Bulked Segregant Analysis For Fine Mapping Of Genes

Cheng Zou, Qi Sun

Bioinformatics Facility

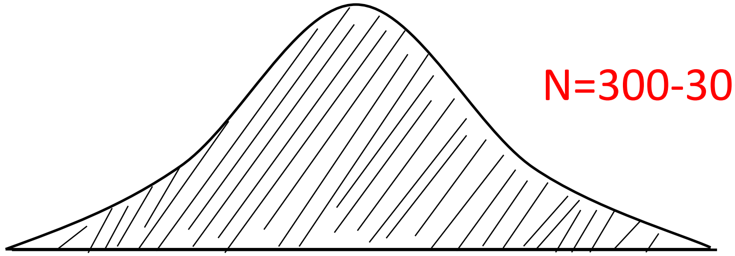
Cornell University

Outline

- What is BSA?
- Keys for a successful BSA study
- Pipeline of BSA
- extended reading

Compare BSA with traditional mapping strategy

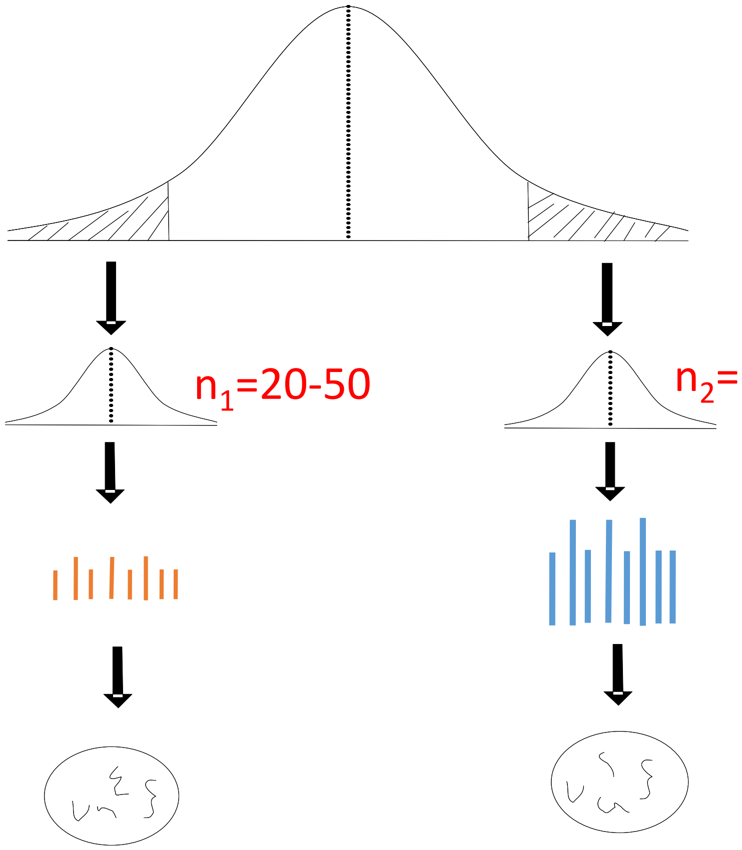
Entire population (all individual) analysis



N=300-3000

GWAS or linkage mapping

Phenotyping	entire population
Genotyping	entire population



Pool the samples with extreme phenotype

n_1=20-50

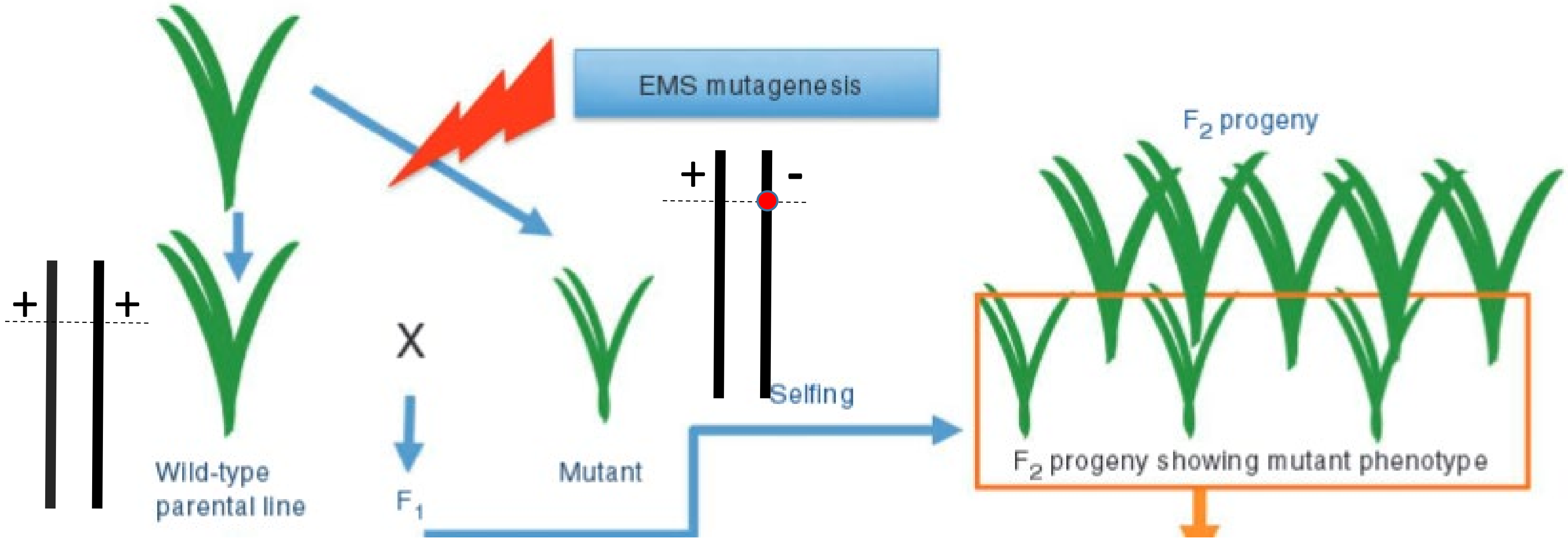
n_2=20-50

Phenotyping	entire population
Genotyping	two samples

Bulked Segregant Analysis (BSA)

rapid discovery of genetic markers and trait mapping

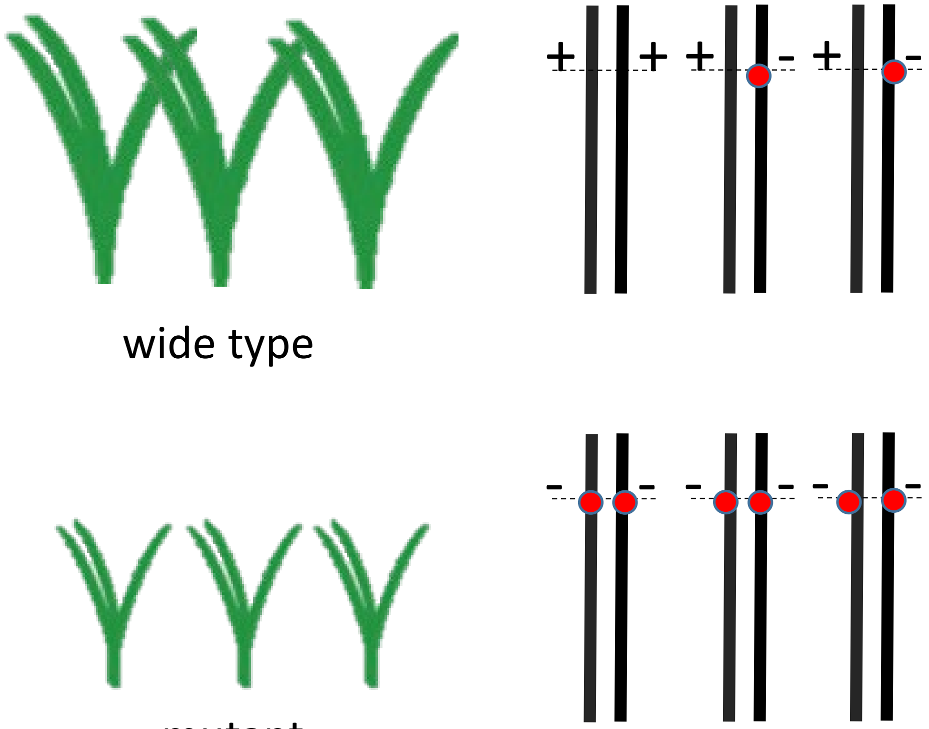
1. Segregation in phenotype



Bulked Segregant Analysis (BSA)

