Zhang-Lab 生信小课堂 第廿二期
Applied Bioinformatics Club (ABC)

水稻基因组变异检测

2025.09.05 二综C102 15:00 欢迎大家交流学习

主讲人：李思远
2025/09/05

软件选择

使用测序数据进行变异检测：

GATK

Deepvariant

两基因组间变异检测：

SYRI

BWA+GATK（二代reads测序数据）

比对-转换

```
module load BWA/0.7.17
module load SAMtools/1.9
bwa mem -t 20 -R '@RG\tID:id\tPL:Illumina\tSM:ng9108'
/public/home/siyli/NG9108/RAWreads/24_10PTG267_5-T250105_0029/bwa/filter_oNG9108_r1
/public/home/siyli/NG9108/RAWreads/24_15CK-T250105_0028/T250105_0028_RUN871_1.fq.gz
/public/home/siyli/NG9108/RAWreads/24_15CK-T250105_0028/T250105_0028_RUN871_2.fq.gz |
samtools view -bS -o r1_ng91082_aligned.bam
```

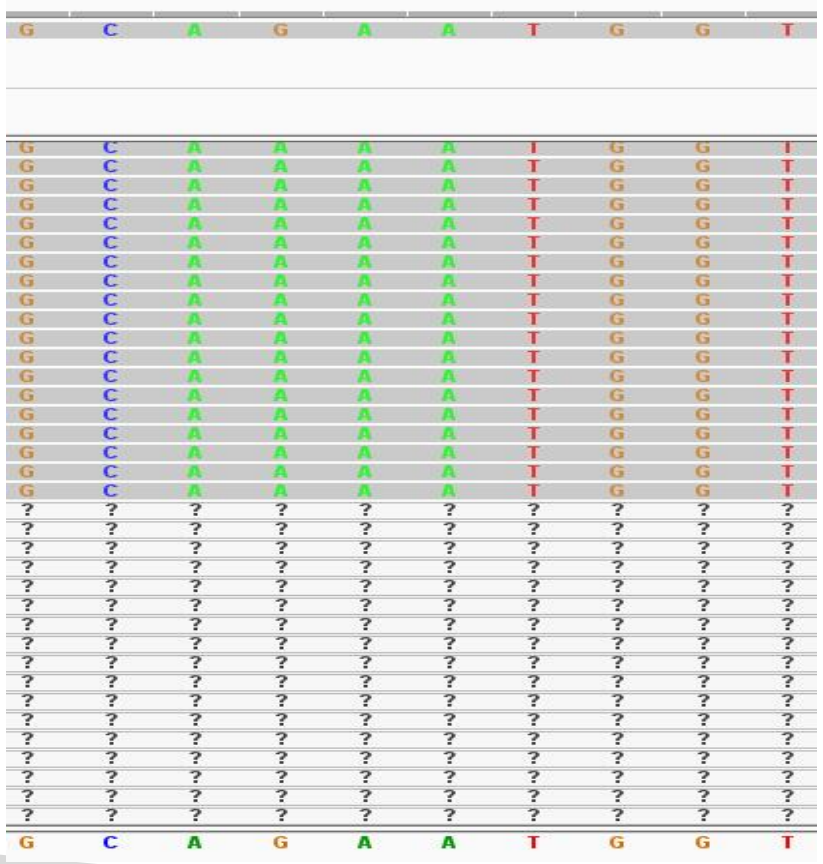
排序

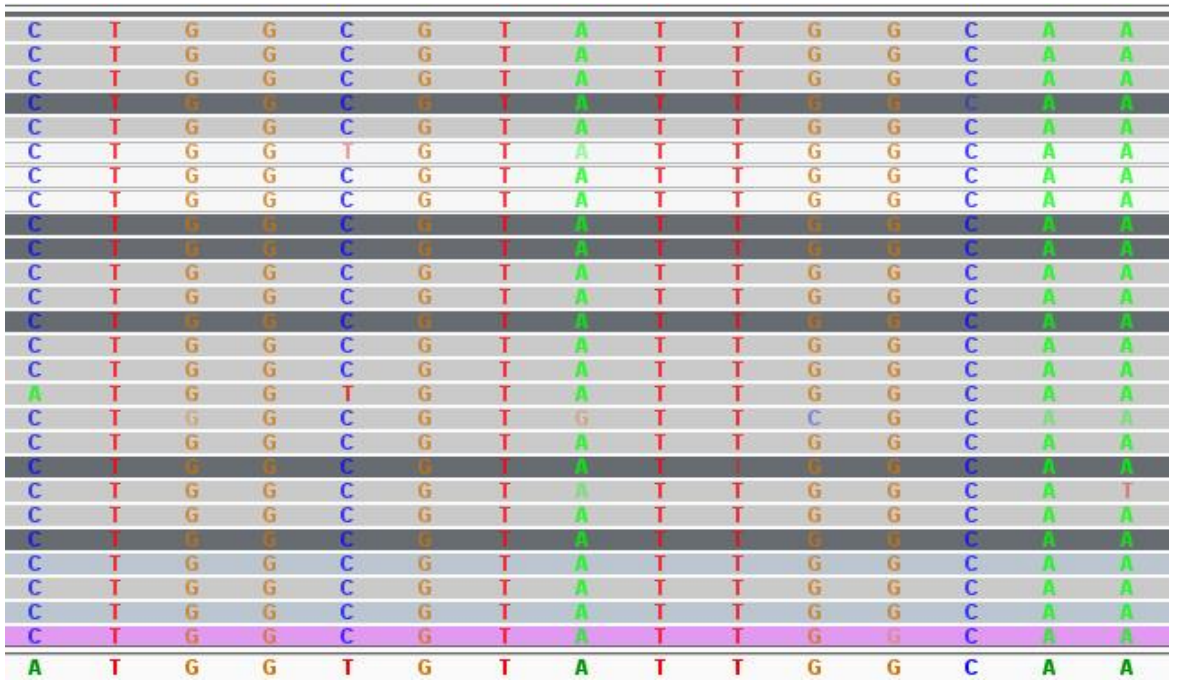
```
module load SAMtools/1.9
samtools sort -o r1_bwa0028_filter_ng9108_sorted_aligned.bam
/public/home/siyli/NG9108/RAWreads/24_15CK-
T250105_0028/bwa/r1_bwa0028_filter_ng9108_aligned.bam
```

过滤BAM

```
samtools view -h -F 0x804 /public/home/siyli/NG9108/gpm/polish/racon2/NG9108_minimap2.sam > NG9108mini_filter1.sam
```

```
samtools view -h /public/home/siyli/NG9108/RAWreads/24_15CK-T250105_0028/bwa/r1_ng91082_aligned.bam | awk '$1 ~ /^@/ || $7 == "=" || $7 == "*"' | samtools view -b -o r1_bwa0028_filtei_ng9108_aligned.bam
