Work Directory:

/fslgroup/cotton\_seq/compute/ReSeq\_Project/AllTetrapliod\_20170911/GvcfList/GATK/

1.Calling variation with 1432 accessions by 500kb region (01.GATK.CallSNP.Loc.sh, 1432 All.GVCF)

VCF: all the variation

SNP: only SNP, filterExpression "QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"

INDEL: --filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0"

2. Mask Low or High coverage region, and recalculate the AN and AC value of VCF files, and calculate the missing rate for each sample (02.SNP.MaskLowHigh.FillAnAc.sh)

--minDP 3 --maxDP 1000

SNP\_Filt/${BASE}.flt.01.vcf.gz: total SNP 116,292,216 (116Mb)

Sample missing rate (All.imiss.indv.sort.txt)

High missing rate samples (>25% missing loci): HighMissingRate

3. Filter high missing rate samples (>0.25), very low MAF (0.01 and 0.99) (Minor Allele Frenquency) and high missing rate loci (0.25).

SNP\_Filt/${BASE}.flt.03.vcf.gz

4. Others filter

--maf 0.01 --max-maf 0.99 --max-missing 0.75

Directory/Path MAF missingRate

SNP\_Filt\_Merge\_M01m25 minF 0.01 maxF 0.99 missing 0.75 36235697

SNP\_Filt\_Merge\_M01m10 minF 0.01 maxF 0.99 missing 0.90 29851575

SNP\_Filt\_Merge\_M05m10 minF 0.05 maxF 0.95 missing 0.90 18663317

SNP\_Filt\_Merge2 minF 0.10 maxF 0.90 missing 0.90 14405777