rapid discovery of genetic markers and trait mapping

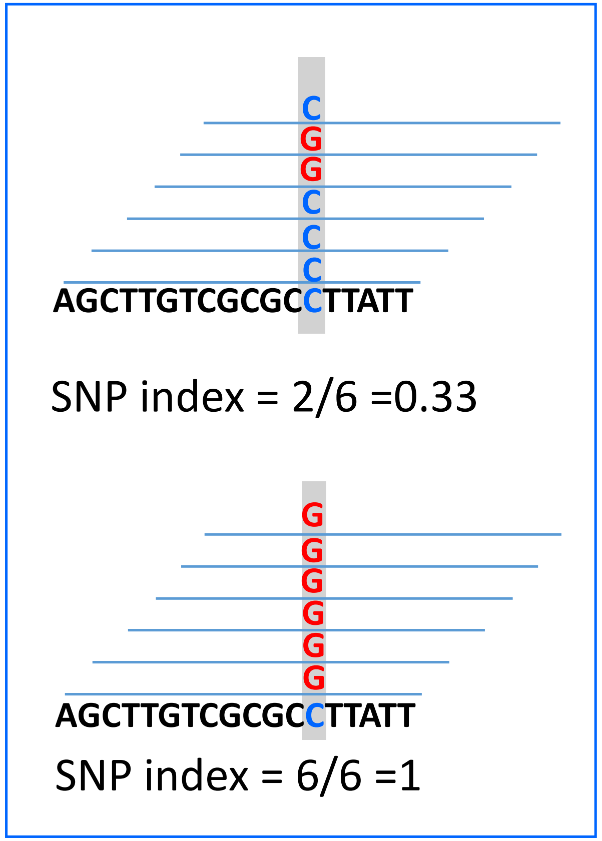
2. Difference in allele frequency



wide type

mutant

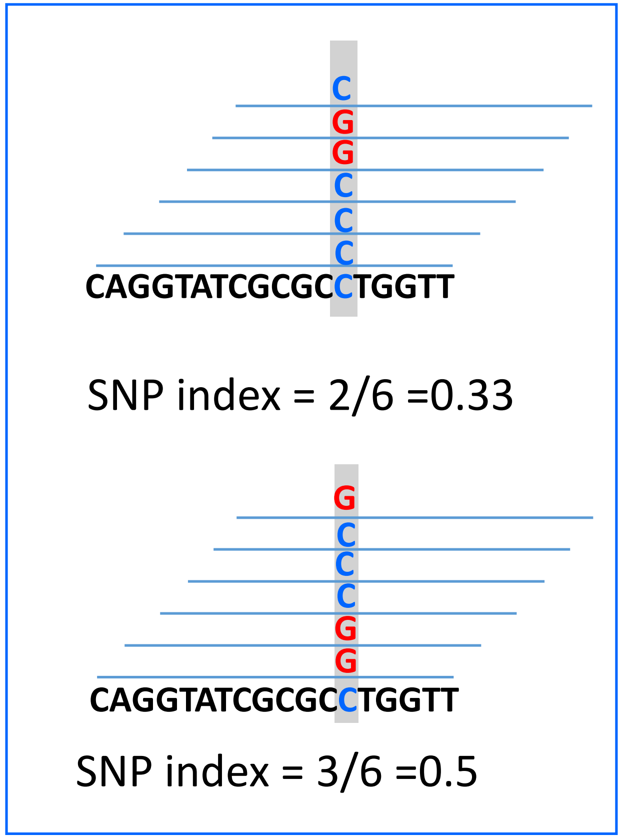
Linked sites



SNP index = 2/6 =0.33

SNP index = 6/6 =1

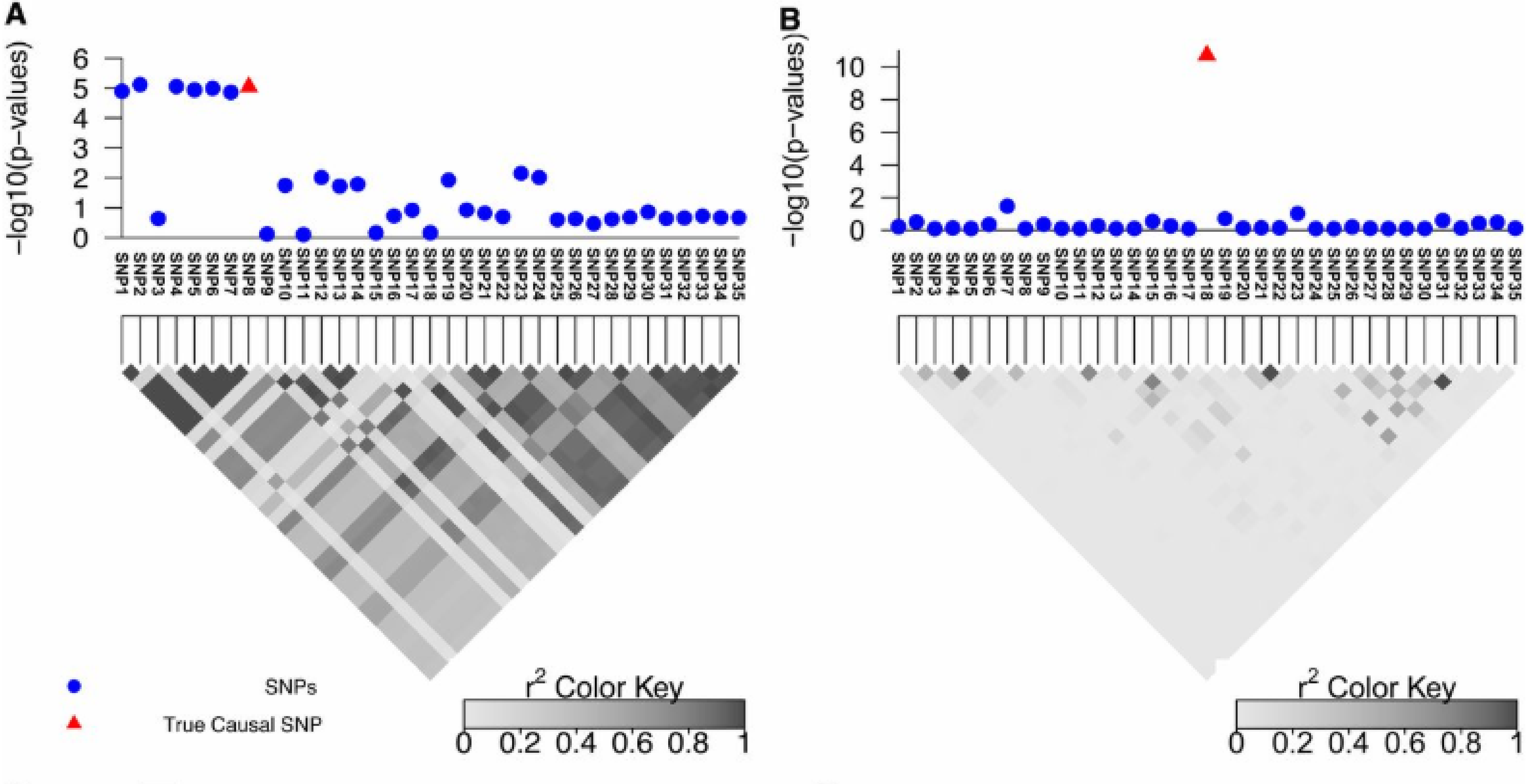
unlinked sites



SNP index = 2/6 =0.33

SNP index = 3/6 =0.5

Causal SNP and SNPs linked with causal SNP

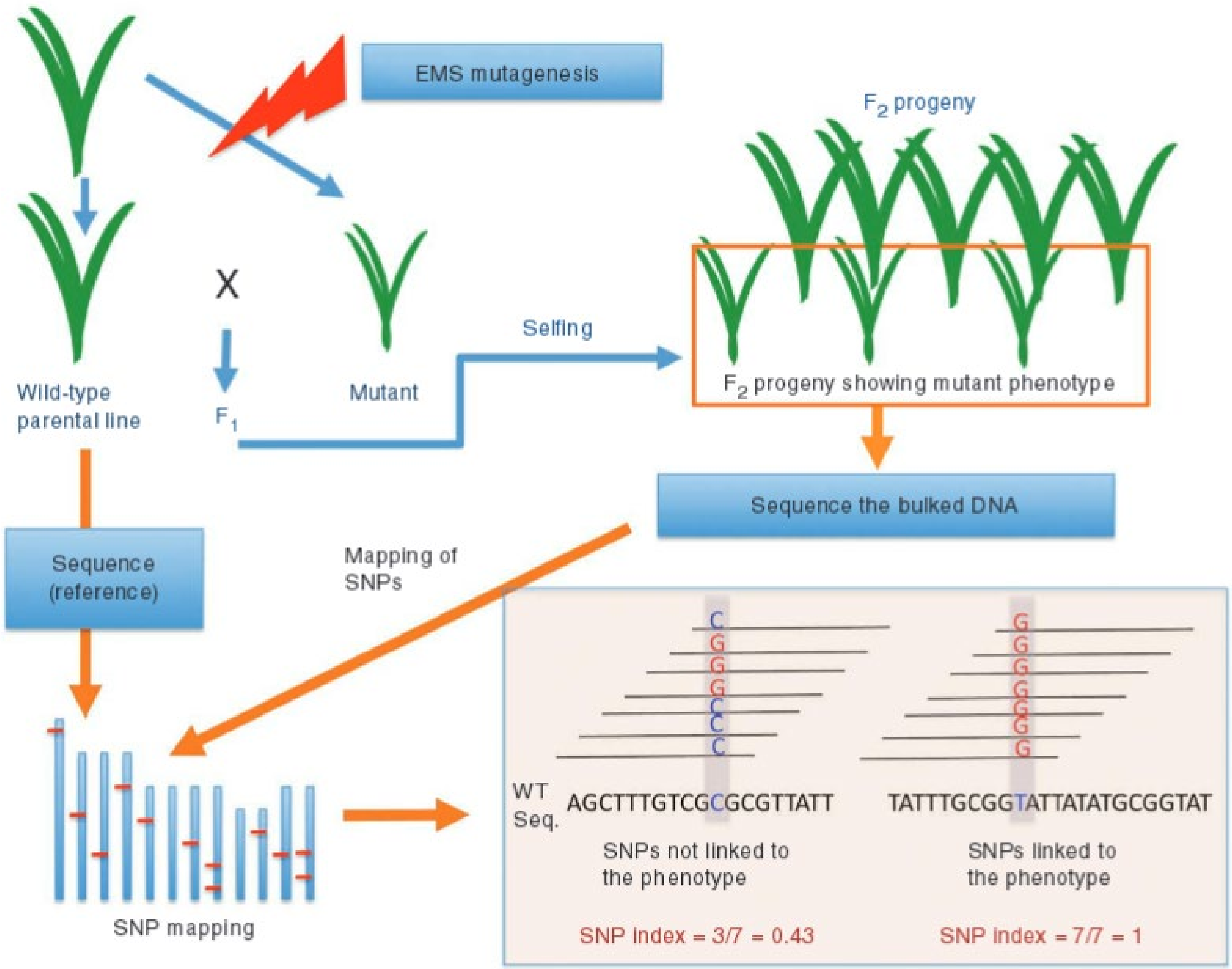


(copy from Hormozdiari, Farhad, et al *Genetics*, 2014)

Applicable populations

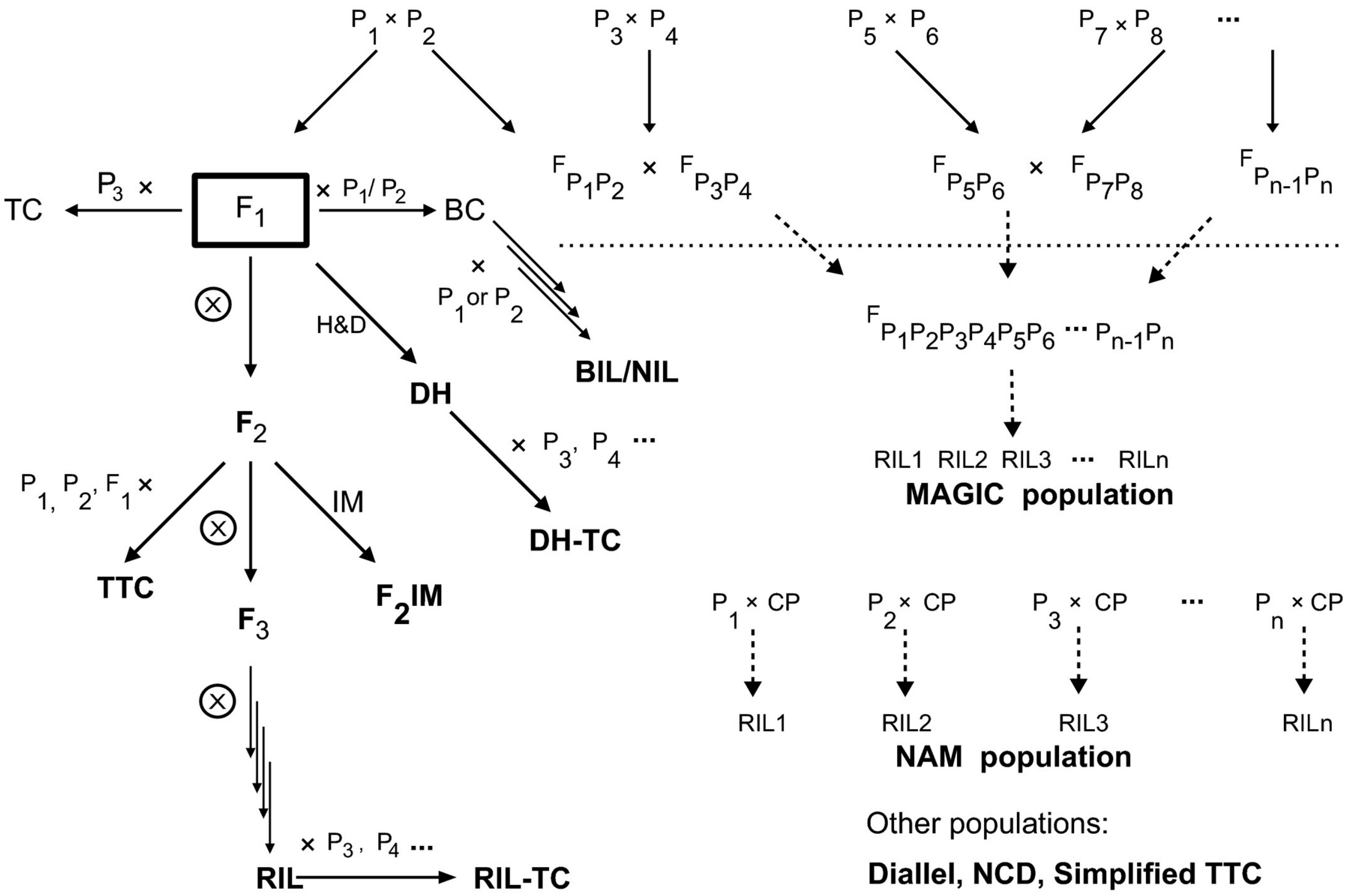
- EMS mutagenized population
- Mapping Population
- Nature Population

EMS mutagenized population



(Abe, 2011, NBT)

Examples of Mapping Populations



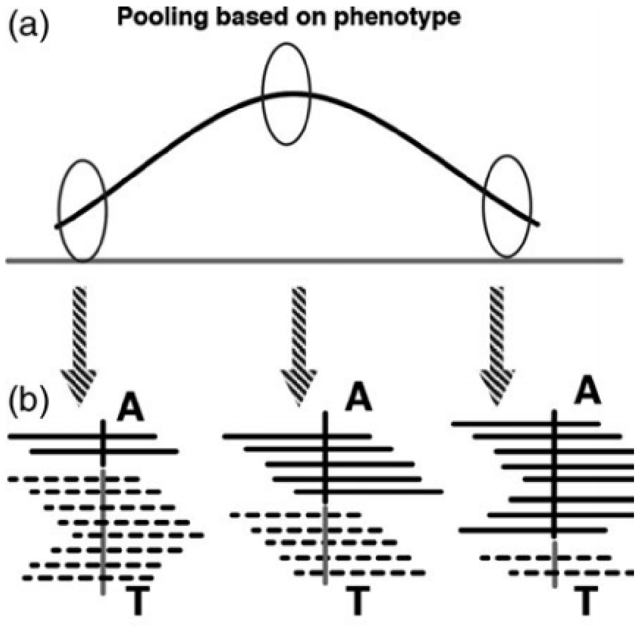
Other populations:

Diallel, NCD, Simplified TTC

(Zou,2016 the Plant Biotechnol J)

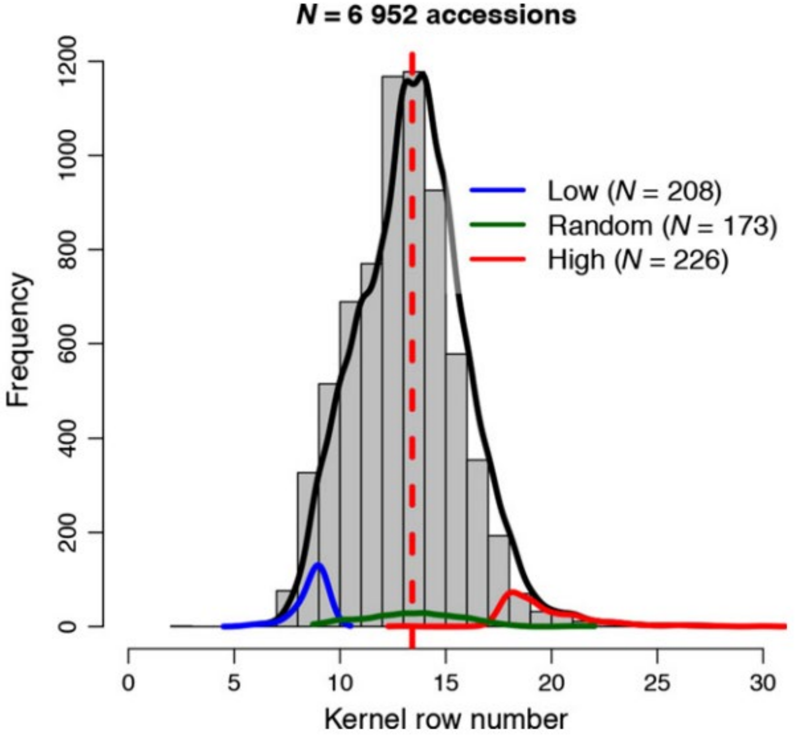
Extreme-phenotype GWAS using pooled samples

1. complex genetic architecture of the trait.
2. complex genetic background and population structure



(c)

Variant	Low	Random	High
Ref(A)	20	50	80
Alt(T)	80	50	20
Total	100	100	100



(Schnable,2015 the Plant Journal)

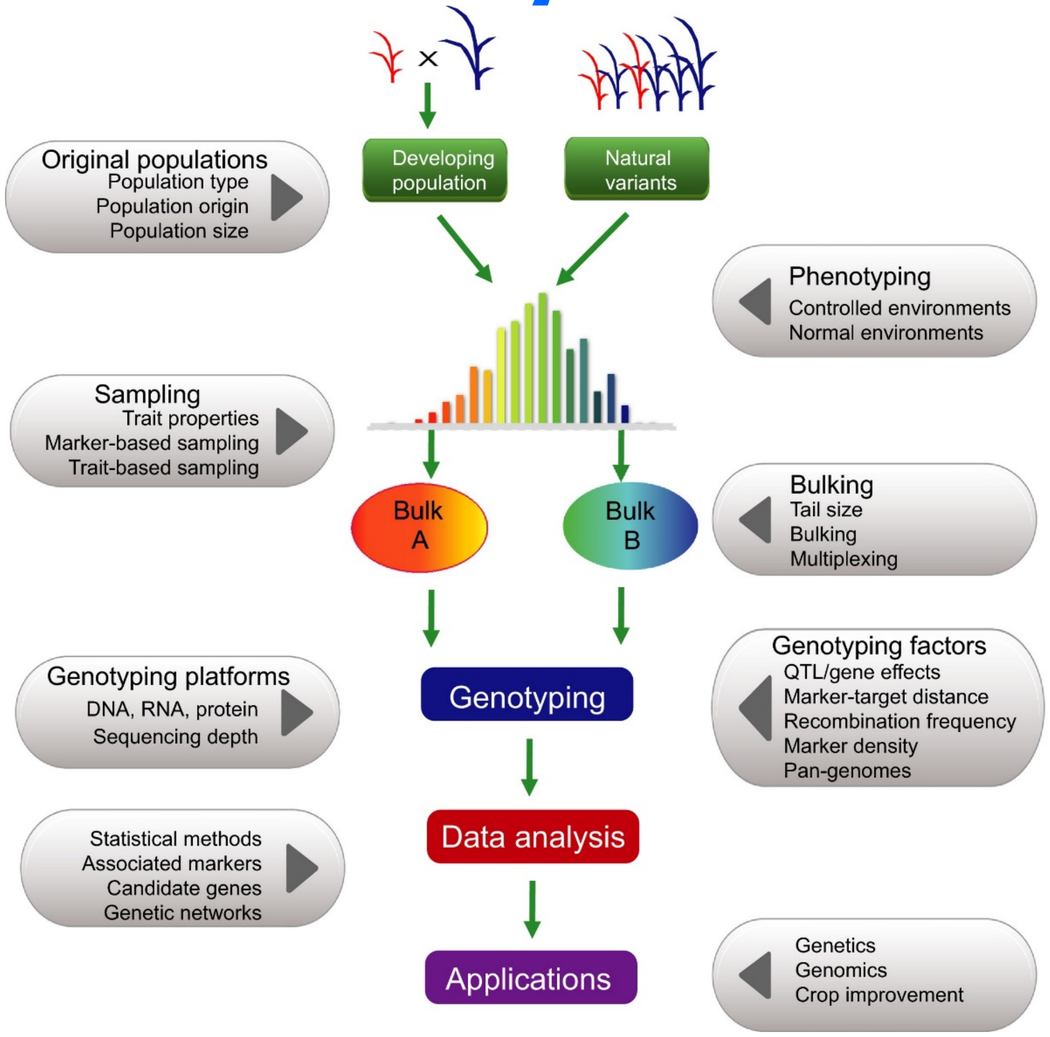
Applicable genotyping platform

- Whole genome sequencing
    - High depth sequencing of each bulk (30 ~ 50 X is recommended)

