具和基本方法。
- 熟悉了linux操作系统的操作