```





GATK变异检测

```
module load GATK/4.6.0.0
gatk HaplotypeCaller -R /public/home/siyli/NG9108/RAWreads/24_10PTG267_5-
T250105_0029/bwa/filter_oNG9108_r1.fasta -I /public/home/siyli/NG9108/RAWreads/24_15CK-
T250105_0028/bwa/r1_bwa0028_filter_ng9108_sorted_aligned.bam -O 0028_filter_variants.g.vcf.gz
```

```
#CHROM  POS ID  REF ALT QUAL  FILTER  INFO  FORMAT  ng9108
Chr01  124049 .  A  C  39.64  .  AC=1;AF=0.500;AN=2;BaseQRankSum=0.081;DP=12;ExcessHet=0.0000;FS=26.646;MLEAC=
1;MLEAF=0.500;MQ=58.26;MQRankSum=-0.230;QD=3.60;ReadPosRankSum=-2.166;SOR=3.892  GT:AD:DP:GQ:PL  0/1:5,6:11:47:47,0,79
Chr01  184527 .  GCCACCCGCACGGCGCCTCCCAGCTGCAGCGCCTCGGCCTCGGCCTCCTCCTCT  G  40.60  .  AC=1;AF=0.500;AN=
2;BaseQRankSum=-2.362;DP=14;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=
2.90;ReadPosRankSum=2.362;SOR=0.269  GT:AD:DP:GQ:PL  0/1:12,2:14:48:48,0,498
Chr01  195242 .  CTGCTCCGAGCAGCCCCGCCGCTCGGGGGGGGAGAGGAGGGGGTGG C  43.60  .  AC=1;AF=0.500;AN=
2;BaseQRankSum=-1.335;DP=14;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=
3.35;ReadPosRankSum=0.473;SOR=0.180  GT:AD:DP:GQ:PL  0/1:11,2:13:51:51,0,448
```