- RNA-seq –based bulk segregant analysis

Checklist for a successful BSA study

- 1. Genetic architecture and the phenotypic segregation
- 2. Population size, bulk size
- 3. Sequencing depth



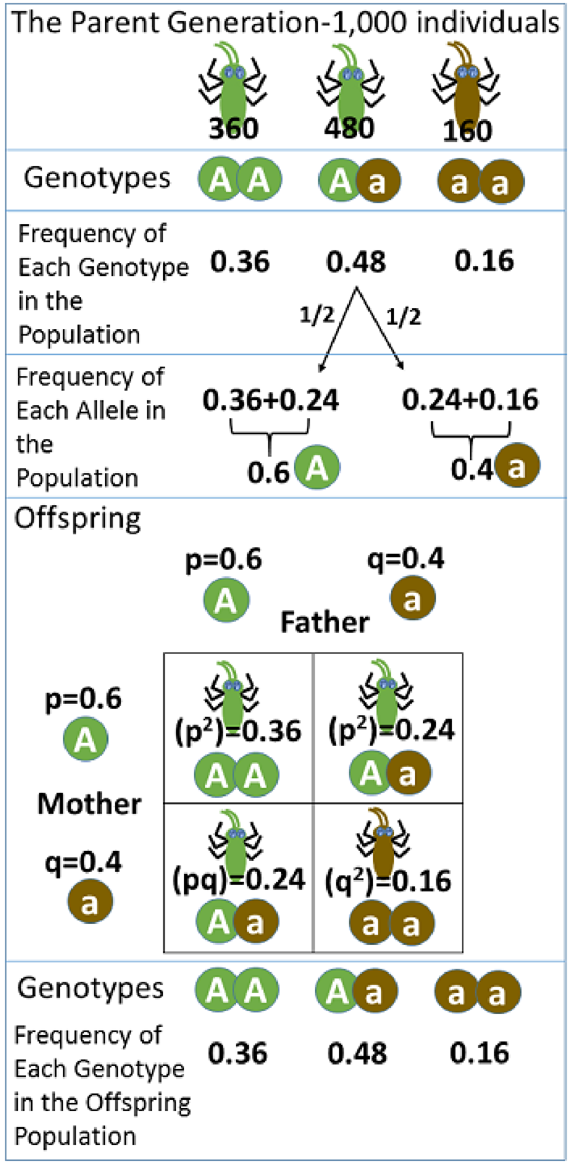
Beware of Variance Callings

Assumptions in Variant callers

for example GATK :

- **assuming Hardy-Weinberg equilibrium**
- **diploidy**

Using read depth directly, not allele calling



The Parent Generation-1,000 individuals

360 480 160

Genotypes A A | A a | a a

Frequency of Each Genotype in the Population 0.36 0.48 0.16

1/2 1/2

Frequency of Each Allele in the Population 0.36+0.24 0.24+0.16

0.6 A 0.4 a

Offspring

p=0.6 A q=0.4 a Father

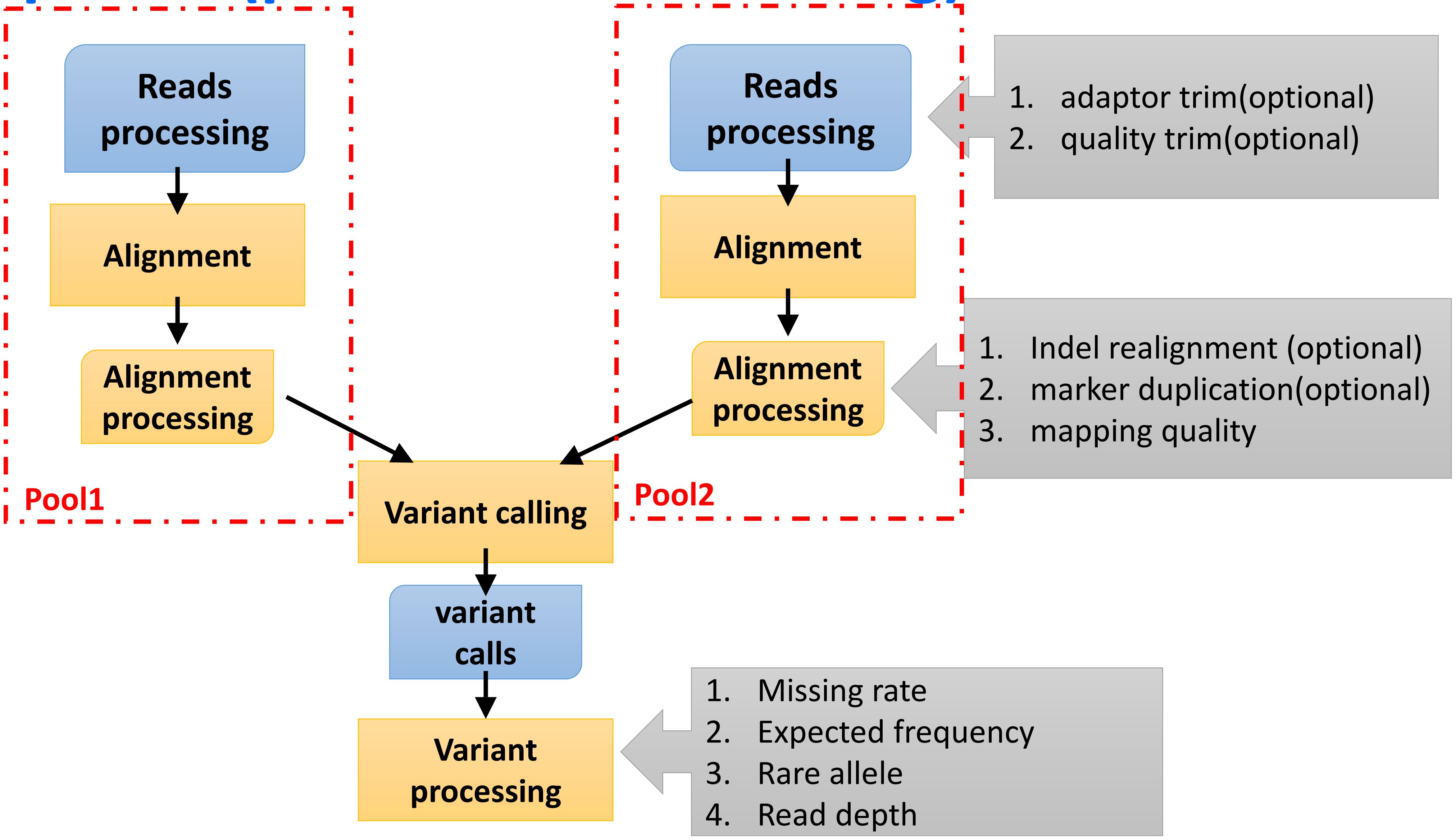
p=0.6 A Mother q=0.4 a

	(p^2)=0.36 A A	(p^2)=0.24 A a
	(pq)=0.24 A a	(q^2)=0.16 a a

Genotypes A A | A a | a a

Frequency of Each Genotype in the Offspring Population 0.36 0.48 0.16

BSA Pipeline (part 1 variants calling)



Pool1

Reads processing

Alignment

Alignment processing

Pool2

Reads processing

Alignment

Alignment processing

Variant calling

variant calls

Variant processing

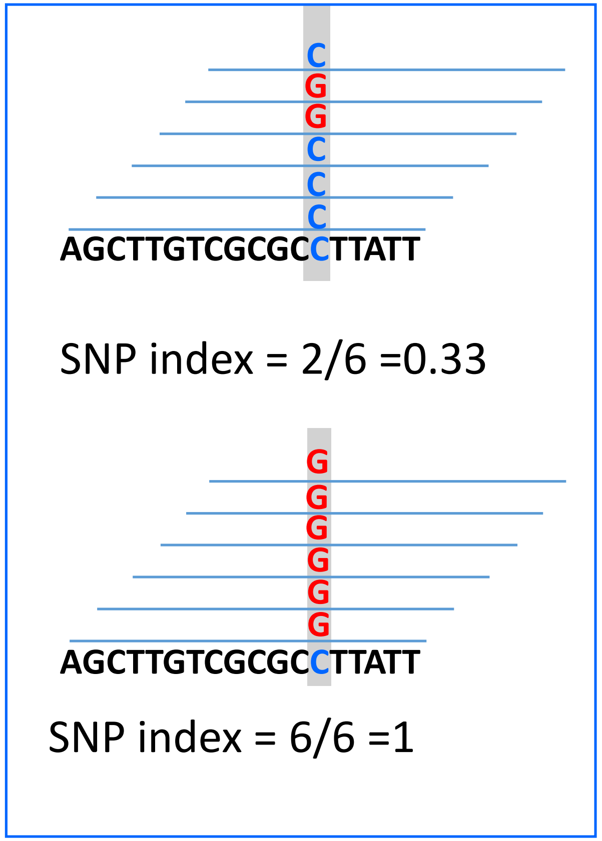
1. adaptor trim(optional)
2. quality trim(optional)

1. Indel realignment (optional)
2. marker duplication(optional)
3. mapping quality

1. Missing rate
2. Expected frequency
3. Rare allele
4. Read depth

BSA Pipeline (part 2 Statistics and sliding window)

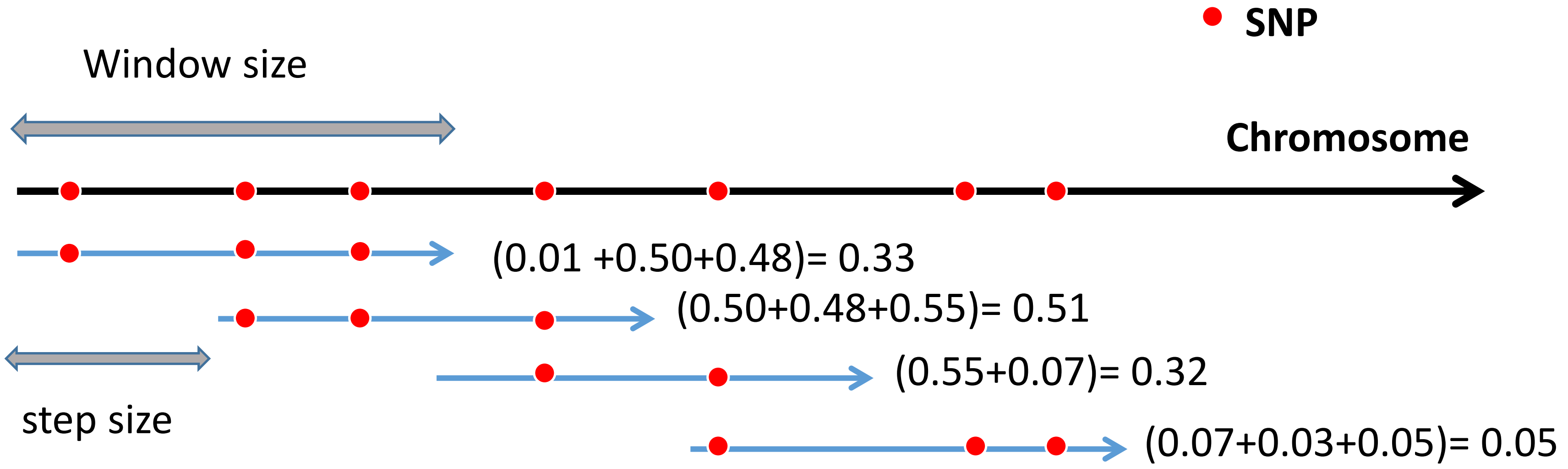
Linked sites



SNP index = 2/6 =0.33

SNP index = 6/6 =1

△SNP index = abs(1-0.33)=0.67

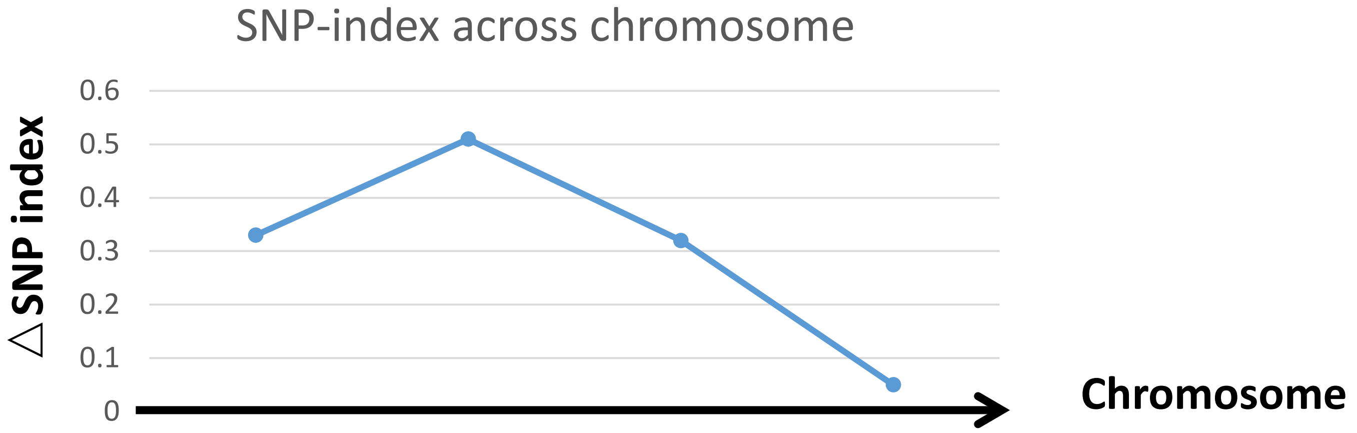


(0.01 +0.50+0.48)= 0.33

(0.50+0.48+0.55)= 0.51

(0.55+0.07)= 0.32

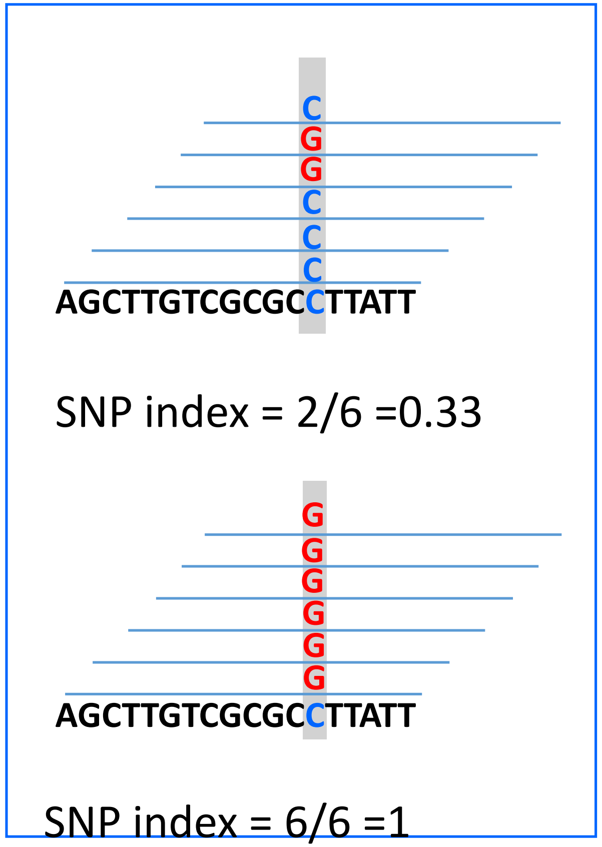
(0.07+0.03+0.05)= 0.05


SNP-index across chromosome

Method 2. fishier exact test

- 2. Compare fishier exact test to test if the read depth in each buck are significantly different or not.

