# GaeaAnnotator 说明文档 (v0.1)

大数据计算组　　黄志博

2017 年 3 月 2 日

## 目录

# 1 概述

## 1.1 开发目的

随着高通量测序技术的发展及其越来越广泛的应用，当今世界正在累积巨量的 NGS 数据，而对 NGS 变异数据的注释和解读已经成为快速理解这些测序数据的一个瓶颈。为了实现对大量变异数据的快速注释，我们决定基于 hadoop 平台开发并行化的注释程序。

## 1.2 开发人员

- 黄志博
- 张勇
- 李胜康
- 石泉
- 肖鹏

# 2 程序特性说明

## 2.1 reference 版本

目前暂只支持 hg19/GRCh37 版本，后续会导入 GRCh38 版本数据库。

- hg19/GRCh37
  - GRCh38

## 2.2 注释变异类型

| Type | What is means | Example |
|---|---|---|
| SNP | Single-Nucleotide Polymorphism | Reference = 'A', Sample = 'C' |
| Ins | Insertion | Reference = 'A', Sample = 'AGT' |
| Del | Deletion | Reference = 'AC', Sample = 'C' |
| MNP | Multiple-nucleotide polymorphism | Reference = 'ATA', Sample = 'GTC' |
| MIXED | Multiple-nucleotide and an InDel | Reference = 'ATA', Sample = 'GTCAGT' |

## 2.3 输入输出格式

- 输入文件: VCF
- 输出文件: TSV, Excel[1]

## 2.4 运行环境及开发语言

- 运行环境: Hadoop 2.6.0 + Hbase 1.2.0
- 开发语言: Java

[1] 使用工具/ifs4/ISDC_BD/huangzhibo/anno/GaeaAnnotator/toExcel.sh 将 TSV 格式结果转换成 Excel，注意 TSV 文件不能过大

### 2.5 性能

- 在 BGI hadoop50 集群注释单个 WGS 变异数据（NA2878 样本 4987513 条变异数据）用时约 11 分钟。

## 3 注释数据库

对 NGS 变异信息进行 Gene 注释、功能预测和致病性预测等。

### 3.1 基因及转录本信息数据库

- UCSC refgene
- ENSEMBL Gene
  - UCSC Known Gene
  - CCDS

### 3.2 其他可用数据库

详细信息见AnnotationDatabase.md,
注释条目列表见/ifs4/ISDC _BD/huangzhibo/Data/database/header/。

- ESP6500
- G1000
- EXAC
- dbSNP
- HGNC
- gwasCatalog
- CLINVAR
- dbNSFP

## 4 使用说明

### 4.1 配置文件

配置文件中需要指明以下几项内容:

ref： reference 版本

GeneInfo： 指定基因及转录本信息文件

GeneInfoType： 指定基因及转录本信息数据库版本

{dbName}.fields： 根据注释条目 List 中的设置各数据库的注释字段

**配置文件示例：** /ifs4/ISDC_BD/GaeaProject/software/GaeaAnnotator/config.properties

```
ref          = hg19
GeneInfo     = file:///ifs4/ISDC_BD/huangzhibo/anno/GaeaAnno/hg19_refGene.txt
GeneInfoType = refGene
#ensGene

# dbName.fields = Header Name
GeneInfo.fields    = EFFECT,IMPACT,GENE,GENEID,TRID,BIOTYPE,RANK,HGVS_DNA,HGVS_P

HGMD.fields        = HGMD_disease,HGMD_gene,HGMD_chrom,HGMD_genename,HGMD_gdbid,HGMD_omimid,
     HGMD_amino,HGMD_deletion,HGMD_insertion,HGMD_codon,HGMD_codonAff,HGMD_descr,HGMD_hgvs,
     HGMD_hgvsAll,HGMD_dbsnp,HGMD_chromosome,HGMD_startCoord,HGMD_endCoord,HGMD_tag,HGMD_dmsupport,
     HGMD_author,HGMD_fullname,HGMD_allname,HGMD_vol,HGMD_page,HGMD_year,HGMD_pmid,HGMD_reftag,
     HGMD_comments,HGMD_acc_num,HGMD_new_date,HGMD_base
ESP6500.fields     = MAF,HGVS_CDNA_VAR,HGVS_PROTEIN_VAR,GWAS_PUBMED,GTC,GTS,EA_AC,AA,AA_AC,AA_AGE,
     AA_GTC,CDS_SIZES,GL
G1000.fields       = AF,EUR_AF,AFR_AF,AMR_AF,EAS_AF,SAS_AF,VT
EXAC.fields        = AC,AF
dbSNP.fields       = RS,DBSNP_CAF,DBSNP_COMMON,dbSNPBuildID
CLINVAR.fields     = CLNACC,CLNDBN,CLNDSDBID,CLNDSDB,CLNSIG,CLNSRCID,CAF,CLNHGVS,GENEINFO,SAO,VC,PM,
     OM,MTP,R3,R5,ALT
gwasCatalog.fields     = riskAllel,riskAlFre,title,pubMedID,trait
HGNC.fields       = hgnc_id,symbol,name,locus_group,locus_type,status,location,location_sortable,
     alias_symbol,alias_name,prev_symbol,prev_name,gene_family,gene_family_id,date_approved_reserved,
     date_symbol_changed,date_name_changed,date_modified,entrez_id,ensembl_gene_id,vega_id,ucsc_id,
     ena,refseq_accession,ccds_id,uniprot_ids,pubmed_id,mgd_id,rgd_id,lsdb,cosmic,omim_id,mirbase,
     homeodb,snornabase,bioparadigms_slc,orphanet,pseudogene.org,horde_id,merops,imgt,iuphar,
     kznf_gene_catalog,mamit-trnadb,cd,lncrnadb,enzyme_id,intermediate_filament_db
dbNSFP.fields    = SIFT_pred,SIFT_score,Polyphen2_HDIV_pred,Polyphen2_HDIV_score,Polyphen2_HVAR_pred,
     Polyphen2_HVAR_score,LRT_pred,LRT_score,MutationTaster_pred,MutationTaster_score,
     MutationAssessor_pred,MutationAssessor_score,FATHMM_pred,FATHMM_score
```

## 4.2 运行

**程序路径：** /ifs4/ISDC_BD/GaeaProject/software/GaeaAnnotator/GaeaAnnotator.jar

**示例：** /ifs4/ISDC_BD/GaeaProject/software/GaeaAnnotator/runGaeaAnnotator.sh

Uasge:

```
 -c,--config <arg>      config file.
    --cacheref          DistributedCache reference sequence file list
    --debug             for debug.
 -h,--help              help information.
 -i,--input <arg>       input file(VCF).
 -m,--mapperNum <arg>   mapper number.
 -o,--output <arg>      output file of annotation results
 -r,--reference <arg>   indexed reference sequence file list [null]
    --verbose           display verbose information.
```

说明：

-i 须指定非压缩 VCF 文件。

-o 本地输出文件。

-r vcf 对应的 reference 序列文件

-c 配置文件

-m mapper 数目，根据集群情况设定。

**注意** 除-o 指定的输出文件为本地路径外，其他参数指定的输入文件默认在 HDFS[2]上，本地文件需在路径前加 “file://” 。

[2]分布式文件系统