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/GaeaProject/software/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
- HGMD

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 header List 中[2]的条目设置各数据库的注释字段。不设置或值为空则视为不对该数据库进行注释。

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

```
ref          = hg19
GeneInfo     = /ifs4/ISDC_BD/GaeaProject/software/GaeaAnnotator/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

执行方式： hadoop jar GaeaAnnotator.jar [options]

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

示例脚本： /ifs4/ISDC_BD/GaeaProject/software/GaeaAnnotator/run.sh

目前只开放 hadoop50 集群作为注释计算平台，使用方法见示例脚本。程序具体参数如下：

```
 -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.
```

[2]见/ifs4/ISDC _BD/huangzhibo/Data/database/header/

说明：

-i 须指定非压缩 VCF 文件。

-o 本地输出文件（gz 压缩）。

-r vcf 对应的 reference 序列 Gaea 索引文件，见/ifs4/ISDC_BD/GaeaProject/reference/

-c 配置文件

-m mapper 数目，根据集群情况设定。

4.3 结果展示

#CHROM	POS	REF	ALT	EFFECT	IMPACT	GENE	GENEID	TRID	BIOTYPE	RANK	HGVS_DNA	HGVS_P	gwasCatalog_riskAllel	gwasCatalog_ris
22	37690808		C	T	intron_variant	MODIFIER		CYTH4	CYTH4	NM_013385		protein_coding	3	c.167+43C>T	.
22	37692024		G	A	intron_variant	MODIFIER		CYTH4	CYTH4	NM_013385		protein_coding	3	c.168-16G>A	.
22	37692092		A	G	missense_variant	MODERATE		CYTH4	CYTH4	NM_013385		protein_coding	4	c.220A>G	
22	37692120		G	A	intron_variant	MODIFIER		CYTH4	CYTH4	NM_013385		protein_coding	4	c.234+14G>A	.

图 1: TSV 格式注释结果

CHRO	POS	REF	ALT	EFFEC	IMPA	GENE	GENE	TRID	BIOTY	RANK	HGVS	HGVS	gwas	gwas	gwas	gwas	gwas	G100	G100
22	37690808	C	T	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	3	c.167+43C							0.00019960	0
22	37692024	G	A	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	3	c.168-16G							0.0029952	0
22	37692092	A	G	missense_v	MODERAT	CYTH4	CYTH4	NM_01338	protein_co	4	c.220A>G	p.Met74Va						0.0213658	0.001
22	37692120	G	A	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	4	c.234+14G							0.0015974	0
22	37693632	C	T	synonymou	LOW	CYTH4	CYTH4	NM_01338	protein_co	5	c.262C>T	p.Leu88Le						0.0061901	0
22	37693646	C	T	synonymou	LOW	CYTH4	CYTH4	NM_01338	protein_co	5	c.276C>T	p.Asp92As						0.0009984	0
22	37693689	A	C	missense_v	MODERAT	CYTH4	CYTH4	NM_01338	protein_co	5	c.319A>C	p.Asn107H						0.0061901	0
22	37693732	G	A	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	5	c.353+9G>							0.0003993	0
22	37693760	A	C	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	5	c.353+37A							0.00019960	0
22	37693761	A	C	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	5	c.353+38A							0.0231629	0.001
22	37695256	T	C	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	5	c.354-11T>							0.679313	0.5765
22	37696906	G	A	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	6	c.435-42G							0.0157748	0.001
22	37696927	G	A	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	6	c.435-21G							0.0003993	0
22	37696934	C	T	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	6	c.435-14C							0.0309505	0.002
22	37696978	C	T	synonymou	LOW	CYTH4	CYTH4	NM_01338	protein_co	7	c.465C>T	p.Gly155Gl						0.0035942	0.004
22	37697110	C	T	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	7	c.547+50C							0.0025958	0
22	37699262	G	C	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	7	c.548-33G							0.0157748	0
22	37699279	G	A	intron_vari	MODIFIER	CYTH4	CYTH4	NM_01338	protein_co	7	c.548-16G							0.0013977	0
22	37699323	C	T	synonymou	LOW	CYTH4	CYTH4	NM_01338	protein_co	8	c.576C>T	p.Ile192Ile						0.0003993	0

图 2: Excel 格式注释结果