Linked sites



SNP index = 2/6 =0.33

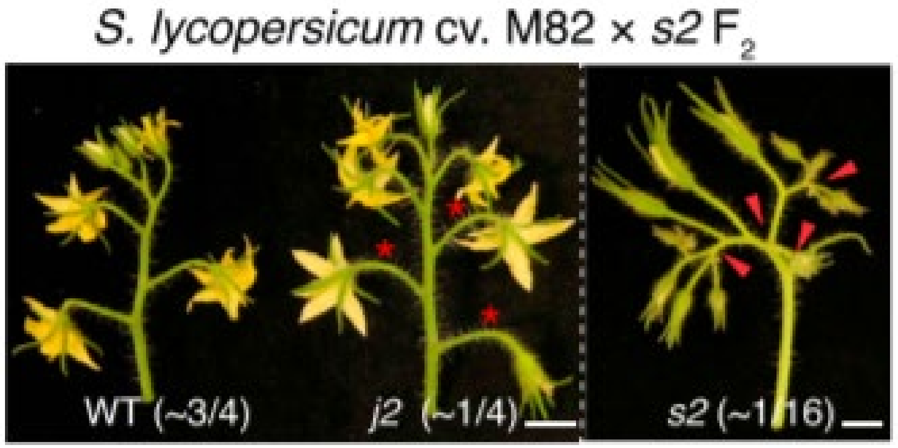
SNP index = 6/6 =1

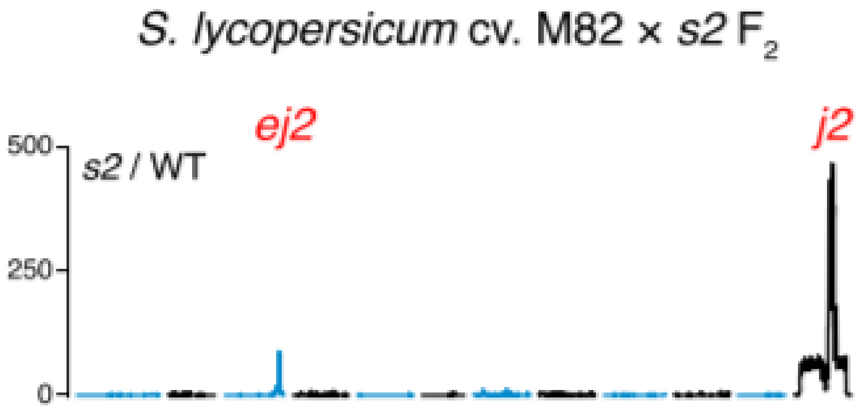
	Ref allele	Alt allele	Row total
WT	4	2	6
Mutant	0	6	6
Column total	4	8	12

$$p=\frac{\binom{6}{4}\binom{6}{0}}{\binom{12}{4}}$$

```
F=fisher.test(rbind(c(4,2),c(0,6)),
         alternative=“two.sided")
F$p.value
 0.06061
```

An exercise of BSA





http://dx.doi.org/10.1016/j.cell.2017.04.032

Download reads and reference genome

The Sequence Read Archive (SRA) on NCBI is the most commonly used website to store the high-throughput sequencing data.

- fastq-dump --split-files --gzip SRR5274882
- fastq-dump --split-files --gzip SRR5274880

- wget ftp://ftp.ensemblgenomes.org/pub/plants/release-35/fasta/solanum_lycopersicum/dna/Solanum_lycopersicum.SL2.50.dna.toplevel.fa.gz

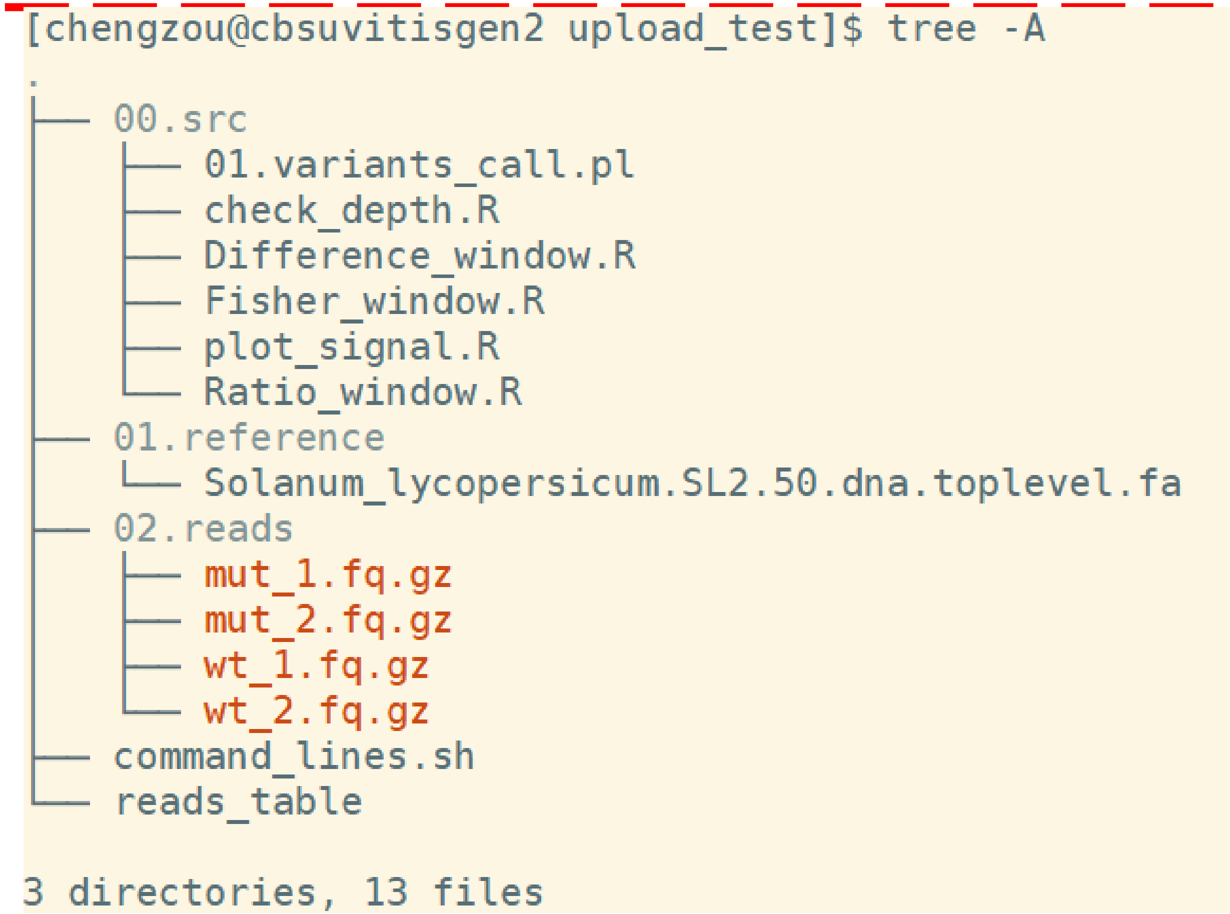
Do not run. Data has been downloaded.

To speed up the calculations, the data has been down-sampled using reads that were mapped to chr3 only in the test data. If you are interested in testing the entire data, you can download it from NCBI.

Copy the data under your directory

```
cp -r /shared_data/BSA_workshop_2018/*   ./

tree -A
```



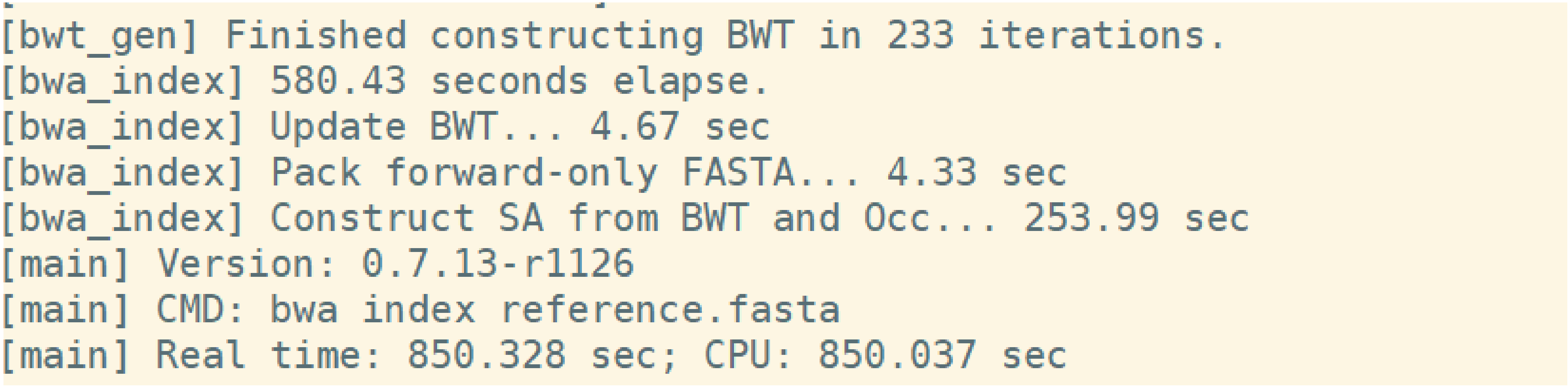
```
[chengzou@cbsuvitisgen2 upload_test]$ tree -A
.
├── 00.src
│   ├── 01.variants_call.pl
│   ├── check_depth.R
│   ├── Difference_window.R
│   ├── Fisher_window.R
│   ├── plot_signal.R
│   └── Ratio_window.R
├── 01.reference
│   └── Solanum_lycopersicum.SL2.50.dna.toplevel.fa
├── 02.reads
│   ├── mut_1.fq.gz
│   ├── mut_2.fq.gz
│   ├── wt_1.fq.gz
│   └── wt_2.fq.gz
├── command_lines.sh
└── reads_table

3 directories, 13 files
```

Index the genome

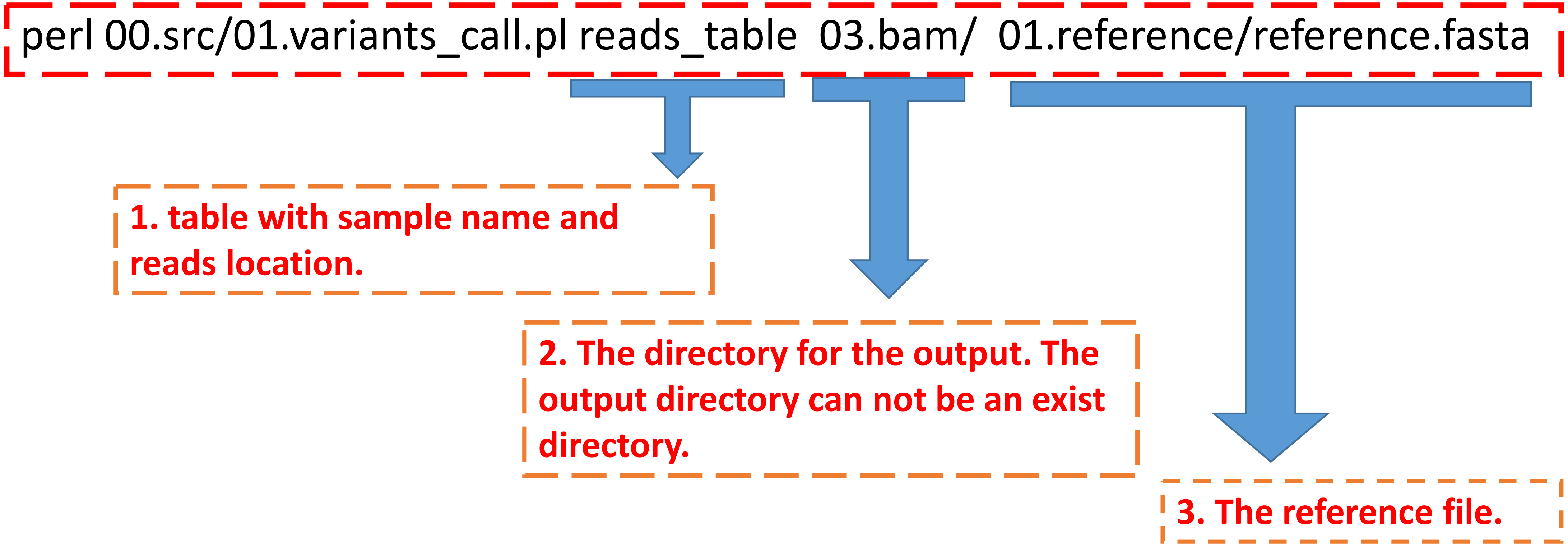
```
cd 01.reference
ln -s Solanum_lycopersicum.SL2.50.dna.toplevel.fa reference.fasta
bwa index reference.fasta
java -jar /programs/picard-tools-2.9.0/picard.jar
CreateSequenceDictionary R=reference.fasta
samtools faidx reference.fasta
```

It takes about ten minutes to finish



```
[bwt_gen] Finished constructing BWT in 233 iterations.
[bwa_index] 580.43 seconds elapse.
[bwa_index] Update BWT... 4.67 sec
[bwa_index] Pack forward-only FASTA... 4.33 sec
[bwa_index] Construct SA from BWT and Occ... 253.99 sec
[main] Version: 0.7.13-r1126
[main] CMD: bwa index reference.fasta
[main] Real time: 850.328 sec; CPU: 850.037 sec
```

Variance calling



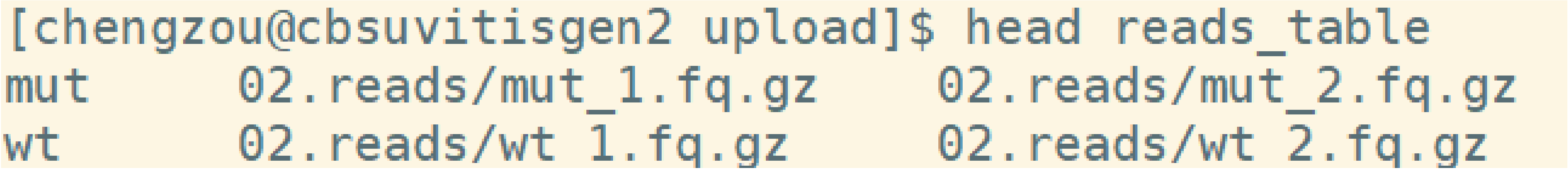
```
perl 00.src/01.variants_call.pl reads_table  03.bam/  01.reference/reference.fasta
```

1. table with sample name and reads location.

