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

GWAS or linkage mapping

N=300-3000

Pool the samples with extreme phenotype

n₁=20-50

n₂=20-50

Phenotyping

<table>
<thead>
<tr>
<th>Phenotyping</th>
<th>entire population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotyping</td>
<td>entire population</td>
</tr>
</tbody>
</table>

Phenotyping

<table>
<thead>
<tr>
<th>Phenotyping</th>
<th>entire population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genotyping</td>
<td>two samples</td>
</tr>
</tbody>
</table>
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**

\[ \text{AGCTTGTCGCGCCTTATT} \]

SNP index = 2/6 = 0.33

\[ \text{CAGGTATCGCGCCTGGTT} \]

SNP index = 3/6 = 0.5

**Unlinked sites**

\[ \text{AGCTTGTCGCGCCTTATT} \]

SNP index = 3/6 = 0.5

\[ \text{CAGGTATCGCGCCTGGTT} \]

SNP index = 2/6 = 0.33
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

(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

(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
BSA Pipeline (part 1 variants calling)

- **Reads processing**
  - 1. adaptor trim (optional)
  - 2. quality trim (optional)

- **Alignment processing**
  - 1. Indel realignment (optional)
  - 2. marker duplication (optional)
  - 3. mapping quality

- **Variant calling**
  - 1. Missing rate
  - 2. Expected frequency
  - 3. Rare allele
  - 4. Read depth

- **Pool1**
- **Pool2**
BSA Pipeline (part 2 Statistics and sliding window)

Linked sites

AGCTTGTCGCGCCTTATT

SNP index = 2/6 = 0.33

AGCTTGTCGCGCCTTATT

SNP index = 6/6 = 1

\[ \triangle \text{SNP index} = \text{abs}(1 - 0.33) = 0.67 \]

Window size

\[ (0.01 + 0.50 + 0.48) = 0.33 \]

\[ (0.50 + 0.48 + 0.55) = 0.51 \]

\[ (0.55 + 0.07) = 0.32 \]

\[ (0.07 + 0.03 + 0.05) = 0.05 \]

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.

<table>
<thead>
<tr>
<th>SNP index</th>
<th>Linked sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/6 = 0.33</td>
<td>AGCTTGTCGCCGCTTTATT</td>
</tr>
<tr>
<td>6/6 = 1</td>
<td>AGCTTGTCGCCGCTTTATT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Ref allele</th>
<th>Alt allele</th>
<th>Row total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Mutant</td>
<td>0</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Column total</td>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

\[
p = \frac{\binom{6}{4}\binom{6}{6}}{\binom{12}{4}}
\]

\[
F = \text{fisher.test(rbind(c(4,2),c(0,6))},
\text{alternative=“two.sided”})
\]

\[
F$p.value = 0.06061
\]
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
  
  **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

```bash
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
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

```bash
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/AD F,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**

```plaintext
#bcftools_viewVersion=1.8+htslib-1.8
#bcftools_viewCommand=view -m2 -M2 -0 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

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

AD -- 15: 16
Step 3: Filtering the variances

```bash
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.
FORMA/T/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

```bash
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
## Final result in txt format -- filter.vcf.txt

<table>
<thead>
<tr>
<th>Chr</th>
<th>Pos</th>
<th>Ref</th>
<th>Alt</th>
<th>total DP</th>
<th>Mut_ref</th>
<th>Mut_alt</th>
<th>WT_ref</th>
<th>WT_alt</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>357</td>
<td>A</td>
<td>C</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>539</td>
<td>A</td>
<td>C</td>
<td>13</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>762</td>
<td>C</td>
<td>T</td>
<td>11</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>860</td>
<td>C</td>
<td>T</td>
<td>35</td>
<td>15</td>
<td>1</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>906</td>
<td>G</td>
<td>T</td>
<td>41</td>
<td>19</td>
<td>1</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>949</td>
<td>T</td>
<td>A</td>
<td>42</td>
<td>22</td>
<td>1</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1369</td>
<td>A</td>
<td>C</td>
<td>25</td>
<td>11</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1449</td>
<td>A</td>
<td>C</td>
<td>29</td>
<td>11</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1454</td>
<td>C</td>
<td>A</td>
<td>30</td>
<td>15</td>
<td>1</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1485</td>
<td>T</td>
<td>G</td>
<td>28</td>
<td>7</td>
<td>2</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1488</td>
<td>T</td>
<td>G</td>
<td>27</td>
<td>9</td>
<td>1</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1524</td>
<td>T</td>
<td>C</td>
<td>27</td>
<td>8</td>
<td>1</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>
The running log

INFO 2018-11-19 13:10:40 MarkDuplicates After output close freeMemory: 13338438088; totalMemory: 13466861568; maxMemory: 19088801792
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
总和 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.bam
-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

cmd 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
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.
```

```
-O z -o filter2.vcf.gz
```
Summary statistics 1. △SNP index

Linked sites

AGCTTGTCGCGCCTTATT

SNP index = 2/6 = 0.33

SNP index = 6/6 = 1

△SNP index = abs(1-0.33) = 0.67

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
Summary statistics 2. ratio of allele frequency

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

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

\[
\text{Linked sites}
\]

\[
\text{SNP index} = 2/6 = 0.33
\]

\[
\text{SNP index} = 6/6 = 1
\]

\[
\text{R --vanilla --slave --args filter.vcf.txt < ../00.src/Ratio_window.R}
\]

\[
\text{R --vanilla --slave --args filter.vcf.txt.ratio_window.txt < ../00.src/plot_signal.R}
\]

\[
\text{Manhattan plot of signal_filter.vcf.txt.ratio_window.txt.jpg}
\]
Summary statistics 3. fishier exact test

Linked sites

AGCTTGTCGC

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

Manhattan plot of signal_filter.vcf.txt.fisher_window.txt.jpg
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)