过滤低置信位点

```
gatk VariantFiltration \
  -R /path/to/reference.fasta \
  -V 0028_filter_variants.g.vcf.gz \
  -O cohort_filtered.vcf.gz \
  --filter-expression "QD < 2.0 || FS > 60.0 || MQ < 40.0" \
  --filter-name "LowQuality"
```

Minimap2+SYRI（完整基因组变异检测）

比对

```
module load minimap2/v2.17
minimap2 -ax asm5 --eqx
/public/home/siyli/NG9108/gpm/polish/racon1/filter_oNG9108_r1.fasta
/public/home/siyli/NG9108/gpm/polish/racon2/filter_oNG9108_r2.fasta >
alnout2_1.sam
```

转换+排序

```
module load SAMtools/1.9
samtools view -b /public/home/siyli/NG9108/gpm/polish/racon2/alnout2_1.sam |
samtools sort -o aln2_1.bam
```

SYRI检测

```
module load syri/1.6
syri -c /public/home/siyli/NG9108/gpm/polish/racon2/alnout2_1.sam -r
/public/home/siyli/NG9108/gpm/polish/racon1/filter_oNG9108_r1.fasta -q
/public/home/siyli/NG9108/gpm/polish/racon2/filter_oNG9108_r2.fasta -k -F S
```

-c 比对文件
-r 参考基因组（fasta)
-q 查询基因组文件(fasta)
-F S输出为sam文件

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
Chr01	1	SYN1	N	<SYN>	.	PASS	END=43725062;ChrB=Chr01;StartB=1;EndB=43723496;Parent=.;VarType=SR;DupType=-
Chr01	1	SYNAL1	N	<SYNAL>	.	PASS	END=43725062;ChrB=Chr01;StartB=1;EndB=43723496;Parent=SYN1;VarType=.;DupType=.
Chr01	66	DEL3	TA	T	.	PASS	END=67;ChrB=Chr01;StartB=66;EndB=66;Parent=SYN1;VarType=ShV;DupType=.
Chr01	492	INS4	T	TA	.	PASS	END=492;ChrB=Chr01;StartB=491;EndB=492;Parent=SYN1;VarType=ShV;DupType=.
Chr01	505	SNP5	T	A	.	PASS	END=505;ChrB=Chr01;StartB=505;EndB=505;Parent=SYN1;VarType=ShV;DupType=.
Chr01	660	DEL6	TAAT	T	.	PASS	END=662;ChrB=Chr01;StartB=660;EndB=660;Parent=SYN1;VarType=ShV;DupType=.
Chr01	1075	INS7	T	TA	.	PASS	END=1075;ChrB=Chr01;StartB=1073;EndB=1074;Parent=SYN1;VarType=ShV;DupType=.
Chr01	2812	INS8	T	TAAACCCTA	.	PASS	END=2812;ChrB=Chr01;StartB=2811;EndB=2819;Parent=SYN1;VarType=ShV;DupType=.

stats

```
bcftools stats cohort_filtered.vcf.gz > vcf_stats.txt
```

```
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      1
SN      0       number of records:      16293
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 9938
SN      0       number of MNPs: 0
SN      0       number of indels:       6355
SN      0       number of others:       1
SN      0       number of multiallelic sites:   63
SN      0       number of multiallelic SNP sites:       1
# TSTV, transitions/transversions:
# TSTV  [2]id   [3]ts   [4]tv   [5]ts/tv        [6]ts (1st ALT) [7]tv (1st ALT) [8]ts/tv (1st ALT)
TSTV    0       5551    4387    1.27    5551    4386    1.27
# SiS, Singleton stats:
# SiS   [2]id   [3]allele count [4]number of SNPs       [5]number of transitions        [6]number of transversions      [7]number of indels     [8]repeat-consistent    [9]r
epeat-inconsistent  [10]not applicable
SiS     0       1       9627    5344    4283    5921    0       0       5921
# AF, Stats by non-reference allele frequency:
# AF    [2]id   [3]allele frequency     [4]number of SNPs       [5]number of transitions        [6]number of transversions      [7]number of indels     [8]repeat-consistent
    [9]repeat-inconsistent  [10]not applicable
AF      0       0.000000        9627    5344    4283    5921    0       0       5921
AF      0       0.990000        311     207     104     495     0       0       495
# QUAL, Stats by quality:
# QUAL  [2]id   [3]Quality      [4]number of SNPs       [5]number of transitions (1st ALT)      [6]number of transversions (1st ALT)    [7]number of indels
QUAL    0       30      76      49      27      152
QUAL    0       31      559     309     250     256
QUAL    0       32      149     99      50      136
```