2. The directory for the output. The output directory can not be an exist directory.

3. The reference file.

Reads_table is a tab delimited txt file



```
[chengzou@cbsuvitisgen2 upload]$ head reads_table
mut     02.reads/mut_1.fq.gz    02.reads/mut_2.fq.gz
wt      02.reads/wt 1.fq.gz     02.reads/wt 2.fq.gz
```

Step 1: Align the reads, sort and index the results

```
bwa mem -t 8  -M -R '@RG\tID:mut\tSM:mut' 01.reference/reference.fasta
03.bam /fixed6.mut_1.fq.gz 04.bam/fixed6.mut_2.fq.gz  | samtools sort -@ 8 -o
03.bam /mut.sorted.bam -  2>> 03.bam/bwalog
java -jar /programs/picard-tools-2.9.0/picard.jar BuildBamIndex INPUT= 03.bam
/mut.sorted.redup.bam QUIET=true VERBOSITY=ERROR
```

```
bwa mem -t 8  -M -R '@RG\tID:wt\tSM:wt' 01.reference/reference.fasta 03.bam
/ fixed.wt_1.fq.gz 04.bam/fixed.wt_2.fq.gz  | samtools sort -@ 8 -o
03.bam/wt.sorted.bam -  2>> 03.bam//bwalog
java -jar /programs/picard-tools-2.9.0/picard.jar BuildBamIndex
INPUT=04.10bam//wt.sorted.redup.bam QUIET=true VERBOSITY=ERROR
```

-M : mark shorter split hits as secondary ***(for Picard compatibility).***

Step 2: Filtering the alignments, mpileup and variance calling

```
samtools  mpileup -t AD,DP \
 -C 50 \
-Q 20 \
-q 40 \
-f 01.reference/reference.fasta  \
 03.bam/mut.sorted.redup.bam \
03.bam/wt.sorted.redup.bam \
  -v  \
| bcftools call --consensus-caller --variants-
only --pval-threshold 1.0 -O z  -o  Out.vcf.gz
```

-t LIST optional tags to output DP,AD,ADF,ADR,SP,INFO/AD,INFO/ADF,INFO/ADR

-C adjust mapping quality; recommended:50 (unique hit of the reads)

-Q skip bases with baseQ/BAQ smaller than INT [13]

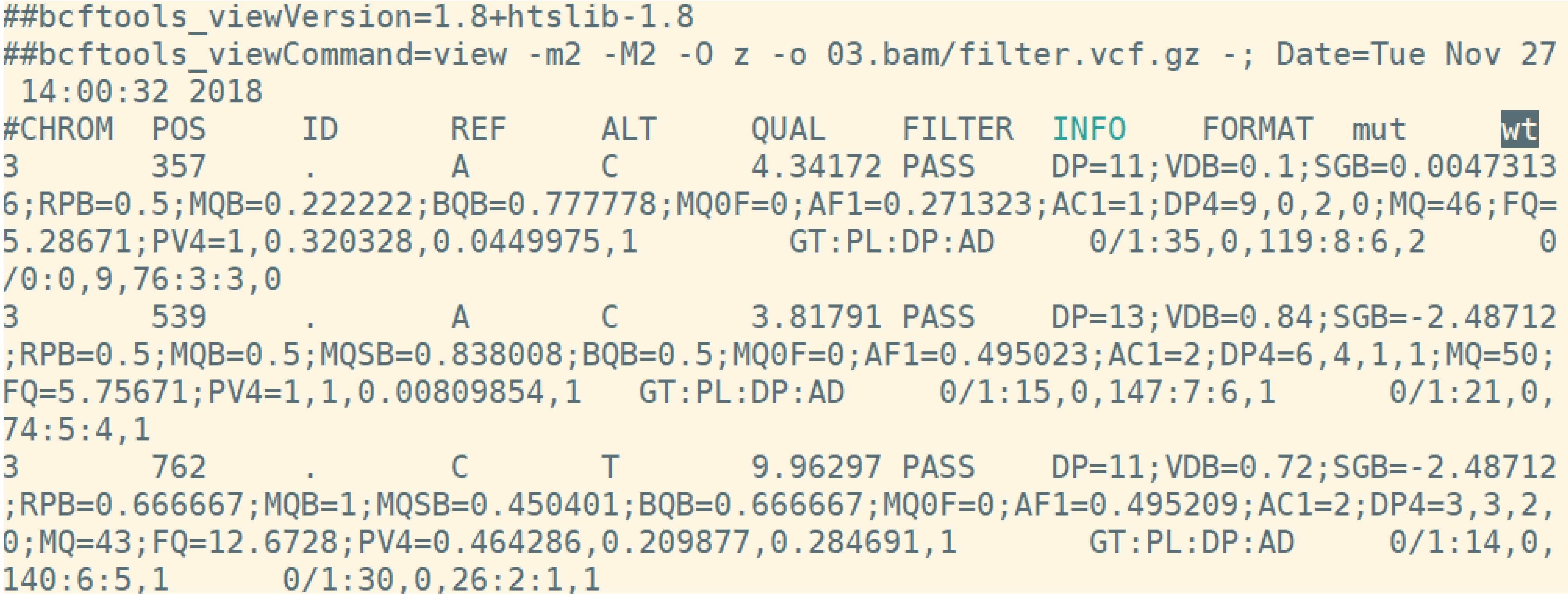
-q skip alignments with mapQ smaller than INT [0]

-f faidx indexed reference sequence file

} input bam files

-v generate genotype likelihoods in VCF format

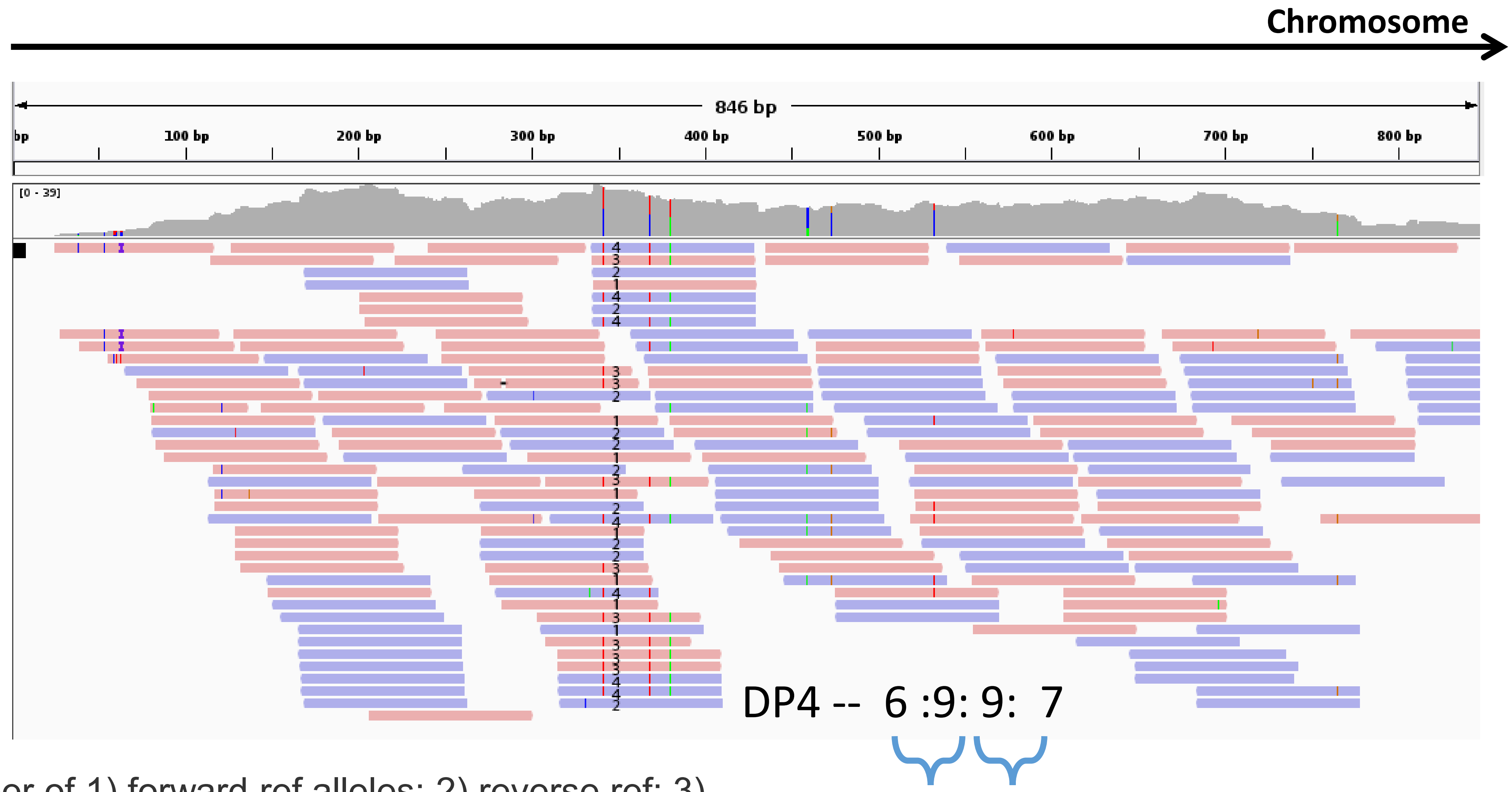
vcf file of variance calling result



```
##bcftools_viewVersion=1.8+htslib-1.8
##bcftools_viewCommand=view -m2 -M2 -O z -o 03.bam/filter.vcf.gz -; Date=Tue Nov 27
 14:00:32 2018
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  mut     wt
3       357     .       A       C       4.34172 PASS    DP=11;VDB=0.1;SGB=0.0047313
6;RPB=0.5;MQB=0.222222;BQB=0.777778;MQ0F=0;AF1=0.271323;AC1=1;DP4=9,0,2,0;MQ=46;FQ=
5.28671;PV4=1,0.320328,0.0449975,1        GT:PL:DP:AD     0/1:35,0,119:8:6,2      0
/0:0,9,76:3:3,0
3       539     .       A       C       3.81791 PASS    DP=13;VDB=0.84;SGB=-2.48712
;RPB=0.5;MQB=0.5;MQSB=0.838008;BQB=0.5;MQ0F=0;AF1=0.495023;AC1=2;DP4=6,4,1,1;MQ=50;
FQ=5.75671;PV4=1,1,0.00809854,1   GT:PL:DP:AD     0/1:15,0,147:7:6,1      0/1:21,0,
74:5:4,1
3       762     .       C       T       9.96297 PASS    DP=11;VDB=0.72;SGB=-2.48712
;RPB=0.666667;MQB=1;MQSB=0.450401;BQB=0.666667;MQ0F=0;AF1=0.495209;AC1=2;DP4=3,3,2,
0;MQ=43;FQ=12.6728;PV4=0.464286,0.209877,0.284691,1       GT:PL:DP:AD     0/1:14,0,
140:6:5,1      0/1:30,0,26:2:1,1
```

GT: Genotype
PL: list of Phred-scaled genotype likelihoods
DP: Number of high-quality bases
AD: Allelic depths

Definition of DP4 and AD



Chromosome

Bam file

DP4 -- 6 :9: 9: 7

AD -- 15: 16

DP4 is Number of 1) forward ref alleles; 2) reverse ref; 3) forward non-ref; 4) reverse non-ref alleles

Step 3: Filtering the variances

```
bcftools filter \
 -g10 \
 -G10 \
-i '(DP4[0]+DP4[1])>1 & (DP4[2]+DP4[3])>1
& FORMAT/DP[]>5'  Out.vcf.gz \
 |   bcftools view \
-m2 -M2
 -
-O z
-o 03.bam/filter.vcf.gz
```

-g filter SNPs within <int> base pairs of an indel

-G filter clusters of indels separated by <int> or fewer base pairs allowing only one to pass

-i expression of Variance that will be included:

(DP4[0]+DP4[1])>1 & (DP4[2]+DP4[3])>1
Both reference allele and alternative allele must be support by at least 2 reads.
FORMAT/DP[]>5 for each sample, there must be more than five reads covering this site.