Deepvariant（深度学习方法）

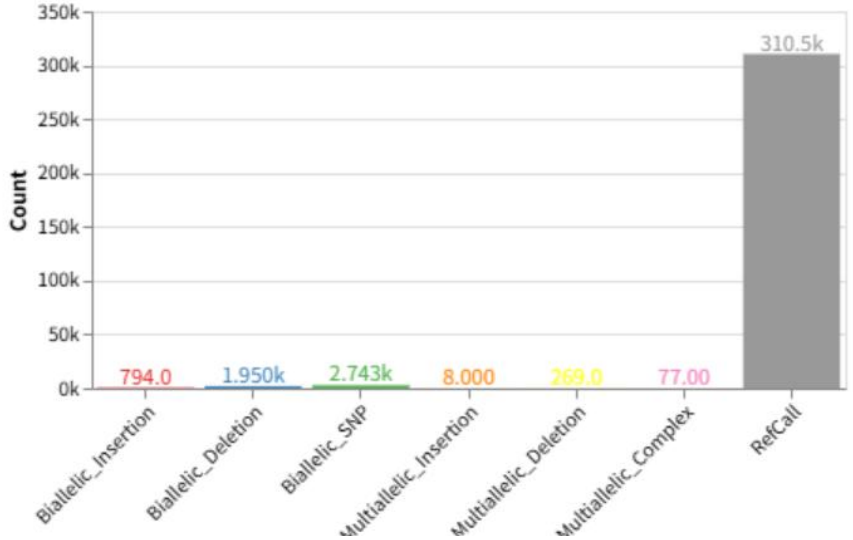
```
#使用singularity安装deepvariant
singularity pull --arch amd64 library://glebus-sasha/repo/deepvariant:1.6.1
#运行命令
#BSUB -J deepvariant
#BSUB -n 10
#BSUB -R span[hosts=1]
#BSUB -o deepvariant.out
#BSUB -e deepvariant.err
#BSUB -q normal
module load Singularity/3.7.3
singularity run deepvariant_1.6.1.sif /opt/deepvariant/bin/run_deepvariant --model_type=WGS --
ref=/public/home/siyli/NG9108/RAWreads/24_10PTG267_5-T250105_0029/bwa/filter_oNG9108_r1.fasta --
reads=/public/home/siyli/NG9108/RAWreads/24_15CK-T250105_0028/bwa/r1_bwa0028_filter_ng9108_sorted_aligned.bam --
output_vcf=/public/home/siyli/NG9108/RAWreads/24_15CK-T250105_0028/bwa/deepvariant/NG9108_pvariant.vcf --
output_gvcf=/public/home/siyli/NG9108/RAWreads/24_15CK-T250105_0028/bwa/deepvariant/NG9108_pvariant.gvcf --
num_shards=10
```

```
##DeepVariant_version=1.6.0
##contig=<ID=Chr01,length=43725062>
##contig=<ID=Chr02,length=36501135>
##contig=<ID=Chr03,length=37904905>
##contig=<ID=Chr04,length=35479058>
##contig=<ID=Chr05,length=30405573>
##contig=<ID=Chr06,length=31344195>
##contig=<ID=Chr07,length=29779047>
##contig=<ID=Chr08,length=28816348>
##contig=<ID=Chr09,length=23585737>
##contig=<ID=Chr10,length=24081125>
##contig=<ID=Chr11,length=30333112>
##contig=<ID=Chr12,length=27745174>
#CHROM  POS ID  REF ALT QUAL  FILTER  INFO  FORMAT  ng9108
Chr01  5276  .  TA  T  2.2  RefCall  .  GT:GQ:DP:AD:VAF:PL  ./.:4:13:11,2:0.153846:0,2,8
Chr01  5324  .  TA  T  5.1  PASS  .  GT:GQ:DP:AD:VAF:PL  1/1:2:5:2,3:0.6:1,1,0
Chr01  5634  .  TA  T  5.9  PASS  .  GT:GQ:DP:AD:VAF:PL  1/1:3:3:1,2:0.666667:2,2,0
Chr01  5641  .  TA  T  0.6  RefCall  .  GT:GQ:DP:AD:VAF:PL  ./.:9:4:1,3:0.75:0,12,10
Chr01  15776  .  CCT C  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:38:27:22,4:0.148148:0,38,52
Chr01  16142  .  G  C  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:43:21:17,4:0.190476:0,43,53
Chr01  16152  .  A  C  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:49:23:19,3:0.130435:0,49,55
Chr01  17874  .  T  C  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:49:23:20,3:0.130435:0,49,54
Chr01  31281  .  GT  G  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:34:21:18,2:0.0952381:0,34,50
Chr01  34335  .  A  AG  19.9 PASS  .  GT:GQ:DP:AD:VAF:PL  0/1:16:9:4,2:0.222222:19,0,17
Chr01  36600  .  G  GT  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:45:43:37,5:0.116279:0,45,53
Chr01  39919  .  G  T  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:33:10:8,2:0.2:0,33,51
Chr01  52297  .  C  CGA  0  RefCall  .  GT:GQ:DP:AD:VAF:PL  0/0:39:28:26,2:0.0714286:0,38,49
```