-m2 -M2 to only view biallelic SNPs

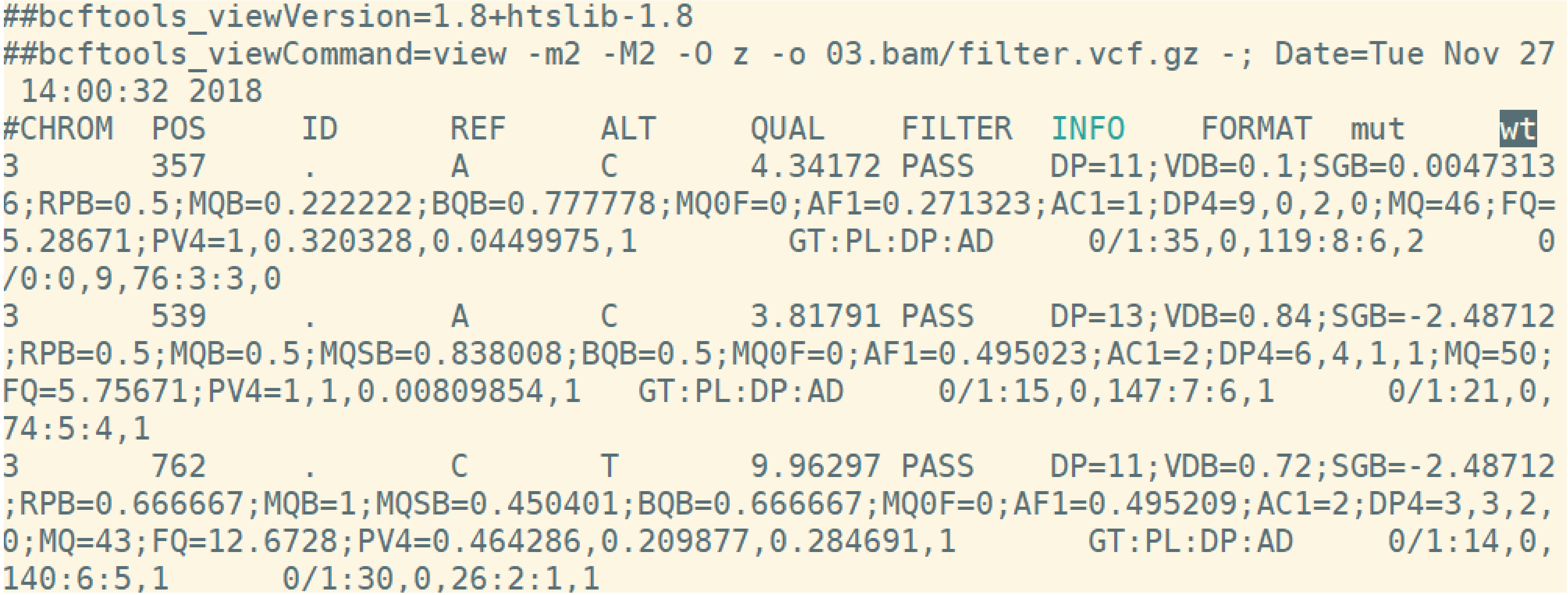
-O format of the output file

-o name of the output file

Step 4: Extract information for downstream analysis

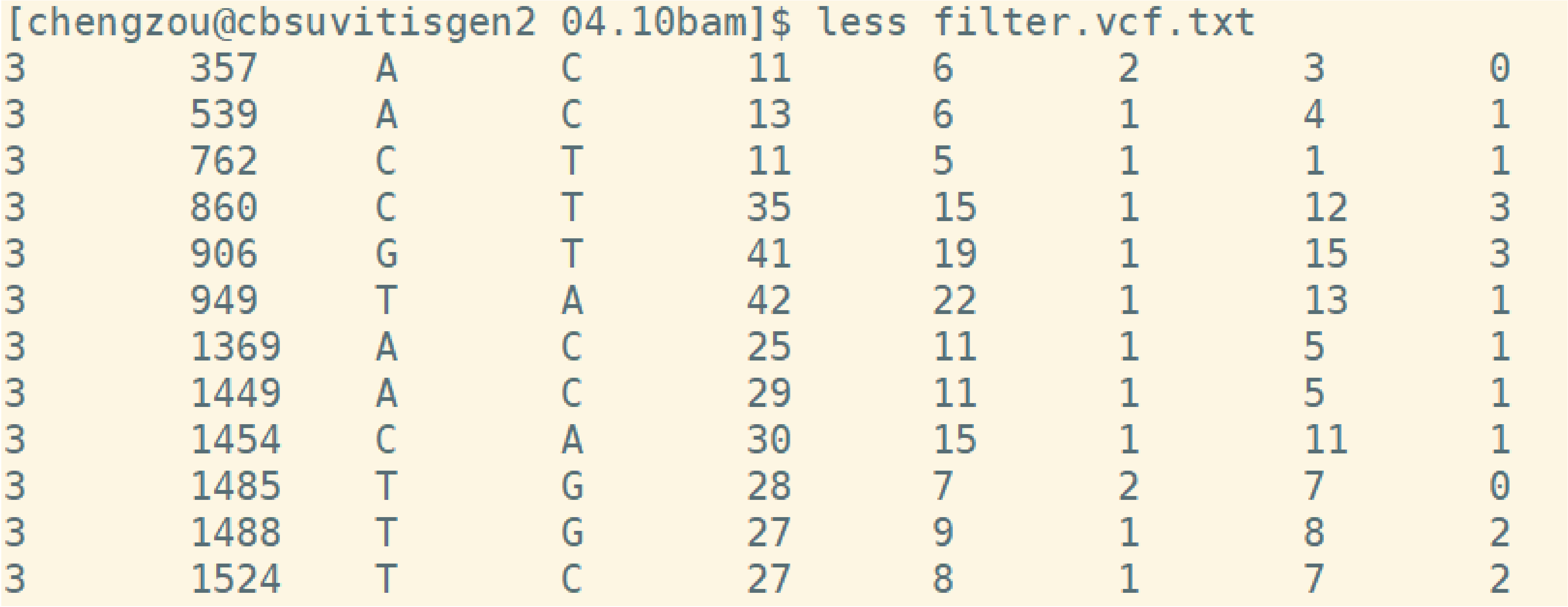
```
bcftools  query  \
 -i 'TYPE="SNP"' \
 -f '%CHROM\t%POS\t%REF\t%ALT{0}\t%DP[\t%AD]\n' \
03.bam/filter.vcf.gz | sed 's/[,]/\t/g' -
>03.bam/filter.vcf.txt
```

Final result in vcf format-- filter.vcf.gz



```
##bcftools_viewVersion=1.8+htslib-1.8
##bcftools_viewCommand=view -m2 -M2 -O z -o 03.bam/filter.vcf.gz -; Date=Tue Nov 27
 14:00:32 2018
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  mut     wt
3       357     .       A       C       4.34172 PASS    DP=11;VDB=0.1;SGB=0.0047313
6;RPB=0.5;MQB=0.222222;BQB=0.777778;MQ0F=0;AF1=0.271323;AC1=1;DP4=9,0,2,0;MQ=46;FQ=
5.28671;PV4=1,0.320328,0.0449975,1        GT:PL:DP:AD     0/1:35,0,119:8:6,2      0
/0:0,9,76:3:3,0
3       539     .       A       C       3.81791 PASS    DP=13;VDB=0.84;SGB=-2.48712
;RPB=0.5;MQB=0.5;MQSB=0.838008;BQB=0.5;MQ0F=0;AF1=0.495023;AC1=2;DP4=6,4,1,1;MQ=50;
FQ=5.75671;PV4=1,1,0.00809854,1   GT:PL:DP:AD     0/1:15,0,147:7:6,1      0/1:21,0,
74:5:4,1
3       762     .       C       T       9.96297 PASS    DP=11;VDB=0.72;SGB=-2.48712
;RPB=0.666667;MQB=1;MQSB=0.450401;BQB=0.666667;MQ0F=0;AF1=0.495209;AC1=2;DP4=3,3,2,
0;MQ=43;FQ=12.6728;PV4=0.464286,0.209877,0.284691,1       GT:PL:DP:AD     0/1:14,0,
140:6:5,1      0/1:30,0,26:2:1,1
```

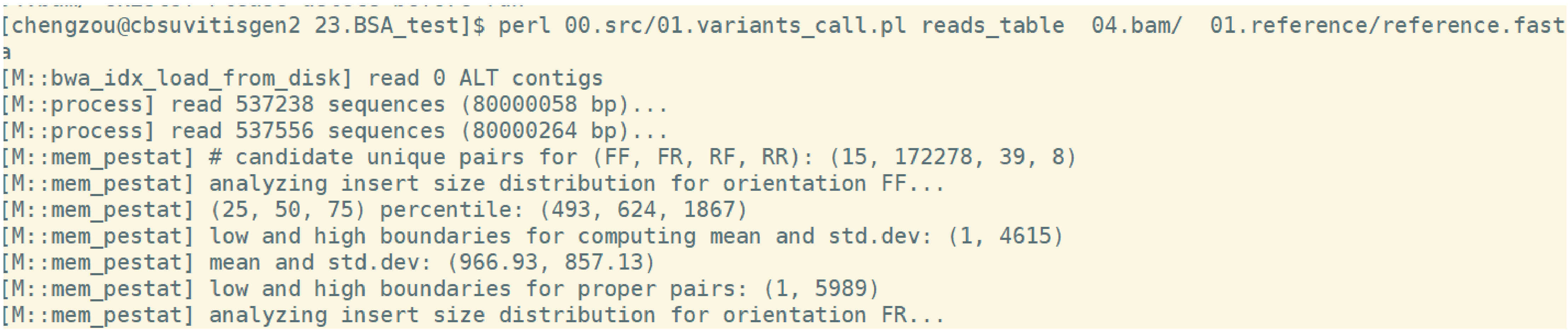
Final result in txt format -- filter.vcf.txt



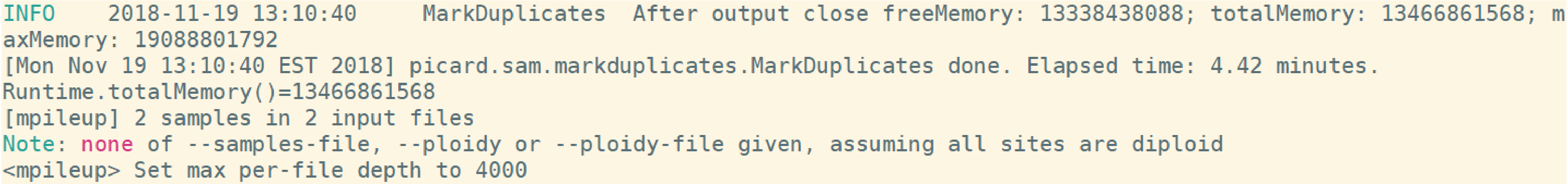
```
[chengzou@cbsuvitisgen2 04.10bam]$ less filter.vcf.txt
3       357     A       C       11      6       2       3       0
3       539     A       C       13      6       1       4       1
3       762     C       T       11      5       1       1       1
3       860     C       T       35      15      1       12      3
3       906     G       T       41      19      1       15      3
3       949     T       A       42      22      1       13      1
3       1369    A       C       25      11      1       5       1
3       1449    A       C       29      11      1       5       1
3       1454    C       A       30      15      1       11      1
3       1485    T       G       28      7       2       7       0
3       1488    T       G       27      9       1       8       2
3       1524    T       C       27      8       1       7       2
```

Chr Pos Ref Alt total DP Mut_ref Mut_alt WT_ref WT_alt

The running log

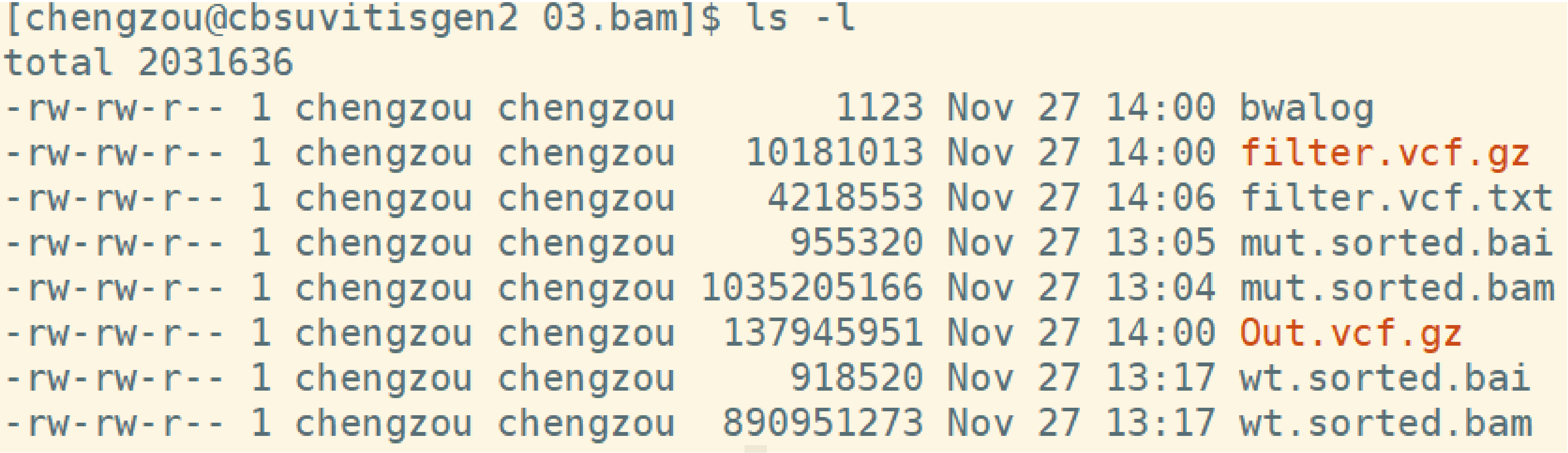


```
[chengzou@cbsuvitisgen2 23.BSA_test]$ perl 00.src/01.variants_call.pl reads_table  04.bam/  01.reference/reference.fast
a
[M::bwa_idx_load_from_disk] read 0 ALT contigs
[M::process] read 537238 sequences (80000058 bp)...
[M::process] read 537556 sequences (80000264 bp)...
[M::mem_pestat] # candidate unique pairs for (FF, FR, RF, RR): (15, 172278, 39, 8)
[M::mem_pestat] analyzing insert size distribution for orientation FF...
[M::mem_pestat] (25, 50, 75) percentile: (493, 624, 1867)
[M::mem_pestat] low and high boundaries for computing mean and std.dev: (1, 4615)
[M::mem_pestat] mean and std.dev: (966.93, 857.13)
[M::mem_pestat] low and high boundaries for proper pairs: (1, 5989)
[M::mem_pestat] analyzing insert size distribution for orientation FR...
```



```
INFO    2018-11-19 13:10:40     MarkDuplicates  After output close freeMemory: 13338438088; totalMemory: 13466861568; m
axMemory: 19088801792
[Mon Nov 19 13:10:40 EST 2018] picard.sam.markduplicates.MarkDuplicates done. Elapsed time: 4.42 minutes.
Runtime.totalMemory()=13466861568
[mpileup] 2 samples in 2 input files
Note: none of --samples-file, --ploidy or --ploidy-file given, assuming all sites are diploid
<mpileup> Set max per-file depth to 4000
```

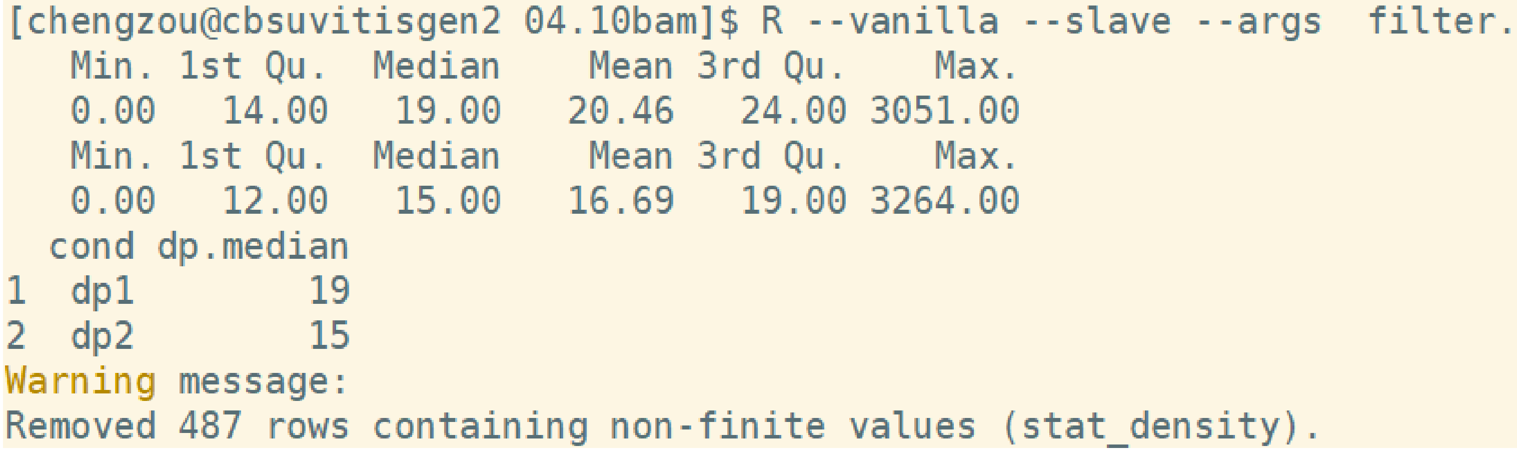
Result of the run



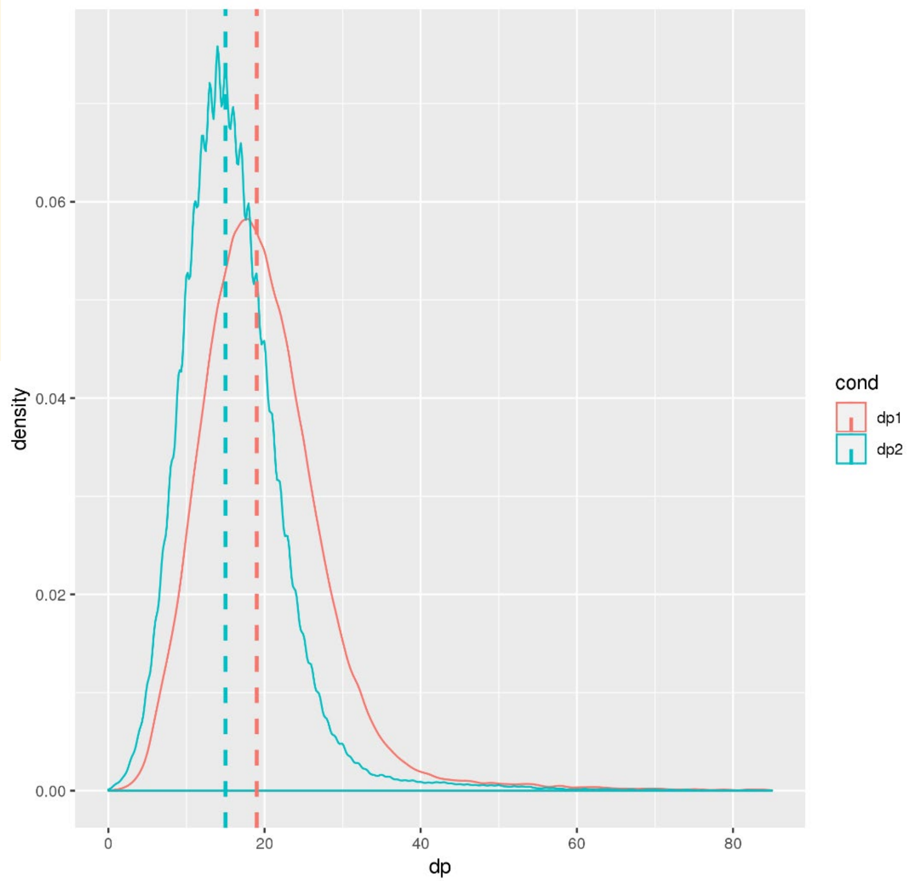
```
[chengzou@cbsuvitisgen2 03.bam]$ ls -l
total 2031636
-rw-rw-r-- 1 chengzou chengzou       1123 Nov 27 14:00 bwalog
-rw-rw-r-- 1 chengzou chengzou   10181013 Nov 27 14:00 filter.vcf.gz
-rw-rw-r-- 1 chengzou chengzou    4218553 Nov 27 14:06 filter.vcf.txt
-rw-rw-r-- 1 chengzou chengzou     955320 Nov 27 13:05 mut.sorted.bai
-rw-rw-r-- 1 chengzou chengzou 1035205166 Nov 27 13:04 mut.sorted.bam
-rw-rw-r-- 1 chengzou chengzou  137945951 Nov 27 14:00 Out.vcf.gz
-rw-rw-r-- 1 chengzou chengzou     918520 Nov 27 13:17 wt.sorted.bai
-rw-rw-r-- 1 chengzou chengzou  890951273 Nov 27 13:17 wt.sorted.bam
```

Check distribution of the depth in each pool

R --vanilla --slave --args filter.vcf.txt < ../00.src/check_depth.R

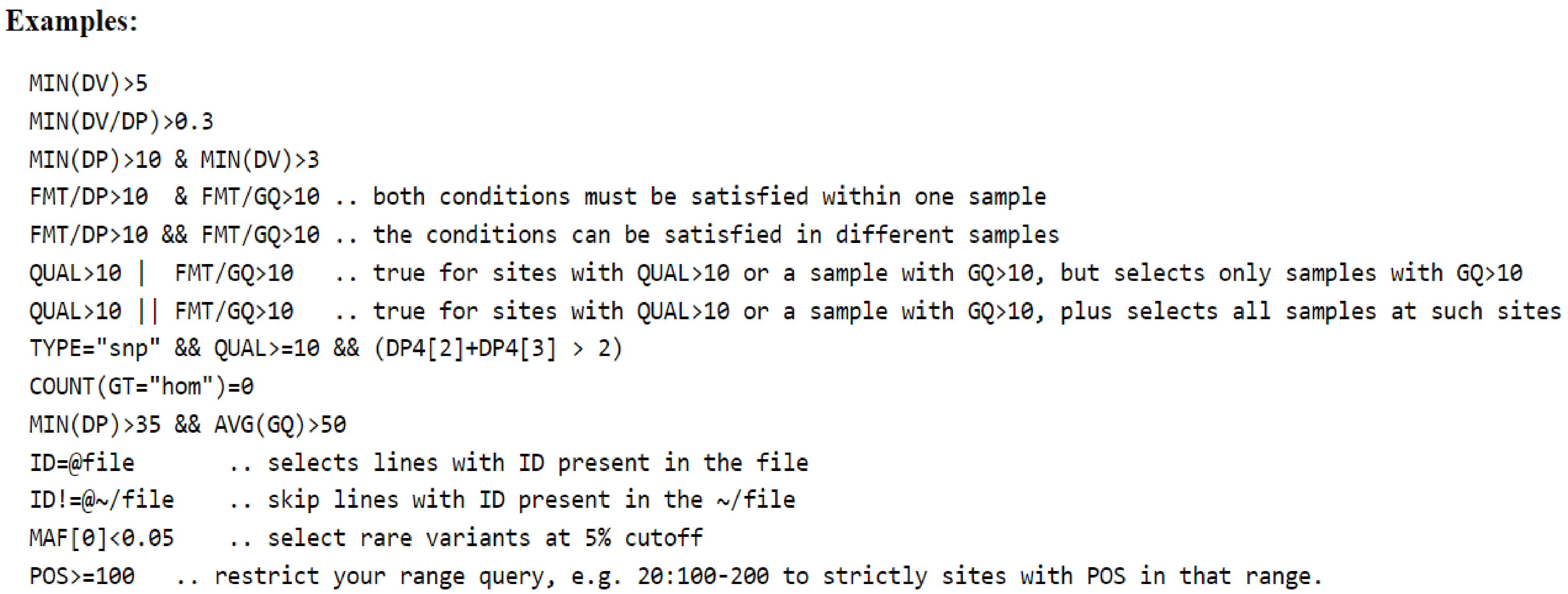


```
[chengzou@cbsuvitisgen2 04.10bam]$ R --vanilla --slave --args  filter.vcf.txt  < ../00.src/check_depth.R
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   14.00   19.00   20.46   24.00 3051.00
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.00   15.00   16.69   19.00 3264.00
  cond dp.median
1  dp1        19
2  dp2        15
Warning message:
Removed 487 rows containing non-finite values (stat_density).
```



SNP with total read depth that is larger than two times of the average is not desired.

Further filtering by depth distribution



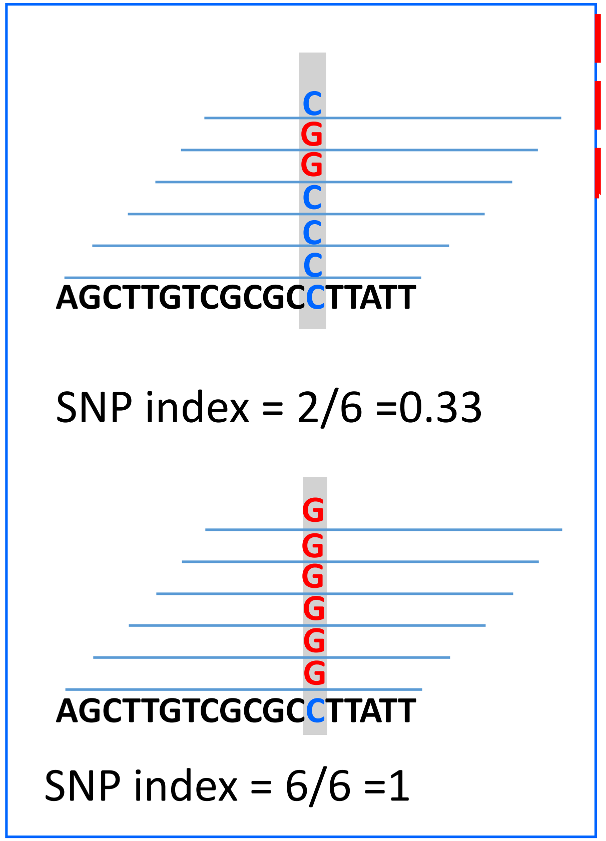
Examples:

```
MIN(DV)>5
MIN(DV/DP)>0.3
MIN(DP)>10 & MIN(DV)>3
FMT/DP>10  & FMT/GQ>10 .. both conditions must be satisfied within one sample
FMT/DP>10 && FMT/GQ>10 .. the conditions can be satisfied in different samples
QUAL>10 |  FMT/GQ>10   .. true for sites with QUAL>10 or a sample with GQ>10, but selects only samples with GQ>10
QUAL>10 || FMT/GQ>10   .. true for sites with QUAL>10 or a sample with GQ>10, plus selects all samples at such sites
TYPE="snp" && QUAL>=10 && (DP4[2]+DP4[3] > 2)
COUNT(GT="hom")=0
MIN(DP)>35 && AVG(GQ)>50
ID=@file       .. selects lines with ID present in the file
ID!=@~/file    .. skip lines with ID present in the ~/file
MAF[0]<0.05    .. select rare variants at 5% cutoff
POS>=100   .. restrict your range query, e.g. 20:100-200 to strictly sites with POS in that range.
```

```
bcftools filter  -i ‘ FORMAT/DP[1]<30 & FORMAT/DP[2]<34’ filter.vcf.gz
 -O z -o filter2.vcf.gz
```

Summary statistics 1. △SNP index

Linked sites

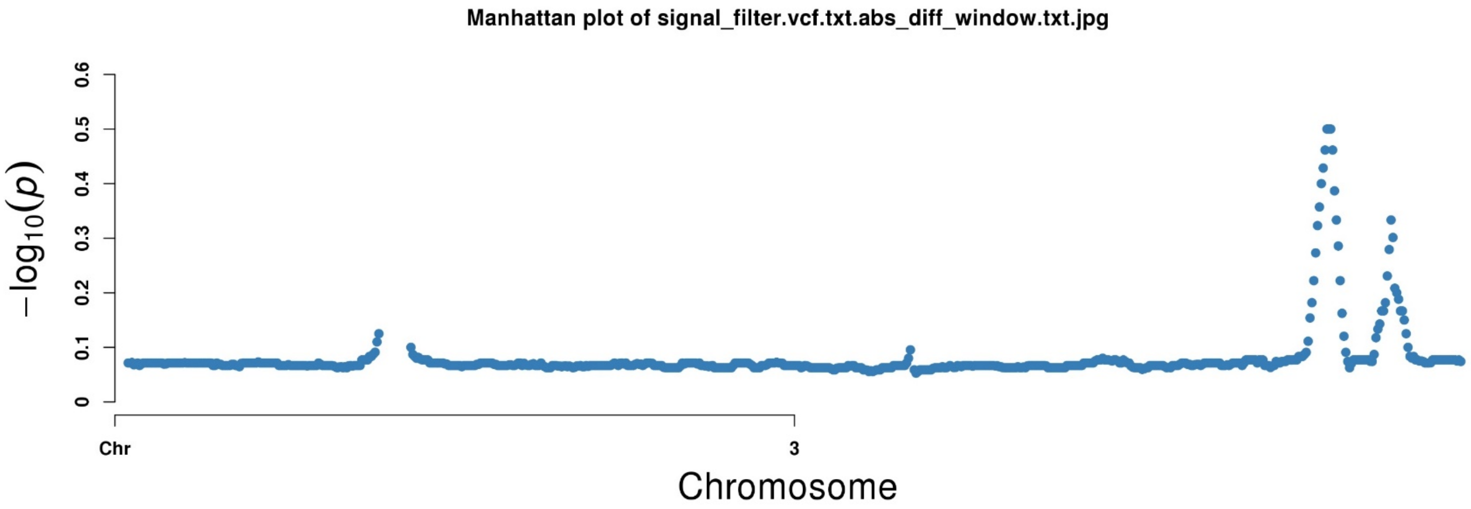


SNP index = 2/6 =0.33

SNP index = 6/6 =1

```
R --vanilla --slave --args  filter.vcf.txt  < ../00.src/Difference_window.R
R --vanilla --slave --args  filter.vcf.txt.abs_diff_window.txt
< ../00.src/plot_signal.R
```