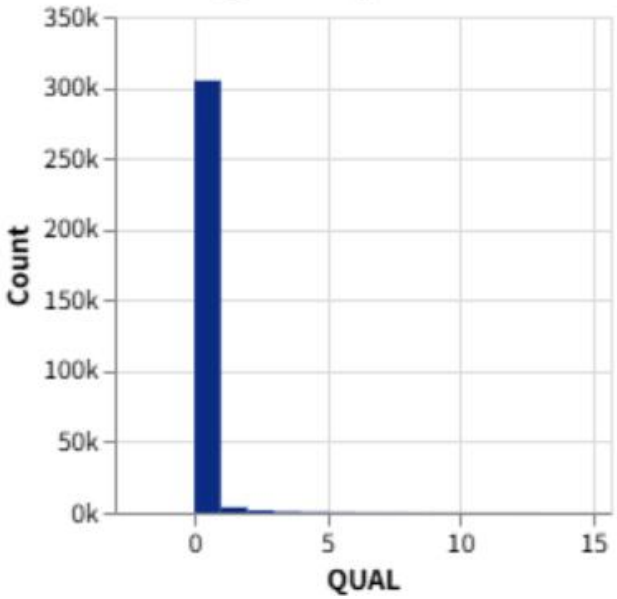
RefCall，表示模型判定该位点无真实变异。

ng9108

Variant types

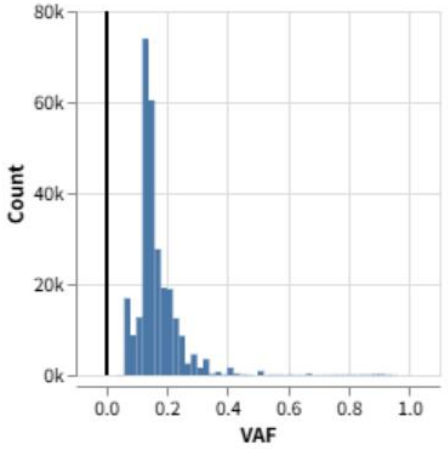


Quality score

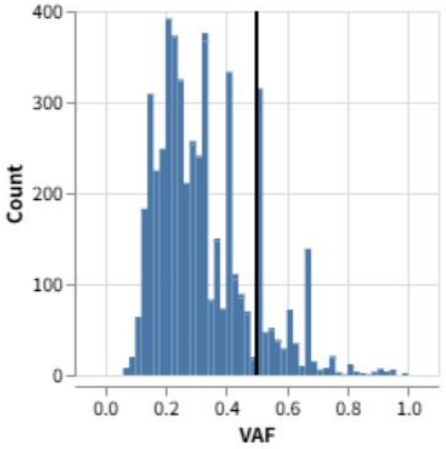


Main genotypes

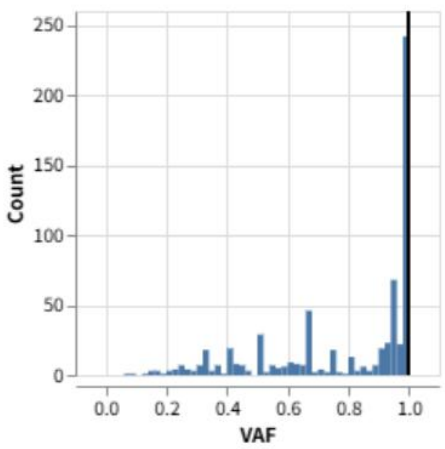
Ref (0/0)



Het (0/x)



Hom (x/x)



质量范围	正确率
10≤QUAL＜20	15.35%
20≤QUAL＜30	53.81%
QUAL≥30	97.55%