△SNP index = abs(1-0.33)=0.67

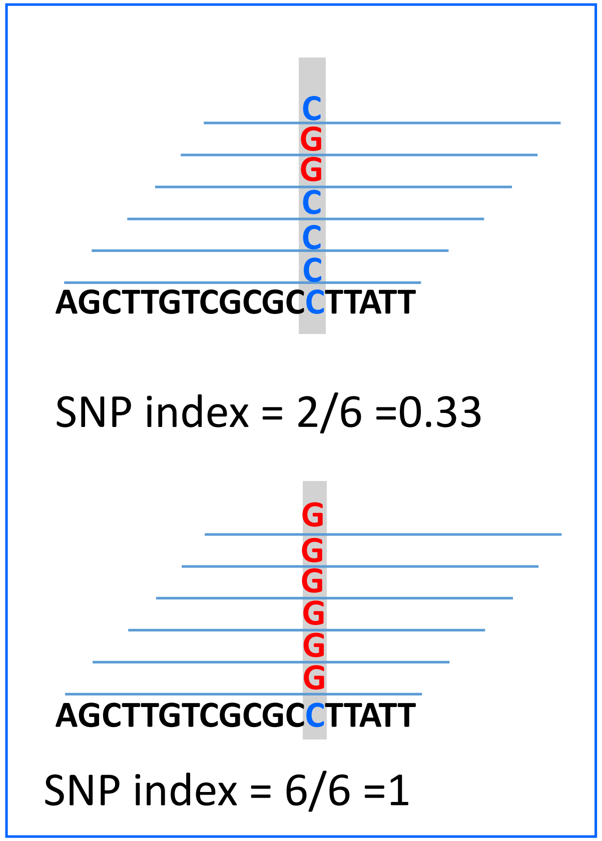

Manhattan plot of signal_filter.vcf.txt.abs_diff_window.txt.jpg

Summary statistics 2. ratio of allele frequency

- 1. Compare the ratio, and sliding window to find the peaks.

```
R --vanilla --slave --args  filter.vcf.txt  < ../00.src/Ratio_window.R
R --vanilla --slave --args  filter.vcf.txt.ratio_window.txt
< ../00.src/plot_signal.R
```

Linked sites

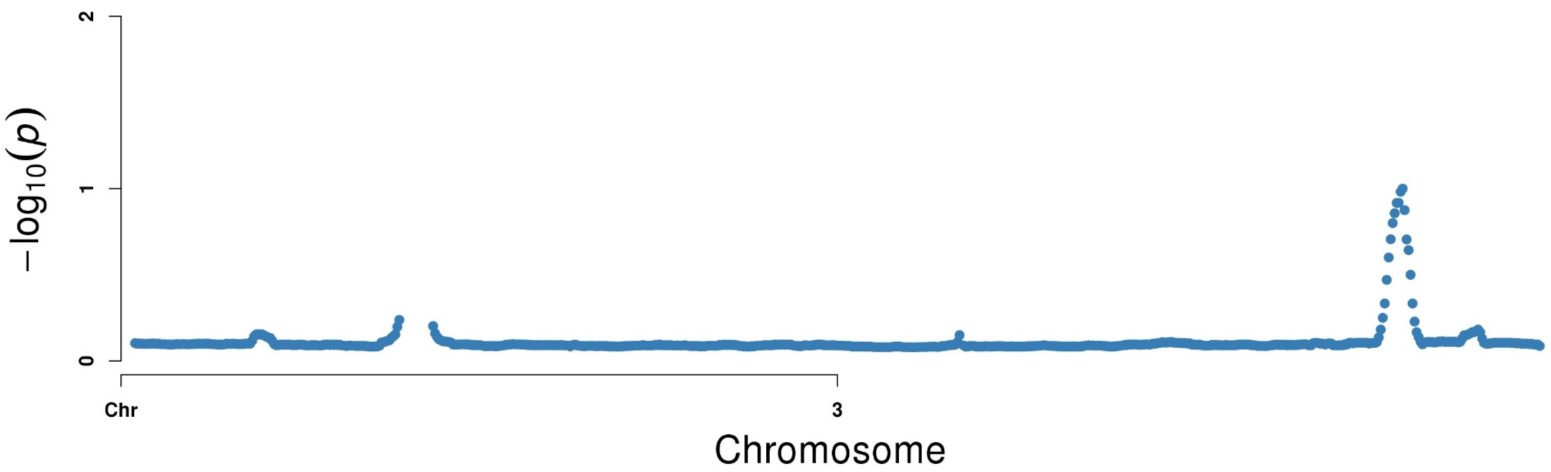


SNP index = 2/6 =0.33

SNP index = 6/6 =1

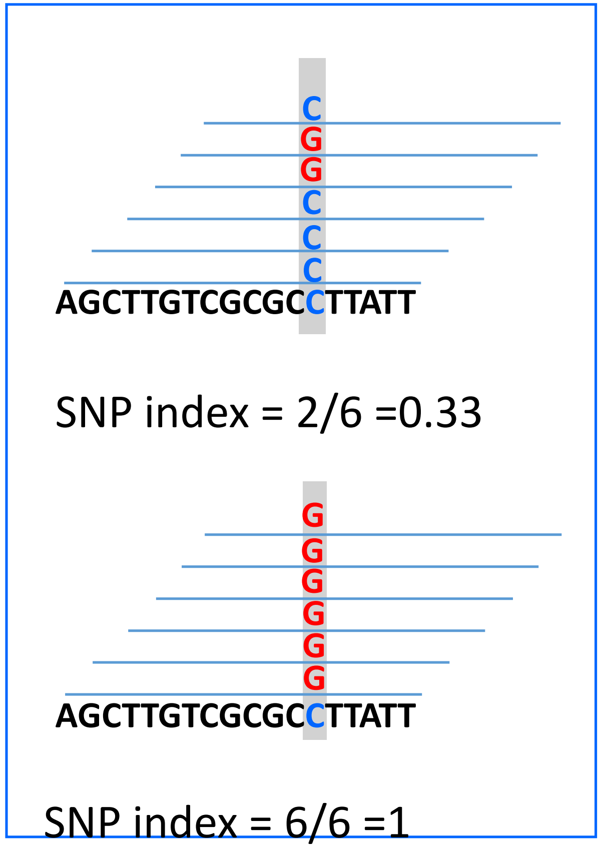
$$\text{Ratio of SNP index} = \frac{1}{0.33} = 3$$

Manhattan plot of signal_filter.vcf.txt.ratio_window.txt.jpg



Summary statistics 3. fishier exact test

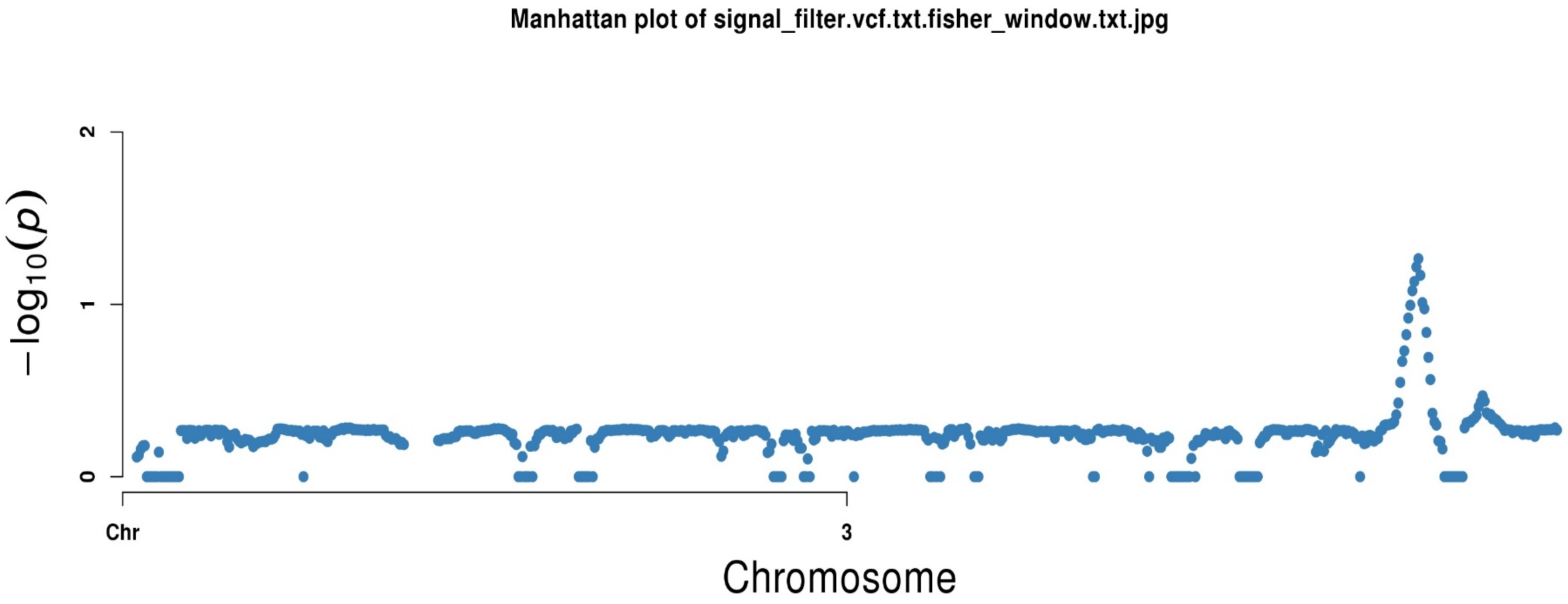
Linked sites



SNP index = 2/6 =0.33

SNP index = 6/6 =1

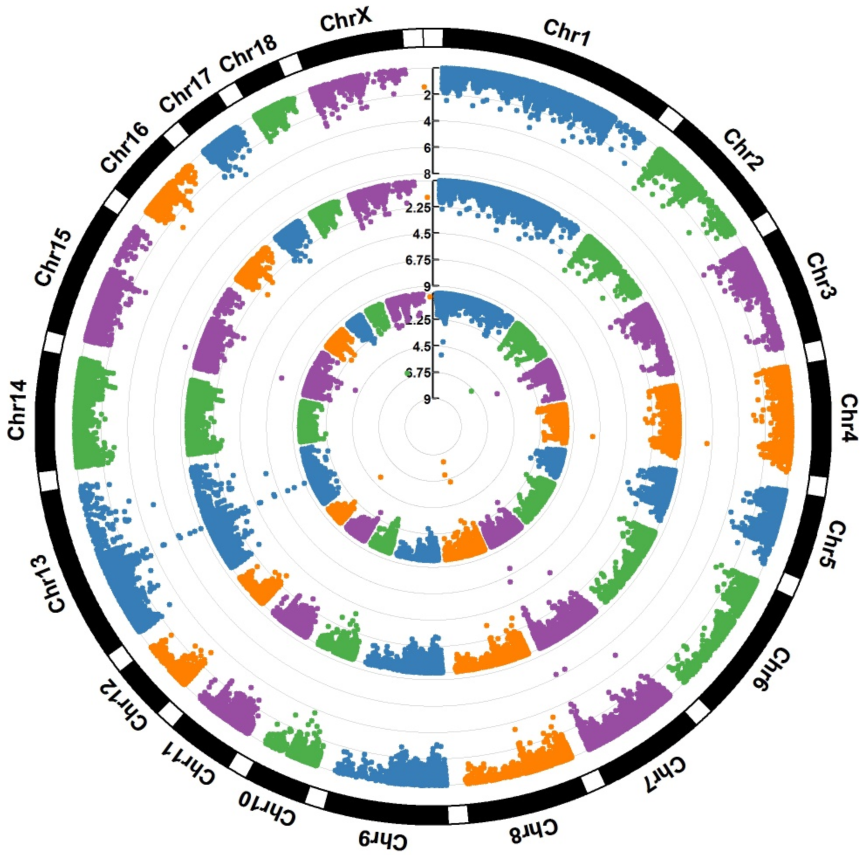
```
R --vanilla --slave --args  filter.vcf.txt  < ../00.src/Fisher_window.R
R --vanilla --slave --args  filter.vcf.txt.fisher_window.txt
< ../00.src/plot_signal.R
```

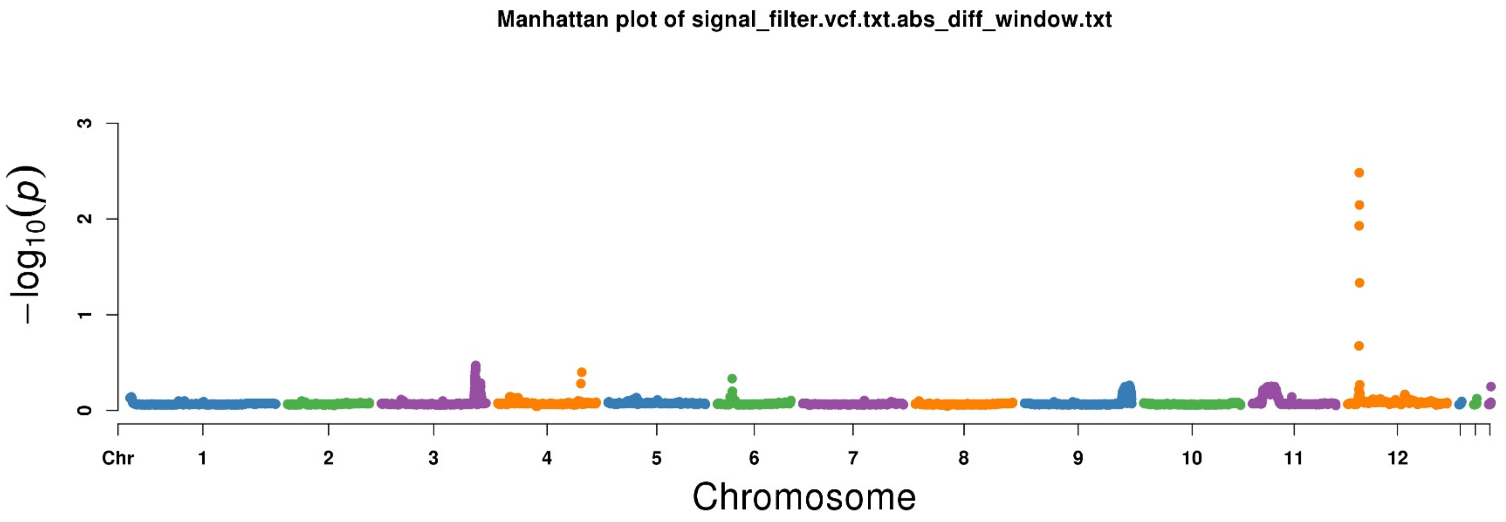


A few notes of the R script:

- Window size in the script is 1 Mbp, steps is 100 kpb
- Only considering contigs >1 Mbp
- Chr name can be any characters, with or without “chr”
- You can manually modify the result (filter.vcf.txt.abs_diff_window.txt) to get rid of undesired scaffolds or contigs.

More plotting options





- https://github.com/YinLiLin/R-CMplot

Further reading

MutMap （Abe, A. et al., 2012）
QTL-seq （Takagi, H. et al., 2013）
MutMap+ （R Fekih et al., 2013）
MutMap-Gap （Takagi, H. et al., 2013）
BSR-Seq (Sanzhen, Liu et al., 2013)